Dropout can Simulate Exponential Number of Models for Sample Selection Techniques

Abstract

Following Coteaching, generally in the literature, two models are used in sample selection based approaches for training with noisy labels. Meanwhile, it is also well known that Dropout when present in a network trains an ensemble of sub-networks. We show how to leverage this property of Dropout to train an exponential number of shared models, by training a single model with Dropout. We show how we can modify existing two model-based sample selection methodologies to use an exponential number of shared models. Not only is it more convenient to use a single model with Dropout, but this approach also combines the natural benefits of Dropout with that of training an exponential number of models, leading to improved results.

1 Introduction

Noisy labels are ubiquitous in practice. For example, noise may appear due to disagreement in crowdsourcing based annotation, Su et al. (2012), or annotations carried out by computer programs on web crawled images, Hu et al. (2017); Ratner et al. (2017). Consequently, it is necessary to research techniques that are robust to noisy labeling. Multiple techniques have been developed to tackle this issue. Among them, sample selection is the one that we will focus on in this paper.

Sample selection can be regarded as a derivative of curriculum learning, Bengio et al. (2009). In sample selection techniques a curriculum is defined/learnt to select a subset of the data in each iteration of the training. MentorNet, Jiang et al. (2018), used a single network to select a mini-batch in each iteration. Self-paced MentorNet only considered samples with a loss lower than a certain threshold for training. Coteaching, Han et al. (2018), upgraded the MentorNet by utilizing two Networks, where the mini-batch used for training one network was decided by the loss obtained on the samples using the second network. Further, authors of Coteaching-plus , Yu et al. (2019), argued that there should be disagreement between the two networks, which can be beneficial for the learning. Multiple techniques have been proposed since then to perform sample selection based training. However, one general trend remains among these techniques, two networks are trained for sample selection Li et al. (2020b); Sachdeva et al. (2021); Feng et al. (2021); Wei et al. (2020a).

We argue that instead of two networks, an exponential number of shared models can be utilized for sample selection. This can be achieved by utilizing Dropout, Srivastava et al. (2014), in a single network. Dropout when present in a network trains an ensemble of sub-networks, thus, it can simulate an exponential number of shared models. We can utilize this property of Dropout to transform existing two-model based approaches to utilize an exponential number of shared models by training just a single model with Dropout. This approach can combine the natural benefits of Dropout, Srivastava et al. (2014); Gal and Ghahramani (2016) with the benefits of training an exponential number of models for performing sample selection, resulting in improved performance for the existing approaches when transformed to use Dropout.

Thus, our main contributions can be listed as follows

- We propose to replace the two model-based sample selection algorithms found in the literature with algorithms using an exponential number of shared models.
- We show how an exponential number of models can be utilized for sample selection using Dropouts in a single network.
- We provide empirical results by transforming existing approaches utilizing two models to use an exponential number of shared models with the help of Dropout. Our results suggest that such transformations lead to better results.

2 Related Works

Various methodologies have been developed to address the problem of learning with noisy labels. Significant efforts have been invested in exploiting a noise transition matrix, Liu and Tao (2016); Hendrycks et al. (2019); Xia et al. (2020); Li et al. (2021). Efforts have also been made using graphical models, Xiao et al. (2015); Li et al. (2017), and meta-learning, Ren et al. (2019); Xu et al. (2021); Shu et al. (2019); Wang et al. (2020).Ciortan et al. (2021) proposed using contrastive pre-training, by utilizing different pseudo-labeling and sample selection strategies, before training with a loss function. In a separate work, Li et al. (2020a) argued that even with poor generalization, good hidden representations can be learned by the model which can be used to train a separate classifier with known correct labels.Northcutt et al. (2021) tried to identify the label errors in the dataset by learning a joint probability distribution for noisy and clean labels under the class-conditional noise process. Meanwhile, SELF, Nguyen et al. (2019), performs self-ensemble to filter out noisy samples from the dataset, which are then used for unsupervised loss.

Authors have also tried developing robust surrogate loss functions that can boost learning in the noisy setting, Patrini et al. (2017); Ma et al. (2018); Cheng et al. (2021); Ziyin et al. (2020). In particular, Wang et al. (2019) added a reverse cross-entropy term with the classical cross-entropy to create a symmetric cross-entropy loss. Whereas, Lyu and Tsang (2020) proposed a curriculum loss(CL) which is a tight upper bound on the 0-1 loss. Moreover, they claimed that the CL can be used to adaptively select samples.

