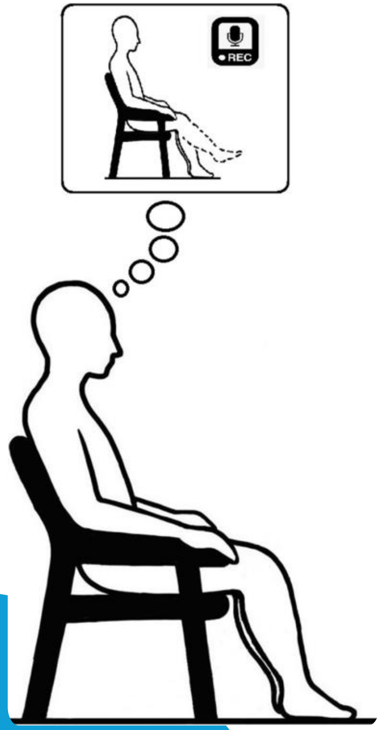
A Transformer-Based Approach Combining Deep Learning Network and Spatial-Temporal Information for Raw EEG Classification

IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022

Presenter: Nooshin Taheri

6/25/2025

Introduction



- **Motor Imagery (MI)** is a widely used paradigm in EEG-based Brain-Computer Interface (BCI) systems.
  It requires subjects to **imagine** movements (e.g., left or right hand), **without actual motion**.
- Accurate classification of MI-EEG signals is **crucial** for enabling BCIs to assist with tasks such as **rehabilitation** and **motor function recovery** in patients.
- However, MI-EEG data is challenging to work with due to:
  - High temporal resolution
  - Low spatial resolution
  - Low signal-to-noise ratio
  - High inter-subject variability
- EEG signals inherently contain **spatial dependencies** (across channels) and **temporal dependencies** (across time), both of which are essential for accurate classification.
- Some methods rely heavily on **Convolutional Neural Networks (CNNs)** to extract both **spatial and temporal features** (depending on the type of kernel used), but CNNs often struggle to capture **global dependencies**, limiting their effectiveness on complex EEG tasks.
- To better model **temporal dynamics**, some models combine CNNs with **Recurrent Neural Networks (RNNs)**.
- **Transformers** can model both **spatial and temporal relationships globally** through an attention mechanism, making them ideal for EEG analysis.

Contributions of This Study

propose an end-to-end Transformer framework that is capable of processing raw EEG data while retaining the spatiotemporal characteristics that are important for model visualization.

- **Novel Transformer-Based Models**
  Designed five architectures to classify raw MI-EEG data:
  - **s-Trans**: Spatial Transformer
  - **t-Trans**: Temporal Transformer
  - **s-CTrans**: Spatial CNN + Transformer
  - **t-CTrans**: Temporal CNN + Transformer
  - **f-CTrans**: Fusion of spatial & temporal CNN + Transformer
- **Integration of Positional Embedding (PE)**
  Explored 3 PE strategies (relative, channel-correlation, learned),
- **Interpretable Attention Visualization**
  Visualized attention weights across electrodes.

Dataset & Preprocessing

Dataset: PhysioNet EEG Motor Movement/Imagery

109 subjects, 64 electrodes, 160 Hz sampling rate

Tasks: left/right fist (L/R), both fists against both feet (F), and rest with eyes open(O)

Each trial lasted 8 seconds

Used **3s** and **6s** EEG segments for 2-(L/R), 3-(L/R/O), and 4-class (L/R/O/F) classification

