Introduction to Knowledge Distillation

2020.12.11

Data Mining & Quality Analytics Lab.

발표자 : 황하은

julyh777@korea.ac.kr

발표자 소개

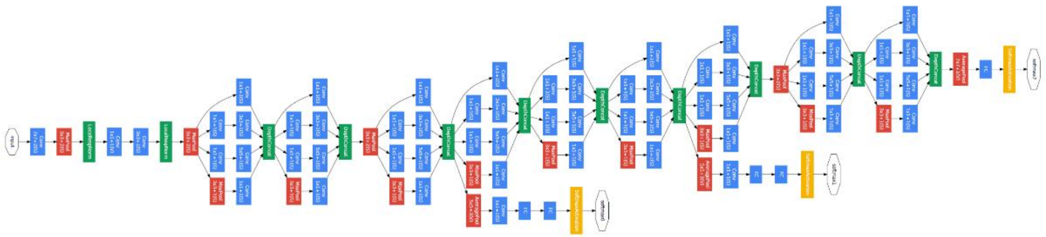
- **황하은 (Haeun Hwang)**
  - 고려대학교 산업경영공학부 재학 중
  - Data Mining & Quality Analytics Lab (김성범 교수님)
  - 석사과정 (2020.03 ~)

- 관심 연구 분야
  - Machine Learning / Deep Learning
  - Knowledge Distillation

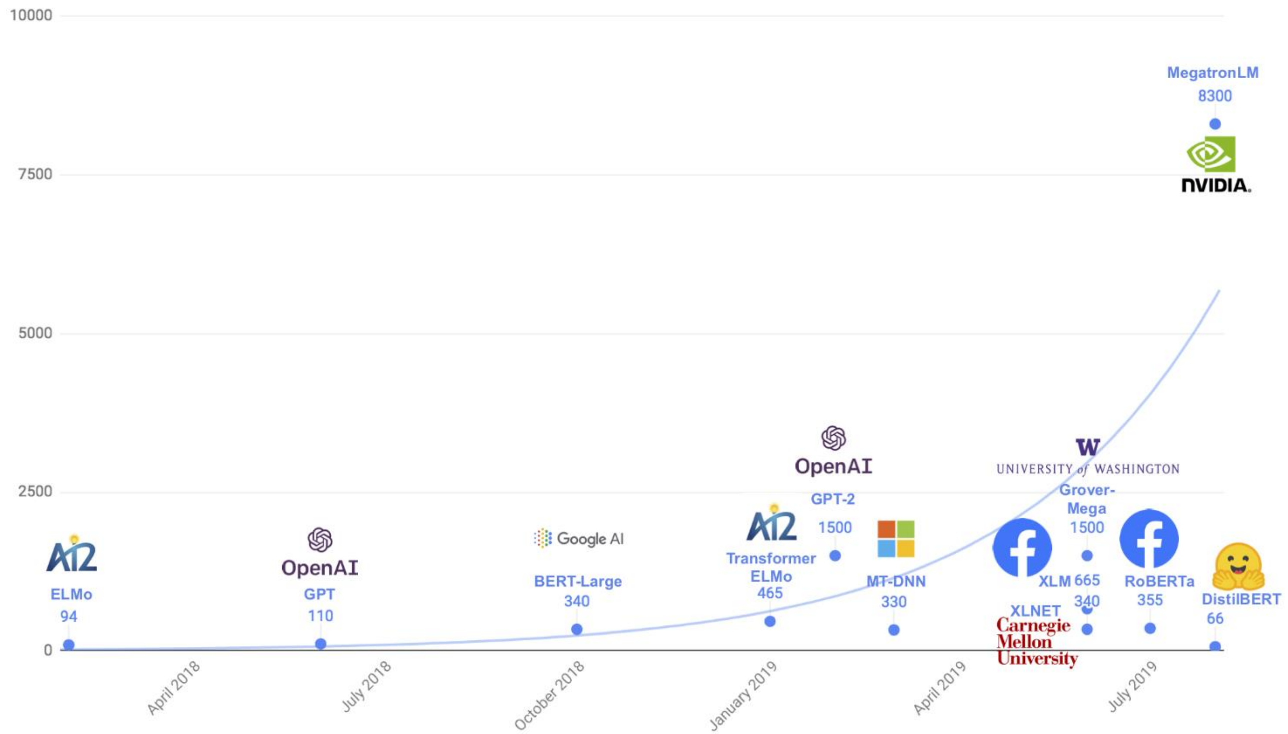
INDEX

1. Introduction

등장배경



복잡한 모델 구조



파라미터수의 기하급수적 증가

Data Mining
Quality Analytics

출처: https://medium.com/huggingface/distilbert-8cf3380435b5

1. Introduction
등장배경

메모리 한계

추론 시간 증가

실용적 측면에서의 한계

⋮

1. Introduction
등장배경

- 경량 딥러닝
  - 경량 알고리즘
    - 모델 구조 변경
    - 합성곱 필터 변경
    - 자동 모델 탐색
  - 알고리즘 경량화
    - 모델 압축
    - 지식 증류
    - 모델 압축 자동 탐색

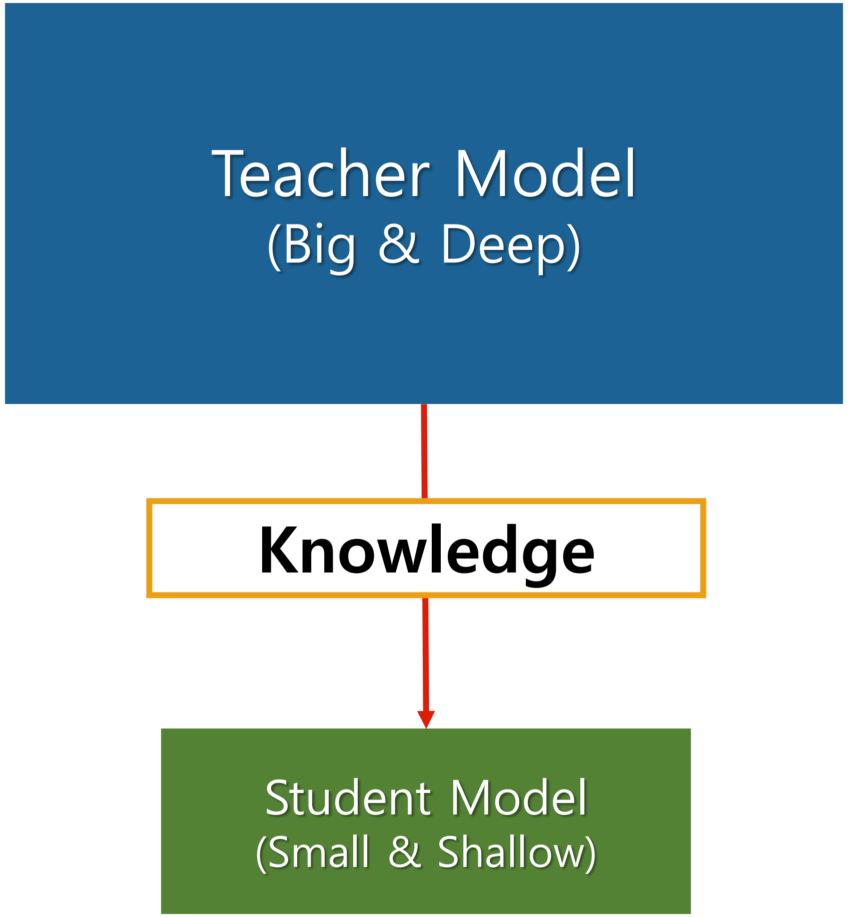
1. Introduction

등장배경

- 경량 딥러닝
  - 경량 알고리즘
    - 모델 구조 변경
    - 합성곱 필터 변경
    - 자동 모델 탐색
  - 알고리즘 경량화
    - 모델 압축
    - 지식 증류
    - 모델 압축 자동 탐색

1. Introduction

Knowledge Distillation이란?



Teacher 모델: 높은 예측 정확도를 가진 복잡한 모델

e.g. 정확도 : 95 %

추론 시간 : 2시간

잘 학습된 Teacher 모델의 지식을 전달하여
단순한 Student모델로 비슷한 좋은 성능을 내고자 함

Student 모델: Teacher 모델의 지식을 받는 단순한 모델

e.g. 정확도 : 90 %

추론 시간 : 5분

1. Introduction

Knowledge Distillation이란?

What

Teacher모델의 **어떠한 지식**을

How

Student 모델에 **어떻게 전달**할 것인가

Knowledge Distillation

2. 기본 Knowledge Distillation

Vanilla Knowledge distillation

❖ Distilling the Knowledge in a Neural Network

- 2014 Neural Information Processing Systems(NeurIPS)에서 발표된 논문
- 2020년 12월 2일 기준 4907회 인용

Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

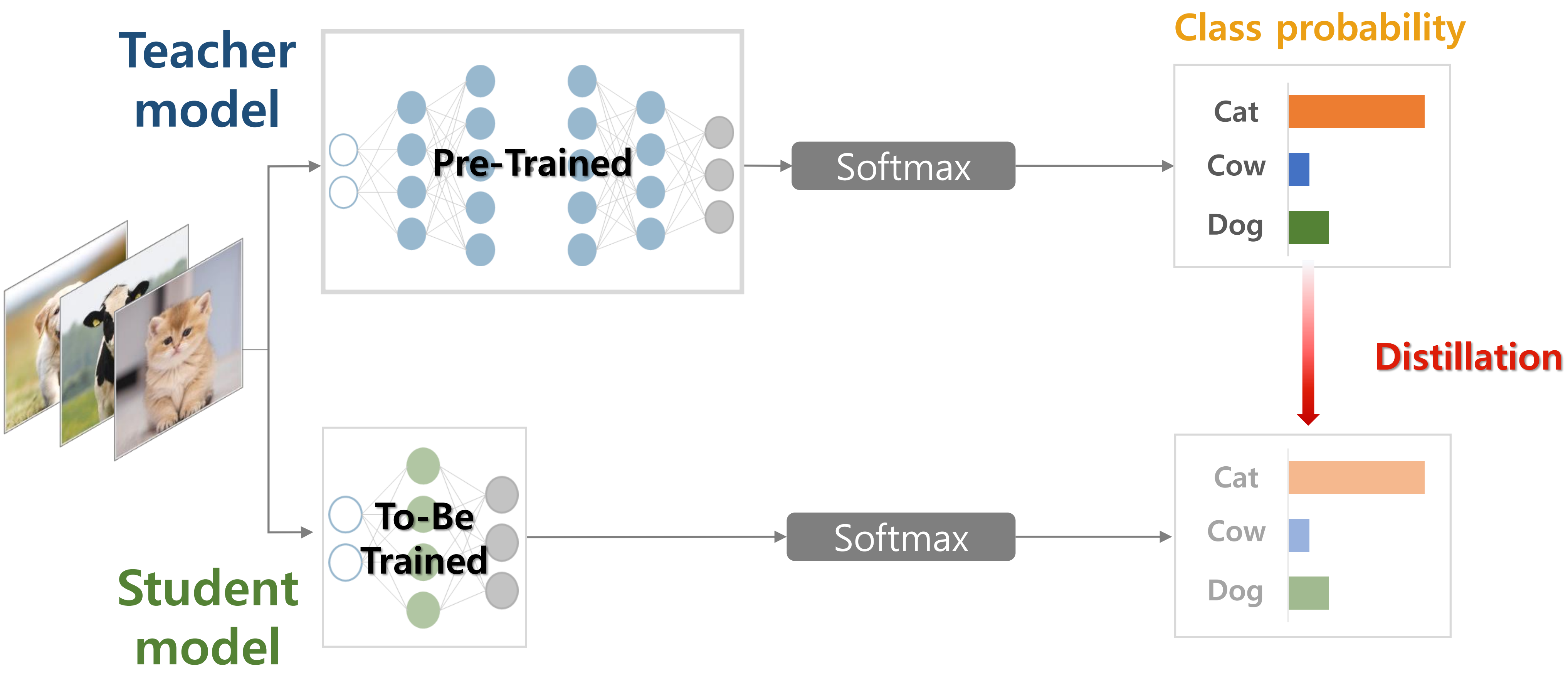
Jeff Dean
Google Inc.
Mountain View
jeff@google.com

Abstract

A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions [3]. Unfortunately, making predictions using a whole ensemble of models is cumbersome and may be too computationally expensive to allow deployment to a large number of users, especially if the individual models are large neural nets. Caruana and his collaborators [1] have shown that it is possible to

