Post-training dimensionality reduction of CNN layers supports better fit to early visual cortex

Navneedh Maudgalya
University of California, Berkeley
navneedhm@berkeley.edu

Joshua C. Peterson
Princeton University
peterson.c.joshua@gmail.com

Abstract

In the following short report, we detail our 4th-place submission to Track 1 of the 2019 Algonauts Project competition aimed at facilitating the use of AI to explain human brain activity. In particular, as a result of little success in model selection and fine-tuning, we focus instead on feature relevance/importance. We find that removing CNN features with low variability across the relevant stimulus set results in robust, layer-agnostic improvements in fit to representational distance matrices based on fMRI-measured brain activity. This effect is particularly pronounced in early visual cortex (EVC) compared to inferotemporal cortex (IT). Using this simple method, we obtain 4th place in Track 1 of the competition with 56-77% fewer overall submissions than the top three teams.

1. Introduction

The success of modern deep neural networks for human-relevant computer vision tasks and the powerful representations they learn have prompted their use as models of human vision [1, 6]. The 2019 Algonauts Project competition [1] is one recent call to aim these networks directly at explaining human brain activity. Training these networks explicitly to better match brain activity, often measured by representational distance matrices (RDMs) [2], is difficult, since brain activity is expensive to record for the viewing of many images, resulting in small datasets for both training and evaluation. One alternative is to perform exhaustive search for the pretrained network architecture and training conditions that better fit the brain, but this method may be time-consuming and fail to generalize.

In the following short report, we detail our first attempt at a simpler method for irrelevant feature removal that may be appropriate for earlier visual brain areas than others. Using our method, we obtain 4th place in the fMRI track of the Algonauts competition (i.e., predicting representations in early visual cortex (EVC) and inferotemporal cortex (IT)) with notably fewer overall submissions.

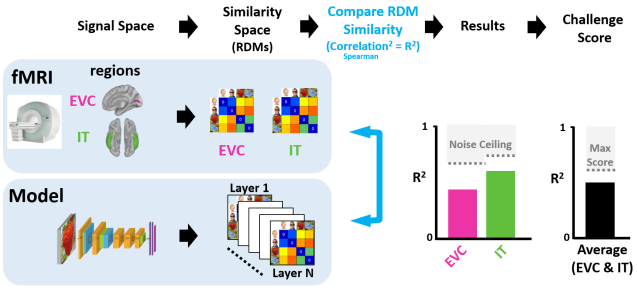

Figure 1. Track 1 (fMRI): RDMs are formed and compared by negating the pairwise correlations between image representations for fMRI activity during viewing (top left) and features extracted from a candidate vision model (bottom left). Figure used with permission from http://algonauts.csail.mit.edu/.

2. Approach

Following the competition baseline, our approach starts with representations extracted from modern convolutional neural networks (CNN) [3]. We then focus on dimensionality reduction as a means to remove irrelevant features as opposed to model/layer search, which we found unreliable. Before explaining this process in more detail, we review our initial failed strategies that motivated the final method.

2.1. Initial failures

We found little success in systematically selecting the best CNN models and layers, as our test performance was not correlated with CNN fit to training data, even though we used raw representations and no additional model parameters were fit. In addition, the 92-image set appeared to be even less relevant to the test set. Going forward, we set out to find other factors that might more robustly influence fit, such as CNN representation transformations or removal of unimportant features. The former was ultimately unsuccessful (using the linear method of [4]), likely due to overfitting despite our use of crossvalidation.

#	User	Entries	Date of Last Entry	Score		EVC		IT	
				Average R.^2 ▲	Noise Normalized Average R.^2 (%) ▲	R.^2 ▲	Noise Normalized R.^2 (%) ▲	R.^2 ▲	Noise Normalized R.^2 (%) ▲
-	Noise Ceiling	1	04/01/2019	0.0644	100.00	0.0640	100.00	0.0647	100.00
1	agustin	248	07/01/19	0.0173 (1)	26.91 (1)	0.0210 (1)	32.88 (1)	0.0136 (1)	20.99 (1)
2	rmldj	188	07/01/19	0.0158 (2)	24.56 (2)	0.0182 (3)	28.40 (3)	0.0134 (2)	20.77 (2)
3	Aakash	224	06/30/19	0.0155 (3)	24.03 (3)	0.0196 (2)	30.56 (2)	0.0114 (6)	17.56 (6)
4	navneedhm	82	07/01/19	0.0135 (4)	20.94 (4)	0.0137 (6)	21.40 (6)	0.0132 (3)	20.48 (3)

Figure 2. Leaderboard for Track 1 (fMRI) at the time of this write-up. (our team is `navneedhm`)

2.2. Choosing a base model and layer

Following the outcomes discussed in the previous section, we decided to freeze our model and layer search, taking the model and layer for our best current leaderboard submission at the time. For this reason, both our base model and layer may be suitable but not optimal starting points. In particular, we settled on the `activation_8` layer from Inception-v3 in the applications module of the Keras Python library but pre-trained using non-standard labels (see [7, 5]). Oddly enough, this single layer did best on both EVC and IT RDMs. Again, we suspect that these choices are not particularly important. In the following section, we detail the method we used to improve these base representations.

