Multimodal Neurons in Pretrained Text-Only Transformers

Sarah Schwettmann[1]*, Neil Chowdhury[1]*, Samuel Klein[2], David Bau[3], Antonio Torralba[1]
[1]MIT CSAIL, [2]MIT KFG, [3]Northeastern University
{schwett, nchow, sjklein, torralba}@mit.edu, d.bau@northeastern.edu

Abstract

Language models demonstrate remarkable capacity to generalize representations learned in one modality to downstream tasks in other modalities. Can we trace this ability to individual neurons? We study the case where a frozen text transformer is augmented with vision using a self-supervised visual encoder and a single linear projection learned on an image-to-text task. Outputs of the projection layer are not immediately decodable into language describing image content; instead, we find that translation between modalities occurs deeper within the transformer. We introduce a procedure for identifying "multimodal neurons" that convert visual representations into corresponding text, and decoding the concepts they inject into the model's residual stream. In a series of experiments, we show that multimodal neurons operate on specific visual concepts across inputs, and have a systematic causal effect on image captioning. Project page: mmns.csail.mit.edu

1. Introduction

In 1688, William Molyneux posed a philosophical riddle to John Locke that has remained relevant to vision science for centuries: would a blind person, immediately upon gaining sight, visually recognize objects previously known only through another modality, such as touch [24, 30]? A positive answer to the *Molyneux Problem* would suggest the existence a priori of 'amodal' representations of objects, common across modalities. In 2011, vision neuroscientists first answered this question in human subjects—*no*, immediate visual recognition is not possible—but crossmodal recognition capabilities are learned rapidly, within days after sight-restoring surgery [15]. More recently, language-only artificial neural networks have shown impressive performance on crossmodal tasks when augmented with additional modalities such as vision, using techniques that leave pretrained transformer weights frozen [40, 7, 25, 28, 18].

Vision-language models commonly employ an image-conditioned variant of prefix-tuning [20, 22], where a separate image encoder is aligned to a text decoder with a learned adapter layer. While *Frozen* [40], MAGMA [7], and FROMAGe [18] all use image encoders such as CLIP [33] trained jointly with language, the recent LiMBeR [28] study includes a unique setting: one experiment uses the self-supervised BEIT [2] network, trained with no linguistic supervision, and a linear projection layer between BEIT and GPT-J [43] supervised by an image-to-text task. This setting is the machine analogue of the Molyneux scenario: the major text components have never seen an image, and the major image components have never seen a piece of text, yet LiMBeR-BEIT demonstrates competitive image captioning performance [28]. To account for the transfer of semantics between modalities, are visual inputs translated into related text by the projection layer, or does alignment of vision and language representations happen inside the text transformer? In this work, we find:

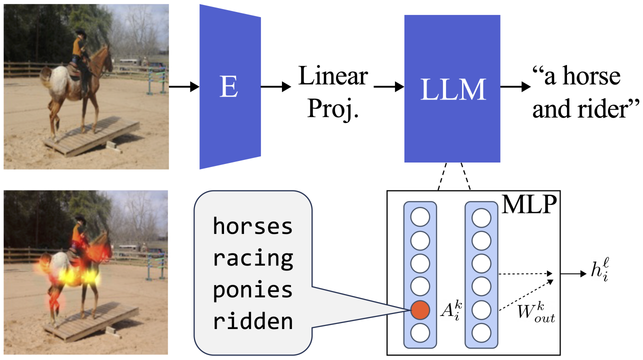
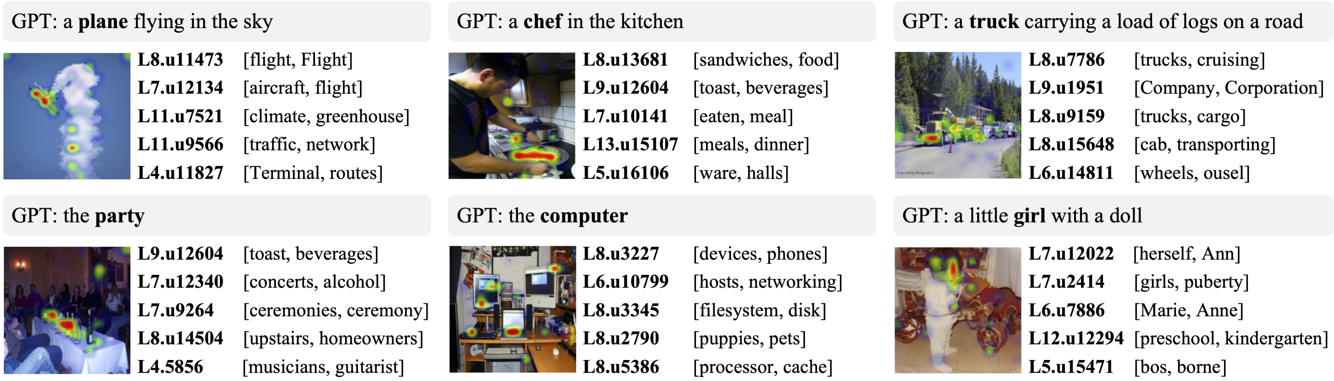


