Journal of Artificial Intelligence Research 47 (2013) 853-899 Submitted 03/13; published 08/13

Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics

Micah Hodosh MHODOSH2@ILLINOIS.EDU
Peter Young PYOUNG2@ILLINOIS.EDU
Julia Hockenmaier JULIAHMR@ILLINOIS.EDU
Department of Computer Science
University of Illinois
Urbana, IL 61801, USA

Abstract

The ability to associate images with natural language sentences that describe what is depicted in them is a hallmark of image understanding, and a prerequisite for applications such as sentence-based image search. In analogy to image search, we propose to frame sentence-based image annotation as the task of ranking a given pool of captions. We introduce a new benchmark collection for sentence-based image description and search, consisting of 8,000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events. We introduce a number of systems that perform quite well on this task, even though they are only based on features that can be obtained with minimal supervision. Our results clearly indicate the importance of training on multiple captions per image, and of capturing syntactic (word order-based) and semantic features of these captions. We also perform an in-depth comparison of human and automatic evaluation metrics for this task, and propose strategies for collecting human judgments cheaply and on a very large scale, allowing us to augment our collection with additional relevance judgments of which captions describe which image. Our analysis shows that metrics that consider the ranked list of results for each query image or sentence are significantly more robust than metrics that are based on a single response per query. Moreover, our study suggests that the evaluation of ranking-based image description systems may be fully automated.

1. Introduction

The ability to automatically describe the entities, events and scenes depicted in an image is possibly the most ambitious test of image understanding. Any advances on this task have significant practical implications, since there are billions of images on the web and in personal photo collections. Our ability to efficiently access the wealth of information they contain is hampered by limitations of standard image search engines, which must rely on text that appears near the image (Datta, Joshi, Li, & Wang, 2008; Popescu, Tsikrika, & Kludas, 2010). There has been a lot of work on the multi-label classification problem of associating images with individual words or tags (see, e.g., Blei & Jordan, 2003; Barnard, Duygulu, Forsyth, Freitas, Blei, & Jordan, 2003; Feng & Lapata, 2008; Deschacht & Moens, 2007; Lavrenko, Manmatha, & Jeon, 2004; Makadia, Pavlovic, & Kumar, 2010; Weston, Bengio, & Usunier, 2010), but the much harder problem of automatically associating images with complete sentences that describe them has only recently begun to attract attention.

1.1 Related Work

Although a few approaches have framed sentence-based image description as the task of mapping images to sentences written by people (Farhadi, Hejrati, Sadeghi, Young, Rashtchian, Hockenmaier, & Forsyth, 2010; Ordonez, Kulkarni, & Berg, 2011), most research in this area has focused on the task of automatically generating novel captions (Kulkarni, Premraj, Dhar, Li, Choi, Berg, & Berg, 2011; Yang, Teo, Daume III, & Aloimonos, 2011; Li, Kulkarni, Berg, Berg, & Choi, 2011; Mitchell, Dodge, Goyal, Yamaguchi, Stratos, Han, Mensch, Berg, Berg, & Daume III, 2012; Kuznetsova, Ordonez, Berg, Berg, & Choi, 2012; Gupta, Verma, & Jawahar, 2012). We argue in this paper that framing image description as a natural language generation problem introduces a number of linguistic difficulties that detract attention from the underlying image understanding problem we wish to address. Since any sentence-based image description or retrieval system requires the ability to associate images with captions that describe what is depicted in them, we argue it is important to evaluate this mapping between images and sentences independently of the generation aspect. Research on caption generation has also ignored the image search task, which is arguably of much greater practical importance.

All of the systems cited above are either evaluated on a data set that our group released in earlier work (Rashtchian, Young, Hodosh, & Hockenmaier, 2010), or on the SBU Captioned Photo Dataset (Ordonez et al., 2011). Our data set consists of 1,000 images from the PASCAL VOC-2008 object recognition challenge that are each annotated with five descriptive captions which we purposely collected for this task. The SBU data set consists of one million images and captions harvested from Flickr. Gupta et al. (2012) is the only system to use Grubinger, Clough, Müller, and Deselaers's (2006) IAPR TC-12 data set, which consists of 20,000 images paired with longer descriptions.

Although details differ, most models rely on existing detectors to define and map images to an explicit meaning representation language consisting of a fixed number of scenes, objects (or 'stuff'), their attributes and spatial relations (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Ordonez et al., 2011; Mitchell et al., 2012). But it is unclear how well these detector-based approaches generalize: the models evaluated on our PASCAL VOC-2008 data set (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Mitchell et al., 2012) all rely on detectors that may have been trained on images contained in this corpus, and Kuznetsova et al. (2012) select a test set of 1,000 images from the SBU data set for which their detectors work well. Moreover, among the systems evaluated on our PASCAL VOC-2008 data set, only Kulkarni et al. (2011), Li et al. (2011), Li et al. (2011) and Mitchell et al.'s (2012) results may be directly comparable, since different research groups report different evaluation metrics and use a different parts of the data set as test or training data. The evaluation of generation systems is generally well known to be difficult (see, e.g., Dale & White, 2007; Reiter & Belz, 2009), and typically requires expensive human judgments that have to consider the quality of both content selection (what is being described) and surface realization (the fluency of the generated text). These syntactic and pragmatic issues confound the purely semantic question of whether the image is correctly described by the caption.

1.2 Our Approach

In this paper, we focus on the task of associating images with sentences drawn from a large, predefined pool of image descriptions. These descriptions are not generated automatically or harvested from the web (Feng & Lapata, 2008; Ordonez et al., 2011), but are written by people who were asked to describe them. We argue that evaluating the ability to select or rank, rather than generate, appropriate captions for an image is the most direct test of the fundamental semantic question of how well we can associate images with sentences that describe them well. Framing image description as a ranking task also has a number of additional advantages. First, it allows us to handle sentence-based image annotation and search in a unified framework, allowing us to evaluate whether advances in one task carry over to the other. Second, framing image description as a ranking problem greatly simplifies evaluation. By establishing a parallel between description and retrieval, we can use the same metrics to evaluate both tasks. Moreover, we show that the rank of the original caption, which is easily determined automatically, leads to metrics that correlate highly with systems rankings obtained from human judgments, even if they underestimate actual performance. We also show that standard automatic metrics such as Bleu (Papineni, Roukos, Ward, & Zhu, 2002) or Rouge (Lin, 2004) that have also been used to evaluate caption generation systems show poor correlation with human judgments, leading us to believe that the evaluation of caption generation system should not be automated. We also perform a large-scale human evaluation, but since the sentences in our data set are image descriptions written by people, we only need to collect purely semantic judgments of whether they describe the images the system associated them with. And since these judgments are independent of the task, we can use them to evaluate both image description and retrieval systems. Since we collect these judgments over image-caption pairs in our publicly available data set, we also establish a common benchmark that enables a direct comparison of different systems. We believe that this is another advantage over the caption generation task. Since there are many possible ways to describe an image, generation systems are at liberty to be more or less specific about what they describe in an image. This makes a direct comparison of independently obtained judgments about the quality of two different systems very difficult, since one system may be aiming to solve a much harder task than the other, and implies that unless system outputs for a common benchmark collection of images were made publicly available, there cannot be any shared, objective evaluation that would allow the community to measure progress on this difficult problem. But since caption generation systems also need to be able to determine how well a caption describes an image, our data set could potentially be used to evaluate their semantic component.

1.3 Contributions and Outline of this Paper

In Section 2, we discuss the need for a new data set for image description and introduce a new, high quality, data set for image description which will enable the community to compare different systems against the same benchmark. Our PASCAL VOC-2008 data set of 1,000 images (Rashtchian et al., 2010) has been used by a number of image description systems (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Gupta et al., 2012), but has a number of shortcomings that limit its usefulness. First, its domain is relatively limited, and the captions are relatively simple. Second, since its images are drawn from the data used for the PASCAL VOC-2008 object classes challenge, it is difficult to guarantee a fair evaluation of description systems which rely on off-the-shelf object detectors (e.g., Felzenszwalb, McAllester, & Ramanan, 2008) on this data set, since it may not be possible to identify which images these detectors have been trained on. The experiments in this paper are therefore based on a larger, and more diverse, data set of 8,000 images. Unlike other data sets that pair images with sentences that are merely related to the image (Feng & Lapata, 2008; Ordonez et al., 2011), each image in our data sets are paired with five different captions that were purposely written to describe the image.

In Section 3, we describe our own image description systems. Because image description is such a novel task, it remains largely unknown what kind of model, and what kind of visual and linguistic features it requires. Instead of a unidirectional mapping from images to sentences that is common to current caption generation systems, we map both images and sentences into the same space. This allows us to apply our system to image search by retrieving the images that are closest to a query sentence, and to image description by annotating images with those sentences that are closest to it. The technique we use, Kernel Canonical Correlation Analysis (KCCA; Bach & Jordan, 2002), has already been successfully used to associate images (Hardoon, Szedmak, & Shawe-Taylor, 2004; Hwang & Grauman, 2012; Hardoon, Saunders, Szedmak, & Shawe-Taylor, 2006) or image regions (Socher & Li, 2010) with individual words or sets of tags, while Canonical Correlation Analysis (Hotelling, 1936) has also been used to associate images with related Wikipedia articles from ten different categories (Rasiwasia, Pereira, Coviello, Doyle, Lanckriet, Levy, & Vasconcelos, 2010). However, the performance of these techniques on the much more stringent task of associating images with sentences that describe what is depicted in them has not been evaluated. We compare a number of text kernels that capture different linguistic features. Our experimental results (discussed in Section 4) demonstrate the importance of robust textual representations that consider the semantic similarity of words, and hence take the linguistic diversity of the different captions associated with each image into account. Our visual features are relatively simple. A number of image description systems (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Kuznetsova et al., 2012) largely rely on trained detectors, e.g. to obtain an explicit intermediate meaning representation of the depicted objects, scenes and events. But this approach would ultimately require separate detectors, and hence labeled training data, for each term or phrase in the chosen meaning representation language. We show here that image features that capture only low-level perceptual properties can in fact work surprisingly well on our larger data set for which no in-domain detectors are available.

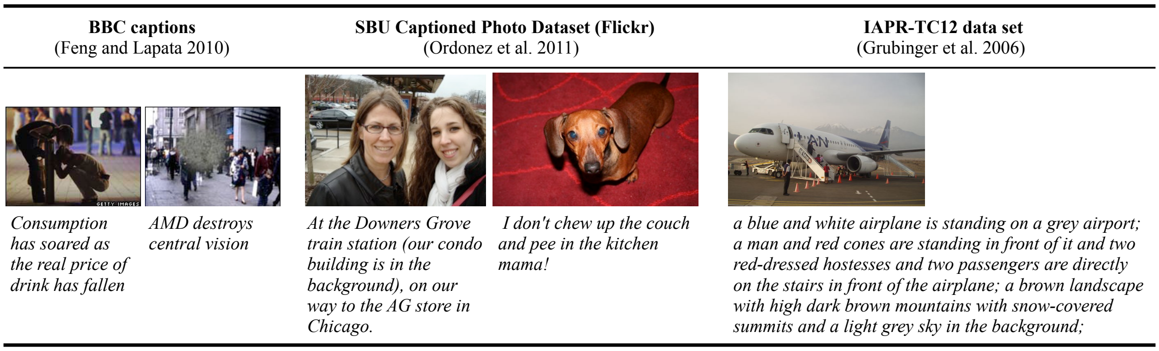
In Section 4, we consider the question of evaluation, and use a number of different metrics to compare our systems. Since we focus on the problem of learning an appropriate mapping between images and captions, we follow standard machine learning practice and evaluate the ability of this function to generalize to unseen examples. Hence, we separate the pool of captions and images used for testing from those used to train our systems. We first consider metrics for the quality of a single image-caption pair, and compare automatically computed scores with detailed human judgments. We then examine metrics that evaluate the ranked lists returned by our systems. Our analysis reveals that, at the current level of performance, differences between models may not become apparent if only a single caption per image is considered, as is commonly done for caption generation systems. But even if two models are equally likely to fail to return a suitable caption as the first result, we should still prefer the one that is more likely to rank good captions higher than the other, since it arguably provides a better approximation of the semantic space in which images are near captions that describe them well. Since the test pool contains a single gold item for each query, we first consider metrics that are based on the rank and recall of this gold item. We then show that simpler, binary judgments of image descriptions that are good approximations of more fine-grained human judgments can be collected on a very large scale via crowdsourcing. We augment the test pool of our data set with these relevance judgments, in the hope that this will add to its usefulness as a community resource and benchmark. These judgments show that the actual performance of our systems is higher than the recall of the gold item indicates. However, a comparison of the system rankings obtained via different metrics also suggests that differences in the rank or recall of the gold item correlate very highly with difference in performance according to the binary relevance judgments.

2. A New Data Set for Image Description

We have used crowdsourcing to collect descriptive captions for a large number of images of people and animals (mostly dogs). Before describing our data set and annotation methodology, we discuss what kind of captions are most useful for image description, and motivate the need to create new data sets for this task.

2.1 What Do We Mean by Image Description?

Since automatic image description is a relatively novel task, it is worth reflecting what it means to describe images, and what we wish to say about an image. There is in fact a substantial body of work on image description related to image libraries (Jaimes, Jaimes, & Chang, 2000; Shatford, 1986) that is useful to revisit for our purpose. We argue that out of the three different kinds of image descriptions that are commonly distinguished, one type, the so-called *conceptual* descriptions, is of most relevance to the image understanding we aim to achieve with automatic captioning. Conceptual image descriptions identify what is depicted in the image, and while they may be *abstract* (e.g., concerning the mood a picture may convey), image understanding is mostly interested in *concrete* descriptions of the depicted scene and entities, their attributes and relations, as well as the events they participate in. Because they focus on what is actually in the image, conceptual descriptions differ from so-called *non-visual* descriptions, which provide additional background information that cannot be obtained from the image alone, e.g. about the situation, time or location in which the image was taken. *Perceptual* descriptions capture low-level visual properties of images (e.g., whether it is a photograph or a drawing, or what colors or shapes dominate) are of little interest to us, unless they link these properties explicitly to the depicted entities. Among concrete conceptual descriptions, a further distinction can be drawn between *specific* descriptions, which may identify people and locations by their names, and *generic* descriptions (which may, e.g., describe a person as a woman or a skateboarder, and the scene as a city street or a room). With the exception of iconic entities that should be recognized as such (e.g., well-known public figures or landmark locations such as the Eiffel Tower) we argue that image understanding should focus on the information captured by



BBC captions (Feng and Lapata 2010)

Consumption has soared as the real price of drink has fallen

AMD destroys central vision

SBU Captioned Photo Dataset (Flickr) (Ordonez et al. 2011)

At the Downers Grove train station (our condo building is in the background), on our way to the AG store in Chicago.

I don't chew up the couch and pee in the kitchen mama!

IAPR-TC12 data set (Grubinger et al. 2006)

a blue and white airplane is standing on a grey airport; a man and red cones are standing in front of it and two red-dressed hostesses and two passengers are directly on the stairs in front of the airplane; a brown landscape with high dark brown mountains with snow-covered summits and a light grey sky in the background;

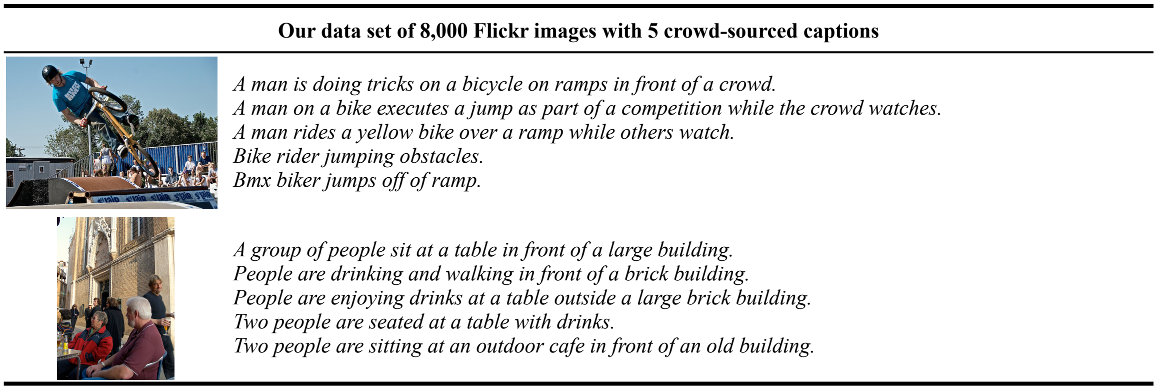
Figure 1: Other data sets of images and captions

generic descriptions. This leaves the question of where to obtain a data set of images paired with suitable descriptions to train automatic description systems on.

2.2 The Need for New Data Sets

While there is no dearth of images that are associated with text available online, we argue that most of this text is not suitable for our task. Some work, notably in the natural language processing community, has focused on images in news articles (Feng & Lapata, 2008, 2010). However, images are often only used to illustrate stories, and have little direct connection to the text (Figure 1, left). Furthermore, even when captions describe the depicted event, they tend to focus on the information that cannot be obtained from the image itself. Similarly, when people provide captions for the images they upload on websites such as Flickr (Figure 1, center), they often describe the situation that the images were taken in, rather than what is actually depicted in the image. That is, these captions often provide non-visual or overly specific information (e.g., by naming people appearing in the image or the location where the image was taken). There is a simple reason why people do not typically provide the kinds of generic conceptual descriptions that are of most use for our purposes: Gricean maxims of relevance and quantity (Grice, 1975) entail that image captions that are written for people usually provide precisely the kind of information that could *not* be obtained from the image itself, and thus tend to bear only a tenuous relation to what is actually depicted. Or, to state it more succinctly, captions are usually written to be seen along with the images they accompany, and users may not wish to bore other readers with the obvious.

