Phrase-based Image Captioning

Rémi Lebret* REMI@LEBRET.CH
Pedro O. Pinheiro* PEDRO@OPINHEIRO.COM
Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Ronan Collobert RONAN@COLLOBERT.COM
Facebook AI Research, Menlo Park, CA, USA

Abstract

Generating a novel textual description of an image is an interesting problem that connects computer vision and natural language processing. In this paper, we present a simple model that is able to generate descriptive sentences given a sample image. This model has a strong focus on the syntax of the descriptions. We train a purely bilinear model that learns a metric between an image representation (generated from a previously trained Convolutional Neural Network) and phrases that are used to described them. The system is then able to infer phrases from a given image sample. Based on caption syntax statistics, we propose a simple language model that can produce relevant descriptions for a given test image using the phrases inferred. Our approach, which is considerably simpler than state-of-the-art models, achieves comparable results in two popular datasets for the task: Flickr30k and the recently proposed Microsoft COCO.

1. Introduction

Being able to automatically generate a description from an image is a fundamental problem in artificial intelligence, connecting computer vision and natural language processing. The problem is particularly challenging because it requires to correctly recognize different objects in images and how they interact. Another challenge is that an image description generator needs to express these interactions in a natural language (*e.g.* English). Therefore, a language model is implicitly required in addition to visual understanding.

Recently, this problem has been studied by many different authors. Most of the attempts are based on recurrent neural networks to generate sentences. These models leverage the power of neural networks to transform image and sentence representations into a common space (Mao et al., 2015; Karpathy & Fei-Fei, 2015; Vinyals et al., 2014; Donahue et al., 2014).

In this paper, we propose a different approach to the problem that does not rely on complex recurrent neural networks. An exploratory analysis of two large datasets of image descriptions reveals that their syntax is quite simple. The ground-truth descriptions can be represented as a collection of noun, verb and prepositional phrases. The different entities in a given image are described by the noun phrases, while the interactions or events between these entities are encoded by both the verb and the prepositional phrases. We thus train a model that predicts the set of phrases present in the sentences used to describe the images. By leveraging previous works on word vector representations, each phrase can be represented by the mean of the representations of the words that compose the phrase. Vector representations for images can also be easily obtained from some pre-trained convolutional neural networks. The model then learns a common embedding between phrase and image representations (see Figure 3).

Given a test image, a bilinear model is trained to predict a set of top-ranked phrases that best describe it. Several noun phrases, verb phrases and prepositional phrases are in this set. The objective is therefore to generate syntactically correct sentences from (possibly different) subsets of these phrases. We introduce a trigram constrained language model based on our knowledge about how the sentence descriptions are structured in the training set. With a very constrained decoding scheme, sentences are inferred with a beam search. Because these sentences are not con-

*These two authors contributed equally to this work.
Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

ditioned to the given image (apart with the initial phrases selection), a re-ranking is used to pick the sentence that is closest to the sample image (according to the learned metric). The quality of our sentence generation is evaluated on two very popular datasets for the task: Flickr30k (Hodosh et al., 2013) and the recently published COCO (Lin et al., 2014). Using the popular BLEU score (Papineni et al., 2002), our results are competitive with other recent works. Our generated sentences also achieve a similar performance as humans on the BLEU metric.

The paper is organized as follows. Section 2 presents related works. Section 3 presents the analysis we conducted to better understand the syntax of image descriptions. Section 4 describes the proposed phrase-based model. Section 5 introduces the sentence generation from the predicted phrases. Section 6 describes our experimental setup and the results on the two datasets. Section 7 concludes.

2. Related Works

The classical approach to sentence generation is to pose the problem as a retrieval problem: a given test image will be described with the highest ranked annotation in the training set (Hodosh et al., 2013; Socher et al., 2014; Srivastava & Salakhutdinov, 2014). These matching methods may not generate proper descriptions for a new combination of objects. Due to this limitation, several generative approaches have been proposed. Many of them use syntactic and semantic constraints in the generation process (Yao et al., 2010; Mitchell et al., 2012; Kuznetsova et al., 2012; Kulkarni et al., 2013). These approaches benefit from visual recognition systems to infer words or phrases, but in contrast to our work they do not leverage a multimodal metric between images and phrases.

More recently, automatic image sentence description approaches based on deep neural networks have emerged with the release of new large datasets. As starting point, these solutions use the rich representation of images generated by Convolutional Neural Networks (LeCun et al., 1998) (CNN) that were previously trained for object recognition tasks. These CNN are generally followed by recurrent neural networks (RNN) in order to generate full sentence descriptions (Vinyals et al., 2014; Karpathy & Fei-Fei, 2015; Donahue et al., 2014; Chen & Zitnick, 2015; Mao et al., 2015; Venugopalan et al., 2014; Kiros et al., 2014). Among these recent works, long short-term memory (LSTM) is often chosen as RNN. In such approaches, the key point is to learn a common space between images and words or between images and sentences, i.e. a multimodal embedding.

