Interpretable fMRI Captioning via Contrastive Learning

Vyacheslav Shen, Kassymzhomart Kunanbayev, Donggon Jang, and Daeshik Kim

KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{shen9910, kkassymzhomart, jdg900, daeshik}@kaist.ac.kr

Abstract. Recent advances in deep learning and generative AI have enhanced our understanding of brain function and enabled brain-computer interfaces to reconstruct stimuli from non-invasive neuroimaging data. In this work, we introduce an efficient two-stage training framework for captioning stimulus images from fMRI data, leveraging the compact representations of vision-language models and incorporating contrastive learning with text embeddings. Our approach demonstrates strong performance in fMRI captioning across multiple evaluation metrics and enables multimodal retrieval, highlighting the advantages of the contrastive learning. Additionally, we conduct an analysis with region-of-interests (ROI) to examine the contributions of specific brain regions to the decoding process, providing interpretable results that align with neuroscience theories. Our findings contribute to advancing brain decoding techniques and improving model interpretability.

Keywords: Neural Decoding · fMRI · Contrastive Learning

1 Introduction

Hierarchical image processing in the brain [6] inspired the development of convolutional neural networks (CNNs) [8]. Visualizing features and saliency maps across CNN layers reveals that early layers detect edges, while deeper layers capture class-specific features [29], similar to the functioning of the visual cortex. Furthermore, CNN-learned representations strongly correlate with neural activity in macaques [28] and humans [26]. Given these similarities, deep neural networks (DNNs) are increasingly used to decode visual representations in the human brain by inversely predicting DNN features from neural activity.

Huth *et al.* [7] showed that hours of narrated stories could be decoded by mapping fMRI data to word embeddings. Recent advances in generative models have further improved decoding accuracy. For instance, Latent Diffusion Models (LDMs) [17] have facilitated the reconstruction of high-resolution stimulus images from fMRI data [13] [22]. Meanwhile, the Transformer architecture [24] and GPT-2 [15] have significantly improved natural language reconstruction from neural activity [23].

However, the quality and semantic coherence of generated outputs require further refinement and alternative approaches. Traditional approaches primarily reconstruct visual stimuli as images from brain activity. Recent advances in multimodal deep learning provide a compelling alternative: decoding neural responses directly into textual descriptions, a process known as fMRI captioning. In this context, multimodal retrieval provides a flexible way to decode both what is seen and the underlying semantic content from brain activity. Despite the progress in fMRI-based decoding, significant challenges remain in efficiently aligning brain activity with meaningful textual descriptions. Existing methods often struggle with computational efficiency, semantic coherence, and retrieval capabilities.

To address these challenges, we propose a novel framework that enhances fMRI captioning through contrastive learning. Our key contributions are as follows:

- We propose a novel and computationally efficient two-stage training algorithm to align fMRI data with the compact latent representations of vision-language models, particularly BLIP-2 [9]. Our approach introduces an improved method for fMRI captioning, demonstrating that contrastive learning enhances decoding performance while extending the capabilities of fMRI-based models to enable direct image and text retrieval.
- We provide an interpretability analysis of the decoding process using synthetic fMRI patterns. Our findings offer insights into the roles of different brain regions in neural decoding, aligning with established theories of hierarchical and modular information processing in the brain.

1.1 Related Work

The CLIP (Contrastive Language-Image Pre-training) model [14], which consists of an image and text encoder, has played a crucial role in advancing multimodal models. CLIP's text embeddings guide the reverse diffusion process in Latent Diffusion models [17] for text-to-image generation, while its image encoder has been widely integrated into Vision-Language Models (VLMs) [11] [25] to align Large Language Models (LLMs) with visual data.

Given its versatility, CLIP has also been actively used in neural decoding. Ferrante et al. [4] and Scotti et al. (MindEye-2) [19] leverage fMRI signals to predict CLIP image embeddings, which are then used to reconstruct visual stimuli via Stable Diffusion [17] or generate captions using the text decoder of the Generative Image Transformer (GIT) [25]. Similarly, Mai et al. [12] propose an fMRI captioning approach that predicts the latents of CLIP-L's text encoder, which are subsequently processed by GPT-2 [15] within Versatile Diffusion's [27] text-to-text pipeline.

