applied sciences

MDPI

Article

Deep Unsupervised Embedding for Remote Sensing Image Retrieval Using Textual Cues

Mohamad M. Al Rahhal [1], Yakoub Bazi [2,*], Taghreed Abdullah [3], Mohamed L. Mekhalfi [4] and Mansour Zuair [2]

1 Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia; mmalrahhal@ksu.edu.sa
2 Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; zuair@ksu.edu.sa
3 Department of Studies in Computer Science, University of Mysore, Mysore 570006, India; taghreed@compsci.uni-mysore.ac.in
4 Department of Information Engineering and Computer Science, University of Trento, 28123 Trento, Italy; mohamed.mekhalfi@alumni.unitn.it
* Correspondence: ybazi@ksu.edu.sa; Tel.: +96-610-1469-6297

Received: 20 November 2020; Accepted: 9 December 2020; Published: 14 December 2020

Abstract: Compared to image-image retrieval, text-image retrieval has been less investigated in the remote sensing community, possibly because of the complexity of appropriately tying textual data to respective visual representations. Moreover, a single image may be described via multiple sentences according to the perception of the human labeler and the structure/body of the language they use, which magnifies the complexity even further. In this paper, we propose an unsupervised method for text-image retrieval in remote sensing imagery. In the method, image representation is obtained via visual Big Transfer (BiT) Models, while textual descriptions are encoded via a bidirectional Long Short-Term Memory (Bi-LSTM) network. The training of the proposed retrieval architecture is optimized using an unsupervised embedding loss, which aims to make the features of an image closest to its corresponding textual description and different from other image features and vise-versa. To demonstrate the performance of the proposed architecture, experiments are performed on two datasets, obtaining plausible text/image retrieval outcomes.

Keywords: remote sensing; big transfer (BiT); text-to-image retrieval; bidirectional long short-term memory network (B-LSTM); unsupervised embedding

1. Introduction

Remote sensing refers to remotely acquiring, interpreting, and possibly pinpointing information about the changes and manifestations that the earth's surface undergoes. It has been possible via observation platforms such as satellites and aerial systems. The significance of remote sensing has seen a rapid rise in the amount of data in several civilian and military applications.

The potential of remote sensing technology has been used in a variety of applications, including environmental assessment and monitoring, precision agriculture, renewable natural resources, military surveillance, meteorology, mapping, and reconnaissance [1]. Information in these applications is acquired via sensors mounted on large satellites, medium aerial vehicles, or even miniaturized drones (which can either be passive or active) [2].

The availability of remote sensing data, especially high-resolution images, has stimulated research in the remote sensing community. Typically, the primary research focus is on image classification and image retrieval [3–5]. Image classification is important because it allows the determination (and often

the spatial localization) of certain objects across a given image volume. Image classification schemes typically attempt to tie meaningful feature representations (i.e., either spatial and/or spectral) to an efficient classification model, which has been shown to produce plausible results. On the other hand, image retrieval retrieves a query image (or an ensemble thereof) from a large-scale archive of image data. Efficient image retrieval methods do not only aim to achieve high retrieval scores but also to cope with the processing overheads due to the size of the retrieval archive. Furthermore, the advent of powerful processing machines has paved the way for deep architectures that have significantly improved the performance of remote sensing image classification and retrieval [6–12].

Traditional image classification schemes, owing to their discrete nature (i.e., single labeling), remain inefficient with respect to the abundant visual information an image may uncover (i.e., where visual content outweighs discrete semantic attributes). As opposed to traditional single-label image classification, image multi-labeling balances the visual attributes of images with respect to their semantic tenor, which is commonly termed the semantic gap. To address this problem, the computer vision community has focused on developing multi-label approaches to reduce the semantic gap and improve the retrieval efficiency. For example, in [13], a multi-label technique for detecting an indoor scene was proposed to help blind people sense nearby objects around them. This technique gave rise to similar approaches in remote sensing [14,15]. Inspired by this, another work [16] proposed a multi-label for remote sensing image retrieval archive, where one or more labels based on visual inspection have been assigned for each image in the archive is manually. However, coherence between the labels themselves remains questionable because multiple labels are assigned to an image without taking in consecration their semantic reciprocity. Therefore, incorporating this element (i.e., relativity among the objects) can elevate the quality of semantic descriptions.

It is evident that image matching has been a major interest [17–19]. However, there have been limited works on text image matching in the field of remote sensing, possibly because of the complexity of remote sensing images. One of the few examples include the work in [18], which is similar to that in [20], that uses convolutional neural networks (CNNs) to extract image descriptors and a sequence long short term memory (LSTM) model to generate a sentence describing the remote sensing images.

Recent works from the general literature have suggested that by addressing the retrieval problem from a cross-modal text-image, instead of image-to-image, the learning perspective seems to offer a more practical solution. This concept has attracted much attention in recent studies, particularly in image captioning, where images are combined with natural language processing. Essentially, it generates a sequential text that describes visual data, which is similar to how humans realize it. For example, in [17], a semantic attention model was proposed to enhance image captioning via convolutional neural network (CNN) and recurrent neural network (RNNs) models to obtain high-level attributes from images, which has improved the description generation of the image and has produced interesting experimental results. Similarly, CNNs and RNNs were combined for image captioning in [20]. On the other hand, a different graph-based pipeline was proposed in [21].

