Journal of Artificial Intelligence Research 55 (2016) 409-442 Submitted 07/15; published 02/16

Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures

Raffaella Bernardi BERNARDI@DISI.UNITN.IT
University of Trento, Italy

Ruket Cakici RUKEN@CENG.METU.EDU.TR
Middle East Technical University, Turkey

Desmond Elliott D.ELLIOTT@UVA.NL
University of Amsterdam, Netherlands

Aykut Erdem AYKUT@CS.HACETTEPE.EDU.TR
Erkut Erdem ERKUT@CS.HACETTEPE.EDU.TR
Nazli Ikizler-Cinbis NAZLI@CS.HACETTEPE.EDU.TR
Hacettepe University, Turkey

Frank Keller KELLER@INF.ED.AC.UK
University of Edinburgh, UK

Adrian Muscat ADRIAN.MUSCAT@UM.EDU.MT
University of Malta, Malta

Barbara Plank BPLANK@CST.DK
University of Copenhagen, Denmark

Abstract

Automatic description generation from natural images is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities. In this survey, we classify the existing approaches based on how they conceptualize this problem, viz., models that cast description as either generation problem or as a retrieval problem over a visual or multimodal representational space. We provide a detailed review of existing models, highlighting their advantages and disadvantages. Moreover, we give an overview of the benchmark image datasets and the evaluation measures that have been developed to assess the quality of machine-generated image descriptions. Finally we extrapolate future directions in the area of automatic image description generation.

1. Introduction

Over the past two decades, the fields of natural language processing (NLP) and computer vision (CV) have seen great advances in their respective goals of analyzing and generating text, and of understanding images and videos. While both fields share a similar set of methods rooted in artificial intelligence and machine learning, they have historically developed separately, and their scientific communities have typically interacted very little.

Recent years, however, have seen an upsurge of interest in problems that require a combination of linguistic and visual information. A lot of everyday tasks are of this nature, e.g., interpreting a photo in the context of a newspaper article, following instructions in conjunction with a diagram or a map, understanding slides while listening to a lecture. In addition to this, the web provides a vast amount of data that combines linguistic and visual information: tagged photographs, illustrations in newspaper articles, videos with subtitles, and multimodal feeds on social media. To tackle combined language and vision tasks and to exploit the large amounts of multimodal data, the CV and NLP communities have moved closer together, for example by organizing workshops on language and vision that have been held regularly at both CV and NLP conferences over the past few years.

In this new language-vision community, *automatic image description* has emerged as a key task. This task involves taking an image, analyzing its visual content, and generating a textual description (typically a sentence) that verbalizes the most salient aspects of the image. This is challenging from a CV point of view, as the description could in principle talk about any visual aspect of the image: it can mention objects and their attributes, it can talk about features of the scene (e.g., indoor/outdoor), or verbalize how the people and objects in the scene interact. More challenging still, the description could even refer to objects that are not depicted (e.g., it can talk about people waiting for a train, even when the train is not visible because it has not arrived yet) and provide background knowledge that cannot be derived directly from the image (e.g., the person depicted is the Mona Lisa). In short, a good image description requires full image understanding, and therefore the description task is an excellent test bed for computer vision systems, one that is much more comprehensive than standard CV evaluations that typically test, for instance, the accuracy of object detectors or scene classifiers over a limited set of classes.

Image understanding is necessary, but not sufficient for producing a good description. Imagine we apply an array of state-of-the-art detectors to the image to localize objects (e.g., Felzenszwalb, Girshick, McAllester, & Ramanan, 2010; Girshick, Donahue, Darrell, & Malik, 2014), determine attributes (e.g., Lampert, Nickisch, & Harmeling, 2009; Berg, Berg, & Shih, 2010; Parikh & Grauman, 2011), compute scene properties (e.g., Oliva & Torralba, 2001; Lazebnik, Schmid, & Ponce, 2006), and recognize human-object interactions (e.g., Prest, Schmid, & Ferrari, 2012; Yao & Fei-Fei, 2010). The result would be a long, unstructured list of labels (detector outputs), which would be unusable as an image description. A good image description, in contrast, has to be comprehensive but concise (talk about all and only the important things in the image), and has to be formally correct, i.e., consists of grammatically well-formed sentences.

From an NLP point of view, generating such a description is a natural language generation (NLG) problem. The task of NLG is to turn a non-linguistic representation into human-readable text. Classically, the non-linguistic representation is a logical form, a database query, or a set of numbers. In image description, the input is an image representation (e.g., the detector outputs listed in the previous paragraph), which the NLG model has to turn into sentences. Generating text involves a series of steps, traditionally referred to as the NLP pipeline (Reiter & Dale, 2006): we need to decide which aspects of the input to talk about (content selection), then we need to organize the content (text planning) and verbalize it (surface realization). Surface realization in turn requires choosing the right words (lexicalization), using pronouns if appropriate (referential expression generation), and grouping related information together (aggregation).

In other words, automatic image description requires not only full image understanding, but also sophisticated natural language generation. This is what makes it such an interesting task that has been embraced by both the CV and the NLP communities.[1] Note that the description task can become even more challenging when we take into account that good descriptions are often user-specific. For instance, an art critic will require a different description than a librarian or a journalist, even for the same photograph. We will briefly touch upon this issue when we talk about the difference between descriptions and captions in Section 3 and discuss future directions in Section 4.

Given that automatic image description is such an interesting task, and it is driven by the existence of mature CV and NLP methods and the availability of relevant datasets, a large image description literature has appeared over the last five years. The aim of this survey article is to give a comprehensive overview of this literature, covering models, datasets, and evaluation metrics.

We sort the existing literature into three categories based on the image description models used. The first group of models follows the classical pipeline we outlined above: they first detect or predict the image content in terms of objects, attributes, scene types, and actions, based on a set of visual features. Then, these models use this content information to drive a natural language generation system that outputs an image description. We will term these approaches *direct generation models*.

The second group of models cast the problem as a retrieval problem. That is, to create a description for a novel image, these models search for images in a database that are similar to the novel image. Then they build a description for the novel image based on the descriptions of the set of similar images that was retrieved. The novel image is described by simply reusing the description of the most similar retrieved image (transfer), or by synthesizing a novel description based on the description of a set of similar images. Retrieval-based models can be further subdivided based on what type of approach they use to represent images and compute similarity. The first subgroup of models uses a *visual space* to retrieve images, while the second subgroup uses a *multimodal space* that represents images and text jointly. For an overview of the models that will be reviewed in this survey, and which category they fall into, see Table 1.

Generating natural language descriptions from videos presents unique challenges over and above image-based description, as it additionally requires analyzing the objects and their attributes and actions in the temporal dimension. Models that aim to solve description generation from videos have been proposed in the literature (e.g., Khan, Zhang, & Gotoh, 2011; Guadarrama, Krishnamoorthy, Malkarnenkar, Venugopalan, Mooney, Darrell, & Saenko, 2013; Krishnamoorthy, Malkarnenkar, Mooney, Saenko, & Guadarrama, 2013; Rohrbach, Qiu, Titov, Thater, Pinkal, & Schiele, 2013; Thomason, Venugopalan, Guadarrama, Saenko, & Mooney, 2014; Rohrbach, Rohrback, Tandon, & Schiele, 2015; Yao, Torabi, Cho, Ballas, Pal, Larochelle, & Courville, 2015; Zhu, Kiros, Zemel, Salakhutdinov, Urtasun, Torralba, & Fidler, 2015). However, most existing work on description generation has used static images, and this is what we will focus on in this survey.[2]

In this survey article, we first group automatic image description models into the three categories outlined above and provide a comprehensive overview of the models in each

1. Though some image description approaches circumvent the NLG aspect by transferring human-authored descriptions, see Sections 2.2 and 2.3.
2. An interesting intermediate approach involves the annotation of image streams with sequences of sentences, see the work of Park and Kim (2015).

category in Section 2. We then examine the available multimodal image datasets used for training and testing description generation models in Section 3. Furthermore, we review evaluation measures that have been used to gauge the quality of generated descriptions in Section 3. Finally, in Section 4, we discuss future research directions, including possible new tasks related to image description, such as visual question answering.

2. Image Description Models

Generating automatic descriptions from images requires an understanding of how humans describe images. An image description can be analyzed in several different dimensions (Shatford, 1986; Jaimes & Chang, 2000). We follow Hodosh, Young, and Hockenmaier (2013) and assume that the descriptions that are of interest for this survey article are the ones that verbalize visual and conceptual information depicted in the image, i.e., descriptions that refer to the depicted entities, their attributes and relations, and the actions they are involved in. Outside the scope of automatic image description are non-visual descriptions, which give background information or refer to objects not depicted in the image (e.g., the location at which the image was taken or who took the picture). Also, not relevant for standard approaches to image description are perceptual descriptions, which capture the global low-level visual characteristics of images (e.g., the dominant color in the image or the type of the media such as photograph, drawing, animation, etc.).

In the following subsections, we give a comprehensive overview of state-of-the-art approaches to description generation. Table 1 offers a high-level summary of the field, using the three categories of models outlined in the introduction: direct generation models, retrieval models from visual space, and retrieval model from multimodal space.

2.1 Description as Generation from Visual Input

The general approach of the studies in this group is to first predict the most likely meaning of a given image by analyzing its visual content, and then generate a sentence reflecting this meaning. All models in this category achieve this using the following general pipeline architecture:

1. Computer vision techniques are applied to classify the scene type, to detect the objects present in the image, to predict their attributes and the relationships that hold between them, and to recognize the actions taking place.

2. This is followed by a generation phase that turns the detector outputs into words or phrases. These are then combined to produce a natural language description of the image, using techniques from natural language generation (e.g., templates, n-grams, grammar rules).

