

Spoken Attributes: Mixing Binary and Relative Attributes to Say the Right Thing

Amir Sadvnik
Cornell University
as2373@cornell.edu

Andrew Gallagher
Cornell University
acg226@cornell.edu

Devi Parikh
Virginia Tech
parikh@vt.edu

Tsuhan Chen
Cornell University
tsuhan@ece.cornell.edu

Abstract

In recent years, there has been a great deal of progress in describing objects with attributes. Attributes have proven useful for object recognition, image search, face verification, image description, and zero-shot learning. Typically, attributes are either binary or relative: they describe either the presence or absence of a descriptive characteristic, or the relative magnitude of the characteristic when comparing two exemplars. However, prior work fails to model the actual way in which humans use these attributes in descriptive statements of images. Specifically, it does not address the important interactions between the binary and relative aspects of an attribute. In this work we propose a spoken attribute classifier which models a more natural way of using an attribute in a description. For each attribute we train a classifier which captures the specific way this attribute should be used. We show that as a result of using this model, we produce descriptions about images of people that are more natural and specific than past systems.

1. Introduction

In computer vision, attributes have earned a position as a valuable tool for capturing characteristics of objects. Of course, in human language, attributes have played a vital role in producing descriptions of objects for far longer. The goal in this paper is to produce better, more natural, attribute-based image descriptions that we learn from human statements and refer to as *spoken attributes*.

When observing a scene to describe, a human can choose from a huge number of attribute-based statements. Any such statement conveys information directly, but also at a deeper level. Take, for example, the attribute “smiling” for describing a pair of faces. One might say, “Tom is smiling more than Susan.” Of course, this direct statement communicates the fact that Tom’s level of the “smiling” attribute is greater than Susan’s, but that is not the end of the story. The statement also gives us the hint that, *at least*, Tom is smiling. We can safely reason (and support with data) that a

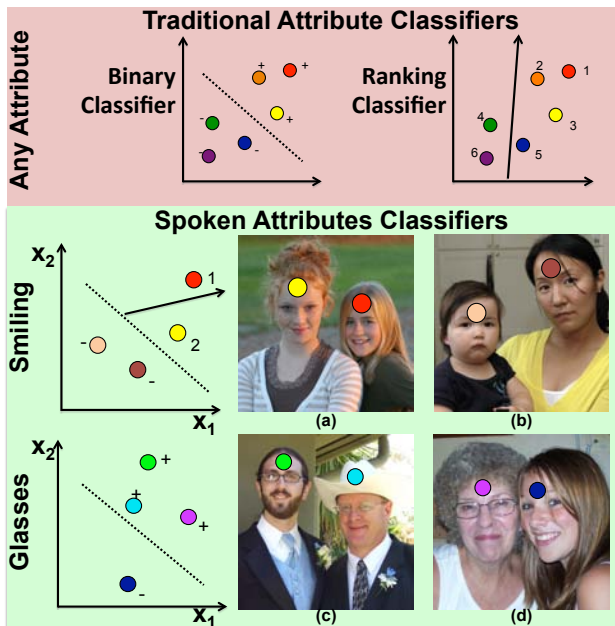


Figure 1. Previous work has focused on attributes as being exclusively either binary or relative. However, real attributes usually do not simply fall into one of those two categories. For example, for image (a) one might say “The person on the right is smiling more than the person on the left”. However, in image (b), where neither face is smiling, using such a relative smiling statement does not make sense. For the glasses attribute it never makes sense to use a relative description even though a ranker output for image (d) would suggest that the person on the left has “more glasses” than the person on the right. In this paper we propose a unified *spoken attribute* classifier that learns this mapping for each attribute and can “say the right thing.”

speaker would not bother to comment that Tom was smiling “more”, unless he was, in fact, smiling. We show that it is probable that both Tom and Susan are smiling, because it is obvious that a smiling person is smiling “more” than a non-smiling person. If Susan were not smiling, “Tom is smiling more than Susan” would typically not be worth mentioning. Therefore, even a simple relative statement about an attribute can convey a great deal of meaning and informa-

tion about the image content.

