Open DMQA Seminar

Deep Neural Networks with Noisy Labels

김상훈

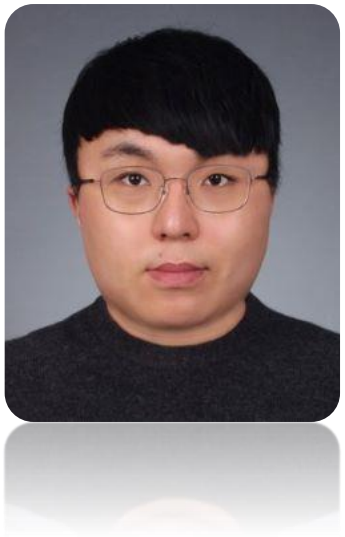
Data Mining & Quality Analytics Lab.

2022.08.28

고려대학교 KOREA UNIVERSITY

Data Mining Quality Analytics

발표자 소개



- 김상훈 (Sanghoon Kim)
  - ✓ 고려대학교 산업경영공학과
  - ✓ Data Mining & Quality Analytics Lab. (김성범 교수님)
  - ✓ Ph.D. Student (2019.09 ~ Present)
- Research Interest
  - ✓ Open Set Recognition / Curriculum Learning
  - ✓ Label Noise Learning
- Contact
  - ✓ E-mail : dawonksh@korea.ac.kr

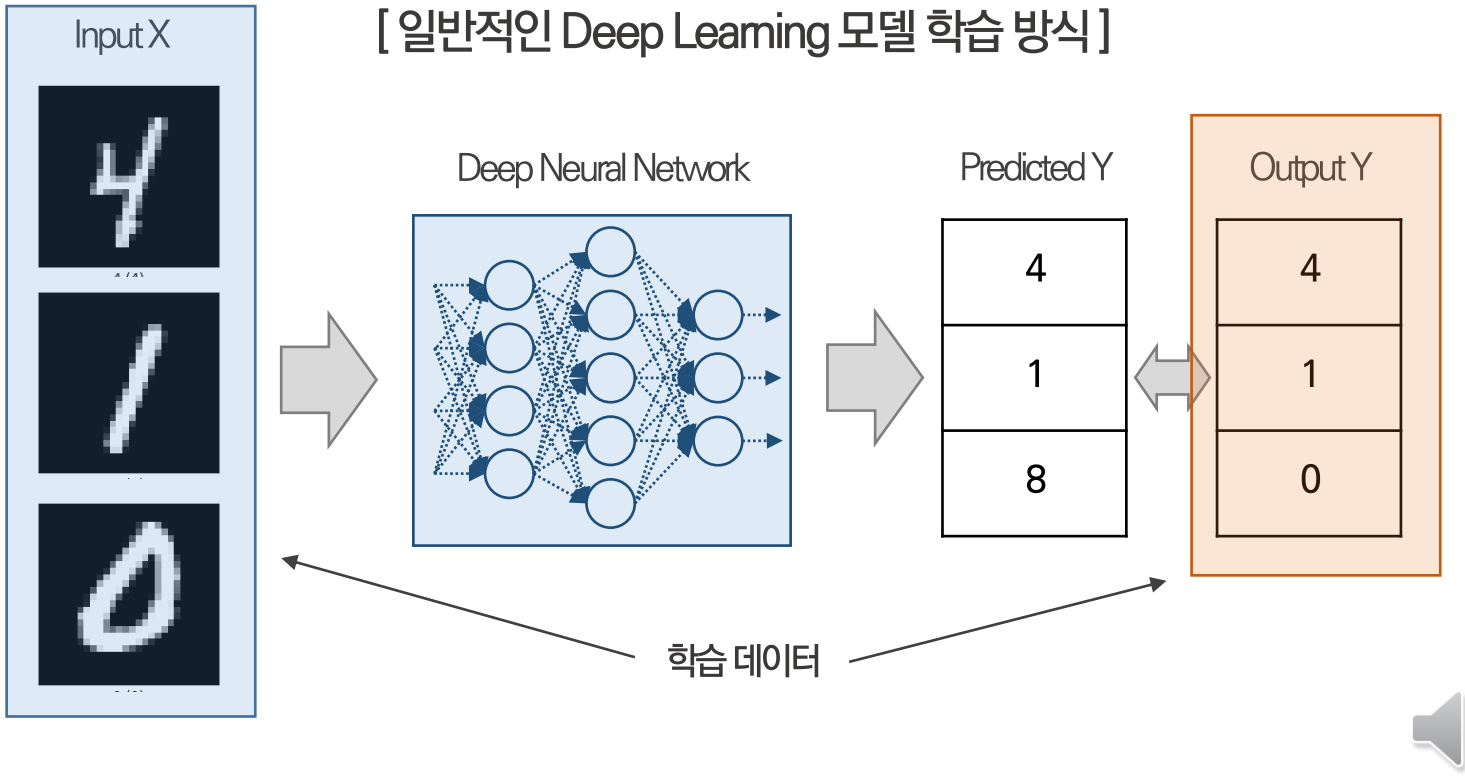
목차

Introduction

Supervised Learning Training Data

❖ 일반적으로 Supervised Learning에서는 학습데이터 모두 정확한 Label이 붙어있음을 가정

- Labeling 작업이 모두 완벽하게 이루어진 정제된 실험 데이터를 이용하여 입력 이미지와 레이블의 관계를 모델이 학습
- MNIST, CIFAR-10, ImageNet 등의 정제된 실험 데이터 활용



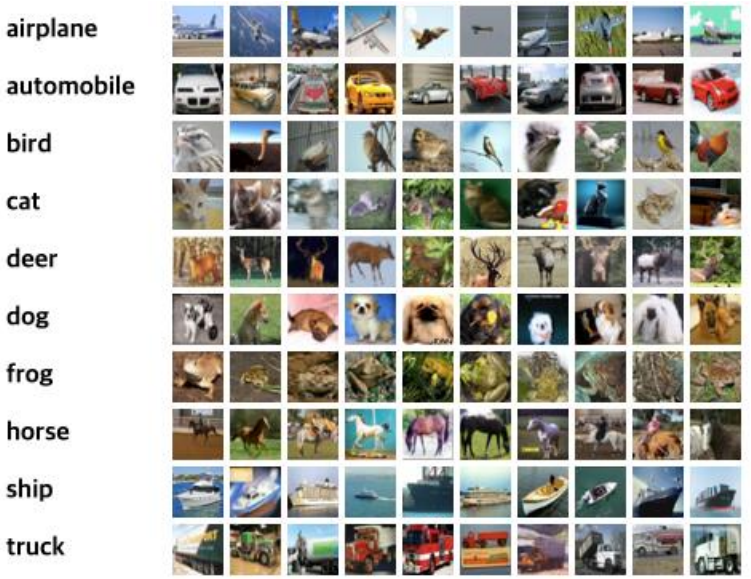
Introduction

Supervised Learning Training Data

❖ 일반적으로 Supervised Learning에서는 학습데이터 모두 정확한 Label이 붙어있음을 가정

- Labeling 작업이 모두 완벽하게 이루어진 정제된 실험 데이터를 이용하여 입력 이미지와 레이블의 관계를 모델이 학습
- MNIST, CIFAR-10, ImageNet 등의 정제된 실험 데이터 활용

MNIST Dataset



CIFAR-10 Dataset

Label이 모두 정확하게 매겨진 학습 데이터를 구축하기 위해서는,
Data Labeling 작업에 많은 비용이 필요

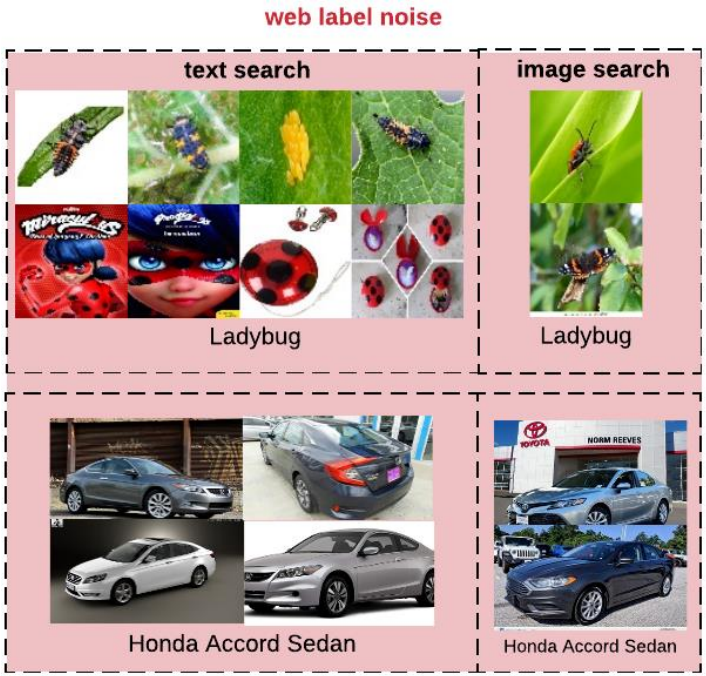
Introduction

Noisy Label Problem – Real World Data

❖ **현실에 존재하는 Label이 정제되지 않은 방대한 양의 데이터가 존재**

- 웹 검색을 통해 수집되는 데이터
- 비전문가에 의해 Label이 매겨진 신뢰도가 낮은 데이터

일반 사용자를 통한 대표적인 Labeling 방법



웹검색을 통해 수집되는 데이터

Labeling 작업의 신뢰도가 낮기에 학습 데이터 내에 Noisy Label이 다수 포함

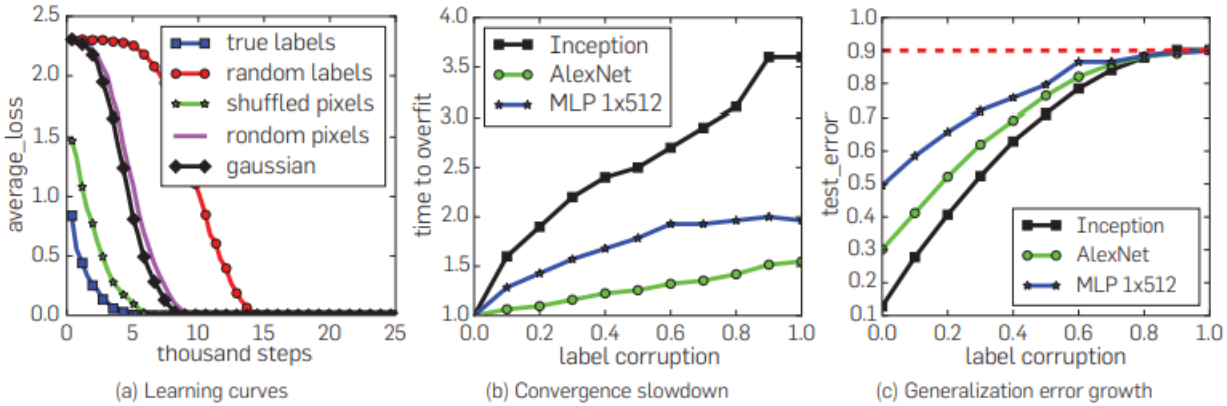
Introduction

Noisy Label로 인한 일반화 성능 저하

❖ Understanding Deep Learning Requires Rethinking Generalization, ICLR 2017

- Noisy Label이 학습데이터에 포함된 경우, 학습 수렴속도가 느림
- 학습 수렴속도가 느리더라도 학습이 진행될수록 모델을 수렴 -> 잘못된 Noisy Label 정보에 Overfitting

Figure 1. Fitting random labels and random pixels on CIFAR10. (a) The training loss of various experiment settings decaying with the training steps. (b) The relative convergence time with different label corruption ratio. (c) The test error (also the generalization error since training error is 0) under different label corruptions.



Deep Neural Network는 학습 데이터의 정보를 모두 학습하기에,
잘못된 Noisy Label 데이터의 정보까지 모두 학습

Introduction

Noisy Label로 인한 일반화 성능 저하

- Understanding Deep Learning Requires Rethinking Generalization, ICLR 2017
  - Noisy Label이 학습데이터에 포함된 경우, 학습 수렴속도가 느림
  - 학습 수렴속도가 느리더라도 학습이 진행될수록 모델을 수렴 → 잘못된 Noisy Label 정보에 Overfitting

금주 세미나 주제 :

학습데이터에 Noisy Label 데이터가 포함되었을 때,

어떻게 하면 모델이 Noisy Label 데이터의 정보를 효과적으로 거르면서

일반화 성능은 높아지도록 학습할 수 있을까?