The approaches reviewed in this section perform an explicit mapping from images to descriptions, which differentiates them from the studies described in Section 2.2 and 2.3, which incorporate implicit vision and language models. An illustration of a sample model is shown in Figure 1. An explicit pipeline architecture, while tailored to the problem at hand, constrains the generated descriptions, as it relies on a predefined sets of semantic classes of scenes, objects, attributes, and actions. Moreover, such an architecture crucially assumes

Reference	Generation	Retrieval from	
		Visual Space	Multimodal Space
Farhadi et al. (2010)			✓
Kulkarni et al. (2011)	✓		
Li et al. (2011)	✓		
Ordonez et al. (2011)		✓	
Yang et al. (2011)	✓		
Gupta et al. (2012)		✓	
Kuznetsova et al. (2012)		✓	
Mitchell et al. (2012)	✓		
Elliott and Keller (2013)	✓		
Hodosh et al. (2013)			✓
Gong et al. (2014)			✓
Karpathy et al. (2014)			✓
Kuznetsova et al. (2014)	✓		
Mason and Charniak (2014)		✓	
Patterson et al. (2014)		✓	
Socher et al. (2014)			✓
Verma and Jawahar (2014)			✓
Yatskar et al. (2014)	✓		
Chen and Zitnick (2015)	✓		✓
Donahue et al. (2015)	✓		✓
Devlin et al. (2015)		✓	
Elliott and de Vries (2015)	✓		
Fang et al. (2015)	✓		
Jia et al. (2015)	✓		✓
Karpathy and Fei-Fei (2015)	✓		✓
Kiros et al. (2015)	✓		✓
Lebret et al. (2015)	✓		✓
Lin et al. (2015)	✓		
Mao et al. (2015a)	✓		✓
Ortiz et al. (2015)	✓		
Pinheiro et al. (2015)			✓
Ushiku et al. (2015)			✓
Vinyals et al. (2015)	✓		✓
Xu et al. (2015)	✓		✓
Yagcioglu et al. (2015)		✓	

Table 1: An overview of existing approaches to automatic image description. We have categorized the literature into approaches that directly generate a description of an image (Section 2.1), approaches that retrieve images via visual similarity and transfer their description to the new image (Section 2.2), and approaches that frame the task as retrieving descriptions and images from a multimodal space (Section 2.3).

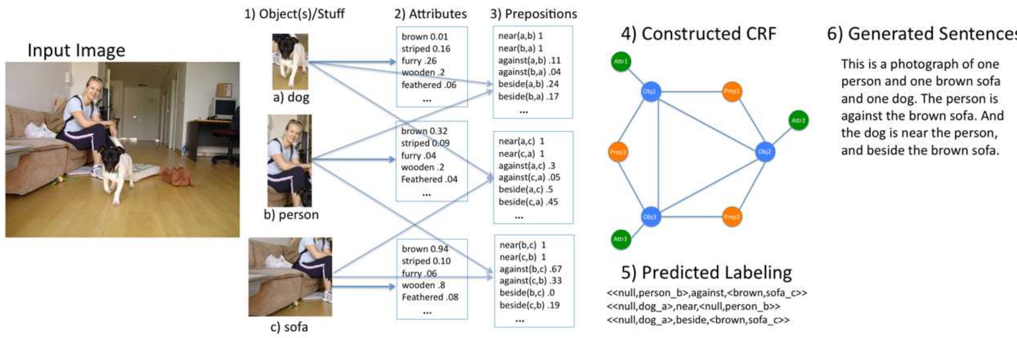


Figure 1: The automatic image description generation system proposed by Kulkarni et al. (2011).

the accuracy of the detectors for each semantic class, an assumption that is not always met in practice.

Approaches to description generation differ along two main dimensions: (a) which image representations they derive descriptions from, and (b) how they address the sentence generation problem. In terms of the representations used, existing models have conceptualized images in a number of different ways, relying on spatial relationships (Farhadi et al., 2010), corpus-based relationships (Yang et al., 2011), or spatial and visual attributes (Kulkarni et al., 2011). Another group of papers utilizes an abstract image representation in the form of meaning tuples which capture different aspects of an image: the objects detected, the attributes of those detections, the spatial relations between them, and the scene type (Farhadi et al., 2010; Yang et al., 2011; Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012). More recently, Yatskar et al. (2014) proposed to generate descriptions from densely-labeled images, which incorporate object, attribute, action, and scene annotations. Similar in spirit is the work by Fang et al. (2015), which does not rely on prior labeling of objects, attributes, etc. Rather, the authors train "word detectors" directly from images and their associated descriptions using multi-instance learning (a weakly supervised approach for the training of object detectors). The words returned by these detectors are then fed into a language model for sentence generation, followed by a re-ranking step.

The first framework to explicitly represent how the structure of an image relates to the structure of its description is the Visual Dependency Representations (VDR) method proposed by Elliott and Keller (2013). A VDR captures the spatial relations between the objects in an image in the form of a dependency graph. This graph can then be related to the syntactic dependency tree of the description of the image.[3] While initial work using VDRs has relied on a corpus of manually annotated VDRs for training, more recent approaches induce VDRs automatically based on the output of an object detector (Elliott & de Vries, 2015) or the labels present in abstract scenes (Ortiz et al., 2015).[4] The idea of explicitly representing image structure and using it for description generation has been picked up

3. VDRs have proven useful not only for description generation, but also for image retrieval (Elliott, Lavrenko, & Keller, 2014).
4. Abstract scenes are schematic images, typically constructed using clip-art. They are employed to avoid the need for an object detector, as the labels and positions of all objects are know. An example is Zitnick and Parikh's (2013) dataset, see Section 3 for details.

by Lin et al. (2015), who parse images into scene graphs, which are similar to VDRs and represent the relations between the objects in a scene. They then generate from scene graphs using a semantic grammar.[5]

Existing approaches also vary along the second dimension, viz., in how they approach the sentence generation problem. At the one end of the scale, there are approaches that use n-gram-based language models. Examples include the works by Kulkarni et al. (2011) and Li et al. (2011), which both generate descriptions using n-gram language models trained on a subset of Wikipedia. These approaches first determine the attributes and relationships between regions in an image as region–preposition–region triples. The n-gram language model is then used to compose an image description that is fluent, given the language model. The approach of Fang et al. (2015) is similar, but uses a maximum entropy language model instead of an n-gram model to generate descriptions. This gives the authors more flexibility in handling the output of the word detectors that are at the core of their model.

Recent image description work using recurrent neural networks (RNNs) can also be regarded as relying on language modeling. A classical RNN is a language model: it captures the probability of generating a given word in a string, given the words generated so far. In an image description setup, the RNN is trained to generate the next word given not only the string so far, but also a set of image features. In this setting, the RNN is therefore not purely a language model (as in the case of an n-gram model, for instance), but it is a hybrid model that relies on a representation that incorporates both visual and linguistic features. We will return to this in more detail in Section 2.3.

A second set of approaches use sentence templates to generate descriptions. These are (typically manually) pre-defined sentence frames in which open slots need to be filled with labels for objects, relations, or attributes. For instance, Yang et al. (2011) fill in a sentence template by selecting the likely objects, verbs, prepositions, and scene types based on a Hidden Markov Model. Verbs are generated by finding the most likely pairing of object labels in the Gigaword external corpus. The generation model of Elliott and Keller (2013) parses an image into a VDR, and then traverses the VDRs to fill the slots of sentence templates. This approach also performs a limited from of content selection by learning associations between VDRs and syntactic dependency trees at training time; these associations then allow to select the most appropriate verb for a description at test time.

Other approaches have used more linguistically sophisticated approaches to generation. Mitchell et al. (2012) over-generate syntactically well-formed sentence fragments and then recombine these using a tree-substitution grammar. A related approach has been pursued by Kuznetsova et al. (2014), where tree-fragments are learnt from a training set of existing descriptions and then these fragments are combined at test time to form new descriptions. Another linguistically expressive model has recently been proposed by Ortiz et al. (2015). The authors model image description as machine translation over VDR–sentence pairs and perform explicit content selection and surface realization using an integer linear program over linguistic constraints.

The systems presented so far aimed at directly generating novel descriptions. However, as argued by Hodosh et al. (2013), framing image description as a natural language generation (NLG) task makes it difficult to objectively evaluate the quality of novel descriptions

5. Note that graphs are also used for image retrieval by Johnson, Krishna, Stark, Li, Shamma, Bernstein, and Fei-Fei (2015) and Schuster, Krishna, Chang, Fei-Fei, and Manning (2015).

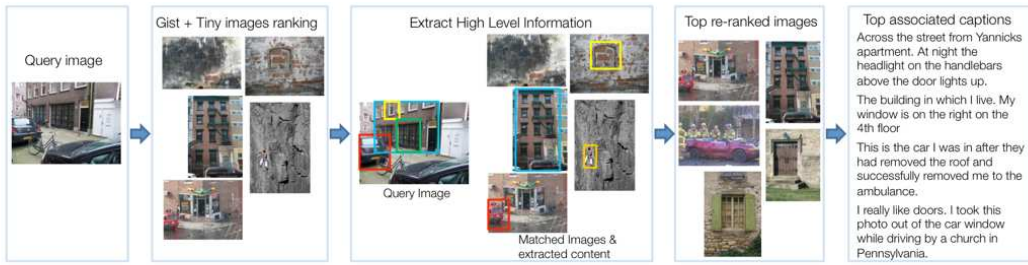


Figure 2: The description model based on retrieval from visual space proposed by Ordonez et al. (2011).

as it "introduces a number of linguistic difficulties that detract attention from the underlying image understanding problem" (Hodosh et al., 2013). At the same time, evaluation of generation systems is known to be difficult (Reiter & Belz, 2009). Hodosh et al. therefore propose an approach that makes it possible to evaluate the mapping between images and sentences independently of the generation aspect. Models that follow this approach conceptualize image description as a *retrieval problem*: they associate an image with a description by retrieving and ranking a set of similar images with candidate descriptions. These candidate descriptions can then either be used directly (description transfer) or a novel description can be synthesized from the candidates (description generation).

The retrieval of images and ranking of their descriptions can be carried out in two ways: either from a visual space or from a multimodal space that combines textual and visual information space. In the following subsections, we will survey work that follows these two approaches.

2.2 Description as a Retrieval in Visual Space

The studies in this group pose the problem of automatically generating the description of an image by retrieving images similar to the query image (i.e., the new image to be described); this is illustrated in Figure 2. In other words, these systems exploit similarity in the visual space to transfer descriptions to the query images. Compared to models that generate descriptions directly (Section 2.1), retrieval models typically require a large amount of training data in order to provide relevant descriptions.

