The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories

Tatiana Tommasi
http://www.idiap.ch/~ttommasi/
Barbara Caputo
http://people.idiap.ch/caputo/

Idiap Research Institute
Martigny, CH
Ecole Polytechnique Federale EPFL
Lausanne, CH

A major goal in object categorisation is learning and recognising effectively thousands of categories, as humans do [1]. To this end, a very promising trend is to develop methods for learning from small samples by exploiting prior experience via knowledge transfer. The basic intuition is that, if a system has already learned N categories, learning the $N+1^{th}$ should be easier, even from one or few training samples, because the algorithm can take advantage of what was learned already. Several approaches have been proposed so far for transferring knowledge, spanning from transferring model parameters [4, 9], to samples [5, 10], to general categorical properties [3], using also information coming from unlabelled data [7, 8]. While all of these approaches proved to work reasonably well in some domain, how to transfer is still an open research question.

This paper presents an algorithm that addresses this issue. We build on recent work on LS-SVM-based model adaptation [6], where a crucial requirement is having available many samples of the new class. Let us assume that we want to learn a new category from a set of labelled training data $\{\mathbf{x}_i\}_{i=1,m}$, taking advantage of what learned so far. Orabona et al. [6] proposes to start the training with a known model and then refine it through adaptation constraining a new model to be close to one of a set of pre-trained models. The proposed method is mathematically formulated changing the classical LS-SVM regularization term and defining the following optimisation problem [6]:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}-\beta\mathbf{w}'\|^2+\frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \quad \text{subject to} \quad y_i=\mathbf{w}\cdot\phi(\mathbf{x}_i)-b+\xi_i \quad (1)$$

where $\mathbf{w}'$ is the parameter describing the old model and β is a scaling factor necessary to control the degree to which the new model is close to the old one. To find the optimal β, the authors take advantage from the possibility of LS-SVM to write the leave-one-out error $r_i^{(-i)}$ in closed form, and use it to evaluate the criterion error [2]:

$$ERR=\sum_{i=1}^{l}\Psi\{y_i r_i^{(-i)}-1\} \quad \text{with} \quad \Psi\{z\}=\frac{1}{1+exp\{-10*z\}}. \quad (2)$$

So for each known model it is possible to find the best β producing the lowest criterion error ERR (2). Moreover, comparing all the criterion errors, the lowest one identifies the best prior knowledge model to use for adaptation. In this way the resulting algorithm determines automatically from where to transfer and how much to rely on the transferred knowledge.

We extended this model in order to enable it to learn a new category even from only one image.

(1) We substituted the criterion error ERR (2) with the leave-one-out cross-validation estimate of the Weighted Error Rate (WERR) [2]:

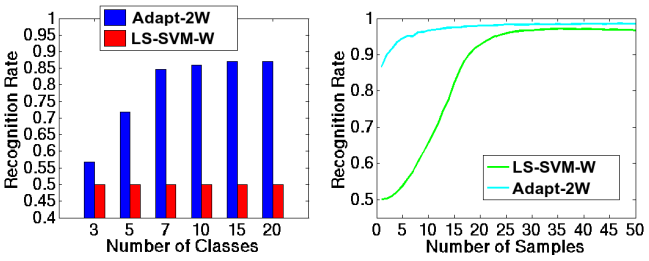
$$WERR=\sum_{i=1}^{l}\zeta_i\Psi\{y_i r_i^{(-i)}-1\} \quad \text{where} \quad \zeta_i=\begin{cases}\frac{l}{2l^+} & \text{if } y_i=+1\\ \frac{l}{2l^-} & \text{if } y_i=-1.\end{cases} \quad (3)$$

Here l^+ and l^- represent the number of positive and negative examples respectively. Introducing the weighting factors ζ_i is asymptotically equivalent to re-sampling the data so that object and non-object samples are balanced [2].

(2) We used a LS-SVM weighted formulation

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}-\beta\mathbf{w}'\|^2+\frac{C}{2}\sum_{i=1}^{l}\zeta_i\xi_i^2 \quad \text{subject to} \quad y_i=\mathbf{w}\cdot\phi(\mathbf{x}_i)-b+\xi_i. \quad (4)$$

In this way the weighting factors ζ_i take into account that the proportion of positive and negative examples in the training data are known not to be representative of the operational class frequencies. More in detail, introducing a weight let the classification model to be built balancing the contribution of penalties coming from different labelled examples. Let's call *LS-SVM-W* the non-adaptive method simply corresponding to (4) with $\beta=0$, and *Adapt-2W* the strategy which combines together the weighted model adaptation technique (4) and the WERR (3).



(a) (b)

Figure 1: (a) one-shot learning performance varying the total number of categories. (b) classification performance as a function of the number of training images when learning on 20 object categories.

We present three set of experiments, designed for studying the behaviour of our algorithm when (a) it knows few categories, and none of them is very similar to the new one; (b) it knows few categories that are very similar to the new one; (c) the number of known categories increases, with a specific focus on the one-shot performance. All the experiments show that the proposed method improves the learning performance when useful information is stored in memory, while it never affects it negatively when the known categories are very different from the new one. Figure 1(a) shows the obtained recognition rate results for *Adapt-2W* and the corresponding *LS-SVM-W* when only one object image is used for training and the number of known categories increases from 3 to 20. The performance of the model improves remarkably, showing a one-shot learning behaviour.

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94.

[2] G.C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted ls-svms. In *proceedings IJCNN*, Vancouver, Canada, July 2006.

[3] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, 28.

[4] N.D. Lawrence and J.C. Platt. Learning to learn with the informative vector machine. In *Proceedings of ICML*, 2004.

[5] X. Liao, Y. Xue, and L. Carin. Logistic regression with an auxiliary data source. In *Proceedings of ICML*, 2005.

[6] F. Orabona, C. Castellini, B. Caputo, E. Fiorilla, and G. Sandini. Model adaptation with least-squares svm for hand prosthetics. In *proceedings ICRA*, 2009.

[7] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.

[8] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[9] A. Schwaighofer, V. Tresp, and K. Yu. Learning gaussian process kernels via hierarchical bayes. In *Proceedings of NIPS*, 2005.

[10] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of ICML*, 2004.