ViT2EEG: Leveraging Hybrid Pretrained Vision Transformers for EEG Data

Ruiqi Yang*
ruiqiyang@ucsb.edu
University of California Santa Barbara
Santa Barbara, CA, USA

Eric Modesitt*
ericjm4@illinois.edu
University of Illinois at Urbana Champaign
Urbana-Champaign, Illinois, USA

ABSTRACT

In this study, we demonstrate the application of a hybrid Vision Transformer (ViT) model, pretrained on ImageNet, on an electroencephalogram (EEG) regression task. Despite being originally trained for image classification tasks, when fine-tuned on EEG data, this model shows a notable increase in performance compared to other models, including an identical architecture ViT trained without the ImageNet weights. This discovery challenges the traditional understanding of model generalization, suggesting that Transformer models pretrained on seemingly unrelated image data can provide valuable priors for EEG regression tasks with an appropriate fine-tuning pipeline.

The success of this approach suggests that the features extracted by ViT models in the context of visual tasks can be readily transformed for the purpose of EEG predictive modeling. We recommend utilizing this methodology not only in neuroscience and related fields, but generally for any task where data collection is limited by practical, financial, or ethical constraints. Our results illuminate the potential of pretrained models on tasks that are clearly distinct from their original purpose.

CCS CONCEPTS

• **General and reference** → *Empirical studies*; • **Computing methodologies** → **Neural networks**; **Transfer learning**; *Machine learning algorithms*; Knowledge representation and reasoning; • **Applied computing** → **Bioinformatics**.

KEYWORDS

Pre-trained Models, Hybrid Vision Transformers, CNN, Transfer Learning, Regression Tasks, EEG, time series, spatio-temporal data.

---

*Both authors contributed equally to this research.

---

1 INTRODUCTION

With its rich, multidimensional structure, electroencephalogram (EEG) data encapsulates a wealth of information about brain activity, offering unique insights into a variety of neurological phenomena [34]. However, the intricacies that define EEG data introduce substantial hurdles towards devising efficient, yet effective predictive models, especially considering the incredibly costly process of data collection [6]. Despite the extensive usage of machine learning regression models for predicting EEG data, these models often fall short of adequately capturing and decoding the complex structures within such data. As an example, the EEGEyeNet dataset [20] provides a baseline for predicting the position of a subject's gaze on a screen given EEG data. Their results clearly demonstrate the inability of machine learning models to accurately interpret the given data. We also explore articles that employ analogous time-series data [11, 25, 26, 28–30, 32, 41].

In recent years, there has been a significant surge in the invention and application of deep learning models to unravel this complexity and extract valuable insights from EEG data. In particular, Convolutional Neural Networks (CNNs) and [23] the Self-Attention operator [36] have demonstrated promising results in managing the complexity of EEG data [1, 31], with several architectures being derived from their successful counterparts in the field of Computer Vision [4, 21, 38]. These deep learning techniques excel at capturing nonlinear patterns and relationships within data, making them particularly suitable for modeling complex EEG signals.

Our study takes the paradigm of converting successful computer vision models to an EEG counterpart a step further and proposes the use of a hybrid Vision Transformer (ViT) [10] pretrained on the ImageNet [8] dataset to take on an EEG regression task. This approach takes a theoretical leap from previous methods, repurposing a model's architecture and weights initially designed for vision tasks to grapple with the high dimensionality and inherent complexity of EEG data. This research is predicated on the observation that EEG data, like image data, possesses rich, high-dimensional structures that can be effectively modeled using Transformer models [36].

Our empirical findings indicate that utilizing a pretrained ViT model significantly outperforms traditional regression models. In addition, we demonstrate that the success of the model comes from not just the model architecture, but also the impact of pretraining the ViT on the model's initial weights. This success underlines the strength of pretraining the model across disciplines, demonstrating how models trained on abundant, easily obtained image data can be effectively repurposed for EEG data analysis, an area where data collection often involves considerable challenges due to practical, financial, and ethical constraints [31].

2 RELATED WORK

2.1 Deep Learning for EEG Tasks

EEG regression and classification tasks, fundamental to neuroscience research and applications like Brain-Computer Interfaces (BCIs), have traditionally relied on Convolutional Neural Networks (CNNs) [7, 44]. While CNNs have proven their use for extracting localized features from time-series EEG data, they struggle with capturing the long-term dependencies, subject-independent and session-independent patterns inherent in EEG data [1]. This limitation motivates the exploration of other model architectures that can replace or support CNNs.

