RETRIEVING IMAGES WITH GENERATED TEXTUAL DESCRIPTIONS

Genc Hoxha[1], Farid Melgani[1] and Begüm Demir[2]

[1]Department of Information Engineering and Computer science, University of Trento, Trento, Italy
[2]Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany
(email: genc.hoxha@unitn.it, farid.melgani@unitn.it, demir@tu-berlin.de)

ABSTRACT

This paper presents a novel remote sensing (RS) image retrieval system that is defined based on generation and exploitation of textual descriptions that model the content of RS images. The proposed RS image retrieval system is composed of three main steps. The first one generates textual descriptions of the content of the RS images combining a convolutional neural network (CNN) and a recurrent neural network (RNN) to extract the features of the images and to generate the descriptions of their content, respectively. The second step encodes the semantic content of the generated descriptions using word embedding techniques able to produce semantically rich word vectors. The third step retrieves the most similar images with respect to the query image by measuring the similarity between the encoded generated textual descriptions of the query image and those of the archive. Experimental results on RS image archive composed of RS images acquired by unmanned aerial vehicles (UAVs) are reported and discussed.

Index Terms—Image retrieval, image textual description generation, semantic gap, unmanned aerial vehicles (UAV).

1. INTRODUCTION

With the fast development of earth observation satellite missions (such as Landsat and Sentinel) and their continuous information acquisition, the amount and the variety of the Remote Sensing (RS) image datasets are exponentially increasing. Therefore, the need for processing and retrieving information carried on those datasets is becoming a big challenge nowadays. As a matter of fact, image retrieval is crucial for expressing big datasets in a structured and comprehensive way for the community.

The most popular image retrieval approach in RS has been the content based image retrieval (CBIR) [1], [2]. It principally focuses on the extraction of low level features (such as color, texture and shape features) of an image. The main problem with the CBIR method is the difficulty that they present in extracting high level semantic content which includes the characteristics of RS images associated with the semantic information (such as the presence of objects or events) within the image. Indeed, bridging the "semantic gap" between the low level features and the high level semantic content remains still a challenge task. To reduce the semantic gap and improve the retrieval accuracy multilabel RS image retrieval has been recently proposed [3], [4]. The main idea is that within the RS images different sub-classes may be found that could enrich the semantic information within the image. Once the labels are obtained, in [4] they use those labels to create regions adjacency graph (RAG) for each image. The created RAG is then used in the graph matching algorithm in order to compute image similarity. Another attempt to reduce the semantic gap could be representing the RS images by textual descriptions. Textual descriptions are also more suitable for humans to describe the content of an image [5]. Also humans prefer to use textual description as query for retrieving the desired images as it allows to express and describe better their thoughts about the query image. However, collecting RS image descriptions is time consuming and costly.

In this work, we propose a system that generates and exploits textual descriptions of RS images for retrieval purposes. In order to overcome the issue of collecting RS image descriptions, we automatically generate the descriptions and exploit them for RS image retrieval. To the best of our knowledge this is the first work in the RS community that uses the generated RS image descriptions for retrieval purposes. The proposed system consists of three main steps: 1) image textual description generation; 2) textual description encoding; and 3) image retrieval using the generated textual descriptions.

2. PROPOSED RS IMAGE RETRIEVAL SYSTEM

Let $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_N}]$ be a dataset consisting of $\boldsymbol{N}$ remote sensing images and $\boldsymbol{X_i}$ 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 $\boldsymbol{X_i}$ and $\boldsymbol{w_p}$ with $\mathrm{l} = 1, 2, \dots, \mathrm{L}$ be the words composing the textual description. Let $\boldsymbol{X_q}$ be the query image for which we want to perform the retrieval**.** The proposed image retrieval methodology consists of: 1) an image caption generator block, 2) a sentence encoding block and 3) a similarity retrieval block using the generated sentence encoding to retrieve the desired number of most

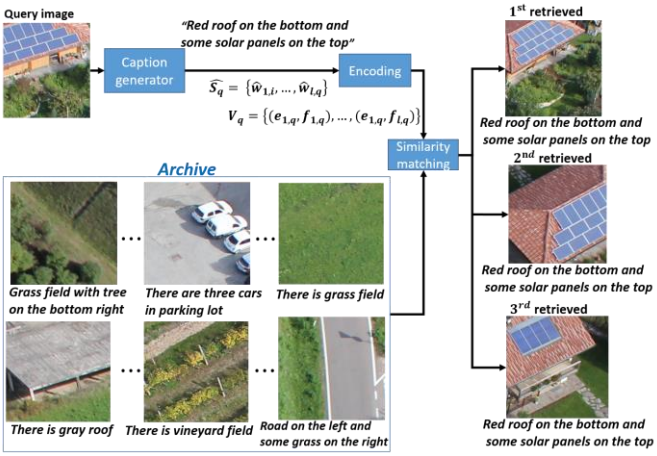

Figure 1. Block diagram of the proposed image retrieval system.

similar images $\boldsymbol{Y} = [\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_r}]$ with respect to the query image $\boldsymbol{X_q}$. In Figure 1 the block diagram of the proposed system during the test phase is illustrated.

