Heterogeneous Domain Adaptation: Learning Visual Classifiers from Textual Description

Mohamed Elhoseiny   Babak Saleh   Ahmed Elgammal
Department of Computer Science, Rutgers University, New Brunswick, NJ
[m.elhoseiny,babaks,elgammal]@cs.rutgers.edu

Introduction

One of the main challenges for scaling up object recognition systems is the lack of annotated images for real-world categories. It is estimated that humans can recognize and discriminate among about 30,000 categories [4]. Typically there are few images available for training classifiers form most of these categories. This is reflected in the number of images per category available for training in most object categorization datasets, which, as pointed out in [12], shows a Zipf distribution.

The problem of lack of training images becomes even more severe when we target recognition problems within a general category, i.e., subordinate categorization, for example building classifiers for different bird species or flower types (estimated over 10000 living bird species, similar for flowers).

In contrast to the lack of reasonable size training sets for large number of real world categories, there are abundant of textual descriptions of these categories. This comes in the form of dictionary entries, encyclopedia entries, and various online resources. For example, it is possible to find several good descriptions of ”Bobolink” in encyclopedias of birds, while there are only few images available for it online.

The main question we address in this work is how to use purely textual description of categories with no training images to learn a visual classifiers for these categories. In other words, we aim at zero-shot learning of object categories where the description of unseen categories comes in the form of typical text such as an encyclopedia entry; see Fig 1. This is a domain adaptation problem between heterogeneous domain (textual and visual). We explicitly address the question of how to automatically decide which information to transfer between classes without the need of any human intervention. In contrast to most related work, we go beyond simple use of tags and image captions, and apply standard Natural Language Processing techniques to typical text to learn visual classifiers.

Similar to the setting of zero-shot learning, we use classes with training data (“seen classes) to predict classifiers for classes with no training data (“unseen classes). Recent works on zero-shot learning of object categories focused on leveraging knowledge about common attributes and shared parts [9, 7]. Typically, attributes are manually defined by humans and are used to transfer knowledge between seen and unseen classes. In contrast, in our work, we do not use any explicit attributes. The description of a new category is purely textual, and the process is totally automatic without human annotation beyond the category labels; see Fig. 1.

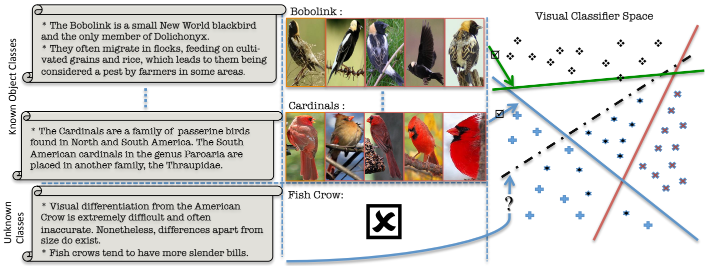
Known Object Classes
* The Bobolink is a small New World blackbird and the only member of Dolichonyx.
* They often migrate in flocks, feeding on cultivated grains and rice, which leads to them being considered a pest by farmers in some areas.
* The Cardinals are a family of passerine birds found in North and South America. The South American cardinals in the genus Paroaria are placed in another family, the Thraupidae.
Unknown Classes
* Visual differentiation from the American Crow is extremely difficult and often inaccurate. Nonetheless, differences apart from size do exist.
* Fish crows tend to have more slender bills.
Bobolink :
Cardinals :
Fish Crow:
Visual Classifier Space
?

Figure 1: Problem Definition: Zero-shot learning with textual description. Left: synopsis of textual descriptions for bird classes. Middle: images for “seen classes”. Right: classifier hyperplanes in the feature space. The goal is to estimate a new classifier parameter given only a textual description [6]

Our work could be seen as another trend for zero shot learning where attribute annotation is not required for each image in Attribute Models. While Attribute Models deals with the dilemma of finding best set of visual attributes for object description [1], our trend could get an unstructured text description of an unseen object category from the wikipedia or the web.