This seemingly simple illustration includes many nuanced details, many of which have no proxy in current vision algorithms. For one thing, humans fluidly move between using *binary* (“Tom and Susan are smiling”) and *relative* (“Tom is smiling more than Susan”) attribute statements. The statement that is selected depends on many factors, including the desire to convey as much information as possible with a single statement. In this work, our goal is to “say the right thing” by choosing between relative and binary attribute statements in the same way that a human describer would. We believe we are the first work to jointly consider both binary and relative attributes, and the first to model *spoken attributes* i.e. model ways in which humans mix binary and relative statements in their spoken language.

Not all attributes are used in the same ways. Many attributes have utility in both binary and relative form (e.g., “smiling”). In past work, either the binary or the relative forms of an attribute were exclusively explored. This is a limitation that humans do not share, and can result in learning attribute relationships that lose correspondence with spoken language. For example, our subjects rarely chose the phrase “more masculine” to describe the male in a male-female couple (preferring “one is male and one is not”). When the relative attribute “more masculine” is learned from male-female training pairs, the resulting model does not relate directly to the phrase humans would actually use.

Not all attributes are equally useful in both binary and relative forms. For example, the “wears glasses” attribute is exclusively binary; either a person is wearing glasses or she isn’t. One person does not have “wear glasses more” than another. Although it is possible to use relative attributes when talking about glasses such as “larger glasses” or “stylish glasses”, these are in effect attributes of glasses and could be learned separately. Our work characterizes the usage patterns of six different attributes, and shows interesting differences and commonalities between the usage of these spoken attributes.

In summary, we learn the relationship between attributes (both binary and relative) and the statements that humans use to describe images, allowing us to produce more natural image descriptions and better understand human-generated attribute-based statements. Our contributions are the following: First, we consider both binary and relative attributes in a unified framework to produce improved image descriptions. Second, we recognize that the attribute statements people make carry more information than is apparent at face value, and we model these relationships between binary attributes, relative attributes, and the statements made by humans to “read between the lines” and improve human-computer communication. Finally, we believe this work represents a first example of joining computer vision attributes with actual spoken attributes used in language to improve computer-generated image descriptions, and to in-

terpret image captions.

2. Related Work

We now describe existing works that use attributes for image understanding and human-machine communication. We relate our work to existing works on automatically generating textual descriptions of images, on modeling importance in images, as well as reading between the lines of what a human explicitly states to describe an image.

Attributes: Attributes have been used extensively, especially in the past few years, for a variety of applications [5, 16, 19, 6, 12, 14, 20, 2, 29, 27, 8, 4, 28, 18]. Binary attributes have been used to reason about properties of objects as opposed to just the object categories [6], for teaching a machine novel concepts simply by describing its properties [16], for interactive recognition with a human-in-the-loop answering questions about a test image [4], for face verification [15], for improved image search by using attributes as keywords [14, 24] and so on. Relative attributes have been used for improved zero-shot learning and textual descriptions [19], for enhanced active learning by allowing the supervisor to provide attribute-based feedback to a classifier [20, 3] and for interactively refining image search results [12]. Scheirer et al. [23] use facial attributes for search in both binary and relative form. However, they use a binary SVM score as a proxy for a relative value and do not attempt to learn which (relative vs. binary) is relevant for each attribute. To the best of our knowledge, we are the first to propose a model that automatically determines whether a categorical or relative statement about an attribute is appropriate for a given image. As a result, we achieve improved performance at human-centric applications like generating textual descriptions of images.

Textual descriptions: Attributes have been used for automatically generating textual description of images [19, 13] that can potentially also point out anomalies in objects [6]. Most recently, efforts are being made to predict entire sentences from image features [5, 17, 30, 13]. Some methods generate novel sentences for images by leveraging existing object detectors [7], attributes predictors [6, 2, 19], language statistics [30] or spatial relationships [13]. Sentences have also been assigned to images by selecting a complete written description from a large set [5, 17]. Generating compact descriptions that helps identify an image from a group of images was studied in [21]. Our work focuses on finding a good description for an image for a specific attribute, and leads to more natural descriptions. Instead of selecting a *true* attribute statement, we select a *spoken* attribute statement: a binary or relative attribute statement that a human would actually use to describe the image.

Importance: Recent works in the community have examined the notion of importance – modeling what people notice and choose to describe in images. The order in which

people are likely to name objects [25] and attributes [26] in an image has been studied. Predicting which objects, attributes, and scenes are likely to be described in an image has been recently studied [1]. Sadovnik et al. [22] examined which attributes should be mentioned in a referring expression. In this work, rather than focusing on which aspects of an image a person describes, we address the orthogonal task of modeling *how* a person is likely to describe a certain attribute in an image.

