Prototypical Priors: From Improving Classification to Zero-Shot Learning

Saumya Jetley
sjetley@robots.ox.ac.uk
Bernardino Romera-Paredes
bernard@robots.ox.ac.uk
Sadeep Jayasumana
sadeep@robots.ox.ac.uk
Philip Torr
phst@robots.ox.ac.uk

University of Oxford
Oxford, UK

Automatic object recognition has witnessed a huge improvement in recent years due to the successful application of convolutional neural networks (CNN). This boost in performance can be explained by the replacement of heuristic parts in the previous feature representation approaches by a methodology [2, 3] based on learning the features straight from the data. The learned feature representation, which is tailored to the given learning scenario, generally outperforms heuristic approaches provided the training data is big enough. When learned over a significant sample variety, this representation captures regularities across samples of a class that help distinguish it from all the other classes.

In an alternative setup, the object recognition problem can be posed as one in which objects in real images are identified by treating them as imperfect and corrupted copies of prototypical concepts. This assumption provides an additional premise that the different samples of a class are not only similar to each other but also resemble a unique prototype. These prototypical concepts are in many cases not available, for example there does not exist a chair that contains only the essence of *chair* and nothing else. However, there are many scenarios where such prototypical instances do exist. An example of this is traffic sign recognition, in which each traffic sign class has its canonical template.

In the present work, we focus on adding this prototypical prior information into convolutional neural networks, as illustrated in Figure 1. The underlying idea is that the high-level representation learned by a CNN should be comparable to the information extracted from the prototypes. An interpretation of this is that layer-by-layer the CNN is able to learn a representation that is invariant to real world factors such as light variation, view point distortion, as described in [1], so that the representation obtained at the end of the network is invariant to all factors appearing in real images, and thus comparable to the prototype.

Prototypical information is introduced by wedging a layer before the output layer, fully connected to the C output neurons using the fixed weights $\phi(p_c) \in \mathbb{R}^k$ for all $c \in \{1,\ldots,C\}$. The new layer and its connections are shown in *blue* (*dark* for grayscale) in Figure 1. Thus, the $k \times C$ weight matrix for the last fully connected layer f_L is defined as a set of $k \times 1$ vectors $\phi(p_c)$ one for each $c \in \{1,\ldots,C\}$. In Figure 1, we use $\phi_1(p_c), \phi_2(p_c), \ldots, \phi_k(p_c)$ to represent the elements of the k-dimensional vector $\phi(p_c)$.

The modified network can now be described using the following formula:

$$\hat{y} = \underset{c\in\{1,\ldots,C\}}{\operatorname{argmax}} \; s\left(f_L\left(f_{L-1}\left(\ldots\left(f_1(x)\right)\ldots\right)\right)\right)_c = \underset{c\in\{1,\ldots,C\}}{\operatorname{argmax}} \langle \phi(p_c), \psi(x) \rangle, \tag{1}$$

where $\psi(.)$ and $\phi(.)$ represent the projections of input images and output labels into the joint feature space, respectively. An interpretation of this approach is that the learnable part of the network, $\psi : \mathbb{R}^d \to \mathbb{R}^k : \psi = f_{L-1} \circ \ldots \circ f_1$, learns a non-linear mapping from the original images to a k-dimensional latent space, which in this case is defined by the prototypes.

Thus, the traditional CNN pipeline is augmented to map both the input and prototypes to a common feature space with the end goal of minimizing the final recognition error. The use of a joint embedding space, as shown in Figure 2 lends the proposed model an interesting possibility of applying it to recognize new classes not present at the training stage. This aligns the approach within the areas of zero and one-shot learning.

Conclusively, this paper makes the following contributions - (a) development of a CNN that is able to use prototypical information to guide its learning process, (b) its application to classification tasks presenting a boost in overall performance, (c) establishment of a new benchmark in logo recognition (on Belga logo dataset), and (d) the seamless application of the proposed model in zero-shot learning scenarios, given the prototypical information of new classes at run time.

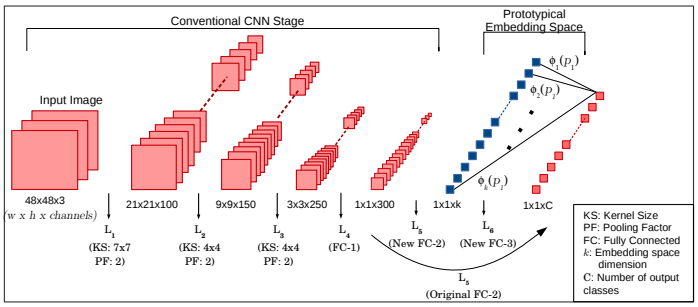

Figure 1: Network architecture with the introduction of prototypical priors. In the current experiments, k-dimensional HoG features extracted over the prototypical templates are used to define the common embedding space.

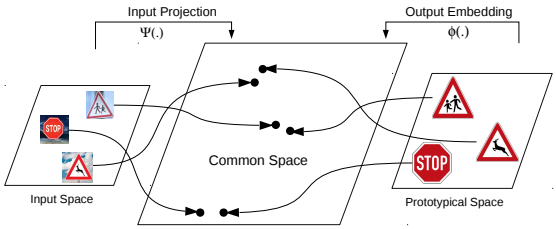

Figure 2: A joint embedding space defined by the prototypes

As observed on two different datasets of traffic signs and brand logos, results of the proposed approach are highly promising. Incorporating the given prototypes improves the classification performance. With regard to zero-shot learning, our model shows better results than a state-of-the-art competitor [4] and we show that it can be more flexibly trained for the required trade-off between seen and unseen class performance at test time.

[1] Ian Goodfellow, Honglak Lee, Quoc V Le, Andrew Saxe, and Andrew Y Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[3] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[4] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.