Other studies have mainly tackled the problem of learning common representations that enable the textual description of visual data and vice versa [22–25]. Frome et al. [22] proposed a deep visual semantic embedding framework to use semantic information and labeled images to recognize visual objects. Karpathy and Li [23] developed a deep model by using inferred latent alignment between the region of the image and the segment of the text that describes it to generate a description of image regions. Wang et al. [24] developed an approach for image and text embedding by using a dual-branch deep network with multiple layers. Wang et al. [25] implemented a two-branch deep network for text-image embedding as well as a similarity network to determine the correspondence of an image-text pair. Yao et al. [26] proposed a hashing technique to build a discrete supervised hashing model for cross-modal retrieval to learn a potent similarity matrix. Kiros et al. [27] proposed a framework that utilizes LSTM to represent the sentence, CNN to learn the representation of the image, and triplet loss to push the matched image-sentence pair closer to each other than the unmatched pairs in the embedding space. Huang et al. [28] introduced a multi-regional multi-label CNN to learn image

semantic concepts and LSTM for representing the sentence. Niu et al. [29] adopted a tree-structured LSTM to learn the hierarchical relations between images and sentences, in addition to learning the relation between phrases and visual objects. Zhang et al. [30] developed a cross-modal projection matching/classification using a CNN to encode visual image features and LSTM to extract text features.

It appears clearly from the literature of computer vision that the core problem in image retrieval is given a query one is interested to retrieve the most similar image in the database. This query can be an image or textual descriptions or a combination of them. Compared to image-image retrieval, text-image retrieval has been less investigated in the remote sensing community, possibly due to the complexity of appropriately tying textual data to respective visual representations. To cope with these limitations, the authors in [31] have proposed a new dataset for text-to-image matching named TextRS. They used a Deep Bidirectional Triplet Network (DBTN) for matching text to images based on CNN and LSTM networks. In this work, we propose an alternative approach based on asymmetric Siamese network. The first branch of this network uses BiT models for image representation, while the second branch relies on bidirectional Long Short-Term Memory (Bi-LSTM) for text encoding. The image and text representations are normalized and projected in a low dimensional space. The embedding features of the pair image-text should be invariant, while the features of different images and text instances should spread-out. The experimental results obtained on TextRS and Merced datasets are reported and discussed.

The paper is structured as follows. In Section 2, we introduce the proposed methodology. While, in Section 3, we present the experimental results. Finally, conclusion and future developments for the proposed work are declared in Section 4.

2. Proposed Methodology

The proposed architecture addresses the task of text-to-image matching. This task aims to retrieve the matching image/sentence that resides in a training set $\mathcal{D}$ prepared offline.

Let us assume a training set $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{N}$ consisting of N images alongside their respective sentences. In the test phase, given a query sentence t_q, we aim to retrieve the most relevant image from the training set $\mathcal{D}$ (Figure 1a). On the side, in the image-to-text retrieval scenario, however, a query image is presented at the text-to-image matching model to retrieve the most likely textual description (Figure 1b). Figure 2 gives the detailed architecture of the proposed method, which is divided into two branches for learning appropriate image and text embedding, i.e., $f(X_i)$ and $g(T_i)$. Further details are provided in the subsequent sections.

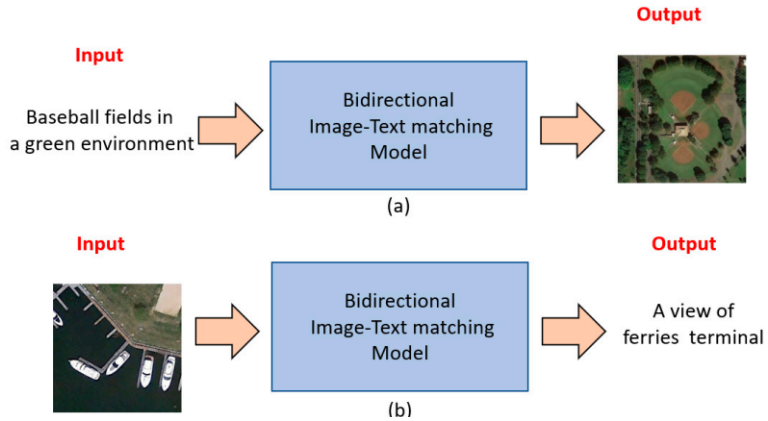

Figure 1. Bidirectional text-image retrieval: (**a**) text-image retrieval; (**b**) image-text retrieval.

