EasyChair Preprint

№ 6763

# Behaviour of Sample Selection Techniques Under Explicit Regularization

Lakshya



October 3, 2021

# Behaviour of Sample Selection Techniques under explicit Regularization

Lakshya[1]

Samsung R&D Institute India - Bangalore, Bangalore, India
lakshya.01@samsung.com

**Abstract.** There is a multitude of sample selection-based learning strategies that have been developed for learning with noisy labels. However, It has also been indicated in the literature that perhaps early stopping is better than fully training the model for getting better performance. It leads us to wonder about the behavior of the sample selection strategies under explicit regularization. To this end, we considered four of the most fundamental sample selection-based models MentorNet, Coteaching, Coteaching-plus and JoCor. We provide empirical results of applying explicit L2 regularization to the above-mentioned approaches. We also compared the results with a baseline - a vanilla CNN model trained with just regularization. We show that under explicit regularization, the preconceived ranking of the approaches might change. We also show several instances where the baseline was able to outperform some or all of the existing approaches. Moreover, we show that under explicit regularization, the performance gap between the approaches can also reduce.

## 1 Introduction

Humans tend to learn much better and more quickly when presented with harder and harder concepts gradually. Yoshua Bengio formalized this notion as Curriculum learning [2]. Not only does Curriculum learning make the training process faster, but it also reaches superior quality minima in the case of non-convex optimization. Building on Curriculum Learning, Kumar [9] proposed Self-Paced Learning(SPL) for learning a latent variable model. Based on the findings of [1], according to which a neural network learns easy patterns first, MentorNet [8] made further progress along this line by using SPL for training with noisy labels.

Noisy labels are ubiquitous in practice. For example, noise may appear due to annotations carried out by computer programs on web crawled images [7, 22] or annotations based on crowdsourcing [27]. Consequently, it is necessary to research techniques that are robust to noisy labeling.

Since MentorNet, a multitude of sample selection-based techniques has emerged. Coteaching [5] upgraded the MentorNet by utilizing two Networks. The mini-batch used for training one network was decided by the loss obtained on the samples using the second network. Coteaching-plus [36], further argued that the two networks should be kept diverged by disagreement in predictions,

which can further benefit the training. JoCoR [30] aimed to reduce the diversity of the two models as opposed to Coteaching-plus. Recently DivideMix [11], EvidentialMix [25] were proposed, which try to incorporate semi-supervised learning on noisy classified labels as opposed to leaving them out of training. Another interesting approach is presented in [35], where a single model has been proposed for doing sample selection by relying on the consistency of predictions.

Meanwhile, it has also been argued that early-stopping might be a better strategy than fully training a network. Thus, in the presence of early stopping regularization, the benefits of MentorNet and other approaches remain unrealized. Although, finding the instance for early stopping or utilizing early stopping is an active area of research itself [14, 31]. This compelled us to wonder about the behavior of the sample selection strategies under explicit regularization. Since, for most of the approaches, either the original results have been provided without explicit regularization or even if regularization was present, less attention was paid to regularization while comparing results.

To provide more insights on these matters, we make the following contributions through this paper.

- We provide empirical results of applying explicit L2 regularization to the sample selection based approaches. We also compared the results with a baseline - a vanilla CNN model trained with just regularization.
- We show that under explicit regularization, the pre-conceived ranking of the approaches might change.
- We also show several instances where the baseline was able to outperform some or all of the existing approaches.
- Moreover, we show that under explicit regularization, the performance gap between the approaches can also reduce.

## 2 Related Works

Various methodologies have been developed to learn with noisy labels. There are invested efforts in exploiting a noise transition matrix [6, 12, 15, 32], using graph models [13, 33]. Progress has also been made using meta-learning [23, 26, 29, 34]. In [4], authors utilized different pseudo-labeling and sample selection strategies for Contrastive pre-training. In separate work, authors of [10] argue that even with overfitting to noise, good hidden representations are learned, which can be used to train a separate classifier with known correct labels. Authors of [19] learned a joint probability distribution for noisy and clean labels under the class-conditional noise process to identify the label errors in the dataset. Meanwhile, SELF [18] performs self ensembling to filter out the noisy label samples from the dataset, which are further used for unsupervised loss.

