

Northumbria Research Link

Citation: Kinghorn, Phil (2017) Deep Learning-based Regional Image Caption Generation with Refined Descriptions. Doctoral thesis, Northumbria University.

This version was downloaded from Northumbria Research Link:
<https://nrl.northumbria.ac.uk/id/eprint/38078/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

Northumbria Research Link

Citation: Kinghorn, Phil (2017) Deep Learning-based Regional Image Caption Generation with Refined Descriptions. Doctoral thesis, Northumbria University.

This version was downloaded from Northumbria Research Link:
<http://nrl.northumbria.ac.uk/id/eprint/38078/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>



Deep Learning-based Regional Image Caption Generation with Refined Descriptions

P Kinghorn

PhD

2017

Deep Learning-based Regional Image Caption Generation with Refined Descriptions

Philip Kinghorn

A thesis submitted to the
University of Northumbria at Newcastle
For the degree of
Doctor of Philosophy

Department of Computer Science and Digital Technologies,
Faculty of Engineering and Environment,
Date: 06/10/17

Acknowledgements

There have been many people who have helped and supported me throughout my PhD studies. I would especially like to thank Molly Elliot for her unending support and encouragement, helping me maintain my focus and always being there when I've needed her most over the past many years. I would like to thank my principal supervisor Dr Li Zhang, for her guidance, help and proof reading throughout my research, my papers and this thesis. I would also like to thank Prof. Ling Shao, for his helpful directions and advice during his time at Northumbria University. I must also thank my family, again for all their help and support not just throughout my studies but the past 25 years.

I would like to thank my fellow researchers within the lab at Northumbria University. Especially, Ben Fielding, Kamlesh Mistry and Diptangshu Pandit for all their help over the past three years of study. Finally, I would also like to thank John Graham and Kerry Veitch for all their time spent proof reading and their support and kindness.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee on 12/2015.

I declare that the Word Count of this thesis is 39,327 words.

Name: Philip Kinghorn

Signature:

Abstract

Image captioning in recent research generally focuses upon small, relatively high-level captions. These captions are generally without detail, or insight. Missing out information which we as humans could easily, and would generally, report. This restricts the usefulness of existing systems within real-world applications. Within this thesis, we propose the following solutions to address these problems.

The first stage proposes a region-based approach, focusing upon regions within images and describing them with attributes. These attributes add more meaning to standard classification labels. Improving the classification label, ‘dog’, produced by existing systems, to the more detailed label ‘white spotted dog’. This adds a large degree of detail when used within template-based description generation. The area of healthcare is also explored in which the system is paired with a visual agent. The agent can describe the environment and report potential hazards, as well as socialising through conversation.

The second stage improves upon the previous architecture, by proposing another region-based architecture which removes the rigidity of templates. Instead sentences are generated through a Recurrent Neural Network. Training this architecture on multiple smaller datasets allows for a quicker training stage, with less computing power required during both training and testing. An encoder-decoder structure is proposed to *translate* the detailed region labels into full image descriptions. This produces natural sounding descriptive phrases that accurately depict the contents of an image.

The third stage proposes a hierarchically trained, end-to-end style system to generate an image description with the same required functionality to describe detections in detail but without the need for multiple models. This system can utilise the humanoid robot’s vision and voice synthesis capabilities. Overall, the above proposed systems within this research outperform many state-of-the-art methods for the refined image description generation task, especially with complex and out-of-domain images, such as images of paintings.

List of Publications

1. **Kinghorn, P.**, Zhang, L. & Shao, L. A Hierarchical and Regional Deep Learning Architecture for Image Description Generation. *Pattern Recognition Letters* (2017).
2. **Kinghorn, P.**, Zhang, L. & Shao, L. A region-based image caption generator with refined descriptions. *Neurocomputing* (2017).
3. **Kinghorn, P.**, Zhang, L. & Shao, L. Deep learning based image description generation. in *2017 International Joint Conference on Neural Networks (IJCNN)* 919–926 (IEEE, 2017).
4. Zhang, L., Fielding, B., **Kinghorn, P.** & Mistry, K. A vision enriched intelligent agent with image description generation. in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* 1488–1490 (2016).
5. Fielding, B., **Kinghorn, P.**, Mistry, K. & Zhang, L. An Enhanced Intelligent Agent with Image Description Generation. in *International Conference on Intelligent Virtual Agents* 110–119 (2016).
6. Neoh, S. C., Zhang, L., Mistry, K., Hossain, M. A., Lim, C. P., Aslam, N., & **Kinghorn, P.** Intelligent facial emotion recognition using a layered encoding cascade optimization model. *Applied Soft Computing*, 34, 72-93. (2015)

Contents

Chapter 1 Introduction	1
1.1 Aims and Objectives	1
1.1.1 Descriptive Captions	2
1.1.2 Cross Dataset Functionality	3
1.1.3 Intelligent Agent and Robotics	3
1.2 Motivation	4
1.2.1 Natural or Human Like Descriptions	5
1.2.2 Social Media and Public Images	5
1.2.3 Healthcare	6
1.2.4 Automated Human Description – CCTV	7
1.3 Contributions.....	7
1.4 Thesis Outline	12
Chapter 2 Background.....	15
2.1 Composite methodology.....	16
2.1.1 Pre-Processing.....	16
2.1.2 Region of Interest Extraction	16
2.1.3 Feature Extraction	16
2.1.4 Classification.....	17
2.1.5 Support Vector Machines/Regressor.....	19
2.1.6 Sentence Generation.....	20
2.2 End-to-End	20
2.2.1 Pre-Processing.....	21
2.2.2 CNN Feature Extraction and Image representation	21
2.2.3 Sentence Generation.....	22
2.3 Transfer Learning.....	25
2.4 Attention Based Methodology with a Multi-Layer Perceptron.....	26

2.5 Evaluation	26
2.5.1 Bleu	27
2.5.2 CIDEr	28
2.5.3 Meteor	28
2.5.4 Rouge	28
2.5.5 SPICE	28
2.6 Deep Learning Frameworks	29
2.7 Summary	30
Chapter 3 Related Work	33
3.1.1 Localisation	33
3.1.2 Feature Extraction	35
3.2 Attribute Prediction	36
3.3 Description Generation	38
3.3.1 Composite Image Caption	38
3.3.2 Video Caption	42
3.3.3 End-to-End Image Caption	43
3.4 Image Synthesis from Descriptions	50
3.5 Robot Image Understanding and Integration	50
3.6 Caption/Description Dataset	51
3.6.1 Image	51
3.6.2 Video	53
3.7 Summary	53
Chapter 4 Composite Deep Learning Image Description Generator	57
4.1 Introduction	57
4.2 The Proposed Image Description Generation Framework	60
4.2.1 Object Detection and Classification	60
4.2.2 Scene Classification	61

4.2.3 Hazard Detection.....	62
4.2.4 Fall Detection	63
4.2.5 Attribute Learning	64
4.2.6 Attribute Datasets.....	65
4.2.7 Prepositions	66
4.2.8 Sentence Generation and Construction	66
4.3 Results and Evaluation	68
4.4 Experiment with an Intelligent Visual Agent.....	71
4.5 Summary	72
Chapter 5 Region-based Image Caption Generator with Refined Descriptions	75
5.1 The Proposed Image Description Generation System	77
5.1.1 Object Detection and Recognition	77
5.1.2 Scene Classification	78
5.1.3 Attribute Prediction Using RNNs	78
5.1.4 Sentence Generation.....	81
5.2 Evaluation	83
5.2.1 Evaluation Metrics	83
5.2.2 Evaluation Results.....	84
5.3 Summary	96
Chapter 6 A Hierarchical and Regional Deep Learning Architecture for Image Description Generation	99
6.1 Introduction.....	99
6.2 The Proposed Deep Network for Image Description Generation	102
6.2.1 Model Training.....	102
6.2.2 Model Datasets.....	104
6.2.3 Architecture.....	104
6.2.4 The Deployment to the Robot Platform.....	109

6.3 Evaluation	111
6.3.1 Experimental Results	113
6.4 Summary	114
Chapter 7 Conclusion and Future work	117
7.1 Summary of Contributions.....	117
7.1.1 Composite methodology	118
7.1.2 Refined Descriptions.....	118
7.1.3 Hierarchical Architecture.....	119
7.1.4 Potential Limitations.....	120
7.2 Future Work.....	121
References.....	125

Glossary of Terms

API – Application Programming Interface

RNN – Recurrent Neural Network

LSTM – Long Short Term Memory

GRU – Gated Recurrent Unit

CRF – Conditional Random Field

CNN – Convolutional Neural Network

DNN – Deep Neural Network

RPN – Region Proposal Network

SVO – Subject – Verb - Object

MLP – Multi-Layer Perceptron

SVM – Support Vector Machine

SVR – Support Vector Regressor

NLP – Natural Language Processing

ROI – Region of Interest

LAN – Local Area Network

CPU – Central Processing Unit

GPU – Graphics Processing Unit

HOG – Histogram of Oriented Gradients

SIFT – Scale Invariant Feature Transform

HCI – Human Computer Interaction

Chapter 1 Introduction

Describing the contents of images is a relatively easy task for humans, and can be achieved with reasonable accuracy from a very young age. Humans can recognize, distinguish and describe many objects, scene or image categories with many external factors, such as occlusions, changes in illumination and pose that make this task incredibly difficult for machine learning algorithms to succeed (Oliva and Torralba, 2007). Computer vision algorithms have struggled to match this capability.

Although recent research indicates that these algorithms have become incredibly powerful in certain vision domains such as image classification and segmentation, producing detailed descriptive sentences based on the contents of an image is still a challenging task. Applying such a task to robotic vision research may also enable humanoid robots to conduct personalized interaction with humans and understand their surroundings, potentially automating care environments.

Each of these areas has recently been an active research area, and has seen a large increase in performance and results due to the resurgence of powerful deep learning algorithms, in particular Convolutional Neural Networks (CNN) (LeCun *et al.*, 1999).

In the domain directly linked to this research, image captions have been widely produced, with systems that can generate paragraphs based on sentences and topics. However, giving a system or model the ability to accurately portray and describe the contents of an image is lacking and is a very active research domain and a difficult challenge for computer vision.

When we, as humans look at an image, we can look at objects, people, scenes and their relationships and describe them in fluid detail. For a machine to achieve similar results, they too should be able to detect objects and people and then describe them to a similar degree.

1.1 Aims and Objectives

The main objective of the research conducted is to propose and improve upon existing image description frameworks, with the ability to automatically classify and describe an image and all of its contents, and improve upon the short captions that are commonplace within this domain.

Introduction

To work towards this overarching aim the following objectives must be explored.

1. Image feature extraction – Experimenting and exploring techniques ranging from low level features, such as edges and colours, to deep learned features from CNN sliding windows.
2. Multiple asset model training – Train respective models and techniques on the relevant datasets for:
 - a. Object detection and classification
 - b. Person detection
 - c. Object and Person attribute prediction
3. Caption/Description training – Train sentence, word or character level RNN predictors for whole system pipeline.
4. Model Construction – Merge, build and manage full system pipeline in order for detailed caption generation to take place on a wide variety of image scenarios and styles.

The secondary objective is to allow such a system to be utilised on any image, creating a system to perform cross dataset improves upon its real world usability, thus allowing the system to be extended to run with devices and applications utilising intelligent agents and robotics.

Existing research has focussed on generating high-level, short captions, with minimal added detail. In order to meet the objective, we propose multiple novel frameworks and methods which explore both composite and end to end style architectures within multiple areas of computer vision and Machine Learning research. The proposed architectures and their internal structures require multiple domains, such as object detection and classification, attribute and scene prediction among others. This allows details and objects that may previously be overlooked to be reported and described. This research, without crossing into these multiple domains, severely limits the overall capability and usability, which could effectively pigeon-hole the proposed frameworks into a specific domain, which is against the aims of this thesis and the research within.

1.1.1 Descriptive Captions

When humans describe an object or a person, they would very rarely simply state one label. E.g. ‘Cup’ ‘Man’ or ‘Car’, instead they can add their attributes almost without thought. This diversifies these labels and creates a very descriptive and vivid piece of

Introduction

information. E.g. ‘Shiny red car, with silver wheels’, manufacturer information and its environment can also be very easily added.

One of the preliminary stages within the methodology is to accurately locate, classify and describe objects and people. Typical systems within this area simply look at the overall image or focus on one Region of Interest (ROI), through the use of an attention mechanism and collected features. Through in-depth detection of multiple regions, we can provide a more complex description of a whole image, rather than one specific section or region, which removes bias from the frameworks and models and produces a more general image description.

1.1.2 Cross Dataset Functionality

There are several disadvantages with typical machine learning systems that are trained upon one dataset. Upon testing the system, the images and/or data tested is generally from the same domain. This could effectively force a system to only function well within that domain. Proposing architectures and models that are hierarchically trained and tested upon different datasets allows the system to describe any image passed through it, with a higher degree of accuracy. Such a system could find applications in more varied areas in the wild or within active research. Similar in effect to transfer learning, in that its domain or dataset is not fixed, however not yet to the degree in which for example, training a system on real world images and testing it on oil paintings.

1.1.3 Intelligent Agent and Robotics

Adding the ability for image description to robots and intelligent agents allows for a fluid and free moving camera that can capture its scenes regularly, analyse them quickly, and interpret and understand the scene and then act or converse appropriately.

Humanoid robots may not yet be commonplace in homes, however exploring and proposing frameworks that can run on smaller readily available counterparts, such as the NAO and Pepper from Aldebaran Robotics (Aldebaran, 2017), show their incredible potential use within the home in the near future.

Intelligent agents such as Siri and Google Now offer conversation and helpful tools with the use of collected data direct from a mobile phone, such as location or device information (Gyroscope, compass etc.) as well as user entered data, such as personal information. Images have been explored in the context as an aid to product searching

Introduction

on the web, e.g. providing or capturing an image of an item, and using the system to search until it locates a website that sells that product. Adding real world images or analysis could allow this to be used on the fly in videos or from microphones that can record natural descriptions.

1.2 Motivation

In recent years, many methods have been proposed for image description generation. However, the majority of the related research relies upon holistic approaches for image understanding and entity recognition, which may lose details relating to important aspects or regions of an image.

Automatic description of images can enable smoother internet navigation for those with visual impairments. As computing hardware becomes more powerful, this technology can become more mainstream allowing the systems, software and models to become much more accessible and dramatically more portable. Systems such as this allow an increased degree of freedom for visually impaired users as well as potentially reducing work load within healthcare and industry.

This ability could allow for the expansion of applications within robotics, as well as healthcare. These domains are explained in further detail in the following sections.

Most of the existing research within the image captioning domain utilise a Convolutional Neural Network (CNN) – Recurrent Neural Network (RNN) style architecture. This works well in most circumstances, especially with the aim of producing short and high-level captions. In this methodology CNNs extract image features from across the entirety of the image, as explained in Chapter 2. These form a very large feature vector, which consists of anywhere between 128 and 4096 dimensions, meaning typical Machine Learned methods cannot cope with the scale or would simply be too slow to be effective, especially in real time or real world scenarios. This paired with the RNN which learns the sequences and patterns within image captions, generates these sentence captions by utilising the large feature vector and the individual generation of each word, this leads to short, relatively nondescript captions.

Introduction

To this end we propose models in which feature vectors are collected for many individual regions within an image. This forms the core of the region-based approach that gives the flexibility for many aspects such as the cross domain, as well as the increased descriptive capability.

1.2.1 Natural or Human Like Descriptions

Image descriptions have generally been referred to and act as image captioning, commonly producing captions of ~10 words in length (Vinyals *et al.*, 2014). This does not provide sufficient word count for much, if any detailing information to be present, unless the tested image is incredibly simple, unlike real world images. This provides the motivation to increase upon the existing research without the wait or need for a complex dataset, which could drain existing resources. Existing research will simply acknowledge the presence of objects and/or people with minimal added detail, providing the motivation in this research to increase upon this fact and describe, in as much detail as humans would within natural conversation.

Existing methods do not provide enough detail and in many cases, do not emulate human judgements. In dataset construction humans are asked to annotate images but this can be extremely short and simple in an effort to reduce the overall difficulty of the problem.

1.2.2 Social Media and Public Images

With the ever-increasing wealth of publicly available images online, due to mainstream and social media, images can be paired with irrelevant or unnecessary user data. User-entered captions on social media could be entirely unrelated to the image content, referring to friends or family that are not present, therefore reducing the effectiveness of those users who rely on screen reading technologies. Recently datasets have been proposed utilising these paired image data (Xiao *et al.*, 2010; Chen *et al.*, 2015; Torabi *et al.*, 2015; Elliott *et al.*, 2016). However, they require an initial cleaning process (Verma and Jawahar, 2011) in order to maintain the desired standard of paired images and captions. This process can be time and resource consuming and difficult, as human annotators can make mistakes over the large number of images, which could sacrifice data validity in the long run.

Automating this process by utilising techniques outlined within this thesis could speed up development and dataset production, while ideally reducing the amount of errors

Introduction

to ensure data precision and integrity. Such a system could accurately annotate, classify and describe these images automatically, allowing screen readers etc. to provide a better experience for those with visual impairments.

Companies that collect or utilise large numbers of images could have their images automatically annotated and described making it easier for natural search queries to locate them or even to describe people (S. Li *et al.*, 2017). Rather than a basic and vague label, users could enter a search query in their natural description. This could return images more accurately based on their request and narrow down the search quickly, yet very effectively, dramatically increasing productivity for the end user.

1.2.3 Healthcare

Adding image description and object attribute prediction functionality within a care providing environment would allow for detailed information regarding changes to living or communal spaces to be automatically collected and verified, especially when paired with aspects from robotics such as automated or remote cameras as well as humanoid robots such as Pepper (Aldebaran, 2017). This could detect and alert users to hazards and potentially even falls depending upon the training and implementation of the system framework.

This, paired with some of the robotics previously mentioned, allows for a free moving camera that can allow an implemented system to understand its environment, by detecting and describing objects within the camera's vision to those within the vicinity of the robot and care providers who can monitor, oversee and ideally prevent any issues.

There is also the opportunity to pair such a system with visual agents, this could allow for a conversation and socialising aspect to be integrated as well as all of the previously discussed benefits. This has the added benefit of being accessible from readily available hardware, such as laptops or tablets, without the need for a care provider or consumer to invest in automation and robotics, such as the previously mentioned Pepper from Aldebaran Robotics.

The previous motivation regarding the widespread use of holistic features can limit these systems within the healthcare domain. If the training data and training process allowed, these systems could in theory be used to describe images such as X-rays and

Introduction

MRIs. Extracted regional information could pinpoint fractures and or breaks better than typical holistic collected features.

1.2.4 Automated Human Description – CCTV

Person description is another avenue of research explored during this thesis. One of the benefits of the proposed frameworks allows such an image captioning and descriptive system to be easily adapted to the description and annotation of a person. A system that could accurately and automatically describe a person in an image, could eventually reduce personnel work load in the relevant domains.

This thesis provides an initial exploration into this ability. Within the Chapter 6 a multistage end-to-end style system is trained in which only one branch of the system describes people, however the attribute prediction covers a wide range of user attributes, and if needed could be retrained or fine-tuned on a specific set, for a specific purpose. Tailoring this aspect of the system for this application could be beneficial for person re-identification tasks such as identifying people on no-fly lists or wanted for convictions so that human resources could be spent on more pressing or urgent matters. This combined with other aspects such as social media could allow for automated posting to alert the general public if required.

1.3 Contributions

Multiple algorithms, structures and architectures have been proposed with the main intention to improve upon the descriptive nature of existing image caption systems. The research conducted has generally focused upon adding regional information and a large vocabulary of attributes, increasing the descriptiveness and accuracy of the automatically generated captions to be more human like. Multiple standard large-scale image datasets such as ImageNet (Russakovsky *et al.*, 2015) for the required subdomains have been used, in tandem with large scale image-description paired databases, such as Microsoft MSCOCO (T.-Y. Lin *et al.*, 2015).

To address these issues, in this research we propose multiple methods and approaches that have been combined to achieve the aim. This has also lead to the exploration of multiple active research areas to increase the descriptive nature of image captioning frameworks.

Introduction

1. The first contribution presented within this thesis is a composite based image captioning framework utilising individual image regions and additional attribute information.

This research utilizes local image regions to initialize image descriptions rather than focus on single stage holistic features of an input test image, therefore image features are extracted from each region. The system contains the core functions of scene classification, object detection and classification, attribute learning, relationship detection and sentence generation. We have also further investigated the applicability of our systems on the fall detection and hazard identification research domains. This is paired with a 3-D intelligent agent for use within care environments.

The proposed system utilises the powerful CNN feature extractors on multiple locations rather than the whole image. This has the added benefit of capturing a dramatically larger area within the image that the remainder of the system pipeline can annotate. This produces many more features which in turn allows for more detail. For example, if the feature extractor collects 4096 features per image, we collect 4096 features for every viable image region. The methodology for determining region viability changes throughout this research however ranges from a Selective Search (Sande, 2011) algorithm to a dedicated Region Proposal Network (Ren *et al.*, 2015).

This possesses a trade-off in which the caption detail, i.e. the aim of this research, is achieved with the sacrifice of execution speed in comparison to some other models and frameworks.

The work also adds attributes to increase the descriptive nature of captions. Our initial work explored the combination of the required research areas and proposed an architecture that is based around local regions, extracting regional features rather than holistic features typically used in existing research. This allowed classified labels and their associated attributes to be placed into a sentence generation methodology which can cover a diverse range of sentence or descriptive outputs.

We propose multiple system frameworks that explore sentence generation from these newly created descriptive labels. The initial research stage involves templates, substituting temporary values for the generated content. The follow up method removes the need for fixed templates, and instead incorporates Recurrent Neural

Introduction

Networks to generate the improved and refined attributes leading to the improved final descriptions.

This system has been reported in Chapter 4 and presented in Kinghorn, Zhang and Shao, (2017c)

2. The second major contribution presented within this research is a fully supervised deep learning framework which combines the regional and attribute aspects of the previous system but improves upon the sentence construction by proposing the use of RNNs to the generated regional information into more refined and useful descriptions.

This proposes a method to translate regional information to Image descriptions. Our proposed architectures can describe image regions in detail, by describing people, their clothes, gender and hair styles, as well as the objects, size colour and shape etc. This research utilises RNNs for attribute prediction rather than the SVMs, due to some limitations within the SVM implementations that are explored later in this thesis, and other methods commonly found within those research domains. RNNs are incredibly useful at generating sequential text data, which inspired their use for attribute prediction. Due to the nature in which they generate based on features and previous generated words, RNNs in theory should stop generating when no other attributes describe the region. The same principle ensures that the generated attributes should not contradict. For example, for a given region both attributes ‘Male’ and ‘Female’ should not be present, therefore it should technically be impossible for these two to be generated for the same region, there is no safeguard against this process with SVMs other than post-processing.

When SVMs are used in this domain, one SVM is required per attribute. When systems require many attributes, this SVM methodology quickly scales, resulting in many SVMs all processing the same feature vector. At run time, all SVMs are tested utilising the same feature vector, consuming resources and, in many cases not being entirely accurate. RNNs allow for one test scenario with one set of CNN extracted features and allow for a multi-word label or caption of a region to be generated. This can be treated as a descriptive region label in our overall system framework.

Introduction

The combination of attributes and generated object labels, that have been classified with the use of trained fully connected layers (fc) of a CNN, produces a multi-word descriptive label, which is useful in many areas that have previously been mentioned.

This stage enables us to treat the image captioning problem as a machine translation problem, effectively translating an image into a descriptive caption through the detailed region. We propose the utilisation of an encoder-decoder description generator embedded with two RNNs to produce refined and detailed descriptions of a given image. This combines the multi label captions, and concatenates them to form a source sentence, consisting of nouns and adjectives. The encoder network then encodes the entire sentence based on the vocabulary into a fixed feature vector of a given length. This feature vector is immediately decoded into a fully descriptive sentence, inferring any detail or words that are missing from the initial label. This can also be viewed as the combined region labels being a source language, and the description being the target, allowing the system to infer the sentence structure and missing words based on those already generated by the system, as well as those present within the encoder-decoder training procedure.

This system is explained in further detail in Chapter 5 and presented in Kinghorn, Zhang and Shao, (2017b).

3. The third and final major contribution of this research proposes a deep neural network architecture utilising a hierarchical training process in order to streamline the overall system pipeline. It has also been deployed and integrated with the vision API of a humanoid robot to indicate its effectiveness in real-life settings.

A novel deep architecture for image region annotation is proposed as the third research contribution. It generates not only regional annotations but also integrates the regional captioning into full image descriptions. The proposed deep networks within this thesis have a more efficient training process and show great robustness and efficiency in dealing with out-of-domain images.

A unique model architecture is proposed to contribute to the solving of this description problem. With the aim of producing a model that can test an image in one pass, while contributing detailed regional information and the previously mentioned machine translation, we propose a hierarchically trained deep neural network. The initial stage

Introduction

utilises the RPN (Ren *et al.*, 2015) to deliver accurate and near cost-free ROIs that the designed model can then use for the relevant labelling and description. The second stage uses a Recurrent Neural Network (RNN)-based encoder-decoder structure to translate these regional descriptions into a full image description. Especially, the proposed deep network model can label scenes, objects, human and object attributes, simultaneously, which is achieved through multiple individually trained RNNs.

The model consists of multiple branches trained on individual datasets. This allows individual branches that specialise on a specific intended output. For example, training certain sections on a scene dataset produces scene labels simultaneously, while another section that is trained on objects produces the corresponding object labels. This process freezes and restricts parts of the network to prevent undesired training. As previously mentioned, this also is more efficient, as it could require much less training data per branch than if the whole network had to be trained in the typical end-to-end fashion. This would require many more samples, each having every single required piece of information annotated in the correct format and accurately.

Image captioning is treated as a machine translation problem rather than a captioning problem as previously described, with the aim of more natural and human sounding outputs. This is achieved with the use of an encoding and decoding RNN.

Many of the above proposed models and systems have been deployed where possible to a robotic platform, utilising photo and voice functionality. A humanoid robot and its application in this domain are proposed. The robot in question is the NAO H25 humanoid robot, equipped with stereo vision, speech synthesis and speech recognition allowing the robot to audibly describe and react to its surroundings. The robot utilisation is based upon LAN connectivity with a powerful GPU workstation. Multiple experiments were conducted on the deployed system and the robot application. The first experiment is conducted on a large image description dataset, the IAPR-TC 12.

This system is reported in Chapter 6 and is presented in Kinghorn, Zhang and Shao, (2017a)

4. Cross-domain deployment and evaluation

Introduction

The final contribution of this research is the ability for these systems to be able to handle and describe a wide range of images, not just those present in the test sets of the datasets within the given training domain, but images collected from other datasets and images collected from in the wild. This is most notable in Chapter 5, when the proposed system is tested on oil paintings of dramatically different styles, still producing accurate descriptions with none to very minor alterations within the training process.

1.4 Thesis Outline

Chapter 1 focuses on the introduction and origins of image captioning and description and the composite areas explored in the proposed methods order to achieve the main objective. For example, object detection, localisation and classification, scene classification as well as person attribute labelling.

A brief overview and background to the research covered and the multiple research areas are covered in Chapter 2.

Chapter 3 discusses and outlines current research within the prominent areas of the research explored, as well as reviewing the state-of-the-art methodology and existing architectures. As the methodology explored covers a diverse range of techniques, this chapter is separated into multiple sections.

In Chapter 4, the initial stage of the proposed research utilising a region-based system with template-based sentence generation description is presented. A composite method is proposed to accommodate all of the required and generated/predicted information and create a system that can be utilised for benefits within a healthcare application. Aspects of Human Computer Interaction (HCI) are also explored.

Chapter 5 consists of the next stage in the research in which machine translation aspects are explored to generate the required descriptions with the intended level of detail, as well as more efficient and accurate attribute prediction models. Aspects of transfer learning are also explored. This chapter directly extends the research conducted in the previous chapter.

End-to-end style model architectures are explored in Chapter 6, presenting a unique model structure requiring less computation resources and training time. This model involved multiple training stages and processes which are further discussed, as well as

Introduction

a thorough evaluation utilising the commonly used metrics available from the MSCOCO (T.-Y. Lin *et al.*, 2015) evaluation scripts.

Finally, chapter 7 summarizes and concludes the thesis with the contributions and methodologies explored. This chapter also expresses some possible paths for future research within the image captioning/description research domain.

