Multimodal neural networks better explain multivoxel patterns in the hippocampus

Anonymous Author(s)
Affiliation
Address
email

Abstract

The human hippocampus possesses "concept cells", neurons that fire when presented with stimuli belonging to a specific concept, regardless of the modality. Recently, similar concept cells were discovered in a multimodal network called CLIP [1]. Here, we ask whether CLIP can explain the fMRI activity of the human hippocampus better than a purely visual (or linguistic) model. We extend our analysis to a range of publicly available uni- and multi-modal models. We demonstrate that "multimodality" stands out as a key component when assessing the ability of a network to explain the multivoxel activity in the hippocampus.

1 Introduction

The field of machine learning has experienced tremendous breakthroughs in the past few years. A hallmark of these breakthroughs are deep neural networks (DNNs) that can solve complex tasks going beyond computer vision to tasks requiring semantic knowledge and understanding, features characteristic of human intelligence (e.g., story completions, context-based question answering, code generation etc.,). This feat has been made possible by both the ability of DNNs to learn expressive representational spaces that enable them to carry out these complex tasks, as well as by the development of improved optimization algorithms required to train the DNNs.

Importantly for the neuroscience community, DNNs also provide a potential model for understanding the human brain. Their mathematical pliability combined with their unprecedented expressivity has opened up novel avenues to investigate the human brain. Efforts are being made to understand the similarities and differences between these two systems due to their architectures, dynamics, behavioral patterns, and representational structures [2–5].

At the same time, DNNs themselves are getting better and better on more human-like tasks. Recently, Radford et al. [1] proposed a model that could simultaneously learn visual and linguistic information from a huge dataset using a constrastive loss function. Importantly, this multimodal model, known as CLIP, was found to possess neurons in its last layer that encoded specific concepts [6]. These artificial neurons are reminiscent of 'concept cells' in the human medial temporal lobe (MTL) [7, 8], biological neurons that appear to represent the meaning of a given stimulus or concept in a manner that is invariant to how that stimulus is actually experienced by the observer. For example, a single neuron in the human hippocampus showed incredible specificity in its response to the actress Halle Berry. This neuron responded to different images of the actress, including to photographs in which she was disguised as Catwoman (her starring role in a movie by the same name). The same neuron also responded to a semantic representation of the concept, i.e. to the letter string "HALLE BERRY". Other studies have since shown that "concept cells" are also activated when stimulus information is provided in other sensory modalities, for example when the name of the person is spoken out loud.

Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

The discovery of concept cells in artificial networks raises a natural question — *Can a multimodal model like CLIP explain the activity of brain regions known to possess concept cells better than a purely visual model?* In this work we investigate this question by using publicly available fMRI data [9], and asking if CLIP can explain the activity of the hippocampus region better than a comparable feedforward visual model, i.e., ResNet. Because fMRI data does not provide us with the spatial resolution to identify individual concept cells, we address this question at the level of multi-dimensional representation spaces rather than at the level of individual neurons. We also extend our analysis to a variety of models from the literature, trained with unimodal or multimodal objectives. Using Representational Similarity Analysis (RSA) [10], we report that multimodal networks consistently rank higher than their unimodal counterparts in their ability to explain fMRI activity in the human hippocampus. We provide all the code for reproducing our results on Github[1].

2 Methods

2.1 RSA

Representational Similarity Analysis (RSA) [10] compares stimulus representations across different high-dimensional spaces (e.g., brain multi-voxel spaces, model latent spaces, etc.). A first step in RSA consists of constructing Representational Dissimilarity Matrices (RDMs) in each space. RDMs are two-dimensional matrices, in which each element measures the pairwise distance between two stimulus conditions. An important property of RDMs is that their size is the same regardless of the initial dimensionality of each space, since the number of elements in an RDM only depends on the number of conditions being compared. RDMs from representational spaces of different dimensionalities can thus be compared to each other. In this work, we use the Pearson correlation distance (defined as 1 - correlation) to construct the RDMs, and subsequently compare them with the Pearson's r correlation coefficient. Results with other choices of metrics are shown in the Appendix.

