Feature Estimation of Global Language Processing in EEG Using Attention Maps

Dai Shimizu[1], Ko Watanabe[2,3], and Andreas Dengel[2,3]

[1] Tokyo Institute of Technology, 2-12-1 Ookayama, 152-8550, Tokyo, Japan
shimizu.d.ab@m.titech.ac.jp
[2] RPTU Kaiserslautern-Landau, Erwin-Schrödinger-Straße 52, 67663, Kaiserslautern, Germany
[3] DFKI GmbH, Trippstadter Str. 122, 67663, Kaiserslautern, Germany
{first.last}@dfki.de

Abstract. Understanding the correlation between EEG features and cognitive tasks is crucial for elucidating brain function. Brain activity synchronizes during speaking and listening tasks. However, it is challenging to estimate task-dependent brain activity characteristics with methods with low spatial resolution but high temporal resolution, such as EEG, rather than methods with high spatial resolution, like fMRI. This study introduces a novel approach to EEG feature estimation that utilizes the weights of deep learning models to explore this association. We demonstrate that attention maps generated from Vision Transformers and EEGNet effectively identify features that align with findings from prior studies. EEGNet emerged as the most accurate model regarding subject independence and the classification of Listening and Speaking tasks. The application of Mel-Spectrogram with ViTs enhances the resolution of temporal and frequency-related EEG characteristics. Our findings reveal that the characteristics discerned through attention maps vary significantly based on the input data, allowing for tailored feature extraction from EEG signals. By estimating features, our study reinforces known attributes and predicts new ones, potentially offering fresh perspectives in utilizing EEG for medical purposes, such as early disease detection. These techniques will make substantial contributions to cognitive neuroscience.

Keywords: EEG · Vision Transformer · Language Processing

1 Introduction

In recent years, electroencephalography (EEG) has emerged as a critical instrument for real-time monitoring of brain activity, owing to its superior temporal resolution and non-invasive nature. However, analyzing EEG data remains a complex task due to inter-individual variability and the subtlety of neural signals.

Traditional signal processing techniques often find it challenging to handle the complexity of EEG data and effectively extract task-specific features. Deep

learning models, especially those equipped with attention mechanisms like the Transformer [45] and Vision Transformer(ViT) [14], have proven to be powerful tools in various domains, including image identification, natural language processing, and complex biological signal analysis. These models offer significant improvements over conventional methods by highlighting relevant features in extensive datasets. The computation of attention mechanism weights, as in Gradient-weighted Class Activation Mapping (Grad-CAM) [40] or Vision Transformer for Attention Map, can identify areas of interest for classification results. Interestingly, these areas can also be computed from the classification results themselves.

This study's novelty lies in using such neural network models, specifically the Vision Transformer, to estimate features from the attention map. These features are not specific to EEG but are rough, task-dependent features. This approach allows us to implement subject-independent analysis and use the entire language-related area [7,20] of the brain for training dataset creation. Interestingly, during speaking and listening activities, brain activity becomes similar [22, 26, 36]. In this context, we aim to estimate the features of speaking and listening using the weights of the neural network model. By leveraging the internal weights of these models to compute attention maps, this study aims to uncover subtle EEG signal patterns indicative of specific brain functions.

The contributions of this study are twofold:

C1 The utility of estimating EEG features using ViT specifically focusing on EEG-based language processing

C2 A comprehensive evaluation of auditory information with participant independent in EEG using ViT

These findings will revolutionize EEG data interpretation, enhancing diagnostic capabilities and personalizing neurotherapeutic approaches. This work is expected to make significant contributions to the fields of neuroimaging and cognitive neuroscience.

2 Related Work

Deep learning techniques have significantly revolutionized the field of EEG analysis. This section first discusses the classification of EEG signals using various methods such as Power Spectral Density (PSD) [2], EEGNet [24], and other CNN models [12]. It then delves into applying Grad-CAM with CNNs and attention maps with Vision Transformers for feature extraction and interpretation in EEG signals.

In the realm of EEG signal classification, several techniques have been employed. PSD and other methods related to EEG frequency have been utilized in the context of emotion recognition, where these were extracted from EEG recorded during a listening task, revealing certain relevant frequency bands [4, 33, 35]. CNN-based architectures, particularly EEGNet [24], have been tailored

for EEG signal processing, enabling automatic feature extraction and classification across various EEG analysis applications [43]. Other CNN models using Net structures have also been widely adopted for EEG classification [29].

Grad-CAM [40] is a technique that enhances interpret-ability in models based on CNNs [46]. It highlights the critical regions within the input that influence the classification outcomes, thereby making the decision-making processes of CNNs transparent and comprehensible. This method has been instrumental in elucidating how CNNs prioritize different regions in an input image or signal during classification tasks. The combination of EEGNet with Grad-CAM has been used to select the most suitable electrodes' channel [27]. Moreover, the Grad-CAM technique in EEGNet was used to determine which brain area was involved in intention [25, 34].

