Can Generalised Divergences Help for Invariant Neural Networks?

Santiago Velasco-Forero

Can generalised divergences help for invariant neural networks?

Santiago Velasco-Forero

MINES Paris - PSL-Research University - Centre de Morphologie Mathématique

Abstract. We consider a framework including multiple augmentation regularisation by generalised divergences to induce invariance for non-group transformations during training of convolutional neural networks. Experiments on supervised classification of images at different scales not considered during training illustrate that our proposed method performs better than classical data augmentation.

1 Introduction

Deep neural networks are the primary model for learning functions from data, in different tasks ranging from classification to generation. Convolutional neural networks (CNNs) have become a widely used method across multiple domains. The *translation* equivariance of convolutions is one of the key aspects to their success [**?**]. This equivariance is induced by applying the same convolutional filter to each area of an image producing learned weights that are independent of the location. Ideally, CNNs should perform equally well regardless of input scale, rotation or reflections. Numerous attempts have been made to address using the formalism of *group-convolutions*[**?**], *steerable filters*[**?**], *moving frames*[**?**], *wavelet*[**?**], *partial differential equations* [**?**], *Gaussian filters*[**?**], *Elementary Symmetric Polynomials*[**?**] among others. Despite all these recent advances, it is still unclear what is the most adequate way to adapt these methods for the case of more general transformations that cannot be considered as a group [**?**,**?**]. The most commonly used solution is to take advantage of *data augmentation*, where the inputs are randomly transformed during training to induce an output (which is) insensible to some given transformations [**?**]. However, data augmentation implies neither equivariance nor invariance.

In this paper, we study the use of contrastive based regularisation on a set of transformations during training. Surprisingly, our proposition presents the best performance considering the power of generalisation outside the interval of values where the transformation has been sampled during training. This phenomenon is illustrated in the case of supervised classification on aerial images and traffic signs at different scales.

2 Proposition

2.1 Motivation

Data augmentation is nowadays one of the main components of the design of efficient training for deep learning models. Initially proposed to improve oversampling on class-imbalanced datasets [**?**] or to prevent overfitting when the model contains more parameters than training points [**?**]. Recent research has shown its interest in increasing generalization ability especially when augmentations yield samples that are diverse [**?**]. We restrict our study to augmentations which act on a single sample and do not modify labels, this means that we do not consider *mixup augmentations* [**?**]. Namely, we study augmentations which can be written as $(t(\mathbf{x}), y)$, where $(\mathbf{x}, y)$ denotes an input-label pair, and $t \in \mathcal{T}$ is a random transformation sampled from a set of possible transformations $\mathcal{T}$. Let f denotes a projection from the input space to a *latent space*. The latent space is said to be *invariant* to $\mathcal{T}$ if for any input $\mathbf{x}$ and any $t \in \mathcal{T}$, $f(t(\mathbf{x})) = f(\mathbf{x})$. Practitioners recommend to use *data augmentation* to induce invariance *by training*[**?**]. Usually, data augmentation consists of randomly applying an element of the set of $\mathcal{T}$ during training. An alternative to data augmentation is possible when $\mathcal{T}$ is a group. One can construct an invariant function $f_\theta(\mathbf{x}; \eta)$ from a non-invariant function $g_\theta(\mathbf{x})$ by integrating over all the group actions. This concept is referred to as *insensitivity* [**?**], *soft-invariance* [**?**], or *deformation stability* [**?**]. The special case in which there exist a subgroup H where the computation can be reduced to summing over H is called *Reynolds design* [**?**]. For topological groups, there is a non-zero, translation invariant measure called *Haar measure* that can be used to define invariant convolution on a group [**?**,**?**] or invariance by integration of kernels [**?**,**?**]. An alternative to define an invariant function, is to use composition of *equivariant* functions followed by an invariant pooling in the *Geometric Deep Learning Blueprint* [**?**]. In both cases, the invariance is defined *by structure*[**?**] which can be seen as a constraint in the model that one is learning.

However, in many applications, the transformations under study is not a group, so that the above arguments are not easily generalizable. In this paper we are interested in using data augmentation to induce invariance during training in deep learning models for the case where the set $\mathcal{T}$ is not a group. The idea is to include a regularisation term that takes into account K realisations of the transformation family, i.e, in the loss function, we will include $\texttt{Loss}_{\mathcal{T}}(\mathbf{x}, t_1(\mathbf{x}), t_2(\mathbf{x})), \ldots, t_K(\mathbf{x})))$ where the t_i denotes a random value of transformations $\mathcal{T}$. To apply this method we do not need any requirements on the transformations $\mathcal{T}$.

2.2 Related work

The idea of using multiple random augmentations (K is our case) during training is also found in the following methods:

Semi-supervised learning In a semisupervised case, [**?**] proposed to learn a classifier penalised for quick changes in its predictions.

