Open DMQA Seminar

# Curriculum Learning

김상훈

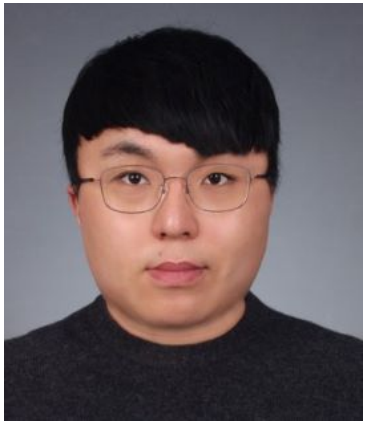
Data Mining & Quality Analytics Lab.

2021.10.08(금)

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
  - ✓ Machine learning / Deep learning Algorithms
  - ✓ Open Set Recognition / Curriculum Learning
- Contact
  - ✓ E-mail : dawonksh@korea.ac.kr

# 목차


- ❖ Curriculum Learning
  - ✓ Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41–48).
- ❖ Self-Paced Learning
  - ✓ Kumar, M., Packer, B., & Koller, D. (2010). Self-paced learning for latent variable models. Advances in neural information processing systems, 23, 1189–1197.
- ❖ Self-Paced Curriculum Learning
  - ✓ Jiang, L., Meng, D., Zhao, Q., Shan, S., & Hauptmann, A. G. (2015, February). Self-paced curriculum learning. In Twenty-Ninth AAAI Conference on Artificial Intelligence.
- ❖ MentorNet
  - ✓ 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.
- ❖ Conclusion


# Curriculum Learning

## 논문 소개

❖ Curriculum Learning

- 2009년 ICML(International Conference on Machine Learning)에 Yoshua Bengio가 쓴 논문
- 2021년 10월 8일 기준 인용 횟수 : 3,216회

**Curriculum Learning**

**Yoshua Bengio**[1] YOSHUA.BENGIO@UMONTREAL.CA
**Jérôme Louradour**[1,2] JEROMELOURADOUR@GMAIL.COM
**Ronan Collobert**[3] RONAN@COLLOBERT.COM
**Jason Weston**[3] JASONW@NEC-LABS.COM
(1) U. MONTREAL, P.O. BOX 6128, MONTREAL, CANADA (2) A2IA SA, 40BIS FABERT, PARIS, FRANCE
(3) NEC LABORATORIES AMERICA, 4 INDEPENDENCE WAY, PRINCETON, NJ, USA

**Abstract**

Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Here, we formalize such training strategies in the context of machine learning, and call them "curriculum learning". In the context of recent research studying the difficulty of training in the presence of non-convex training criteria (for deep deterministic and stochastic neural networks), we explore curriculum learning in various set-ups. The experiments show that significant improvements in generalization can be achieved. We hypothesize that curriculum learning has both an effect on the speed of convergence of the training process to a minimum and, in the case of non-convex criteria, on the quality of the local minima obtained: curriculum learning can be seen as a particular form of continuation method (a general strategy for global optimization of non-convex functions).

training and remarkably increase the speed at which learning can occur. This idea is routinely exploited in *animal training* where it is called **shaping** (Skinner, 1958; Peterson, 2004; Krueger & Dayan, 2009).

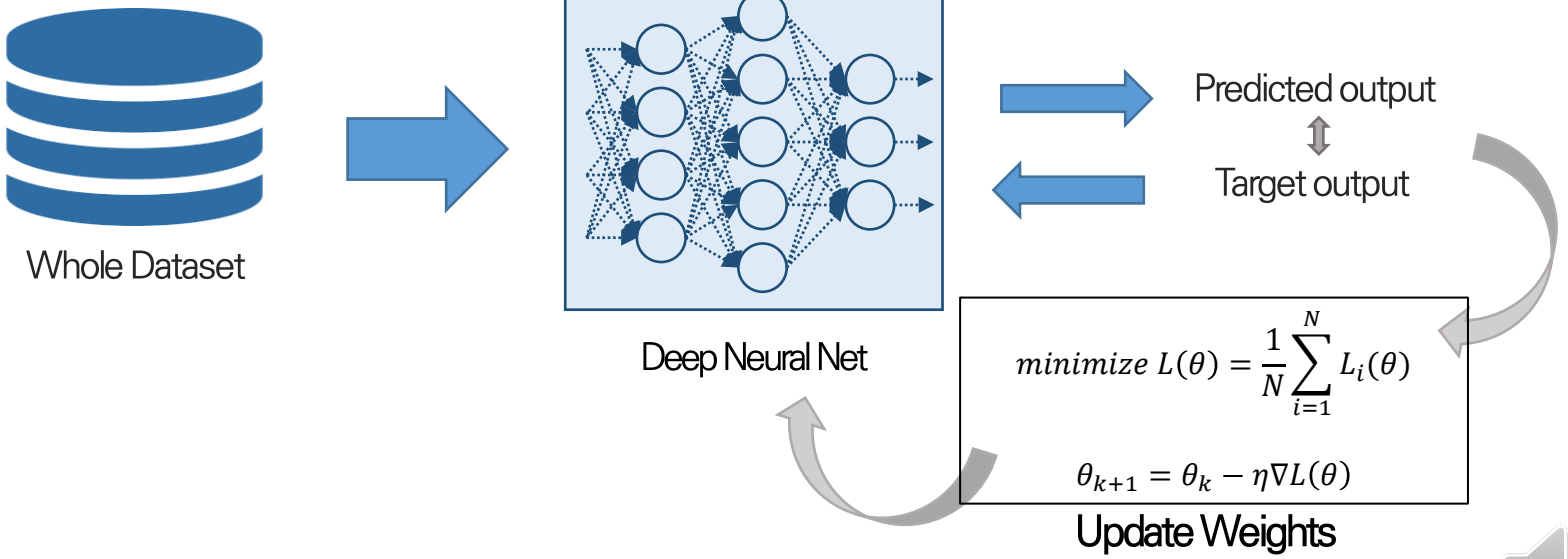
Previous research (Elman, 1993; Rohde & Plaut, 1999; Krueger & Dayan, 2009) at the intersection of cognitive science and machine learning has raised the following question: can machine learning algorithms benefit from a similar training strategy? The idea of training a learning machine with a curriculum can be traced back at least to Elman (1993). The basic idea is to *start small*, learn easier aspects of the task or easier subtasks, and then gradually increase the difficulty level. The experimental results, based on learning a simple grammar with a recurrent network (Elman, 1993), suggested that successful learning of grammatical structure depends, not on innate knowledge of grammar, but on starting with a limited architecture that is at first quite restricted in complexity, but then expands its resources gradually as it learns. Such conclusions are important for developmental psychology, because they illustrate the adaptive value of starting, as human infants do, with a simpler initial state, and then building on that to develop more and more sophis-