ViT [14] has been applied to EEG studies [5,10,17], marking a significant shift in image classification. ViTs use attention maps to illustrate how different image parts influence classification, providing insights into decision-making processes. In EEG studies, these maps reveal brain region activations during cognitive tasks, enriching our understanding of brain function. Extensive studies have focused on delineating specific brain regions involved in auditory information processing [7]. Techniques, including EEG, have been pivotal in activating and studying various cerebral regions in response to auditory stimuli. Insights from this research are critical for comprehending auditory system functions and have profound implications for diagnosing and treating auditory-related disorders. The application of advanced deep learning techniques such as Grad-CAM and Vision Transformers has markedly enriched EEG analysis.

These methodologies boost the analytical capabilities and enhance the interpretability of EEG-based models, paving the way for significant neurological discoveries. Ongoing and future studies are expected to further harness the potential of these innovative techniques in complex EEG signal analysis.

3 Methodology

In this study, our primary objective is to investigate whether the attention mechanisms in neural network models can capture the characteristics of brain waves depending on the task. Specifically, we aim to compare the brain waves recorded while listening to speech and while speaking the same speech heard. By examining these two conditions, we seek to identify broad differences in neural activity patterns associated with auditory perception and speech production.

To achieve this, we utilize several models: a pre-trained Vision Transformer (pre-trained ViT) [15], a customized Vision Transformer (Custom ViT) [14], EEGNet [24], and a Support Vector Machine (SVM) [39]. These models classify data during the listening and speaking phases. We aim to compute which aspects each neural network focuses on by analyzing the weights of these models. Subsequent sections will describe the detailed dataset types, data processing, and methods for creating attention maps for each model.

3.1 EEG Data from OpenNEURO

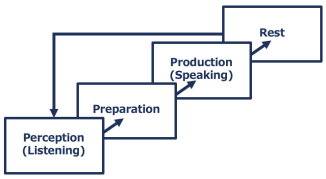


Fig. 1: Experimental protocol of the dataset. Subjects listen to and then repeat one of 30 randomly selected Spanish sentences, forming 30 perception-production pairs. Each sentence lasts approximately two seconds. Subjects perform between 360 and 420 trials, with each figure representing one trial.

This study uses a public dataset available from OpenNEURO [9] based on EEG recordings. The EEG data utilized in this study were obtained from Spanish participants. The dataset comprises 60 sessions, each recorded from 64 EEG channels and Electrocardiogram(ECG) and Electrooculogram (EOG) channels at a sample rate of 1000 Hz. This dataset was collected from 56 healthy participants.

The experimental paradigm presents participants with one of 30 different Spanish sentences, selected randomly for each trial. After listening to the sentence, participants are asked to repeat it aloud. Each sentence lasts approximately 2 seconds, and subjects perform between 360 and 420 trials. Figure 1 illustrates protocol ensures a comprehensive set of perception-production pairs for analysis.

3.2 Channel Selection for Classification

This study investigates how broad auditory information is represented in brain waves. To achieve this, we utilized multiple channels from the EEG recordings to form a comprehensive dataset. The selection of specific channels is informed by existing literature, which indicates heightened activity in the left hemisphere, particularly the left temporal lobe, during auditory tasks [19, 37, 38]. Therefore, we focused on channels located in these regions to capture relevant neural activities.

As shown in Figure 2, we specifically extracted EEG data from the following channels: 'F7', 'F5', 'F3', 'FT7', 'FC5', 'FC3', 'T7', 'C5', 'C3', 'TP7', 'CP5', 'CP3', 'P7', 'P5', and 'P3'. These channels were chosen due to their significance in language processing tasks [7, 20], which are the primary focus of our analysis. By isolating these particular channels, we aim to capture EEG features that are most indicative of the cognitive functions associated with language and auditory processing.

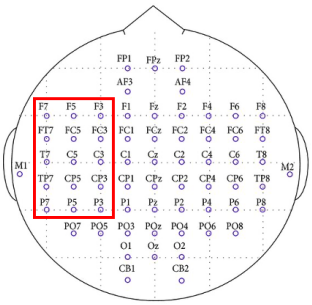


Fig. 2: Electrode placement following the 64-channel international 10–20 system. Electrodes framed in red were used.

Table 1: Datasets: ALL indicates all participants, S indicates the number of trials applied, and C indicates the number of channels used.

Model	Training Set	Validation Set	Test Set
`Custom-ViT`	$(\text{ALL}-1)\times S\times C$	$\frac{1\times S\times C}{2}$	$\frac{1\times S\times C}{2}$
`Pretrained-ViT`	$(\text{ALL}-1)\times S\times C$	$\frac{1\times S\times C}{2}$	$\frac{1\times S\times C}{2}$
`EEGNet`	$(\text{ALL}-1)\times S$	$\frac{1\times S}{2}$	$\frac{1\times S}{2}$
`SVM`	$(\text{ALL}-1)\times S\times C$	$\frac{1\times S\times C}{2}$	$\frac{1\times S\times C}{2}$

3.3 Pre-processing and Data Processing

We applied a band-pass filter to isolate frequencies from 1 to 40 Hz in EEG data, capturing the most relevant waves for cognitive and neural processes [6]. We implemented artifact removal procedures for EOG and ECG signals to enhance the clarity of neural signal interpretation [6]. We focused on language-related channels, such as Broca's and Wernicke's areas, to concentrate our analysis on the neural substrates of language function [7, 8, 20]. These preprocessing steps were uniformly applied across all computational models employed in our study to establish a consistent foundation for downstream analyses.

