Spatial transformations in convolutional networks and invariant recognition

Ylva Jansson, Maksim Maydanskiy, Lukas Finnveden, and Tony Lindeberg
Computational Brain Science Lab, Division of Computational Science and Technology
KTH Royal Institute of Technology, Stockholm, Sweden

I. Introduction

Convolutional neural networks (CNNs) that are *invariant* to certain groups of image transformations have fewer parameters, can learn from smaller datasets and enable *generalization outside the training distribution*. A number of current methods use spatial transformations of CNN feature maps or filters to enhance the ability of CNNs to handle different types of image transformations [1], [2], [3], [4], [5], [6], [7], [8]. For example, *spatial transformer networks* (STNs) [8] were designed to enable CNNs to learn invariance to image transformations by transforming *CNN feature maps* as well as input images. Clearly, if a network learns to align transformed input images to a common pose, this can enable invariant recognition. The original work [8], however, simultaneously claims the ability of STNs to learn invariance from data and that the spatial transformer layers (STs) can be inserted into the network "anywhere" (i.e. at any depth). There is no mention of whether the key motivation for the framework – the ability to learn invariance – is still supported when transforming feature maps deeper in the network.

This seems to have left some confusion about whether spatially transforming CNN feature maps can support invariant recognition, with a number of subsequent works advocating image alignment by *transforming feature maps* [1], [2], [3], [4]. Other commonly used methods that are based on transforming CNN feature maps or filters are spatial pyramid pooling [5], dilated convolutions [6] and deformable convolutions [7]. Such methods are often motivated by the need for CNNs to better deal with variability in object pose. There is, however, no discussion about the difference between pose normalizing the input image and spatially transforming feature maps, or the implications this choice has for the ability to achieve e.g. affine or scale invariance [5], [6], [7], [8].

We, here, aim to clear this confusion and elucidate *under what conditions* it is possible to achieve invariance to affine image transformations by means of purely *spatial transformations* of CNN feature maps. We show that these conditions are very restrictive, implying network filters or features that are *already invariant* to the relevant image transformations. This since, spatial transformations of CNN feature maps *cannot*, for general affine transformations, align the feature maps of a transformed image with those of an original.

These facts have, in the single-layer case, some parallels with the work in [9] and [10]. Our contribution is to provide a simple proof for the single layer case and to build on it to give an analysis of the general multi-layer case, using only elementary analysis and without relying on any covariance assumptions about the individual layers. A preprint with details and the full proofs is available [11].

Presented at DeepMath2020 Conference on the Mathematical Theory of Deep Neural Networks Nov 5 - Nov 6, 2020. The support from the Swedish Research Council (contract 2018-03586) is gratefully acknowledged.

II. Formalism and results

We work with a continuous model of the image space. We consider both an **image** f and a convolutional **filter** λ to be a map from $\mathbb{R}^N$ to $\mathbb{R}$ (with $f \in L^1_{loc}$ and $\lambda \in L^1_{comp}$). We use notation V for the function space to which the images f belong, and V^k for the space of maps that have each of their k components in V.

Then a *continuous CNN* with k layers and M_k feature channels in the final layer is a map $\Lambda : V \to V^{M_k}$ given inductively in components by

$$(\Lambda^{(i)} f)_c(x) = \sigma\left(\sum_{m=1}^{M_{i-1}} \int_y (\Lambda^{(i-1)} f)_m(x-y)\, \lambda^{(i)}_{m,c}(y)\, dy + b_{i,c}\right)$$

with Λ^0 being the identity map, σ_i the chosen non-linearity and $b_{i,c}$ the bias constant for c-th component of the i-th layer. In the single-layer ($k = 1$) case the corresponding single component operator Λ^1_c is denoted Λ_λ, where λ is the convolution kernel $\lambda^1_{1,c}$.