Vinyals et al. (2014) consider the problem in a similar way as a machine translation problem. The authors propose an encoder/decoder (CNN/LSTM networks) system that is trained to maximize the likelihood of the target description sentence given a training image. Karpathy & Fei-Fei (2015) propose an approach that is a combination of CNN, bidirectional RNN over sentences and a structured objective responsible for a multimodal embedding. They then propose a second RNN architecture to generate new sentences. Similarly, Mao et al. (2015) and Donahue et al. (2014) propose a system that uses a CNN to extract image features and a RNN for sentences. The two networks interact with each other in a multimodal common layer.

Our model shares some similarities with these recent proposed approaches. We also use a pre-trained CNN to extract image features. However, thanks to the phrase-based approach, our model does not rely on complex recurrent networks for sentence generation, and we do not fine-tune the image features.

As our approach, Fang et al. (2015) proposes to not use recurrent networks for generating the sentences. Their solution can be divided into three steps: (i) a visual detector for words that commonly occur are trained using multiple instance learning, (ii) a set of sentences are generated using a Maximum-Entropy language model and (iii) the set of sentences is re-ranked using sentence-level features and a proposed deep multimodal similarity model. Our work differs from this approach in two different important ways: our model infers phrases present in the sentences instead of words and we use a considerably simpler language model.

3. Syntax Analysis of Image Descriptions

The art of writing sentences can vary a lot according to the domain. When reporting news or reviewing an item, not only the choice of the words might vary, but also the general structure of the sentence. In this section, we wish to analyze the syntax of image descriptions to identify whether images have their own structures. We therefore proceed to an exploratory analysis of two recent datasets containing a large amount of images with descriptions: Flickr30k (Hodosh et al., 2013) and COCO (Lin et al., 2014).

3.1. Datasets

The Flickr30k dataset contains 31,014 images where 1,014 images are for validation, 1,000 for testing and the rest for training (i.e. 29,000 images). The COCO dataset contains 123,287 images, 82,783 training images and 40,504 validation images. The testing images has not yet been released. We thus use two sets of 5,000 images from the validation images for validation and test, as in Karpathy & Fei-Fei (2015)[1]. In both datasets, images are given with five (or six) sentence descriptions annotated using Amazon

[1]Available at http://cs.stanford.edu/people/karpathy/deepimagesent/

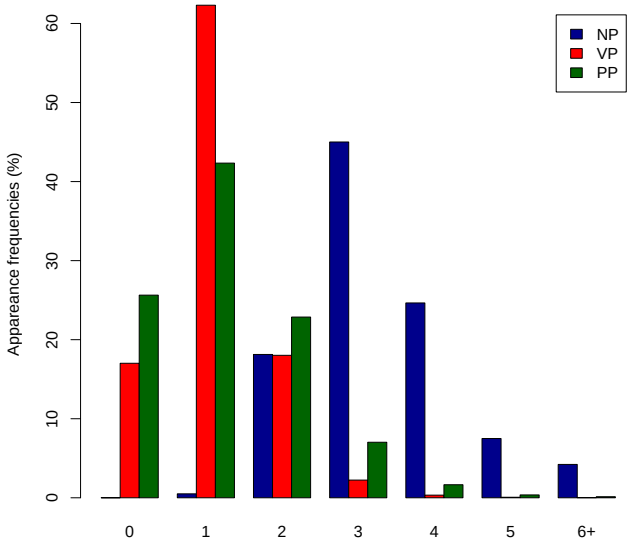


Figure 1. Statistics on the number of phrase chunks (NP, VP, PP) per ground-truth descriptions in Flickr30k and COCO training datasets.

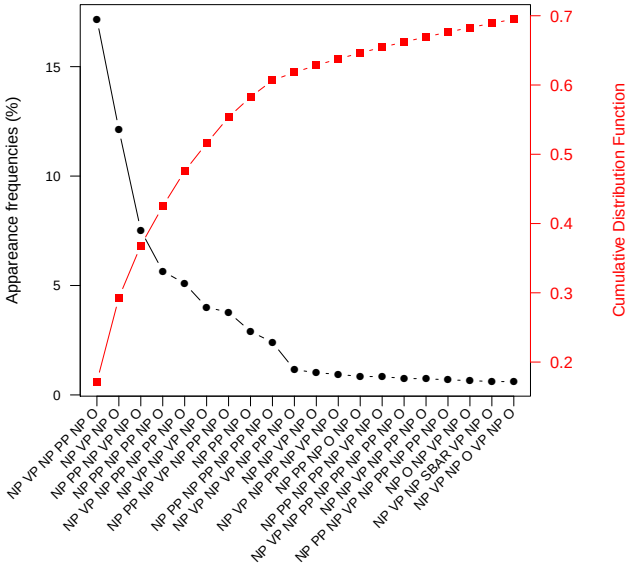


Figure 2. The 20 most frequent sentence structures in Flickr30k and COCO training datasets. The black line is the appearance frequency for each structure, the red line is the cumulative distribution.

