Derivative Manipulation for General Example Weighting

Reviewer #1: Borderline

1. "The sentiment analysis dataset is not a vision dataset. The video retrieval dataset/setting is not very common ..."

This is because we try to make our experiments have diverse contexts (document, image, video), so that DM is widely evaluated and tested. In our view, this can be a merit.

2. "(i) The performances on the CIFAR are not on par with the SOTA .... the weaker network backbone is used ... The result on Clothing1M is promising but it is still worse than the SOTA ...; (ii) Generally, the method looks sound, new, and is supported by probabilistic explanations. My biggest concern is the empirical performance stands not very strong when compared with SOTA numbers."

(i) Our main goal is not to beat the SOTA. Generally, the SOTA number benefits from better network backbones, data augmentation, and training strategies. DM can be easily combined with them to make performance better.

(ii) Instead, our evaluation focuses on comparing fairly with highly relevant baselines to prove DM's effectiveness. As you may understand, the reported results from different papers may be trained with different settings. Therefore, in our tables, we highlight and group them by marking "Our Trained Results" and "Results From some paper".

(iii) Your concern makes sense. For any CVPR paper, beating the SOTA is great. However, as we do not focus on specific industrial applications, e.g., face detection and recognition, to be SOTA on CIFAR may not be very desirable.

3. "In Clothing1M, did you use the ImageNet pre-trained representation? This setting seems to be different from the CIFAR in which the networks are trained from scratch."

(i) Yes, as it is in other work. Please see Lines #642-643.

(ii) This denotes DM works in both cases: either random initialisation (CIFAR) or with pre-training (Clothing 1M).

Reviewer #2: Weak accept

1. "..., a linear classifier is assumed. If a non-linear classifier or non-linear activation function is used in the last layer, can the method in this paper still work?"

Yes. (i) As shown in Figure 1(a), a deep net and a linear classifier are wrapped as a model box to be updated. We put no assumption on this model box, as long as this box outputs a logit vector to be processed by softmax and loss layers. (ii) We focus on analysing the last two layers, i.e., softmax and loss. Softmax is a non-linear activation.

2. "For some existing in-differentiable loss functions, could the method in this paper be extended?"

Yes, it should be interesting to extend non-differentiable losses. Since we design the derivative other than the loss, as long as a derivative function is known, we can address it in a similar fashion, e.g., as noted in Lines #382-384, DM can even express some non-elementary loss functions too.

3. "Using derivative manipulation to process, will it increase computational burden or difficulty ... ?"

No. Instead, the computational burden can be reduced. In the loss and softmax layers, back-propagation is dropped. The derivative with respect to the logit vector is directly computed, whose cost is negligible.

4. "Does it make sense to compare some representatives of existing class weighting methods?"

Yes. Additionally, in our view, DM is a method for instance weighting. Presumably, DM can be an add-on of class weighting methods to make a hierarchical weighting scheme, other than being a competitor of class weighting.

5. "Other minor format issues."

Thanks. We will address them in the revision.

Reviewer #3: Weak reject

1. "Can authors provide derivative forms of more robust losses, e.g., SL[56], Forward [41], PENCIL, and so on?"

Yes, we can add this in the revision. (i) $L_{\mathrm{SL}} = \mu L_{\mathrm{CCE}} + \nu L_{\mathrm{RCE}}$. Since $L_{\mathrm{RCE}} = -\frac{A}{2} L_{\mathrm{MAE}}$, then $L_{\mathrm{SL}} = \mu L_{\mathrm{CCE}} - \frac{A\nu}{2} L_{\mathrm{MAE}}$. Following Eq. (6), $w_i^{\mathrm{SL}} = 2\mu(1-p_i)(1-\frac{A\nu}{2\mu}p_i) = 2\mu + A\nu p_i(p_i - \frac{2\mu+A\nu}{A\nu})$. We omit the constant 2μ. Then $w_i^{\mathrm{SL}} = A\nu p_i(p_i - \frac{2\mu+A\nu}{A\nu})$. If exponential transformation is added, it becomes a normal distribution as in Eq. (9). (ii) Forward focuses on the noise-transition matrix estimation while PENCIL does label estimations.

2. "With two hyper-parameters in DM, is there a guideline for hyper-parameter selection in practical scenarios?"

(i) Yes, in Figure 2(c), Section 4.1 and Appendix E, we discuss insights on how to choose them in different scenarios.

(ii) Analogously with Gaussian distributions which have mode and variance, a weight distribution has mode and variance too. Therefore, we should have two hyper-parameters.

3. "Can author provide more elaboration on the introduced concepts, especially "Emphasis Mode and Variance"?"

Yes. In a weight distribution, the mode denotes "what samples get the highest weights" while emphasis variance indicates "how big variation among all examples' weights". We can provide more elaboration in the revision.

4. "(i) It is not novel to regard the loss as an implicit reweighting (ii) Can the author provides a unified understanding on the performance of robust losses and reweighting methods? Lines #616-619 seems too rough."

(i) We focus on derivative other than loss. (ii) Yes, we explain the derivative angle as a unification of loss and weighting in Lines #041-053. For more results analysis, please see Section 4.1, Figure 2(c), Lines #611-619 and Appendix E.

5. "Is there any evidence to support "A noisy example should have smaller derivative magnitude" in Line 115?"

Generally speaking, the impact on a model's update of a noisy example is decided by its derivative magnitude.