Self-training Self-training also known as decision-directed or self-taught learning machine, is one of the earliest approach in semi-supervised learning [**?**,**?**]. The idea of these approaches is to start by learning a supervised classifier on the labelled training set, and then, at each iteration, the classifier selects a part of the unlabelled data and assigns pseudo-labels to them using the classifier's predictions. These pseudo-labeled examples are considered as additional labeled examples in the following iterations. The function loss includes a trade-off term to balance the influence of pseudo labels.

Self-supervised learning Most of these works are placed in a *joint-embedding framework* [**?**,**?**], where augmented views (usually two) are generated from a source image. These two views are then projected to an encoder, giving representations, and then through a projection back to an embedding space. Finally, a loss minimises the distance between the embeddings, i.e. makes them invariant to the augmentations, and is combined with a regularisation loss to spread embeddings in space.

Data augmentation regularisations A negative aspect of data augmentation has been illustrated in [**?**] which is the slow down of training speed and a minimal effect on the variance of the model. The idea of using multiple augmentation per image in the same minibatch has been used to solve that problem, and it has been used to improve at the same time the classifier's generalisation performance [**?**]. This simple modification computes an average of the minibatch on different augmentations that asymptotically approaches a Reynolds operator (**??**) when the number of considered augmentations gets as large as possible. Recently, [**?**,**?**] proposed the use of a regularisation term for multiple augmentations, which is the mechanism that we will evaluate in this paper.

2.3 Supervised regularisation by generalised divergences

Fig. 1. Scheme of our proposition. We propose to use a regularisation that considers multiple realisations of the transformation family, this regularisation uses generalised divergences. Since you want to evaluate the invariance of a classification problem, the model uses only the classification of the original image (not of the transformations). The probability distributions are obtained in the output of a softmax layer.

We propose to use multiple data augmentations in the target transformation, and use a generalised divergence as a regularisation term. The idea follows those

presented in [**?**], and is contrary to the usual mechanism of *data augmentation*, where the network is trained to classify in the same class each of the augmentations, but never considers a term related to the divergence produced by the transformation. Since we use K augmentations, we must consider a divergence from multiple probability distributions, which is called *generalised divergences*. We consider the classical framework of training deep learning models from N samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and as objective minimising the following loss function:

$$\texttt{Loss}(\mathbf{x}, y) = \sum_{i}^{N} \texttt{Loss}_{class}(y_i, \hat{y}_i) + \alpha \sum_{i}^{N} \texttt{Loss}_{\mathcal{T}}(\mathbf{x}_i, t_0(\mathbf{x}_i), t_1(\mathbf{x}_i), \ldots, t_K(\mathbf{x}_i)), \tag{1}$$

where $\hat{y}_i$ denotes the prediction of the model, and y_i the ground-truth class of the i-th sample $\mathbf{x}_i$.

The first term is a supervised classification term, and the second term $\texttt{Loss}_{\mathcal{T}}$ is the main interest of our proposition. We propose to use statistical divergences to compare the outputs produced by model f applied to the original data $\mathbf{x}$ and $K+1$ random augmentations of $\mathbf{x}$, i.e $\{\mathbf{x}_i, t_0(\mathbf{x}_i), t_1(\mathbf{x}_i), \ldots, t_K(\mathbf{x}_i)\}$. In our supervised case, we use the last layer of the model f, which is usually a sum-one layer (softmax) indicating the probability of belonging to a given class. For two probability distributions P, Q, the most renowned statistical divergence rooted in information theory [**?**] is the *Kullback–Leibler divergence*,

$$D_{KL}(P||Q) = \sum P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

Defining divergence between more than two distributions has been studied for many authors called often *generalised divergences* or *dissimilarity coefficient* in [**?**]. Let $K \geq 1$ be a fixed natural number. Each generalised divergence R that we consider here, satisfies the following properties:

1. $R(P_0, P_1, P_2, \ldots, P_K) \leq 0$
2. $R(P_0, P_1, P_2, \ldots, P_K) = 0$ whenever $P_0 = P_1 = \ldots = P_K$
3. R is invariant to permutation of input components.

These three properties are important for the minimisation of this divergence to induce the invariance during training in the case we are studying. Accordingly, we consider the following two generalised divergences, the *Average Divergence* [**?**]

$$R_1(P_0, P_1, P_2, \ldots, P_K) = \frac{1}{K(K+1)} \sum_{i,j=0, i \neq j}^{K} D_{KL}(P_i||P_j) \tag{2}$$

the *Information radius* [**?**] which is the generalised mean of the Rényi's divergences between each of the P_i's and the generalised mean of all the P_i's,

$$R_2(P_0, P_1, P_2, \ldots, P_K) = \frac{1}{K+1} \sum_{i}^{K} D_{KL}((K+1)^{-1} \sum_{j}^{K} P_j||P_i) \tag{3}$$

In the following section, we compare the use of (**??**), considering as $\texttt{Loss}_{\mathcal{T}}$ the average divergence in (**??**) or the information radius (**??**).

