Multi-view and Cross-view Brain Decoding

Subba Reddy Oota, Jashn Arora, Manish Gupta, Raju Surampudi Bapi

Multi-view and Cross-view Brain Decoding

Subba Reddy Oota[1,3]*, **Jashn Arora**[1*] **Manish Gupta**[1,2] **and Bapi Raju Surampudi**[1]
[1]IIIT Hyderabad, India; [2]Microsoft, India; [3]INRIA, Bordeaux, France
subba-reddy.oota@inria.fr, jashn.arora@research.iiit.ac.in
gmanish@microsoft.com, raju.bapi@iiit.ac.in

Abstract

Can we build multi-view decoders that can decode concepts from brain recordings corresponding to any view (picture, sentence, word cloud) of stimuli? Can we build a system that can use brain recordings to automatically describe what a subject is watching using keywords or sentences? How about a system that can automatically extract important keywords from sentences that a subject is reading?

Previous brain decoding efforts have focused only on single view analysis and hence cannot help us build such systems. As a first step toward building such systems, inspired by Natural Language Processing literature on multilingual and cross-lingual modeling, we propose two novel brain decoding setups: (1) multi-view decoding (MVD) and (2) cross-view decoding (CVD). In MVD, the goal is to build an MV decoder that can take brain recordings for any view as input and predict the concept. In CVD, the goal is to train a model which takes brain recordings for one view as input and decodes a semantic vector representation of another view. Specifically, we study practically useful CVD tasks like image captioning, image tagging, keyword extraction, and sentence formation.

Our extensive experiments lead to MVD models with ∼0.68 average pairwise accuracy across view pairs and CVD models with ∼0.8 average pairwise accuracy across tasks. Analysis of the contribution of different brain networks reveals exciting cognitive insights: (1) Models trained on picture or sentence view of stimuli are better MV decoders than a model trained on word cloud view. (2) Our extensive analysis across 9 broad brain regions, 11 language sub-regions, and 16 visual sub-regions of the brain help us localize, for the first time, the parts of the brain involved in cross-view tasks like image captioning, image tagging, sentence formation, and keyword extraction. We make the code publicly available[1].

1 Introduction

Brain decoding models aim to understand what a subject is thinking, seeing, and perceiving by analyzing neural recordings. Thus, in the context of language, it may be beneficial to learn mappings between linguistic representation and the associated brain activation, and how we compose the linguistic meaning from different stimuli such as text (Pereira et al., 2018; Wehbe et al., 2014a), images (Eickenberg et al., 2017; Beliy et al., 2019), videos (Huth et al., 2016; Nishimoto et al., 2011), or speech (Zhao et al., 2014) by analyzing the evoked brain activity. Also, decoding the functional activity of the brain has numerous applications in education and healthcare.

Brain recordings can be obtained by providing stimuli to a subject in various forms. For example, a concept (like *apartment*) can be presented using: (1) Word Picture (WP) view: picture along with the concept word, (2) Sentence (S) view: sentence containing the word, or (3) Word cloud (WC) view: word cloud containing the word along with other semantically related words. Recent studies have made much progress using functional magnetic resonance imaging (fMRI) brain activity to reconstruct semantic vectors corresponding to linguistic items, including words (Mitchell et al., 2008; Pereira et al., 2018), phrases, sentences, and paragraphs (Wehbe et al., 2014a). However, all such studies have been limited to single-view analysis. Separate models are trained to process different views. Also, the decoding target is typically a semantic vector of the concept word.

In the Natural Language Processing (NLP) community, researchers have recently started focusing on building multi-lingual and cross-lingual systems (Conneau et al., 2018; Conneau and Lample, 2019; Xue et al., 2021). Multi-lingual systems

*The first two authors made equal contribution.

[1]https://tinyurl.com/MVCVBD

Proceedings of the 29th International Conference on Computational Linguistics, pages 105–115
October 12–17, 2022.

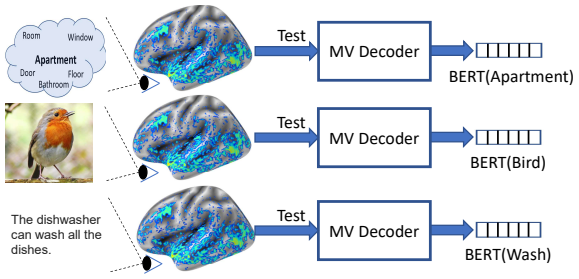

Figure 1: A multi-view decoder can be used to decode concepts using brain recordings for any view. Target is BERT representation of the concept word.

improve accuracy for low-resource languages and enable applications even in the absence of training data for low-resource languages. Cross-lingual systems take input in one language and produce output (e.g., summary) in another language. Inspired by this multi-lingual/cross-lingual shift in NLP, we propose two novel brain decoding setups: multi-view decoding (MVD) and cross-view decoding (CVD). Such setups are critical to build MV decoders which can decode concepts from brain recordings corresponding to any view (picture, sentence, word cloud) of stimuli or systems that can automatically describe using sentences or keywords what a subject is watching or automatically extract important keywords from sentences that a subject is reading.

In MVD, the goal is to build an MV decoder that can take brain recordings for any view as input and predict the concept. Fig. 1 shows examples of using an MV decoder. Such an MV decoder can be trained on data for any specific view. Multi-lingual models have shown huge zero-shot accuracy improvements for inference on low-resource language inputs across many NLP tasks (Conneau and Lample, 2019). Similarly, can we improve decoding accuracy using an MV decoder model for some views?

