

Article

TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images

Taghreed Abdullah ¹, Yakoub Bazi ^{2,*}, Mohamad M. Al Rahhal ³, Mohamed L. Mekhalfi ⁴, Lalitha Rangarajan ¹ and Mansour Zuair ²

¹ Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore 570006, India; taghreed@compsci.uni-mysore.ac.in (T.A.); lalithar@compsci.uni-mysore.ac.in (L.R.)

² Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; zuair@ksu.edu.sa

³ Information System Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia; mmalrahhal@ksu.edu.sa

⁴ Department of Information Engineering and Computer Science, University of Trento, Disi Via Sommarive 9, Povo, 28123 Trento, Italy; mohamed.mekhalfi@alumni.unitn.it

* Correspondence: ybazi@ksu.edu.sa; Tel.: +966-1014696297

Received: 15 December 2019; Accepted: 23 January 2020; Published: 27 January 2020



Abstract: Exploring the relevance between images and their respective natural language descriptions, due to its paramount importance, is regarded as the next frontier in the general computer vision literature. Thus, recently several works have attempted to map visual attributes onto their corresponding textual tenor with certain success. However, this line of research has not been widespread in the remote sensing community. On this point, our contribution is three-pronged. First, we construct a new dataset for text-image matching tasks, termed TextRS, by collecting images from four well-known different scene datasets, namely AID, Merced, PatternNet, and NWPU datasets. Each image is annotated by five different sentences. All the five sentences were allocated by five people to evidence the diversity. Second, we put forth a novel Deep Bidirectional Triplet Network (DBTN) for text to image matching. Unlike traditional remote sensing image-to-image retrieval, our paradigm seeks to carry out the retrieval by matching text to image representations. To achieve that, we propose to learn a bidirectional triplet network, which is composed of Long Short Term Memory network (LSTM) and pre-trained Convolutional Neural Networks (CNNs) based on (EfficientNet-B2, ResNet-50, Inception-v3, and VGG16). Third, we top the proposed architecture with an average fusion strategy to fuse the features pertaining to the five image sentences, which enables learning of more robust embedding. The performances of the method expressed in terms Recall@K representing the presence of the relevant image among the top K retrieved images to the query text shows promising results as it yields 17.20%, 51.39%, and 73.02% for $K = 1, 5$, and 10, respectively.

Keywords: remote sensing; text image matching; triplet networks; EfficientNets; LSTM network

1. Introduction

The steady accessibility of remote sensing data, particularly high resolution images, has animated remarkable research outputs in the remote sensing community. Two of the most active topics in this regard refer to image classification and retrieval [1–5]. Image classification aims to assign scene images to a discrete set of land use/land cover classes depending on the image content [6–10]. Recently, with rapidly expanded remote sensing acquisition technologies, both quantity and quality of remote sensing data have been increased. In this context, content-based image retrieval (CBIR) has become a paramount research subject in order to meet the increasing need for the efficient organization and

management of massive volumes of remote sensing data, which has been a long lasting challenge in the community of remote sensing.

In the last decades, great efforts have been made to develop effective and precise retrieval approaches to search for interest information across large archives of remote sensing. A typical CBIR system involves two main steps [11], namely feature extraction and matching, where the most relevant images from the archive are retrieved. In this regard, both extraction of features as well as matching play a pivotal role in controlling the efficiency of a retrieval system [12].

Content-based remote sensing image retrieval is a particular application of CBIR, in the field of remote sensing. However, the remote sensing community seems to put the emphasis more on devising powerful features due to the fact that image retrieval systems performance relies greatly on the effectiveness of the extracted features [13]. In this respect, remote sensing image retrieval approaches rely on handcrafted features and deep-learning.

As per handcrafted features, low-level features are harnessed to depict the semantic tenor of remote sensing images, and it is possible to draw them from either local or global regions of the image. Color features [14,15], texture features [2,16,17], and shape features [18] are widely applied as global features. On other hand, local features tend to emphasize the description on local regions instead of looking at the image as a whole. There are various algorithms for describing local image regions such as the scale-invariant feature transform (SIFT) and speed up robust features (SURF) [19,20]. The bag-of-word (BOW) model [21], and the vector of aggregated local descriptors (VLAD) [22] are generally proposed to encode local features into a fixed-size image signature via a codebook/dictionary of keypoint/feature vectors.

Recently, remote sensing images have been witnessing a steady increase due to the prominent technological progress of remote sensors [23]. Therefore, huge volumes of data with various spatial dimensions and spectral channels can be availed [24]. On this point, handcrafted features may be personalized and successfully tailored to small chunks of data; they do not meet, however, the standards of practical contexts where the size and complexity of data increases. Nowadays, deep learning strategies, which aim to learn automatically the discriminative and representative features, are highly effective in large-scale image recognition [25–27], object detection [28,29], semantic segmentation [30,31], and scene classification [32]. Furthermore, recurrent neural networks (RNNs) have achieved immense success with various tasks in sequential data analysis as recognition of action [33,34] and image captioning [35]. Recent research shows that image retrieval approaches work particularly well by exploiting deep neural networks. For example, the authors in [36] introduced a content-based remote sensing image retrieval approach depending on deep metric learning using a triplet network. The proposed approach has shown promising results compared to prior state-of-the-art approaches. The work in [37] presented an unsupervised deep feature learning method for the retrieval task of remote sensing images. Yang et al. [38] proposed a dynamic kernel with a deep convolutional neural network (CNN) for image retrieval. It focuses on matching patches between the filters and relevant images and removing the ones for irrelevant pairs. Furthermore, deep hashing neural network strategies are adopted in some works for large-scale remote sensing image retrieval [39]. Li et al. [40] presented a new unsupervised hashing method, the aim of which is to build an effective hash function. In another work, Li et al. [41], investigated cross-source remote sensing image retrieval via source-invariant deep hashing CNNs, which automatically extract the semantic feature for multispectral data.

It is worthwhile mentioning that the aforementioned image retrieval methods are single label retrieval approaches, where the query image and the images to be retrieved are labelled by a single class label. Although these approaches have been applied with a certain amount of success, they tend to abstract the rich semantic tenor of a remote sensing image into a single label.

In order to moderate the semantic gap and enhance the retrieval performance, recent remote sensing research proposed multi-label approaches. For instance, the work in [12] presented a multi-label method, making use of a semi-supervised graph-theoretic technique in order to improve the region-based retrieval method [42]. Zhou et al. [43] proposed a multi-label retrieval technique

by training a CNN for semantic segmentation and feature generation. Shao et al. [11] constructed a dense labeling remote sensing dataset to evaluate the performance of retrieval techniques based on traditional handcrafted feature as well as deep learning-based ones. Dai et al. [44] discussed the use of multiple hyperspectral image retrieval labels and introduced a multi-label scheme that incorporates spatial and spectral features.

It is evident that the multi-label scenario is generally favored (over the single label case) on account of its abundant semantic information. However, it remains limited due to the discrete nature of labels pertaining to a given image. This suggests a further endeavor to model the relation among objects/labels using an image description. With the rapid advancement of computer vision and natural language processing (NLP), machines began to understand, slowly but surely, the semantics of images.

Current computer vision literature suggests that, instead of tackling the problem from an image-to-image matching perspective, cross-modal text-image learning seems to offer a more concrete alternative. This concept has manifested itself lately in the form of image captioning, which stems as a crossover where computer vision meets NLP. Basically, it consists of generating a sequential textual narration of visual data, similar to how humans perceive it. In fact, image captioning is considered as a subtle aid for image grasping, as a description generation model should capture not only the objects/scenes presented in the image, but it should also be capable of expressing how the objects/scenes relate to each other in a textual sentence.

The leading deep learning techniques, for image captioning, can be categorized into two streams. One stream adopts encoder-decoder, an end-to-end fashion [45,46] where a CNN is typically considered as the encoder and an RNN as the decoder, often a Long-Short Term Memory (LSTM) [47]. Rather than translating between various languages, such techniques translate from a visual representation to language. The visual representation is extracted via a pre-trained CNN [48]. Translation is achieved by RNNs based language models. The major usefulness of this method is that the whole system adopts end to end learning [47]. Xu et al. [35] went one step further by introducing the attention mechanism, which enables the decoder to concentrate on specific portions of the input image when generating a word. The other stream adopts a compositional framework, such as [49] for instance, which divided the task of generating the caption into various parts: detection of the words by a CNN, generating the caption candidates, and re-ranking the sentence by a deep multimodal similarity model.

