Received: July 21, 2020. Revised: August 11, 2020. 539

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Learning to Hash with Convolutional Network for Multi-label Remote Sensing Image Retrieval

Marwa Sayed Moustafa[1] **Sayed Ahmed**[1]* **Amal Ahmed Hamed**[1]

[1]*National Authority for Remote Sensing and Space Sciences, Egypt.*
* Corresponding author's Email: se.sayedahmed@gmail.com

Abstract: Recently, deep hashing dominated single label image retrieval approaches. However, the complex nature of remote sensing images, which likely contains multi-labels, hardly benefits from the above approaches. To overcome single-label image retrieval limitations in remote sensing domain, we address this problem by proposing a multi-label remote sensing image retrieval (MLRSIR-NET) framework. Specifically, the proposed MLRSIR-NET composed of two main sub-networks: multi-level feature extraction and deep hash. The multi-level feature extraction network predicts multi-level features to exploit different levels of Convolution Neural Network (CNN (characteristics. To suppress discriminative feature representation, the multi-level features are aggregated and feed to Convolutional Block Attention Module (CBAM) to amplify the representation of relevant multi-label features. CBAM is flexibly integrated into our network with end-to-end training. The hash network stacked two fully connected layers aimed to learn multiple hashing functions to encode the feature vector into a compact hash code. Finally, we conduct experiments on two benchmarks for multi-label images: Dense Labelling Remote Sensing Dataset (DLRSD) and Wuhan Dense Labeling Dataset (WHDLD) to systematically assess the performance. The results show that the proposed framework improved the accuracy in terms of Mean Average Precision (MAP) by a considerable margin of 85.4%, 87.2%, 90.8% and 92.9% for 12-bit, 24-bit, 32-bit and 48-bit code lengths respectively on DLRSD. For WHDLD, it can be noted that the proposed framework supers the DCH by 93.8%, 98.7%, 91.9%, and 94.6% on average respectively.

Keywords: Deep hashing, Remote sensing image retrieval, Multi-label, CBAM attention module.

1. Introduction

The modern, sophisticated earth observation satellite instruments capture almost a daily significant volume of heterogeneous remote sensing data which introduce a new challenge in the automatic fast retrieval of large-scale images databases. Recently, significant efforts had been introduced from remote sensing community to image retrieval topic due to its importance. The previous efforts mainly focused on developing a robust handcraft feature extraction to enhance the performance of the retrieval system [1]. Traditional Remote Sensing Image Retrieval (RSIR) method uses handcraft features to represent the content of the images [1-4]. These features could be categorized based on features type into global and local. The global features are extracted from the entire images while the local features are extracted from image patches. Remote Sensing (RS) community investigates handcraft features for different applications such as change detection, image registration, urban planning, and image classification. Despite their robustness to occlusion, view angle, and light conditions, the burden of computation time which often involves ad-hoc or heuristic design decisions, making handcraft features extractor hardly optimal for retrieval task.

Content-Based Image Retrieval (CBIR) [2] can be used to search visually-similar images. The key to image retrieval is to design or learn representative descriptions especially for satellite images. Different natural image descriptions have been applied for satellite image retrieval [4]. These conventional descriptions include shape, color, texture and other features. Recently, CNNs achieved state-of-art in different fields, such as image classification, object

detection, and natural language processing. RS researchers adopted different CNNs to benefit from their efficient feature representation compared with handcraft learning- based features [5]. By leveraging the pre-trained convolutional architectures, the deeply learned features promote the retrieval accuracy dramatically. As a result, designing convolutional networks with hashing function provides an effective solution for the task of satellite image retrieval.

Deep dense features [3, 6, 7] of the satellite images are considered shortlisted from high-dimensional database. However, the memory cost and computational time of searching the deep dense features increase linearly with database size. It becomes very difficult to search these huge satellite image inventories in real or near real time. Recently, image hashing techniques have become more attractive due to its compact representation, which converts images into binary hash codes to save the computational time and storage cost. In particular, deep hashing methods leverage the merits of deep learning and image hashing and proven to be efficient for visual search whereas, these hashing methods were designed mainly for natural scene images, and for satellite images. Due to the substantial gap between natural scene image and satellite image, deep hashing models trained on natural scene images cannot be applied directly to the task of satellite image retrieval. Hence, it is necessary to explore a new deep hashing method for satellite images.

The aforementioned methods were able to achieve comparable performance for single label remote sensing datasets. However, Single label is hardly sufficient to address the complex nature of RS images which likely contains multiple classes. Thus, in the case of RSIR problem with such complex image categories, multi-label RSIR approaches are needed. Several attempts had been introduced to address the multilabel problem for image classification and retrieval [8]. Recently, RS community directed their efforts to propose multilabel approaches to tackle the complex nature of RS image. Ample of work was introduced to effectively develop multi-labelled classification and image retrieval methods that overcome the single-label RSIR methods limitation. In this paper, we consider the problem of multi-label image retrieval for large scale remote sensing images.

The main contributions of this work can summarized as follows:

We propose efficient multi-label image retrieval framework based on attention learning and deep hash for multi-label satellite image.

