Image retrieval method based on metric learning for convolutional neural network

Jieyuan Wang, Ying Qian, Qingqing Ye, Biao Wang

School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract. At present, the research of content-based image retrieval (CBIR) focuses on learning effective feature for the representations of origin images and similarity measures. The retrieval accuracy and efficiency are crucial to a CBIR. With the rise of deep learning, convolutional network is applied in the domain of image retrieval and achieved remarkable results, but the image visual feature extraction of convolutional neural network exist high dimension problems, this problem makes the image retrieval and speed ineffective. This paper uses the metric learning for the image visual features extracted from the convolutional neural network, decreased the feature redundancy, improved the retrieval performance. The work in this paper is also a necessary part for further implementation of feature hashing to the approximate-nearest-neighbor (ANN) retrieval method.

1 Introduction
The crucial technology of the content-based image retrieval (CBIR) is using image feature extraction method to obtain the origin images, then using the feature of visual information as the target, using the similarity measurement between image features, returning the retrieval results from image database. Recent years, users using their personal terminal productions taking a massive photos and pictures. These images are stored and shared in the network, so that the accurate and fast retrieval of a massive-images-stored image retrieval system has become the core focus of the study.

Image feature extraction is widely used to represent the image content. The key challenge has been attributed to the “semantic-gap” which is well-known. Its issue exists that exists between low-level image pixels captured by image processing methods and high-level semantic concepts perceived by user. Typical image features are colour features (Colour Histogram, Colour Moments, Colour Sets, etc.), texture features (Local Binary Patterns, Co-occurrence Matrix, etc.), edge features (Canny and Sobel Operator Edge Detection, etc.), as well as SIFT, GIST, BoVW [1] features, these are the characteristics of artificial design. These features will Extract the original image information and dimensionality reduction, the purpose is to get a better representation of the original image, shorten the gap between the user's retrieval needs and the actual search results, reducing the semantic gap. With the development of deep neural network technology in the field of machine learning, convolutional neural network [2] has made important progress in the field of image recognition. Through the pre-trained convolutional neural network (CNN) to extract the feature of the image in the image database, the extracted feature is better than the traditional manual designed features [3]. In recent years, the CNN structure proposed by scholars in ILSVRC (ImageNet Large Scale Visual Recognition Challenge) has achieved remarkable results in the Alex-Net [4], making the research of convolutional neural network become a research focus. However, there is an obvious situation that the feature extracted from pre-trained CNN model has a high dimension. This situation brings the performance of the retrieval speed and accuracy not effective. The disaster of dimension need to be improved. The
extracted feature from image database becomes a feature set, this set could be regarded as a collection in a metric space [5]. Theoretically, this metric space could be improved to a lower dimension metric space, keeps the origin information the similarity measure faster to raise the speed of the retrieval. And, this work is also an important step for the implementation for hash coding to the approximate-nearest-neighbor (ANN) retrieval method. The method requires the original feature coded transform into a binary codes.

2 Convolutional Neural Network
Convolutional Neural Network (CNN) is not really a new method of machine learning, as early as the 1990s, LeCun et al. proposed a multi-layers artificial neural network architecture for classification of handwritten numbers, the architecture named LeNet-5\(^6\), which have a typical structure of CNN and did well performance on recognition of handwritten characters (Figure 1 from [6]). CNN is a discriminative deep architecture among various Deep Learning techniques. It belongs to the Deep Neural Networks (DNN). But limited by the computer performance at that time, the CNN couldn’t have much more hidden layers and people prefer the shallow machine learning architecture like Logical Regression (LR), Support Vector Machines (SVM), which have the same performance as CNN but have faster learning speed. Along with the scholar’s research for decades, and the raising processing speed of computer, CNN could have more hidden layer (deeper structure). CNN has found state-of-the-art performance on various tasks and competitions in computer vision and image recognition, such as ILSVRC (ImageNet Large Scale Visual Recognition Challenge).

![Figure 1: LeNet-5 \(^6\), A Convolutional Neural Network of architecture.](image)

2.1 Typical Structure of CNN
The typical structure of CNN developed from the concept of neocognitron, which was inspired by the structure of animal visual cortical cells. CNN including 3 kind of artificial neurons layer, convolutional layer, pooling layer and activation layer, they stack up with one on top of another. Different from BP (Backpropagation) neural network, the weights sharing and local receptive field are the features of convolutional layers (Figure 2) and pooling layer (Figure 3) doing sub-sampling the output of the convolutional layer, reducing the data rate.

