UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

ICT International Doctoral School

IMAGE CAPTIONING FOR REMOTE SENSING IMAGE ANALYSIS

Genc Hoxha

Advisor

Prof. Farid Melgani

Università degli Studi di Trento

August 2022

Acknowledgment

First and foremost, I would like to express my deepest thanks of gratitude to my supervisor Professor Farid Melgani. I would like to thank him for the dedication, professionalism and mentoring during this journey. In particular, I would like to thank him for being always there in critical times regarding the research and personal life. You have been very helpful and you are truly an inspiration to me.

I want to thank Professors Yakoub Bazi, Mauro Dalla Mura and Nicola Conci for serving in my examination committee, and for the precious feedbacks and comments they already provided about my work. I am truly honoured to have them in my committee.

I would like to thank Mesay for his friendship and support during the time we passed together in our SPR laboratory along with the visiting PhD students, in particular Seloua, Youyou and Sofia for their intercultural and research exchange and the newly joining members Riccardo and Praveen for the last year passed together.

I would like to thank Professor Begüm Demir and the members of her research group at TU Berlin. The period abroad that I spent with her and the members of RSiM group has been an important time for my professional and personal growth. I would like to thank her for hosting me in her group during critical times. I would like to thank RSiM group members Mahdyar, Gencer, Tom, Tai, Adina, Kerem, Georgii and Steve for the multiple conversations we had about research, politics and life.

I would like to thank the ICT doctoral school at the University of Trento and all the people, colleagues and friends I met here during these years.

A special thanks goes to my friends.

A special thanks goes to Enrico and Nicoletta.

Finally, I would like express my deepest gratitude to my parents Vjollca and Veri, my brother Gerti, my uncle Jorgji, my aunt Lindita and my fiancée Giulia for always supporting and believing in me.

To Giulia.

Abstract

Image Captioning (IC) aims to generate a coherent and comprehensive textual description that summarizes the complex content of an image. It is a combination of computer vision and natural language processing techniques to encode the visual features of an image and translate them into a sentence. In the context of remote sensing (RS) analysis, IC has been emerging as a new research area of high interest since it not only recognizes the objects within an image but also describes their attributes and relationships. In this thesis, we propose several IC methods for RS image analysis. We focus on the design of different approaches that take into consideration the peculiarity of RS images (e.g. spectral, temporal and spatial properties) and study the benefits of IC in challenging RS applications.

In particular, we focus our attention on developing a new decoder which is based on support vector machines. Compared to the traditional decoders that are based on deep learning, the proposed decoder is particularly interesting for those situations in which only a few training samples are available to alleviate the problem of overfitting. The peculiarity of the proposed decoder is its simplicity and efficiency. It is composed of only one hyperparameter, does not require expensive power units and is very fast in terms of training and testing time making it suitable for real life applications.

Despite the efforts made in developing reliable and accurate IC systems, the task is far for being solved. The generated descriptions are affected by several errors related to the attributes and the objects present in an RS scene. Once an error occurs, it is propagated through the recurrent layers of the decoders leading to inaccurate descriptions. To cope with this issue, we propose two post-processing techniques with the aim of improving the generated sentences by detecting and correcting the potential errors. They are based on Hidden Markov Model and Viterbi algorithm. The former aims to generate a set of possible states while the latter aims at finding the optimal sequence of states. The proposed post-processing techniques can be injected to any IC system at test time to improve the quality of the generated sentences.

While all the captioning systems developed in the RS community are devoted to single and RGB images, we propose two captioning systems that can be applied to multitemporal and multispectral RS images. The proposed captioning systems are able at describing the changes occurred in a given geographical through time. We refer to this new paradigm of analysing multitemporal and multispectral images as change captioning (CC). To test the proposed CC systems, we construct two novel datasets composed of bi-temporal RS images. The first one is composed of very high-resolution RGB images while the second one of medium resolution multispectral satellite images. To advance the task of CC, the constructed datasets are publically available in the following link: https://disi.unitn.it/~melgani/datasets.html.

Finally, we analyse the potential of IC for content based image retrieval (CBIR) and show its applicability and advantages compared to the traditional techniques. Specifically, we focus our attention on developing a CBIR systems that represents an image with generated descriptions and uses sentence similarity to search and retrieve relevant RS images. Compare to traditional CBIR systems, the proposed system is able to search and retrieve images using either an image or a sentence as a query making it more comfortable for the end-users.

The achieved results show the promising potentialities of our proposed methods compared to the baselines and state-of-the art methods.

Keywords

Change Captioning, Content-based image retrieval, Hidden Markov Models, Image Captioning, Post-processing, Support Vector Machines, Viterbi Algorithm.

Contents

List of Tables

List of Figures

Chapter 1

1. Introduction

Describing the environment that surrounds us is a relatively easy task for us, humans. Given an image, it is natural for a human to describe with immeasurable details the salient contents of an image at a glance [1]. However, computers struggle to describe even the simplest scenarios. Making computers understand their surroundings with the same natural ability as humans do, has been the main focus of researchers in the field of artificial intelligence (AI).

Remote sensing (RS) is defined as the acquisition of information about an object without being in physical contact with it [2]. This is done by detecting and measuring changes that an object imposes on the surrounding field (electromagnetic, acoustic or potential). Most commonly, the term 'remote sensing' is used in connection with electromagnetic techniques of information acquisition that cover the whole electromagnetic spectrum [2]. Though RS data can be of various types, in this work we will focus only on optical RS imagery.

The automatic interpretation of RS images has been mainly concentrated on techniques, such as image classification/segmentation and object recognition. Such techniques aim to represent images with a set of spatialized semantic land-cover classes (labels). Due to their intrinsic nature, these techniques do not capture the attributes and the relationships that exist between the different land-cover classes. It is worth mentioning that the attributes and the relationships are part of the high-level semantic information of an RS scene. In particular, with the fast development of RS technology, we are now able to acquire images characterized by high spatial resolution with an abundant quantity of high-level semantic information. Accordingly, conventional techniques such the image classification/segmentation or object recognition might not be appropriate for analysing such images.

Recently, to have a better understanding of an RS scene, image captioning (IC) has attracted the attention of the community. IC is a difficult but fundamental task in AI that aims to generate a textual description (i.e., sentence or caption) of the content of an image. It requires both the knowledge of computer vision (CV) and natural language processing (NLP) fields to better understand the image content and express this knowledge through a sentence description. Unlike previous techniques, IC not only provides the labels of different land-cover classes and their locations but is also able to capture and describe the attributes and the relationships with a sentence following linguistics rules. This richer representation can be useful in a variety of remote sensing applications such as image retrieval [3]–[5], change detection [6], image generation [7], [8] and so on. The importance of IC is reflected in the incremented number of articles recently published in the RS community [9]. However, the initial previous works were mostly devoted to adopting the IC techniques from natural images to RS images and constructing the relative datasets [9].

IC techniques in the RS community are inspired by the seminal works of the CV community [10]–[19] and can be divided into three main categories: 1) template-based, 2) retrieval-based and 3) encoder-decoder frameworks. Template-based IC systems are composed of fixed sentence templates. First, object detection algorithms are exploited to detect objects and actions and then the fixed templates are filled with the detected objects. The only example of a template-based IC method in the RS community is the work of Shi et al. [20] where a fully convolutional network is used to detect the objects of an image and a language model based on fixed templates is used to generate the descriptions. The sentences generated by template-based IC are in general correct from a grammatical and content viewpoint. However, they are heavily hand-designed and, because of the fixed templates, the generated descriptions tend to be less natural compared to human descriptions.

The second type is retrieval-based IC. In this methodology, the generation of a sentence is treated as a retrieval problem. First given a target image, the retrieval-based IC methods search and retrieve from an archive the most similar images (to the target image) along with their descriptions. Then to the target image is assigned one or more descriptions of its most similar images. As an example, Wang et al. [21] mapped images and descriptions in the same semantic space to learn a distance metric to quantify the similarity between images and descriptions. At inference time, to a target image are assigned five descriptions that have the smallest distance from the considered image. Retrieval-based IC systems cannot generate novel descriptions and they assume that there is always a related image-text pair in the archive for a target image. This assumption might not always be true leading to descriptions that are uncorrelated to the image content.

The most widely used IC in the RS community is the encoder-decoder framework [22]–[29] which is inspired by the progress in deep learning (DL) and machine translation (ML) [30]–[33]. Most of the encoder-decoder IC methods exploit pre-trained convolutional neural networks (CNNs) to encode the visual feature of an image and sequential models such as recurrent neural networks (RNNs) or long-short term memory (LSTM) [34] to translate the encoded visual features in a sentence description [9]. Encoder-decoder frameworks can be divided as: 1) simple encoder-decoder and 2) attention-based encoder-decoder. The former represents an image with a fixed-length vector and the same feature vector is used as input to the decoder to generate the description one word at a time. The latter focus its attention on different parts of an image and extract different feature vectors for each part. In this case, the decoder uses feature vectors extracted by those portions of an image that are more related to the generated words.

The first work in the RS community that uses a simple encoder-decoder framework is exploited in [22] where different CNNs [35]–[37] are used to encode the visual features of images and RNN or LSTM [34] is used to generate the captions. Zhang et al. [23] detected the main objects from an RS scene and forwarded the detected objects into an RNN model to generate the descriptions. Hand-crafted features [38]–[40] in addition to deep features and attention mechanism [18], [32] are explored in [24] to generate sentences. An attribute attention mechanism is introduced in [25]. A multiscale cropping and training mechanism is introduced in [26] for data augmentation to alleviate the problem of overfitting. To deal with the large scale variation present in RS images, a multiscale feature fusion combined with a de-noising mechanism is introduced in [27] and two different multiscale methods based on feature pyramid networks [41] are presented in [42]. Lu et al. [28] introduced an active attention mechanism where the sound of the name of different objects uttered by humans is used as guiding information to generate descriptions, and a retrieval topic recurrent memory is introduced in [29] where sentence topics are used to guide the description generation process. Sumbul et al. [43] introduced a summarization driven RS IC system where a pre-trained pointer generator [44] is used to summarize the ground-truth captions to keep only the relevant information which is then integrated into the IC system through an attention mechanism to generate coherent descriptions. A combination of a simple encoder-decoder and a retrieval-based IC framework is explored in [45] to alleviate the misrecognition problem. Li et al. [46] introduced a truncation cross-entropy loss to alleviate the overfitting whereas Zhao et al. [47] proposed a structured attention mechanism that can exploit structured spatial relationships widely present in an RS scene in contrast to previous unstructured attention methods. More recently Transformers [33] have been exploited in the RS community to boost the performances of IC systems [9]. Compared to RNNs, Transformers have several advantages such as removing the recurrence mechanism of RNNs allowing parallel computations and a more sophisticated attention mechanism [33], [48]. Shen et al. [49], [50] exploited Transformers to perform RS IC. In these works, Transformers are used as a decoder and are combined with self-critical sequence training [51] and variational autoencoder [52] to cope with insufficient training data.

In general, encoder-decoder architectures generate novel sentences that are very similar (from syntax and lexical viewpoint) to the sentences produced by humans. However, because they are based on deep learning architectures their performances strongly depend on the number of annotated training samples (the larger

the training set, the lower the risk of overfitting). Indeed, in the CV community, the datasets that are used to train and test IC systems have a very large amount of annotated samples. An example of such a dataset is the MS COCO dataset [53] which is composed of more than 300.000 images where each image is annotated with 5 different descriptions. In contrast to the CV community, in the RS community, the datasets used to perform IC are very small. This is due to the fact that creating large datasets is an expensive process in terms of time and resources. This is reflected on the size of the largest dataset present in the RS community. This dataset is composed of only 10.921 images. Each image is annotated with 5 descriptions but only 7% of them have 5 different ones. The majority of the images (48%) have only one unique description and the rest of them has from two to four different ones. This makes it hard to build IC models based on deep learning. Strategies like self-critical sequence training or variational autoencoders are usually used to cope with the data limitation problem [50], [51]. Another issue of the deep learning based models is also the high number of hyperparameters to be carefully tuned to have good performances. Furthermore, deep learning methodologies demand expensive computational power units, such are the graphics processing units to have reasonable training and testing time. It is worth noting that the more complex the system, the more acute the aforementioned issues.

1.1. Objective of The Thesis

The aim of this thesis is twofold:

1. Development of new image captioning techniques in the context of RS image analysis.
2. Study the potential of IC in challenging RS applications.

The first and major objective is the one of developing novel IC techniques that take into consideration the peculiarities of the RS images, such are the spectral, spatial and temporal properties. In this context we propose three main solutions:

- A novel decoder for RS IC based on support vector machines (SVMs). The peculiarity of the proposed decoder is its simplicity and efficiency. The proposed SVM based decoder is injected in the simple encoder-decoder architecture instead of RNNs or Transformers. Compared to the deep learning based solutions, the proposed decoder is particularly interesting for situations in which only a few training samples are available to alleviate the overfitting problem. It is characterized by a very short training and inference time, has only one hyperparameter and does not require expensive computational power units such are the GPUs, making it suitable in real-time applications.
- Regardless of the efforts made in designing reliable IC systems, the task is far from being solved. In general, the generated descriptions are affected by several errors related to the objects and their attributes. Once an error occurs, this is propagated through the recurrent layers of the decoder leading to non-accurate descriptions. To cope with this problem, we propose two post-processing strategies. The proposed post-processing strategies aim to improve the quality of an IC system by detecting and correcting potential errors in the generated sentences. They are based on Hidden Markov Models (HMMs) and the Viterbi algorithm. The former aims to generate a set of possible states while the latter aims at finding the optimal sequence of states efficiently. The proposed post-processing strategies are applied at test time and can be injected into any IC system to improve the quality of a generated sentence.
- While all of the IC systems developed in the RS community are dedicated to single and RGB images, in this thesis we propose two captioning systems that are applied to multitemporal and multispectral image. The proposed captioning systems have the goal of describing the changes occurred in the multi-temporal and multi-spectral images. We refer to this new paradigm as change captioning (CC). Compared to the traditional change detection (CD) systems, that produce a binary change map highlighting the changes in a given geographical area or a semantic change map, a CC

system describes the changes with sentence descriptions including high-level semantic information such are the relationships and the attributes of the changed areas. To test the proposed CC systems, we construct two novel bi-temporal datasets. The first one consists of very high resolution (VHR) RGB images while the second one consists of medium resolution multispectral satellite images. To advance the task of CC, the constructed dataset are publically available in the following link: https://disi.unitn.it/~melgani/datasets.html.

The second objective of this thesis is to study the benefits of the IC on challenging RS applications. In particular, we study the benefits of the IC for content-based remote sensing image retrieval (CBIR). CBIR systems are very important in the RS community as they allow one to query and retrieve relevant information from massive archives of RS images. A CBIR system aims to search and retrieve the most relevant images to a query image from a massive archive and consists of two main steps: 1) image representation where the goal is to describe an RS image with discriminative features that model the primitives (such are the different land-cover classes) and 2) retrieval of the most similar images to the query image by evaluating the similarity between the extracted features. The most crucial step of a CBIR system is image representation. While most of the CBIR systems use the hand-crafted or deep feature for image representation, we propose to use sentence descriptions to represent the images. We argue that the visual image descriptors might have limitation in modelling primitives such are attributes and the relationships of different objects present in an RS scene. These limitations strongly affect the performances of a CBIR system. To cope with this problem, we propose to use textual descriptors (sentences) that are a default container of such high-level semantic information. Then, we perform retrieval using the generated textual descriptors. This not only leads to a better but also to more user-friendly CBIR system. Our system allows one to use images or text as query depending on the needs of the users. In the former, an IC system is used to generate the textual descriptions while in the latter users are allowed to formulate the query as they wish using text. Overall, our contribution in this field is that, while most of the CBIR methods use hand-crafted or deep feature to represent the image content we propose to represent the images by generated sentence descriptions and use sentence similarity to search and retrieve the similar images. Compared to the existing CBIR methods, our proposed method is able to search and retrieve images using an image as a query or using a sentence as a query. The former is obtained by generating a sentence description utilizing an IC system, while the latter allow users to simply type a text and search for similar images that best represent their needs. The proposed CBIR system is more user-friendly compared to the traditional ones.

1.2. Structure of the Thesis

In Chapter 2, we start first with the second objective described above to explore and understand the conventional IC systems and their importance in a CBIR system. This will also help familiarizing the reader with standard IC techniques before moving to the new proposed methodological developments described in the successive chapters. We briefly review the recent successful approaches for CBIR. We then describe in detail the proposed CBIR system that performs query and retrieval using generated sentence descriptions. The proposed CBIR system is composed of three main blocks: 1) an IC system that aims to generate a sentence description of the content of a RS image; 2) a sentence encoder that converts the generated descriptions into semantically meaningful feature vectors. This is achieved using recent word embedding techniques; 3) metric similarity steps the estimates the similarity between the vectors of the generated descriptions and the ones of the archive, and then retrieves the most similar images to the query image.

Chapter 3 presents the proposed IC system based on the SVM decoder. Compared to the traditional IC systems, we propose the SVM decoder to generate the sentence descriptions. The work is part of simple encoder-decoder systems that uses global feature to represent the RS scene. The extracted features are then forwarded to the decoder to generate the coherent captions. Compared to deep learning solutions, the

proposed decoder alleviates the overfitting problem, is faster and does not require expensive computation power units such are the GPUs.

In Chapter 4 we describe the two proposed post-processing strategies that can improve the outcome of an IC system. These strategies are based on HMMs and Viterbi algorithm to propose and find an optimal sequence of states in an efficient way, respectively. The proposed post-processing strategies are applied at test time and can be injected to any IC system to improve the generated sentence quality.

In Chapter 5 we introduce the proposed change captioning systems. These systems are able to describe the changes occurred in a given geographical area with sentence descriptions. They can be applied to multitemporal and multispectral images and constitute a new paradigm in analysing multitemporal RS images.

Finally, Chapter 6 concludes the thesis and proposes some future directions.

Chapter 2

2. Application of IC to Content-Based Image Retrieval

In this chapter, we present the benefits of IC captioning to content-based image retrieval (CBIR). We propose a solution that exploits IC to perform CBIR and demonstrate its effectiveness. The main idea of the work consists in representing an RS image with a textual description to include high-level semantic information and performing a text-based retrieval. The main advantages of the proposed system are the inclusion of high-level semantic information through sentence description generation and a more user-friendly query solution based on either text or image depending on the users' needs. To this end, the proposed retrieval system consists of three main steps. The first step aims to encode the image's 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 an encoder-decoder IC system. 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 retrieves the most similar images to the query image.

2.1. Motivation

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, great attention is devoted to CBIR which 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 [54].

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 chapter 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). Textual descriptions not only are a default container of high-level semantic information but also are more user-friendly [55]. Hence in this work, we allow users to query images from a massive archive by either using image or text as a query depending on the users' needs.

2.2. Introduction and Literature Review

A CBIR system consists of 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 [54]. 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 [56]. In [57], a histogram of local binary patterns that models the relationship of each pixel

in a given image with its neighbours (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 [3], [58], [59]. Descriptors of bag of spectral values are introduced in [60] to model the spectral information content of high dimensional RS images. After defining the image's 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 neighbour 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 [58] can be used. Supervised methods require the 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 [61]. If the training images are annotated by low-level land cover class labels (i.e., multi-labels), multi-label image retrieval methods are required. In [62], a sparse reconstruction-based method that generalizes the standard sparse classifier to the case of multi-label 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 compared to traditional systems. Deep learning (DL)-based RSIR systems simultaneously optimize feature learning and image retrieval [62]–[68]. Deep feature representations based on convolutional neural networks (CNNs) are introduced in the framework of the RSIR in [65], [68]. In [66], 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 that represents each RS image with CNN features and then applies image-to-class distance measures for retrieval problems is presented in [67]. 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 [63]. Image representations obtained through binary codes are discussed particularly for scalable image search and retrieval in [62], [64]. To obtain the binary codes, in [62] a deep hashing neural network that exploits the cross-entropy loss is presented, whereas in [64] 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 modelling 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 paper 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: 1) image captioning; 2) a sentence encoding, and 3) 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 based on 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. In 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 a query to retrieve the most similar images. Figure 2.1. shows the block diagram of the proposed retrieval system.

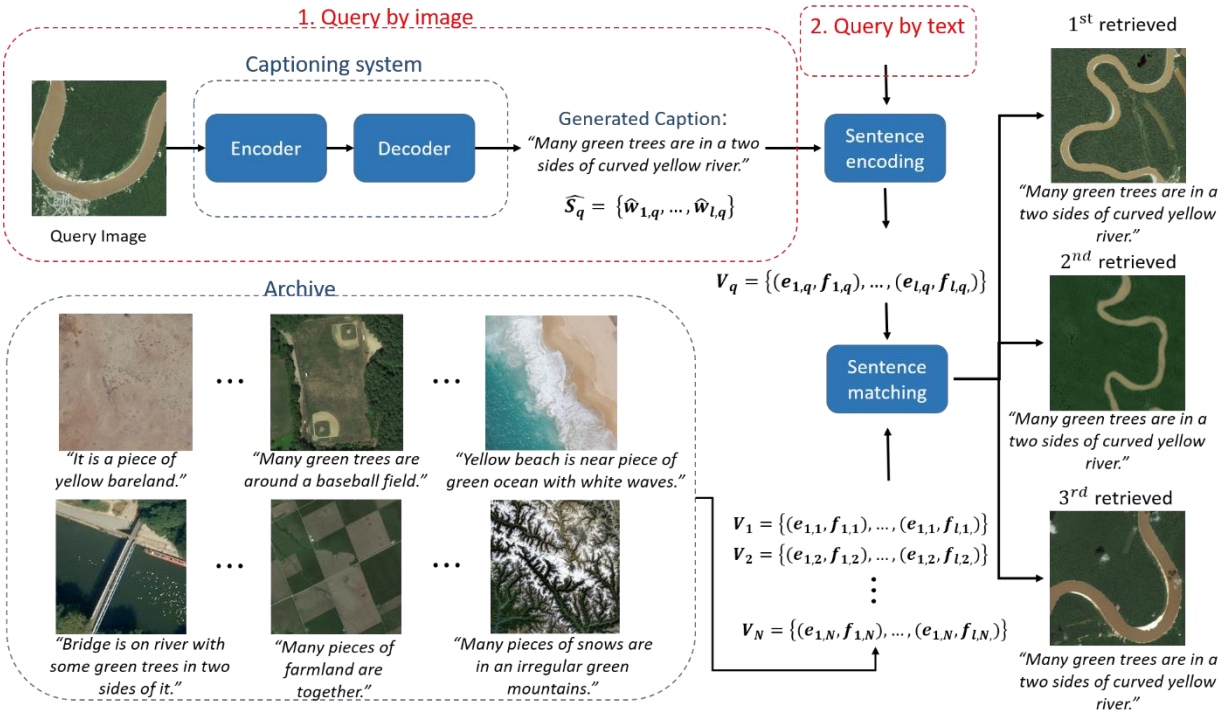


Figure 2.1 Block diagram of the proposed retrieval system. Configuration 1 allows users to query and retrieve images from the archive using image as query. Configuration 2 lets users use directly a textual description to query and retrieve images from the archive. In this work, the default scenario is configuration 1.

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 textual descriptions. The proposed RSIR system has been briefly presented in [69] with limited experimental analysis. This paper extends our work by introducing a detailed description of the proposed approach with a thorough experimental analysis. Another work recently published is in [4], 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 [4]. 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 [4] the query can be only in the form of a textual description.

2.3. Methodology

Let $\mathrm{X} = \{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{N}\}$ be an archive of N images and X_i be the i^{th} image present in the archive. Each image in the archive is associated with J ground truth textual descriptions (i.e., captions). Let $S_{i,j} = \{w_{1,j}, w_{2,j}, \dots, w_{p,j}\}$, $\mathrm{j} = 1,2,\dots,\mathrm{J}$ be the jth textual description of the image X_i and w_p, $p = 1,2,\dots,P$ 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 $Y = \{Y_1, Y_2, \dots, Y_r\}$ of the most similar images to X_q from the archive with high accuracy. To this end, the proposed methodology consists of three main steps: 1) image caption generation; 2) sentence encoding; and 3) image retrieval based on the encoded sentences of images. The block diagram is shown in Figure. 2.1.

2.3.1 Image Caption Generation

Due to the success of the encoder-decoder IC systems in the RS community, in this chapter, we focus our attention on the use of encoder-decoder 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

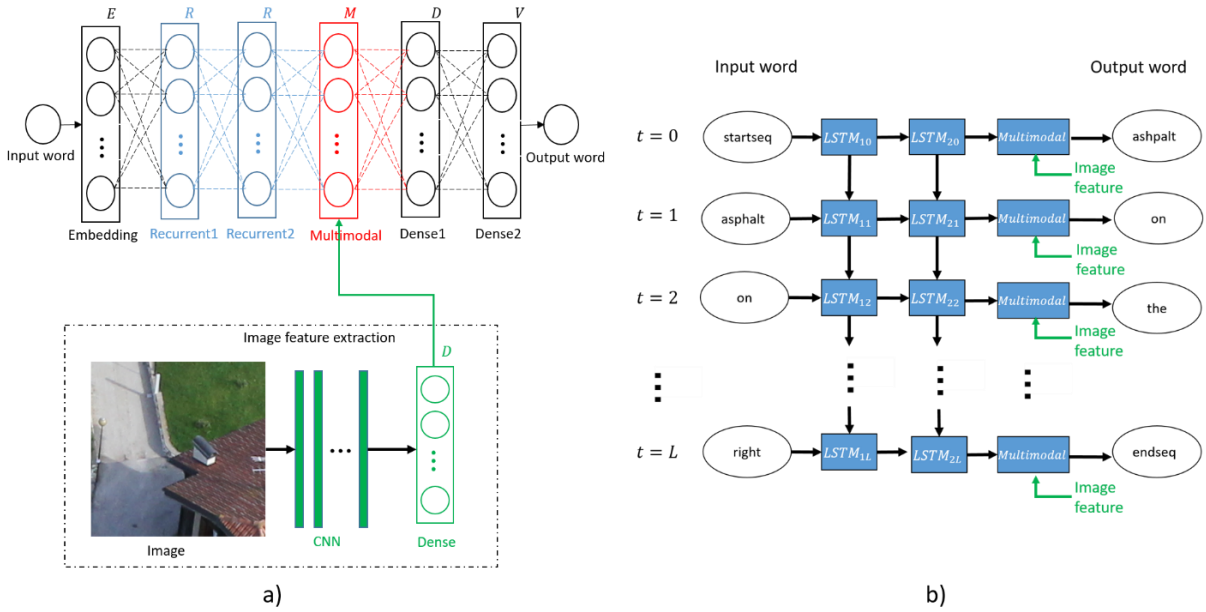


Figure 2.2 The multimodal recurrent neural network. a) multimodal recurrent neural network architecture; and b) the word prediction at each time stamp t regarding the input image and its related sentence description (e.g. asphalt on the left and a red roof on the bottom right and some grass on the top 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.

and a CNN to model the language descriptions and the image visual content in a unique multimodal layer [15]. 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 to generate a word-by-word description of the RS image content. In this work, we have used ResNet 50 [70] to extract representative image features and the LSTM network to generate textual descriptions. The multimodal RNN is shown in Figure 2.2.

