Turning Noise into Treasure: A Complementary Label Learning Approach for Noisy Labels

Meng Pan
School of Computer Science and Engineering, Sun Yat-sen University, China
{panm9}@mail2.sysu.edu.cn

Abstract

1. Introduction

Monocular Depth Estimation is a crucial 3D scene perception algorithm. Trapped by the lack of sufficient geometric constraints, MDE is still an ill-posed problem. With the rapid development of Deep Neural Networks(DNNs), MDE has emerged many excellent works in recent years. However, according to Szegedy's research, DNNs are prone to be attacked by adversarial examples: when a imperceptibly small perturbation is added to an input image, the classifier based on DNNs will classify the image into a wrong category with high probability. This property has been verified not only on classification tasks, but also on logical regression tasks, such as semantic segmentation and object recognition. MDE also failed to get rid of this dilemma.

As a kind of 3D scene perception algorithm, MDE needs to be robust and reliable, so that when an attack occurs in automatic driving, the vehicle can still accurately perceive the depth of the scene. Otherwise, the damage is much more serious than the percentage loss in accuracy. On MDE, this problem was first confirmed and researched by Wong. They conducted comprehensive and rigorous experiments to explore the impact of pixel attacks on MDE. Including attacking the whole depth map of a scene, attacking the depth map of one single object in the scene, and even making a specific object vanished completely on the depth map by adding perturbations. This research is pioneering and enlightening for attacking MDE.

However, the study of Wong can hardly apply to physical world, it didn't take physical attack into account. The perturbations used in experiments are elaborately designed, which are very hard to realize in the physical world. While attacking a scene by patch is an appropriate solution: sticking a patch into the scene to be attacked is much more reasonable and implementable. Some researches exploit patches to attack DNNs, but most of them didn't pay much attention on inconspicuousness. Abrupt and unreal patch in scenes will reveal the attack intention. Indeed, some works noticed this problem and utilize GAN to make patches more naturalistic. This will increase time and space complexity obviously. In this study, we take physical realizability and inconspicuousness into accout, design an algorithm to attack MDE.

Specifically, we design a learnable matrix, which can generate reasonable patches according to depth map automatically. The generation processes are driven by three learnable coefficients and don't need any extra generation networks.

2. Related Work

In this work, patch attacks of MDE model are concentrated on, so a brief review of MDE and patch attacks will be introduced in this section.

2.1. Monocular Depth Estimation

2.1.1 Supervised

DNNs learn through the supervision of ground truth labels annotated artificially. Since the mapping from a single image to a depth map is very complicated, this supervised learning manner is very suitable for MDE tasks. The initial work adopts this intuitive endoder-decoder concept. Eigen et al designed two hierarchical neural networks to predict the depth map in coarse and fine respectively. A network was proposed subsequently by them, in which three different tasks, depth estimation, segmentation and normals prediction was integrated into one model. Lee coined novel local planar guidance layers and locate them at different scale decoding phases. As a result of this, final fine depth maps are densely restored from small to big. Current MDE networks are bulky and inefficient, which can not be equipped on embedding systems. Wofk et

al. proposed Fashdepth, a light, fast MDE network. Some researches address this problem by formulizing it as a classification problem. Discrete depth labels are prepared by separating continuous depth value, which are then used to supervise the training process. While using relative depth information to predict depth indirectly is also a effective solution. Lee proposed a method to predict depth from relative depth. They first estimate relative depth by utilizing rank-1 property, then the final depth maps are reconstructed through these relative depth maps.

2.1.2 Unsupervised

Although there exist some datasets with ground truth depth annotation, dense and accurate depth maps are labor-consuming and rare. In that case, unsupervised methods dominate recently. Godard et al. exploit the left-right consistency of binocular image pairs to tackle this issue: with a decently predicted disparity, the two images sampled from a binocular camera can synthesize each other. A consecutive vedio sequence can provide a constrain for MDE. Given a middle frame and its former, later frame sampled from a video, DNNs can estimate the camera pose and the depth, with which we can restore the other two frames. Based on this constrain, camera pose and depth can be learned jointly. However, there is a flaw in both video and binocular solutions. On the one hand, binocular image pairs methods struggle in occlusion and texture-copy problems, yet, on the other hand, as an alternative method, predicting depth from a video performs unsatisfactorily when it comes to relatively stationary objects. To address this, Godard et al. proposed a multi-scale reconstruction loss and a automasking approach to ignore relatively stationary objects.

