The Use of Object Labels and Spatial Prepositions as Keywords in a Web-Retrieval-Based Image Caption Generation System

Brandon Birmingham and **Adrian Muscat**
Communications & Computer Engineering
University of Malta
Msida MSD 2080, Malta
{brandon.birmingham.12, adrian.muscat}@um.edu.mt

Abstract

In this paper, a retrieval-based caption generation system that searches the web for suitable image descriptions is studied. Google's search-by-image is used to find potentially relevant web multimedia content for query images. Sentences are extracted from web pages and the likelihood of the descriptions is computed to select one sentence from the retrieved text documents. The search mechanism is modified to replace the caption generated by Google with a caption composed of labels and spatial prepositions as part of the query's text alongside the image. The object labels are obtained using an off-the-shelf R-CNN and a machine learning model is developed to predict the prepositions. The effect on the caption generation system performance when using the generated text is investigated. Both human evaluations and automatic metrics are used to evaluate the retrieved descriptions. Results show that the web-retrieval-based approach performed better when describing single-object images with sentences extracted from stock photography websites. On the other hand, images with two image objects were better described with template-generated sentences composed of object labels and prepositions.

1 Introduction

The automatic generation of concise natural language descriptions for images is currently gaining immense popularity in both Computer Vision and Natural Language Processing communities (Bernardi et al., 2016). The general process of automatically describing an image fundamentally involves the visual analysis of the image content such that a succinct natural language statement, verbalising the most salient image features, can be generated. In addition, natural language generation methods are needed to construct linguistically and grammatically correct sentences. Describing image content is very useful in applications for image retrieval based on detailed and specific image descriptions, caption generation to enhance the accessibility of current and existing image collections and most importantly as an assistive technology for visually impaired people (Kulkarni et al., 2011). Research work on automatic image description generation can be organised in three categories (Bernardi et al., 2016). The first group generates textual descriptions from scratch by analysing the composition of an image in terms of image objects, attributes, scene types and event actions, extracted from image visual features. The other groups describe images by retrieving sentences either from *visual* space composed of image-description pairs or from a *multimodal* space that combines image and sentences in one single space. As opposed to direct-generation-based methods, the latter two approaches generate less verbose and more human-like descriptions. In this paper, a web-retrieval-based system that exploits the ever-growing vision-text content is studied while exploring how object labels and prepositions affect the retrieval of image descriptions.

This paper is organised as follows: section 2 gives an overview of existing image caption algorithms. Section 3 outlines the problem definition and section 4 presents a web-retrieval-based framework followed by its implementation details in section 5. The dataset and evaluation are discussed in sections 6 and 7 respectively. The results are presented in section 8 followed by a discussion in section 9. Finally, section 10 concludes with the main observations and the future direction.

Proceedings of the 6th Workshop on Vision and Language, pages 11–20,
Valencia, Spain, April 4, 2017.

2 Related Work

Direct-generation models (Fang et al., 2015; Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011), exploit the image visual information to derive an image description by driving a natural language generation model such as *n*-grams, templates and grammar rules. Despite producing correct and relevant image descriptions, this approach tends to generate verbose and non-human-like image captions. The second and most relevant group of models to this paper, tackles the problem of textually describing an image as a *retrieval* problem. There are attempts that make use of pre-associated text or meta-data to describe images. For instance, Feng and Lapata (2010) generated captions for news images using an extractive and abstractive generation methods that require relevant text documents as input to the model. Similarly, Aker and Gaizauskas (2010) relied on GPS metadata to access relevant text documents to be able to generate captions for geo-tagged images. Other models formulate descriptions by finding visually similar images to the query images from a collection of already-annotated images. Query images are then described either by (a) reusing the whole description of the most visually similar retrieved image, or by (b) associating relevant phrases from a large collection of image and description pairs (Ordonez et al., 2016). Retrieval models can be further subdivided, based on the technique used for representing and computing image similarity. The first subgroup uses a *visual space* for finding related images, while the second subgroup uses a *multimodal space* for combining both textual and visual image information. The first subgroup (Ordonez et al., 2011; Ordonez et al., 2016; Gupta et al., 2012; Mason and Charniak, 2014; Yagcioglu et al., 2015), is intended to first extract visual features from the query images. Based on a visual similarity measure dependent on the extracted features, a candidate set of related images is retrieved from a large collection of pre-annotated images. Retrieved descriptions are then re-ranked by further exploiting the visual and textual information extracted from the retrieved candidate set of similar images. Conversely, retrieving descriptions from a multimodal space is characterised by the joint space between visual and textual data constructed from a collection of image-description pairs. For example, in Farhadi et al. (2010), image descriptions were retrieved from a multimodal space consisting of $< object, action, scene >$ tuples. More recently, deep neural networks were introduced to map images and corresponding descriptions in one joint multimodal space (Socher et al., 2014; Kiros et al., 2014; Donahue et al., 2015; Karpathy and Li, 2015; Chen and Zitnick, 2015).

