Shared Hierarchical Representations Explain Temporal Correspondence Between Brain Activity and Deep Neural Networks

Eric Lützow Holm[1,2]*, **Giovanni Marraffini, Diego Fernandez Slezak**[1,3,4], **Enzo Tagliazucchi**[1,2,5]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CABA, Argentina.
[2] Instituto de Física Interdisciplinaria y Aplicada, Departamento de Física, UBA, CABA, Argentina.
[3] Departamento de Computación, FCEyN, UBA, CABA, Argentina.
[4] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET–UBA, CABA, Argentina.
[5] Latin American Brain Health (BrainLat), Universidad Adolfo Ibáñez, Santiago, Chile.

Abstract

The visual cortex and artificial neural networks both process images hierarchically, progressing from low-level features to high-level semantic representations. We investigated the temporal correspondence between activations from multiple convolutional and transformer-based neural network models (AlexNet, MoCo, ResNet-50, VGG-19, and ViT) and human EEG responses recorded during visual perception tasks. Leveraging two EEG datasets of images presented at different durations, we assessed whether this correspondence reflects general architectural principles or model-specific computations. Our analysis revealed a robust mapping: early EEG components correlated with activations from initial network layers and low-level visual features, whereas later components aligned with deeper layers and semantic content. Moreover, for images presented during longer times, the extent of correspondence correlated with the semantic contribution to the EEG response. These findings highlight a consistent temporal alignment between biological and artificial vision, suggesting that this correspondence is primarily driven by the hierarchical transformation of visual to semantic representations rather than by idiosyncratic computational features of individual network architectures.

1 Introduction

Humans interact with a wide variety of objects in daily life, recognizing and generalizing complex sensory patterns. Object recognition relies on hierarchical neural processing in the ventral visual pathway, where neurons progressively transform low-level visual input into abstract semantic representations [1, 2]. The neural dynamics underlying this transformation have been studied using non-invasive techniques such as electroencephalography (EEG) and magnetoencephalography (MEG) [3]. Early evoked neural activity primarily reflects low-level image features—such as luminance, contrast, orientation, and color—while later components encode object identity and conceptual content, indicating a neural semantic representation [4].

Artificial neural networks (ANNs) that achieve high performance on object recognition tasks exhibit a similar hierarchical organization, with multiple layers transforming visual features at increasing levels of abstraction [5]. While the specific computations differ across models, human neuroimaging functional magnetic resonance imaging (fMRI) studies have revealed anatomical correspondences between biological and artificial representations during object recognition [6, 7, 8, 9, 10, 11]. Similar

*Correspondence to: elholm90 at gmail dot com

Preprint.

results have been obtained in monkey electrophysiological studies [12, 9]. These correspondences also extend to the temporal domain: early brain responses (measured via EEG/MEG) are best predicted by early ANN layers, while late responses align with deeper layers [13, 14, 15].

The representational similarity between brain and ANN representations depends on multiple factors, including training data, classification task and performance [6]. Importantly, this similarity does not imply that artificial systems mimic the brain's computational mechanisms. Human fMRI and macaque electrophysiology studies show that widely different ANN models are in correspondence with brain activity along the ventral pathway, regardless of their biological interpretability [16]. For example, convolutional neural networks (CNNs) share several similarities with cortical processing, with a hierarchy of neurons presenting increasingly large receptive fields [17, 2]. Other models, such as Vision Transformers (ViTs) and multilayer perceptrons (MLPs), employ computations with a less clear neurobiological interpretation. However, both have been equally supported as plausible models to predict fMRI activity during object perception and recognition [16, 6].

We tested the hypothesis that brain-ANN temporal correspondence reflects two aspects commonly found in artificial vision models: the hierarchical organization of DNNs [18] and their transformation of low-level visual features to semantic representations [19]. Using representational similarity analysis (RSA) [20], we compared EEG-based decoding similarities with layer-specific representational similarities extracted from five object-recognition models (AlexNet [21], MoCo [22], ResNet-50 [23], VGG-19 [24], and ViT [25]). Leveraging two independent EEG datasets allowed us to investigate the effect of stimulus duration (50 ms vs. 150 ms) on brain-ANN temporal correspondence. Finally, we investigated how this correspondence depended on the contribution of visual and semantic content to ANN activations and EEG responses.

2 Related Work

Brain-ANN Temporal Correspondence. Extensive fMRI work revealed a hierarchical correspondence between multiple ANNs and anatomical areas involved in vision and object recognition [8, 9, 10, 11, 6, 16]. CNNs capture the dynamics of human visual processing along the dorsal and ventral streams, as revealed using MEG [13, 15]. Deeper layers of the CNN predict late EEG responses and vice-versa for early layers [14]. However, being limited to a single ANN model (CNNs) limits the perspective of these studies on how model-specific computations contribute to this correspondence.

Contributions of Low-level Visual Features. The categorical organization of the ventral visual stream can be partially explained by the low-level image properties of different object categories [26]. Luminance and contrast alone suffice to categorize images into abstract concepts with significant accuracy [27]. Decoding of visual information from EEG/MEG signals also depends on visual regularities in the stimuli: low-level visual features dominate object decodability early after stimulus onset (50 ms), with more abstract semantic representations appearing at later times (>150 ms) [4]. However, it remains to be addressed whether brain-ANN temporal correspondence is driven by their shared transformation of visual to semantic representation.

3 Methods

3.1 Datasets

We analyzed two versions of the THINGS-EEG dataset, consisting of EEG data recorded during visual perception (Fig. 1A). The THINGS-EEG 1 dataset includes EEG data from 50 participants (see A.2 for further details) presented with images spanning 1,854 concepts, each presented for 50 ms in a rapid serial visualization paradigm (RSVP) Fig. 1B) [28]. The THINGS-EEG 2 dataset includes 10 participants presented with images spanning 1,654 concepts, each presented for 100 ms with 100 ms of interstimuli interval [29]. From both datasets, we considered a subset of 200 unique images, each representing a distinctive concept.

