Noname manuscript No.
(will be inserted by the editor)

Large Scale Retrieval and Generation of Image Descriptions

Vicente Ordonez · Xufeng Han · Polina Kuznetsova · Girish Kulkarni · Margaret Mitchell · Kota Yamaguchi · Karl Stratos · Amit Goyal · Jesse Dodge · Alyssa Mensch · Hal Daumé III · Alexander C. Berg · Yejin Choi · Tamara L. Berg

Received: date / Accepted: date

Abstract What is the story of an image? What is the relationship between pictures, language, and information we can extract using state of the art computational recognition systems? In an attempt to address both of these questions, we explore methods for retrieving and generating natural language descriptions for images. Ideally, we would like our generated textual descriptions (captions) to both sound like a person wrote them, and also remain true to the image content. To do this we develop data-driven approaches for image description generation, using retrieval-based techniques to gather either: (a) whole captions associated with a visually similar image, or (b) relevant bits of text (phrases) from a large collection of image+description pairs. In the case of (b), we develop optimization algorithms to merge the retrieved phrases into valid natural language sentences. The end result is two simple, but effective, methods for harnessing the power of big data to produce image captions that are altogether more general, relevant, and human-like than previous attempts.

Keywords Retrieval · Image Description · Data Driven · Big Data · Natural Language Processing

1 Introduction

Our overarching goal is to better understand the complex relationship between images, computer vision, and the natural language people write to describe imagery. Successful mapping from photographs to natural language descriptions could have significant impacts on information retrieval, and failures can point toward future goals for computer vision. Studying collections of existing natural language descriptions of images and how to compose descriptions for novel queries will also help advance progress toward more complex visual recognition recognition goals, such as how to *tell the story behind an image*. These goals include determining the relative importance of content elements within an image and what factors people use to construct natural language to describe imagery [50], as well as tasks related to how people name content in images [41]. For example, in Figure 1, 2^{nd} photo from left, the user describes the girl, the dog, and their location, but selectively chooses not to describe the surrounding foliage and hut.

Producing a relevant and accurate caption for an arbitrary image is an extremely challenging problem because a system needs to not only estimate what image content is depicted, but also predict what a per-

V. Ordonez (🖂) · X. Han · A. C. Berg · T. L. Berg (🖂)
University of North Carolina, Chapel Hill, NC, USA
E-mail: vicente@cs.unc.edu,tlberg@cs.unc.edu

P. Kuznetsova · G. Kulkarni
Stony Brook University, Stony Brook, New York, USA

M. Mitchell
Microsoft Research, Redmond, Washington, USA

K. Yamaguchi
Tohoku University, Sendai, Japan

K. Stratos
Columbia University, New York, New York, USA

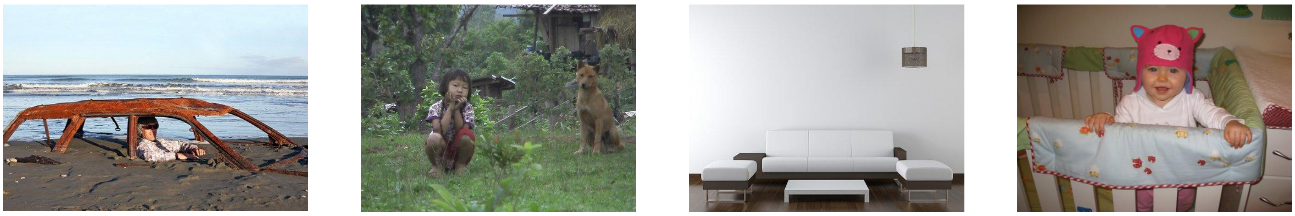
A. Goyal
Yahoo! Labs, Sunnyvale, California, USA

J. Dodge
Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

A. Mensch
University of Pennsylvania, Philadelphia, Pennsylvania, USA

H. Daumé III
University of Maryland, College Park, Maryland, USA

Y. Choi
University of Washington, Seattle, Washington, USA



Man sits in a rusted car buried in the sand on Waitarere beach

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

Interior design of modern white and brown living room furniture against white wall with a lamp hanging.

Emma in her hat looking super cute

Fig. 1 SBU Captioned Photo Dataset: Photographs with user-associated captions from our web-scale captioned photo collection. We collect a large number of photos from Flickr and filter them to produce a data collection containing over 1 million well captioned pictures.

son would describe about the image. However, there are already many images with relevant associated descriptive text available in the noisy vastness of the web. The key is to find the right images and make use of them in the right way! In this paper, we present two techniques to effectively skim the top of the image understanding problem to caption photographs by taking a *data driven* approach. To enable data driven approaches to image captioning we have collected a large pool of images with associated visually descriptive text. We develop retrieval algorithms to find good strings of text to describe an image, ultimately allowing us to produce natural-sounding and relevant captions for query images. These data-driven techniques follow in the footsteps of past work on internet-vision demonstrating that big data can often make challenging problems, see examples in image localization [21], retrieving photos with specific content [52], or image parsing [51], amenable to simple non-parametric matching methods.

A key potential advantage to making use of existing human-written image descriptions is that these captions may be more natural than those constructed directly from computer vision outputs using hand written rules. Furthermore we posit that many aspects of natural human-written image descriptions are difficult to produce directly from the output of computer vision systems, leading to unnatural sounding captions (see e.g. [25]). This is one of our main motivations for seeking to sample from existing descriptions of similar visual content. Humans make subtle choices about what to describe in an image, as well as how to form descriptions, based on image information that is not captured in, for instance, a set of object detectors or scene classifiers. In order to mimic some of these human choices, we carefully sample from descriptions people have written for images with some similar visual content, be it the pose of a human figure, the appearance of the sky, the scene layout, etc. In this way, we implicitly make use of human judgments of content importance and of some aspects of human composition during description generation. Another advantage of this type of method is that we can produce subtle and varied natural language for images without having to build models for every word in a vast visual vocabulary – by borrowing language based on visual similarity.

This paper develops and evaluates two methods to automatically map photographs to natural language descriptions. The first uses global image feature representations to retrieve and transfer whole captions from database images to a query image [42]. The second retrieves textual phrases from multiple visually similar database images, providing the building blocks, phrases, from which to construct novel and content-specific captions for a query image.

For the second method, finding descriptive phrases requires us to break the image down into constituent content elements, e.g. object detections (*e.g.*, person, car, horse, etc.) and coarse regions from image parsing (*e.g.*, grass, buildings, sky, etc.). We then retrieve visually similar instances of these objects and regions as well as similar scenes and whole images from a very large database of images with descriptions. Depending on what aspect of the image is being compared, we sample appropriate phrases from the descriptions. For example, a visual match to a similar sky might allow us to sample the prepositional phrase, "on a cloudless day." Once candidate phrases are retrieved based on matching similar image content, we evaluate several collective selection methods to examine and rerank the set of retrieved phrases. This reranking step promotes consistent content in the matching results up while pushing down outliers both in the image and language domain. In addition to intrinsic evaluation, the final set of reranked phrases are then evaluated in two applications. One tests the utility of the phrases for generating novel descriptive sentences. The second uses the phrases as features for text based image search.

Data-driven approaches to generation require a set of captioned photographs. Some small collections of captioned images have been created by hand in the past. The UIUC sentence data sets contain 1k [47] and 30k [57] images respectively each of which is associated with 5 human generated descriptions. The ImageClef[1] image retrieval challenge contains 20k images with associated human descriptions. Most of these collections are relatively small for retrieval based methods, as demonstrated by our experiments on captioning with varying collection size (Sec 4). Therefore, we have collected and released the SBU Captioned Photo Dataset [42] containing 1,000,000 Flickr images with natural language descriptions. This dataset was collected by performing a very large number of search queries to Flickr, and then heuristically filtered to find visually descriptive captions for images. The resulting dataset is large and varied, enabling effective matching of whole or local image content. The very large dataset also facilitates automatic tuning methods and evaluation that would not be possible on a dataset of only a few thousand captioned images. In addition this is the first – to our knowledge – attempt to mine the internet for general captioned images.

