Abstract

When we look at an image, some properties of the image stand out more than others. When describing an image, people are likely to describe these dominant attributes first. Attribute dominance is a result of a complex interplay between the various properties present or absent in the image. Which attributes in an image are more dominant than others reveals rich information about the content of the image. In this paper we tap into this information by modeling attribute dominance. We show that this helps improve the performance of vision systems on a variety of human-centric applications such as zero-shot learning, image search and generating textual descriptions of images.

1. Introduction

When we look at an image, some properties of the image pop out at us more than others. In Figure 1 (a), we are likely to talk about the puppy as being white and furry. Even though the animal in Figure 1 (b) is also white and furry, that is not what we notice about it. Instead, we may notice its sharp teeth. This is also true at the level of categories. We are likely to talk about bears being furry but wolves being fierce even though wolves are also furry. While all attributes are – by definition – semantic visual concepts we care about, different attributes dominate different images or categories. The same attribute present in different images or categories may dominate in some but not in others.

An attribute may be dominant in a visual concept due to a variety of reasons such as strong presence, unusualness, absence of other more dominant attributes, etc. For example, Figure 1 (c) depicts a person with a very wide smile with her teeth clearly visible. Figure 1 (d) is a photograph of a person wearing very bright lipstick. Hence smiling and wearing lipstick are dominant in these two images respectively. It is relatively uncommon for people to have a beard and wear glasses, making these attributes dominant in Figure 1 (f). When neither of these cases are true, attributes that are inherently salient (e.g. race, gender, etc. for people) is what one would use to describe an image or category (Figure 1 (e)) and turn out to be dominant. Correlations among attributes can also affect dominance. For instance, bearded people are generally male, and so “not bearded” is unlikely to be noticed or mentioned for a female. In general, attribute dominance is different from the relative strength of an attribute in an image. Relative attributes [24] compare the strength of an attribute across images. Attribute dominance compares the relative importance of different attributes within an image or category. Attribute dominance is an image- or category-specific phenomenon – a manifestation of a complex interplay among all attributes present (or absent) in the image or category.

Why should we care about attribute dominance? Because attribute dominance affects how humans perceive and describe images. Humans are often users of a vision system as in image search where the user may provide an attribute-based query. Humans are often supervisors of a vision system as in zero-shot learning where the human teaches the machine novel visual concepts simply by describing them in terms of its attributes. Attribute dominance affects which attributes humans tend to name in these scenarios, and in which order. Since these tendencies are image- and category-specific, they reflect information about the visual content – they provide identifying information about an im-
2. Related Work

We now describe existing works that use attributes for image understanding. We elaborate more on approaches geared specifically towards the three applications that we evaluate our approach on: zero-shot learning, image search and automatically generating textual descriptions of images. We also relate our work to existing works on modeling saliency and importance in images, as well as reading between the lines of what a user explicitly states.

Attributes: Attributes have been used extensively, especially in the past few years, for a variety of applications [2, 4, 7, 8, 11, 16, 18, 20, 23–25, 31–33]. Attributes have been used to learn and evaluate models of deeper scene understanding [8, 20] that reason about properties of objects as opposed to just the object categories. Attributes can provide more effective active learning by allowing the supervisor to provide attribute-based feedback to a classifier [25], or at test time with a human-in-the-loop answering relevant questions about a test image [4]. Attributes have also been explored to improve object categorization [8] or face verification performance [19]. Attributes being both machine detectable and human understandable provide a mode of communication between the two. In our work, by modeling the dominance of attributes in images, we enhance this channel of communication. We demonstrate the resultant benefits on three human-centric computer vision applications that we expand on next: zero-shot learning, image search and generating textual descriptions of image.

Zero-shot learning: Attributes have been used for alleviating human annotation efforts via zero-shot learning [8, 20, 24] where a supervisor can teach a machine a novel concept simply by describing its properties and without having to provide example images of the concept. For instance, a supervisor can teach a machine about zebras by describing them as being striped and having four legs. Works have looked at allowing supervisors to provide more fine-grained descriptions such as “zebras have shorter necks than giraffes” to improve zero-shot learning [24]. Our work takes an orthogonal perspective: rather than using a more detailed (and hence likely more cumbersome) mode of supervision, we propose to model attribute dominance to extract more information from existing modes of supervision.

