Distilling the Knowledge in a Neural Network

Geoffrey Hinton, Oriol Vinyals, Jeff Dean

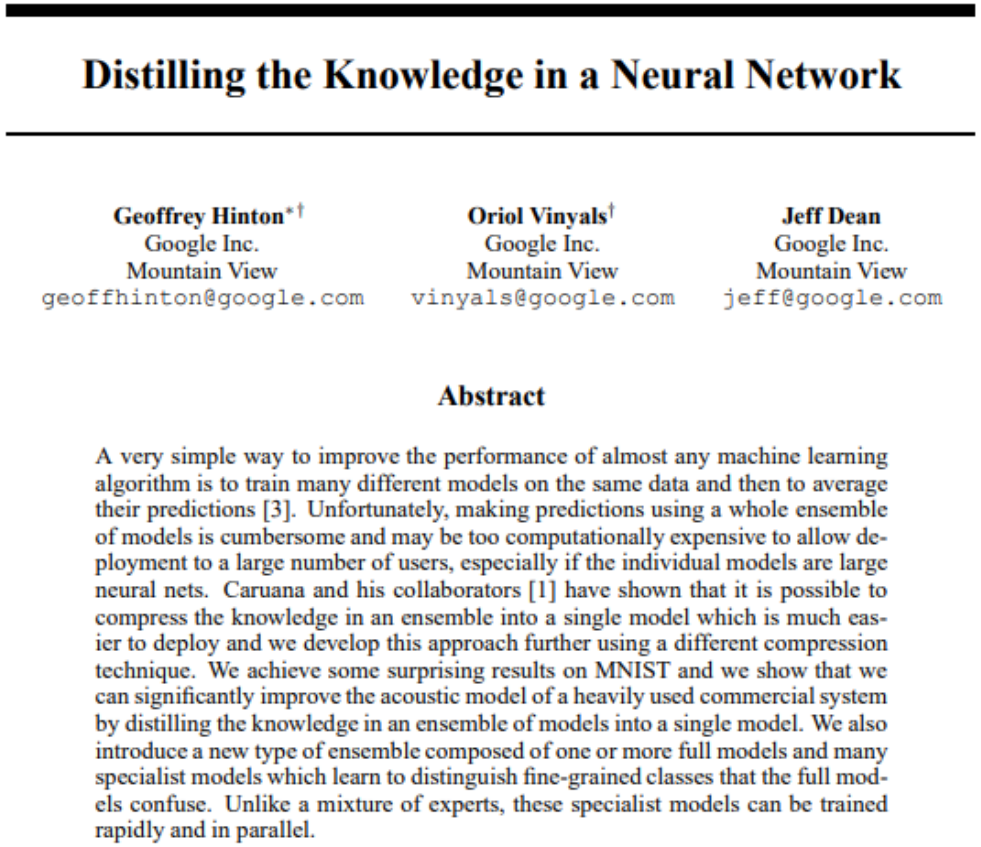
Paper review

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

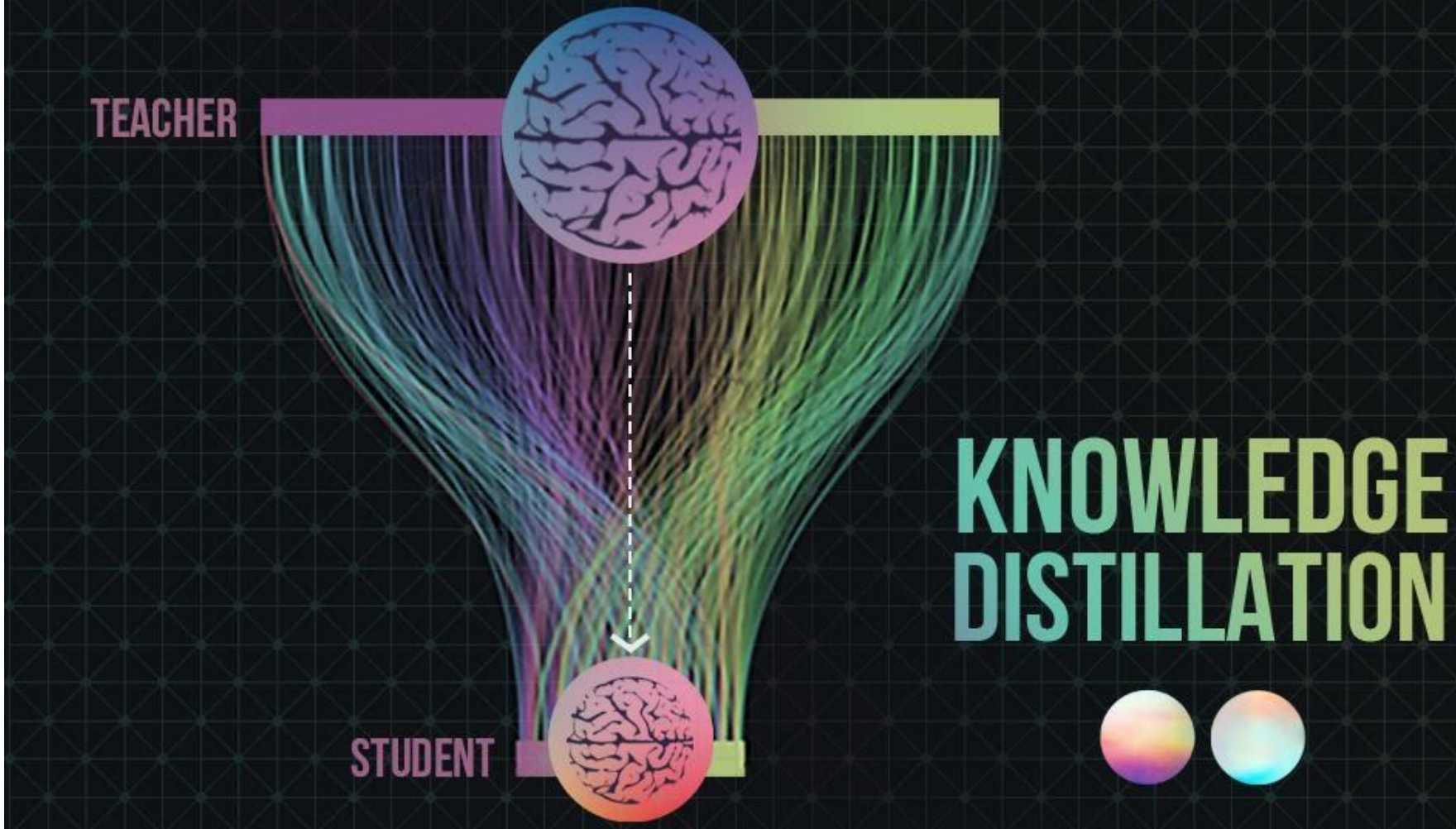
Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

