Supplementary Material — SSR: An Efficient and Robust Framework for Learning with Unknown Label Noise

Chen Feng
https://sites.google.com/view/mr-chenfeng
Georgios Tzimiropoulos
https://ytzimiro.github.io/
Ioannis Patras
https://sites.google.com/view/ioannispatras

School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK

Supplementary A: Dataset details

Synthetic noisy dataset CIFAR10 and CIFAR100 both consist of 50K images. Following the standard practice, for CIFAR10 and CIFAR100, we evaluate our method with two types of artificial noise: *symmetric noise* by randomly replacing labels of all samples using a uniform distribution; and *asymmetric noise* by randomly exchanging labels of visually similar categories, such as Horse $\leftrightarrow$ Deer and Dog $\leftrightarrow$ Cat. For closed-set noise only dataset, we test with 20%, 50%, 80% and 90% symmetric noise and 40% asymmetric noise following DivideMix [4]. For datasets including also open-set noise, following settings in EDM [7], we test with 30%, 60% total noise ratio and 50%, 100% open-set noise ratio on CIFAR10 dataset. The total noise ratio denotes the total proportion of noisy samples in the population of samples while the open-set noise ratio denotes the proportion of open-set noise in the noisy samples. The closed-set noise is generated as *symmetric noise* while the open-set noise is randomly sampled from CIFAR100.

Real-world noisy dataset WebVision [5] is a large-scale dataset of 1000 classes of images crawled from the Web. Following previous work [3, 4, 6], we compare baseline methods on the top 50 classes from Google images Subset of WebVision. The noise ratio is estimated to be around 20%. ANIMAL-10N [9] is a smaller and recently proposed real-world dataset consisting of 10 classes of animals, that are manually labelled with an error rate that is estimated to be approximately 8%. ANIMAL-10N has similar size characteristics to the CIFAR datasets, with 50000 train images and 10000 test images. Clothing1M [10] is a large-scale dataset of 14 classes of clothing images crawled from online shopping websites, consisting of 1 million noisy images. The noise ratio is estimated to be around 38.5%.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Supplementary B: Implementation details

We use a PresActResNet-18 [7] as the backbone for all CIFAR10/100 experiments following previous works. Unlike previous methods that use specific warmup settings for CIFAR10/CIFAR100, we train the model from scratch with $\theta_s = 1.0$ in all experiments. We set $\theta_r = 0.8$ for higher noise ratio — sym50, sym80 and sym90 noise, $\theta_r = 0.9$ for remain settings in all CIFAR experiments except in the corresponding ablation part. We train all modules with the same SGD optimizer for 300 epochs with a momentum of 0.9 and a weight decay of 5e-4. The initial learning rate is 0.02 and is controlled by a cosine annealing scheduler. The batchsize is fixed as 128.

For WebVision, we use InceptionResNetv2 following [4]. We train the network with SGD optimizer for 150 epochs with a momentum of 0.9 and a weight decay of 1e-4. The initial learning rate is 0.01 and reduced by a factor of 10 after 50 and 100 epochs. The batchsize is fixed as 32. For Clothing1M, we use ResNet50 following [4] with ImageNet pretrained weights. We train the network with SGD optimizer for 150 epochs with a momentum of 0.9 and weight decay of 1e-3. The initial learning rate is 0.002 and reduced by a factor of 10 after 50 and 100 epochs. The batchsize is fixed as 32. For ANIMAL-10N, we use VGG-19 [8] with batch-normalization following [9]. We train the network with SGD optimizer for 150 epochs with a momentum of 0.9 and weight decay of 5e-4. The initial learning rate is 0.02 and reduced by a factor of 10 after 50 and 100 epochs. The batchsize is fixed as 128. For all real-world noisy datasets, we train the model from scratch with $\theta_s = 1$, while θ_r is fixed as 0.95.