2.2 fMRI data

For our investigations, we use publicly available data from [9]. This dataset consists of fMRI data collected on five healthy participants viewing images from a subset of categories available in ImageNet. Participants performed a one-back test in the scanner in which they had to press a button when the same image was repeated on two consecutive presentations. The data were collected on 1200 training images that were presented once, and 50 test images presented 35 times each. For our experiments, we restrict ourselves to the subset of test images since the higher number of repetitions provides a more robust estimate of the multi-voxel representation of each image.

After downloading the raw fMRI data, we preprocessed them with a standard pipeline using SPM12 [11]: slice-time correction, realignment, and coregistration to the T1W anatomical images. We performed a GLM using regressors for each image (the onset and duration), along with regressors for 'fixation' and 'one-back'. For each subject, the beta coefficients obtained from the GLM were transformed into a common MNI305 space [12] using FreeSurfer (`http://surfer.nmr.mgh.harvard.edu`) to allow analysis across subjects. We defined four regions of interest (ROIs) using the Deskian-Killiany atlas [13] for both the left and right hemispheres: a visual ROI comprising the lateraloccipital and pericalcarine regions, a fusiform ROI, a hippocampal ROI and a parahippocampal ROI. fMRI RDMs were built using the beta values in each ROI for each subject. Since 50 image conditions were compared, each RDM was 50x50 in squareform.

2.3 Models

We include a variety of models in our analysis to facilitate interpretation and discovery of underlying trends in different classes of models. All the included models are publicly available, and possess a ResNet50 backbone to minimize architectural differences.

For CLIP, we used the visual CLIP-RN50 backbone (called CLIP hereafter), which was jointly trained along with a linguistic head (called CLIP-L hereafter) on a contrastive learning task on 400M image-caption pairs [1]. Additionally we also considered visual features from TSM [14], another multimodal model that is trained with a contrastive objective on the HowTo100M dataset [15], in

[1] Link hidden to preserve anonymity

a task that comprises three modalities (video, text, and audio). The impact of contrastive learning objectives on the features of these models can be compared to VirTex and ICMLM, multimodal networks trained with different objectives. For Virtex, the visual backbone is trained on an image captioning task [16], while for ICMLM, the visual features are trained on a text-unmasking task [17]. Both VirTex and ICMLM are trained on MS-COCO [18], a much smaller dataset compared to those used for CLIP or TSM.

To tease apart the effect of multimodal training, we also included visual-only models in our comparisons. Since dataset size has been suggested to affect the quality of representations learned by a network, we considered two visual-only models trained on different datasets. We used the standard ResNet50 model (the control visual model) trained on ImageNet-1K, as well as BiT-M, a ResNet50 backbone trained on the significantly larger ImageNet-21K dataset [19]. We also included adversarially robust models (AR-L2, AR-L4, AR-L8) from [20] in our comparisons. These models are trained to be robust to minute perturbations to the input images by explicitly incorporating such perturbed (adversarial) images [21] in the training dataset. These models have been observed to possess more human-like features compared to standard feedforward versions [22], making them particularly relevant to our analysis.

Unlike human observers who rely on shape information, standard ImageNet models are strongly biased by the texture of images [23]. Therefore, Geirhos et al. [23] designed a stylized version of ImageNet to train models that have a stronger bias towards shape than texture. To assess whether representations optimized for human-like biases are better at explaining brain activity in MTL regions, we included three StylizedImageNet models in our comparisons: (i) a model pretrained on only StylizedImageNet (SIN) images, (ii) a model trained on SIN images and ImageNet combined (SIN-IN), and (iii) the SIN-IN model further fine-tuned on ImageNet (SIN-IN+FIN).

Finally, apart from visual and multimodal models, we also included language models: GPT-2 [24], BERT [25], as well as CLIP-L. Although these models are not trained to process visual data, they provide an important basis for comparison along with visual and multimodal networks.

