# IMPROVING BRAIN DECODING METHODS AND EVALUATION

*Damián Pascual, Béni Egressy, Nicolas Affolter, Yiming Cai, Oliver Richter, Roger Wattenhofer*

ETH Zurich, Zurich, Switzerland

## ABSTRACT

Brain decoding, understood as the process of mapping brain activities to the stimuli that generated them, has been an active research area in the last years. In the case of language stimuli, recent studies have shown that it is possible to decode fMRI scans into an embedding of the word a subject is reading. However, such word embeddings are designed for natural language processing tasks rather than for brain decoding. Therefore, they limit the model's ability to recover the precise stimulus. In this work, we propose to directly classify an fMRI scan, mapping it to the corresponding word within a fixed vocabulary. Unlike existing work, we evaluate on scans from previously unseen subjects. We argue that this is a more realistic setup and we present a model that can decode fMRI data from unseen subjects with $2.62\%$ Top-1 and $9.76\%$ Top-5 accuracy in this challenging task. Moreover our model can be fine-tuned on data from the test subject to achieve $4.22\%$ Top-1 and $12.87\%$ Top-5 accuracy, significantly outperforming all the considered competitive baselines.

***Index Terms*—** fMRI, classification, word embeddings

## 1. INTRODUCTION

Recent advances in brain imaging suggest that it may be possible to infer what a person is perceiving from their brain scans. The ability of decoding brain signals has important applications in medicine, e.g., assisting handicapped people who cannot move or talk, as well as in the consumer industry, e.g., producing content that adapts to what a person is seeing, feeling or thinking. In this context, language is of particular interest, since it is the vehicle we use to express our thoughts. Consequently, a body of research has focused on decoding functional Magnetic Resonance Imaging (fMRI) scans into a representation of the word a person is reading while being scanned [1, 2, 3, 4, 5, 6, 7].

Most previous work, and notably [5], evaluate the quality of their models for brain decoding by measuring the similarity between predicted word representations and true word representations. They show that in most cases their predicted representation is closer to the corresponding word than to a randomly selected word from the data. Although an important first step in showing that inferring such information from brain scans is at all possible, this task, called pairwise classification, is rather simple. Furthermore, as [8] shows, this task is fundamentally limited by the choice of word representation the decoder is trained to produce. Word representations are often derived from models optimized for very different tasks and contain extraneous information, e.g., word frequencies in the training set. The inference models used to solve it are equally simple, normally based on ridge regression or simple Multi-Layer Perceptrons (MLP), while they rely on complex subject-specific pre-processing and feature selection [5, 6].

In this work, we argue that a more demanding setup needs to be considered in order to understand the extent to which we can currently map brain activities to words. In particular, we propose direct classification, i.e. to directly classify a brain scan as one of the $v$ words within the considered vocabulary, as opposed to pairwise classification. Furthermore, we address brain decoding on unseen subjects, i.e., we consider the setting where the training data does not contain any data from the test subject. This is known to be a remarkably hard problem, since fMRI scans are very different across subjects and even across recording sessions, among other reasons due to variable numbers of voxels and lack of alignment between scans. Thus, the challenge with this setup is twofold, the evaluation task is more demanding and strong generalization is required since subject-specific pre-processing is not possible.

On the bright side, in this setup we can exploit a larger training set consisting of scans from multiple subjects in order to train more complex models. Specifically, we propose a neural autoencoder model that takes as input a complete fMRI scan and encodes it as a hidden representation. This hidden representation can then be decoded into the stimulus word. We use no external knowledge and we let the model learn features that generalize to all subjects. We validate our model on the classical pairwise classification task and then demonstrate its performance in direct classification. We further analyze the model behavior in terms of the amount of data available and conclude that larger datasets are of paramount importance for brain decoders to be applicable in real-world settings.

## 2. RELATED WORK

Since the publication of the seminal work [1], decoding brain activity into words has attracted a lot of attention from the research community. In recent years, a large number of studies have tackled this problem from different angles, including

deep learning methods, which are expected to help in improving our understanding of the brain [9]. In particular, [4] proposes a model to learn new classes unseen during training, [2, 3] build brain decoders that help draw conclusions about how the brain processes language. [5] presents a model that decodes brain activity into word embeddings. [7] decodes text passages rather than single words and, similarly, [6] decodes sentences using distributed representations. All of these works represent just a part of a large body of research that has strongly contributed to the progress of decoding and understanding brain activities. Here, we build on previous studies to further advance the state-of-the-art of brain decoding.

