Inferring Analogous Attributes: Large-Scale Transfer of Category-Specific Attribute Classifiers

Chao-Yeh Chen and Kristen Grauman

Abstract

The appearance of an attribute can vary considerably from class to class, causing standard class-independent attribute models to break down. Yet, training object-specific models for each attribute is impractical, and defeats the purpose of using attributes to bridge category boundaries. We propose a novel form of transfer learning that addresses this dilemma. Given a sparse set of class-specific attribute classifiers, our tensor factorization approach can infer new ones for object-attribute pairs unobserved during training. We apply our idea to learn over 25,000 analogous attribute classifiers on SUN and ImageNet.[1]

1. Introduction

Attributes are visual properties that describe objects or scenes, such as "fluffy" or "formal". A major appeal of attributes is the fact that they appear across category boundaries. But are attributes really category-independent? Does fluffiness on a dog look the same as fluffiness on a towel? While the *linguistic* semantics are preserved across categories, the *visual* appearance of the property may be transformed to some degree. This suggests that the standard approach [1, 2, 4, 3]—pooling training images from any category and learning a discriminative classifier—will weaken the learned model to account for the "least common denominator" of the attribute's appearance.

Taking the other extreme, one might attempt to learn *category-sensitive*[2] attribute classifiers, by gathering positive exemplar images for each category+attribute combination (e.g., separate sets of fluffy dog images, fluffy towel images). However, learning attributes in this manner is quite costly in terms of annotations. In fact, even in the era of Big Vision, the long-tailed distribution of object/scene/attribute occurrences in the real world means that some object-attribute pairs will have inadequate exemplars to build a statistically sound model. Furthermore, naively training each attribute in an object-specific manner would fail to leverage the common semantics of attributes.

[1]Per the call for papers, we are submitting single-blind because this work appears in the main conference at CVPR 2014.

[2]We use "category" to refer to either an object or scene class.

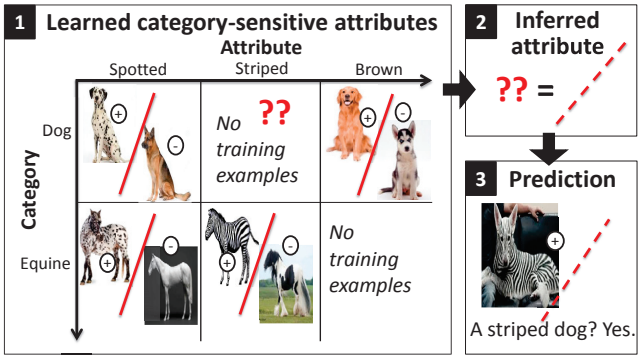

Figure 1. Having learned a sparse set of object-specific attribute classifiers, our approach infers *analogous attribute classifiers*. The inferred models are object-sensitive, despite having no object-specific labeled images of that attribute during training.

To resolve this problem, we propose a novel form of transfer learning to infer *analogous* category-sensitive attribute models. Intuitively, even though an attribute's appearance may be specialized for a particular object, there likely are latent variables connecting it to other objects' manifestations of the property. So, having learned some category-sensitive attributes, we aim to predict how the attribute might look on a new object, *even without labeled examples depicting that object with the attribute*. See Fig. 1.

2. Approach: Main Idea

Given training images labeled by their category and attributes, our method produces as output a series of category-sensitive attribute classifiers. Some of those classifiers are explicitly trained with the labeled data, while the rest are "analogous" attributes inferred by our method using no additional training images.

For each attribute for which we *do* have category-specific labeled examples, we train an importance-weighted support vector machine (SVM). It uses images from all categories, but places a higher penalty on violating attribute label constraints for the specified category. Only a fraction of possible attributes can be explicitly trained in this manner—for example, only $\sim$25% for ImageNet or SUN.

Next we define a tensor to capture the structure underlying many such category-sensitive models. Let m index the

	Trained explicitly		Trained via transfer		
	Category-sens.	Universal	Inferred	Adopt	One-shot
ImageNet	**0.7589**	0.7037	**0.7428**	0.6194	0.6309
SUN	**0.6505**	0.6343	**0.6429**	N/A	N/A

Table 1. mAP accuracy for thousands of attribute models.

