RESEARCH INSTITUTE

SIMPLE IMAGE DESCRIPTION GENERATOR VIA A LINEAR PHRASE-BASED MODEL

Rémi Lebret[a]   Pedro H. O. Pinheiro

Ronan Collobert

Idiap-RR-22-2015

JUNE 2015

[a]Idiap

SIMPLE IMAGE DESCRIPTION GENERATOR VIA A LINEAR PHRASE-BASED MODEL

Rémi Lebret[*] **& Pedro O. Pinheiro**[*]
Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
`remi@lebret.ch, pedro@opinheiro.com`

Ronan Collobert
Idiap Research Institute, Martigny, Switzerland
`ronan@collobert.com`

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 sentences given a sample image. Our model learns to embed an image representation (generated from a previously trained Convolutional Neural Network) into a multimodal space that is common to the image and the phrases that are used to described them. The system is able to infer phrases from a given image sample. Based on the sentence descriptions, we propose a simple language model that is able to produce relevant descriptions for a given test image using the feature representation and the phrases inferred. We achieve promising first results on the recently released COCO dataset.

1 INTRODUCTION

Being able to automatically generate a description from an image is a fundamental problem in artificial intelligent, connecting computer vision and natural language processing. The problem is particularly challenging because it requires to correctly recognize different objects in images and also how they interact.

Convolutional Neural Networks (CNN) have achieved state of the art results in different computer vision tasks in the last few years. More recently, different authors proposed automatic image sentence description approaches based on deep neural networks. All the solutions use the representation of images generated by CNN that was previously trained for object recognition tasks as start point.

Vinyals et al. (2014) consider the problem in a similar way as a machine translation problem. The authors propose a 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 (2014) propose an approach that is a combination of CNN, bidirectional recurrent neural networks over sentences and a structured objective responsible for a multimodal embedding. They then propose a second recurrent neural network architecture to generate new sentences. Similar to the previous works, Mao et al. (2014) and Donahue et al. (2014) propose a system that uses a CNN to extract image features and a deep recurrent neural network for sentences. The two networks interact with each other in a multimodal common layer.

Fang et al. (2014) proposes a different approach to the problem that does not rely on recurrent neural networks. 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.

---

[*]These two authors contributed equally to this work.

A man in a helment skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
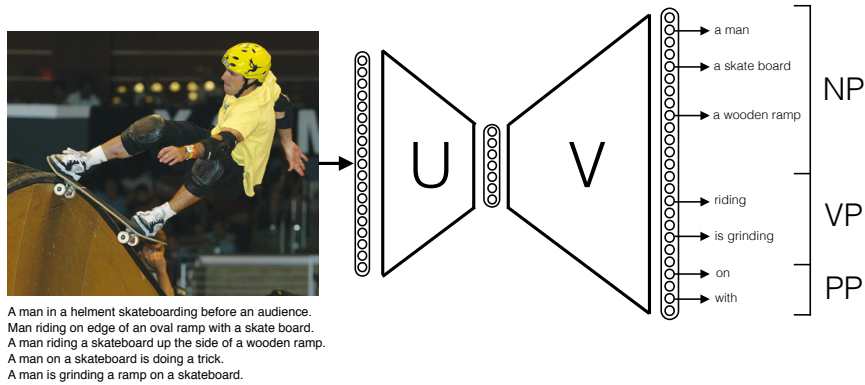A man is grinding a ramp on a skateboard.

Figure 1: Schematic illustration of our phrase-based model for image descriptions.

This paper proposes a different approach for the problem. We propose a system that at the same time: (i) automatically generates a sentence describing a given scene and (ii) is relatively simpler than the recently proposed approaches. Our model shares some similarities with previous proposed deep approaches. We also use a pre-trained CNN to extract image features and we also consider a multimodal embedding. However, thanks to the phrase-based approach, we do not use any complex recurrent network for sentence generation.

