Leveraging over prior knowledge for online learning of visual categories

Tatiana Tommasi[12]
ttommasi@idiap.ch

Francesco Orabona[3]
francesco@orabona.com

Mohsen Kaboli[1]
mkaboli@idiap.ch

Barbara Caputo[1]
bcaputo@idiap.ch

[1] Idiap Research Institute,
Martigny, CH

[2] École Polytechnique Fédérale EPFL
Lausanne, CH

[3] Toyota Technological Institute
Chicago, USA

The underlying main goal of all research in visual recognition is to enable vision-based artificial systems to operate autonomously in the real world. However, even the best system we can currently engineer is bound to fail whenever the setting is not heavily constrained. This is because the real world is generally too complicated and too unpredictable to be summarized within a limited set of specifications. This calls for algorithms able to support open ended learning of visual classes which can process continuously new data guided by past experience. The main issues of open ended learning has been typically addressed in a fragmented fashion in the literature. A first component is that of transfer learning, i.e. the ability to leverage over prior knowledge when learning a new class, especially in presence of few training data [3]. A second component is that of updating the learned visual class, as new samples arrive sequentially. The dominant approach in the literature here is that of online learning [1]: predictions are made on the fly and the model is progressively updated at each step, on the basis of the given true label. In this paper we propose to merge together these two components, using prior knowledge sources for initializing the online learning process on a new target task through transfer learning.

We consider binary object-vs-background problems where each image is represented by a vector $\boldsymbol{x} \in \mathbb{R}^d$ associated to a unique label $y \in \{-1, 1\}$ and the prediction mechanism is based on a hyperplane which divides the instance space into two parts. This hyperplane is defined by its orthogonal vector $\boldsymbol{w} \in \mathbb{R}^d$ and the predicted label is given by $\text{sign}(\boldsymbol{w} \cdot \boldsymbol{x})$. We assume without loss of generality that $\|\boldsymbol{x}_t\| \leq 1$ and we define the hinge loss with margin 1 of a classifier $\boldsymbol{w}$ over an instance / label pair $(\boldsymbol{x}, y)$ as $\ell^H(\boldsymbol{w} \cdot \boldsymbol{x}, y) = \max\{0, 1 - y\boldsymbol{w} \cdot \boldsymbol{x}\}$.

We adopt the Passive Aggressive (PA) algorithm [2] as our basic online learning method. A sequence of instances are presented to the learner $\boldsymbol{x}_t$, $t = 1, \ldots, T$ which generates the corresponding prediction and then receives the true label y_t which is used to update its hypothesis for future trials. Starting from an arbitrary hypothesis, $\boldsymbol{w}_1$, at the t−th round PA is updated solving the following optimization problem

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + C\xi \quad \text{s.t.} \quad \ell^H(\boldsymbol{w} \cdot \boldsymbol{x}_t, y_t) \leq \xi \quad \text{and} \quad \xi \geq 0 , \tag{1}$$

where C is the aggressiveness parameter that trades off the two quantities in (1).

Among the existing transfer learning approaches we consider the Multi-KT algorithm [4]. We suppose to have k binary source tasks and a discriminative model learned for each of them in terms of a linear function $h_j(\boldsymbol{x}) = \hat{\boldsymbol{w}}_j \cdot \boldsymbol{x}$ for $j = 1, \ldots, k$. For a novel target task with T available training samples $(\boldsymbol{x}_t, y_t)$ $t = 1, \ldots, T$, Multi-KT solves the following optimization problem [4]:

$$\min_{\boldsymbol{w}, b} \frac{1}{2} \left\| \boldsymbol{w} - \sum_{j=1}^{k} \beta_j \hat{\boldsymbol{w}}_j \right\|^2 + \frac{C}{2} \sum_{t=1}^{T} (y_i - \boldsymbol{w} \cdot \boldsymbol{x}_t - b)^2 . \tag{2}$$

Here the weights β_j assigned to each prior knowledge are found by minimizing $\sum_{t=1}^{T} \ell^H(\tilde{y}_t, y_t)$ subject to $\|\boldsymbol{\beta}\|_2 \leq 1$, where $\tilde{y}_t$ is the leave-one-out prediction for the t−th sample, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$.

Thus we define a learning algorithm based on two phases: at the beginning n target training samples are given as input to Multi-KT which outputs the corresponding target model, and as second step, this model is used to initialize the online learning process. This has several advantages: by using a principled transfer learning process we can study the relation between the old sources and the new target. Within this framework, few
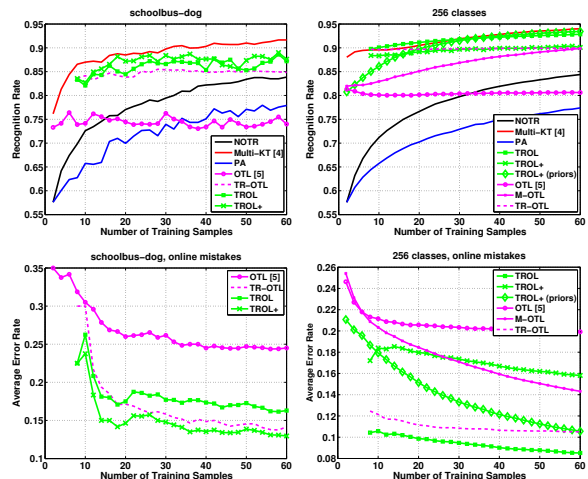

Figure 1: Left column: single source experiments. Right column: 255 object classes are considered as sources, while the remaining one defines the target problem. Top line: recognition rate results on the test set as a function of the number of training samples. Bottom line: corresponding rate of mistakes for the online learning methods.

samples might be sufficient to indicate in which part of the original space the correct solution (the best in term of generalization capacity) should be sought. At the same time, by using the transfer process only at the beginning we limit its computational burden. Then PA guarantees that the updated solution is at each step close to the previous one: this helps keeping the positive effect produced by Multi-KT together with the proper introduction of new information when necessary. We show theoretically that a good initialization for the online learning process produces a tighter mistake bound compared to previous work (OTL [5]), while empirically improving the recognition performance on an unseen test set. We name this algorithm TROL: TRansfer initializes Online Learning and we also consider the possibility to reweight at each step prior and new knowledge defining the variant TROL+.

We ran experiments on the Caltech 256 database selecting related / unrelated object classes and one or multiple prior knowledge sources, beside considering the full dataset (see Figure 1). Over all the experiments TROL and TROL+ present better results than PA, never showing negative transfer, and they are able to match the batch performance of Multi-KT on the test set. In terms of online mistakes, TROL and TROL+ outperform all the other considered baselines.

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

[2] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[3] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[4] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] P. Zhao and S. C. H. Hoi. OTL: A Framework of Online Transfer Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.