Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.027275
Article

Tech Science Press

Triplet Label Based Image Retrieval Using Deep Learning in Large Database

K. Nithya[1,*] and V. Rajamani[2]

[1]Research Scholar, Department of Information and Communication Engineering, Anna University, Chennai, 600025, India
[2]Department of Electronics and Communication Engineering, Veltech Multitech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, 600062, India
*Corresponding Author: K. Nithya. Email: nithyakmaha@gmail.com
Received: 13 January 2022; Accepted: 23 February 2022

Abstract: Recent days, Image retrieval has become a tedious process as the image database has grown very larger. The introduction of Machine Learning (ML) and Deep Learning (DL) made this process more comfortable. In these, the pair-wise label similarity is used to find the matching images from the database. But this method lacks of limited propose code and weak execution of misclassified images. In order to get-rid of the above problem, a novel triplet based label that incorporates context-spatial similarity measure is proposed. A Point Attention Based Triplet Network (PABTN) is introduced to study propose code that gives maximum discriminative ability. To improve the performance of ranking, a correlating resolutions for the classification, triplet labels based on findings, a spatial-attention mechanism and Region Of Interest (ROI) and small trial information loss containing a new triplet cross-entropy loss are used. From the experimental results, it is shown that the proposed technique exhibits better results in terms of mean Reciprocal Rank (mRR) and mean Average Precision (mAP) in the CIFAR-10 and NUS-WIPE datasets.

Keywords: Image retrieval; deep learning; point attention based triplet network; correlating resolutions; classification; region of interest

1 Introduction

In the Content based retrieval the relevant images are retrieved from the image database for a given input image. The semantic gap is large as there is a much difference between the human interpretation of an image and the machine representation. This made the image retrieval a difficult task. But the Convolutional Neural Network and deep learning methods reduce this semantic gap by representing the images as list of high level features. Then these features are likened with the features in the image database and the relevant images are extracted. The basic method used to find relevant images is pair-wise label similarity in which the distance between the features of the images is found by either Manhattan distance or knn.

But, these methods have their own limitations. For example, the similar images with different background and dissimilar images with same background. In this scenario, the hash codes generated are less use [1]. So a triplet loss function is used to rank the labels. This function is made with 3 images. (a)

Anchor image, (b) a positive, which is close to anchor image and (c) a negative image that is different from the anchor image.

The input of this study is as ensued:

A new triplet label incorporating context-spatial resemblance measure competent of apprehending the contextual-spatial info of images devoid of employing label information is proposed. An end-to-end deep triplet hashing framework to study hash codes having the maximum ability, and this is known as the Point Attention-based Triplet Hashing (PABTN) network is used. Feature extraction by several stages is incorporated in the pre-processing phases. Later, the context-based spatial-attention module data having the feature is optimized to the convolutional neural network (CNN) architecture for advancing discriminative ability by apprehending the Triplet label info and mapping this into the hashing space.

The subsequent is how the rest of the paper is organized: The second section is devoted to a review of the literature. The proposed methodology is described in Section 3. The results and comments are summarized in Section 4, and the paper is concluded in Section 5.

2 Literature Survey

Bag of Visual Words is proposed [2] to retrieve the images from the database. For this Bag of Visual words are used to represent the extracted features of the image. Clustering and Image indexing were also used and Cosine similarity was used to retrieve the images. In [3] considered scale feature transform oriented Fast associated and BRIEF to detect features. They used K-means clustering algorithm to analyse the data. To further improve the performance, locality preservative projection was used to reduce the number of feature vectors.

A multimedia feature vector graph was proposed by joining the existing image and video analysis algorithm, Unified semantic annotation, indexing and retrieval model. The proposed model provided the accurate semantic indexing and retrieval with the help of machine learning.

Bag of Visual Words (BoVW) are used in CBIR for classification and image recognition. This method is enhanced by [4] to increase the accuracy of the retrieved candidates. They used Approximate Nearest Neighbourhood algorithm to reduce the signature construction time. apply

To improve the semantic gap, a radix trie indexing model was proposed by [5] on the basis of visual semantic indexing. They identified the visual descriptor space and joint and the RTI model is applied to integrate the visual semantic space. Then, a spark dispersed perfect is applied. Based on mean values of top ranked images' features, [6] reformulated the new expand query image. Then they selected only the most important features to improve the relevancy and retrieval.

