CorText: large language models for cross-modal transformations from visually evoked brain responses to text captions.

Victoria Bosch
victoria.bosch@uos.de
Institute of Cognitive Science, University of Osnabrück
Osnabrück, NI 49090, Germany

Dirk Gütlin
Department of Education and Psychology, Freie Universität
Berlin, 12249, Germany

Adrien Doerig
Institute of Cognitive Science, University of Osnabrück
Osnabrück, NI 49090, Germany

Daniel Anthes
Institute of Cognitive Science, Osnabrück University
Osnabrück, NI 49090, Germany

Sushrut Thorat
Institute of Cognitive Science, Osnabrück University
Osnabrück, NI 49090, Germany

Peter König
Institute of Cognitive Science, Osnabrück University
Osnabrück, NI 49090, Germany

Tim C Kietzmann
tim.kietzmann@uos.de
Institute of Cognitive Science, University of Osnabrück
Osnabrück, NI 49090, Germany

Abstract

An emerging trend in cognitive neuroscience is to investigate neural responses to complex natural scenes. While more ecologically valid, the complexity of these stimuli requires analysis techniques capable of studying not only the neural responses to object categories that constitute a given scene, but also their rich spatial and semantic interactions. Here, we present a generative brain-to-text decoder, CorText, that produces linguistic descriptions of natural scenes based on visually-evoked fMRI responses. At no point does the decoder have access to the visual stimulus, it operates solely on brain data. This cross-modal transformer, consisting of a linear encoder for neural data and a partly frozen pre-trained language decoder, enables us to harness the powerful features of language models to study neural representations. As a proof of concept, we analyse the neural regions most informative for generating specific words by visualising the transformer's attention patterns. This approach reproduces known functional organisation: elevated attention in the ventral stream and, accordingly, attention in cortical regions involved in category-specific processing. This work thus marks an important first advance into end-to-end generative language transformers for investigating complex neural data.

Keywords: vision; transformer; NSD; scene perception; cross-modal alignment; semantics; neural decoding

Introduction

Experimental work in cognitive neuroscience has been expanding from simple object stimuli towards the study of natural scenes (Doerig et al., 2022; Peelen, Berlot, & de Lange, 2024; Bartnik & Groen, 2023; Epstein & Baker, 2019). These stimulus materials are complex and move beyond single categories by showing groups of objects embedded in scene context, as well as people interacting with them and with each other. Classical decoding techniques fall short in capturing this richness of information, due to their focus on single object categories and the constraints that come with predefining fixed classifier categories. For more holistic semantic decoding of scene information from neural data, recent work has explored the usage of sentence embedding vectors, derived from image captions (Doerig et al., 2022; Güçlü & van Gerven, 2015; Zhang, Han, Worth, & Liu, 2020). Here we expand this line of research to leverage the full generative capacity that large language models have to offer. We present *CorText*, a cross-modal brain-to-text transformer that decodes fMRI data, recorded while people viewed natural scenes, into captions of the perceived image in natural language. This work is based on the insight that pre-trained, largely frozen language transformers are capable of cross-modal fine-tuning (Lu, Grover, Abbeel, & Mordatch, 2022). Consequently, we adapt the FlanT5 language transformer (Chung et al., 2022) to take neural data as input by constructing a linear brain encoder and partially fine-tuning the decoder to map from brain to text. CorText is trained on the Natural Scenes Dataset (NSD), a large-scale 7T fMRI dataset of neural responses to images depicting complex natural scenes (Allen et al., 2022), as well as the corresponding image captions that are available for the stimuli (Chen et al., 2015). To our knowledge, this work is the first approach to adapt pre-trained language transformers to yield a generative model that maps from the brain to textual captions, without any access to the underlying stimulus materials. We demonstrate the feasibility of this approach by showing that the transformer's attention on the neural embeddings aligns with known functional organisation of cortex. Future work will further exploit the capabilities of current language models for neuroscience, such as question-answering and language-based hypothesis testing.