3.2 Data Pre-processing

Pre-processing of the recorded EEG signals (Fig. 1C) included a Hamming windowed finite impulse response filter (0.1 Hz high-pass and 100 Hz low-pass), re-referencing to the average reference, and

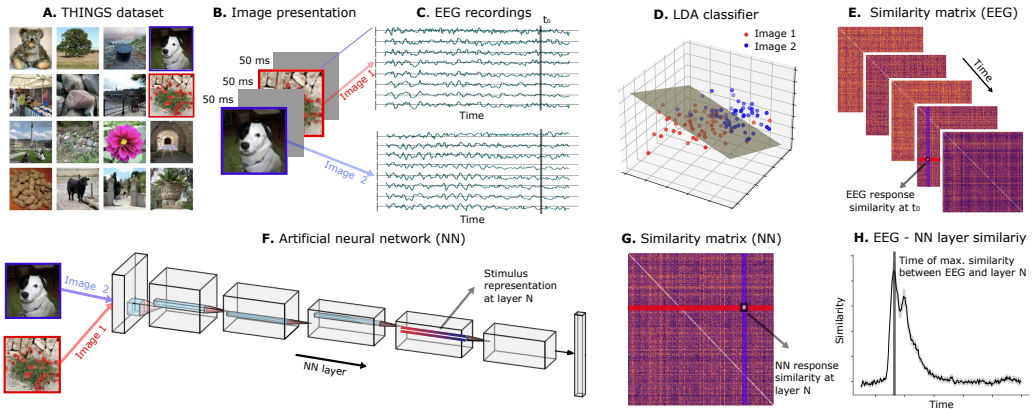

Figure 1: Overview of analysis methodology. A) Public domain images similar to the ones from the THINGS dataset (from PublicDomainPictures.net). B) Images were presented for 50 ms and 100 ms in THINGS-EEG 1 and THINGS-EEG 2, respectively, with inter-stimulus intervals of the same duration. C) For each image, EEG signals were acquired at 64 channels at different scalp locations. D) An LDA model was trained to classify all pairs of images at each time point using EEG signals as features. E) EEG RDMs were obtained using LDA accuracy as a distance metric. F) Images were used as inputs to different ANNs and the activations at all layers were extracted. G) For each ANN model and layer, RDMs were constructed using the cosine distance between activations. H) For each time point, RDMs from EEG and each ANN model were compared using Spearman correlation as similarity metric.

down-sampling to 250 Hz. EEG data were preprocessed by baseline correction, where the mean of the pre-stimulus interval was subtracted for each trial and channel separately. For all this operations, we used MNE-Python v.1.9.0 [30].

3.3 EEG Decoding Distance

We used regularized Linear Discriminant Analysis (LDA) implemented in *scikit-learn* [31] to classify object pairs based on EEG signals at each time point (Fig. 1D). For each image pair, trials were split using a 70/30 stratified split. A regularized Linear Discriminant Analysis (LDA) classifier was fit on the training set with tolerance 0.01 and evaluated on the held-out test set [2]. The resulting accuracies were used to construct a 200 $\times$ 200 representational dissimilarity matrix (RDM) for each subject in each point in time (Fig. 1E).

3.4 Artificial Neural Network Architectures

Images corresponding to both datasets were fed to ANNs to obtain layer-specific activations (Fig. 1F) after reshaping to a common size of 224 $\times$ 224 pixels. ANNs included three models previously used by Gifford and colleagues: AlexNet (CNN with 5 convolutional layers followed by 3 fully-connected layers), MoCo (feedforward architecture with self-supervised training), and ResNet-50 (supervised feedforward neural network with 50 layers and shortcut connections between layers at different depths) [29]. All of these networks were pre-trained on object categorization using the ILSVRC-2012 training image partition. We also included VGG-19 and ViT. VGG-19 was pre-trained on the ImageNet dataset, which contains over 3 million images and their corresponding concepts. ViT was pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224 $\times$ 224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224 $\times$ 224.

3.5 Artificial Neural Network Activations

The selection of layers for AlexNet, MoCo and ResNet-50 was based on Gifford et al [29]. For VGG-19, we extracted activations from the five pooling layers (*pool1* to *pool5*) and the second

[2]Refer to the code for further details A.3

fully-connected layer (*fc2*). From ViT, we utilized the activations from all 12 layers, extracting the class token from each layer, which represents the entire input with 768 dimensions [32]. We then flattened each layer, then re-scaled it and finally performed an iterative poly kernel Principal Component Analysis (PCA) algorithm with 200 components. We computed the cosine distance between activations to create a 200 $\times$ 200 RDM for each layer (Fig. 1G).

3.6 Temporal Correspondence

We compared each subject's RDM at a given time with the RDM of each layer across ANN architectures. Then, we investigated whether the peak correlation times correlated with the order of the layers in each architecture, and whether there were differences between architectures (Fig. 1H). All the correlations between architectures were calculated using Spearman rho and all the comparisons between architectures were done using Wilcoxon test. Finally, we calculated the slope of correlation between LDA-semantic correlation and LDA-activation correlation through layers or each architecture for each dataset. We employed the Mann-Whitney U to test if there were differences between both datasets in each layer.

3.7 Low-level Visual Features

We used the methods described by Lützow Holm et al. to quantify the statistics of low-level visual features [26]. We extracted the descriptors computed by Harrison [27] and added values for the red, green, and blue channels. Using a recursive procedure [26], we then determined the 8 most important variables: 3 related to spectral features (SFE), 2 related to object edges (OE), luminance, and the colors green and blue, followed by PCA with retention of the first principal component. By correlating the absolute difference between each pair of vectors, we obtained 200 $\times$ 200 RDMs for each dataset.

3.8 Semantic Distance

For the semantic distance between concepts, we used FastText embeddings [3] between concepts and computed the cosine distance between every pair of embeddings, resulting in 200 $\times$ 200 RDMs for each dataset [33]. This RDM was validated using another set of embeddings from a state-of-the-art large language model, NV-Embed-v2 [34] and Stella [35]. See A.1 for further details.

