Toward Remote Sensing Image Retrieval Under a Deep Image Captioning Perspective

Genc Hoxha, *Student Member, IEEE*, Farid Melgani, *Fellow, IEEE*, and Begüm Demir, *Senior Member, IEEE*

***Abstract*—The performance of remote sensing image retrieval (RSIR) systems depends on the capability of the extracted features in characterizing the semantic content of images. Existing RSIR systems describe images by visual descriptors that model the primitives (such as different land-cover classes) present in the images. However, the visual descriptors may not be sufficient to describe the high-level complex content of RS images (e.g., attributes and relationships among different land-cover classes). To address this issue, in this article, we present an RSIR system that aims at generating and exploiting textual descriptions to accurately describe the relationships between the objects and their attributes present in RS images with captions (i.e., sentences). To this end, the proposed retrieval system consists of three main steps. The first step aims to encode the image visual features and then translate the encoded features into a textual description that summarizes the content of the image with captions. This is achieved based on the combination of a convolutional neural network with a recurrent neural network. The second step aims to convert the generated textual descriptions into semantically meaningful feature vectors. This is achieved by using the recent word embedding techniques. Finally, the last step estimates the similarity between the vectors of the textual descriptions of the query image and those of the archive images, and then retrieve the most similar images to the query image. Experimental results obtained on two different datasets show that the description of the image content with captions in the framework of RSIR leads to an accurate retrieval performance.**

***Index Terms*—Convolutional neural network, deep learning, image captioning, image retrieval, recurrent neural network, remote sensing, semantic gap.**

I. Introduction

RECENT advances in satellite technology result in an explosive growth of remote sensing (RS) image archives. Thus, one of the important research topics is the development of accurate RS image retrieval (RSIR) systems to retrieve the most relevant images to a query image from such massive archives. To this end, in the RS community, a great attention is devoted to content-based image retrieval that aims to search and retrieve the most similar images to a query image based on two main steps: 1) description of images by a set of visual features that model the primitives (such as different land-cover classes) present in the images; and 2) retrieval of images that are similar to the query image by evaluating the similarity between the features of the query image and those of the archive images [1]. The traditional content-based RSIR systems rely on hand-crafted features to describe the semantic content of images. To this end, several visual descriptors are presented in RS. As an example, bag of-visual-words representations of the scale invariant feature transform features are introduced in [2]. In [3], histogram of local binary patterns that models the relationship of each pixel in a given image with its neighbors (which are located on a circle around that pixel) by a binary code is presented. Graph-based image representations, where the nodes model region properties and the edges represent the spatial relationships among the regions, are introduced in [4]–[6]. Descriptors of bag of spectral values are introduced in [7] to model the spectral information content of high-dimensional RS images. After defining the image visual features (i.e., visual descriptors), image retrieval can be achieved by considering unsupervised or supervised retrieval methods. Unsupervised methods compute the similarity between the visual features of the query image and those of the archive images and then retrieve the most similar images to the query. To this end, one can simply use the k-nearest neighbor algorithm. If the images are represented by graphs, graph matching techniques can be used. As an example, an inexact graph matching strategy that jointly exploits a subgraph isomorphism algorithm and a spectral embedding algorithm [5] can be used. Supervised methods require an availability of a set of annotated images for the training of the classifier. If the training images are annotated by single high-level category labels, any binary classifier could be exploited [8]. If the training images are annotated by low-level land cover class labels (i.e., multilabels), multilabel image retrieval methods are required. In [9], a sparse reconstruction-based method that generalizes the standard sparse classifier to the case of multilabel RS image retrieval problems is introduced.

Recent advances in deep neural networks have led to a significant performance gain in terms of content-based RSIR with respect to traditional systems. Deep learning-based RSIR systems simultaneously optimize feature learning and image retrieval [9]–[15]. Deep feature representations based on convolutional neural networks (CNNs) are introduced in the framework of the RSIR in [12] and [15]. In [13], a retrieval method that exploits a weighted distance measure that is applied to the image features obtained by a CNN is presented. A re-ranking method

Manuscript received January 23, 2020; revised April 19, 2020, June 18, 2020, and July 22, 2020; accepted July 27, 2020. Date of publication August 3, 2020; date of current version August 20, 2020. This work was supported by the European Research Council through the ERC-2017-STG BigEarth Project under Grant 759764. *(Corresponding author: Farid Melgani.)*

Genc Hoxha and Farid Melgani are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: genc.hoxha@unitn.it; melgani@disi.unitn.it).

Begüm Demir is with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10623 Berlin, Germany (e-mail: demir@tu-berlin.de).

Digital Object Identifier 10.1109/JSTARS.2020.3013818

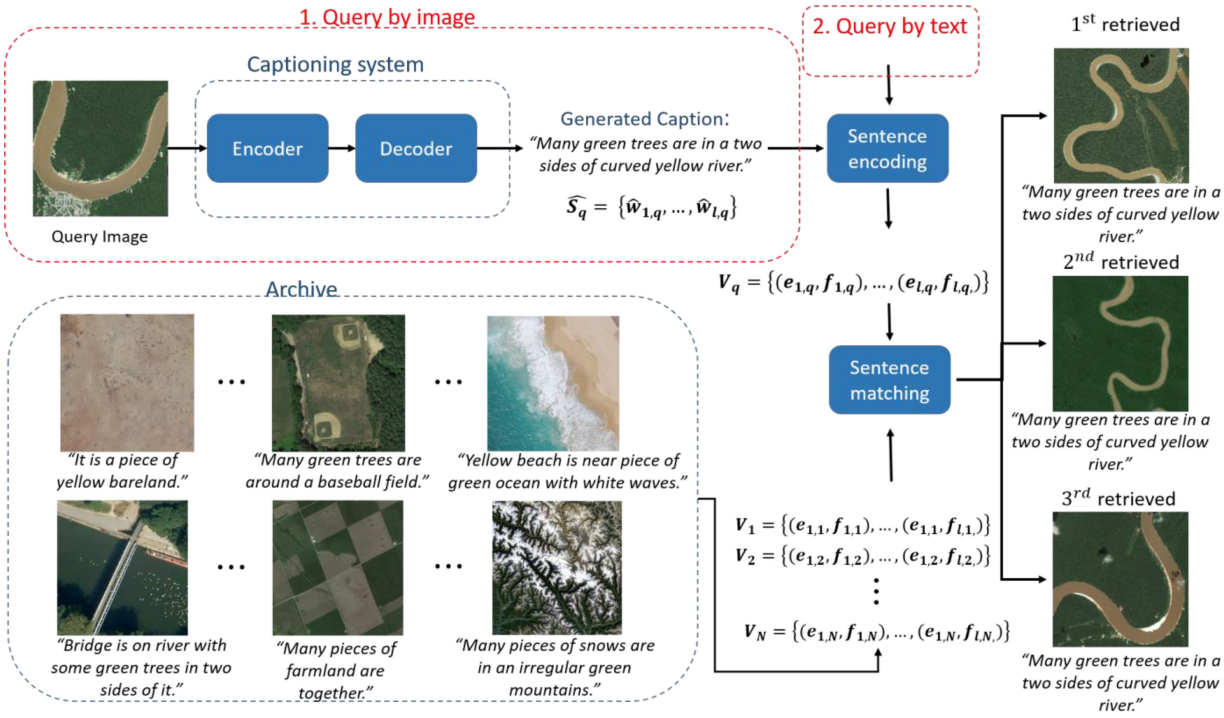

Fig. 1. Block diagram of the proposed retrieval system. Configuration 1 allows users to query and retrieve images from the archive using image as query. Configuration 2 lets users use directly a textual description to query and retrieve images from the archive. In this work, the default scenario is configuration 1. Sentence encoding block produces a tuple vector $\boldsymbol{V_q} = \{(e_{1,q}, \boldsymbol{f}_{1,q}), \ldots, (e_{l,q}, \boldsymbol{f}_{l,q})\}$ of the generated sentence (textual description) $\widehat{S_q} = \{\widehat{w_{1,q}}, \widehat{w_{2,q}}, \ldots, \widehat{w_{l,q}}\}$ of the query image where $e_{l,q}$ is the word embedding and $f_{l,q}$ is the normalized frequency of the l-th word in the generated sentence of the query image. Sentence matching measures the similarity between the generated sentence vector $\boldsymbol{V_q}$ of the query image and those of the archive $\boldsymbol{V_N} = \{(e_{1,N}, \boldsymbol{f}_{1,N}), \ldots, (e_{l,N}, \boldsymbol{f}_{l,N})\}$ where N is the total number of images in the archive. Note that the descriptions of the images found in the archive are all generated.

that represents each RS image with CNN features and then applies image-to-class distance measures for retrieval problems is presented in [14]. A Siamese graph convolution network that assesses the similarity between a pair of graphs that can be trained with the contrastive loss function is introduced in [10]. Image representations obtained through binary codes are discussed particularly for scalable image search and retrieval in [9] and [11]. To obtain the binary codes, in [9], a deep hashing neural network that exploits the cross-entropy loss is presented, whereas in [11] a metric learning-based deep hashing network that uses triplet loss function (instead of the cross-entropy loss) is presented.

