Distilling the Knowledge in a Neural Network

Geoffrey Hinton, Oriol Vinyals, Jeff Dean

Paper review

KAERI-UST

Jungmin Kim

Distilling the knowledge in a Neural Network

- NIPS 2014 Deep Learning Workshop
- 제프리 힌튼, 오리올 비니알스, 제프 딘
- 2021년 04월 08일 기준 5830회 인용

Distilling the Knowledge in a Neural Network

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

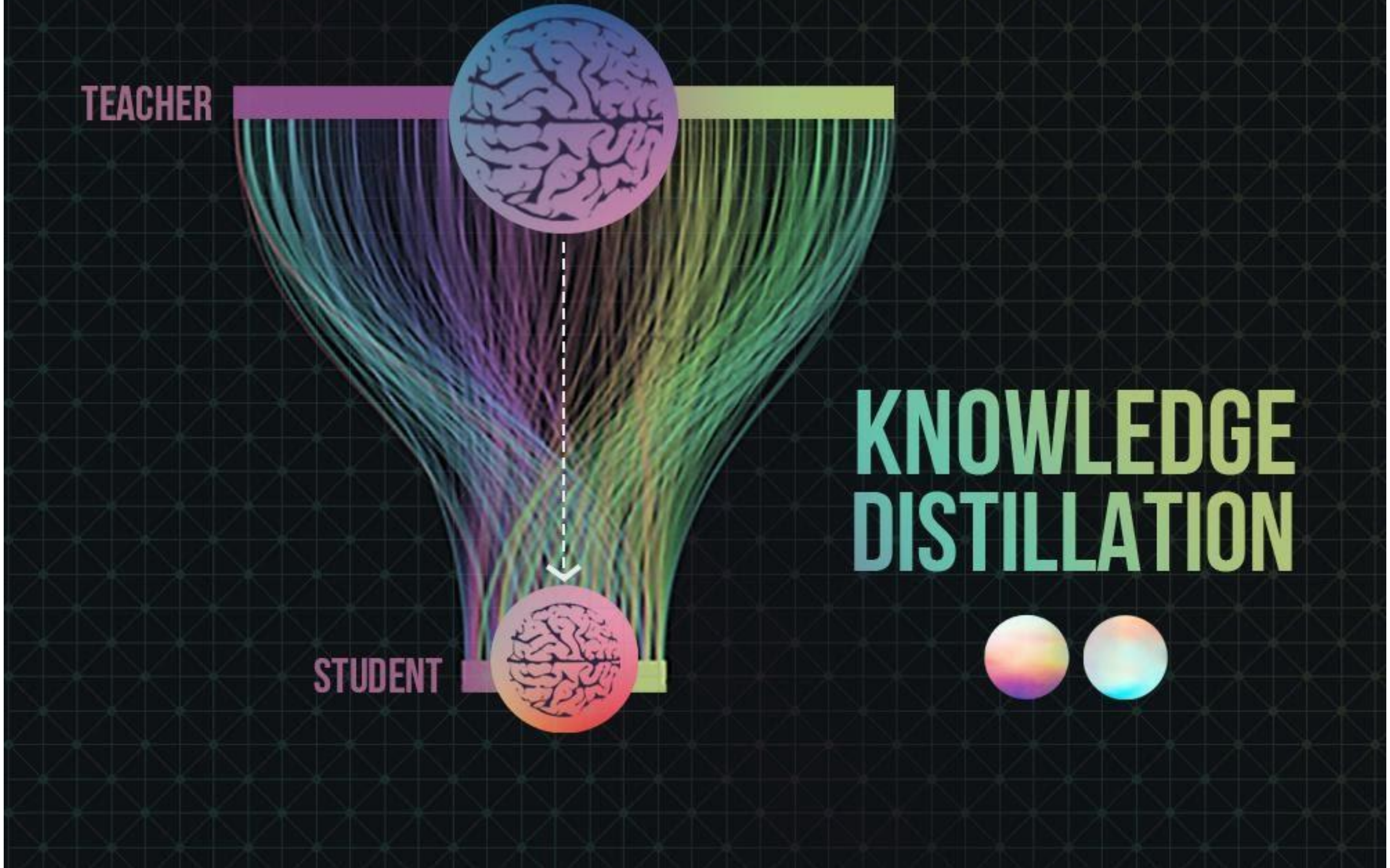
Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