# Curriculum Learning

## 핵심 아이디어

❖ **일반적인 딥러닝 모델 학습 방식**

- 데이터가 큰 경우, 배치 단위로 나누어서 학습 : 모델 가중치 업데이트
- 각 배치는 랜덤한 순서로 학습됨

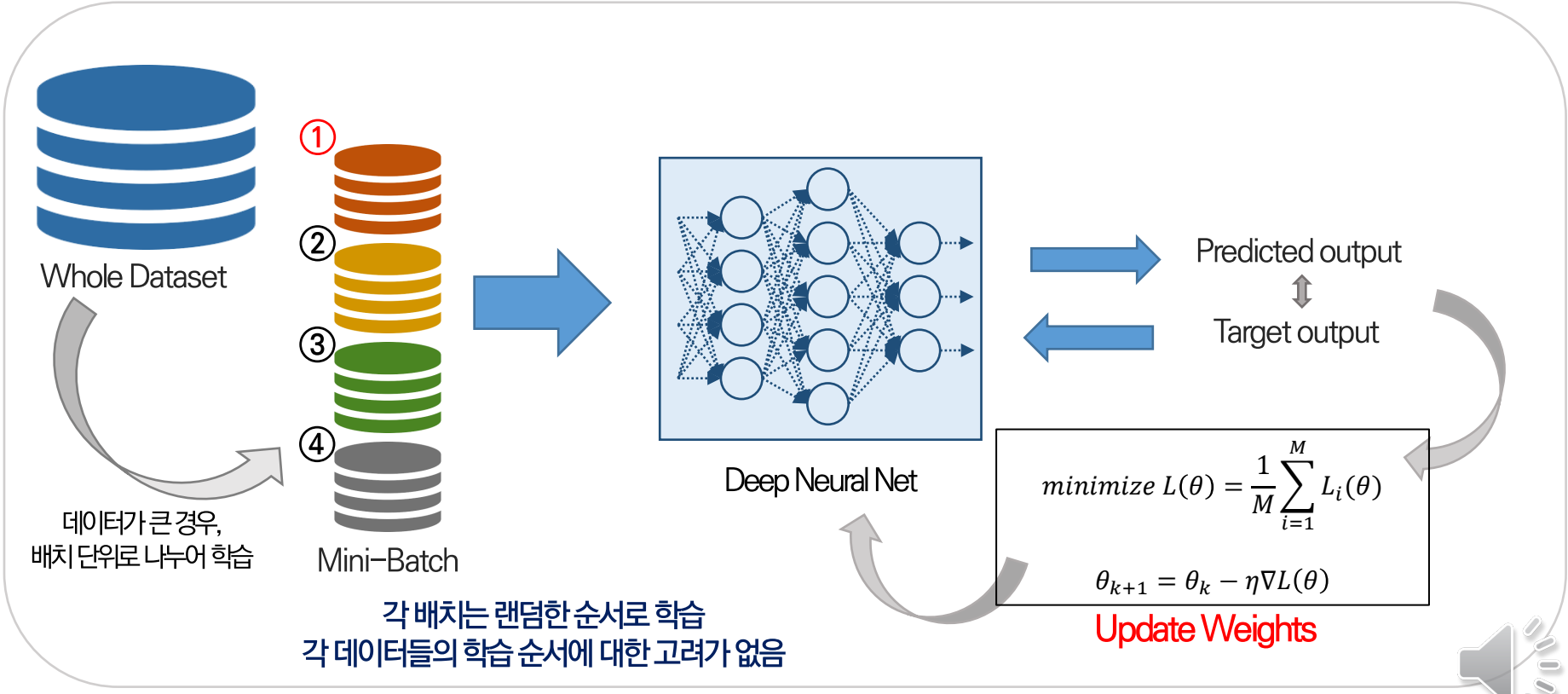


# Curriculum Learning

## 핵심 아이디어

❖ **일반적인 딥러닝 모델 학습 방식**

- 데이터가 큰 경우, 배치 단위로 나누어서 학습 : 모델 가중치 업데이트
- 각 배치는 랜덤한 순서로 학습됨

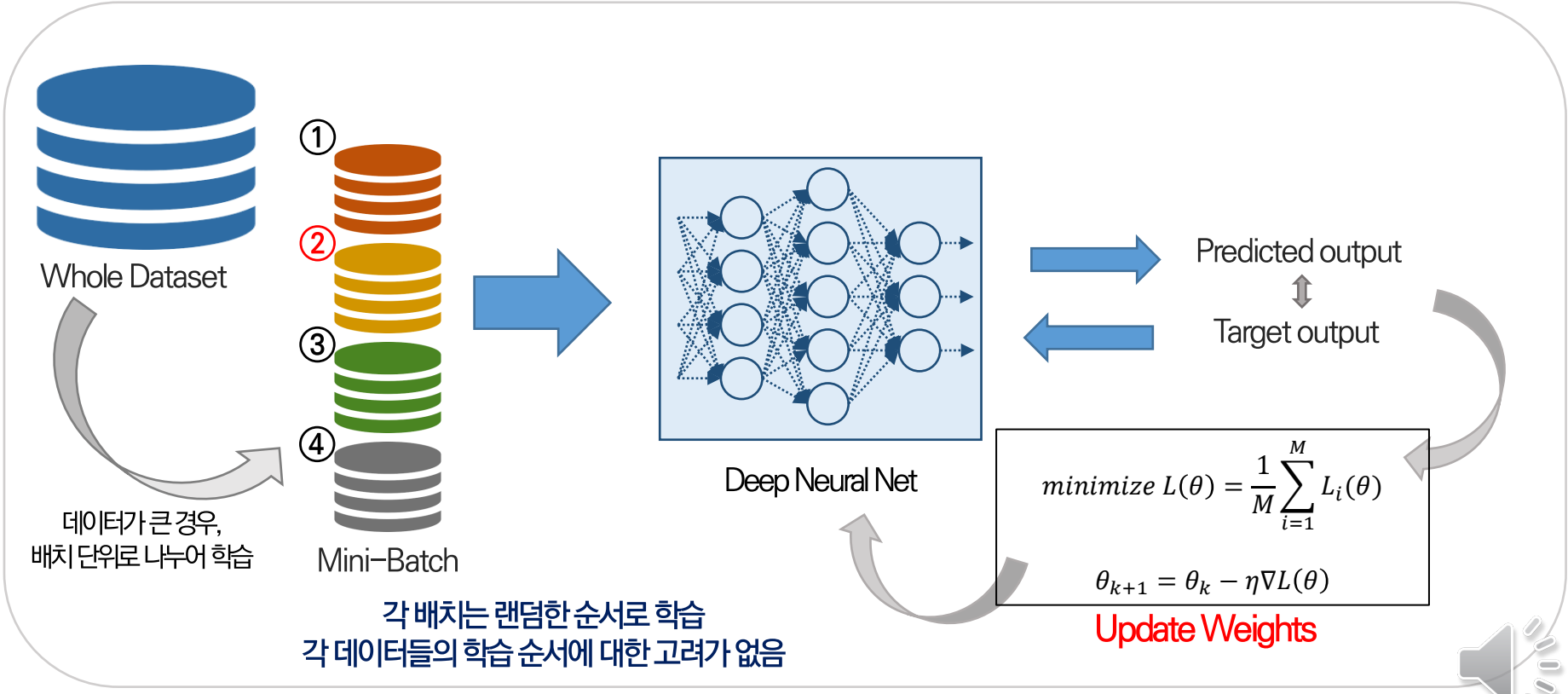


# Curriculum Learning

## 핵심 아이디어

❖ **일반적인 딥러닝 모델 학습 방식**

- 데이터가 큰 경우, 배치 단위로 나누어서 학습 : 모델 가중치 업데이트
- 각 배치는 랜덤한 순서로 학습됨

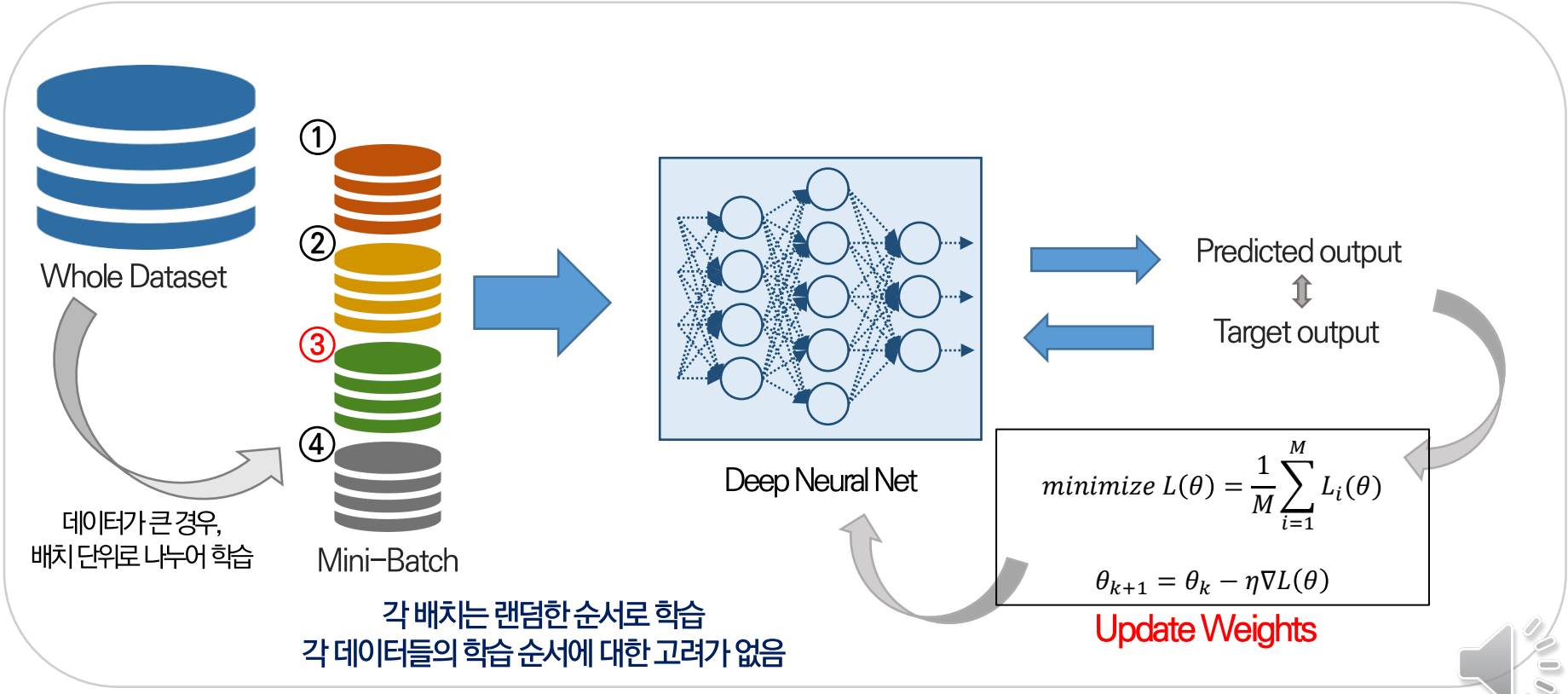


# Curriculum Learning

## 핵심 아이디어

❖ **일반적인 딥러닝 모델 학습 방식**

- 데이터가 큰 경우, 배치 단위로 나누어서 학습 : 모델 가중치 업데이트
- 각 배치는 랜덤한 순서로 학습됨

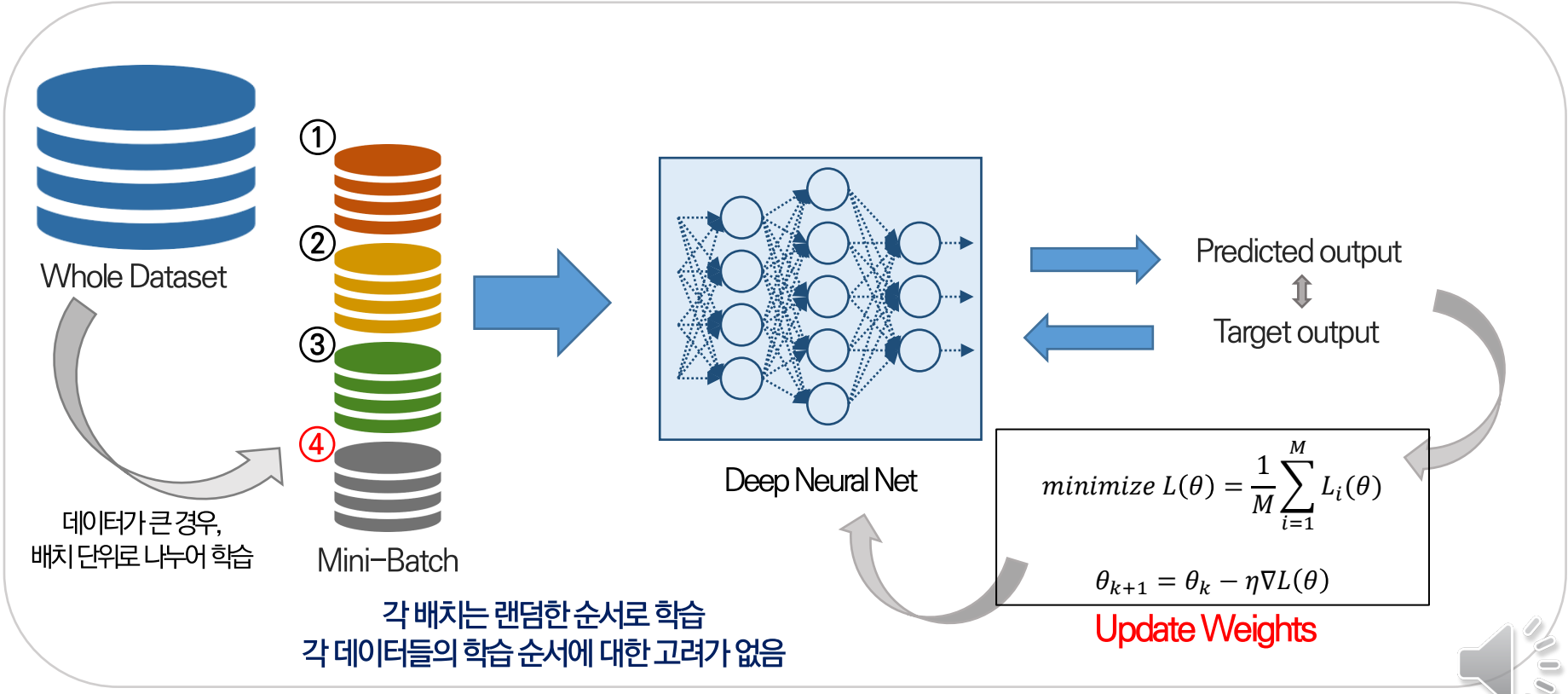


# Curriculum Learning

## 핵심 아이디어

❖ **일반적인 딥러닝 모델 학습 방식**

- 데이터가 큰 경우, 배치 단위로 나누어서 학습 : 모델 가중치 업데이트
- 각 배치는 랜덤한 순서로 학습됨

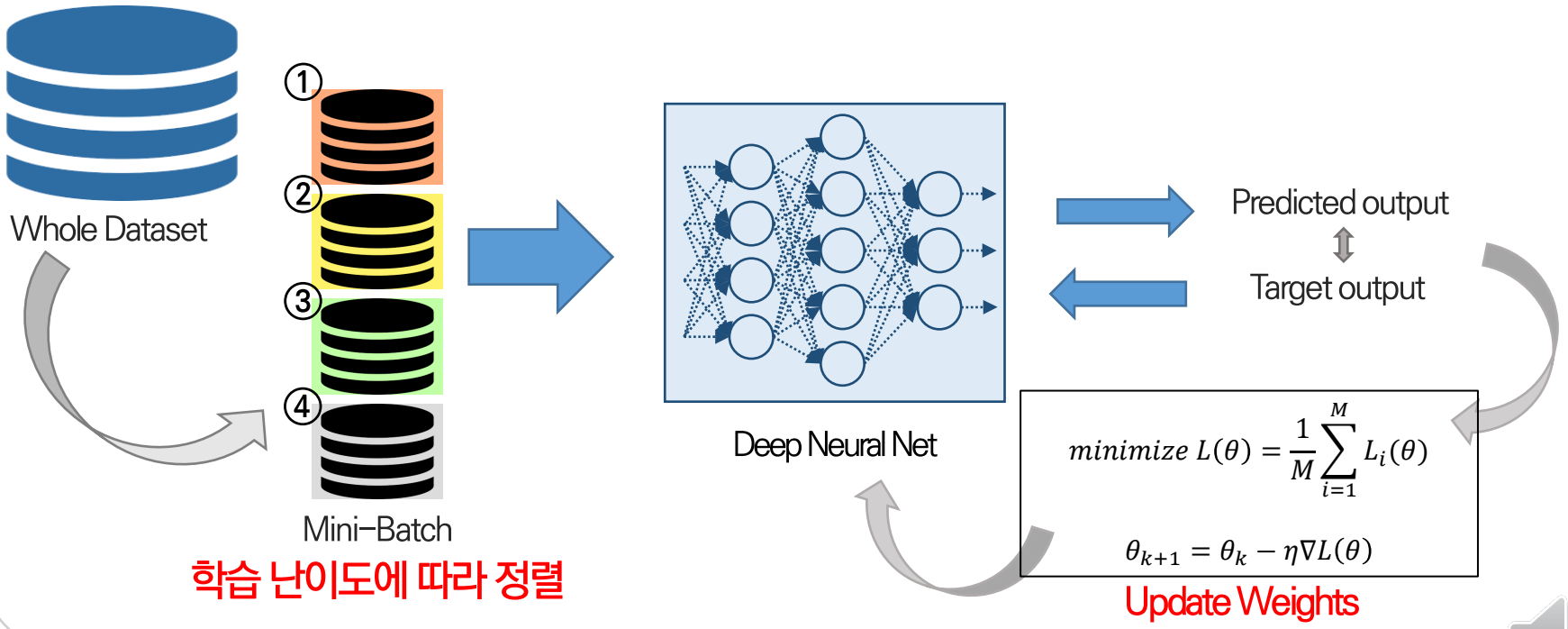


# Curriculum Learning

## 핵심 아이디어

❖ **쉬운 샘플부터 학습하자!**

- 인간의 학습 프로세스를 모방 : 쉬운 것부터 학습하고, 차츰 어려운 것까지 학습
- 빠른 수렴 속도 확보 및 Local minima에 빠지는 경향성 감소
- 최종 학습 모델을 효율적으로 학습하여 고성능 성과 확보

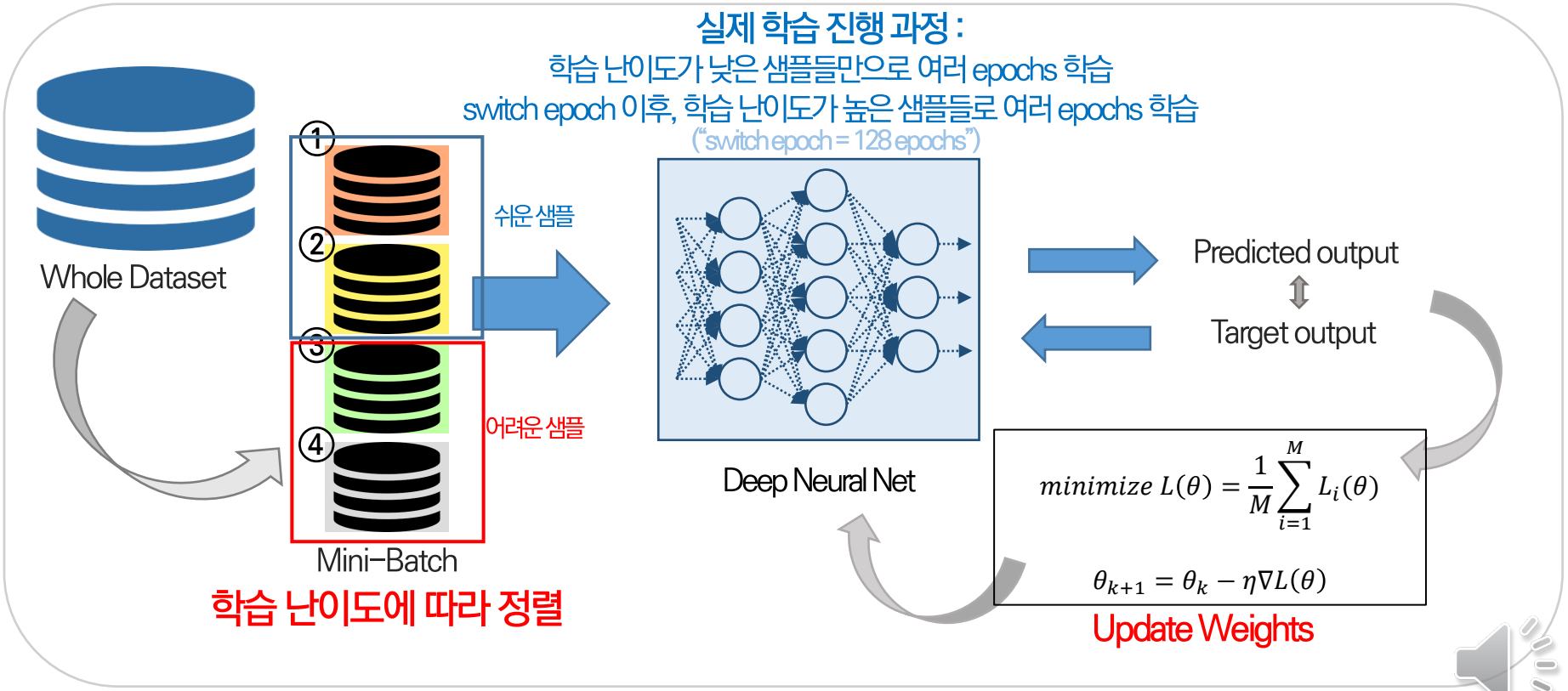


# Curriculum Learning

## 핵심 아이디어

❖ **쉬운 샘플부터 학습하자!**

- 인간의 학습 프로세스를 모방 : 쉬운 것부터 학습하고, 차츰 어려운 것까지 학습
- 빠른 수렴 속도 확보 및 Local minima에 빠지는 경향성 감소
- 최종 학습 모델을 효율적으로 학습하여 고성능 성과 확보



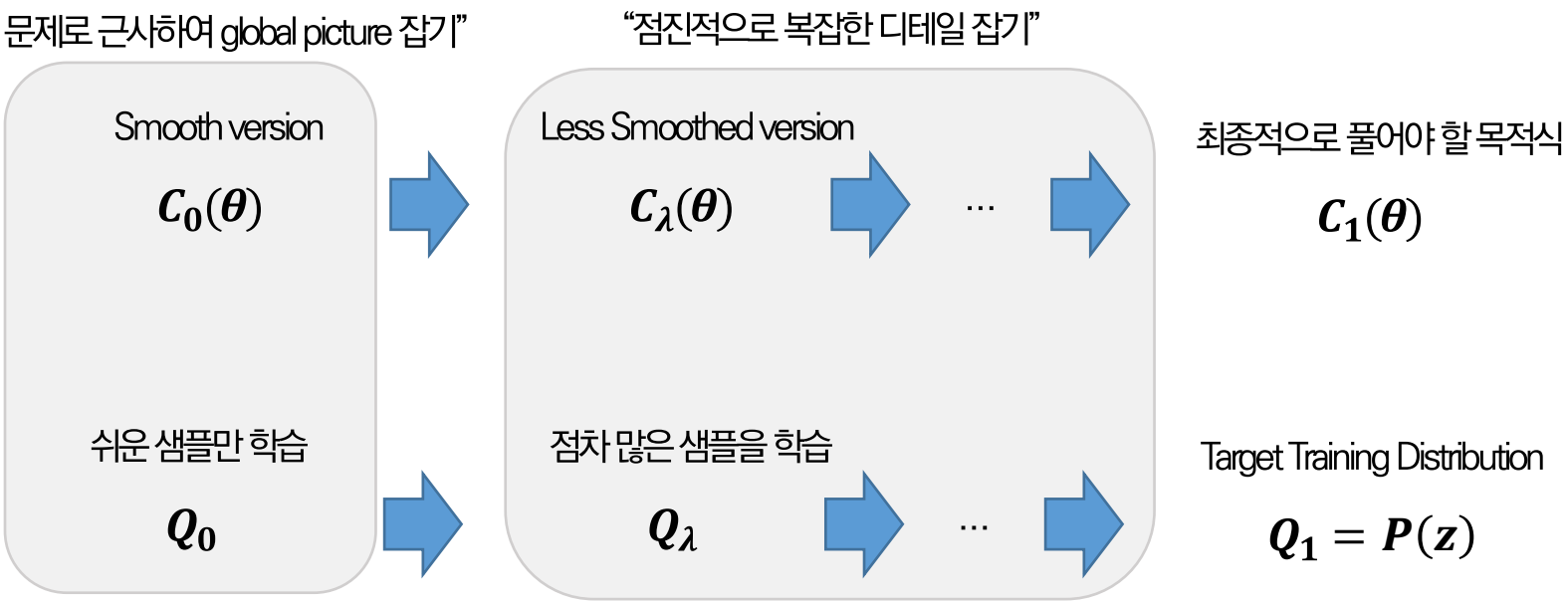
