Towards a quantitative measure of rareness

Tatiana Tommasi and Barbara Caputo**

Idiap Research Institute
Centre Du Parc, Rue Marconi 19
P.O. Box 592, CH-1920 Martigny, Switzerland
{ttommasi, bcaputo} @idiap.ch

Abstract. Within the context of detection of incongruent events, an often overlooked aspect is how a system should react to the detection. The set of all the possible actions is certainly conditioned by the task at hand, and by the embodiment of the artificial cognitive system under consideration. Still, we argue that a desirable action that does not depend from these factors is to update the internal model and learn the new detected event. This paper proposes a recent transfer learning algorithm as the way to address this issue. A notable feature of the proposed model is its capability to learn from small samples, even a single one. This is very desirable in this context, as we cannot expect to have too many samples to learn from, given the very nature of incongruent events. We also show that one of the internal parameters of the algorithm makes it possible to quantitatively measure incongruence of detected events. Experiments on two different datasets support our claim.

1 Introduction

The capability to recognize, and react to, rare events is one of the key features of biological cognitive systems. In spite of its importance, the topic is little researched. Recently, a new theoretical framework has emerged [7], that defines rareness as an incongruence compared to the prior knowledge of the system. The model has shown to work on several applications, from audio-visual persons identification [7] to detection of incongruent human actions [5].

A still almost completely unexplored aspect of the framework is how to react to the detection of an incongruent event. Of course, this is largely influenced by the task at hand, and by the type of embodiment of the artificial system under consideration: the type of reactions that a camera might have are bound to be different from the type of actions a wheeled robot might take. Still, there is one action that is desirable for every system, regardless of their given task and embodiment: to learn the detected incongruent event, so to be able to recognize it correctly if encountered again in the future.

In this paper we propose a recently presented transfer learning algorithm [6] as a suitable candidate for learning a newly detected incongruent event. Our method is able to learn a new class from few, even one single labeled example by

** This work was supported by the DIRAC project.

exploiting optimally the prior knowledge of the system. This would correspond, in the framework proposed by Weinshall et al, to transfer from the general class that has accepted. Another remarkable feature of our algorithm is that the internal parameter, that controls the amount of transferred knowledge, shows different behaviors depending on how similar the new class is to the already known classes. This suggests that it is possible to derive from this parameter a quantitative measure of incongruence for new detected events. Preliminary experiments on different databases support our claims.

2 Multi Model Transfer Learning

Given k visual categories, we want to learn a new $k+1$ category having just one or few labeled data. We can use only the available samples and train on them, or we can take advantage of what already learned. The Multi model Knowledge Transfer algorithm (Multi-KT) addresses this latter scenario in a binary, discriminative framework based on LS-SVM [6]. In the following we describe briefly the Multi-KT algorithm. The interested reader can find more details in [6].

Suppose to have a binary problem and a set of l samples $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is an input vector describing the i^{th} sample and $y_i \in \mathcal{Y} = \{-1, 1\}$ is its label. We want to learn a linear function $f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$ which assigns the correct label to an unseen test sample $\mathbf{x}$. $\phi(\mathbf{x})$ is used to map the input samples to a high dimensional feature space, induced by a kernel function $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$ [2].

If we call $\mathbf{w'_j}$ the parameter describing the old models of already known classes ($j = 1, \ldots, k$), we can write the LS-SVM optimisation problem slightly changing the regularization term [6]. The idea is to constrain a new model to be close to a weighted combination of pre-trained models:

$$\min_{\mathbf{w},b} \frac{1}{2} \left\| \mathbf{w} - \sum_{j=1}^{k} \beta_j \mathbf{w'_j} \right\|^2 + \frac{C}{2} \sum_{i=1}^{l} \zeta_i (y_i - \mathbf{w} \cdot \phi(\mathbf{x}_i) - b)^2 \ . \tag{1}$$

Here $\boldsymbol{\beta}$ is a vector containing as many elements as the number of prior models k, and has to be chosen in the unitary ball, i.e. $\|\boldsymbol{\beta}\|_2 \leq 1$. Respect to the original LS-SVM, we are also adding the weighting factors ζ_i, they help to balance the contribution of the sets of positive (l^+) and and negative (l^-) examples to the data misfit term:

$$\zeta_i = \begin{cases} \frac{l}{2l^+} \text{ if } y_i = +1 \\ \frac{l}{2l^-} \text{ if } y_i = -1 \ . \end{cases} \tag{2}$$

With this new formulation the optimal solution is

$$\mathbf{w} = \sum_{j=1}^{k} \beta_j \mathbf{w'}_j + \sum_{i=1}^{l} \alpha_i \phi(\mathbf{x}_i) \ . \tag{3}$$

Hence $\mathbf{w}$ is expressed as a sum of the pre-trained models scaled by the parameters β_j, plus the new model built on the incoming training data.

