VISIO-LINGUISTIC BRAIN ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Enabling effective brain-computer interfaces needs understanding how the human brain encodes stimuli across modalities such as visual, language (or text), etc. Brain encoding aims at constructing fMRI brain activity given a stimulus. There exists a plethora of neural encoding models which study brain encoding for single mode stimuli: visual (pretrained CNNs) or text (pretrained language models). Few recent papers have also obtained separate visual and text representation models and performed late-fusion using simple heuristics. However, previous work has failed to explore: (a) the effectiveness of image Transformer models for encoding visual stimuli, and (b) co-attentive multi-modal modeling for visual and text reasoning. Further, as pretrained image Transformers and multi-modal Transformers have continued to evolve, it is important to understand if they are becoming more brain-like and hence lead to improved brain encoding. In this paper, we systematically explore the efficacy of image Transformers (ViT, DEiT, and BEiT) and multi-modal Transformers (VisualBERT, LXMERT, ViLBERT, and CLIP) for brain encoding. Extensive experiments on two popular datasets, BOLD5000 and Pereira, provide the following insights. (1) To the best of our knowledge, we are the first to investigate the effectiveness of image and multi-modal Transformers for brain encoding. (2) Surprisingly, we observe a better encoding correlation between Transformer model layers and the levels of visual processing in the human brain when compared to CNN architectures. (3) We find that multi-modal Transformers significantly outperform previously proposed single-mode CNNs, image Transformers as well as other previously proposed multi-modal models, thereby establishing new state-of-the-art. The supremacy of visio-linguistic models raises the question of whether the responses elicited in the visual regions are affected implicitly by linguistic processing even when passively viewing images. Future fMRI tasks can verify this computational insight in an appropriate experimental setting. We make our code publicly available[1].

1 INTRODUCTION

In the past decade, artificial neural networks have witnessed a remarkable performance in the computational neuroscience community in understanding how the brain effortlessly performs information perception and processing given various forms of sensory inputs like visual processing in object recognition tasks (Yamins et al., 2014; Cadieu et al., 2014; Eickenberg et al., 2017). This line of work, namely brain encoding, aims at constructing neural fMRI (functional magnetic resonance imaging) brain activity given an input stimulus. The two most studied forms of stimuli include vision and language.

Since the discovery of the relationship between language/visual stimuli and functions of brain networks (Constable et al., 2004; Thirion et al., 2006), researchers have been interested in understanding how the neural encoding models predict the fMRI brain activity. Recently, several brain encoding models have been developed to (i) understand the ventral stream in biological vision (Yamins et al., 2014; Kietzmann et al., 2019; Bao et al., 2020), and (ii) to study the higher-level cognition like language processing (Gauthier & Levy, 2019; Schrimpf et al., 2020a; Schwartz et al., 2019). Previous work has mainly focused on independently understanding vision and text stimuli. However, the biological systems perceive the world by simultaneously processing high-dimensional inputs

[1]https://www.dropbox.com/s/qxvxahaknzigr4s/Visio_linguistic_Brain_Encoding.zip?dl=0

from diverse modalities such as vision, auditory, touch, proprioception, etc. (Jaegle et al., 2021). In particular, how the brain effectively processes and provides its visual understanding through natural language and vice versa is still an open question in neuroscience.

There exist a plethora of neural encoding models which predict the fMRI brain activity using representations of single-mode stimuli: visual or text. Convolutional neural networks (CNNs) have been shown to encode semantics from visual stimuli effectively. Interestingly, intermediate layers in deep CNNs trained on ImageNet (Deng et al., 2009) categorization task can partially account for how neurons in intermediate layers of the visual system respond to any given image (Yamins et al., 2013; 2014; Güçlü & van Gerven, 2015; Yamins & DiCarlo, 2016; Wang et al., 2019). However, the more recent and deeper CNNs have not been shown to further improve on measures of brain-likeness, even though their ImageNet performance has vastly increased (Russakovsky et al., 2015). Recently, Kubilius et al. (2019) proposed a shallow recurrent anatomical network, CORnet, which provided state-of-the-art results on the Brain-score (Schrimpf et al., 2020b) benchmark. Similar to visual encoding models, neural models like deep recurrent neural networks (RNNs), Transformer (Vaswani et al., 2017) based language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019) have been leveraged to predict fMRI brain activity corresponding to semantic vectors of linguistic items including words, phrases, sentences, and paragraphs (Gauthier & Levy, 2019; Schrimpf et al., 2020a). However, these pieces of work suffer from the following drawbacks: (1) Although these neural encoding models have demonstrated promising results of processing in one of the two brain regions (visual cortex and pre-frontal cortex), they still need plenty of efforts to improve on brain encoding for other parts of the brain. Brain encoding for more brain regions is important since input stimuli elicit diverse and distributed representations in the brain, and these activation responses could be repurposed for several novel tasks. (2) They manually choose particular CNN layers whose activations are used for accurately predicting brain fMRIs specific to the datasets they work with; thus, generalization to other datasets is unclear.

Unlike previous studies, which focus on single-modality (either visual or language stimuli), some authors demonstrated that multi-modal models formed by combining text-based distributional information with visual representations provide a better proxy for human-like intelligence (Anderson et al., 2015; Oota et al., 2019). However, these methods extract representations from each mode separately (image features from CNNs and text features from pretrained embeddings) and then perform a simple late-fusion. Thus, they cannot exploit semantic correspondence across the two modes at different levels effectively. Such late-fusion based multi-modal models are the closest to our work, and our experiments show that our models outperform them.

Recently, image-based transformer models like ViT (Dosovitskiy et al., 2020), DEiT (Touvron et al., 2021), and BEiT (Bao et al., 2021) have been shown to provide excellent results compared to traditional CNNs on image classification tasks. Also, multi-modal Transformers like VisualBERT (Li et al., 2019), LXMERT (Tan & Bansal, 2019), ViLBERT (Lu et al., 2019) and CLIP (Radford et al., 2021) have shown awesome results on visio-linguistic tasks like visual question answering, visual common-sense reasoning, etc. Inspired by the success of language, image, and multi-modal Transformers, we build multi-modal transformer models to learn the joint representations of image content and natural language and use them for brain encoding.

In this work, we investigate several fundamental questions: do vision Transformers act like the human brain visual system? Does the hierarchy of Transformer layers match with the visual cortical hierarchy? Do multi-modal Transformers act more or less brain-like (perceiving multi-modal inputs simultaneously)? Moreover, can we use multi-modal Transformers to perform fMRI encoding on the whole brain? In this paper, we study these questions, uncovering insights about the association between fMRI voxels and representations of multi-modal/image Transformers and CNNs. Fig. 1 illustrates our brain encoding methodology.

Specifically, we make the following contributions in this paper. (1) We present state-of-the-art encoding results using multi-modal Transformers. We also study the effectiveness of our models in a cross-validated setting. (2) Our approach generalizes the use of Transformer-based architectures, removing the need to manually select specific layers as in existing CNN-based fMRI encoding architectures. (3) We observe a better encoding correlation between transformer layer activations and the levels of visual processing in the brain compared to CNN layer activations.

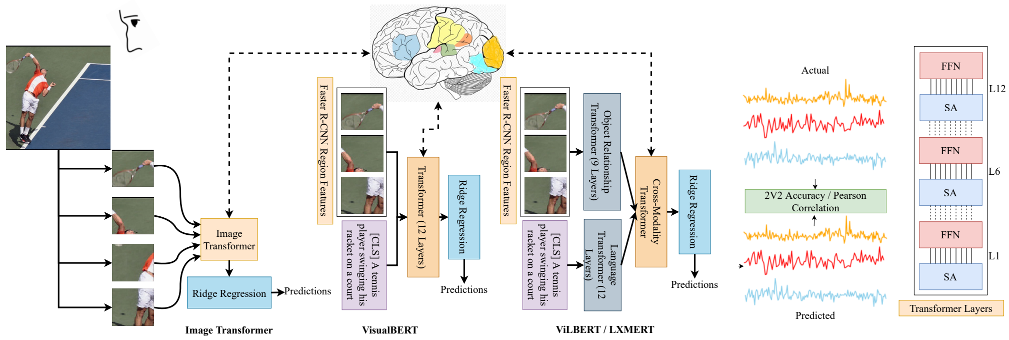

Figure 1: Brain encoding methodology. We use features from image/multi-modal Transformers (like ViT, VisualBERT and ViLBERT/LXMERT) as input to the regression model to predict the fMRI activations for different brain regions. Brain encoding results are evaluated by computing 2V2 accuracy and Pearson correlation between actual and predicted activations. We also perform layer-wise correlation analysis between transformer layers and brain regions.