Let x_t be the one-hot encoding that represents a given word $w_t \in V$ where V is the vocabulary (whose size is the total number of words). Let h_t be the RNNs' memory. At each time step, RNNs take as input x_t and produces an output y_t which is a combination of h_{t-1} and the current input x_t. Equations 2.1-2.3 describe the RNN, where e_t is the so-called embedding layer which projects the one-hot encodings of the words into a semantic space where words having similar meanings are scattered near to each other, g_1 is element-wise logistic sigmoid function and U, V, W represent the weights to be learned and b the relative biases.

$$e_t = W_x x_t \tag{2.1}$$

$$h_t = g_1(U_w e_t + U_h h_{t-1} + b) \tag{2.2}$$

$$y_t = V_h h_t \tag{2.3}$$

The output is then combined through a multimodal layer with the image features obtained by ResNet 50 as follows:

$$m_t = g_2(U_y y_t + U_F F_i + b) \tag{2.4}$$

where g_2 is the Rectified Linear Unit (ReLU). The last layer of the change captioning model consists of a fully connected layer of size V (vocabulary size) with a softmax activation function (Equation 2.5.) which estimates the probability of the next word of the sentence given the combination of the image features and the previously predicted words.

$$s(m_t) = e^{m_t} / \sum_{k=1}^{V} e^{m_k} \tag{2.5}$$

In the training phase, the model receives the image X_i along with one of the reference change descriptions and the goal of the model is to learn the parameters θ that maximize the probability of constructing the correct sentence that describes the changes between the different acquisitions as follows:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{X_i,S} \log p(S|X_i;\theta) \tag{2.6}$$

Decomposing the sentence descriptions into words Equation 2.6 becomes:

$$\log p(S|X_i) = \sum_{l=1}^{L} \log p(w_k|X_i, w_1, \dots, w_{k-1}) \tag{2.7}$$

where $p(w_k|X_i, w_1, \dots, w_{k-1})$ is the probability of predicting w_k (k^{th} word) given the image information and the previously predicted words. Two special tokens are inserted into each sentence to account for the start and the end of a sentence. Finally, the best parameters of the model are estimated by minimizing the following standard cross-entropy loss function

$$Loss(X,S) = -\sum_{i=1}^{l} log p_i(w_i). \tag{2.8}$$

In the test phase, the images are given as input to the model to generate the sentence description. The prediction process ends when the special token indicating the end of the sentence is predicted. Different studies have shown that the simple RNNs are affected by gradient vanishing/exploding problems which means that RNNs tend to forget faraway previous information that might be relevant for the prediction of subsequent words. To overcome this problem researchers have come up with LSTM [34] and their variation Gated Recurrent Unit (GRU) [71]. These special RNNs are specifically designed to cope with gradient vanishing/exploding problems. They have a more complex internal structure composed of different gates that allow the backpropagation of the gradients through time with uninterrupted gradient flow [34], [71]. In particular, in our captioning system, we have used the LSTM to generate descriptions. The LSTM architecture is illustrated in Figure 2.3. 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 (see Figure 2.3.). 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 embedding e_t are first passed through the forget gate f_t 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 e_t are passed to the input gate represented again by a sigmoid function and to a

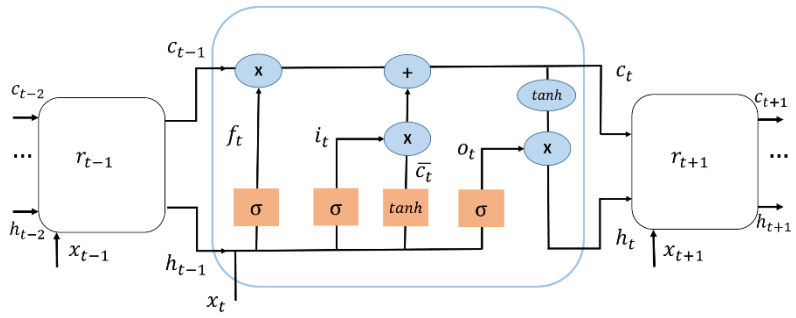


Figure 2.3 LSTM architecture.

tanh layer to decide the value to be updated and the candidates to such values, respectively. The outputs of input gate i_t and of tanh layer $\bar{c_t}$ 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, h_{t-1} and e_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 tanh layer. Equations 2.9-2.14 describe the LSTM inner layers and information update where W represents the weight parameters to be learned and $*$ represents the Hadamard product. These Equations substitute Equation 2.2 of the multimodal network.

$$f_t = \sigma(W_f[e_t, h_{t-1}] + b) \tag{2.9}$$

$$i_t = \sigma(W_i[e_t, h_{t-1}] + b) \tag{2.10}$$

$$\bar{c_t} = \tanh(W_c[e_t, h_{t-1}] + b) \tag{2.11}$$

$$c_t = f_t * c_{t-1} + i_t * \bar{c_t} \tag{2.12}$$

$$o_t = \sigma(W_o[e_t, h_{t-1}] + b) \tag{2.13}$$

$$h_t = o_t * \tanh(c_t) \tag{2.14}$$

2.3.2 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 this work, these vectors are used as features 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; [72] and 2) GloVe [73]. Based on the word co-occurrence both techniques are capable of 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 [72]. Word2vec comes with two different predictive models; 1) the Continuous Bag of Words model (CBOW), 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 [74] 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 on 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 $\widehat{S_i} = \{\widehat{w_{1,i}}, \widehat{w_{2,i}}, \dots, \widehat{w_{p,i}}\}$ representing the image X_i are encoded as $V_i = \{(e_{1,i}, f_{1,i}), \dots, (e_{p,i}, f_{p,i})\}$, 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 subsection.

2.3.3 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) [75], which is a special case of the well-known earth mover distance (EMD) [76] metric.

The WMD uses the word vectors scattered in the semantic vector space 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 (nBOW), where $f_i = w_i / \sum_{k=1}^{n} w_k$ is the number of times the word w_i appears in S divided by the total number of words composing S. Let $c(i,j) = \left\| e_i - e_j \right\|_2$ 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 the word i in S should be transferred to the word j in S', [75] 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:

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{i,j} \cdot c(i,j) \quad i,j \in \{1,2,\dots,n\} \tag{2.15}$$

$$subject\ to: \sum_{j=1}^{n} T_{i,j} = f_i \quad \sum_{i=1}^{n} T_{i,j} = f_j \tag{2.16}$$

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.

2.4. Data Description and Experimental Setup

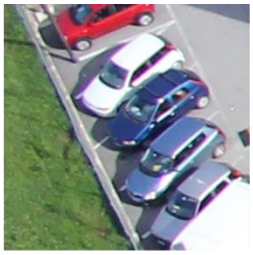
2.4.1 Dataset Description

To evaluate the proposed method, we used two different RS datasets:

1) Unmanned aerial vehicles (UAV) image captioning dataset: The first dataset consists of images acquired by unmanned aerial vehicles (UAVs) with EOS 550D camera near the city of Civezzano, Italy on October 17, 2012. The dataset is composed of 10 RGB images of pixel size 5184×3456 characterized by a spatial resolution of $2\ cm$, of which 6 images are used for training, 1 image for validation and 3 images for the 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 descriptions written by three different human annotators. Examples of frames along with their descriptions are shown in Figure 2.4. The vocabulary size V of the dataset is 185. Since we make a comparison between our method and multilabel image retrieval, each image is labelled with one or more labels based on the ground truth descriptions. The total number of 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 [24]. It is composed of more than $10,000$ RS images gathered from different maps with various resolutions. Thus, it

1. Vineyard with red roof on the left.
2. A vineyard and a red roof on the left.
3. A small red roof in the upper left is close to vineyards.
4. Vineyard with red roof on the top.



1. There are six cars in the parking lot near to a small grass field.
2. Six cars in the parking lot near a field of grass on left.
3. Five cars in the center are next to grass at the left and asphalt at right.
4. Four cars in the parking lot near grass field.

1. Building facade with red roof on the top.
2. A building facade and a red roof on the top.
3. A red roof at upper is close to a building facade.
4. Building facade and red roof on the top.

Figure 2.4 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.

is the largest dataset used for RS image captioning. Each image has a 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 is 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.

2.4.2 Experimental Settings

As 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. To avoid overfitting, dropout is also applied. After 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 dimensions $V = 185$ and $V = 3323$ for UAV and RISCD datasets, respectively. We randomly selected i) 60% of images to derive the training set; ii) 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 to the query image X_q.

Sentence encoding is performed using GloVe and fasText. GloVe vectors are pre-trained on Wikipedia 2014 + Gigaword 5 corpus. They are available at the Stanford website [77]. 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 trade-off between the accuracy and computational complexity.

2.4.3 Multilabel Image Retrieval System

To evaluate the performance of the proposed method we compare it with the multilabel image retrieval. As 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 [78] with the difference in the last layer in which we use a dense layer with 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 sigmoid activation. The output of $sigmoid$ 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 $\theta_{th} = 0.5$.

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 to the query image. The images having the lowest Hamming distance to the query one are retrieved.

2.4.4 Evaluation metrics

The effectiveness of the proposed image retrieval system is quantified using three different metrics: BLEU score [79], F-score [80] and user evaluation. To define the different metrics, let X_q be the query image along with its j different descriptions $S_{q,j} = \{w_{1,j}, w_{2,j}, \dots, w_{p,j}\}$ 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 $S_{r,j} = \{w_{1,j}, w_{2,j}, \dots, w_{p,j}\}$ 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 $\mathrm{P(N, G, R)}$ and the brevity penalty $BP(G, R)$ as follows:

$$BLEU(N, G, R) = P(N, G, R) \times BP(G, R) \tag{2.17}$$

where $\mathrm{P(N, G, R)}$ is the geometric mean of n-gram precision defined as:

$$P(N, G, R) = \left(\prod_{n=1}^{N} p_n \right)^{1/N} \tag{2.18}$$

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:

$$BP(G, R) = \min\left(1.0, \exp\left(1 - \left(\frac{len(R)}{len(G)}\right)\right)\right) \tag{2.19}$$

where $len(R)$ is the length of the reference translation and $len(G)$ 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 ($e.g.\ n = 4$), the 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 [81], which replaces the 0 scores 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 the 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 by 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 the case of multilabel information [80]. 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 β = 1,2. The equations of precision, recall and F-score are given in Equation 2.20, 2.21 and 2.22 respectively, where N_r is the number of retrieved images.

$$Precision = \frac{1}{N_r}\sum_{r=1}^{N_r}\left|\frac{C_q \cap C_r}{C_r}\right| \quad 2.20$$

$$Recall = \frac{1}{N_r}\sum_{r=1}^{N_r}\frac{\left|C_q \cap C_r\right|}{\left|C_q\right|} \quad 2.21$$

$$F_\beta = \frac{(\beta^2+1)Precision \times Recall}{\beta^2 \times Precision + Recall} \quad 2.22$$

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 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, and 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 images 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.

2.5. Experimental Results

2.5.1 Experimental Results on UAV Dataset

In this subsection, we evaluate the proposed retrieval system in terms of the 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 the query and retrieving the desired most similar images to a query image. As the dataset has more than one ground truth description we randomly pick one of them to use as a query. We repeat this process 10 times and average the results. Table 2.1 and Table 2.2 show the upper bound results and the proposed retrieval system results, respectively. In terms of the word embedding technique, the results of each table are rather similar. We can notice an average gap of 10% in terms of the mean BLEU score between the two tables. We believe that the reason for having this gap is that the proposed retrieval system is affected by several errors, one of which, is the captioning block shown in Figure 2.1 Indeed, observing Figure 2.4 we can notice some errors in the generated sentences. Thus, one way to reduce the gap is to improve the image captioning block.

TABLE 2-1 UPPER BOUND RESULTS IN TERMS OF MEAN BLEU SCORE (B): GROUND TRUTH DESCRIPTIONS ARE USED TO QUERY AND RETRIEVE THE IMAGES.

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

TABLE 2-2 PROPOSED RETRIEVAL SYSTEM RESULTS IN TERMS OF MEAN BLEU SCORE: GENERATED DESCRIPTIONS ARE USED TO QUERY AND RETRIEVE THE IMAGES.

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**

TABLE 2-3 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 [77]	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 2-4 COMPARISON RESULTS BETWEEN THE PROPOSED RETRIEVAL SYSTEM AND MULTILABEL METHOD FROM END USERS ON 20 RETRIEVED IMAGES.

Method	*User Evaluation (%)*
Multilabel image retrieval [77]	48
Proposed retrieval system	**52**

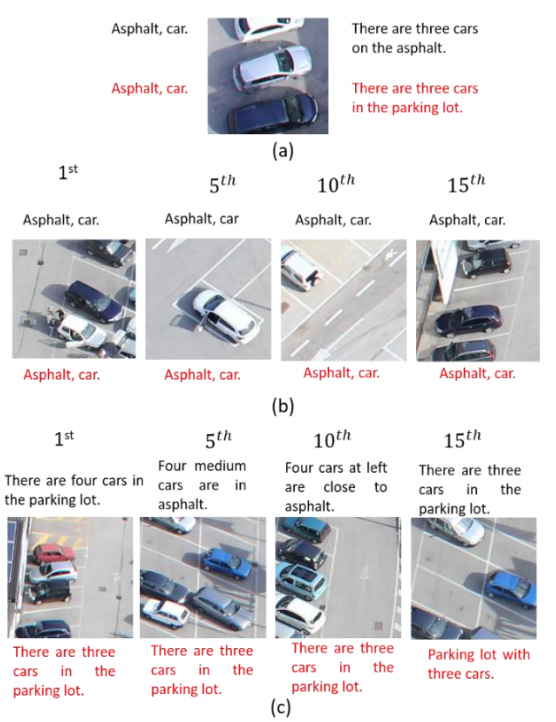


Figure 2.6 Retrieval example. (a) Query image where the ground truth and generated labels are shown on the top and bottom left of the image, respectively; ground truth and generated sentences are shown on the top and bottom right, respectively. (b) Images retrieved using multilabel image retrieval system. 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

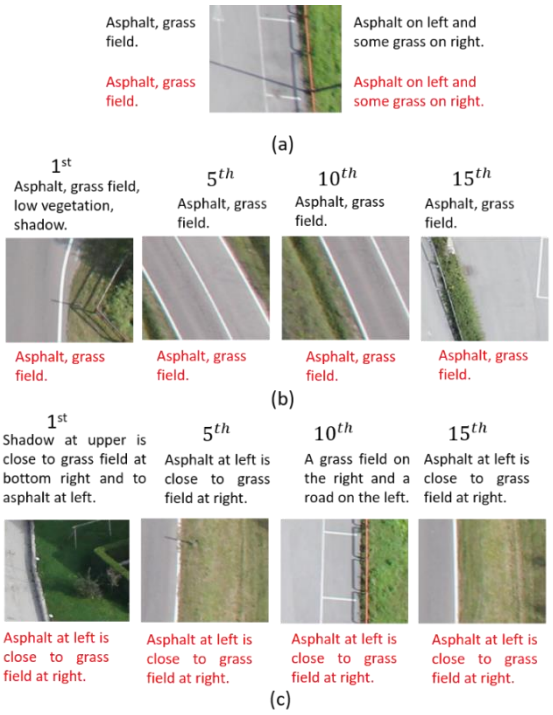


Figure 2.5 Retrieval example. (a) Query image where the ground truth and generated labels are shown on the top and bottom left of the image, respectively; ground truth and generated sentences are shown on the top and bottom right, respectively. (b) Images retrieved using multilabel image retrieval system. 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 2.3. 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 analysing Table 2.3 one can observe that in terms of recall, F-1 and F-2 score our proposed retrieval system achieves slightly better values compared to the multilabel image retrieval system. However, in terms of precision, the multilabel image retrieval shows slightly better results. From the comparison between the proposed retrieval system and the multilabel one, we observed that the results are quite similar as can be seen in Table 2.3. To have a better understanding of the behaviour of the proposed retrieval system we also made a comparison from an end-user perspective between the proposed retrieval system and the multilabel one. Table 2.4 reports the results of the 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 compared to the multilabel image retrieval method [78].

Figure 2.5. 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 image shown in Figure 2.5. (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 below each image in Figure 2.5. (b) and (c), respectively. The retrieved images by the multilabel and proposed retrieval system are shown also in Figure 2.5. (b) and (c), respectively. Both the retrieval systems can find similar images to the query image.

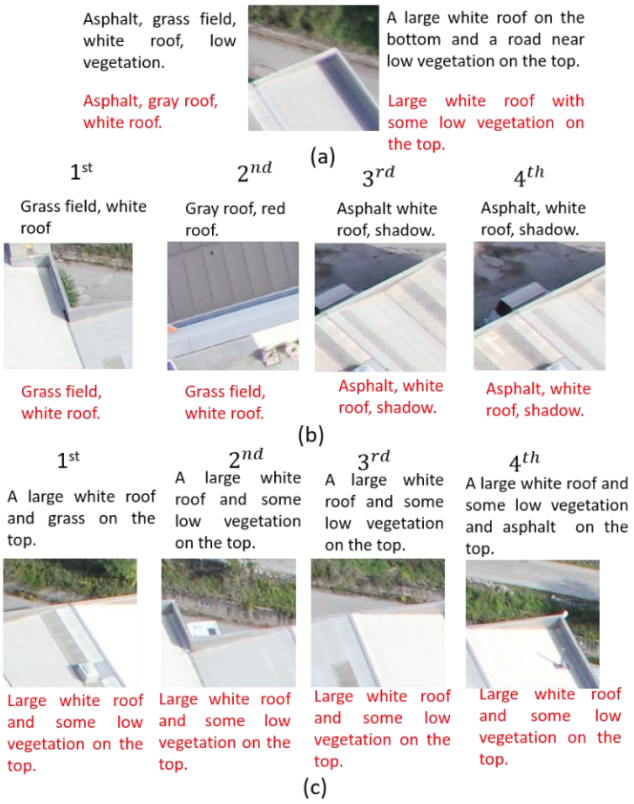


Figure 2.7 Retrieval example. (a) Query image where the ground truth and generated labels are shown on the top and bottom left of the image, respectively; ground truth and generated sentences are shown on the top and bottom right, respectively. (b) Images retrieved using multilabel image retrieval system. 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.

However, the proposed retrieval system is more accurate in finding similar images. Even though the first retrieved image missed "shadow", all the retrieved images of the proposed retrieval system (see Figure 2.6. (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 though the results of the multilabel retrieval system (see Figure 2.6. (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. Figure 2.6. 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 the multilabel retrieval system (see Figure 2.6. (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 Figure 2.6. (c)), show cars parked in 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 Figure 2.7. By looking at the images retrieved by both systems we can see that the proposed retrieval system can accurately find very similar images (see Figure 2.7. (c)) to the query image. On the contrary, the multilabel retrieval system misses different primitive classes and adds others (see Figure 2.7. (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.

2.5.2 Experimental Results on RSICD Dataset

Table 2. V. and Table 2.VI. report the upper bound and the proposed retrieval system results, respectively. As 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. Unlike the results obtained in the UAV dataset, here we observe a reduction in terms of the 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 the two tables varies with the number of retrieved images, from 0.15 to 0.3.

TABLE 2-5 UPPER BOUND RESULTS IN TERMS OF MEAN BLEU SCORE: GROUND TRUTH DESCRIPTIONS ARE USED TO QUERY AND RETRIEVE THE IMAGES.

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**

TABLE 2-6
PROPOSED RETRIEVAL SYSTEM RESULTS IN TERMS OF MEAN BLEU SCORE: GENERATED DESCRIPTIONS ARE USED TO QUERY AND RETRIEVE THE IMAGES.

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**

Figure 2.8 shows an example of images retrieved by the proposed retrieval system when a query image is selected from the 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 descriptions. The generated descriptions are highlighted in red and are given below the retrieved image. From visual inspection of the retrieved images, we can observe that all the retrieved images are very similar to the query image. The 15^{th} retrieved image (see Fig. 2.8(b)) contains some bare land that 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. Figure 2.9. shows another example of retrieved images when the query image is selected from the sparse residential category of the archive. We can observe that all the retrieved images are very similar to the query

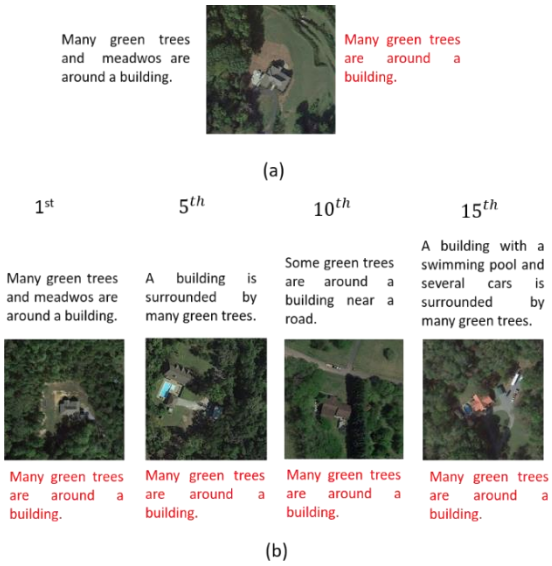


Figure 2.9 Sparse residential image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image

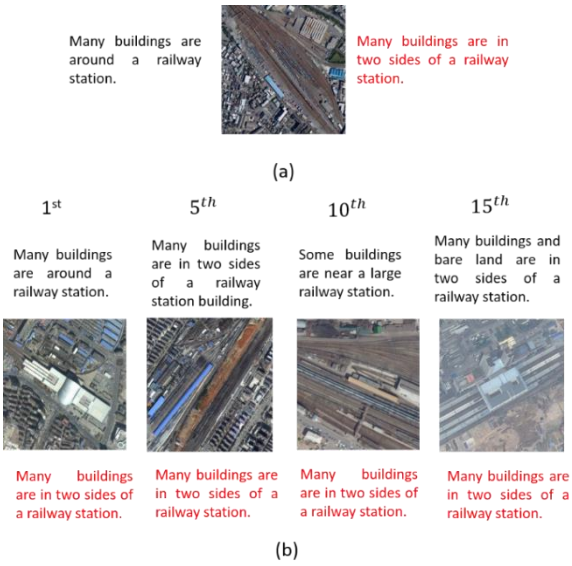


Figure 2.8 Railway station image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

image. Even if the 'meadow' primitive class is missing, all the retrieved images show a single building surrounded by trees.

Figure 2.10 shows another example of the retrieved images where the query image is selected from the port category of the archive. The ground truth descriptions of the query image are in total three:

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 1^{st} retrieved image (see Fig. 2.11(b)) are:

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

We notice 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 are 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. 2.10 (a) and(b)). Furthermore, one can see that all the retrieved images, from a visual inspection, are highly correlated to the query image.

Figure 2.11 shows another example of the retrieved image when the query image is selected from the dense residential category of the archive. The ground truth descriptions of the query images are in total three:

1. "The roof of residential buildings is red"
2. "The wide have a lot of people walking on the road"

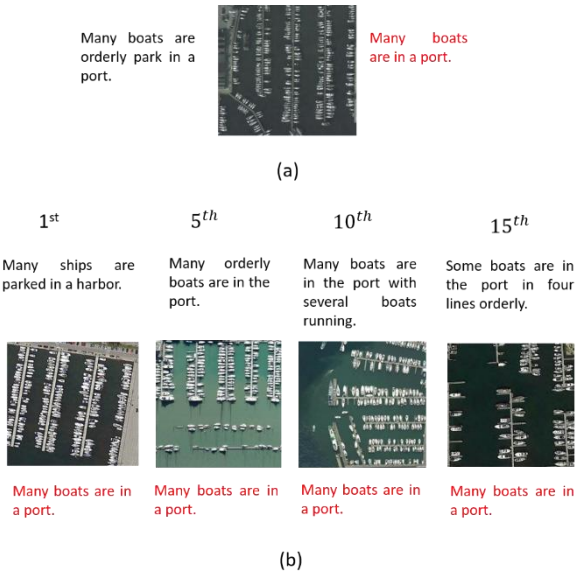


Figure 2.10 Port image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

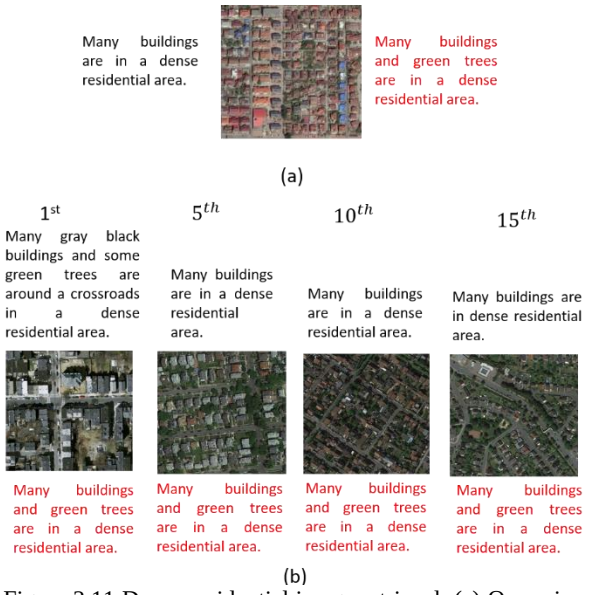


Figure 2.11 Dense residential image retrieval. (a) Query image. (b) Images retrieved using the proposed retrieval system. Generated textual descriptions (highlighted by red) are reported below the related images of (b). One ground truth description is reported above the related images of (b). The order of the retrieved images is reported above each image.

3. "Many buildings are in a dense residential area"

Even in this example, we can notice that the first two ground truth descriptions of the query image 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 to calculate the BLEU score when the ground truth description presents some ambiguity the score will be low. Indeed, for the given example we have BLEU 1, BLEU 2, BLEU 3 and BLEU 4 scores of 0.333, 0.230, 0.207 and 0.163, respectively. However, for the example shown in Figure 2.9. we have BLEU 1, BLEU 2, BLEU 3 and BLEU 4 scores of 0.637, 0.551, 0.514 and 0.456, respectively. This may be one of the reasons why the results reported in Table. 2.6 are much lower compared to Table. 2.5.

From different examples that we have seen from the RSICD, we can conclude that the reason why the results of Table 6 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.

2.6. Final Remarks

In this chapter, we have presented a novel image retrieval system that represents the high-level semantic content of the images by generated sentences and performs image retrieval based on the generated sentences. The main idea and contribution of the chapter is the combination of RS and NLP techniques to perform RS image retrieval. Representing the image content by generated sentences allows expressing better the complex content of an RS scene compared to 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, the 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. In future work, we plan to improve the captioning block.

Chapter 3

3. SVM-Based Decoder

3.1. Motivation

As we saw in Chapter 1 most of the IC systems in the RS community are based on the encoder-decoder framework. In this system, CNNs are used to represent the images with discriminative features and RNNs are used to translate the features into a sentence description. Besides the RNNs, Transformers lately are being used as decoders. In particular, the use of Transformers is combined with self-critical sequence training [51] or variational autoencoder [52] to cope with insufficient training data.

The advantages of the encoder-decoder frameworks that use RNNs or Transformers as decoders is their ability to generate human-like descriptions. However, they are affected by various issues. For example, the performance of these systems depends on the number of annotated training samples (the larger the training set, the lower the risk of overfitting). Indeed, in the CV community, the datasets that are used to train and test IC systems are characterized by a very large amount of annotated samples. An example of such a dataset is the MS COCO dataset [53] which is composed of more than 300,000 images where each image is annotated with 5 descriptions. In contrast to the CV community, in the RS community, datasets used to perform IC are typically small since creating big datasets is not always possible as it is an expensive process in terms of time and resources. Another issue is the high number of hyperparameters that need to be carefully chosen to have good performance. Furthermore, deep learning methodologies demand expensive computational power units like graphic processing units (GPUs) to have reasonable training and testing time. It is worth mentioning that the more complex the system the more acute the aforementioned issues.

To cope with the aforementioned problems, in this Chapter, we propose a novel decoder that is based on a network of support vector machines (SVMs) [82] for those situations in which it is possible to only have a limited number of training samples. SVMs are well-known classifiers in the RS community [83]. They are based on the margin maximization principle that renders them less sensitive to overfitting compared to deep learning methodologies [83]. In this work, a network of SVMs is used as a decoder instead of RNNs or Transformers to alleviate the problem of overfitting and to speed up training and inference time. Another advantage of SVMs is the low number of hyperparameters that need to be chosen to yield an accurate system. The proposed IC framework is shown in Figure 3.1. A CNN extracts image features and represents them with a fixed-length feature vector and a network of k SVM multiclass classifiers in cascade translates the feature vector into a sentence description (i.e., caption). The last SVM multiclass classifier is rendered recurrent to model the dependency on the previous words while generating the new words of a sentence. Note that this work is part of simple encoder-decoder networks that do not explore any kind of attention mechanism.

