View Reviews

Paper ID
11437

Paper Title
ProSelfLC: Progressive Self Label Correction for Training Robust Deep Neural Networks

Reviewer #1

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
The paper is an interesting survey of target modification techniques. Several entropy and divergence modifications have been analyzed in this survey. The paper has proposed a limitation analysis with respect to these information measures. The balancing effects between predicted and annotated labels depending on the cost function are convincing.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
1) Several entropy and divergence modifications have been analyzed.
2) The paper has proposed a limitation analysis with respect to these information measures.
3) The balancing formulas between predicted and annotated labels are convincing.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).
There aren't really any very original theoretical contribution, but rather a theoretically consistent and pedagogical dissertation.

4. [Overall rating] Paper rating (pre-rebuttal)
Weak accept

5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.
The paper is an interesting survey of target modification techniques in a lovely mathematical presentation.

6. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestion to make the submission stronger)
The paper is an interesting survey of target modification techniques. Several entropy and divergence modifications have been analyzed and the paper has proposed a limitation analysis with respect to these information measures. I do not believe the superiority of any strategy (and this even if it fuses several relevant strategy): everything has to be written in a relative way if one would like a safe language. Indeed, selecting an information measure and not another relates not only on the information structures, but also on the decision issue with respect to available data and labels. Seeking the best among several information measures has a also given cost, etc. So, maybe the conclusion of the paper can be more realistic. Indeed, the drawbacks of existing literature are point out, but what are the drawbacks of the proposed approach (ProSelfLC)? This should be addressed.

Reviewer #2

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.

The authors analyzed the issues in existing label modification approaches that aim to deal with training deep neural networks (DNNs) on data with label noise and proposed a progressive self label correction (ProSelfLC) method to address them. ProSelfLC can automatically adjust the degree of trust on model prediction again the provided ground truth (labels) as training proceeds by considering both the length of training and the confidence (measured via entropy) of model prediction. Experiments on synthetic and real data sets containing label noise demonstrate the superiority of the proposed method over several existing label modification approaches.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
This work analyzed the intrinsic relations between output regularization (OR) and label correction (LC) methods and proposed a trust adaptation strategy based on both the length of training and the entropy of the model prediction result to address the issue of label noise in DNN training.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).
Existing works described (and compared) under the categories of OR and LC are conventional ones. It is unclear whether some works have studied trust adaption instead of using the fixed trust (via epsilon). If so, these works (more relevant to the proposed method) should be elaborated and experimentally compared. Further, in Section 2, self knowledge distillation methods were mentioned. However, none of them have been included in experiments. The current description of existing works (especially those used in experiments) is less self-contained.

Similarity structure in labels is frequently mentioned but less understandable because it was described in the too abstract way.

Table 2 is very hard to understand and there lacks enough text description on it. Also, the meanings of horizontal and vertical axes of Figures 2 and 3 were not clearly described.

To verify the effectiveness of the proposed trust adaptation strategy (Eq. 7), it is more convincing to compare the performance of merely using the global trust or the local trust with using both. Also, how to search/specify optimal parameter values in the proposed method should be more clearly described. Further, it is necessary to have multiple runs and report the mean and standard deviation of the results.

4. [Overall rating] Paper rating (pre-rebuttal)
Borderline

5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.
The authors proposed a new way of label correction to handle training DNNs in the presence of label noise. However, descriptions and experiments need to be improved according to the above comments.

6. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestion to make the submission stronger)
The English presentation should be improved to increase readability.

Reviewer #3

Questions

1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.
This paper compares methods for adjusting the training targets for classification algorithms with the goal to reduce over-fitting and improve robustness to noisy training labels. Although the approaches are general, the paper focuses on deep neural networks for image classification. The paper begins by proposing a unified mathematical framework for several existing target modification approaches, e.g. label smoothing, confidence

penalties, knowledge distillation, pseudo-labeling. The key difference between methods is how to weigh existing labels over pseudo-labels (model predictions) when creating targets. The paper proposes a new method that places a high weight on existing labels early in training. The weight placed on model predictions (pseudo-labels) is then increased globally as training progresses, and on a per-example basis in proportion to the model's confidence for that example. The paper compares the proposed method against several baselines and shows improved accuracy for both clean and noisy labels on CIFAR-100 and ImageNet.

2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.

* The paper is topical and relevant since it is becoming increasingly important to deal with noisy labels as datasets become larger and labeling is becoming more automated or derived from noisy sources such as the web.

* The comparison/unification of existing methods is useful and clear.

* The proposed ProSelfLC method is simple and makes intuitive sense, which will help analysis and adoption.

* The experiments provide convincing evidence that the method improves over baselines in terms of model accuracy.

3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).

* Despite the gradual increase of the weight of pseudo-labels, the approach may suffer from a positive-feedback amplification of early errors, causing the final model to make highly confident but wrong predictions for some examples. Figure 3 (e) shows some evidence for this: The final entropy for the noisy subset is smallest for ProSelfLC, which means that the model is highly confident on examples that were incorrectly labeled during training. On average, the benefits of ProSelfLC seem to outweigh this issue. However, it is likely that some of these highly confident predictions are wrong, which is a problem for safety-critical applications that require robust classifiers. It would be good to also report model calibration (e.g. ECE, see Guo 2017, https://arxiv.org/abs/1706.04599) and plot reliability diagrams (see Guo Figure 1 bottom). I recommend 100 bins for ECE and reliability diagrams when sufficient test examples are available (as is the case for ImageNet and CIFAR-100).

* The approach is largely heuristic: Increasing model weight as training progresses makes sense, but the approach does not theoretically motivate the proposed schedule.

* The method introduces additional hyperparameters that need to be tuned.

4. [Overall rating] Paper rating (pre-rebuttal)

Weak accept

5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.

Overall, I think this is an interesting, relevant and useful contribution. My main concern is about the risk for amplification of early errors, which is not sufficiently addressed. I would increase the rating if a convincing analysis of model calibration and errors is provided.

6. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestion to make the submission stronger)

* Add analysis of model calibration, specifically the confidence of incorrect predictions.

* Perhaps try to use the number of abbreviations or spell them out more often, since it can be hard to follow them.

10. [Final rating] Paper rating (post-rebuttal)

Strong accept

11. [Justification of final rating] Explanation of final rating. Describe the rationale for your final rating, including notes based on the rebuttal, discussion, and other reviews.

The rebuttal cleared up one of my main concerns, so I am increasing my rating accordingly.