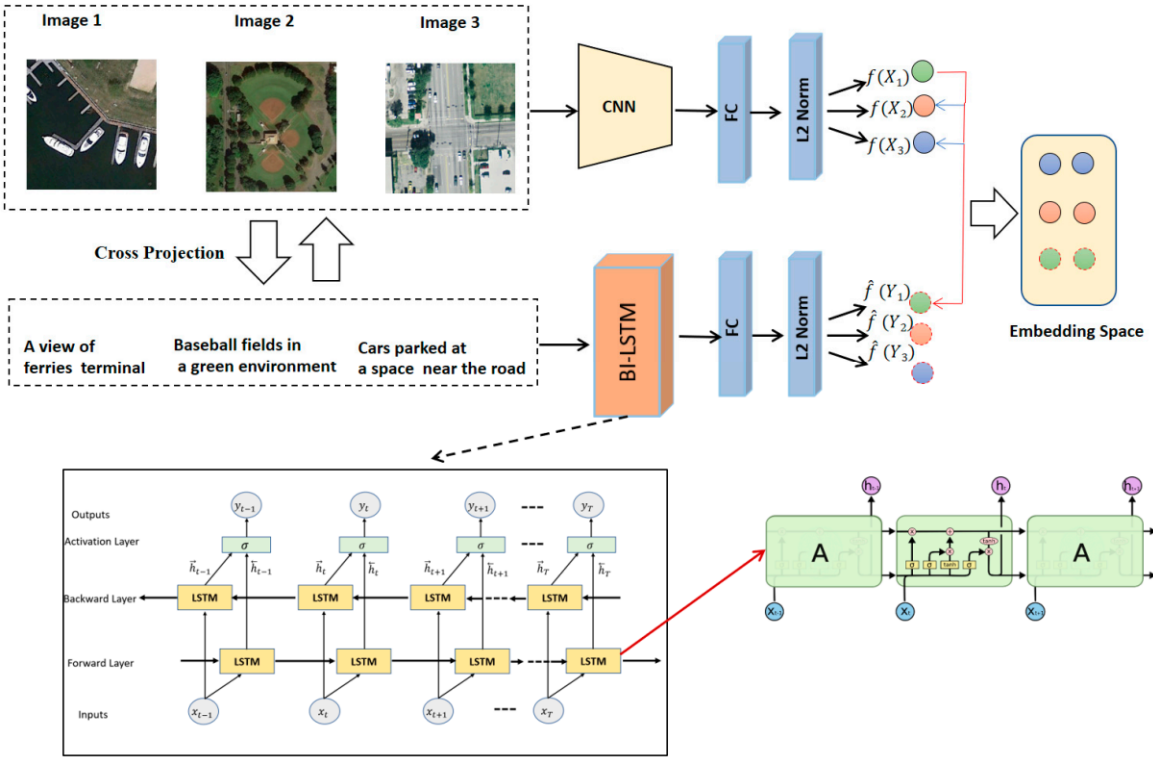

Figure 2. Flowchart of the proposed method.

2.1. Image Representation Using (BiT) Models

The backbone of image-embedding module is based on BiT models (i.e., BiT-S, BiT-M, and BiT-L) [32]. These BiT models are trained on three upstream datasets with different scales: ImageNet-1k [33] (BiT-S), ImageNet-21k (BiT-M), and JFT-300 M [34] (BiT-L). While ImageNet-1k is a dataset designed for ILSVRC image classification task, which is composed of more than 1.28 M images and 1k classes. On the other hand, ImageNet-21k is a large-scale few-shot dataset that contains 14 M images and 21k classes; it is also called the full ImageNet. The JFT-300 M dataset is a subsequent version of the dataset introduced in [35,36]. It has 300 M real-world images and 18k classes with each image approximately having 1.26 labels, resulting in a total of 375 M labels. Note that these BiT models yield state-of-the-art performances on several benchmark datasets for transfer learning with 928 M parameters.

BiT models adopt a standard ResNet-v2 [37] architecture with different sizes (i.e., a ResNet-50 (R50 × 1), a ResNet-50 that is three times wider (R50 × 3), a ResNet-101 × 1, a ResNet-101 that is three times wider (R101 × 3), and a ResNet-152 that is four times wider (R152 × 4)), and some updates. Unlike the standard ResNet architecture, BiT models are based on the idea that two new layers, called group normalization (GN) and weight normalization (WN), supersede batch normalization (BN) in all convolutional layers (Figure 3). Note that there are other normalization methods, such as layer norm (LN) and instance norm (IN), which can be considered as extreme cases of GN. Figure 4 illustrates the various normalization techniques and the relations between BN, LN, IN, and GN.

