LEARNING ROBUST VISUAL REPRESENTATIONS USING DATA AUGMENTATION INVARIANCE

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

Peter König
Institute of Cognitive Science
University of Osnabrück, Germany
Dept. of Neurophysiology and Pathophysiology
University Medical Center Hamburg-Eppendorf
pkoenig@uos.de

Tim Kietzmann
Institute for Brain, Cognition and Behaviour
Radboud University, Netherlands
t.kietzmann@donders.ru.nl

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 sufficiently 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

1 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 et al., 2017; Kubilius et al., 2018). Yet, despite consensus about the benefits of a closer integration of deep learning and neuroscience (Bengio et al., 2015; Marblestone et al., 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 (V1-V2) are strongly affected by variation in e.g. object size, position or illumination, later levels of processing are increasingly robust to such changes, as measured first in single neurons of the inferior temporal (IT) cortex of macaques Booth & Rolls (1998) and then in humans' (Quiroga et al., 2005; Isik et al., 2013). The cascaded achievement of invariance to such identity-preserving transformations has been proposed as a key mechanism for robust object recognition (DiCarlo & Cox, 2007; Tacchetti et al., 2018).

Learning such invariant representations has been an objective since the early days of artificial neural networks (Simard et al., 1992). Accordingly, a myriad of techniques have been proposed to attempt to achieve tolerance to different types of transformations (see Cohen & Welling (2016) for a review).

Interestingly, recent theoretical work (Achille & Soatto, 2018) has shown that invariance to "nuisance factors" should naturally emerge from the optimization of deep models.

Nevertheless, DNNs are still not robust to identity-preserving transformations, including simple image translations (Zhang, 2019). A remarkable extreme example are adversarial attacks (Szegedy et al., 2013), in which small changes, imperceptible to the human brain, can alter the model predictions. Extending this line of research, we use the data augmentation framework (Hernández-García & König, 2018) to show that the representations learned by DNNs, despite being trained on augmented data, are not more robust than in the input space.

This is likely related to the growing evidence that DNNs may exploit highly discriminative features that do not match human perception (Jo & Bengio, 2017; Ilyas et al., 2019; Wang et al., 2019). Here, we postulate that this is due to the combination of their large capacity and the highly unconstrained learning setup of typical supervised deep models. We argue that incorporating elements from human visual perception and biological constraints can add a positive inductive bias that may yield better, more robust representations.

In particular, inspired by the increasing invariance observed along the primate ventral visual stream, we subsequently propose *data augmentation invariance*, a simple, yet effective and efficient mechanism to improve the robustness of the representations: we include an additional, unsupervised term in the objective function that encourages the similarity between augmented examples within each batch, while achieving the same classification performance or higher.

2 METHODS

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

2.1 EVALUATION OF INVARIANCE

To measure the invariance, 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 square difference between $f^{(l)}(x_i)$ and a random transformation of x_i, that is $d^{(l)}(x_i, G(x_i))$. $G(x)$ refers to the stochastic function that transforms the input images according to a pre-defined data augmentation scheme. We then normalize the distance by the average distance in the (test) set. We define the invariance score $S_i^{(l)}$ of the transformation $G(x_i)$ at layer l of a model, with respect to a data set of size N, as follows:

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

2.2 DATA AUGMENTATION INVARIANCE

Most DNNs 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. 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 receives 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 conjecture that this information may help learn better representations, inspired by biological vision, where the high temporal correlation of the stimuli that reach the visual cortex may play a crucial role in the creation of robust connections (Wyss et al., 2006).

In order to make use of this information and improve the robustness, we first propose to perform *in-batch* data augmentation by constructing the batches with M transformations of each example (see Hoffer et al. (2019) for a similar idea). Second, we 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 $\mathcal{L}_{inv}$ depends neither 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 both, certain representational invariance at L layers of the network and high object recognition performance at the network output are desired, 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 between the true and predicted labels. All the results we report in this paper were 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 to simulate a probable response along the ventral visual stream, where higher regions are more invariant than the early visual cortex[1].

2.3 Architectures and data sets

As test bed for our hypotheses and proposal we trained the all convolutional network, All-CNN-C (Springenberg et al., 2014), on the object recognition data sets CIFAR-10 (Krizhevsky & Hinton, 2009), with 50,000 32x32 px training images and 10 classes, and *tiny* ImageNet, a subset of ImageNet (Russakovsky et al., 2015) with 100,000 64x64 px training images and 200 classes. The invariance loss defined in Equation 3 was computed after the ReLU activation of each convolutional layer.

All models are trained using a data augmentation scheme that consists in affine transformations, contrast adjustment and brightness adjustment (see the details in Appendix A). Importantly, we adopt the conservative approach of keeping the training hyperparameters (learning rate, number of epochs, etc.) as in the original paper, except that, following the recommendation by Hernández-García & König (2018) we do not use explicit regularization (weight decay and dropout) since comparable performance is obtained without them if data augmentation is used. In Appendices B and C we include results on two additional architectures and further analyses.

