# Learning representational invariance instead of categorization

Alex Hernández-García
Institute of Cognitive Science, University of Osnabrück
alexhernandezgarcia.github.io

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

## Abstract

*The current most accurate models of image object categorization are deep neural networks trained on large labeled data sets. Minimizing a classification loss between the predictions of the network and the true labels has proven an effective way to learn discriminative functions of the object classes. However, recent studies have suggested that such models learn highly discriminative features that are not aligned with visual perception and might be at the root of adversarial vulnerability. Here, we propose to replace the classification loss with the joint optimization of invariance to identity-preserving transformations of images (data augmentation invariance), and the invariance to objects of the same category (class-wise invariance). We hypothesize that optimizing these invariance objectives might yield features more aligned with visual perception, more robust to adversarial perturbations, while still suitable for accurate object categorization.*

## 1. Introduction

Image object categorization performance dramatically increased with the successful training of deep artificial neural networks. Instead of using handcrafted features, DNNs automatically learn highly discriminative features from large labeled data sets. Such impressive performance, reminiscent of human visual object categorization, and the fact that some similarities have been found [7] between the features learned by DNNs and the activations measured in the high-level visual cortex can make us think that neural networks solve visual object recognition in a similar way to how the brain does. However, important differences remain.

A remarkable example of the mismatch between DNNs and primate visual perception is the well-known vulnerability of the former to adversarial perturbations [13], which make DNNs classify instances in a perceptually implausible way. Recent work [6] has suggested that adversarial vulnerability might be caused by highly discriminative features present in the data yet incomprehensible to humans. Notably, this is only one example of the differences between current artificial and biological visual object perception.

While for many applications a high discriminative performance is enough, disciplines such as computational neuroscience demand models that reasonably match some aspects of human perception. Besides, we believe that exploring the connections between computer vision and biological vision [10] and pushing the development of artificial neural networks towards more perceptually aligned solutions, can help us better understand the generalization properties of DNNs and, potentially, obtain better, more robust models.

One step towards the integration of deep learning and neuroscience is to incorporate properties of visual perception and the visual cortex into the computer vision algorithms. Rather than the architectural aspects, here we focus on the learning objective. While the most accurate models for image object recognition are trained by minimizing a loss between the predicted and the true class of the image samples, it has been argued that the visual brain develops with little supervised information [1]. Although the specific mechanisms that yield robust object recognition in the brain are yet to be well understood, a well established theory is that invariance may play an important role.

It has been proposed [14] that a major property of biological vision is the increasing invariance of neural populations along the processing hierarchy towards identity-preserving transformations of the objects. Moreover, it is widely accepted that higher areas of the visual cortex form similar patterns of activation within relevant object categories [3].

In this paper, we propose to incorporate these invariances into the optimization of DNNs trained on object categorization data sets. In particular, we combine data augmentation invariance [5] and class-wise invariance [2] as a single semi-supervised objective. We hypothesize that the features obtained through *invariance learning* may be more aligned with visual perception, less vulnerable to adversarial perturbations, while still suitable for object categorization.

## 2. Related work

The semi-supervised learning (SSL) literature (see [11] for a review of recent methods) offers a wide range of approaches that aim at exploiting desirable invariance properties in the data and the learning algorithm. Ladder networks [12] jointly optimize the classification objective and a layer-wise denoiser. Data augmentation has been used before as a source of stochastic variability during training, together with dropout, random max-pooling and other sources of randomness [9]. Typically in SSL, the unsupervised objective is used to complement the classification objective to more efficiently learn from fewer labeled examples.

In contrast, our focus is on learning representations with desirable properties inspired by biological vision and perception, by fully replacing the classification objective with data augmentation and class-wise invariance. Thus, we do not use perceptually irrelevant sources of variability, such as dropout. Our method may also be able to efficiently learn from fewer data and, for that purpose, we explore the trade off between the unsupervised (data augmentation) and the supervised (class-wise) invariance objectives.

## 3. Methods

This section introduces the two learning objectives that we propose as an alternative to the classification loss: data augmentation and class-wise invariance.