In this study, we leveraged Mel-spectrograms [41] to extract the spatio-temporal characteristics of EEG data for the training and evaluation of ViTs. The Mel-spectrogram transformation [13] was selected due to its effectiveness in encapsulating the dynamic changes in EEG signal power across both time and frequency domains [1], which is essential for our models to learn the intricate patterns associated with different cognitive states. For the EEGNet architecture, which requires input in Channel and Time series, we downsampled the EEG data from 1000 Hz to 125 Hz. This preprocessing step was implemented to align with the Nyquist criterion [42], ensuring the capture of all pertinent information below the 40 Hz frequency threshold, which encompasses the delta, theta, alpha, and beta wavebands known to be most relevant for brain-computer interface applications. Finally, for the SVM classifier, we utilized PSD [4, 35] estimates as the dataset to effectively reduce the dimensionality of the EEG signals. By

transforming the data into the power frequency domain, we aim to highlight the most discriminative features for classification while simultaneously reducing computational complexity and enhancing model interpretability.

Additionally, a leave-one-subject-out (subject independence) approach was employed for all models to prevent the overlap of participant data during model training, ensuring that the training sets were participant-independent [3]. Table 1 shows the data split.

3.4 Models Used for Data Analysis

Our study adopted the Custom ViT to generate attention maps that span both temporal and frequency domains. The Custom ViT utilized the same structure described in the original paper [14] and we also integrated a pre-trained ViT [15], utilizing its pre-trained weights and the same architecture to explore the interpretative capabilities of a network trained on extensive datasets with the following specific configurations:

- pre-trained ViT: 12 layers, 12 heads, 16 patch size, 14x14 patches, 224x224 input, and 1024 MLP dimensions.
- Custom ViT: 3 layers, 4 heads, 4 patch sizes, 8x8 patches, 32x32 input, and 256 MLP dimensions.

For Custom ViT, all layers were considered for training, and for pre-trained ViT, only the final layer of Linear was considered for training.

For baseline comparisons, we employed EEGNet and SVM as standard models for EEG classification. EEGNet enables the extraction of temporal attention maps by applying Grad-CAM on the convolutional weights of the final layer.

During the model training phase, we employed a subject-wise cross-validation approach [23]. This involved using the data from a single subject as the validation and test set, while the remaining subjects' data constituted the training set. Such a strategy ensures that the model learns to generalize features of EEG data across different tasks and individuals, rather than overfitting to the characteristics of a single subject's data. This methodological choice is pivotal for developing robust EEG-based models that can reliably perform across diverse population samples, thereby enhancing the universality and applicability of the findings.

3.5 Attention Maps

This section describes the methodology employed to compute the attention maps for each model used in this study. Attention maps were utilized to explain the classification decisions made by the models, highlighting the features that contributed most significantly to their predictions. For the Custom ViT and the pre-trained ViT, attention maps were derived from the weights of the final layer. This involved extracting the attention weights corresponding to the most significant parts of the input data, as identified by the model during classification.

Specifically, the following steps were performed to compute the attention maps for the Vision Transformers:

1. The EEG data were transformed into mel-spectrograms, which were then fed into the ViTs.
2. The attention weights from the final layer were extracted, representing the importance of different time-frequency regions in the input data.
3. These weights were visualized to create the attention maps, illustrating the areas the model focused on during classification.

For EEGNet, we used Grad-CAM to generate attention maps. Grad-CAM provides a visual explanation by highlighting the regions of the input that are most influential for the model's prediction. The following steps outline the process:

1. EEG data were input into the EEGNet model, which processes them through its convolutional layers.
2. Grad-CAM was applied to the convolutional weights of the final layer, identifying the most critical features for classification.
3. The resulting attention maps display the temporal regions and channels that contributed most to the model's decision.

To ensure the reliability of the attention maps, we calculated them from the top 10 participants with the highest classification accuracy, denoted as '@10' in Table 2. This selection criterion helps focus on the dataset's most informative and consistent patterns. All attention maps were derived from the final layer of the models [47]. By utilizing attention maps from both ViTs and EEGNet, we aim to gain insights into the classification criteria used by each model, providing a clearer understanding of how neural network models interpret EEG data for task-related cognitive processes.