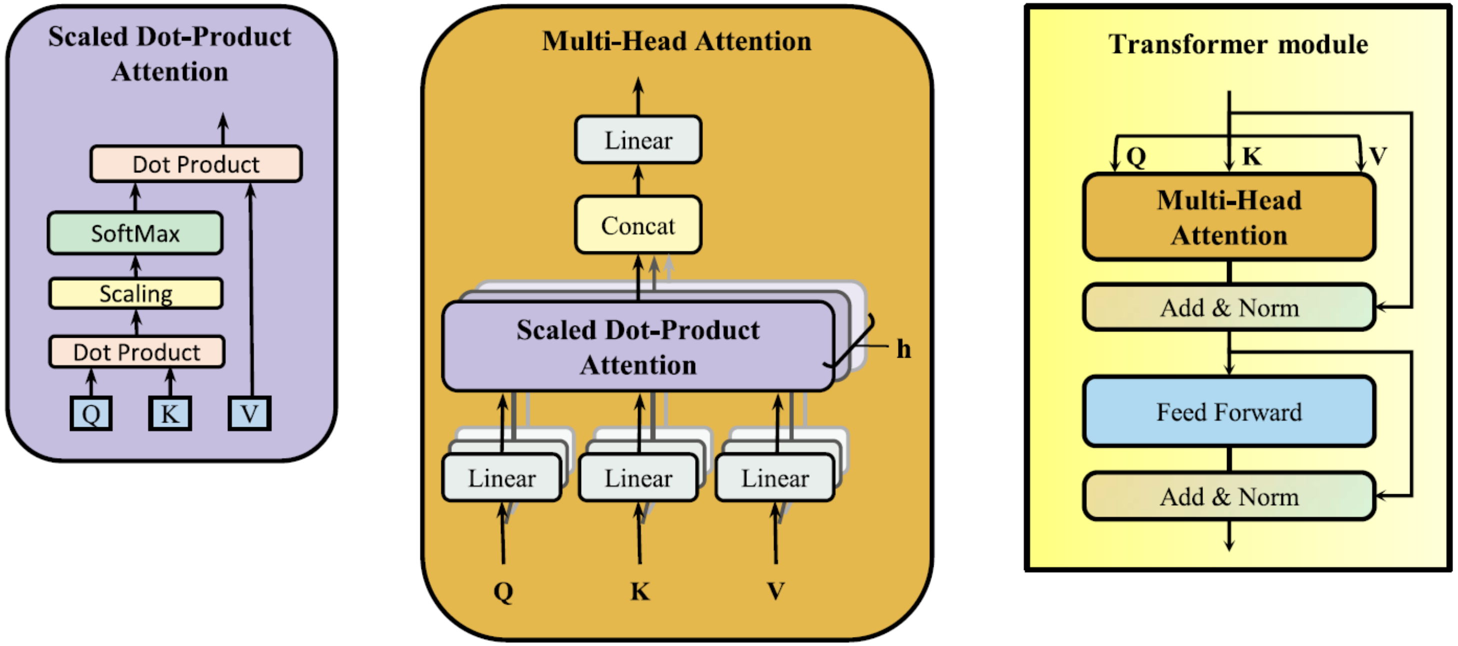
Preprocessing:

Z-score normalization applied to each EEG trial

Added small **random noise (α = 0.01)** to improve generalization and avoid overfitting

Data segmented from the motor imagery period

Structure of the transformer module



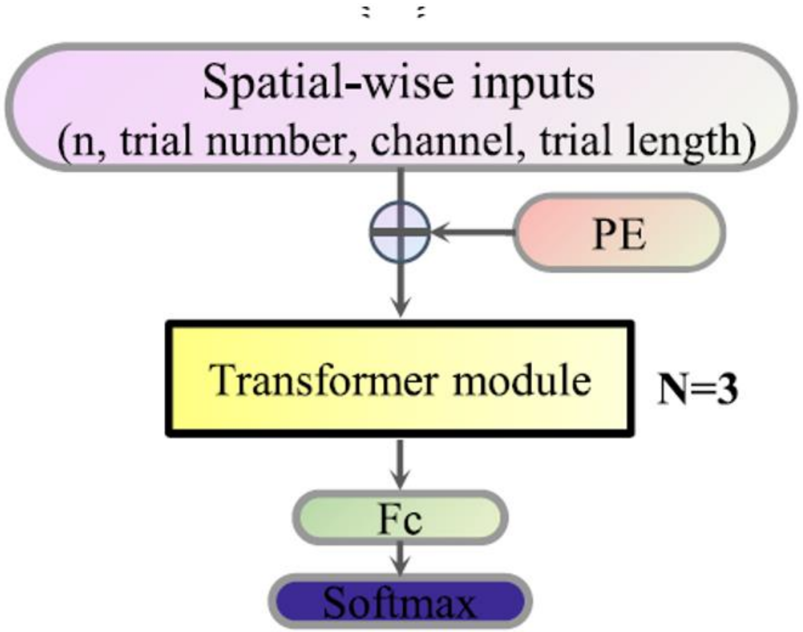
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

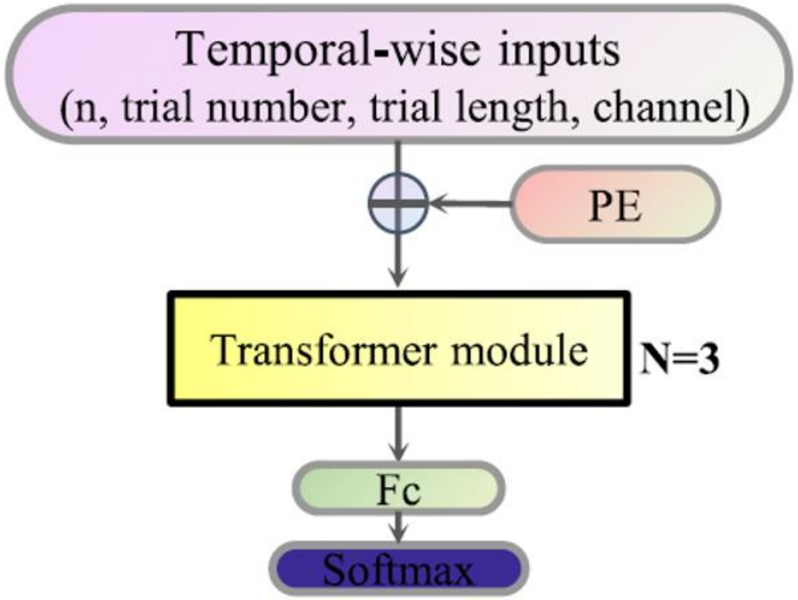
- Multi-head attention consisted of several "Scaled Dot-Product Attention" layers, allowing the model to jointly focus on information from different representation subspaces at different locations.

