Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Choosing Linguistics over Vision to Describe Images

Ankush Gupta, Yashaswi Verma, C. V. Jawahar
International Institute of Information Technology, Hyderabad, India - 500032
{ankush.gupta@research., yashaswi.verma@research., jawahar@}iiit.ac.in

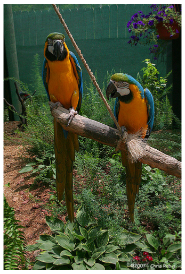
Abstract

In this paper, we address the problem of automatically generating human-like descriptions for unseen images, given a collection of images and their corresponding human-generated descriptions. Previous attempts for this task mostly rely on visual clues and corpus statistics, but do not take much advantage of the semantic information inherent in the available image descriptions. Here, we present a generic method which benefits from all these three sources (i.e. visual clues, corpus statistics and available descriptions) simultaneously, and is capable of constructing novel descriptions. Our approach works on syntactically and linguistically motivated phrases extracted from the human descriptions. Experimental evaluations demonstrate that our formulation mostly generates lucid and semantically correct descriptions, and significantly outperforms the previous methods on automatic evaluation metrics. One of the significant advantages of our approach is that we can generate multiple interesting descriptions for an image. Unlike any previous work, we also test the applicability of our method on a large dataset containing complex images with rich descriptions.

1 Introduction

An image can be described either by a set of keywords (Guillaumin et al. 2009; Feng , Manmatha, and Lavrenko 2004; Makadia, Pavlovic, and Kumar 2010), or by a higher level structure such as a phrase (Sadeghi and Farhadi 2011) or sentence (Aker and Gaizauskas 2010; Kulkarni et al. 2011; Yang et al. 2011). The keyword based approach is inspired by the web search engines but has its own limitations, e.g. an image tagged with $\{black, car, dog\}$ does not convey the complete meaning (whether it has a *black car* and *dog*, or a *car* and a *black dog*, and what are their state(s) and relative position); whereas, the sentence *“a dog is sitting on a black car”* implicitly encodes the relationships between words.

Previous attempts for generating descriptions for unseen images (Kulkarni et al. 2011; Yang et al. 2011; Li et al. 2011; Farhadi et al. 2010; Ordonez, Kulkarni, and Berg 2011) rely mostly on few object detectors (a detector locates in an image one or more instances of a specific semantic category), classifiers and corpus statistics, but do not utilize the semantic information encoded in available descriptions of images. Either they use these descriptions to restrict the set of prepositions/verbs (Kulkarni et al. 2011; Yang et al. 2011; Li et al. 2011; Yao et al. 2008), or pick one or more complete sentences and transfer them to a test image unaltered (Farhadi et al. 2010; Ordonez, Kulkarni, and Berg 2011). While the former may result in quite verbose and non-humanlike descriptions, in the latter it is very unlikely that a retrieved sentence would be as descriptive of a particular image as a generated one (Kulkarni et al. 2011). This is because a retrieved sentence is constrained in terms of objects, attributes and spatial relationship between objects; whereas a generated sentence can more closely associate the semantics relevant to a given image (Figure 1).



§ “This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the road.” (Kulkarni et al. 2011)

§ “The person is showing the bird on the street.” (Yang et al. 2011)

§ “Black women hanging from a black tree. Colored man in the tree.” (Li et al. 2011)

§ **“An American eagle is perching on a thick rope.” (Ours)**

Figure 1: Descriptions generated by four different approaches for an example image from the UIUC Pascal sentence dataset.

Image descriptions not only contain information about the different objects present in an image, but also tell about their states and spatial relationships. Even for complex images, this information can be easily extracted, hence leveraging the gap between visual perception and semantic grounding. With this motivation, we present a generative approach that gives emphasis to textual information rather than just relying on computer vision techniques. Instead of using object detectors, we *estimate* the content of a new image based on its similarity with available images. To minimize the impact of encountering noisy and uncertain visual inputs, we extract syntactically motivated patterns from known descriptions and use only those for composing new descriptions. Extracting dependency patterns from descriptions rather than us-

