CYCLIC ANNEALING TRAINING CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE CLASSIFICATION WITH NOISY LABELS

Jiawei Li, Tao Dai, Qingtao Tang, Yeli Xing, Shu-Tao Xia

Department of Computer Science and Technology, Tsinghua University, China
{li-jw15, dait14, tqt15, xyl16}@mails.tsinghua.edu.cn, xiast@sz.tsinghua.edu.cn

ABSTRACT

Noisy labels modeling makes a convolutional neural network (CNN) more robust for the image classification problem. However, current noisy labels modeling methods usually require an expectation-maximization (EM) based procedure to optimize the parameters, which is computationally expensive. In this paper, we utilize a fast annealing training method to speed up the CNN training in every M-step. Since the training is repeated executed along the entire EM optimization path and obtain many local minimal CNN models from every training cycle, we name it as the Cyclic Annealing Training (CAT) approach. In addition to reducing the training time, CAT can further bagging all the local minimal CNN models at the test time to improve the performance of classification. We evaluate the proposed method on several image classification datasets with different noisy labels patterns, and the results show that our CAT approach outperforms state-of-the-art noisy labels modeling methods.

Index Terms— Image Classification, Noisy Labels, Cyclic Annealing Training, EM algorithm, Bagging CNNs.

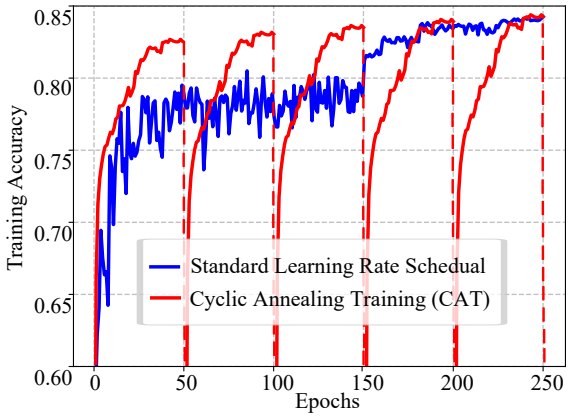

Fig. 1: Training DenseNet-40 on CIFAR-10 using *standard learning rate schedule* (blue) and *cyclic annealing training* (red). To tackle the noisy labels problem, current EM based approaches require an entire training in every M-step. Fast annealing the learning rate in every M-step cycle is able to speed up the convergence and obtain many intermediate models (denoted by the dotted red lines). At the test time, we can further bagging these models to improve the performance.

1. INTRODUCTION

Convolutional neural network (CNN) has been successfully used in many supervised learning tasks, such as image classification or object recognition. In general, the labels used to train models are assumed to be accurate. However, in practice, labeling image dataset by hand is a subjective task and easily induce noise to vary degrees, thus leading to deteriorating the performance [1, 2]. To improve the robustness of CNN models, there exist many practicable noise modeling methods [3, 4, 5, 6] to tackle the noise in the feature. Beyond the noise in feature, it is common that noise in label (label noise) [7] and still remains to be addressed [8].

To cope with the label noise, a series of noise modeling approaches [9, 10, 11, 12] have been proposed based on the expectation-maximization (EM) framework, which assumes that the unknown true label could be regarded as a hidden random variable. In these EM-based approaches, every E-step estimates the true label, while every M-step updates the model parameters. These EM-based iterative methods can render CNN models robust to label noise, however, the repeated training in each M-steps is very expensive, especially when the training of modern deep CNN already requires many computing resource.

To reduce the training cost, one recent work, named by Noise Adaption Layer (NAL) [13], uses an additional plugged noise layer to learn the latent pattern of noisy labels and it optimizes all parameters within a standard training procedure of neural networks. However, due to the degrees of freedom in noise layer, NAL approach suffers the problem of convergence, and the performance is sensitive to the initialization of parameters [13]: only a careful initialization of parameters can successful converge to a robust model and extra training cost of this initialization is still unavoidable.

In this paper, we propose a novel robust CNN model by

This work is supported by the National Natural Science Foundation of China under grant Nos. 61771273. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of Titan X GPUs for this research.

embedding cyclic annealing training (CAT) into EM framework to speed up the convergence. To be specific, we train the CNN classifier with a fast cosine annealing learning rate schedule in every M-step cycle and also alternately update the noise pattern in every E-step.