3 Problem Definition

Image caption generators are designed to associate images with corresponding sentences, hence they can be viewed in terms of an affinity function $f(i, s)$ that measures the degree of correlation between images and sentences. Based on a set of candidate images $\mathbf{I}_{cand}$ annotated with corresponding candidate sentences $\mathbf{S}_{cand}$, typical retrieval-based caption generation methods describe an image by reusing sentence $s \in \mathbf{S}_{cand}$. The selected sentence is the one that maximises the affinity function $f(i_q, s)$ for a given query image i_q. On the contrary, generation-based image descriptors attempt to construct a novel sentence s_n composed of image entities and attributes.

The system described in this paper extracts sentences from a collection of web pages $\mathbf{W}$, rather than from a limited set of candidate human-authored image descriptions $\mathbf{S}_{cand}$, as done in most existing retrieval-based studies. Websites containing visually similar images to the query image are found using search-by-image technology. The intuition to this method is based on the fact that the evergrowing Internet-based multimedia data is a readily-available data source as opposed to the purposely constructed and limited image-description datasets used in many studies. The search for a query image can be thought of as providing a dynamic and specialised small dataset for a given query image.

The suggested framework starts by generating a simple image description based on the image visual entities and their spatial relationship. This simple description is then used as keywords to drive and optimise a web-data-driven based retrieval process. The latter is primarily intended to retrieve the most relevant sentence from the set of candidate web pages $\mathbf{W}$ by utilising the functionality offered by a search-by-image algorithm. This strategy is adopted under the assumption that web pages featuring visually similar images to a query image i_q, can contain sentences which can be effectively re-used to describe image i_q.

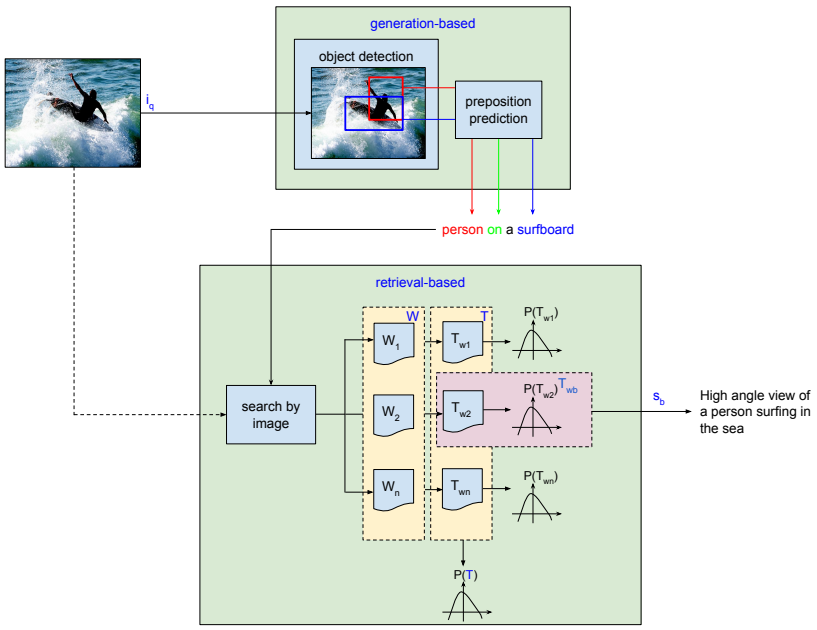


Figure 1: The proposed web-retrieval-based system designed in two stages. The query image i_q is first described by the keywords generated by the first stage. These are then used to retrieve image descriptions from a collection of web pages $\mathbf{W}$. The best sentence s_b is extracted from the best text document $\mathbf{T}_{w_b}$, with respect to the global word probability distribution $P(\mathbf{T})$ and the query image i_q.