Authors have also tried developing robust surrogate loss functions that can help to learn in noisy labels setting [3, 17, 20, 37]. In particular, in [16], authors proposed a curriculum loss(CL) which is a tight upper bound on the 0-1 loss and can also be used to adaptively select samples. Whereas authors of [28] added

a reverse cross-entropy element with classical cross-entropy to create metric cross-entropy loss.

## 3 Experimentation

We considered four different sample selection algorithms and analyzed their results under explicit L2 regularization. We first conducted experiments to find the optimal weight decay value for each combination of algorithm, dataset, noise type, and noise rate. Next, we compared the results of the algorithms with their optimal weight decay values, which are provided in this section.

**Existing Approaches.** We used four different approaches for these experiments, Self-paced MentorNet, Coteaching, Coteaching-plus, and JoCor. We also considered a Baseline approach - a vanilla model trained only with weight decay.

**Datasets.** We used four different simulated noisy datasets for benchmarking, three vision-based datasets, MNIST, CIFAR-10, CIFAR-100, and one text-based dataset, NEWS. Although, we could only do the testing with JoCor on the CIFAR-10 and CIFAR-100 datasets, since we also had to find the optimal co-lambda [30] value for the experiments. We used three different simulated noise settings for our experiments. Namely, Symmetric noise [24] with 0.2 noise rate, Symmetric noise with 0.5 noise rate, and Pair-flipping [5] Noise with 0.45 noise rate.

**Hyperparameters.** Experiments were run for 200 epochs with three different seeds. All the other Hyperparameters, including warm-up schedule, were kept same as the original algorithm.

**Network architecture.** For all our experiments, we used the following models(similar to Coteaching-plus).

- MNIST-MLP for MNIST: a 2 layer MLP with ReLU activation
- CNN-small for CIFAR-10: A CNN model with 2 convolutional layers and 3 Dense layers with ReLU activation.
- CNN-large for CIFAR-100: A CNN model with 6 convolutional layers and 1 Dense layer with ReLU activation.
- NEWS-MLP for NEWS: a 3 layer MLP with Softsign activation function on top of pre-trained word embeddings from GloVe [21].

Table-1 shows the details of these networks(This table is motivated by Coteaching-plus [36]).

Table 1: Different architectures used.

| MNIST-MLP | CNN-small | CNN-large | NEWS-MLP |
|---|---|---|---|
| Dense 28x28 -> 256 | 5x5 Conv 6<br>2x2 Max-pool | 3x3 Conv 64, BN,<br>3x3 Conv 64, BN<br>2x2 Max-Pool | 300-D Embedding<br>Flatten => 1000x300<br>Adaptive avg-pool -> 16x300 |
| | 5x5 Conv 16<br>2x2 Max-pool | 3x3 Conv 128, BN<br>3x3 Conv 128, BN<br>2x2 Max-Pool | Dense 16x300 -> 4x300<br>BN, Softsign |
| | Dense 16x5x5 -> 120<br>Dense 120 ->84 | 3x3 Conv 196, BN<br>3x3 Conv 196, BN<br>2x2 Max-Pool | Dense 4x300 -> 300<br>BN, SoftSign |
| Dense 256 -> 10 | Dense84 -> 10/100 | Dense 256->10/100 | Dense 300 -> 7 |

Table 2: Average last ten epoch accuracy for different Models at their optimal weight decay values on the CIFAR-10

| dataset | type | rate | model | Test Accuracy | error(±) |
|---|---|---|---|---|---|
| CIFAR-10 | sym | 0.2 | jocor | 62.544 | 0.986 |
| | | | coteaching | 60.364 | 3.317 |
| | | | baseline | 59.187 | 4.667 |
| | | | mentornet | 58.392 | 4.107 |
| | | | coteaching-plus | 58.35 | 0.946 |
| | pairflip | 0.45 | baseline | 48.221 | 1.083 |
| | | | coteaching-plus | 39.766 | 0.356 |
| | | | mentornet | 39.666 | 1.137 |
| | | | coteaching | 38.753 | 3.941 |
| | | | jocor | 38.733 | 0.396 |
| | sym | 0.5 | jocor | 51.688 | 1.36 |
| | | | coteaching-plus | 49.881 | 0.789 |
| | | | baseline | 49.2 | 1.144 |
| | | | coteaching | 48.589 | 4.684 |
| | | | mentornet | 45.423 | 2.498 |

