Image Specificity

Mainak Jas
Aalto University
mainak.jas@aalto.fi

Devi Parikh
Virginia Tech
parikh@vt.edu

Abstract

For some images, descriptions written by multiple people are consistent with each other. But for other images, descriptions across people vary considerably. In other words, some images are specific – they elicit consistent descriptions from different people – while other images are ambiguous. Applications involving images and text can benefit from an understanding of which images are specific and which ones are ambiguous. For instance, consider text-based image retrieval. If a query description is moderately similar to the caption (or reference description) of an ambiguous image, that query may be considered a decent match to the image. But if the image is very specific, a moderate similarity between the query and the reference description may not be sufficient to retrieve the image.

In this paper, we introduce the notion of image specificity. *We present two mechanisms to measure specificity given multiple descriptions of an image: an automated measure and a measure that relies on human judgement. We analyze image specificity with respect to image content and properties to better understand what makes an image specific. We then train models to automatically predict the specificity of an image from image features alone without requiring textual descriptions of the image. Finally, we show that modeling image specificity leads to improvements in a text-based image retrieval application.*

1. Introduction

Consider the two photographs in Figure 1. How would you describe them? For the first, phrases like "people lined up in terminal", "people lined up at train station", "people waiting for train outside a station", *etc*. come to mind. It is clear what to focus on and describe. In fact, different people talk about similar aspects of the image – the train, people, station or terminal, lining or queuing up. But for the photograph on the right, it is less clear how it should be described. Some people talk about the the sunbeam shining through the skylight, while others talk about the alleyway, or the people selling products and walking. In other words, the photograph on the left is *specific* whereas the photograph on the right is *ambiguous*.

"people lined up in terminal"
"people lined up at train station"
"long line at a station"
"people waiting for train outside a station"

"alleyway in a small town"
"People sitting and walking"
"man walking in shopping area with others selling products"
"sunbeam shining through skylight"

Figure 1. Some images are *specific* – they elicit consistent descriptions from different people (left). Other images (right) are ambiguous.

The computer vision community has made tremendous progress on recognition problems such as object detection [12, 16], image classification [26], attribute classification [48] and scene recognition [50, 52]. Various approaches are moving to higher-level semantic image understanding tasks. One such task that is receiving increased attention in recent years is that of automatically generating textual descriptions of images [5, 8, 13, 14, 23, 25, 28, 31, 35, 37, 38, 47, 51] and evaluating these descriptions [11, 32, 38, 39]. However, these works have largely ignored the *variance in descriptions* produced by different people describing each image. In fact, early works that tackled the image description problem [14] or reasoned about what image content is important and frequently described [3] claimed that human descriptions are consistent. We show that there is in fact variance in how consistent multiple human-provided descriptions of the same image are. Instead of treating this variance as noise, we think of it as a useful signal that if modeled, can benefit applications involving images and text.

We introduce the notion of *image specificity* which measures the amount of variance in multiple viable descriptions of the same image. Modeling image specificity can benefit a variety of applications. For example, computer-generated image description and evaluation approaches can benefit from specificity. If an image is known to be am-

biguous, several different descriptions can be generated and be considered to be plausible. But if an image is specific, a narrower range of descriptions may be appropriate. Photographers, editors, graphics designers, *etc*. may want to pick specific images – images that are likely to have a single (intended) interpretation across viewers.

Given multiple human-generated descriptions of an image, we measure specificity using two different mechanisms: one requiring human judgement of similarities between two descriptions, and the other using an automatic textual similarity measure. Images with a high average similarity between pairs of sentences describing the image are considered to be specific, while those with a low average similarity are considered to be ambiguous. We then analyze the correlation between image specificity and image content or properties to understand what makes certain images more specific than others. We find that images with people tend to be specific, while mundane images of generic buildings or blue skies do not tend to be specific. We then train models that can predict the specificity of an image just by using image features (without associated human-generated descriptions). Finally, we leverage image specificity to improve performance in a real-world application: text-based image retrieval.

2. Related work

Image properties: Several works study high-level image properties beyond those depicted in the image content itself. For instance, unusual photographs were found be interesting [17] and images of indoor scenes with people were found to be memorable, while scenic, outdoor scenes were not [18, 19]. Other properties of images such as aesthetics [7], attractiveness [30], popularity [24], and visual clutter [42] have also been studied[1]. In this paper, we study a novel property of images – specificity – that captures the degree to which multiple human-generated descriptions of an image vary. We study what image content and properties make images specific. We go a step further and leverage this new property to improve a text-based image retrieval application.

Importance: Some works have looked at what is worth describing in an image. Bottom-up saliency models [20, 22] study which image features predict eye fixations. Importance [3, 44] characterizes the likelihood that an object in an image will be mentioned in its description. Attribute dominance [45] models have been used to predict which attributes pop out and the order in which they are likely to be named. However, unlike most of these works, we look at the *variance* in human perception of what is worth mentioning in an image and how it is mentioned.