Following recent works, in this work, we define three augmentation strategies: original image which we denote with 'none' augmentation for testing, random cropping+horizontal flipping which we denote as 'weak' augmentation, and 'strong' augmentation the one that further combines the augmentation policy from [1]. For L_{ce} we use 'strong' augmentation with mixup interpolations [11] while for L_{fc}, we use 'weak' augmentation for $\boldsymbol{x}_2$ and 'strong' augmentation for $\boldsymbol{x}_1$ in eq.(4). For mixup interpolation, following DivideMix [4], we set α, β as 4 for beta mixture for the CIFAR10/CIFAR100 datasets, and as 0.5 for the real-world noisy dataset.

Supplementary C: Evaluation on different choices for sample selection

In this section, we extensively compare the performance of different sample selection mechanisms under different noise ratios and modes. More specifically, with the PMC g_p and NPK g_q, we compare the performance of these two classifiers in two different selection modes. Pre-defined mode means the number of noisy samples is known. That is, given a symmetric noisy dataset with noise ratio as τ, we select the top $1-\tau$ percent of the samples as clean according to its loss value or prediction confidence. Automatic mode means that there is no information about the ratio of noise. In this case, we utilize the consistency measure c using NPK in our method and the GMM-based loss modelling for PMC. Note that here we only considered two simple modes for sample selection and one variant in each mode respectively. There are many different variants proposed, however, we just aim to show the robustness of NPK over PMC here. For a fair and clear comparison, we also show results of training with the whole dataset and clean subset as the bottom baseline and top baseline, respectively. Note, that in order to compare only the effect of the sample selection part, we

exclude sample relabeling, strong data augmentation and optional feature consistency loss here.

Methods			50% sym	80% sym	40% asym	60% all(50% open)	60% all(100% open)
Automatic	PMC	Last	88.41	52.19	48.88	83.91	**85.24**
		Best	88.52	64.03	79.24	84.08	**85.32**
	NPK(Ours)	Last	**88.82**	**68.67**	**88.20**	**85.27**	84.72
		Best	**88.88**	**69.57**	**88.99**	**85.30**	85.17
Pre-defined	PMC	Last	87.85	65.93	**84.33**	85.75	86.16
		Best	87.98	66.33	**85.41**	85.76	86.48
	NPK	Last	**88.20**	**72.25**	75.47	**85.84**	**86.43**
		Best	**88.27**	**72.66**	83.36	**86.10**	**86.53**
Whole dataset		Last	58.93	27.59	76.68	59.92	80.99
		Best	80.86	61.55	85.41	79.26	84.18
Clean subset		Last	92.46	87.64	93.40	90.70	90.62
		Best	92.53	87.85	93.56	90.87	90.82

Table 1: Results with different sample selection mechanisms.

In table 1, we can see that the NPK-based selection achieved better performance compared to PMC-based selection regardless of the mode, and that our choice can significantly improve the baseline (Whole dataset) without knowing the noise ratio.

Supplementary D: Distance metric for feature consistency loss

In Section 3.3, we use the cosine distance as the distance metric for feature consistency loss. Here we also experiment with the use of the L2 distance. The results are shown in table 2, where it can be seen that there are small differences, but that the cosine similarity is in general better.

Training	50% sym	90% sym	40% asym	60% all (50% open-set)
SSR+(negative cosine similarity)	**96.1**	**95.2**	95.5	**95.2**
SSR+(L2 distance)	96.0	94.7	**96.1**	94.3

Table 2: SSR+ with different distance metric for feature consistency loss

Supplementary E: Computational cost analysis

In table 3, we report the running time of each step of our model on the datasets that we have experimented with. It can be seen that the time of sample selection and relabelling is almost negligible compared to the time of gradient propagation.

Dataset(Size)	Model training	Sample selection and relabelling		
		Feature extraction	sample selection	sample relabelling
CIFAR(50K)	112s	9s	1.23s	$\sim$ 0s
WebVision(65K)	587s	109s	1.48s	$\sim$ 0s
Clothing1M(32K)	575s	57s	0.79s	$\sim$ 0s

Table 3: Computational cost analysis.

References

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[3] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR, 2018.

[4] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[5] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[6] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615, 2021.

[7] Ragav Sachdeva, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Evidentialmix: Learning with combined open-set and closed-set noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3607–3615, 2021.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019.

[10] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[11] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.