With respect to image captioning, the computer vision literature suggests several contributions mainly based on deep learning. For instance, You et al. [50] combined top-down (i.e., image-to-words) and bottom-up (i.e., joining several relevant words into a meaningful image description) approaches via CNN and RNN models for image captioning, which revealed interesting experimental results. Chen et al. [51] proposed an alternative architecture based on spatial and channel-wise attention for image captioning. In other works, a common deep model called a bi-directional spatial–semantic attention network was introduced [52,53], where an embedding and a similarity network were adopted to model the bidirectional relations between pairs of text and image. Zhang and Lu [54] proposed a projection classification loss that classified the vector projection of representations from one form to another by improving the norm-softmax loss. Huang et al. [52] addressed the problem of image text matching in bi-direction by making use of attention networks.

So far, it can be noted that computer vision has been accumulating a steady research basis in the context of image captioning [47,50,55]. In remote sensing, however, contributions have barely begun to move in this direction, often regarded as the ‘next frontier’ in computer vision. Lu et al. [56] for instance, proposed a similar concept as in [51] by combining CNNs (for image representation) and LSTM network for sentence generation in remote sensing images. Shi et al. [57] leveraged a fully convolutional architecture for remote sensing image description. Zhang et al. [58] adopted an attribute attention strategy to produce remote sensing image description, and investigated the effect of the attributes derived from remote sensing images on the attention system.

As we have previously reviewed, the mainstream of the remote sensing works focuses mainly on scenarios of single label, whereas in practice images may contain many classes simultaneously.

In the quest for tackling this bottleneck, recent works attempted to allocate multiple labels to a single query image. Nevertheless, coherence among the labels in such cases remains questionable since multiple labels are assigned to an image regardless of their relativity. Therefore, these methods do not specify (or else model) explicitly the relation between the different objects in a given image for a better understanding of its content. Evidently, remote sensing image description has witnessed rather scarce attention in this sense. This may be explained by the fact that remote sensing images exhibit a wide range of morphological complexities and scale changes, which render text to/from image retrieval intricate.

In this paper we propose a solution based DBTN for solving the text-to-image matching problem. It is worth mentioning that this work is inspired from [53]. The major contributions of this work can be highlighted as follows:

- Departing from the fact that the task of text-image retrieval/matching is a new topic in the remote sensing community, we deem it necessary to build a benchmark dataset for remote sensing image description. Our dataset will constitute a benchmark for future research in this respect.
- We propose a DBTN architecture to address the problem of text image matching, which to the best of our knowledge, has never been posed in remote sensing prior-art thus far.
- We tie the single models into fusion schemes that can improve the overall performance through adopting the five sentences.

The paper includes five sections, where the structure of the paper is as follows. In Section 2, we introduce the proposed DBTN method. Section 3 presents the TextRS dataset and the experimental results followed by discussions in Section 4. Finally, Section 5 provides conclusions and directions for future developments.

2. Description of the Proposed Method

Assume a training set $\mathcal{D} = \{X_i, Y_i\}_{i=1}^N$ composed of N images with their matching sentences. In particular, to each training image X_i we associated a set of M matching sentences $Y_i = \{y_i^1, \dots, y_i^K\}$. In the test phase, given a query sentence t_q , we aimed to retrieve the most relevant image in the training set \mathcal{D} . Figure 1 shows a general description of the proposed DBTN method composed of image and text encoding branches that aimed to learn appropriate image and text embeddings $f(X_i)$ and $g(T_i)$, respectively, by optimizing a bidirectional triplet loss. Detailed descriptions are provided in the next sub-sections.

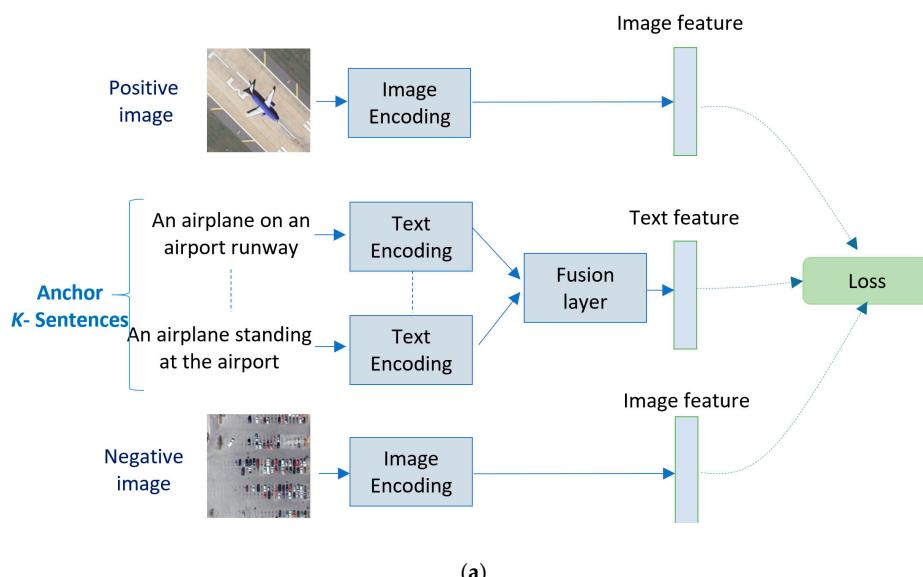


Figure 1. Cont.

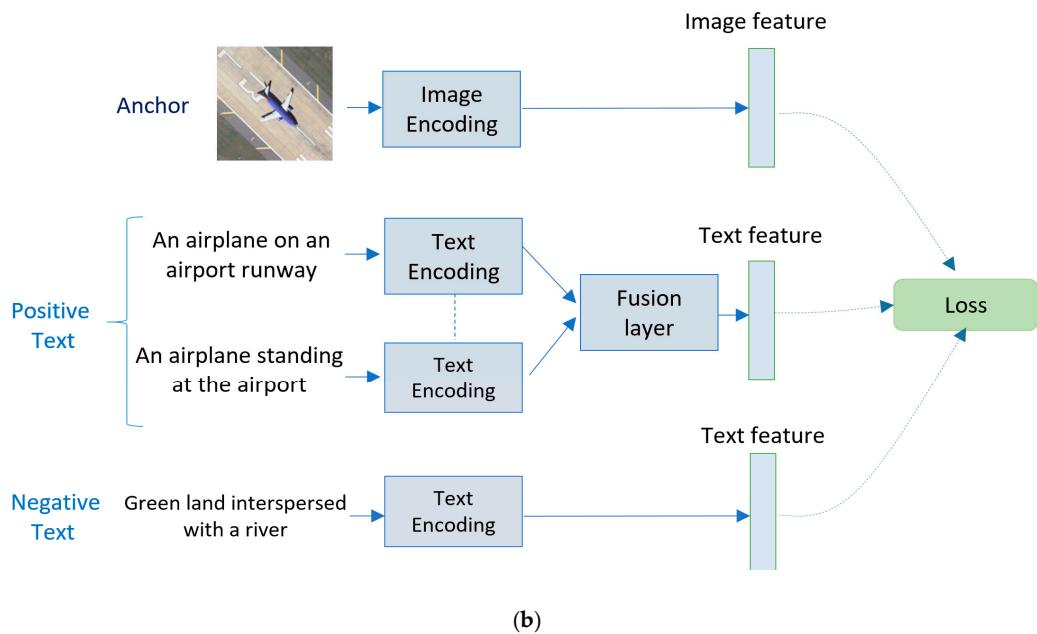


Figure 1. Flowchart of the proposed Deep Bidirectional Triplet Network (DBTN): (a) text as anchor, (b) image as anchor.