Overall, the main contributions of this chapter can be summarized as follows:

- A novel decoder architecture based on SVM is introduced for the first time in the framework of IC. It is suitable for situations in which only a few training samples are available to alleviate the problem of overfitting.
- The proposed framework achieves better results compared to simple encoder-decoder frameworks in terms of accuracy and shows comparable and sometimes better results compared with the more sophisticated encoder-decoders that exploit attention mechanisms.
- The proposed method is characterized by an extremely short training and inference time.

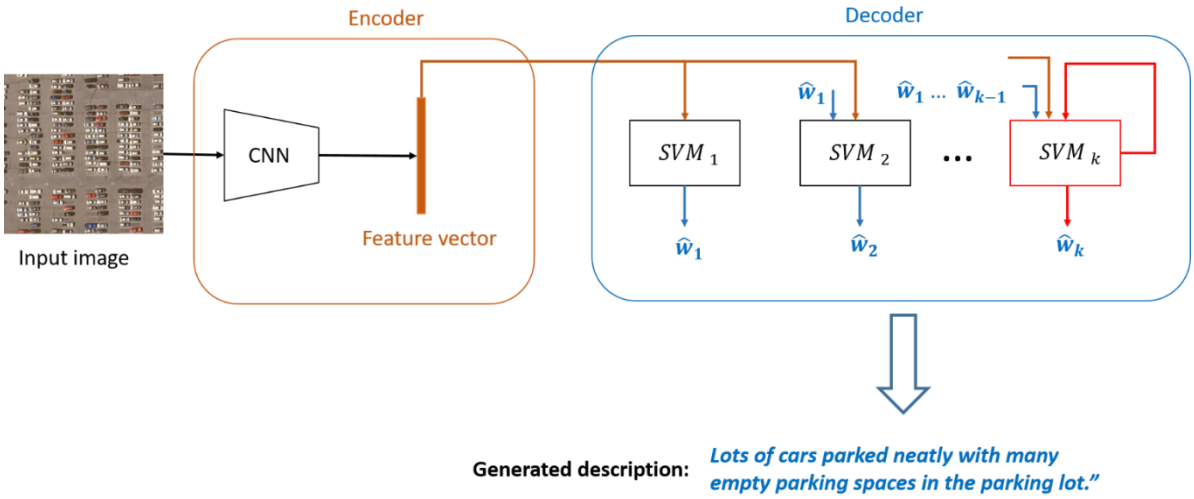


Figure 3.1 An overview of the proposed captioning method. The proposed method consists of two parts: an encoder, which maps the images into feature space and a decoder composed of a network of K SVM multiclass classifiers that generates the captions. The k^{th} classifier (highlighted in red) is rendered recurrent. The prediction process stops when a particular words indicating the end of the sequence is predicted.

3.2. Methodology

Let $\mathrm{X} = \{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{M}\}$ be a training set consisting of M images and X_i be the i^{th} image. Let us assume that each image X_i is annotated with one or more sentence descriptions (or captions). Let $S_i = \{s_{i,j}\}_{j=1}^{J}$ be the set of sentences associated with the image X_i and $s_{i,j}$ be the j^{th} sentence description in the set. Each sentence $s_{i,j}$ can be formulated as a set of ordered words $s_{i,j} = \{w_{ij,l}\}_{l=1}^{L}$ where $w_{i,j,l}$ is the l^{th} word of the sentence $s_{i,j}$ and L is the maximum length of $s_{i,j}$. As with any encoder-decoder IC framework, our proposed IC system is composed of two steps: 1) image representation and 2) sentence generation. The first step aims to represent the input image with discriminative features while the second one is focused on the translation of the features into a sentence description. In this work, we have used a pre-trained CNN to deal with the first part and a network of k SVM multiclass classifiers to deal with the language part. In particular, the k^{th} SVM is rendered recurrent to model the dependency of the previously predicted words while generating the successive words of the sentence description.

3.2.1 Image Representation

The first step of an IC system is to represent the images with discriminative features. To this end, we rely on CNNs since they have shown to be able to overcome the need of hand-crafted features [84]. To be in the same line as in most previous encoder-decoder IC systems in the RS community, in our work we exploit the VGG16 [36] CNN architecture pre-trained on ImageNet [85]. The image features are obtained passing each image X_i through the pre-trained CNN architecture (omitting the last fully connected layer) as follows:

$$f_i = VGG16(X_i) \tag{3.1}$$

3.2.2 SVM decoders

The main difference between the proposed IC system and the previous works is the sentence generation part or decoding stage. While most of IC systems use sequential models such as RNN and LSTM as a decoder, in this work for the first time we develop a network of k SVMs in cascade to decode the features

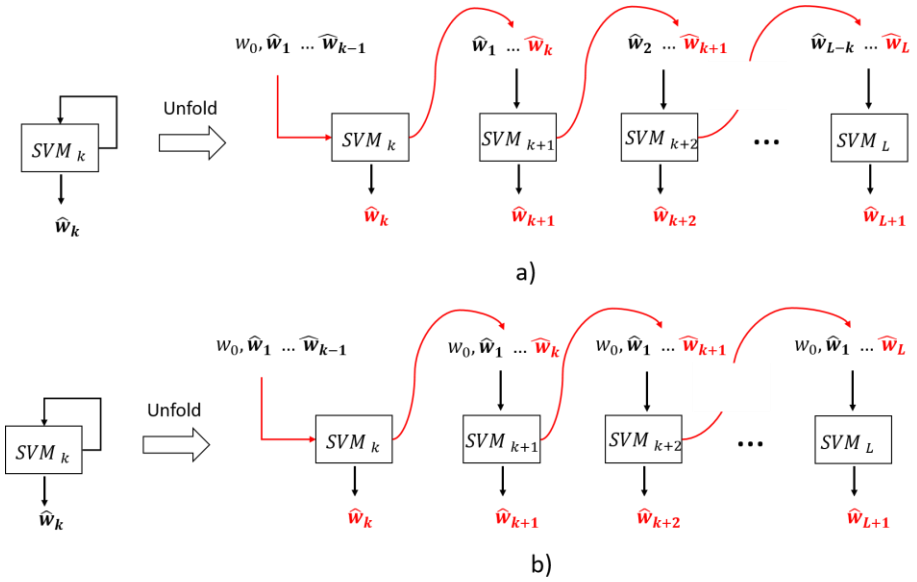


Figure 3.2 Recurrent SVM multiclass classifier a) with word concatenation (SVM-D CONC) and b) with BoW (SVM-D BOW). w_0 and w_{L+1} are special tokens indicating the start and the end of a sentence, respectively.

into a sentence as shown in Figure 3.1. More precisely, given an image X_i and one of its sentence descriptions $s_{i,j}$, the first SVM namely SVM_1 learns the mapping between the feature vector f_i of the considered image X_i and the first word $w_{i,j,1}$ of the sentence $s_{i,j}$ represented by the following formulation:

$$w_{i,j,1} = SVM_1(f_i) \tag{3.2}$$

The second SVM multiclass classifier (SVM_2) in turn learns the mapping between, on the one hand, the feature vector f_i and the first word $w_{i,j,1}$ of sentence $s_{i,j}$ and, on the other hand, the second word $w_{i,j,2}$ as shown in (3):

$$w_{i,j,2} = SVM_2(f_i, w_{i,j,1}) \tag{3.3}$$

Following the same logic, the subsequent $k-1$ SVM multiclass classifiers (SVM_{k-1}) will learn the mapping between, on the one hand, the image features f_i and the previous $k-2$ words $w_{i,j,l}$ (with l= 1,2 ... $k-2$) and, on the other hand, the subsequent word $w_{i,j,k-1}$:

$$w_{i,j,k-1} = SVM_{k-1}(f_i, w_{i,j,1} \dots w_{i,jk-2}). \tag{3.4}$$

The last multiclass SVM classifier namely SVM_k is a particular classifier as it is rendered recurrent. In a recurrent manner, this classifier learns the mapping between the image features f_i and $k-1$ previous words on the one hand and the subsequent $L-k$ words on the other hand where L is the length of the considered sentence. To each sentence are added two special words w_0 'startseq' and w_{L+1} 'endseq' indicating the start and the end of a sentence, respectively. Each word $w_{i,j,l}$ of the sentence $s_{i,j}$ is encoded using one-hot encoding with dimension V which is the vocabulary size. To represent the previous words at a given point l in the sentence, we encode the part of the sentence up to l in two ways: 1) by concatenating the word vectors or 2) by relying on a bag of words (BoW) representation.

1) *Sentence Encoding with word concatenation:* The word concatenation allows for the preservation of the sequential order of the words. The k^{th} SVM multiclass classifier recurrently learns the mapping between the image features and a fixed size of $k-1$ previous words on the one hand and the subsequent $L-k$ words on the other hand. Image features and the $k-1$ previous word vectors are concatenated together. More precisely, at each step (iteration) we have a fixed window size shift of

$k-1$ words while learning the mapping of the subsequent words $w_{i,j,l}$. The recurrent SVM multiclass classifier can capture temporal sequences in an explicit way up to order k, but the input vector may be large. This however is well handled by SVM, which tolerates high-dimensional inputs.

2) *Sentence encoding using Bow:* Bag of words (BoW) is an encoding technique used in the NLP field where each sentence is represented as a vector of a fixed length of vocabulary size V and each entry of the vector represents the number of times that each word appears in the considered sentence. To have the BoW representation of part of a sentence, we simply sum up the one-hot vector representations of the words composing the considered part of the sentence. An advantage of this encoding is that, to learn the mapping of a subsequent word, we exploit all the previous words of the sentence (and not just a subset) while keeping unchanged the size of the generated code (V). A drawback is that the word order is lost. As in the previous sentence encoding strategy, image features and word vectors are concatenated together.

3.3. SVM Training and Inference

For simplicity, let us consider a binary classification. Let us assume to have a training set consisting of M vectors from d-dimensional feature space $x_i \in \Re^d$ $(i = 1,2,\dots,M)$ where each training sample is associated to a positive or negative class $y_i \in \{1,-1\}$. The SVM consists in mapping the data into a higher dimensional feature space i.e., $\Phi(x) \in \Re^{d'}$ $(d' \gg d)$ to find a hyperplane that separates the two classes by minimizing the following cost function:

$$\Psi(\omega,\xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{M}\xi_i \tag{3.5}$$

that is a combination of two criteria, margin maximization and error minimization. ξ_i are the so-called slack variables and C is a regularization parameter. The cost minimization function $\Psi(\omega,\xi)$ is subjected to the following constraints:

$$y_i\left(\bar{\omega}\cdot\Phi(x_i)\right) + b \geq 1-\xi_i, \qquad i = 1,2,\dots,M \tag{3.6}$$

and

$$\xi_i \geq 0, \qquad i = 1,2,\dots,M \tag{3.7}$$

The optimization problem can be transformed into a dual formulation and kernelized, leading to a Quadratic Programming (QP) solution [82]. In the end, the following discriminant function is obtained:

$$h(x) = \sum_{i\in S}\alpha_i y_i \mathrm{K}(x_i,x) + \bar{b} \tag{3.8}$$

where $\mathrm{K}(\cdot,\cdot)$ is a kernel function and S a subset of indices $\{i = 1,2,\dots,M\}$. The binary classification case can be easily extended to multiclass classification following one vs one or one vs all strategies [83].

During the test phase, the features of the test images are given as input to the network of k SVM classifiers to generate the sentence descriptions of the images one word at a time. The sentence generation process of a test image X_t is described with the following equations:

$$\widehat{w_{t,1}} = SVM_1(f_i) \tag{3.9}$$

$$\widehat{w_{t,2}} = SVM_2\left(f_i,\widehat{w_{i,1}}\right) \tag{3.10}$$

$$\vdots$$

$$\widehat{w_{t,k}} = SVM_k(f_i, \widehat{w_{t,1}}, \dots, \widehat{w_{t,k-1}}) \quad (3.11)$$

The recurrent SVM (SVM_k), depending on the sentence encoding (see Fig.3.2.), will recurrently predict the subsequent words based on the image features and the previous words. The sentence generation process stops when predicting the special word 'endseq' indicating the end of the sentences.

3.4. Experiments

3.4.1 Dataset Description

To validate the proposed RS IC system, we conducted experiments on four different datasets: UAV, UCM, Sydney and Remote Sensing Image Captioning dataset (RSICD). The first three datasets are characterized by a small number of annotated images and are more suitable for our IC system. The UAV and RSICD datasets have been introduced in Chapter 2. In the following, we describe Sydney and UCM datasets.

1) UCM caption dataset is based on the UC Merced Land Use Dataset [86] and proposed in [22]. It contains 2100 images of size $256\ x\ 256$ characterized by a spatial resolution of 30.48 cm. Each image is annotated with five different sentences.
2) Sydney caption dataset originates from the Sydney dataset [87] and is proposed in [22]. The dataset is composed of 613 images characterized by a spatial resolution of $50\ cm$. Each image is annotated with five different descriptions.

3.4.2 Evaluation Metrics

The metrics used to evaluate the performances of our RSCC systems are BiLingual Evaluation Understudy (BLEU) [79], Recall-Oriented Understudy for Gisting Evaluation (ROUGE_L) [88], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [89] and Consensus-based Image description Evaluation (CIDEr) [90]. They measure how close the output of an IC system (generated description) is to the descriptions provided by human experts (reference descriptions). In particular, the BLEU score uses the n-gram (n-consecutive words) precision to quantify the similarity between the generated change caption and the reference ones. In this paper, n ranges from 1 to 4. ROUGE_L calculates the F-score with respect to the longest common subsequence between the generated descriptions and the reference ones. METEOR is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. Finally, CIDEr computes the similarity between a generated description and reference descriptions based on the Term Frequency Inverse Document Frequency (TF-IDF). The main idea of CIDEr is to weigh differently the words present in the corpus. The most frequent words are likely to be characterized by low information while the rarest ones have more information.

3.4.3 Experimental Setup

The experiments were performed maintaining the default splitting for the three publically available datasets (UCM, Sydney and RSICD): 80%, 10% and 10% for training, validation and test, respectively. Whereas for the UAV dataset, the splitting is 60% for training, 10% for validation and 30% for test. The vocabulary sizes V for each dataset are 127, 338, 216 and 3323 words for UAV, UCM, Sydney and RSCID, respectively. As previously discussed, the peculiarity of this work is a decoding process based on $K\ SVM$ multiclass classifiers. In our experiments, we adopted $K = 4$, that is we have 4 SVM multiclass classifiers in total where the last one is recurrent. We believe this value captures satisfactorily within-sentence correlation while model complexity keeps contained. It is also noteworthy that in the literature the BLEU metric typically does not go beyond n=4 because within-sentence correlation drops significantly. To provide an in-depth analysis, we conduct an additional thorough experimental study which is reported in subsection 3.5.5.

We rely on a linear SVM multiclass classifier that has only one free parameter which is the value of the regularization parameter C. The best values found on the validation set are $C = 10^{-3}, 10^{-1}, 10^{-2}$ and 10^{-2} for SVM_1, SVM_2, SVM_3 and SVM_4, respectively for all datasets. The SVM implementation is based on LIBLINEAR [91] and is implemented in python. The experiments were conducted on an Intel(R) Xeon(R) CPU E5-1620 v3 @3.50 GHz machine.

The image features are obtained using VGG-16 as a backbone pre-trained on ImageNet. VGG-16 produces a fixed feature vector of 4096 dimensions. Each SVM takes as input the image features and the previously predicted words (if present) to predict the subsequent words. All image and word features are scaled to be in the range [0, 1] using a minimax scaler characterized by the following formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.12}$$

where x is the original feature value, x_{min} and x_{max} are the minimum and the maximum feature values respectively, and x' is the normalized feature value. Note that also the recurrent SVM at each step (iteration) takes as input both image features and previously predicted words to recurrently predict the subsequent ones. Depending on the sentence encoding process, we have a different dependency on the previously predicted words. In the case of sentence encoding with word concatenation, the prediction of a new word depends on a fixed window size of $k = 4$ previously predicted words. This strategy allows for maintaining the word order. Conversely, the sentence encoding is performed using BoW, the prediction of a new word depends on all previous words (from sentence start), but the word order is not preserved. The iteration process of the recurrent SVM ends when a special token indicating the end of the sentence is predicted or the maximum possible length of a sentence is reached. The maximum length of a sentence is chosen based on the longest sentence present in the training set and it is different for each dataset.

3.4.4 Description of Reference Methods

To show the effectiveness of the proposed method, we compare it with some state-of-the-art methods. The majority of these works are based on the encoder-decoder frameworks. Among them, we can find encoder-decoder frameworks without and with attention mechanisms. It is worth noting that our SVM-based IC system belongs to the category of encoder-decoder methods, which do not exploit any kind of attention mechanism. Besides these two major categories, it is noteworthy to mention a retrieval method proposed in [21] and called CSMLF. In this method, images and sentences are mapped into the same latent semantic space and a distance metric is developed to measure the similarity between images and sentences. In the following, we describe briefly the methods used for comparison.

1) *Simple encoder-decoder frameworks*

   a) VLAD+RNN and VLAD+LSTM introduced in [24]: these methodologies are based on hand-crafted features where the encoder is the well-known vector of locally aggregated descriptors (VLAD) [40] and the decoder is the simple RNN and LSTM [34], respectively. LSTM is a more sophisticated version of the simple RNN.

   b) mRNN, mGRU and mLSTM introduced in [22]: these methodologies use deep features where the pre-trained VGG 16 CNN architecture is employed as an encoder to represent the image with a fixed-length feature vector and as decoder are used the simple RNN, LSTM, and GRU respectively. Gated Recurrent Unit (GRU) [92] is a variant of the simple RNN.

   c) mGRU-embed word is the same multimodal framework used in [22] in which images are encoded using the pre-trained VGG16 architecture to obtain the image features represented by a fixed-length

vector and GRU is used as a decoder to generate the descriptions. The difference is that, instead of training the word vectors from scratch, the authors exploit the GloVe pre-trained word vectors [73].

d) Merge GRU-D [93]: this method uses VGG-16 to encode the image features into a fixed-length feature vector and GRU to generate the descriptions. Different from the previous methods, GRU deals only with the sentence part. The image features are concatenated with the GRU output in a subsequent layer to condition the sentence generation with the image information.

2) *Attention-based encoder-decoder frameworks*

a) Soft attention and hard attention are two attention-based methods [18], [32] introduced in the RS community in [24]. The image features are obtained using VGG16 CNN architecture. Unlike the simple encoder-decoder framework that exploits the penultimate fully connected layer to represent an image, here convolutional layers are trained to produce convolutional maps of different parts of the image. The different parts of the images are weighted differently by the LSTM decoder to decide where to focus the attention while generating the words composing the sentence. In the hard attention model, a sampling strategy is used to decide the focus of the attention.

b) ConvCap is an attention-based model that is based on CNNs as encoders and decoders. In particular, VGG-16 is used to encode the image and CNN architecture designed by [94] is employed as a decoder to generate the descriptions.

c) Retrieval Topic Recurrent Memory Networks (RTRMN) [29] is an attention-based method based on ResNet-101 CNN architecture as encoder and a memory network as the decoder. Topic words are extracted and retrieved from the reference descriptions and are used to guide the decoder in generating the descriptions. “RTRMN” semantic and “RTRMN “statistical are two variants of the RTRMN that are based on the semantic topic words and the statistical topic words, respectively.

d) Sound Active Attention (SAA) [28] is another attention-based method that exploits sound information to guide the decoder in generating the description of an image. It uses an encoder VGG-16 and sound GRU to encode the image information and the sound information, respectively, as well as a separate GRU as a decoder to generate the descriptions.

e) SD-RSIC [43] is an attention-based method that exploits the summary of the ground truth captions to guide the decoder to generate the description of an image. It uses different pre-trained CNN and LSTM to encode the image information and generate the descriptions, respectively. To be coherent with the other IC systems, we will consider only the results where VGG-16 is used as an encoder.

3.5. Experimental results

In this section, we discuss the experimental results achieved on four different datasets. The comparison with previously mentioned reference methods is done only for the three publicly available datasets. Regarding our UAV dataset, we have reported only the results achieved by our IC systems and merge GRU-D [93] which we implemented following the instructions in [93]. More details about the implementation of the merge GRU-D method are provided in Section 3.5.5. The SVM-D CONC and SVM-D BOW are the two proposed IC systems that are based on the word concatenation and BoW model to encode the sentences, respectively. The results of most of the methods in terms of accuracy, training and test times are taken from the experiments performed in [28]. The experiments provided by [28] are conducted on Ubuntu 14.04.5 LTS with 48 Intel(R) CPU E5-2650 v4 @ 2.20, which has a better performance compared to our machine. See Table 3-1 for more details. It is worth noting that even though we have trained the merge GRU-D for 50 epochs, on each dataset (Section 3.5.5), the reported results in terms of accuracy, training and testing

times are based on the model which achieves the highest validation accuracy. The epoch number in which the highest validation accuracy is reached differs from dataset to dataset and it is much less than 50. In particular, the epoch number with the highest validation accuracy is $13, 18, 10$ and 3 for UAV, Sydney, UCM and RSICD datasets, respectively.

TABLE 3-1 PERFORMANCE COMPARISON OF THE MACHINES USED FOR THE EXPERIMENTS

Method	Ours (CPU E5-1620)	[28] (CPU E5-2650)
CPU Speed	$4\ x\ 3.6\ GHz$	$\mathbf{8\ x\ 2\ GHz}$
CPU Threads	8	**16**
PassMark	9154	**10953**
Maximum memory size	$375GB$	$\mathbf{750GB}$

3.5.1 Experimental Results on UAV Dataset

Table 3-2 reports the quantitative results of our two IC systems and the merge GRU-D [93] in terms of different metrics, and training and test times. We can see that both of our IC systems show better results compared with merge GRU-D [93]. In particular, we can notice that SVM-D BOW achieves the highest results in almost all the metrics. In terms of training and testing time, we can see that our two proposed decoder is much faster compared to merge GRU-D [93], in particular the testing time. Fig. 4 depicts four examples of images where the first description of each image is one of the reference descriptions, whereas the second, third and fourth ones are generated by GRU-D [93], SVM-D BOW and SVM-D CONC, respectively. We can notice that all the generated descriptions of our two models, even though with some errors, are in line with the image semantic content. In particular, the descriptions of the first and fourth images (see Figures 3.3. a) and 3.3. d)) contain all the semantic information of the image and are very similar to the reference descriptions whereas the descriptions of the second and third images (see Figures 3.3. b and 3.3. c) are affected by some errors or they miss some semantic information. The descriptions generated by the merge GRU-D [93] seem to be more affected by errors. In fact, except for the first image (see Figures 3.3. a) where the description is very accurate, the rest of the descriptions contain some errors related to the position and orientation of the objects. In particular, the generated descriptions seem to be biased towards the word 'bottom'. In the last column of Table 3-2 are reported training and inference times of the models. Our two models are faster than merge GRU-D [93], in particular the test time. Note that the test time represents the time to generate all the descriptions of the images found on the test set.

TABLE 3-2 EVALUATION SCORES (%) AND TIMES NEEDED FOR TRAINING (MINUTES) AND TESTING (SECONDS) ON UAV DATASET. IN BOLD ARE REPRESENTED THE BEST RESULTS WHILE IN ITALIC THE SECOND BEST RESULTS.

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D	Training time (minutes)	Test time (seconds)
Merge GRU-D [93]	*66.03*	55.08	44.87	35.37	31.31	66.76	368.36	4.9	20.20
SVM-D BOW	**68.84**	**58.05**	**48.33**	*39.22*	**32.81**	**69.63**	**391.31**	*3.8*	**1.73**
SVM-D CONC	65.13	*56.53*	*48.15*	**39.69**	*32.17*	*69.31*	*389.45*	**3.3**	*1.87*

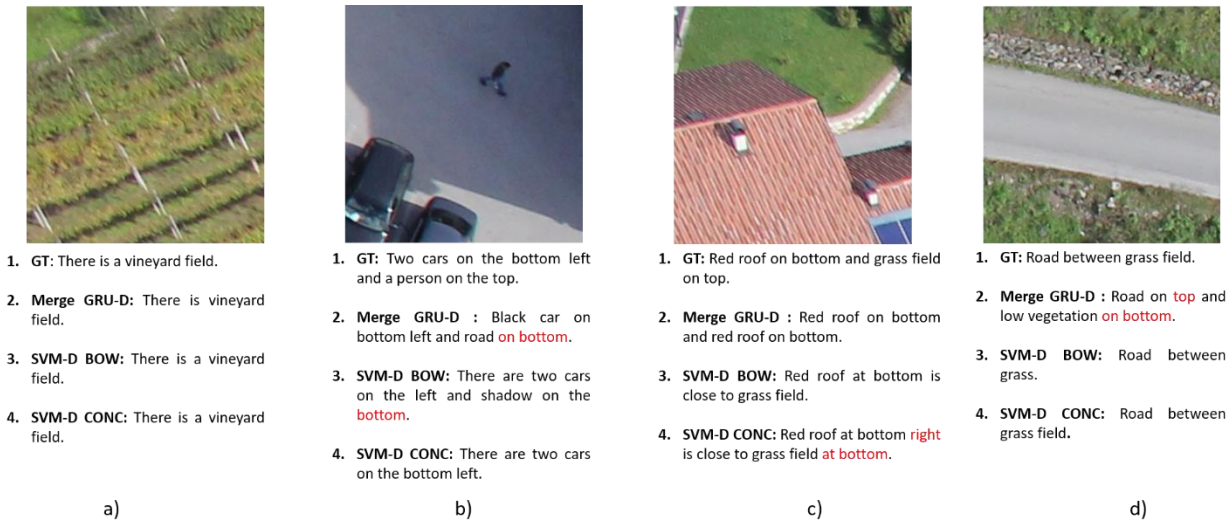


Figure 3.3 Captioning examples of test images from UAV dataset. The first description corresponds to one of the ground truth descriptions while the second, third and fourth descriptions are generated by Merge GRU-D, SVM-D BOW and SVM-D CONC models, respectively. The words in red indicate errors in the generated descriptions.

3.5.2 Experimental Results on Sydney Dataset

In this dataset, we compare the results of our IC systems with thirteen different state-of-the-art IC systems that include retrieval based, encoder-decoder and attention-based encoder-decoder IC frameworks. In Table 3-3 are reported the results of each method where the best and the second-best results are in bold and italic respectively, whereas the "-" symbol indicates that the corresponding metrics are not available for the considered models. In particular, our two proposed IC systems not only can outperform all the simple encoder-decoder frameworks with a good margin in all the metrics but also show comparable or better results compared with attention-based IC systems. It is noteworthy that the results in terms of BLEU-1 and BLEU-2 achieved by SVM-D BOW are the best.

TABLE 3-3 EVALUATION SCORES (%) AND TIMES NEEDED FOR TRAINING (MINUTES) AND TESTING (SECONDS) ON SYDNEY CAPTION DATASET. "- "INDICATES THAT THE CORRESPONDING METRIC ARE NOT AVAILABLE FOR THAT MODEL. IN BOLD ARE REPRESENTED THE BEST RESULTS WHILE IN ITALIC THE SECOND BEST RESULTS.

