Learning robust visual representations using *data augmentation invariance*

Alex Hernandez-Garcia[1], Peter König[1,2], Tim C. Kietzmann[3,4]

[1]Institute of Cognitive Science, University of Osnabrück, Germany • [2]Institute of Neurophysiology und Pathophysiology, University Medical Center Hamburg-Eppendorf, Germany • [3]MRC Cognition and Brain Sciences Unit, University of Cambridge, United Kingdom • [4]Donders Institute for Brain Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

QUESTIONS

Invariance to identity-preserving transformations is a key mechanism in the primate visual ventral stream [1, 2].

Are the representations of DNNs also invariant?

Can data augmentation help improve the robustness?

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

ahernandez@uos.de • @alexhdezgcia

alexhernandezgarcia.github.io

ABSTRACT

Deep convolutional neural networks trained for image object categorization have shown remarkable similarities with representations found across the primate ventral visual stream [3]. 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 [1, 2]. Despite theoretical evidence that invariance should emerge naturally from the optimization process [4], 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 [5]. 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.

METHODS

Evaluation of invariance

We want to assess the invariance of DNN's features under the influence of **identity-preserving transformations**.

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

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

Data augmentation invariance score

Learning invariance

- Data augmentation within the batches
- Layer-wise terms in the loss function:

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

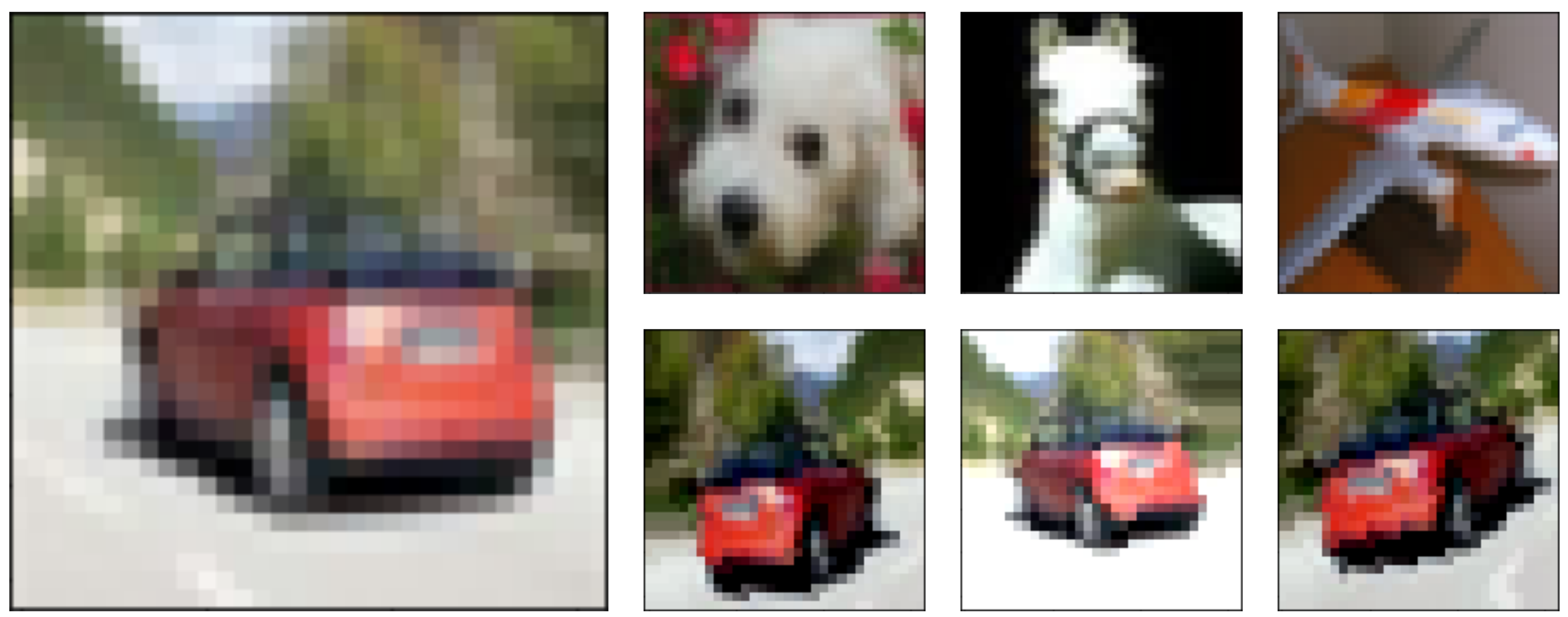
$$\mathcal{L} = (1-\alpha)\mathcal{L}_{obj} + \sum_{l=1}^{L} \alpha^{(l)} \mathcal{L}_{inv}^{(l)}$$

- Higher layers are set to be exponentially more invariant than early layers.