The unlabeled-data-based indexing and retrieval method was proposed by [7] in remote sensing data retrieval. They used self-supervised learning process to learn the visual features from the large datasets.

In the [8] proposed a neighbourhood rough set approximation definition concept to measure the relevancy of the visual features. Thus the reduced visual features made the retrieval process faster. [9] added the image copyright protection and traitor tracking in the CBIR. For this they used one way hashing algorithm and XOR operation to defend the copyright of images. In order to track the traitors, a reversible information hiding algorithm was used.

In order to efficiently retrieve the encrypted images from the cloud, an improved CNN based hashing method was proposed by [10]. Initially, the image size is increased to make all the images are of equal in size. Then the parameter values are reduced and in the second phase, a compact binary hash code is generated by the addition of hash layer at the end of the network.

The Encryption alongside Compression methodology was proffered by [11] for retrieving images safely.

The size of the Remote sensing data is becoming larger. The introduction of Deep CNN model made the Content based sensing image retrieval more comfortable. But learning the features of RS images and retrieving from a large scale database are very difficult. In order to overcome these difficulties, [12] introduced feature and hash based learning. In this method, they learnt deep feature by using CNN model and they combined antagonistic hash learning model. The dense features of RS images are learnt by DLFM with higher retrieval precision. The dense features are mapped onto compact hash codes by AHLM maps.

In the [13] proposed a hybrid approach to enhance the relevancy. For this they used user interaction and hybrid features of images. It supports text based and context based image retrieval process. The retrieved images are repositioned depending on their visual and textual similarities of the inquiry images.

Classification based CBIR is used popularly to reduce the search space. But the large number of images in each classification makes the retrieval process as more time consuming. To overcome this, [14] proposed a post dynamic clustering methods to create clusters within class based on their semantic order. A semantic cluster ordering technique was also used to improve the efficiency.

Feature weighting based algorithm was proposed by [15] to improve the image classification. In this method, they used C-Means clustering algorithm to group the input images and a local feature weighting method was used to evaluate the feature weights. First the corresponding cluster is found by matching with the query image and then the most similar images are found.

Most of the existing works focused only on the semantic information by extracting the last fully connected layer. A multi-level supervised hashing was used in [16] to learn the deep image features in multiple levels. A Multiple hash table method comprises of semantic contents and structured information.

In order to generate a robust features, a dictionary learning (DL) approach was proposed by [17] by using CNN model. For this, the initial value of the dictionaries is taken from the middle layers. Then the DL structure produces the lambda vectors and is converted in to binary values. The hash codes of the features are generated by DL.

Even though Deep hashing methods are used to generate more effective features, finding the similarity between the images are uncertain. In order to minimize this problem, a novel hashing method was proposed by [18]. In order to learn more effective binary codes, they make use of the fuzzy rules to model the uncertainties in data. A generated hamming is formulated in the convolutional layers and fully connected layers. Thus they combined fuzzy logic with DNN to improve the retrieval accuracy and training speed.

3 Proposed Methodology

Learning minimal hash code from the image is the purpose of the proposed work. This learnt code must meet the following criteria. (i) The positive images should be represented very close to the hash code generated. (ii) the negative images must have larger distance from the hash codes. (iii) The Region Of Interest has to be represented efficiently in discriminative hash codes. (iv) the hash codes should have complete representation for minor samples and their classification. The image features and the hash codes are learnt at the same time by training the PABTN in an complete manner. Fig. 1 Illustrates the comprehensive block diagram of the proposed method.

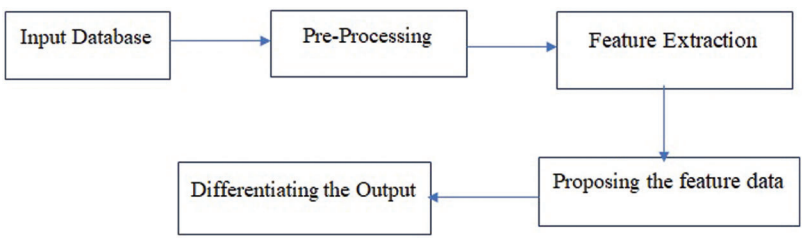

Figure 1: Comprehensive block schematic illustration of the proposed methodology

The Pre-process experiences several phases for extracting the featured data context-based feature optimization technique for giving resistance contra noise. Point Attention-based network is employed for extremely imbalanced high-resolution images. The propose codes framework employed in this work depends upon a Point Attention-based network (PABTN) for the execution. To improve the ranking efficiency, a correlating resolution for classification, Region of Interest, a small space information loss, a new triplet cross-entropy-loss and a spatial care device are proposed in [19]. Random high-resolution images are given input to the network. Fig. 2 illustrates high-resolution image data. Our proposed technique would be a pre-processing module as illustrated in Fig. 3.