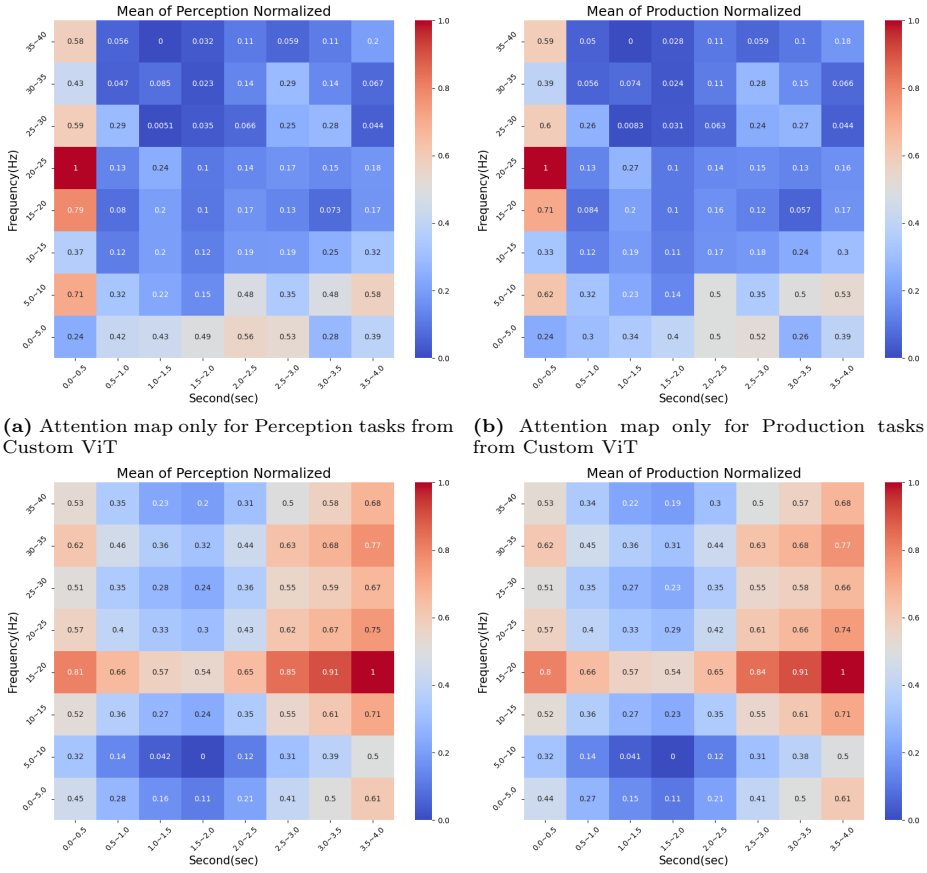
3.6 Software and Tools

All processing tasks, excluding data collection, were executed using Python. We utilized the PyTorch framework for deep learning algorithms, which is renowned for its flexibility and efficiency in building complex neural network architectures. EEG signal processing was conducted using the MNE library [18], which is specifically designed for advanced electrophysiological data analysis and provides robust tools for EEG data manipulation and visualization.

4 Result

4.1 Accuracy of Classification

As presented in Table 2, EEGNet attained the highest classification accuracies among all models tested, recording values of 0.7248 for all participants and 0.8433 for the top 10 participants. While the ViTs, both Custom ViT and pre-trained ViT, did not achieve the highest overall accuracies, Custom ViT was notably the second most accurate model in the binary classification task: listening versus speaking across all participants.



(a) Attention map only for Perception tasks from Custom ViT

(b) Attention map only for Production tasks from Custom ViT

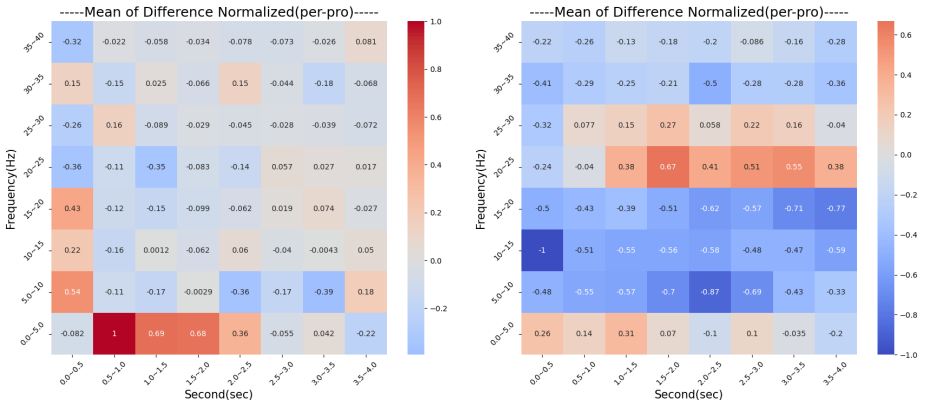
(c) Attention map only for Perception tasks from pre-trained ViT

(d) Attention map only for Production tasks from pre-trained ViT

Fig. 3: Attention Maps of the models during classification. Lower values (indicated by blue) represent regions where the models allocate less attention, whereas higher values, indicated by red, signify areas of focused attention. The x-axis represents the time series from 0 to 4 seconds, and the y-axis represents the frequency series from 0 to 40 Hz. Normalized attention maps, averaged from data collected during (a) the perception task (listening) using the Custom ViT, (b) the production task (speaking) using the Custom ViT, (c) the perception task using the pre-trained ViT, and (d) the production task using the pre-trained ViT.

