Spatial-Temporal Fusion of Electroencephalography Data for Transformer-Based Gaze Prediction

Names Redacted
Department of Computer Science, The George Washington University
2134 G St NW, Washington, DC 20052

Introduction

Electroencephalography (EEG) is a non-invasive technique to record the electrical activity generated by the brain. Owing to its relative accessibility and superior temporal resolution compared to other neuroimaging techniques, EEG's potential extends to many different fields.

One such field is the complementary applications with eye-tracking. Among the many uses, one task stands out for investigation: eye-tracking using EEG data (Montenegro and Argyriou 2016; Sun et al. 2023; Kastrati et al. 2023).

EEGViT (Yang and Modesitt 2023) is the current state-of-the-art (SOTA) model on EEG-based gaze prediction accuracy on the EEGEyeNet dataset (Kastrati et al. 2021). It employs a hybrid transformer model fine-tuned with EEG data (Khan et al. 2022; Vaswani et al. 2017).

Research Questions

In this paper, we propose two methods that attempt to answer these two questions:

- The SOTA performs convolution on a fixed-size subset of the EEG channels each time. Can convolution over all channels improve accuracy?
- Would a vision transformer model with spatial embeddings yield similar or better performance than the spatial convolution layer used in SOTA?

Related Works

Dataset

EEGEyeNet (Kastrati et al. 2021) is a dataset that offers EEG and eye tracking data collected simultaneously using a 128-channel EEG Geodesic Hydrocel system shown in Figure 1. In one of the experimental paradigms, the participants fixate on specific dots on a "large grid" on a 600×800 screen as seen in Figure 2. The collected gaze positions can be seen in Figure 3 (Kastrati et al. 2021).

EEGEyeNet also proposes a benchmark where a model predicts the 2-dimensional gaze position from 128-channel, 500-time-step EEG signals (Kastrati et al. 2021). The model is then evaluated by the Root Mean Squared Error (RMSE) in pixels or millimeters (where 1 millimeter equals 2 pixels).

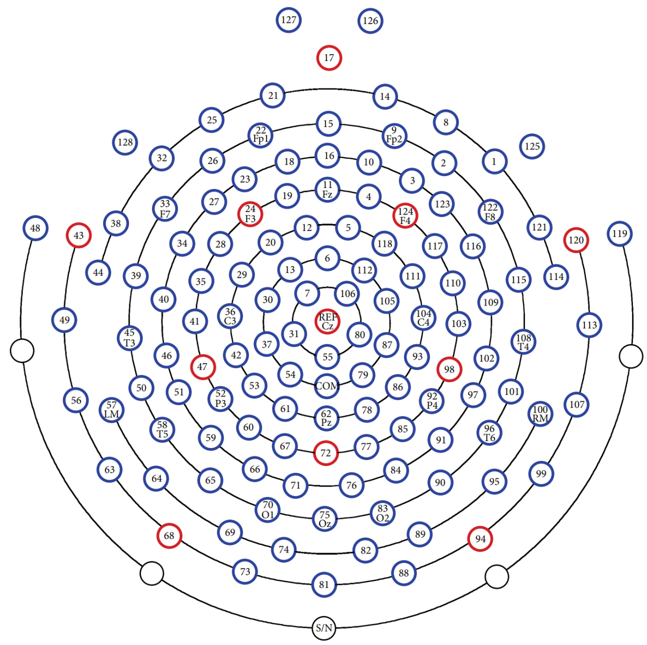

Figure 1: Electrode Layout of the 128-channel EEG Geodesic Hydrocel system (Bamatraf et al. 2016)

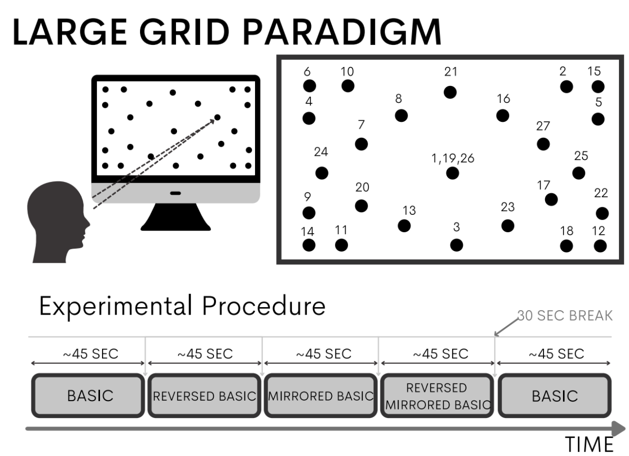

Figure 2: The Large Grid Paradigm of EEGEyeNet (Kastrati et al. 2021)

