Discovering Shades of Attribute Meaning with the Crowd

Adriana Kovashka and Kristen Grauman

The University of Texas at Austin
{adriana,grauman}@cs.utexas.edu

Abstract. To learn semantic attributes, existing methods train one discriminative model for each word in a vocabulary of nameable properties. This "one model per word" assumption is problematic: while a word might have a precise linguistic definition, it need not have a precise visual definition. We propose to discover *shades* of attribute meaning. Given an attribute name, we use crowdsourced image labels to discover the latent factors underlying how annotators perceive the named concept. Structure in those latent factors helps reveal shades, i.e., interpretations for the attribute shared by some group of annotators. Using these shades, we train classifiers to capture the variants of the attribute. The resulting models are both semantic and visually precise, and improve attribute prediction accuracy on novel images.

1 Introduction

Attributes are semantic properties of objects and scenes. By injecting language into visual analysis, attributes broaden the visual recognition problem—from labeling images, to *describing* them. Typically one defines a vocabulary of attribute words relevant to the domain at hand. Then one gathers labeled images depicting each attribute, and trains a model to recognize each word [1–10].

The problem with this standard approach, however, is that there is often a gap between language and visual perception. In particular, *the words in an attribute vocabulary need not be visually precise.* An attribute word may connote multiple "shades" of meaning—whether due to polysemy, variable context-specific meanings, or differences in humans' perception. For instance, the attribute *open* can describe a door that's ajar, a fresh countryside scene, a peep-toe high heel, or a backless clog. Each shade may require dramatically different visual cues to correctly capture. Thus, the standard approach of learning a single classifier for the attribute as a whole may break down.

Unfortunately, neither bottom-up attribute "discovery" nor relative attributes solve the problem. Unsupervised discovery methods detect clusters or splits in the low-level image descriptor space [11–16], but they need not be human-nameable (semantic). Further, discovery methods are intrinsically biased by the choice of features. Relative attributes [8] do not address the existence of shades, either. Just like categorical attributes, relative attributes assume that there is some single interpretation of the property—namely, that a single ordering of images from least to most [attribute] is possible.

Fig. 1. Our method discovers factors responsible for an attribute's presence, then learns predictive models based on those visual cues. For example, for the attribute *open*, our method will discover peep-toed (*open* at toe) vs. slip-on (*open* at heel) vs. sandal-like (*open* at toe *and* heel), which are three visual definitions of openness. Since these shades are not coherent in terms of their global descriptors, they'd be difficult to discover using traditional image clustering.

Our goal is to automatically discover the shades of an attribute. An attribute "shade" is a visual interpretation of an attribute name that one or more people apply when judging whether that attribute is present in an image. See Figure 1.

Note that work in automatically finding the multiple "senses" of a polysemous word [17–20] is orthogonal to our goal, as it focuses on nouns/object categories, not descriptive properties. Further, the visual differences of polysemous nouns are stark (e.g., a river *bank* or financial *bank*). In contrast, attribute shades are often subtle differences in interpretation. Unlike noun senses, attribute shades cannot be easily enumerated in a dictionary.

2 Approach

Given a semantic attribute name, we want to discover and model its multiple visual interpretations. Rather than attempt to manually enumerate the possible shades, we propose to learn them indirectly from the crowd. First we ask many annotators to label various images. We then estimate latent factors that represent the annotators in terms of the visual cues they associate with the attribute. By clustering in the low-dimensional latent space, we identify the "schools of thought" underlying the labels. (We use the terms "school" and "shade" interchangeably.) Finally, we use the positive exemplars in each school to train a predictive model.

We use two datasets: Shoes [21, 9] and SUN Attributes [10]. To focus our study on plausibly "shaded" words, we select 12 attributes (see Table 1) that can be defined concisely in language, yet may vary in their visual instantiations. We show definitions of these attributes to our workers on Amazon Mechanical Turk. We sample $N = 250$ to 1000 representative images to be labeled per attribute, and obtain annotations from $M = 195$ workers on average. We show a random subset of 50 images to each worker, and ask him to state whether a given attribute is present in the images. For a random set of 5 images, the worker must also explain his label in free-form text. These questions slow the worker down, helping quality control, and provide ground truth data for evaluation.