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

Abstract

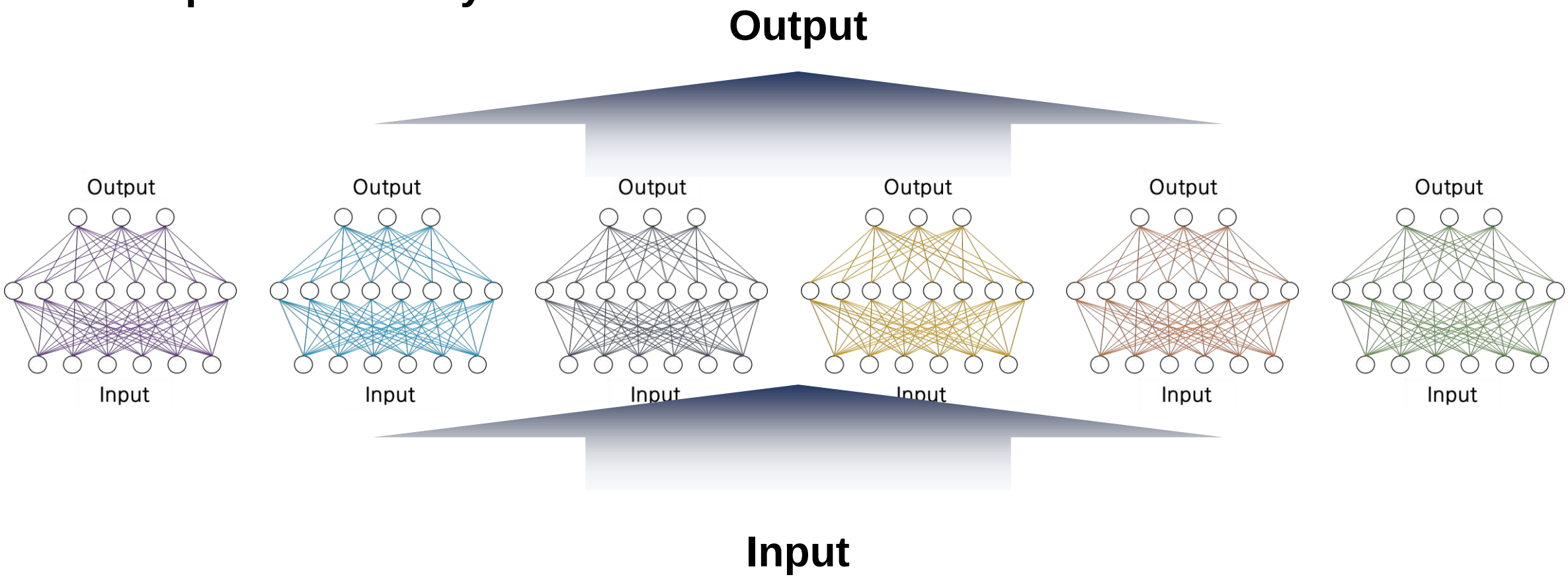
A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions [3]. Unfortunately, making predictions using a whole ensemble of models is cumbersome and may be too computationally expensive 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 compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique. We achieve some surprising results on MNIST and we show that we can significantly improve the acoustic model of a heavily used commercial system by distilling the knowledge in an ensemble of models into a single model. We also introduce a new type of ensemble composed of one or more full models and many specialist models which learn to distinguish fine-grained classes that the full models confuse. Unlike a mixture of experts, these specialist models can be trained rapidly and in parallel.