Mechanical Turk (see Figure 3). This results in 559,113 sentences when combining both training datasets.

3.2. Chunking-based Approach

A quick overview over these sentence descriptions reveals that they all share a common structure, usually describing the different entities present in the image and how they interact between each other. This interaction among entities is described as actions or relative position between different objects. The sentence can be short or long, but it generally respects this process. To confirm this claim and better understand the description structures, we used a chunking (also called shallow parsing) approach which identifies the phrase chunks of a sentence (i.e., the non-recursive cores of various phrase types in text). These chunks are usually noun phrases (NP), verb phrases (VP) and prepositional phrases (PP). We extract them from the training sentences with the SENNA software[2]. Pre-verbal and post-verbal adverb phrases are merged with verb phrases to limit the number of phrase types. Table 1 presents an example sentence with its chunking analysis.

a man	is grinding	a ramp	on	a skateboard	.
NP	VP	NP	PP	NP	O

Table 1. Chunking analysis of an image description.

Statistics reported in Figure 1 and Figure 2 confirm that image descriptions possess a simple and distinct structure.

[2]Available at http://ml.nec-labs.com/senna/

These sentences do not have much variability. All the key elements in a given image are usually described with a noun phrase (NP). Interactions between these elements can then be explained using prepositional phrases (PP) or verb phrases (VP). A large majority of sentences contain from two to four noun phrases. Two noun phrases then interact using a verb or prepositional phrase. Describing an image is therefore just a matter of identifying these chunks. We thus propose to train a model which can predict the phrases which are likely to be in a given image.

4. Phrase-based Model for Image Descriptions

By leveraging previous works on word and image representations, we propose a simple model which can predict the phrases that best describe a given image. For this purpose, a metric between images and phrases is trained, as illustrated in Figure 3. The proposed architecture is then just a low-rank bilinear model $U^T V$.

4.1. Image Representations

For the representation of images, we choose to use a Convolutional Neural Network. CNN have been widely used in different vision domains and are currently the state-of-the-art in many object recognition tasks. We consider a CNN that has been pre-trained for the task of object classification (Simonyan & Zisserman, 2014). We use a CNN solely to the purpose of feature extraction, that is, no learning is

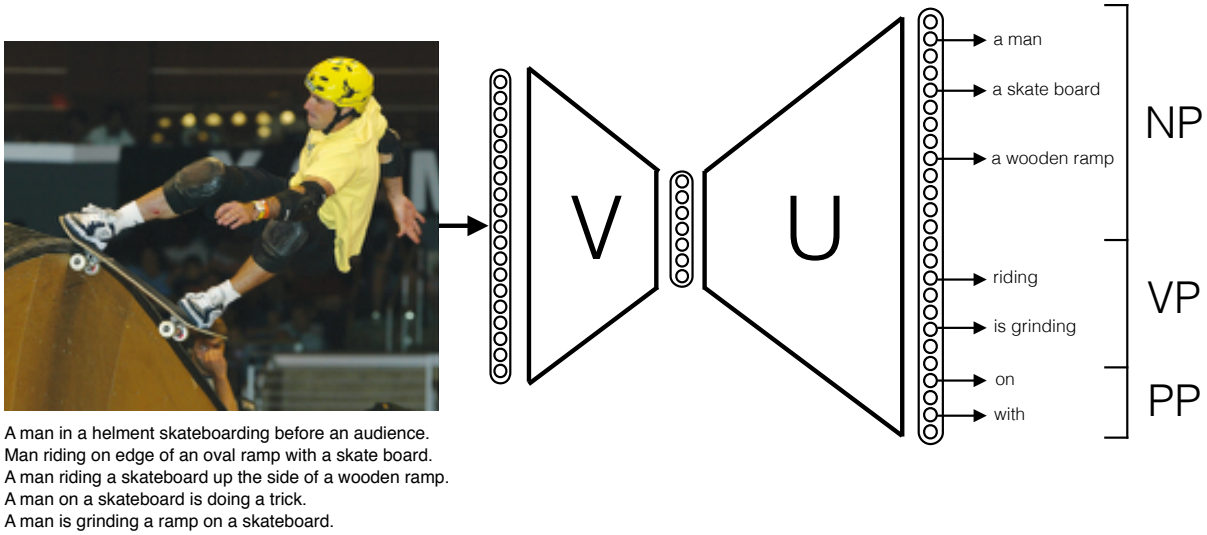


Figure 3. Schematic illustration of our phrase-based model for image descriptions.

done in the CNN layers.

