Improving MAE's Fitting Ability: Fundamental and Thorough Analysis with A Simple Solution

We thank all reviewers for their helpful comments. First, we clarify the common concerns from all reviewers together:
Focus/novelty: Our work is *a further study of robust losses following MAE [9] and GCE [31]*. They proved MAE is more robust than CCE when noise exists. *However, MAE's underfitting phenomenon is not exposed and studied in the literature*. We analysed it thoroughly and proposed a simple solution to embrace both high fitting ability (accurate) and test stability (robust). Our main purpose is *not a proposal to push current best performance under label noise*. Instead, we focus on *analysing how different losses perform differently and why,* which is a fundamental research question. IMAE is suitable for cases where inputs and labels may be unmatched.
Experimental design: Our focus is to analyse why CCE overfits while MAE underfits as presented in ablation studies in Table 2. Under unknown real-world noise in Table 3, we only compared with GCE [31] as it is the most related and demonstrated to be the state-of-the-art. Thank you all for suggesting adding more evaluation. We will add Table A to make it more comprehensive. We tried $T = 8$ and $T = 16$.

Table A: Classification accuracy (%) on Clothing 1M [a] with ResNet-50. We follow exactly the same setup as others.

CCE [b]	Forward [b]	Bilevel [g]	[c]	[f]	[e]	IMAE (default: $T = 8$)	IMAE ($T = 16$)
68.9	69.8	69.9	70.4	71.1	72.2	**72.9**	**72.9**

For reviewer #1, we agree it is greatly interesting to combine with mixup [30], which may create noisy training data when augmenting data. But it is beyond our focus of analysis here.
R#1–Q1. More accessible presentation is better: the meanings of sample's weight and probability in Fig. 1 are defined in Sec. 3. Your advice is insightful. We plan to replace sample's probability with input-to-label relevance score (more conceptual level) and move Fig. 1 to Sec. 3.
R#2–Q1. It is not sufficient to use resnets only: On real-world video retrieval task [32] with unknown noise, we also applied GoogLeNet V2 [12] following [21].
R#2–Q2. Our work is quite close to [c]: They both intend to change the weight of each sample before sending to softmax. Their *critical differences* are: 1) [c] explicitly estimates latent true labels by an additional softmax layer while our IMAE reweights examples based on their input-to-label relevance scores; 2) IMAE reweights samples ***after softmax***, i.e., scaling their gradients as shown in Eq. (22).
R#2–Q3. My biggest concern is whether their method, in fact, is just a majority voting. What about adding more class dependent noise? The questions are great: 1) We choose uniform noise because it is more challenging than asymmetric (class-dependent) noise which was verified in [d]; 2) Interestingly and surprisingly, in fact, the majority voting is our reasonable assumption. We believe that if the noise accounts the majority, DNNs is hard to learn meaningful patterns. Being natural and intuitive, the majority voting decides the meaningful data patterns to learn.
R#3–Q1. Some claims are out of logic. 1) Thank you for pointing out the non-parallel terms, we will fix it. 2) *'Two claims conflict with each other?'* Sorry, we clarify their different viewpoints: The robustness/sensitive to noise is from the angle of test accuracy stability/trend, i.e., CCE's final test accuracy drops a lot versus its best one while MAE's final one is almost the same as its best one; Instead, the claim 'MAE works worse than CCE' is from the aspect of best test accuracy since we generally apply early stopping to help CCE. That is, MAE's fitting ability is much worse than CCE. In other words, CCE overfits to incorrect labels while MAE underfits to correct labels. 3) It is correct and a common practice to use one-hot vectors as the ground-truth.
R#3–Q2. The study from the gradient perspective is not new, e.g., Truncated Cauchy Non-Negative Matrix Factorization. We agree the perspective itself is not new. However, we find *how we analyse fundamentally and go to the simple solution via the gradient viewpoint is novel*. This TPAMI-2017 work truncates large errors to drop extreme outliers. The related work GCE [31] is similar to it. Instead, our IMAE adjusts weighting variance without dropping samples.
R#3–Q3. The robustness is not specific for label noise. I think the method works well for general noise, e.g., outliers. Yes, you are right. Our IMAE is suitable for all cases where inputs and their labels are not semantically matched, which may come from noisy data or labels. Since we only evaluated on label noise, we did not exaggerate its efficacy.
R#3–Q4. Clean validation data, more experiments. Following the ML literature, a validation set should be clean as we should not expect a ML model to predict noisy data well. In other words, we cannot evaluate a model's performance on noisy validation/test data. Our goal is to avoid learning faults from noisy data and generalise better during inference. We tested on real-world video retrieval task in Table 3. We will add comparison on Clothing 1M presented in Table A.

References: [a,..] is used to avoid confusion with the paper.

[a] Xiao et al. Learning From Massive Noisy Labeled Data for Image Classification. In CVPR, 2015.
[b] Patrini et al. Making deep neural networks robust to label noise: A loss correction approach. In CVPR, 2017.
[c] Goldberger et al. Training deep neural-networks using a noise adaptation layer. In ICLR, 2017.
[d] Vahdat et al. Toward robustness against label noise in training deep discriminative neural networks. In NeurIPS, 2017.
[e] Tanaka et al. Joint optimization framework for learning with noisy labels. In CVPR, 2018.
[f] Han et al. Masking: A new perspective of noisy supervision. In NeurIPS, 2018.
[g] Jenni et al. Deep bilevel learning. In ECCV, 2018.