Brain-Like Object Recognition Neural Networks are more robustness to common corruptions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Previous work [12, 13] have shown that there exists a correlation between the performance of neural networks in object recognition tasks and its ability to match behavioral and neural recordings. We expanded on this work to ask the question: Does the behavioral and neural recordings are also correlated to the robustness of neural networks to common corruptions (e.g ImageNet-C). We selected several models from the leaderboard in Brain-Score, a platform that hosts neural and behavioral benchmarks for brain-model similarity, and tested their robustness to the corruption from ImageNet-C. We showed that higher brain-score is correlated with lower mean corruption error across models. Particularly, we show a correlation between the V4 and Behavioral datasets and the model's robustness to ImageNet-C. These finds suggest that explicitly modeling/matching data from V4 might be a good strategy for developing robust models to common corruptions.

1 Introduction

Perceptual Robustness is a key component of the human vision system, however, this robustness is not present in current state-of-the-art deep learning models [7, 5, 15, 14]. Furthermore, these models are not robust to small image perturbations such as fog, snow, blur, pixelation, etc, which humans are not confused by. This discrepancy between humans and computer vision models has to be addressed if we want current deep learning models to generalize on natural settings beyond training set statistics.

Nonetheless, improvements have been made to improve the robustness of deep learning models to common corruptions, mostly by training the models with different data augmentation techniques [4, 8, 11, 6]. However, there is still large room to match human level performance on ImageNet-C [7] and other robustness benchmarks. If we want to decrease this gap between humans and deep learning models, one strategy is to make computer vision models more brain-like. Previous work [16, 1, 2, 9], have studied how similar are these deep learning models to humans [], and found that high performing networks have similar representation to different visual cortical areas, and that the hierarchical structure of the visual representation from neural recordings is shared with the hierarchy of deep learning models.

Recently [12, 13], established a benchmark to compare different deep learning models in their ability to predict the activity of different visual cortical areas (V1, V2, V4 and IT). This was done by doing a linear regression from the features of the model given an image x with neural responses:

$$y = Xw + \epsilon, \tag{1}$$

.

where w denotes linear regression weights and ϵ is the noise in the neural recordings.

Submitted to 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM).

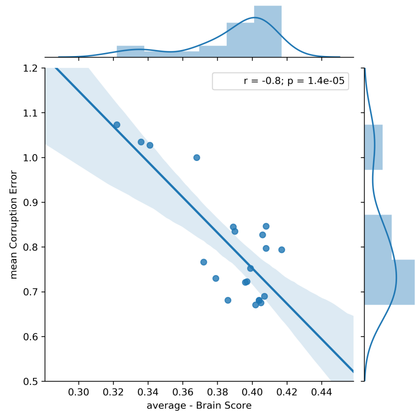

Figure 1: Scatter Plot of mean Corruption Error (mCE) with respect to the average Brain-Score. We observe a correlation (r=-0.8, p=1.4e-05) between these two variables. This indicates that higher brain score model have lower mCE values

They were able to reproduce previous work [16], and expand the predictability for behavioral tasks, by using the metric of the image-by-image patterns of difficulty, broken down by the object choice alternatives. We decided to expand on this work and ask the question: Is there a correlation between this Brain-Score benchmark and the robustness of a deep learning model to common corruptions? and if there is, what cortical areas are responsible for this correlation?

Main Contributions We showed that V4 and behavioral predictability are positively correlated with lower mean Corruption Error in ImageNet-C. Furthermore, we found that the predictability of V1 is anticorrelated with robustness to these common corruptions.

2 Experimental Results

To test this hypothesis, we used 20 deep learning models currently ranked on the Brain-Score website (See Table 1 for specific models), and extracted their brain score for each brain area (V1, V2, V4 and IT), behavioral score and their average score. Furthermore, we evaluate each model on the ImageNet-C benchmark. This dataset consists of 19 common corruptions (c) (See Sup. Figure **??** for examples) with 5 different severity levels (s) added into the validation set images of ImageNet. We evaluated all the 20 models in this benchmark by calculating the mean Corruption Error (mCE), which was computed by:

$$CE_c^f = (\sum_{s,c}^{5} E_{s,c}^f)/(\sum_{s,c}^{5} E_{s,c}^{Alexnet}) \tag{2}$$

Where the error for each corruption is normalized against AlexNet performance to measure the improvement in robustness with respect to the stablished deep learning model. Then, we did the Spearman's correlation between the average score for each model with their corresponding mCE from ImageNet-C. In Figure 1, we observe the scatter plot we found a negative correlation between the average brain score and the mean corruption error (Lower corruption error means higher performance). Now that we established that the average brain-score is correlated with mCE, we asked, are all components of the brain predictability negatively correlated with the mCE?