We present an on-going research on the task of learning visual classifiers from purely textual description with zero or very few visual examples. To the best of our knowledge, this problem is not explored in the compute vision community. In [6], we investigated this new problem, we proposed two baseline formulations based on regression and domain adaptation. Then we proposed a new constrained optimization formulation that combines a regression function and a knowledge transfer function with additional constraints to solve the problem. *In this document, we present and compare different formulations that we have investigated for the problem, with results surpassing that which were reported in [6]. We also report other interesting formulations that was not studied there* [2] Quantitative and qualitative evaluation of these different formulations gives insight about the problem of domain adaptation between such heterogeneous domains (textual and visual).

[1] which is an art and requires human design

[2] We recommend reading [6] before reading this document since it includes more details about the motivation, the related work and the settings

Problem Definition: We denote the visual domain by $\mathcal{V}$ and the textual domain by $\mathcal{T}$. We are given training data in the form $V = \{(x_i, l_i)\}_N$, where x_i is an image and $l_i \in \{1 \cdots N_{sc}\}$ is its class label among N_{sc} training classes. Depending on the domain we might find a few, a couple, or as little as one narrative for each class. We denote the textual training data for class k by $\{t_i \in \mathcal{T}\}^k$. In this study, we assumed we are dealing with the extreme case of having only one narrative per class, which makes the problem even more challenging. Let us denote the visual feature vector for a given image, amended with 1, by $\mathbf{x} \in \mathbb{R}^{d_v+1}$. Let's denote the extracted textual features by $T = \{\mathbf{t}_k \in \mathbb{R}^{d_t}\}_{k=1 \cdots N_{sc}}$. We are given textual description $\mathbf{t}_*$ for a new class, and the goal is to predict a classifier for that class with no training images.

Formulations: We investigated several formulations for predicting classifier parameters $c(\mathbf{t}_*)$ for a new class with textual description $\mathbf{t}_*$, in the form of a linear one-vs-all classifier, *i.e.*, $c(\mathbf{t}_*)^\mathsf{T} \cdot \mathbf{x} > 0$ if $\mathbf{x}$ belongs to $\mathcal{C}$ and $c(\mathbf{t}_*)^\mathsf{T} \cdot \mathbf{x} < 0$ otherwise. Recall that the features are amended with 1 such that $c(\mathbf{t}_*)$ is hyperplane paramterization. Table 1 shows results of different formulations, which will be details next.

Regression Models: Given classifiers learned on seen classes, with hyperplanes $\{\mathbf{c}_k\}$, and the corresponding textual features $\{\mathbf{t}_k\}$, we can learn a regression function $c(\cdot) : \mathbb{R}^{d_t} \to \mathbb{R}^{d_v}$ from $\mathcal{T}$ to $\mathcal{V}$. We investigated Gaussian Process Regression (GPR) [11] and structured regression using the Twin Gaussian Processes (TGP) [5]. However such regression model will only learn the correlation between the textual and visual domain through the information available in its input-output pairs, *i.e.* $(\mathbf{t}_k, \mathbf{c}_k)$. Here the visual domain information is encapsulated in the pre-learned classifiers, and prediction does not have access to the original data in the visual domain. Another problem, is the sparsity of the data; the number of training classes is typically much less than the dimension of the visual and textual feature spaces. We also investigated formulations that use regression to predict an initial hyperplane $\tilde{c}(\mathbf{t}_*)$, which is then optimized to put all seen data in one side, *i.e.*

$$\hat{c}(\mathbf{t}_*) = \underset{\mathbf{c},\zeta_i}{\operatorname{argmin}}[\mathbf{c}^\mathsf{T}\mathbf{c} + \alpha\, \phi(\mathbf{c}, \tilde{c}(\mathbf{t}_*)) + \beta \sum_{i=1}^{N} \zeta_i]$$
$$s.t. : -\mathbf{c}^\mathsf{T}\mathbf{x}_i \geq \zeta_i,\ \zeta_i \geq 0, i = 1, \cdots, N$$

where $\phi(\cdot,\cdot)$ is a similarity function between hyperplanes, *e.g.* a dot product, or other functions incorporating the predictive variance. We call this class of methods *constrained GPR/TGP* in Table 1.