2.1. Image Encoding Module

The image encoding module uses a pre-trained CNN augmented with an additional network to learn the visual features $f(X_i)$ of the image (Figure 2). To learn informative features and suppress less relevant ones, this extra network applies a channel attention layer termed squeeze excitation (SE) to the activation maps layer obtained after the 3×3 convolution layer. The goal is to enhance further the representation of the features by grasping the significance of each feature map among all extracted feature maps. As illustrated in Figure 2, the squeeze operation produces features of dimension (1,1,128) by means of global average pooling (GAP), which are then fed to a fully connected layer to reduce the dimension by 1/16. Then the produced feature vector s calibrates the feature maps of each channel (V) by channel-wise scale operation. SE works as shown below [59]:

$$s = \text{Sigmoid}(W_2(\text{ReLU}(W_1(V)))) \quad (1)$$

$$V_{SE} = s \odot V \quad (2)$$

where s is the scaling factor, \odot refers to the channel-wise multiplication, and V represents the feature maps obtained from a particular layer of the pre-trained CNN. Then the resulting activation maps V_{SE} are fed to a GAP followed by a fully connected and l_2 -normalization for feature rescaling yielding the features $f(X_i)$.

As pre-trained CNNs, we adopted in this work different CNNs including VGG16, inception_v3, ResNet50, and EfficientNet. The VGG16 was proposed in 2014 and has 16-layers [27]. Such network was trained on the imangenet dataset to classify 1.2 million RGB images of size 224×224 pixel into 1000 classes. The inception-v3 network [60], introduced by Google, contains 42 layers as well as three kinds of inception modules, which comprise convolution kernels with sizes of 5×5 to 1×1 . Such modules seek to reduce the parameters number. The Residual network (ResNet) [25] is a 50-layer network with shortcut connection. This network was proposed for deeper networks to solve the problem of vanishing gradients. Finally, EfficientNets, which are new state-of-the-art models with up to 10 times better efficiency (faster as well as smaller), were developed recently by a research team from Google [61] to scale up CNNs using a simple compound coefficient. Differently from traditional approaches that scale network dimensions (width, depth, and resolution) individually, EfficientNet tries to scale each

dimension in a balanced way using a stationary set of scaling coefficients evenly. Practically, the performance of the model can be enhanced by scaling individual dimensions. Further, enhancing the entire performance can be achieved through scaling each dimension uniformly, which leads to higher accuracy and efficiency.

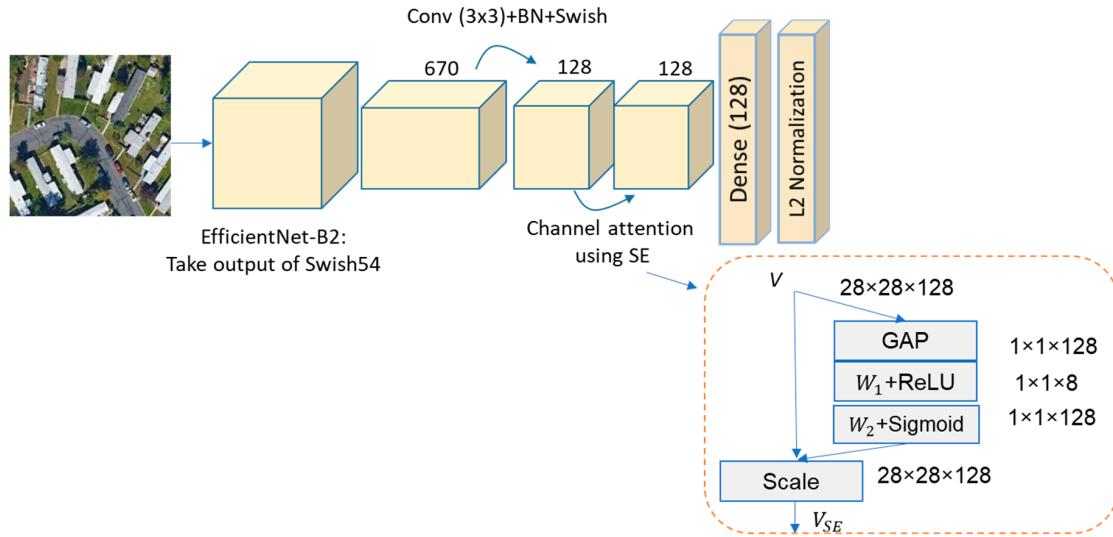


Figure 2. Image encoding branch for extracting the visual features.

2.2. Text Encoding Module

Figure 3 shows the text encoding module, which is composed of K symmetric branches, where each branch is used to encode one sentence describing the image content. These sub-branches use a word embedding layer followed by LSTM, a fully-connected layer, and l_2 -normalization.

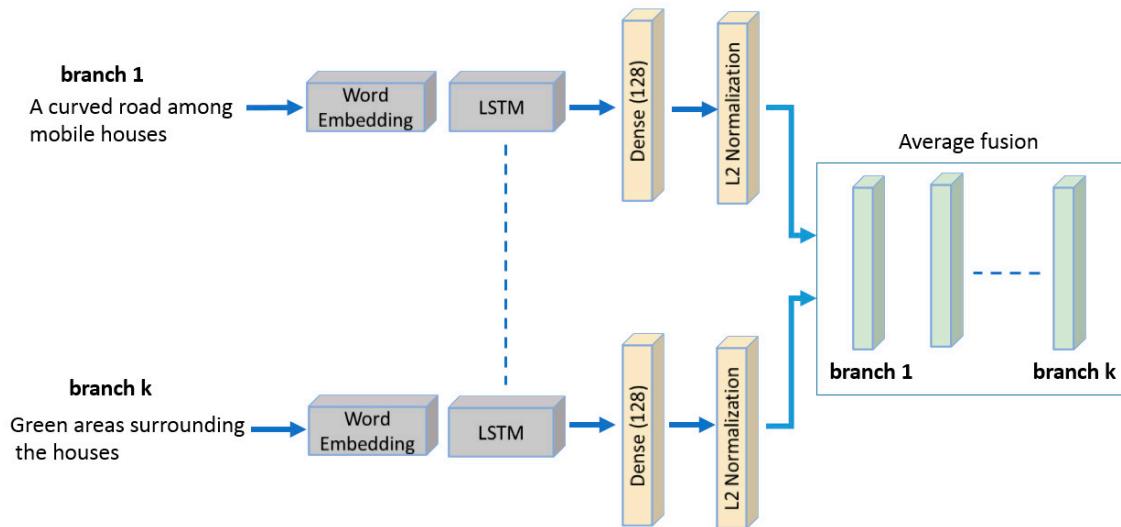


Figure 3. Text embedding branch: The five sentences describing the content of an image are aggregated using an average fusion layer. LSTM is Long-Short Term Memory.

The word embedding layer receives a sequence of integers representing the words in the sentence and transforms them into representations, where similar words should have similar encodings. Then the outputs of this layer are fed to LSTM [62] for modeling the entire sentence based on their long-term dependency learning capacity. Figure 4 shows the architecture of LSTM, with its four types of gates at each time step t in the memory cell. These gates are the input gate i_t , the update gate c_t , the output

gate o_t , and the forget gate f_t . For each time step, these gates receive as input the hidden state h_{t-1} and the current input y_t . Then, the cell memory recursively updates itself based on its previous values and forget and update gates.

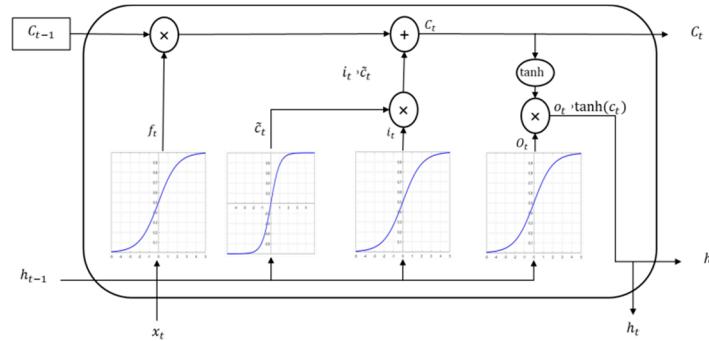


Figure 4. LSTM structure.

