Northumbria Research Link

Northumbria University NEWCASTLE

University Library

Interpreting Deep Learning based Cerebral Palsy Prediction with Channel Attention

Manli Zhu[1], Qianhui Men[2], Edmond S. L. Ho[1], Howard Leung[2], and Hubert P. H. Shum[3*]

[1]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK
[2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[3]Department of Computer Science, Durham University, Durham, UK

Abstract—**Early prediction of cerebral palsy is essential as it leads to early treatment and monitoring. Deep learning has shown promising results in biomedical engineering thanks to its capacity of modelling complicated data with its non-linear architecture. However, due to their complex structure, deep learning models are generally not interpretable by humans, making it difficult for clinicians to rely on the findings. In this paper, we propose a channel attention module for deep learning models to predict cerebral palsy from infants' body movements, which highlights the key features (i.e. body joints) the model identifies as important, thereby indicating why certain diagnostic results are found. To highlight the capacity of the deep network in modelling input features, we utilize raw joint positions instead of hand-crafted features. We validate our system with a real-world infant movement dataset. Our proposed channel attention module enables the visualization of the vital joints to this disease that the network considers. Our system achieves 91.67% accuracy, suppressing other state-of-the-art deep learning methods.**

Index Terms—**Cerebral palsy, deep learning, artificial neural network, channel attention**

I. INTRODUCTION

Cerebral palsy (CP) is one of the most common childhood neurodevelopmental disorders in the United States [1]. It affects people's movements, and has an impact on growth progression and life quality. It is estimated that between 3 to 10 out of 1,000 children develop some form of cerebral palsy depending on gestational age at birth [2]. Therefore, early diagnosis is essential as it leads to taking the treatment as early as possible for preventing bigger health issues. However, a definitive diagnosis can be challenging for clinicians as years of practical experience and training is required.

Early diagnosis of CP has been investigated by the diagnostic tool General Movements Assessment (GMA), however, significant time and resource investment is needed to train an assessor. Automated GMA has been studied in some works. Stahl et al. [3] applied optical flow and statistical pattern recognition on early video recordings of infants' spontaneous movements for later CP prediction. Wu et al. [4] extracted the infant body 2D key points from RGB images. From the depth information in RGB-D videos, they obtained the 3D

movements of infants in the supine position, and extracted the limb angle features for infant CP prediction.

Machine learning and deep learning-based automated systems for CP diagnosis have been proposed. McCay et al. [5] proposed the histograms of joint orientation 2D (HOJO2D) and joint displacement 2D (HOJD2D) features, for CP prediction by using machine learning methods such as K-nearest neighbors (KNN) and linear discriminant analysis (LDA). They further proposed five deep learning models [6] to demonstrate the prospect of using deep learning in disease prediction. As manually crafted features require considerable domain expertise and may not be transferable to another domain, we opt for raw features, i.e. the joint positions, and demonstrate the same level of accuracy with our model.

Although deep learning systems achieved impressive performance in disease diagnosis, it is hard for humans to understand how the decisions are made due to their black-box nature [7]. A medical diagnostic system needs to be interpretable that allowing physicians and regulators to have the confidence on the diagnostic result.

To this end, we propose a channel attention module that tells what features have been considered important in our deep learning model during predictions. Attention mechanisms [8] have been widely used in deep learning for feature modelling, but there are limited researches on using them as an interface for interpreting deep neural networks. We implement our attention module with the squeeze-and-excitation framework [8], which allows a single attention value to be learned for each channel. We propose a novel design to represent each joint as a channel, such that the attention value can be interpreted by humans.

We validate our system with an infant movement dataset [9] targeting for CP prediction. We show that the inclusion of our module not only allows users to interpret the network at a feature level, but also achieves state-of-the-art performance comparing to existing deep learning methods that use hand-crafted features, demonstrating a reliable framework to analyze and predict CP on infants. Our source code can be downloaded at *bit.ly/InterpretableDLforCP*.