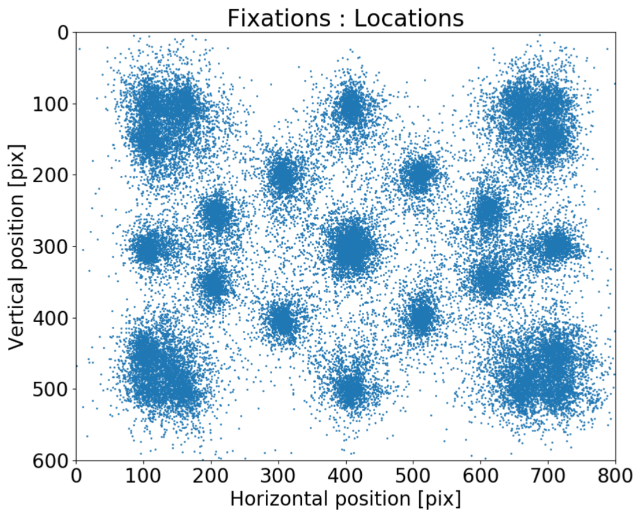

Figure 3: Distribution of the Fixation Positions in the Large Grid Paradigm (Kastrati et al. 2021)

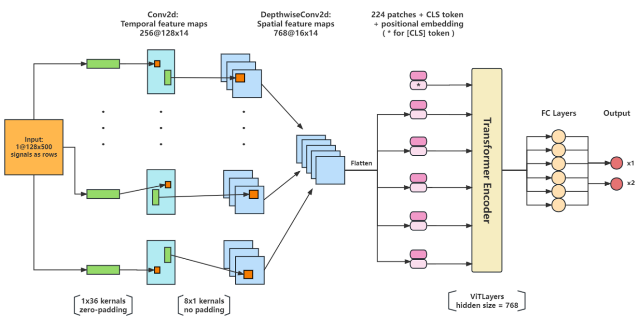

Figure 4: EEGViT Model Architecture (Yang and Modesitt 2023)

Prior Works

Baseline EEGEyeNet established the naive baseline of 123.3 millimeters RMSE by predicting the mean position of the training set, the Convolutional Neural Network (CNN) baseline of 70.2 millimeters with a standard 1D CNN with max pooling, and an EEGNet result of 81.7 millimeters (Kastrati et al. 2021). A comparison can be found in Table 1.

Spatial-Temporal Fusion of EEG Data A two-level convolution feature extraction method was first proposed in EEGNet (Lawhern et al. 2018) and Filter Bank Common Spatial Patterns (Schirrmeister et al. 2017) which enables efficient extraction of spatial (EEG electrodes) features for each temporal (frequency) channel.

State-of-the-Art on EEG Gaze Prediction Combining the convolution layers of EEGNet and a vision transformer using the ViT-Base model (Dosovitskiy et al. 2020) pre-trained with ImageNet (Deng et al. 2009; Ridnik et al. 2021) as shown in Figure 4, EEGViT by (Yang and Modesitt 2023) achieves an RMSE of 55.4 ± 0.2 millimeters on the EEGEyeNet dataset.

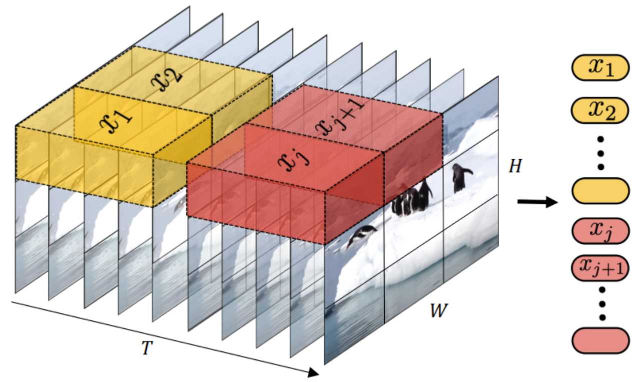

Figure 5: Tubelet Embedding for Videos (Arnab et al. 2021)

Methods

We plan to evaluate the research questions by testing the methods below.

