Brain2Word: Improving Brain Decoding Methods and Evaluation

Nicolas Affolter[*]
ETH Zurich
nicolaff@student.ethz.ch

Béni Egressy[*]
ETH Zurich
begressy@ethz.ch

Damián Pascual[*]
ETH Zurich
dpascual@ethz.ch

Roger Wattenhofer[*]
ETH Zurich
wattenhofer@ethz.ch

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 our 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. Our model achieves 5.22% Top-1 and 13.59% Top-5 accuracy in this challenging task, significantly outperforming all the considered competitive baselines.

1 Introduction

Since the publication of the seminal work [7], decoding brain activity into words has attracted a lot of attention from the research community [4, 5, 6, 9, 10, 12, 13, 16]. Most previous works, and notably [10], evaluate the quality of their decoders 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. However, this task is rather simple and, as [2] shows, it 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 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. Hence, we propose doing direct classification, i.e., directly classifying a brain scan as one of the v words within the considered vocabulary. Since we are mapping the scan to a word-class rather than mapping it to a representation of the word, this task does not suffer from limitations associated with a particular vector representation.

Furthermore, we address the more challenging and realistic scenario of decoding for unseen subjects, i.e., the training data does not contain any data from the test subject. This is known to be a remarkably hard problem, due to the lack of alignment in fMRI scans across subjects and even across recording sessions. Recent work [8, 15] has studied this problem in controlled settings and approached it from

[*] Authors in alphabetic order.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

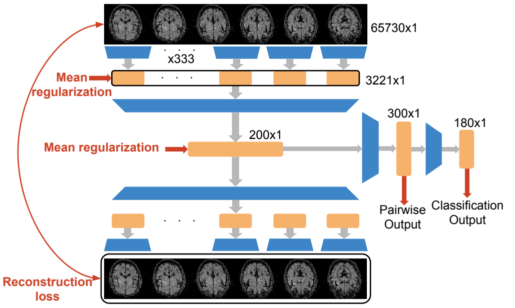

Figure 1: Architecture of the decoder. Blue trapezoids represent dense layers, orange rectangles feature maps and solid black lines concatenation.

Model	Pairwise	Direct
Base	0.8268	4.07%
+ ROI	0.8336	4.25%
+ Reconstruction	0.8411	4.81%
+ Mean reg.	0.8464	5.55%
+ Pretraining	**0.8637**	**6.29%**

Table 1: Ablation study. The extensions are progressively added to the model.

an algorithmic perspective. Here, we take a data-driven approach and successfully generalize brain decoding to unseen subjects.

Thus, the challenge in our setup is twofold, the evaluation task is more demanding and strong generalization is required since subject-specific pre-processing is not possible. We propose a model that decodes brain activities in this setup, we validate our model on the classical pairwise classification task, and we demonstrate its performance in direct classification.

2 Background

We use the dataset from [10]. This dataset contains fMRI scans from 15 subjects recorded while reading 180 words; it also contains $4,530$ scans from a subset of the subjects reading sentences[2]. To evaluate our model on data from a subject not present in the training set we follow a leave-one-out approach and train our model with the data from $n-1$ subjects and test it on the remaining subject; we repeat this process for each subject. We use one subject for validation (hyperparameter search). We consider two evaluation tasks:

Pairwise classification [10] The decoder predicts a vector representation from an fMRI scan. For each pair of words the correlation between the predicted vectors and the actual embedding vectors of both words is computed. If the predicted vectors are more similar to their corresponding word embeddings than to the alternatives, the decoding is deemed correct. The random baseline is 50%.

Direct classification The decoder classifies the input fMRI scan into one of the v words of the vocabulary, $v = 180$ for our dataset.The random baseline is $1/v$ for the Top-1 score (0.6% in our case).

3 Brain Decoding Model

Our decoder[3] takes as input a one dimensional vector of the fMRI scan with size $65,730 \times 1$ voxels; padding is applied as required. The model can be used in both tasks, pairwise and direct classification by simply changing the output layer and the loss function. For pairwise classification we follow [10] and use GloVE embeddings of size 300×1; our loss is calculated as:

$$\mathcal{L}_{reg} = \sum_{i}^{v} (\cos{(y_{pr,i}, y_{true,i})} - \sum_{j \neq i}^{v} \cos{(y_{pr,i}, y_{true,j})}) \tag{1}$$

Where $y_{pr,i}$ is the predicted word embedding for word i, $y_{true,j}$ is the real word embedding for word j and $\cos()$ is the cosine distance. This loss is inspired by the triplet loss [11] and aims at guiding the