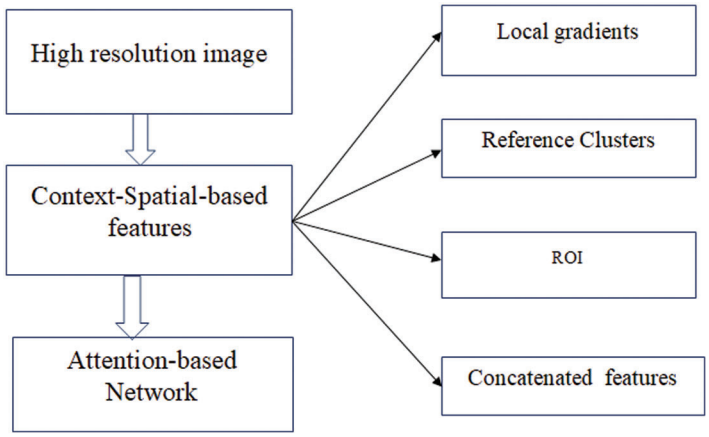

Figure 2: Outline sketch of featured data out of the pre-processing phase of deep learning

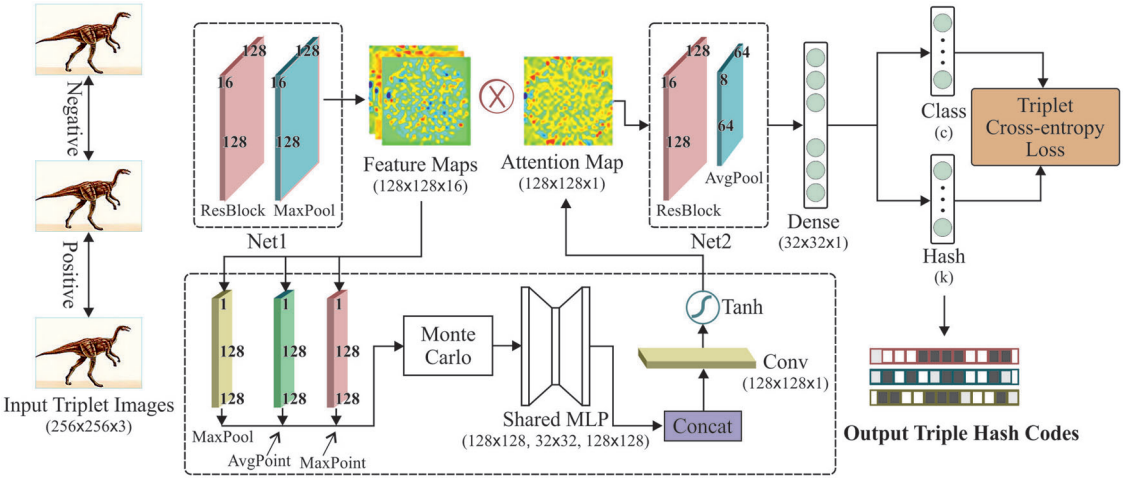

Figure 3: Proposed PABTN network structure

3.1 Four-Phase Technique

1. Local gradients remain feature vectors of disparities. The distance between the pivotal pixel p and the neighboring pixels with 3 × 3 × 3 is calculated and is taken as feature vectors of each pixels in hyperspectral cube.

2. The high threshold and low threshold values of feature vectors are present in reference clusters. For every pixel, the feature vectors are calculated with 9 × 9 neighborhood pixels. This is the reference cluster of p.
3. Region of Interest (ROI). In this phase, we compute these feature vectors for every pixel p in the hyperspectral cube. This remains a subset of the image and is sketched at the provided borders on an object's image.

To perceive ROI, the object detection method in [20] is used. A high-resolution image is given as input for which the outputs are multiple predictions. A grid with m × m size is applied on image and for which the following predictions are done.

(i) confidence (c). The confidence value states that the degree of certainty that ROI prevails concerning the grid cell. That is the certainty that the grid cell consists of the ROI center x intersection on union ratio and the grid region.