The performance of the above-mentioned retrieval methods depends on the image descriptors that model the visual semantics of the considered images in the archives. However, these descriptors can have limitations in modeling the primitives (i.e., attributes and relationships between different land-cover classes) present in the images. It is important to note that there are usually several areas within each RS image associated with different land-cover classes. Thus, describing an RS image with a visual image descriptor may lead to limited retrieval accuracy particularly when high-level semantic content is present in the images. To address this issue, in this article, we present an image retrieval system that generates and exploits textual descriptions through image captions of RS images. The proposed retrieval system consists of three main blocks: image captioning; a sentence encoding; and similarity matching. In the first step, a CNN is initially used to extract the visual features of RS images and then a recurrent neural network (RNN) is employed to generate a textual feature from the visual features. In the second step, the semantic meaning of the generated sentences is encoded on the basis of recent word embedding techniques that are capable of producing semantically rich word vector representations. Finally, in the last step, the semantically rich word vectors are exploited to search and retrieve the most similar images to the query image from the archive. By this way, image retrieval is applied through the estimation of similarities among the generated textual descriptions instead of considering the visual descriptors. The proposed system can also be configured to allow one to use directly the textual descriptors as query to retrieve the most similar images. Fig. 1 shows the block diagram of the proposed retrieval system. Experiments carried out on two different archives that include satellite and aerial unmanned aerial vehicle (UAV) images together with their captions demonstrate the effectiveness of the proposed system in terms of retrieval accuracy.

It is worth noting that image captioning methods have been recently introduced in the RS community to generate a coherent and comprehensive description of the complex semantic content of an image [16]–[23]. The contribution of this article consists in presenting and testing the effectiveness of the textual descriptors (i.e., image captions) in the framework of the RSIR problems to provide accurate search capability within big data archives in RS.

To the best of our knowledge, this is the first work in the RS community that achieves querying and retrieving images from the archive based on the textual descriptions. The proposed RSIR system has been briefly presented in [24] with limited experimental analysis. This article extends our work introducing a detailed description of the proposed approach with a thorough experimental analysis. Another work recently published is in [25], which proposes a deep bidirectional triplet loss to learn the similarity between an image and its descriptions in a common feature space. The basic idea is that the related image-text pairs should be closer than the unrelated pairs in the common feature space. The query is performed using one or multiple sentence descriptions. Note that our proposed work is different from the work in [25]. In our work one can search for similar images using either an image (by automatically generating a description of its content) or using directly a textual description as a query, whereas in [25] the query can be only in the form of a textual description.

The rest of this article is organized as follows. Section II discusses the related work about image caption, and Section III introduces the proposed retrieval system. Section IV describes the datasets used in the experiments and the experimental setup. Section V illustrates the experimental results, while Section VI draws the conclusion of this article.

II. Image Captioning Related Work

Image Captioning (IC) aims at automatically generating natural language descriptions that are capable of describing the content of an image [26]. This is achieved based on machine learning and natural language processing (NLP) techniques that initially extract the visual features of an image and then generate textual description from the visual features. Image captioning has been recently introduced in the RS community [16]–[23] and is still a topic which calls for further development and consolidation.

Unlike RS, in the computer vision and multimedia communities, the use of captioning is more extended and widely studied for the representation of the semantic content of an image [27]–[32]. From a methodological point of view, RS image captioning methods can be divided into three main categories:

1) template-based;
2) retrieval-based; and
3) generation-based methods.

Template-based IC methods are composed of prefixed textual (sentence) templates with empty slots. Detection algorithms are first used to detect different objects present in the images along with their attributes. Then, the empty slots are filled with the detected entities forming a sentence description. As an example, Shi and Zou [18] proposed to leverage different fully convolutional neural networks in order to detect different objects present in the images and a language model based on fixed templates to generate the image descriptions. In general template-based IC methods produce textual description that might be correct from a grammatical and content viewpoint. However, the generated descriptions tend to be simple due to the prefixed templates and the performance of this method highly depends on the object detection algorithm used in the first stage.

A second methodology of IC in RS is retrieval-based IC. In this methodology, the captioning task is treated as a retrieval problem. Given a target image, this method first looks into the archive for the most similar images together with their descriptions and then assigns to the target image one (or more) existing description(s) of the retrieved most similar images. As an example, Wang *et al*. [20] mapped images and sentences in the same semantic space and developed a distance metric to learn the similarity between images and sentences. To a target image, the sentences that have the smallest distance are assigned to describe its content. In general, retrieval-based IC methods produce sentences that are correct from a syntax and grammatical point of view (if the archive is properly built) but they are not unseen-generated captions. Furthermore, as the sentences are retrieved from similar images, their content may be irrelevant to the target image.

A third methodology is generation-based IC. It is the most used methodology in RS and computer vision communities as it produces novel descriptions, which are very similar to those written by humans. Generation-based IC is usually based on the encoder-decoder framework, where a CNN is used to extract the image visual features, and then a sequence model such as an RNN is used to generate the image content description [16], [17], [19], [21]–[23]. As an example, an approach that combines different CNNs and RNNs to generate textual descriptions for high spatial resolution RS images is presented in [16].

III. Proposed Method

A. Problem Formulation

Let $\mathrm{X} = \{\mathrm{X}_1, \mathrm{X}_2, \ \ldots, \mathrm{X}_\mathrm{N}\}$ be an archive of N images and X_i be the ith image present in the archive. Each image in the archive is associated to J ground truth textual descriptions (i.e., captions). Let $S_{i,j} = \ \{w_{1,j}, w_{2,j}, \ldots, w_{p,j}\},\ j = 1, 2, \ldots, J$ be the jth textual description of the image X_i and $w_p,\ p = 1, 2, \ldots, P$ be the words of the textual description. Let X_q be the query image that can be selected by the user. Given a query image X_q, we aim to find a set $Y = \{Y_1,\ Y_2, \ldots, Y_r\}$ of the most similar images to X_q from the archive with a high accuracy. To this end, the proposed methodology consists of three main steps, which are as follows:

1) image caption generation;
2) sentence encoding; and
3) image retrieval based on the encoded sentences of images.

The block diagram is shown in Fig. 1.

B. Image Caption Generation

Due to the success of the generation-based image captioning systems in RS community, in this article, we focus our attention on the use of the generation-based systems in the framework of RSIR. In detail, we define the textual descriptions of the RS images based on a multimodal RNN. Multimodal RNN is a combination of an RNN and a CNN in order to model the language descriptions and the image visual content in a

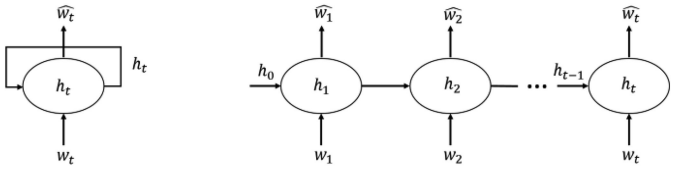

Fig. 2. RNN architecture.

unique multimodal layer [29]. The RNN learns the dense feature embedding of the words in the dictionary and keeps track of the semantic temporal context using its recurrent layers. The CNN extracts the visual features of the RS images. The multimodal layer combines the previously extracted word feature with the image features in a unique layer representation in order to generate word-by-word description of the RS image content. As RNNs are affected by gradient vanishing/exploding problem, they have limitations in the cases in which the prediction of a new word is related to a faraway previous information. To overcome the long-terms dependency problem, in this work, we exploit the long short-term memory (LSTM) [33] that is a type of RNNs that enable the long-range learning. In the following, we introduce the simple RNN and the LSTM. Later, we present the multimodal RNN.

1) Simple Recurrent Neural Networks: In NLP community, RNNs have shown great success in different tasks such as speech recognition [34], [35]. Through their recurrent connections they have the ability to explore sequence data characterized by an inner temporal relationship. An example of sequence data are sentences. Sentences are composed of sequence of words correlated to each other within a semantic context. Thus, to generate a sentence, the knowledge of previous words is required. RNNs have feedback loops that allow the flow and the storage of the semantic temporal context in order to produce meaningful sentences. The RNN architecture is illustrated in Fig. 2. At each time step t , an input word w_t is passed to the RNN hidden state h_t. The hidden state h_t acts as a "memory" containing all the past information. The output (predicted) word $\widehat{w_t}$ is a function of previous information stored in the hidden state h_t and the current input word w_t. The equation involved in the RNN is reported in (1) and (2), in which $f_1\,(.), g_1(.)$ are element-wised sigmoid and softmax function, respectively, and U, V are weights to be learned through backpropagation