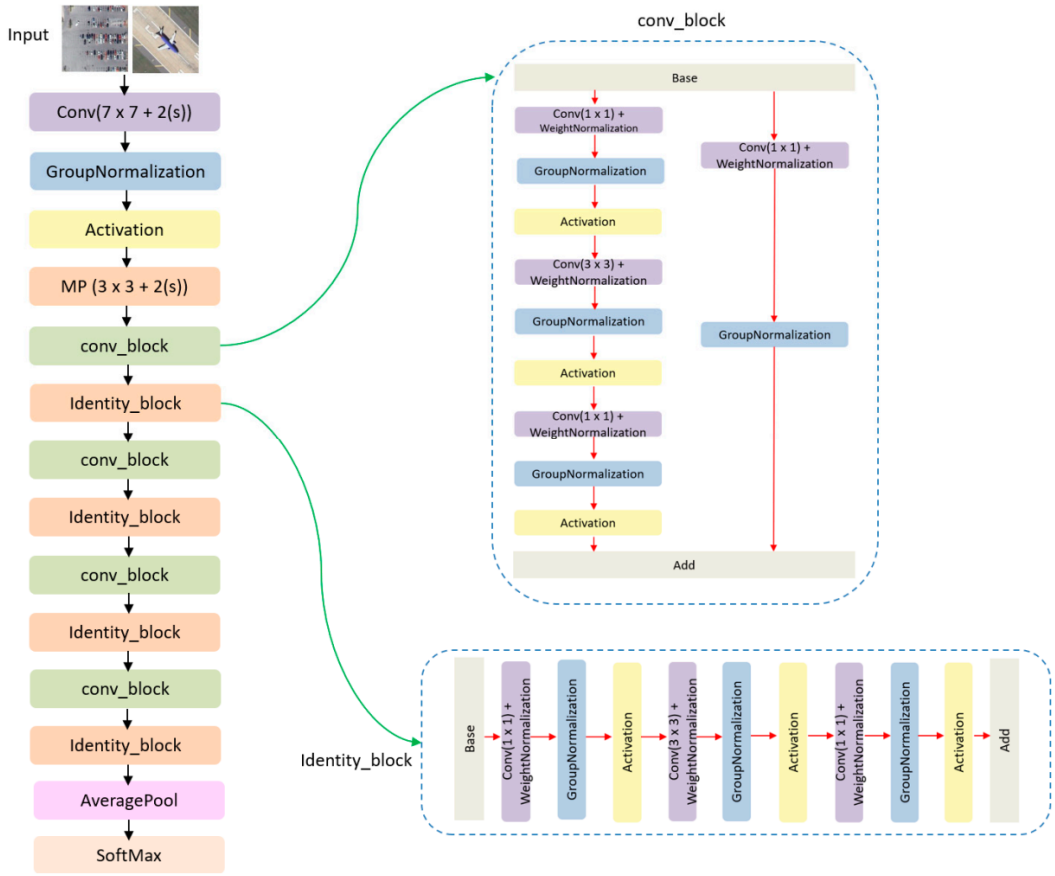

Figure 3. Architecture of Big transfer (BiT) model.

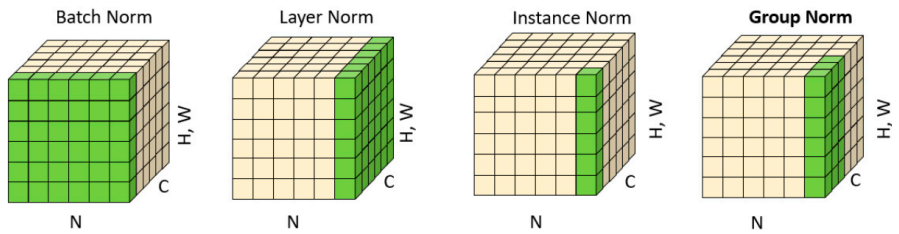

Figure 4. Demonstration of various normalization techniques.

GN has proven to be effective in many applications, such as detection and segmentation [38] and video classification [39], which makes it a better alternative for BN. It has been proposed with a layer that divides the channels into groups, and then the normalized features in each group according to the mean and the variance of the group. It has been shown that GN is more stable than BN with respect to batch size. This is because in GN, calculations of the batch statistics are inherently avoided. Furthermore, the batch dimension is not exploited in GN and it can be transferred to fine-tuning from pre-training irrespective of batch size changes.

In this work, we use these BiT models for image representation. In particular, we feed the image as input to the network and extract the corresponding feature representation before the classification layer. Then these feature are further projected and normalized into a low dimensional feature space using a dense layer as shown in Figure 2.

2.2. Text Representation Using Bi-LSTM

The sentence is fed through a word embedding layer followed by Bi-LSTM. Note that Bi-LSTM [40] is an extension of conventional LSTMs, while an LSTM network is a modified version of RNNs.

These last are based on the idea of connecting current and previous information, which enables understanding of the sequence of data [41]. RNNs tries to remember information they learned during the training procedure as well as what they learned from the previous input. This is achieved by repeatedly applying transformations to the input sequence data. After the output has been generated, it is copied and returned to the recurrent network.

Although RNNs have been used in various tasks, they encounter major problems such as gradient vanishing and exploding. To cope with these limitation the LSTM network [42] has been introduced as an alternative solution. The LSTM network depends on the so-called memory cell, which can learn (make decisions) how to allow data to be entered, left, or removed from the cell state through an iterative process. This is done at a time step through special structures called gates. The LSTM has three gates, i.e., input, forget, and output gates. The input gate controls how the input data would change the state of the memory cell. The forget gate controls the cell of the previous state, whether it has to remember or forget it. The output gate enables the memory cell to influence the outputs. The equations for the gates are as follows:

As mentioned previously, in this work, we use Bi-LSTM which comprises two LSTMs. During the training process, one of the LSTMs is trained on the input sequence and the other one is trained on the reversed copy of the input sequence. In other words, the input sequence is processed in both forward (past to future) and backward (future to past) directions by using two recurrent networks (two separate hidden layers, a forward state sequence and a backward state sequence). Both of the networks then connect to the same output layer to generate the output. Similar to the image branch, the output of Bi-LSTM is fed to a fully connected layer followed by l_2-normalization yielding a feature representation $f(Y_i)$ for the sentence Y_i.

2.3. Optimization

Retrieval tasks are usually solved by learning a distance metric [43]. Inspired by accomplishments of deep learning in computer vision [26], deep neural networks have been used to learn how to embed discriminative features useful for learning these distances [44,45]. In this case, the embedded features of similar samples should be closer, while those of dissimilar samples should be farther. The literature of computer vision convoys several loss functions such as triplet [45], quadruplet [46], lifted structure [47], N-pairs [48], and angular [49] losses. In this work, we extend the softmax embedding variant mainly proposed for image-to-mage retrieval to the case of text/image retrieval. Because of memory requirements, we learn iteratively these distances on small batches sampled from the complete dataset. The aim is to make the features of an image closest to its corresponding textual description and different from other image features and vise-versa.