We perform extensive evaluation of our proposed methods, including evaluation of the sentences produced by our baseline and phrasa-based composition methods as well as evaluation of collective phrase selection and its application to text based image search. As these are relatively new and potentially subjective tasks, careful evaluation is important. We use a variety of techniques, from direct evaluation by people (using Amazon's Mechanical Turk) to indirect automatic measures like BLEU [43] and ROUGE [34] scores for similarity to ground truth phrases and descriptions. Note that none of these evaluation metrics are perfect for this task [25, 22]. Hopefully future research will develop better automatic methods for image description evaluation, as well as explore how descriptions should change as a function of task, e.g. to compose a description for image search vs image captioning for the visually impaired.

The reminder of the paper describes:

- A large data set containing images from the web with associated captions written by people, filtered so that the descriptions are likely to refer to visual content (Sec 3). This was previously published as part of [42].
- A description generation method that utilizes global image representations to retrieve and transfer captions from our data set to a query image (Sec 4). This was previously published as part of [42].
- New methods to utilize local image representations and collective selection to retrieve and rerank relevant phrases for images (Sec 5).
- New applications of phrase-based retrieval and reranking to: description generation (Sec 6.1), and complex query image search (Sec 6.2).
- New evaluations of our proposed image description methods, collective phrase selection algorithms, and image search prototype (Sec 7).

2 Related Work

Associating natural language with images is an emerging endeavor in computer vision. Some seminal work has looked at the task of mapping from images to text as a translation problem (similar to translating between two languages) [14]. Other work has tried to estimate correspondences between keywords and image regions [2], or faces and names [3,4]. In a parallel research goal, recent work has started to move beyond recognition of leaf-level object category terms toward mid-level elements such as attributes [5,15,19,26,29], or hierarchical representations of objects [10,12,13].

Image description generation in particular has been studied in recent papers [16,18,22,25,27,31,38,42,56, 36,20]. Some approaches [25,31,55], generate descriptive text from scratch based on detected elements such as objects, attributes, and prepositional relationships. This results in descriptions for images that are sometimes closely related to image content, but that are also often quite verbose, non-human-like, or lacking in creativity. Other techniques for producing descriptive image text, e.g. [56], require a human in the loop for image parsing (except in specialized circumstances) and various hierarchical knowledge ontologies. The recent work of Hodosh *et al* [22] argues in favor of posing the image-level sentence annotation task as a sentence ranking problem, where performance is measured by the rank of the ground truth caption, but does not allow for composing new language for images.

Other attempts to generate natural language descriptions for images have made use of pre-associated text or other meta-data. For example, Feng and Lapata [18] generate captions for images using extractive and abstractive generation methods, but assume relevant documents are provided as input. Aker *et al.* [1] rely on GPS meta data to access relevant text documents.

The approaches most relevant to this paper make use of existing text for caption generation. In Farhadi *et al.* [16], the authors produce image descriptions via a retrieval method, by translating both images and text descriptions to a shared meaning space represented by a single $< object, action, scene >$ tuple. A description for a query image is produced by retrieving whole image descriptions via this meaning space from a set of image

[1] http://www.imageclef.org/2011

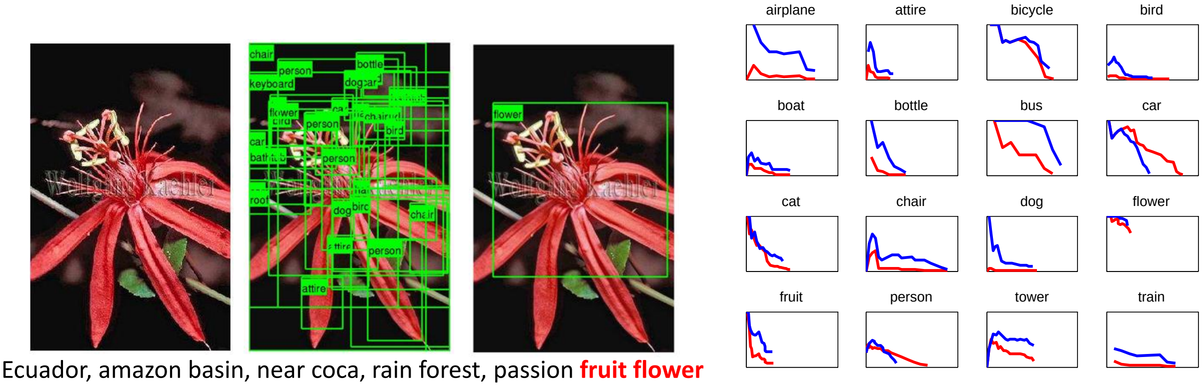


Fig. 2 Left: Blindly running many object detectors on an image produces very noisy results. Running object detectors mentioned in a caption can produce much cleaner results. **Right:** Improvement in detection is measured with precision-recall (red shows raw detector performance, blue shows caption triggered). For some categories (e.g., airplane, dog) performance is greatly improved, for others not as much (e.g., cat, chair).

descriptions (the UIUC Pascal Sentence data set [47]). This results in descriptions that sound very human – since they were written by people – but which may not be relevant to the specific image content. This limited relevancy often occurs because of problems of sparsity, both in the data collection – 1000 images is too few to guarantee similar image matches – and in the representation – only a few categories for 3 types of image content are considered.

In contrast, we attack the caption generation problem for more general images (images found via thousands of paired-word Flickr queries) and a larger set of object categories (89 vs 20). In addition to extending the object category list considered, we also include a wider variety of image content aspects in our search terms, including: non-part-based region categorization, attributes of objects, activities of people, and a larger number of common scene classes. We also generate our descriptions via an extractive method with access to a much larger and more general set of captioned photographs from the web (1 million vs 1 thousand).

Compared to past retrieval based generation approaches such as Farhadi *et al.* [16] and our work [42], which retrieve whole *existing captions* to describe a query image, here we develop algorithms to associated bits of text (phrases) with parts of an image (e.g. objects, regions, or scenes). As a product of our phrase retrieval process, we also show how to use our retrieved phrases (retrieved from multiple images) to compose *novel captions*, and to perform *complex query retrieval.* Since images are varied, the likelihood of being able to retrieve a complete yet relevant caption is low. Utilizing bits of text (*e.g.*, phrases) allows us to directly associate text with part of an image. This results in better, more relevant and more specific captions if we apply our phrases to caption generation. A key subroutine in the process is reranking the retrieved phrases in order to produce a shortlist for the more computationally expensive optimization for description generation, or for use in complex query retrieval. In this paper we explore two techniques for performing this reranking collectively – taking into account the set of retrieved phrases. Our reranking approaches have close ties to work in information retrieval including PageRank [24] and TFIDF [48].

Producing a relevant and human-like caption for an image is a decidedly subtle task. As previously mentioned, people make distinctive choices about what aspects of an image's content to include or not include in their description. This link between visual importance and descriptions, studied in Berg *et al.* [50], leads naturally to the problem of text summarization in natural language processing. In text summarization, the goal is to produce a summary for a document that describes the most important content contained in the text. Some of the most common and effective methods proposed for summarization rely on extractive summarization [32, 37, 39, 46, 53] where the most important or relevant text is selected from a document to serve as the document's summary. Often a variety of features related to document content [39], surface [46], events [32] or feature combinations [53] are used in the selection process to compose sentences that reflect the most significant concepts in the document. Our retrieval based description generation methods can be seen as instances of extractive summarization because we make use of existing text associated with (visually similar) images.

