Proceedings of ICSP2018

Combining CNN with Hand-Crafted Features for Image Classification

Zhou Tianyu, Miao Zhenjiang, Zhang Jianhu
School of Computer and Information technology, Beijing Jiaotong University, Beijing, China
Email: {13112064, zjmiao, 15120382}@bjtu.edu.cn

***Abstract*—Convolutional neural networks (CNN) have achieved outstanding performance in image recognition tasks, but many hand-crafted features still play important roles in some areas. Hand-crafted features are designed to describe image content from specific aspects, which may provide complementary information for CNN in image classification tasks. This paper explores feature fusion methods and proposes a novel framework for combining CNN with hand-crafted features. The framework has two main advantages. First, feature encoder can encode non-normalized features in CNN, which takes advantage of some good edge, texture and local features. Second, joint training strategy makes features fuse better in CNN. We validate that many hand-crafted features help to improve the performance of origin CNN. Experiments show our method outperforms the origin CaffeNet on Cifar10 dataset with 79.16% accuracy.**

Keywords—convolutional neural network; hand-crafted feature; feature fusion; image classification

I. INTRODUCTION

In the early years, hand-crafted image features such as Haar-like feature, HOG, LBP, and SIFT were widely used in most image recognition tasks. These features describe image from different aspects but have limited description capabilities. The past few years have witnessed a breakthrough of Convolutional Neural Networks in the field of computer vision. CNN automatically learn features from big data, and have strong description capability, which far exceed hand-crafted features. The recent development of CNN [1] make people believe that it is the future of computer vision. However, one problem is its low interpretability. It is difficult to describe what features are learned. Some researchers try to understand network interpretability [8], and study the relationship between hand-crafted features and CNN [7][8]. But little study was reported on combining CNN with hand-crafted features. This paper demonstrates that the combination of traditional features can achieve higher accuracy in image classification tasks. It is necessary to find out what combination method is proper for the CNN architecture. In the experiments, we connect some hand-crafted features to the end of CNN feature directly and find some features provide a little improvement. We also present a novel framework for combining image features in CNN architecture. It has a feature encoder structure to normalize features in different forms. And joint training strategy is used in this framework to improve feature fusion performance. Experiments show that our framework can improve classification accuracy of origin CNN. In another aspect, this work can help us understand the relationship between common hand-crafted features and CNN.

II. RELATED WORK

In the past decades, designing efficient and robust image features was the primary goal of image classification tasks. These hand-crafted features describe image from different perspectives. Color Histogram reflects the color distribution of the image. Canny edge, Gabor response and LBP descriptor show the edge and texture information of the image. HOG can describe the shape of objects which is very useful in human detection. SIFT and SURF descriptors provide local features of image, which reflect the details of the target. BoW builds these descriptors as local feature dictionary for object recognition. But hand-crafted features cannot describe image content in all aspects, so the recognition accuracy is limited in complicated tasks. In recent years CNN have become a hotspot of research in computer vision. AlexNet [2] achieved an accuracy of 83.6% top-5 accuracy in the 2012 ImageNet large-scale image recognition competition (ILSVRC2012). After that, CNN continued to evolve and achieved higher accuracy in the tasks of image classification. VGG [3], GooLeNet [4] and ResNet [5] expanded the depth and width of CNN, which made CNN features contain more detailed information and have stronger discrimination ability. As CNN automatically learn features, it is not easy to understand what features are learned. Controlling the composition of CNN features is also a difficult problem. Some researchers try to understand network interpretability and explain ability [6], and some other researchers study the relationship between hand-crafted features and CNN. This paper [7] compares SIFT and CNN features and find that they share common characteristics. This paper [8] demonstrates that low level features help to provide complementary information for CNN. LIFT [9] proposes a neural network model using CNN as feature detector and descriptor to replace SIFT.

Although some work provides ideas of the link between hand-crafted features and CNN, there is little study on combining CNN with hand-crafted features for classification tasks.

III. METHOD

In this chapter, we will introduce different feature fusion methods. In section A we analyze common hand-crafted features and choose some of the most representative ones, which can describe image from different aspects. We use the

978-1-5386-4673-1/18/$31.00 ©2018 IEEE

combinations of these features for image classification. In section B we connect some statistic features to the end of CNN feature directly. In section C, we present a novel framework to encode non-normalized features, and use joint training strategy to fuse features in CNN. In all the three parts, we use fully connected neural network layers as classifier.

A. *Combination of Hand-Crafted Features*

Commonly used features describe image content from several aspects such as color, shape, texture and local feature. We choose RGB, LAB and HSV histogram as color feature, HOG as shape feature, LBP histogram as texture feature, and BoW based on dense-SIFT as local feature. We give details of the extraction of these features as follows.

