ATTENTION-GUIDED DEEP LEARNING FRAMEWORK FOR MOVEMENT QUALITY ASSESSMENT

Aditya Kanade\textsuperscript{a} Mansi Sharma\textsuperscript{a,b} Manivannan Muniyandi\textsuperscript{c}

\textsuperscript{a} Department of Electrical Engineering, Indian Institute of Technology Madras, India
\textsuperscript{b} Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, Amrita Vishwa Vidyapeetham, India
\textsuperscript{c} Department of Applied Mechanics, Indian Institute of Technology Madras, India

ABSTRACT

Physical rehabilitation programs frequently begin with a brief stay in the hospital and continue with home-based rehabilitation. Lack of feedback on exercise correctness is a significant issue in home-based rehabilitation. Deep learning-based movement quality assessment (MQA) can assist with home-based rehabilitation by providing the necessary quantitative feedback. However, systems developed for home-based settings must be fast and offer interpretability of the generated assessment. In this paper, we explore an attention-guided transformer-based architecture for MQA. A comparative analysis against the current state-of-the-art methods is undertaken to establish the validity of the proposed model. Further, we show that the proposed model offers significant performance improvement in training and inference time, which is pivotal for any real-time system. Finally, we show that analysis of the attention maps of the proposed model can give critical insights into the decision-making process of the deep-learning model, thus improving the overall interpretability of predicted assessment scores.

Index Terms— Movement Quality Assessment, deep learning, transformer, performance score generation, attention mechanism

1. INTRODUCTION

Physical rehabilitation is a part of the treatment plan for diseases such as stroke, Parkinson’s, and numerous musculoskeletal injuries. Patients undergoing physical rehabilitation sessions are assessed considering improvements in movement quality. These programs begin under the clinician’s supervision and slowly transition into home-based rehabilitation. Literature surveys indicate that more than 90% of rehabilitation sessions are carried out at home [1]. In a home-based rehabilitation setting, it is logistically and economically impossible to take aid of clinicians for movement quality assessment. To monitor their progress in the rehabilitation program, patients must either self-monitor or rely on family members. Current technological solutions for patients undergoing home-based rehabilitation [2, 3, 4] fail to inform patients of their incremental improvements in movement quality. Thus, developing systems capable of assessing movement quality in a home-based rehabilitation setting are critical.

Recent research is being directed toward developing deep-learning based movement quality assessment systems, which are trained using movement data captured on depth imaging devices. Movement data is temporal; hence, popular temporal deep-learning architectures such as multi-scale CNN-LSTM/Spatio-temporal GCNs have been adopted to predict movement quality scores [5, 6]. The neural network models are trained to predict movement quality scores based on raw skeletal movement data input. The performance of these models is measured in terms of the deviation from the ground-truth scores. In a home-based setting, accuracy is not the only criterion for evaluating the practicality of a neural network model. The explainability of the predicted score and the ability to scale well with more data are equally important when evaluating model efficacy.

The attention mechanism was initially developed for machine translation tasks allowing inference on the importance of source text sequence in translating to a text sequence in the target language [7]. It allows a better under-the-hood view of the mechanisms of a deep-learning model. The attention mechanism is increasingly being used to improve interpretability of neural network models [8, 9]. Existing neural network models developed for movement quality assessment are recurrent. A significant drawback of these models is the lack of parallel processing due to the inherent recurrent nature of execution [10]. Transformer is a neural network architecture capable of operating on temporal data without recurrence relation, allowing models to be trained on higher amounts of data in less time [10].

In this study, we examine the use of transformer architecture for predicting movement quality assessment scores. We also explore the use of the attention mechanism as a way to improve model explainability. The attention scores can provide insights into patient performance and indicate compensatory motion.
In this paper, we propose a novel transformer based architecture for movement quality score prediction. We show that the proposed model performs as well as existing state-of-the-art methods for the skeletal data captured on the Vicon system. While, the proposed model outperforms existing state-of-the-art methods on noisy data captured on the Kinect system. We use CNN-based feature extractors for vector representation of temporal window slices, allowing the transformer network to operate on the skeletal data. We show that embedding the attention mechanism into feature extraction can allow the transformer network to pay attention to specific body parts contributing to the performance of an exercise.