$$h_t = f_1\left(U_w w_t + U_h h_{t-1}\right) \tag{1}$$

$$\widehat{w_t} \;= g_1\left(V_h h_t\right)\;. \tag{2}$$

However, as mentioned before, RNNs are affected by long-terms dependency problem. To overcome this problem, LSTMs have been introduced in [33]. The structure of the LSTM is more complex than the simple RNN. Within the LSTM are found a cell state and three gates to control the information flow through the network as illustrated in Fig. 3. The first step of the LSTM is to decide which information to cancel from the previous cell state c_{t-1}. The previous hidden state h_{t-1} together with the current input word w_t are first passed through the forget gate represented by a sigmoid function which outputs a number between 0 and 1 stating that if the output of forget gate $f_t = 0$ information has to be completely forgotten otherwise it has to be kept. Then h_{t-1} and w_t are passed to the input gate represented again by a sigmoid function and to a tan*h* layer to decide the value to be updated and the candidates to such values, respectively. The outputs of input gate i_t and of the tan*h* layer $\overline{c_t}$ multiplied together are added to the multiplication between forget gate output f_t with the previous cell state c_{t-1} to finally update the current cell state c_t. Finally, the previous hidden state h_{t-1} and the current word w_t are passed to another sigmoid function to output o_t. The new state h_t of the LSTM will be formed as the multiplication of o_t with the filtered version of current cell state c_t. Filtering of the cell state c_t is done by a tan*h* layer. Equations (3)–(9) describe the LSTM inner layers and information update, where W represents the weight parameters to be learned and $*$ represents the Hadamard product

$$f_t = \sigma\left(W_{wf} \cdot w_t + W_{hf} \cdot h_{t-1}\right) \tag{3}$$

$$i_t = \sigma\left(W_{wi} \cdot\; w_t + W_{hi} \cdot h_{t-1}\right) \tag{4}$$

$$\overline{c_t} = \tanh\left(W_{wc} \cdot w_t + W_{hc} \cdot h_{t-1}\right) \tag{5}$$

$$c_t = f_t * c_{t-1} + i_t * \overline{c_t} \tag{6}$$

$$o_t = \sigma\left(W_{wo} \cdot w_t + W_{ho} \cdot h_{t-1}\right) \tag{7}$$

$$h_t = o_t * \tanh(c_t) \tag{8}$$

$$\widehat{w_t} \;= \mathrm{softmax}\left(W_h h_t\right). \tag{9}$$

2) Multimodal Recurrent Neural Network: The multimodal recurrent network is shown in Fig. 4(a). The inputs of multimodal RNN are the image features extracted using a CNN architecture and their related textual descriptions to estimate the probability distribution of the next word given the image features and the previous words. The architecture of the multimodal RNN is composed of an embedding layer which converts the input words from one-hot encoding into a dense vector representation allowing to encode the semantic meaning of the words. Then the embedding of words is passed in the recurrent layer of the LSTM to store the temporal semantic context. The multimodal layer combines the LSTM new state with the image features extracted by the CNN. For image representation, deep learning features have shown to overcome the need of hand-crafted feature [36]. Hence, we exploit ResNet50 [37] to extract the visual features of the image. As ResNet50 is a fully connected CNN pretrained on ImageNet for image classification task, we remove the last layer and use the penultimate layer to represent the image. Before inputting the extracted features to the multimodal layer, we pass them first to a fully connected layer (dense layer) of dimension *D* having ReLu activations $f(x) = \max(0, x)$ [38]. The reason of choosing the ReLu activation is due to the fact that it is less affected from the gradient vanishing problem during backpropagation [39]. The combination in the multimodal layer is done by adding both the LSTM hidden state with the image features forming the mixed vector of dimension M as the sum of the LSTM hidden state vector and image features vector dimensions. Another dense layer of dimension *D* and $ReLu$ activation is added before the last layer. The output of this layer is passed to the dense layer having dimension *V* of vocabulary

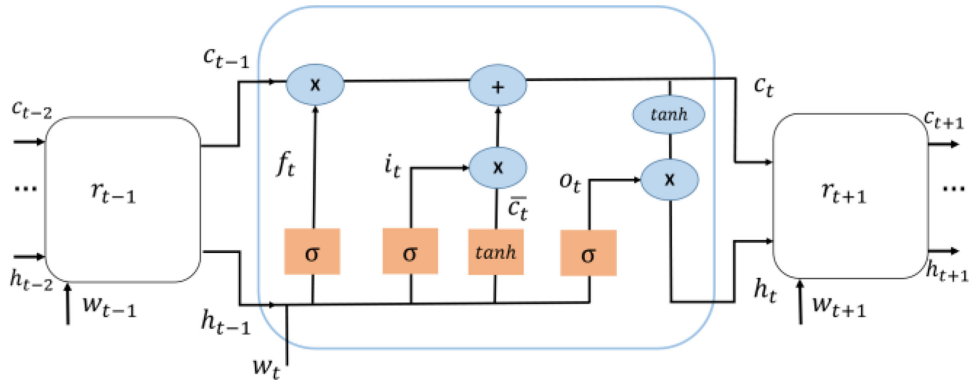

Fig. 3. LSTM architecture.

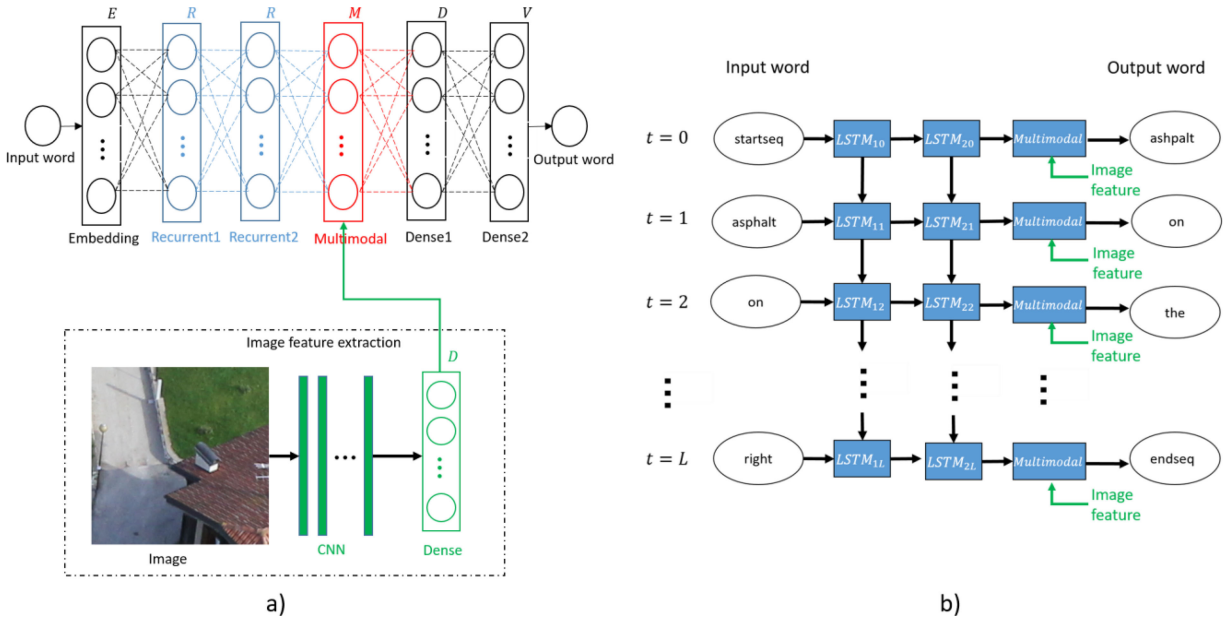

Fig. 4. Multimodal recurrent neural network. (a) Multimodal recurrent neural network architecture; and b) the word prediction at each time stamp t regarding the input image and its related sentence description (e.g. asphalt on the left and a red roof on the bottom right and some grass on the top right). "startseq" and "endseq" are special tokens denoting the start and the end of the sentence. L is the sequence (sentence) length. During test time only RS image is inputted to the model and word-by-word prediction is made regarding the image content until sampling "endseq" token.

size and softmax activation $f(x_i) = e^{x_i} / \sum_k^V e^{x_k}$ to predict the next word conditioned to the image features and the previous words. At inference stage, the RS image is fed into the model to predict word by word the textual description of its content.

C. Sentence Encoding

Once the generated descriptions for each image are obtained, they need to be scattered into a vector space able at exploring the semantic content within each description. This is achieved by representing each word with a real-valued vector. In our work these vectors are used as features in order to retrieve the most similar images in the archive to a query image. The word representation in the semantic vector space in this work is done using two different word embedding techniques: 1) *word2vec*; [40] and 2) *GloVe* [41]. Based on the word co-occurrence both techniques are capable at producing semantically rich word vectors. *Word2vec* is trained on a shallow neural network language model composed of an input layer, projection layer, and output layer to learn the word vector representations based on the nearby words [40]. *Word2vec* comes with two different predictive models: 1) the Continuous Bag of Words model; and 2) the Skip-gram model. The former attempts to predict a word given its context (nearby words), while the latter attempts to predict the context given a target word. In this work, we used *fastText* [42] which is a faster version of *word2vec* that takes into account the word morphology. This technique is based on the skip-gram model and the words are represented as a sum of their n-gram characters. However, *word2vec* does not take into account the global co-occurrence of words in the whole text corpus. *GloVe* technique combines the Skip-gram model with the global matrix factorization to explore the global statistical co-occurrence of the words in the whole corpus. Instead of focusing only the probability of words within a

context it also takes into account the ratio of co-occurrence probabilities in the whole corpus extracting information from the data repetition within a text corpus. The generated sentences $\widehat{S}_i = \{\widehat{w_{1,i}}, \widehat{w_{2,i}}, \ldots, \widehat{w_{p,i}}\}$ representing the image X_i are encoded as $V_i = \{(e_{1,i}, f_{1,i}), \ldots, (e_{p,i}, f_{p,i})\}$, where $e_{p,i}$ is the word embedding obtained by the two embedding techniques and $f_{p,i} = \widehat{w_{p,i}} / \sum_{k=1}^{p} \widehat{w_{k,i}}$ is the word frequency normalized by the total number of words in the sentence. The reason behind this representation is explained in the following section.

