Deep Learning from Noisy Labels with Some Adjustments of a Recent Method

István Fazekas[1], László Fórián[2], and Attila Barta[2]

Abstract—**In this paper we have used JoCoR, a fairly recent method for learning with label noise, that makes use of two neural networks with a joint loss function using an additional contrastive loss to increase the agreement between them. This method can be extended to more than two networks in a straightforward way. We have carried out experiments on the CIFAR-10 and CIFAR-100 datasets (contaminated by synthetic label noise) with this kind of extension using several contrastive losses. We have concluded that it makes a significant improvement if we use a third network, especially when we use Kullback-Leibler terms for all possible pairs of softmax outputs. Further extension also means some kind of improvement, but in the case of the CIFAR datasets, those were not so significant, maybe except the cases with lower ratio of label noise.**

Index Terms—**Deep Learning, Noisy Labels, Classification, Neural Networks, Supervised Learning**

I. Introduction

DEEP neural networks have excellent performance in image classifications tasks, but they are in need of large sets of training data with correct labels. This is a drawback, since labeling is difficult or too expensive in many cases. The available datasets are often contaminated by label noise, that is why the challenge of learning with noisy labels has become an important research topic with several directions [1], [5]. Even though deep neural networks tend to learn the simple, consistent patterns first, they can easily overfit to noisy labels [2]. If we are able to prevent this overfitting and treat the label noise during the training process, we can obtain models with good generalization ability.

In this work, we have investigated the possibilities of the improvement of a recent method in the topic of learning with label noise. We have applied some modifications to the training process, evaluated those adjusted models and drawn conclusions from the results.

JoCoR [6] is one of the recent state-of-the-art techniques for learning with label noise. It uses the idea of the selection of small-loss samples along with the utilization of two neural networks and it gradually increases the agreement between them. This model is trained with two classifiers in the background and a joint loss function which contains an additional term to reduce the divergence of the two networks, they are forced to make similar predictions. This scheme has a regularization effect during the training, it plays an important role in preventing overfitting. The parameters of the networks are updated simultaneously by the joint loss function, which is a weighted sum of the supervised losses and the contrastive loss term. JoCoR shows very impressive performance on several datasets with label noise, including CIFAR-10 and CIFAR-100 with symmetric and asymmetric label noise.

The method of JoCoR can be considered as a special ensemble of the two classifiers. Unlike the techniques using a disagreement strategy ([3], [4], [7]), JoCoR can be naturally extended to more than two networks. This raises the question: is it worth to use JoCoR with three neural networks if we have the computational capacity?

One of our results is that the answer for the above question is yes; we were able to make a significant improvement in the considered symmetric and asymmetric noise cases on CIFAR-10 and CIFAR-100 using three networks and totally six Kullback-Leibler terms (for every possible pair of softmax outputs). Similar results were obtained by using only three KL terms in a circular manner, but the improvement of the model over the training process was slightly slower and the test performance seemed to have a larger variance. Cross-Entropy contrastive losses were also applied, however they led to moderately weaker performance with larger variance as well.

We have also experimented with the utilization of more networks. They made a slight improvement, too, but the further increase comes with the cost of larger computational needs and the benefits are not as significant as in the case of three networks. However, the improvement is relatively larger when we have a lower amount of label noise.

II. CIFAR-10 and CIFAR-100 with Synthetic Label Noise

The dataset CIFAR-10 consists of images from 10 classes with 32 × 32 RGB pixels. The size of the training set is 50000 examples and the test set has 10000 samples. For CIFAR-100, the size and quantity of the images are the same, but the number of classes equals 100. We also have 20 superclasses, each of them contains 5 classes.

The CIFAR-10 and CIFAR-100 datasets are used with synthetic label noise of two types: symmetric and asymmetric. We have experimented with two types of synthetic label noise:

- **Symmetric label noise**: a given proportion of the labels is flipped to one of the other classes according to a discrete uniform distribution.
- **Asymmetric label noise**: it is generated by taking pairs of classes (which are similar to each other, for which humans make some mistakes, too), and a proportion of the data labels are flipped between these class pairs.