2. 기본 Knowledge Distillation

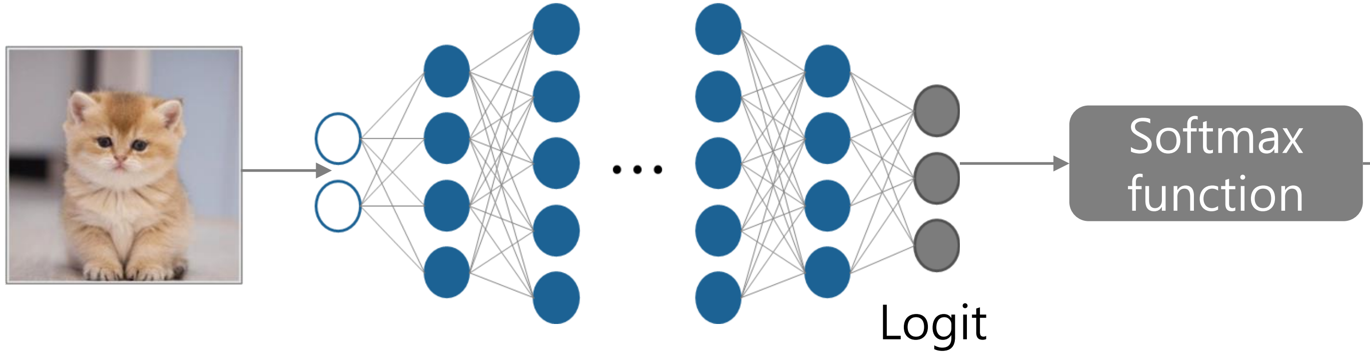
전체 프레임워크



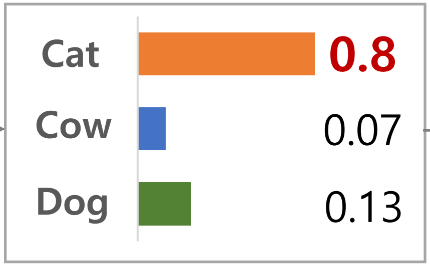
2. 기본 Knowledge Distillation

Hard Target

Input



Class probability



Output

Cat	1
Dog	0
Cow	0

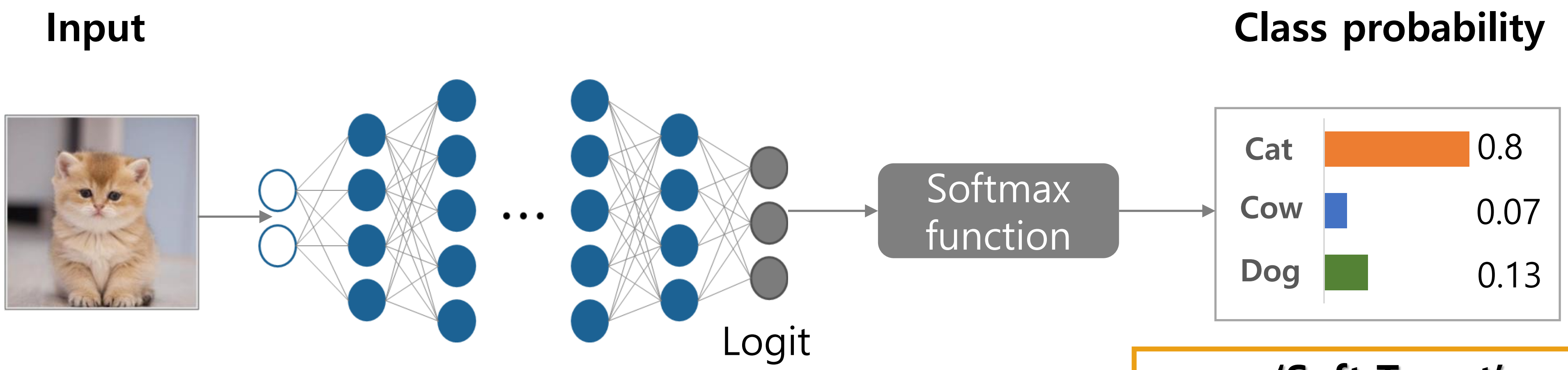
One-hot Encoding

'Hard Target'

2. 기본 Knowledge Distillation

Soft Target

Knowledge : Soft Target 사용



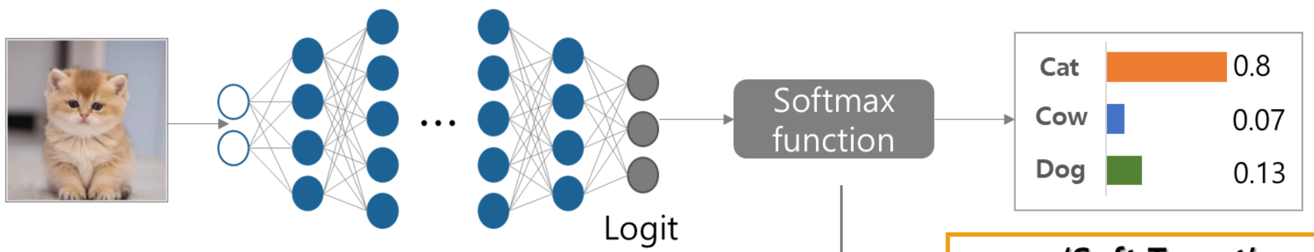
'Soft Target'

= 예측결과의 확률분포

2. 기본 Knowledge Distillation

Temperature

Knowledge : Soft Target 사용



'Soft Target'
= 예측결과의 확률분포

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad \Rightarrow \quad Softmax(z_i) = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}$$

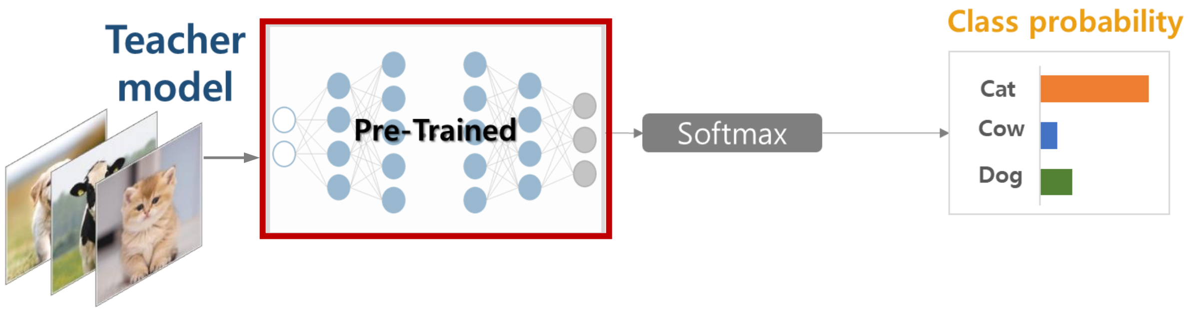
τ $(Temperature)$: Scaling 역할의 하이퍼 파라미터

- τ = 1일 때, 기존 softmax function과 동일
- τ클수록, 더 soft한 확률분포

2. 기본 Knowledge Distillation

지식 전달 방법

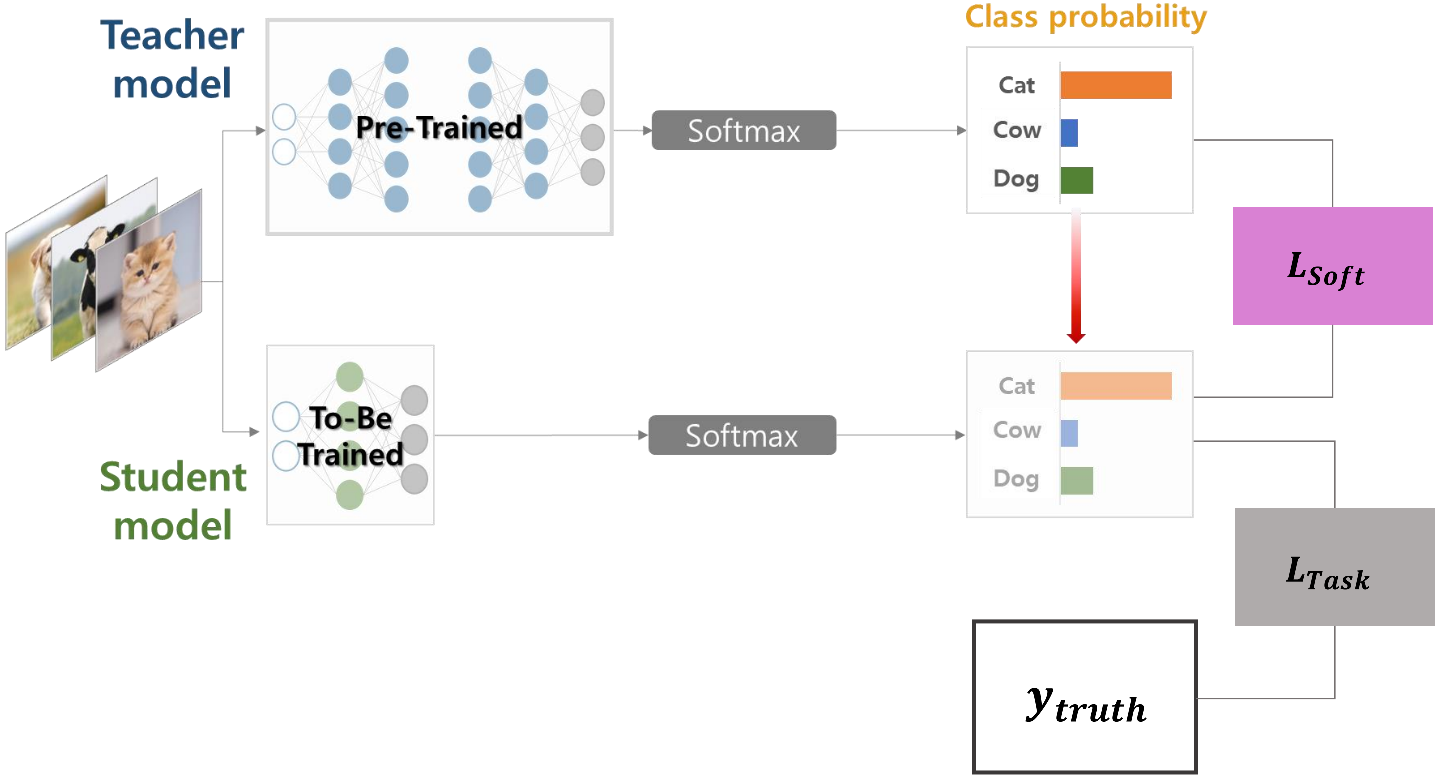
Distillation 방법 : Offline - distillation



Teacher 모델을 미리 학습

2. 기본 Knowledge Distillation
지식 전달 방법

Distillation 방법 : Offline - distillation



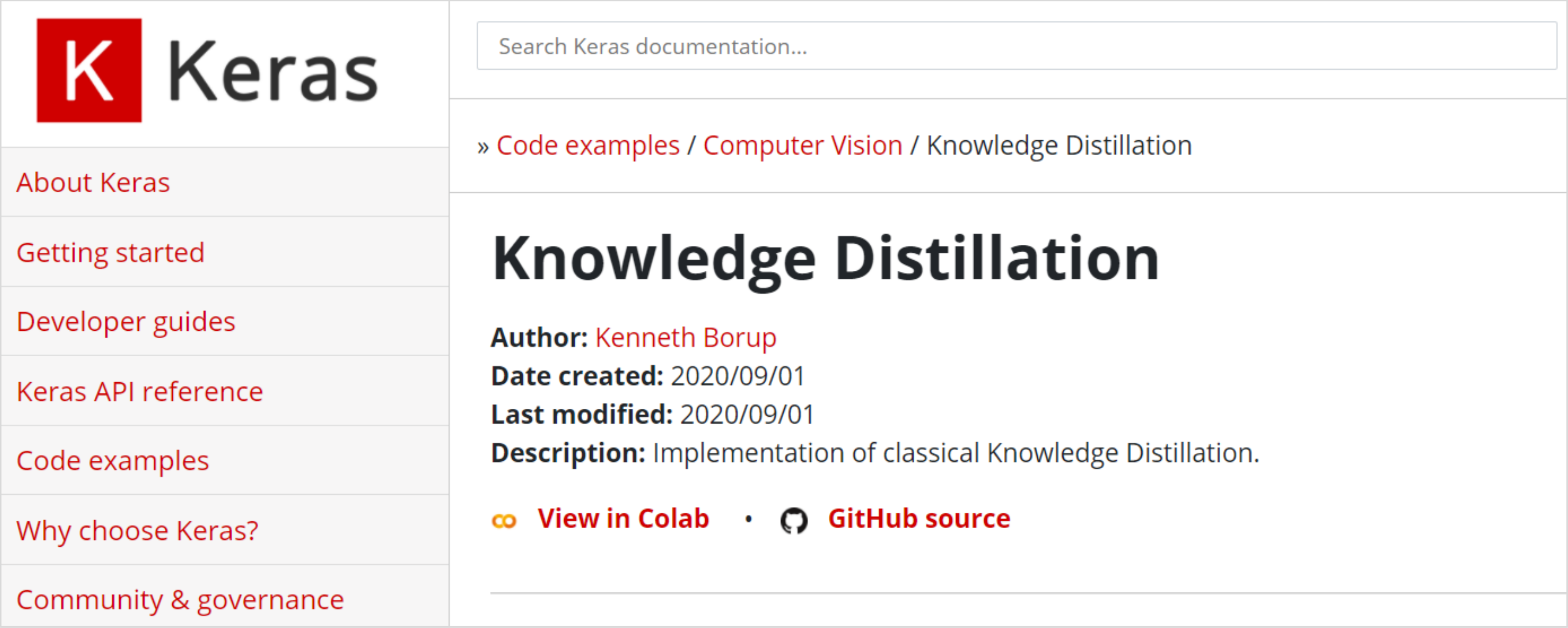
- $f_T(x_i)$: Teacher 모델의 logit 값
- $f_T(x_i)$: Student 모델의 logit 값
- τ :Scaling 역할의 하이퍼 파라미터