[1]Our work is complementary to visual metamers [15]. In visual metamers, different images are perceived similarly but in specificity, we study how the same image can be perceived differently, and how this variance in perception differs across images.

Image description: Several approaches have been proposed for automatically describing images. This paper does not address the task of generating descriptions. Instead, it studies a property of how humans describe images – some images elicit consistent descriptions from multiple people while others do not. This property can benefit image description approaches. Some image description approaches are data-driven. They retrieve images from a database that are similar to the input image, and leverage descriptions associated with the retrieved images to describe the input image [14, 38]. In such approaches, knowledge of the specificity of the input image may help guide the range of the search for visually similar images. If the input image is specific, perhaps only highly similar images and their associated descriptions should be used to construct its description. Other approaches analyze the content of the image and then compose descriptive sentences using knowledge of sentence structures [28, 37]. For images that are ambiguous, the model can predict multiple diverse high-scoring descriptions of the image that can all be leveraged for a downstream application. Finally, existing automatic image description evaluation metrics such as METEOR [1], ROUGE [32], BLEU [39] and CIDEr [46] compare a generated description with human-provided reference descriptions of the image. This evaluation protocol does not account for the fact that some images have multiple viable ways in which they can be described. Perhaps the penalty for not matching reference descriptions of ambiguous images should be less than for specific ones.

Image retrieval: Query- or text-based image and video retrieval approaches evaluate how well a query matches the content of [2, 10, 34, 43] or captions (descriptions) associated with [2, 29] images in a database. However, the fact that each image may have a different match score or similarity that is sufficient to make it relevant to a query has not been studied. In this work, we use image specificity to fill this gap. While the role of lexical ambiguity in information retrieval has been studied before [27], reasoning about inherent ambiguity in images for retrieval tasks has not been explored.

3. Approach

We first describe the two ways in which we measure the specificity of an image. We then describe how we use specificity in a text-based image retrieval application.

3.1. Measuring Specificity

We define the specificity of an image as the average similarity between pairs of sentences describing the image. For each image i, we are given a set S^i of N sentence descriptions $\{s_1^i, \ldots, s_N^i\}$. We measure the similarity between all possible $\binom{N}{2}$ pairs of sentences and average the scores. The similarity between two sentences can either be judged by humans or computed automatically.

Figure 2. Example images with very low to very high human-annotated specificity scores.

3.1.1 Human Specificity Measurement

M different subjects on Amazon Mechanical Turk (AMT) were asked to rate the similarity between a pair of sentences s_a^i and s_b^i, on a scale of 1 (very different) to 10 (very similar). Note that subjects were not shown the corresponding image and were not informed that the sentences describe the same image. This ensured that subjects rated the similarity between sentences based solely on their textual content. We shift and scale the similarity scores to lie between 0 and 1. We denote this similarity, as assessed by the m-th subject to be $sim^m_{\text{hum}}(s_a^i, s_b^i)$

The average similarity score across all pairs of sentences and subjects gives us the specificity score $spec^i_{\text{hum}}$ for image i based on human perception. For ease of notation, we drop the superscript i when it is clear from the context.

$$spec_{\text{hum}} = \frac{1}{M\binom{N}{2}} \sum_{\forall\{s_a,s_b\}\subset S} \sum_{m=1}^{M} sim^m_{\text{hum}}(s_a, s_b) \quad (1)$$

Figure 2 shows images with their human-annotated specificity scores. Note how the specificity score drops as the sentence descriptions become more varied.

3.1.2 Automated Specificity Measurement

To measure specificity automatically given the N descriptions for image i, we first tokenize the sentences and only retain words of length three or more. This ensured that semantically irrelevant words, such as 'a', 'of', *etc*., were not taken into account in the similarity computation (a standard stop word list could also be used instead). We identified the synsets (sets of synonyms that share a common meaning) to which each (tokenized) word belongs using the Natural Language Toolkit [4]. Words with multiple meanings can belong to more than one synset. Let $Y_{au} = \{y_{au}\}$ be the set of synsets associated with the u-th word from sentence s_a.

Every word in both sentences contributes to the automatically computed similarity $sim_{\text{auto}}(s_a, s_b)$ between a pair of sentences s_a and s_b. The contribution of the u-th word from sentence s_a to the similarity is c_{au}. This contribution is computed as the maximum similarity between this word, and all words in sentence s_b (indexed by v). The similarity between two words is the maximum similarity between all pairs of synsets (or senses) to which the two words have been assigned. We take the maximum because a word is usually used in only one of its senses. Concretely,

$$c_{au} = \max_{v} \max_{y_{au}\in Y_{au}} \max_{y_{bv}\in Y_{bv}} sim_{\text{sense}}(y_{au}, y_{bv}) \quad (2)$$

The similarity between senses $sim_{\text{sense}}(y_{au}, y_{bv})$ is the shortest path similarity between the two senses on WordNet [36]. We can similarly define c_{bv} to be the contribution of v-th word from sentence s_b to the similarity $sim_{\text{auto}}(s_a, s_b)$ between sentences s_a and s_b.

