Deep neural networks trained with heavier data augmentation learn features closer to representations in hIT

Alex Hernández-García (ahernandez@uos.de)
Institute of Cognitive Science, University of Osnabrück
27 Wachsbleiche, 49090 Osnabrück, Germany

Johannes Mehrer (johannes.mehrer@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge
15 Chaucer Road, CB2 7EF Cambridge, UK

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Zuckerman Mind Brain Behavior Institute, Columbia University
3227 Broadway, L3-064 New York City, NY, USA

Peter König[*] (pkoenig@uos.de)
Institute of Cognitive Science, University of Osnabrück
27 Wachsbleiche, 49090 Osnabrück, Germany

Tim C. Kietzmann[*] (tim.kietzmann@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge
15 Chaucer Road, CB2 7EF Cambridge, UK

[*] shared senior authorship

Abstract

Modern artificial neural networks have been shown to learn representations comparable to the human visual cortex. However, the degree of representational similarity greatly differs between network architectures, training data sets and other factors. Understanding what makes a deep neural network learn representations closer to the human brain and subsequently developing models that reduce the gap helps computational neuroscientists investigate the underlying mechanisms that shape neural representations. Furthermore, understanding information processing in the brain paves the way for better artificial intelligence algorithms, as human vision is known to be highly robust. In this work, we investigate the relationship between augmentation of training data and the representational similarity of convolutional neural networks with high-level visual representations in human inferior temporal cortex. Our results suggest that networks trained with heavier augmentation yield representations that are more similar between deep neural networks and the brain.

Keywords: deep neural networks; visual cortex; representational similarity analysis; data augmentation

Introduction

One of the central goals of computational neuroscience is to develop better models of the human brain. The re-emergence of deep artificial neural networks, which now excel at many artificial intelligence tasks by automatically learning hierarchical representations (Girshick, Donahue, Darrell, & Malik, 2014), has also had a positive impact on computational neuroscience. For instance, the deep features learned by networks trained for image object classification have been found to correlate better with the representations in the human inferior temporal cortex (hIT) than traditional hand-crafted features or shallow models (Khaligh-Razavi & Kriegeskorte, 2014). Further, convolutional neural networks (CNN) are currently the most accurate models for multiple regions across the primate visual cortex (Kietzmann, McClure, & Kriegeskorte, 2017; Yamins & DiCarlo, 2016). However, while the similarity between artificial and human neural networks is promising, the crucial question remains: what makes CNNs learn representations that more closely mirror the ones in hIT?

Previous work has revealed that networks performing better in classification tasks correlate more strongly with neural representations in high level areas (Yamins et al., 2014). Moreover, the network architecture plays a crucial role (Storrs, Mehrer, Walther, & Kriegeskorte, 2017) and Mehrer, Kietzmann, and Kriegeskorte (2017) recently showed that training with more ecologically relevant input statistics yields more similar representations. Inspired by the apparent importance of the training data, we here explore the influence of data augmentation on the representational similarity of CNNs and hIT.

Data augmentation in machine learning refers to synthetically expanding a training set by applying transformations on existing examples such that they reflect plausible variations of the real objects and serve as additional training data. Although data augmentation has been used for a long time, systematic explorations of its benefits compared to other popular techniques have only recently gained traction (Hernández-

García & König, 2018; Perez & Wang, 2017). Further, additional data augmentation techniques have been recently proposed (Ratner, Ehrenberg, Hussain, Dunnmon, & Ré, 2017). The impact of data augmentation on the similarity with representations in hIT is, to our knowledge, still unknown.

Here, we train two CNN architectures for image object classification with two different data augmentation schemes: *light* transformations are limited to random crops, horizontal flips of the images and translations of 10 % of the image size; and *heavier* transformations perform additionally contrast and brightness adjustment as well as a larger range of affine transformations. Our analyses reveal that, regardless of the performance, the networks trained with heavier data augmentation learn features more similar to representations in the human inferior temporal cortex.

Methods

This section presents the experimental setup to analyze the role of data augmentation on the similarity between artificial neural networks and neural representations in hIT. We describe the network architectures, the augmentation schemes and the methodology employed to compare both systems.

Network architectures

To ensure the generality of the effects observed, we analyze two distinct, well-known CNNs, which reach high-performance on image object-classification: the all convolutional network, All-CNN (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014) and the wide residual network, WRN (Zagoruyko & Komodakis, 2016):

- **All-CNN** consists only of 12 convolutional layers, each followed by batch normalization and a ReLU activation. It has a total of 9.4 million parameters.

- **WRN** is a modification of ResNet (He, Zhang, Ren, & Sun, 2016) that achieves better performance with fewer layers, but more units per layer. We choose the WRN-28-10 version of the original paper, which has 28 layers and about 36.5 million parameters.

Following our previous work (Hernández-García & König, 2018), we remove weight decay and dropout, since, in the currently tested architectures, explicit regularization techniques do not contribute to better performance when enough data augmentation is applied. The rest of the hyperparameters are identical to the original papers.