For this, we computed the scatter plot and correlation for each individual component of the brain-score: V1 predictability, V2 predictability, V4 predictability, IT predictability and behavioral predictability against the mCE. In Figure 2, we observe that most of the Brain scores are correlated with lower mCE (V2, V4, IT, and behavioral scores). Particularly, V4 and behavioral predictability have p values lower than 0.05. Interestingly, we have a positive correlation between V1 predictability and mCE, which suggests that V1-like have less robustness to common corruption compared to other areas. This is a surprising result, however, previous work [3] has shown that a V1-Like model is not more

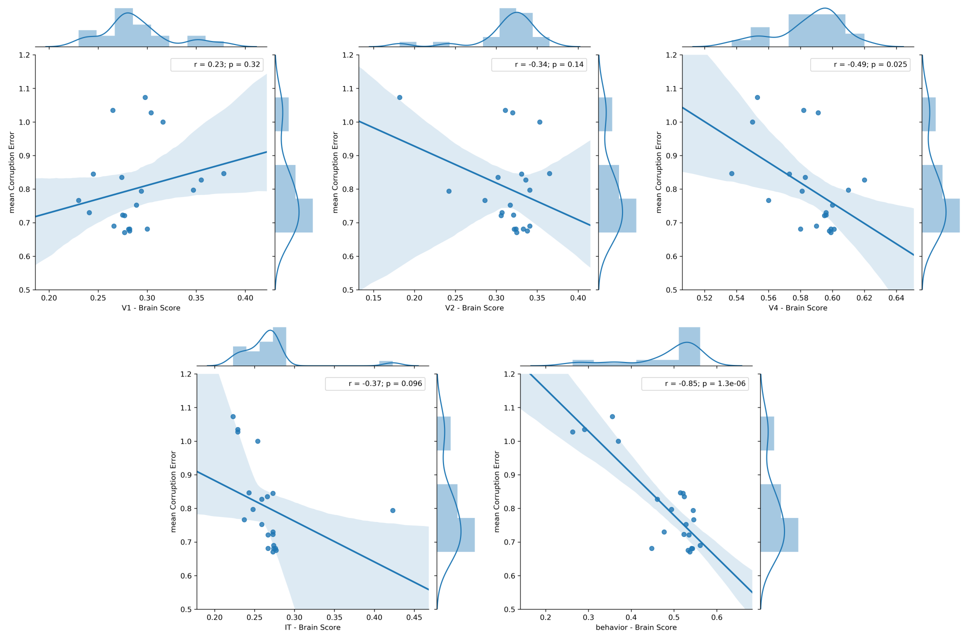

Figure 2: Scatter plot of mean Corruption Error (mCE) with respect to brain-score for Top:(left) V1 neural predictability, (middle) V2 neural predictability and (right) V2 predictability. Bottom: (Left) IT Predictability, (Right) Behavioral Predictability.

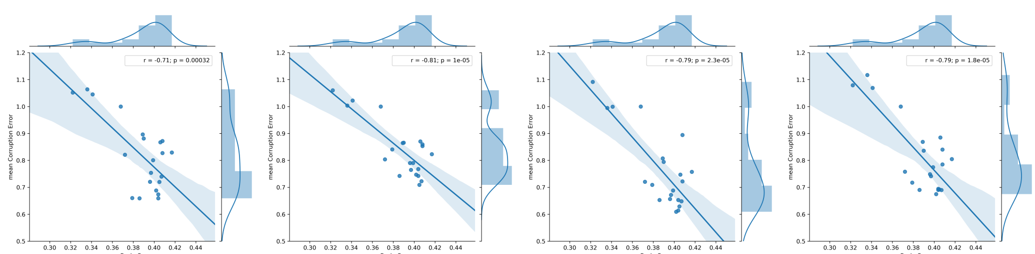

Figure 3: Scatter plot of mean Corruption Error (mCE) with respect to brain-score for different corruption Families. Left: Noise, Middle Left: Blur, Middle Right: Weather, Right: Digital. We found that all corruption families have a negative correlation with the average brain-score.

robust than other models to common corruptions, however, they were more robust to adversarial perturbations [10].

Given these results, another aspect we decided to explore whether higher brain-score was correlated with robustness against all common corruptions or was it correlated with an specific family of image corruption presented on Imagenet-C? Within, ImageNet-C there are 4 corruption families: Noise, Blur, Weather and Digital. Each of these families have different properties and therefore you could obtain robustness to one without gaining robustness on the other ones. To test this, we calculated the correlation between different models and the mCE to the different common corruption families. In Figure 3, we observe the scatter plot between mCE for each specific corruption family and the average brain-score. We observe that the correlation between each corruption family and the average brain-score is the same as with the mean Corruption Error (with $p < 0.05$ for all corruptions). This shows that higher correlation between mCE and brain-score is not due to improvement in an specific corruption family but an improvement for across all corruptions. This is an interesting result because in theory brain-like models should be equally robust to all these types of common corruptions and we observe that there is not bias in performance towards an specific corruption type.