If we consider $B_k = \{X_i, Y_i\}_{i=1}^{m}$ as the k th mini-batch of size m sampled instances from the full dataset $\mathcal{D}$, then for a sentence Y_i, its corresponding image X_i should be classified into instance i, while other images X_j, with $j \neq i$ are not classified into instance i. Therefore, the probability of the image X_i being recognized as instance i is defined by:

$$P(i|X_i) = \frac{exp\left(f^T(Y_i)f(X_i)/\tau\right)}{\sum_{k=1}^{m} exp(f^T(Y_k)f(X_i)/\tau)} \tag{1}$$

where τ is a temperature parameter controlling the sharpness of the distribution. Similarly, the probability of a sentence Y_j not being assigned to instance i can be defined by $1 - P\left(i\middle|Y_j\right)$ where

$$P\left(i\middle|Y_j\right) = \frac{exp\left(f^T(Y_i)f\left(Y_j\right)/\tau\right)}{\sum_{k=1}^{m} exp\left(f^T(Y_k)f\left(X_j\right)/\tau\right)},\ k \neq j \tag{2}$$

Then under the assumption that different sentences being recognized as instance i are independent, the joint probability image X_i being recognized as instance i and the sentence Y_j not assigned to instance i where $j \neq i$ is simply $P(i|X_i) \prod_{j \neq i} \left(1 - P\left(i \middle| Y_j\right)\right)$.

Then the corresponding negative log-likelihood for the min-batch B_k can be given as follows:

$$J_i = -log(P(i|X_i)) - \sum_{j \neq i} log\left(1 - P\left(i \middle| Y_j\right)\right) \tag{3}$$

The corresponding total log-likelihood J over the entire dataset is:

$$J_{Sentence} = -\sum_i log(P(i|X_i)) - \sum_i \sum_{j \neq i} log\left(1 - P\left(i \middle| Y_j\right)\right) \tag{4}$$

Similarly, we can extend this formulation in the reverse direction that form image X_i to sentence Y_i yielding the following log-likelihood:

$$J_{Image} = -\sum_i log(P(i|Y_i)) - \sum_i \sum_{j \neq i} log\left(1 - P\left(i \middle| X_j\right)\right) \tag{5}$$

Then the total log-likelihood used as loss for learning the parameters of the proposed asymmetric Siamese network is given by:

$$J = \lambda_1 J_{Sentence} + \lambda_2 J_{Image} \tag{6}$$

where λ_1 and λ_2 are balancing weights.

3. Experimental Results

3.1. Dataset Description

In this work we use two different benchmark datasets to validate the performances of the proposed method. The first one is the TextRS dataset [50], which consists of 2144 images collected from several scene datasets (i.e., AID [31], Merced [51], PatternNet [52], and NWPU [53]) (see Figure 5). In particular, this dataset is built by selecting randomly 16 images from each class of four popular heterogeneous scene datasets: AID (30 classes), UC Merced (21 classes), PatternNet (38 classes), and NWPU (45 classes). TextRS has 2144 images: 720 images with spatial resolution 0.2 to 30 m, 608 images with spatial resolution 0.062 to 4.7 m, 480 images with spatial resolution 0.5 to 8 m, and 336 images with spatial resolution 30 m as shown in Table 1. Each remote sensing image is annotated by five various sentences; therefore, the total number of sentences is 10,720. The second dataset is the Merced Land-Use dataset, which consists of remote sensing images with 21 classes [54]. Each class has 100 RGB images (256×256 pixels). The total number of images is 2100, with every image labeled also with five different sentences (see Figure 5b).

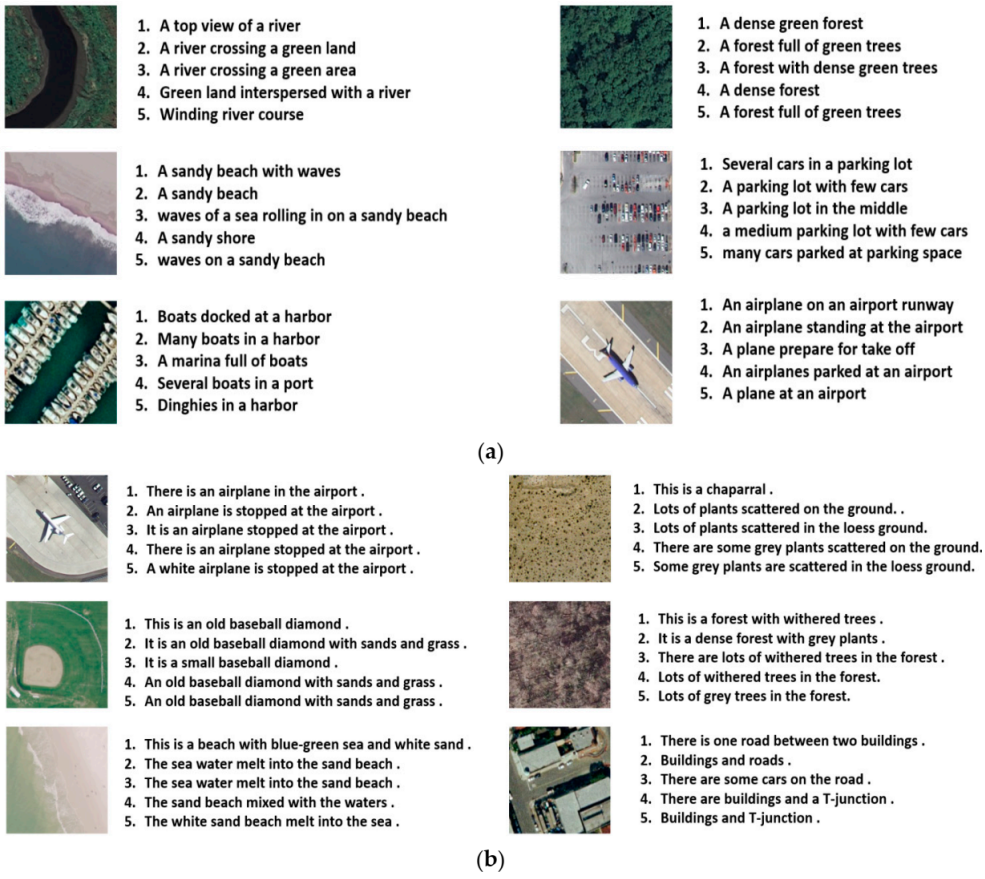

Figure 5. Example of images with five sentences in the following datasets: (**a**) TextRS and (**b**) Merced.