However, brain decoding research face a significant challenge due to the high dimensionality of conditional embeddings. For instance, Brain Diffuser [13], MindEye-2 [19], and the work of Ferrante et al. [4] attempt to predict high-dimensional embeddings of size 257×768 and 257×1024, respectively, from an

already high-dimensional fMRI voxel vector of length $15,724$. These approaches impose high computational demands.

To mitigate this limitation, our work leverages the BLIP-2 [9] multimodal model, which utilizes more compact visual embeddings of size 32×768. BLIP-2 employs a Querying Transformer (Q-Former) to map image encoder features into the LLM embedding space. Acting as a compression network, Q-Former encodes large frozen image features (257×1024) into compact query tokens (32×768), preserving text-relevant and semantically rich image representations that are well-suited for brain decoding.

2 Methodology

2.1 Dataset

Our work utilizes the Natural Scenes Dataset (NSD) [1], a 7 Tesla fMRI dataset collected from eight subjects who viewed images from the COCO dataset [10] for three seconds each. Consistent with prior fMRI captioning studies, we focus on data from a single subject (*subj1*) for quantitative analysis. Subject 1 completed all experimental trials, resulting in a dataset of $8,859$ training images corresponding to $24,980$ fMRI trials (with up to three repetitions per image) and 982 test images with $2,770$ fMRI trials. For images with multiple presentations, we averaged the corresponding fMRI trials in our experiments.

Following Ozcelik et al. [13], we processed the fMRI data using single-trial beta weights obtained from a GLM with ridge regression (*betas_fithrf_GLMdenoise_RR*). We applied z-normalization along the time dimension and extracted a $15,764$-voxel vector using the *NSDGeneral* Regions-of-Interest (ROI) masks. For interpretability analysis, we employed ROI masks corresponding to various visual cortex regions.

2.2 fMRI Captioning with BLIP-2

Our method leverages the pretrained BLIP-2 model[1] to generate textual descriptions from fMRI activity. We chose BLIP-2 over other Vision-Language Models due to its more compact and language-aligned image representations (32×768) compared to the larger embeddings used in previous fMRI captioning methods (257×768 or 257×1024).

As shown in Figure 1, our two-stage framework begins with feature extraction and Brain Model training. A stimulus image is processed by the BLIP-2 Image Encoder, and its extracted features undergo cross-attention with learned query vectors in the BLIP-2 Q-Former, producing a final representation of size 32×768. Next, we train the Brain Model using Ridge Regression, mapping fMRI activity ($15,764$ voxels) to each channel of the BLIP-2 Q-Former embeddings. To optimize performance, we evaluate the model across a range of regularization parameters (α) and select the best-performing configuration.

[1] https://huggingface.co/Salesforce/blip2-opt-2.7b

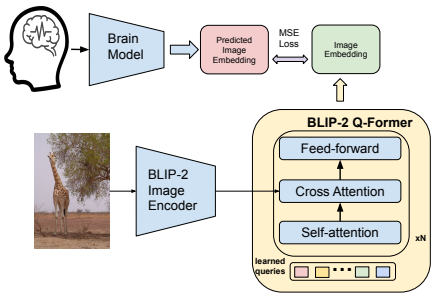

Fig. 1: Stage 1: fMRI voxels are mapped to BLIP-2's internal representations through a Brain Model. The image encoder extracts features, which interact with query tokens in the BLIP-2 Q-Former to produce the final embedding (green).

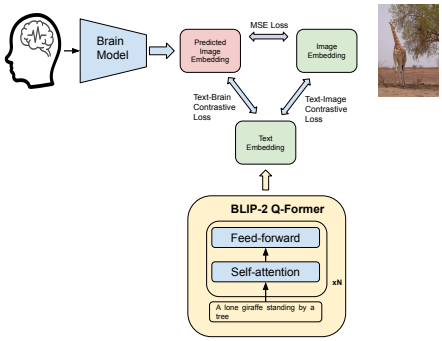

Fig. 2: Stage 2: The Brain Model predicts embeddings aligned with both image and text representations using MSE and contrastive losses. The BLIP-2 Q-Former generates text embeddings from COCO captions via self-attention layers.