D. Image Retrieval Based on Generated Textual Descriptions

The final step of the proposed methodology consists of exploiting the generated sentences of each RS image to retrieve the desired number of most similar RS images in the archive given a query image. To this end, we need a metric that is capable of exploiting the semantic content encoded in each word using the two embedding techniques of the previous section. To this end, we exploit the word mover's distance (WMD) [43], which is a special case of the well-known earth mover distance [44] metric.

The WMD uses the word vectors scattered in the semantic vector space in order to create a dissimilarity measurement of any two sentences as the minimum distance needed to convert the words of one sentence into the words of another sentence. In detail, let $E \in R^{d \times n}$ be the word embedding matrix with a vocabulary size n. Let $e_i \in R^d$ be the d-dimensional encoding vector of word i. Let S and S' be two documents (or sentences) represented as a normalized Bag of Words vector, where $f_i = w_i / \sum_{k=1}^{n} w_k$ is the number of times word w_i appears in S divided by the total number of words composing S. Let $c(i,j) = \|e_i - e_{j_2}\|$ be the Euclidean distance in the semantic vector space between any two words i and j representing the word dissimilarity. Introducing an auxiliary matrix $T \in R^{n \times n}$ such that $T_{i,j} \geq 0$ denotes how much of word i in S should be transferred to word j in S', the work in [43] defines the distance between any two documents as the minimum cumulative cost necessary to move all the words from sentence S to S' solving the following linear problem:

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{i,j} \cdot c(i,j) \quad i,j \in \{1,2,\ldots,n\} \tag{10}$$

$$\text{subject to}: \sum_{j=1}^{n} T_{i,j} = f_i \qquad \sum_{i=1}^{n} T_{i,j} = f_j \tag{11}$$

where $\sum_{j=1}^{n} T_{i,j} = f_i$ states that the total flow from word w_i in S is fully transported to word w_j in S' and $\sum_{i=1}^{n} T_{i,j} = f_j$ states that the word w_j in S' receives all the incoming flow. Once the WMD distance between the generated description of the query image X_q and all of the other images in the archive is calculated, the images with the smallest distance to the query image are retrieved.

IV. Dataset Description and Experimental Setup

A. Dataset Description

In order to evaluate the proposed method, we used two different datasets:

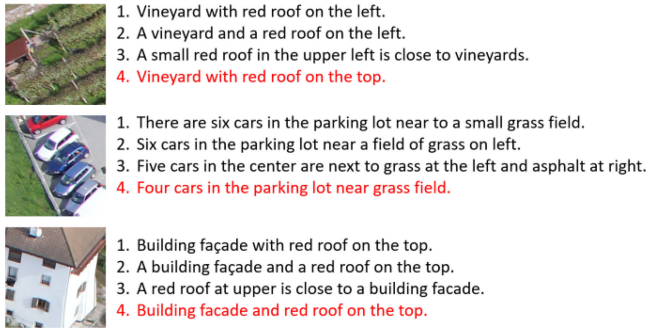

Fig. 5. Example of three images from the UAV dataset. The sentences from 1 to 3 correspond to ground truth sentence and sentence 4 (highlighted in red) is the generated sentence.

1) Unmanned Aerial Vehicles (UAV) Image Captioning Dataset: The first dataset consists of images acquired by UAVs with EOS 550D camera near the city of Civezzano, Italy, on October 17, 2012. The dataset is composed of ten RGB images of pixel size 5184×3456 characterized by a spatial resolution of 2 cm, of which six images are used for training, one image for validation, and three images for test. For the purpose of this work, frames of size 256×256 for training, validation and test sets are generated. In total there are 2940 frames and each of them is composed of three textual description written by three different human annotators. Examples of frames along with their descriptions are shown in Fig. 5. The vocabulary size V of the dataset is 185. Since we make a comparison between our method and multilabel image retrieval, each image is labeled with one or more labels based on the ground truth descriptions. The total number of the labels associated with the archive is $C = 16$. The labels composing the archive are: "Asphalt," "Grass," "Tree," "Vineyard," "Low Vegetation," "Car," "Gray Roof," "Red Roof," "White Roof," "Solar Panel," "Soil," "Gravel," "Rock," "Person," "Shadow," and "Building Facade."

2) Remote Sensing Image Captioning Dataset (RSICD): The second dataset is the RSICD dataset [19]. It is composed of more than 10 000 RS images gathered from different maps with various resolutions. Thus, it is the largest dataset used for RS image captioning. Each image has different number of descriptions varying from one to five. The images are fixed to 224×224 pixel size. The vocabulary size of this dataset is 3323. It has shown to be very useful for image captioning problems despite being affected by numerous misspellings. This popular benchmark dataset is unfortunately not suited for a straightforward conversion into a multilabel version. We therefore did not consider it for multilabel experiments.

B. Experimental Settings

As it was discussed in the previous section, our proposed method consists of image captioning, sentence encoding, and image retrieval blocks. The dimension of the embedding, recurrent and multimodal layers that compose the image captioning block are $E = R = M = 256$. The features of each image are obtained using the ResNet50. The obtained features are passed to a dense layer (fully connected layer) of dimension $D = 256$

with ReLu activations. In order to avoid overfitting, drop_out is also applied. Subsequent to the multimodal layer a dense layer having a dimension $D = 256$ with activation ReLu is applied. The output consists of a dense layer having softmax activations with vocabulary size dimension $V = 185$ and $V = 3323$ for UAV and RISCD datasets, respectively. We randomly selected 1) 60% of images to derive the training set; 2) 10% of images to derive the validation set and 30% of images for the test set. In the retrieval stage we unite the training and validation sets to construct the image archive and all the images in the test set for each dataset are used as query images to retrieve the most similar images from the archives with respect to query image X_q.

Sentence encoding is performed using *GloVe* and *fasText*. GloVe vectors are pretrained on Wikipedia 2014 + Gigaword 5 corpus. They are available at Stanford website [45]. The *fasText* vectors were trained separately in the two datasets corpus. The word vectors dimensionality is chosen as 50 for both the UAV and the RSICD dataset as a tradeoff between the accuracy and computational complexity.

C. Multilabel Image Retrieval System

In order to evaluate the performance of the proposed method, we compare with the multilabel image retrieval. As it was already mentioned before, the comparison with multilabeling is only done in the UAV dataset. The architecture of the multilabel method is the same as [46] with the difference on the last layer in which we use a dense layer with $\mathrm{sigmoid}$ activation $f(x) = 1/(1 + e^{-x})$ instead of radial basis function neural network. To be fair in the comparison, we use the same features extracted with the ResNet50 as in the proposed caption retrieval method. The features are then passed to a dense layer of dimension $D = 256$ with ReLu activation and then to the final dense layer with dimension $C = 16$, the number of classes/labels of the UAV dataset with $\mathrm{sigmoid}$ activation. The output of $\mathrm{sigmoid}$ function is a probability score for each label. During the inference stage, to determine the presence/absence of a label in the image, we fix a threshold value θ_{th} and check whether the output of each neuron exceeds the threshold value. The neuron output of each label exceeding θ_{th} are considered active determining the presence of the labels for a given image. The threshold value is empirically decided as $\theta_{\mathrm{th}} = 0.5$.

Once the label prediction is made, for each image we obtain a binary vector of dimension $C = 16$, where 1 is associated with the presence of a given label in the image and 0 with the absence of a given label. The retrieval is performed by computing the Hamming distance with respect to the query image. The images having the lowest Hamming distance to the query one are retrieved.

D. Evaluation Metrics

The effectiveness of the proposed image retrieval system is quantified using three different metrics: BLEU score [47], *F*-score [48], and user evaluation. In order to define the different metrics, let X_q be the query image along with its j different descriptions $S_{q,j} = \{w_{1,j}, w_{2,j}, \ldots, w_{p,j}\}$ and with the set $C_q \in C$ of labels present in X_q. Similarly, let $Y_r \in Y$ be the retrieved image along with its j different descriptions $S_{r,j} = \{w_{1,j}, w_{2,j}, \ldots, w_{p,j}\}$ and with the set $C_r \in C$ of labels present in X_r.