An advantage of the LS-SVM formulation is that it gives the possibility to write the LOO error in closed form [1]. The LOO error is an unbiased estimator of the classifier generalization error and can be used for model selection [1]. A closed form for the LOO error can be easily written even for the modified LS-SVM formulation:

$$r_i^{(-i)} = y_i - \tilde{y}_i = \frac{\alpha_i}{\mathbf{G}_{ii}^{-1}} - \sum_{j=1}^{k} \beta_j \frac{\alpha'_{i(j)}}{\mathbf{G}_{ii}^{-1}}, \tag{4}$$

where $\alpha'_{i(j)} = \mathbf{G}_{(-i)}^{-1}[\hat{y}_1^j, \ldots, \hat{y}_{i-1}^j, \hat{y}_{i+1}^j, \ldots, \hat{y}_l^j, 0]^T$, $\hat{y}_i^j = (\mathbf{w'_j} \cdot \phi(\mathbf{x}_i))$ and $\tilde{y}_i$ are the LOO predictions. The $\mathbf{G}$ matrix is $[\mathbf{K} + \frac{1}{C}\mathbf{W}, \mathbf{1}; \mathbf{1}^T, 0]$, $\mathbf{K}$ is the kernel matrix, $\mathbf{W} = diag\{\zeta_1^{-1}, \zeta_2^{-1}, \ldots, \zeta_l^{-1}\}$, and $\mathbf{G}_{(-i)}$ is obtained when the i^{th} sample is omitted in $\mathbf{G}$.

If we consider as loss function $loss(y_i, \tilde{y}_i) = \zeta_i \max\left[1 - y_i \tilde{y}_i, 0\right]$, to find the best $\boldsymbol{\beta}$ vector we need to minimise the objective function:

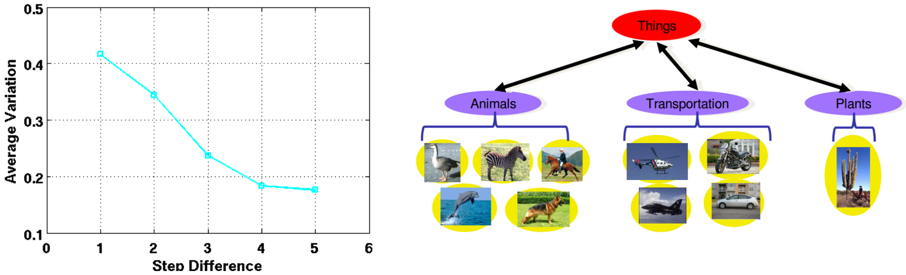
$$J = \sum_{i=1}^{l} \max\left[y_i \zeta_i \left(\frac{\alpha_i}{\mathbf{G}_{ii}^{-1}} - \sum_{j=1}^{k} \beta_j \frac{\alpha'_{i(j)}}{\mathbf{G}_{ii}^{-1}} \right), 0 \right] \qquad \text{s.t.} \quad \|\boldsymbol{\beta}\|_2 \leq 1 \ . \tag{5}$$

3 Stability as a Quantitative Measure of Incongruence

An important property of Multi-KT is its stability. Stability here means that the behaviour of the algorithm does not change much if a point is removed or added. This notion is closely related to the LOO error, which is exactly calculated measuring the performance of the model every time a point is removed. From a practical point of view, this should correspond to a graceful decreasing of the variations in $\boldsymbol{\beta}$ as new samples arrive. This decrease of variations as the training data for the new class arrives should also be related to how difficult it is to learn it. Indeed, if the algorithm does not transfer much, we expect that $\boldsymbol{\beta}$ will stabilize slowly. This corresponds to the situation where the new class is very different from all the classes already learned– in other words, we expect that the stability of $\boldsymbol{\beta}$ is correlated to the rareness of the incoming class.