Now we use the label data to discover latent factors, which are needed to recover the shades of meaning. Let $\mathbf{L}$ be the $M \times N$ label matrix, where $L_{ij} \in \{0, 1, ?\}$ is a binary attribute label for image j by annotator i. A ? denotes an unlabeled example. We suppose there is a small number D of unobserved

factors that influence the annotators' labels. This reflects that their decisions are driven by some mid-level visual cues. The label matrix $\mathbf{L}$ can be factored as the product of an $M \times D$ annotator latent factor matrix $\mathbf{A}^T$ and a $D \times N$ image latent factor matrix $\mathbf{I}$: $\mathbf{L} = \mathbf{A}^T\mathbf{I}$. While a number of existing methods can be used to factor this partially observed matrix, we use a probabilistic matrix factorization algorithm (PMF) [22, 23] due to its efficiency. We represent each annotator in terms of his association with each discovered factor. Details can be found in our technical report [24].

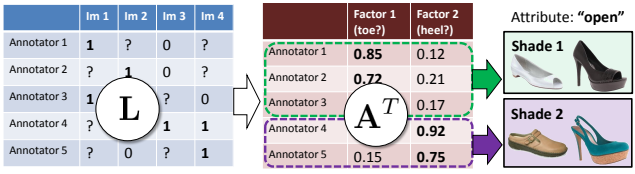

Fig. 2. Given an attribute label matrix (left), we recover its latent factors and their influence on each annotator (middle). We discover shades as clusters in this space (right).

Figure 2 illustrates with a toy example. Some annotators tended to label images 1 and 2 as having the attribute, whereas others labeled 3 and 4 as positive. Suppose we discover $D = 2$ latent factors. Though nameless, they align with semantic visual cues; suppose here they are "toe is open" and "heel is open". In this example, we see the first three annotators labeled images 1 and 2 as open due to factor 1, whereas the others focused on factor 2 in other images.

After recovering each user's latent factor vector, we apply K-means to the columns of $\mathbf{A}$ to obtain clusters $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$. Each cluster is a shade. Annotators in the same cluster show similar labeling behavior, meaning they interpret similar combinations of mid-level visual cues as salient for the attribute at hand. Depending on the visual precision of the word, some attributes may have only one shade; others may have many. To automatically select K, we use the silhouette coefficient [25].

Now we can use shades to improve attribute prediction accuracy. We represent each shade $\mathcal{S}_k$ as the total pool of images that its annotators labeled as positive. If multiple annotators in the shade labeled an image, we perform a majority vote within the shade to decide on the label. We use the images to train a classifier, using the Adapt-SVM objective [26] to regularize its parameters to be similar to those of a *generic* model for this attribute trained with majority vote labeled examples from any annotator [1–7, 10]. Then we apply the adapted shade model for the cluster to which a user belongs, to predict the presence or absence of the attribute in novel images.

Prior work on attribute learning uses one of two extremes—either a GENERIC classifier, or a USER-ADAPTIVE classifier trained by adapting that generic model to satisfy an individual user's training labels [27]. We propose an approach between these extremes. With shades, we can account for differing perceptions of an attribute, yet avoid specializing predictions down to the level of each individual user. In contrast to [27], we can "borrow" labels from the user's neighbors in the crowd, and leverage the robustness of the intra-shade majority vote.

Fig. 3. Top words and images for two shades per attribute. Best seen zoomed in PDF.

3 Results

Attribute	Shades	Generic	User-adp
Pointy	**76.3** (0.3)	74.0 (0.4)	74.8 (0.3)
Open	**74.6** (0.4)	66.5 (0.5)	71.6 (0.3)
Ornate	**62.8** (0.7)	56.4 (1.1)	61.1 (0.6)
Comfort.	**77.3** (0.6)	75.0 (0.7)	75.5 (0.6)
Formal	**78.8** (0.5)	76.2 (0.7)	77.1 (0.4)
Brown	**70.9** (1.0)	69.5 (1.2)	68.5 (0.9)
Fashion.	**62.2** (0.9)	58.5 (1.4)	62.0 (1.4)
Cluttered	**64.5** (0.3)	60.5 (0.5)	63.1 (0.4)
Soothing	**62.5** (0.4)	61.0 (0.5)	61.5 (0.4)
Open area	**64.6** (0.6)	62.9 (1.0)	63.5 (0.5)
Modern	**57.3** (0.8)	51.2 (0.9)	56.2 (1.1)
Rustic	**67.4** (0.6)	66.7 (0.5)	67.0 (0.5)

Table 1. Accuracy of predicting perceived attributes.

Table 1 shows the accuracy of predicting personalized attributes. Our method outperforms all other methods. For "open", we achieve an 8 point gain over Generic, which indicates how distinct user perceptions of this attribute are. We also outperform User-adaptive [27], while requiring the exact same labeling effort. While that method learns personalized models, shades leverage *common perceptions* and thereby avoid overfitting to a user's few labeled instances.

We next perform an image search test. Given a multi-attribute query (e.g. "white formal shoes"), we measure the fraction of test images where a method's predictions on all requested attributes agree with the ground truth user labels. For queries with 2 to 6 attributes, we see similar relative gain of our shades approach as in Table 1.