Deep Neural Network는 학습 데이터의 정보를 모두 학습하기에,
잘못된 Noisy Label 데이터의 정보까지 모두 학습

MentorNet

논문 소개

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

- 2018년 ICML (International Conference on Machine Learning)에 Lu Jiang가 발표한 논문
- 2022년 08월 28일 기준 인용 횟수 : 935회

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.

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,

MentorNet

논문 소개

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

- 2018년 ICML (International Conference on Machine Learning)에 Lu Jiang가 발표한 논문
- 2022년 08월 28일 기준 인용 횟수 : 935회

Algorithm 1 SPADE for minimizing Eq. (1)

Input : Dataset $\mathcal{D}$, a predefined G or a learned $g_m(\cdot;\Theta)$
Output : The model parameter $\mathbf{w}$ of StudentNet.

Initialize $\mathbf{w}^0, \mathbf{v}^0, t = 0$
while *Not Converged* **do**
  Fetch a mini-batch Ξ_t uniformly at random
  For every $(\mathbf{x}_i, y_i)$ in Ξ_t compute $\phi(\mathbf{x}_i, y_i, \mathbf{w}^t)$
  if *update curriculum* **then**
    $\Theta \leftarrow \Theta^*$, where Θ^* is learned in Sec. 3.1
  end
  if G *is used* **then**
    $\mathbf{v}_\Xi^t \leftarrow \mathbf{v}_\Xi^{t-1} - \alpha_t \nabla_{\mathbf{v}} \mathbb{F}(\mathbf{w}^{t-1}, \mathbf{v}^{t-1})|_{\Xi_t}$
  end
  else $\mathbf{v}_\Xi^t \leftarrow g_m(\phi(\Xi_t, \mathbf{w}^{t-1}); \Theta)$;
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \alpha_t \nabla_{\mathbf{w}} \mathbb{F}(\mathbf{w}^{t-1}, \mathbf{v}^t)|_{\Xi_t}$
  $t \leftarrow t + 1$
end
return $\mathbf{w}^t$

MentorNet Algorithm

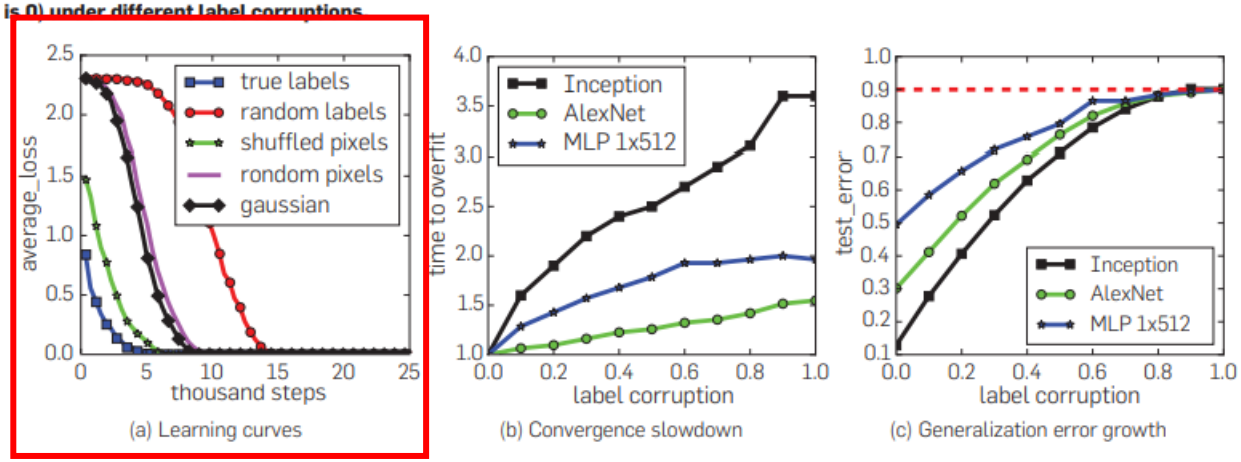
MentorNet

Method

❖ [1] Deep Neural Network는 전체 데이터 중 쉬운 패턴의 데이터를 먼저 학습

- 초기 학습 단계에서 올바른 Label을 가진 데이터가 Noisy Label 데이터에 비해 빠르게 학습
- 올바른 Label을 가진 데이터가 Noisy Label 데이터보다는 Training Loss가 상대적으로 빨리 감소

Figure 1. Fitting random labels and random pixels on CIFAR10. (a) The training loss of various experiment settings decaying with the training steps. (b) The relative convergence time with different label corruption ratio. (c) The test error (also the generalization error since training error is 0) under different label corruptions.



데이터와 라벨에 일관성이 없어 상대적으로 어려운 패턴인,

Noisy Label 데이터가 올바른 Label 데이터에 비해 큰 Training Loss 값을 가짐

[1] Arpit, Devansh, et al. "A closer look at memorization in deep networks." International Conference on Machine Learning. PMLR, 2017.

MentorNet

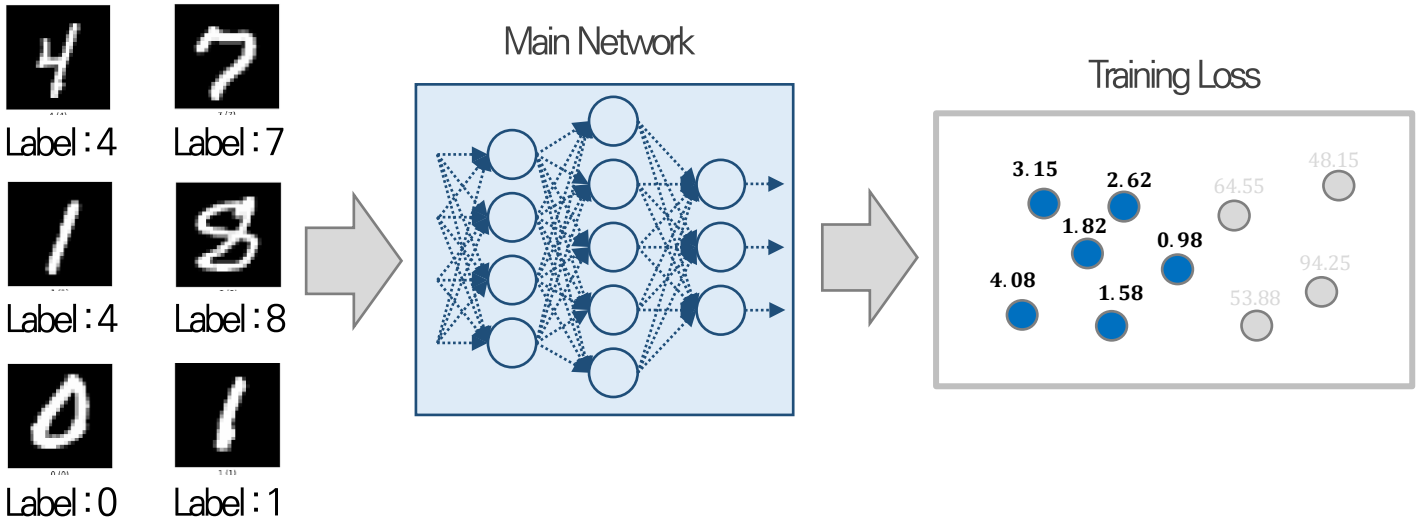
Method

❖ Training Loss 값에 대한 가정

- Training Loss 값이 클수록, 모델 학습에 방해되는 Noisy Label Data

〈Deep Neural Network 학습과정〉

올바른 Label 학습 데이터



데이터와 Label이 일관성이 존재하는 쉬운 패턴의 데이터이므로

올바른 Label이 매겨진 학습 데이터의 경우 Training Loss 값이 상대적으로 작음

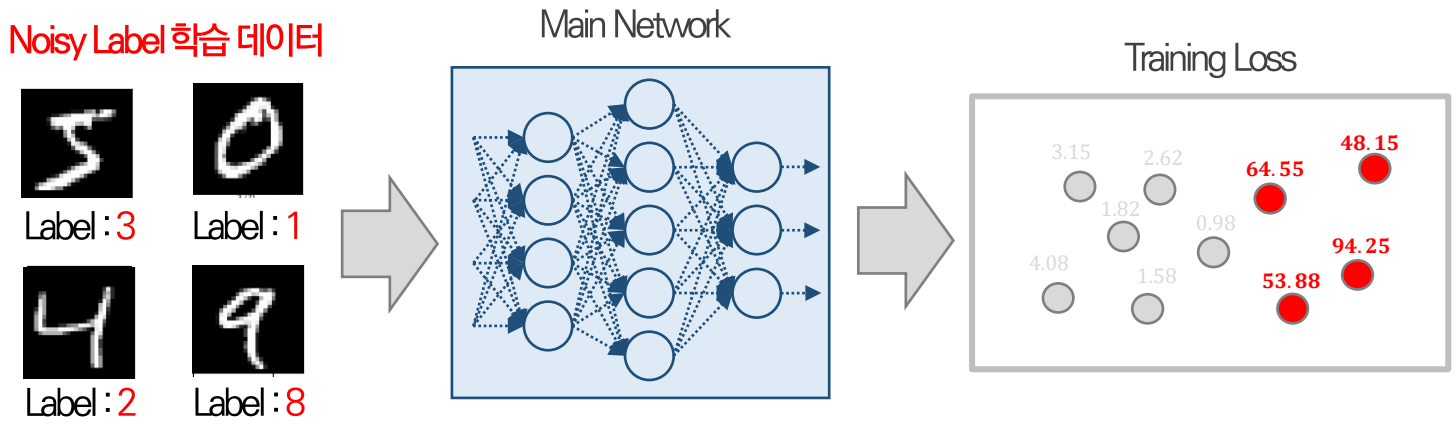
MentorNet

Method

❖ Training Loss 값에 대한 가정

- Training Loss 값이 클수록, 모델 학습에 방해되는 Noisy Label Data

〈Deep Neural Network 학습과정〉



모델 학습에 방해가 되는 샘플이므로

Noisy Label 학습 데이터는 상대적으로 큰 Training Loss 값을 가짐

MentorNet

Method

❖ Training Loss값에 따라 학습 반영 가중치를 조절

〈Self-pace Learning 전략〉

$$minimize\ L(\theta) = \frac{1}{N}\left(\sum_{i=1}^{N} v_i L_i(\theta) - \lambda \sum_{i=1}^{N} v_i\right) + r(\theta)$$

$$= \frac{1}{N}\left(\sum_{i=1}^{N} v_i \{L_i(\theta) - \lambda\}\right) + r(\theta)$$

: 각 단계에서 변수 v_i 와 θ에 대해 한 변수를 고정한 상태로 나머지 변수 최적화

- ✓ 첫 번째 학습 단계 : 변수 θ 고정, **변수 v_i 최적화**
- ✓ 두 번째 학습 단계 : 변수 v_i 고정, **변수 θ 최적화**

〈학습반영 가중치 최적화 v_i 〉

변수 v_i ($v_i = 0\ or\ 1$)는 현재 데이터 포인트 (x_i, y_i) 가 Noisy Label 데이터인지 여부 판단

$$\begin{cases} L_i(\theta) \geq \lambda \rightarrow v_i = 0 \\ L_i(\theta) < \lambda \rightarrow v_i = 1 \end{cases}$$