In Stage 2, we apply contrastive learning to align the Brain Model outputs with text embeddings for retrieval. Ground truth COCO captions are processed through only the BLIP-2 Q-Former's self-attention layers to generate text embeddings. We use a pretrained BLIP-2 checkpoint[2], which includes vision and language projection weights optimized for image-text retrieval. To align the Brain Model's output with BLIP-2's shared image-text space, we introduce a linear projection layer.

Figure 2 illustrates this training process, where the Brain Model from Stage 1 is optimized alongside contrastive objectives. The loss function is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}\text{MSE}(b, i) + \lambda_2 \mathcal{L}\text{CLIP}(b, t) + \lambda_3 \mathcal{L}_{\text{CLIP}}(i, t) \tag{1}$$

where $\mathcal{L}_{\text{CLIP}}$ refers to the InfoNCE loss used in CLIP training [14]. The total loss consists of three terms, each weighted by a corresponding λ:

1. Mean Squared Error (MSE) loss: Preserves alignment between the Brain Model's predicted embeddings b and the ground truth image embeddings i from Stage 1.
2. Brain-text contrastive loss: Encourages the Brain Model's outputs b to align with text embeddings t, improving text retrieval.
3. Image-text contrastive loss: Prevents catastrophic forgetting in the BLIP-2 Q-Former and ensures robust image-text alignment by reinforcing consistency between t and i.

[2] https://huggingface.co/Salesforce/blip2-itm-vit-g

For Stage 2, we implement our model using PyTorch Lightning [3] and leverage its automatic learning rate tuning mechanism [20]. All experiments were conducted on an NVIDIA RTX A6000 GPU.

Our code is available on GitHub[3].

3 Results & Discussion

3.1 Retrieval

For image and brain retrieval, we first convert candidate images into BLIP-2 Q-Former representations (Figure 1). Using the predicted representation from the input fMRI data, we compute cosine similarity scores between the fMRI-derived representation and candidate image embeddings. Following the MindEye-2 evaluation protocol, we compute top-1 retrieval accuracy across 300 samples, where random chance is 0.33%. Reported results reflect the mean accuracy over 30 trials.

For text/brain retrieval, we predict an text-aligned image embedding from fMRI data using the Brain Model from the second stage (Figure 2) and obtain caption embeddings via BLIP-2 Q-Former. To identify the correct caption from fMRI data—and vice versa we use cosine similarity, following the image/brain protocol, with top-1 accuracy averaged over 50 trials.

Our brain decoding pipeline (Table 1) excels in multimodal retrieval. Stage 1 surpasses Brain Diffuser in image retrieval but slightly underperforms in brain retrieval. Stage 2 significantly improves accuracy, demonstrating the benefits of contrastive learning. While not matching MindEye-2's near-perfect performance, our model is capable of text/brain retrieval, achieving 49.6% accuracy in retrieving text from fMRI and 45.0% for brain signals from text (among 300 candidates). This multimodal retrieval unlocks natural-language querying of fMRI data, resulting in a more comprehensive interpretation of brain activity. We attribute MindEye-2's superior results to its direct image-retrieval training, larger embeddings (256×1664), and multi-subject data, whereas our approach emphasizes compact, multimodal representations.

3.2 fMRI Captioning

For fMRI captioning, we generate textual descriptions using the OPT-2.7B decoder-only language model [30], as implemented in BLIP-2. The process maps fMRI data to query embeddings, which are then projected and fed into the language model. As shown in Figure 3, this task requires only the Brain Model, allowing us to utilize outputs from both Stage 1 and Stage 2. To evaluate performance, following previous works, we use linguistic metrics (Meteor, Rouge-1, Rouge-L) and cosine similarity scores between predicted and ground truth captions using the *all-MiniLM-L6-v2* Sentence Transformer [16], **CLIP-B** and **CLIP-L** text encoders.