Ordonez et al. (2011) harvested images and their captions from Flickr to create the SBU Captioned Photo Dataset, but had to discard the vast majority of images because their captions were not actually descriptive. Further analysis of a random sample of 100 images of their final data set revealed that the majority (67/100) of their captions describe information that cannot be obtained from the image itself (e.g., by naming the people or locations appearing in the image), while a substantial fraction (23/100) only describe a small detail of the image or are otherwise just commentary about the image. Examples of these issues are shown in Figure 1 (center). This makes their data set less useful for the kind of image understanding we are interested in: unless they refer to specific entities one may actually wish to identify (e.g., celebrities or famous landmarks that appear in the image), proper nouns are of little help in learning about visual properties of entity types unless one



Our data set of 8,000 Flickr images with 5 crowd-sourced captions

A man is doing tricks on a bicycle on ramps in front of a crowd.
A man on a bike executes a jump as part of a competition while the crowd watches.
A man rides a yellow bike over a ramp while others watch.
Bike rider jumping obstacles.
Bmx biker jumps off of ramp.

A group of people sit at a table in front of a large building.
People are drinking and walking in front of a brick building.
People are enjoying drinks at a table outside a large brick building.
Two people are seated at a table with drinks.
Two people are sitting at an outdoor cafe in front of an old building.

Figure 2: Our data set of images paired with generic conceptual descriptions

can infer what kind of entity they refer to.[1] The IAPR TC-12 data set (Grubinger et al., 2006), which consists of 20,000 photographs is potentially more useful for our purposes, since it contains descriptions of "what can be recognized in an image without any prior information or extra knowledge." However, the descriptions, which consist often of multiple sentences or sentence fragments, have a tendency to be lengthy (average length: 23.1 words) and overly detailed, instead of focusing on the salient aspects of the photograph. For example, in the photo of an airplane in Figure 1 (right), the *'two hostesses'* are barely visible but nevertheless described in detail.

2.3 Our Data Sets

Since the kinds of captions that are normally provided for images do not describe the images themselves, we have collected our own data sets of images and captions. The captions are obtained by using the crowdsourcing service provided by Amazon Mechanical Turk to annotate each image with five descriptive captions. By asking people to describe the people, objects, scenes and activities that are shown in a picture without giving them any further information about the context in which the picture was taken, we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone. Our annotation process and quality control are described in detail in Rashtchian et al. (2010)'s paper. We have annotated two different data sets in this manner:

2.3.1 The PASCAL VOC-2008 Data Set

The first data set we produced is relatively small, and consists of only 1,000 images randomly selected from the training and validation set of the PASCAL 2008 object recognition challenge (Everingham, Gool, Williams, Winn, & Zisserman, 2008). It has been used by a large number of image description systems (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Gupta et al., 2012), but since almost all of these systems (the only exception being Gupta et al., 2012) rely on detectors trained

[1] The data set of Ordonez et al. (2011) also differs significantly in content from ours: while our collection focuses on images of eventualities, i.e. people or animals doing something, the majority of Ordonez et al.'s images (60/100) do not depict people or animals (e.g., still lifes, landscape shots).

on images from the same data set (Felzenszwalb et al., 2008), it is unclear how well these approaches would generalize to other domains where no labeled data to train detectors is available. The captions in the PASCAL data set are also relatively simple. For example, since the data set contains many pictures that do not depict or focus on people doing something, 25% of the captions do not contain any verb, and an additional 15% of the captions contain only the common static verbs *sit*, *stand*, *wear*, or *look*.

2.3.2 The Flickr 8K Data Set

For the work reported in this paper we therefore collected a larger, more diverse data set consisting of 8,092 images from the Flickr.com website. Unlike the more static PASCAL images, the images in this data set focus on people or animals (mainly dogs) performing some action. Examples from this data set are shown in Figure 2. The images were chosen from six different Flickr groups,[2] and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations. In order to avoid ungrammatical captions, we only allowed workers from the United States who had passed a brief spelling and grammar test we devised to annotate our images. Because we were interested in conceptual descriptions, annotators were asked to write sentences that describe the depicted scenes, situations, events and entities (people, animals, other objects). We collected multiple captions for each image because there is a considerable degree of variance in the way many images can be described. As a consequence, the captions of the same images are often not direct paraphrases of each other: the same entity or event or situation can be described in multiple ways (*man* vs. *bike rider*, *doing tricks* vs. *jumping*), and while everybody mentions the bike rider, not everybody mentions the crowd or the ramp. The more dynamic nature of the images is also reflected in how they are being described: Captions in this data set have an average length of 11.8 words, compared to 10.8 words in the PASCAL data set, and while 40% of the PASCAL captions contain no verb other than *sit*, *stand*, *wear*, or *look*, only 11% of the captions for the Flickr 8K set contain no verb, and an additional 10% contain only these common verbs. Our data sets, the Flickr training/test/development splits and human relevance judgments used for evaluation of the test items (Section 4) are publicly available.[3] The online appendix to this paper contains our instructions to the workers, including the qualification test they had to pass before being allowed to complete our tasks.

3. Systems for Sentence-Based Image Description

Since image description requires the ability to associate images and sentences, all image description systems can be viewed in terms of an affinity function $f(\mathbf{i}, \mathbf{s})$ which measures the degree of association between images and sentences. We will evaluate our ability to compute such affinity functions by measuring performance on two tasks that depend directly on them. Given a candidate pool of sentences S_{cand} and a candidate pool of images I_{cand}, sentence-based image retrieval aims to find the image $\mathbf{i}^* \in I_{\text{cand}}$ that maximizes $f(\mathbf{i}, \mathbf{s}_q)$ for a query sentence $\mathbf{s}_q \in S_{\text{cand}}$. Conversely, image annotation aims to find the sentence $\mathbf{s}^* \in S_{\text{cand}}$ that

[2] These groups were called strangers!, Wild-Child (Kids in Action), Dogs in Action (Read the Rules), Outdoor Activities, Action Photography, Flickr-Social (two or more people in the photo)

[3] `http://nlp.cs.illinois.edu/HockenmaierGroup/data.html`

maximizes $f(\mathbf{i}_q, \mathbf{s})$ for a query image $\mathbf{i}_q \in I_{\text{cand}}$. In both cases, $f(\mathbf{i}, \mathbf{s})$ should of course be maximized for image-sentence pairs in which the sentence describes the image well:

$$\begin{aligned} \textbf{Image search:} \quad & \mathbf{i}^* = \arg\max\nolimits_{\mathbf{i} \in I_{\text{cand}}} f(\mathbf{i}, \mathbf{s}_q) \\ \textbf{Image annotation:} \quad & \mathbf{s}^* = \arg\max\nolimits_{\mathbf{s} \in S_{\text{cand}}} f(\mathbf{i}_q, \mathbf{s}) \end{aligned} \tag{1}$$

This formulation is completely general: although we will, for evaluation purposes, define S_{cand} as the set of captions originally written for the images in I_{cand}, this does not have to be the case, and S_{cand} could also, for example, be defined implicitly via a caption generation system. In order to evaluate how well f generalizes to unseen examples, we will evaluate our system on test pools I_{test} and S_{test} that are drawn from the same domain but are disjoint from the training data $D_{\text{train}} = (I_{\text{train}}, S_{\text{train}})$ and development data $D_{\text{dev}} = (I_{\text{dev}}, S_{\text{dev}})$.

The challenge in defining f lies in the fact that images and sentences are drawn from two different spaces, I and S. In this paper, we present two different kinds of image description systems. One is based on nearest-neighbor search (NN), the other uses a technique called Kernel Canonical Correlation Analysis (KCCA; Bach & Jordan, 2002; Hardoon et al., 2004). Both rely on a set of known image-sentence pairs $D_{\text{train}} = \{\langle \mathbf{i}, \mathbf{s} \rangle\}$.

3.1 Nearest-Neighbor Search for Image Description

Nearest-neighbor based systems use unimodal text and image similarity functions directly to first find the image-sentence pair in the training corpus D_{train} that contains the closest item to the query, and then score the items in the other space by their similarity to the other item in this pair:

$$\begin{aligned} \textbf{Image retrieval:} \quad & f_{\text{NN}}(\mathbf{i}, \mathbf{s}_q) = f_I(\mathbf{i}^{\text{NN}}, \mathbf{i}) \quad \text{for } \langle \mathbf{i}^{\text{NN}}, \mathbf{s}^{\text{NN}} \rangle = \underset{\langle \mathbf{i}^t, \mathbf{s}^t \rangle \in D_{\text{train}}}{\arg\max} f_S(\mathbf{s}_q, \mathbf{s}^t) \\ \textbf{Image annotation:} \quad & f_{\text{NN}}(\mathbf{i}_q, \mathbf{s}) = f_S(\mathbf{s}^{\text{NN}}, \mathbf{s}) \quad \text{for } \langle \mathbf{i}^{\text{NN}}, \mathbf{s}^{\text{NN}} \rangle = \underset{\langle \mathbf{i}^t, \mathbf{s}^t \rangle \in D_{\text{train}}}{\arg\max} f_I(\mathbf{i}_q, \mathbf{i}^t) \end{aligned} \tag{2}$$

Despite their simplicity, such nearest-neighbor systems are non-trivial baselines: for the task of annotating images with tags or keywords, methods which annotate unseen images with the tags of their nearest neighbors among training images are known to achieve competitive performance (Makadia et al., 2010), and similar methods have recently been proposed for image description (Ordonez et al., 2011). Since the task we address here does not allow us to return items from the training data, but requires us to rerank a pool of unseen captions or images, our nearest-neighbor search requires two similarity functions. All of our nearest-neighbor systems use the same image representation as our KCCA-based systems, described in Section 3.3. Our main nearest-neighbor system, NN ($\text{NN5}^{\text{idf}}_{\text{F1}}$), treats the five captions associated with each training image as a single document. It then reweights each token by its inverse document frequency (IDF) λ_w, and defines the similarity of two sentences as the F1-measure (harmonic mean of precision and recall) computed over their IDF-reweighted bag-of-words representation. If $D_{\text{train}}(w)$ is the subset of training images in whose captions word w appears at least once, the inverse document frequency (IDF) of w is defined as $\lambda_w = \log \frac{|D_{\text{train}}|}{|D_{\text{train}}(w)|+1}$. IDF-reweighting is potentially helpful for our task, since words that describe fewer images may be particularly discriminative between captions.

In the appendix, we provide results for NN systems that use the same text representation as two of our KCCA systems.

3.2 Kernel Canonical Correlation Analysis for Image Description

Most of the systems we present are based on a technique called Kernel Canonical Correlation Analysis (Bach & Jordan, 2002; Hardoon et al., 2004). We first provide a brief introduction, and then explain how we apply it to our task.

3.2.1 Kernel Canonical Correlation Analysis (KCCA)

KCCA is an extension of Canonical Correlation Analysis (Hotelling, 1936), which takes training data consisting of pairs of corresponding items $\langle \mathbf{x}_i, \mathbf{y}_i \rangle$ drawn from two different feature spaces ($\mathbf{x}_i \in \mathcal{X}, \mathbf{y}_i \in \mathcal{Y}$), and finds maximally correlated linear projections $\alpha\mathbf{x}$ and $\beta\mathbf{y}$ of both sets of items into a newly induced common space $\mathcal{Z}$. Since linear projections of the raw features may not capture the patterns that are necessary to explain the pairing of the data, KCCA implicitly maps the original items into higher-order spaces $\mathcal{X}'$ and $\mathcal{Y}'$ via kernel functions $K_{\mathcal{X}} = \langle \phi_{\mathcal{X}}(\mathbf{x}_i) \cdot \phi_{\mathcal{X}}(\mathbf{x}_j) \rangle$, which compute the dot product of two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in a higher-dimensional space $\mathcal{X}'$ without requiring the explicit computation of the mapping $\phi_{\mathcal{X}}$. KCCA then operates on the two resulting kernel matrices $\mathbf{K}_{\mathcal{X}}[i,j] = \langle \phi_{\mathcal{X}}(\mathbf{x}_i) \cdot \phi_{\mathcal{X}}(\mathbf{x}_j) \rangle$ and $\mathbf{K}_{\mathcal{Y}}[i,j] = \langle \phi_{\mathcal{Y}}(\mathbf{y}_i) \cdot \phi_{\mathcal{Y}}(\mathbf{y}_j) \rangle$ which evaluate the kernel functions on pairwise combinations of items in the training data. It returns two sets of projection weights, α^* and β^*, which maximize the correlation between the two (projected) kernel matrices:

$$(\alpha^*, \beta^*) = \arg\max_{\alpha,\beta} \frac{\alpha' \mathbf{K}_X \mathbf{K}_Y \beta}{\sqrt{(\alpha' \mathbf{K}_X^2 \alpha + \kappa \alpha' \mathbf{K}_X \alpha)(\beta' \mathbf{K}_Y^2 \beta + \kappa \beta' \mathbf{K}_Y \beta)}} \tag{3}$$

This can be cast as a generalized eigenproblem $(\mathbf{K}_X + \kappa\mathbf{I})^{-1}\mathbf{K}_Y(\mathbf{K}_Y + \kappa\mathbf{I})^{-1}\mathbf{K}_X\alpha = \lambda^2\alpha$, and solved by partial Gram-Schmidt orthogonalization (Hardoon et al., 2004; Socher & Li, 2010). The regularization parameter κ penalizes the size of possible solutions, and is used to avoid overfitting, which arises when the matrices are invertible.

One disadvantage of KCCA is that it requires the two kernel matrices of the training data to be kept in memory during training. This becomes prohibitive with very large data sets, but does not cause any problems here, since our training data consists of only 6,000 items (see Section 4.1).

3.2.2 Using KCCA to Associate Images and Sentences

KCCA has been successfully used to associate images (Hardoon et al., 2004; Hwang & Grauman, 2012; Hardoon et al., 2006) or image regions (Socher & Li, 2010) with individual words or sets of tags. In our case, the two original spaces $\mathcal{X} = I$ and $\mathcal{Y} = S$ correspond to images and sentences that describe them. Images $\mathbf{i} \in I$ are first mapped to vectors $K_I(\mathbf{i})$ whose elements $K_I(\mathbf{i})(t) = K_I(\mathbf{i}_t, \mathbf{i})$ evaluate the image kernel function K_I on $\mathbf{i}$ and the t-th image in D_{train}. Similarly, sentences $\mathbf{s} \in S$ are mapped to vectors $K_S(\mathbf{s})$ that evaluate the sentence kernel function K_S on $\mathbf{s}$ and the sentences in D_{train}. The learned projection weights (α^*, β^*) then map $K_I(\mathbf{i})$ and $K_S(\mathbf{s})$ into our induced space $\mathcal{Z}$, in which we expect images to appear near sentences that describe them well. In a KCCA-based image annotation or search system, we therefore define f as the cosine similarity (sim) of points in this new space:

$$f_{\text{KCCA}}(\mathbf{i}, \mathbf{s}) = \text{sim}(\alpha K_I(\mathbf{i}), \beta K_S(\mathbf{s})) \tag{4}$$

We now describe the image and text kernels used by our KCCA systems.

3.3 Image Kernels

In contrast to much of the work done on image description, which assumes the existence of a large number of preexisting detectors, the image representations used in this paper are very basic, in that they rely only on three different kinds of low-level pixel-based perceptual features that capture color, texture (Varma & Zisserman, 2005) and shape information in the form of SIFT descriptors (Lowe, 2004; Vedaldi & Fulkerson, 2008). We believe that this establishes an important baseline, and leave the question of how more complex image representations affect performance to future work. We use two different kinds of kernels: a histogram kernel K^{Histo}, which represents each image as a single histogram of feature values and computes the similarity of two images as the intersection of their histograms, and a pyramid kernel K^{Py} (Lazebnik, Schmid, & Ponce, 2009), which represents each image as a pyramid of nested regions, and computes the similarity of two images in terms of the intersection of the histograms of corresponding regions. In both cases, we compute a separate kernel for each of the three types of image features and average their result.

