Modeling Attributes from Category-Attribute Proportions

Felix X. Yu[1] Liangliang Cao[2] Michele Merler[2]
Noel Codella[2] Tao Chen[1] John R. Smith[2] Shih-Fu Chang[1]
[1]Columbia University [2]IBM Research
{yuxinnan, taochen, sfchang}@ee.columbia.edu
{liangliang.cao, mimerler, nccodell, jsmith}@us.ibm.com

ABSTRACT

Attribute-based representation has been widely used in visual recognition and retrieval due to its interpretability and cross-category generalization properties. However, classic attribute learning requires manually labeling attributes on the images, which is very expensive, and not scalable. In this paper, we propose to model attributes from category-attribute proportions. The proposed framework can model attributes without attribute labels on the images. Specifically, given a multi-class image datasets with N categories, we model an attribute, based on an N-dimensional category-attribute proportion vector, where each element of the vector characterizes the proportion of images in the corresponding category having the attribute. The attribute learning can be formulated as a learning from label proportion (LLP) problem. Our method is based on a newly proposed machine learning algorithm called $\propto$SVM. Finding the category-attribute proportions is much easier than manually labeling images, but it is still not a trivial task. We further propose to estimate the proportions from multiple modalities such as human commonsense knowledge, NLP tools, and other domain knowledge. The value of the proposed approach is demonstrated by various applications including modeling animal attributes, visual sentiment attributes, and scene attributes.

Categories and Subject Descriptors

I.5.4 [**Applications**]: Computer Vision

Keywords

visual attribute; visual recognition; learning from label proportions

1. INTRODUCTION

Attributes often refer to human nameable properties that are shared across categories. Some examples are animal attributes (furry, striped, black), scene attributes (open, natural, indoor), visual sentiment attributes (happy, sad, lovely), and human attributes (long hair, round face, blue eye). Due to the interpretability and cross-category generalization properties, attributes have been used in various applications including face verification [4], image retrieval [11, 14], action recognition [7], and recognition with few or zero examples [6, 13]. Attributes are sometimes referred as "concepts" in multimedia [12].

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654993.

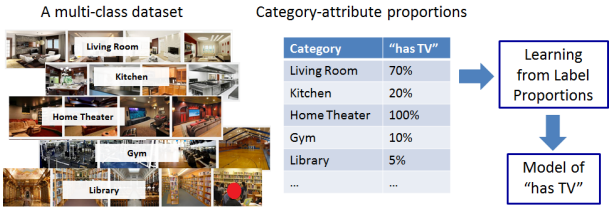

Figure 1: Illustration of the proposed framework for modeling the attribute "has TV". The input includes a multi-class dataset, and a category-attribute proportion vector. The output is an attribute model to predict "has TV" for new images.

Unfortunately, conventional attribute modeling requires expensive human efforts to label the attributes on a set of images. In this paper, we propose *attribute modeling based on category-attribute proportions*, an efficient attribute modeling framework, which requires no attribute labeling on the images. Figure 1 illustrates our framework by a conceptual example of modeling the attribute "has TV". The input includes two parts:

- A multi-class image datasets of N categories, i.e. a set of images, each with a category label. Such datasets are widely available in various visual domains, such as objects, scenes, animals, human faces etc.
- An N-dimensional category-attribute proportion vector, where the i-th dimension of the vector characterizes the *proportion* of positive images of the attribute in the i-th category.

Given the above input, the attribute learning problem naturally fits the machine learning framework called *learning from label proportions* (LLP) [10, 16]. We can then use the existing LLP techniques to train an attribute classifier, whose output can be used to predict whether the attribute is present in a new image. The above framework requires no attribute labels on the images for training. Intuitively, it is more efficient to collect the category-attribute label proportions than image-level attribute labels. For example, based on statistics, or commonsense, "80% bears are black", "90% Asians are with black hair", and "70% living rooms have a TV".