Another area of research is based on the early stopping criterion, for instance, Liu et al. (2020) proposed early learning regularization. Similarly, Xia et al. (2021) argued that the parameters of a model can be divided into critical and non-critical params, which can help reduce the side effects of early learning noisy labels before early stopping.

Arpit et al. (2017) suggested that a neural network learns easy patterns first. Multiple curriculum-based sample selection approaches have been proposed based on this observation. MentorNet by Jiang et al. (2018), Coteaching by Han et al. (2018) and Coteaching-plus by Yu et al. (2019). Similarly, JoCoR, Wei et al. (2020b), aims to reduce the diversity of the two models in contrast to Coteaching-plus. Recent upgrades to these models include DivideMix by Li et al. (2020b), EvidentialMix by Sachdeva et al. (2021). They try to utilize semi-supervised learning on noisy classified labels. Another interesting approach is presented by Yi and Huang (2021), where a single model has been proposed for doing sample selection based on consistency of predictions by the model. However, unlike our methodology of using dropout that reduce two model networks to a single one for any two-model based approach, they rely on consistency of predictions to perform sample selection as a new approach.

3 Using Dropout to Realize Exponential Number of Models

In this section, we describe a strategy that can be utilized to convert an existing two model-based approach to realize an exponential number of shared models.

First, we discuss the modification scheme for a model.

- Generally, ML models terminate with a stack of Dense Layers. Add a Dropout unit in front of the each Dense Layer. In case there is only one or no Dense layer, add the Dropout as the pen-ultimate layer(in our experience, it worked equally well).
- Increase the model width for all the layers by a factor of $\frac{1}{(1-p)}$, where p is the dropout probability. This step is intuitive since dropout reduces the expected width of a layer by a factor of $(1-p)$, which means that the effective width of a single network out of the exponential possibilities has reduced.

Let's refer to a model obtained by the above modifications as DropoutNet. While, let NetA and NetB refer to two unmodified models. Note that when we forward pass with DropoutNet, a different instance of DropoutNet is obtained, based on the retained units across all the dropout layers.

Given that, a sample selection learning algorithm(which uses NetA and NetB) can be modified as follows.

- In each training iteration of the unmodified algorithm, replace NetA and NetB with different instances of DropoutNet. To get the instance for NetA, pass the mini-batch meant for NetA through the DropoutNet, similarly for NetB.
- Only perform backward pass through the instance corresponding to NetA.

Since, during each training iteration, different instances of the DropoutNet acts as NetA and NetB, this strategy effectively converts an existing approach to utilize an exponential number of shared models corresponding to different instances of the DropoutNet. We didn't try to perform the backward pass twice to promote implementation simplicity as when the second instance of DropoutNet is sampled, the operation graph for

previous instance is removed from the memory. Moreover, since no backward pass is required for the first instance, we can direct the underlying implementation framework to not to keep track of the gradients.

We provide an example by modifying the Coteaching-plus algorithm using the above-mentioned steps. Let's call the resultant algorithm as Coteaching-plus-ours(in general we will use the suffix '-ours' to indicate an existing algorithm modified by our strategy). Algo-1 shows the corresponding algorithm of Coteaching-plus-ours(symbols have the same meaning as in Coteaching-plus). Figure-1 shows the flow difference between Coteaching-plus and Coteaching-plus-ours.

Algorithm 1: Coteaching-plus-ours

Input: DropoutNet, training set D, batch size B, learning rate η, estimate noise rate τ, epoch E_k and E_{max}

for $e = 1, .., E_{max}$ **do**

  1: Shuffle D into $\frac{|D|}{B}$ mini-batches ;

  for $n = 1, .., \frac{|D|}{B}$ **do**

    2: get the n_{th} mini-batch $\bar{D}$ from D ;

    3: $w^{(1)} \leftarrow$ DropoutNet instance by forward passing on $\bar{D}$;

    4: $w^{(2)} \leftarrow$ DropoutNet instance by re-forward passing on $\bar{D}$;

    5: select small-loss instance $\bar{D}'^{(2)}$ based on $w^{(2)}$; `// refer to original, Yu et al. (2019)`

    6: Update: $w^{(1)} = w^{(1)} - \eta \Delta \ell(\bar{D}'^{(2)}; w^{(1)})$; `// update Dropout instance` $w^{(1)}$

  end

  7: Update: $\lambda(e)$; `// refer to original, Yu et al. (2019)`

end

4 Experimentation

Existing approaches. We modified three different algorithms in our experiments, Coteaching-plus, JoCor, and DivideMix. We believe these algorithms form a good representative set of existing algorithms utilizing two model-based sample selection.