# Curriculum Learning

## 이론적 배경

### ❖ Continuation methods (Allgower & Georg, 1980)

- 목적식이 non-convex한 조건에서의 global optimum에 근사한 해를 찾는 최적화 방법론
- 쉬운 문제로 근사하여 global picture 잡기 : 목적식의 smooth version(convex)을 찾아 최적화
- 점진적으로 복잡한 디테일 잡기 : highly smoothed version에서 less smoothed version까지 순차적 진행

### ❖ A curriculum as a continuation method




Step $\lambda$ 마다 난이도를 점차 높여 최종 목적식의 global optimum에 근접한 해를 찾는 방식

# Curriculum Learning

## 이론적 배경

### ❖ Continuation methods (Allgower & Georg, 1980)

- 목적식이 non-convex한 조건에서의 global optimum에 근사한 해를 찾는 최적화 방법론
- 쉬운 문제로 근사하여 global picture 잡기 : 목적식의 smooth version(convex)을 찾아 최적화
- 점진적으로 복잡한 디테일 잡기 : highly smoothed version에서 less smoothed version까지 순차적 진행

### ❖ A curriculum as a continuation method

**Step $\lambda$** $0 \leq \lambda \leq 1$ (starting from $\lambda = 0$ and ending at $\lambda = 1$)

현재 학습단계에서 학습에 포함시킨 데이터의 비율

최종적으로 학습해야 하는 Target Training Distribution

$$Q_\lambda \propto W_\lambda(z)P(z) \; \forall z$$

$\lambda$ 단계에서 학습한 Distribution

Random Variable : (x,y) pair

$\lambda$가 증가할수록, 점차 많은 샘플을 학습

: 쉬운 샘플 먼저 학습하여, Global Picture 잡아나가기

Entropy 증가 : $H(Q_\lambda) < H(Q_{\lambda+\epsilon}) \; \forall \epsilon > 0$
데이터비율증가 : $W_{\lambda+\epsilon}(z) \geq W_\lambda(z) \; \forall z, \forall \epsilon > 0$