M attributes in the vocabulary, let n index the N possible object/scene categories, and let D be the image descriptor dimensionality. Let $\boldsymbol{w}(n, m)$ denote a category-sensitive linear SVM weight vector trained for the n-th object and m-th attribute. We construct a tensor $\mathbf{W} \in \Re^{N \times M \times D}$ using all available category-sensitive models. Each entry w^d_{nm} contains the value of the d-th dimension of the classifier $\boldsymbol{w}(n, m)$. The resulting tensor is quite sparse; we can only fill entries for which we have class-specific positive and negative training examples for the attribute of interest.

Rather than resort to universal models for the "missing" combinations, we propose to use the latent factors for the observed classifiers to synthesize analogous models for the unobserved classifiers. Let $\mathbf{O} \in \Re^{K \times N}$, $\mathbf{A} \in \Re^{K \times M}$, and $\mathbf{C} \in \Re^{K \times D}$ denote matrices whose columns are the K-dimensional latent feature vectors for each object, attribute, and classifier dimension, respectively, discovered with Bayesian tensor factorization [5]. These factors affect how the various attributes, objects, and image descriptors covary (e.g., one might capture how "spots" appear on something "flat" vs. how they appear on something "bumpy"). We suppose that an analogous attribute $\boldsymbol{w}(n, m)$ can be expressed as an inner product of latent factors: $w^d_{nm} \approx \langle O_n, A_m, C_d \rangle$. In this way, we infer how an attribute will look for another object category that lacks any images labeled for that attribute.

3. Example Experimental Results

We evaluate our approach on ImageNet [4](384 object categories and 25 attributes) and SUN Attributes [3](280 categories and 59 attributes) and use standard GIST, color, SIFT descriptors. See our CVPR 2014 paper for all details.

First we test whether category-sensitive attributes are even beneficial. We explicitly train 1,498 and 6,118 category-sensitive attribute classifiers for ImageNet and SUN, respectively, and compare them to the standard universal class-independent approach. Both models have access to the exact same set of images in training and testing. Table 1 (cols 2 and 3) indicates it is indeed worthwhile to tailor attributes to specific categories when possible. Category-sensitive models surpass universal ones in 76% of the cases, with average increases of 0.15 in AP.

We stress that the explicit models (above) *are impossible to train for 18K of the ∼26K possible attributes in these datasets.* This is where our method comes in. We infer all remaining $18K$ attribute models with our method, in a leave-one-out manner.

Figure 2. Analogous attribute examples for ImageNet (top) and SUN (bottom). Words above each neighbor indicate the 3 most similar attributes (learned or inferred) between leftmost query category and its neighboring categories in latent space (*not* the image's attribute prediction). Query category:neighbor category = 1.Bottle: filter, syrup, bullshot, gerenuk. 2.Platypus: giraffe, ungulate, rorqual, patas. 3.Airplane cabin: aquarium, boat deck, conference center, art studio. 4.Courtroom: cardroom, florist shop, performance arena, beach house.

Table 1 (col 4) shows this key result, with comparisons to standard transfer methods where applicable (cols 5 and 6). Our inferred analogous attributes are nearly as accurate as the "upper bound" category-sensitive results, yet use no category-specific labeled images. Critically, our inferred models are more accurate than the status quo universal approach. We infer models for *all* missing attributes; whereas the category-sensitive method would require 20 labeled examples per classifier—about $384K$ additional labeled images—to train those models, our method uses zero.

Fig. 2 illustrates how analogous attributes enable transfer. We take a category j and identify its neighboring categories in the latent feature space. Then, for each neighbor i, we sort its attribute classifiers ($\boldsymbol{w}(i, :)$, real or inferred) by their maximal cosine similarity to any of category j's attributes $\boldsymbol{w}(j, :)$. The resulting shortlist shows which attribute+category pairs our method expects to transfer to category j. We show 4 examples, with one representative image for each category. Neighboring categories in the latent space are often semantically related (e.g., syrup/bottle) or visually similar (e.g., airplane cabin/conference center). Our method receives no explicit side information on semantic distance, yet it discovers such ties via the observed attribute classifiers. Some semantically more distant neighbors (e.g., platypus/rorqual, courtroom/cardroom) are also amenable to transfer.

References

[1] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009.

[2] C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009.

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

[4] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *ECCV Workshop on Parts and Attributes*, 2010.

[5] L. Xiong, X. Chen, T. Huang, J. Schneider, and J. Carbonell. Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In *SDM*, 2010.