The similarity between the two sentences is defined as the average contribution of all words in both sentences, weighted by the importance of each word. Let the importance of the u-th word from sentence s_a be t_{au}. This importance is computed using term frequency-inverse document frequency (TF-IDF) using the scikit-learn software package [40]. Words that are rare in the corpus but occur frequently in a sentence contribute more to the similarity of that sentence with other sentences. So we have

$$sim_{\text{auto}}(s_a, s_b) = \frac{\sum_u t_{au}c_{au} + \sum_v t_{bv}c_{bv}}{\sum_u t_{au} + \sum_v t_{bv}} \quad (3)$$

The denominator in Equation 3 ensures that the similarity between two sentences is independent of sentence-length and is always between 0 and 1. Finally, the automated specificity score $spec_{\text{auto}}$ of an image i is computed by averaging these similarity scores across all sentence pairs:

$$spec_{\text{auto}} = \frac{1}{\binom{N}{2}} \sum_{\forall\{s_a,s_b\}\subset S} sim_{\text{auto}}(s_a, s_b) \quad (4)$$

The reader is directed to the supplementary material [21] for a pictorial explanation of automated specificity computation.

3.2. Application: Text-based image retrieval

We now describe how we use image specificity in a text-based image retrieval application.

3.2.1 Setup

There is a particular image the user is looking for from a database of images. We call this the target image. The user inputs a query sentence q that describes the target image. Every image in the database is associated with a single reference description r_i (not to be confused with the "training" pool of sentences S^i described in Section 3.1 used to define the specificity of an image). This can be, for example, the caption in an online photo database such as Flickr. The goal is to sort the images in the database according to their relevance score rel^i from most to least relevant, such that the target image has a low rank.

3.2.2 Baseline Approach

The baseline approach automatically computes a similarity $sim_{\text{auto}}(q, r_i)$ between q and r_i using Equation 3. All images in the database are sorted in descending order using this similarity score. That is,

$$rel^i_{\text{baseline}} = sim_{\text{auto}}(q, r_i). \tag{5}$$

The image whose reference sentence has the highest similarity to the query sentence gets ranked first while the image whose reference sentence has the lowest similarity to the query sentence gets ranked last.

3.2.3 Proposed Approach

In the proposed approach, instead of ranking just by similarity between the query sentence and reference descriptions in the database, we take into consideration the specificity of each image. The rationale is the following: a specific image should be ranked high only if the query description matches the reference description of that image well, because we know that sentences that describe this image tend to be very similar. For ambiguous images, on the other hand, even mediocre similarities between query and reference descriptions may be good enough.

This suggests that instead of just sorting based on $sim_{\text{auto}}(q, r_i)$, the similarity between the query description q and the reference description r_i of an image i (which is what the baseline approach does as seen in Equation 5), we should model $P(\text{match}|sim_{\text{auto}}(q, r_i))$ which captures the probability that the query sentence matches the reference sentence *i.e.*, the query sentence describes the image. We use Logistic Regression (LR) to model this.

$$\begin{aligned} rel^i_{\text{gt-specificity}} &= P(\text{match}|sim_{\text{auto}}(q, r_i)) \\ &= \left\{ \frac{1}{1 + e^{-\beta^i_0 - \beta^i_1 sim_{\text{auto}}(q,r_i)}} \right\} \end{aligned} \tag{6}$$

For each image in the database, we train the above LR model. Positives examples of this model are the similarity scores between pairs of sentences both describing the image i taken from the set S^i described in Section 3.1. Negative examples are similarity scores between pairs of sentences where one sentence describes the image i but the other does not. If there are N descriptions available for each image during training, we have $\binom{N}{2}$ positive examples (all pairs of N sentences). We generate a similar number of negative examples by pairing each of the N descriptions with $\lceil \frac{N-1}{2} \rceil$ descriptions from other images. $\lceil . \rceil$ is the ceiling function.

The parameters of this LR model, β^i_0 and β^i_1, inherently capture the specificity of the image. Note that a separate LR model is trained for each image to model the specificity for that image. After these models have been trained, given a new query description q, the similarity $sim_{\text{auto}}(q, r_i)$ is computed with every reference description r_i in the dataset. The trained LR for each image, i.e. the parameters β^i_0 and β^i_1, can be used to compute $P(\text{match}|sim_{\text{auto}}(q, r_i))$ for that image. All images can then be sorted by their corresponding $P(\text{match}|sim_{\text{auto}}(q, r_i))$ values. In our experiments, unless mentioned otherwise, the query and reference descriptions being used at test time were not part of the training set used to train the LRs.