**Final Step**

$$Q_1 = P(z) \; \forall z$$

최종적으로 학습하고자 하는 Target Training Distribution 학습

: 점차 어려운 샘플을 학습하여, Target Training Distribution 학습

# Curriculum Learning

## 이론적 배경

❖ Continuation methods (Allgower & Georg, 1980)

- 목적식이 non-convex한 조건에서의 global optimum에 근사한 해를 찾는 최적화 방법론
- 쉬운 문제로 근사하여 global picture 잡기 : 목적식의 smooth version(convex)을 찾아 최적화
- 점진적으로 복잡한 디테일 잡기 : highly smoothed version에서 less smoothed version까지 순차적 진행

❖ A curriculum as a continuation method

**실제 실험에서는 2 Step의 $\lambda$로 구성**

(전체 데이터를 쉬운 샘플/어려운 샘플 2종류로 구분)

$Q_0$ : 쉬운 샘플만 학습

$Q_1$ : 전체 데이터셋 학습

쉬운 샘플만 학습한 모델을 특정 학습 시점 이후부터 전체 데이터 모두 학습

수렴속도 ↑

Local Minimum에 빠질 경향성 ↓

모델 성능 ↑

# Curriculum Learning

## 실험 소개

❖ Experiments on shape recognition

- 동그라미, 삼각형, 사각형을 분류하는 문제
- Curriculum Learning 적용 : **사전지식**을 통해 데이터의 학습 난이도 구분
  - ✓ 쉬운 샘플 : 원, 정삼각형, 정사각형
  - ✓ 어려운 샘플 : 타원, 삼각형(정삼각형 x), 직사각형
- Multi-layer neural network (3 hidden layers) with SGD

정형화된 패턴
쉬운 샘플

정형화되지 않은 패턴
어려운 샘플

*Figure 2.* Sample inputs from `BasicShapes` (top) and `GeomShapes` (bottom). Images are shown here with a higher resolution than the actual dataset (32x32 pixels).

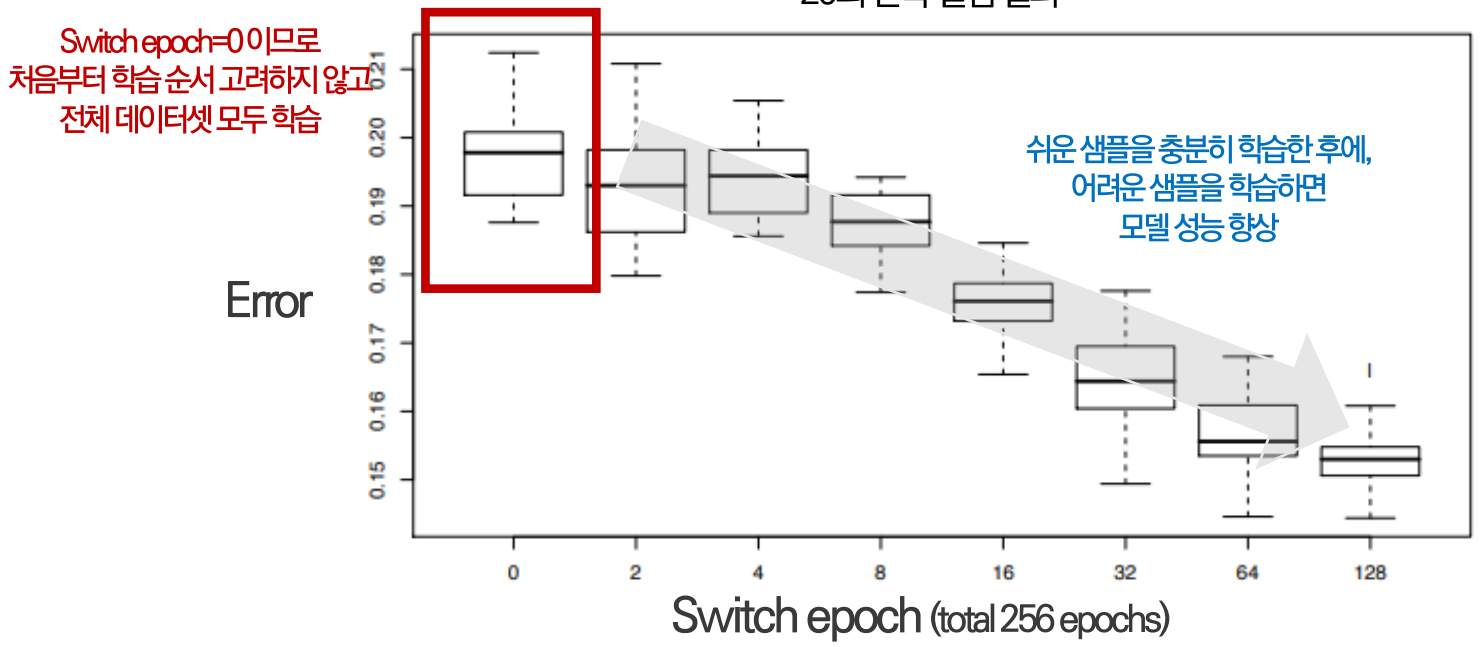
# Curriculum Learning

## 실험 소개

❖ Experiments on shape recognition

- 동그라미, 삼각형, 사각형을 분류하는 문제
- Curriculum Learning 적용 : **사전지식**을 통해 데이터의 학습 난이도 구분
  - ✓ 쉬운 샘플 : 원, 정삼각형, 정사각형
  - ✓ 어려운 샘플 : 타원, 삼각형(정삼각형 x), 직사각형
- Multi-layer neural network (3 hidden layers) with SGD

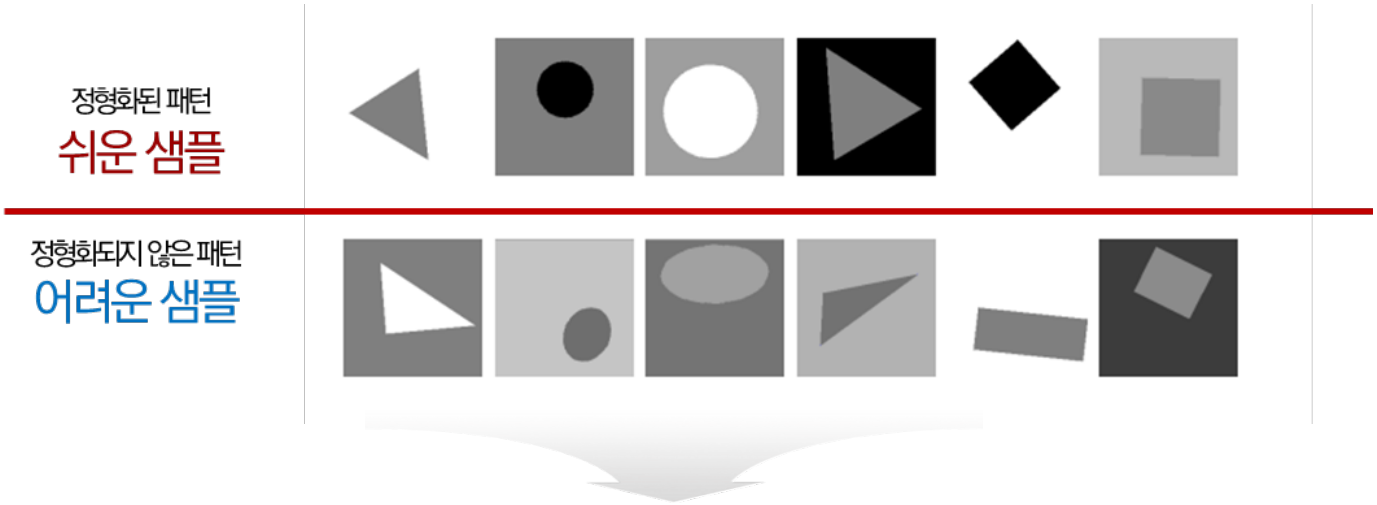
20회 반복 실험 결과



# Curriculum Learning

## 한계점

❖ Curriculum Learning을 적용하기 위해서는 **데이터의 학습 난이도 사전지식이 필요**




현실 문제의 실제 데이터는 **학습 난이도에 대한 사전지식이 없음**

# Self-paced Learning

## 논문 소개

❖ Self-Paced Learning for Latent Variable Models

- 2010년 NeurIPS (International Conference on Machine Learning)에 DeepMind 소속 M.Kumar가 발표한 논문
- 2021년 10월 8일 기준 인용 횟수 : 955회

**Self-Paced Learning for Latent Variable Models**

**M. Pawan Kumar** **Ben Packer** **Daphne Koller**
Computer Science Department
Stanford University
{pawan,bpacker,koller}@cs.stanford.edu

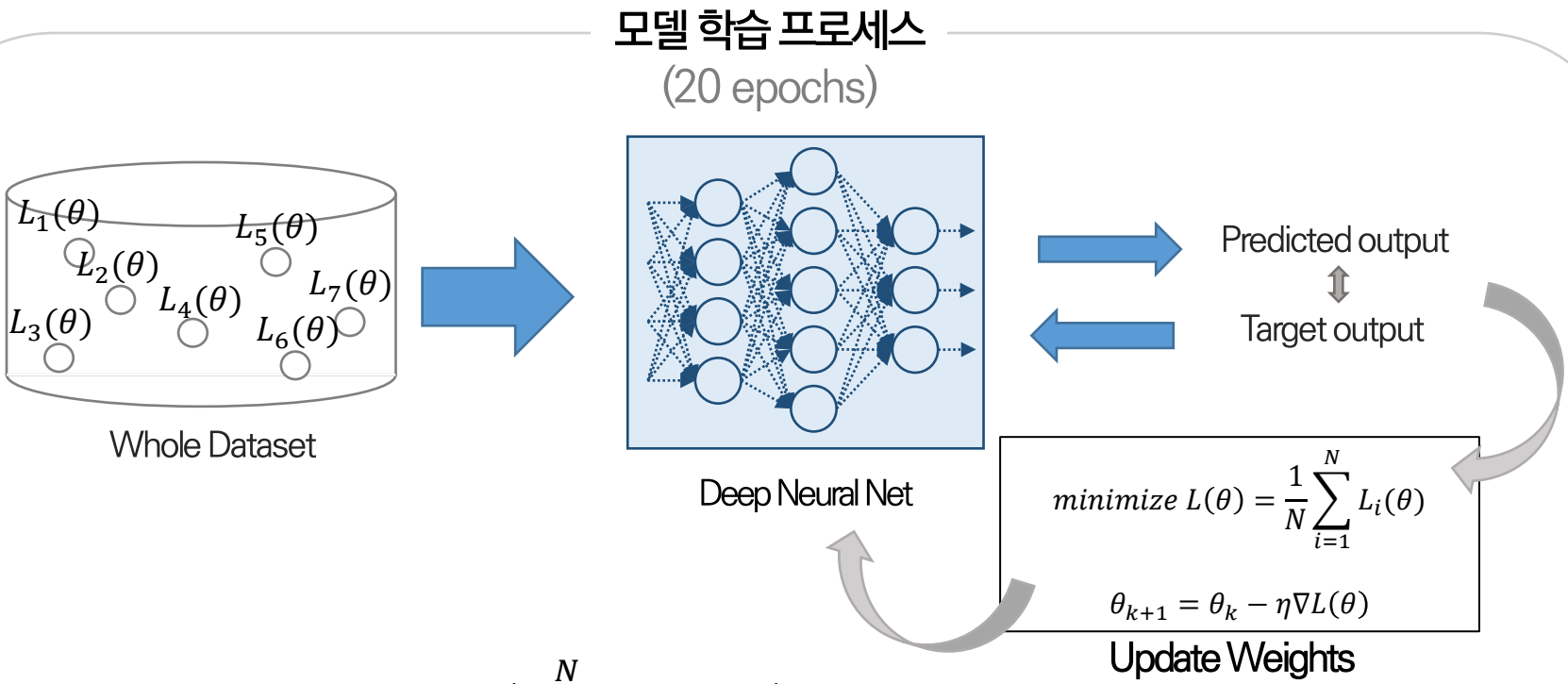
**Abstract**

Latent variable models are a powerful tool for addressing several tasks in machine learning. However, the algorithms for learning the parameters of latent variable models are prone to getting stuck in a bad local optimum. To alleviate this problem, we build on the intuition that, rather than considering all samples simultaneously, the algorithm should be presented with the training data in a *meaningful* order that facilitates learning. The order of the samples is determined by how *easy* they are. The main challenge is that typically we are not provided with a readily computable measure of the easiness of samples. We address this issue by proposing a novel, iterative *self-paced learning* algorithm where each iteration simultaneously selects easy samples and learns a new parameter vector. The number of samples selected is governed by a weight that is annealed until the entire training data has been considered. We empirically demonstrate that the self-paced learning algorithm outperforms the state of the art method for learning a latent structural SVM on four applications: object localization, noun phrase coreference, motif finding and handwritten digit recognition.

