2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Gated Transformer for Decoding Human Brain EEG Signals

Yunzhe Tao[1], Tao Sun[1], Aashiq Muhamed[2], Sahika Genc[†1], Dylan Jackson[2],
Ali Arsanjani[3], Suri Yaddanapudi[2], Liang Li[2] and Prachi Kumar[2]

***Abstract*— In this work, we propose to use a deep learning framework for decoding the electroencephalogram (EEG) signals of human brain activities. More specifically, we learn an end-to-end model that recognizes natural images or motor imagery by the EEG data that is collected from the corresponding human neural activities. In order to capture the temporal information encoded in the long EEG sequences, we first employ an enhanced version of Transformer, i.e., gated Transformer, on EEG signals to learn the feature representation along a sequence of embeddings. Then a fully-connected Softmax layer is used to predict the classification results of the decoded representations. To demonstrate the effectiveness of the gated Transformer approach, we conduct experiments on the image classification task for a human brain-visual dataset and the classification task for a motor imagery dataset. The experimental results show that our method achieves new state-of-the-art performance compared to multiple existing methods that are widely used for EEG classification.**

I. INTRODUCTION

Recently, the research on brain-computer interfaces (BCIs) has been an area of high public awareness. The main goal of BCI is to restore or provide assistance on some useful functions for the disabled or injured people, such as the spelling system or the control of cursors, wheelchairs and other devices. The electroencephalogram (EEG) machine is widely used in building BCIs. It records brain signals that encode neural intention with high temporal resolution. However, the EEG machine alone is not a BCI. The signal processing techniques are required for EEG-based BCIs, which focus on the feature extraction, selection and classification. The recent emerging development of machine learning (ML), or deep learning, has suggested us that using ML techniques to decode human EEG signals is a good option [15]. In particular, with the development of the computing capability, we are able to learn a high-performing ML system from large datasets for building the BCI.

Many traditional ML approaches have been applied in EEG classification tasks, which include k-nearest neighbors [19], logistic regression [27], linear discriminant analysis (LDA) [25], support-vector machines (SVMs) [9], and the discrete wavelet transform [20]. More recently, the availability of large EEG datasets and advances in ML have both led to the deployment of deep learning architectures, which enable large-scale ML systems to achieve higher accuracy in EEG classification tasks. The deep learning approaches in literature vary from the auto-encoder [16], convolutional neural network (CNN) [26], recurrent neural network (RNN) [11], long short-term memory (LSTM) [2], to more advanced architectures, such as SyncNet [17], EEGNet [14], EEG-ChannelNet [21], and graph convolutional network (GCN) [18]. However, there are still a couple of limitations in the aforementioned methods that prevent us from building high-performing BCIs. First, due to high temporal resolution, EEG signals are usually extremely long sequences. The sequence models, e.g., RNNs and LSTMs, process the EEG signals sequentially, namely, they train the data at each time step one by one, which largely increases the training time for convergence. In addition, although some deep learning frameworks can capture temporal dependencies, such as RNN-based models for long-term dependencies and CNN-based models for neighboring interactions, they can only achieve limited performance when the sequences are extremely long (see, e.g., [35]).

In this work, we propose to use the Transformer-like architecture for EEG classification. Transformer models are attention-based models, which process the entire signal as a whole. Theoretically, the attention mechanism naturally enables the model to capture long term dependencies with no limitation of the sequence length. In the experiments, we investigate two variants of vanilla Transformer, i.e., Pre-LN Transformer and Post-LN Transformer that differ in the placement of layer normalization [3]. Different from the vanilla Transformer architectures, we employ the gating mechanism [22] instead of the residual connection [28]. We show empirically that the gating mechanism can further improve the model performance. We assess the gated Transformer method by conducting classification experiments on two datasets, i.e., the brain-visual dataset and the motor imagery dataset. We also compare its performance against multiple cutting-edge models in EEG data processing. The results demonstrate that in the brain-visual EEG data classification, the gated Transformer achieves an accuracy of 61.11%, which greatly outperforms current state-of-the-art accuracy (52.20%), and in the motor imagery EEG data classification, the performance of gated Transformer (55.40%) is comparable to the best accuracy among baseline models (55.46%), but is significantly better than the performance of most of the state-of-the-art models.