4 Results

4.1 Analysis of ANN representations

Figure 2A displays the correlation between RDMs corresponding to all layers for the THINGS-EEG 1 dataset, both within and between ANNs. The architectures exhibited a high internal correlation, except for the last layer of the ViT, which differed from the other layers within the same architecture and from the layers of the other architectures. There was also correspondence between the first four architectures in terms of the relative order of their layers. As a general trend, the correlation with image statistics decreased as a function of the layer (Fig. 2B), while the correlation with semantic information increased (Fig. 2C). ViT differed from the rest as the decline of its correlation with image statistics was subtler than the rest and the slope of the semantic correlation did not differ significantly from zero. Similar results are shown for THINGS-EEG 2 in Figs. 2D-F. Additional analyses using a held-out validation image dataset are provided in Appendix A.1, Figure A.1.

4.2 Brain-ANN temporal correspondence

The time of maximum correlation (T_{max}) between the ANN and EEG RDMs is shown in Fig. 3A (THINGS EEG 1) and B (THINGS EEG 2). T_{max} increased as a function of the layer, with exceptions for the last layer of VGG-19 and ViT in THINGS-EEG 1, as well as the last layer in both MoCo and ViT for THINGS-EEG 2. Also, the peak correlation between brain and ANN RDMs (ρ_{max}) decreased as a function of the layer, except for the first layer of VGG-19 for both datasets and ViT in

[3] Training Corpus: Common Crawl (600 billion tokens). 2 million word vectors. 300 dimensions embeddings

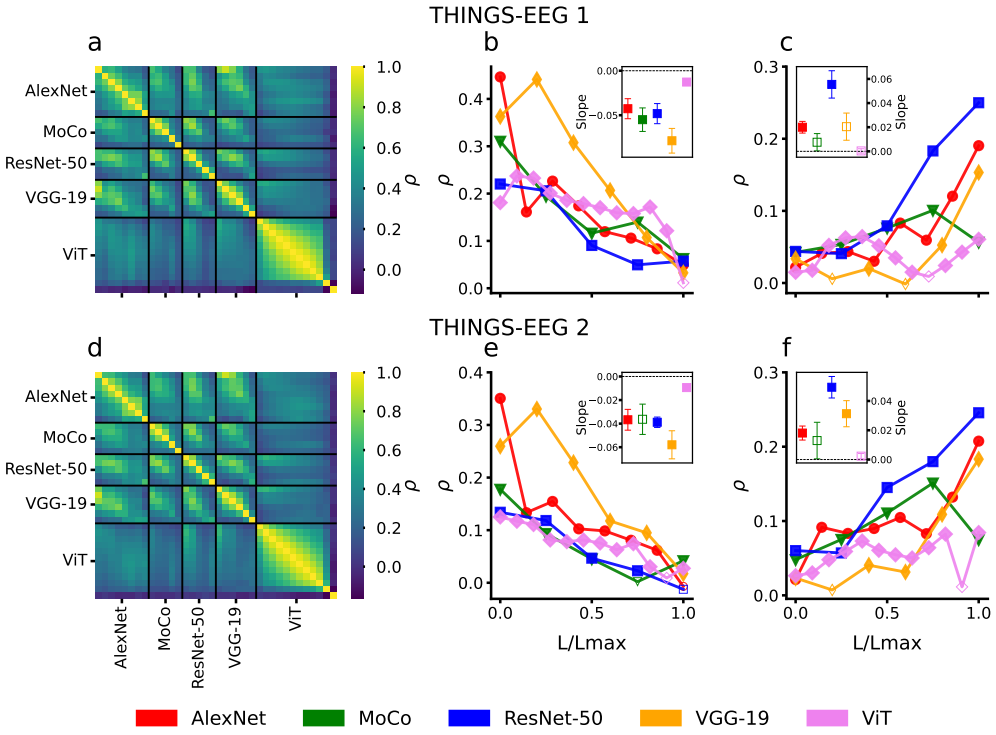

Figure 2: Transformation of visual to semantic representations in ANNs. A) Correlation between activations within and between ANNs for THINGS-EEG 1. B) Spearman correlation between RDMs based on low-level image statistics and ANN activations for THINGS-EEG 1. C) Spearman correlation between RDMs based on semantic similarity and ANN activations for THINGS-EEG 1. D, E, F) Same as panels A, B, C but for the THINGS-EEG 2 database. Full symbols indicate statistically significant Spearman correlation (p-value < 0.05).

THINGS-EEG 1. Appendix Figures A.2–A.3 show the same $T_{\max}$ and $\rho_{\max}$ analyses computed over an independent validation set.

Next, we investigated the correlation between the layer order and the time of maximum correlation for each ANN (Figs. 5A and B for THINGS-EEG 1 and 2, respectively). For all models, there was a correspondence between T_{max} and layer order. Using the median value as a reference, ResNet-50 exhibited the highest time-layer correspondence, although there were no significant differences compared to the other architectures. Additionally, we investigated whether the peak correlation ρ_{max} behaved in a similar way (Figs. 5C and D). Except for ViT with THINGS-EEG 1, there was a negative correlation between ρ_{max} and layer order for all ANNs, indicating that the strength of brain-ANN correspondence decreased as a function of the layer. We found that ρ_{max} was closer to -1 and presented less variance for THINGS-EEG 2 compared to THINGS-EEG 1.

4.3 Contributions of visual and semantic information

We then compared the RDMs obtained from image statistics and semantics with those computed from the EEG data (Fig. 6). We observed two significant peaks for the image statistics correlation in THINGS-EEG 1 (100 and 200 ms) and a very conspicuous peak for THINGS-EEG 2 (near 100 ms) followed by two smaller peaks at 200 and 300 ms. For THINGS-EEG 1, there was no significant correlation between the EEG and semantic RDMs. For THINGS-EEG 2, significant correlations between EEG and semantic RDMs were observed after the peak of correlation with the image statistics RDMs. As expected, the LDA curve for THINGS-EEG 1 was aligned with the visual component of the EEG, suggesting that for shorter stimuli, low-level features drove almost exclusively the discrimination between concepts. In contrast, the LDA curve for THINGS-EEG 2