We represent the ground-truth sentences as a collection of noun, verb and prepositional phrases. Each phrase is represented by the mean of the vector representations of the words that compose the phrase. We then train a simple *linear* embedding model that transform an image representation into a multimodal space that is common to the image and the phrases that are used to describe them. To automatically generate sentences in inference time, we (i) infer the phrases that correspond to the sample image and (ii) use a simple language model based on the statics of ground-truth sentences in the corpus.

2 PHRASE-BASED MODEL FOR IMAGE DESCRIPTIONS

2.1 UNDERSTANDING STRUCTURES 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. Sentence structures used for describing images can therefore be identified.

They possess a very distinct structure, usually describing the different objects present on the scene and how they interact between each other. This interaction among objects is described as actions or relative position between different objects. The sentence can be short or long, but it generally respects this process. This statement is illustrated with the ground-truth sentence descriptions of the image in Figure 1.

Chunking-based approach 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). Describing an image is therefore just a matter of identifying these constituents to describe images. We propose to train a model which can predict the phrases which are likely to be in a given image.

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

From phrases to sentence   After identifying the most likely constituents in the image, we propose to use a statistical language model to combine these and generate a proper description. A general framework is defined to reduce the total number of combination and thus speed up the process for generating sentences. In general, a noun phrase is always followed by a verb phrase or a prepositional phrase, and both are then followed by another noun phrase. Figure 2 illustrates the Bayesian network used to generate such sentences. The nodes are the phrases, and the edges are the transition probabilities which are obtained with an n-gram model. This process is repeated L times until reaching the end of a sentence (characterized by a period).
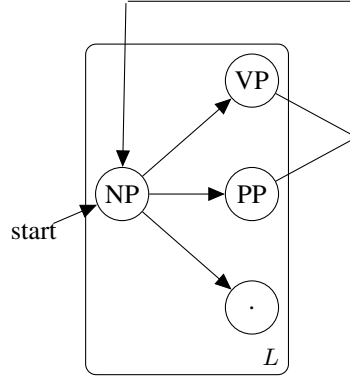

Figure 2: A Bayesian network for generating description given the predicted phrases for an image.

2.2   A MULTIMODAL REPRESENTATION

Image representations   For the representation of images, we choose to use a Convolutional Neural Network. CNNs have been widely used for many 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. We use a CNN solely to the purpose of feature extraction, that is, no learning is done in the CNN layers.

Learning of a common space for image and phrase representations   Let $\mathcal{I}$ be the set of training images, $\mathcal{D}$ the set of all sentence descriptions for $\mathcal{I}$, $\mathcal{C}$ the set of all phrases occuring in $\mathcal{D}$, and θ the trainable parameters of the model. $\mathcal{D}_i$ is the set of sentences describing a given image $i \in \mathcal{I}$, and $\mathcal{C}_d$ is the set of phrases which compose a sentence description $d \in \mathcal{D}_i$. The training objective is to find the phrases c that describe the images i by maximizing the log probability:

$$\sum_{i \in \mathcal{I}} \sum_{d \in \mathcal{D}_i} \sum_{c \in \mathcal{C}_d} \log p(c|i) \tag{1}$$

Each image $i \in \mathcal{I}$ is represented by a vector $x_i \in \mathbb{R}^n$ thanks to a pre-trained CNN. Each phrase c is composed of K words w which are represented by a vector $x_w \in \mathbb{R}^m$ thanks to another pre-trained model for word representations. A vector representation z_c for a phrase $c = \{w_1, \ldots, w_K\}$ is then calculated by averaging its word vector representations:

$$z_c = \frac{1}{K} \sum_{k=1}^{K} x_{w_k} . \tag{2}$$

Vector representations for all phrases $c \in \mathcal{C}$ can thus be obtained to build a matrix $V = \left[z_{c_1}, \ldots, z_{c_{|\mathcal{C}|}} \right] \in \mathbb{R}^{m \times |\mathcal{C}|}$. In general, $m \ll n$. An encoding function is therefore defined to map image representations $x_i \in \mathbb{R}^n$ in the same vector space than phrase representations $z_c \in \mathbb{R}^m$:

$$g_\theta(i) = x_i U , \tag{3}$$

where $U \in \mathbb{R}^{n \times m}$ is initialized randomly and trained to encode images in the same vectorial space than the phrases used for their descriptions. Because representations of images and phrases are in a common vector space, similarities between a given image i and all phrases can be calculated:

$$f_\theta(i) = g_\theta(i) V , \tag{4}$$

where V is fine-tuned to incorporate other features coming from the images. By denoting $[f_\theta(i)]_j$ the score for the jth phrase, this score can be interpreted as the conditional probability $p(c = c_j|i, \theta)$ by applying a softmax operation over all the phrases:

$$p(c = c_j|i, \theta) = \frac{e^{[f_\theta(i)]_j}}{\sum_{k=1}^{|\mathcal{C}|} e^{[f_\theta(i)]_k}} . \tag{5}$$

In practice, this formulation is often impractical due to the large set of possible phrases $\mathcal{C}$.

3
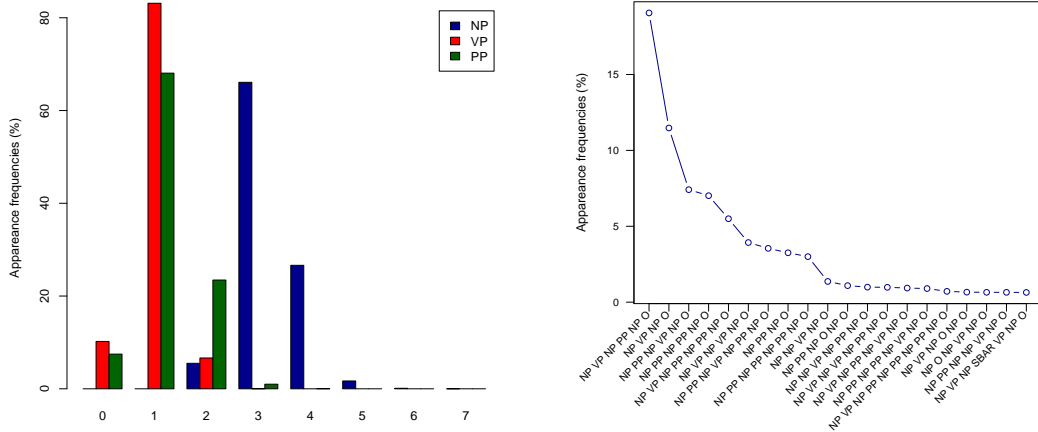
(a) The number of phrases per sentence.

(b) The 20 most frequent sentence syntactic structures.

Figure 3: Sentence structure statistics of COCO datasets.

Training with negative sampling With $\theta = \{U, V\}$ and a negative sampling approach, we instead minimize the following logistic loss function with respect to θ:

$$\theta \mapsto \sum_{i \in \mathcal{I}} \sum_{d \in \mathcal{D}_i} \sum_{c_j \in \mathcal{C}_d} \left(\log\left(1 + e^{[f_\theta(i)]_j} \right) + \sum_{k=1}^{N} \log\left(1 + e^{-[f_\theta(i)]_k} \right) \right). \tag{6}$$

Thus the task is to distinguish the target phrase from draws from the noise distribution, where there are N negative samples for each data sample. The model is trained using stochastic gradient descent.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Dataset We validate our model on the recently proposed COCO dataset (Lin et al., 2014), which contains complex images with multiple objects. The dataset contains a total of 123,000 images, each of them with 5 human annotated sentences. We use 4,000 images for both validation and testing, considering the two sets to have the same statistics. We measure the quality of of the generated sentences using the popular, yet controversial, BLEU score (Papineni et al., 2002).