Reading between the lines: We have argued that the way in which a person chooses to describe a property in an image reveals information about the content of the image. This can be thought of as reading between the lines of what a person is saying, and not simply taking the description – listing which attributes are present or absent or more or less – at face value. At a high level, this is related to an approach that uses the order that a user tags objects in an image to determine the likely scales and locations of those objects in the image leading to improved object detection [11].

3. Understanding Spoken Attributes

To motivate our work and explore our hypothesis that limiting attributes to either binary or relative cannot faithfully capture how people actually use the attribute in descriptions, we collect a new dataset of spoken attributes. With this dataset, we can examine the contrast between conventional methods of collecting ground truth (for building either binary classifiers or relative rankers) and what people truly believe makes a good, descriptive statement for an image. We work with facial attributes because they have been extensively studied and have shown utility for exploring ideas related to attribute research. However, we believe these ideas extend to attributes in general, beyond faces.

To allow both binary and relative statements, we collect data on images with pairs of people from [9]. We use 400 images with pairs of people, and study six separate attributes (bald, beard, glasses, male, smiling, teeth visible) from [15]. Because some of these attributes are more rare (for example, the dataset does not contain many bald people, let alone pairs of bald people), we add manually-constructed images which are presented as side-by-side faces. We create 50 images for each of the rare occurrences (both faces glasses, both faces bald, both faces bearded, both faces not smiling) for a total of 600 images. For each image and attribute we collect three types of input from three Amazon Mechanical Turk workers each for a total of $600 \times 6 \times 3 \times 3 = 32400$ data points.

First, we collect traditional binary classification labels. That is, for each face we ask the worker if it has/does not have the specific attribute. Second, for each pair of faces we collect the relative magnitude of the attribute between the two faces. We do this in a similar fashion to [19] where a worker must select one of three options:

- x has more of attribute a than y .

- y has more of attribute a than x .
- x and y have attribute a equally.

These two questions provide the traditional binary and relative ranking annotations that have been used in previous attribute studies.

For the third question, we collect spoken attribute ground truth by asking the worker to select one of six statements, spanning the space of both binary and relative attributes:

- Both x and y have attribute a .
- Neither x nor y has attribute a .
- x has attribute a but y does not.
- y has attribute a but x does not.
- x has more of attribute a than y .
- y has more of attribute a than x .

We desire that the worker selects a statement that is both natural and specific. Our statement domain is still somewhat limited because it has only six statements instead of being completely free-form. However, it expands over previous work with binary or relative attributes. We conducted a pilot experiment to explore the effects of the exact question posed to the worker on the selected spoken attribute answers. We explored three questions: “Select the statement that is [most natural, most accurate, the best description to use to search for this image]”. No significant differences were observed with respect to the question phrasing, so we used the “most accurate” variation for the remainder of our experiments. We believe that the worker-chosen statements represent the most correct way to describe the people in the image using that specific attribute. Each image is labeled three times for each type of data collection (binary/relative/accurate), and we use the majority vote for each (the few examples which resulted in three way ties were not used in the experiments). To ensure worker diversity, we allowed no worker to label more than 50 images.

We recognize that not every attribute would naturally be used to describe a specific pair of people. For example, for an image of two people who are female, it seems unnatural to comment on them (not) having a beard, as it is obvious that females don’t have beards. However, our focus is specifically on choosing the best statement for a given attribute. Other works e.g., [22] have focused on selecting which attribute to describe to produce a referring expression for a specific individual in a group.

We show detailed analysis of our collected data in Fig. 2. This figure provides a visualization of the relationship between conventional binary and relative attributes, and the workers’ spoken attributes. For each attribute (rows), and for each binary and relative attribute ground truth label (columns) we show the distribution of the type of selected spoken attribute. For example, for the teeth visible attribute, when both faces have visible teeth (the left-most cell in the teeth visible row), about 40% of the images were described