Method 1: Large Spatial Convolution Kernel

The architecture of our Method 1 can be seen in Figure 6. Similar to prior works (Yang and Modesitt 2023; Lawhern et al. 2018; Schirrmeister et al. 2017), we employ two convolution layers which filters the temporal and spatial (channel) dimensions respectively.

In the first layer, a 1×16 kernel scans across the 1-second 128×500 input which is zero-padded to 128×512. The kernels effectively function as band-pass filters on the raw input signals. Our choice of 1×16 kernel is smaller than that of EEGViT at 1×36 (Yang and Modesitt 2023) and that of EEGNet at 1×64 (Lawhern et al. 2018). This provides a greater resolution of temporal features to be learned. Batch normalization is then applied on the 128×32 output (Ioffe and Szegedy 2015).

In the second layer, a depth-wise 128×1 kernel scans over all EEG channels of each temporal filter. We hypothesize that better results are achievable with our kernels of shape $(C, 1)$ where $C = 128$ is the number of EEG channels. This kernel will be able to learn any spatial relationships between any two EEG channels at the same point in time.

The model is trained for 15 epochs on a NVIDIA V100 in batches of 64 samples, with an initial learning rate of 1e-4 which is dropped by a factor of 10 every 6 epochs.

Method 2: Tubelet Embedding of Temporal Features

The architecture of our Method 2 can be seen in Figure 7. Since EEG data is recorded by attaching electrodes to the scalp in a layout shown in Figure 1, we hypothesize that it is possible to model the electrodes as recordings on a two-dimensional plane. We then consider the temporal-filtered features as a 3D volume of two spatial dimensions and one temporal dimension. Then, inspired by ViViT by (Arnab et al. 2021), a video vision transformer, we will extract spatial-temporal ”tubes” from the input volume as seen in Figure 5,

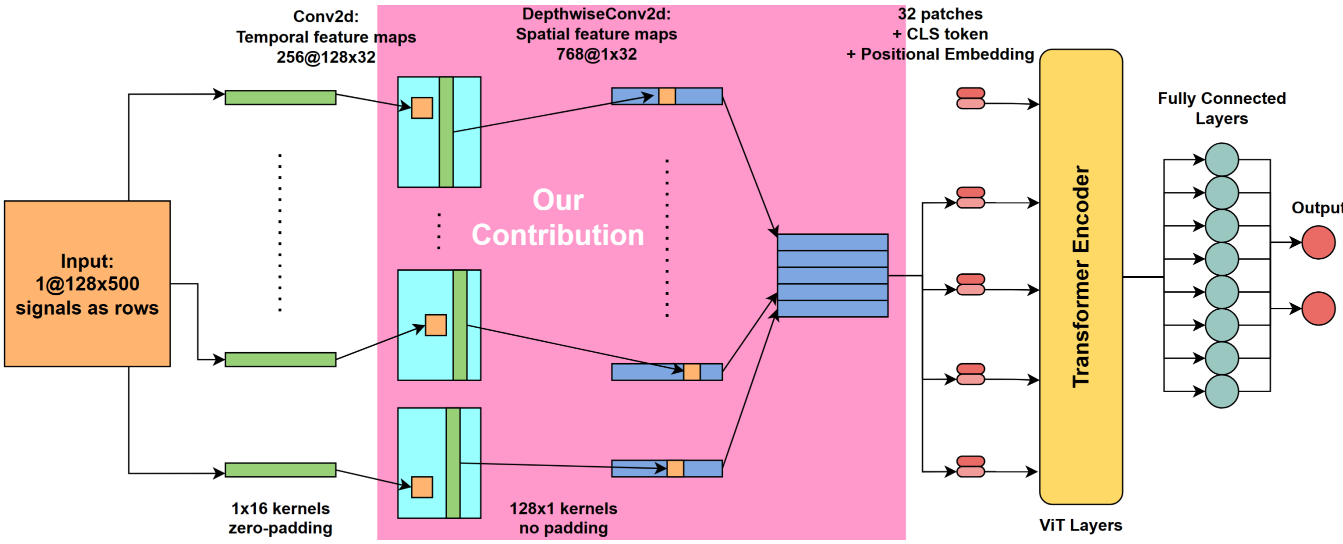

Figure 6: Method 1 Model Architecture, modified from (Yang and Modesitt 2023)

