A REVIEW ON MULTILABEL IMAGE RETRIEVAL

Surbhi D. Bhosale[1] N.M.Shahane[2]

[1]M.E. Student [2]Associate Professor
[1,2]Department of Computer Engineering
[1,2] K. K. W. I. E. E. R., Nashik
Savitribai Phule Pune University, Maharashtra, India
[1]surbhi.bhosale11@gmail.com

ABSTRACT

Images on web have become one of the most important information for browsers however, the large number of results retrieved from images search engine increases the difficulty in finding the intended images. Images with more than one tag, i.e. multi label images make this task even more difficult. Geodesic Object Proposal method is proposed for generating regions followed by spatial pyramid pooling to learn features for each of these regions. This feature vector then can be used to in multiple ways such as clustering or hashing to retrieve similar images given a query image.

Keywords: *multi-label image annotation; convolutional neural network, clustering, hashing*

1. INTRODUCTION

The modern information society is overloaded with a huge amount of data. Various search engines such as Google, Bing, and Yahoo make it easier for mass amounts of people to access information according to their need. Since there are a huge number of images on the web, it is hard to find the intended images by simple query. The results of a query may belong to some image which has more tags apart from query image. For example, a query Pluto in Google Images, the results should contain two different types of images: one is Pluto in the solar system; the other is just the image of Pluto planet. It is worth noting that the results of the two scopes are mingled. Generally search engines give results of only second type, but if some other multi labelled image which is relevant to image should definitely be included in result. This is where multi-label image retrieval comes into picture.

Recently, web image search engine, like Google image, only returns a small fixed number of images. In the perspective of users, it is important to give some other relevant images as results as well to make informed decision about which is the most desired and accurate result. To solve the problem, a frame work is proposed which first generates region proposals using Geodesic Object Proposal followed by deep convolutional neural network with the aid of spatial pyramid pooling which learns feature for given proposals. Clustering technique can the be used to find nearest tags to given image. After label probability is calculation, we can use direct threshold mapping or hashing to retrieve most relevant results.

2. RELATED WORK

Feature analysis and similarity measure:-many-early years studies of CBIR focus on feature analysis and similarity measure. Similarity matching is significant issue in CBIR. So many image retrieval applications are based on shape feature and color feature [1]. G. Hinton and R. Salakhutdinov [2] proposed the most recent research work in deep learning based hashing technique is Semantic Perseverating Hashing. This is Semantic Perseverating Hashing method constructs a deep learning model which is used to explore binary units that are not visible, which can be model input text data. Such a deep learning based model is prepared as a layer of Restricted Boltzmann Machines (RBMs) technique. Once the learning RBM multiple layer model is prepared through fine-tuning and pre-training step which is performed on a collection of various digital documents, then the hash code of any collected document is gained by a thresholding point as a output of the deepest layer. Such hash codes are provided by the deep RBM layer model that shows semantically similar relationships between documents, in which each hash code or hash key is used as a memory address of collected documents that locate its associated documents. Semantically similar between the documents are mapped into nearest neighborhood memory addresses, hence it performs efficient searching a documents using hash lookup table.

Among all the previous supervised and unsupervised hashing techniques including the previously mentioned deep neural networks, this techniques generate the hash code that takes a hand-crafted visual features vector representation which are extracted feature from an input image. Hence, the generated hash code values quality are not independent on the quality of hand-crafted visual features vector representation that are extracted feature from an input image. Hence this problem is minimize with a most recent hashing technique called as convolutional Neural Network Hashing was introduced by R. Xia, Y. Pan, H. Lai [4] to combine both image feature learning as well as hash code learning. The supervised learning information, this model consists of a two stage. In this one stage for learning approximate hash codes and a one stage for training a deep convolutional Neural Network (CNN) which was proposed by A. Krizhevsky, I. Sutskever [3], that produces outputs as a hash values. In CNNs, it continuously learning image features vector representation and hash values representation that are directly working on a pixels image. R. Xia, H. Lai, Y. Pan, Y. Liu[4] proposed the most recent method Deep Convolution Neural Network Hashing technique, in which its first step it learns the image information representation and hash code bits representation. So that image information representation learning and hash code bits representation learning are taking advantages with each other. This technique is same as Deep Semantic Ranking Hashing technique, the Deep Convolution Neural Network Hashing technique combines list wise supervised learning information to train a deep convolution neural network. The deep hashing architecture contains following blocks:

1. A triplet of images are passed to the convolution neural network module and then image triplet ranking loss function is calculated to divide or to categorized the list wise supervised learning information.

2. A Shared sub network model of a convolution neural network layers is to produce the image features vector representation.