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



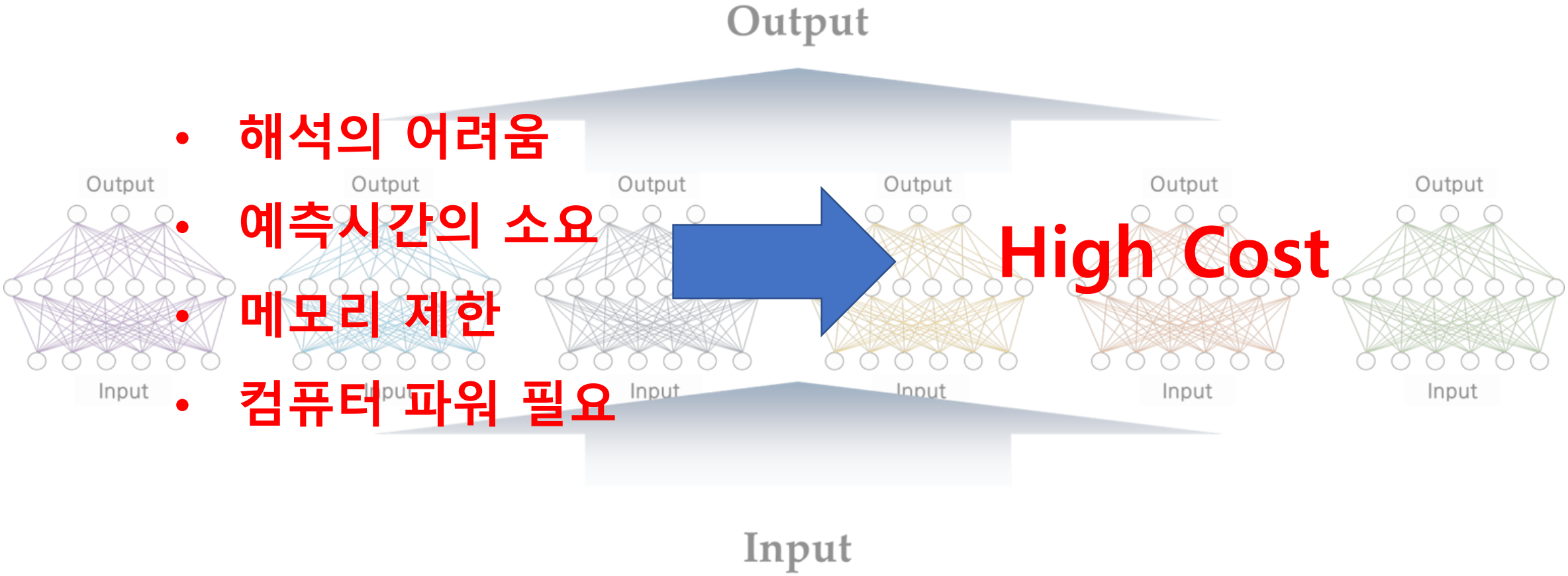
Ensemble

- 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

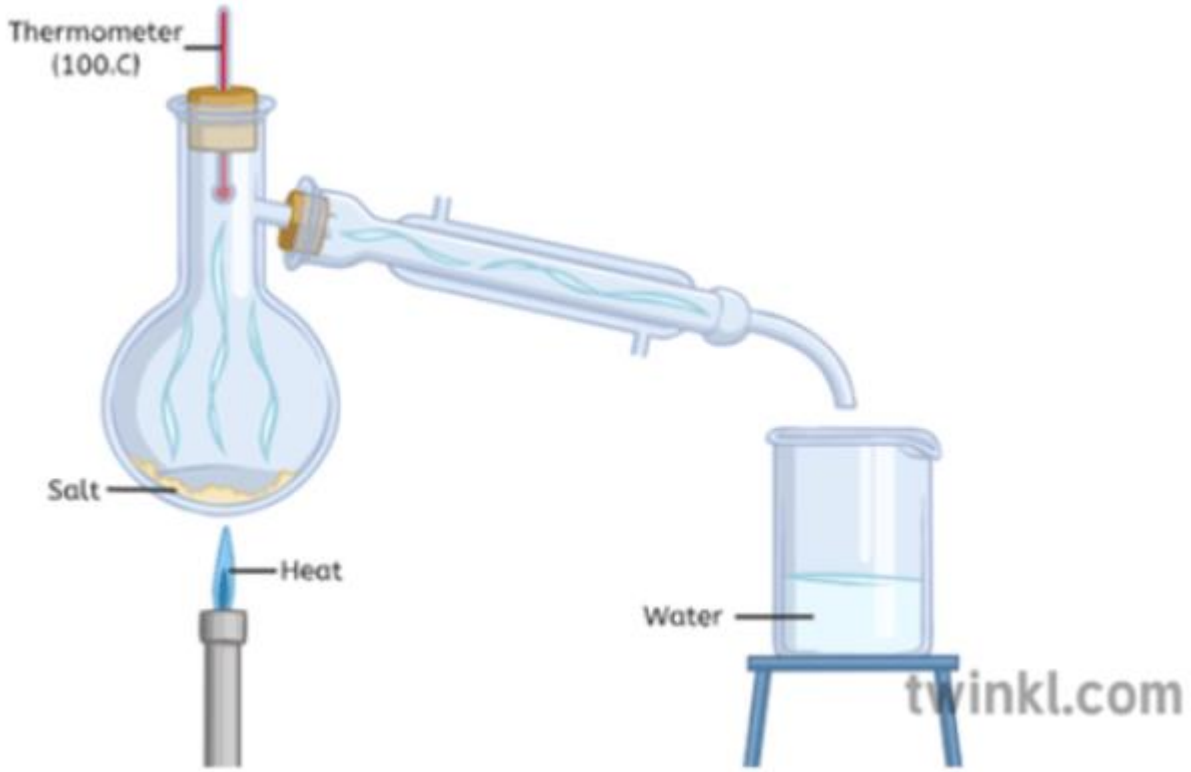
High Cost

Distillation

Distillation (증류)

불순물이 섞여 있는 혼합물에서 원하는 특정 성분을 추출

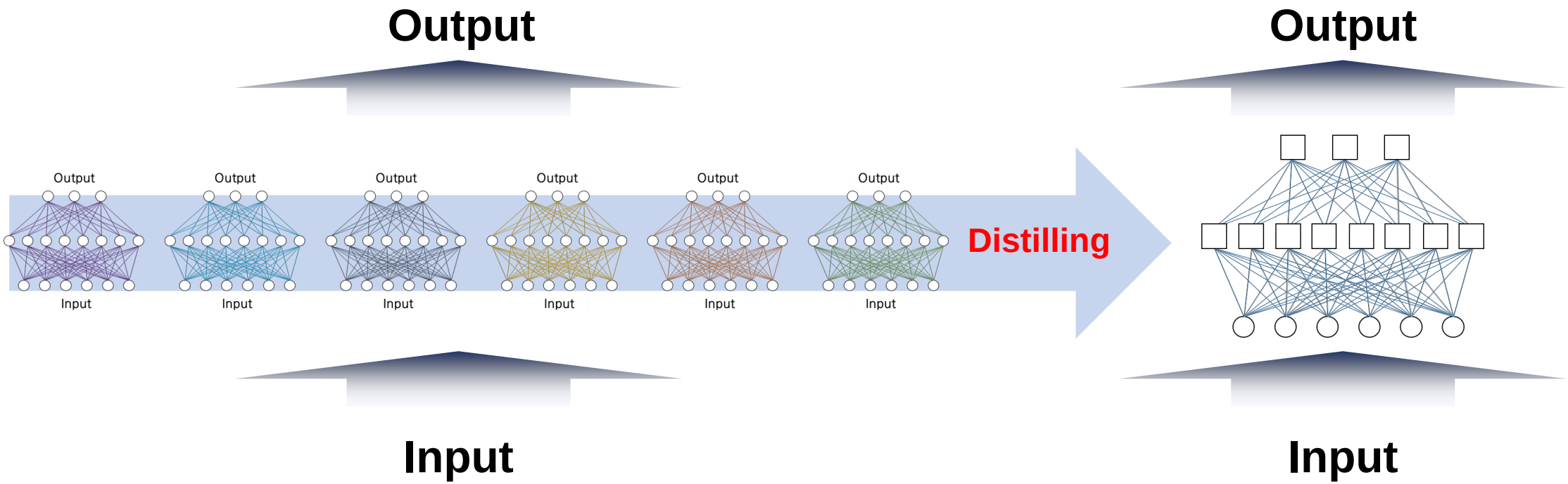
Ensemble model로부터, generalization 성능을 향상시킬 수 있는 knowledge를 추출



Distilling Ensemble to Single Model

큰 네트워크의 지식(일반화 능력)을 작은 네트워크에게 전달하여

작은 네트워크의 성능을 높이는 것이 목적



Knowledge Distillation

큰 모델(Teacher Network)로부터 증류한 지식을 작은 모델(Student Network)로 transfer하는 일련의 과정

Abstract

Deploy models

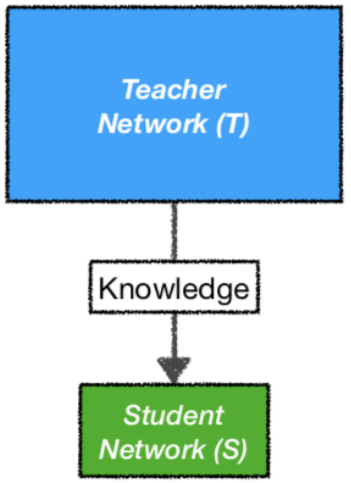
Q) But Could we use Ensemble when we deploy models?