For multimodal and visual backbones, we used the test images shown to the human participants and obtained their feature representations from the final average pooling layer. For language models, for each image, we encoded the text 'a photo of {ImageNet label of the image}' to obtain the latent representations. For each model, these latent representations were then used to obtain the RDMs, of shape 50x50.

2.4 Voxel Selection from anatomical ROIs based on Noise ceilings

We start by evaluating the signal of the selected beta coefficients. In each ROI, we calculated the noise-ceiling, defined as the average inter-subject correlation between RDMs. The noise ceiling provides an estimate of the reliability of the fMRI signal in a given ROI across subjects. Due to the visual nature of the task, the more visual regions (visual ROI, fusiform and parahippocampus) unsurprisingly showed higher values for the noise-ceiling (between 0.2 and 0.6). In contrast, the noise ceiling in the hippocampus was relatively low, and not significantly different from zero (-0.012 ± 0.012). This could be due, in part, to the fact that the fMRI signal in the hippocampus is generally less reliable. However, single neuron recordings in the human hippocampus have also revealed that only a small proportion of cells ($\approx$15% of recorded cells) is responsive to visual stimuli, and even fewer ($\approx$5%) qualify as "concept cells". In fact, the hippocampus is well-known for its implication in non-visual tasks, i.e. spatial navigation or memory retrieval and consolidation. If only a small subset of voxels in the hippocampus respond to visual stimuli, it stands to reason that a noise ceiling computed across all voxels would not capture any meaningful visual information.

To circumvent this issue, we defined a quantifiable criterion to select a limited number of voxels from each ROI. Specifically, we selected the N voxels with the highest beta value (for any of the 50 stimuli), and calculated the noise ceiling based on this voxel selection. We varied N systematically. As a control, we used random selections of N voxels. As Figure 1 shows, the noise ceiling in the more visual regions (visual region, fusiform, parahippocampus) increased rapidly and then stabilized after the inclusion of $\approx$20% of the total voxels. This was true, even when the voxels were randomly selected, indicating that most voxels in these regions carry information about visual stimuli. In the hippocampus however, the noise ceiling was virtually zero when based on random voxel selections: most hippocampal voxels do not appear to encode visual information. Nonetheless, when selecting

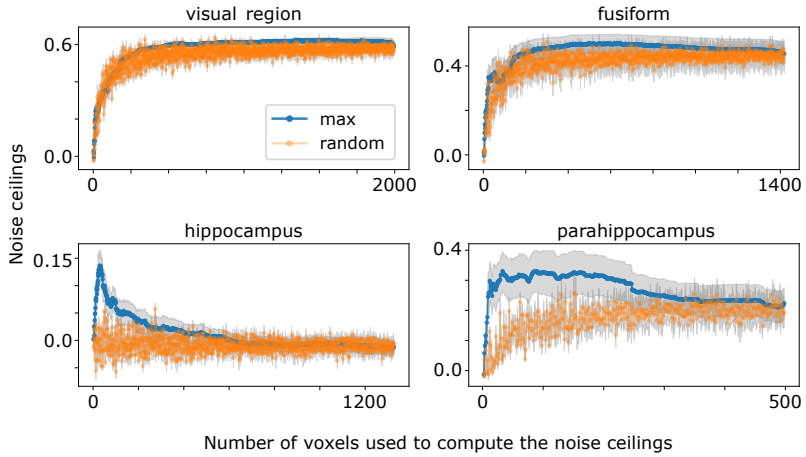

Figure 1: **Noise ceilings after selecting subsets of voxels from each region** The panels show the noise ceilings (i.e., inter-subject correlation) calculated after selecting different numbers of voxels from each region of interest. The noise ceilings were computed using either voxels with the highest beta values (blue) or via a random sampling of voxels (orange). The gray regions denote the standard error of mean. For certain ROIs (visual region, fusiform), most voxels are informative about the visual stimulus, and the two selection methods yield similar results. For other ROIs (hippocampus, parahippocampus), the noise ceiling depends on the selection method, implying that some voxels (with the highest betas) are more informative than others (randomly selected). The hippocampus shows an improved noise ceiling when 30 voxels with the highest beta values are selected, with additional voxels degrading the signal.

