Addressing Noisy Label Problem: Iterative Cross Majority Learning with CNNs

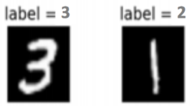
Honghao Qiu
honq@stanford.edu

Abstract

Data collected in real world often has incorrect labels, such noise in training data can lead to significant model performance degrade on test set. We propose a simple But effective method, Iterative Cross Majority Learning (ICML), to train multiple convolutional neural networks independently on noisy data, and update data labels based on majority vote across predicted outputs from the trained models. We then repeat the process in multiple stages iteratively, and use the last stage trained CNNs to perform ensemble learning for making test set prediction. Our experiments on MNIST shows that the proposed ICML method is able to achieve near state of the art result of 97+% accuracy on test set after trained on data with 70% noise.

Data

Data: we use MNIST dataset. There are 60,000 training images (resolution: 28*28*1), and 10,000 test images in total. We partition training data into 55,000 training set and 5,000 validation set, introduce noise to labels at different levels: 30%, 50%, 70% labels in training set are set to wrong labels randomly. We set 10,000 intact true label images aside for testing. *Example:*



Pre-processing: for each image, we normalize them to 0-1 scale, subtracted mean of training images, then divide by standard deviation of training images.

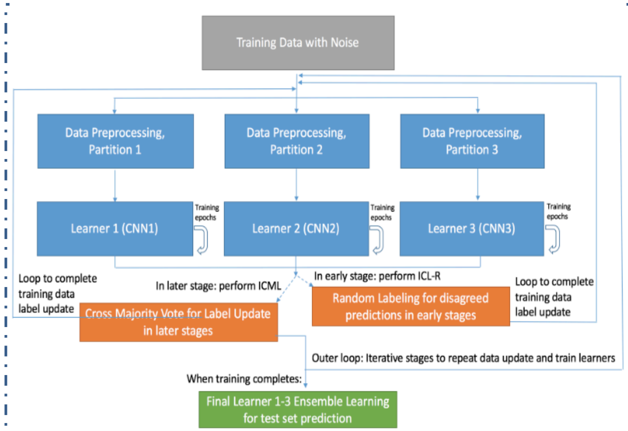
Future Work

1. Check effectiveness on CIFAR10: we only used MNIST dataset to check the effectiveness of our proposed approach, it is necessary to cross check effectiveness on more extensive and complicated image dataset such as CIFAR10;
2. Theoretical proof: beyond current analysis of the results, we would like to seek theoretical proof of the effectiveness of our proposed methods to build a more solid understanding on effectively training CNNs with noisy labels;
3. True label recovery: extend our work to true label recovery and auto-labeling by assigning random label first and then use ICML to correct labels iteratively.

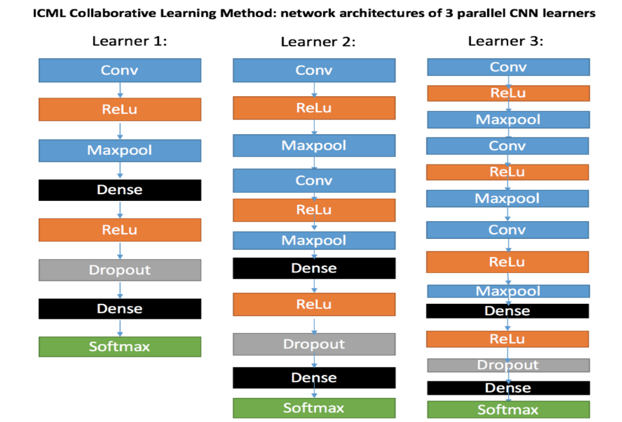
References

[1] T. Xiao. Learning from massive noisy labeled data for image classification, *CVPR, 2015*
[2] V. Mnih and G. E. Hinton. Learning to label aerial images from noisy data, *ICML, 2012*
[3] Y. Bodi, and S. McMains. Iterative cross learning on noisy labels, *WACV, 2018*
[4] I. Goodfellow and Yoshua Beingo, Generative adversarial nets, *NIPS, 2014*
[5] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, *http://yann.lecun.com/exdb/mnist, 1998*

➢ Our approach:



Iterative Cross Majority Learning (ICML) workflow



Network architecture of 3 CNNs trained in parallel

➢ Objective (Loss):

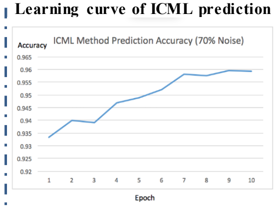
Extends cross entropy loss from binary classification: $loss = -(ylog(p) + (1-y)log(1-p))$

Into multi-class classification cross entropy loss: $loss = -\sum_{j=1}^{C} y_{o,j} log(p_{o,j})$

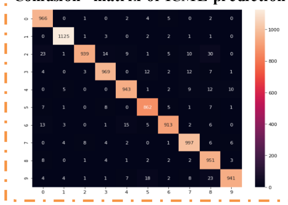
Categorical output with softmax: $y_k = \dfrac{\exp(\phi_k)}{\sum_j^C \exp(\phi_j)},$

Experiment Results

Learning curve of ICML prediction



Confusion matrix of ICML prediction



Experiments Summary: average results from 10 runs. ICML method performs label update based on majority vote, while iterative cross learning – random (ICL-R) method performs random update when CNNs outputs do not agree.

Noise level	0.3	0.5	0.7
Baseline	0.9771	0.9622	0.9324
Oracle	0.9875	0.9875	0.9875
ICML	0.9791	0.9695	0.9596
ICL-R	0.9833	0.9740	0.9657
ICML+ICLR hybrid*	**0.9864**	**0.9810**	**0.9709**
ICML+ICLR staged	0.9845	0.9773	0.9695

Experiments show that hybrid label update strategy combining ICML majority vote and ICL-R random labeling (when all CNNs outputs do not agree) achieves the best near state-of-the-art result

Discussions

1. This paper proposes to train multiple (>=3) CNNs in parallel and perform a hybrid update strategy with majority vote and random label update based on CNNs outputs for training data. Test time performance proves the effective of this method, which is able to achieve near state of the art 97% accuracy with noise level as high as 70% in training set.
2. Our proposed approach is a generalized, flexible, and performant method that can serve as a base method to plug in other noise label learning techniques.
3. It can also be used to automatically label data by assigning random label and training CNNs with this method on these data along with a portion of true label data provided to perform label update.