The working mechanism of LSTM is given below (for simplicity, we omit the image index i) [62]:

$$i_t = \text{sigmoid}(W_i \cdot [h_{t-1}, y_t]) \quad (3)$$

$$f_t = \text{sigmoid}(W_f \cdot [h_{t-1}, y_t]) \quad (4)$$

$$\tilde{c}_t = \tanh(W_g \cdot [h_{t-1}, y_t]) \quad (5)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (6)$$

$$o_t = \text{sigmoid}(W_o \cdot [h_{t-1}, y_t]) \quad (7)$$

$$h_t = o_t * \tanh(c_t) \quad (8)$$

where $*$ denotes the Hadamard product, and W_i , W_f , W_g , and W_o are learnable weights. In general, we can model the hidden state h_t of the LSTM as follows [62]:

$$h_t = \text{LSTM}(h_{t-1}, y_t, r_{t-1}) \quad (9)$$

where r_{t-1} indicates the memory cell vector at time step $t - 1$.

For each branch, the output of LSTM is fed to an additional fully-connected layer yielding K feature representation $g(y_i^k)$, $k = 1, \dots, K$. Then, the final outputs of different branches are fused using an average fusion layer to obtain a feature of dimension 128 [7]:

$$g(T_i) = \frac{\sum_{k=1}^K g(y_i^k)}{K} \quad (10)$$

2.3. DBTN Optimization

Many machine learning and computer vision problems are based on learning a distance metric for solving retrieval problems [63]. Inspired by achievements of deep learning in computer vision [26], deep neural networks were used to learn how to embed discriminative features [64,65]. These methods learn to project images or texts into a discriminative embedding space. The embedded vectors of similar samples are closer, while they are farther to those of dissimilar samples. Then several loss functions were developed for optimization such as triplet [65], quadruplet [66], lifted structure [67], N-pairs [68], and angular [69] losses. In this work, we concentrate on the triplet loss, which aims to learn a discriminative embedding for various applications such as classification [64], retrieval [70–74], and person re-identification [75,76]. It is worth recalling that a standard triplet in image-to-image retrieval is composed of three samples: an anchor, a positive sample (from the same category to the

anchor), and a negative sample (from the different category to the anchor). The aim of the triplet loss is to learn an embedding space, where anchor samples are closer to positive samples than to negative ones by a given margin.

In our case, the network is composed of asymmetric branches, unlike standard triplet networks, as the anchor; positive and negative samples are represented in a different way. For instance, triplets can be formed using a text as an anchor, its corresponding image as a positive sample in addition to an image with a different content image as a negative. Similarly, one can use an image as an anchor associated with positive and negative textual descriptions. The aim is to learn discriminative features for different textual descriptions and discriminative features for different visual features as well. In addition, we should learn similar features to each image and its corresponding textual representation. For such purpose, we propose a bidirectional triplet loss as a possible solution to the problem. The bidirectional triplet loss is given as follows:

$$l_{DBTN} = \lambda_1 L_1 + \lambda_2 L_2 \quad (11)$$

$$L_1 = \sum_{i=1}^N \left[\|g(T_i^a) - f(X_i^p)\|_2^2 - \|g(T_i^a) - f(X_i^n)\|_2^2 + \alpha \right]_+ \quad (12)$$

$$L_2 = \sum_{i=1}^N \left[\|f(X_i^a)\|_2^2 - \|f(X_i^a) - g(T_i^p)\|_2^2 - \|f(X_i^a) - g(T_i^n)\|_2^2 + \alpha \right]_+ \quad (13)$$

where $|z|_+ = \max(z, 0)$, and α is the margin that ensures the negative is farther away than the positive. $g(T_i^a)$ refers to the embedding of the anchor text, $f(X_i^p)$ is the embedding of the positive image, and $f(X_i^n)$ refers to the embedding of the negative image. On the other side, $f(X_i^a)$ refers to the embedding of the anchor image, $g(T_i^p)$ is the embedding of the positive text, and $g(T_i^n)$ refers to the embedding of the negative text. λ_1 and λ_2 are parameters of regularization controlling the contribution of both terms.

The performance of DBTN heavily relies on triplet selection. Indeed, the process of training is often so sensitive to the selected triplets, i.e., selecting the triplets randomly leads to non-convergence. To surmount this problem, the authors in [77] proposed triplet mining, which utilized only semi-hard triplets, where the positive pair was closer than the negative. Such valid semi-hard triplets are scarce, and therefore semi-hard mining requires a large batch size to search for informative pairs. A framework named smart mining was provided by Harwood et al. [78] to find out hard samples from the entire dataset that suffered from the burden of off-line computation. Wu et al. [79] discussed the significance of sampling and proposed a sampling technique called distance weighted sampling, which uniformly samples negative examples by similarity. Ge et al. [80] built a hierachal tree of all the classes to find out hard negative pairs, which were collected via a dynamic margin. In this paper, we proposed to use a semi-hard mining strategy, as shown in Figure 5, although other sophisticated selection mechanism could be investigated as well. In particular, we selected triplets in an online mode based on the following constraint [77]:

$$d(g(T^a), f(X^p)) < d(g(T^a), f(X^n)) < d(g(T^a), f(X^p)) + \alpha \quad (14)$$

where $d(\cdot)$ is the cosine distance.

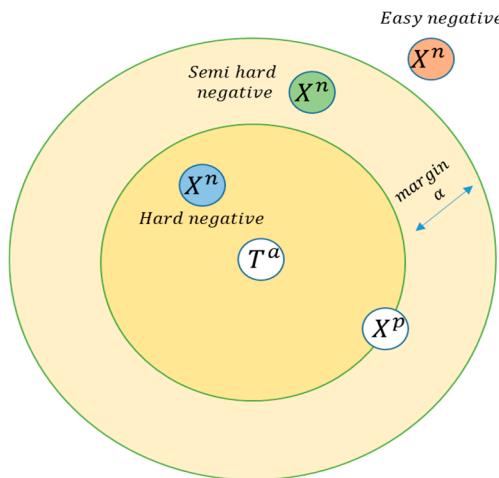


Figure 5. Semi-hard triplet selection scheme.

3. Experimental Results

3.1. Dataset Description

We built a dataset, named TextRS, by collecting images from four well-known different scene datasets, namely the AID dataset, which consists of 10,000 aerial images of size 600×600 pixels within 30 classes collected from Google Earth imagery by different remote sensors. The Merced dataset contains 21 classes; each class has 100 images of size 256×256 pixels with a resolution of 30 cm and RGB color. Such dataset was collected from USGS. The PatternNet was gathered from high-resolution imagery and includes 38 classes; each class contains 800 images of size 256×256 pixels. The NWPU dataset is another scene dataset, which has 31,500 images and is composed of 45 scene classes.

TextRS is composed of 2144 images selected randomly from the above four scene datasets. In particular, 480, 336, 608, and 720 images were selected from AID, Merced, PatternNet, and NWPU, respectively (16 images were selected from each class of such datasets). Then each remote sensing image was annotated by five different sentences; therefore, the total number of sentences was 10,720, and all the captions of this dataset were generated by five people to prove the diversity. It is worth recalling that the choice of the five sentences was mainly motivated by other datasets developed in the general context of computer vision literature [47,81]. During, the annotation we took into consideration some rules that had to be followed during generation of the sentences:

- Focus on the main dominating objects (tiny ones may be useless).
- Describe what exists instead of what does not exist in the scene.
- Try not to focus on the number of objects too much but use generic descriptions such as several, few, many, etc.
- Try not to emphasize the color of objects (e.g., blue vehicles) but rather on their existence and density.
- When mentioning, for instance, a parking lot (in an airport), it is important to mention the word ‘airport’ as well to distinguish it from any generic parking lot (downtown for example).
- Avoid using punctuation and conjunctions.

Some samples from our dataset are shown in Figure 6.

