MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels

Lu Jiang [1] **Zhengyuan Zhou** [2] **Thomas Leung** [1] **Li-Jia Li** [1] **Li Fei-Fei** [1 2]

Abstract

Recent deep networks are capable of memorizing the entire data even when the labels are completely random. To overcome the overfitting on corrupted labels, we propose a novel technique of learning another neural network, called MentorNet, to supervise the training of the base deep networks, namely, StudentNet. During training, MentorNet provides a curriculum (sample weighting scheme) for StudentNet to focus on the sample the label of which is probably correct. Unlike the existing curriculum that is usually predefined by human experts, MentorNet learns a data-driven curriculum dynamically with StudentNet. Experimental results demonstrate that our approach can significantly improve the generalization performance of deep networks trained on corrupted training data. Notably, to the best of our knowledge, we achieve the best-published result on WebVision, a large benchmark containing 2.2 million images of real-world noisy labels. The code are at https://github.com/google/mentornet.

1. Introduction

Zhang *et al.* (2017a) found that deep convolutional neural networks (CNNs) are capable of memorizing the entire data even with corrupted labels, where some or all true labels are replaced with random labels. It is a consensus that deeper CNNs usually lead to better performance. However, the ability of deep CNNs to overfit or memorize the corrupted labels can lead to very poor generalization performance (Zhang et al., 2017a). Recently, Neyshabur *et al.* (2017) and Arpit *et al.* (2017) proposed deep learning generalization theories to explain this interesting phenomenon.

This paper studies how to overcome the corrupted label for deep CNNs, so as to improve generalization performance on the clean test data. Although learning models on weakly labeled data might not be novel, improving deep CNNs on corrupted labels is clearly an under-studied problem and worthy of exploration, as deep CNNs are more prone to overfitting and memorizing corrupted labels (Zhang et al., 2017a). To address this issue, we focus on training very deep CNNs from scratch, such as resnet-101 (He et al., 2016) or inception-resnet (Szegedy et al., 2017) which has a few hundred layers and orders-of-magnitude more parameters than the number of training samples. These networks can achieve the state-of-the-art result but perform poorly when trained on corrupted labels.

Inspired by the recent success of Curriculum Learning (CL), this paper tackles this problem using CL (Bengio et al., 2009), a learning paradigm inspired by the cognitive process of human and animals, in which a model is learned gradually using samples ordered in a meaningful sequence. A curriculum specifies a scheme under which training samples will be gradually learned. CL has successfully improved the performance on a variety of problems. In our problem, our intuition is that a curriculum, similar to its role in education, may provide meaningful supervision to help a student overcome corrupted labels. A reasonable curriculum can help the student focus on the samples whose labels have a high chance of being correct.

However, for the deep CNNs, we need to address two limitations of the existing CL methodology. First, existing curriculums are usually predefined and remain fixed during training, ignoring the feedback from the student. The learning procedure of deep CNNs is quite complicated, and may not be accurately modeled by the predefined curriculum. Second, the alternating minimization, commonly used in CL and self-paced learning (Kumar et al., 2010) requires alternative variable updates, which is difficult for training very deep CNNs via mini-batch stochastic gradient descent.

To this end, we propose a method to learn the curriculum from data by a network called *MentorNet*. MentorNet learns a data-driven curriculum to supervise the base deep CNN, namely *StudentNet*. MentorNet can be learned to approximate an existing predefined curriculum or discover new data-driven curriculums from data. The learned data-driven

[1]Google Inc., Mountain View, United States [2]Stanford University, Stanford, United States. Correspondence to: Lu Jiang <lujiang@google.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

curriculum can be updated a few times taking into account of the StudentNet's feedback. Whenever MentorNet is learned or updated, we fix its parameter and use it together with StudentNet to minimize the learning objective, where MentorNet controls the timing and attention to learn each sample. At the test time, StudentNet makes predictions alone without MentorNet.

The proposed method improves existing curriculum learning in two aspects. First, our curriculum is learned from data rather than predefined by human experts. It takes into account of the feedback from StudentNet and can be dynamically adjusted during training. Intuitively, this resembles a "collaborative" learning paradigm, where the curriculum is determined by the teacher and student together. Second, in our algorithm, the learning objective is jointly minimized using MentorNet and StudentNet via mini-batch stochastic gradient descent. Therefore, the algorithm can be conveniently parallelized to train deep CNNs on big data. We show the convergence and empirically verify it on large-scale benchmarks.

We verify our method on four benchmarks. Results show that it can significantly improve the performance of deep CNNs trained on both controlled and real-world corrupted training data. Notably, to the best of our knowledge, it achieves the best-published result on WebVision (Li et al., 2017a), a large benchmark containing 2.2 million images of real-world noisy labels. To summarize, the contribution of this paper is threefold:

- We propose a novel method to learn data-driven curriculums for deep CNNs trained on corrupted labels.
- We discuss an algorithm to perform curriculum learning for deep networks via mini-batch stochastic gradient descent.
- We verify our method on 4 benchmarks and achieve the best-published result on the WebVision benchmark.

