Collaborative Learning for Deep Neural Networks

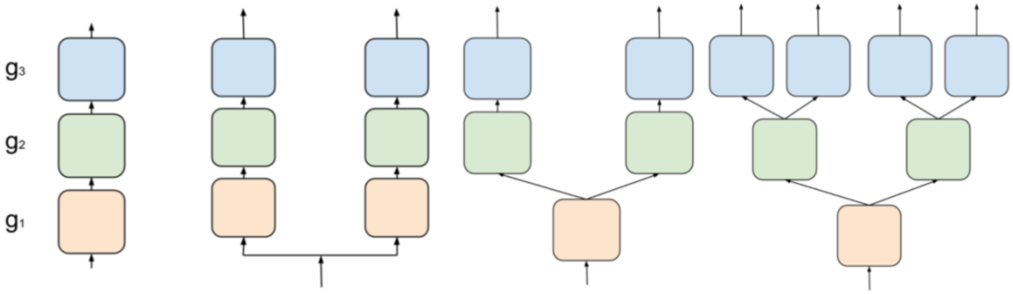
Overview

Collaborative learning is introduced where in multiple classifier heads of the same network are simultaneously trained on the same training data. Through this method, the paper aims to improve generalization and robustness to label noise, with no extra inference cost. Methods such as SGD cannot guarantee to converge to a global minimum, and the non-convex optimization problem must be taken into account. An ensemble of multiple instances of a neural network trained with random seeds would perform better than a single trained instance. However, training multiple models will be computationally expensive. Collaborative learning aims to achieve this better performance while keeping the model complexity same for inference. It trains several classifier heads of the same network simultaneously on the same training data. Multiple views on the same example have diversity of predictions, collaborative learning is by nature more robust to label noise than individual learning. Therefore, benefits are:

1. Consensus of multiple views on the same example increases generalization.
2. Intermediate level representation (ILR) with backpropagation rescaling reduces training computational complexity.

Three major parts to collaborative learning

1. Generation of training graph
   Add several new classifier heads into the original network graph (z = g(x; θ)) during training time.



(a) Target network (b) Multiple instances (c) Simple ILR sharing (d) Hierarchical ILR sharing

(a) is the target network to train, cascaded as: $g(x; \theta) = g_3(g_2(g_1(x; \theta_1); \theta_2); \theta_3)$
(b) each head is a new instance of the original network
(c) ILRs in the same low layers can be shared
(d) Multiple hierarchical ILRs of dependent nodes, similar to a binary tree
For inference, we just need to keep one head with its dependent nodes and discard the rest. Therefore, the inference graph is identical to the original graph.

2. Learning objective
   For each target head $z^{(h)}$, the softmax function is:

$$\sigma_i(\boldsymbol{z}^{(h)}; T) = \frac{\exp\left(z_i^{(h)}/T\right)}{\sum_{j=1}^{m} \exp\left(z_j^{(h)}/T\right)}$$

And the "soft" objective function is:

$$J_{soft}(\boldsymbol{q}^{(h)}, \boldsymbol{z}^{(h)}) = -\sum_{i=1}^{m} q_i^{(h)} \log(\sigma_i(\boldsymbol{z}^{(h)}; T))$$

Where $q^{(h)}$ for H heads is:

$$\boldsymbol{q}^{(h)} = \sigma\left(\frac{1}{H-1}\sum_{j \neq h} \boldsymbol{z}^{(j)}; T\right)$$

T is the temperature measure, *T* = 1 is normal softmax, increase in *T* softens the probability distribution.

3. Optimization
    a. Simultaneous SGD: Instead of the slow “alternatively update the parameters associated with each head one-by-one”, apply SGD and update all the head parameters at once by calculating total loss of the heads along with regularization.
    b. Backpropagation rescaling: Normalize the backpropagation flow in a subnet and keep that in the next subnet, and stabilize the flow for ILR sharing.
    c. Balance objectives by multiplying with T^2

Results and Conclusions

Datasets used were CIFAR-10 and CIFAR-100 using ResNets and DenseNets to obtain the ILR sharing from the split points. It was concluded that in a given training graph, the more classifier heads, the lower the generalization error. Along with this, ILR sharing reduces GPU memory consumption and training time. Simultaneous optimization also provides an accuracy and speed boost over alternative optimization. Backpropagation rescaling was proved necessary as while no scaling suffers from too large gradients in the shared layers, loss scaling results in a too small factor for updating the parameters of independent layers.

Similar experiments were performed for ResNet-50 on ImageNet dataset. As collaborative learning generates more classifier heads during training, there is some extra training cost as a tradeoff for ILR sharing which improves generalization error and speeds up training, reduces memory consumption while keeping inference complexity constant.

Overall, collaborative learning provides a powerful approach for deep neural networks to achieve better performance.