[†]Corresponding author
[1]Yunzhe Tao, Tao Sun and Sahika Genc are with AI Labs, Amazon Web Services, Seattle, WA 98121, USA {yunzhet,suntao,sahika}@amazon.com
[2]Aashiq Muhamed, Dylan Jackson, Suri Yaddanapudi, Liang Li and Prachi Kumar are with AI Devices, Amazon Web Services, East Palo Alto, CA 94303, USA {muhaaash,jacdylan,yaddas,mzliang,kumprach}@amazon.com
[3]Ali Arsanjani is with AI/ML Specialist Solution Architecture Group, Amazon Web Services, San Diego, CA 92121, USA arsanjan@amazon.com

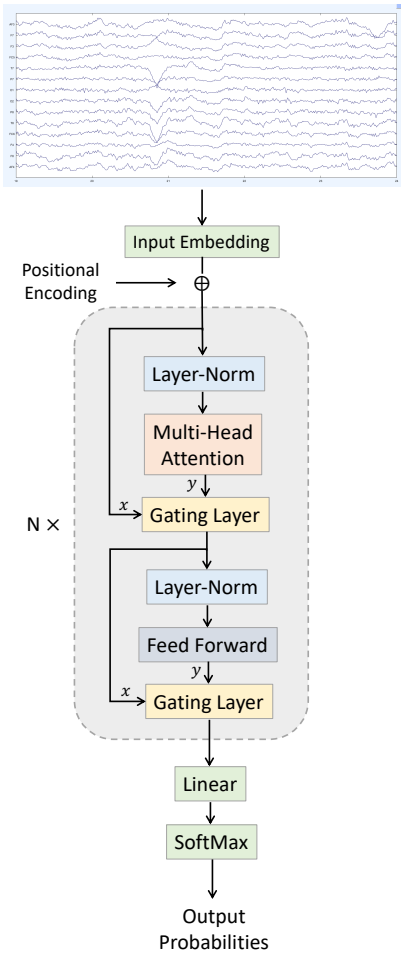


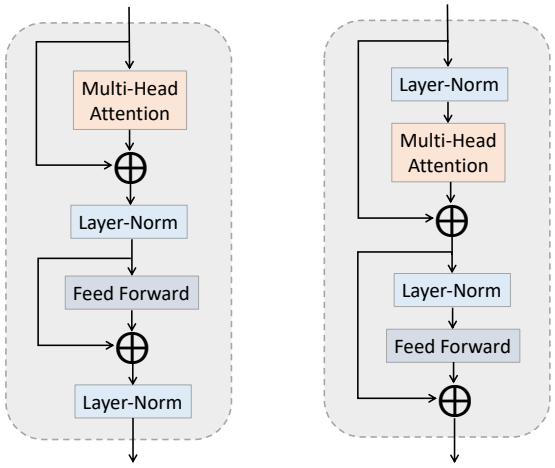
Fig. 1: The graphical illustration of the gated Transformer architecture.

The remaining of the paper is organized as follows. In the next section, we introduce two forms of the vanilla Transformer as well as the architecture of the Gated Transformer. Then in Section III we present the experimental results as well as some discussion on them. Lastly, Section IV concludes this paper with some concluding remarks and future directions.

II. GATED TRANSFORMER FOR CLASSIFICATION

In this section, we introduce the gated Transformer architecture in detail. The model is adapted from [22], where a gating mechanism is used to stabilize the Transformers for reinforcement learning. In this paper, we aim to show that the gated Transformer model is also helpful in the EEG classification tasks. Fig. 1 shows the details of the overall architecture for classification.

The input raw EEG data first goes through an input embedding. Because the Transformer model processes the sequential data as a whole, to give the model a sense of the order of the value at each time step, it adds positional encoding vectors upon the input embedding. The values of encoding vectors usually follow a specific pattern. Here we use sine and cosine functions following [28].



(a) Post-LN Transformer (b) Pre-LN Transformer

Fig. 2: The encoder architectures of two variants of vanilla Transformers.

Then the model tries to encode the input vectors by a set of encoder blocks (in Fig. 1 we set the number to be N). The encoder block consists of two major sub-layers, namely, multi-head attention layer and feed forward layer. The EEG data at each time step first passes through a self-attention process. By self-attention, the model can encode any non-local correlation of EEG data along a long sequence. In the implementation, we usually use multi-head attention layer for improving the performance of self-attention layers. Then, the feature vectors pass through a feed-forward neural network for further embedding.