Figure 1. Multimodal neurons in transformer MLPs activate on specific image features and inject related text into the model's next token prediction. Unit 2019 in GPT-J layer 14 detects horses.

1. Image prompts cast into the transformer embedding space do not encode interpretable semantics. Translation between modalities occurs inside the transformer.
2. Multimodal neurons can be found within the transformer, and they are active in response to particular image semantics.
3. Multimodal neurons causally affect output: modulating them can remove concepts from image captions.

[*] Indicates equal contribution.



GPT: a **plane** flying in the sky

L8.u11473	[flight, Flight]
L7.u12134	[aircraft, flight]
L11.u7521	[climate, greenhouse]
L11.u9566	[traffic, network]
L4.u11827	[Terminal, routes]

GPT: a **chef** in the kitchen

L8.u13681	[sandwiches, food]
L9.u12604	[toast, beverages]
L7.u10141	[eaten, meal]
L13.u15107	[meals, dinner]
L5.u16106	[ware, halls]

GPT: a **truck** carrying a load of logs on a road

L8.u7786	[trucks, cruising]
L9.u1951	[Company, Corporation]
L8.u9159	[trucks, cargo]
L8.u15648	[cab, transporting]
L6.u14811	[wheels, ousel]

GPT: the **party**

L9.u12604	[toast, beverages]
L7.u12340	[concerts, alcohol]
L7.u9264	[ceremonies, ceremony]
L8.u14504	[upstairs, homeowners]
L4.5856	[musicians, guitarist]

GPT: the **computer**

L8.u3227	[devices, phones]
L6.u10799	[hosts, networking]
L8.u3345	[filesystem, disk]
L8.u2790	[puppies, pets]
L8.u5386	[processor, cache]

GPT: a little **girl** with a doll

L7.u12022	[herself, Ann]
L7.u2414	[girls, puberty]
L6.u7886	[Marie, Anne]
L12.u12294	[preschool, kindergarten]
L5.u15471	[bos, borne]

Figure 2. Top five multimodal neurons (layer **L**, unit **u**), for sample images from 6 COCO supercategories. Superimposed heatmaps (0.95 percentile of activations) show mean activations of the top five neurons over the image. Gradient-based attribution scores are computed with respect to the logit shown in bold in the GPT caption of each image. The two highest-probability tokens are shown for each neuron.

2. Multimodal Neurons

Investigations of individual units inside deep networks have revealed a range of human-interpretable functions: for example, color-detectors and Gabor filters emerge in low-level convolutional units in image classifiers [8], and later units that activate for object categories have been found across vision architectures and tasks [44, 3, 31, 5, 16]. *Multimodal neurons* selective for images and text with similar semantics have previously been identified by Goh *et al.* [12] in the CLIP [33] visual encoder, a ResNet-50 model [14] trained to align image-text pairs. In this work, we show that multimodal neurons also emerge when vision and language are learned *entirely separately*, and convert visual representations aligned to a frozen language model into text.