Chapter 2 Background

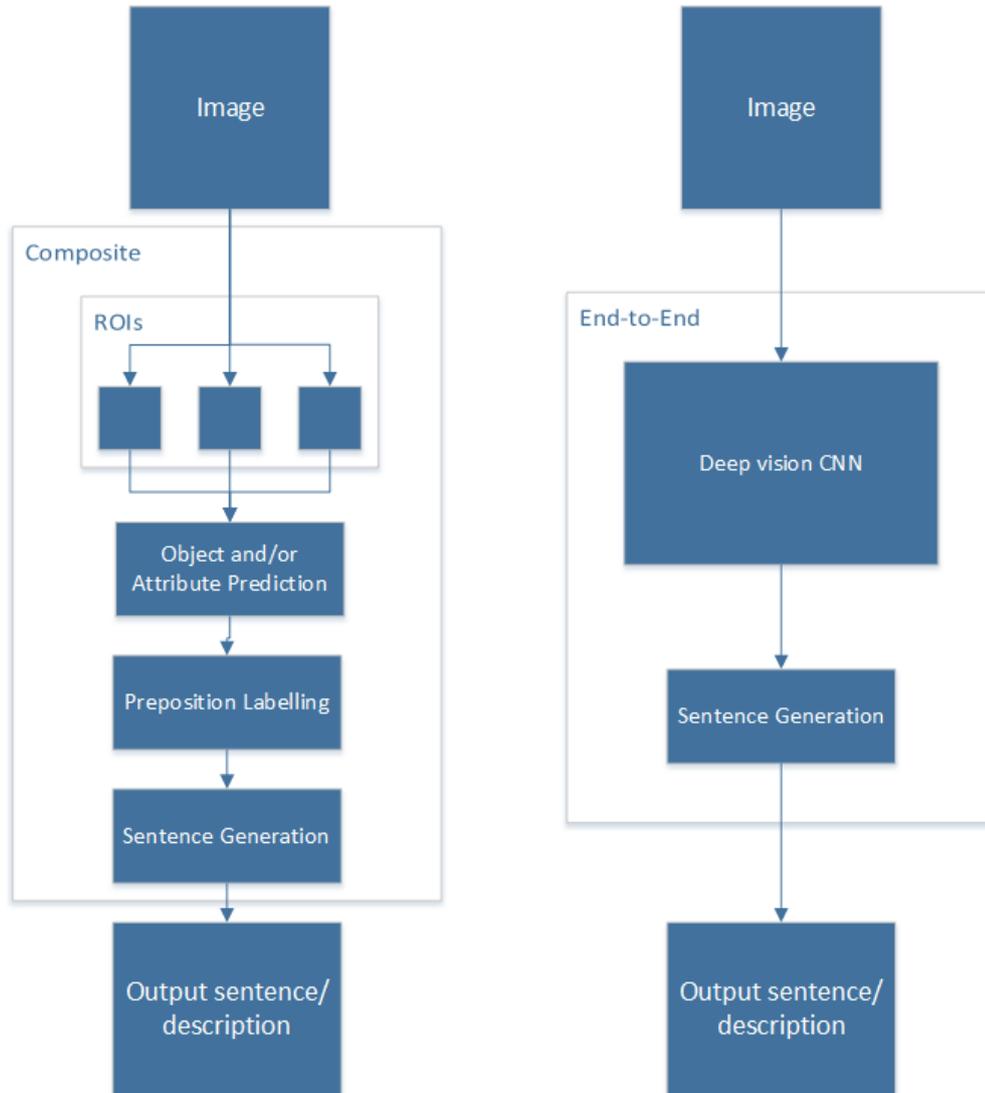


Figure 2.1 Example comparison between Composite and End-to-End style image caption architectures

Image captioning/description has been an active research area for some time, recently gaining traction from the deep learning community. Image captioning is the direct result of a computer system that can receive an input image, then correctly and automatically annotate and describe its contents. Image captioning can therefore cover many existing computer vision domains in order to fulfil this criterion. For example, for these systems to recognize and locate objects, object classification must be explored, and for natural sounding output descriptions, utilising techniques and methods within the field of Natural Language Processing (NLP) becomes necessary.

Background

The main concepts used throughout the research are CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network). General research background on these methodologies are presented throughout this chapter.

There are two commonly implemented architectures used within the field of image captioning/description, composite and end-to-end. High level examples of these architectures are shown in Figure 2.1. The initial sections of this chapter explore the features and stages of each of these architectures and their position within the active research environment.

2.1 Composite methodology

The composite framework for image captioning generally produces longer descriptions. This framework generally consists of multiple separate stages that are run procedurally, one stage after the other.

2.1.1 Pre-Processing

As previously stated, a composite framework consists of multiple separate stages. These stages can differ dramatically depending upon the final aim of the system. In this background section, we outline some of the initial and pre-processing stages that the image undergoes in order to deliver the desired output sentence or description.

2.1.2 Region of Interest Extraction

Methods such as object localisation and recognition, as well as many composite image caption frameworks depend upon ROI extraction. This process involves searching over an image, and producing bounding boxes or regions that have the possibility to contain something interesting. These are boxes where the features or pixel values are generally more varied and/or different enough from the rest of the image. The most common methods of generating these regions for further processing in the pipeline are the Selective Search (Sande, 2011) algorithm, Edge-Boxes, and the newest methods, the RPN (Ren *et al.*, 2015). These are further discussed in detail in the following chapter.

2.1.3 Feature Extraction

In early computer vision research, a number of existing techniques allow for feature extraction, these generally differ from the CNN extraction commonly implemented in

Background

recent works. Some of the most commonly used low level features that have been utilised in this and related work are briefly discussed.

Variations of the HOG descriptor are used for textual descriptors, and canny edge detectors are used for edge detection. SIFT features are extracted to allow for scale invariant feature use. Each of these techniques has strengths and weaknesses for certain types of classification. For example, texture descriptors can be used exclusively or in tandem with other features for attributes that relate specifically to texture, such as metallic or spotted. CNN feature extraction could/can also be used, this technique is explained further in section 2.2.2

2.1.4 Classification

A Multilayer Perceptron or more traditionally named Artificial Neural Networks (ANN) have been widely used for classification problems (Chaudhuri and Bhattacharya, 2000; Orhan, Hekim and Ozer, 2011; Talukdar and Mehta, 2018) Multilayer perceptrons were created in an attempt to overcome the limitations of earlier perceptrons in which the classes were not linearly separable meaning the models were unable to classify correctly or accurately (Kotsiantis, 2007). Neural networks generally consist of a number of nodes or neurons connected to each other. The layers of these neurons are typically separated into three categories forming the structure, typically these neural networks consist of one input layer of multiple neurons, one or more hidden layers of multiple neurons and an output layer consisting of a number of neurons equal to the number of classes (High level representation shown in Figure 2.2).

To train these neural networks these neurons will be equipped with initially random values or multipliers that the input values will pass over before an output is given (Hemmat Esfe *et al.*, 2015), this is the feedforward part of the neural network model. To adjust the weights and the values within the network the error is calculated based on the target output and the actual output, and during backpropogation updates them based on the comparison of target to actual outputs. How these values are updated based on these values is dependant upon the activation function set before training begins, typically these are tangent sigmoid for hidden layers, and linear for the output layers. Training is ceased either upon an exeptable error rate upon evaluation. I.e.

Background

Mean squared error (MSE) or upon reaching a set number of iterations over the training set (epochs).

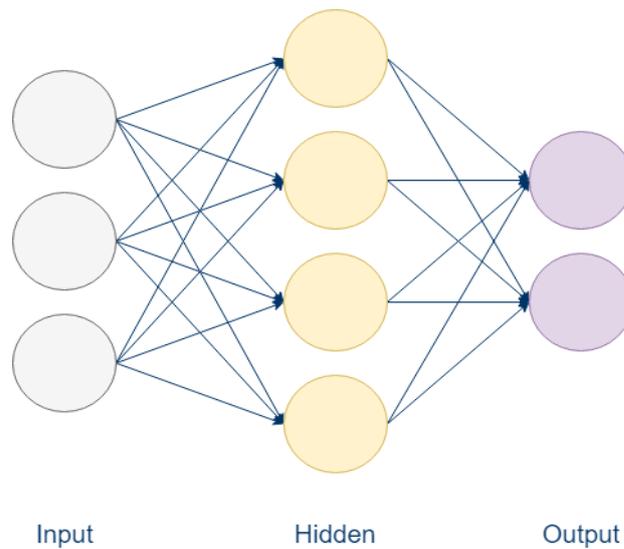


Figure 2.2. A high level representation of the neurons within an Artificial Neural Network.

2.1.4.1 Objects

Composite methods for object classification are typically different to conventional and recent object classification techniques such as fully connected convolutional networks. One of the early, successful object detectors trained 200 SVMs, one for each possible object label (Girshick *et al.*, 2016).

2.1.4.2 Attributes

Where attributes are present in composite methods they have generally utilised SVM classifiers trained on low level extracted features as previously discussed. One of the earliest works into object attribute prediction explored selecting features that were discriminable for a specific attribute (Farhadi *et al.*, 2009). For example, to train a specific detector, the features which distinguish between those with and without that attribute are used, not all of the available/collected features.

Attributes add a substantial amount of information to an otherwise basic label. For example, tying a colour, texture, or style as a pre-requisite much more information about that object has been given. When attempting to deliver upon the research aim of increasing the descriptive nature of these captions, this is of high priority to ensure that these labels are paired with the correct and detailed label attributes.

2.1.5 Support Vector Machines/Regressor

There have also been some more traditional tools used in image captioning in order to increase the accuracy of these deep learned methodologies. In addition to SVMs having their use within attribute prediction, as described in the previous section, they have been utilised by (S. Liang *et al.*, 2017) to determine which differently trained model is used to infer for a given test case image. Combining a relatively cheap machine learning model to enhance the capability of deep models allows for added confidence measures from the deep model and produces a more relevant caption.

SVMs (Cortes and Vapnik, 1995) are based on the concept of mapping multiple input vectors to a very high dimensional feature space, and in terms of classification the aim is to identify the plane between multiple classes (Meyer and Technikum Wein, 2001). For linearly separable SVMs this process can be seen within figure 2.3.

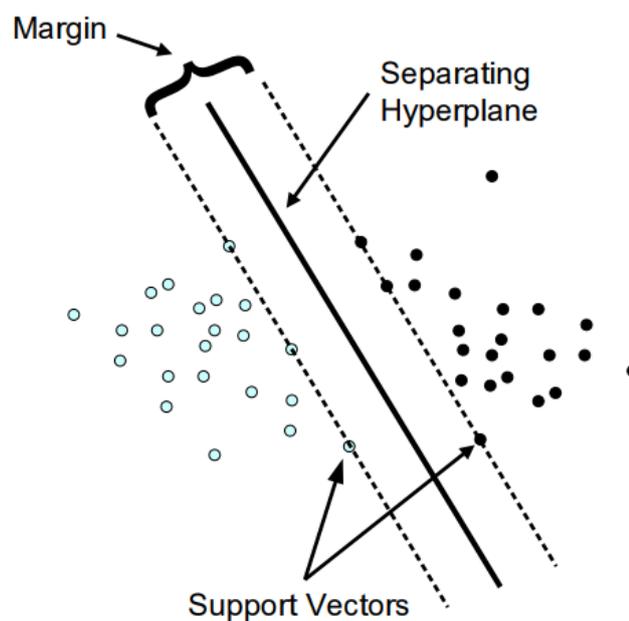


Figure 2.3. Linear separable data and the separating hyperplane produced based on the supporting edge of class vectors (Meyer, 2007).

Support Vector Machines produce their classification result as either a 0 or a 1. Which can translate to a query being part of the tested class or not, it is entirely binary with no immediately available confidence measure. This limitation had been addressed by (Drucker *et al.*, 1997) in which a regression technique based on “*Vapniks concept of support vectors*” outlined in the original paper is implemented. These machines produce a range of outputs between 0-1 providing a confidence score and a more usable output depending upon the scenario. SVRs have been implemented for a

Background

number of active domains such as sales forecasting (Yu, Qi and Zhao, 2013), user gaze tracking (Lu *et al.*, 2010a) and stock market prediction (Yang, Chan and King, 2002).

2.1.6 Sentence Generation

One of the earliest examples of sentence construction within image description frameworks is that of Li *et al.*, (2011) among other methods, templates are stated to be a popular and effective method when adjectives, nouns and prepositions are present. For example, two objects each with an adjective, and a preposition can construct a sentence such as “There is a [adjective] [object] [preposition] [adjective] [object]” producing something like “There is a green apple on top of the wooden table” (Xu *et al.*, 2015).

Templates provide a fast prototyping option when other aspects of the system are already in use, the system can be set up with very little compute cost and time. However, the sentences can all sound very similar, they would all be structured identically which could lead to grammatical errors, and over a large number of images potentially miss a large amount of detail. There have been numerous efforts in recent research to overcome some of these limitations of templates by combining them with other approaches, these are discussed in more detail in the following chapter.

Another approach to sentence generation has been sentence/caption retrieval. This as the name suggests retrieves a sentence from a large corpus that the system believes is the closest matching sentence to the image, rather than generating or constructing a sentence from previous or existing detections. This has the benefit of producing a grammatically correct sentence every time, however this method could lead to the possibility of producing two identical descriptions of images whose content differ substantially.

2.2 End-to-End

The simplified high-level architecture of this style of image caption system consists of a very deep convolutional neural network (LeCun *et al.*, 1999) that acts as the feature extractor paired with a sentence/description generation algorithm such as a Recurrent Neural Network (Lipton, Berkowitz and Elkan, 2015). These models are trained in an end-to-end fashion on paired image-caption datasets such as Flickr8k/30k and MSCOCO, further details on these systems are presented in Chapter 3.

Background

2.2.1 Pre-Processing

Within these models, data-loaders and image pre-processing, where images are resized and scaled, is required in order for the initial CNN to accept and run the selected image. In general, many of the existing pre-trained nets require this, unless the dataset images are small such as the MNIST or CIFAR dataset, any user captured or standard size image must be scaled to around 224 x 224 pixels, in a 3-dimensional format.

2.2.2 CNN Feature Extraction and Image representation

In this style of system, to represent the image for the subsequent stages, a form of feature extraction takes place with the use of a very deep CNN, sometimes referred to as a DNN (Deep Neural Network). These large models are generally pre-trained to a high degree of accuracy in their desired field, such as scene or object recognition.

To utilise a CNN as a feature extractor it must first be trained. The process to train these models is conceptually similar to the methodology for typical Artificial Neural Networks. However, with a lot more data. Typical architectures for CNNs consist of Convolutional layers, Pooling layers and Fully Connected layers (Krizhevsky, Sulskever and Hinton, 2012).

Convolutional layers apply a mathematical convolution to the layers input and pass the result to the following layer. The next layer is typically a pooling layer. These layers take clusters of nodes from the previous network layer and output them as a single node. This style typically repeats until the fully connected layers are added where the final classification takes place. CNNs share the same architecture traits as ANNs in which the layers are connected via weights that are updated through the forward and backward passes of the model, as well as utilising the fully connected architecture for the models last layers.

As a CNN is generally applied for the classification application, it must be used as a feature extractor, by removing the output layers, this would therefore mean that the network would produce a high dimensional vector rather than label confidence scores. For example, the AlexNet CNN with the output (or fully connected) layers of the network removed, can produce a 4096-dimensional feature vector. It has been commonplace for CNNs pre-trained on the extremely large ImageNet dataset to be used as a generic feature extractor in many other areas, such as scene and object recognition.

Background

Features of images can be collected from something as simple as pixel values, collecting each value in an image would lead to the feature vectors being incredibly large due to the increase in mainstream camera quality etc. CNNs can be used as a feature extractor, by learning the features from a small grid and using this trained detector at different locations across the whole of the image, dramatically scaling down the size of the collected features making them much more manageable, combining this with pooling layers, and then alternating, creates a very discriminative feature extractor that can be utilised on high resolution images. This can be shown briefly in Figure. 2.4, for a trained 3x3 feature extractor over a 5x5 image. The yellow 3x3 grid shows the trained feature extractor during its first two steps over the blue 5x5 grid image. The feature extractor has stored methods to execute over those pixel values (the small red multipliers). The sum of the calculations returns the results in a new reduced dimensional vector, as shown in the green box on the newly created 3x3 vector to the right of the original blue image. The extractor is moved one step and the process continues.

Pooling would be done in a very similar fashion however simply collecting the highest value within its sliding window or adding them all together or averaging among others. This style of feature extraction technique is used throughout the conducted research, starting in chapter 4.

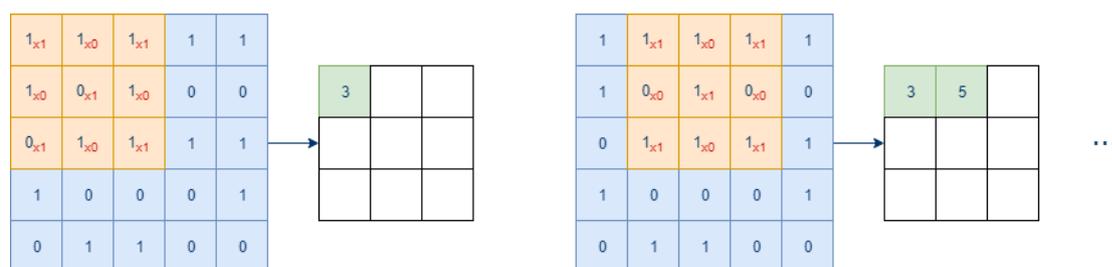


Figure 2.4. (Orange) Trained feature extractor on each 3x3 patch collecting smaller feature vector as it passes over each pixel of the image. (Stanford CNN tutorial, available at: <http://cs231n.stanford.edu/>)

2.2.3 Sentence Generation

As mentioned in the previous sections, sentences and descriptions have been constructed with methods such as templates, substituting entities into the variables within a fixed sentence and Conditional Random Fields (CRFs) which can produce longer sentences but can sound robotic and rigid. There have also been works in which the closest relating sentence/description has been retrieved from the corpus based on

Background

the image features. In a bid to overcome this issue in recent research the use of recurrent neural networks has been the main methodology for generating sentences from images.

2.2.3.1 Recurrent Neural Networks

Recurrent Neural Networks are based on the use of time steps, as most neural networks accept the current input example, RNNs take the current example and the input at the previous time step. This can be simply defined as the decision a RNN makes. The timestep, $t-1$ affects the decision of the network at timestep t . This is the RNNs two inputs, the present moment (t), and the immediate past ($t-1$). This can act as a form of context and allow the correct reactions to new data, much in the same way humans also respond to similar circumstances.

This is different from typical neural networks due to this feedback loop, revising their own outputs as inputs immediately after producing them. This adds memory to neural networks, collecting and storing the hidden information and patterns within sequences of data.

This can be described mathematically as:

$$h_t = \phi(W_{X_t} + Uh_{t-1}) \quad (1)$$

This defines the hidden state at timestep t (h_t). This is a function of the input at the given timestep, modified by the matrix containing the weights in addition to the hidden state of the timestep on previous multiplied by its own hidden state matrix. The generated error will be calculated within the back-propagation algorithm to adjust the weights to minimise this loss. ϕ is generally either a logistic sigmoid or tanh function depending upon the task, and this is utilised with the sum of the weight input and the hidden state.

RNNs can suffer from a number of issues which can dramatically affect the training and in the end how successful the model is at its given task. The largest issue is the vanishing and exploding gradients problem. These mathematical problems are on the basis that the matrices within these networks go through the multiplication process with a multiplier larger than 1, the values can quickly become immense. The reverse in the case of vanishing gradients is when the multiplier is less than 1. This makes the weights either immeasurably large, which can be truncated or squashed. However

Background

incredibly small (vanishing) gradients propose a much larger challenge (Karpathy, 2015). These RNNs can generate sentences and descriptions word by word, or on an individual character based level, and are used to overcome the limitation of the fixed input and output vector length of neural networks, examples shown in Figure 2.5.

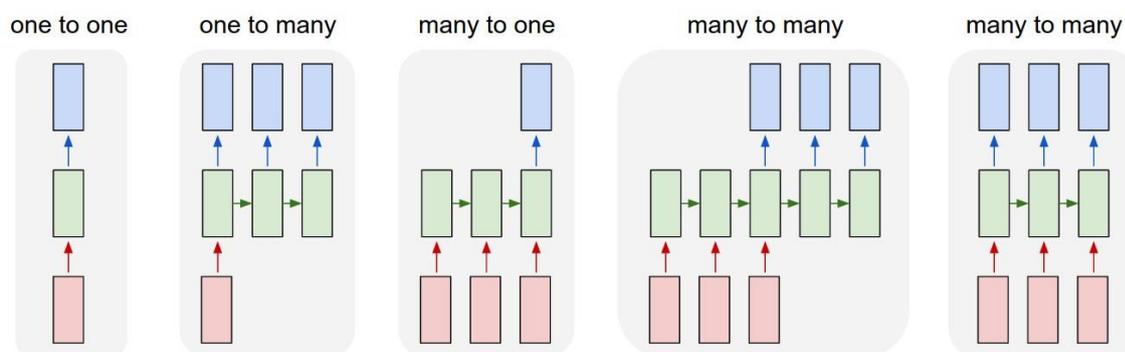


Figure 1.5 With each box a sequence, this above displays some of the versatility of RNNs (Karpathy, 2015)

2.2.3.2 Long Short-Term Memory Units (LSTM) and Gated Recurrent Units (GRU)

LSTMs were proposed after RNNs as a potential and usable solution to the problems of vanishing gradients as previously discussed. This is achieved by their memory that can help to store and maintain a more constant error. This allows the RNN it is linked with to iterate and learn over many more time steps. LSTMs differ slightly to those nodes typically found within an RNN, as the gated cells within LSTMs allow data to be stored, written and read from, the decision to perform each action is performed from within the cell. The gates within the LSTM are activated or blocked based on the input and its confidence which is determined and if necessary filtered based on its own set of internal weights, these weights are also adjusted during the RNN training procedure (Mikolov *et al.*, 2010; Mao *et al.*, 2014; Karpathy, 2015; Lipton, Berkowitz and Elkan, 2015).

The other unit in this section is the GRU, this cell is essentially identical to the LSTM as previously described however lacks an output gate. Instead, the GRU writes its contents to the overall neural network rather than the cell at each timestep.

2.2.3.3 Character Based RNN

Recurrent Neural Networks that follow the character level modality (Hwang and Sung, 2017) have a number of benefits over the word-based equivalents. For example, modelling out of vocabulary words is a large benefit. However, in recent research their

Background

overall performance does not compare to the word based methods. This could be down to their need to consider many more previously generated tokens in order to predict the next one, compared to relatively few words that need to be taken into account. There are a number of available character models both in research and in public repositories, such as the popular Char-RNN (Karpathy, 2016) package, which can generate anything from Baby names to Linux source code.

2.2.3.4 Word Based RNN

The word based RNNs are typically the modality used within the image caption domain, a high level example is shown in Figure 2.6. Pairing these with a CNN feature extractor as presented earlier provides a solid baseline for improving and innovating within this domain. The word based RNNs, are based on the principle that given a sequence of words, i.e. a sentence or caption, the probability of each word should be predicted, given the previous words. This allows the next word to be accurately predicted, when given the features and starting words, this combination allows for words to be continuously generated at each time step until the caption and the stop word has finished being generated.

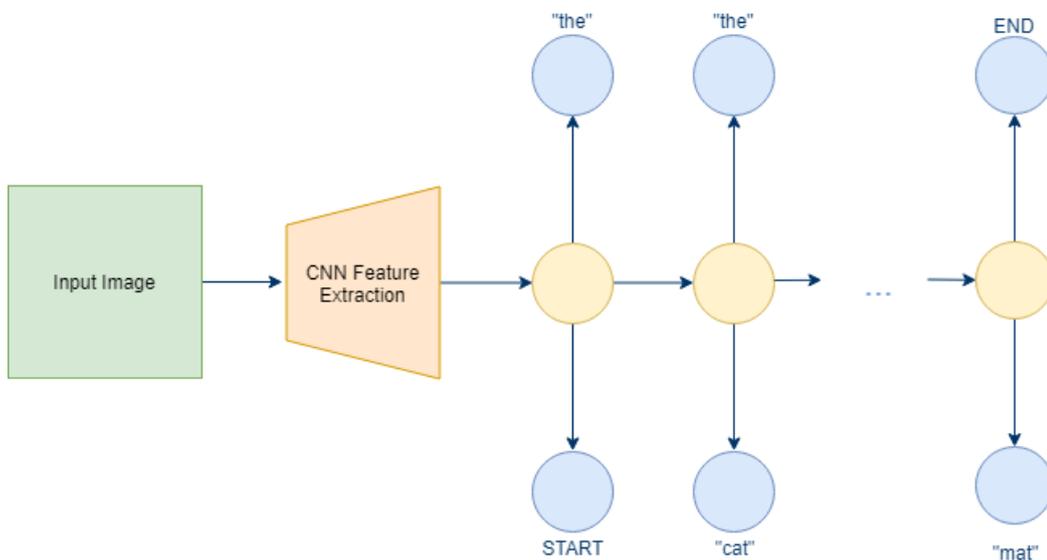


Figure 2.6. A high level representation of the word based RNNs producing words that would eventually form a full sentence.

2.3 Transfer Learning

Where possible, aspects of transfer learning have been explored. Transfer learning is utilised when there is not sufficient data in an intended domain, that when the typical training procedure is adhered to over fitting can occur and the model performs

Background

unsatisfactorily. To this end, related data can be utilised from either the target or another related domain to include information learned from both domains. This is similar to the belief in which children learn, for example a child can only store and recognise a certain number of objects within a day, anything new they see or learn is heavily weighted on knowledge and understanding from those existing learned classes.

Transfer learning can take many forms and has many different approaches in order to execute the knowledge transfer. For example, Cross-domain knowledge transfer and Cross-view knowledge transfer. Cross domain identifies the gap between source and target data that can be entirely different to another, for example cars. This domain could utilise data from wheels or car radiators to the unrelated such as laptops. Whereas cross-view aims to account for the near infinite available viewpoints of any one object (Shao, Zhu and Li, 2015).

2.4 Attention Based Methodology with a Multi-Layer Perceptron

Attention over the wording of an image, is considered a form of textual context between generated words and/or characters. Within the literature, this is can be accomplished with the use of a more traditional MLP (Multilayer Perceptron) (Shen and Huang, 2016; Yang *et al.*, 2016). In which the weight of each word within the overall sentence is calculated, this is achieved by passing each word and each extracted entity into the MLP. These MLPs are generally used in combination with CNNs for the feature extraction and provide the information required for an RNN to generate the correct words in the correct order.

2.5 Evaluation

There has been active research into automatically determining the similarity between generated and reference annotation descriptions. A number of methods have been utilised throughout this research, some of which were originally intended for other domains. Possibly the most accurate method for determining the quality of the generated captions is for a human to evaluate and score based on a number of criteria. However, this can be expensive and require additional clearing stages to ensure good

Background

data. Throughout research some of the common/popular criteria are as follows (Bernardi *et al.*, 2017):

- *Does the description accurately describe the image?*
- *Is the caption/description grammatically correct?*
- *Whether the caption/description contains no incorrect information*
- *Is the caption/description relevant for the given image?*
- *Is the caption diversely constructed (creative, no rigidity in templates for example)*
- *Does the caption/description sound human like?*

In recent research, most if not all methods refer to the MSCOCO evaluation script when comparing between relating or competing works. This standard script contains 4 evaluation metrics, these are Bleu (Papineni *et al.*, 2002), CIDEr (Vedantam, Lawrence Zitnick and Parikh, 2015), Meteor (Denkowski and Lavie, 2014) and Rouge (Lin, 2004). A modified script with a relatively newer metric, SPICE (Anderson *et al.*, 2016), is also available. These metrics are discussed in more detail later in the chapter.

2.5.1 Bleu

The Bleu score is possibly the earliest evaluation metric that has become ‘standard’ in the image captioning domain. The Bleu score was originally utilised for evaluating machine translation systems, between source and target languages. The score presents multiple methods outlining the number of matching n-grams (1-4). The basic measure of BLEU has been shown to favour smaller captions, as even if a caption is for example two words long, as long as those two words are in the reference, the score will be unreasonably high.

BLEU-1 works on an individual word basis. For each individual word within the reference caption, Bleu takes its frequency. The frequency of the word becomes W_{\max} .

Within the generated captions, the frequency of each word is capped at W_{\max} , meaning that any frequency below W_{\max} would be treated as its own value, whereas any frequency above W_{\max} will remain equal to W_{\max} . To calculate the unigram precision score for BLEU, W_{\max} is divided by the length of the generated caption. This is applied to each word within the generated caption and produces an overall score between 0-1, where 1 represents a 100% relation between reference and generated caption.

Background

A similar method is applied over more than one word, i.e. bigram or n-grams in order to generate the score for the different BLEU scores.

2.5.2 CIDEr

The CIDEr metric works at its best when used with many reference sentences, the dataset the authors provide contains 50 reference sentences for 1000 pascal VOC images (Everingham *et al.*, 2010). Their evaluation metric is based on determining the human consensus within a large amount of data. This is achieved by a triplet based method in which human annotated descriptions are collected, and an automated metric in which the consensus is captured.

2.5.3 Meteor

Meteor is a thorough search method that claims to align with human rankings stronger than other metrics. This metric evaluates generated and reference sentences by exhaustively identifying matches in a number of areas. These areas are as follows:

- Scoring words which match identically in their current 'in-sentence' form. This is achieved with the use of a word stemmer, such as the Stanford parser (Schuster and Manning, 2016).
- This is extended to determine if any of the matching words share a synonym within the WordNet (Miller, 1995) dataset.
- Finally recognising matching phrases to a pre-processed internal language paraphrasing table.

2.5.4 Rouge

The rouge metrics contains many individual metrics within the package, including multiple n-gram based methods, very similar to the Bleu score.

In recent research, the most commonly reported Rouge metric is Rouge-L This metric is a measure of the longest common subsequence, in which the score would increase depending upon the longer the length in which the two sentences are similar. Slightly differing LCS evaluation methods have been used to compare large text summarization applications.

2.5.5 SPICE

SPICE, or Semantic Propositional Image Caption Evaluation is a metric specifically designed to evaluate the similarity of automatically generated image descriptions more

Background

closely to human judgements. Their work states that the metric scores 0.88 in correlation with human judgements on MSCOCO as opposed to 0.43 for CIDEr and 0.53 for Meteor. This is achieved by initially converting both the reference and candidate captions into an intermediate representation, in this case a scene graph, this is achieved by semantically parsing the generated captions.

There has been recent research into the viability and meaning of these metrics (van Miltenburg and Elliott, 2016). As simply comparing text outputs to text outputs does not directly relate to where a given improvement within a system has occurred. They provide a very good benchmark for simple comparison, but contain little to no information regarding the strengths and weaknesses of the given model. They state within their research that when checking a system's outputs, they check for accuracy and then categorize the errors. They state that many systems do improve upon other frameworks, however only 20% of the generated outputs are free from errors while 26% are irrelevant or just wrong in comparison to the source image.

2.6 Deep Learning Frameworks

There are also a number of frameworks commonly used within machine and deep learning research. These frameworks are typically used to develop and implement these deep and powerful models. The frameworks range in functionality, complexity and ease of use, however all have very similar end goals to be powerful, reliable and efficient and creating and running these very complex models. Some of the more popular frameworks are briefly described below:

Caffe (Jia *et al.*, 2014) was widely used among Deep Learning research and has gained popularity among senior researchers, before being replaced with newer more efficient alternatives. It was initially produced with the aim to enable the production of state-of-the-art deep learning algorithms with a growing collection of existing pre-trained models. The framework is equipped with bindings that allows communication with C++, python and MATLAB allowing for both training and distribution of the generated models. At the time of the research Caffe was stated to be able to process > 40 million images in a single day on a Nvidia TITAN GPU.