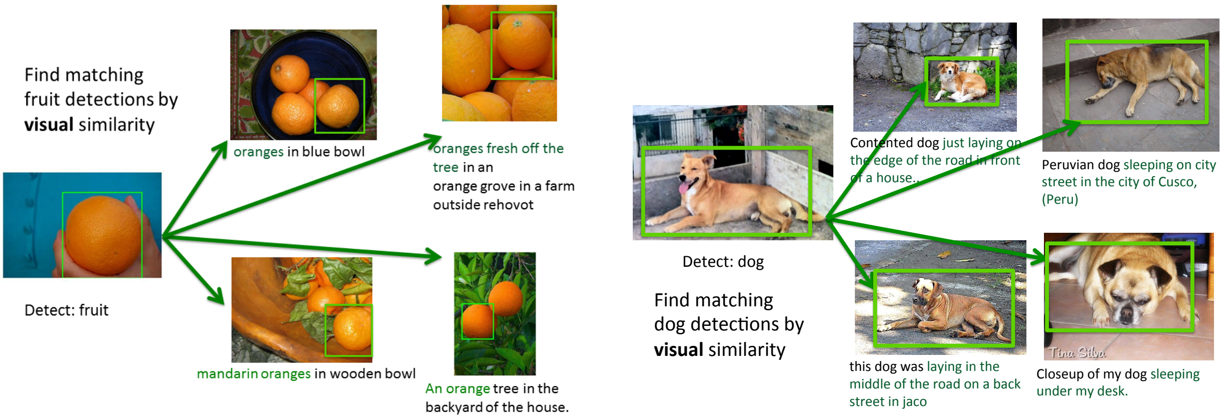

Find matching
fruit detections by
visual similarity
Detect: fruit
oranges in blue bowl
oranges fresh off the tree in an orange grove in a farm outside rehovot
mandarin oranges in wooden bowl
An orange tree in the backyard of the house.
Detect: dog
Find matching
dog detections by
visual similarity
Contented dog just laying on the edge of the road in front of a house..
Peruvian dog sleeping on city street in the city of Cusco, (Peru)
this dog was laying in the middle of the road on a back street in jaco
Closeup of my dog sleeping under my desk.
Tina Silva

Fig. 3 **Left:** For a query "fruit" detection, we retrieve similar looking "fruit" detections (including synonyms or holonyms) from the database and transfer the referring noun-phrase (NP). **Right:** For a query "dog" detection, we retrieve similar looking "dog" detections (including synonyms or holonyms) from the database and transfer the referring verb-phrase (VP).

3 Web-Scale Captioned Image Collection

One key requirement of this work is a web-scale database of photographs with associated descriptive text. To enable effective captioning of novel images, this database must satisfy two general requirements: 1) It must be large so that visual matches to the query are reasonably similar, 2) The captions associated with the database photographs must be visually relevant so that transferring captions between pictures driven by visual similarity is useful. To achieve the first requirement we queried Flickr using a huge number of pairs of query terms (objects, attributes, actions, stuff, and scenes). This produced a very large, but noisy initial set of photographs with associated text (hundreds of millions of images). To achieve our second requirement we filtered this set so that the descriptions attached to a picture are likely to be relevant and visually descriptive. To encourage visual descriptiveness, we select only those images with descriptions of satisfactory length, based on observed lengths in visual descriptions. We also enforce that retained descriptions contain at least 2 words belonging to our term lists and at least one prepositional word, e.g. "on", "under" which often indicate visible spatial relationships.

This resulted in a final collection of over 1 million images with associated text descriptions – the *SBU Captioned Photo Dataset*. These text descriptions generally function in a similar manner to image captions, and usually directly refer to some aspects of the visual image content (see Fig 1 for examples).

To evaluate whether the captions produced by our automatic filtering are indeed relevant to their associated images, we performed a forced-choice evaluation task, where a user is presented with two photographs and one caption. The user must assign the caption to the most relevant image (care is taken to remove biases due to temporal or left-right placement in the task). In this case we present the user with the original image associated with the caption and a random image. We perform this evaluation on 100 images from our web-collection using Amazon's Mechanical Turk service, and find that users are able to select the ground truth image *96%* of the time. This demonstrates that the task is reasonable and that descriptions from our collection tend to be fairly visually specific and relevant. One possible additional pre-processing step for our dataset would be to use sentence compression by eliminating overly specific information as described in our previous work [28].

4 Global Generation of Image Descriptions

Past work has demonstrated that if your data set is large enough, some very challenging problems can be attacked with simple matching methods [21,52,51]. In this spirit, we harness the power of web photo collections in a non-parametric approach. Given a query image, I_q, our goal is to generate a relevant description. In our first baseline approach, we achieve this by computing the global similarity of a query image to our large web-collection of captioned images. We find the closest matching image (or images) and simply transfer over the description from the matching image to the query image.

For measuring visual similarity we utilize two image descriptors. The first is the well known gist feature, a global image descriptor related to perceptual dimensions – naturalness, roughness, ruggedness etc – of scenes [40]. The second descriptor is also a global image descriptor, computed by resizing the image into a "tiny image", essentially a thumbnail of size 32x32.

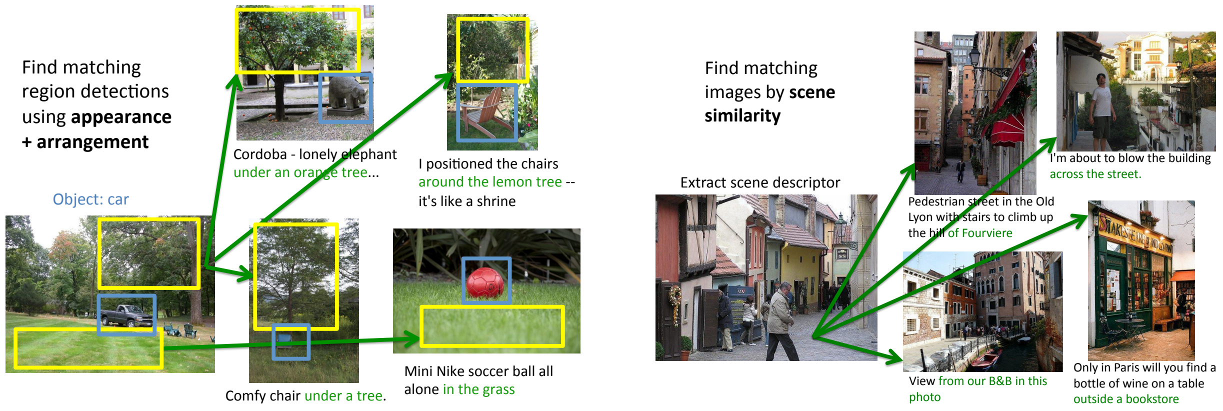

Find matching region detections using appearance + arrangement
Object: car
Cordoba - lonely elephant under an orange tree...
I positioned the chairs around the lemon tree -- it's like a shrine
Comfy chair under a tree.
Mini Nike soccer ball all alone in the grass
Find matching images by scene similarity
Extract scene descriptor
Pedestrian street in the Old Lyon with stairs to climb up the hill of Fourviere
I'm about to blow the building across the street.
View from our B&B in this photo
Only in Paris will you find a bottle of wine on a table outside a bookstore

Fig. 4 Left: For query object-stuff detection pairs, e.g.,"car" and "tree," we retrieve relevant object-stuff detections from the database using visual and geometric configuration similarity (where the database match can be e.g., "any object" and "tree" pair) and transfer the referring prepositional-phrase (PP). **Right:** We use whole image scene classification descriptors to transfer contextual scene prepositional-phrases (PPs).

This helps us match not only scene structure, but also the overall color of images. To find visually relevant images we compute the similarity of the query image to images in the captioned photo dataset using a sum of gist similarity and tiny image color similarity (equally weighted).