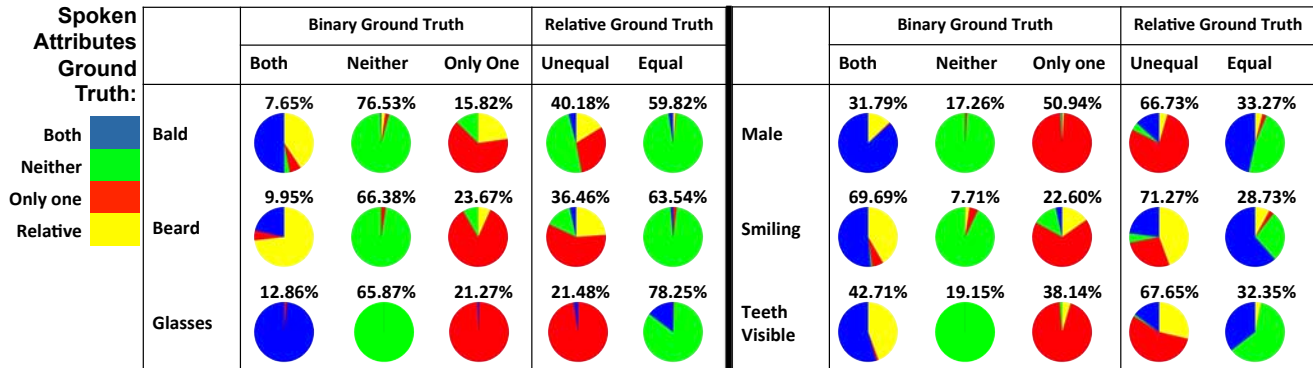


Figure 2. An analysis of our spoken attribute data collection from Sec. 3. Each row represents a specific attribute, and each column indicates the ground truth relative and binary label that is true for an image. Each pie chart represents the distribution of the statement selected by subjects as the spoken attribute. Note that each attribute is used in a different manner to describe images. For example, the glasses attribute row has no yellow, meaning that glasses was never used in the relative form in the spoken attribute experiment. In contrast, the male attribute was only used in relative form (“more masculine”) when both people were male. See text in Sec. 3 for more analysis.

by the spoken attribute “ x ’s teeth are more visible than y ’s”, while for about 55% of images the spoken attribute “both faces have visible teeth” was selected. Clearly, when not forced to make either a binary or a relative statement, our workers select different statements – sometimes relative and sometimes binary – as the spoken attribute depending on the face pair they are describing.

This chart clearly shows the differences between simply using either relative or binary statements and the spoken attribute statements that people actually prefer to say. For example, the glasses attribute never receives a relative statement as the spoken attribute (no yellow in that row), indicating that people simply do not say “ x has more glasses than y .” However, about 20% of the images receive a relative statement for glasses when using the traditional method to collect relative attribute ranking labels. This occurs in situations where one person has glasses and the other does not, and the workers are forced to select the only ranking statement that seems logically consistent. The traditional method of collecting data “forces” the worker to select a relative term even when it is not natural. This effect is visible for some of the other attributes as well (male, teeth visible). Although intuitive, this phenomenon has not been previously documented and exploited.

Another example that shows limitations of using solely binary descriptions is the portion of relative statements selected when both people have a certain attribute. Relative statements were chosen for about two-thirds of both the teeth-visible and the smiling attributes when both people in the image possess that attribute. These descriptive relative relationships cannot be expressed by simply using a binary classifier. For none of the attributes was a relative statement selected as the spoken attribute when neither of the faces had the attribute. This supports our hypothesis that relative statements cannot be used throughout the entire attribute space (as shown in Fig. 1), and conversely, shows

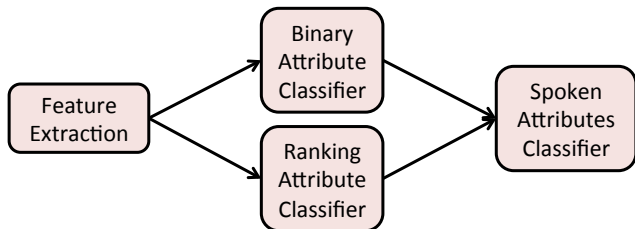


Figure 3. Our two-level classifier for predicting the spoken attribute for an image of a pair of faces.

that relative statements, when unforced, also contain information about the binary labels for the pair of exemplars.

Not all attributes have similar usage patterns. For example, glasses is never used in relative form, male is used in relative form about a third of the time both people are male, while smiling and teeth-visible are used in relative form much more often. These differences, which are not apparent from looking only at the binary or relative labels, represent certain inherent characteristics of the attribute, and show that there are different types or classes of attributes. Some are used solely in a binary sense (e.g., glasses) to indicate presence or absence of an attribute, and others are used in both the binary and relative senses, but in different proportions (smiling and teeth visible) or situations (e.g., “more masculine” is rarely used as the spoken attribute unless both faces are male).