Table 2: Comparison of model accuracies for classification employed a leave-one-participant-out cross-validation approach, with one participant left out in each fold

Model	Accuracy	Accuracy@10
`SVM`	0.5884	0.7048
`Custom ViT`	0.6153	0.6704
`pre-trained ViT`	0.5633	0.6222
`EEGNet`	**0.7248**	**0.8433**



(a) Attention map for Whole tasks from Custom ViT

(b) Attention map for Whole tasks from pre-trained ViT

Fig. 4: Contrasts of attention maps between production and perception tasks. Lower values (blue) indicate greater attention during production tasks, whereas higher values (red) highlight areas of intensified focus during perception tasks. Both axes are consistent with those in Figure 3. Normalized attention maps are obtained by calculating the difference between the Perception and Production attention maps from (a) the Custom ViT and (b) the pre-trained ViT.

4.2 Attention Maps

Figure 3 highlights that ViTs predominantly focus on the initial stages of the task. The Custom ViT, shown in Figures 3a and 3b, consistently emphasizes the delta and theta bands throughout the task duration, reflecting its sensitivity to lower frequency ranges. In contrast, the pre-trained ViT, depicted in Figures 3c and 3d, exhibits a marked preference for beta waves, particularly the high beta frequencies, and additionally shifts its attention significantly towards the task's conclusion.

Further nuances in the attention distribution are evident from the comparative analyses presented in Figure 4. The Custom ViT shows the greatest variance between perception and production tasks within the delta band: 0.5 to 4.0Hz, with perceptual tasks showing increased activity in the beta band: 16.5 to 20.0Hz, low beta band: 12.5 to 16Hz, alpha band (8.0 to 12.0Hz), and theta band: 4.0 to

7.0Hz, while production tasks predominantly engage the high beta band: 20.5 to 28Hz and gamma band: over 30Hz. This indicates a complex interplay of frequency bands depending on the cognitive demands of the task, as illustrated in Figure 4a.

Conversely, the pre-trained ViT demonstrates the largest disparities in the alpha band when contrasting perception and production tasks. During perceptual tasks, there is a notable increase in beta band activity, whereas production tasks see heightened activity in the theta, alpha, low beta, and gamma bands. These findings, presented in Figure 4b, suggest the differential engagement of task-dependent brain rhythms, highlighting the adaptability of neural network models to varying cognitive requirements.

These activity patterns underscore the intricate relationship between task-specific cognitive processes and neural focus, as represented by frequency band engagement. The attention maps, particularly those derived from the Custom ViT and pre-trained ViT, validate the hypothesis that neural networks can adaptively highlight relevant EEG features that signify distinct cognitive states associated with specific tasks.

4.3 Grad-CAM

The attention maps derived from EEGNet via Grad-CAM analysis, as shown in Figure 5, reveal distinct patterns of focus depending on the task and timing. Specifically, when analyzing EEG data associated with perception tasks, the model predominantly concentrates on the initial phase of the task. However, a dominant shift in attention occurs between 2.5 and 3.0 seconds, indicating a temporal transition in neural engagement. This shift suggests that the model identifies critical periods of neural activity that correspond to key moments in the cognitive process, highlighting the dynamic nature of brain function during these tasks.

4.4 Validity of Usage of ViT for Feature Estimation

We further evaluated the EEG data from the previous Grad-CAM and ViT analyses by visualizing the differences between perception and production tasks. This was done by subtracting the EEG signals during the production phase from those during the perception phase, which is the object of classification.

Figure 6 shows the results of bandpass filtering the EEG data between 1-40 Hz, capturing the broad range of cognitive and neural processes. Each trial is aligned to start at 0 seconds, with an average of 0.2 seconds before and 4 seconds after the task onset across all participants. Significant amplitude differences are observed up to 0.5 seconds after the task starts, indicating early task-specific neural engagement as shown in Figure 4b and Figure 5.

To focus on specific frequency bands and validate the attention map's findings, EEG data from 1 to 5 Hz, including the delta and theta bands, were visualized in Figure 4 and Figure 7. Each trial is similarly aligned to start at 0 seconds, with 0.2 seconds before and 4 seconds after the task onset averaged

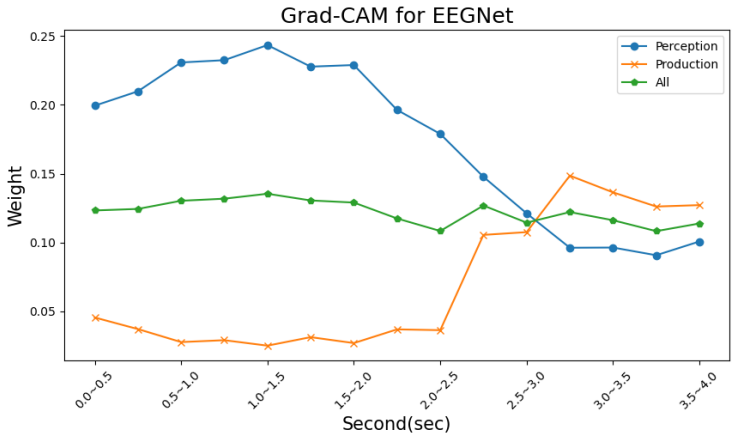


Fig. 5: The features extracted from the final layer of EEGNet using Grad-CAM.