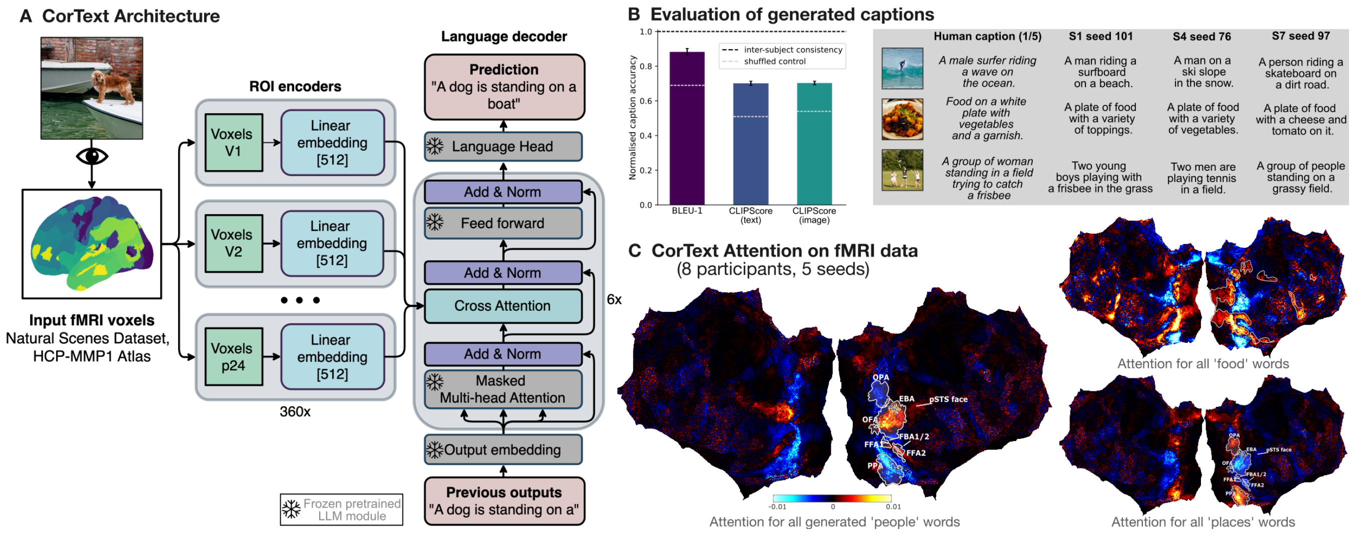
Results

CorText successfully captions fMRI data

To investigate neural decoding with a partly frozen language decoder, we first parcellate the brain data using the HCP-MMP1 atlas. The resulting regions are linearly embedded and fed as tokens into the language decoder (Fig. 1A). The final output of the transformer is then compared to a human caption of the corresponding scene, and gradient descent is used to adjust the encoder weights, cross-attention, and layer normalisation. After training, we find that CorText's captions of perceived visual content match ground-truth human captions when trained on only a single participant's data (Fig. 1B). These results show that with little training data, pre-trained language transformers can be adapted to produce sensible captions of neural data without access to the image stimuli.

CorText exhibits word-specific attention that aligns with the functional organisation of visual cortex

To utilise the trained transformer for insights into brain computations, we investigate the model's cross-attention patterns, which weigh the input data to select the most relevant information for the task at hand. For both language and vision transformers, visualisation of the attention weights in input data space has shown to be informative about the levels of semantic processing in the model (Dosovitskiy et al., 2020; Vig & Belinkov, 2019). For CorText, this means that we can inspect how the attention mechanism for the generation of each word of a given caption weighs the encoded brain regions. Because our encoder is linear, we can reverse the encoding and retrieve the attention-weighted beta values. We find that the attention for the generation of words of several categories is matched to known functional organisation of cortex (Fig. 1C; data averaged over all subjects and seeds of final layer). For words describing people (e.g., 'man', 'woman'), attention matches regions of interest (ROIs) involved in face perception and processing of social interactions, e.g. FFA and EBA. Likewise, words depicting places (e.g., 'street', 'kitchen') increase attention to relevant ROIs such as PPA (Allen et al., 2022). Lastly, we find that for words that describe food (e.g., 'pizza', 'donut') the attention map matches the food ROIs as defined



	Human caption (1/5)	S1 seed 101	S4 seed 76	S7 seed 97
	A male surfer riding a wave on the ocean.	A man riding a surfboard on a beach.	A man on a ski slope in the snow.	A person riding a skateboard on a dirt road.
	Food on a white plate with vegetables and a garnish.	A plate of food with a variety of toppings.	A plate of food with a variety of vegetables.	A plate of food with a cheese and tomato on it.
	A group of woman standing in a field trying to catch a frisbee	Two young boys playing with a frisbee in the grass	Two men are playing tennis in a field.	A group of people standing on a grassy field.

Figure 1: A - The CorText architecture *fMRI data is parcellated into 360 regions which are linearly encoded into 512-dim. embeddings each. The partially frozen language decoder decodes the neural embeddings and produces a caption of the perceived stimulus, without having access to the image. Each network is trained on one out of 8 NSD subjects, we train 5 seed instances of each.* **B - Evaluation** *Left: Generated captions match ground-truth human captions as measured by Bleu-1 and CLIPscore (text and image). Right: Example human and model captions for 3 stimuli.* **C - Attention on fMRI data** *The cross-attention-weighted betas for specific generated word categories align with well-known ROIs, here visualized for the final attention layer across all trained networks.*

by (Pennock et al., 2023). These results indicate that CorText can accurately target and exploit structured neural representations to produce natural language descriptions of the perceived stimulus. Furthermore, the attention maps show that transformer attention on neural data is informative and justifies their use in novel neural decoding applications.

