# Data-Driven Image Captioning via Salient Region Discovery

Mert Kilickaya[1], Burak Kerim Akkus[2], Ruket Cakici[2], Aykut Erdem[1], Erkut Erdem[1,*], Nazli Ikizler-Cinbis[1]

[1]Hacettepe University, Dept. of Computer Engineering, Ankara, Turkey
[2]Middle East Technical University, Dept. of Computer Engineering, Ankara, Turkey

**Abstract:** In the past few years, automatically generating descriptions for images has attracted a lot of attention in computer vision and natural language processing research. Among the existing approaches, data-driven methods have been proven to be highly effective. These methods compare the given image against a large set of training images to determine a set of relevant images, then generate a description using the associated captions. In this study, we propose to integrate an object-based semantic image representation into a deep features-based retrieval framework to select the relevant images. Moreover, we present a novel phrase selection paradigm and a sentence generation model which depends on a joint analysis of salient regions in the input and retrieved images within a clustering framework. We demonstrate the effectiveness of our proposed approach on Flickr8K and Flickr30K benchmark datasets and show that our model gives highly competitive results compared to the state-of-the-art models.

## 1. Introduction

Image captioning, the problem of automatically generating descriptions from images, is a new task at the intersection of computer vision and natural language processing which has gained much attention in recent years [4]. Our focus in this paper is to improve the existing data-driven image captioning models. Given an image, these models retrieve a number of relevant images and their captions from a large set of images from which they generate a description. The pioneering study [24] considers a two-step strategy, which includes an initial retrieval of globally similar images and a consecutive image analysis to select the most relevant image. This last step relies on running detectors; responses of these detectors are used to re-rank the images so that the caption of the most similar image is selected as the description of the input image. This strategy, however, does not consider an open-world assumption and depends on a predefined set of classifiers. This limits the representation capacity, and affects overall performance. Extending detection models is computationally expensive as it requires both additional labeled samples and training. Moreover, the object detectors may output noisy results (see Figure 1(a)), which could affect the captioning process. In addition, even if all the objects are accurately detected, this does not always mean that all are worth mentioning [3] (see Figure 1(b)).

When the captions of retrieved images are used only to determine which visual detectors are fired and the caption is selected by considering the visual features alone (as in [24, 27]), some inaccuracies may occur. Some studies have eliminated this drawback by employing visual features along with the captions [39, 9]. They replace the detector-based analysis step with a simpler yet effective text-based alternative where the captions of the retrieved set of images are examined directly.

Another critical issue is that sometimes the caption of a single selected image falls short of


(a) (b)

Fig. 1: Problems with detection-based image captioning. (a) A state-of-the-art object detector [11] detects the boy as a "person", but it also results in many noisy detections. (b) The object detector detects the child and the potted plants in the image but only the child appears in the description of the image.

Two people are under a blue umbrella while a woman riding a bike has a green umbrella .

Two people stand by the water 's edge .

Man and woman riding bikes on paved , fenced path near ocean .

Two blond models pose with a man under an umbrella .

Fig. 2: Problems with selecting a single caption from the retrieved images. The selected caption might not express the whole scene with all its complexities.

describing the input image with all its complexities. The problem is that all the aforementioned models analyze the images and the captions in a global manner. However, global caption transfer is only accurate when there is an exact similar image in the database, which may not always be the case. On the other hand, local information in images and captions are as valuable (Figure 2).

Our first contribution is a novel mechanism to select more relevant images from a large pool by combining deep features with an object-specific image representation. Specifically, we first compare a given image against a large set of images using the state-of-the-art deep features that encode scene and object characteristics. We further improve this initial retrieval by reranking the images considering a novel visual representation, called Bag-of-Object-Proposals, which encodes images in terms of objects. This improves the accuracy of the relevancy of retrieved images, and from here, we can generate more accurate descriptions.

Our second contribution is an effective local image analysis method which is used to explore the similarities among local image regions. Motivated by the observation that people mention different salient image regions when describing an image, we utilize a state-of-the-art visual saliency model to select image regions that are likely to attract attention. In particular, using a single caption from the retrieved captions might fall short in describing every detail of the input (Figure 2). We use similarities among salient image regions to collect phrases from the relevant set of images, and develop a phrase-based sentence generation approach. This allows us to consider local image characteristics within a data-driven strategy, and leads to the generation of more accurate descriptions, as we illustrate on Flickr8K [14] and Flickr30K [40] datasets.

The paper is organized as follows: We first provide a review of the related work in Section 2. Next, we present our approach in Section 3, where we discuss the proposed image retrieval and reranking scheme (Section 3) and our image captioning model which first discovers salient image regions (Section 3.1) and then employ them to collect a set of relevant phrases that are used to generate a novel description (Section 3.2). Section 4 presents our experimental results on benchmark datasets and discusses the proposed framework.

## 2. Related Work

Data-driven image captioning approaches can be divided into three groups: (1) models that select and transfer a caption based on visual information alone [24, 27], (2) models that select and transfer a caption based on a combination of visual and textual information [23, 39, 9], and (3) models that generate a novel description by combining fragments of the retrieved descriptions [18, 12]. In the following, we provide a brief summary of these approaches along with a review on the relation between attention and image captioning.

**Transfer-based captioning using visual information.** The early examples of data-driven models [24, 27] pose image captioning problem as a visual retrieval problem. From a large set of images they find the image that is visually closest to the input image and then transfers its caption as the final description.