3. A divide-and-encrypt : This block is divide the intermediate image features vector into the n number of channels, each intermediate image features vector encrypted into a one bit of hash value. In divide-and-encrypt block consist of two layers a fully-connected layer for classification purpose and another layer is hash layer to produce hash value. The Deep Neural Network Hashing technique was shown to exceed the Convolutional Neural Network Hashing method along with various learning based supervised hashing technique in order to improve image retrieval accuracy.

3. OVERVIEW

The multilabel image retrieval framework consist of five modules as follows,

1. First module generates region proposal for an input image. Geodesic Object Proposal method is used in proposed module to automatically generate these regions.

2. Second module extracts features from regions proposed by the first module. Spatial Pyramid Polling (SPP) method which computes feature map of entire image only once is used here.

3. Third module calculates the label probability of each proposed region. This module outputs a probability matrix which has labels along row and proposed regions along column and each data point C_{ij} in the matrix represents probability of label i being assigned to proposed region j of the image.

4. Fourth module actually generates hash codes for the proposed representations. Firstly instance aware representation is generated and which can be further encoded with wither sematic hashing or category aware hashing.

5. Geodesic distance between the learnt feature vector and the centroid of class category can then be computed. It repeat this procedure for all class categories and select top m class categories whose geodesic distance from feature vector of query image is minimum and give their members as result.

Input Image

Input Image Input image is a multi labelled image of size 224 × 224.

N Region Proposal

Geodesic Object Proposals(GOP) method is used to identify the proposals in input image. The method is divided in four stages, In first stage, for each input image, over segmentation into superpixels (an image segments) and boundary probability map corresponding to each superpixel is computed by representing the image as weighted graph G(I) = (V(I), E(I)), where V is a superpixel and E is an edge connecting two superpixels and weights being likelihood of object boundary at corresponding image edge. Then a nearest boundary (superpixel at boundary) is computed using dijkstra algorithm. The second step is to identify set of superpixels that are likely to be located inside objects. For this seeds are placed inside image (seed is basically a super pixel in image). The first seed is placed at the geodesic center of the image which lies halfway on the longest geodesic path of the graph. Now each successive seed is placed so as maximize the geodesic distance to previous seeds. In third step, the foreground and background masks are generated from each seed. Initial approach for mask generation is to label each seed as a foreground and remaining empty area as background mask. However, this approach is further improved by a learning based approach where a pre-trained classifier is used to identify foreground and background mask. Features used by classifiers for mask generation are location relative to the seed, distance to the image boundary edges and color similarity in multiple color spaces. Finally signed geodesic distance transform (SGDT) is computed for both background and foreground masks over the image. The geodesic distance between two nodes is defined as the length of the shortest paths between the nodes in geodesic space. Good proposals, are extracted by identifying the particular critical level sets (stationary points in geodesic function) of the SGDT. Eventually non-maxima suppression is done to remove any near duplicate object proposals.

Spatial Pyramid Pooling

Previous deep convolutional neural networks method (CNNs) require a fixed-size (e.g., 224 × 224) of input image. This requirement is unreal and it may be reduce the recognition accuracy for the input images or its sub-images of

an absolute size or scale. The networks with pooling scheme, is called as spatial pyramid pooling, which is used to eliminate the fixed size of input image or its sub image. The network structure, is called as SPP-net, that can be generate a fixed-length of image representation irrespective of image size or image scale. Pyramid pooling is also powerful to object deformations. With the help of these advantages, SPP-net improve all CNN learning based image classification methods. The SPP-net rise the accuracy of a large variety of CNN architectures. On the Pascal datasets, SPP-net achieves state-of-the art categorization results using a single full-image representation. The SPP-net is also important in object detection. With the help of SPP-net, it compute the feature maps from the entire image only once, and then pool features in absolute regions or patches (sub-images) to generate fixed-length of image representations. This method avoids continuously computing the convolutional features. A spatial pyramid pooling (SPP) layer is to eliminate the fixed-size image constraint of the network. Specifically, SPP layer present on top of the last convolutional layer. The SPP layer pools the features and it generates fixed length of image data, which are then union into the fully connected layers. It perform some information aggregation at a bottom layer of the CNN network layer hierarchy i.e., in between CNN layers and fully-connected layers which is used to avoid the need for cropping or warping at the beginning. It uses new network structure SPP-net. Spatial pyramid pooling, commonly known as spatial pyramid matching or SPM, which is as an extension of the Bag-of-Words (BoW) model, is one of the most widely used methods in computer vision. It divides the image representation from finer to coarser layer, and then it combines features. SPP is a key component for image categorization and image detection. SPP is not considered in the CNN based learning method.

