View Reviews

Paper ID
2786

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

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
In this paper the authors introduce a self correcting label adjusting method in the context of supervised image classification. The authors name the new method "ProSelfLC" which adaptively follows a self correcting label scheme that redefines the training example labels as a linear interpolation between the model's current prediction for the example and the traditional one hot label target for the example. This type of self label correction has been studied before but previous approaches did not use a dynamic parameter to vary the interpolation during the training. In addiction to introducing ProSelfLC, the authors summarize other label-modification approaches in the literature and give a nice summary of connections between such methods (not new but nice). ProSelfLC improves classification accuracy on both ImageNet and CIFAR-10

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The paper describes a dynamic self label correction method that anneals the targets from the model as an interpolation between the current prediction and the one hot vector. Based on my knowledge and the description in the paper the authors are the only ones to modify the annealing parameter epsilon as training progresses.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
The presented result is a simple extension of previous label correcting methods (seems very straight forward extension) and the experimental results are not compelling (+0.5% increase compared to other methods). Namely the main novelty of ProSelfLC is applying an adaptive epsilon that changes during training.

The writing of the paper needs additional work. The paper is poorly organized, scattered with fragmented sentences and has a lot of grammatical and syntax issues to the point that it makes it difficult to follow at times. The "theory" of the paper is mostly re-iterating previous results and is not rigorous, making largely unbacked claims multiple times mentioning (e.g. mentioning "sub-optimality" without rigorous explanation).

Also the model for CIFAR-100 is not specified in the experiment paragraph.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
See Weaknesses, some of the claims in the paper are largely unbacked by valid explanations. They are not the main claims of the papers yet they are still distracting. Other than the main idea of ProSelfLC seems correct to me, based on my review

5. Clarity: Is the paper well written?
The paper writings need to be revised, this includes the organization of the sections and fixing grammar issues throughout the paper.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes, the authors connect their work clearly with relation to other works (the mention of two groups of label modifications) and also give a nice detailed summary of other works in the literature.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
3: A clear reject.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Only partially, more discussion is needed.

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This work studies an interesting problem, i.e., considering the label noise exists in a dataset, can the classification probabilities help correct the label noise? To this end, the authors formulate the label correction as a learning problem that is a function of learning time step and the classification probabilities. The proposed method is evaluated within three experimental settings, that is, with no label noise, synthetic label noise, and real-world label noise. The experimental results show the improvement produced by the proposed method to a certain degree.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
1. The research problem studied in this work is interesting and useful.

2. This work evaluates the proposed method with synthetic label noise and real-world label noise.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
1. This work implies that the models trained with the proposed method can yield robust predictions against the label noise, whereas the models trained with the cross entropy cannot. However, this point is not well justified. The cross entropy is widely-used in deep learning methods and is also an important part of the proposed method (Equation (10)). With the hypothesis that the dataset exhibits some label noise, it would be better to have a theoretical analysis that demonstrates how the proposed method and cross entropy react differently to the label noise.

2. There is no direct empirical evidence to support the claim. This work claim the proposed method is accredited by long learning time and low entropy. However, how the learning time steps and entropy (i.e., H(p) rather than H(p)/H(u)) correlates to the classification performance is still unclear to me.

3. The key question that motivates this work, i.e., "human annotations and predicted label distributions, which should we trust more?", may be vague and pointless without restrictions. Let's assume "we can trust predicted label distributions more". How many training samples are required so that this statement holds? Will the statement be always true as the number of training samples increases? In this light, the raised question is not thoughtful.

4. The objective function (10) may not be well designed. Global trust score g(t) and local trust score l(p) may not be intuitive. The global trust score only depends the learning time step. However, the convergence rate would vary from domain to domain and from task to task. It could depend on other factors, e.g., the state of the loss minimization. Similarly, l(p) is not well defined. Here is a counterexample. Given two different images, is it possible that a model accidentally yields the same classification probabilities?

5. The experiments with different types of noises can be improved by adding one more model, e.g., mobilenet. As the classification probabilities p could vary between different models, in the experiments in Table 4 and 5, only one model is used for evaluation and this set-up cannot demonstrate that the proposed method generalize to other models.

6. The experiments in standard image classification doesn't make sense to me. As we know the labels in CIFAR and ImageNet are 100% accurate (i.e., without noises) and the proposed method is designed to correct noisy labels, why and how does the proposed method improve the classification performance? The improvement shown in Table 3 looks marginal to me. It seems to suggest that the proposed method may leverage the randomness of the stochastic process in learning.