# Self-paced Learning

핵심 아이디어

❖ 일반적인 딥러닝 모델 학습 프로세스



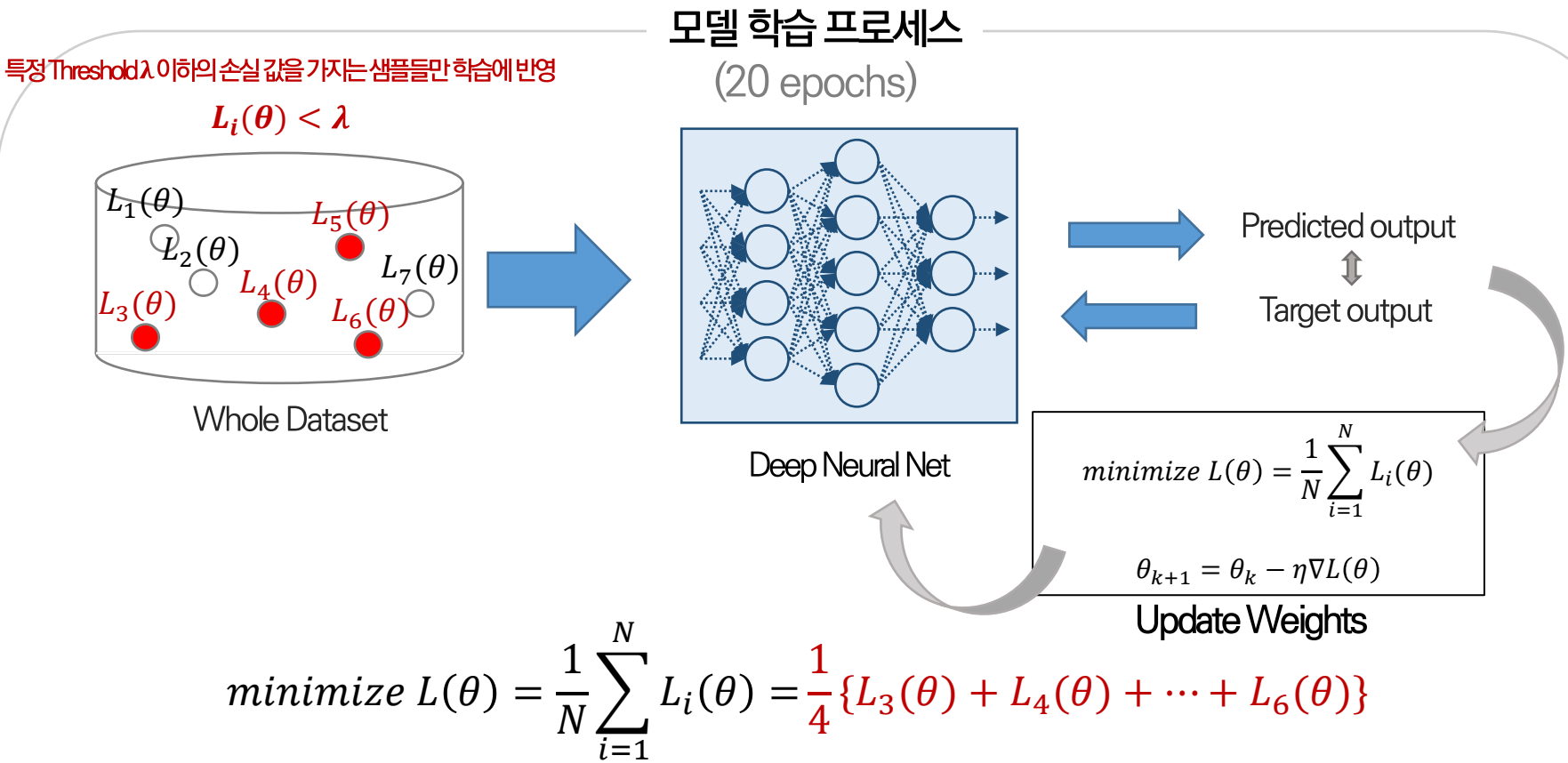

$$minimize\ L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta) = \frac{1}{7}\{L_1(\theta) + L_2(\theta) + \cdots + L_7(\theta)\}$$

Loss function을 최소화하는 방향으로 모델 파라미터 학습

# Self-paced Learning

## 핵심 아이디어

❖ 사전지식없이 모델의 학습 과정에서 학습 난이도를 계산하여, 쉬운 것부터 학습하자!




$$minimize\ L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta) = \frac{1}{4}\{L_3(\theta) + L_4(\theta) + \cdots + L_6(\theta)\}$$

Loss 값이 클수록, 현재 학습 단계에서 모델이 잘 못 맞추는 어려운 데이터 (학습에서 배제)

# Self-paced Learning

## 핵심 아이디어

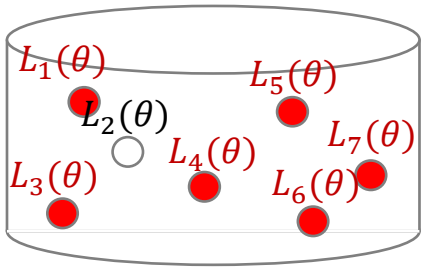
❖ 사전지식없이 모델의 학습 과정에서 학습 난이도를 계산하여, 쉬운 것부터 학습하자!

### 모델 학습 프로세스

(50 epochs)

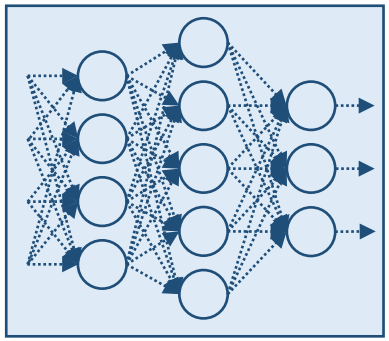
특정 Threshold $\lambda$ 이하의 손실 값을 가지는 샘플들만 학습에 반영

$L_i(\theta) < \lambda \uparrow$ Monotonically Increase : $\lambda \rightarrow \mu^k\lambda$




Whole Dataset

Threshold $\lambda$를 단조 증가시켜 학습 단계에 따라 점차적으로 어려운 샘플을 학습할 수 있게 함



Deep Neural Net

Predicted output

Target output

$$minimize\ L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta)$$

$$\theta_{k+1} = \theta_k - \eta\nabla L(\theta)$$

Update Weights

$$minimize\ L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta) = \frac{1}{6}\{L_1(\theta) + L_2(\theta) + L_3(\theta) + \cdots + L_6(\theta)\}$$

Loss 값이 큰 어려운 데이터를 학습 단계가 지날수록 학습에 포함

# Self-paced Learning

## 알고리즘

❖ Self-paced Learning 손실함수 $L(\theta)$

- 손실 함수는 각 단계에서 변수 $v_i$ 와 $\theta$에 대해 한 변수를 고정한 상태로 나머지 변수 최적화
  - ✓ 첫 번째 학습 단계 : 변수 $\theta$ 고정, 변수 $v_i$ 최적화 (threshold $\lambda$)
  - ✓ 두 번째 학습 단계 : 변수 $v_i$ 고정, 변수 $\theta$ 최적화 (Gradient descent)

〈일반적인 손실함수〉

$$minimize\ L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta) + r(\theta)$$

Regularization term

〈Self-paced Learning 손실함수〉

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

Regularization term

변수 $v_i$ 는 현재 데이터 포인트 $(x_i, y_i)$ 가 학습하기에 충분히 쉬운지 여부를 결정
$(v_i = 0\ or\ 1)$

# Self-paced Learning

## 알고리즘

❖ Self-paced Learning 손실함수 $L(\theta)$

- 손실 함수는 각 단계에서 변수 $v_i$ 와 $\theta$에 대해 한 변수를 고정한 상태로 나머지 변수 최적화
  - ✓ 첫 번째 학습 단계 : 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** (threshold $\lambda$)
  - ✓ 두 번째 학습 단계 : 변수 $v_i$ 고정, **변수 $\boldsymbol{\theta}$ 최적화** (Gradient descent)

〈Self-paced Learning 손실함수 – 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** 과정〉

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

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

$(v_i = 0\ or\ 1)$

변수 $v_i$ 는 현재 데이터 포인트 $(x_i, y_i)$ 가 학습하기에 충분히 쉬운지 여부를 결정

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

**특정 Threshold $\lambda$ 이하의 손실 값을 가지는 샘플들만 학습에 반영**

# Self-paced Learning

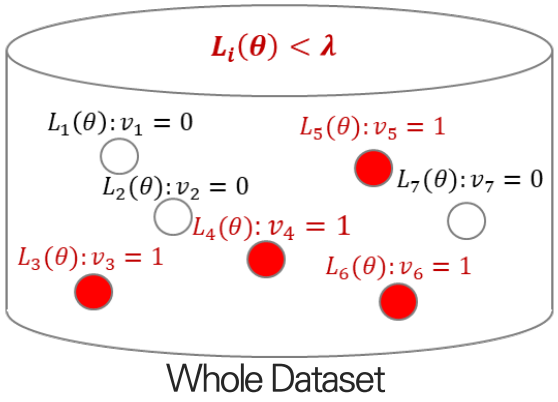
## 알고리즘

❖ Self-paced Learning 손실함수 $L(\theta)$

- 손실 함수는 각 단계에서 변수 $v_i$ 와 $\theta$에 대해 한 변수를 고정한 상태로 나머지 변수 최적화
  - ✓ 첫 번째 학습 단계 : 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** (threshold $\lambda$)
  - ✓ 두 번째 학습 단계 : 변수 $v_i$ 고정, **변수 $\boldsymbol{\theta}$ 최적화** (Gradient descent)

〈Self-paced Learning 손실함수 – 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** 과정〉

(20 epochs)