BLEU metric is a machine translation (MT) evaluation metric that measures how close the output of the MT system (candidate translation) is to the translation of a human expert (reference translation). The evaluation is based on the *precision* measure. Precision is calculated as the number of consecutive words (*n*-grams) in the candidate translation that occur in the reference translation divided by the total number of words of the candidate translation. BLEU score between a reference translation *R* and a candidate translation *G* is computed as a product of precision $P(N, G, R)$ and the brevity penalty $BP(G, R)$ as follows:

$$BLEU\,(N, G, R) = P\,(N, G, R) \times BP\,(G, R) \tag{12}$$

where $P(N, G, R)$ is the geometric mean of *n*-gram precision defined as follows:

$$P\,(N, G, R) = \left(\prod_{n=1}^{N} p_n\right)^{1/N} \tag{13}$$

and $p_n = m_n/l_n$, where m_n is the number of *n*-grams between *G* and *R*, l_n is the total number of *n*-grams in *G*. The brevity penalty penalizes the shorter translations and is calculated as follows:>

$$BP\,(G, R) = \min\left(1.0, \exp\left(1 - \left(\frac{\mathrm{len}\,(R)}{\mathrm{len}\,(G)}\right)\right)\right) \tag{14}$$

where $\mathrm{len}(R)$ is the length of the reference translation and $\mathrm{len}(G)$ is the length of the candidate translation. Due to the geometric mean of *n*-gram precision when there is no higher order *n*-gram precision (e.g. $n = 4$), BLEU score of the whole sentence is 0 independently of the low-order *n*-gram precisions ($n = 1, 2, 3$). To overcome this issue, we use a smoothing technique proposed in [49], which replaces the 0 score in presence of low-order *n*-grams with a small value ε. BLEU scores range from 0 to 1, where 1 is good. In this work, for the *n*-gram precision we used $n = 1, 2, 3, 4$. In our image retrieval system the reference translations are the ground truth descriptions $S_{q,j}$ of the query image X_q and the candidate translations are the ground truth descriptions $S_{r,j}$ of the retrieved image X_r. Before calculating the BLEU score, we apply WMD distance between each description $S_{q,j}$ of X_q and all the descriptions $S_{r,j}$ of retrieved image X_r to determine the closest description to $S_{q,j}$ and then calculate the BLEU score between the closest descriptions. The BLEU score for a query image X_q is determined averaging the BLEU score between each description of the query image and the closest description of the retrieved image. Finally, the BLEU score is averaged over all the retrieved images.

Since we are comparing the proposed image retrieval system with multilabel image retrieval system, we also evaluate the performances of the proposed image retrieval system using *F*-score, which is an adequate metric in case of multilabel information [48]. *F*-score is defined as the weighted harmonic mean of precision (Pr) and recall (Rec), where precision is defined as the fraction of identical labels of X_q and X_r in the label set C_r and recall is defined as the fraction of identical labels of X_q and X_r in the label set C_q. Depending on the parameter β

the F-score gives more importance to precision or recall. In this work, we have used $\beta = 1{,}2$. The equations of precision, recall and F-score are given in (15)–(17), respectively, where N_r is the number of retrieved images.

A third metric used to measure the effectiveness of the proposed retrieval system is the end-user evaluation. Each of the end-users is asked to evaluate the performances of our retrieval system and multilabel retrieval system based on a

$$\text{Precision} = \frac{1}{N_r}\sum_{r=1}^{N_r}\left|\frac{C_q \cap C_r}{C_r}\right| \tag{15}$$

$$\text{Recall} = \frac{1}{N_r}\sum_{r=1}^{N_r}\frac{|C_q \cap C_r|}{|C_q|} \tag{16}$$

$$F_\beta = \frac{\left(\beta^2+1\right)\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{17}$$

simple question: "If you were to choose between the two retrieval systems, which one satisfies you the most in terms of retrieved images?" The term "satisfied" is interpreted as the similarity between the query image and retrieved image in terms of the relationship of different entities, the position, orientation present in the query image and in the retrieved images. Users also considered the ranking produced by each retrieval system. The users are required to choose one of the retrieval systems. This evaluation is only done on our UAV dataset. In total we randomly take 100 query image out of 882 query images and for each query image we retrieve 20 images with both our retrieval system and multilabel content-based image retrieval method. In total, 16 users performed the evaluation.

V. Experimental Results

A. Experimental Results on UAV Dataset

1) BLEU Evaluation (B-1,2,3,4): In this section, we evaluate the proposed retrieval system in terms of mean BLEU score. In absence of works that use generated textual descriptions to query and retrieve images, we also report the upper bound results regarding the dataset. The upper bound results are obtained using the ground truth descriptions for query and retrieve the desired most similar images to a query image. As datasets has more than one ground truth descriptions, we randomly pick one of them to use as a query. We repeat this process 10 times and average the results. Tables I and II show the upper bound results and the proposed retrieval system results, respectively. In terms of word embedding technique, the results of each table are rather similar. We can notice an average gap of 10% in terms of mean BLEU score between the two tables. We believe that the reason of having this gap is due to the fact that the proposed retrieval system is affected by several errors, one of which is the captioning block as shown in Fig. 1. Indeed, observing Fig. 5, we can notice some errors in the generated sentences. Thus, one way to reduce the gap is to improve the image captioning block.

2) Comparison With Multilabel Image Retrieval Method [46]: Table III shows the results in terms of precision, recall, F-1 and F-2 scores when multilabel image retrieval and the proposed retrieval system are used. By analyzing Table III one can observe that in terms of recall, F-1 and F-2 score our proposed retrieval system achieves slightly better values respect to the multilabel image retrieval system. However, in terms of precision, the multilabel image retrieval shows slightly better results.

3) End User Evaluation: From the comparison between the proposed retrieval system and the multilabel one, we observed that the results are quite similar as it can be seen from Table III. In order to have a better understanding of the behavior of the proposed retrieval system, we also made a comparison from an end-user prospective between the proposed retrieval system and the multilabel one. Table IV reports the results of end-user evaluation. The results show that the proposed retrieval system overcomes the multilabel retrieval system by 4% from the end-users point of view. The users were also required to give some general comments about the two retrieval systems. In summary, the users confirmed that both algorithms retrieve similar images to a query image, however the proposed retrieval system shows better visual results in terms of orientation, number, and position of the objects with respect to the multilabel image retrieval method [46].

Fig. 6 shows an example of images retrieved by the multilabel retrieval system and the proposed retrieval system. The predicted primitive classes and the generated descriptions of the query

TABLE I
Upper Bound Results in Terms of Mean BLEU Score (B)

Embedding	*Nr of retrieved images*	*B1*	*B2*	*B3*	*B4*
GloVe	1	**0.738**	**0.651**	**0.595**	**0.524**
	5	**0.721**	**0.633**	**0.578**	**0.509**
	10	**0.707**	**0.618**	**0.563**	**0.494**
	15	**0.699**	**0.609**	**0.553**	**0.484**
	20	**0.690**	**0.599**	**0.543**	**0.474**
fasText	1	0.734	0.649	0.594	0.524
	5	0.717	0.631	0.577	0.508
	10	0.705	0.617	0.562	0.490
	15	0.697	0.607	**0.553**	**0.484**
	20	0.688	0.598	**0.543**	0.473

Note: Ground truth descriptions are used to query and retrieve the images.

TABLE II
Proposed Retrieval System Results in Terms of Mean BLEU Score

Embedding	*Nr of retrieved Images*	*B1*	*B2*	*B3*	*B4*
GloVe	1	0.626	0.529	0.472	0.408
	5	0.609	0.514	0.461	**0.397**
	10	0.608	0.515	0.463	0.402
	15	0.607	0.513	0.463	0.402
	20	0.604	0.511	0.461	**0.400**
fasText	1	**0.627**	**0.530**	**0.474**	**0.409**
	5	**0.611**	**0.517**	**0.464**	0.389
	10	**0.609**	**0.516**	**0.465**	**0.403**
	15	**0.609**	**0.516**	**0.465**	**0.404**
	20	**0.605**	**0.512**	**0.462**	**0.400**

Note: Generated descriptions are used to query and retrieve the images.

TABLE III
COMPARISON RESULTS BETWEEN THE PROPOSED RETRIEVAL SYSTEM AND MULTILABEL METHOD

Method	*Nr of retrieved images*	*Precision*	*Recall*	*F-1 Score*	*F-2 score*
Multilabel Retrieval System [46]	1	**0.823**	0.794	0.777	0.780
	5	**0.821**	0.797	**0.779**	0.780
	10	**0.799**	0.793	0.762	0.772
	15	**0.779**	0.793	0.749	0.765
	20	**0.789**	**0.790**	0.752	0.764
Proposed Retrieval System	1	0.778	**0.828**	**0.781**	**0.802**
	5	0.778	**0.809**	0.767	**0.783**
	10	0.778	**0.801**	**0.763**	**0.777**
	15	0.776	**0.794**	**0.759**	**0.772**
	20	0.773	0.788	**0.754**	**0.766**

TABLE IV
COMPARISON RESULTS BETWEEN THE PROPOSED RETRIEVAL SYSTEM AND MULTILABEL METHOD FROM END USERS ON 20 RETRIEVED IMAGES

Method	*User Evaluation (%)*
Multilabel image retrieval [46]	48
Proposed retrieval system	**52**