The proposed framework benefits from multi-level features, CBAM and deep hashing to empower relevant multi-label features representation due to their effectiveness.

The proposed framework was systematically assessed on two public multi-label satellite datasets. Ample experiments evaluate the effectiveness and efficiency of the proposed framework. The results demonstrated that the proposed framework surpass other retrieval methods.

The rest of this paper is organized as follows. Section 2 outlines the related work briefly. The proposed deep hashing method for satellite images is presented in Section 3. Experimental results for the proposed method evaluation are introduced in Section 4. Section 5 describes conclusions.

2. Related work

Image retrieval [9] explores and searches large-scale database to access precise information efficiently. This search can be done using metadata or content-based data. RSIR performance depends mainly on the effectiveness of the feature representations. Significant work had been undertaken to develop powerful feature representations over the past few decades. Feature representation is categorized into two main groups: handcrafted features or shallow and deep learning features. Recently, a combined approach was considered. Content based image retrieval method had mainly focused on three main issues: visual feature, similarity metric and relevance feedback. Several comprehensive reviews had been recently published [1-4, 8].

2.1 Satellite image retrieval

Various approaches have been presented for satellite images retrieval. Features are vitally important in searching satellite images; different approaches are devoted to explore the representative features. Several features adopted in image retrieval purposes, color features [10], shape features [11], texture features [12] and local invariant features [13]. However, the low-level features hardly represent the images accurately, high-level features such as Scale Invariant Feature Transformation method (SIFT) and Speeded-Up Robust Features (SURF) [14], Bag of Words (BOW), Histograms of Oriented Gradients (HOG) and Local Binary Pattern (LBP), Gray Level Coocurrence Matrix (GLCM), and Maximal Response 8 (MR8) [15] had been introduced to enhance the image representation.

Recently, attention models [16] had been extensively explored in natural language processing

[17], computer vision [18, 19] and others. The multi-label image retrieval problem aims to learn the discriminative features to distinguish different labels. Commonly, a satellite image is often annotated to several labels due to its complex nature and spatial resolution, which become critical cues for classification, localization, and retrieval problems. Residual attention [20] is integrated to train a stacked deep neural network to leverage the classification accuracy. Several attention-based image retrievals had been explored in different studies [18, 21-24, 36-39].

2.2 Deep hashing

Deep hashing approaches had attracted much attention in image retrieval field, as deep hashing is joined with the deep features. These deep models have achieved superior performance for natural scene images and outperformed the hand-crafted features. Two existing approaches for deep hashing [25]: supervised and unsupervised. In unsupervised deep hashing, stacked autoencoder and Boltzmann machine were adopted and minimize reconstruction error function is usually utilized to learn the parameters of the nonlinear projections. In supervised deep hashing, CNNs are adopted as backbones to extract the generic image features. Then hashing layer is appended to learn binary hash codes. Different loss function [18, 21-24, 36-39] had been explored including point-wise similarity, pair-wise similarity [37, 39] and triplet-wise supervision [38] to learn binary hash codes. In [26], different loss had been considered to train to generic feature. Different deep hashing neural networks (DHNNs) were introduced in [27] to overcome the limitation of the handcrafted features, different deep architectures were considered to extract deep features. Uni-source and cross-source satellite images inventory had been investigated and deep hashing convolutional neural networks named "Source-Invariant Deep Hashing Convolutional Neural Networks" (SIDHCNNs) [28], which can be optimized in an end-to-end manner using a series of well-designed optimization constraints. Instead of keeping fixed deep hashing function, online hashing method was introduced to learn the hashing functions with respect to the newly incoming RS images [27]. The introduced hash model is updated in a sequential mode. Supervised discrete hashing called "Fast Supervised Discrete Hashing" (FSDH) [29], and Partial Randomness Hashing (PRH) [30, 31] were introduced to cope with big data problems.

In [36], bayesian learning framework was introduced, which utilized the Cauchy distribution to reduce the inconsistency between features and binary codes. The approaches in [37, 39] utilized the pair-wise similarity to optimize the similarity of correlated hash codes.

The loss functions integrated with the above deep hashing methods have initially been designed for natural images. However, satellite images contain more than one class label with different object size, orientation, sensor noise, which results in a degraded and inconsistent hash codes learned via the above methods. In our work, we introduced a multi-label image retrieval framework based on attention learning to tackle satellite images complex nature. Compared with other works, the proposed framework learns compact and effective hash codes for multi-label satellite images.

3. Proposed method

This section formulates the multi-label image retrieval problem using deep hashing. In addition, a detailed description of the proposed framework is presented.

3.1 Problem definition

Let $\{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ be a set of multi-labelled training image samples, where $x \epsilon \mathbb{R}^w$, w is feature vector dimension, n is the number of training samples. Each image x_i is associated to its ground-truth label vector $y_i = [y_i^1, y_i^2, \dots . y_i^L]$. Any image x_i can be assigned to one or more label based on its content. Thus, the label vector $(y_i^k, k = 1, \dots, L)$ is set to $y_i^k = 1$ if the image contains the k^{th} label and $y_i^k = 0$ otherwise, where L referring to the total number of labels in the dataset.

