Transformer Encoder with Temporal Convolutions for Wrist Motion Classification

Zhenyuan Lu

Abstract

In this report, we adapt our previous work PAN model [9], a novel transformer-encoder deep learning framework for classifying pain intensities using physiological signals. PAN model integrates convolutional networks to capture multimodal features, a squeeze-and-excitation residual network emphasizing interdependencies among relevant features, and a transformer encoder block for optimal feature extraction and temporal dependency analysis. Supplemental information and source code are available at: `https://github.com/zhenyuanlu/WristMotion-Pison`.

1. Exploration

The workflow Fig. 2 is designed to handle multimodal data features such as timestamps, raw channel data, high-pass filtered channel data, quaternion, gyroscope, accelerometer data, body movement labels, and repetition numbers from Pison Wrist Motion Data (PWMD).

A crucial step in our preprocessing pipeline was the standardization of sample lengths across the dataset. This was achieved by either truncating or padding the data samples to achieve a uniform target length (we used truncating for the experiment). In our procedure, the length of each sample was truncated upon on the minimum sample length 967 in the first step. While we also noticed that there is a couple timestamps of inactivity at the begining of each motion Fig. 3, we also tested on the length of 900 Fig. 5.

Next, we have implemented a strategy to segregate data based on body movement labels and repetition numbers. Each unique combination of body movement label and repetition number was treated as a separate subject, e.g. Standing 2 with repetition number 1 as one subject. In total we have 15 subjects from the original dataset. This approach lends itself to a more practical and realistic representation of the data. For instance, in a real-world scenario, a subject may perform different wrist motions with the different body movements. Therefore, the data should be treated as a collection of unique subjects, each with a distinct body movement label and repetition number. Whereas, we found out that the wrist motion classes are based on the repetition numbers, e.g. 1, 2, 3, as three different wrist classes repeatedly performed during different body movement Fig. 4. During the training, the dataset was split into k-fold cross validation, where k is 5, each fold has three samples, each one has a unique body movement label with one unique repetition number.

2. Methodology

2.1. Convolutions

We initially employ a convolutional network to extract multimodal features. These extracted features can capture important information about the overall trend and variations in the features, providing valuable insight into the wrist motions Fig. 1.

2.2. SEResNet

Next, we use Squeeze-and-Excitation Residual Network (SEResNet) to learn the interdependencies among the extracted features to enhance the representation capability of the features [7]. SEResNet consist of two main components: a squeeze operation, which reduces the number of channels in the feature maps by taking their spatial average, and an excitation operation, which scales the channel-wise feature maps using a weighted sum of the squeezed features. This allows the network to selectively weight the importance of different channels and adaptively recalibrate the feature maps.

2.3. Transformer Encoder

Finally, to capture the temporal representations of the extracted features, we use a multi-head attention mechanism in conjunction with a temporal (causal) convolutional network.

Temporal Convolutional Network (TCN) TCN framework, inspired by the studies of Lea *et al*. [8] and Van den Oord *et al*. [10, 11], has been used effectively for processing and generating sequential data, *e.g*. audio or images. In