Recently, hybrid methods complementing CNNs with Self-Attention modules [3] have shown promising improvements in performance by capturing both local patterns and long-term dependencies in EEG signals [33, 40]. In addition, recent studies in Computer Vision have demonstrated the superiority of Transformer-focused models over predictive models focused on convolutional layers [35, 37]. Following this trend, our study aims to test the standalone capabilities of Transformers (specifically Vision Transformers), expecting the ViT's ability to capture global dependencies in the data to deliver superior performance in the complex, high-dimensional EEG data present in eye-tracking tasks.

2.2 EEG Eye-Tracking with EEGEyeNet

Using EEG data for applications in eye-tracking presents unique challenges and applications, ranging from assisting with cognitive studies to developing assistive technologies. The recent introduction of the EEGEyeNet dataset [20], containing 47 hours of high-density 128-channel EEG data synchronized with eye-tracking recordings from 356 healthy adults, offers a valuable resource for studying such complex relationships. This dataset not only aids in studying attention and reaction time, but it also provides a preprocessing pipeline and benchmarks for gaze estimation, facilitating reproducible research every step of the way.

2.3 Vision Transformers (ViTs)

Introduced by [10], Vision Transformers (ViTs) have demonstrated impressive results in image classification tasks. Unlike CNNs, which use spatial convolution operations for feature extraction, ViTs divide the input image into a grid of patches and leverage transformer architecture to process these patches as a sequence. This design allows ViTs to capture global dependencies in the input data without convolution operations, making them more suited for the globally-correlated nature of EEG data.

Research in similar domains indicates the applicability of ViTs. For example, studies have demonstrated ViT's effectiveness in tasks involving audio [14] and video processing [2, 43], hinting at its potential for spatial-temporal analysis of EEG data. Despite its potential, the usage of ViTs in EEG analysis is largely unexplored, prompting the need for research like the current study.

2.4 Pretrained ViTs

The innate capacity of the self-attention mechanism in Vision Transformer (ViT) models to possess a global receptive field facilitates
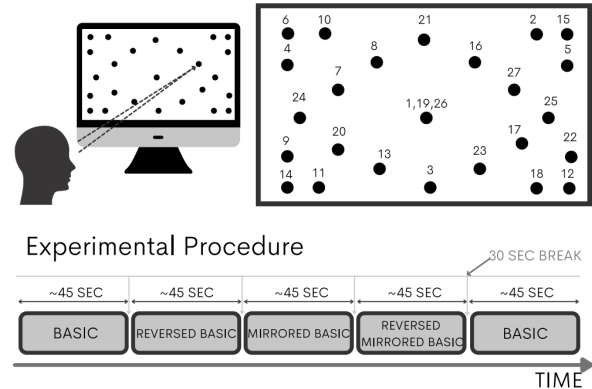

Figure 1: EEGEyeNet Large Grid Paradigm [20]. Participants are asked to fixate on particular dots in a given period.

simultaneous contemplation of all input fragments during decision-making procedures. This particular ability is critical in capturing intricate and distant dependencies, a characteristic frequently encountered in Electroencephalogram (EEG) data. Moreover, the transferability of pretrained ViT models [15, 24, 42], which are originally trained on vast datasets, permits their fine-tuning for subsequent tasks. This property is particularly advantageous for EEG-related endeavors, which often grapple with a dearth of labeled data.

Nonetheless, despite the demonstrable success and potential of pretrained ViT models, it is imperative to recognize their limitations when deployed directly to realms that are substantially distinct from their original training context. One such example is the inherently high-dimensional and noisy nature of EEG data. Hence, the present investigation critically assesses the efficacy of employing a pretrained ViT model for an EEG regression task. This exploration could consequently lay a foundation for future scientific inquiries in this domain.