[3] https://github.com/slavaheroes/brain-decoding-with-blip2

Table 1: Top-1 retrieval accuracies. All values are computed for subject 1, except Brain Diffuser, whose numbers are the mean over subjects 1, 2, 5, 7 [19]. **I, B, T** denote Image, Brain, and Text, respectively. **I** $\rightarrow$ **B** refers to retrieving the correct image given its corresponding fMRI signal, and so on.

Model	I $\rightarrow$ B	B $\rightarrow$ I	T $\rightarrow$ B	B $\rightarrow$ T
Stage 1	0.437	0.222	-	-
Stage 2	0.722	0.549	**0.496**	**0.450**
MindEye-2 [19]	**1.000**	**0.997**	-	-
MindEye-1 [18]	0.972	0.947	-	-
Brain Diffuser [13]	0.188	0.263	-	-

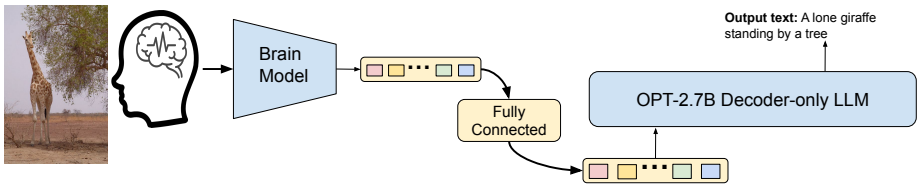

Fig. 3: Inference pipeline for fMRI captioning. The Brain Model maps fMRI data to BLIP-2 Q-Former embeddings, which are projected and processed by the OPT-2.7B [30] language model to generate textual descriptions.

Table 2 compares fMRI captioning performance across different models. Our Stage 2 model achieves the best performance in 5 out of 6 evaluation metrics, excelling in all linguistic metrics and sentence transformer and CLIP-Base embedding similarities. The only exception is CLIP-L similarity, where UniBrain performs better, likely due to its direct use of CLIP-L text encoder embeddings in the decoding process.

Notably, our Stage 1 model is already competitive with the state-of-the-art MindEye-2, outperforming it in all linguistic metrics while achieving similar results in semantic similarity scores. The consistent improvement from Stage 1 to Stage 2 across all metrics highlights the effectiveness of contrastive learning in

Table 2: Comparison of fMRI captioning performance with other works for subject 1.

Model	Meteor	Rouge-1	Rouge-L	Sentence	CLIP-B	CLIP-L
Stage 1	0.303	0.443	0.407	0.447	0.742	0.639
Stage 2	**0.327**	**0.467**	**0.430**	**0.515**	**0.771**	0.674
MindEye-2	0.248	0.326	0.353	0.479	0.737	0.638
UniBrain	0.170	0.247	0.225	-	-	**0.861**

generating more precise and detailed captions, mirroring its impact in retrieval tasks.

Figure 4 provides qualitative evidence of these improvements. The Stage 2 model generates more accurate and detailed descriptions, distinguishing fine-grained details. For instance, , it correctly identifies "a wave" instead of the more general "a beach" in a surfing image, recognizes "zebras" rather than "horses" and "buses" instead of "a train", and generates more descriptive adjectives, showing the better understanding of the scenes.

Fig. 4: Qualitative comparison of caption generation between Stage 1 and Stage 2 models. The top row shows stimuli images with captions generated by both models. Red text highlights key content differences.

3.3 Interpretability Analysis of ROI-Specific fMRI Signals

To analyze the role of different brain regions in visual processing and their contribution to brain decoding, we conducted an ROI-based interpretability analysis following Brain Diffuser [13]. We generated synthetic fMRI signals by setting voxel values of a Region of Interest (ROI) to 1 while zeroing out others. These signals were processed through our Brain Model, normalized, scaled by 11, and passed to the language model for caption generation.

Table 3 presents ROI-specific captions across four subjects who completed all trials. The results align with neuroscientific findings [2] [5] [6] [21], reflecting the hierarchical and modular nature of visual processing. Captions generated from V1 highlight basic black-and-white features, particularly for subjects 1 and 5, while descriptions become increasingly detailed from V1 to V4. Higher visual areas exhibit clear functional specialization: floc-words regions generate captions related to text and signs, floc-faces and floc-bodies consistently describe people and animals, and floc-places produces location-specific descriptions, such as "a chair in front of a building" or "a bathroom with a sink and a toilet on a wooden floor." These findings are consistent with ROI-specific image reconstructions

Table 3: Captions generated from synthetic fMRI signals that maximize activity in specific ROIs using Stage 1 models. Results are shown for subjects who completed all NSD sessions.

