11

Deep Learning for Image Search and Retrieval in Large Remote Sensing Archives

Gencer Sumbul, Jian Kang, and Begüm Demir

11.1 Introduction

With the unprecedented advances in the satellite technology, recent years have witnessed a significant increase in the volume of remote sensing (RS) image archives. Thus, the development of efficient and accurate content-based image retrieval (CBIR) systems in massive archives of RS images is a growing research interest in RS. CBIR aims to search for RS images of the similar information content within a large archive with respect to a query image. To this end, CBIR systems are defined based on two main steps: (i) image description step (which characterizes the spatial and spectral information content of RS images); and (ii) image retrieval step (which evaluates the similarity among the considered descriptors and then retrieve images similar to a query image in the order of similarity). A general block scheme of a CBIR system is shown in Figure 11.1.

Traditional CBIR systems extract and exploit hand-crafted features to describe the content of RS images. As an example, bag-of-visual-words representations of the local invariant features extracted by the scale invariant feature transform (SIFT) are introduced in Yang and Newsam (2013). In Aptoula (2014), a bag-of-morphological-words representation of the local morphological texture features (descriptors) is proposed in the context of CBIR. Local Binary Patterns (LBPs), which represent the relationship of each pattern (i.e., pixel) in a given image with its neighbors located on a circle around that pixel, are found very efficient in RS. In Tekeste and Demir (2018), a comparative study that analyzes and compares different LBPs in RS CBIR problems is presented. To define the spectral information content of high-dimensional RS images the bag-of-spectral-values descriptors are presented in Dai et al. (2018). Graph-based image representations, where the nodes describe the image region properties and the edges represent the spatial relationships among the regions, are presented in Li and Bretschneider (2007); Chaudhuri et al. (2016, 2018). Hashing methods that embed high-dimensional image features into a low-dimensional Hamming (binary) space by a set of hash functions are found very effective in RS (Demir and Bruzzone 2016; Li and Ren 2017; Reato et al. 2019). By this method, the images are represented by binary hash codes that can significantly reduce the amount of memory required for storing the RS images with respect to the other descriptors. Hashing methods differ from each other on how the hash functions are generated. As an example, in Demir and Bruzzone (2016);

Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences,
First Edition. Edited by Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein.

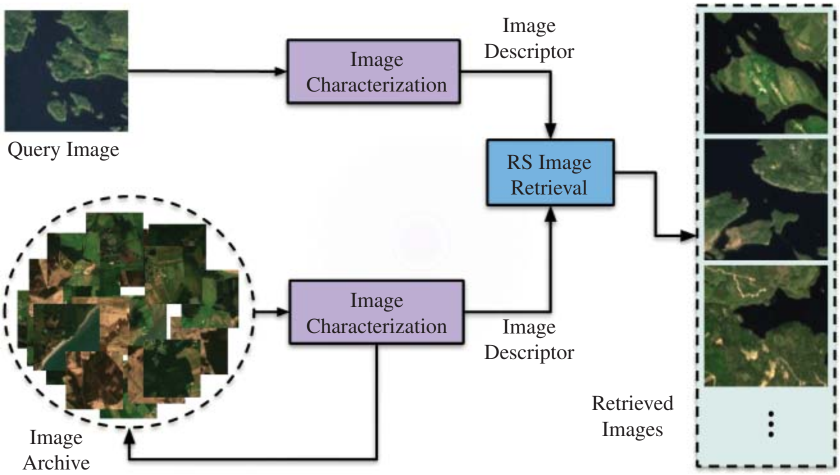

Figure 11.1 General block scheme of a RS CBIR system.

Reato et al. (2019) kernel-based hashing methods that define hash functions in the kernel space are presented, whereas a partial randomness hashing method that defines the hash functions based on a weight matrix defined using labeled images is introduced in Li and Ren (2017). More details on hashing for RS CBIR problems are given in section 11.3.

