Learning robust visual representations using data augmentation invariance

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

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

Deep convolutional neural networks trained for image object categorization have shown remarkable similarities with representations found across the primate ventral visual stream. Yet, 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. Despite theoretical evidence that invariance should emerge naturally from the optimization process, we present empirical evidence that the activations of convolutional 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 increasing the invariance of the models while obtaining similar categorization performance.

Keywords: deep neural networks; visual cortex; invariance; data augmentation

Introduction

Deep artificial neural networks (DNNs) have borrowed much inspiration from neuroscience and are, at the same time, the current best model class for predicting neural responses across the visual system in the brain (Kietzmann, McClure, & Kriegeskorte, 2017; Kubilius et al., 2018). Yet, despite consensus about the benefits of a closer integration of deep learning and neuroscience (Bengio, Lee, Bornschein, Mesnard, & Lin, 2015; Marblestone, Wayne, & Kording, 2016), important differences remain.

Here, we investigate a representational property that is well established in the neuroscience literature on the primate visual system: the increasing robustness of neural responses to identity-preserving image transformations. While early areas of the ventral stream are strongly affected by variation in e.g. object size, position or illumination, later levels of processing are increasingly robust to such changes (Isik, Meyers, Leibo, & Poggio, 2013). The cascaded achievement of invariance to such identity-preserving transformations has been proposed as a key mechanisms for obtaining robust object recognition (DiCarlo & Cox, 2007; Tacchetti, Isik, & Poggio, 2018).

Learning such invariant representations has been a desired objective since the early days of artificial neural networks (Simard, Victorri, LeCun, & Denker, 1992). Accordingly, a myriad of techniques have been proposed to attempt to achieve tolerance to different types of transformations (see Cohen and Welling (2016) for a review). Interestingly, recent theoretical work has shown that invariance to "nuisance factors" should naturally emerge from the optimization process (Achille & Soatto, 2018).

Nevertheless, DNNs are still not robust to identity-preserving transformations, including simple image translations (Zhang, 2019), or more elaborate adversarial attacks (Szegedy et al., 2013), in which small changes, imperceptible to the human brain, can alter the classification output of the network. In this regard, there is growing evidence that DNNs may exploit highly discriminative features that do not match human perception (Ilyas et al., 2019). Extending this line of research, we use image perturbations using the data augmentation framework (Hernández-García & König, 2018) to show that DNNs, despite being trained on augmented data, are not sufficiently robust to such transformations.

Inspired by the increasing invariance observed along the primate ventral visual stream, we subsequently propose a simple, yet effective and efficient mechanism to improve the robustness of the representations: we include an additional term in the objective function that encourages the similarity between augmented examples within each batch.

Methods

This section presents the procedure to empirically measure the invariance of the representations of a convolutional neural network and our proposal to improve the invariance.

Model, data and training parameters

As a test bed for our hypotheses and proposal we use the all convolutional network, All-CNN (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014), a well-known architecture which achieves good performance in spite of being much shallower than other architectures, and thus faster to train and more convenient for the analysis. It consists of 9 convolutional layers, with a total of 1.3 million parameters. Our model is identical to All-CNN-C in the original paper, except that we remove the explicit regularizers—weight decay and dropout—following the conclusions from Hernández-García and König (2018). We also keep the original training hyperparameters: 350 epochs, initial learning rate of 0.01 and batch size of 128.

We train on the highly benchmarked data set for object recognition CIFAR-10 (Krizhevsky & Hinton, 2009) and apply heavier data augmentation than in the original paper. Specifically, we use the *heavier* training and evaluation scheme described by Hernández-García and König (2018), which includes random affine transformations and contrast and brightness adjustment.

Evaluation of invariance

To measure the invariance of the learned features 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.