Data augmentation

We define two data augmentation schemes, *light* and *heavier*. Both are applied to the highly benchmarked ImageNet ILSVRC 2012 data set (Russakovsky et al., 2015). The dataset contains almost 1.3 million high resolution images, which were resized into 150×200 pixels. The data augmentation schemes are the following:

- The **light** augmentation scheme is adopted from the literature, for instance (Springenberg et al., 2014). It performs only random horizontal flips and horizontal and vertical translations of maximum 10% of the image size, as well as random crops of 128×128 pixels.

- The **heavier** scheme performs a larger range of random affine transformations such as scaling, rotations and shear mapping, as well as contrast and brightness adjustment and random crops. Further details can be found in (Hernández-García & König, 2018).

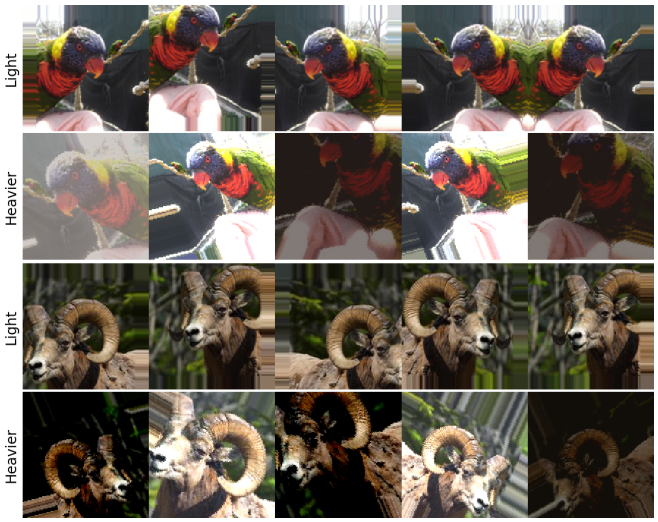

Figure 1: Illustration of the transformations performed by the light and heavier augmentation schemes on two example images. Note that the five transformations of each image have been produced by setting extreme values of the parameters, in order to highlight the characteristics of the schemes and the differences between them.

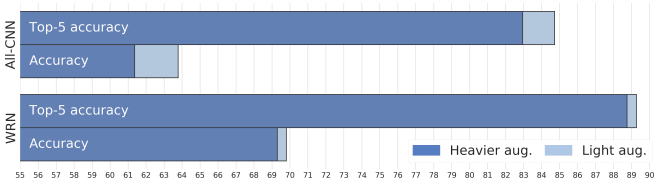

Figure 2: Test performance of All-CNN and WRN trained with light and heavier data augmentation.

The performance of All-CNN and WRN, trained with either light or heavier augmentation is shown in Figure 2. Note that training with light augmentation provides better results, specially on All-CNN. As pointed out in (Hernández-García & König, 2018), this is likely explained by the limited capacity of All-CNN, that prevents it from exploiting the heavy transformations of the already large ImageNet data set as improved classification accuracy. Besides, the heavier scheme was not

designed to optimize classification, but rather as an arbitrary larger set of plausible transformations. On smaller data sets, such as CIFAR, All-CNN trained with the same heavier augmentation scheme does outperform the model trained with light augmentation.

Representational similarity analysis

We make use of representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014) to characterize how similar the features learned by a CNN are to the representations in hIT. RSA has the benefit of allowing for direct comparisons across different model systems without having to explicitly align the different measurement types. This is accomplished by constructing representational dissimilarity matrices (RDMs). Across a large set of input images, RDMs characterize the internal representations of a given system by storing all pairwise distances. The resulting matrix therefore expresses the representational geometry in the high-dimensional activation space. By relying on distances, RDMs remain unchanged, if the space over which they are computed is rotated.

To characterize the representations in hIT, functional magnetic resonance imaging (fMRI) was used to measure BOLD responses while the participants (n=15) were presented with 92 images of isolated objects. The images originate from a wide variety of categories and levels of abstraction. On the broadest level, they can be separated into animate and inanimate. Inanimates can either be natural or artificial objects, whereas animates are divided into human stimuli (heads and body parts) and animals (full body and heads only). See (Kriegeskorte, Mur, Ruff, et al., 2008) for further details.

To compare DNNs and hIT representations, the network activation profiles for the 92 images were extracted. In particular, we compute the activations at the outputs of the 12 ReLU layers of All-CNN and at the outputs of the residual blocks of WRN. We then compute the RDM of both the responses in hIT, as well as at each layer of the CNNs using correlation distance. To obtain a more compact representation of the CNN models, we obtain a single RDM as a linear combination of the individual layer RDMs with respect to the hIT RDM using non-negative least squares and a cross-validation procedure, which avoids overfitting the image set.