2.1 Image textual description generation

The task of image textual description generation is to generate natural language description of the content of an image. For the text generation in this work, we resort to the long short-term memory (LSTM) [6] which is a special case of the recurrent neural networks (RNN).

RNNs have shown great success in natural language process (NLP) field in word prediction task. Sentence generation is based on the human thoughts where the prediction of new words depends on the previous ones. The main feature of the RNN is that the aforementioned property is satisfied by means of feedback loops which make the information to persist through the network. However, RNN suffers the long-term dependency which occurs when the prediction of a new word is related to a faraway previous information. To address this problem in [6] the LSTM is introduced. LSTM is composed of a cell state which allows the unchanged flowing of information through the network and three gates which are used to control the information flow through the cell. In our system, the word predictions are also conditioned on the image content. Thus, we extract the image feature using a pre-trained convolutional neural network (CNN). In particular, we use the ResNet50 model [7]. The words (composing the sentences) are encoded using one-hot encoding having dimension of the vocabulary size and then projected to an embedding layer that is able to explore their semantic content. The sentences are represented as a sequence of individual word embedding. The word embedding (composing the sentences) are given as input to the LSTM that stores and learns the semantic temporal context of words through its recurrent layers. The final output of the LSTM is concatenated with image features in a 'multimodal' feedforward layer to generate textual descriptions of the content of an image. At inference stage we input the image to the model and obtain the generated description of its content.

2.2 Sentence encoding

Each word of the generated descriptions is transformed into a vector of numbers using two different recent word embedding techniques: *word2vec* and *GloVe* [8]. Both techniques are based on co-occurrence of words in order to take into account context in a text represented by the neighboring words.

The *word2vec* is trained on a feed-forward neural network using two predictive models, continuous bag of words (CBOW) and skip-gram model to learn the embedding of the words. CBOW model attempts to predict a word given its context, while skip-gram attempts to predict the context from a given word. In this work we use *fastText* [9], a faster version of *word2vec* which takes into account the word morphology. This technique is based on the skip-gram model and each word is represented as a sum of its n-gram character vectors. However, *word2vec* embedding technique has the limitation of not taking into account the global co-occurrence of the words in the whole corpus. In order to capture the global statistical information of a text corpus, *GloVe* combines the global matrix factorization with the skip-gram model. In addition to the probability of words in the context, it takes into account the ratio of co-occurrence probabilities. In this phase the generated sentences $\widehat{\boldsymbol{S_q}} = \{\hat{\boldsymbol{w}}_{\boldsymbol{1,i}}, \ldots, \hat{\boldsymbol{w}}_{\boldsymbol{l,q}}\}$ are encoded as $\boldsymbol{V_q} = \{(\boldsymbol{e_{1,q}}, \boldsymbol{f_{1,q}}), \ldots, (\boldsymbol{e_{l,q}}, \boldsymbol{f_{l,q}})\}$ where $\boldsymbol{e_{l,q}}$ is the word embedding and $\boldsymbol{f_{l,q}}$ is the word frequency in the sentence normalized by the number of unique words composing the sentence.

2.3 Image retrieval using the generated descriptions

Using the encoded vector of the generated description, the similarity of any two images could be measured calculating the distance between their generated encoded vector. In order to explore the semantic information embedded in generated description vectors in the retrieval process we adopt the word mover's distance (WMD) which is a special case of Earth Mover's Distance [10] applied to space documents. The WMD [11] takes the advantage of *word2vec* and *GloVe* capability to embed the semantic information of words in the vector space to create a dissimilarity measurement of two sentences (or documents) as the minimum distance in the embedding space to transform the words of one sentence to the words of another sentence. Assuming that text documents are represented as normalized bag of words (nBOW) where the frequency of the $\boldsymbol{p}$-th word in the document is given as $\boldsymbol{f_p} = \frac{\boldsymbol{w_p}}{\sum_{k=1}^{n} \boldsymbol{w_k}}$ where w_p represents the frequency that word $\boldsymbol{p}$ appears in the document and $\boldsymbol{n}$ is the number of unique words of the document. Let $d(p,k) = \|v_p - v_k\|_2$ be the Euclidean distance in the embedding space of any two words indicating the word dissimilarity. The WMD extends the word dissimilarity to document dissimilarity. More specifically, let