II. METHODOLOGY

The overview of our proposed framework is illustrated in Fig. 1. First, the input of the network is a temporal
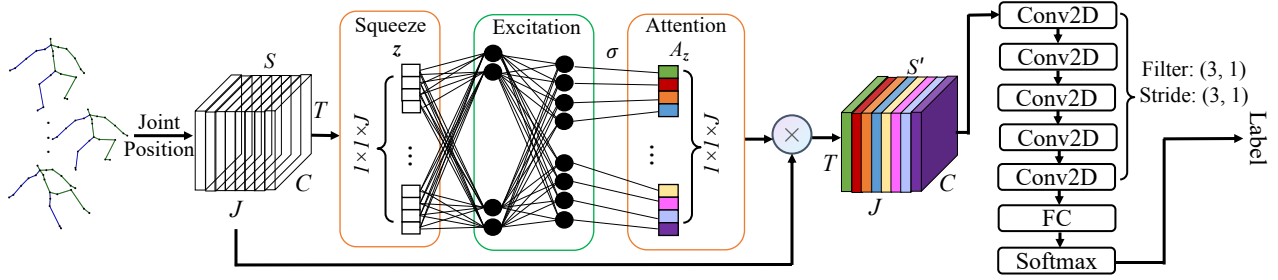
Fig. 1. The overview of our proposed framework.

sequence of joint positions represented as a tensor (i.e. a multi-dimensional matrix). It is then fed into an attention module including Squeeze and Excitation blocks, and the importance of each joint is represented by its respective channel attention value. The attention value is multiplied with the input feature, forming an attended feature tensor. Such a tensor is fed into five 2D convolutional layers (Conv2D) to further extract the high-level information. Finally, a fully-connected layer (FC) and a softmax layer are applied to perform the classification.

A. The Dataset

We conduct experiments on the public dataset Moving Infants In RGB-D (MINI-RGBD) [9], in which twelve sequences of real infant movements were captured. To protect the privacy of recorded infants, such movements are represented by synthetic infant characters generated using the Skinned Multi-Infant Linear (SMIL) model [10]. In this research, we use the provided 3D joint positions of the sequences as our input.

McCay et al. [5] further annotated the dataset, and each sequence was labeled as either *normal* or *abnormal* by an independent expert using the GMA method. We use such annotations as ground-truth.

To ensure that the classifier is not affected by the global translation of the infant, we normalize the joint positions frame by frame. This is done by translating the spine joint of each frame to the global origin, and representing other joints as a relative value to the spine.

B. Channel Attention

We propose an attention module based on the channel-wise attention mechanism [8] to provide insight into the network's decision process. As shown in our overall framework in Fig. 1, we first propose a squeeze step that aims at compressing the joint information, then an excitation operation is applied to identify the importance of each joint to the network, such that different weights are assigned to different joints.

Unlike traditional research to use channel attention for modelling features, we model the joints dimension intuitively as channels as a novel way for interpreting the network. On the one hand, the joint dimension provides a more meaningful interpretation to the users compared to the coordinate dimension. On the other hand, using a single attention value for the whole 3D trajectory of a joint in the entire video results

in a much more compact representation, making it easier for humans to follow and more stable to train the network. As a result, the designed attention module helps humans understand which joints contribute more to the diagnostic outcomes.

More specifically, the system input is represented as a feature tensor $S \in R^{T \times C \times J}$, in which T is the total number of frames, $C = 3$ represents the 3 dimensions of a joint position, J is the total number of joints. In S, each feature value is represented as $S(t, c, j)$, where $t \in T$, $c \in C$ and $j \in J$.

In the squeeze operation, the tensor from each joint is compressed into a representative joint descriptor using Global Average Pooling. For a specific joint $j' \in J$, its descriptor $z_{j'}$ is calculated as:

$$z_{j'} = \frac{1}{T \times C} \sum_{t=1}^{T} \sum_{c=1}^{C} S(t, c, j') \tag{1}$$

These descriptors form a joint-level embedding $\mathbf{z} = [z_1, \ldots, z_J]$ from the whole movements. The obtained global embedding enables the aggregation of the local joint features which provides expressive joint-level statistics that triggers the following excitation steps.

