
Supplementary Material for “Generative Noisy-Label Learning by Implicit Discriminative Approximation with Partial Label Prior”

Anonymous Author(s)

Affiliation

Address

email

1 Implementation details

In Tab. 1, we describe the implementation details of our method for each dataset, including the main reference for the respective baseline. In addition, for the Clothing1M, we sample 1000 mini-batches from the training set, where in every mini batch we ensure that the 14 classes are evenly sampled to form a pseudo-balanced learning problem. Also for Clothing1M, we first resize the image to 256×256 and then random crop to 224×224 with random horizontal flipping. The number of epochs for warmup is 10 for datasets containing 10 classes, and 15 for datasets containing 100 classes. For Clothing1M, we run 1 epoch of warmup.

	IDN CIFAR10/100	CIFAR10/100 N	Red Mini-ImageNet	Animal-10N	Clothing1M
Baseline reference	kMEIDTM [2]	Real-world [3]	FaMUS [4]	Nested [1]	CausalNL [5]
Backbone	ResNet-34	ResNet-34	Pre-act ResNet-18	VGG-19BN	ResNet-50 *
# Training epochs	150	120	150	100	40
Batch size	128	128	128	128	64
Learning rate	0.02	0.02	0.02	0.02	0.002
Weight decay	5e-4	5e-4	5e-4	5e-4	1e-3
LR decay at epochs	0.1/100	0.1/80	0.1/100	0.1/50	0.1/20
Data augmentation		Random Crop / Random horizontal flip			
β			0.9		
K			1		

Table 1: Implementation detail of our method in each dataset. *: Uses ImageNet pre-trained model.

2 Additional ablation study

In Tab. 2, we perform a hyper-parameter sensitivity test for our method on CIFAR10-IDN, including coverage and uncertainty for prior label construction. To test label coverage, we first examine the model performance as function of $\beta \in \{0.5, 0.7, 0.8, 0.9\}$ in Eq. 11, where the default value for $\beta = 0.9$. We observe that performance does not change much for $\beta \in \{0.7, 0.8, 0.9\}$, which indicates our model’s robustness with respect to that hyper-parameter. For $\beta = 0.5$, the performance drops significantly, indicating that using a moving average with a relatively high value for β is important for estimating model prediction and avoiding overfitting. We test $K \in \{1, 3\}$ in Eq. 13 by sampling multiple times $\{\hat{\mathbf{y}}_{i,j}\}_{j=1}^K \sim \text{Cat}(g_\theta(\mathbf{x}_i))$. We observe no significant changes to model performance with this higher value of K . Therefore, we choose $K = 1$ for simplicity.

We also test our model by shutting down the moving average, either by making $\beta = 0$ or by completely relying on the model’s current prediction, which is an experiment denoted by “arg max” in Tab. 2. Note that the model shows a major performance drop in both cases. This happens because the model overfits to the inaccurate model predictions, biasing the training procedure. Furthermore,

		CIFAR10			
		20%	30%	40%	50%
$\beta = 0.9, K = 1$		92.65	91.96	91.02	89.94
Hyper-parameter	$\beta = 0.8$	92.49	91.88	90.83	88.81
	$\beta = 0.7$	91.55	90.87	90.62	88.40
	$\beta = 0.5$	89.13	87.98	87.48	85.73
	$K = 3$	92.30	91.83	90.83	89.75
Coverage	$\beta = 0$	84.57	81.59	68.88	61.47
	arg max	20.19	18.56	16.09	15.26
	No Cov	85.57	81.00	72.42	66.61
Uncertainty	Uniform w	90.10	89.66	86.25	84.28
	No Unc	84.96	83.19	81.88	78.38

Table 2: Ablation study on hyper-parameter sensitivity, including β , K , coverage and uncertainty.

arg max performs the worst because inaccurate model prediction in prior label in the early stage, causing confirmation bias and leading to wrong optimisation goal. We also test our model without using the coverage term c_i in Eq. 10 – this experiment is denoted as “No Cov”. In this case, the performance of the model drops significantly for all noise rates, compared to the default model in the first row, which indicates the importance of having a coverage term in our prior label construction.

Furthermore, we study the uncertainty aspect for the prior label construction. We first experiment by setting w_i to a uniform value (“Uniform w ” row) instead of a GMM weight that represents the probability that the sample is carrying a clean label. The result is not competitive with the one that uses the GMM weight, which indicates the importance of having w representing a clean-label sample probability. We also test our model without the uncertainty component u_i in Eq. 10 (see row “No Unc”). This case shows a significant performance drop in all noise rates.

References

- [1] Y. Chen and et al. Boosting co-teaching with compression regularization for label noise. In *CVPR*, pages 2688–2692, 2021.
- [2] D. Cheng and et al. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16630–16639, 2022.
- [3] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [4] Y. Xu, L. Zhu, L. Jiang, and Y. Yang. Faster meta update strategy for noise-robust deep learning. In *CVPR*, pages 144–153, 2021.
- [5] Y. Yao, T. Liu, M. Gong, B. Han, G. Niu, and K. Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021.