Uncertainty Based Detection and Relabeling of Noisy Image Labels

Jan M. Köhler
BCAI

Maximilian Autenrieth
BCAI, University of Ulm

William H. Beluch
BCAI*

Abstract

Deep neural networks (DNNs) are powerful tools in computer vision tasks. However, in many realistic scenarios label noise is prevalent in the training images, and overfitting to these noisy labels can significantly harm the generalization performance of DNNs. We propose a novel technique to identify data with noisy labels based on the different distributions of the predictive uncertainties from a DNN over the clean and noisy data. Additionally, the behavior of the uncertainty over the course of training helps to identify the network weights which best can be used to relabel the noisy labels. Data with noisy labels can therefore be cleaned in an iterative process. Our proposed method can be easily implemented, and shows promising performance on the task of noisy label detection on CIFAR-10 and CIFAR-100.

1. Introduction

In the last decade Deep neural networks (DNNs) have proven their predictive power in many supervised learning tasks with complex data patterns. Especially when the training set is large, representative, and correctly labeled, DNNs are the current state-of-the-art on several learning tasks. Unfortunately, the latter assumption does not hold in many realistic cases (e.g. medical imaging, crowd-sourced labeling), and DNNs have been shown to overfit on noisy labels, leading to poor generalization performance. For example, [33] shows that DNNs can easily fit randomly assigned labels on the training set, which leads to poor test performance. Therefore, it is important to detect and correct for noisy labels in the training set.

We propose an iterative label noise filtering process, based on the predictive uncertainty of the training images. Ensembles [17] and MC dropout [8] are used to obtain uncertainty estimates for each image. We show that the uncertainties of the noisy images and the uncertainties of the clean images follow two different distributions, enabling the detection of potentially noisy labels. After the detection of the noisy labels, the detected image could be taken out of the training set, its weight on the loss could be decreased, or it could be relabeled through an oracle or any appropriate relabeling approach.

*Bosch Center for Artificial Intelligence. First name.last name@de.bosch.com

2. Related Work

In the literature various approaches have been proposed to deal with label noise. [26] and [15] utilize sample weights, derived from the performance of the network on a validation set, to reduce the influence of noisy labels. Other approaches exploit additional networks to assign sample weights by learning the structure of the label noise [30, 16, 11]. [2, 23, 13, 34, 22, 25] propose adjusted loss functions to diminish the influence of noisy labels during training. [4] excludes potential noisy labels from training, with the disadvantage that information from the data is thrown away. Further approaches have recently been proposed to tackle the issue of label noise in image classification tasks, [28, 31, 10, 29, 20, 18, 32, 21]; however, to the best of our knowledge, none of the proposed methods utilize model uncertainty to detect and filter out label noise in image classification tasks. [1] concurrently explore a very similar method to the one proposed in this paper by using a mixture of beta distributions over the training loss of noisy vs clean images.

3. Methodology

In this section we explain our iterative, uncertainty-based, noise filtering process. In Section 4 the proposed method is then evaluated on two different noise patterns: Symmetric and Pair noise. In the former, k$\%$ of the training labels are randomly flipped to another label ($k \in [20, 40, 60]$). In the latter, $k\%$ of the labels are systematically flipped to the subsequent class label. Both noise patterns are very realistic in image classification tasks and currently discussed in the literature [11]. The Pair setting is generally more difficult, especially when k is greater than 50%, as for a given real class, the class with the majority of labels is not the real class (e.g. a majority of the images with real label 'dog' are labeled cat).

Uncertainty acquisition Let $y = g^W(x)$ be the output of a neural network with weights W and input x, and $u = h(y|g^W, x)$ be the uncertainty of the model for its prediction y given the input data x and the model g^W. Since it has been shown that a single softmax score of one classifier does not serve well as an uncertainty measure [8, 7, 12], we use three recent methods to easily obtain uncertainty estimates: Deep Ensembles [17] with M members, Monte-Carlo dropout (MC-dropout [7]), using T forward passes, and a combination of both [27].

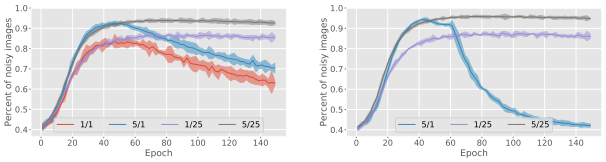
Having the predictions $y = g_m^{W_t}(x)$, with $t = 1, \cdots, T$ forward passes and $m = 1, \cdots, M$ classifiers, one needs a statistic $u(y)$ to quantify the uncertainty. The goal is to find an uncertainty statistic over the predictions $g_m^{W_t}(x)$ which depicts a high uncertainty value if x has a noisy label and a low value if x has a clean label. Many such statistics exist which have been successfully used in different settings [3, 27, 6, 9, 19, 24].

We compare different statistics over the predictions $y = g_m^{W_t}(x)$ given by the multiple forward passes, including BALD [14], Variation-Ratio, the standard deviation over the predictions $std(g_m^t(x))$ (averaged over all classes), and the mean over the predictions. For the mean we take the most likely of the k classes of the softmax vector, *i.e.* for brevity we denote $g_m^t(x) = \max_k g_m^{W_t}(x)$. [12] show that this maximum softmax probability is useful to distinguish between correctly and wrongly classified images.