3.3.1 The Histogram Kernel (K^{Histo})

Each image $\mathbf{x_i}$ is represented as a histogram H_i of discrete-valued features, such that $H_i(v)$ is the fraction of pixels in $\mathbf{x_i}$ with value v. The similarity of two images $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as the intersection of their histograms, i.e. the percentage of pixels that can be mapped onto a pixel with the same feature value in the other image:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{v=1}^{V} \min(H_i(v), H_j(v)) \tag{5}$$

We combine three kernels based on different kinds of visual features: K_C captures color, represented by the three CIELAB coordinates. K_T captures texture, represented by descriptors which capture edge information at different orientations centered on the pixel (Varma & Zisserman, 2005). K_S is based on SIFT descriptors, which capture edge and shape information in a manner that is invariant to changes in rotation and illumination, and have been shown to be distinct across possible objects of an image Lowe, 2004; Vedaldi & Fulkerson, 2008. We use 128 color words, 256 texture words and 256 SIFT words, obtained in an unsupervised fashion by K-means clustering on 1,000 points of 200 images from the PASCAL 2008 data set (Everingham et al., 2008). Our final histogram kernel K^{Histo} is the average of the responses of the three kernels $K_C^{\text{Histo}}, K_T^{\text{Histo}}, K_S^{\text{Histo}}$, taken to the pth power:

$$K^{\text{Histo}}(\mathbf{x}_i, \mathbf{x}_j) = \left[\frac{1}{3} \sum_{F \in \{\text{C,S,T}\}} K_F^{\text{Histo}}(\mathbf{x}_i, \mathbf{x}_j)\right]^p \tag{6}$$

3.3.2 The Pyramid Kernel K^{Py}

The spatial pyramid kernel (Lazebnik et al., 2009) is a generalization of the histogram kernel that captures similarities not just at a global, but also at a local level. Each image $\mathbf{x}_i$ is represented at multiple levels of scale l ($l \in \{0,1,2\}$) such that each level partitions the image into a smaller and smaller grid of $C_l = 2^l \times 2^l$ cells ($C_0 = 1$, $C_1 = 4$, $C_2 = 16$), and each cell c is represented as a histogram H_{ic}. The similarity of images $\mathbf{x}_i$ and $\mathbf{x}_j$ at level l, I^l_{ij}, is in turn defined as the sum of the histogram similarities of their corresponding cells $0_l, ..., C_l$ at this level:

$$I^l_{ij} = \sum_{c=0_l}^{C_l} \sum_{v=1}^{V} \min(H_{ic}(v), H_{jc}(v)) \tag{7}$$

Although similarities at level l subsume those at a more fine-grained level $l+1$ ($I^l_{ij} \geq I^{l+1}_{ij}$), similarities that hold at a more fine-grained level are deemed more important, since they indicate a greater local similarity. The pyramid kernel therefore proceeds from the most fine-grained ($l = L$) down to the coarsest (whole-image) scale ($l = 0$), and weights the similarities first encountered at level l ($I^l_{ij} - I^{l+1}_{ij}$) by $\frac{1}{2^{L-l}}$:

$$\begin{aligned} K^{\mathrm{Py}}(\mathbf{x}_i, \mathbf{x}_j) &= I^L_{ij} + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (I^l_{ij} - I^{l+1}_{ij}) \\ &= \frac{1}{2^L} I^0_{ij} + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l_{ij} \end{aligned} \tag{8}$$

We again compute three separate pyramid kernels $K^{\mathrm{Py}}_C, K^{\mathrm{Py}}_T, K^{\mathrm{Py}}_S$ based on the same color, texture and SIFT features as described above, and combine them into a single pyramid kernel K^{Py}, as in equation 6.

3.4 Basic Text Kernels

We examine three different basic text kernels: a bag of words (BoW) kernel, Hwang and Grauman's (2012) TagRank kernel, and a truncated string kernel (Tri).

3.4.1 The Bag of Words Kernel (BoW)

Since bag-of-words representations have been successfully used for other tasks involving text and images (e.g., Grangier & Bengio, 2008; Hardoon et al., 2006), we include a basic bag of words kernel, which ignores word order and represents each caption simply as a vector of word frequencies. The BoW kernel function is defined as the cosine similarity of the corresponding bag of words vectors. We either merge the five captions of each training item into a single document (BoW5), or reduce each training item to a single, arbitrarily chosen, caption (BoW1). A word's frequency can also be reweighted by its IDF-score. As in the nearest neighbor approach, the IDF-weight of a word w is defined as $\lambda_w = \log \frac{|D_{\mathrm{train}}|}{|D_{\mathrm{train}}(w)|+1}$, where $D_{\mathrm{train}}(w)$ is the subset of training images in whose captions word w appears at least once. We found the square root of λ_w (BoW5$^{\sqrt{\text{idf}}}$) to give better results than the standard IDF-score λ_w (BoW5$^{\text{idf}}$).

3.4.2 The Tag Rank Kernel (TagRank)

Hwang and Grauman (2012) apply KCCA to keyword-based image annotation and retrieval. They focus on a data set where each image is paired with a list of tags ranked by their importance, and propose a new kernel for this kind of data. This so-called tag rank kernel (TagRank) is a variant of the bag of words kernel that aims to capture the relative importance of tags by reweighting them according to their position in this list. Although Hwang and Grauman do not evaluate the ability of their system to associate images with entire sentences, they also consider another data set in which the lists of "tags" correspond to the words of descriptive captions, and argue that the linear order of words in these captions also reflects the relative importance of the corresponding objects in the image, so that words that appear at the beginning of the sentence describe more salient aspects of the image.

In the TagRank kernel, each sentence is represented as two vectors, $\vec{a}$ and $\vec{r}$. In $\vec{a}$, the weight of each word is based on its absolute position, so that the first words in each sentence are always assigned a high weight. In this "absolute tag rank" representation, each caption $\mathbf{s}$ is mapped to a vector $\vec{a} = [\vec{a}(1) \ldots \vec{a}(|V|)]$, where $|V|$ is the size of the vocabulary. $\vec{a}(i)$ depends on the absolute position p_i of w_i in $\mathbf{s}$ (if w_i occurs multiple times in $\mathbf{s}$, p_i is averaged over all its positions). If w_i does not occur in $\mathbf{s}$, $\vec{a}(i) = 0$. Otherwise,

$$\vec{a}(i) = \frac{1}{log_2(1 + p_i)} \tag{9}$$

In $\vec{r}$, the weight of a word depends on how its current position compares to the distribution of positions it occupies in the training data. The intuition behind this "relative rank" representation is that words should have a higher weight when they occur earlier in the sentence than usual. Here, each caption $\mathbf{s}$ is mapped to a vector $\vec{r} = [\vec{r}(1) \ldots \vec{r}(V)]$ of relative tag ranks. Again, when w_i does not appear in $\mathbf{s}$, $\vec{r}(i) = 0$. Otherwise w_i's relative tag rank $\vec{r}(i)$ indicates what percent of its occurrences in the training data appear after position p_i. Defining n_{ik} as the number of times word w_i appears in position k in the training data, and $n_i = \sum_k n_{ik}$ as the total frequency of w_i in the training data:

$$\vec{r}(i) = 1 - \frac{\sum_{k=1}^{p_i} n_{ik}}{n_i} \tag{10}$$

The final kernel K_T is given by the average of two χ^2 kernels computed over $\vec{r}$ and $\vec{a}$ (Ω and Ω' are normalization terms):

$$K_T(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2}\left[\exp\left(\frac{-1}{2\Omega}\sum_{k=1}^{V}\frac{(\vec{r}_i(k) - \vec{r}_j(k))^2}{\vec{r}_i(k) + \vec{r}_j(k)}\right) + \exp\left(\frac{-1}{2\Omega'}\sum_{k=1}^{V}\frac{(\vec{a}_i(k) - \vec{a}_j(k))^2}{\vec{a}_i(k) + \vec{a}_j(k)}\right)\right] \tag{11}$$

Since each image in our training data is associated with multiple, independently generated captions, we evaluate the kernel separately on each sentence pair and average the response, instead of treating the multiple sentences as a single document.

The TagRank kernel is relatively sensitive to overall sentence length, especially in cases where the subject is preceded by multiple adjectives or other modifiers (*'a very large brown dog'* vs. *'a dog'*). In English, the absolute tag rank will generally assign very high weights to the subjects of sentences, lower weight to verbs, and even lower weight to objects or scene descriptions, which tend to follow the main verb. The relative tag rank may not downweight verbs, objects and scene descriptions as much (as long as they are always used in similar positions in the sentence).

3.4.3 The Trigram Kernel (Tri)

Since bag-of-words representations ignore which words appear close to each other in the sentence, they lose important information: an image of *'a small child with red hair playing with a large brown dog on white carpet'* looks quite different from one of *'a small white dog playing with a large red ball on brown grass'*, although both descriptions share the majority of their words. To capture this information, we define a trigram kernel as a truncated variant of string kernels (Shawe-Taylor & Cristianini, 2004) that considers not just how many single words two captions share, but also how many short sequences (pairs and triples) of words occur in both.

A word sequence $\mathbf{w} = w_1 ... w_k$ is an ordered list of words. A sentence $\mathbf{s} = s_1 ... s_n$ contains $\mathbf{w}$ ($\mathbf{w} \in \mathbf{s}$) as long as the words in $\mathbf{w}$ appear in $\mathbf{s}$ in the order specified by $\mathbf{w}$. That is, the sentence *'A large white dog runs and catches a red ball on the beach'* (when lemmatized) contains both the subject-verb-object triple *'dog catch ball'* and the subject-verb-location triple *'dog run beach'*. Formally, every substring $(i, j) = s_i ... s_j$ in $\mathbf{s}$ that starts with $s_i = w_1$, ends in $s_j = w_k$, and contains $\mathbf{w}$ is considered a match between $\mathbf{s}$ and $\mathbf{w}$. $M_{\mathbf{s},\mathbf{w}}$ is the set of all substrings in $\mathbf{s}$ that match the sequence $\mathbf{w}$:

$$M_{\mathbf{s},\mathbf{w}} = \{(i, j) \mid \mathbf{w} = w_1 ... w_k \in s_i ... s_j, w_1 = s_i, w_k = s_j\} \tag{12}$$

When $\mathbf{w}$ is restricted to individual words ($k = 1$), string kernels are identical to the standard BoW kernel.

A match between strings $\mathbf{s}$ and $\mathbf{s}'$ is a pair of substrings $(i, j) \in \mathbf{s}$ and $(i', j') \in \mathbf{s}'$ that both match the same word sequence $\mathbf{w}$. Standard string kernels $K(\mathbf{s}, \mathbf{s}')$ weight matches by a factor $\lambda^{(j-i+1)+(j'-i'+1)}$ that depends on an adjustable parameter λ and the respective length of the matching substrings:

$$K(\mathbf{s}, \mathbf{s}') = \sum_{\mathbf{w}} \sum_{(i,j) \in M_{\mathbf{s},\mathbf{w}}} \sum_{(i',j') \in M_{\mathbf{s}',\mathbf{w}}} \lambda^{(j-i+1)+(j'-i'+1)} \tag{13}$$

In order to distinguish between the length of the matching subsequence, $l(\mathbf{w})$, and the length of the gaps in (i, j) and (i', j'), we replace λ by two parameters λ_{m}, λ_{g}, and reformulate this as:

$$K(\mathbf{s}, \mathbf{s}') = \sum_{\mathbf{w}} \sum_{(i,j) \in M_{\mathbf{s},\mathbf{w}}} \sum_{(i',j') \in M_{\mathbf{s}',\mathbf{w}}} \lambda_{\mathrm{m}}^{2l(\mathbf{w})} \lambda_{\mathrm{g}}^{(j-i+1)+(j'-i'+1)-2l(\mathbf{w})} \tag{14}$$

We found that a gap score of $\lambda_{\mathrm{g}} = 1$, which means that gaps are not penalized, and a match score of $\lambda_{\mathrm{m}} = 0.5$ perform best on our task.

Although string kernels are generally defined over sequences of arbitrary length ($k \leq \infty$), we found that allowing longer sequences did not seem to impact performance on our task but incurred a significant computational cost. Intuitively, word pairs and triplets represent most of the linguistic information we need to capture beyond the BoW representation, since they include head-modifier dependencies such as *large-dog* vs. *small-dog* and subject-verb-object dependencies such as *child-play-dog* vs. *dog-play-ball*. We therefore consider only sequences up to length $k \leq 3$. With $\mathbf{w}$ restricted to sequences of length $k \leq 3$ and $m_{\mathbf{s},\mathbf{w}} = |M_{\mathbf{s},\mathbf{w}}|$, this yields the following 'trigram' kernel (Tri):

$$K_{\text{TRI}}(\mathbf{s}, \mathbf{s}') = \sum_{\mathbf{w}:k\leq 3} m_{\mathbf{s},\mathbf{w}} m_{\mathbf{s}',\mathbf{w}} \lambda_{\text{m}}^{2l(\mathbf{w})} \tag{15}$$

To deal with differences in sentence length, we normalize the kernel response between two examples by the geometric mean of the two example responses with themselves.

Since the trigram kernel also captures sequences that are merely coincidental, such as *'large white red'*, it may seem advantageous to use richer syntactic representations such as dependency tree kernels (Moschitti, Pighin, & Basili, 2008), which only consider word tuples that correspond to syntactic dependencies. However, such kernels are significantly more expensive to compute, and initial experiments indicated that they may not perform as well as the trigram kernel. We believe that this is due to the fact that our image captions contain little syntactic variation, and that hence surface word order may be sufficient to differentiate e.g. between the agent of an action (whose mention will be the subject of the sentence) and other participants or entities (whose mentions will appear after the verb). On the other hand, many of our image captions contain a lot of syntactic ambiguity (e.g. multiple prepositional phrases), and a vocabulary that is very distinct from what standard parsers are trained on. It may be that we were not able to benefit from using a richer representation simply because we were not able to recover it with sufficient accuracy.

In order to capture the relative importance of words, we can also reweight sequences by the IDF (or $\sqrt{\text{idf}}$) weight of their words. With λ_w defined as before, the IDF-weight of a sequence $\mathbf{w} = w_i ... w_j$ is $\lambda_{\mathbf{w}} = \prod_{k=i}^{j} \lambda_{w_k}$. The $\sqrt{\text{idf}}$-weighted trigram kernel $K_{\text{TRI}^{\sqrt{\text{idf}}}}$ (Tri5$^{\sqrt{\text{idf}}}$) is therefore

$$K_{\text{TRI}^{\sqrt{\text{idf}}}}(\mathbf{s}, \mathbf{s}') = \sum_{\mathbf{w}:k\leq 3} \lambda_{\mathbf{w}} m_{\mathbf{s},\mathbf{w}} m_{\mathbf{s}',\mathbf{w}} \lambda_{\text{m}}^{2l(\mathbf{w})} \tag{16}$$

3.5 Extending the Trigram Kernel with Lexical Similarities

One obvious shortcoming of the basic text kernels is that they require exact matches between words, and cannot account for the fact that the same situation, event, or entity can be described in a variety of ways (see Figure 2 for examples). One way of capturing this linguistic diversity is through lexical similarities which allow us to define partial matches between words based on their semantic relatedness. Lexical similarity have found success in other tasks, e.g. semantic role labeling (Croce, Moschitti, & Basili, 2011), but have not been fully exploited for image description. Ordonez et al. (2011) define explicit equivalence classes of synonyms and hyponyms to increase the natural language vocabulary corresponding to each of their object detectors (e.g. the word *"Dalmatian"* may trigger the dog detector), but do not change the underlying, pre-trained detectors themselves, ignoring the potential variation of appearance between, e.g., different breeds of dog. Similarly, Yang et al.'s (2011) generative model can produce a variety of words for each type of detected object or scene, but given an object or scene label, the word choice itself is independent of the visual features. We therefore also investigate the effect of incorporating different kinds of lexical similarities into the trigram kernel that allow us to capture partial matches between words. We did not explore the effect of incorporating lexical similarities into the tag-rank kernel, since it is unclear how they should affect the computation of ranks within a sentence.

3.5.1 String Kernels with Lexical Similarities

Since standard lexical similarities $\mathrm{sim}_S(w, w_i)$ do not necessarily yield valid kernel functions, we follow Bloehdorn, Basili, Cammisa, and Moschitti (2006) and use these similarities to map each word w to vectors $\vec{w}_S$ in an N-dimensional space, defined by a fixed vocabulary of size N. Each vector component $\vec{w}_S(i)$ corresponds to the similarity of w and w_i as defined by S:

$$\vec{w}_S(i) = \mathrm{sim}_S(w, w_i) \tag{17}$$

We then define the corresponding word kernel function $\kappa_S(w, w')$, which captures the partial match of words w and w' according to S, as the cosine of the angle between $\vec{w}_S$ and $\vec{w}'_S$:

$$\kappa_S(w, w') = \cos(\vec{w}_S, \vec{w}'_S) \tag{18}$$

S may only be defined over a subset of the vocabulary. The similarity of words outside of its vocabulary is defined by the identify function, as in the standard string kernel.

The similarity of sequences $\mathbf{w}$ and $\mathbf{w}'$ of length l is defined as the product of the word kernels over the corresponding pairs of sequence elements w_i, w'_i:

$$\sigma_S(\mathbf{w}, \mathbf{w}') = \prod_{i=1}^{l} \kappa_S(w_i, w'_i) \tag{19}$$

If $\sigma_S(\mathbf{w}) = \{\mathbf{w}' | \sigma_S(\mathbf{w}', \mathbf{w}) > 0, l(\mathbf{w}') = l(\mathbf{w})\}$ is the set of sequences that have a non-zero match with $\mathbf{w}$, the string kernel K_S with similarity S is:

$$K_S(\mathbf{s}, \mathbf{s}') = \sum_{\mathbf{w}} \sum_{\mathbf{w}' \in \sigma_S(\mathbf{w})} m_{\mathbf{s},\mathbf{w}} m_{\mathbf{s}',\mathbf{w}'} \lambda_{\mathrm{m}}^{2l(\mathbf{w})} \sigma_S(\mathbf{w}', \mathbf{w}) \tag{20}$$

To obtain the IDF-weighted version of this kernel, $K_S^{\sqrt{\mathrm{idf}}}(\mathbf{s}, \mathbf{s}')$, the inner term is multiplied by $\sqrt{\lambda_{\mathbf{w}} \lambda_{\mathbf{w}'}}$:

$$K_S(\mathbf{s}, \mathbf{s}') = \sum_{\mathbf{w}} \sum_{\mathbf{w}' \in \sigma_S(\mathbf{w})} \sqrt{\lambda_{\mathbf{w}} \lambda_{\mathbf{w}'}} m_{\mathbf{s},\mathbf{w}} m_{\mathbf{s}',\mathbf{w}} \lambda_{\mathrm{m}}^{2l(\mathbf{w})} \sigma_S(\mathbf{w}', \mathbf{w}) \tag{21}$$

In our experiments, we use the trigram variants of these kernels, and restrict $\mathbf{w}$ again to sequences of length $k \leq 3$.