### 3.1 Observations

Tables-2, 3, 4, and 5 show the results of experimentation's on the CIFAR-10, CIFAR-100, MNIST and NEWS datasets, resp. Test accuracy's mentioned are average over last ten epoch across all the seeds . 'type' column refers to the noise type and 'rate' represents the noise rate. In these tables, for ease of analyzing,

Table 3: Average last ten epoch accuracy for different Models at their optimal weight decay values on the CIFAR-100 dataset.

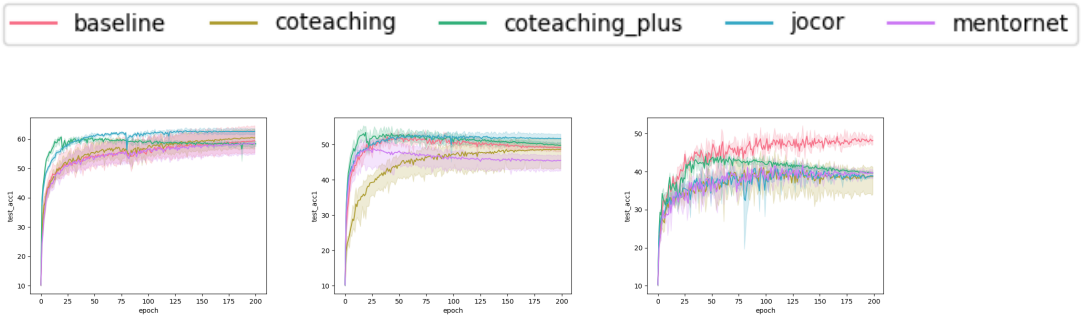
| dataset | type | rate | model | Test Accuracy | error(±) |
|---|---|---|---|---|---|
| CIFAR-100 | sym | 0.2 | jocor | 53.626 | 0.212 |
| | | | coteaching-plus | 49.332 | 0.32 |
| | | | coteaching | 47.812 | 0.537 |
| | | | mentornet | 47.437 | 0.527 |
| | | | baseline | 37.561 | 0.434 |
| | pairflip | 0.45 | coteaching-plus | 30.116 | 0.374 |
| | | | jocor | 29.562 | 0.351 |
| | | | coteaching | 28.811 | 0.155 |
| | | | mentornet | 27.333 | 0.42 |
| | | | baseline | 25.119 | 0.478 |
| | sym | 0.5 | jocor | 43.41 | 0.401 |
| | | | coteaching-plus | 40.445 | 0.429 |
| | | | coteaching | 38.384 | 0.271 |
| | | | mentornet | 37.507 | 0.485 |
| | | | baseline | 22.872 | 0.472 |

Table 4: Average last ten epoch accuracy for different Models at their optimal weight decay values on the MNIST dataset.

| dataset | type | rate | model | Test Accuracy | error(±) |
|---|---|---|---|---|---|
| MNIST | sym | 0.2 | coteaching-plus | 97.776 | 0.111 |
| | | | baseline | 97.52 | 0.119 |
| | | | mentornet | 97.496 | 0.062 |
| | | | coteaching | 97.49 | 0.062 |
| | pairflip | 0.45 | coteaching | 91.894 | 0.41 |
| | | | mentornet | 91.852 | 0.538 |
| | | | coteaching-plus | 86.403 | 4.367 |
| | | | baseline | 77.084 | 0.29 |
| | sym | 0.5 | coteaching | 96.311 | 0.104 |
| | | | mentornet | 96.265 | 0.098 |
| | | | coteaching-plus | 95.995 | 0.113 |
| | | | baseline | 95.799 | 0.062 |