- Color histogram in RGB color space is calculated as a 51-d vector, 16 bins for each channel and 3 average values according to this paper [8]. Histograms in HSV and LAB space are calculated with the same standard. The final color histogram feature is a 51*3= 153-d vector.
- HOG feature is calculated as a 324-d vector. Divide each image to 4*4 regions and each block has 2*2 cells, so the number of block locations is 9. The histogram of gradients in each cell contains 9 bins corresponding to angles in the range of 0-180 degrees.
- LBP histogram is a 256-d vector, 16*16 pixels for each cell.
- BoW based on dense-SIFT is a 128-d vector, which has the dictionary of 300 words.

We connect these features directly but by different combinations in order to find which features perform better after fusion. After that, we connect the fusion feature to a fully connected neural network classifier to see the result. The combination architecture is shown in figure 1: (a) illustrates color hist feature in CNN architecture as one example of single feature. (b) is the combination of all 4 hand-crafted features in CNN architecture. The 4 hand-crafted features have 15 combinations, and we will show all these results in the experiment.

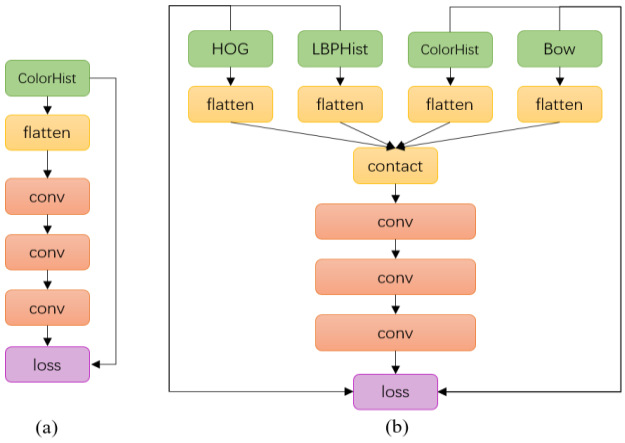

Fig. 1. Hand-crafted features combination architecture in CNN.

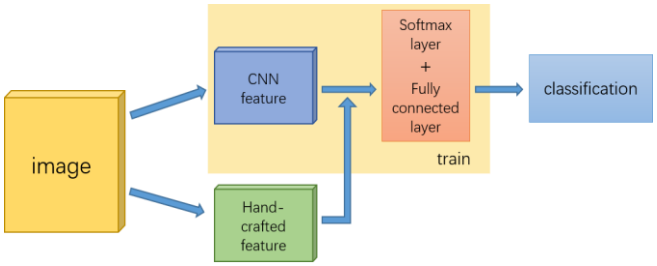

Fig. 2. The framework of direct combination of CNN feature and hand-crafted features.

B. *Direct Combination of CNN Feature and Hand-Crafted Features*

We use CaffeNet to extract CNN features for image classification. CaffeNet is a modified AlexNet model, which has 3 convolution layers, 3 pooling layers and 2 fully connected layers. The output of the last pooling layer is the CNN feature. We keep the main parts of CaffeNet architecture, and connect hand-craft features in section A to the end of CNN feature as the input of the fully connected layers. Figure 2 illustrates the architecture of direct combination of CNN feature and hand-crafted features. Hand-crafted features are extracted separately from CNN. These hand-craft features mainly reflect color, shape, texture and local features of the image, so we can see which part of information is complementary to CNN feature.

C. *Encoding and Joint Training Features in CNN*

Some original hand-crafted features are not normalized vectors but describe valuable image information like edge and texture feature. These features have different forms so we need to encode them before combination. Here we choose 4 representative features: Canny edge, LBP feature map, Gabor filter response map and dense-SIFT descriptor. For each feature we design a feature encoder in the CNN architecture, and train the total network to produce fusion feature.

- Canny edge feature is a 2-D map of original image which highlights the edge location. It has the same size with the original image, so it can be represented with a similar network as a feature encoder.
- LBP feature map is a texture feature. It is also a 2-D map so the encoder architecture is just the same as Canny edge feature.
- Gabor feature has 8 rotation directions in 5 scales. The size of its feature map differs from CNN feature map, so we use multiple 1*1 convolutional layers as encoder to reduce the dimensionality.
- Dense-SIFT descriptor is the most suitable local feature according to Cifar10 image size (32*32). We use 9 dense-SIFT descriptors to calculate a feature vector, and encode it with 1*1 convolutions.