In CVD, the goal is to train a model which takes brain recordings for one view as input and decodes a semantic vector representation of another view. Fig. 2 shows examples of four such CVD tasks. Given an fMRI activation corresponding to a picture view of the stimuli, how accurately can we decode a sentence representing the picture? Which parts of the brain are involved in CVD tasks like image captioning, image tagging, keyword extraction, and sentence formation?

Historically, the fMRI brain activity has been decoded to a semantic vector representation of a view (word picture, sentence, word cloud) using a ridge-regression decoder (Pereira et al., 2018; Sun et al., 2019). In particular, earlier brain decoding works focused on hand-crafted features to train such decoder models (Mitchell et al., 2008; Wehbe et al., 2014a). Recently, many studies have shown accurate results in mapping the brain activity using neural distributed word embeddings for linguistic stimuli (Anderson et al., 2017; Pereira et al., 2018; Oota et al., 2018; Nishida and Nishimoto, 2018; Sun et al., 2019). To represent meaning, these studies use either word or sentence level embeddings extracted from the models trained on large corpora. Unfortunately, none of these addresses the open questions around multi-view decoding and cross-view decoding. Recently, Transformer-based models have been explored for brain encoding (Hollenstein et al., 2019), which inspires us to harness Transformer-based models like BERT (Devlin et al., 2019) for our brain decoding tasks.

Our main contributions are as follows. (1) We propose two novel brain decoding settings: multi-view decoding and cross-view decoding. (2) We build decoder models using Transformer-based methods and analyze brain network contributions across multi-view and cross-view tasks. (3) We augment the popular Pereira et al. (2018)'s dataset with pairwise-view relationships and use it to demonstrate the efficacy of our proposed methods. We make the code publicly available[2].

Our experiments lead us to the following insights: (1) Models trained on picture or sentence view are better MV decoders than models trained on word cloud view. Surprisingly, the MV decoder trained on sentence view leads to a zero-shot accuracy for word cloud stimuli, which is better than that obtained using the same-view word cloud model. (2) For the first time, we show language and visual sub-regions involved in four cross-view tasks. (3) High pairwise accuracies of 0.78, 0.83, 0.84, and 0.75 for image captioning, image tagging, keyword extraction, and sentence formation resp., help us conclude that cross-view decoding tasks using fMRI data are practically feasible.

2 Related Work

Advances in functional neuroimaging tools such as fMRI have made it easier to study the relationship between language/visual stimuli and functions of brain networks (Constable et al., 2004; Thirion et al., 2006; Fedorenko et al., 2010).

[2] https://tinyurl.com/MVCVBD

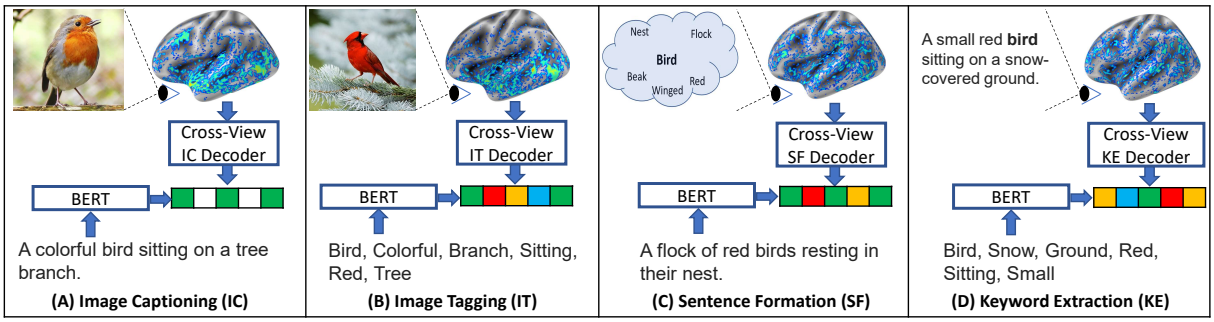

Figure 2: Cross-View Decoding Task (Input, output) Examples.

Initial brain decoding experiments studied the recovery of simple concrete nouns and verbs from fMRI brain activity (Mitchell et al., 2008; Palatucci et al., 2009; Nishimoto et al., 2011; Pereira et al., 2011) where the subject watches either a picture or a word. Unlike the earlier work, Wehbe et al. (2014a); Huth et al. (2016) built a model to decode the text passages instead of individual words. However, these studies used either simple or constrained sets of stimuli, which poses a question of generalization of these models. Recently, Pereira et al. (2018) explicitly decoded both words and sentences when subjects were shown both concrete and abstract stimuli. Affolter et al. (2020) reconstructed the sentences along with categorizing words or predicting the semantic vector representation from fMRI brain activity. Schwartz et al. (2019); Wang et al. (2020a) focused on understanding how multiple tasks activate associated regions in the brain.

