Mixture of Deep CNN-based Ensemble Model for Image Retrieval

Hsin-Kai Huang[1], Chien-Fang Chiu[1], Chien-Hao Kuo[1], Yu-Chi Wu[1], Narisa N.Y. Chu[2], Pao-Chi Chang[1]

[1]Department of Communication Engineering
National Central University, Zhungli, Taiwan
{hkhuang, cfchiu, chkuo, ycwu,
pcchang}@vaplab.ce.ncu.edu.tw

[2]CWLab International
Thousand Oaks, CA, USA
Narisa.chu@ieee.org

***Abstract*—This paper proposes an aggregate (or mixture) of ensemble models for image retrieval based on deep Convolutional Neural Networks (CNN). It utilizes two kinds of deep learning networks, AlexNet and Network In Network (NIN), to obtain image features, and to compute weighted average feature vectors for image retrieval. Based on experimental results, the aggregate ensemble architecture effectively enhances learning with higher accuracy than single CNN in image classification. When the proposed aggregate of deep CNN-based ensemble model is applied to CIFAR-10 and CIFAR-100 datasets, it is shown to achieve 0.867 and 0.526 mean average precision in image retrieval, respectively.**

Keywords—Content-based image retrieval; Ensemble learning, Deep learning; Neural networks; Convolutional neural networks

I. INTRODUCTION

Many challenges are encountered in Big Data accumulated by prevailing usage of tablets, smartphones, digital cameras, and various multimedia devices. Image dataset management and retrieval customization rank on the top of the challenge list [1]. Our study proposes an Aggregate Ensemble model based on deep CNN (AECNN), which sums up results from two distinguishable CNN ensemble architectures, to yield the best classification and precision for image retrieval among many existing approaches.

II. RELATED WORK

Past research on image retrieval authored by Kuo [2] used a single CNN to do image retrieval on CIFAR-10 and CIFAR-100 datasets. CNN is known for its contribution to image retrieval because it can

- learn abstract semantic concepts of each image,
- transfer concepts to formulate feature vectors, and
- measure three distinguishable distance metrics to correlate image similarities and assess the efficacy of image retrieval.

These capabilities have deemed CNN a rather reliable image retrieval method.

As for image classification, Ciresan [7] integrated several CNN to form a multi-column DCNN architecture for Deep-learning CNN. Following the DCNN architecture, the average of all predictions resulting from each independent CNN is used to derive the final image classification. This kind of modeling has shown to yield easier feature descriptions and better performance than any single CNN.

III. FEATURE LEARNING FRAMEWORK

The framework used in our method includes pre-processing, training and testing stages, as shown in Fig. 1. Sections below explain the block diagrams in detail.

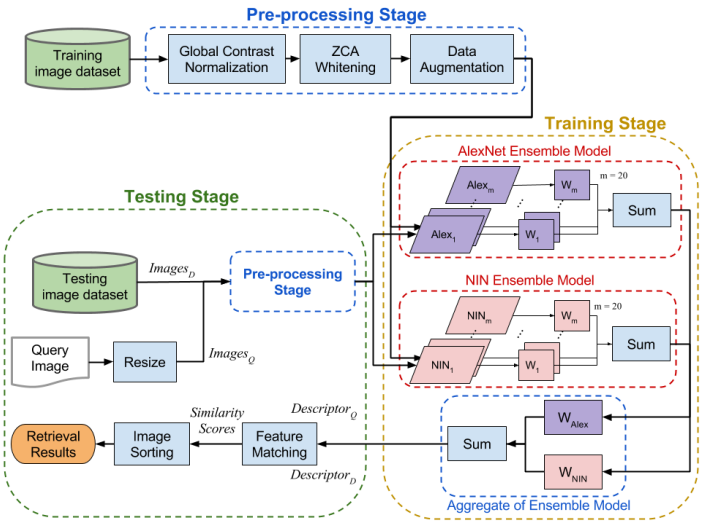

Fig. 1. Architecture of proposed AECNN for image retrieval

A. Pre-processing Stage

In the pre-processing stage, there are two phases. The first phase is based on Goodfellow[3], which uses Global Contrast Normalization and Zero-phase Component Analysis (ZCA) Whitening for the purpose of normalization and reduction of data redundancy. The second phase uses Data Augmentation which is an image processing method to increase the training dataset while prevent it from over-fitting. The AECNN also takes advantage of multi-dimensional color attributes, i.e.:

(1) three-dimensional attributes: RGB: red, green, and blue
(2) four-dimensional attributes: gray scale and histogram equalization, plus their respective mirror images.