3 METHODS

3.1 Structuring Images from EEG Time Series Data

The data utilized in this research comes directly from the large grid paradigm of EEGEyeNet (see Fig. 1). Participants of this paradigm are instructed to fixate on a sequence of 25 dots appearing at distinct screen positions, for 1.5 to 1.8 seconds each. We primarily used data

Participants			
Total	Train	Validation	Test
27	19	4	4
Samples			
Total	Train	Validation	Test
21464	14706	3277	3481

Table 1: Distribution of participants and samples in the Abs. Position dataset.
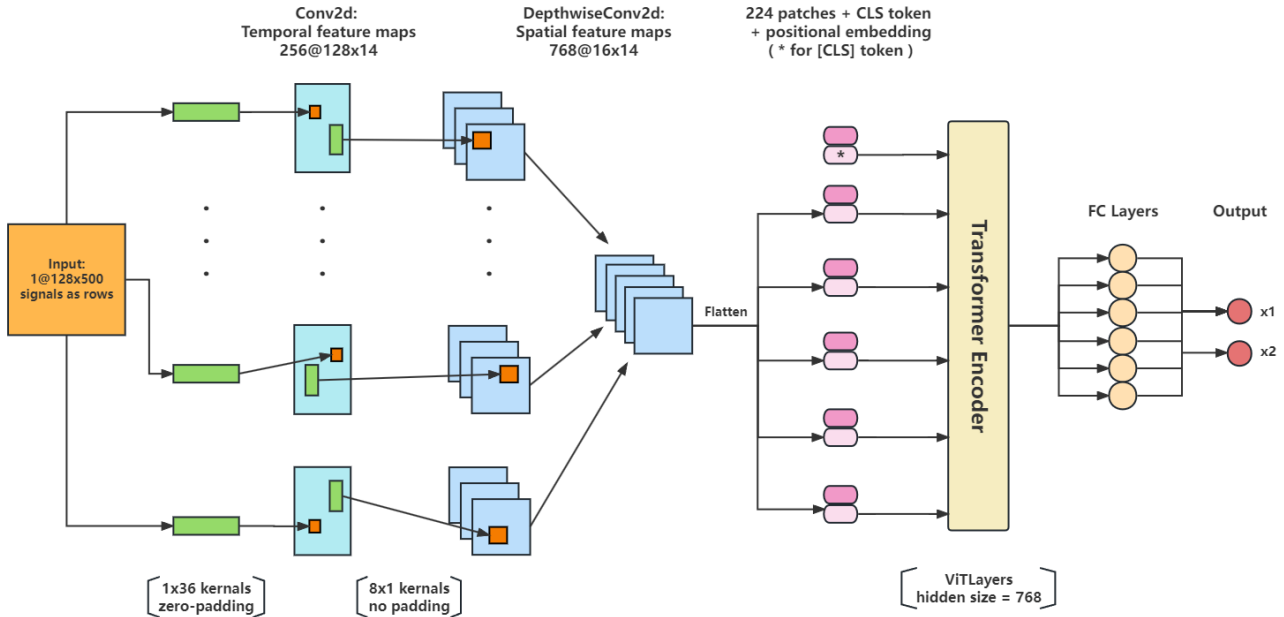
Figure 2: Proposed EEGViT, a hybrid ViT (Vision Transformer) architecture designed specifically for EEG raw signal as input. A two-step convolution operation is applied to generate patch embeddings. Then we add positional embeddings and pass the resulting sequence into ViT layers. The illustration of positional embedding and ViT layer is based on [10].

from EEGEyeNet's Absolute Position Task, which concentrates on pinpointing the precise XY-coordinates of the subject's gaze on the screen [20].

The dataset consists of raw EEG signals collected by a 128-channel Geodesic Hydrocel system at sampling rate of 500Hz, and it undergoes a preprocessing process recommended by [16]. Initially, defective electrodes that could potentially introduce noise were identified and handled. Following this, the data underwent frequency filtering, with a 40 Hz high-pass filter and a 0.5 Hz low-pass filter applied, which allowed us to retain useful frequencies and eliminate potential noise. Eventually, the resulted signals were segmented into 1-second trials including 500 time samples for each of 128 channels, and we stacked channels row by row to form a 128×500 matrix as one input trial.

The processed data was then divided into training, validation, and test sets in a 70%, 15%, 15% distribution. The validation and test data sets were made up of distinct subjects, as depicted in Table 1. This approach allowed us to assess the model's ability to generalize and adapt to new data, providing a more robust and accurate evaluation of its potential real-world application.

3.2 Model Architecture

The model presented in this paper, named EEGViT, is a hybrid Vision Transformer architecture that integrates a two-step convolution operation during the patch embedding process (Fig. 2).

3.2.1 Two-Step Convolution Block. We adopt the two-step convolution block utilized in EEG analysis from the early layers of EEGNet

[22]. This block consists of two convolutional layers, with one layer filtering the temporal dimension and the other layer filtering the channel (spatial) dimension.