Once image descriptors are obtained, one can use the k-nearest neighbor (*k*-NN) algorithm, which computes the similarity between the query image and all archive images to find the k most similar images to the query. If the images are represented by graphs, graph matching techniques can be used. As an example, in Chaudhuri et al. (2016) an inexact graph matching approach, which is based on the sub-graph isomorphism and spectral embedding algorithms, is presented. If the images are represented by binary hash codes, image retrieval can be achieved by calculating the Hamming distances with simple bit-wise XOR operations that allow time-efficient search capability (Demir and Bruzzone 2016). However, these unsupervised systems do not always result in satisfactory query responses due to the semantic gap, which is occurred among the low-level features and the high-level semantic content of RS images (Demir and Bruzzone 2015). To overcome this problem and improve the performance of CBIR systems, semi-supervised and fully supervised systems, which require user feedback in terms of RS image annotations, are introduced (Demir and Bruzzone 2015). Most of these systems depend on the availability of training images, each of which is annotated with a single broad category label that is associated to the most significant content of the image. However, RS images typically contain multiple classes and thus can simultaneously be associated with different class labels. Thus, CBIR methods that properly exploit training images annotated by multi-labels are recently found very promising in RS. As an example, in Dai et al. (2018) a CBIR system that exploits a measure of label likelihood based on a sparse reconstruction-based classifier is presented in the framework of multi-label RS CBIR problems. Semi-supervised CBIR systems based on graph matching algorithms are proposed in Wang et al. (2016); Chaudhuri et al. (2018). In detail, in Wang et al. (2016) a three-layer framework in the context graph-based learning is proposed for query expansion and fusion of global and local features by using the label information of query images. In Chaudhuri et al. (2018) a correlated label propagation algorithm, which operates on a neighborhood graph for automatic labeling of images by using a small number of training images, is proposed.

The above-mentioned CBIR systems rely on shallow learning architectures and hand-crafted features. Thus, they can not simultaneously optimize feature learning and image retrieval, resulting in limited capability to represent the high-level semantic content of RS images. This issue leads to inaccurate search and retrieval performance in practice. Recent advances in deep neural networks (DNNs) have triggered substantial performance gain for image retrieval due to their high capability to encode higher-level semantics present in RS images. Differently from conventional CBIR systems, deep learning (DL)-based CBIR systems learn image descriptors in such a way that feature representations are optimized during the image retrieval process. In order words, DNNs eliminate the need for human effort to design discriminative and descriptive image descriptors for the retrieval problems. Most of the existing RS CBIR systems based on DNNs attempt to improve image retrieval performance by: (i) learning discriminative image descriptors; and (ii) achieving scalable image search and retrieval. The aim of this chapter is to present different DNNs proposed in the literature for the retrieval of RS images. The rest of this chapter is organized as follows. Section 11.2 reviews the DNNs proposed in the literature for the description of the complex information content of RS images in the framework of CBIR. Section 11.3 presents the recent progress on the scalable CBIR systems defined based on DNNs in RS. Finally, section 11.4 draws the conclusion of this chapter.

11.2 Deep Learning for RS CBIR

The DL-based CBIR systems in RS differ from each other in terms of: (i) the strategies considered for the mini-batch sampling; (ii) the approaches used for the initialization of the parameters of the considered DNN model; (iii) the type of the considered DNN; and (iv) the strategies used for image representation learning. Figure 11.2 illustrates the main approaches utilized in DL-based CBIR systems in RS. In detail, a set of training images is initially selected from the considered archive to train a DNN. Then, the selected training images are divided into mini-batches and fed into the considered DNN. After initializing the model parameters of the network, the training phase is conducted with an iterative estimation of the model parameters based on a loss function. The loss function is selected on the basis of the characteristics of the considered learning strategy.