A) Unfortunately, cumbersome & too computationally expensive

- to a large number of users
- the individual models are large NNs

Introduction

Knowledge Distillation



1. **Teacher Network (T)**
   - **cumbersome model**
     ex) ensemble / a large generalized model
   - (pros) excellent performance
   - (cons) computationally expansive
   - can not be deployed when limited environments

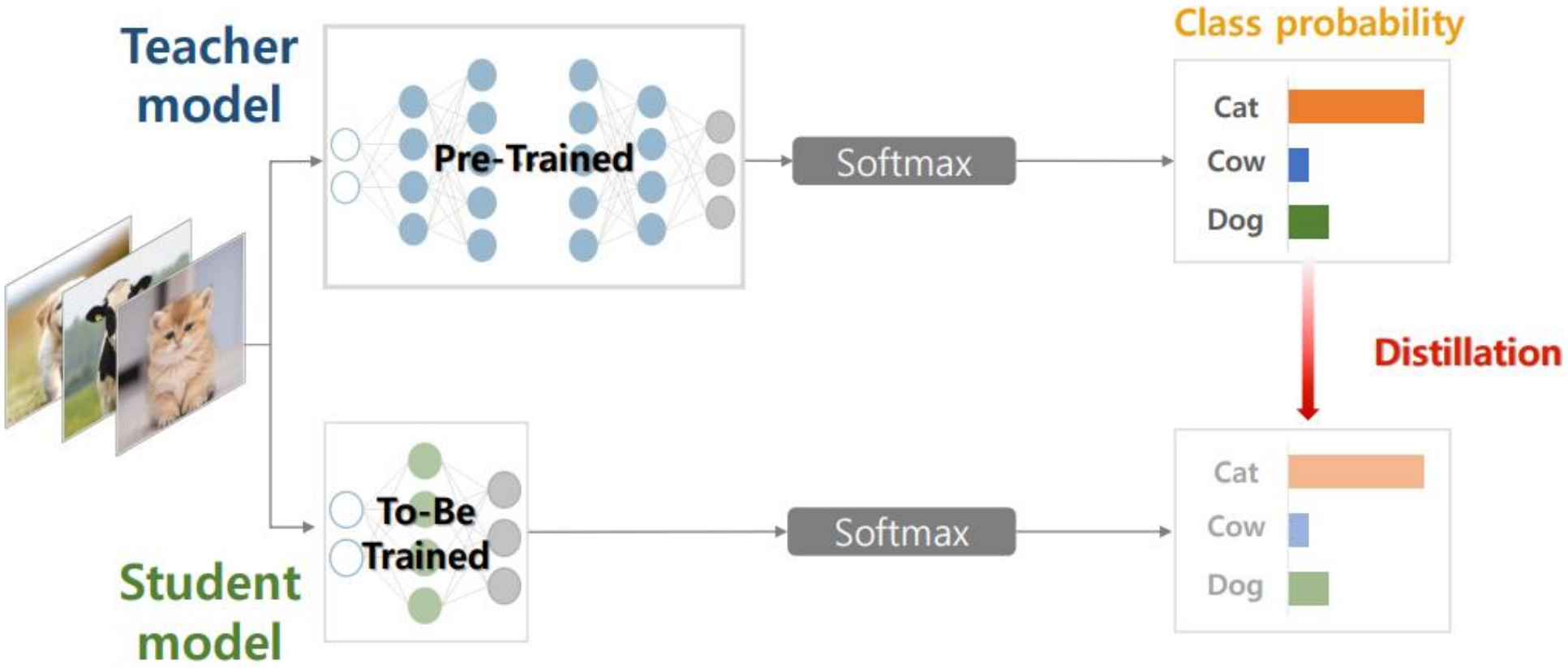
2. **Student Network (S)**
   - **small model**
   - suitable for deployment
   - (pros) fast inference
   - (cons) lower performance than T

