CONVOLUTIONAL NEURAL NETWORKS AS A MODEL OF VISUAL ACTIVITY IN THE BRAIN: GREATER CONTRIBUTION OF ARCHITECTURE THAN LEARNED WEIGHTS

Anna Truzzi & Rhodri Cusack *
Trinity College Institute of Neuroscience
Trinity College Dublin
Dublin 2, Ireland
truzzia@tcd.ie, cusackrh@tcd.ie

ABSTRACT

Convolutional neural networks (CNNs) have proven effective as models of visual responses in the inferior temporal cortex (IT). The belief has been that training a network for visual recognition leads it to represent discriminative features similar to those the brain has learned. However, a CNN's response is affected by its architecture and not just its training. We therefore explicitly measured the effect of training different CNN architectures on their representational similarity with IT. We evaluated two versions of AlexNet and two training regimes, supervised and unsupervised. Surprisingly, we found that the representations in a random-weight variant of AlexNet, reflect brain representations in IT better than the benchmark supervised AlexNet and also better than the corresponding network trained in either a supervised or unsupervised manner. These results require a re-evaluation of the explanation of why CNNs act as an effective model of IT.

1 INTRODUCTION

When a human or non-human primate sees an object, activity cascades from the primary visual cortex at the back of the brain forwards through a set of regions to the inferior temporal (IT) cortex, leading to object recognition. The best current models of IT activity use convolutional neural networks (CNNs) pre-trained for object recognition (Lindsay, 2020) - often AlexNet trained on ImageNet (Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015; Cichy et al., 2016). Trained CNNs have been found to reflect the processing hierarchy in the brain, with representations in earlier layers more similar to those in the primary visual cortex and later layers with IT (Yamins & DiCarlo, 2016; Güçlü & van Gerven, 2015; Cichy et al., 2016; Wen et al., 2018). During training, CNNs learn discriminative features, which emphasise inter-class differences while becoming invariant to within-class variation (Achille & Soatto, 2017). It has often been assumed that the correspondence of pre-trained CNNs with IT happens because they learn similar discriminative features to the brain (Yamins & DiCarlo, 2016; Lindsay, 2020). However, while this idea is seductive, representations in CNNs are shaped not just by trained weights but also by model architecture (Pinto et al., 2009). For example, there is evidence that random networks can extract visual features that are surprisingly effective for visual recognition tasks (Saxe et al., 2011; Gaier & Ha, 2019; Jarrett et al., 2009). But, are such features present in IT? Furthermore, although the correspondence between the brain and trained network representations has been tested in supervised training regime we lack information about the effects of unsupervised training regime on the networks' activation patterns. Here we aim to distinguish the effects of training and architecture, by calculating the similarity of representations in a random CNNs with the brain, and by testing how training changes the correspondence between the CNNs and IT. The results suggest that architecture may play a larger role than training for some CNNs.

*www.cusacklab.org; code and data available at: https://annatruzzi.github.io/BrainVsCNNs/

2 CHOICE OF NETWORKS, TRAINING, AND ANALYSIS

Standard AlexNet. We used the standard AlexNet architecture (Krizhevsky et al., 2012) from torchvision.models, as used in previous studies (Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015; Cichy et al., 2016). The model was trained on ImageNet (top 1-accuracy: 56.55) and was not fine-tuned to the 92 visual stimuli. We examined representations at the five convolutional layers and the two fully connected layers.

DeepCluster with unsupervised training. As unsupervised algorithm, we evaluated DeepCluster (Caron et al., 2018), trained on ImageNet and not fine-tuned to the 92 visual stimuli (top 1-accuracy from 5th convolutional layer: 36.1). The underlying convolutional network of DeepCluster is an AlexNet architecture (Krizhevsky et al., 2012) modified for unsupervised learning Caron et al. (2018), with the local response normalisation layers removed and batch normalisation used instead (Ioffe & Szegedy, 2015) and an initial linear transformation based on Sobel filters applied on the input to remove colour and increase local contrast.

DeepCluster AlexNet with supervised training. Any difference in the results between standard AlexNet architecture and DeepCluster might be due either to their difference in architecture or to the different training process, supervised vs unsupervised. To control for this, we repeated the experiments using the same AlexNet variant as in DeepCluster, but trained on ImageNet with a supervised regime and not fine-tuned to the 92 visual stimuli (top 1-accuracy: 35.9).

