Supplementary Material for Joint Optimization Framework for Learning with Noisy Labels

Daiki Tanaka Daiki Ikami Toshihiko Yamasaki Kiyoharu Aizawa
The University of Tokyo, Japan
{tanaka, ikami, yamasaki, aizawa}@hal.t.u-tokyo.ac.jp

A. Detailed Architecture

Table 1 details the network architecture used in the experiments on the CIFAR-10 dataset. It is based on PreAct ResNet-32 [1].

B. Dependency on Hyper Parameters

We show the hyper parameters used in the experiments on SN-CIFAR in Table 2. If the noise rate is high, the optimal learning rate also tends to be high.

The prediction accuracy is not so sensitive to the hyper parameters and our method demonstrated good performance with a different set of the hyper parameters as shown in Table 3, 4, 5, 6. In addition, Table 7, 8 show the validation accuracy with different t_1 and t_2, where t_1 is the value at which to start label-updating, and t_2 is the value at which to stop label-updating. When we train the network with a high learning rate, the prediction accuracy retains high value, and thus we can start label-updating when the validation accuracy once reach high value. Label-updating should be stopped after the training loss converge.

Table 1. The network architecture used in the experiments on CIFAR-10.

NAME	DESCRIPTION
input	32×32 RGB imgae
conv	32 filters, 3×3, pad=1, stride=1
unit1	(pre-activation Residual Unit 32→32)×5
unit2a	pre-activation Residual Unit 32→64
unit2b	(pre-activation Residual Unit 64→64)×4
unit3a	pre-activation Residual Unit 64→128
unit3b	(pre-activation Residual Unit 128→128)×4
pool	Batch Normarization, ReLU, Global average pool (8×8→1×1 pixels)
dense	Fully connected 128→10
output	Softmax

Table 2. The hyper parameters used in the experiments on SN-CIFAR.

noise rate (%)	0	10	30	50	70	90
α	1.2	1.2	1.2	1.2	1.2	0.8
β	0.8	0.8	0.8	0.8	0.8	0.4
learning rate	0.01	0.02	0.03	0.04	0.08	0.12

Table 3. Validation accuracy with different hyper parameters in the triple test (experimented on AN-CIFAR with noise rate = 0.4).

$\beta = 0.4$, learning rate = 0.03							
α	0.1	0.2	0.5	**0.8**	1.0	2.0	5.0
val (%)	91.9	92.0	91.7	92.0	92.1	92.1	88.8

$\alpha = 0.8$, learning rate = 0.03							
β	0.05	0.1	0.2	**0.4**	0.5	1.0	2.0
val (%)	90.8	91.7	91.8	92.0	91.6	89.5	91.1

$\alpha = 0.8, \beta = 0.4$							
learning rate	0.005	0.01	0.02	**0.03**	0.05	0.1	0.2
val (%)	90.6	90.9	91.3	92.0	92.1	91.3	88.5

Table 4. Validation accuracy with different hyper parameters in the triple test (experimented on AN-CIFAR with noise rate = 0.2).

$\beta = 0.4$, learning rate = 0.03							
α	0.1	0.2	0.5	**0.8**	1.0	2.0	5.0
val (%)	92.9	92.9	93.0	93.2	93.1	93.2	89.7

$\alpha = 0.8$, learning rate = 0.03							
β	0.05	0.1	0.2	**0.4**	0.5	1.0	2.0
val (%)	92.6	93.0	93.2	93.2	93.1	92.8	92.8

$\alpha = 0.8, \beta = 0.4$							
learning rate	0.005	0.01	0.02	**0.03**	0.05	0.1	0.2
val (%)	92.5	92.7	92.7	93.2	92.7	91.8	89.2

Table 5. Validation accuracy with different hyper parameters in the triple test (experimented on SN-CIFAR with noise rate = 0.7).

$\beta = 0.8$, learning rate = 0.08							
α	0.1	0.2	0.5	1.0	**1.2**	2.0	5.0
val (%)	85.7	86.0	85.5	85.9	85.5	85.7	83.8

$\alpha = 1.2$, learning rate = 0.08							
β	0.05	0.1	0.2	0.5	**0.8**	1.0	2.0
val (%)	82.0	82.3	83.1	85.3	85.5	85.2	30.3

$\alpha = 1.2, \beta = 0.8$							
learning rate	0.005	0.01	0.02	0.05	**0.08**	0.1	0.2
val (%)	79.5	80.7	82.8	85.4	85.5	85.4	83.8

Table 6. Validation accuracy with different hyper parameters in the triple test (experimented on SN-CIFAR with noise rate = 0.3).

$\beta = 0.8$, learning rate = 0.03							
α	0.1	0.2	0.5	1.0	**1.2**	2.0	5.0
val (%)	91.6	91.7	91.5	91.8	91.8	91.8	89.9

$\alpha = 1.2$, learning rate = 0.03							
β	0.05	0.1	0.2	0.5	**0.8**	1.0	2.0
val (%)	90.0	90.4	91.2	91.8	91.8	91.9	91.0

$\alpha = 1.2$, $\beta = 0.8$							
learning rate	0.005	0.01	0.02	**0.03**	0.05	0.1	0.2
val (%)	90.1	90.7	91.0	91.8	92.1	91.1	89.0

Table 7. Validation accuracy with different t_1 (start epoch) and t_2 (stop epoch) in the triple test (experimented on AN-CIFAR with noise rate = 0.4, $\alpha = 0.8$, $\beta = 0.4$, learning rate = 0.1).

start epoch	0	50	**70**	100	150
val (%)	58.4	90.3	91.3	91.4	91.6

stop epoch	100	150	**200**	250	300
val (%)	91.8	91.5	91.3	90.8	90.7

Table 8. Validation accuracy with different t_1 (start epoch) and t_2 (stop epoch) in the triple test (experimented on SN-CIFAR with noise rate = 0.7, $\alpha = 1.2$, $\beta = 0.8$, learning rate = 0.08).

start epoch	0	50	**70**	100	150
val (%)	38.0	84.7	85.5	86.1	85.6

stop epoch	100	150	**200**	250	300
val (%)	85.0	85.6	85.5	85.9	85.6

C. Effect of Soft-Labeling

We show the analysis of the effect of soft-labeling on the noisy CIFAR-10 dataset in Table 9, 10. The soft-labels with high probability are almost correct. Conversely, when the probability is low, the label seems to be updated incorrectly. As opposed to the hard-labels, the soft-labels contain the probabilities of each class in themselves, and thus the network can consider the incorrectly updated labels as not important.

Table 9. Recovery accuracies of the updated soft-labels whose maximum probabilities p are within each range (experimented on AN-CIFAR with noise rate = 0.4).

p	$1-0.99$	$0.99-0.95$	$0.95-0.9$	$0.9-0$	$1-0$
acc (%)	99.8	96.9	91.3	73.1	95.1
number	27046	8647	3484	5823	45000

Table 10. Recovery accuracies of the updated soft-labels whose maximum probabilities p are within each range (experimented on SN-CIFAR with noise rate = 0.7).

p	$1-0.99$	$0.99-0.95$	$0.95-0.9$	$0.9-0$	$1-0$
acc (%)	97.5	82.2	70.6	53.3	86.4
number	27591	7368	3351	6690	45000

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.