4 Experiments

This Section presents three set of experiments designed to test our claim that the stability of $\boldsymbol{\beta}$ is related to the rareness of the incoming class. We first show that, as expected, $\boldsymbol{\beta}$ gets stable smoothly when the number of training samples grows (Section 4.1). We then explore how this behavior changes when considering prior knowledge related or unrelated to the new class. This is done first on an easy task (Section 4.2) and then in a more challenging scenario (Section 4.3).



(a) (b)

Fig. 1. (a) Norm of the difference between two $\boldsymbol{\beta}$ vectors correspondent two subsequent step in time. The norms are averaged both on the classes and on the splits; (b) Classes extracted from the Caltech-256 database: goose, zebra, horse, dophin, dog, helicopter, motorbike, fighter-jet, car-side, cactus.

For the experiments reported in Section 4.1 and 4.3 we used subsets of the Caltech-256 database [4] together with the features described in [3], available on the authors' website[1]. For the experiments reported in Section 4.2 we used the audio-visual database and features described in [7] using only the face images. All the experiments are defined as "object vs background" where the background corresponds respectively to the Caltech-256 clutter class and to a synthetically defined non-face, obtained scrumbling the face feature vector elements.

4.1 A Stability Check

As a first step we want to show that the variation in the $\boldsymbol{\beta}$ vector is small when the algorithm is stable. We consider the most general case of prior knowledge consisting of of a mix of related and unrelated categories. We therefore selected ten classes from the Caltech-256 database (see Figure: 1(b)). We run experiments ten times considering in turn one of the classes as the new one and all the other as prior knowledge. We defined 6 steps in time corresponding to a new sample entering the training set. For each couple of subsequent steps we calculated the difference between the obtained $\boldsymbol{\beta}$ vectors. Figure 1(a) shows the average norm of these differences and demonstrates that the algorithm stability does translate in a smooth decrease in the $\boldsymbol{\beta}$ vector of Multi-KT.

4.2 Experiments on Visual Data: Easy Learning Task

In the second set of experiments we dealt with the problem of learning male/female faces when prior knowledge consisted of only female/male faces. A scheme of the two experiments is shown in Figure 2.

For the first experiment, prior knowledge consisted of four women; the task was to learn three new men and three new women. Results are reported in Figure

[1] http://www.vision.ee.ethz.ch/ pgehler/projects/iccv09/

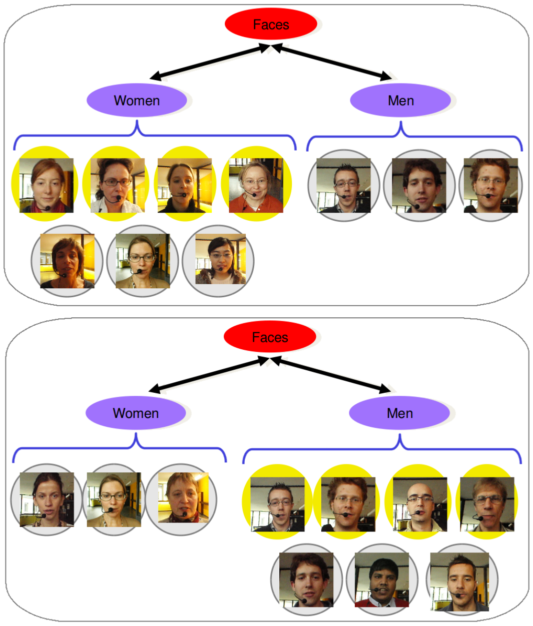


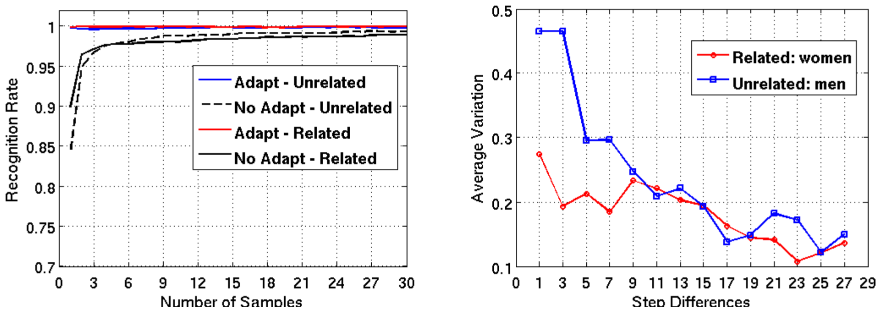
Fig. 2. Top: four women faces used as prior knowledge while three men and three women faces are considered in learning; Bottom: four men faces used as prior knowledge while three men and three women faces are considered in learning.