3.2.4 Predicting Specificity of Images

The above approach needs several sentences per image to obtain positive and negative examples to train the LR. But in realistic scenarios, it may not be viable to collect multiple sentences for every image in the database. Hence, we learn a mapping from the image features $\{\boldsymbol{x_i}\}$ to the LR parameters estimated using sentence pairs. We call these parameters ground-truth LR parameters. We train two separate ν-Support Vector Regression (SVR) models (one for each β term) with Radial Basis Function (RBF) kernel. The learnt SVR model is then used to predict the LR parameters $\hat{\beta}^i_0$ and $\hat{\beta}^i_1$ of any previously unseen image. Finally, these predicted LR parameters are used to compute $\hat{P}(\text{match}|sim_{\text{auto}}(q, r_i))$ and sort images in a database according to their relevance to a query description q. Of course, each image in the database still needs a (single) reference description r_i (without which text-based image retrieval is not feasible).

$$\begin{aligned} rel^i_{\text{pred-specificity}} &= \hat{P}(\text{match}|sim_{\text{auto}}(q, r_i)) \\ &= \left\{ \frac{1}{1 + e^{-\hat{\beta}^i_0 - \hat{\beta}^i_1 sim_{\text{auto}}(q,r_i)}} \right\} \end{aligned} \tag{7}$$

Notice that the baseline approach (Equation 5) is a special case of our proposed approaches (Equations 6 and 7) with $\beta^i_0 = \hat{\beta}^i_0 = \text{constant}_0$ and $\beta^i_1 = \hat{\beta}^i_1 = \text{constant}_1 \; \forall i$, where the parameters for each image are the same.

3.2.5 Summary

Let's say we are given a new database of images and associated (single) reference descriptions that we want to search using query sentences. SVRs are used to predict each of the two LR parameters using image features for every image in the database. This is done offline. When a query is issued, its similarity is computed to each reference description in

the image. Each of these similarities are substituted into Equation 7 to calculate the relevance of each image using the LR parameters predicted for that image. This query-time processing is computationally light. The images are then sorted by the probability outputs of their LR models. The quality of the retrieved results using this (proposed) approach is compared to the baseline approach that sorts all images based on the similarity between the query sentence and reference descriptions. Of course, in the scenario where multiple reference descriptions are available for each image in the database, we can directly estimate the (ground-truth) LR parameters using those descriptions (as described in Section 3.2.3) instead of using the SVR to predict the LR parameters. We will show results of both approaches (using ground-truth LR parameters and predicted LR parameters).

4. Experimental Results

4.1. Datasets and Image Features

We experiment with three datasets. The first is the MEM-5S dataset containing 888 images from the memorability dataset [19], which are uniformly spaced in terms of their memorability. For each of these images, we collected 5 sentence descriptions by asking unique subjects on AMT to describe them. Figure 2 shows some example images and their descriptions taken from the MEM-5S dataset. Since specificity measures the variance between sentences, and more sentences would result in a better specificity estimate, we also experiment with two datasets with 50 sentences per image in each dataset. One of these is the ABSTRACT-50S dataset [46] which is a subset of 500 images made of clip art objects from the Abstract Scene dataset [53] containing 50 sentences/image (48 training, 2 test). We use only the training sentences from this dataset for our experiments. The second is the PASCAL-50S dataset [46] containing 50 sentences/image for the 1000 images from the UIUC PASCAL dataset [41]. These datasets allow us to study specificity in a wide range of settings, from real images to non-photorealistic but semantically rich abstract scenes. All correlation analysis reported in the following sections was performed using Spearman's rank correlation coefficient.

For predicting specificity, we extract 4096D DECAF-6 [9] features from the PASCAL-50S images. Images in the ABSTRACT-50S dataset are represented by the occurrence, location, depth, flip angle of objects, object co-occurrences and clip art category (451D) [53].

4.2. Consistency analysis

In Section 3.1, we described two methods to measure specificity. In the first, humans are involved in annotating the similarity score between the sentences describing an image and in the second, this is done automatically. We first analyze if humans agree on their notions of specificity, and then study how well human annotation of specificity correlates with automatically-computed specificity.

Do humans rate sentence pair similarities consistently? The similarity of each pair of sentences in the MEM-5S dataset was rated by 3 subjects on AMT. This means that every image is annotated by $\binom{5}{2} * 3 = 30$ similarity ratings. The average of the similarity ratings gives us the specificity scores. These ratings were split thrice into two parts such that the ratings from one subject was in one part and ratings from the other two subjects were in the other part. The specificity score computed from the first part was correlated with the specificity score of the other part. This gave an average correlation coefficient of 0.72, indicating high consistency in specificity measured across subjects.