Noisy label detection We investigate the ability of the aforementioned uncertainty metrics to distinguish between noisy and clean labels. Our goal is to identify an epoch $T1$ in which the uncertainty of the noisy labels is significantly higher compared to the uncertainty of the clean labels. Fig 1 exhibits the ability of a given model type and a given uncertainty measure to detect noisy labels. At each training epoch, the training images with the highest $p\%$ uncertainty are selected, i.e. $X_p := \{x : g_m^{W_t}(x) \geq x_p\}$. The proportion of images in this subset $(p = 0.9)$ with a noisy label is plotted on the y-axis. Note that this value should be scaled against the baseline noise level, i.e. if the baseline noise level is 40%, the baseline ratio given by random sampling would be 40%.

For both the Symmetric and Pair setting on CIFAR-10, the averaged softmax value of the predicted class and the Variation Ratio result in the highest selectivity (other uncertainty measures not shown). Interestingly, the combination of MC-Dropout with an Ensemble is essential to obtain both a high selectivity and a robustness to long training times; using just MC-Dropout results in a lower peak ratio than an Ensemble, but when using an Ensemble with no stochastic forward passes the high selectivity lasts for only an epoch or two before rapidly tapering off.

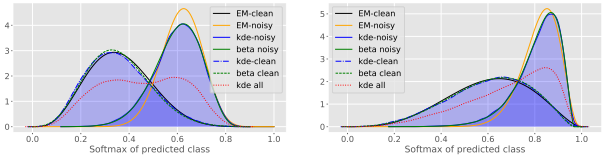
The optimal uncertainty threshold is highly dependent



(a) 40%; S; 10%; Softmax (b) 40%; S; 10%; VR

(c) 40%; P; 10%; Softmax (d) 40%; P; 10%; VR

Figure 1: Ability to detect CIFAR-10 training images with noisy labels over the course of training. The Y-axis is the percent of images in the subset of the X% most uncertain images that have noisy labels. The different colored-curves correspond to the model type used, and follow the format # of classifiers / # of stochastic forward passes per classifier. The subcaptions follow the format: noise level; type of noise (P for pair, S for symmetric); uncertainty threshold (i.e. for 10% on CIFAR-10, this would be the 5000 most uncertain images); uncertainty measure (VR=variation ratio).

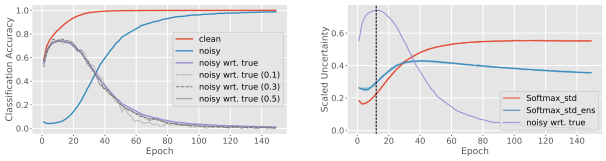


(a) CIFAR-10 (b) CIFAR-100

Figure 2: Expectation Maximization (EM) fits of two beta distributions (EM-clean / EM-noisy) to the averaged softmax values of images with clean and noisy labels. The overall distribution and its mixture components are plotted via kernel density estimations (kde all, kde-noisy, kde-clean) and their fit with a parametrized beta distribution (beta noisy, beta clean).

on many things, e.g. the data set, noise type, noise level, and model architecture. Within one experiment, as there are less noisy labels as the iterative process progresses, it makes sense to be more selective with the threshold. Instead of taking a fixed $p\%$ of most uncertain images, a possible extension is to model the distributions of the uncertainties of the noisy and clean images. Empirically, a mixture of two beta distributions fits rather well to this task , and after using the Expectation Maximization (EM) algorithm [5] to fit the distributions (Fig 2), the threshold can be chosen to be more selective, i.e. the number of clean images identified as noisy can be explicitly controlled for.

Relabeling of noisy labels After noisy labels are detected, the goal is to relabel these images with the true label. The simplest approach is to use the network's prediction for a noisy image before the network has overfit to the noisy labels. Looking at Fig 3, this begins to happen at a relatively early phase, and ultimately, the network learns to predict the



(a) Accuracy of the noisy and clean (b) Mean uncertainty (std. dev. of CIFAR-10 training images over the softmax vectors) over all correctly course of training (40% symmetric classified CIFAR-10 training images noise). Black curves correspond to over the course of training (40% training subsets with known noise. symmetric noise).

Figure 3: Relabeling of noisy images.

noisy labels perfectly. However, around epoch 15 the network has learned enough from the clean images to correctly relabel 80% of the noisy images (purple curve), before the overfitting to the noisy labels has begun.

Of course this information is not available during training, as it requires knowledge of the true label for noisy images. Thus two alternative approaches are proposed: the first uses a random subset of training images (e.g. 1000 images), in which the label noise is known (e.g. through expert relabeling; this is a common setting in the literature, e.g. [26]), and that the network is trained on. Fig 3a shows that it is not required to know the true noise level in the data set; the curves behave similarly, and are sufficient proxies to identify at which epoch to use the predictions for relabeling.