### 3.1. Data augmentation invariance

Data augmentation invariance has been recently proposed [5] as a simple way of learning features robust to identity-preserving transformations. The authors showed that the deep features learned by a standard convolutional neural network are not more robust than in the pixel space to the transformations used in data augmentation schemes, such as rotation, scaling or brightness adjustment. However, adding a term to the loss function that promotes the similarity between the representations of transformations of the same object enables learning robust features while keeping the same categorization performance or higher.

We believe that learning such invariant representations is a desirable property and is motivated by the invariance observed along the visual ventral pathway of the primate brain [14]. Interestingly, data augmentation invariance is a fully unsupervised objective, since it does not require labeled data. Yet, data augmentation invariance alone may bias the model towards learning trivial, useless features. We believe that some degree of supervision might be necessary and this can be provided by class-wise invariance.

### 3.2. Class-wise invariance

Class-wise invariant representation learning [2] was introduced as a regularization term that encourages similarity in the representations of objects from the same class. The authors showed that class-wise invariance helps improve generalization, especially when few examples are available.

Class-wise invariance is interesting because, in spite of being a supervised algorithm, it sets the learning objective on the intermediate features, rather than solely on the classification with the top-most features. However, used on its own it would possibly be subject to some of the same undesirable properties of purely supervised methods. We hypothesize that combined, data augmentation and class-wise invariance alone may learn robust, discriminative features.

### 3.3. Learning objective

Let $\mathcal{L}_{DA}^{(l)}$ be the data augmentation invariance loss and $\mathcal{L}_{C}^{(l)}$ the class-wise invariance loss at layer $l$ of a neural network model with $L$ layers. We propose to optimize, through stochastic gradient descent, the following overall objective:

$$\mathcal{L} = \sum_{l=1}^{L} \alpha^{(l)} \mathcal{L}_{DA}^{(l)} + \sum_{l=1}^{L} \beta^{(l)} \mathcal{L}_{C}^{(l)} \tag{1}$$

where $\alpha^{(l)}$ and $\beta^{(l)}$ are scalars that control the degree of similarity between the features of augmented samples and of objects of the same category, respectively, at each layer $l$ of the architecture. Given a data set of size $N$, we construct each mini-batch $\mathcal{B}$ by randomly sampling $K$ images and generating $M$ stochastic augmentations for each of them. Thus, each batch consists of $|\mathcal{B}| = K \times M$ data points.

We define the data augmentation invariance loss identically as in [5]:

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

where $\mathcal{S}_k$ are the subsets from $\mathcal{B}$ formed by augmented versions of the same seed sample $x_k$. For the class-wise invariance loss, instead of using the exact definition from [2], we rather use a parallel definition to Equation 2, for convenience, which keeps the same spirit—to promote the similarity of representations of images from the same class:

$$\mathcal{L}_{C}^{(l)} = \frac{\sum_r \frac{1}{|\mathcal{T}_r|^2} \sum_{x_i, x_j \in \mathcal{T}_r} d^{(l)}(x_i, x_j)}{\frac{1}{|\mathcal{B}|^2} \sum_{x_i, x_j \in \mathcal{B}} d^{(l)}(x_i, x_j)} \tag{3}$$

where $\mathcal{T}_r$ are the subsets from $\mathcal{B}$ formed by images of the same object class $r$. Regarding the similarity metric,

$d^{(l)}(x_i, x_j)$, the authors of [5] tested the mean squared difference and in [2] three metrics were tested, but the squared Euclidean distance was identified as the most successful. Therefore, we consider the mean squared difference a reasonable, meaningful and computationally efficient choice. Nonetheless, the proposed method allows for other metrics, which would be interesting to explore in the future.

## 4. Experimental protocol

This section first describes the essential experiments that we will perform to assess our proposal and then some other desirable tests that would shed more light on our algorithm.

### 4.1. Essential experiments

To the extent possible and applicable, we will follow the guidelines in [11] to assess SSL algorithms. Initially, we will test our invariance learning on two architectures, Wide ResNet and All-CNN, and train on CIFAR-10/100. First, we will need to verify that the objective defined in Equation 1 is optimized and thus the model converges. Ideally, the model should learn representations such that the classes form separate clusters and, in turn, transformations of the same data point are close to each other (see Figure 1).