Is specificity consistent for a new set of descriptions? Additionally, 5 more sentences were collected for a subset of 222 images in the memorability dataset. With these additional sentences, specificity was computed using human annotations and the correlation with the specificity from the previous set of sentences was found to be 0.54. Inter-human agreement on the same set of 5 descriptions for 222 images was 0.76. We see that specificity measured across two sets of five descriptions each is not highly consistent. Hence, we hypothesize that measuring specificity using more sentences would be desirable (thus our use of the PASCAL-50S and ABSTRACT-50S datasets)[2].

How do human and automated specificity compare? We find that the rank correlation between human-annotated and automatically measured specificity (on the same set of 5 sentences for 888 images in MEM-5S) is 0.69 which is very close to the inter-human correlation of 0.72. Note that this automated method still requires textual descriptions by humans. In a later section, we will consider the problem of predicting specificity just from image features if textual descriptions are also not available. Note that in most realistic applications (*e.g.* image search, that we explore later), it is practical to measure specificity by comparing descriptions automatically. Hence, the automated specificity measurement may be the more relevant one.

Are some images more specific than others? Now that we know specificity is well-defined, we study whether some images are in fact more specific than others. Figure 2 shows some examples of images whose specificity values range from low to very high values. Note how the descriptions become more varied as the specificity value drops. Figure 3 shows a histogram of specificity values on all three datasets. In the MEM-5S dataset, the specificity values range from 0.11 to 0.93[3]. This indicates that indeed, some images are specific and some images are ambiguous. We can exploit this fact to improve applications such as text-based image retrieval (Section 4.4).

[2]It was prohibitively expensive to measure human specificity for all pairs of 50 sentences to verify this hypothesis

[3]Specificity values can fall between 0 and 1.

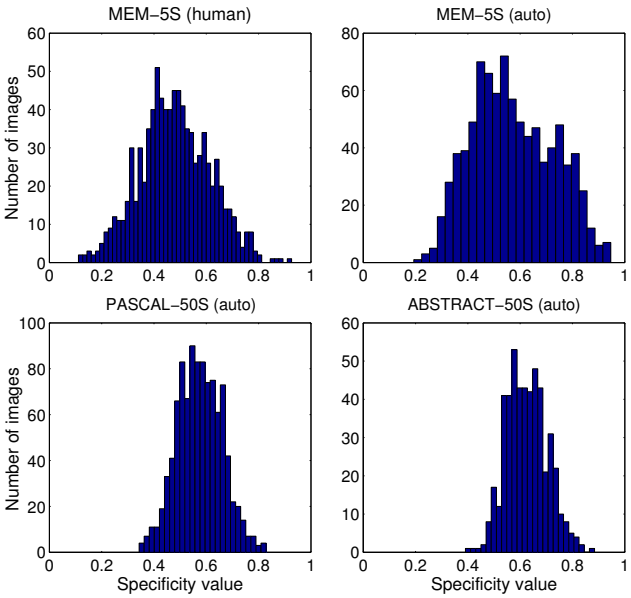

Figure 3. A histogram of human-annotated specificity values for the MEM-5S dataset (top left) and automated specificity values for all three datasets (rest).

4.3. What makes an image specific?

We now study what makes an image specific. The first question we want to answer is whether longer sentence descriptions lead to more variability and hence less specific images. We correlated the average length of a sentence (measured as the number of words in the sentence) with specificity, and surprisingly, found that the length of a sentence had no effect on specificity (ρ=-0.02, p-value=0.64). However, we did find that the more specific an image was, the less was the variation in length of sentences describing it (ρ=–0.16, p-value$<$0.01).

Next, we looked at image content to unearth possibly consistent patterns that make an image specific. We correlated publicly available attribute, object and scene annotations [18] for the MEM-5S dataset with our specificity scores. We then sorted the annotations by their correlation with specificity and showed the top 10 and bottom 10 correlations as a bar plot in Figure 4. We find that images with people tend to be specific, while mundane images of generic buildings or blue skies tend to not be specific. Note that if a category (*e.g.* person) and its subcategory (*e.g.* caucasian person) both appeared in the top 10 or bottom 10 list and had very similar correlations, the subcategory was excluded in favour of the main category since the subcategory is redundant.

Next, we hypothesized that images with larger objects in them may be more specific, since different people may all talk about those objects. Confirming this hypothesis, we found a correlation of 0.16 with median object area and 0.14 with mean object area.

We then investigated how importance [3] relates to specificity. Since important objects in images are those that tend to be mentioned often, perhaps an image containing an important object will be more specific because most people will talk about the object. We consider all sentences corresponding to all images containing a certain object category. In each sentence, we identify the word (*e.g.* vehicle) that best matches the category (*e.g.* car) using the shortest-path similarity of the words taken from the WordNet database [36]. We average the similarity between this best matching word in each sentence to the category name across all sentences of all images containing that category. This is a proxy for how frequently that category is mentioned in descriptions of images containing the category. A similar average was found for randomly chosen sentences from other categories as a proxy for how often a category gets mentioned in sentences *a priori*. These two averages were subtracted to obtain an importance value for each category. Now, the specificity scores of all images containing an object category was averaged and this score was found to be significantly correlated (ρ=0.31, p-value=0.05) with the importance score. This analysis was done only for categories that were present in more than 10 images in the MEM-5S dataset. This shows that images containing important objects do tend to be more specific.