4. Spoken Attributes Classifier

To predict the spoken attribute statement, we formalize the problem as a multiclass classification problem. The input to our classifier is an image with a pair of people, and the output is one of the six statements we presented in Sec. 3 for a given attribute. To be able to capture both the relative and absolute information about the people in the image we propose a two-level classifier as shown in Fig. 3.

Feature Extraction. We use the face pyramid features described in [10]. This is a 21504 dimension feature vector which is composed of densely sampled SIFT descriptors at each of the 21 spatial grid cells on a 3-level pyramid. The SIFT descriptors are first encoded using Locality-constrained Linear Coding (LLC) to a 1024-dimension vector and then concatenated to a 21504 length vector.

First Stage Classifiers. We build our first layer classifiers using the traditional method used previously for both binary and relative attributes. The outputs (SVM scores) of these classifiers are fed as inputs (features) to the second layer classifier, whose role is to predict the correct spoken attribute statement from among the binary and relative attribute statement choices. Although we can directly feed the low-level face pyramid features to an alternatively trained spoken attribute classifier (we compare to this method in Sec. 5), we hypothesize that extracting this semantic information in layers better aids the second layer classifier.

For our binary classifier we use a linear SVM applied to each face. We use each face’s features individually to produce two SVM decision scores for the pair of faces in the image. For the rankSVM used for modeling relative attributes, we use the algorithm provided by [19] which allows training on pairwise ranked data that includes pairs that are judged to be equal. This algorithm returns a weight vector w . By projecting each face’s features onto it and taking the difference, the relative strength of the attribute between the pair is found.

Spoken Attributes Classifier. For our second stage classifier we use a multiclass (one-vs.-one) linear SVM with the six spoken attribute statements described in Sec. 3 as the classes. We extract a 5 dimensional vector from the first stage classifiers that includes each face’s scores from the binary classifier, each face’s projection from the relative classifier, and the difference between the relative classifier scores for the pair. The binary scores indicate the belief of the presence or absence of that attribute for each specific face. The ranking scores indicate the uncalibrated absolute strength of the presence of the attribute and the difference indicates the relative attribute strength between the two faces. This short input vector makes the result easy to analyze since every number has a semantic meaning. We tried using as features a 30 dimensional vector that stores these 5 features across all 6 attributes to train the spoken attribute classifier for each attribute, but no improvement in performance was observed, so we did not include them in our final model.

5. Experiments and Results

In this Section, we present the results from a number of experiments that explore our model’s ability to “say the right thing” by predicting correct spoken attributes, and to

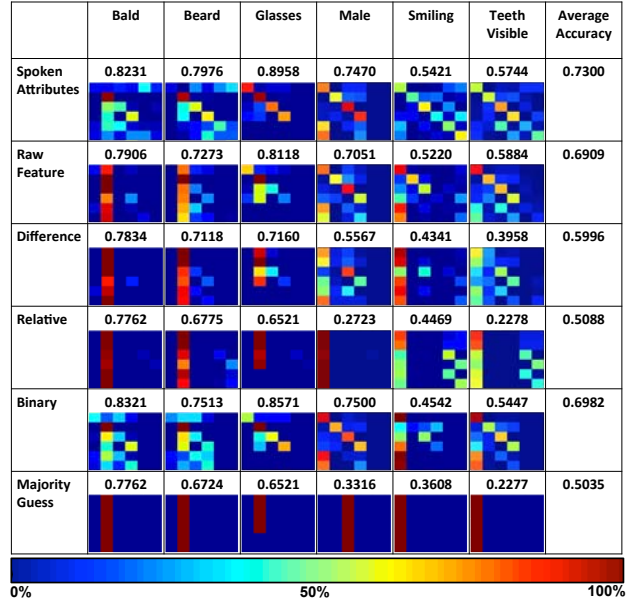


Figure 4. The results of our approach. We show the confusion matrices for each attribute for each of the baseline methods, in addition to the average accuracy across all attributes. The classes in the confusion matrix are ordered as presented in Sec. 3 (4 binary statements followed by 2 relative statements).

interpret spoken attributes to better understand the binary attributes that the images possess.