On the one hand, CAT reduces the training cost and does not have any limitation for the parameters initialization. The comparison of different training schedules is illustrated in Figure 1, from which we find that CAT significantly speed up the convergence of CNN. On the other hand, CAT enables the usage of ensemble [14] strategy. It is known that bagging [15] is more robust than boosting [16] in the presence of noisy labels [17, 18]. As CAT learns the noise pattern from every E-step, the latest noise pattern can be used as a prior condition for every E-step CNN models. Thus, inspired by the Snapshot Ensemble [19], at the test time, we aggregate many intermediate models to further improve the performance.

We evaluate the performance and efficacy of our CAT approach on image classification with different label noise patterns. In contrast to current EM based approaches [9, 10, 11, 12], when a dataset has a given noisy pattern, CAT is significantly faster to learn it and has a better classification performance. Compared with the state-of-art NAL approach [13], CAT is able to train a better classifier in all of the noise levels.

It is worthwhile to highlight two main contributions of our proposed CAT approach:

- Existing EM based noisy labels modeling approaches require too many training time. We utilize a fast annealing training method to speed up the learning progress, without any limitation for the parameters initialization.
- As we can obtain many intermediate models from all of the M-step cycles, we bagging all of them to further improve the robustness when the noise pattern is obtained from all of the E-steps.

2. CYCLIC ANNEALING TRAINING

2.1. Noisy Labels Modelling

To solve the noisy labels problem, it is natural to consider the noise from a statistical point of view [8]. Figure 2 shows two possible statistical models of the label noise pattern.

For reasons of brevity, here we first consider the left situation and the right situation will be discussed in our experiments, where the noisy label z only depends on true label y with a transfer probability $\Theta(\theta_{ij} = p(z = j|y = i))$. Then a robust CNN classifier $p(y = i|x; W)$ can be trained on a n samples noisy labels dataset $D = (x_i, z_i), i = 1, ..., n$ by different training approaches. Specifically, W is the network parameters, the dataset has n samples, x is the feature of sample, and i is one of k class index.

Figure 3 illustrates a high-level view of the noisy labels model architecture: the noise can be modeled by different

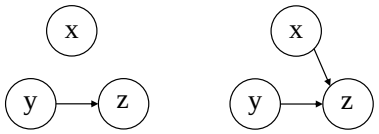

Fig. 2: **Left:** noisy label z only depends on true label y. **Right:** z depends on both of true label y and feature x.

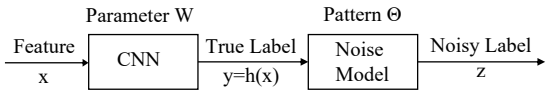

Fig. 3: A typical label noise modeling framework.

noise model layer [10, 13] inserted between the CNN softmax layer and the cross-entropy cost layer. The log likelihood of model parameters will be:

$$L(W, \Theta) = \sum_{t=1}^{n} log(\sum_{i=1}^{k} p(z_t|y_t = i; \Theta)p(y_t = i|x_t; W)). \tag{1}$$

Then, as y is a hidden random variable, it is natural to utilize an EM based method [12] to find the maximum-likelihood parameters W and Θ.

2.1.1. EM Approach

In every E-step, all parameters are fixed and hidden true label y can be estimated as:

$$q_{ti} = p(y_t = i|x_t, z_t) \propto p(y_t = i|x_t; W)p(z_t|y_t = i; \Theta), \tag{2}$$

where q_{ti} is a soft estimates of sample x_t belonging to true label $i = 1, ..., k$.

In every following M-step, at first the noise parameter Θ can be updated by q_{ti} in a closed-form function:

$$\Theta_{ij} = \frac{\sum_t q_{ti} 1_{\{z_t = j\}}}{\sum_t q_{ti}}. \tag{3}$$

Then the CNN parameters W will be trained with an optimization algorithm such as SGD. In this way, the log likelihood (1) will have a soft version:

$$\arg\max_W L(W) = \sum_{t=1}^{n} \sum_{i=1}^{k} q_{ti} \log p(y_t = i|x_t, W). \tag{4}$$

Once the iteration conditions are met, we will have the noise pattern Θ and the CNN parameters W.