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D	Training time (minutes)	Test time (seconds)
VLAD+RNN [24]	56.58	45.14	38.07	32.79	26.72	52.71	93.72	-	-
VLAD +LSTM [24]	49.13	34.12	27.60	23.14	19.30	42.01	91.64	-	-
mRNN [22]	51.30	37.50	20.40	19.30	18.50	-	161.00	-	-
mLSTM [22]	54.60	39.50	22.30	21.20	20.50	-	186.00	-	-
mGRU [22]	69.64	60.92	52.39	44.21	31.12	59.17	171.55	-	-
mGRU embedword [22]	68.85	60.03	51.81	44.29	30.36	57.47	168.94	-	-
Merge GRU-D [93]	73.07	63.37	56.41	49.87	33.09	63.34	193.93	4.2	2.45
CSMLF [21]	59.98	45.83	38.69	34.33	24.75	50.18	75.55	-	-
ConvCap [94]	74.72	65.12	57.25	50.12	34.76	66.74	214.84	-	-
Soft-attention [24]	73.22	66.74	**62.23**	**58.20**	**39.42**	*71.27*	**249.93**	-	-
Hard-attention [24]	*75.91*	66.10	58.89	52.58	*38.98*	**71.89**	218.19	-	-
SAA [28]	68.82	60.73	52.94	45.39	30.49	58.20	170.52	-	-
SD-RSIC [95]	72.4	62.1	53.2	45.1	34.2	63.6	139.5	-	-
SVM-D BOW	**77.87**	**68.35**	*60.23*	53.05	37.97	69.92	227.22	*3.37*	*0.39*
SVM-D CONC	75.47	*67.11*	59.70	*53.08*	36.43	67.46	*222.22*	**2.26**	**0.23**

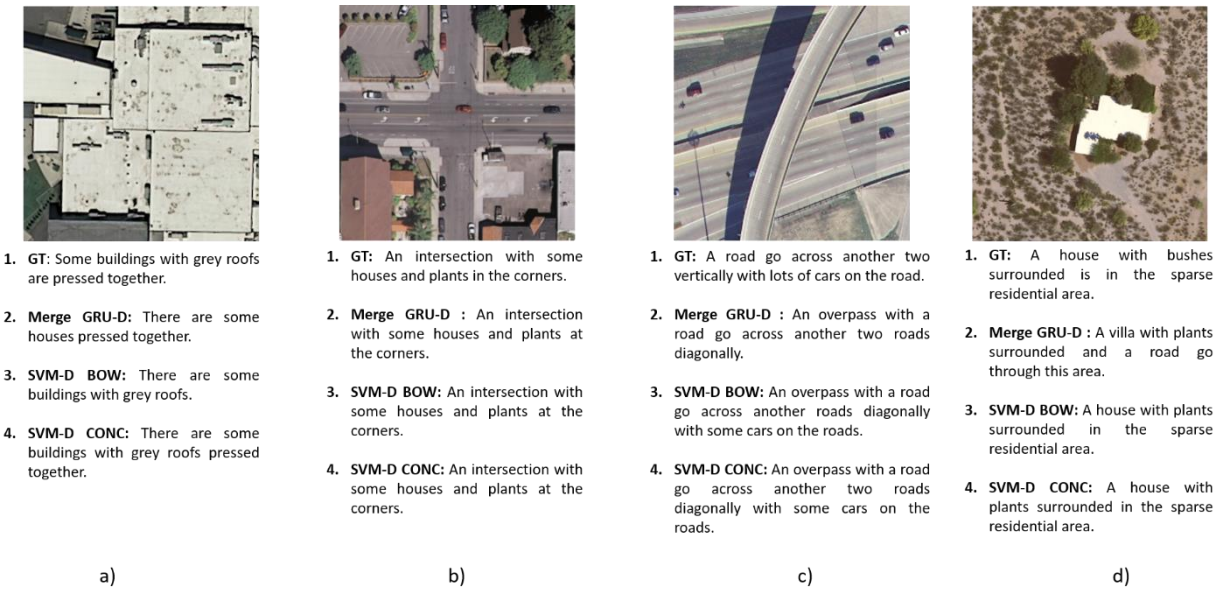


Figure 3.4 Captioning examples of test images from UCM dataset. The first description corresponds to one of the ground truth descriptions while the second, third and fourth descriptions are generated by Merge GRU-D, SVM-D BOW and SVM-D CONC models, respectively. The words in red indicate errors in the generated descriptions.

3.5.3 Experimental Results on UCM Dataset

Here, we compare the results of our IC systems with fifteen different state-of-the-art IC methods that comprise retrieval based, encoder-decoder and attention-based encoder-decoder IC frameworks. In Table IV are reported the results of each method. The proposed IC systems can outperform the simple encoder-decoder and CSMLF methods in all the metrics. Furthermore, we can see that the results achieved by our IC systems, in particular by SVM-D CONC, are also higher compared with some methods that exploit attention mechanisms, such as ConvCap [94], soft attention [24] and RTRMN (statistical) [29]. In the last two columns of Table, 3-4 is reported the training and inference times of each method. We can see that our method is the fastest one.

TABLE 3-4 EVALUATION SCORES (%) AND TIMES NEEDED FOR TRAINING (MINUTES) AND TESTING (SECONDS) ON UCM CAPTION DATASET.
"-" INDICATES THAT THE CORRESPONDING METRIC ARE NOT AVAILABLE FOR THAT MODEL.
IN BOLD ARE REPRESENTED THE BEST RESULTS WHILE IN ITALIC THE SECOND BEST RESULTS.

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D	Training time (minutes)	Test time (seconds)
VLAD+RNN [24]	63.11	51.93	46.06	42.09	29.71	58.78	200.66	121.43	2.57
VLAD +LSTM [24]	70.16	60.85	54.96	50.30	34.64	65.20	231.31	291.18	2.89
mRNN [22]	60.10	50.70	32.80	20.80	19.30	-	214.00	18.64	6.39
mLSTM [22]	63.50	53.20	37.50	21.30	20.30	-	222.50	23.58	5.83
mGRU [22]	42.56	29.99	22.91	17.98	19.41	37.97	124.82	34.00	32.43
mGRU embedword [22]	75.74	69.83	64.51	59.98	36.85	66.74	279.24	31.75	29.76
Merge GRU-D [93]	75.74	67.16	60.63	55.29	37.81	69.11	274.85	14	7.2
CSMLF [21]	36.71	14.85	7.63	5.05	9.44	29.86	13.51	-	-
ConvCap [94]	70.34	56.47	46.24	38.57	28.31	59.62	190.15	1567.32	56.21
Soft-attention [24]	74.54	65.45	58.55	52.50	38.86	72.37	261.24	1251.51	140.02
Hard-attention [24]	**81.57**	73.12	67.02	61.82	**42.63**	*76.98*	*299.47*	1310.45	14.79
SAA [28]	79.62	**74.01**	**69.09**	**64.77**	38.59	69.42	294.51	38.08	37.01
SD-RSIC [95]	74.8	66.4	59.8	53.8	39.0	69.5	213.2	-	-
RTRMN (semantic) [29]	55.26	45.15	39.62	35.87	25.98	55.38	180.25	-	-
RTRMN (statistical) [29]	*80.28*	*73.22*	*68.21*	*63.93*	*42.58*	**77.26**	**312.70**	-	-
SVM-D BOW	76.35	66.64	58.69	51.95	36.54	68.01	271.42	*10.80*	**1.73**
SVM-D CONC	76.53	69.47	64.17	59.42	37.02	68.77	292.28	**9.80**	*1.90*

In particular, our IC systems are 2 to 10 times faster in terms of training time compared to simple encoder-decoder frameworks and 4 to 170 times faster compared with more complicated IC systems that exploit attention mechanisms. We can notice that the same ratios are similar for the inference time.

Figure 3.4 depicts four examples of images from the UCM dataset where the first description of each image is one of the reference descriptions whereas the second, third and fourth descriptions are generated by GRU-D [93], SVM-D BOW and SVM-D CONC, respectively. From a visual inspection, we can see that all the descriptions generated by all the models are highly correlated with the content of the images. In particular, one can notice that the SVM-D CONC seems to produce more complete descriptions compared to SVM-D BOW or to merge GRU-D [93] (see Figures 3.4. a) and 3.4. c)). SVM-D CONC can detect and describe the fact that the buildings are in close contact in Figure 3.4.a) or that there are two roads in Figure 3.4.c) whereas the generated descriptions from SVM-D BOW do not capture this information. Indeed, we can see that this analysis is clearly reflected in Table IV where we have a similar BLEU-1 score for both the systems while regarding BLEU-2, BLEU-3, and BLEU-4 SVM-D CONC shows better results compared to SVM-D BOW. SVM-D CONC is also able to generate more complete descriptions than merge GRU-D [93] as it can be seen from Figure 3.4.c) where merge GRU-D misses the cars on the road.

3.5.4 Experimental Results on RSICD Dataset

The RSICD dataset is the biggest one in the RS community. It is about five times larger than the three small datasets presented so far. It is worth noting that our two SVM-based decoder IC systems are explicitly developed for those situations in which only small datasets are present. RSICD dataset is not part of those situations considering the number of images (more than 10.000) and the number of descriptions (more than 50.000) [24]. Furthermore, this dataset has a vocabulary size of dimension 3323 words that can be translated into 3323 unique classes which become rather complex for an SVM multiclass classifier to deal with. For this reason, we have significantly reduced the vocabulary size of the dataset by taking only the most 430 frequent words in our experiments. This reduction of the vocabulary size resulted in better results and also acceptable training and inference time. We have used the same configuration also with the merge GRU-D.

TABLE 3-5 EVALUATION SCORES (%) AND TIMES NEEDED FOR TRAINING (MINUTES) AND TESTING (SECONDS) ON RSICD CAPTION DATASET. "-" INDICATES THAT THE CORRESPONDING METRIC ARE NOT AVAILABLE FOR THAT MODEL. IN BOLD ARE REPRESENTED THE BEST RESULTS WHILE IN ITALIC THE SECOND BEST RESULTS.

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D	Training time (minutes)	Test time (seconds)
VLAD+RNN [24]	49.38	30.91	22.09	16.77	19.96	42.42	103.92	-	-
VLAD +LSTM [24]	50.04	31.95	23.19	17.78	20.46	43.34	118.01	-	-
mRNN [22]	45.58	28.25	18.09	12.13	15.69	31.26	19.15	-	-
mLSTM [22]	50.57	32.42	23.19	17.46	17.84	35.02	31.61	-	-
mGRU [22]	42.56	29.99	22.91	17.98	19.41	37.97	124.82	-	-
mGRU embedword [22]	60.94	46.24	36.80	29.81	26.14	48.20	159.54	-	-
Merge GRU-D [93]	60.30	42.48	32.03	25.20	22.94	4383	65.90	98.33	90.51
CSMLF [21]	51.06	29.11	19.03	13.52	16.93	37.89	33.88	-	-
ConvCap [94]	63.36	51.03	41.74	34.52	33.25	57.70	166.48	-	-
Soft-attention [24]	*67.53*	**53.08**	*43.33*	*36.17*	**32.55**	**61.09**	**196.43**	-	-
Hard-attention [24]	66.69	*51.82*	41.64	34.07	*32.01*	*60.84*	179.25	-	-
SAA [28]	**67.60**	44.33	**44.33**	**36.45**	31.09	55.36	*193.96*	-	-
SD-RSIC [95]	64.5	47.1	36.4	29.4	24.9	51.9	77.5	-	-
RTRMN (semantic) [29]	62.01	46.23	36.44	29.71	28.29	55.39	151.46	-	-
RTRMN (statistical) [29]	61.02	45.14	35.35	28.59	27.51	54.52	148.20	-	-
SVM-D BOW	61.12	42.77	31.53	24.11	23.03	45.88	68.25	*41.25*	**11.36**
SVM-D CONC	59.99	43.47	33.55	26.89	22.99	45.57	68.54	**35.82**	*23.32*

In Table 3-5 are reported the results of each state-of-the-art method and our two SVM-D IC solutions. We can see that the proposed methods outperform the simple encoder-decoder frameworks (except for the mGRU embedword) and CSMLF. However, the results are lower compared with more sophisticated systems that exploit attention mechanisms. The results of SVM-D BOW and SVM-D CONC are very

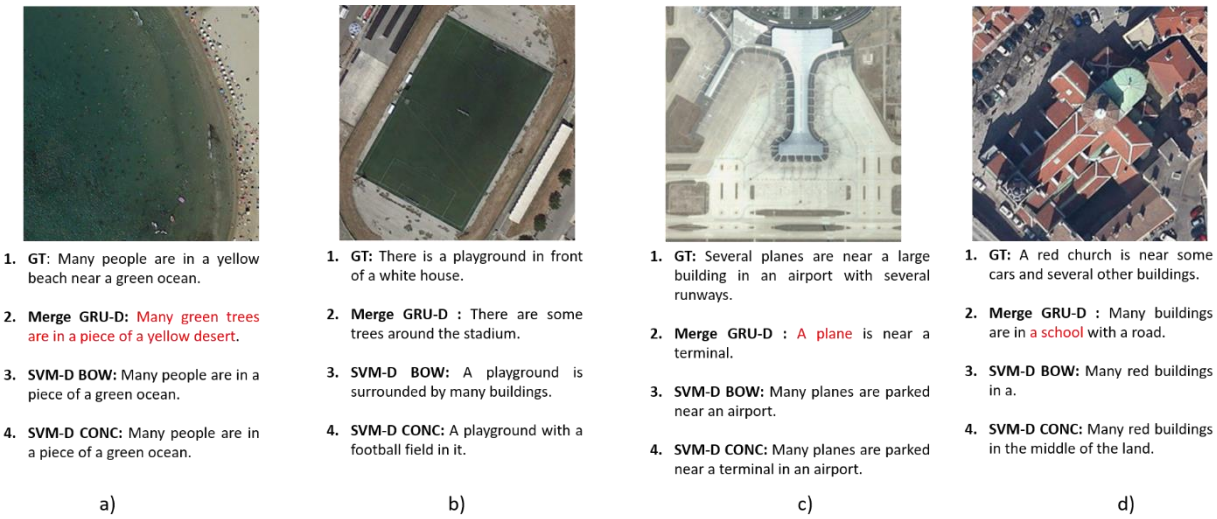


Figure 3.5 Captioning examples of test images from RSICD dataset. The first description corresponds to one of the ground truth descriptions while the second, third and fourth descriptions are generated by Merge GRU-D, SVM-D BOW and SVM-D CONC models, respectively. The words in red indicate errors in the generated descriptions.

similar and we can note the same behaviour as on the other datasets where in terms of BLEU-1 and BLEU-2 SVM-D BOW shows better results while in terms of BLEU-3 and BLEU-4 SVM-D CONC appears the best. Furthermore, they are characterized by short training and inference times. Training and inference time results of other methods are not provided in [28]. Considering the training and inference times of other methods in the UCM dataset, we can expect that the same ratio would be applied also to the RSCID dataset. Figure 3.5 depicts four examples of images from the RSCID dataset where the first description of each image is one of the reference descriptions whereas the second, third and fourth are generated by merge GRU-D, SVM-D BOW, and SVM-D CONC. It comes out that the generated descriptions of our two methods are more in line with the image content compared to merge GRU-D. In particular, the description generated by merge GRU-D of the first image (see Figure. 3.5.a) completely misses the semantic content of the scene.

3.5.5 Impact of Parameter K and Number of Training samples

Since, in the previous experiments, SVM-D CONC has performed slightly better compared to SVM-D BOW on all the datasets (see Tables 3-2, 3-3, 3-4 and 3-5), we will analyse further the proposed decoding approach by running a set of additional experiments on SVM-D CONC decoder. In particular, we will focus on two aspects. The first one is the importance of the parameter K, which controls the number of SVMs to construct the cascaded decoder. The second aspect is related to the impact of the number of training samples on its generalization capability.

For the sake of comparison, we implemented merge GRU-D [93]. Its encoder is the VGG-16 pre-trained on ImageNet (same as ours) and the decoder is GRU. The decoder deals only with the language part and the image features are introduced in a subsequent layer by concatenating them with the GRU output. As in [93], the image features are mapped in an embedding space whose dimension is 128 which is the same as the word embedding layer and the GRU hidden state. Adam optimizer [96] with the default parameters and cross-entropy loss function is used to train the network in 50 epochs [93]. The learning rate is reduced by 10% if there are 2 epochs without any improvement on the validation set performance. The batch size is set to 128.

The parameter K determines the number of SVMs in cascade that compose our SVM decoders. In particular, in SVM-D CONC, it determines the window size of the previously considered words while predicting the successive ones. We varied the parameter K from 1 to 6. Figure 3.6 depicts the results obtained for each dataset. As expected, setting K=1 has a drastic effect on the accuracy. The reason behind this is the fact that

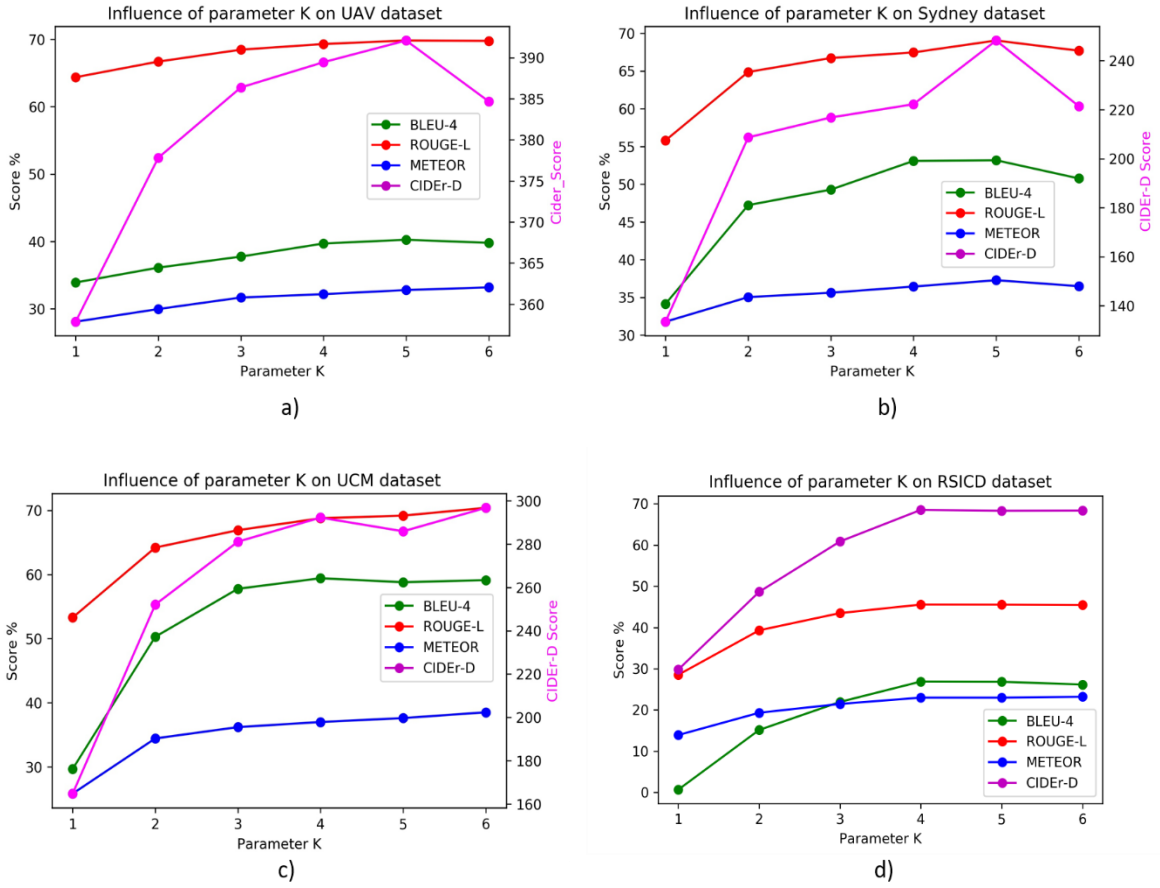


Figure 3.6 Effect of parameter K of SVM-D CONC in all the four explored datasets.

the generation of the subsequent words is limited to the previous word reducing to the minimum the word correlation within a sentence. On the other hand, increasing K extends the window size of the previous words that are considered to predict the subsequent ones leading to better exploitation of a context within a sentence, and in general as a consequence better accuracy. We also can note that when the parameter K goes beyond 4, the improvement ceases to be significant and in some cases start to drop.

To assess further the generalization capability of our method, we run experiments by reducing further the number of training samples. In particular, we randomly generate five subsets of training samples composed of a half (50%) and another five subsets consisting of a quarter (25%) of the original set of training samples. The training of both SVM-D CONC and GRU-D [93] is performed on each dataset. Table 3-6 reports the average metrics as well as the related standard deviation (on the five runs). In particular, our method shows more accurate and stable compared to merge GRU-D except for the Sydney dataset where 50% of training samples are considered. Furthermore, one can also observe that the drop in accuracy of our method is lower compared to the merge GRU-D in almost all the metrics. This confirms, as expected from the intrinsic generalization properties of SVM, that our method is less sensitive to the size of the training set.

TABLE 3-6 MEAN EVALUATION SCORES (%) AND STANDARD DEVIATION ($\mu \pm \sigma$) IN FUNCTION OF AMOUNT OF TRAINING SAMPLES (%) USED TO TRAIN MERGE GRU-D [93] AND SVM-D CONC (OURS).

Dataset	Method	TRAIN %	$\mu \pm \sigma$ B-1	$\mu \pm \sigma$ B-2	$\mu \pm \sigma$ B-3	$\mu \pm \sigma$ B-4	$\mu \pm \sigma$ METEOR	$\mu \pm \sigma$ ROUGE-L	$\mu \pm \sigma$ CIDEr-D
UAV	Merge GRU-D	25	60.01±3.20	49.19±3.54	39.81±3.45	31.68±2.98	28.91±1.68	61.30±4.44	231.74±140.06
	SVM-D CONC	25	59.77 ±0.92	51.83 ±0.67	44.04 ±0.78	37.09 ±1.26	29.28 ±0.53	65.85 ±0.53	365.25 ± 5.7
	Merge GRU-D	50	59.27±3.46	49.05±3.22	39.99±3.13	32.06±2.85	28.78±1.47	60.89±4.21	299.48±66.07
	SVM-D CONC	50	62.33±1.5	53.91 ±1.19	45.70 ±1.05	37.80 ±0.93	30.55 ±0.67	67.56 ±0.69	378.23 ±5.07
	Merge GRU-D	100	**66.03**	55.08	44.87	35.37	31.31	66.76	368.36
	SVM-D CONC	100	65.13	**56.53**	**48.15**	**39.69**	**32.17**	**69.31**	**389.45**
Sydney	Merge GRU-D	25	69.78±2.03	59.91±1.73	52.75±1.65	46.65±1.76	33.41±1.23	62.84±2.15	195.27±15.25
	SVM-D CONC	25	73.88±0.83	64.71±0.87	57.32±0.76	50.90±0.64	35.81±1.01	65.81±1.56	213.02±12.37
	Merge GRU-D	50	70.97±1.63	61.4±1.25	53.72±1.34	47.12±1.69	34.79±1.50	64.26±1.47	197.10±9.6
	SVM-D CONC	50	75.09±2.21	66.03±2.42	58.15±2.73	51.18±2.96	36.01±0.90	67.08±1.29	215.93±13.49
	Merge GRU-D	100	73.07	63.37	56.41	49.87	33.09	63.34	193.93
	SVM-D CONC	100	**75.47**	**67.11**	**59.70**	**53.08**	**36.43**	**67.46**	**222.22**
UCM	Merge GRU-D	25	69.12±1.72	59.66±2.23	52.66±2.69	46.92±3.00	32.56±1.19	63.51±2.01	233.21±14.58
	SVM-D CONC	25	71.50±2.64	62.93±3.07	56.72±3.54	51.33±3.92	33.08±1.95	63.49±2.93	253.26±21.54
	Merge GRU-D	50	72.73±0.80	63.51±0.91	56.92±1.03	51.39±1.11	35.12±1.11	66.24±1.10	251.86±7.20
	SVM-D CONC	50	75.40±1.02	67.49±1.25	61.73±1.53	56.64±1.86	36.37±0.55	67.50±1.03	281.82±10.70
	Merge GRU-D	100	75.74	67.16	60.63	55.29	37.81	69.11	274.85
	SVM-D CONC	100	**76.53**	**69.47**	**64.17**	**59.42**	**37.02**	**68.77**	**292.28**
RSICD	Merge GRU-D	25	54.46±1.33	36.96±1.03	25.29±3.02	19.95±0.89	23.99±8.83	39.60±0.93	51.68±1.60
	SVM-D CONC	25	58.15±0.18	41.13±0.27	31.24±0.38	24.69±0.40	21.74±0.07	43.63±0.38	61.16±0.90
	Merge GRU-D	50	56.25±1.28	38.57±1.09	28.10±0.92	21.28±0.87	21.13±0.71	41.35±1.18	56.04±2.77
	SVM-D CONC	50	59.57±0.65	42.52±0.72	32.56±0.79	25.91±0.84	22.47±0.39	44.76±0.56	65.20±1.08
	Merge GRU-D	100	**60.30**	42.48	32.03	25.20	22.94	43.83	65.90
	SVM-D CONC	100	**59.99**	**43.47**	**33.55**	**26.89**	**22.99**	**45.57**	**68.54**

3.6. Final Remarks

In this chapter, we have presented a novel remote sensing image captioning system that is based on a network of SVM multiclass classifiers. The proposed IC system is part of the well-known encoder-decoder family and uses convolutional neural networks to represent the image with a set of discriminative features and a network of SVMs as the decoder (instead of the RNNs or Transformers) to generate the image descriptions. In particular, the last SVM multiclass classifier is rendered recurrent to model the dependency of the past words while generating the subsequent words of a sentence. In particular, the dependency on the previous words is modelled using a fixed window size of previous words (SVM-D CONC) or considering all the past words (SVM-D BOW). The former has the advantage of preserving the word order but within a fixed window size while the latter has no constraint on window size but does not preserve the word order. The proposed system is particularly interesting in those situations characterized by the availability of only a few training samples. It exhibits very short training and inference times. Moreover, it requires the setting of just one hyperparameter, namely the regularization parameter C. The experiments carried out on four different remote sensing captioning datasets confirm the effectiveness of the proposed IC system, especially on small datasets. For future work, we think it worth exploring more sophisticated NLP strategies that can capture better the word dependencies without resorting to a predefined fixed window size while preserving the word order.

Chapter 4

4. Post-Processing Strategies

4.1. Motivation

Designing reliable IC systems is very difficult. Just like any machine learning system, also IC systems are prone to errors. The outcome of an IC system might be affected by misrecognition problems related to the object and their attributes or relationships. Some of these features might be wrongly part of a generated sentence and some other might be omitted. Hence, post-processing can be a useful technique to rectify the generated sentences and improving their quality.

To improve the quality of the outcome of an IC system, we propose two post-processing strategies. The proposed post-processing strategies aim to rectify a generated sentence by detection and correction of the potential errors. These strategies are based on Hidden Markov Models (HMMs) and the Viterbi algorithm. The former aims to generate a set of possible states while the latter aims at finding the optimal sequence of states efficiently. The proposed post-processing strategies are applied at test time and can be injected into any IC system to improve the quality of a generated sentence.

Overall the main contribution of this chapter can be summarized as follows:

- Two post-processing strategies are introduced for the first time in the framework of IC to correct the potential mistakes present in the generated sentences of an IC system. The two proposed postprocessing strategies are based on HMMs and Viterbi algorithm. Sentence rectification can be done either once the sentence is fully generated (post-generation strategy) or during the generation process (in-generation strategy).
- The two proposed post-processing strategies are able to correct the potential errors in the generated descriptions improving the performances of IC models.
- The achieved results of the post-processing strategies applied to a simple encoder-decoder network compete the complex state of the art IC systems.