4.2. Learning a Common Space for Image and Phrase Representations

Let $\mathcal{I}$ be the set of training images, $\mathcal{C}$ the set of all phrases used to describe $\mathcal{I}$, and θ the trainable parameters of the model. By representing each image $i \in \mathcal{I}$ with a vector $\mathrm{z}_i \in \mathbb{R}^n$ thanks to the pre-trained CNN, we define a metric between the image i and a phrase c as a bilinear operation:

$$f_\theta(c, i) = \mathrm{u}_c^T V \mathrm{z}_i \,, \tag{1}$$

with $U = (\mathrm{u}_{c_1}, \ldots, \mathrm{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$ and $V \in \mathbb{R}^{m \times n}$ being the trainable parameters θ. Note that $U^T V$ could be a full matrix, but a low-rank setting eases the capacity control.

4.3. Phrase Representations Initialization

Noun phrases or verb phrases are often a combination of several words. Good word vector representations can be obtained very efficiently with many different recent approaches (Mikolov et al., 2013b; Mnih & Kavukcuoglu, 2013; Pennington et al., 2014; Lebret & Collobert, 2015). Mikolov et al. (2013a) also showed that simple vector addition can often produce meaningful results, such as *king - man + woman* $\approx$ *queen*. By leveraging the ability of these word vector representations to compose by simple summation, representations for phrases are easily computed with an element-wise addition.

Each phrase c composed of K words w_k is therefore represented by a vector $\mathrm{x}_{w_k} \in \mathbb{R}^m$ thanks to a word representation model pre-trained on large unlabeled text corpora. A vector representation u_c for a phrase $c = \{w_1, \ldots, w_K\}$ is then calculated by averaging its word vector representations:

$$\mathrm{u}_c = \frac{1}{K} \sum_{k=1}^{K} \mathrm{x}_{w_k} \,. \tag{2}$$

Vector representations for all phrases $c \in \mathcal{C}$ can thus be obtained to initialized the matrix $U \in \mathbb{R}^{m \times |\mathcal{C}|}$. $V \in \mathbb{R}^{m \times n}$ is initialized randomly and trained to encode images in the same vector space than the phrases used for their descriptions.

4.4. Training with Negative Sampling

Each image i is described by a multitude of possible phrases $\mathcal{C}^i \subseteq \mathcal{C}$. We consider $|\mathcal{C}|$ classifiers attributing a score for each phrase. We train our model to discriminate a target phrase c_j from a set of negative phrases $c_k \in \mathcal{C}^- \subseteq \mathcal{C}$, with $c_k \neq c_j$. With $\theta = \{U, V\}$, we minimize the following logistic loss function with respect to θ:

$$\theta \mapsto \sum_{i \in \mathcal{I}} \sum_{c_j \in \mathcal{C}^i} \left(\log\left(1 + e^{-\mathrm{u}_{c_j}^T V \mathrm{z}_i}\right) + \sum_{c_k \in \mathcal{C}^-} \log\left(1 + e^{+\mathrm{u}_{c_k}^T V \mathrm{z}_i}\right)\right) . \tag{3}$$

The model is trained using stochastic gradient descent. A new set of negative phrases $\mathcal{C}^-$ is randomly picked from the training set at each iteration.

5. From Phrases to Sentence

After identifying the L most likely constituents c_j in the image i, we propose to generate sentences out of them. From this set, $l \in \{1, \ldots, L\}$ phrases are used to compose a syntactically correct description.

5.1. Sentence Generation

Using a statistical language modeling framework, the likelihood of a certain sentence is given by:

$$P(c_1, c_2, \ldots, c_l) = \prod_{j=1}^{l} P(c_j | c_1, \ldots, c_{j-1}) \quad (4)$$

Keeping this system as simple as possible and using the second order Markov property, we approximate Equation 4 with a trigram language model:

$$P(c_1, c_2, \ldots, c_l) \approx \prod_{j=1}^{l} P(c_j | c_{j-2}, c_{j-1}) \,. \quad (5)$$

The best candidate corresponds to the sentence $P(c_1, c_2, \ldots, c_l)$ which maximizes the likelihood of Equation 5 over all the possible sizes of sentence. Because we want to constrain the decoding algorithm to include prior knowledge on chunking tags $t \in \{NP, VP, PP\}$, we rewrite Equation 5 as:

$$\prod_{j=1}^{l} \sum_{t} P(c_j | t_j = t, c_{j-2}, c_{j-1}) P(t_j = t | c_{j-2}, c_{j-1})$$
$$= \prod_{j=1}^{l} P(c_j | t_j, c_{j-2}, c_{j-1}) P(t_j | c_{j-2}, c_{j-1}) \,. \quad (6)$$

Both conditions $P(c_j | t_j, c_{j-2}, c_{j-1})$ and $P(t_j | c_{j-2}, c_{j-1})$ are probabilities estimated by counting trigrams in the training datasets.