The originally designed Post-LN Transformer from [28] (see Fig. 2(a)) places the layer normalization [3] between the residual blocks, which has been shown that the expected gradients of the parameters near the output layer are large [30]. Therefore, a learning rate warm-up stage is required for avoiding the problem. On the other hand, some works [4], [5], [29] put the layer normalization inside the residual blocks (recently proposed as Pre-LN Transformer, see Fig. 2(b)). Then the gradients are well-behaved at initialization. We adopt the Pre-LN architecture for the gated Transformer in this paper. Hence, before each sub-layer (self-attention or feed-forward), the input is normalized by the Layer-Norm operation, which can be viewed as a regularization approach.

The main difference between the gated Transformer and Pre-LN Transformer is that the vectors pass through a gating layer after each sub-layer. [22] has listed several variants of gating layers. We investigate all of them in the experiments. For simplicity, in this section, we only provide the formulation of the best gating layer, which is extended from the gated recurrent unit (GRU) approach [7], and can be empirically shown to be more stabilized than the residual connection in the vanilla Transformer. Assume the vectors x and y are given as indicated in Fig. 1. The superscript l is used to represent the l-th gating layer in the model. Then the gating

layer computes $g^{(l)}(x, y)$ as:

$$
\begin{aligned}
r &= \sigma(W_r^{(l)} y + U_r^{(l)} x)\,, \\
z &= \sigma(W_z^{(l)} y + U_z^{(l)} x + b_g^{(l)})\,, \\
\hat{h} &= \tanh(W_g^{(l)} y + U_g^{(l)} (r \odot x))\,, \\
g^{(l)}(x, y) &= (1 - z) \odot x + z \odot \hat{h}\,.
\end{aligned}
$$

Here, $\sigma(\cdot)$ and $\tanh(\cdot)$ represent the sigmoid and tanh functions, respectively. W's and U's are the parameter matrices, and b_g's are the bias terms to be learned. The $\odot$ denotes Hadamard product (element-wise product).

Finally, in order to output the probabilities for each class, the output of encoder is fed into a linear layer with a Softmax function. The output vector has a dimension equal to the number of classes and sums up to 1, which represents the probability of predicting the EEG signal as each class.

III. EXPERIMENTS

In order to assess the gated Transformer and compare with other state-of-the-art methods, we conduct experiments on two classification tasks, i.e., the brain-visual dataset [21] and the motor imagery dataset [10]. We first introduce the detail and pre-processing of the datasets, the models we use for comparison in the experiments, and then present the extensive results for the EEG data classification.

A. Datasets

a) Brain-visual dataset: We use the same dataset that was introduced in [21], which consists of 40 classes and was collected from 6 subjects (1 female and 5 males). Each class has 50 different images, which were taken from the ImageNet dataset [8]. During the experiment, 2,000 images were shown in bursts for 0.5 second each. The bursts for each class last for 25 seconds, followed by a 10-second pause where a black image was shown. The EEG signals were recorded using a 128-channel cap with active, low-impedance electrodes (actiCAP 128Ch). Sampling frequency and data resolution were set, respectively, to 1000 Hz and 16 bits. After some data cleaning, the brain-visual dataset contains 11,964 EEG segments in total, each segment has a length of around 500 time steps. In experiments, the first 20 samples (20 ms) were discarded, and each segment was cut to 440 samples (20 ms to 460 ms). We also report the performance of different EEG time intervals in the experiments.

For data pre-processing, a second-order band-pass Butterworth filter was first set up. In the experimental results presented later, we investigated several low and high cut-off frequencies for the filtering. The filtered signal is then normalized to zero mean and unitary standard deviation. In addition, a notch filter was also used around the power line frequency at 50 Hz. In order to replicate and compare the performance in literature, we follow [21] and use the same training, validation and test splits of the brain-visual dataset, which consists of 1600 (80%), 200 (10%), 200 (10%) images with associated EEG signals, respectively.

b) Motor imagery dataset: We then use the PhysioNet dataset [10] of the subject-wise scenario to evaluate the performance of models. The dataset has been widely used in recent works [31], [33], [34], [35], which was collected using BCI2000 instrumentation with 64 electrode channels and 160 Hz sampling rate. It consists of EEG recordings of movement intention with 109 subjects, each performed experimental runs for baselines (open/closed eyes), motor execution (open and close left/right fist, and open and close either both fists or both feet) and motor imagination (the same as motor execution, but pure imagination).