Caffe 2 (*Facebook Research*, 2017) is the direct successor and intended replacement to Caffe. This framework aims to provide an easy way to experiment and utilise the same created models within the cloud and mobile platforms with minimal effort.

Background

Google has released a proprietary framework named TensorFlow (Abadi *et al.*, 2016). This framework is based on the utilisation and manipulation of mathematical Tensors and data flow graphs. This system allows for deployment and utilisation again on both CPU and GPU, however it holds a major advantage in its ability to be ported to mobile devices within the same API.

Theano (Theano Development Team, 2016) is a similar framework to TensorFlow, however integrally relies upon the python Numpy framework for most of its matrix manipulations.

Keras (Chollet, 2015) is a relatively new framework, and aims to be a high level neural network API. This particular framework runs on top of either TensorFlow, Theano or CNTK. The high level nature again, allows for rapid prototyping due to the ability to build and remove layers quickly and easily within code. Allowing the created models to be run on both CPU and GPU seamlessly adds to its functionality.

Torch (Collobert, Farabet and Kavukcuoğlu, 2008) is a machine learning framework designed specifically for use on GPUs and is designed for speed due to the nature of the LUA scripting language. When used in tandem with the CUDA implementation required for GPU processing and optimization it is an incredibly fast and efficient tool. Torch has heavily influenced a python framework producing a *Python Torch* framework - PyTorch. This has much of the same functionality as Torch, however allows many of the familiar python packages to be included within the models and overall extend the functionality of PyTorch.

2.7 Summary

Image description is a multi-disciplinary research area depending upon the methodology explored. This research utilises both composite and end-to-end style architectures. Most of the domains utilise powerful CNN feature extractors paired with classifiers for a given purpose. This can be image or object classification, and modified for the use of localisation. We explore utilising these differing domains to improve upon existing systems. Many systems produce captions without explicitly recognising objects or people, and rely heavily on holistic features. We propose that combining

Background

these techniques allows for a more descriptive and potentially more useful image description.

The composite methodology requires research from areas such as object detection, classification and localisation, each relying on deep learned CNNs. Combining these in an efficient manner could allow for more accurate and descriptive outputs. As discussed in later chapters, combining these algorithms can allow for prediction and classification of longer and more accurate regional descriptions, which in our composite methodology equals better descriptive outputs.

For sentence and/or caption generation, RNNs in various forms are utilised, using them as a sequential classification framework, as well as a machine translation framework is explored throughout this thesis. These are areas that RNNs have been underused, currently focusing on their uses and modifications within the end-to-end style architectures utilising the holistic image features.

The end-to-end architectures in which CNNs + RNNs produce captions are becoming increasingly popular, this thesis proposes extensions and modifications to this ‘standard’ pipeline in order to fulfil the aims and motivations as expressed in Chapter 1.

The metrics previously discussed are utilised throughout the experimentation within this thesis where applicable. Due to some requirements and the nature of some of the metrics, wide scale testing is difficult or impossible. For example, the CIDEr evaluation essentially only allows for testing on its dataset due to the amount of reference sentences available (50), collecting or adding reference sentences up to 50 on other datasets would consume a large amount of resources, making testing difficult.

Background

Chapter 3 Related Work

As mentioned in previous chapters, depending upon the style of system, image captioning/description requires several varying techniques from multiple research domains. Composite frameworks generally require an object detection/localisation method, paired with some form of sentence or description generation framework, while end-to-end methods focus on the CNN+RNN architecture. In this chapter, the active and recent research works within the fields that are required and have been utilised throughout the life of this research are explored further.

3.1.1 Localisation

A Regional Convolutional Neural Network (R-CNN) was proposed by Girshick *et al.*, (2013). This traditional R-CNN is based upon collecting many regions that could potentially contain objects and/or people, before extracting machine learned features and classifying each region. The R-CNN process starts with the selective search (SS) (Sande, 2011) region proposal computation. These regions are then manipulated individually to size before the system extracts a set of 4096 features from each region. The extracted features are then passed to multiple category-specific linear SVMs with each SVM representing a single object class, which produces an object classification label for a region. Although powerful this initial research had a number of limitations, with possibly the largest being the model inference time during the testing phase. Processing a single image can take minutes, which for modern models and requirements would make this model almost unusable for production systems. It also relies heavily on some of the more traditional methodologies, in this case the SVM. This initial R-CNN can classify 200 categories of images, and this requires training and inference of all 200 SVMs once the features have been extracted. This could potentially be amended with the use of a more efficient model replacing all 200 SVMs into one.

Girshick, (2016) improves on the original R-CNN and proposes a much more efficient regional detector, i.e., fast R-CNN, to reduce computational cost and improve detection accuracy. The fast R-CNN implements Spatial Pyramid Pooling networks (SPPnet) (He *et al.*, 2014). This modification enables the whole image to be transformed into a deep feature map. In other words, rather than manipulating individual regions, the network now takes an image and all of the generated region

Related Work

proposals as inputs, and pools each region proposal into a fixed size feature map and then maps it to a feature vector through the fully connected layers of the CNN. However, this new version still relies on pre-computed image regions. This exposes computation of image regions to be the bottleneck during this style of object detection. A further bottleneck with this proposed system is also the feature that improved the accuracy and speed – SPPnet. This network is stated within their work to possess the same drawbacks as the initial work, in that it is still a multi-stage, compositional framework that involves CNN feature extraction, SVM training and bounding box regression. This SPPnet also prohibits the convolutional layers in the framework that come before the SPPnet from being updated, which can have a large effect on the potential accuracy.

To overcome the above bottleneck problem of fast R-CNN, a faster R-CNN (Ren *et al.*, 2015) is combined with a new technique called Region Proposal Network (RPN). The RPN is trained on selective search proposals to create proposals at a much higher speed than the previous efforts. This implementation allows for near cost-free region proposal generation. There are also other proposed alternative region generators. E.g., the EdgeBoxes algorithm (Zitnick and Dollár, 2014) has been mentioned in recent research as a compromise between the exhaustive nature of SS and a segmentation technique (Sande, 2011). The RPN, although more accurate within the test circumstances in this research domain, requires more training data than existing models. This method of training could also lead to overfitting and bias, which may mean that the generated region proposals may not be accurate in all test domains.

The recent work of Redmon *et al.*, (2016) produces a real-time object detector, called YOLO, trained on the PASCAL VOC 20 (Everingham *et al.*, 2010) class detection challenge. This detector is able to run in real time on a webcam at frame rates of up to 45 FPS. Further research from Redmon *et al.* has increased the speed of this detector to 155 FPS, however the main limitation of this is that the model sacrifices its overall accuracy, delivering a significantly lower Mean Average Precision (mAP) than the related models.

The main reason for the increased speed in comparison to other methods such as the R-CNN is the network architecture and structure. For example, existing pipelines rely on collecting and generating region proposals as an initial stage from across the entire

Related Work

span of the image, these regions are then passed to a classifier in order to tag the appropriate object label to each generated region. Post processing stages involve Non-Maximum Suppression (NMS) to remove duplicate or similar image regions and rescore the object labels.

Object detection has been extended by Mao *et al.*, (2015) to allow for their system to generate unambiguous descriptions of objects or image regions. This work also provides a new large scale paired image-description dataset based on MSCOCO (T.-Y. Lin *et al.*, 2015). This work also notes that most existing image captioning systems have an underlying problem, i.e. the fact that a given caption can be very subjective. If one caption describes one aspect of an image, then another caption that describes a different aspect, will not match this initial sentence. Even if it still correctly captions the image however such situations owing to subjectivity would not score highly on any evaluation metric.

3.1.2 Feature Extraction

Mopuri, Athreya and Babu, (2017) state that utilising pre-trained CNN feature extractors that have become widespread, are limited in their capability due to the limited information given to these feature extractors during training. Their research states that areas such as scene recognition fall into this problem, and that the knowledge transfer needs to be stronger. To this end, the authors utilise features learned from existing caption generators, (Johnson, Karpathy and Fei-Fei, 2016; Vinyals *et al.*, 2017) to learn novel image specific representations. This shows that the features extracted when paired with textual caption information perform better than those traditional CNNs which are simply trained on a single label. Most of these models exhibit some form of trade off for their added accuracy. Further details will be presented within the relevant article discussion.

Karpathy *et al.*, (2014) aim to extensively evaluate the power of CNNs on large scale video stream classification due to their performance on image recognition problems. They conduct multiple experiments in order to allow CNNs to take advantage of the spatio-temporal information that is present within these video streams. Their research also highlights a deep learning process in which a low-resolution stream and a high-resolution stream are passed into the architecture as a method to increase the runtime performance of these models, with minimal to no cost to their accuracy. It is interesting

Related Work

to note that, within this related research, models are designed and implemented based on multiple frames from within a video. It shows little performance increase on those models that are based on images and only take into account single frames at a time, stating that this fine grained inter frame information may not be wholly important to be successful within this task.

3.2 Attribute Prediction

Attributes can be described as high level features, which were widely employed by many computer vision applications. High level features in images consist of information such as age, gender and hair colour etc. for people, and size, colour and texture for detected objects, which can be used to provide powerful descriptive capabilities to image description generation systems.

Attributes have been used in numerous ways, either by adding high level information to labels or by using them to aid in other computer vision tasks, i.e. semantic or data driven attributes. Some early works utilising object attributes (Farhadi *et al.*, 2009) aim to describe objects by their attributes allowing for objects that do not or cannot be classified into a specific label to be described by their shape, size colour or texture. In their work they propose an example of a goat/sheep, rather than attach this label they detect ‘has horns’, ‘has legs’, ‘has head’ and ‘has wool’ which could potentially be used to infer or guess the label by users. The obvious limitation to such a system is that it does not actually categorise or label the subject, meaning that an additional stage must be present to annotate an image fully.

Attributes have been used by Dhar, Ordonez and Berg, (2011) for predicting Aesthetics and Interestingness, their work demonstrates predictors for compositional attributes which relate to how an image is laid out. Content attributes are those attributes specifically relating to the description or presence of objects, animals and/or people. And finally, Sky-Illumination attributes which aims to detect the characteristics of natural illumination in certain images. They also indicate that their interpretation of high level attributes indicate that these are the kinds of attributes that humans may use during the description of the same images. This is achieved with the use and extraction of multiple low-level features, such as colour histograms, colour spatial distributions and haar features. These are combined based on the individual use cases to represent either compositional, content or sky illumination attributes. The

Related Work

combination of these represents an images 'interestingness'. The potential limitations of such a system is that these attributes can be subjective. For example, an attribute classified as 'blue' would represent a feature of the subject that is blue. However, taking into account interestingness is per user, what one user may find interesting does not translate to all users. Dhar, Ordonez and Berg, (2011) have conducted research into the psychological aspect of this as means to overcome these limitations.

(Zheng *et al.*, 2014) achieve semantic image segmentation with objects and attributes by "*formulating the problem of joint visual attribute and object class image segmentation as a dense multi-labelling problem*". The aim of their approach is to jointly label attributes and segment objects from the image. Each image pixel is labelled with a class label, i.e. car, bike, dog etc. Visual semantic attributes are object materials such as wooden, or metallic and object properties such as shiny. This overcomes the main limitations of the previous research from Dhar, Ordonez and Berg, (2011). However, they state that their main research limitation is the lack of GPU support and this current implementation lacking the ability to parse 3D scenes due to the lack of available data. This is due to the model and its ability to parse an existing 2D image. The authors aim to overcome this barrier via constantly adding annotations to the data hoping to create a large-scale 3D database for this purpose.

Research from Bourdev, Maji and Malik, (2011) uses pose information to determine people attributes such as gender, hair style and clothing style. To achieve this, they implement a part-based approach using what the authors refer to as poselet information. Poselet is the term for the detection and segmentation of a user's pose within an image utilising part based image detectors. They implement the method from their previous work to train many poselets based on the training and validation sets, each poselet during training has its own soft mask which determines the probability of each body component at each location. The main limitation of using these poselets on an image level is that these poselets are focused on a particular region. For example, research from the authors (Bourdev *et al.*, 2010), states that a poselet of a user's feet does not hold any information relating to the position of the user, e.g. are they sitting or standing? Although the poselets can classify attributes with substantial occlusions, a further stated limitation is that they cannot classify the occlusion. E.g. if a table is blocking most of a person, the person could not be identified.

3.3 Description Generation

As in the previous chapter, image description generation can be defined as two different strategies, composite and end-to-end. There are many related research articles which cover these two areas, a number of these are reviewed as follows.

Devlin *et al.*, (2015) provides an overview into recent image captioning systems and their different results and structures and how they define their state-of-the-art results within image captioning. They highlight two works, the first in which a set of words is generated from a CNN trained on a large dataset of images, and then applies Maximum Entropy language model to arrange these generated words into a caption. The second, is the most common method within research, the CNN as input into a RNN that generates the final caption sequence, word by word. Some highlighted concerns within this research are relating to the evaluation and the associated metrics. The authors state that the models and approaches explored gain in the BLEU score, however in this case it does not translate directly into human judgement. Further discussion on these limitations is explored later in this thesis.

3.3.1 Composite Image Caption

Another framework from Fang *et al.*, (2015) is proposed for image description generation. Their system is trained with a weakly supervised method as there was no bounding box information to learn from. It then generates sentences with a statistical model that works effectively as a search for the most likely sentence given a set of words. Although early within the active research of image captioning they state that their methodology improves upon the BLEU score. The human judgements only gauge the generated outputs as an improvement 34% of the time. Their method involves using two deep networks (Krizhevsky, Sulskever and Hinton, 2012) to map images and text fragments to a common space.

Recent research Fang *et al.*, (2015) proposed a caption generation system utilizing a bag of words method. Their work implements multiple instance learning and uses visual classifiers for words that commonly appear in existing captions. In addition, their system treats the caption/description generation as an optimization problem. It takes the previously generated words and then finds a sentence/caption that has the highest likelihood to caption the image that contains every word it has detected.

Related Work

Hendricks and Venugopalan, (2016) proposed a Deep Compositional Caption framework in which they aim to describe and caption images wherein they detect objects which are not present or labelled within their dataset corpus of images. They achieve this by transferring knowledge between large scale object recognition and text datasets. This allows their model to describe novel objects and their interactions and relationships with other novel objects. Furthermore, the same methodology can be extended to generate descriptions of video. The main issues and errors within this implementation are when it incorrectly classifies the new class of object in combination with the inferred textual data, this will result in the output captions being highly inaccurate.

3.3.1.1 Templates

Sentence generation has been implemented in numerous ways over the years, ranging from templates (Khan, Al Harbi and Gotoh, 2015), graphical models (Kulkarni *et al.*, 2013) and neural networks (Vinyals *et al.*, 2014; Venugopalan, Xu, *et al.*, 2015; Yao *et al.*, 2016).

Template matching is applied by Liu *et al.*, (2013) to describe video streams. In their work, semantic level representations are used to describe events in the video and classify human activities. Bounding boxes are adopted to collect attribute information such as location, size and velocity. From this information, ‘subject-verb-object’ could be determined and a description constructed.

Templates have also been implemented by Khan, Harbi and Gotoh, (2015) in which the high level features collected come from initial stages of human and object detection and scene classification. From this initial processing, age, gender and emotion information is collected, along with actions and any detected objects. The work also relies on spatial relations to specify where an object or a human is located in a scene compared to one another in order to provide more descriptive sentences.

Within this domain of image caption/description generation templates can be useful but have a severe limitation when comparing system generated outputs to Ground Truth outputs, and suffer when human judgements are taken into account. This limitation is derived from the rigid structures generally imposed on the generated words. As previously mentioned the SVO – Subject, verb and object structure is one of the

Related Work

simplest methods. Abiding by this format can lead to the descriptions becoming short and repetitive, regardless of the content, meaning that metrics and judgements can be adversely affected.

3.3.1.2 Other Machine Learned Methods

Kulkarni, Premraj and Ordonez, (2013) develop a system for image understanding and description generation. It employs object detectors, attribute classifiers and preposition relationship functions and a Conditional Random Field (CRF) based sentence generator. Their object detectors are trained upon PASCAL 2010 with 24 categories in total. For attribute classification, they implement 21 SVMs to classify a set of attributes. However, they only conduct attribute classification for objects, without considering classification of human attributes. The number of available attributes are also lower than a number of existing works, limiting the systems ability to directly compare to human judgements.

Kong *et al.*, (2014) utilise a Markov Random field as their method to merge both visual and textual information. This method takes the full sentence descriptions in order to help annotate and locate objects within a 3D scene. This system wholly relies on a 3D RGB-D dataset and relies on the depth information to achieve their results.

Lin *et al.*, (2015) propose a method in which similar methods utilising 3-D image data are used. They state that this allows for the generation of multi-sentence coherent image captions. This differs from typical frameworks due to the use of a 3D parsing process in which objects, attributes and relationships are jointly inferred. As well as sentence and grammar being considered, the sentence generation and construction utilises coherence methods. These methods first generate scene graphs from the collected data, before forming a semantic tree, from which grammar is added before forming a full multi-sentence image caption of indoor scenes.

3.3.1.3 Sentence Retrieval

Sentence retrieval is another method that is occasionally utilised when describing or captioning images, however rather than generate sentences, whether with templates or machine learned methods, a sentence is retrieved from a given set where the two images are closest. This has a number of benefits in that the speed and resource requirement is much lower, and this will always produce grammatically correct

Related Work

sentences as long as the corpus is clean and collected properly. However, this has the downside in that similar images could still differ widely, for example the scene could be identical however the focal object or point in the image could be different, if the images are similar enough this could lead to the incorrect objects being labelled and identified in the description.

Ordonez *et al*, (2016) state that ideally, generated image captions would sound like and be structured as if a human had written them. In an effort to achieve this, they propose a data driven approach that utilises retrieval-based methodologies. This involves retrieving either whole captions when the image is visually comparable, parts of a caption or description or even text from a large corpus. In the second case they propose an optimisation method which merges parts of sentences into full natural sounding descriptions. This research states that this method leads to more general, relevant and human-like image captions. However, as previously stated being more general can have a negative impact upon the generated metrics, and be limited by the complexity and diversity of captions already within the systems dataset.

Smeaton and Quigley, (1996) proposed an early look into the semantic distances between words in order to retrieve the most relevant image caption.

Aslandogan and Yu, (1999) review methods in which rather than retrieve sentences and descriptions based on an image, they retrieve images and videos based on the textual descriptions and captions etc. They also explore visual features to achieve the same result, i.e. by extracting visual features for colour, texture, shape and spatial relations, these have been utilised in earlier works for tasks such as attribute prediction. In terms of the non visual features, they acknowledge the difficulty within this methodology relating to the natural language ambiguity normally found with these kinds of queries. To address this, they quote 3 differing works in which they restrict the type and style of sentences utilising inference rules, relevance feedback methods and structured description queries. For example, they state some of their queries as follows in order to retrieve relevant images:

“Retrieve images which contain 25 percent red, 50 percent blue, 25 percent yellow.”

“Retrieve images with sky blue at the upper half and green at the bottom half.”

Related Work

The above examples place restrictions and specifications that can mathematically apply to images and their features, as with colour features it can be easier to determine the percentage of an image which is a specific colour, with a similar practise with added natural language being used within the second example. Similar features can be utilised for the video retrieval task however used on individual video frames and combining them over the entire length or a specific number of frames.

3.3.2 Video Caption

Thomason *et al.*, (2014) propose a Factor Graph Model, FGM, to select content from within video streams. This is achieved by integrating visual and textual linguistic information to select the most relevant subject-verb-object-place description within a video. They also claim to be the first work that annotates video to include the use of scene classification. Their initial framework utilises previously developed state-of-the-art applications for confidence measures on entities, activities and scenes, this is combined with their proposed model to estimate the most likely subject, verb, object and place to generate a description. Other more complex methodologies could potentially have been explored to increase the diversity. As previously stated the SVO methodology can be severely limiting in regards to description generation. RNNs could achieve better results, as they would be 'aware' of the structure of sentences and generate to a more fluid structure. This could have its trade off in computational time and power, in that SVO methodologies are cheaper and faster compared to deep learned methodologies.

Yu *et al.*, (2015) propose a machine learned approach in which hierarchical RNNs are used to address the video captioning problem, to generate single or multiple captions to describe a real life video. Their proposed methodology allows for single caption generation as well as a separate model for paragraph generation. The single caption exploits temporal and spatial attention mechanisms to focus on desired objects. The paragraph generator functions by collecting the inter sentence relationships through sentiment embedding. They state that this methodology outperforms current state-of-the-art within this field. A potential limitation from such a structure, is that if a generated sentence is incorrect, it could dramatically skew the overall paragraph results, delivering an incorrect overall description. A potential method to overcome

Related Work

this would be to investigate the inter sentence relationships at an earlier stage and disregard sentences whose relationships are too far apart.

3.3.3 End-to-End Image Caption

3.3.3.1 Dense Labelling

One other aspect that utilises the end-to-end nature of these models is dense labelling. Recent work of Johnson, Karpathy and Fei-Fei, (2016) introduced a dense labelling system to this field. Rather than captioning an image, their work captions many individual regions with rich annotations, e.g. objects and attributes. This is achieved with a localization layer which acts as a region proposal generator to annotate image regions. This layer was developed based upon the research of Ren *et al.*, (2015), in which a Region Proposal Network was trained to generate the regional proposals rather than relying on the existing less efficient techniques such as EdgeBoxes (Zitnick and Dollár, 2014) or Selective Search (Sande, 2011).

(Tan and Chan, 2017) proposed a system similar to the existing work of Johnson, Karpathy and Fei-Fei, (2016). However, rather than utilizing the typical word based approaches that RNNs tend to adopt, their model encodes the sentence as a combination of both phrases and words for image caption generation. Their system, although competitive, is trained on the same data as other existing systems and their metric scores within the higher n-grams are beaten by some older models. They state that this is due to their system producing different variations of the same phrase. E.g. ‘a man’ and ‘a person’. This does however show how much of an effect gender attributes can have on the overall metric scores of these systems.

Within this dense labelling domain, (Krishna, Hata, *et al.*, 2017) apply similar techniques to dense captioning events but within series of images, i.e. videos. Existing work in this area has been explored with action classification, for example outputting such labels as ‘dancing’. These methods work well however lack depth and detail that a human would generally use when describing an action within a video. Their work aims to generate multiple descriptions for multiple events within a video and localise them to their specific start and end times. This has a number of challenges over the dense labelling/captioning as actions have dramatically varying times, as well as the potential to overlap, to this end a method to encode the varying degrees of video frames is proposed. Context is also explored in order to maintain the required detail of

Related Work

the labels. This work utilises a similar process to the dense labelling, utilising similar modules within their network. i.e. to generate proposals, albeit for video sequences. The captioning module is addressed to form an overall pipeline.

3.3.3.2 Attention/Saliency

The attention mechanisms presented within this section are intended to overcome some limitations of typical methods. I.e. the models and output change minimally once the model has started producing the description, and the limitation in which the models typically only describe the whole scene (Pedersoli *et al.*, 2017). Xu *et al.*, (2015) proposes a method to caption images through the use of an attention model. This attention model is able to improve the system outputs, especially for the cases where there is a large amount of clutter in a given image. The system also uses the image as input and collects a feature vector from a deep CNN. The feature vector is passed to an RNN which employs attention over the image to generate image descriptions.

Li *et al.*, (2017) state similar limitations as this thesis regarding the limitations presented by the common CNN+RNN image captioning framework. That is due to the use of only global representations at an image level. Their work proposes a GLA method, i.e. Global-Local Attention. This method integrates object information from a localised image representation in combination with the overall global, whole image representation. This aims to label individual objects within the image, while maintaining the overall scene, image and object context.

3.3.3.3 Recurrent Neural Networks

To address the image captioning task, Mao *et al.*, (2016) implements a system framework which takes as inputs, the whole image and a single bounding box, then outputs a description or a sentence using word based RNNs relating specifically to that given region. This work is very similar to existing caption and description generation methods however it simply annotates one region at a time. The baseline model is very similar to other existing and state-of-the-art captioning systems, e.g CNN + LSTM. To achieve this, they modify the training objective, called MMI (Maximum Mutual Information) this aims to penalize the model if it believes the sentence could also be used to describe other aspects/regions of the image.

Chen and Zitnick, (2014) proposed a bi-directional model capable of generating sentence-based descriptions from images and visual representations from descriptions.

Related Work

The recurrent visual memory enabled the RNN to learn to reconstruct visual features from the previous words and learn long-term visual concepts while the description was generated, i.e. latent variables that encoded the visual interpretation of the generated words were used as a long-term visual memory of the words that have been generated previously. In comparison to image description generation tasks, their model conducted the maximization of the likelihood of both the next word and the visual features given the previous words and their corresponding visual interpretation. Their system was evaluated with the tasks of image description generation, image retrieval and sentence retrieval across diverse datasets and achieved impressive performances.

Recent work by (Mathews, Xie and He, (2015) also attempts to address a shortfall in image captioning systems. They aim to describe an image with sentiments present in the description. For example, to describe an image with positive or negative sentiments. The downsides to this style of system are very related to the limitations presented in this research in regard to the effects of these systems on the datasets that are available in the public domain. As this system adds sentiment vocabulary that is not present within the corpus, the outputs are lower than other competing datasets in certain metrics. However, the authors crowd source their own data in an attempt to compare better to human judgements.

The architecture proposed is built upon the increasingly popular approach of combining CNN extracted features with a RNN (Convolutional Neural Network + Recurrent Neural Network), and allows for the generated descriptions containing sentences to be deemed as least as descriptive as the fact or ground truth sentences. For example, the output of this system could change from ‘a black and white cat lying on a bed’ to ‘a close up of an adorable cat lying on a couch’. The architecture of this system is still based around the CNN+RNN structure however combines the two of them in parallel, allowing one architecture to generate ‘factual’ descriptions and the other generating sentiment words.

Elliott *et al.*, (2016) extends the methods of image captioning with extracted machine learned features and sentence generating sequence models by combining them with a machine translation approach. They aim to use the information present within the sentences as well as image features/contexts to remove much of the ambiguity present in translation. E.g in "Ein Rad steht neben dem Haus", "Rad" could refer to either

Related Work

“bicycle” or “wheel”. However, when paired with visual data, the intended translation can be extracted more easily.

Tran *et al.*, (2016) have produced a system that could richly caption images. Their research claims to be able to detect and classify a large range of visual concepts. This includes specific locations, as well as specific persons such as celebrities or people of influence. Their framework consists of a compositional approach. It combines a feature extracting CNN, which passes features to their visual concept network that was trained on 700 visual concepts, such as celebrities and landmarks. It then follows similar research in the field and passes these into a language model. Their work has an advantage, i.e. if the system’s confidence is low, instead of generating a rich caption, it can produce a simpler caption that can essentially annotate/list the objects within the image. This approach utilises a composite style approach combining rich detail classifiers and generic scene information. A possible limitation to this approach at this stage within the research could mean that the generated outputs could be less relevant. In the example when a person is named, ‘Barack Obama’, correctly within the description, the caption is highly accurate. However, if ‘Barack Obama’ had been labelled incorrectly, the outputs could arguably be worse than the proposed attribute based approach.

Jia *et al.*, (2015) propose an extension to the LSTM framework, dubbed gLSTM. This extension utilises semantic information that is extracted from within an image. This semantic information is then added into each LSTM unit as an extra input. Their aim is to tighten the descriptions to be more closely related to the actual contents of the image rather than a high level caption. This can have similar limitations and restrictions to the previously discussed methodologies.

Liang *et al.*, (2017) present a possible flaw in existing methodology, in which they are restricted by the small corpus the images are paired with, and fail to annotate the rich information that is present and required in order for a description to be meaningful. To this end, they present a semi-supervised paragraph generation system that can generate varied, coherent descriptions by utilising semantic regional information as well as the textual information within a large corpus. This proposed framework utilises Generative Adversarial Networks, that uses a structured paragraph generator and a multi-level paragraph discriminator. The generator utilises region based information

Related Work

and language attention at each time step to produce paragraphs in a typical recurrent fashion, while the discriminator, assesses the quality of those paragraphs for plausibility at a sentence level and a topic level over the whole generated paragraph.

Yang *et al.*, (2017) propose a method for image captioning that specifically utilises two networks for object detection and localisation methods. These models collect information regarding the objects information as well as spatial relationship in terms of the objects within the overall image. They claim that this proposed model is similar to the attention method used within the human vision system. The authors state the main difference between their model and the typical soft attention approach is the nature in which the attention is applied. Typical soft attention methods are utilised on the generated output words, whereas within this work it is focussed upon the objects and their surrounding spatial information and relationships.

Socher *et al.*, (2014) propose utilising dependency trees with an RNN, called DT-RNN. This allows this newly proposed model to focus on the action and agents within a sentence. This model is able to learn the vector representations for a given sentence based upon the dependency trees. Learning the outputs from CNNs applied to the same images, and mapping them into the same space allows both sentences and images to be directly compared. They are also shown to outperform standard and other RNNs, and a bag of words baseline for image retrieval from a natural language caption as well as retrieving a caption for a given image.

Yu *et al.*, (2015) present a novel approach in which hierarchical Recurrent Neural Networks are used in the video captioning domain. In their work, this is described as producing one or more sentences to describe a *realistic* video. To achieve this, they propose two generators, one to produce short sentence that can describe or annotate a short video or sequence of frames, and the other to generate full paragraph descriptions of the image by capturing the inter-sentence dependency by utilising the outputs of the sentence generator.

3.3.3.4 Natural Language Processing (NLP)

3.3.3.4.1 Long Short-Term Memory (LSTM)

Vinyals *et al.*, (2014) propose an image caption generator by integrating a deep CNN with Recurrent Neural Networks (RNN). CNNs are used to generate features as

Related Work

previously described and are fed into a RNN. Their work employs the Long Short Term Memory (LSTM) to decode the CNN feature representation into a sequence of words for image description. This is stated to be used in this area due to its ability to cope with the vanishing and exploding gradient problems which are common within RNNs. This particular style of model is subject to the main limitations this thesis intends to move towards overcoming. The nature in which the CNN is deployed in these scenarios extracts features over the entirety of an image. This could result in the system overlooking a number of details that may be helpful in forming a human like description.

