A NEW CNN-RNN FRAMEWORK FOR REMOTE SENSING IMAGE CAPTIONING

Genc Hoxha, Farid Melgani and Jacopo Slaghenauffi

Department of Information Engineering and Computer Science
University of Trento, 38123 Trento, Italy
(email: genc.hoxha@unitn.it; farid.melgani@unitn.it; jacopo.slaghenauffi@studenti.unitn.it)

ABSTRACT

Remote sensing (RS) image captioning has been recently attracting the attention of the community as it provides more semantic information with respect to the traditional tasks such as scene classification. Image captioning aims to generate a coherent and comprehensive description that summarizes the content of an image. The description can be obtained directly from the ground truth descriptions of similar images (retrieval based image captioning) or can be generated through the encoder-decoder framework. The former has the limitation of not generating new descriptions. The latter may be affected by misrecognition of scenes or semantic objects. In this paper we try to address these issues by proposing a new framework which is a combination of generation and retrieval based image captioning. First a CNN-RNN framework combined with beam-search generates multiple captions for a target image. Then the best caption is selected on the basis of its lexical similarity with the reference captions of most similar images. Experimental results on RSCID dataset are reported and discussed.

Index Terms— Beam-search algorithm, encoder-decoder framework, remote sensing image captioning, retrieval based image captioning.

1. INTRODUCTION

Most of the remote sensing applications are concentrated on image segmentation, object recognition and scene classification. While these applications have been very pervasive in the last years, they are limited to explore the images by only providing information about the class labels or recognizing the objects present in the image and do not take into account the attributes of the objects and their relation. In order to model the attributes and the relationships between different objects of an image, image captioning is recently introduced in the remote sensing community [1]–[5]. Remote sensing image captioning aims at generating a textual description (caption) that summarizes the content of an image. Complete and correct sentences include by default the attributes and the relationships between the objects within the image and give a richer representation of the image semantic content with respect to other techniques. This richer representation can be useful in a variety of remote sensing applications such as for example image retrieval [6].

In the RS literature generally there are two different image captioning methods: retrieval-based image captioning and generation-based image captioning. Given a query image, the retrieval-based image captioning method first searches for similar images along with their descriptions in an archive and then assigns to the query image one (or more) pre-existing description(s) of the most similar images. This method produces sentences that may be correct in syntax (if the archive is properly built) but have the limitation of not generating new captions. Also the method assumes that within the archive there is always a relevant image-text pair for the query image, which is not always the case. Nevertheless, the method remains still interesting. Delvin *et al.* [7] explored nearest neighbor approach to assign a caption to a target image. After finding the similar images to a target image, they choose the best caption from the retrieved images captions implementing a consensus caption score. The consensus caption score aims at measuring lexical similarity between the captions of the retrieved images. The caption that maximizes this score with respect to all the captions of the retrieved images is assigned to the target image. They showed that their simple mechanism could outperform several state-of- the art image caption approaches.

Many generation-based image captioning methods are usually based on the encoder-decoder framework. First using a pre-trained convolutional neural network (CNN), the visual features of the image are encoded in a raw vector. The encoded features are then forwarded to a language model (decoder). The language model is usually made of recurrent neural networks (RNNs) which generate word by word descriptions of the content of the images. Most of the generation-based image captioning models produce only one caption. The produced caption is generated one word at a time and is conditioned on the image features and the previously predicted words. If at a given time step an object or scene is misrecognized during the sentence generation, this error will be propagated in the successive predicted words which may lead to an incorrect sentence.

To overcome the aforementioned problems, in this paper we propose to combine the generation based image

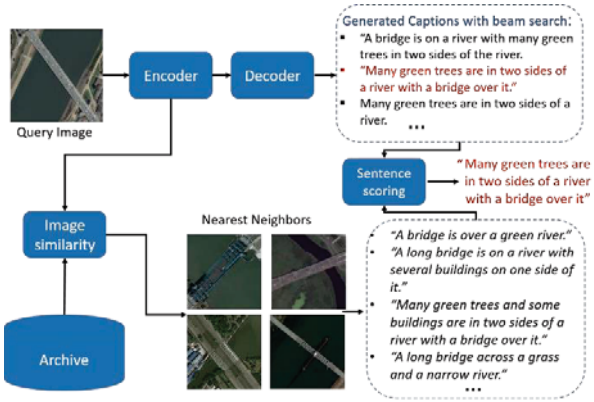

Figure 1. Scheme block of the proposed image captioning system.

captioning method with retrieval based image captioning method in order to: 1) generate multiple captions for a target image using the beam-search algorithm and 2) choose the best caption among the multiple generated captions on the basis of the lexical similarity with reference captions of similar images from the archive.