2.1. Detecting multimodal neurons

We analyze text transformer neurons in the multimodal LiMBeR model [28], where a linear layer trained on CC3M [36] casts BEIT [2] image embeddings into the input space ($e_L = 4096$) of GPT-J 6B [43]. GPT-J transforms input sequence $x = [x_1, \ldots, x_P]$ into a probability distribution y over next-token continuations of x [42], to create an image caption (where $P = 196$ image patches). At layer ℓ, the hidden state h_i^ℓ is given by $h_i^{\ell-1} + \mathbf{a_i}^\ell + \mathbf{m_i}^\ell$, where $\mathbf{a_i}^\ell$ and $\mathbf{m_i}^\ell$ are attention and MLP outputs. The output of the final layer L is decoded using W_d for unembedding: $y = \mathrm{softmax}(W_d h^L)$, which we refer to as $\mathrm{decoder}(h^L)$.

Recent work has found that transformer MLPs encode discrete and recoverable knowledge attributes [11, 6, 26, 27]. Each MLP is a two-layer feedforward neural network that, in GPT-J, operates on $h_i^{\ell-1}$ as follows:

$$\mathbf{m_i}^\ell = W_{out}^\ell \mathrm{GELU}(W_{in}^\ell h_i^{\ell-1}) \tag{1}$$

Motivated by past work uncovering interpretable roles of individual MLP neurons in language-only settings [6], we investigate their function in a multimodal context.

Attributing model outputs to neurons with image input. We apply a procedure based on gradients to evaluate the contribution of neuron u_k to an image captioning task. This approach follows several related approaches in neuron attribution, such as Grad-CAM [35] and Integrated Gradients [39, 6]. We adapt to the recurrent nature of transformer token prediction by attributing neuron effects from image patches to generated tokens in the caption, which may be several transformer passes later. We assume the model is predicting c as the most probable next token t, with logit y^c. We define the **attribution score** g of u_k on token c after a forward pass through image patches $\{1, \ldots, p\}$ and pre-activation output Z, using the following equation:

$$g_{k,c} = Z_p^k \frac{\partial y^c}{\partial Z_p^k} \tag{2}$$

This score is maximized when both the neuron's output and the effect of the neuron are large. It is a rough heuristic, loosely approximating to first-order the neuron's effect on the output logit, compared to a baseline in which the neuron is ablated. Importantly, this gradient can be computed efficiently for all neurons using a single backward pass.

2.2. Decoding multimodal neurons

What effect do neurons with high $g_{k,c}$ have on model output? We consider $u_k \in U^\ell$, the set of first-layer MLP units ($|U^\ell| = 16384$ in GPT-J). Following Equation 1 and the formulation of transformer MLPs as key-value pairs from [11], we note that activation A_i^k of u_k contributes a "value" from W_{out} to h_i. After the first layer operation:

$$\mathbf{m_i} = W_{out} A_i \tag{3}$$

As A_i^k grows relative to A_i^j (where $j \neq k$), the direction of $\mathbf{m_i}$ approaches $W_{out}^k A_i^k$, where W_{out}^k is one row of weight matrix W_{out}. As this vector gets added to the residual stream, it has the effect of boosting or demoting

	BERTScore	**CLIPScore**
random	.3627	21.74
multimodal neurons	.3848	23.43
GPT captions	.5251	23.62

Table 1. Language descriptions of multimodal neurons correspond with image semantics and human annotations of images. Scores are reported for a random subset of 1000 COCO validation images. Each BERTScore (F1) is a mean across 5 human image annotations from COCO. For each image, we record the max CLIPScore and BERTScore per neuron, and report means across all images.

certain next-word predictions (see Figure 1). To decode the *language contribution* of u_k to model output, we can directly compute $\text{decoder}(W^k_{out})$, following the simplifying assumption that representations at any layer can be transformed into a distribution over the token vocabulary using the output embeddings [11, 10, 1, 34]. To evaluate whether u_k translates an image representation into semantically related text, we compare $\text{decoder}(W^k_{out})$ to image content.

Do neurons translate image semantics into related text? We evaluate the agreement between visual information in an image and the text multimodal neurons inject into the image caption. For each image in the MSCOCO-2017 [23] validation set, where LiMBeR-BEIT produces captions on par with using CLIP as a visual encoder [28], we calculate $g_{k,c}$ for u_k across all layers with respect to the first noun c in the generated caption. For the 100 u_k with highest $g_{k,c}$ for each image, we compute $\text{decoder}(W^k_{out})$ to produce a list of the 10 most probable language tokens u_k contributes to the image caption. Restricting analyses to interpretable neurons (where at least 7 of the top 10 tokens are words in the English dictionary containing $\geq$ 3 letters) retains 50% of neurons with high $g_{k,c}$ (see examples and further implementation details in the Supplement).