The first layer employs a $1 \times T$ kernel and scans across the entire input, capturing temporal events that occur over the same channels. The kernels in this temporal convolutional layer can be regarded as various band-pass filters applied to the raw signals. The resulting output feature maps represent the frequency bands of interest for the EEG. Batch normalization is applied on the output [18].

Following this, a depthwise convolutional layer [17] containing a $C \times 1$ kernel is used. This layer scans over multiple channels, given the same point in time, filtering the inputs separately. This depthwise convolution also acts as a frequency-specific spatial filter when applied to EEG feature extraction in this context.

The overall operation can be viewed as splitting the input images into $C \times T$ patches, followed by a row-by-row linear projection, with the resulting column vector transformed into a scalar feature.

Layers	Kernel	Channels	Options
Conv2D	1×36	256	padding = (0,2)
			stride = (1,36)
BatchNorm2D		256	
DepthwiseConv2D	8×1	768	stride = (8,1)

Table 2: Details of Two-Step Convolution Block in EEGViT, implemented with PyTorch library [27]
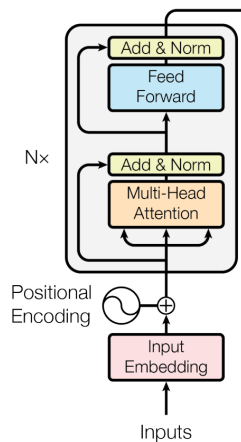
Figure 3: Transformer Encoder block by [36] We follow this paradigm to implement the Transformer block in EEGViT.

The decision to separate the projection for each dimension and incorporate the depthwise mechanism proves advantageous.

To align with the ViT-base model's encoder layers, which are pretrained on images divided into 192 patches [10], we carefully select hyperparameters for an output of 224 patch embedding vectors. These patches, each representing the activity of eight adjacent channels over 36 timestamps (72ms at a 500Hz sampling rate), are set to be non-overlapping, based on experimental findings. Hyperparameters are further tuned for optimal feature preservation and computational efficiency (Table 2).

Lastly, to integrate our patch sequence into the pretrained ViT layers, which use 192 patches and corresponding pretrained positional embeddings, we reinitialize a sequence of positional embeddings to match our patch sequence length."

3.2.2 Transformer Block. We adopt the transformer encoder paradigm (see Fig. 3) and utilize the ViT-Base model architecture in [10] to implement the transformer block in our model. This block consists of 12 modules, each comprising a Multi-Head Attention layer and a Multi-Layer Perceptron, wrapped by residual connection and layer normalization. Detailed hyperparameters for these layers can be found in Table 3.

For the pretrained version of the transformer block, we load the weights of the encoder layers from the ImageNet-pretrained ViT-Base model, consisting of approximately 86M parameters, and readily available on HuggingFace [39].

The input to the transformer block is a sequence of 1D patch embeddings obtained by flattening the output feature maps of the preceding Two-Step Convolution block. We incorporate positional embeddings into this sequence. Additionally, inspired by ViT, we

	Modules	Hidden Size	MLP Size	Heads
ViT-Base	12	768	3072	12

Table 3: Details of ViT-Base model [10] adopted for the transformer block in EEGViT

adopt BERT's approach of appending a CLS token to the sequence. This CLS token aids in aggregating information across the sequence [9]. The hidden state of the CLS token ultimately serves as the representation of the entire sequence of patch embeddings and is subsequently transformed into the target space, which corresponds to 2D coordinates on the grid, using fully connected layers.

3.3 Training

Training machine learning models effectively requires a meticulous and strategic approach to setting various parameters and configurations. The goal is to optimize the model's learning process, enabling it to understand complex patterns in the data and deliver reliable predictions on new, unseen data. The EEGViT model was trained with such attention to detail.

For our study, we adopted the Mean Squared Error (MSE) loss function as the primary measure to guide our model's learning. MSE loss function is widely used in regression problems due to its mathematical simplicity and effectiveness. In our specific context of EEG signal analysis, it allows the model to minimize the prediction error of continuous output values, essentially improving the model's ability to predict the exact EEG signal values. Although we employed MSE for training, we decided to use Root Mean Square Error (RMSE) to measure our results.

It's essential to mention that using MSE as the loss function and subsequently using RMSE as the performance metric is a commonly adopted approach in various scientific studies, particularly those involving EEG and eye-tracking data [20]. The use of these metrics is not only restricted to our work but has also been used extensively in similar studies.