2.3. Dimension reduction strategies

The primary assumption of our final method is that many of potentially thousands of CNN features may not be relevant to either the brain representation or the images in the stimulus sets (since CNNs are often trained with many more images/categories). While removing features constitutes loss of information about the image, including features not relevant to the images or brain data may result in suboptimal representational distances.

One way to combat such features is to learn weights for each feature, but this method risks overfitting (see section 2.1). Instead, we adopted a simpler strategy for removing irrelevant features based on their variation observed across the training images. Specifically, we ranked each feature column by L1 norm, and removed those features with the lowest scores. By varying N, we can determine the optimal number of features to remove.

3. Analysis & Competition Results

Results are shown for each brain area and layer on the last page in Figure 3 for the best 8 layers for each brain area. In most cases, R^2 increases for EVC, often until many fewer features are left, but not for IT.

For Track 1 of the competition, our final submission used the network and best layer above, which 80% of the features removed for EVC, and 30% removed for IT. Our final R^2 scores are shown in the fourth column of the leaderboard in Figure 2.

4. Discussion & Conclusion

Increasing the correspondence between CNN representations and human brain activity is a non-trivial task given the scarcity of training data to learn from. In this report, we highlight one method that may reliably increase this correspondence with a potentially smaller risk of overfitting: simple dimensionality reduction, where the only free parameters are the reduction method and level of reduction. While our success in the competition lends some credence to the method, future work will be needed to verify the extent to which our analysis generalizes.

References

[1] R. M. Cichy, G. Roig, A. Andonian, K. Dwivedi, B. Lahner, A. Lascelles, Y. Mohsenzadeh, K. Ramakrishnan, and A. Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*, 2019. 1

[2] N. Kriegeskorte and R. A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013. 1

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[4] J. C. Peterson, J. T. Abbott, and T. L. Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018. 1

[5] J. C. Peterson, P. Soulos, A. Nematzadeh, and T. L. Griffiths. Learning hierarchical visual representations in deep

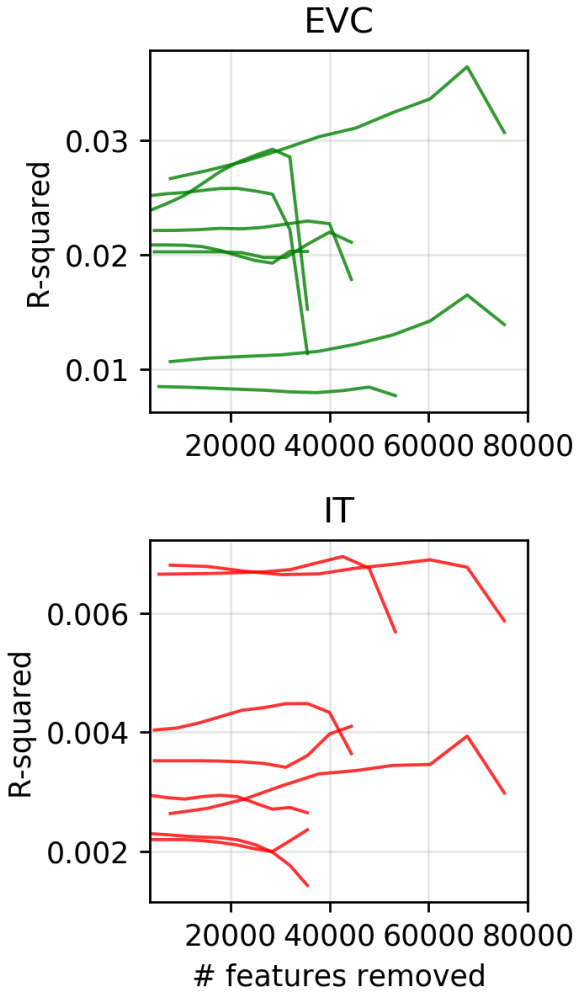

Figure 3. Effect of removing N features from the CNN representations for each brain area (EVC and IT). Each line is one CNN layer.

neural networks using hierarchical linguistic labels. *ArXiv*, abs/1805.07647, 2018. 2

[6] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, K. Schmidt, et al. Brain-score: which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018. 1

[7] P. Wang and G. W. Cottrell. Basic level categorization facilitates visual object recognition. *arXiv preprint arXiv:1511.04103*, 2015. 2