(ii) The coordinates (x, y) of the center of the ROI, of a grid cell. A grid cell is solely in charge of the ROI of which the center remains within the grid cell. Thus, $0 \leq x \leq 1$ and $0 \leq y \leq 1$. ROI's width (w) and height (h) remain corresponding to the dimension of the image and a class label vector (P).

4. Concatenated features remain entire feature vectors. For a high resolution image, a single context based feature vector is formulated by combining reference clusters and local gradients for every p.

Spatial and context data are added by a context spatial-attention unit. This is done by combining channel axis with MaxPool, MaxPoint and element-wise mean. The layer in which the hash code is generated, is mapped with the dense layer. The dense layer is also mapped with classification layer. Finally the classification and hash code generation layer produces triplet cross-entropy loss.

3.2 Point Attention-Based Triplet Network (PABTN)

Small sample ranking is one of the complex task in the case based image retrieval. Because the hash code contains small sample information loss, ROI and the classifications. If the rank of the same class images is below the expected value, then the relevancy prediction is very low.

To improve the efficiency of the ranking, a new triplet entropy loss is combined with a spatial attention method and triplet labels in the proposed method. To extract the visual features and map compact hash codes, Point Attention Based Triplet Network (PABTN) is proposed. The proposed method is depicted in Fig. 3. It consists of two stages. The first stage consists of attention map which is responsible for adding spatial and contextual information to the image. The second stage consists of dense layer to produce the compact hash codes and classified outputs.

In the first module, the residual block is used to improve the learning rate by skipping connections, which is followed by max pooling layer. Then a spatial attention module is also present to extract context and spatial data. In the second stage, another residual block is added followed by average pooling layer. Tab. 1 provides the network setup information.

Table 1: Network setup information

Size of the filters	3 × 3
Stride	2
Normalization	Batch Normalization
Activation Function	ReLU

In PABTN, the size of the feature map for the input is 128 × 128 × 16. This input feature is reduced into 128 × 128 × 1 by considering the inter-spatial association of features. We assert that every ROI area chiefly contains informative portions and prominent phases, both react to the gradient back-propagation. To produce two spatial context descriptors, each element is calculated for MaxPoint and AvgPoint operations along the channel axis. They are denoted as F_{avg} and F_{max}

$$F_{max} = [f_1 \ldots f_i \ldots f_{128\times128}] \qquad fi = \max_{1\le c\le 16} \chi_i(c)$$

$$F_{avg} = [f_1 \ldots f_i \ldots f_{128\times128}] \qquad fi = \overline{\chi_i(C)} \tag{1}$$

in which Zi(c) represents the response of the ith point in the cth channel. For concentrating on instructive sections, the max-pooling (MaxPool) procedure is employed for creating a max-pooled feature descriptor: fmaxp. The entire three describes are individually fed forward towards a shared embedded Monte-Carlo (MC)-multi-layer perceptron (MLP) to denoise.

3.3 Temporal Rendering Monte Carlo

Monte Carlo is rendered alongside the neural networks for creating a sequence of images to retrieve. The above-said focuses on the rendering quality and the rest extra heed towards the equilibrium betwixt quality and speed for attaining an interactive processing rate. In addition to speed, a requisite contemplation is required for increasing temporal stability betwixt frames for evading low-frequency variances, which might result in flicking artifacts in image retrieval. In image retrieval rendering, the temporal consistency limitations could be inflicted upon a temporal window in which the Spatial features out of the former and upcoming independent frames could be excerpted and warped employing motion vectors for complementing the center frame. We proposed an adaptive rendering methodology, which dispenses samples through the spatiotemporal joint practicing of neural network-based sample predictors and MC denoisers over multiple consecutive frames, that enhances temporal steadiness and image fidelity. An enhanced s predictor allows the learning of Spatio-temporal sampling schemes that aid the rendering engine for adaptively positioning the additional sample in un-concluded areas or tracking specular highlights in which high-frequency particulars are tough to rebuild.

The Maximum point response over feature maps is produced by MaxPoint procedure. A hidden layer is present in the SharedMLP and the size is equal to the dense layer. Three input descriptors shares the MLP weights. Then the attention map is generated by convolving the three descriptors. This attention map is used to encode the regions. A filter of size 3 × 3 is used in the convolutional layer. The tangent activation function is used.