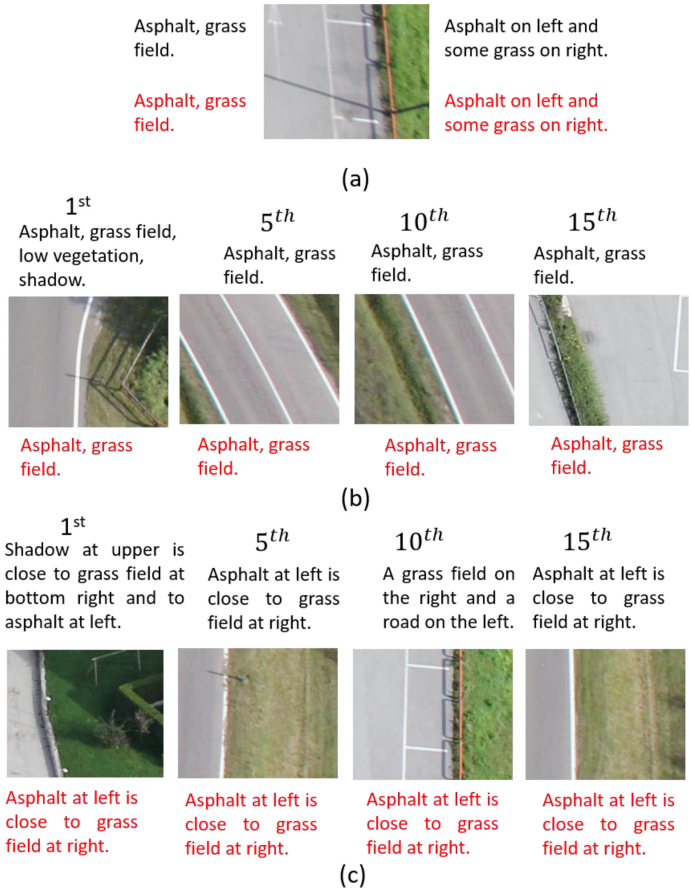

Fig. 6. Retrieval example. (a) Query image where the ground truth and generated labels are shown on the top and bottom left of the image, respectively; ground truth and generated sentences are shown on the top and bottom right, respectively. (b) Images retrieved using multilabel image retrieval system. Above are shown the ground truth labels and below highlighted in red are shown the predicted labels. (c) Images retrieved using the proposed retrieval system. Above is shown one of the ground truth description and below highlighted in red are shown the generated descriptions. The order of the retrieved images is reported above each image.

image shown in Fig. 6(a) are reported on the left and right of the image, respectively. The predicted primitive classes and the generated descriptions of the retrieved images are shown in Fig. 6(b) and (c), respectively. The retrieved images by the multilabel and proposed retrieval system are also shown in Fig. 6(b) and (c), respectively. Both the retrieval systems are able to find similar images with respect to the query image. However, the proposed retrieval system is more accurate in finding similar images. Even if the first retrieved images missed "shadow," all the retrieved images of the proposed retrieval system [see Fig. 6(c)] have the same spatial arrangement as the query image, the asphalt is on the left and the grass field is on the right. On the contrary, even if the results of multilabel retrieval system [see Fig. 6(b)] include the same primitive classes as the ones of the query image their spatial arrangement is not accurate, except for the first retrieved image. Fig. 7 shows another example of images retrieved by multilabel retrieval system and proposed retrieval system. We can notice that the results of both the retrieval systems show cars parked in a parking lot. However, the fifth and tenth retrieved images from multilabel retrieval system [see Fig. 7(b)] show only one car each, while the query image describes a parking lot with three cars. Moreover, the first image retrieved by the multilabel retrieval system shows an additive primitive class, namely "person." On the other hand, the images retrieved by the proposed retrieval system, even though the generated descriptions are affected by some errors [see Fig. 7(c)], show cars parked on the parking lot (from 3 to 4) and do not add any other primitive classes. Another example of images retrieved by the two retrieval systems is shown in Fig. 8. By looking at the images retrieved by both systems, we can see that the proposed retrieval system is able to accurately find very similar images [see Fig. 8(c)] to the query image. On the contrary, the multilabel retrieval system misses different primitive classes and adds others [see Fig. 8(b)]. By visual analysis of all the obtained results regarding the UAV dataset, we can conclude that even though the generated descriptions are affected by some errors, the proposed method detects and retrieves visually most similar images from the archive to a query image.

B. Experimental Results on the RSICD Dataset

Tables V and VI report the upper bound and the proposed retrieval system results, respectively. As it was mentioned in the previous section, for this dataset, the multilabels of the images are not available. We thus only provide the results of the proposed retrieval system and the upper bound in terms of mean BLEU scores.

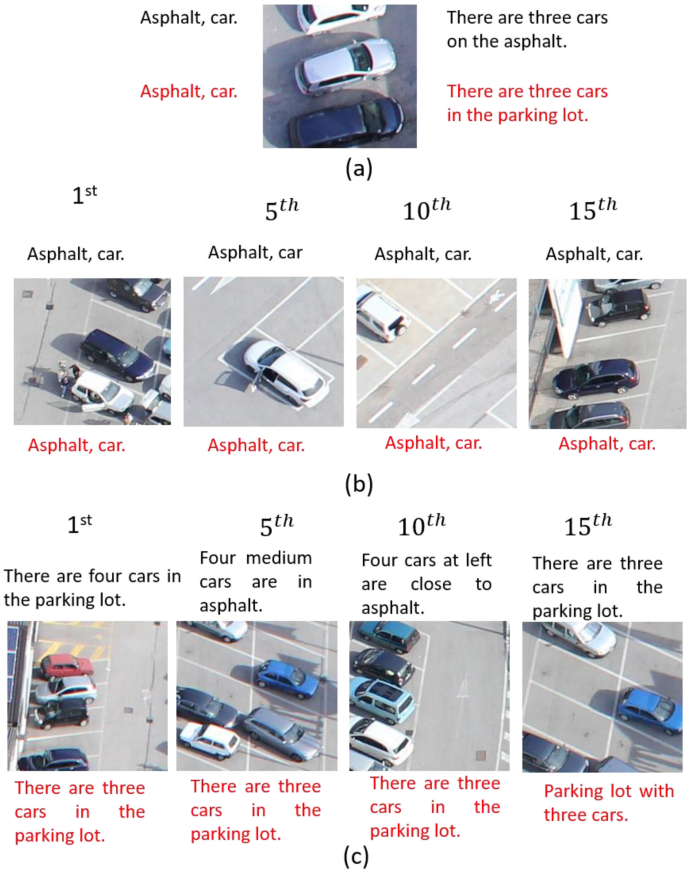

Fig. 7. Retrieval example. (a) Query image where the ground truth and generated labels are shown on the top and bottom left of the image, respectively; ground truth and generated sentences are shown on the top and bottom right, respectively. (b) Images retrieved using multilabel image retrieval system. Above are shown the ground truth labels and below highlighted in red are shown the predicted labels. (c) Images retrieved using the proposed retrieval system. Above is shown one of the ground truth description and below highlighted in red are shown the generated descriptions. The order of the retrieved images is reported above each image.

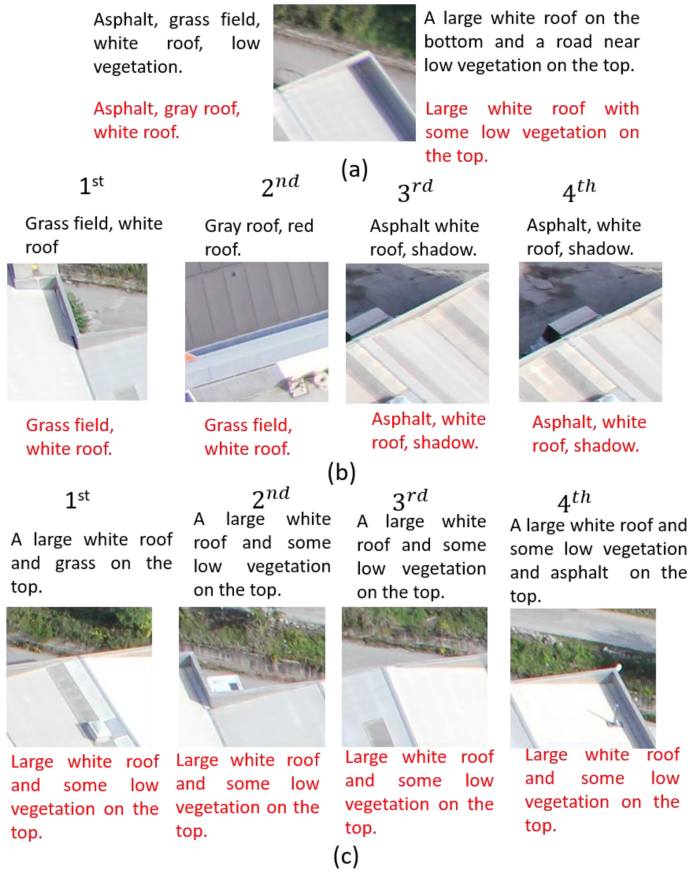

Fig. 8. Retrieval example. (a) Query image where the ground truth and generated labels are shown on the top and bottom left of the image, respectively; ground truth and generated sentences are shown on the top and bottom right, respectively. (b) Images retrieved using multilabel image retrieval system. Above are shown the ground truth labels and below highlighted in red are shown the predicted labels. (c) Images retrieved using the proposed retrieval system. Above is shown one of the ground truth description and below highlighted in red are shown the generated descriptions. The order of the retrieved images is reported above each image.