The second approach leverages the behavior of certain uncertainty estimates over the course of training, and requires no ground truth subset. The std. dev. over the softmax outputs of the multiple forward passes, averaged over all training examples, first briefly sinks, before beginning to rise as the predicted class of the noisy labeled-images switches from the true label to the noisy label (Fig 3b). This trend can be leveraged to roughly identify a good point to relabel using the predicted class. Averaging the forward passes within one classifier and taking the std. dev. over the resulting 5 vectors (blue curve) provides a better heuristic for identifying the ideal relabel time than taking the standard deviation over all 125 total forward passes.

4. Experimental results

As an initial proof of concept experiment, we tackle the task of noisy label detection on CIFAR-10 and CIFAR-100, using the simple convolutional network described in [3]. The noisy images are identified by taking the top 10% of most uncertain images ($p = 0.9$) (as in Fig1a). In Table 1 the relabeling is based on the predicted softmax at an epoch determined by the criteria presented in Fig 3b. The results of 5 iterations of this process are shown, starting with 40% symmetric noise. The algorithm is able to reduce the number of noisy labels by almost half. However, the accuracy of the trained networks do not rise (on a full clean data set this

Iter.	Acc.	# Noisy Images	Noise Prop.	Det. Prec.
1	0.775	20000	0.400	0.917
2	0.775	16803	0.336	0.852
3	0.775	13979	0.280	0.722
4	0.773	11804	0.236	0.576
5	0.767	10773	0.215	-

Table 1: Iterative relabeling based on predicted softmax on CIFAR-10. Det.Prec = Detection Precision

Iter.	Acc.	# Noisy Images	Noise Prop.	Det. Prec.
1	0.773 (0.477)	20000 (20000)	0.400 (0.400)	0.943 (0.868)
2	0.796 (0.513)	15284 (15660)	0.306 (0.313)	0.902 (0.756)
3	0.812 (0.535)	10775 (11881)	0.216 (0.237)	0.796 (0.611)
4	0.826 (0.557)	6797 (8825)	0.136 (0.177)	0.625 (0.451)
5	0.847 (0.572)	3671 (6572)	0.074 (0.131)	-

Table 2: Iterative relabeling with oracle relabeling on CIFAR-10 and CIFAR-100 (in parentheses). Det.Prec = Detection Precision

network achieves 87% accuracy). Further analysis reveals that the images identified as noisy and correctly relabeled simply are not helpful in making the network generalize and be more robust to the label noise; the images that remain noisy, which are more difficult to relabel, are those important for increasing the classifier's performance.

To highlight the effectiveness of the approach at detecting noisy labels, the experiment is repeated with oracle relabeling, in which all identified noisy images are given the correct true label (Table 2). Now the accuracy of the network rises as the number of noisy labels drops. At the end of the process, 16329 out of 20000 detected images were correctly identified as noisy and relabeled. As expected, the precision of the detection drops as there are fewer noisy labels in the data set, yet at each iteration a fixed 10% of images (i.e. 5000) with the highest uncertainty are selected.

5. Conclusion

We have shown that the predictive uncertainty given by a combination of an ensemble and MC-Dropout is very effective at identifying noisy labels under multiple noise settings and different datasets. Further work is focused on improving both the detection and the relabeling. For the former case, the beta-distribution fit will be further extended to increase the precision of detection at later stages in the algorithm. Possible extensions to the relabeling include using a majority vote instead of the predicted softmax, or assigning multiple labels to an image that is difficult to relabel, with the idea that at least one of them is the correct one. This can be done for a single model, or across model (i.e. each member of the ensemble gets a different label for an image). Finally, the methods will be tested on more data sets and network architectures, and compared to the state-of-the-art results from the literature.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. *International Conference on Machine Learning (ICML)*, 2019.

[2] Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep cnns with noisy labels. *arXiv preprint arXiv:1511.07069*, 2015.

[3] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9368–9377, 2018.

[4] Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[6] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[7] Yarin Gal. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning (ICML)*, 2016.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *International Conference on Machine Learning (ICML)*, 2017.

[10] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. *International Conference on Learning Representations (ICLR)*, 2017.

[11] 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 *Advances in Neural Information Processing Systems*, pages 8527–8537, 2018.

[12] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[13] Dan Hendrycks, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[14] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[15] Simon Jenni and Paolo Favaro. Deep bilevel learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–633, 2018.

[16] 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. *arXiv preprint arXiv:1712.05055*, 2017.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[18] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2018.

[19] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):17816, 2017.

[20] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017.

[21] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

[22] 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. *arXiv preprint arXiv:1806.02612*, 2018.

[23] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1196–1204, 2013.

[24] Ambrish Rawat, Martin Wistuba, and Maria-Irina Nicolae. Adversarial phenomenon in the eyes of bayesian deep learning. *arXiv preprint arXiv:1711.08244*, 2017.

[25] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[26] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.

[27] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

[28] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.

[29] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[30] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8688–8696, 2018.

[31] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 2691–2699, 2015.

[32] Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922, 2019.

[33] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR)*, 2017.

[34] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, pages 8778–8788, 2018.