Learning with Noisy labels via Self-supervised Adversarial Noisy Masking Supplementary Material

Yuanpeng Tu[1*] Boshen Zhang[2*] Yuxi Li[2] Liang Liu[2] Jian Li[2]
Yabiao Wang[2] Chengjie Wang[2,3†] Cai Rong Zhao[1†]
[1]Dept. of Electronic and Information Engineering, Tongji Univeristy, Shanghai
[2]YouTu Lab, Tencent, Shanghai, [3]Shanghai Jiao Tong University
{2030809, zhaocairong}@tongji.edu.cn
{boshenzhang, yukiyxli, leoneliu, swordli, caseywang, jasoncjwang}@tencent.com

1. Detailed Experimental Results

Quantitative Results. To further verify the regularization effect of SANM on the activation map, we visualize the activation maps of DivideMix and SANM (DivideMix) for both correctly and incorrectly labeled samples. As shown in Fig. 1, for DivideMix, due to the influence of noisy samples, the activation maps of correctly labeled samples are focused on specific areas of the target instead of covering the whole object, while the results of mislabeled ones pay more attention to the meaningless background area, resulting in poor representation quality. By contrast, the results of SANM (DivideMix) cover the central regions of the objects for both samples and more information can be covered in our features, even for the samples that are not object-centric (see Fig. 3), indicating that SANM can alleviate confirmation bias and prevent models from over-fitting to noisy labels by imposing explicit regularization on the peak locations. Finally, we also provide the visualization results by the reconstruction component within the proposed SANM on CIFAR-10, as shown in Fig. 2. The results indicate that the model trained with the proposed masking reconstruction objective is capable of generating high-quality reconstructed images that cover most of the semantic information, and this in turns helps the model alleviate the negative effect of noisy labels.

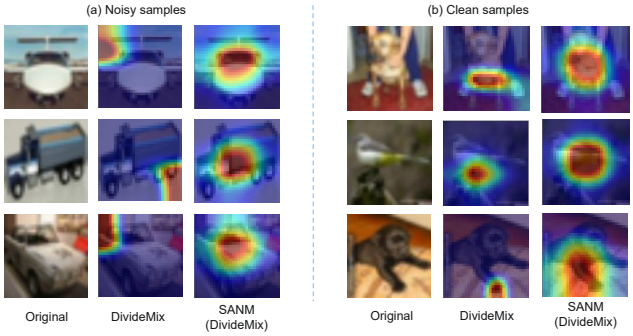

Figure 1. Activation maps for samples with noisy and clean labels between DivideMix and SANM (DivideMix).

Figure 2. The reconstruction result of SANM on CIFAR-10 of SANM (DivideMix). (a) Original images. (b) The corresponding reconstruction results.

Visualization of Clothing1M. CIFAR-10/100 are relatively small-scale and lack of diversity, therefore, besides that the Fig.1 provided in the manuscript, we further visualize more examples on real-world Clothing1M dataset to verify our motivation. As shown in Fig. 4, similar phenomena to the results on CIFAR-10 can be observed as well. Activation maps of mis-predicted samples by the model trained with clean labels are usually focused on their foreground areas. By contrast, when the model is trained with noisy labels, it tends to generate results focused on the meaningless background area for the mis-predicted samples. And even for cases where the prediction is correct, the high-activated region drifts to irrelevant areas of the object, rather than the main regions of the object. It supports the empirical findings in our manuscript that model trained with mislabeled samples is likely to overfit to some corner parts of the object from noisy data or remember the less informative regions (i.e., background).