During the last years, several DL-based CBIR systems that consider different strategies for the above-mentioned factors are presented. As an example, in Zhou et al. (2015) an unsupervised feature learning framework that learns image descriptors from a set of unlabeled RS images based on an autoencoder (AE) is introduced. After random selection of mini-batches and initialization of the model parameters, SIFT-based image descriptors are encoded into sparse descriptors by learning the reconstruction of the descriptors. The learning strategy relies on minimization of a reconstruction loss function between the SIFT descriptors and the reconstructed image descriptors in the framework of the AE. A CBIR system that applies a multiple feature representation learning and a collaborative affinity metric fusion is presented in Li et al. (2016b). This system randomly selects RS images

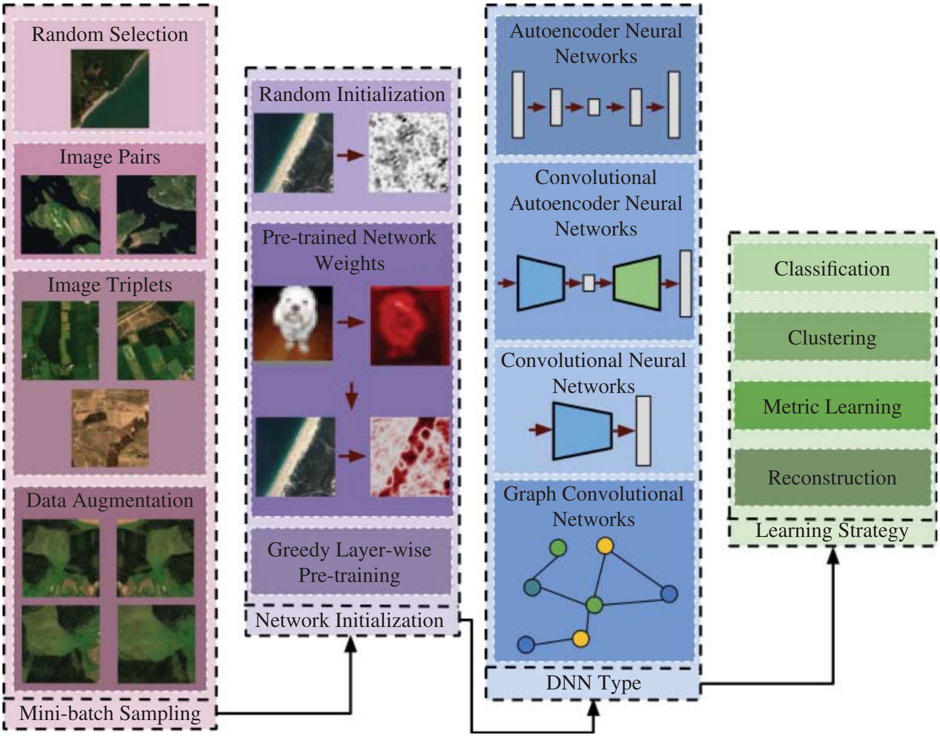

Figure 11.2 Different strategies considered within the DL-based RS CBIR systems.

for mini-batches and initializes the model parameters of a Convolutional Neural Network (CNN). Then, it employs the CNN for k-means clustering (instead of classification). To this end, a reconstruction loss function is utilized to minimize the error induced between the CNN results and the cluster assignments. Collaborative affinity metric fusion is employed to incorporate the traditional image descriptors (e.g., SIFT, LBP) with those extracted from different layers of the CNN. A CBIR system with deep bag-of-words is proposed in Tang et al. (2018b). This system employs a convolutional autoencoder (CAE) for extracting image descriptors in an unsupervised manner. The method first encodes the local areas of randomly selected RS images into a descriptor space and then decodes from descriptors to image space. Since encoding and decoding steps are based on convolutional layers, a reconstruction loss function is directly applied to reduce the error between the input and constructed local areas for the unsupervised reconstruction based learning. Since this system operates on local areas of the images, bag-of-words approach with k-means clustering is applied to define the global image descriptor from local areas. Although this system has the same learning strategy as Zhou et al. (2015), its advantages are two-fold compared to Zhou et al. (2015). First, model parameters are initialized with greedy layer-wise pre-training that allows more effective learning procedure with respect to the random initialization approach. Second, the CAE model has better capability to characterize the semantic content of images since it considers the neighborhood relationship through the convolution operations. The reader is referred to Chapter 2 for the detailed discussion on unsupervised feature learning in RS.