모델 배포(model deployment) 에서 지식 증류의 필요성

- 복잡한 모델 T : 예측 정확도 99% + 예측 소요 시간 3시간
- 단순한 모델 S : 예측 정확도 90% + 예측 소요 시간 3분

출처: https://baeseongsu.github.io/posts/knowledge-distillation/

Distillation 프레임워크



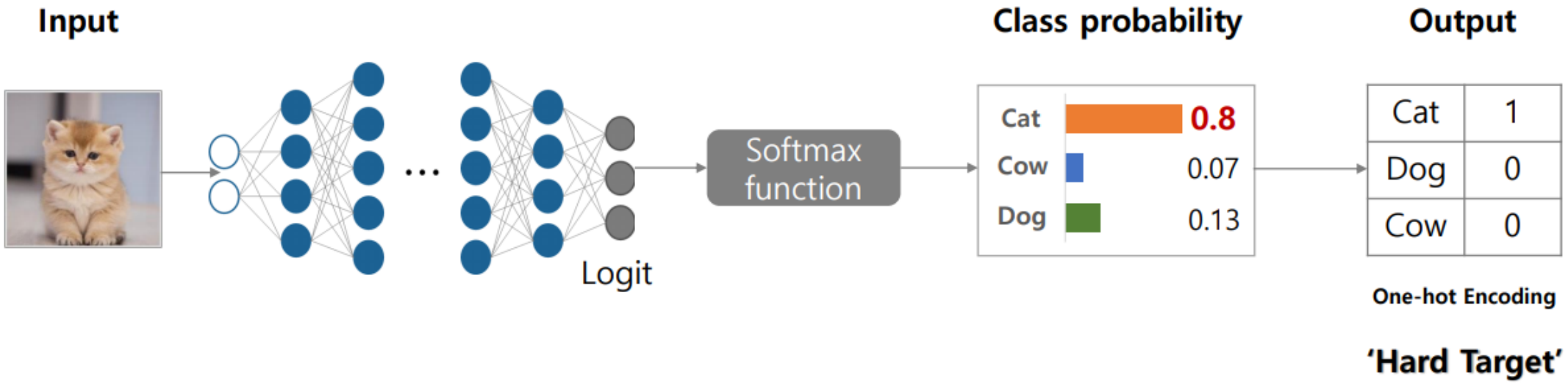
Distillation 프레임워크

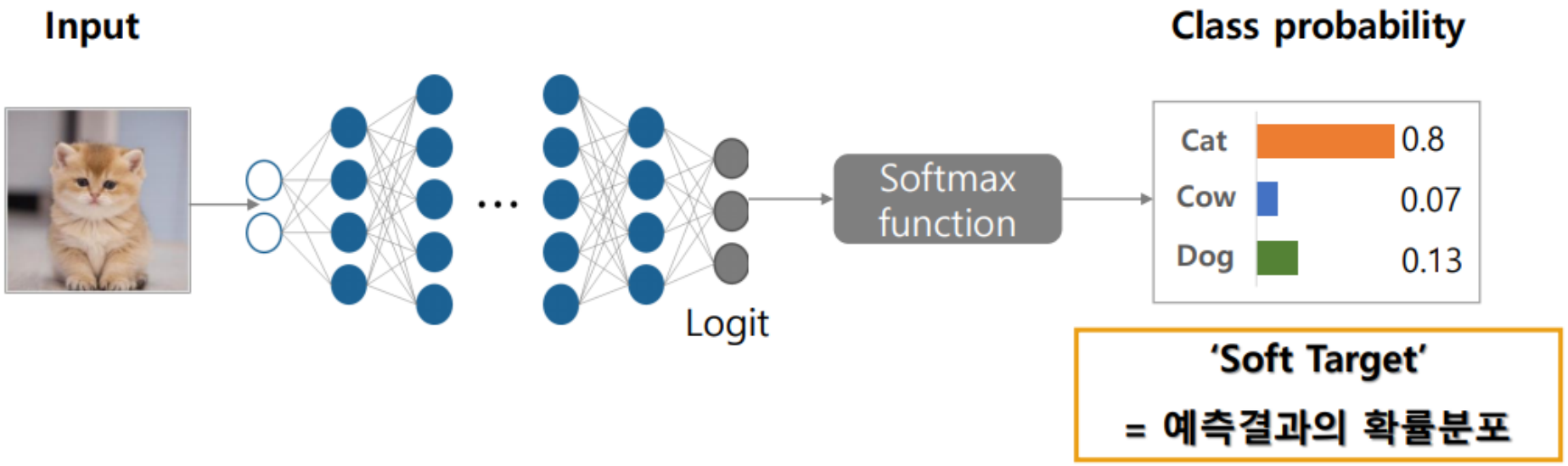
1.training set (x, hard target)을 사용해 large model을 학습한다.

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

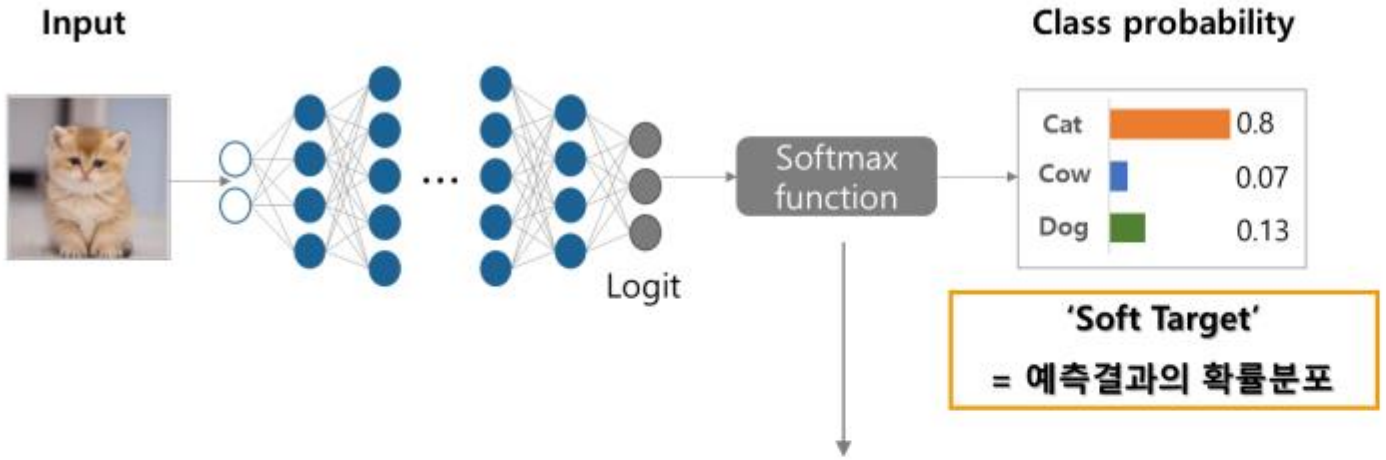
3.transfer set을 사용해 small model을 학습한다.

Target 종류


Input
Class probability
Output
Softmax function
Logit
Cat 0.8
Cow 0.07
Dog 0.13
Cat 1
Dog 0
Cow 0
One-hot Encoding
'Hard Target'


Input
Class probability
Softmax function
Logit
Cat 0.8
Cow 0.07
Dog 0.13
'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