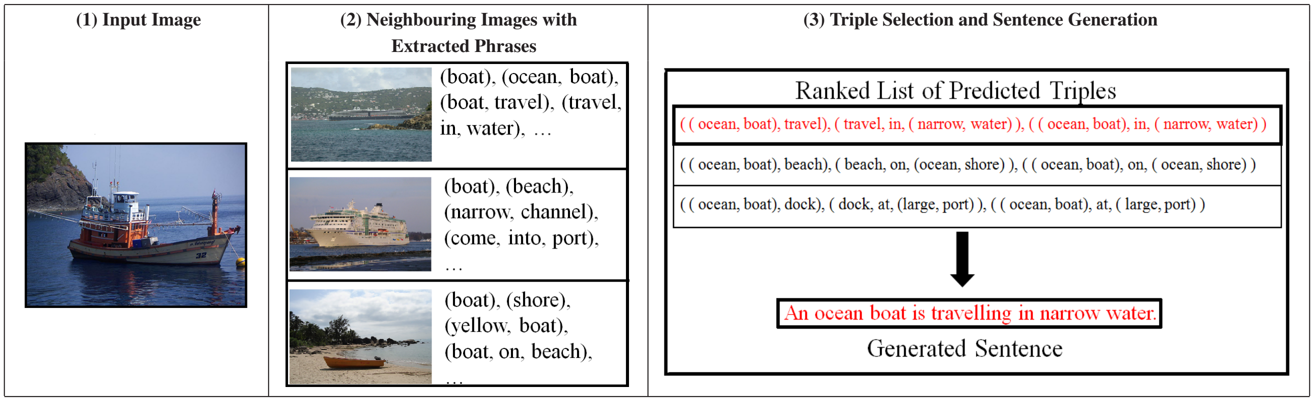

(1) Input Image
(2) Neighbouring Images with Extracted Phrases
(boat), (ocean, boat), (boat, travel), (travel, in, water), …
(boat), (beach), (narrow, channel), (come, into, port), …
(boat), (shore), (yellow, boat), (boat, on, beach), …
(3) Triple Selection and Sentence Generation
Ranked List of Predicted Triples
((ocean, boat), travel), (travel, in, (narrow, water)), ((ocean, boat), in, (narrow, water))
((ocean, boat), beach), (beach, on, (ocean, shore)), ((ocean, boat), on, (ocean, shore))
((ocean, boat), dock), (dock, at, (large, port)), ((ocean, boat), at, (large, port))
An ocean boat is travelling in narrow water.
Generated Sentence

Figure 2: Overview of our approach for an example image from the Pascal dataset. (1) Given an unseen image, (2) find K images most similar to it from the training images, and using the phrases extracted from their descriptions, (3) generate a ranked list of *triples* which is then used to compose description for the new image.

ing it as an n-gram language model is inspired by Aker and Gaizauskas (2010). These patterns have a predefined structure (e.g. $(subject, verb)$, $(attribute, subject)$, etc), and can easily be mapped to generate a syntactically and grammatically correct description.

The main strength of our approach is that it works on these patterns which carry a bigger chunk of information, compared to predicting individual bits such as objects, attributes, verb, preposition, etc. in a piece-wise manner and then combining them at a later stage as done by previous methods.

To summarize, our contributions are: (i) A novel approach for generating human-like descriptions for images that effectively uses available textual information. Instead of relying on trained object detectors or classifiers, our method captures the semantics of an image using the information encoded in its description. (ii) Extensive evaluations to test the applicability of our model on the IAPR TC-12 benchmark[1] (to our best knowledge, this is the first study devoted to compose descriptions for complex images with rich and complicated descriptions). (iii) Producing state-of-the-art performance on the Pascal sentence data set[2], and setting a baseline for the IAPR TC-12 benchmark.

2 Related Work

Generating natural language descriptions for images is an emerging area of research with few attempts directly addressing this problem. Most of these approaches (Kulkarni et al. 2011; Yang et al. 2011; Li et al. 2011; Ordonez, Kulkarni, and Berg 2011; Farhadi et al. 2010) rely on trained detectors and classifiers to determine the content of an image. Some of these approaches (Kulkarni et al. 2011; Yang et al. 2011; Li et al. 2011) explore the use of corpus to smooth the errors in the detections. In Yang et al. (2011), these detections are used as parameters of an HMM where the hidden nodes are related to the sentence structure and the emissions are related to the image detections. In Kulkarni et al. (2011), a CRF-based model; whose nodes correspond to image entities (such as objects, attributes and prepositions); is used to predict the best labelling for an image. For sentence generation, either templates (Kulkarni et al. 2011; Yang et al. 2011; Yao et al. 2008) are used, or complete sentences from the available descriptions (Farhadi et al. 2010; Ordonez, Kulkarni, and Berg 2011) are transferred.

Aker and Gaizauskas (2010) use GPS meta data to access web-documents relevant to an image, and generate image description by summarizing these documents. But their domain is limited to static objects such as buildings and mountains, and cannot be applied to dynamic objects in daily life like people, cars, etc. In Feng and Lapata (2010), assuming that a relevant document is available for a given image, the output (keywords) of an image annotation model is combined with the document properties to generate captions for images in the news domain.