4 Image Description Framework

The proposed generation-retrieval-based approach is centrally decomposed into two phases. The first *generation* stage of the framework is mainly intended to generate simple image descriptions that will serve as keywords for the second *retrieval* phase. By exploiting the vast amount of image-text data found on the Web, the latter will then extract the most likely sentence for a given query image. A high-level overview of the proposed image description framework is presented in Figure 1.

4.1 Generation-based Image Description

The first stage of the image description generation framework analyses the image visual content and detects the most important image objects. Therefore, the aim of this step is to detect and annotate image objects with corresponding high-level image labels and corresponding bounding boxes. In order to describe the spatial relationship between the predominant image objects, various predictive models based on different textual and geometric feature sets, were investigated as described in section 4.2. From this simple generated image description, in the form of an *object-preposition-article-object* keyword structure, the framework is then designed to drive a web-retrieval-based process. This process exploits both the visual aspect of the query image, as well as the linguistic keywords generated by the first stage of the pipeline.

4.2 Preposition Predictive Model

The generation of prepositions was cast as a prediction-based problem through geometrical and encoded textual features. Four different predictive models based on separate feature sets were analysed. This experiment confirmed that the Random Forest model obtained the best preposition prediction accuracy rate. This was achieved when predicting prepositions via word2vec (Mikolov et al., 2013) textual labels combined with the geometric feature sets used by Muscat and Belz (2015) and Ramisa et al. (2015). This setup marginally outperformed the best preposition prediction accuracy achieved by Ramisa et al. (2015) when trained and evaluated on the same Visen's MSCOCO Prepositions[1] testing set having original object labels. Results can be found in Table 1.

4.3 Retrieval-based Image Description

The aim of the second phase of the proposed framework is to retrieve descriptions based on the visual aspect of a query image and its correspond-

[1] http://preposition.github.io

Table 1: The accuracies obtained from the Visen's MSCOCO original object labels. The accuracies for different configuration setups are presented, based on different geometric feature sets, in relation to different textual label encoding. LE stands for the Label Encoder which encodes object labels with corresponding integers, IV for Indicator Vectors and W2V for Word2Vec.

	Geometric + Textual Features											
	Ramisa et al.				Muscat & Belz				All Geometric Features			
Model	LE	IV	W2V	GF	LE	IV	W2V	GF	LE	IV	W2V	GF
SVM	0.03	0.42	0.77	0.60	0.01	0.42	0.77	0.60	0.08	0.44	0.74	0.63
Decision Tree	0.53	0.66	0.75	0.69	0.52	0.65	0.76	0.67	0.53	0.64	0.75	0.69
Random Forest	0.60	0.65	**0.81**	0.72	0.56	0.62	**0.81**	0.69	0.59	0.68	**0.82**	0.71
Logistic Regression	0.64	0.50	0.80	0.64	0.61	0.50	0.81	0.61	0.65	0.51	0.80	0.64

ing simple generated image description, as discussed in Section 4.1. This phase is designed to find a set of web pages composed of images that are visually related to the query image. This search functionality is freely available by the current two dominant search-engines, Google[2] and Bing[3]. These two proprietary image-search algorithms are able to retrieve visually similar images, which may therefore be used for collecting web pages with featured visually similar images. From the retrieved collection of web pages characterised with visually similar images to the query image, this phase is designed to extract the best sentence that can be used to describe the query image. Based on the idea that websites usually describe or discuss the embedded images, it is assumed that this stage is capable of finding human-like sentences describing the incorporated images which can be re-used to describe the query images.

Given a collection of candidate web pages $\mathbf{W}$ with embedded visually similar images, this phase is intended to extract the main text $\mathbf{T}_{w_i}$ from each corresponding web page $w_i \in \mathbf{W}$. This is carried out by analysing the Document Object Model (DOM) of each web page as well as by statistically distinguishing between HTML and textual data. Moreover, this stage is intended to discard any boilerplate text that is normally found in web pages, including navigational text and advertisements by exploiting shallow text features (Kohlschütter et al., 2010). After transforming the set of web pages $\mathbf{W}$ to the corresponding text documents $\mathbf{T}$, this stage computes the word probability distribution $P(\mathbf{T}_{w_i})$ for each $\mathbf{T}_{w_i}$, disregarding any stop words in the distribution. The text found in each text document $\mathbf{T}_{w_i}$ is combined in one text collection $\mathbf{T}$ and the probability distribution $P(\mathbf{T})$, representing all the probabilities for the words contained in collection $\mathbf{T}$, is calculated. The top k most probable words from each generated probability distribution $P(\mathbf{T}_{w_i})$ are considered to find the most probable relevant text document $\mathbf{T}_{w_b}$, for the extraction of the best sentence s_b that describes the query image i_q. Specifically, the best text document is selected by the following maximising function over each text document probability distribution $P(\mathbf{T}_{w_i})$, with respect to the global word probability distribution $P(\mathbf{T})$:

$$\mathbf{T}_{w_b} = \arg\max_{w_i} \sum_{n=1}^{k} P(\mathbf{T}_{w_{i,n}}) P(\mathbf{T} = \mathbf{T}_{w_{i,n}}), \tag{1}$$

where n represents the n^{th} most probable word of the probability distribution.

This strategy is used to eliminate documents that are probably irrelevant to provide correct descriptions for query images. A similar approach is carried out to retrieve the best sentence s_b that could potentially describe the query image. The technique used to select the most appropriate sentence from $\mathbf{T}_{w_b}$ is initiated by extracting the set of candidate sentences $\mathbf{S}_{cand}$ from the selected best file $\mathbf{T}_{w_b}$. The second step is to weight each sentence $s_i \in \mathbf{S}_{cand}$ by the summation over how probable each word is, with respect to the global word probability distribution $P(\mathbf{T})$. Therefore, s_b is retrieved by maximising the following formula:

$$s_b = \arg\max_{s_i} \sum_{n=1}^{|s_i|} P(\mathbf{T} = s_{i,n}), \tag{2}$$

where n represents the n^{th} word found in sentence $s_i \in \mathbf{S}_{cand}$ extracted from the best file

[2] https://images.google.com

[3] https://www.bing.com/images/explore?FORM=ILPSTR

$\mathbf{T}_{w_b}$, and $|s_i|$ represents the number of words found in sentence s_i.

To further enhance the contextual reliability of the selected sentence, the approach used to retrieve image descriptions is combined with the image visual aspect. This is accomplished by weighting the visible object class labels in accordance to their corresponding image predominance level. The area of the visible image entities, with respect to the entire query image i_q, was used to prioritise visible image objects. Therefore, the best sentence s_b is retrieved by combining the knowledge extracted from the most probable words found in $P(\mathbf{T})$ and the visual aspect of the query image i_q, by the following formula:

$$s_b = \arg\max_{s_i} \sum_{n=1}^{|s_i|} P(\mathbf{T} = s_{i,n}) R(i_q, s_{i,n}), \quad (3)$$

where R is a function which computes the area of the object class label $s_{i,n}$ found in the n^{th} word of sentence s_i in the context of image i_q.

5 Implementation

The image description generation framework was modularised and implemented in two stages. To detect the main image objects, the first stage employs the two-phased fast region-based convolutional neural network (R-CNN) proposed by Ren et al. (2015). The first module of the R-CNN is a deep fully convolutional neural network designed to propose regions, while the second module is a detector that uses the proposed regions for detecting image objects enclosed in bounding boxes. This architecture is trained end-to-end into a single network by sharing convolutional features. The deep VGG-16 model (Simonyan and Zisserman, 2014) pre-trained on MSCOCO (Lin et al., 2014) dataset, was utilised to detect image objects with corresponding class labels and bounding boxes. These were then used to infer the spatial relationship between the detected image objects as discussed in section 4.2.

By using the linguistic keywords generated from the first stage, the second part of the framework is designed to retrieve the most probable sentence from a set of relevant web pages that feature visually similar images. The set of web pages is collected by using the free functionality offered by Google's Search By Image[4] proprietary technology. For a given uploaded query image, this functionality is intended to return visually similar images. Based on extracted image visual features and automatically generated textual keywords by the same functionality, Google's Search by Image retrieves visually similar images. The websites of the visually returned images are then retrieved from the corresponding URLs binded with each visually similar image. By using Selenium[5] to automate the headless PhantomJS browser, query images were automatically uploaded to retrieve websites featuring visually similar images. In this study, it was shown how object labels connected with spatial prepositions affect the retrieval search performed by Google's search-by-image algorithm. This was accomplished by replacing Google's keywords with object labels and preposition generated by the first stage of the proposed framework. Furthermore, this study also investigated whether stock photography websites could improve the retrieval search of the designed framework. The retrieval of websites featuring stock photos was achieved by concatenating the phrase "stock photos" with the keywords extracted from the visual aspect of the query image. To detect and extract the main textual content of each respective web page, the boilerpipe[6] toolkit was employed. From the set of extracted text documents, the most probable sentence that best describes the query image is then retrieved, as discussed in Section 4.3.