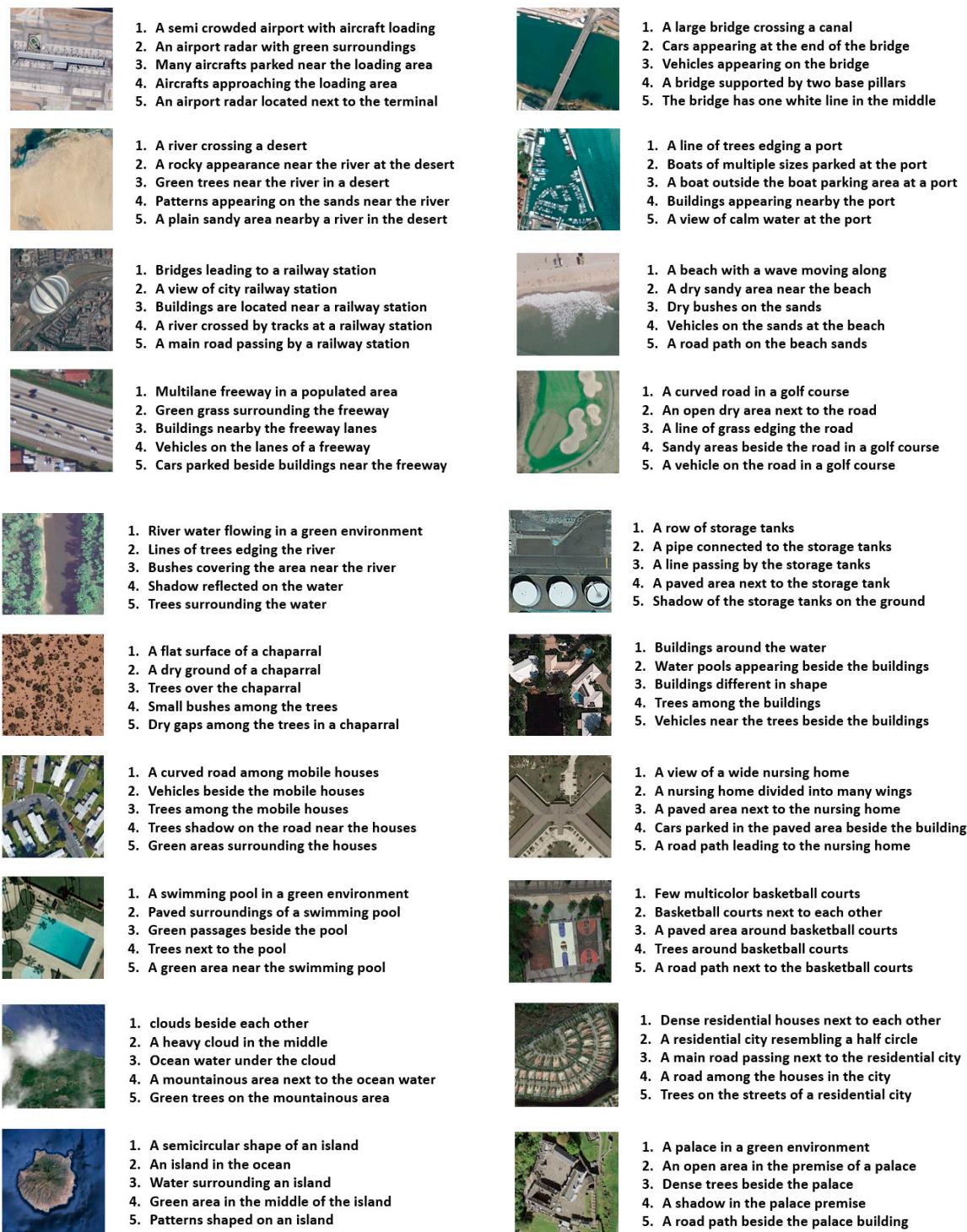


Figure 6. Example of images with the five sentences for each image.

3.2. Performance Evaluation

We implemented the method using the keras open-source library for deep learning written in python. For training the network, we randomly select 1714 images as training and the remaining 430 images as the test corresponding to approximately to 80% for training and 20% for testing. For training the DBTN, we used a mini-batch size of 50 images with the Adam optimization method with a fixed learning rate equal to 0.001 and exponential decay rates for the moment estimates equal to 0.9 and 0.999. Additionally, we set the regularization parameters to the default values of $\lambda_1 = \lambda_2 = 0.5$. To evaluate

the performance of the method, we used the wide recall measure, which is suitable for text-to-image retrieval problems. In particular, we presented the results in Recall@K (R@K) terms for different values of K (1, 5, 10), which are the percentage of ground-truth matches shown in the top K-ranked results. We conducted the experiments on a station with an Intel Core i9 processor with a speed of 3.6 GHz and 32 GB of memory, and a Graphical Processing Unit (GPU) with 11 GB of GDDR5X memory.

3.3. Results

As mentioned in the previous sections, we used four different pre-trained CNNs for the image encoding branch, which were EfficientNet, ResNet50, Inception_v3, and VGG16. Figure 7 illustrates the evolution of the triplet loss function during the training phase for these different networks. We can see that the loss function decreased gradually with an increase in the number of iterations. In general, the model reached stable values after 40 iterations. In Figure 8 we show examples of features obtained by the image and text encoding branches at the end of the training process.

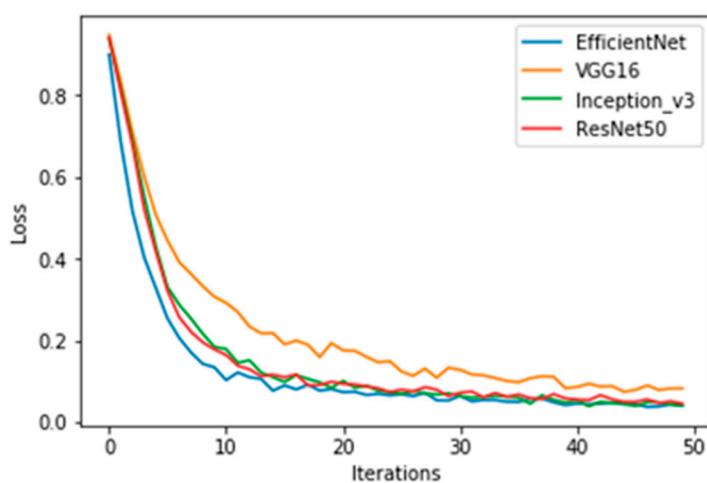


Figure 7. Evolution of loss function for the EfficientNet, ResNet50, Inception_v3, and VGG16.

Table 1 illustrates the performance of DBTN using EfficientNet as a pre-trained CNN for encoding the visual features. It could be observed with one sentence (Sent.1). The method achieved 13.02%, 40%, and 59.30% in R@1, R@5, and R@10, respectively. In contrast, when the five sentences are fused, the performance was further improved to 17.20%, 51.39%, and 73.02% of R@1, R@5, and R@10, respectively. Further, we computed the average of R@1, R@5, and R@10 for each sentence, and for fusion, we observed that the average of fusion had the highest score. Table 2 shows the results obtained using ResNet50 as the image encoder to learn the image features. We can see that the performances in R@1, R@5, and R@10 were 10.93%, 38.60%, and 54.41%, respectively, for Sent.1, while the method achieved 13.72%, 50.93%, and 69.06% of R@1, R@5, and R@10, respectively, with the fusion. Similarly, from Table 3 we observed that with Inception_v3, considering the fusion, the performance was also better than that of individual sentences. Finally, the results of using VGG16 are shown in Table 4. We can see that for Sent.1, our method achieved 10%, 36.27%, and 51.62% of R@1, R@5, and R@10, respectively, whereas the fusion process yielded 11.86%, 44.41%, and 63.72% of R@1, R@5, and R@10, respectively.

According to these preliminary results, one can notice that the fusing of the representations of the five sentences produced better matching results than did using one sentence. Additionally, EfficientNet seemed to be better compared to the other three pre-trained networks. This indicates that learning visual features by EfficientNet was quite effective and allowed better scores to be obtained compared to the other pre-trained CNNs.