Our goal is to map multiple hashing functions $H_i: \mathrm{X}_i\{0,1\}^{\kappa}$ to encode each input image X_i into a κ-bit binary code that preserving the semantic data structure inherent in both visual contents and its multi-label information. Each bit in the binary code is calculated as pairwise similarity $S = \{s_{ij}\}_{i=1}^{N}$. The similarity label is set $s_{ij} = 1$ if x_i and x_j have shared a semantic label, otherwise $s_{ij} = 0$.

3.2 Proposed image retrieval framework

We introduce a detail description of the proposed Multi-Label Remote Sensing Image Retrieval (MLRSIR-NET) framework. Fig. 1 illustrates the graphical representation for the components of the proposed framework. The proposed framework composed of two sub-networks: deep feature extraction network and hash network.

3.2.1. Deep feature extraction network

A traditional CNN consists of several convolutional blocks whereas each block contains several convolutional layers followed by pooling learning. The multi level feature extraction network is designed to facilities transfer learning from natural image domain to remote sensing domain, fine-tuned using specific samples. The output of multi-level layer of the feature extraction network is feed to an CBAM [32] to allow a more robust feature, and this is believed to render it capable of achieving the optimal features with regard to multi-label problem. Let us assume that each satellite image $x_{\mathrm{i}} \in R^{w}$ is associated with any of y_i^k vector contains L possible labels. y_i^k is set to 1 to indicate the image contains the k^{th} label and 0 otherwise. The proposed network aims to learn multiple non-linear embedding H_i: $\mathrm{X}_i\{0,1\}^{\kappa}$ that maps each x_{i} onto a compact feature space R^{κ} where $\kappa \ll w$. In this subspace, the Euclidean distance among groups of similar images should be small, and conversely the distance between dissimilar images should be large. The distance should be robust to different variability such as geometric distortions and noise. In addition, the proposed network is designed to depend only upon a learnable parameter vector θ. Consequently, the learned distance, $d_\theta(f_\theta(x_i), f_\theta(x_j))$, also depends only on θ.

In this context, we assumed that x_i and x_j are positive with respect to each other only when they share at least one label, i.e. when $y_i \cap y_j \neq \emptyset$; conversely, both images are interpreted as negative when the equality is not satisfied.

Recently, attention mechanism had become a popular concept and improved deep learning architectures. We added CBAM [32] to refine the spatial and channel-wise features to improve salient regions and extract more distinctive features. In order to capture more significant features and improve the discrimination of the multi-scale features, CBAM is utilized after aggregating of the multi level features. In CBAM, both channel and spatial attention mechanisms are integrated. The detailed architecture of CBAM is shown in Fig. 2. The input feature vector is multiplied sequentially by both the channel attention module and spatial attention module outputs. The channel attention module architecture consists of Max-Pooling and Avg-Pooling layers operate along the width and the height dimensions of feature maps in parallel fashion. Then, a multiple layer perceptron (MLP) is adopted to calculate the weights along the channel dimension. The channel attention is formulated as shown in Eq. (1).

$$Att_{ch}=f(H(Maxpool(x))+ H(Avgpool(x))) \quad (1)$$

where f and H refer to the sigmoid function, multilayer perceptron respectively.

The spatial attention module shares the same architecture as the channel attention module, but Max-Pooling and Avg-Pooling layers are adopted along channel dimension of the input features. The spatial attention module is formulated as shown in Eq. (2).

$$Att_{sp} = g(cov2d\big(Maxpool(Att_{ch}.x)\big) + g\big(Avgpool(Att_{ch}.x)\big) \quad (2)$$

where $g, cov2d$ refer to the sigmoid function, and convolutional with filter size 7×7.

The final output of CABM module is formulated as shown in Eq. (3).

$$CABM(x) = x.Att_{ch}.Att_{sp} \quad (3)$$

where Att_{ch} and Att_{sp} refer to sigmoid function, and output vector from the channel and the spatial modules respectively.

3.2.2. Deep hash network

The proposed deep hashing network composed of two fully-connect layers whereas the last layer contains κ neuron. Let x identify the feature vectors outputs from $CABM$ attached to the deep feature extraction network. The hash vector for vector x can be obtained as shown in Eq. (4) and Eq. (5).

$$h(x) = tan(W_2 . tan(W_1 x + bais)) \quad (4)$$

$$b = sign\big(h(x)\big) = \begin{cases} 1, h(x) > 0 \\ 0, h(x) \leq 0 \end{cases} \quad (5)$$

where W_1, W_1 are the computed weight for the two layers.