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

Abstract

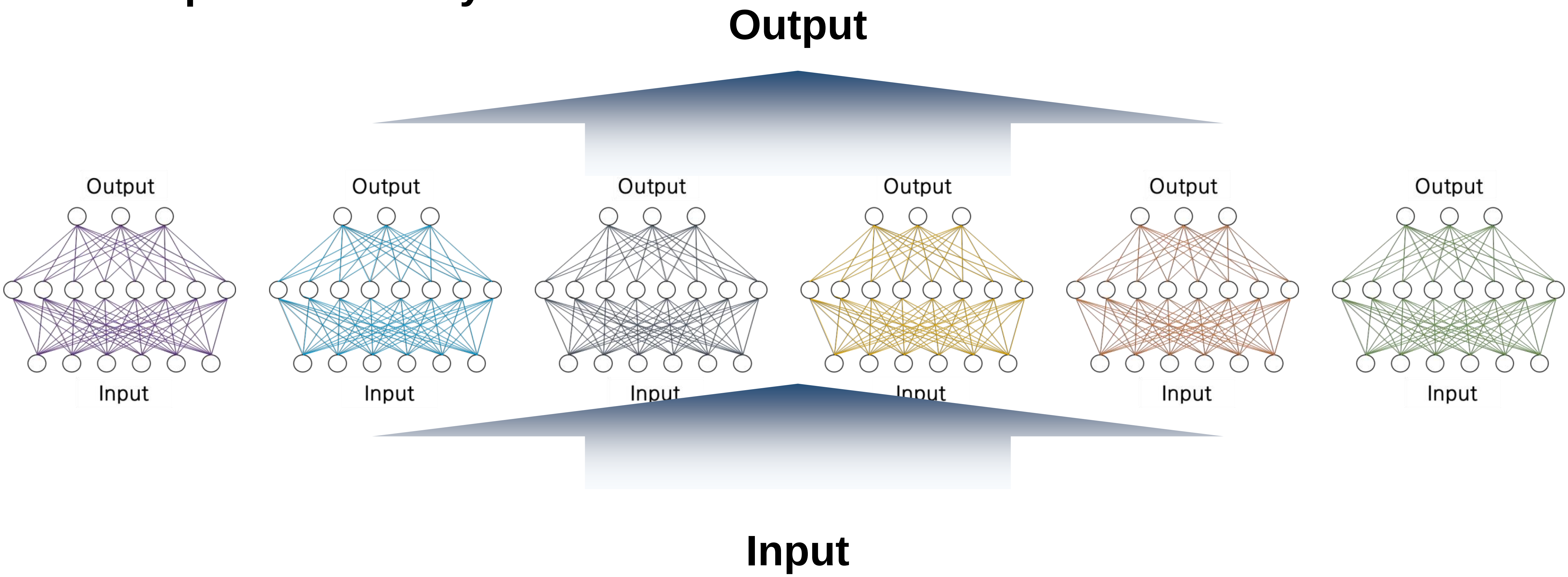
A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions [3]. Unfortunately, making predictions using a whole ensemble