To train ridge regression decoder models, earlier works focused on hand-crafted features (Mitchell et al., 2008; Wehbe et al., 2014a), which suffer from various drawbacks like inability to capture the context and sequential aspects of a sentence, inability to extract signals for abstract stimuli, etc. With the success of deep learning based word representations, multiple researchers have used distributed word embeddings for brain decoding models in place of carefully hand-crafted feature vectors (Huth et al., 2016; Anderson et al., 2017; Pereira et al., 2018; Oota et al., 2018; Nishida and Nishimoto, 2018; Sun et al., 2019; Wang et al., 2020b). Using the distributed sentence representations, Wehbe et al. (2014b); Jain and Huth (2018); Abnar et al. (2019); Sun et al. (2019) demonstrated that neural sentence representations are better for decoding whole sentences from brain activity patterns. Recently, Transformer models like BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019) have been found to be very effective for decoding (Gauthier and Levy, 2019; Toneva and Wehbe, 2019; Affolter et al., 2020). Inspired by such studies, we leverage BERT representations. Inspired by such studies, we leverage BERT representations. Unlike single-view analysis done in previous studies, multi-view and cross-view setups are the main focus of our work.

3 Methodology

3.1 Brain Imaging Dataset

We experiment with the popular dataset from (Pereira et al., 2018). It is obtained from 11 subjects (P01, M01, M02, M04, M07, M09, M10, M13, M15, M16, M17) where each subject read 180 concept words (abstract + concrete) in three different paradigms or views while functional magnetic resonance images (fMRI) were acquired. These contain 128 nouns, 22 verbs, 29 adjectives and adverbs, and 1 function word. In paradigm-1 (WP), participants were shown concept word along with picture with an aim to observing brain activation when participants retrieved relevant meaning using visual information. In paradigm-2 (S), the concept word presented in a sentence allows us to probe activity in the language areas associated with contextual information and meaning of a sentence. In paradigm-3 (WC), the concept word was presented in a word cloud format, surrounded by five semantically similar words. These paradigms provide brain representation of 180 concepts in three different views.

For each of the 180 concepts, the dataset contains five pictures, six sentences each containing the concept word, and a word cloud. For example, for a concept 'bird', dataset has (1) a picture p showing a red bird sitting on a tree branch, (2) sentence s like "A green bird flying in the sky", and (3) word cloud c with words "bird, purple, flock, winged, nest, beak". The dataset also has fMRIs for each of these three views. This dataset was

Task	Input	Output (View type)
Image captioning	Word+Picture fMRI	Caption (Sentence)
Image tagging	Word+Picture fMRI	Image tags (Word Cloud)
Keyword extraction	Sentence fMRI	Keywords (Word Cloud)
Sentence formation	Word-cloud fMRI	Sentence

Table 1: Cross-View Decoding Task Definitions

meant for single-view decoding and hence follows a star schema (concept at the center and specific views like word+picture, sentence, and word cloud around it). Clearly, we cannot use this dataset as is for cross-view decoding (CVD). For example, for the image captioning CVD task, it is wrong to take an fMRI with the stimuli being a picture showing a red bird sitting on a tree branch, and use it to decode a sentence "A green bird flying in the sky".

To enable cross-view decoding tasks, it was critical to build direct pairwise-view relationships (picture-sentence, picture-word cloud, sentence-word cloud, and word cloud-sentence). In other words, it was necessary to have captions and tags for image-view, keywords for sentence-view, and 3-4 sentences corresponding to wordcloud-view. Hence, we augment the dataset in Pereira et al. (2018) by obtaining target annotations manually. For example, for the fMRI associated with picture p, we manually annotated it with target sentence s'="A red bird sitting on a tree branch". Pairs like (p, s') are then used to train model for image captioning. Note that these manual annotations do not involve obtaining more fMRIs.

Fig. 2 shows the input and output examples for the four cross-view decoding tasks. We make the augmented dataset publicly available[2]. Note that we do not experiment with CVD tasks like image generation from sentences or word clouds since obtaining target annotations would mean that we need to draw images to augment the dataset. We leave it as part of future work.

3.2 Task Descriptions

We train the decoder regression models on 5000 informative voxels selected from fMRI brain activations and evaluate all the models using pairwise accuracy and rank-based decoding. Details of the informative voxel selection, the regression model, and metrics are discussed in the subsequent sections. The main goal of each decoder model is to predict a semantic vector representation of the stimuli in each experiment. The input view (word+picture, sentence, or word-cloud) and output representation (word, sentence, or word-cloud) differ across experiments. We follow K-fold cross-validation, in which all the data samples from K-1 folds were used for training, and the model was tested on samples of the left-out fold. We use the BERT-pooled output for obtaining output semantic representations. We also experimented with RoBERTa, but the results were very similar to BERT, and hence we omit them for lack of space.

Multi-View Concept Decoding For each subject in the dataset, for each of the three input views, we trained K=18 models (one for each fold) where each model is trained on the brain activity of 170 concepts and tested on left-out 10 concepts to predict vector representation of the concept word. The 5000 informative voxels were selected for 170 concepts in each fold, and the same voxel locations were chosen for test datasets. At test time, the input to each model can belong to any of the three views. Thus, for each subject, for each fold, we perform (1) three same-view train-test experiments and (2) six multi-view zero-shot train-test experiments with different input views at train and test time. Target is always fixed as a vector representation of the concept word. We use pairwise accuracy to report results.

Cross-view Decoding Tasks For each subject in the dataset, we learn models for the four cross-view decoding tasks (IC, IT, KE, SF) using 18 fold cross-validation. The input and output for each of these tasks is shown in Table 1. Fig. 2 shows an example for each task. As before, we use 5000 informative voxels, computed separately for each of the 11 subjects and each of the four tasks. The regression target is semantic vector representation.