Image search: Attributes have been exploited for image search by using them as keywords [18, 28] or as intermediate mid-level semantic representations [5, 21, 26, 29, 34, 36] to reduce the well known semantic gap. Statements about relative attributes [24] can be used to refine search results [16]. Modeling attribute dominance allows us to inject the user’s subjectivity into the search results without explicitly eliciting feedback or more detailed queries.

Textual descriptions: Attributes have been used for automatically generating textual description of images [17, 24] that can also point out anomalies in objects [8]. Efforts are also being made to predict entire sentences from image features [7, 17, 22, 35]. Some methods generate novel sentences for images by leveraging existing object detectors [10], attributes predictors [2, 8, 24], language statistics [35] or spatial relationships [17]. Sentences have also been assigned to images by selecting a complete written description from a large set [9, 22]. Our work is orthogonal to these directions. Efforts at generating natural language sentences can benefit from our work that determines which attributes ought to be mentioned in the first place, and in what order.

Saliency: Dominance is clearly related to saliency. A lot of works [6, 14, 15] have looked at predicting which regions of an image attract human attention, i.e., humans fixate on. We look at the distinct task of modeling which high-level semantic concepts (i.e. attributes) attract human attention.

Importance: A related notion of importance has also been examined in the community. The order in which people are likely to name objects in an image was studied in [30]. We study this for attributes. Predicting which objects, attributes, and scenes are likely to be described in an image has recently been studied [1]. We focus on the constrained problem of predicting which attributes are dominant or important. Unlike [1], in addition to predicting which attributes are likely to be named, we also model the order in which the attributes are likely to be named. Note that attribute dominance is not meant to capture user-specific preferences (e.g. when searching for a striped dark blue tie, a particular user may be willing to compromise on the stripes but not on the dark blue color). Similar to saliency and importance, while there may be some user-specific aspects to attribute dominance, we are interested in modeling the user-independent signals.

Reading between the lines: We propose exploiting the order in which humans name attributes when describing an image. This can be thought of as reading between the lines of what the user is saying, and not simply taking the description – which attributes are stated to be present or absent –
3. Approach

We first describe how we annotate attribute dominance in images (Section 3.1). We then present our model for predicting attribute dominance in a novel image (Section 3.2). Finally, we describe how we use our attribute dominance predictors for three applications: zero-shot learning (Section 3.3), image search (Section 3.4) and textual description (Section 3.5).

3.1. Annotating Attribute Dominance

We annotate attribute dominance in images to train our attribute dominance predictor, and use it as ground truth at test time to evaluate our approach. We conduct user studies on Amazon Mechanical Turk to collect the annotations. We collect dominance annotations at the category-level, although our approach trivially generalizes to image-level dominance annotations as well.

We are given a vocabulary of $M$ binary attributes \( \{ a_m \}, m \in \{ 1, \ldots, M \} \), images from $N$ categories \( \{ C_n \}, n \in \{ 1, \ldots, N \} \) and the ground truth presence or absence of each attribute in each category: \( g_m^n = 1 \) if attribute \( a_m \) is present in category \( C_n \), otherwise \( g_m^n = 0 \).

We show subjects example images from a category \( C_n \), along with a pair of attributes \( a_m \) and \( a_{m'} \), \( m', m' \in \{ 1, \ldots, M \} \). Without loss of generality, let us assume that both attributes are present in the image i.e. \( g_m^n = 1 \) and \( g_{m'}^n = 1 \). If the user had to describe the category using one of these two attributes "\( a_m \) is present" or "\( a_{m'} \) is present", we want to know which one (s)he would use i.e. which one is more dominant. For quality control, instead of showing subjects only the two options that correspond to \( g_m^n \), we show them all four options including in this case "\( a_m \) is absent" and "\( a_{m'} \) is absent" (see Figure 2). If a worker picks a statement that is inconsistent with \( g_m^n \) on several occasions, we remove his responses from our data.