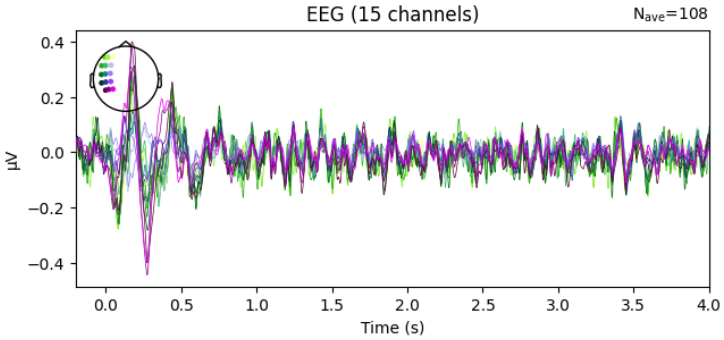


Fig. 6: Average EEG data across all frequency bands: 1 to 40 Hz, highlighting differences between perception and production tasks. Each trial is aligned to start at 0 seconds, with 0.2 seconds before and 4 seconds after the task onset.

across all participants. Differences between perception and production tasks are evident throughout the task duration, underscoring the distinct neural dynamics associated with these cognitive processes.

One potential reason for the early differences observed is the event-related potential (ERP), which captures time-locked neural responses to specific sensory, cognitive, or motor events [31]. ERPs provide a precise temporal measure of brain activity, which can be effectively captured by EEG. In our study, the ViT model successfully identifies these ERPs, highlighting their significance in distinguishing between perception and production tasks. These visualizations confirm that the attention maps accurately highlight the task-specific neural activity patterns, reinforcing the models' ability to distinguish between different cognitive states based on EEG data.

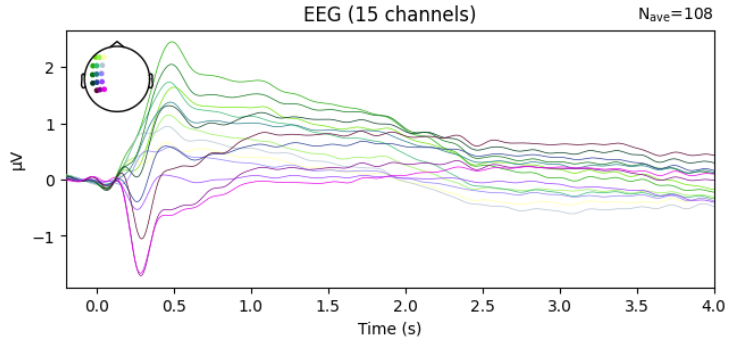


Fig. 7: Average EEG data within the low-frequency band (1-5 Hz) highlighting differences between perception and production tasks. Each trial is aligned to start at 0 seconds, with 0.2 seconds before and 4 seconds after the task onset.

5 Discussion

Our findings underscore the efficacy of neural networks, particularly ViTs, in interpreting EEG features, demonstrating their capacity to recognize EEG characteristics for listening and speaking tasks. Neural network models, especially those equipped with attentional mechanisms, excel at extracting and visualizing salient features from distinct brain activities. This proficiency is derived from their capability [32] to systematically process inputs across both time and frequency dimensions, thereby preserving the structural integrity of the data.

ViTs demonstrate a unique ability to identify specific features of EEG signals that vary with the cognitive task, enabling precise dissection of frequency and temporal information(C1). This highlights the adaptability and accuracy of ViTs in neuroscientific research, making them invaluable for tasks requiring a nuanced understanding of brain functions. Our comprehensive evaluation of auditory information in EEG using ViT (C2) further emphasizes its robustness and effectiveness in capturing task-specific neural dynamics.

However, the performance of each model varies slightly, a phenomenon primarily attributed to the large datasets utilized and the inherent EEG variability among individuals [11,44]. This variability complicates the generalization of findings across different populations and emphasizes the need for models to accommodate individual differences. These observations suggest ample opportunities for further advancements in neural network architectures, potentially enhancing their effectiveness and precision in analyzing complex biological signals.

Interestingly, our study found specific attention areas consistent with previous studies [16,21,22,28,30]. The synchronization of brain activity, the emergence of features at the start and end of tasks, and the distinctions in frequency bands were clearly illustrated by the attention maps derived from ViTs and EEGNet. These attention maps highlight critical evaluative points for task classification, focusing on distinct EEG features relevant to different cognitive states.

When results were not normalized, performance differences between the Vision Transformers for different tasks may be attributed to the model's insufficient training and the coarse granularity of the training data. While the models aimed to estimate emergent features throughout the tasks, a more detailed, channel-by-channel analysis could potentially improve accuracy.

This study confirms that neural networks can effectively leverage model weights to pinpoint specific EEG features, thereby distinguishing between different cognitive tasks (C1). This capability validates neural networks' potential to parse EEG data accurately and opens avenues for discovering new insights into EEG features. Such advancements underscore the potential of neural networks to deepen our understanding of the neural bases of cognitive tasks through sophisticated pattern recognition and feature extraction methods.