The *Im2Text* model by Ordonez et al. [24] consists of two main steps. First, a number of images are retrieved from a large pool of images using simple global image descriptors. Next, a more thorough visual analysis is performed by using retrieved captions to fire some classifiers. Each image is represented in terms of the responses of these semantic classifiers, which can be used to rerank the retrieved images. Finally, the caption of the top-ranked image is transferred to the input image as the description. Patterson et al. [27] simply replace the global image features in the initial retrieval of [24] with semantic features encoding scene characteristics. As demonstrated, using better global image descriptors gives better image retrieval performance, improving the description quality.

**Transfer-based captioning using visual and textual information.** Mason and Charniak [23] formulate the caption transfer problem as an extractive summarization problem. After retrieving a set of visually relevant images as in [27], they rerank them by utilizing their captions. The image captions are encoded via a word frequency-based representation, which in return provides an approximate textual representation for the input image. Finally, the caption which better summarizes the textual content of the given image is selected and transfered.

Yagcioglu et al. [39] employ an average query expansion approach which depends on compositional distributional semantics of sentences. Images are represented by deep features [6] and the relevant images are retrieved. Then, the captions of these retrieved images are represented in terms of a distributed word model [29], and the weighted average of these sentence representations is taken.Devlin et al. [9] consider a similar strategy. They employ CNN activations as a global image descriptor.Then, they determine the so-called “consensus caption” among the retrieved captions, which corresponds to the description that has the highest textual similarity with all of the retrieved captions on average.

**Generation-based captioning.** Success of the data-driven image captioning models highly depends on the size of the image set. As dataset size increases, it is more likely that for a given image, one can find similar images, and hence use their captions to describe the input image. However, this is not always possible, as depicted in Figure 2. A potential solution could be generating novel descriptions for a given image using parts of captions of the visually similar images. Li et al. [20] generate image descriptions by selecting certain phrases and integrating them. The phrases are used in the form of triplets containing two objects and a preposition, which are obtained by off-the-shelf object and stuff detectors. They combine phrases using n-gram scores. Kuznetsova et al. [18] extend this model by obtaining the phrases by performing a non-parametric analysis. Specifically, they represent images by the responses of the detectors and the classifiers used in [24], but use these responses separately to collect noun, verb phrases, and two types of preposition phrases for region/stuff and scene. Finally, a description is generated from collected phrases by treating caption generation as a constraint optimization problem which considers word ordering and redundancy.

Gupta et al. [12] again propose to retrieve visually similar images using a set of hand-crafted visual features. Then, they perform dependency parsing on the descriptions of these retrieved images to segment these descriptions into phrases, and measure the relevance of these phrases using a joint probability model that considers image similarity and Google search counts. Finally, they generate a set of descriptions by using the SimpleNLG [10] tool.

In our study, we follow a similar strategy, but we do not use any hand-crafted object detector or scene classifier as in [18]. We rather formulate the problem as a phrase selection and ordering problem. We make a thorough analysis of the retrieved images by considering an object proposals-based scheme, which allows to obtain a small but visually more relevant subset of object regions. The phrases extracted from them are then used to generate the final description, as opposed to [12] which extracts the phrases from all of the retrieved images and scores them by considering visual similarity and priors based Google. We generate sentences by using language models based on the phrases and the words extracted from image descriptions. In particular, we maximize the score that combines unary and pairwise phrase scores.

**Visual attention and visual descriptions.** In [41], the authors show that there is a high correlation between where humans look in an image and what they mention while they are describing images. In [3], the authors analyze the importance of objects via image descriptions, and show that they can learn to predict whether an object is worth mentioning or not by considering its internal and semantic characteristics. Our work builds upon these findings. In particular, we sample certain regions from images in a category-independent manner and we measure the importances of these regions using a bottom-up saliency model [22] within an Inhitibion-of-return [15] based framework.

## 3. Proposed Approach

Figure 3 illustrates the steps of our proposed approach which are described in the following sections.

We propose a two-step coarse-to-fine procedure for image retrieval. First, we utilize convolutional neural network (CNN) features to retrieve the images of similar scenes. Then, we rerank those candidate images using a novel object-based semantic representation.

Initially, we utilize HYBRID-CNN model [42] to represent images. In particular, we use the output of the fc7 layer of the HYBRID-CNN as our global image descriptor and express each image by a 4096-dimensional feature vector. Sharing the same architecture with AlexNet[17], HYBRID-CNN is trained on two large datasets, the scene-centric PLACES dataset [42] and the object-centric ImageNet dataset [8] For a query image, the initial set of relevant images is obtained as the set of N-nearest neighbors where ($N = 100$) and the distance is calculated via $L_2$ distance.

HYBRID-CNN features have the ability to encode both scene and object information. However, such a global representation may not always be sufficient enough to fully capture the image content in all its complexity. Hence, we introduce an additional reranking step in which the candidate images are reranked by taking into account the local object content in a more explicit way. The reranking strategy in this study is quite different than the previous work, particularly that of [25], in that we deliberately avoid incorporating any pretrained object detectors for the reasons mentioned above. The key to our proposed reranking approach is a new object-based semantic image representation, which we refer to Bag-of-Object-Proposals (*BoOP*). For each