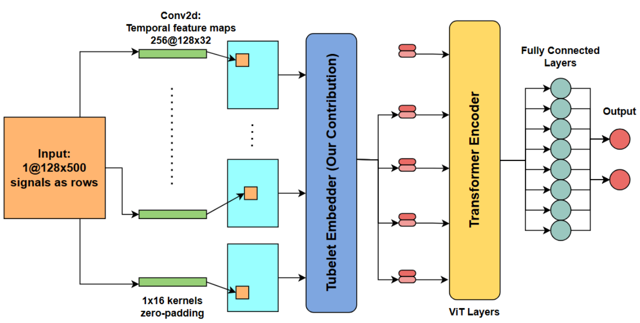

Figure 7: Method 2 Model Architecture, modified from (Yang and Modesitt 2023)

which can then be flattened and directly sent to a transformer encoder.

Results

A comparison can be found in Table 1. We achieved an RMSE of average 51.3 millimeters and standard deviation of 0.4 millimeters across 5 runs[1]. A figure of losses during training and a figure of its predictions in the testing set can be seen in Figure 8 and Figure 9.

We also hypothesize that applying method 2 separately

Model	RMSE [mm]	Study
Naive Baseline	123.3 ± 0	Kastrati et al. 2021
KNN	119.7 ± 0	Kastrati et al. 2021
RBF SVR	123 ± 0	Kastrati et al. 2021
Linear Regression	118.3 ± 0	Kastrati et al. 2021
Ridge Regression	118.2 ± 0	Kastrati et al. 2021
Lasso Regression	118 ± 0	Kastrati et al. 2021
Elastic Net	118.1 ± 0	Kastrati et al. 2021
Random Forest	116.7 ± 0.1	Kastrati et al. 2021
Gradient Boost	117 ± 0.1	Kastrati et al. 2021
AdaBoost	119.4 ± 0.1	Kastrati et al. 2021
XGBoost	118 ± 0	Kastrati et al. 2021
CNN	70.2 ± 1.1	Kastrati et al. 2021
PyramidalCNN	73.6 ± 1.9	Kastrati et al. 2021
EEGNet	81.7 ± 1.0	Kastrati et al. 2021
InceptionTime	70.8 ± 0.8	Kastrati et al. 2021
Xception	78.7 ± 1.6	Kastrati et al. 2021
ViT-Base	61.5 ± 0.6	Yang et al. 2023
- Pre-trained	58.1 ± 0.6	Yang et al. 2023
EEGViT	61.7 ± 0.6	Yang et al. 2023
- Pre-trained	55.4 ± 0.2	Yang et al. 2023
Ours (Method 1)	$\mathbf{51.3 \pm 0.4}$	-
Ours (Method 2)	-	-

Table 1: Existing EEGEyeNet Gaze Position RMSE Means and Standard Deviation across 5 Runs

[1]The source code of the 5 runs can be found at:
https://colab.research.google.com/drive/1E64a4uimCl9l8ETMK2zxyFs5wGAOtbkq
https://colab.research.google.com/drive/1Gql1N0WvsODaxBbbXWKMgVbp2cKNPw_E
https://colab.research.google.com/drive/1JV2YyRQFO_i0YwCcjQVJr6g1cJh4Iar
https://colab.research.google.com/drive/1UDB5O9qH-GVKI-E3FHWEZmR5nRHvMNtW
https://colab.research.google.com/drive/1pZdMuiTf4NDk44W5ssjCv4FZoynMhqvl