	Noise			Blur				Weather				Digital				
Models	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	**mCE**
Alexnet	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
CORnet$_S$	82	83	86	79	88	81	83	84	80	73	63	76	84	81	79	79
CORnet$_Z$	105	104	104	107	105	106	102	104	107	108	116	106	108	103	113	107
Densenet121	71	72	74	76	87	77	78	74	71	59	56	62	87	73	76	72
Densenet169	66	67	70	71	84	76	79	69	67	56	52	55	82	68	70	68
Densenet201	67	70	70	70	82	73	77	68	66	57	50	57	79	63	69	67
Inception v3	65	66	66	81	88	81	89	76	73	70	63	69	87	59	71	72
Mobilenet v2	89	90	90	82	94	85	88	87	87	79	69	81	89	93	83	84
Resnet101	73	75	76	67	81	70	74	73	70	62	53	66	77	64	67	69
Resnet152	72	73	76	66	81	65	74	70	67	62	50	67	75	68	65	67
Resnet18	87	88	91	83	91	86	88	86	84	78	68	78	90	80	85	83
Resnet34	81	83	84	76	86	79	84	79	77	69	61	71	86	70	74	76
Resnet50	79	81	82	74	88	78	79	77	74	66	56	71	84	76	76	75
Resnet50SIN	66	66	68	69	81	69	80	68	70	64	57	66	78	61	69	68
Resnet50SIN-IN	66	66	68	69	81	69	80	68	70	64	57	66	78	61	69	68
Resnet50SIN-IN-IN	75	76	77	71	86	73	79	74	72	66	55	67	80	74	73	72
Resnet50-robust	86	86	92	92	79	85	81	81	83	111	81	105	79	60	68	84
Squeezenet1.0	105	104	103	100	103	105	101	101	103	97	97	97	109	107	113	102
Squeezenet1.1	107	106	105	99	102	100	100	100	102	96	97	97	105	109	133	103
VGG-16	86	87	89	83	94	86	87	83	79	72	63	75	95	94	88	82
VGG-19	82	82	88	81	93	83	86	80	78	68	61	73	93	85	83	79

Table 1: mCE, and Corruption Error values of different corruptions and architectures on IMAGENET-C. The mCE value is the mean Corruption Error of the corruptions in Noise, Blur, Weather, Extra, and Digital columns.

3 Conclusions and Future Work

We found a correlation between V4 neural predictability and behavioral predictability, and performance on ImageNet-C. However, this correlation is not found for V2 and IT (See Figure 2), this is perplexing given that a more brain-like model should have high predictability across brain areas and low mean corruption error. Furthermore, we found an anti-correlation between V1 predictability and mean corruption error. For future work, we want to expand this work to other robustness dataset such as ImageNet-P and CIFAR100-C to see if our results also hold for these datasets. Also, given previous work on the adversarial robustness of V1-Like models [3], we want to explore the correlation between brain-score and adversarial robustness. In addition, we want to generate models that have explicit high predictability of V4 and see if this model outperforms other models on ImageNet-C and other common corruption datasets. Finally, we want to expand the brain-score evaluation for models that are robust to ImageNet-C such as the ones from [8], [6] and [11].

References

[1] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[2] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol*, 10(12):e1003963, 2014.

[3] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. D. Cox, and J. J. DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

[4] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[5] K. Gu, B. Yang, J. Ngiam, Q. Le, and J. Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019.

[6] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[7] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[8] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[9] A. Kell, D. Yamins, S. Norman-Haignere, D. Seibert, H. Hong, J. DiCarlo, and J. McDermott. Computational similarities between visual and auditory cortex studied with convolutional neural networks, fmri, and electrophysiology. *Journal of vision*, 15(12):1093–1093, 2015.

[10] Z. Li, W. Brendel, E. Walker, E. Cobos, T. Muhammad, J. Reimer, M. Bethge, F. Sinz, Z. Pitkow, and A. Tolias. Learning from brains how to regularize machines. In *Advances in Neural Information Processing Systems*, pages 9529–9539, 2019.

[11] E. Rusak, L. Schott, R. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel. Increasing the robustness of dnns against image corruptions by playing the game of noise. *arXiv preprint arXiv:2001.06057*, 2020.

[12] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018.

[13] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020.

[14] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[15] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[16] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. 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.