5 Retrieving and Reranking Phrases Describing Local Image Content

In this section we present methods to retrieve natural language phrases describing local and global image content from our large database of captioned photographs. Because we want to retrieve phrases referring to specific objects, relationships between objects and their background, or to the general scene, a large amount of image and text processing is first performed on the collected database (Sec 5.1). This allows us to extract useful and accurate estimates of local image content as well as the phrases that refer to that content. For a novel query image, we can then use visual similarity measures to retrieve sets of relevant phrases describing image content (Sec 5.2). Finally, we use collective reranking methods to select the most relevant phrases for the query image (Sec 5.3).

5.1 Dataset Processing

We perform 4 types of dataset processing: object detection, rough image parsing to obtain background elements, scene classification, and caption parsing. This provides textual phrases describing both local (e.g. objects and local object context) and global (e.g. general scene context) image content.

Object detection: We extract object category detections using deformable part models [17] for 89 common object categories [33,42]. Of course, running tens or hundreds of object detectors on an image would produce extremely noisy results (e.g., Fig 2). Instead, we place priors on image content – by only running detectors for objects (or their synonyms and hyponyms, e.g., Chihuahua for dog) mentioned in the caption associated with a database image. This produces *much cleaner results* than blindly running all object detectors. Though captions can provide a semi-weak annotation signal (e.g. an image captioned "A horse outside my car window" probably does not depict a car), we are able to obtain a fairly accurate pool of object localizations with associated text phrases without requiring a fully annotated dataset. Figure 2 shows precision-recall curves for raw detectors in red and caption triggered detectors in blue for 1000 images from the SBU Dataset covering a balanced number of categories with hand labeled bounding box annotations for evaluation. Detection performance is greatly improved for some categories (e.g., bus, airplane, dog), and less improved for others (e.g. cat, person). From the million photo database we obtain a large pool of (up to 20k) high scoring object detections for each object category.

Image parsing: Image parsing is used to estimate regions of background elements in each database image. Six categories are considered: sky, water, grass, road, tree, and building, using detectors [42] which compute color, texton, HoG [9] and Geometric Context [23] as input features to a sliding window based SVM classifier. These detectors are run on all database images.

Scene Classification: The scene descriptor for each image consists of the outputs of classifiers for 26 common scene categories. The features, classification method,

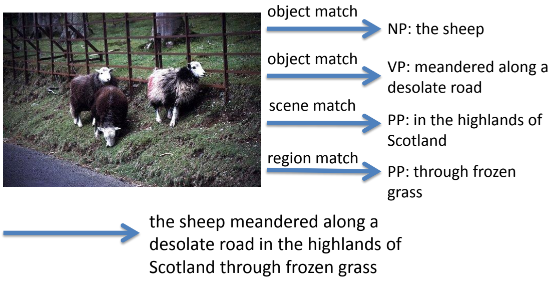

object match
NP: the sheep
object match
VP: meandered along a desolate road
scene match
PP: in the highlands of Scotland
region match
PP: through frozen grass
the sheep meandered along a desolate road in the highlands of Scotland through frozen grass

Fig. 5 For a query image, we take a data-driven approach to retrieve (and optionally rerank) a set of visually relevant phrases based on local and global image content estimates. We can then construct an image caption for the query using phrasal description generation. Our optimization approach to generation maximizes both visual similarity and language-model estimates of sentence coherence. This produces captions that are more relevant, and human-sounding than previous approaches.

and training data are from the SUN dataset [54]. This descriptor is useful for capturing and matching overall global scene appearance for a wide range of scene types. Scene descriptors are computed on 700,000 images from the database to obtain a large pool of scene descriptors for retrieval.

Caption Parsing: The Berkeley PCFG parser [44,45] is used to obtain a hierarchical parse tree for each caption. From this tree we gather constituent phrases, (e.g., noun phrases, verb phrases, and prepositional phrases) referring to each of the above kinds of image content in the database.

5.2 Retrieving Phrases

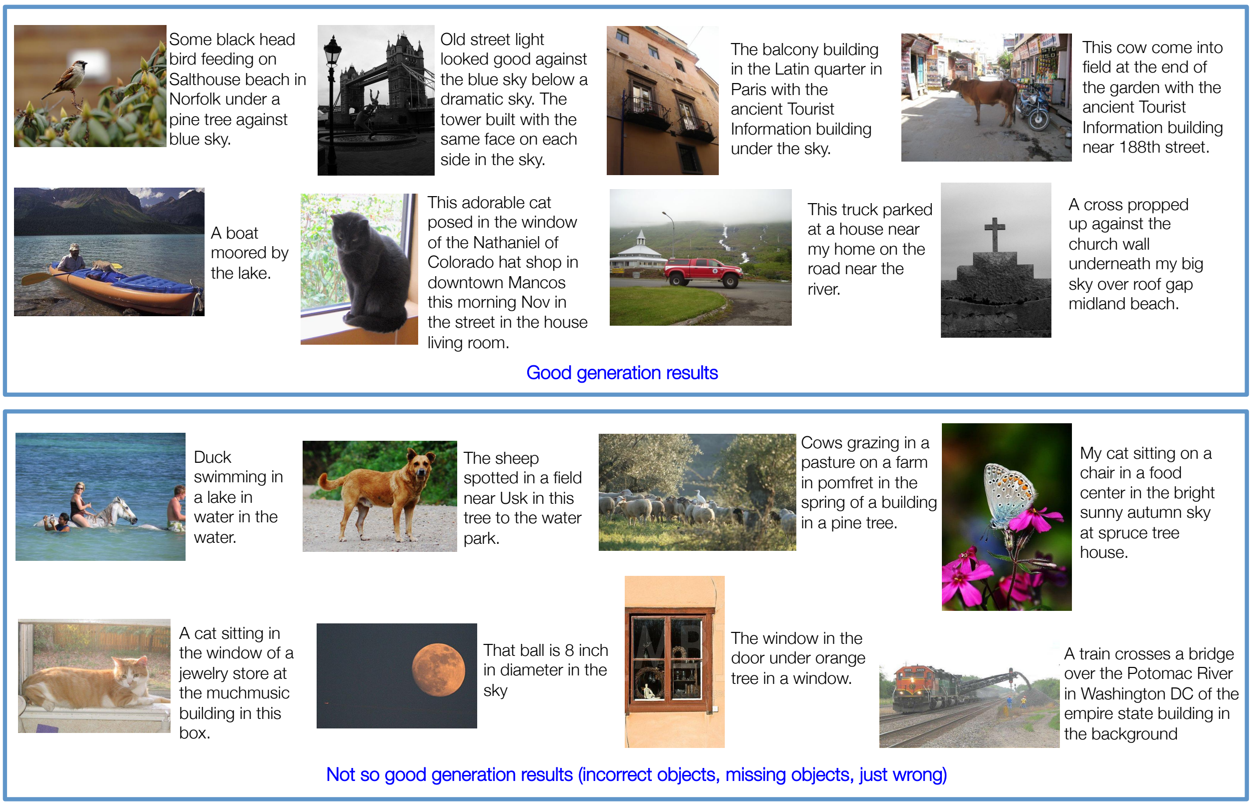
For a query image, we retrieve several types of relevant phrases: noun-phrases (NPs), verb-phrases (VPs), and prepositional-phrases (PPs). Five different features are used to measure visual similarity: **Color** – LAB histogram, **Texture** – histogram of vector quantized responses to a filter bank [30], **SIFT Shape** – histogram of vector quantized dense SIFT descriptors [35], **HoG Shape** – histogram of vector quantized densely computed HoG descriptors [9], **Scene** – vector of classification scores for 26 common scene categories. The first 4 features are computed locally within an (object or stuff) region of interest and the last feature is computed globally.