3 Results

One of the contributions of this paper is to empirically test in how far DNNs produce invariant representations under the influence of identity-preserving transformations of the input images. Figure 1 shows the distribution of invariance scores of the test images, as defined in Equation 2 with 5 transformations $G(x_i)$ for each image i, across the network layers.

[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.

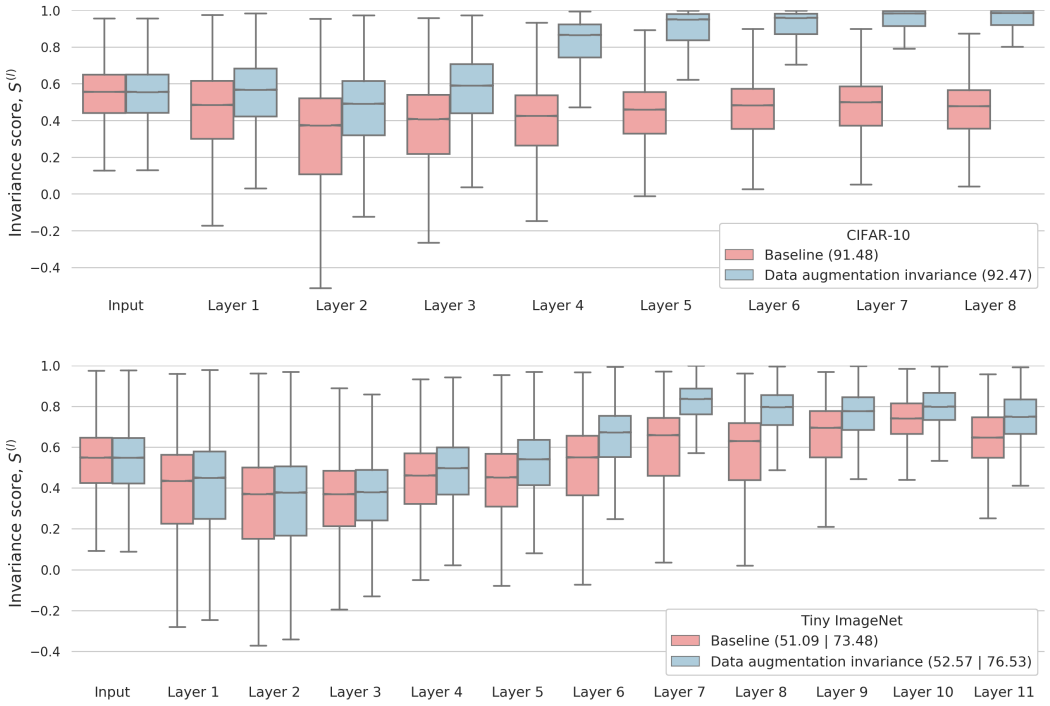

Figure 1: Distribution of the invariance score of the CIFAR-10 (top) and Tiny ImageNet (bottom) test images at each layer of the baseline All-CNN and the model trained data augmentation invariance (higher is better). The values in the legend indicate the test accuracy (and top5 accuracy on ImageNet)

In spite of training with data augmentation, which implies that the networks *see* and may learn augmentation-invariant transformations, the representations of the baseline models (red boxes) do not increase substantially beyond the invariance observed in 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 the plots (blue boxes), our data augmentation mechanism pushed network representations to become increasingly more robust with network depth.

Importantly, the improved robustness comes at no cost in the categorization performance, as the models trained with data augmentation invariance achieved similar or even higher accuracy than the baseline model, as reported in the legend of Figure 1. In terms of efficiency, adding terms to the objective function implies an overhead of computations. However, since the pairwise distances can be efficiently computed through matrix operations, the training time was only increased by about 10 %.

4 Conclusions

In this work, we have first proposed an invariance score that assesses the robustness of the features learned by a neural network towards identity-preserving transformations (see Equation 2). Using this score, we have shown that the invariance of the features learned by a prototypical DNN hardly increased with respect to the original pixel space. 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 learning objective to encourage learning more robust features, using data augmentation as the framework to transform the input data. Data augmentation invariance effectively produced more robust representations at no cost in classification performance and with an affordable, slight increase (10 %) in training time. We hope this work contributes to the growing body of evidence indicating that inspiration from biological vision can provide useful inductive biases for deep learning.

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.

MC Booth and Edmund T Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex (New York, NY: 1991)*, 8(6):510–523, 1998.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning, ICML*, pp. 2990–2999, 2016.

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

Alex Hernández-García and Peter König. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. IEEE, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Leyla Isik, Ethan M Meyers, Joel Z Leibo, and Tomaso Poggio. The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, 111(1):91–102, 2013.

Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

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.

Rodrigo Q Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102, 2005.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, IJCV*, 2015.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems, NIPS*, pp. 895–903, 1992.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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.

Haohan Wang, Xindi Wu, Pengcheng Yin, and Eric P Xing. High frequency component helps explain the generalization of convolutional neural networks. *arXiv preprint arXiv:1905.13545*, 2019.

Reto Wyss, Peter König, and Paul FM J Verschure. A model of the ventral visual system based on temporal stability and local memory. *PLoS biology*, 4(5):e120, 2006.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference, BMVC*, 2016.

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

A DATA AUGMENTATION SCHEME

As specified in Section 2.3, the data augmentation scheme used to trained all the models consists of affine transformations, contrast adjustment and brightness adjustment. Specifically, we apply the following image transformations, with the parameters defined in Table 1:

- Affine transformations:
$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} f_h z_x \cos(\theta) & -z_y \sin(\theta + \phi) & t_x \\ z_x \sin(\theta) & z_y \cos(\theta + \phi) & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$
- Contrast adjustment: $x' = \gamma(x - \overline{x}) + \overline{x}$
- Brightness adjustment: $x' = x + \delta$

For the computation of the invariance score we use exactly the same transformations, but instead of randomly sampling for the parameter ranges defined in Table 1, we halve the range and sample from one of the extreme values.

Table 1: Description and range of possible values of the parameters used for the data augmentation scheme. $B(p)$ denotes a Bernoulli distribution and $\mathcal{N}(a, b)$ a truncated normal distribution centered at $\frac{a+b}{2}$ and with standard deviation $\frac{b-a}{4}$.

Parameter	**Description**	**Range**
f_h	Horiz. flip	$1 - 2B(0.5)$
t_x	Horiz. translation	$\mathcal{N}(-0.1, 0.1)$
t_y	Vert. translation	$\mathcal{N}(-0.1, 0.1)$
z_x	Horiz. scale	$\mathcal{N}(0.85, 1.15)$
z_y	Vert. scale	$\mathcal{N}(0.85, 1.15)$
θ	Rotation angle	$\mathcal{N}(-22.5°, 22.5°)$
ϕ	Shear angle	$\mathcal{N}(-0.15, 0.15)$
γ	Contrast	$\mathcal{N}(0.5, 1.5)$
δ	Brightness	$\mathcal{N}(-0.25, 0.25)$

B ADDITIONAL EXPERIMENTS

In this appendix we present similar results to those shown in Section 3, with additional architectures. Figure 2 shows the distribution of the invariance of a baseline wide residual network, WRN-28-10 (Zagoruyko & Komodakis, 2016) as well as the model trained with our proposed data augmentation invariance. Figure 3 shows the equivalent results of training DenseNet-BC (Huang et al., 2017).

The results on the two additional architectures are consistent with the results presented in the main body of the paper, on All-CNN. Therefore, they reinforce our conclusion that models trained with standard data augmentation do not learn features invariant to the transformations used during training. However, by optimizing a data augmentation invariant learning objective, the models effectively learn more invariant features, while preserving the classification performance.

C LEARNING DYNAMICS

In order to better understand the effect of the data augmentation invariance, we analysed the learning dynamics of the invariance loss at each layer of All-CNN trained on CIFAR-10. In Figure 4, 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. Perhaps 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 & Tishby (2017).

