Learning from Noisy Labels via Discrepant Collaborative Training
Supplementary Material

Yan Han[1,2], Soumava Kumar Roy[1,2], Lars Petersson[1,2], Mehrtash Harandi[2,3]
[1]The Australian National University, [2]DATA61-CSIRO, Australia, [3]Monash University
yan.han@anu.edu.au, soumava.kumarroy@anu.edu.au
Lars.Petersson@data61.csiro.au, mehrtash.harandi@monash.edu

In this supplementary material, we study the robustness of our DCT design with respect to its parameters, namely λ_2 and λ_3 (see Eq. (1), below). We will first report the value of the hyper parameters used in our experiments. Then we will analyze and investigate the effect of the hyper-parameters over the performance of the DCT algorithm.

Table 1. The accuracy (%) of DCT for various setups (using CIFAR10 dataset contaminated by symmetric noise with rate of 50%.)

λ_2	λ_3	**DCT(%)**	**Descriptions**
0	0	74.02	Baseline
0	0.0001	74.35	Consistency loss only
0.001	0	77.11	Diversity loss only
0.001	0.0001	**78.50**	Optimal case
0.001	0.00001	77.88	-
0.001	0.0005	78.32	-
0.001	0.001	77.79	-
0.01	0.0001	77.94	-
0.005	0.0001	78.24	-
0.0001	0.0001	77.23	-

A. Hyper-Parameters

We recall that the loss of the DCT algorithm as:

$$\mathrm{Loss} = \mathrm{L}_1 + \lambda_3 \mathrm{L}_3 - \lambda_2 \mathrm{L}_2. \tag{1}$$

Here, L_1 is the classification loss (for each network), L_2 and L_3 are diversity and consistency losses, respectively. Furthermore, λ_2 and λ_3 denote the associated weights of the diversity and consistency loss, respectively. In all of our experiments, we set $\lambda_2 = 1e-3$. We set $\lambda_3 = 1e-3$ for MNIST, CUB200-2011 and CARS196. For CIFAR10 and CIFAR100, we set $\lambda_3 = 1e-4$. In all our experiments, we used a Gaussian kernel for MMD with $\sigma = 0.05$. We stress that the hyper-parameters reported above are used across all noise settings.

B. Robustness of the DCT algorithm

In this part, we analyze the robustness of the DCT algorithm with respect to its parameters. By doing so, we evaluate the performance of the DCT algorithm on the CIFAR10 dataset by varying the values of λ_2 and λ_3. Table 1 shows the accuracy of the DCT algorithm for the 50% symmetric noise (which is a challenging setup) for various values of λ_2 and λ_3. First note that for $\lambda_2 = \lambda_3 = 0$, we recover the vanilla co-training framework with an accuracy of 74.02%. Setting $\lambda_3 = 1e-4$ and $\lambda_2 = 0$ results in adding only the consistency loss and a modest increase in the performance (0.33% to be exact). Interestingly, by just adding the diversity loss ($\lambda_2 = 1e-3$ and $\lambda_3 = 0$), the accuracy soars to 77.11%, a significant improvement over the vanilla co-training solution. This confirms the premise of our work, *i.e.*, the importance of diversity in co-training.

Another observation in favor of the DCT algorithm is its robustness to the variation of λ_3 and λ_2. For example, by fixing $\lambda_2 = 0.001$ and varying λ_3 in a wide range from $1e-5$ to $1e-3$, the accuracy varies in the range $[77.79\%, 78.50\%]$. Similar trends can be observed if we pick λ_2 reasonably. For example, with $\lambda_3 = 1e-4$, changing λ_2 from $1e-4$ to $1e-2$ results in accuracies in the range of $[77.23\%, 78.50\%]$.