Overview of the ImageCLEF 2014 Domain Adaptation Task

Barbara Caputo[1] and Novi Patricia[2]

[1] University of Rome La Sapienza, Italy
[2] Idiap Research Institute, Switzerland

Abstract. This paper describes the first edition of the Domain Adaptation Task at ImageCLEF 2014. Domain adaptation refers to the challenge of leveraging over knowledge acquired when learning to recognize given classes on a database, when using a different data collection. We describe the scientific motivations behind the task, the research challenge on which the 2014 edition focused, the data and evaluation metric and results obtained by participants. After a discussion on the lesson learned during this first edition, we conclude with possible ideas for future editions of the task.

1 Introduction and Motivation

The amount of freely available and annotated image collections is dramatically increased over the last years, thanks to the diffusion of high-quality cameras, and also to the introduction of new and cheap annotation tools such as Mechanical Turk [3]. Attempts to leverage over and across such large data sources has proved challenging. Indeed, tools like Google GoggleS[3] are able to recognize reliably limited classes of objects, like books or wine labels, but are not able to generalize across generic objects like food items, clothing items and so on. Several authors showed that, for a given task, training on a dataset (e.g. Pascal VOC 07) and testing on another (e.g. ImageNet) produces very poor results, although the set of depicted object categories is the same [10,13,6,12]. In other words, existing object categorization methods do not generalize well across databases.

This problem is known in the literature as the domain adaptation challenge, as known in machine learning for speech and language processing [1,5]. A source domain (S) usually contains a large amount of labeled images, while a target domain (T) refers broadly to a dataset that is assumed to have different characteristics from the source, and few or no labeled samples. Formally, two domains differ when their probability distributions differ: $P_S(x, y) \neq P_T(x, y)$, where $x \in \mathcal{X}$ indicates the generic image sample and $y \in \mathcal{Y}$ the corresponding class label. Within this context, the across dataset generalization problem stems from an intrinsic difference between the underlying distributions of the data.

Addressing this issue would have a tremendous impact on the generality and adaptability of any vision-based annotation system. Current research in domain adaptation focuses on a scenario where

[3] http://www.google.com/mobile/goggles

- (a) the prior domain (source) consists of one or maximum two databases;
- (b) the labels between the source and the target domain are the same, and
- (c) the number of annotated training data for the target domain are limited.

The goal of the Domain Adaptation Task, initiated in 2014 under the ImageCLEF umbrella [4], is to push the state of the art in domain adaptation towards more realistic settings, relaxing these assumptions. Our ambition is to provide, over the years, stimulating problems and challenging data collections that might stimulate and support novel research in the field.

In the rest of the paper we describe the 2014 Domain Adaptation Task (section 2.1), the data and features provided to the participants (section 2.2), and the evaluation metric adopted (section 2.3). Section 3 describes the results obtained while section 4 provides an in depth discussion of the results obtained and identifies possible new directions for the 2015 edition of the task. Conclusions are given in section 5.

2 The 2014 Domain Adaptation Task

In this section we describe the Domain Adaptation Task proposed in the ImageCLEF 2014 lab. We first outline the research challenge we aimed at addressing (section 2.1). Then, we describe the data collection used and the features provided to all participants (section 2.2) and we describe the evaluation metric used (section 2.3).

2.1 The Research Challenge

In the 2014 version (first edition) of the Domain Adaptation Task, we focused on the number of sources available to the system. Current experimental settings, widely used in the community, consider typically one source and one target [10], or at most two sources and one target [6,11]. This scenario is unrealistic: with the wide abundance of annotated resources and data collections that are made available to users, and with the fast progress that is being made in the image annotation community, it is likely that systems will be able to access more and more databases, and therefore to leverage over a much larger number of sources than two, as considered in the most challenging settings today.

To push research towards more realistic scenarios, the 2014 edition of the Domain Adaptation Task has proposed an experimental setup with four sources, where such sources were built by exploiting existing available resources like the ImageNet, Caltetch256 [7] databases and so on. Participants were thus requested to build recognition systems for the target classes by leveraging over such source knowledge. We considered a semi-supervised setting, i.e. a setting where the target data, for each class, is limited but annotated. In the next section we describe in details the data used for the sources, the classes contained both in the source and the target, and the target data provided to participants.