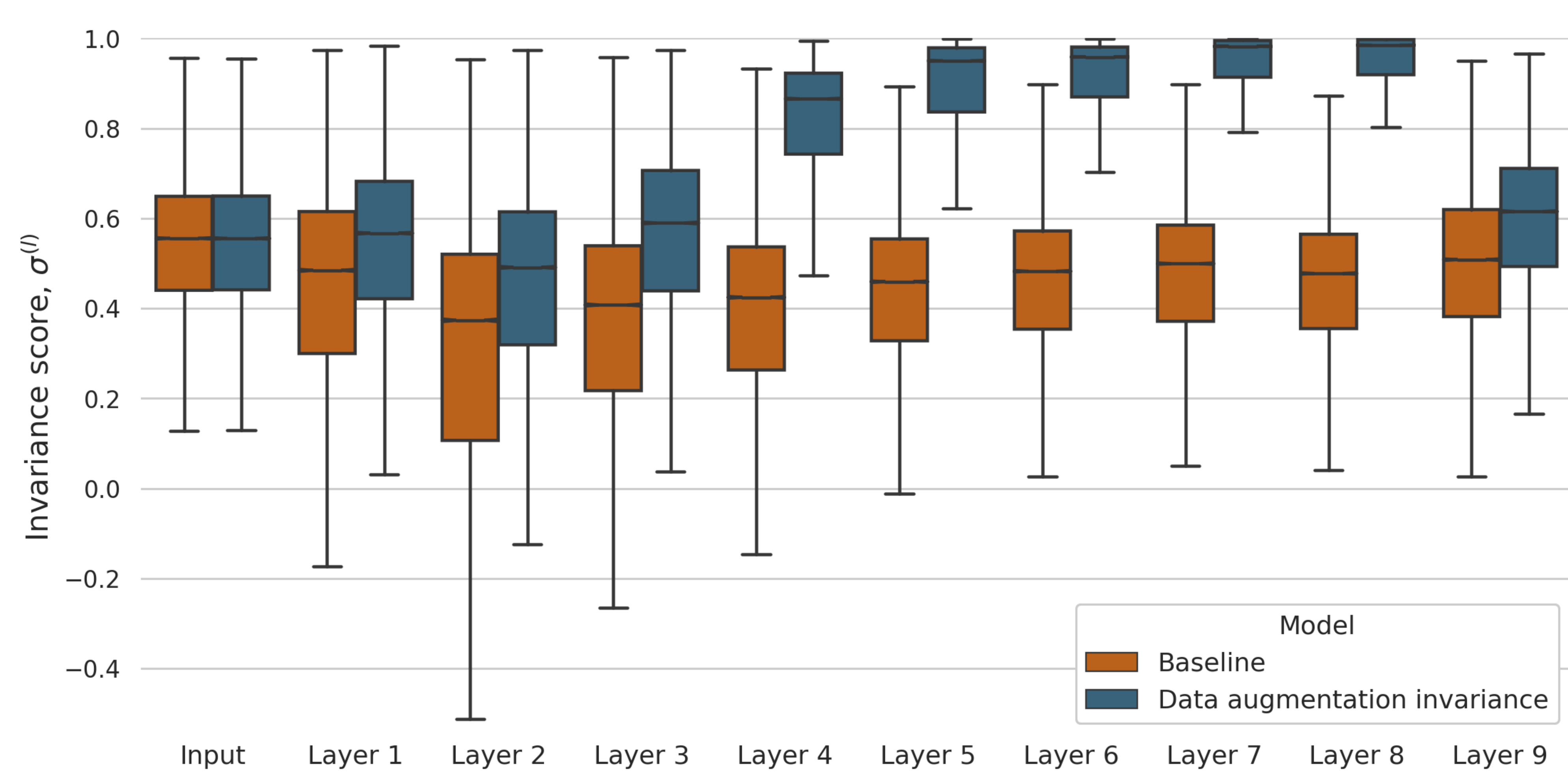
Experiments

- Network architecture: All-CNN
- Data set: CIFAR-10
- Hyperparameters as in original model

The high level representations of the six images on the right learned by a standard DNN are equally similar to the reference image (left).