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. We used the same simulated noise for our experiments as done by the original approaches. Namely, Symmetric and Pair flipping or Asymmetric.

Hyperparameters. For all the experiments in this section, Dropout with $p = 0.7$ was used. Experiments were run for 200 epochs. All the other hyperparameters, including warm-up schedule, were kept same as the original algorithm.

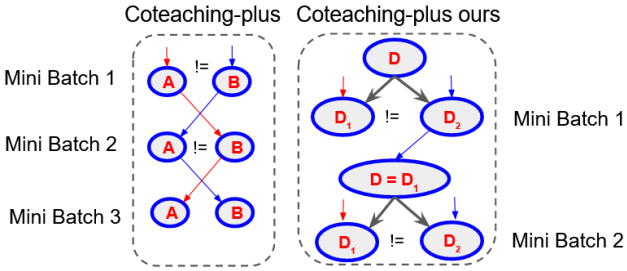


Figure 1: Flow difference between Coteaching-plus-ours and Coteaching-plus. In Coteaching-plus, A chooses the small loss-data for B to train on and vice-versa. However, in Coteaching-plus-ours, two instances of D are created based on the forward pass of the mini-batch. Then, D2 chooses the small-loss data for D1 on which back propagation is performed. The updated network acts as the new D.

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

- MNIST-MLP: a 2 layer MLP with ReLU activation.
- CNN-small: A CNN model with 2 convolutional layers and 3 Dense layers with ReLU activation.
- CNN-large: A CNN model with 6 convolutional layers and 1 Dense layer with ReLU activation.
- NEWS-MLP: a 3 layer MLP with Softsign activation function on top of pre-trained word embeddings from GloVe, Pennington et al. (2014).

Table-1 in Section-1 of supplementary materials shows the details of these networks(This table is motivated by Coteaching-plus). Table-2 in Section-1 of supplementary materials shows the details of these networks modified as per the strategy in Section-3, for a dropout with $p = 0.7$.

4.1 Results with Coteaching plus

Table-1 summarizes the the average last ten epoch test accuracy across five different seeds for these experiments. We used MNIST-MLP for the MNIST dataset, CNN-small for the CIFAR-10 dataset, CNN-large for the CIFAR-100 dataset, and MLP-NEWS for the NEWS dataset. Figure-2, Figure-3 show the corresponding test accuracy vs epoch plots for CIFAR-10 and CIFAR-100 dataset(additional figures are provided in Section-2.1 of supplementary materials). The

Table 1: Average last 10 epoch test accuracy for various algorithms

Algorithm	Dataset	Noise type	Noise rate	unmodified	ours
Coteaching-plus	MNIST	sym	0.2	**97.75 ± 0.09**	96.34 ± 0.14
		sym	0.5	**95.88 ± 0.20**	95.435 ± 0.19
		pairflip	0.45	78.09 ± 7.78	**87.06 ± 8.6**
	CIFAR-10	sym	0.2	58.07 ± 0.88	**58.38 ± 2.11**
		sym	0.5	49.22 ± 1.04	**53.87 ± 1.14**
		pairflip	0.45	38.13 ± 1.0	**46.72 ± 2.99**
	CIFAR-100	sym	0.2	49.22 ± 0.50	**53.47 ± 0.33**
		sym	0.5	40.06 ± 0.63	**42.72 ± 0.70**
		pairflip	0.45	28.78 ± 0.32	**37.13 ± 0.88**
	NEWS	sym	0.2	42.13 ± 0.51	**46.54 ± 0.38**
		sym	0.5	34.17 ± 0.65	**39.49 ± 0.43**
		pairflip	0.45	29.89 ± 0.46	**31.39 ± 0.66**
JoCor	CIFAR-10	sym	0.2	78.91 ± 0.19	**83.89 ± 0.36**
		sym	0.5	72.15 ± 0.58	**76.89 ± 0.232**
		pairflip	0.45	62.97 ± 0.486	**65.16± 1.53**
	CIFAR-100	sym	0.2	44.76 ± 0.54	**56.58 ± 0.21**
		sym	0.5	36.22 ± 0.84	**46.55 ± 0.20**
		pairflip	0.45	27.18 ± 0.66	**34.00 ± 0.31**
DivideMix	CIFAR-10	sym	0.2	84.11	**89.00**
		sym	0.5	88.82	**91.01**
		sym	0.7	87.41	**89.23**
	CIFAR-100	sym	0.2	63.21	**67.17**
		sym	0.5	59.49	**62.38**