2.2 Data and Features

Source and Target Data To define the source and target data, we considered five publicly available databases:

- the *Caltech-256* database, consisting of 256 object categories, with a total of 30.607 images;
- the *ImageNet ILSVRC2012* database, organized according to the WordNet hierarchy, with an average of 500 images per node;
- the *PASCAL VOC2012* database, an image data set for object class recognition with 20 object classes;
- the *Bing* database, containing all 256 categories from the Caltech-256 one, and augmented with 300 web images per category that were collected through textual search using Bing;
- and the *SUN* database, a scene understanding database that contains 899 categories and 130.519 images.

We then selected twelve classes, common to all the dataset listed above: aereoplane, bike, bird, boat, bottle, bus, car, dog, horse, monitor, motorbike, and people. Figure 1 illustrates the images contained for each class in each of the considered datasets. As sources, we considered 50 images representing the classes listed above from the databases Caltech-256, ImageNet, PASCAL and Bing. The 50 images were randomly selected from all those contained in each of the data collection, for a total of 600 images for each source. As target, we used images taken from the SUN database for each class. We randomly selected 5 images per class for training, and 50 images per class for testing. These data were given to all participants as validation set. The test set consisted of 50 images for each class, for a total of 600, manually collected by us using the class names as textual queries with standard search engines.

Features Instead of making available directly the images to participants, we decided to release pre-computed features only, in order to keep the focus on the learning aspects of the algorithms in this year's competition. Thus, we represented every image with dense SIFT descriptors (PHOW features) at points on a regular grid with spacing 128 pixels [2]. At each grid point the descriptors were computed over four patches with different radii, hence each point was represented by four SIFT descriptors. The dense features have been vector quantized into 256 visual words using k-means clustering on a randomly chosen subset of the Caltech-256 database. Finally, all images were converted to 2×2 spatial histograms over the 256 visual words, resulted in 1024 feature dimension. The software used for computing such features is available at www.vlfeat.org.

2.3 Evaluation Metrics

We asked participants to provide the class name for each of the 600 test images released. Results were compared with the ground truth, and a score was assigned as follows:

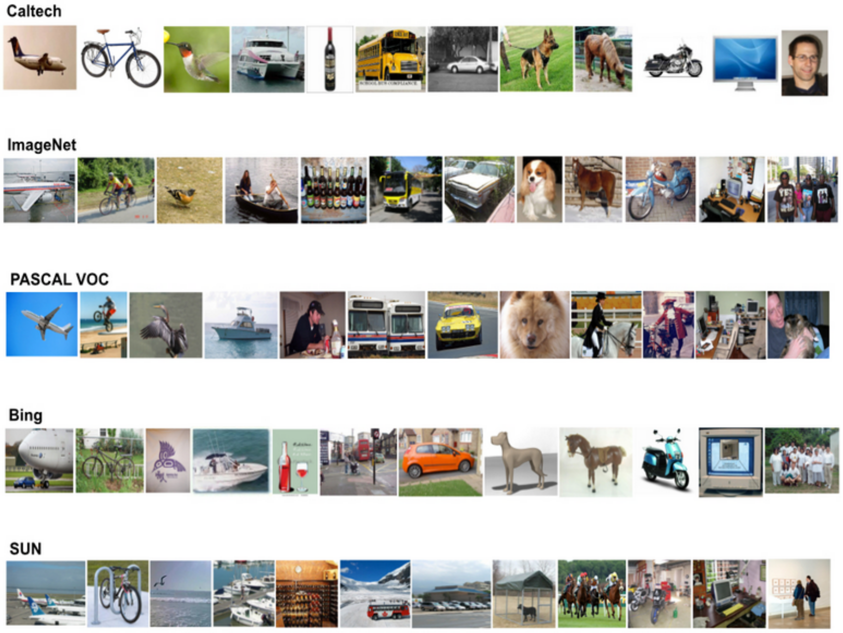

Fig. 1. Exemplar images for the 12 classes from the five selected public databases.

- For each correctly classified image will receive 1 point;
- For each misclassified image will receive 0 point.

We provided to all participants, together with the validation data, a matlab script for evaluating the performance of their algorithms before the official submission, i.e. on the validation data. The script had been tested under Matlab (ver 8.1.0.64) and Octave (ver 3.6.2).

3 Results

While 19 groups registered to the domain adaptation task to receive access to the training and validation data, only 3 groups eventually submitted runs: the XRCE group, the Hubert Curien Lab group and the Idiap group (organizers). They submitted the following algorithms:

- the XRCE group submitted a set of methods based on several heterogeneous methods for domain adaptation, whose predictions were subsequently fused. By combining the output of instance based approaches and metric learning one with a brute force SVM prediction, they obtained a set of heterogeneous classifiers all producing class prediction for the target domain instances. These were combined through different versions of majority voting in order to improve the overall accuracy.