2.1.2. Noise Adaption Layer Approach

Recently, a noise adaption layer approach [13] is proposed to optimize the log-likelihood function (1) directly within the procedure of CNN training.

Denote the parameter of last fully connected layer as w and b, the other modules of CNN as function $h = h(x)$. Then the softmax prediction loss of true label y is:

$$p(y = i|x; h, w, b) = \frac{exp(w_i^\top h(x) + b_i)}{\sum_{l=1}^{k} exp(w_l^\top h(x) + b_l)}. \quad (5)$$

The label noise is modeled by an additional noise adaptation softmax layer between the original classification softmax and cross-entropy cost layer. This noise adaptation softmax layer describes the transfer probability as $\theta_{ij} = \frac{exp(q_{ij})}{\sum_l exp(q_{il})}$. As this noise layer is modelled with a differentiable layer parameter q_{ij}, all of the model parameters can be optimized within a typical network training (e.g. SGD), as below:

$$\arg\max_{h,w,b,q} L(h, w, b, q) =$$

$$\sum_{t=1}^{n} \log(\sum_{i=1}^{k} p(z_t|y_t = i, x_t; q) p(y_t = i|x_t; h, w, b)). \quad (6)$$

2.2. Cyclic Annealing Learning Rate

The noise adaptation layer approach does not need any iterative CNN training, but it gives rise to a complex parameter initialization. Because only a very careful initialization can make a successful converge, too much extra training costs of the initialization procedure are still unavoidable. To directly reduce the training cost without raising any other problems, our CAT approach benefits from the *cyclic annealing learning rate* [20] and is able to speed up the learning of every M-step.

The state-of-art CNN architectures for image classification such as ResNet [21] and DenseNet [22] usually have millions of parameters. A study [23] demonstrates that the more parameters, the more possible local minima could be visited in the training phase. The cyclic annealing learning rate method aims to generate many local optima CNN models from a single training process. Specifically, it abruptly raises the learning rate α and then quickly decreases it with a cosine function:

$$\alpha(t) = \frac{\alpha_0}{2}(\cos(\frac{\pi mod(t-1, \lceil T/C \rceil)}{\lceil T/C \rceil}) + 1), \quad (7)$$

where t is current epoch number, T is the total epoch number, α_0 is the initial learning rate, and the total training epochs are divided to equal C cycles.

While the training cost for each current M-step is an entire CNN training, we can align every annealing learning rate cycle to an M-step and then use the obtained local minimal CNN model to update the following E-step. By this way, our CAT approach can directly achieve a C times faster convergence than current EM based approaches. In order to further accelerate the convergence, we can also faster update the learning rate at every iteration rather than at each epoch [20].

2.3. Bagging CNNs

We can obtain many local minima CNN models from different M-steps. As those models usually make different mistakes, bagging all of them is able to significantly improve the classification performance [19].

In the presence of the noisy labels, our CAT approach is able to iteratively learn a noise pattern Θ through many E-steps. At the test time, we can use this noise pattern to divide the noise layer from local minimal CNN models. In this way, the other part of CNN will become a robust classifier, which predicts the true label y but not the observed noisy label z.

2.4. Cyclic Annealing Training

For the noisy labels problem, the details of our proposed CAT approach is given in Algorithm 1.

Algorithm 1 Cyclic Annealing Training on Noisy Labels

1: Given n samples training data $X(x_1, ..., x_n)$ with noisy label $Z(z_1, ..., z_n)$, the true label $Y(y_1, ..., y_n)$ are unknown. The transfer probability between true label and noisy label is denoted as $\Theta(\theta_{ij} = p(z = j|y = i))$.
2: We first generate a random matrix $\hat{\Theta}_0$ to be the initialization of the noise pattern.
3: Then we repeatedly do C times the following:

(1) For every training cycle c ranges from 1 to C, initiate the learning rate with a constant value α_0.

(2) With the learning rate annealing from α_0 to 0 as function $\alpha(t) = \frac{\alpha_0}{2}(cos(\frac{\pi \cdot mod(t-1,T)}{T}) + 1)$, where t is current iteration number, train the CNN $p(y|x; W^c)$ with a fixed follow-up noise layer (linear or softmax) $p(z|y; \hat{\Theta}_{c-1})$ for total T iterations.

(3) Update the learned noise pattern $\hat{\Theta}_c$ with the closed-form function (3).