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

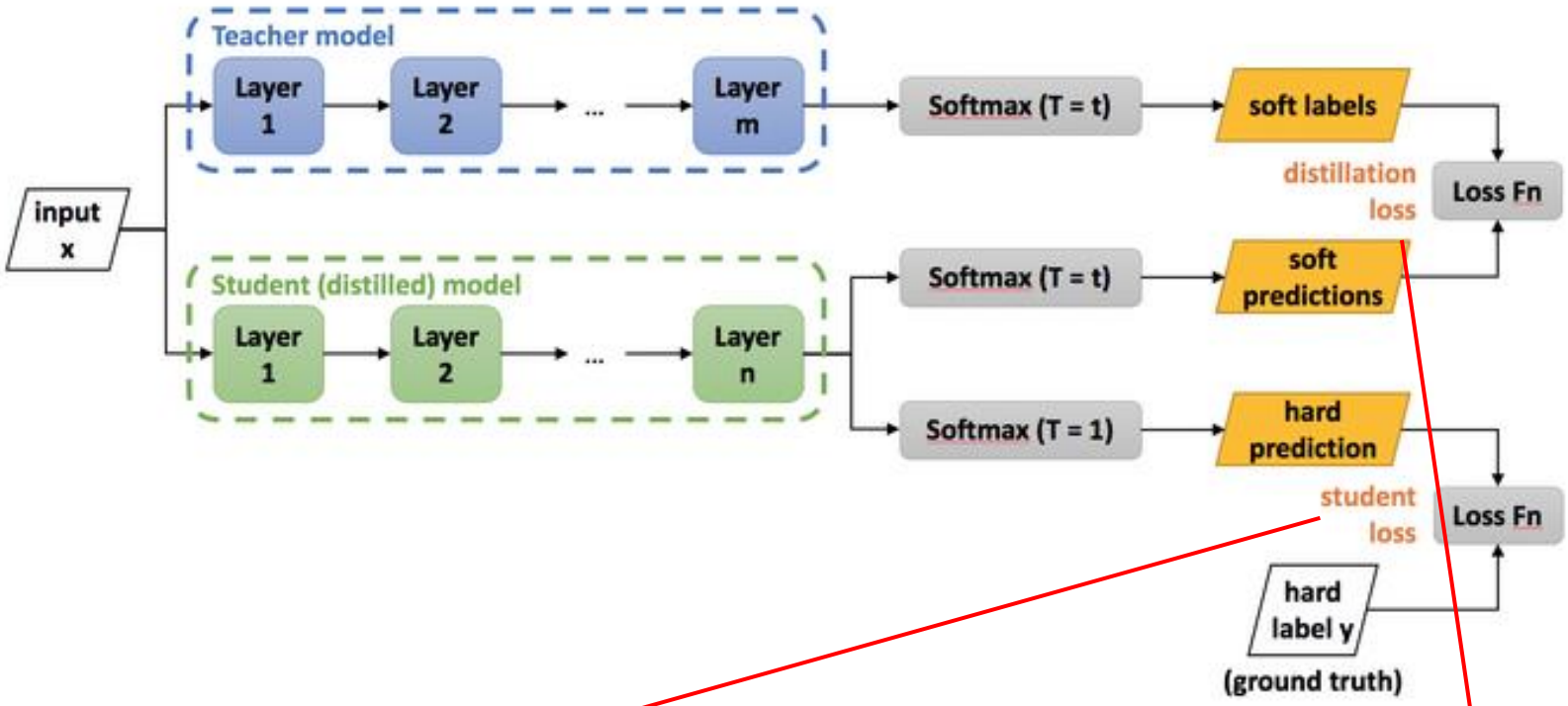
Softened output of Softmax

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

Soft targe과 hard target을 비교하기 위해 임의로 나타낸 값

Loss function



$$\mathcal{L}(x;W) = \alpha * \mathcal{H}(y, \sigma(z_s; T=1)) + \beta * \mathcal{H}(\sigma(z_t; T=\tau), \sigma(z_s, T=\tau))$$

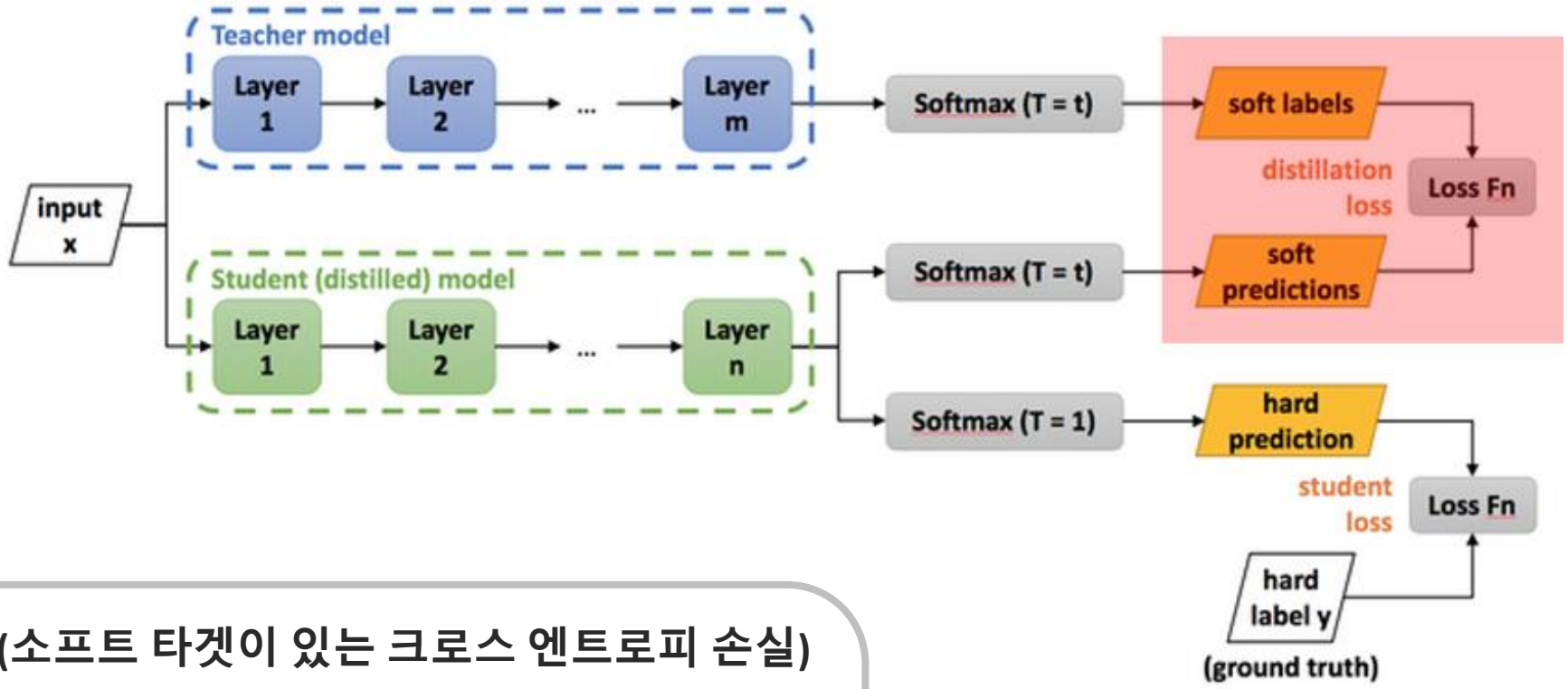
x: input, W: student model parameters, y: ground truth label(one-hot),

H: cross-entropy loss function, σ: softmax function parameter, T: Temperature,

α and β: coefficients, Zs :logits of the student, Zt: logits of the teacher

출처: "Similiarity preserving knowledge distillation", https://intellabs.github.io/distiller/knowledge_distillation.html

Loss function



Loss1 (소프트 타겟이 있는 크로스 엔트로피 손실)

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \qquad p_i = \frac{exp(v_i/T)}{\sum_j exp(v_j/T)}$$