Reconstruction based unsupervised learning of RS image descriptors is found effective particularly when annotated training images are not existing. However, minimizing a reconstruction loss function on a small amount of unannotated images with a shallow neural network limits the accurate description of the high-level information content of RS images. This problem can be addressed by supervised DL-based CBIR systems that require a training set that consists of a high number of annotated images to learn effective models with several different parameters. The amount and the quality of the training images determine the success of the supervised DL models. However, annotating RS images at large scale is time-consuming, complex, and costly in operational applications. To overcome this problem, a common approach is to exploit DL models with proven architectures (such as ResNet or VGG), which are pre-trained on publicly available general purpose computer vision (CV) datasets (e.g., ImageNet). The existing models are then fine-tuned on a small set of annotated RS images to calibrate the final layers (this is known as transfer learning). As an example, in Hu et al. (2016) model parameters of a CNN are initialized with the parameters of a CNN model that is pre-trained on ImageNet. In this work, both initial training and fine-tuning are applied in the framework of the classification problems. To this end, the cross-entropy loss function is utilized to reduce the class prediction errors. Image descriptors learned with the cross-entropy loss function encode the class discrimination instead of similarities among images. Thus, it can limit the performance of CBIR systems. A data augmentation technique for mini-batch sampling is utilized in Hu et al. (2016) to improve the effectiveness of the image descriptors. To this end, different scales of RS images in a mini-batch are fed into the CNN. Then, the obtained descriptors are aggregated by using different pooling strategies to characterize the final image descriptors. A low-dimensional convolutional neural network (LDCNN) proposed in Zhou et al. (2017a) also utilizes parameters of a pre-trained DL model to initialize the network parameters. However, it randomly selects RS images for the definition of mini-batches and adopts a classification based learning strategy with the cross-entropy loss function. This system combines convolutional layers with cross channel parametric pooling and global average pooling to characterize low-dimensional descriptors. Since fully connected (FC) layers are replaced with pooling layers, LDCNN significantly decreases the total number of model parameters required to be estimated during the training phase. This leads to significantly reduced computational complexity and also reduced risk of over-fitting (which can occur in the case of training CNNs with a small amount of training images). A CBIR system based on a CNN with weighted distance is introduced in Ye et al. (2018). Similar to Hu et al. (2016) and Zhou et al. (2017a), this system also applies fine-tuning on a state-of-the-art CNN model pre-trained on ImageNet. In addition, it enhances the conventional distance metrics used for image retrieval by weighting the distance between a query image and the archive images based on their class probabilities obtained by a CNN. An enhanced interactive RS CBIR system, which extracts the preliminary RS image descriptors based on the LDCNN by utilizing the same mini-batch sampling, network initialization and learning strategy (based on the cross-entropy loss function), is introduced in Boualleg and Farah (2018). Labeled training images are utilized to obtain the preliminary image descriptors. Then, a relevance feedback scheme is applied to further improve the effectiveness of the image descriptors by considering the user feedbacks on the automatically retrieved images. The use of aggregated deep local features for RS image retrieval is proposed in Imbriaco et al. (2019). To this end, the VLAD representation of local convolutional descriptors from multiplicative and additive attention mechanisms are considered to characterize the descriptors of the most relevant regions of the RS images. This is achieved based on three steps. In the first step, similar to Ye et al. (2018) and Zhou et al. (2017a), the system operates on randomly selected RS images and applies fine-tuning to a state-of-the art CNN model while relying on a classification based learning strategy with the cross-entropy loss function. In the second step, additive and multiplicative attention mechanisms are integrated into the convolutional layers of the CNN and thus are retrained to learn their parameters. Then, local descriptors are characterized based on the attention scores of the resized RS images at different scales (which is achieved based on data augmentation). In the last step, the system transforms VLAD representations with Memory Vector (MV) construction (which produces the expanded query descriptor) to make the CBIR system sensitive to the selected query images. In this system, the query expansion strategy is applied after obtaining all the local descriptors. This query-sensitive CBIR approach further improves the discrimination capability of image descriptors, since it adapts the overall learning procedure of DNNs based on the selected queries. Thus, it has a huge potential for RS CBIR problems.