The excitation operation calculates the attention value that indicates the importance of each channel (i.e. joint) by encoding the correlations from the embedded output $\mathbf{z}$ in the squeeze step. This is done by allocating the inner dependencies among joints with a nonlinear gating mechanism:

$$A_{\mathbf{z}} = \sigma\left(\mathbf{W_2}\delta\left(\mathbf{W_1}\mathbf{z}\right)\right) \tag{2}$$

where δ is the ReLU activation, σ is the sigmoid activation function, $\mathbf{W_1}$ and $\mathbf{W_2}$ are parameters of the fully-connected layers.

The obtained $A_{\mathbf{z}} \in R^{1 \times 1 \times J}$ from the squeeze-and-excitation steps provides the joints attention, and each value in $A_{\mathbf{z}}$ indicates the attention weight of the corresponding joint. The attended feature tensor S' is obtained by multiplying $A_{\mathbf{z}}$ with the original feature tensor S.

C. The CP Prediction Network

We propose a 2D convolutional neural network to encode the tensor S' for CP prediction. The network consists of five convolutional layers, one fully-connected layer and one softmax layer, as shown in Fig. 1. We found that having multiple convolutional layers help to encode high-level features with

a suitable receptive field, and achieve the best classification accuracy. The fully-connected layer maps the learned feature with the CP labels (i.e. normal or abnormal), and the Softmax layer is to generate probability distributions for classification.

The filter K in the convolutional layers has the dimension of $F_T \times F_C \times J$, where $F_T = 3$ and $F_C = 1$ in our experiment. This means that the filter covers all joints at once, and moves in the frame and coordinate dimensions. This design ensures that the intrinsic correlations among joints can be discovered, and that the temporal correlations of nearby frames are modelled.

D. Loss Functions

We propose a novel loss function known as attention loss. Unlike the traditional usage of attention, we apply attention for deep neural network interpretation. To facilitate better human understanding, the number of attended features should be minimized, while keeping the classification accuracy unchanged. We define the attention loss as:

$$L_{\text{att}} = \sum_{j=1}^{J} A_{z_j} \tag{3}$$

The final loss function consists of three parts: the cross-entropy loss L_{cep} to encourage classification accuracy, the attention loss L_{att} to minimize the number of attended features, and the regularization loss $\|w\|_2$ (where w is the network parameters) to discourage overfitting:

$$L = L_{\text{cep}} + \gamma L_{\text{att}} + \lambda \|w\|_2 \tag{4}$$

where γ and λ are the loss weights set as 0.0005 and 0.0001 respectively. We deliberately chose a small γ such to minimize the attention loss effect to the classification accuracy.

As the data is biased with fewer abnormal samples (i.e. 4 abnormal, 8 normal), we introduce a class balancing weight in L_{cep}. The idea is to give a higher weight to the abnormal samples to facilitate training. Without the weights, the model would have poor predictive performance for the minority class (i.e. abnormal) when an imbalanced dataset is given, resulting in more false-negatives. In particular, we assign a weight α_i to the class i by the following equation:

$$\alpha_i = \sqrt{\frac{n}{n_c \times n_i}} \tag{5}$$

where n is the total number of the samples, n_i is the number of samples of class i, and $n_c = 2$ is the number of classes.

In our implementation, the whole network was trained in an end-to-end manner using the PyTorch platform with the Adams optimizer. The hyper-parameters *epoch*, *learning rate*, and *batch size* were set as 400, 0.0003 and 3 respectively for all cross-validations.

III. Experimental Results

A. Quantitative Evaluation

We perform the leave-one-out cross-validation on the MINI-RGBD dataset [9], which is also used in the baseline works [4], [6], and the averaged result for all cross-validations is presented as the final accuracy.

