SAMPLING FOR UNSUPERVISED DOMAIN ADAPTIVE OBJECT DETECTION

Fatemeh Mirrashed, Vlad I. Morariu, Larry S. Davis

Department of Computer Science, University of Maryland at College Park

ABSTRACT

We explore the problem of extreme class imbalance present when performing fully unsupervised domain adaptation for object detection. The main challenge arises from the fact that images in unconstrained settings are mostly occupied by the background (negative class). Therefore, random sampling will not typically result in a sufficient number of positive samples from the target domain, which is required by domain adaptation methods. Motivated by traditional semi-supervised learning algorithms that aim for better classification using both labeled and unlabeled data, we propose a variation of co-learning technique that automatically constructs a more balanced set of samples from the target domain. We evaluate the effectiveness of our approach using a vehicle detection task in an urban surveillance dataset. Furthermore, we compare the performance of our technique with two other approaches–one based on unbiased learning on multiple training data sets and the other on self-learning.

Index Terms— Object Detection, Domain Adaptation, Semi-supervised Learning

1. INTRODUCTION

Discriminative learning algorithms for classification perform well when training and test data are drawn from the same distribution. Often, however, we have sufficient labeled training data from single or multiple source domains but wish to learn a classifier which performs well on a target domain with a different distribution and no labeled training data. In object detection, for example, where the goal is to determine the position and size of all of the objects within one category appearing in a given image, it may be infeasible to collect training data to model the enormous variety of possible combinations of pose, background, resolution, and lighting conditions affecting object appearance. Thus, in realistic applications, we expect to encounter domains at test time for which we have seen little or no training data.

For this reason, domain adaptation techniques have gained considerable attention in computer vision applications with some promising results [1, 2, 3, 4, 5]. Previous works have addressed the case in which a few positive and negative examples, with or without their labels, are available from the target domain. Even in the case where labels are not provided for target samples [4], some weak information about class labels is still used in adapting to the new domain simply because the number of positive and negative samples are roughly of the same order, i.e, the two classes are balanced. Here, in contrast, we focus on the extreme case where no samples from the new domain are given and so they have to be obtained automatically for a *fully unsupervised* adaptation approach.[1]

Unfortunately, fully unsupervised domain adaptation for object detection is a chicken and egg problem: in order to best adapt to the target domain, class labels are needed to balance the set of positive and negative samples; but, class labels can only be obtained automatically with a model that works sufficiently well on the target domain. Here, the main challenge arises from the fact that there are only a limited number of object instances but almost an infinite number of samples of the background class, since any portion of an image that does not contain the object of interest is considered an example of the background class. If training samples were obtained by randomly sampling the image, the positive and negative samples would be highly unbalanced and the model would only adapt to the appearance of the more pervasive background class.

To address this problem, Mirrashed et al. [5] proposed an approach for bootstrapping the target domain with the source trained detector, assuming that the source and target domain are sufficiently close to obtain a more balanced training set from the target domain. However their work is limited to adaptation from only a single source domain. In real world applications, especially with the ever-increasing numbers of public data sets, it would be desirable to make use of classifiers pre-trained on multiple independent datasets (regarding each dataset as a source domain) when adapting to a new domain. Our target application is the scenario of vehicle detection in urban surveillance videos, where videos are collected from cameras in multiple locations and each camera is treated as a separate domain, representing varying viewpoint, illumination conditions (e.g. sunlight, shadows, reflections, rain, snow) and traffic patterns. Our goal, then, is to leverage all available domains (camera locations) for which we have labeled data to adapt to a new domain without labeled data.

Traditionally, semi-supervised learning methods are employed to address problems where both labeled and unlabeled

[1] An adaptive approach where the samples from the target domain are obtained automatically and used without their class labels.