4: Once all of the training finished, drop the noise layer according to the final $\hat{\Theta}_C$. The remaining CNN parameters W^c will be used to predict the true labels, as $\hat{f}_c = p(y|x; W^c), c = 1, ..., C$.
5: For any prediction sample $\langle x_0, z_0 \rangle$ with a hidden true label y_0, the aggregating output $\hat{f}$ of bagging CNNs is the simple averaging $\hat{f}^{AVG}(x_0) = \frac{1}{C}\sum_{c=1}^{C} \hat{f}_c(x_0)$.
6: The prediction error is given by counting the proportion of prediction mistakes $\hat{f}(x_0) \neq y_0$ among the test dataset.

Compared to the state-of-art noisy labels modeling approaches [10, 11, 13], the proposed CAT approach is not only faster, but also more accurate and without any limitation for the parameters initialization.

3. EXPERIMENTS

In the following experiments, we compare our CAT approach with other state-of-art approaches [10, 11, 13] in three image classification datasets (e.g. MNIST [24], CIFAR-10 and CIFAR-100 [25]), in the presence of noisy labels.

3.1. Settings

Figure 2 illustrates two possible label noise patterns [8]. Different to other approaches, we use the same CAT approach (Algorithm 1) to tackle both of the two noise patterns.

Several recently proposed noisy labels modelling methods are used to be the baselines: (1) Original CNN: The CNN is trained in the normal way, without any noisy labels modeling methods; (2) Hard Bootstrap EM [11]: The CNN is iteratively trained on a hard convex combination of the noisy label z and currently predicts label y, which is a state-of-art EM nosiy modelling approach; (3) Simple NAL [13]: The CNN is trained with a noise adaptation layer, in which the noisy label z depends on only the true label y; and (4) Complex NAL [13]: The CNN is trained with a noise adaptation layer, in which the noisy label z is depending on feature x and true label y. We also include a method without bagging: in step 4 of Algorithm 1, we directly predict the true labels with a single CNN $\hat{f}_C$ but not the bagging CNNs $\hat{f}^{AVG}$.

3.2. Effective of CAT

We first generate a noisy MNIST dataset with label flipping operation, where a label is erroneously given another label within the dataset. Specifically, we randomly flip the labels as a noise pattern [7,9,0,4,2,1,3,5,6,8], which means digital 0 will be labeled by 7, 1 by 9, and so on. Then our CAT approach is applied on the noisy MNIST dataset to train an 8-layer CNN (Conv-ReLU-Max-Conv-ReLU-Max-FC-FC) [13]. When up to 46 percent of the labels are flipped, the simple NAL approach can achieve a 99.68% classification accuracy, while our CAT approach can still achieve a 99.77% classification accuracy. As Figure 4 illustrated, the transfer probability $\hat{\Theta}$ acquired by CAT is more consistent to the prior noise pattern. Therefore, our CAT approach is still effective when the noise level is high.

3.3. Comparison of Efficiency

To show the efficiency, we adopt different approaches to train a DenseNet-40 [22] on the noisy CIFAR-10 dataset, which has 10% randomly flipped labels. The result is illustrated in Figure 1. As CAT needs less time to converge, it is more effective than other EM based approaches.

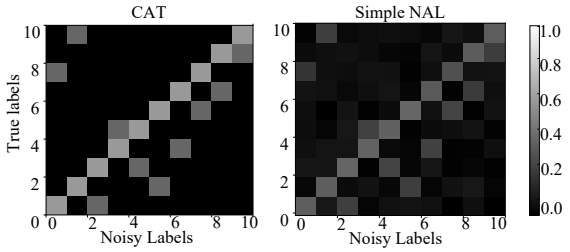

Fig. 4: 46% noisy labels on MNIST with noise pattern [7,9,0,4,2,1,3,5,6,8]. The acquired transfer probability $\hat{\Theta}$ of CAT and Simple NAL are visualized.