Consider the activations of an input image x at layer l of a neural network, which can be described by a function $f^{(l)}(x) \in \mathbb{R}^{D^{(l)}}$. We can define the distance between the activations of two input images x_i and x_j by their mean square difference:

$$d^{(l)}(x_i, x_j) = \frac{1}{D^{(l)}} \sum_{k=1}^{D^{(l)}} (f_k^{(l)}(x_i) - f_k^{(l)}(x_j))^2 \tag{1}$$

Following this, we compute the mean squared difference between every $f^{(l)}(x_i)$ and a random transformation of x_i, that is $d^{(l)}(x_i, G(x_i))$. In this case, we define $G(x)$ as the data augmentation scheme that can take any of the extreme values of each transformation in the *heavier* scheme, after halving the parameter ranges. This is to ensure the same level of augmentation in all comparisons, while preventing too extreme transformations.

The assessment of the similarity between the activations of an image x_i and of its augmented versions $G(x_i)$ was normalised by the similarity with the other, different images, reminiscent of an image identification problem. We define the invariance score $\sigma_i^{(l)}$ of the transformation $G(x_i)$ at layer l of a model, with respect to a data set of size N, as follows::

$$\sigma_i^{(l)} = 1 - \frac{d^{(l)}(x_i, G(x_i))}{\frac{1}{N}\sum_{j=1}^{N} d^{(l)}(x_i, x_j)} \tag{2}$$

The invariance $\sigma_i^{(l)}$ takes the maximum value of 1 if the activations of x_i and its transformed version $G(x_i)$ are identical. To assess the overall invariance of a model post training, we calculate $\sigma_i^{(l)}$ for the 10,000 test images of CIFAR-10, with respect to five different random transformations. In Figure 1 we show the distribution of $\sigma_i^{(l)}$ at each layer of All-CNN-C.

Data augmentation invariance

Most CNNs trained for object categorization are optimized through mini-batch gradient descent (SGD), that is the weights are updated iteratively by computing the loss of a batch $\mathcal{B}$ of examples, instead of the whole data set at once. The models are typically trained for a number of *epochs*, E, which is a whole pass through the entire training data set of size N. That is, the weights are updated $K = \frac{N}{|\mathcal{B}|}$ times each epoch.

Data augmentation introduces variability into the process by performing a different, stochastic transformation of the data every time an example is fed into the network. However, with standard data augmentation, the model has no information about the *identity* of the images, that is, that different augmented examples, seen at different epochs, separated by $\frac{N}{|\mathcal{B}|}$ iterations on average, correspond to the same seed data point. We believe this information may be valuable and useful to learn better representations in a self-supervised manner. For example, the high temporal correlation of the stimuli that reach the visual cortex may play a crucial role in the creation of robust connections (Wyss, König, & Verschure, 2006).

In order to make use of this information in an unsupervised way, we propose to perform data augmentation within the batches by constructing the batches to include M transformations of each example (see Hoffer et al. (2019) for a similar idea). Additionally, we propose to modify the loss function to include an additional term that accounts for the invariance of the feature maps across multiple image samples. Considering the difference between the activations at layer l of two images, $d^{(l)}(x_i, x_j)$, defined in Equation 1, we define the data augmentation invariance loss at layer l for a given batch $\mathcal{B}$ as follows:

$$\mathcal{L}_{inv}^{(l)} = \frac{\sum_k \frac{1}{|\mathcal{S}_k|^2} \sum_{x_i, x_j \in \mathcal{S}_k} d^{(l)}(x_i, x_j)}{\frac{1}{|\mathcal{B}|^2} \sum_{x_i, x_j \in \mathcal{B}} d^{(l)}(x_i, x_j)} \tag{3}$$

where $\mathcal{S}_k$ is the set of samples in the batch $\mathcal{B}$ that are augmented versions of the same seed sample x_k. This loss term intuitively represents the average difference of the activations between the sample pairs that correspond to the same source image, relative to the average difference of all pairs. A convenient property of this definition is that $\mathcal{L}_{inv}$ does not depend on the batch size nor the number of in-batch augmentations $M = |\mathcal{S}_k|$. Furthermore, it can be efficiently implemented using matrix operations.