The spatial-attention is calculated as follows:

$$A(f)) = T(F([MLP(fAvg); MLP(fMax); MLP(fmaxp)])) \tag{2}$$

T is the tangent function, F is the convolution procedure and MPL represents the shared MLP operator. By using the Max pooling, Avg point and Max point operations, the ROI related information is accumulated to get optimized spatial attention mapping. The marginal value of ROI is prevented by MaxPooling. To optimize the noise value, the MaxPoint and the AvgPoint are used independently and given input to the Shared MLP. The ROI feature extraction from raw pixels, is focused in the learning by using the attention-based CNN. The hierarchical non linear function is used to capture the ROI information.

3.4 Denoised PABTN (Denoised Point Attention-Based Triplet Network) Proposing Framework

This segment concisely assesses general sections of administered attribute proposing frameworks defined in [19] and [20] employing neural networks for image retrieval assignments. The triplet cross-

entropy loss attains maximum-category severability and maximum propose code-differentiability by penalizing likeness and classification losses concurrently. The main concept after the triplet cross-entropy loss remains that the likeness and classification must be concurrently conserved while paradigm practicing out of triplet input images.

The paradigm f :IMG−⟶ H' excerpts a pseudo K-bit hash h '∈ [−1, 1] K out of a provided input image Img,

$$H' = g(x) \tag{3}$$

and a quantized K-bit binary propose h ∈ {−1, 1} K is acquired by getting the indication of h ',

$$h = sign(h'). \tag{4}$$

To retrieve the images, it could be advantage out of Euclidean distance betwixt two hash codes that are described as,

$$distEucllidean(hm, hn) = 1/2(K - < hm, hn >) \tag{5}$$

in which hm and hn portray the hash codes of mth and nth images and <·, ·> portrays an inner product.

Denoised PABTN are particularly favoured for making attribute proposing structures owing to their capacity in capturing images' contexts.

To describe mathematically, provided M practicing images I = {I1, . . ., Im}, category labels CL = {1, . . ., c}, and triplet labels TL = {(Q1, P1, N1), . . .,(Qi, Pi, Ni), . . .,(Qm, Pm, Nm)} are produced by haphazardly choosing two images as a input and a positive image out of the similar category (Qi and Pi) and haphazardly choosing a negative image out of disparate categories (Qi and Ni). In the triplet labels, the input image of index Qi remains alike the positive image Pi and unlike the negative image Ni in which the index i∈ {1, . . ., m} is haphazardly chosen out of the m practicing images. In order to sample triplet labels, tiny samples are chosen as the negative image of a huge sample. That is to say, tiny samples are reutilized on many occasions during paradigm practicing for conserving the data in the propose codes.

As illustrated in Fig. 2, the last but one layer of PABTN z is transferred towards a proposing layer with a view to output h' in which the proposing layer remains a completely linked layer having tanh activation. For notational simplicity, we contributed the PABTN section as fPABTN and the proposing layer as fpropose. The practicing process of feature proposeing paradigm too could divide into twin sections: (i) PABTN section and (ii) practicing the proposeing layer. The criteria θPABTN of the PABTN section are pre-practiced alongside cross-entropy loss betwixt the outputs of the softmax z '∈ (0, 1)n and the one-hot encoded labels l ∈ [0,1]n as common categorizer practicing alongside several categories. The similar z' and the loss LPABTN are described as

$$z' = softmax(Wlogitz + blogit) \tag{6}$$

$$LPABTN = - < l, logz' > . \tag{7}$$

To practice the criteria θpropose of the proposeing section, proposed Bayesian architecture having a provided Triplet labels similarity sijk. The Maximal Posterior assessment logp(H' |S) ∝ log p(S|H')p(H') having the provided Triplet labels resemblance is evolved as,

$$\log p(S|H')p(H') = \sum_{s_{ij} \in S} \log p\left(s_{ij}|h'_i, h'_j\right) p(h'_i) p\left(h'_j\right) \tag{8}$$

in which the Triplet label similarity sijk remains

$$S_{ij} = \begin{cases} 1, & if\ x_i\ and\ x_j\ has\ the\ same\ label \\ 0, & Otherwise \end{cases} \tag{9}$$

The paradigm describes the conditional similarity of sijk as a Triplet logistic function and the loss function is described as a Triplet cross-entropy as,

$$p(S_{ij}|h'_i, h'_j) = \begin{cases} \sigma\left(\langle h'_i, h'_j\rangle\right), & S_{ij} = 1 \\ 1 - \sigma\left(\langle h'_i, h'_j\rangle\right), & S_{ij} = 0 \end{cases} \tag{10}$$

$$L_{hash} = \log(1 + exp\langle h'_{i,} h'_j\rangle - S_{ij}\langle h'_{i,} h'_j\rangle \tag{11}$$

Notice that Lhash employs the continuous h' rather than the quantized propose code h. The quantization error issue could be lessened by standardizing the prior p(h') alongside the bimodal Laplacian dispensation [3]. During the practicing of θpropose alongside Lpropose, the practiced criteria θPABTN could too be enhanced with modified learning rates.

