# A Transformer Model with Spatiotemporal Input Embedding for fNIRS data-Driven Neural Decoding

Hyunmin Lee, Taehun Kim, and Jinung An, *Member, IEEE*

***Abstract*— We propose a neural decoding model with input embedding reflecting fNIRS data's spatial and temporal traits. Our model embeds fNIRS channels into new channels based on the fNIRS optode layout. The embedded data is passed to the classifier through the 2D convolution layer and the transformer encoder. The open dataset (mental arithmetic, BNCI Horizon 2020) was trained to verify model performance. As a data preprocessing process, the 0.09Hz low-pass filter, data segmentation was performed, and then z-score standardization was applied. The average accuracy of leave-one-subject-out (LOSO) cross-validation (CV) was 90.19%, and the average accuracy of k-fold cross-validation was 81.78%. The proposed model resulted in the k-fold accuracy being lower than the LOSO accuracy. We checked the training loss graph and revealed an overfitting problem. One reason could be that the model parameter size is too large compared to the dataset. Another could be that the trained feature patterns are not very different because fNIRS data were measured only in the forebrain area. Therefore, the proposed model should be further applied and updated to other datasets measured in the whole brain region or different cortical areas.**

## I. Introduction

In the fNIRS-based neural decoding problem, fNIRS measures local cortico-activity, so it is necessary to configure the input space to encompass all spatiotemporal features. In the course of performing a task, activation of a particular neural region results in a change in cerebral hemodynamics, the amount of oxygen metabolism in that region. Through the fNIRS device, changes in cerebral hemodynamics can be confirmed by measuring the oxy-hemoglobin concentration (HbO) and deoxy-hemoglobin concentration (HbR) [1]. Therefore, the spatial location information of the fNIRS channel is essential information that can confirm the structural characteristics of brain activity in the area.

Attempts have been made to learn spatial information of brain signals in a deep learning model. One research [2] classifies 72 visual stimulation tasks using the topography of multi-channel EEG signals as input data for artificial intelligence models. This research uses CNN to learn spatial information but fails to consider the correlation between distant channels. Another research [3] uses multi-channel fNIRS signals as input in channel-wise and spatial-wise formats. In the spatial-wise format, consecutively numbered channels were embedded into one channel using CNN, and the channel was numbered regardless of the structural position of the brain. Therefore, existing multi-channel EEG/fNIRS -based deep learning methods do not consider connectivity between all channels, so there is a limit to expressing cerebral cortex networks.

We propose a neural decoding model with input embedding reflecting fNIRS data's spatial and temporal traits. Our model uses a transformer encoder and spatial embedding, considering the optode layout. The transformer-based model is better for evaluating the correlated features between non-adjacent and distant channels because it inherently possesses the multi-head self-attention layer. Input embeddings that reflect the layout form of the optode are expected to improve model performance because the channel of fNIRS consists of a pair of optodes called transmitters and receivers.

## II. Method and Data

### A. Open dataset

We decided to use the mental arithmetical dataset of BNCI Horizon 2020 [4] among the three datasets used by Wang [3]. Wang used the flooding technique to increase the accuracy of the model. Since Wang did not apply flooding to only one of the three datasets, using this dataset would be appropriate for comparison.

The dataset consists of fNIRS data obtained from eight subjects. In the experiment, a mental arithmetical task was performed for 12 seconds, and then a rest was performed for 28 seconds. The subject performed a task by looking at the formula appearing on the black screen, and a green bar was displayed on the screen before and after the formula appeared. A total of 174 mental arithmetical tasks and 174 rest sections were performed.

### B. Preprocessing

As a data preprocessing process [3], the frequency band of 0.09Hz or more was removed through the 4th-order low-pass Butterworth filter, data segmentation was performed, and then z-score standardization was applied.

The task and rest segments were set to 14 seconds [3], respectively. The task segment included 2 seconds after the 12-second task was completed considering the delay in which the effect of the task was reflected in the fNIRS signal. The rest period was set from 4 seconds to 18 seconds after completing the task.

### C. Model Structure

The data after preprocessing is the channel-level representation. Dimension of the data is (Batch, fNIRS

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2023R1A2C2006752).

H. Lee Author is with the School of Interdisciplinary Studies, DGIST, Republic of Korea (e-mail: chaos7ds@dgist.ac.kr).

T. Kim Author is with the School of Interdisciplinary Studies, DGIST, Republic of Korea (e-mail: t.kim@dgist.ac.kr).

J. An Author is with the School of Interdisciplinary Studies, DGIST, Republic of Korea (corresponding author to provide phone: +82-53-785-4610; fax: +82-53-785-4779; e-mail: robot@dgist.ac.kr).

Channel, Time, HbO/HbR). The data is used as an input to the embedding models for the spatial-level representation. Two types of embedding modules are used: optode-wise embedding modules and region-wise embedding modules.

The two modules embed some neighbor fNIRS channels into a new channel using the depth-wise convolution layers corresponding to 3x3 and 5x5 sized grids, respectively, based on fNIRS optodes.

The embedded data and channel-level representation are passed to the classifier through the transformer module. Each transformer module comprises a 2D convolution layer and the transformer encoder. Kernel size and stride of 2D convolution is (channels=1, time=30), (channels=1, time=4).

The classifier module combines the CLS-token of several transformer modules into a single vector in the concat layer. The vector is converted into a predicted value through a linear layer and a softmax layer.