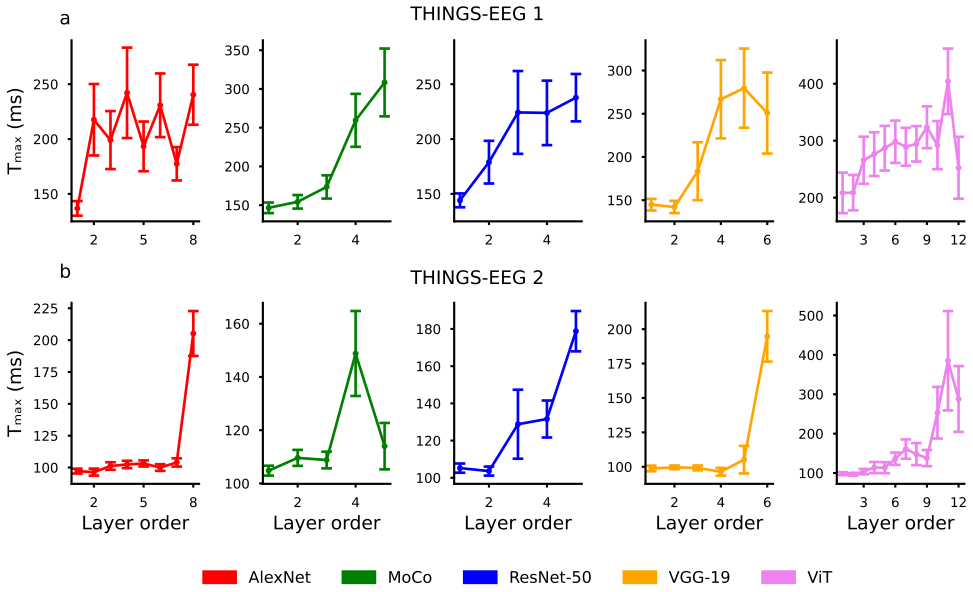

Figure 3: Time of maximum correlation between EEG and ANN activations. A) Time corresponding to the peak Spearman correlation between RDMs based on EEG and ANN activations (T_{max}), for each layer of the models (layer order) and obtained using THINGS-EEG 1. B) Same as panel A, but for THINGS-EEG 2. Error bars correspond to standard error of the mean.

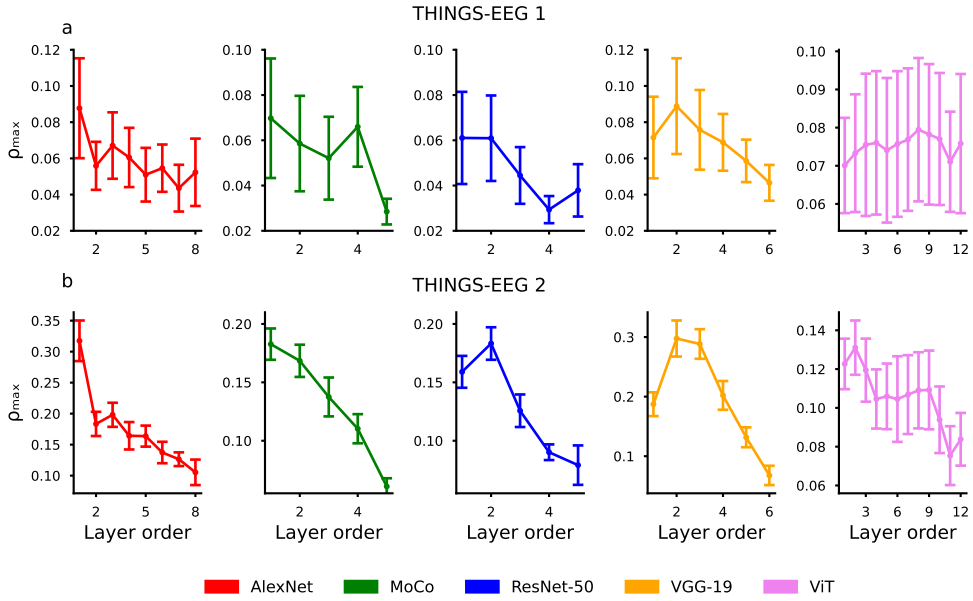

Figure 4: Peak correlation between EEG and activations. A) Peak value of the Spearman correlation between RDMs based on EEG and ANN activations (ρ_{max}), for each layer of the models (layer order) and obtained using THINGS-EEG 1. B) Same as panel A, but for THINGS-EEG 2. Error bars correspond to standard error of the mean.

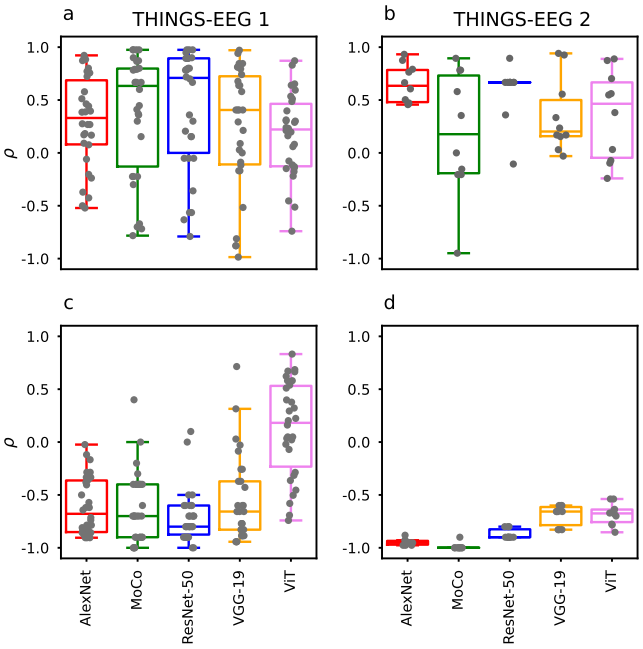

Figure 5: Monotonous behavior of T_{max} and ρ_{max} vs. layer order. A) Spearman correlation between layer order and T_{max} for THINGS-EEG 1. B) Same as in panel A, but for THINGS-EEG 2 dataset. C) Correlation between ρ_{max} and layer order for THINGS-EEG 1. D) Same as in panel C, but for THINGS-EEG 2.

had peak aligned with the visual component, but with a slower decay, as the semantic component was more relevant to distinguish image pairs. We also evaluated the correlation between the RDMs based on image statistics and semantic embeddings, without obtaining significant results (THINGS-EEG 1: r= -0.009, p-value < 0.05; THINGS-EEG 2: r = 0.011, p-value > 0.05). See Appendix Figure A.4 for similar results obtained in an independent validation dataset.