3.3 Informative Voxel Selection

Inspired by the voxel selection method in (Pereira et al., 2018), we chose the informative voxels for our linear regression models as follows. The regression models are trained on each voxel and its 26 neighboring voxels to predict the semantic vector representation. For each voxel in the training part, the mean correlation was calculated between "true" (text-derived) and predicted representations, and the voxels corresponding to the top 5000 mean cor-

relation values were selected as informative voxels. Target semantic representations are word embeddings for multi-view zero-shot concept decoding and 'word or sentence or word-cloud' embedding for cross-view decoding experiments. Voxel selection provides meaningful cognitive insights across brain networks.

3.4 Model Architecture

We trained a ridge regression based decoding model to predict the semantic vector representation associated with the fMRI informative voxels for a type (view) of each language stimulus. Each dimension is predicted using a separate ridge regression model. Formally, we are given the informative voxel matrix $X \in \mathbb{R}^{N \times V}$ and stimuli vector representation $Y \in \mathbb{R}^{N \times D}$, where N denotes the number of training examples, V denotes the number of informative voxels (we fix it to 5000), and D denotes the embedding dimension of language stimuli. For BERT, D=768. The ridge regression objective function is $f(X_i) = \min_{W_{io}} \|Y_o - X_i W_{io}\|_F^2 + \lambda \|W_{io}\|_F^2$ where, X_i denotes the input voxels for view i (out of {word+picture, sentence, wordcloud}), Y_o denotes the matrix with embeddings o (out of {word, sentence, word cloud}), W_{io} denotes the learned weight coefficients for each input view i and output embedding o, $\|.\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tunable hyper-parameter representing the regularization weight. Besides ridge regression, of course, various other models could be used. However, the goal of this paper is to analyze novel decoding setups using the most popular decoding model in neuro-science literature, namely, ridge regression. We leave exploration of complex models as part of future work. **Hyper-parameter Settings**: We used sklearn's ridge regression with default parameters, 18-fold cross-validation, Stochastic-Average-Gradient Descent Optimizer, Huggingface for BERT, MSE loss function and L2-decay (λ):1.0.

3.5 Brain Networks Selection

Inspired by Pereira et al. (2018) and based on the resting-state functional networks, we focused on four brain networks: Default Mode Network (DMN) (linked to the functionality of semantic processing) (Buckner et al., 2008; Binder et al., 2009), Language Network (related to language processing, understanding, word meaning, and sentence comprehension) (Fedorenko et al., 2011), Task Positive Network (related to attention, salience information) (Binder et al., 2009; Duncan, 2010; Power et al., 2011), and Visual Network (related to the processing of visual objects, object recognition) (Buckner et al., 2008; Power et al., 2011). We report the distribution of 5000 informative voxels across the four brain networks across various experiments in Section 4. Across all participants, voxel distribution across networks is as follows: 4670 (Language), 6490 (DMN), 11630 (TP), and 8170 (Visual). Note that the reported distributions in Section 4 do not add up to 1 because the contribution of the remaining brain networks is not considered.

4 Results and Cognitive Insights

Since we are the first to propose multi-view and cross-view tasks, unfortunately, there are no baselines to compare with. For the sake of comparison, we design a "chance-level" BERT (Random) baseline where models are trained using BERT embeddings of randomly chosen words as a target rather than BERT embeddings of the actual target word. For same-view experiments, our results are in line with that reported in (Pereira et al., 2018).

5 Evaluation Metrics

We use the popular pairwise and rank accuracy metrics for evaluation. **Pairwise Accuracy** To measure the pairwise accuracy, the first step is to predict all the test stimulus vector representations using a trained decoder model. Let S = [S_0, S_1,$\cdots$,S_n], $\hat{S}$ = [$\hat{S}_0$, $\hat{S}_1$,$\cdots$,$\hat{S}_n$] denote the "true" (text-derived) and predicted stimulus representations for n test instances resp. Given a pair (i, j) such that $0 \leq i, j \leq n$, score is 1 if *corr*(S_i,$\hat{S}_i$) + *corr*(S_j,$\hat{S}_j$) > *corr*(S_i,$\hat{S}_j$) + *corr*(S_j,$\hat{S}_i$), else 0. Here, *corr* denotes the Pearson correlation. Final pairwise matching accuracy per participant is the average of scores across all pairs of test instances.

Rank Accuracy We compared each decoded vector to all the "true" text-derived semantic vectors and ranked them by their correlation. The classification performance reflects the rank r of the text-derived vector for the correct word: $1 - \frac{r-1}{\#instances-1}$. The final accuracy value for each participant is the average rank accuracy across all instances.

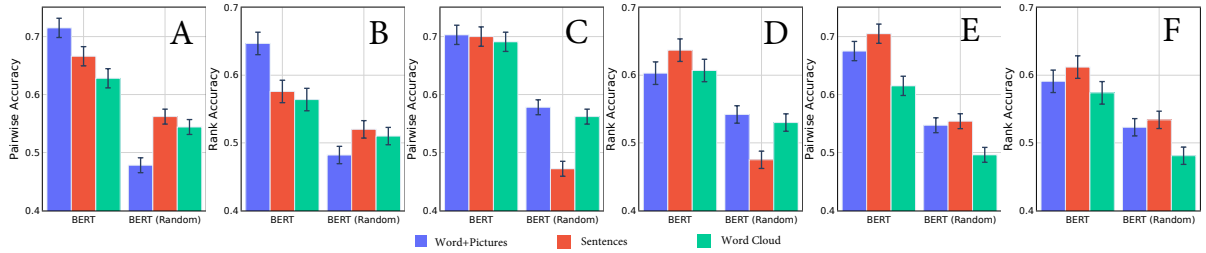

Figure 3: Model trained on *Word+Pictures* (A and B), *Sentences* (C and D), and *Word-Cloud* (E and F) view. MVD Pairwise and Rank accuracy when tested on Word+Picture/Sentence/Word-cloud views, averaged across all subjects.