작은 네트워크도 큰 네트워크와 비슷한 성능을 낼 수 있도록,
학습과정에서 큰 네트워크의 지식을 작은 네트워크에게 전달



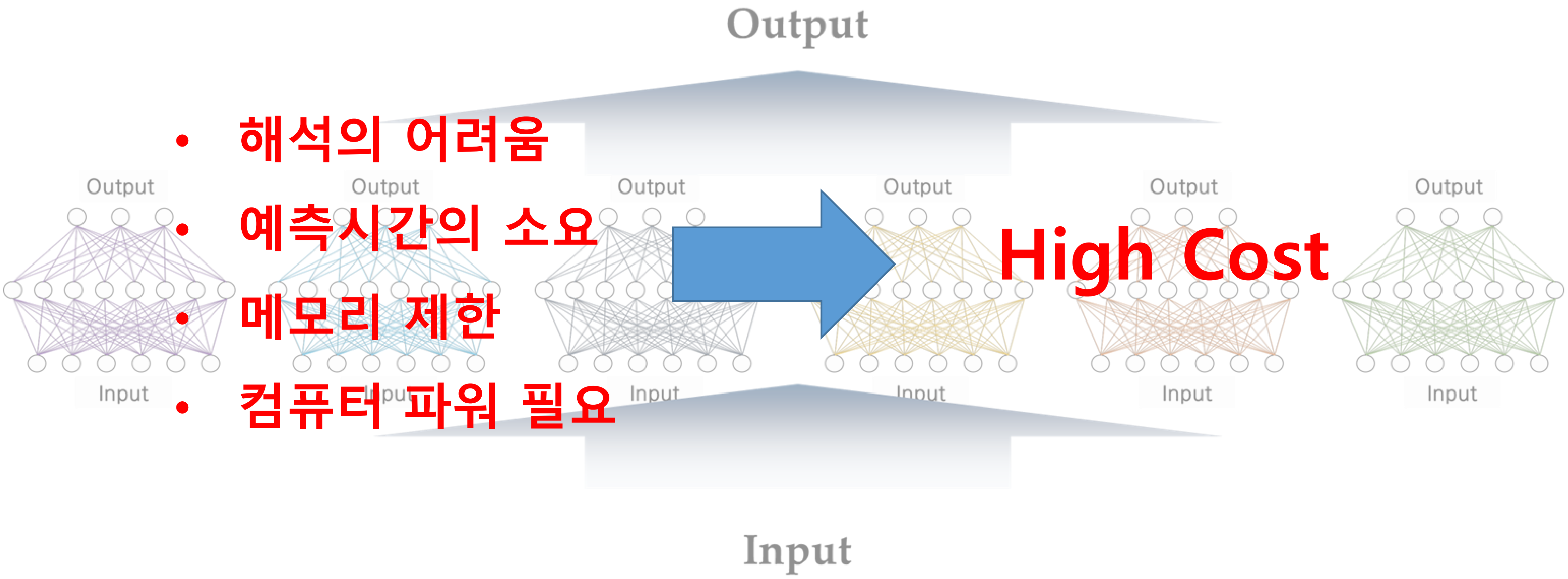
Ensemble

- To improve generalization performance
- Composed of many models



Ensemble

- ensemble of models is cumbersome and too computationally expensive
- especially if the models are large neural nets