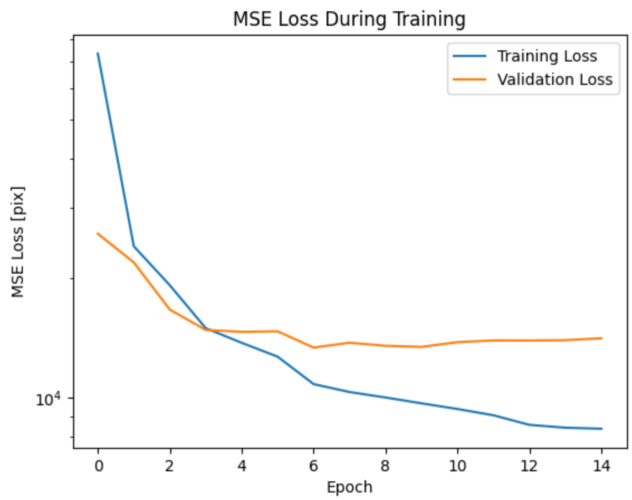

Figure 8: Method 1 MSE Loss During Training

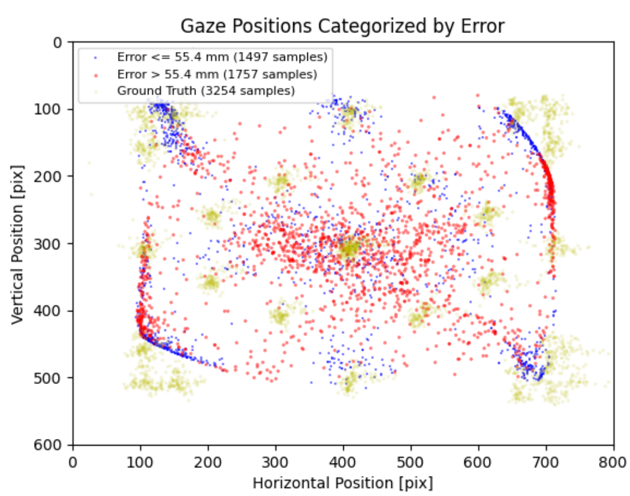

Figure 9: Method 1 Gaze Position Coordinates, where predictions with lower error than the mean of SOTA are colored blue, and higher ones are colored red, and the ground truths are colored yellow.

or in combination with method 1 may further improve the accuracy. This will be our topic of focus in the following two months.

Discussion

Our method 1 outperforms the existing state-of-the-art OTA (Yang and Modesitt 2023) by a clear margin. This is achieved by choosing the spatial convolution kernel to cover all EEG channels, which is able to learn stronger spatial features than SOTA's $(8, 1)$ spatial convolution kernel.

Conclusion

In this paper, we proposed two methods of EEG-based gaze prediction that potentially outperform the SOTA. We evaluated changes to the spatial-temporal fusion of previous works and then presented our preliminary results from our first method, which already surpasses current SOTA in accuracy.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bamatraf, S.; Hussain, M.; Aboalsamh, H.; Qazi, E.-U.-H.; Malik, A. S.; Amin, H. U.; Mathkour, H.; Muhammad, G.; and Imran, H. M. 2016. A system for true and false memory prediction based on 2d and 3d educational contents and eeg brain signals. *Computational Intelligence and Neuroscience* 2016:45–45.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.

Kastrati, A.; Płomecka, M. B.; Pascual, D.; Wolf, L.; Gillioz, V.; Wattenhofer, R.; and Langer, N. 2021. Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. *arXiv preprint arXiv:2111.05100*.

Kastrati, A.; Plomecka, M. B.; Küchler, J.; Langer, N.; and Wattenhofer, R. 2023. Electrode clustering and bandpass analysis of eeg data for gaze estimation. In *Annual Conference on Neural Information Processing Systems*, 50–65. PMLR.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54(10s):1–41.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering* 15(5):056013.

Montenegro, J. M. F., and Argyriou, V. 2016. Gaze estimation using eeg signals for hci in augmented and virtual reality headsets. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1159–1164. IEEE.

Ridnik, T.; Ben-Baruch, E.; Noy, A.; and Zelnik-Manor, L. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Schirrmeister, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; and Ball, T. 2017. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping* 38(11):5391–5420.

Sun, R.; Cheng, A. S.; Chan, C.; Hsiao, J.; Privitera, A. J.; Gao, J.; Fong, C.-h.; Ding, R.; and Tang, A. C. 2023. Tracking gaze position from eeg: Exploring the possibility of an eeg-based virtual eye-tracker. *Brain and Behavior* e3205.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Yang, R., and Modesitt, E. 2023. Vit2eeg: Leveraging hybrid pretrained vision transformers for eeg data. *arXiv preprint arXiv:2308.00454*.