2. Preliminary on Curriculum Learning

We formulate our problem based on the model in (Kumar et al., 2010) and (Jiang et al., 2015). Consider a classification problem with the training set $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$, where $\mathbf{x}_i$ denotes the i^{th} observed sample and $\mathbf{y}_i \in \{0,1\}^m$ is the noisy label vector over m classes. Let $g_s(\mathbf{x}_i, \mathbf{w})$ denote the discriminative function of a neural network called *StudentNet*, parameterized by $\mathbf{w} \in \mathbb{R}^d$. Further, let $\mathbf{L}(\mathbf{y}_i, g_s(\mathbf{x}_i, \mathbf{w}))$, a m-dimensional column vector, denote the loss over m classes. Introduce the latent weight variable, $\mathbf{v} \in \mathbb{R}^{n\times m}$, and optimize the objective:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{v}\in[0,1]^{n\times m}} \mathbb{F}(\mathbf{w},\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i^T \mathbf{L}(\mathbf{y}_i, g_s(\mathbf{x}_i,\mathbf{w})) + G(\mathbf{v};\lambda) + \theta\|\mathbf{w}\|_2^2 \quad (1)$$

where $\|\cdot\|_2$ is the l_2 norm for weight decay, and data augmentation and dropout are subsumed inside g_s. $\mathbf{v}_i \in [0,1]^{m\times 1}$ is a vector to represent the latent weight variable for the i-th sample. The function G defines a *curriculum*, parameterized by λ. This paper focuses on the one-hot label. For notation convenience, denote the loss $\mathbf{L}(\mathbf{y}_i, g_s(\mathbf{x}_i,\mathbf{w})) = \ell_i$, $\mathbf{v}_i$ as a scalar v_i, and $\mathbf{y}_i$ as an integer $y_i \in [1, m]$.

In the existing literature, alternating minimization (Csiszar, 1984), or its related variants, is commonly employed to minimize the training objective, e.g. in (Kumar et al., 2010; Ma et al., 2017a; Jiang et al., 2014). This is an algorithmic paradigm where $\mathbf{w}$ and $\mathbf{v}$ are alternatively minimized, one at a time while the other is held fixed. When $\mathbf{v}$ is fixed, the weighted loss is typically minimized by stochastic gradient descent. When $\mathbf{w}$ is fixed, we compute $\mathbf{v}^k = \arg\min_{\mathbf{v}} \mathbb{F}(\mathbf{v}^{k-1}, \mathbf{w}^k)$ using the most recently updated $\mathbf{w}^k$ at epoch k. For example, Kumar *et al.* (2010) employed $G(\mathbf{v}) = -\lambda\|\mathbf{v}\|_1$. When $\mathbf{w}$ is fixed, the optimal $\mathbf{v}$ can be easily derived by:

$$v_i^* = \mathbb{1}(\ell_i \leq \lambda), \forall i \in [1, n], \quad (2)$$

where $\mathbb{1}$ is the indicator function. Eq. (2) intuitively explains the predefined curriculum in (Kumar et al., 2010), known as self-paced learning. First, when updating $\mathbf{v}$ with a fixed $\mathbf{w}$, a sample of smaller loss than the threshold λ is treated as an "easy" sample, and will be selected in training ($v_i^* = 1$). Otherwise, it will not be selected ($v_i^* = 0$). Second, when updating $\mathbf{w}$ with a fixed $\mathbf{v}$, the classifier is trained only on the selected "easy" samples. The hyperparameter λ controls the learning pace and corresponds to the "age" of the model. When λ is small, only samples of small loss will be considered. As λ grows, more samples of larger loss will be gradually added to train a more "mature" model.

As shown, the function G specifies a curriculum, *i.e.*, a sequence of samples with their corresponding weights to be used in training. When $\mathbf{w}$ is fixed, its optimal solution, *e.g.* Eq. (2), computes the time-varying weight that controls the timing and attention to learn every sample. Recent studies discovered multiple predefined curriculums and verified them in many real-world applications, *e.g.*, in (Fan et al., 2017; Ma et al., 2017a; Sangineto et al., 2016; Fan et al., 2017; Chang et al., 2017).

This paper studies learning curriculum from data. In the rest of this paper, Section 3 presents an approach to learn data-driven curriculum by *MentorNet*. Section 4 discusses an algorithm to optimize Eq. (1) using MentorNet and StudentNet together via mini-batch training.

3. Learning Curriculum from Data

Existing curriculums are either predetermined as an analytic expression of G or a function to compute sample weights. Such predefined curriculums cannot be adjusted accordingly, taking into account of the feedback from the student. This