B. Training Stage

The training stage is consisted of two phases. In the first phase, multiple, e.g., 20 in our case, AlexNet [4] and NIN [5] are placed for processing. The ensemble model follows (1):

$$S_i = \sum_{j=1}^{m} W_j F_j(i) \quad (1)$$

where S_i is the ith image feature vector in this ensemble learning, m is the total number of networks, j is the run number [1..20] of processing for AlexNet or NIN, W_j is the weighting factor of the jth network, and $F_j(i)$ is the ith image feature vector of the jth network. The weighting factor W_j is calculated based on the classification accuracy of the training dataset for the network j as in (2):

$$W_j = \frac{R(A_j)}{m} \quad (2)$$

where m is the total number of networks, A_j is the accuracy of jth network. $R(A_j)$ is assigned in a descending order starting from m based on the classification accuracy, and decreases by 1 accordingly. The proposed integrated model is generated by combining the ensemble models from AlexNet and NIN, respectively. The final feature vectors S_i^{mix} are calculated based on the weighted summation as (3).

$$S_i^{mix} = W_{Alex} S_i^{Alex} + W_{NIN} S_i^{NIN} \quad (3)$$

Empirically, the average training accuracy of the NIN ensemble model is as twice as that of the AlexNet. Hence, W_{Alex} is set to 0.5 for AlexNet and W_{NIN} is set to 1 for NIN.

IV. Experimental Results

The aggregate ensemble model was exercised on the specified image datasets for evaluation of image classification and retrieval using accuracy and Mean Average Precision (MAP) as measurement, respectively

A. Datasets

The CIFAR-10 and CIFAR-100 datasets [6] were used for evaluation of performance in this experiment. They are composed of labeled subsets of 80 million tiny images. The CIFAR-10 image dataset consists of 60,000 color images with size 32*32 in 10 classes. The training set has 5000 samples per class, and the testing set has 1000 samples per class.

The CIFAR-100 dataset, similar to CIFAR-10, has 100 classes with a training set consisting of 500 samples per class, and a testing set: 100 samples per class.

TABLE I. Image classification accuracy on CIFAR-10 and CIFAR-100 datasets

Methods	CIFAR-10	CIFAR-100
Kuo [2]	79.3%	44.0%
MCDNN [7]	89.79%	
Our CNN model		
Base : AlexNet	79.08%	44.01%
20 AlexNet ensemble	83.75%	50.0%
Base : NIN	86.46%	57.65%
20 NIN ensemble	90.42%	66.70%
AECNN	90.19%	67.28%

B. Image classification accuracy

Table I shows the image classification accuracy on CIFAR-10 and CIFAR-100 datasets. Both 20 NIN ensemble and AECNN yielded very high accuracy on CIFAR-10, over 90%. The AECNN model yields the highest accuracy among all tested methods in image classification on CIFAR-100 dataset as shown in Table I.

C. Image retrieval

MAP is used to evaluate image retrieval. Feature vectors of each image are extracted from the output of AlexNet and NIN. The similarity between images was calculated by measuring the cosine distance between two feature vectors. According to Table II, AECNN model shows the best MAP among all tested methods on CIFAR-100.

TABLE II. Image retrieval MAP on CIFAR-10 and CIFAR-100

Methods	CIFAR-10	CIFAR-100
Kuo [2]	0.707	0.244
Our CNN model		
Base : AlexNet	0.697	0.262
20 AlexNet ensemble	0.772	0.327
Base : NIN	0.817	0.396
20 NIN ensemble	0.874	0.523
AECNN	0.867	0.526

V. Conclusion

The Aggregate Ensemble CNN Model (AECNN) combines two distinguishable Convolutional Neural Networks (CNN) architectures for large runs (e.g., 20 iterations) to reach deep-learning effect. Two kinds of deep-learning network: AlexNet and NIN, are combined in the computation to obtain the weighted average of feature vectors. Based on the AECNN modeling runs, we observe the aggregate model can yield higher classification accuracy and better retrieval precision on images than the single CNN ensemble model. In our experiment with CIFAR-10 and CIFAR-100 datasets, image classification presents an accuracy of **90.19%** and **67.28%,** respectively; image retrieval yields a MAP of **0.867** and **0.526** MAP, respectively.

References

[1] J. Wan, D. Wang, C.H. Hoi, P.C. Wu, "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study," ACM International Conference on Multimedia, pp.157-166, 2014.

[2] C.H. Kuo, Y.H. Chou, and P.C. Chang, "Using Deep Convolutional Neural Networks for Image Retrieval," Visual Information Processing and Communication VI, February 2016

[3] I. Goodfellow, D.Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," International Conference on Machine Learning, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[5] M. Lin, Q. Chen, and S. Yan, "Network In Network," Computing Research Repository, 2013

[6] A. Krizhevsky, G. E. Hinton, "Learning Multiple Layers of Features from Tiny Images," Master thesis, Department of Computer Science, University of Toronto, 2009.

[7] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012.