6 Dataset

To evaluate the proposed image description framework, a specific subset of human-annotated images featured in MSCOCO[7] testing set was used. Since the preposition prediction task is targeted to generate prepositions between two image objects, describing images having exactly two image objects was of particular interest to this study. Therefore, the following steps were carried out to select images consisting of two image objects. From the ViSen's MSCOCO testing set, 1975 instances having strictly one single preposition between two image objects were found and extracted. Finally, 1000 images were randomly selected from the latter subset. Since images may contain background image objects, the same object detector employed in the proposed framework was used for detecting

[4] https://images.google.com

[5] http://docs.seleniumhq.org

[6] https://boilerpipe-web.appspot.com

[7] http://mscoco.org

Table 2: Configuration Setups

Setup	Name	Image Descriptions
G	Generation	Descriptions consisting of object labels
GP	Generation-Preposition	Descriptions consisting of object labels connected with spatial prepositions
R	Retrieval	Descriptions retrieved based on Google's automatic generated keywords
GR	Generation-Retrieval	Descriptions retrieved based on the generated keywords by G
GRS	Generation-Retrieval-Stock	Descriptions retrieved based on the generated keywords by G from stock photography websites
GPR	Generation-Preposition-Retrieval	Description retrieved based on the generated keywords by GP
GPRS	Generation-Preposition-Retrieval-Stock	Descriptions retrieved from stock photography websites based on the descriptions generated by GP

objects. The fast R-CNN found 128 images containing one image object, 438 images containing exactly two image objects, while the remaining 434 images contained more than two image objects. For the evaluation of this framework, images composed of one and two image objects were only considered. Therefore, the framework was evaluated on a dataset consisting of 566 images, where 128 images contain one single object, while the other remaining 438 images contain exactly two image objects.

7 Evaluation

Both human and computational evaluation were used to evaluate the web-retrieval-based framework. The automatic evaluation was performed by using existing metrics, intended to measure the similarity between generated descriptions and corresponding human ground truth descriptions. The measures include BLEU (Papineni et al., 2002), ROUGE_L (Lin and Hovy, 2003), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015). To complement the automatic evaluation, human judgments for image descriptions were obtained from a qualified English teacher. Since the human evaluation process is considerably time-consuming, human judgments were collected for a sample of 200 images split equally for single and double-object images. The same human evaluation criteria proposed by Mitchell et al. (2012) was used to evaluate the generated descriptions. Human evaluation was conducted by rating the grammar, main aspects, correctness, order and the human-likeness of descriptions using a five-point Likert scale.

8 Results

The framework was evaluated in each phase of its pipeline as described in Table 2. The results are given in Tables 3 and 4 for single and double-object images respectively. The generation phase of the framework that describes images with just object labels is represented by G, while the standalone retrieval-based approach which uses Google's automatic generated keywords is represented by R. Furthermore, when describing single-object images, the joint generation-retrieval stage that uses the prototype's keywords is represented by GR. When describing double-object images, the generation-retrieval process is denoted by GPR given that it uses both object labels and prepositions as keywords. Moreover, the results obtained when the retrieval phase considers stock photography websites are denoted by the letter S. The retrieval-based stages are specified by the two parameters, W and F. The latter represents the number of text files analysed from the corresponding websites, whereas W represents the number of most probable words used for the selection of the best sentence from a set of web pages. A grid search was performed to find these parameters for each configuration. The same notation was used for the human evaluation results. Typical image descriptions generated by the proposed web-retrieval-based image caption generation system can be found in Figure 2.

9 Discussion

The automatic evaluation showed that single-object images were best described by the generation-retrieval from stock photography websites (GRS). This outperformed the one-word description of the generation-based configuration

Table 3: Automatic evaluation of single-object images.