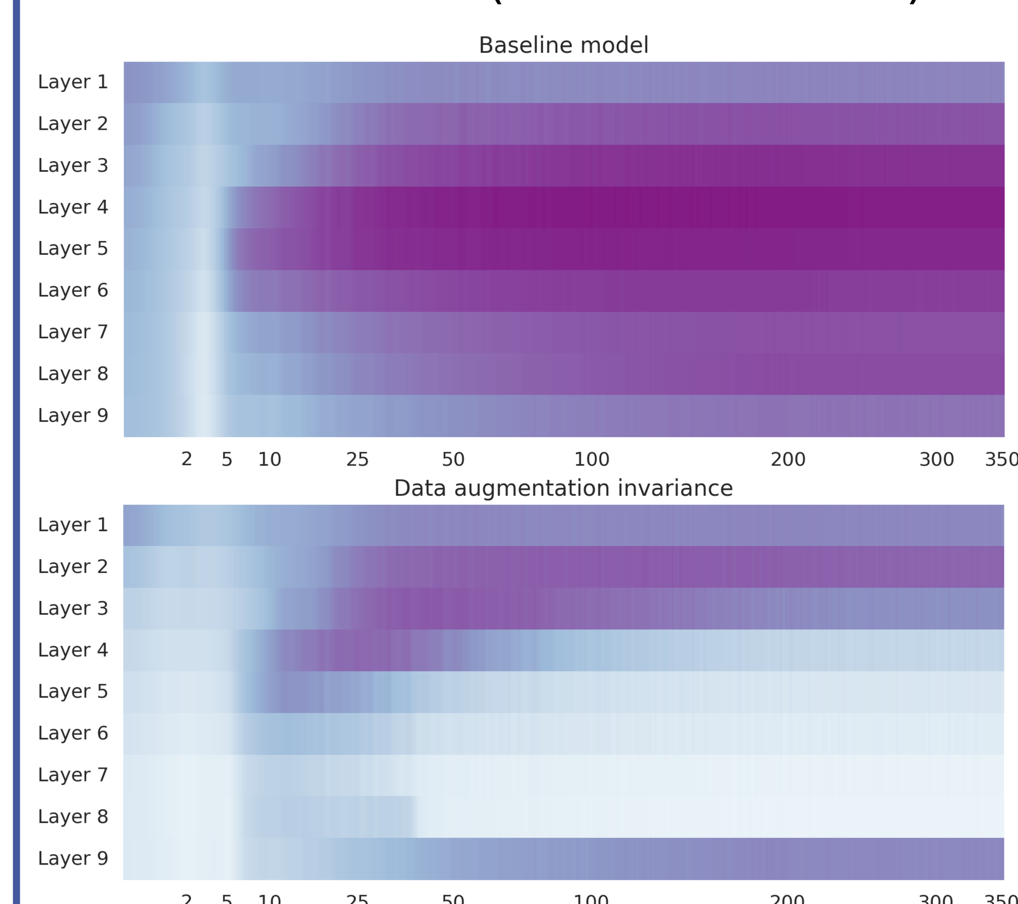


By applying our data augmentation invariance, the high level representations of transformations of the reference image become indeed similar (robust) and the categorization performance stays as good or better.



RESULTS

- The high-level representations learned by a standard DNN are not more robust to identity-preserving transformations than at pixel space (red boxes).
- Our proposal, **data augmentation invariance**, successfully learns invariant representations (blue boxes).
- The model needs only 10 % more training time to learn the invariance.
- The classification performance is not affected: 92.2 % (baseline: 91.5 %)



Dynamics of the invariance loss

- In the baseline model (top), the invariance loss increases during training.
- With **data augmentation invariance** (bottom), the loss smoothly decreases.

CONCLUSIONS

- We have empirically shown that prototypical DNNs are not invariant to identity-preserving transformations.
- This property is fundamentally different to the primate visual ventral stream.
- Taking this as inspiration, we have proposed an unsupervised objective that encourages learning robust features.
- We create mini-batches with augmented examples and modify the loss function to maximize their similarity.
- Our method, data augmentation invariance effectively produces more robust representations, at no cost in performance and only 10 % increase of training time.

[1] DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. Trends in Cognitive Sciences.
[2] Tacchetti, A., Isik, L., & Poggio, T. A. (2018). Invariant recognition shapes neural representations of visual input. Annual review of vision science.
[3] Kietzmann, T. C. et al (2017). Deep neural networks in computational neuroscience. Oxford Research Encyclopedia of Neuroscience.
[4] Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. JMLR.
[5] Hernández-García, A., & König, P. (2018). Data augmentation instead of explicit regularization. arXiv preprint.