4.2. Methodology

4.2.1 Image Captioning Architecture

The tested IC architecture is based on multimodal neural network. We have introduced the multimodal neural network in subsection 2.3.3. In this chapter, the decoder is the GRU [71] instead of the LSTM [34]. This because compared to the LSTM, GRU is more simple and present the same advantages. The GRU decoder is shown in Figure 4.1. The information flow in the GRU is controlled by two gates, a reset gate r_t that decides how much previous information stored in h_{t-1} to keep and an update gate z_t that combines the previous information with the new input to update the new state h_t. Equation 2.2 changes with the following equations that describes the GRU:

$$h_t = z_t \cdot h_{t-1} + (1 - z_t) \cdot h'_t \tag{4.1}$$

$$z_t = \sigma(W_z[e_t, h_{t-1}] + b) \tag{4.2}$$

$$h'_t = \tanh(W_h[e_t, r_t \cdot\ h_{t-1}] +\ b) \tag{4.3}$$

$$r_t = \sigma(W_r[e_t, h_{t-1}] + b). \tag{4.4}$$

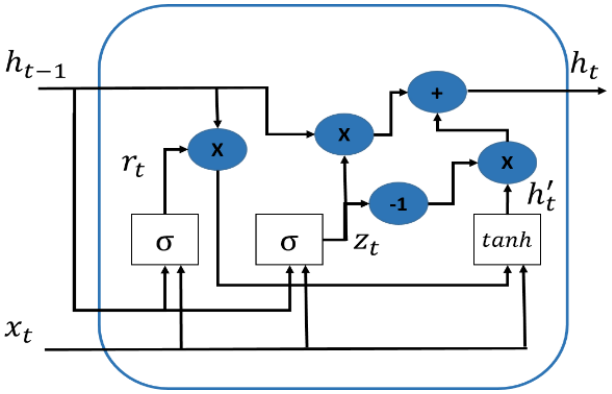
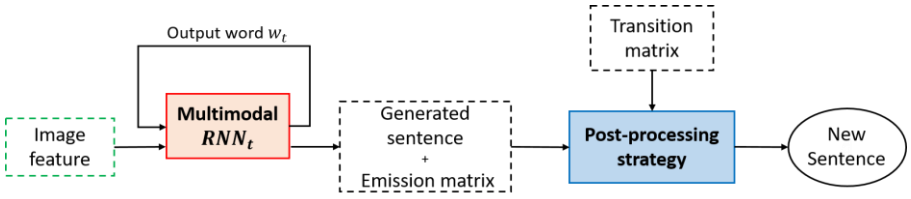
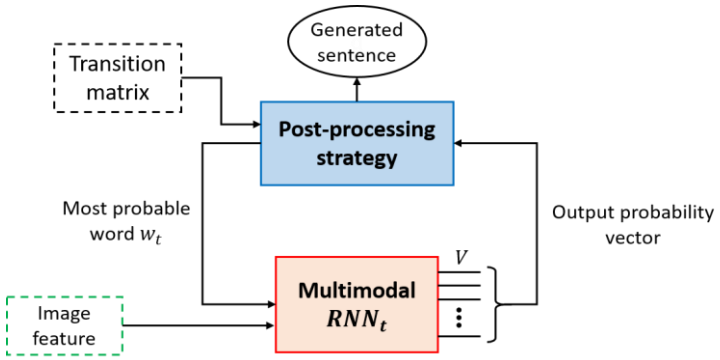


Figure 4.1 GRU architecture.



a)



b)

Figure 4.2 Two proposed post-processing strategies: a) Post-generation strategy and b) in-generation strategy. In a) the detection and correction of potential errors is done once the sentence is fully generated from the captioning system. In b) the detection and correction of potential errors is done sequentially during the generation process.

4.2.2 Proposed Post-Processing Strategies

In this subsection, we explain the two proposed post-processing strategies. These strategies are applied in the test phase, once the model is fully trained. We aim to detect and correct the potential errors of the generated sentences. This is done following two main strategies, the first one namely the post-generation strategy is applied once the sentence is fully generated by the IC model. The second one, namely the in-generation strategy, aims at detecting and correcting potential errors during the sentence generation process. To this end, the in-generation strategy alters at each time steps the probability of the generated words in the

sentence by penalizing the potential error words or giving more weight to more correct words. We will see that this leads not only to the detection and the correction of the potential errors within the sentence but also to the generation of more representative sentences that agree more with the image's visual content. The information used to detect errors is found in the training set itself. Based on the training set and through HMMs we are able to define a set of possible states and model the transition from one state to another. Then through the Viterbi algorithm, we find the best sequential of states, thus forming the best possible path that represents the sentence that possibly describes best the content of the considered image.

4.2.3 HMM

HHMs are widely utilized to analyse time-series data where a series of unknown states is inferred from a sequence of observations [97], [98]. In the context of IC, the finite number of states is the vocabulary of words $w \in V$ (of the training/validation sentence corpus) where V is the vocabulary size and the sequence of observations are the image information (feature vector f) along with the generated words $w_{1:t-1}$ up to time $t-1$. Our goal is to use HMMs to improve the IC model in our hand in generating more coherent sentences given an image. In particular, we seek to rectify the generated sentence from a captioning model by reducing and correcting the potential errors. Formally, let u_t denote the mixed vector of image features and words $u_t = (f, w_{1:t})$ collecting the observed features at time t. Further, let $U = \{u_1, u_2, \dots, u_L\}$ and $\Omega = \{\omega_1, \omega_2, \dots, \omega_L\}$ denote a set of observed feature vectors and possible states, respectively for all timestamps $t = 1, \dots L$ where L is the maximum possible sentence length. The Bayes rule for Markovian models can be expressed as:

$$P(U, \Omega) = P(U|\Omega)P(\Omega) = \prod_{t=1}^{L} P(u_t|\omega_t)P(\omega_t|\omega_{t-1})) \tag{4.5}$$

where $P(\omega_1|\omega_0) = P(\omega_1)$. In the context of IC, the states ω are simply the words $w \in V$. The term $P(u_t|\omega_t)$ in our case represents the output of the IC model. In particular, $P(u_t|\omega_t) = P(w_t|f, w_{1:t-1})$ which is the probability of predicting the word w_t at time t given the image features f and the previously predicted words $w_{1:t-1}$ up to time $t-1$ (Eq. 4.5.). It can be viewed as the emission probability of an HMM. The term $P(\omega_t|\omega_{t-1})$ becomes $P(w_t|w_{t-1})$ and it is the transition probability that models the transition from state w_{t-1} at time $t-1$ to the state w_t at time t and $P(w_1|w_0) = P(w_1)$ is negligible being the first word equal for all the sentences as it indicates the start of a sequence (sentence). The transitions probabilities $P(w_t|w_{t-1})$ are obtained from the sentence training corpus as count of the co-occurrence $C(w_t, w_{t-1})$ of the word (state) w_{t-1} followed by the word w_t over the total count number $C(w_{t-1})$ of w_{t-1}:

$$P(w_t|w_{t-1}) \approx \frac{C(w_{t-1}, w_t)}{C(w_{t-1})}. \tag{4.6}$$

Taking into account all the above considerations Eq13. becomes

$$P(U|\Omega)P(\Omega) = \prod_{t=1}^{L} P(w_t|f, w_{1:t-1})P(w_t|w_{t-1})). \tag{4.7}$$

Applying most likely sequence criterion the most probable (best), or maximum likelihood, sequence of states given the observed data is given by the following equation

$$\widehat{w_{1:L}} = argmax\left\{\sum_{t=1}^{L} \log P(w_t|f, w_{1:t-1}) + \log P(w_t|w_{t-1})\right\}. \tag{4.8}$$

The expression is expressed on a logarithmic scale for practical computational purposes. We find the most likely sequence of states efficiently using the Viterbi algorithm. To control the weight between the emission probabilities $P(w_t|f, w_{1:t-1})$ and the transition probabilities $P(w_t|w_{t-1})$ we introduce a parameter $\beta \in [0,1]$

$$\widehat{w_{1:L}} = argmax\left\{\sum_{t=1}^{L} \beta \log P(w_t|f, w_{1:t-1}) + (1-\beta)\log P(w_t|w_{t-1})\right\}. \quad (4.9)$$

The closer the parameter β to 1 the more important are the transition probabilities and vice versa, the closer to 0 the parameter β, the more important are probabilities emitted by the captioning system. With this parameter, we aim at finding a good trade-off between the contribution coming from the IC models and the HMMs model. We believe that this trade-off, once found, would allow the correction of potential errors in the generated sentence leading to more correct ones.

To further extend the dependence order on previous information we propose to include it in Eq. 17. the transition probability $P(w_t|w_{t-2})$ from state w_{t-2} at time $t-2$ to the state w_t at time t where $P(w_2|w_0) = P(w_2)$. The second-order dependency is incorporated in Eq. 17 as follows:

$$\widehat{w_{1:L}} = argmax\left\{\sum_{t=1}^{L} \beta \log P(w_t|f, w_{1:t-1}) + (1-\beta)\left[\log P(w_t|w_{t-1}) + \log P(w_t|w_{t-2})\right]\right\}. \quad (4.10)$$

where $P(w_t|w_{t-2})$ are obtained from the sentence training corpus as the count of the co-occurrence $C(w_t, w_{t-2})$ of the word (state) w_{t-2} followed by the word w_t (skipping w_{t-1}) over the total count number $C(w_{t-2})$ of w_{t-2}:

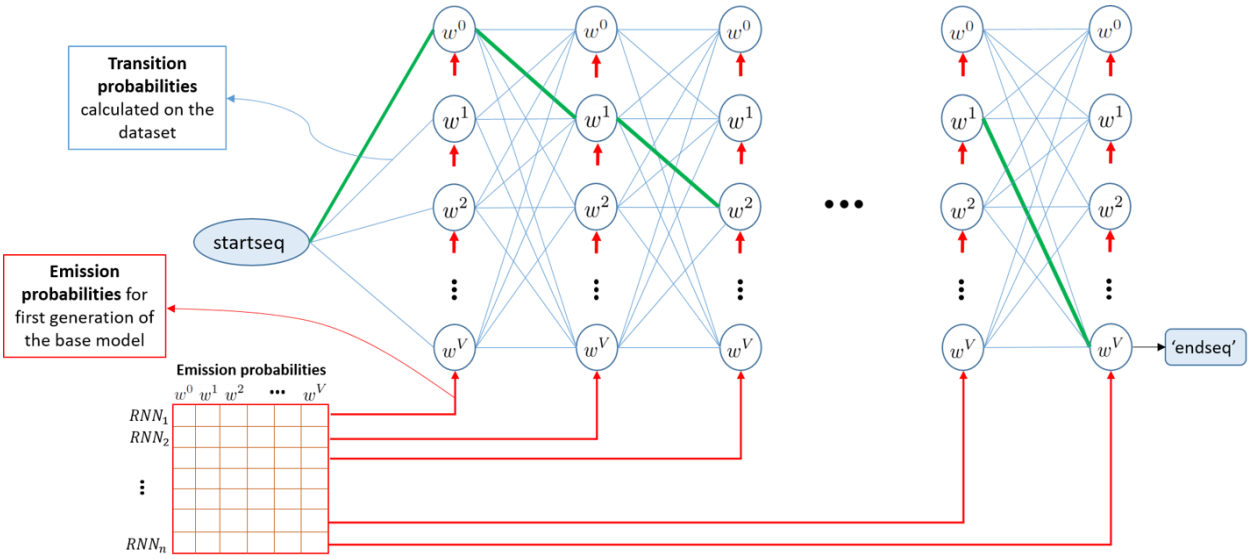
$$P(w_t|w_{t-2}) \approx \frac{C(w_{t-2}, w_t)}{C(w_{t-2})}. \quad (4.11)$$

4.2.4 Post-generation strategy

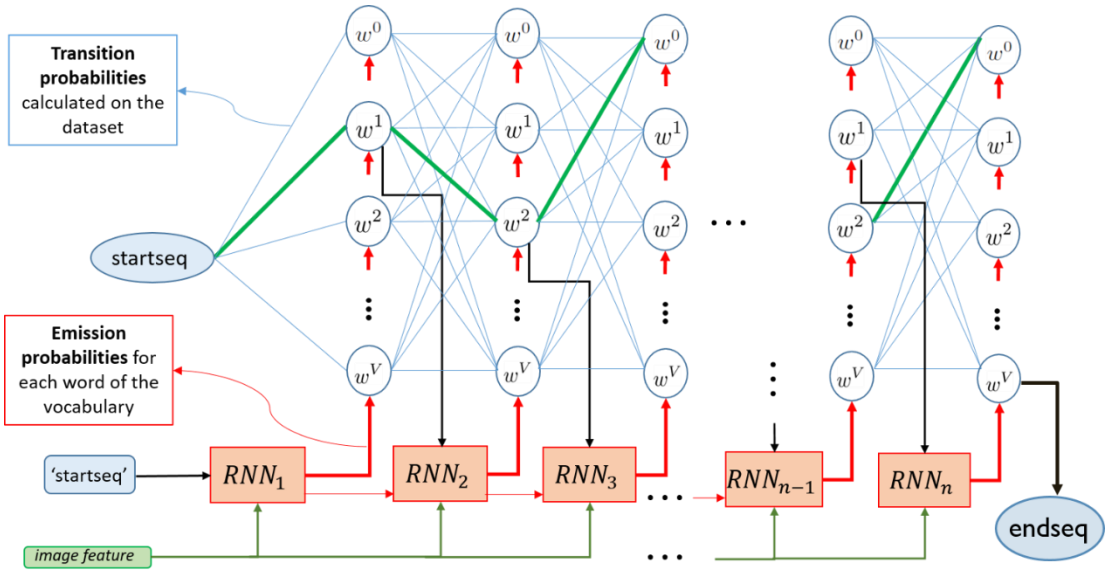
The post-generation strategy aims at correcting the potential errors once the sentence descriptions is fully generated from the IC models. Fig.4.2 a) depicts the scheme block of the post-generation strategies. As it can be seen, to detect and correct the potential errors in the generated sentence we need the emission and transition probabilities saved into matrices. Combining these two items through Eq.4.9. and Eq.4.10. and utilizing the Viterbi algorithm we seek to find the most likely sequence of words that best describes a given image. Thus, the post-generation strategy seeks to correct the potential errors once the sentence is fully generated. An illustration of this process is depicted in Fig. 4.3 a) where the green trellis represents the best path (likely sequence) calculated through Viterbi algorithm.

4.2.5 In-generation strategy

Different from the post-generation strategy, in-generation strategy looks for potential errors while the sentence is being generated instant per instant as depicted in Fig. 4.2. b) and Fig. 4.3.b). The main idea of this strategy is that if a sentence is fully generated it is hard to detect correct the potential errors. This because, the IC generation process is done sequentially and the prediction of the next words depends on image features and the previously predicted words. Hence, once an error word is generated it is going to affect all the subsequent instants and as a consequence it might become hard to detect and correct the potential errors. Furthermore, these errors could be many. Reflecting on the above considerations, the in-generation strategy acts at each time-instant to detect the potential errors. As a consequence, if an error is corrected, the future time-instants will depend on the corrected word which will then lead to a more correct sentence. Fig. 4.3. b) depicts this process. The word prediction at a certain time-instant will thus depend



a)



b)

Figure 4.3 Illustration of two proposed post-processing strategies: a) Post-generation strategy and b) in-generation strategy. In a) the detection and correction of potential errors is done once the sentence is fully generated from the captioning system. In b) the detection and correction of potential errors is done sequentially during the generation process. In the former the Viterbi algorithm is applied only at the end of the generation process, while in the latter, it is applied at each time step of the generation process.

not only on the emission probability of the captioning system but also on the transition probabilities. The combination of the two is done through Eq.4.9. and Eq.4.10. and the most likely sequence (trellis) is found using a simplified version of the Viterbi-algorithm that is applied at each time step. An illustration of this process is depicted in Fig.4.3 b).

4.3. Experimental Results

4.3.1 Experimental Settings

To test the proposed post-processing methods, we have maintained the default split for the publically available datasets (Sydney, UCM and RSICD): 80%,10% and 10% for training, validation and test,

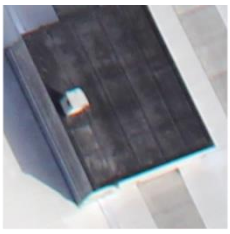
respectively. In the UAV dataset instead the splitting is 60%,10% and 30% for training, validation and test, respectively. The vocabulary sizes V for each dataset are 127, 338, 216 and 3323 words for UAV, UCM, Sydney and RSCID, respectively. The image features are obtained using VGG-16 pre-trained on ImageNet [85], [99]. The choice of VGG-16 is simply because it is the most used CNN in the RS community for image captioning and allows to make a fair comparison with other works. It produces a fixed feature vector of dimension 4096. In our IC system, the original image features are reduced to 256 through a projection layer (dense layer) with activation function ReLu. The same dimensions are fixed for the word embedding layer e_t and the hidden memory of the RNN h_t. The output of the RNN and the image features are fused together through the multimodal layer m_t by elementwise addition preserving the dimension to 256. Another learnable fully-contented layer with the same dimension and ReLu activation function is added before the softmax layer. Adam optimizer [96] with a learning rate of 10^{-4}, cross-entropy loss function and a batch size of 128 are used to train the model. Dropout regularization is applied in different parts of the model to avoid overfitting. In particular, dropout with a rate of 0.5 is applied to the original image features, to the word embedding layer and to the output of the RNN.

The proposed post-processing strategies have only one free parameter that is $\beta = [0,1]$. This parameter weighs the importance of the emission probabilities provided by the IC model and the transition probabilities provided by the HMM. The best β parameter is chosen using a grid search of step size equal to 0.05 on the validation set. The number of states, for each dataset, corresponds to its vocabulary size with the exception the of RSICD dataset. For this dataset, we have used the most frequent words to train both the IC system and to calculate the transition matrix. In particular, we have used the words which have a frequency of 45 (number of appearances in the training set) reducing the vocabulary size (number of states) from 3323 to 368. From our experiments, we noticed that it not only provides the best results but also allows the IC system to have a reasonable testing time.

TABLE 4-1 EVALUATION SCORES (%) AND TESTING TIME (SECONDS) ON DIFFERENT DATASETS.

Dataset	Method	β	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	Testing Time (seconds)
UAV	Baseline	0	64.47	53.71	44.11	35.33	32.91	67.83	374.47	50.22
	Post-generation first order	0.65	66.85	56.08	46.18	37.13	32.45	67.03	373.12	77.95
	Post-generation second order	0.45	66.64	55.85	46.13	37.26	32.64	67.43	373.60	91.11
	In-generation first order	0.25	67.01	56.63	47.04	38.23	**33.24**	69.07	387.59	64.51
	In-generation second order	0.25	**67.66**	**57.40**	**47.97**	**39.27**	32.99	**69.20**	**388.42**	76.40
Sydney	Baseline	0	77.22	67.70	59.86	53.25	37.31	68.87	233.83	3.26
	Post-generation first order	0.35	77.23	68.08	60.55	54.24	37.29	68.68	235.57	16.12
	Post-generation second order	0.55	78.31	69.09	61.69	55.23	37.37	69.07	254.11	20.57
	In-generation first order	0.3	77.47	68.97	62.33	56.45	39.26	70.07	247.43	17.13
	In-generation second order	0.3	**78.37**	**69.85**	**63.22**	**57.17**	**39.49**	**71.06**	**255.53**	21.30
UCM	Baseline	0	77.37	70.22	64.56	59.71	40.42	73.54	298.22	11.27
	Post-generation first order	0.55	78.08	71.11	65.43	60.50	40.26	73.51	300.44	85.48
	Post-generation second order	0.45	78.18	71.17	65.56	60.78	40.20	73.68	303.66	150
	In-generation first order	0.25	79.13	72.15	66.72	62.05	40.96	74.03	307.89	89.49
	In-generation second order	0.2	**79.73**	**72.98**	**67.44**	**62.62**	**40.80**	**74.06**	**309.64**	111
RSICD	Baseline	0	**63.30**	**46.04**	35.62	28.56	**25.42**	47.29	75.21	116.08
	Post-generation first order	0.2	62.91	45.97	35.64	28.63	25.36	**47.34**	**75.61**	706.29
	Post-generation second order	0.2	62.90	45.99	**35.68**	**28.68**	25.30	47.34	75.56	893.20
	In-generation first order	0.05	62.89	45.92	35.59	28.56	25.33	47.28	74.81	720.34
	In-generation second order	0.05	62.92	45.97	35.61	28.57	25.29	47.28	74.47	1063.86

1. **GT**: Asphalt with shadow and three cars.
2. **Baseline:** There are two cars on the asphalt and five cars on the right.
3. **Post-generation order 1:** There are three cars on the asphalt.
4. **Post-generation order 2:** There are two cars on the right.
5. **In-generation order 1:** There are two cars on the asphalt and shadow on the bottom right.
6. **In-generation order 2:** There are two cars on the asphalt and shadow on the bottom right.



1. **GT:** There is white roof.
2. **Baseline:** There is red roof.
3. **Post-generation order 1** There is white roof.
4. **Post-generation order 2:** There is white roof.
5. **In-generation order 1:** There is white roof.
6. **In-generation order 2:** There is white roof.

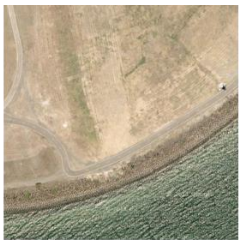
1. **GT:** Parking lot with shadow and three cars on the top right.
2. **Baseline:** Parking lot with three cars on the left.
3. **Post-generation order 1** Parking lot with three cars on the right.
4. **Post-generation order 2:** Parking lot with three cars on the right.
5. **In-generation order 1:** There are three cars on the asphalt.
6. **In-generation order 2:** There are three cars on the asphalt.

1. **GT:** There is low vegetation.
2. **Baseline:** There are several trees.
3. **Post-generation order 1** There is low vegetation.
4. **Post-generation order 2:** There is low vegetation.
5. **In-generation order 1:** There are several trees.
6. **In-generation order 2:** There are several trees.

a)

1. **GT**: There are some green bushes and paths on the meadow while some houses beside.
2. **Baseline:** There are some white bushes on the meadow with some curved paths on it.
3. **Post-generation order 1:** There are some green bushes on the meadow with some curved paths on it.
4. **Post-generation order 2:** There are some green bushes on the meadow with some curved paths on it.
5. **In-generation order 2:** There are some green bushes on the meadow with some curved paths on it.
6. **In-generation order 2:** There are some green bushes on the meadow with some curved paths on it.

1. **GT:** An industrial area with many white buildings densely arranged while a lawn beside.
2. **Baseline:** A small river with many white buildings on it while a lawn beside.
3. **Post-generation order 1** An industrial area with many white buildings on it while a lawn beside.
4. **Post-generation order 2:** A residential area with many white buildings densely arranged while a lawn beside
5. **In-generation order 1:** An industrial area with many white buildings and a lawn beside.
6. **In-generation order 2:** An industrial area with many white buildings and a lawn beside.



1. **GT:** A narrow runway on the river bank.
2. **Baseline:** There are some white lines on the wide river with a river beside.
3. **Post-generation order 1:** There are some mark lines on the wide river with a river beside.
4. **Post-generation order 2:** There are some marking lines on the wide river with some lawns beside.
5. **In-generation order 1:** A part of deep green sparkling sea with a lawn beside.
6. **In-generation order 2:** A part of deep green sparkling sea with a lawn beside.

1. **GT:** A residential are with many houses arranged densely and some roads go through this area.
2. **Baseline:** Lots of houses with different colours of roofs arranged neatly.
3. **Post-generation order 1:** Lots of houses with different colours of roofs arranged neatly.
4. **Post-generation order 2:** Lots of houses with different colours of roofs arranged neatly.
5. **In-generation order 1:** A residential area with houses arranged neatly and divided into rectangles by some roads.
6. **In-generation order 2:** A residential area with houses arranged neatly and some roads go through this area.

b)

Figure 4.4 Examples of the generated descriptions of the baseline methods and the proposed post-processing strategies from the a) UAV dataset and b) Sydney dataset. In red are highlighted the heavy mistakes present in the generated description while in blue their corrections by the proposed post-generation strategies. In green are highlighted light mistakes in the generated captions.

1. **GT**: There is a white airplane in the airport.
2. **Baseline:** There is one airplanes in the airport.
3. **Post-generation order 1:** There is an airplane in the airport.
4. **Post-generation order 2:** There is a piece of the airport.
5. **In-generation order 2:** There is a white airplane in the airport.
6. **In-generation order 2:** There is a white airplane in the airport.

1. **GT:** A tennis court is surrounded by some trees and a road beside.
2. **Baseline:** A tennis tennis is surrounded by some plants and a road beside.
3. **Post-generation order 1** A small tennis courts surrounded by some plants.
4. **Post-generation order 2:** A small tennis court surrounded by some plants and a road beside.
5. **In-generation order 1:** A tennis court is surrounded by some plants and a road beside.
6. **In-generation order 2:** A tennis court is surrounded by some plants and a road beside.

1. **GT:** Two white straight freeways closed together with some cars on them.
2. **Baseline:** Some cars are on the freeways and some cars on the roads.
3. **Post-generation order 1:** Some cars on the roads.
4. **Post-generation order 2:** Some cars on the roads.
5. **In-generation order 1:** There are two straight freeways closed to each other with some cars on the roads.
6. **In-generation order 2:** There are two straight freeways closed to each other with some cars on the roads.

1. **GT:** Lots of boats docked at the harbor and the water is deep blue.
2. **Baseline:** Lots of boats docked neatly at the harbor and the boats are closed to each other.
3. **Post-generation order 1:** Lots of boats docked neatly at the harbor and the boats are closed to each other.
4. **Post-generation order 2:** Lots of boats docked neatly at the harbor and the boats are closed to each other.
5. **In-generation order 1:** Lots of boats docked at the harbor and the water is deep blue.
6. **In-generation order 2:** Lots of boats docked at the harbor and the water is deep blue.

a)

1. **GT**: A playground with two basketball fields in it is surrounded by some green trees and buildings.
2. **Baseline:** A playground with basketball fields next to it is surrounded by many green trees and many buildings.
3. **Post-generation order 1:** A playground with basketball fields next to it is surrounded by many green trees and many buildings.
4. **Post-generation order 2:** A playground with basketball fields next to it is surrounded by many green trees and many buildings.
5. **In-generation order 2:** A playground with two basketball fields next to it is surrounded by many green trees and many buildings.
6. **In-generation order 2:** A playground with two basketball fields next to it is surrounded by many green trees and many buildings.

1. **GT:** Some snows cover part of a piece of irregular green mountains.
2. **Baseline:** Some white snows cover the mountain.
3. **Post-generation order 1** Some white snows cover the mountain.
4. **Post-generation order 2:** Some white snows cover a mountain.
5. **In-generation order 1:** Some white snows cover the mountain.
6. **In-generation order 2:** Some white snows cover a piece of green mountain.

1. **GT:** A large number of trees are planted on both sides of the river.
2. **Baseline:** Many green trees are in two sides of a curved river.
3. **Post-generation order 1:** Many green trees are in two sides of a curved river.
4. **Post-generation order 2:** Many green trees are in two sides of a curved river.
5. **In-generation order 1:** Many green trees are in two sides of a curved river.
6. **In-generation order 2:** Many green trees are in two sides of a curved river with a bridge over it.

1. **GT:** Some white ships are parked in the harbor .
2. **Baseline:** Many boats are in a port near some buildings.
3. **Post-generation order 1** Many boats are in a port near some buildings.
4. **Post-generation order 2:** Many boats are in a port near some buildings.
5. **In-generation order 1:** Many boats are in a port near a wharf.
6. **In-generation order 2:** Many boats are in a port near a wharf.

b)

Figure 4.5 Examples of the generated descriptions of the baseline methods and the proposed post-processing strategies from the a) UAV dataset and b) Sydney dataset. In red are highlighted the heavy mistakes present in the generated description while in blue their corrections by the proposed post-generation strategies. In green are highlighted some light mistakes in the generated captions.

4.3.2 Quantitative Results

Table 4-1 reports the quantitative results on different datasets of the baseline method (no post-processing strategy) and the results of the two post-processing strategies, post-generation and in-generation of the first and second order. We can see that the two proposed post-processing strategies improve the performance of the baseline model with a good margin, in almost all the metrics and datasets. The only exception is the RSICD dataset. In this dataset the obtained quantitative results of the proposed post-processing strategies remain the same as the one achieved with the baseline model. We noticed that the optimal path of the post-processing strategies generally is similar to the output of the baseline system. This is due to the fact that RSICD dataset contains many similar sentences for different scenarios. Furthermore, in this dataset 48% of the images are only describe by an unique sentence and only 7% of the images are described with 5 different descriptions [24]. In general, the in-generation strategy achieves the highest results. The rectification of the errors in this strategy is done at each time step. Once a potential error is corrected, this affects the future time steps leading to better sentences. On the other hand, rectification in the post-generation strategy is done once the sentence is fully generated by the baseline system. This means that the strategy corrects the potential mistakes while maintaining the form of the starting sentence. This is clearly reflected in the β values. We can see that the β values of the post-generation strategy are in general higher compared to the in-generation strategy. This means that is harder and more onerous (in terms of time) to rectify a sentence once it is fully generated compared to a step by step correction. The best results of each post-processing strategy are reached when extending the memory from first to second order. However, this improvement is marginal compared to the first-order memory and it is more onerous in terms of time.