Recent work from Matsuo *et al.*, (2016) showed initial exploration of quantitative natural language descriptions utilizing human brain activities. Owing to the lack of brain activity datasets for deep learning research, the work re-uses frameworks of Vinyals *et al.*, (2014) and Xu *et al.*, (2015) Overall, the work synchronizes image datasets from movies and brain activity data from an fMRI scanner, and relies on the MRI data to generate descriptions. The main limitations of this style of system is the lack of available brain activity – text training data. Initial deep learning experiments with this data were conducted when the model was trained on 3,600 activities – text pairs. This meant that pre-training on an entirely unrelated domain was required for the deployed CNN. Overall, this could result in the generations to be biased towards its original training, with the textual generations not being as accurate as possible.

Users' vision has also been considered in attempt to improve existing image captioning research. Human gaze has been explored for tasks such as localization in the form of attention. For instance, Sugano and Bulling, (2016) explored gaze-assisted image captioning by examining the relationship between human gaze and attention mechanisms. Also, the attention mechanism has been adopted by the work of Xu *et al.*, (2015)

Recent work from Microsoft, (Gan *et al.*, 2017). propose StyleNet, generating image captions with differing styles, such as *romantic* or *humorous*. This is achieved through their proposal of a novel model component, dubbed factored LSTM. The module automatically extracts the style information within the textual corpus, then allowing the style generation to be chosen at runtime. This is achieved by utilising 2 sets of data, the image/video that is paired with captions, as well as stylised textual data. The

Related Work

limitations regarding this type of stylised outputs are related to the systems outputs not being directly comparable to human judgements, as the subject and style can change dramatically, and although grammatically correct may have no truth or relevance to the image. For example, a ‘romantic’ output of ‘A dog runs through the grass to meet his lover’ does contain an image of a dog running on grass, and with the romantic added style, the truth and accuracy of the description are jeopardised.

Person Search with Natural Language Description has also been explored by S. *Li et al.*, (2017). This research takes natural textual description of a person, and ranks all images within its corpus to return the most relevant image to the textual description. This is effectively the exact reverse of the image captioning problem. An extension to the RNN, a Gated Neural Attention RNN is proposed, this model takes a description and a person image as input and outputs the similarity. A word-based LSTM is used to pre-process each word along with its attention mechanism. The aim of the model is to search word image relations effectively, allowing a model to determine if particular regions are being described within the textual description.

Similarly, Hu *et al.*, (2016) uses comparable techniques to retrieve objects based on natural language descriptions. Rather than typical object retrieval based on natural language or even just object labels, this research aims to localize the desired target object from within the larger image. This involves collecting and determining spatial information within global scene contextual information. They present a novel Spatial context recurrent convolutional network, in which the query, global and spatial context as well as local descriptors are accepted as input, to score candidate regions.

3.3.3.5 Gated Recurrent Unit (GRU)

Many of these methods can be limited in their complexity and descriptiveness due to their collection of holistic image features combined with the image-caption dataset, which typically only utilise relatively short captions. This is further explored in section 3.6.

Research from Chen, Dong and Li, (2014) proposes a simplified GRU framework that can be implemented as a new style of GRU layer within different popular ML frameworks. This layer achieves comparable results to the LSTM, however with less learned parameters, increasing training speed while reducing memory consumption. This research essentially follows the same system pipeline as many of the previous

Related Work

systems, i.e. CNN+RNN, however the RNN takes on its form of a simplified GRU. This achieves similar performance to state-of-the-art results, however may not perform as well as other larger models due to the reduced parameters and feature vectors in their simplified GRU.

GRUs have also been utilised in research domains such as visual and textual Question and Answering, this is achieved through new modules which incorporate attention gates that utilise the global information from within the image, this new model does not require the supporting facts, i.e. the facts that are relevant to answering a particular question, the model instead learns required facts from a much larger set.

3.4 Image Synthesis from Descriptions

There has also been related research in not only generating descriptions of images, but the reverse. Generating or synthesising images from a textual or linguistic description. Reed *et al.*, (2016) proposes a method for automating the process of synthesising realistic looking images from a textual description. This utilises the powerful and discriminative text feature representations with the use of deep Generative Adversarial Networks, GANs. Their work proposes a novel GAN architecture in which the pixels in the image are generated from the visual concepts, words and characters within the initial description. This work is able to generate lifelike images of pictures and birds with incredibly small details, such as petal and stem colour as well as beak and crown styles.

3.5 Robot Image Understanding and Integration

There are many works in which a humanoid robot utilises its vision capabilities for scene understanding, allowing it to determine where, and what certain objects are. There is however very little work into the combination of natural language description of scenes, which could play a large role in the aforementioned works.

Robots have played many roles within Artificial Intelligence, including within the domain of human computer interaction. Hameed, (2016) proposes a method which utilises facial detection and recognition to learn user profiles, before communicating with the users based on a chat functioning neural network in order to learn a person's basic information, hobbies and interests in a bid to learn and store as much as possible

Related Work

to aid within this robot interaction domain. The downsides to this methodology are in regards to the time and resources required before the robot can converse with the desired user.

Johnson-Roberson and Bohg, (2011) propose a methodology in which robot scene understanding is presented, this work has the aim of allowing the robot to count the number of separate entities there are within the image, and then describe them with the use of multiple adjectives (attributes). This is achieved with the use of segmentation algorithms. Once this process has been conducted, the objects and scenes can be appropriately labelled and stored within memory. The human in the loop approach allows the proposed novel HCI framework to combine description and dialog frameworks within state-of-the-art computer vision systems, that can allow the human to intervene allowing for rapid training.

Ye *et al.*,(2016) state that for robots to have true capability to interact with their scenes, it takes a lot more than scene or object classification. The robot should have a functional understanding of the image of the scene. To achieve this Ye *et al.*, (2016) propose a two stage deep learning based framework. This initially follows similar processes to extract region proposals and CNN based feature extraction.

3.6 Caption/Description Dataset

3.6.1 Image

There are an increasing number of datasets that are designed and released for the purpose of image caption/description generation. These generally provide many images each paired with multiple image captions, depending upon the dataset other potential beneficial aspects are also labelled and provided, such as object locations and classification labels. These are discussed further below:

Possibly the most popular image caption dataset is MSCOCO (T.-Y. Lin *et al.*, 2015). This dataset is extremely diverse providing annotations for image context, object recognition and segmentation as well as 5 captions for each of the 300,000 images.

The PASCAL VOC (Everingham *et al.*, 2010) is another popular dataset, and is generally considered slightly easier when it comes to object classification, as this dataset only contains segmentation and classification labels of 20 object classes. The dataset does however contain action labelling which is something many others do not

Related Work

include. The pascal sentences is a collection of 5 user captions per image composed from Amazon Mechanical Turk workers. In regards to newer datasets, this could be overlooked in regards to deep learning due to the relatively small number of available classes.

The Visual Genome (Krishna, Zhu, *et al.*, 2017) dataset is an incredibly large dataset, similar to the very popular ImageNet dataset. This dataset was proposed in the Dense Captioning work presented earlier.

Flickr is a social media site in which vast numbers of images are posted, annotated and stored by the user. These annotations are not guaranteed to be 100% representative of the image contents, or maybe too specific for use in the image captioning domain. For example, an image caption could be a user memory, or the caption could describe specific places or people, which could mean that the images and its features would be too specific to be generalised over an entire collection.

FlickrStyle10k (Gan *et al.*, 2017) is a subset of the Flickr30k (Young *et al.*, 2014) dataset, that contains stylised textual information, such as romantic or humorous image captions. The training and validation set is paired with one humorous and one romantic caption, and the test set is paired with five of each stylised caption.

The SBU (Verma and Jawahar, 2014) dataset collects 1 million images from Flickr, and collects its associated user entered caption. Due to the nature of this dataset there is no validation or test sets. However, the Flickr8k/30 datasets are also built upon the Flickr website and provide reliable and correctly annotated image captions/descriptions, again collected from Amazon Mechanical Turk workers. There are two datasets as suggested earlier each with 8000, and 30,000 images respectively.

NYUv2 (D. Lin *et al.*, 2015) is an RGB 3D parsing dataset that is used for scene description. The dataset is equipped with 3D images of a given scene collected from a Microsoft Kinect sensor, this area is out of scope for this research. However, each image in the dataset is equipped with a single detailed and descriptive sentence regarding the entire contents of the image and a scene label. Although this sacrifices the multiple reference captions, it provides a challenge to scene understanding and natural language description.

Related Work

3.6.2 Video

The research for image description generation has been extended to describe small video sequences. This has been achieved by Yao *et al.*, (2016) through the application of a 3D CNN that considers spatio-temporal information. This work adopts a traditional CNN to collect a single feature representation from videos. This work uses LSTM within the RNN to generate sentences. The 3D CNN is used to collect the local motion descriptions of short sequences, these features are then combined with the features from the normal 2D CNN.

Venugopalan *et al.*, (2015) also describes video sequences through the use of large scale deep CNNs combined with RNNs. The networks were originally trained on static images but fine-tuned on video description. A potential limitation to this method is that it generates a very simple single caption, which may not contain sufficient details to describe a video.

The dense captioning events within video (Krishna, Hata, *et al.*, 2017) also introduces a new video caption dataset, ActivityNet, this dataset contains 20,000 videos accumulating 849 hours of video content, these are paired with a total of 100,000 descriptions that are also equipped with start and end times of the associated caption.

3.7 Summary

To conclude there are many research works and domains that are required for successful image captioning. The general deep learning concepts have been in the machine learning field for many years, and are not relatively recent. However, the widespread use and increase in interest is current.

There have been many related works which focus on the generation of image captions through similar framework architectures, with small variances, such as CNN + RNN. There are also a number of works in which the fundamentals or the general concepts of the layers, connections and inputs are changed in order to compete with existing state-of-the-art methodologies.

Related Work

Related work	Methodologies	Contributions	Limitations
Ren	Region Proposal Network (RPN) – Allowing accurate and near cost free region proposals. Faster and more reliable than methods such as Selective Search.	Constructing a RPN, which is a fully connected network trained end-to-end specifically on region proposals. Proposing a deep model that alternates between fine-tuning the RPN and the object detector.	Requires more training data than existing methods. Nature of training could restrict the use case on different test sets.
Xu et al.	Saliency/Attention model – Focusing on a region similar to that in which the human vision behaves	Introducing both soft and hard attention mechanisms, as well as showing how ‘where’ and ‘what’ can be used to gain insight and interpret the results from the framework by visualizing where the model was ‘looking’.	Intended to overcome the limitations of whole image feature extraction. However, generally focuses on one large central region. Smaller details could still be missed,
Johnson et al.	Regional description – Applying caption techniques to individual image regions	Introducing a dense localization layer that can be implanted into existing CNN models. Introducing a new large-scale dataset (i.e. Visual Genome).	The method in which individual regions are described, means that each description is independent and may not necessarily be fully coherent in regards to a full description/
Tan and Chan	Phrase based LSTM – Encoding sequences of phrases and words	Proposing a novel phrase based LSTM in which the image is encoded in three stages, i.e. chunking of the image, phrase composition as a vector representation and encoding the sentence based on the image, words, and phrases.	Similar to the research performed throughout this thesis. The generated outputs may contain different, yet similar meaning attributes or labels. This is a limitation in regards to the used evaluation metrics.
Fang et al.	Determining salient content and knowing which image contents are interesting or novel using contextual common-sense knowledge.	Re-ranking word detectors that capture global semantics.	Human judgements score the results with a smaller improvement despite a large BLEU score increase.
Tran et al.	Rich description – Adding specifics to image, such as person and location	Presenting a caption model for open domain images, which utilizes a composite approach. Enriching existing frameworks with visual concepts such as landmarks and celebrity identification.	Incorrect classifications within the visual classifiers could result in caption more irrelevant than a misclassification from a typical classifier, e.g. classifying a specific individual as someone else.
Sugano and Bulling	Employing gaze annotated image inputs to generate gaze assisted captioning	Providing an analysis of the relation between object and scene recognition models and human gaze, as well as presenting a novel gaze assisted attention framework.	Similar to the limitations of Xu et al. in which the attention models used can still miss very relevant and potentially required image information.

Table 3.1. This table highlights some of the most notable works and their contributions within the image captioning domain and its required relating domains.

Many of the related works focus on novel model architectures, or utilising layers and modules in novel ways. This achieves a great deal of success when images and captions are trained and tested within domain. Out of domain image testing is rare within this field, however due to the increasing complexity of the newest paired

Related Work

datasets this may reduce the impact of this downside on these particular model architectures.

Cross domain is something which is not generally explored within this field as datasets are becoming more accessible and easier to collect due to frameworks and facilities such as Amazon Mechanical Turk and Crowdfunder.

Recent research from MacLeod *et al.*, (2017) encourages scepticism within these machine learned image captioning techniques. This is due to the nature in which people are exposed to them, for example, social media can automatically annotate images for blind and visually impaired users, and these users place a lot of trust in these captions being correct. However, these systems are trained and tested with a sighted individual as their target audience, as the generated outputs are compared with that of sighted persons. Visually impaired users cannot use the image as a reference point, any experiments with these individuals would be without control. To overcome this, this research recommends generating captions that reinforce the possibility that the captions could be incorrect, using negative framing. This style of captioning allows visually impaired users to use their own judgement, and could encourage 2nd opinions making their experience more pleasant and accurate. The proposed stages of this research and each key stage within the development are further discussed in the following chapters.

Related Work

Chapter 4 Composite Deep Learning Image Description Generator

The following chapters within this thesis introduce and explain the proposed systems and research conducted. This chapter takes an initial exploration into image captioning, utilising powerful object detectors, trained Support Vector Regressors (SVR) attribute prediction both combined with a simple yet effective template generation framework. This pipeline forms a powerful composite image description framework capable of outperforming many existing captioning systems. The following Chapters 5 and 6 introduce an improved and refined description framework as well as a novel end-to-end style description generation architecture.

4.1 Introduction

The first proposed system therefore aims to deal with the above challenges by generating detailed description of image contents using natural language. It will describe the user's immediate environment as well as information about the user, provide alerts and questions as well as general descriptions, through the application of people and object detectors. Most importantly, our research focuses on a local region based approach, which improves over holistic techniques for scene classification and thus relates more specifically to image regions of people and objects in a given image in order to recover more detailed information of the overall image. Our approach differs from state-of-the-art comparable methods which relied on holistic approaches to collect features from the whole image rather than from image regions which may potentially lose details relating to the most important aspects in a scene. Proposing a

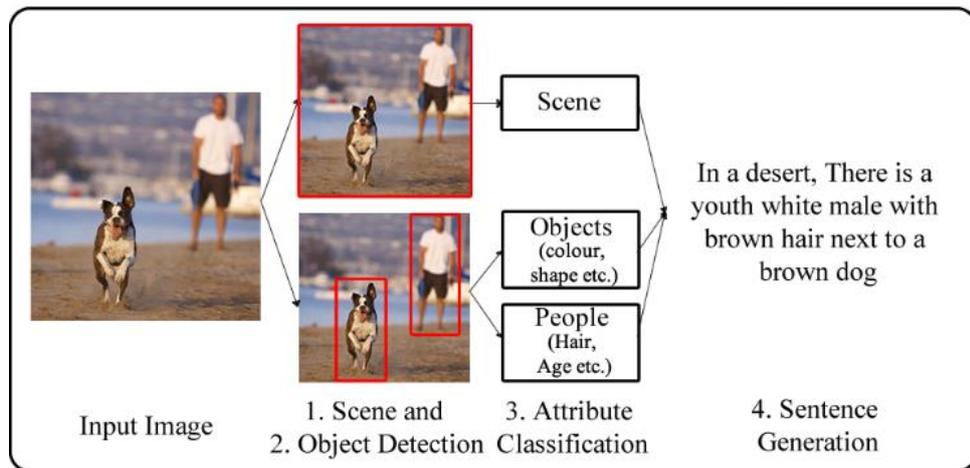


Fig 4.1: Proposed system's pipeline including processes from object, scene detection and classification, attribute classification and sentence generation

Composite Image Captions

local region based approach minimizes the likelihood of missing important information. Through the use of object detectors and classifiers, our approach enables relationships and attributes of detections to be described in detail.

The proposed system consists of 4 main stages, shown in Figure 4.1: (1) scene classification, (2) object detection and classification, (3) attribute learning and (4) sentence generation, as shown in Figure. 4.1. First of all, in step 1, the system collects machine learned holistic image features from large scale Convolutional Neural Networks (CNN) that are computed at test time to classify the scene and provide some additional descriptive data for the subsequent stages of the system pipeline. Secondly, in step 2, object detectors are implemented with the use of a deep CNN to locate people and objects within an image, and provide bounding boxes and an object detection label. The outputs from this step are also collected for implementation of fall and hazard detection to generate alerts. In step 3, new features extracted by the CNN from step 1 are subsequently used for attribute learning. This allows for classifying attributes of the local regions using multiple trained Support Vector Regressors (SVRs) creating a set of descriptive attribute labels relating to each specific object region. e.g., this enables the system to describe a region containing an apple as ‘a smooth green apple’ rather than the basic classification label, ‘apple’. The same principle has been applied to the person description, e.g. to extend the classification label of ‘a person’ to ‘a young white male with brown hair’. The system overall identifies 26 and 25 attributes respectively for people and object description. Finally, in step 4, a template-based sentence generation method is applied to concatenate the previous outputs into one or multiple natural sentences.

The proposed system makes use of pre-trained deep networks, and fine-tuning for object and bounding box regression trained on the ILSVRC13 (Krizhevsky, Sulskever and Hinton, 2012) dataset. The model can classify 200 individual object classes, and is trained on ~400,000 images. The data is also annotated with Pascal (Everingham *et al.*, 2010) style bounding boxes, that are used for the training of the bounding box regressors. This CNN is paired with 200 SVMs, one for each possible classification result from the data. For attribute classification, the algorithms are trained on PubFig (Kumar *et al.*, 2009) for human attributes whereas a subset of the ImageNet (Russakovsky and Fei-Fei, 2012) dataset is applied for the training of object attributes.

Composite Image Captions

Research contributions in this chapter are summarized below:

- This research utilizes local image regions to initialize image descriptions rather than focuses on single holistic features of an input image. For example the existing work of (Karpathy and Fei-Fei, 2015; Xu *et al.*, 2015)
- It generates more descriptive labels and attributes for both objects and people for subsequent sentence generation when compared to dedicated attribute prediction works such as (Bourdev, Maji and Malik, 2011; Dhar, Ordonez and Berg, 2011).
- The system possesses the capability to successfully describe images from outside the domain of training. Most state-of-the-art previous methods were only applicable on the datasets that they have been trained on. By combining multiple datasets, this system can be used on a much wider variety of images for testing.
- It combines the detection and alerts for fall and hazards with the image description problem.

The proposed system has also been integrated with an intelligent user interface fronted with a 3D intelligent conversational agent to conduct health well-being monitoring of the elderly. This experiment is intended to show a use case of this style of system if it were to be implemented into a care environment. Utilising methodologies discussed further in this chapter, the designed 3D agent is equipped with the image processing capability of the presented work via a computer attached camera, to detect users' falls and provide conversation and web lookup functionality. Equipping the intelligent agent with the above discussed system functions such as object and hazard detection, attribute classification, and background scene description generation, the system is

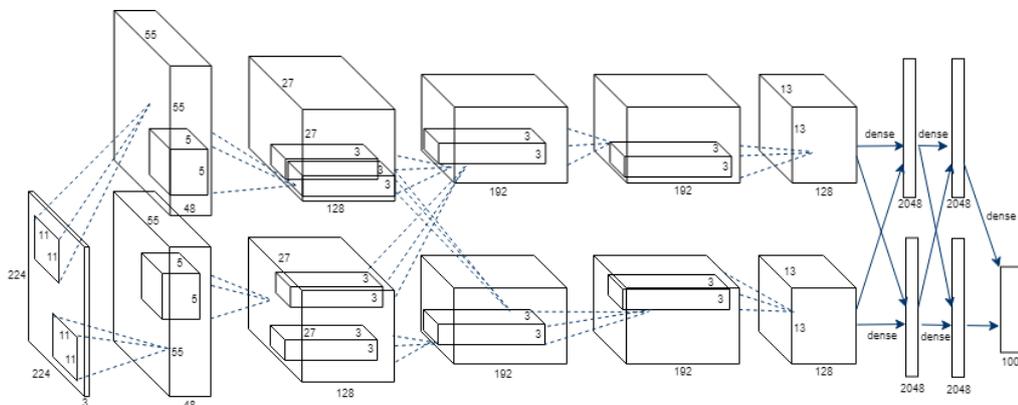


Figure 4.2: Illustration of the CNN implemented, which shows the network architecture across multiple GPUs

able to gain information of users' direct surroundings to warn of hazards and generate alerts if critical events such as falling have occurred.

4.2 The Proposed Image Description Generation Framework

As discussed earlier, most previous methods that attempt to annotate or caption images rely upon holistic methods and deep learning techniques to collect features and information from the entire image (Vinyals *et al.*, 2014; Karpathy and Fei-Fei, 2015). This implies that information could potentially be lost. In order to overcome this problem, our proposed research uses local regions classified using R-CNN under diverse circumstances. In addition, many previous state-of-the-art applications (Karpathy and Fei-Fei, 2015) depended largely on the fact that training and testing images were extracted from the same domain, otherwise the results generated can become heavily compromised. However, in our work, using the local R-CNN based object detectors enables cross-domain images or images outside of the scope to be tested upon.

Therefore, in this research, we propose an image description generation system to generate detailed sentences through the use of scene and object detection and classification, local attribute labelling and template-based sentence generation for the description of a given scenario in static images.

4.2.1 Object Detection and Classification

The object detector implemented in this work is based upon the R-CNN. The R-CNN is a traditional CNN (Krizhevsky, Sulskever and Hinton, 2012) (but with added outputs that can predict bounding box coordinates). In this research, the R-CNN is trained upon the ImageNet ILSVRC13 dataset that can detect, localize and classify 200 object categories. This CNN network structure consists of eight learned layers - five convolutional layers and three fully connected layers (see Figure. 4.2).

The first stage of the proposed system is to detect objects within an image using R-CNN. It applies an SS algorithm (Sande, 2011) to collect regions of images that possibly contain objects and/or people. The input image is first scaled to a maximum width of 500px. The SS algorithm combines the advantages of an exhaustive search and a segmentation technique. It utilises size and appearance features to generate

Composite Image Captions

locations. A greedy algorithm iteratively groups similar regions and measures similarity between regions and their neighbours. This process is repeated until the whole image becomes a region. The similarity between regions a and b in the greedy search algorithm is defined as:

$$(2) \quad S(a, b) = S_{\text{size}}(a, b) + S_{\text{texture}}(a, b),$$

which returns a value between $[0, 1]$. In Equation (2), $S_{\text{size}}(a, b)$ is a proportion that a and b both occupy, which persuades small regions to merge, and $S_{\text{texture}}(a, b)$ is the intersection between SIFT like measurements.

The regions are warped to fit the required input of 227×227 . For each region proposal, a feature vector with the size of 4096 is extracted. The features are extracted with the Krizhevsky style CNN with the final fully connected layers removed, providing a network output consisting of very discriminative features. During testing, each feature vector is passed to each of the 200 SVMs representing each object class to deliver a score for each region. Non Maximum Suppression (NMS) is used to dispose of a region if the region has an Intersection over Union (IoU) with a comparatively higher score than a predefined threshold.

The R-CNN implementation (Girshick *et al.*, 2016) is also modified in our research to crop and store images, bounding boxes, their respective classification categories and confidence scores. The classification labels are concatenated to form a list of detections. The detections and the regions are then cropped and labelled again, simply as an object or a person. This new label is simply used to determine which set of attribute detectors (e.g., object or person attribute detectors) the image will be subjected to.

4.2.2 Scene Classification

The scene classification is implemented with the application of the hybrid Alex-Net deep CNN (Krizhevsky, Sutskever and Hinton, 2012), which is trained to detect 1183 categories, including 205 scene labels and 978 object categories. This network, however, cannot be applied as an object detector in our work as this hybrid CNN network does not provide the object localization. This network, however, is implemented on a GPU and takes approximately 0.2s per image on our GPU. If speed is not a requirement, this network can be implemented using ‘cold brewed’ Caffe (Jia

Composite Image Captions

et al., 2014), i.e., CPU mode classification, which takes ~2 seconds per image to produce an output.

Scene classification is used to increase the descriptive nature of the system. Therefore, rather than simply stating indoor or outdoor, this research is able to attach a semantic label to the scene to provide a more refined description. The system can identify a total of 205 scene labels using the above hybrid deep network, which provides efficient accuracy for a relatively low computational cost while dramatically increasing the descriptive nature of the sentence.

4.2.3 Hazard Detection

Object detection in this work has been further extended to deal with hazard identification. As mentioned earlier, in the elderly care situation, identifying potential hazards could aid in the prevention of falls and alert care providers. To achieve this, the object classes and bounding box coordinates are obtained from previous object detection and classification, and collected for the implementation of hazard detection. For hazard detection, we assume that the camera in a care home environment would be static.

For the elderly care application, the system hosts a 3D intelligent conversational agent that can provide speech-based interaction, this is further highlighted in the evaluation section of this chapter. Users may sit close to or stay away from the camera. When a user sits in front of the webcam, the system must not label held objects as a hazard. To overcome this potential issue, the hazard detection will not be triggered if a person is detected in a large portion of the background viewable scene whereas the hazard detection would be enabled when the camera is surveying an overall scene that may not contain people. This assumption enables a threshold value to be set initializing where the floor meets the wall. This however does restrict the use to fixed terminal or camera locations. An example of this is shown in Figure 4.3. If the system was portable, for example on a laptop, additional methods utilising possible edge detection algorithms would be required as the floor may not be in the same location relative to the angle and positioning of the webcam or humanoid robot head. The object detectors are operable on the overall scene image as usual. From the bounding box coordinates, the top of the object can be determined. If this is below the floor threshold, it is

Composite Image Captions

determined to be a hazard and subsequent attribute classification provides more details of the hazardous object.

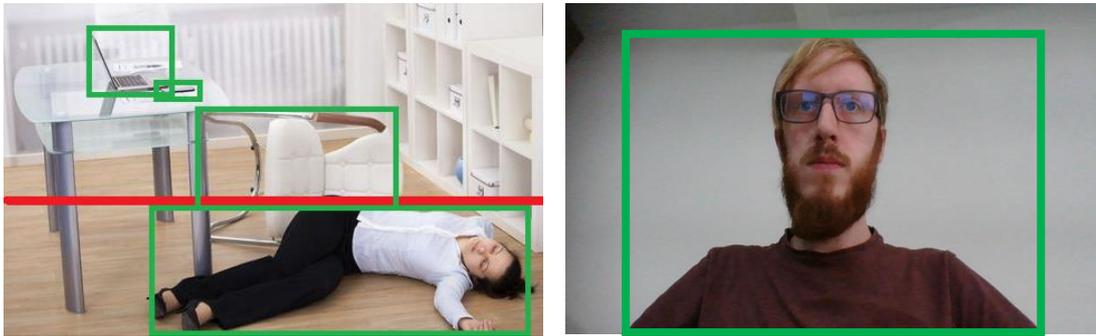


Figure 4.3. The left image shows where the floor threshold is set on this image. Identifying a person below this triggers the fall detection framework. The objects on the desk are not below this threshold, therefore will not be triggered or identified as a hazard.

The right image would not trigger the hazard or fall framework due to the user occupying more than the determined percentage of the image.

To overcome the issue that some objects should be present on the floor such as free standing lamps or tables, the system has simply been implemented with the use of an exclusion list constructed based on common sense knowledge identifying which objects will and will not be labelled as hazards. As our system implements R-CNN, the framework can currently detect 200 object categories, some of which are expected to be located on the floor, each of the 200 objects has been labelled with a ‘True’ or ‘False’ tag, based on general hazard or non-hazard cases respectively. When the system detects a hazard, it first consults this exclusion list to determine if this ‘hazard’ should be reported during sentence construction or left out.

4.2.4 Fall Detection

The R-CNN based object detector in this research is also able to detect people within images. Utilizing a similar methodology employed for hazard detection, the system is developed to detect a fallen person lying on the ground and thus responds accordingly. To achieve this, a person is again labelled as a hazard in the aforementioned exclusion list, but is dealt with differently to the existing ‘object hazards’. If this falling scenario occurs, a dedicated sentence, such as asking the fallen subject if any assistance is required, alongside all other constructed sentences will be used as outputs, see Figure 4.3. The fall and hazardous object detection could be applied to an elderly care environment to generate alerts for fall prevention or inform necessary individuals if critical events such as falling have occurred.

4.2.5 Attribute Learning

To increase the descriptive nature of the sentences, high level information such as attributes are employed to add details and context to the basic object labelling. Attributes are required to describe colour, texture and shape of detected objects as well as to estimate a person's age or describe his/her hair colour, gender and other equipped accessories. To achieve attribute classification to aid the sentence description, regional features must be first extracted.

Handcrafted features such as SIFT were popular in computer vision applications. Instead of applying handcrafted features, in this research, machine learned features are used in order to produce better trained algorithms. In order to extract such machine learned features, another faster large scale CNN (Chatfield *et al.*, 2014) that has previously been trained on ImageNet is utilized. This network has five convolutional layers and three fully connected layers. This network is modified to produce a large, discriminate feature vector of a given image. In this research, the given image is a cropped image produced from the bounding box regression performed earlier in the pipeline. A smaller set of 1024 features is collected for attribute learning rather than the existing 4096 features extracted for object detection due to the balance of training and testing time of the employed attribute regression methods. This is achieved via running the same CNN feature extractor as earlier within the system pipeline, however equipped with reduced dimensionality fully connected output layers.

Results obtained from this small set of features were as competitive as using the full set of features, but with much improved computational cost. It indicates that the comparatively small set of 1024 features is discriminative enough to enable the effective training of multiple Support Vector Regressors SVRs for attribute classification. For object and human attribute classification, multiple SVRs were used, with one SVR dedicated to each semantic attribute. There are 25 semantic attributes for objects, which lead to 11 SVRs for colour, 2 for pattern, 4 for shape and 8 for texture. Human attributes, on the other hand, are more diverse. There are originally more than 70 attributes in the dataset. In this research, 26 attributes are used based on a subset of this dataset. This subset is chosen based on the textual label only and manually chosen by the author. The human attribute classifiers are grouped into 5 categories representing hair colour, age, gender, ethnicity and accessories. Each category consists of a diverse number of SVRs. The SVR with the highest output

Composite Image Captions

probability score from each group is taken as the attribute that best represents the person within that group. E.g., there are 5 SVRs for hair colour prediction, i.e., 1 for blonde hair, 1 for brown, etc. The same extracted input features are passed through each of these 5 SVRs, with each returning a value between -1 and 1. The highest prediction value is used to describe the subject's hair colour during sentence generation. This process is repeated for other attribute group, and unless all SVRs within a given group produce a low certainty, there would be 5 attributes used to describe a person, making the descriptions more detailed.