Retrieving Noun-Phrases (NPs): For each proposed object detection in a query image, we retrieve a set of relevant noun-phrases from the database. For example, if "fruit" is detected in the query, then we retrieve NPs from database image captions with visually similar "fruit" detections (including synonyms or holonyms, e.g. "apples" or "oranges"). This process is illustrated in Fig 3, left, where a query fruit detection is matched to visually similar database fruit detections (and their referring NPs in green). Visual similarity is computed as an unweighted combination of color, texture, SIFT, and HoG similarity, and produces visually similar and conceptually relevant NPs for a query object.

Retrieving Verb-Phrases (VPs): For each proposed object detection in a query image, we retrieve a set of relevant verb-phrases from the database. Here we associate VPs in database captions to object detections in their corresponding database images if the detection category (or a synonym or holonym) is the head word in an NP from the same sentence (e.g. in Fig 3 bottom right dog picture, "sleeping under my desk" is associated with the dog detection in that picture). Our measure of visual similarity is again based on equally weighted combination of color, texton, SIFT and HoG feature similarities. As demonstrated in Fig 3 (left), this measure often captures similarity in pose.

Retrieving Image parsing-based Prepositional-Phrases (PPStuff): For each proposed object detection and for each background element detection in a query image (e.g. sky or road), we retrieve relevant PPs according to visual and spatial relationship similarity (illustrated on the left in Fig 4 for car plus tree and grass detections). Visual similarity between a background query region and background database regions is computed based on color, texton, and SIFT co-sine similarity. Spatial relationship similarity is computed based on the similarity in geometric configuration between the query object-background pair and object-background pairs observed in the database (where the object in the database pairs need not be the same object category as the query). This spatial relationship is measured in terms of the normalized distance between the foreground object and the background region, the normalized overlap area between the foreground object and the background region, and the absolute vertical position of the foreground object. Visual similarity and geometric similarity measures are given equal weights and produce appealing results (Fig 4).

Retrieving Scene-based Prepositional-Phrases (PPScene): For a query image, we retrieve PPs referring to the overall setting or scene by finding the most similar global scene descriptors from the database. Here we retrieve the last PP in a sentence since it is most likely to describe the scene content. As shown on the right in Fig 4, useful matched phrases often correspond to places (e.g., "in Paris") or general scene context (e.g., "at the beach").



Some black head bird feeding on Salthouse beach in Norfolk under a pine tree against blue sky.

Old street light looked good against the blue sky below a dramatic sky. The tower built with the same face on each side in the sky.

The balcony building in the Latin quarter in Paris with the ancient Tourist Information building under the sky.

This cow come into field at the end of the garden with the ancient Tourist Information building near 188th street.

A boat moored by the lake.

This adorable cat posed in the window of the Nathaniel of Colorado hat shop in downtown Mancos this morning Nov in the street in the house living room.

This truck parked at a house near my home on the road near the river.

A cross propped up against the church wall underneath my big sky over roof gap midland beach.

Duck swimming in a lake in water in the water.

The sheep spotted in a field near Usk in this tree to the water park.

Cows grazing in a pasture on a farm in pomfret in the spring of a building in a pine tree.

My cat sitting on a chair in a food center in the bright sunny autumn sky at spruce tree house.

A cat sitting in the window of a jewelry store at the muchmusic building in this box.

That ball is 8 inch in diameter in the sky

The window in the door under orange tree in a window.

A train crosses a bridge over the Potomac River in Washington DC of the empire state building in the background

Not so good generation results (incorrect objects, missing objects, just wrong)

Fig. 6 Using our retrieved, reranked phrases for description generation (Sec 6.1). Reasonably good results are shown on top and less good results (with incorrect objects, missing objects, or just plain wrong descriptions) are shown at the bottom.

5.3 Reranking Phrases

Given a set of retrieved phrases for a query image, we would like to rerank these phrases using collective measures computed on the entire set of retrieved results. Related reranking strategies have been used for other retrieval systems. Sivic and Zisserman[49] retrieve images using visual words and then rerank them based on a measure of geometry and spatial consistency. Torralba *et al.*[52] retrieve a set of images using a reduced representation of their feature space and then perform a second refined reranking phase on top matching images to produce exact neighbors.

In our case, instead of reranking images, our goal is to rerank retrieved phrases such that the relevance of the top retrieved phrases is increased. Because each phrase is retrieved independently in the phrase retrieval step, the results tend to be quite noisy. Spurious image matches can easily produce irrelevant phrases. The wide variety of Flickr users and contexts under which they capture their photos can also produce unusual or irrelevant phrases.

As an intuitive example, if one retrieved phrase describes a dog as “the brown dog” then the dog *may* be brown. However, if several retrieved phrases describe the dog in similar ways, e.g., “the little brown dog”, “my brownish pup”, “a brown and white mutt”, then it is much more likely that the query dog is brown and the predicted relevance for phrases describing brown attributes should be increased.

In particular, for each type of retrieved phrase (see Sec 5.2), we gather the top 100 best matches based on visual similarity. Then, we perform phrase reranking to select the best and most relevant phrases for an image (or part of an image in the case of objects or regions). We evaluate two possible methods for reranking: 1) PageRank based reranking using visual and/or text similarity, 2) Phrase-level TFIDF based reranking.

5.3.1 PageRank Reranking

PageRank [7] computes a measure for the relative importance of items within a set based on the random walk probability of visiting each item. The algorithm was originally proposed as a measure of importance for web pages using hyperlinks as connections between pages [7], but has also been applied to other tasks such as reranking images for product search [24]. For our task, we use PageRank to compute the relative importance of phrases within a retrieved set on the premise

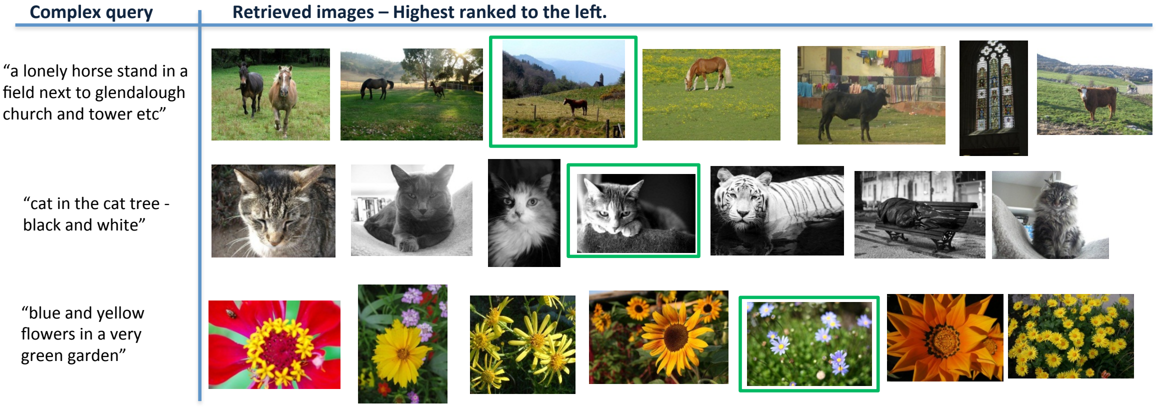

Complex query
Retrieved images – Highest ranked to the left.
"a lonely horse stand in a field next to glendalough church and tower etc"
"cat in the cat tree - black and white"
"blue and yellow flowers in a very green garden"

Fig. 7 Complex query image retrieval. For a complex natural language text query (left), we retrieve images displaying relevant content (right). The image originally associated with the complex text query is highlighted in green.

that phrases displaying strong similarity to other phrases within the retrieved set are more likely to be relevant to the query image.