Influence of Mask Sizes. We further investigate the effect of different aspect ratios δ (i.e., Eq. 3) in the manuscript) on CIFAR-10/100, the results are shown in Table. 1 and 2, where "Fixed" denotes adopting a fixed aspect

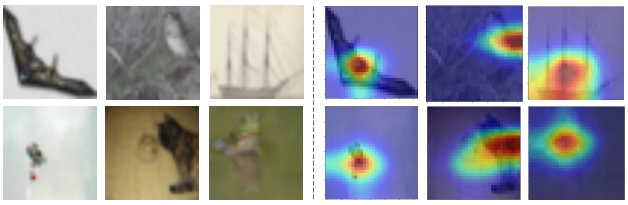

Figure 3. The activation maps of images where objects are not in the centre on CIFAR-10 dataset by SANM (DivideMix).

Table 1. Influence of different aspect ratios δ on CIFAR-10.

Dataset	CIFAR-10			
Method/Noise ratio	20%	50%	80%	90%
Fixed (0.5)	95.5	94.4	93.2	90.5
Fixed (0.8)	96.0	94.8	93.7	91.0
Fixed (0.9)	96.0	95.0	93.8	90.8
Uniform (0.5)	96.2	95.5	94.4	91.8
Uniform (0.8)	**96.4**	**95.8**	**94.6**	**92.3**
Uniform (0.9)	**96.4**	95.7	**94.6**	92.0

Table 2. Influence of different aspect ratios δ on CIFAR-100.

Dataset	CIFAR-100			
Method/Noise ratio	20%	50%	80%	90%
Fixed (0.5)	79.7	76.7	67.8	42.8
Fixed (0.8)	80.5	76.9	68.1	42.5
Fixed (0.9)	80.3	77.2	68.0	42.7
Uniform (0.5)	80.6	77.5	68.4	43.0
Uniform (0.8)	**81.2**	**78.2**	**68.7**	**43.5**
Uniform (0.9)	80.9	78.0	68.3	43.3

ratio for all samples while "Uniform" represents that δ is sampled from a uniform distribution for each sample. As the results show, uniformly-sampled δ performs better than using a fixed aspect ratio. And the results indicate that the performance of SANM is not sensitive to the choice of δ within a certain coarse range.

Experimental Details and Results on Real-world Noisy Datasets. On Clothing1M, we adopt ResNet-50 as the backbone, which is trained for 80 epochs with a batch size of 64. The Adam optimizer is used and the initial learning rate is 0.002 with a reduction factor of 10 after 40 epochs. On Animal-10N, we follow the previous method SSR [2] to utilize VGG-19 as backbone and train for 150 epochs with a batch size of 64. The SGD optimizer is adopted and the learning rate is initially set as 0.02 with a reduction factor of 10 after 50 and 100 epochs. Table 3 describes the results on the Animal-10N. It shows that compared with previous LNL competitors, i.e., ActiveBias [1], PLC [9], Co-teaching [3], SELFIE [6], CREMA [8] and SSR [2], SANM achieves high classification accuracy of

Table 3. Comparison with state-of-the-art methods in test accuracy on Animal-10N.

Method	Test Accuracy (%)
Cross-Entropy	79.4
ActiveBias [1]	80.5
PLC [9]	83.4
Co-teaching [3]	80.2
SELFIE [6]	81.8
CREMA [8]	84.2
SSR [2]	88.5
SANM(SSR)	**89.3**

89.3% and sets the new state-of-the-art, indicating that the proposed method is able to handle fine-grained noisy datasets (i.e., the noisy samples within Animal-10N are mostly caused by misjudgement on semantically similar sample pairs, e.g., cat and lynx) as well.

Generality of SANM. To demonstrate the generality of SANM, we investigate more applications leveraging SANM to boost current mainstream LNL frameworks. Specifically, the naive CE, Co-teaching [4], CDR [7], and ELR+ [5] are chosen as baselines in our experiments. Here we list both the last and best performance of these baselines and their corresponding SANM-improved versions in Table. 4, where consistent performance boost can be observed for all the corresponding baselines across all the noisy cases. Even when training with the naive CE loss, leveraging the proposed SANM framework can already achieve a considerable performance compared with other LNL frameworks (e.g., CO-teaching, CDR). Therefore, all the results verify the generalization of SANM to boost existing LNL approaches.