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

[3] Code: `https://github.com/nicolaffETHZ/Brain2Word_paper`

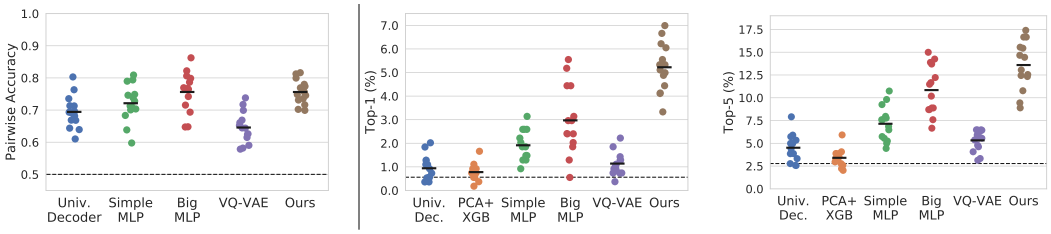

Figure 2: Performance comparison for pairwise (left) and direct classification (right). Each point represents a subject, solid lines are the mean across subjects and dashed lines the random baselines

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 one dimensional vector of probabilities of the size of the vocabulary v. The classification loss is the categorical cross-entropy.

Our model is a symmetric autoencoder (decoder-encoder) that uses cosine distance for the *reconstruction* loss. In the first layer of the decoder, we partition the fMRI scans into *Regions Of Interest (ROIs)* following the atlas from [3], and use a different dense layer to process each region. After concatenating the result we map it to a latent vector of size 200×1, that is given as input to the encoder. Each layer has 0.4 dropout, batch normalization and Leaky ReLU activation ($\alpha = 0.3$). Since the model should produce the same output for scans from different subjects when exposed to the same word, the latent representations inside the model should converge. Thus, using the same structure as in Eq. 1 we regularize the output of each layer of the decoder to be similar to the *mean* representation for a given word across subjects at that layer, and dissimilar to the mean representation of the other words. To exploit general language-related fMRI features, we use the $4,530$ sentence scans from the dataset to *pretrain* our model using exclusively the reconstruction loss for 30 epochs. Fig. 1 depicts the model and in Tab. 1 we ablate our design decisions on the validation subject.

4 Results

Pairwise Classification To put our model into context with respect to existing work, we evaluate it on the pairwise classification task and compare it with four competitive baselines. First, the Universal Decoder from [10], which uses ridge regression. Given the reduced capacity of this model, training it on subjects different than the test subject produced close to random performance. Therefore, we train and evaluate on the same subject as in the original work [10]. Second, a simple MLP consisting of a non-linear layer that maps the input to a feature vector of size 2000×1 followed by a linear layer that outputs the GloVe embedding. Third, a big MLP with one non-linear dense layer per ROI, as in our complete model, followed by a linear layer that outputs the GloVe embedding. Last, the VQ-VAE model from [14] adapted to regression-based decoding of fMRI. This model discretizes the latent space, thus, we hypothesize that it may naturally separate the scans according to the word that they encode. We report the results in Fig. 2 (left). First, we see that the VQ-VAE has the worst performance, which rejects our hypothesis about the discrete latent space. All the other models outperform the Universal Decoder even with the disadvantageous training setup (unseen test subject). These results show that neural network-based decoders successfully generalize to unseen subjects and even clearly outperform classical models trained on the target subject.

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 decomposition (PCA) for dimensionality reduction, followed by XGBoost [1]. We present the results for Top-1 and Top-5 accuracy in Fig. 2 (right). In this more complicated task (the random baseline is 0.6% for Top-1 and 2.8% for Top-5) the Universal Decoder mean accuracy is 0.94% for the Top-1 score and 4.5% for Top-5, slightly above random. We see that in this challenging setup, our complete model is clearly the best for both Top-1 and Top-5 scores. In particular, its Top-5 mean accuracy is above 13.59%, 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. The good performance of our decoder on this realistic scenario shows the potential of using brain decoding in real life applications.

5 Conclusion

In this work we have presented a model for decoding fMRI scans into words, and shown that it outperforms existing models by a big margin. Furthermore, we have shown that a more realistic task is necessary to understand the performance of decoding models and to this end proposed direct classification. We have run our experiments on the extremely demanding scenario where no data from the target subject is available at training time and demonstrated that our model successfully generalizes to unseen subjects. The good performance of our model in this setup opens the door to real-life applications using brain decoding interfaces.

References

[1] 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*, pages 785–794, 2016.

[2] 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.

[3] 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*, 26(1):288–303, 2016.

[4] Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, Giovanna Marotta, and Pietro Pietrini. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, 135:232–242, 2016.

[5] 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*, 532(7600):453–458, 2016.

[6] 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*, 5(1):e8622, 2010.

[7] 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*, 320(5880):1191–1195, 2008.

[8] 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*, page 116865, 2020.

[9] 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*, pages 1410–1418, 2009.

[10] 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.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

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

[13] 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*, volume 33, pages 7047–7054, 2019.

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

[15] Cara E Van Uden, Samuel A Nastase, Andrew C Connolly, Ma Feilong, Isabella Hansen, M Ida Gobbini, and James V Haxby. Modeling semantic encoding in a common neural representational space. *Frontiers in neuroscience*, 12:437, 2018.

[16] 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*, 9(11):e112575, 2014.