2 Brain Imaging Datasets

The following datasets are popularly used in the literature for studying brain encoding: Vim-1 (Kay et al., 2008), Harry Potter (Wehbe et al., 2014), Pereira et al. (2018), BOLD5000 (Chang et al., 2019), Algonauts (Cichy et al., 2019) and SS-fMRI (Beliy et al., 2019). Vim-1 has only black and white images and is only related to object recognition, and is subsumed by BOLD5000. SS-fMRI is smaller and very similar to BOLD5000. Harry Potter dataset does not have images. Lastly, fMRIs have not been made publicly available for the Algonauts dataset. Hence, we experiment with BOLD5000 and Pereira datasets in this work.

BOLD5000: BOLD5000 dataset was collected from four subjects where three subjects engaged in 5254 natural images (ImageNet: 2051, COCO: 2135, Scenes: 1068) while fMRIs were acquired. The fourth subject was shown 3108 images only. Details of the visual stimuli and fMRI protocols of the dataset have been discussed in (Chang et al., 2019). We only briefly summarize the details of the dataset in Table 1 in the Appendix. The data covers ten visual areas in the human visual cortex, i.e., early visual area (LHEarlyVis, RHEarlyVis); object-related areas such as the lateral occipital complex (LHLOC, RHLOC); and scene related areas such as the occipital place area (LHOPA, RHOPA), the parahippocampal place area (LHPPA, RHPPA), and the retrosplenial complex (LHRSC, RHRSC). Each image also has corresponding text labels: ImageNet has a few out of 1000 possible tags per image, COCO has five captions per image, and Scenes has one out of 250 possible categories per image.

Pereira: For the Pereira dataset, participants were shown concept word along with a picture with an aim to observe brain activation when participants retrieved relevant meaning using visual information. Sixteen subjects were presented images (six per concept) corresponding to 180 concepts (abstract + concrete), while fMRIs were acquired. Out of 180 concepts, 116 are concrete, and others are abstract. Here, we augmented the image captions using the concept word associated with each image in the picture view. As in (Pereira et al., 2018), we focused on nine brain regions corresponding to four brain networks: Default Mode Network (DMN) (linked to the functionality of semantic processing), Language Network (related to language processing, understanding, word meaning, and sentence comprehension, Task Positive Network (related to attention, salience information), and Visual Network (related to the processing of visual objects, object recognition). We briefly summarize the details of the dataset and the number of voxels corresponding to each region in Table 2 in the Appendix.

3 Task Descriptions

For both the datasets, we train fMRI encoding models using Ridge regression on stimuli representations obtained using a variety of models as shown in Fig. 1. The main goal of each fMRI encoder model is to predict fMRI voxel values for each brain region given a stimuli. In all cases, we train a model per subject separately. Different brain regions are involved in the processing of stimuli involving objects and scenes. Similarly, some regions specialize in understanding vision inputs while others interpret linguistic stimuli better. To understand the generalizability of our models across these cognitive aspects (objects vs. scenes, language vs. vision), we conduct the following experiments. Whenever we train and test on the same dataset, we follow K-fold (K=10) cross-validation. All the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold.

Full dataset fMRI Encoding: For each dataset, we perform K-fold (K=10) cross-validation.

Cross-validated fMRI Encoding: In the BOLD5000 dataset, we have three sub-datasets: COCO, ImageNet, and Scenes. For each of the three sub-datasets, we perform K-fold (K=10) cross-validation within the sub-dataset.

ImageNet images mainly contain objects. Scenes images are about natural scenes, while COCO images relate to both objects and scenes. To evaluate the generalizability of our models across objects vs. scenes understanding, we also perform cross-validated experiments where the train images belong to one sub-dataset while the test images belong to the other sub-dataset. Thus, for each subject, we perform (1) three same-sub-dataset train-test experiments and (2) six cross-sub-dataset train-test experiments.

Abstract vs Concrete Concept fMRI Encoding: Similarly, in the Pereira dataset, we have two sub-datasets: abstract and concrete. Intuitively, concrete images can be interpreted mainly using visual processing, while abstract images may require linguistic processing. Hence, we experiment with two different settings for each subject: (train on abstract, test on concrete) and (train on concrete, test on abstract).

4 Methodology

We trained a ridge regression based encoding model to predict the fMRI brain activity associated with the stimuli representation for each brain region. Each voxel value is predicted using a separate ridge regression model. Formally, we encode the stimuli as $X \in \mathbb{R}^{N \times D}$ and brain region voxels $Y \in \mathbb{R}^{N \times V}$, where N denotes the number of training examples, D denotes the dimension of input stimuli representation, and V denotes the number of voxels in a particular region.

The input stimuli representation can be obtained using any of the following models: (i) pretrained CNNs, (ii) pretrained text Transformers (iii) image Transformers, (iv) late-fusion models, or (v) multi-modal Transformers. The ridge regression objective function for the i^{th} example is given by Eq. 1.

$$f(X_i) = \min_W \|Y_i - X_i W\|_F^2 + \lambda \|W\|_F^2 \tag{1}$$

Here, W are the learnable weight parameters, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. λ was tuned on a small disjoint validation set obtained from the training part.

In the following text, we discuss different input stimuli representation methods. Pretrained CNNs and Image Transformers encode image stimuli only, while Pretrained text Transformers encode text stimuli only. Late fusion models and Multi-modal Transformers encode both text and image stimuli.

Pretrained CNNs: Inspired by the Algonauts challenge (Cichy et al., 2019), we extract the layer-wise features from different pretrained CNN models such as VGGNet19 (Simonyan & Zisserman, 2014) (MaxPool1, MaxPool2, MaxPool3, MaxPool4, MaxPool5, FC6, FC7, FC8), ResNet50 (He et al., 2016) (Block1, Block2, Block3, Block4, FC), InceptionV2ResNet (Szegedy et al., 2017) (Conv2D5, Conv2D50, Conv2D100, Conv2D150, Conv2D200, Conv2D_7b), and EfficientNetB5 (Tan & Le, 2019) (Conv2D2, Conv2D8, Conv2D16, Conv2D24, FC) use in predicting fMRI brain activity. Here, we use adaptive average pooling on each layer to get feature representation for each image.

Pretrained text Transformers: RoBERTa (Liu et al., 2019) builds on BERT's language masking strategy and has been shown to outperform several other text models on the popular GLUE NLP benchmark. We use the average-pooled representation[2] from RoBERTa to encode text stimuli.

[2] Average-pooled representation gave us better results compared to using the CLS representation.

Image Transformers: We used three image Transformers: Vision Transformer (ViT), Data Efficient Image Transformer (DEiT), and Bidirectional Encoder representation from Image Transformer (BEiT). Given an image, image Transformers output two representations: pooled and patches. We experiment with both representations.

Late-fusion models: In these models, the stimuli representation is obtained as a concatenation of image stimuli encoding obtained from pretrained CNNs and text stimuli encoding obtained from pretrained text Transformers. Thus, we experiment with these late-fusion models: VGGNet19+RoBERTa, ResNet50+RoBERTa, InceptionV2ResNet+RoBERTa and EfficientNetB5+RoBERTa. Previously proposed methods like StepEnCog (Oota et al., 2019) is also a late-fusion model. These models do not incorporate information fusion across modalities.

Multi-modal Transformers: We experimented with these multi-modal Transformer models: Contrastive Language-Image Pre-training (CLIP), Learning Cross-Modality Encoder Representations from Transformers (LXMERT), Vision-and-Language BERT (ViLBERT), and VisualBERT. These Transformers take both image and text stimuli as input and output a joint visio-linguistic representation. Specifically, the image input for these models comprises of region proposals as well as bounding box regression features extracted from Faster R-CNN (Ren et al., 2015) as input features as shown in Fig. 1. These models incorporate information fusion across modalities at different levels of processing using co-attention and hence are expected to result in high quality visio-linguistic representations.

Hyper-parameter Setting: We used sklearn's ridge-regression with default parameters, 10-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for Transformer models, MSE loss function, and L2-decay (λ) as 1.0. We used Word-Piece tokenizer for the linguistic Transformer input and Faster-RCNN (Ren et al., 2015) for extracting region proposals. All experiments were conducted on a machine with 1 NVIDIA GEFORCE-GTX GPU with 16GB GPU RAM. We make our code publicly available[1].

5 Experiments

5.1 Evaluation Metrics

We evaluate our models using popular brain encoding evaluation metrics described in the following. Given a subject and a brain region, let N be the number of samples. Let $\{Y_i\}_{i=1}^N$ and $\{\hat{Y}_i\}_{i=1}^N$ denote the actual and predicted voxel value vectors for the i^{th} sample. Thus, $Y \in R^{N\times V}$ and $\hat{Y} \in R^{N\times V}$ where V is the number of voxels in that region. Let $\{\hat{E}_i\}_{i=1}^N$ denote the stimuli representation for the i^{th} sample. Thus, $E \in R^{N\times R}$ where R is the dimensionality of the encoded representation.

