Supplementary Material for "Towards Discovering the Effectiveness of Moderately Confident Samples for Semi-Supervised Learning"

Hui Tang and Kui Jia*
South China University of Technology
381 Wushan Road, Tianhe District, Guangzhou, Guangdong, China
eehuitang@mail.scut.edu.cn, kuijia@scut.edu.cn

A. More Explanations on Technical and Experimental Details

Performance on larger datasets. We also do experiments on the larger, challenging STL-10 [1] dataset, which has a distribution gap between labeled and unlabeled data. STL-10 consists of $5,000$ labeled RGB images of size 96×96 and $100,000$ unlabeled images. The labeled set is split into ten pre-defined folds of $1,000$ images each. We evaluate on five of these ten folds. Under the same baseline condition, our proposed GSF improves over FixMatch [2] by a large margin (2.89%), verifying its efficacy on the difficult task. Due to time limit, it is difficult to present here additional experiments on more large datasets (e.g., ImageNet); we expect that performance improvement over FixMach would be similarly obtained. We will try to acquire additional results soon.

Time cost of model training. Under the same baseline condition and on STL-10, the training efficiencies of GSF, PPF, and FixMatch are respectively 1.05 it/s, 1.16 it/s, and 1.20 it/s ('it' means a training iteration); their memory consumption is almost identical.

Clarification on the design of GSF. GSF uses the gradient of features outputted by the feature extractor, since **1)** they are at the highest abstraction level and thus contain rich information of global context and semantics, **2)** they are in the feature space that we aim to learn and are thus closely related to optimization dynamics, and **3)** their dimension is much lower than that of input space, enabling lower computation.

Clarification on the design of PPF. We use Euclidean distance that gives improved performance over cosine similarity (*e.g.*, by 3.87% on CIFAR-10 with 40 labels).

Spectrum of confidences for samples selected by GSF and PPF. We show here the confidence spectrum in Fig. S1.

Relationship between GSF and PPF. Features used by PPF are *static* data points in the learned feature space, which to some extents characterize the semantics of particular classes; in contrast, feature gradients used by GSF represent the update direction of these feature points (*i.e.*, optimization *dynamic*). They can be combined to further improve the performance since they contain complementary information in that the samples selected by GSF and PPF form different sets. We will study the scheme in future work.

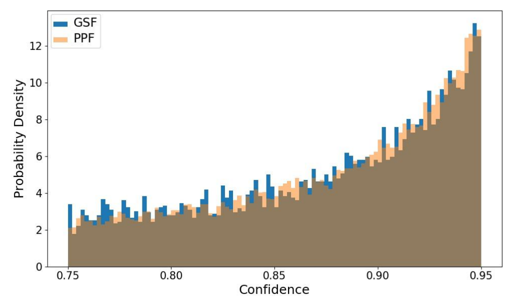

Figure S1. Confidence spectrum (128^{th} epoch, CIFAR10@40).

References

[1] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011. 1

[2] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, volume 33, pages 596–608, 2020. 1

*Corresponding author.