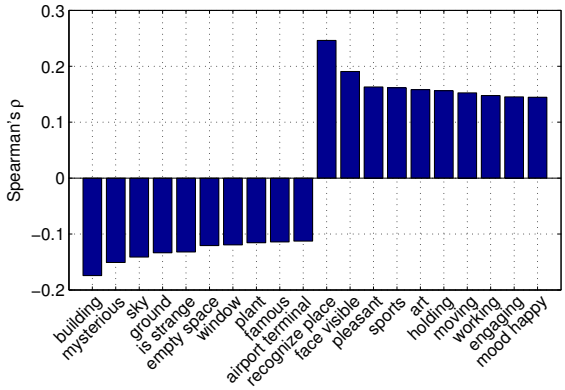

Figure 4. Spearman's rank correlation of human specificity with attributes, objects and scene annotations for the MEM-5S dataset.

In another study, Isola *et al.* [19] measured image memorability. Images were flashed in front of subjects who were asked to press a button each time they saw the same image again. Interestingly, repeats of some images are more reliably detected across subjects than other images. That is, some images are more memorable than others. We tested if memorability and specificity are related by correlating them and found a high correlation (ρ=0.33, p-value$<$0.01) between the two. Thus, specificity can explain memorability to some extent. However, the two concepts are distinct. For instance, peaceful, picture-perfect scenes that may appear on a postcard or in a painting were found to be negatively correlated with memorability [18] (ρ_{peaceful}=–0.35, ρ_{postcard}=–0.31, ρ_{painting}=–0.32). But these attributes have no correlation with specificity (ρ_{peaceful}=–0.05, ρ_{painting}=–

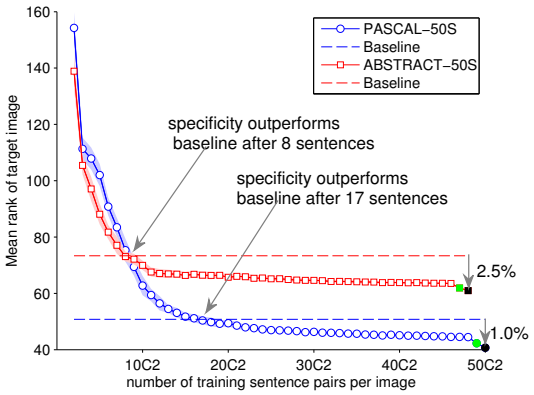

Figure 5. Image search results: Increasing the number of training sentences per image improves the mean target rank obtained with ground truth LR parameters (specificity). As expected, there is a sharp improvement when the reference sentence (green fill) or both the reference and query sentences (black fill) are included when estimating the LR parameters. The results are averaged across 25 random repeats and the error intervals are shown in shaded colors. Annotations indicate the number of sentences required to beat baseline and the maximum improvement possible over baseline using all available sentences. Lower mean rank of target means better performance.

0.02, ρ_{postcard}=–0.05). In the supplementary material [21], we include examples of images that are memorable but not specific and vice-versa. Additional scatter plots for a subset of the computed correlations are also included in the supplementary material [21]. Finally, correlation of mean color of the image with specificity was ρ_{red}=0.01, ρ_{green}=0.02 and ρ_{blue}=0.01.

Overall, specificity is correlated with image content to quite an extent. In fact, if we train a regressor to predict automated specificity directly from DECAF-6 features in the PASCAL-50S and MEM-5S dataset, we get a correlation of 0.2 and 0.25. The correlation using semantic features in the ABSTRACT-50S dataset was 0.35. More details in supplementary material [21]. The reader is encouraged to browse our datasets through the websites on the authors' webpages.

4.4. Image search

4.4.1 Ground truth Specificity

Given a database of images with multiple descriptions each, Section 3.2.3 describes how we estimate parameters of a Logistic Regression (LR) model, and use the model for image search. In our experiments, the query sentence corresponds to a known target image (known for evaluation, not to the search algorithm). The evaluation metric is the rank of the target images, averaged across multiple queries.

We investigate the effect of number of training sentences per image used to train the LR model on the average rank of target images. Figure 5 shows that the mean rank of the target image decreases with increasing number of training sentences. The baseline approach (Section 3.2.2) simply sorts the images by the similarity value between the query sentence and all reference sentences in the database (one per image). For the PASCAL-50S dataset, 17 training sentences per image were required to estimate an LR model that can beat this baseline while for ABSTRACT-50S dataset, 8 training sentences per image were enough. The improvement obtained over the baseline by training on all 50 sentences was 1.0% of the total dataset size for PASCAL-50S and 2.5% for the ABSTRACT-50S dataset[4]. With this improvement, we bridge 20.5% of the gap between baseline and perfect result (target rank of 1) for PASCAL-50S and 17.5% for ABSTRACT-50S.