In terms of their algorithmic components, visual retrieval approaches typically follow a pipeline of three main steps:

1. Represent the given query image by specific visual features.
2. Retrieve a candidate set of images from the training set based on a similarity measure in the feature space used.
3. Re-rank the descriptions of the candidate images by further making use of visual and/or textual information contained in the retrieval set, or alternatively combine fragments of the candidate descriptions according to certain rules or schemes.

One of the first model to follow this approach was the Im2Text model of Ordonez et al. (2011). GIST (Oliva & Torralba, 2001) and Tiny Image (Torralba, Fergus, & Freeman, 2008) descriptors are employed to represent the query image and to determine the visually similar images in the first retrieval step. Most of the retrieval-based models consider the result of this step as a baseline. For the re-ranking step, a range of detectors (e.g., object, stuff, pedestrian, action detectors) and scene classifiers specific to the entities mentioned in the candidate descriptions are first applied to the images to better capture their visual content, and the images are represented by means of these detector and classifier responses. Finally, the re-ranking is carried out via a classifier trained over these semantic features.

The model proposed by Kuznetsova et al. (2012) first runs the detectors and the classifiers used in the re-ranking step of the Im2Text model on a query image to extract and represent its semantic content. Then, instead of performing a single retrieval by combining the responses of these detectors and classifiers as the Im2Text model does, it carries out a separate image retrieval step for each visual entity present in the query image to collect related phrases from the retrieved descriptions. For instance, if a dog is detected in the given image, then the retrieval process returns the phrases referring to visually similar dogs in the training set. More specifically, this step is used to collect three different kinds of phrases. Noun and verb phrases are extracted from descriptions in the training set based on the visual similarity between object regions detected in the training images and in the query image. Similarly, prepositional phrases are collected for each stuff detection in the query image by measuring the visual similarity between the detections in the query and training images based on their appearance and geometric arrangements. Prepositional phrases are additionally collected for each scene context detection by measuring the global scene similarity computed between the query and training images. Finally, a description is generated from these collected phrases for each detected object via integer linear programming (ILP) which considers factors such as word ordering, redundancy, etc.

The method of Gupta et al. (2012) is another phrase-based approach. To retrieve visually similar images, the authors employ simple RGB and HSV color histograms, Gabor and Haar descriptors, GIST and SIFT (Lowe, 2004) descriptors as image features. Then, instead of using visual object detectors or scene classifiers, they rely only on the textual information in the descriptions of the visually similar images to extract the visual content of the input image. Specifically, the candidate descriptions are segmented into phrases of a certain type such as (*subject*, *verb*), (*subject*, *prep*, *object*), (*verb*, *prep*, *object*), (*attribute*, *object*), etc. Those that best describe the input image are determined according to a joint probability model based on image similarity and Google search counts, and the image is represented by triplets of the form $\{((attribute1, object1), verb), (verb, prep, (attribute2, object2)), (object1, prep, object2)\}$. In the end, the description is generated using the three top-scoring triplets based on a fixed template. To increase the quality of the descriptions, the authors also apply syntactic aggregation and some subject and predicate grouping rules before the generation step.

Patterson et al. (2014) were the first to present a large-scale scene attribute dataset in the computer vision community. The dataset includes 14,340 images from 707 scene categories, which are annotated with certain attributes from a list of 102 discriminative attributes related to materials, surface properties, lighting, affordances, and spatial layout. This allows them to train attribute classifiers from this dataset. In their paper, the authors also demonstrate that the responses of these attribute classifiers can be used as a global image descriptor which captures the semantic content better than the standard global image descriptors such as GIST. As an application, they extended the baseline model of Im2Text by replacing the global features with automatically extracted scene attributes, giving better image retrieval and description results.

Mason and Charniak's (2014) description generation approach differs from the models discussed above in that it formulates description generation as an extractive summarization problem, and it selects the output description by considering only the textual information in the final re-ranking step. In particular, the authors represented images by using the scene attributes descriptor of Patterson et al. (2014). Once the visually similar images are identified from the training set, in the next step, the conditional probabilities of observing a word in the description of the query image are estimated via non-parametric density estimation using the descriptions of the retrieved images. The final output description is then determined by using two different extractive summarization techniques, one depending on the SumBasic model (Nenkova & Vanderwende, 2005) and the other based on Kullback-Leibler divergence between the word distributions of the query and the candidate descriptions.

Yagcioglu et al. (2015) proposed an average query expansion approach which is based on compositional distributed semantics. To represent images, they use features extracted from the recently proposed Visual Geometry Group convolutional neural network (VGG-CNN; Chatfield, Simonyan, Vedaldi, & Zisserman, 2014). These features are the activations of the last layer of a deep neural network trained on ImageNet, which have been proven to be effective in many computer vision problems. Then, the original query is expanded as the average of the distributed representations of retrieved descriptions, weighted by their similarity to the input image.

The approach of Devlin et al. (2015) also utilizes CNN activations as the global image descriptor and performs k-nearest neighbor retrieval to determine the images from the training set that are visually similar to the query image. It then selects a description from the candidate descriptions associated with the retrieved images that best describes the images that are similar to the query image, just like the approaches by Mason and Charniak (2014) and Yagcioglu et al. (2015). Their approach differs in terms of how they represent the similarity between description and how they select the best candidate over the whole set. Specifically, they propose to compute the description similarity based on the n-gram overlap F-score between the descriptions. They suggest to choose the output description by finding the description that corresponds to the description with the highest mean n-gram overlap with the other candidate descriptions (k-nearest neighbor centroid description) estimated via an n-gram similarity measure.

2.3 Description as a Retrieval in Multimodal Space

The third group of studies casts image description generation again as a retrieval problem, but from a multimodal space (Hodosh et al., 2013; Socher et al., 2014; Karpathy et al., 2014). The intuition behind these models is illustrated in Figure 3, and the overall approach can be characterized as follows:

1. Learn a common multimodal space for the visual and textual data using a training set of image–description pairs.

2. Given a query, use the joint representation space to perform cross-modal (image–sentence) retrieval.

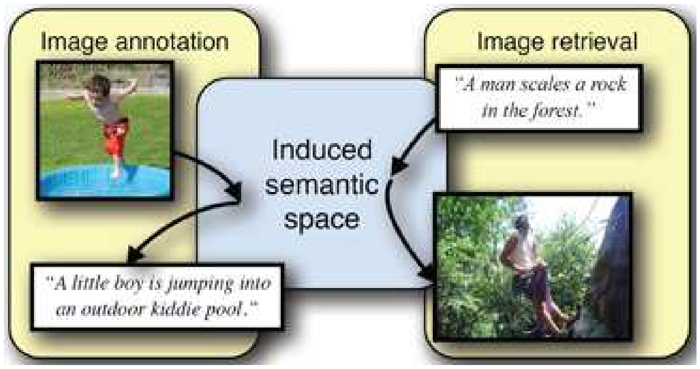


Figure 3: Image descriptions as a retrieval task as proposed in the works by Hodosh et al. (2013), Socher et al. (2014), and Karpathy et al. (2014)[6].

In contrast to the retrieval models that work on a visual space (Section 2.2), where unimodal image retrieval is followed by ranking of the retrieved descriptions, here image and sentence features are projected into a common multimodal space. Then, the multimodal space is used to retrieve descriptions for a given image. The advantage of this approach is that it allows bi-directional models, i.e., the common space can also be used for the other direction, retrieving the most appropriate image for a query sentence.

In this section, we first discuss the seminal paper of Hodosh et al. (2013) on description retrieval, and then present more recent approaches that combine a retrieval approach with some form of natural language generation. Hodosh et al. map both images and sentences into a common space. The joint space can be used for both image search (find the most plausible image given a sentence) and image annotation (find the sentence that describes the image well), see Figure 3. In an earlier study the authors proposed to learn a common meaning space (Farhadi et al., 2010) consisting of a triple representation of the form $\langle object, action, scene\rangle$. The representation was thus limited to a set of pre-defined discrete slot fillers, which was given as training information. Instead, Hodosh et al. use KCCA, a kernelized version of CCA, Canonical Correlation Analysis (Hotelling, 1936), to learn the joint space. CCA takes a training dataset of image-sentence pairs, i.e., $D_{train} = \{\langle \mathbf{i}, \mathbf{s}\rangle\}$, thus input from two different feature spaces, and finds linear projections into a newly induced common space. In KCCA, kernel functions map the original items into higher-order space in order to capture the patterns needed to associate image and text. KCCA has been shown previously to be successful in associating images (Hardoon, Szedmak, & Shawe-Taylor, 2004) or image regions (Socher & Fei-Fei, 2010) with individual words or set of tags.

Hodosh et al. (2013) compare their KCCA approach to a nearest-neighbor (NN) baseline that uses unimodal text and image spaces, without constructing a joint space. A drawback of KCCA is that it is only applicable to smaller datasets, as it requires the two kernel matrices

6. Source `http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/`

to be kept in memory during training. This becomes prohibitive for very large datasets. Some attempts have been made to circumvent the computational burden of KCCA, e.g., by resorting to linear models (Hodosh & Hockenmaier, 2013). Alternatively, Sun, Gan, and Nevatia (2015) have used automatically discovered concepts from images to form a semantic space, and performed sentence retrieval accordingly. However, recent work on description retrieval has instead utilized neural networks to construct a joint space for image description generation.

Socher et al. (2014) use neural networks for building sentence and image vector representations that are then mapped into a common embedding space. A novelty of their work is that they use compositional sentence vector representations. First, image and word representations are learned in their single modalities, and finally mapped into a common multimodal space. In particular, they use a DT-RNN (Dependency Tree Recursive Neural Network) for composing language vectors to abstract over word order and syntactic difference that are semantically irrelevant. This results in 50-dimensional word embeddings. For the image space, the authors use a nine layer neural network trained on ImageNet data, using unsupervised pre-training. Image embeddings are derived by taking the output of the last layer (4,096 dimensions). The two spaces are then projected into a multi-modal space through a max-margin objective function that intuitively trains pairs of correct image and sentence vectors to have a high inner product. The authors show that their model outperforms previously used KCCA approaches such as the work by Hodosh and Hockenmaier (2013).