Most of the above-mentioned DL-based supervised CBIR systems learn an image feature space directly optimized for a classification task by considering entropy-based loss functions. Thus, the image descriptors are designed to discriminate the pre-defined classes by taking into account the class based similarities rather than the image based similarities during the training stage of the DL models. The absence of positive and negative images with respect to the selected query image during the training phase can lead to a poor CBIR performance. To overcome this limitation, metric learning is recently introduced in RS to take into account image similarities within DNNs. Accordingly, a Siamese graph convolutional network is introduced in Chaudhuri et al. (2019) to model the weighted region adjacency graph (RAG) based image descriptors by a metric learning strategy. To this end, mini-batches are first constructed to include either similar or dissimilar RS images (Siamese pairs). If a pair of images belongs to the same class, they are assumed as similar images, and vice versa. Then, RAGs are fed into two graph convolutional networks with shared parameters to model image similarities with the contrastive loss function. Due to the considered metric learning strategy (which is guided by the contrastive loss function) the distance between the descriptors of similar images is decreased, while that between dissimilar images is increased. The contrastive loss function only considers the similarity estimated among image pairs, i.e., similarities among multiple images are not evaluated, which can limit the success of similarity learning for CBIR problems.

To address this limitation, a triplet deep metric learning network (TDMLN) is proposed in Cao et al. (2020). TDMLN employs three CNNs with shared model parameters for similarity learning through image triplets in the content of metric learning. Model parameters of the TDMLN are initialized with a state-of-the-art CNN model pre-trained on ImageNet. For the mini-batch sampling, TDMLN considers an anchor image together with a similar (i.e., positive) image and a dissimilar (i.e., negative) image to the anchor image at a time. Image triplets are constructed based on the annotated training images (Chaudhuri et al. 2019). While anchor and positive images belong to the same class, the negative image is associated to a different class. Then, similarity learning of the triplets is achieved based on the triplet loss function. By the use of triplet loss function, the distance estimated between

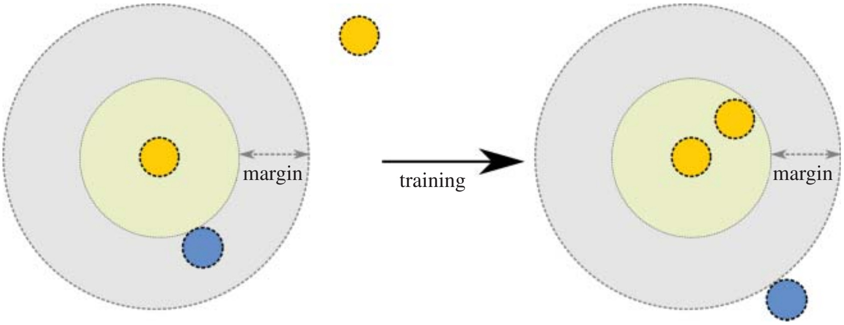

Figure 11.3 The intuition behind the triplet loss function: after training, a positive sample is moved closer to the anchor sample than the negative samples of the other classes.

the anchor and positive images in the descriptor (i.e., feature) space is minimized, whereas that computed between the anchor and negative images is separated by a certain margin. Figure 11.3 illustrates intuition behind the triplet loss function. Metric learning guided by the triplet loss function learns similarity based on the image triplets and thus provides highly discriminative image descriptors in the framework of CBIR. However, how to define and select image triplets is still an open question. Current methods rely on the image-level annotations based on the land-cover land-use class labels, which do not directly represent the similarity of RS images. Thus, metric learning-based CBIR systems need further improvements to characterize retrieval specific image descriptors. One possible way to overcome this limitation can be an identification of image triplets through visual interpretation instead of defining triplets based on the class labels. Tabular overview of the recent DL-based CBIR systems in RS is presented in Table 11.1.