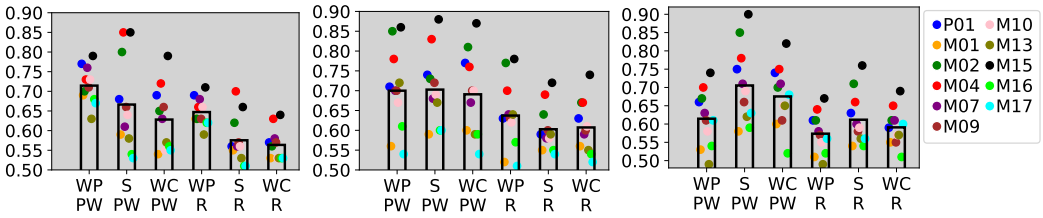

Figure 4: Model trained on *Word+Pictures* view (left), *Sentences* view (middle), *Word-Cloud* view (right). MVD Pairwise (PW) and Rank (R) accuracy when tested on Word+Picture (WP)/Sentence (S)/Word-cloud (WC) views. Each colored dot represents a subject. The bar plot shows averages.

5.1 Multi-View Concept Decoding

5.1.1 Pairwise and Rank Accuracy Results

Fig. 3 and Table 2 show detailed results for models trained on word+picture (WP), sentence (S), and word-cloud (WC) views and tested on each of the three views. Specifically, Fig. 3(A) shows pairwise accuracy results when we train using the WP view but infer using voxels corresponding to any of the three views. Ground-truth is the BERT embedding vector. In comparison to the "chance-level" BERT (Random) baseline with random target vectors, our proposed BERT embedding-based method is much better. Fig. 4 shows subject wise results.

Test↓/Train→	WP	S	WC
WP	0.72/0.65	0.70/0.60	0.68/0.59
S	0.67/0.58	0.70/0.64	0.71/0.61
WC	0.63/0.56	0.69/0.61	0.62/0.57

Table 2: Multi-View Zero-shot Concept Decoder Results (Pairwise/Rank Accuracy)

Same view versus MV zero-shot: In most cases, same-view results are better than multi-view zero-shot results. However, this does not hold for the WC view, where a model trained on sentence view performs better (Left green bars in Fig. 3 (C and D) vs. Fig. 3 (E and F)).

Can we train MV decoders that can decode concepts from brain recordings for any view? We experimented with three different MV decoders, each trained on one of the three views. Fig. 3 and the statistical significance test results in Table 3 show that either of the WP and sentence (S) views can be used to train MV decoders. This means that if we train a model with WP or S view fMRIs, and test it using any of the three views, the results are better or equivalent to any other model. This does not hold for the WC view. Thus, an MV decoder trained with a WC view is not very effective.

Setting 1	Setting 2	p-value
Train(WP)-Test(WP)	Train(S)-Test(WP)	0.098
Train(WP)-Test(WP)	Train(WC)-Test(WP)	0.026*
Train(S)-Test(WP)	Train(WC)-Test(WP)	0.474
Train(WP)-Test(S)	Train(S)-Test(S)	0.485
Train(WC)-Test(S)	Train(S)-Test(S)	0.469
Train(WP)-Test(S)	Train(WC)-Test(S)	0.420
Train(WP)-Test(WC)	Train(WC)-Test(WC)	0.691
Train(WP)-Test(WC)	Train(S)-Test(WC)	0.134
Train(S)-Test(WC)	Train(WC)-Test(WC)	0.045*

Table 3: p-values for measuring if setting 1 is stat significantly better than setting 2. Only rows with * mark denote statistically significant improvements.

5.1.2 Cognitive Insights based on Distribution of Informative Voxels

Table 4 and Fig. 5 show the distribution of informative voxels among four brain networks for various MV models. In this figure, (WP, D) means input view=WP (Word+picture), brain network=DMN (D). The figure clearly shows that a lot of informative voxels belong to the visual brain region for the WP view. Also, for sentence view, a large percentage of informative voxels are from the language region.

Figs. 6 to 8 show more distribution details by zooming further into language and visual regions.

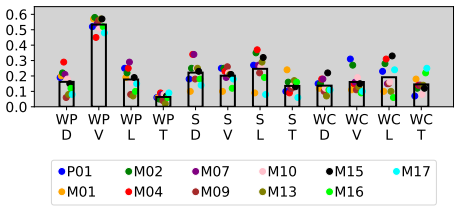

Figure 5: Distribution of informative voxels among four brain networks: DMN (D), Visual (V), Language (L), Task Positive (T). Models trained on Word+Picture (WP), Sentence (S) or Word-Cloud (WC) views.

When the model is trained on the WP view (unlike other views), Table 4 and Fig. 6 show that most informative voxels (about 53%) lie in the visual brain network, which is expected for the predominantly visual information-driven task.