We measure how well decoded tokens (*e.g.* `horses, racing, ponies, ridden, ...` in Figure 1) correspond with image semantics by computing CLIPScore [17] relative to the input image and BERTScore [45] relative to COCO image annotations (*e.g. a cowboy riding a horse*). Table 1 shows that tokens decoded from multimodal neurons perform competitively with GPT image captions on CLIPScore, and outperform a baseline on BERTScore where pairings between images and decoded multimodal neurons are randomized (we introduce this baseline as we do not expect BERTScores for comma-separated token lists to be comparable to GPT captions, *e.g. a horse and rider*).

Figure 2 shows example COCO images alongside top-scoring multimodal neurons per image, and image regions where the neurons are maximally active. Most top-scoring neurons are found between layers 5 and 10 of GPT-J ($L = 28$; see Supplement), consistent with the finding from [26] that MLP knowledge contributions occur in earlier layers.

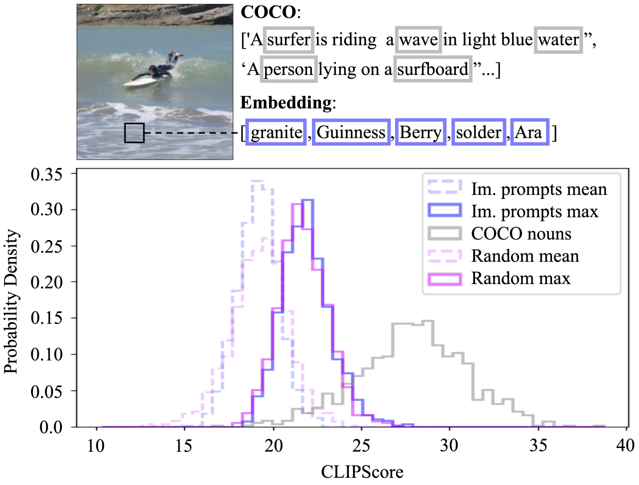


Figure 3. CLIPScores for text-image pairs show no significant difference between decoded image prompts and random embeddings. For image prompts, we report the mean across all image patches as well as the distribution of max CLIPScores per image.

	Random	**Prompts**	**GPT**	**COCO**
CLIPScore	19.22	19.17	23.62	**27.89**
BERTScore	.3286	.3291	**.5251**	.4470

Table 2. Image prompts are insignificantly different from randomly sampled prompts on CLIPScore and BERTScore. Scores for GPT captions and COCO nouns are shown for comparison.

3. Experiments

3.1. Does the projection layer translate images into semantically related tokens?

We decode image prompts aligned to the GPT-J embedding space into language, and measure their agreement with the input image and its human annotations for 1000 randomly sampled COCO images. As image prompts correspond to vectors in the embedding space and not discrete language tokens, we map them (and 1000 randomly sampled vectors for comparison) onto the five nearest tokens for analysis (see Figure 3 and Supplement). A Kolmogorov-Smirnov test [19, 37] shows no significant difference ($D = .037, p > .5$) between CLIPScore distributions comparing real decoded prompts and random embeddings to images. We compute CLIPScores for five COCO nouns per image (sampled from human annotations) which show significant difference ($D > .9, p < .001$) from image prompts.

We measure agreement between decoded image prompts and ground-truth image descriptions by computing BERTScores relative to human COCO annotations. Table 2 shows mean scores for real and random embeddings alongside COCO nouns and GPT captions. Real and random prompts are negligibly different, confirming that inputs to GPT-J do not readily encode interpretable semantics.