The MATLAB implementation of LibSVM (Chang and Lin, 2011) is used for the training of SVRs. The SVRs utilized are nu-SVRs using the Radial Basis Function and were trained with a subset of the object or people datasets introduced below.

4.2.6 Attribute Datasets

The training of the attribute classifiers is achieved through the use of two datasets discussed below:

4.2.6.1 Human Attributes

The dataset used for training the person attribute detectors is the PubFig dataset (Kumar *et al.*, 2009), which consists of approximately 16,000 training images with 200 celebrity faces. The dataset consists of a total of 73 attributes - only attributes relating to a person's age, gender and general appearance are taken into account. The whole dataset has many attributes that are either irrelevant or potentially too detailed for use in our given scenario. The training data for these attributes contain images that close crops a person's face, meaning that attributes relating to ear rings etc. can be classified. This is due to the fact that the ear ring would take up a much larger proportion of the image, opposed to the image regions generated in this chapter which contain a human as a whole body region. The dataset has a number of links that have been removed, modified or no longer active, which leads to the overall number of images used during training to be around 10,000.

4.2.6.2 Object Attributes

Attributes for objects from ImageNet (Russakovsky and Fei-Fei, 2012) are collected for describing properties of objects. ImageNet partitions its images into multiple categories called synsets. Attributes are available upon a relatively small subset of ImageNet of around 10,000 images from 400 synsets. Each synset has 25 images that

Composite Image Captions

are each annotated with 25 attributes. These attributes are grouped into 4 categories, including Colour, Pattern, Shape and Texture. This small dataset is provided with bounding boxes highlighting the areas where the attributes are present.

4.2.7 Prepositions

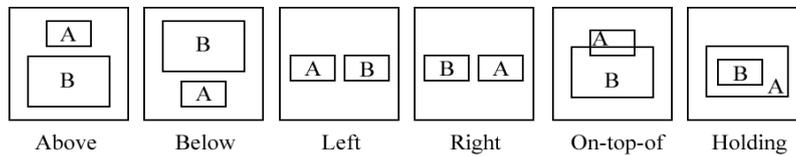


Figure 4.4: Spatial relations between multiple humans and objects.

Relationships between people and/or objects in the scene can be an important aspect to describe within a given image. Multiple relationships are taken into account in this research (see Figure. 4.4). These relationship functions are mainly applied to detect pair-wise relationships between people-people and people-object. These relationships are computed through the use of the bounding box coordinates. Each person has all of the relationships computed in relation to each other object and/or person. The chosen relationship is simply the output with the highest score. Moreover, some of these relationships can be modified based upon the objects in question. In an attempt to use these relationships as actions, if a person is ‘on-top-of’ a horse or a bike the relationship will be modified to ‘riding’. Similarly, this can be applied to a person wearing a hat or helmet, etc.

4.2.8 Sentence Generation and Construction

Image Description New
Image Description New 1/20



Describe the above image in a detailed sentence or two. E.g. Scene, persons age, hair colour, object colour, texture etc

Continue »

Figure 4.5. Showing an example set up of the sentence collection methodology utilised in this chapter’s research.

Composite Image Captions

The final stage in the pipeline is the creation of sentences, and utilises a similar approach to the work of Kulkarni, Premraj and Ordonez, (2013). The current system implements template based sentence generation that utilises the detailed labels collected from the combination of the scene detectors, the object and person detectors and the related attribute regressors.

Depending upon the situation, the system is able to output between 1-3 sentences for each image. The system always first produces a sentence that describes what it ‘sees’ in a scene, i.e., people and objects, their descriptions, relevant attributes and relationships to other items of interest within the scene. This sentence can be one of a number of template structures depending upon how many people there are, or how many objects or a combination of the two.

The next two sentences are generated depending upon results of further analysis and can be skipped if the analysis does not return the information that these additional sentences require. These sentences are generated based on the determination if a user has fallen and/or hazardous objects are detected on the floor or in an environment they should not be residing in. For example, if the system determines that there are multiple hazards that need to be removed, or should not be located in a room at all, it produces sentences advising on what to move, or even asks related questions (e.g., to see if the test subject is aware of the hazardous situation).

Attribute	Semantic Label
Person	True/False
Gender	Male/Female
Age	Youth – Senior
Accessories	Glasses, scarf, necktie/necklace, etc.
Hair Color	Brown, Blonde, Black, etc.
Ethnicity	White, Indian, Chinese, etc.
Objects	200 object categories from ILSVRC13 including Car, Dog, Orange, Desk, etc.
Scenes	205 scene labels from MIT scenes including playground, living room and bedroom, etc.
Relationships	Above, Below, Left, Right, Holding, and On-top-of.

Table 4.1. Information captured for sentence generation

The final ‘optional’ sentence to be generated is related to coping strategies for people falling. If the system detects a fallen person lying on the floor, the system outputs a sentence that may alert those nearby, as well as collects scene information that could be used as the source to link with how the fall occurred such as hazards and/or to provide information about the scene for easier locating of the fallen individual. For the fall and hazard sentence generation, the system chooses at random one out of multiple sentences in order to diversify sentence description generation, and prevent the

Composite Image Captions

description from all being similar sounding, which is a downfall that template based systems can succumb to.

Information considered during sentence construction is summarized in Table 4.1.

4.3 Results and Evaluation

Our system is trained on images of objects and people but not necessarily the whole images whereas other existing methods required the whole images as inputs. Which could lead to local and specific detailed information could be missing from the high level outputs produced from such systems. Moreover, these state-of-the-art related systems were trained specifically on a dataset that they were tested on, e.g., Flick8k, Flickr30k and MSCOCO, whereas our system is trained upon multiple seemingly



Human Sentence 1: Two off-road cyclists in red riding in a forest

Human Sentence 2: 2 men on cycle-cross cycles racing in a competition in a Forrest setting

Human Sentence 3: In front , a middle aged female wearing a red jacket, helmet and sunglasses riding a bike through the woods with brown leaves on the ground. Behind a middle aged male with a brown beard wearing a red jacket and silver helmet riding a bike. There are two males standing nearby and a female with a young child in the distance.

Figure 4.6: Showing the diversity in human collected sentences

General Image



Machine Gen: In a sandbar, there is a furry dog, and a black dog

NeuralTalk: a group of people riding horses on a beach

Human Gen:

2 dogs, one black one white playing in shallow water on the sandy shore

In a lake a white dog chases a black dog out of the water

Two dogs, one brown, one white, playing by the beach in the water

Healthcare related Image



Machine Gen: There is a child white male with black hair in a plaza. There is a person on the floor. Do you require assistance?

NeuralTalk: a man standing in front of a door holding a skateboard

Human Gen:

A middle aged woman has fallen in the street carrying her shopping outside of some shops

A woman in a pink jacket falling over in a street with shopping bags

Woman falling on paved ground in front of shop while holding carrier bags

Figure 4.7: Example outputs from general and healthcare images

Composite Image Captions

unrelated datasets (i.e. the ILSVRC13 dataset for object detection (Russakovsky *et al.*, 2015), the object attribute dataset from ImageNet and the PubFig dataset for object and human attribute learning respectively) and tested upon randomly selected images from multiple datasets. In addition, empirical results indicate that the sentences generated by this research are a lot more descriptive than those provided by Flickr or COCO. This indicates that any BLEU (Papineni *et al.*, 2002) or ROUGE (Lin, 2004) scores generated by our system may not be directly comparable, as the BLEU score has been proven in recent research to dramatically favour shorter captions. To this end, we employ the comparison between machine generated sentences and some human generated ‘Ground Truth (GT)’ sentences for evaluation. A BLEU score is therefore produced to measure the similarity between the machine generated sentence and human annotated ‘GT’ sentences. The BLEU score has been commonly employed as a machine translation measure and can be applied to determine the semantic similarity between sentences. Moreover, human performance has been stated in the literature to achieve a performance benchmark of 69% on the BLEU score (Vinyals *et al.*, 2014).

The evaluation of our system is conducted with a small set of 45 images randomly selected from the Flickr8k and the Pascal VOC 2012 datasets which are unrelated to the datasets employed for training to show the diversity of our approach. To collect user descriptions of images without the use of Amazon Mechanical Turk, this research utilised Google Forms. An example of the interface provided to our human annotators is shown in Figure 4.5. These images also do not include scenes of people falling or hazardous objects. The test images have been shown to three human evaluators, who were asked to describe each image in detail, i.e., relating to a person’s gender, age and accessories, etc., as well as object colour and texture, etc. This allowed for a direct comparison between long detailed human annotations and our systems generated outputs. These generated descriptions by human annotators are used as GT for comparison with sentences generated by our system in order to produce BLEU scores.

Method	BLEU Score			
	<i>Flickr and VOC</i>		<i>Detailed Sentences</i>	
	<i>All Sentences</i>	<i>Detailed</i>	<i>Flickr</i>	<i>VOC</i>
Avg. Human	0.69			
This research	0.46	0.30	0.31	0.29
NeuralTalk	0.59	0.27	0.25	0.30

Table 4.2. Bleu score comparison between average human judgements, this research and NeuralTalk.

Moreover, we employ the popular and public models of the state-of-the-art related research, NeuralTalk (Karpathy and Fei-Fei, 2015) for comparison with our approach.

Composite Image Captions

NeuralTalk was also tested upon the same 45 images that this proposed method was tested upon. BLEU scores have also been generated in comparison to human generated sentences. Their system was trained upon MSCOCO. Experimental results indicate that their algorithm shows great limitations and does not perform satisfactorily when tested upon images from unrelated datasets. This work has not been trained on any of these datasets either but shows improved performance when tested with the test set images from different domains. The evaluation results are demonstrated in Table 4.2.

In Table 4.2, ‘all sentences’ indicate the three GT full length sentences provided by the three human evaluators whereas ‘detailed sentences’ refer to the longest or most detailed sentences among the three human generated sentences. This could be as little as one annotation and as many as all 3, if the detail and length was present within the descriptions. Figure 4.6 shows the diversity in the collected images for the example image. The BLEU scores are generated by comparing ‘all sentences’ and ‘detailed sentences’ GT against the machine generated outputs for all the test images. The ‘detailed sentences’ GT is also used to generate BLEU scores for images from Flickr and VOC datasets respectively.

Overall, when tested upon ‘all sentences’ GT, NeuralTalk achieves a higher BLEU score. This leads to the assumption that these images are very much like those contained within MSCOCO that NeuralTalk has previously been trained upon. Therefore, it causes a spike in its BLEU score. Moreover, the GT sentences collected may vary in detail, which has caused some of the BLEU scores to favour other shorter captioned work, although the description generated by our system has also been correct. When evaluated against ‘detailed sentences’ GT, our system outperforms NeuralTalk for the overall test set images from both databases as well as images extracted purely from Flickr whereas it shows equivalent performance for images from VOC to that of NeuralTalk. Inspection of images indicates that the VOC images are less complex than the Flickr images. This implies that even the most descriptive sentences from human annotators are short which may cause slightly higher BLEU scores for NeuralTalk, despite the sentences generated by our system being correct.

	BLEU Score
	<i>Falls and Hazard</i>
This research	0.38
NeuralTalk	0.28

Table 4.3. Bleu score for images with fall and hazard situations.

Composite Image Captions

Generally, the above evaluation results could indicate that the more complex an image is, the better the generated sentence description by our system and the better BLEU score. This has also been further proved by using images randomly selected from the web containing falls and hazardous situations in the following experiment.

A second experiment (Table 4.3) has also been conducted to evaluate our system performance with images containing fall and hazardous situations. We randomly extract a small dataset of 6 images from web resources for testing, evenly split between fall and hazardous scenarios. These test images are again from new, out-of-scope application domains. GT sentences are also collected from the three human annotators with a BLEU score subsequently computed to indicate system performance.

Our system shows great robustness, by scoring well when tested with these images from other indirect sources with hazards and falling cases, and outperforms NeuralTalk by a significant margin of 10 BLEU score. Example output comparisons between this research and NeuralTalk for images from both general and healthcare domains are provided in Figure. 4.5. All experiments in this chapter were implemented on a computer with the Ubuntu OS an Intel XEON, 8GB DDR4 RAM and an NVIDIA Quadro K5200. The models were implemented with a combination of Caffe and Keras. 3D modelling software Blender was used in the creation of the agent. For the agent, the communication was integrated with web sockets in python, utilising API end points for google maps and Wikipedia integration.

4.4 Experiment with an Intelligent Visual Agent



Figure. 4.8. Example interface of the intelligent agent experiment.

This system was also utilised in a third and final experiment, that has the benefit to aid disabled or impaired users, especially due to the nature in which this framework

generates descriptions, labels, regions and potential falls and hazards. To this end, the framework presented within this chapter is paired with a user facing 3D intelligent agent. This agent is able to hold general conversation, as well as simple question and answering, with the use of the ALICE framework (*A. L. I. C. E. The Artificial Linguistic Internet Computer Entity*, 1995). An example of this set up is shown within Figure 4.8. As a real-life application of such a system, the agent was equipped with aspects such as the ability to search Wikipedia, emotion recognition and geo-location, in addition to those already in the proposed framework. This set up was tested on a small subset of images from within the IAPR dataset previously described, as well as a small preliminary evaluation with user annotated real-life images. In this experiment, it was found that the descriptions produced by the proposed framework were more relating to the image, and improved user experience in the experiment set up. This system was tested in the wild with the use of a computer laptop and its webcam. The method of testing was simply asking colleagues to converse and communicate with the agent, followed by brief discussion. This overall system performed well, in regard to the implemented similarity metrics, however this initial system pipeline took multiple seconds to generate a system description output. This in a real life care environment is not ideal and would require amendments in this particular use case.

The proposed experiment in combination with the proposed framework and the agent could allow for real use or test deployment within a care scenario. If a user had this system on a laptop, or a fixed terminal within a common area, when capturing and annotating an image, the proposed framework can verbally express concern for potential hazards and raise questions to fallen individuals. Then from the lack of/or the response from an individual, an alert could be passed to the relevant care providers.

4.5 Summary

In this research, we proposed a local region based approach to develop a multi-sentence image description generation system. The region based object detection, attribute classification and relationship identification allow for efficient descriptive capabilities and incorporate regional details into template-based image description generation. Using the local R-CNN based object detectors also enables images outside of the scope to be tested upon. The system has also been equipped with fallen subject and hazard detection to alert danger and risks to aid elderly care.

Composite Image Captions

In comparison to related state-of-the-art research, experimental results indicate that our system is able to produce more detailed descriptions of image contents than those that utilise holistic image features, rather than the regional information captured within this system. Empirical results also indicate that it shows great flexibility and robustness when tested with cross-domain image datasets whereas the performance of other related work is heavily compromised when tested with such unrelated domain images.

Especially, in the realm of elderly care, our system shows superior performance for out-of-scope hazardous and falling situation description generation whereas other related methods fail to generate any acceptable outputs. The issue with the system within this domain is the time in which an output is produced from the initial capturing of an image. Attempts to address the time constraints are in subsequent chapters, as well as the future work section.

Although our system achieves impressive performance, we also identify the following directions for future improvements. We will replace R-CNN with fast R-CNN, because of its impressive computational efficiency and further train the fast R-CNN network with ILSVRC13 200 object dataset to better deal with real-time image description generation tasks. We also aim to further test our system's efficiency by employing a large-scale cross-domain dataset for system evaluation. In the next chapter of this thesis an extension to this method is proposed with a new model architecture to address the image captioning problem. As stated in this chapter, the template sentence generation can lead to similar non-diverse image descriptions. The next stage in this research would be to address this shortfall by replacing this generation methodology. This would require additional or different pre-requisite stages in order for accurate and diverse captions and descriptions to be automatically output. Later within this thesis we also explore the integration of the proposed system with a humanoid robot to provide more personalized services and better deal with the challenging open-ended natural human robot interaction.

Composite Image Captions

Chapter 5 Region-based Image Caption Generator with Refined Descriptions

This chapter introduces the second proposed system for image description generation. In order to achieve refined and detailed descriptions, this chapter proposes a novel local deep learning architecture for image description generation. It employs a regional object detector, recurrent neural network (RNN)-based attribute classification, and a pair of encoder-decoder based RNNs to generate detailed descriptions of image contents. Most importantly, the proposed system focuses on a local based approach to improve upon existing holistic methods, which relates to image regions of people and objects in a given image.

The proposed system consists of four key stages and aims to tackle the major shortfall of generating the image captions as described in the previous chapter: (1) object detection and recognition; (2) attribute prediction; (3) scene classification; and (4) diverse and refined description generation. The overall system architecture is shown in Figure 5.1. In the first stage, an object detector is implemented with the use of a large scale deep Convolutional Neural Network (CNN) to locate and classify people

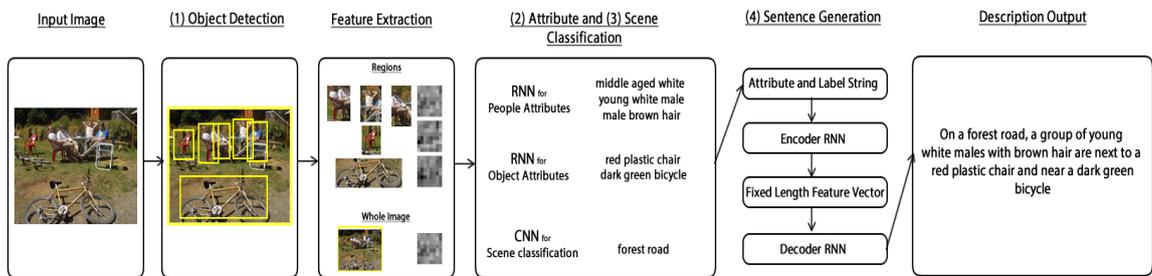


Figure 5.1. The overall system architecture, which consists of (1) object detection and recognition, (2) attribute prediction, (3) scene classification and (4) sentence generation

and objects within an image. This CNN provides localization with bounding boxes and object class labels as outputs. In the second stage, the above CNN is applied to the detected regions to extract features for subsequent attribute prediction. Two RNNs are implemented for attribute classification from the local regions, with one dedicated to human attribute prediction and the other applied to object attribute prediction. In the third stage, machine learned holistic image features are extracted using the same CNN for scene classification. In the fourth stage, the recognized objects, scene, people and

Refined Descriptions

their associated attribute labels are passed to an encoder-decoder structure, which consists of two RNNs to translate detected class (human or object) and attribute labels to full image descriptions.

The main contributions of our research are two-fold, as follows:

A local region-based deep learning architecture is proposed for image description generation, in extension to the previous chapter. In order to overcome limitations of existing holistic methods, it employs a regional object detector, RNN-based attribute prediction, and encoder-decoder based caption/description generation. Especially, the challenge of sentence-based captioning and description generation in this research is treated as a machine translation problem, rather than an image annotation problem. This contribution directly addresses the shortcomings within the research conducted within the previous chapter, although the results of the evaluation are promising, the template based sentence generation and construction methodology proved to be a limiting factor, as most if not all of the generated sentences shared an identical structure, despite methods to limit this. This may work over a small number of images, however if the system were implemented within real world applications, the captions would begin to sound repetitive, providing a less than ideal environment for the user. The methodology within this chapter aims to combat this issue by allowing a dramatically diverse range of sentence and caption structures.

A comprehensive evaluation of the proposed system is conducted using the IAPR TC-12 dataset (Grübinger *et al.*, 2006). The empirical results indicate that the proposed system shows impressive performance and outperforms state-of-the-art related research on nearly all the evaluated metrics. In particular, the system shows great superiority and efficiency when dealing with cross-domain indoor scene images extracted from the NYUv2 (D. Lin *et al.*, 2015) sentence dataset, which is wholly different from the dataset used for training the system and is regarded as a great challenge to existing research due to the nature of the descriptions and its intended use case of 3-D parsing of paired Kinect data.

This chapter is organized as follows. In Section 5.1, we present the proposed system, including object detection and recognition, scene classification, RNN-based attribute prediction, and encoder-decoder based sentence generation. A comprehensive

evaluation is provided in Section 5.4. Section 5.5 summarizes our work and identifies potential areas for extension and future research.

5.1 The Proposed Image Description Generation System

As discussed earlier, state-of-the-art methods largely rely on holistic techniques to extract image features for caption generation (Vinyals *et al.*, 2014; Karpathy and Fei-Fei, 2015). The major drawback of such methods (e.g. NIC) is that local information could be overlooked. In order to overcome such limitations, this research proposes a region-based deep learning approach in extension to the system in the previous chapter, to capture local details by incorporating object detection and recognition, scene classification, attribute prediction and description generation. Moreover, the captioning and description generation challenge is treated as a machine translation problem, therefore an encoder-decoder structure for sentence generation is proposed. The proposed system possesses superiority over existing state-of-the-art methods, especially when dealing with image description tasks for out-of-domain images. We introduce each major step of the proposed system in detail in the following subsections.

5.1.1 Object Detection and Recognition

For robust object detection, we implement the R-CNN object detector. The R-CNN is a traditional CNN but with extra outputs that predict bounding box coordinates. The R-CNN is trained using the ILSVRC13 dataset. It is able to detect, localize, and classify 200 object categories. The R-CNN (Girshick *et al.*, 2016) uses the ImageNet (Russakovsky and Fei-Fei, 2012) database for the training of the CNN (Krizhevsky, Sutskever and Hinton, 2012).

Specifically, the R-CNN employed in this research for object detection applies a Selective Search (SS) algorithm to collect regions of interest (ROIs) from images that possibly contain objects and/or people. A greedy search algorithm then groups similar regions and measures similarity between regions and their neighbours. This process is repeated until the whole image becomes a region. The same greedy search algorithm, (i.e. Equation 2) as in the previous chapter is implemented in regard to the bounding box proposal generation. These generated region proposals are warped to fit the required input of 227×227 pixels.

Refined Descriptions

For each region proposal, a feature vector with the size of 4096 is extracted. During test, every feature vector is passed to each of the 200 SVMs representing each object class to deliver a score for each region. Non-Maximum Suppression (NMS) is also used to dispose of a region if the region has an Intersection over Union (IoU) with a comparatively higher score than a predefined threshold.

The R-CNN implementation by Girshick *et al.*, (2013) is modified in this research to crop and store images, bounding boxes, their respective class labels, and confidence scores. The class labels are concatenated to form a list of detections. The detections and the regions are then cropped and labelled again, as an object or a person. This additional label is used to determine which attribute predictors (e.g., object or person attribute predictor) the system utilizes in the subsequent processing.

5.1.2 Scene Classification

Scene classification is employed in this research to further increase the descriptive nature of the proposed system. Instead of simply stating indoor or outdoor, our system produces a semantic scene label (e.g. a park or a shopping mall) obtained by scene classification, to provide a more refined description.

This scene classification in this chapter is a mirror of the scene classification integrated into the system pipeline in Chapter 4. To recap, this scene CNN employs the hybrid Alex-Net deep CNN (Krizhevsky, Sulskever and Hinton, 2012), that can only classify scene information.

5.1.3 Attribute Prediction Using RNNs

Traditionally, attributes are normally classified using large clusters of classifiers, with one classifier dedicated to one attribute (Bourdev, Maji and Malik, 2011; Sheshadri, Aashish, Endres, 2012). These are typically SVMs or SVRs as used in the previous chapter. Such classifier clusters not only require more resources for training, but also tend to only indicate the presence or absence of an attribute, without any confidence measure.

In this research, we employ RNN-based attribute classification to address the above drawbacks. Research has indicated that RNNs are useful in many areas of machine learning as discussed in chapter 2, including image captioning and machine translation (Bahdanau, Cho and Bengio, 2014; Vinyals *et al.*, 2014). RNNs are used in this work due to their ability to not only classify an arbitrary number of human and object

Refined Descriptions

attributes due to their success with sequence generation, but to also effectively determine which attributes should be reported for a given set of features dynamically. I.e. determining which attributes should or should not be reported, provide a diverse and concise required number of outputs. In this research, two RNNs are used for attribute classification, with one dedicated to human attribute prediction and the other applied to object attribute classification. Splitting the people and object attribute classification aims to reduce the amount of training data required, as well as preventing human attributes from being generated for the description of an object, and vice versa.

Also, RNNs have been used for attribute prediction owing to their significance and flexibility in dealing with natural language processing tasks. The initial comparison between RNNs and other alternative methods also indicates that RNNs show better accuracy than those of other methods such as SVMs for attribute prediction.

Moreover, the RNNs employed in this work are ‘word based’ and adopt a Long Short Term Memory (LSTM) architecture, which are further discussed in Chapter 2 (Graves and Schmidhuber, 2005). At test time, the previous CNN is used to extract image features. The extracted features are then used to predict multiple attributes one by one. The nature of the RNNs and the LSTM ensures that the previously generated attribute is considered when generating the next. This should theoretically stop contradictory attributes from being generated, as the attributes within the training data would all be annotated correctly. The attribute prediction is based on the extracted image features combined with the previously generated words. This process continues to generate relevant attributes until the designated STOP word has been generated. I.e. the generation of this STOP word occurs when the RNN determines that no other attributes could be used to describe the cropped image in question, based on the features and all of the previously generated attributes.

In this research, PubFig (Kumar *et al.*, 2009) and a subset of ImageNet (Russakovsky and Fei-Fei, 2012) are also used for training of RNN-based human and object attribute prediction, respectively. A slightly modified version of AlexNet (Krizhevsky, Sutskever and Hinton, 2012) without classification layers is implemented to extract CNN features for training of RNNs. This network extracts 4096 image features from previously cropped images of the desired objects or people, which are then paired with the attribute labels from the respective attribute dataset for training.

5.1.3.1 Human Attributes

The RNN for human attribute prediction in this research is trained with the PubFig (Kumar *et al.*, 2009) dataset, which consists of ~10,000 images in total. It has 200 unique celebrity faces, each labelled with 73 attributes, such as age, gender, and hair style. The attributes used to describe people in this research are a selected subset of those used in the PubFig dataset. Overall 26 human attributes are chosen for this work, which are shown in Table 5.1. This subset was chosen as it covers a large range of diverse individuals without overcrowding the network. The dataset original included many attributes that are irrelevant in circumstances such as these due to the style of images and the captions desired. For example, ‘bags under eyes’ as the images would rarely be close enough to a test subjects face, as well as attributes such as ‘Fully Visible Forehead, Partially Visible Forehead, Obstructed Forehead’. Some close attributes were chosen as they should stand out on a user’s face such as ‘wearing lipstick’ or ‘wearing sunglasses’.

5.1.3.2 Object Attributes

The ImageNet dataset is widely used in many computer vision challenges. It also has a small subset of images that are fully annotated with object bounding boxes and their respective attributes. This subset consists of around 10,000 images collected from 400 synsets. Each synset represents a group of ImageNet images associated with a specific WordNet (Miller, 1995) ID. Each image is paired with 25 object attributes, and all of these 25 attributes are employed in this research. They are illustrated in Table 5.2.

Overall, in comparison with traditional SVM-based attribute classification, the RNN-based object and human attribute prediction shows great efficiency for the classification of highly associated attributes, and provides efficient flexibility and diversity to aid subsequent sentence generation.

Hair Colour	Brown, Blonde, Black, Grey, etc.
Hair Style	Wavy, Curly, Straight, etc.
Age	Child, Youth, Middle aged, Senior
Gender	Male, Female
Ethnic	Asian, White, Indian
Accessories	Glasses, Sunglasses, Lipstick, Necklace, Necktie
Facial Hair	Goatee, Moustache

Table 5.1 Human attributes employed in this research

Refined Descriptions

Colour	Black, Blue, Brown, Grey, Green, Orange, Pink, Red, Violet, White, Yellow
Pattern	Spotted, Striped
Shape	Long, Round, Rectangular, Square
Texture	Furry, Smooth, Rough, Shiny, Metallic, Vegetation, Wooden, Wet

Table 5.2 Object attributes employed in this research

5.1.4 Sentence Generation

In this research, we regard language generation as a machine translation problem. This section of the framework is a direct improvement and extension of the proposed work in the previous chapter. The RNNs in this section are utilised to remove the dependency for template description generation. The generated sentences are intended to be more descriptive than those generated by other existing research, with attribute and labelling details owing to the proposed local based approach. To this end, the IAPR TC-12 (Grübinger *et al.*, 2006) dataset captions are used for both training and evaluation of the proposed system owing to its detailed captions in comparison to those provided by other popular databases (such as Flickr8k/30k (Young *et al.*, 2014)).

In this research, an encoder-decoder RNN structure is proposed, to overcome typical challenges in the machine learning field, such as variable input length, which plague typical machine learned networks. It consists of two RNNs, with one serving as an encoder and the other as a decoder, to overcome the input length variation problem. This encoder-decoder architecture was originally proposed by Bahdanau, Cho and Bengio, (2014) for machine translation problems. We transform it to deal with sentence generation and use the encoder RNN to encode the noun (objects/people) and adjective (attributes) keywords generated from the previous stage in the pipeline, into a fixed-length vector. The decoder RNN is employed to transform this fixed-length vector into a full descriptive sentence. The architecture of the encoder-decoder language generator is illustrated in Figure 5.2.