This common usage is mainly due to the advantage that these metrics offer in evaluating the model's ability to accurately predict continuous outcomes while minimizing the overall deviation. Moreover, since these metrics are squared, they place a higher penalty on larger errors, thereby pushing the model to improve its overall accuracy. This attribute is extremely beneficial in sensitive analyses like EEG signal interpretation, where the aim is to minimize prediction discrepancies as much as possible.

Furthermore, the use of RMSE as a performance metric provides us a direct comparison with other research findings in the field of EEG eye-tracking. This allows for a standardized evaluation across various studies, enhancing the generalizability and transferability of our findings within the field. Given that other studies have successfully employed the same method, it strengthens our confidence in the adopted approach and provides more credibility to our results.

In order to compare different strategies and evaluate the effectiveness of our proposed approach, we trained two different model architectures. The first was the unmodified, non-pretrained ViT-Base model. This model, serving as a benchmark, was a simple architecture comprised of a single convolutional layer that used 8x36 kernels (i.e. patches) to generate patch embeddings. Alongside this plain version, we used a pretrained version of the ViT-Base model trained on the ImageNet dataset [8], using the weights from [39]. This pre-training approach aimed to utilize any cross-image features the ViT learned from the vast ImageNet dataset that may be applicable to EEG data.

Secondly, we modified the ViT-Base model into the proposed EEGViT model by adding the Two-Step Convolution Block. We trained two versions of the EEGViT model, one using randomly initlized weights and the other relying on the weights of ViT-Base obtained on the ImageNet dataset. This in effect allows for our EEGViT model to be comparable to both versions of the ViT-Base model.

Each of these models underwent training for 15 epochs with one RTX 4090 GPU. Each epoch takes about one minute to iterate through all the batches of 64 samples. For the models without pre-training, we kept the learning rate at a steady value of 10^{-4} throughout the training process. For the pretrained model, we dropped the initial learning rate of 1e-4 by a factor of 0.9 every 6 epochs. This ensured that the models were exposed to a sufficient level of detail in the data to capture intricate patterns without overwhelming the finetuning process with too much noise.

4 BASELINE METHODS

4.1 Naive Guessing

In an attempt to compare our performance against a general naive, untrained model, we adopt the Naive Guessing baseline from EEGEyeNet, where the mean position from the training set is used as the constant prediction for every test data point. According to the original EEGEyeNet paper, this yields a RMSE distance of 123.3 mm, a standard we aim to improve upon.

4.2 Machine Learning Methods

Machine Learning methods tried in the original EEGEyeNet study did not achieve significant improvements over the naive baseline when estimating absolute position [20]. In this study, we re-evaluated these methods in the following sections. Details regarding these models can be found in Appendix.

4.3 Deep Learning Methods

In this study, we focus on a comparison of the proposed EEGViT with state-of-the-art CNN-based architectures, as they are widely recognized for their effectiveness in handling EEG data and have demonstrated state of the art performance on the EEGEyeNet benchmark.

Convolutional Neural Network (CNN) We implemented a standard one-dimensional CNN, applying max pooling in line with EEGEyeNet.

PyramidalCNN The PyramidalCNN model leverages varying time granularity in the CNN architecture to capture temporal features at multiple scales [19].

EEGNet EEGNet, an EEG-specific CNN architecture that employs depthwise and separable convolutions, was designed for versatility across Brain-Computer Interface (BCI) paradigms [22].

InceptionTime InceptionTime, a scalable Time Series Classification (TSC) model, employs an ensemble of deep CNNs engineered for TSC tasks [13].

Xception Xception, an efficient CNN architecture employing depthwise separable convolutions, was designed to capture spatial correlations while enabling faster training [5].

Model	Absolute Position RMSE (mm)
Naive Guessing	123.3 ± 0.0
KNN	119.7 ± 0
RBF SVR	123 ± 0
Linear Regression	118.3 ± 0
Ridge Regression	118.2 ± 0
Lasso Regression	118 ± 0
Elastic Net	118.1 ± 0
Random Forest	116.7 ± 0.1
Gradient Boost	117 ± 0.1
AdaBoost	119.4 ± 0.1
XGBoost	118 ± 0
CNN	70.4 ± 1.1
PyramidalCNN	73.9 ± 1.9
EEGNet	81.3 ± 1.0
InceptionTime	70.7 ± 0.8
Xception	78.7 ± 1.6
ViT-Base	61.5 ± 0.6
ViT-Base Pre-trained	58.1 ± 0.6
EEGViT	61.7 ± 0.6
EEGViT Pre-trained	**55.4 ± 0.2**

Table 4: Comparison of Root Mean Squared Error (RMSE) loss in millimeters for different models on the Absolute Position Task. Original error is in pixels, and we convert it into millimeters by 2 pixels/mm for better interpretation. Lower RMSE values indicate better performance as they represent closer estimations to the actual values. The values represent the mean and standard deviation of 5 runs.

