Examining Representational Similarity in ConvNets and the Primate Visual Cortex

Abhimanyu Dubey, Jayadeva and Sumeet Agarwal

Indian Institute of Technology Delhi

Abstract. We compare several ConvNets with different depth and regularization techniques with multi-unit macaque IT cortex recordings and assess the impact of the same on representational similarity with the primate visual cortex. We find that with increasing depth and validation performance, ConvNet features are closer to cortical IT representations.

1 Introduction

With the success of convolutional neural networks in object image classification [1–4], it is of interest to examine how similar the representations learnt by these networks are to the visual representations learnt by the human brain [5–7]. In Cadieu *et al.*'s pioneering work on understanding representational similarity [8], a comparison of the pre-final layer activations of deep neural net models with multi-unit IT cortex responses is presented, which confirms a significant correspondence between the two representations for the task of core object recognition. We aim to shed more light on the role of engineered aspects of these deep networks, examining the effect of variation in regularization, network depth and model size on the representational similarity to multi-unit IT responses and classification accuracy. This can help understand whether sparsity and increased network depth create representations which are closer to representations employed by the primate visual cortex, and also validate the hypothesis that the mammalian visual cortex is the best known object detector, by checking whether increased proximity to cortical representations implies an increase in recognition performance.

2 Methodology

For each ConvNet, we supply input images from n output classes and perform a feed-forward operation to obtain the layer-wise activations, and extract the activation at the penultimate fully-connected layer. For any network architecture N, we denote the set of activations for the image set by $\mathbf{A}_N$. For any particular target class $i \in [1, n]$, the set of activations is given by $\mathbf{A}_N^i$. For any set of activations, we can define the *average activation* as $\phi(\mathbf{A}_N^i) = (\sum_{a \in \mathbf{A}_N^i} a)/(|\mathbf{A}_N^i|)$.

Fig. 1.   (Reproduced from [8].) Samples from the dataset introduced by [8].

Given these representations, we can compute the *representational dissimilarity matrix* $\mathbf{R}_N$ as ([8])

$$\mathbf{R}_N = (r_{ij}) = 1 - \frac{\sigma(\phi(\mathbf{A}_N^i), \phi(\mathbf{A}_N^j))}{\sqrt{\sigma^2(\phi(\mathbf{A}_N^i)) \times \sigma^2(\phi(\mathbf{A}_N^j))}}; \mathbf{R}_N \in \mathbb{R}^{n \times n}$$

Where σ and σ^2 denote covariance and variance respectively. We compute this matrix for the cortical responses as well, treating the responses as the set of activations as $\mathbf{R}_{IT}$. To measure representational similarity, we use "similarity to IT dissimilarity matrix" (s_{IT}) as the default metric [8], which is as the Spearman's rank correlation between the upper triangular, non-diagonal elements of the two matrices $\mathbf{R}_N$ and $\mathbf{R}_{IT}$. In essence, this metric would be expected to encapsulate the nature of variation across classes for each representation [8].

To create the image set for activations, we randomly sample images from the total set in a manner identical to [8]. We also add noise to the ConvNet activations in order to account for the measurement noise present in the IT cortical responses following the experimental noise matched model in [8].

Dataset Cadieu *et al.* [8] introduce a dataset of multi-unit V4 and IT cortex responses across two male rhesus monkeys while they were presented synthetically generated object image samples. These recordings were taken from the 128 most visually driven neural measurement sites determined via a separate pilot dataset. We employ these recordings as the inferior-temporal (IT) representations (features) for our comparative experiments, and we have a total of 1,960 images for 7 categories (see Figure 1).

Evaluation We evaluate the best performing ConvNet architectures on ImageNet LSVRC and employ their publicly available pre-trained weights. We experiment on both the original architectures and retrained architectures using different regularization schemes - L1, L2 (default) and DeCov [9]. DeCov tries to learn independent, generalizable filters by minimizing the covariances of filter activations, with the additional DECOV loss $\mathcal{L}_{DeCov}$ at each hidden layer given by ([9])

$$\mathcal{L}_{DeCov} = \frac{1}{2}(\|C\|_F^2 - \|diag(C)\|_2^2)$$

$$C = (c_{i,j}) = \frac{1}{|\mathbf{A}_N^{batch}|} \sum_k (h_i^k - \mu_i)(h_j^k - \mu_j)$$
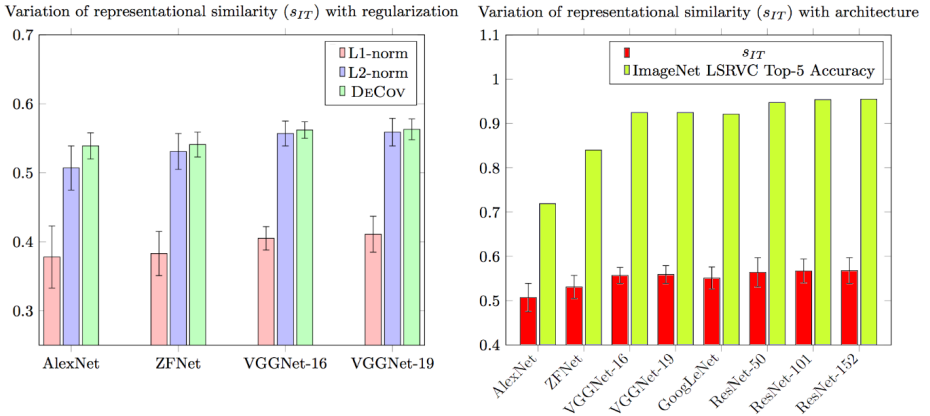
Fig. 2. Variation of representational similarity s_{IT} with model regularization (left) and validation performance (right). We see that representational similarity consistently increases with validation performance across architectures.