- The Hubert Curien Lab group did not submit any working notes, neither sent any detail about their algorithm. We are therefore not able to describe it.
- The Idiap group submitted a baseline run using a recently introduced learning to learn algorithm [9]. The approach considers source classifiers as experts, and it combines their confidence output with a high-level cue integration scheme, as opposed to a mid-level one as proposed in [8]. The algorithm is called High-level Learning to Learn (H-L2L). As our goal was not to obtain the best possible performance but rather to provide an off the shelf baseline against which to compare results of the other participants, we did not perform any parameter tuning.

Table 1 reports the final ranking among groups. Table 2 reports the results obtained by the best run submitted by each group, for each of the 12 target classes. We see that XRCE obtained the best score, followed by the Hubert Curien lab. The Idiap baseline obtained the worst score, clearly pointing towards the importance of parameter selection in these kind of benchmark evaluations.

Rank	**Group**	**Score**
1	XRCE	228
2	Hubert Curien Lab Group	158
3	Idiap	45

Table 1. Ranking and best score obtained by the three groups that submitted runs.

class	**Score XRCE**	**Score Hubert Curien**	**Score Idiap**
aereoplane	41	36	3
bike	12	7	1
bird	15	15	0
boat	18	5	4
bottle	20	25	3
bus	23	10	6
car	17	13	7
dog	8	8	3
horse	17	6	2
monitor	28	15	3
motorbike	12	7	3
people	17	11	10

Table 2. Class by Class score obtained by the three groups that submitted runs.

4 Analysis and Discussion

The clear success of the XRCE group, obtained by combining several domain adaptation methods presented in the literature, seems to indicate that current methods are not able to address effectively the problem of leveraging over multiple sources. Ensemble methods, chosen by at least two teams, appear instead to be a viable option in this setting, whether used to combine the output of various domain adaptation algorithms, whether used to combine several source output confidences.

The choice made to provide to participants only the features computed from each image, and not the images itself, forced groups to focus on the learning aspects of the problems, but perhaps did not allow for enough flexibility in attacking the problem. We don't plan to repeat this choice in the future editions of the task.

A last remark should be made on the scarce participation to the task. Even though only three groups eventually submitted runs, 19 groups expressed interest and registered, in order to access the training and validation data. We believe that this is an indicator of enough interest to push us to organize again the task next year, also collecting feedbacks from the participating and registered groups in order to identify possible problems in the current edition and to offer a more engaging edition of the task in the future.

5 Conclusions

The first edition of the Domain Adaptation Task, organized under the ImageCLEF umbrella, focused on the problem of building a classifier in a target domain while leveraging over four different sources. nineteen groups registered for the task, and eventually three groups submitted runs, with the XRCE winning the competition with an ensemble learning based method. For the 2015 edition of the task, we plan to make available to participants the raw images, as opposed to pre-computed features as done in 2014, so to allow for a wider generality of approaches. We will continue to propose data supporting the problem of leveraging from multiple sources, possibly by augmenting the number of classes (which was 12 in the 2014 edition), and/or allowing for a partial overlap of classes between sources and between sources and target, as proposed in [12]. In order to significantly increase the number of participants to the task next year, we will contact all groups that registered to the task and ask their preferences among these different options.

Acknowledments

This work was partially supported by the Swiss National Science Foundation project Situated Vision (SIVI).

References

1. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2006)
2. Bosch, Anna ad Zisserman, A.: Image classification using random forests and ferns. In: Proc. CVPR (2007)
3. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6(1), 3–5 (2011)
4. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2014)
5. Daumé III, H.: Frustratingly easy domain adaptation. In: Association for Computational Linguistics Conference (ACL) (2007)
6. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: Proc. CVPR. Extended version considering its additional material
7. Griffin, G., Holub, A., Perona, P.: Caltech 256 object category dataset. Tech. Rep. UCB/CSD-04-1366, California Institue of Technology (2007)
8. Jie, L., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: Proc. ICCV (2011)
9. Patricia, N., Caputo, B.: Learning to learn, from transfer learning to domain adaptation: a unifying perspective. In: Proc. CVPR (2014)
10. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Proc. ECCV (2010)
11. Tommasi, T., Caputo, B.: Frustratingly easy nbnn domain adaptation. In: Proc. ICCV (2013)
12. Tommasi, T., Quadrianto, N., Caputo, B., Lampert, C.: Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In: Proc. ACCV (2012)
13. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Proc. CVPR (2011)