Moreover, our research suggests that a data-driven approach can reveal how EEG reflects underlying brain activity characteristics (C2). By analyzing attention maps and model weights, we can infer which aspects of neural activity are most informative for different cognitive states. This approach could lead to identifying biomarkers for specific mental processes, enhancing EEG's diagnostic and therapeutic capabilities in clinical settings.

These findings could significantly improve brain-computer interfaces (BCIs), neurofeedback systems, and other EEG-based diagnostic tools in practical medical applications. For instance, more accurate and individualized EEG analysis could lead to better detection and monitoring of neurological conditions such as epilepsy, depression, and sleep disorders. By providing a clearer understanding of the neural dynamics associated with different tasks, our study paves the way for developing more targeted and effective interventions in cognitive and neurological health.

6 Conclusion

This study highlights the potential of neural networks, specifically ViTs and EEGNet, in EEG data interpretation for cognitive task classification. These models recognize established EEG features and uncover new information crucial for understanding brain function. Both Custom ViT and pre-trained ViT demonstrate proficiency in focusing on specific temporal stages of cognitive tasks, with attention to different frequency bands (C2). EEGNet, analyzed through Grad-CAM, reveals variable attention allocation depending on the task, indicating the temporal complexity involved in processing different cognitive activities.

The attention maps generated across models are instrumental in understanding how neural networks prioritize certain features for task classification. They identify the EEG signal regions most relevant for distinguishing between cognitive states (C1). Our study validates the capability of these models in EEG data analysis and suggests that a data-driven approach can reveal significant insights into brain activity patterns (C2). This paves the way for further enhancements in neural network designs to accommodate individual variability and generalize findings across diverse populations.

In conclusion, using neural networks in EEG analysis offers a transformative approach to understanding the neural bases of cognitive tasks, providing deep insights into the temporal and frequency-related dynamics of brain activity. This research holds promise for improving diagnostic and therapeutic applications in clinical settings, potentially leading to better brain-computer interfaces and neurofeedback systems.

Acknowledgement

This work was supported by the Nakatani Foundation.

References

1. Abdul, Z.K., Al-Talabani, A.K.: Mel frequency cepstral coefficient and its applications: A review. IEEE Access **10**, 122136–122158 (2022)
2. Al-Fahoum, A.S., Al-Fraihat, A.A.: Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains. International Scholarly Research Notices **2014**(1), 730218 (2014)
3. Albawi, S., Bayat, O., Al-Azawi, S., Ucan, O.N.: Social touch gesture recognition using convolutional neural network. Computational Intelligence and Neuroscience **2018**(1), 6973103 (2018)
4. Alsolamy, M., Fattouh, A.: Emotion estimation from eeg signals during listening to quran using psd features. In: 2016 7th International Conference on computer science and information technology (CSIT). pp. 1–5. IEEE (2016)
5. Arjun, A., Rajpoot, A.S., Panicker, M.R.: Introducing attention mechanism for eeg signals: Emotion recognition with vision transformers. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 5723–5726. IEEE (2021)
6. Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.M., Robbins, K.A.: The prep pipeline: standardized preprocessing for large-scale eeg analysis. Frontiers in neuroinformatics **9**, 16 (2015)
7. Binder, J.R., Frost, J.A., Hammeke, T.A., Cox, R.W., Rao, S.M., Prieto, T.: Human brain language areas identified by functional magnetic resonance imaging. Journal of Neuroscience **17**(1), 353–362 (1997)
8. Blank, S.C., Scott, S.K., Murphy, K., Warburton, E., Wise, R.J.: Speech production: Wernicke, broca and beyond. Brain **125**(8), 1829–1838 (2002)
9. Carlos Valle, Carolina Méndez-Orellana, M.R.F., Herff, C.: Subject-independent decoding of perceived sentences from eeg signals using artificial neural networks p. 2826
10. Chen, C., Wang, H., Chen, Y., Yin, Z., Yang, X., Ning, H., Zhang, Q., Li, W., Xiao, R., Zhao, J.: Understanding the brain with attention: A survey of transformers in brain sciences. Brain-X **1**(3), e29 (2023)
11. Chowdhury, R.R., Muhammad, Y., Adeel, U.: Enhancing cross-subject motor imagery classification in eeg-based brain–computer interfaces by using multi-branch cnn. Sensors **23**(18), 7908 (2023)
12. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering **16**(3), 031001 (2019)