Refined Descriptions

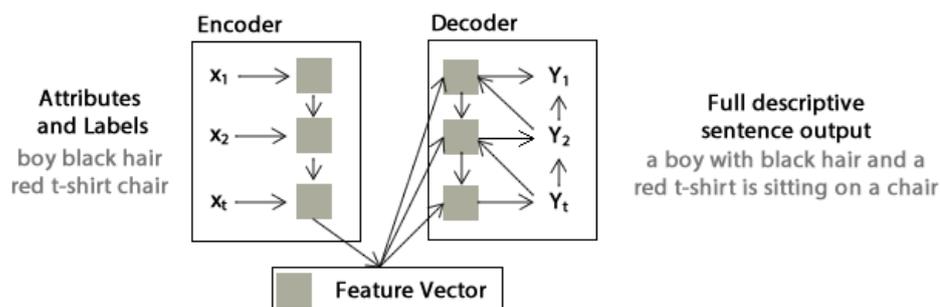


Figure 5.2 The encoder-decoder architecture with example attribute and label inputs and the expected sentence generation output

The proposed language generation component is trained using caption data from MSCOCO and a small subset of caption data from IAPR TC-12 (Grübinger *et al.*, 2006), in order to infer varying types of sentence structures. Small captions from MSCOCO and larger more descriptive captions from IAPR TC-12. Specifically, each image in the MSCOCO (T.-Y. Lin *et al.*, 2015) dataset is paired with five captions. The IAPR TC-12 dataset only has one sentence per image, however, it is more descriptive than captions provided by other datasets (e.g. Flickr30k (Young *et al.*, 2014)). Our proposed sentence generation structure (i.e., encoder-decoder) is therefore trained on nouns (object labels) and adjectives (human and object attributes) from the captions provided by the above training datasets. Each caption is first Part of Speech (POS) tagged with the Python NLTK (Natural Language Tool Kit) [<http://www.nltk.org/>], enabling nouns and adjectives to be extracted and stored. This leads to a source training set, which consists of ~30,000 unique words, and a target training set, which consists of ~20,000 unique words. The overall size of the training set is more than 500,000 captions, with more than 20,000 used as a validation set.

Recognized Entities & Attributes	Semantic Labels
Scenes	205 scene labels from MIT scenes including playground, living room and bedroom, etc.
People	Present/Non-Present
Objects	200 object categories from ILSVRC13 including car, dog, orange, desk, etc.
Person Attributes	A subset of 26 human attribute labels taken from the PubFig dataset
Object Attributes	All 25 object attribute labels present in the ImageNet dataset

Table 5.3 Information used for sentence generation

Both the encoder and decoder RNNs are implemented based on (Bahdanau, Cho and Bengio, 2014). All sequences longer than 50 words are ignored, or removed from the training set. Each RNN consists of 1000 hidden units in a single hidden layer to

Refined Descriptions

compute the probability of the next target word. These RNN models are trained with Stochastic Gradient Descent (SGD) with the use of Adadelta (Zeiler, 2012). During training, each update is computed with minibatches of 80 sentences. When an RNN has been fully trained, beam search is implemented to generate a description that maximizes the conditional probability of the source matching the target, i.e. $\text{argmax}_y P(y/x)$. During the test stage, object/people/scene and attribute labels generated in earlier stages of the pipeline and are used as the input source language with the full caption or description as the intended output. Overall, information taken into account during sentence construction is summarized in Table 5.3.

5.2 Evaluation

To evaluate the effectiveness of the overall system for image description generation, existing research, including NeuralTalk (Karpathy and Fei-Fei, 2015), NIC (Vinyals *et al.*, 2014), Show, Attend and Tell (Xu *et al.*, 2015) and Adaptive Attention (Lu *et al.*, 2017), has been employed for comparison. Besides that, we also evaluate the efficiency of the RNN-based attribute prediction. We first introduce the evaluation metrics applied in this research.

5.2.1 Evaluation Metrics

In this research, BLEU, ROUGE, and METEOR are employed for description generation evaluation. All of these methods use a similarity based measure between machine generated and ground truth sentences. We introduce each of these evaluation methods in the following sub-sections. The metrics have been explored in detail in Chapter 2, the metrics present and the reason behind their use are summarised below.

5.2.1.1 BLEU

The BLEU score (Papineni *et al.*, 2002) has been commonly employed as a machine translation measure. It can be applied to determine the semantic similarity between sentences. In this research, it is used to determine the closeness between a machine generated description and multiple high-quality human generated sentences. Moreover, in the literature, human performance has been stated to achieve a performance benchmark of 0.69 on the BLEU score (Vinyals *et al.*, 2014).

5.2.1.2 ROUGE

ROUGE-L is a subset of all of the available ROUGE (Lin, 2004) metrics. It takes two input sequences into account and identifies the Longest Common Subsequence (LCS).

Refined Descriptions

This LCS is the subsequence that occurs in both sequences with the maximum length. In sentence level ROUGE-L, the perception is that the longer the LCS of generated and GT sentences, the more similar the sentences.

5.2.1.3 METEOR

The METEOR score (Denkowski and Lavie, 2014) has been widely used owing to its high correlation to human subjects' annotations in comparison with other metrics (Denkowski and Lavie, 2014). It not only computes sentence similarity scores between reference and machine generated sentences, but also exhaustively identifies all matches between sentences based on certain matching criteria, such as exact word, synonym, and paraphrase matching.

5.2.2 Evaluation Results

To thoroughly test the proposed system, a comprehensive evaluation study is conducted. We first of all evaluate the RNN-based attribute classification using the test sets of their respective databases, i.e. the ImageNet (Russakovsky and Fei-Fei, 2012) and PubFig (Kumar *et al.*, 2009) datasets. We then test the system functionality on image captioning and description generation with a substantial evaluation using the IAPR TC-12 (Grüninger *et al.*, 2006) dataset, owing to its caption size and complexity. A small-scale experiment is also conducted using the NYUv2 sentence dataset (Wilde, 2007) to further test the system's superiority and efficiency for dealing with out-of-scope images. State-of-the-art methods, such as NeuralTalk (Karpathy and Fei-Fei, 2015), NIC (Vinyals *et al.*, 2014), Show, Attend and Tell (Xu *et al.*, 2015) and Adaptive Attention (Lu *et al.*, 2017), have also been employed for image captioning performance comparison. These first two methods are incredibly popular and were among the initial breakthrough in this field, they still perform extremely well due to their architecture and training styles. The other comparison research works utilise an attention model within their architectures, which could allow for a more direct comparison between the regional descriptions of this research.

5.2.2.1 Evaluation of attribute prediction

To compare the outputs of the RNN-based attribute classification of the proposed system with the annotations provided in the attribute datasets of PubFig and ImageNet, the BLEU score is used since multiple attributes can be treated as a sequence of words, resembling a sentence or a caption.

Refined Descriptions

The PubFig dataset and the subset of ImageNet are used for evaluation of human and object attribute classification, respectively. For evaluation of object attribute prediction, the subset of ImageNet is split into ~7000 training images with ~1500 images for validation and test, respectively. Evaluated with ~1500 test images, the proposed object attribute RNN classifier achieves an impressive BLEU score of 0.62 for BLEU-1. Some example object attribute prediction results are shown in Figure 5.3. We also divide the PubFig dataset into ~6000, ~1000, and ~1000 images for training, validation, and test, respectively, for evaluation of human attribute classification. Based on the testing of ~1000 images, the proposed RNN achieves a BLEU score of 0.61 for BLEU-1. As observed, both scores for object and human attribute prediction are very close to the benchmark BLEU score of 0.69 pertaining to human performance (Vinyals *et al.*, 2014). Figure 5.4 shows some example outputs of human attribute classification.

Good	OK	Not good
 Round shiny wet yellow	 Long metallic rough	 Red
 Furry spotted	 Square	 Brown rough

Figure 5.3 Object attribute classification results using object attribute RNN

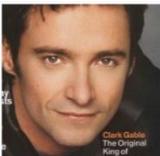
Good	OK	Not good
 Female White Youth Wavy Hair Wearing Lipstick	 Female White Brown Hair Wearing Necklace	 Female White Brown Hair Wearing Lipstick
 Male White Brown Hair Sunglasses Curly Hair	 Male White Curly Hair	 Female White Youth Wearing Lipstick

Figure 5.4 Person attribute classification results using human attribute RNN

5.2.2.2 Evaluation using the IAPR TC-12 Database

Evaluation of image description generation is conducted using multiple metrics, i.e., BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004), and METEOR. The main dataset used for evaluation is IAPR TC-12 (Grübinger *et al.*, 2006), consisting of 20,000 images. This dataset is used for evaluation over existing datasets such as MSCOCO (T.-Y. Lin *et al.*, 2015) or Flickr8k/30k (Young *et al.*, 2014), due to the characteristics of our generated descriptions. The above popular datasets such as MSCOCO and Flickr8k/30k typically consist of multiple short captions. However, the proposed system tends to produce descriptions that are more than twice the length of typical captions provided by these databases, which makes the comparison with these datasets less relevant. Moreover, the images from IAPR TC-12 are considerably larger with more objects and human actions included, paired with longer and more descriptive annotations than those in other datasets such as PASCAL (Everingham *et al.*, 2010), Flickr, and MSCOCO. Therefore, it is selected in this research for evaluation. The methodology and the framework presented in this chapter could not be trained on the same data as these related works. The architecture requires a training style and a dataset that covers multiple domains, in an aim to overcome the relative simplicity of the existing training methodology. This simplicity is in regard to one set of paired data, meaning that a lot of information will have to be inferred at test time. This highlights that the performance of these systems relating to the metrics are a result of the proposed system architecture.

Moreover, the proposed pipeline processing of this research also has added benefits regarding the training data required. The proposed system has been trained on a large text corpus, however the number of training images required is considerably less than those for existing methods (e.g. NIC). As mentioned earlier, we also conduct experiments with NIC (Vinyals *et al.*, 2014), NeuralTalk (Karpathy and Fei-Fei, 2015), as well as the attention based methods, such as Adaptive Attention (Lu *et al.*, 2017) and Show, Attend and Tell (Xu *et al.*, 2015), for performance comparison. All methods are tested with their best pre-trained models. Moreover, NIC is trained on Flickr whereas the other three comparative methods are all trained on MSCOCO, no images from the IAPR TC-12 dataset have been seen by any of the models. However, our proposed system is trained on multiple seemingly unrelated databases (such as the ILSVRC13 dataset for object detection, the ImageNet and PubFig datasets for object

Refined Descriptions

and human attribute prediction, respectively). To ensure a fair comparison, an unseen test subset of 2400 images from the IAPR TC-12 dataset is used for the evaluation of all methods.

Utilising a unique test set creates a more level field when testing and comparing to existing research. This reduces training time and resources for all models, as best case pre-trained models can be utilised. Each model may have its strengths and weaknesses in any given scenario or test data which provides the motivation for a new unique test set. This aims to minimize bias towards a particular models' dataset. This is highlighted in the nature of the images within the IAPR TC-12 dataset. These images are much more complex, varying dramatically in illumination, image focus, subject and finally their paired description.

The generated sentences by all models are compared with the reference sentences associated with each image provided by the evaluation database, and passed through the MSCOCO evaluation script (T.-Y. Lin *et al.*, 2015). This evaluation script uses all the above-mentioned metrics and enables comparison between the four baseline methods and the proposed research. Since each image in the IAPR TC-12 dataset is only paired with one description, this makes the image captioning task more challenging for all methods to score well with respect to the above metrics due to a dramatic lack of diversity.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Chapter 4	0.220	0.041	-	-	-	0.210
This Research	0.231	0.099	0.046	0.024	0.067	0.183
NeuralTalk	0.133	0.072	0.041	0.025	0.067	0.222
NIC	0.098	0.048	0.025	0.015	0.060	0.209
Show, Attend and Tell	0.117	0.067	0.037	0.021	0.077	0.204
Adaptive Attention	0.052	0.027	0.014	0.008	0.060	0.196

Table 5.4 Evaluation results for the IAPR TC-12 dataset using the MSCOCO evaluation script

The results for the IAPR TC-12 dataset are shown in Table 5.4. As illustrated in Table 5.4, in comparison with NeuralTalk (Karpathy and Fei-Fei, 2015), NIC (Vinyals *et al.*, 2014), Show, Attend and Tell (Xu *et al.*, 2015) and Adaptive Attention (Lu *et al.*, 2017), our system achieves superior performance, and outperforms other baseline methods for nearly all evaluation metrics. Specifically, in comparison with the

Refined Descriptions

attention-based methods, i.e. Adaptive Attention and Show, Attend and Tell, our work outperforms these two methods considerably for the BLEU metrics and remains competitive in the other metrics with Show, Attend and Tell beating our score in METEOR. Moreover, it is worth pointing out that all the baseline methods, especially NeuralTalk and NIC, rely on the typical CNN + RNN structure as the foundation for caption generation, while the proposed system employs a local-based approach to carry out not only local object detection and recognition, but also associated attribute prediction to inform subsequent encoder-decoder based description generation. The empirical results indicate that the sentences generated by our system are thus more descriptive. Comparatively, all baseline methods generate shorter descriptions using their respective methods. This accounts for the largest reason for the increase in metric score, where the score does suffer but remains comparable it is most likely due to the nature of the metrics and the nature of the reference GT description. For example, if the reference provided one attribute label, and this research provided 4, even if they were all correct this model would be punished. Overall, the empirical results and human inspection indicate that the generated descriptions from our system show more descriptive capabilities than those generated by NeuralTalk, NIC, Show, Attend and Tell and Adaptive Attention. As previously mentioned and presented the BLEU-1 score dramatically favours smaller generated captions, this further highlights the results from the proposed work due to the generated descriptions being longer than the comparative work. Table 5.5 shows some example outputs of our system and the four baseline methods based on the evaluation of the IAPR TC-12 dataset.

The results on this test set are also compared to the work presented in the previous chapter. This system shows an improvement over this previous work, highlighting the positives of building upon the existing framework. The ROUGE-L score is reportedly higher than this presented framework, this could again be a limitation of the metric, specifically the issues surrounding the different sentence and caption structures. E.g. This systems outputs would include additional describing words or attributes, limiting the effectiveness of this measure.

Referring to Table 5.4, it is worth mentioning that, despite the comparatively high BLEU scores and comparable METEOR scores yielded by our system, it achieves a comparatively lower ROUGE-L score than those of the other systems.

Refined Descriptions

IAPR TC-12		
Our system	View from above of a valley.	a sea cliff with a brown, red and brown cliffs and a wet seal behind it
Reference annotation	A river in a brown valley with many terraces; a yellowish-brown bush in the foreground.	grey and brown rocks in the foreground a light brown sandy beach at the sea and wooded hills behind it a blue sky in the background
NeuralTalk	a dog is running through the woods	a person on a beach with a surfboard
NIC	a herd of elephants walking across a lush green field	a man is standing on a rock overlooking a lake
Adaptive Att.	a group of people standing on top of a lush green field	a rocky beach with rocks on it
Show, Attend and Tell	A woman is walking down a rocky hill	A man is standing on a rock in the desert
IAPR TC-12		
Our system	Three people walking across a rope bridge	A hotel room with brown walls, a rectangular bed and a wooden sofa with a black, rectangular lamp behind it
Reference annotation	A woman is walking on a narrow rope bridge in the middle of a forest with many green trees and bushes	A single bed made of wood with a red pillow and a striped blanket two small bedstands made of wood with a bedside lamp each a light blue carpet and a wooden wall in the background
NeuralTalk	A man is climbing a rock	A dog is laying on the floor with a blue ball in its mouth
NIC	A man is standing in the woods holding a frisbee	A bed with a white comforter and a white blanket
Adaptive Att.	A brown and white dog standing next to a fence	A bedroom with a bed and a dresser
Show, Attend and Tell	A man in a white shirt is standing in a wooded area	A woman in a pink striped shirt is sleeping on a bed

Refined Descriptions

Table 5.5.Cont.		
IAPR TC-12		
Our system	A construction site with a thunderous person at a round, brown wooden bicycle on a long wooden cart	A kitchen with red walls and brown vegetation on the white table
Reference annotation	A man with a blue cap and a greenish brown overalls is standing with a red machine in front of a grey wall a black iron door on the right	A tray on bar with cut melons oranges mangos and a coconut in form of a mouse head a man behind it three brown round doors and a white wall in the background
NeuralTalk	A man sitting on a bench with a dog	A woman is sitting on a couch with a man in a black shirt
NIC	A man in a red shirt is sitting on a bench in a park	A table with a vase of flowers and a vase
Adaptive Att.	A man standing next to a bicycle on a sidewalk	A woman is holding a stuffed animal in her hands
Show, Attend and Tell	A man in a blue shirt and blue hat is riding a bicycle	A woman in a white shirt is sitting on a table with flowers and flowers

Table 5.5 Example system outputs as compared with reference descriptions and the outputs of all baseline methods for the evaluation of the IAPR TC-12 dataset

This could be owing to certain characteristics of our generated descriptions. As discussed earlier, the ROUGE-L score looks for the longest common subsequence between a machine generated caption and the reference. In the proposed system, since attributes are predicted to enrich the description, it has occasionally used additional or different attributes to precede object labels. This could have a negative impact with respect to the ROUGE metric, which leads to cutting the affected sequence at that point for score calculation, therefore reducing the overall ROUGE-L result. For future work, we aim to produce a dataset similar to Flickr or MSCOCO, but with more detailed descriptive sentences to enable further evaluation of our proposed system and other similar methods.

5.2.2.3 Additional Experiment using NYUv2

A small-scale additional experiment is also conducted using cross-domain images extracted from a scene description challenge, i.e. the NYUv2 sentence dataset (Wilde, 2007). The objective is to determine how well our system is able to cope with scene

Refined Descriptions

description generation for out-of-scope images. This NYUv2 sentence dataset describes purely indoor scenes with annotated descriptions containing objects, their attributes and relationships between multiple objects. Each image in this dataset is paired with one description. Each description consists of up to three sentences. Therefore, the available reference annotations for this dataset could range from very brief (e.g. one sentence) to very descriptive (e.g. multiple sentences) which further increases the challenge of this dataset.

In this small-scale experiment, we train the encoder-decoder based sentence generation component purely on captions and descriptions provided by the MSCOCO dataset. The newly generated model and all other baseline systems are then used to test upon ~1400 images extracted from the NYUv2 sentence dataset. Without depth information, the experiments indicate that this is a challenging dataset for the proposed system and related methods. Although the overall metric scores obtained for this dataset are comparatively lower for all models than those achieved using the IAPR TC-12 dataset, the results indicate that for the BLEU-1 metric, descriptions generated by our system outperform those generated by NeuralTalk, NIC, and Adaptive Attention by 30%, 70% and 71%, respectively. It also scores equally against Show, Attend and Tell under the same metric. Some example outputs generated by our system and related methods are shown in Table 5.5.

5.2.2.4 Real-life Deployment

In order to determine how the proposed system would perform under real-life or in the wild scenarios, some initial experiments were conducted using the vision system of a humanoid NAO robot. Because of the computationally exhaustive process of the proposed system, instead of deploying it to the robot platform, we utilize a GPU server. The robot is therefore performed as a client which communicates with our GPU server via a wireless network. This enables the robot’s vision API to capture real-life images, which are subsequently transferred to and analysed in the remote GPU server. The generated outputs are then communicated back to the robot, which enables it to describe the environment. We have also conducted some initial testing with the robot using real-life scenes. The results show that the robot can identify many objects and describe the environmental layouts and people accurately, however struggles with more complex scenes, owing to the longer processing time and the captured image resolution of the robot’s front facing camera. In future work, we endeavour to enable the robot to react to scenes quicker and more accurately and pair it with conversation capabilities and internet access (similar to Siri or Amazon Alexa) however more

NYUv2 Sentence		
Our system	A hotel room with a long rectangular draped sofa and a very long, rectangular bookshelf awaits	An office setting with long rectangular red chairs next to a wooden table.
Reference annotation	This is a living room with wooden floor. There is a big beige sofa on the left of the room with two pillows on top. Near the sofa is a black armchair. There is a book cabinet behind the sofa and the room is separated by a door.	This is a conference room with a long wooden table with red chairs around it, a projector mounted to the ceiling, and red window blinds. A multiline phone sits on the table.
NeuralTalk	A man and a woman are sitting on a bench	A man is sitting on a bench in a room
NIC	A living room with a couch and a tv	A living room with a couch and a tv
Adaptive Att.	A living room filled with furniture and a bookshelf	A room with two windows and a window
Show, Attend and Tell	A woman sitting on a couch in a library	Two men are sitting on a couch

Table 5.5 Example system outputs compared with reference descriptions and the outputs of all baseline methods for the evaluation of the NYUv2 sentence dataset

Refined Descriptions

mobile, visual and social, which could be used for healthcare purposes, e.g. to alert care providers of a fallen person or if the fallen subject requires aid.

5.2.2.5 Transfer Learning

Three further experiments were conducted with this proposed architecture. One of the initial aims of the proposed study was to allow a created system and/or model to be utilized on any given image. This can be interpreted in many ways, we conduct further experimentation in order to address some of these possibilities.

Any image, within computer vision can vary depending upon the domain. For example, it could refer to the ability to cope with images in the wild, meaning any image captured on a camera, or that it is within the domain of images that the system is already trained on. Another possibility is that the system can cope with any image, meaning photographs, artwork, or even domain specific applications. Therefore, in this work we conduct some additional experiments on dramatically differing datasets, comparing the system as it currently stands, to some small fine-tuning transfer learning techniques on the given domains.

The first additional experiment conducted is on the Pandora dataset (Florea *et al.*, 2016), this database is available in either 7k or 18k formats, each equipped with approximately the number of images in its name. In this experiment, a test set of ~1000 images of the 7740 in the first database are taken and passed through the entirety of the architecture as described in this chapter. We conduct multiple experiments on this dataset, investigating the effects of transfer learning.

Our initial experiment consists of just the model as described with no changes and no further training. Examples of these results are shown in Tables 5.6, 5.7 and 5.8.

For a different, larger, more diverse painting dataset, the application is tested in similar fashion as the previous experiment. However, this time on the wikiart dataset (Tan *et al.*, 2016). This dataset, contains a diverse range of art styles, ranging from realism to abstract works of art. Examples of the generations provided by this system are provided in Table 5.6.

This research was also extended with the regional descriptive capabilities on the Person re-identification dataset (iLids) (Wang *et al.*, 2014, 2016; Ma *et al.*, 2017). This dataset contains readily cropped images of people on relatively low resolution CCTV

Refined Descriptions

images, from 2 cameras at 2 different angles. This dataset also consists of 300 different pedestrians observed from 2 distinct camera positions, the dataset is also split into static and image sequences. The image sequences are a series of images ranging from 23 to 192 individual captured camera frames, with an average sequence length of 73.

This dataset is claimed to be extremely challenging due to the clothing similarities among the subjects, as well as lighting and viewpoint changes due to the multiple angles, cluttered and busy backgrounds and random unavoidable image occlusions.

For evaluation of this dataset we collect a random 400 images from both cam1 and cam2. This creates a relatively small test subset of 800 images. These images are then passed into the architecture described within this chapter, however the final translation stage is removed. This is simply as there is no need for a ‘sentence’ to be inferred, as the regional person descriptions are just as accurate in this given domain. As the previous experiment on the Pandora dataset, the effects of fine-tuning and transfer learning are explored. The initial experiment consists of the model and architecture as described. These results can be seen in Table 5.7.

The initial results of the transfer learning are displayed in the following tables and figures. In addition to the experimental framework highlighted in the previous section, further information regarding each experiment is provided along with these results.

The iLids video reidentification dataset was tested with this description capability, in a bid to address how such a system could process and annotate relatively low resolution images of people. To this end, the whole system framework was not utilized. This is due to the nature of the sentences and descriptions produced, these would more than likely be irrelevant or include information that is distracting as well as potentially incorrect. To this end, the whole system with the absence of the sentence/descriptive layers was utilized for this particular experiment.

5.2.2.5.1 Transfer Learning Results

The results from the Pandora dataset show that the system is detecting and reporting many of the entities within the image correctly. For example, describing the scene and scope of the valley, as well as the colour of a person’s clothes and gender within a full and rounded descriptive sentence. However, the system fails to annotate other aspects at all. This shows that there is promise within this technique and its subsequent

Refined Descriptions

methodology to allow such a system the ability to annotate and accurately describe these kinds of paintings.

In Table 5.8, the results from the iLids person re-identification experiment without the description framework show the success of the framework with minimal alterations.

	
A flat, green and brown landscape with a brown hill in the background	A green coast with the leaves of a palm tree on hill in the background

Table 5.6: Initial transfer learning experiments on the WikiArt dataset in which paintings were passed through this chapter's framework and their associated result.

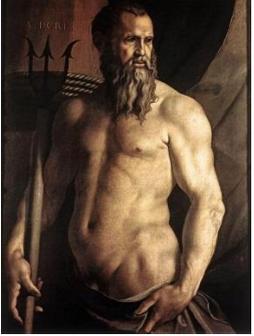
	
Gravel road in a brown, dry valley with a steep, bald mountain on the right and a wooded mountain on the left	A local man with a red costume is kneeling on the ground and holding a dish in his hands at night
	
Photo of four dark-haired children on a grey wall in front of a yellow building made of grey bricks and with a thatched roof sand behind it	A man is sitting on a wooden chair at an entrance to a workshop

Table 5.7: Initial transfer learning experiments on the Pandora dataset in which paintings were passed through this chapter's framework and their associated results

Refined Descriptions

	
profilepose standing casualjacket Person	black upperbody Person
	
turnedback standing kid	tshirt Person female frontalpose standing armsbent middleaged Person

Table 5.8: Initial transfer learning experiments in which person re-identification datasets were tested upon.

5.3 Summary

In this research, we have proposed a local based deep learning architecture for image description generation, in order to describe and annotate multiple objects and people within the given image. It consists of object detection and recognition, scene classification, RNN-based attribute prediction, and encoder-decoder RNNs for sentence generation. The proposed system mitigates the problems associated with holistic methods by relating specifically to image regions of people and objects in a given image in order to gain more detailed and longer descriptions in comparison to most notable works within this domain. The experimental results have indicated the impressive performance of the proposed RNN-based object and human attribute prediction, both standalone and their use within the image caption and description framework. Furthermore, the overall system also showed its significance for image description generation. Evaluated with the IAPR TC-12 dataset, in comparison with several baseline methods, i.e. NeuralTalk, NIC, Show, Attend and Tell and Adaptive Attention, the proposed system produces more detailed and descriptive captions, and outperforms these state-of-the-art methods significantly for nearly all of the evaluation metrics. The empirical results also indicated the superiority of our proposed system

Refined Descriptions

over existing methods when dealing with out-of-domain indoor scene description generation for images from the NYUv2 sentence dataset.

This system has also been tested within a transfer learning experiment, to determine its ability to annotate and describe images from an irrelevant domain, as well as a re-identification task. The out of domain images consisted of paintings from different artists, periods and styles making it a difficult task.

The re-identification task was conducted with only part of the overall system pipeline. With no additional training and the removal of the encoder-decoder RNNs, we are able to show the power of the attribute predicting RNNs. The system can accurately label genders, clothes and estimate age bands all with the training previously conducted on the PubFig dataset, no images from the iLids dataset were used during training.

For future work, more detailed attributes, such as those related to garments (Shen *et al.*, 2014), could be considered to further improve descriptive capabilities of the proposed system. We also aim to explore saliency detection to further improve the system's outputs with the emphasis of the potential focus of the images. In the longer term, we also aim to equip the proposed system with transfer learning (Shao, Zhu and Li, 2015) to deal with image description generation for images such as cartoons and oil paintings.

Refined Descriptions

Chapter 6 A Hierarchical and Regional Deep Learning Architecture for Image Description Generation

6.1 Introduction

This chapter introduces the third proposed system for image description generation. Image captioning is one of the upcoming but also most challenging research areas for deep learning. As previously stated, a system, that can not only accurately label image regions but also scale to whole image description, shows great potential in diverse

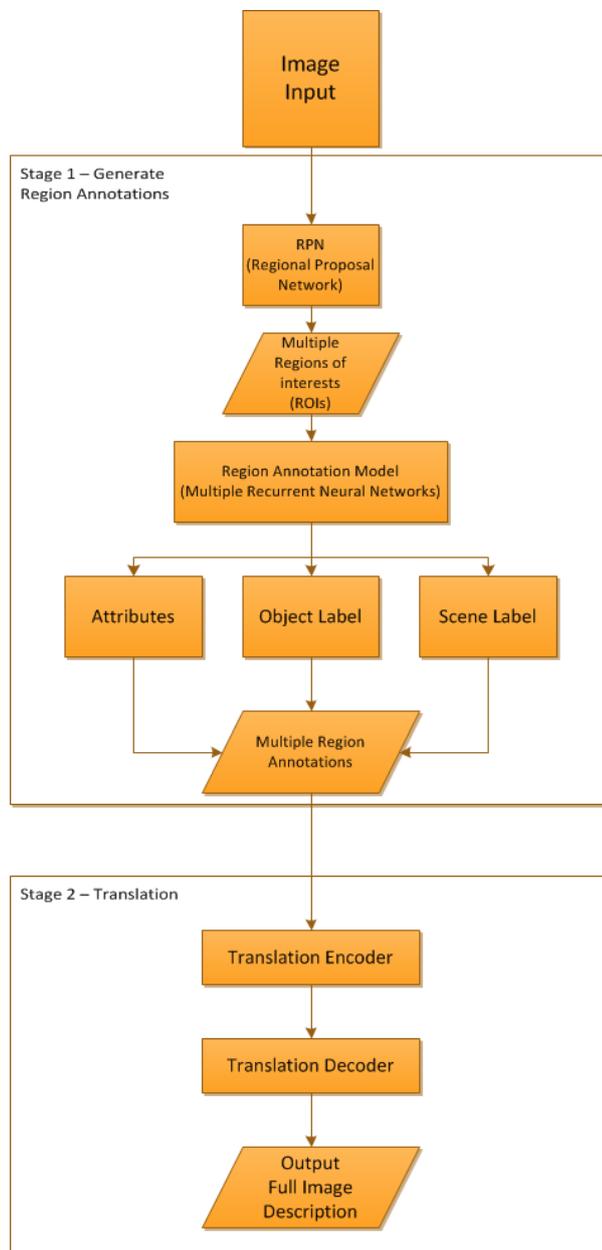


Figure 6.1. The high-level overview of the proposed image description generation system.

applications, such as news or medical image annotation and automatic scripts generation for movies. Many existing research and publicly available datasets were tailored for brief image captioning. So far it is still a challenging task to generate detailed and refined descriptions for both relevant regions and the whole image.