[1] University of Debrecen, Faculty of Informatics, Debrecen, Hungary (E-mail: fazekas.istvan@inf.unideb.hu)

[2] University of Debrecen, Faculty of Informatics and Doctoral School of Informatics, Debrecen, Hungary (E-mail: forian.laszlo@inf.unideb.hu, barta.attila@inf.unideb.hu)

DOI: 10.36244/ICJ.2023.5.2

III. A Recent Method for Learning with Noisy Labels: JoCoR

JoCoR utilizes the idea of small-loss selection and uses two neural networks. The agreement between those networks is gradually increased during the training process, this was inspired by some semi-supervised learning methods. The model is trained using two classifiers and a joint loss function which contains an additional regularization term to reduce the divergence of the two networks, so they are forced to agree with each other. This setup also has a regularization effect during the training, and it helps to prevent overfitting, too.

JoCoR uses convolutional neural networks (CNNs) with several convolutional and batch-normalization layers in the background, but it can be changed to any other neural network. This backbone CNN can be seen on Fig. 1, the source is [6].

CNN on *CIFAR-10*& *CIFAR-100*
32×32 RGB Image
3×3, 64 BN, ReLU 3×3, 64 BN, ReLU 2×2 Max-pool
3×3, 128 BN, ReLU 3×3, 128 BN, ReLU 2×2 Max-pool
3×3, 196 BN, ReLU 3×3, 196 BN, ReLU 2×2 Max-pool
Dense 256 $\longrightarrow$ 100

Fig. 1. The network in the background of JoCoR

The loss function of JoCoR is a weighted sum of the supervised loss of the two networks (two Cross-Entropy terms) and a contrastive loss term. The latter quantity is a symmetric Kullback-Leibler divergence (the sum of two KL terms). Here the dataset is given with N samples from M classes as $D = \{\mathbf{x_i}, y_i\}_{i=1}^{N}$. $\mathbf{x}_i$ is the i-th instance with its observed label $y_i \in \{1, \ldots, M\}$. The formulas:

$$L(\mathbf{x}_i) = (1-\lambda) * L_{sup}(\mathbf{x_i}, y_i) + \lambda * L_{con}(\mathbf{x}_i),$$

$$L_{sup}(\mathbf{x_i}, y_i) = L_{C_1}(\mathbf{x_i}, y_i) + L_{C_2}(\mathbf{x_i}, y_i),$$

$$L_{con} = D_{KL}(p||q) + D_{KL}(q||p),$$

if p and q are the two discrete probability distributions obtained from the softmax outputs.

Images considered as clean are selected with the small loss criterion using this joint loss function. At the start, the whole training dataset is used, then fewer training examples are selected in the upcoming epochs until it gradually reaches the ratio $1-\tau$, where τ is the known or estimated ratio of the noisy labels in the training dataset.

IV. Extending the Method to Three Networks

Since JoCoR can be considered as a special ensemble of the two classifiers and it can be extended to more than two networks in a natural way, we wanted to investigate whether it is worth to use three neural networks instead of the original two.

We have investigated the performance of JoCoR with three networks (we have used copies of the same CNN as JoCoR) and several types of contrastive loss (where p_1, p_2, p_3 are the softmax outputs). We have carried out experiments with the following contrastive loss setups:

Totally six KL-terms (for every possible pair of softmax outputs):

$$L_{con}(\mathbf{x}_i) = D_{KL}(p_1||p_2) + D_{KL}(p_2||p_1) + D_{KL}(p_1||p_3) + D_{KL}(p_3||p_1) + D_{KL}(p_2||p_3) + D_{KL}(p_3||p_2).$$

This can be considered the extension of JoCoR's contrastive loss to three networks and the force is quite strong for the classifiers to predict similarly.

Using three KL-terms in a circular manner:

$$L_{con} = D_{KL}(p_1||p_2) + D_{KL}(p_2||p_3) + D_{KL}(p_3||p_1).$$

This function came into consideration because the effect of using these three terms only, may lead to the same situation in the long run: the predictions of the classifiers should be quite similar at the end of the training process.

Three Cross-Entropy terms:

$$L_{con} = L_{CE}(p_1||p_2) + L_{CE}(p_1||p_3) + L_{CE}(p_2||p_3).$$

This loss function came up as an idea since the Cross-Entropy can also be considered as a distance between two discrete probability distributions.

A. Results on CIFAR-10

Table I contains our results on CIFAR-10 (and the results of the original JoCoR). We have implemented our experiments using PyTorch as the authors of JoCoR. We have used the Adam optimizer with momentum 0.8. The initial learning rate was 0.001 and the mini-batch size was set to 128. The number of epochs was 200 and the learning rate has started to decrease from the 80-th epoch and it was linearly decreased to 0 until the end of the training. The parameter λ was set to 0.5 in the case of 6 Kullback-Leibler terms and 0.7 for the setups with 3 Kullback-Leibler and 3 Cross-Entropy terms.