5.2. Sentence Decoding

At decoding time, we prune the graph of all possible sentences made out of the top L phrases with a beam search, according to three heuristics: (i) we consider only the transitions which are likely to happen (we discard any sentence which would have a trigram transition probability inferior to 0.01). This thresholding helps to discard sentences that are semantically incorrect; (ii) each predicted phrases c_j may appear only once[3]; (iii) we add syntactic constraints which are illustrated in Figure 4. The last heuristic is based on the analysis of syntax in Section 3. In Figure 2, we see that a noun phrase is, in general, always followed by a verb phrase or a prepositional phrase, and both are then followed by another noun phrase. A large majority of the sentences contain three noun phrases interleaved with verb phrases or prepositional phrases. According the statistics reported in Figure 1, sentences with two or four noun phrases are also common, but sentences with more than four noun phrases are marginal. We thus repeat this process $N = \{2, 3, 4\}$ times until reaching the end of a sentence (characterized by a period).

[3]This is easy to implement with a beam search, but intractable with a full search.

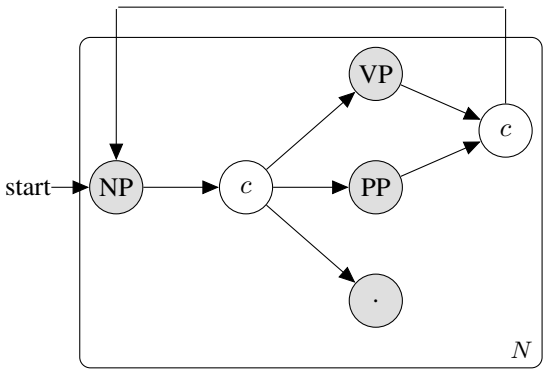


Figure 4. The constrained language model for generating description given the predicted phrases for an image.

5.3. Sentence Re-ranking

For each test image i, the proposed model will generate a set of M sentences. Sentence generation is not conditioned on the image, apart from phrases which are selected beforehand. Some phrase sequences might be syntactically good, but have low match with the image. Consider, for instance, an image with a cat and a dog. Both sentences *"a cat sitting on a mat and a dog eating a bone"* and *"a cat sitting on a mat"* are correct, but the second is missing an important part of the image. A ranking of the generated sentences is therefore necessary to choose the one that has the best match with the image.

Because a generated sentence is composed from l phrases predicted by our system, we simply average the phrase scores given by Equation 1. For a generated sentence s composed of l phrases c_j, a score between s and i is calculated as:

$$\frac{1}{l} \sum_{c_j \in s} f_\theta(c_j, i) \,. \quad (7)$$

The best candidate is the sentence which has the highest score out of the M generated sentences. This ranking helps the system to chose the sentence which is closer to the sample image.

6. Experiments

6.1. Experimental Setup

6.1.1. Feature Selection

Following Karpathy & Fei-Fei (2015), the image features are extracted using VGG CNN (Simonyan & Zisserman, 2014). This model generates image representations of dimension 4096 from RGB input images.

For each training set, only phrases occurring at least ten times are considered. This threshold is chosen to fulfil two objectives: (i) limit the number of phrases $\mathcal{C}$ and therefore

the size of the matrix U and (ii) exclude rare phrases to better generalize the descriptions. Statistics on the number of phrases are reported in Table 2. For Flickr30k, this threshold covers about 81% of NP, 83% of VP and 99% of PP. For COCO, it covers about 73% of NP, 75% of VP and 99% of PP. Phrase representations are then computed by averaging vector representations of their words. We obtained word representations from the Hellinger PCA of a word co-occurrence matrix, following the method described in Lebret & Collobert (2015). The word co-occurrence matrix is built over the entire English Wikipedia[4], with a symmetric context window of ten words coming from the 10,000 most frequent words. Words, and therefore also phrases, are represented in 400-dimensional vectors.

	Flickr30k	**COCO**
Noun Phrase (NP)	4818	8982
Verb Phrase (VP)	2109	3083
Prepositional Phrase (PP)	128	189
Total $\|\mathcal{C}\|$	7055	12254

Table 2. Statistics of phrases appearing at least ten times.