We also find that the clusters our method discovers are more coherent than alternative shade formation methods. For each given shade, we pool all the workers' (held-out) textual explanations for their labels, perform pLSA [28], and examine the distribution of topics in each school. Low-entropy clusters are better as they correspond to more coherent sets of descriptions. Our method produces more coherent clusters than an existing Attribute discovery method [14] that mines for non-semantic attributes, and Image clusters: an image descriptor clustering approach inspired by [18]. The corresponding figure can be found in our technical report.

In Figure 3, we see that shades capture nuanced visual sub-definitions of the attribute words. For "open", one shade includes open-heeled shoes (top), while another includes sandals open at the front *and* back (bottom). For "comfortable", one shade emphasizes a low arch (top), while the other requires soft materials (bottom). The "open area" attribute can be either outside (top) or inside (bottom).

Our approach to discover attribute shades brings together language, crowdsourcing, human perception, and visual representations in a new way. Shades show great promise to separate the (sub-)attributes involved in a person's use of an attribute vocabulary during search or organization of image content.

References

1. Ferrari, V., Zisserman, A.: Learning Visual Attributes. In: NIPS. (2007)
2. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Describable Visual Attributes for Face Verification and Image Search. PAMI (2011)
3. Lampert, C., Nickisch, H., Harmeling, S.: Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In: CVPR. (2009)
4. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing Objects by Their Attributes. In: CVPR. (2009)
5. Vaquero, D., Feris, R., Tran, D., Brown, L., Hampapur, A., Turk, M.: Attribute-based People Search in Surveillance Environments. In: WACV. (2009)
6. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual Recognition with Humans in the Loop. In: ECCV. (2010)
7. Wang, Y., Mori, G.: A Discriminative Latent Model of Object Classes and Attributes. In: ECCV. (2010)
8. Parikh, D., Grauman, K.: Relative Attributes. In: ICCV. (2011)
9. Kovashka, A., Parikh, D., Grauman, K.: WhittleSearch: Image Search with Relative Attribute Feedback. In: CVPR. (2012)
10. Patterson, G., Hays, J.: SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In: CVPR. (2012)
11. Parikh, D., Grauman, K.: Interactively Building a Discriminative Vocabulary of Nameable Attributes. In: CVPR. (2011)
12. Mahajan, D., Sellamanickam, S., Nair, V.: A Joint Learning Framework for Attribute Models and Object Descriptions. In: ICCV. (2011)
13. Duan, K., Parikh, D., Crandall, D., Grauman, K.: Discovering Localized Attributes for Fine-grained Recognition. In: CVPR. (2012)
14. Rastegari, M., Farhadi, A., Forsyth, D.: Attribute Discovery via Predictable Discriminative Binary Codes. In: ECCV. (2012)
15. Sharmanska, V., Quadrianto, N., Lampert, C.: Augmented Attribute Representations. In: ECCV. (2012)
16. Yu, F., Cao, L., Feris, R., Smith, J., Chang, S.F.: Designing Category-Level Attributes for Discriminative Visual Recognition. In: CVPR. (2013)
17. Barnard, K., Yanai, K., Johnson, M., Gabbur, P.: Cross Modal Disambiguation. Toward Category-Level Object Recognition (2006)
18. Loeff, N., Alm, C., Forsyth, D.: Discriminating Image Senses by Clustering with Multimodal Features. In: ACL. (2006)
19. Saenko, K., Darrell, T.: Unsupervised Learning of Visual Sense Models for Polysemous Words. In: NIPS. (2008)
20. Berg, T.L., Forsyth, D.A.: Animals on the Web. In: CVPR. (2006)
21. Berg, T.L., Berg, A.C., Shih, J.: Automatic Attribute Discovery and Characterization from Noisy Web Data. In: ECCV. (2010)
22. Salakhutdinov, R., Mnih, A.: Probabilistic Matrix Factorization. In: NIPS. (2007)
23. Salakhutdinov, R., Mnih, A.: Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo. In: ICML. (2008)
24. Kovashka, A., Grauman, K.: Discovering Attribute Shades of Meaning with the Crowd. Technical Report AI13-02, The University of Texas at Austin (November 2013)
25. Rousseeuw, P.: Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics **20** (1987) 53–65

26. Yang, J., Yan, R., Hauptmann, A.G.: Adapting SVM Classifiers to Data with Shifted Distributions. In: ICDM Workshops. (2007)
27. Kovashka, A., Grauman, K.: Attribute Adaptation for Personalized Image Search. In: ICCV. (2013)
28. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: UAI. (1999)