Table 1. TextRS Images Distribution.

No. of Images	Images from	No. of Sentences	Spatial Resolution	Image Size
720	NWPU	3600	0.2 to 30 m	256 × 256
608	PatternNet	3040	0.062 to 4.7 m	256 × 256
480	AID	2400	0.5 to 8 m	600 × 600
336	UC Merced	1680	30 m	256 × 256
Total: 2144		Total: 10,720		

3.2. Experimental Setup

We implemented the proposed method using Tensorflow-keras. We divided the TextRS dataset according to [32], while we randomly split the Merced dataset into: 80% is for training and 20% is for testing. We use a mini-batch size of 50 images for training the network. For optimization, we use Adam optimizer with its default parameters. In addition, we set the regularization parameters λ_1 *and* λ_2 controlling the contribution of the two losses to 1. We train the models for 50 iterations. For performance evaluation of the proposed method, we use the Recall@K (R@K) metric, which is widely used to match the scores of an image with a query sentence and vice versa. In the subsequent sections, we present the results in terms of R@K for different values of (K = 1, 5, and 10) calculated as follows:

$$\text{R@k} = (\text{true positives @k})/((\text{true positives @k}) + (\text{false negative @k})) \quad (7)$$

All experiments are conducted on a workstation with an Intel Core i9 processor with a speed of 3.6 GHz, 64 GB of memory, and a GPU (with 11 GB GDDR5X memory).

3.3. Results

3.3.1. Results on TextRS and Merced Datasets

Table 2 shows the results obtained with our model using m-R50x1 as a pre-trained CNN for the image branch. In the case of text-to-image retrieval, the scores (R@1, R@5, and R@10) are 19.02%, 55.25%, and 71.72%, respectively, on the TextRS dataset and they are 21.86%, 60.00%, and 75.58%, respectively, for the Merced dataset. On the other hand, the image-to-text matching scores are 22.95%, 59.52%, and 77.23%, respectively, on the TextRS dataset and 25.47%, 59.76%, and 72.61%, respectively, on the Merced dataset. We observe that retrieval accuracy increases significantly when going from R@1 to R@10, which indicates the difficulty of getting the exact match for the first retrieved query.

Table 2. Bidirectional retrieval results on the TextRS and Merced datasets.

Pretrained	Architecture	TextRS			Merced		
		R@1	R@5	R@10	R@1	R@5	R@10
Text-to-image	m-R50x1	19.02	55.25	71.72	21.86	60.00	75.58
Image-to-text	m-R50x1	22.95	59.52	77.23	25.47	59.76	72.61

In Figure 6, we show two successful scenarios for query sentences with their corresponding ground-truth images. The retrieval results show that the output images almost have the same objects. On the other hand, Figure 7 shows two unsuccessful scenarios, where the true matched image was not retrieved correctly. It is worth recalling that during training, we learn an embedding loss that aims to learn close representations of images with the same descriptions. However, this task is very challenging as the dataset is not very large (each image is associated with a textual description).