11.3 Scalable RS CBIR Based on Deep Hashing

Due to the significant growth of RS image archives, an image search and retrieval through linear scan (which exhaustively compares the query image with each image in the archive) is computationally expensive and thus impractical. This problem is also known as large-scale CBIR problem. In large-scale CBIR, the storage of the data is also challenging as RS image contents are often represented in high-dimensional features. Accordingly, in addition to the scalability problem, the storage of the image features (descriptors) also becomes a critical bottleneck. To address these problems, approximate nearest neighbor (ANN) search has attracted extensive research attention in RS. In particular, hashing-based ANN search schemes have become a cutting-edge research topic for large-scale RS image retrieval due to their high efficiency in both storage cost and search/retrieval speed. Hashing methods encode high-dimensional image descriptors into a low-dimensional Hamming space where the image descriptors are represented by binary hash codes. By this way, the (approximate) nearest neighbors among the images can be efficiently identified based on the Hamming distance with simple bit-wise operations. In addition, the binary codes can significantly reduce the amount of memory required for storing the content of images. Traditional hashing-based RS CBIR systems initially extract hand-crafted image descriptors and then utilize hash functions that map the original high-dimensional

Table 11.1 Main characteristics of the DL-based CBIR systems in RS.

Reference	Mini-batch Sampling	Network Initialization	DNN Type	Learning Strategy	Loss Function
Zhou et al. (2015)	Random selection	Random initialization	Auto-encoder	Reconstruction (unsupervised)	Reconstruction
Hu et al. (2016)	Data augmentation	Pre-trained network weights	Convolutional neural network	Classification (supervised)	Cross-entropy
Li et al. (2016b)	Random selection	Random initialization	Convolutional neural network	Clustering (unsupervised)	Reconstruction
Zhou et al. (2017a)	Random selection	Pre-trained network weights	Convolutional neural network	Classification (supervised)	Cross-entropy
Ye et al. (2018)	Random selection	Pre-trained network weights	Convolutional neural network	Classification (supervised)	Cross-entropy
Tang et al. (2018b)	Random selection	Greedy layer-wise pre-training	Convolutional auto-encoder	Reconstruction (unsupervised)	Reconstruction
Boualleg and Farah (2018)	Random Selection	Pre-trained network weights	Convolutional neural network	Classification (supervised)	Cross-entropy
Imbriaco et al. (2019)	Random selection Data augmentation	Pre-trained network weights	Convolutional neural network	Classification (supervised)	Cross-entropy
Chaudhuri et al. (2019)	Image pairs	Random initialization	Graph convolutional network	Metric learning (supervised)	Contrastive
Cao et al. (2020)	Image triplets	Pre-trained network weights	Convolutional neural network	Metric learning (supervised)	Triplet

representations into low-dimensional binary codes, such that the similarity to the original space can be well preserved (Demir and Bruzzone 2016; Li and Ren 2017; Reato et al. 2019; Fernandez-Beltran et al. 2020). Thus, descriptor extraction and binary code generation are applied independently from each other, resulting in sub-optimal hash codes. Success of DNNs in image feature learning has inspired research on developing DL-based hashing methods (i.e., deep hashing methods).

Recently, several deep hashing-based CBIR systems that simultaneously learn image representations and hash functions based on the suitable loss functions are introduced in RS (see Table 11.2). As an example, in Li et al. (2018b) a supervised deep hashing neural network (DHNN) that learns deep features and binary hash codes by using the contrastive and quantization loss functions in an end-to-end manner is introduced. The contrastive loss function can also be considered as the binary cross-entropy loss function, which is optimized to classify whether an input image pair is similar or not. One advantage of the contrastive loss function is its capability of similarity learning, where similar images can be grouped together, while moving away dissimilar images from each other in the feature space. Due to the ill-posed gradient problem, the standard back-propagation of DL models to directly optimize hash codes is not feasible. The use of the quantization loss mitigates the performance degradation of the generated hash codes through the binarization

Table 11.2 Main characteristics of the state-of-the-art deep hashing-based CBIR systems in RS.