The models were evaluated on the 10000-element test dataset and these values are the averages (and standard deviations) of test accuracies of 10 inependent runs. We can see that it is worth using 3 networks with 6 Kullback-Leibler divergence terms, because we got higher accuracies and lower standard deviations. If we have the capacity, it may also be worth using the 3 Kullback-Leibler version: the results are similar, just the standard deviatons are slightly higher. We are also able to improve JoCoR's results with Cross-Entropy regularization terms, but the difference is smaller in that case.

TABLE I
THE CIFAR-10 AND CIFAR-100 TEST PERFORMANCE OF JOCOR WITH 2 NETWORKS AND WITH 3 NETWORKS USING THE MODIFIED CONTRASTIVE LOSSES

CIFAR-10				
	JoCoR	6 Kullback-Leibler	3 Kullback-Leibler	3 Cross-Entropy
Symm. 20%	85.73 ± 0.19%	86.95 ± 0.19%	86.75 ± 0.19%	85.90 ± 0.27%
Symm. 40%	79.41 ± 0.25%	80.49 ± 0.21%	80.46 ± 0.33%	79.96 ± 0.29%
Symm. 80%	27.78 ± 3.16%	29.02 ± 3.11%	28.10 ± 3.19%	28.02 ± 3.39%
Asymm. 40%	76.36 ± 0.49%	77.27 ± 0.41%	77.43 ± 0.57%	77.28 ± 0.57%
CIFAR-100				
	JoCoR	6 Kullback-Leibler	3 Kullback-Leibler	3 Cross-Entropy
Symm. 20%	53.01 ± 0.44%	54.15 ± 0.39%	54.13 ± 0.43%	53.18 ± 0.51%
Symm. 40%	43.49 ± 0.46%	44.01 ± 0.37%	44.01 ± 0.40%	43.51 ± 0.54%
Symm. 80%	15.49 ± 0.98%	16.23 ± 0.91%	16.15 ± 0.96%	16.07 ± 1.05%
Asymm. 40%	32.70 ± 0.35%	33.35 ± 0.27%	33.34 ± 0.33%	32.71 ± 0.37%

B. Results on CIFAR-100

Our results using three networks on CIFAR-100 are also summarized in Table I. JoCoR's original results are also present for comparison. The value of the hyper-parameters were the same as for the models using CIFAR-10 previously.

The models were evaluated on the 10000-element CIFAR-100 test dataset and these values are the averages (and standard deviations) of test accuracies of 10 runs. The 6 Kullback-Leibler version makes a significant improvement again, but the circular 3-term KL contrastive loss is not far behind, especially in the cases with lower noise ratio. The Cross-Entropy-type loss also means an improvement, but it results in slightly higher standard deviation and it seems to be more suitable for symmetric noise than asymmetric.

V. FURTHER INCREASE OF THE NUMBER OF NETWORKS

We have also investigated the possibilities of improvement by using more than 3 networks and contrastive losses with Kullback-Leibler divergence for all possible pairs of the softmax outputs. In table II we only report the averages of test accuracies of 10 runs. The parameters of the training were the same as before except the value of λ. It was set to 0.3 in the case of the 4-network model. Its value was 0.2 and 0.1 for 5 and 6 classifiers, respectively.

We can see that we were able to make a significant improvement with the third network, and the fourth classifier also makes an improvement, but the difference is not so large. The fifth network could also improve our results for smaller noise ratios, but these accuracies seem to be almost constant for larger models. It is also important to note that there was no significant decrease despite the complexity of computations.

Table II contains the results for more than 3 networks and Kullback-Leibler terms for all possible softmax output pairs, using the CIFAR-100 test dataset. We have used the same set of hyper-parameters that were used for CIFAR-10.

We can observe similar results on the CIFAR-100 dataset, too. The third network gave a significant improvement, but for more than 3 networks, the upgrade was not so significant (maybe except for the 4-part version with smaller amount of noise). The difference between the performance of the models is generally smaller for this dataset since this classification is a more difficult task.

The increase of the computational needs, execution costs of the larger models are approximately linear, so is has to be taken it into consideration when trying to increase the performance of the model.