2V2 Accuracy is computed as shown in Eq. 2.

$$\text{2V2Acc} = \frac{1}{N_{C_2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I[cosD(Y_i, \hat{Y}_i) + cosD(Y_j, \hat{Y}_j) < cosD(Y_i, \hat{Y}_j) + cosD(Y_j, \hat{Y}_i)] \quad (2)$$

where $cosD$ is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if c is true, else it is 0. The higher the 2V2 accuracy, the better.

Pearson Correlation (PC) is computed as PC=$\frac{1}{N}\sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$ where corr is the correlation function.

Representation Similarity Analysis (RSA) is computed as RSA=$corr(YY^T, EE^T)$ where YY^T and EE^T are called Representation Similarity Matrices (RSMs).

5.2 Do multi-modal Transformers outperform other models?

Unfortunately, there is no previous work that uses image Transformers or multi-modal Transformers for brain encoding. StepEnCog (Oota et al., 2019) is a late-fusion method, but it has a different setting where the model expects voxel values per brain slice rather than per brain region. Besides performing extensive evaluation using a large variety of models, we also compare our results with those obtained by two previously proposed baselines that leverage pretrained CNN models: (Blauch et al., 2019) and (Wang et al., 2019) which use VGGNet.

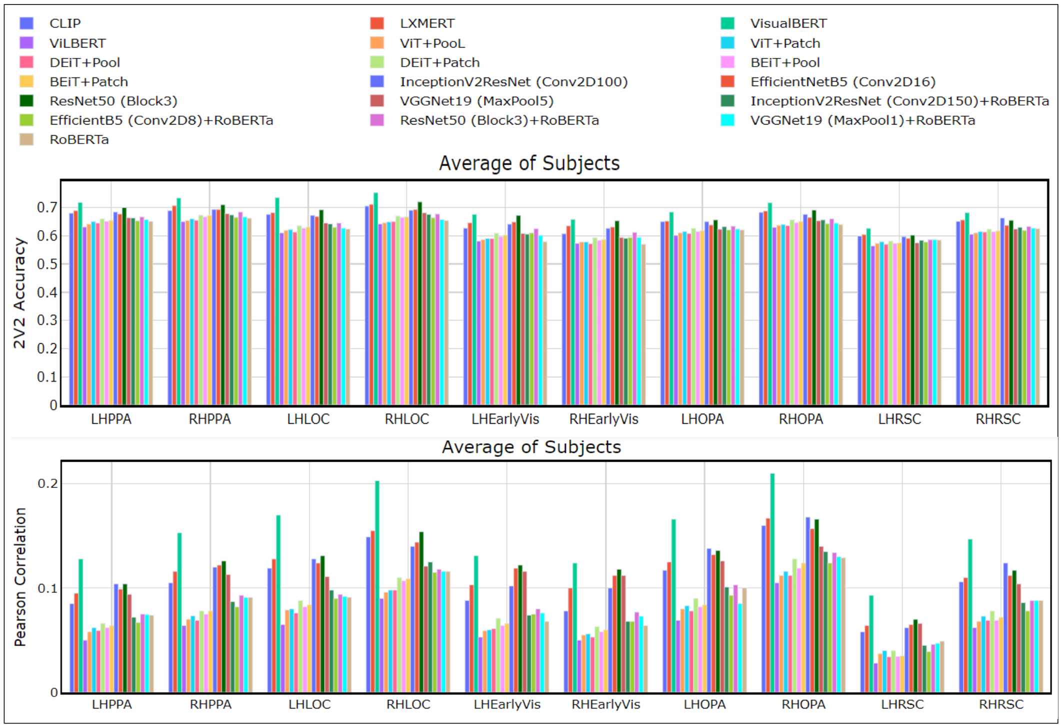

Figure 2: BOLD5000 Results: 2V2 (top figure) and Pearson correlation coefficient (bottom figure) between predicted and true responses across different brain regions using a variety of models. Results are averaged across all participants. VisualBERT, LXMERT, and CLIP perform the best.

We present the 2V2 accuracy and Pearson correlation results for models trained with different input representations (features extracted from the best performing layer of every pretrained CNN model and last output layer of transformer model) on the two datasets: BOLD5000 and Pereira in Figs. 2 and 3, respectively. We also present the results using intermediate layer activations for all the models in the Appendix (Please refer Figs. 9, 10, 11, and 12).

BOLD5000: We make the following observations from Fig. 2: (1) Multi-modal Transformers such as VisualBERT, LXMERT, and CLIP show better performance than uni-modal image Transformers, pretrained CNNs, late-fusion representations, and RoBERTa. (2) On both 2V2 accuracy and Pearson correlation, VisualBERT is better across all the models. (3) Late visual areas such as OPA (scene-related) and LOC (object-related) display a higher correlation with multi-modal Transformers which is inline with the visual processing hierarchy. In general, a higher correlation with all the visual brain ROIs with multi-modal Transformers demonstrates the power of jointly encoding visual and language information. (4) The Patch representation of image Transformers shows an improved 2V2 accuracy and Pearson correlation compared to the Pooled representation. (5) Both InceptionV2ResNet and ResNet-50 have better performance among uni-modality models.

Pereira: We make the following observations from Fig. 3: (1) Similar to BOLD5000, multi-modal Transformers such as VisualBERT and LXMERT perform better. (2) Lateral visual areas such as Vision_Object, Vision_Body, Vision_Face, and Vision areas display higher correlation with multi-modal Transformers. A higher correlation with all the visual brain regions, Language regions, DMN, and TP with multi-modal Transformers, demonstrate that the alignment of visual-language understanding helps.

Further, we show RSA analysis for both the datasets in Fig. 4. On average, the VisualBERT (VB) model has a high correlation across various brain regions. This is why VisualBERT leads to the highest brain encoding accuracy across both datasets. In Fig. 5, we show the mean absolute error (MAE) between the actual and predicted voxels across brain regions using VisualBERT. Comparing with similar brain charts for other models (shown in Figs. 13 and 14 in the Appendix), we notice that the error magnitudes are very small for the majority of the voxels. We also report the layer-wise RSA scores for both the datasets using all the models in the Appendix (Please refer Figs. 15 and 16).

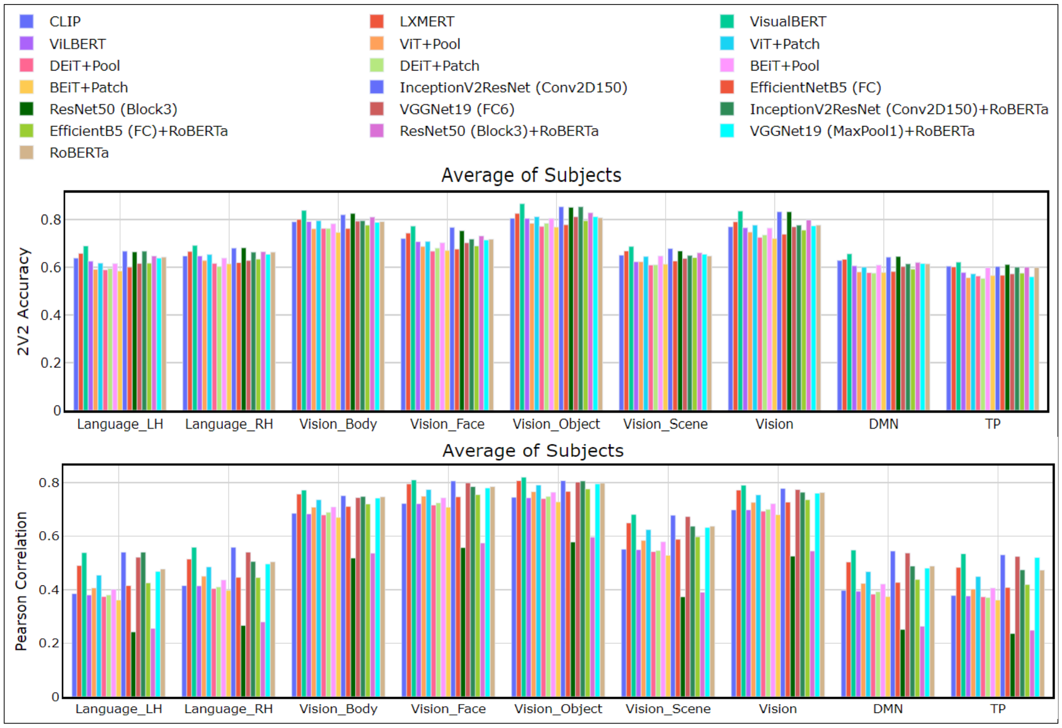

Figure 3: Pereira Results: 2V2 (top figure) and Pearson correlation coefficient (bottom figure) between predicted and true responses across different brain regions using a variety of models. Results are averaged across all participants. VisualBERT, LXMERT, and InceptionV2ResNet perform the best.