Following [35], we pre-processed the dataset by a band-pass filter with [0.5-55] Hz cut-off, and only considered the closed-eye baseline and the four motor imaginations. In order to balance the data for the five classes, for the closed-eye baseline, we randomly picked number of chunks from each experimental run which has the same number of trials as in other classes. As a result, we ended with 11,354 data points of size 64 (channels) by 656 (time steps). Different from [35] and the brain-visual experiments, we considered *subject-independent* experiments, where we split the data by subjects to train:validation:test=4:1:1.

B. Models for Comparison

In this paper, we implement and compare the performance of the gated Transformer with multiple cutting-edge models for EEG classification. All models are implemented with either TensorFlow [1] (Bi-LSTM, CRAM, and Mesh-Cascade) or PyTorch [23] (SyncNet, EEGNet, EEG-ChannelNet, and Transformers) framework, and trained from scratch in a fully-supervised manner. The Adam optimization approach [12] is used to minimize the cross-entropy loss function, with tuned learning rates for each model. In addition, the details of comparison models are given as follows.

a) Bidirectional LSTM (Bi-LSTM): RNNs are widely used for processing sequential data, where all inputs and outputs are not explicitly dependent. However, in order to predict the next value of the sequence, RNN learns a representation of the current and historical data. Long Short-Term Memory (LSTM) networks are a variant of RNN, which is designed for learning long-term dependencies by introducing several gating units. The bi-directional LSTM (Bi-LSTM) model allows us to learn the temporal representation of EEG data from two directions, namely, a forward path from past to future and a backward path from future to past. In the experiments, we also add an attention layer for Bi-LSTM to better select the useful time steps.

b) SyncNet [17]: Unlike Bi-LSTM, SyncNet is built upon Convolutional Neural Networks (CNNs), which learn the interaction among EEG data from different time steps by convolution operations. SyncNet performs structured (parameterized) 1D convolution for jointly modeling the power, frequency and phase relationships among EEG channels. In addition, we also evaluate the baseline CNN model in the motor imagery task for the completeness.

c) EEGNet [14]: EEGNet is a fully convolutional network, which performs a couple of 2D convolution layers

along different dimension of the EEG data. It starts with a temporal convolution to learn frequency filters, then uses a depthwise convolution to learn frequency-specific spatial filters. Then through another combination of depthwise convolutions, EEGNet learns a temporal summary for each feature map individually and mixes them for the prediction.

d) EEG-ChannelNet [21]: EEG-ChannelNet is another CNN-based model developed most recently. The EEG signal is first processed by a set of concatenated 1D convolutions along temporal direction (temporal block), followed by a set of concatenated 1D convolutions across channels (spatial block). Then the resulting features are processed by residual layers, which leads to the representation for classification.

e) Convolutional Recurrent Attention Model (CRAM) [32]: CRAM combines the use of convolutional network and recurrent network in one model. It first splits the EEG signal into temporal slices and leverages a CNN to encode each temporal slice for extracting its spatio-temporal features. Then an attention-based recurrent network is further used to explore the temporal dynamics among different slices. In addition, the attention mechanism concentrates the temporal dynamics on the most relative slices to the classification.

f) Mesh-Cascade [35]: Similar to CRAM, Mesh-Cascade consists of spatial feature extraction by a set of CNNs and temporal feature extraction by stacked LSTM layers. However, Mesh-cascade conducts a special processing for the inputs, where it maps 1D input vectors along channels to 2D matrices (meshes). The 2D meshes contain information of the electrodes placement, hence the convolution operations later can better capture the local spatial interactions. Due to lack of the 2D mapping with the 128-channel dataset, we only evaluate Mesh-Cascade in the motor imagery task.

g) Vanilla and Gated Transformers: In experiments, we train the classification tasks on both vanilla Transformers and gated Transformers. For vanilla Transformers, we assess both Pre-LN and Post-LN Transformers. For gated Transformers, in addition to the GRU gates mentioned in last section, we also test a few of other gating mechanisms, including InputGate, OutputGate, HighwayGate and SigTanhGate. The details of each gate formulation can be found in [22].