	Word+Picture	Sentence	Word-Cloud
DMN	0.162	0.222	0.137
Visual	0.534	0.202	0.161
Language	0.177	0.246	0.192
Task-Positive	0.064	0.135	0.145

Table 4: Distribution of informative voxels among four brain networks for various Multi-View models

We also observe that DMN and Language network voxels are higher in the sentence view than in the word cloud view. Compared to the model trained on WP view, the distribution of voxels among the four brain networks shows that the model trained on sentence view has a higher percentage of voxels among the Language, DMN, and Task-positive networks and lower in the visual network. This is in line with our understanding that linguistic and attention skills are essential for understanding sentence stimuli. As for the model trained on the WC view compared to other views, we see that the informative voxels are spread equally among all the networks. From Fig. 6, we observe that in all the views, the region corresponding to language processing in the left hemisphere (Language_LH) has higher informative voxels than that of the right hemisphere (Language_RH). This is in line with the left hemisphere dominance for language processing (Binder et al., 2009). When the visual network dominates as in the case of WP view, the majority of these are located in the object processing area, followed by face and body processing areas. In the following, we investigate these two regions in detail.

In the language network, the distribution of informative voxels in the sub regions (LPTG, LMTG, LATG, LFus, LPar, LAngG, LIFGorb,

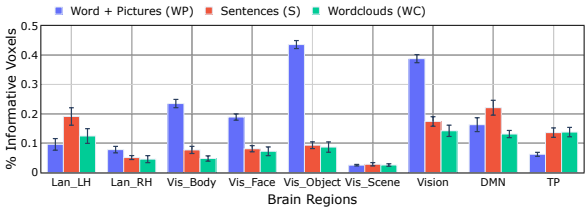

Figure 6: Distribution of informative voxels among nine brain regions for Multi-view Decoding

LIFG, LaMFG, LpMFG, and LmMFG) are shown in Fig. 7. We find that regions in the posterior (LPTG), middle (LMTG), and anterior (LATG) temporal gyrus share a higher percentage of informative voxels than other regions in the language network, such as those in the middle and inferior frontal areas. This indicates that the language functions sub-served by the temporal cortex, such as comprehension and semantic processing, are critical for processing sentences as well as multi-modal integration and thus are important for decoding across multiple views. Further, brain regions in the angular gyrus (LAngG) and parietal (LPar) each have >5% of informative voxels. These areas may be involved in attention, self-processing, and visio-linguistic integration.

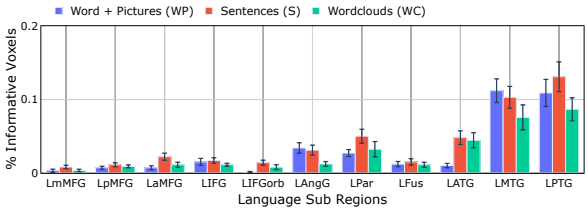

Figure 7: Distribution of informative voxels among eleven sub regions of Language network for MVD

Similarly, we explored the distribution of informative voxels across sub regions of the visual network, as shown in Fig. 8. In the visual sub regions, voxels in the bilateral occipital cortex (LLOC and RLOC) have more informative voxels than in other sub regions. In particular, the scene regions in the parahippocampal place area (such as RSC and PPA) display very few informative voxels, while the bilateral body area (REBA and LEBA) captures more

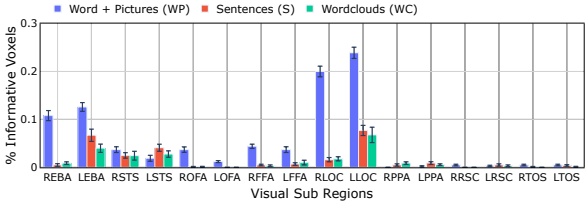

Figure 8: Distribution of informative voxels among sixteen sub regions of Visual network for MVD

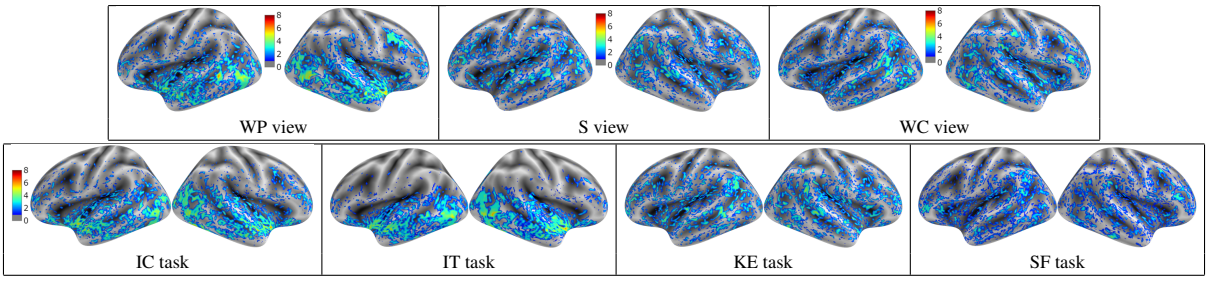

Figure 9: Brain Maps for Multi-View and Cross-View Decoding Tasks (plotted using nilearn Python library).