Finally, we computed the linear regression between EEG-semantic and EEG-activation correlations, and calculated the slopes for each layer of each architecture in THINGS-EEG 1 and 2 (Fig. 7). For AlexNet, Moco, ResNet-50 and VGG-19, THINGS-EEG 2 showed a significantly higher slope than THINGS-EEG 1 for some layers (one tailed t-test, p-value < 0.05). In these cases, brain-ANN correspondence was predicted by the semantic contribution to EEG responses across participants.

5 Discussion and Conclusions

We demonstrated a robust and consistent temporal correspondence between the temporal hierarchy of artificial neural network activations and human EEG responses during visual perception. Across multiple ANN architectures, we observed that early EEG components aligned with layers encoding low-level visual features, whereas later EEG components corresponded with deeper layers that capture semantic content. This alignment was modulated by stimulus duration, with longer exposures revealing stronger correspondence to semantic representations. Finally, for longer presentations times we observed a link between brain-ANN correspondence and the semantic contribution to EEG decoding (with the exception of the vision transformer). This result suggests that brain-ANN is driven by low-level visual representations for shorter stimulus duration, while semantic representations drive brain-ANN for longer durations, especially at deeper layers.

Our findings suggest that the observed brain–ANN reflects a shared computational principle: the progressive transformation from visual input to abstract semantic representations over time. Notably, even architectures without clear biological plausibility (e.g., ViT) exhibited this correspondence, supporting the idea that temporal alignment is driven more by representational hierarchy than by

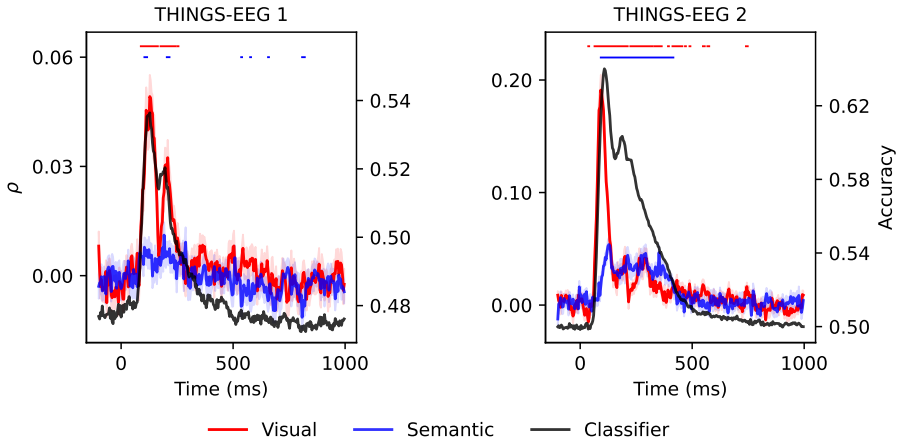

Figure 6: Contribution of low-level visual features and semantics to the EEG RDMs. Correlation between image statistics (red) and semantic (blue) RDMs, and the EEG RDM vs. time, for THINGS-EEG 1 (left) and THINGS-EEG 2 (right). The black curves correspond to LDA classifier accuracy. Lines above the plots indicate points in time with statistically significant Spearman correlation (red for statistics and blue for semantic, respectively; one sample t-test, p-value < 0.05). Shaded areas indicate standard error of the mean.

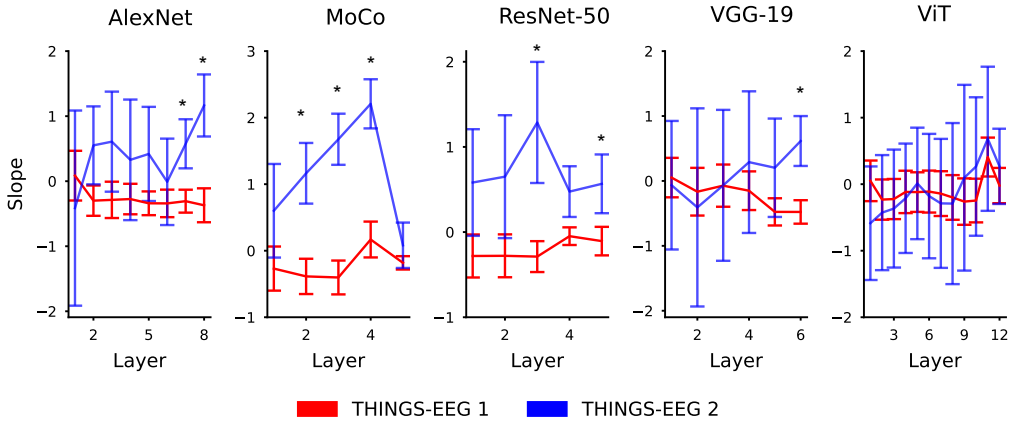

Figure 7: Regression of LDA-semantic vs. LDA-activation correlations. Slope of the best linear regression vs. layer order for all ANNs, comparing THINGS-EEG 1 (red) vs. THINGS-EEG 1 (blue). Asterisks indicate statistically significant differences between the slopes corresponding to both datasets (one tailed t-test, p-value < 0.05). Error bars correspond to standard error of the mean.

biological fidelity. This result is in agreement with fMRI and electrophysiological studies showing that architectural differences in AANs have little consequence in the prediction of brain responses during object recognition [16, 6].

Both the brain and ANNs process low-level visual information to obtain semantic representations. In the brain, the receptive field of neurons becomes increasingly larger and complex along the ventral pathway [36, 2]. Similarly, the hierarchical organization of CNNs results in units capable of invariant activation to specific objects [17]. Our analysis showed that this principle extends to other architectures: as layer order increases, activation similarity becomes more unrelated to low-level visual similarity and is increasingly predicted by semantic similarity. Since a similar progression is manifest in EEG responses, it is plausible that brain-ANN correspondence is primarily driven by the shared hierarchical transformation of visual to semantic representations.