We consider three different kinds of lexical similarities: the WordNet-based Lin similarity (Lin, 1998) (σ_{Lin}), a distributional similarity metric (σ_D), and a novel alignment-based similarity metric (σ_{A}), which takes advantage of the fact that each image is associated with five independently generated captions. All metrics are computed on our training corpus. Distributional similarity is also computed on the British National Corpus (BNC Consortium, 2007). Both corpora are lemmatized, and stop words are removed before similarities are computed. Since almost any pair of words will have a non-zero similarity, the word kernel matrices are very dense, but since most of these similarities are very close to zero, they have very little effect on the resulting kernel. We therefore zero out entries smaller than 0.05 in the alignment-based kernel κ_{A} and less than 0.01 in any distributional kernel $\kappa_{D_{\mathcal{C}}}$.

3.5.2 The Lin Similarity Kernel (σ_{Lin})

Lin's (1998) similarity relies on the hypernym/hyponym relations in WordNet (Fellbaum, 1998) as well as corpus statistics. WordNet is a directed graph in which the nodes ("synsets") represent word senses and the edges indicate is-a relations: a parent sense (e.g., dog_1) is a hypernym of its children (e.g., poodle_1 or dachshund_1). Kernels based on Lin's similarity have been found to perform well on tasks such as text categorization (Bloehdorn et al., 2006). But with the exception of Farhadi et al. (2010), who incorporate Lin's similarity into their model, but do not evaluate what benefit they obtain from it, WordNet's hypernym-hyponym relations have only been used superficially for associating images and text (Weston et al., 2010; Ordonez et al., 2011; Gupta et al., 2012). The Lin similarity of two word senses s_i, s_j is defined as

$$\text{sim}_{\text{Lin}}(\text{s}_i, \text{s}_j) = \frac{2 \log P(LCS(\text{s}_i, \text{s}_j))}{\log P(\text{s}_i) + \log P(\text{s}_j)} \tag{22}$$

$LCS(\text{s}_1, \text{s}_2)$ refers to the lowest common subsumer of s_1 and s_2 in WordNet, i.e. the most specific synset that is an ancestor (hypernym) of both s_1 and s_2. $P(\text{s})$ is the probability that a randomly drawn word is an instance of synset s or any of its descendants (hyponyms). We use our training data to estimate $P(\text{s})$, and follow Bloehdorn et al. (2006) in assigning each word w its most frequent (first) noun sense s_w in WordNet 3.0. Hence, we represent each word w with WordNet sense s as a vector $\vec{w}_{Lin}$ of Lin similarities over its hypernyms $H(\text{s}_w)$:

$$\vec{w}_{Lin}(i) \quad = \quad \begin{cases} \frac{2 \times log(f(\text{s}_i))}{log(f(\text{s})) + log(f(\text{s}_i))} & \text{s}_i \in H(\text{s}) \\ 1 & \text{s}_w = \text{s}_i \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

3.5.3 Distributional Similarity ($\sigma_{D_{\mathcal{C}}}$)

Distributional similarity metrics are based on the observation that words that are similar to each other tend to appear in similar contexts (Jurafsky & Martin, 2008). The components of $\vec{w}_{D_{\mathcal{C}}}$ are the non-negative pointwise mutual information scores (PMI) of w and w_i, computed on the corpus $\mathcal{C}$:

$$\vec{w}_{D_{\mathcal{C}}}(i) = \max\left[0, \log_2 \frac{P_{\mathcal{C}}(w, w_i)}{P_{\mathcal{C}}(w) P_{\mathcal{C}}(w_i)}\right] \tag{24}$$

$P_{\mathcal{C}}(w)$ is the probability that a random sentence in $\mathcal{C}$ contains w, and $P_{\mathcal{C}}(w, w_i)$ is the probability that a random sentence in $\mathcal{C}$ contains both w and w_i. We compute two variants of the same metric: $\sigma_{D_{\mathrm{ic}}}$ is computed on the image captions in our training corpus, and is defined over the cooccurrences of the 1,928 words that appear at least 5 times in this corpus, while $\sigma_{D_{\mathrm{BNC}}}$ uses the British National Corpus (BNC Consortium, 2007), and is defined for the 1,874 words that appear at least 5 times in both corpora, but considers their PMI scores against the 141,656 words that appear at least 5 times in the BNC.

3.5.4 Alignment-Based Similarity (σ_{A})

We also propose a novel, alignment-based, similarity metric (σ_{A}), which takes advantage of the fact that each image is associated with five independently generated captions, and is specifically designed to capture how likely two words are to describe the same event or entity in our data set. We borrow the concept of alignment from machine translation (Brown, Pietra, Pietra, & Mercer, 1993), but instead of aligning the words of sentences in two different languages, we align pairs of captions that describe the same image. This results in a similarity metric that has better coverage on our data set than WordNet based metrics, and is much more specific than distributional similarities which capture broad topical relatedness rather than semantic equivalence. Instead of aligning complete captions, we have found it beneficial to align nouns and verbs independently of each other, and to ignore all other parts of speech. We create two versions of the training corpus, one consisting of only the nouns of each caption, and another one consisting only of the verbs of each caption. We then use Giza++ (Och & Ney, 2003) to train IBM alignment models 1–2 (Brown et al., 1993) over all pairs of noun or verb captions of the same image to obtain two sets of translation probabilities, one over nouns ($P_n(\cdot|w)$) and one over verbs ($P_v(\cdot|w)$). Finally, we combine the noun and verb translation probabilities as a sum weighted by the relative frequency with which the word w was tagged as a noun ($P_n(w)$) or verb ($P_v(w)$) in the training corpus. The ith entry in w_{A} is therefore:

$$\vec{w}_{\mathrm{A}}(i) = P_n(w_i|w)P_n(w) + P_v(w_i|w)P_v(w) \tag{25}$$

We define the noun and verb vocabulary as follows: words that appear at least 5 times as a noun, and are tagged as a noun in at least 50% of their occurrences, are considered nouns. But since verbs are more polysemous than nouns (leading to broader translation probabilities) and are often mistagged as nouns in our domain, we only include those words as verbs that are tagged as verbs at least 25 times, and in at least 25% of their occurrences. This results in 1180 noun and 143 verb lemmas, including 11 that can be nouns or verbs. We use the OpenNLP POS tagger before lemmatization.

3.5.5 Comparing the Similarity Metrics (Figure 3)

Figure 3 illustrates the different similarity metrics, using the words *rider* and *swim* as examples. While distributional similarities are high for words that are topically related (e.g., *swim* and *pool*), the alignment similarity tends to be high for words that can be used to describe the same entity (usually synonyms or hyper/hyponyms) or activity such as *swim* or *paddle*. Distributional similarities that are obtained from the image captions are very specific to our domain. The BNC similarities are much broader and help overcome data sparsity, although the BNC has relatively low coverage of the kinds of sports that occur in our data set. The Lin similarity associates *swim* with hypernyms such as *sport* and *activity*,

Comparing similarity metrics: The five words most similar to *rider* and *swim*								
	Alignment		**Distributional**				**Lin**	
Corpus	S_{train}		S_{train}		BNC		S_{train}	
w	w_i	$\sigma_{\mathbf{A}}$	w_i	$\sigma_{D_{\mathbf{ic}}}$	w_i	$\sigma_{D_{\mathbf{BNC}}}$	w_i	$\sigma_{\mathbf{Lin}}$
rider	*biker*	0.86	*bike*	0.41	*ride*	0.21	*traveler*	0.94
	bicyclist	0.82	*dirt*	0.35	*horse*	0.20	*cyclist*	0.89
	cyclist	0.79	*motocross*	0.33	*race*	0.19	*bicyclist*	0.89
	bmx	0.75	*motorcycle*	0.33	*bike*	0.17	*horseman*	0.84
	bicycler	0.73	*ride*	0.33	*jockey*	0.16	*jockey*	0.84
swim	*retrieve*	0.56	*pool*	0.53	*fish*	0.21	*bathe*	0.85
	paddle	0.54	*trunk*	0.35	*water*	0.18	*sport*	0.85
	dive	0.52	*water*	0.34	*sea*	0.18	*football*	0.77
	come	0.38	*dive*	0.30	*pool*	0.18	*activity*	0.75
	wade	0.31	*goggles*	0.29	*beach*	0.17	*soccer*	0.73

Figure 3: A comparison of lexical similarities for the noun *rider* and the verb *swim*

or other kinds of sport such as *football* or *soccer*. This makes it the least suitable similarity for our task (see also Section 4.3.4 for experimental results), since these terms should not be considered similar for our purposes of identifying the different ways in which visually similar events or entities can be described.

3.5.6 Combining Different Similarities

Combining the different distributional and the alignment-based similarities allows us to capture the different strengths of each method. We define an averaged similarity which captures aspects of the distributional similarities computed over both corpora:

$$\kappa_{D_{\text{BNC,ic}}}(w, w') = \frac{\kappa_{D_{\text{BNC}}}(w, w') + \kappa_{D_{\text{ic}}}(w, w')}{2} \tag{26}$$

For every distributional kernel $\kappa_D(w, w')$, we also define a variant $\kappa_{D+A}(w, w')$ which incorporates alignment-based similarities by taking the maximum of either kernel:[4]

$$\kappa_{D+A}(w, w') = \max(\kappa_A(w, w'), \kappa_A(w, w')) \tag{27}$$

4. Evaluation Procedures and Metrics for Image Description

In order to evaluate scoring functions $f(\mathbf{i}, \mathbf{s})$ for image-caption pairs, we need to evaluate their ability to associate previously unseen images and captions with each other. In analogy to caption generation systems, we first examine metrics that aim to measure the quality of a single image-description pair (Section 4.2). Here, we focus on the image annotation task, and restrict our attention to the first caption returned for each test item, and a subset of our systems. We collect graded human judgments from small number of native speakers of American English, and investigate whether these “expert” judgments can be approximated

[4] This operation may not preserve the positive definiteness of the matrix required to be a valid kernel, but this simply means we effectively use (plain) CCA with this representation.

with automatically computed Bleu (Papineni et al., 2002) or Rouge (Lin & Hovy, 2003) scores, or with simpler crowdsourced human judgments that can be collected on a much larger scale. In Section 4.3, we consider approaches to evaluation that aim to measure the quality of the ranked list of image-caption pairs returned by each system, and allow us to evaluate a large number of systems. For reasons of space, we focus most of our discussion again on only a subset of our systems, and refer the interested reader to Appendix B for complete results. Since the candidate pool contains one sentence or image that was originally associated with the query image or sentence, we first compare systems by the rank and recall of this original item. These metrics can be computed automatically, but should only be considered lower bounds on actual performance, since each image may be associated with a number of captions that describe it well or perhaps with only minor errors. We then show that the crowdsourced human judgments can be mapped to binary relevance judgments that correlate well with the more fine-grained expert judgments, and consider metrics based on these relevance judgments.

4.1 Experimental Setup

We now describe the data, the tasks, and the systems we evaluate in our experiments.

4.1.1 The Data

Since the PASCAL 2008 data set contains only a total of 1,000 images, we perform our experiments exclusively on the Flickr 8K set. We split 8,000 images from this corpus (see Section 2.3) into three disjoint sets. The training data $D_{\text{train}} = \langle I_{\text{train}}, S_{\text{train}} \rangle$ consists of 6,000 images, each associated with five captions, whereas the test and development data, D_{test} and D_{dev}, each consist of 1,000 images associated with one, arbitrarily chosen, caption. All captions are preprocessed by spellchecking with Linux *spell*, normalizing compound words (e.g., *t-shirt, t shirt, and tee-shirt* → *t-shirt*), stop word removal, and lemmatization.

4.1.2 The Tasks

We evaluate our systems on two tasks, sentence-based image annotation (or description) and sentence-based image search. For image search, the task is to return a ranked list of the 1,000 images in I_{test} for each of the captions (queries) in S_{test}. Image annotation is defined analogously as a retrieval problem: the task is to return a ranked list of the 1,000 captions in S_{test} for each of the 1,000 test (query) images in I_{test}. In both cases, the ranked lists are produced independently for each of the 1,000 possible queries.

4.1.3 The Systems

We have a total of 30 different systems, each of which uses either a nearest-neighbor approach or KCCA, paired with a different combination of image and text representations. But for the purposes of discussing different evaluation metrics, we will focus on only a small number of these systems: the best-performing nearest-neighbor-based system, NN (NN5$^{\text{idf}}_{\text{F1}}$), and a small number of KCCA-based systems with with different text kernels: BoW1 and BoW5 both use the simple bag-of-words kernel. TagRank uses Hwang and Grauman's (2012) kernel, Tri5 uses the trigram kernel, and Tri5Sem (Tri5$^{\sqrt{\text{idf}}}_{\text{A},D_{\text{BNC+ic}}}$ in Appendix B) uses the

$\sqrt{\text{idf}}$-reweighted trigram kernel with all distributional and the alignment-based similarities. With the exception of BoW1, where we have arbitrarily selected a single caption for each training image, all other models use all five captions for the training images. For BoW5, we merge them into a single document. In all other cases, we follow Moschitti (2009) and sum the kernel responses over the cross product of sentences before normalization. All of these systems (including NN) use the pyramid kernel as their image representation. For the large-scale evaluations in Section 4.3, the scores of all models are given in Appendix B.

All our systems use Hardoon et al.'s (2004) KCCA implementation, which allows us to vary the regularization parameter κ. We also vary n, the number of dimensions (largest eigenvalues) in the learned projection The allowable values for these parameters were based on early exploratory experiments. In the experiments reported in this paper, κ is sampled from 4 possible values (0.1, 0.5, 1, 5), and n is chosen from 46 possible values in the range of (10, 6000). There are two additional parameters that are fixed in advance for each text image kernel pair: the image kernels are either squared or cubed, and the text kernels are regularized by multiplying the values on the diagonal by a factor d in the range of (1, 15).

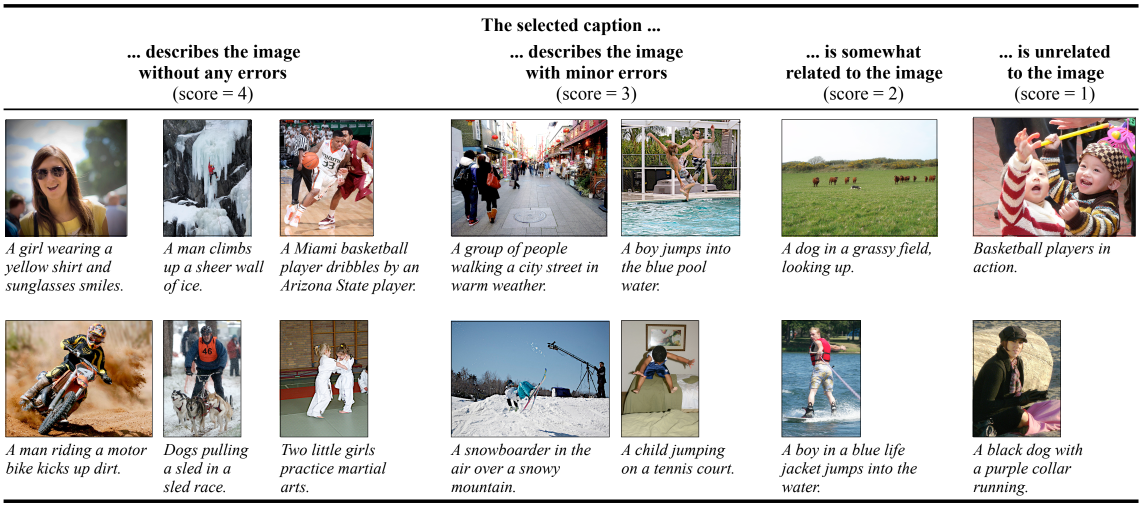
For each kernel and for each of the two tasks (image annotation and search), we then use the development set to pick five settings of n and κ that maximize the recall of the original item as the first result, five settings that maximize its recall among the first five results, and five settings that maximize its recall among the first ten results, yielding a total of 15 different models for each pair of kernels and each task. For each query image (annotation) or caption (search) in the test set, each of these 15 models returns a ranking of all 1,000 test items (sentences or images). To combine these 15 rankings, we use Borda counts (van Erp & Schomaker, 2000), a simple, deterministic method for rank aggregation: with N items to be ranked, each system assigns a score of $N - r$ to the item it ranks in position $r = 0...N-1$, and the final rank of each item is determined by the sum of its scores across all systems. We break ties between items by the median of their ranks across all models.