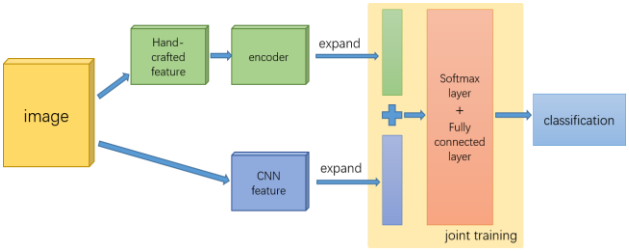

Fig. 3. The framework of combination of CNN feature and encoded hand-crafted features.

Joint training strategy can make better global optimization for two models. For image feature fusion, it is a suitable method to fuse different feature in CNN. The encoded hand-crafted features and CNN feature are expended and connected together. The two parts of features are trained in CNN using joint training strategy. Through these feature encoders we fuse CNN feature with the hand-crafted features in different forms. The architecture is shown in figure 3. To use the fusion feature, we lock the network parts before the connection of features, and train the rest parts only. The training details will be shown in next chapter.

IV. EXPERIMENTS AND DISCUSSION

We use Cifar10 as our experiment dataset. Cifar10 is a broadly used dataset which contains 60,000 color images in 10 different object classes. The size of each image is 32*32 pixels, with a single object in the center of the image. We choose Cifar10 because it has normalized images and balanced sample distribution, which is suitable for analyzing different kinds of image feature. This section describes the details of our experiments in three parts.

A. Combination of Hand-Crafted Features

We calculate color histogram in RGB, HSV and LAB color space as color feature, HOG as shape feature, LBP histogram as texture feature, and BoW based on dense-SIFT as local feature as representatives. The classifier we use is from CNN model which has 3 fully connected layers and a softmax layer in accordance with the following experiments. We train the classifier with stochastic gradient descent method, using adaptive learning rate. The mini-batch size is 100 for 550 epochs.

The 4 kind of hand-crafted features have 15 combinations. Table 1 shows the detailed accuracy of all combinations. According to the results, we see the best single feature is HOG which achieves an accuracy of 52.71%. It is easy to explain because shape feature has an advantage in representing a single object. The other features seem not that effective on this dataset. The combinations of them perform better than single feature, and the combination of all the 4 features achieves top accuracy of 61.18%. We call it "4H-Feature" in later parts. It demonstrates that different hand-crafted features can be used together. More aspects of image information make better classification results.

TABLE I. COMBINATION OF HAND-CRAFTED FEATURES

Number	*HOG*	*LBPHist*	*ColorHist*	*Bow*	*accuracy*
1	√				52.71
2		√			33.94
3			√		31.27
4				√	39.56
5	√	√			57.11
6	√		√		57.66
7	√			√	53.93
8		√	√		42.73
9		√		√	45.79
10			√	√	38.47
11	√	√	√		61.12
12	√	√		√	57.18
13	√		√	√	57.73
14		√	√	√	45.78
15	√	√	√	√	61.18

B. Direct Combination of CNN Feature and Hand-Crafted Features

We train CaffeNet to learn CNN feature. The optimization algorithm is stochastic gradient descent method. During the training process, we use adaptive learning rate. The base learning rate is 0.01 and decreases with the number of iterations grow. The mini-batch size is 100, number of epochs is 90, and the weight decay is 0.004. The momentum is 0.9 for the reliability of loss decreasing. After training, the output of the third pooling layer is the learned CNN feature. We directly connect hand-crafted features listed in the former section to the end of CNN feature. Then we set the learning rate of forward layers to 0, and train the fully connected layers and a softmax layer only. We combine the 4 hand-crafted features with CNN feature separately, and combine 4H-Feature with CNN feature in additional.

Figure 4 shows the results of 5 sets of experiment: HOG, LBP Hist, Color Hist, BoW based on dense-SIFT and all four features. Blue is the single hand-crafted feature classification accuracy. Yellow is the CaffeNet accuracy. Red is the direct combination of of CNN Feature and hand-crafted features accuracy.

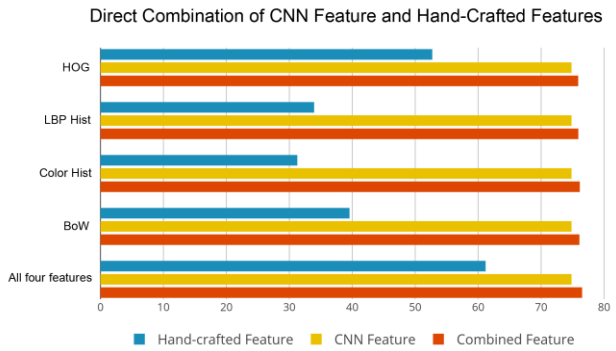

Fig. 4. Results of combination of CNN feature and encoded hand-crafted features.