We observed that the similarity between RDMs computed from EEG and AAN activations (ρ_{max}) decreased as a function of the layer order. As shown in Fig. 6 and consistent with previous reports, the low-level visual features present a larger contribution to early EEG responses compared to semantic content [4]. Thus, this gradual reduction of ρ_{max} may parallel the shared shift from visual to semantic representations in the brain and ANNs. Moreover, Fig. 5 shows that ρ_{max} decreased at a faster rate in the dataset with shortest presentation times. The lack of significant semantic contribution to EEG responses observed in this dataset suggests that, in this case, brain-ANN correspondence could have been solely dominated by low-level visual features during the first $\sim$200 ms after stimulus presentation (corresponding to the first layers of all models except ViT, as shown in Fig. 3. This result is aligned with known limits to post-perceptual processing for very short stimulus presentation times in the 200 ms post-stimuli [37] (see also Fig. 6).

While the ViT model behaved similarly to other ANNs in terms of its temporal correspondence with EEG activity, it also presented some distinct features. As seen in the off-diagonal elements of Figs. 2A and D, a hierarchical correspondence existed between the intermediate layers of all ANNs, with the exception of the ViT. This result is in agreement with previous reports and is indicative of transformer-specific representations [38], which reinforces the independence of brain-ANN correspondence with regards to the idiosyncratic computational features of individual network architectures. The ViT model also behaved differently (i.e. non-monotonously) regarding the relative contribution of visual and semantic information, which could stem from the more uniform nature of representations across ViT layers [39].

In conclusion, our findings support the view that artificial neural networks resemble the brain not because they replicate specific low-level operations, but because they share representational goals—most notably, the transformation of visual input into semantic abstractions. This perspective de-emphasizes biologically motivated interpretations of CNNs and challenges the explanatory power of model-specific mechanisms such as convolutions, spatial locality, or receptive fields. Future work should investigate the role of hierarchical organization in the emergence of brain-like representations by systematically comparing models with varying degrees of hierarchical structure, including shallow or non-hierarchical architectures.

6 Limitations

EEG spatial resolution. The limited spatial resolution of EEG restricted our ability to localize the neural sources of observed effects onto specific anatomical substrates, limiting our analysis to the temporal aspects of brain-ANN correspondence. Future work using source-localized MEG or intracranial recordings could provide more fine-grained spatial insights.

Variance in the training set and tasks. Generalization from THINGS to naturalistic or dynamic visual environments (e.g., videos, scenes) remains untested. In particular, the limited sub-set of images may constrain the generalizability of the observed temporal correspondences. Also, while we employed a range of ANN architectures, all were trained on variants of ImageNet and optimized for object classification. Thus, our findings may not extend to models trained on other tasks.

RSA. Similarity in representational space does not necessarily imply that the compared systems encode the same features or use similar representational formats. RSA is also sensitive to low-level confounds in the stimuli, which may inflate apparent correspondences driven by shared biases rather

than meaningful alignment [40]. Finally, RSA offers only an indirect view of the temporal dynamics or transformations that occur within or across systems.

Semantic similarity. Word embeddings provide a language-informed approximation of semantic structure based on distributional statistics of word co-occurrence, but they may not fully capture the distinctions most relevant to human visual perception. While restricting the analysis to images with a single dominant concept helps mitigate this concern, caution is warranted when extending this approach to more complex stimuli with ambiguous or multi-object content. Future work could address this limitation by incorporating more nuanced descriptions derived from large language models with multimodal capabilities [41].

References

[1] Nikos K Logothetis and David L Sheinberg. "Visual object recognition." In: *Annual review of neuroscience* 19 (1996), pp. 577–621.

[2] Maximilian Riesenhuber and Tomaso Poggio. "Models of object recognition". In: *Nature neuroscience* 3.11 (2000), pp. 1199–1204.

[3] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. "Resolving human object recognition in space and time". In: *Nature neuroscience* 17.3 (2014), pp. 455–462.

[4] Erika W Contini, Susan G Wardle, and Thomas A Carlson. "Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions". In: *Neuropsychologia* 105 (2017), pp. 165–176.

[5] Ravpreet Kaur and Sarbjeet Singh. "A comprehensive review of object detection with deep learning". In: *Digital Signal Processing* 132 (2023), p. 103812.

[6] Colin Conwell et al. "A large-scale examination of inductive biases shaping high-level visual representation in brains and machines". In: *Nature communications* 15.1 (2024), p. 9383.

[7] Nikolaus Kriegeskorte. "Deep neural networks: a new framework for modeling biological vision and brain information processing". In: *Annual review of vision science* 1.1 (2015), pp. 417–446.

[8] Umut Güçlü and Marcel AJ Van Gerven. "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream". In: *Journal of Neuroscience* 35.27 (2015), pp. 10005–10014.

[9] Charles F Cadieu et al. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition". In: *PLoS computational biology* 10.12 (2014), e1003963.

[10] Michael Eickenberg et al. "Seeing it all: Convolutional network layers map the function of the human visual system". In: *NeuroImage* 152 (2017), pp. 184–194.

[11] Yaoda Xu and Maryam Vaziri-Pashkam. "Limits to visual representational correspondence between convolutional neural networks and the human brain". In: *Nature communications* 12.1 (2021), p. 2065.

[12] Daniel LK Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[13] Radoslaw Martin Cichy et al. "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence". In: *Scientific reports* 6.1 (2016), p. 27755.

[14] Nathan CL Kong et al. "Time-resolved correspondences between deep neural network layers and EEG measurements in object processing". In: *Vision Research* 172 (2020), pp. 27–45.

[15] Katja Seeliger et al. "Convolutional neural network-based encoding and decoding of visual object recognition in space and time". In: *NeuroImage* 180 (2018), pp. 253–266.

[16] M Schrimpf et al. *Brainscore: Which artificial neural network for object recognition is most brain-like?[Pages: 407007 Section: New Results]*. 2018.

[17] Daniel LK Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3 (2016), pp. 356–365.

[18] Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018.1 (2018), p. 7068349.

[19] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.

[20] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in systems neuroscience* 2 (2008), p. 249.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[22] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[23] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[24] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[25] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[26] Eric Lützow Holm, Diego Fernández Slezak, and Enzo Tagliazucchi. "Contribution of low-level image statistics to EEG decoding of semantic content in multivariate and univariate models with feature optimization". In: *NeuroImage* 293 (2024), p. 120626.