Reference	Loss Functions	Learning Type	Hash Layer
Li et al. (2018b)	Contrastive, Quantization	supervised	linear
Li et al. (2018a)	Contrastive, Quantization	supervised	linear
Roy et al. (2020)	Triplet, Bit balance, Quantization	supervised	sigmoid
Song et al. (2019)	Contrastive, Quantization, Cross-entropy	supervised	linear
Tang et al. (2019)	Cross-entropy, Contrastive, Reconstruction, Quantization, Bit balance	semi-supervised	linear
Liu et al. (2019)	Adversarial, Quantization, Contrastive, Cross-entropy	supervised	sigmoid

on the CNN outputs. In Li et al. (2018a) the quantization and contrastive loss functions are combined in the framework of the source-invariant deep hashing CNNs for learning a cross-modality hashing system. Without introducing a margin threshold between the similar and dissimilar images, a limited image retrieval performance can be achieved based on the contrastive loss function. To address this issue, a metric-learning based supervised deep hashing network (MiLaN) is recently introduced in Roy et al. (2020). MiLaN is trained by using three different loss functions: (i) the triplet loss function for learning a metric space (where semantically similar images are close to each other and dissimilar images are separated); (ii) the bit balance loss function (which aims at forcing the hash codes to have a balanced number of binary values); and (iii) the quantization loss function. The bit balance loss function makes each bit of hash codes to have a 50% chance of being activated, and different bits to be independent from each other. As noted in Roy et al. (2020), the learned hash codes based on the considered loss functions can efficiently characterize the complex semantics in RS images. A supervised deep hashing CNN (DHCNN) is proposed in Song et al. (2019) in order to retrieve the semantically similar images in an end-to-end manner. In detail, DHCNN utilizes the joint loss function composed of: (i) the contrastive loss function; (ii) the cross-entropy loss function (which aims at increasing the class discrimination capability of hash codes); and (iii) the quantization loss. In order to predict the classes based on the hash codes, a FC layer is connected to the hash layer in DHCNN. As mentioned above, one disadvantage of the cross-entropy loss function is its deficiency to define a metric space, where similar images are clustered together. To address this issue, the contrastive loss function is jointly optimized with the cross-entropy loss function in DHCNN. A semi-supervised deep hashing method based on the adversarial autoencoder network (SSHAAE) is proposed in Tang et al. (2019) for RS CBIR problems. In order to generate the discriminative and similarity preserved hash codes with low quantization errors, SSHAAE exploits the joint loss function composed of: (i) the cross-entropy loss function; (ii) a reconstruction loss function; (iii) the contrastive loss function; (iv) the bit balance loss function; and (v) the quantization loss function. By minimizing the reconstruction loss function, the label vectors and hash codes can be obtained as the latent outputs of the AEs. A supervised deep hashing method based on a generative adversarial network (GAN) is proposed in Liu et al. (2019). For the generator of the GAN, this method introduces a joint loss function that

Table 11.3 Comparison of the DL loss functions considered within the deep hashing-based RS CBIR systems. Different marks are provided: "×" (no) or "✓" (yes).

Loss Function	Similarity Learning Capability	Mini-batch Sampling Requirement	Bit Balance Capability	Binarization Capability	Annotated Image Requirement
Contrastive	✓	Image pairs	×	×	✓
Triplet	✓	Image triplets	×	×	✓
Adversarial	×	Random selection	✓	×	×
Reconstruction	×	Random selection	×	×	×
Cross-entropy	×	Random selection	×	×	✓
Bit balance	×	Random selection	✓	×	×
Quantization	×	Random selection	×	✓	×

composed of: (i) the cross-entropy loss function; (ii) the contrastive loss function; and (iii) the quantization loss function. For the discriminator of the GAN, the sigmoid function is used for the classification of the generated hash codes as true codes. This allows the learned hash codes following the uniform binary distribution to be restricted. Thus, the bit balance capability of hash codes can be achieved. It is worth noting that the above-mentioned supervised deep hashing methods preserve the discrimination capability and the semantic similarity of the hash codes in the Hamming space by using annotated training images.