$$L_{Soft} = \sum_{x_i \in X} KL(softmax(\frac{f_T(x_i)}{\tau}), softmax(\frac{f_s(x_i)}{\tau}))$$

$$L_{Task} = CrossEntropy\ (softmax(f_s(x_i)), y_{truth})$$

$$Student\ L_{Total} = L_{Task} + \lambda \cdot L_{Soft}$$

딥러닝 라이브러리 Keras에서 함수 제공



https://keras.io/examples/vision/knowledge_distillation/

2. 기본 Knowledge Distillation

다양한 Knowledge Distillation 알고리즘

Teacher모델의 **어떠한 지식**을 | Student 모델에 **어떻게 전달**할 것인가

Knowledge Distillation

Response - Based

Feature - Based

Relation - Based

Offline - Distillation

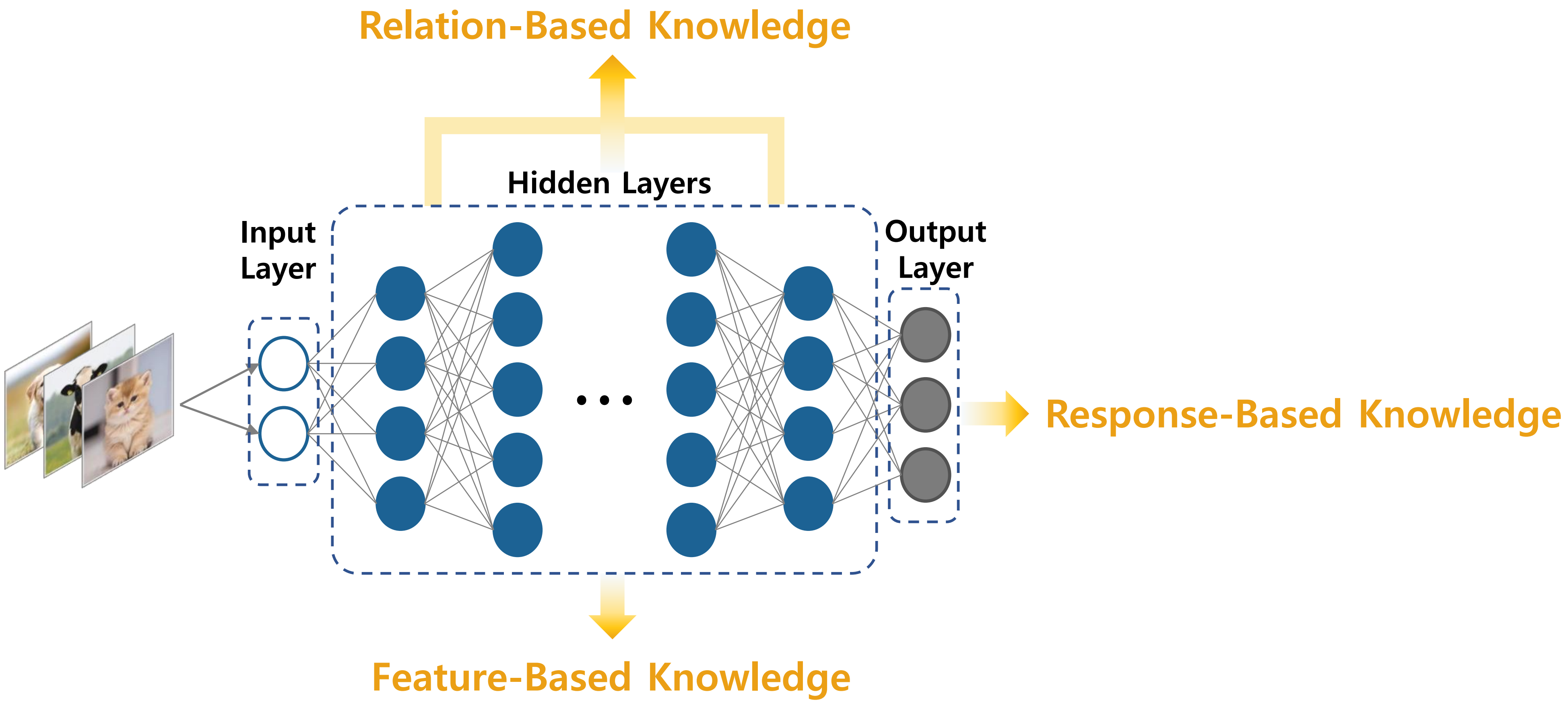
Online - Distillation

Self – Distillation

3. Knowledge 관점 연구

어떠한 지식을 넘길 것 인가

Knowledge Distillation

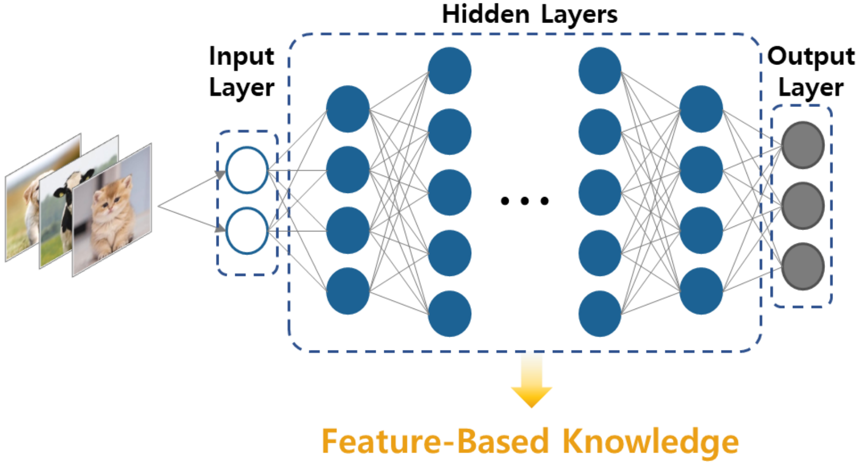


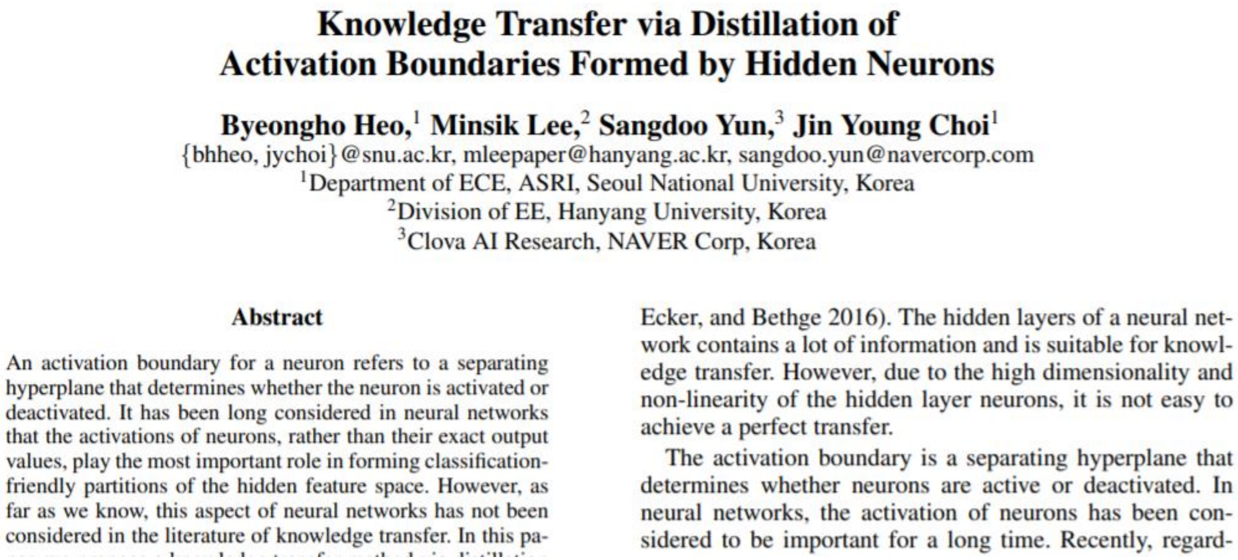
3. Knowledge 관점 연구

Feature – Based knowledge

❖ Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons

- 2019 Association for the Advancement of Artificial Intelligence(AAAI)에 발표된 논문
- 2020년 12월 2일 기준 53회 인용





Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons

Byeongho Heo,[1] Minsik Lee,[2] Sangdoo Yun,[3] Jin Young Choi[1]
{bhheo, jychoi}@snu.ac.kr, mleepaper@hanyang.ac.kr, sangdoo.yun@navercorp.com
[1]Department of ECE, ASRI, Seoul National University, Korea
[2]Division of EE, Hanyang University, Korea
[3]Clova AI Research, NAVER Corp, Korea

Abstract

An activation boundary for a neuron refers to a separating hyperplane that determines whether the neuron is activated or deactivated. It has been long considered in neural networks that the activations of neurons, rather than their exact output values, play the most important role in forming classification-friendly partitions of the hidden feature space. However, as far as we know, this aspect of neural networks has not been considered in the literature of knowledge transfer. In this pa-

Ecker, and Bethge 2016). The hidden layers of a neural network contains a lot of information and is suitable for knowledge transfer. However, due to the high dimensionality and non-linearity of the hidden layer neurons, it is not easy to achieve a perfect transfer.

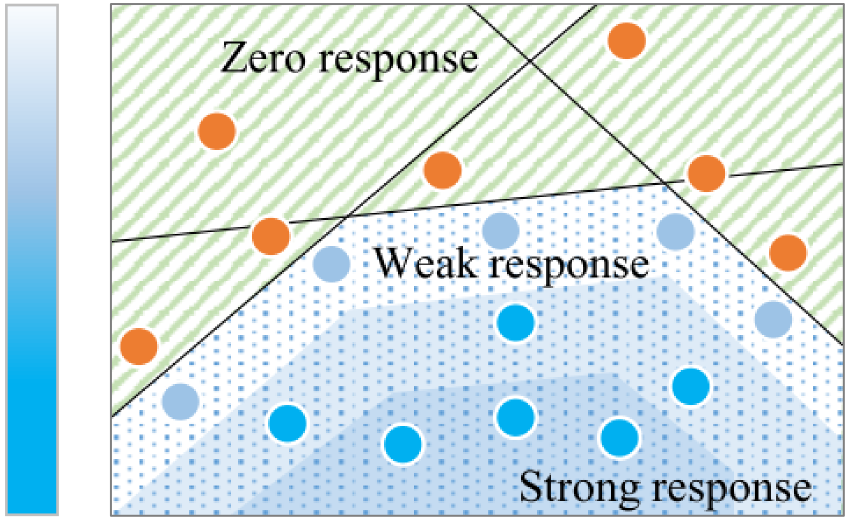
The activation boundary is a separating hyperplane that determines whether neurons are active or deactivated. In neural networks, the activation of neurons has been considered to be important for a long time. Recently, regard-

3. Knowledge 관점 연구

Feature – Based knowledge

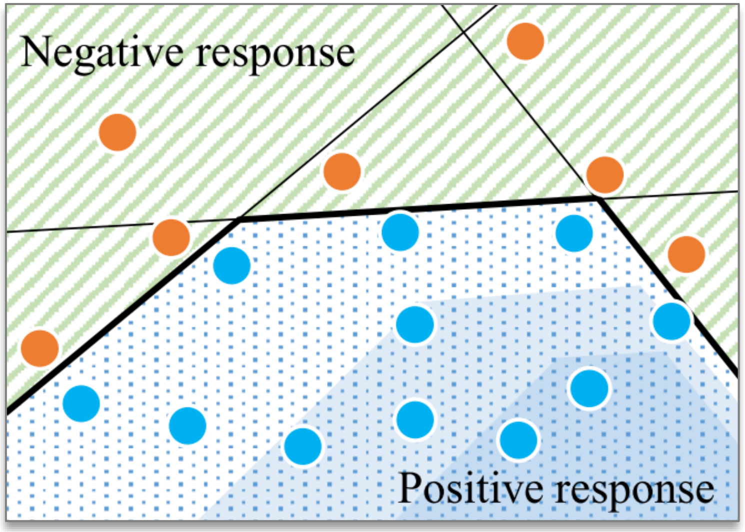
Knowledge: Teacher 모델의 Activation boundary