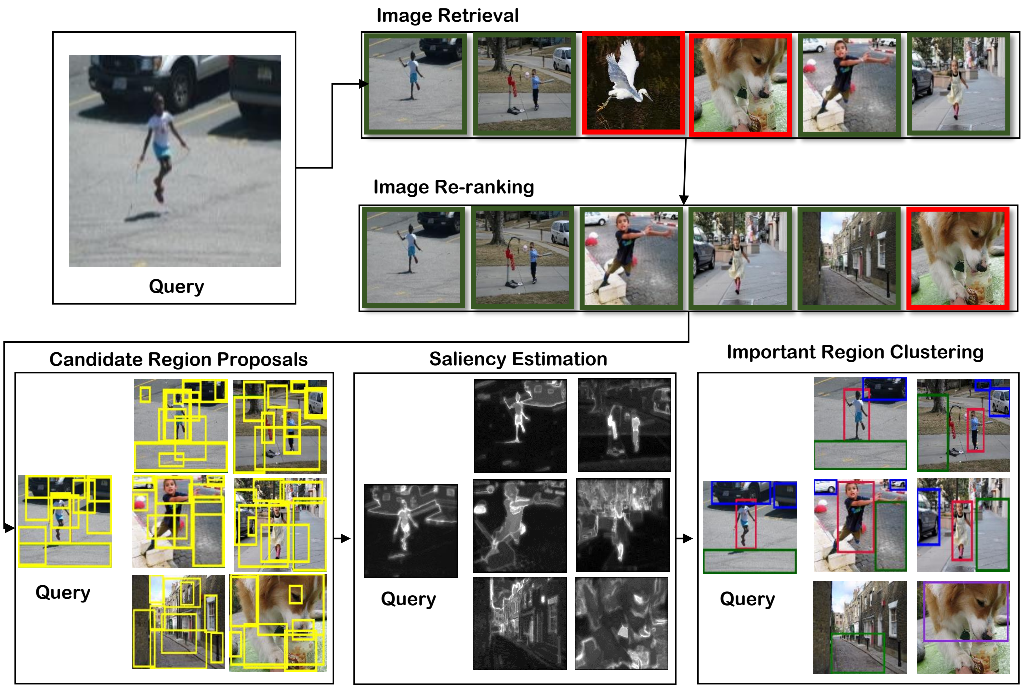



Fig. 3: Steps of the proposed image captioning framework. First, our approach employs a novel strategy to retrieve visually similar images by making use of both global and local image features (Section 3). Next, it expands its analysis to retain only the most relevant ones, determined by focusing on the salient parts of the images (Section 3.1). Once the final set of candidate images are collected, the approach generates a description of the query image by exhaustively combining phrases from the retrieved captions of relevant images (Section 3.2).

image, we subsample object-like windows using [38], and represent each window with fc7 features of VGGNet [31]. Similar to a Bag-of-words (BoW) approach, we then learn a dictionary (of size 1000) and apply LLC coding [37] to obtain the codeword vector of any given image through quantization over candidate image regions.

While image CNN features provide a holistic image representation, our BoOP representation yields complementary information about the potential objects in the image. To integrate the strengths of both views and enhance the retrieval quality, we apply the VisualRank[16] algorithm as follows. First, we assume that an initial ranking $\mathbf{r}^{(0)}$ is given by HYBRID-CNN. Then, we compute the final ranking as the PageRank of the similarity graph $\mathbf{W}$ generated according to BoOP codeword vectors. This process can be formulated as the solution of the following eigenvalue problem:

$$\left(\alpha\mathbf{P} + (1-\alpha)\frac{1}{n}\mathbf{e}^T\right)\mathbf{r} = \mathbf{r}. \tag{1}$$

where $\alpha \in (0,1)$ is a fixed teleportation parameter, $\mathbf{P}$ is the column normalized version of the matrix $\mathbf{W}$, and $\mathbf{e}$ is the column vector of all ones.

### 3.1. Identifying salient regions for caption generation

Inspired by [3, 41], we introduce an attention-driven refinement step to further retain the images whose salient regions are similar to those of the query image.

The proposed approach is a joint clustering-based approach where we aim at simultaneously identifying the salient regions having similar characteristics in the retrieved set of training images and the query image. This problem is somehow related to the co-detection/co-localization task [13, 33], but it is inherently different since it is unsupervised and it utilizes saliency to guide multiple joint detection tasks. The steps of the algorithm are summarized in the second row of Figure 3. First, several candidate object proposals are extracted. Then, we filter them out by considering their saliency. Finally, we jointly cluster the remaining set of image regions to identify which training images are similar to the query in terms of the salient regions.

Similar to our BoOP representation, we again resort to a generic objectness method [1] to extract an initial set of candidate image regions. Specially, for each image, we consider the top $K$ proposals with highest objectness scores. We additionally utilize R-CNN detector [11], which better extracts the objects in an image. Note that, even though R-CNN is trained on ImageNet to detect objects form 200 different classes, we do not make use of the class labels.

Our goal is to identify important image regions that are likely to be mentioned by humans. Even though we sample relatively less number of proposals, the number is still very high and clearly not all of those image regions should appear in the description of the corresponding image. Hence, we need a way to select a subset of these candidate regions to keep only those that are worth mentioning. Inspired by the study in [41], our importance definition is based on the notion of visual saliency [5] where we adopt the Inhitibion-of-Return (*IoR*) mechanism [15]. For this, we first employ the superpixel-based method of [22] to compute the visual saliency maps. Then, we use the mean saliency value within a region to associate an importance score to each candidate region. Giving more priority to candidate regions from R-CNN detections, at each iteration we apply non-maximum suppression to choose the most salient region. After a region is selected, we subtract a local Gaussian distribution to darken the corresponding region in the saliency map. We continue in this manner after updating the importance scores of the remaining windows. This process stops when saliency scores fall below a certain threshold. Figure 4 shows the stages of this IoR mechanism on some sample images.