Each category is shown with all \( \binom{M}{2} \) pairs of attributes. Each question is shown to 6 different subjects.

Note that absence of an attribute e.g. "does not have eyebrows" can also be dominant. Since the presence of an attribute may be dominant in some images, but its absence may be dominant in others, we model them separately. This effectively leads to a vocabulary of \( 2M \) attributes \( \{ a_m^1, a_m^0 \}, m \in \{ 1, \ldots, M \} \), where \( a_m^1 \) corresponds to \( a_m = 1 \) i.e. attributes \( a_m \) is present, and \( a_m^0 \) corresponds to \( a_m = 0 \). For ease of notation, from here on, we replace \( a_m^1 \) and \( a_m^0 \) with just \( a_m \), but let \( m \in \{ 1, \ldots, 2M \} \). We refer to this as the expanded vocabulary.

The dominance \( d_m^a \) of attribute \( a_m \) in category \( C_n \) is defined to be the number of subjects that selected \( a_m \) when it appeared as one of the options. Each attribute appears as an option \( M \) times. So we have

\[
d_m^a = \sum_{s=1}^{M} \sum_{o=1}^{S} \mathbb{I}[m_s^o]
\]

where \( S \) is the number of subjects (6 in our case), \( \mathbb{I} \) is 1 if the argument is true, and \( \mathbb{I}[m_s^o] \) indicates that subject \( s \) selected attribute \( a_m \) the \( o^{th} \) time it appeared as an option.

We now have the ground truth dominance value for all \( 2M \) attributes in all \( N \) categories. We assume that when asked to describe an image using \( K \) attributes, users will use the \( K \) most dominant attributes. This is consistent with the instructions subjects were given when collecting the annotations (see Figure 2). The data we collected is publicly available on the authors’ webpage. We now describe our approach to predicting dominance of an attribute in a novel image.
3.2. Modeling Attribute Dominance

Given a novel image $x_t$, we predict the dominance $d^m_t$ of attribute $m$ in that image using

$$d^m_t = \mathbf{w}_m^T \phi(x_t) \quad (2)$$

We represent image $x_t$ via an image descriptor. We use the output scores of binary attribute classifiers to describe the image. This exposes the complex interplay among attributes discussed in the introduction that leads to the dominance of certain attributes in an image and not others. The relevant aspects of the interplay are learnt by our model. $\phi(x_t)$ can be just $x_t$ or an implicit high-(potentially infinite-) dimensional feature map implied by a kernel.

For training, we project the category-level attribute dominance annotations to each training image. If we have $P$ training images $\{x_p, y_p\}, p \in \{1, \ldots, P\}$, along with their class label indices $\{y_p\}, y_p \in \{1, \ldots, N\}$, the dominance of attribute $m$ in image $x_p$ is $d^m = d^m_p$. This gives us image and attribute dominance pairs $\{(x_p, d^m_p)\}$ for each attribute $m$. Using these pairs as supervision, we learn $\mathbf{w}_m$ using a regressor that maps $x_p$ to $d^m_p$. We experimented with a linear and RBF kernel. Linear regression performed better and was used in all our experiments.

The learnt parameters $\mathbf{w}_m$ allow us to predict the dominance value of all attributes in a new image $x_t$ (Equation 2). We sort all $2M$ attributes in descending order of their dominance values $d^m_t$. Let the rank of attribute $m$ for image $x_t$ be $r^m(x_t)$. Then the probability $pd^m_k(x_t)$ that attribute $m$ is the $k^{th}$ most dominant image $x_t$ is computed as

$$pd^m_k(x_t) = \frac{s^m_k(x_t)}{s^m_k(x_t)} \quad (3)$$

and

$$s^m_k(x_t) = \frac{1}{\log(|r^m(x_t) - k| + 1) + 1} \quad (4)$$

$s^m_k(x_t)$ is a score that drops as the estimated rank $r^m(x_t)$ of the attribute in terms of its dominance in the image is further away from $k$. Equation 3 simply normalizes these scores across $k$ to make it a valid distribution i.e. each attribute is one of the $2M$ most dominant in an image, since there are only $2M$ attributes in the vocabulary. From here on we drop the subscript $t$ for a novel test image.