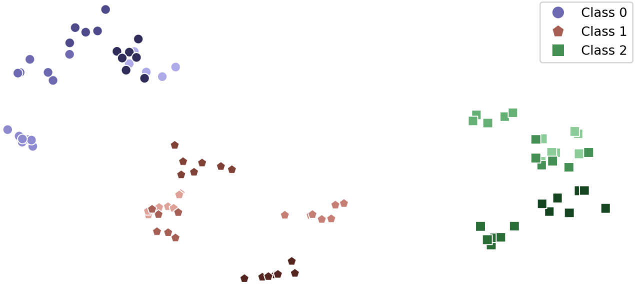


Figure 1: Simulation of a desirable projection of the features in two dimensions. Augmented versions of the same data point are plotted with exactly the same color.

Such a visualization of the learned features could be obtained through techniques such as t-SNE. Yet, we plan to perform additional tests. To assess the robustness of the features, we will compute the similarity of augmented test images with alternative metrics, for instance the recently proposed centered kernel alignment (CCK) [8]. To test whether the learned features are indeed useful for categorization, we will train both a linear model and a neural network with one hidden layer with the features of the last layer ($L$) of our invariance learning model. A successful model should not perform significantly worse than the baseline model trained with the cross entropy loss.

Although it is improbable that our proposal completely solves the adversarial vulnerability, it may help increase the robustness. Therefore, we will assess the adversarial robustness of our invariance learning model by creating both white- and black-box attacks, using the fast gradient sign method (FGSM) and projected gradient descent (PGD).

Regarding the hyperparameters, we will explore which $\alpha^{(l)}$ and $\beta^{(l)}$ guarantee the joint optimization of $\mathcal{L}_{DA}^{(l)}$ and $\mathcal{L}_{C}^{(l)}$ and good classification performance. A reasonable approach would be to set $\alpha = \sum_l \alpha^{(l)}$ and $\beta = \sum_l \beta^{(l)}$ such that $\alpha = 1 - \beta$ and progressively increase the value of $\beta$ during training, such that the class-wise invariance becomes more important only provided the features are sufficiently robust. Similarly to [5], both $\alpha^{(l)}$ and $\beta^{(l)}$ could be distributed exponentially, such that higher layers become more invariant, as it is thought to occur along the visual cortex.

We will use the data augmentation schemes used in [5, 4] and, in the spirit of [4], we will not include any explicit regularization (e.g. weight decay and dropout) in our models.

### 4.2. Other desirable experiments

In addition to the essential experimental setup described in 4.1, other tests would shed more light on the benefits and limitations of the proposal. In particular, the priority should be to train the models on ImageNet.

Interesting, yet out of the scope of this paper, would be to compare the representations learned with the proposed model to the activations measured through fMRI in the visual cortex [7], since one motivation for this proposal is to learn more human-like features.

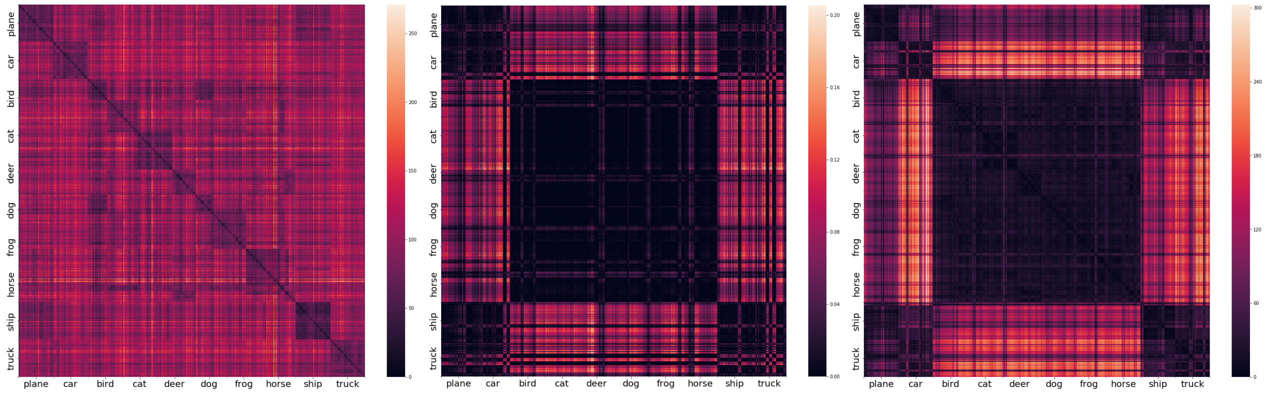
## 5. Results

The purpose of this section is to briefly outline the main findings of this work, both positive and negative, so as to facilitate future research in this direction.