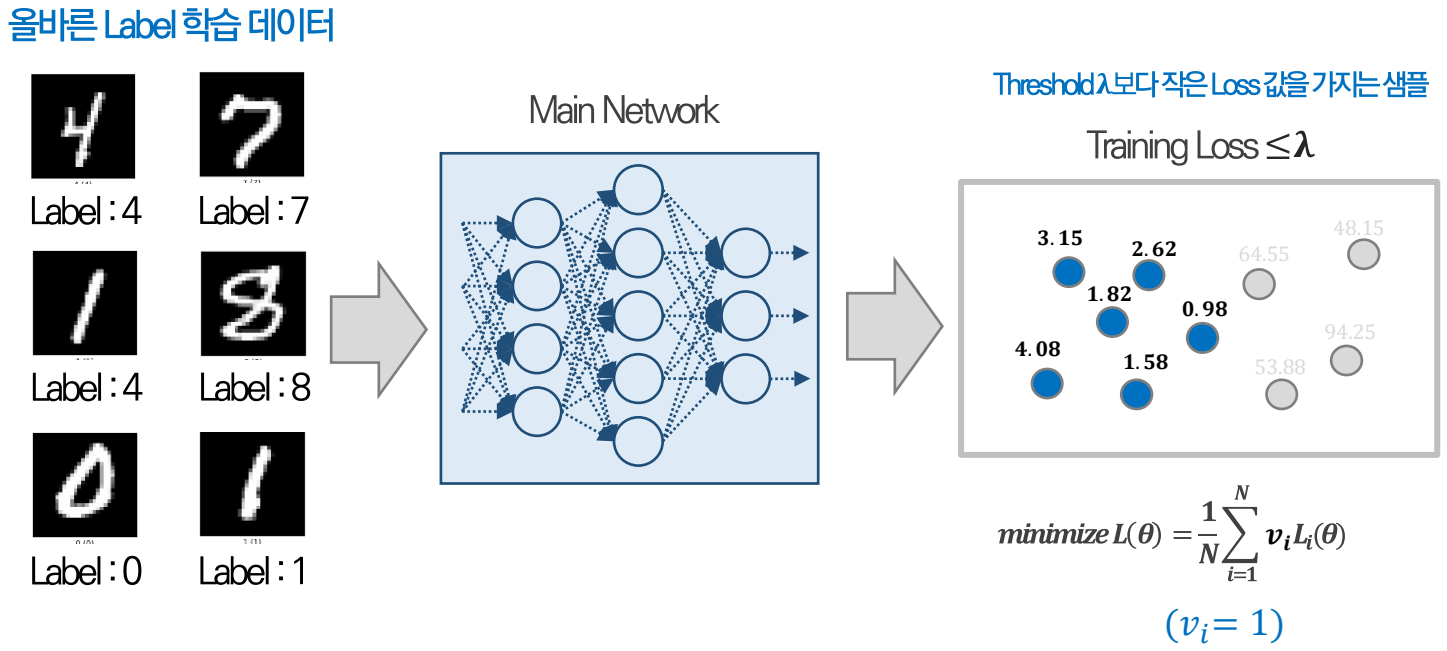
특정 Threshold λ 이하의 손실 값을 가지는 샘플들만 학습에 반영

MentorNet

Method

❖ **Training Loss 값에 따른 학습 반영 가중치 조절**

- Training Loss 값이 큰 샘플들을 Noisy Label 데이터로 가정하여 학습에서 배제
- Training Loss 값이 작은 샘플들만 올바른 Label이 매겨진 데이터로 가정하고 학습에 반영



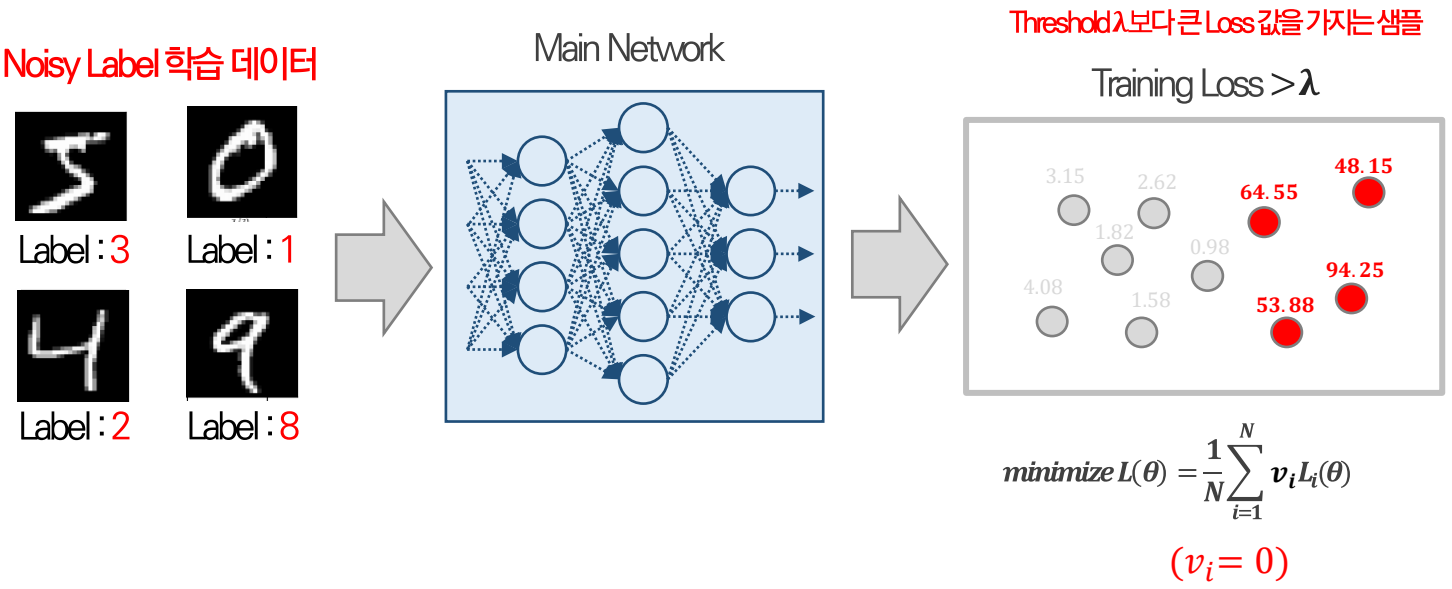
Training Loss 값이 Threshold λ 이하의 **올바른 Label이 매겨진 학습 데이터**만 모델 학습에 반영

MentorNet

Method

❖ **Training Loss 값에 따른 학습 반영 가중치 조절**

- Training Loss 값이 큰 샘플들을 Noisy Label 데이터로 가정하여 학습에서 배제
- Training Loss 값이 작은 샘플들만 올바른 Label이 매겨진 데이터로 가정하고 학습에 반영



Training Loss 값이 Threshold λ보다 큰 Noisy Label 데이터는 모델 학습에 미반영

MentorNet

Experiments

❖ 실험 데이터 : CIFAR-10N, CIFAR-100N

- 모든 라벨이 정확히 매겨진 CIFAR-10 & CIFAR-100에 의도적인 조작을 가해 Noisy Label 데이터 생성
- MentorNet의 경우 Noisy Label 데이터의 학습 반영을 줄이기에 오정보에 의한 성능 저하가 적음

Table 2. Comparison of validation accuracy on CIFAR-10 and CIFAR-100 under different noise fractions.

	Resnet-101 StudentNet						Inception StudentNet					
	CIFAR-100			CIFAR-10			CIFAR-100			CIFAR-10		
Method	0.2	0.4	0.8	0.2	0.4	0.8	0.2	0.4	0.8	0.2	0.4	0.8
FullModel	0.60	0.45	0.08	0.82	0.69	0.18	0.43	0.38	0.15	0.76	0.73	0.42
Forgetting	0.61	0.44	0.16	0.78	0.63	0.35	0.42	0.37	0.17	0.76	0.71	0.44
Self-paced	0.70	0.55	0.13	0.89	0.85	0.28	0.44	0.38	0.14	**0.80**	0.74	0.33
Focal Loss	0.59	0.44	0.09	0.79	0.65	0.28	0.43	0.38	0.15	0.77	0.74	0.40
Reed Soft	0.62	0.46	0.08	0.81	0.63	0.18	0.42	0.39	0.12	0.78	0.73	0.39
MentorNet PD	0.72	0.56	0.14	0.91	0.77	0.33	0.44	0.39	0.16	0.79	0.74	0.44
MentorNet DD	**0.73**	**0.68**	**0.35**	**0.92**	**0.89**	**0.49**	**0.46**	**0.41**	**0.20**	0.79	**0.76**	**0.46**

[Noisy Label Data 비율에 따른 성능]

MentorMix

논문 소개

❖ Beyond synthetic noise: Deep learning on controlled noisy labels.

- 2020년 ICML (International Conference on Machine Learning)에 MentorNet 저자 Lu Jiang가 발표한 논문
- 2022년 08월 28일 기준 인용 횟수 : 91회

Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels

Lu Jiang [1] Di Huang [2] Mason Liu [3] Weilong Yang [1]

Abstract

Performing controlled experiments on noisy data is essential in understanding deep learning across noise levels. Due to the lack of suitable datasets, previous research has only examined deep learning on controlled synthetic label noise, and real-world label noise has never been studied in a controlled setting. This paper makes three contributions. First, we establish the first benchmark of controlled real-world label noise from the web. This new benchmark enables us to study the web label noise in a controlled setting for the first time. The second contribution is a simple but effective method to overcome both synthetic and real noisy labels. We show that our method achieves the best result on our dataset as well as on two public benchmarks (CIFAR and WebVision). Third, we conduct the largest study by far into understanding deep neural networks trained on noisy labels across different noise levels, noise types, network architectures, and training settings. The data and code are released at the following link http://www.lujiang.info/cnlw.html.

However, due to the lack of suitable datasets, previous work has only examined DNNs on controlled synthetic label noise, and real-world label noise has never been studied in a controlled setting. This leads to two major issues. First, as synthetic noise is generated from an artificial distribution, a tiny change in the distribution may lead to inconsistent or even contradictory findings. For example, contrary to the common understanding that DNNs trained on synthetic noisy labels generalize poorly (Zhang et al., 2017), Rolnick et al. (2017) showed that DNNs can be robust to massive label noise when the noise distribution is made slightly different. Due to the lack of datasets, these findings, unfortunately, have not yet been verified beyond synthetic noise in a controlled setting. Second, the vast majority of previous studies prefer to verify robust learning methods on a spectrum of noise levels because the goal of these methods is to overcome a wide range of noise levels. However, current evaluations are limited because they are conducted only on synthetic label noise. Although there do exist datasets of real label noise *e.g.* WebVision (Li et al., 2017a), Clothing-1M (Xiao et al., 2015), *etc*, they are not suitable for controlled evaluation in which a method must be systematically verified on multiple different noise levels, because the training images in these datasets are not manually labeled and hence their data noise level is fixed and unknown.

MentorMix

논문 소개

❖ **Beyond synthetic noise: Deep learning on controlled noisy labels.**

- 2020년 ICML (International Conference on Machine Learning)에 MentorNet 저자 Lu Jiang가 발표한 논문
- 2022년 08월 28일 기준 인용 횟수 : 91회

MentorMix

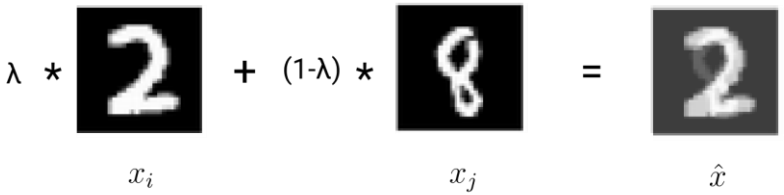
= MentorNet + Mix-up

MentorMix

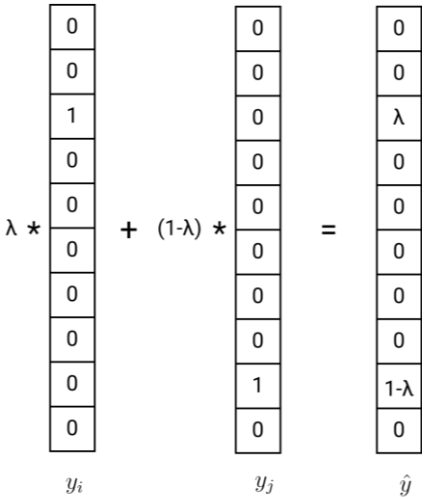
논문 소개

❖ mixup: Beyond empirical risk minimization

- Deep Neural Network의 Memorization 문제와 Adversarial Examples 문제를 해결하기 위한 Data Augmentation 기법
- Mixup 기법을 활용하여 증강된 데이터로 모델을 학습하면, Noisy Label에 Overfitting되는 문제를 해결



〈Input Mixup〉



〈Label Mixup〉

Mixup을 통해 새로운 데이터를 생성하여 학습 데이터 증강

MentorMix

논문 소개

❖ mixup: Beyond empirical risk minimization

- Deep Neural Network의 Memorization 문제와 Adversarial Examples 문제를 해결하기 위한 Data Augmentation 기법
- Mixup 기법을 활용하여 증강된 데이터로 모델을 학습하면, Noisy Label에 Overfitting되는 문제를 해결

Label corruption	Method	Test error		Training error	
		Best	Last	Real	Corrupted
20%	ERM	12.7	16.6	0.05	0.28
	ERM + dropout ($p = 0.7$)	8.8	10.4	5.26	83.55
	mixup ($\alpha = 8$)	**5.9**	6.4	2.27	86.32
	mixup + dropout ($\alpha = 4, p = 0.1$)	6.2	**6.2**	1.92	85.02
50%	ERM	18.8	44.6	0.26	0.64
	ERM + dropout ($p = 0.8$)	14.1	15.5	12.71	86.98
	mixup ($\alpha = 32$)	11.3	12.7	5.84	85.71
	mixup + dropout ($\alpha = 8, p = 0.3$)	**10.9**	**10.9**	7.56	87.90
80%	ERM	36.5	73.9	0.62	0.83
	ERM + dropout ($p = 0.8$)	30.9	35.1	29.84	86.37
	mixup ($\alpha = 32$)	25.3	30.9	18.92	85.44
	mixup + dropout ($\alpha = 8, p = 0.3$)	**24.0**	**24.8**	19.70	87.67

Table 2: Results on the corrupted label experiments for the best models.