![Figure 2: Processing demo of convolutional layer](image)
Figure 3: Processing demo of pooling layer

Activation layer is an important part to the network structure, this layer get the output of convolutional layer and output the activated feature to next pooling layer. It essentially provides a kind of non-linear transformation to solve the problem of the expression of linear models. Common activation functions are Sigmoid (Eq. 1), ReLu (Eq. 2), softmax(Eq. 3).

\[ f(x) = \frac{1}{1+e^{-x}} \]  

\[ f(x) = \max(x, 0) \]  

\[ f(x) = \log(1 + \exp(x)) \]

At the end of the network structure, the output of the abstracted features will get full connected to a vector and enter to a classifier to finish the network training supervised.

3 Metric Learning

Metric learning for image retrieval has been extensively studied in both machine learning and multimedia retrieval communities. In the domain of metric learning, metric learning methods can be used to transform the feature space into another metric space which has lower dimension, can retain the original information of the feature as far as possible, and even make the transformed new features more effective to meet specific measurement requirements. The metric learning methods can be divided into two way: linear transformation and non-linear transformation [7].

The methods can also be divided into supervised distance metric learning and unsupervised distance metric learning. For the supervised learning, algorithm uses the class labels or marks; for the unsupervised metric learning, algorithm usually through the transformation which is theoretical supported, or uses the measurement between origin feature and transformed feature to complete the learning. Here are some typical metric learning algorithms used in this paper.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) [8] is a non-parametric method in pattern recognition. A group of variables which may be related to each other (compute the Covariance Eq. 4) are transformed into a
set of linearly uncorrelated variables by means of orthogonal transformation, then the transformed set of variables is called the principal component, which can achieve the purpose of dimension reduction.

\[
\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}
\]

(4)

3.2 Simple Principal Component Analysis
Simple Principal Component Analysis (SPCA) [9] is a data oriented method. The algorithm does not explicitly calculate nor diagonalise the covariance matrix. Also SPCA does not require the tuning of learning parameters and convergence is obtained with very few iterations. SPCA is intended to be a fast approximator for the principal components.

3.3 Kernel Principal Component Analysis
Kernel Principal Component Analysis (KPCA) method requires only the solution of an eigenvalue problem. In fact, it essentially requires only standard linear algebra, and can handle a wide range of nonlinearities due to its ability to employ different kernels. The goal of KPCA is to map the input space into a feature space via nonlinear mapping and then to extract whitened principal components in that feature space such that their covariance structure is an identity matrix. That is, we can avoid performing the nonlinear mappings and computing inner products in the feature space by introducing a kernel function of form Eq. 5.

\[
k(x, y) = \langle \Phi(x), \Phi(y) \rangle
\]

(5)

3.4 Laplacian Eigenmaps
Laplacian Eigenmaps algorithm is a kind of efficient streamline learning algorithm. The basic idea is that points that are very close in a high dimensional space whose image should be close too in a low dimensional space. By using the weighted distance between two points as the loss function and using spectral properties of image Laplacian operator to derive the optimal low dimensional data set to keep some local representation.

3.5 Locality Preserving Projections
Locality preserving projections (LPP) is based on the construction of the sample space between the distance relationship, and this relationship is maintained in the projection. It preserves the local neighborhood structure of the sample in the space, that is to say, minimize the distance weighted square sum of the nearest neighbor samples in the low dimensional space. It can also be understood as to avoid the divergence of the sample set and keep the original structure.

3.6 Auto Encoder
Auto encoder neural network is a kind of unsupervised learning algorithm and it uses the back propagation algorithm and the target value is equal to the input value. Auto encoder is a kind of neural network that can be viewed as two consisting parts: an encoder function \(H=f(x)\) and a reconstruction decoder \(r=g(H)\). Traditionally, automatic encoders have been used for dimensionality reduction or feature learning.