Conceptually, our work closely relates to Sadeghi and Farhadi (2011). They hypothesize that a *visual phrase* (e.g. *"person riding horse"*) is more meaningful than individual objects. To detect phrases, they use specific phrase detectors trained on few hundreds of images. In contrast, we extract this information from image descriptions. Also, we work on significantly larger number of phrases compared to them.

3 Our Approach

Given a dataset of images and their corresponding human-generated descriptions, our task is to describe an unseen image. We extract linguistically motivated phrases from these descriptions; and given a new image, the phrases present in its neighbouring images are ranked based on image similarity. These are then integrated to get triples of the form $(\ ((attribute_1, object_1), verb),\ (verb, prep, (attribute_2, object_2)), (object_1, prep, object_2)\ )$, which are used for sentence generation (Figure 2).

3.1 Phrase Extraction

The key component of our approach is to effectively use the information in the ground-truth descriptions, and for that we

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

[2] http://vision.cs.uiuc.edu/pascal-sentences/

Sentence	Phrases w/o Synonym	Phrases w/ Synonym
A black and white pug is looking at the camera.	$\left(\textbf{pug}_{(subj)}\right)$, $\left(\textbf{camera}_{(obj)}\right)$, $\left(\textbf{pug}_{(subj)}, \textbf{look}_{(verb)}\right)$, $\left(\textbf{black}_{(attr)}, \textbf{pug}_{(subj)}\right)$, $\left(\textbf{white}_{(attr)}, \textbf{pug}_{(subj)}\right)$, $\left(\textbf{look}_{(verb)}, \textbf{at}_{(prep)}, \textbf{camera}_{(obj)}\right)$	$\left(\textbf{dog}_{(subj)}\right)$, $\left(\textbf{camera}_{(obj)}\right)$, $\left(\textbf{dog}_{(subj)}, \textbf{look}_{(verb)}\right)$, $\left(\textbf{black}_{(attr)}, \textbf{dog}_{(subj)}\right)$, $\left(\textbf{white}_{(attr)}, \textbf{dog}_{(subj)}\right)$, $\left(\textbf{look}_{(verb)}, \textbf{at}_{(prep)}, \textbf{camera}_{(obj)}\right)$

Table 1: Example sentence with extracted phrases. In "Phrases w/ Synonym", 'pug' is replaced by its most frequently used form 'dog' (determined using WordNet).

need to extract relation tuples from text automatically. For this purpose, we process the descriptions using the Stanford CoreNLP toolkit[3]. We use *"collapsed-ccprocessed-dependencies"* which is intended to be more useful for relation extraction task (Marneffe and Manning 2008), as dependencies involving prepositions and conjuncts are collapsed to reflect direct relation between content words.

We extract syntactically motivated phrases from human-generated image descriptions, and map each sentence to a list of phrases like $(subject, verb)$, $(object, verb)$, $(verb, prep, object)$, etc. We look at an image as a collection of such phrases in the visual domain, and hypothesize that similar appearing images have identical phrases. Previous approaches obtain relations of the form $(object, action, scene)$ (Farhadi et al. 2010), $(object_1, object_2, verb, scene, prep)$ (Yang et al. 2011), or $((attribute_1, object_1), prep, (attribute_2, object_2))$ (Kulkarni et al. 2011) by combining the outputs of individual detectors with some heuristic and/or corpus statistics to predict the involved action/preposition(s). But such predictions can be quite noisy (e.g., $(person, under, road)$ (Kulkarni et al. 2011)), resulting in absurd sentences. In contrast, our phrases implicitly encode ordering preference information, and hence generate semantically meaningful descriptions.

In practice, we extract 9 distinct types of phrases from human-generated descriptions : $(subject)$, $(object)$, $(subject, verb)$, $(object, verb)$, $(subject, prep, object)$, $(object, prep, object)$, $(attribute, subject)$, $(attribute, object)$, $(verb, prep, object)$. Each noun (subject/object) is expanded up to at most 3 hyponym levels using its corresponding WordNet synsets. To explore the possibilities of generating varied and interesting descriptions, we also experiment without considering synonyms (e.g. Figure 3, row 3). Table 1 shows phrases extracted from a sample sentence.