Once we identify the most salient regions in the query image and the retrieved set of visually similar images, what remains is to keep only the training images which share common important regions. This is different than reranking since we return not a ranking but a set of images relevant to the query image. As will be described in Section 3.2, we put all descriptions that belong to each cluster to the same bag, and use textual relations with these descriptions to generate a novel description of the query image.

To group the set of salient regions extracted from the query image and its neighboring training images, we utilize the graph-based clustering method of Dominant Sets (DS) [28], which basically finds the maximal cliques in the corresponding affinity graph. DS results in good compact clusters and quite robust to outliers in the data, which is central to our purpose. Given the salient regions of the query $q_i$ and their cluster assignments $c_i \in C_q$, we keep the retrieved images $r_i$ only if they have at least one important region in the query clusters $r_i \in C_q$. The rationale behind picking the relevant images in this way is that we end up with a small number of similar training images, which are not only similar globally, but also share similar local regions. The captions of these visually similar images are then further exploited to form a good description of the query image.

### 3.2. Combining Relevant Phrases to Generate Descriptions

In this step, the captions collected through the salient regions of the input image are syntactically analysed and decomposed into phrases and tokens in order to be combined again to generate the novel description for the query image. We propose three different generation methods. The results for each of these methods are presented in Section 4.2.

*3.2.1. Phrase Combination Models (PHRASE):* Phrase combination method utilizes a phrase-structure parser in order to isolate phrases from the collected descriptions. We use the

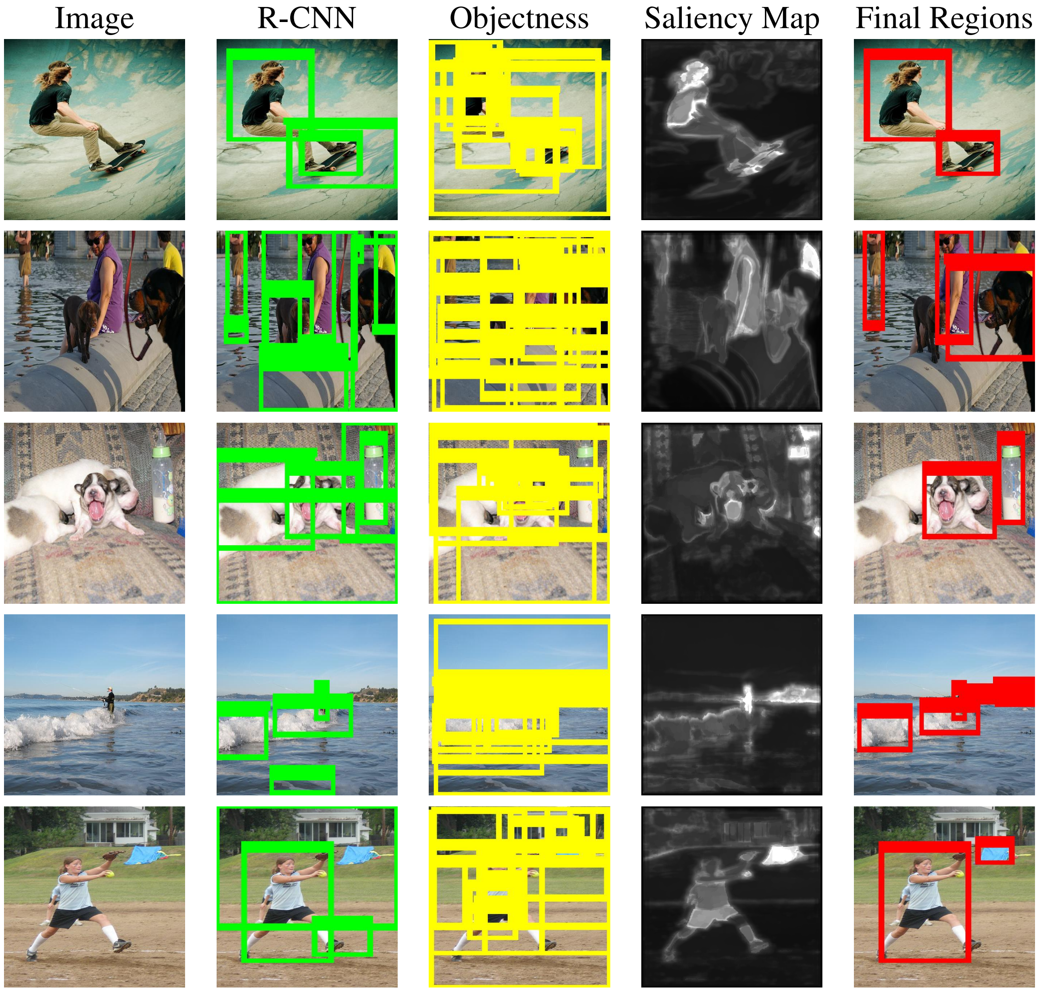



Fig. 4: Identifying important image regions. From left to right: Image, R-CNN proposals, Objectness proposals, Saliency Map and the important image regions. Candidate region proposals from two different methods that have a high overlap are suppressed using non-maximum supression, taking R-CNN proposals as references.

Stanford Parser[32] to parse the sentences. The phrase structure trees are then traversed from root to the leaves and each node corresponding to an NP (noun phrase), VP (verb phrase) or PP (prepositional phrase) are extracted. This sometimes results in the parts of the same phrase being repeated if it is inside another phrase. For instance, both *a black dog with a red collar* and *a black dog* are extracted from the same sentence. From another sentence, *a black dog in shallow water* and *a black dog* maybe extracted. Frequencies of these NPs are later used to choose the best candidate phrase to be used in a description and this means the repeated phrases such as *black dog* in this example will have higher scores in the generation process.