Comparison with the brain. Activity in response to 92 images, as used by Khaligh-Razavi & Kriegeskorte (2014), was measured in the CNNs and in the brains of 15 adults. In the CNNs we recorded activity at the output to the ReLU of the five convolutional layers and the two fully connected layers. For each CNN layer, we characterised the representations through the representational dissimilarity matrix (RDM) (Diedrichsen & Kriegeskorte, 2017; Khaligh-Razavi & Kriegeskorte, 2014). For the brain, we used the human IT RDMs provided by Cichy et al. (2019). The correlation between each CNN layer's RDM and each human subject's RDM was calculated using the Mantel procedure with 10,000 permutations and the Kendall's Tau as the correlation statistic. The use of Kendall's Tau is recommended because, being a rank correlation coefficient, it does not assume the presence of a linear relationship between the RDM values (Khaligh-Razavi & Kriegeskorte, 2014), and because it proved to be more likely than Spearman's Rho to prefer the true model over a simplified categorical model (Nili et al., 2014). A repeated-measures ANOVA was then calculated with the Kendall's Tau values from every subject as the dependent variable and the network type and layer as within-subject factors. As post-hocs, Student's t-tests were used to calculate whether the corresponding layers of different CNNs correlated with IT to a different extent, and whether within each CNN the representation in the last layer better correlated to IT compared with the first layer. Moreover, the noise ceiling of the MRI data was calculated in order to evaluate the amount of variability that a model could possibly explain. The calculation of the noise ceiling shows how much of the data variability is explained by our model by taking into account the noise intrinsic in the MRI data. If the correlation between the model and the MRI data is close to the noise ceiling, the model explains the data well (Khaligh-Razavi & Kriegeskorte, 2014).

3 RESULTS

Between networks. Surprisingly, we found that representations within the *random* DeepCluster correlated with the brain as well as or better than the *trained* standard AlexNet [Fig.1e] (Network: $F(1,14) = 6.51$, $p < 0.05$; Layer: $F(6,84) = 10.45$, $p < 0.001$; Network x Layer: $F(6,84) = 5.87$, $p < 0.001$). Post-hoc tests showed that representations within layers 1, 2, 3, and 5 of the random DeepCluster performed significantly better than the corresponding layers in the trained standard AlexNet (for the five layers $t(28)$=3.43, 2.73, 2.57, 1.25, 2.38, 0.75, 0.73; p <0.01, 0.05, 0.05, ns, 0.05, ns, ns. The random DeepCluster was overall a better model of IT than the trained AlexNet. These results show that the architecture of a CNN can extract visual features that partially explain the variance of the brain. However, training could still improve the correlation between the CNN and the brain representations. We therefore tested how the three CNNs correlated with the brain before and after training [Fig.2]. Training significantly improved the correlation between standard AlexNet activations and IT representations (Network: $F(1,14) = 29.18$, $p < 0.001$; Layer: $F(6,84) = 7.25$, $p < 0.001$; Network x Layer: $F(6,84) = 14.53$, $p < 0.001$). Specifically a significant improve-

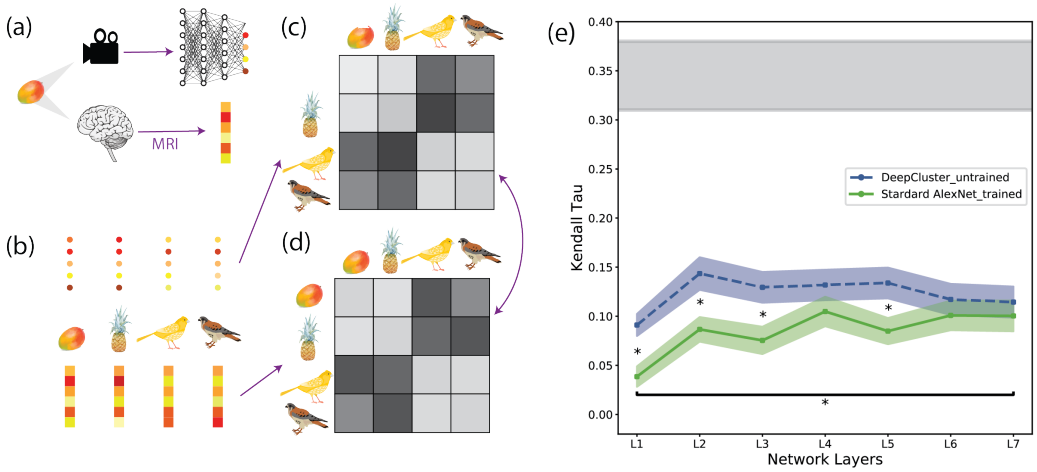

Figure 1: (a) The patterns of response evoked by an object in the neural network and the brain. (b) Some objects evoke more similar patterns than others. (c) The representation of the network can be characterised by its representational dissimilarity matrix (RDM) in which darker colours denote pairs of objects that evoked more dissimilar responses. (d) The representation of the brain can be characterised in the same way. The correlation between these two RDMs can then be calculated, to compare the similarity of the representations of the network and brain. (e) Results of the correlation between each subject's RDM and each networks' layer RDMs. The blue dashed line represents the correlations with the random DeepCluster. The green solid line represents the correlations with the trained standard Alexnet. The grey band represents the noise ceiling for the MRI data