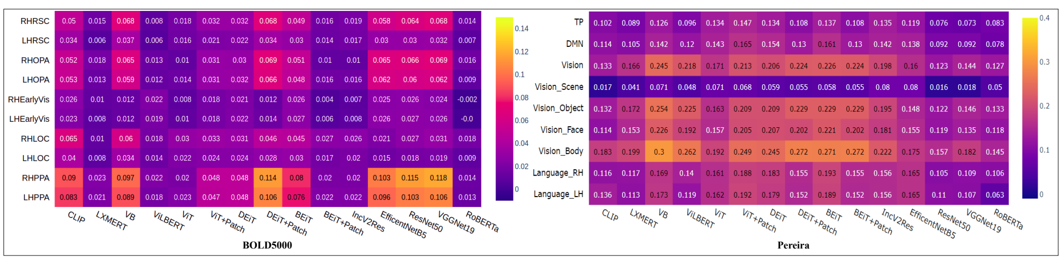

Figure 4: Representation Similarity Analysis: Pearson correlation between RSA matrices of model activations and brain voxel values.

5.3 Do image Transformers act like human visual system?

Given the hierarchical processing of visual stimulus across the image-based Transformers (ViT, DEiT, and BEiT), we further examine how these layer-wise representations correlate with voxel representations of each brain region using RSA scores. We show the RSA scores for three models: best for visio-linguistic Transformers (VisualBERT), best for image Transformers (DEiT), best for pretrained CNNs (InceptionV2ResNet) in Figs. 6 and 7 for BOLD5000 and Pereira resp. as a heatmap. We observe clear differences between the internal representation structure between the three model architectures in the BOLD5000 experiments: (1) VisualBERT shows a much more uniform similarity structure from lower to intermediate layers (L1-L9) across the brain regions. (2) Higher layers in DEiT show much greater RSA similarity than lower layers except EarlyVisual areas. (3) The early visual areas (LHEarlyVis and RHEarlyVis) show higher RSA scores with initial layers across the three models. (4) The intermediate blocks of InceptionV2ResNet (Conv2D100/L3 and Conv2D150/L4) have higher RSA with brain regions. (5) VisualBERT has higher RSA with every brain region compared to DEiT and InceptionV2ResNet. Qualitatively similar findings are obtained in the experiments with the Pereira dataset: (i) All three models have higher RSA scores for visual areas (vision, object, face, and body), possibly indicating their ability to capture visual information

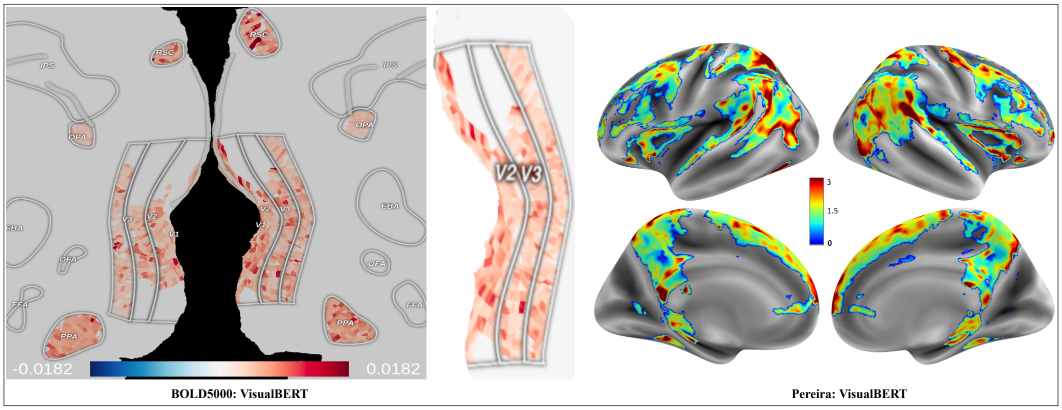

Figure 5: MAE between actual and predicted voxels: (a) left figure is zoomed on V2 and V3 brain areas for VisualBERT on BOLD5000. Note that V1 and V2 are also called EarlyVis area, while V3 is also called LOC area. (b) the right figure is for VisualBERT on the Pereira dataset.

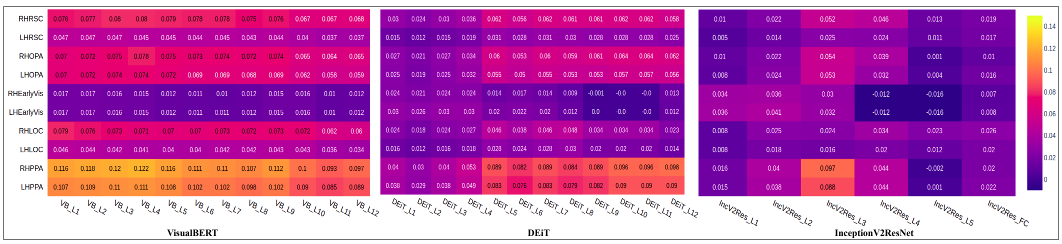

Figure 6: Representation Similarity Analysis for BOLD5000: Pearson correlation between RSA matrices of model activations and brain voxel values.

(ii) On the other hand, only VisualBERT shows additionally higher RSA scores in the Language areas (Language_LH/RH) too, indicating the joint alignment between vision and language.

5.4 Cross-Validated fMRI Encoding

Fig. 8(a) illustrates Pearson correlation for cross-validated encoding on BOLD5000 using three multi-modal Transformers (VisualBERT, LXMERT, and CLIP). We also show results for a baseline method (Blauch et al., 2019). We observe that (1) multi-modal Transformers outperform the baseline results across all the ten brain regions for all the cross-validated tasks. (2) The Pearson correlation score is higher for train on COCO and test on ImageNet in the object-selective visual area LOC (lateral occipital cortex), which makes sense since COCO has many objects. (3) Similarly, the scene-selective brain areas such as RSC and OPA have a higher correlation for the COCO-Scenes, ImageNet-Scenes, and Scenes-Scenes tasks. (4) EarlyVisual areas have a lower correlation compared to other brain regions across the three tasks. (5) Overall, the models trained on COCO or ImageNet report higher correlation rather than those trained on Scenes.

5.5 Abstract-Concrete fMRI Encoding

The results for the abstract-train-concrete-test and concrete-train-abstract-test encoder models are presented across brain regions using two best multi-modal Transformers (VisualBERT and LXMERT), and the best pretrained CNN model (InceptionV2ResNet) in Fig. 8(b). We observe that the concrete-train-abstract-test model provides a better Pearson correlation compared to the abstract-train-concrete-test model. This matches our expectation that our brain can learn much better from concrete concepts than abstract concepts. The Pearson correlation analysis across brain regions provides the following insights. (1) The visual brain areas such as Vision_Body, Vision_Face, Vision_Object, and Vision have superior performance for both concrete and abstract concepts; surprisingly, this is not the case for the Vision_Scene area. (2) The language, DMN, and Task Positive (TP) brain networks have a higher correlation in the concrete-train-abstract-test than the abstract-train-

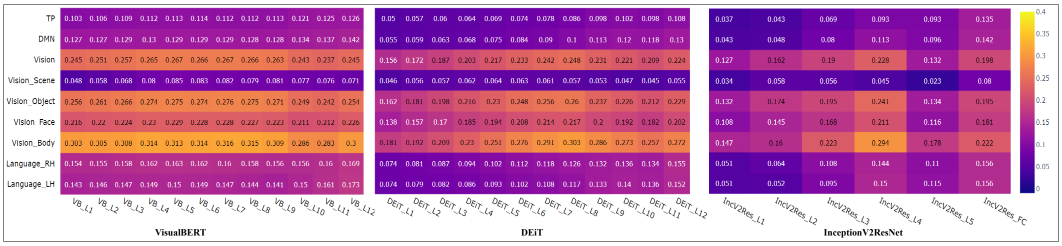

Figure 7: Representation Similarity Analysis for Pereira: Pearson correlation between RSA matrices of model activations and brain voxel values.

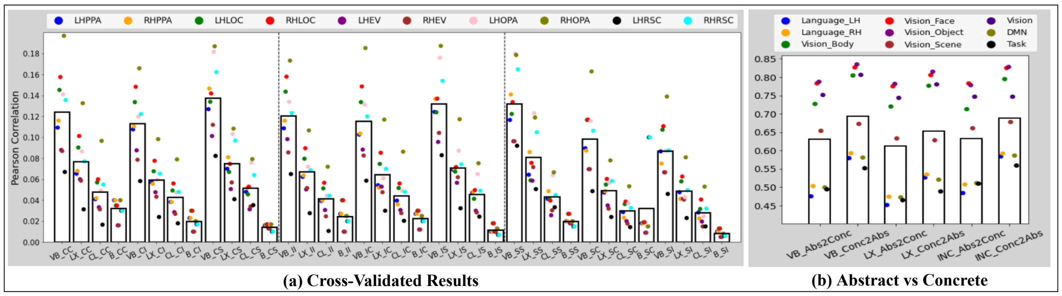

Figure 8: (a) Cross-Validated Results for BOLD5000 dataset. (b) Abstract-Concrete Results for Pereira dataset. VB=VisualBERT, LX=LXMERT, CL=CLIP, B=Baseline (Blauch et al., 2019), INC=InceptionV2ResNet. CC=Train and test on COCO, CI=Train on COCO and test on ImageNet, CS=Train on COCO and test on Scenes, and so on.)

concrete-test model. (3) Overall, VisualBERT and InceptionV2ResNet report similar performance with a slight edge over the LXMERT model.