3.2 Image Features

We assign each image a unique signature using a set of global (colour, texture and scene) and local (shape) features similar to Makadia, Pavlovic, and Kumar (2010) and Guillaumin et al. (2009). Each feature is represented by a multi-dimensional vector. The colour features include histograms in the RGB and HSV colourspaces. While RGB is the standard colourspace used in the digital displays, HSV encodes some visual properties important for humans such as brightness and colour saturation. To extract texture properties, we use Gabor and Haar descriptors. For scene characteristics, we use the GIST feature. This feature entails a set of perceptual properties such as roughness, ruggedness, naturalness, etc. Finally, for shape we use the SIFT descriptors. These are well-known for extracting shape properties that are invariant to object size, scale, translation, rotation and illumination. In order to extract some information about the spatial layout of an image, we also compute all but the GIST features over 3 equal horizontal as well as vertical partitions of an image, which are then concatenated into a single vector. We found that such features were useful in distinguishing between images which differ in their spatial layout (e.g., the images in the first and third column of Figure 3). To compute distance between two feature vectors, we use L_1 distance for colour, L_2 for texture and scene, and χ^2 for shape features.

3.3 Google Counts

In data-driven learning techniques, text corpus is employed to estimate the statistical behaviour of different n-grams. In our case, the number and diversity of phrases is huge, and it is unlikely to predict their general behaviour using only the available descriptions. To address this, we use the number of approximate search results reported by Google for an exact match query on each phrase similar to Kulkarni et al. (2011).

3.4 Model for Predicting Phrase Relevance

Let $\mathcal{T}$ be the set of images with their descriptions and $\mathcal{Y}$ be the set of distinct phrases extracted from these descriptions (Section 3.1). Each image $J \in \mathcal{T}$ is represented by a set of feature vectors $\{f_{1,J}, \ldots, f_{n,J}\}$ (Section 3.2), and associated with phrases $Y_J \subset \mathcal{Y}$ computed from its description.

Given any test image I, our goal is to determine the joint probability $P(y_i, I)$ of associating a phrase $y_i \in \mathcal{Y}$, $\forall i \in \{1, \ldots, |\mathcal{Y}|\}$ with it. As per our hypothesis, an image inherits the characteristics of images similar to it. This similarity is defined based on distance of I from any other image J in the feature space, which is calculated as

$$D_{I,J} = w_1 d_{1I,J} + \ldots + w_n d_{nI,J} = \mathbf{w} \cdot \mathbf{d}_{I,J}, \tag{1}$$

where $D_{I,J}$ is the distance between image I and J, $d_{1I,J}, \ldots, d_{nI,J}$ are real-valued base distances between the corresponding features of both images (using distance functions discussed in Section 3.2), and $w_1, \ldots, w_n$ are the (positive) real-valued weights in their linear combination.

If $\mathcal{T}_I^K \subset \mathcal{T}$ is the set of $K \leq |\mathcal{T}|$ nearest neighbours of I obtained using Eq. 1, we define $P(y_i, I)$ as (Jeon, Lavrenko,

[3] http://nlp.stanford.edu/software/corenlp.shtml

(object$_1$)	(attribute$_1$, object$_1$)	(object$_1$, verb)	(verb, prep, object$_2$)	(object$_2$)	(attribute$_2$, object$_2$)	(object$_1$, prep, object$_2$)
boat	ocean ship	boat come	travel in water	port	mountain village	boat at port
ocean	ocean boat	boat travel	sit on water	sea	tropical beach	ship near coast
channel	lone boat	ship position	come into port	water	ocean shore	boat in water
riverboat	canal boat	boat beach	sail in water	ocean	narrow water	boat in city
ship	vintage ship	boat sail	park at harbor	coast	dirty shore	boat in river

⬇

((ocean, boat), travel), (travel, in, (narrow, water)), ((ocean, boat), in, (narrow, water))

Table 2: An illustration of the Phrase Integration algorithm.

and Manmatha 2003):

$$P(y_i, I) = \sum_{J \in \mathcal{T}_I^K} P_{\mathcal{T}}(J) P_{\mathcal{F}}(I|J) P_{\mathcal{Y}}(y_i|J). \tag{2}$$

Our prediction model constitutes of three parts: $P_{\mathcal{T}}$, $P_{\mathcal{F}}$ and $P_{\mathcal{Y}}$. $P_{\mathcal{T}}(J)$ denotes the probability of selecting some image $J \in \mathcal{T}_I^K$. We model it as a uniform prior, with $P_{\mathcal{T}}(J) = \frac{1}{K}$. $P_{\mathcal{F}}(I|J)$ is the density function that gives the likelihood of *generating* image I given its neighbouring image J. This is defined as a function of distance between the two images:

$$P_{\mathcal{F}}(I|J) = \frac{exp(-D_{I,J})}{\sum_{J' \in \mathcal{T}_I^K} exp(-D_{I,J'})}. \tag{3}$$

Such a definition of image similarity provides two advantages: first it smoothly varies with the distance; and second, it helps in optimizing the weight vector $\mathbf{w}$ (Section 3.5). In our experiments, we also tried other similarity measures (such as $(1 - D_{I,J})$) but did not achieve any improvements.