Karpathy et al. (2014) extend the previous multi-modal embeddings model. Rather than directly mapping entire images and sentences into a common embedding space, their model embeds more fine-grained units, i.e., fragments of images (objects) and sentences (dependency tree fragments), into a common space. Their final model integrates both global (sentence and image-level) as well as finer-grained information and outperforms previous approaches, such as DT-RNN (Socher et al., 2014). A similar approach is pursued by Pinheiro et al. (2015), who propose a bilinear phrase-based model that learns a mapping between image representations and sentences. A constrained language model is then used to generate from this representation. A conceptually related approach is pursued by Ushiku et al. (2015): the authors use a common subspace model which maps all feature vectors associated with the same phrase into nearby regions of the space. For generation, a beam-search based decoder or templates are used.

Description generation systems are difficult to evaluate, therefore the studies reviewed above treat the problem as a retrieval and ranking task (Hodosh et al., 2013; Socher et al., 2014). While such an approach has been valuable because it enables comparative evaluation, retrieval and ranking is limited by the availability of existing datasets with descriptions. To alleviate this problem, recent models have been developed that are extensions of multimodal spaces; they are able to not only rank sentences, but can also generate them (Chen & Zitnick, 2015; Donahue et al., 2015; Karpathy & Fei-Fei, 2015; Kiros et al., 2015; Lebret et al., 2015; Mao et al., 2015a; Vinyals et al., 2015; Xu et al., 2015).

Kiros et al. (2015) introduced a general encoder-decoder framework for image description ranking and generation, illustrated in Figure 4. Intuitively the method works as follows. The encoder first constructs a joint multimodal space. This space can be used to rank images and descriptions. The second stage (decoder) then uses the shared multimodal

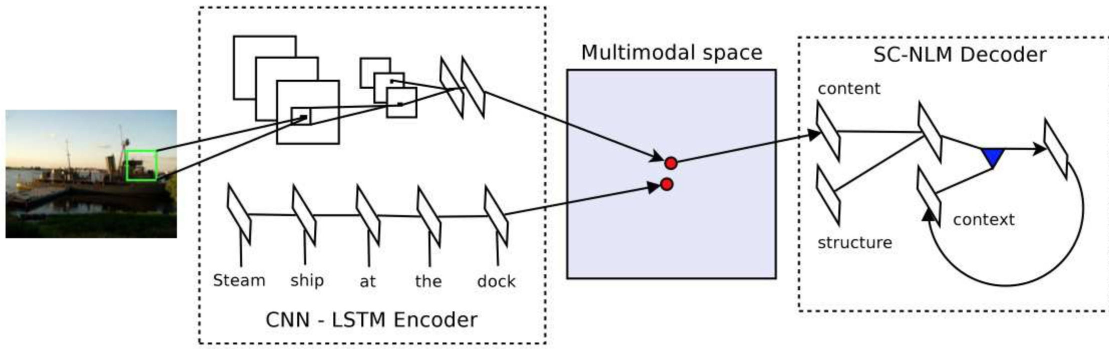


Figure 4: The encoder-decoder model proposed by Kiros et al. (2015).

representation to generate novel descriptions. Their model, directly inspired by recent work in machine translation, encodes sentences using a Long–Short Term Memory (LSTM) recurrent neural network, and image features using a deep convolutional network (CNN). LSTM is an extension of the recurrent neural network (RNN) that incorporates built-in memory to store information and exploit long range context. In Kiros et al.'s (2015) encoder-decoder model, the vision space is projected into the embedding space of the LSTM hidden states; a pairwise ranking loss is minimized to learn the ranking of images and their descriptions. The decoder, a neural-network-based language model, is able to generate novel descriptions from this multimodal space.

Another work that has been carried out at the same time and is similar to the latter is described in the paper by Donahue et al. (2015). The authors propose a model that is also based on the LSTM neural architecture. However, rather than projecting the vision space into the embedding space of the hidden states, the model takes a copy of the static image and the previous word directly as input, that is then fed to a stack of four LSTMs. Another LSTM-based model is proposed by Jia et al. (2015), who added semantic image information as additional input to the LSTM. The model by Kiros et al. (2015) outperforms the prior DT-RNN model (Socher et al., 2014); in turn, Donahue et al. report that they outperform the work of Kiros et al. (2015) on the task of image description retrieval. Subsequent work includes the RNN-based architectures by Mao et al. (2015a) and Vinyals et al. (2015), who are very similar to the one proposed by Kiros et al. (2015) and achieve comparable results on standard datasets. Mao, Wei, Yang, Wang, Huang, and Yuille (2015b) propose an interesting extension of Mao et al.'s (2015a) model for the learning of novel visual concepts.

Karpathy and Fei-Fei (2015) improve on previous models by proposing a deep visual-semantic alignment model with a simpler architecture and objective function. Their key insight is to assume that parts of the sentence refer to particular but unknown regions in the image. Their model tries to infer the alignments between segments of sentences and regions of images and is based on convolutional neural networks over image regions, bidirectional RNN over sentences and a structured objective that aligns the two modalities. Words and image regions are mapped into a common multimodal embedding. The multimodal recurrent neural network architecture uses the inferred alignments to learn and generate novel descriptions. Here, the image is used as condition for the first state in the recurrent neural network, which then generates image descriptions.

Another model that can generate novel sentences is proposed in (Chen & Zitnick, 2015). In contrast to the previous work, their model dynamically builds a visual representation of the scene as a description is being generated. That is, a word is read or generated and the visual representation is updated to reflect the new information. They accomplish this with a simple RNN. The model achieves comparable or better results than most prior studies, except for the recently proposed deep visual-semantic alignment model (Karpathy & Fei-Fei, 2015). The model of Xu et al. (2015) is closely related in that it also uses an RNN-based architecture in which the visual representations are dynamically updated. Xu et al.'s (2015) model incorporates an attentional component, which gives it a way of determining which regions in an image are salient, and it can focus its description on those regions. While resulting in an improvement in description accuracy, it also makes it possible to analyze model behavior by visualizing the regions that were attended to during each word that was generated by the model.

The general RNN-based ranking and generation approach is also followed by Lebret et al. (2015). Here, the main innovation is on the linguistic side: they employ a bilinear model to learn a common space of image features and syntactic phrases (noun phrases, verb phrases, and prepositional phrases). A Markov model is then utilized to generate sentences from these phrase embedding. On the visual side, standard CNN-based features are used. This results in an elegant modeling framework, whose performance is broadly comparable to the state of the art.

Finally, two important directions that are less explored are: portability and weakly supervised learning. Verma and Jawahar (2014) evaluate the portability of a bi-directional model based on topic models, showing that performance significantly degrades. They highlight the importance of cross-dataset image description retrieval evaluation. Another interesting observation is that all of the above models require a training set of fully-annotated image-sentence pairs. However, obtaining such data in large quantities is prohibitively expensive. Gong et al. (2014) propose an approach based on weak supervision that transfers knowledge from millions of weakly annotated images to improve the accuracy of description retrieval.

2.4 Comparison of Existing Approaches

The discussion in the previous subsections makes it clear that each approach to image description has its particular strengths and weaknesses. For example, the methods that cast the task as a generation problem (Section 2.1) have an advantage over other types of approaches in that they can produce novel sentences to describe a given image. However, their success relies heavily on how accurately they estimate the visual content and how well they are able to verbalize this content. In particular, they explicitly employ computer vision techniques to predict the most likely meaning of a given image; these methods have limited accuracy in practice, hence if they fail to identify the most important objects and their attributes, then no valid description can be generated. Another difficulty lies in the final description generation step; sophisticated natural language generation is crucial to guarantee fluency and grammatical correctness of the generated sentences. This can come at the price of considerable algorithmic complexity.

In contrast, image description methods that cast the problem as a retrieval from a visual space problem and transfer the retrieved descriptions to a novel image (Section 2.2) always produce grammatically correct descriptions. This is guaranteed by design, as these systems fetch human-generated sentences from visually similar images. The main issue with this approach is that it requires large amounts of images with human-written descriptions. That is, the accuracy (but not the grammaticality) of the descriptions reduces as the size of the training set decreases. The training set also needs to be diverse (in addition to being large), in order for visual retrieval-based approaches to produce image descriptions that are adequate for novel test images (Devlin et al., 2015). Though this problem can be mitigated by re-synthesizing a novel description from the retrieved ones (see Section 2.2).

Approaches that cast image description as a retrieval from a multimodal space problem (Section 2.3) also have the advantage of generating human-like descriptions as they are able to retrieve the most appropriate ones from a pre-defined large pool of descriptions. However, ranking these descriptions requires a cross-modal similarity metric that compares images and sentences. Such metrics are difficult to define, compared to the unimodal image-to-image similarity metrics used by retrieval models that work on a visual space. Additionally, training a common space for images and sentences requires a large training set of images annotated with human-generated descriptions. On the plus side, such a multimodal embedding space can also be used for the reverse problem, i.e., for retrieving the most appropriate image for a query sentence. This is something generation-based or visual retrieval-based approaches are not capable of.

3. Datasets and Evaluation

There is a wide range of datasets for automatic image description research. The images in these datasets are associated with textual descriptions and differ from each other in certain aspects such as in size, the format of the descriptions and in how the descriptions were collected. Here we review common approaches for collecting datasets, the datasets themselves, and evaluation measures for comparing generated descriptions with ground-truth texts. The datasets are summarized in Table 2, and examples of images and descriptions are given in Figure 5. The readers can also refer to the dataset survey by Ferraro, Mostafazadeh, Huang, Vanderwende, Devlin, Galley, and Mitchell (2015) for an analysis similar to ours. It provides a basic comparison of some of the existing language and vision datasets. It is not limited to automatic image description, and it reports some simple statistics and quality metrics such as perplexity, syntactic complexity, and abstract to concrete word ratios.

3.1 Image-Description Datasets

The Pascal1K sentence dataset (Rashtchian et al., 2010) is a dataset which is commonly used as a benchmark for evaluating the quality of description generation systems. This medium-scale dataset, consists of 1,000 images that were selected from the Pascal 2008 object recognition dataset (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010) and includes objects from different visual classes, such as humans, animals, and vehicles.