C. Experimental Results

a) Brain-visual dataset: In the brain-visual data classification task, we first assess all state-of-the-art models using whole dataset and complete EEG time course, i.e., 20-460 ms. Moreover, we process the data using a band-pass filter with different cut-off frequencies, and test the classification performance on these frequency ranges. Similar to [21], the bands we have used include high gamma ([55-95] Hz), beta to mid gamma ([14-70] Hz), and all frequency ([5-95] Hz). The results are presented in Table I. We can first observe that the best accuracy or each frequency band is always achieved by the gated Transformer, two from GRUGate and the other from SigTanhGate. In particular, for the high gamma band, GRUGate Transformer outperforms non-Transformer models by at least $\sim 9\%$ regarding the classification accuracy. In average, the performance of gated Transformers are noticeably higher than that of other state-of-the-art models. By comparing to the vanilla Transformers, we can also get that introducing the gating mechanisms increases the Transformer performance by a few percent, which is most obvious when running experiments on all-frequency band data. Additionally, the comparison among filtering frequency bands indicate that better performance is always achieved on high gamma band for each model, which is consistent with the literature on cognitive neuroscience (e.g., [6]) and the conclusion in [22].

We then evaluate the performance of GRUGate Transformer on the data that consists of temporal EEG subsequences, namely, we use the EEG signals in different time intervals as inputs. Table II shows that the best performance

TABLE I: EEG classification performance of state-of-the-art models using band-pass filters with different cut-off frequencies. The results are averaged across three runs. We use the whole EEG time course (20-460 ms) and data from all subjects for this experiment.

Models	High gamma ([55-95] Hz)	Beta - gamma ([14-70] Hz)	All freq. ([5-95] Hz)
Bi-LSTM	52.20%	45.30%	44.50%
SyncNet	30.39%	24.18%	26.64%
EEGNet	45.36%	34.51%	32.35%
EEG-ChannelNet	50.95%	40.64%	35.90%
CRAM	43.10%	35.60%	37.50%
Post-LN Transformer	58.07%	42.77%	37.11%
Pre-LN Transformer	57.90%	39.80%	34.60%
InputGate Transformer	58.99%	47.36%	48.32%
OutputGate Transformer	58.42%	47.28%	46.29%
HighwayGate Transformer	56.35%	46.58%	47.25%
SigTanhGate Transformer	59.34%	45.13%	**49.13%**
GRUGate Transformer	**61.11%**	**47.53%**	46.42%

TABLE II: EEG classification performance of GRUGate Transformer using different EEG time intervals with data filtered in the [55-95] Hz band. The results are averaged across three runs.

EEG time interval (ms)	Classification accuracy
20-240	52.22%
20-350	56.84%
20-460	**61.11%**
130-350	54.37%
130-460	56.06%
240-460	53.44%

TABLE III: EEG classification performance of state-of-the-art models using the data from each individual subject. The results are averaged across three runs. We use the entire EEG time course (20-460 ms) and high gamma band-pass filter ([55-95] Hz). For gated Transformers, we only evaluate GRUGate Transformer as a representative.

Models	Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Average (±std)
Bi-LSTM	30.20%	52.20%	57.60%	72.90%	53.80%	49.40%	52.68% (±12.59%)
SyncNet	40.21%	52.50%	46.25%	63.44%	45.11%	40.83%	48.06% (±7.98%)
EEGNet	19.58%	49.17%	48.85%	62.61%	39.06%	46.56%	44.31% (±13.06%)
EEG-ChannelNet	10.00%	59.90%	48.44%	62.50%	45.63%	44.69%	45.19% (±17.15%)
CRAM	31.40%	50.90%	57.70%	66.00%	45.10%	38.00%	48.18% (±11.62%)
Post-LN Transformer	40.31%	68.02%	60.42%	72.18%	51.56%	64.06%	59.43% (±10.69%)
Pre-LN Transformer	42.60%	64.69%	60.00%	68.33%	53.23%	62.71%	58.59% (±8.52%)
GRUGate Transformer	**43.02%**	**70.52%**	**63.75%**	**73.65%**	**56.25%**	**64.58%**	**61.96%** (±10.09%)

is achieved by using the entire time course. Leaving out the data from any time intervals will affect the classification results. We remark that similar observations can be obtained with other models, hence the results are omitted here.