Our work makes the following contributions. We propose a framework to model attributes based on category-attribute proportions, in which no image-level attribute labels are needed (Section 1). Finding the category-attribute proportions is still not a trivial task. To this end, we propose methods for efficiently estimating the category-attribute proportions from different modalities, such as automatic NLP tools, and manual efforts with minimal human interactions (Section 3). The effectiveness of the proportion method is verified by various applications including modeling animal attribute, sentiment attributes, and scene attributes (Section 4).

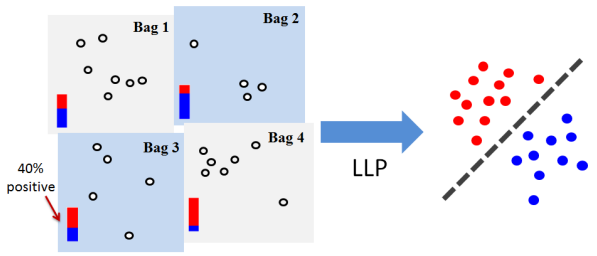

Figure 2: Illustration of learning from label proportions (LLP). In this examples, the training data is provided in 4 bags, each with its label proportion. The learned model is a separating hyperplane to classify the individual instances.

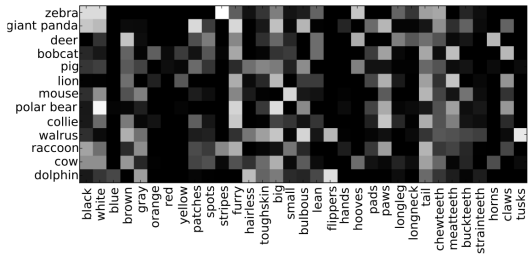

Figure 3: Manually defined category-attribute similarity matrix copied from [6]: the rows are the categories, and the columns are the attributes. This matrix is obtained from human judgments on the "relative strength of association" between attributes and animal categories.

2. LEARNING FROM LABEL PROPORTIONS

The learning setting. Key to the proposed approach is a machine learning setting called learning from label proportions (LLP). LLP is a binary learning setting, where the training data is provided in "groups" or "bags", and for each bag, only the proportion of positive instances is given. The task is to learn a model to predict the labels of the individual instances. Figure 2 illustrates a toy example of LLP. Compared to supervised learning, where the exact labels of all the training instances are known, only the label proportions for the bags are given in LLP. Therefore, LLP is a very challenging problem. Recently there have been several LLP algorithms proposed with encouraging results [10, 16]. The feasibility of this learning setting has also been verified from a theoretical perspective [15]. LLP has broad applications in political science, marketing, and healthcare. Very recently, in the multimedia and computer vision communities, LLP has been successfully applied in video event detection [5].

Applying LLP to attribute modeling. As introduced in Section 1, our problem can be viewed as a learning with label proportion setting. Here the bags are defined by the N categories, each containing some corresponding images (instances). The proportions are represented by a category-attribute proportion vector. The task is to model the attribute based on such information. In the case of modeling k attributes, we will need an $N \times k$ category-attribute proportion matrix, where the i-th column characterizes the proportion for the i-th attribute, $i = 1, \cdots, k$. For simplicity, the k attributes are modeled independently in this work.

The $\propto$SVM algorithms. Among the LLP algorithms, a recently proposed method called $\propto$SVM has been shown to outperform the alternatives [16]. We therefore use this algorithm in our work. $\propto$SVM is based on a generalization of SVM. Compared to SVM, $\propto$SVM models the unknown instance labels as latent variables. It also includes one additional loss function of the label proportions. The algorithm jointly optimizes both the model parameters and the latent labels in a large-margin framework. In other words, $\propto$SVM tries to find a large-margin classifier which is compatible with the given label proportions, with the help of latent labels. Our implementation of $\propto$SVM is based on the alter-$\propto$SVM algorithm[1], with liblinear as the underlying QP solver.

3. COLLECTING CATEGORY-ATTRIBUTE PROPORTIONS

Collecting the *exact* category-attribute proportion is still a challenging problem. In this section, we propose several ways of efficiently estimating the proportions.

