ERTNet: an interpretable transformer-based framework for EEG emotion recognition

Ruixiang Liu1 , Yihu Chao1ǂ , Xuerui Ma1 , Xianzheng Sha1 , Limin Sun2 , Shuo Li3 and Shijie Chang1

[1]Department of Electrical Engineering

Introduction

The paper, ERTNet: an interpretable transformer-based framework for EEG emotion recognition1 pro- poses a new emotion recognition frame- work using a hybrid CNN and Transformer architecture.

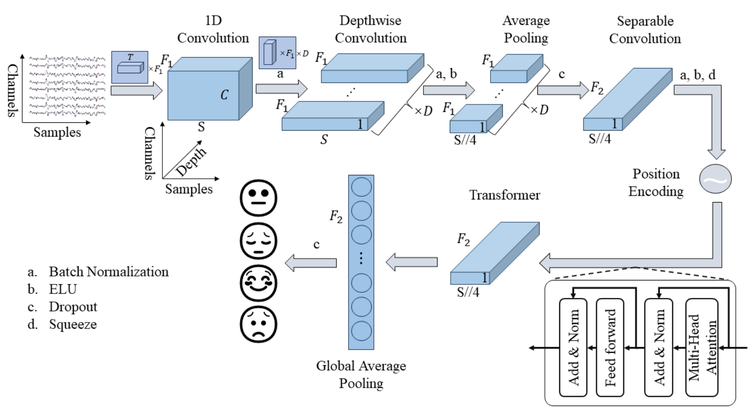


Figure 1. Model Architecture

Here we try to make some changes in the existing architecture provided in [1], and compare the results in Subject Dependent and Subject Indepen- dent testing.

Changes Proposed

The following changes in the model:

- **Activation Function:** Change the **ELU** activation with **GELU**
- **Dilation Convolution:** Adding Dilation to some Convolution layers
- **Data Preprocessing:** Segmenting data with overlaps
- **Positional Encoding:** Replace the sinusoidal encoding with learnable positional encoding

Methodology

1. **Activation Function:**GELU (Gaussian Error Linear Unit) is generally shown to be superior to ELU (Exponential Linear Unit) and RELU (Rectified Linear Unit) on a number of tasks like computer vision, natural language processing, and speech tasks which are also temporal data as the EEG data in our case.
2. **Dilation Convolution:**We incorporated dilation steps of (2,2) into the Separable 2D Convolution, which is the convolution applied before the positional encoding in the transformer block. This adjustment leverages the larger receptive field of dilated CNN networks,
3. **Data Preprocessing:** The paper directed us to segment the data into 4s (Sample Rate = 128Hz) non overlapping segments. But we added a function to create 4s segments with 25% overlap. The overlap reveals some new EEG contiguous segments that will help the models generalize better.
   It also increased the training data size by about 33%.
4. **Positional Encoding:** The paper is using fixed sinusoidal positional encoding, which are universal and is very generalized.
   Learnable positional encoding help us achieve better in our specific application, by learning a positional encoding that is best for our case. We made the Learnable Positional Encoding as a learnable matrix layer of size (F_{dim}, max_len), where F_{dim} is the number of output channels from the Separable Convolution Network (32 here) and max len is the maximum length of input.

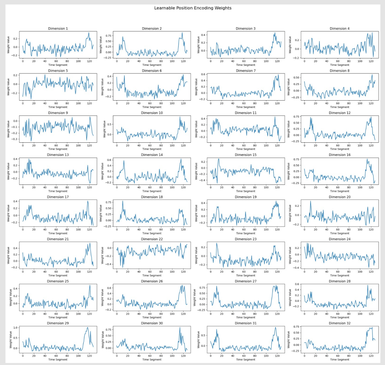
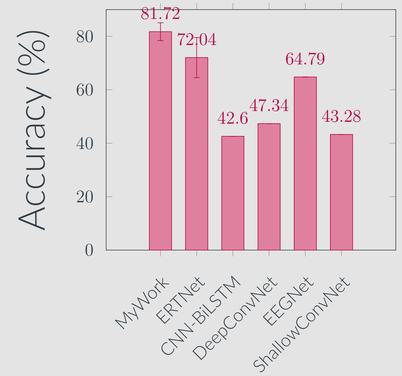


Figure 2. Learned Positonal encoding for all dimensions

Results

- **Subject Dependent Testing:** There is around 9% accuracy gain in the subject dependent training.



DEAP

Accuracy of Subject-Dependent DEAP

- **Subject Independent Training Results** The accuracy of the model on Subject Independent Training did not change much with the new model.

$$ERTNetAccuracy = 33.4 + -(7.4)\%$$
$$NewModelAccuracy = 32.12 + -(3.95)\%$$

Conclusion

Here we observed how the our model was superior to the original model in the Subject Independent Testing, and comparable in the Subject inde- pendent Testing. So we can conclude that the changes proposed helped improve the existing architecture given by [1].

References

[1] R Liu, Y Chao, X Ma, X Sha, L Sun, S Li, and S Chang. Ertnet: an interpretable transformer-based framework for eeg emotion recognition. *Frontiers in Neuroscience*, 18:1320645, 2024.