Note that although the dominance of each attribute is predicted independently, the model is trained on an attribute-based representation of the image. This allows the model to capture correlations among the attributes. More sophisticated models and features as explored in [12] can also be incorporated. As our experiments demonstrate, even our straight forward treatment of attribute dominance results in significant improvements in performance in a variety of human centric applications. We describe our approach to these applications next.

3.3. Zero-shot Learning

In zero-shot learning [20], the supervisor describes novel $N'$ previously unseen categories in terms of their attribute signatures $\{g^m_n\}, n' \in \{1, \ldots, N'\}$. With a pre-trained set of $M$ binary classifiers for each attribute and Lambert et al.’s [20] Direct Attribute Prediction (DAP) model, the probability that an image $x$ belongs to each of the novel categories $C_{n'}$ is

$$pa_{n'}(x) \propto \prod_{m=1}^M pa^m(x) \quad (5)$$

where $pa^m(x)$ is the probability that attribute $a_m$ takes the value $g^m_n \in \{0, 1\}$ in image $x$ as computed using the binary classifier for attribute $a_m$. The image is assigned to the category with the highest probability $pa_{n'}(x)$. This approach forms our baseline. It relies on an interface where a supervisor goes through every attribute in a pre-defined arbitrary order and indicates its presence or absence in a test category. We argue that this is not natural for humans.

People are likely to describe a zebra as a horse-like animal with stripes, an elephant as a grey large animal with a trunk and tusks, and a hippopotamus as a round animal often found in or around water. It is much more natural for humans to describe categories using only a subset of attributes. These subsets are different for each category. Moreover, even within the subsets people consistently name some attributes before others (more on this in the results section). Our approach allows for this natural interaction. More importantly, it exploits the resultant patterns revealed in human behavior when allowed to interact with the system naturally, leading to improved classification of a novel image.\footnote{Recall, our vocabulary of $2M$ attributes is over-complete and redundant since it includes both the presence and absence of attributes. The supervisor only needs to specify half the attribute memberships.}

It assumes that since being striped is a dominant attribute for zebras, a test image is more likely to be a zebra if it is striped \textit{and} being striped is dominant in that image.

Let’s say the supervisor describes the category $C_{n'}$ using $K$ attributes in a particular order $\{g^m_{n'_1}, \ldots, g^m_{n'_k}, \ldots, g^m_{n'_{K'}}\}, m_k \in \{1, \ldots, 2M\}$. To determine how likely an image is to belong to class $C_{n'}$, our approach not only verifies how well its appearance matches the specified attributes presence / absence, but also verifies how well the predicted ordering of attributes according to their dominance matches the order of attributes used by the supervisor when describing the test category. We compute the probability of an image $x$ belonging to a class $C_{n'}$ as:

$$p_{n'}(x) = pa_{n'}(x)pd_{n'}(x) \quad (6)$$

where $pa_{n'}(x)$ is the appearance term computed using Equation 5 and the dominance term $pd_{n'}(x)$ is
\[ \frac{pd_n(x)}{\prod_{k=1}^{K} \frac{p_d^m(x)}{\prod_{k=1}^{K} \frac{p_a^m(x)}{\prod_{k=1}^{K} \frac{p_d^m(x)}{\prod_{k=1}^{K} pd^m(x)}}} \] (7)

\[ pd_k^m(x) \] is the probability that attribute \(a_{m_k}\) is the \(k^{th}\) most dominant attribute in image \(x\) and is computed using Equations 3 and 4. The test instance is assigned to the category with the highest probability \(p_n(x)\). In our experiments we report results for varying values of \(K\).

### 3.4. Image Search

We consider the image search scenario where a user has a target category in mind, and provides as query a list of attributes that describe that category. It is unlikely that the user will provide the values of all \(M\) attributes when describing the query. (S)he is likely to use the attributes dominant in the target concept, naming the most dominant attributes first.