2. PROPOSED METHODOLOGY

Let $\boldsymbol{I} = [\boldsymbol{I_1}, \boldsymbol{I_2}, \dots, \boldsymbol{I_N}]$ be an archive consisting of $\boldsymbol{N}$ remote sensing images and $\boldsymbol{I_i}$ be the $\boldsymbol{i}$-th image. Each image is composed of $\boldsymbol{J}$ textual descriptions (or sentences). Let $\boldsymbol{S_{i,j}} = \{\boldsymbol{w_{1,i}}, \boldsymbol{w_{2,i}} \dots, \boldsymbol{w_{l,i}}\}$ with $\mathrm{j} = 1, 2, \dots, \mathrm{J}$ be the $\boldsymbol{j}$-th textual description of image $\boldsymbol{I_i}$ and $\boldsymbol{w_p}$ with $\mathrm{l} = 1, 2, \dots, \mathrm{L}$ be the words composing the textual description. Let $\boldsymbol{I_t}$ be a target image for which we want to generate a textual description summarizing its content. The proposed captioning system consists of two stages: 1) encoder-decoder framework to generate multiple captions for a target image $\boldsymbol{I_t}$ exploring beam-searching algorithm and 2) a retrieval-based image captioning system that first searches for similar images to $\boldsymbol{I_t}$ from the archive and then measures the lexical similarity between previously multiple generated captions and the reference captions of target similar images to choose the best generated caption. The lexical similarity is measured using consensus caption score [7].

2.1. Encoder-decoder framework with beam search

The architecture of the encoder-decoder framework is similar to [8]. A CNN architecture is used to encode the visual features of an image in a raw vector. The encoded features are then forwarded to a RNN decoder that translates them into semantic description.

During training, the task of the model is to maximize the probability of the correct description given an image I by using the following formulation:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{I,S} \log p(S|I;\theta) \tag{1}$$

in which θ are the parameters of the model, I is an image and S is one of the J correct descriptions of the image. Breaking the sentence into words and applying the chain rule the inner part of the summation of Eq. 1 becomes:

$$\log p(S|I) = \sum_{i=0}^{l} \log p(w_i|I, w_0, \dots, w_{i-1}) \tag{2}$$

where the dependency on θ is dropped for convenience.
As sentences are sequential data, the RNNs are a common choice to model them. The goal is to model $p(w_i|I, w_0, \dots, w_{i-1})$. Using RNN the probability $p(w_i|I, w_0, \dots, w_{i-1})$ can be modeled with a fixed length hidden state or memory $\boldsymbol{h_i}$ conditioned on the previous time steps $\boldsymbol{i-1}$. When a new input $\boldsymbol{x_i}$ arrives the memory is updated as following:

$$h_{i+1} = f(h_i, x_i) \tag{3}$$

The linear function $f(\cdot)$ used in this article is the Gated Recurrent Unit (GRU)[9] . This choice is made to deal with the vanishing and exploiting gradients which is a common problem of the RNN. The practical effect of vanishing and exploiting gradients is the tendency to forget far away previous information which may be of particular importance when predicting a new word. Another issue of Equation 3 is the representation of image information and its combination with words to form input $\boldsymbol{x_i}$. For image representation in this work we rely on a particular CNN architecture, *InceptionV3,* proposed in [10]. First let us explain the simple GRU and then its combination with InceptionV3 to generate a description of a content of an image.