	Images	Texts	Judgments	Objects
Pascal1K (Rashtchian et al., 2010)	1,000	5	No	Partial
VLT2K (Elliott & Keller, 2013)	2,424	3	Partial	Partial
Flickr8K (Hodosh & Hockenmaier, 2013)	8,108	5	Yes	No
Flickr30K (Young et al., 2014)	31,783	5	No	No
Abstract Scenes (Zitnick & Parikh, 2013)	10,000	6	No	Complete
IAPR-TC12 (Grubinger et al., 2006)	20,000	1–5	No	Segmented
MS COCO (Lin et al., 2014)	164,062	5	Collected	Partial
BBC News (Feng & Lapata, 2008)	3,361	1	No	No
SBU1M Captions (Ordonez et al., 2011)	1,000,000	1	Collected[7]	No
Déjà-Image Captions (Chen et al., 2015)	4,000,000	Varies	No	No

Table 2: Image datasets for the automatic description generation models. We have split the overview into image *description* datasets (top) and *caption* datasets (bottom) – see the main text for an explanation of this distinction.

Each image is associated with five descriptions generated by humans on Amazon Mechanical Turk (AMT) service.

The Visual and Linguistic Treebank (VLT2K; Elliott & Keller, 2013) makes use of images from the Pascal 2010 action recognition dataset. It augments these images with three, two-sentence descriptions per image. These descriptions were collected on AMT with specific instructions to verbalize the main action depicted in the image and the actors involved (first sentence), while also mentioning the most important background objects (second sentence). For a subset of 341 images of the Visual and Linguistic Treebank, object annotation is available (in the form of polygons around all objects mentioned in the descriptions). For this subset, manually created Visual Dependency Representations (see Section 2.1) are also included (three VDRs per images, i.e., a total of 1023).

The Flickr8K dataset (Hodosh et al., 2013) and its extended version Flickr30K dataset (Young et al., 2014) contain images from Flickr, comprising approximately 8,000 and 30,000 images, respectively. The images in these two datasets were selected through user queries for specific objects and actions. These datasets contain five descriptions per image which were collected from AMT workers using a strategy similar to that of the Pascal1K dataset.

The Abstract Scenes dataset (Zitnick & Parikh, 2013; Zitnick, Parikh, & Vanderwende, 2013) consists of 10,000 clip-art images and their descriptions. The images were created through AMT, where workers were asked to place a fixed vocabulary of 80 clip-art objects into a scene of their choosing. The descriptions were then sourced for these worker-created scenes. The authors provided these descriptions in two different forms. While the first group contains a single sentence description for each image, the second group includes two alternative descriptions per image. Each of these two descriptions consist of three simple sentences with each sentence describing a different aspect of the scene. The main advantage of this dataset is it affords the opportunity to explore image description generation without

7. Kuznetsova et al. (2014) ran a human judgments study on 1,000 images from this dataset.

1. One jet lands at an airport while another takes off next to it.
2. Two airplanes parked in an airport.
3. Two jets taxi past each other.
4. Two parked jet airplanes facing opposite directions.
5. two passenger planes on a grassy plain

(a) Pascal1K[8]

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

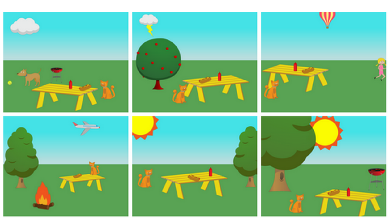
(b) VLT2K[9]

1. A man is snowboarding over a structure on a snowy hill.
2. A snowboarder jumps through the air on a snowy hill.
3. a snowboarder wearing green pants doing a trick on a high bench
4. Someone in yellow pants is on a ramp over the snow.
5. The man is performing a trick on a snowboard high in the air.

(c) Flickr8K[10]

1. a yellow building with white columns in the background
2. two palm trees in front of the house
3. cars are parking in front of the house
4. a woman and a child are walking over the square

(d) IAPR-TC12[11]



1. A cat anxiously sits in the park and stares at an unattended hot dog that someone has left on a yellow bench

(e) Abstract Scenes[12]

1. A blue smart car parked in a parking lot.
2. Some vehicles on a very wet wide city street.
3. Several cars and a motorcycle are on a snow covered street.
4. Many vehicles drive down an icy street.
5. A small smart car driving in the city.

(f) MS COCO[13]

Figure 5: Example images and descriptions from the benchmark image datasets.

the need for automatic object recognition, thus avoiding the associated noise. A more recent version of this dataset has been created as a part of the visual question-answering (VQA) dataset (Antol, Agrawal, Lu, Mitchell, Batra, Zitnick, & Parikh, 2015). It contains 50,000 different scene images with more realistic human models and with five single-sentence descriptions.

The IAPR-TC12 dataset introduced by Grubinger et al. (2006) is one of the earliest multi-modal datasets and contains 20,000 images with descriptions. The images were originally retrieved via search engines such as Google, Bing and Yahoo, and the descriptions were produced in multiple languages (predominantly English and German). Each image is associated with one to five descriptions, where each description refers to a different aspect of the image, where applicable. The dataset also contains complete pixel-level segmentation of the objects.

The MS COCO dataset (Lin et al., 2014) currently consists of 123,287 images with five different descriptions per image. Images in this dataset are annotated for 80 object categories, which means that bounding boxes around all instances in one of these categories are available for all images. The MS COCO dataset has been widely used for image description, something that is facilitated by the standard evaluation server that has recently become available[14]. Extensions of MS COCO are currently under development, including the addition of questions and answers (Antol et al., 2015).

One paper (Lin et al., 2015) uses an the NYU dataset (Silberman, Kohli, Hoiem, & Fergus, 2012), which contains 1,449 indoor scenes with 3D object segmentation. This dataset has been augmented with five descriptions per image by Lin et al.

3.2 Image-Caption Datasets

Image descriptions verbalize what can be seen in the image, i.e., they refer to the objects, actions, and attributes depicted, mention the scene type, etc. Captions, on the other hand, are typically texts associated with images that verbalize information that cannot be seen in the image. A caption provides personal, cultural, or historical context for the image (Panofsky, 1939). Images shared through social networking or photo-sharing websites can be accompanied by descriptions or captions, or a mixtures of both types of text. The images in a newspaper or a museum will typically contain cultural or historical texts, i.e., captions not descriptions.

The BBC News dataset (Feng & Lapata, 2008) was one of the earliest collections of images and co-occurring texts. Feng and Lapata (2008) harvested 3,361 news articles from the British Broadcasting Corporation News website, with the constraint that the article includes an image and a caption.

8. Source `http://nlp.cs.illinois.edu/HockenmaierGroup/pascal-sentences/index.html`
9. Source `http://github.com/elliottd/vlt`
10. Source `https://illinois.edu/fb/sec/1713398`
11. Source `http://imageclef.org/photodata`
12. Source `http://research.microsoft.com/en-us/um/people/larryz/clipart/SemanticClassesRender/Classes_v1.html`
13. Source `http://mscoco.org/explore`
14. Source `http://mscoco.org/dataset/#captions-eval`

The SBU1M Captions dataset introduced by Ordonez et al. (2011) differs from the previous datasets in that it is a web-scale dataset containing approximately one million captioned images. It is compiled from data available on Flickr with user-provided image descriptions. The images were downloaded and filtered from Flickr with the constraint that an image contained at least one noun and one verb on predefined control lists. The resulting dataset is provided as a CSV file of URLs.

The Déjà-Image Captions dataset (Chen et al., 2015) contains 4,000,000 images with 180,000 near-identical captions harvested from Flickr. 760 million images were downloaded from Flickr during the calendar year 2013 using a set of 693 nouns as queries. The image captions are normalized through lemmatization and stop word removal to create a corpus of the near-identical texts. For instance, the sentences *the bird flies in blue sky* and *a bird flying into the blue sky* were normalized to *bird fly IN blue sky* (Chen et al., 2015). Image–caption pairs are retained if the captions are repeated by more than one user in normalized form.

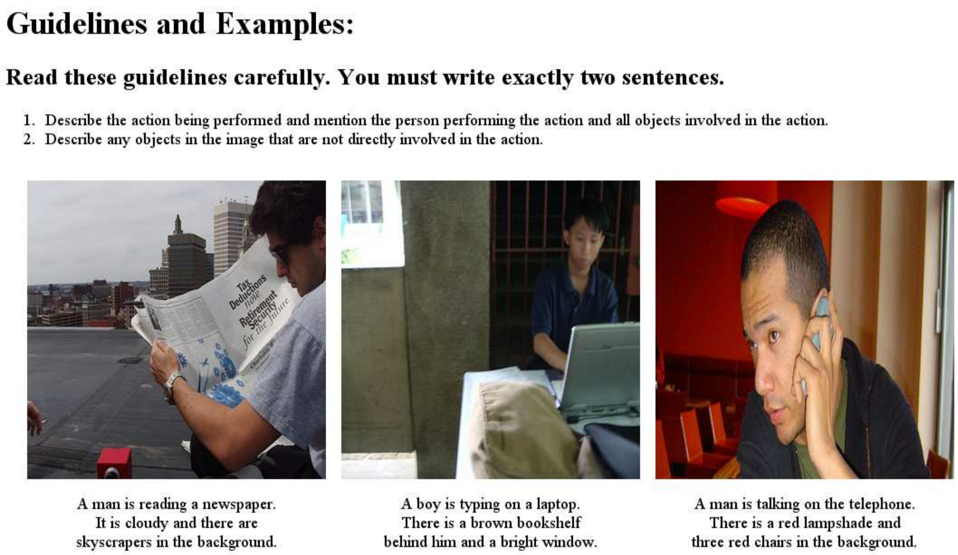
3.3 Collecting Datasets

Collecting new image–text datasets is typically performed through crowd-sourcing or harvesting data from the web. The images for these datasets have either been sourced from an existing task in the computer vision community – the Pascal challenge (Everingham et al., 2010) was used to the Pascal1K and VLT2K datasets – directly from Flickr, in the case of Flickr8K/30K, MS COCO, SBU1M Captions, and Déjà-Image Captions datasets, or crowdsourced, in the case of the Abstract Scenes dataset. The texts in image–*description* datasets are usually crowd-sourced from Amazon Mechanical Turk or Crowdflower; whereas the texts in image–*caption* datasets have been harvested from photo-sharing sites, such as Flickr, or from news providers. Captions are usually collected without financial incentive because they are written by the people sharing their own images, or by journalists.

