This is an unofficial letter from me to let peers know better about our work. Here, I declaim:

1. In this letter, `we' means `I', all replies only represent my personal viewpoint, not my co-authors.
2. If there is something improper, please kindly let me know. It is greatly appreciated.

-Xinshao Wang, 15/08/2020

#2786: ProSelfLC: Progressive Self Label Correction for Training Robust Deep Neural Networks

We thank all reviewers #1, #2, #3, #4 for their comments. Some of them are detailed, constructive and insightful. Unfortunately, we have to remark:

1. Most comments from the reviewer#1 make no sense. Please see Lines 7-15 of this letter.
2. Most comments from the reviewer#3 are harsh. Please see Lines 38-46 of this letter.

R#1 (S3-F3): "The presented result is a simple extension of previous label correcting methods..." Sorry, we absolutely disagree. Although the mathematical modelling and format are simple, our study presents some significant analysis: (1) relying on a learner's knowledge versus human annotations; (2) rewarding or penalising a low entropy status? Please kindly read Lines 52-61, 63-69 and 232-253 in more detail, and clarify the reasons if you still disagree. **"The paper is poorly organized, scattered with fragmented sentences and has a lot of grammatical and syntax issues"** Sorry, this strong negative comment is too abstract and unconvincing without details, thus being irresponsible. **"The "theory" of the paper is mostly re-iterating previous results and is not rigorous..."** If something is already surrounded by prior work, then we suppose there is no need to clarity it. For example, please see Lines 60-61, **"Also the model for CIFAR-100 is not specified ..."** Sorry, you missed it. Please see Table 4.

R#2 (S5-F4): "...It would be better to have a theoretical analysis that demonstrates how ProSelfLC and CCE react differently to the label noise." Our purpose is to provide a practical way for training DNNs robustly. It is quite natural to believe that noise exists in large-scale datasets. We empirically compare their learning dynamics in Figure 3. Theoretical comparison generally requires many assumptions, which we try to avoid in this work. Please refer to Lines 60-61 and Table 1 for comparing the underlying ideas and mathematical expression analysis. **"There is no direct empirical evidence to support the claim...accredited by long learning time and low entropy..."** The empirical intuitions are (long and short are relative terms in this work): (1) A learner should be unreliable at the beginning or an earlier learning phase; (2) After a model is trained for a longer time, we change to believe its confident predictions instead of human annotations. *Direct empirical evidences:* Figures 2(b) and 3(c) display label correction dynamics. Figures 3(d) and 3(e) show entropy dynamics. Figures 2(c) and 3(f) show generalisation dynamics. **""Human annotations and predicted label distributions, which should we trust more?", may be vague and pointless without restrictions."** A very thoughtful point, thanks. Basically, we assume large-scale applications and test on them. We will discuss restrictions, e.g., data requirements, for this statement if our work is accepted. **"The objective function (10) may not be well designed."** (1) Regarding g(t), yes, the convergence rate depends on many factors. Therefore, *the number of total training iterations varies when dataset changes, i.e., we do not propose to fix the learning time everywhere*; (2) It is fine for l(p) to be identical for two different inputs. **"The experiments can be improved by adding one more model, e.g., mobilenet... "** Thanks, we will add another model (e.g., MobileNet) if this work is accepted. We tested ResNet-44 and ResNet-50. **"Standard image classification: CIFAR and ImageNet are 100% accurate?"** (1) CIFAR and ImageNet are too large, we cannot make sure they are 100% accurate; (2) Please see our definition of label correction in Section 4.1.

R#3 (S4-F3): "Method seems arbitrary: ...motivate the presented method from some intuitive change...ablation study..." *This comment is too harsh.* Sorry, we absolutely disagree: (1) For motivations, please see Lines 169-195; (2) Experimentally, the method is compared with its variants in Figure 2, and other counterparts (baselines) in Figure 3. **"Only modest improvements, especially given the additional complexity introduced."** Sorry, we definitely disagree: (1) ProSelfLC outperforms the others in all cases, sometimes significantly as shown in Table 4. *It is too harsh to require a big gap everywhere*; (2) At the loss layer, the computation time of $g(t)$ and $l(\mathbf{p})$ is negligible. **"How does a 1 pt gain in noisy ImageNet affect the broader research community and beyond?"** *Again, this comment is really harsh.* As deep learning (DL) becomes the most popular approach in AI/ML systems, a fundamental analysis about DL definitely has great impact. As highlighted, we present new insights on training better DNNs.

R#4 (S7-F3): "The related work could be more thorough and balanced...knowledge distillation..." We discussed knowledge distillation in related work, please read Lines 88-93. We will make it more thorough as suggested. **"CCE is undefined. I assume it's class cross entropy?"** Yes, CCE = categorical 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?"** Really sorry, it's a typo/mistake. In fact, we would like to remark: "a learner is **NOT GIVEN** whether an example is clean or not." **"(1) A simple way to reduce the proposed loss is to always produce confident predictions ...simply not change its predictions ...; (2) Have you considered adding a loss that penalized large epsilons, to avoid such a degenerate solution?"** (1) You are partially right here. Rewarding confident predictions is a selling point in this work, as it challenges a recently popular practice–confidence penalty. However, rewarding confidence and without changing confident predictions ***happens only when a model becomes confident at the later training phase***; (2) Large epsilons will not lead to a degenerated solution. Epsilon becomes large when a learner becomes confident at the later training phase, which is what we desire to properly correct labels by exploiting a model's confident knowledge.