Small Sample Size Training

Hongzhi Liu

May 11, 2018

1 Talking about SSF-CNN

As is known to all, Convolutional Neural Networks (CNN) have provided advanced results in several computer vision problems. And of course, it is a multilayer representation learning architecture which has received immense success in multiple applications such as object classification and image segmentation.

However, large training data is also a limiting requirement for applications with small sample size and many of these architectures easily overfit on small training samples. To address the challenge of small sample size, Professor Keshari presents a novel approach, termed as Structure and Strength Filtered CNN (SSF-CNN) [1]. Specifically, in case of newborn face recognition, remarkable improvement in accuracy is achieved with the proposed approach.

It is well known that matrix factorization or dictionary learning allows us to learn the dictionary that helps encoding the representative features. If one can represent CNN filters using dictionary, it can provide the "structure" but may not be well optimized for the classification task.

In the thesis, Professor Keshari proposes to use dictionary learning algorithm for learning the structure of the filters. The algorithm can be divided into two steps. One is learning hierarchical dictionary filters and utilize trained dictionary filters to initialize the CNN, and the other is training CNN with dictionary initialized filters. As shown in Figure 1, filters learned from the dictionary learning technique show more "structure" than traditional approaches, particularly with small training data.

Learning structure of filters is mentioned above. And the proposed concept of learning strength of the filter is illustrated in Figure 2. The team freeze
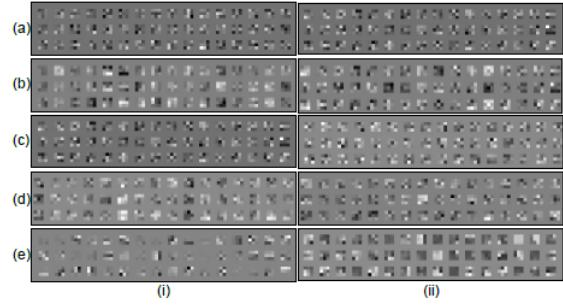

Figure 1: Filter visualization of the (i) 1st layer and (ii) 2nd layer of the ResNet architecture on CIFAR10 dataset. (a) Xavier [2] initialized filters at zero epoch, (b) Xavier initialized filters are trained on 1000 training samples, (c) MSRA [3]initialized filters at zero epoch, (d) MSRA initialized filters are trained on 1000 training samples, and (e) Dictionary initialized filters at zero epoch. For better visualization, only 16 filters are used from the 2nd layer.

the values of filters obtained from dictionary learning technique and update only the strength of the filter. As shown in Figure 2, this significantly reduces the number of learning parameters. For l^{th} layer, the strength parameter 't^l' is learned using stochastic gradient descent method; i.e. a scalar value is learned rather than learning the complete filter. The proposed process can be written as Equation. (1):

$$f(X, W, b, t) = X \times (t \odot W) + b \qquad (1)$$

where, $(t \odot W)$ represents element-wise multiplication. The pre-trained filters learned from dictionary learning or pre-trained model are selected and
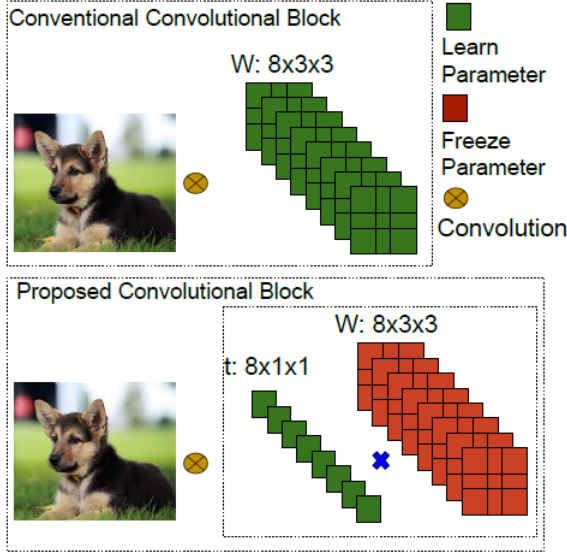
Figure 2: Illustrating the concept of learning the strength of a filter which significantly reduces the number of training parameters.

the only variable to be learned is t which can be learned using SGD. Since $|W| \gg |t|$, even small training data can be used to train the network. In literature, various regularization techniques have been utilized for better convergence. Existing regularization techniques such as dropconnect and ℓ_1 regularization can also be used while learning t.

2 Evaluation of SSF-CNN

The effectiveness of the proposed algorithm is evaluated on multiple databases with state-of-the-art CNN architectures. Since the proposed architecture is for small size trainingdata, the experiments are performed with varying training sizes on three databases that is MNIST [4], CIFAR10 [5], and NORB [6].

More specifically, as shown in Table 1, the experiments are performed with 14 data sizes, 100, 200,··· , 1000, 2000,··· , 5000. The proposed algorithm is also tested with the complete/standard training set. Furthermore, experiments are performed on an interest-

Table 1: Experimental protocols for MNIST, CIFAR-10 and NORB databases.

Databases	Small Training Data	Standard Training	Standard Testing
MNIST	100 : 100 : 1k; 1k : 1k : 5k	50k	10k
CIFAR-10	100 : 100 : 1k; 1k : 1k : 5k	40k	10k
NORB	100 : 100 : 1k; 1k : 1k : 5k	20k	24.3k

ing and small sample size problem of newborn face recognition. I can also learn that the proposed algorithm by Professor Keshari, in general yields higher performance compared to several existing algorithms.

References

[1] Rohit Keshari, Mayank Vatsa, Richa Singh, and Afzel Noore. Learning structure and strength of CNN filters for small sample size training. *CoRR*, abs/1803.11405, 2018.

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256, 2010.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[4] S Haykin and B Kosko. *GradientBased Learning Applied to Document Recognition*. Wiley-IEEE Press, 2009.

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[6] Yann Lecun, Fu Jie Huang, and L¨¦on Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, pages II–97–104 Vol.2, 2004.