4.3.3 Qualitative Results

Fig. 4.3. and 4.4. show four examples of test images from all the considered datasets and their generated captions. Fig. 4.3. a), and b) depicts four examples of the UAV, and Sydney datasets while Fig. 4.4. a) and b) depicts four examples of the UCM and RSICD datasets, respectively. In the first three examples of the UAV dataset (Fig.4.3 a), we can notice that the generated descriptions of the baseline model are affected by several errors related to the objects and their attributes. It is worth noticing that the UAV dataset is rich of attributes related to the color, size, direction and the location of the objects. The post-processing strategies are able at detecting and correcting most of these errors present in the generated captions of the baseline method related to the color attribute (Fig.4.3 a) second column) and relative positions (Fig.4.3 b) third column). The in-generation strategy has the capability to completely change the generated description of the baseline method as the rectification is done sequentially at each time step (see third column of Fig. 4.3. a)), while the post-generation strategy changes the only error present in the generated description of the baseline model related to the position of the cars. In the rightest column of Fig. 4.3 a), we can notice that the baseline model is able at generating a coherent description which is not disrupted by the two proposed post-processing strategies. In general, passing from first to second order does not change significantly the rectification of the sentences. This is also reflected on the quantitative results present in Table 4-I.

TABLE 4-2 COMPARISON WITH THE STATE OF THE ART METHODS ON DIFFERENT DATASETS.

Dataset	Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr-D
UAV	Merge GRU-D [93]	66.03	55.08	44.87	35.37	31.31	66.76	368.36
	SVM-D BOW	**68.84**	**58.05**	**48.33**	39.22	32.81	**69.63**	**391.31**
	SVM-D CONC	65.13	56.53	48.15	**39.69**	32.17	69.31	389.45
	Proposed post-processing strategy	67.66	57.40	47.97	39.27	**32.99**	69.20	388.42
Sydney	VLAD+RNN [24]	56.58	45.14	38.07	32.79	26.72	52.71	93.72
	VLAD +LSTM [24]	49.13	34.12	27.60	23.14	19.30	42.01	91.64
	mRNN [22]	51.30	37.50	20.40	19.30	18.50	-	161.00
	mLSTM [22]	54.60	39.50	22.30	21.20	20.50	-	186.00
	mGRU [22]	69.64	60.92	52.39	44.21	31.12	59.17	171.55
	mGRU embedword [22]	68.85	60.03	51.81	44.29	30.36	57.47	168.94
	Merge GRU-D [93]	73.07	63.37	56.41	49.87	33.09	63.34	193.93
	CSMLF [21]	59.98	45.83	38.69	34.33	24.75	50.18	75.55
	ConvCap [94]	74.72	65.12	57.25	50.12	34.76	66.74	214.84
	Soft-attention [24]	73.22	66.74	62.23	**58.20**	39.42	71.27	249.93
	Hard-attention [24]	75.91	66.10	58.89	52.58	38.98	**71.89**	218.19
	SAA [28]	68.82	60.73	52.94	45.39	30.49	58.20	170.52
	SD-RSIC [95]	72.4	62.1	53.2	45.1	34.2	63.6	139.5
	SVM-D BOW	77.87	68.35	60.23	53.05	37.97	69.92	227.22
	SVM-D CONC	75.47	67.11	59.70	53.08	36.43	67.46	222.22
	Proposed Post-Processing strategy	**78.37**	**69.85**	**63.22**	57.17	**39.49**	71.06	**255.53**
UCM	VLAD+RNN [24]	63.11	51.93	46.06	42.09	29.71	58.78	200.66
	VLAD +LSTM [24]	70.16	60.85	54.96	50.30	34.64	65.20	231.31
	mRNN [22]	60.10	50.70	32.80	20.80	19.30	-	214.00
	mLSTM [22]	63.50	53.20	37.50	21.30	20.30	-	222.50
	mGRU [22]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
	mGRU embedword [22]	75.74	69.83	64.51	59.98	36.85	66.74	279.24
	Merge GRU-D [93]	75.74	67.16	60.63	55.29	37.81	69.11	274.85
	CSMLF [21]	36.71	14.85	7.63	5.05	9.44	29.86	13.51
	ConvCap [94]	70.34	56.47	46.24	38.57	28.31	59.62	190.15
	Soft-attention [24]	74.54	65.45	58.55	52.50	38.86	72.37	261.24
	Hard-attention [24]	**81.57**	73.12	67.02	61.82	**42.63**	76.98	299.47
	SAA [28]	79.62	**74.01**	**69.09**	**64.77**	38.59	69.42	294.51
	SD-RSIC [95]	74.8	66.4	59.8	53.8	39.0	69.5	213.2
	RTRMN (semantic) [29]	55.26	45.15	39.62	35.87	25.98	55.38	180.25
	RTRMN (statistical) [29]	80.28	73.22	68.21	63.93	42.58	**77.26**	**312.70**
	SVM-D BOW	76.35	66.64	58.69	51.95	36.54	68.01	271.42
	SVM-D CONC	76.53	69.47	64.17	59.42	37.02	68.77	292.28
	Proposed Post-Processing strategy	79.73	72.98	67.44	62.62	40.80	74.06	309.64
RSICD	VLAD+RNN [24]	49.38	30.91	22.09	16.77	19.96	42.42	103.92
	VLAD +LSTM [24]	50.04	31.95	23.19	17.78	20.46	43.34	118.01
	mRNN [22]	45.58	28.25	18.09	12.13	15.69	31.26	19.15
	mLSTM [22]	50.57	32.42	23.19	17.46	17.84	35.02	31.61
	mGRU [22]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
	mGRU embedword [22]	60.94	46.24	36.80	29.81	26.14	48.20	159.54
	Merge GRU-D [93]	60.30	42.48	32.03	25.20	22.94	4383	65.90
	CSMLF [21]	51.06	29.11	19.03	13.52	16.93	37.89	33.88
	ConvCap [94]	63.36	51.03	41.74	34.52	**33.25**	57.70	166.48
	Soft-attention [24]	67.53	**53.08**	43.33	36.17	32.55	**61.09**	**196.43**
	Hard-attention [24]	66.69	51.82	41.64	34.07	32.01	60.84	179.25
	SAA [28]	**67.60**	44.33	**44.33**	**36.45**	31.09	55.36	193.96
	SD-RSIC [95]	64.5	47.1	36.4	29.4	24.9	51.9	77.5
	RTRMN (semantic) [29]	62.01	46.23	36.44	29.71	28.29	55.39	151.46
	RTRMN (statistical) [29]	61.02	45.14	35.35	28.59	27.51	54.52	148.20
	SVM-D BOW	61.12	42.77	31.53	24.11	23.03	45.88	68.25
	SVM-D CONC	59.99	43.47	33.55	26.89	22.99	45.57	68.54
	Proposed Post-Processing strategy	62.90	45.99	35.68	28.68	25.30	47.34	75.56

In the Sydney dataset (see Fig.3 b) we can see that the trend is same as with the UAV dataset. The proposed post-processing techniques are able at rectifying the errors present in the captions generated by the baseline system. In particular, we can see that the proposed post-processing strategies are not only able to correct the errors related to the attributes (first column of Fig. 3.b)), but also more significant errors related to the objects as it can be seen in the second and third column of the of Fig. 3. b).

Fig. 4.4 a) and b) shows example of images and generated description from the UCM and RSICD datasets, respectively. Also here we can notice that the post-processing strategies are able at rectifying the generated sentences by correcting the errors and providing a more coherent description. In particular, in the examples taken from the UCM dataset we can see that the rectified sentences contain more attributes and are more complete compared to the baseline model. This can be seen in the first a second column of Fig. 4.4 a) in which the attributes related to the color of the airplane and the size of tennis court are included in the rectified version provided by the ingeneration post-processing strategy. The same happens also in the RSICD dataset (Fig. 4.4 b)). Even though in this dataset the quantitative results did not improve with respect of the baseline method, the strategies are able at correcting mistakes in the sentence. However, because the evaluation metrics measure the similarity between the generated description and the reference ones based on the word co-occurrence this is not reflected in the results of Table 4-I. In general, we can conclude that our post-processing strategies are helpful in rectifying the errors present in the generated sentences. The strategies seem to work best with small datasets that are more affected to overfitting.

4.3.4 Comparison with State of the Art methods

In this subsection we compare the integration of the post-processing strategies into the simple encoder-decoder IC system with some state of the art methods [22], [24], [28], [29], [93]–[95], [100]. In Table 4-2 we depict the comparison results in all the used dataset. From the results one can notice that the integration of the post-processing strategies in the encoder-decoder framework achieve very competitive results with the state of the arts methods, with the exception of RSICD. In particular, we can see that for UAV and UCM datasets, the proposed post-processing strategy is able to compete not only with simple encoder-decoder frameworks but also with very sophisticated method that use attention mechanism. We can see that for the Sydney dataset, the achieved results are the best. Regarding the RSICD dataset, we can achieve moderate results.

4.4. Final Remarks

In this chapter, we have presented two post-processing strategies to improve the sentence generation of an IC system. The postprocessing strategies based on the combination of HMMs and Viterbi algorithm. They are able to rectify errors present in the generated sentences of an IC system and as a consequence improve the outcome of the IC system. The post-processing strategies are applied at test time, once an IC system is fully trained. In particular, we propose the post-generation processing strategy that corrects the error present in a generated sentence, once this is fully generated, and the in-generation processing strategy that is able to detect and correct potential errors during the sentence generation process. While the first strategy maintains the same form as the originally generated sentence the second strategy is able at altering the output of the IC system at each time step, influencing the next ones. This allows to not only provide better sentences but also different compared to the original ones. The experiments carried out on four different RS captioning datasets confirm the effectiveness of the proposed post-processing strategies in rectifying the outcome of an IC system. The post-processing strategies tested on a simple encoder-decoder IC system not only are able to improve the baseline method but also produce competitive and sometimes better results than sophisticated state of the art methods. As future work, we aim to combine the postprocessing strategy with attention mechanism to correct the mismatch between the words and the related image parts.

Chapter 5

5. Change Captioning

In this chapter, we present a new paradigm to analyse multitemporal RS data. In particular, we focus on change detection (CD). Most CD systems, provide an output that is either a binary change map that highlights the changed regions or a semantic change map that indicates the type of change for each pixel. In this chapter, we propose to describe the changes in the multitemporal images. The benefits of such a system are multiple. For instance, important high-level semantic information such as the attributes and the relationships of the changed areas are omitted by the change maps. Through a change caption system, we are able to include this high-level semantic information in the sentence description. Furthermore, the interpretation of change maps might be challenging for end-users who do not have expertise in the topic. User-friendly information about the changed areas would widen the impact of CD applications in the community. Sentence change description are much easy to interpret for end-users.

5.1. Introduction and Literature Review

Observing Earth's surface evolution is one of the main purposes of remote sensing (RS). Through multi-temporal images acquired by sensors on-board satellites or aerial platforms, we can continually observe and track changes that occur around our globe. As a consequence, change detection (CD) is among the most important applications in RS. The task of a CD system is to automatically detect changes over time in a given geographical area by analysing two or more co-registered images [101], [102]. Singh et al. defined CD as *"the process of identifying differences in the state of an object or phenomenon by observing it at different times"* [101]. Accordingly, the CD is indispensable in a wide range of RS applications based on land cover and land use analysis such as urban planning, environmental change assessment, disasters monitoring, deforestation and agricultural investigations [101].

In the RS literature, several approaches have been proposed for change detection. These approaches can be grouped into two main broad categories: 1) unsupervised and 2) supervised approaches. Unsupervised approaches include CD methods that do not rely on prior information provided by reference data. In contrast, supervised approaches need the a priori information of reference data as most of them are based on the use of supervised classifiers. The majority of the unsupervised approaches in the literature are based on the so-called "difference image" (DI) generation and analysis [103]–[110]. DI can be generated by applying pixel by pixel subtraction [103] or ratio operator [105] in the co-registered input images where the assumption is that the pixels associated with the land cover changes have different values compared to the ones associated with the unchanged areas. Threshold or clustering methods can be used to analyse the DI and produce the final change map [102]–[104]. In general, unsupervised CD approaches are appealing because they do not need ground truth information. However, defining the optimal threshold to produce the final change map is not an easy task and its effectiveness strongly depends on the statistical characteristics of the difference image [104], [106]. Furthermore, the outcome change map is, in most cases, binary (change/no change) [102]–[110]. Even though there are specific cases in which different kinds of changes can be detected, the exact land-cover transition associated with the change cannot be explicitly identified since the ground truth information is missing [102].

Supervised approaches assume that the ground-truth information is available for all the multi-temporal images. This information is injected into machine learning algorithms to learn automated decision models to distinguish between changed and non-changed areas and, possibly the type of change in the multi-temporal images [6], [111]–[119]. This is achieved thanks to the use of supervised classifiers such as support vector machines (SVM) [82], [83], [117], decision trees and random forest [120]. Post-

classification comparison is a traditional supervised technique that independently classifies the multi-temporal images and then compares the individual classification maps to define the changed areas and the exact land-cover transitions associated with the change [111], [112], [114], [117]. The advantage of this technique is that it is straightforward. However, post-classification comparison techniques suffer the propagation of the classification error which leads to an error accumulation in the final change map [101]. To alleviate the classification error accumulation present in post-classification comparison techniques, compound classification simultaneously analyses the multi-temporal data by taking into account temporal contextual information. Markov and conditional random fields have shown to be very effective in incorporating such information and improving the change detection results in terms of accuracy and reliability [6], [113], [115], [118]. Although supervised change detection approaches are very effective in analysing the multi-temporal images and detecting the changes present, they assume that the ground truth information is available for each of the multi-temporal images. Such an assumption is not always possible in real scenarios as collecting ground-truth information is time-consuming and costly. To address this, semi-supervised change detection approaches assume that the labelled information is available for at least one of the multi-temporal images and the learning paradigm of these approaches benefits both labelled and unlabeled samples making them more suitable in real-life scenarios [121], [122].

One common limitation of the aforementioned conventional change detection approaches is the use of hand-crafted features. These features are tedious to obtain and unable to capture high-level feature representation for multi-temporal images. Recently, deep neural networks have shown to be very effective in automatically learning discriminative and representative features directly from raw images. Improvement in the quality and quantity of RS data along with more accessible computational resources have oriented the researchers in the RS community toward the exploitation of deep learning (DL) algorithms to improve the accuracy of CD systems [123]–[126]. The DL algorithms exploited for CD includes restricted Boltzmann machines [127], autoencoders [128], convolutional neural networks [129]–[134] and recurrent neural networks (RNNs) [135], [136]. A comprehensive and extensive review of these strategies can be found in [126].

Despite the efforts made in developing reliable and accurate CD systems, the obtained output is either a binary change map that highlights the changed regions, or a semantic change map that indicates the type of change for each pixel. Important high-level semantic information such as the attributes and the relationships of the changed areas are omitted by the change maps. This information needs to be inferred and interpreted directly by end-users. It is worth noting that the interpretation of change maps might be challenging for end-users who do not have expertise in the topic. User-friendly information about the changed areas would widen the impact of CD applications in the community. To address this issue, in this article, we present a CD system that automatically describes the changes in bi-temporal images with sentence descriptions (i.e., captions). In doing so we are inspired by the recent advancements in IC in the RS community [9]. As we saw in the previous chapters, compared to traditional RS image analysis techniques, like classification or object recognition, IC not only provides a more user-friendly representation but also a richer one including relationships and attributes of the detected objects. In this work, we aim at expanding the remote sensing image captioning (RSIC) systems from single-date images to bi-temporal images where the goal is transformed from describing the image content to describing the changes that have occurred between the two acquisitions. We named the system change captioning (CC).

The proposed CC system takes as input bi-temporal images instead of single-date images and describes the possible changes. As shown in Figure 5.1, it is based on an encoder-decoder framework. The encoder is composed of three main blocks: 1) a spectral dimensionality reduction block that aims at reducing the spectral dimension of the images, 2) a fusion block that combines the multi-temporal information, and 3) a pre-trained CNN to extract discriminative features from the multi-temporal images. We propose two main fusion strategies: early fusion which is applied at the image level (before the CNN) and a late fusion which is applied at the feature level (after the CNN) as can be seen in Figure. 5.1.a) and Figure. 1.b), respectively.

In this chapter, we present a new paradigm to analyse multitemporal RS data. In particular, we focus on change detection (CD). Most CD systems, provide an output that is either a binary change map that highlights the changed regions or a semantic change map that indicates the type of change for each pixel. In this chapter, we propose to describe the changes in the multitemporal images. The benefits of such a system are multiple. For instance, important high-level semantic information such as the attributes and the relationships of the changed areas are omitted by the change maps. Through a change caption system, we are able to include this high-level semantic information in the sentence description. Furthermore, the interpretation of change maps might be challenging for end-users who do not have expertise in the topic. User-friendly information about the changed areas would widen the impact of CD applications in the community. Sentence change description are much easy to interpret for end-users.

5.2. Introduction and Literature Review

Observing Earth's surface evolution is one of the main purposes of remote sensing (RS). Through multi-temporal images acquired by sensors on-board satellites or aerial platforms, we can continually observe and track changes that occur around our globe. As a consequence, change detection (CD) is among the most important applications in RS. The task of a CD system is to automatically detect changes over time in a given geographical area by analysing two or more co-registered images [101], [102]. Singh et al. defined CD as *"the process of identifying differences in the state of an object or phenomenon by observing it at different times"* [101]. Accordingly, the CD is indispensable in a wide range of RS applications based on land cover and land use analysis such as urban planning, environmental change assessment, disasters monitoring, deforestation and agricultural investigations [101].

In the RS literature, several approaches have been proposed for change detection. These approaches can be grouped into two main broad categories: 1) unsupervised and 2) supervised approaches. Unsupervised approaches include CD methods that do not rely on prior information provided by reference data. In contrast, supervised approaches need the a priori information of reference data as most of them are based on the use of supervised classifiers. The majority of the unsupervised approaches in the literature are based on the so-called "difference image" (DI) generation and analysis [103]–[110]. DI can be generated by applying pixel by pixel subtraction [103] or ratio operator [105] in the co-registered input images where the assumption is that the pixels associated with the land cover changes have different values compared to the ones associated with the unchanged areas. Threshold or clustering methods can be used to analyse the DI and produce the final change map [102]–[104]. In general, unsupervised CD approaches are appealing because they do not need ground truth information. However, defining the optimal threshold to produce the final change map is not an easy task and its effectiveness strongly depends on the statistical characteristics of the difference image [104], [106]. Furthermore, the outcome change map is, in most cases, binary (change/no change) [102]–[110]. Even though there are specific cases in which different kinds of changes can be detected, the exact land-cover transition associated with the change cannot be explicitly identified since the ground truth information is missing [102].

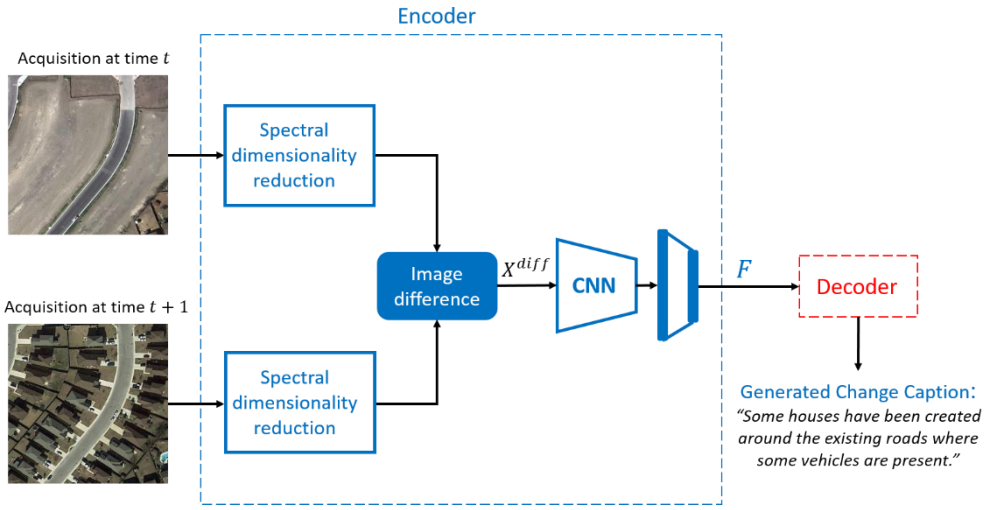
Supervised approaches assume that the ground-truth information is available for all the multi-temporal images. This information is injected into machine learning algorithms to learn automated decision models to distinguish between changed and non-changed areas and, possibly the type of change in the multi-temporal images [6], [111]–[119]. This is achieved thanks to the use of supervised classifiers such as support vector machines (SVM) [82], [83], [117], decision trees and random forest [120]. Post-classification comparison is a traditional supervised technique that independently classifies the multi-temporal images and then compares the individual classification maps to define the changed areas and the exact land-cover transitions associated with the change [111], [112], [114], [117]. The advantage of this technique is that it is straightforward. However, post-classification comparison techniques suffer the propagation of the classification error which leads to an error accumulation in the final change map [101].

To alleviate the classification error accumulation present in post-classification comparison techniques, compound classification simultaneously analyses the multi-temporal data by taking into account temporal contextual information. Markov and conditional random fields have shown to be very effective in incorporating such information and improving the change detection results in terms of accuracy and reliability [6], [113], [115], [118]. Although supervised change detection approaches are very effective in analysing the multi-temporal images and detecting the changes present, they assume that the ground truth information is available for each of the multi-temporal images. Such an assumption is not always possible in real scenarios as collecting ground-truth information is time-consuming and costly. To address this, semi-supervised change detection approaches assume that the labelled information is available for at least one of the multi-temporal images and the learning paradigm of these approaches benefits both labelled and unlabeled samples making them more suitable in real-life scenarios [121], [122].

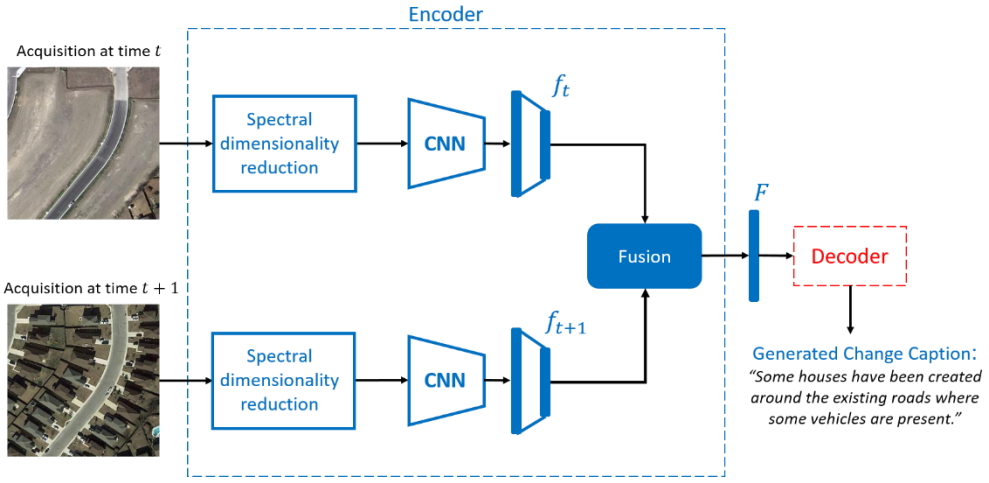
One common limitation of the aforementioned conventional change detection approaches is the use of hand-crafted features. These features are tedious to obtain and unable to capture high-level feature representation for multi-temporal images. Recently, deep neural networks have shown to be very effective in automatically learning discriminative and representative features directly from raw images. Improvement in the quality and quantity of RS data along with more accessible computational resources have oriented the researchers in the RS community toward the exploitation of deep learning (DL) algorithms to improve the accuracy of CD systems [123]–[126]. The DL algorithms exploited for CD includes restricted Boltzmann machines [127], autoencoders [128], convolutional neural networks [129]–[134] and recurrent neural networks (RNNs) [135], [136]. A comprehensive and extensive review of these strategies can be found in [126].

Despite the efforts made in developing reliable and accurate CD systems, the obtained output is either a binary change map that highlights the changed regions, or a semantic change map that indicates the type of change for each pixel. Important high-level semantic information such as the attributes and the relationships of the changed areas are omitted by the change maps. This information needs to be inferred and interpreted directly by end-users. It is worth noting that the interpretation of change maps might be challenging for end-users who do not have expertise in the topic. User-friendly information about the changed areas would widen the impact of CD applications in the community. To address this issue, in this article, we present a CD system that automatically describes the changes in bi-temporal images with sentence descriptions (i.e., captions). In doing so we are inspired by the recent advancements in IC in the RS community [9]. As we saw in the previous chapters, compared to traditional RS image analysis techniques, like classification or object recognition, IC not only provides a more user-friendly representation but also a richer one including relationships and attributes of the detected objects. In this work, we aim at expanding the remote sensing image captioning (RSIC) systems from single-date images to bi-temporal images where the goal is transformed from describing the image content to describing the changes that have occurred between the two acquisitions. We named the system change captioning (CC).

The proposed CC system takes as input bi-temporal images instead of single-date images and describes the possible changes. As shown in Figure 5.1, it is based on an encoder-decoder framework. The encoder is composed of three main blocks: 1) a spectral dimensionality reduction block that aims at reducing the spectral dimension of the images, 2) a fusion block that combines the multi-temporal information, and 3) a pre-trained CNN to extract discriminative features from the multi-temporal images. We propose two main fusion strategies: early fusion which is applied at the image level (before the CNN) and a late fusion which is applied at the feature level (after the CNN) as can be seen in Figure. 5.1.a) and Figure. 1.b), respectively. Once the encoding process is done, the obtained bi-temporal image representations are forwarded to the decoder to generate coherent change descriptions. In particular, we implement two different decoders: the



a)



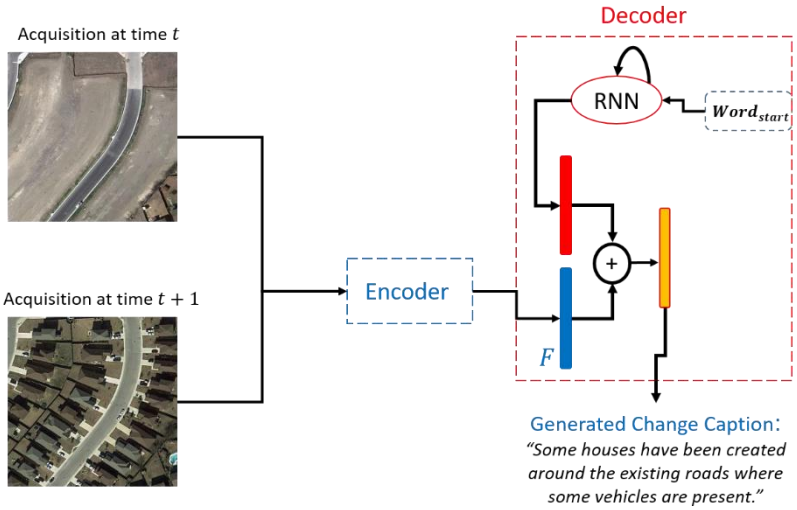
b)

Figure 5.1 Overview of the proposed change captioning system based on: a) image-based level fusion, and b) feature-based level fusion.