Whole Dataset

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

$$= \frac{1}{7}\{0 \times L_1(\theta) + 0 \times L_2(\theta) + L_3(\theta) + L_4(\theta) + L_5(\theta) + L_6(\theta) + 0 \times L_7(\theta)\} - 4\lambda + r(\theta)$$

**변수 $\boldsymbol{\theta}$ 최적화** : $\theta_{k+1} = \theta_k - \eta \nabla L(\theta)$ 쉬운 샘플들만 학습에 반영

# Self-paced Learning

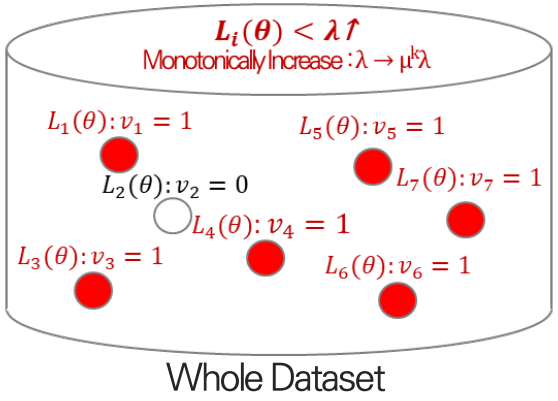
## 알고리즘

❖ Self-paced Learning 손실함수 $L(\theta)$

- 손실 함수는 각 단계에서 변수 $v_i$ 와 $\theta$에 대해 한 변수를 고정한 상태로 나머지 변수 최적화
  - ✓ 첫 번째 학습 단계 : 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** (threshold $\lambda$)
  - ✓ 두 번째 학습 단계 : 변수 $v_i$ 고정, **변수 $\boldsymbol{\theta}$ 최적화** (Gradient descent)

〈Self-paced Learning 손실함수 – 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** 과정〉

(50 epochs)




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

$$= \frac{1}{7}\{L_1(\theta) + 0 \times L_2(\theta) + L_3(\theta) + L_4(\theta) + L_5(\theta) + L_6(\theta) + L_7(\theta)\} - 6\lambda + r(\theta)$$

**변수 $\boldsymbol{\theta}$ 최적화** : $\theta_{k+1} = \theta_k - \eta\nabla L(\theta)$ 점차 어려운 샘플 포함

# Self-paced Curriculum Learning

## 논문 소개

❖ Self-Paced Curriculum Learning

- 2015년 AAAI (Association for the Advancement of Artificial Intelligence)에 Lu Jiang가 발표한 논문
- 2021년 10월 8일 기준 인용 횟수 : 345회

**Self-Paced Curriculum Learning**

**Lu Jiang**[1], **Deyu Meng**[1,2], **Qian Zhao**[1,2], **Shiguang Shan**[1,3], **Alexander G. Hauptmann**[1]
[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 15217
[2] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, P. R. China, 710049
[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P. R. China, 100190
lujiang@cs.cmu.edu, dymeng@mail.xjtu.edu.cn,
timmy.zhaoqian@gmail.com, sgshan@ict.ac.cn, alex@cs.cmu.edu

**Abstract**

Curriculum learning (CL) or self-paced learning (SPL) represents a recently proposed learning regime inspired by the learning process of humans and animals that gradually proceeds from easy to more complex samples in training. The two methods share a similar conceptual learning paradigm, but differ in specific learning schemes. In CL, the curriculum is predetermined by prior knowledge, and remain fixed thereafter. Therefore, this type of method heavily relies on the quality of prior knowledge while ignoring feedback about the learner. In SPL, the curriculum is dynamically determined to adjust to the learning pace of the leaner. However, SPL is unable to deal with prior knowledge, rendering it prone to overfitting. In this paper, we discover the missing link between CL and SPL, and propose a unified framework named self-paced curriculum leaning (SPCL). SPCL is formulated as a concise optimization problem that takes into account both prior knowledge known before training and the learning progress during training. In comparison to human education, SPCL is analogous to "instructor-student-collaborative" learning mode, as opposed to "instructor-driven" in CL or "student-driven" in SPL. Empirically, we show that the advantage of SPCL on two tasks.

between *curriculum learning* (CL) and *self-paced learning* (SPL) lies in the derivation of the curriculum. In CL, the curriculum is assumed to be given by an oracle beforehand, and remains fixed thereafter. In SPL, the curriculum is dynamically generated by the learner itself, according to what the learner has already learned.

The advantage of CL includes the flexibility to incorporate prior knowledge from various sources. Its drawback stems from the fact that the curriculum design is determined independently of the subsequent learning, which may result in inconsistency between the fixed curriculum and the dynamically learned models. From the optimization perspective, since the learning proceeds iteratively, there is no guarantee that the predetermined curriculum can even lead to a converged solution. SPL, on the other hand, formulates the learning problem as a concise biconvex problem, where the curriculum design is embedded and jointly learned with model parameters. Therefore, the learned model is consistent. However, SPL is limited in incorporating prior knowledge into learning, rendering it prone to overfitting. Ignoring prior knowledge is less reasonable when reliable prior information is available. Since both methods have their advantages, it is difficult to judge which one is better in practice.

# Self-paced Curriculum Learning

## 핵심 아이디어

❖ Self-paced Curriculum Learning = Self-Paced Learning + Curriculum Learning

- 데이터에 대한 사전지식을 활용하면서, 학습 과정에서 데이터 학습 난이도 또한 활용!

〈Self-paced Learning 손실함수〉

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

Regularization term

변수 $v_i$ 는 현재 데이터 포인트 $(x_i, y_i)$ 가 학습하기에 충분히 쉬운지 여부를 결정

$(\boldsymbol{v_i = 0\ or\ 1})$

〈Self-paced Curriculum Learning 손실함수〉

① Self-paced function implementation :
$v_i$ 가 0 ~ 1 사이의 실수 출력

$$minimize\ L(\theta) = \frac{1}{N}\sum_{i=1}^{N} v_i L_i(\theta) + f(v;\lambda) + r(\theta)$$

Regularization term

② Curriculum region : 사전지식 반영

$$\boldsymbol{s.t.\ v_i \in \Psi}$$

변수 $v_i$ 는 0 ~ 1 사이의 실수, 현재 데이터 포인트 $(x_i, y_i)$의 학습 가중치 반영

# Self-paced Curriculum Learning

## 핵심 아이디어

❖ Self-paced function implementation

- Binary scheme

$$f(v;\lambda) = -\lambda \sum_{i=1}^{N} v_i, \qquad 0 < \lambda$$

- Linear scheme

$$f(v;\lambda) = \frac{1}{2}\lambda \sum_{i=1}^{N} (v_i^2 - 2v_i), \qquad 0 < \lambda$$

- Logarithmic scheme

$$f(v;\lambda) = \sum_{i=1}^{N} \zeta v_i - \frac{\zeta^{v_i}}{log\zeta}, \qquad \zeta = 1 - \lambda, \qquad 0 < \lambda < 1$$

- Mixture scheme

$$f(v;\lambda) = -\zeta \sum_{i=1}^{N} \log(v_i + \frac{1}{\lambda_1}\zeta), \qquad \zeta = \frac{\lambda_1\lambda_2}{\lambda_1 - \lambda_2}, \qquad 0 < \lambda_2 < \lambda_1$$

Self-paced Learning

변수 $\theta$ 고정, 변수 $v_i$ 최적화 결과, $\boldsymbol{v_i = 0\ or\ 1}$의 값을 출력

Self-paced Curriculum Learning

변수 $\theta$ 고정, 변수 $v_i$ 최적화 결과,

$\boldsymbol{v_i}$는 0~1 사이의 실수 값을 출력

# Self-paced Curriculum Learning

핵심 아이디어

❖ Self-paced function implementation

〈Logarithmic scheme – 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** 과정〉

$$minimize\ L(\theta) = \frac{1}{N}\left(\sum_{i=1}^{N} v_i L_i(\theta) + \sum_{i=1}^{N}\left(\zeta v_i - \frac{\zeta^{v_i}}{log\zeta}\right)\right) + r(\theta), \quad \zeta = 1-\lambda, \quad 0 < \lambda < 1$$

**($\mathbf{L(\theta)}$를 $\mathbf{v_i}$에 대해 편미분)**

$$\frac{\partial L(\theta)}{\partial v_i} = L_i(\theta) + (\zeta - \zeta^{v_i}) = 0,$$

$$\log(L_i(\theta) + \zeta) = v_i log\zeta.$$

Optimal solution $$v_i^* = \begin{cases} \dfrac{1}{log\zeta}\log(L_i(\theta) + \zeta), & L_i(\theta) < \lambda \\ 0, & L_i(\theta) \geq \lambda \end{cases}$$

**$\mathbf{v_i}$**는 0~1 사이의 실수 값을 출력

현재 데이터 포인트 $(x_i, y_i)$의 학습 가중치 반영

# Self-paced Curriculum Learning

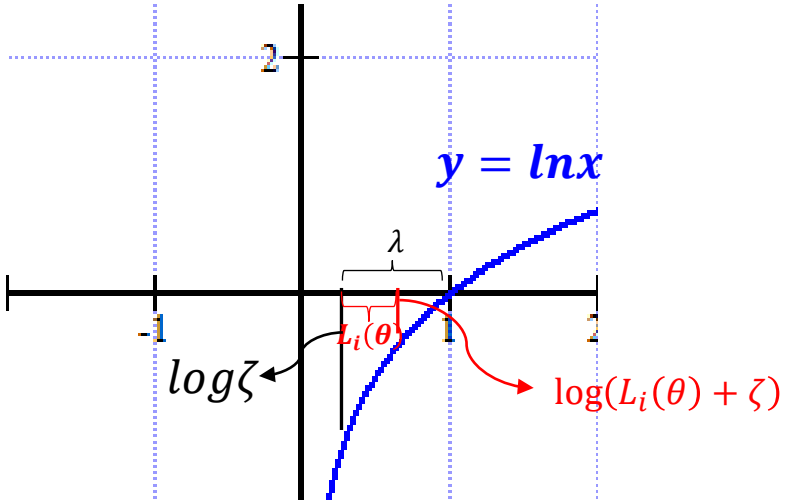
핵심 아이디어

❖ Self-paced function implementation

〈Logarithmic scheme – 변수 $\theta$ 고정, **변수 $\mathbf{v_i}$ 최적화** 과정〉

Optimal solution $$v_i^* = \begin{cases} \dfrac{1}{log\zeta}\log(L_i(\theta)+\zeta), & L_i(\theta) < \lambda \\ 0, & L_i(\theta) \geq \lambda \end{cases}$$

$$(\zeta = 1-\lambda, \quad 0 < \lambda < 1)$$




$\boldsymbol{L_i(\theta)}$의 값이 커질수록, 적은 가중치 반영

# Self-paced Curriculum Learning

## 핵심 아이디어

❖ Curriculum region

〈Curriculum region : **사전지식** 반영〉

$$\gamma: X \rightarrow \{1,2,\ldots,N\},$$

(사전지식 : Training Samples X의 학습 난이도 순위 매기기)

$$\Psi = \{v | a^T v \leq c\}$$

γ에 의해 정의된 a　　상수

예제 :

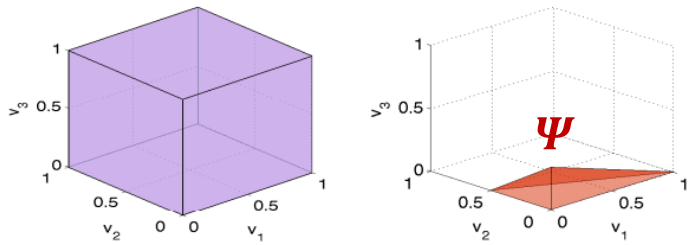
$$x_1, x_2, x_3$$
$$c = 1$$

$$\gamma_1 = 1,\quad \gamma_2 = 2,\quad \gamma_3 = 3$$

(사전지식을 통한 순위 매기기)

$$v_1 \leq 1,\quad 2v_2 \leq 1,\quad 3v_3 \leq 1$$

(Curriculum Region 도출)




(a) SPL　　(b) SPCL

Figure 1: Comparison of feasible regions in SPL and SPCL.