2.2. Patch Attack

Adding perturbations to images has a limit: the imperceptive noise can not be captured by camera. Recently, researchers turn to generating adversarial examples with patches. It can be captured by camera and it can implement to physical world. The only question is, how to make it concealed.

3. Approach

In this section, we first introduce some preliminaries about our noisy label learning. Subsequently, we give a brief introduction to our baseline method DivideMix [6]. Finally, we illustrate our proposed complementary label learning approach.

3.1. Preliminaries

Problem formulation. We define the task of K-category classification with noisy labels. Formally, we denote the labeled dataset $\mathbb{D} = \{(x_i, y_i) \mid 1 \leq i \leq N\}$, in which $x_i \in \mathcal{X}$ is the i-th training sample and $y_i \in \mathcal{Y}$ is the corresponding label over K classes with noise. Normally, a supervised classification learns a function $f : \mathcal{X} \to \mathcal{Y}$ that maps the input space to label space. The common objective loss function that used to define the empirical risk is the Cross-Entropy(CE) loss function:

$$\mathcal{L}_{CE}(f, y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_i^k \log \left(p_i^k\right) \tag{1}$$

in which p_i^k is the predicted probability of sample x_i for class k. The objective of training classifier is to find the optimal parameters θ that minimize the objective loss function. And our target is to train a classifier robust to the noisy label in training set.

Sample-selection-based methods. A common idea of dealing with noisy labels is sample-selection-based method, which aims to determine which samples may be noisy. Such methods need to find a reliable selector to distinguish the clean samples and noisy samples, and then either discard noisy samples directly or reassign labels for them. The challenging issue is to design a reliable criteria to select clean samples. Previous works have shown that deep neural networks(DNNs) tend to learn simple pattern first before fitting label noise [1]. Therefore, lots of methods treat samples with large loss as noisy samples and construct new pseudo labels for them like semi-supervised learning.

3.2. Baseline: DivideMix

Our baseline method is DivideMix [6] which is a typical sample selection based method. DivideMix uses a mixture model to model the per-sample loss distribution so that it can divide the training data into a labeled set with clean samples and an unlabeled set with noisy samples dynamically. Besides, to avoid confirmation bias of self-training, that is, the model would accumulate its errors, DivideMix train two networks at the same time to filter errors through epoch-level implicit teaching and batch-level explicit teaching. Specifically, at each epoch, DivideMix perform co-divide, in which one network divides the noisy training dataset into clean labeled set and noisy unlabeled set, which are then used by the other network. At each mini-batch, utilizes both labeled and unlabeled samples are utilized by one network to perform semi-supervised learning guided by the other network.

3.3. Our negative losses

DivideMix has achieved remarkable performance on many large-scale datasets with both symmetric and asymmetric label noise. However, it does not make good use of the labels of noisy samples, but directly assign pseudo labels for them. In order to further utilize the label of noisy samples, we propose to adopt complementary label learning [4] approach that utilize noisy labels as complementary labels to serve as an additional robust supervised signal after sample selection. Our approach is based on two basic assumptions:

1) Designing a reliable sample detector is feasible because DNNs tend to learn common patterns first and then memorize the noisy data [1].

2) Given a reliable sample selector, we are more sure of what an image isn't than what it is.

Based on the two assumptions, we apply the complementary label learning after sample selection. For noisy sample, the original label now serves as the complementary label and the output probability of it is suppressed explicitly. Specially, given the label $\tilde{y}_i$ of a noisy sample , we can define $q_i^k = 1 - p_i^k$, which means the probability that the sample does not belong to class k. Then we treat the label $\tilde{y}_i$ of a noisy sample as the label that it does not belong to this class and calculate the loss with q_i^k.