Model Architecture

- 8 attention heads were employed in this study, and solely embedded the encoder part of the Transformer into the EEG classification.
- Three types of PE were explored:
  - **Relative Positional Encoding** – uses sine & cosine functions to represent positions.
  - **Channel Correlation Encoding** – based on cosine distance between electrodes.
  - **Learned Positional Encoding** – trainable embedding matrix updated during training.
- The number of Transformer layers was varied from 1 to 6, and **using 3 layers achieved the best classification performance**.

Spatial and Temporal Transformer Models



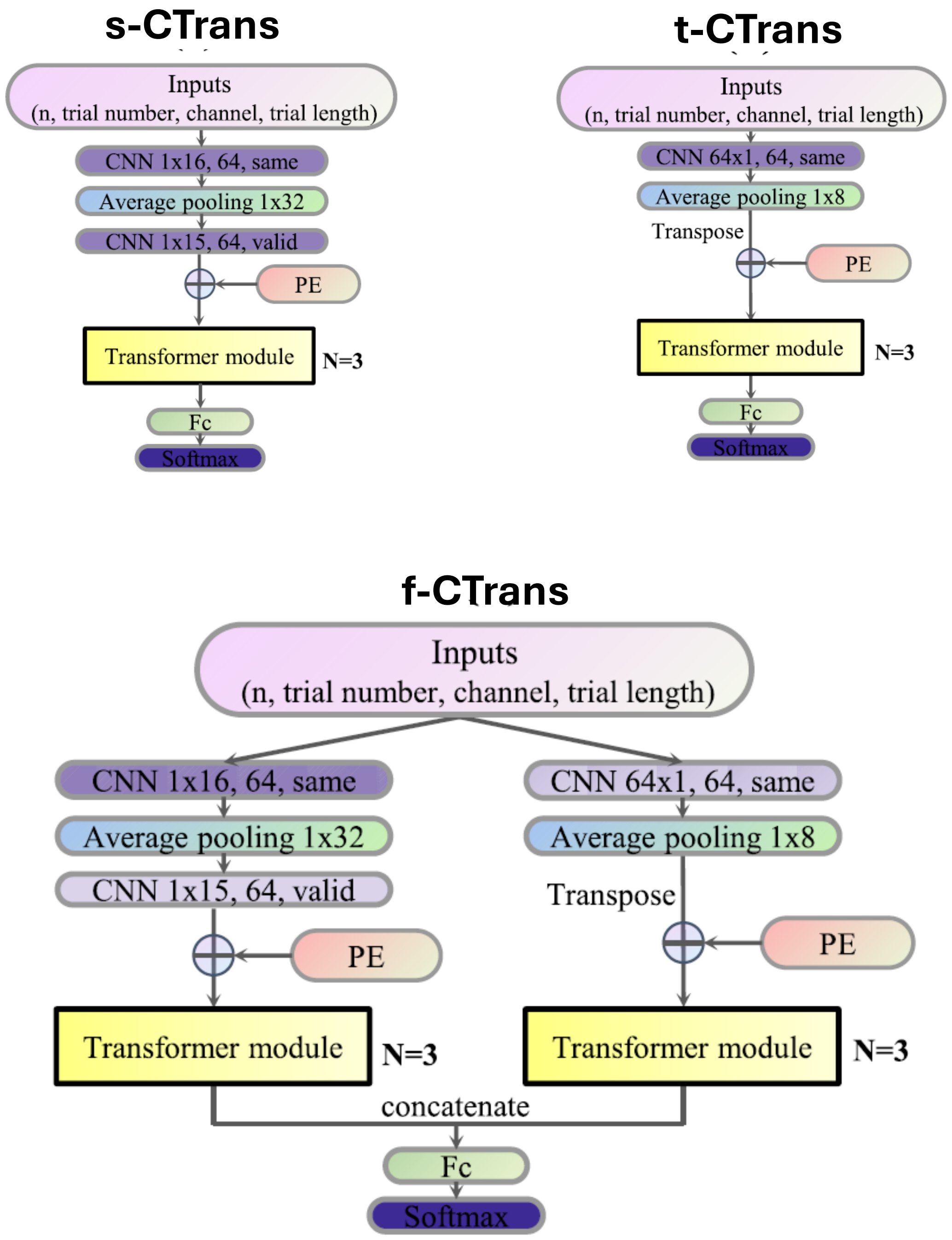


Spatial Transformer (s-Trans)
EEG data along the time axis from each channel were regarded as features, and the Transformer module calculated the correlations between different channels.

