Few Shot Network Compression via Cross Distillation

Abhishek Kumar
Computer Science and Engineering
IIT PATNA,2nd year
Bihta, India

I. Introduction

As we know, Deep Neural Networks(DNNs) have achieved remarkable success in a wide range of applications. However, there are some areas where it lacks in efficiency. Some of those areas are large usage of computations resources, high storage consumptions etc. So, In order to obtain light-weighted DNNs, network compression techniques have been widely developed in recent years, including network pruning, quantization and knowledge distillation. Despite the success of previous efforts, a majority of them suffer from security and privacy issues. In order to obtain scalable network compression algorithms, a compromise between privacy and performance is to compress the network with few shot training instances, e.g., 1-shot for one training instances per class. Prevalent works along this line extend knowledge distillation by minimizing layerwise estimation errors between the teacher and student network. Nevertheless, a key challenge in few shot network compression is rarely investigated in previous efforts: as there are few shot training samples available, the student network tend to over-fit on the training set and consequently suffer from high estimation errors from the teacher network during inference. Moreover, the estimation errors could propagate and accumulate layer-wisely and finally deteriorate the student network. To deal with the above challenge, we proceed along with few shot network compression and propose cross distillation, a novel layer-wise knowledge distillation approach. Cross distillation can effectively reduce the layer-wisely accumulated errors in the few shot setting, leading to a more powerful and generalizable student network.

II. Methods and their description

A. Cross Distillation

Before going into this, here are some symbols and their meaning which are going to be used later on.
Symbol 's' denotes student network, 'l' denotes layer index,'t' denotes teacher network ,'h_s' denotes feature map of student network,h_t denotes the feature map of teacher network and 'L' denotes estimation error. Our goal is to obtain a compact student network s from the over-parameterized teacher network t . Given few shot training instances xn, yn from n=1 to n=N, we denote their correspond- ing l-th convolutional feature map of the teacher network t as h_t(l). Unlike standard knowledge distillation approaches, here we adopt layer-wise knowledge distillation which can take layer-wise supervision from the teacher network.When there are only few shot training instances, the student network s tends to suffer from high estimation errors on the test set as a result of over-fitting. Moreover, the errors propagate and en- large layer-wisely and finally lead to a large performance drop on s.
To address the above issue, we propose cross distillation, a novel layer-wise distillation method targeting at few shot network compression. Since the estimation errors are accumulated on the student network s and h are taken as the target during layer-wise distillation, we direct h to s in substitution of h_S to reduce the historically accumulated errors. In the forward pass of s, however, directing h_t to s results in inconsistency because s takes h_t from t in the training while it is expected to behave along during inference. Therefore, mini- mizing the regularized l(estimation error) could lead to a biasedly-optimized student net. In order to maintain the consistency during forward pass for s and simultaneously make the teacher aware of the accumulated errors on the student net, we can inverse the strategy by guiding h_s to t . Despite the teacher network now can provide error-aware supervised signal, such connection brings inconsistency on the teacher network. The errors in h_s is be enlarged during layer-wise propagation, leading to deviated supervision for s that deteriorates the distillation. Consequently, the correction loss 'l_c' and the imitation loss 'l_i' compensate each other, and it is necessary to find a proper balance between them. There is a theorem which shows that the gap of cross entropy between the student network s and teacher network t is upper bounded by l(estimation error), and therefore layer-wise minimization of the constrained optimization problem could decrease the gap of cross entropy and finally improve s.

III. Baseline

For structured pruning, the proposed methods are compared against a number of baselines: 1) L1-norm pruning (Li et al. 2016), a data-free approach; 2) Back- propagation (BP) based fine-tuning on L1-norm pruned models; 3) FitNet (Romero et al. 2014) and 4) FSKD (Li et al. 2018), both of which are knowledge distillation methods; 5) ThiNet (Luo, Wu, and Lin 2017) and 6) Channel Pruning (CP) (He, Zhang, and Sun 2017), both of which are layer-wise regression based channel pruning methods. For unstructured pruning, 1 is modified) to element-wise L1-norm based pruning (Zhu and Gupta 2017).

Besides, 4) FSKD, 5) ThiNet and 6) CP are removed since they are only applicable in channel pruning

IV. Pruning scheme

The structured pruning schemes are similar to those used in (Li et al. 2016; 2018). For the VGG- 16 network, we denote the three pruning schemes in (Li et al. 2018) in the ascending order of sparsity as VGG-A, VGG- B and VGG-C respectively. We further prune 50% channels layer-wisely and denote the resulting scheme as VGG-50For ResNet-34, we remove r% channels in the middle layer of the first three residual blocks with some sensitive layers skipped (e.g., layer 2, 8, 14, 16). The last residual block is kept untouched. The resulting structured pruning schemes are denoted as Res-r%. Besides, we further remove 50channels for the last block to reduce more FLOPs when r = 70%, denoted as Res-70%+. In terms of unstructured pruning, we follow a similar pattern in (Zhu and Gupta 2017) by removing r = 50%, 70%, 90%, 95% parameters for both the VGG network and ResNet, and each layer is treated equally.

V. Results

We evaluate structured pruning with VGG-16 on CIFAR-10 and ResNet-34 on ILSVRC-12. It can be observed that both Ours and Ours-S generally outperform the rest baselines on both networks, whereas Ours enjoys a larger advantage on VGG-16 while Ours-S is superior on ResNet-34. Meanwhile, as the training size decreases, cross distillation brings more advantages comparing to the rest baselines, indicating that the layer-wise regression can benefit more from cross distillation when the student network over-fits more seriously on fewer training samples. In the next, we fix the training size and change the pruning schemes. Again on both datasets our proposed cross distillation performs consistently better comparing to the rest approaches. Besides, the gain from cross distillation becomes larger as the sparsity of the student network increases (e.g., VGG-C and ResNet-70%+). We suspect that networks with sparser structures tend to suffer more from higher estimation errors, which poses more necessity for cross distillation to reduce the errors.

One potential issue troubles us is the generalization of cross distillation, since the training of Ours and Ours-S is somehow biased comparing to Ours-NC that directly minimizes the estimation error l. Since estimation errors l among feature maps and cross entropy l_c of logits directly reflect the closeness between t and s during inference, we compare both results among student nets obtained by Ours-NC, Ours and Ours-S respectively. We again take unstructured pruning with VGG-90on the test set of CIFAR-10, and the rest settings are kept unchanged. For ease of comparison, we similarly divide values of Ours and Ours-S by those obtained by Ours-NC. Ratios smaller than 1 indicate a more generalizable student net. In summary, cross distillation can indeed generalize well when t and s are properly mixed in the few shot setting

VI. Conclusion

In this paper, we present cross distillation, a novel knowledge distillation approach for learning compact student net.The accuracies are 83.44% and 83.32% respectively on VGG-16, and 84.93% and 86.63% respectively on ResNet-56. By reducing estimation errors between the student network and teacher network, cross distillation can bring a more powerful and generalizable student network. Extensive experiments on benchmark datasets demonstrate the superiority of our method against various competitive baselines