In addition, we also assess the models with the data from each individual subject. The results are shown in Table III. For gated Transformers, we only evaluate on GRUGate Transformer as a representative. From the table, we can see that the model performance varies among subjects. The best and worst accuracies are consistently achieved from Subject 4 and Subject 1, respectively, which to some degree indicates the "data quality" for subjects. In particular, EEGNet and EEG-ChannelNet cannot learn features for Subject 1 well. More careful tuning or treatment of these two models on the data of Subject 1 is required. The performance of GRUGate Transformer is consistently higher than that of other models on the data of any single subject. When comparing the average accuracy with the performance that shown in Table I, we can find that some models are robust to the data size and distribution, including Bi-LSTM and three Transformer models, where the difference is within 2%.

b) Motor imagery dataset: We now present results for the classification task on the motor imagery (movement intention) dataset. Unlike the brain-visual classification task, we consider the cross-subject scenario in this experiment, namely, the validation data and test data are from subjects that are unseen in the training set. To this end, we randomly generate six different splits of training, validation and test data. Table IV shows the average accuracy and the variance across six runs for each model.

Among non-Transformer models, EEGNet has achieved the best mean accuracy and SyncNet also performs better than other models. The performance for Mesh-Cascade model is very low in the subject-wise experiments, which is much different from the performance reported in [35]. In [35], Mesh-Cascade outperforms EEGNet and SyncNet, and achieves the state-of-the-art results, where the cross-subject data is mixed, namely, data from the same subject can be seen in training, validation and test sets. On the other hand, for Transformer architectures, GRUGate Transformer has achieved comparable results with lower variance in contrast to EEGNet. However, we do not see improvement of Transformer models in this case. We suspect that Transformer-like models are more powerful in higher frequencies, but perform similarly to other state-of-the-art models in lower-frequency brainwaves, e.g., delta to beta. The similar observation can be obtained from Table I as well, but the hypothesis requires further investigation, so we leave it as a future work.

TABLE IV: EEG classification performance of state-of-the-art models on motor imagery task. The results are averaged across six runs with different subject-wise data splits.

Models	Classification accuracy
CNN	49.52% (±2.91%)
Bi-LSTM	50.66% (±1.47%)
SyncNet	54.21% (±2.84%)
EEGNet	**55.46%** (±2.30%)
EEG-ChannelNet	52.77% (±2.04%)
CRAM	46.68% (±2.43%)
Mesh-Cascade	31.60% (±2.01%)
Post-LN Transformer	54.28% (±2.95%)
Pre-LN Transformer	53.88% (±2.95%)
InputGate Transformer	54.48% (±3.11%)
OutputGate Transformer	53.61% (±3.52%)
HighwayGate Transformer	54.79% (±3.15%)
SigTanhGate Transformer	54.32% (±2.78%)
GRUGate Transformer	**55.40%** (±2.09%)

IV. CONCLUSION

In this work, we have used the Transformer models with gating mechanism for decoding human brain EEG signals. The gated Transformers apply the attention mechanism to learn long-term temporal dependencies of the EEG signals, and also employ the gating mechanism to stabilize the training process. Experiments on two datasets, i.e., brain-

visual dataset and motor imagery dataset, have demonstrated the effectiveness of the gated Transformer models over multiple classes and subjects. The variants of Transformer architectures and gating layer formulations also provide customizability when developing BCIs in practice.

A natural extension of this work in the future is to further evaluate the gated Transformers in other applications, and incorporate them in end-to-end BCI development. In addition, we want to investigate the factors that can impact the performance of gated Transformers, such as band-pass filtering frequencies and other pre-processing approaches, the electrodes placement and device used to collect EEG data. Moreover, it is also of interest to improve the explanability of our models. For example, when predicting a class, the models can also indicate which part of electrodes and which range of the time course contribute most to the prediction.

ACKNOWLEDGMENT

We want to thank Concetto Spampinato and Isaak Kavasidis from PeRCeiVe Lab for the help on the brain-visual data classification, including the data processing and the implementation for several baseline models.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," In 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283) (2016).

[2] S. Alhagry, A.A. Fahmy, and R.A El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," Emotion, 8(10), pp.355-358 (2017).

[3] J.L. Ba, J.R. Kiros, and G.E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450 (2016).

[4] A. Baevski, and M. Auli, "Adaptive input representations for neural language modeling," arXiv preprint arXiv:1809.10853 (2018).

[5] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," arXiv preprint arXiv:1904.10509 (2019).