- 해석의 어려움
- 예측시간의 소요
- 메모리 제한
- 컴퓨터 파워 필요

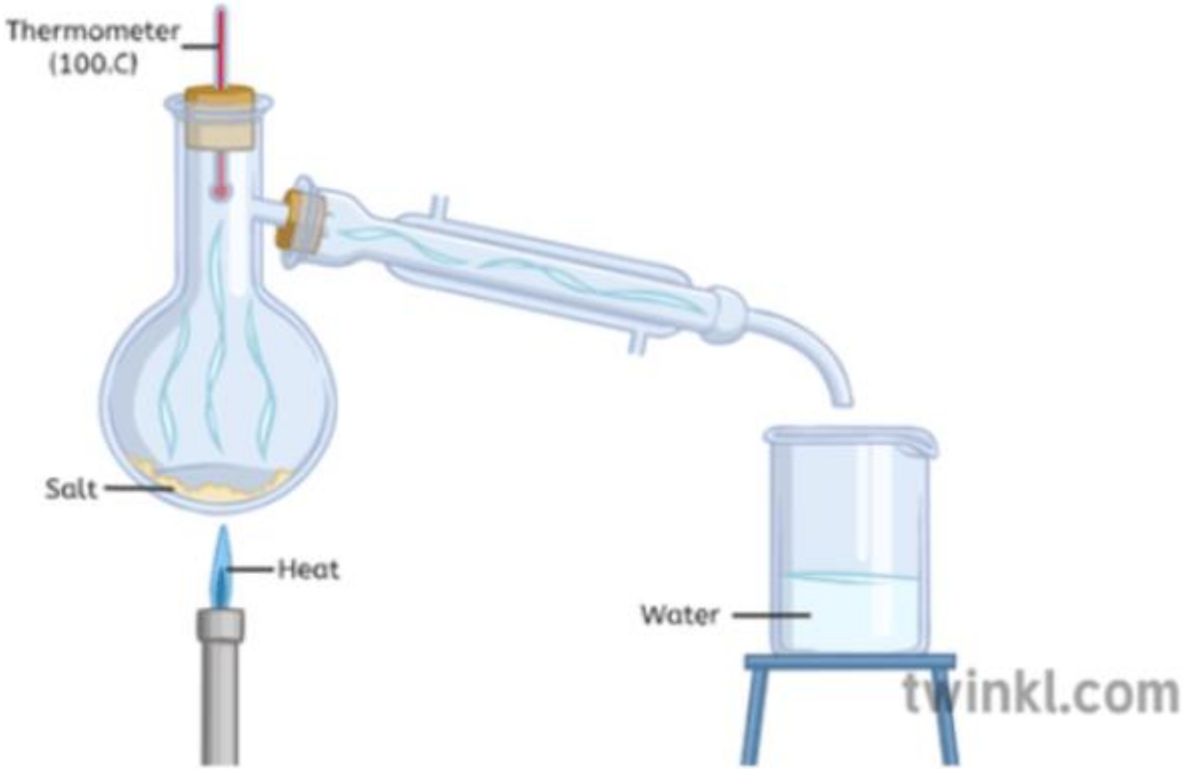
High Cost

Distillation

Distillation (증류)

불순물이 섞여 있는 혼합물에서 원하는 특정 성분을 추출

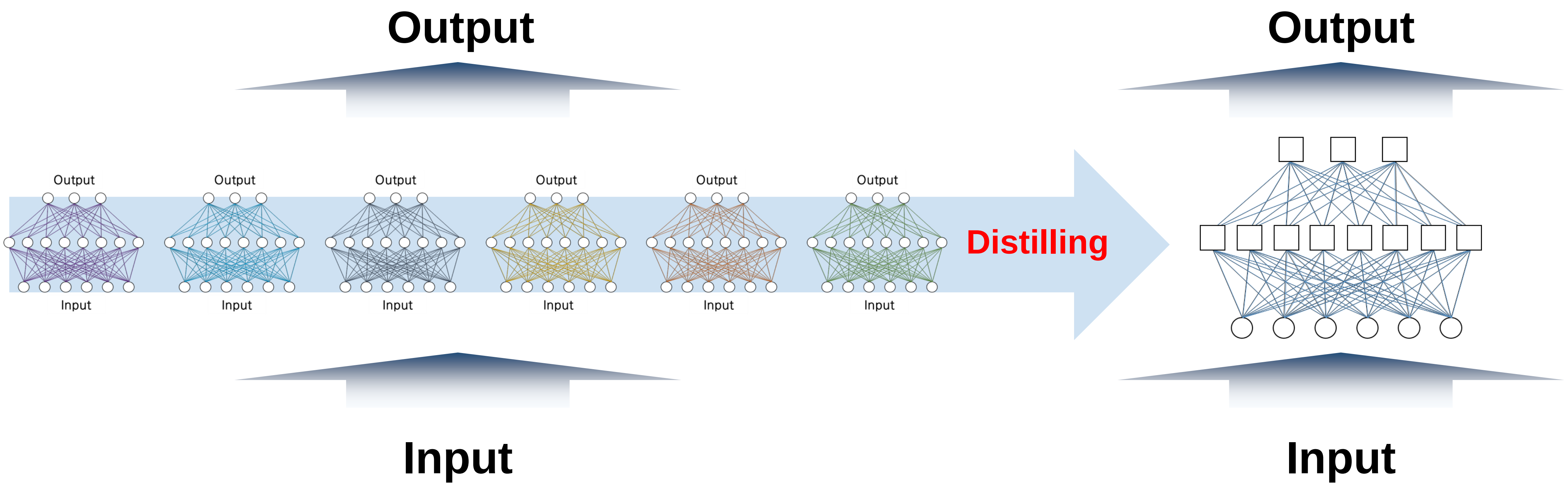
Ensemble model로부터, generalization 성능을 향상시킬 수 있는 knowledge를 추출



