COMP6258 Reproducibility Challenge: Does Label Smoothing Help Deep Partial Label Learning?

Yanzhe S. Zhang[*], William Toland[†], Jack Booth[‡], Winston McCarthy [§]

February 26, 2025

Paper Information:

Openreview URL	https://openreview.net/pdf?id=drjjxmi2Ha
Is code available:	No, the repository is either private or deleted.

Plan:

What is the motivation of the paper (i.e. context)? (∼50-100 words)

To reduce training-time labeling costs for machine learning classification tasks, researchers have taken an interest in partial label learning (PLL) . This is a process in which a machine learns to identify the single true label from a set of multiple candidate labels. DNNs perform poorly on this task due to the overconfidence in it's predictions reinforcing the incorrect labels. The paper offers label smoothing (LS) as a solution to this problem by forcing the DNN to be less certain of any single prediction.

What are the key claimed contributions of the paper? (∼50-100 words)

The paper provides a mathematical basis and experimental results concluding that LS is effective for deep PLL with superior classification performance when compared to without LS. They claim that LS encourages activations of the penultimate layer to be close to the template of the correct class, and equally distant to the templates of incorrect classes on PLL datasets, regardless of architecture. Meaning LS is effective in learning distinguishable representations for deep PLL, and on pre-logits is independent of architecture, datasets, or noise levels. The optimal smoothing rate performance is consistent with the best effect on pre-logit representations.

What experiments are you aiming to do (be specific: datasets, models, how long will you train for, etc.)? (∼300-400 words)

We first aim to reproduce the implementation of the models used in the paper. These are models trained on four datasets: Fashion-MNIST, Kuzushiji-MNIST, CIFAR-10 and CIFAR-100. We must produce a new label space constituted by the top-K highest probabilities predicted by a neural network trained on the original clean dataset excluding the ground truth label. We use a LeNet-5 architecture for the Fashion-MNIST and Kuzishiji-MNIST datasets, ResNet-18 for the CIFAR-10 dataset and ResNet-56 for the CIFAR-100 dataset.

For the initial experiments, we use the same model architectures and datasets used to produce the new label space. Each model will be trained using stochastic gradient descent (SGD) with a momentum of 0.9, and weight decay of 1×10^{-3}. We will use a mini-batch size of 128, an initial learning rate of 0.01, and train for 200 epochs to ensure adequate convergence. This setup mirrors the original experiments, enabling a direct comparison with the reported results.

The first experiment aims to explore the effect of LS for deep PLL on classification performance. For this, we will compare test accuracies for these datasets and architectures both with and without label smoothing on the PLL datasets under various noise levels, as in the original paper.

The second experiment compares the pre-logits of the architectures trained both with and without label smoothing under various noise levels, to visualise the effect of LS on pre-logits.

[*]yz25g21@soton.ac.uk
[†]welt1g21@soton.ac.uk
[‡]jgb1g21@soton.ac.uk
[§]wm2g20@soton.ac.uk

Upon testing the machine on the aforementioned datasets, we tune the smoothing parameter (0.1, 0.3, 0.5, 0.7 and 0.9) to check whether the optimal smoothing rate is consistent with the pre-logit representations as well as checking if it is possible to mathematically determine whether the result is consistent with the mathematically determined optimal smoothing rate.

In addition to replicating the experiments reported, we also plan to test on Tiny ImageNet, using the same architecture and parameters as used for CIFAR-100, to further assess the generalisation ability and robustness of the result on a larger dataset. We also plan an experiment generating a random set of labels for the label space, rather than the top-K highest probabilities.