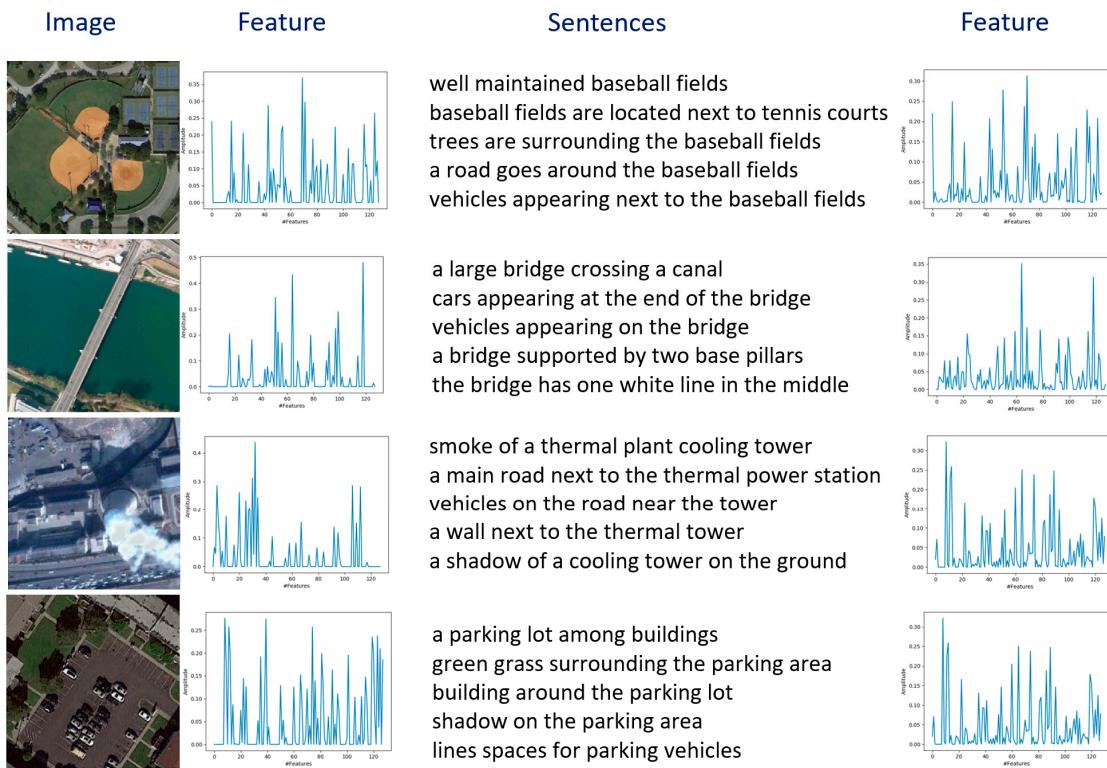


Figure 8. Image and text feature generated by the image and text encoding branches.

Table 1. Bidirectional text image matching results on our dataset by using EfficientNet-B2.

	Sent. 1	Sent. 2	Sent. 3	Sent. 4	Sent. 5	Fusion
R@1	13.02	13.48	14.18	13.02	10.09	17.20
R@5	40.00	44.18	44.18	40.93	38.60	51.39
R@10	59.30	60.46	62.55	57.67	58.37	73.02
Average	37.44	39.37	40.30	37.21	35.69	47.20

Table 2. Bidirectional text image matching results on our dataset by using ResNet50.

	Sent. 1	Sent. 2	Sent. 3	Sent. 4	Sent. 5	Fusion
R@1	10.93	12.79	12.32	12.32	11.86	13.72
R@5	38.60	38.37	42.58	43.02	38.19	50.93
R@10	54.41	56.27	61.16	60.93	55.58	69.06
Average	34.65	35.81	38.69	38.76	35.21	44.57

Table 3. Bidirectional text image matching results on our dataset by using Inception_v3.

	Sent. 1	Sent. 2	Sent. 3	Sent. 4	Sent. 5	Fusion
R@1	8.13	11.86	10.46	10.69	11.16	13.95
R@5	34.88	36.97	36.04	35.58	36.51	46.74
R@10	54.18	55.34	56.27	54.18	55.11	67.44
Average	32.40	34.72	34.26	33.48	34.26	42.71

Table 4. Bidirectional text image matching results on our dataset by using VGG16.

	Sent. 1	Sent. 2	Sent. 3	Sent. 4	Sent. 5	Fusion
R@1	10.00	9.06	11.86	8.13	7.67	11.86
R@5	36.27	35.11	36.51	34.41	33.25	44.41
R@10	51.62	51.16	56.51	51.16	47.90	63.72
Average	32.63	31.78	38.84	31.23	29.60	40.00

To analyze the performance in detail for image retrieval given a query text, we showed many successful and failure scenarios. For example, we could see (Figure 9) a given query text (five sentences) with its image, and the top nine relevant retrieved images (from left to right); the image in red box is the ground truth image of the query text (true match). We could observe that our method output reasonable relevant images, where all nine images had almost the same content (objects). In these four scenarios, the rank of the retrieved true images was 1, 6, and 1, respectively.

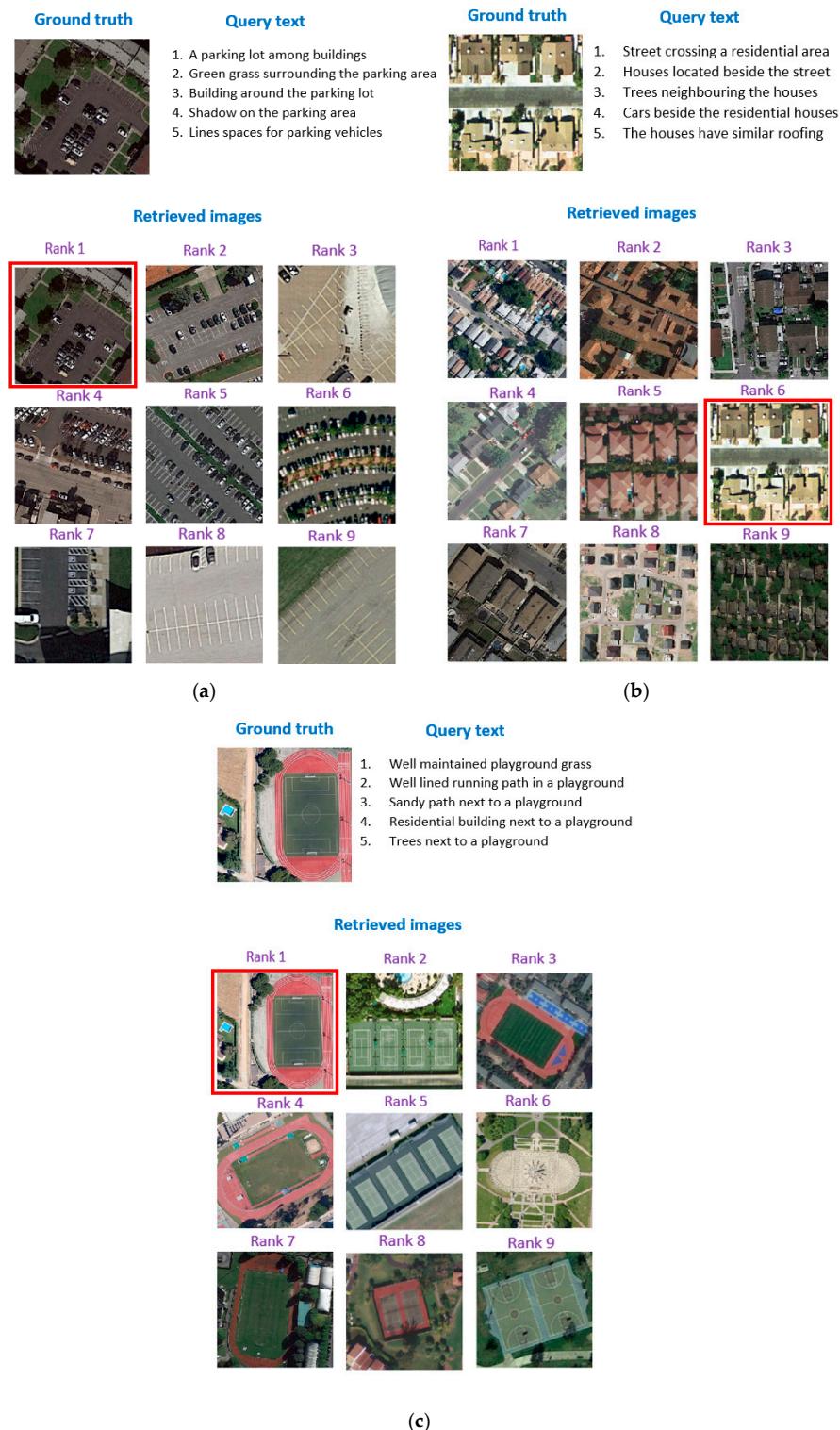


Figure 9. Successful scenarios (a, b and c) of text-to-image retrieval.