$P_{\mathcal{Y}}(y_i|J)$ is formulated as a multiple-Bernoulli distribution (Feng , Manmatha, and Lavrenko 2004) as

$$P_{\mathcal{Y}}(y_i|J) = \frac{\mu_i \delta_{y_i,J} + N_i}{\mu_i + N}. \tag{4}$$

Here, $\delta_{y_i,J} = 1$ if the phrase y_i is present among the phrases associated with J and zero otherwise, N_i is the Google count of the phrase y_i, N is the sum of Google counts of all phrases of the same type as that of y_i (Section 3.3), and $\mu_i > 0$ is a smoothing parameter. For $\mu_i \ll N_i$, $P_{\mathcal{Y}}(y_i|J) \approx \frac{N_i}{N}$ and hence depends only on the Google counts. Whereas, for $\mu_i \gg N$, $P_{\mathcal{Y}}(y_i|J) \approx \delta_{y_i,J}$ which means that only the information in the image descriptions is relied upon. We will discuss in Section 3.5 how to optimally determine μ_i to make the best use of both these sources of textual information.

3.5 Parameter Learning

The two types of parameters in our phrase prediction model are feature weights w_k's (Eq. 1) and smoothing weights μ_l's (Eq. 4). Given an image I with its computed phrases $Y_I \subset \mathcal{Y}$, we want to learn these parameters such that (i) the probability of predicting any phrase not in Y_I should be minimized, and (ii) the probability of predicting phrases present in Y_I should be greater than any other phrase. With this aim, we define the error function as follows:

$$e = \sum_{I,y_j} P(y_j, I) + \lambda \sum_{(I,y_j,y_i) \in \mathcal{M}} (P(y_j, I) - P(y_i, I)). \tag{5}$$

Here, $y_i \in Y_I$, $y_j \in \mathcal{Y} \setminus Y_I$, $\mathcal{M}$ is the set of all triplets of the form (I, y_j, y_i) which violate the second condition stated above, and $\lambda > 0$ takes care of the trade-off between the two competing error terms. To estimate the parameters (i.e., $\mathbf{w}$ and μ_l's), we use a gradient descent method. Note that the set $\mathcal{M}$ is not fixed and may change at each iteration. In practice, we determine the parameters $\mathbf{w}$ and μ_l's $\forall l \in \{1, \ldots, |\mathcal{Y}|\}$ in an alternating manner.

3.6 Phrase Integration

Using Eq. 2, we get probability scores for all the phrases in $\mathcal{Y}$ for a given test image I. We integrate these phrases to get triples of the form $t = \{ ((attribute_1, object_1), verb)$, $(verb, prep, (attribute_2, object_2))$, $(object_1, prep, object_2$) } (Table 2). Score of each triple is calculated as

$$t_{score} = \prod_{y_i \in S_t} P(y_i, I). \tag{6}$$

where $S_t = \{$ $(object_1)$, $(attribute_1, object_1)$, $(object_1, verb)$, $(verb, prep, object_2)$, $(object_2)$, $(attribute_2, object_2)$, $(object_1, prep, object_2)$ $\}$. During integration, we look for matching elements in different phrases, as is apparent by the indices; e.g. $(object_1)$, $(attribute_1, object_1)$, $(object_1, verb)$, and $(object_1, prep, object_2)$ have the same object. This restricts the number of feasible triples. These triples are then used for sentence generation. Note that the distinction between object and subject (Section 3.1) will make no difference here; but during generation they are treated separately.

3.7 Sentence Generation

The output of our phrase integration step is a ranked list of triples. One major challenge in generation is to determine the appropriate content. While Yang et al. (2011) perform content selection to deal with noisy inputs from detectors, Li et al. (2011) use n-gram frequencies for correct ordering preference of words. In our approach, a triple consists of the phrases extracted from the *human-generated* descriptions.

Hence all such phrases are likely to be clean and relevant, and so we require no explicit content selection or word reordering. Once we have determined the content of our sentences (i.e. triple), the task of generation is to frame it into a grammatical form and output the result.