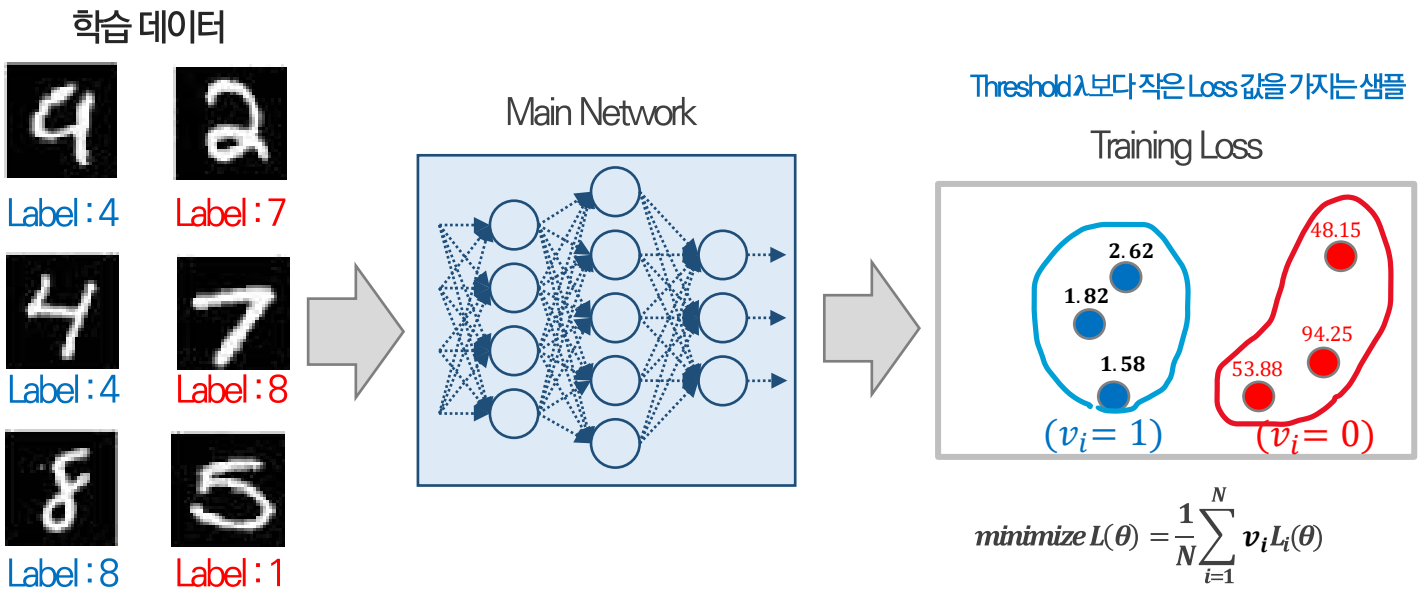
모델이 Noisy Label 정보에 Overfitting되는 현상을 방지하여 모델 일반화 성능 확보

MentorMix

Method

❖ MentorMix = MentorNet + Mixup

- Training Loss 값을 Threshold 기준으로 학습 반영 여부 결정

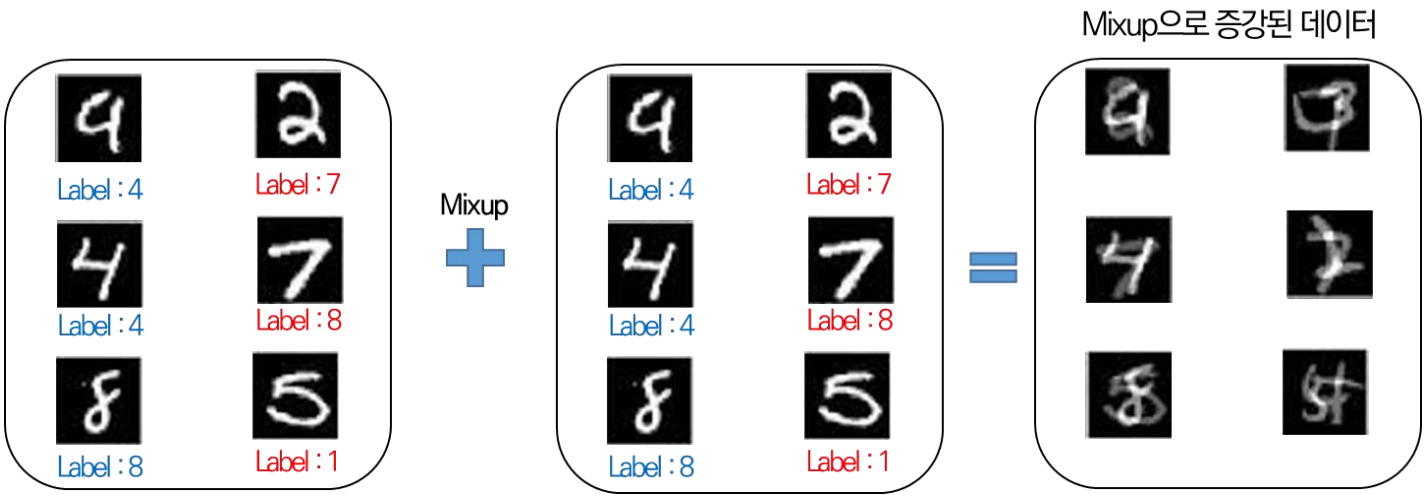


MentorMix

Method

❖ MentorMix = MentorNet + Mixup

- 배치 안의 데이터 중 mixup 적용할 데이터를 일부 샘플링 (학습 반영 데이터 비율이 높게)
- 올바른 라벨 데이터와 노이지 라벨 데이터 mixup 시에는 mixup 가중치를 학습 반영 데이터에 더 높게 부여
- Loss 값 크기로 학습반영 여부를 결정하면, Noisy Label 데이터임에도 Loss 값 크기가 작아 학습에 반영될 수 있음



Mixup을 통해 학습에 반영된 Noisy Label 데이터의 영향력을 줄이고,

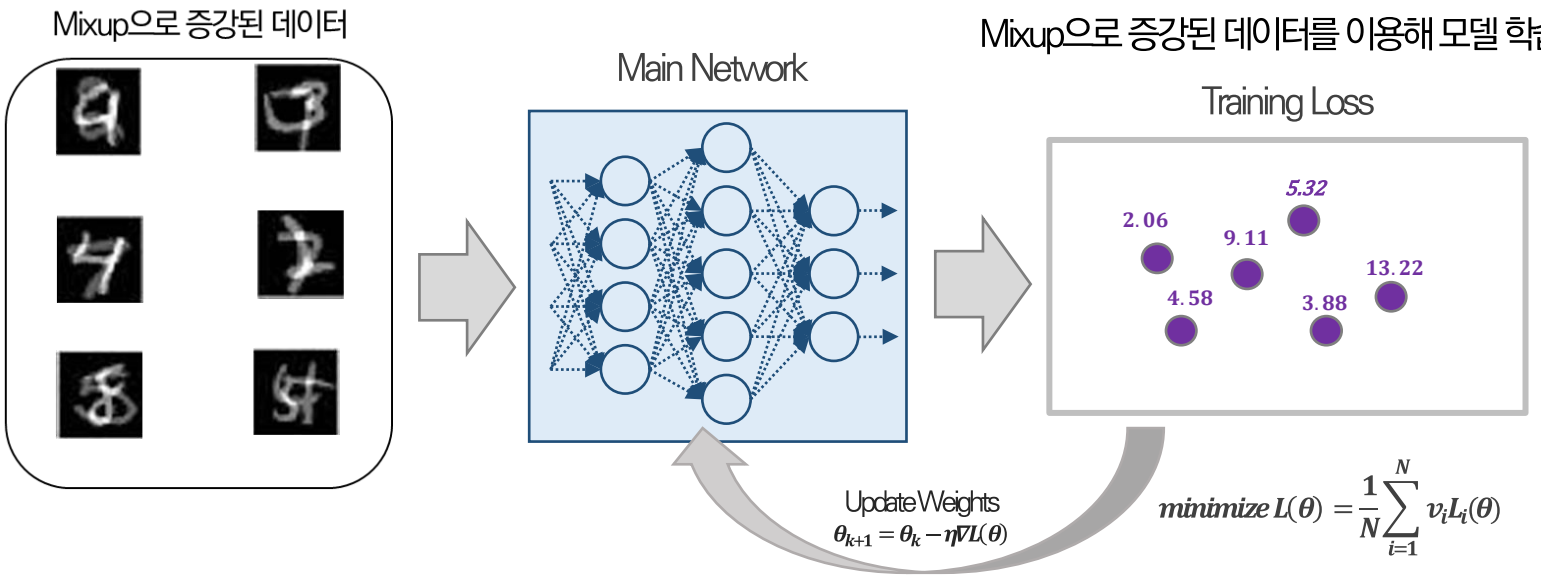
학습 샘플 수 증강을 통해 모델 일반화 성능을 확보할 수 있다.

MentorMix

Method

❖ MentorMix = MentorNet + Mixup

- 배치 안의 데이터 중 mixup 적용할 데이터를 일부 샘플링 (학습 반영 데이터 비율이 높게)
- 올바른 라벨 데이터와 노이지 라벨 데이터 mixup 시에는 mixup 가중치를 학습 반영 데이터에 더 높게 부여
- Loss 값 크기로 학습반영 여부를 결정하면, Noisy Label 데이터임에도 Loss 값 크기가 작아 학습에 반영될 수 있음



Mixup을 통해 학습에 반영된 Noisy Label 데이터의 영향력을 줄이고,

학습 샘플 수 증강을 통해 모델 일반화 성능을 확보할 수 있다.

MentorMix

Experiments

❖ 실험 데이터 : Mini-ImageNet & Stanford Cars

- 공개된 Mini ImageNet Dataset과 Stanford Cars 데이터에 있는 레이블을 웹에서 검색하여 데이터 수집하여 구성
- Google Cloud Labeling Service를 사용하여 주어진 웹 레이블이 올바른지 여부를 식별

Google AI Blog : Understanding Deep Learning on Controlled Noisy Labels

MentorMix

Experiments

❖ 실험 데이터 : Mini-ImageNet & Stanford Cars

- 공개된 Mini ImageNet Dataset과 Stanford Cars 데이터에 있는 레이블을 웹에서 검색하여 데이터 수집하여 구성
- Google Cloud Labeling Service를 사용하여 주어진 웹 레이블이 올바른지 여부를 식별

Table 2. Peak accuracy (%) of the best trial of each method averaged across 10 noise levels. – denotes the method failed to train.