Experimental results with statistical significance. Besides the results shown in the main text, we provide detailed results with Standard Deviations (STD) to show the statistical significance of proposed method on CIFAR-10/100. Which are shown in Table 5.

References

[1] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active Bias: Training more accurate neural networks by emphasizing high variance samples. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1002–1012, 2017. 2

[2] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. S3: Supervised self-supervised learning under label noise. *arXiv preprint arXiv:2111.11288*, 2021. 2

[3] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy

(a)Clean-trained / Mis-predicted (b) Noisy-trained / Mis-predicted (c) Noisy-trained / Correctly-predicted

Figure 4. The activation maps of the trained base model on Clothing1M dataset.

Table 4. Comparison between the LNL methods and their SANM applications with symmetric noise on CIFAR-10/100. Specifically, the 9-layer CNN is adopted as the backbone network of Co-teaching.

Dataset		CIFAR-10				CIFAR-100			
Method/Noise ratio		20%	50%	80%	90%	20%	50%	80%	90%
CE	Best	86.8	79.4	62.9	42.7	62.0	46.7	19.9	10.1
	Last	82.7	57.9	26.1	16.8	61.8	37.3	8.8	3.5
SANM(CE)	Best	**92.4**	**89.7**	**72.1**	**51.5**	**70.9**	**53.1**	**34.8**	**18.6**
	Last	**92.1**	**89.0**	**69.6**	**47.3**	**70.5**	**50.9**	**32.0**	**18.1**
Co-teaching [4]	Best	82.6	73.0	24.0	14.6	50.5	38.2	11.8	4.9
	Last	81.9	72.6	23.5	11.7	50.3	38.0	11.3	4.3
SANM(Co-teaching)	Best	**89.2**	**78.2**	**36.4**	**20.7**	**58.2**	**51.3**	**19.4**	**13.4**
	Last	**88.6**	**76.7**	**35.2**	**18.4**	**56.9**	**50.1**	**17.9**	**12.7**
CDR [7]	Best	90.4	85.0	47.2	12.3	63.3	39.5	29.2	8.0
	Last	82.7	49.4	16.6	10.1	62.9	39.5	9.7	4.5
SANM(CDR)	Best	**92.6**	**91.6**	**55.3**	**16.7**	**72.7**	**56.4**	**36.6**	**20.8**
	Last	**91.8**	**90.8**	**48.6**	**15.5**	**71.2**	**53.2**	**30.0**	**19.7**
ELR+ [5]	Best	94.6	93.8	91.1	75.2	77.5	72.4	58.2	30.8
	Last	94.4	93.7	90.5	73.5	76.2	72.2	56.8	30.6
SANM(ELR+)	Best	**96.3**	**95.7**	**94.1**	**82.9**	**79.8**	**77.3**	**65.0**	**38.7**
	Last	**96.2**	**95.4**	**94.0**	**81.7**	**79.2**	**77.1**	**64.1**	**37.9**

Table 5. Comparison on CIFAR-10/100 with symmetric noise.

Dataset	CIFAR-10			
Method	20%	50%	80%	90%
SANM(DivideMix)	94.41±0.11	95.79±0.09	94.62±0.16	92.28±0.13
Dataset	CIFAR-100			
Method	20%	50%	80%	90%
SANM(DivideMix)	81.21±0.10	78.22±0.14	68.71±0.11	43.49±0.18

labels. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. 2

[4] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3

[5] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[6] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 5907–5915, 2019. 2

[7] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[8] Boshen Zhang, Yuxi Li, Yuanpeng Tu, Jinlong Peng, Yabiao Wang, Cunlin Wu, Yang Xiao, and Cairong Zhao. Learning from noisy labels with coarse-to-fine sample credibility modeling. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 21–38. Springer, 2023. 2

[9] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent

label noise: A progressive approach. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 2