the N most-activated voxels, the noise ceiling peaked at ≈30 voxels, before sharply dropping down to random levels. This is consistent with our hypothesis that although a relatively small number of hippocampal voxels are reliably activated by visual inputs, the signal in these voxels (as measured by the noise ceiling) is reliably above chance. For the main RSA analysis, we thus considered only these top-30 hippocampus voxels. Note that this selection criterion only ensures that the considered brain responses are meaningful, but does not bias the outcome of the RSA with neural network models (i.e., there is no danger of circular reasoning). For the other ROIs, we also considered the top-30 voxels for a fair comparison; yet we also report a different selection procedure (based on a fixed beta threshold) in the Supplementary Material.

3 Results

To investigate whether CLIP explains multivoxel activity patterns in MTL regions better than its visual (or linguistic) counterparts, we computed RSA between the brain RDMs and each model RDM. The noise ceiling places an upper limit on brain-model comparisons because it is an estimate of inter-subject variability. Thus, we normalized the RSA values by the noise-ceiling to allow for comparisons across models and across regions. The normalized RSA values for each model in each ROI are shown in Figure 2A, and averaged across groups of models in Figure 2B. (The corresponding non-normalized values are shown in the Appendix.)

RSA values for the majority of models and brain regions were positive. However, comparisons between individual models (e.g., CLIP vs. ResNet in hippocampus) were not significant (Wilcoxon signed-rank test), possibly because of the small number of fMRI participants. Thus, we grouped the models according to their modalities (for example, BERT, GPT2 and CLIP-L as the *language models*). We thus obtained three classes of models (visual, language, multi-modal), and asked whether one class outperformed the others in explaining brain activity in each ROI. In the hippocampus, in line with our main hypothesis, multimodal models significantly better explained activity patterns compared to both visual and linguistic models (Welch's t-test. Figure 2B). In fact, the multimodal networks reached the noise ceiling in the hippocampus, meaning that they could explain all of the explainable variance in brain responses–this result did not happen for any other model group in any other ROI. A similar trend was observed in other regions (even reaching statistical significance in

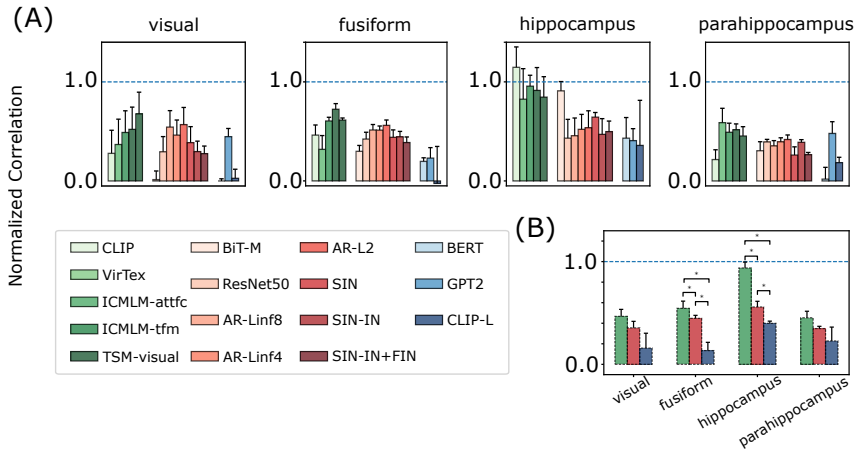

Figure 2: **Multimodal models better explain fMRI response patterns in the hippocampus**: Panel A shows the correlation values obtained with different models across selected regions of interest (ROI). Only 30 voxels were selected from each ROI. The values are normalized with the noise-ceilings to facilitate comparisons across regions. Panel B shows the correlation values after aggregating them over mutlimodal (green), visual (red) and language (blue) models. Statistical significance is calculated by using Welch's t-test and is denoted by an asterisk.

the fusiform ROI), but the RSA values were lower and more variable compared to the hippocampus. Finally, the visual and vision-language models performed systematically better than the linguistic models–as expected since all stimuli were visual.