TABLE V
UPPER BOUND RESULTS IN TERMS OF MEAN BLEU SCORE

Embedding	*Nr of retrieved images*	*B1*	*B2*	*B3*	*B4*
GloVe	1	0.643	0.539	0.490	0.424
	5	0.596	0.486	0.435	0.366
	10	0.570	0.455	0.403	0.334
	15	0.554	0.438	0.385	0.315
	20	**0.**543	0.426	0.374	0.303
fasText	1	**0.646**	**0.543**	**0.494**	**0.430**
	5	**0.600**	**0.489**	**0.438**	**0.370**
	10	**0.574**	**0.459**	**0.407**	**0.338**
	15	**0.558**	**0.441**	**0.389**	**0.319**
	20	**0.546**	**0.428**	**0.376**	**0.306**

Note: Ground truth descriptions are used to query and retrieve the images.

TABLE VI
PROPOSED RETRIEVAL SYSTEM RESULTS IN TERMS OF MEAN BLEU SCORE

Embedding	*Nr of retrieved images*	*B1*	*B2*	*B3*	*B4*
GloVe	1	0.386	0.257	0.209	0.147
	5	0.385	0.259	0.214	**0.153**
	10	0.384	0.259	0.214	**0.155**
	15	0.381	0.256	0.212	**0.153**
	20	0.380	0.255	0.211	0.152
fasText	1	**0.388**	**0.259**	**0.210**	**0.148**
	5	**0.387**	**0.260**	**0.214**	**0.153**
	10	**0.386**	**0.260**	**0.215**	**0.155**
	15	**0.384**	**0.258**	**0.213**	**0.153**
	20	**0.382**	**0.256**	**0.218**	**0.153**

Note: Generated descriptions are used to query and retrieve the images.

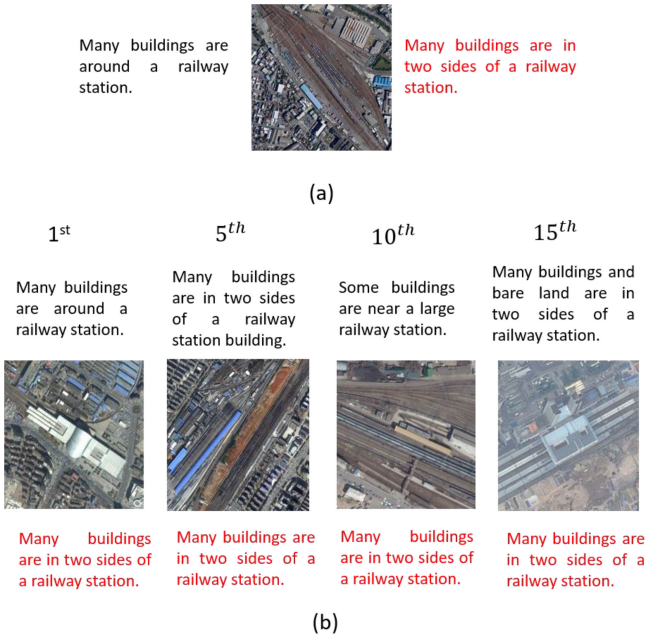

Fig. 9. Railway station image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

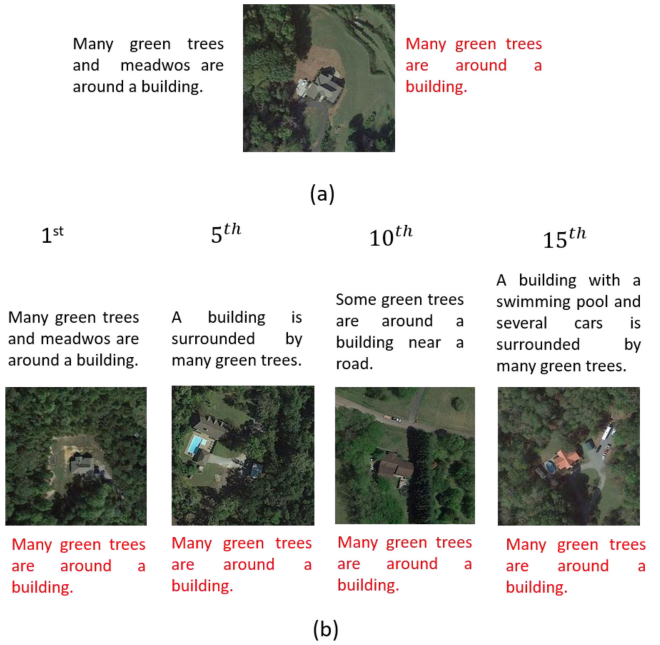

Fig. 10. Sparse residential image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

Unlike the results obtained in the UAV dataset, here we observe a reduction in terms of mean BLEU score for the two tables and in particular the results of the proposed retrieval system are lower. The gap in terms of mean BLEU score between two tables varies with the number of retrieved images, from 0.15 to 0.3.

Fig. 9 shows an example of images retrieved by the proposed retrieval system when the query image is selected from railway station category of the RSCID archive. The retrieval order of each image is given above the related image together with one of the ground truth description. The generated descriptions are highlighted by red and are given below the retrieved image. From the visual inspection of the retrieved images, we can observe that all the retrieved images are very similar to the query image. The 15th retrieved image [see Fig. 9(b)] contains some bare land which is not captured by the generated description. However, the bare land even if not included in all the ground truth descriptions is present in all the retrieved images.

Fig. 10 shows another example of retrieved images when the query image is selected from sparse residential category of the archive. We can observe that all the retrieved images are very similar to the query image. Even if "meadow" primitive class is missing, all the retrieved images show a single building surrounded by trees.

Fig. 11 shows another example of the retrieved image when the query image is selected from dense residential category of the archive. The ground truth descriptions of the query images are in total three, which are as follows.

1) "The roof of residential buildings is red."
2) "The wide have a lot of people walking on the road."
3) "Many buildings are in a dense residential area."

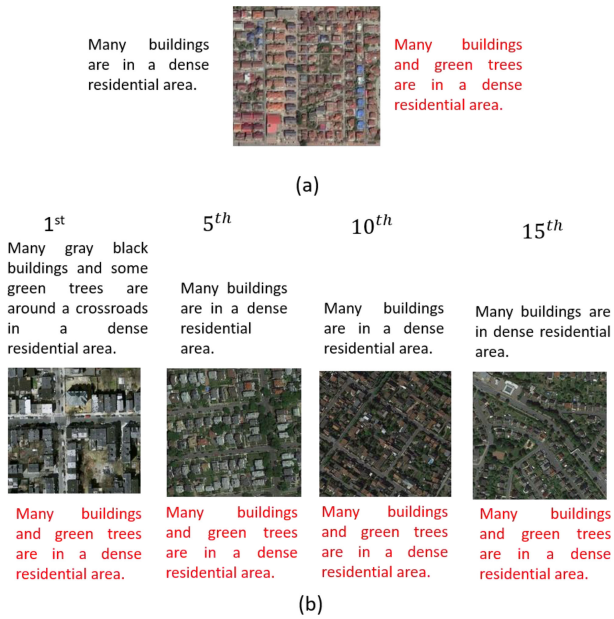

Fig. 11. Dense residential image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

From the ground truth descriptions, we can notice that the first two descriptions may not be very accurate, as they are missing some classes and adding some others not present in the query image. The third description instead is found almost in all the images within the dense residential category. As we use all the ground truth descriptions of the query image in order to calculate

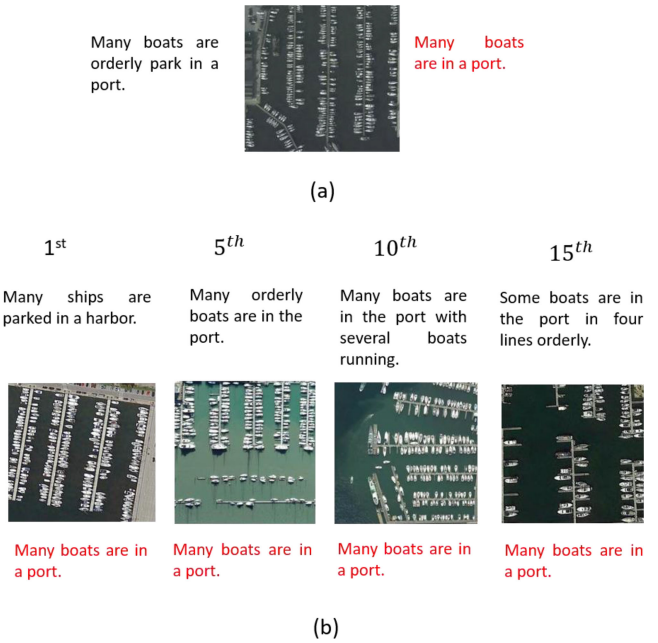

Fig. 12. Port image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

the BLEU score, when the ground truth description presents some ambiguity, the score will be low. Indeed, for the given example, we have a BLEU 1, BLEU 2, BLEU 3, and BLEU 4 score of 0.333, 0.230, 0.207, and 0.163, respectively. However, for the example shown in Fig. 9, we have BLEU 1, BLEU 2, BLEU 3, and BLEU 4 score of 0.637, 0.551, 0.514, and 0.456, respectively. This may be one of the reasons why the results reported in Table VI are much lower with respect to Table V.