ment happened in all the layers but the first one (Standard AlexNet: $t(28)$= 1.59, 3.47, 2.29, 3.76, 3.28, 4.62, 5.41; p <ns, 0.01, 0.05, 0.001, 0.01, 0.001, 0.001). In contrast, in the modified AlexNet training resulted in a significant worsening of the network correlation with the IT cortex, both for the unsupervised (Network: $F(1,14)$ = 9.17, p <0.001; Layer: $F(6,84)$ = 8.30, p <0.001 ; Network x Layer: $F(6,84)$ = 14.07, p <0.001) and supervised learning strategies (Network: $F(1,14)$ = 6.93, p <0.05; Layer: $F(6,84)$ = 10.88, p <0.001 ; Network x Layer: $F(6,84)$ = 23.77, p <0.001). Specifically, DeepCluster layers 1, 2, 3, and 4 correlated significantly less strongly with IT's representation after training (DeepCluster: $t(28)$= -2.36, -3.88, -3.13, -3.45, -2.02, -0.43, 0.42; p <0.05, 0.001, 0.01, 0.01, ns, ns, ns). In the modified AlexNet, this was true for layers 2, 3, 4, and 5 (DeepCluster AlexNet supervised: $t(28)$=-2.01, -4.84, -3.14, -2.30, -2.36, 0.45, 0.47; p <ns, 0.001, 0.01, 0.05, 0.05, ns, ns).

Correspondence of Hierarchy Between Brain and CNN Layers Within networks. A hierarchical correspondence between the network and the brain has been previously observed. We therefore expected that after training, the upper layer would show a higher correlation with IT than the first layer. This was indeed found for all the trained networks - Standard AlexNet (Trained: $t(28)$: 3.28, p <0.001; Random: $t(28)$: -0.78, ns), DeepCluster (Trained: $t(28)$: 3.49, p <0.01; Random: $t(28)$: 1.20, ns), and the modified AlexNet (Trained: $t(28)$: 3.61, p <0.01; Random: $t(28)$: 1.60, ns). However, while in the standard AlexNet the difference between the first and last layer was driven by an improvement in the last layer, in the modified AlexNet architecture, either unsupervised or supervised, the difference was driven by a worsening in the correlation between the first layer and IT representations.

4 Discussion

Here we showed that the representations elicited by a visual stimuli in an random network may correlate with the IT cortex activity patterns more strongly than in a trained network. Moreover, in some cases the training process can worsen, rather then improve, the correspondence between the network and the brain. These surprising results suggest that the architecture of some CNNs, rather than the weights learned during the training process, might most strongly explain their explanatory power for activity in IT. However, none of the architectures, either trained or random,

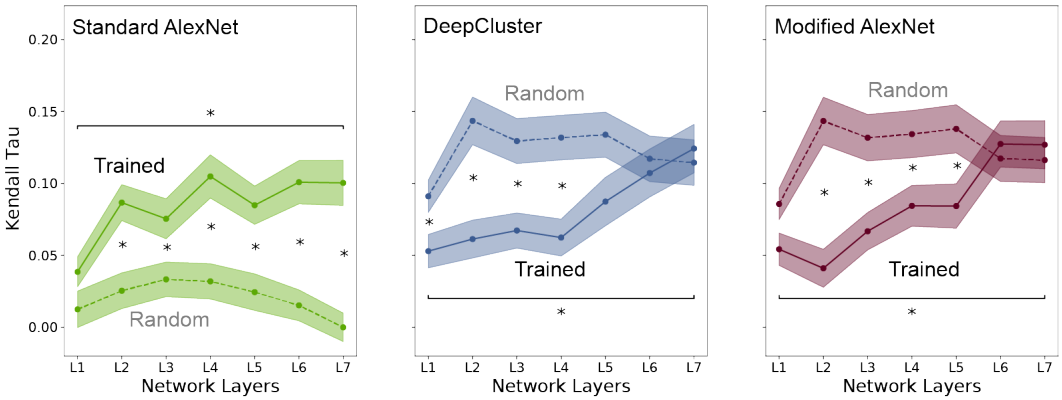

Figure 2: Results of the correlation between each subject's RDM and each network layer's RDMs. The dashed lines represent the correlations with the random networks. The solid lines represent the correlations with the trained networks. None of the models reached the noise ceiling (for reference see figure 1). (a) Correlations between the brain and standard AlexNet. (b) Correlations between the brain and DeepCluster. (c) Correlations between the brain and the modified AlexNet.