기존 방법



Magnitude : 해당 클래스에 속하는 정도

제안 방법

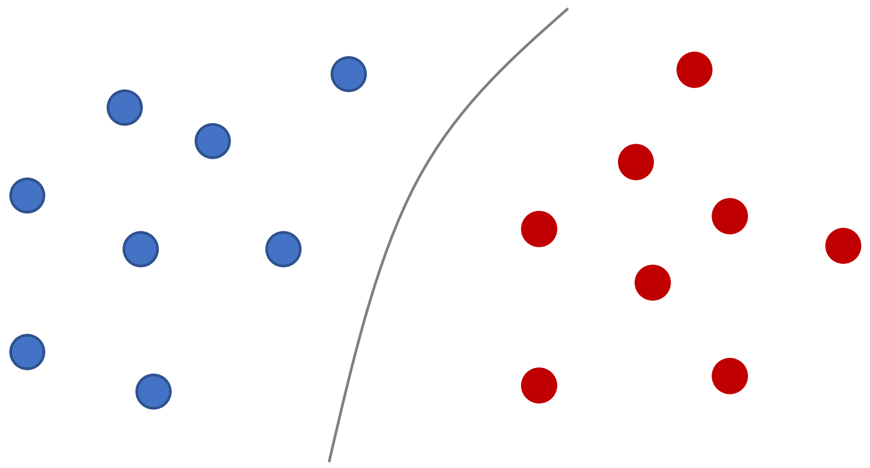


3. Knowledge 관점 연구

Feature – Based knowledge

❖ Activation boundary를 가져오는 이유

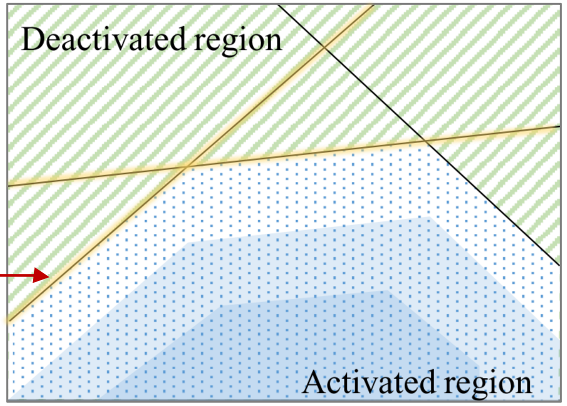
좋은 Decision boundary → **좋은 일반화 성능**



❖ Activation boundary를 가져오는 이유

좋은 **Decision boundary** → 좋은 일반화 성능

각 Hidden layer의
Activation boundary 조합으로 구성



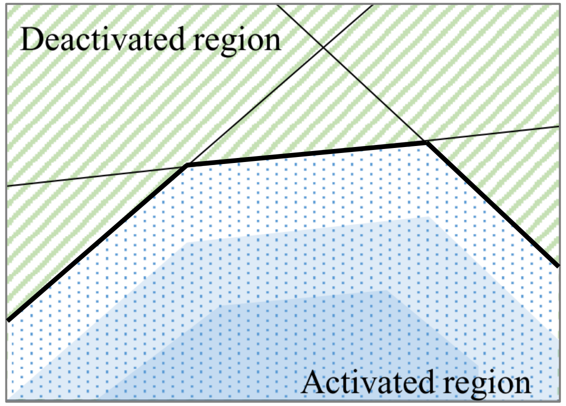
3. Knowledge 관점 연구

Feature – Based knowledge

❖ Activation boundary를 가져오는 이유

좋은 Decision boundary → 좋은 일반화 성능

각 Hidden layer의
Activation boundary 조합으로 구성



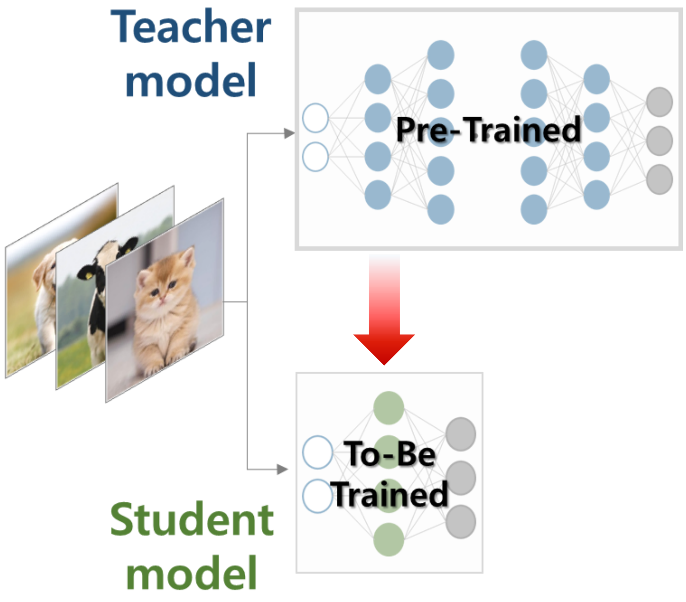
3. Knowledge 관점 연구

Feature – Based knowledge

목표: Teacher와 Student의 Activation Boundary만 같아지도록 학습

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

- $\mathbf{1}$: 모든 성분이 1로 구성된 벡터
- $\odot$: element-wise 곱
- $T(x_i)$: Teacher 모델의 히든 레이어의 반응벡터
- $S(x_i)$: Student 모델의 히든 레이어의 반응벡터
- μ : 분류경계면의 margin



$$L_{activation} = \left\| \rho\left(T(x_i)\right) - \rho\left(S(x_i)\right) \right\|_1$$

미분 불가능

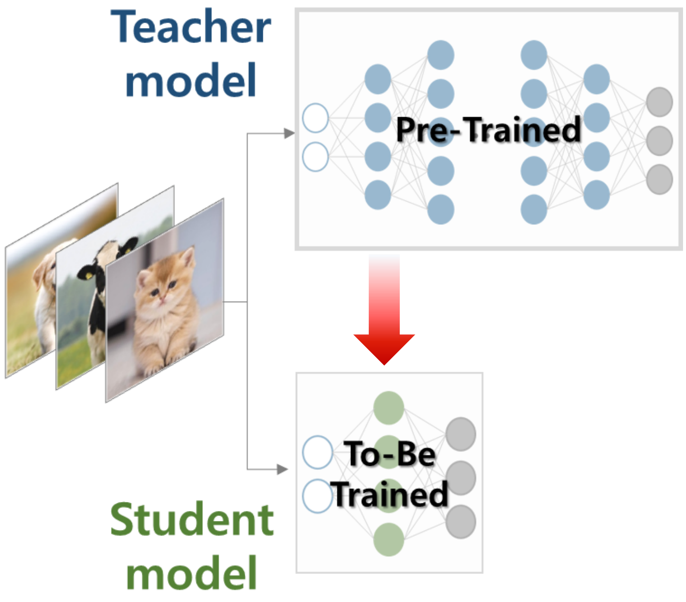
3. Knowledge 관점 연구

Feature – Based knowledge

목표: Teacher와 Student의 Activation Boundary만 같아지도록 학습

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

- $\mathbf{1}$: 모든 성분이 1로 구성된 벡터
- $\odot$: element-wise 곱
- $\sigma(x)$: ReLU 함수
- $T(x_i)$: Teacher 모델의 히든 레이어의 반응벡터
- $S(x_i)$: Student 모델의 히든 레이어의 반응벡터
- μ : 분류경계면의 margin



$$L_{activation} = \left\| \rho\big(T(x_i)\big) \odot \sigma\big(\mu\mathbf{1} - S(x_i)\big) + \big(\mathbf{1} - \rho\big(T(x_i)\big)\big) \odot \sigma\big(\mu\mathbf{1} + S(x_i)\big) \right\|_2^2$$

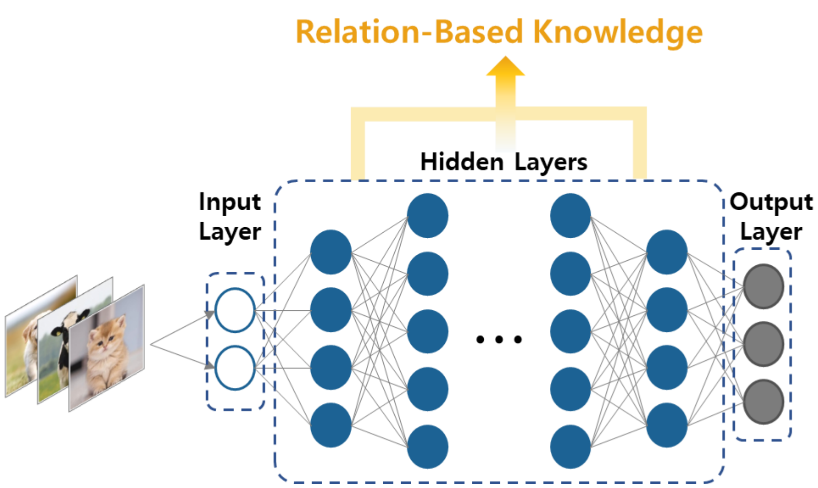
미분 가능

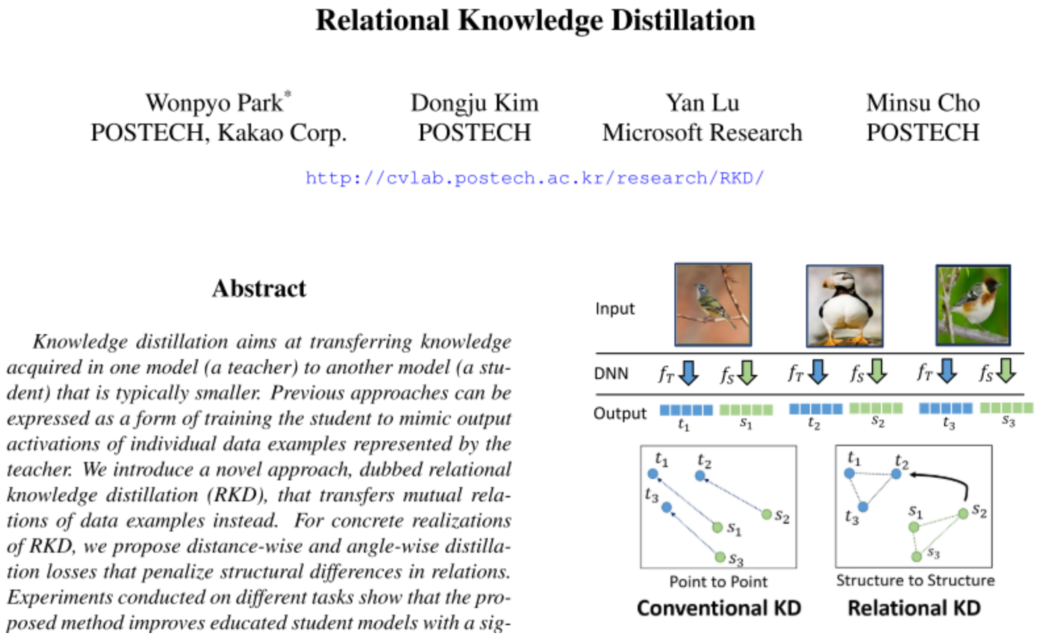
3. Knowledge 관점 연구

Relation – Based knowledge

❖ Relational Knowledge Distillation

- 2019 Computer Vision and Pattern Recognition (CVPR)에 발표된 논문
- 2020년 12월 2일 기준 110회 인용





Relational Knowledge Distillation

Wonpyo Park* POSTECH, Kakao Corp.
Dongju Kim POSTECH
Yan Lu Microsoft Research
Minsu Cho POSTECH

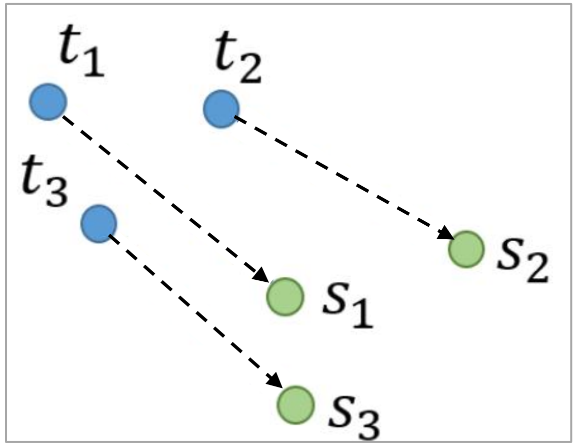
http://cvlab.postech.ac.kr/research/RKD/

Abstract

Knowledge distillation aims at transferring knowledge acquired in one model (a teacher) to another model (a student) that is typically smaller. Previous approaches can be expressed as a form of training the student to mimic output activations of individual data examples represented by the teacher. We introduce a novel approach, dubbed relational knowledge distillation (RKD), that transfers mutual relations of data examples instead. For concrete realizations of RKD, we propose distance-wise and angle-wise distillation losses that penalize structural differences in relations. Experiments conducted on different tasks show that the proposed method improves educated student models with a sig-