The CaffeNet classification accuracy is 74.83%. From figure 4 we can see combining CNN feature with any hand-crafted feature can increase the accuracy, but not much. The combination with 4H-Feature performs best with the accuracy 76.51% with 1.68% increase. According to these results, we can learn that CNN feature is sufficient to describe image in many aspects, but still has room for improvement. Hand-crafted features are helpful to assist CNN feature. The effect of direct combination method is limited.

C. *Encoding and Joint Training Features in CNN*

Some edge, texture and local features are not normalized vectors. To use these features in CNN, we design encoders in the neural network for each feature. The main part of this network is the same as CaffeNet, but we add some convolutional layers as encoders and a fusion layer. We train the whole network together. The base learning rate is 0.01 and adaptively decreases. The mini-batch size is 100 for 90 epochs. The weight decay is 0.004 and the momentum is 0.9.

We use this framework to encode and combine Canny edge, LBP feature map, Gabor filter response and dense-SIFT descriptor separately and together.

Figure 5 shows the results of 5 sets of experiment: Canny edge, LBP feature map, Gabor filter response, dense-SIFT descriptor and all four features. Yellow is the CaffeNet classification accuracy. Green is the combination of CNN feature and encoded feature accuracy. Red is the 4H-Feature accuracy. Blue is the combination of CNN feature, encoded feature, and 4H-Feature accuracy. The accuracy of each set of results increases after fusing CNN feature with hand-crafted features. The top accuracy of combination of CNN and a single feature is Gabor feature, 78.14%. When we use all four features in this framework, the accuracy reaches to 78.47%.

We also use 4H-Feature in this framework, which means combining and joint training CNN feature, encoded feature and 4H-Feature together. The accuracy of each set is higher than the 4H-Feature result in section B. The top accuracy of all is 79.16% when we combine CNN feature, the Gabor filter response map and 4H-Feature together. The increase is 4.43% compared to CNN baseline. These results demonstrate that our framework is effective. The encoded features provide complementary information. Joint training strategy makes these features fuse better. But when we use all these four features, the accuracy is 78.24%. This result tells that the features still have redundancy.

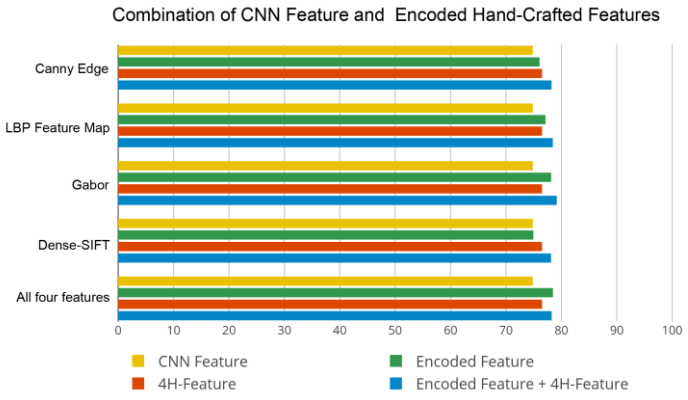

Fig. 5. Results of combination of CNN feature and encoded hand-crafted features

V. CONCLUSION

In this paper, we discussed the relationship of hand-crafted features and CNN. We compared some feature fusion methods, and showed the performances of them in the classification tasks. Through experiments, the direct combination of CNN feature and hand-crafted features made improvements but not outstanding. Although CNN feature is dominant, other hand-crafted features provide complementary information. We also presented a novel framework to fuse image features in CNN architecture. This framework can encode non-normalized hand-crafted features, and joint train all features together in CNN. From the results of experiments, our method shows good performance of feature fusion. But combining all these features together cannot get highest accuracy. The fusion of features is still redundant in this framework. In the future work, we are planning to find which part of features is redundant, and how to reduce the influence of it.

ACKNOWLEDGMENT

This work is supported by the NSFC 61672089 , 61273274 , 61572064, and National Key Technology R&D Program of China 2012BAH01F03.

REFERENCES

[1] Gu, Jiuxiang, et al. "Recent advances in convolutional neural networks." arXiv preprint arXiv:1512.07108 (2015).

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[4] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Bau, David, et al. "Network dissection: Quantifying interpretability of deep visual representations." arXiv preprint arXiv:1704.05796 (2017).

[7] Zheng, Liang, Yi Yang, and Qi Tian. "SIFT meets CNN: A decade survey of instance retrieval." IEEE transactions on pattern analysis and machine intelligence 40.5 (2018): 1224-1244.

[8] Lee, Gayoung, Yu-Wing Tai, and Junmo Kim. "Deep saliency with encoded low level distance map and high level features." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[9] Yi, Kwang Moo, et al. "Lift: Learned invariant feature transform." European Conference on Computer Vision. Springer, Cham, 2016.