3.2.3. Loss function

In order to learn a compact hash code, we utilized pairwise loss function based on the following criterions: 1) Multi-label similarity preserving. The distance between the learned features for a pair of images should be very small to indicate the similarity between them; however, the distance should be large to indicate the dissimilarity between the two images. 2) Semantic hash coding. To generate a

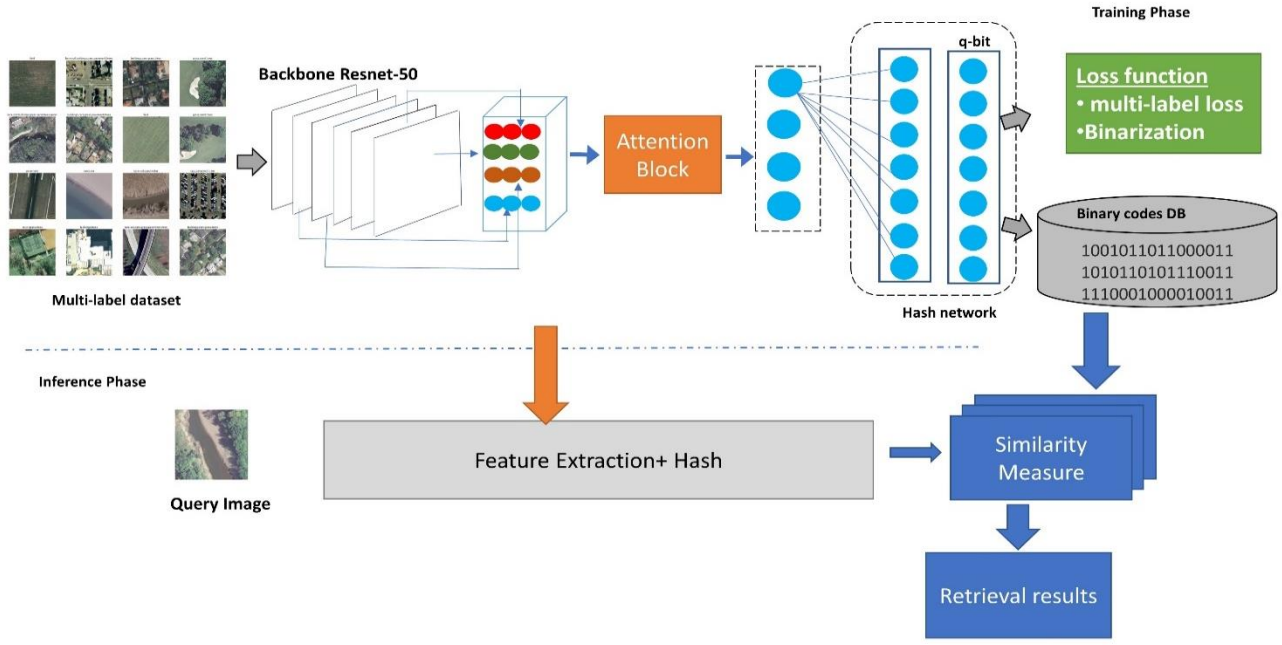

Figure. 1 Graphical representation for the proposed multi-label remote sensing image retrieval (MLRSIR-NET) framework

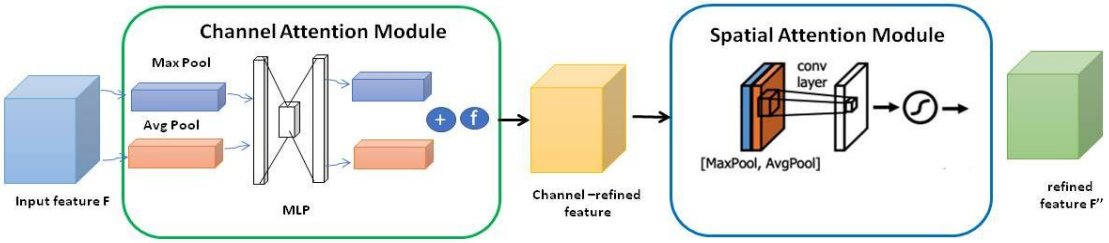

Figure. 2 Snapshot of CBAM components. The module has two consecutive modules channel and spatial

distinguishable hash code for each of the reweighted semantic vectors of the pair image. Accordingly, the overall loss of the proposed network is written as shown in Eq. (6).

$$\min_{\emptyset} l_{feat\_att\_s} + H_{quan} \tag{6}$$

where $l_{feat\_att\_s}, l_{quan}$ are the Multi-label similarity preserving for the two images.

In the retrieval procedure, the learned model if used directly to generate a hash code for the query image. Euclidean distance is calculated between the generated hash code for the query image and each code in the binary database to retrieve the most similar images and the retrieved matched with those in the dataset by computing the hamming distances between the query images.

4. Experiments

This section is organized as follows: Section 4.1 provides description of the dataset. Section 4.2 presents the experiments setup and protocol. Section 4.3 highlights the experimental results and findings.

4.1 Datasets

The experiments in this work were conducted on two challenging multi-label datasets in the remote sensing domain. DLRSD introduced by Chaudhuri et al. [33] as the first multi-label dataset based on UC Merced archive. Each image was manually annotated per pixel with the following 17 classes, i.e., airplane, bare soil, building, car, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tank, tree, and water. The dataset consists of total 2100 RGB images of size 256×256 and the spatial resolution is 0.3 m. the percentages of each class are shown in Fig. 3.