In contrast, Figure 10 shows two failure scenarios. In this case, we obtained relevant and irrelevant images, but the true matched image was not retrieved. This gives an indication that the problem was not easy and requires further investigations in improving the alignment of the descriptions to the image content.

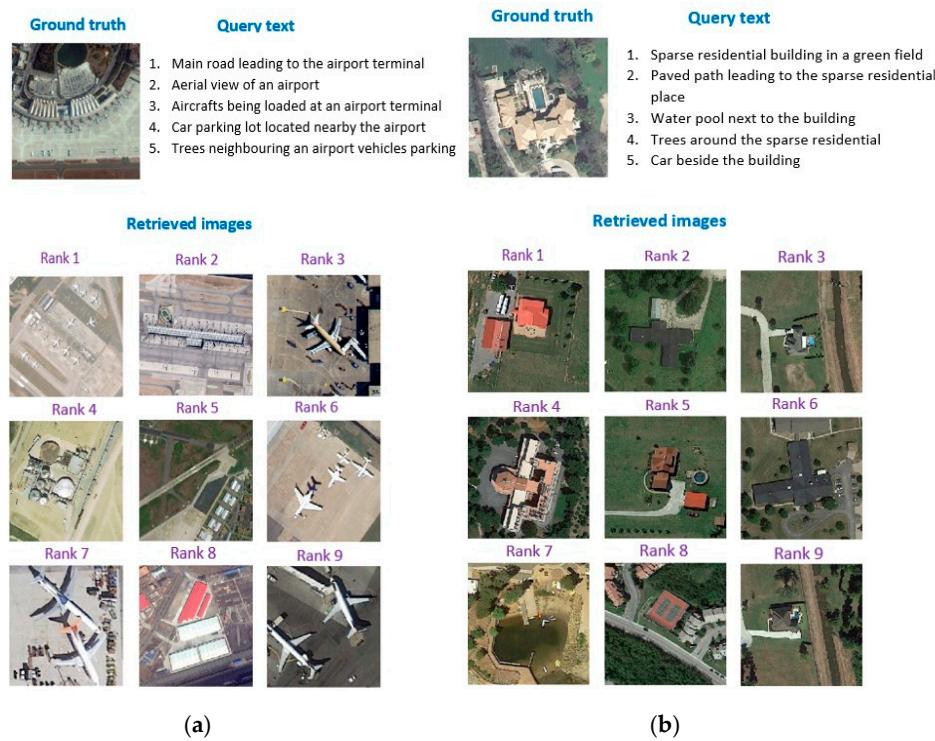


Figure 10. Unsuccessful scenarios (a and b) of text-to-image retrieval.

4. Discussion

In this section, we analyze further the performances of DBTN using different versions of EfficientNets, which are B0, B3, and B5. The version B0 contains 5.3 M parameters, while B3 and B5 are deeper and have 12M and 30M parameters, respectively. The results reported in Table 5 show that using B2 yields slightly better results compared to the other models. On the other side, B0 seems to be less competing as it provides an average recall of 45.65 compared to 47.20 for B2.

Table 5. Bidirectional text image matching results on our dataset using different EfficientNets.

	B0	B2	B3	B5
R@1	16.74	17.20	16.74	16.51
R@5	51.62	51.39	50.23	51.39
R@10	68.60	73.02	72.09	71.62
Average	45.65	47.20	46.35	46.51

Table 6 shows sensitivity analysis for bidirectional text image matching at multiple margin values. We can observe that setting this parameter to $\alpha = 0.5$ seems to be the most suitable choice. Increasing further this value leads to a decrease in the average recall as the network tends to select easy negative triplets.

In Table 7, we report the recall results obtained by using only one direction instead of bidirectional training. That is, we use text-to-image (Anchor text) and image-to-text (Anchor image). Obviously, the performance with bidirectional achieves the best results where relative similarity in one direction is useful for retrieval in the other direction, in the sense that the model trained with text-to-image triplets obtains a reasonable result in an image-to-text retrieval task and vice-versa. Nevertheless, the model

trained with bi-directional triplets achieves the best result, indicating that the triplets organized in bidirectional provide more overall information for text-to-image matching.

Table 6. Sensitivity with respect to the margin parameter α .

	α		
	0.1	0.5	1
R@1	13	17.20	5
R@5	37.67	51.39	22.09
R@10	54.18	73.02	37.83

Table 7. Comparison between unidirectional and bidirectional loss.

	Anchor Text	Anchor Image	Bidirectional
R@1	12.55	12.55	17.20
R@5	41.62	39.53	51.39
R@10	62.09	59.53	73.02

5. Conclusions

In this work, we proposed a novel DBTN architecture for matching textual descriptions to remote sensing images. Different from traditional remote sensing image-to-image retrieval, our network seeks to carry out a more challenging problem, which is text-to-image retrieval. Such a network is composed of an image and text encoding branches and is trained using a bidirectional triplet loss. In the experiments, we validated the method on a new benchmark data set termed TextRS. Experiments show in general promising results in terms of the recall measure. In particular, better recall scores were obtained by fusing the textual representations rather than using one sentence for each image. In addition, EfficientNets allows better visual representations to be obtained compared to the other pre-trained CNNs. For future developments, we propose to investigate image-to-text matching and propose advanced solutions based on attention mechanisms.

Author Contributions: T.A., Y.B. and M.M.A.R. designed and implemented the method, and wrote the paper. M.L.M., M.Z. and L.R. contributed to the analysis of the experimental results and paper writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research at King Saud University through the Local Research Group Program, grant number RG-1435-050.

Acknowledgments: This work was supported by the Deanship of Scientific Research at King Saud University through the Local Research Group Program under Project RG-1435-050.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Al Rahhal, M.M.; Bazi, Y.; Abdullah, T.; Mekhalfi, M.L.; AlHichri, H.; Zuair, M. Learning a Multi-Branch Neural Network from Multiple Sources for Knowledge Adaptation in Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1890. [[CrossRef](#)]
- Aptoula, E. Remote Sensing Image Retrieval With Global Morphological Texture Descriptors. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3023–3034. [[CrossRef](#)]
- Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 120–147. [[CrossRef](#)]
- Schroder, M.; Rehrauer, H.; Seidel, K.; Datcu, M. Interactive learning and probabilistic retrieval in remote sensing image archives. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 2288–2298. [[CrossRef](#)]
- Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 606–617. [[CrossRef](#)]
- Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [[CrossRef](#)]