Table 2: Average last 10 epoch test accuracy for Coteaching+Dropout and Dropout on the CIFAR-10 dataset

Noise type	Noise rate	Coteaching+Dropout	Dropout
sym	0.2	**58.38 ± 2.11**	51.41 ± 1.16
sym	0.5	**53.86 ± 1.14**	30.37 ± 0.46
pairflip	0.45	**46.72 ± 2.99**	40.19 ± 1.80

Table 3: Average last 10 epoch test accuracy for MentorNet-ours and Coteaching-plus-ours on the CIFAR-100 dataset

Noise type	Noise rate	Coteaching-plus-ours	MentorNet-ours
sym	0.2	**53.47 ± 0.33**	48.96 ± 0.51
sym	0.5	**42.72 ± 0.70**	38.36 ± 0.62
pairflip	0.45	**37.31 ± 0.88**	28.17 ± 0.67

plots compare the Coteaching(unmodified), Coteaching-plus(unmodified) and Coteaching-plus-ours(modified Coteaching-plus).

As can be seen in Table-1, except in MNIST-symmetric-0.2 and MNIST-symmetric-0.5, our approach could elevate Coteaching-plus on every other setting by values as high as 8.97%(MNIST-pairflip-0.45), 8.59%(CIFAR-10-pairflip-0.45), 8.35%(CIFAR-100-pairflip-0.45).

4.2 Results with JoCor

Similar to Coteaching, we calculated the average last ten epoch test accuracy across five different seeds. Table-1 summarizes the results for these experiments with CIFAR-10 and CIFAR-100. For these experiments, we used CNN-large for both CIFAR-10 and CIFAR-100 datasets. Figure-4 and 5 show the corresponding test accuracy vs epoch plots for CIFAR-10 and CIFAR-100 datasets.

Again, as can be seen in Table-1, our approach could elevate JoCor on every setting by values as high as 11.82%(CIFAR-100-symmetric-0.2).

4.3 Results with DivideMix

For these experiments, we used only a single seed to save computation. Table-1 summarizes the results for these experiments with CIFAR-10 and CIFAR-100. Similar to JoCor, we used CNN-large for both CIFAR-10 and CIFAR-100 datasets. Figure-6 and 7 show the corresponding test accuracy vs epoch plots for CIFAR-10 and CIFAR-100 datasets.

Again, as can be seen in Table-1, our approach could elevate DivideMix on every setting by a value as high as 4.89%(CIFAR-10-Symmetric-0.2).

5 Additional Experiments

5.1 Ablation Study: Dropout without Coteaching-plus

In these experiments, we trained a CNN-small, modified by the strategies in Section-3, without any sample selection and compared it to the same model trained by Coteaching-plus-ours. The former has been referred to as the Dropout model and later as the Coteaching+Dropout model. We used the CIFAR-10 dataset and three different seeds for these experiments. Dropout with $p = 0.7$ was used. Table-2 summarizes the average last ten epoch test accuracy across all the seeds while Figure-3 in Section-2.2 of supplementary material shows the plot of test-accuracy vs epoch.

By these experiments, we wanted to confirm that it is indeed the natural benefits of Dropout combined with the training algorithm that explains the test accuracy results and not just the Dropout regularization introduced to the model. From these results, it is clear that while the Dropout regularization is helpful, our approach is also able to harness the benefits of the original algorithm(Coteaching-plus).

5.2 MentorNet-ours vs Coteaching-plus-ours

In these experiments, we compared a modified CNN-large model, trained with self-paced MentorNet[1] with the same model trained with Coteaching-plus-ours. We refer to the former as MentorNet-ours. Experiments are done on the CIFAR-100 dataset for five different seeds. Again, we used a dropout with a $p = 0.7$ for these experiments. Table-3 summarizes the average last ten epoch test accuracy across all the seeds while Figure-4 in Section-2.3 of supplementary material shows the plot of test-accuracy vs epoch.