4.2 Metrics for the Quality of Individual Image-Caption Pairs

Before we consider metrics that consider the quality of the ranked list of results (Section 4.3), we first examine metrics that measure the quality of individual image-caption pairs.

4.2.1 Human Evaluation with Graded 'Expert' Judgments

'Expert' scores The decision of how well a caption describes an image ultimately requires human judgment. For the caption generation task, a number of different evaluation schemes have been proposed for image description: Ordonez et al. (2011) presented judges with a caption produced by their model and asked them to make a forced choice between a random image and the image the caption was produced for, and Kuznetsova et al. (2012) asked judges to choose between captions from two of their models for a given test image. Such forced choice tasks may give a clear ranking of models, but cannot be compared across different experiments unless the output of each system is made publicly available. One advantage of framing image description as a ranking task is that different systems can be compared directly on the same test pool. Forced choice evaluations also do not directly measure the quality of the captions. Following common practice in natural language generation, Yang et al. (2011) and Kulkarni et al. (2011) evaluated captions on a graded scale for relevance


The selected caption ...
... describes the image without any errors (score = 4)
... describes the image with minor errors (score = 3)
... is somewhat related to the image (score = 2)
... is unrelated to the image (score = 1)

A girl wearing a yellow shirt and sunglasses smiles.

A man climbs up a sheer wall of ice.

A Miami basketball player dribbles by an Arizona State player.

A group of people walking a city street in warm weather.

A boy jumps into the blue pool water.

A dog in a grassy field, looking up.

Basketball players in action.

A man riding a motor bike kicks up dirt.

Dogs pulling a sled in a sled race.

Two little girls practice martial arts.

A snowboarder in the air over a snowy mountain.

A child jumping on a tennis court.

A boy in a blue life jacket jumps into the water.

A black dog with a purple collar running.

Figure 4: Our 1–4 rating scale for the fine-grained expert judgments, with actual examples returned by our best model (Tri5Sem)

and readability, while Li et al. (2011) added a “creativity” score, and Mitchell et al. (2012) compared systems based on whether the captions describe the “main aspects” of the images, introduce objects in an appropriate order, are semantically correct, and seemed to have been written by a human.

Since the captions in our test pool are all produced by people, we do not need to evaluate their linguistic quality, and can focus on their semantic correctness. In order to obtain a fine-grained assessment of description quality, we asked three different judges to score image-caption pairs returned by our systems on a graded scale from 1 to 4. The judges were 21 adult native speakers of American English, mostly recruited from among the local graduate student population. In contrast to the anonymous crowdsourcing-based evaluation described in Section 4.3.2, we will refer to them as ‘experts’. The rating scale is illustrated in Figure 4 with actual examples returned by our models. A score of 4 means that the caption describes the image perfectly (without any mistakes), a score of 3 that the caption almost describes the image (minor mistakes are allowed, e.g. in the number of entities), whereas a score of 2 indicates that the caption only describes some aspects of the image, but could not be used as its description, and a score of 1 indicates that the caption bears no relation to the image. The online appendix to this paper contains our annotation guidelines. Annotators took on average ten minutes per 50 image-caption pairs, and all image-caption pairs were judged independently by three different annotators. Inter-annotator agreement, measured as Krippendorff’s (2004) α, is high ($\alpha = 0.81$) (Artstein & Poesio, 2008). The final score of each image-caption pair was obtained by averaging the three individual scores. Since this is the most time-consuming evaluation, we only judged the highest-ranked caption for each test image on the annotation task, and only focused on the subset of our models described above. To gauge the difficulty of this task on our data set, we also include a random baseline. Since we only evaluate a single caption for each image, we are interested in the percentage of images for which a suitable caption was returned. We therefore show each model’s cumulative distribution of test items with scores at or above thresholds ranging

Quality of first caption (image annotation)							
	Cumulative distribution of expert scores (% $\geq$ X)						
	= 4.0	**$\geq$ 3.66**	**$\geq$ 3.33**	**$\geq$3.0**	**$\geq$ 2.66**	**$\geq$ 2.33**	**$\geq$ 2.0**
Random	0.5^{***}	0.6^{***}	0.7^{***}	1.1^{***}	1.5^{***}	2.9^{***}	7.8^{***}
NN	3.4^{***}	4.1^{***}	5.2^{***}	8.5^{***}	11.4^{***}	16.3^{***}	27.1^{***}
BoW1	6.6^{***}	8.1^{***}	9.9^{***}	18.3^{***}	22.9^{**}	29.7^{***}	44.2^{**}
BoW5	9.7	11.8	13.6^{*}	19.8^{***}	24.7^{***}	33.0^{***}	46.9^{***}
TagRank	9.6	12.3	14.2	21.1	25.8^{*}	32.9^{**}	46.2^{***}
Tri5Sem	11.0	13.3	15.7	23.0	28.1	36.9	53.0

Table 1: Cumulative distribution of expert judgments on our 1–4 scale (Figure 4), indicating what percentage of image-caption pairs are judged to be at or above a given score. Scores are averaged over three judges. Superscripts indicate statistically significant difference to Tri5Sem ($^{*}: p \leq 0.1$, $^{**}: p \leq 0.05$, $^{***}: p \leq 0.01$).

from 4.0 to 2.0. Each threshold can be interpreted as a more or less strict mapping of the fine-grained scores into binary relevance judgments. In order to assess whether the difference between models at any given threshold reaches statistical significance, we use McNemar's significance test, a paired, non-parametric test that has been advocated for the evaluation of binary classifiers (Dietterich, 1998). Given the output of models A and B on the same set of items, McNemar's test considers only the items on which A and B's output differ (the 'discordant pairs' of output) to test the null hypothesis that both outputs are drawn from the same underlying population. Among these discordant pairs, it compares the proportion of items for which model A is successful but model B is not with the proportion of items for which Model B is successful but model A is not. In our results tables, $*$ superscripts indicate whether the difference between a model and Tri5Sem is statistically significant ($^{*}: p \leq 0.1$, $^{**}: p \leq 0.05$, $^{***}: p \leq 0.01$) .

'Expert' results (Table 1) We first interpret the 'expert' scores as binary relevance judgments, and therefore show their cumulative distribution for different thresholds in. We see very clear differences between the random baseline, NN, and the KCCA models at all thresholds. The differences between NN and the random model, as well as between any KCCA model and NN are highly significant ($p < 0.001$) at any threshold. While the random baseline returns a perfect caption for 0.5% of the images, and a good caption (assuming a threshold of $\geq$ 2.66) for 1.5% of the images, our best KCCA model, Tri5Sem, returns a perfect caption for 11.0% and a good caption for 28.1% of the images. However, the differences among the KCCA models are more subtle, and may only become apparent at lower thresholds. There is no significant difference between BoW5 and TagRank at any threshold, but they are both significantly better than BoW1 ($p < 0.001$) at thresholds of 3.33 and above. Tri5Sem outperforms all other models, but the differences to BoW5 and TagRank only reach statistical significance when the threshold of what is considered a suitable caption is lowered to either 3.33 ($p = 0.06$) or 3.0 ($p = 0.01$), or to 2.66 ($p = 0.08$) or to 2.33 ($p = 0.005$). This lack of statistical significance can be partially explained by the fact that McNemar's test has relatively low power when the percentage of items on which the two models are successful is very low, as is the case for the higher thresholds here.

We will show in Sections 4.3.1 and 4.3.2 below that there is a very significant difference between Tri5Sem and these two models on image annotation once we extend the analysis beyond the highest-ranked caption. This shows that evaluations that are only based on a single caption returned per image may fail to uncover significant differences between models that become apparent once multiple results are considered. It may also be important to consider performance on both annotation and retrieval. On the image retrieval task, we will see that Tri5Sem significantly outperforms all other models even when only the first result is considered. Table 1 also reveals another artefact of McNemar's test: since it is not based on absolute differences in performance but on the number of discordant pairs, the difference between BoW1 and Tri5Sem at thresholds 2.66 and 2.0 is considered less significant than that between BoW5 and Tri5Sem at the same thresholds, even though BoW1's scores are lower than BoW5's. In Table 2, we present the systems' average expert scores, and use Fisher's Randomization Test to determine statistical significance. According to this evaluation, Tri5Sem is very significantly better than all other models ($p \leq 0.0001$ in all cases), but since the average score of Tri5Sem is only 2.08, this difference is not reflected at the higher thresholds in the cumulative distribution shown in Table 1.

4.2.2 Automatic Evaluation with Bleu and Rouge

Since human judgments are expensive and time-consuming to collect, we now examine how well they can be approximated by Bleu (Papineni et al., 2002) and Rouge (Lin, 2004), two standard metrics for machine translation and summarization.

Bleu and Rouge scores Bleu and Rouge scores can be computed automatically from a number of reference captions, and have been used to evaluate a number of caption generation systems (Kulkarni et al., 2011; Ordonez et al., 2011; Li et al., 2011; Kuznetsova et al., 2012; Yang et al., 2011; Gupta et al., 2012), although it is unclear how well they correlate with human judgments on this task.

Given a caption $\mathbf{s}$ and an image $\mathbf{i}$ that is associated with a set of reference captions $R_{\mathbf{i}}$, the Bleu score of a proposed image-caption pair ($\mathbf{i}$, $\mathbf{s}$) is based on the n-gram precision of $\mathbf{s}$ against $R_{\mathbf{i}}$, while Rouge is based on the corresponding n-gram recall. As is common for image description, we only consider unigram-based scores (only 3.5% of all possible image-caption pairs in the test have a non-zero bigram-based Bleu-2 score, but 39.4% set have a non-zero Bleu-1 score). We also ignore Bleu's brevity penalty, since our data set has relatively little variation in sentence length, and we would like to avoid penalizing short, but generic captions that include few details but are otherwise correct. Hence, if $c_{\mathbf{s}}(w)$ is the number of times word w occurs in $\mathbf{s}$:

$$\begin{aligned}
\textsc{Bleu}(\mathbf{i},\mathbf{s}) &= \frac{\sum_{w\in\mathbf{s}} min(c_{\mathbf{s}}(w), max_{\mathbf{r}\in R_{\mathbf{i}}} c_{\mathbf{r}}(w))}{\sum_{w\in\mathbf{s}} c_{\mathbf{s}}(w)} \\
\textsc{Rouge}(\mathbf{i},\mathbf{s}) &= \frac{\sum_{\mathbf{r}\in R_{\mathbf{i}}} \sum_{w\in\mathbf{r}} min(c_{\mathbf{s}}(w), c_{\mathbf{r}}(w))}{\sum_{\mathbf{r}\in R_{\mathbf{i}}} \sum_{w\in\mathbf{r}} c_{\mathbf{r}}(w)}
\end{aligned} \qquad (28)$$

Both reference and candidate captions are preprocessed. We first tokenize the sentences with the OpenNLP[5] tools. Then we break up hyphenated words, stripping out non-alphanumeric

[5] http://opennlp.apache.org

Avg. score of first caption (image annotation)			
	Expert	**Bleu**	**Rouge**
Random	1.22^{***}	0.31^{***}	0.04^{***}
NN	1.57^{***}	0.35^{***}	0.11^{***}
BoW1	1.90^{***}	0.43^{***}	0.14^{***}
BoW5	1.98^{***}	0.46^{*}	0.15^{***}
TagRank	1.99^{***}	0.46^{**}	0.15^{***}
Tri5Sem	2.08	0.48	0.17

Table 2: Comparison of averaged scores according to the 4-point 'expert' evaluation (Figure 4), Bleu and Rouge, using all five test captions as reference. Superscripts indicate statistically significant difference to Tri5Sem ($^{*} : p \leq 0.1$, $^{**} : p \leq 0.05$, $^{**} : p \leq 0.01$).

and hyphen characters, and converting all words to lower case. Following the work of Lin (2004), we use a stemmer (Porter, 1980) and remove stopwords before compute Rouge scores. We compute the Bleu and Rouge score of a system as the average Bleu or Rouge scores of all items in the test set.[6]

We use Fisher's Randomization Test (Fisher, 1935; Smucker, Allan, & Carterette, 2007) to assess the statistical significance of the difference between models. This is a paired, sampling-based test that evaluates the null hypothesis that the results of models A and B are produced by the same underlying distribution. In each sample, the scores that A and B assign to each test item are randomly reassigned to the two models, and p-values are obtained by comparing the actual difference between A and B's performance to the fraction of samples with equal or greater difference between the models. We sample 100,000 reassignments of the entire test set.

Bleu and Rouge results (Table 2) Table 2 shows the average Bleu and Rouge scores of the highest ranked caption pairs returned by each image annotation systems, computed against a reference pool consisting of the five original captions for each test image (including the caption that was randomly selected to be part of the candidate pool). These scores lead to the same broad conclusions as the average expert scores: all metrics find very clear differences ($p < 0.0001$) between the random baseline and any of the other models, as well as between NN and any of the KCCA models, and none find any significant difference between BoW5 and TagRank. Tri5Sem outperforms the other KCCA models according to all metrics, but both the expert evaluation and Rouge find a much larger difference to BoW5 (Experts: $p \leq 0.0001$, Rouge: $p < 0.001$) than to TagRank (Experts: $p = 0.001$, Rouge: $p = 0.005$). Bleu only finds a significant difference to TagRank ($p < 0.05$), but not to BoW5 ($p < 0.05$), which indicates Bleu may be less well suited to identify more subtle differences between systems.

Agreement of Bleu and Rouge with 'expert' scores Since it is difficult to measure directly how well the Bleu and Rouge scores agree with the expert judgments, we consider

[6] A system's Bleu score is usually computed at the corpus level, but since we are only dealing with unigram scores and evaluate all systems on sentences from the same corpus, the averaged sentence-level Bleu scores of our systems we report are almost identical ($r > 0.997$) to their corpus-level Bleu scores.

a number of different relevance thresholds for each type of score (θ_B, θ_R, and θ_E), and turn them into binary relevance judgments. This allows us to use Cohen's (1960) κ to measure the agreement between the corresponding binarized scores. Since Bleu and Rouge both require a set of reference captions for each test image, we compare four different ways of defining the set of reference captions (for detailed scores, see Tables 8 and 9 in the appendix).

Since our data set contains multiple descriptions for each image, we first use all five captions as reference. In this setting, Bleu reaches the best agreement ($\kappa = 0.72$) against $\theta_E = 4.0$ with $\theta_B = 1.0$ or against $\theta_E \geq 3.6$ with $\theta_B \geq 0.8$. However, such high Bleu scores are generally only obtained when the system proposes the original caption. Rouge has much lower agreement ($\kappa = 0.54$) against the expert scores, obtained at $\theta_R \geq 0.4$ vs. $\theta_E \geq 4.0$ or $\theta_E \geq 3.6$, or $\theta_R \geq 0.3$ against $\theta_E \geq 3.0$. Since other data sets may have only one caption per image, we also evaluate against a reference corpus that consists only of the single caption in the test pool. In this case, both metrics reach again the highest agreement against an expert threshold of $\theta_E = 4.0$ (Bleu: $\kappa = 0.71$, Rouge: $\kappa = 0.69$), with thresholds of $\theta_B \geq 0.8$, and $\theta_R \geq 0.9$. We conclude that neither Bleu nor Rouge are useful in this scenario, since they require such high thresholds that they only capture how often the system returned the reference caption.

When Bleu and Rouge are used to evaluate caption generation systems, we cannot assume that the generated caption is identical to one of the reference captions. We therefore examine to what extent Bleu and Rouge scores agree with human judgments when the candidate pool contains human generated captions, but is disjoint from the reference captions. We first use a reference corpus of four captions per image, excluding the caption we use in the candidate pool. In this case, all three metrics show significantly lower agreement with human judgments than when the candidate pool contains the reference caption. Bleu reaches only $\kappa = 0.52$ (with $\theta_B \geq 0.7$ against $\theta_E \geq 3.3$) and Rouge reaches only $\kappa = 0.51$ (with $\theta_R \geq 0.2$ against $\theta_E \geq 2.6$). To simulate the case where only a single caption per image is available, we also evaluate against a reference corpus consisting of only one of these four captions. In this case, agreement with human judgments is even lower: Bleu reaches $\kappa = 0.36$, and Rouge reaches $\kappa = 0.42$. These results suggest that Bleu and Rouge are not appropriate metrics when the pool of candidate captions does not contain the reference captions, and lead us to question their usefulness for the evaluation of caption generation systems. This is consistent with the findings of Reiter and Belz (2009), who have studied Bleu and Rouge scores to evaluate natural language generation systems, and concluded that they may be useful as metrics of fluency, but are poor measures of content quality.