Since we want to achieve image invariance at L layers of the network and jointly train for object recognition, we define the total loss as follows:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{obj} + \sum_{l=1}^{L} \alpha^{(l)} \mathcal{L}_{inv}^{(l)} \tag{4}$$

where $\sum_{l=1}^{L} \alpha^{(l)} = \alpha$ and $\mathcal{L}_{obj}$ is the loss associated with the object recognition objective, typically the cross-entropy be-

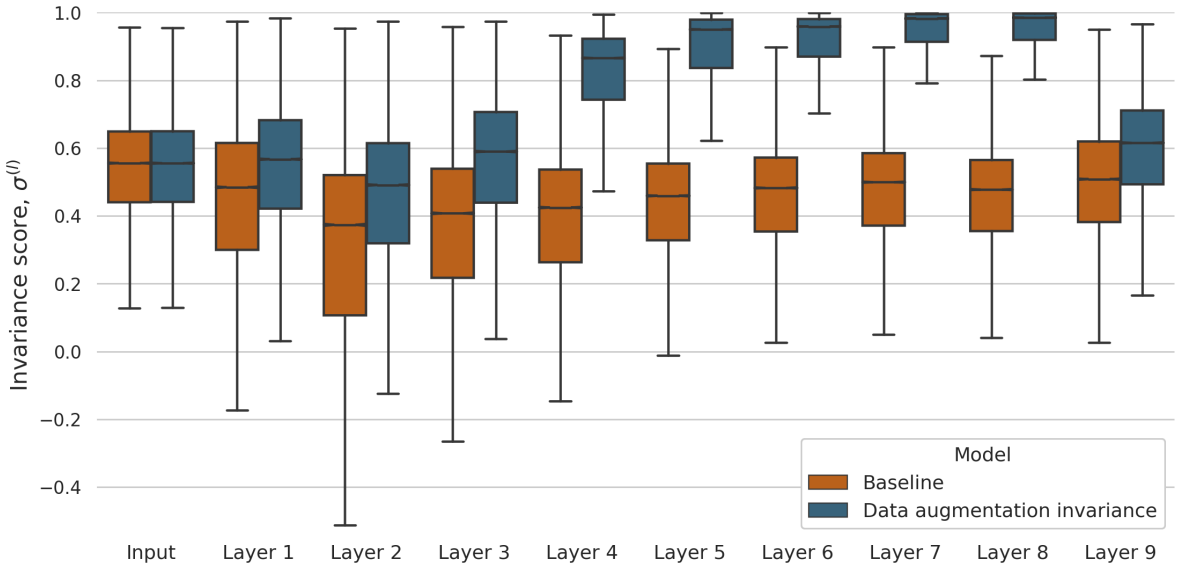

Figure 1: Distributon of tnvariance score at each layer of the baseline model and the model trained data augmentation invariance.

tween the object labels and the output of a softmax layer. All the results we report in this paper have been obtained by setting $\alpha = 0.1$ and distributing the coefficients across the layers according to an exponential law, such that $\alpha^{(l=L)} = 10\alpha^{(l=1)}$. This aims at simulating a probable response along the ventral visual stream, where higher regions are more invariant than the early visual cortex[1].

Results

One of the contributions of this paper is to empirically test in how far convolutional neural networks produce invariant representations under the influence of identity-preserving transformations of the input images. Figure 1 shows the invariance scores, as defined in Equation 2, across network layers.

Despite the presence of data augmentation during training, which implies that the network may learn augmentation-invariant transformations, the representations of the baseline model (red boxes) do not increase in invariance beyond the pixel space. As a solution, we have proposed a simple, unsupervised modification of the loss function to encourage the learning of data augmentation-invariant features. As can be seen in Figure 1 (blue boxes), our data augmentation mechanism pushed network representations to become increasingly more robust with network depth. One exception is the top, 'readout' layer, likely because the features are dominated by the categorization objective.