In our main analysis, the RSA was performed using a subset of 30 voxels that showed the highest beta values in each ROI. While this threshold is reasonable in the hippocampus based on our noise-ceiling calculations (Fig 1), visual regions did have a larger number of voxels with reliable beta values. Thus, in a control analysis, in each ROI we selected the N voxels that had beta values greater than a common threshold (determined so as to yield 30 hippocampal voxels). Figure 4 in the appendix shows that including a larger number of voxels had little impact on the main results shown in Figure 2. Finally, we assessed if the trends observed in Figure 2 are robust to the choice of distance metrics by using other metrics commonly used for fMRI data (see Appendix).

4 Discussion and Conclusion

We applied RSA to study the ability of different neural network models – multi- or uni-modal – to explain the fMRI activity patterns in various brain regions. Based on recent findings [6], our hypothesis was that CLIP (and similar multimodal networks) would be specifically adept at explaining brain activity in the hippocampus–where 'concept cells' are found. This hypothesis was supported by the data: the *multimodal* nature of a model was as a key component in explaining the activity in the human hippocampus – a trend that proved robust to different methods of voxel selection and distance metrics.

Our findings could provide a novel insight for making more brain-like models. While many studies have reported a reliable correspondence between deep neural network representations and neural activity along the ventral visual pathway, the similarity is far from absolute [26]. In particular, Xu and Vaziri-Pashkam [27] casted doubt on the utility of DNNs for explaining higher regions in the brain. Our findings provide a potential way forward to address this limitation: building models that explain higher regions in the brain might require using datasets spanning different modalities. This can be further combined with bio-plausible architectural changes to the DNNs. For example, it would be interesting to investigate the effects of training a bio-inspired recurrent neural network [28] using multimodal objectives. Combining these architecture- and objective-based approaches could potentially have synergistic effects in learning human-like representations.

We hope the findings in this work further contribute to the efforts in bridging the gap between machine learning and neuroscience.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed in the Appendix Section A.5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All the code to reproduce the experimental results will be provided on Github.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] The experiments involved analysis of already pretrained models in the field. Apart from a small GPU to get representations of different images, no heavy computer resources were required.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] The licenses are mentioned in the Appendix Section A.4
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data was made publicly available by the authors.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data doesn't contain any personally identifiable information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[2] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[3] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35 (27):10005–10014, 2015.

[4] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

[5] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11): e1003915, 2014.

[6] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3): e30, 2021.

[7] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[8] Leila Reddy and Simon J Thorpe. Concept cells through associative learning of high-level representations. *Neuron*, 84(2):248–251, 2014.

[9] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[10] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/article/10.3389/neuro.06.004.2008.

[11] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

[12] D Louis Collins, Peter Neelin, Terrence M Peters, and Alan C Evans. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. *Journal of computer assisted tomography*, 18(2):192–205, 1994.

[13] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[14] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020.

[15] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[16] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021.

[17] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision (ECCV)*, 2020.

[18] 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. *Lecture Notes in Computer Science*, page 740–755, 2014. ISSN 1611-3349. doi: 10.1007/978-3-319-10602-1_48. URL http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019.

[20] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

[21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[22] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/24357dd085d2c4b1a88a7e0692e60294-Paper.pdf.

[23] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020.

[27] Yaoda Xu and Maryam Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1): 1–16, 2021.

[28] Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, Benjamin Ador, Andrea Alamia, and Rufin VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Neural Information Processing Systems (NeurIPS)*, 2021.