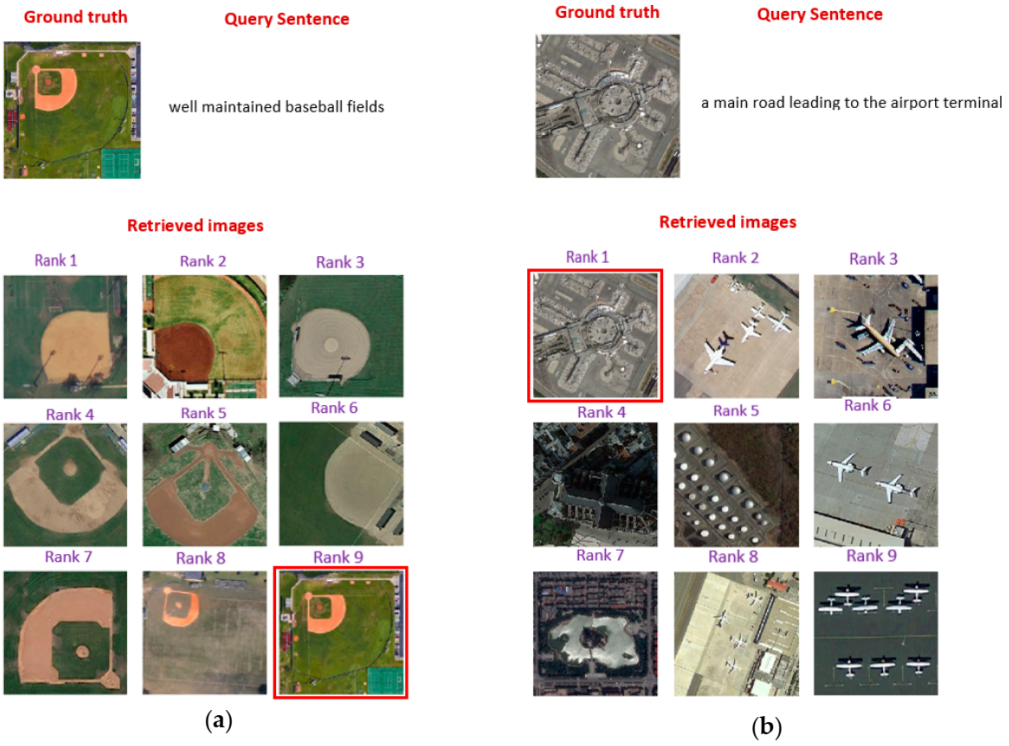

Figure 6. Successful scenarios of text-to-image retrieval.

Figure 7. Failed scenarios of text-to-image retrieval.

3.3.2. Sensitivity Analysis

We next investigated the results using different models pre-trained on two types of ImageNet datasets (i.e., ImageNet-21k and ImageNet-1k). Table 3 shows the results on the TextRS dataset using three different BiT architectures (m-R50x1, m-R50x3, and m-R101x1). As can be seen the models pre-trained on ImageNet-21k yields better results compared to the model pre-trained on ImageNet-1k. In particular, the obtained R@1, R@5, and R@10 values for the text branch are 19.02%, 55.25%, and 71.72%, respectively. On the other hand, the obtained values for the image branch are 21.86%, 60.00%, and 75.58%, respectively. These results suggest that the modes pre-trained on ImageNet-21 are more suitable compared to the one pre-trained on ImageNet-1k. In terms of computation complexity and accuracy, the m-R50 x 1 seems a good a choice compared to the other models.

Table 3. Text-to-image and image-to-text retrieval results on the TextRS dataset.

Pre-Trained	Architecture	Text-to-Image			Image-to-Text		
		R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ImageNet-21k	m-R50x1	19.02	55.25	71.72	21.86	60.00	75.58
	m-R50x3	18.27	52.23	68.79	23.25	60.93	74.41
	m-R101x1	15.11	51.96	67.76	18.13	53.25	71.86
ImageNet-1k	s-R50x1	15.53	46.97	63.25	18.13	49.53	68.13

Regarding Merced dataset (Table 4), we observe that for the largest model m-R101x1 it yields 23.68%, 60.38%, and 78.52%, respectively for text-to-image retrieval. While for image-to-text retrieval, the scores are 21.86%, 60.00%, and 75.58%, respectively. In comparison, using ImageNet-1k, the obtained R@1, R@5, and R@10 values are 19.71%, 56.04%, and 74.76%, respectively. Here again, the models pre-trained on ImageNet-2k exhibits a better behavior with m-R50x1 as a good compromise between accuracy and computation complexity.

Table 4. Text-to-image and image-to-text retrieval results on the Merced dataset.

Pre-Trained	Architecture	Text-to-Image			Image-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10
ImageNet-21k	m-R50x1	22.95	59.52	77.23	25.47	59.76	72.61
	m-R50x3	23.23	58.95	77.00	25.47	59.76	72.61
	m-R101x1	23.68	60.38	78.52	21.19	56.90	75.23
ImageNet-1k	s-R50x1	19.71	56.04	74.76	21.66	53.57	72.82

In order to assess further the proposed model, we propose to carry additional experiments by changing the size of the fully connected layer from 256 to 512. The results obtained in Tables 5 and 6 suggest that setting the number of neurons 256 yields in general better results.

Table 5. Retrieval results on the TextRS dataset using two different configurations for the hidden layer.

Pre-Trained	Hidden Layer Size	Text-to-Image			Image-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10
ImageNet-21k m-R50x1	256	19.02	55.25	71.72	21.86	60.00	75.58
	512	18.18	52.10	67.53	21.86	54.88	73.25

Table 6. Retrieval results on the Merced dataset using two different configurations for the hidden layer.

Pre-Trained	Hidden Layer Size	Text-to-Image			Image-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10
ImageNet-21k m-R50x1	256	22.95	59.52	77.23	25.47	59.76	72.61
	512	22.76	58.47	78.38	25.00	56.66	70.47

Finally, we compare our results to the method based on triplet networks proposed recently in [31] (Table 7). As can be seen the proposed method yields better retrieval results in terms of R@1, R@5, and R@10 scores. For instance, our method yields on TextRS dataset scores of 19.02%, 55.25%, and 71.72% versus 14.18%, 44.18%, and 62.55% for the method based on triplet networks. Similarly, our method obtains on Merced dataset scores of 22.76%, 58.47%, and 78.38%, while the one based on triplet networks gives 18.52%, 50.12%, and 69.20%.

Table 7. Comparison of the proposed method with other methods.

-	-	R@1	R@5	R@10
TextRS	Triplet [31]	14.18	44.18	62.55
	Proposed	19.02	55.25	71.72
Merced	Triplet [31]	18.52	50.12	69.20
	Proposed	22.76	58.47	78.38

4. Conclusions

In this work, we have proposed an unsupervised learning method for image retrieval in remote sensing imagery. Unlike traditional remote sensing image-to-image retrieval, this approach addresses the problem of text-to-image retrieval. The network consists of two asymmetric branches for image and sentence encoding, respectively. The experiments provided on two different benchmark datasets show interesting results compared to a recent method based on triplet networks. For future developments, we propose to investigate other embedding models in addition to more robust losses to increase the retrieval accuracy.