For the Pascal dataset, we use the triple with the highest score (Eq. 6) for generation. As the available descriptions for IAPR TC-12 images contain multiple sentences, we pick the top 3 triples; and instead of generating a separate sentence for each triple, we aggregate them by applying Syntactic Aggregation (Reape and Mellish 1999) to enhance text fluency and readability. For aggregation, the following two rules are used: (a) the subject grouping rule, and (b) the predicate grouping rule. The descriptions shown in Figure 4 are generated by applying aggregation. The description generated without aggregation for the first image will be "In this image, a *dark-skinned person* is climbing with a pickaxis. The *dark-skinned person* is posing with a green chile. A green garden is surrounded with a striped chair." Aggregation combines the actions associated with '*dark-skinned person*' in a single sentence, hence making the description more readable. (The datasets are discussed in Section 4.1.)

For *Linguistic Realisation*, we use SimpleNLG (Gatt and Reiter 2009). It is a surface realizer for a simple grammar and has significant coverage of English syntax and morphology. As our triples have a syntactically and linguistically motivated structure, their mapping to a sentence is straightforward using SimpeNLG. It offers several advantages for our task. These include setting various features such as tense (verb), voice (sentence), aspect (verb), etc. Though the sentences generated in our case follow a template-like structure, use of SimpleNLG saves the manual effort of writing individual templates. Note that our sentence generation approach is domain independent unlike Yao et al. (2008) which requires domain specific hand-written grammar rules.

4 Experiments

4.1 Datasets

We use the UIUC Pascal Sentence dataset and the IAPR TC-12 Benchmark to test the performance of our approach. The UIUC Pascal sentence dataset (Rashtchian et al. 2010) was first used by Farhadi et al. (2010) for the image description task, and since then it has been used as a test-bed by most of the previous methods addressing the same problem. It comprises of $1,000$ images each containing 5 independent human-generated sentences.

The IAPR TC-12 benchmark was first published for cross-language retrieval by Grubinger et al. (2006). It has $20,000$ images each captioned with free-flowing text of up to 5 sentences. Contrary to the Pascal dataset, it has varied images with large number of distinct object categories and complicated descriptions which makes it an extremely challenging dataset for our task.

4.2 Experimental Details

Similar to Yang et al. (2011), we partition the dataset into 90% training and 10% testing set to determine the parameters (for each dataset). This is repeated to generate results for all the images. Note that Kulkarni et al. (2011) use a different partitioning. They rely on object detectors which are trained using thousands of images, whereas our method uses only the available data.

Dataset	B-1	B-2	B-3	Rouge-1
Pascal w/ syn.	0.41	0.11	0.02	0.28
Pascal w/o syn.	0.36	0.09	0.01	0.21
Pascal Human (std.)	0.64	0.42	0.24	0.50
IAPR TC-12 w/ syn.	0.21	0.07	0.01	0.14
IAPR TC-12 w/o syn	0.15	0.06	0.01	0.11

Table 3: Our automatic evaluation results for sentence generation. Higher score means better performance. B-n means n-gram BLEU score.

For each dataset, we extract all possible phrases from the available descriptions and perform two experiments. In the first experiment, we compute triples from all these phrases and use them to generate image descriptions. In the second experiment, each object/subject in the above phrases is replaced by its synonym (the most frequently used form of the word; determined using WordNet synsets). These phrases are then used to obtain triples and hence image descriptions. Table 1 shows phrases extracted from a sample description. For Pascal dataset, we extract $12,865$ distinct phrases. After considering synonyms, these reduce to $10,429$. Similarly, for IAPR TC-12 dataset, $38,123$ phrases are extracted which map to $29,985$ phrases after using synonyms.

4.3 Evaluation

Automatic Evaluation BLEU (Papineni et al. 2002) and Rouge (Lin and Hovy 2008) are popular metrics in the field of machine translation and text summarization respectively. These compare system generated sentences w.r.t. human generated sentences. As our task can be viewed as summarizing an image and translating it into text, we use these metrics to evaluate our approach. We consider the BLEU n-gram (n=1,2,3) and Rouge-1 precision scores because the descriptions generated are short. Since there is a large scope of variation in description of the same image by different people as compared to translating or summarizing text, these two metrics could penalize many correctly generated descriptions. However, we report our results on these metrics as a standard evaluation method in Table 3. For Pascal dataset, we also show the average BLEU and Rouge scores using the available human generated descriptions for each image in a leave-one-out manner.