SPP has various characteristics for deep CNN learning method are as follows:

1. SPP is generate a fixed length of image data irrespective of the input size of image data, whereas the sliding window pooling method which is used in the previous deep CNN networks cannot generate a fixed length of image data.

2. SPP uses multiple level spatial bins, whereas the sliding window pooling method uses only a single window size. These Multi-level pooling is powerful to object distortion.

3. SPP can pool features extracted at variable scales this is because of input scales flexibility.

Label Probability Calculation Module :

This module learns probability of each region belonging to each class. If there are c class labels and it generate a probability vector for each proposal, then in Pi = ($P1^i$,..., Pc^i) Pj^i signifies the probability of proposed region i belonging to class j. First it for each proposed region i compression will be done, its corresponding feature d-dimensional intermediate feature vector which corresponds to i^{th} row of N x d feature matrix into c-dimensional vector. Then consolidate all this compressed individual vectors into one c-dimensional vector. For this cross hypothesis max polling function will be used such that for each of class its probability of being assigned to input image is maximum probability of it being assigned to any of proposed region. m_j = max { M^1_j, M^2_j , ... ,M^N_j}, for all j = 1,..., c where m_j is the consolidated probability corresponding to category j. It will use each of this mj for all class categories to calculate probability distribution p = (p1 , p2, . . . ,pc) given as,

$$m_j = \frac{exp(m_j)}{\sum_{k=1}^{c} exp(m_k)}$$

This predicted probability to compute loss function which is based on cross entropy between the probability scores and correct labels, This soft-max loss function is will be used to train back propagation convolutional neural network.

Clustering

Clustering images and identifying central nodes in a graph are complex and computationally expensive tasks. It utilize feature vectors learnt from spatial pyramid for each of region pooling as efficient representation of the hidden structure of the clustering problem. Initial cluster centers are determined by graph centrality measures. Cluster centers are fine-tuned repeatedly by minimizing fuzzy-weighted geodesic distances as more and train images are added. The shortest-path based representation is parallel to the concept of identifying similar images visually.

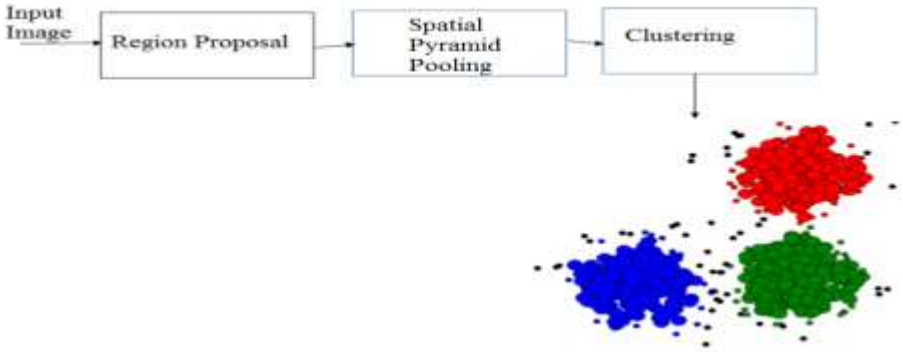

Figure-1 : Retrieval using Clustering

For each the proposed region of query image will first learn the feature vector. It then compute a simple geodesic distance between this feature vector and centroid of each the cluster which is representative of individual class category. If that distance is below some threshold value it will include that particular cluster in image labels and give top m results from that cluster. It will repeat this procedure for all the clusters. As evident this procedure can be computationally very heavy as it repeating the procedure for each of region proposal and in it for each of cluster. Solution to this problem is hashing based retrieval showed below.

Retrieval Based On Label Probability

Once it learnt max-pooled label probability of given image, it will just decide some threshold value of probability to assign given tag to image and if that probability value falls above that threshold value will include top m results from that tag in result. As evident this task can be computationally both heavy and inaccurate, so the hashing based strategy to overcome this problem.

Hash Coding Module

It will first convert input image into an instance aware representation using cross proposal fusion. Then it will be used to do either of the category-aware hashing or semantic hashing. Initially N x d matrix which was output of label probability calculation module is compressed into b dimensional matrix using fully-connected layer of neural net.