As deep learning has become the mainstream in pattern recognition due to its capacity to modelling complicated data, we compare the classification performance of our proposed deep learning method with the existing deep learning methods [6] as well as the most recent GMA-based method [4] that is also based on 3D skeletal data as input. The results are reported in Table I. It can be seen that our full system (i.e. the bottom row) achieves the same state-of-the-art performance, i.e. 91.67% prediction accuracy, in addition to the interpretability of our network which tells users which of the joints of an infant body are the most important for its diagnostic outcome. Compared with [6], we utilize raw features instead of hand-crafted ones (i.e. HOJO2D + HOJD2D that generates the best accuracy), which makes our method easy to be transferred to another domain as there is no feature engineering involved. Compared with [4], which does not require any training for the model but a threshold has to be defined manually to separate the *positive* and *negative* samples. In contrast, our model is trained end-to-end without the need for any human intervention.

TABLE I
Classification accuracy comparison between our proposed method and other baseline methods.

Feature	Method	Accuracy (%)
HOJO2D	FCNet [6]	83.33
HOJD2D	FCNet [6]	**91.67**
HOJO2D / HOJD2D	Conv1D-1 [6]	83.33
HOJO2D / HOJD2D	Conv1D-2 [6]	83.33
HOJO2D / HOJD2D	Conv2D-1 [6]	83.33
HOJO2D / HOJD2D	Conv2D-2 [6]	83.33
HOJO2D + HOJD2D	Conv1D-1 [6]	**91.67**
HOJO2D + HOJD2D	Conv1D-2 [6]	**91.67**
Limb Angle	GMA-based method [4]	**91.67**
Joint Position	Ours - Without Attention	**91.67**
Joint Position	Ours - Full System	**91.67**

We conducted an ablation study to evaluate if the attention module, which was responsible for providing the interpretation functionality, had any negative effect on classification accuracy. The results are shown at the bottom two rows of Table I. The attention module does not affect the overall classification accuracy, and it also provides us with an interface to interpret the outcome of the deep learning model in a human-understandable way. Notice that while traditional attention modules tend to improve classification accuracy, the target of ours is for interpretation. We tailor the way a channel is defined such that the results can be as human-understandable as possible. This limits the module's potential to fully optimize the attention for accuracy improvement.

We further studied the effectiveness of the proposed attention loss (see Table II). The present numbers are averaged by the total number of cross-validations. We find that the loss term effectively reduced the average per-joint attention values. The average number of joints with high attention values is also reduced. This allows the results to be more easily understood by humans, while maintaining the same prediction accuracy.

Method	Without L_{att}	With L_{att}
Avg. Per-joint Attention Value	0.492	**0.439**
Avg. # of Joints with Attention ≥ 0.5	11.039	**9.910**
Avg. # of Joints with Attention ≥ 0.6	10.856	**9.861**
Avg. # of Joints with Attention ≥ 0.7	10.750	**9.702**
Avg. # of Joints with Attention ≥ 0.8	10.566	**9.583**
Avg. # of Joints with Attention ≥ 0.9	10.074	**9.084**
Avg. # of Joints with Attention $= 1.0$	6.583	**6.250**
Accuracy (%)	**91.67**	**91.67**

Fig. 3. The visualization of joints attention in two cross-validations with a normal sample (upper row) and an abnormal sample (lower row) as validation.

B. Quantitative Evaluation

Fig. 2 presents a box plot to visualize the distributions of joint attention values across the twelve cross-validations. For each joint, the interquartile range (IQR) is represented by the blue bar, and the yellow line denotes the median value. The generally high median lines indicate most of the cross-validations tend to assign a higher attention value to the joint.

From the median lines, we can see that the joints 'Left Thigh', 'Right Hand', 'Thoracic Spine', 'Right Shoulder' and 'Left Upper Arm' have the highest attention values in almost all the cross-validations. Most of these joints are located in the upper body, including those on the arms and shoulders. This aligns with the results reported in [5] that the features extracted from the arms are more discriminative and resulted in higher accuracy in CP prediction on the MINI-RGBD dataset.
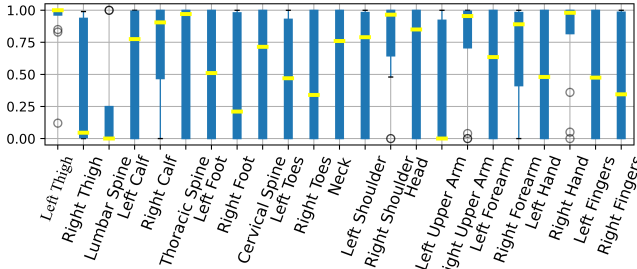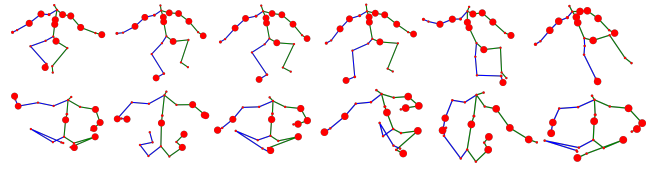