3 Experiments

In this experimental section, we have followed the training protocol presented in [**?**] on two datasets, *Aerial* and *Traffic Signs*, which contains images 64×64 RGB-color images on 48 different scales. The objective is to obtain a scale and translation invariant model for supervised classification on nine (resp. 16) classes on *Aerial* (resp. *Traffic Signs*) dataset. Keen readers are referred to [**?**,**?**,**?**,**?**,**?**] for a deeper understanding of different propositions for scale invariant convolutional networks. Following [**?**] the model is a CNN with two layers using categorical cross-entropy as a supervised term in (**??**). An example per dataset at different scales is shown in Figure **??**. The models are trained on the middle interval of the transformation parameterisation and the performance of the models are evaluated outside this interval. This is called *Mid2Rest* scenario in [**?**]. The value of K in (**??**) and (**??**) has been set equal to three in our experiments. Quantitative comparison of results are found in Table **??** for both considered datasets. The reported result is the average and standard deviation of performance on the scales and images that were not considered during training. On the considered datasets, the information radius (**??**) presents better results in terms of performance over the unseen scales, with respect to both the average divergence (**??**), and the classical data augmentation method. Finally, for a better illustration of the difficulty of the task, the best value of the lambda and regularisation function is compared with the data augmentation in five random training runs, and compared across the different scales for the two databases in Figure **??**.

Fig. 2. Examples of images at different parameter transformation in the two considered datasets. From left to right: Scale 1, 3, 23, 25, 45 and 47. Training is done considering only images of intermediate scales (17 to 32) in both training and validation. Evaluation is performed on both small (0 to 16) and large (33-48) scales. In first row: An example of *Traffic Sign* dataset. In second row: An example of *Aerial* dataset.

Fig. 3. Detailed plots of scale generalisation on *Mid2Rest* scenario in *Aerial* datasets (Left) and *Traffic Sign* dataset (Right). Five repetitions of the training is illustrated per method. Our proposition performs clearly better than classical data augmentation.

4 Conclusions

In this paper we present a proposal for the use of regularisation from multiple data augmentation with generalised divergences. Quantitative results show

Table 1. Results of the Generalised Divergence in Aerial and Traffic Sign dataset on scales non-considered during training. A visual comparison of results are shown in Fig.**??**

Aerial		Small Scales		Large Scales	
Method	λ	test acc.	±std	test acc.	±std
Data Aug.	0.0	0.776	±0.014	0.845	±0.008
Av. Div.(**??**)	0.5	0.854	±0.013	0.889	±0.011
	1.0	0.852	±0.016	0.888	±0.009
	1.5	**0.858**	± 0.013	**0.889**	±0.011
	2.0	0.841	±0.028	0.880	±0.020
	2.5	0.846	±0.015	0.881	±0.009
	3.0	0.834	±0.023	0.877	±0.018
	3.5	0.833	±0.008	0.873	±0.017
	4.0	0.822	±0.021	0.864	±0.016
	5.0	0.828	±0.016	0.872	±0.022
	10.0	0.824	±0.019	0.865	±0.012
Inf. Rad.(**??**)	0.5	0.845	±0.016	0.878	±0.017
	1.0	0.847	±0.012	0.885	±0.011
	1.5	0.853	±0.010	0.881	±0.009
	2.0	0.853	±0.009	0.884	±0.011
	2.5	0.850	±0.014	**0.887**	±0.005
	3.0	**0.859**	±0.013	0.885	±0.010
	3.5	0.841	±0.008	0.886	±0.009
	4.0	0.839	±0.018	0.877	±0.016
	5.0	0.845	±0.017	0.883	±0.010
	10.0	0.829	±0.014	0.870	±0.011

Traffic Sign		Small Scales		Large Scales	
Method	λ	test acc.	±std	test acc.	±std
Data Aug.	0.0	0.721	± 0.021	0.824	±0.024
Av. Div. (**??**)	0.5	0.821	±0.014	0.898	±0.020
	1.0	**0.829**	±0.012	**0.921**	±0.018
	1.5	0.820	±0.015	0.902	±0.012
	2.0	0.806	±0.027	0.886	±0.025
	2.5	0.797	±0.044	0.883	±0.034
	3.0	0.789	±0.025	0.882	±0.018
	3.5	0.754	±0.026	0.842	±0.028
	4.0	0.743	±0.026	0.832	±0.042
Inf. Rad.(**??**)	0.5	0.806	±0.012	0.908	±0.017
	1.0	0.825	±0.022	**0.921**	±0.016
	1.5	0.829	±0.014	0.918	±0.015
	2.0	**0.830**	±0.016	0.918	±0.007
	2.5	0.815	±0.020	0.906	±0.016
	3.0	0.828	±0.013	0.909	±0.018
	3.5	0.818	±0.019	0.898	±0.014
	4.0	0.823	±0.020	0.917	±0.016

the interest of our method in the case of generalisation to scales that have not been considered during training. Future studies may include the study of multi-parametric transformations, as these are used to avoid overfitting in large neural networks. Additionally, generalised divergence considering barycenters for probability distributions in [**?**] seems a promising direction to generalise the results of this article.

Acknowledgements This work was granted access to the Jean Zay supercomputer under the allocation 2022-AD011012212R2.