Domain Adaptation (DA) Models: Another formulation is to pose the problem as domain adaptation from the textual to the visual domain. In particular, in [8] an approach for learning cross domain transformation was introduced, by learning a regularized asymmetric transformation between points in two domains. The approach was applied to transfer learned categories between different visual domains. A particular attractive characteristic of [8] is that the source and target domains do not have to share the same feature spaces or dimensionality. Inspired by [8], we adapt a model that learns a linear transfer function $\mathbf{W}$ between $\mathcal{T}$ and $\mathcal{V}$. The matrix $\mathbf{W}$ can be learned by optimizing, with a suitable regularizer, over constraints of the form $\mathbf{t}^\mathsf{T}\mathbf{W}\mathbf{x} \geq l$ if $\mathbf{t} \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{V}$ belong to the same class, and $\mathbf{t}^\mathsf{T}\mathbf{W}\mathbf{x} \leq u$ otherwise. Here l and u are model parameters. This transfer function acts as a compatibility function between the textual and visual features. Given a textual feature $\mathbf{t}_*$ and a test image, represented by feature vector $\mathbf{x}$, a classification decision can be obtained by $\mathbf{t}_*^\mathsf{T}\mathbf{W}\mathbf{x} \gtrless b$ where b is a decision boundary which can be set to $(l+u)/2$. Therefore, $c(\mathbf{t}_*) = \mathbf{t}_*^\mathsf{T}\mathbf{W}$ is the desired predicted classifier. There is no guarantee that such a classifier will put all the seen data on one side and the new unseen class on the other side of that hyperplane.

Regression+Domain Adaptation: Each of the regression and domain adaptation models captures partial information about the problem. Therefore, we investigated several objective functions that combines a learned domain correlation matrix $\mathbf{W}$ and a structure predictor to generate a classifier predictor. The new classifier has to be consistent with the seen classes and put all the seen instances at one side of the hyperplane. It has also to be consistent with the learned domain transfer function. This leads to the following constrained optimization problem

$$\hat{c}(\mathbf{t}_*) = \underset{\mathbf{c},\zeta_i}{\operatorname{argmin}}[\mathbf{c}^\mathsf{T}\mathbf{c} - \alpha\mathbf{t}_*{}^\mathsf{T}\mathbf{W}\mathbf{c} - \beta \ln p_{reg}(\mathbf{c}|\mathbf{t}_*) + \gamma \sum_{i=1}^{N} \zeta_i]$$
$$s.t. : -\mathbf{c}^\mathsf{T}\mathbf{x}_i \geq \zeta_i,\ \zeta_i \geq 0,\ \mathbf{t}_*{}^\mathsf{T}\mathbf{W}\mathbf{c} \geq l \qquad (1)$$

The first term is a regularizer over the classifier $\mathbf{c}$. The second term enforces that the predicted classifier has high correlation with $\mathbf{t}_*^\mathsf{T}\mathbf{W}$. The third term favors a classifier that aligns with the prediction of the regressor $\tilde{c}(\mathbf{t}_*)$. The constraints $\mathbf{c}^\mathsf{T}\mathbf{x}_i \geq \zeta_i$ enforce that all seen data instances are at the negative side of the predicted hyperplane with some missclassification allowed through the slack variables ζ_i. The constraint $\mathbf{t}_*{}^\mathsf{T}\mathbf{W}\mathbf{c} \geq l$ enforces that the correlation between the predicted classifier and $\mathbf{t}_*{}^\mathsf{T}\mathbf{W}$ is no less than l, *i.e.* a minimum correlation between the text and visual features. Given $\mathbf{W}$, and the form of the probability estimate $p_{reg}(\mathbf{c}|\mathbf{t}_*)$, the optimization reduces to a quadratic program on $\mathbf{c}$ with linear constraints.

Constrained-DA We also investigated constrained-DA formulations that learns a transfer matrix $\mathbf{W}$ and enforce $\mathbf{t}_k^\mathsf{T}\mathbf{W}$ to be close to the classifiers learned on seen data, $\{\mathbf{c}_k\}$,*i.e.*

$$\min_{\mathbf{W}} r(\mathbf{W}) + \lambda_1 \sum_i c_i(\mathbf{T}\mathbf{W}\mathbf{X}^\mathsf{T}) + \lambda_2 \sum_k (\mathbf{c}_k - \mathbf{t}_k^\mathsf{T}\mathbf{W})^T(\mathbf{c}_k - \mathbf{t}_k^\mathsf{T}\mathbf{W})$$

A classifier can be obtained by optimizing an objective similar to Eq 1 without the regression term.

