Learning robust visual representations using data augmentation invariance

Alex Hernández-García
Institute of Cognitive Science
University of Osnabrück, Germany
ahernandez@uos.de

Peter König
Institute of Cognitive Science
University of Osnabrück, Germany
pkoenig@uos.de

Tim C. Kietzmann
MRC Cognition and Brain Sciences Unit
University of Cambridge, UK
tim.kietzmann@mrc-cbu.cam.ac.uk

Deep artificial neural networks (DNNs) trained for image object categorization have shown remarkable similarities with representations found across the primate ventral visual stream (Kietzmann et al., 2017; Kubilius et al., 2018). Yet, in spite of the consensus about the benefits of a closer integration of deep learning and neuroscience (Bengio et al., 2015; Marblestone et al., 2016), artificial and biological networks still exhibit important differences. Here we investigate one such property: increasing invariance to identity-preserving image transformations found along the ventral stream, proposed as a key mechanism for developing robust object recognition (DiCarlo & Cox, 2007; Tacchetti et al., 2018). Despite theoretical evidence that invariance should emerge naturally from the optimization process (Achille & Soatto, 2018), we present empirical evidence that the activations of prototypical artificial neural networks trained for object categorization are not robust to identity-preserving image transformations commonly used in data augmentation. As a solution, we propose *data augmentation invariance*, an unsupervised learning objective which improves the robustness of the learned representations by promoting the similarity between the activations of augmented image samples. Our results show that this approach is a simple, yet effective and efficient (10 % increase in training time) way of improving the invariance of the models while obtaining similar categorization performance.

To assess the invariance of the features learned by a DNN under the influence of identity-preserving image transformations we compare the activations of a given image with the activations of a data augmented version of the same image. As a metric of similarity we use the mean squared difference of the activations and define a *invariance score* as one minus the ratio between the similarity of augmented samples and the average similarity of the images in a data set. In order to promote the learning of robust features, we modify the objective function by adding a *data augmentation invariance loss*, which enforces the similarity between the activations of augmented samples within a batch, and is optimized jointly with the categorical cross entropy.

As a test bed for our hypotheses and proposal we use the all convolutional network, All-CNN (Springenberg et al., 2014), trained on the highly benchmarked data set for object recognition CIFAR-10 (Krizhevsky & Hinton, 2009). Figure 1 shows that the representations learned by the baseline model do not become more robust than at the pixel space after training. On the contrary, with our data augmentation invariance, the learned representations become increasingly more invariant to data augmentation transformations along the feature hierarchy. This comes at no cost in categorization performance and just a 10 % increase in training time. Further details of the methodology and extended results can be found in Hernández-García et al. (2019).

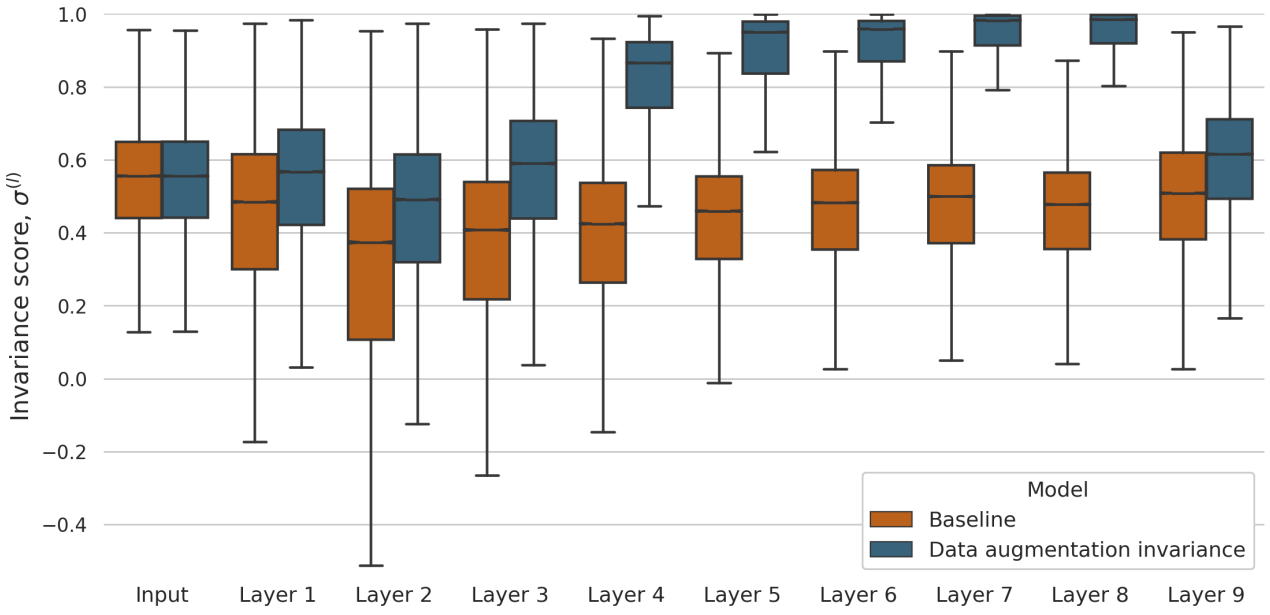

Figure 1: Distribution of the invariance score at each layer of the models.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research, JMLR*, 19(1):1947–1980, 2018.

Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*, 2015.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007.

Alex Hernández-García, Peter König, and Tim C Kietzmann. Learning robust visual representations using data augmentation invariance. *arXiv preprint arXiv:1906.04547*, 2019.

Tim Christian Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep neural networks in computational neuroscience. *bioRxiv:133504*, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv:408385*, 2018.

Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94, 2016.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations, ICLR, arXiv:1412.6806*, 2014.

Andrea Tacchetti, Leyla Isik, and Tomaso A Poggio. Invariant recognition shapes neural representations of visual input. *Annual review of vision science*, 4:403–422, 2018.