In our approach, the probability that a target image satisfies the given query depends on whether its appearance matches the presence/absence of attributes specified, and whether the predicted dominance of attributes in the image satisfies the order used by the user in the query. If the user used \(K\) attributes to describe his/her query \((g_1, \ldots, g_{m^*}, \ldots, g_{n^*})\) the probability that \(x\) is the target image is computed as:

\[ p(x) \propto \prod_{k=1}^{K} p_d^m(x)pd^m(x) \] (8)

All images in the database are sorted in descending order of \(p(x)\) to obtain the retrieval results for a given query. The approach of Kumar et al. [18] corresponds to ignoring the \(pd^m(x)\) term from the above equation, and using the appearance term alone, which forms our baseline approach. Again, we report results for varying values of \(K\).

### 3.5. Textual Description

The task at hand is to describe a new image in terms of the attributes present / absent in it. Again, if humans are asked to describe an image, they will describe some attributes before others, and may not describe some attributes at all. If a machine is given similar abilities, we expect the resultant description to characterize the image better than an approach that lists attributes in an arbitrary order [8] and chooses a random subset of \(K\) out of \(M\) attributes to describe the image [24].

Given an image \(x\), we compute \(\tilde{d}^m\) using Equation 2. We sort all attributes in descending order of their predicted dominance score for this image. If the task is to generate a description with \(K\) attributes, we pick the top \(K\) attributes from this ranked list to describe the image. We report results with varying values of \(K\). Note that since dominance is predicted for the expanded vocabulary, the resultant descriptions can specify the presence as well as absence of attributes.

### 4. Results

We first describe the datasets we experimented with. We then provide an analysis of the dominance annotations we collected to gain better insights into the phenomenon and validate our assumptions. We then describe our experimental setup and report results on the three applications described above.

#### 4.1. Datasets

We experimented with two domains: faces and animals. For faces, we used 10 images from each of the 200 categories in the Public Figures Face Database (PubFig) [19]. We worked with a vocabulary of 13 attributes (26 in the expanded vocabulary including both presence and absence of attributes). These attributes were selected to ensure (1) a variety in their presence / absence across the categories and (2) ease of use for lay people on MTurk to comment on. We combined some of the attributes of [19] into one e.g. mustache, beard and goatee were combined to form facial hair. We used the pre-trained attribute classifiers provided by Kumar et al. [19] as our appearance based attribute classifiers. We used 180 categories for training, and 20 for testing. We report average results of 10-fold cross validation.

For animals, we used the Animals with Attributes dataset (AWA) [20] containing a total of 30475 images from 50 categories. We worked with a vocabulary of 27 attributes (54 in expanded vocabulary). These were picked to ensure that lay people on MTurk can understand them. We used the pre-trained attribute classifiers provided by Lampert et al. [20]. These attributes were trained on 21866 images from 40 categories. We used a held out set of 2429 validation images from those 40 categories to train our dominance predictor. We tested our approach on 6180 images from 10 previously unseen categories (as did Lampert et al. [20]).

We collected attribute dominance annotation for each attribute across all categories as described in Section 3.1. We represent each image with the outputs of all 73 and 85 attribute classifiers provided by Kumar et al. [19] and Lam...
consistently dominance predictors instead of tend to name some at-

et al. [20], especially for the PubFig term in Equation 6). To demonstrate the need toattribute dominance predictor described in Section 3.2. The ground truth is 0.68. The rank correlation between a fixed
dom ordering of attributes with the dominance score is 0.93. The rank correlation between our predicted dominance score and the
correlation between our predicted dominance score and the
ground truth is 0.93. The rank
the dominance score obtained using responses from half the
The inter-human rank correlation computed by comparing
the attributes in the categories. The Spearman rank corre-
score as well as using the ground truth relative strength of
attributes are in common with their 29. For a given category,
ent attributes that are dominant in them (2) Even when the
same attribute is present in different categories, it need not be

dominant in all of them. For instance, “Has tough skin” is
present in 23 animal categories but has high dominance values in only 12 of them. (3) Absence of attributes can on
occasion be dominant. For instance, since most animals
walk on four legs, animals who don’t walk on four legs have
“Does not walk on four legs” as a dominant attribute.