Human Evaluation To quantify the aspects that are not addressed by automatic evaluation metrics, human evaluation becomes necessary for our problem. We collect human judgements on 100 and 500 images from the Pascal and IAPR TC-12 datasets respectively. Two aspects are verified in human evaluation : *Readability* of descriptions and *Relevance* of (generated) text with given image. Human evaluators assign a score on a likert scale of $\{1, 2, 3\}$ for each aspect per image, where 1 is good, 2 is ok and 3 is bad. We adopt the definition and guideline used by Li et al. (2011):

A car with a canoe on top is parked on the street near a moped.	A brown cat sleeping on a sofa.	A man and woman are posing for the camera.	A man walking a dog on the beach near large waves.	A black and white photo of a glass bottle of Coca Cola.
A black ferrari is parked in front of a green tree.	An adult hound is laying on an orange couch.	A blond woman is posing with an elvis impersonator.	*An osprey is flying over a dirty water.*	*A motor racer is speeding through a splash mud.*
A sporty car is parked on a concrete driveway.	A sweet cat is curling on a pink blanket.	An orange fixture is hanging in a messy kitchen.	*A water cow is grazing along a roadside.*	*A motor person is covering in a splash mud.*

Figure 3: Example images from the Pascal dataset alongwith their descriptions. The descriptions in the second row are (one of the five) human-generated descriptions, in the third row are those generated by our method without considering synonyms, and those in the fourth row after considering synonyms. The ones in *"italics"* are bad results as judged by human evaluators.

Dataset	Readability	Relevance
Pascal w/ syn.	**1.19**	**1.57**
Pascal w/o syn.	1.24	1.76
IAPR TC-12 w/ syn.	**1.38**	**2.32**
IAPR TC-12 w/o syn	1.41	2.55

Table 4: Our human evaluation results for sentence generation. Lower score means better performance.

Approach	B-1	B-2	B-3	Rouge-1
BabyTalk	0.30	-	-	-
Ours	**0.47**	0.19	0.06	0.33
CorpusGuided	0.41	0.13	0.03	0.31
Ours	**0.54**	**0.23**	**0.07**	**0.41**

Table 5: Comparison of our approach with *BabyTalk* (Kulkarni et al. 2011) and *CorpusGuided* (Yang et al. 2011) following their protocol. B-n means n-gram BLEU score.

Readability: How grammatically correct is the generated sentence?

(1) Mostly perfect English phrase or sentence.

(2) There are some errors, but mostly comprehensible.

(3) Terrible.

Relevance: How relevant is the generated description to the given image?

(1) Very relevant.

(2) Reasonably relevant.

(3) Totally off.

Table 4 summarizes our human evaluation results. The scores given by two human evaluators were identical on 82% and 64% of the instances on the two test sets respectively.

4.4 Comparison with Previous Methods

To compare our approach with Kulkarni et al. (2011), we generate triples of the form $((attribute_1, object_1)$, $prep$, $(attribute_2, object_2))$ following their protocol (i.e., by considering same sets of objects, attributes and prepositions). Similarly, triples of the form $((object_1, verb)$, $(verb, prep, object_2)$, $(object_1, prep, object_2))$ are generated to compare with Yang et al. (2011). Results of comparison (using automatic evaluation) are shown in Table 5.

4.5 Discussion

In our experiments (Table 3, Table 4), we found that considering synonyms always perform better than without synonyms. A possible reason is using synonyms reduces the number of distinct phrases, and hence improves learning. As evident in Table 5, our method significantly outperforms Yang et al. (2011) in terms of automatic evaluation measures [4]. Though we cannot compare our approach directly with Kulkarni et al. (2011) as they generate multiple sentences, we still report our scores. Interestingly, even after considering all possible phrases with synonyms (Table 3), we score better or comparable to previous methods. This is because the phrases are extracted from the available descriptions, hence resulting in close-to human descriptions. As the triples used for generation have a syntactically well-defined structure, our sentences generated are mostly grammatically correct. This is the reason why we get high "readability" scores when judged by human evaluators (Table 4).

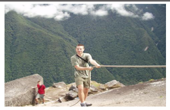
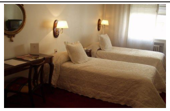
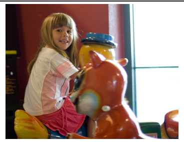
	In this image, a dark-skinned person is climbing with a pick-axis and posing with a green chile. A green garden is surrounded with a striped chair.
	In this image, a king-size bed is made with a white bedcover. A black cabinet is made with a brown telephone and standing on left.

Figure 4: Example images from IAPR TC-12 dataset, and their description generated by our method (w/o synonyms).

Figure 3 and Figure 4 show examples of image descriptions generated by our approach on Pascal and IAPR TC-12 datasets respectively. Note that by considering all phrases (without using synonyms), we are able to generate interest-

[4] Statistically significant at a level of $p < 0.0001$.



(1) A white toddler is playing with a condiment package.
(2) A wide-eyed baby is holding in a highchair.
(3) A female child is holding in a sling.

Figure 5: Illustration of generating multiple interesting descriptions for an example image from the Pascal dataset.