S and S' be two sentences represented as nBOW. Each word $\boldsymbol{p}$ in S is transformed into any word in S' in total or in parts. In [11] the flow matrix $T\epsilon\, R^{n\times n}$ is introduced where $T_{p,k} \geq 0$ determines how much of word $\boldsymbol{p}$ in S travels to word $\boldsymbol{k}$ in S'. Basically, the flow matrix measures the effort needed to transport the histogram weight of one word from one document to every word in the other document. Then the minimum cumulative cost of moving one document to another under constraints is given by solving the following linear problem,

$$\min_{T\geq 0} \sum_{p,k=1}^{n} T_{p,k} \cdot d(p,k) \quad p,k \,\epsilon\, \{1,2,\dots,n\} \tag{1}$$

$$subject\ to: \sum_{k=1}^{n} T_{p,k} = f_p \qquad \sum_{p=1}^{n} T_{p,k} = f_k \tag{2}$$

where the first constraint $\sum_{k=1}^{n} T_{p,k} = f_p$ assures that the total flow from word **p** in S is totally transported to word **k** in S' and the second constraint $\sum_{p=1}^{n} T_{p,k} = f_k$ assures that word **k** receives all the incoming flow. After estimating the WMD between the generated description of the query image and those of all the images in the archive, the images that have the lowest distance with respect to the query image are retrieved.

3. EXPERIMENTAL RESULTS

In order to validate the proposed method, we used images acquired by unmanned aerial vehicles (UAVs) with EOS 550D camera near the city of Civezzano, Italy on October 17, 2012. This dataset has 10 RGB images of pixel size 5184×3456 characterized by a spatial resolution of 2 cm. The dataset is split into training (7 images) and test (3 images) sets. For the purpose of this work, we generated non-overlapping frames of size 256×256 for both the training and test sets. In total there are 2058 and 882 frames in the training and test sets, respectively and each frame is composed of three text descriptions written by three different human annotators. Example of frames along with the description is shown in Figure 2.

The metric used in this paper is BLEU [12]. BLEU metric is based on the *precision* measure. Precision is computed as the number of consecutive words (n-grams) occurring in the reference sentence divided by the total number of words in the candidate sentence. More precisely, supposed to have a generated description G and a real description (reference) R, BLEU score between G and R is computed as follows:

$$BLEU(N,G,R) = P(N,G,R) \times BP(G,R) \tag{3}$$

where $P(N,G,R) = (\prod_{n=1}^{N} p_n)^{1/N}$ is the geometric mean of n-gram precision, $p_n = m_n/l_n$, m_n is the number of matched n-grams between G and R, l_n is the total number of n-grams

1. A grass field with some asphalt on the bottom right.
2. There is a grass field.
3. Asphalt at bottom right is close to a grass field.
4. Grass field with some low vegetation on the bottom right.

1. A gray roof on the top and solar panels on the bottom.
2. A gray roof on top and a solar panel on bottom.
3. A gray roof at upper left is close to rocks at upper right and to solar panels at center.
4. Gray roof topped by solar panels.

1. There are four cars on the asphalt and a white building facade on the top.
2. Four cars parked on an asphalt and some grass on top left.
3. A white building facade is next to asphalt and to two cars.
4. There are two cars on asphalt and some grass on the right.

Figure 2. An example of three images from the dataset. The sentences from 1 to 3 correspond to Ground Truth data and sentence 4 (highlighted by red) is the generated sentence.

in G and $BP(G,R) = \min\left(1.0, \exp\left(1 - \left(\frac{len(R)}{len(G)}\right)\right)\right)$ is a brevity penalty if the length of the generated sentence G $len(G)$ is smaller than the one of reference $len(R)$. When there is no higher order n-gram precision $(e.g.\ n = 4)$ in a sentence, the entire BLEU score of the sentence is 0 independently from the quantity of the lower n-grams $(n = 1,2,3)$ matching found in the sentence. Therefore we use a smoothing technique proposed in [13] which replaces the 0 score, in presence of low order n-grams, to a small positive value ε. BLEU score ranges from 0 to 1 where 1 is good. For the n-gram precision we used $n = 1,2,3,4$.