13. Daube, C., Ince, R.A., Gross, J.: Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Current Biology **29**(12), 1924–1937 (2019)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Face, H.: google/vit-base-patch16-224 (2023), `https://huggingface.co/google/vit-base-patch16-224`, accessed: 2024-03-15
16. Giglio, L., Ostarek, M., Sharoh, D., Hagoort, P.: Diverging neural dynamics for syntactic structure building in naturalistic speaking and listening. Proceedings of the National Academy of Sciences **121**(11), e2310766121 (2024)
17. Gong, L., Li, M., Zhang, T., Chen, W.: Eeg emotion recognition using attention-based convolutional transformer neural network. Biomedical Signal Processing and Control **84**, 104835 (2023)
18. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al.: Meg and eeg data analysis with mne-python. Frontiers in Neuroinformatics **7**, 267 (2013)
19. de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E.: The hierarchical cortical organization of human speech processing. Journal of Neuroscience **37**(27), 6539–6557 (2017)
20. Hollenstein, N., Renggli, C., Glaus, B., Barrett, M., Troendle, M., Langer, N., Zhang, C.: Decoding eeg brain activity for multi-modal natural language processing. Frontiers in Human Neuroscience **15**, 659410 (2021)
21. Kubetschek, C., Kayser, C.: Delta/theta band eeg activity shapes the rhythmic perceptual sampling of auditory scenes. Scientific Reports **11**(1), 2370 (2021)
22. Kuhlen, A.K., Allefeld, C., Haynes, J.D.: Content-specific coordination of listeners' to speakers' eeg during communication. Frontiers in human neuroscience **6**, 266 (2012)
23. Kwon, O.Y., Lee, M.H., Guan, C., Lee, S.W.: Subject-independent brain–computer interfaces based on deep convolutional neural networks. IEEE transactions on neural networks and learning systems **31**(10), 3839–3852 (2019)
24. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
25. Leong, D., Do, T.T.T., Lin, C.T.: Ventral and dorsal stream eeg channels: Key features for eeg-based object recognition and identification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 4862–4870 (2023)
26. Li, J., Hong, B., Nolte, G., Engel, A.K., Zhang, D.: Eeg-based speaker–listener neural coupling reflects speech-selective attentional mechanisms beyond the speech stimulus. Cerebral Cortex **33**(22), 11080–11091 (2023)
27. Li, Y., Yang, H., Li, J., Chen, D., Du, M.: Eeg-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-cam. Neurocomputing **415**, 225–233 (2020)
28. Lin, Y., Liu, B., Liu, Z., Gao, X.: Eeg gamma-band activity during audiovisual speech comprehension in different noise environments. Cognitive neurodynamics **9**, 389–398 (2015)
29. Liu, X., Hui, Q., Xu, S., Wang, S., Na, R., Sun, Y., Chen, X., Zheng, D.: Tacnet: task-aware electroencephalogram classification for brain-computer interface through a novel temporal attention convolutional network. In: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous

Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers. pp. 660–665 (2021)
30. Lopez-Bernal, D., Balderas, D., Ponce, P., Molina, A.: A state-of-the-art review of eeg-based imagined speech decoding. Frontiers in human neuroscience **16**, 867281 (2022)
31. Luck, S.J.: An introduction to the event-related potential technique. MIT press (2014)
32. Martínez-Cañada, P., Ness, T.V., Einevoll, G.T., Fellin, T., Panzeri, S.: Computation of the electroencephalogram (eeg) from network models of point neurons. PLOS Computational Biology **17**(4), e1008893 (2021)
33. Mihajlović, V.: Eeg spectra vs recurrence features in understanding cognitive effort. In: Proceedings of the 2019 ACM International Symposium on Wearable Computers. pp. 160–165 (2019)
34. Orima, T., Motoyoshi, I.: Spatiotemporal cortical dynamics for visual scene processing as revealed by eeg decoding. Frontiers in Neuroscience **17**, 1167719 (2023)
35. Park, Y., Luo, L., Parhi, K.K., Netoff, T.: Seizure prediction with spectral power of eeg using cost-sensitive support vector machines. Epilepsia **52**(10), 1761–1770 (2011)
36. Pérez, A., Carreiras, M., Duñabeitia, J.A.: Brain-to-brain entrainment: Eeg interbrain synchronization while speaking and listening. Scientific reports **7**(1), 4190 (2017)
37. Price, C.J.: The anatomy of language: contributions from functional neuroimaging. The Journal of Anatomy **197**(3), 335–359 (2000)
38. Price, C.J.: A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. Neuroimage **62**(2), 816–847 (2012)
39. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. vol. 3, pp. 32–36. IEEE (2004)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
41. Stevens, S.S., Volkmann, J., Newman, E.B.: A scale for the measurement of the psychological magnitude pitch. The journal of the acoustical society of america **8**(3), 185–190 (1937)
42. Sun, J.: Impedance-based stability criterion for grid-connected inverters. IEEE transactions on power electronics **26**(11), 3075–3078 (2011)
43. Sun, Y., Liu, X., Na, R., Wang, S., Zheng, D., Fan, S.: Cross-domain feature distillation framework for enhancing classification in ear-eeg brain-computer interfaces. In: Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing. pp. 706–711 (2023)
44. Thakor, N.V., Sherman, D.L.: Eeg signal processing: Theory and applications. In: Neural engineering, pp. 259–303. Springer (2012)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
46. Watanabe, K., Sathyanarayana, T., Dengel, A., Ishimaru, S.: Engauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network. IEEE Access **11**, 52886–52898 (2023). https://doi.org/10.1109/ACCESS.2023.3279428

47. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)