Temporal Transformer (t-Trans)
EEG data along the channel axis at the same time point were regarded as features, and the model calculated the correlations between different time points.

CNN + Transformer Models

- Combined CNN's **local feature extraction** with Transformer's **global attention** to enhance EEG classification.
- Three hybrid models were proposed:
  - **s-CTrans**: CNN for **temporal features**, Transformer for **spatial attention**
  - **t-CTrans**: CNN for **spatial features**, Transformer for **temporal attention**
  - **f-CTrans: Parallel fusion** of spatial and temporal branches
- CNN layers reduce dimensionality and extract robust features before passing them to the Transformer.



Classification Results

- Using **3-second EEG data**, the best accuracies achieved were:
  - **83.31%** (2-class), **74.44%** (3-class), **64.22%** (4-class)
    → Outperformed all baseline models.
- Using **6-second data**, performance improved further:
  - **87.80%**, **78.98%**, and **68.54%** for 2-, 3-, and 4-class tasks respectively.
- **f-CTrans** performed best on 3s data (3/4-class),
  while **t-CTrans** was best on 6s data.
- The EEG data with a longer period produced higher classification accuracy.

ACCURACY (%) COMPARISON BETWEEN OUR MODELS AND OTHER SOTA MODELS IN THE PHYSIONET DATASET FOR CROSS-INDIVIDUAL CLASSIFICATION

Models	**3s**			**>= 4s**		
	L/R	L/R/O	L/R/O/F	L/R	L/R/O	L/R/O/F
Our s-Trans	81.11	70.25	59.35	87.46	75.41	64.04
Our t-Trans	80.77	70.31	58.21	86.10	75.24	62.15
Our s-CTrans	**83.31**	72.88	63.25	87.80	77.09	68.10
Our t-CTrans	82.56	72.87	63.48	87.80	**78.98**	**68.54**
Our f-CTrans	82.95	**74.44**	**64.22**	87.26	78.44	67.96
CNN (2018) [5]	80.38	69.82	58.58	**87.98**	76.61	65.73
EEGNet (2020) [13]	82.43	72.33	63.16	--	--	--
EEGNet Fusion (2020) [60]	--	--	--	83.80	--	--
DG-CRAM (2020) [61]	74.71	--	--	--	--	--
MAML-CNN (2021) [62]	80.60	--	--	--	--	--
BENDR (2021) [45]	--	--	--	86.70	--	--

Effect of Positional Embedding (PE)

CLASSIFICATION RESULTS OF SPATIAL-TRANSFORMER MODEL USING DIFFERENT POSITIONAL EMBEDDING METHODS

Methods	480 (3s)			960 (6s)		
	L/R	L/R/O	L/R/O/F	L/R	L/R/O	L/R/O/F
relative PE	81.11%	**70.25%**	59.35%	**87.46%**	75.41%	64.04%
Channel correlation PE	**81.49%**	69.48%	**59.47%**	87.14%	75.26%	64.05%
learned PE	81.47%	70.02%	59.08%	87.07%	**75.52%**	**64.06%**
No PE	81.13%	68.25%	57.23%	86.83%	73.15%	61.43%

- Three PE methods (relative, channel-correlation, learned) were tested using the **s-Trans model**.
- All PE methods **outperformed the no-PE baseline** for both **3s and 6s EEG data**.
- **Learned PE** showed slightly better accuracy but required **more training parameters**.
- **Adding positional embeddings improves classification accuracy**, even if modestly.

Conclusion & Future Directions

- Developed five **Transformer-based models** for motor imagery EEG classification.
- Achieved **state-of-the-art accuracy** across **2-, 3-, and 4-class** tasks using raw EEG.
- **Fusion model (f-CTrans)** performed best on short input (3s), showing robustness and efficiency.
- Models are suitable for **real-time BCI applications** and can be extended to other EEG tasks like **disease diagnosis** or **neurorehabilitation**.
- **Future Optimizations**:
  - Use **multi-scale attention** to better capture EEG features with varying time-scales.
  - **Prune uninformative attention heads** to reduce computational cost and enhance model robustness.