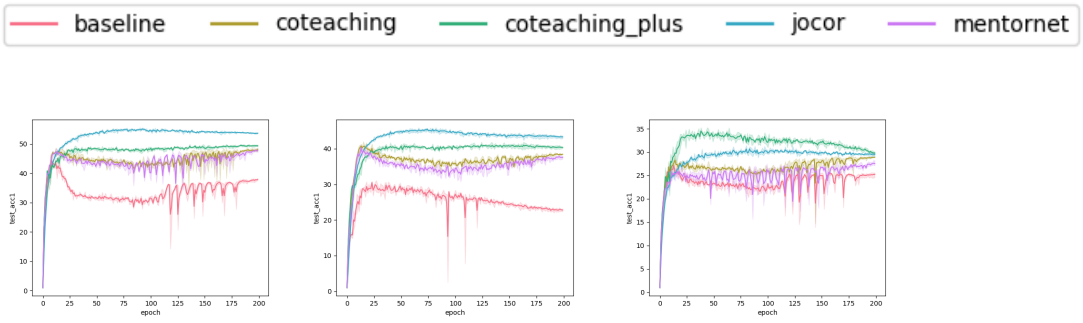
Table 5: Average last ten epoch accuracy for different Models at their optimal weight decay values on the NEWS dataset.

| dataset | type | rate | model | Test Accuracy | error(±) |
|---|---|---|---|---|---|
| NEWS | sym | 0.2 | coteaching-plus | 42.266 | 0.2 |
| | | | coteaching | 38.768 | 0.21 |
| | | | mentornet | 38.596 | 0.73 |
| | | | baseline | 36.558 | 0.383 |
| | pairflip | 0.45 | coteaching-plus | 30.195 | 0.611 |
| | | | mentornet | 29.669 | 0.322 |
| | | | coteaching | 29.054 | 0.425 |
| | | | baseline | 27.356 | 0.4 |
| | sym | 0.5 | coteaching-plus | 34.916 | 0.436 |
| | | | coteaching | 33.919 | 0.562 |
| | | | mentornet | 32.857 | 0.226 |
| | | | baseline | 26.217 | 0.521 |



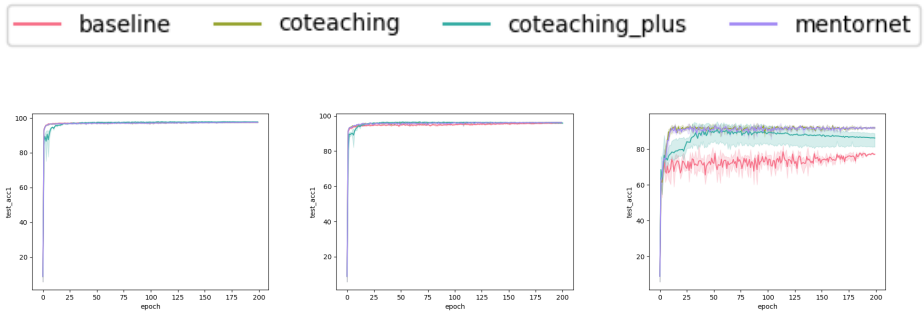

(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Fig. 1: Results on the CIFAR-10 dataset for different Models at their optimal weight decay values



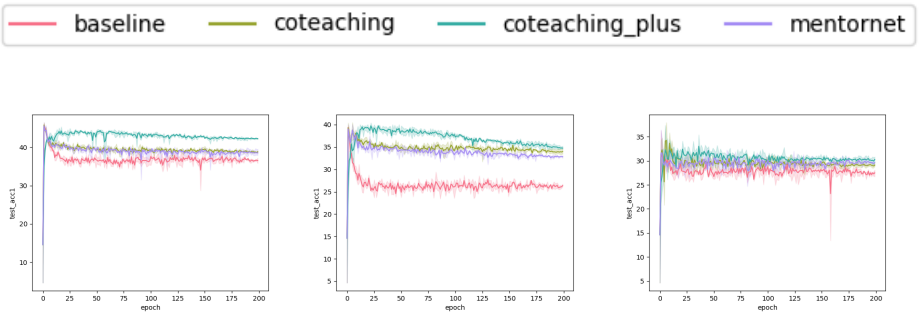

(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Fig. 2: Results on the CIFAR-100 dataset for different Models at their optimal weight decay values