We realize the generation with four sentence templates, namely; NP, NP+PP, NP+VP and NP+VP+PP. In particular, we generate four different sentences using these templates, and then select the one that scores closest to the collected descriptions. The score is computed by using the following equation:

$$scr_s = \frac{\sum\limits_{c \in s} f_c}{l_s} + \frac{\sum\limits_{h \in c} f_h}{l_c} \tag{2}$$

The first part of Equation (2) uses the phrase frequency score $f_c$ for the phrase or chunk $c$

in sentence $s$. Frequencies of the phrases (NP, VP and PP only) are divided by the total of sum of each individual category, then all scores are normalized, so that they are between 0 and 1. Therefore, each type is given equal weight in the combination. The second part of Equation (2) reflects the dependencies between the phrases. For this, we consider the co-occurrences of the heads of phrases. This is because phrases are less likely to be seen in pairs frequently. The heads of phrases are extracted in a simplified way as follows. For NPs and PPs the last word is the head, and for VPs the first word except auxiliaries is the head. $h$ denotes a pair of head words in the set $c$ which contains all the pairs in the sentence $s$. The co-occurrence frequencies in this part are also normalized to be in the 0-1 interval. $l_s$ and $l_c$ show the lengths of the sets or the number of phrases and head pairs respectively.

*3.2.2. Word-based Language Models (LM-WORD):* The second generation method is rule-based and it uses a language model. First, important objects and actions in the images are extracted. For this purpose, frequencies of the words that are seen in the retrieved sentences are used. The most frequent 10 nouns and verbs are selected as keywords or representatives of subjects, objects and verbs after cleaning determiners, auxiliaries, adjectives and so on.

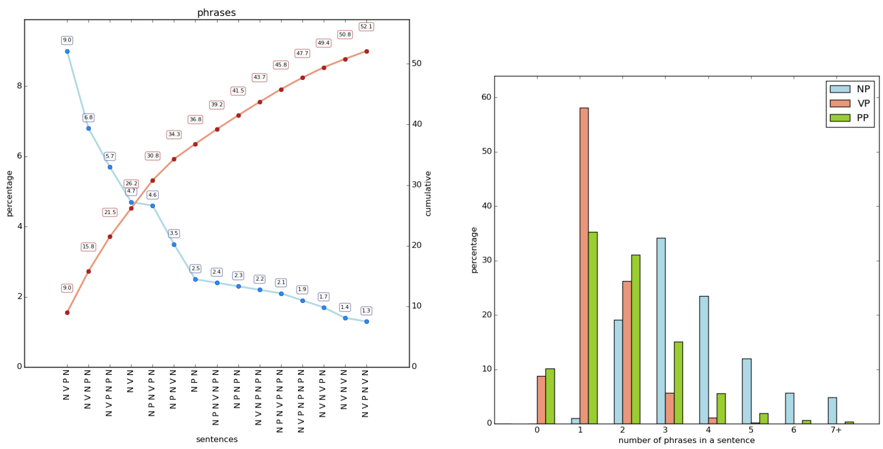
Using these keywords, preceding and following words are added with the help of the most common bi-grams in the language model. Several rules are used to handle determiners, auxiliaries and plurality while generating candidate sentences. The sentence with the highest probability according to the language model is selected.

*3.2.3. Phrase-based Language Models (LM-PHRASE):* LM-PHRASE method uses phrases extracted with shallow parsing (chunking) to build the language model used here. For each query image, language models extracted from the retrieved descriptions of the visually similar images from the same data set are used to generate candidate sentences and the best sentences are selected after reordering them according to their probabilities and their lengths.

A shallow parser is trained in order to chunk the sentences. The problem is considered as a sequence labeling task and Conditional Random Fields [19] are used to build a shallow parser similar to Sha and Pereira [30] with the training data of CoNLL 2000 shared task [34]. The word, its POS tag, whether it contains digits, capital letters, the following and preceding words and their part-of-speech (POS) tags are used as features. POS tags are assigned using the Stanford POS Tagger [35].

Trigram language models are trained after the extraction of phrases. The models use unigram, bigram and trigram probabilities in order to predict the next phrase. Syntactic structures of the sentences are also learned from the same data used for building the language model. The most common syntactic structure and the ones following it down to the 20% of the frequency of the most common one are used as templates for generating sentences. Beam search is used with the language model to generate 5 sentences for each template.

The most common syntactic structures observed in the Flickr30K data set are presented in Figure 5(a). We conjecture that selecting a fixed set of templates for all images is not suitable. Evidence of this is seen in Figure 5(a) . The cumulative percentage of the most common 15 syntactic structures barely cover 50% of all structures. There is not a small set of templates that can cover most of the sentences, therefore, we choose templates for each image individually. The numbers of each type of phrase per sentence are also presented in Figure 5(b). We can see that the sentences contain at least one NP, but, there are sentences without VPs or PPs. In addition, more than half of the sentences contain only one VP. Language models use common syntactic structures extracted for each image and fill them with phrases extracted from retrieved descriptions of visually similar images.