Method	Mini-ImageNet				Stanford Cars			
	Fine-tuned		Trained from scratch		Fine-tuned		Trained from scratch	
	Blue	Red	Blue	Red	Blue	Red	Blue	Red
Vanilla	82.3±1.9	81.6±1.9	58.3±10.3	64.9±5.2	70.0±16.8	82.4±6.9	53.8±24.4	77.7±10.4
WeightDecay	81.9±1.8	81.5±1.8	—	—	72.2±17.5	84.3±6.6	—	—
Dropout	82.8±1.3	81.8±1.8	59.3±9.5	65.7±5.0	71.7±16.9	83.8±6.6	62.8±23.5	84.1±6.7
S-Model	82.3±1.8	82.0±1.9	58.7±10.2	64.6±5.1	69.7±16.8	82.4±7.1	53.9±23.5	77.6±10.2
Bootstrap	83.1±1.6	82.7±1.8	60.1±9.7	65.5±4.9	71.7±16.9	82.8±6.7	55.6±23.9	78.9±9.6
Mixup	81.7±1.8	82.4±1.7	60.7±9.8	66.0±4.9	73.1±16.6	85.0±6.2	64.2±21.6	82.5±8.0
MentorNet	82.9±1.7	82.4±1.7	61.8±10.3	65.1±5.0	75.9±16.8	82.6±6.6	56.8±23.1	78.9±8.9
Our MentorMix	**84.2±0.7**	**83.3±1.9**	**70.9±3.4**	**67.0±5.0**	**78.2±16.2**	**86.9±5.5**	**67.7±23.0**	**83.6±7.5**

Mixup 적용만으로도 기존 MentorNet 대비 성능 향상

Co-teaching

논문 소개

❖ Co-teaching: Robust training of deep neural networks with extremely noisy labels

- 2018년 NeurIPS (Neural Information Processing Systems)에 Bo Han이 발표한 논문
- 2022년 08월 28일 기준 인용 횟수 : 1010회

Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels

Bo Han[*1,2], **Quanming Yao**[*3], **Xingrui Yu**[1], **Gang Niu**[2], **Miao Xu**[2], **Weihua Hu**[4], **Ivor W. Tsang**[1], **Masashi Sugiyama**[2,5]

[1]Centre for Artificial Intelligence, University of Technology Sydney; [2]RIKEN; [3]4Paradigm Inc.; [4]Stanford University; [5]University of Tokyo

Abstract

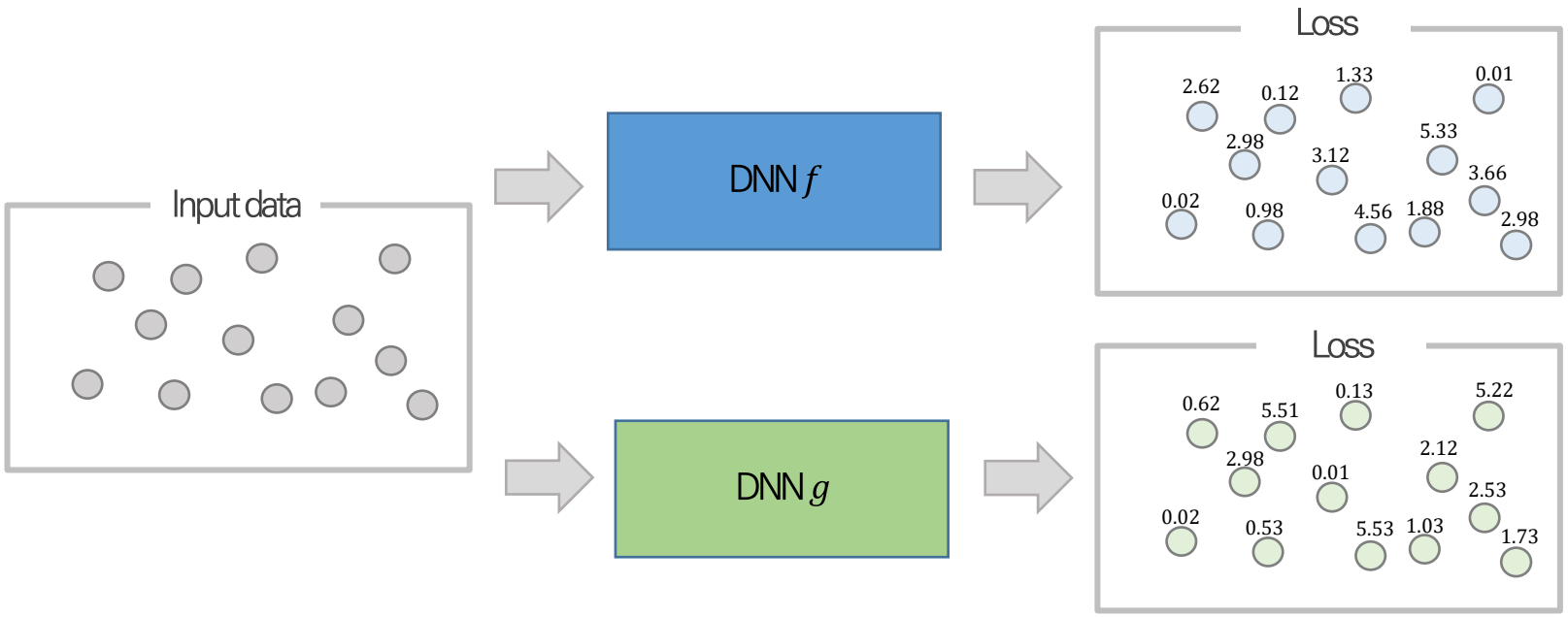
Deep learning with noisy labels is practically challenging, as the capacity of deep models is so high that they can totally memorize these noisy labels sooner or later during training. Nonetheless, recent studies on the *memorization effects* of deep neural networks show that they would first memorize training data of clean labels and then those of noisy labels. Therefore in this paper, we propose a new deep learning paradigm called "*Co-teaching*" for combating with noisy labels. Namely, we train two deep neural networks simultaneously, and let them *teach each other* given every mini-batch: firstly, each network feeds forward all data and selects some data of possibly clean labels; secondly, two networks communicate with each other what data in this mini-batch should be used for training; finally, each network back propagates the data selected by its peer network and updates itself. Empirical results on noisy versions of *MNIST*, *CIFAR-10* and *CIFAR-100* demonstrate that Co-teaching is much superior to the state-of-the-art methods in the robustness of trained deep models.

Co-teaching

논문 소개

❖ **2개의 모델을 동시에 Cross-update**

- 초기 값이 다른 두 모델 학습 시, 미니배치에서 상대적으로 Loss 값이 작은 데이터들을 일정 비율 만큼 선별
- 선별된 작은 Loss 값의 데이터들을 모델 간 교환 하여 학습(Cross-update)



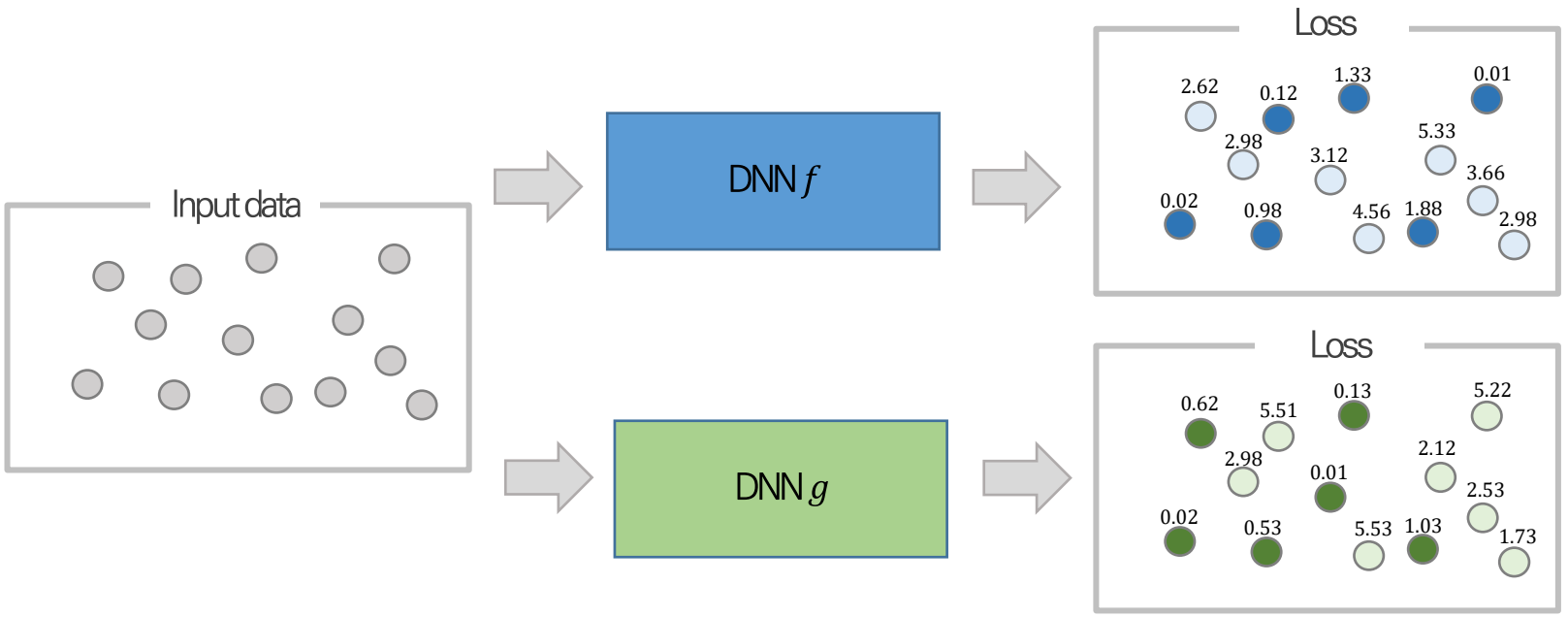
1. 같은 mini-batch에 대해 두 모델(모델f, 모델g) 이 각각 Loss 값 계산

Co-teaching

논문 소개

❖ **2개의 모델을 동시에 Cross-update**

- 초기 값이 다른 두 모델 학습 시, 미니배치에서 상대적으로 Loss 값이 작은 데이터들을 일정 비율 만큼 선별
- 선별된 작은 Loss 값의 데이터들을 모델 간 교환 하여 학습(Cross-update)



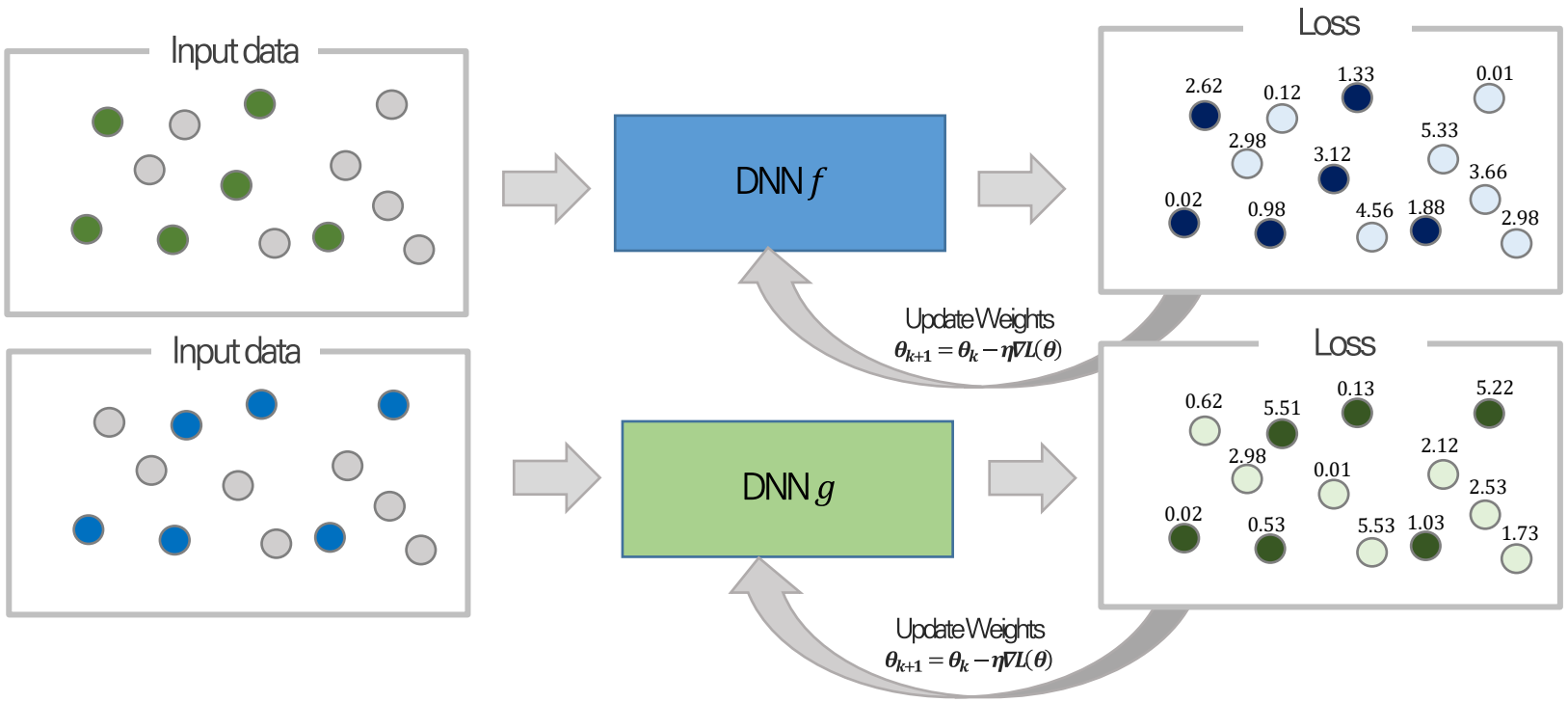
2. Loss 값이 작은 순으로 일정 비율 R(T)만큼 데이터 선별

Co-teaching

논문 소개

❖ 2개의 모델을 동시에 Cross-update

- 초기 값이 다른 두 모델 학습 시, 미니배치에서 상대적으로 Loss 값이 작은 데이터들을 일정 비율 만큼 선별
- 선별된 작은 Loss 값의 데이터들을 모델 간 교환 하여 학습(Cross-update)