6.1.2. Learning the Multimodal Metric

The parameters θ are $V \in \mathbb{R}^{400\times 4096}$ (initialized randomly) and $U \in \mathbb{R}^{400\times|\mathcal{C}|}$ (initialized with the phrase representations) which are tuned on the validation datasets. They are trained with 15 randomly chosen negative samples and a learning rate set to 0.00025. It takes 2.5 hours on single CPU (Intel i7 4930K 3.4 GHz) to train on the COCO training dataset.

6.1.3. Generating Sentences from the Predicted Phrases

Transition probabilities for our constrained language model (see Figure 4) are calculated independently for each training set. No smoothing has been used in the experiments. Concerning the set of top-ranked phrases for a given test image, we select only the top five predicted verb phrases and the top five predicted prepositional phrases. Since the average number of noun phrases is higher than for the two other types of phrases (see Figure 1), more noun phrases are needed. The top twenty predicted noun phrases are thus selected.

6.2. Experimental Results

As a first evaluation, we consider the task of retrieving the ground-truth phrases from test image descriptions. Results reported in Table 3 show that our system achieves a recall of around 50% on this task on the test set of both datasets, assuming the threshold considered for each type of phrase (see 6.1.3). Note that this task is extremely difficult, as semantically similar phrases (*the women* / *women* / *the little girls*) are classified separately. Despite the possible number of noun phrases being higher, results in Table 3 reveal that noun phrases are better retrieved than verb phrases. This shows that our system is able to detect different objects in the image. However, finding the right verb phrase seems to be more difficult. A possible explanation could be that there exists a wide choice of verb phrases to describe interactions between the noun phrases. For instance, we see in Figure 3 that two annotators have used the same noun phrases (*a man*, *a skateboard* and *a (wooden) ramp*) to describe the scene, but they have then chosen a different verb phrase to link them (*riding* versus *is grinding*). Therefore, we suspect that a low recall for verb phrases does not necessarily mean that the predictions are wrong. Finding the right prepositional phrase seems, on the contrary, much easier. The high recall for prepositional phrase can be explained by much lower variability of this type of phrase compared to the two others (see Table 2).

	Flickr30k	**COCO**
Noun Phrase (NP)	38.14	45.44
Verb Phrase (VP)	20.61	27.83
Prepostional Phrase (PP)	81.70	84.49
Total	44.92	52.49

Table 3. Recall on phrase retrieval. For each test image, we take the top 20 predicted NP, the top 5 predicted VP, and the top 5 predicted PP.

As a second evaluation, we consider the task of generating full descriptions. We measure the quality of the generated sentences using the popular, yet controversial, BLEU score (Papineni et al., 2002). Table 4 shows our sentence generation results on the two datasets considered. BLEU scores are reported up to 4-gram. Human agreement scores are computed by comparing the first ground-truth description against the four others[5]. For comparison, we include results from recently proposed models. Our model, despite being simpler, achieves similar results to state of the art results. It is interesting to note that our results are very close to the human agreement scores.

We show examples of full automatic generated sentences in Figure 5. The simple language model used is able to generate sentences that are in general syntactically correct. Our

[4]Available at http://download.wikimedia.org. We took the January 2014 version.

[5]For all models, BLEU scores are computed against five reference sentences which give a slight advantage compared to human scores.

	Flickr30K				COCO			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
Human agreement	0.55	0.35	0.23	0.15	0.68	0.45	0.30	0.20
Mao et al. (2015)	0.60	0.41	0.28	0.19	0.67	0.49	0.35	0.25
Karpathy & Fei-Fei (2015)	0.57	0.37	0.24	0.16	0.62	0.45	0.32	0.23
Vinyals et al. (2014)	0.66	0.42	0.28	0.18	0.67	-	-	-
Donahue et al. (2014)	0.59	0.39	0.25	0.16	0.63	0.44	0.30	0.21
Fang et al. (2015)	-	-	-	-	-	-	-	0.26
Our model	0.60	0.37	0.22	0.14	0.73	0.50	0.34	0.23

Table 4. Comparison between human agreement scores, state of the art models and our model on both datasets. Note that there are slight variations between the test sets chosen in each paper.

model produces sensible descriptions with variable complexity for different test samples. Due to the generative aspect of the model, it can occur that the sentence generated is very different from the ground-truth and still provides a descent description. The last row of Figure 5 illustrates failure samples. We can see in these failure samples that our system has however outputted relevant phrases. There is still room for improvement for generating the final description. We deliberately choose a simple language model to show that competitive results can be achieved with a simple approach. A more complex language model could probably avoid these failure samples by considering a larger context. The probability for *a dog* to stand on top of *a wave* is obviously very low, but this kind of mistake cannot be detected with a simple trigram language model.