To evaluate the performance of fMRI-to-text decoders, one common approach is to decode the input into a textual representation [5, 6], such as Global Vectors (GloVE) [10]. However, such representations contain information beyond semantics. For instance, the word frequency in the data used to train the embedding is implicitly encoded [11]. [8] shows that decoding brain images into representations derived from models optimized to solve very different tasks, e.g., image captioning or machine translation, produce similar results as the baseline decoder from [5]. We propose to classify fMRI scans directly into words (or classes) thereby avoiding noisy textual representations.

One open problem in this field is to learn patterns from fMRI scans that generalize to subjects not present in the training set. This problem is rooted in the way the human brain represents semantic meaning [12, 13] and to what extent it differs across subjects. To address this challenge in brain decoding, different methods for aligning fMRI scans across datasets [14] and across subjects have recently been developed. [15] uses an autoencoder to aggregate data from different subjects, [16] obtain promising results with hyperalignment and more recently, [17] proposes a transfer learning approach to align data across subjects. In this work, we show that using no inductive bias, a deep neural network is able, to a certain extent, to generalize across fMRI scans of different subjects without the need for new alignment techniques.

## 3. EVALUATION SETUP

We use the dataset from [5], which contains fMRI scans from 15 subjects. Each subject was recorded reading 180 different words, one at a time. Each word, was shown to the subject following three different paradigms that ensure that all subjects focus on the same meaning, i.e., supporting the word with a word cloud, with sentences and with images. As such, the dataset we use consists of 15 subjects with 540 scans each (180 words, three paradigms).[1]

Most previous work on brain decoding [5, 6] considered the scenario where the model is trained with data from the same subject that is being evaluated. In said scenario, for each new subject a new training set needs to be recorded in order to train a personalised decoder model. However, recording fMRI scans is a costly and slow process. Furthermore, the amount of data that can be recorded for one subject is limited, which restricts the complexity of the decoder and forces the model to rely on subject-specific pre-processing. By using data from all recorded subjects, we can build larger neural network-based decoders that learn common features and generalize across subjects.

In this work, we consider two scenarios, (1) only the data from other subjects is available for training and (2) some data from the target subject is also available. In (1), a highly challenging setup, we follow a leave-one-out strategy in our evaluation, i.e., we train our model with data from $n-1$ subjects and test it on the remaining subject; we repeat this process for each subject. In (2), we *additionally* fine-tune the model on the scans of 170 of the words (out of 180) from the target subject. Following [5], the resulting model is tested on the remaining 10 words; this process is repeated 18 times such that all words are included in the test set once.

We evaluate our model in two different tasks:

**Pairwise classification.** In this task, a regression-based decoder is trained to produce a vector representation from a brain image (fMRI); following [5], we use GloVe embeddings [10]. Then, for each possible pair of words the correlation between the decoded vectors and the actual embedding vectors of both words is computed, i.e., four values. If the decoded vectors are more similar to their corresponding word embeddings than to the alternatives, the evaluation is considered correct. As such, the random baseline for this task is 50%. The final result is the mean across test instances.

**Direct classification.** In this task, which we propose, a classification-based decoder receives as input a brain scan and produces as output a vector of size $v$, where $v$ is the size of the vocabulary. In our case $v = 180$. This vector contains the predicted probability for each word in the vocabulary. This way, the decoder is effectively a classifier that infers which word was seen by the subject when the scan was taken. This task is significantly more challenging than the pairwise classification task, with the random baseline being $1/v$ for the Top-1 score (0.56% in our case). On the other hand, it does not suffer from limitations associated with the chosen vector representation. We report Top-1 and Top-5 scores, i.e., the classification is correct if the stimulus word is within the Top-X predictions.