Metric	Model			
	G	R (20W, 30F)	GR (5W, 35F)	GRS (5W, 35F)
CIDEr	0.134	0.066	0.099	**0.154**
BLEU@4	0.000	0.000	0.010	**0.013**
BLEU@3	0.000	0.007	0.022	**0.032**
BLEU@2	0.001	0.026	0.058	**0.074**
BLEU@1	0.001	0.080	0.148	**0.173**
ROUGE_L	0.124	0.101	0.133	**0.164**
METEOR	0.062	0.060	0.078	**0.089**

Table 4: Automatic Evaluation of double-object images.

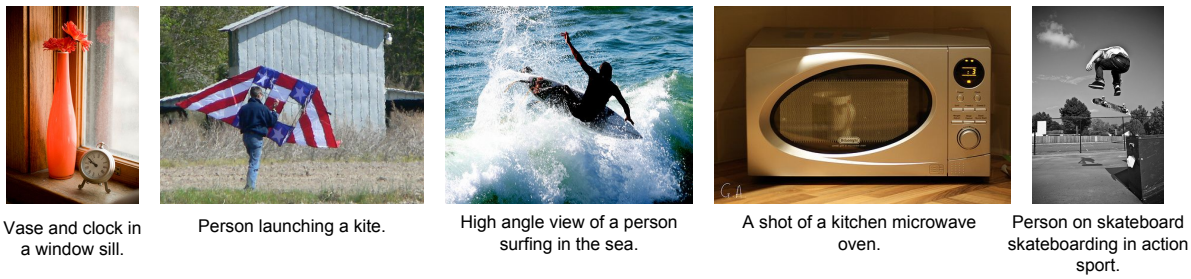
Metric	Model						
	G	GP	R (20W, 30F)	GR (5W, 35F)	GRS (5W, 25F)	GPR (10W, 15F)	GPRS (10W, 15F)
CIDEr	0.482	**0.604**	0.082	0.148	0.176	0.132	0.152
BLEU@4	0.033	**0.132**	0.005	0.014	0.018	0.013	0.017
BLEU@3	0.085	**0.187**	0.015	0.030	0.036	0.028	0.035
BLEU@2	0.165	**0.241**	0.038	0.069	0.081	0.067	0.077
BLEU@1	0.252	**0.292**	0.125	0.190	0.199	0.175	0.190
ROUGE_L	0.340	**0.413**	0.130	0.185	0.210	0.174	0.198
METEOR	0.152	**0.177**	0.078	0.109	0.117	0.100	0.113

(G), as well as the retrieval-based (R) setup. The latter result confirms that the replacement of Google's Search by Image captions improved the retrieved descriptions. This concludes that more relevant images were returned by Google when replacing its automatic caption with object labels.

Conversely, double-object images were best described via the generation-preposition (GP) configuration. Although replacing Google's Search By Image keywords improved the results, the simple descriptions based on object labels connected with spatial prepositions were more accurate. Automatic evaluation also confirmed that the web-retrieval approach (GRS) performs better on double-object images. This study also showed that the retrieval process performs better without using prepositions as keywords. This resulted from the fact that prepositions constrain the search result performed by Google when indexing web pages, since most descriptive text available on the Web includes verbs rather than prepositions.

The human evaluation results for the single-object images are presented in Table 5. Particularly, generation-based (G) descriptions obtained a grammatical median score of 1, confirming that one-word descriptions do not produce grammatically correct sentences. The results also confirm that the used object detector accurately describes the dominant objects in an image. By considering the improbability of one-word human derived descriptions, this stage resulted in a low human likeness score of 2. The retrieval method applied on stock photography websites (RS) lead to grammatical improvement in the generated descriptions. Such descriptions were grammatically rated with a median score of 3. However, results show that the retrieval method decreases the relevancy of the retrieved descriptions. Despite generating grammatically sound sentences with better human-likeness, the human evaluation showed a degree of inconsistency between the descriptions and their corresponding images. When combining the generation (G) and retrieval (RS) proposed approaches, the grammar, order and the human likeness improved for single-object images.

Table 5 also demonstrates that the generation-preposition (GP) configuration generated the best descriptions when describing double-object images. Furthermore, these results also confirmed that the retrieval (RS) approach improves when replacing Google's caption with object labels. The human evaluation also established the ineffectiveness of the retrieval stage when combined with the generation-prepositions (GPRS) stage. This table also confirmed that the web-retrieval approach described double-object images better than single-object images.