(a) (b)

Fig. 5: (a) Most common syntactic structures in Flickr30K sentences.(b) Percentages as phrase count for individual sentences. Here *N* shows noun phrases, *V* is used for verb and *P* is used for prepositional phrases. The plot shows both individual and cumulative percentages.

$$scr_s = \frac{\ln P(s)}{l_s} - \frac{(l_s - \mu)^2}{2\sigma^2} \tag{3}$$

After generation, sentences are ranked using the score function in Equation (3) also considering the sentence length. In the equation, $P(s)$ shows the probability of the sentence according to the language model, $l_s$ is the number of phrases in the sentence, $\mu$ and $\sigma$ shows the mean and the standard deviation of the number of phrases in the retrieved sentences that the model is extracted. The score is a combination of language model probability and a penalty score based on sentence length. Figure 6 summarizes the flow of the generation process for an example query image.

## 4. Experiments

### *4.1. Datasets*

We evaluate the proposed algorithms on Flickr30K [40] and Flickr8K [14] datasets, which consist of 30K images and 8K images respectively. Each of the images in these datasets are coupled with 5 descriptions that are collected from Amazon Mechanical Turkers. In Flickr30K dataset, ground truth information about the image regions are also available; *i.e.* the corresponding annotations within descriptions for each image region.

We evaluate the performance of captions using BLEU [26], ROUGE [21], METEOR [2] and CIDEr [36] metrics, using the evaluation codes for MS COCO Caption Generation [7].

### *4.2. Experimental Results*

First, we evaluate the performance of the transfer-based methods for description generation. Transfer-based methods simply transfer the description of visually closest image as the description of the query image, and they only differ in the image similarity metrics they utilize.

The first transfer-based method (HYBRID) employs HYBRID-CNN features and transfers the caption of the nearest training image based on $L_2$ distance. BoOP method uses the aforementioned BOOP codeword vectors instead, and employs histogram intersection similarity to find the nearest image. Reranking approach is based on our reranking strategy where top-

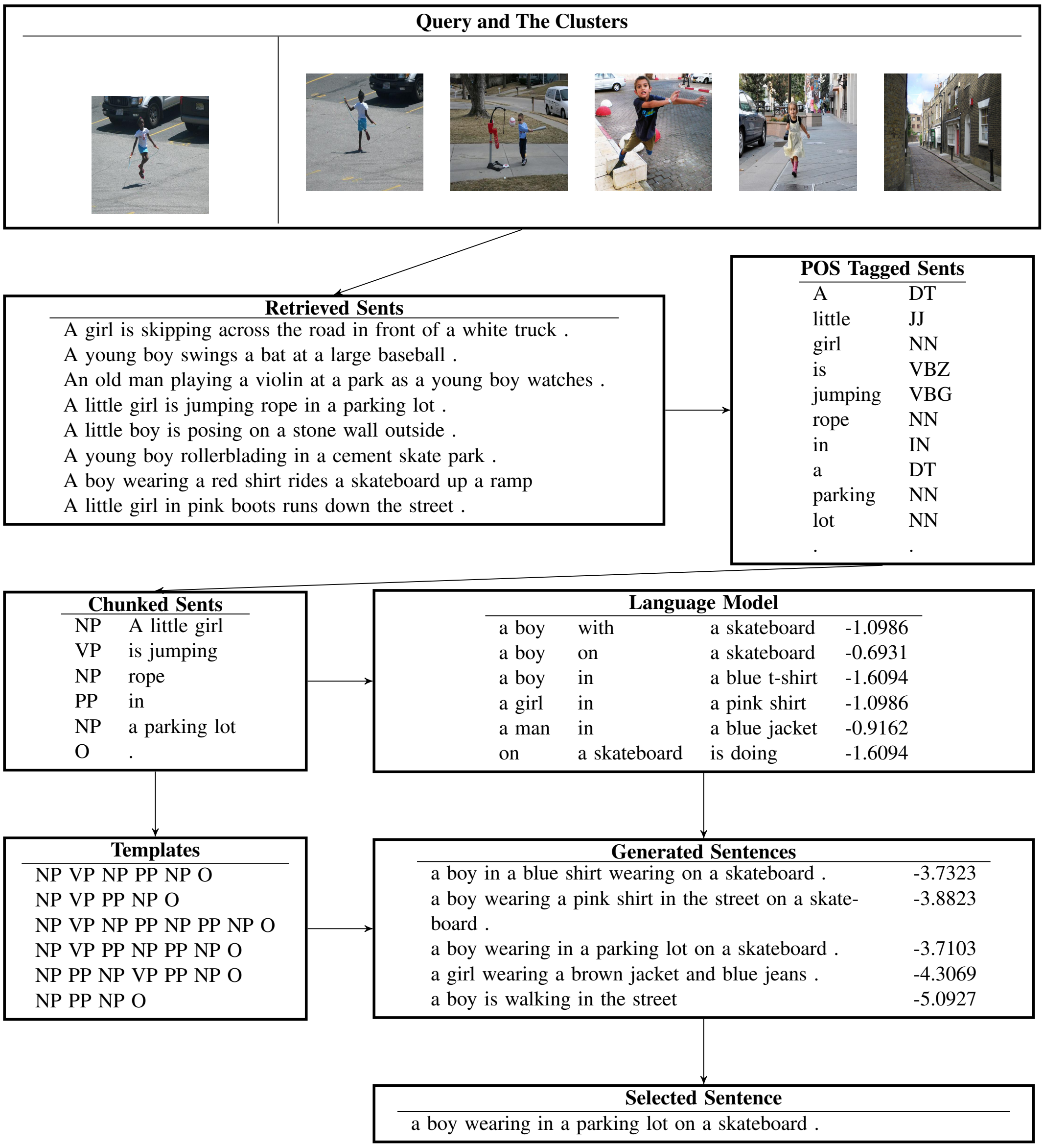



Fig. 6: Steps of the description generation system with phrase based language model (LM-PHRASE)

100 visually closest images from the training set using HYBRID-CNN features are retrieved first, and then reranked using the BoOP descriptors and the VisualRank method. Concatenation method uses a simple rank aggregation which combines HYBRID and BoOP rankings by simple averaging.