3. 서로 다른 네트워크에서 선별된 small-loss instance들로 모델 업데이트

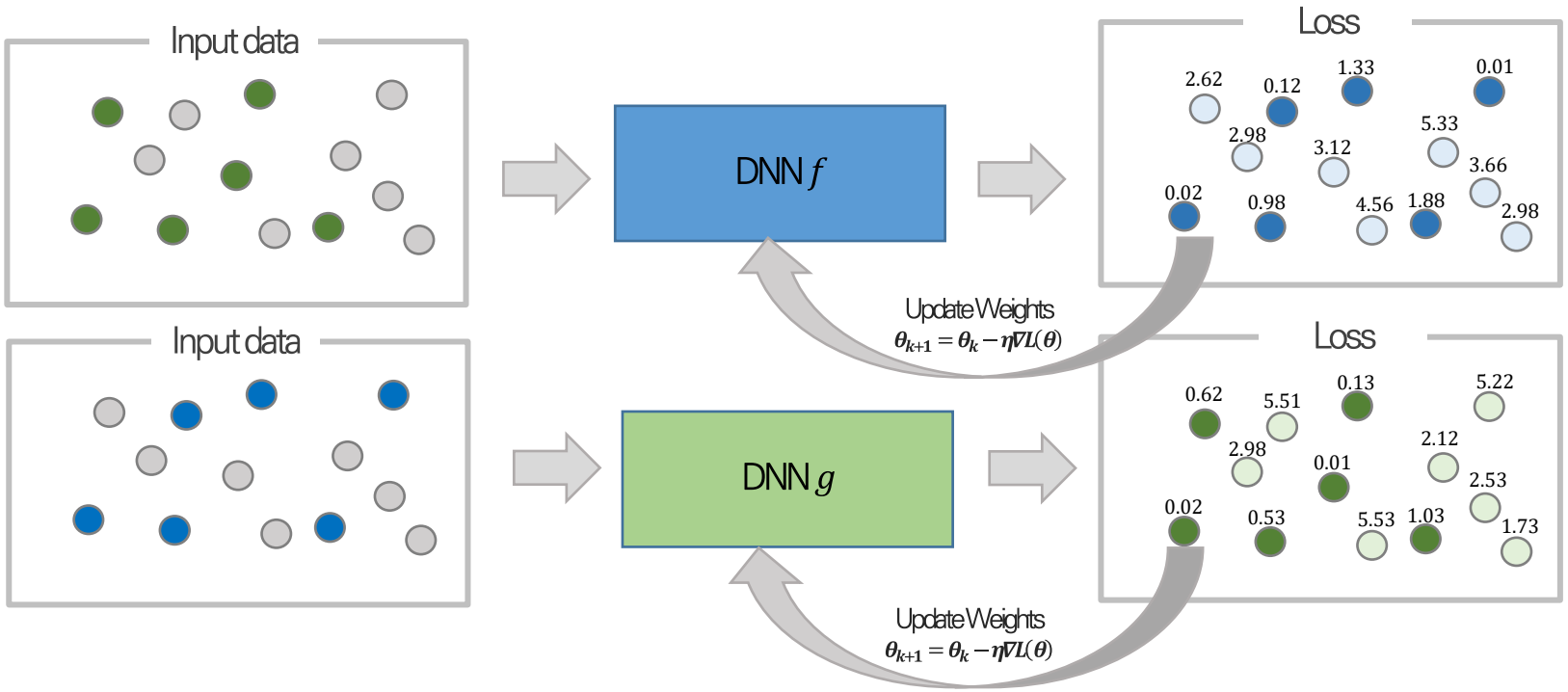
g모델의 small-loss instances들로 f모델이 업데이트, f모델의 small-loss instances들로 g모델이 업데이트,

Co-teaching

논문 소개

❖ 2개의 모델을 동시에 Cross-update

- 초기 값이 다른 두 모델 학습 시, 미니배치에서 상대적으로 Loss 값이 작은 데이터들을 일정 비율 만큼 선별
- 선별된 작은 Loss 값의 데이터들을 모델 간 교환 하여 학습(Cross-update)



1개 모델만을 사용하여 학습 시, 라벨이 올바르더라도 한번 학습에서 제외된 데이터는 추후 학습에서도 학습에서 제외
-> 2개의 모델을 사용함으로써, 특정 데이터에 Overfitting되는 것을 방지

Dividemix

논문 소개

❖ Dividemix: Learning with noisy labels as semi-supervised learning.

- 2020년 ICLR (International Conference on Learning Representations)에 발표된 논문
- 2022년 08월 28일 기준 인용 횟수 : 354회

DIVIDEMIX: LEARNING WITH NOISY LABELS AS SEMI-SUPERVISED LEARNING

Junnan Li, Richard Socher, Steven C.H. Hoi
Salesforce Research
{junnan.li,rsocher,shoi}@salesforce.com

ABSTRACT

Deep neural networks are known to be annotation-hungry. Numerous efforts have been devoted to reducing the annotation cost when learning with deep networks. Two prominent directions include learning with noisy labels and semi-supervised learning by exploiting unlabeled data. In this work, we propose DivideMix, a novel framework for learning with noisy labels by leveraging semi-supervised learning techniques. In particular, DivideMix models the per-sample loss distribution with a mixture model to dynamically divide the training data into a labeled set with clean samples and an unlabeled set with noisy samples, and trains the model on both the labeled and unlabeled data in a semi-supervised manner. To avoid confirmation bias, we simultaneously train two diverged networks where each network uses the dataset division from the other network. During the semi-supervised training phase, we improve the MixMatch strategy by performing label co-refinement and label co-guessing on labeled and unlabeled samples, respectively. Experiments on multiple benchmark datasets demonstrate substantial improvements over state-of-the-art methods. Code is available at https://github.com/LiJunnan1992/DivideMix.

Dividemix

논문 소개

❖ Dividemix: Learning with noisy labels as semi-supervised learning.

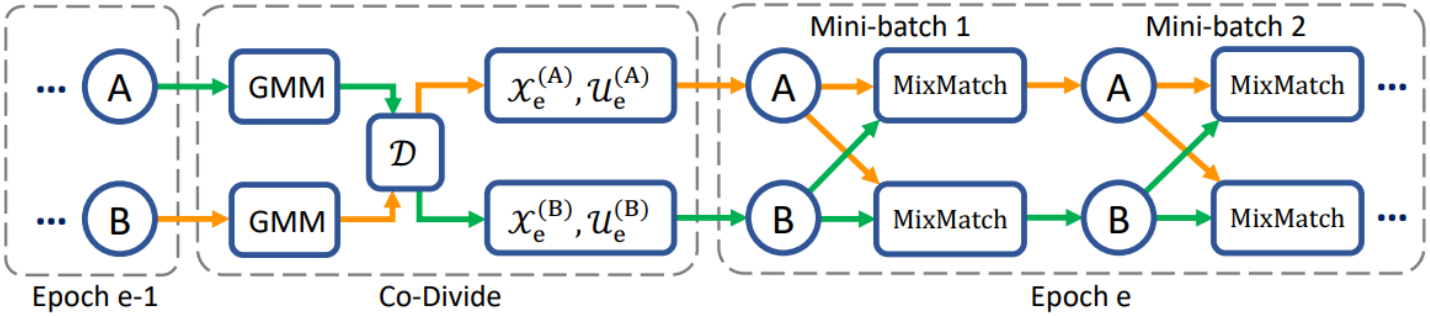
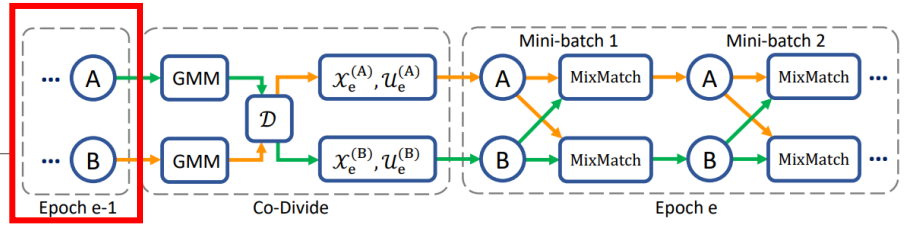


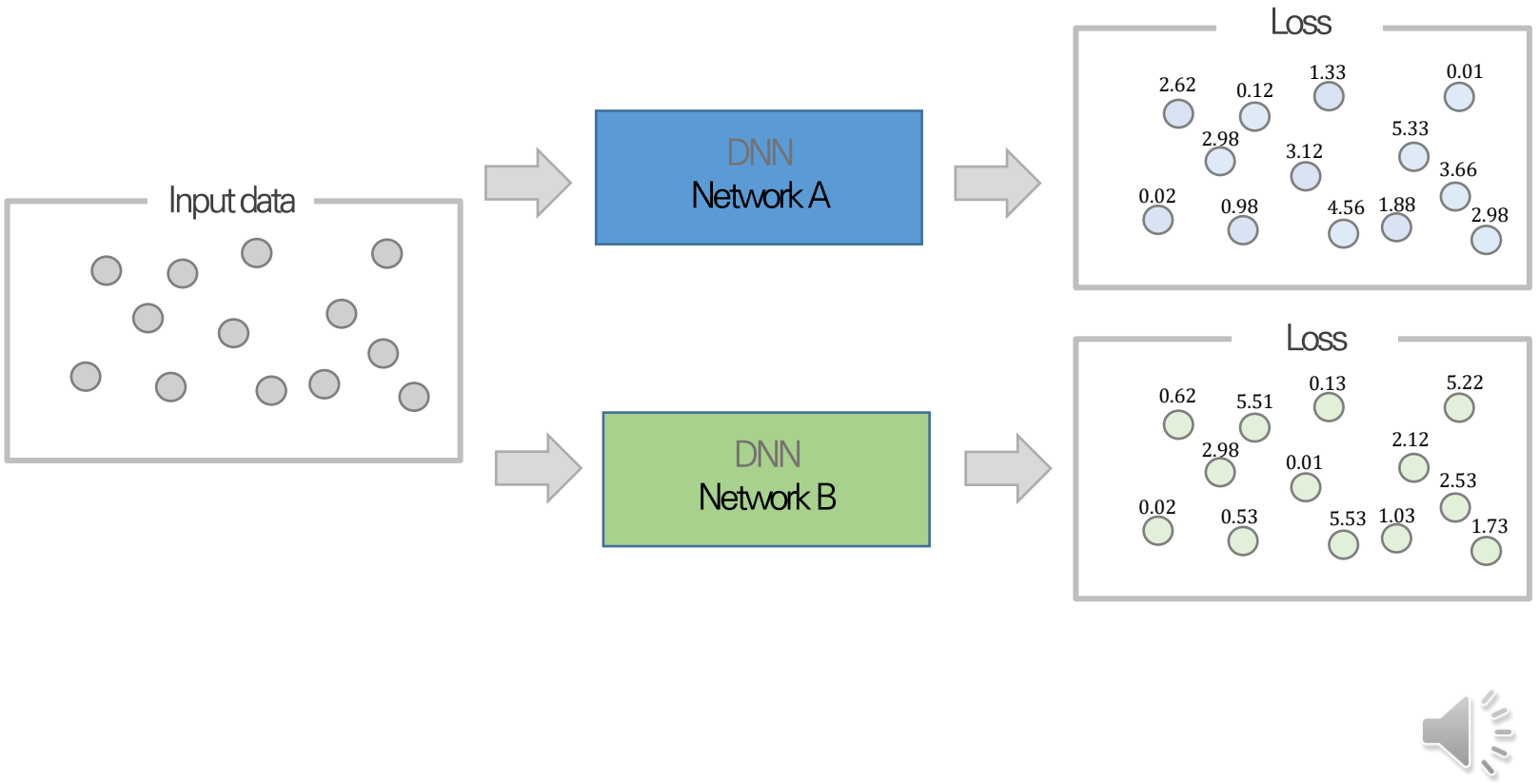
Figure 1: DivideMix trains two networks (A and B) simultaneously. At each epoch, a network models its per-sample loss distribution with a GMM to divide the dataset into a labeled set (mostly clean) and an unlabeled set (mostly noisy), which is then used as training data for the other network (*i.e.* co-divide). At each mini-batch, a network performs semi-supervised training using an improved MixMatch method. We perform label co-refinement on the labeled samples and label co-guessing on the unlabeled samples.