The architecture of GRU is depicted in Figure 2. It consists of two gates to control the information flow through the network. A reset gate $\boldsymbol{r_i}$ which decides how much of the previous information (stored in $\boldsymbol{h_{i-1}}$) to keep and an update gate $\boldsymbol{z_i}$ that decides the way to combine the previous information with the new input to update the new state $\boldsymbol{h_i}$. The equation of the GRU are the following:

$$h_i = z_i \cdot h_{i-1} + (1 - z_i) \cdot h'_i \tag{4}$$

$$z_i = \sigma(W_z[x_i, h_{i-1}] + b_z) \tag{5}$$

$$h'_i = \tanh(W_h[x_i, r_i \cdot h_{i-1}] + b_h) \tag{6}$$

$$r_i = \sigma(W_r[x_i, h_{i-1}] + b_i) \tag{7}$$

where W and b are the weights and biases, respectively; $\sigma(\cdot)$ and $tanh(\cdot)$ are sigmoid and hyperbolic tangent, respectively. The multiplications with gates makes it possible to deal with exploding and vanishing gradients.

Once the features of the images are extracted through InceptionV3 CNN architecture, they are fed to GRU only

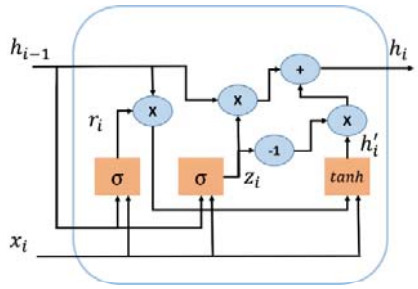

Figure 2. GRU architecture.

once, at time $i = -1$. The equations describing the whole process are as follows:

$$x_{-1} = CNN\ (I) \tag{8}$$

$$x_i = W_e w_i, \quad i \in \{0, \dots, l-1\} \tag{9}$$

$$x_{i+1} = GRU(x_i)\ , \quad i \in \{0, \dots, l-1\} \tag{10}$$

$$p(x_{i+1}) = \frac{e^{x_{i+1}}}{\sum_k^V (e^{x_{i+1}})_k} \tag{11}$$

where W_e is the word embedding layer that takes as input a sequence of integers representing the words of a sentence and convert them into an embedding space where words with similar meaning will be scattered near each other. Eq. 11 is the probability distribution of all the words in the vocabulary from which the next word will be sampled. The words and the image are mapped together in the same space using the word embedding and output of the CNN for words and image, respectively. w_0 and w_l are special tokens indicating the start and the end of a sentence, respectively. Finally, the best parameters of the model can be estimated by minimizing the following loss function:

$$Loss(I, S) = -\sum_{i=1}^{l} log p_i(w_i) \tag{12}$$

At inference stage, an image is input to the model and a sentence is built one word at a time. Most of the models, at each time step, sample the word with the highest probability. The process stops when the special token signaling the end of a sentence is sampled. This usually is referred as greedy search and produces only one description.

A more sophisticated model to generate image descriptions is beam-search algorithm. The peculiarity of this algorithm is the capability to generate more than one description for a target image. In particular, at each time step, instead of sampling the word with the highest probability, beam-search algorithm samples B best words where B is the beam size. At successive time steps other B words will be sampled starting from the first B ones. Up to time i this will generate $H_i = B^i$ word sequences (candidate descriptions) which is computationally prohibitive. That is why at each time step beam-search algorithm prunes the set of candidate descriptions by keeping only the best B ones. Finally, the algorithm selects the best description as the word sequence with maximum cumulative log-likelihood and discards the other $B - 1$ generated descriptions.

In this work we adopt the beam-search algorithm to generate more than one descriptions. Unlike the default implementation of the algorithm, we keep all the generated descriptions and choose the best one applying the caption consensus score as it is described in the following subsection.

2.2. Caption consensus score

Once the candidate captions are generated, to choose the best candidate, we apply nearest neighbor approach first to find similar images to a target image and then caption consensus score between the candidate captions and reference captions of the most K similar images to measure the lexical similarity. The candidate captions with the highest lexical similarity will be chosen. The image similarity between any two images is given by the Euclidean distance $d(I_i, I_j) = \|V_i - V_j\|_2$ between the feature vectors V of the images. The lexical similarity between the candidate captions and the reference captions of the most similar images is measured by applying BLEU score [11] which will be explained in the next section.

3. DATASET AND EXPERIMENTAL SETUP

To validate the proposed image captioning architecture, we utilized RSICD dataset [4]. Composed of more than 10000 images, is the biggest RS dataset utilized for image captioning. Each image is composed of five descriptions. The descriptions are written by five different people under particular instructions related to the characteristics of RS images. The images size is fixed to 224×224.