(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Fig. 3: Results on the MNIST dataset for different Models at their optimal weight decay values




(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Fig. 4: Results on the NEWS dataset for different Models at their optimal weight decay values

entries for a particular combination of dataset, noise type and noise rate are sorted by the Test Accuracy. Meanwhile, Figures-1, 2, 3, and 4 show the test accuracy vs epoch plots during the training.

Before we analyze the results, we present a ranking order for existing approaches and the baseline. Based on the existing claims in the literature, we can assume the following ranking order, JoCor($1^{st}$) > Coteaching-plus($2^{nd}$) > Coteaching($3^{rd}$) > MentorNet($4^{th}$) > Baseline($5^{th}$), where '>' implies better in terms of test-accuracy. Moreover, it is expected that if explicit regularization doesn't have any relative effect on the approaches, then this pre-defined rank order should be maintained in our experiments as well.

Please note that there are 12 different groups of experiments, corresponding to the twelve different combinations of the dataset, noise type, and noise rate. Moreover, in each of these groups, a ranking order of the approaches can be observed(Each Table-2, 3, 4, and 5 show 3 groups based on noise-type and noise-rate for a particular dataset.). We have made the following observations by comparing these ranking orders with the pre-defined ranking order.

- In 8 out of the 12 groups, the ranking order was different than the pre-defined ranking order. This gives a clear indication that explicit regularization can indeed change the relative ranking of the approaches.
- Among all the combinations, Pairflip-0.45 proved to get most affected by regularization, where ranking order broke for every dataset value. Pairflip-0.45 is the toughest noise category as can be seen by lowest test accuracy for any approach-dataset pair. Thus, higher amount of overfitting to noisy labels happens in Pairflip-0.45 case, thus, the effect of L2-regularization is more profound.
- There were 4 different groups in which Baseline wasn't at the bottom of the ranking order. This includes the group CIFAR-10-Pairflip-0.45, where Baseline ranked $1^{st}$ with a difference of 8.455% between Baseline and the $2^{nd}$ ranked approach.
- We also observed that the performance gap between the MentorNet and the Coteaching was reduced significantly(please check the plots). On 3 different groups, MentorNet was even able to outperform Coteaching. Moreover, on the remaining 9 groups, the average performance difference between the Coteaching and the MentorNet was only 1.0212%.
- Following observations were made regarding the individual performance of each approach. (We denote the group as a failure if the ranking of the approach in the group was lower than the pre-defined ranking. Similarly, a win if it was higher than the pre-defined ranking.)
  - JoCor failed on 2(33.33%) groups out of 6(since we only experimented with JoCor on the CIFAR-10 and the CIFAR-100).
  - Coteaching-plus failed 3 times and won once, which implies the ranking order was changed 4(33.33%) times for it.
  - Coteaching failed 4 times and won 3 times i.e. 7(58.33%) times the ranking order was changed.
  - MentorNet failed once and won 4 times.

- Baseline won 4 times as well.

Based on this data, we can say that while the domination of one approach over the other might not be altered with explicit regularization (for instance, JoCor still ranks $1st$ in 4 out of 6 groups), in many instances, it can alter the outcome of the experiments and change the believed ranking of the approaches. Moreover, it can also reduce the performance gap between the algorithms as observed in the case of MentorNet and Coteaching.

## 4 Conclusion

In this paper, we showed that under explicit regularization, the pre-conceived ranking of the approaches might change. We also showed several instances where a vanilla CNN trained with just L2 regularization was able to outperform some or all of the existing approaches. Moreover, under explicit regularization, the performance gap between the approaches can also reduce. All these points suggest that special attention should be given to explicit regularization. Since explicit regularization can significantly alter the outcome, we suggest that it should be made sure that the comparison between the two approaches is done with their optimal regularization values.