We construct 4 graphs, one for each type of retrieved phrase (NP, VP, PPStuff, or PPScene), from the set of retrieved phrases for that type. Nodes in these graphs correspond to retrieved phrases (and the corresponding object, region, or image each phrase described in the SBU database). Edges between nodes are weighted using visual similarity, textual similarity, or an unweighted combination of the two – denoted as Visual PageRank, Text PageRank, or Visual + Text PageRank respectively. Text similarity is computed as the cosine similarity between phrases, where phrases are represented as a bag of words with a vocabulary size of approximately 100k words, weighted by term-frequency inverse-document frequency (TFIDF) score [48]. Here IDF measures are computed for each phrase type independently rather than over the entire corpus of phrases to produce IDF measures that are more type specific. Visual similarity is computed as cosine similarity of the visual representations used for retrieval (Sec 5.2).

For generating complete image descriptions (Sec 6.1), the PageRank score can be directly used as a unary potential for phrase confidence.

5.3.2 Phrase-level TFIDF Reranking

We would like to produce phrases for an image that are not only relevant, but specific to the particular depicted image content. For example, if we have a picture of a cow a phrase like "the cow" is always going to be relevant to any picture of a cow. However, if the cow is mottled with black and white patches then "the spotted cow" is a much better description for this specific example. If both of these phrases are retrieved for the image, then we would prefer to select the latter over the former.

To produce phrases with high description specificity, we define a phrase-level measure of TFIDF. This measure rewards phrases containing words that occur frequently within the retrieved phrase set, but infrequently within a larger set of phrases – therefore giving higher weight to phrases that are specific to the query image content (e.g., "spotted"). For object and stuff region related phrases (NPs, VPs, PPStuff), IDF is computed over phrases referring to that object or stuff category (e.g., the frequency of words occurring in a noun phrase with "cow" in the example above). For whole image related phrases (PPScene), IDF is computed over all prepositional phrases. To compute TFIDF for a phrase, the TFIDF for each word in the phrase is calculated (after removing stop words) and then averaged. Other work that has used TFIDF for image features (we use it for text associated with an image) include Sivic and Zisserman [49], Chum *et al.* [8], and Ordonez *et al.* [42].

For composing image descriptions (Sec 6.1), we use phrase-level TFIDF to rerank phrases and select the top 10 phrases. The original visual retrieval score (Sec 5.2) is used as the phrase confidence score, effectively merging ideas of visual relevance with phrase specificity (denoted as Visual + TFIDF).

6 Applications of Phrases

Once we have retrieved (and reranked) phrases related to an image we can use the associated phrases in a number of applications. Here we demonstrate two potential applications: phrasal generation of image descriptions (Sec 6.1), and complex query image search (Sec 6.2).

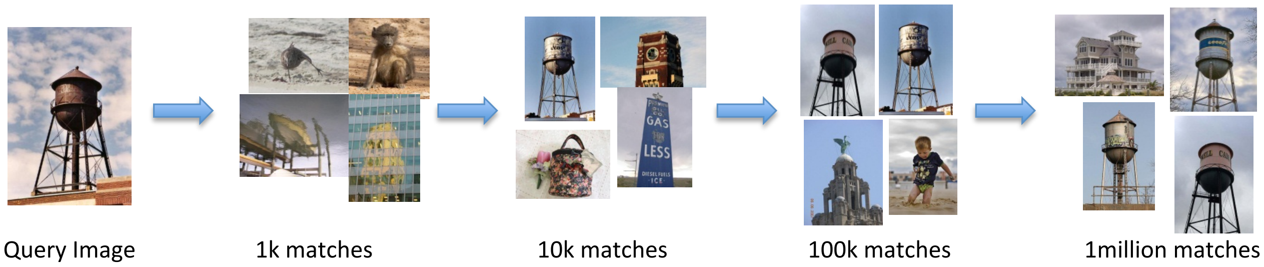


Fig. 8 **Size Matters:** Example matches to a query image for varying data set sizes.

6.1 Phrasal Generation of Image Descriptions

We model caption generation as an optimization problem in order to incorporate two different types of information: the confidence score of each retrieved phrase provided by the original retrieval algorithm (Sec 5.2) or by our reranking techniques (Sec 5.3), and additional pairwise compatibility scores across phrases computed using observed language statistics. Our objective is to select a set of phrases that are visually relevant to the image and that together form a reasonable sentence, which we measure by compatibility across phrase boundaries.

Let $X = \{x_{obj},\ x_{verb},\ x_{stuff},\ x_{scene}\}$ be a candidate set of phrases selected for caption generation. We maximize the following objective over possibilities for X:

$$E(X) = \Phi(X) + \Psi(X) \tag{1}$$

Where $\Phi(X)$ aggregates the unary potentials measuring quality of the individual phrases:

$$\Phi(X) = \phi(x_{obj}) + \phi(x_{verb}) + \phi(x_{stuff}) + \phi(x_{scene}) \tag{2}$$

And $\Psi(X)$ aggregates binary potentials measuring pairwise compatibility between phrases:

$$\Psi(X) = \psi(x_{obj}, x_{verb}) + \psi(x_{verb}, x_{stuff}) + \psi(x_{stuff}, x_{scene}) \tag{3}$$

Unary potentials, $\phi(x)$, are computed as the confidence score of phrase x determined by the retrieval and reranking techniques discussed in Sec 5.3. To make scores across different types of phrases comparable, we normalize them using Z-score (subtract mean and divide by standard deviation). We further transform the scores so that they fall in the [0,1] range.

Binary potentials: N-gram statistics are used to compute language naturalness – a frequent n-gram denotes a commonly used, “natural”, sequence of words. In particular, we use n-gram frequencies provided by the Google Web 1-T dataset [6], which includes frequences up to 5-grams with counts computed from text on the web. We use these counts in the form of normalized point-wise mutual information scores to incorporate language-driven compatibility scores across different types of retrieved phrases. The compatibility score $\psi(x_i, x_j)$ between a pair of adjacent phrases x_i and x_j is defined as follows: $\psi(x_i, x_j) = \alpha \cdot \psi^{\rm L}_{ij} + (1-\alpha) \cdot \psi^{\rm G}_{ij}$. Where ψ^{L}_{ij} and ψ^{G}_{ij} are the local and the global cohesion scores defined below.[2]

Local Cohesion Score: Let L_{ij} be the set of all possible n-grams ($2 \leq n \leq 5$) across the boundary of x_i and x_j. Then we define the n-gram local cohesion score as:

$$\psi^{\rm L}_{ij} = \frac{\sum\limits_{l \in L_{ij}} \mathrm{NPMI}(l)}{\|L_{ij}\|} \tag{4}$$

Where $\mathrm{NPMI}(v) = (\mathrm{PMI}(v) - a)/(b - a)$ is a normalized point-wise mutual information (PMI) score where a and b are normalizing constants computed across n-grams so that the range of $\mathrm{NPMI}(v)$ is between 0 and 1. This term encourages smooth transitions between consecutive phrases. For instance the phrase “The kid on the chair” will fit better preceding “sits waiting for his meal” than “sleeps comfortably”. This is because the words at the end of the first phrase including “chair” are more compatible with the word“sit” at the beginning of the second phrase than with the word “sleep” at the beginnining of the third phrase.