3(a). The learning curves clearly indicate that the task becomes very easy when using the transfer learning mechanism: we obtain 100 % accuracy even with just one training sample, regardless of the gender. It is interesting to note that the information coming from the female face models is helpful for learning models of male faces. This is understandable, as they all are faces. Nevertheless, the difficulty in relying on faces of the opposite gender is still readable in Figure 3(b) which reports the norm of the differences between two $\boldsymbol{\beta}$ vectors for two subsequent steps in time.

We repeated the experiment using four men as prior knowledge for the task to learn the faces of three new men and three new women. Figure 4(a) show again that there is no significative difference between the two transfer learning curves obtained when learning man and woman faces, and they correspond both to 100 % accuracy. Looking at Figure 4(b) we notice that the $\boldsymbol{\beta}$ vector results more stable when learning a face of the same gender of those contained in the prior knowledge.

4.3 Experiments on Visual Data: Difficult Learning Task

In the third experiment we consider two different scenarios. In the first, we have a set of animals as prior knowledge and the task is to learn a new animal. In the



(a) (b)

Fig. 3. Women as prior knowledge. (a) Classification performance as a function of the number of training images. The results shown correspond to average recognition rate considering each class out experiments repeated ten times. (b) Norm of the difference between two $\boldsymbol{\beta}$ vectors correspondent two subsequent step in time. The norms are averaged both on the classes and on the splits.

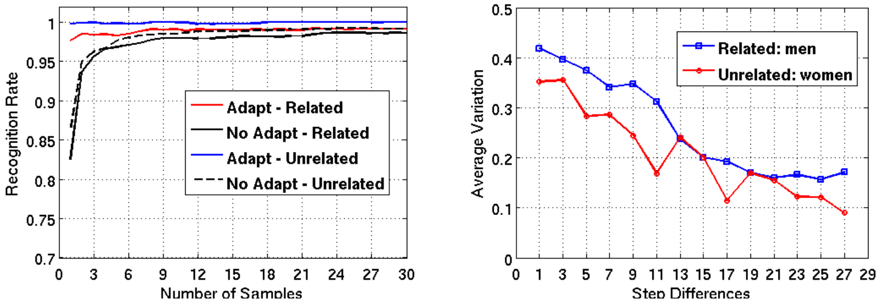
second we have a mix of unrelated categories and the task is to learn a new one. From the point of view of transfer learning we expect the first problem to be easier than the second. Namely, in the first case only 1-2 labeled samples should be necessary, while in the second case the algorithm should need more samples.

To verify this hypothesis we extracted six classes from the Caltech-256 general category "Animal, land" and another group of six was defined picking each class from a different general category (see Figure 5). Two different experiments were run: one with only the animal related classes, considering in turn 5 classes as known and one as new. The second, following the same setting on the six unrelated classes. Even if the two experiments were run separately, the non-transfer learning curve for the problems do not present a significative difference (see Figure 6(a)). This allow us to benchmark the corresponding results for learning with adaptation.

Figure 6(a) shows that when prior knowledge is not informative the algorithm needs more labeled data to learn the new class, demonstrating our initial intuition. In Figure 6(b) the corresponding norm of the differences between two $\boldsymbol{\beta}$ vectors for two subsequent steps in time is reported. We can compare the curves supposing to choose a treshold in the $\boldsymbol{\beta}$ variation: to reach $\Delta\boldsymbol{\beta} < 0.15$ it is necessary to have at least 3 samples when using related prior knowledge and 6 samples for unrelated prior knowledge. For $\Delta\boldsymbol{\beta} < 0.1$, 6 samples are required using related prior knowledge and 12 for unrelated, while to have $\Delta\boldsymbol{\beta} < 0.075$, 10 samples are needed using related prior knowledge and 18 for unrelated.