6 Cognitive Insights: Does Language Influence Vision?

We discussed various insights in detail in Section 5. We summarize cognitive insights in the following. BOLD5000 dataset comprises brain responses from visual areas (early visual, scene-related, and object-related) when visual stimuli are presented to the subjects. Although only visual information is present in the stimuli, it is conceivable that participants implicitly invoke appropriate linguistic representations that in turn influence visual processing (Lupyan et al., 2020). Thus, it is not surprising that computational models such as multi-modal Transformers (VisualBERT, and LXMERT) that learn joint representation of language and vision show superior performance on the 'purely' visual response data in BOLD5000 (see Figs. 2, 5(a) and 6). Further, the performance of these models is naturally good in the case when text and image are shown to the participants, and whole brain responses are captured as in the case of the Pereira dataset (see Figs. 3, 5(b) and 7). Based on the intuition from the computational experiments, we make the following testable prediction for future fMRI experiments. Instead of a passive viewing task, if participants were to perform a naming task/decision-making task on the objects/scenes, we expect to see more pronounced and focused activation in the visual areas with the explicit top-down influence of Language areas during the language-based task as compared to passive viewing.

7 Conclusions

In this paper, we studied the effectiveness of multi-modal modeling for brain encoding. We found that multi-modal visio-linguistic Transformers, which jointly encode text and visual input using cross-modal attention at multiple levels, perform the best. Our experiments on BOLD5000 and Pereira datasets lead to interesting cognitive insights. These insights indicate that fMRIs reveal reliable responses in scenes and object selection visual brain areas, which shows that cross-view translation tasks like image captioning or image tagging are practically possible with good accuracy. We plan to explore this as part of future work.

REFERENCES

Andrew James Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120:309–322, 2015.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: self-supervision in natural-image reconstruction from fmri. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 6517–6527, 2019.

Nicholas M Blauch, Filipe De Avila Belbute Peres, Juhi Farooqui, Alireza Chaman Zar, David Plaut, and Marlene Behrmann. Assessing the similarity of cortical object and scene representations through cross-validated voxel encoding models. *Journal of Vision*, 19(10):188d–188d, 2019.

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18, 2019.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv e-prints*, pp. arXiv–1905, 2019.

R Todd Constable, Kenneth R Pugh, Ella Berroya, W Einar Mencl, Michael Westerveld, Weijia Ni, and Donald Shankweiler. Sentence complexity and input modality effects in sentence comprehension: an fmri study. *NeuroImage*, 22(1):11–21, 2004.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017.

Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 529–539, 2019.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32:12805–12816, 2019.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 13–23, 2019.

Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark. Effects of language on visual perception. *Trends in cognitive sciences*, 2020.

Subba Reddy Oota, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Stepencog: A convolutional lstm autoencoder for near-perfect fmri encoding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *BioRxiv*, 2020a.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2020b.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, 32:14123–14133, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32:15501–15511, 2019.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *in press*, 2014.

Daniel Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. 2013.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

A DATASET STATISTICS

We show number of instances and voxel distribution across various brain regions for the BOLD5000 and Pereira datasets in Tables 1 and 2 respectively.

		Number of Voxels in Each ROI									
ROIs→		PPA		LOC		EarlyVis		OPA		RSC	
↓Subjects	#Instances	LH	RH	LH	RH	LH	RH	LH	RH	LH	RH
Subject-1	5254	131	200	152	190	210	285	101	187	86	143
Subject-2	5254	172	198	327	561	254	241	85	95	59	278
Subject-3	5254	112	161	430	597	522	696	187	205	78	116
Subject-4	3108	157	187	455	417	408	356	279	335	51	142

Table 1: BOLD5000 Dataset Statistics. LH=Left Hemisphere. RH - Right Hemisphere.

	Number of Voxels in Each ROI								
ROIs→	Language		Vision					DMN	Task Positive
↓Subj	LH	RH	Body	Face	Object	Scene	Vision	RH	LH
P01	5265	6172	3774	4963	8085	4141	12829	17190	35120
M01	5716	5561	3934	4246	7357	3606	12075	17000	34582
M02	4930	5861	3873	4782	7552	3173	11729	15070	30594
M03	3616	4247	2838	3459	5956	2822	9074	12555	24486
M04	5906	5401	3867	4803	7812	3602	12278	18011	34024
M05	4607	4837	2961	4023	6609	3135	10417	14096	28642
M06	4993	5099	3424	4374	7300	4058	11986	16289	30109
M07	5629	5001	4190	4993	8617	3721	12454	17020	30408
M08	5083	5062	2624	4082	6463	3503	10439	14950	29972
M09	3513	3650	2876	3343	5992	2815	9003	12469	25167
M10	5458	5581	3232	4844	7445	3474	11530	16424	29400
M13	4963	4811	2675	4008	5809	3323	9848	14489	30608
M15	5315	6141	4112	4941	8323	3496	12383	15995	31610
M16	4726	5534	4141	4669	8060	4142	12503	15104	31758
M17	5854	5698	4416	4801	8831	4521	13829	16764	37463

Table 2: Pereira Dataset Statistics

B Do multi-modal Transformers perform better encoding compared to intermediate layer representations from pretrained CNNs?

We present the 2V2 accuracy and Pearson correlation for models trained with representations extracted from the last layer of multi-modal Transformers and all the lower to higher-level representations from pretrained CNNs on the two datasets: BOLD5000 and Pereira in Figs. 9 and 10, respectively.

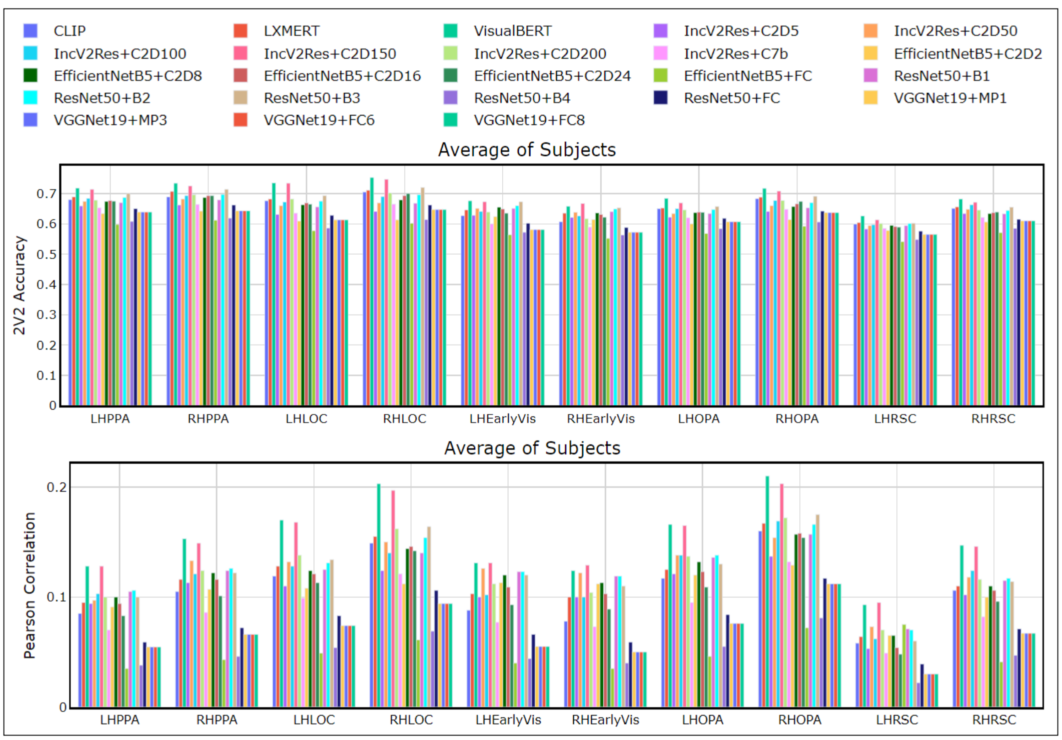

Figure 9: BOLD5000: 2V2 (top Fig.) and Pearson correlation coefficient (bottom Fig.) between predicted and true responses across different brain regions using variety of models. Results are averaged across all participants. Pretrained CNN results are shown for all layers while multi-modal Transformer results are shown for last layers only.