We consider the group of *affine image transformations*, i.e. of linear maps $T_h : \mathbb{R}^N \to \mathbb{R}^N$. The corresponding operator $\mathcal{T}^j_h : V^j \to V^j$ is defined by the "contragradient" representation, that is by precomposing with T_h^{-1} i.e. $(\mathcal{T}^j_h F)_c(x) := (F)_c(T_h^{-1}x)$ for $F \in V^j$. We ask whether the transformation $\mathcal{T}^1_h$ applied to the image f can be "undone" by transforming features Λf by $\mathcal{T}_g^{M_k}$ for some (possibly different) affine map T_g, i.e. under what conditions one can have $\mathcal{T}_g^{M_k} \Lambda \mathcal{T}_h f \stackrel{?}{=} \Lambda f$. If possible, this would enable invariant recognition.

In the single layer case ($k = 1$), we show that the above is only possible if $T_g = (T_h)^{-1}$ and only if the convolutional filters are themselves invariant to the relevant transformation:

Theorem. *Equality $\mathcal{T}_g \Lambda_\lambda \mathcal{T}_h = \Lambda_\lambda$ implies $T_g = T_h^{-1}$ and also $\lambda = (\det T_h)\mathcal{T}_h^{-1}\lambda$.*

By analysing the dynamics of T_h we show that the condition on λ is very restrictive, even ignoring rescaling (the proposition below is for $N = 2$ but similar statements can be made for arbitrary N).

Theorem. *The equality $\lambda = C\mathcal{T}_h^{-1}(\lambda)$ can hold for λ with support on a set of finite but non-zero measure only if T_h is conjugate to some rotation or, if T_h is orientation reversing, a reflection matrix; and in those cases only if (i) $T_h^n = Id$ for some n and λ is symmetric with respect to this finite set of transforms, or (ii) if λ is constant on a collection of concentric ellipses along which T_h rotates things.*

Thus, invariance will only possible for transformations that correspond to rotations or reflections in some basis and if using *invariant filters*. Moreover, we have a similar result for the multi-layer case:

Theorem. *If not all eigenvalues (real or complex) of T_h have absolute value equal to 1, equation $\mathcal{T}_g \Lambda \mathcal{T}_h = \Lambda$ implies that Λ is the trivial operator that outputs the same constant signal for all inputs.*

To prove the single layer statement, we use commutation relations between T_h and translations, equivariance of convolution, together

with an analysis of the dynamics of T_h. To handle the multilayer case, we then extract properties of Λ_λ that underlie most of this proof – which are continuity, translation-covariance and what we call semi-locality – and show that they hold for the multilayer operator Λ as well (in fact the last proposition is true for any continuous, semi-local and translation-covariant operator Λ). Details are available in [11].

Our results have straightforward implications for STNs and other methods that perform spatial transformations of CNN feature maps. An experimental evaluation of the practical consequences of these limitations on STNs is presented in [12].

REFERENCES

[1] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Advances in Neural Information Processing Systems*, 2016, pp. 2414–2422.

[2] J. Li, Y. Chen, L. Cai, I. Davidson, and S. Ji, "Dense transformer networks," *arXiv preprint arXiv:1705.08881*, 2017.

[3] S. Kim, S. Lin, S. R. JEON, D. Min, and K. Sohn, "Recurrent transformer networks for semantic correspondence," in *Advances in Neural Information Processing Systems*, 2018, pp. 6126–6136.

[4] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," *CoRR, abs/1703.06211*, vol. 1, no. 2, p. 3, 2017.

[8] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 2017–2025.

[9] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*, 2016, pp. 2990–2999.

[10] T. S. Cohen, M. Geiger, and M. Weiler, "A general theory of equivariant CNNs on homogeneous spaces," in *Advances in Neural Information Processing Systems (NIPS)*, 2019, pp. 9142–9153.

[11] Y. Jansson, M. Maydanskiy, L. Finnveden, and T. Lindeberg, "Inability of spatial transformations of CNN feature maps to support invariant recognition," *arXiv preprint arXiv:2004.14716*, 2020.

[12] ——, "Understanding when spatial transformer networks do not support invariance, and what to do about it," *International Conference on Pattern Recognition (ICPR2020) (to appear)*, 2021, extended version arXiv preprint arXiv:2004.11678.