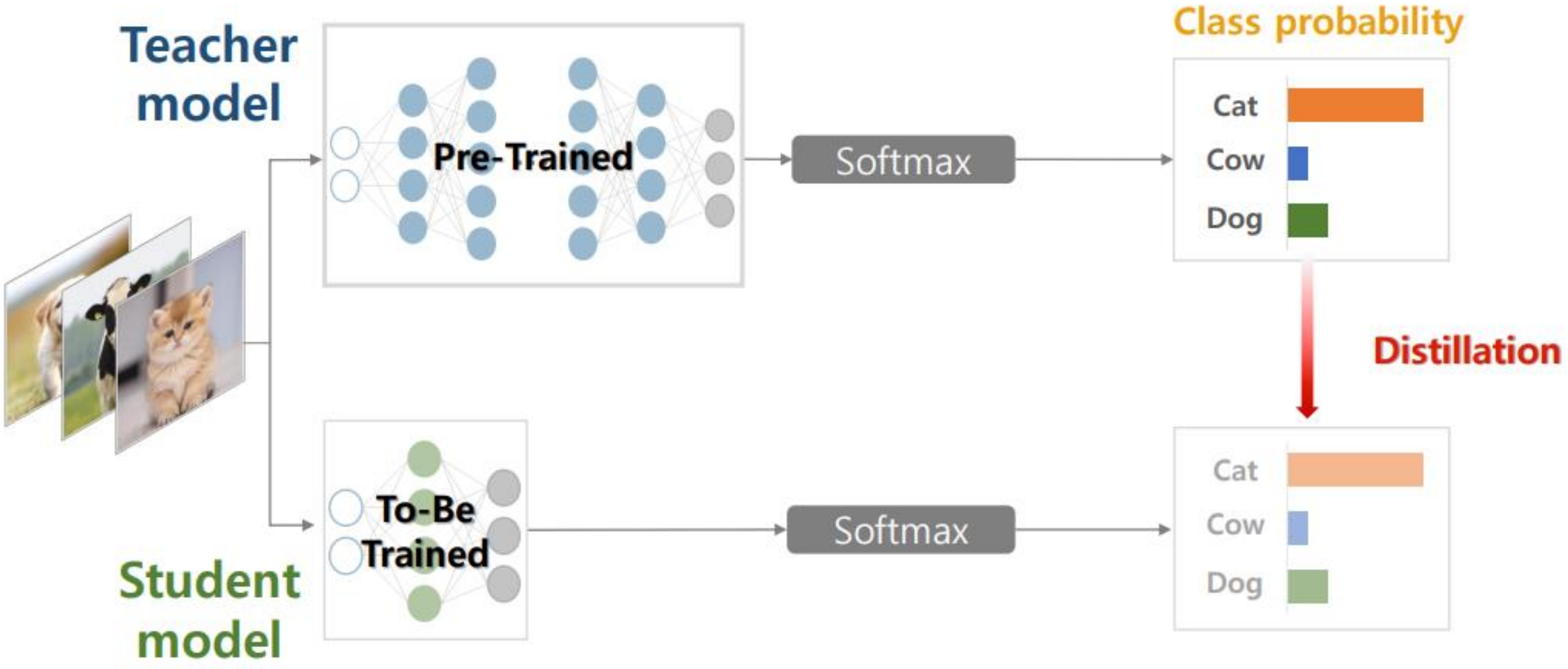
Distilling Ensemble to Single Model

큰 네트워크의 지식(일반화 능력)을 작은 네트워크에게 전달하여

작은 네트워크의 성능을 높이는 것이 목적



Distillation 프레임워크



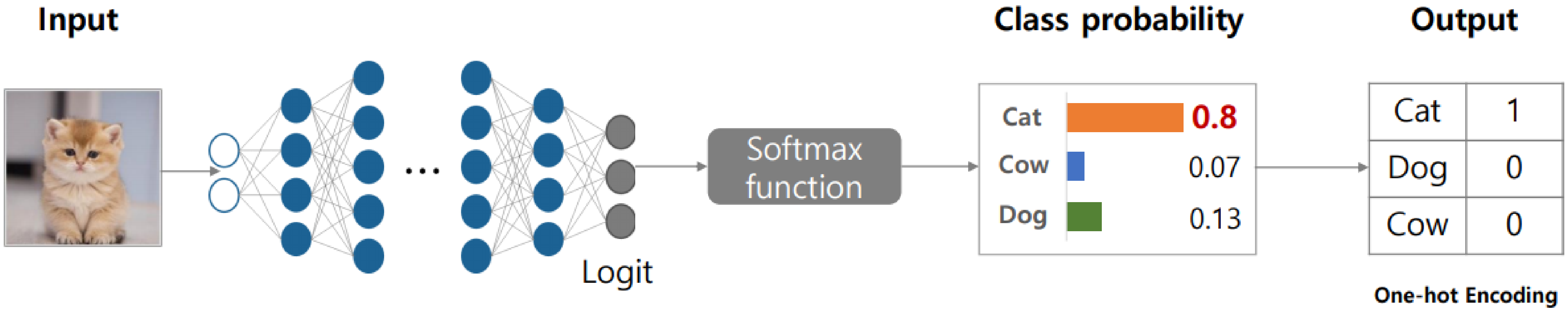
Distillation 프레임워크