**실제 학습 프로세스에서 약한 영향**

초기 학습 단계에서 데이터별 학습 가중치 $v_i$의 Initialization 역할

# 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
- 2021년 10월 8일 기준 인용 횟수 : 637회

**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.

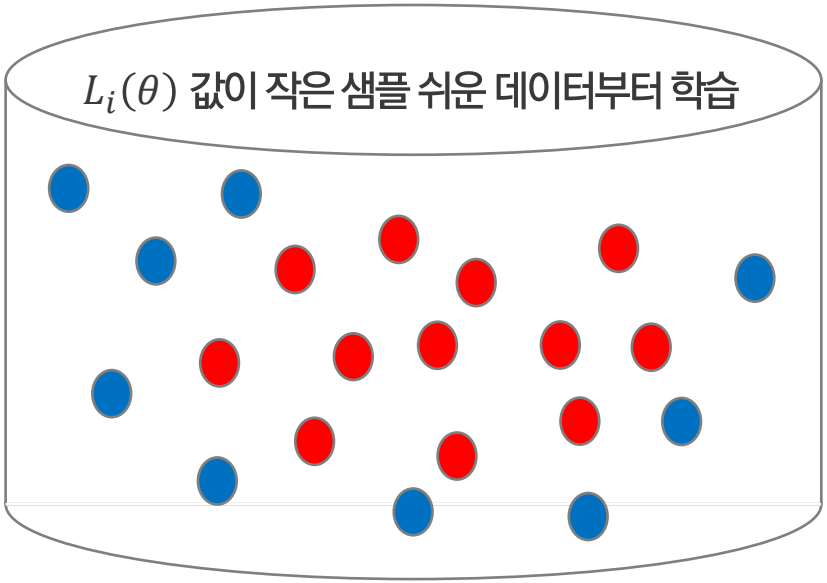
Inspired by the recent success of Curriculum Learning (CL), this paper tackles this problem using CL (Bengio et al., 2009), a learning paradigm inspired by the cognitive process of human and animals, in which a model is learned gradually using samples ordered in a meaningful sequence. A curriculum specifies 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

## 핵심 아이디어

❖ Curriculum Learning의 아이디어를 활용해 잘못 매겨진 라벨(Corrupted samples)의 오정보를 걸러내자!

〈Self-paced Curriculum Learning〉




: $L_i(\theta)$ 값이 작은 데이터 (Easy samples ) ⇨ 학습에 큰 가중치로 반영

: $L_i(\theta)$ 값이 큰 데이터 (Hard samples) ⇨ 학습 미반영 또는 작은 가중치로 반영

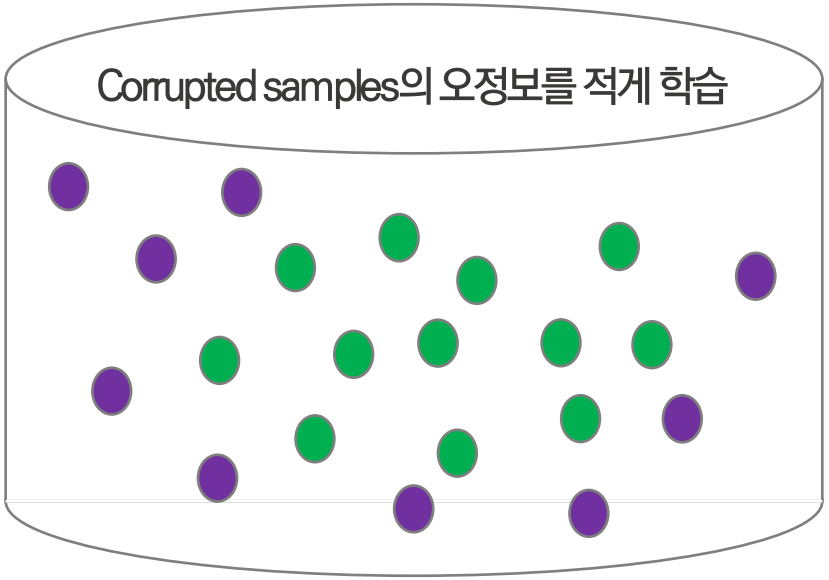
목적 : 쉬운 샘플에서 차츰 어려운 샘플 학습하여 모델 성능 향상

# MentorNet

## 핵심 아이디어

❖ Curriculum Learning의 아이디어를 활용해 잘못 매겨진 라벨(Corrupted samples)의 오정보를 걸러내자!

〈MentorNet〉




: $L_i(\theta)$ 값이 작은 데이터 (Clean samples) ⇨ 학습에 큰 가중치로 반영

: $L_i(\theta)$ 값이 큰 데이터 (Corrupted samples) ⇨ 학습 미반영 또는 작은 가중치로 반영

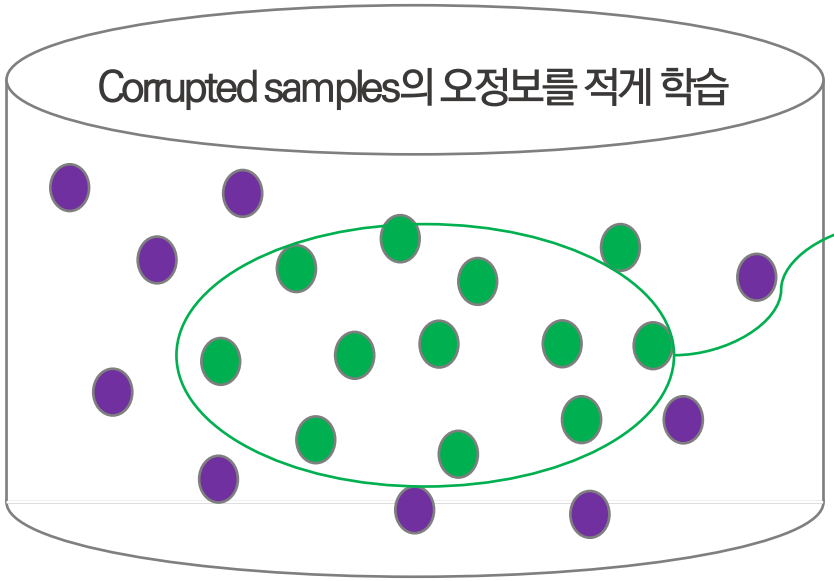
목적 : 오정보(Corrupted samples)에 대한 과적합 방지

# MentorNet

## 핵심 아이디어

❖ Curriculum Learning의 아이디어를 활용해 잘못 매겨진 라벨(Corrupted samples)의 오정보를 걸러내자!

〈MentorNet〉




MentorNet의 가정 :
학습 데이터 중 정확히 라벨이 매겨진
Clean samples을 반드시 확보하고 있어야 함
(전체 데이터의 10%)

: $L_i(\theta)$ 값이 작은 데이터 (Clean samples) ⇨ 학습에 큰 가중치로 반영

: $L_i(\theta)$ 값이 큰 데이터 (Corrupted samples) ⇨ 학습 미반영 또는 작은 가중치로 반영

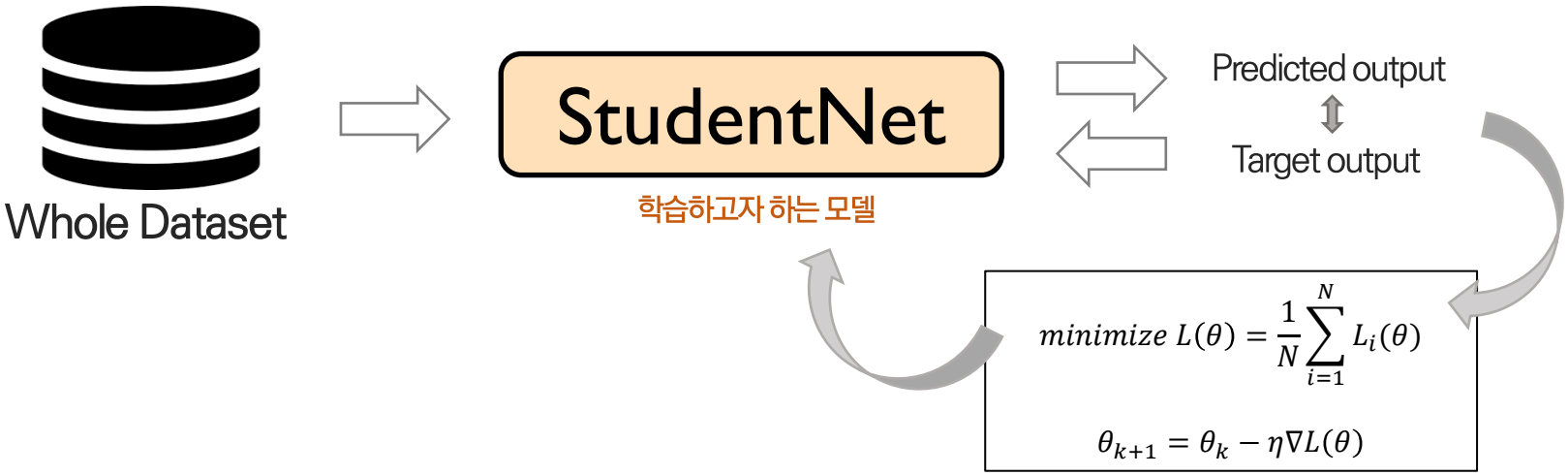
목적 : 오정보(Corrupted samples)에 대한 과적합 방지

# MentorNet

## 모델 구조

❖ MentorNet이라는 추가적인 모델을 통해서 샘플 noise 판별

- 학습 전체의 21%, 75% 에 해당하는 epoch에서 StudentNet을 고정하고 MentorNet을 Clean set을 이용하여 학습
- Clean set에서 임의로 noise 생성 (학습 데이터와 동일한 Corrupted samples 비율)
- MentorNet을 학습하기 전인 20%의 epoch동안은 $v_i \sim Bernoulli(p)$ 적용 (확률 p로 데이터 샘플을 dropout하는 것과 동일)