We make the following observations from Fig. 9: (1) With respect to 2V2 and Pearson correlation, the multi-modal Transformer, VisualBERT, performs better than all the internal representations of pretrained CNNs. (2) In the pretrained CNNs, intermediate blocks have better correlation scores as compared to lower or higher level layer representations. (3) Other multi-modal Transformers, CLIP,

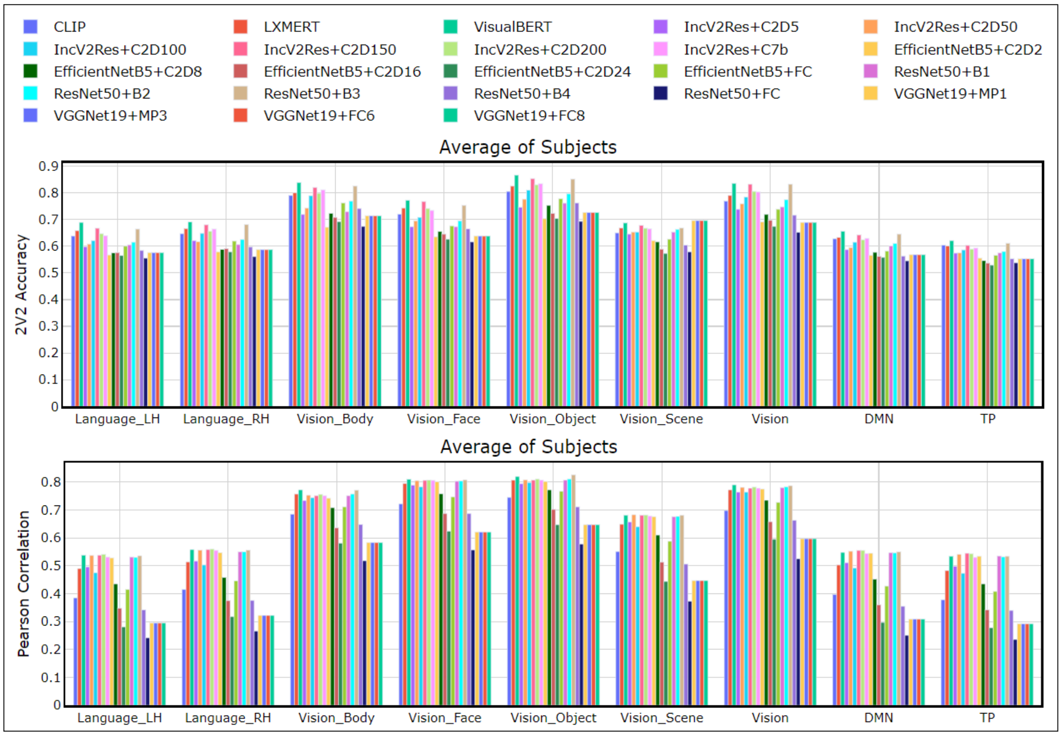

Figure 10: Pereira dataset: 2V2 (top Fig.) and Pearson correlation coefficient (bottom Fig.) between predicted and true responses across different brain regions using variety of models. Results are averaged across all participants. Pretrained CNN results are shown for all layers while multi-modal Transformer results are shown for last layers only.

and LXMERT, have marginal improvements over all the models except intermediate blocks such as Conv2D150 in InceptionV2ResNet.

We make the following observations from Fig. 10: (1) With respect to 2V2 and Pearson correlation, the multi-modal Transformer, VisualBERT, performs better than all the internal representations of pretrained CNNs. (2) Similar to BOLD5000, the intermediate blocks have better correlation scores as compared to lower or higher level layer representations in the pretrained CNNs on Pereira Dataset. (3) Other multi-modal Transformer, LXMERT, have equal performance with intermediate blocks of each pretrained CNN model.

C Do multi-modal Transformers perform better encoding in their layers?

Given the hierarchical processing of visual or visual-language information across the Transformer layers, we further examine how these Transformer layers encode fMRI brain activity using image and mulit-modal Transformers. We present the layer-wise encoding performance results on two datasets: BOLD5000 and Pereira in Figs. 11 and 12, respectively.

We make the following observations from Fig. 11: (i) The multi-modal Transformer, VisualBERT, have consistent performance across the layers from 1 to 10. (ii) The LXMERT model have marginal decreasing performance from intermediate layer (L7) to higher layers. (iii) The image Transformers have higher Pearson correlation for early visual areas in the lower layers whereas higher visual areas such as LOC, OPA, and PPA have an increasing correlation in higher layers. (iv) This is clearly indicate that the hierarchy of processing of visual stimulus in the human brain is similar to image Transformer layers.

We make the following observations from Fig. 12: (i) The multi-modal Transformers, VisualBERT, have consistent performance across the layers from 1 to 10. (ii) The LXMERT model have marginal decreasing performance from lower to higher layers. (iii) The image Transformer, ViT, has higher

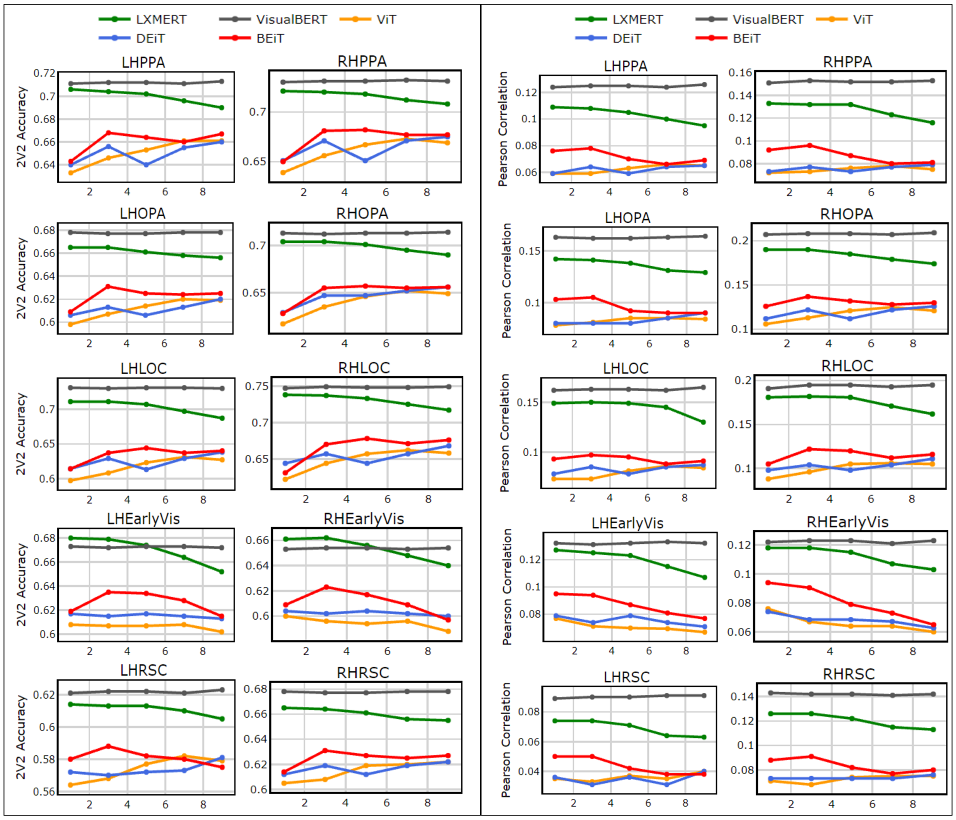

Figure 11: BOLD5000: 2V2 (left) and Pearson correlation coefficient (right) between predicted and true responses across different brain regions using Transformer models. Results are averaged across all participants. The results are shown for all layers of image and multi-modal Transformers.

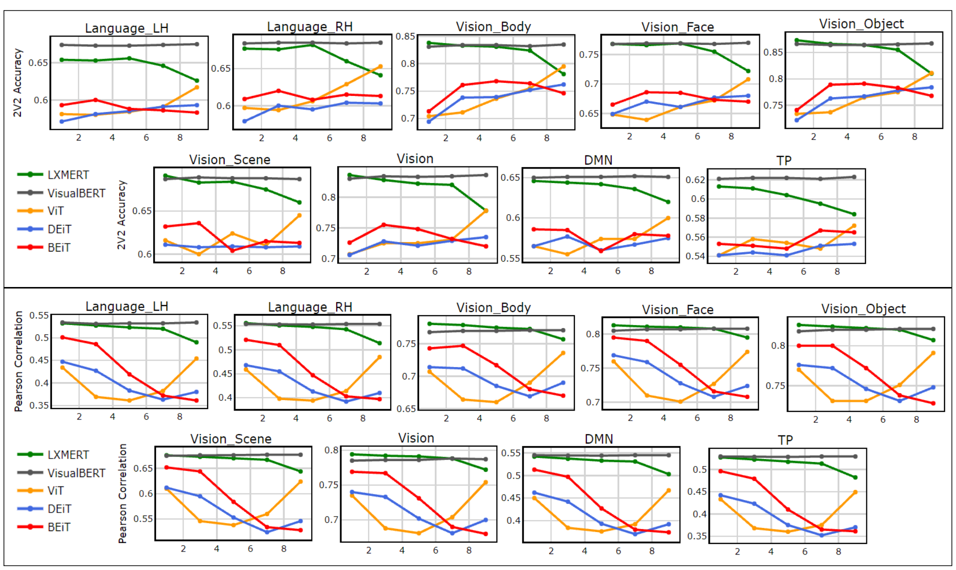

Figure 12: Pereira: 2V2 (left) and Pearson correlation coefficient (right) between predicted and true responses across different brain regions using Transformer models. Results are averaged across all participants. The results are shown for all layers of image and multi-modal Transformers.