## References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., Lacoste-Julien, S.: A closer look at memorization in deep networks (2017)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. p. 41–48. ICML '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1553374.1553380, https://doi.org/10.1145/1553374.1553380
3. Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., Liu, Y.: Learning with instance-dependent label noise: A sample sieve approach (2021)
4. Ciortan, M., Dupuis, R., Peel, T.: A framework using contrastive learning for classification with noisy labels (2021)
5. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels (2018)
6. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise (2019)
7. Hu, M., Yang, Y., Shen, F., Zhang, L., Shen, H.T., Li, X.: Robust web image annotation via exploring multi-facet and structural knowledge. IEEE Transactions on Image Processing **26**(10), 4871–4884 (2017). https://doi.org/10.1109/TIP.2017.2717185
8. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels (2018)

9. Kumar, M., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems. vol. 23. Curran Associates, Inc. (2010), https://proceedings.neurips.cc/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf
10. Li, J., Zhang, M., Xu, K., Dickerson, J.P., Ba, J.: Noisy labels can induce good representations (2020)
11. Li, J., Socher, R., Hoi, S.C.H.: Dividemix: Learning with noisy labels as semi-supervised learning (2020)
12. Li, X., Liu, T., Han, B., Niu, G., Sugiyama, M.: Provably end-to-end label-noise learning without anchor points (2021)
13. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation (2017)
14. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels (2020)
15. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(3), 447–461 (Mar 2016). https://doi.org/10.1109/tpami.2015.2456899, http://dx.doi.org/10.1109/TPAMI.2015.2456899
16. Lyu, Y., Tsang, I.W.: Curriculum loss: Robust learning and generalization against label corruption (2020)
17. Ma, X., Wang, Y., Houle, M.E., Zhou, S., Erfani, S.M., Xia, S.T., Wijewickrema, S., Bailey, J.: Dimensionality-driven learning with noisy labels (2018)
18. Nguyen, D.T., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Beggel, L., Brox, T.: Self: Learning to filter noisy labels with self-ensembling (2019)
19. Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: Estimating uncertainty in dataset labels (2021)
20. Patrini, G., Rozza, A., Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: a loss correction approach (2017)
21. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://www.aclweb.org/anthology/D14-1162
22. Ratner, A., Sa, C.D., Wu, S., Selsam, D., Ré, C.: Data programming: Creating large training sets, quickly (2017)
23. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning (2019)
24. van Rooyen, B., Menon, A.K., Williamson, R.C.: Learning with symmetric label noise: The importance of being unhinged (2015)
25. Sachdeva, R., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: Evidentialmix: Learning with combined open-set and closed-set noisy labels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3607–3615 (January 2021)
26. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting (2019)
27. Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing annotations for visual object detection pp. 40–46 (01 2012)
28. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

29. Wang, Z., Hu, G., Hu, Q.: Training noise-robust deep neural networks via meta-learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4523–4532 (2020). https://doi.org/10.1109/CVPR42600.2020.00458
30. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization (2020)
31. Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., Chang, Y.: Robust early-learning: Hindering the memorization of noisy labels. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=Eql5b1_hTE4
32. Xia, X., Liu, T., Han, B., Wang, N., Deng, J., Li, J., Mao, Y.: Extended t: Learning with mixed closed-set and open-set noisy labels (2020)
33. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2691–2699 (2015). https://doi.org/10.1109/CVPR.2015.7298885
34. Xu, Y., Zhu, L., Jiang, L., Yang, Y.: Faster meta update strategy for noise-robust deep learning (2021)
35. Yi, R., Huang, Y.: Transform consistency for learning with noisy labels (2021)
36. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I.W., Sugiyama, M.: How does disagreement help generalization against label corruption? (2019)
37. Ziyin, L., Chen, B., Wang, R., Liang, P.P., Salakhutdinov, R., Morency, L.P., Ueda, M.: Learning not to learn in the presence of noisy labels (2020)