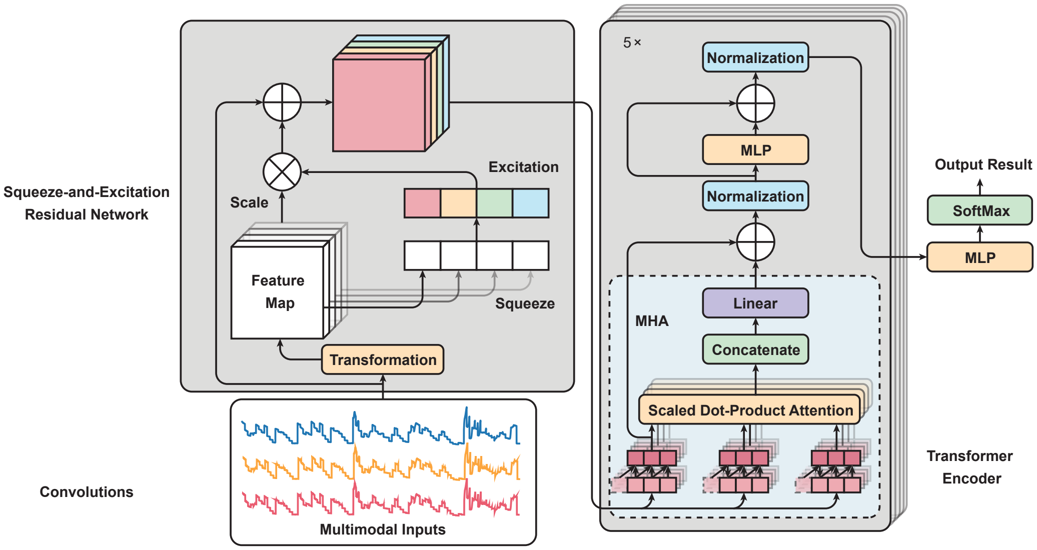

Figure 1. Outline framework of our proposed model. Left bottom: Convolutions. Left top: Squeeze-and-Excitation Residual Network (SEResNet). Right: Transformer Encoder.

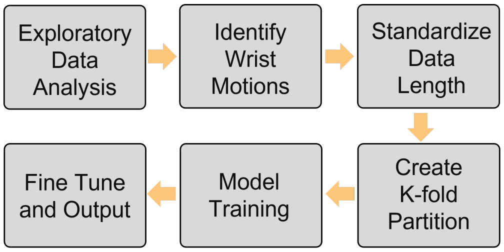

Figure 2. Workflow.

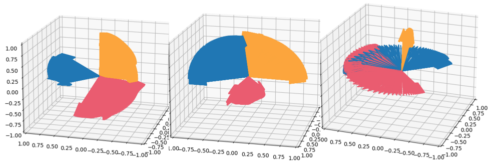

Figure 3. The transformation of quaternion data to rotation objects: X,Y, Z. Three wrist motion patterns present in the chart. Left: repetition number 1 with movement 0. Middle: repetition number 2 with movement 0. Right: repetition number 3 with movement 0.

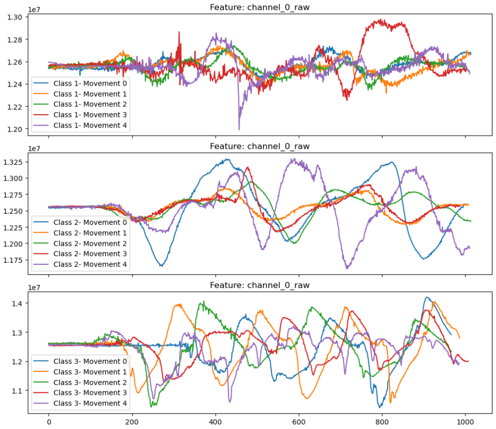

Figure 4. Different repetition numbers of raw data during different body movements. Here we only show the channel 0 raw signal.

contrast to a regular convolutional network, the TCN's output at time t depends only on the inputs before t. TCN only permits the convolutional layer to look back in time by masking future inputs.

Multi-Head Attention (MHA) MHA is the main part of the Transformer Encoder. It is a popular method for learning long-term relationships in sequences of features. We adapt this algorithm from Dosovitskiy *et al*. [5], Vaswani *et al*. [12], and Bahdanau *et al*. [2]. It has significant performance in different fields, *e.g*. GPT [3] and BERT [4] models in natural language process, and physiological signals classification for sleep Eldele *et al*. [6], Zhu *et al*. [13].

In particular, let the output feature maps from SEResNet, $\mathbf{X} = \{x_1, \ldots, x_N\} \in \mathbb{R}^{N \times L}$. Then we take three duplicates of $\mathbf{X}$ such that $\tilde{\mathbf{X}} = \varphi(\mathbf{X})$, where $\varphi(\cdot)$ is the function of TCN, and $\tilde{\mathbf{X}}$ is the output of TCN. Next we send the three outputs, $\tilde{\mathbf{X}}^{(Q)}, \tilde{\mathbf{X}}^{(K)}, \tilde{\mathbf{X}}^{(V)}$ to attention layers to calculate the weighted sum of the input, the attention scores $\mathbf{z}_i$:

$$\mathbf{z}_i = \sum_{j=1}^{L} \alpha_{ij} \varphi\left(\tilde{\mathbf{x}}_j^{(V)}\right), \tag{1}$$

the weight α_{ij} of each $\varphi(\tilde{\mathbf{x}}_j)$ is computed by:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{r=1}^{L} \exp(e_{ir})}, \tag{2}$$