## 4. BRAIN DECODING MODEL

Our model is a symmetric autoencoder that generates a latent representation $z$ of size 1 000. The input is a one dimensional vector of the fMRI scan with 65 730 voxels. To align the fMRI scans between subjects, we rely on the *Region Of Interest (ROI)* information following the Gordon atlas [18]. The voxels of each ROI are zero-padded to a common size

[1] For more details on the dataset refer to `https://osf.io/crwz7`

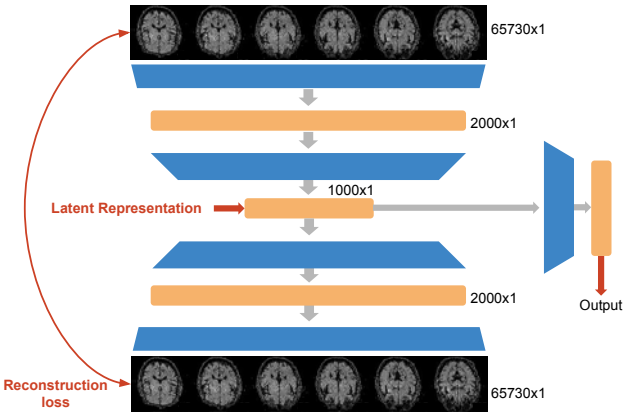
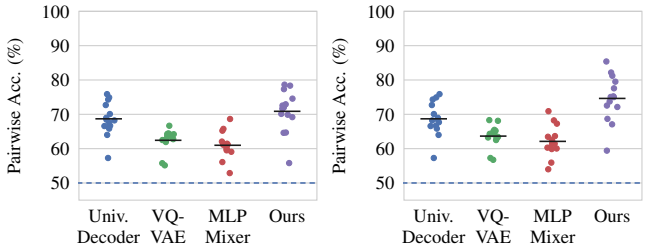



**Fig. 1**: Architecture of the decoder. Blue trapezoids represent dense layers and orange rectangles represent intermediate representations.




(a) No fine-tuning (b) With fine-tuning

**Fig. 2**: Performance comparison for pairwise classification without (left) and with (right) fine-tuning on the target subject. Each point represents a subject, solid lines are the mean across subjects and dashed lines the random baselines.

across subjects and concatenated in a predefined order to give a partial alignment of the inputs. Cosine distance is used for the *reconstruction* loss and each layer has $0.4$ dropout, batch normalization and Leaky ReLU activation ($\alpha = 0.3$). Besides the autoencoder, we attach an output head to the latent vector. The details of this output head depend on the task, as explained next.

For pairwise classification we follow [5] and use GloVE embeddings of size $300$; our loss is calculated as:

$$\mathcal{L}_{reg} = \sum_{i}^{v} \left( \cos\left(\tilde{y}_i, y_i\right) - \sum_{j \neq i}^{v} \cos\left(\tilde{y}_i, y_j\right) \right) \quad (1)$$

where $\tilde{y}_i$ is the predicted word embedding for word $i$, $y_j$ is the true word embedding for word $j$ and $\cos$ is the cosine distance. This loss is inspired by the triplet loss [19] and aims at guiding the model's output as close as possible to the true embedding while keeping it as far as possible from the rest.

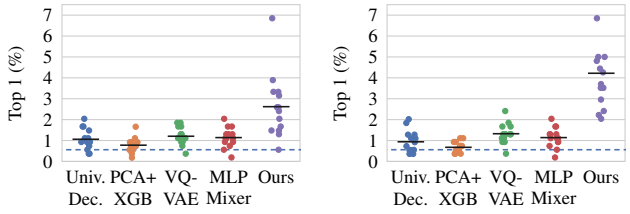
For direct classification the model outputs a vector of probabilities of the size of the vocabulary $v$ and we use categorical cross-entropy loss. A detailed depiction of the model can be seen in Fig. 1.

## 5. EVALUATION RESULTS

We compare our proposed decoder with three state-of-the-art architectures. First, we take the model from [5] which uses ridge regression. In the following we refer to this model as Universal Decoder. Second, we evaluate the VQ-VAE model from [20] adapted to regression-based decoding of fMRI. This model discretizes the latent space, which may naturally help it separate the scans according to the words they encode. Finally, we use the recent MLP-mixer architecture with $8$ mixer layers and $30$ patches [21], which reaches state-of-the-art performance in different computer vision tasks. In the direct classification task we consider as an additional baseline a model consisting of Principal Component Analysis (PCA) decomposition for dimensionality reduction, followed by XGBoost [22] for classification.