7. This work misses two related works [1,2].

References:
[1] Yan, Yan, and Yuhong Guo. "Partial Label Learning with Batch Label Correction." AAAI. 2020.
[2] Yuan, Li, et al. "Revisit knowledge distillation: a teacher-free framework." arXiv preprint arXiv:1909.11723 (2019).

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
As discussed in point 2 in the weaknesses section, the claim is based on the question which is not well-defined. Therefore, I think it is not fully correct.

I don't see any noticeable problems in the method, but the proposed Sobolev gradient seems not optimal as the proposed re-weighted L^2 gradient in terms of cost-performance tradeoff.

As discussed in point 6 in the weaknesses section, the experiments of image classification make less sense to me.

5. Clarity: Is the paper well written?
I think this paper is easy to follow, but it can be further improved.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes, I think the contributions of this work are clearly discussed by comparing to the cited related works.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
5: Marginally below the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Only partially, more discussion is needed.

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper presents a method to tune the epsilon parameter label correction (LC) [7-9]. The proposed algorithm is evaluated on Cifar100, ImageNet under synthetic label noise distributions, and Clothing 1M using the existing noise in the dataset.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
+ Good organization of related work
+ Good evaluation

The prior work is quite well organized, and put into a common structure that help understand and relate the proposed method to existing work.

The evaluation is quite extensive spanning three datasets and multiple different settings.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
- Method seems arbitrary
- Only modest improvements.
- Broader impacts weak

While the problem is well motivated, the solution (sec 4) seems arbitrary. The authors propose one closed form expression for epsilon, without properly relating it to what it does to the learning objective, or even ablating it in the experiments. The work would be much stronger if the paper could :
* motivate the presented method from some intuitive change in the objective, or similar and
* experimentally show that the presented choice in weighting is in-deep at least locally optimal (through an ablation study comparing it to various minor changes to the expression. E.g. what is the impact of g(t), what is the impact of l(p), ...)

Compared to some baselines (e.g. LS) the improvement of ProSelfLC seem modest at best, especially given the additional complexity introduced.

Finally, the broader impacts statement does not actually talk about broader impacts. I'd highly encourage the authors to think about what the broader impacts of their work are. How does the presented work aim to change to world, how does a 1 pt gain in noisy ImageNet affect the broader research community and beyond?

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
yes

5. Clarity: Is the paper well written?
yes

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
yes

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
4: An okay submission, but not good enough; a reject.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
No

Reviewer #4

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This paper systematically analyses different types of label smoothing, both mathematically and empirically. It further proposes a new variant of label correction. As is commonly done, the target label is a weighted combination of the one-hot ground truth label and a predicted label. Thereby the novelty lies in that the relative weighting is not global nor fixed. Instead, it is dependent on the training progression and the entropy of individual examples. Through empirical evaluation on CIFAR-100, ILRVC2012 and Clothing 1M the papers shows the benefits of the proposed method, where it outperforms all compared methods.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

This is a strong paper which has value as a survey as well as a novel technique. It
- Gives an interesting mathematical comparison and survey of standard label smoothing techniques.
- Proposes a novel per-example adaptive smoothing technique.
- Achieves good empirical results

3. Weaknesses: Explain the limitations of this work along the same axes as above.
- The related work could be more through and balanced. In particular the popular technique of knowledge distillation is only discussed in the introduction. But a discussion on the effects of using the models own predictions or the predictions from another model as a target would make sense and is not explored in this paper.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Yes, as far as I understand.

5. Clarity: Is the paper well written?
Yes, it is clearly written. But there are some incorrect sentences, e.g. L. 128 and L. 200. Please have your submission proof-read for English style and grammar issues.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes, the systematic discussion in the intro and method section is well done. The related work section could be improved, maybe by moving it to the supplementary material, giving the necessary space to give a more complete discussion.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
Some things remain unclear to me after reading.
- CCE is undefined. I assume it's class cross entropy?
- Figure 3: "We remark at training, a learner is given whether an example is clean or not." How is that information used in your method?
- A simple way to reduce the proposed loss in Eq. (10) is to always produce confident predictions. If it wasn't for the global term that keeps epsilon small at the beginning of training, the target labels could become p, hence the model would be encouraged to simply not change its predictions. I assume that is also why the entropy is always low for ProSelfLC (Fig. 3). Can you please comment on that? Have you considered adding a loss that penalized large epsilons, to avoid such a degenerate solution?

9. Please provide an "overall score" for this submission.
7: A good submission; accept.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
No

/