The $diag(\cdot)$ operator extracts the main diagonal of a matrix into a vector. The function C is the matrix of covariances of all pairs of activations h_i^k, h_j^k at a hidden layer k, and $\mathbf{A}_N^{batch} \subset \mathbf{A}_N$ is the sample batch during training of architecture N, and $\mu_i = \frac{1}{|\mathbf{A}_N^{batch}|} \sum_k h_i^k$ is the sample mean of activation i over the batch. We add this loss to each layer of the DeCov network [9] and tune the hyperparameter via cross-validation. For L1, we replace weight-decay (L2) with L1-norm, and tune via cross-validation. We maintain the original weight decay formulation and weights for L2. All networks are trained with dropout.

3   Results and Conclusions

While Cadieu *et al.* [8] validated the plausibility of deep neural net representations being similar to cortical representations, our work seeks to examine a wider range of variation in the characteristics of the deep learning models used. Our results show that by increasing network depth, we observe a marginal increase in representational similarity, which is also consistent with an increase in validation accuracy. Another interesting result is that by decorrelating representations, we observe a higher representational similarity, even though validation performance is similar. Our experiments offer a preliminary understanding of the effect of network depth and model complexity control on the similarity between deep neural net and cortical representations. Such an approach can help provide some pointers to the learning architectures and mechanisms employed by the brain. Substantial further work needs to be done: comparison of lower-layer activations of the network with preliminary regions of object recognition (V1, V2), visualization of the effect of context on representational similarity and understanding the impact of dropout and structural risk minimization on networks are some extensions we have initiated. The ultimate aim is to create models

Deep Network	Layers	s_{IT}	Accuracy on [8]	ImageNet Top-5 Error Rate
Original Architectures				
AlexNet [1]	8	0.507 ($\pm$0.032)	0.523 ($\pm$0.025)	18.1%
ZFNet [10]	8	0.531 ($\pm$0.026)	0.568 ($\pm$0.019)	16.0%
VGGNet-16 [2]	16	0.557 ($\pm$0.018)	0.580 ($\pm$0.021)	7.5%
VGGNet-19 [2]	19	0.559 ($\pm$0.020)	0.582 ($\pm$0.023)	7.5%
GoogLeNet [3]	22	0.551 ($\pm$0.025)	0.575 ($\pm$0.021)	7.89%
ResNet-50 [4]	50	0.564 ($\pm$0.033)	0.601 ($\pm$0.011)	5.25%
ResNet-101 [4]	101	0.567 ($\pm$0.027)	0.603 ($\pm$0.010)	4.60%
ResNet-152 [4]	152	0.568 ($\pm$0.029)	0.612 ($\pm$0.013)	4.49%
Modified Architectures				
AlexNet-L1[1]	8	0.378 ($\pm$0.045)	0.413 ($\pm$0.019)	27.2%
ZFNet-L1[1]	8	0.383 ($\pm$0.038)	0.431 ($\pm$0.017)	23.9%
VGGNet-16-L1[1]	16	0.405 ($\pm$0.017)	0.434 ($\pm$0.029)	13.5%
VGGNet-19-L1[1]	19	0.411 ($\pm$0.026)	0.437 ($\pm$0.015)	12.9%
AlexNet-DeCov [9]	8	0.539 ($\pm$0.019)	0.521 ($\pm$0.027)	20.0%
ZFNet-DeCov [9]	8	0.541 ($\pm$0.018)	0.528 ($\pm$0.025)	18.8%
VGGNet-16-DeCov [9]	16	0.562 ($\pm$0.012)	0.556 ($\pm$0.019)	11.6%
VGGNet-19-DeCov [9]	19	0.563 ($\pm$0.015)	0.561 ($\pm$0.024)	11.2%

Table 1. A comparison of the representational similarity of ConvNets to the IT cortex (s_{IT}), as per [8]. Original accuracies on ImageNet are reported directly. Accuracy on the data set of [8] is obtained using a linear SVM trained on pre-final activations.

similar to humans in both representation and performance on complex vision tasks, as a means of better understanding, or 'reverse engineering', the human visual system itself.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
3. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1–9
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
5. 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) (2014) 8619–8624
6. Güçlü, U., van Gerven, M.: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. The Journal of neuroscience: the official journal of the Society for Neuroscience **35**(27) (2015) 10005–10014
7. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A.: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports **6** (2016)
8. Cadieu, C.F., Hong, H., Yamins, D.L., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J.: Deep neural networks rival the representation of primate it cortex for core visual object recognition. PLoS Comput Biol **10**(12) (2014) e1003963
9. Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., Batra, D.: Reducing overfitting in deep networks by decorrelating representations. arXiv preprint arXiv:1511.06068 (2015)
10. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, Springer (2014) 818–833