[27] William J Harrison. "Luminance and contrast of images in the THINGS database". In: *Perception* 51.4 (2022), pp. 244–262.

[28] Tijl Grootswagers et al. "Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams". In: *Scientific Data* 9.1 (2022), p. 3.

[29] Alessandro T Gifford et al. "A large and rich EEG dataset for modeling human visual object recognition". In: *NeuroImage* 264 (2022), p. 119754.

[30] Alexandre Gramfort et al. "MEG and EEG Data Analysis with MNE-Python". In: *Frontiers in Neuroscience* 7.267 (2013), pp. 1–13. DOI: 10.3389/fnins.2013.00267.

[31] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[32] Tianyang Lin et al. "A survey of transformers". In: *AI open* 3 (2022), pp. 111–132.

[33] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146.

[34] Chankyu Lee et al. "Nv-embed: Improved techniques for training llms as generalist embedding models". In: *arXiv preprint arXiv:2405.17428* (2024).

[35] Dun Zhang et al. "Jasper and Stella: distillation of SOTA embedding models". In: *arXiv preprint arXiv:2412.19048* (2024).

[36] Shaul Hochstein and Merav Ahissar. "View from the top: Hierarchies and reverse hierarchies in the visual system". In: *Neuron* 36.5 (2002), pp. 791–804.

[37] Joachim Bellet et al. "Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing". In: *Neuroscience of consciousness* 2022.1 (2022), niac005.

[38] Shikhar Tuli et al. "Are convolutional neural networks or transformers more like human vision?" In: *arXiv preprint arXiv:2105.07197* (2021).

[39] Maithra Raghu et al. "Do vision transformers see like convolutional neural networks?" In: *Advances in neural information processing systems* 34 (2021), pp. 12116–12128.

[40] Marin Dujmović et al. "Some pitfalls of measuring representational similarity using Representational Similarity Analysis". In: *bioRxiv* (2022), pp. 2022–04.

[41] Duzhen Zhang et al. "Mm-llms: Recent advances in multimodal large language models". In: *arXiv preprint arXiv:2401.13601* (2024).

[42] Tijl Grootswagers, Amanda K. Robinson, and Thomas A. Carlson. "The representational dynamics of visual objects in rapid serial visual processing streams". In: *NeuroImage* 188 (2019), pp. 668–679. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2018.12.046. URL: https://www.sciencedirect.com/science/article/pii/S1053811918321906.

A Technical Appendices and Supplementary Material

A.1 Correlation using embedding models

Correlation between fastText and Stella for THINGS-EEG 1: PearsonRResult(statistic = 0.47, p-value = 0.0) Correlation between fastText and NV-Embed-v2 for THINGS-EEG 1: PearsonRResult(statistic = 0.53, p-value = 0.0) Correlation between NV-Embed-v2 and Stella for THINGS-EEG 1: PearsonRResult(statistic = 0.55, p-value = 0.0)

Correlation between fastText and Stella for THINGS-EEG 2: PearsonRResult(statistic = 0.41, p-value = 0.0) Correlation between fastText and NV-Embed-v2 for THINGS-EEG 2: PearsonRResult(statistic = 0.55, p-value = 0.0) Correlation between NV-Embed-v2 and Stella for THINGS-EEG 2: PearsonRResult(statistic = 0.56, p-value = 0.0)

A.2 Datasets

THINGS-EEG 1 [4] This study was approved by the University of Sydney Ethics Committee, and informed consent was obtained from all participants. Data are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Of the 50 subjects, we included the 30 with the highest peak LDA values, as many of the remaining participants exhibited notably poor or noisy signals.

THINGS-EEG 2[5] This study was approved by the Ethics Committee of Freie Universität Berlin (conducted in accordance with the Declaration of Helsinki). License: Creative Commons Attribution 4.0 International (CC BY 4.0) Repository: Open Science Framework

All validation tests were conducted using the dataset from Grootswagers et al. [42] , which consists of EEG recordings from 16 adults viewing 200 visual objects presented at 20 Hz and then 5 Hz on a screen. The study was approved by the Ethics Committee of the University of Sydney. EEG recordings and images used as stimuli are available in the Open Science Framework repository [6]. Repository: Open Science Framework. Image License: Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

A.3 Implementation details

All computations were performed using Python 3.11.7 on a Ubuntu 22.04 Intel® Core™ i9-10900 CPU @ 2.80GHz × 20, Mesa Intel® UHD Graphics 630, 64 GB RAM with 1 TB storage. Our code is publicly available at `https://anonymous.4open.science/r/0BF6/README.md`.

[4] https://openneuro.org/datasets/ds003825/versions/1.2.0

[5] https://osf.io/3jk45/

[6] https://osf.io/a7knv/

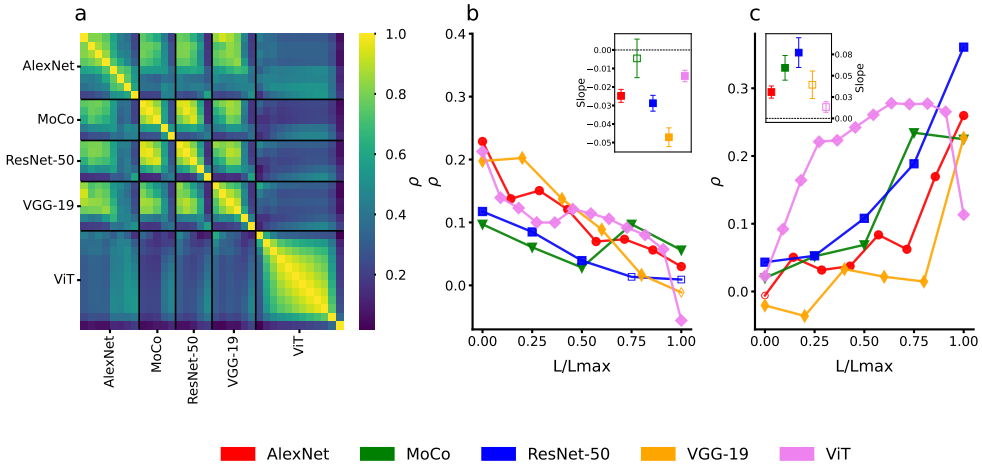

Figure A.1: Comparison of activations for each layer of the architectures. A) Heatmap of correlation between RDMs for each layer of each architecture using the images of the validation dataset. B) Spearman correlation between image statistics RDM and activations RDM for the validation dataset. C) Spearman correlation between semantic correlation RDM and activations RDM for the validation dataset. Insets: slope for each curve with standard deviation. Full markers indicate statistical significance from zero (p-value < 0.05).