Feature selection For image features, we choose the public available Overfeat[1] feature extractor, proposed by Sermanet et al. (2014), as our CNN of choice (more precisely, the *slow* model). This model generates image representations of dimension 4096 from RGB input images. For sentence features, we extract phrases from the 576,737 training sentences with the SENNA software[2]. Statistics reported in Figure 3 confirm the hypothesis that image descriptions have a simple syntactic structure. A large majority of sentences contain from two to four noun phrases. Two noun phrases then interact using a verb or prepositional phrase. Only phrases occuring at least ten times in the training set are considered. This results in 11,688 noun phrases, 3,969 verb phrases[3] and 219 prepositional phrases. Phrase representations are then computed by averaging vector representations of their words. We obtained word vector representations from the Hellinger PCA of a word co-occurence matrix, following the method described in Lebret & Collobert (2014). The word co-occurence matrix is built over the entire English Wikipedia[4], with a symmetric context window of

---

[1] Available at `http://cilvr.nyu.edu/doku.php?id=software:overfeat:start`

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

[3] Pre-verbal and post-verbal adverb phrases are merged with verb phrases.

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

Method of generating text	B-1	B-2	B-3
Human agreement	0.66	0.44	0.29
Karpathy & Fei-Fei (2014)	0.50	0.25	0.12
Vinyals et al. (2014)	0.67	-	-
Donahue et al. (2014)	0.63	0.44	0.30
Our model	0.67	0.42	0.26

Table 1: Comparison between human agreement scores, state of the art models and our model on the COCO dataset. All the results were done in different test set.

ten words coming from the 10,000 most frequent words. Words, and therefore also phrases, are represented in 400-dimensional vectors.

Learning multimodal representation The parameters θ are $U \in \mathbb{R}^{4096 \times 400}$ and $V \in \mathbb{R}^{400 \times 15876}$ which is initialized with the phrase representations. They are trained with $N = 15$ negative samples and a learning rate set to 0.00025.

Generating sentences from the predicted phrases According to statistics on sentence structures, we set $L = \{2, 3, 4\}$. As nodes, we consider only the top twenty predicted noun phrases, the top ten predicted verb phrases and the top five predicted prepositional phrases. A trigram language model is used for the transition probabilities between two nodes. Probabilities are calculated using lexical phrases $P(c_j | c_{j-2}, c_{j-1})$. In order to reduce the number of sentences generated, we just consider the transitions which are likely to happen (we discard any sentence which would have a trigram transition probability inferior to 0.01). This thresholding also helps to discard sentences that are semantically incorrect.

Ranking generated sentences Our final step consists on ranking the sentences generated and choosing the one with the highest score as the final output. For each test image i, we generate a set of M sentence candidates using the proposed language model. For each sentence s_m ($m \in \{1, ..., M\}$), we compute its vector representation z_{s_m} by averaging the representation of the phrases $z_c \in V$ that make the sentence. The final score for each sentence s_m is computed by doing a dot product between the sentence vector representation and the encoded representation of the sample image i:

$$f_\theta(i, m) = g_\theta(i) z_{s_m} \,. \tag{7}$$

The output of the system is the sentence which has the highest score. This ranking helps the system to chose the sentence which is closer to the sample image.

3.2 EXPERIMENTAL RESULTS

Table 1 show our sentence generation results on the COCO dataset. BLEU scores are reported at a unigram, bigram and trigram level. Human agreement scores are computed by comparing one of the ground-truth description against the others. For comparison, we include results from recently proposed models based on recurrent neural networks. Note that for the moment, the COCO dataset does not provide a test set. For that reason, the test set used at each model is not the same. Our model gives competitive results at unigram, bigram and trigram level. It is interesting to note that our results are very close to the human agreeemnt scores. Examples of full automatic generated sentences can be found in Figure 4.

4 CONCLUSION AND FUTURE WORKS

In this paper, we propose a simple model that is able to automatically generate sentences from an image sample. Our model is considerably simpler than the current state of the art, which uses complex recurrent neural networks. We predict phrase components that are likely to describe a

Many people sitting under umbrellas on a sunny beach.
A group of people sitting on the beach.