Pearson correlation for early visual areas in the lower layers whereas higher visual areas such as Vision_Body, Vision_Face, and Vision_Obj have an increasing correlation in higher layers.

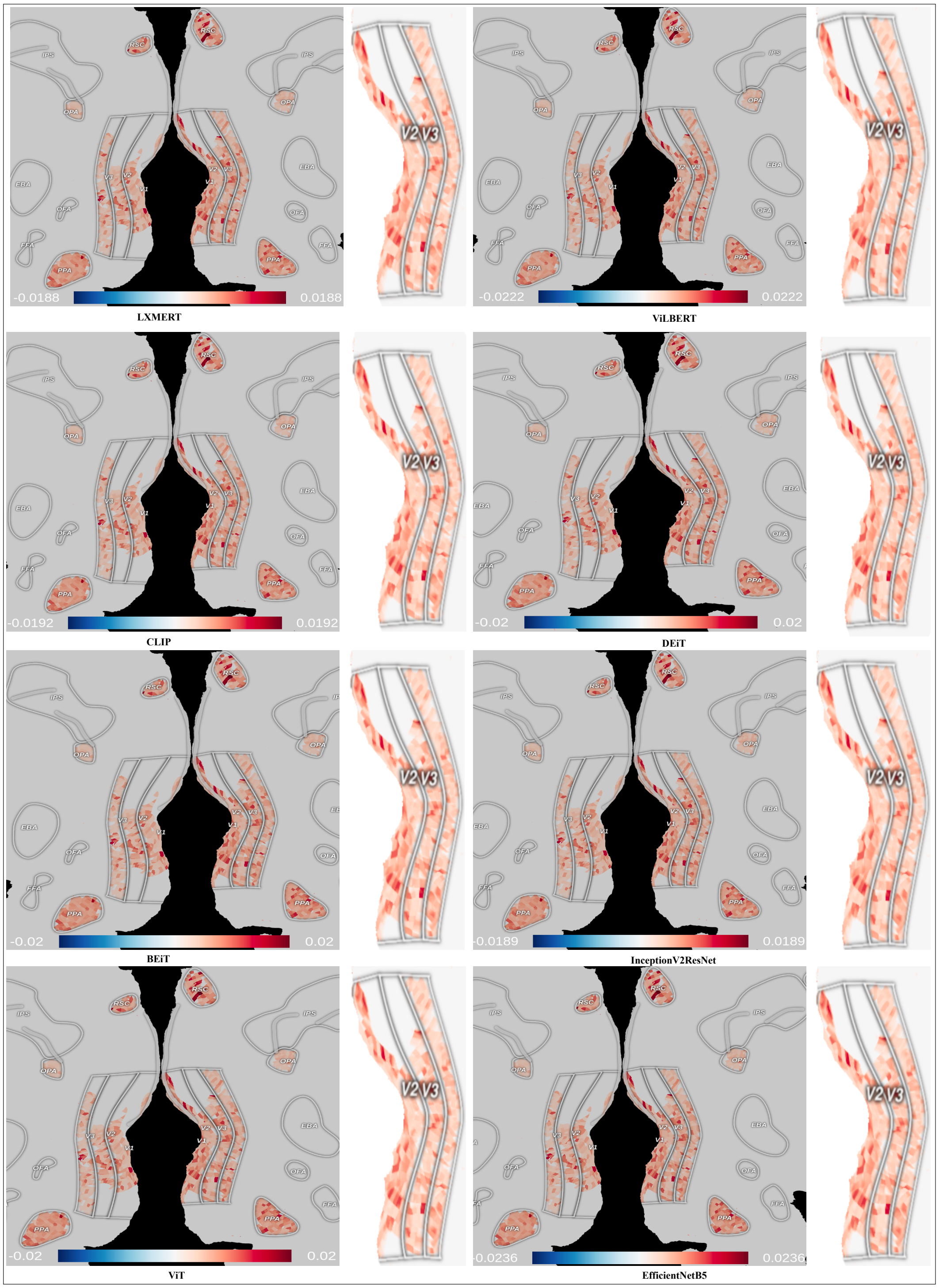

Figure 13: MAE between actual and predicted voxels zoomed on V2 and V3 brain areas for various models. Note that V1 and V2 are also called EarlyVis area, while V3 is also called LOC area.

D Brain Maps for various models for BOLD5000 Dataset

Fig. 13 shows mean absolute errors (MAE) between actual and predicted voxels for various models on the BOLD5000 dataset. Notice that the magnitude of errors is much higher for a majority of voxels, compared to that with the VisualBERT model as shown in Fig. 5(a). Also, the multi-modal

Transformers, VisaulBERT (MAE range: 0 to 0.0181) and LXMERT (MAE range: 0 to 0.0188), have lower MAE compared to both image Transformers (MAE range: 0 to 0.02) and pretrained CNNs (MAE range: 0 to 0.0236).

E Brain Maps for various models for Pereira Dataset

Fig. 14 shows mean absolute errors (MAE) between actual and predicted voxels for various models on the Pereira dataset. Notice that the magnitude of errors is much higher for a majority of voxels, compared to that with the VisualBERT model as shown in Fig. 14(a). Also, the multi-modal Transformers, VisaulBERT and LXMERT, and InceptionV2ResNet+Conv2D150 have lower MAE compared to both image Transformers and other pretrained CNNs.

F RSA for various models on BOLD5000 dataset

In Fig. 15, we show RSA for various models (like LXMERT, ViT, BEiT, EfficientNetB5, ResNet, and VGGNet) on the BOLD5000 dataset. In Fig. 17, we show RSA for the VisualBERT model on the COCO, ImageNet and Scenes sub-datasets of the BOLD5000 dataset.

G RSA for various models on Pereira dataset

In Fig. 16, we show RSA for various models (like LXMERT, ViT, BEiT, EfficientNetB5, ResNet, and VGGNet) on the Pereira dataset.