2. PROPOSED ARCHITECTURE

Transformers have been consistently successful in tasks such as image classification, automatic speech recognition, and natural language processing [11, 12, 13]. The transformer architecture is scalable, and a higher representation capacity can be obtained by increasing the number of layers in the architecture [14]. Additionally, parallel execution is also possible since the transformer architecture is not recurrent. This architectural difference significantly reduces training time and allows for training on higher amounts of data. Training on more extensive data can pave the way for more complicated and scalable modeling of human movements as more movement data becomes available.

Movement data is more difficult to train on a transformer directly. For example, data for a single repetition of an exercise performed for 10 secs, captured on a Vicon system at 90fps will result in a tensor \(X\) of size \(R^{900 \times 117}\). Passing this tensor through a single attention block of the transformer will cost \(900^2\) operations. Thus, passing the raw movement data input directly to the transformer does not seem reasonable. It leads to a quadratic increase in the number of parameters as the number of attention blocks increases.

Taking inspiration from the adaptation of the transformer architecture in the field of image classification and automatic speech recognition, we make use of an embedding layer between the raw movement data and the transformer for tokenizing the input data. The raw input \(X \in R^{T \times D}\) is split into \(N\) temporal window slices of size \(W\) each, such that \(T = W \times N\), given by the set \(S\), where each \(s_i\) represents the \(i^{th}\) temporal window.

\[
S = (s_1, s_2, s_3, \ldots, s_N), \quad s_i \in R^{W \times D} \tag{1}
\]

An embedding layer \(E\) projects each temporal window into a latent subspace \(Z\), where \(E: s_i \rightarrow z_i, i \in 1\ldots N\). This operation reduces the dimensionality to \(R^K\) such that each vector \(z_i\) is a \(K\) dimensional vector representation of the \(i^{th}\) temporal window \(s_i\). The transformer architecture does not have a way to capture positional information of the input tokens by design. We use the technique described in [10] to add positional information to the input. Each vector \(z_i\) is added with a positional embedding vector \(p_i \in R^K\). The set \(Z_p\) is the result of this operation.

\[
Z_p = (z_1 + p_1, z_2 + p_2, z_3 + p_3, \ldots, z_N + p_N), \quad z_i \in R^K \tag{2}
\]

![Fig. 1: Network Architecture](image)

The set \(Z_p\) is passed on to the lowest encoder layer of the transformer encoder. The output at the topmost layer of the transformer encoder is projected through a series of dense layers which finally predict the movement quality score.

The set \(Z_p\) is passed on to the lowest encoder layer of the transformer encoder. The output at the topmost layer of the transformer encoder is projected through a series of dense layers, which finally predict a score value for the exercise. The network uses the binary cross entropy loss between the predicted \(\hat{y}\) and the ground truth assessment score to jointly learn the parameters of the transformer encoder and the embedding layer. The architecture of this network is shown in Fig 1. Previous body of work has shown better performance exploiting the hierarchical structure of the movement data for action recognition [5, 15]. Hierarchical processing is achieved by designing a feature extractor to exploit the spatial characteristics of human movements by dedicating sub-networks for processing joint displacements of individual body parts. We propose a novel feature extractor termed the Attention guided Hierarchical Feature Extractor (HFE-A). This feature extractor consists of a multi-head attention block after hierarchical
The study of mixture weights for an input movement data sample can help assess compensatory movements and offer more interpretability of predicted assessment scores.

3. RESULTS AND DISCUSSIONS

3.1. Setup

The proposed model is trained on raw joint orientation data of the skeletal landmarks. The network is designed to predict the ground truth movement quality scores in a supervised regression setting. The model was implemented on an HP desktop computer with an i7 processor, 16GB RAM, and an NVIDIA-2080Ti GPU card. A separate model is trained for each of the ten exercises in the UI-PRMD dataset. We report the model performance in terms of the average absolute deviation (lower the better) between the ground truth movement quality scores and the network prediction; these values are averaged over five runs for generating the results. The network is trained on a 0.8/0.2 train/validation split. The model is trained using the Adam optimizer, setting the learning rate to 0.0005 [16]. Early stopping is used to avoid overfitting the training data. A patience value of 100 epochs is set to monitor the validation loss. The network is trained on binary cross-entropy loss between the predicted score and the ground truth scores. An extensive grid search was carried out for selecting hyperparameters and fine-tuning the transformer encoder and the embedding layer. Following is a list of the best configuration we found for the proposed model.