5 RESULTS

In this study, we set out to evaluate and compare the performance of an array of models for an EEG regression task. Our approach included an assessment of traditional Machine Learning models, Deep Neural Networks, as well as variants of ViT models, both with and without pretraining and the Two-Step convolution block. We gauged the models' performance using the root mean square error (RMSE) on a dedicated test set. Our comprehensive findings, which also encompass the naive baseline, are outlined in Table 4.

5.1 Baseline Regression Models

The outcomes from these methods establish a baseline that outperforms the naive baseline considerably. However, it is evident that there is ample scope for enhancement. The best-performing model of this group, the Convolutional Neural Network (CNN), could only achieve an average RMSE of 70.4 mm, indicating potential room for improvement.

5.2 Transformer Models

Both the basic versions of ViT-Base and our EEGViT demonstrated a similar level of performance, registering a significant 12% improvement over the preceding state-of-the-art models. This marked
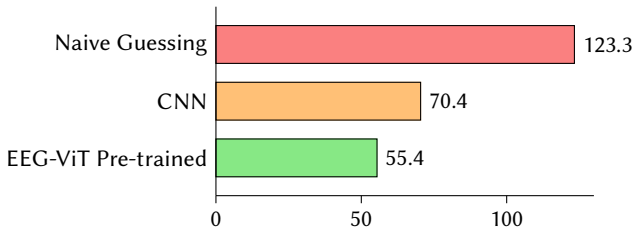
Figure 4: Comparison of Root Mean Squared Error (RMSE) loss in millimeters for random guessing, previous state of the art (CNN) and our novel method (EEG-ViT Pre-trained). Lower is better.

improvement underlines the utility of the Attention mechanism in capturing the temporal intricacies in EEG data.

5.3 Pretrained Transformer Models

The pretrained versions of both ViT-Base and EEGViT exceeded the performance of the previous models by significantly enhancing accuracy. More specifically, EEGViT demonstrated an exceptional performance boost of 10% compared to the non-pretrained ViT, and an impressive advancement of about 21% over the preceding state-of-the-art model, as shown in Figure 4.

Interestingly, even though ViT-Base underwent an identical pretraining process as EEGViT, it did not showcase the same level of prowess. ViT-Base's loss on the test set was roughly 5% higher than EEGViT's, which deviates from their comparative performance when trained from scratch. This disparity can likely be attributed to the positive impact of pretraining.

When the model can draw from a substantial knowledge base, the advantages of enhanced representations become more noticeable. Consequently, the Two-Step convolution representations in our model likely lead to better leverage within the Attention mechanism and Transformer architecture. However, this difference in performance highlights the need for further research into optimizing the training process and leveraging domain-specific knowledge in EEG signal processing.

6 DISCUSSION

Our EEGViT model significantly outperforms conventional regression models, Convolutional Neural Networks (CNNs), and the baseline ViT-Base, establishing the utility of transfer learning and the enhancement potential of image data for EEG analysis.

Despite incorporating convolutional layers, our EEGViT model's superior performance is primarily attributable to the Transformer block. This is highlighted when compared with baseline models that do not use a Transformer block, such as EEGNet and Xception, which also employ depthwise convolutions in their early layers, yet underperform relative to EEGViT.

Our findings accentuate the significance of reduced inductive bias in transitioning from Convolutional Neural Networks (CNNs) to Transformer-based architectures for processing extensive datasets. This echoes recent studies within the Computer Vision field, suggesting that the inherent strong inductive bias of CNNs may impede their efficiency, particularly with large-scale data, thereby making

Transformers a preferable alternative [10, 12]. Our investigation mirrors this, demonstrating that an unmodified Vision Transformer (ViT) exceeds CNN performance when trained exclusively on the EEGEyeNet dataset, a trend that is amplified with pretraining. Consequently, future research might explore strategies to mitigate the intrinsic inductive bias of prediction models, especially in data-rich environments.