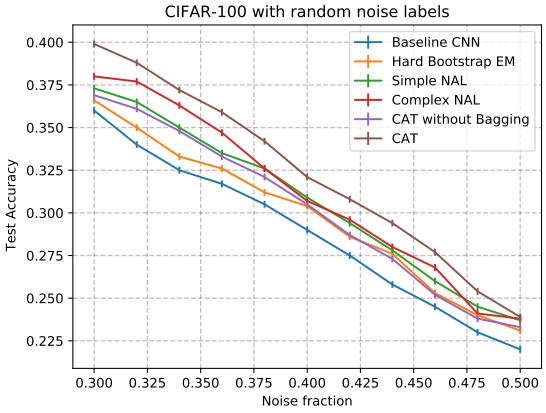

Fig. 5: Our CAT approach achieve a state-of-art classification accuracy on CIFAR-100 in the presence of noisy labels with different noise fraction.

3.4. Comparison of Performance

We also evaluate how the different noise fraction influences our proposed CAT approach. Figure 5 depicts that the test accuracy is a monotone decreasing function of the noise fractions generally, with the fraction stepping from 0.3 to 0.5 with the step value 0.02. It shows that our approach is more robust than other state-of-art noisy labels modeling approaches.

4. CONCLUSION

In this paper, we propose a Cyclic Annealing Training (CAT) approach to train the CNN for image classification in the presence of noisy labels. While current EM based noisy labels modeling approaches require too much time costs, CAT is able to directly reduce the training cost in every M-step, without bringing in a complex parameter initialization. At the test time, we bagging all of the intermediate models, which comes from many M-step cycles. The experiments show that these proposed strategies make the CNN model more robust and effective.

5. REFERENCES

[1] David F Nettleton, Albert Orriols-Puig, and Albert Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artificial intelligence review*, vol. 33, no. 4, pp. 275–306, 2010.

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

[3] Deyu Meng and Fernando De La Torre, "Robust matrix factorization with unknown noise," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1337–1344.

[4] Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Lei Zhang, "Robust principal component analysis with complex noise," in *International Conference on Machine Learning (ICML)*, 2014, pp. 55–63.

[5] Xiangyong Cao, Yang Chen, Qian Zhao, Deyu Meng, Yao Wang, Dong Wang, and Zongben Xu, "Low-rank matrix factorization under general mixture noise distributions," in *ICCV*. IEEE, 2015, pp. 1493–1501.

[6] Xiangyong Cao, Qian Zhao, Deyu Meng, Yang Chen, and Zongben Xu, "Robust low-rank matrix factorization under general mixture noise distributions," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4677–4690, 2016.

[7] José A Sáez, Mikel Galar, Julián Luengo, and Francisco Herrera, "Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition," *Knowledge and information systems*, vol. 38, no. 1, pp. 179–206, 2014.

[8] Benoit Frenay and Michel Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.

[9] Volodymyr Mnih and Geoffrey E Hinton, "Learning to label aerial images from noisy data," in *ICML*, 2012, pp. 567–574.

[10] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus, "Training convolutional networks with noisy labels," in *ICLR Workshop*, 2015.

[11] Dragomir Anguelov Christian Szegedy Dumitru Erhan Scott E. Reed, Honglak Lee and Andrew Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," in *ICLR Workshop*, 2015.

[12] Jacob Bekker, Alan Joseph Goldberger, "Training deep neural-networks based on unreliable labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2682–2686.

[13] Jacob Goldberger and Ehud Ben-Reuven, "Training deep neural networks using a noise adaption layer," in *International Conference on Learning Representations (ICLR)*, 2017, pp. 11–22.

[14] Zhi-Hua Zhou, *Ensemble methods: foundations and algorithms*, CRC press, 2012.

[15] Leo Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] Yoav Freund, Robert E Schapire, et al., "Experiments with a new boosting algorithm," in *ICML*. Bari, Italy, 1996, vol. 96, pp. 148–156.

[17] Thomas G Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.

[18] Joaquín Abellán and Andrés R Masegosa, "Bagging decision trees on data sets with classification noise," in *International Symposium on Foundations of Information and Knowledge Systems (FoIKS)*. Springer, 2010, pp. 248–265.

[19] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger, "Snapshot ensembles: Train 1, get m for free," *ICLR*, 2017.

[20] Ilya Loshchilov and Frank Hutter, "Sgdr: stochastic gradient descent with restarts," in *ICLR*, 2017.

[21] 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 (CVPR)*, 2016, pp. 770–778.

[22] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[23] Kenji Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing System (NIPS)*, 2016, pp. 586–594.

[24] Yann LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[25] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," *Technical report, University of Toronto*, 2009.