Datasets and Features: We performed experiments on two datasets: The Oxford Flower (102 classes) [10] and the CUB-UCSD Bird (200 classes) [15]. We generated narrative for each of the datasets using Wikipedia, Plant DataBase [2], Plant Encyclopedia [3], and BBC articles [1]. For this preliminary study, we represented the textual description using tf-idf (Term Frequency-Inverse Document Frequency) [13], then the Clustered Latent Semantic Indexing (CLSI) [16, 17] was used to reduce the dimensionality. We represented each image using the Classeme [14]

Table 1: Comparative Evaluation of Different Formulations on the Flower and Bird Datasets

Approach	Oxford Flowers Avg AUC (+/- std)	UC-UCSD Birds Avg AUC (+/- std)
Regression - GPR [6]	0.54 (+/- 0.02)	0.52 (+/- 0.001)
Structured Regression - TGP [6]	0.58 (+/- 0.02)	0.61 (+/- 0.02)
Constrained GPR	0.621(+/- 0.005)	-
Constrained TGP	0.629(+/- 0.007)	-
Domain Adaptation [6]	0.62(+/- 0.03)	0.59 (+/- 0.01)
Constrained Domain Adaptation (CDA)	0.638 (+/- 0.006)	-
Regression+DA + constraints [6]	0.68 (+/- 0.01)	0.62 (+/- 0.02)

Top-5 Classes with highest combined improvement

class	TGP (AUC)	DA (AUC)	TGP+DA+C	% Improv.
2	0.51	0.55	0.83	57%
28	0.52	0.54	0.76	43.5%
26	0.54	0.53	0.76	41.7%
81	0.52	0.82	0.87	37%
37	0.72	0.53	0.83	35.7 %

features, which encodes each image by the response of 2569 weak classifiers, trained independently on various categories. We used Classeme features since they offer a representation that is closer to the semantic level. We used the same classifiers provided by [14], which are trained and optimized on a term list and images independent of our fine-grained datasets.

Experiments and Conclusions: We computed the ROC curves and report the area under that curve (AUC) as a comparative measure[3] Five-fold cross validation over the classes were performed, within each of these class-folds, the data of the seen classes are further split into training and test sets. Table 1 shows the average AUCs for different formulations. The results prove our hypothesis. Even though the visual features and textual features were independently extracted, by learning correlation between them, we can predict classifiers for new categories. GPR performed poorly, while, as expected, TGP performed better. Adding constraints to GPR/TGP improved their performance. Combining regression and DA gave significantly better results for classes where both approaches individually perform poorly, as can be seen in Table 1-right. We performed an additional experiment, where **W** is firstly computed using Constrained Domain adaptation (CDA). Then, the unseen classifier is predicted using equation 1 . The average AUC of this experiment is 0.64 on Birds dataset which is $\approx 2\%$ better than the quadratic program in [6].

References

[1] Bbc science articles. http://www.bbc.co.uk/science/0/. 2

[2] Plant database. http://plants.usda.gov/java/. 2

[3] Plant encyclopedia. http://www.theplantencyclopedia.org /wiki/Main_Page. 2

[4] Irving Biederman et al. Recognition-by-components: A theory of human image understanding. *Psychological review*, 1987. 1

[5] Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *Int. J. Comput. Vision*, 87(1-2):28–52, March 2010. 2

[6] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero shot learning using purely textual descriptions. In *ICCV*, 2013. 1, 3

[7] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1

[8] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[9] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In *CVPR*, 2009. 1

[10] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 2

[11] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. 2

[12] Ruslan Salakhutdinov, Antonio Torralba, and Joshua B. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 1

[13] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. 2

[14] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, 2010. 2, 3

[15] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 2

[16] Dimitrios Zeimpekis and Efstratios Gallopoulos. Clsi: A flexible approximation scheme from clustered term-document matrices. In *In SDM*, 2005. 2

[17] Dimitrios Zeimpekis and Efstratios Gallopoulos. Linear and non-linear dimensional reduction via class representatives for text classification. In *Proceedings of International Conference on Data Mining*, 2006. 2

[3] In zero-shot learning setting the test data from the seen class are typically very large compared to those from unseen classes. This makes other measures, such as accuracy, useless since high accuracy can be obtained even if all the unseen class test data are wrongly classified; hence we used ROC curves, which are independent of this problem.