7. Mekhalfi, M.L.; Melgani, F.; Bazi, Y.; Alajlan, N. Land-Use Classification With Compressive Sensing Multifeature Fusion. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2155–2159. [[CrossRef](#)]
8. Mekhalfi, M.L.; Melgani, F. Sparse modeling of the land use classification problem. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 3727–3730.
9. Weng, Q.; Mao, Z.; Lin, J.; Liao, X. Land-use scene classification based on a CNN using a constrained extreme learning machine. *Int. J. Remote Sens.* **2018**, *39*, 6281–6299. [[CrossRef](#)]
10. Wu, H.; Liu, B.; Su, W.; Zhang, W.; Sun, J. Deep Filter Banks for Land-Use Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1895–1899. [[CrossRef](#)]
11. Shao, Z.; Yang, K.; Zhou, W. Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sens.* **2018**, *10*, 964. [[CrossRef](#)]
12. Chaudhuri, B.; Demir, B.; Bruzzone, L.; Chaudhuri, S. Multi-label Remote Sensing Image Retrieval using a Semi-Supervised Graph-Theoretic Method. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1144–1158. [[CrossRef](#)]
13. Shao, Z.; Yang, K.; Zhou, W. Correction: Shao, Z.; et al. A Benchmark Dataset for Performance Evaluation of Multi-Label Remote Sensing Image Retrieval. *Remote Sens.* **2018**, *10*, 1200. [[CrossRef](#)]
14. Bosilj, P.; Aptoula, E.; Lefèvre, S.; Kijak, E. Retrieval of Remote Sensing Images with Pattern Spectra Descriptors. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 228. [[CrossRef](#)]
15. Sebai, H.; Kourgli, A.; Serir, A. Dual-tree complex wavelet transform applied on color descriptors for remote-sensed images retrieval. *J. Appl. Remote Sens.* **2015**, *9*, 095994. [[CrossRef](#)]
16. Bouteldja, S.; Kourgli, A. Multiscale texture features for the retrieval of high resolution satellite images. In Proceedings of the 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), London, UK, 10–12 September 2015; pp. 170–173.
17. Shao, Z.; Zhou, W.; Zhang, L.; Hou, J. Improved color texture descriptors for remote sensing image retrieval. *J. Appl. Remote Sens.* **2014**, *8*, 083584. [[CrossRef](#)]
18. Scott, G.J.; Klaric, M.N.; Davis, C.H.; Shyu, C. Entropy-Balanced Bitmap Tree for Shape-Based Object Retrieval From Large-Scale Satellite Imagery Databases. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 1603–1616. [[CrossRef](#)]
19. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
20. Bay, H.; Tuytelaars, T.; Gool, L.V. SURF: Speeded Up Robust Features. In Proceedings of the Computer Vision-ECCV 2006, Berlin, Heidelberg, 7–13 May 2006; Springer; pp. 404–417.
21. Yang, J.; Liu, J.; Dai, Q. An improved Bag-of-Words framework for remote sensing image retrieval in large-scale image databases. *Int. J. Digit. Earth* **2015**, *8*, 273–292. [[CrossRef](#)]
22. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE computer society conference on computer vision and pattern recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
23. Zhang, F.; Du, B.; Zhang, L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [[CrossRef](#)]
24. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [[CrossRef](#)]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
27. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:14091556 Cs. Available online: <https://arxiv.org/abs/1409.1556> (accessed on 24 January 2020).
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [[CrossRef](#)] [[PubMed](#)]
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
31. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
32. Han, W.; Feng, R.; Wang, L.; Cheng, Y. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 23–43. [CrossRef]
33. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef] [PubMed]
34. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
35. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 2048–2057.
36. Cao, R.; Zhang, Q.; Zhu, J.; Li, Q.; Li, Q.; Liu, B.; Qiu, G. Enhancing remote sensing image retrieval using a triplet deep metric learning network. *Int. J. Remote Sens.* **2020**, *41*, 740–751. [CrossRef]
37. Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Unsupervised Deep Feature Learning for Remote Sensing Image Retrieval. *Remote Sens.* **2018**, *10*, 1243. [CrossRef]
38. Yang, J.; Liang, J.; Shen, H.; Wang, K.; Rosin, P.L.; Yang, M.-H. Dynamic Match Kernel With Deep Convolutional Features for Image Retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 5288–5302. [CrossRef]
39. Li, Y.; Zhang, Y.; Huang, X.; Zhu, H.; Ma, J. Large-Scale Remote Sensing Image Retrieval by Deep Hashing Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 950–965. [CrossRef]
40. Li, P.; Ren, P. Partial Randomness Hashing for Large-Scale Remote Sensing Image Retrieval. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 464–468. [CrossRef]
41. Li, Y.; Zhang, Y.; Huang, X.; Ma, J. Learning Source-Invariant Deep Hashing Convolutional Neural Networks for Cross-Source Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6521–6536. [CrossRef]
42. Chaudhuri, B.; Demir, B.; Bruzzone, L.; Chaudhuri, S. Region-Based Retrieval of Remote Sensing Images Using an Unsupervised Graph-Theoretic Approach. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 987–991. [CrossRef]
43. Zhou, W.; Deng, X.; Shao, Z. Region Convolutional Features for Multi-Label Remote Sensing Image Retrieval. *arXiv* **2018**, arXiv:180708634 Cs. Available online: <https://arxiv.org/abs/1807.08634> (accessed on 24 January 2020).
44. Dai, O.E.; Demir, B.; Sankur, B.; Bruzzone, L. A Novel System for Content-Based Retrieval of Single and Multi-Label High-Dimensional Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2473–2490. [CrossRef]
45. Wu, Q.; Shen, C.; Liu, L.; Dick, A.; Hengel, A.v.d. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 203–212.
46. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:14061078 Cs Stat. Available online: <https://arxiv.org/abs/1406.1078> (accessed on 24 January 2020).
47. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 652–663. [CrossRef] [PubMed]
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
49. Fang, H.; Gupta, S.; Iandola, F.N.; Srivastava, R.K.; Deng, L.; Dollar, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. From captions to visual concepts and back. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.

50. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image Captioning with Semantic Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vega, NV, USA, 27–30 June 2016; pp. 4651–4659.
51. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.
52. Huang, F.; Zhang, X.; Li, Z.; Zhao, Z. Bi-directional Spatial-Semantic Attention Networks for Image-Text Matching. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **2018**, *28*, 2008–2020. [CrossRef] [PubMed]
53. Wang, L.; Li, Y.; Lazebnik, S. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 394–407. [CrossRef]
54. Zhang, Y.; Lu, H. Deep Cross-Modal Projection Learning for Image-Text Matching. In Proceedings of the European Conference on Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Springer; pp. 686–701.
55. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting image captioning with attributes. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4904–4912.
56. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
57. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
58. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description Generation for Remote Sensing Images Using Attribute Attention Mechanism. *Remote Sens.* **2019**, *11*, 612. [CrossRef]
59. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
60. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
61. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946 Cs Stat. Available online: <https://arxiv.org/abs/1905.11946> (accessed on 24 January 2020).
62. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
63. Weinberger, K.Q.; Saul, L.K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.* **2009**, *10*, 207–244.
64. Wang, J.; Song, Y.; Leung, T.; Rosenberg, C. Learning Fine-Grained Image Similarity with Deep Ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1386–1393.
65. Hoffer, E.; Ailon, N. Deep Metric Learning Using Triplet Network. In Proceedings of the Similarity-Based Pattern Recognition, Copenhagen, Denmark, 12–14 October 2015; Feragen, A., Pelillo, M., Loog, M., Eds.; Springer International Publishing: Cham, Switzerland; pp. 84–92.
66. Law, M.T.; Thome, N.; Cord, M. Quadruplet-Wise Image Similarity Learning. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 249–256.
67. Oh Song, H.; Xiang, Y.; Jegelka, S.; Savarese, S. Deep Metric Learning via Lifted Structured Feature Embedding. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4004–4012.
68. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016; pp. 1857–1865.
69. Wang, J.; Zhou, F.; Wen, S.; Liu, X.; Lin, Y. Deep Metric Learning with Angular Loss. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2612–2620.
70. Huang, J.; Feris, R.; Chen, Q.; Yan, S. Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1062–1070.

71. Lai, H.; Pan, Y.; Liu, Y.; Yan, S. Simultaneous Feature Learning and Hash Coding With Deep Neural Networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3270–3278.
72. Zhuang, B.; Lin, G.; Shen, C.; Reid, I. Fast Training of Triplet-Based Deep Binary Embedding Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5955–5964.
73. Gordo, A.; Almazan, J.; Revaud, J.; Larlus, D. Deep image retrieval: Learning global representations for image search. VI. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9910, pp. 241–257.
74. Yuan, Y.; Yang, K.; Zhang, C. Hard-Aware Deeply Cascaded Embedding. In Proceedings of the IEEE international conference on computer vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 814–823.
75. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference (BMVC); British Machine Vision Association: Swansea, September, 2015; pp. 41.1–41.12.
76. Wang, L.; Li, Y.; Lazebnik, S. Learning Deep Structure-Preserving Image-Text Embeddings. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5005–5013.
77. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
78. Harwood, B.; VijayKumar, B.G.; Carneiro, G.; Reid, I.; Drummond, T. Smart Mining for Deep Metric Learning. In Proceedings of the IEEE international conference on computer vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2821–2829.
79. Wu, C.-Y.; Manmatha, R.; Smola, A.J.; Krähenbühl, P. Sampling Matters in Deep Embedding Learning. In Proceedings of the IEEE international conference on computer vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2840–2848.
80. Ge, W.; Huang, W.; Dong, D.; Scott, M.R. Deep Metric Learning with Hierarchical Triplet Loss. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 269–285.
81. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; IEEE; pp. 2641–2649.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).