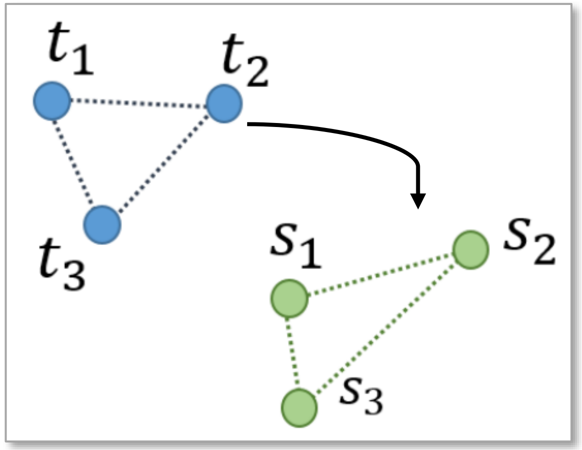
Knowledge: Teacher 모델을 통해 학습된 Activation의 구조

기존 방법



Individual wise

제안 방법

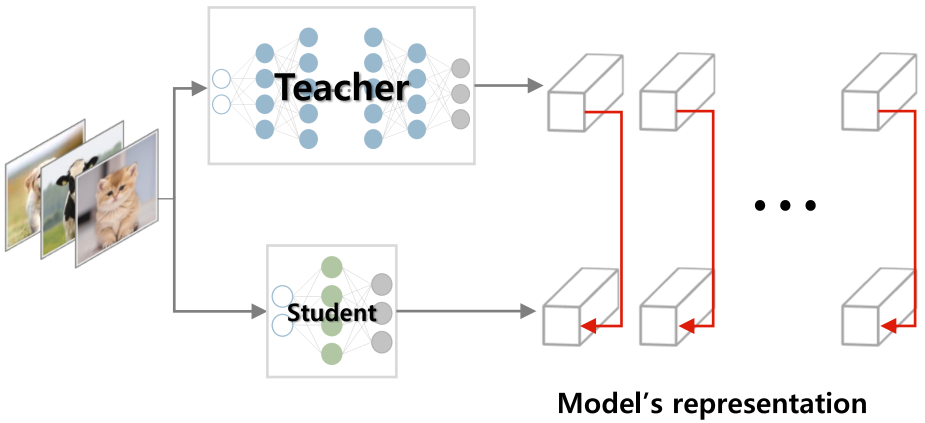


Structure wise

3. Knowledge 관점 연구
Relation – Based knowledge

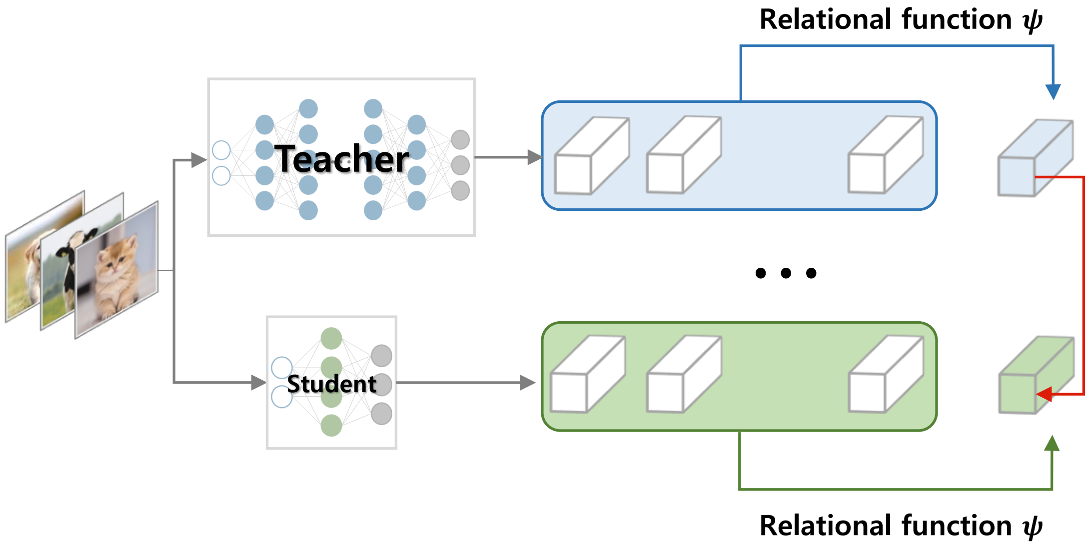
Knowledge: Teacher 모델을 통해 학습된 Activation의 구조

기존 방법



Individual wise

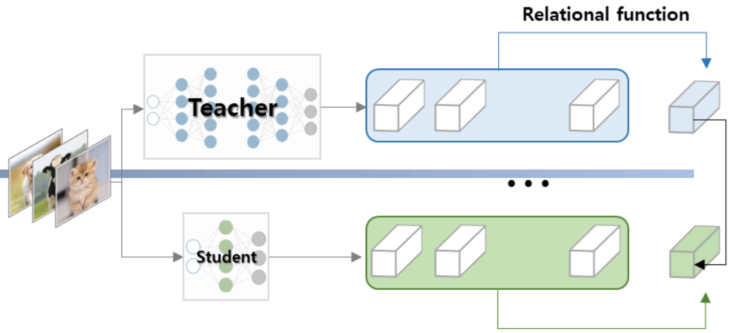
제안 방법



Structure wise

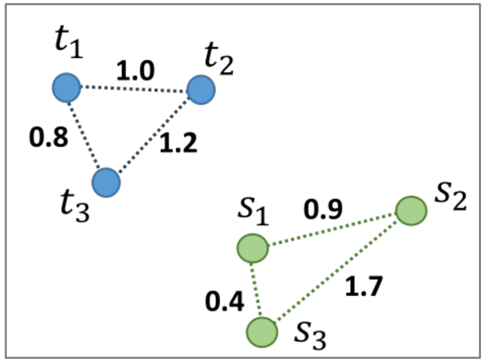
3. Knowledge 관점 연구

Relation – Based knowledge



Knowledge: Teacher 모델을 통해 학습된 Activation의 구조

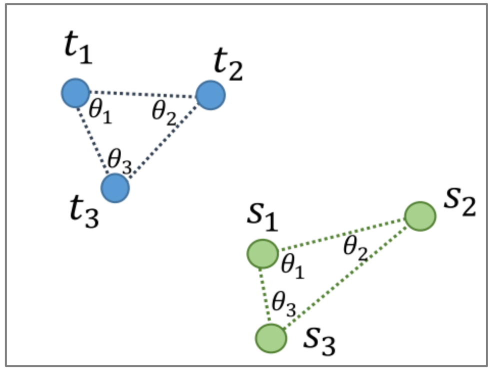
Distance-wise Relational function



$$\psi_D(t_i, t_j) = \frac{1}{\mu}\left\|t_i - t_j\right\|_2$$

$$\mu = \frac{1}{|\chi^2|}\sum_{(x_i, x_j)\in\chi^2}\left\|t_i - t_j\right\|_2$$

Angel-wise Relational function



$$\psi_A(t_i, t_j, t_k) = cos\angle t_i t_j t_k = \langle e^{ij}, e^{kj}\rangle$$

$$where\ e^{ij} = \frac{t_i - t_j}{\left\|t_i - t_j\right\|_2},\ e^{kj} = \frac{t_k - t_j}{\left\|t_k - t_j\right\|_2}$$

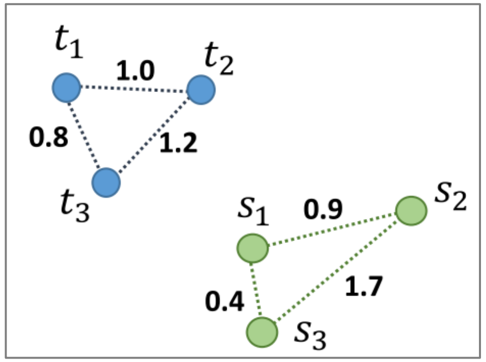
3. Knowledge 관점 연구

Relation – Based knowledge

Knowledge: **Teacher 모델을 통해 학습된 Activation의 구조**

$$Huber\ loss(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \le 1 \\ |x-y| - \frac{1}{2}, & \text{otherwise} \end{cases}$$

Distance-wise Relational function

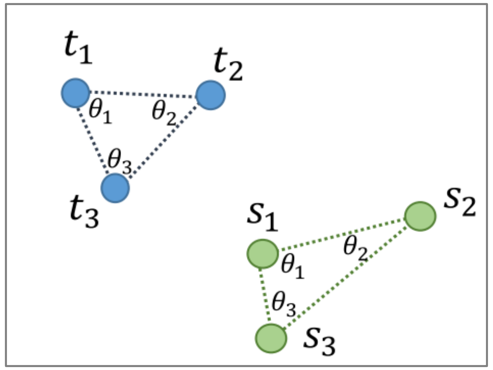


$$\psi_D(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2$$

$$\mu = \frac{1}{|\chi^2|}\sum_{(x_i, x_j)\in\chi^2}\|t_i - t_j\|_2$$

$$L_{RKD-D} = \sum_{(x_i, x_j)\in\chi^2} Huber\ loss(\psi_D(t_i, t_j), \psi_S(s_i, s_j))$$

Angel-wise Relational function



$$\psi_A(t_i, t_j, t_k) = cos\angle t_i t_j t_k = \langle e^{ij}, e^{kj}\rangle$$

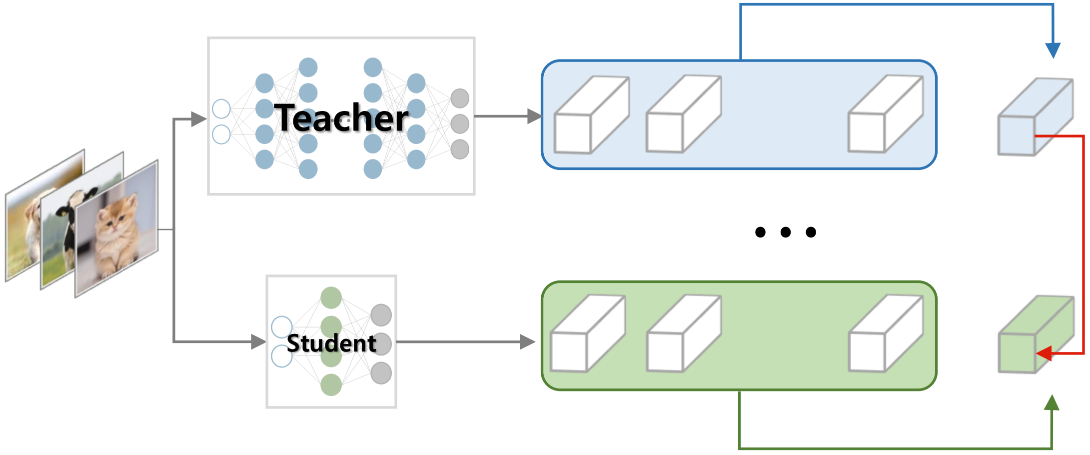
$$where\ e^{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2},\ e^{kj} = \frac{t_k - t_j}{\|t_k - t_j\|_2}$$

$$L_{RKD-A} = \sum_{(x_i, x_j, x_k)\in\chi^2} Huber\ loss(\psi_A(t_i, t_j, t_k), \psi_A(s_i, s_j, s_k))$$

3. Knowledge 관점 연구

Relation – Based knowledge

Knowledge: Teacher 모델을 통해 학습된 Activation의 구조



$$L_{RKD-D} = \sum_{(x_i, x_j) \in \chi^2} Huber\ loss(\psi_D(t_i, t_j), \psi_S(s_i, s_j))$$

or

$$L_{RKD-A} = \sum_{(x_i, x_j, x_k) \in \chi^2} Huber\ loss(\psi_A(t_i, t_j, t_k), \psi_A(s_i, s_j, s_k))$$