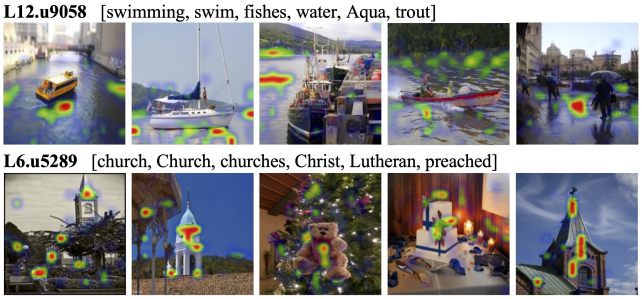

L12.u9058 [swimming, swim, fishes, water, Aqua, trout]
L6.u5289 [church, Church, churches, Christ, Lutheran, preached]

Figure 4. Top-activating COCO images for two multimodal neurons. Heatmaps (0.95 percentile of activations) illustrate consistent selectivity for image regions translated into related text.

3.2. Is visual specificity robust across inputs?

A long line of interpretability research has shown that evaluating alignment between individual units and semantic concepts in images is useful for characterizing feature representations in vision models [4, 5, 46, 16]. Approaches based on visualization and manual inspection (see Figure 4) can reveal interesting phenomena, but scale poorly.

We quantify the selectivity of multimodal neurons for specific visual concepts by measuring the agreement of their receptive fields with COCO instance segmentations, following [3]. We simulate the receptive field of u_k by computing A_i^k on each image prompt $x_i \in [x_1, ..., x_P]$, reshaping A_i^k into a 14 × 14 heatmap, and scaling to 224 × 224 using bilinear interpolation. We then threshold activations above the 0.95 percentile to produce a binary mask over the image, and compare this mask to COCO instance segmentations using Intersection over Union (IoU). To test specificity for individual objects, we select 12 COCO categories with single object annotations, and show that across all categories, the receptive fields of multimodal neurons better segment the object in each image than randomly sampled neurons from the same layers (Figure 5). While this experiment shows that multimodal neurons are reliable detectors of concepts, we also test whether they are selectively active for images containing those concepts, or broadly active across images. Results in the Supplement show preferential activation on particular categories of images.

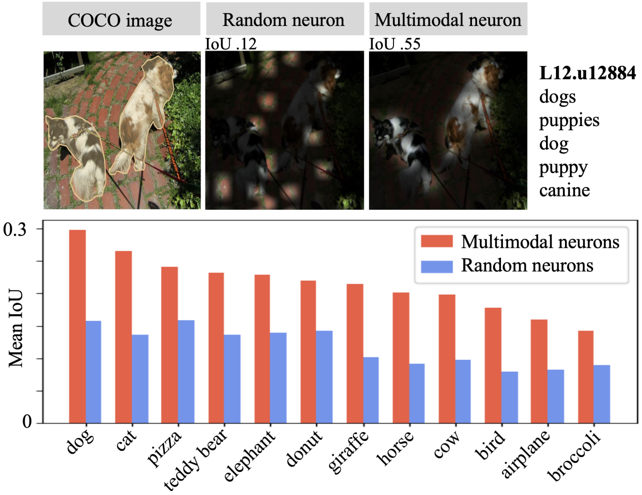


Figure 5. Across 12 COCO categories, the receptive fields of multimodal neurons better segment the concept in each image than randomly sampled neurons in the same layers. The Supplement provides additional examples.

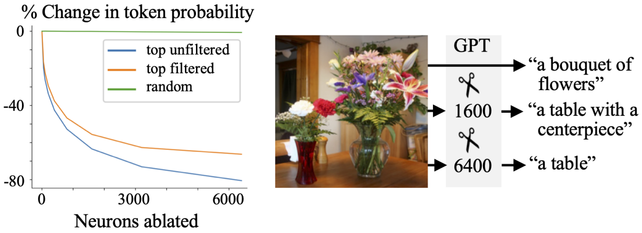


Figure 6. Ablating multimodal neurons degrades image caption content. We plot the effect of ablating multimodal neurons ordered by $g_{k,c}$ and randomly sampled units in the same layers (left), and show an example (right) of the effect on a single image caption.

3.3. Do multimodal neurons causally affect output?

To investigate how strongly multimodal neurons causally affect model output, we successively ablate units sorted by $g_{k,c}$ and measure the resulting change in the probability of token c. Results for all COCO validation images are shown in Figure 6, for multimodal neurons (filtered and unfiltered for interpretability), and randomly selected units in the same layers. When up to 6400 random units are ablated, we find that the probability of token c is largely unaffected, but ablating the same number of top-scoring units decreases token probability by 80% on average. Ablating multimodal neurons also leads to significant changes in the semantics of GPT-generated captions. Figure 6 shows one example; additional analysis is provided in the Supplement.