However, there is a compelling need to make these models more interpretable. Future work should consider transforming our model into an interpretable version. For instance, visualization techniques could be incorporated to understand the activation patterns and correlations in the data.

Our EEGViT model holds promise for a wide range of EEG datasets. The EEGEyeNet dataset, upon which this study is based, is an extensive compilation of synchronized EEG and eye-tracking data. Given the very nature of the input data it learns, we strongly believe that the methodology presented in this paper can be applied to other, similar EEG datasets.

The practical value of our findings lies in the model's ability to supplement scarce EEG data with readily available image data, thereby reducing ethical and financial barriers to EEG data collection and analysis. This approach could streamline research in the field, facilitating robust, scalable EEG studies.

Additionally, our study highlights the benefits of pre-trained models in the EEG domain, traditionally hampered by data scarcity and complexity. The demonstrated superior performance of the pre-trained EEGViT model suggests that the prior knowledge learned from vast image datasets can effectively help when handling EEG data complexity. Future work may include exploring pretraining transformers for EEG data using other modalities of data. By providing an open-source version of our code[1], we encourage further exploration and enhancements to our approach.

Nevertheless, we recognize the limitations of our study, particularly the use of a relatively small Vision Transformer due to resource limitations. Future research with increased resources could potentially scale up the model by a factor of 10 or more, potentially leading to further performance gains. It is critical to balance model complexity, computational resources, and performance gains as we progress in this research direction.

7 CONCLUSION

In conclusion, the implementation of a Hybrid Vision Transformer (ViT) pretrained on ImageNet in an EEG regression task demonstrates the potential of Transformer-based transfer learning in neuroscience. The ViT, initially created for visual tasks, notably outperformed state-of-the-art convolution-based models in EEG data analysis.

These findings challenge the prevalent use of convolution in EEG neural networks and imply that cross-domain pretrained Transformer models may serve as valuable priors for EEG tasks with minor tweaks. This insight, especially relevant where data collection is difficult, suggests that repurposing available image data for EEG can enrich future research in computer science and neuroscience.

---

[1]Code is available at: https://github.com/ruiqiRichard/EEGEyeNet-vit.

REFERENCES

[1] Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. 2021. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Computing and Applications* (2021), 1–42.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.

[3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. 2019. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3286–3295.

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 0–0.

[5] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

[6] Alexander Craik, Yongtian He, and Jose L. Contreras-Vidal. 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering* 16, 3 (2019), 031001.

[7] Guanghai Dai, Jun Zhou, Jiahui Huang, and Ning Wang. 2020. HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification. *Journal of neural engineering* 17, 1 (2020), 016025.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Guangyao Dou, Zheng Zhou, and Xiaodong Qu. 2022. Time Majority Voting, a PC-Based EEG Classifier for Non-expert Users. In *HCI International 2022-Late Breaking Papers. Multimodality in Advanced Interaction Environments: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*. Springer, 415–428.

[12] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*. PMLR, 2286–2296.

[13] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34, 6 (2020), 1936–1962.

[14] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[15] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021. Transformer in transformer. *Advances in Neural Information Processing Systems* 34 (2021), 15908–15919.

[16] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data* 5, 1 (2018), 1–13.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

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

[19] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 562–570.

[20] Ard Kastrati, Martyna Beata Płomecka, Damián Pascual, Lukas Wolf, Victor Gillioz, Roger Wattenhofer, and Nicolas Langer. 2021. EEGEyeNet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. *arXiv preprint arXiv:2111.05100* (2021).

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.

[22] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[25] Yingzhou Lu, Huazheng Wang, and Wenqi Wei. 2023. Machine Learning for Synthetic Data Generation: a Review. *arXiv preprint arXiv:2302.04062* (2023).

[26] Nathan Koome Murungi, Michael Vinh Pham, Xufeng Dai, and Xiaodong Qu. 2023. Trends in Machine Learning and Electroencephalogram (EEG): A Review for Undergraduate Researchers. *arXiv preprint arXiv:2307.02819* (2023).

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[28] Xiaodong Qu, Peiyan Liu, Zhaonan Li, and Timothy Hickey. 2020. Multi-class time continuity voting for EEG classification. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*. Springer, 24–33.

[29] Xiaodong Qu, Saran Liukasemsarn, Jingxuan Tu, Amy Higgins, Timothy J Hickey, and Mei-Hua Hall. 2020. Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on EEG patterns. *Frontiers in psychiatry* 11 (2020), 541659.