The PABTN tries to learn triplet hash codes from 3 input images. The produced hash codes H_{Query} = {h1, k, . . ., hm,k}, H_{Pos} = {h1,k, . . ., hm,k}, and H_{Neg} = {h1,k, . . ., hm,k}. If the database is larger, the hash codes should be compact to make it as scalable. The distance between the H_{query} and H_{pos} should be lesser than the distance between the H_{query} and H_{neg}. The ground truth labels for the triplet labels T are, G = {(g_{Q1},g_{P1},g_{N1}), (g_{Qi},g_{Pi},g_{Ni}), (g_{Qm},g_{Pm},g_{Nm})}. Where g_{Qi},g_{Pi},g_{Ni} ∈ L. The ranking loss form is created such that it reduces the distance between related images and increase the distance between irrelated images. The loss is given as follows.

$$L(T) = max\left\{r.m - D\left(H_{query}, H_{Neg}\right) + D\left(H_{query}, H_{pos}\right), 0\right\}. \tag{12}$$

Where D is signified distance between the produced hash codes. m is the length and r ∈ [0,1]. r controls the penalty strength between dissimilar images that differentiating degrees. Only the different pairs that falls within a specified radius are considered to calculate the loss. If r = 0, then no penalty is given. If r = 1, the generated hash codes from different images should be unique.

When the ground truth labels are given and to penalize the classification loss, a cross entropy loss is defined as follows.

$$\mathrm{L(T,\ G)} = \sum_{i=1}^{m}\{CEL(gQi', gQi) + CEL(gPi', gPi) + CEL(gNi', gNi)\} \tag{13}$$

Where CEL denotes Common-Cross-Entropy-Loss, g' represents the predicted class. L(T) is the similarity loss and L(T, G) is the classification loss.

4 Performance Analysis

For assessing the proposed methodology, six types of public image datasets are applied. Later, the query images are chosen randomly as test images out of the randomly chosen images out of the dataset as target images. The three private datasets are described as:

4.1 Dataset Description

The proposed method is compared with the existing models with the following parameters. mAP, precision-recall curve and the precision is plotted against the hamming distance. Then the hamming

distance between the hash codes are evaluated using histograms. The proposed method is experimented with the following datasets.

i) CIFAR-10 : This includes 60,000 images having 10 classes. Every image consists of three channels and the dimension of 32 × 32. We rescaled every image to the dimension of 256 × 256 and center-cropped it to the dimension of 224 × 224. From the image set, 5000 images are chosen for training the network. Approximately 0.5% images are chosen as random for input images and rest of the images are in the database.

ii) NUS-WIDE: It consists of totally 2,69,648 images which are having multi labels. All images are grouped under 81 concepts in the web.

iii) MIRFLICKR-25k: This comprises 25,000 images which were downloaded from Flickr social photography site.

4.2 Assessment on Limits of Triplet Label Resemblance

The following assessment exhibits that our methodology is curative for the two limits of employing solely pairwise label similarity; Scarce propose code diversity and weak execution on misclassification images.

Three metrics were employed for calculating the precision and retrieval excellence in our experimentations.

a) Hit Ratio (H). This is defined as the sum of images in the retrieved images are relevant to the input image.

b) Average Precision (A). This is defined as, in the retrieved image list, the average of the rank position of the relevant images.

c) Reciprocal Rank (RR). The RR is defined as the mutual of the rank of top1 relevant image in the retrieved list.

The precision vs recall curve is portrayed in Fig. 4.