To analyze whether dominance simply captures the rela-
tive strength of an attribute in an image, we compare the
ground truth dominance of an attribute across categories
with relative attributes [24]. Relative annotations for 29 at-
tributes in 60 categories in the development set of the Pub-
Fig dataset [19] were collected in [3]. Six of our 13 at-
tributes are in common with their 29. For a given category,
we sort the attributes using our ground truth dominance
score as well as using the ground truth relative strength of
the attributes in the categories. The Spearman rank corre-
lation between the two was found to be 0.46. To put this
number in perspective, the rank correlation between a ran-
dom ordering of attributes with the dominance score is 0.01.
The inter-human rank correlation computed by comparing
the dominance score obtained using responses from half the
subjects with the scores from the other half is 0.93. The rank
correlation between our predicted dominance score and the
ground truth is 0.68. The rank correlation between a fixed

4.2. Dominance Analysis

In Figure 3 we show the ground truth dominance scores
of all attributes (expanded vocabulary) in all categories as
computed using Equation 1. We also show the ground truth
attribute presence / absence of the attributes. We make three
observations (1) Different categories do in fact have differ-
ent attributes that are dominant in them (2) Even when the
same attribute is present in different categories, it need not be
dominant in all of them. For instance, “Has tough skin” is
present in 23 animal categories but has high dominance values in only 12 of them. (3) Absence of attributes can
on occasion be dominant. For instance, since most animals
walk on four legs, animals who don’t walk on four legs have
“Does not walk on four legs” as a dominant attribute.

We compare our approach of using appearance and dom-
inance information both (Equation 6) to the baseline ap-
proach of Lampert et al. [20] that uses appearance informa-
tion alone (Equation 5). We also compare to an approach
that uses dominance information alone (i.e. uses only the
\( pd_{\omega}(x) \) term in Equation 6). To demonstrate the need
to model dominance of attribute presence and absence sep-
arately, we report results using a compressed vocabulary
where the ground truth dominance score (Equation 1) of
the presence and absence of an attribute is combined (sum), and
we learn only \( M \) dominance predictors instead of \( 2M \).
The results are shown in Figures 4a and 4d. Since AW A has
a pre-defined train/test split, we can report results only on
one split. The baseline curve is noisy across different values
of \( K \). This is because not all attribute predictors are equally
accurate. If the prediction accuracy of an attribute is poor,
it can reduce the overall appearance-only zero-shot learning
performance. This leads to lower accuracy after \( K > 20 \).
Note that our approach is significantly more stable. We see
that the incorporation of dominance can provide a notable
boost in performance compared to the appearance-only ap-
proach of Lampert et al. [20], especially for the PubFig
data set. We also see that the expanded vocabulary for mod-
eling dominance performs better than the compressed ver-
sion. To evaluate the improvement in performance possible
by improved modeling of dominance, we perform zero-shot
learning using the responses of half the subjects to com-
pute the ground truth dominance score and responses from
the other half to compute the “predicted” dominance score,
4.4. Image Search

To run our experiments automatically while still using queries generated by real users, we collected the queries for all possible target categories offline (Figure 2). When experimenting with a scenario where the user provides queries containing $K$ attributes, for each target, we use the $K$ attributes selected most often by the users to describe the target category (Equation 1). As the evaluation metric, we use the log of the rank of the true target category when images in the dataset are sorted by our approach (Section 3.4) or the baselines. Lower is better. We compare to the same baseline, global dominance based (same attributes for all categories), global dominance based (same attributes for all categories) most dominant attributes using the ground truth dominance score of attributes (Equation 1). The results are shown in Figures 4c and 4f. We compare to a strong baseline (global) that always predicts the same $K$ attributes for all images. These are the $K$ attributes that are on average (across all training categories) most dominant. We also compare to an approach that predicts $K$ random attributes for an image. To make the baselines even stronger, we first predict the presence / absence of attributes in the image using attribute classifiers, and then pick $K$ attributes from those randomly or using the compressed dominance regressor. We see that our approach significantly outperforms these baselines. Our improved performance over the global baseline demonstrates that our approach reliably captures image-specific dominance patterns. We also report inter-human agreement as an upper-bound performance for this task.