4.3 Metrics for the Large-Scale Evaluation of Image Description Systems

Metrics that only consider the first caption returned for each image cannot capture the fact that a better model should score good captions higher than most other captions, even if fails to consider them the best possible caption. Since our systems return a ranked list of results for each item, we now examine metrics that allow us to evaluate the quality of this list. In contrast to the human evaluations described in Section 4.2 above, we now also evaluate our image retrieval systems. We first consider metrics that can be computed automatically: recall and median rank of the item (image or sentence) that was originally associated with the query sentence or image (Section 4.3.1). We then show how to use crowdsourcing to

Performance: Rank of the original item
R@k: percentage of queries with the original item among top X responses.
Median r: median rank of the original item

	Image annotation				**Image retrieval**			
	R@1	**R@5**	**R@10**	**Median** r	**R@1**	**R@5**	**R@10**	**Median** r
NN	2.5^{***}	7.6^{***}	9.7^{***}	251.0^{***}	2.5^{***}	4.7^{***}	7.2^{***}	272.0^{***}
BoW1	4.8^{***}	13.5^{***}	19.7^{***}	64.0^{***}	4.5^{***}	14.3^{***}	20.8^{***}	67.0^{***}
BoW5	6.2^{**}	17.1^{***}	24.3^{***}	58.0^{***}	5.8^{**}	16.7^{***}	23.6^{***}	60.0^{***}
TagRank	6.0^{***}	17.0^{***}	23.8^{***}	56.0^{***}	5.4^{***}	17.4^{***}	24.3^{***}	52.5^{***}
Tri5	7.1	17.2^{***}	23.7^{***}	53.0^{***}	6.0^{***}	17.8^{***}	26.2^{***}	55.0^{***}
Tri5Sem	8.3	21.6	30.3	34.0	7.6	20.7	30.1	38.0

Table 3: Model performance as measured by the rank of the original image or caption (= 'correct response'). R@k: percentage of queries for which the correct response was among the first X results. Median r: Median position of the correct response in the ranked list of results. Superscripts indicate statistically significant difference to Tri5Sem ($^{**}: p \leq 0.05$, $^{***}: p \leq 0.01$).

collect a very large number of human judgments (Section 4.3.2), and use these relevance judgments to define two additional metrics: the 'rate of success', which is akin to recall, and R-precision, an established information retrieval metric (Section 4.3.3). Although these metrics allow us to evaluate all of our systems, we will focus our discussion on the small set of systems considered so far, and refer the interested reader to Section B of the appendix for the scores of all systems.

4.3.1 Recall and Median Rank of the Original Item

One advantage of our ranking framework is that the position of the original caption or image among the complete list of 1,000 test items can be determined automatically. Since a better system should, on average, assign a higher rank to the original items than a worse system, we can use their ranks to define a number of different evaluation metrics.

Recall (R@k) and median rank scores Since each query is only associated with a single gold result, we need not be concerned with precision. However, recall at position k (R@k), i.e. the percentage of test queries for which a model returns the original item among the top k results, is a useful indicator of performance, especially in the context of search, where a user may be satisfied if the first k results contain a single relevant item. We focus on $k = 1, 5, 10$ (R@1, R@5, R@10). Since this is a binary metric (for each query, the gold item is either found among the top k results or not), we use again McNemar's test to identify statistically significant differences between models. Conversely, the median rank indicates the k at which a system has a recall of 50% (i.e. the number of results one would have to consider in order to find the original item for half the queries). Here, we use Fisher's randomization to identify significant differences between models.

Recall (R@k) and median rank results (Table 3) The results in Table 3 confirm our earlier observation that the NN baseline is clearly beaten by all KCCA models ($p < 0.001$ for all metrics and models, except for R@1 search, where the difference to BoW1 has a

p-value of $p < 0.01$). Since the R@1 annotation scores are based on the same image-caption pairs as the expert scores in Table 1, we can compare them directly. The difference between the R@1 and expert scores, even at the strictest threshold of 4.0 for the experts, indicates that measures which capture how often the original caption was returned should be viewed as a lower bound on actual performance: while Tri5Sem returns the original caption first for 8.3% of the images, our human judges found that these captions describe 11.0% of the images without any errors. This discrepancy is even larger for BoW5 (6.2% vs. 9.7%) and TagRank (6.0% vs. 9.6%). As a consequence, the automatically computed R@1 scores indicate erroneously that there is a statistically significant difference between the quality of the first captions returned by Tri5Sem and those returned by BoW5 or TagRank, even though these differences are not significant according to the human evaluation. However, metrics that are only based on the first caption may fail to identify differences between models that become very apparent under all other metrics. For example, R@1 reveals no significant difference between Tri5 and Tri5Sem on the annotation task, although their difference is highly significant according to all other metrics. In Section 4.3.3, we present the results of a large-scale human evaluation which confirm that the actual differences between Tri5Sem and Tri5 on annotation can only be identified when more than the first caption is taken into account.

Table 11 in Section B provides recall and median rank scores for all models.

4.3.2 Collecting Binary Relevance Judgments on a Large Scale

In order to perform a human evaluation of a system that goes beyond measuring the quality of the highest ranked result, we would have to obtain relevance judgments for all image-caption pairs among the top k results for each query. Since we have two tasks, and a total of 30 different systems, this set consists of 113,006 distinct image-caption pairs for $k = 10$, rendering an exhaustive evaluation on the four-point scale described in Section 4.2.1 infeasible. We therefore needed to reduce the total number of judgments needed, and to define a simpler annotation task that could be completed in less time. Crowdsourcing platforms such as Amazon Mechanical Turk offer new possibilities for evaluation because they enable us to collect a large number of human judgments rapidly and inexpensively, and a number of researchers have evaluated caption generation systems on Mechanical Turk (Ordonez et al., 2011; Yang et al., 2011; Kuznetsova et al., 2012; Kulkarni et al., 2011; Li et al., 2011). But these experiments have not been performed at the scale of our analysis, and have also not evaluated how well crowdsourced judgments for this task approximate what can be obtained from a smaller pool of judges that can be given more detailed instructions. We examine here whether crowdsourcing allows us to collect reliable relevance judgments for a large scale evaluation of all of our image description systems.

The crowdsourcing task We presented workers with images that were paired with ten different captions, and asked them to indicate (via checkboxes) which of the captions describe the image. We adapted the guidelines developed for the fine-grained annotation such that a caption that describes the image with minor errors (corresponding to a score of 3 on our 4-point scale) would still be permitted to receive a positive score. These guidelines can also be found in the online appendix to this paper. Each individual task consisted of six different images, each paired with ten captions, and included a copy of the guidelines. We accessed

Amazon Mechanical Turk through a service provided by Crowdflower.com, which makes it easy to include control items for quality control. One of the six images in each task was such a control item, which we generated by taking random images from the development set, using between one and three of their original captions as correct responses, and adding another nine to seven randomly selected captions (which we verified manually that they did not describe the image) as incorrect responses. We only used workers who judged 70% of their control items correctly. Each image-caption pair was annotated by three different annotators (at a total cost of 0.9¢), and the final score of each image-caption pair was computed as the average number of positive judgments it received.

Filtering unlikely image-caption pairs In order to reduce the number of annotations needed, we devised a filter based on Bleu scores (Papineni et al., 2002) to filter out image-caption pairs whose caption is so dissimilar from the five captions originally written for the image that it is highly unlikely it describes the image. We found that a filter based on unigram Bleu-1 scores in combination with the stemming and stop word removal that is standardly done by Lin's (2004) Rouge script (Bleu_{pre}) proved particularly effective: a threshold of $\text{Bleu}_{\text{pre}} \geq 0.25$ filters out 86.0% of all possible (1,000×1,000) image-caption pairs in our test set, but eliminates only 6.7% of the pairs with an expert score of $2\frac{2}{3}$ or greater, and 3.5% of the pairs with an expert score of 3 or greater. A slightly higher cutoff of $\text{Bleu}_{\text{pre}} \geq 0.26$ would filter out 90.4% of all image caption pairs, but discard 12.3% of all image-caption pairs with an expert score of $\geq 2\frac{2}{3}$ and 7.5% of all image-caption pairs with an expert score of ≥ 3. Among the 113,006 image-caption pairs that we actually wished to obtain judgments for, the 0.25 filter eliminates 72.8%, reducing the number of pairs we needed to annotate to 30,781. Since our setup required us to pair each image with a number of captions that was a multiple of 10, we also annotated an additional 10,374 image caption pairs that had been filtered out, allowing us to evaluate the performance of our filter. For 98.3% of these filtered out pairs, all Mechanical Turk judges decided that the caption did not describe the image, and for 99.8% of them, the majority of annotators thought so. We also found that standard Bleu-1 without preprocessing is not a very effective filter: a threshold of Bleu ≥ 0.330 misses 6.9% of the good captions (with an expert score of $\geq 2\frac{2}{3}$), while only filtering out 55% of the entire data set, whereas a threshold of Bleu ≥ 0.333 filters out 65% of the entire data set, but misses 11.9% of the good captions.

Agreement of crowdsourced and expert judgments We again use Cohen's κ to measure the agreement between the crowdsourced and the expert judgments (Table 10 in the appendix). The best agreement is obtained between crowdsourced scores with a threshold of 0.66 or above (i.e. at least two of the three judges think the caption describes the image) and expert scores with a threshold of 3.33 (one expert thinks the caption describes the image perfectly and the other two agree or think it describes the image with only minor errors, or two experts think it describes the image perfectly and the other one thinks it is at least related). At $\kappa = 0.79$, this is a significantly better approximation to the expert scores than was possible with either Bleu or Rouge. We also examine the precision, recall and f-scores that these approximate relevance judgments achieve when compared against relevance judgments obtained from binarizing expert judgments (Table 10). 98.6% of all items with a perfect expert score (and 95.0% of all items with an almost perfect expert score of 3.7) are identified, and at least 94.7% of the items that pass this threshold have an expert score of 2.7 or greater (i.e. the majority of experts agreed that the caption describes the image perfectly or with minor errors). Using a threshold of 0.66 adds 2,031 suitable image-caption pairs to the 1,000 test images paired with their original caption. Among the 1,000 test captions, 446 still describe only a single image, 202 describe two test images, 100 three, and 252 describe four or more images. Among the 1,000 test images, 331 have only a single (i.e. the original) caption, 202 have two possible captions, 100 have three possible captions, and 317 have four or more captions.

4.3.3 Large-Scale Evaluation with Relevance Judgments

The crowdsourced relevance judgments allow us to define two new metrics, the 'rate of success' (S@k) and *R*-precision. We believe *R*-precision to be the more reliable indicator of overall performance, since it summarizes the human judgments in a single number that does not depend on an arbitrary cutoff. We therefore use it in Section 4.3.4 for an in-depth analysis of the impact of the different linguistic features our models incorporate. The S@k rate of success scores are motivated by the fact that search engines commonly return multiple results at once. Since users may be satisfied as long as these results contain at least one relevant item, S@k scores provide a more direct measure of utility for hypothetical users.

Rate of success (S@k) scores The 'rate of success' metric (S@k) is analogous to the recall-based R@k-scores used in Table 3, and is intended to measure the utility of our system for a hypothetical user. It indicates the percentage of test items for which at least one relevant result is found among the highest ranked k results. Following the analysis in Section 4.3.2, an image-caption pair is considered relevant if the majority of the judges say that the caption describes the image.

Rate of success results (Table 4) Table 4 confirms again that NN performs clearly worse than any of the KCCA models. The differences between Tri5Sem and the other models shown in Table 4 are highly statistically significant ($p < 0.001$) for all metrics except for the S@1 annotation scores, where, in agreement with the expert scores from Table 1, only the differences to NN and BoW1 are significant. It is unclear why the quality of the first caption that Tri5Sem returns for annotation is not significantly better than those returned by the other models, since it outperforms them on all other metrics. The S@k scores in Table 4 indicate that Tri5Sem returns a relevant caption among the top 10 responses for 49.1% of the images, and a relevant image for 48.5% of the captions. A comparison with the expert scores in Table 1 shows that all S@1 annotation scores lie between expert scores with a threshold of 3.66 and 3.0, while a comparison with the R@k results in Table 3 shows that the S@1 scores are at least twice as high as the corresponding R@1 scores. That is, the highest ranked response is just as often a relevant item that was not originally associated with the query as it is the original gold item itself.

***R*-precision scores** Given the crowdsourced relevance judgments, each test image may now be associated with multiple relevant captions, and each test caption may have been deemed relevant for multiple images besides the one it was originally written for. When queries have a variable number of relevant answers, the performance of retrieval systems is commonly measured in terms of *R*-precision (Manning, Raghavan, & Schütze, 2008). Unlike the S@k scores, this metric does not depend on an arbitrary cutoff, but summarizes the

Rate of success (S@k)
(Percentage of items with relevant response among top X results)

	Image annotation			**Image retrieval**		
	S@1	**S@5**	**S@10**	**S@1**	**S@5**	**S@10**
NN	5.8^{***}	15.4^{***}	20.2^{***}	5.0^{***}	13.3^{***}	18.4^{***}
BoW1	12.2^{***}	30.3^{***}	39.7^{***}	11.4^{***}	30.5^{***}	40.2^{***}
BoW5	15.0	34.1^{**}	42.7^{***}	12.1^{***}	31.5^{***}	40.8^{***}
TagRank	16.2	34.2^{**}	42.9^{***}	12.4^{***}	31.5^{***}	41.6^{***}
Tri5	16.4	32.9^{***}	43.4^{***}	13.1^{**}	33.1^{**}	43.8^{***}
Tri5Sem	16.6	37.7	49.1	15.7	36.9	48.5

Table 4: The rate of success (S@k) indicates the percentage of test items for which the top X results contain at least one relevant response. Superscripts indicate statistically significant difference to Tri5Sem ($^*: p \leq 0.1$, $^{**}: p \leq 0.05$, $^{***}: p \leq 0.01$)

***R*-precision**

	Annotation	**Search**	**Total**
NN	5.2^{***}	3.8^{***}	4.5
BoW1	10.7^{***}	9.6^{***}	10.1
BoW5	11.1^{***}	10.5^{***}	10.8
TagRank	11.7^{***}	10.5^{***}	11.1
Tri5	11.6^{***}	11.0^{***}	11.3
Tri5Sem	13.7	13.4	13.5

Table 5: Model performance as measured by R-precision, with statistically significant differences to Tri5Sem ($^*: p \leq 0.1$, $^{**}: p \leq 0.05$, $^{***}: p \leq 0.1$)

performance of each system in a single number, allowing us to rank models according to their overall performance (see Section 4.3.4 below). And while the S@k scores measure only whether at least one of the relevant items is ranked highly, R-precision requires all relevant items to be ranked highly. It is therefore a better indicator of the quality of the mapping between images and sentences, since a better mapping should prefer all relevant captions or images over any irrelevant caption or image.

The R-precision of system s on a query q_i with r_i known relevant items in the test data is defined as its precision at rank r_i (i.e. the percentage of relevant items among the top r_i responses returned by s). The R-precision of s is obtained by averaging over all test queries. We again use Fisher's randomization test to assess whether the differences between models reaches statistical significance.

***R*-precision results (Table 5)** Table 5 gives the R-precision of the model types that were used when collecting expert judgments (Section 4.2.1). We see that the nearest neighbor baseline is again very clearly below all KCCA models ($p < 0.001$). R-precision indicates that there is little difference between BoW1, BoW5 and TagRank in terms of their overall performance. Although TagRank and Tri5 outperform BoW1 slightly on search ($p = 0.062$), the only statistically significant difference among these three models is that between BoW1 and Tri5 on search ($p = 0.01$). In contrast to the human evaluation that considered only

R-precision								
	Tri5		**+IDF**		**+Align**		**+Align&IDF**	
	Ann.	**Search**	**Ann.**	**Search**	**Ann.**	**Search**	**Ann.**	**Search**
Tri5	**11.6**	**11.0**	12.5^{ii}	11.3	13.4^{aaa}	12.3^{aa}	13.4^{a}	$13.2^{aaa,ii}$
$+D_{\mathbf{BNC}}$	12.7^{dd}	12.1^{dd}	12.9	12.2^{ddd}	13.2	12.8^{a}	12.9	12.9^{a}
$+D_{\mathbf{ic}}$	12.7^{dd}	12.8^{ddd}	12.8	13.1^{ddd}	13.0	12.8	13.3	13.4
$+D_{\mathbf{BNC+ic}}$	12.5^{dd}	12.7^{ddd}	13.3^{d}	13.0^{ddd}	13.4^{aa}	13.2^{dd}	**13.7**	**13.4**

Table 6: The effect of adding IDF weighting (i), alignment-based similarities (a) and distributional similarities (d) to the Tri5 model. The bolded scores indicate Tri5 (top left) and Tri5Sem (Tri5 +Align&IDF+$D_{\text{BNC+ic}}$; bottom right). Superscripts indicate statistically significant differences that result from the addition of the corresponding feature ($x : p \leq 0.1$, $xx : \leq 0.05$, $xxx : p \leq 0.01$). D_c = distributional similarities computed over corpus c (the BNC, our training corpus of image captions ('ic'), or both)

the first result (Table 1), Tri5Sem clearly outperforms all other models on both annotation and retrieval (for all differences $p \leq 0.0001$). Table 12 in Appendix B shows scores for all models.

4.3.4 Measuring the Impact of Linguistic Features (Table 6)

The results presented so far indicate clearly that Tri5Sem outperforms the simpler Tri5 model, but have not considered the impact of the individual text features that distinguish the two models. Since R-precision summarizes the performance of each system in a single number, it allows us to easily perform this analysis.