Fig. 2. The visualization of attention values of all joints across twelve validations.

We also visualize the joint attention values in two cross-validations, one with a normal sample (Fig. 3 upper) and one with an abnormal sample (Fig. 3 lower) as validation. The sizes of the red circles visualize the attention values of the corresponding joints. It can be seen that the example motions from different classes (i.e. *normal* and *abnormal*) resulted in different sets of attention values. This highlights the network has different focuses on the body joints to differentiate motions from the two classes. Compared with the previous works [4], [6], in which manually select some joints for analysis, our attention module is able to identify the more important joints automatically.

IV. CONCLUSION

We have presented an interpretable deep learning framework with channel attention for CP prediction. The attention module designed in this paper has shown how the network analyzes the movements by looking at the body joints that are more important to the final diagnostic results. We further demonstrated the effectiveness of the proposed attention loss for minimizing the number of attended joints such that the user can focus on a smaller set of joints for more in-depth analysis. Compared with the state-of-the-art methods, our proposed model achieves the same level of classification accuracy, even though we do not use hand-crafted features nor any manual interventions.

A future direction is to incorporate both spatial and temporal attention into our channel attention module, which will allow interpreting when an infant performs abnormal movements with which joints. This can facilitate a user to focus the analysis on the video segment with high attention. A key challenge is the large number of frames, which results in a large number of temporal attention values. Further research on how to effectively visualize such information is needed.

REFERENCES

[1] I. Novak, C. Morgan, L. Adde, J. Blackman, R. N. Boyd, J. Brunstrom-Hernandez, G. Cioni, D. Damiano, J. Darrah, A.-C. Eliasson *et al.*, "Early, accurate diagnosis and early intervention in cerebral palsy: advances in diagnosis and treatment," *JAMA pediatrics*, vol. 171, no. 9, pp. 897–907, 2017.

[2] P. Anderson, L. W. Doyle, V. I. C. S. Group, V. I. C. S. Group *et al.*, "Neurobehavioral outcomes of school-age children born extremely low birth weight or very preterm in the 1990s," *jama*, vol. 289, no. 24, pp. 3264–3272, 2003.

[3] A. Stahl, C. Schellewald, Ø. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerod, "An optical flow-based method to predict infantile cerebral palsy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 4, pp. 605–614, 2012.

[4] Q. Wu, G. Xu, F. Wei, L. Chen, and S. Zhang, "Rgb-d videos-based early prediction of infant cerebral palsy via general movements complexity," *IEEE Access*, vol. 9, pp. 42 314–42 324, 2021.

[5] K. D. McCay, E. S. L. Ho, C. Marcroft, and N. D. Embleton, "Establishing pose based features using histograms for the detection of abnormal infant movements," in *IEEE EMBC*, July 2019, pp. 5469–5472.

[6] K. D. McCay, E. S. L. Ho, H. P. H. Shum, G. Fehringer, C. Marcroft, and N. D. Embleton, "Abnormal infant movements classification with deep learning on pose-based features," *IEEE Access*, vol. 8, pp. 51 582–51 592, 2020.

[7] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 7132–7141.

[9] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. Sebastian Schroeder, "Computer vision for medical infant motion analysis: State of the art and rgb-d data set," in *Proceedings of ECCV Workshops*, 2018, pp. 0–0.

[10] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an infant body model from rgb-d data for accurate full body motion analysis," in *Proceedings of MICCAI*. Springer, 2018, pp. 792–800.