Negative-CE(Neg-CE). We can use Cross-Entropy loss function to measure the difference between predicted q_i^K and label $\tilde{y}_i$ of a noisy sample:

$$\mathcal{L}_{Neg-CE}(f, \tilde{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\tilde{y}_i^k \log\left(q_i^k\right) \qquad (2)$$

Negative-MSE(Neg-MSE). We can also use mean square error(MSE) instead of Cross-Entropy loss, the Neg-MSE loss function for noisy labels can be defined as:

$$\mathcal{L}_{Neg-MSE}(f, \tilde{y}) = -\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(q_i^k - \tilde{y}_i^k)^2 \qquad (3)$$

Our proposed Neg-CE or Neg-MSE loss function can be combined with arbitrary existing sample-selection methods by simply adding the Neg-CE/MSE loss term.

4. Experiments

In this section, we would like to answer the following questions: (1) how does the loss function compare to the baseline? (2) Is the loss function robust and efficient enough for different noisy label ratios and different noisy label types? To answer these question, we evaluate our method on CIFAR-10 datasets [5]. We apply our loss function to DivideMix [6] and compare our model to the naive model in different sets.

4.1. Datasets and Label Noise Types

CIFAR-10 [5]. To demonstrate the capability of the loss function, we conduct experiments on CIFAR-10, consisting of 32×32 color images arranged in 10 classes. The datasets contains 50,000 training and 10,000 test images.

Label Noise Types. For CIFAR-10, we use two types of label noise to mix up: symmetric and asymmetric. Symmetric: The label has equal probability to flip to another class. Following [2]and [7], we generate noisy label by convert the labels of a give training data to one of the other classes using the set proportion(the ground truth labels are randomly scrambled). Asymmetric: The label was only changed to a specific class. For example, we choose the train sample randomly with a set probability and change the label to its similar class (e.g. BIRD → AIRPLANE), CAT → DOG, HORSE → DEER, same to the setting in [8].

For comparing the methods, we mix up the datasets under asymmetric noise rate of $\eta_{asymm} \in \{0.4\}$ and the symmetric noise rates of $\eta_{symm} \in \{0.2, 0.5, 0.8, 0.9\}$.

4.2. Implementation Details

Hyperparameters. For implementation on CIFAR-10, we set the batch size to 64 and use an 18-layer Pre-Train Resnet [3]. Training it use stochastic gradient descent (SGD), momentum of 0.9, weight decay of $5e^{-4}$, and a batch size of 64. Besides, we train the two networks for 300 epochs. The initial learning rate is set to 0.02, and was reduced by a factor 10 after 150 epochs. For Warm up the network, the period is 10 epochs for CIFAR-10.

Data Argumentation. In order to enhance the diversity and richness of the data and enable the network to learn more robust features, we performed normalize, random horizontal, and random crops for CIFAR-10.

The Hyperparameters and Data Augmentation are the same with baseline [6].

4.3. Quantitative Results

Table 1 shows the results of our method and baseline method in various noise types and ratios on CIFAR-10. The result shows baseline method has a high accuracy in the low noisy environment, but the performance decreases drastically as the noise ratio increases. Our

Method/Noise		symmetric				asymmetric
		0.2	0.5	0.8	0.9	0.5
DivideMix [6]	Avg	95.7	94.4	92.9	75.4	92.1
	Best	96.1	94.6	93.2	76.0	93.4
+Neg-MSE	Avg	95.6	94.3	93.2	77.8	92.9
	Best	95.8	94.6	93.5	79.2	93.4
+Neg-CE	Avg	95.2	93.5	92.4	77.0	92.1
	Best	95.5	94.1	92.6	78.5	92.6

Table 1. Comparison with baseline method on CIFAR-10 mixed with various types and ratios of noise.

method (Neg-MSE) obtained an improvement compared to the baseline method when the noise ratio is high. But the performance is not good enough when the set comes to low noise ratio (e.g. 0.2,0.5). In addition, compared to the cross entropy, MSE Loss is much more robust to various label noise ratio.