The second dataset is WHDLD which was introduced by Shao et al. [34] in 2020. The image was cropped to 256×256 for Wuhan urban area. Then, experts manually annotated each pixel to one of the following six classes: building, road, pavement, vegetation, bare soil, and water. The dataset consists

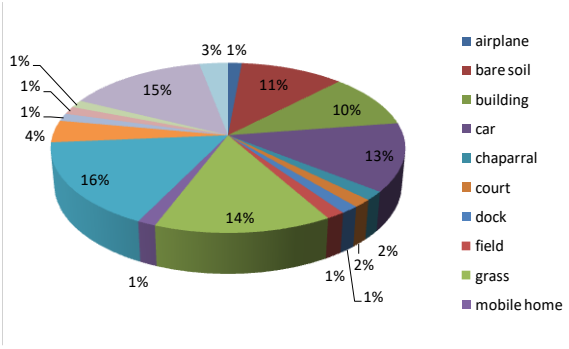

Figure. 3 The distribution of 17 classes in DLRSD dataset

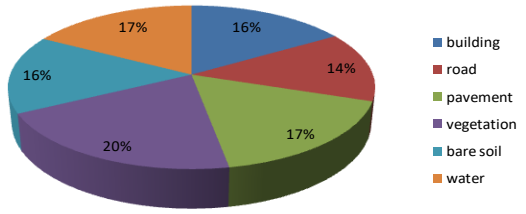

Figure. 4 The distribution 6 classes in WHDLD dataset

of total 4940 RGB images with the spatial resolution of 2 m. Fig. 4 shows the percentages of each class in WHDLD dataset.

4.2 Experimental setup and protocols

In the implementation phase, all of the models were built using Tensorflow [35], in addition, the proposed hashed sub-network contains two fully-connected layers. The hash sub-network contains 500 and κ neurons respectively in its two layers, where κ is the length of the binary code. ReLU activation function was used for all the convolution layers and fully connected layer. Also, we set the learning rate, momentum and weight decay to 10-5, 0.9 and 0.0005 respectively in Stochastic Gradient Descent (SGD) optimizer. The stopping criterion for objective value is set to 200 iterations; the mini-batch size was set to 32 as it is limited by the memory of GPUs (NVIDIA GeForce 1080Ti).

For all datasets, 70% of each class is used as training set and the rest 30% as the test set. All images are squared and resized to 100×100 for extracting fixed-dimensional features. The images are fed into the proposed retrieval model to obtain the hash codes. To overcome over-fitting, different data augmentations were employed to train the proposed framework.

The precision-recall curve is used to evaluate different components of the proposed framework. Precision is the fraction of positive satellite images among the retrieved images, while recall is the fraction of positive satellite images that have been retrieved over the total amount of positive images. On the other side, to evaluate the performance of hashing approaches, the MAP is used to measure the performance of searching images on satellite image datasets. MAP can be calculated as show in Eq. (7).

$$MAP = \frac{\sum_{q=1}^{Q} avg(P(q))}{Q} \tag{7}$$

Where Q is the number of queries. The precision P(q) for query q is calculated by dividing the total number of images by the number of images which are similar to the query image, and $avg(.)$ computes the average precision. Furthermore, MAP represents the evaluation when only top n returned images are considered to calculate the average precision.

4.3 Results

In the training phase, each image rescaled to the size of 100×100 and feed to the CNN model. We compared the proposed retrieval model with Deep Cauchy Hashing for Hamming Space Retrieval (DCH) [36], Deep Hashing Network for Efficient Similarity Retrieval (DHN) [37], Deep Triplet Quantization (DTQ) [38], and Deep Quantization Network for Efficient Image Retrieval (DQN) [39]. MAP metric is used as a performance evaluation metric in comparison. Table 1 shows the MAP-based performance comparison using different code lengths k as 12-bit, 24-bit, 32-bit, and 48-bit. From Table 1, we can observe that the proposed method outperform the conventional deep hashing methods on DLRSD, and WHDLD datasets with a significant margin.

This demonstrates superiority of multi-label feature representations compared with other approaches. Furthermore, the proposed method outperforms the DCH. Compared with the DCH, the proposed framework improved the accuracy in terms of MAP by a considerable margin of 85.4%, 87.2%, 90.8% and 92.9% for 12-bit, 24-bit, 32-bit and 48-bit code lengths respectively on DLRSD. For WHDLD, it can be noted that the proposed framework supers the DCH by 93.8%, 98.7%, 91.9%, and 94.6% on average respectively. Also, for DLRSD, it can be noted that the proposed framework achieves average accuracy for 12-bit and 24-bit code length 2% higher than the highest accuracy achieved, and 3% higher than the highest accuracy achieved for 32-bit and 48-bit code length. For WHDLD, it can be noted that the proposed framework achieves average accuracy 12% higher than the highest average accuracy achieved for

Table 1. Performance comparison of the proposed framework with other deep hash methods in terms of MAP on DLRSD and WHDLD datasets