3.1 Human knowledge based

Perhaps the most straightforward method is to estimate the proportions based on human commonsense. For example, it is easy to know things like "80% bears are black", "100% home theater have TVs". To alleviate the bias of individual user, one can estimate the proportion based on averaged value of multiple persons. In addition, to make the human interaction task easier, one can also discrete the proportion values, for example, into 0, 1/3, 2/3, 1.

Similar method has been used in modeling the category-level attributes in [6]. Figure 3 shows a subset of the category-attribute similarity matrix on the AwA dataset. [6] treats this matrix as a "visual similarity" matrix. When training the attributes, they binarize the matrix, and treat the images of all positive categories as positive for the corresponding attribute. Unfortunately, the binarization will lead to huge information loss. Different from the above work, we treat this matrix as a proportion matrix. Based on the animal dataset provided in [6]. we will show by experiments (based on user study) that LLP provides better results than the binarization approach.

3.2 NLP tools based

To make the above commonsense based approach more efficient and scalable, we can also use NLP tools to automatically build such a category-attribute proportion matrix. For example, the ConceptNet[2] can be used to construct a category-attribute similarity matrix. The ConceptNet is a hypergraph that links a large amount of concepts (words) by the knowledge discovered from Wikipedia and WordNet, or provided by community contributors. The concept similarity can be computed by searching the shortest path in this graph. We apply the association function of the Web API of ConceptNet 5 to get the semantic similarity between the categories and attributes. After getting this semantic similarity matrix, we use the normalized similarity as the estimation of the proportions. We demonstrate this NLP tool based approach in the experiment section by modeling visual sentiment attributes.

3.3 Transferring domain knowledge and domain statistics

The proportions can also be borrowed from knowledge of other domains. For example, we can get the "black hair" proportion of different ethnic groups of people based on genetic research. And by combining such statistics with a multi-ethnic group human face dataset, we can model the attribute "black hair". In addition, the proportions can be borrowed from another dataset whose proportion is available. We will show one application of this approach, modeling scene attributes, in the experiment section.

[1] https://github.com/felixyu/pSVM

[2] http://conceptnet5.media.mit.edu/

Categories	Attributes	Source of Proportions	Evaluation Method
Animals	Animal visual properties	Commonsense by human	User study
Sentiment attributes	Sentiment attributes	Commonsense by ConceptNet	Test on a labeled set
Scenes	Scene properties	Borrowed from another dataset	Application-based (event detection)

Table 1: Summary of the three applications explored in our experiments.

3.4 Discussion

The proposed attribute modeling technique provides an efficient alternative to the classic approaches. However there are some limitations. First, the estimated category-attribute proportion has to be close to the *exact* proportion of the dataset. This may not be true if the multi-class dataset is very biased, or if the number of images in each category is very small. Second, enough number of categories are needed. For example, the method is not going to work in the worst case scenario where only a single category and a single category-attribute proportion value is available.

4. EXPERIMENTS

We demonstrate the power of the proposed framework in three different applications: modeling animal attributes, modeling sentiment attributes, and modeling scene attributes. The three applications are summarized in Table 3.1. The parameters of the LLP algorithm are tuned based on cross-validation in terms of the proportion loss. The algorithm we use has the same computational complexity of linear SVM (it scales linearly to the number of images). In practice, the LLP algorithm is several times slower than linear SVM due to the alternating minimization process [16].

4.1 Modeling Attributes of Animals

Setting. Our first experiment consists in modeling animal attributes on the AwA dataset [6], which contains 30,475 images of 50 animal categories. Each image is uniquely labeled as one of the 50 categories. Associated with the dataset, there is a category-attribute similarity matrix of 85 attributes based on manual efforts mentioned in Section 3.1. A subset of the category-attribute matrix is shown in Figure 3. We use the same set of low-level features provided in [6]. In order to model the attributes, [6] first thresholds the matrix to a 0/1 matrix. They then train 85 attribute classifiers, where for each attribute, all the images belonging to the positive categories are treated as positive, and all images belonging to negative categories are treated as negative. The binarization step obviously leads to a big information loss.

Method. In this work, we treat the similarity matrix as a category-attribute proportion matrix, and train the attribute models with $\propto$SVM. We use 50% images for training and 50% for testing.