[6] M.S. Clayton, N. Yeung, and R.C. Kadosh, "The roles of cortical oscillations in sustained attention," Trends in cognitive sciences, 19(4), pp.188-195 (2015).

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555 (2014).

[8] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE (2009).

[9] G.N. Garcia, T. Ebrahimi, and J.M. Vesin, "Support vector EEG classification in the Fourier and time-frequency correlation domains," First International IEEE EMBS Conference on Neural Engineering (pp. 591-594) (2003).

[10] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," circulation, 101(23), pp.e215-e220 (2000).

[11] N.F. Güler, E.D. Übeyli, and I. Güler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," Expert systems with applications, 29(3), pp.506-514 (2005).

[12] D.P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).

[13] T.N. Kipf, and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907 (2016).

[14] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," Journal of Neural Engineering, vol. 15, no. 5, p. 056013 (2018).

[15] G. Li, C.H, Lee, J.J. Jung, Y.C. Youn, and D. Camacho, "Deep learning for EEG data analytics: A survey," Concurrency and Computation: Practice and Experience, 32(18), p.e5199 (2020).

[16] J. Li, Z. Struzik, L. Zhang, and A. Cichocki, "Feature learning from incomplete EEG with denoising autoencoder," Neurocomputing, 165, pp.23-31 (2015).

[17] Y. Li, M. Murias, S. Major, G. Dawson, K. Dzirasa, L. Carin, and D.E. Carlson, "Targeting EEG/LFP Synchrony with Neural Nets," NIPS (pp. 4620-4630) (2017).

[18] X. Lun, S. Jia, Y. Hou, Y. Shi, Y. Li, H. Yang, S. Zhang, and J. Lv, "GCNs-Net: A Graph Convolutional Neural Network Approach for Decoding Time-resolved EEG Motor Imagery Signals," arXiv preprint arXiv:2006.08924 (2020).

[19] R.M. Mehmood, and H.J. Lee, "Emotion classification of EEG brain signal using SVM and KNN," IEEE international conference on multimedia & expo workshops (ICMEW) (pp. 1-5) (2015).

[20] M. Murugappan, N. Ramachandran, and Y. Sazali, "Classification of human emotion from EEG using discrete wavelet transform," Journal of biomedical science and engineering, 3(04), p.390 (2010).

[21] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[22] E. Parisotto, F. Song, J. Rae, R. Pascanu, C. Gulcehre, S. Jayakumar, M. Jaderberg, R.L. Kaufman, A. Clark, S. Noury, and M. Botvinick, "Stabilizing transformers for reinforcement learning," International Conference on Machine Learning (pp. 7487-7498). PMLR (2020).

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," arXiv preprint arXiv:1912.01703 (2019).

[24] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," IEEE Transactions on biomedical engineering, 51(6), pp.1034-1043 (2004).

[25] A. Subasi, and M.I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," Expert systems with applications, 37(12), pp.8659-8666 (2010).

[26] Z. Tang, C. Li, and S. Sun, "Single-trial EEG classification of motor imagery using deep convolutional neural networks," Optik, 130, pp.11-18 (2017).

[27] R. Tomioka, K. Aihara, and K.R. Müller, "Logistic regression for single trial EEG classification," Advances in neural information processing systems (pp. 1377-1384) (2007).

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, 30, pp.5998-6008 (2017).

[29] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D.F. Wong, and L.S. Chao, "Learning deep transformer models for machine translation," arXiv preprint arXiv:1906.01787 (2019).

[30] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.Y. Liu, "On layer normalization in the transformer architecture," arXiv preprint arXiv:2002.04745 (2020).

[31] D. Zhang, K. Chen, D. Jian, and L. Yao, "Motor imagery classification via temporal attention cues of graph embedded EEG signals," IEEE journal of biomedical and health informatics, 24(9), 2570-2579 (2020).

[32] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent eeg signal analysis," IEEE Signal Processing Letters, 26(5), pp.715-719 (2019).

[33] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for EEG-based human intention recognition," IEEE transactions on cybernetics, 50(7), 3033-3044, (2019).

[34] D. Zhang, L. Yao, K. Chen, S. Wang, P.D. Haghighi, and C. Sullivan, "A graph-based hierarchical attention model for movement intention detection from EEG signals," IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27(11), 2247-2253 (2019).

[35] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, R. Boots, and B. Benatallah, "Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface," In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1) (2018).