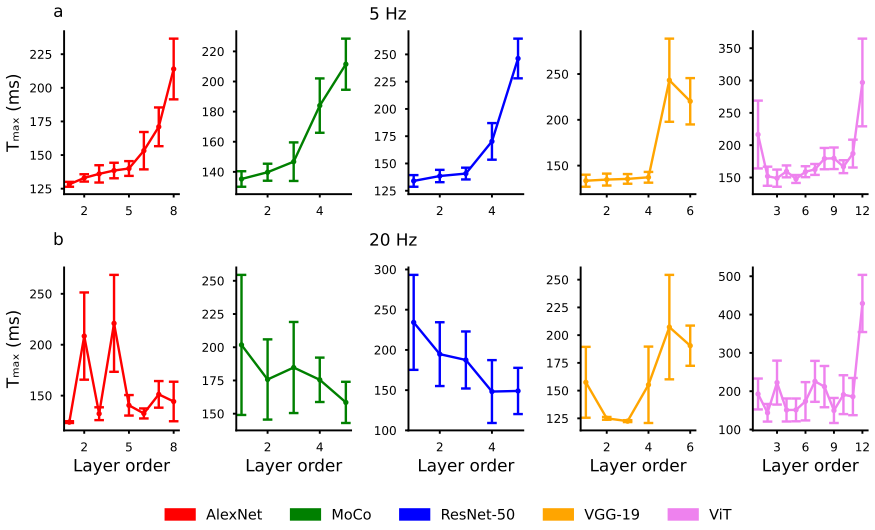

Figure A.2: Time of maximum correlation between EEG and activations. A) Average time of maximum Spearman correlation between RDM of each layer and LDA for the validation dataset of 5 Hz. B) Same as A, but for the validation dataset of 20 Hz. Error bars correspond to standard error of the mean.

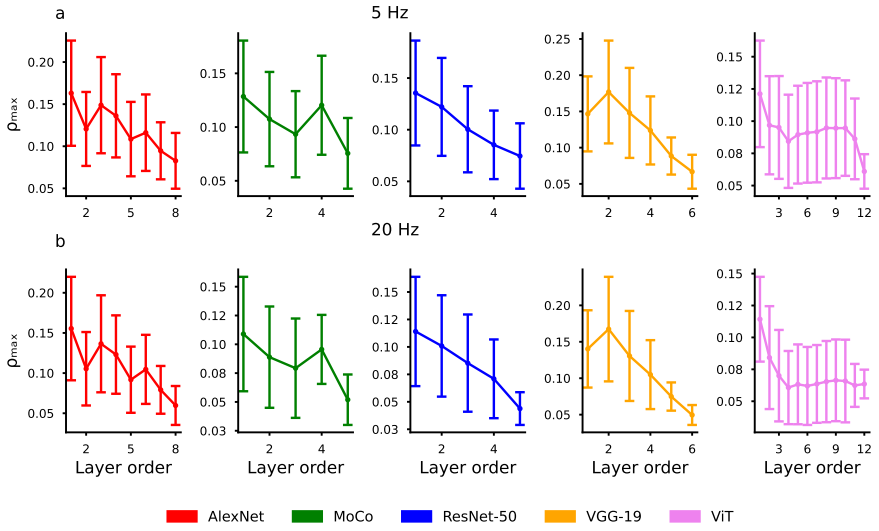

Figure A.3: Value of maximum correlation between EEG and activations. A) Average maximum value of Spearman correlation between RDM of each layer and LDA for the validation dataset of 5 Hz. B) Same as A, but for the validation dataset of 20 Hz. Error bars correspond to standard error of the mean.

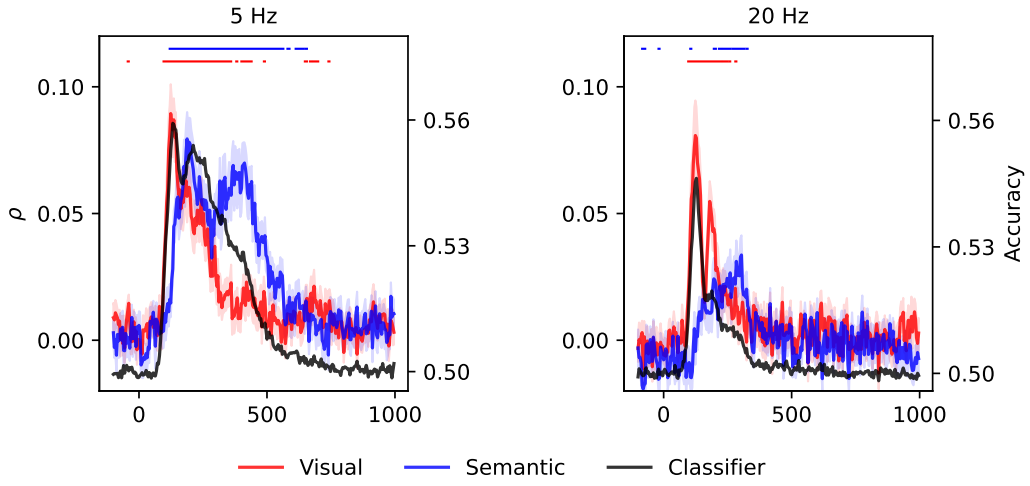

Figure A.4: Contribution of low-level visual features and semantics to the EEG RDMs. Correlation between image statistics (red) and semantic (blue) RDMs, and the EEG RDM vs. time, for the 5 Hz (left) and 20 Hz (right) validation dataset. The black curves correspond to the LDA classifier used in each dataset. Lines above the plots indicate points in time with statistically significant Spearman correlation (red for statistics and blue for semantic, respectively; one sample t-test, p-value < 0.05). Shaded areas indicate standard error of the mean.