Evaluation. As there is no labeled images for the 85 attributes, it is hard to directly compare our method with the baselines quantitatively. For evaluation, we perform a preliminary user study. For each attribute, 20 images are randomly selected from the top-100 ranked images for each method. Human subjects are then asked to determine which method produces better attribute modeling results. The subjects are 5 graduate students majoring in engineering and business, who are not aware of the underlying learning framework. In this experiment, the users prefer our results over the baseline ones 74% of the time. This clearly demonstrates the plausibility of the newly proposed framework.

4.2 Modeling Sentiment Attributes

Setting. We consider the task of modeling object-based sentiment attributes such as "happy dog", and "crazy car". Such attributes are defined in [1, 3]. In this work we consider three nouns: dog, car, face and the sentiment attributes associated with them. This results in 77 sentiment attributes (or adjective-noun pairs, ANPs) to be modeled. Such ANPs appear widely in social media. We use the data and features provided by [1]: for each ANP, there is a set of images collected by querying that ANP on Flickr. [1] uses such labels to train one-vs-all linear SVMs to model the ANPs. One critical problem for this approach is that many sentiment attributes are intrinsically ambiguous. For example, a "cute dog" can also be a "lovely dog". Therefore, when modeling "lovely dog", some images belonging to "cute dog" should also be positive.

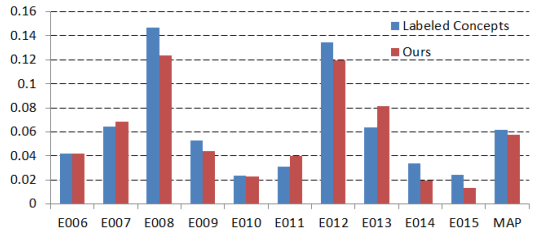

Figure 6: Event detection APs based on our attribute models, and the manual concept models. The modeled attributes (without manual labeling process) provides very competitive results compared to the concepts (with manual labeling process).

Method. To solve the above problem, we first use the method of Section 3.2 to collect the semantic similarity between every pairs of ANPs. We then use our framework to model the ANPs. Different from other applications, both the categories and the attributes are ANPs. The proposed framework is used to improve the modeling of existing attributes, rather than learning new ones.

Evaluation. To evaluate the ANP modeling performance, for each ANP, we manually label 40 positive images, and 100 negative images from a separate set. Multiple people are involved in the labeling process, and images with inconsistent labels are discarded. Figure 4 compares our ANP modeling performance with [1]. Our approach dramatically outperforms the baseline for most of the sentiment attributes. Our method provides a relative performance gain of 30% in terms of the Mean Average Precision.

4.3 Modeling Scene Attributes

Setting. Concept classifiers have been successfully used in video event detection [2, 8]. In such systems, concept classifiers (about scenes, objects, activities etc.) are trained based on a set of labeled images visually related to the event detection task. The trained classifiers are used as feature extractors to obtain mid-level semantic representations of video frames for event detection. Key to the above event detection paradigm is a set of comprehensive concepts. In order to model one additional concept, traditional approaches require an expensive manual labeling process to label the concepts on the images. The objective of this experiment is to model the 102 scene attributes defined by [9] with the IMARS images [2, 8], without the requirement of manual labeling. Some examples of the scene attributes are "cold", "dry", and "rusty". Existing IMARS concepts do not cover the 102 attributes.

Method. We first compute the empirical category-attribute proportions based on a separate multi-class scene datasets with 717 categories ("SUN attribute dataset") [9], in which each image comes with the attribute labels. Such proportions are then used on the IMARS set to model the attributes. For each of the 717 categories, we can find the corresponding set of images on IMARS based on