- Temporal Window (W) Size - 40
- Feature Vector (z_i) Size - 256
- Number of Heads in Transformer Encoder Block - 4
- Number of Transformer Encoder Blocks - 2

3.2. Model Performance

|          | Proposed Model | Liao et al. [5] | Deb et al. [6] |
|----------|----------------|-----------------|----------------|
| E1       | 0.0175         | 0.011           | 0.009          |
| E2       | 0.0321         | 0.028           | 0.006          |
| E3       | 0.0369         | 0.039           | 0.013          |
| E4       | 0.0342         | 0.012           | 0.006          |
| E5       | 0.0321         | 0.019           | 0.008          |
| E6       | 0.0331         | 0.018           | 0.006          |
| E7       | 0.0544         | 0.038           | 0.011          |
| E8       | 0.0379         | 0.023           | 0.016          |
| E9       | 0.0252         | 0.023           | 0.008          |
| E10      | 0.0613         | 0.042           | 0.031          |

We compare the performance of proposed transformer model to current state-of-the-art deep learning models for movement assessment. [6] is a graph convolutional architecture, while, [5] is a CNN-LSTM based architecture. Table 1 shows the comparative analysis of the proposed model against the current state-of-the-art methods on the UI-PRMD dataset [17]. The dataset contains movement data with corresponding ground truth labels. The data was collected on a vicon optical system. The results show that the proposed model performs at par with the comparative models. We also show results on KIMORE dataset, which was collected on the Kinect V2 system in Table 2.

|          | Proposed Model | Liao et al. [5] | Deb et al. [6] |
|----------|----------------|-----------------|----------------|
| E1-E5    | **0.1093**     | 0.1183          | 0.1458         |
| E1       | **0.0826**     | 0.0981          | 0.1265         |
| E2       | 0.1624         | **0.1503**      | 0.1732         |
| E3       | **0.0767**     | 0.1219          | 0.1364         |
| E4       | **0.1097**     | 0.1113          | 0.1585         |
| E5       | 0.1153         | **0.1098**      | 0.1345         |

The UI-PRMD [17] dataset is captured on a vicon optical system which provides an extremely clean capture of the human movements. This system however is impractical for use in a home-based rehabilitation setting. A better option is to use a Kinect system which is markerless movement data.
capturing system and can be deployed in a home-based setting easily. An issue with kinect system is the system introduced noise into the captured movement data. The KIMORE dataset [18] is captured using the Kinect V2 system. It is much more practical to compare deep-learning models on their performance on data captured on noisy devices such as kinect. From Table 1 and 2 it is clear that the proposed model performs almost equivalent on the UI-PRMD dataset, however, it outperforms existing state-of-the-art methods on the noisy KIMORE dataset. Thus, making the proposed model suitable for home-based rehabilitation setting. At this point, it is important to note that finding the right hyperparameters for the Transformer based architecture is more difficult than CNN-LSTM based architectures. However, the increased accuracy, scalability and speed of the Transformer-based network architecture compensates for this. Currently, the field of movement assessment lacks a large-scale dataset containing a large number of diverse examples of various exercises prescribed for physical rehabilitation. We believe that the proposed model can aid in scaling to much larger datasets as a result of the Transformer based architecture.

3.3. Model Analysis and Interpretability

We evaluate the training and inference time of current state-of-the-art methods with that of the proposed model; the results are shown in Table 3. Results show superior performance of the proposed transformer based architecture, this is due to the absence of any form of recurrence in the deep-learning model. Model training and inference times are critical when developing systems for a home-based settings due to non-availability of powerful computing devices and hence faster movement quality assessment systems can be developed using the proposed model.