Dividemix

Method

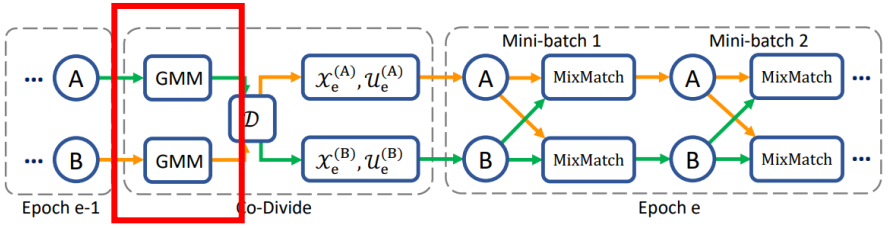


❖ STEP 1 : 동시에 2개의 네트워크를 학습하여 각 epoch마다 Loss 값을 산출



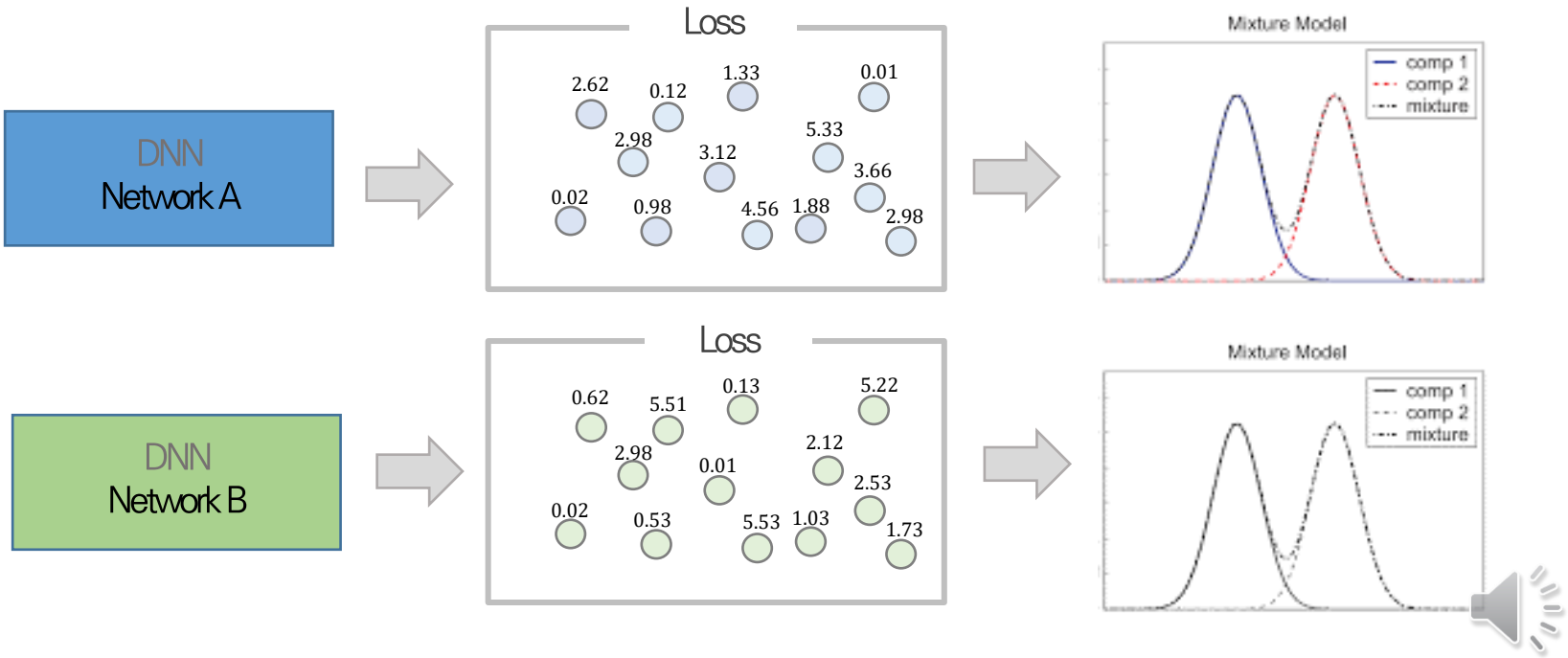
Dividemix

Method



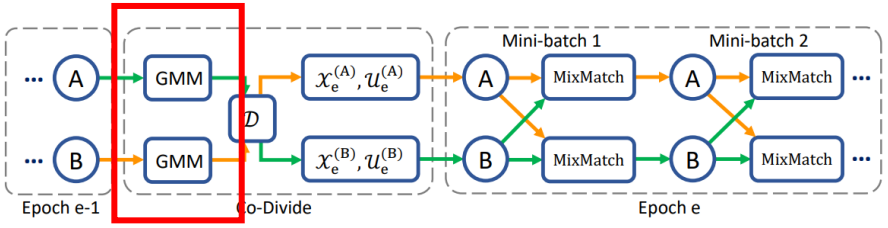
❖ **STEP2 :** 각 네트워크의 Loss 값을 이용해 각각 Gaussian Mixture Model을 Fitting

- Loss 값을 이용해 2개의 Gaussian Distribution으로 표현된 Gaussian Mixture Model Fitting
- Gaussian Mixture Model을 통해 Clean Data 분포에 속할 확률을 산출
- Clean Data 군집에 속할 확률이 Threshold τ보다 큰 경우, 올바르게 Labeling된 데이터로 판정



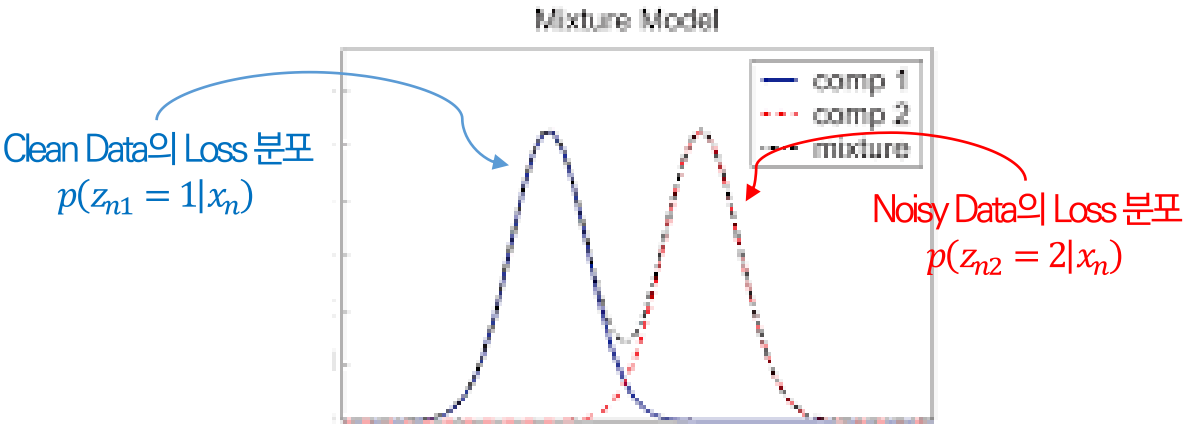
Dividemix

Method



❖ **STEP2 :** 각 네트워크의 Loss 값을 이용해 각각 Gaussian Mixture Model을 Fitting

- Loss 값을 이용해 2개의 Gaussian Distribution으로 표현된 Gaussian Mixture Model Fitting
- Gaussian Mixture Model을 통해 Clean Data 분포에 속할 확률을 산출
- Clean Data 군집에 속할 확률이 Threshold τ보다 큰 경우, 올바르게 Labeling된 데이터로 판정

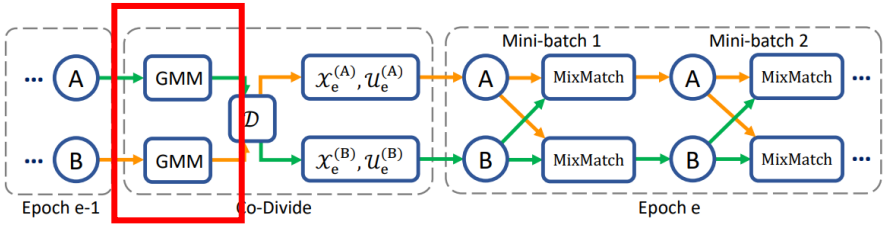


Clean Probability를 이용해 Noisy Data 판정

$p(z_{n1} = 1|x_n) \geq \tau$: Clean Data
$p(z_{n1} = 1|x_n) < \tau$: Noisy Data

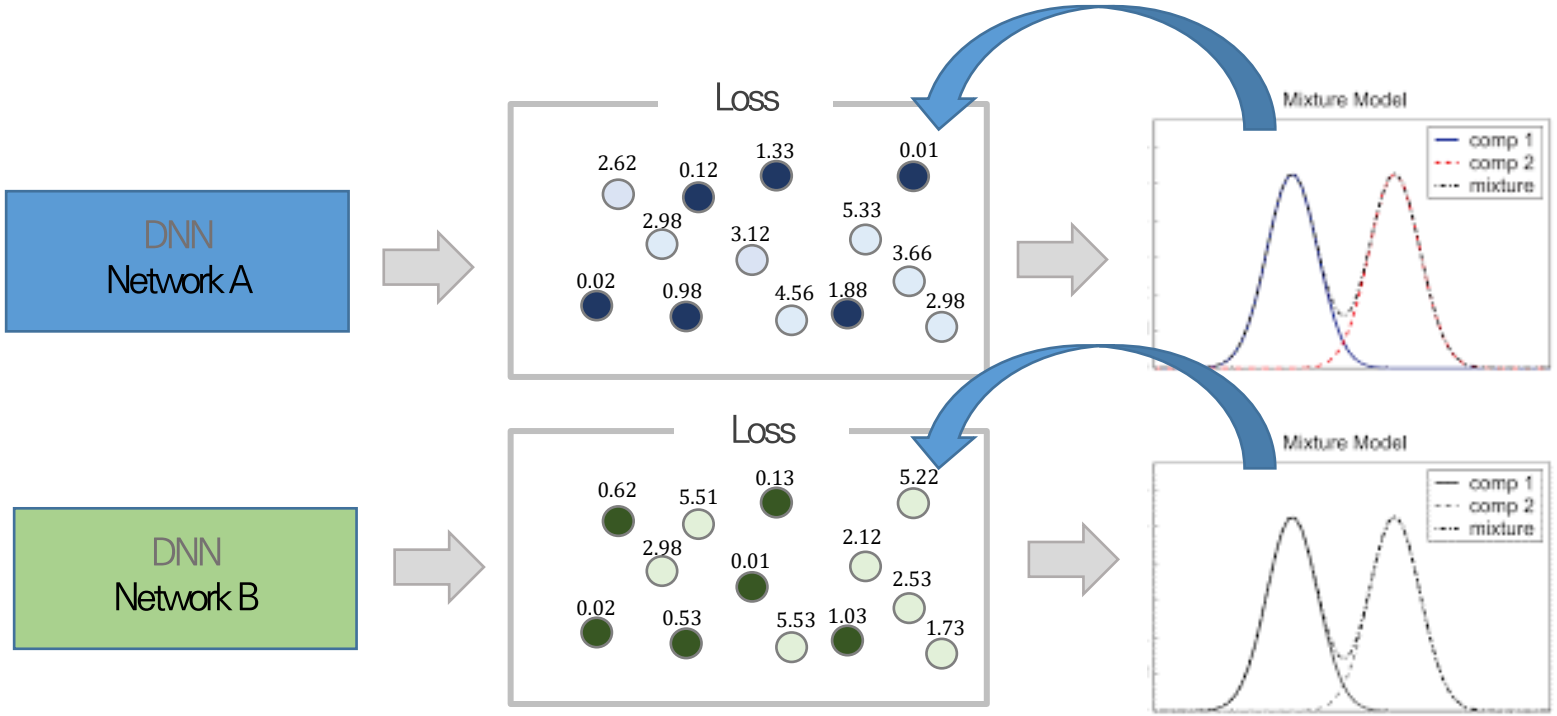
Dividemix

Method



❖ **STEP2 :** 각 네트워크의 Loss 값을 이용해 각각 Gaussian Mixture Model을 Fitting

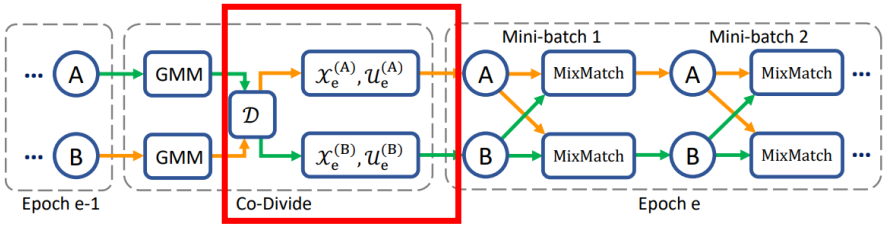
- Loss 값을 이용해 2개의 Gaussian Distribution으로 표현된 Gaussian Mixture Model Fitting
- Gaussian Mixture Model을 통해 Clean Data 분포에 속할 확률을 산출
- Clean Data 군집에 속할 확률이 Threshold τ보다 큰 경우, 올바르게 Labeling된 데이터로 판정