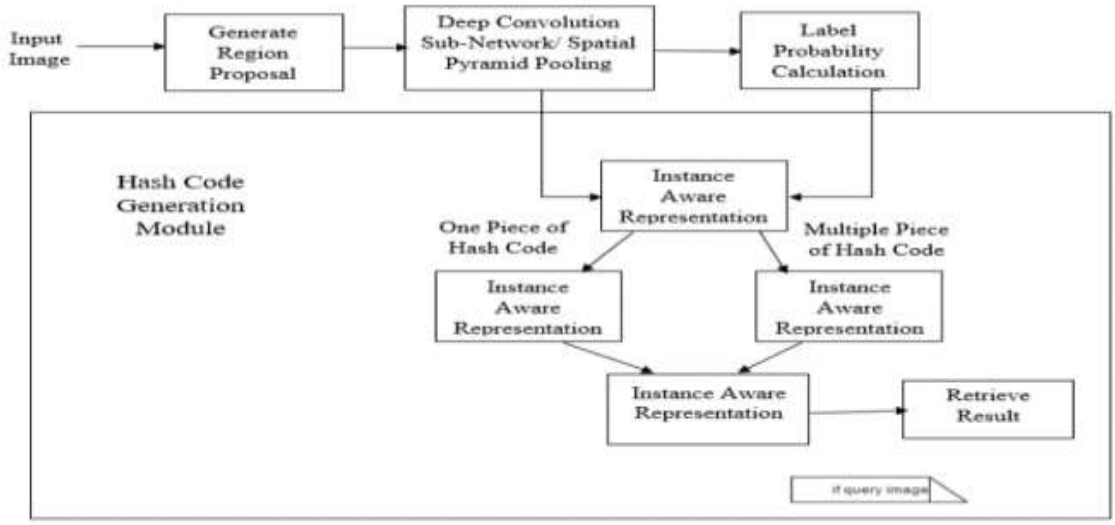

Figure-2 : Reetrieval using Hashing

Cross Proposal Fusion

N x c probability matrix and above N x b matrix is fused into a c x b matrix using Kronecker product such that each row corresponds to b-dimensional feature corresponding to each class category. This is called as c x b matrix as f,

$$f = \frac{1}{N}\sum_{i=1}^{N} P^i \bigotimes H^i$$

Hash Representation

For each image I it will compute c triplet loss functions corresponding to each class category. This triplet loss function is given as,

$$\begin{aligned}&\ell_{Triplet}(f^{(j)}(I), f^{(j)}(I^+), f^{(j)}(I^-))\\ &= max(0, 1 - \|f^{(j)}(I) - f^{(j)}(I^-)\|_2^2\\ &+ \|f^{(j)}(I) - f^{(j)}(I^+)\|_2^2))\end{aligned}$$

Where, I^+ is any image that belongs to the same category as image I. I^- is any image that doesn't belongs to the same category as image I. Then each of modified f^i will be converted into b-bit binary code given as, b(i) = sign(f (i)) where, sign(x) = 1 if x > 0 otherwise sign(x) = 0 Final category aware hash code for image I is of the form b(I) = (b (1) I, b (2) I, b (3) I, , b (c) I).

Category Aware Retrieval

First it create a hash table from images in database for retrieval. This hash table has c columns each corresponding to one of the c classes. For each image in retrieval dataset, it will first encode it into b-bit binary hash code for each of c class categories. Now this b-bit hash code for images j th class category is added to corresponding j th column in hash table if the respective value in the matrix output of label probability module is greater than certain threshold value (0.2 here). Similarly during retrieval, using test query image, c pieces b-bit codes will be generated and drop

ones which have probability values less than 0.2. Now the search will be conducted in corresponding columns of hash table to obtain a list of retrieved images.

Conclusion

The problem of multi label image retrieval can be solved effectively with the combined use of deep convolutional neural network for feature learning and either clustering or hashing for image retrieval. Even though clustering would not be a feasible option, hashing drastically reduces computation time required for retrieval. In future if relevance of multi labelled image result compared to a single labelled image can be measured, even better results can be achieved with this strategy.

REFERENCES

[1] Youngeum An, Junguks Baek AND Minyuk Chang Classification of Feature Set Using K-Means Clustering From Histogram Refinement method (Korea Electronics technology Institute, South Korea)-2008. Learning Research 3, 45, 9931022.

[2] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313(5786):504507, 2006

[3] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems (NIPS), volume 25, pages 11061114, 2012.

[4] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan. Supervised hashing for image retrieval via image representation learning. In Proc. AAAI Conference on Artifi- cial Intelligence (AAAI), pages 21562162, 2014.K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361

[5] J X. Zhang, S. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in Proc. Eur. Conf. Comput. Vis., 2014, pp. 346361.

[6] P. Krhenbhl and V. Koltun, Geodesic object proposals, in Proc. Eur. Conf. Comput. Vis., 2014, pp. 725739..