5.1. Spoken Attribute Prediction

In this first experiment our goal is to predict which one of the 6 statements a human would use to describe an image of a pair of people. In other words, we predict the spoken attribute description for the image. We compare our classifier described in Sec. 4 to the following baselines:

- **Majority.** As our spoken attribute prediction, we simply choose the attribute statement that is most commonly used as the spoken attribute for the specific attribute at hand. This represents a prior probability.
- **Binary.** We use a linear SVM trained on traditional binary labels on individual faces. The output of the SVM on a pair of faces corresponds to one of the four binary statements and is used as the spoken attribute prediction. This classifier cannot output relative statements.
- **Relative.** We use a rankSVM trained on relative labels as in [19] and output a relative statement. When the relative difference is smaller than some threshold, (i.e., nearly equal attribute levels), we output one of the binary statements that indicate agreement (either “both” or “neither” faces have the attribute, choosing the statement with the larger prior for the attribute). For each attribute we use the threshold which would produce the highest accuracy, thus unfairly benefiting this baseline. This method can not produce binary statements where one face has the attribute but the other doesn’t.

- **Difference.** We use a multiclass linear SVM on the difference of the raw LLC feature vector between the faces as input and the ground-truth “spoken attribute” statement as labels.
- **Raw Feature.** Same as above, but using a concatenation of the raw LLC features as input.

We use leave-one-out cross validation. Since we manually created pairs from people who exist in the database (see Sec. 3) we remove all images of the individuals in the test pair from the training set. For each image, we divide the remaining dataset samples in two. We use the first half to train the first layer classifiers, which we apply to the second half, and use these scores to train the second layer classifier. Finally, we test the model on the test image using both layers to get our predicted spoken attribute statement.

Results are presented in Fig. 4. These results show the strength of our two layer classifier. First, the average accuracy using our approach is higher than all other methods. Specifically, we perform better than just using a “one layer classifier” (second and third row), which just uses the input feature vector for classification. Looking at the confusion matrices reveals where the higher accuracy comes from. Using only the raw features is not able to capture the relative statements very well, and therefore our algorithm performs much better for these two classes (last two classes in the confusion matrix). This provides strong evidence of the advantage of using a two-stage classifier.

It is interesting that the binary classifier performs well on average. This is because many of the spoken attributes are mostly binary (for example glasses and male) and so good performance at selecting the correct binary statement provides the correct spoken attribute in many cases. However, this obviously fails to produce any relative attribute statements as the spoken attribute, missing a significant portion of the vocabulary that humans actually use. Our approach produces both relative and binary attribute statements as the spoken attribute and has the best overall performance.

We present examples of our predictions in Fig. 6 as compared to the **Binary** and **Relative** results. The first row of images shows examples where our algorithm correctly predicts one of the relative statements. Although **Binary** predicts “correct” results as well for these examples, it is clear our relative statements are more specific for the respective images. It is interesting to note that for many of these images the trivial majority solution is selected as the **Relative** result since the difference is below the threshold which gives the highest accuracy.

The second row presents examples where we correctly predict one of the binary statements. The **Relative** results show the effects of using the traditional ranking data collection method. For many of these images relative statements are predicted but are unsuitable. It is interesting to note that in some cases we predict the correct binary statement even

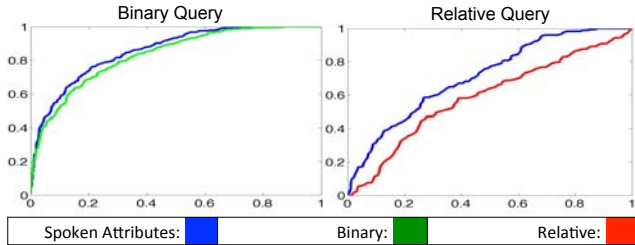


Figure 5. ROC curves for our experiment on the smiling attribute as presented in Sec. 5.2 (other attributes in supplemental material).

when the **Binary** classifier is wrong. This is due to the additional information we receive from the first layer ranker.

The final row presents examples of common errors which occur in our predictions. Since we allow relative statements (which are not very common) we end up using it inappropriately at times as compared to the binary method.