By the ablation study in Section-5.1, we were able to confirm that our strategy is able to draw out the benefits of the Curriculum learning. By the favorable results in this section, we can affirm that our approach can also harness the benefits of two-model based algorithms just with a single model. Since, if it couldn't the results would have been similar.

6 Discussion

A notable caveat of the proposed approach is that the width of the model has to be increased such that the expected size of each instance of the modified Dropout-based model(referred to as DropoutNet in Section-3) remains equivalent to the original model.

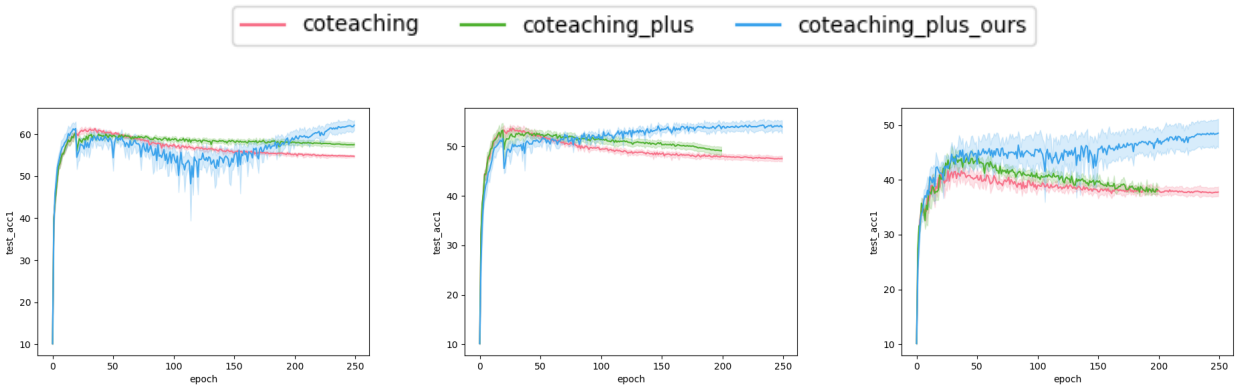
7 Conclusion

In this paper, we provided a strategy to elevate the existing two model-based sample selection algorithms to utilize an exponential number of shared models. We showed that this can be achieved by using a Dropout in a single model. We further provided empirical results by modifying Coteaching-plus, JoCor, and DivideMix that suggest such modification leads to better performance.

References

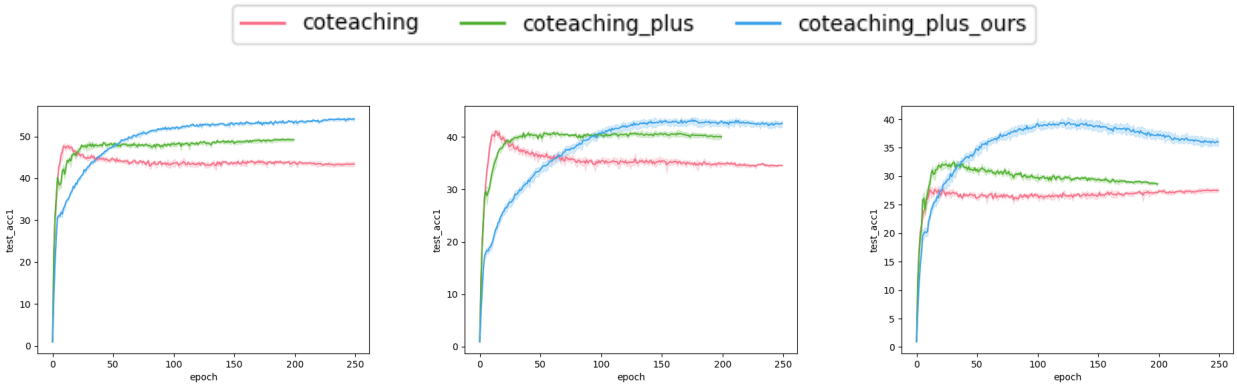
Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017.

[1]self-paced MentorNet algorithm uses a single model. Hence, modification by the strategy in Section-3 doesn't alter the algorithm.