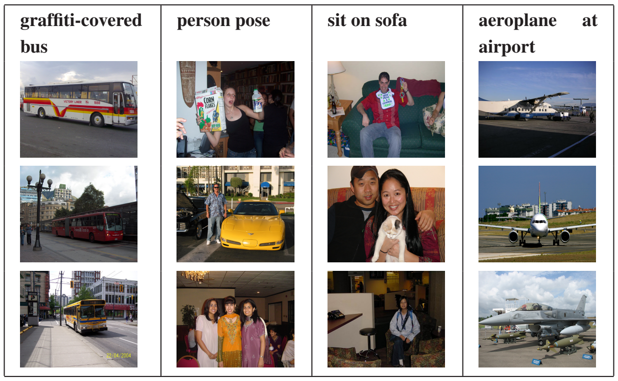


Figure 6: Images annotated with four phrases: (i) “graffiti-covered bus” $(attribute_1, subject_1)$, (ii) “person pose” $(subject_1, verb)$, (iii) “sit on sofa” $(verb, prep, object_2)$, and (iv) “aeroplane at airport” $(subject_1, prep, object_2)$.

ing words like “ferrari”, “hound”, “couch”, “cabinet”, etc.

Interesting Descriptions One of the potential applications of our approach is that we can generate *multiple* interesting descriptions for an image (Figure 5), which none of the previous approaches has focussed on.

Phrase Annotation Another interesting application of our approach is *phrase annotation*; i.e., we can annotate images with phrase(s) instead of just keywords (Makadia, Pavlovic, and Kumar 2010; Guillaumin et al. 2009). This can significantly improve the quality of images retrieved (Figure 6).

5 Conclusion

We proposed a novel approach for generating relevant, fluent and human-like descriptions for images without relying on any object detectors, classifiers, hand-written rules or heuristics. Even with simple Computer Vision and Machine Learning techniques, we achieved significantly better results than state-of-the-art by analyzing and efficiently extracting the semantic information encoded in the image descriptions.

Acknowledgment We would like to thank Prashanth Mannem from LTRC, IIIT Hyderabad for helpful discussions and pointers during the course of this work.

References

Aker, A., and Gaizauskas, R. 2010. Generating image descriptions using dependency relational patterns. In *ACL*.

Feng, S.; Manmatha, R.; and Lavrenko, V. 2004. Multiple Bernoulli relevance models for image and video annotation. In *CVPR*, pages 1002–1009.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29.

Feng, F., and Lapata, M. 2010. How many words is a picture worth? Automatic caption generation for news images. In *ACL*, pages 1239–1249.

Gatt, A., and Reiter, E. 2009. SimpleNLG: A realisation engine for practical applications. In *ENLG*, pages 91–93.

Grubinger, M.; Clough, P. D; Müller, H.; and Thomas, D. 2006. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *LREC*.

Guillaumin, M.; Mensink, T.; Verbeek, J.; and Schmid, C. 2009. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*.

Jeon, J.; Lavrenko, V.; and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, pages 119–126.

Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2011. Baby talk: Understanding and generating simple image descriptions . In *CVPR*.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics In *NAACLHLT*, pages 71–78.

Li, S.; Kulkarni, G.; Berg, T. L.; Berg, A. C.; and Choi, Y. 2011. Composing simple image descriptions using webscale n-grams. In *CoNLL*, pages 220–228.

Makadia, A.; Pavlovic, V.; and Kumar, S. 2010. Baselines for image annotation. In *IJCV*, 90(1):88–105.

Marneffe, M.-C. de, and Manning, C. D. 2008. The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2Text: Describing images using 1 million captioned photographs. In *NIPS*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. BLEU: A method for automatic evaluation of machine translation. In ACL, pages 311–318.

Rashtchian, C.; Young, P; Hodosh, M; and Hockenmaier, J. 2010. Collecting Image Annotations Using Amazon’s Mechanical Turk. In *NAACLHLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*.

Reape, M., and Mellish, C. 1999. Just what is aggregation anyway?. In *EWLG*, pages 20–29.

Reiter, E., and Dale, R. eds. 2000. *Building Natural Language Generation Systems.* New York, NY, USA: Cambridge University Press.

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *CVPR*, pages 1745–1752.

Yang, Y.; Teo, C. L.; Daumé, H. (III); and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. In *EMNLP*, pages 444–454.

Yao, B. Z.; Yang, X.; Lin, L.; Lee, M. W.; and Zhu, S.-C. 2008. I2T: Image parsing to text description. In *Proceedings of the IEEE*, 98(8):1485–1508.