4 Metric Learning for CNN

4.1 Pre-training of the CNN
In this paper, the structure of Convolutional Neural Network (Table 1, layers was abbreviated and numbered) has 5 convolution layers, 3 pooling layers and 3 full connection layers, and use softmax as the classifier to finish the learning. Noted that activation layer is set after every convolution layer and full connection layer. Two pooling layers is set after the first two convolution layers and one pooling layers is set between the last
convolution layers and the first full connection layer. This structure is referring the part of the architecture of VGG-Net [10] to train the Convolutional Neural Network for CBIR. Besides, we used normalization unit after the first two layers and one drop-out unit to prevent the overfitting. In the pre-training stage, the training of the CNN model based the object image database “Caltech101”, trained 200 epochs and select the finest epoch for experiments.

Table 1, CNN structure in this paper

| Layer     | Size     | Act     | Pooling |
|-----------|----------|---------|---------|
| conv1     | 11*11*64 | act1    | pooling1 |
|           | 4*4 stride | norm1  | (3*3)  |
|           |          |         | 2*2 stride |
| conv2     | 5*5*256  | act2    | pooling2 |
|           |          | norm2   | (3*3)  |
|           |          |         | 2*2 stride |
| conv3     | 3*3*256  | act3    |         |
| conv4     | 3*3*256  | act4    |         |
| conv5     | 3*3*256  | act5    | pooling5 |
|           |          |         | (3*3)  |
|           |          |         | 2*2 stride |
| fc6       | 6*6*4096 | act6    |         |
| fc7       | 1*1*4096 | act7 dropout7 |         |

The situation of model fitting in the training process is shown in Figure 4.

4.2 Dimension Reducing using Metric Learning methods

After pre-training of the CNN model, we get a fine-tuned model to extract the features for image representation. The extracted features come from the last full connection layer, which has 4096 dimension and need to be dimension reduced. In this paper, we used methods PCA, SPCA, KPCA, Laplacian Eigenmaps described in the previous section.

4.3 Evaluation of CBIR System

The performance of evaluation is MAP (Mean Average Precision), which is widely used to evaluate the global performance of a CBIR system. The results are shown in the following figure:

“Metric Learning” for experiments, and reducing features in 8, 16, 32, 64, 128 dimension, to see the performances of the new feature in different dimension and different methods.
Figure 6: The MAP of features under variety of dimensions.

The dotted line shows the baseline of original feature, and the other lines show the performance of dimension reduced new features. In the reduction of the dimension, the Matric Learning methods usually make the original feature information lost to some extent. But in the dimension of 128 and 64, the MAP of PCA and SPCA is above the origin features, this explains that there is some information redundancy in the original features.

5 Summary
Through the experiment, we noticed that the Matric Learning methods usually make the original feature information lost to some extent. But in the dimension of 128 and 64, the MAP of PCA and SPCA is above the origin features, this explains that there is some information redundancy in the original features. Other non-linear method in this paper shown a not-well result, possible reason is CNN feature is a multi-level abstracted feature, after further non-linear transformation would destroy the information of the original feature. We can use the linear Metric Learning methods to raise the performance of CNN feature such as PCA, SPCA.

Acknowledgements
This work was supported by Project Supported by the National Natural Science Foundation of China under Grants 61171060.

References
[1] C.Dance, J.Willamowski, L.Fan, C.Bray, G.Csurka. “Visual categorization with Bags of Keypoints”, ECCV International Workshop on Statistical Learning in Computer Vision, (2004).
[2] Wan J, Wang D, Hoi S.C.H., et al. “Deep Learning for Content-Based Image Retrieval: A Comprehensive Study”, ACM International Conference on Multimedia. ACM, pp. 157-166, (2014).
[3] Cun L, Boser B. “Handwritten digit recognition with a back-propagation network”, Advances in Neural Information Processing Systems, pp. 396-404, (1990).
[4] Alex K, Ilya S. “Imagenet classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems, pp. 1097-1105, (2012).
[5] Xing E P, Jordan M I, Russell S, et al. “Distance metric learning with application to clustering with side-information”, Advances in neural information processing systems, pp. 505-512, (2002).
[6] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner. “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, volume 86, pp. 2278-2324, (1998).
[7] Yang L, Jin R. “Distance metric learning: A comprehensive survey”, Michigan State University, (2006).
[8] Fielding, Alan. “Principal Component Analysis.” Business Research Methods, volume 23, pp. 41-61, (2010).
[9] Matthew Partridge, Rafael A. Calvo. “Fast Dimensionality Reduction and Simple PCA”, Intelligent Data Analysis, volume 2: pp. 203-214, (1998).
[10] Simonyan K, Zisserman A. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. Computer Science, (2014).