Crowd-sourcing the descriptions of images involves defining a simple task that can be performed by untrained workers. Examples of the task guidelines used by Hodosh et al. (2013) and Elliott and Keller (2013) are given in Figure 6. In both instances, care was taken to clearly inform the potential workers about the expectations for the task. In particular, explicit instructions were given on how the descriptions should be written, and examples of good texts were provided. In addition, Hodosh et al. provided more extensive examples to explain what would constitute unsatisfactory texts. Further options are available to control the quality of the collected texts: a minimum performance rate for workers is a common choice; and a pre-task selection quiz may be used to determine whether workers have a sufficient grasp of the English language (Hodosh et al., 2013).

The issue of remuneration for crowd-sourced workers is controversial, and higher payments do not always lead to better quality in a crowd-sourced environment (Mason & Watts, 2009). Rashtchian et al. (2010) paid \$0.01/description, Elliott and Keller (2013) paid \$0.04 for an average of 67 seconds of work to produce a two-sentence description. To the best of our knowledge, such information is not available for the other datasets.

(a) Mechanical Turk Interface used to collect Flickr8K dataset[15].



(b) Mechanical Turk Interface used to collect VLT2K dataset.

Figure 6: Examples of Mechanical Turk interfaces for collecting descriptions.

3.4 Evaluation Measures

Evaluating the output of a natural language generation (NLG) system is a fundamentally difficult task (Dale & White, 2007; Reiter & Belz, 2009). The most common way to assess the quality of automatically generated texts is the subjective evaluation by human experts.

15. Source Appendix of the work by Hodosh et al. (2013)

NLG-produced text is typically judged in terms of grammar and content, indicating how syntactically correct and how relevant the text is, respectively. Fluency of the generated text is sometimes tested as well, especially when a surface realization technique is involved during the generation process. Automatically generated descriptions for images can be evaluated using the same NLG techniques. Typically, judges are provided with the image as well as with the description during evaluation tasks. Subjective human evaluations of machine generated image descriptions are often performed on Mechanical Turk with the help of questions. So far, the following Likert-scale questions have been used to test datasets and user groups of various sizes.

- The description accurately describes the image (Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013; Hodosh et al., 2013).
- The description is grammatically correct (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013, inter alia).
- The description has no incorrect information (Mitchell et al., 2012).
- The description is relevant for this image (Li et al., 2011; Yang et al., 2011).
- The description is creatively constructed (Li et al., 2011).
- The description is human-like (Mitchell et al., 2012).

Another approach for evaluating descriptions is to use automatic measures, such as BLEU (Papineni, Roukos, Ward, & Zhu, 2002), ROUGE (Lin & Hovy, 2008), Translation Error Rate (Feng & Lapata, 2013), Meteor (Denkowski & Lavie, 2014), or CIDEr (Vedantam, Lawrence Zitnick, & Parikh, 2015). These measures were originally developed to evaluate the output of machine translation engines or text summarization systems, with the exception of CIDEr, which was developed specifically for image description evaluation. All these measures compute a score that indicates the similarity between the system output and one or more human-written reference texts (e.g., ground truth translations or summaries). This approach to evaluation has been subject to much discussion and critique (Kulkarni et al., 2011; Hodosh et al., 2013; Elliott & Keller, 2014). Kulkarni et al. found weakly negative or no correlation between human judgments and unigram BLEU on the Pascal 1K Dataset (Pearson's ρ = -0.17 and 0.05). Hodosh et al. studied the Cohen's κ correlation of expert human judgments and binarized unigram BLEU and unigram ROUGE of retrieved descriptions on the Flickr8K dataset. They found the best agreement between humans and BLEU ($\kappa = 0.72$) or ROUGE ($\kappa = 0.54$) when the system retrieved the sentences originally associated with the images. Agreement dropped when only one reference sentence was available, or when the reference sentences were disjoint from the proposal sentences. They concluded that neither measure was appropriate for image description evaluation and subsequently proposed image–sentence ranking experiments, discussed in more detail below. Elliott and Keller analyzed the correlation between human judgments and automatic evaluation measures for retrieved and system-generated image descriptions in the Flickr8K and VLT2K datasets. They showed that sentence-level unigram BLEU, which at that point in time was the de facto standard measure for image description evaluation, is only weakly correlated with human judgments. Meteor (Banerjee & Lavie, 2005), a less frequently used translation evaluation measure, exhibited the highest correlation with human judgments. However, Kuznetsova et al. (2014) found that unigram BLEU was more strongly correlated with human judgments than Meteor for image *caption* generation.

The first large-scale image description evaluation took place during the MS COCO Captions Challenge 2015,[16] featuring 15 teams with a dataset of 123,716 training images and 41,000 images in a withheld test dataset. The number of reference texts for each testing image was either five or 40, based on the insight that some measures may benefit from larger reference sets (Vedantam et al., 2015). When automatic evaluation measures were used, some of the image description systems outperformed a human–human upper bound,[17] whether five or 40 reference descriptions were provided. However, none of the systems outperformed human–human evaluation when a judgment elicitation task was used. Meteor was found to be the most robust measure, with the systems beating the human text on one and two submissions (depending on the number of references); the systems outperformed humans seven or five times measured with CIDEr; according to ROUGE and BLEU, the system nearly always outperformed the humans, further confirming the unsuitability of these evaluation measures.

The models that approach the description generation problem from a cross-modal retrieval perspective (Hodosh & Hockenmaier, 2013; Hodosh et al., 2013; Socher et al., 2014; Gong et al., 2014; Karpathy et al., 2014; Verma & Jawahar, 2014) are also able to use measures from information retrieval, such as median rank (mRank), precision at k (S@k), or recall at k (R@k) to evaluate the descriptions they return, in addition to the text-similarity measures reported above. This evaluation paradigm was first proposed by Hodosh et al., who reported high correlation with human judgments for image–sentence based ranking evaluations.

In Table 3, we summarize all the image description approaches discussed in this survey, and list the datasets and evaluation measures employed by each of these approaches. It can be seen that more recent systems (starting in 2014) have converged on the use of large description datasets (Flickr8K/30K, MS COCO) and employ evaluation measures that perform well in terms of correlation with human judgments (Meteor, CIDEr). However, the use of BLEU, despite its limitations, is still widespread; also the use of human evaluation is by no means universal in the literature.

4. Future Directions

As this survey demonstrates, the CV and NLP communities have witnessed an upsurge in interest in automatic image description systems. With the help of recent advances in deep learning models for images and text, substantial improvements in the quality of automatically generated descriptions has been registered. Nevertheless, a series of challenges for image description research remain. In the following, we discuss future directions that this line of research is likely to benefit from.

16. Source `http://mscoco.org/dataset/cap2015`
17. Calculated by collecting an additional human-written description, which was then compared to the reference descriptions.

Reference	Approach	Datasets	Measures
Farhadi et al. (2010)	MultRetrieval	Pascal1K	BLEU
Kulkarni et al. (2011)	Generation	Pascal1K	Human, BLEU
Li et al. (2011)	Generation	Pascal1K	Human, BLEU
Ordonez et al. (2011)	VisRetrieval	SBU1M	
Yang et al. (2011)	Generation	IAPR, Flickr8K/30K, COCO	BLEU, ROUGE, Meteor, CIDEr, R@k
Gupta et al. (2012)	VisRetrieval	Pascal1K, IAPR	Human, BLEU, ROUGE
Kuznetsova et al. (2012)	VisRetrieval	SBU1M	Human, BLEU
Mitchell et al. (2012)	Generation	Pascal1K	Human
Elliott and Keller (2013)	Generation	VLT2K	Human, BLEU
Hodosh et al. (2013)	MultRetrieval	Pascal1K, Flickr8K	Human, BLEU, ROUGE, mRank, R@k
Gong et al. (2014)	MultRetrieval	SBU1M, Flickr30K	R@k
Karpathy et al. (2014)	MultRetrieval	Flickr8K/30K, COCO	BLEU, Meteor, CIDEr
Kuznetsova et al. (2014)	Generation	SBU1M	Human, BLEU, Meteor
Mason and Charniak (2014)	VisRetrieval	SBU1M	Human, BLEU
Patterson et al. (2014)	VisRetrieval	SBU1M	BLEU
Socher et al. (2014)	MultRetrieval	Pascal1K	mRank, R@k
Verma and Jawahar (2014)	MultRetrieval	IAPR, SBU1M, Pascal1K	BLEU, ROUGE, P@k
Yatskar et al. (2014)	Generation	Own data	Human, BLEU
Chen and Zitnick (2015)	MultRetrieval	Flickr8K/30K, COCO	BLEU, Meteor, CIDEr, mRank, R@k
Donahue et al. (2015)	MultRetrieval	Flickr30K, COCO	Human, BLEU, mRank, R@k
Devlin et al. (2015)	VisRetrieval	COCO	BLEU, Meteor
Elliott and de Vries (2015)	Generation	VLT2K, Pascal1K	BLEU, Meteor
Fang et al. (2015)	Generation	COCO	Human, BLEU, ROUGE, Meteor, CIDEr
Jia et al. (2015)	Generation	Flickr8K/30K, COCO	BLEU, Meteor
Karpathy and Fei-Fei (2015)	MultRetrieval	Flickr8K/30K, COCO	BLEU, Meteor, CIDEr, mRank, R@k
Kiros et al. (2015)	MultRetrieval	Flickr8K/30K	R@k
Lebret et al. (2015)	MultRetrieval	Flickr30K, COCO	BLEU, R@k
Lin et al. (2015)	Generation	NYU	ROUGE
Mao et al. (2015a)	MultRetrieval	IAPR, Flickr30K, COCO	BLEU, mRank, R@k
Ortiz et al. (2015)	Generation	Abstract Scenes	Human, BLEU, Meteor
Pinheiro et al. (2015)	MultRetrieval	COCO	BLEU
Ushiku et al. (2015)	Generation	Pascal1K, IAPR, SBU1M, COCO	BLEU
Vinyals et al. (2015)	MultRetrieval	Pascal1K, SBU1M, Flickr8K/30K	BLEU, Meteor, CIDEr, mRank, R@k
Xu et al. (2015)	MultRetrieval	Flickr8K/30K, COCO	BLEU, Meteor
Yagcioglu et al. (2015)	VisRetrieval	Flickr8K/30K, COCO	Human, BLEU, Meteor, CIDEr

Table 3: An overview of the approaches, datasets, and evaluation measures reviewed in this survey and organised in chronological order.