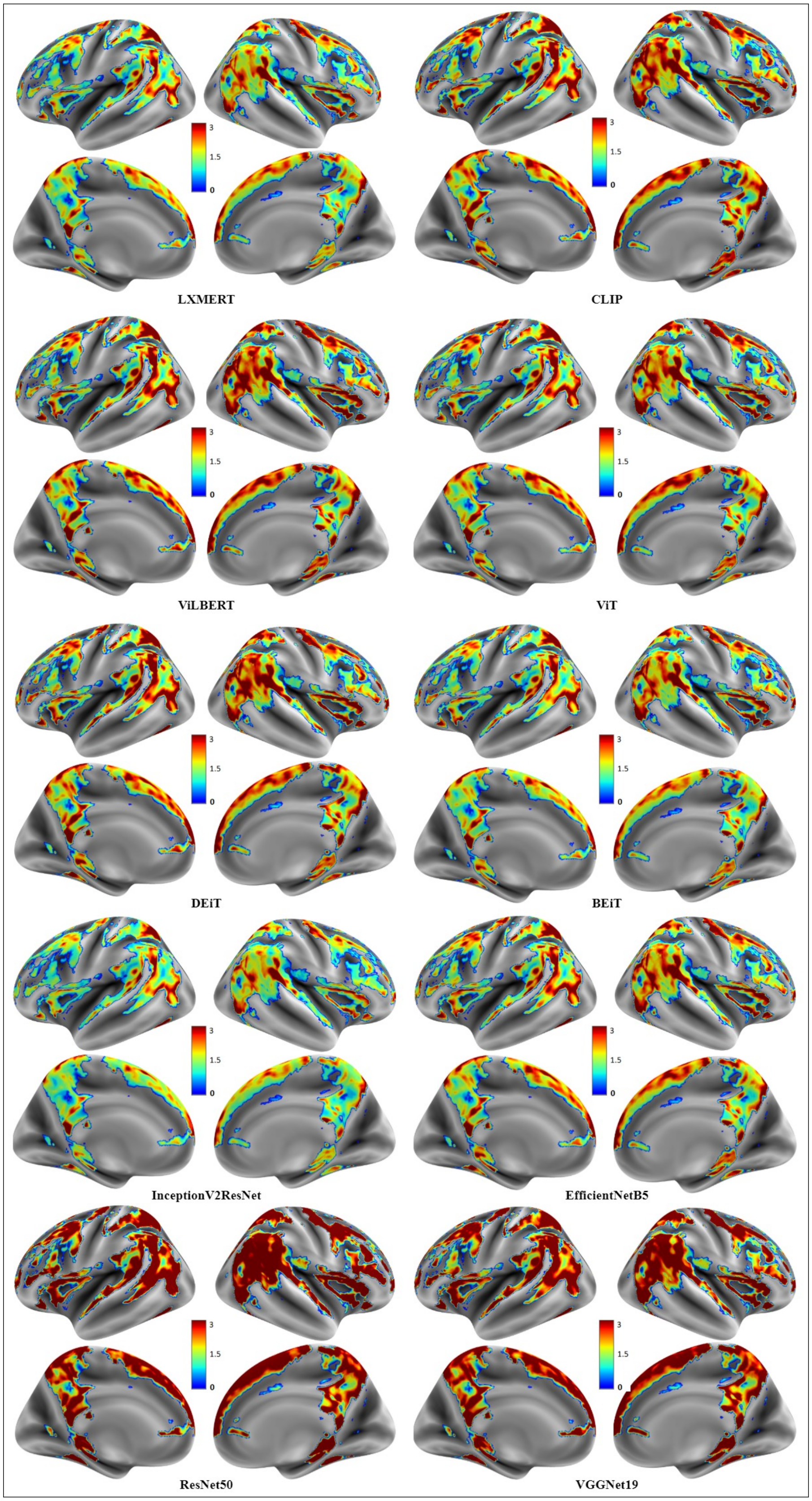

Figure 14: MAE between actual and predicted voxels zoomed on V2 and V3 brain areas for various models. Note that V1 and V2 are also called EarlyVis area, while V3 is also called LOC area.

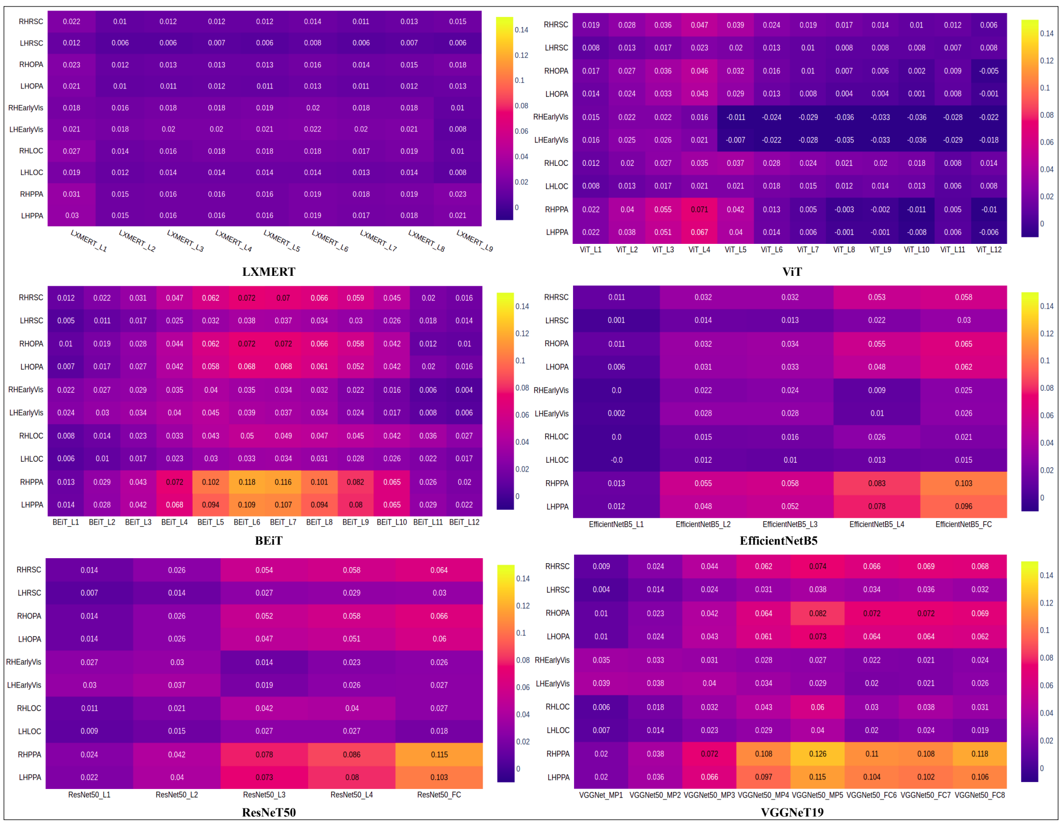

Figure 15: Representation Similarity Analysis for BOLD5000: Pearson correlation between RSA matrices of different models activations and brain voxel values.

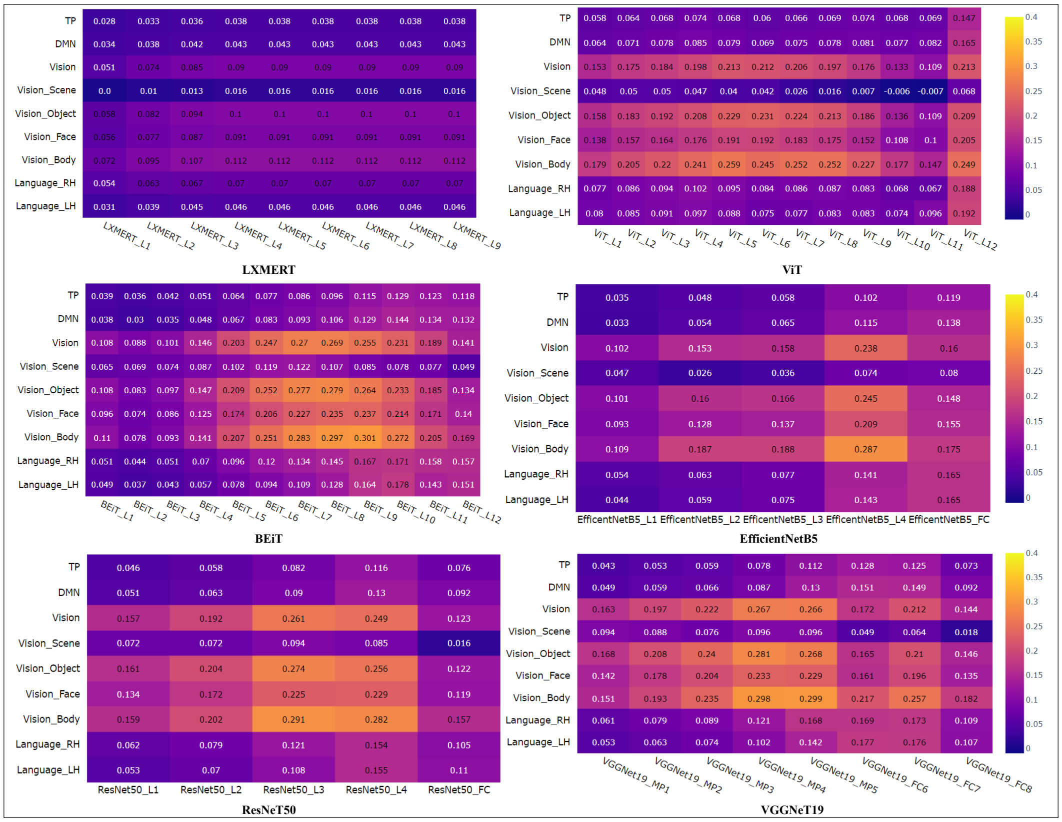

Figure 16: Representation Similarity Analysis for Pereira: Pearson correlation between RSA matrices of different models activations and brain voxel values.

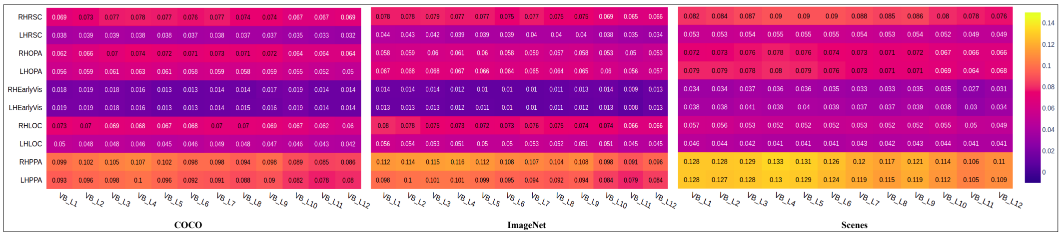

Figure 17: BOLD5000: VisualBERT COCO vs ImageNet vs Scenes RSA