View Reviews

Paper ID	4552
Paper Title	Online Knowledge Distillation with Diverse Peers

Reviewer #1

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words.
The authors proposed an online knowledge distillation method with auxiliary peers trained in an ensemble manner. The diversity of the peers is considered, and the final model is deployed to a leader network. Attention-based mechanism is incorporated to also capture relative importance of peers.

2. [Relevance] Is this paper relevant to an AI audience?
Relevant to researchers in subareas only

3. [Significance] Are the results significant?
Moderately significant

4. [Novelty] Are the problems or approaches novel?
Somewhat novel or somewhat incremental

5. [Soundness] Is the paper technically sound?
Technically sound

6. [Evaluation] Are claims well-supported by theoretical analysis or experimental results?
Sufficient

7. [Clarity] Is the paper well-organized and clearly written?
Good

8. [Detailed Comments] Please elaborate on your assessments and provide constructive feedback.
The paper is generally easy to follow and understand. My major concerns are summarized as follows:

1) 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"?

2) It is claimed that the benefits of group-based methods include "eliminating the necessity of pre-training a large teacher model". However, 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?

3) 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?

4) Some recent papers of KD-based models need to be included and compared: "A Gift from Knowledge Distillation", "Coupled end-to-end transfer learning with generalized Fisher information" and "Variational Information Distillation for Knowledge Transfer". It looks like the relative improvement of these methods over KD is fairly close to the relative improvement of OKDDip over KD?

9. [QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.
Pleas see "Detailed Comments" above.

10. [OVERALL SCORE]

7 - Accept

15. Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons below.

Most of my concerns are sufficiently addressed in the rebuttal, and thus I improved my rating accordingly. The authors are highly encouraged to revise the paper according to all reviewers' suggestions to make the paper more comprehensive and substantial.

Reviewer #2

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words.

This work targets at preventing fast homogenization in teacher-free online knowledge distillation. Authors apply a self-attention mechanism to group-derived targets for each individual student, which increases the diversity across students' target distributions. The experimental results, compared with other state-of-the-art models, including DML, CL-ILR, and ONE, have shown nice performance improvement.

2. [Relevance] Is this paper relevant to an AI audience?

Likely to be of interest to a large proportion of the community

3. [Significance] Are the results significant?

Significant

4. [Novelty] Are the problems or approaches novel?

Novel

5. [Soundness] Is the paper technically sound?

Technically sound

6. [Evaluation] Are claims well-supported by theoretical analysis or experimental results?

Sufficient

7. [Clarity] Is the paper well-organized and clearly written?

Excellent

8. [Detailed Comments] Please elaborate on your assessments and provide constructive feedback.

The paper is well written and well organized. The author provides a clear intuition about mechanism design. Empirical results of inference accuracy are also promising, compare with the four chosen mechanisms: DML, CL-ILR, ONE, and Ind. Authors have also provided a nice demonstration of peer diversity, which shows consistency with their intuition.

9. [QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.

There are minor details might require further explanations or explorations:

1. In Figure.2 and Appendix Figure.1 showing the diversity evolutions, 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?

2. As demonstrated empirically, peer-diversity has indeed been increased, and the ensemble prediction (at least for experiments in the main paper) has also been improved. 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.

10. [OVERALL SCORE]

7 - Accept

Reviewer #3

Questions

1. [Summary] Please summarize the main claims/contributions of the paper in your own words.

This paper proposed a two-level distillation algorithm, which performs group learning with diverse peers. The first level works as a diversity maintained group distillation with several auxiliary peers while the second level distillation transfers the fused diverse group knowledge to the group leader modle which is the final output.

The experimental results show that keeping a relative higher peer diversity leading to effective knowledge transfer from the base level to the second level. The proposed method is validated on several data sets and the performance is better than the baseline approaches.

2. [Relevance] Is this paper relevant to an AI audience?

Relevant to researchers in subareas only

3. [Significance] Are the results significant?

Significant

4. [Novelty] Are the problems or approaches novel?

Novel

5. [Soundness] Is the paper technically sound?

Technically sound

6. [Evaluation] Are claims well-supported by theoretical analysis or experimental results?

Sufficient

7. [Clarity] Is the paper well-organized and clearly written?

Good

8. [Detailed Comments] Please elaborate on your assessments and provide constructive feedback.

The strengths of the proposed method
1) This paper is well written and organized. The proposed approach is clearly demonstrated.
2) The experimental results are impressive which shows the proposed approach outperforms the state-of-the-art online knowledge distillation approaches without additional training and inference cost.
3) Two-level distillation framework is successfully avoiding the degradation problem by maintenance the diversity of group learning with multiple auxiliary peers and one group leader.

The drawbacks of this paper
1) 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?
2) The diversity level of auxiliary peers should change during training. Initially, it is better to generate a relatively large diversity while in the stage near to converge, a small diversity is preferred. Are there any parameters to control the diversity for different training stages?

9. [QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.

Please see the drawbacks of the paper.

10. [OVERALL SCORE]

7 - Accept