Understanding image representations by measuring their equivariance and equivalence

Karel Lenc[1], Andrea Vedaldi[1]
[1]Department of Engineering Science, Oxford University.

Image representations have been a key focus of the research in computer vision for at least two decades. Notable examples include textons [6], histogram of oriented gradients (SIFT [7] and HOG [2]), bag of visual words [1][12], sparse [15] and local coding [14], super vector coding [18], VLAD [4], Fisher Vectors [8], and the latest generation of deep convolutional networks [5, 10, 16]. However, despite their popularity, our theoretical understanding of representations remains limited. It is generally believed that a good representation should combine invariance and discriminability, but this characterisation is rather vague; for example, it is often unclear what invariances are contained in a representation and how they are obtained.

In this work, we propose a new approach to study image representations. We look at a *representation* ϕ as an abstract function mapping an image $\mathbf{x}$ to a vector $\phi(\mathbf{x}) \in \mathbb{R}^d$ and we *empirically establish key mathematical properties* of this function. We focus in particular on three such properties. The first one is **equivariance**, which looks at how the representation *changes upon transformations of the input image*. We demonstrate that most representations, including HOG and most of the layers in deep neural networks, change in a *easily predictable* manner with the input. The key result is that the CNN features globally change in an easily predictable way in term of linear transformations; importantly, the *same* linear transformation works for *any* input image, and hence *any* object category, suggesting that geometry is factored in a uniform way for all of them. This is observable mainly for the first three convolutional layers as the latter layers start to be more class specific.

We show that such equivariant transformations can be learned empirically from data and that, importantly, they amount to simple linear transformations of the representation output. In the case of convolutional networks, we obtain this by introducing and learning a new *transformation layer* which allows us to e.g. use a method based on back-propagation to visualise the filters, as shown in figure 1. By analyzing the learned equivariant transformations we are also able to find and characterize the **invariances** of the representation, our second property. This allows us to quantify invariance and show how it builds up with depth in deep models.

The third property, **equivalence**, looks at whether the information captured by heterogeneous representations is in fact the same. CNN models, in particular, contain millions of redundant parameters [3] that, due to non-convex optimization in learning, may differ even when retrained on the same data. The question then is whether the resulting differences are genuine or just apparent. To answer this question we learn *stitching layers* that allow swapping parts of different networks. Equivalence is then obtained if the resulting "Franken-CNNs" perform as well as the original ones. We show that a very good level of equivalence can be established between networks with different weights trained for the same task. We also observe that for networks trained for different tasks (such as ILSVRC 2012 [9] and Places dataset [17]) it is harder to find the projection between the image representations with increasing depth as shown in table 1. This corroborates the intuition that the representation generated by lower convolutional layers are generic image codes, whereas the higher layers are task-specific.

As a complement of the theoretical investigation we show a direct practical application of the learned equivariant mappings to structured-output regression [13] on the task of pose estimation. We show that the equivariant map can significantly accelerate the regressor in a simple and elegant manner for both HOG and CNN features.

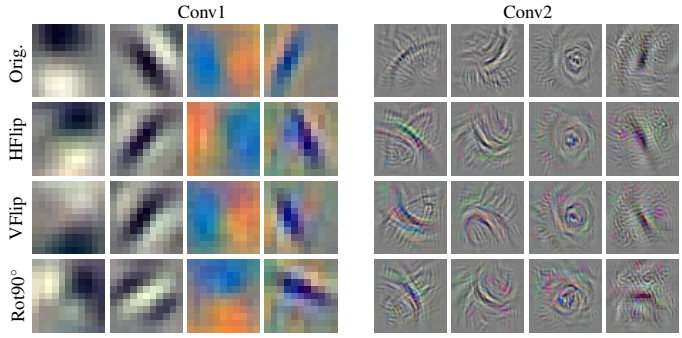

Figure 1: **Equivariant transformation of CNN filters.** Top: Conv1 and Conv2 filters of a convolutional neural network visualised with the method of [11]. Other rows: geometrically warped filters reconstructed from an equivariant transformation of the network output learned using the proposed method for Horizontal flip, Vertical flip and Rotation 90°.

Layer	IMNETA → IMNETB		PLCS → IMNETB	
	Top1	Top5	Top1	Top5
Conv1	0.43	0.20	0.43	0.20
Conv2	0.46	0.22	0.47	0.23
Conv3	0.46	0.22	0.50	0.25
Conv4	0.46	0.22	0.54	0.29
Conv5	0.50	0.25	0.65	0.39

Table 1: **CNN equivalence.** Performance on the ILSVRC12 validation set of several "Franken-CNNs" obtained by stitching the first portion of IMNETA, PLCS up to a certain convolutional layer and the last portion of IMNETB where IMNETA and IMNETB are different networks trained on the ILSVRC 2012 object classification task and PLCS on the Places dataset - a scene recognition task. For reference, the top-1 and top-5 error of the unmodified IMNETB are 0.43 and 0.20 respectively and without the stitching layer in all cases the top-1 error is $> 99\%$.

[1] G. Csurka, C. R. Dance, L. Dan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Stat. Learn. in Comp. Vision*, 2004.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[3] M. Denil, , B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas. Predicting parameters in deep learning. In *Proc. NIPS*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[4] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[6] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1), 2001.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004.

[8] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorizaton. In *Proc. CVPR*, 2006.

[9] 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, 2014.

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. volume abs/1312.6229, 2014.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2013.

[12] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[13] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Proc. NIPS*, 2003.

[14] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *Proc. CVPR*, 2010.

[15] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *Proc. CVPR*, 2010.

[16] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.

[18] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *Proc. ECCV*, 2010.