ROI	subj1	subj2	subj5	subj7
V1	a black and white sign	a man and a woman are sitting on a bench	a black and white cow	a man and a woman are walking on a street
V2	a man is standing next to a small animal	a close up of a large bowl with food on it	a group of people sitting on a street	a table with a bunch of food on it
V3	a white and blue building	a close up of a bathroom with a white and yellow striped bed	a man holding a black and white striped shirt	a bathroom with a sink and a few black and white chairs
V4	a group of people standing on a field with a ball	a small group of people standing on a white playing field	a group of people have on helmets on top of a skateboard	a group of white and yellow colored baseballs
floc-words	a sign that says "the park"	a sign that says "the black and white"	two brown colored cartoon animals	an object with the words inside stands under the blue sky
floc-bodies	a man in a blue shirt	a man holding a surfboard in a field	a man and a giraffe standing on a field	a man holding a surfboard
floc-faces	a man sitting on a chair with a dog	a woman sitting on a chair with a cat sitting on her lap	a woman sitting on a couch with a cat	a man and his very furry friend
floc-places	a woman sitting in a chair in front of a building	a room with a pool table and a couch	a bathroom with a window and a table with a clock	a bathroom with a sink and a toilet sitting on a wooden floor

reported in Brain Diffuser [13], where floc-faces and floc-words regions showed similar category-specific responses.

4 Conclusion

Our work introduces a compute-efficient two-stage training framework that integrates contrastive learning with text embeddings to generate accurate captions from fMRI signals. Additionally, our approach enables multimodal retrieval, demonstrating that contrastive training effectively aligns neural activity with

Vision-Language model representations, leading to improved performance. Furthermore, our ROI-optimal stimuli analysis improves interpretability by identifying the contributions of specific brain regions in the decoding process.

Future work will focus on cross-subject decoding to improve generalizability and explore multimodal generation to further enhance the applicability of our approach.

Acknowledgments. This work was supported by the Engineering Research Center of Excellence (ERC) Program supported by National Research Foundation (NRF), Korean Ministry of Science & ICT (MSIT) (Grant No. NRF-2017R1A5A101470823).

Disclosure of Interests. The authors have no competing interests.

References

1. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience **25**(1), 116–126 (2022)
2. Epstein, R., Kanwisher, N.: A cortical representation of the local visual environment. Nature **392**(6676), 598–601 (1998)
3. Falcon, W.A.: Pytorch lightning. GitHub **3** (2019)
4. Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., Toschi, N.: Brain captioning: Decoding human brain activity into images and text. arXiv preprint arXiv:2305.11560 (2023)
5. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science **293**(5539), 2425–2430 (2001)
6. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology **160**(1), 106 (1962)
7. Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural speech reveals the semantic maps that tile human cerebral cortex. Nature **532**(7600), 453–458 (2016)
8. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: 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. Springer (2014)
11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
12. Mai, W., Zhang, Z.: Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. arXiv preprint arXiv:2308.07428 (2023)

13. Ozcelik, F., VanRullen, R.: Natural scene reconstruction from fmri signals using generative latent diffusion. Scientific Reports **13**(1), 15666 (2023)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
16. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813 (2020)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
18. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al.: Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems **36** (2024)
19. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207 (2024)
20. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). pp. 464–472. IEEE (2017)
21. Spiridon, M., Kanwisher, N.: How distributed is visual category information in human occipito-temporal cortex? an fmri study. Neuron **35**(6), 1157–1165 (2002)
22. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14453–14463 (2023)
23. Tang, J., LeBel, A., Jain, S., Huth, A.G.: Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience **26**(5), 858–866 (2023)
24. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
25. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)
26. Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z.: Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral cortex **28**(12), 4136–4160 (2018)
27. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7754–7765 (2023)
28. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences **111**(23), 8619–8624 (2014)
29. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)

30. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)