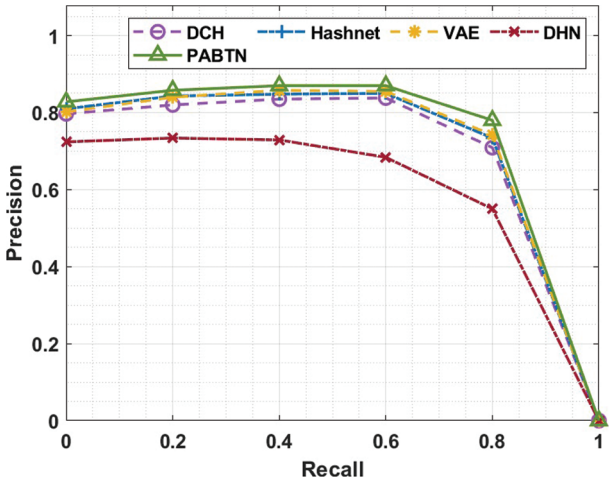

Figure 4: Recall *vs*. precision

From the Tab. 2, it is inferred that, the proposed method exhibits the better mAP. This means that more number of applicable images in the retrieved image list are ranked ahead. The av_H of the projected method is higher than the existing approaches. This denotes that the retrieval list contains more number of relevant images. This is portrayed in Fig. 5. The mRR values of the proposed method for all datasets are above

0.92 which indicates that the retrieved images consists of same class as of input image. The mAP score values are plotted in Fig. 6.

Table 2: The executions (AV_HR, mAP, mRR) on the CIFAR-10, NUS-WIDE, and MIRFLICKR-25k datasets

Dataset	Methodology	Av_HR	mAP	mRR
CIFAR-10	DHN	0.6063	0.5182	0.8077
	DCH	0.6761	0.5889	0.8555
	Proposenet	0.6146	0.5291	0.8187
	DSHSD	0.7074	0.6376	0.8918
	CSQ	0.7180	0.6359	0.8441
	Greedy Propose	0.7220	0.7220	0.9033
	Proposed	0.8342	0.8271	0.9256
NUS-WIDE	DHN	0.6273	0.6382	0.8177
	DCH	0.6961	0.6589	0.8655
	Proposenet	0.7156	0.6691	0.8287
	DSHSD	0.7284	0.6776	0.9018
	CSQ	0.7460	0.6359	0.8541
	Greedy Propose	0.7540	0.7420	0.9143
	Proposed	0.8642	0.8471	0.9656

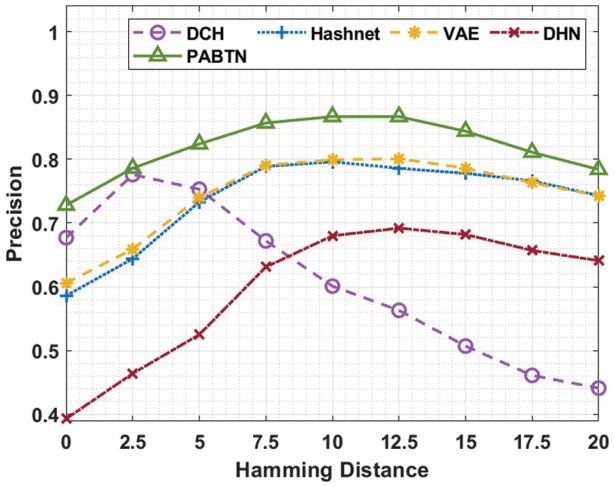

Figure 5: Hamming distance *vs*. precision

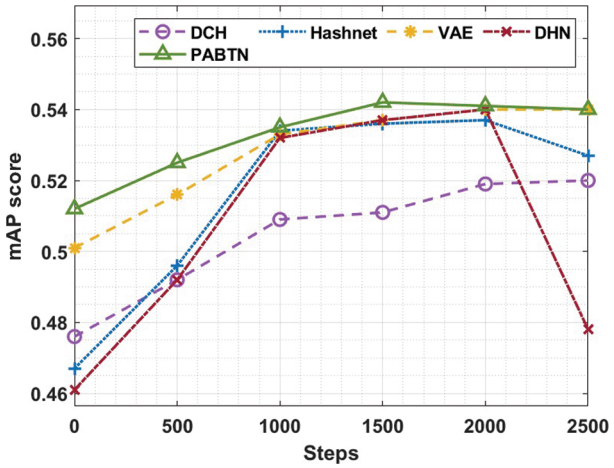

Figure 6: mean Average Precision (mAP)

5 Conclusion

In order to efficiently retrieve the images in the large database, a novel triplet based label that incorporates context-spatial similarity measure is proposed. To minimize the misclassification errors, a Point Attention Based Triplet Network (PABTN) is also introduced. Correlating resolutions for the classification, and slight sample info loss containing a new triplet cross-entropy loss, ROI, a spatial-attention instrument and triplet labels based on findings are used to recover the ranking mechanism. The proposed method is simulated and the results were compared with the state-of-the-art retrieval methods. The results show that the performance of the proposed method is better. The proposed method only considered the ROI, sample information loss and spatial attention mechanism as prime parameters. In the future, the ranking mechanism can be further improved by considering more number of parameters.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 815–823, 2015.

