View Author Feedback

Paper ID	4552
Paper Title	Online Knowledge Distillation with Diverse Peers

AUTHOR FEEDBACK QUESTIONS

1. Rebuttal

We thank all reviewers for the helpful comments and response to each of the reviewers' concerns below.

----------reviewer #1----------

Q1: If it's after all an ensemble approach, should there be any error bound established? That is, theoretically, how do you expect each of the auxiliary peer to perform and how would that influence the final aggregated "soft target"?

Our response: Existing work may shed some light on theoretical analysis of our approach. It has been proved according to VC theory in "Improvement Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher" that it is beneficial for the student to learn from a less powerful assistant. Since each auxiliary peer distills from the soft target, which plays a similar role of teacher assistant in a teacher-absent manner, we may expect the auxiliary peer to perform better than the baseline according to the above work. The influence of each auxiliary peer to the soft target shall be analyzed with the ensemble learning theory in "learning with ensembles: how over-fitting can be useful" and "managing diversity in regression ensembles". We consider to provide a detailed analysis in the future work.

Q2: In this paper, (m-1) auxiliary networks need to be trained with the same architecture as the final deployment network. Wouldn't that involve much longer training time?

Our response: One merit of group-based methods is that it is easy to train the multiple student networks by parallelization as shown in "large scale distributed neural network training through online distillation". Even without parallelization, we found that when the number of branches is less than 4, our method needs less computations as well as training time than KD.

Q3: In Figure 3, it looks like the trend is still going down at 8 branches, then how did you just stop at 8 and have no further comparisons till some kind of convergence? Also, is it possible that CL-ILR would perform better than ONE with a larger branch number?

Our response: In fact, we have tested 9 branches but found no significant changes as those of 8 branches for all the methods. We would try larger number of branches and update the results in the final version.

Q4: Some recent papers of KD-based models need to be included and compared. It looks like the relative improvement of these methods over KD is fairly close to the relative improvement of OKDDip over KD?

Our response: Thanks for your suggestion. We will add them into related work. We obtained the results of the first mentioned approach called FSP with KD and OKKDip for CIFAR-10 as: 6.05, 6.08, 5.58, and similarly for CIFAR-100 as 26.48, 26.51, 25.63. The three approaches are implemented with the same experimental settings as the results in Table 6 for fair comparison. In fact, the proposed approach and those mentioned ones aim to improve the classic KD in orthogonal directions. The former emphasizes on more effective group-based distillation in a teacher-absent scenario, while the latter work with the classic one-teacher and one-student framework with new formulation of teacher-learned knowledge. It is an interesting future exploration to incorporate both types of work into a unified model. However, in the current work, since we use the logits to represent the knowledge to be distilled as the classic KD, comparing with KD seems to be more appropriate for a direct demonstration.

----------reviewer #2----------

Q1: There is a drop in diversity during training each time the learning rate reduced. Since the attention is involved during training across the peers, is there any explanation for this diversity drop?

Our response: We may explain this phenomenon as follows. The reduction of learning rates leads to an immediate scale-down of the parameter space, which directly causes large diversity drop. The attention mechanism only helps to increase diversity gradually after some iterations with the new learning rate.

Q2: However, there seems to be a relatively large gap between the ensemble prediction performance and the leader's performance. This seems to suggest a problem existing in the design of the second-stage. This gap is much smaller in other mechanisms, such as DML and CL-ILR. It would be clear if some discussion on this can be added.

Our response: Since OKDDip maintains a large diversity among base classifiers, applying ensemble brings a larger benefit than DML or CL-ILR, which converge with similar base classifiers. Currently we use simple average to compute t_m (refer to the lines above Eq.(8)) for the second level distillation, which gives a final performance that is already better than others. Nevertheless, we still appreciate your suggestion to seek for more sophisticated formulation of the second-stage distillation for a better performance.

----------reviewer #3----------

Q1: Since there is no teacher network available in the training stage, the distillation targets are generated by the network itself, which will keep changing during training thus very unstable. Group learning with a large diversity will make the training targets even more unstable. Is it equal to add some types of noise to distillation targets, which is just as a distillation regularization?

Our response: Although there is no teacher network available during training, each student is also guided by the ground-truth labels as the first term in the overall loss function, i.e., Eq. (8). Additionally, the two distillation terms would be multiplied by a weighting function as in the footnote below Eq. (8) to only allow substantial contribution of distillation in later iterations when the status is good enough. These two strategies help to enhance the stableness of training.

Q2: Are there any parameters to control the diversity for different training stages?

Our response: No, we haven’t set an explicit parameter to control the changes of diversity. It is learned automatically.