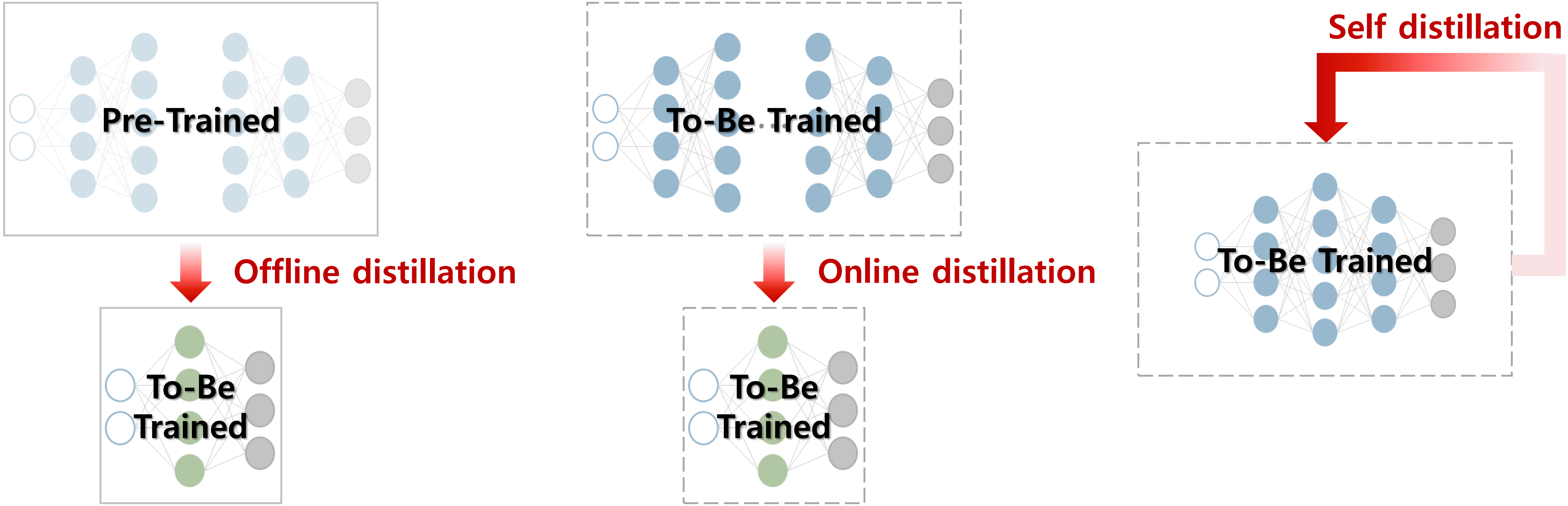
$$L_{Task} = CrossEntropy\ (softmax(f_s(x_i)), y_{truth})$$

$$Student\ L_{Total} = L_{Task} + \lambda \cdot L_{RKD}$$

4. Distillation 관점 연구

어떻게 지식을 넘길 것인가

Knowledge **Distillation**

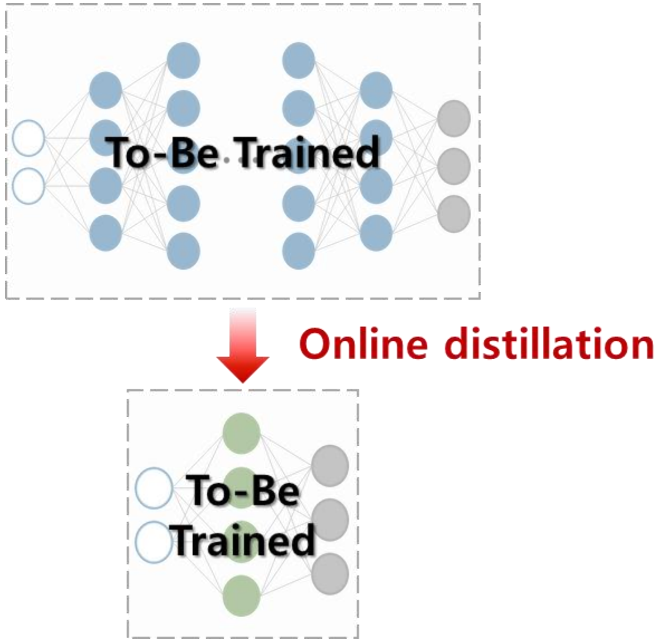


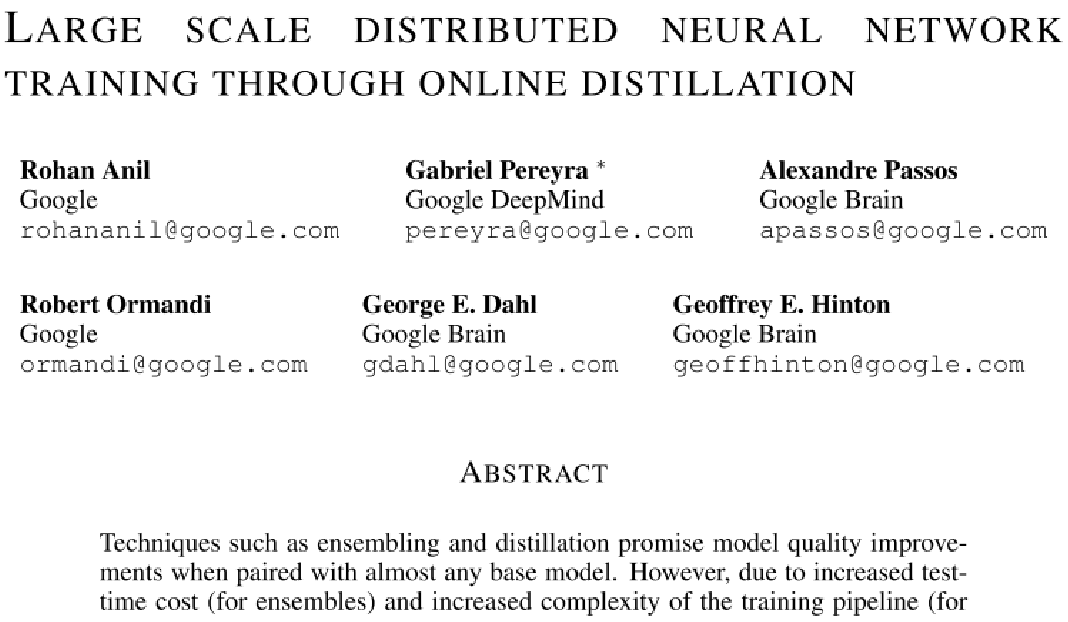
4. Distillation 관점 연구

Online – distillation

❖ Large scale distributed neural network training through online distillation

- 2018년 International Conference on Learning Represntations(ICLR)에서 발표된 논문
- 2020년 12월 2일 기준 110회 인용





LARGE SCALE DISTRIBUTED NEURAL NETWORK TRAINING THROUGH ONLINE DISTILLATION

Rohan Anil
Google
rohananil@google.com

Gabriel Pereyra *
Google DeepMind
pereyra@google.com

Alexandre Passos
Google Brain
apassos@google.com

Robert Ormandi
Google
ormandi@google.com

George E. Dahl
Google Brain
gdahl@google.com

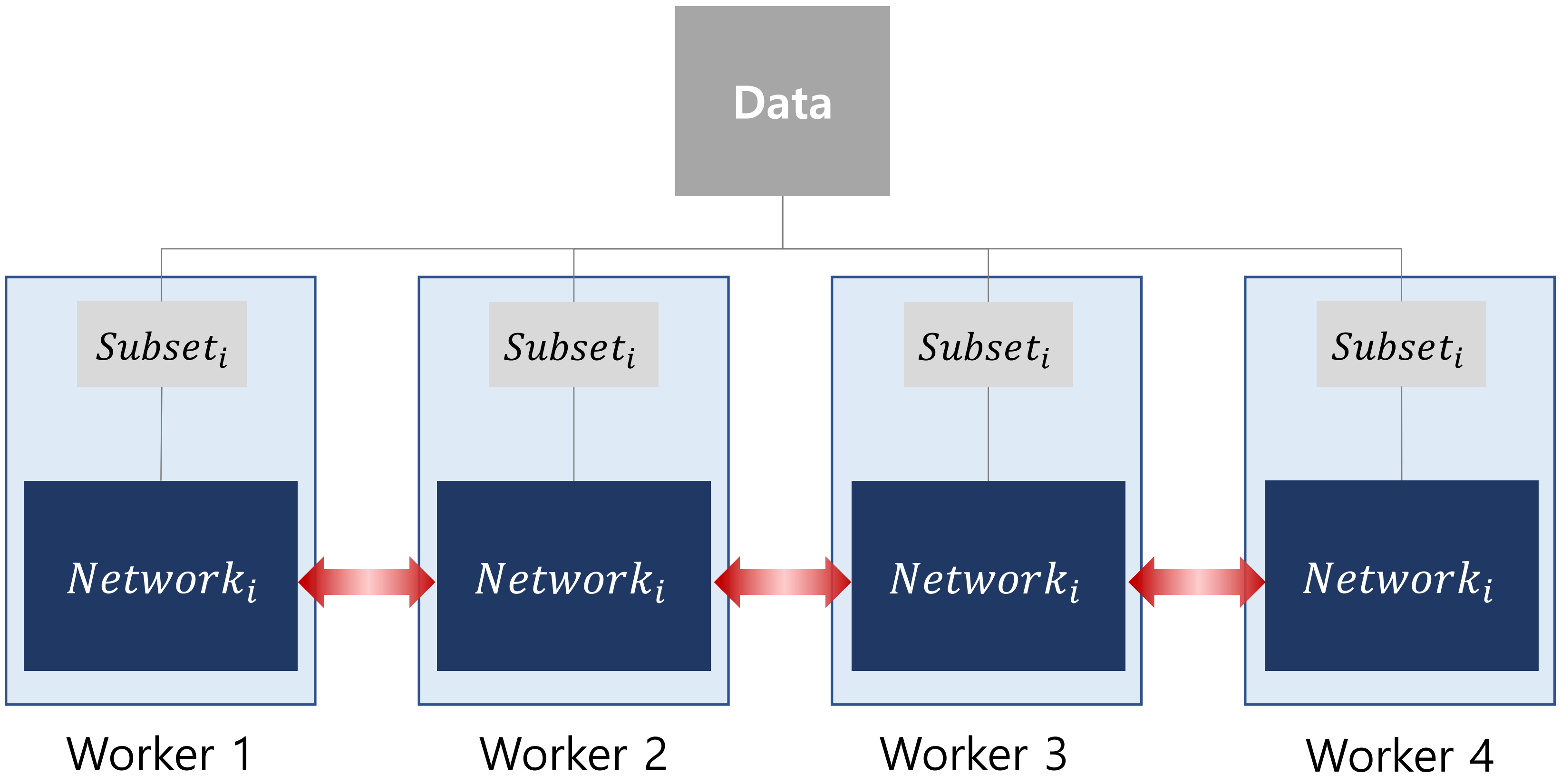
Geoffrey E. Hinton
Google Brain
geoffhinton@google.com

ABSTRACT

Techniques such as ensembling and distillation promise model quality improvements when paired with almost any base model. However, due to increased test-time cost (for ensembles) and increased complexity of the training pipeline (for

4. Distillation 관점 연구
Online – distillation

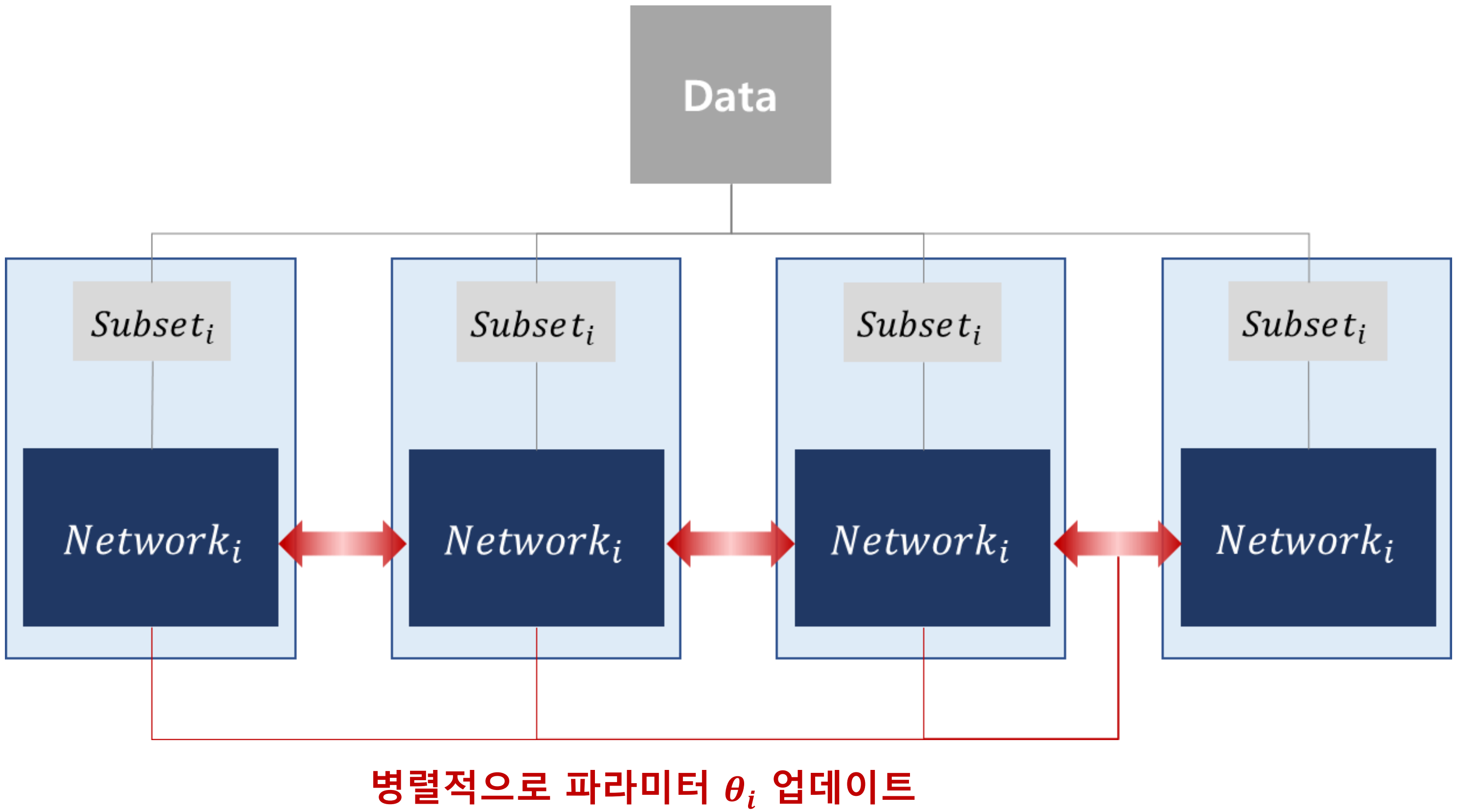
Distillation 방법 : 멀티 GPU를 통한 데이터 병렬처리와 더불어 복사된 네트워크끼리 서로 지식을 전달