1. training set (x, hard target)을 사용해 large model을 학습한다.

2. large model이 충분히 학습된 뒤에, large model의 output을 soft target으로 하는 transfer set(x, soft target)을 생성해낸다.

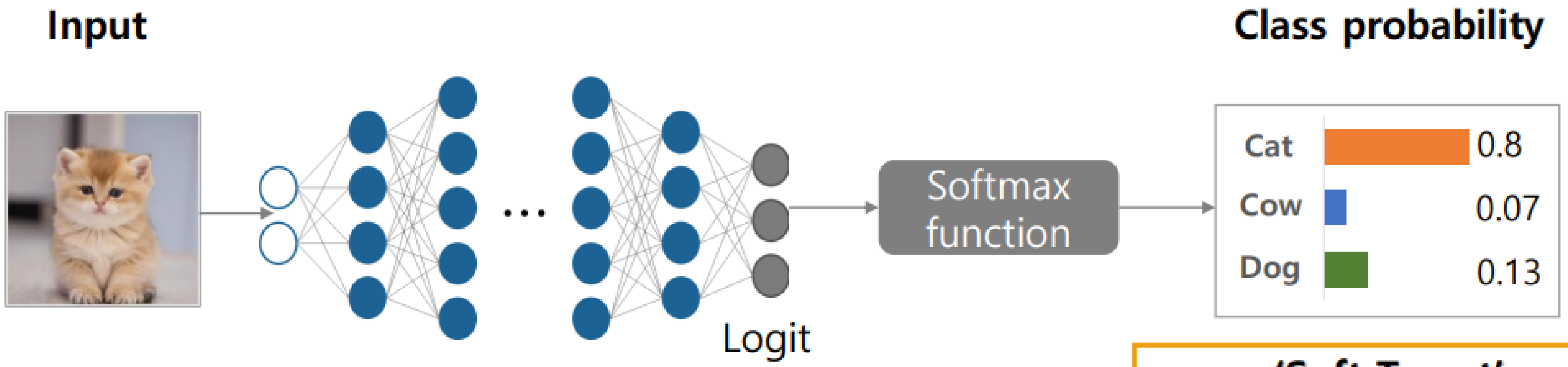
3. transfer set을 사용해 small model을 학습한다.