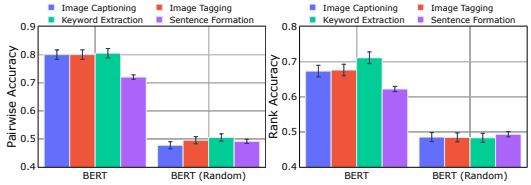

Figure 10: Cross-View Decoding Pairwise and Rank accuracy for Image Captioning (IC), Image Tagging (IT), Keyword Extraction (KE), and Sentence Formation (SF) averaged across all the subjects.

voxels in the WP view. Interestingly, activation in the superior temporal sulcus (RSTS and LSTS) in all views point out its role in visio-linguistic integration. Lastly, Fig. 9 shows the spatial distribution of informative voxels (plotted using nilearn Python library) across models trained on different forms of stimuli (WP, S, and WC). The value of each voxel is the fraction of 11 participants for whom that voxel was among the 5000 most informative.

5.1.3 Informative Voxel Overlap across Views

Given the distribution of informative voxels across four brain networks, we further examine how these voxels from one view overlap with those from another view. Table 5 shows that (1) the language network has a very high overlap compared to other brain networks in the WC-S pair. (2) 29% (and 25%) of visual voxels for the S (and WC) view are shared with visual voxels of the WP view. This makes sense since a large percentage of informative voxels for WP view are from the visual network.

	DMN	Visual	Language	Task Positive
WP-S	0.24/0.17	0.11/0.29	0.25/0.17	0.09/0.05
WC-S	0.25/0.16	0.25/0.20	0.30/0.22	0.07/0.07
WP-WC	0.14/0.16	0.08/0.25	0.15/0.15	0.06/0.03

Table 5: For each pair of views and each brain network, we show coverage ratios (second task on first/first task on second) of the voxels.

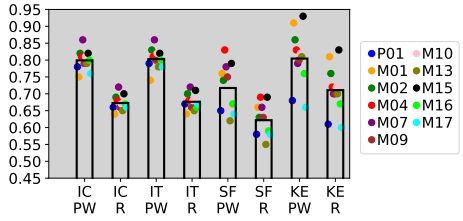

Figure 11: CVD Pairwise (PW) and Rank (R) accuracy for IC, IT, KE and SF tasks. Each colored dot represents a subject. The bar plot shows averages.

	IC	IT	SF	KE
DMN	0.114	0.067	0.152	0.214
Visual	0.572	0.736	0.154	0.236
Language	0.116	0.081	0.182	0.275
Task Positive	0.045	0.007	0.141	0.118

Table 6: Distribution of informative voxels among four brain networks for all 4 CVD Tasks.

5.2 Cross-View Decoding

5.2.1 Pairwise and Rank Accuracy Results

Fig. 10 illustrates pairwise and rank accuracy for Image Captioning (IC), Image Tagging (IT), Sentence Formation (SF), and Keyword Extraction (KE). Subject wise results are reported in Fig. 11. We observe that (1) our proposed BERT embedding-based method is much better compared to the "chance-level" baseline with random target vectors. (2) For all the four tasks, pairwise accuracy is ∼80%, and rank-based accuracy is ∼70% (except for SF), which shows that CVD is possible with good accuracy.

5.2.2 Cognitive Insights based on Distribution of Informative Voxels

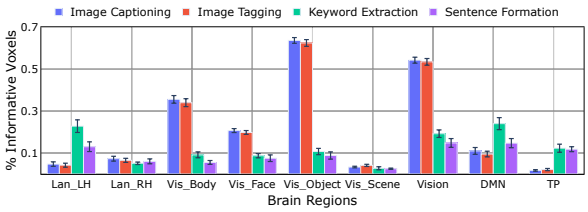

Figure 12: Distribution of informative voxels among nine brain regions for CVD tasks.

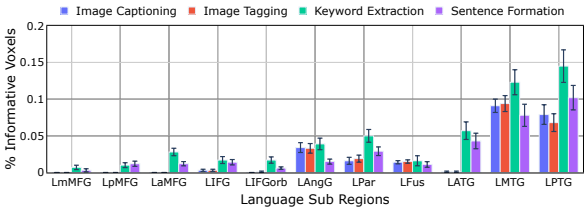

Figure 13: Distribution of informative voxels among 11 sub regions of Language network for CVD tasks.

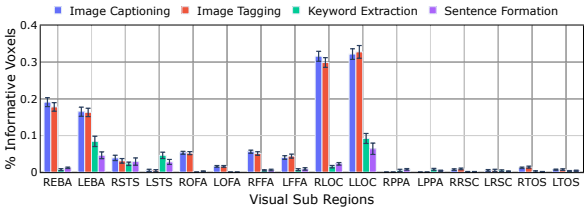

Figure 14: Distribution of informative voxels among 16 sub regions of Visual network for CVD tasks.

Fig. 12 shows the distribution of informative voxels among nine brain regions across all four tasks. As expected, a high percentage of visual voxels are involved in IC and IT tasks, and a high percentage of language voxels are involved in the SF and KE tasks, especially in the left hemisphere. Further, from Table 6, we observe that IC involves relatively higher language voxels compared to IT. This could be because generating a caption involves a higher level of language (sequence) skills than generating a set of keywords.