4.4. Further Analysis

Results on CIFAR-10. Fig.1 shows the accuracy graph of Negmse (our) and DivideMix with CIFAR-10 mixed with 20%, 50%, 80%, and 90% symmetric noise. As can be seen in Fig.1, training with negative Mse can improve the AUC of clean/noiey selection in high noise radio setting (bottom), but reduce it in low noise radio setting (top). That's to say, the negative Mse tends to let the network makes more false positive predictions than the baseline under low noise radio.

we have two Hypothesis to explain the performance of our loss function on CIFAR-10. Firstly, false positive samples dominate the gradient of negative MSE loss. Assuming that, the similarity between distribution of prediction and label of false positive data is larger than that of true noisy data. So, the negative MSE would produce greater gradients on these false positive samples then the true label samples. But the optimization direction for these samples is most likely to be wrong. When the noisy ratio comes to low, the nework in a much stronger wrong direction will have a lower accuracy in true positive data then before. Secondly, training in low noise ratio is more sensitive to false positive samples compared to high loss ratio,since the number of true noisy sample is less, which will enhance the influence of false positive samples.

5. Conclusion

In this paper, we apply the complementary label learning to further utilize the noisy labels to strengthen the training of existing sample-selection methods, instead of discarding them. From the empirical results, we observe that it works well on high noise ratio, but gets trouble on low noise rate. We think the underlying reason is related to the influence of FP samples. In a word, the results on high noise ratio show the effectiveness of our approach, but the influence of FP samples needs to be alleviated, which is the focus of our future work.

We then speculate that the reason for poor performance under low noise rate is the influence of false positive (FP) samples(i.e.,clean samples which are misjudged as noise).

References

[1] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In International Conference on Machine Learning, pages 233–242. PMLR, 2017. 2, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016. 3

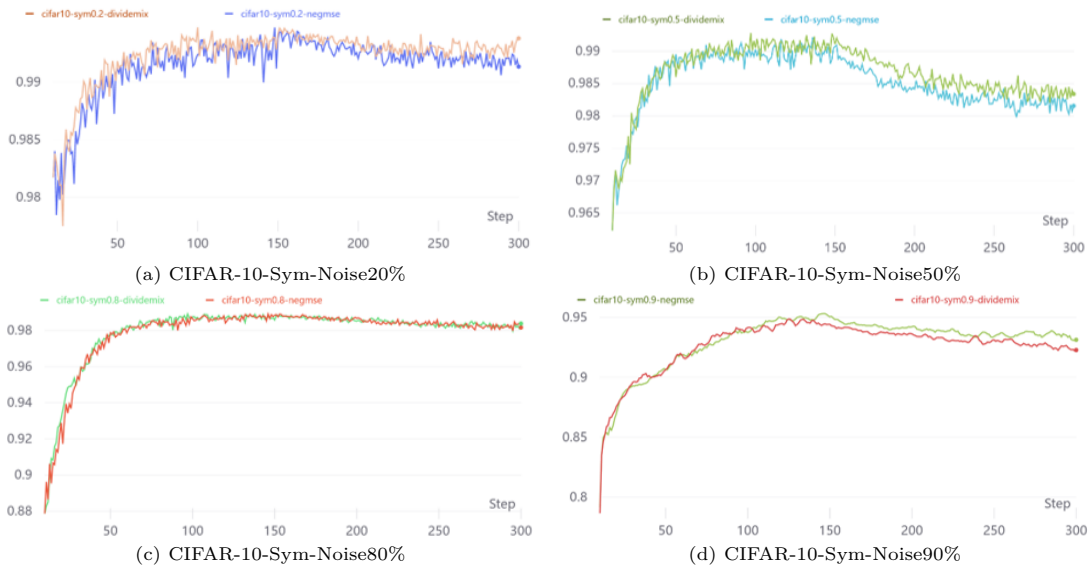
[4] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. arXiv preprint arXiv:1705.07541, 2017. 3

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[6] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394, 2020. 2, 3, 4

[7] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5552–5560, 2018. 3

[8] Jiangchao Yao, Hao Wu, Ya Zhang, Ivor W Tsang, and Jun Sun. Safeguarded dynamic label regression for noisy supervision. In Proceedings of the AAAI Conference



(a) CIFAR-10-Sym-Noise20%

(b) CIFAR-10-Sym-Noise50%

(c) CIFAR-10-Sym-Noise80%

(d) CIFAR-10-Sym-Noise90%

Figure 1. Accuracy graph of Negmse(our) and DivideMix with CIFAR-10 mixed with 20%, 50%, 80%, and 90% symmetric noise.

on Artificial Intelligence, volume 33, pages 9103–9110, 2019. 3