4. Distillation 관점 연구

Online – distillation

Distillation 방법 : 멀티 GPU를 통한 데이터 병렬처리와 더불어 복사된 네트워크끼리 서로 지식을 전달



다른 모델들의 평균 예측값과 일치하도록 학습

Distillation 방법 : Online - distillation

for n steps do

    for θ_i in model − set do

        $y_{truth}, x = \text{get\_train\_example}()$

        $\theta_i = \theta_i - \eta \nabla_{\theta_i}\{\phi(y_{truth}, F(\theta_i, x))\}$

    end for

end for

while not converged do

    for θ_i in model − set do

        $y_{truth}, x = \text{get\_train\_example}()$

Distillation loss term

        $\theta_i = \theta_i - \eta \nabla_{\theta_i}\{\phi(y_{truth}, F(\theta_i, x)) + \psi((\underbrace{\frac{1}{(N-1)}\sum_{j \neq i} F(\theta_j, x)}_{\text{나머지 네트워크의 예측값의 평균}}), F(\theta_i, x))\}$

    end for

end while

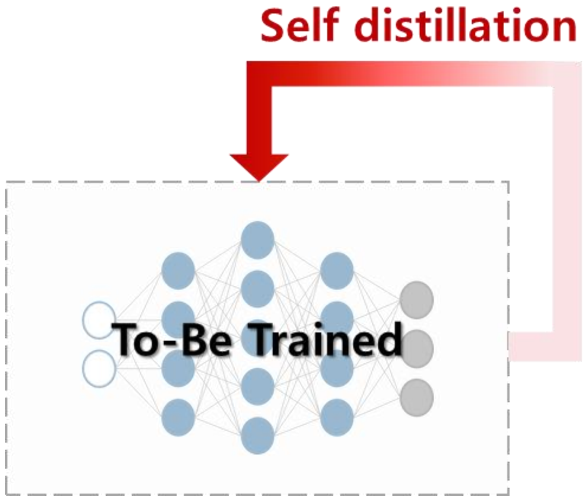
- ϕ(label, prediction) : Task에 대한 loss term
- ψ(aggregated_label, prediction): Distillation loss term
- $F(\theta_i, x)$: i번째 모델에 대한 soft target
- η : Learning rate

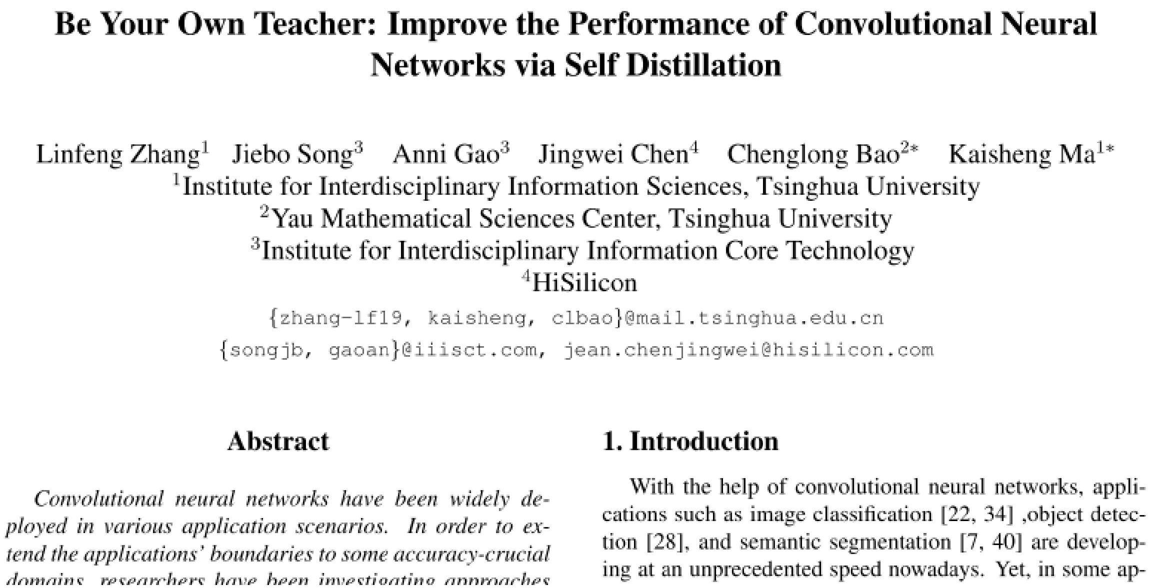
병렬적으로 학습

4. Distillation 관점 연구
Self – distillation

❖ Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation

- 2019 International Conference on Computer Vision (ICCV)에서 발표된 논문
- 2020년 12월 2일 기준 40회 인용





Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation

Linfeng Zhang[1] Jiebo Song[3] Anni Gao[3] Jingwei Chen[4] Chenglong Bao[2*] Kaisheng Ma[1*]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Yau Mathematical Sciences Center, Tsinghua University
[3]Institute for Interdisciplinary Information Core Technology
[4]HiSilicon
{zhang-lf19, kaisheng, clbao}@mail.tsinghua.edu.cn
{songjb, gaoan}@iiisct.com, jean.chenjingwei@hisilicon.com

Abstract

Convolutional neural networks have been widely deployed in various application scenarios. In order to extend the applications' boundaries to some accuracy-crucial

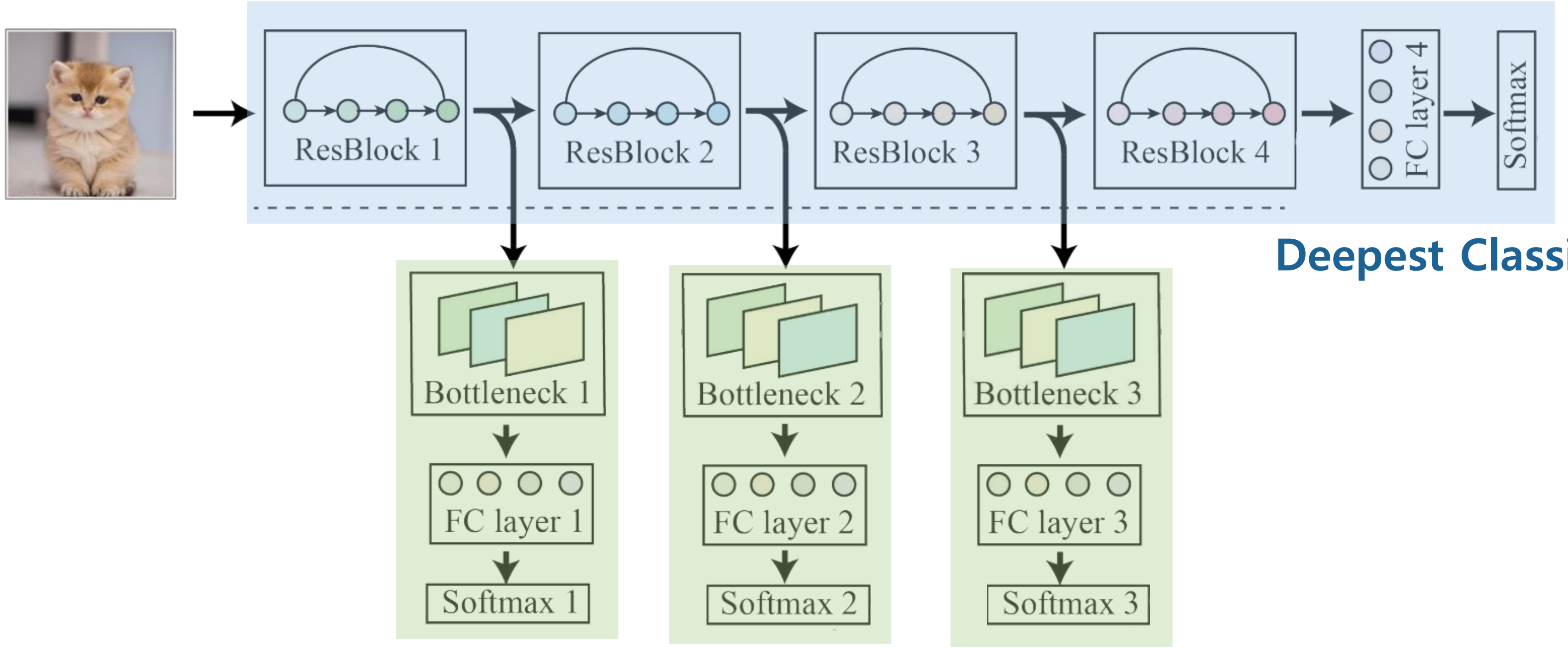
1. Introduction

With the help of convolutional neural networks, applications such as image classification [22, 34] ,object detection [28], and semantic segmentation [7, 40] are developing at an unprecedented speed nowadays. Yet, in some ap-

4. Distillation 관점 연구

Self – distillation

Distillation 방법 : 하나의 네트워크 안에서 지식이 전달되면서 학습



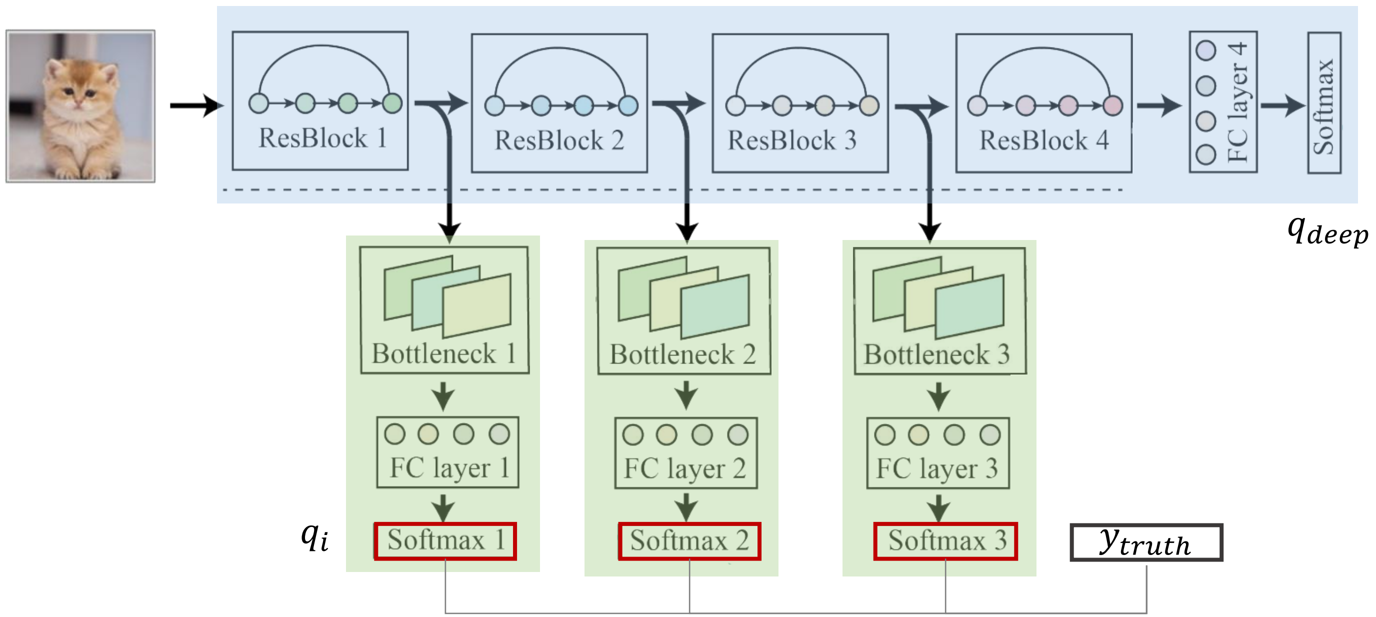
Deepest Classifier: Teacher model 역할

Shallow Classifier: Student model 역할

4. Distillation 관점 연구

Self – distillation

Distillation 방법 : 하나의 네트워크 안에서 지식이 전달되면서 학습



- q_{deep}: Deep classifier의 soft target
- q_i : 각 Shallow classifier의 soft target
- F_{deep}: Deep classifier의 마지막 feature map
- F_i: 각 Shallow classifier의 feature map