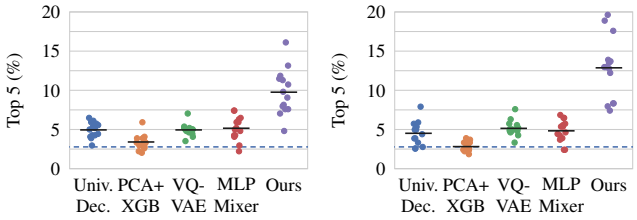
**Pairwise Classification.** To put our model into context with respect to existing work, we first evaluate it on the pairwise classification task. We report the results in Fig. 2 with (right) and without (left) fine-tuning on the target subject. The results show that the Universal Decoder, which was build for brain decoding, outperforms the other baselines, giving credit to its setup. However, all baselines struggle in our more challenging setup which asks to generalize across subjects. No baseline reaches an average accuracy over 70%, and are therefore outperformed by our model. In fact our model without fine-tuning ($70.88\%$) even outperforms the other models with fine-tuning. Our model with fine-tuning achieves the best results with $74.63\%$.

**Direct Classification.** We compare our model against five competitive baselines, the same four as above, but adapted to the classification task, and additionally, against a model consisting of Principal Component Analysis (PCA) for dimensionality reduction, followed by XGBoost [22]. We present the results for Top-1 and Top-5 accuracy in Figures 3 and 4 respectively. In this more complicated task the random baseline is $0.56\%$ for Top-1 and $2.8\%$ for Top-5 accuracy. We see that even with data from the target subject the Universal Decoder has a mean score of $0.94\%$ for Top-1 and $4.5\%$ for Top-5 accuracy, only slightly above random. Furthermore, in this challenging setup, our model is clearly the best for both Top-1 and Top-5 scores both with and without fine-tuning. In particular, without fine-tuning the Top-1 mean accuracy is $2.62\%$, almost 5 times the random baseline. This result is outstanding given the difficulty of the task, i.e., decoding the exact word corresponding to the fMRI scan of an unseen subject. This score improves to $4.22\%$ with fine-tuning on the target subject. The good performance of our decoder on this realistic scenario shows the potential of using brain decoding in real life applications.




(a) No fine-tuning (b) With fine-tuning

**Fig. 3**: Comparison of Top-1 performance on *Direct classification* with (left) and without (right) fine-tuning.




(a) No fine-tuning (b) With fine-tuning

**Fig. 4**: Comparison of Top-5 performance on *Direct classification* with (left) and without (right) fine-tuning.

## 6. DATA ANALYSIS

In order to understand the contribution of data from additional subjects better, we consider removing subjects from the training set. We hypothesize that removing subjects that are more dissimilar to the test subject will damage the performance less severely than both, removing at random or removing more similar subjects. Moreover removing dissimilar subjects might even improve the performance.

We use our validation subject (M15) as our test subject. We look at performance on the direct classification task without fine-tuning. As a similarity measure we use the sum of the cosine similarities between scans corresponding to the same word and paradigm between two different subjects, i.e., for subjects $a$ and $b$ the similarity is calculated as

$$sim(a,b) = \sum_{i}^{v} \sum_{p=1}^{3} 1 - \cos\left(x_{i,p}^{a}, x_{i,p}^{b}\right),$$

where $x_{i,p}^{a}$ is the fMRI scan of subject $a$ looking at word $i$ in paradigm $p$ and cos is the cosine distance.

To test our hypothesis we remove train subjects in three different orders.

- *Least*: we order the train subjects by similarity to the test subject and remove them one by one, least similar first.
- *Random*: we order the train subjects randomly and remove them one by one.
- *Most*: as with *Least*, but in reverse order, removing the most similar subject first.