## III. Result

As a result of the learning, the average accuracy of leave-one-subject-out (LOSO) cross-validation (CV) was 90.19%, and the average accuracy of k-fold cross-validation was 81.78%. A five-fold CV was performed five times, and the average accuracy of 25 results was used.

TABLE I. LOSO CV Result

| | Accuracy | Precision | Recall | F1-score | Kappa |
|---|---|---|---|---|---|
| Our model | 90.19 ± 8.65 | 89.48 ± 9.35 | 91.49 ± 8.05 | 90.40 | 0.80 |
| Wang (Reproduced) | 90.02 ± 8.61 | 92.88 ± 7.76 | 86.63 ± 12.52 | 89.34 | 0.80 |
| Wang (Paper) | 93.84 ± 5.99 | 96.30 ± 4.22 | 91.15 ± 10.09 | 93.42 | 0.88 |

TABLE II. K-fold CV Result

| | Accuracy | Precision | Recall | f1-score | Kappa |
|---|---|---|---|---|---|
| Our model | 81.78 ± 8.97 | 86.33 ± 10.21 | 77.97 ± 22.49 | 78.64 | 0.62 |
| Wang (Reproduced) | 94.19 ± 2.44 | 94.25 ± 3.87 | 94.20 ± 3.08 | 94.16 | 0.88 |
| Wang (Paper) | 95.29 ± 3.17 | | | | |

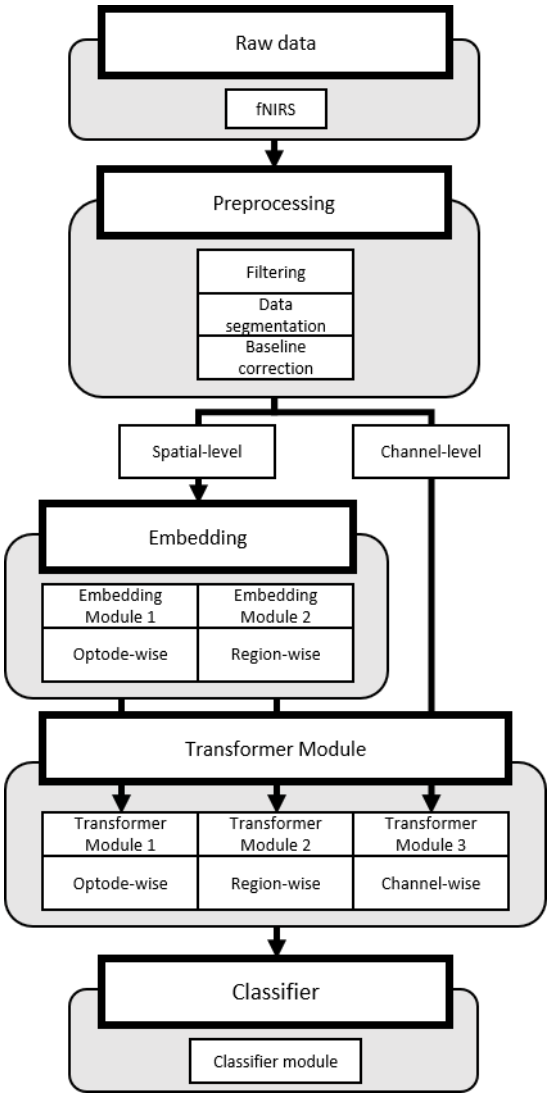
## IV. Discussion

Typically, the accuracy of the k-fold CV is supposed to be higher than that of the LOSO CV. Still, the proposed model resulted in the k-fold accuracy being lower than the LOSO accuracy.

We checked the training loss graph and revealed an overfitting problem. The reason could be that the model parameter size is too large compared to the complexity of the problem (i.e., binary classification between rest state and task state) to be solved in the open dataset. Another could be that the trained feature patterns are similar because fNIRS data were measured only in the forebrain area, a functionally and structurally identical cortical region.

Therefore, the proposed model should be further applied and updated to other datasets [5, 6] measured in the whole brain region or different cortical areas.

Figure 1. Overall Model Structure




## References


[1] Y. Kwak, W.-J. Song, and S.-E. Kim, "FGANet: Fnirs-guided attention network for hybrid EEG-FNIRS brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 329–339, 2022. doi:10.1109/tnsre.2022.3149899

[2] S. Bagchi and D. R. Bathula, "EEG-convtransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognition*, vol. 129, p. 108757, 2022. doi:10.1016/j.patcog.2022.108757

[3] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, "Transformer model for functional near-infrared spectroscopy classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2559–2569, 2022. doi:10.1109/jbhi.2022.3140531

[4] G. Bauernfeind, R. Scherer, G. Pfurtscheller, and C. Neuper, "Single-trial classification of antagonistic oxyhemoglobin responses during mental arithmetic," *Medical & Biological Engineering & Computing*, vol. 49, no. 9, pp. 979–984, 2011. doi:10.1007/s11517-011-0792-5

[5] J. Shin et al., "Open access dataset for EEG+NIRS single-trial classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1735–1745, 2017. doi:10.1109/tnsre.2016.2628057

[6] S. Bak, J. Park, J. Shin, and J. Jeong, "Open-access fNIRS dataset for classification of unilateral finger- and foot-tapping," *Electronics*, vol. 8, no. 12, p. 1486, 2019. doi:10.3390/electronics8121486