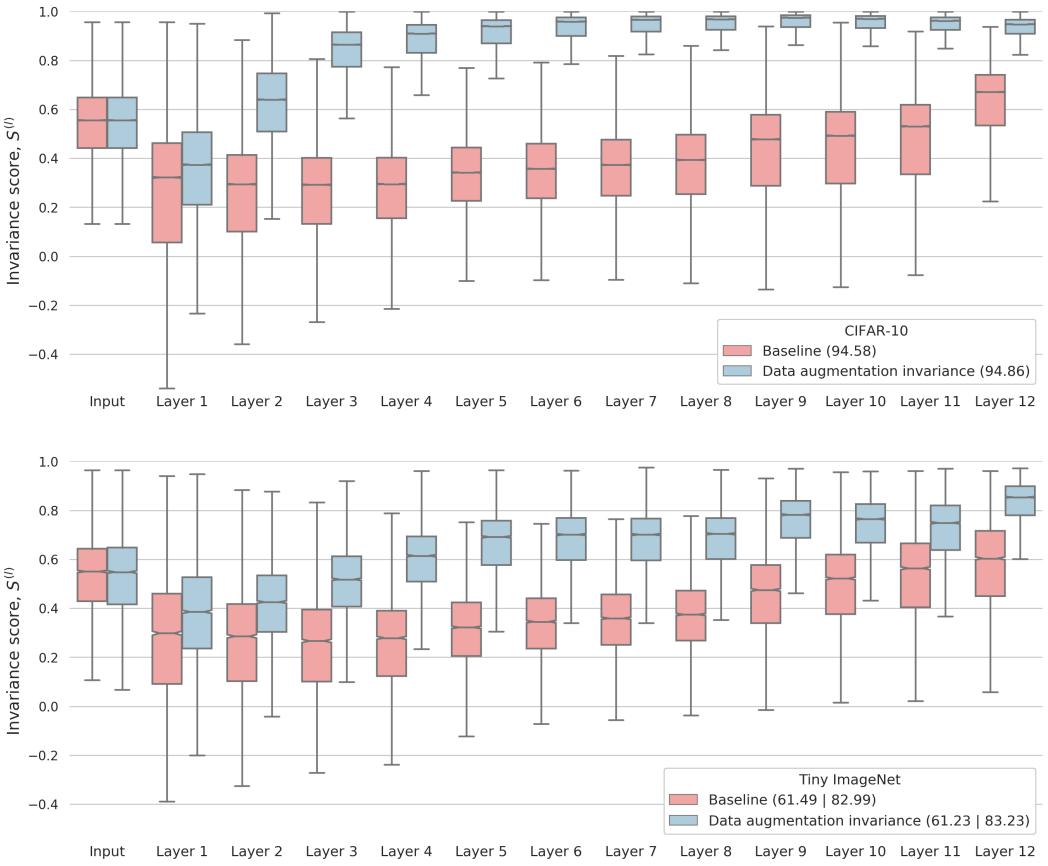

Figure 2: Distribution of the invariance score of the CIFAR-10 (top) and Tiny ImageNet (bottom) test images at each layer of the baseline WRN and the model trained data augmentation invariance (higher is better). The values in the legend indicate the test accuracy (and top5 accuracy on ImageNet)

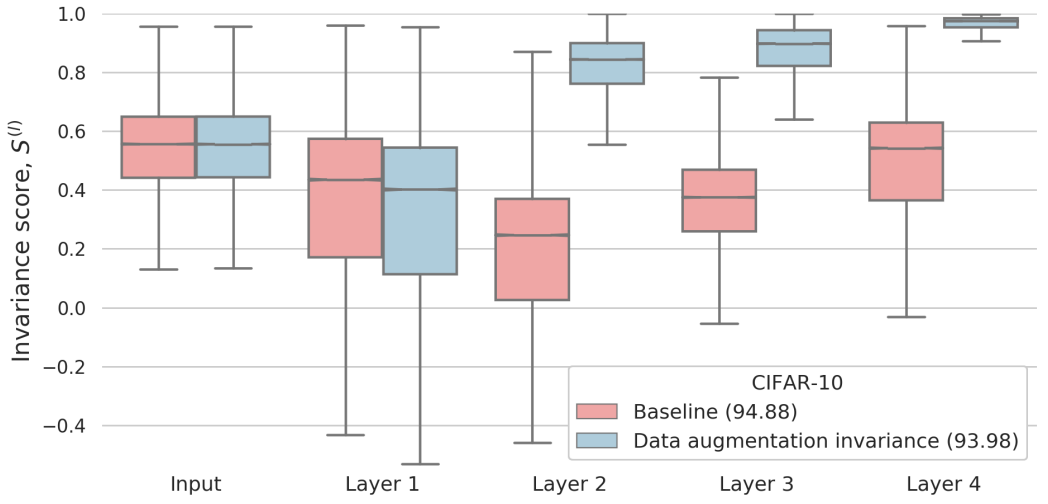

Figure 3: Distribution of the invariance score of the CIFAR-10 test images at each layer of the baseline DenseNet and the model trained data augmentation invariance (higher is better). The values in the legend indicate the test accuracy

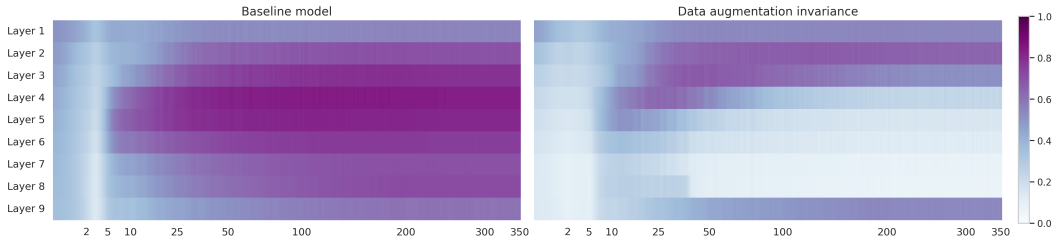

Figure 4: Dynamics of the data augmentation invariance loss $\mathcal{L}_{inv}^{(l)}$ during training (lower is better). The axis of abscissas (epochs) is scaled quadratically to better appreciate the dynamics at the first epochs. The same random initialization was used for both models.