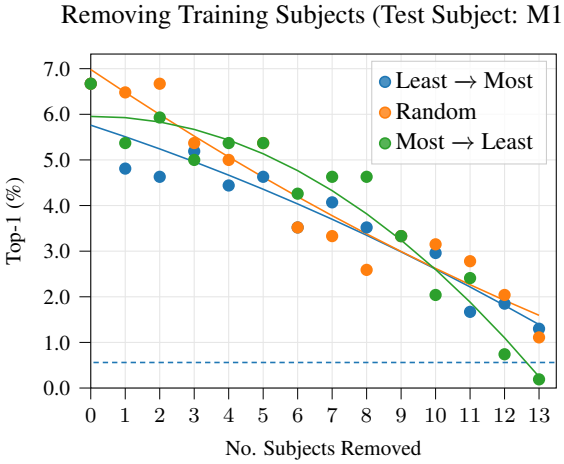



**Fig. 5**: Top-1 accuracy when training with fewer subjects. Quadratic trend lines are shown. Test subject is M15.

The results are shown in Fig. 5. We first note that the accuracy does not exceed $6.67\%$ when removing subjects and we observe an almost monotone decreasing trend. We can also see that the results do not support our hypothesis that removing less similar subjects damages the performance less severely. In fact in more than half the cases (8) removing the most similar subjects leads to a higher accuracy than removing the least similar ones and in 8 cases removing subjects in a random order performs best. This suggests that the specific subjects we use is of lesser importance, whereas the almost monotone decrease suggests that the number of training subjects is of much greater importance. Therefore, we conclude that more data could significantly improve our ability to decode brain activities.

## 7. CONCLUSION

In this work we have shown that to evaluate and understand current brain decoding models a more demanding setup is necessary and to this end, we propose *direct classification*. This task does not require to approximate noisy textual representations and its low random baseline makes apparent the differences in performance across models. Furthermore, we have presented a neural model for decoding fMRI scans into words, and shown that it outperforms existing models by a big margin. We have also run our experiments on the extremely demanding scenario where no data from the target subject is available at training time, demonstrating that our model successfully generalizes to unseen subjects. Finally, we have identified the lack of large enough datasets as one of the main obstacles for further advancing brain decoding. All in all, we believe that the good performance shown by our neural model in the proposed setup brings real-life applications of brain decoding interfaces one step closer.

## 8. REFERENCES


[1] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just, "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[2] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, pp. 453–458, 2016.

[3] Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell, "A neurosemantic theory of concrete noun representation based on the underlying brain codes," *PloS one*, vol. 5, no. 1, pp. e8622, 2010.

[4] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell, "Zero-shot learning with semantic output codes," in *Advances in neural information processing systems*, 2009, pp. 1410–1418.

[5] 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*, vol. 9, no. 1, pp. 1–13, 2018.

[6] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong, "Towards sentence-level brain decoding with distributed representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[7] 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," *PloS one*, 2014.

[8] Jon Gauthier and Anna Ivanova, "Does the brain represent words? an evaluation of brain decoding studies of language understanding," *arXiv preprint arXiv:1806.00591*, 2018.

[9] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al., "A deep learning framework for neuroscience," *Nature neuroscience*, vol. 22, no. 11, pp. 1761–1770, 2019.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014.

[11] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu, "Learning semantic hierarchies via word embeddings," in *52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[12] Karalyn Patterson, Peter J Nestor, and Timothy T Rogers, "Where do you know what you know? the representation of semantic knowledge in the human brain," *Nature reviews neuroscience*, 2007.

[13] Alex Martin, "The representation of object concepts in the brain," *Annu. Rev. Psychol.*, 2007.

[14] Hejia Zhang, Po-Hsuan Chen, and Peter Ramadge, "Transfer learning on fmri datasets," in *International Conference on Artificial Intelligence and Statistics*, 2018.

[15] Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S Turek, Janice Chen, Theodore L Willke, Uri Hasson, and Peter J Ramadge, "A convolutional autoencoder for multi-subject fmri data aggregation," *arXiv preprint arXiv:1608.04846*, 2016.

[16] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Kenneth A Norman, and Uri Hasson, "Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space," *NeuroImage*, 2020.

[17] Muhammad Yousefnezhad, Alessandro Selvitella, Daoqiang Zhang, Andrew Greenshaw, and Russell Greiner, "Shared space transfer learning for analyzing multi-site fmri data," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[18] Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, Jeremy F Huckins, William M Kelley, and Steven E Petersen, "Generation and evaluation of a cortical area parcellation from resting-state correlations," *Cerebral cortex*, vol. 26, no. 1, pp. 288–303, 2016.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE conference on computer vision and pattern recognition*, 2015.

[20] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[21] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al., "Mlp-mixer: An all-mlp architecture for vision," *arXiv preprint arXiv:2105.01601*, 2021.

[22] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.