Finally, we characterize the similarity between the CNNs and hIT by computing the Kendall's rank correlation coefficient τ_A between the RDM of the hIT representations and the RDM of the convolutional models. Standard errors were obtained from the similarity estimates to the 15 human subjects.

Results and discussion

The results presented in Figure 3 show that the correlation with the hIT representations is significantly higher for the models trained with heavier data augmentation. Notably, this is true despite its lower performance in terms of classification accuracy. In the case of the wide residual network (WRN) the difference between the two levels of augmentation is considerably larger, while in the All-CNN models, although statistically significant ($p < 0.05$), the difference is smaller. This may be explained by the worse classification accuracy of the model trained with heavier augmentation, as discussed above. The overall lower correlation of WRN replicates Storrs et al. (2017), who showed that residual networks exhibit a particularly low correlation with hIT compared to other architectures.

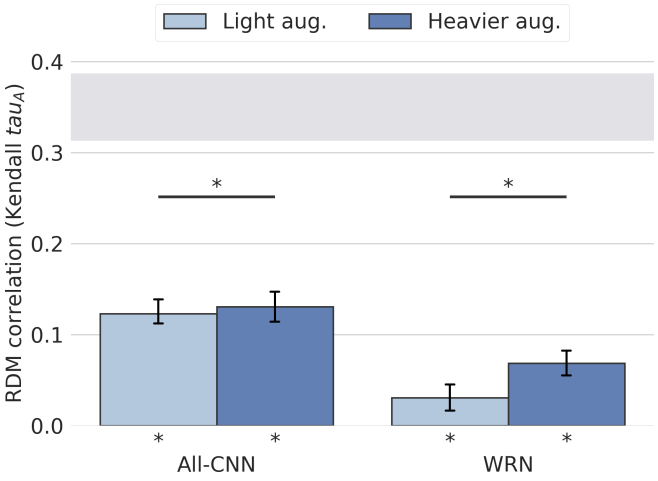

Figure 3: Comparison of the Kendall's τ_A coefficient of the hIT RDM and the RDM of the networks trained with light and heavier data augmentation. Both on All-CNN and WRN, the correlation of the model trained with heavier transformations is significantly higher than the light counterpart. The gray shaded area indicates the maximum possible correlation of a model given the noise in the measured data.

Given the exploratory nature of this project, it is not yet clear what exact mechanisms lead to the better match between representational geometries in higher level visual cortex and networks trained with heavier data augmentation. One possibility is that the larger variety during training may be more biologically plausible than training with constant images or very light transformations. Humans develop robust object representations based on highly variable input, while freely exploring the world. Sources of variation include different orientations, lighting conditions, backgrounds and occlusion. Eye-movements, including drifts and microsaccades, may further contribute to the variability in the sensory input to which the brain has to be invariant. Finally, developmental aspects of vision during early infancy lead to drastic changes in the input and may further facilitate robustness.

Our experiments address the question as to which factors drive computational models to learn representations closer to human brain. Given the superiority in visual robustness of the human brain, these insights may have implications for artificial vision systems based on deep neural networks, and for DNNs as a model system for visual processing in the brain. Finding that heavier training data transformations leads to more IT-like representations furthermore supports the notion that the input distribution can play a crucial role during the learning of representations in the brain (Mehrer et al., 2017).

Conclusion

We here explored how far light and heavier augmentation of the training set can affect the internal representations of deep neural networks and their alignment with human IT. To compare the neural and model system, we used representational similarity analysis, which allows for straight forward comparisons across different modalities (here fMRI BOLD and DNN activations). RSA revealed that the CNNs trained with heavier transformations learn representations more similar to those observed in higher visual cortex.

Future work should analyze a larger range of network architectures and data sets to gain better insights into the mechanisms driving the internal representations. It will be also interesting to study the different components of data augmentation in order to understand which particular transformations play a bigger role in better explaining hIT.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 641805; from the Cambridge Commonwealth, European and International Trust; and a DFG research fellowship to TCK.

References

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hernández-García, A., & König, P. (2018). Do deep nets really need weight decay and dropout? *arXiv preprint arXiv:1802.07042*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *bioRxiv*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*.

Mehrer, J., Kietzmann, T. C., & Kriegeskorte, N. (2017). Deep neural networks trained on ecologically relevant categories better explain human it. *Conference on Cognitive Computational Neuroscience (CCN)*.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Ratner, A. J., Ehrenberg, H. R., Hussain, Z., Dunnmon, J., & Ré, C. (2017). Learning to compose domain-specific transformations for data augmentation. *Advances in Neural Information Processing Systems (NIPS)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *International Conference on Learning Representations (ICLR)*.

Storrs, K., Mehrer, J., Walther, A., & Kriegeskorte, N. (2017). Architecture matters: How well neural networks explain it representation does not depend on depth and performance alone. *Conference on Cognitive Computational Neuroscience (CCN)*.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *Proceedings of the British Machine Vision Conference (BMVC)*.