Table 3: Computational Cost Analysis - A comparative analysis of training and inference times of the current state-of-the-art methods with the proposed model. We see that the proposed model performs much faster than the competitive models.

|                | Time per Batch (sec) | Inference Time (ms) |
|----------------|----------------------|---------------------|
| Proposed       | 0.5                  | 13                  |
| Liao et al. [5]| 3                    | 100                 |
| Deb et al. [6] | 9                    | 350                 |

Fig 3 shows attention map generated at the multi-head attention layer of HFE-A. The attention map was generated by training the proposed transformer network with HFE-A for the embedding layer on Standing Shoulder Abduction exercise. The exercise is primarily an upper-body exercise. The results from Fig 3 support our intuitive assumption that the multi-head attention block introduced in the design of HFE-A allows the proposed transformer network to attend specific body parts that contribute significantly to an exercise while predicting movement quality score. The attention maps of HFE-A can be used for insights into the scoring process, in terms of joint sub-networks that contributed significantly for a particular score. After the model is trained on performances by varied categories of subjects, comparing attention maps of patients and experts can reveal compensatory behaviour of movements.

4. CONCLUSION

This paper aims to address the problem of movement quality assessment using deep learning. We proposed a novel Transformer based architecture for movement quality assessment. We introduced four novel feature extractors for the embedding layer, aiding the transformer to operate on the continuous skeletal data. We compared the proposed network against current state-of-the-art methods on the UI-PRMD and KIMORE datasets. We saw that the proposed model outperformed on the noisy KIMORE dataset while performed equivalently on the clean UI-PRMD dataset. We demonstrated that the proposed model has a much lower training and inference time, making it more suitable for home-based movement quality assessment systems. Finally, improvements in the movement quality scores are aided by enhanced explainability of the score prediction process due to the addition of the attention layer.

5. ACKNOWLEDGEMENT

This work has been carried out at the Dept. of Electrical Engineering and Applied Mechanics, Indian Institute of Technology Madras in collaboration with the Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, Amrita Vishwa Vidyapeetham, India. We would also like to thank Prof. Alexandar Vakanski from University of Idaho for providing valuable comments and proofreading the manuscript.
References

[1] Ravi Komatireddy, “Quality and quantity of rehabilitation exercises delivered by a 3-d motion controlled camera: A pilot study,” *International Journal of Physical Medicine & Rehabilitation*, vol. 02, 08 2014.

[2] Sandeep Guguloth, Sivakumar Balasubramanian, and Sujatha Srinivasan, “A novel robotic device for shoulder rehabilitation,” 07 2018.

[3] Lynne Gauthier, Chelsea Kane, Alexandra Borstad, Nancy Strahl, Gitendra Uswatte, Edward Taub, David Morris, Alli Hall, Melissa Arakelian, and Victor Mark, “Video game rehabilitation for outpatient stroke (vigorous): Protocol for a multi-center comparative effectiveness trial of in-home gamified constraint-induced movement therapy for rehabilitation of chronic upper extremity hemiparesis,” *BMC Neurology*, vol. 17, 06 2017.

[4] Adyaasha Dash, Anand Yadav, Anand Chauhan, and Uptama Lahiri, “Kinect-assisted performance-sensitive upper limb exercise platform for post-stroke survivors,” *Frontiers in Neuroscience*, vol. 13, 2019.

[5] Yalin Liao, Aleksandar Vakanski, and Min Xian, “A deep learning framework for assessing physical rehabilitation exercises,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. PP, pp. 1–1, 01 2020.

[6] Swakshar Deb, Md Fokhrul Islam, Shafin Rahman, and Sejuti Rahman, “Graph convolutional networks for assessment of physical rehabilitation exercises,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 410–419, 2022.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” 2014.

[8] Hila Chefer, Shir Gur, and Lior Wolf, “Transformer interpretability beyond attention visualization,” 2020.

[9] Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer, “On identifiability in transformers,” 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *CoRR*, vol. abs/2010.11929, 2020.

[13] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *CoRR*, vol. abs/2006.11477, 2020.

[14] Alessandro Raganato and Jörg Tiedemann, “An analysis of encoder representations in transformer-based machine translation,” 01 2018, pp. 287–297.

[15] Yong Du, Wei Wang, and Liang Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” 06 2015, pp. 1110–1118.

[16] Diederik Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 12 2014.

[17] Aleksandar Vakanski, Hyung-pil Jun, David Paul, and Russell Baker, “A data set of human body movements for physical rehabilitation exercises,” *Data*, vol. 3, no. 1, 2018.

[18] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriù, L. Romeo, and Federica Verdini, “The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, pp. 1436–1448, 2019.