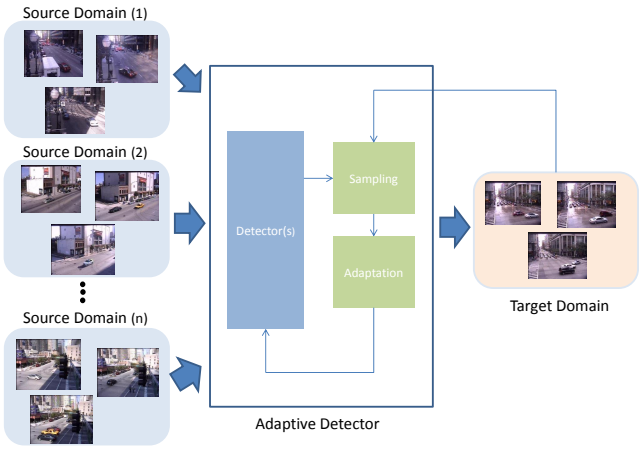

Fig. 1. Our framework for fully unsupervised adaptive object detection. With detectors trained on training data from multiple source domains, we bootstrap the target domain for positive and negative samples. We then retrain the detectors with training data adapted to samples from the target domain. And the whole process is reiterated.

data are used to yield improved classification [6]. The authors in [7, 8] proposed a co-training strategy, which unlike the original work of Blum et. al in [9], does not assume independence and redundancy in the feature space. Instead, an ensemble of learners with different inductive biases (e.g. decision trees, naive Bayes, SVMs, etc.) are trained separately on the same labeled data set. They then make predictions on the unlabeled data. If a majority of learners confidently agree on the class of an unlabeled sample, that sample with its predicted label is added to the training data. All learners are retrained on the updated training set. The final prediction is made with a variant of a weighted majority vote among all the learners.

While the intuition behind this procedure is similar to our problem of sampling for adaptive object detection, there are three main complications in our setting compared to that of [7, 8] and semi-supervised approaches in general. First, it is assumed that both labeled and unlabeled data are sampled from an identical underlying distribution which does not hold for our problem setting; second, we have different sets of labeled data, from multiple source domains; and third, we do not necessarily have different types of classifiers with different inductive biases. To address the first problem we use a domain adaption technique, such as TCA [10], to project all the training data into a common space before re-training the classifiers. To address the second and third situations, we show that training the same learning algorithm, in particular linear SVM's which are the most common classifiers used in computer vision, over different domains or data sets with generally different biases [11, 12] will result in the different inductive biases needed for successful co-training. So in fact, in our case, having labeled data from different source domains with generally different biases substitutes for having different classifiers with different inductive biases.

Another important difference between our adaptive co-learning method and the original version [9] or other variations of co-learning algorithm [7, 8] is that we do not use the new labeled data in retraining the classifiers. Instead, we ignore their labels and use the new samples from the target domain as a representative of the data distribution in the feature space for that domain. We then iteratively learn a common latent subspace (using TCA) underlying source and target domains in which we train and run our classifiers. Therefore, since we do not explicitly require the labels of samples from the target domain, the labeling noise in the machine-labeled predictions will not be detrimental as it would be to noise sensitive supervised learning algorithms.

We compare our approach to two baseline methods: (1) an adaptive approach based on a recently proposed discriminative framework [13] that explicitly estimates a bias for each source domain and approximates an unbiased classifier for an unseen target domain (referred to as *visual world*); and, (2) the adaptive approach in [5] where a single classifier trained on all the labeled data from multiple source domains is used to bootstrap the detection in target domain for adaptation.

The remainder of the paper is organized as follows. We describe the two alternative methods in sections 2.1 and 2.2 and explain our proposed approach in section 3. A detailed description of the experiments and results are given in section4, followed by a conclusion in section 5.

2. BASELINE APPROACHES

The following sections describe the two baseline methods that we employed to address the problem of automatic sampling from the target domain in a fully unsupervised domain adaptive object detection framework. We analyze a setting in which we have plentiful labeled training data drawn either from multiple source domains with uncontrolled various distributions or data sets with naturally different biases [11, 12] but no labeled training data is available from the target domain of interest.

2.1. Unbiased Learning

Similar to our proposed method in 3, this approach relies on the assumption that the bias between datasets (or domains) can be identified in the feature space, i.e. the features used to describe the images are rich enough to capture the bias in the data distribution of a domain. With that assumption, Khosla et al in [13] present an algorithm, which is largely based on max-margin learning (SVM), to explicitly model the bias vector in the feature space for each training dataset. Based on the