	Method	Mean rank	% of queries meet or beat BL
PASCAL-50S	BL	50.80	–
	GT-Spec	44.70	67.3
	P-Spec	49.72	73.2
ABSTRACT-50S	BL	73.34	–
	GT-Spec	63.30	61.0
	P-Spec	69.41	61.6

Table 1. Image search results for different ranking algorithms: baseline (BL), specificity using ground truth (GT-Spec) LR parameters, and specificity (P-Spec) using predicted LR parameters. The column with header Mean rank gives the rank of the target image averaged across all images in the database. The final column indicates the percentage of queries where the method does better than or as good as baseline.

4.4.2 Predicted Specificity

We noted in the previous section that as many as 17 training sentences per image are needed to estimate specificity accurately enough to beat baseline on the PASCAL-50S dataset. In a real application, it is not feasible to expect these many sentences per image. This leads us to explore if it is possible to predict specificity directly from images accurately enough to beat the baseline approach.

As described in Section 3.2.4, we train regressors that map image features to the LR parameters. These regressors are then used to predict the LR parameters that are used for ranking the database of test images in an image search application. We do this with leave-one-out cross-validation (ensuring that none of the textual descriptions of the predicted image were included in the training set) so that we have predicted LR parameters on our entire dataset.

Table 1 shows how often our approach does better than or matches the baseline. It can be seen that specificity using the LR parameters predicted directly using image features does better than or matches the baseline for 73.2% of the queries. This is especially noteworthy since *no sentence descriptions* were used to estimate the specificity of the images. Specificity is predicted using purely image features.

[4]Our approach is general and can be applied to different automatic text similarity metrics. For instance, cosine similarity (dot product of the TF-IDF vectors) also works quite well with a 3.5% improvement using ground-truth specificity for ABSTRACT-50S.

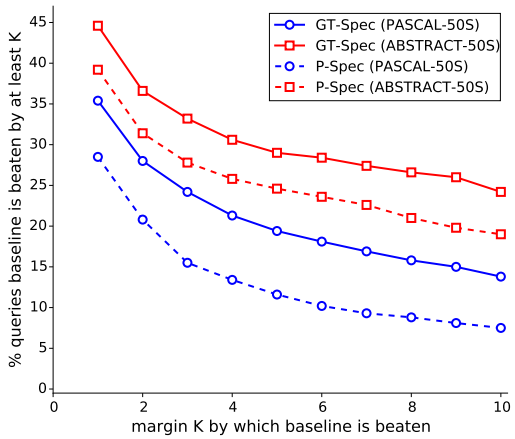

Figure 6. Image search results: On the x-axis is plotted K, the margin in rank of target image by which baseline is beaten, and on the y-axis is the percentage of queries where baseline is beaten by at least K.

From Table 1, we note that predicted specificity (P-Spec) loses less often to baseline as compared to ground-truth specificity (GT-Spec), but GT-Spec still has a better average rank of target images compared to P-Spec. The reason is that GT-Spec does *much* better than P-Spec on queries where it wins against baseline. Therefore, we would like to know that when an approach beats baseline, how often does it beat baseline by a low or high margin? Figure 6 shows the percentage of queries which beat baseline by different margins. The x-axis is the margin by at least which the baseline is beaten and on the y-axis is the percentage of queries. As expected, ground-truth specificity performs the best amongst the three methods. But even predicted specificity often beats the baseline by large margins.

Many approaches [6, 49] retrieve images based on text-based matches between the query and reference sentences, and then re-rank the results based on image content. This content-based re-ranking step can be performed on top of our retrieved results as well. Note that in our approach, the image features are used to modulate the similarity between the query and reference sentences – and not to better assess the match between the query sentence and image content. These are two orthogonal sources of information.

Finally, Figure 7 shows a qualitative example from the ABSTRACT-50S dataset. In the first image, the query and the reference sentence for the target image do not match very closely. However, since the image has a low automated specificity, this mediocre similarity is sufficient to lower the rank of the target image.

5. Discussion and Conclusion

We introduce the notion of specificity. We present evidence that the variance in textual descriptions of an image, which we call specificity, is a well-defined phenomenon. We find that even abstract scenes which are not photorealistic capture this variation in textual descriptions. We study