VI. SOME CONSIDERATIONS ON THE λ HYPER-PARAMETER

The λ hyper-parameter is the weight of the contrastive loss in the overall loss function, hence it controls the force that pulls the predictions together. It also provides the regularization effect in our models. The larger the λ is, the less the divergence of the softmax outputs of the networks. However, if we set it too high, the classifiers make almost the same predictions which is not favourable, especially if we have several neural nets. The best λ depends on the dataset and the model as well. When obtaining the results in our previous tables, we have used the most suitable λ values out of $0.1, 0.2, \ldots 0.9$. As an illustration on CIFAR-10, we present the case of the model with 3 networks and 6 Kullback-Leibler terms and 20% symmetric label noise in Table III. That table also contains the average of the test performance of 10 runs.

VII. CONCLUSIONS

We have carried out experiments with JoCoR, a recent technique for learning with label noise, which can be naturally extended to more than the two networks it originally uses.

We were able to make a significant improvement in accuracy by utilizing a third network with a 6-term Kullback-Leibler contrastive loss. Despite the other types of used regularization losses had some drawbacks, they could also improve the original JoCoR model, especially in the scenarios with lower proportion of label noise.

If we use a contrastive loss with all the possible pairs of softmax outputs, we can further improve our results by increasing the number of neural networks on both CIFAR datasets, but those improvements can be considered not as significant as the case of the 3-classifier model.

ACKNOWLEDGEMENT

This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

TABLE II
THE CIFAR-10 AND CIFAR-100 TEST PERFORMANCE OF JOCOR WITH 2-6 NETWORKS (AVERAGE OF 10 INDEPENDENT RUNS)

CIFAR-10					
	JoCoR	3 networks	4 networks	5 networks	6 networks
Symm. 20%	85.73%	86.95%	87.19%	87.26%	87.29%
Symm. 40%	79.41%	80.49%	80.80%	80.82%	80.86%
Symm. 80%	27.78%	29.02%	29.11%	29.11%	29.12%
Asymm. 40%	76.36%	77.27%	77.49%	77.49%	77.48%
	JoCoR	3 networks	4 networks	5 networks	6 networks
Symm. 20%	53.01%	54.15%	54.34%	54.38%	54.41%
Symm. 40%	43.49%	44.01%	44.15%	44.18%	44.20%
Symm. 80%	15.49%	16.23%	16.26%	16.29%	16.30%
Asymm. 40%	32.70%	33.35%	33.42%	33.44%	33.44%

TABLE III
THE CIFAR-10 TEST PERFORMANCE OF JOCOR WITH 3 NETWORKS AND 6 KULLBACK-LEIBLER TERMS FOR DIFFERENT λ VALUES, IN THE CASE OF 20% SYMMETRIC LABEL NOISE

λ	Performance
0.1	75.92%
0.2	76.21%
0.3	81.84%
0.4	84.97%
0.5	86.95%
0.6	86.43%
0.7	85.75%
0.8	84.24%
0.9	83.56%

REFERENCES

[1] G. Algan and I. Ulusoy, Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey, *Knowledge-Based Systems* 215, 106771, 2021, **DOI**: 10.1016/j.knosys.2021.106771

[2] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio and S. Lacoste-Julien, A Closer Look at Memorization in Deep Networks *Proceedings of Machine Learning Research* vol. 70, 2017, pp. 233–242.

[3] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang and M. Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, *NeurIPS*, 2018, 8535–8545.

[4] E. Malach and S. Shalev-Shwartz, Decoupling "when to update" from "how to update", Advances in *Neural Information Processing Systems*, 2017, pp. 960–970.

[5] H. Song, M. Kim, D. Park, Y. Shin and J. Lee, Learning from Noisy Labels with Deep Neural Networks: A Survey *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] H. Wei, L. Feng, X. Chen and B. An, Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization, *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2020, 13723–13732, **DOI**: 10.1109/CVPR42600.2020.01374

[7] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang and M. Sugiyama, How does Disagreement Help Generalization against Label Corruption?, *International Conference on Machine Learning*, 2019, 7164–7173.

István Fazekas graduated from Kossuth Lajos University, Debrecen, Hungary in 1978. He is currently a full professor at Faculty of Informatics, University of Debrecen, Hungary. He had been head of the Department of Applied Mathematics and Probability Theory. His main research interests are asymptotic theorems of probability theory and mathematical statistics, network theory and machine learning.

László Fórián graduated from University of Debrecen in 2019 as a mathematician. He is currently a PhD student at the Doctoral School of Informatics and an assistant lecturer at the University of Debrecen, Hungary. His main research areas consist of neural networks and random graphs.

Attila Barta is an assistant lecturer at University of Debrecen, Hungary. He holds an MSc degree in applied mathematics from the University of Debrecen and he is also a PhD candidate at the Doctoral School of Informatics in the institution. His main research fields are network science and neural networks.