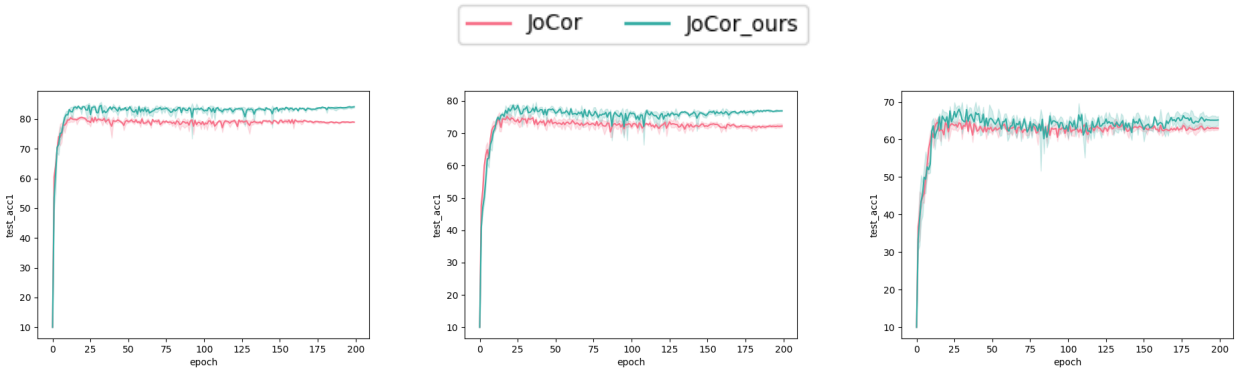
(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Figure 2: Results on CIFAR-10 for experimentation with Coteaching-plus



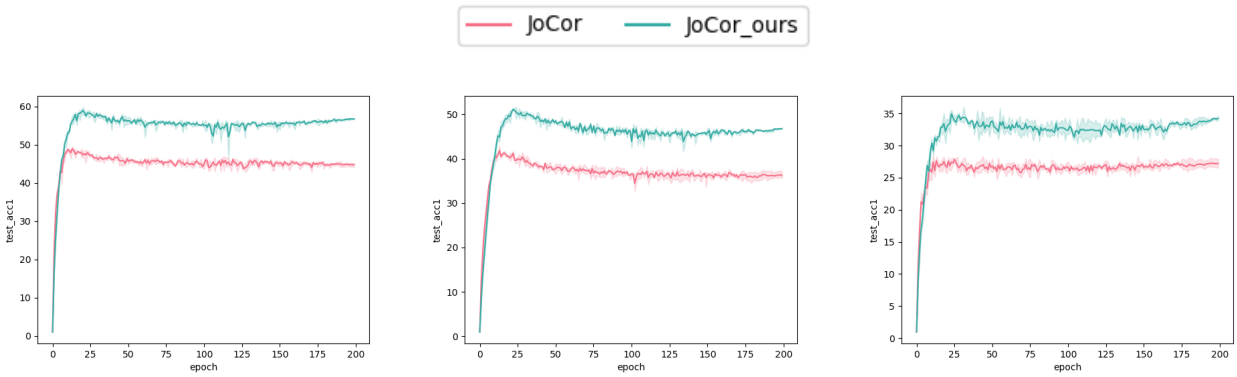
(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Figure 3: Results on CIFAR-100 for experimentation with Coteaching-plus



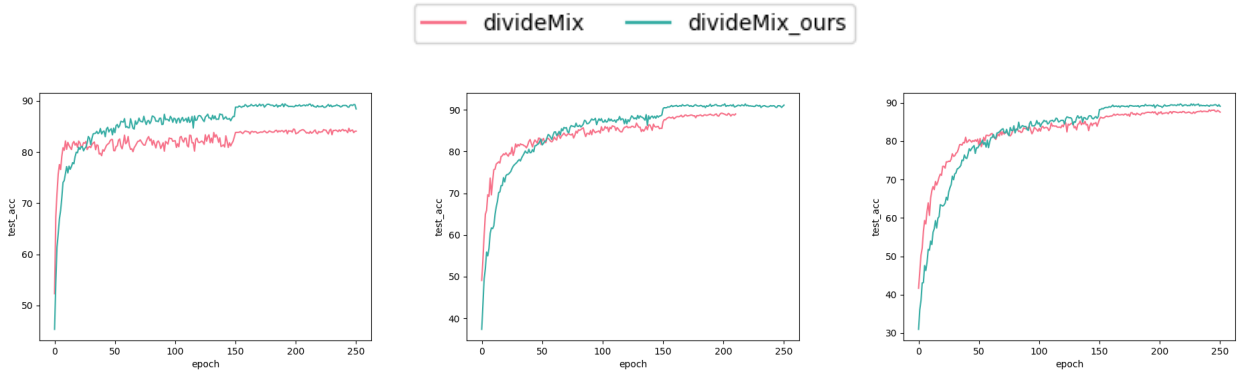
(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Figure 4: Results on CIFAR-10 for experimentation with JoCor