Therefore, in this research, we aim to address the above challenges and propose a novel compact deep network architecture for detailed image description generation. we propose a hierarchically trained deep network in order to increase the fluidity and descriptive nature of the generated image captions. The proposed deep network consists of initial regional proposal generation and two key stages for image description generation. The initial regional proposal generation is based upon the Region Proposal Network from the Faster R-CNN (Ren *et al.*, 2015). This process generates regions of interest that are then used to annotate and classify human and object attributes. The first key stage of the proposed system conducts detailed label description generation for each region of interest. The second stage uses a Recurrent Neural Network (RNN)-based encoder-decoder structure to translate these regional descriptions into a full image description. The proposed deep network is composed of multiple Convolutional Neural Networks (CNNs) (LeCun *et al.*, 1998) in combination with Recurrent Neural Networks (RNNs) (Mikolov *et al.*, 2010), specifically Long Short Term Memory networks (LSTM) (Graves and Schmidhuber, 2005) and Gated Recurrent Network (GRU) (Chung *et al.*, 2014), for image caption generation. It is capable of performing object and scene classification, as well as human and object attribute prediction, simultaneously. Especially, the simultaneous generation of human and object attributes provides rich and detailed descriptions of image regions, and equips the proposed system with impressive capabilities to deal with out-of-domain queries.

The overall architecture of the proposed deep network is presented in Figure. 6.1. The system is composed of multiple stages (including pre-processing and two key stages), the combination of which allows for the full functionality. The initial processing of the system uses the Region Proposal Network (RPN) (Ren *et al.*, 2015) to generate multiple regional proposals (i.e. regions of interest), that are likely to contain objects or people. The first key stage of the model conducts object and scene classification and attribute prediction. This stage combines the extracted regional features with word vectors in an RNN for attribute prediction, and utilizes the same regional features for

Hierarchical Description Architecture

scene and object labelling. The second stage is used for language ‘conversion’. It converts the generated attributes and other class labels into fully descriptive image captions.

In order to enhance the system’s generalization and scalability, instead of training the system using existing image caption datasets such as Flickr (Young *et al.*, 2014) or MSCOCO (T.-Y. Lin *et al.*, 2015), the overall proposed deep architecture is hierarchically trained upon multiple different datasets from different domains. These datasets have their individual dedications to be used within a particular domain, e.g. attribute datasets are originally purely dedicated to attribute prediction applications.

Using multiple datasets allows for several benefits. As an example, the proposed system ensures that there is reduced offline training, as compared to other end-to-end and composite methods. Our proposed model ideally would require one dataset that would be annotated with all the functionality of the system. The closest dataset currently is *Visual Genome (VG)* (Krishna, Zhu, *et al.*, 2017), which is very large with ~110,000 images, however no full image captions are provided, despite all other desired features. Therefore, the proposed system is trained on multiple smaller image datasets that are not typically associated with image caption, yet still provides a competitive outcome to systems that are trained on a single image captioning dataset. Our proposed region-based method allows for increased functionality, including longer image descriptions with regional details, and simultaneous region and full image description generation. Overall, the system shows great diversity for image caption/description generation with more efficient training and testing in comparison to existing methods.

Moreover, as the proposed system is not specifically trained on image-to-caption datasets, such as Flickr or MSCOCO, another main advantage of the system is that it can handle out-of-domain images efficiently. This ensures that a reasonable detailed description can be generated for most images passed on to the system regardless of its source to increase the system’s robustness. Finally, the main contributions of this chapter are summarized as follows:

- A novel deep architecture for image region annotation is proposed. It generates not only regional annotations but also integrates the regional captioning into full image descriptions.

- The proposed deep network has a more efficient training process and shows great robustness and efficacy in dealing with out-of-domain images. It has also been deployed and integrated with the vision API of a humanoid robot to indicate its effectiveness in real-life settings.

6.2 The Proposed Deep Network for Image Description Generation

This research proposes a hierarchical deep network with the intention to produce image description with a great level of detail. It integrates multiple CNNs with particular types of RNNs such as LSTM and GRU, for image description generation. We introduce each key stage of the proposed network below.

6.2.1 Model Training

Our proposed framework is loosely categorized as an end-to-end system. It presents a unified model that can generate not only descriptive region annotations, similar to DenseCap (Johnson, Karpathy and Fei-Fei, 2016), but also full image descriptions, as Google NIC and NeuralTalk, or even attention based methodologies as described within the previous chapters. Due to the large and complex nature of the proposed model, and the fact that the model is trained on multiple datasets, the proposed system has to be trained hierarchically.

This process first involves freezing multiple sections and branches of the model, before training and fine-tuning the desired weights on the relevant branches with the relevant data. This leads to a large amount of training data being utilized, across multiple datasets, including the generation and use of dummy data for the unaltered or frozen branches. The freezing and training of certain layers at the training stage depend upon the dataset currently in use, and prevent any unnecessary and unhelpful data from interrupting the way the model should learn.

VGG (Krizhevsky, Sulskever and Hinton, 2012) is a popular CNN and commonly used for object classification. It has a high top-5 accuracy which means that these weights of its layers can directly be inserted into the corresponding layers of our model for object classification, as shown in Figure. 6.2. However, VGG's success can be in part related to the discriminative features it extracts before applying its fully connected layers. We intend to utilize these features for more than just object and scene classification, but also subsequent attribute prediction.

Hierarchical Description Architecture

This is achieved as previously stated by freezing and preventing the updating of the weights within the convolution layers of VGG, so that the extracted features will be extracted in the same way, maintaining the discriminative features. These features are then used to train the scene classification using the fully connected layers on the given dataset. A list of datasets used for the training of each key stage of the proposed model is provided in Section 6.2.2.

Human and object attributes are also trained in our current model configuration. This stage uses word encoding in combination with the extracted features in order to generate regional attribute labels, which can be viewed similar to the methodology within Densecap (Johnson, Karpathy and Fei-Fei, 2016). Again, all layers which are irrelevant during training are isolated or unaltered.

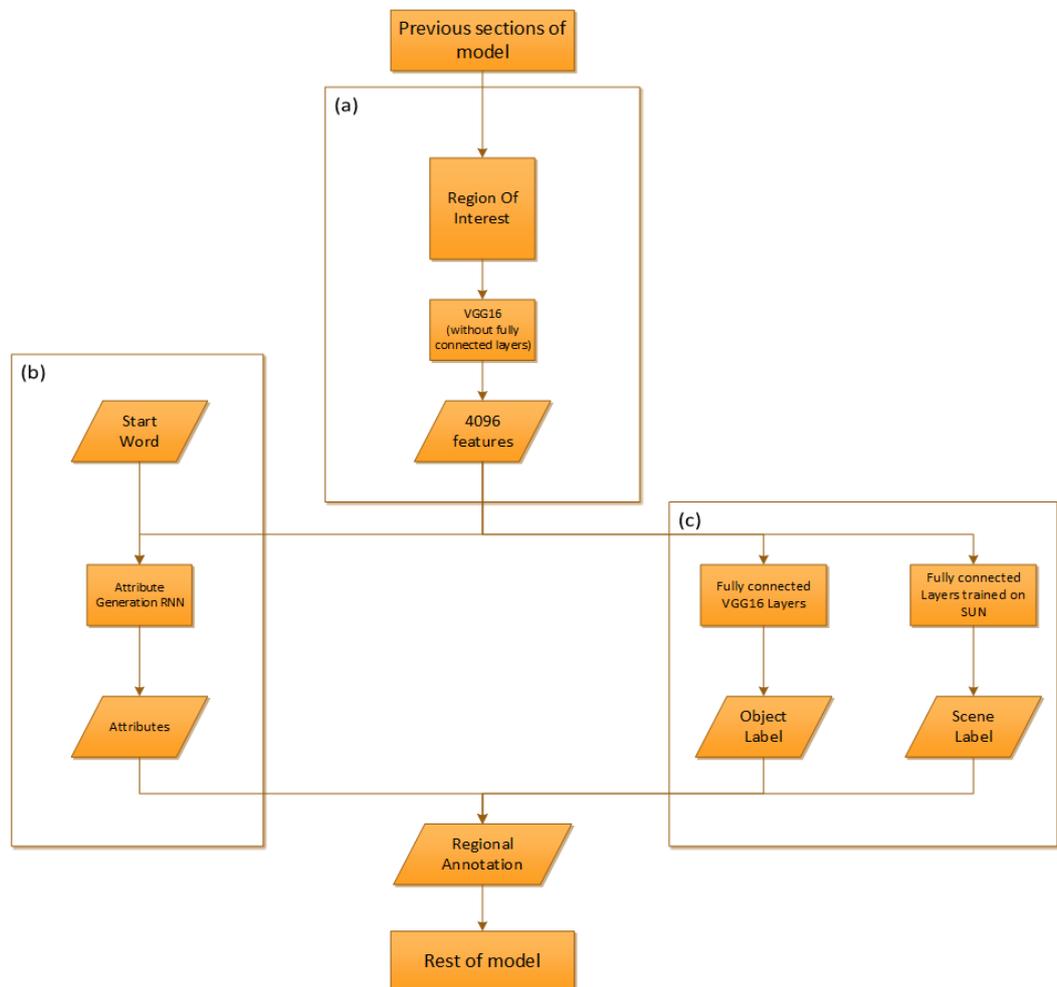


Figure. 6.2. (a) The initial process of extracting features from the generated region of interest. (b) The process of combining the features and word vectors to generate region attributes. (c) Using the re-added VGG layers to generate object labels and the trained scene layers to generate scene labels.

Hierarchical Description Architecture

When training the deep neural networks to obtain a target output, we are only training certain layers, while the frozen layers are given dummy data. These will not update the weights in any way yet still allow the training of the necessary layers to take place. This process is repeated for all relevant branches of the model.

Although the model is currently trained hierarchically, in future work, if a dataset exists that covers all needed outcomes, the model could theoretically be trained in an end-to-end fashion.

6.2.2 Model Datasets

In Figure. 6.2, the model has been split into a number of components. We highlight the datasets used to train the relevant layers below.

ImageNet (objects labels) (Russakovsky and Fei-Fei, 2012)– This dataset consists of around half a million images, for 200 objects. VGG is also pre-trained on this dataset.

PubFig (Kumar *et al.*, 2009) – This large facial image dataset initially has 79 attributes for each of the ~60,000 images. We only use a small subset of the available human attributes, although a large subset of the images is used in this research.

ImageNet (object attributes) (Russakovsky and Fei-Fei, 2012) – A small portion of 10,000 images from the full ImageNet dataset is paired with 10 object attributes. All of these images are used for training in our work.

MSCOCO/IAPR TC-12 (Grübinger *et al.*, 2006; T.-Y. Lin *et al.*, 2015)captions – We collect captions from MSCOCO and IAPR TC-12 for the training of the caption generator in this work.

SUN scene dataset (Xiao *et al.*, 2010) – This dataset consists of more than 100,000 images of 397 scene categories. We use a subset of the available images, i.e. 10,000 images, for training.

6.2.3 Architecture

The initial processing of the system requires regions to be collected and cropped. These regions are likely to contain objects or people for the system to annotate. This stage involves region proposal generation. In this research, it is implemented by the Faster R-CNN (Ren *et al.*, 2015). The RPN within the Faster R-CNN is essentially a powerful neural network that generates bounding box regions and confidence scores.

Hierarchical Description Architecture

It produces a high score when the system believes the region contains an object or something of interest. The number of regional proposals passed on to the next stage is determined by several factors however capped at a max of 300, dependent upon the size and the complexity of the image as well as the generated confidence measures.

After generating regional proposals, the rest of the proposed model is split into two key stages. The first stage generates detailed regional labels, and the second stage translates these region labels into a full description. We introduce these two stages in detail in the following sections.

The first key stage accepts two inputs, i.e. a start word and an image. It generates attribute and object labels word by word and this generated sequence is fed back into the network after each word to produce all attribute and object labels. The second stage is based upon a common machine translation approach to ‘translate’ the labels from source (i.e. region labels) to a full description. Therefore, the final generated description is expected to be more detailed than that of existing research.

These two stages, described for the rest of this chapter as regional and translation models, can be further broken down into a number of branches that are responsible for a specific task, i.e. attribute prediction and scene and object classification. These branches are explained in detail below.

The regional model in the first stage can be split into three branches, as shown in Figure. 2.2. The first section, as shown in Figure. 6.2(a), shows the region of interest being passed into the modified CNN-based, VGG feature extraction network to generate the discriminative features as used in the subsequent stages of the model architecture. The left branch, as indicated in Figure. 6.2(b), learns word feature vectors, which are ultimately used for attribute generation. This branch converts word integer positions within the vocabulary into a fixed size vector. In our work, this is a 128-dimensional vector.

Hierarchical Description Architecture

The vector is passed through a GRU (Chung *et al.*, 2014) accompanied with fully connected layers to generate an output at each time step. The next stage in this branch involves the image features. The outputs of GRU and sequence image features are then merged and combined before being passed through the subsequent language generating neural network. This network consists of LSTM (Graves and Schmidhuber, 2005) layers in combination with a fully connected layer in which the next word in the sequence is generated. The aim of this branch is to generate human and object attribute labels of the input regions which are later used in the translation stage.

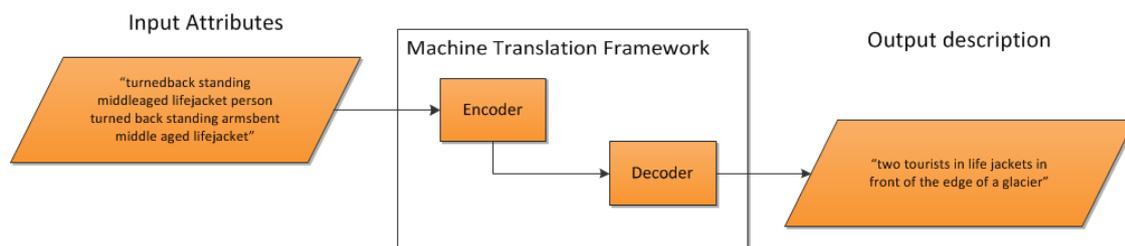


Figure. 6.3. An example of the generated input regional attributes and its translated output description

The first part of the right branch, as illustrated in Figure. 6.2(c), specifically classifies object labels. This branch uses 4096 image features extracted from the VGG16 CNN. This branch initially has the last two layers removed, which consist of a dropout and a fully connected layer, in order to deliver the features. However, these layers are re-added subsequently in the architecture so that captions and object labels can be simultaneously generated, and later concatenated with the attribute labels to deliver an overall output. On the same branch, the extracted image features are used to classify a scene label. Furthermore, the far right section, in Figure. 6.2(c), is used to classify the scene and it generally uses the whole image as a region. To achieve classification, the layers of this branch are fine-tuned on the SUN dataset (Xiao *et al.*, 2010). The initial stage of this process is the same as the object classification as it collects 4096 discriminative features from the VGG16 CNN.

This aforementioned process relies on multiple training stages. However, the second stage of the model, i.e. the 'translation' or 'conversion' stage, only requires one training step, while the entire top section, previously discussed, remains as is, unaltered by this stage.

Hierarchical Description Architecture

Good Results:



Ours: three grown-ups in a flat landscape with a mountain in the background

NTalk: a man standing on a beach holding a surfboard

NIC: a man and a woman are sitting on a rock overlooking a lake



Ours: a man is standing on a black rock with a sandy desert, a green valley and a mountain range and clouds in

NTalk: a man sitting on a bench in the middle of a field

NIC: a man and a woman are sitting on a rock overlooking a lake



Ours: four female tourists are posing with a large, dark brown mountain with a snow covered peak in the background

NTalk: a man and a woman are walking on a beach

NIC: a man is standing on a snow covered mountain



Ours: a grey and light brown house with a small very dense vegetation behind it

NTalk: a man and a woman are sitting on a bench in a park

NIC: a man sitting on a bench with a dog

Figure 6.4. Example outputs generated by the proposed system, Google NIC and NeuralTalk (referred as NTalk) on the IAPR TC-12 dataset

Reasonable Results:



Ours: a man in a black jacket with a grey rain jacket on a head on a grey brick on a bridge over mountains in the background

NTalk: a man and a woman are standing on a rocky path

NIC: a man and a woman standing next to each other



Ours: a man green mountain landscape with a few towns with a brown, bald mountain range in the background

NTalk: a young girl in a pink dress is walking on a path

NIC: a man and a woman standing next to a man

Figure. 6.4. Example outputs generated by the proposed system, Google NIC and NeuralTalk (referred as NTalk) on the IAPR TC-12 dataset

This translation model in the second stage shown in Figure. 6.3 follows the work of Bahdanau, Cho and Bengio, (2014) in which a single neural network architecture consists of an improved encoder-decoder. This encoder-decoder architecture is originally designed for machine translation, and outperforms existing statistical machine translation approaches (Lopez, 2008; Sutskever, Vinyals and Le, 2014)

This model structure of Encoder-Decoder (Cho *et al.*, 2014) is still used in the typical fashion of encoding the source text into a vector, and decoding the vector to generate the target text. In this research, the source consists of extracted attribute labels together with scene and image information, with the target being a detailed image description.

As previously discussed, the encode structure is used to encode the attribute and class labels into a vector. It initially encodes into a sequence of vectors, of which a subset is adaptively chosen for use during the decoding stage. This is followed by the decoder which uses this subset to generate an image description. This encoder-decoder processing is opposed to a fixed length vector, which is determined to be a bottle neck problem in existing research. An example of generated regional attributes and its associated translated output is shown in Figure. 6.3. As can be seen in Figure. 6.3, the

generated input attributes describe two middle aged persons in life jackets. This is translated and the system generates ‘two tourists in life jackets in front of the edge of a glacier’, owing to the training of the caption generation system where the glacier information is inferred in the output description. The scene information is omitted in the attributes when a confidence value is not reached. The above two-stage deep network implementation is utilized in this research owing to the stated improvements over other existing work for the generation of sentences that are longer and more descriptive than other works.

6.2.4 The Deployment to the Robot Platform

A real-life application of this system has also been initially explored, by combining the proposed model with the vision SDK of a humanoid robot, NAO NextGen H25. The NAO robot has a powerful CPU processor and camera sensors to allow for real-time image processing and better low light perception. The integration of the proposed system with the robot’s SDK enables the robot to conduct health monitoring, e.g. to identify falling subjects and describe users’ environment to promote personalized human robot interaction. It also enables the evaluation of this proposed system in diverse real-world settings. The robot begins the interaction and image description generation process upon being verbally asked by the user or a tap on the robot’s head, utilizing either the touch feedback API or the speech recognition API as available within the robot’s framework. Although the NAO robot has a powerful CPU processor for its size, it does not have the capacity to run complex and deep models like the proposed system in real time. To this end, the robot acts essentially as a front end that interacts with the user and with the system deployed on a more powerful GPU workstation. The processing procedures are as follows: The robot captures an image from one of its two cameras, sends it via a LAN to the remote server to be processed. The process passes the image through the proposed network model and produces

Hierarchical Description Architecture

outputs in the form of raw text.

This is then sent again via LAN to the robot, who then receives a response from the web sockets in raw text and finally verbally outputs the generated description of the captured image using its text-to-speech API. Preliminary experimentation shows that the robot is capable of observing, recognizing and describing diverse objects (such as cups, fruits, furniture etc.) and people, as well as their attributes within multiple environments, such as ‘stairway’, ‘library’, ‘kitchen’, or ‘office’. The server used throughout the experiment is based on the Nvidia Deep Learning DevBox [https://developer.nvidia.com/devbox] equipped with 4 GTX TITANS. We will also conduct more experiments to explore the efficiency of the proposed system integrated with the NAO robot for diverse real-life settings as one of the future directions.

As an initial indication of the system efficiency, we deploy the robot platform

Table 6.1. Our results and comparison with related work on the IAPR TC-12 dataset

IAPR TC-12	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	SPICE
Chapter 5	0.215	0.101	0.048	0.016	0.075	0.216	0.037
Our Work	0.201	0.105	0.053	0.024	0.073	0.216	0.038
NeuralTalk	0.129	0.069	0.038	0.022	0.065	0.216	0.061
Google NIC	0.094	0.046	0.034	0.013	0.059	0.205	0.06

integrated with the proposed deep network to a real-life application scenario. Table 6.1 shows images fed through the robot’s network and the aspects in which the robot

			
Description output	“Two women and six men are sitting behind a wooden table in a room with a light yellow wall”		“A man and three women are walking on a slope with a white ladder and bushes behind them, snow covered mountain range in the background”

Table 6.2. Example images and outputs produced by the system deployed on the NAO robot

reports based on image description generation output. The empirical results indicate the efficiency of the proposed system in dealing with real-life images via the robot platform. In future work, we aim to incorporate floor detection methods with the existing object and scene recognition to allow the system to detect hazards and audibly present this information via the robot platform to benefit e.g. healthcare application scenarios. We also aim to equip the proposed system with the capabilities of dealing with low resolution images to further enhance performance, as well as conduct a large scale real world test utilizing images captured from a local camera to prevent resolution becoming an issue, or directly from the robots vision cameras.

6.3 Evaluation

In order to evaluate the efficiency of the proposed system, we implement two popular baseline methods, i.e. Google NIC and NeuralTalk, for comparison. The IAPR TC-12 dataset has been used for evaluation. The IAPR TC-12 dataset consists of 20,000 images, with each image paired with one descriptive sentence or a short paragraph. In the test stage, we ran our system on a random selection of ~10,000 images.

We have trained our RNN-based language models purely on two small caption subsets of IAPR TC-12 and MSCOCO, respectively, without using the associated images. That is, we only use a small number of captions from each dataset for training and the training process has not used any images from either of these datasets. The baseline methods, i.e. NeuralTalk and Google NIC, however, have been trained on these datasets (using both captions and images), therefore leading to higher evaluation scores. We still provide the evaluation results using these datasets in order to indicate the efficiency of the proposed model.

To quantify the performance of the system, the MSCOCO evaluation script, is utilized. The MSCOCO script contains four BLEU metrics (i.e. BLEU-1, BLEU-2, BLEU-3 and BLEU-4) based on the n -gram method of determining string/sentence similarity. It is also equipped with other evaluation metrics such as, METEOR, ROUGE-L, and SPICE.

The detailed results using all of the above metrics for the evaluation of the IAPR TC-12 dataset are shown in Table 6.1. We also illustrate example images from the test dataset, along with their generated paired descriptions in Figure. 6.4. An interesting caveat within the presented results show the systems pros and possible cons over

existing frameworks. For example, the bottom left image successfully identifies the subjects of the image as female. Whereas the bottom right does not identify people. This could be the presented image captioning pipeline and its ability to annotate or not annotate depending upon its confidence measures and the overall complexity of the image. I.e. the system being unsure on the gender of the subjects and avoiding reporting people because of this. Or could be a limitation in the training set used, I.e. Having seen similar sentence structures in regards to 'X' females, but not in regards to multiple genders in different locations and poses.

The proposed architecture is designed and motivated to expand upon the short captions produced by existing research. The framework from Chapter 5 is also presented in these evaluation results. This system is tested upon the larger test set utilized throughout this chapter, which accounts for the difference between the 2 chapters results. As indicated in Table 6.1, our Chapter 5 system shows the largest improvement over existing work is in the BLEU-1 metric. However, this framework also shows a substantial improvement over other metrics and related 3rd party works. This could be attributed to the addition of attribute prediction, or the prediction of more attributes, in comparison with those of the existing methods. On top of this, generating individual words that are more likely to be present in the reference sentence would increase the lower n -gram BLEU score (e.g. BLEU-1), which takes the frequency of the words and the length of the description into account for score generation.

On the other hand, generating words or attributes, which are correct, but may not be present within the reference description, reduces the score within the higher n -gram metrics, such as BLEU-3 or BLEU-4. This effect is most noticeable within the ROUGE-L score. Our system has the capability to generate multiple attributes for a given object. If the reference description only contains one or two attributes, and our system generates more than that, our score in the higher n -grams would be penalized. A simplified example is given below, which illustrates example theoretical descriptions generated by the proposed model and a typical existing framework such as Google NIC.

The proposed model: "a young man wearing a red striped shirt"

An existing method: "a man wearing a shirt"

Reference: "a man wearing a red shirt"

The BLEU-1 score is calculated for each word, so, for example, each generated word would score a precision of $x/(\text{the length of the corresponding generated sentence by a specific method})$, depending upon its frequency. Only the word ‘*red*’ in the above example of the proposed model would receive the precision score of this word, whereas the existing method would receive 0 owing to the fact that the attribute ‘*red*’ is not generated by the existing method. Therefore, our model scores well for the BLEU-1 metric, however the score suffers with the higher n -grams, since the correct ‘*red striped*’ does not appear in the reference. This is attributed to the nature in which our system is trained and built. Having trained separately and solely in sections of large scale attribute datasets, this enables our model to put more focus into attribute annotating than existing methods, even if some attributes are not present in the reference corpus.

6.3.1 Experimental Results

The empirical results for the evaluation of the IAPR TC-12 dataset indicate that the proposed system outperforms the two baseline methods of similar structures. The BLEU-1 score obtained by all the methods is lower than the human performance of around 0.6, however this is to be expected due to the cross-dataset evaluation. The BLEU metric provides an insight into the similarity comparison of the words’ and short phrases’ levels. The higher levels of BLEU metrics indicate the comparison of longer strings in the source and target sentences.

The comparatively higher scores throughout all of the BLEU metrics show how our work outperforms its competition. Specifically, for the BLEU evaluation, our work outperforms NeuralTalk by an average of 0.035 and NIC by an average of 0.053. The proposed system also outperforms the two baseline methods for the METEOR measures. In comparison to BLEU, the METEOR metric has gained increasing popularity owing to the closer correlation between the sentence level information, to human performance.

Moreover, the ROUGE-L metric is also used for evaluation. This metric is looking for the longest matching subsets between the automatically generated captioning and the human annotation. Although the captions generated by the different methods show great distinctions, the three systems achieve identical ROUGE-L scores.

The SPICE metric is also used for evaluation. It determines the semantic similarity between the pair of a generated description and its ground truth annotation. The proposed system achieves the lower score for the SPICE metric, in comparison to those of other metrics. This could be caused by the lengthy descriptions generated by the proposed system which may challenge and affect the semantic similarity score calculation in SPICE. SPICE utilizes scene graphs. The longer reference and generated descriptions could make a greater difference between these graphs, making a high similarity harder to achieve.

Overall, our proposed deep network outperforms systems of a similar structure, when all methods have been tested on different images to their training sets. This shows that our system has sufficient diversity and possesses the ability to generate descriptive captions for real-life and staged images. As can be seen in Figure. 6.4, our results are also considerably longer and more descriptive, and in many cases correct, in comparison to those generated by related methods.

6.4 Summary

In this research, we have proposed a novel deep network architecture for region annotations and full image description generation. The proposed model consists of a set of deep networks, including the regional proposal generator, CNNs and RNN-based encoder-decoder, to achieve a high level of quality for image description generation. By employing a regional approach, the proposed system is able to collect, annotate and describe a large number of details missed by other typical methods. It also requires fewer training images that a typical end-to-end system would require to achieve the same level of descriptive output. For example, our system is trained on subsets of existing datasets with a total of less than 200,000 image description pairs, as opposed too many existing image caption sets which consist of 200,000 image caption pairs and continuing upwards of one million image caption pairs. The proposed framework has also shown its significance in dealing with out-of-domain datasets, which challenge other state-of-the-art methods significantly, as shown in the evaluation of the IAPR TC-12 dataset. The overall architecture of our model is complex, combining multiple techniques and procedures to deliver effective image description generation. In future work, exploration in reducing the number of layers, model stages, and the feature complexity will be conducted to improve the system efficiency and runtime, and potentially the results. We were also able to pinpoint

potential issues with the evaluation metrics in regards to the aims and objectives of this research. For example, adding an attribute or multiple attributes before an object or a person label that are not present within the reference could lead to a large deterioration within the metric scores. This makes the scores even more impressive in comparison to the related state-of-the-art methodologies.

In future work, we aim to explore another advanced deep network, i.e. Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2016), because of its superior capability for image generation. We aim to explore its adaptation for image description generation owing to its unique style of training. Specifically, GANs are composed of two models i.e. the generative and discriminative models, which are trained simultaneously. The latter estimates that some data belongs to the training set, or some generated by the generative model. The generative model is trained to maximize the probability of the discriminative model making a mistake. Such training mechanisms may benefit image description generation tasks as well. We also intend to incorporate an attention or saliency mechanism into the region proposal generation stage to improve upon the quality of the generated regions of interest.

Finally, we also aim to produce a large-scale image description dataset that is more descriptive and discriminative than existing publicly available datasets. This would allow for not only a more accurate representation and evaluation of our model, but also further research into the descriptive caption generation rather than the typically available short captions.

Chapter 7 Conclusion and Future work

In this thesis, a number of research contributions are presented through the multiple chapters and research work undertaken.

The research within this thesis spans many active domains; however, all are needed to successfully caption images in the domains and to achieve the functionality desired. The scope of this thesis has aimed to stay focused around the regional based approach to adding more detail to this description and captioning domain, however highlighting where needed the importance and the work within the required related areas.

7.1 Summary of Contributions

The proposed research has made several contributions within the field of image description and conducted initial experimentation into its applicability in fields such as robotics and connected healthcare. First, we have proposed a composite pipeline in which longer and more detailed image captions are produced. This is based on a regional approach in which multiple regional features are extracted that can be utilised for independent attribute prediction to improve upon the relationship between human and generated captions. We also propose a novel fully supervised deep learning framework which combines the regional and attribute aspects of the system within Chapter 5, improves upon the sentence construction, and directly extends and improves upon the architecture by proposing the use of RNNs to translate the generated regional information into more refined and useful descriptions. The evaluation results show that the proposed system outperforms other state-of-the art methodologies, such as NeuralTalk by 30% and NIC by 70%, utilising both holistic and attention based features, in both in-domain and cross-dataset functionality evaluation, such as oil paintings. Highlighting our research towards our fourth main contribution highlighted in Chapter 1. We have also proposed, for our final major contribution a novel deep learning architecture for image region annotation in Chapter 6. It succeeds in not only generating regional annotations but also integrating the regional captioning into full image descriptions. This proposed deep network has a more efficient hierarchical training process and shows great robustness and efficiency in dealing with out of and cross domain images. The empirical evaluation results show how this novel methodology can compete with attention based regional models that utilize the typical end-to-end fashion of image captioning. Where possible these

Conclusion

models have been deployed upon the existing Linux based desktop machines, as well as the Nao Humanoid Robot, showing the potential for these systems to be deployed across multiple systems while maintaining their accuracy.