5.2. Search Results

As described in Sec. 3 one of the original questions we asked workers when presenting the six statements was “If you were searching for a photograph of a couple on Google image search that looks like the one above, what query would you use?”. Therefore, the results of our spoken attribute classifier are useful in producing relevant images to these types of search queries. For each statement we rank all the images based on their individual classifier confidence. We calculate the confidence as the number of votes received for that class from the “one vs. one” classifiers. We compare to the **Binary** and **Relative** results. For the **Relative** classifier we rank the pairs by the difference in their ranking score. For the **Binary** results, we use the product of the pair’s normalized binary SVM scores.

We show the ROC curves for the smiling attribute (other attributes produce similar results) in Figure 5. These results show that our unified framework can handle both relative and binary statements well. In addition, our *spoken attribute* classifier can better capture additional meaning which is not explicit in the statement, but is expected by the user. For example, when a user searches for an image in which person A is more bearded than person B, our algorithm will return images where both people are bearded. More ROC curves are in the supplemental materials.

5.3. Binary Classifier Prediction by Reading Between the Lines

We show that the spoken attribute statements can carry more information than what is explicitly said. For example, when the spoken attribute statement is a relative statement, it is possible to “read between the lines” and gain information about the binary presence or absence of attributes in the image. We examine the subset of the images that have a relative statement as the *spoken attribute*. We then predict the binary classifier results using three different methods. First,

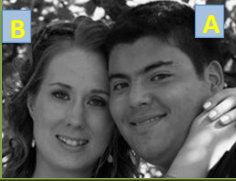
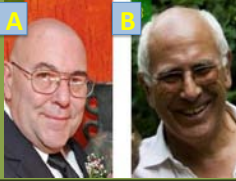

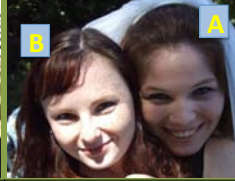





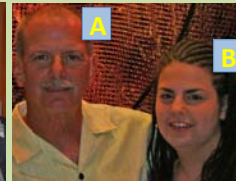




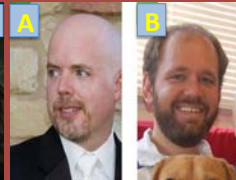
					
Binary	Both People's teeth are visible	Both People are bald	Person A is bearded and person B is not	Both People are smiling	Both People's teeth are visible
Relative	Both People's teeth are visible	Neither person is bald	Neither person is bearded	Both People are smiling	Person B's teeth are more visible than person's A's
Spoken	Person B's teeth are more visible than person's A's	Person A is more bald than person B	Person A is more bearded than person B	Person A is more smiling than person B	Person B's teeth are more visible than person's A's
					
Binary	Person B has glasses and person A does not	Neither person is smiling	Person A is male and person B is not	Neither person has glasses	Both people are male
Relative	Person B has glasses more than person A	Person A is smiling more than person B	Person A is more masculine than person B	Person B has glasses more than person A	Both people are male
Spoken	Person B has glasses and person A does not	Neither person is smiling	Person A is male and person B is not	Person B has glasses and person A does not	Person A is male and person B is not
					
Binary	Person B has a beard and person A does not	Both people are smiling	Neither person is bald	Both people's teeth are visible	Person A is bald and person B is not
Relative	Neither person is bearded	Both people are smiling	Neither person is bald	Person B's teeth are more visible than person A's	Neither person is bald
Spoken	Person B is more bearded than person A	Neither person is smiling	Person A is bald and person B is not	Both people's teeth are visible	Person A is bald and person B is not
Ground Truth	Person B has a beard and person A does not	Person A is smiling more than person B	Neither person is bald	Person B's teeth are more visible than person A's	Person A is more bald than person B

Figure 6. Qualitative examples of our prediction results in comparison to **Binary** and **Relative**. The first two image rows present correct predictions, while the bottom row presents failure cases.

we simply use the binary classifier predictions. Second, we train a new classifier with the binary classifier scores of the two faces as features. However, we train it solely on the subset of images with relative *spoken attribute* statements. This represents the information we gain by knowing the *spoken attribute*. Finally we try an additional method which simply selects the most common binary statement for that specific relative statement. This represents the information we have only from the statement, without any appearance features.