Clean Probability를 이용해 각각 Clean Data & Noisy Data 판정

Dividemix

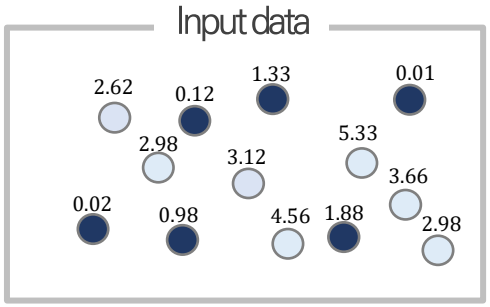
Method



❖ **STEP3 :** 서로 다른 네트워크에서 선별된 Clean Data & Noisy Data로 Mixup 데이터 구성

- GMM으로 선별된 데이터를 활용하여 서로 다른 네트워크의 업데이트에 활용
- GMM에 의해 Noisy Data로 판별된 데이터는 Unlabeled Data와 같이 취급

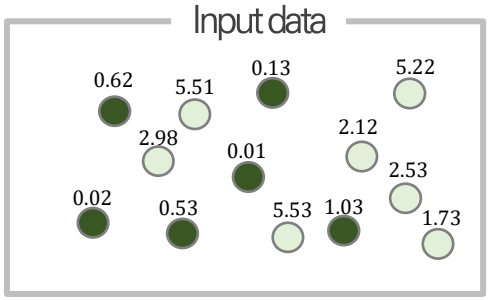
Network A



Network B 업데이트에 활용

- $X_e^{(B)}$: Network A의 GMM에 의해 Clean Data로 판정된 데이터
- $U_e^{(B)}$: Network A의 GMM에 의해 Noisy Data로 판정된 데이터

Network B

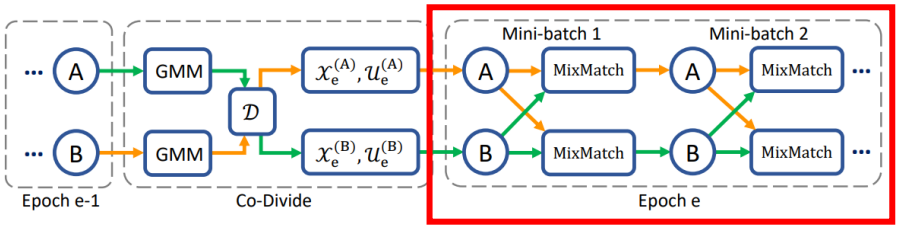


Network A 업데이트에 활용

- $X_e^{(A)}$: Network B의 GMM에 의해 Clean Data로 판정된 데이터
- $U_e^{(A)}$: Network B의 GMM에 의해 Noisy Data로 판정된 데이터

Dividemix

Method



❖ STEP4 : MixMatch 준비 과정 – Label 정제 및 구성

- $\{X_e^{(B)}, U_e^{(B)}\}$: Network A의 GMM에 의해 판정된 데이터 (Network B 업데이트에 활용)
- y_b : Ground Truth label
- p_b : Network's Prediction
- w_b : Clean Probability

Label Co-Refinement (Clean Data)

- Labeled sample : $\{(x_b, y_b, w_b); b \in (1, \ldots, B)\}$

$$\tilde{y}_b = w_b y_b + (1 - w_b) p_b$$

$$\tilde{y}_b = Sharpen(\tilde{y}_b, \mathrm{T})$$

실제 Label과 네트워크의 예측 결과를 이용해 Label 정제

Label Co-Guessing (Noisy Data)

- Unlabeled sample : $\{u_b; b \in (1, \ldots, B)\}$

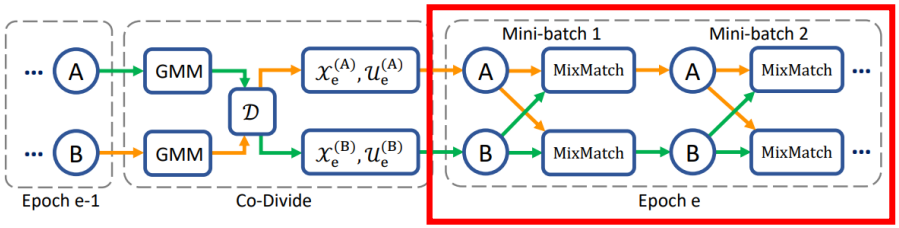
$$\tilde{q}_b = \frac{1}{2M} \sum_m \left(p_{model}(\hat{u}_{b,m}; \theta^A) + p_{model}(\hat{u}_{b,m}; \theta^B) \right)$$

$$\tilde{q}_b = Sharpen(\tilde{q}_b, \mathrm{T})$$

두 네트워크의 예측 결과의 평균을 이용하여
Unlabeled Data의 Label 구성

Dividemix

Method



❖ STEP5 : MixMatch & Network Training

number of augmentations M

- Clean Data $X_e^{(B)}$ 간의 Mixup을 통해 데이터 증강-> 증강된 $\hat{X}^{(B)} = \{(\hat{x}_{b,m}, \hat{y}_b); b \in (1,\ldots,B), m \in (1,\ldots,M)\}$ 확보
- Noisy Data $U_e^{(B)}$ 간의 Mixup을 통해 데이터 증강-> 증강된 $\hat{U}^{(B)} = \{(\hat{u}_{b,m}, q_b); b \in (1,\ldots,B), m \in (1,\ldots,M)\}$ 확보

Mixup을 통해 증강된 데이터를 이용해 B 네트워크 업데이트

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_u \mathcal{L}_{\mathcal{U}} + \lambda_r \mathcal{L}_{\text{reg}}.$$

$$\mathcal{L}_{\mathcal{X}} = -\frac{1}{|\mathcal{X}'|} \sum_{x,p \in \mathcal{X}'} \sum_c p_c \log(\mathrm{p}^c_{\text{model}}(x;\theta)),$$

〈Clean Data에 대해 Cross-entropy Loss〉

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{|\mathcal{U}'|} \sum_{x,p \in \mathcal{U}'} \|p - \mathrm{p}_{\text{model}}(x;\theta)\|_2^2 .$$

〈Noisy Data에 대해 L2-Loss〉

$$\mathcal{L}_{\text{reg}} = \sum_c \pi_c \log\left(\pi_c \Big/ \frac{1}{|\mathcal{X}'| + |\mathcal{U}'|} \sum_{x \in \mathcal{X}' + \mathcal{U}'} \mathrm{p}^c_{\text{model}}(x;\theta)\right).$$

〈Noisy Data가 많을 경우, 1개의 클래스로만 예측하는 것을 방지하기 위한 Regularization Term〉

Dividemix

Method

〈Co-teaching〉

- 각 네트워크에서 작은 Loss 값을 가지는 Clean Data 선별
- 선별된 Clean Data만을 이용해 서로 다른 네트워크의 업데이트에 이용

〈Dividemix〉

- 각 네트워크에서 Loss 값을 이용해 GMM을 Fitting하여 Clean/Noisy Data 구분
- 선별된 Clean/Noisy Data를 Mixup을 통한 데이터 증강 후, 서로 다른 네트워크의 업데이트에 이용

Dividemix

Method

❖ 실험 데이터 : CIFAR-10N & CIFAR-100N

Dataset		CIFAR-10				CIFAR-100			
Method/Noise ratio		20%	50%	80%	90%	20%	50%	80%	90%
Cross-Entropy	Best	86.8	79.4	62.9	42.7	62.0	46.7	19.9	10.1
	Last	82.7	57.9	26.1	16.8	61.8	37.3	8.8	3.5
Bootstrap (Reed et al., 2015)	Best	86.8	79.8	63.3	42.9	62.1	46.6	19.9	10.2
	Last	82.9	58.4	26.8	17.0	62.0	37.9	8.9	3.8
F-correction (Patrini et al., 2017)	Best	86.8	79.8	63.3	42.9	61.5	46.6	19.9	10.2
	Last	83.1	59.4	26.2	18.8	61.4	37.3	9.0	3.4
Co-teaching+* (Yu et al., 2019)	Best	89.5	85.7	67.4	47.9	65.6	51.8	27.9	13.7
	Last	88.2	84.1	45.5	30.1	64.1	45.3	15.5	8.8
Mixup (Zhang et al., 2018)	Best	95.6	87.1	71.6	52.2	67.8	57.3	30.8	14.6
	Last	92.3	77.6	46.7	43.9	66.0	46.6	17.6	8.1
P-correction* (Yi & Wu, 2019)	Best	92.4	89.1	77.5	58.9	69.4	57.5	31.1	15.3
	Last	92.0	88.7	76.5	58.2	68.1	56.4	20.7	8.8
Meta-Learning* (Li et al., 2019)	Best	92.9	89.3	77.4	58.7	68.5	59.2	42.4	19.5
	Last	92.0	88.8	76.1	58.3	67.7	58.0	40.1	14.3
M-correction (Arazo et al., 2019)	Best	94.0	92.0	86.8	69.1	73.9	66.1	48.2	24.3
	Last	93.8	91.9	86.6	68.7	73.4	65.4	47.6	20.5
DivideMix	Best	**96.1**	**94.6**	**93.2**	**76.0**	**77.3**	**74.6**	**60.2**	**31.5**
	Last	**95.7**	**94.4**	**92.9**	**75.4**	**76.9**	**74.2**	**59.6**	**31.0**

Table 1: Comparison with state-of-the-art methods in test accuracy (%) on CIFAR-10 and CIFAR-100 with symmetric noise. Methods marked by * denote re-implementations based on public code.

Conclusion

연구 동향 및 결론

- Noisy Label Problem에서는 Training Loss 값이 큰 데이터를 Noisy Label 데이터라고 가정
- Mixup을 이용한 모델의 일반화 성능 확보가 Noisy Label이 포함된 데이터 학습에도 효과적
- 두 개의 Network를 각각 학습하여 Cross update하는 전략이 Noisy Label이 포함된 데이터 학습에 효과적
- 금일 세미나에서 소개한 방식 외에도 다양한 접근 방식으로 Noisy Label Problem이 연구되고 있음
  - Robust Loss Function Design
  - Robust Regularization
  - Robust Network Architecture
  - Sample Selection

Thank you

본 세미나 내용에 대한 문의 사항이 있으시면
아래의 이메일 주소로 연락주시길 바랍니다.

E-mail : dawonksh@korea.ac.kr

참고 문헌

- ✓ Jiang, L., Zhou, Z., Leung, T., Li, L. J., & Fei-Fei, L. (2018, July). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International Conference on Machine Learning (pp. 2304-2313). PMLR.
- ✓ Jiang, L., Huang, D., Liu, M., & Yang, W. (2020, November). Beyond synthetic noise: Deep learning on controlled noisy labels. In International Conference on Machine Learning (pp. 4804-4815). PMLR.
- ✓ Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., ... & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems, 31.
- ✓ Li, J., Socher, R., & Hoi, S. C. (2020). Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394.
- ✓ Chen, P., Liao, B. B., Chen, G., & Zhang, S. (2019, May). Understanding and utilizing deep neural networks trained with noisy labels. In International Conference on Machine Learning (pp. 1062-1070). PMLR.
- ✓ Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.
- ✓ Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3), 107-115.