The second evaluation task consists of human studies. We presented the three descriptions: dominance-based (our approach), global dominance based (same attributes for all images) and random, along with the image being described to human subjects on Amazon Mechanical Turk. We asked them which description is the most appropriate. We conducted this study using 200 images for PubFig and 50 images for AwA with 10 subjects responding to each image. For PubFig & AwA, subjects preferred our description 73%, while still using trained attribute classifiers for appearance. At the highest value of $K$, PubFig achieves 69% accuracy and AwA achieves 68% accuracy. We see that better prediction of dominance values would lead to a huge improvement in accuracies. Note that for a fixed value of $K$ (x-axis), different categories use their respective $K$ most dominant attributes that a user is likely to list, which are typically different for different categories. Our accuracies on the AwA dataset are not directly comparable to the numbers in Lampert et al. [20] because we use only 27 attributes instead of 85 used in [20]. We see that by incorporating dominance, we achieve 83.7% of their performance while using only 31.7% of the attributes.

4.5. Textual Description

We evaluate the textual descriptions generated by our approach in two ways. In the first case, we check what percentage of the attributes present in our descriptions are also present in the ground truth descriptions of the images. The ground truth descriptions are generated by selecting the $K$ most dominant attributes using the ground truth dominance score of attributes (Equation 1). The results are shown in Figures 4e and 4f. We compare to a strong baseline (global) that always predicts the same $K$ attributes for all images. These are the $K$ attributes that are on average (across all training categories) most dominant. We also compare to an approach that predicts $K$ random attributes for an image. To make the baselines even stronger, we first predict the presence / absence of attributes in the image using attribute classifiers, and then pick $K$ attributes from those randomly or using the compressed dominance regressor. We see that our approach significantly outperforms these baselines. Our improved performance over the global baseline demonstrates that our approach reliably captures image-specific dominance patterns. We also report inter-human agreement as an upper-bound performance for this task.

The second evaluation task consists of human studies. We presented the three descriptions: dominance-based (our approach), global dominance based (same attributes for all images) and random, along with the image being described to human subjects on Amazon Mechanical Turk. We asked them which description is the most appropriate. We conducted this study using 200 images for PubFig and 50 images for AwA with 10 subjects responding to each image. For PubFig & AwA, subjects preferred our description 73%
& 64% of the times as compared to global (22% & 28%) and random (5% & 8%). Clearly, modeling attribute dominance leads to significantly more natural image descriptions. We repeated this study, but this time with ground truth dominance and ground truth presence / absence of attributes. For PubFig & AWA, subjects preferred our description 73% & 84% of the times as compared to global (25% & 16%) and random (2% & 0%). This validates our basic assumption that users use dominant attributes when describing images. This is not surprising because we collected the dominance annotations by asking subjects which attributes they would use to describe the image (Figure 2).

5. Conclusion and Future Work

In this paper we make the observation that some attributes in images pop out at us more than others. When people naturally describe images, they tend to name a subset of all possible attributes and in a certain consistent order that reflects the dominance of attributes in the image. We propose modeling these human tendencies, i.e., attribute dominance and demonstrate resultant improvements in performance for human-centric applications of computer vision such as zero-shot learning, image search and automatic generation of textual descriptions of images in two domains: faces and animals.

Future work involves incorporating the notion of dominance for relative attributes [24]. Relative attributes allow users to provide feedback during image search [16] or while training an actively learning classifier [25]. When the user says “I want shoes that are shinier than these” or “This image is not a forest because it is too open to be a forest”, perhaps users name attributes that are dominant in the images. Incorporating this when updating the search results or re-training the classifier may prove to be beneficial. Moreover, when collecting pairwise annotations for relative attributes where a supervisor is asked “does the first image have more/less/equal amount of attribute $a_m$ than the second image?”, the responses from human subjects may be more consistent if we ensure that the two images being compared have equal dominance of attribute $a_m$.

References