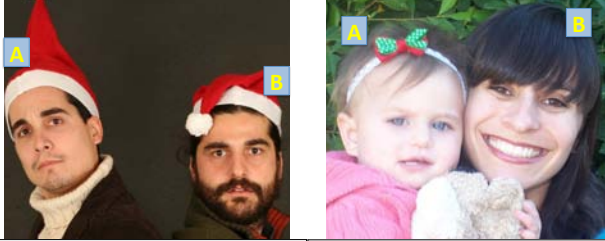
Results are shown in Table 1. We show results for attributes that have enough (> 40) images where the relative statement was the ground truth spoken attribute. A few interesting conclusions can be drawn from this table. First, it appears that adding the spoken attribute statement in some cases can significantly help the binary classifier (as is shown for teeth-visible and beard). We show examples of images

	Bald	Beard	Smiling	Teeth
Binary Classifiers	0.55	0.43	0.93	0.84
Binary Trained for SA	0.53	0.81	0.92	0.92
Most Common for SA	0	0.84	0.91	0.90

Table 1. Results of the “reading between the lines” experiment described in Sec. 5.3. Higher accuracies are better. SA = Spoken Attributes.

which were predicted correctly only after adding the *spoken attribute* statement in Fig. 7.

The bald binary prediction does not improve when adding the *spoken attribute* statement. The reason for this is evident in the results of using the bald statement alone (without appearance features). Although for most attributes, using a relative statement already implies that both people have that attribute (as in bearded and smiling), this is not true for the bald attribute. A relative “more bald”



	Beard		Teeth Visible	
	A	B	A	B
Binary Prediction	No Beard	Beard	Teeth not visible	Teeth visible
Spoken Attribute Statement	Person B is more bearded than person A		Person B's teeth are more visible than person A's	
Binary Prediction With Statement	Beard	Beard	Teeth visible	Teeth visible

Figure 7. Examples which show how having the relative spoken attribute statement contains implicit binary information and can help binary attribute classification. In each image, one people clearly has the attribute (person B) while the other has it much less (person A). The binary classifier misclassifies these images as “B has the attribute and A does not”. However, since for these attributes the relative expression is almost only used when both people have the attribute, we correct the binary classification.

statement can be made when one or even both people are not bald. Therefore, the fact that a relative statement was selected does not contain any information about the binary results, strengthening our claim that for each attribute, the model that is learned is unique to that attribute.

6. Conclusion

In this paper we show that the traditional way of modeling attributes as either binary or relative is not sufficient to fully exploit the expressiveness of the attributes, and can lead to unnatural descriptions. We present a novel way of thinking of attributes. We construct a *spoken attribute* classifier that is able to capture the interactions between the relative and binary aspects of an attribute and better model the way this attribute would be used in an actual description. Experiments show the utility of our model for image descriptions, image search and “reading between the lines.”

There are interesting ways in which this work may be extended. For example, here we model the correct way an attribute would be used in a description given that it was chosen to be discussed. However, we have not addressed the problems of choosing which attributes to mention in the first place. For example, we might not wish to talk about the beard attribute if we are presented with an image of a pair of females (unless one of them is bearded). In fact, the statement “both people do not have a beard” already implies that both people are male. We plan to investigate this problem of selecting which attributes to use, and using these correlations between attributes to read between the lines.

References

- [1] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *CVPR*, 2012.
- [2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.
- [3] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013.
- [4] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.
- [5] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010.
- [6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.
- [7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9), 2010.
- [8] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.
- [9] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.
- [10] B. G. H. Chen, A. Gallagher. What’s in a name: First names as facial attributes. In *Proc. CVPR*, 2013.
- [11] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *PAMI*, 2012.
- [12] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012.
- [13] G. Kulkarni, V. Premraj, S. L. Sagnik Dhar and, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011.
- [14] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2010.
- [15] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.
- [16] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.
- [17] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.
- [18] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.
- [19] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.
- [20] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012.
- [21] A. Sadovnik, Y. Chiu, N. Snavey, S. Edelman, and T. Chen. Image description with a goal: Building efficient discriminating expressions for images. In *CVPR*, 2012.
- [22] A. Sadovnik, A. Gallagher, and T. Chen. Its not polite to point: Describing people with uncertain attributes. In *CVPR*, 2013.
- [23] W. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2013.
- [24] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.
- [25] M. Spain and P. Perona. Measuring and predicting object importance. *IJCV*, 91(1), 2011.
- [26] N. Turakhia and D. Parikh. Attribute dominance: What pops out? In *ICCV*, 2013.
- [27] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009.
- [28] G. Wang, D. Forsyth, and D. Hoiem. Comparative object similarity for improved recognition with few or no examples. In *CVPR*, 2010.
- [29] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.
- [30] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011.