Using R-precision for model comparison Table 6 shows the results of an ablation study which compares the R-precision of Tri5 and Tri5Sem with that of other trigram-based KCCA models that use a subset of Tri5Sem's additional features. The basic Tri5 model yields the bolded scores shown in the top left corner. Tri5Sem's scores are given in the bottom right corner. The top row contains models that do not capture any distributional similarities, while each of the bottom three rows corresponds to the addition of one kind of distributional similarity (computed on the BNC, on the image captions in our training corpus, or on both corpora) to the corresponding model in the top column. The first column contains models that do not capture any IDF reweighting or alignment-based similarities. The second column corresponds to the addition of IDF reweighting to models in the first column, while the third column adds alignment-based similarities to the models in the first column. The last column adds both IDF-reweighting and alignment-based similarities, and these scores should be compared to both the second and third column. Superscripts indicate that the addition of a particular feature leads to a statistically significant improvement over the model that does not include this feature but is otherwise identical. That is, d superscripts show that the addition of a distributional similarity metric leads to a significant improvement over the model in the top cell of the same column. The i superscripts indicate that the addition of IDF reweighting leads to a significant improvement over the corresponding model without IDF reweighting in the immediately preceding cell in the same row. The *a* superscripts in the third column show that the addition of the alignment-based similarity leads to a significant improvement over the model without IDF reweighting shown in the first column of the same row, and *a* superscripts in the fifth column show that the addition of the alignment-based similarity to the model with IDF reweighting shown in the second column of the same row leads to a significant improvement.

The impact of IDF weighting, distributional and alignment-based similarities While IDF weighting is almost always beneficial, the improvements obtained by adding IDF weighting to a given text kernel reach statistical significance (indicated by *i* superscripts in Table 6) in only two cases: the performance of the basic Tri5 model on image annotation, and the performance of the alignment-based Tri5 model on image search. By contrast, adding lexical similarities leads almost always to a significant or highly significant improvement. Distributional similarities (*d* superscripts) are very beneficial for the basic Tri5 model on both tasks, and help the IDF weighted Tri5 model on image search. Distributional similarities computed on both corpora also significantly improve the performance of the alignment-based Tri5 model that does not incorporate IDF weighting. Adding them to the alignment-based Tri5 model without IDF weighting leads to further improvement on search (while not helping or slightly decreasing performance on annotation, albeit not significantly so). The improvements on search only reach statistical significance when the similarities computed over both corpora are added. Conversely, adding alignment-based similarities to the non-IDF weighted Tri5 model with distributional similarities from both corpora leads to a significant improvement on annotation. Finally, the top cell of the last column shows that adding alignment-based similarities to the IDF-weighted Tri5 model leads to a significant improvement on both tasks, although the impact on search is even greater. Comparing this model's performance to the alignment-based Tri5 model without IDF weighting shows that in this case, IDF weighting only helps on search. The bottom cells of this column show that adding alignment-based similarities to models that already use IDF weighting and distributional similarities, or adding IDF weighting to models with distributional and alignment-based similarities generally lead to minor improvements.

Table 6 shows only whether the difference in performance obtained by the addition of one kind of feature reaches statistical significance, but it is worth noting that any model that captures lexical similarities of any kind is significantly better than the basic Tri5 model on both tasks ($p \leq 0.02$ search; $p < 0.0001$ annotation), while IDF-reweighting by itself only leads to a significant improvement on the annotation task ($p < 0.03$). Moreover, the difference between Tri5Sem (13.7 search; 13.4 annotation) and the basic Tri5 kernel with IDF-reweighting (12.5 search; 11.3 annotation) are highly significant ($p < 0.03$ search; $p < 0.0001$ annotation).

The impact of Lin's similarity Not shown in Table 6 is the performance of Tri5_{Lin}, the model which augments the trigram kernel with Lin's (1998) WordNet-based similarity. Tri5Sem does not include Lin's similarity, since we found during development that Tri5_{Lin} performed similarly to or worse than the basic Tri5 model on the automatic R@k and median rank scores. This is also reflected in Tri5_{Lin}'s *R*-precision scores of 11.7 for annotation (Tri5: 11.6) and 10.7 for search (Tri5: 11.0). Lin's similarity may simply be too coarse for our purposes. As shown in Table 3, the hypernym relations in WordNet lead it to associate terms such as swimming and football with each other. But even though these are semantically

Correlation of system rankings between S@k and R@k				
	Annotation		**Search**	
	ρ	τ	ρ	τ
@1	0.86	0.69	0.97	0.87
@5	0.92	0.76	0.97	0.88
@10	0.96	0.82	0.97	0.87

(a) S@k vs. R@k

Correlation of system rankings between *R*-precision				
	Annotation		**Search**	
	ρ	τ	ρ	τ
R@1	0.85	0.68	0.94	0.83
R@5	0.93	0.78	0.96	0.87
R@10	0.94	0.79	0.97	0.89
Median rank	-0.92	-0.79	-0.97	-0.89

(b) *R*-precision vs. R@k and median rank

Table 7: Correlation (Spearman's ρ and Kendall's τ) of system rankings obtained from human metrics (S@k and *R*-precision) and automated scores (R@k and median rank)

related by the fact that they are both different kinds of sports or activities, they are visually very dissimilar, and should not be considered related by our systems.

4.3.5 Can Human Evaluations Be Approximated by Automatic Techniques?

R-precision and the S@k scores require human judgements, and therefore cannot be applied to datasets where these judgements have not yet been collected or whose scale may prohibit ever creating a definitive set of judgements. However, if the evaluation is intended to measure relative progress on image description rather than absolute performance, our automatic metrics may be a sufficient approximation, since they yield a similar ranking of systems to *R*-precision and the S@k scores. Table 7(a) shows the correlations between the rankings of all of our NN and KCCA systems ($n = 30$) obtained from the S@k scores and those obtained from the corresponding R@k scores. Table 7(b) shows the correlations between R-precision and our automatic metrics. We report two rank correlation coefficients, Spearman's ρ and Kendall's τ. We first observe that system rankings obtained via R@1 do not correlate highly with either *R*-precision or S@1 based rankings. On the other hand, we also observe that R@5, R@10, and the median rank scores correlate well with *R*-precision and that R@5 and R@10 correlate well with their corresponding S@k metrics. This suggests that ranking-based metrics are significantly more robust than metrics that consider only the quality of the first result. Moreover, these results indicate that our framework, in which systems are expected to rank a pool of images or sentences written by people, may enable a large-scale, fully automated, evaluation of image description systems that does not require an equally large-scale effort to collect human judgments.

5. Summary of Contributions and Conclusions

In this paper, we have proposed to frame image description as the task of selecting or ranking descriptions among a large pool of descriptions provided by people, because this framework provides a direct test of the purely semantic aspects of image description and does not need to be concerned with the difficulties involved in the automatic generation of syntactically correct and pragmatically appropriate sentences. We have also introduced a new data set of images paired with multiple captions, and have used this data set to evaluate a number of nearest-neighbor and KCCA-based models on sentence-based image annotation as well as on the converse task of sentence-based image search. Our experiments indicate the importance of capturing lexical similarities. Finally, we have performed an in-depth analysis of different evaluation metrics for image description.

5.1 The Advantages of Framing Image Description as a Ranking Task

One of our main motivations for framing image description as a ranking rather than a generation problem was the question of an objective, comparable evaluation of our ability to understand what is depicted in images. In order to make progress on this challenging task, it is important to define tasks and evaluation metrics that allow for an objective comparison of different approaches. We have argued that the task of ranking a pool of captions written by people is attractive for a number of reasons: first, results obtained on the same data set can be compared directly; second, human evaluation is easier than for generated captions since it needs to only focus on factual correctness of the description rather than grammaticality, fluency, or creativity; third, statistically significant differences between systems may not become apparent when only a single caption per image is considered; and finally, ranking makes it possible to automate evaluation, e.g. by considering the position of the original caption. Moreover, framing image description as a ranking task also establishes clear parallels to image retrieval, allowing the same metrics to be used for both tasks.

5.2 Our Data Set

Our Flickr 8k data set of over 8,000 images, each paired with five crowdsourced captions, is a unique resource for image description. Although it is much smaller than the SBU corpus (Ordonez et al., 2011), we believe that the generic conceptual descriptions in our corpus are more useful for image understanding than the original Flickr captions in the SBU data set. The data set that is perhaps most similar to ours is the IAPR data set (Grubinger et al., 2006), but the captions in our corpus are shorter, and focus on the salient aspects of the image. And while we focus on images of people and animals, the IAPR data set covers a slightly different domain, including city pictures and landscape shots which typically do not depict or focus on people. A distinct advantage of our corpus is that it pairs each image with multiple, independently written captions. Our results indicate that using more than a single caption at training time leads to a significant increase in performance. We have also shown how to use these multiple captions to define an alignment-based lexical similarity that may be more useful for image description than standard distributional or WordNet-based similarities.

5.3 Our Models

This paper is the first to apply Kernel Canonical Correlation analysis (KCCA) to sentence-based image description. Our results show that KCCA significantly outperforms nearest neighbor-based approaches on our data set of 6,000 training images and 1,000 test images (although these may scale up better to very large data sets such as Ordonez et al.'s (2011) SBU corpus, where the memory requirements to train KCCA may be prohibitive). One advantage of KCCA-based approaches over other image description systems that are geared specifically towards caption generation is that they can not only be applied to image description, but also to image retrieval, and our results indicate that performance on both tasks is fairly similar.

An important difference between our approach taken in this paper and most other image description systems is that all the features used by the models presented here can be computed with minimal supervision. The only feature that relies on a supervised classifier is the alignment-based similarity, which uses a POS-tagger to identify nouns and verbs. Despite the simplicity of the underlying features, our models achieve relatively high performance, considering the difficulty of the task: Although there is only a 1.5% chance that a randomly chosen test caption will describe a test image well, fine-grained human judgments reveal that for image annotation, the first caption returned by our best KCCA system is a good description for 28% of the test images. Furthermore, our large-scale evaluation shows that with our best system, there is an almost 50% chance that a suitable image or caption will be returned among the first ten results. Our results indicate that there are two main reasons for this high performance: the availability of multiple captions for each image at training time, and the use of robust text representations that capture lexical similarities rather than requiring strict equality between words. However, it is also clear that this task remains far from being solved, and we leave the question of how KCCA may benefit from models that rely on richer visual or linguistic features such as detector responses or rich syntactic analyses for future work.

5.4 Evaluating Ranking-Based Image Description Systems

The main advantage of framing image description as a ranking problem is that it allows for a direct comparison of different approaches, since they can be evaluated on the same data set. It also makes it possible to borrow established evaluation metrics from information retrieval, and to use the same metrics and data sets for sentence-based image annotation and image search.

On the one hand, we have shown that crowdsourcing can be used to collect a large number of binary judgments of image-caption pairs for a relatively low price, and that these crowdsourced judgments correlate well with more fine-grained judgments. Being able to collect human judgments on a large scale is particularly important for retrieval-based approaches to image description, since the number of relevance judgments that need to be collected for a test collection may be significantly larger than the number of judgments commonly used to evaluate a single caption generation system. However, our experiments on image annotation have provided an example where human judgments of a first caption returned for each test image did not reveal differences between systems that become apparent when more results are taken into account. Our fine-grained evaluation also indicates that evaluations that are based on a single result may require a potentially much larger number of test items in order to reveal robust statistically significant differences. Among the human evaluation metrics we have compared, we believe that R-precision computed over the crowdsourced relevance judgments is the most robust. R-precision is a standard metric for evaluating ranked retrieval results when items have a varying number of relevant responses, and since it yields a single score, it also makes it particularly easy to compare systems. However, the S@k scores, which measure the percentage of items for which the top k responses contain a relevant result, are perhaps a more direct measure of how useful a system may be in practice. We will release the crowdsourced relevance judgments we have collected in order to enable others to evaluate their image description system on our data. We hope that this will establish a benchmark that can be used for a direct and fair large-scale comparison between an arbitrary number of image description systems.

On the other hand, we have also shown that our framework in which systems are evaluated on their ability to rank a pool of images or sentences may make it possible to perform a fully automated evaluation. Contrary to current practice, our analysis indicates clearly that standard metrics such as Bleu or Rouge are not very reliable indicators for how well captions describe images, even if Bleu with Rouge-style preprocessing can be used as an effective filter of implausible image-caption pairs. Although we only consider human-generated captions, we stipulate that similar observations may hold for automatically generated captions, since similar criticisms about Bleu's appropriateness for generation and machine translation evaluation are well known (Reiter & Belz, 2009; Callison-Burch, Osborne, & Koehn, 2006). However, in a ranking-based framework each test query is associated with a 'gold' response that it was originally associated with, and our results indicate that metrics based on this rank of the 'gold' item lead to very similar conclusions as human judgments. This suggest that the evaluation of the ranking-based image description task can be automated, and performed on a potentially much larger scale than we have examined here.

5.5 Implications for the Evaluation of Caption Generation Systems

Image description can, and should, also be treated as a problem for the natural language generation community. But automatically generating captions that are indistinguishable from captions written by people (an evaluation criterion used by Mitchell et al. (2012) for their comparison of caption generation systems) requires much more than the ability to provide factually correct information about the image. We believe that the linguistic issues that need to be solved in a generation setting need to be evaluated separately from ability to decide whether a given caption describes an image. It is unclear that the kinds of evaluations performed by e.g. Mitchell et al. could ever be automated, since the question of how natural an automatically produced caption seems may always require human judgment. But human experiments are expensive, and since each system generates its own captions, such judgments have to be collected anew for each system and experiment. Since there is no consensus on what constitutes a good image description, independently obtained human assessments of different caption generation systems should not be compared directly. This means that a direct comparison of systems, e.g. as performed by Mitchell et al., is typically only possible within one research group, since there is no common data set for which different system outputs are publicly available. Although automatic scores such as Bleu and Rouge may still be useful for caption generation as measures of fluency (Reiter & Belz, 2009), we have shown that they are not reliable metrics for how well a caption describes an image, especially when the candidate pool is disjoint from the reference captions. This suggests that the evaluation of the syntactic and pragmatic aspects of the caption generation task should not be automated, and may have to rely on human judgments. However, it may be possible to use the framework proposed in this paper to evaluate the semantic affinity functions $f(\mathbf{i}, \mathbf{s})$ that are implicitly used in caption generations systems.

Acknowledgments

We gratefully acknowledge support for this project from the National Science Foundation through IIS Medium grant 0803603, CAREER award 1053856, and CNS-1205627 CI-P.

Appendix A. Agreement Between Approximate Metrics and Expert Human Judgments

Tables 8 and 9 use Cohen's Kappa (κ) to measure the agreement between Bleu and Rouge scores and expert judgments. We have selected a few thresholds that yield optimal results. Table 10 (a) shows the agreement between the crowdsourced judgments and the expert judgments. Since the best agreement to the expert scores is obtained with the crowdsourced judgments using a threshold of 0.6, Table 10 (b) measures precision and recall of the resulting binary relevance judgments against binarized expert judgments obtained with varying thresholds.

Appendix B. Performance of All Systems

The following tables give results for all models. In Section 4 of the body of the paper, NN corresponds to $\text{NN5}^{\text{idf}}_{\text{F1}}$, while Tri5Sem corresponds to $\text{Tri5}^{\sqrt{\text{idf}}}_{\text{A},D_{\text{BNC+ic}}}$.

R@k and median rank scores Table 11 gives the recall and median rank of the original item (Section 4.3.1) for all of our models.

Agreement between expert and Bleu/Rouge scores (Cohen's κ) **Case 1: $S_{\text{cand}} \subseteq S_{\text{ref}}$**						
5 reference captions/test image ($S_{\text{cand}} \subset S_{\text{ref}}$; $\mathbf{R_5}$)						
Expert	**Bleu θ_B**			**Rouge θ_R**		
$\theta_{\mathbf{E}}$	**≥0.9**	**≥0.8**	**≥0.7**	**≥0.4**	**≥0.3**	**≥0.2**
=4.0	**0.72**	0.70	0.59	**0.54**	0.47	0.29
≥3.6	0.71	**0.72**	0.61	**0.54**	0.50	0.33
≥3.3	0.64	**0.67**	0.63	**0.50**	**0.50**	0.37
≥3.0	0.45	0.54	**0.57**	0.45	**0.54**	0.49
≥2.6	0.35	0.45	**0.51**	0.38	0.51	**0.53**
1 reference caption/test image ($S_{\text{cand}} = S_{\text{ref}}$; $\mathbf{R_1}$(gold))						
Expert	**Bleu**			**Rouge**		
$\theta_{\mathbf{E}}$	**≥ 0.8**	**≥ 0.6**	**≥ 0.5**	**≥ 0.9**	**≥ 0.7**	**≥ 0.3**
=4.0	**0.71**	0.70	0.52	**0.69**	0.67	0.35
≥3.6	**0.68**	**0.68**	0.56	**0.65**	0.64	0.39
≥3.3	**0.60**	0.59	0.56	**0.57**	0.56	0.40
≥3.0	0.41	0.42	**0.48**	0.39	0.40	**0.45**
≥2.6	0.32	0.32	**0.42**	0.30	0.32	**0.43**

Table 8: Agreement (Cohen's κ) between binarized expert and Bleu/Rouge scores when the pool of candidate captions contains each test image's reference caption(s).