To further investigate the informative voxel distribution across Language and Visual networks, we display the sub region voxels distribution for the Language network in Fig. 13, and for the Visual network in Fig. 14. In all the tasks, the left hemisphere language network activation is dominated by activity in the temporal gyrus (middle: LMTG and posterior: LPTG) but more in the KE task. This clearly demonstrates the importance of language comprehension and semantic process common across the cross-view tasks. Further, the common activation in the angular gyrus (LAG) in all tasks points out the role of visio-linguistic integration critical for all the tasks. The activation profile of the vision network, in contrast, shows distinct activation differences across the tasks (IC & IT vs. KE & SF). IC and IT tasks are related to a higher proportion of informative voxels in the primary visual regions in the lateral occipital areas (LLOC, RLOC) and bilateral extrastriate body-related areas (REBA and LEBA). Domination of activation in the vision network in captioning and tagging tasks (IC and IT) as compared to predominantly sentence processing based tasks (KE and SF) is along expected lines.

	DMN	Visual	Language	Task Positive
IC-IT	0.27/0.44	0.70/0.54	0.32/0.45	0.07/0.32
IC-KE	0.31/0.17	0.11/0.27	0.28/0.12	0.12/0.05
IC-SF	0.16/0.12	0.07/0.25	0.14/0.09	0.08/0.03
IT-KE	0.27/0.08	0.08/0.25	0.22/0.07	0.05/0.01
IT-SF	0.13/0.05	0.06/0.27	0.10/0.05	0.04/0.00
KE-SF	0.19/0.26	0.20/0.29	0.22/0.32	0.09/0.08

Table 7: For each pair of CVD tasks and each brain network, we show coverage ratios (second task on first/first task on second) of the voxels.

The brain maps (see Fig. 9) corresponding to the IC and IT tasks clearly activate the visual cortex and the temporal cortex, the areas known for visual processing and object identification. On the other hand, the brain maps of KE and SF exhibit diffuse activation that includes the temporal and frontal regions known to be related to the sentence semantics. None of the maps show a left-hemisphere bias, which is often found in such semantic-related maps. Lack of frontal-lobe activation and the concentration of informative voxels in the sensory cortex suggest that the cross-view embedding may rely on some non-abstract domain-specific encoding rather than higher-level semantic concept encoding.

5.2.3 Informative Voxel Overlap across Tasks

Given the distribution of informative voxels across four brain networks, we further examine how these voxels from one task overlap with those from another task. Table 7 shows that (1) Many voxels overlap across different brain networks for IC and IT tasks. This is expected since the two tasks are very related. Interestingly, 44% of DMN voxels needed for IT are shared with IC. Similarly, as high as 70% of visual voxels needed for IC are shared with IT. (2) Similarly, KE and SF share a very good overlap across different brain networks, which is expected given the textual nature of the two tasks.

6 Conclusion

We studied brain decoding in the context of zero-shot multi-view concept decoding and cross-view decoding tasks. We studied four cross-view decoding tasks: image captioning, image tagging, sentence formation, and keyword extraction. We show that cross-view decoding is feasible with good accuracy. Brain network distribution analysis reveals insights about the importance of various parts of the brain for each of these tasks.

7 Ethical Statement

We reused publicly available Pereira dataset, downloadable from https://osf.io/crwz7/, for this work. Please read their terms of use[3] for more details. We did not collect any new dataset. We do not foresee any harmful uses of this technology.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Jelle Zuidema. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. In *Proceedings of the ACL-Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. 2020. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *arXiv preprint arXiv:1907.02431*.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

RL Buckner, JR Andrews-Hanna, and DL Schacter. 2008. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124:1–38.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

John Duncan. 2010. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179.

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

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194.

Jon Gauthier and Roger Levy. 2019. 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)*, pages 529–539.

Nora Hollenstein, A de la Torre, Nicolas Langer, and Ce Zhang. 2019. Cognival: A framework for cognitive word embedding evaluation. In *Proceedings of The SIGNLL Conference on Computational Natural Language Learning 2019*.

Alexander G Huth, Tyler Lee, Shinji Nishimoto, Natalia Y Bilenko, An T Vu, and Jack L Gallant. 2016. Decoding the semantic content of natural movies from human brain activity. *Frontiers in systems neuroscience*, 10:81.

Shailee Jain and Alexander G Huth. 2018. Incorporating context into language encoding models for fmri. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6629–6638.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

[3] https://github.com/CenterForOpenScience/cos.io/blob/master/TERMS_OF_USE.md

Satoshi Nishida and Shinji Nishimoto. 2018. Decoding naturalistic experiences from human brain activity via distributed representations of words. *Neuroimage*, 180:232–242.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646.

Subba Reddy Oota, Naresh Manwani, and Raju S Bapi. 2018. fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings. In *International Conference on Neural Information Processing*, pages 3–15. Springer.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *NIPS*.

Francisco Pereira, Greg Detre, and Matthew Botvinick. 2011. Generating text from functional brain images. *Frontiers in human neuroscience*, 5:72.

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

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. 2011. Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. *arXiv preprint arXiv:1911.03268*.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7047–7054.

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

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv preprint arXiv:1905.11833*.

Shaonan Wang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2020a. Probing brain activation patterns by dissociating semantics and syntax in sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9201–9208.

Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. 2020b. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256–272.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Shijie Zhao, Xi Jiang, Junwei Han, Xintao Hu, Dajiang Zhu, Jinglei Lv, Tuo Zhang, Lei Guo, and Tianming Liu. 2014. Decoding auditory saliency from fmri brain imaging. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 873–876.