5 Conclusions

In this paper we addressed the problem of what action an artificial cognitive system can take, upon detection of an incongruent event. We argued that learning



(a) (b)

Fig. 4. Men as prior knowledge. (a) Classification performance as a function of the number of training images. The results shown correspond to average recognition rate considering each class out experiments repeated ten times. (b) Norm of the difference between two $\boldsymbol{\beta}$ vectors correspondent two subsequent step in time. The norms are averaged both on the classes and on the splits.

the new event from few labeled samples is one of the most general and desirable possible actions, as it does not depend on the embodiment of the system, nor its task. We showed how a recently introduced transfer learning algorithm could be used for this purpose, and also how its internal parameter regulating transfer learning could be used for evaluating the degree of incongruence of the new event. Future work will explore further this intuition, with the goal to derive a principled foundation for these results.

References

1. G.C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *IJCNN*, 2006.
2. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
3. P. Gehler and S. Nowozin. Let the kernel figure it out: Principled learning of preprocessing for kernel classifiers. In *Proc. CVPR*, 2009.
4. G. Griffin, A. Holub, and P. Perona. Caltech 256 object category dataset. Technical Report UCB/CSD-04-1366, California Institue of Technology, 2007.
5. Fabian Nater, Helmut Grabner, and Luc van Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. In *CVPR*, 2010.
6. T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proc. CVPR*, 2010.
7. D. Weinshall, H. Hermansky, A. Zweig, J. Luo, H. Brgge Jimison, F. Ohl, and M. Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In *Proc. NIPS*, 2008.

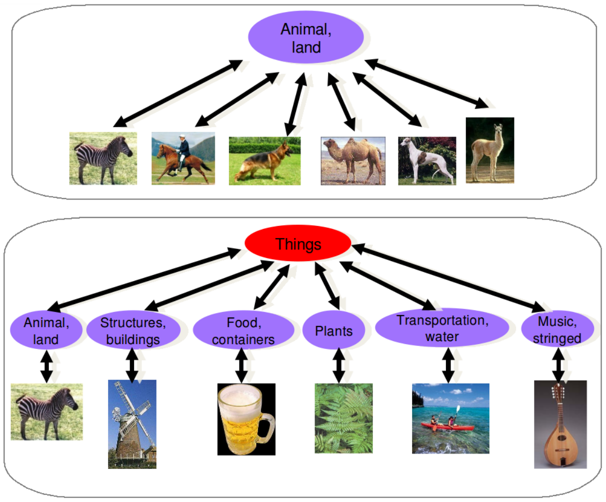
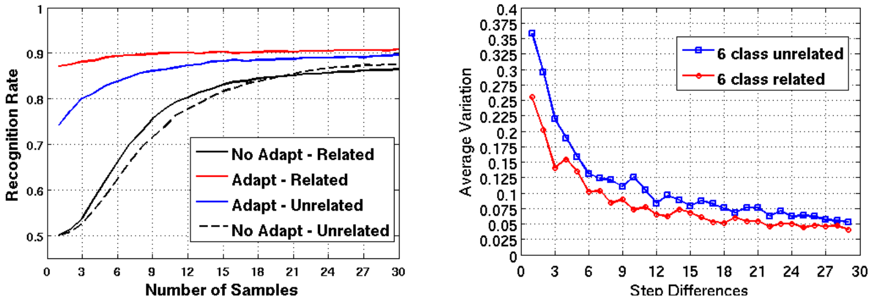


Fig. 5. Top: six classes from the Caltech-256 general category “Animal, land” (zebra, horse, dog, camel, llama, greyhound). Bottom: six classes extracted each form a general category of the Caltech-256 (zebra from “Animal, land”, windmill from “Structures, building”, beermug from “Food,containers”, fern from “Plants”, canoe from “Transportation, water” and mandolin from “Music, stringed”).



(a)

(b)

Fig. 6. (a) Classification performance as a function of the number of training images. The results shown correspond to average recognition rate considering each class out experiments repeated ten times. (b) Norm of the difference between two $\boldsymbol{\beta}$ vectors correspondent two subsequent step in time. The norms are averaged both on the classes and on the splits.