6.3. Diversity of Image Descriptions

In contrast to RNN-based models, our model is not trained to match a given image i with its ground-truth descriptions s, i.e., to give $P(s|i)$. Because our model outputs instead a set of phrases, this is not really surprising that only 1% of our generated descriptions are in the training set for Flickr30k, and 9.7% for COCO. While a RNN-based model is generative, it might easily overfit a small training data. Vinyals et al. (2014) report, for instance, that the generated sentence is present in the training set 80% of the time. Our model therefore offers a good alternative with the possibility of producing unseen descriptions with a combination of phrases from the training set.

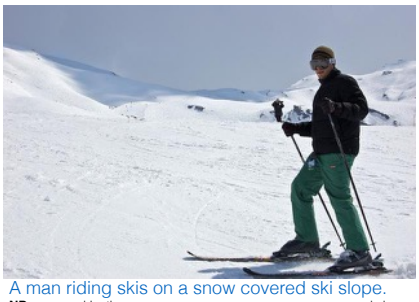
6.4. Phrase Representation Fine-Tuning

Before training the model, the matrix U is initialized with phrase representations obtained from the whole English Wikipedia. This corpus of unlabeled text is well structured and large enough to provide good word vector representations, which can then produce good phrase representations. However, the content of Wikipedia is clearly different from the content of the image descriptions. Some words used for describing images might be used in different contexts in Wikipedia, which can lead to out-of-domain representations for certain phrases. This becomes thus crucial to adapt these phrase representations by fine-tuning the matrix U during the training[6]. Some examples of noun phrases are reported in Table 5 with their nearest neighbors before and after the training. These confirm the importance of fine-tuning to incorporate visual features. In Wikipedia, *cat* seems to occur in the same context than *dog* or other animals. When looking at the nearest neighbors of a phrase such as *a grey cat*, other *grey* animals arise. After training on images, the word *cat* becomes the important feature of that phrase. And we see that the nearest neighbors are

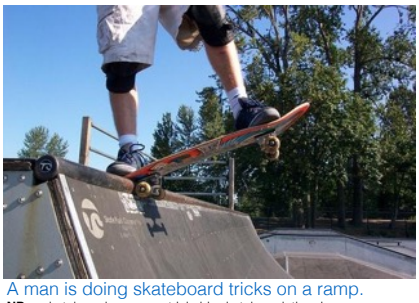
PHRASES	#	NEAREST NEIGHBORS BEFORE	NEAREST NEIGHBORS AFTER
A GREY CAT	1	A GREY DOG	A GRAY CAT
	2	A GREY AND BLACK CAT	A GREY AND BLACK CAT
	3	A GRAY CAT	A BROWN CAT
	4	A GREY ELEPHANT	A GREY AND WHITE CAT
	10	A YELLOW CAT	GREY AND WHITE CAT
HOME PLATE	1	A HOME PLATE	A HOME PLATE
	4	A PLATE	HOME BASE
	6	ANOTHER PLATE	THE PITCH
	9	A RED PLATE	THE BATTER
	10	A DINNER PLATE	A BASEBALL PITCH
A HALF PIPE	1	A PIPE	A PIPE
	2	A HALF	THE RAMP
	5	A SMALL CLOCK	A HAND RAIL
	9	A LARGE CLOCK	A SKATE BOARD RAMP
	10	A SMALL PLATE	AN EMPTY POOL

Table 5. Examples of three noun phrases from the COCO dataset with five of their nearest neighbors before and after learning.

[6]Experiments with a fixed U phrase representations matrix significantly hurt the general performance. We observe about a 50% decrease in both datasets with the BLEU metric. Since the number of trainable parameters is reduced, the capacity of V should be increased to guarantee a fair comparison.



A man riding skis on a snow covered ski slope.
NP: a man, skis, the snow, a person, a woman, a snow covered slope, a slope, a snowboard, a skier, man.
VP: wearing, riding, holding, standing on, skiing down.
PP: on, in, of, with, down.
A man wearing skis on the snow.



A man is doing skateboard tricks on a ramp.
NP: a skateboard, a man, a trick, his skateboard, the air, a skateboarder, a ramp, a skate board, a person, a woman.
VP: doing, riding, is doing, performing, flying through.
PP: on, of, in, at, with.
A man riding a skateboard on a ramp.

The girl with blue hair stands under the umbrella.
NP: a woman, an umbrella, a man, a person, a girl, umbrellas, that, a little girl, a cell phone.
VP: holding, wearing, is holding, holds, carrying.
PP: with, on, of, in, under.
A woman is holding an umbrella.