Author Contributions: Y.B. and M.M.A.R. designed and implemented the method, and wrote the paper. M.L.M., M.Z. and T.A. 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 work was supported by the Deanship of Scientific Research at King Saud University through the Local Research Group Program under Project 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 conflict of interest.

References

1. Schowengerd, R.A. *Remote Sensing: Models and Methods for Image Processing*; Elsevier: Amsterdam, The Netherlands, 2007; ISBN 978-0-12-369407-2.
2. Al-Doski, J.; Mansor, S.B.; Shafri, H.Z.M. Change detection process and techniques. *Civ. Environ. Res.* **2013**, *3*, 10.
3. Al Rahhal, M.M.; Bazi, Y.; Abdullah, T.; Mekhalfi, M.L.; Al Hichri, 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]
4. Aptoula, E. Remote sensing image retrieval with global morphological texture descriptors. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3023–3034. [CrossRef]
5. 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]
6. Kampffmeyer, M.; Salberg, A.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688.
7. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [CrossRef]
8. Tao, C.; Pan, H.; Li, Y.; Zou, Z. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2438–2442. [CrossRef]
9. Li, W.; Fu, H.; Yu, L.; Gong, P.; Feng, D.; Li, C.; Clinton, N. Stacked autoencoder-based deep learning for remote-sensing image classification: A case study of African land-cover mapping. *Int. J. Remote Sens.* **2016**, *37*, 5632–5646. [CrossRef]
10. He, Z.; Liu, H.; Wang, Y.; Hu, J. Generative adversarial networks-based semi-supervised learning for hyperspectral image classification. *Remote Sens.* **2017**, *9*, 1042. [CrossRef]
11. Atkinson, P.M.; Tatnall, A.R.L. Introduction neural networks in remote sensing. *Int. J. Remote Sens.* **1997**, *18*, 699–709. [CrossRef]
12. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
13. Mekhalfi, M.L.; Melgani, F.; Bazi, Y.; Alajlan, N. Fast indoor scene description for blind people with multiresolution random projections. *J. Vis. Commun. Image Represent.* **2017**, *44*, 95–105. [CrossRef]
14. Moranduzzo, T.; Mekhalfi, M.L.; Melgani, F. LBP-based multiclass classification method for UAV imagery. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milano, Italy, 26–31 July 2015; pp. 2362–2365.
15. Moranduzzo, T.; Melgani, F.; Mekhalfi, M.L.; Bazi, Y.; Alajlan, N. Multiclass coarse analysis for UAV imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6394–6406. [CrossRef]
16. 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]
17. 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 Vegas, NV, USA, 27–30 June 2016.
18. 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]
19. 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]
20. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016.

21. Pan, J.Y.; Yang, H.J.; Faloutsos, C.; Duygulu, P. GCap: Graph-based automatic image captioning. In Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004; p. 146.
22. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.A.; Mikolov, T. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2013; pp. 2121–2129.
23. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
24. Wang, L.; Li, Y.; Lazebnik, S. Learning deep structure-preserving image-text embeddings. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5005–5013.
25. 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]
26. Yao, T.; Zhang, Z.; Yan, L.; Yue, J.; Tian, Q. Discrete Robust supervised hashing for cross-modal retrieval. *IEEE Access* **2019**, *7*, 39806–39814. [CrossRef]
27. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:14112539.
28. Huang, Y.; Wu, Q.; Song, C.; Wang, L. Learning semantic concepts and order for image and sentence matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 6163–6171.
29. Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G. Hierarchical multimodal LSTM for dense visual-semantic embedding. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1881–1889.
30. Zhang, Y.; Lu, H. *Deep Cross-Modal Projection Learning for Image-Text Matching*; Springer: Cham, Switzerland, 2018; pp. 686–701.
31. Abdullah, T.; Bazi, Y.; Al Rahhal, M.M.; Mekhalfi, M.L.; Rangarajan, L.; Zuair, M. TextRS: Deep bidirectional triplet network for matching text to remote sensing images. *Remote Sens.* **2020**, *12*, 405. [CrossRef]
32. Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; Houlsby, N. Big transfer (BiT): General visual representation learning. *arXiv* **2020**, arXiv:191211370.
33. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
34. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 843–852.
35. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
36. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:150302531.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. *Identity Mappings in Deep Residual Networks*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 630–645.
38. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P.; et al. Microsoft COCO: Common objects in context. *arXiv* **2015**, arXiv:14050312.
39. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:170506950.
40. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
41. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
43. Weinberger, K.Q.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **2009**, *10*, 207–244.

44. Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; Wu, Y. 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.
45. Hoffer, E.; Ailon, N. *Deep Metric Learning Using Triplet Network*; Feragen, A., Pelillo, M., Loog, M., Eds.; Springer: Cham, Switzerland, 2015; pp. 84–92.
46. Law, M.T.; Thome, N.; Cord, M. Quadruplet-wise image similarity learning. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 249–256.
47. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 4004–4012.
48. Sohn, K. Improved deep metric learning with multi-class N-pair loss objective. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2016; pp. 1857–1865.
49. 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.
50. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
51. Yang, Y.; Newsam, S. 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, San Jose, CA, USA, 2–5 November 2010; Association for Computing Machinery: San Jose, CA, USA, 2010; pp. 270–279.
52. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
53. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
54. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016; pp. 1–5.