4.1 Datasets

The earliest work on image description used relatively small datasets (Farhadi et al., 2010; Kulkarni et al., 2011; Elliott & Keller, 2013). Recently, the introduction of Flickr30K, MS COCO and other large datasets has enabled the training of more complex models such as neural networks. Still, the area is likely to benefit from larger and diversified datasets that share a common, unified, comprehensive vocabulary. Vinyals et al. (2015) argue that the collection process and the quality of the descriptions in the datasets affect performance significantly, and make transfer learning between datasets not as effective as expected. They show that learning a model from MS COCO and applying it to datasets collected in different settings such as SBU1M Captions or Pascal1K, leads to a degradation in BLEU performance. This is surprising, since MS COCO offers a much larger amount of training data than Pascal1K. As Vinyals et al. put it, this is largely due to the differences in vocabulary and in the quality of descriptions. Most learning approaches are likely to suffer from such situations. Collecting larger and comprehensive datasets and developing more generic approaches that are capable of generating naturalistic descriptions across domains therefore is an open challenge.

While supervised algorithms are likely to take advantage of carefully collected large datasets, lowering the amount of supervision in exchange of access to larger unsupervised data is also an interesting avenue for future research. Leveraging unsupervised data for building richer representations and description models is another open research challenge in this context.

4.2 Measures

Designing automatic measures that can mimic human judgments in evaluating the suitability of image descriptions is perhaps the most urgent need in the area of image description (Elliott & Keller, 2014). This need can be dramatically observed at the latest evaluation results of MS COCO Challenge. According to existing measures, including the latest CIDEr measure (Vedantam et al., 2015), several automatic methods outperform the human upper bound (this upper bound indicates how similar human descriptions are to each other). The counterintuitive nature of this result is confirmed by the fact that when human judgments are used for evaluation, the output of even the best system is judged as worse than a human generated description for most of the time (Fang et al., 2015). However, since conducting human judgment experiments is costly, there is a major need for improved automatic measures that are more highly correlated with human judgments. Figure 7 plots the Epanechnikov probability density estimate (a non-parametric optimal estimator) for BLEU, Meteor, ROUGE, and CIDEr scores per subjective judgment in Flickr8K dataset. The human judgments were obtained from human experts (Hodosh et al., 2013). BLEU is once again confirmed to be unable to sufficiently discriminate between the lowest three human judgments, while Meteor and CIDEr show signs of moving towards a useful separation.

4.3 Diversity and Originality

Current algorithms often rely on direct representations of the descriptions they see at training time, making the descriptions generated at test time very similar. This results in many

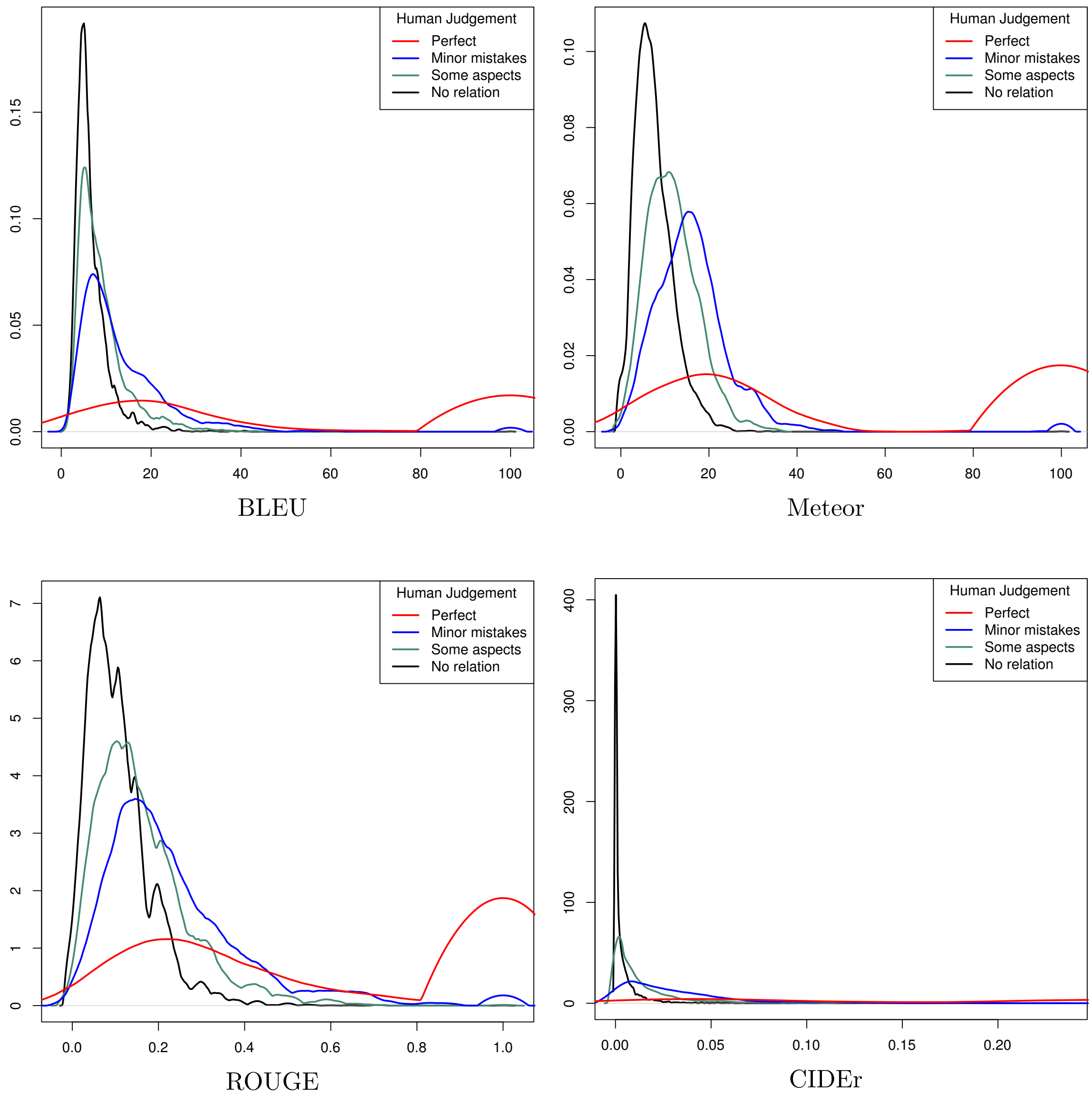


Figure 7: Probability density estimates of BLEU, Meteor, ROUGE, and CIDEr scores against human judgments in the Flickr8K dataset. The y-axis shows the probability density, and the x-axis is the score computed by the measure.

repetitions and limits the diversity of the generated descriptions, making it difficult to reach human levels of performance. This situation has been demonstrated by Devlin et al. (2015), who show that their best model is able to generate only 47.0% of unique descriptions. Systems that generate diverse and original descriptions that do not just repeat what is already seen, but also infer the underlying semantics therefore remain as an open challenge. Chen and Zitnick (2015) and related approaches take a step towards addressing such limitations by coupling description and visual representation generation.

Jas and Parikh (2015) introduces the notion of image specificity, arguing that the domain of image descriptions is not uniform, certain images being more specific than others. Descriptions of non-specific images tend to vary a lot as people tend to describe a non-specific scene from different aspects. This notion and its effects to description systems and measures should be investigated in further detail.

4.4 Further Tasks

Another open challenge is visual question-answering (VQA). While natural language question-answering based on text has been a significant goal of NLP research for a long time (e.g., Liang, Jordan, & Klein, 2012; Fader, Zettlemoyer, & Etzioni, 2013; Richardson, Burges, & Renshaw, 2013; Fader, Zettlemoyer, & Etzioni, 2014), answering questions about images is a task that has recently emerged. Towards achieving this goal, Malinowski and Fritz (2014a) propose a Bayesian framework that connects natural language question-answering with the visual information extracted from image parts. More recently, image question answering methods based on neural networks have been developed (Gao, Mao, Zhou, Huang, & Yuille, 2015; Ren, Kiros, & Zemel, 2015; Malinowski, Rohrbach, & Fritz, 2015; Ma, Lu, & Li, 2016). Following this effort, several datasets on this task are being released: DAQUAR (Malinowski & Fritz, 2014a) was compiled from scene depth images and mainly focuses on questions about the type, quantity and color of objects; COCO-QA (Ren et al., 2015) was constructed by converting image descriptions to VQA format over a subset of images from the MS COCO dataset; the Freestyle Multilingual Image Question Answering (FM-IQA) Dataset (Gao et al., 2015), Visual Madlibs dataset (Yu, Park, Berg, & Berg, 2015) and the VQA dataset (Antol et al., 2015), were again built for images from MS COCO, but this time question-answer pairs are collected via human annotators in a freestyle paradigm. Research in this emerging field is likely to flourish in the near future. The ultimate goal of VQA is to build systems that can pass the (recently developed) Visual Turing Test by being able to answer arbitrary questions about images with the same precision as a human observer (Malinowski & Fritz, 2014b; Geman, Geman, Hallonquist, & Younes, 2015).

Having multilingual repositories for image description is an interesting direction to explore. Currently, among the available benchmark datasets, only the IAPR-TC12 dataset (Grubinger et al., 2006) has multilingual descriptions (in English and German). Future work should investigate whether transferring multimodal features between monolingual description models results in improved descriptions compared to monolingual baselines.

It would be interesting to study different models and new tasks in a multilingual multimodal setting using larger and more syntactically diverse multilingual description corpora.[18]

Overall, image understanding is the ultimate goal of computer vision and natural language generation is one of the ultimate goals of NLP. Image description is where these both goals are interconnected and this topic is therefore likely to benefit from individual advances in each of these two fields.

5. Conclusions

In this survey, we discuss recent advances in automatic image description and closely related problems. We review and analyze a large body of the existing work by highlighting common characteristics and differences between existing research. In particular, we categorize the related work into three groups: (i) direct description generation from images, (i) retrieval of images from a visual space, and (iii) retrieval of images from multimodal (joint visual and linguistic) space. In addition, we provided a brief review of the existing corpora and automatic evaluation measures, and discussed some future directions for vision and language research.