$$\text{Loss}_1 = \alpha * \text{CE}(q_i, p_i)$$

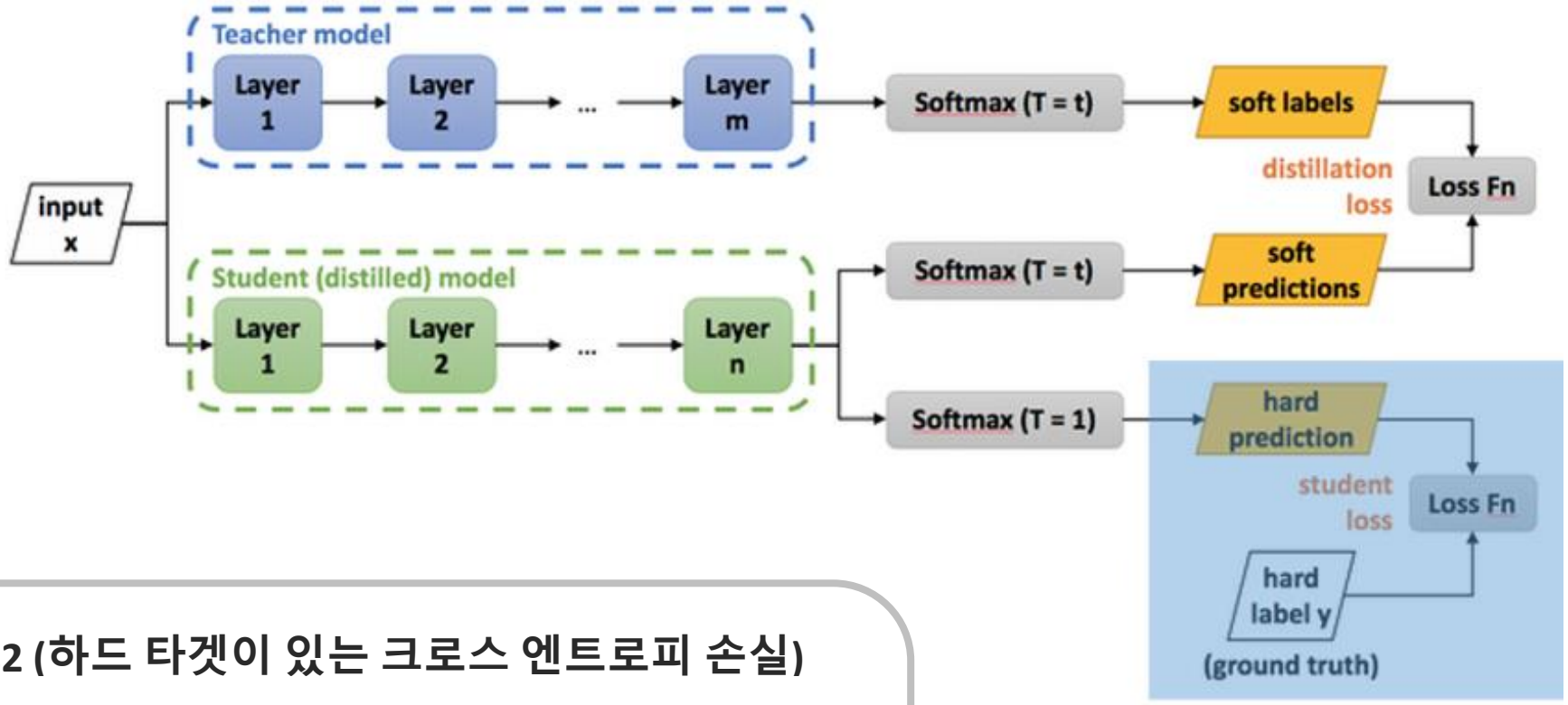
온도 T>1에서 가중치 매개 변수α를 곱한

교사(Q)와 학생(P)에 대한

두 온도 소프트 맥스의 교차 엔트로피 (CE)

출처: https://towardsdatascience.com/distilling-knowledge-in-neural-network-d8991faa2cdc

Loss function



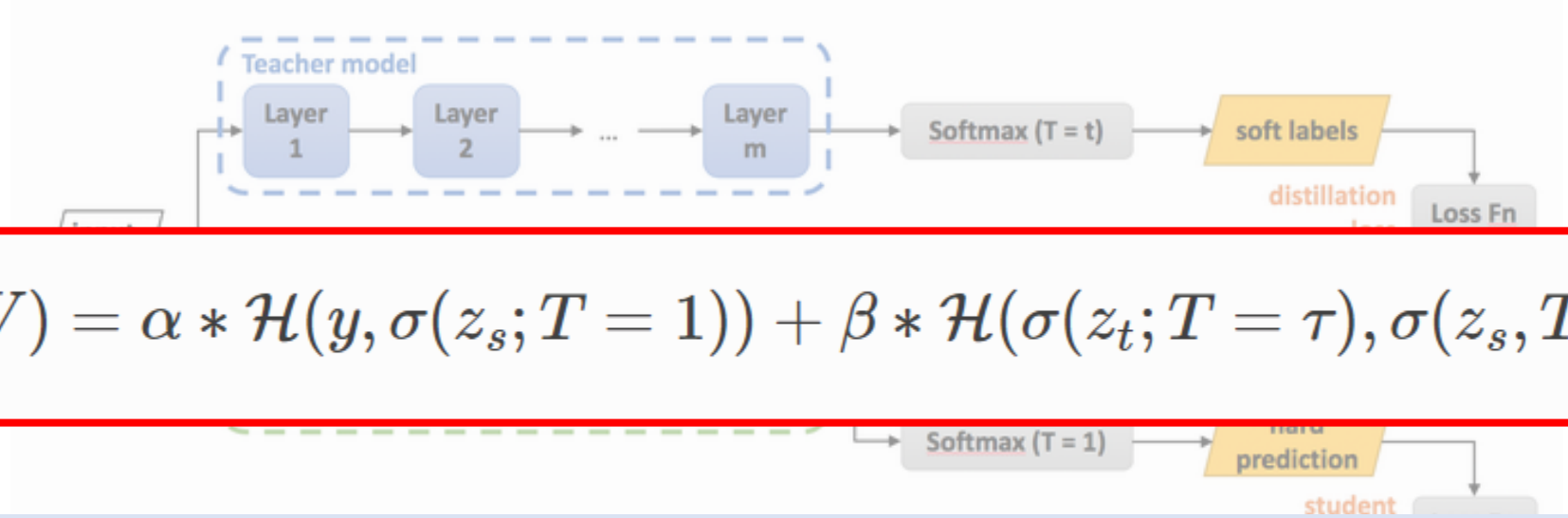
Loss2 (하드 타겟이 있는 크로스 엔트로피 손실)

$$\text{student\_pred} = argmax\left(\frac{exp(v_i/T)}{\sum_j exp(v_j/T)}\right)$$

$$\text{Loss}_2 = (1-\alpha) * \text{CE}(\text{student\_pred}, \text{y\_true})$$

Loss2는 T = 1 인 올바른 레이블과
학생 하드 타겟(y)의 교차 엔트로피 (CE)

Loss function



$$\mathcal{L}(x;W) = \alpha * \mathcal{H}(y, \sigma(z_s; T=1)) + \beta * \mathcal{H}(\sigma(z_t; T=\tau), \sigma(z_s, T=\tau))$$

Hinton의 논문에서는 증류 손실과 학생 손실 사이의 가중 평균을 사용하여 $\beta = 1-\alpha$ 를 사용함

(실험에서는 $\alpha = \beta = 0.5$를 사용)