Fig. 12 shows another example of the retrieved images where the query image is selected from port category of the archive. The ground truth descriptions of the query image are in total three, which are as follows.

1) "There are many places to a relatively large port."
2) "The lake is green above a lot of ship."
3) "Many boats are orderly in a port."

The ground truth descriptions for the first retrieved image [see Fig. 11(b)] are as follows.

1) "Many small black fish are in the pond."
2) "The pond is surrounded by light green lawns and vegetation."
3) "Many cars are parked on the street."
4) "Many ships are parked in the harbor."

Even in this example we can see that in both the query and the first retrieved image we find some ambiguity in the ground truth descriptions. For instance, the first ground truth description of the first retrieved image is completely wrong. Indeed, the BLEU scores 1, 2, 3, and 4 between the query image and the first retrieved image is 0.316, 0.064, 0.055, and 0.033, respectively. We also can notice that the generated descriptions, even if very short and simple, are in line with what it is shown in the query image and the retrieved images [see Fig. 12(a) and (b)]. Furthermore, one can see that all the retrieved images, from a visual inspection are highly correlated to the query image.

From different examples that we have seen from the RSICD we can conclude that the reason why the results of Table VI are low may be mainly related to the ambiguity of the ground truth descriptions. We would like to emphasize that this phenomenon occurs throughout all the RSICD dataset. Despite this, we can conclude that the caption generator block concentrates more on the most frequent ground truth examples during training to learn and to generate during test time highly correlated captions with the image visual contents. We also can conclude that no matter the low results we have obtained in terms of mean BLEU score per query image, the similarity between the query image and all the retrieved images shown in different examples is considerably high.

VI. Conclusion

In this article, we have presented a novel image retrieval system that represents the high-level semantic content of the images by generated sentences and perform image retrieval based on the generated sentences. The main idea and contribution of the article is the combination of remote sensing and natural language processing techniques to perform RS image retrieval. Representing the image content by generated sentences allows to express better the complex content of an RS image instead of using descriptors that only model the primitives. As a consequence, the retrieval system might be more accurate if proper sentences are generated and used to query and retrieve images from an archive. Hence, image captioning block is crucial. We have tested our system in two different RS archives. From the qualitative and quantitative results, using generated sentences as a query to perform image retrieval could be a promising direction for the community to improve the CBIR techniques. As a future work we plan to improve the captioning block.

References

[1] D. Peijun, C. Yunhao, T. Hong, and F. Tao, "Study on content-based remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sensing Symp.*, Jul. 2005, vol. 2, pp. 707–710, doi: 10.1109/IGARSS.2005.1525204.

[2] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.

[3] I. Tekeste and B. Demir, "Advanced local binary patterns for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 6855–6858, doi: 10.1109/IGARSS.2018.8518856.

[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] B. Chaudhuri, B. Demir, L. Bruzzone, and S. Chaudhuri, "Region-based retrieval of remote sensing images using an unsupervised graph-theoretic approach," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 987–991, Jul. 2016.

[6] K. Amiri and M. Farah, "Graph of concepts for semantic annotation of remotely sensed images based on direct neighbors in RAG," *Can. J. Remote Sens.*, vol. 44, no. 6, pp. 551–574, Nov. 2018.

[7] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[8] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2323–2334, May 2015.

[9] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2018.

[10] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Comput. Vis. Image Understanding*, vol. 184, pp. 22–30, Jul. 2019.

[11] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Deep metric and hash-code learning for content-based retrieval of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 4539–4542, doi: 10.1109/IGARSS.2018.8518381.

[12] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, May 2017, Art. no. 489.

[13] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018.

[14] F. Ye, M. Dong, W. Luo, X. Chen, and W. Min, "A new re-ranking method based on convolutional neural network and two image-to-class distances for remote sensing image retrieval," *IEEE Access*, vol. 7, pp. 141498–141507, 2019, doi: 10.1109/ACCESS.2019.2944253.

[15] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 318–328, Jan. 2020, doi: 10.1109/JSTARS.2019.2961634.

[16] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst.*, Jul. 2016, pp. 1–5, doi: 10.1109/CITS.2016.7546397.

[17] 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 *Proc. IEEE Int. Geoscience Remote Sens. Symp.*, Jul. 2017, pp. 4798–4801, doi: 10.1109/IGARSS.2017.8128075.

[18] 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.

[19] 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.

[20] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[21] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, Jan. 2019, Art. no. 612, doi: 10.3390/rs11060612.

[22] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 10039–10042, doi: 10.1109/IGARSS.2019.8900503.

[23] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2020.

[24] G. Hoxha, F. Melgani, and B. Demir, "Retrieving images with generated textual descriptions," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5812–5815, doi: 10.1109/IGARSS.2019.8899321.

[25] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, "TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sens.*, vol. 12, no. 3, Jan. 2020, Art. no. 3.

[26] R. Bernardi *et al.*, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, 2016.

[27] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.

[28] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, 2014, vol. 32, pp. II-595–II-603.

[29] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," 2014, *arXiv:1410.1090*.

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.

[31] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[32] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2422–2431, doi: 10.1109/CVPR.2015.7298856.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[34] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 5528–5531.

[35] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2010, pp. 1045–1048, [Online]. Available: http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html

[36] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519, doi: 10.1109/CVPRW.2014.131.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:151203385 Cs*, [Online]. Available: http://arxiv.org/abs/1512.03385. Accessed: Jan. 05, 2020

[38] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, Jun. 2000, Art. no. 6789, doi: 10.1038/35016072.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Assoc., Inc., 2012, pp. 1097–1105.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[41] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proc. Conf. Empirical Methods Natural Language Process.*, Oct. 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162. Accessed: Jan. 7, 2019

[42] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.

[43] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.

[44] 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.

[45] "GloVe: Global vectors for word representation." Oct. 2015. [Online]. Available: https://nlp.stanford.edu/projects/glove/. Accessed: Jan. 7, 2019.

[46] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318, doi: 10.3115/1073083.1073135.

[48] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. Pac.-Asia Conf. Knowledge Discovery Data Mining*, 2004, pp. 22–30.

[49] B. Chen and C. Cherry, "A systematic comparison of smoothing techniques for sentence-level BLEU," in *Proc. 9th Workshop Statist. Mach. Translation*, Jun. 2014, pp. 362–367, [Online]. Available: http://www.aclweb.org/anthology/W14-3346. Accessed: Jan. 7, 2019

Genc Hoxha (Student Member, IEEE) received the B.S degree in electronics and telecommunications engineering and the M.Sc. degree in telecommunications engineering from the University of Trento, Trento, Italy, in 2014 and in 2018, respectively. He is currently working toward the Ph.D. degree in signal processing and pattern recognition in the ICT Doctoral School, University of Trento.

His research interests include machine learning and image processing with applications to remote-sensing image analysis.

Farid Melgani (Fellow, IEEE) received the State Engineer degree in electronics from the University of Batna, Batna, Algeria, in 1994, the M.Sc. degree in electrical engineering from the University of Baghdad, Baghdad, Iraq, in 1999, and the Ph.D. degree in electronic and computer engineering from the University of Genoa, Genoa, Italy, in 2003.

He is a Full Professor of telecommunications with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he teaches pattern recognition, machine learning, and digital transmission. He is the Head of the Signal Processing and Recognition Laboratory, and the Coordinator of the Doctoral School in Industrial Innovation, University of Trento. His research interests include the areas of remote sensing, signal/image processing, pattern recognition, machine learning, and computer vision. He has coauthored more than 220 scientific publications.

Dr. Melgani is currently an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *International Journal of Remote Sensing*, and IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS.

Begüm Demir (Senior Member, IEEE) received the B.S., M.Sc., and Ph.D. degrees in electronic and telecommunication engineering from Kocaeli University, Kocaeli, Turkey, in 2005, 2007, and 2010, respectively.

Since 2018, she has been a Full Professor and the Head of the Remote Sensing Image Analysis (RSiM) group with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany. Before joining TU Berlin, from 2013 to 2017, she was an Assistant Professor with the Department of Computer Science and Information Engineering, University of Trento, Italy, while in 2017, she became an Associate Professor with the Department of Computer Science and Information Engineering. Her research activities lie at the intersection of machine learning, remote sensing, and signal processing. Specifically, she performs research on developing innovative methods for addressing a wide range of scientific problems in the area of remote sensing for Earth observation.

Dr. Demir is a scientific committee member of several international conferences and workshops, such as: Conference on Content-Based Multimedia Indexing, Conference on Big Data from Space, Living Planet Symposium, International Joint Urban Remote Sensing Event, SPIE International Conference on Signal and Image Processing for Remote Sensing, and Machine Learning for Earth Observation Workshop organized within the ECML/PKDD. She is a Referee for several journals such as the PROCEEDINGS OF THE IEEE, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition*, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, the *International Journal of Remote Sensing*), and several international conferences. She is currently an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and *MDPI Remote Sensing*. She was a recipient of a Starting Grant from the European Research Council (ERC) with the project "BigEarth-Accurate and Scalable Processing of Big Data in Earth Observation" in 2017, and the "2018 Early Career Award" presented by the IEEE Geoscience and Remote Sensing Society.