Method	DLRSD				WHDLD			
	12-bit	24-bit	32-bit	48-bit	12-bit	24-bit	32-bit	48-bit
DCH	0.761	0.803	0.848	0.873	0.779	0.832	0.861	0.890
DHN	0.577	0.604	0.638	0.694	0.581	0.628	0.671	0.698
DTQ	0.834	0.857	0.878	0.898	0.814	0.869	0.884	0.902
DQN	0.696	0.705	0.744	0.766	0.722	0.732	0.756	0.826
Proposed framework	0.854	0.872	0.908	0.929	0.938	0.987	0.919	0.946

Table 2. The comparison of mean precision of the top K returned examples for different methods on DLRSD dataset with varied hash bits

Method	Top 10				Top-100	
	12-bit	36-bit	48-bit	12-bit	36-bit	48-bit
DCH	0.761	0.721	0.771	0.531	0.582	0.646
DHN	0.577	0.599	0.625	0.508	0.512	0.527
DTQ	0.804	0.844	0.868	0.688	0.735	0.759
DQN	0.696	0.732	0.759	0.503	0.601	0.623
Proposed framework	0.884	0.914	0.935	0.736	0.752	0.802

Table 3. The comparison of mean precision of the top K returned examples for different methods on WHDLD dataset with varied hash bits

Method	Top-10			Top-100		
	12-bit	36-bit	48-bit	12-bit	36-bit	48-bit
DCH	0.799	0.817	0.829	0.573	0.607	0.632
DHN	0.585	0.609	0.62	0.509	0.518	0.526
DTQ	0.847	0.862	0.861	0.707	0.753	0.787
DQN	0.722	0.748	0.771	0.621	0.646	0.664
Proposed framework	0.901	0.924	0.941	0.754	0.781	0.825

12-bit and 24-bit code length, and an average accuracy 3% and 4% higher than the highest accuracy achieved for the 32-bit and 48-bit code length respectively.

Next, Tables 2 and 3 shows the average precision of the Top-10 and Top-100 retrieved image samples by applying different hashing methods on the two datasets. We can observe that the DTQ and DCH methods achieve relative better results among the batch-based hashing methods under varied hash bits. For the online hashing methods, the proposed method achieves better results compared with the competitors in most cases. By comparing the proposed framework with other baseline hashing methods, it can be noted that our proposed framework achieves a comparable performance on DLRSD and WHDLD datasets while sometimes achieves even better results than all of the other compared approaches on WHDLD datasets, which has indicated the effectiveness of the proposed framework.

The average precision with respect to different retrieved samples and the precision-recall curves of compared hashing methods on the two datasets are shown in Fig. 5 (a)-(l). It can be observed that proposed framework consistently outperforms other methods when the retrieved images increase and the improvements are more notable for long code length. Precision–recall curve reflects the overall image retrieval performance of different hashing approaches. In Fig. 5 (a)-(l), it can be also finding that the proposed framework achieves the best results among the compared methods. The proposed framework has comparable and much better overall performance than other compared approaches on the two datasets.

5. Conclusion

In this paper, we proposed multi-level features attention learning framework for multi-label remote sensing image retrieval. The proposed framework acquired discriminative features for multi-label image retrieval. The multi-level features attention was performed to enhance the discriminative relevant features. Extensive experiments were conducted on two benchmark image sets: WHDLD and DLRSD, to verify the effectiveness and efficiency of our proposed framework. The results obtained showed that the proposed framework outperforms the compared techniques mentioned in section 4.3. It