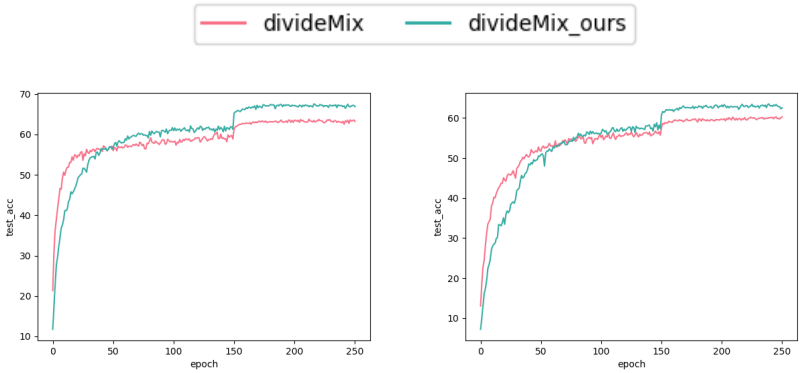
(a) 20% symmetric noise (b) 50% symmetric noise (c) 45% pairflip noise

Figure 5: Results on CIFAR-100 for experimentation with JoCor



(a) 20% symmetric noise (b) 50% symmetric noise (c) 70% symmetric noise

Figure 6: Results on CIFAR-10 for experimentation with DivideMix



(a) 20% symmetric noise (b) 50% symmetric noise

Figure 7: Results on CIFAR-100 for experimentation with DivideMix

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach, 2021.

Madalina Ciortan, Romain Dupuis, and Thomas Peel. A framework using contrastive learning for classification with noisy labels, 2021.

Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

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, 2018.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise, 2019.

Mengqiu Hu, Yang Yang, Fumin Shen, Luming Zhang, Heng Tao Shen, and Xuelong Li. Robust web image annotation via exploring multi-facet and structural knowledge. *IEEE Transactions on Image Processing*, 26(10):4871–4884, 2017. doi: 10.1109/TIP.2017.2717185.

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, 2018.

Jingling Li, Mozhi Zhang, Keyulu Xu, John P. Dickerson, and Jimmy Ba. Noisy labels can induce good representations, 2020a.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning, 2020b.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points, 2021.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation, 2017.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels, 2020.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, Mar 2016. ISSN 2160-9292. doi: 10.1109/tpami.2015.2456899. URL http://dx.doi.org/10.1109/TPAMI.2015.2456899.

Yueming Lyu and Ivor W. Tsang. Curriculum loss: Robust learning and generalization against label corruption, 2020.

Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels, 2018.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling, 2019.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels, 2021.

Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach, 2017.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly, 2017.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning, 2019.

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 (WACV)*, pages 3607–3615, January 2021.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Hao Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. pages 40–46, 01 2012.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zhen Wang, Guosheng Hu, and Qinghua Hu. Training noise-robust deep neural networks via meta-learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4523–4532, 2020. doi: 10.1109/CVPR42600.2020.00458.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization, 2020b.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Jiankang Deng, Jiatong Li, and Yinian Mao. Extended t: Learning with mixed closed-set and open-set noisy labels, 2020.

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 *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Eql5b1_hTE4.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2699, 2015. doi: 10.1109/CVPR.2015.7298885.

Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning, 2021.

Rumeng Yi and Yaping Huang. Transform consistency for learning with noisy labels, 2021.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption?, 2019.

Liu Ziyin, Blair Chen, Ru Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Learning not to learn in the presence of noisy labels, 2020.