Target 종류



One-hot Encoding

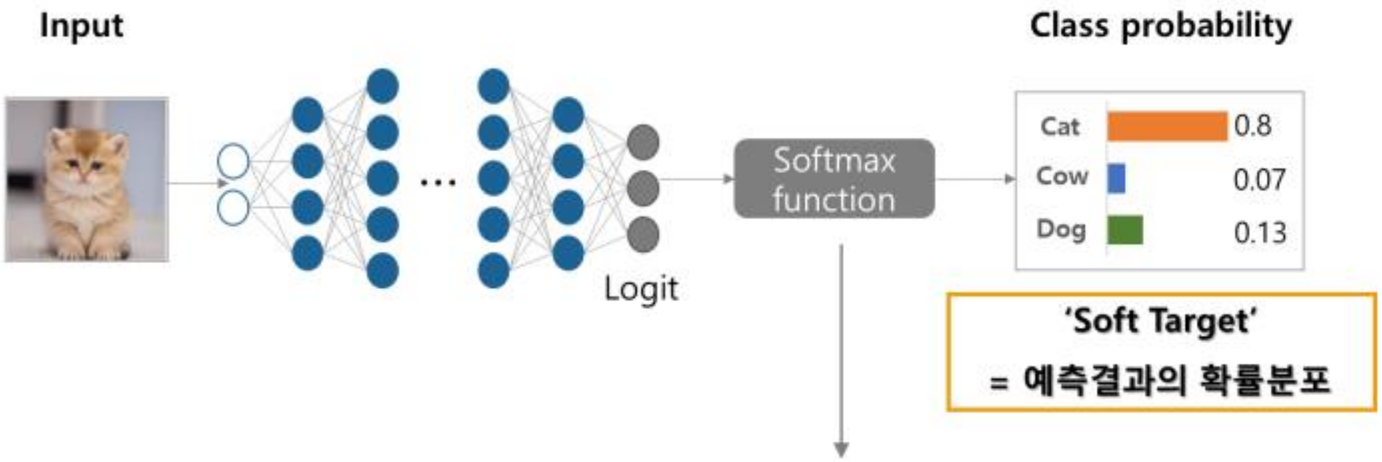
'Hard Target'



'Soft Target'

= 예측결과의 확률분포

Softened output of Softmax



Softmax ouput

- 특정 범주가 0에 매우 가까움
- 지식전달에 어려움

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

τ $(Temperature)$: Scaling 역할의 하이퍼 파라미터

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

$$Softmax\begin{pmatrix}1\\2\\9\end{pmatrix} = \begin{pmatrix}0.000335\\0.000911\\0.998754\end{pmatrix}, \quad \underset{T=3}{Softmax}\begin{pmatrix}1\\2\\3\end{pmatrix} = \begin{pmatrix}0.059\\0.083\\0.857\end{pmatrix}$$

Softened output of Softmax

dog

cow	dog	cat	car	
0	1	0	0	original hard targets
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble
.05	.3	.2	.005	softened output of ensemble

MNIST-1

Teacher (large model) → 67 test errors

Student (small model) →146 test errors

Distilled (small model에 transfer set)

→ 74 test errors

Model	Architecture	Test errors	Temperature
Teacher (Hard targets)	2 FC layer with 1200 hidden units	67	1
Student (Hard targets)	2 FC layer with 800 hidden units	146	1
Distilled model (Hard + soft targets)	2 FC layer with 800 hidden units	74	20