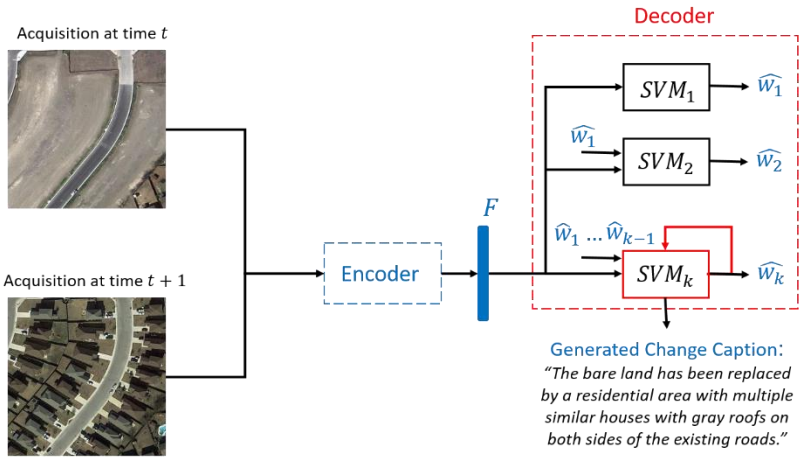
first one is based on the classical RNNs. The second one relies on a network of k multiclass support vector machines (SVMs), where the last one is rendered recurrent. Figure 5.2 shows the block schemes of the two proposed decoders. To test the proposed RSCC systems, in this work we have built two different datasets. The first one is based on VHR images while the second one is based on medium resolution satellite images. Both datasets contain 500 images and to each image, are assigned 5 different descriptions. The two datasets will be made publicly available to boost research in this new area.

Overall the main contribution of this chapter can be summarized as follows:

- We propose two baseline methods for RSCC based on the encoder-decoder framework. The difference between the baseline methods consists of the decoder part. A classical RNNs is developed in the first baseline method while a decoder based on the SVMs is proposed in the second baseline method.



a)



b)

Figure 5.2 Overview of the decoding phase based on a) RNNs, and b) on SVMs.

- Two datasets for RSCC are created. The first one consists of VHR images while the second one consists of medium resolution satellite images. Each dataset is composed of 500 bi-temporal images and to each bi-temporal image are assigned 5 different change descriptions.

We have briefly presented an RSCC system in [137]. The present article extends our previous work [137] by proposing novel RSCC methods and disclosing the two datasets

5.3. Proposed Change-Captioning Datasets

In this section we propose two novel datasets to caption the changes in the bi-temporal RS images. It is worth noting that these datasets are the first-ever made to tackle the problem of change detection through textual descriptions that carefully summarize the changes in a given geographical area between two different acquisitions. The datasets are different from each other. The first dataset is obtained by exploiting and annotating an existing dataset for RS change detection. The considered dataset is LEVIR and is originally used for detecting changes related to buildings [138]. The second dataset is entirely built in this work and consists of multispectral Landsat 7 images acquired over the region of Dubai. In the following subsection, we describe each dataset and the annotation process.

1. Many similar houses with gray roofs appear on the right instead of the bare land.
2. The bare land on the right has been replaced by many similar houses with gray roofs.
3. Many trees appear near the houses present on the left and no vehicle remains on the right road.
4. New roads and houses have been created, many trees appear near the houses and different vehicles are present.
5. Multiple similar house with gray roofs has been added on the sides of the existing road on the right of the residential area.

a)

1. The grass appears more green and the trees have increased in size.
2. There is no change except the color of the grass and the size of the trees.
3. Only the trees and the grass appear different.
4. Slight changes related to the color of the grass and the size of the trees.
5. There is no significant change to mention expect for the grass that have grown.

b)

Figure 5.3 Two examples from LEVIR CCD along with the reference descriptions. a) significant change and b) slight changes between the acquisitions.

5.3.1 LEVIR Change Captioning Dataset (LEVIR CCD)

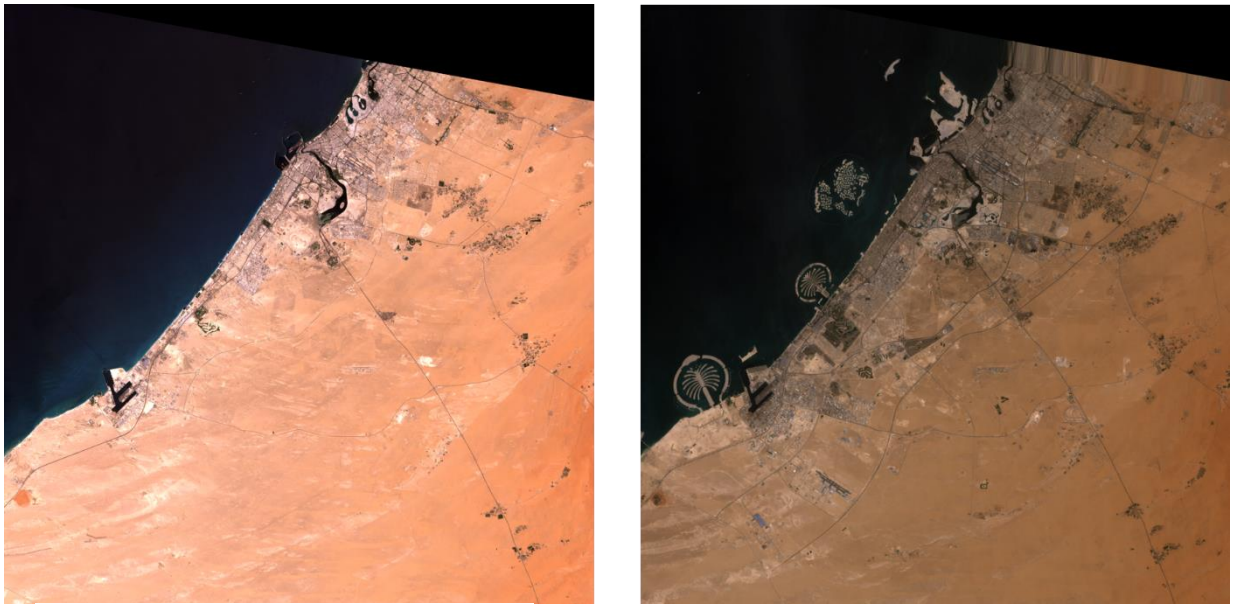
The LEVIR dataset is originally constructed for building change detection and consists of 637 VHR ($0.5\ m/pixel$) bi-temporal images of dimension $1024x1024$ acquired over 20 different regions over different cities in Texas US [138]. The collection of the image pair is obtained from Google Earth API with a period of 5~14 years. A full description of the original dataset for building CD and the terms of use are given by the authors in [138].

To make the dataset appropriate for RSCC, we have exploited and annotated it with sentence descriptions that describe the nature of changes that occurred between the two acquisitions. In particular, the annotation process is performed by annotators with RS backgrounds. First, a visual inspection of the scenes in the bi-temporal images is performed to identify the nature of the changes present. Then, the attributes (i.e., colour, size) and relationships between different changes are considered, in terms of position and relative position to other objects. After, the original image pairs are cropped into 256x256 pixels patches and finally, to each image pair 5 different change sentence descriptions are given by the annotators. The total number of image pairs is 500 leading to a total number of 2500 change descriptions.

From the visual inspection, we have identified 13 types of objects that undergo a change. These objects include sparse/dense vegetation areas bare land, residential area, houses, roofs, roads/streets, vehicles, constructions, trees, grass and swimming pools. The relationships between the objects and their positions and relative positions include mainly the directions such as: on the right/left, on the upper or bottom part, on the upper left/right and bottom left/right. The attributes of the objects include colour (i.e., red, grey, white), size (big/small) and their similarity (similar/different). Starting from the visual inspections the annotators are asked to formulate a coherent change sentence description including the attributes and the relationships of different objects. In the dataset, the minimum length of the sentence is 3 and it refers to a situation in which there is no change between the two acquisitions: “nothing has changed”. The maximum length of the sentence is 37 words while the average is 15.12 words. The total vocabulary size of the dataset is 279 unique words. Two examples from the dataset are depicted in Fig. 5.3. The terms of use for the dataset follow [138].

5.3.2 Dubai Change Captioning Dataset (Dubai CCD)

From 2000 to 2010 the area of Dubai has significantly changed in terms of urbanization. To understand and automatically describe the urbanization phenomena in Dubai, we decided to create the Dubai change captioning dataset (Dubai CCD). To this end, we needed images of the years 2000 and 2010. From our research, the only free available images that cover the area of Dubai in the time of interest are multispectral images acquired by Enhanced Thematic Mapper Plus (ETM+) sensor on-board Landsat 7 [139]. ETM+ acquires images in different bands with different spatial resolutions: 30 m visible and near-infrared bands; $60\,m$ thermal band; and $15\,m$ panchromatic band [139]. Fig. 4. shows the RGB representation of the images. The bi-temporal images are acquired on $19.05.2000$ and $16.06.2010$, respectively and have a dimension of $2569x2468$ pixels. To create the Dubai CDD we have considered only visible and infrared bands which have a spatial resolution of 30m. Considering the medium spatial resolution of the images here we extracted 500 tiles of sizes $50x50$ from the bi-temporal images. In order to properly identify and describe the nature of the changes in the small bi-temporal tiles, we also resorted to Google Maps to better inspect the area and publically available documents to understand the physical composition of the area. From the visual inspection we identified six different broad categories: 1) roads (i.e., crossroads, roundabout), 2) houses (i.e., neighborhood, residence areas) 3) buildings (any structure different from



a) b)

Figure 5.4 Images acquired over the region of Dubai from Landsat 7 on a) 19.05.2000 and b) 16.06.2010. For visualization purposes, the RGB combination is shown.

a) b)

Figure 5.5 Two examples from DUBAI CCD along with the reference descriptions. a) significant change and b) slight changes between the acquisitions.

house), 4) green areas, 5) lakes and 6) islands (i.e., archipelagos, peninsula). The annotators were asked to describe the changes on the bi-temporal images with a minimum of 3 words accounting for spatial distribution and the attributes of the changes. Each image was annotated by 5 different change descriptions yielding to 2500 descriptions, a maximum length of 23 words, an average length of 7.35 words and a vocabulary size of 246 unique words. Fig. 5.5. depicts two examples from the dataset along with the reference descriptions.

5.4. Proposed Change-Captioning System

Let $X_i = (I_t, I_{t+1})$ be a pair of RS images (or patches) acquired over the same geographical area at time t and $t+1$. Let us assume that we have M pairs of such images $X = \{X_i\}_{i=1}^{M}$ and each pair is associated with at least one textual description (sentence/caption) that describes the most salient changes that have occurred between the two acquisitions. Let $S_i = [w_1, w_2, \dots, w_L]$ be the sentence change description of the image pair X_i composed of L ordered words w. Inspired by the captioning systems developed for single images, our change-captioning systems are composed of two main steps: 1) deep bi-temporal image feature extraction and fusion (i.e., encoder) and 2) sentence generation (i.e., decoder). The first step aims to fuse and represent the bi-temporal images with discriminative features while the second one targets the translation of the features into a coherent textual description that describes the nature of the changes detected among the two acquisitions. The encoder is composed of two branches in the case of bi-temporal images (and more branches in the case of multi-temporal images). Each branch conveys three main blocks: 1) spectral dimension reduction to reduce the spectral dimension of the images, 2) a CNN to extract discriminative features, and 3) a fusion block to integrate the bi-temporal (or multi-temporal) information. The fusion operation can be applied either at the image level before the CNN or at the feature level after the CNN. To generate the sentences that describe the possible changes, we exploit both DL based methods based on RNNs, as well as the SVM-based decoder introduced in chapter 3. In particular, we use the multimodal RNN based on GRU [71] presented in Section 2.3.1. We present the details about the multimodal RNN in subsection 2.3.1. In the following we focus only on multi-temporal image representation.

5.4.1 Multi-temporal image representation and fusion

1) *Spectral dimensionality reduction:* This is the first block of the encoder. It is used to reduce the spectral dimensionality of multispectral images. In particular, we rely on the simple and popular principal component analysis (PCA) transformation [67] to reduce the original spectral dimension into three channels to fit the pre-trained CNN input requirement (see next subsection).

2) *Multi-temporal image representation:* the scope of feature generation, we rely on the powerful VGG16 CNN architecture pre-trained on ImageNet [85] to extract discriminative features [36]. The image features are obtained passing the input information from the bi-temporal images through the pre-trained VGG-16 (omitting the last fully connected layer) as follows:

$$f_i = VGG16(X_i). \tag{6.1}$$

3) *Multi-temporal image fusion:* Different from single image captioning, our scenario is composed of bi-temporal images. To encode the changes, we consider two different fusion strategies as shown in Fig. 5.1. In the first one, the fusion is done before the feature extraction process (early fusion). We apply channel-wise image difference $X_i^{diff} = (I_{t+1} - I_t)$ and afterwards we forward the difference image to the CNN to extract the image features. In the second one, fusion is performed once the individual features are extracted by the pre-trained CNN (late fusion). In this configuration we have considered two simple but very effective operators, feature concatenation $F = (f_t, f_{t+1})$ or element-wise feature subtraction $F = (f_{t+1} - f_{t-1})$.

5.5. Experiments

5.5.1 Experimental Set-Up

In this section, we evaluate the two RSCC systems on the proposed LEVIR and Dubai RSCC datasets. The split of the datasets is as follows: 60%, 10% and 30% for training, validation and testing respectively. We obtain the image features using VGG-16 pre-trained on ImageNet. VGG-16 produces a fixed feature vector of dimension 4096. In this work, we have considered two different configurations of the systems based on how image processing is performed. In the first configuration, we first obtain a difference image by applying the channel wise difference of the individual images composing the bi-temporal image and then apply the CNN to extract the features. In this configuration, the obtained feature vector is of 4096 dimensions. In the second configuration, the individual image features are extracted from each bi-temporal image and later feature fusion is applied. The fusion operator is either concatenation, doubling the size of the feature vector, or elementwise subtraction of the individual feature vectors obtaining a fused feature vector of dimension 4096. To the system based on the RNNs, the original features are projected through a projection layer (dense layer) with activation function ReLu and dimension 256. The same dimensions are fixed for the word embedding layer and the internal memory of the RNN. The output of the RNN and the image features are fused by applying element-wise addition. Adam optimizer [96] with a learning rate of 0.0001, cross-entropy loss function and a batch size of 128 are used to train the model. Dropout regularization is applied in different parts of the model to avoid overfitting. In particular, dropout with a rate of 0.5 is applied to the original image features, to the embedding layer, and to the output of the RNN. Note that this configuration yields the best results. We also tried to concatenate the image features and the words features forming the multimodal layer without applying any reduction and the results were poor.

The SVM decoder is composed of K multiclass classifiers. In our experiments, we adopted K=6, for the LEVIR CC dataset and K=4 for the Dubai CC dataset, respectively. This choice depends on the average length of the change sentence descriptions which is 15.12 and 7,25 for LEVIR CC and Dubai CC datasets, respectively. Because of the intrinsic properties of the SVM decoder, we need the number of SVM to be almost half of the average sentence length to model the word dependency while generating the change captions. This means that in total the system is composed of six and four SVM multiclass classifiers for the LEVIR CC dataset and Dubai CC dataset, respectively. The last one is rendered recurrent to model the dependency of the generated change caption on the previously predicted words. In particular, we rely on a linear SVM multiclass classifier that has only one free parameter, the value of the regularization C. The best values found on the validation set are C=0.001 for SVM_1 and C=0.01 for the rest of the SVMs. Each

SVM takes as input the concatenation between the bi-temporal image features and words features (if present). In contrast to the CC system based on the RNN, no reduction is applied to the image feature vector. This is well handled by the SVM that tolerates high-dimensional inputs. All image and words features are scaled to be in the range [0, 1] using a minimax scaler (see Equation 3.12).

5.6. Experimental Results on different Datasets

5.6.1 Quantitative Resutls

Table 5-1 and Table 5-2 report the quantitative results of the two proposed methods on LEVIR and Dubai CC Datasets, respectively. In particular, in the tables we can see the behaviour of the proposed methods according to the different combination of bi-temporal images. The achieved results by the two proposed

TABLE 5-1 EVALUATION SCORES (%) AND TIMES NEEDED FOR TRAINING/TESTING ON LEVIR CC DATASET.

Method	Fusion operator	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	Training Time (minutes)	Testing Time (seconds)
CC System based on RNN	Feature Concatenation	73.19	61.70	52.77	45.88	26.07	52.94	59.97	29.9	16.92
	Feature Subtraction	71.85	60.40	52.18	45.94	**27.43**	**54.13**	**71.64**	48.3	19.29
	Difference Image	68.20	54.95	45.64	38.79	25.34	49.86	60.27	48.75	16.89
CC System based on SVM	Feature Concatenation	70.89	60.19	51.59	44.80	24.19	50.60	54.87	9.88	2.36
	Feature Subtraction	**73.83**	**63.18**	**54.10**	**46.97**	25.09	50.48	51.42	36.4	**1.21**
	Difference Image	66.94	54.81	45.25	37.99	22.69	45.61	42.49	**2.58**	1.29

TABLE 5-2 EVALUATION SCORES (%) AND TIMES NEEDED FOR TRAINING/TESTING ON DUBAI CC DATASET.

Method	Fusion operator	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	Training Time (minutes)	Testing Time (seconds)
CC System based on RNN (RGB channels)	Feature Concatenation	64.41	49.39	38.24	28.03	25.85	50.65	68.85	12.4	8.47
	Feature Subtraction	67.19	52.23	40.00	28.54	25.51	51.78	69.73	16.57	8.23
	Difference Image	65.54	51.51	40.13	28.70	25.34	51.90	69.57	15.24	11.52
CC System based on RNN (PCA)	Feature Concatenation	**71.44**	56.80	45.36	34.38	**28.07**	**56.67**	80.73	10.25	5.57
	Feature Subtraction	70.71	**57.58**	**46.10**	**35.50**	27.60	56.61	**82.96**	13.69	5.53
	Difference Image	64.84	49.98	39.57	30.22	24.36	51.57	67.69	12	8.16
CC System based on SVM de-coder (RGB channels)	Feature Concatenation	65.78	51.25	40.34	30.74	24.94	50.66	69.68	5.88	0.76
	Feature Subtraction	66.50	50.36	38.02	27.79	23.78	50.07	64.44	15.75	0.38
	Difference Image	65.80	51.21	40.17	31.12	24.01	48.84	68.53	1.27	0.4
CC System based on SVM de-coder (PCA)	Feature Concatenation	66.34	51.89	41.34	31.78	25.62	52.12	75.70	5.09	0.75
	Feature Subtraction	68.99	54.18	43.32	33.46	26.47	51.46	72.40	11.51	0.39
	Difference Image	65.62	51.17	39.80	29.98	24.01	50.35	67.86	1.32	0.44

methods are similar to each other. The best combination of bi-temporal images is their individual feature subtraction while the worst one is the channel wise difference of the bi-temporal images. The Dubai CC dataset is composed of multispectral images. In particular, we have used only those spectral bands that are

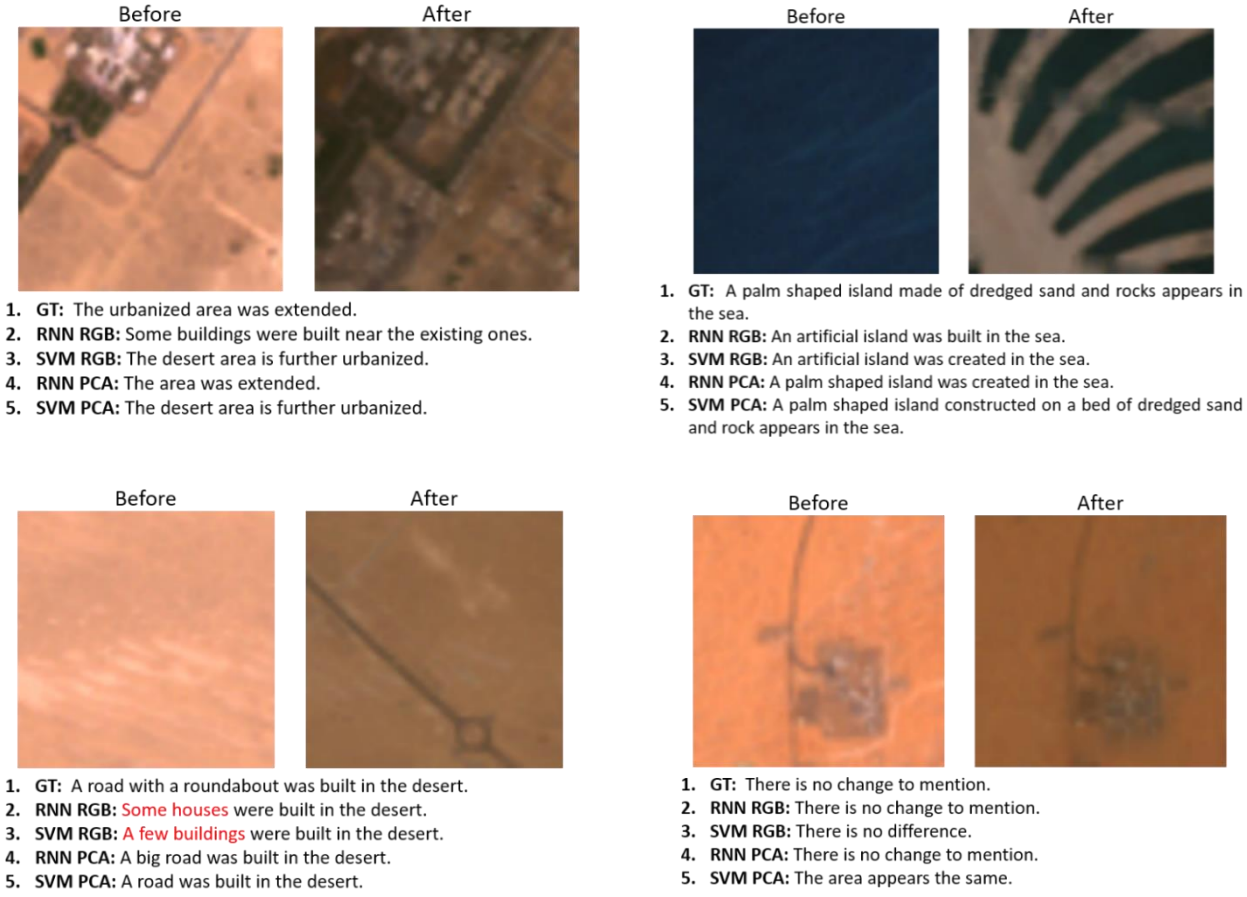


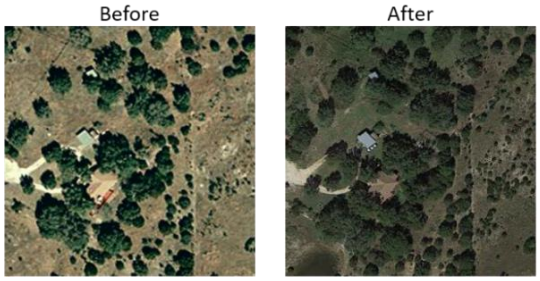
Figure 5.6 Change captioning examples of test images from the Dubai CC dataset. The first description corresponds to one of the ground-truth change descriptions, while the second and third are generated by the CC system based on RNN and SVM, respectively, when subtraction operator is applied to RGB channels. The fourth and fifth descriptions are generated by the CC system based on RNN and SVM, respectively, when the subtraction operator is applied to the three principal components of PCA. In red are highlighted the errors in the generated change captions.

characterized by a spatial resolution of 30 m (visible and infrared channels). Table 5-2 reports the results obtained using only the RGB channels and the combination of all the 6 available bands, to better exploit the spectral information. The latter is achieved through a spectral dimensionality reduction obtained through PCA. In Table II, we can see that both methods have an important gain in terms of accuracy when PCA is applied with respect to the use of only RGB channels. This confirms the importance of spectral information in RS applications. Last two columns of Tables I and II show the training and testing times of the proposed CC methods. One can notice that the CC system based on SVM decoder is in general much faster to train and to test compared to the CC system based on the RNN. In particular, the testing time to generate the change descriptions for the whole test set is 7 to 13 times faster on Levir CC dataset and to 14 to 25 times faster on Dubai CC dataset making it very suitable in real life scenarios.

5.6.2 Qualitative Results

Figure 5.6 shows four examples of the generated change captions from the proposed systems of bi-temporal images on Dubai CC. In particular, we report the qualitative results obtained by the two methods applied to the images considering the RGB channels only and to the three principal components of the PCA that are applied to all the 6 available bands. The fusion operator considered in these examples is the individual feature subtraction. From the images in Figure 5.6 we can notice that all the generated change descriptions generated by the two systems utilizing all the available spectral information are in line with the changes of the bi-temporal images. The generated change captions are more complete and coherent with the changes of bi-temporal images compared to the ones generated utilizing only the RGB bands. In particular, utilizing

1. **GT:** The bare land has been replaced by a residential area with multiple similar houses with gray roofs on both sides of the existing roads and a swimming pool appears.
2. **RNN FeatConc:** The low vegetation has been replaced by a residential area with multiple similar houses with gray roofs on both sides of the existing roads.
3. **RNN FeatSub:** Multiple similar houses with gray roofs appears on the sides of the existing roads
4. **SVM FeatConc:** The low vegetation has been replaced by a residential area with multiple similar houses with gray roofs on both sides of the existing roads.
5. **SVM FeatSub:** The similar houses with gray roofs on both sides of the existing roads.



1. **GT:** There is no significant change to mention.
2. **RNN FeatConc:** The color of the grass and the trees have increased in size.
3. **RNN FeatSub:** There is no significant change to mention.
4. **SVM FeatConc:** The trees have increased in size.
5. **SVM FeatSub:** The trees have increased in size.

Figure 5.7 Change captioning examples of test images from the LEVIR CC dataset. The first description corresponds to one of the ground-truth change descriptions, while the second and third are generated by the CC system based on RNN when feature concatenation and subtraction operator are applied, respectively. The fourth and fifth descriptions are generated by the CC system based on SVM when feature concatenation and subtraction operator are applied, respectively. In red are highlighted the errors in the generated change captions.

only the RGB bands some generated change descriptions are affected by errors as shown in the first image of the second raw in Figure 5.6.

Figure 5.7 shows two examples of bi-temporal images from the Levir CC dataset along with the generated change captions from the two proposed systems. In particular, are shown the generation change captions when the fusion operator is concatenation and subtraction. The generated captions by the two systems are in line with the visual changes between the bi-temporal images. We notice an error in the change generated description of the systems regarding the land cover transition when concatenation is applied as fusion operator (Figure 5.7, first image). In the 'before' image bare land is present instead of the low vegetation. However, the post-event changes are well-described by two methods.

5.7. Final Remarks

In this chapter, we present two different RSCC systems able at describing the changes from bi-temporal images with sentence descriptions. We showed that sentence descriptions not only include more high-level semantic information about the changes but also are more user-friendly and could help in widening the impact of CD applications for the community.

In absence of datasets, we constructed two different change captioning datasets. One is based on very high-resolution RGB RS images and the other one is based on multi-spectral images. The obtained experimental results show the promising capabilities of the proposed CC systems to generate complete and coherent change sentence descriptions that summarize the changes in bi-temporal images. In particular, we show that spectral information is very important in generating reliable change captions. The best strategy to fuse the image information seems to be late fusion where information is integrated at the feature level. From our results, it comes out that early fusion, which is applied at the image level as channel-wise difference, is not the best solution for RSCC. This might be explained by the stronger expressive power of high-level representations comparedlow-level ones. We also showed that the two decoders obtain very similar results in all the metrics. Furthermore, we show that the decoder based on SVMs is faster in terms of training and testing compared to that based on RNN. In the future, we aim at expanding the two proposed datasets with more images. Furthermore, we aim at developing more sophisticated CC methods based on attention mechanisms to better exploit the peculiarities of RS images such as scale and semantic ambiguities as well as spectral information.

Chapter 6

6. Conclusions and Future Directions

6.1. Conclusions

This thesis has presented distinct methodologies to describe the content of remote sensing images with sentence descriptions, namely remote sensing image captioning. We have studied and proposed different image captioning methods in the context of remote sensing image analysis. We have shown the advantages of image captioning to perform content based image retrieval. Specifically, we have shown that using generated sentences as a query not only allows to include high-level semantic information increasing the accuracy of the retrieval process but also renders it more comfortable for end-users.

We have proposed a novel decoder based on support vector machines which is particularly suitable for those situations in which only a small number of annotated samples is available. Furthermore, it is characterized by a very short training and testing and does not require expensive computational power making it very suitable for real-time applications.