(a)

(b)

Figure 2: (a) Correct and (b) incorrect descriptions generated by the web-retrieval-based framework.

Table 5: Human evaluation of single and double-object images with scores (1-5) obtained for each stage of the proposed framework: median, mean and standard deviation in parentheses.

	single-object images				
Model	Grammar	Main Aspects	Correctness	Order	Humanlike
G	1 (1.11, 0.31)	4 (3.82, 0.89)	5 (4.84, 0.68)	5 (4.38, 1.04)	2 (1.79, 0.65)
RS	3 (3.31, 1.50)	2 (2.27, 1.26)	2 (2.07, 1.35)	3 (2.90, 1.63)	2.5 (2.68, 1.46)
GRS	4 (3.56, 1.25)	2 (2.31, 1.02)	2 (2.00, 1.14)	4 (3.26, 1.60)	3 (2.75, 1.22)

	double-object images				
Model	Grammar	Main Aspects	Correctness	Order	HumanLike
G	4 (3.80, 0.65)	5 (4.42, 0.97)	5 (4.69, 0.75)	5 (4.63, 0.79)	4 (3.77, 0.72)
GP	5 (4.44, 0.97)	5 (4.53, 0.81)	5 (4.81, 0.63)	5 (4.69, 0.81)	5 (4.43, 0.90)
RS	4 (3.39, 1.24)	2 (2.50, 1.25)	2 (2.20, 1.14)	2 (2.27, 1.26)	3 (2.93, 1.45)
GRS	3 (3.00, 1.41)	3 (3.14, 1.24)	2.5 (2.71, 1.32)	3 (2.93, 1.40)	3 (2.69, 1.52)
GPRS	3 (2.70, 1.32)	3 (2.87, 1.13)	2 (2.42, 1.16)	2 (2.45, 1.31)	2.5 (2.38, 1.31)

10 Conclusion and Future Work

This paper investigated the use of object labels and prepositions as keywords in a web-retrieval-based image caption generator. By employing object detection technology combined with a preposition prediction module, keywords were extracted in the form of object class labels and prepositions. The proposed retrieval approach is independent of any purposely human-annotated image datasets. Images were described by extracting sentences found in websites, featuring visually similar images to the query image. The search is aided with the use of the generated keywords. This approach was particularly effective when describing single-object images, and especially so when extracting sentences from stock photography websites.

Despite the retrieval of relevant descriptions for both single and double-object images, object labels connected with spatial prepositions obtained better accuracies when describing double-object images. Although Google's Search By Image was enhanced by the replacement of its predicted image annotations with object labels, further work in using a wider variety of keywords such as verbs can be carried out to improve the results. It is also worth studying whether linguistic parsing can be used to assess the quality of sentences during the caption extraction phase to increase the likelihood of choosing better sentences.

References

Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1250–1258, Uppsala, Sweden, July. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55(1):409–442, January.

Xinlei Chen and Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2422–2431.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2625–2634.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1473–1482.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 15–29, Heraklion, Crete, Greece. Springer-Verlag.

Yansong Feng and Mirella Lapata. 2010. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249, Uppsala, Sweden, July. Association for Computational Linguistics.

Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. 2012. Choosing linguistics over vision to describe images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 606–612, Toronto, Ontario, Canada. AAAI Press.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA. ACM.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1601–1608, Washington, DC, USA. IEEE Computer Society.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 71–78, Edmonton, Canada. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer International Publishing, Cham.

Rebecca Mason and Eugene Charniak. 2014. Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Baltimore, Maryland, June. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, Lake Tahoe, Nevada. Curran Associates Inc.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 747–756, Avignon, France. Association for Computational Linguistics.

Adrian Muscat and Anja Belz. 2015. Generating descriptions of spatial relations between objects in images. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 100–104, Brighton, UK, September. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 1143–1151, Granada, Spain. Curran Associates Inc.

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, and Tamara L. Berg. 2016. Large scale retrieval and generation of image descriptions. *Int. J. Comput. Vision*, 119(1):46–59, August.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal, September. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 91–99, Montreal, Canada. MIT Press.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Richard Socher, Andrej Karpathy, Quoc Le, Christopher Manning, and Andrew Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakici. 2015. A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 106–111, Beijing, China, July. Association for Computational Linguistics.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 444–454, Edinburgh, United Kingdom. Association for Computational Linguistics.