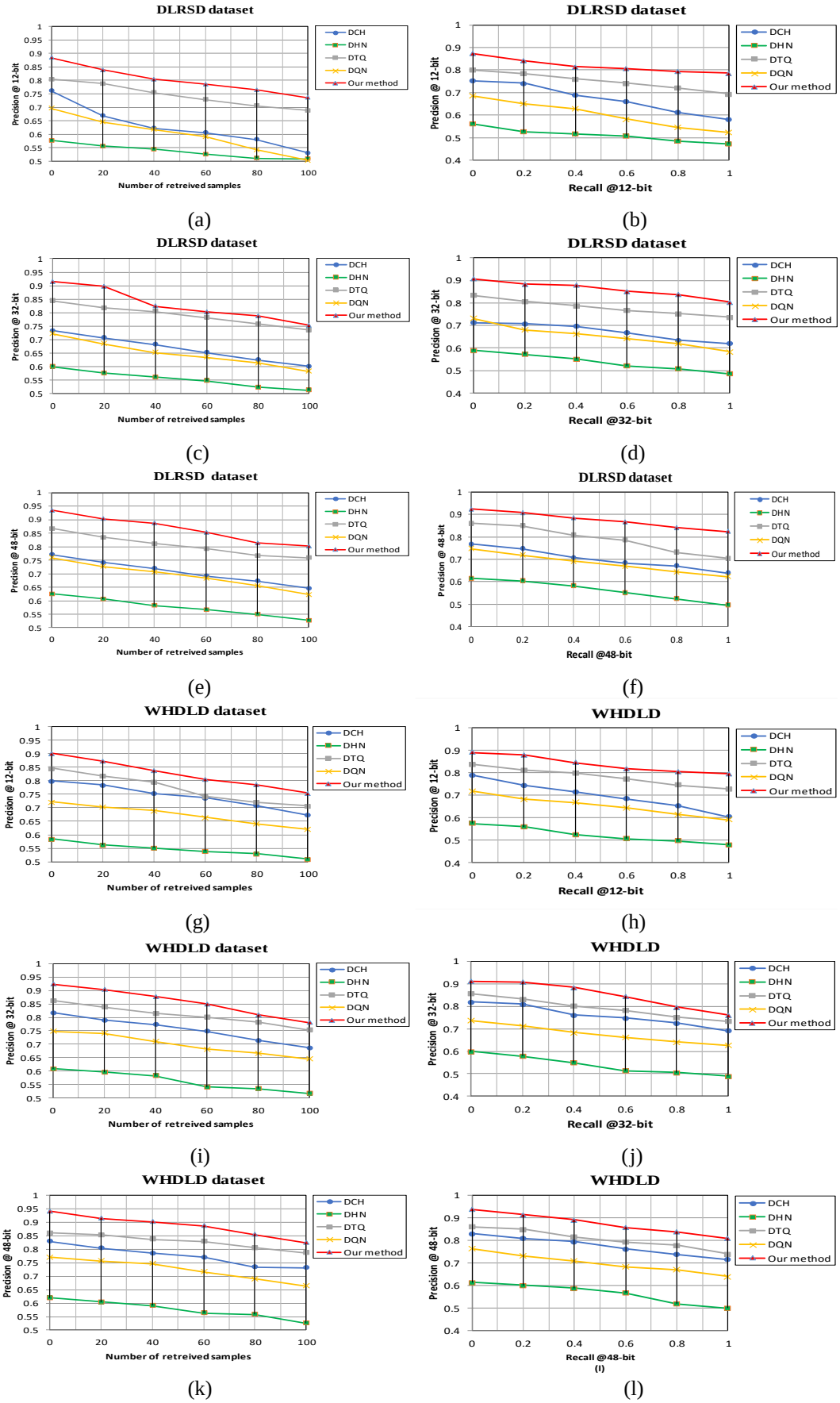

Figure. 5 The average precision with respect to different retrieved samples and precision-recall curves for the compared methods on the two datasets: (a)–(f) DLRSD and (g)–(l) WHDLD

achieved higher accuracy (2%) for 12-bit and 24-bit code lengths and higher accuracy (3%) for 32-bit and 48-bit code lengths, when it was applied on DLRSD. It also achieved higher accuracy (12%) for 12-bit and 24-bit code lengths and higher accuracy (3% and 4%) for 32-bit and 48-bit code lengths respectively, when it was applied on WHDLD.

In the future, a multi-label remote sensing image retrieval method with weak supervision will be investigated to overcome the expensive cost and time of manual annotation.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Marwa Sayed Moustafa. and Sayed Ahmed; methodology, Sayed Ahmed, Amal Ahmed Hamed, and Marwa Sayed Moustafa; software, Marwa Sayed Moustafa and Amal Ahmed Hamed; validation, Amal Ahmed Hamed and Marwa Sayed Moustafa; formal analysis, Amal Ahmed Hamed; writing—original draft preparation, Marwa Sayed Moustafa and Amal Ahmed Hamed; writing—review and editing, Marwa Sayed Moustafa, Sayed Ahmed, and Amal Ahmed Hamed; visualization, Amal Ahmed Hamed and Sayed Ahmed.

References

[1] S. Sudha and S. Aji, "A Review on Recent Advances in Remote Sensing Image Retrieval Techniques", *Journal of the Indian Society of Remote Sensing*, pp. 1-11, 2019.

[2] L. Aswathi and K. Anoop, "A Review on Various Content Based Remote Sensing Image Retrieval", *Advancement and Research in Instrumentation Engineering*, Vol. 2, No. 3, 2020.

[3] H. H. Bhatt and A. P. Mankodia, "A Comprehensive Review on Content-Based Image Retrieval System: Features and Challenges", In: *Proc. of Data Science and Intelligent Applications, Springer*, Singapore, pp. 63-74.

[4] L. B. Damahe and N. V. Thakur, "Review on Image Representation Compression and Retrieval Approaches", In: *Proc. of Technological Innovations in Knowledge Management and Decision Support, IGI Global*, pp. 203-231, 2019.

[5] G. Sumbul, J. Kang, and B. Demir, "Deep Learning for Image Search and Retrieval in Large Remote Sensing Archives", *arXiv preprint, arXiv*:2004.01613, 2020.

[6] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 2, pp. 818-832, 2012.

[7] G.-S. Xia, X.-Y. Tong, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation", *IEEE Transactions on Big Data*, 2019.

[8] J. Rodrigues, M. Cristo, and J. G. Colonna, "Deep hashing for multi-label image retrieval: a survey", *Artificial Intelligence Review*, pp. 1-47, 2020.