Global Cohesion Score: These local scores alone are not sufficient to capture semantic cohesion across very long phrases, because Google n-gram statistics are limited to 5 word sequences. Therefore, we also consider compatibility scores between the head word of each phrase, where the head word corresponds semantically to the most important word in a given phrase (last word or main verb of the phrase). For instance the phrase “The phone in the hall” is more compatible with the phrase “rings loudly all the time” than with the phrase “thinks about philosophy everyday” because the head word “phone” is more compatible with the head word “rings” than with the head word “thinks”. Let h_i and

[2] The coefficient α can be tuned via grid search, and scores are normalized $\in [0, 1]$.

h_j be the head words of phrases x_i and x_i respectively, and let $f_\Sigma(h_i, h_j)$ be the total frequency of all n-grams that start with h_i and end with h_j. Then the global cohesion is computed as:

$$\psi^{\mathrm{G}}_{ij} = \frac{f_\Sigma(h_i, h_j) - \min(f_\Sigma)}{\max(f_\Sigma) - \min(f_\Sigma)} \quad (5)$$

Inference by Viterbi decoding: Notice that the potential functions in the objective function (Equations 1 & 3) have a linear chain structure. Therefore, we can find the argmax, $X = \{x_{obj}, x_{verb},\ x_{stuff}, x_{scene}\}$, efficiently using Viterbi decoding.[3]

6.2 Complex Query Image Search

Image retrieval is beginning to work well. Commercial companies like Google and Bing produce quite reasonable results now for simple image search queries, like “dog” or “red car”. Where image search still has much room for improvement is for complex search queries involving appearance attributes, actions, multiple objects with spatial relationships, or interactions. This is especially true for more unusual situations that cannot be mined directly by looking at the meta-data and text surrounding an image, *e.g.*, “little boy eating his brussels sprouts”.

We demonstrate a prototype application, showing that our approach for finding descriptive phrases for an image can be used to form features that are useful for complex query image retrieval. We use 1000 test images (described in Sec 7) as a dataset. For each image, we pick the top selected phrases from the vision+text PageRank algorithm to use as a complex text descriptor for that image – note that the actual human-written caption for the image is not seen by the system. For evaluation we then use the original human caption for an image as a complex query string. We compare it to each of the automatically derived phrases for images in the dataset and score the matches using normalized correlation. For each matching image we average those scores for each retrieved phrase. We then sort the scores and record the rank of the correct image – the one for which the query caption was written. If the retrieved phrases matched the actual human captions well, then we expect the query image to be returned first in the retrieved images. Otherwise, it will be returned later in the ranking. Note that this is only a demo application performed on a very small dataset of images. A real image retrieval application would have access to billions of images.

Method	BLEU
Global Description Generation (1k)	0.0774 +- 0.0059
Global Description Generation (10k)	0.0909 +- 0.0070
Global Description Generation (100k)	0.0917 +- 0.0101
Global Description Generation (1million)	0.1177 +- 0.0099

Table 1 Global Matching Performance with respect to data set size (BLEU score measured at 1)

7 Evaluation

We perform experimental evaluations on each aspect of the proposed approaches: global description generation (Sec 7.1), phrase retrieval and reranking (Sec 7.2), phrase based description generation (Sec 7.3), and phrase based complex query image search (Sec 7.4).

To evaluate global generation, we randomly sample 500 images from our collection. As is usually the case with web photos, the photos in this set display a wide range of difficulty for visual recognition algorithms and captioning, from images that depict scenes (e.g. beaches), to images with relatively simple depictions (e.g. a horse in a field), to images with much more complex depictions (e.g. a boy handing out food to a group of people). For all phrase based evaluations (except where explicitly noted) we use a test set of 1000 query images, selected to have high detector confidence scores. Random test images could also be sampled, but for images with poor detector performance we expect the results to be much the same as for our baseline global generation methods. Therefore, we focus on evaluating performance for images where detection is more likely to have produced reasonable estimates of local image content.

7.1 Global Generation Evaluation

Results – Size Matters! Our global caption generation method often performs surprisingly well. As reflected in past work [21, 52] image retrieval from small collections often produces spurious matches. This can be seen in Fig 8 where increasing data set size has a significant effect on the quality of retrieved global matches and their corresponding transferred caption relevance. Quantitative results also reflect this observation. As shown in Table 1 data set size has a significant effect on automatic measures of caption quality, specifically on BLEU score; more data provides more similar and relevant matched images (and captions). BLEU scores are a measure of precision on the number of n-grams matched of a given candidate text against a set of reference ground truth texts. For our task we use BLEU at 1, meausuring uni-gram performance. This measure also incorporates a penalty on the length of the candidate text.

[3] An interesting but non-trivial extension to this generation technique is allowing re-ordering or omission of phrases [27].

Method	Noun Phrases $K = 1, 5, 10$	Verb Phrases $K = 1, 5, 10$	Prepositional Phrases(stuff) $K = 1, 5, 10$	Prepositional Phrases(scenes) $K = 1, 5, 10$
No reranking	$0.24, 0.24, 0.23$	$0.15, 0.14, 0.14$	$0.30, 0.29, 0.27$	$0.28, 0.26, 0.25$
Visual PageRank	$0.23, 0.23, 0.23$	$0.13, 0.14, 0.14$	$0.28, 0.28, 0.27$	$0.26, 0.25, 0.25$
Text PageRank	$0.30, 0.29, 0.28$	$0.20, 0.19, 0.17$	$0.38, 0.37, 0.36$	$0.34, 0.30, 0.27$
Visual+Text PageRank	$0.28, 0.27, 0.26$	$0.17, 0.17, 0.16$	$0.32, 0.30, 0.28$	$0.27, 0.28, 0.27$
TFIDF Reranking	$0.29, 0.28, 0.27$	$0.19, 0.19, 0.18$	$0.38, 0.37, 0.36$	$0.40, 0.36, 0.32$

Table 2 Average BLEU@1 score for the top K retrieved phrases against Flickr captions.

Method	Noun Phrases	Verb Phrases	Prepositional Phrases(stuff)	Prepositional phrases(scenes)
No reranking	0.2633	0.0759	0.1458	0.1275
Visual PageRank	0.2644	0.0754	0.1432	0.1214
Text PageRank	0.3286	0.1027	0.1862	0.1642
Visual + Text PageRank	0.2262	0.0938	0.1536	0.1631
TFIDF Reranking	0.3143	0.1040	0.2096	0.1912

Table 3 Average BLEU@1 score evaluation K=10 against MTurk written descriptions.

7.2 Phrase Retrieval & Reranking Evaluation

We calculate BLEU scores (without length penalty) for evaluating the retrieved phrases against the original human associated captions from the SBU Dataset [42]. Scores are evaluated for the top K phrases for $K = 1, 5, 10$ for each phrase type in Table 2. We can see that except for Visual PageRank all other reranking strategies yield better BLEU scores than the original (unranked) retrieved phrases. Overall, Text PageRank and TFIDF Reranking provide the best scores.

One possible weakness in this initial evaluation is that we use a single caption as reference – the captions provided by the owners of the photos – which often include contextual information unrelated to visual content. To alleviate this effect we collect 4 additional human written descriptions using Amazon Mechanical Turk for a subset of 200 images from our test set (care was taken to ensure workers were located in the US and filtered for quality control). In this way we obtain good quality sentences referring to the image content, but we also notice some biases like rich noun-phrases while very few verb-phrases within those sentences. Results are provided in Table 3, further supporting our previous observations. TFIDF and Text PageRank demonstrate the most increase in BLEU score performance over the original retrieved ranking.

7.3 Application 1: Description Generation Evaluation

We can also evaluate the quality of our retrieved set of phrases indirectly by using them in an application to compose novel full image descriptions (Sec 6.1). Automatic evaluation is computed using BLEU score [43] (including length penalty), and we additionally compute ROUGE scores [34] (analogous to BLEU scores, ROUGE scores are a measure of recall often used in machine translation and text summarization). The original associated captions from Flickr are used as reference descriptions. Table 4 shows results. For BLEU, all of our reranking strategies except visual PageRank outperform the original image based retrieval on the generation task and the best method is Visual + TFIDF reranking. For ROUGE, the best reranking strategy is Visual + Text PageRank.

We also evaluate our results by collecting human judgments using two-alternative forced choice tasks collected using Amazon's Mechanical Turk. Here, users are presented with an image and two captions (each generated by a different method) and they must select the caption which better describes the image. Presentation order is randomized to remove user bias for choosing the first or second option. Table 5 shows results. The top 3 rows show our methods are preferred over unranked phrases. Row 4 shows our top 2 methods are comparable. Finally, row 5 shows one of our methods is strongly preferred over the whole sentence baseline provided with the SBU dataset [42]. We also show some qualitative results in Fig. 6 showing successful cases of generated captions and different failure cases (due to incorrect objects, missing objects, incorrect grammar or semantic inconsistencies) for our top performing method.