reached the noise ceiling. This means that IT is representing some feature of the images which is not currently captured by the considered models. Although the difference between the model and noise ceiling is different from the original ones in Khaligh-Razavi & Kriegeskorte (2014), a correction has been published that makes them similar to ours (Storrs et al., 2020). A possible limitation of our analysis is that we used only representational similarity analysis (RSA) to compare CNNs and the brain. Some results may not generalise to an alternative measures. However, a key previous study to find a correspondence between CNNs and the brain also used RSA (Khaligh-Razavi & Kriegeskorte, 2014). Moreover, a recent study using single-cell electrophysiology in the mouse and an encoding model, found that random weights performed similarly in predicting brain activity as trained weights (Cadena et al., 2019). This suggests that the unreasonable effectiveness of random weighs generalises to quite a different experiment. Another possible limitation is that the DeepCluster model included a "hand-crafted" Sobel edge-detection front end. However, previous evidence suggests that this alone cannot explain brain representations. Khaligh-Razavi & Kriegeskorte (2014) tested many hand-crafted visual-feature extraction methods, including an edge detector, and found they performed substantially less well than AlexNet. In light of our unexpected findings, we may need to reconsider two aspects of how CNNs predict the brain. First, we found that the architecture of CNNs contributes substantially to their brain-like representations. Therefore, a relatively quick architecture search may be as important as the lengthy and computationally expensive training process, in developing good models of the brain. Second, the features that CNNs learn during training can sometimes drive them away from the way the brain represents the world rather then bringing it closer. Finding ways to encourage CNNs to learn other features, such as global shape rather than texture (Geirhos et al., 2019) might make their representations more brain-like. In the future we will also obtain the brain-score of each network through the online platform (Schrimpf et al., 2018). Improving our understanding of the network training process is likely to impact both neuroscience and AI. In neuroscience we need testable computational models to investigate how the brain learns in both typical and atypical developmental contexts. On the other hand, making the networks' learning process more similar to how the human brain learns may improve performance of the network and effectiveness of the training, for example by shrinking the size of the necessary data sets and shortening training time.

ACKNOWLEDGMENTS

Funded by the European Research Council Advanced Grant FOUNDCOG 787981 We thank Radoslaw Martin Cichy for making the fMRI data available through the Algonauts project.

REFERENCES

Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *CoRR*, abs/1706.01350, 2017. URL http://arxiv.org/abs/1706.01350.

S. A. Cadena, F. H. Sinz, T. Muhammad, E. Froudarakis, E. Cobos, E. Y. Walker, J. Reimer, M. Bethge, A. Tolias, and A. S. Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? 2019. URL https://openreview.net/forum?id=rkxcXmtUUS.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. *arXiv:1807.05520 [cs]*, July 2018. URL http://arxiv.org/abs/1807.05520. arXiv: 1807.05520.

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*, 6:27755, June 2016. ISSN 2045-2322. doi: 10.1038/srep27755. URL https://www.nature.com/articles/srep27755.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*, 2019.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, 13(4):e1005508, 2017.

Adam Gaier and David Ha. Weight agnostic neural networks. *CoRR*, abs/1906.04358, 2019. URL http://arxiv.org/abs/1906.04358.

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. May 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

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.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pp. 2146–2153. IEEE, 2009.

S M Khaligh-Razavi and N Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 2014. ISSN 1553-734X. URL http://journals.plos.org/ploscompbiol/article/figures?id=10.1371/journal.pcbi.1003915.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future, 2020.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4), 2014.

Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11), 2009.

Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *ICML*, volume 2, pp. 6, 2011.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018.

Katherine R Storrs, Seyed-Mahdi Khaligh-Razavi, and Nikolaus Kriegeskorte. Noise ceiling on the crossvalidated performance of reweighted models of representational dissimilarity: Addendum to khaligh-razavi & kriegeskorte (2014). *bioRxiv*, 2020.

Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization. *Scientific Reports*, 8(1):1–17, February 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-22160-9. URL https://www.nature.com/articles/s41598-018-22160-9.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, March 2016. ISSN 1097-6256. doi: 10.1038/nn.4244. URL http://dx.doi.org/10.1038/nn.4244.