Methods

Datasets: The Natural Scenes Dataset (NSD) contains 7T fMRI measurements of 8 participants who have each viewed 9000 unique images sourced from the MS COCO dataset (Lin et al., 2014). In addition, all participants have seen (up to) 1000 shared images, of which we use 515 as test set. We use the beta values of the 1.8-mm volume preparation in fsaverage space. Each image in NSD has five human captions from MS COCO, which are used for supervised training.

Model architecture: CorText is an encoder-decoder transformer, inspired by the insight that with finetuning, frozen pretrained language transformers are capable of generalizing to other modalities (Lu et al., 2022). As a basis for our exploration, we rely on a pre-trained language transformer (FlanT5-Small) with strong capabilities in various text-based tasks (Chung et al., 2022). We freeze the pre-trained language decoder and replace the FlanT5 encoder with 360 linear encoders, one for each brain region (180 per hemisphere) as defined by the HCP-MMP1 atlas (Glasser et al., 2016). These ROIs are processed similarly to language tokens, in analogy to the approach taken by vision transformers (Dosovitskiy et al., 2020). We replace softmax in cross-attention layers by the following: $Att_{norm} = \frac{att_i - \min(\vec{att})}{max(\vec{att}) - min(\vec{att})}$. Note that the encoder used in CorText is linear. While limited in computational expressivity, this approach enables us to subsequently analyse cross-attention maps in the original voxel space.

Training: Each encoder learns a linear mapping between a parcel of NSD beta values and a 512-dimensional embedding for the decoder. Only the encoders, the cross-attention heads and layer normalisation throughout the decoder are trainable, resulting in a relatively low number of trainable parameters for the sampled dataset (37M). The model minimizes cross-entropy between generated and randomly selected human captions for each trial. Models are trained with Adagrad for 70 epochs, with a learning rate of 1e-3, and L2 encoder weight regularisation of 5e-1. For each model, we train 5 seeds to account for inter-model variation. Importantly, we used FlanT5 as, to our knowledge, it was not trained on COCO captions, unlike most language and multimodal models (e.g., CLIP). This avoids memorization of captions, forcing our model to learn beyond simple look-up.

Metrics: To evaluate the quality of the captions generated from neural data, we use several metrics. BLEU evaluates the n-gram match between words in the predicted caption and its ground truth. RefCLIPscore captures semantic correspondences, dealing well with sentences that have different syntax but similar meanings. Furthermore, we use CLIPScore (Hessel, Holtzman, Forbes, Bras, & Choi, 2021) to evaluate the correspondence with the seen image. The ceiling for each metric is set by the average correspondence of all MS COCO human captions.

Attention maps: To investigate the cross-attention, we retrieve the average attention of the attention heads in each layer on the test dataset. Because cross-attention maps from each brain 'token' to each generated word, we can retrieve the attention for specific categories of words. We do so for people, places and food. To retrieve the voxel-level attention maps, we weigh the brain embeddings for the stimulus condition for which a word of interest was generated with the cross-attention on that word. Following this, we invert the linear embedding to retrieve an attention-weighted voxel brain map.

Acknowledgments

The authors acknowledge support by the ERC stg grant 101039524 TIME (Bosch, Doerig, Kietzmann), SNF grant n.203018 (Doerig), RTG GRK2340 DFG (Anthes). Compute resources for this project are in part funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 456666331.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Bartnik, C. G., & Groen, I. I. (2023). Visual perception in the human brain: How the brain perceives and understands real-world scenes. In *Oxford research encyclopedia of neuroscience.*

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... others (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual review of vision science*, *5*, 373–397.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., ... others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Güçlü, U., & van Gerven, M. A. (2015). Semantic vector space models predict neural responses to complex visual stimuli. *arXiv preprint arXiv:1510.04738*.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Lu, K., Grover, A., Abbeel, P., & Mordatch, I. (2022). Frozen pretrained transformers as universal computation engines. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 7628–7636).

Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, *3*(1), 13–26.

Pennock, I. M., Racey, C., Allen, E. J., Wu, Y., Naselaris, T., Kay, K. N., ... Bosten, J. M. (2023). Color-biased regions in the ventral visual pathway are food selective. *Current Biology*, *33*(1), 134–146.

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, *11*(1), 1877.