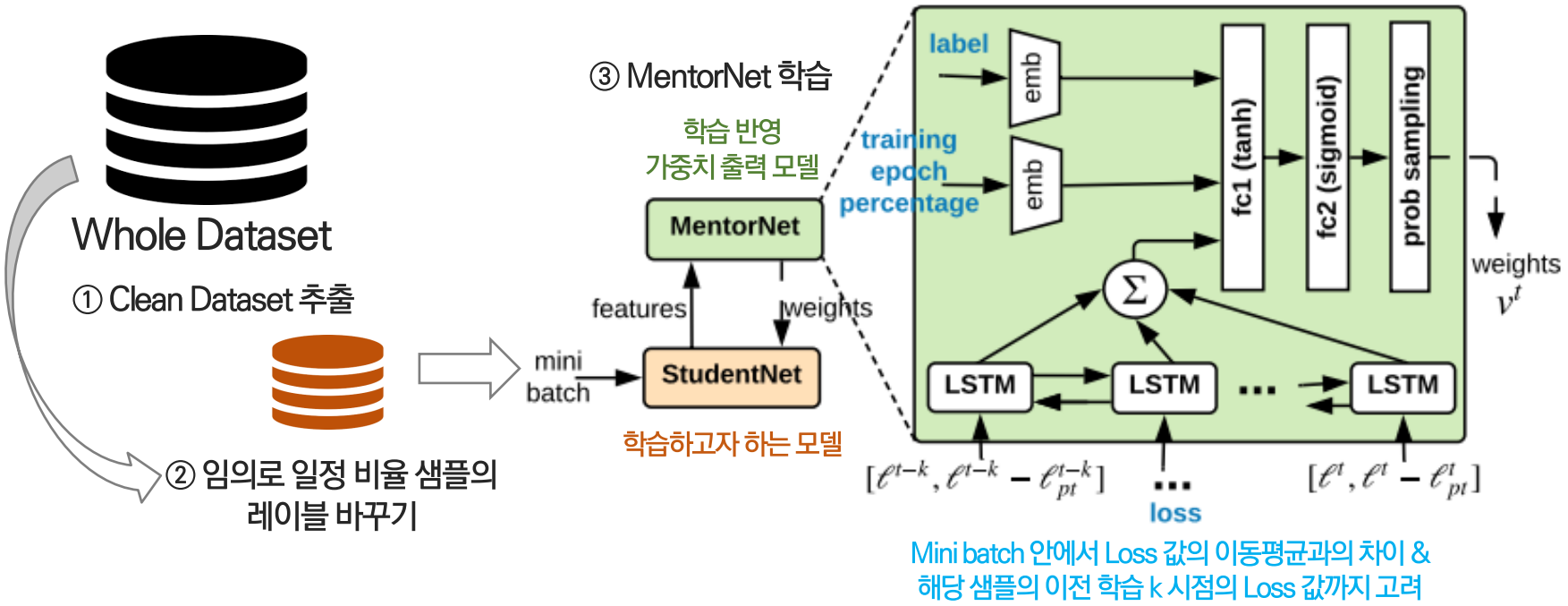


# MentorNet

## 모델 구조

❖ MentorNet이라는 추가적인 모델을 통해서 샘플 noise 판별

- 학습 전체의 21%, 75% 에 해당하는 epoch에서 StudentNet을 고정하고 MentorNet을 Clean set을 이용하여 학습
- Clean set에서 임의로 noise 생성 (학습 데이터와 동일한 Corrupted samples 비율)
- MentorNet을 학습하기 전인 20%의 epoch동안은 $v_i \sim Bernoulli(p)$ 적용 (확률 p로 데이터 샘플을 dropout하는 것과 동일)

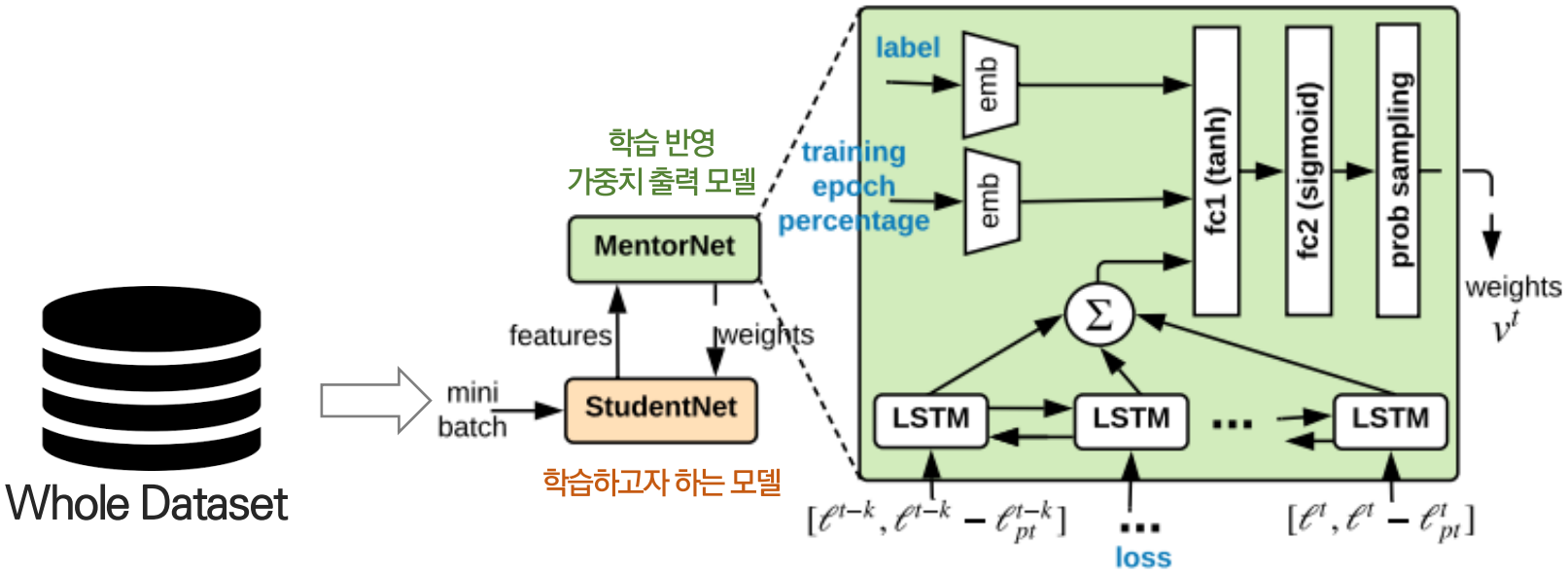


# MentorNet

## 모델 구조

❖ MentorNet이라는 추가적인 모델을 통해서 샘플 noise 판별

- 학습 전체의 21%, 75% 에 해당하는 epoch에서 StudentNet을 고정하고 MentorNet을 Clean set을 이용하여 학습
- Clean set에서 임의로 noise 생성 (학습 데이터와 동일한 Corrupted samples 비율)
- MentorNet을 학습하기 전인 20%의 epoch동안은 $v_i \sim Bernoulli(p)$ 적용 (확률 p로 데이터 샘플을 dropout하는 것과 동일)




학습된 MentorNet을 통해 학습 가중치 $v_i^*$ 산출

# MentorNet

## 실험 – Noisy CIFAR-10 & CIFAR-100

- ❖ 모든 라벨이 정확히 매겨진 CIFAR-10 & CIFAR-100에 일정 비율로 라벨을 잘못 매겨 Corrupted samples 생성
- ❖ MentorNet의 경우 Corrupted samples의 학습 반영을 줄이기에 오정보에 의한 성능 저하가 적음

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

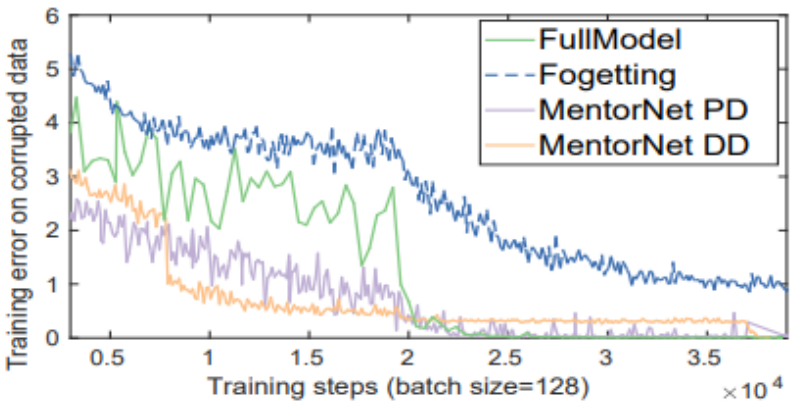
| | Resnet-101 StudentNet | | | | | | Inception StudentNet | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CIFAR-100 | | | CIFAR-10 | | | CIFAR-100 | | | CIFAR-10 | | |
| Method | 0.2 | 0.4 | 0.8 | 0.2 | 0.4 | 0.8 | 0.2 | 0.4 | 0.8 | 0.2 | 0.4 | 0.8 |
| FullModel | 0.60 | 0.45 | 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** |

[Corrupted samples 비율에 따른 성능]

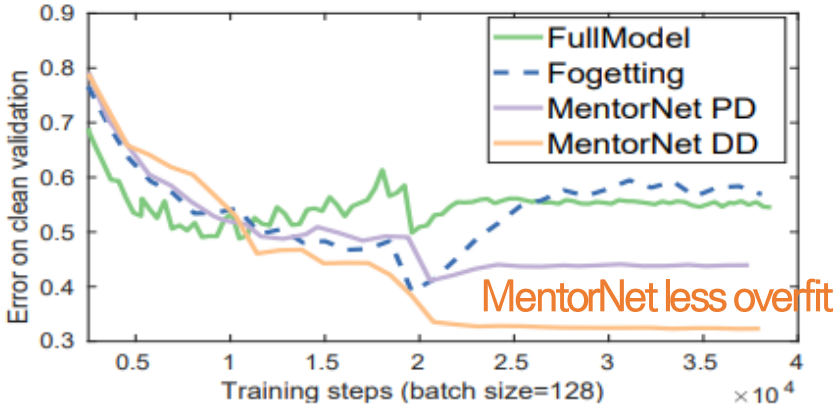
# MentorNet

## 실험 – Noisy CIFAR-10 & CIFAR-100

- ❖ 모든 라벨이 정확히 매겨진 CIFAR-10 & CIFAR-100에 일정 비율로 라벨을 잘못 매겨 Corrupted samples 생성
- ❖ MentorNet의 경우 Corrupted samples의 학습 반영을 줄이기에 오정보에 의한 성능 저하가 적음




(b) Training error on corrupted labels




(a) Test error on clean validation

[epochs에 따른 Training error 및 Test error]

# Conclusion

연구 동향 및 결론

❖ 잘못 매겨진 라벨(Corrupted samples)이 존재하는 상황에 적용

- Robust Curriculum Learning: from clean label detection to noisy label self-correction. (ICLR 2021)
- Robust Early-Learning: Hindering The Memorization of Noisy Labels. (ICLR 2021)

❖ 클래스 불균형 문제에 적용

- Dynamic Curriculum Learning for Imbalanced Data Classification. (ICCV 2019)
- Semi-Supervised Semantic Segmentation via Dynamic Self-Training and Class-Balanced Curriculum. (arXiv)

❖ 최근 강화학습 분야에서 활발히 연구

- Evolutionary Population Curriculum for Scaling Multi-Agent Reinforcement Learning. (ICLR 2020)
- Self-Paced Deep Reinforcement Learning. (NeurIPS 2020)
- Self-Paced Context Evaluation for Contextual Reinforcement Learning (ICML 2021)

❖ 낮은 성능을 보이는 실제 현실 문제 데이터에서 학습 난이도에 따른 학습 전략 반영은 효과적으로 활용될 수 있음

# Thank you

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

E-mail : dawonksh@korea.ac.kr

# 참고 문헌


- ✓ Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41–48).

- ✓ Kumar, M., Packer, B., & Koller, D. (2010). Self-paced learning for latent variable models. Advances in neural information processing systems, 23, 1189–1197.

- ✓ Jiang, L., Meng, D., Zhao, Q., Shan, S., & Hauptmann, A. G. (2015, February). Self-paced curriculum learning. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

- ✓ 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.