	D_{t1}	D_{t2}	D_{t3}	D_{t4}	D_{t5}	D_{t6}	D_{t7}	D_{t8}	D_{t9}	D_{t10}	Average
Single source (no adaptation)	.48	.27	.21	.28	.35	.22	.50	.21	.40	.06	.30
Multi source (no adaptation)	.65	.48	**.31**	.27	.51	.15	.67	.35	.36	.01	.37
Unbiased learning (W_{vw})	.57	.38	.20	.17	.50	.13	.57	.26	.31	.01	.31
Unbiased learning ($W_{vw} + \Delta_{tar}$)	.68	.51	.15	.20	.47	.19	.64	.70	.47	**.22**	.42
Adaptive self-learning	.74	.58	.22	.37	.64	**.24**	**.76**	.70	.54	.19	.50
Adaptive co-learning	**.76**	**.69**	.20	**.46**	**.65**	.22	**.76**	**.77**	**.55**	.21	**.53**

Table 1. Vehicle detection results. The detector was trained on labeled data from one of the 4 groups of 3 source domains and tested on one of the 10 target domains. The numbers are average precision. The performance for each target domain is averaged over 4 possible scenarios resulting from the 4 different multi-source domain groups

observation that different image datasets are biased samples of a more general dataset (the visual world), they model the weight vector (W_i) learned for a specific dataset (d_i) as a linear additive function of the corresponding bias term (Δ_i) and the weight vector for the visual world (W_{wv}).

For our adaptive detection framework, we employ this method in an iterative mode. In the first iteration, using labeled data from multiple source domain, we learn W_{wv} and bootstrap the (unseen) target domain to obtain high confidence positive and negative samples. Then considering those samples along with their predicted labels as a new "source" domain, we learn the new bias vector for that domain, Δ_{tar}, in the second and following iterations.

2.2. Adaptive Self-learning

The most common method of semi-supervised learning is self-learning (also known as self-training or bootstrapping) [6], in which a given model predicts the classes for the unlabeled portion of the data. The automatically labeled examples are then added to the training set, the model is retrained, and the whole process is iterated.

However, in our setting the labeled and unlabeled data are not drawn from the same probability distribution. So similar to [5], we use TCA[10], as a domain adaptation technique to adapt the learned model on training data to the target distribution of interest. In other words, unlike semi-supervised learning algorithm, we do not use the new self-predicted labels in retraining the classifier. We instead use these new predictions as a sample of the data distribution in the target domain to learn a common latent subspace for both the source and target domains. Since TCA does not use class labels for training, the labeling noise of self-learning will not be detrimental (note that most machine learning approaches perform poorly in the presence of labeling noise). Also to prevent a classification mistake from reinforcing itself over iterations, only the most and least confident predictions by the baseline detectors are used as positive and negative samples for the target domain.

3. PROPOSED METHOD: ADAPTIVE CO-LEARNING

In this approach, instead of training a single classifier on all the labeled data from all the source domains, we train a classifier separately on the training dataset from each of the source domains. Then, each of the classifiers make predictions separately on the data from the target domain. The final prediction is made with a variant of a weighted majority vote among all the classifiers. Using a vote weighted by a measure of confidence eliminates the possibility that a majority of learners make the same wrong predictions each with very low confidence.

The rest of the process, including adaptation and relearning of the classifiers, is the same as in 2.2. Once again, the most and least confident predictions by these baseline detectors are used correspondingly as positive and negative samples of the target domain for TCA training. Then all the baseline classifier are re-trained and tested on the target domain in the common latent subspace learned by TCA and the whole process is iterated. To compare the confidence values of predictions between these different classifiers, we use the score calibration method described in [14].

The idea behind this strategy is that since multiple classifiers are trained on different training datasets with different biases, they learn diverse models with different inductive biases that can complement each other. The hope is that using the consensus between these hypotheses with different biases would result in more accurate predictions on the target domain with the unknown and possibly different bias.

4. EXPERIMENTS AND RESULTS

We evaluated our proposed methods and baseline approaches for vehicle detection on the dataset used in [5]. This dataset consists of videos from 50 different traffic surveillance cameras, located in a large North American city. From each camera viewpoint, frames were collected at different times of the day and contain large variations in illumination due to the changes in the direction of sunlight and the resulting reflections and shadows from buildings. Apart from the viewpoint,

which changes significantly across the cameras, the amount and type of traffic also varies. On average each test image contains one to three vehicles. We chose a subset of 22 domains, 12 as source and 10 as target domains, so that there is no overlap between the location of the cameras or the intersection that are looked at between the source and target domains. There are at least 100 annotated frames within each target domain. Dividing the 12 source domains into 4 groups of 3 multi-source domains, we perform experiments on 4×10 possible testing scenarios.