Compared to traditional keyword-based image annotation (using object recognition, attribute detection, scene labeling, etc.), automatic image description systems produce more human-like explanations of visual content, providing a more complete picture of the scene. Advancements in this field could lead to more intelligent artificial vision systems, which can make inferences about the scenes through the generated grounded image descriptions and therefore interact with their environments in a more natural manner. They could also have a direct impact on technological applications from which visually impaired people can benefit through more accessible interfaces.

Despite the remarkable increase in the number of image description systems in recent years, experimental results suggest that system performance still falls short of human performance. A similar challenge lies in the automatic evaluation of systems using reference descriptions. The measures and the tools currently in use are not sufficiently highly correlated with human judgments, indicating a need for measures that can deal with the complexity of the image description problem adequately.

Acknowledgments

We thank the anonymous reviewers for their useful comments. This work has been partially supported by the European Commission ICT COST Action "iV&L Net: The European Network on Integrating Vision and Language" (IC1037). RC, AE, EE, NIC was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) research grant 113E116. FK would like to acknowledge ERC funding through starting grant 203427 "Synchronous Linguistic and Visual Processing". DE was supported by ERCIM ABCDE Fellowship 2014-23.

18. The Multimodal Translation Shared Task at the 2016 Workshop on Machine Translation will use an English and German translated version of the Flickr30K corpora. See `http://www.statmt.org/wmt16/multimodal-task.html` for more details.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *International Conference on Computer Vision*.

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Annual Meeting of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Berg, T. L., Berg, A. C., & Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*.

Chen, J., Kuznetsova, P., Warren, D., & Choi, Y. (2015). Déjà image-captions: A corpus of expressive descriptions in repetition. In *North American Chapter of the Association for Computational Linguistics*.

Chen, X., & Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

Denkowski, M., & Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Conference of the European Chapter of the Association for Computational Linguistics Workshop on Statistical Machine Translation*.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., & Mitchell, M. (2015). Language Models for Image Captioning: The Quirks and What Works. In *Annual Meeting of the Association for Computational Linguistics*.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Elliott, D., & de Vries, A. P. (2015). Describing images using inferred visual dependency representations. In *Annual Meeting of the Association for Computational Linguistics*.

Elliott, D., & Keller, F. (2013). Image Description using Visual Dependency Representations. In *Conference on Empirical Methods in Natural Language Processing*.

Elliott, D., & Keller, F. (2014). Comparing Automatic Evaluation Measures for Image Description. In *Annual Meeting of the Association for Computational Linguistics*.

Elliott, D., Lavrenko, V., & Keller, F. (2014). Query-by-Example Image Retrieval using Visual Dependency Representations. In *International Conference on Computational Linguistics*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fader, A., Zettlemoyer, L., & Etzioni, O. (2013). Paraphrase-driven learning for open question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Fader, A., Zettlemoyer, L., & Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, C. L., & Zweig, G. (2015). From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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 *European Conference on Computer Vision*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Feng, Y., & Lapata, M. (2008). Automatic Image Annotation Using Auxiliary Text Information. In *Annual Meeting of the Association for Computational Linguistics*.

Feng, Y., & Lapata, M. (2013). Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(4), 797–812.

Ferraro, F., Mostafazadeh, N., Huang, T., Vanderwende, L., Devlin, J., Galley, M., & Mitchell, M. (2015). A survey of current datasets for vision and language research. In *Conference on Empirical Methods in Natural Language Processing*.

Gao, H., Mao, J., Zhou, J., Huang, Z., & Yuille, A. (2015). Are you talking to a machine? dataset and methods for multilingual image question answering. In *International Conference on Learning Representations*.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, *112*(12), 3618–3623.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., & Lazebnik, S. (2014). Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *European Conference on Computer Vision*.

Grubinger, M., Clough, P., Muller, H., & Deselaers, T. (2006). The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation*.

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., & Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *International Conference on Computer Vision*.

Gupta, A., Verma, Y., & Jawahar, C. V. (2012). Choosing linguistics over vision to describe images. In *AAAI Conference on Artificial Intelligence*.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664.

Hodosh, M., & Hockenmaier, J. (2013). Sentence-based image description with scalable, explicit models. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *0*, 321–377.

Jaimes, A., & Chang, S.-F. (2000). A conceptual framework for indexing visual information at multiple levels. In *IST SPIE Internet Imaging*.

Jas, M., & Parikh, D. (2015). Image specificity. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Jia, X., Gavves, E., Fernando, B., & Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *International Conference on Computer Vision*.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

Karpathy, A., Joulin, A., & Fei-Fei, L. (2014). Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*.

Khan, M. U. G., Zhang, L., & Gotoh, Y. (2011). Towards coherent natural language description of video streams. In *International Conference on Computer Vision Workshops*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2015). Unifying visual-semantic embeddings with multimodal neural language models. In *Advances in Neural Information Processing Systems Deep Learning Workshop*.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., & Guadarrama, S. (2013). Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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 *IEEE Conference on Computer Vision and Pattern Recognition*.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012). Collective Generation of Natural Image Descriptions. In *Annual Meeting of the Association for Computational Linguistics*.

Kuznetsova, P., Ordonezz, V., Berg, T. L., & Choi, Y. (2014). TREETALK: Composition and compression of trees for image descriptions. In *Conference on Empirical Methods in Natural Language Processing*.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Lebret, R., Pinheiro, P. O., & Collobert, R. (2015). Phrase-based image captioning. In *International Conference on Machine Learning*.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *The SIGNLL Conference on Computational Natural Language Learning*.

Liang, P., Jordan, M. I., & Klein, D. (2012). Learning dependency-based compositional semantics. *Computational Linguistics*, *39*(2), 389–446.

Lin, C.-Y., & Hovy, E. (2008). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Lin, D., Fidler, S., Kong, C., & Urtasun, R. (2015). Generating multi-sentence natural language descriptions of indoor scenes. In *British Machine Vision Conference*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(4), 91–110.

Ma, L., Lu, Z., & Li, H. (2016). Learning to answer questions from image using convolutional neural network. In *AAAI Conference on Artificial Intelligence*.

Malinowski, M., & Fritz, M. (2014a). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*.

Malinowski, M., & Fritz, M. (2014b). Towards a visual turing challenge. In *Advances in Neural Information Processing Systems Workshop on Learning Semantics*.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *International Conference on Computer Vision*.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2015a). Deep captioning with multimodal recurrent neural networks (m-RNN). In *International Conference on Learning Representations*.

Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., & Yuille, A. L. (2015b). Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *International Conference on Computer Vision*.

Mason, R., & Charniak, E. (2014). Nonparametric Method for Data-driven Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Mason, W. A., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". In *ACM SIGKDD Workshop on Human Computation*.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A. C., Yamaguchi, K., Berg, T. L., Stratos, K., Daume, III, H., & III (2012). Midge: generating image descriptions from computer vision detections. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Nenkova, A., & Vanderwende, L. (2005). The impact of frequency on summarization. Tech. rep., Microsoft Research.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

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

Ortiz, L. M. G., Wolff, C., & Lapata, M. (2015). Learning to Interpret and Describe Abstract Scenes. In *Conference of the North American Chapter of the Association of Computational Linguistics*.

Panofsky, E. (1939). *Studies in Iconology*. Oxford University Press.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Parikh, D., & Grauman, K. (2011). Relative attributes. In *International Conference on Computer Vision*.

Park, C., & Kim, G. (2015). Expressing an image stream with a sequence of natural sentences. In *Advances in Neural Information Processing Systems*.

Patterson, G., Xu, C., Su, H., & Hays, J. (2014). The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision*, *108*(1-2), 59–81.

Pinheiro, P., Lebret, R., & Collobert, R. (2015). Simple image description generator via a linear phrase-based model. In *International Conference on Learning Representations Workshop*.

Prest, A., Schmid, C., & Ferrari, V. (2012). Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 601–614.

Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, *35*(4), 529–588.

Reiter, E., & Dale, R. (2006). *Building Natural Language Generation Systems*. Cambridge University Press.

Ren, M., Kiros, R., & Zemel, R. (2015). Image question answering: A visual semantic embedding model and a new dataset. In *International Conference on Machine Learningt Deep Learning Workshop*.

Richardson, M., Burges, C. J., & Renshaw, E. (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*.

Rohrbach, A., Rohrback, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In *International Conference on Computer Vision*.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013). Translating Video Content to Natural Language Descriptions. In *International Conference on Computer Vision*.

Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., & Manning, C. D. (2015). Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Conference on Empirical Methods in Natural Language Processing Vision and Language Workshop*.

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

Silberman, N., Kohli, P., Hoiem, D., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*.

Socher, R., & Fei-Fei, L. (2010). Connecting modalities: Semi-supervised segmentation and annotation of im- ages using unaligned text corpora. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*, *2*, 207–218.

Sun, C., Gan, C., & Nevatia, R. (2015). Automatic concept discovery from parallel text and visual corpora. In *International Conference on Computer Vision*.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., & Mooney, R. (2014). Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild. In *International Conference on Computational Linguistics*.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1958–1970.

Ushiku, Y., Yamaguchi, M., Mukuta, Y., & Harada, T. (2015). Common subspace for model and similarity: Phrase learning for caption generation from images. In *International Conference on Computer Vision*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Verma, Y., & Jawahar, C. V. (2014). Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval. In *British Machine Vision Conference*.

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

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.

Yagcioglu, S., Erdem, E., Erdem, A., & Cakici, R. (2015). A Distributed Representation Based Query Expansion Approach for Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Yang, Y., Teo, C. L., Daume, III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Conference on Empirical Methods in Natural Language Processing*.

Yao, B., & Fei-Fei, L. (2010). Grouplet: A structured image representation for recognizing human and object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In *International Conference on Computer Vision*.

Yatskar, M., Galley, M., Vanderwende, L., & Zettlemoyer, L. (2014). See No Evil, Say No Evil: Description Generation from Densely Labeled Images. In *Joint Conference on Lexical and Computation Semantics*.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, *2*, 67–78.

Yu, L., Park, E., Berg, A. C., & Berg, T. L. (2015). Visual madlibs: Fill in the blank description generation and question answering. In *International Conference on Computer Vision*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *International Conference on Computer Vision*.

Zitnick, C. L., Parikh, D., & Vanderwende, L. (2013). Learning the visual interpretation of sentences. In *International Conference on Computer Vision*.

Zitnick, C. L., & Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition*.