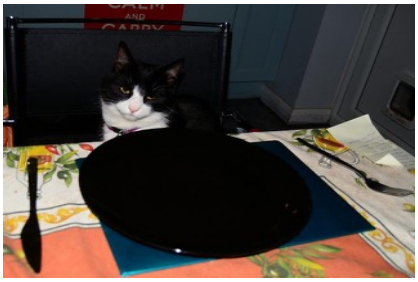
A slice of pizza sitting on top of a white plate.
NP: a plate, a white plate, a table, pizza, it, a pizza, food, a sandwich, top, a close.
VP: topped with, has, is, sitting on, is on.
PP: of, on, with, in, up.
A table with a plate of pizza on a white plate.

A baseball player swinging a bat on a field.
NP: the ball, a game, a baseball player, a man, a tennis court, a ball, home plate, a baseball game, a batter, a field.
VP: swinging, to hit, playing, holding, is swinging.
PP: on, during, in, at, of.
A baseball player swinging a bat on a baseball field.

A bunch of kites flying in the sky on the beach.
NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group.
VP: flying, flies, is flying, flying in, are.
PP: on, of, with, in, at.
People flying kites on the beach.

People gather around a truck parked on a boat.
NP:a man, a bench, a boat, a woman, a person, luggage, that, a train, water, the water.
VP: sitting on, carrying, riding, sitting in, sits on.
PP: of, on, with, in, next to.
A man sitting on a bench with a woman carrying luggage.

A person on a surf board in the ocean.
NP: a dog, a wave, a person, the water, a man, the ocean, top, that, the snow, a surfboard.
VP: riding, standing on, wearing, laying on, sitting on.
PP: on, of, in, with, near.
A dog standing on top of a wave on the ocean.



A cat sitting in a chair staring at a plate on a table.
NP: a table, top, a desk, a cat, front, it, that, a laptop, a laptop computer, the table.
VP: sitting on, is, sitting in, sitting next to, has.
PP: of, on, with, in, next to.
A cat sitting on top of a desk with a laptop.

Figure 5. Quantitative results for images on the COCO dataset. Ground-truth annotation (in blue), the NP, VP and PP predicted from the model and generated annotation (in black) are shown for each image. The last row are failure samples.

now cats with different colours. In some cases, averaging word vectors to represent phrases is not enough to capture the semantic meaning. Fine-tuning is thus also important to better learn specific phrases. Images related to baseball games, for example, have enabled the phrase *home plate* to be better defined. This is also true for the phrase *a half pipe* with images about skateboarding. This leads to interesting phrase representations, grounded in the visual world, which could be possibly used in natural language applications in future work.

7. Conclusion

In this paper, we propose a simple model that is able to infer different phrases from image samples. From the phrases predicted, our model is able to automatically generate sentences using a statistical language model. We show that the problem of sentence generation can be effectively achieved without the use of complex recurrent networks. Our algorithm, despite being simpler than state-of-the-art models, achieves similar results on this task. Also, our model generate new sentences which are not generally present in training set. Future research directions will go towards leveraging unsupervised data and more complex language models to improve sentence generation. Another interest is assessing the impact of visually grounded phrase representations into existing natural language processing systems.

Acknowledgements

This work was supported by the HASLER foundation through the grant "Information and Communication Technology for a Better World 2020" (SmartWorld).

References

Chen, X. and Zitnick, C. L. Minds Eye: A Recurrent Visual Representation for Image Caption Generation. In *IEEE International Concference on Computer Vision and Patter Recognition (CVPR)*, 2015.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *arXiv preprint arXiv:1411.4389*, 2014.

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

Hodosh, M., Young, P., and Hockenmaier, J. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013.

Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE International Concference on Computer Vision and Patter Recognition (CVPR)*, 2015.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint arXiv:1411.2539*, 2014.

Kulkarni, G., Premraj, V., Dhar, S., Li, Siming, Choi, Yejin, Berg, A. C., and Berg, T. L. Baby Talk: Understanding and Generating Simple Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (12):2891–2903, 2013.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. Collective Generation of Natural Image Descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 359–368. Association for Computational Linguistics, July 2012.

Lebret, R. and Collobert, R. Rehabilitation of count-based models for word vector representations. In Gelbukh, Alexander (ed.), *Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, pp. 417–429. Springer International Publishing, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. L. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In *International Conference on Learning Representations (ICLR)*, 2015.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119. 2013b.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., and Daumé, III, H. Midge: Generating Image Descriptions from Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756. Association for Computational Linguistics, 2012.

Mnih, A. and Kavukcuoglu, Koray. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pp. 2265–2273. 2013.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Pennington, J., Socher, R., and Manning, C. D. GloVe: Global Vectors for Word Representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12, 2014.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

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

Srivastava, N. and Salakhutdinov, R. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, 2014.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. J., and Saenko, K. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. *arXiv preprint arXiv:1412.4729*, 2014.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

Yao, B. Z., Yang, X., Lin, L., Lee, M. W., and Zhu, S. C. I2T: Image Parsing to Text Description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.