Agreement between expert and Bleu/Rouge scores (Cohen's κ)
Case 2: $S_{\mathbf{cand}} \not\subseteq S_{\mathbf{ref}}$

4 reference captions/ test image ($\mathbf{R_4}$)

Expert	**Bleu**			**Rouge**		
$\theta_{\mathbf{E}}$	≥ 0.7	≥ 0.6	≥ 0.5	≥ 0.4	≥ 0.3	≥ 0.2
=4.0	**0.50**	0.40	0.23	**0.44**	0.40	0.26
≥3.6	**0.51**	0.43	0.28	**0.43**	**0.43**	0.30
≥3.3	**0.52**	0.46	0.32	0.40	**0.44**	0.34
≥3.0	0.47	**0.48**	0.41	0.39	**0.50**	0.47
≥2.6	0.41	**0.47**	0.44	0.33	0.48	**0.51**

1 reference caption/test image ($\mathbf{R_1}$(other))

Expert	**Bleu**			**Rouge**		
$\theta_{\mathbf{E}}$	**≥ 0.5**	**≥ 0.4**	**≥ 0.3**	**≥ 0.4**	**≥ 0.3**	**≥ 0.2**
=4.0	**0.33**	0.27	0.16	**0.33**	0.30	0.18
≥3.6	**0.34**	0.29	0.19	**0.34**	0.32	0.21
≥3.3	**0.34**	0.32	0.22	**0.35**	**0.35**	0.24
≥3.0	0.32	**0.36**	0.29	0.39	**0.42**	0.34
≥2.6	0.30	**0.35**	0.31	0.37	**0.41**	0.38

Table 9: Agreement (Cohen's κ) between binarized expert and Bleu/Rouge scores when the pool of candidate captions may not contain each test image's reference caption(s).

Agreement between expert and lay scores (Cohen's κ)

Expert		**Lay θ_L**	
$\theta_{\mathbf{E}}$	**=1.0**	**≥ 0.6**	**≥ 0.3**
=4.0	0.75	0.69	0.49
≥3.6	0.78	0.76	0.57
≥3.3	0.74	**0.79**	0.65
≥3.0	0.56	0.71	0.74
≥2.6	0.45	0.62	0.73

(a) Agreement (Cohen's κ) between relevance judgments obtained from expert scores (relevance = score $\geq \theta_E$) and lay scores (relevance = score $\geq \theta_L$)

Lay vs. expert relevance judgments ($\theta_L = 0.66$)

θ_E	**Precision**	**Recall**	**F1**
=4.0	55.9	98.6	71.4
≥3.6	65.4	95.0	77.5
≥3.3	75.2	88.0	81.1
≥3.0	90.0	64.7	75.3
≥2.6	94.7	53.4	68.3
≥2.3	98.2	40.1	57.0

(b) Precision, recall, and F1 scores of binarized lay scores ($\theta_L = 0.66$) against binarized 'expert' scores with varying thresholds θ_E.

Table 10: Comparing the relevance judgments obtained from the lay scores against those obtained from expert scores

S@k and *R*-precision scores Table 12 gives the S@k success rate (Section 4.3.3) and *R*-precision scores (Section 4.3.3) for all of our models, based on the crowdsourced human judgments (Section 4.3.2).

Performance of all models (automatic evaluation)
(R@k: percentage of queries with original item in top X results
Median r: median rank of original item)

	Image annotation				Image search			
	R@1	**R@5**	**R@10**	**Median r**	**R@1**	**R@5**	**R@10**	**Median r**
NN5$_{\mathbf{F1}}$	1.9	5.9	8.7	251.0	2.1	5.2	7.1	278.0
NN5$^{\mathbf{idf}}_{\mathbf{F1}}$	2.5	7.6	9.7	251.0	2.5	4.7	7.2	272.0
NN5$_{\text{BoW5}}$	2.1	5.9	9.6	258.5	2.8	6.4	9.1	266.0
NN5$_{\text{Tri}\mathbf{(best)}}$	2.1	5.9	9.4	248.0	2.3	6.1	9.0	240.0
BoW1	4.8	13.5	19.7	64.0	4.5	14.3	20.8	67.0
Tri1	4.6	14.4	21.0	68.0	4.5	14.0	22.5	71.0
BoW5$^{\mathbf{Histo}}$	5.9	14.9	21.2	69.0	4.8	14.2	20.8	74.0
BoW5	6.2	17.1	24.3	58.0	5.8	16.7	23.6	60.0
BoW5$^{\mathbf{idf}}$	6.1	17.0	23.2	60.5	6.4	16.5	24.5	59.5
BoW5$^{\sqrt{\mathbf{idf}}}$	6.1	17.3	23.9	56.0	6.1	16.9	24.5	60.5
TagRank	6.0	17.0	23.8	56.0	5.4	17.4	24.3	52.5
Tri5$^{\mathbf{Histo}}$	6.0	15.0	21.7	63.5	5.7	14.5	22.1	67.0
Tri5	7.1	17.2	23.7	53.0	6.0	17.8	26.2	55.0
Tri5$_{\mathbf{Lin}}$	6.2	16.7	23.7	53.5	6.0	16.7	24.4	61.0
Tri5$_{D_{\mathbf{BNC}}}$	7.5	19.8	26.1	40.0	7.2	18.4	27.4	44.5
Tri5$_{D_{\mathbf{ic}}}$	7.0	19.5	27.1	36.0	7.0	19.3	27.5	41.0
Tri5$_{D_{\mathbf{BNC+ic}}}$	7.3	20.0	27.0	36.0	6.9	19.2	28.0	42.0
Tri5$_{\mathbf{A}}$	7.2	20.2	28.0	41.0	6.8	18.5	27.7	41.5
Tri5$_{\mathbf{A},D_{\mathbf{BNC}}}$	7.9	20.3	28.4	39.0	7.8	19.0	27.4	39.0
Tri5$_{\mathbf{A},D_{\mathbf{ic}}}$	6.9	20.2	29.5	35.0	7.3	19.9	28.7	39.5
Tri5$_{\mathbf{A},D_{\mathbf{BNC+ic}}}$	7.6	20.7	30.0	35.0	7.4	19.4	29.2	38.0
Tri5$^{\sqrt{\mathbf{idf}}}$	7.6	18.8	25.1	46.0	6.2	18.0	26.5	52.0
Tri5$^{\sqrt{\mathbf{idf}}}_{D_{\mathbf{BNC}}}$	6.8	18.7	28.9	40.0	6.7	18.2	27.6	45.0
Tri5$^{\sqrt{\mathbf{idf}}}_{D_{\mathbf{ic}}}$	7.3	20.4	27.5	38.0	8.1	19.1	28.4	40.5
Tri5$^{\sqrt{\mathbf{idf}}}_{D_{\mathbf{BNC+ic}}}$	6.7	20.0	28.9	35.0	7.0	19.7	28.8	39.0
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A}}$	7.3	21.1	28.3	37.0	7.5	19.2	28.7	38.0
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A},D_{\mathbf{BNC}}}$	7.5	21.3	30.0	38.0	7.8	18.9	29.0	41.5
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A},D_{\mathbf{ic}}}$	7.2	20.8	29.6	34.0	7.4	21.4	30.1	37.5
Tri5$^{\sqrt{\mathbf{idf}},\mathbf{Histo}}_{\mathbf{A},D_{\mathbf{BNC+ic}}}$	6.5	18.0	26.7	45.0	6.0	18.0	24.2	48.5
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A},D_{\mathbf{BNC+ic}}}$	8.3	21.6	30.3	34.0	7.6	20.7	30.1	38.0

Table 11: Performance of all models, measured as the percentage of test items for which the original item was returned among the top 1, 5 or 10 results, as well as as the median rank of the original item. In Section 4, NN5$^{\text{idf}}_{\text{F1}}$ = NN, Tri5$^{\sqrt{\text{idf}}}_{\text{A},D_{\text{BNC+ic}}}$ = Tri5Sem.

Performance of all models (human evaluation)
S@k: Percentage of items with relevant response among top X results
R-prec: R-precision computed over relevant responses

	Image annotation				Image search			
	S@1	**S@5**	**S@10**	**R-prec.**	**S@1**	**S@5**	**S@10**	**R-prec.**
NN5$_{\mathbf{F1}}$	4.9	13.3	19.1	4.2	4.9	13.2	17.8	3.8
NN5$_{\mathbf{F1}}^{\mathbf{idf}}$	5.8	15.4	20.2	5.2	5.0	13.3	18.4	3.8
NN5$_{\text{BoW5}}$	6.4	14.8	20.6	5.4	5.7	13.4	18.4	4.6
NN5$_{\text{Tri}\mathbf{(best)}}$	7.2	17.4	23.1	6.2	4.4	13.5	19.8	4.3
BoW1	12.2	30.3	39.7	10.7	11.4	30.5	40.2	9.6
Tri1	12.8	32.2	40.2	10.5	12.2	30.6	41.5	9.9
BoW5$^{\mathbf{Histo}}$	13.9	29.8	39.6	9.9	11.5	28.0	38.1	9.3
BoW5	15.0	34.1	42.7	11.1	12.1	31.5	40.8	10.5
BoW5$^{\mathbf{idf}}$	13.9	32.7	42.4	11.0	13.3	30.8	41.8	10.6
BoW5$^{\sqrt{\mathbf{idf}}}$	15.0	34.0	42.7	11.3	12.9	31.3	41.0	10.7
TagRank	16.2	34.2	42.9	11.7	12.4	31.5	41.6	10.5
Tri5$^{\mathbf{Histo}}$	15.0	29.0	38.9	9.9	12.9	28.9	39.9	10.5
Tri5	16.4	32.9	43.4	11.6	13.1	33.1	43.8	11.0
Tri5$_{\mathbf{Lin}}$	15.5	34.1	43.8	11.7	12.7	32.5	41.7	10.7
Tri5$_{D_{\mathbf{BNC}}}$	16.8	37.4	45.5	12.7	14.5	35.3	44.9	12.1
Tri5$_{D_{\mathbf{ic}}}$	15.8	36.7	47.0	12.7	14.5	36.6	46.1	12.8
Tri5$_{D_{\mathbf{BNC+ic}}}$	16.4	37.2	47.1	12.5	14.8	36.3	45.8	12.7
Tri5$_{\mathbf{A}}$	17.3	36.9	47.4	13.4	14.3	35.4	46.6	12.3
Tri5$_{\mathbf{A},D_{\mathbf{BNC}}}$	16.6	36.5	47.4	13.2	15.3	35.0	45.8	12.8
Tri5$_{\mathbf{A},D_{\mathbf{ic}}}$	15.8	37.0	48.2	13.0	15.2	37.4	47.6	12.8
Tri5$_{\mathbf{A},D_{\mathbf{BNC+ic}}}$	16.4	37.4	48.3	13.4	15.4	37.0	46.8	13.2
Tri5$^{\sqrt{\mathbf{idf}}}$	16.9	35.4	44.2	12.5	13.0	33.4	43.9	11.3
Tri5$^{\sqrt{\mathbf{idf}}}_{D_{\mathbf{BNC}}}$	16.1	36.0	47.5	12.9	15.0	34.0	44.6	12.2
Tri5$^{\sqrt{\mathbf{idf}}}_{D_{\mathbf{ic}}}$	15.9	36.9	46.8	12.8	16.0	35.8	47.5	13.1
Tri5$^{\sqrt{\mathbf{idf}}}_{D_{\mathbf{BNC+ic}}}$	16.2	37.4	47.5	13.3	15.3	34.8	46.7	13.0
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A}}$	17.4	37.6	46.3	13.4	15.8	36.3	47.3	13.2
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A},D_{\mathbf{BNC}}}$	15.7	36.9	48.1	12.9	15.8	35.3	47.2	12.9
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A},D_{\mathbf{ic}}}$	15.7	37.3	47.3	13.3	15.5	38.3	47.8	13.4
Tri5$^{\sqrt{\mathbf{idf}},\mathbf{Histo}}_{\mathbf{A},D_{\mathbf{BNC+ic}}}$	13.6	30.6	41.9	11.1	14.4	33.8	42.7	12.2
Tri5$^{\sqrt{\mathbf{idf}}}_{\mathbf{A},D_{\mathbf{BNC+ic}}}$	16.6	37.7	49.1	13.7	15.7	36.9	48.5	13.4

Table 12: Performance of all models, measured as the percentage of test items for which they return an item that was deemed relevant according to the crowdsourced judgments among the top 1, 5 or 10 results, and as R-precision computed over these judgments. In Section 4, NN5$^{\text{idf}}_{\text{F1}}$ = NN, Tri5$^{\sqrt{\text{idf}}}_{\text{A},D_{\text{BNC+ic}}}$ = Tri5Sem.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Bach, F. R., & Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, *3*, 1–48.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. D., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 127–134, Toronto, Ontario, Canada.

Bloehdorn, S., Basili, R., Cammisa, M., & Moschitti, A. (2006). Semantic kernels for text classification based on topological measures of feature similarity. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006)*, pp. 808–812, Hong Kong, China.

BNC Consortium (2007). The British National Corpus, version 3 (BNC XML edition). `http://www.natcorp.ox.ac.uk`.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluation the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 249–256, Trento, Italy.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Croce, D., Moschitti, A., & Basili, R. (2011). Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1034–1046, Edinburgh, UK.

Dale, R., & White, M. (Eds.). (2007). *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation: Position Papers*, Arlington, VA, USA.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, *40*(2), 5:1–5:60.

Deschacht, K., & Moens, M.-F. (2007). Text analysis for automatic image annotation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 1000–1007, Prague, Czech Republic.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Everingham, M., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. `http://www.pascal-network.org/challenges/VOC/voc2008/workshop/`.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV), Part IV*, pp. 15–29, Heraklion, Greece.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Anchorage, AK, USA.

Feng, Y., & Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pp. 272–280, Columbus, OH, USA.

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1239–1249, Uppsala, Sweden.

Fisher, R. A. (1935). *The Design of Experiments*. Olyver and Boyd, Edinburgh, UK.

Grangier, D., & Bengio, S. (2008). A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*, 1371–1384.

Grice, H. P. (1975). Logic and conversation. In Davidson, D., & Harman, G. H. (Eds.), *The Logic of Grammar*, pp. 64–75. Dickenson Publishing Co., Encino, CA, USA.

Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006). The IAPR benchmark: A new evaluation resource for visual information systems. In *OntoImage 2006, Workshop on Language Resources for Content-based Image Retrieval during LREC 2006*, pp. 13–23, Genoa, Italy.

Gupta, A., Verma, Y., & Jawahar, C. (2012). Choosing linguistics over vision to describe images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, Toronto, Ontario, Canada.

Hardoon, D. R., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2006). A correlation approach for automatic image annotation. In Li, X., Zaïane, O. R., & Li, Z.-H. (Eds.), *Advanced Data Mining and Applications*, Vol. 4093 of *Lecture Notes in Computer Science*, pp. 681–692. Springer Berlin Heidelberg.

Hardoon, D. R., Szedmak, S. R., & Shawe-Taylor, J. R. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*, 2639–2664.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321–377.

Hwang, S., & Grauman, K. (2012). Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision*, *100*(2), 134–153.

Jaimes, A., Jaimes, R., & Chang, S.-F. (2000). A conceptual framework for indexing visual information at multiple levels. In *Internet Imaging 2000*, Vol. 3964 of *Proceedings of SPIE*, pp. 2–15, San Jose, CA, USA.

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing* (2nd edition). Prentice Hall.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1601–1608.

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. (2012). Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 359–368, Jeju Island, Korea.

Lavrenko, V., Manmatha, R., & Jeon, J. (2004). A model for learning the semantics of pictures. In Thrun, S., Saul, L., & Schölkopf, B. (Eds.), *Advances in Neural Information Processing Systems 16*, Cambridge, MA, USA.

Lazebnik, S., Schmid, C., & Ponce, J. (2009). Spatial pyramid matching. In S. Dickinson, A. Leonardis, B. S., & Tarr, M. (Eds.), *Object Categorization: Computer and Human Vision Perspectives*, chap. 21, pp. 401–415. Cambridge University Press.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL)*, pp. 220–228, Portland, OR, USA.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S. (Ed.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, Barcelona, Spain.

Lin, C.-Y., & Hovy, E. H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 71–78, Edmonton, AB, Canada.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pp. 296–304, Madison, WI, USA.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Internationa Journal of Computer Vision*, *60*(2), 91–110.

Makadia, A., Pavlovic, V., & Kumar, S. (2010). Baselines for image annotation. *International Journal of Computer Vision*, *90*(1), 88–105.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daume III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 747–756, Avignon, France.

Moschitti, A. (2009). Syntactic and semantic kernels for short text pair categorization. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 576–584, Athens, Greece.

Moschitti, A., Pighin, D., & Basili, R. (2008). Tree kernels for semantic role labeling. *Computational Linguistics*, *34*(2), 193–224.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24*, pp. 1143–1151.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, Philadelphia, PA, USA.

Popescu, A., Tsikrika, T., & Kludas, J. (2010). Overview of the Wikipedia retrieval task at ImageCLEF 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, Padua, Italy.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon's Mechanical Turk. In *NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, pp. 139–147, Los Angeles, CA, USA.

Rasiwasia, N., Pereira, J. C., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia (MM)*, pp. 251–260, New York, NY, USA.

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, *35*(4), 529–558.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, *6*, 39–62.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM)*, pp. 623–632, Lisbon, Portugal.

Socher, R., & Li, F.-F. (2010). Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 966–973, San Francisco, CA, USA.

van Erp, M., & Schomaker, L. (2000). Variants of the Borda count method for combining ranked classifier hypotheses. In *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 443–452, Nijmegen, Netherlands.

Varma, M., & Zisserman, A. (2005). A statistical approach to texture classification from single images. *International Journal of Computer Vision*, *62*, 61–81.

Vedaldi, A., & Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

Weston, J., Bengio, S., & Usunier, N. (2010). Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, *81*(1), 21–35.

Yang, Y., Teo, C., Daume III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 444–454, Edinburgh, UK.