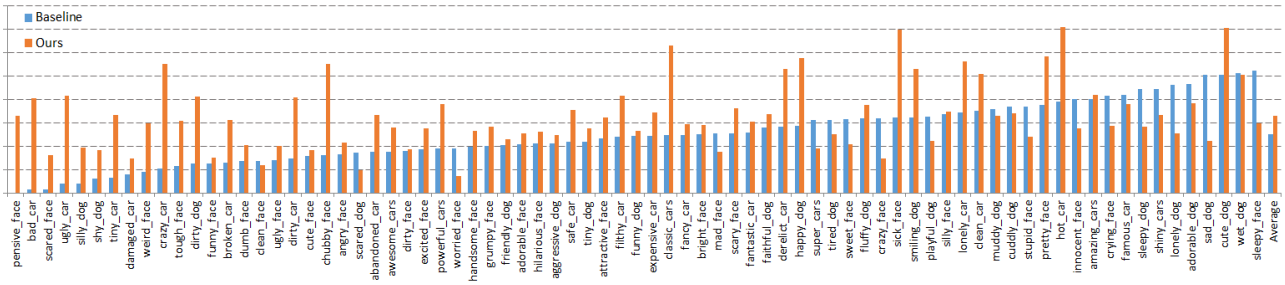

Figure 4: Experiment result of sentiment attribute modeling. The figure shows AP@20 of our method and the baseline (binary SVM). The AP is computed based on a labeled evaluation set.

(a) bathing (b) cold (c) dry (d) reading (e) rusty

Figure 5: Top ranked images of the learned scene attributes classifiers on IMARS.

the concept labels[3]. The supervised information is a 717 $\times$ 102 dimensional category-attribute proportion matrix. We then train 102 $\propto$SVM classifiers to model the attributes.

Evaluation. Figure 5 visualizes a few modeled attributes by the top ranked images of IMARS. We can qualitatively see that the attribute classifiers can successfully capture the corresponding visual properties. We further apply the learned attribute models in the event detection task. The evaluation is based the TRECVID MED 2011 events with the evaluation set and pipeline described in [2, 8]. The baseline method is the concepts trained by manual labels. The attributes modeled by category-attribute proportions are very competitive compared to the manually labeled concepts in terms of the average precision on the event detection task. For certain events, e.g., E007, E013, the performance of the attributes even outperforms the manual concepts. In summary, the proposed technique provides an efficient way of expanding IMARS concept classifiers for event detection.

5. CONCLUSION

We proposed a novel framework of modeling attributes based on category-attribute proportions. The framework is based on a machine learning setting called learning from label proportions (LLP). We showed that the category-attribute proportion can be efficiently estimated by various methods. The effectiveness of the proposed scheme has been demonstrated by various applications including modeling animal attributes, sentiment attributes, and scene attributes.

Acknowledgement This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of IARPA, DoI/NBC, or the U.S. Government. We thank Quoc-Bao Nguyen and Matthew Hill for their help. Felix Yu is partly supported by the IBM PhD Fellowship.

[3]Note that one could train the attribute models based on SUN attribute dataset directly. But such attribute models do not lead to satisfactory result on IMARS due to cross-domain issues. Instead, the proportions of the two datasets are empirically very similar.

6. REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia*, 2013.

[2] L. Brown et al. IBM Research and Columbia University TRECVID-2013 Multimedia Event Detection (MED), Multimedia Event Recounting (MER), and Semantic Indexing (SIN) Systems. In *NIST TRECVID Workshop*, 2013.

[3] T. Chen, F.X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang. Object-based visual sentiment concept analysis and application. In *ACM Multimedia*, 2014.

[4] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *CVPR*, 2009.

[5] K.-T. Lai, F.X. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014.

[6] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[7] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[8] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, 2012.

[9] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[10] N. Quadrianto, A.J. Smola, T.S. Caetano, and Q.V. Le. Estimating labels from label proportions. In *ICML*, 2008.

[11] B. Siddiquie, R.S. Feris, and L.S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

[12] J.R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *ICME*, 2003.

[13] F.X. Yu, L. Cao, R.S Feris, J.R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[14] F.X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012.

[15] F.X. Yu, S. Kumar, T. Jebara, and S.-F. Chang. On learning with label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

[16] F.X. Yu, D. Liu, Sanjiv K., T. Jebara, and S.-F. Chang. $\propto$SVM for learning with label proportions. In *ICML*, 2013.