4. Conclusion

We find multimodal neurons in text-only transformer MLPs and show that these neurons consistently translate image semantics into language. Interestingly, soft-prompt inputs to the language model do not map onto interpretable tokens in the output vocabulary, suggesting translation between modalities happens *inside* the transformer. The capacity to align representations across modalities could underlie the utility of language models as general-purpose interfaces for tasks involving sequential modeling [25 , 13, 38, 29], ranging from next-move prediction in games [21, 32] to protein design [41, 9]. Understanding the roles of individual computational units can serve as a starting point for investigating how transformers generalize across tasks.

5. Limitations

We study a single multimodal model (LiMBeR-BEIT) of particular interest because the vision and language components were learned separately. The discovery of multimodal neurons in this setting motivates investigation of this phenomenon in other vision-language architectures, and even models aligning other modalities. Do similar neurons emerge when the visual encoder is replaced with a raw pixel stream such as in [25], or with a pretrained speech autoencoder? Furthermore, although we found that the outputs of the LiMBeR-BEIT projection layer are not immediately decodable into interpretable language, our knowledge of the structure of the vector spaces that represent information from different modalities remains incomplete, and we have not investigated how concepts encoded by individual units are assembled from upstream representations. Building a more mechanistic understanding of information processing within transfomers may help explain their surprising ability to generalize to non-textual representations.

6. Acknowledgements

We are grateful for the support of the MIT-IBM Watson AI Lab, and ARL grant W911NF-18-2-0218. We thank Jacob Andreas, Achyuta Rajaram, and Tazo Chowdhury for their useful input and insightful discussions.

References

[1] J Alammar. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations*, pages 249–257, 2021. 3

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 2, 4

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017. 4

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. 2, 4

[6] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arxiv:2104.08696*, 2022. 2

[7] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021. 1

[8] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 2

[9] Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence* , 4(6):521–532, 2022. 4

[10] Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. *arXiv preprint arXiv:2204.12130*, 2022. 3

[11] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020. 2, 3

[12] Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. https://distill.pub/2021/multimodal-neurons. 2

[13] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[15] Richard Held, Yuri Ostrovsky, Beatrice de Gelder, Tapan Gandhi, Suma Ganesh, Umang Mathur, and Pawan Sinha. The newly sighted fail to match seen with felt. *Nature neuroscience*, 14(5):551–553, 2011. 1

[16] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2022. 2, 4

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 3

[18] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 1

[19] Andrej N Kolmogorov. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933. 3

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1

[21] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022. 4

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[24] John Locke. *An Essay Concerning Human Understanding*. London, England: Oxford University Press, 1689. 1

[25] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 1, 2021. 1, 4, 5

[26] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. 2, 3

[27] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass editing memory in a transformer. *arXiv preprint arxiv:2210.07229*, 2022. 2

[28] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022. 1, 2, 3

[29] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. In *arXiv preprint arXiv:2307.04721*, 2023. 4

[30] Michael J. Morgan. *Molyneux's Question: Vision, Touch and the Philosophy of Perception*. Cambridge University Press, 1977. 1

[31] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 2

[32] Vishal Pallagani, Bharath Muppasani, Keerthiram Murugesan, Francesca Rossi, Lior Horesh, Biplav Srivastava, Francesco Fabiano, and Andrea Loreggia. Plansformer: Generating symbolic plans using transformers. *arXiv preprint arXiv:2212.08681*, 2022. 4

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[34] Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. What are you token about? dense retrieval as distributions over the vocabulary. *arXiv preprint arXiv:2212.10380*, 2022. 3

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. 2

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[37] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948. 3

[38] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 4

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. 2

[40] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 1

[41] Serbulent Unsal, Heval Atas, Muammer Albayrak, Kemal Turhan, Aybar C Acar, and Tunca Doğan. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245, 2022. 4

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[43] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021. 1, 2

[44] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 2

[45] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 3

[46] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018. 4