The most ambitious hypothesis of this paper, that is that it may be possible to fully replace the standard classification loss by the joint optimization of class-wise and data augmentation invariance, has not yet been possible to demonstrate through the methods presented in Section 3: we could not achieve comparable classification accuracy on CIFAR-10/100 by training a single-layer neural network with the output features learned by the models trained solely with the invariance objectives.

Note, however, that this was an ambitious and challenging task. In order to better understand the effect of the invariance losses, we looked into the representations of the top layer of All-CNN, trained on CIFAR-10, as proposed in Section 4.1. We compared a baseline, purely categorization model, a purely invariance learning model (with $\alpha = 0.1$, $\beta = 0.9$) and a model trained with both invariance and categorical losses. Figure 2 shows the dissimilarity matrices of the top-layer features of the three models.

Interestingly, clear clusters (groups of classes) are



(a) Only categorical loss (b) Only invariance loss (c) Both invariance and categorical losses

Figure 2: Dissimilarity matrix of models trained with different losses, constructed with 100 examples of each class.

formed when the models are trained with invariance learning, while the dissimilarity matrix of the standard model is fairly homogeneous. Remarkably, the main clusters formed through invariance learning correspond to the animate and inanimate classes, a separation consistently observed in the primate visual cortex [7]. Furthermore, although the mixed-losses model trained achieves comparable classification performance to the baseline model (91.9 % and 93.2 % respectively), its representational organization is more similar to the pure invariance learning model (21 % accuracy).

Since both its accuracy and its representational pattern (Figure 2b) indicate that pure invariance learning manages to separate only two groups of classes, we also computed the explained variance by the first principal component of the top-layer representations and found it to be extremely high, 0.996, while the standard model's is 0.085. Surprisingly, it is also high in the model with mixed losses: 0.743. This suggests that the current form of the invariance losses constrain the model towards using one or few dimensions, what prevents it from discriminating multiple classes.

To further evaluate this idea, we trained models to perform binary classification on pairs of classes from CIFAR and found that, in this case, not only the invariance learning models do match the categorization models in terms of classification accuracy, but they achieve a better class separation which results in a largely improved adversarial robustness.

## 6. Conclusion

In sum, although the results revealed that the present proposal of invariance learning is not yet comparable to categorization models for multivariate classification, the representational organization of the features, which resembles patterns observed in the visual cortex, and the improved adversarial robustness suggest that invariance instead of categorization may be a promising research avenue to follow.

## References

[1] J. Atkinson. *The developing visual brain*. Oxford Scholarship Online, 2002. 1

[2] S. Belharbi, C. Chatelain, R. Hérault, and S. Adam. Neural networks regularization through class-wise invariant representation learning. *arXiv:1709.01867*, 2017. 2, 3

[3] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 2007. 1

[4] A. Hernández-García and P. König. Data augmentation instead of explicit regularization. *arXiv:1806.03852*, 2018. 3

[5] A. Hernández-García, P. König, and T. C. Kietzmann. Learning robust visual representations using data augmentation invariance. *arXiv:1906.04547*, 2019. 2, 3

[6] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *arXiv:1905.02175*, 2019. 1

[7] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 2014. 1, 3, 4

[8] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. *arXiv:1905.00414*, 2019. 3

[9] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *arXiv:1610.02242*, 2016. 2

[10] A. H. Marblestone, G. Wayne, and K. P. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 2016. 1

[11] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 2, 3

[12] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015. 2

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013. 1

[14] A. Tacchetti, L. Isik, and T. A. Poggio. Invariant recognition shapes neural representations of visual input. *Annual review of vision science*, 2018. 1, 2