7.1.1 Composite methodology

In order to increase the length and descriptive capabilities of existing systems, we initially proposed a composite methodology, that is able to add multiple attributes and label most/all entities within an image if the parameters of the pipeline allow.

In **Chapter 4** an initial image captioning framework is proposed that had been equipped with the ability to annotate and describe entities within images that other systems would generally miss, due to the nature of the architecture in which only holistic image features are extracted. This employs a 4-stage system pipeline. The first stage implements scene classification, which is paired with subsequent object detection, classification and localisation with the use of bounding boxes. With this information, the third stage can allow for features to be extracted before attributes are classified with multiple SVRs. These generations produce a number of labels. When these are structured with the use of templates, a full image description can automatically be generated. This system pipeline had also been paired in additional experiments with a 3D intelligent agent. Two experiments were conducted to show its proficiency at image description and hazard labelling. The first explored image description on a small sample of images from two large scale datasets. This small set up was chosen as it allowed human annotators to describe the image in detail. The detailed descriptions were more in tune to our developments rather than the incredibly short captions available within the original datasets. The second experiment was conducted on healthcare images, to determine that the system could help identify and annotate its surroundings. It could be used to detect if a person had fallen and potentially require assistance, as well as detecting and reporting objects and/or hazards present on the floor that could be used to identify how a user had fallen or to be taken as a preventative measure.

7.1.2 Refined Descriptions

Chapter 5 proposes a direct improvement upon the previous chapter. The previous proposal works well under its given circumstances, whereas the use of templates can have a negative effect when testing on images out of domain. To this end, we employ

Conclusion

a machine translation approach to generate the image captions rather than fixed templates. The addition of two RNNs in the structure of an encoder-decoder allows for the translation of source to target language. This typically takes on a sentence-to-sentence format. However, in our proposal system, we treat the concatenated detailed regional annotations as the source and produce a descriptive and correctly structured output description. This method proposes a regional based method similar to chapter 4, however utilises the generated dense labels from each individual region in which the system is confident in and ‘translates’ this into full image descriptions. This has a number of benefits over the existing template based approach such as, removing the rigidity from the descriptions and making them sound more natural and ‘human like’. This however is more costly to train and execute, as well as requiring additional stages when medical or healthcare images are required to be analysed. The initial experiments and training of this system show that this pipeline can outperform similar and state-of-the-art competitors who rely on holistically extracted features and approaches on most of the implemented metrics, while remaining competitive on all other metrics. Transfer learning was also explored, with the proposed methodology showing its ability to annotate and caption images in an entirely different domain with minimal network modifications.

7.1.3 Hierarchical Architecture

Chapter 6 proposes another novel model architecture for image caption or description generation. Building upon the aspects learned from the previous chapters, this model employs a hierarchical training procedure, which is the unique way in which it is trained on multiple separate unrelated datasets. This allows multiple domains to be covered, and allows the generated descriptions to be more closely related to its contents regardless of the image content. The system is split into two distinct sections, the first utilising region generation, and simultaneous region annotation with the use of RNN attribute predictors, CNN object classifiers and scene label generation. With this collected information, the model can proceed to translate these regional annotations into full output image descriptions. The system can generate regional annotations and integrate the regional captioning into full image descriptions. This proposed deep network has a more efficient hierarchical training process and shows great robustness throughout the evaluation. The evaluation results show that our

Conclusion

framework can outperform state-of-the-art methodologies by an average of 0.044 on the BLEU score.

We also propose the use of such a system that can utilise the visual power of a humanoid robot, that could effectively collect images within its vision and describe them. This utilises the NAO H25s visual and textual APIs in combination with the deep learning algorithms running as required on a high-end workstation. Pairing the two over a LAN allows for minimal network communication, simply transferring the image one way and the textual information as its return. The initial experimentation indicates that the time required for processing would not be of detriment to its use.

7.1.4 Potential Limitations

Due to the nature of the aims of this research, and the training methodologies therein, the proposed systems could lack certainty in specific domains, as our systems are not trained and tested within one domain.

In terms of the initial system proposed in Chapter 4, the main limitation is the size of the model, which is addressed later in this section. However, another limiting factor is the time taken to process images. This can have a number of downsides both in research but especially real life deployment as in one of the conducted experiments. The initial methodology can take 10 seconds to produce an output. This would require dramatic improvements were this particular system to be utilised elsewhere.

The nature of our generated descriptions and the methods that are undertaken in order to train the proposed models make meeting the requirements of the metrics very difficult, and that directly comparing relating works requires both to be testing on out of domain images. Therefore, the IAPR dataset, has been widely used as an evaluation set. No images from this dataset were used during training, and the average length and nature of the descriptions are much more in line with the initial aims of this research.

Another potential limitation of this research also applies to the general deep learning domain, the sheer requirement of computational resources. These deep learning models require powerful and expensive GPUs in order to train and test. Even with powerful hardware, the computational time required in order to train them to the desired performance can take several weeks. Reducing the overall computational cost would be a potential direction, which is further discussed in the following section.

Conclusion

A potential limitation could come down to the metrics, due to reasons already highlighted in Chapter 4, but also as mentioned within the research from (van Miltenburg and Elliott, 2016). This would direct researchers to try and understand further where individual models differ and to what aspects are the individual strengths and weaknesses. This research highlights some of the errors which our work has been prone to, and provides motivation to overcome these issues. For example, even though our attribute predictors can correctly label gender, when translated into full image descriptions, this can be altered based on the sentence structure, meaning that gender can be incorrectly labelled.

7.2 Future Work

There are several directions this research can take within the future. Some of those directions as well as possible challenges and benefits are outlined below. As an extension to some of the previously conducted experiments, training an end-to-end style model with both a discriminator and a generator, or a GAN could be an interesting and powerful method to generate image captions. In the literature, it is stated that producing textual based outputs, or pairing a GAN with RNNs is a difficult challenge, in very recent work RNNs have been utilised in tandem with this technology to some degree, however not in the terms of image captioning or description.

Most of the experiments conducted within this research were conducted and tested upon images that are out of domain, or those with which the system is trained, but are not tailored to one specific cause. Healthcare is an aspect that has been mentioned a number of times within this thesis, with other areas such surveillance and police use briefly cited. This could lead to the possibility of training these models on a specific function to provide a niche real-world application. Image captioning has a varied number of uses, however many of the images explored initially have been real-world and in the wild captured images.

If the systems were trained on x-ray images, a similar technique could be used, especially if the data was available to describe and annotate these images. This could lead to X-ray images being described in layman terms without medical jargon. This could also be used as an assistant to a medical professional in which areas of interest are noted and described for a thorough follow up by a trained medical individual.

Conclusion

The same method could apply to person re-identification. Training the attribute detectors on factors that are more specific or relating to an individual could again reduce workload for those professionals. In the presented research, low resolution images have led to inaccurate proposals resulting in inaccurate captions. In terms of CCTV images, low resolution may be standard. Proposing a method that is less susceptible to the resolution could provide many benefits across all image captioning domains. Further to this, creating a system that could work entirely out of domain, or recognising and correcting itself with the use of novel class detection unsupervised learning could create a system that could generate for any image, regardless of source, of any type for minimal compute cost.

Another area of interest, in which this application could be taken forward was mentioned previously in Chapter 4, i.e. annotating and describing entirely out of domain images, such as paintings or drawings. This could also cover areas of transfer learning which could improve and have an effect on the way these kinds of systems are utilised. For example, utilising transfer learning techniques, it could be possible to train an image caption system on real life images, as well as some form of painting dataset and allow fully descriptive captions to be produced of a given painting. This is a difficult challenge due to the nature of art and paintings. There are a number of styles, artists and representations, as an abstract drawing of a person may have little to no similarity with that of a real life image. This again would provide a large amount of benefit to those visually impaired users, which could lead to an easily accessible tool to annotate and describe anything a user intended, whether it be an image captured of a real life scene, or of a work of art in a traditional museum.

Another potential direction for future work would be improving the utilisation of the humanoid robot. This use case would require the use of low resolution images, depending upon the hardware within the robot at test time. Currently this thesis has only covered relatively small scale experiments in which the robot has been used, extending this as well as exploring the possibility of modifying the code and frameworks to newer technologies which are optimised for small processors such as those found in mobile phones, such as TensorFlow.

Producing a ‘one for all’ style dataset that could be used for detailed description generation. A large scale paired image caption dataset which contains attributes,

Conclusion

objects, regions etc would allow for the creation of very powerful and complex models. This would remove the requirement of the hierarchical training methodology implemented in Chapter 6 and allow for an end-to-end approach which could further reduce the compute cost of such a system. Collecting the dataset would be an immense challenge but could be achieved in a number of ways, such as: crowd sourced through systems such as Amazon Mechanical Turk or Crowdfunder, collecting all of the above functionality would be time consuming and expensive, however would produce accurate and correct annotations once verified. An alternate route would be to build upon existing datasets, combining images, attributes, captions and utilising systems as previously mentioned to fill in any gaps or to verify the content and the paired data.

There are many research domains that are reliable, such as object classification and localisation as well as face detection etc. This could lead to automating the collection of some of these areas and labels, which could reduce the overall load and cost impact of creating such a dataset.

Conclusion

References

- A. L. I. C. E. *The Artificial Linguistic Internet Computer Entity* (1995). Available at: <http://www.alicebot.org/about.html> (Accessed: 28 September 2017).
- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X. and Brain, G. (2016) ‘TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning’, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pp. 265–284.
- Aldebaran (2017) *Aldebaran Robotics / Humanoid robotics & programmable robots, 2015*. Available at: <https://www.aldebaran.com/en> (Accessed: 28 September 2017).
- Anderson, P., Fernando, B., Johnson, M. and Gould, S. (2016) ‘Spice: Semantic propositional image caption evaluation’, in *European Conference on Computer Vision*, pp. 382–398.
- Aslandogan, Y. A. and Yu, C. T. (1999) ‘Techniques and systems for image and video retrieval’, *IEEE Transactions on Knowledge and Data Engineering*, 11(1), pp. 56–63.
- Bahdanau, D., Cho, K. and Bengio, Y. (2014) ‘Neural Machine Translation by Jointly Learning to Align and Translate’.
- Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A. and Plank, B. (2017) ‘Automatic description generation from images: A survey of models, datasets, and evaluation measures’, in *IJCAI International Joint Conference on Artificial Intelligence*, pp. 4970–4974.
- Bourdev, L., Maji, S., Brox, T. and Malik, J. (2010) ‘Detecting people using mutually consistent poselet activations’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Berlin, Heidelberg, pp. 168–181.
- Bourdev, L., Maji, S. and Malik, J. (2011) ‘Describing people: A poselet-based approach to attribute classification’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1543–1550.

Caffe2 – *Facebook Research* (2017). Available at: <https://research.fb.com/downloads/caffe2/> (Accessed: 22 July 2018).

Chang, C. and Lin, C. (2011) ‘A Library for Support Vector Machines’, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), p. 39.

Chatfield, K., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) ‘Return of the Devil in the Details: Delving Deep into Convolutional Nets’, *robots.ox.ac.uk*.

Chaudhuri, B. B. and Bhattacharya, U. (2000) ‘Efficient training and improved performance of multilayer perceptron in pattern classification’, *Neurocomputing*, 34, pp. 11–27.

Chen, J., Dong, W. and Li, M. (2014) ‘Image Caption Generator Based On Deep Neural Networks’.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P. and Zitnick, C. L. (2015) ‘Microsoft COCO Captions: Data Collection and Evaluation Server’.

Chen, X. and Zitnick, C. L. (2015) ‘Mind’s eye: A recurrent visual representation for image caption generation’, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2422–2431.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) ‘Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation’.

Chollet, F. (2015) ‘Keras (2015)’, URL <http://keras.io>.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014) ‘Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling’, *arXiv preprint arXiv:1412.3555*.

Collobert, R., Farabet, C. and Kavukcuoğlu, K. (2008) *Torch / Scientific computing for LuaJIT.*, *NIPS Workshop on Machine Learning Open Source Software*. Available at: <http://torch.ch/>.

Cortes, C. and Vapnik, V. (1995) ‘Support-Vector Networks’, *Machine learning*, 20(3), pp. 273–297.

Denkowski, M. and Lavie, A. (2014) ‘Meteor Universal: Language Specific

Translation Evaluation for Any Target Language’, in *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380.

Dhar, S., Ordonez, V. and Berg, T. L. (2011) ‘High level describable attributes for predicting aesthetics and interestingness’, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1657–1664.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V. (1997) ‘Support vector regression machines’, in *Advances in Neural Information Processing Systems*, pp. 155–161.

Elliott, D., Frank, S., Sima’an, K. and Specia, L. (2016) ‘Multi30K: Multilingual English-German Image Descriptions’, *arXiv preprint arXiv:1605.00459*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. (2010) ‘The pascal visual object classes (VOC) challenge’, *International Journal of Computer Vision*, 88(2), pp. 303–338.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L. and Zweig, G. (2015) ‘From captions to visual concepts and back’, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1473–1482.

Farhadi, A., Endres, I., Hoiem, D. and Forsyth, D. (2009) ‘Describing objects by their attributes.’, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE*, pp. 1778–1785.

Florea, C., Condorovici, R., Vertan, C., Butnaru, R., Florea, L. and Vrânceanu, R. (2016) ‘Pandora: Description of a painting database for art movement recognition with baselines and perspectives’, in *European Signal Processing Conference*, pp. 918–922.

Gan, C., Gan, Z., He, X. and Gao, J. (2017) ‘StyleNet: Generating Attractive Visual Captions with Styles’, in *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pp. 3137–3146.

Girshick, R. (2016) ‘Fast R-CNN’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2016) ‘Region-Based Convolutional Networks for Accurate Object Detection and Segmentation’, *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 38(1), pp. 142–158.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2016) ‘Generative adversarial nets’, in *Advances in neural information processing systems*, pp. 2672–2680.

Graves, A. and Schmidhuber, J. (2005) ‘Framewise phoneme classification with bidirectional LSTM and other neural network architectures’, *Neural Networks*, 18(5–6), pp. 602–610.

Grübinger, M., Clough, P., Müller, H. and Deselaers, T. (2006) ‘The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems’, in *LREC Workshop OntoImage Language Resources for Content-Based Image Retrieval*, pp. 13–23.

Hameed, I. A. (2017) ‘Using natural language processing (NLP) for designing socially intelligent robots’, in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2016*. IEEE, pp. 268–269.

He, K., Zhang, X., Ren, S. and Sun, J. (2014) ‘Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition’, pp. 346–361.

Hemmat Esfe, M., Afrand, M., Yan, W. M. and Akbari, M. (2015) ‘Applicability of artificial neural network and nonlinear regression to predict thermal conductivity modeling of Al₂O₃-water nanofluids using experimental data’, *International Communications in Heat and Mass Transfer*, 66, pp. 246–249.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K. and Darrell, T. (2016) ‘Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K. and Darrell, T. (2016) ‘Natural Language Object Retrieval’, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4555–4564.

Hwang, K. and Sung, W. (2017) ‘Character-level language modeling with hierarchical recurrent neural networks’, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5720–5724.

Jia, X., Gavves, E., Fernando, B. and Tuytelaars, T. (2015) ‘Guiding the long-short term memory model for image caption generation’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2407–2415.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T. (2014) ‘Caffe: Convolutional Architecture for Fast Feature Embedding’, in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675–678.

Johnson-Roberson, M., Bohg, J., Skantze, G., Gustafson, J., Carlson, R., Rasolzadeh, B. and Kragic, D. (2011) ‘Enhanced visual scene understanding through human-robot dialog’, in *IEEE International Conference on Intelligent Robots and Systems*, pp. 3342–3348.

Johnson, J., Karpathy, A. and Fei-Fei, L. (2016) ‘Densecap: Fully convolutional localization networks for dense captioning’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) ‘Large-scale video classification with convolutional neural networks’, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

Karpathy, A. (2015) *The Unreasonable Effectiveness of Recurrent Neural Networks*.

Karpathy, A. (2016) *Char-Rnn*. Available at: <https://github.com/karpathy/char-rnn> (Accessed: 28 September 2017).

Karpathy, A. and Fei-Fei, L. (2015) ‘Deep visual-semantic alignments for generating image descriptions’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137.

Khan, M. U. G., Al Harbi, N. and Gotoh, Y. (2015) ‘A framework for creating natural language descriptions of video streams’, *Information Sciences*, 303, pp. 61–82.

Kinghorn, P., Zhang, L. and Shao, L. (2017a) ‘A hierarchical and regional deep learning architecture for image description generation’, *Pattern Recognition Letters*.

Kinghorn, P., Zhang, L. and Shao, L. (2017b) ‘Deep learning based image description generation’, in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 919–926.

Kinghorn, P., Zhang, L. and Shao, L. (2018) ‘A region-based image caption generator with refined descriptions’, *Neurocomputing*, 272, pp. 416–424.

Kotsiantis, S. B. (2007) ‘Supervised Machine Learning: A Review of Classification Techniques’, *Informatica*, 31, pp. 249–268.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L. and Niebles, J. C. (2017) ‘Dense-Captioning Events in Videos’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. and others (2017) ‘Visual genome: Connecting language and vision using crowdsourced dense image annotations’, *International Journal of Computer Vision*. Springer US, 123(1), pp. 32–73.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) ‘ImageNet Classification with Deep Convolutional Neural Networks’, *Advances in Neural Information and Processing Systems (NIPS)*, 60(6), pp. 84–90.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. and Berg, T. (2013) ‘Babytalk: Understanding and generating simple image descriptions’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), pp. 2891–2903.

Kumar, N., Berg, A. C., Belhumeur, P. N. and Nayar, S. . (2009) ‘Attribute and simile classifiers for face verification’, in *2009 IEEE 12th International Conference on Computer Vision*.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) ‘Gradient-based learning applied to document recognition’, *Proceedings of the IEEE*, 86(11), pp. 2278–2323.

LeCun, Y., Haffner, P., Bottou, L. and Bengio, Y. (1999) ‘Object Recognition with Gradient-Based Learning’, in Springer, Berlin, Heidelberg, pp. 319–345.

Li, L., Tang, S., Deng, L., Zhang, Y. and Tian, Q. (2017) ‘Image Caption with Global-Local Attention’, in *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 4133–4139.

Li, S., Kulkarni, G., Berg, T., Berg, A. and Choi, Y. (2011) ‘Composing simple image descriptions using web-scale n-grams’, in *Conference on Computational Natural Language Learning. Association for Computational Linguistics*, pp. 220–228.

- Li, S., Xiao, T., Li, H., Zhou, B., Yue, D. and Wang, X. (2017) ‘Person search with natural language description’, in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 5187–5196.
- Liang, S., Li, X., Zhu, Y., Li, X. and Jiang, S. (2017) ‘ISIA at the ImageCLEF 2017 image caption task’, in *CEUR Workshop Proceedings*.
- Liang, X., Hu, Z., Zhang, H., Gan, C. and Xing, E. P. (2017) ‘Recurrent Topic-Transition GAN for Visual Paragraph Generation’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3382–3391.
- Lin, C. Y. (2004) ‘Rouge: A package for automatic evaluation of summaries’, in *Proceedings of the workshop on text summarization branches out (WAS 2004)*, pp. 25–26.
- Lin, D., Fidler, S., Kong, C. and Urtasun, R. (2015) ‘Generating Multi-sentence Natural Language Descriptions of Indoor Scenes’, in *Proceedings of the British Machine Vision Conference 2015*, p. 93.1-93.13.
- Lin, T.-Y., Marie, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P. and Zitnick, C. L. (2015) ‘Microsoft COCO : Common Objects in Context’, in *European conference on computer vision*, pp. 740–755.
- Lipton, Z., Berkowitz, J. and Elkan, C. (2015) ‘A Critical Review of Recurrent Neural Networks for Sequence Learning’, *arXiv preprint arXiv:1506.00019*.
- Liu, C., Hu, C., Liu, Q. and Aggarwal, J. K. (2013) ‘Video event description in scene context’, *Neurocomputing*, 119, pp. 82–93.
- Lopez, A. (2008) ‘Statistical machine translation’, *ACM Computing Surveys*, 40(3), pp. 1–49.
- Lu, H. C., Fang, G. L., Wang, C. and Chen, Y. W. (2010b) ‘A novel method for gaze tracking by local pattern model and support vector regressor’, *Signal Processing*, 90(4), pp. 1290–1299.
- Lu, J., Xiong, C., Parikh, D. and Socher, R. (2017) ‘Knowing when to look: Adaptive attention via a visual sentinel for image captioning’, in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 3242–3250.

Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K. M. and Zhong, Y. (2017) ‘Person re-identification by unsupervised video matching’, *Pattern Recognition*, 65, pp. 197–210.

MacLeod, H., Bennett, C. L., Morris, M. R. and Cutrell, E. (2017) ‘Understanding Blind People’s Experiences with Computer-Generated Captions of Social Media Images’, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. New York, New York, USA: ACM Press, pp. 5988–5999.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuille, A. (2014) ‘Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)’.

Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z. and Yuille, A. L. (2015) ‘Learning like a child: Fast novel visual concept learning from sentence descriptions of images’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2533–2541.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. and Murphy, K. (2016) ‘Generation and Comprehension of Unambiguous Object Descriptions’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20.

Mathews, A., Xie, L. and He, X. (2016) ‘SentiCap: Generating Image Descriptions with Sentiments’, in *Thirtieth AAAI Conference on Artificial*, pp. 3574–3580.

Matsuo, E., Kobayashi, I., Nishimoto, S., Nishida, S. and Asoh, H. (2016) ‘Generating Natural Language Descriptions for Semantic Representations of Human Brain Activity’, in *ACL 2016*, pp. 22–29.

Meyer, D. and Technikum Wein, F. (2001) ‘Support Vector Machines *’, *R News*, 1(3), pp. 23–26.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J. and Sanjeev Khudanpur (2010) ‘Recurrent neural network based language model’, in *Eleventh Annual Conference of the International Speech Communication Association*.

Miller, G. A. (1995) ‘WordNet: a lexical database for English’, *Communications of the ACM*, 38(11), pp. 39–41.

van Miltenburg, E. and Elliott, D. (2016) ‘Room for improvement in automatic image

description: an error analysis’, *arXiv*.

Mopuri, K. R., Athreya, V. B. and Babu, R. V. (2017) ‘Deep image representations using caption generators’, *arXiv preprint arXiv:1705.0914*.

Ordonez, V., Han, X., Kuznetsova, P., Kulkarni, G., Mitchell, M., Yamaguchi, K., Stratos, K., Goyal, A., Dodge, J., Mensch, A., Daumé, H., Berg, A. C., Choi, Y. and Berg, T. L. (2016) ‘Large Scale Retrieval and Generation of Image Descriptions’, *International Journal of Computer Vision*. Springer US, 119(1), pp. 46–59.

Orhan, U., Hekim, M. and Ozer, M. (2011) ‘EEG signals classification using the K-means clustering and a multilayer perceptron neural network model’, *Expert Systems with Applications*. Pergamon, 38(10), pp. 13475–13481.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) ‘BLEU: a method for automatic evaluation of machine translation’, in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318.

Pedersoli, M., Lucas, T., Schmid, C. and Verbeek, J. (2017) ‘Areas of Attention for Image Captioning’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1251–1259.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2015) ‘You Only Look Once: Unified, Real-Time Object Detection’, pp. 779–788.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H. (2016) ‘Generative Adversarial Text to Image Synthesis’, *arXiv preprint arXiv:1605.05396*.

Ren, S., He, K., Girshick, R. and Sun, J. (2015) ‘Faster r-cnn: Towards real-time object detection with region proposal networks’, in *Advances in neural information processing systems*, pp. 91–99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and others (2015) ‘Imagenet large scale visual recognition challenge’, *International Journal of Computer Vision*. Springer US, 115(3), pp. 211–252.

Russakovsky, O. and Fei-Fei, L. (2012) ‘Attribute learning in large-scale datasets’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 1–14.

Sande, K. Van de (2011) ‘Segmentation as selective search for object recognition’, *(ICCV), 2011, (2)*, pp. 1879–1886.

Schuster, S. and Manning, C. D. (2016) ‘Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks’, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2371–2378.

Shao, L., Zhu, F. and Li, X. (2015) ‘Transfer Learning for Visual Categorization: A Survey’, *IEEE transactions on neural networks and learning systems*, 26(5), pp. 1019–1034.

Shen, J., Liu, G., Chen, J., Fang, Y., Xie, J., Yu, Y. and Yan, S. (2014) ‘Unified structured learning for simultaneous human pose estimation and garment attribute classification’, *IEEE Transactions on Image Processing*, 23(11), pp. 4786–4798.

Shen, Y. and Huang, X. (2016) ‘Attention-Based Convolutional Neural Network for Semantic Relation Extraction’, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2526–2536.

Sheshadri, Aashish, Endres, I. (2012) ‘Describing Objects by their Attributes’, *Computer Vision and*, pp. 1778–1785.

Smeaton, A. F. and Quigley, I. (1996) ‘Experiments on using semantic distances between words in image caption retrieval’, *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, New York, USA: ACM Press, pp. 174–180.

Socher, R., Karpathy, A., Le, Q., Manning, C. and Ng, A. (2014) ‘Grounded Compositional Semantics for Finding and Describing Images with Sentences’, *Transactions of the Association for Computational Linguistics*, 2(1), pp. 207–218.

Sugano, Y. and Bulling, A. (2016) ‘Seeing with Humans: Gaze-Assisted Neural Image Captioning’, *arXiv preprint arXiv:1608.05203*.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014) ‘Sequence to Sequence Learning with Neural Networks’, *In Advances in Neural Information Processing Systems (NIPS 2014)*, pp. 1–9.

Talukdar, J. and Mehta, B. (2018) ‘Human action recognition system using good

features and multilayer perceptron network’, in *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, pp. 317–323.

Tan, W. R., Chan, C. S., Aguirre, H. E. and Tanaka, K. (2016) ‘Ceci n’est pas une pipe: A deep convolutional network for fine-art paintings classification’, in *Proceedings - International Conference on Image Processing, ICIP*, pp. 3703–3707.

Tan, Y. H. and Chan, C. S. (2017) ‘Phi-LSTM: A phrase-based hierarchical LSTM model for image captioning’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 101–117.

Theano Development Team (2016) ‘Theano: A Python framework for fast computation of mathematical expressions’, *arXiv e-prints*, p. 19. Available at: <http://arxiv.org/abs/1605.02688>.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K. and Mooney, R. J. (2014) ‘Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild.’, in *Coling*, p. 9.

Torabi, A., Pal, C., Larochelle, H. and Courville, A. (2015) ‘Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research’, *arXiv preprint arXiv:1503.01070*.

Tran, K., He, X., Zhang, L. and Sun, J. (2016) ‘Rich Image Captioning in the Wild’, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 434–441.

Vedantam, R., Lawrence Zitnick, C. and Parikh, D. (2015) ‘Cider: Consensus-based image description evaluation’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K. (2015) ‘Sequence to sequence - Video to text’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4534–4542.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. and Saenko, K. (2015) ‘Translating Videos to Natural Language Using Deep Recurrent Neural

Networks’, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1494–1504.

Verma, Y. and Jawahar, C. V (2014) ‘Im2Text and Text2Im : Associating Images and Texts for Cross-Modal Retrieval’, *Bmvc*, p. 89.1-89.13.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2014) ‘Show and Tell: A Neural Image Caption Generator’, *arXiv:1411.4555 [cs]*, abs/1411.4, pp. 3156–3164.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2017) ‘Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), pp. 652–663.

Wang, T., Gong, S., Zhu, X. and Wang, S. (2014) ‘Person re-identification by video ranking’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 688–703.

Wang, T., Gong, S., Zhu, X. and Wang, S. (2016) ‘Person Re-Identification by Discriminative Selection in Video Ranking’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12), pp. 2501–2514.

Wilde, E. (2007) ‘What are you talking about?’, in *2007 IEEE International Conference on Services Computing*, pp. 256–261.

Xiao, J., Hays, J., Ehinger, K. A. and Torralba, A. (2010) ‘SUN Database : Large-scale Scene Recognition from Abbey to Zoo’, in *Computer vision and pattern recognition (CVPR)*, pp. 3485–3492.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. and Bengio, Y. (2015) ‘Show, Attend and Tell: Neural Image Caption Generation with Visual Attention’, in *International conference on machine learning*, pp. 2048–2057.

Yang, H., Chan, L. and King, I. (2002) ‘Support Vector Machine Regression for Volatile Stock Market Prediction’, *Intelligent Data Engineering and Automated*, pp. 391–396.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016) ‘Hierarchical Attention Networks for Document Classification’, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pp. 1480–1489.

Yang, Z., Zhang, Y. J., ur Rehman, S. and Huang, Y. (2017) ‘Image captioning with object detection and localization’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 109–118.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A. (2016) ‘Describing videos by exploiting temporal structure’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4507–4515.

Ye, C., Yang, Y., Mao, R., Fermuller, C. and Aloimonos, Y. (2017) ‘What can i do around here? Deep functional scene understanding for cognitive robots’, in *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 4604–4611.

Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. (2014) ‘From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions’, *Transactions of the Association for Computational Linguistics (TACL)*, 2(April), pp. 67–78.

Yu, H., Wang, J., Huang, Z., Yang, Y. and Xu, W. (2015) ‘Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks’.

Yu, X., Qi, Z. and Zhao, Y. (2013) ‘Support vector regression for newspaper/magazine sales forecasting’, in *Procedia Computer Science*. Elsevier, pp. 1055–1062.

Zeiler, M. D. (2012) ‘ADADELTA: An Adaptive Learning Rate Method’.

Zheng, S., Cheng, M.-M., Warrell, J., Sturgess, P., Vineet, V., Rother, C. and Torr, P. H. S. (2014) ‘Semantic Image Segmentation with Objects and Attributes’, in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3214–3221.

Zitnick, C. L. and Dollár, P. (2014) ‘Edge boxes: Locating object proposals from edges’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Cham, pp. 391–405.