[2] A. Rudrawar, "Content based remote-sensing image retrieval with bag of visual words representation," in *2018 2nd Int. Conf. on I-SMAC*, Palladam, India, pp. 162–167, 2018.

[3] K. Munish, C. Payal and G. Naresh Kumar, "An efficient content based image retrieval system using BayesNet and K-NN," *Multimedia Tools Applications*, vol. 77, no. 16, pp. 21557–21570, 2018.

[4] A. Ouni, "A machine learning approach for image retrieval tasks," in *2020 35th Int. Conf. on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–5, 2020.

[5] N. Krishnaraj, M. Elhoseny, E. L. Lydia, K. Shankar and O. ALDabbas, "An efficient radix trie -based semantic visual indexing model for large-scale image retrieval in cloud environment," *Software: Practice and Experience*, vol. 51, no. 3, pp. 489–502, 2021.

[6] A. Ahmed and S. J. Malebary, "Query expansion based on top-ranked images for content-based medical image retrieval," *IEEE Access*, vol. 8, pp. 194541–194550, 2020.

[7] K. Walter, M. J. Gibson and A. Sowmya, "Self-supervised remote sensing image retrieval," in *IGARSS, 2020 IEEE Int. Geoscience and Remote Sensing Symp.*, Waikoloa, HI, USA, pp. 1683–1686, 2020.

[8] J. Zhao, C. Zhang and F. Yao, "Efficient selection of visual features in automatic image retrieval," in *2020 12th Int. Conf. on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 361–365, 2020.

[9] Z. Wang, J. Qin, X. Xiang and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systems*, vol. 27, no. 3, pp. 403–415, 2021.

[10] W. Pan, M. Wang, J. Qin and Z. Zhou, "Improved CNN-based hashing for encrypted image retrieval," *Security and Communication Networks*, vol. 2021, pp. 5556634:1–5556634:8, 2021.

[11] K. Iida and H. Kiya, "A content-based image retrieval scheme using compressible encrypted images," in *2020 28th European Signal Processing Conf. (EUSIPCO)*, Amsterdam, Netherlands, pp. 730–734, 2021.

[12] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang *et al.,* "Deep hash learning for remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3420–3443, 2021.

[13] B. J. Dange, S. K. Yadav and D. B. Kshirsagar, "Enhancing image retrieval and re-ranking efficiency using hybrid approach," in *2020 Int. Conf. on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, Aurangabad, India, pp. 20–26, 2020.

[14] J. Pradhan, A. K. Pal, M. S. Obaidat and S. H. Islam, "A post dynamic clustering approach for classification-based image retrieval," in *2020 Int. Conf. on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, Sharjah, United Arab Emirates, pp. 1–7, 2020.

[15] S. Ghodratnama and H. A. Moghaddam, "Content-based image retrieval using feature weighting and C-means clustering in a multi-label classification framework," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 1–10, 2021.

[16] W. W. Y. Ng, J. Li, X. Tian, H. Wang, S. Kwong *et al.,* "Multi-level supervised hashing with deep features for efficient image retrieval," *Neurocomputing*, vol. 399, no. 4, pp. 171–182, 2020.

[17] Ş. Öztürk, "Convolutional neural network based dictionary learning to create hash codes for content-based image retrieval," *Procedia Computer Science*, vol. 183, no. 4, pp. 624–629, 2021.

[18] H. Lu, M. Zhang, X. Xu, Y. Li and H. T. Shen, "Deep fuzzy hashing network for efficient image retrieval," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 166–176, 2021.

[19] Y. Cao, B. Liu, M. Long and J. Wang, "HashGAN: Deep learning to hash with pair conditional wasserste in GAN," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 1287–1296, 2018.

[20] Y. Cao, M. Long, B. Liu and J. Wang, "Deep cauchy hashing for hamming space retrieval," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 1229–1237, 2018.