As it was mentioned in Section 2 we have used two different encoding techniques to convert the words of the generated sentences into vectors. The *GloVe* vectors we used are pre-trained on Wikipedia 2014 + Gigaword 5 corpus and are available on the Stanford website [14] for free. The *fastText* vectors instead are trained in our own corpus. From our empirical results we have chosen the word vectors to be of size 50 as a tradeoff between the computational time and accuracy. In Table 1 we report the results in terms of mean BLEU score per query image in which we use the image "Ground Truth" sentences for retrieval purposes. The methodology applied is the same as the one in Figure 1 without the caption generator block. The obtained results represent the upper bound of the proposed methodology regarding the considered dataset. In Table 2 we report the results using the automatically generated sentences for retrieval purposes. As there is no other work to perform a comparative study, we make a comparison between the results of the two tables. We can notice, in terms of mean BLEU score, an average gap of 0.3 with respect to the upper bound. The two encoding techniques show rather similar results. The reported results are affected by many factors, for example by the caption generator block shown in Figure 1. Indeed, observing Figure 2, we can see that for the top and bottom images, the generated sentences are affected by some errors. One way to get closer to the upper bound result could be improving the caption block. An example of the retrieved

images is shown in Figure 3. The query image is highlighted in red and the order of retrieved images is given above each retrieved image. Though the automatically generated descriptions are characterized by some errors as already stressed before, we can see that the five retrieved images are semantically similar to the query. They all show cars (from one to three) parked in a parking lot.

Table 1. Upper bound results in terms of mean BLEU score per query image. For the query and retrieved images the "Ground Truth" sentences are used**.**

Embedding	# of Retriev.	Bleu 1	Bleu 2	Bleu 3	Bleu 4
GloVe	1	0.991	0.851	0.806	0.741
	5	0.880	0.811	0.761	0.690
	10	0.859	0.786	0.735	0.663
fastText	1	0.895	0.846	0.806	0.746
	5	0.858	0.801	0.759	0.694
	10	0.839	0.773	0.729	0.664

Table 2. Results obtained by the proposed system in terms of mean BLEU score per image. For the query and retrieved images the generated sentences are used.

Embedding	# of Retriev.	Bleu 1	Bleu 2	Bleu 3	Bleu 4
GloVe	1	0.605	0.519	0.473	0.417
	5	0.574	0.487	0.439	0.382
	10	0.559	0.469	0.422	0.361
fastText	1	0.605	0.519	0.473	0.417
	5	0.575	0.487	0.439	0.382
	10	0.558	0.469	0.421	0.360

4. CONCLUSION

In this paper, we have presented a semantic image retrieval method based on generated textual descriptions which attempt to explore the high level semantic content incorporated in the generated descriptions. A comparison between using the real descriptions and the generated descriptions for RS image retrieval purpose is made, from which we can notice that there is an average gap of 0.3 in terms of mean BLEU score. In order to reduce this gap and improve the retrieval performances, in a future work we plan to improve the caption generation block.

5. REFERENCES

[1] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, Sep. 2003.

[2] Y. Li and T. R. Bretschneider, "Semantic-Sensitive Satellite Image Retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 853–860, Apr. 2007.

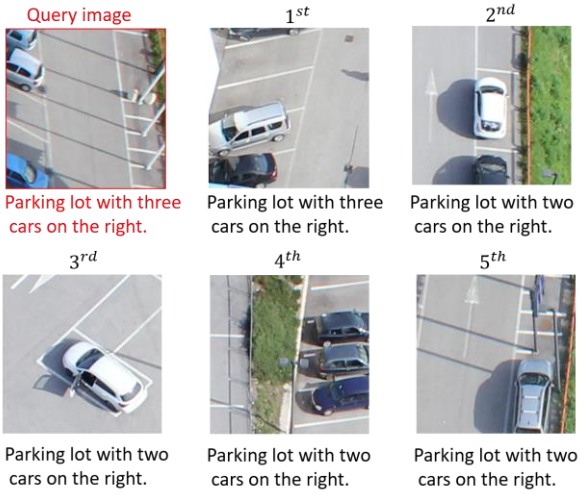

Figure 3. Example of a query image and five retrieved images.

[3] M. Wang and T. Song, "Remote Sensing Image Retrieval by Scene Semantic Matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2874–2886, May 2013.

[4] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[5] M. L. Kherfi, D. Ziou, and A. Bernardi, "Image Retrieval from the World Wide Web: Issues, Techniques, and Systems," *ACM Comput Surv*, vol. 36, no. 1, pp. 35–67, Mar. 2004.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ArXiv14091556 Cs*, Sep. 2014.

[8] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, Dec. 2017.

[10] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[11] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From Word Embeddings To Document Distances," p. 10.

[12] 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.

[13] B. Chen and C. Cherry, "A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, 2014, pp. 362–367.

[14] "GloVe: Global Vectors for Word Representation." [Online]. Available: https://nlp.stanford.edu/projects/glove/. [Accessed: 07-Jan-2019].