[30] Xiaodong Qu, Qingtian Mei, Peiyan Liu, and Timothy Hickey. 2020. Using EEG to distinguish between writing and typing for the same cognitive task. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*. Springer, 66–74.

[31] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. 2019. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering* 16, 5 (2019), 051001.

[32] Yang Tang, Shuang Song, Shengxi Gui, Weilun Chao, Chinmin Cheng, and Rongjun Qin. 2023. Active and Low-Cost Hyperspectral Imaging for the Spectral Analysis of a Low-Light Environment. *Sensors* 23, 3 (2023), 1437.

[33] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. 2020. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing* (2020).

[34] Michal Teplan. 2002. Fundamentals of EEG measurement. *Measurement science review* 2, 2 (2002), 1–11.

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*. PMLR, 10347–10357.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 568–578.

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[40] Guowen Xiao, Meng Shi, Mengwen Ye, Bowen Xu, Zhendi Chen, and Quansheng Ren. 2022. 4D attention-based neural network for EEG emotion recognition. *Cognitive Neurodynamics* (2022), 1–14.

[41] Long Yi and Xiaodong Qu. 2022. Attention-Based CNN Capturing EEG Recording's Average Voltage and Local Change. In *International Conference on Human-Computer Interaction*. Springer, 448–459.

[42] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*. 558–567.

[43] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. 2020. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 528–543.

[44] Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. 2018. Epileptic seizure detection based on EEG signals and CNN. *Frontiers in neuroinformatics* 12 (2018), 95.

A   APPENDIX

A.1   Baseline Machine Learning Models

KNN K-Nearest Neighbors (KNN), a non-parametric method for regression, was implemented with a leaf size of 50. The selection of the leaf size aims to balance computational cost and prediction accuracy.

RBF SVC/SVR SVMs with Radial Basis Function (RBF) kernels were used, tuned with a regularization parameter C of 1, tolerance of 1e-05 for stopping criteria, and a gamma of 0.01. These parameters were chosen to balance the trade-off between model complexity and overfitting.

Linear Regression Linear Regression, a fundamental statistical method, was employed with its default parameters, providing a benchmark against more complex models.

Lasso Regression Lasso Regression uses shrinkage to limit the complexity of the model and prevent overfitting. We adopted the default parameters for the Absolute Position task, given the absence of task-specific parameters.

Elastic Net Elastic Net, a blend of Ridge Regression and Lasso Regression, was implemented with an alpha of 1, L1 ratio of 0.6, tolerance of 1e-05, and a gamma of 0.01. These parameters were chosen to optimize the balance between Ridge and Lasso penalties.

Random Forest The Random Forest ensemble learning method was deployed with a maximum depth of 50 and 250 estimators to balance the trade-off between model complexity and computational feasibility.

AdaBoost AdaBoost, an ensemble method that enhances the performance of base learners, was employed with a learning rate of 0.01 and 50 estimators. These parameters aim to avoid overfitting while maintaining a reasonable learning rate.

XGBoost XGBoost, an efficient gradient boosting library, was applied with a learning rate (eta) of 0.1, 250 estimators, and a maximum depth of 10. These settings were chosen to balance training speed and model performance.

Hyperparameters can be found in [20] and EEGEyeNet GitHub repository.

A.2   Baseline Deep Learning Model Hyperparameters

We reused the configurations and hyperparameters suggested by EEGEyeNet benchmark (see Table 5).

A.3   Guide for reproducibility

We make our code publicly available at our GitHub repository to facilitate further research. Details on the EEGEyeNet dataset and benchmark can be found in [20] and their GitHub repository. All training data used for experimentation is publicly available on OSF.io. Further information on the implementation and reproducibility of our work can be found in our GitHub repository.

Parameter	CNN	PyramidalCNN	EEGNet	InceptionTime	Xception
Depth	12	6	2	12	18
Number of filters	16	16	16/256	16	64
Kernel size	64	16	256	64	40
Batch size	64	64	64	64	64
Epochs	50	50	50	50	50
Early stopping patience	20	20	20	20	20
Residual connections	True	False	False	True	True
Bottleneck size	-	-	-	16	-
Dropout rate	-	-	0.5	-	-

Table 5: Hyperparameters of baseline deep learning models, suggested by EEGEyeNet [20].