7.4 Application 2: Complex Query Image Retrieval Evaluation

We test complex query image retrieval using 200 captions from the dataset described in Sec. 6.2 as queries.

Method	No Reranking	Visual PageRank	Text PageRank	Visual + Text PageRank	Visual + TFIDF Rerank
BLEU[43]	0.1192	0.1133	0.1257	0.1224	**0.1260**
ROUGE[34]	0.2300	0.2236	0.2248	**0.2470**	0.2175

Table 4 BLEU and ROUGE score evaluation of full image captions generated using HMM decoding with our strategies for phrase retrieval and reranking.

Method	Percentage
Text PageRank **vs.** No Reranking	54%/46%
Visual + Text PageRank **vs** No Reranking	57%/43%
Visual + TFIDF Reranking **vs** No Reranking	61%/39%
Text + Visual PageRank **vs** Visual + TFIDF Reranking	49%/51%
Text + Visual PageRank **vs** Global Description Generation	71%/29%

Table 5 Human forced-choice evaluation between various methods.

For 3 queries, the corresponding image was ranked first by our retrieval system. For these images the automatically selected phrases described the images so well that they matched the ground truth captions better than the phrases selected for any of the other 999 images. Overall 20% of queries had the corresponding image in the top 1% of the ranked results (top 10 ranked images), 30% had the corresponding image in the top 2%, and 43% had the corresponding image in the top 5% of ranked retrievals. In addition to being able to find the image described out of a set of 1000, the retrieval system produced reasonable matches for the captions as shown in Fig. 7.

8 Conclusion

We have described explorations into retrieval based methods for gathering visually relevant natural language for images. Our methods rely on collecting and filtering a large data set of images from the internet to produce a web-scale captioned photo collection. We present two variations on text retrieval from our captioned collection. The first retrieves whole existing image descriptions and the second retrieves bits of text (phrases) based on visual and geometric similarity of objects, stuff, and scenes. We have also evaluated several methods for collective reranking of sets of phrases and demonstrated the results in two applications, phrase based generation of image descriptions and complex query image retrieval. Finally, we have presented a thorough evaluation of each of our presented methods through both automatic and human-judgment based measures.

In future work we hope to extend these methods to a real time system for image description and incorporate state of the art methods for large-scale category recognition [11,12]. We also plan to extend our prototype complex query retrieval algorithm to web-scale. Producing human-like and relevant descriptions will be a key factor for enabling accurate and satisfying image retrieval results.

Acknowledgment

Support of the 2011 JHU-CLSP Summer Workshop Program. Tamara L. Berg and Kota Yamaguchi were supported in part by NSF CAREER IIS-1054133; Hal Daumé III and Amit Goyal were partially supported by NSF Award IIS-1139909.

References

1. Aker, A., Gaizauskas, R.: Generating image descriptions using dependency relational patterns. In: ACL (2010)
2. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.: Matching words and pictures. JMLR (2003)
3. Berg, T., Berg, A., Edwards, J., Forsyth, D.: Who's in the picture? In: NIPS (2004)
4. Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Learned-Miller, E., Teh, Y., Forsyth, D.: Names and faces. In: CVPR (2004)
5. Berg, T.L., Berg, A.C., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: ECCV (2010)
6. Brants, T., Franz., A.: Web 1t 5-gram version 1. In: LDC (2006)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: WWW (1998)
8. Chum, O., Philbin, J., Zisserman, A.: Near duplicate image detection: min-hash and tf-idf weighting. In: BMVC (2008)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
10. Deng, J., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: CVPR (2011)
11. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: ECCV (2010)
12. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: CVPR (2012)

13. Deng, J., Satheesh, S., Berg, A.C., Fei-Fei, L.: Fast and balanced: Efficient label tree learning for large scale object recognition. In: NIPS (2011)
14. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation. In: ECCV (2002)
15. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.A.: Describing objects by their attributes. In: CVPR (2009)
16. Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.A.: Every picture tells a story: generating sentences for images. In: ECCV (2010)
17. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/∼pff/latent-release4/
18. Feng, Y., Lapata, M.: How many words is a picture worth? automatic caption generation for news images. In: ACL (2010)
19. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
20. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV (2013)
21. Hays, J., Efros, A.A.: im2gps: estimating geographic information from a single image. In: CVPR (2008)
22. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. JAIR (2013)
23. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: ICCV (2005)
24. Jing, Y., Baluja, S.: Pagerank for product image search. In: WWW (2008)
25. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T.: Babytalk: Understanding and generating simple image descriptions. TPAMI (2013)
26. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
27. Kuznetsova, P., Ordonez, V., Berg, A., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: ACL (2012)
28. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Generalizing image captions for image-text parallel corpus. In: ACL (2013)
29. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
30. Leung, T.K., J., M.: Recognizing surfaces using three-dimensional textons. In: ICCV (1999)
31. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: CoNLL (2011)
32. Li, W., Xu, W., Wu, M., Yuan, C., Lu, Q.: Extractive summarization using inter- and intra- event relevance. In: Int Conf on Computational Linguistics (2006)
33. Li-Jia Li Hao Su, E.P.X., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: NIPS (2010)
34. Lin, C.Y.: Rouge: A Package for Automatic Evaluation of Summaries. In: ACL (2004)
35. Lowe, D.G.: Distinctive image features from scale invariant keypoints. IJCV (2004)
36. Mason, R., Charniak, E.: Nonparametric method for data-driven image captioning (2014)
37. Mihalcea, R.: Language independent extractive summarization. In: AAAI (2005)
38. Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Sratos, K., Han, X., Mensch, A., Berg, A., Berg, T.L., III, H.D.: Midge: Generating image descriptions from computer vision detections. In: EACL (2012)
39. Nenkova, A., Vanderwende, L., McKeown, K.: A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In: SIGIR (2006)
40. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV (2001)
41. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: From large scale image categorization to entry-level categories. In: ICCV (2013)
42. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NIPS (2011)
43. Papineni, K., Roukos, S., Ward, T., jing Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
44. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: COLING/ACL (2006)
45. Petrov, S., Klein, D.: Improved inference for unlexicalized parsing. In: HLT-NAACL (2007)
46. Radev, D.R., Allison, T.: Mead - a platform for multi-document multilingual text summarization. In: LREC (2004)
47. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon's mechanical turk. In: NAACL Workshop Creating Speech and Language Data With Amazon's Mechanical Turk (2010)
48. Roelleke, T., Wang, J.: Tf-idf uncovered: a study of theories and probabilities. In: SIGIR (2008)
49. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV (2003)
50. Stratos, K., Sood, A., Mensch, A., Han, X., Mitchell, M., Yamaguchi, K., Dodge, J., Goyal, A., III, H.D., Berg, A., Berg, T.L.: Understanding and predicting importance in images. In: CVPR (2012)
51. Tighe, J., Lazebnik, S.: Superparsing: Scalable nonparametric image parsing with superpixels. In: ECCV (2010)
52. Torralba, A., Fergus, R., Freeman, W.: 80 million tiny images: a large dataset for non-parametric object and scene recognition. TPAMI (2008)
53. Wong, K.F., Wu, M., Li, W.: Extractive summarization using supervised and semi-supervised learning. In: COLING (2008)
54. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
55. Yang, Y., Teo, C.L., III, H.D., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: EMNLP (2011)
56. Yao, B., Yang, X., Lin, L., Lee, M.W., Zhu, S.C.: I2t: Image parsing to text description. Proc. IEEE (2010)
57. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions (2014)