[9] L. Fan, H. Zhao, and H. Zhao, "Distribution consistency loss for large-scale remote sensing image retrieval", *Remote Sensing*, Vol. 12, No. 1, p. 175, 2020.

[10] T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content", In: *Proc. of the IEEE International Geoscience and Remote Sensing Symposium*, Vol. 4, pp. 2253-2256, 2002.

[11] A. Ma and I. K. Sethi, "Local shape association based retrieval of infrared satellite images", In: *Proc. of Seventh IEEE International Symposium on Multimedia (ISM'05), IEEE*, p. 7, 2005.

[12] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, pp. 837-842, 1996.

[13] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 2, pp. 818-832, 2013.

[14] S. K. Sundararajan, B. S. Gomathi, and D. S. Priya, "Continuous set of image processing methodology for efficient image retrieval using BOW SHIFT and SURF features for emerging image processing applications", In: *Proc. of International Conf. on Technological Advancements in Power and Energy (TAP Energy), IEEE*, pp. 1-7, 2017.

[15] Y. Li, Y. Zhang, C. Tao, and H. Zhu, "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion", *Remote Sensing*, Vol. 8, No. 9, p. 709, 2016.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, "Attention is all you need", *Advances in Neural*

Information Processing Systems, pp. 5998-6008, 2017.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, ... and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention", In: *Proc. of International Conf. on Machine Learning*, pp. 2048-2057, 2015.

[18] A. Li, J. Chen, B. Kang, W. Zhuang, and X. Zhang, "Adaptive Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition", In: *Proc. of IEEE Globecom Workshops (GC Wkshps), IEEE*, pp. 1-5, 2019.

[19] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4438-4446, 2017.

[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, ... and X. Tang, "Residual attention network for image classification", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3156-3164, 2017.

[21] Q. Guan and Y. Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning", *Pattern Recognition Letters*, Vol. 130, pp. 259-266, 2020.

[22] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning", In: *Proc. of the IEEE /CVF Conf. on Computer Vision and Pattern Recognition*, pp. 8776-8786, 2020.

[23] P. Li, P. Chen, Y. Xie, and D. Zhang, "Bi-modal learning with channel-wise attention for multi-label image classification", *IEEE Access*, Vol. 8, pp. 9965-9977, 2020.

[24] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A Multi-Level Attention Model for Remote Sensing Image Captions", *Remote Sensing*, Vol. 12, No. 6, p. 939, 2020.

[25] D. Cai, X. Gu, and C. Wang, "A revisit on deep hashings for large-scale content based image retrieval", *CoRR*, Vol. abs/1711.06016, 2017.

[26] S. Roy, E. Sangineto, B. Demir, & N. Sebe, "Deep metric and hash-code learning for content-based retrieval of remote sensing images", In: *Proc. of IGARSS IEEE International Geoscience and Remote Sensing Symposium, IEEE*, pp. 4539-4542, 2018.

[27] P. Li, X. Zhang, X. Zhu, and P. Ren, "Online hashing for scalable remote sensing image retrieval", *Remote Sensing*, Vol. 10, No. 5, p. 709, 2018.

[28] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning Source-Invariant Deep Hashing Convolutional Neural Networks for Cross-Source Remote Sensing Image Retrieval", *IEEE Transactions on Geoscience and Remote Sensing*, No. 99, pp. 1-16, 2018.

[29] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Fast supervised discrete hashing", *IEEE transactions on pattern analysis and machine intelligence*, Vol. 40, No. 2, pp. 490-496, 2018.

[30] P. Li and P. Ren, "Partial randomness hashing for large-scale remote sensing image retrieval", *IEEE Geoscience and Remote Sensing Letters*, Vol. 14, No. 3, pp. 464-468, 2017.

[31] D. Ye, Y. Li, C. Tao, X. Xie, and X. Wang, "Multiple feature hashing learning for large-scale remote sensing image retrieval", *ISPRS International Journal of Geo-Information*, Vol. 6, No. 11, p. 364, 2017.

[32] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module", In: *Proc. of the European Conf. on Computer Vision (ECCV)*, pp. 3-19, 2018.

[33] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 2, pp. 1144-1158, 2017.

[34] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 13, pp. 318-328, 2020.

[35] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems", *arXiv preprint*, arXiv:1603.04467, 2016.

[36] Y. Cao, M. Long, B. Liu, and J. Wang, "Deep cauchy hashing for hamming space retrieval", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1229-1237, 2018.

[37] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval", In: *Proc. of Thirtieth AAAI Conf on Artificial Intelligence*, 2016.

[38] B. Liu, Y. Cao, M. Long, J. Wang, and J. Wang, "Deep triplet quantization", In: *Proc. of the 26th ACM international Conf. on Multimedia*, pp. 755-763, 2018.

[39] Y. Cao, M. Long, J. Wang, H. Zhu, and Q. Wen, "Deep quantization network for efficient image retrieval", In: *Proc. of Thirtieth AAAI Conf. on Artificial Intelligence*, 2016.