In order to better understand the effect of the data augmentation invariance, we plotted the learning dynamics of the invariance loss at each layer. In Figure 2, we can see that in the baseline model, the invariance loss keeps increasing over the course of training. In contrast, when the loss is added to the optimization objective, the loss drops for all but the last layer. Unexpectedly, the invariance loss increased during the first epochs and only then started to decrease. While further investigations are required, these two phases may correspond to the compression and diffusion phases proposed by Shwartz-Ziv and Tishby (2017).

In terms of efficiency, adding terms to the objective function implies an overhead of the computations. However, since the pairwise distances can be efficiently computed at each batch through matrix operations, the training time is only increased by about 10 %. Finally, the improved invariance comes at no cost in the categorization performance, as the network trained with data augmentation invariance achieves similar classification performance to the baseline model—test accuracy baseline: 91.5 %; test accuracy data augmentation invariance: 92.2 %).

Conclusions

In this work we have empirically shown that the features learned by a prototypical convolutional neural networks are not invariant to identity-preserving image transformations despite being part of the training procedure. This property is fundamentally different to the primate ventral visual stream, where neural populations have been found to be increasingly robust to changes in view or lighting conditions of the same object (DiCarlo & Cox, 2007).

Taking inspiration from this property of the visual cortex, we have proposed an unsupervised objective to encourage

[1] It is beyond the scope of this paper to analyze the sensitivity of the hyperparameters $\alpha^{(l)}$, but we have not observed a significant impact in the classification performance by using other distributions.

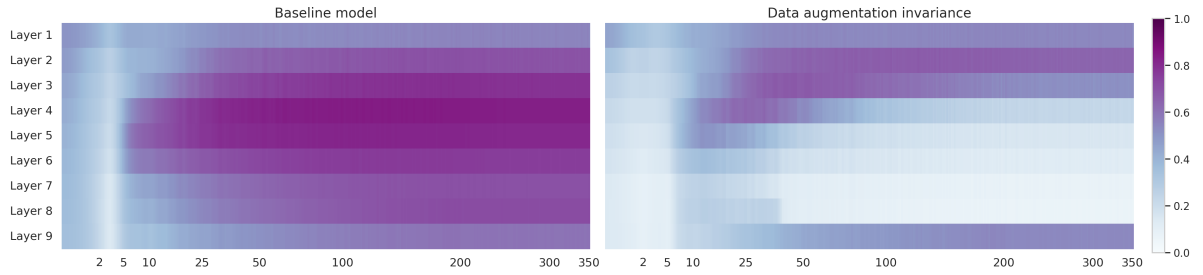

Figure 2: Dynamics of the data augmentation invariance loss $\mathcal{L}_{inv}^{(l)}$ during training. The axis of abscissas (epochs) is scaled quadratically to better appreciate the dynamics at the first epochs. The same random initialization was used for both models.

learning more robust features, using data augmentation as the framework to perform identity-preserving transformations on the input data. We created mini-batches with M augmented versions of each image and modified the loss function to maximize the similarity between the activations of the same seed images.

Data augmentation invariance effectively produces more robust representations, unlike standard models optimized only for object categorization, at no cost in classification performance. Future work will investigate whether this increased robustness also allows for better modelling of neural data.

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 the DFG.

References

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

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

Cohen, T., & Welling, M. (2016). Group equivariant convolutional networks. In *International Conference on Machine Learning, ICML* (pp. 2990–2999).

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

Hernández-García, A., & König, P. (2018). Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*.

Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., & Soudry, D. (2019). Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2013). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *bioRxiv:133504*.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Technical report, University of Toronto*.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv:408385*.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94.

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Simard, P., Victorri, B., LeCun, Y., & Denker, J. (1992). Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems, NIPS* (pp. 895–903).

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations, ICLR, arXiv:1412.6806*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tacchetti, A., Isik, L., & Poggio, T. A. (2018). Invariant recognition shapes neural representations of visual input. *Annual review of vision science*, *4*, 403–422.

Wyss, R., König, P., & Verschure, P. F. J. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS biology*, *4*(5), e120.

Zhang, R. (2019). Making convolutional networks shift-invariant again. In *International Conference on Machine Learning, ICML*.