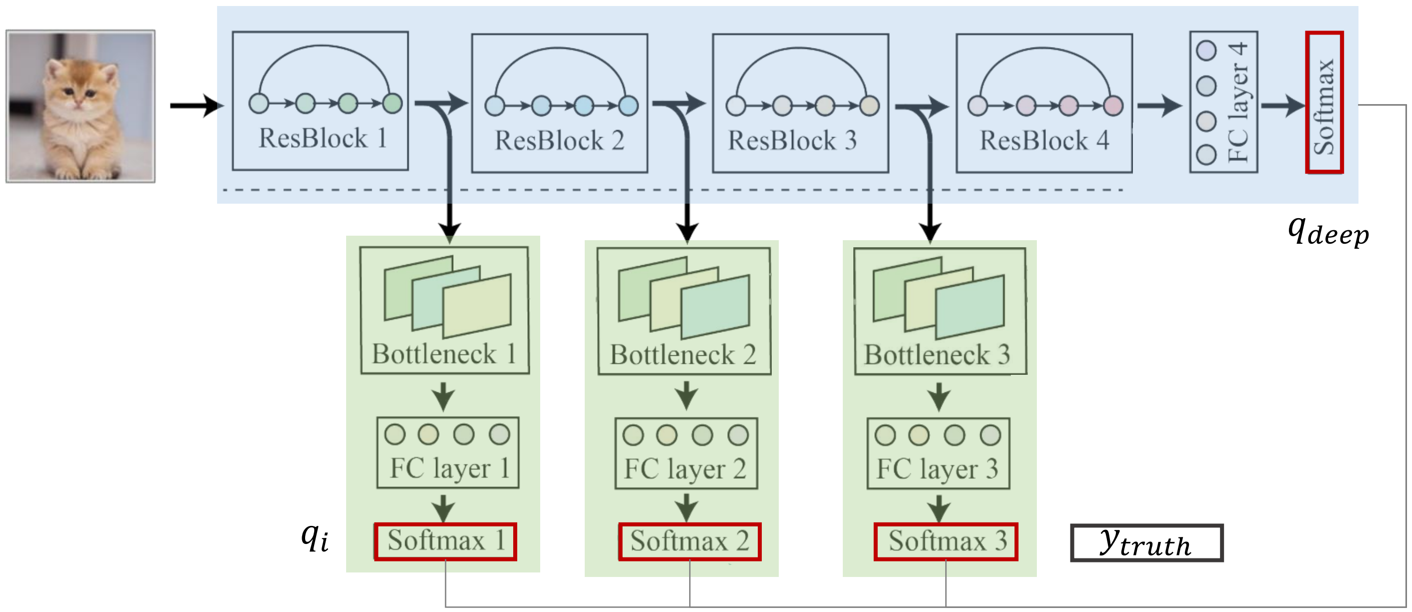
$$L_{Task} = CrossEntropy\,(softmax(q_i), y_{truth})$$

$$L_{total} = (1-\alpha)L_{task} + \alpha \cdot L_{soft} + \lambda \cdot L_{feature}$$

4. Distillation 관점 연구

Self – distillation

Distillation 방법 : 하나의 네트워크 안에서 지식이 전달되면서 학습



- q_{deep}: Deep classifier의 soft target
- q_i : 각 Shallow classifier의 soft target
- F_{deep}: Deep classifier의 마지막 feature map
- F_i: 각 Shallow classifier의 feature map

$$L_{Task} = CrossEntropy\ (softmax(q_i), y_{truth})$$

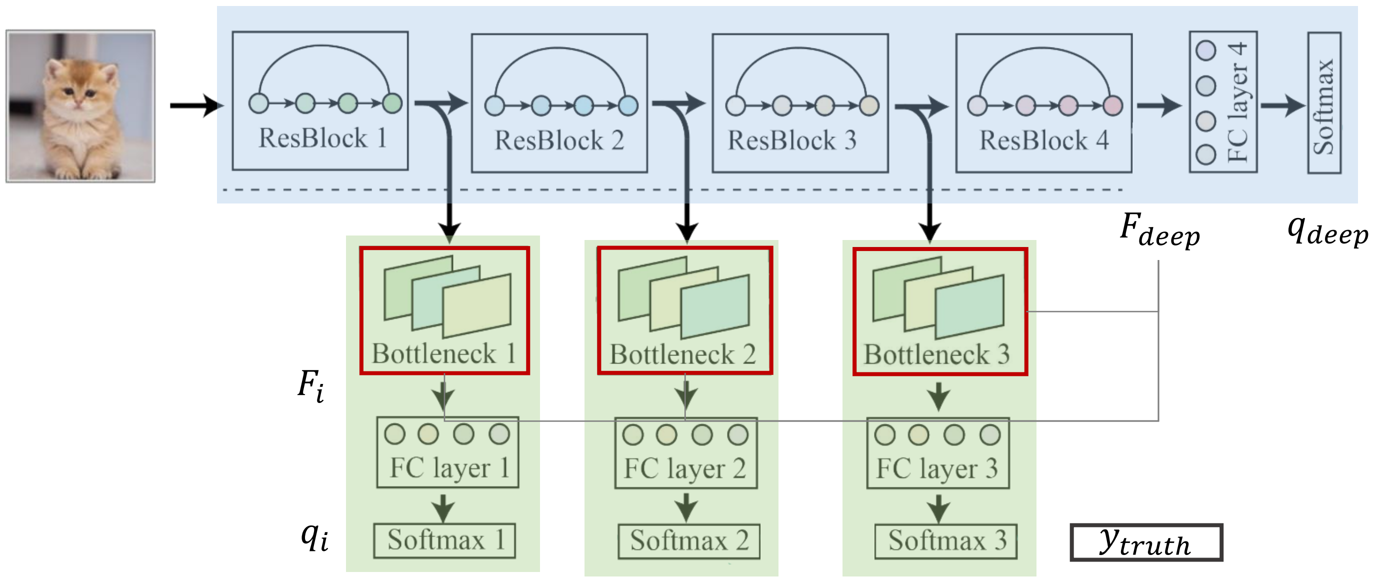
$$L_{soft} = KL(q_i, q_{deep})$$

$$L_{total} = (1-\alpha)L_{task} + \alpha \cdot L_{soft} + \lambda \cdot L_{feature}$$

4. Distillation 관점 연구

Self – distillation

Distillation 방법 : 하나의 네트워크 안에서 지식이 전달되면서 학습



- q_{deep}: Deep classifier의 soft target
- q_i : 각 Shallow classifier의 soft target
- F_{deep}: Deep classifier의 마지막 feature map
- F_i: 각 Shallow classifier의 feature map

$$L_{Task} = CrossEntropy\,(softmax(q_i), y_{truth})$$

$$L_{soft} = KL(q_i, q_{deep})$$

$$L_{feature} = \left|\left|F_i - F_{deep}\right|\right|_2^2$$

$$L_{total} = (1-\alpha)L_{task} + \alpha \cdot L_{soft} + \lambda \cdot L_{feature}$$

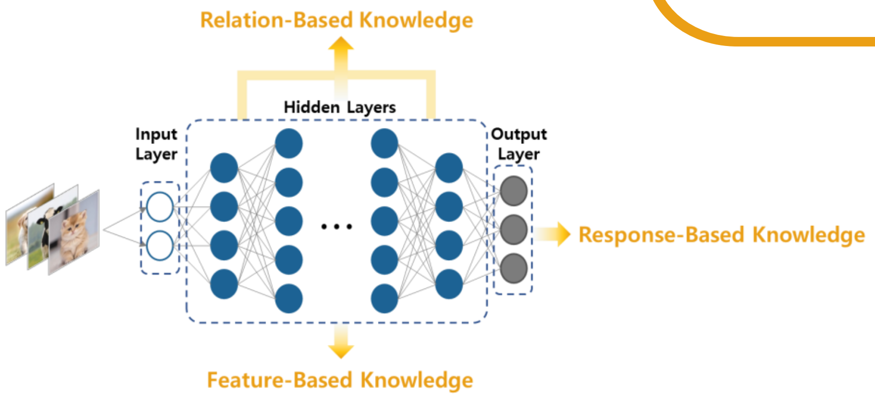
6. 결론

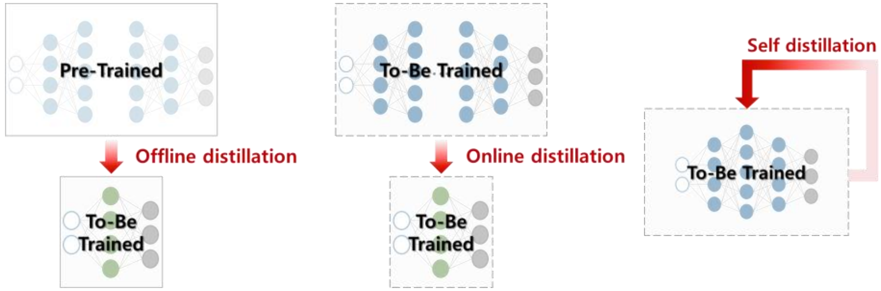
다양한 Knowledge Distillation 알고리즘

Teacher모델의 **어떠한 지식**을

Student 모델에 **어떻게 전달**할 것인가

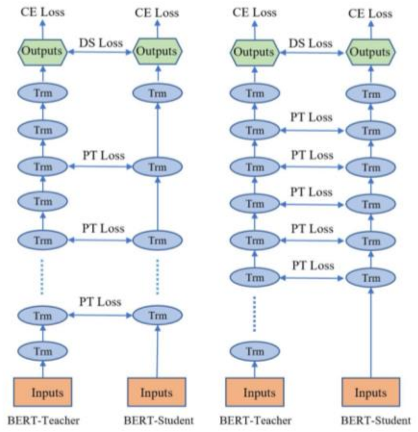
Knowledge Distillation



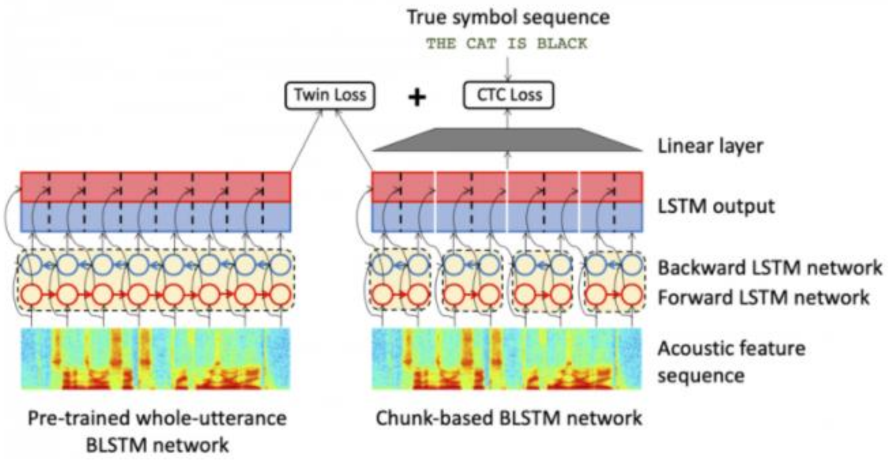


6. 결론

- 더 빠르고 가벼운 딥러닝 모델을 가능하게 하는 Knowledge Distillation 분야의 연구가 활발하게 진행 중
- 기존 딥러닝 알고리즘에서 아이디어를 차용하여 변형되는 추세
- 기존 연구들은 대부분 이미지 데이터 기준으로 진행, 최근 음성인식 분야 및 NLP 분야의 모델에도 적용됨
- 시계열 데이터에 적합한 Knowledge distillation 기법 연구 계획



BERT 모델에 적용된 사례



음성인식 분야에 적용된 사례

6. 참고 문헌

- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network.
- Heo, B., Lee, M., Yun, S., & Choi, J. Y. (2019). Knowledge transfer via distillation of activation boundaries formed by hidden neurons.
- Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation.
- Anil, R., Pereyra, G., Passos, A., Ormandi, R., Dahl, G. E., & Hinton, G. E. (2018). Large scale distributed neural network training through online distillation.
- Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation.
- GOU, Jianping, et al. Knowledge Distillation: A Survey.
- https://blog.lunit.io/2018/03/22/distilling-the-knowledge-in-a-neural-network-nips-2014-workshop/
- https://m.blog.naver.com/PostView.nhn?blogId=hist0134&logNo=221525718843&proxyReferer=https:%2F%2Fwww.google.com%2F

Thank You