MNIST-2

- knowledge distillation을 통해 학습
- Student 학습 dataset에서 숫자 “3”이 없음
- test 결과 →109 test error
- test set에 1010개의 “3”중 14개만 틀림 (98.6% accuracy)

Model	Architecture	Test errors	Temperature
Teacher (Hard targets)	2 FC layer with 1200 hidden units	67	1
Student (Hard targets)	2 FC layer with 800 hidden units	146	1
Distilled model (Hard + soft targets)	2 FC layer with 800 hidden units	74	20
without “3” in MNIST data		109	

MNIST-2

- knowledge distillation을 통해 학습
- Student 학습 dataset에서 숫자 "3"이 없음
- test 결과 →109 test error
- test set에 1010개의 "3"중 14개만 틀림 (98.6% accuracy)

- **Softmax data를 통해 학습한 Distilled model에서 "3"을 본적은 없지만 soft label을 통해 "3"을 유추하고 test 과정에서 등장한 "3"을 구분함**

Soft Targets as Regularizers

- soft target은 regularization 효과
- hard target에는 없는 유용한 정보들이 overfitting을 방지

System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

- Hard target으로 모든 data에 대해 학습을 수행했을 때에 최종 test accuracy는 58.9%가 도출
- 3%로 학습을 진행한 결과 최종 test accuracy는 44.5%가 도출됐고 학습 도중 early stopping을 사용했음에도 overfitting이 발생
- 100%의 training set에서 soft target을 추출해내 그 중 3%만을 갖고 학습을 진행했을 때에는 test accuracy가 accuracy가 57.0% 수렴

Conclusion

- **Distilling은 앙상블 모델에서 작은 모델로 일반화 지식을 전달**
- **Softmax 함수값을 이용해 Knowledge Distillation**
- **Softmax값을 Temperature로 soften (일반적 2≤T≤4)**
- **Soft label을 통해 소실된 데이터를 유추**
- **Soft target을 사용하는 것은 overfitting을 방지 Regularizer**

Application

K Keras

Search Keras documentation...

- About Keras
- Getting started
- Developer guides
- Keras API reference
- Code examples
- Why choose Keras?
- Community & governance
- Contributing to Keras

» Code examples / Computer Vision / Knowledge Distillation

Knowledge Distillation

Author: Kenneth Borup
Date created: 2020/09/01
Last modified: 2020/09/01
Description: Implementation of classical Knowledge Distillation.

View in Colab • **GitHub source**

Introduction to Knowledge Distillation

Knowledge Distillation is a procedure for model compression, in which a small (student) model is trained to match a large pre-trained (teacher) model. Knowledge is transferred from the teacher model to the student by minimizing a loss function, aimed at matching softened teacher logits as well as ground-truth labels.

The logits are softened by applying a "temperature" scaling function in the softmax, effectively smoothing out the probability distribution and revealing inter-class relationships learned by the teacher.

Reference:

- Hinton et al. (2015)

Reference

- **논문:** * Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531
- Distilling the Knowledge in a Neural Network 발표: https://github.com/YoungJaeChoung/Presentation
- **고려대학교 DMQA 세미나, Introduction to knowledge distillation: http://dmqm.korea.ac.kr/activity/seminar/304**
- **논문리뷰 블로그: https://blog.lunit.io/2018/03/22/distilling-the-knowledge-in-a-neural-network-nips-2014-workshop/**
- **논문리뷰 블로그: https://medium.com/@arvindwaskarthik/knowledge-distillation-in-a-neural-network-6f469066be7e**
- **논문리뷰 블로그: https://towardsdatascience.com/distilling-knowledge-in-neural-network-d8991faa2cdc**
- **논문리뷰 블로그: https://cpm0722.github.io/paper-review/distilling-the-knowledge-in-a-neural-network**
- **Keras: https://keras.io/examples/vision/knowledge_distillation/**

Thank You