Table 1 shows the results. Amongst the four transfer-based alternatives, we observe that concatenating HYBRID with BoOP (denoted with Concatenation) gives the best result with respect to all evaluation metrics, and re-ranking is the second best. We observe that, while singleton BoOP yields inferior results than utilizing HYBRID-CNN features, it provides valuable information, since its combination with HYBRID increases the overall performance.

Contrary to transfer-based approaches, Table 2 presents the results of the generation based description methods using different types of retrieval strategies. Three different sentence generation methods and five retrieval strategies are evaluated. Here, *IOR* refers to the case where we apply saliency estimation procedures to both object-like region sampling methods, namely objectness and R-CNN. For *IOR-5*, saliency estimation is only applied to objectness windows, and

Table 1: Comparison of transfer-based methods on Flickr8K and Flickr30K datasets.

| Dataset | Metric | BoOP | HYBRID | Concatenation | Reranking |
|---|---|---|---|---|---|
| Flickr8K | CIDEr | 16.30 | 18.92 | **19.48** | 19.37 |
| | METEOR | 14.02 | 14.53 | **14.79** | 14.67 |
| | ROUGEL | 34.26 | 35.49 | **35.77** | 35.55 |
| | BLEU | 20.51 | 21.94 | **22.30** | 21.90 |
| Flickr30k | CIDEr | 11.70 | 12.43 | **13.74** | 13.09 |
| | METEOR | 12.15 | 12.60 | **13.11** | 12.86 |
| | ROUGEL | 31.86 | 32.27 | 32.72 | **33.15** |
| | BLEU | 18.33 | 19.15 | **19.94** | 19.60 |

fused with top-5 R-CNN windows, taking R-CNN as references and applying non-maximum suppression. Third retrieval strategy (*IOR-ranked*) apply re-ranking as described above. In this setting, similar to IOR-5, top-5 R-CNN windows are merged with objectness candidates. The procedure where IOR is opted out and the descriptions are used directly, is referred with (*All-ranked*) in Table 2. In all of the methods, we first retrieve an initial set of visually similar images, via N-nearest neighbour retrieval with HYBRID-CNN features, where $N = 100$.

According to the results in Table 2, we observe that, amongst different retrieval strategies, IOR-5 achieves the highest performance in Flickr30K dataset, whereas All-ranked strategy is better for Flickr8K dataset. This performance difference is likely due to the difference in the size of the data; local analysis benefits from the presence of larger amount of data. Flickr8K has a more compact set of queries (dogs running, children playing), while Flickr30K is more unconstrained. For Flickr30K, some queries are ambiguous, having only 2-5 similar images. The novelty of our algorithm arises for this type of queries: we eliminate all irrelevant images though they are the dominant set in the initial retrieval stage.

We evaluate the proposed image retrieval methods with three generation approaches with increasing order of complexity, namely (PHRASE, LM-WORD, LM-PHRASE) in Table 2. Among these three generation models, LM-PHRASE consistently performs better than the other two. Rouge scores of PHRASE is higher than LM-WORD, but the rest of the evaluation metrics suggest that LM-WORD is better than PHRASE.

Figure 7 shows sample generation results using IOR-5 retrieval strategy along with LM-PHRASE generation method. Finally, we compare our method against data-driven approaches in the literature in Table 3 and also against human performance. Our method competes well with [39] on Flickr8K dataset and outperforms their method on Flickr30K on three metrics, indicating the superiority of our method. Our method also outperforms Visually Closest approach [24], indicating the benefit of our generic local and textual analysis.

## 5. Conclusion

In this paper, we propose data-driven methods to transfer and generate natural descriptions of images. First, we explore ways of retrieving relevant images from a large set of images by combining deep image features with an object-based local representation. Our results show that taking into account both scene and object information is crucial for transfer-based image captioning. Second, we propose a sentence generation framework based on a novel phrase selection paradigm, where we utilize the phrases collected from the captions of retrieved images. The semantically related phrases are identified by a novel clustering-based method which jointly discovers important salient image regions and parses the corresponding captions. For sentence generation, we investigate three different generation methods which employ combina-

Table 2: Performances of retrieval strategies (*IOR, IOR-5, IOR-Ranked, All-Ranked, Oracle*) and generation models (*PHRASE, LM-WORD, LM-PHRASE*) on Flickr8K and Flickr30K.