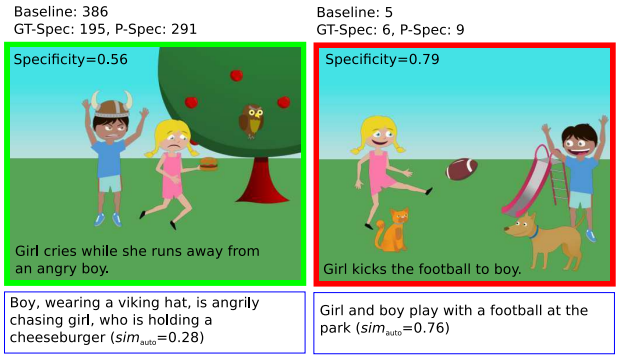

Figure 7. Qualitative image search results from the ABSTRACT-50S dataset. The images are the target images. Query sentences are shown in the blue box below each image (along with their automated similarity to the reference sentence). The reference sentence of the image is shown on the image. The automated specificity value is indicated at the top-left corner of the images. Green border (left) indicates that both predicted and ground-truth specificity performed better than baseline, and red border (right) indicates that baseline did better than both P-Spec and GT-Spec. The rank of the target image using baseline, GT-Spec and P-Spec is shown above the image.

various object and attribute-level properties that influence specificity. More importantly, modeling specificity can benefit various applications. We demonstrate this empirically on a text-based image retrieval task. We use image features to predict the *parameters* of a classifier (Logistic Regression) that modulates the similarity between the query and reference sentence differently for each image. Future work involves exploring robust measures of specificity that consider only representative sentences (not outliers), investigating other similarity measures such as Lin's similarity [33] or word2vec[5] when measuring specificity, exploring the potential of low-level saliency and objectness maps in predicting specificity, studying specificity in more controlled settings involving a closed set of visual concepts and using image specificity in various applications such as image tagging to determine how many tags to associate with an image (few for specific images and many for ambiguous images), image captioning, *etc*. Our data and code are publicly available on the authors' webpages.

Glossary

automated specificity Specificity computed from image textual descriptions by averaging automatically computed sentence similarities (Section 3.1.2) [3, 5–8] **ground-truth LR/specificity** Specificity computed using Logistic Regression parameters estimated from image textual descriptions (Section 3.2.3) [4, 5, 7, 8] **human specificity** Specificity measured from image textual descriptions by averaging human-annotated sentence similarities (Section 3.1.1) [3, 5, 6] **predicted LR/specificity** Specificity computed using Logistic Regression parameters predicted from image features (Section 3.2.4) [4, 5, 7, 8]

[5] http://code.google.com/p/word2vec/

Acknowledgements: This work was supported in part by The Paul G. Allen Family Foundation Allen Distinguished Investigator award to D.P.

References

[1] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. 2005. 2

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 408–415, 2001. 2

[3] A. C. Berg, T. L. 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3562–3569, 2012. 1, 2, 6

[4] S. Bird. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006. 3

[5] X. Chen and C. L. Zitnick. Learning a Recurrent Visual Representation for Image Caption Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[6] J. Cui, F. Wen, and X. Tang. Real time google and live image search re-ranking. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 729–732. ACM, 2008. 8

[7] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1657–1664, 2011. 2

[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014. 1

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 5

[10] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 745–752, 2011. 2

[11] D. Elliott and F. Keller. Comparing Automatic Evaluation Measures for Image Description. 1

[12] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge–a Retrospective. 1

[13] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[14] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010. 1, 2

[15] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201, 2011. 2

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013. 1

[17] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. The interestingness of images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1633–1640, 2013. 2

[18] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, 2011. 2, 6

[19] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011. 2, 5, 6

[20] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 2

[21] M. Jas and D. Parikh. Image Specificity. *arXiv preprint arXiv:1502.04569*, 2015. 3, 7

[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to Predict Where Humans Look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 2

[23] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[24] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. International World Wide Web Conferences Steering Committee, 2014. 2

[25] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Transactions of the Association for Computational Linguistics (TACL)*, 2015. 1

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[27] R. Krovetz and W. B. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, 10(2):115–141, 1992. 2

[28] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1608, 2011. 1, 2

[29] M. S. Lew. Next-generation web searches for visual content. *Computer*, 33(11):46–53, 2000. 2

[30] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski. Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics (TOG)*, 27(3):38, 2008. 2

[31] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-

grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011. 1

[32] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004. 1, 2

[33] D. Lin. An information-theoretic definition of similarity. 8

[34] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2657–2664, 2014. 2

[35] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014. 1

[36] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. 3, 6

[37] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012. 1, 2

[38] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011. 1, 2

[39] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 1, 2

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011. 3

[41] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010. 5

[42] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of vision*, 7(2):17, 2007. 2

[43] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 801–808, 2011. 2

[44] M. Spain and P. Perona. Measuring and predicting importance of objects in our visual world. 2007. 2

[45] N. Turakhia and D. Parikh. Attribute dominance: What pops out? In *IEEE International Conference on Computer Vision (ICCV)*, pages 1225–1232, 2013. 2

[46] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5

[47] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[48] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 537–544, 2009. 1

[49] X. Wang, K. Liu, and X. Tang. Query-specific visual semantic spaces for web image re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 857–864, 2011. 8

[50] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010. 1

[51] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. 1

[52] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014. 1

[53] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3016, 2013. 5