In Table 11.3, we analyze and compare all the above-mentioned loss functions based on their: (i) capability on similarity learning, (ii) requirement on the mini-batch sampling; (iii) capability of assessing the bit balance issues; (iv) capability of binarization of the image descriptors; and (v) requirement on the annotated images. For instance, the contrastive and triplet loss functions have the capabilities to learn the relationship among the images in the feature space, where the semantic similarity of hash codes can be preserved. Regarding to the requirement of mini-batch sampling, pairs of images should be sampled for the contrastive loss function, image triplets should be constructed for the triplet loss function. The bit balance and adversarial loss functions are exploited for learning the hash codes with the uniform binary distribution. It is worth noting that an adversarial loss function can be also exploited for other purposes, such as for image augmentation problems to avoid overfitting (Cao et al. 2018). The quantization loss function enforces the produced low-dimensional features by the DNN models to approximate the binary hash codes. With regard to the requirement on image annotations, the contrastive and triplet loss functions require the semantic labels to construct the relationships among the images.

11.4 Discussion and Conclusion

In this chapter, we presented a literature survey on the most recent CBIR systems for efficient and accurate search and retrieval of RS images from massive archives. We focused our attention on the DL-based CBIR systems in RS. We initially analyzed the recent DL-based CBIR systems based on: (i) the strategies considered for the mini-batch sampling; (ii) the approaches used for the initialization of the parameters of the considered DNN models; (iii) the type of the considered DNNs; and (iv) the strategies used for image representation learning. Then, the most recent methodological developments in RS related to scalable image search and retrieval were discussed. In particular, we reviewed the deep hashing-based CBIR systems and analyzed the loss functions considered within these systems based on their: (i) capability of similarity learning, (ii) requirement on the mini-batch sampling; (iii) capability of assessing the bit balance issues; (iv) capability of binarization; and (v) requirement on the annotated images. Analysis of the loss functions under these factors provides a guideline to select the most appropriate loss function for large-scale RS CBIR problems.

It is worth emphasizing that developing accurate and scalable CBIR systems is becoming more and more important due to the increased number of images in the RS data archives. In this context, the CBIR systems discussed in this chapter are very promising. Despite the promising developments discussed in this chapter (e.g., metric learning, local feature aggregation, and graph learning), it is still necessary to develop more advanced CBIR systems. For example, most of the systems are based on the direct use of the CNNs for the retrieval tasks, whereas the adapted CNNs are mainly designed for learning a classification problem and thus model the discrimination of pre-defined classes. Thus, the image descriptors obtained through these networks cannot learn an image feature space that is directly optimized for the retrieval problems. Siamese and triplet networks are defined in the context of metric learning in RS to address this problem. However, the image similarity information to train these networks is still provided based on the pre-defined classes, preventing to achieve retrieval specific image descriptors. Thus, CBIR systems that can efficiently learn image features optimized for retrieval problems are needed. Furthermore, the existing supervised DL-based CBIR systems require a balanced and complete training set with annotated image pairs or triplets, which is difficult to collect in RS. Learning an accurate CBIR model from imbalanced and incomplete training data is very crucial and thus there is a need for developing systems addressing this problem for operational CBIR applications. Furthermore, the availability of an increased number of multi-source RS images (multispectral, hyperspectral and SAR) associated to the same geographical area motivates the need for effective CBIR systems, which can extract and exploit multi-source image descriptors to achieve rich characterization of RS images (and thus to improve image retrieval performance). However, multi-source RS CBIR has not been explored yet (i.e., all the deep hashing-based CBIR systems are defined for images acquired by single sensors). Thus, it is necessary to study CBIR systems that can mitigate the aforementioned problems.

Acknowledgement

This work is supported by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764.