To improve the quality of an IC system we have proposed two post-processing strategies that are able to rectify the generated sentences by detecting and correcting the potential errors. They are applied at test time and can be injected into any IC system. We have shown that by injecting the post-processing strategies to a simple encoder-decoder IC system, it can improve the quality of the generated sentences. The achieved results not only improve the simple baseline but also compare and sometimes are better than more sophisticated IC systems based on attention mechanisms.

Finally, we have proposed a novel track named change captioning applied to multitemporal remote sensing images. Specifically, our methods can be applied not only to the conventional single RGB images but also to multispectral and multitemporal images. Compared to traditional change detection algorithm, our proposed solution does not produce a change map highlighting the changed areas but is able to describe the changes in a given geographical area with sentence descriptions. This allows to widen the benefit of remote sensing change detection not only to expert users but also to non-expert ones. To test the proposed change captioning systems, we have constructed two bi-temporal RS datasets. The first one is composed of very high spatial resolution RGB images while the second one is composed of medium spatial resolution multispectral images. In the latter, we showed the importance of the spectral information to generate coherent change descriptions. To advance the task of CC, the constructed dataset are publically available in the following link: https://disi.unitn.it/~melgani/datasets.html.

6.2. Future Directions

We saw that the proposed content based image retrieval system is composed of two main blocks: the image captioning and the caption matching block. We noticed that the proposed system heavily depends on the image captioning block. Furthermore, the two blocks are trained in different stages. A future work could be to have a unified training that on the same time generates the textual descriptions and performs retrieval. This could be done by embedding in the same space image and sentence representation and learning the similarity directly on the unified space. By doing so the unified framework would take advantages of the both tasks improving the retrieval accuracy and on the same time become less dependent on the captioning system.

The IC architecture proposed in the RS literature are done in a supervised way. This means that there is a need of a training set composed of samples (images) and labels (sentences). During the training phase, the

system should learn the inherent relationship between the training samples and the corresponding labels to be able to generalize on unseen samples. However, obtaining the labels of the images is not a trivial task, especially when these labels are sentence descriptions in which the annotator should carefully analyse the content of an image and summarize this content into a concise sentence. Furthermore, there is a need to provide more than a sentence to have a robust RSIC system as the descriptions are very subjective. Hence, more than one annotator is needed to create a good training set demanding for more resources in terms of cost and time. To alleviate the need of large annotated samples, we proposed the SVM decoder which is particularly suitable when only a small number of annotated samples is available. However, more should be done in this direction. For instance, creating an IC system completely based on SVMs can be a future direction to further alleviate the overfitting problem. To this end, convolutional SVM (CSVM) can be exploit instead of the CNN as encoders [140]. Another future direction could be active learning based solutions. Given a suboptimal training set an active learning system iteratively selects the most relevant samples from a large amount of unlabeled data (learning set), to use for training with the aim of limiting the number of training samples as much as possible while maintaining the system's accuracy as high as possible. Active learning solutions are not new in the RS community. They have been proposed in several RS applications such as image classification [141]–[143] or image retrieval [61]. We believe that this could be a future direction worthy to exploit also for RSIC to alleviate the need of large annotated samples.

The proposed post-processing strategies are based on the combination of Hidden Markov Models and Viterbi algorithm to determine a set of possible states and to find the optimal sequence of states, respectively. They are applied to simple encoder decoder IC systems. As future work, we propose to combine the post-processing strategies with attention mechanism to correct the mismatch between the words and the related image parts. We also believe that is worthy exploring graph neural network for post-processing as they are now a default choice when dealing with structured data such are the sentences. Furthermore, they can be applied also to the images in order to better relate the image parts with relevant words [144].

The proposed change captioning systems are applied to bi-temporal images. Even though they constitute a novel track in the RS community to analyse multitemporal and multispectral data, for now the temporal order is two. As future, we propose to verify the effectiveness of the proposed change captioning systems on multitemporal data of order bigger than two. To this end, a future direction could be to construct datasets of such kind. Furthermore, it would be worthy to exploit visual questioning answering [145] and generation [146] do aid the single or multiple image captioning systems in generating more coherent descriptions that describe the content of the image or their changes.

Finally, in this thesis we considered only optical images (RGB or multispectral) characterized by a high or medium spatial resolution. The experiments presented in Chapter 5, showed that using all the spectral information significantly improves the results of a change captioning system compared to only using the RGB channels. This highlights the importance of spectral information in generating more reliable change descriptions. We believe that it would be interesting to apply image captioning (single or multi date) on hyperspectral data or even on other type of data such as images acquired by active systems (i.e., Radar or Lidar). To this end, it would be first important to interact with potential end-users in order to define the captioning tasks and its peculiarities, and then construct the related datasets.

Publications and Awards

The contents of this thesis are based on several peer-reviewed papers published during my PhD studies together with some work that are under review process. Among the published and under review papers the most relevant ones to this thesis are listed as follows.

Journal Articles

- **Improving Image Captioning Systems with Post-Processing Strategies**
  G. Hoxha, and F. Melgani
  *IEEE Transactions on Geoscience and Remote Sensing (**TGRS**), (Under Revision).*
- **Change Captioning: A new Paradigm for Multitemporal Remote Sensing Image Analysis**
  G. Hoxha, S. Chouaf,, F. Melgani and Youcef Smara,
  *IEEE Transactions on Geoscience and Remote Sensing (**TGRS**), **(2022),** (Accepted, in press).*
- **A Novel SVM-Based Decoder for Remote Sensing Image Captioning**
  G. Hoxha and F. Melgani
  *IEEE Transactions on Geoscience and Remote Sensing (**TGRS**), (**2022**).*
- **Toward Remote Sensing Image Retrieval Under a Deep Image Captioning Perspective**
  G. Hoxha, F. Melgani and B. Demir,
  *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **(JSTARS), (2020)**.*

Conference Proceedings

- **Captioning Changes in Bi-Temporal Remote Sensing Images**
  S. Chouaf, G. Hoxha, F. Melgani and Youcef Smara,
  IEEE International Geoscience and Remote Sensing Symposium (**IGARSS**)
  Brussels, Belgium (**2021**).
- **Remote Sensing Image Captioning with SVM-Based Decoding**
  G. Hoxha and F. Melgani
  IEEE International Geoscience and Remote Sensing Symposium (**IGARSS**)
  Waikoloa, Hawaii, USA (**2020**).
- **Retrieving images with generated descriptions**
  G. Hoxha, F. Melgani and B. Demir,
  IEEE International Geoscience and Remote Sensing Symposium (**IGARSS**)
  Yokohama, Japan, (**2019**).

Awards

- **Best Oral Paper Award, First Place**
  In recognition of an Outstanding Oral Paper Contribution at the Mediterranean and Middle-East Geoscience and Remote Sensing Symposium **(M2GARSS 22')**
- **Best Oral Paper Award, First Place**
  In recognition of an Outstanding Oral Paper Contribution at the Mediterranean and Middle-East Geoscience and Remote Sensing Symposium **(M2GARSS 20')**

Bibliography

[1] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *Journal of vision*, vol. 7, no. 1, p. 10, 2007, doi: 10.1167/7.1.10.

[2] "Introduction," in *Introduction to the Physics and Techniques of Remote Sensing*, John Wiley & Sons, Ltd, 2006, pp. 1–21. doi: 10.1002/0471783390.ch1.

[3] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018, doi: 10.1109/TGRS.2017.2760909.

[4] 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 Sensing*, vol. 12, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/rs12030405.

[5] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4284–4297, 2021, doi: 10.1109/JSTARS.2021.3070872.

[6] F. Melgani and S. B. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 11, pp. 2478–2487, Nov. 2003, doi: 10.1109/TGRS.2003.817269.

[7] M. B. Bejiga, G. Hoxha, and F. Melgani, "Retro-Remote Sensing With Doc2Vec Encoding," in *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, Mar. 2020, pp. 89–92. doi: 10.1109/M2GARSS47143.2020.9105139.

[8] M. B. Bejiga, G. Hoxha, and F. Melgani, "Improving Text Encoding for Retro-Remote Sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 4, pp. 622–626, Apr. 2021, doi: 10.1109/LGRS.2020.2983851.

[9] B. Zhao, "A Systematic Survey of Remote Sensing Image Captioning," *IEEE Access*, vol. 9, pp. 154086–154111, 2021, doi: 10.1109/ACCESS.2021.3128140.

[10] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 444–454.

[11] A. Farhadi *et al.*, "Every Picture Tells a Story: Generating Sentences from Images," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*, Berlin, Heidelberg, 2010, pp. 15–29. Accessed: Oct. 22, 2019. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888089.1888092

[12] G. Kulkarni *et al.*, "Baby talk: Understanding and generating simple image descriptions," in *CVPR 2011*, Jun. 2011, pp. 1601–1608. doi: 10.1109/CVPR.2011.5995466.

[13] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing Simple Image Descriptions using Web-scale N-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, Portland, Oregon, USA, Jun. 2011, pp. 220–228. Accessed: Oct. 23, 2019. [Online]. Available: https://www.aclweb.org/anthology/W11-0326

[14] V. Ordonez, G. Kulkarni, and T. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," in *Advances in Neural Information Processing Systems*, 2011, vol. 24. Accessed: Mar. 31, 2022. [Online]. Available: https://papers.nips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

[15] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks," *arXiv:1410.1090 [cs]*, Oct. 2014, Accessed: May 13, 2019. [Online]. Available: http://arxiv.org/abs/1410.1090

[16] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3128–3137. doi: 10.1109/CVPR.2015.7298932.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.

[18] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *arXiv:1502.03044 [cs]*, Apr. 2016, Accessed: Oct. 30, 2019. [Online]. Available: http://arxiv.org/abs/1502.03044

[19] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022, doi: 10.1109/TPAMI.2022.3148210.

[20] Z. Shi and Z. Zou, "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017, doi: 10.1109/TGRS.2017.2677464.

[21] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic Descriptions of High-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2019, doi: 10.1109/LGRS.2019.2893772.

[22] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Jul. 2016, pp. 1–5. doi: 10.1109/CITS.2016.7546397.

[23] X. Zhang, X. Li, J. An, L. Gao, B. Hou, and C. Li, "Natural language description of remote sensing images based on deep learning," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2017, pp. 4798–4801. doi: 10.1109/IGARSS.2017.8128075.

[24] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, doi: 10.1109/TGRS.2017.2776321.

[25] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description Generation for Remote Sensing Images Using Attribute Attention Mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, Jan. 2019, doi: 10.3390/rs11060612.

[26] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-Scale Cropping Mechanism for Remote Sensing Image Captioning," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2019, pp. 10039–10042. doi: 10.1109/IGARSS.2019.8900503.

[27] W. Huang, Q. Wang, and X. Li, "Denoising-Based Multiscale Feature Fusion for Remote Sensing Image Captioning," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020, doi: 10.1109/LGRS.2020.2980933.

[28] X. Lu, B. Wang, and X. Zheng, "Sound Active Attention Framework for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020, doi: 10.1109/TGRS.2019.2951636.

[29] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 256–270, 2020, doi: 10.1109/JSTARS.2019.2959208.

[30] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[31] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. Accessed: Oct. 29, 2019. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[32] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” *arXiv:1409.0473 [cs, stat]*, May 2016, Accessed: Oct. 29, 2019. [Online]. Available: http://arxiv.org/abs/1409.0473
[33] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[34] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. Accessed: May 16, 2019. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
[36] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
[37] C. Szegedy *et al.*, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.
[38] Sivic and Zisserman, “Video Google: a text retrieval approach to object matching in videos,” in *Proceedings Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1470–1477 vol.2. doi: 10.1109/ICCV.2003.1238663.
[39] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, “Large-scale image retrieval with compressed fisher vectors,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3384–3391.
[40] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.
[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.
[42] X. Ma, R. Zhao, and Z. Shi, “Multiscale Methods for Optical Remote-Sensing Image Captioning,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020, doi: 10.1109/LGRS.2020.3009243.
[43] G. Sumbul, S. Nayak, and B. Demir, “SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020, doi: 10.1109/TGRS.2020.3031111.
[44] A. See, P. J. Liu, and C. D. Manning, “Get To The Point: Summarization with Pointer-Generator Networks,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, Jul. 2017, pp. 1073–1083. doi: 10.18653/v1/P17-1099.
[45] G. Hoxha, F. Melgani, and J. Slaghenauffi, “A New CNN-RNN Framework For Remote Sensing Image Captioning,” in *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, Mar. 2020, pp. 1–4. doi: 10.1109/M2GARSS47143.2020.9105191.
[46] X. Li, X. Zhang, W. Huang, and Q. Wang, “Truncation Cross Entropy Loss for Remote Sensing Image Captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2020, doi: 10.1109/TGRS.2020.3010106.
[47] R. Zhao, Z. Shi, and Z. Zou, “High-Resolution Remote Sensing Image Captioning Based on Structured Attention,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021, doi: 10.1109/TGRS.2021.3070383.

[48] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 10575–10584. doi: 10.1109/CVPR42600.2020.01059.

[49] X. Shen, B. Liu, Y. Zhou, and J. Zhao, "Remote sensing image caption generation via transformer and reinforcement learning," *Multimed Tools Appl*, vol. 79, no. 35–36, pp. 26661–26682, Sep. 2020, doi: 10.1007/s11042-020-09294-7.

[50] X. Shen, B. Liu, Y. Zhou, J. Zhao, and M. Liu, "Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning," *Knowledge-Based Systems*, vol. 203, p. 105920, Sep. 2020, doi: 10.1016/j.knosys.2020.105920.

[51] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-Critical Sequence Training for Image Captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1179–1195. doi: 10.1109/CVPR.2017.131.

[52] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *ICLR*, 2014.

[53] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 740–755.

[54] Du Peijun, Chen Yunhao, Tang Hong, and Fang Tao, "Study on content-based remote sensing image retrieval," in *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05.*, Jul. 2005, vol. 2, p. 4 pp.-. doi: 10.1109/IGARSS.2005.1525204.

[55] 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, doi: 10.1145/1013208.1013210.

[56] Y. Yang and S. Newsam, "Geographic Image Retrieval Using Local Invariant Features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818–832, Feb. 2013, doi: 10.1109/TGRS.2012.2205158.

[57] I. Tekeste and B. Demir, "Advanced Local Binary Patterns for Remote Sensing Image Retrieval," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2018, pp. 6855–6858. doi: 10.1109/IGARSS.2018.8518856.

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

[59] K. Amiri and M. Farah, "Graph of Concepts for Semantic Annotation of Remotely Sensed Images based on Direct Neighbors in RAG," *Canadian Journal of Remote Sensing*, vol. 44, no. 6, pp. 551–574, Nov. 2018, doi: 10.1080/07038992.2019.1569507.

[60] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018, doi: 10.1109/JSTARS.2018.2832985.

[61] B. Demir and L. Bruzzone, "A Novel Active Learning Method in Relevance Feedback for Content-Based Remote Sensing Image Retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2323–2334, May 2015, doi: 10.1109/TGRS.2014.2358804.

[62] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-Scale Remote Sensing Image Retrieval by Deep Hashing Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 950–965, Feb. 2018, doi: 10.1109/TGRS.2017.2756911.

[63] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Computer Vision and Image Understanding*, vol. 184, pp. 22–30, Jul. 2019, doi: 10.1016/j.cviu.2019.04.004.

[64] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Deep Metric and Hash-Code Learning for Content-Based Retrieval of Remote Sensing Images," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2018, pp. 4539–4542. doi: 10.1109/IGARSS.2018.8518381.
[65] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval," *Remote Sensing*, vol. 9, no. 5, p. 489, May 2017, doi: 10.3390/rs9050489.
[66] 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 Geoscience and Remote Sensing Letters*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018, doi: 10.1109/LGRS.2018.2847303.
[67] 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.
[68] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020, doi: 10.1109/JSTARS.2019.2961634.
[69] G. Hoxha, F. Melgani, and B. Demir, "Retrieving Images with Generated Textual Descriptions," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2019, pp. 5812–5815. doi: 10.1109/IGARSS.2019.8899321.
[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[71] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
[72] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv:1301.3781 [cs]*, Jan. 2013, Accessed: May 13, 2019. [Online]. Available: http://arxiv.org/abs/1301.3781
[73] 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, Oct. 2014, pp. 1532–1543. Accessed: Jan. 07, 2019. [Online]. Available: http://www.aclweb.org/anthology/D14-1162
[74] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.
[75] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*, 2015, pp. 957–966.
[76] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, Nov. 2000, doi: 10.1023/A:1026543900054.
[77] "GloVe: Global Vectors for Word Representation." https://nlp.stanford.edu/projects/glove/ (accessed Jan. 07, 2019).
[78] A. Zeggada, F. Melgani, and Y. Bazi, "A Deep Learning Approach to UAV Image Multilabeling," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 694–698, May 2017, doi: 10.1109/LGRS.2017.2671922.
[79] 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. doi: 10.3115/1073083.1073135.

[80] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Pacific-Asia conference on knowledge discovery and data mining*, 2004, pp. 22–30.

[81] 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, Jun. 2014, pp. 362–367. Accessed: Jan. 07, 2019. [Online]. Available: http://www.aclweb.org/anthology/W14-3346

[82] V. Vapnik and V. Vapnik, "Statistical learning theory Wiley," *New York*, pp. 156–160, 1998.

[83] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004, doi: 10.1109/TGRS.2004.831865.

[84] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," 2014, pp. 806–813. Accessed: Jan. 07, 2019. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_workshops_2014/W15/html/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.html

[85] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[86] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10*, San Jose, California, 2010, p. 270. doi: 10.1145/1869790.1869829.

[87] F. Zhang, B. Du, and L. Zhang, "Saliency-Guided Unsupervised Feature Learning for Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015, doi: 10.1109/TGRS.2014.2357078.

[88] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81. Accessed: Jul. 20, 2020. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[89] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, Jun. 2005, pp. 65–72. Accessed: Jul. 20, 2020. [Online]. Available: https://www.aclweb.org/anthology/W05-0909

[90] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4566–4575. doi: 10.1109/CVPR.2015.7299087.

[91] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[92] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv:1412.3555 [cs]*, Dec. 2014, Accessed: Sep. 23, 2019. [Online]. Available: http://arxiv.org/abs/1412.3555

[93] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Nat. Lang. Eng.*, vol. 24, no. 3, pp. 467–489, May 2018, doi: 10.1017/S1351324918000098.

[94] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 5561–5570. doi: 10.1109/CVPR.2018.00583.

[95] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021, doi: 10.1109/TGRS.2020.3031111.

[96] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *ICLR*, 2015.
[97] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989, doi: 10.1109/5.18626.
[98] S. P. Abercrombie and M. A. Friedl, “Improving the Consistency of Multitemporal Land Cover Maps Using a Hidden Markov Model,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 2, pp. 703–713, Feb. 2016, doi: 10.1109/TGRS.2015.2463689.
[99] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556 [cs]*, Sep. 2014, Accessed: Jan. 07, 2019. [Online]. Available: http://arxiv.org/abs/1409.1556
[100] G. Hoxha and F. Melgani, “A Novel SVM-Based Decoder for Remote Sensing Image Captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3105004.
[101] A. Singh, “Review Article Digital change detection techniques using remotely-sensed data,” *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, Jun. 1989, doi: 10.1080/01431168908903939.
[102] L. Bruzzone and F. Bovolo, “A Novel Framework for the Design of Change-Detection Systems for Very-High-Resolution Remote Sensing Images,” *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013, doi: 10.1109/JPROC.2012.2197169.
[103] L. Bruzzone and D. F. Prieto, “Automatic analysis of the difference image for unsupervised change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, May 2000, doi: 10.1109/36.843009.
[104] F. Melgani, G. Moser, and S. B. Serpico, “Unsupervised change-detection methods for remote-sensing images,” *OE*, vol. 41, no. 12, pp. 3288–3297, Dec. 2002, doi: 10.1117/1.1518995.
[105] Y. Bazi, L. Bruzzone, and F. Melgani, “An unsupervised approach based on the generalized Gaussian model to automatic change detection in multitemporal SAR images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 874–887, Apr. 2005, doi: 10.1109/TGRS.2004.842441.
[106] F. Melgani and Y. Bazi, “Markovian Fusion Approach to Robust Unsupervised Change Detection in Remotely Sensed Imagery,” *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 4, pp. 457–461, Oct. 2006, doi: 10.1109/LGRS.2006.875773.
[107] G. Moser and S. B. Serpico, “Generalized minimum-error thresholding for unsupervised change detection from SAR amplitude imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 10, pp. 2972–2982, Oct. 2006, doi: 10.1109/TGRS.2006.876288.
[108] T. Celik, “Multiscale Change Detection in Multitemporal Satellite Images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 820–824, Oct. 2009, doi: 10.1109/LGRS.2009.2026188.
[109] Y. Bazi, F. Melgani, and H. D. Al-Sharari, “Unsupervised Change Detection in Multispectral Remotely Sensed Imagery With Level Set Methods,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 8, pp. 3178–3187, Aug. 2010, doi: 10.1109/TGRS.2010.2045506.
[110] O. Yousif and Y. Ban, “Improving Urban Change Detection From Multitemporal SAR Images Using PCA-NLM,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 2032–2041, Apr. 2013, doi: 10.1109/TGRS.2013.2245900.
[111] P. Serra, X. Pons, and D. Saurí, “Post-classification change detection with data from different sensors: Some accuracy considerations,” *null*, vol. 24, no. 16, pp. 3311–3340, Jan. 2003, doi: 10.1080/0143116021000021189.
[112] F. Yuan, K. E. Sawaya, B. C. Loeffelholz, and M. E. Bauer, “Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat

remote sensing," *Remote Sensing of Environment*, vol. 98, no. 2, pp. 317–328, Oct. 2005, doi: 10.1016/j.rse.2005.08.006.

[113] P. Zhong and R. Wang, "A Multiple Conditional Random Fields Ensemble Model for Urban Area Detection in Remote Sensing Optical Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 3978–3988, Dec. 2007, doi: 10.1109/TGRS.2007.907109.

[114] O. Ahlqvist, "Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 U.S. National Land Cover Database changes," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1226–1241, Mar. 2008, doi: 10.1016/j.rse.2007.08.012.

[115] C. Benedek and T. Sziranyi, "Change Detection in Optical Aerial Images by a Multilayer Conditional Mixed Markov Model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009, doi: 10.1109/TGRS.2009.2022633.

[116] T. Habib, J. Inglada, G. Mercier, and J. Chanussot, "Support Vector Reduction in SVM Algorithm for Abrupt Change Detection in Remote Sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 606–610, Jul. 2009, doi: 10.1109/LGRS.2009.2020306.

[117] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised change detection in VHR images using contextual information and support vector machines," *International Journal of Applied Earth Observation and Geoinformation*, vol. 20, pp. 77–85, Feb. 2013, doi: 10.1016/j.jag.2011.10.013.

[118] P. Singh, Z. Kato, and J. Zerubia, "A Multilayer Markovian Model for Change Detection in Aerial Image Pairs with Large Time Differences," in *2014 22nd International Conference on Pattern Recognition*, Aug. 2014, pp. 924–929. doi: 10.1109/ICPR.2014.169.

[119] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sensing of Environment*, vol. 199, pp. 241–255, Sep. 2017, doi: 10.1016/j.rse.2017.07.009.

[120] Z. Shao, H. Fu, P. Fu, and L. Yin, "Mapping Urban Impervious Surface by Fusing Optical and SAR Data at the Decision Level," *Remote Sensing*, vol. 8, no. 11, Art. no. 11, Nov. 2016, doi: 10.3390/rs8110945.

[121] N. Ghoggali and F. Melgani, "Genetic SVM Approach to Semisupervised Multitemporal Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 2, pp. 212–216, Apr. 2008, doi: 10.1109/LGRS.2008.915600.

[122] B. Demir, F. Bovolo, and L. Bruzzone, "Updating Land-Cover Maps by Classification of Image Time Series: A Novel Change-Detection-Driven Transfer Learning Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 300–312, Jan. 2013, doi: 10.1109/TGRS.2012.2195727.

[123] X. X. Zhu *et al.*, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, Dec. 2017, doi: 10.1109/MGRS.2017.2762307.

[124] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.015.

[125] Y. You, J. Cao, and W. Zhou, "A Survey of Change Detection Methods Based on Remote Sensing Images for Multi-Source and Multi-Objective Scenarios," *Remote Sensing*, vol. 12, no. 15, Art. no. 15, Jan. 2020, doi: 10.3390/rs12152460.

[126] L. Khelifi and M. Mignotte, "Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020, doi: 10.1109/ACCESS.2020.3008036.

[127] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change Detection in Synthetic Aperture Radar Images Based on Deep Neural Networks," *IEEE Transactions on Neural Networks*

and Learning Systems, vol. 27, no. 1, pp. 125–138, Jan. 2016, doi: 10.1109/TNNLS.2015.2435783.

[128] J. Geng, H. Wang, J. Fan, and X. Ma, "Change detection of SAR images based on supervised contractive autoencoders and fuzzy clustering," in *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*, May 2017, pp. 1–3. doi: 10.1109/RSIP.2017.7958819.

[129] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting Large-Scale Urban Land Cover Changes from Very High Resolution Remote Sensing Images Using CNN-Based Classification," *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, Art. no. 4, Apr. 2019, doi: 10.3390/ijgi8040189.

[130] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019, doi: 10.1109/TGRS.2018.2886643.

[131] D. Peng, Y. Zhang, and H. Guan, "End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++," *Remote Sensing*, vol. 11, no. 11, Art. no. 11, Jan. 2019, doi: 10.3390/rs11111382.

[132] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred Deep Learning-Based Change Detection in Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019, doi: 10.1109/TGRS.2019.2909781.

[133] R. Hedjam, A. Abdesselam, and F. Melgani, "Change Detection in Unlabeled Optical Remote Sensing Data Using Siamese CNN," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4178–4187, 2020, doi: 10.1109/JSTARS.2020.3009116.

[134] Z. Wang, C. Peng, Y. Zhang, N. Wang, and L. Luo, "Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss," *Neurocomputing*, vol. 457, pp. 155–167, Oct. 2021, doi: 10.1016/j.neucom.2021.06.059.

[135] R. Liu, Z. Cheng, L. Zhang, and J. Li, "Remote Sensing Image Change Detection Based on Information Transmission and Attention Mechanism," *IEEE Access*, vol. 7, pp. 156349–156359, 2019, doi: 10.1109/ACCESS.2019.2947286.

[136] L. Li, H. Ma, and Z. Jia, "Change Detection from SAR Images Based on Convolutional Neural Networks Guided by Saliency Enhancement," *Remote Sensing*, vol. 13, no. 18, Art. no. 18, Jan. 2021, doi: 10.3390/rs13183697.

[137] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, "Captioning Changes in Bi-Temporal Remote Sensing Images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, pp. 2891–2894. doi: 10.1109/IGARSS47720.2021.9554419.

[138] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," *Remote Sensing*, vol. 12, no. 10, Art. no. 10, Jan. 2020, doi: 10.3390/rs12101662.

[139] Earth Resources Observation And Science (EROS) Center, "Collection-1 Landsat 7 Enhanced Thematic Mapper Plus (ETM+) Level-1 Data Products." U.S. Geological Survey, 2018. doi: 10.5066/F7WH2P8G.

[140] Y. Bazi and F. Melgani, "Convolutional SVM Networks for Object Detection in UAV Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3107–3118, Jun. 2018, doi: 10.1109/TGRS.2018.2790926.

[141] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active Learning Methods for Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009, doi: 10.1109/TGRS.2008.2010404.

[142] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM Active Learning Approach for Image Classification Using Spatial Information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2217–2233, Apr. 2014, doi: 10.1109/TGRS.2013.2258676.

[143] E. Pasolli, F. Melgani, and Y. Bazi, "Support Vector Machine Active Learning Through Significance Space Construction," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 431–435, May 2011, doi: 10.1109/LGRS.2010.2083630.

[144] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10677–10686. doi: 10.1109/CVPR.2019.01094.

[145] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual Question Answering for Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020, doi: 10.1109/TGRS.2020.2988782.

[146] C. Patil and M. Patwardhan, "Visual Question Generation: The State of the Art," *ACM Comput. Surv.*, vol. 53, no. 3, p. 47:1-47:22, May 2020, doi: 10.1145/3383465.