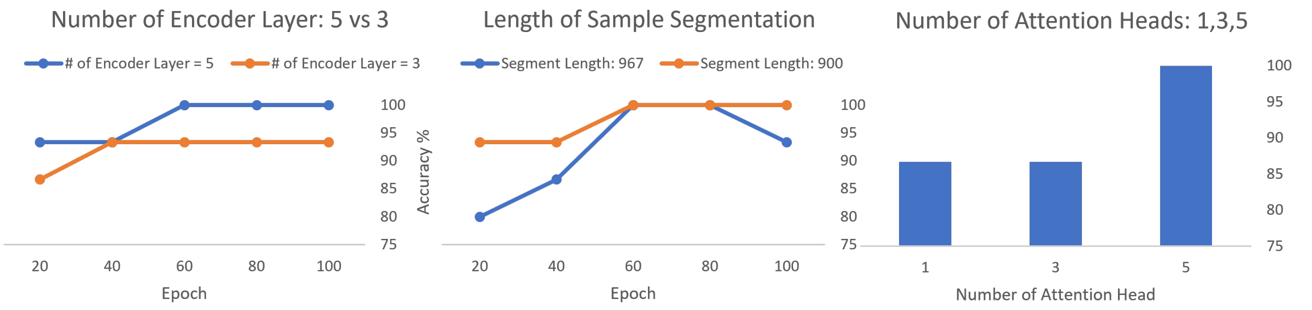

Figure 5. Left: the performance between the number of encoder layer - 5 and 3. Middle: the performance between different sample segmentation - minimum length 967 and arbitrary length 900. Right: the performance between different number of attention heads - 1,3, and 5.

here,

$$e_{ij} = \frac{1}{\sqrt{L}} \cdot \tilde{\mathbf{x}}_i^{(Q)} \cdot \tilde{\mathbf{x}}_j^{(K)\top}. \tag{3}$$

then the output of one attention layer is $\mathbf{z} = \{z_0, \ldots, z_L\} \in \mathbb{R}^{N \times L}$.

Next, MHA calculates all the attention scores $\mathbf{Z}^{(H)}$ from multiple attention layers parallelly, and then concatenate them into $\tilde{\mathbf{Z}}_{\text{MHA}} \in \mathbb{R}^{N \times HL}$, where H is the number of attention heads, and HL is the overall length of the concatenated attention scores.

We apply a linear transformation with learnable weight $W \in \mathbb{R}^{HL \times L}$ to make the input and output dimensions the same so that we can easily process the subsequent stages. The overall equation for MHA is as follows:

$$\tilde{\mathbf{Z}}_{\text{MHA}} = \text{Concat}(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(H)}) \cdot W \in \mathbb{R}^{N \times L}. \tag{4}$$

After concatenating these attention scores, we process them with the original $\tilde{\mathbf{X}}$ using an addition operation and layer normalization adopted from [1]. Finally, the results are obtained from another two fully connected networks, which are then followed by a Softmax function.

3. Experimental Results

In our experiment, we used the PWMD for modeling. This dataset consists of 17 columns provided in the dataset. 14 of them are features we used in our model for prediction of the repetition number (wrist motion), e.g. 1, 2, 3 among body movements. The results with different settings showed in Fig. 5. We have the best performance with the following settings: 5 encoder layers, 5 attention heads, and minimum length 967. The performance is 100% accuracy, 100% macro F1, and 100% Cohen's Kappa.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021. 2

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 1

[8] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to ac-

tion segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 47–54. Springer, 2016. 1

[9] Zhenyuan Lu, Burcu Ozek, and Sagar Kamarthi. Transformer encoder with multiscale deep learning for pain classification using physiological signals. *arXiv preprint arXiv:2303.06845*, 2023. 1

[10] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 1

[11] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 1

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[13] Tianqi Zhu, Wei Luo, and Feng Yu. Convolution-and attention-based neural network for automated sleep stage classification. *International Journal of Environmental Research and Public Health*, 17(11):4152, 2020. 2