We used HOG features (as implemented by [15]) with a dimension of 55,648 to represent detection windows. We used a sliding-window detector where an SVM with linear kernel (as implemented by *LibLinear* [16]) is applied as the binary classifier to each window location at multiple scales. Classifiers were trained on a a fixed number of 300 positive and 300 negative samples from each of the source domains. Samples were drawn randomly from all source domains and the performance was averaged over 10 iterations. The number of positive samples automatically sampled from target domains for adaptation was set to 50 for all the methods. Following the result of pilot experiments in [5], the dimension of the subspace in TCA was set to 15 for all the experiments. Performance is measured by average precision, the area under the precision-recall curve.

Table 1 summarizes the results of our experiments for each target domain and for each of the baseline and proposed methods. The reported average precision per each target domain is averaged over the 4 testing scenarios with 4 sets of multi-source domains. The last column in the table shows the overall performance averaged over all the target domains.

The first and second rows in table 1 show the results of baseline detectors with no adaptation, where the training labeled data comes from either only one out of three source domains (single source) or is accumulated from all three source domains (multi-source). While on average more training data can result in slightly better classification, in 4 out of 10 cases (Dt4, Dt6, Dt9 , and Dt10) that assumption does not hold, likely due to the degree of difference between the training sample distributions.

The last two rows show the results for adaptive self-learning and adaptive co-learning detection, where the performance is increased respectively 35% and 43% over the multi-source baseline method with no adaptation. While our proposed method of adaptive co-learning outperforms adaptive self-learning in the majority of cases, the simplicity and lower computational cost of self-learning can still make it an attractive and competitive choice for time-sensitive applications.

The results of unbiased learning are reflected in rows 3 and 4. Row 3 shows the scenario where an unbiased weight vector (W_{wv}) learned in the first iteration is used for detection on the unseen target domain,as suggested by [13]. However in our case, it significantly underperforms the multi-source baseline detector (row 2) over all of the 10 test domains. Row 4 shows the results of our extension to the algorithm where a bias term specific to samples obtained from the target domain in the first iteration is learned and used for classification at the second iteration ($W_{vw} + \Delta_{tar}$). This time, while the performance increases with respect to the baseline detectors (first iteration), it still falls short compared to the other two methods of adaptive self-learning and co-learning.

The sampling and re-training iteration was repeated five times for each algorithm. On average the performances changed only by $.9\%$ across all methods from iteration 1 to iteration 5. Consequently, the reported results in table 1 indicate those from the first iteration for all methods.

5. CONCLUSION

We have presented and evaluated an approach for fully unsupervised domain adaptive vehicle detection from multiple source domains in traffic surveillance videos and showed its superior performance compared to some alternative methods. Although we have demonstrated the effectiveness of our approach on the task of vehicle detection, it can be potentially applied to other classes of object. The generality of our proposed adaptive object detection framework also extends to employment of any domain adaptation technique or supervised learning algorithm.

6. REFERENCES

[1] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010.

[2] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[3] V. Jain and E. Learned-Miller, "Online domain-adaptation of a pre-trained cascade of classifiers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[4] R. Gopalan ad R. Li and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *IEEE International Conference on Computer Vision*, 2011.

[5] F. Mirrashed, V. Morariu, B. Siddiquie, R. Feris, and L. Davis, "Domain adaptive object detection," in *IEEE Workshop on the Applications of Computer Vision*, 2013.

[6] Xiaojin Zhu, "Semi-Supervised Learning Literature Survey," Tech. Rep., Computer Sciences, University of Wisconsin-Madison, 2005.

[7] Sally Goldman and Yan Zhou, "Enhancing supervised learning with unlabeled data," in *International Conference on Machine Learning*, 2000.

[8] Yan Zhou and Sally A. Goldman, "Democratic Co-Learning," in *International Conference on Tools with Artificial Intelligence*, 2004.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Workshop on Computational Learning Theory*, 1998.

[10] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, 2011.

[11] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[12] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman, "Dataset issues in object recognition," in *Toward Category-Level Object Recognition, volume 4170 of LNCS*, 2006.

[13] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba, "Undoing the damage of dataset bias," in *European Conference on Computer Vision*, 2012.

[14] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2011.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[16] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, 2008.