To evaluate the accuracy of the generated captions we used BLEU metric [11]. BLEU metric measures the correlation of a generated caption with respect to one or more reference captions. The correlation is measured as the geometric average of n-gram (n-consecutive words) precision between a generated caption and reference captions. In this paper we have used n = 1, 2, 3, 4.

As it is mentioned in Section 2, we first generate multiple captions for a target image and then choose the best caption based on the lexical similarity with the reference captions of most similar images. We used a beam size $B = 5$ in the beam-search algorithm yielding 5 different candidate

Table 1. Comparison results in terms of BLEU score between the simple encoder-decoder (without beam-search) and the proposed image captioning framework.

Model	BLEU 1	BLEU 2	BLEU 3	BLEU4
Encoder-decoder without beam search	0.657	0.469	0.361	0.288
Proposed framework	**0.661**	**0.476**	**0.375**	**0.302**

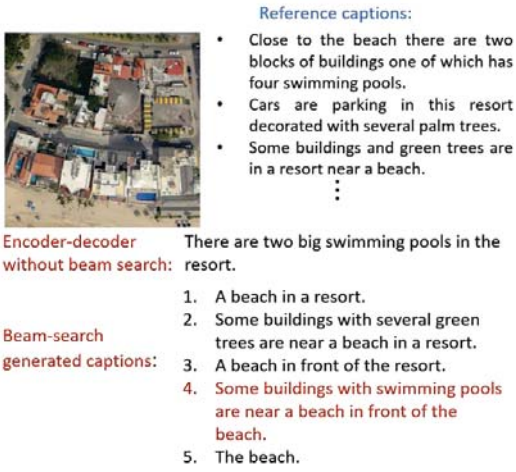

Figure 3. Example of generated captions through encoder-decoder with and without the beam-search algorithm. The caption highlighted in red is chosen by the proposed model to describe the image.

captions for each target image. We used four most similar images along with their 5 reference captions. Lexical similarity between the candidate captions and the reference ones is measured by applying BLEU score with $n = 2$.

Table 1. reports a comparison between the results of simple encoder-decoder framework and the proposed framework in terms of BLEU score. The proposed framework results are slightly higher with respect to the simple encoder-decoder architecture. Figure 3. shows an example of the captions generated through the encoder-decoder framework with and without beam-search algorithm. The ground truth descriptions are shown on the right of the image while the generated descriptions are shown below. We can see that all the generated captions describe, to some extent, well the content of the image. In this example the caption chosen by the proposed model is the fourth one (highlighted in red). Even though this caption is affected by some errors it still describes the image more accurately with respect to the first caption (see Figure 3. caption number 1.) that would have been the choice of beam-search algorithm or the caption generated by the simple encoder-decoder model (greedy search) without the beam-search algorithm. Even though this framework does not consider sophisticated image retrieval mechanisms and sentence scoring, it shows promising results.

4. CONCLUSION

In this paper we have presented a novel image captioning framework as a combination of generation and retrieval based image captioning. A CNN-RNN architecture combined with beam-searching generates multiple descriptions. A consensus score measures the lexical similarity between the generated descriptions and the reference descriptions of the similar images to choose the best description. A comparison between the results of the proposed framework and the simple CNN-RNN architecture are reported which shows that the proposed framework is promising. As a future work we plan to develop more sophisticated lexical similarity scoring and image retrieval mechanism.

11. REFERENCES

[1] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 2016, pp. 1–5.

[2] X. Zhang, X. Li, J. An, L. Gao, B. Hou, and C. Li, "Natural language description of remote sensing images based on deep learning," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 4798–4801.

[3] Z. Shi and Z. Zou, "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[4] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[5] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic Descriptions of High-Resolution Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, 2019.

[6] G. Hoxha, F. Melgani, and B. Demir, "Retrieving Images with Generated Textual Descriptions," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5812–5815.

[7] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring Nearest Neighbor Approaches for Image Captioning," *ArXiv150504467 Cs*, May 2015.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," *ArXiv14114555 Cs*, Nov. 2014.

[9] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, 2014, pp. 103–111.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818–2826.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2002, pp. 311–318.