View Reviews

Paper ID

3003

Paper Title

Improving MAE against CCE under Label Noise

Reviewer #1

Questions

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

The authors propose an improvement over training with standard Mean Absolute Error (MAE) in the case where labels on the training set are noisy.
MAE is known to be more robust to label noise than the Categorical Cross Entropy (CCE), but it generally leads to worse performance.
The authors provide a theoretical explanation of this phenomenon based on the gradient norm of the logits (sample weights) under both loss functions.
The proposed modification IMAE is then a re-weighting that is performed on the logit gradients during back-propagation.
The superiority of IMAE over CCE is demonstrated in experiments on synthetic label noise on CIFAR-10 and also on the MARS dataset for video-based person re-identification under realistic label noise.

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

The theoretical analysis of CCE and MAE is thorough and provides an explanation of the tendency of CCE to overfit to incorrect labels and the underfitting of MAE to the correct labels.
The proposed modification IMAE is quite simple and should be considerably more efficient than other methods that deal with label noise.
The experiments show a significant improvement over CCE in the case of noisy labels which validates the approach. I also appreciate the experiment on MARS with realistic label noise.

3. What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak.

I found the presentation of the method not so easily accessible. In Figure 1, for example, the exact meaning of sample weight (logit gradient magnitude wrt loss) and sample probability (p(y|x) as predicted by the network) were only really clear after reading Section 3. The same can be said

The experiments lack comparison to prior work on training with label noise.
I think it would be important to see a comparison on the same data and task to state-of-the-art methods (e.g., [1, 2]). A combination with [1] for example could also be interesting.

[1] mixup: Beyond Empirical Risk Minimization
[2] Deep Bilevel Learning

4. Paper rating (pre-rebuttal)

Weak accept

5. Justification of rating. What are the most important factors for your overall recommendation?

The proposed method is simple, backed up by theory and performs well in experiments with noisy labels.
However, it lacks comparisons to other works that deal with label noise.
I hope that the authors can improve the presentation and writing a bit, so as to make the paper more accessible.

10. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal)

Weak accept

11. Final justification (post-rebuttal)

After reading the other reviews and author response I would still tend accept the paper although I would not be upset if the paper were rejected.
I agree with the criticism raised by R3 that clean validation sets are required. A comparison to SotA on CIFAR-10 is sadly not provided (I expect it to be quite a bit worse than SotA there).
I appreciate the comparison on Clothing-1M provided by the authors however. The results there suggest that under realistic label noise the method actually works well when compared to SotA methods.

Reviewer #2

Questions

1. Summary. In 3-5 sentences, describe the key ideas, experiments, and their significance.
This paper addresses the noise labeling problem. Based on the mean absolute error (MAE), they propose an Improved MAE (IMAE) method to resist the noise from the noisy labels. Authors carefully investigate and analyze the difference between MAE and CCE, and introduce their IMAE very naturally. The experiments show that their algorithm is better than CCE and MAE.

2. What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
They provide a cost function to help the neural network to resist the noise. Their paper is also well written and organized.

3. What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak.
I have several concerns about this paper.
1. Authors only provide results by using resnet as backbone neural network, which is not sufficient.
2. The idea of this paper is quite close to "training deep neural-networks using a noise adaptation layer", they both intend to change the weight of each sample before sending to softmax, definitely they do in different ways. It decreases the novelty of this paper.
3. My biggest concern is whether their method, in fact, is just a majority voting. When they designed their experiment on Cifar by adding uniform noise, even up to 80%, the correct portion is still the majority, since the 80% are relocated to other 9 classes evenly. What about adding more class dependent noise?
4. They hardly use other methods to compare with their method, except in person-reid they used one method.

4. Paper rating (pre-rebuttal)
Weak reject

5. Justification of rating. What are the most important factors for your overall recommendation?
There are several limitations on the design of their experiments, I also would like to see more comparison with other methods.

10. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal)
Weak reject

11. Final justification (post-rebuttal)
My concerns about the experiment part still exist. I would like authors to add more experiments and comparison in their next submission.

Reviewer #3

Questions

1. Summary. In 3-5 sentences, describe the key ideas, experiments, and their significance.
This paper studies the robustness of different loss functions to noise. I have to mention that the robustness is for general noise, including instance noise, rather than specifically for label noise. The robustness is analyzed by comparing the magnitude of gradients w.r.t. the logit vector or learned feature rather than the class posterior probability as previous work [31] did.

2. What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.
The authors presented sufficient details making the paper easy to read.

3. What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak.
The paper needs proofreading. Some claims are out of logic. For example, (1) the authors mentioned "computer vision, speech recognition, and reinforcement learning" as different tasks of deep learning. They are not parallel: the first two are of applications while the last one is about methodology. (2) The authors stated that "CCE is sensitive to label noise while MAE owns robustness under label noise" and then asked, "why does MAE work much worse than CCE although it is noise -robust?" Those two claims conflict with each other. One claim must be wrong. (3) "Let q(j|X_i) be the ground-truth probability of x_i belonging to class j, i.e., q(j|x_i)=1 if j=y_i, q(j|x_i)=0 otherwise". This is a wrong claim. If j=y_i, q(j|x_i) can be any value larger than 0.5 if the label is assigned by a Bayes classifier.

The study of the loss function robustness from the gradient perspective is not new.

4. Paper rating (pre-rebuttal)
Weak reject

5. Justification of rating. What are the most important factors for your overall recommendation?
The study of the loss function robustness from the gradient perspective is not new. There are papers comparing the magnitude of the gradient w.r.t. to input feature to compare the robustness of loss functions. For example, Truncated Cauchy Non-Negative Matrix Factorization. Although the authors study the deep learning based case, the contribution is marginal.

The robustness is not specific for label noise. Although the experiments are conducted for label noise. I think the method works well for general noise, e.g., outliers. Thus, the claim of the robustness to label noise is not accurate.

6. Additional comments.
For the hyperparameter T, I still have some concerns. The author mentioned to use validation data. Is the data clean or not? If clean, this would greatly reduce the contribution of the paper. If not, what the philosophy to choose T? Having the largest variance? Seem not reasonable.

For learning with label noise, we also care about the experiments on real-world datasets, e.g., clothing1m. It may be problematic to choose T on those real-world datasets.

Many state-of-the-art label noise learning methods haven't be compared with， making it unclear if the proposed method is superior to them or not.

10. Final recommendation based on ALL the reviews, rebuttal, and discussion (post-rebuttal)
Weak reject

11. Final justification (post-rebuttal)
Some of my concerns still exist. For example, the method proposed is not specifically designed for label noise and works for different kinds of noise (including feature noise) as well. This makes the claim misleading. Also, the validation sets are required to be clean, which greatly decrease the contribution. Many existing methods employ noisy validation set to choose hyper-parameters, e.g., when the risk is consistent. As minimizing risks on the noisy validation set is asymptotically equal to minimizing risk on the clean data. The authors also failed to compare the proposed methods with state-of-the-art ones.