| | | Flickr8K | | | | Flickr30K | | | |
|---|---|---|---|---|---|---|---|---|---|
| RETRIEVAL STRATEGY | GENERATION MODEL | CIDEr | METEOR | ROUGE | BLEU | CIDEr | METEOR | ROUGE | BLEU |
| IOR | PHRASE | 18.71 | 13.19 | 35.59 | 23.72 | 14.94 | 12.38 | 33.67 | 22.39 |
| | LM-WORD | 22.39 | 14.13 | 32.60 | 28.51 | 16.79 | 12.16 | 31.83 | 27.20 |
| | LM-PHRASE | 24.94 | 16.84 | 38.62 | 28.82 | 18.90 | 15.32 | 36.30 | 27.91 |
| IOR-5 | PHRASE | 20.19 | 13.44 | 36.20 | 24.78 | 15.24 | 12.51 | 33.88 | 23.13 |
| | LM-WORD | 24.13 | 14.53 | 33.22 | 29.92 | 18.03 | 12.41 | 31.90 | 27.47 |
| | LM-PHRASE | 24.14 | 16.74 | 38.64 | 28.68 | **20.13** | **15.38** | **36.75** | **28.73** |
| IOR-Ranked | PHRASE | 21.82 | 13.89 | 36.79 | 25.70 | 11.18 | 11.68 | 32.89 | 20.85 |
| | LM-WORD | 22.96 | 14.43 | 32.69 | 28.65 | 14.35 | 11.84 | 31.40 | 26.26 |
| | LM-PHRASE | 26.55 | 16.92 | 38.58 | 29.20 | 15.94 | 14.63 | 35.67 | 27.12 |
| All-Ranked | PHRASE | 26.25 | 14.34 | 38.02 | 27.64 | 12.62 | 11.64 | 32.23 | 21.54 |
| | LM-WORD | 25.45 | 14.84 | 32.98 | 29.90 | 15.42 | 12.11 | 31.90 | 27.05 |
| | LM-PHRASE | **29.12** | **17.62** | **40.01** | **30.58** | 17.13 | 14.93 | 36.62 | 28.49 |


Fig. 7: Qualitative examples. The first two rows include successful sentences generated by our model, whereas the last row demonstrate cases of failure. As can be seen, the proposed approach is quite competent at generating relevant and high quality captions. However, due to the noise in the object identification process, some captions can be off track.

Table 3: Comparison with state-of-the-art. **Bold** denotes the best performing method while second best method is underlined.

| | Flickr8K | | | | Flickr30K | | | |
|---|---|---|---|---|---|---|---|---|
| | CIDEr | METEOR | ROUGE | BLEU | CIDEr | METEOR | ROUGE | BLEU |
| OUR-BEST | <u>29.12</u> | **17.62** | <u>40.01</u> | **30.58** | <u>20.13</u> | **15.38** | **36.75** | **28.73** |
| QE [39] | **31.04** | <u>17.04</u> | **40.90** | <u>25.67</u> | **20.26** | <u>15.18</u> | <u>33.89</u> | <u>23.31</u> |
| MC-KL [23] | 14.85 | 16.31 | 30.96 | 18.83 | 6.73 | 14.44 | 27.91 | 15.44 |
| MC-SB [23] | 26.71 | 14.00 | 32.34 | 24.42 | 20.40 | 12.70 | 32.47 | 24.72 |
| VC [24] | 20.02 | 13.65 | 36.73 | 20.57 | 13.64 | 11.74 | 28.20 | 18.88 |
| HUMAN | 82.11 | 25.52 | 53.08 | 40.79 | 66.44 | 23.21 | 45.27 | 37.32 |

tions of rule based, template based and probability driven language models. This whole pipeline results in relatively realistic and accurate phrase/word usages and grammatical constructions. Experiments demonstrate that our combined framework provides much more effective results compared to the transfer-based image description approaches. We believe that with larger data sets, a larger number of similar images can be retrieved, yielding more accurate phrases to be generated.

## Acknowledgements

This research was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), Scientific and Technological Research Projects Funding Program Award 113E116.

## 6. References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE TPAMI*, 34(11):2189–2202, 2012.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, volume 29, pages 65–72, 2005.

[3] Alexander C Berg, Tamara L Berg, Hal Daume III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569, 2012.

[4] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *JAIR*, 55:409–442, 2016.

[5] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. *IEEE TIP*, 24(12):5706–5722, 2015.

[6] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[9] Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[10] Albert Gatt and Ehud Reiter. Simplenlg: A realisation engine for practical applications. In *ENLG*, pages 90–93, 2009.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[12] Ankush Gupta, Yashaswi Verma, and CV Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

[13] Zeeshan Hayder, Mathieu Salzmann, and Xuming He. Object co-detection via efficient inference in a fully-connected crf. In *ECCV*, pages 330–345, 2014.

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, pages 853–899, 2013.

[15] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, (11):1254–1259, 1998.

[16] Yushi Jing and Shumeet Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE TPAMI*, 30(11):1877–1890, 2008.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[18] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *ACL*, pages 359–368, 2012.

[19] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[20] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *CoNLL*, pages 220–228, 2011.

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: ACL Workshop*, volume 8, 2004.

[22] Ran Margolin, Avishay Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *CVPR*, pages 1139–1146, 2013.

[23] Rebecca Mason and Eugene Charniak. Nonparametric Method for Data-driven Image Captioning. In *ACL*, 2014.

[24] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, et al. Large scale retrieval and generation of image descriptions. *IJC*, 119(1):46–59, 2013.

[25] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[27] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *IJCV*, 108(1-2):59–81, 2014.

[28] Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE TPAMI*, 29(1):167–172, 2007.

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. *EMNLP*, 2014.

[30] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *NAACL HLT*, pages 134–141, 2003.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[32] Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. Parsing with compositional vector grammars. In *ACL*, 2013.

[33] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *CVPR*, pages 1464–1471, 2014.

[34] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *LLL and CoNLL*, ConLL '00, pages 127–132, 2000.

[35] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL HLT*, pages 173–180, 2003.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*, 2014.

[37] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[38] Chen Xiaozhi, Huimin Ma, Xiang Wang, and Zhichen Zhao. Improving object proposals with multi-thresholding straddling expansion. In *CVPR*, 2015.

[39] Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakici. A Distributed Representation Based Query Expansion Approach for Image Captioning. In *ACL*, 2015.

[40] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[41] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR*, pages 739–746, 2013.

[42] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.