A lady sitting down in a living room full of furniture.
A woman sitting in a living room.

A guy on a surf board riding a wave.
A man riding a wave on surfboard.

Large sandwich plate with onion rings on the side.
A white plate topped with sandwich.

Large crowd moving together down a city street.
People watching a man standing in front of a group.

A man and a woman sitting on a bench at a tennis court.
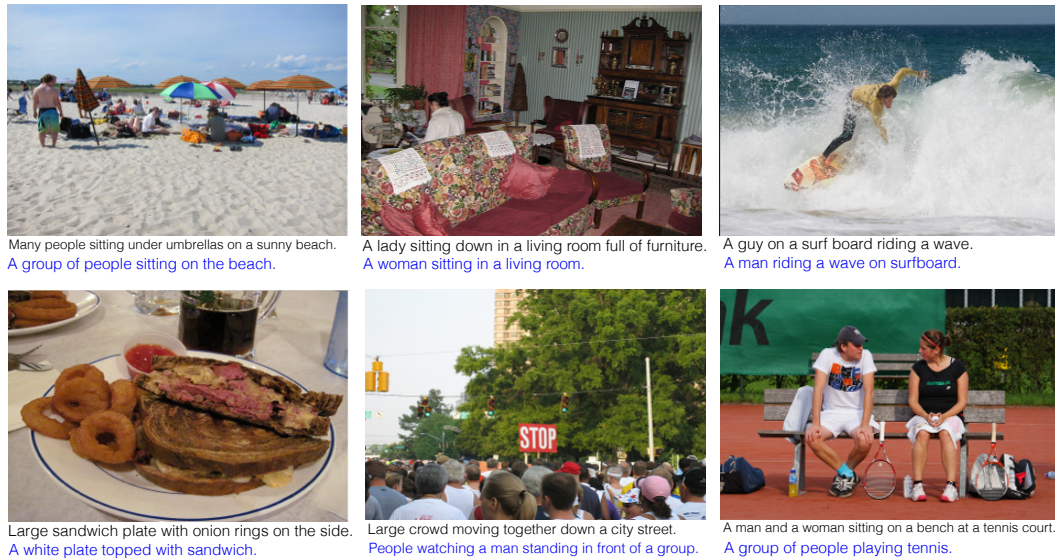A group of people playing tennis.

Figure 4: Quantitative results for images on the COCO dataset. Ground-truth (black) and generated (blue) captions are shown for each image. The two last are failure samples.

given image and use the knowledge of how the sentences are structured to generate sentences. Our model achieve promising first results. Future works include extend the model to different datasets (Flickr8k, Flickr30k and final COCO version for benchmarking), do image-sentence ranking experiments and improve the language model used.

REFERENCES

Donahue, Jeff, Hendricks, Lisa Anne, Guadarrama, Sergio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014.

Fang, Hao, Gupta, Saurabh, Iandola, Forrest N., Srivastava, Rupesh, Deng, Li, Dollár, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, Margaret, Platt, John C., Zitnick, C. Lawrence, and Zweig, Geoffrey. From captions to visual concepts and back. *CoRR*, abs/1411.4952, 2014.

Karpathy, Andrej and Fei-Fei, Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

Lebret, Remi and Collobert, Ronan. Rehabilitation of Count-based Models for Word Vector Representations. *CoRR*, abs/1412.4930, 2014.

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan L. Explain images with multimodal recurrent neural networks. *CoRR*, abs/1410.1090, 2014.

Mikolov, T., Chen, K., Corrado, G., and Dean, Jeff. Efficient Estimation of Word Representations in Vector Space. *ICLR Workshp*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 2013b.

Mnih, A. and Kavukcuoglu, Koray. Learning word embeddings efficiently with noise-contrastive estimation. In *NIPS*. 2013.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

Pennington, J., Socher, Richard, and Manning, C. D. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*, 2014.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.