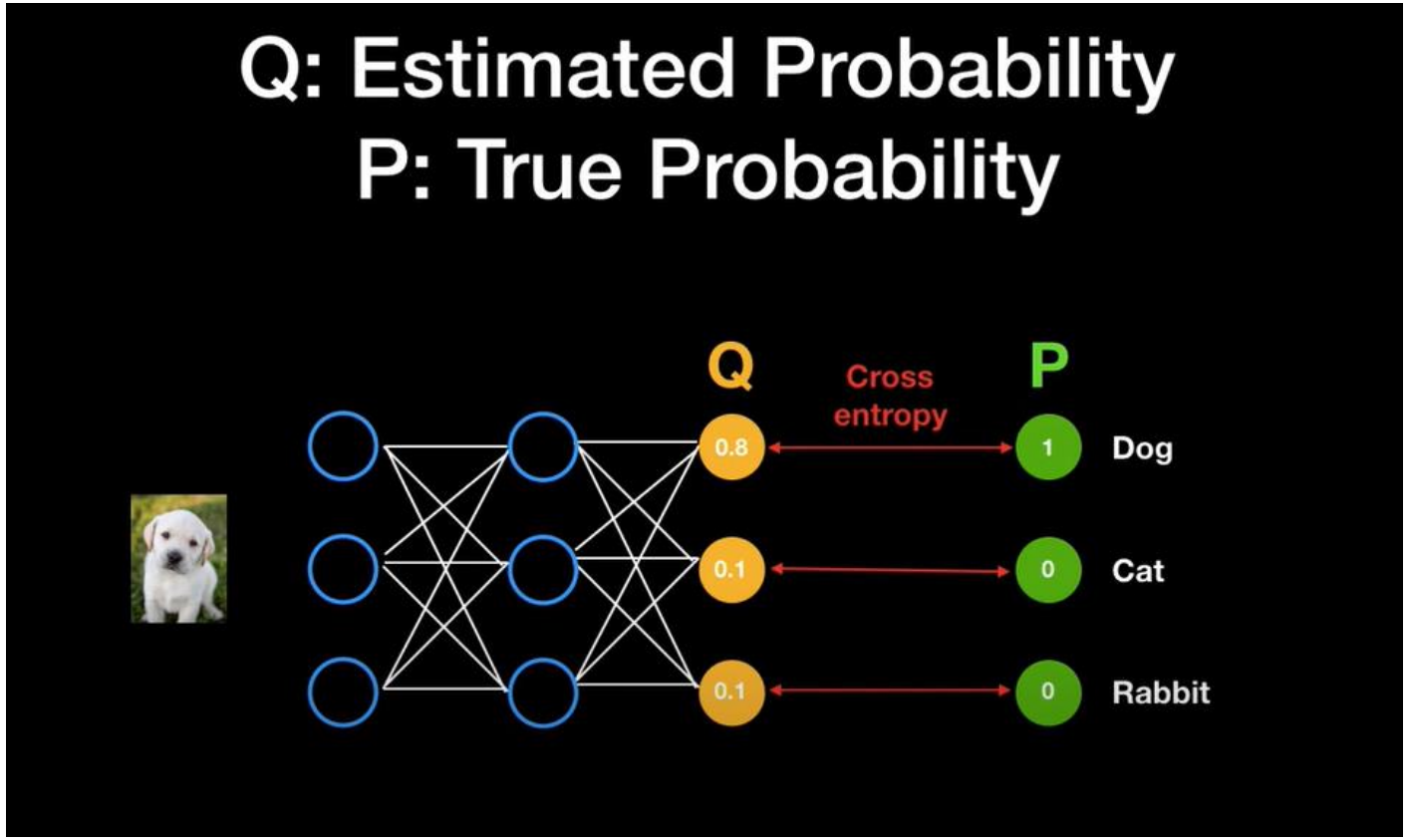
일반적으로 α를 β보다 훨씬 작게 설정할 때 좋은 결과가 나옴

지식 증류를 활용하는 다른 연구에서 가중 평균을 사용하지 않고 β 조정 가능, $\alpha = 1$을 설정

Cross-Entropy

실제 값과 예측 값의 사이에 엔트로피

불확실한 정보를 가지고 구한 엔트로피

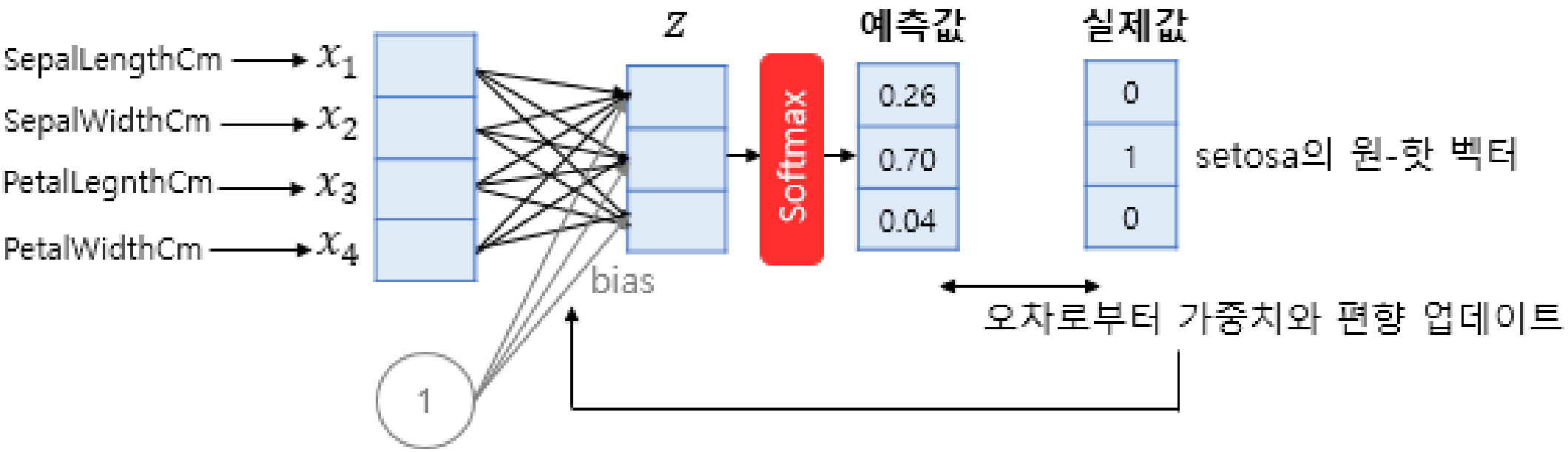


출처: 유튜브,**크로스 엔트로피**, https://www.youtube.com/watch?v=Jt5BS71uVfI

Cross-Entropy

어떻게 사용되는가?

딥러닝 모델의 Cost function



Cross-entropy를 Cost function으로 하여 deep learning 모델을 최적화할 수 있다.

출처: 딥 러닝을 이용한 자연어 처리 입문 9)소프트 맥스 회귀 **(Softmax Regression)** - **다중 클래스 분류,** https://wikidocs.net/35476

Entropy

엔트로피는 **불확실성의 척도**

$$H(x) = -\sum_{i=1}^{n} p(x_i) log p(x_i)$$

엔트로피는 불확실성을 나타내며, 엔트로피가 높다는 것은 정보가 많고, 확률이 낮다는 것을 의미

- 동전을 던졌을 때, 앞/뒷면이 나올 확률을 1/2

$$H(x) = -\left(\frac{1}{2} log \frac{1}{2} + \frac{1}{2} log \frac{1}{2}\right) \approx 0.693$$

- 주사위를 던졌을 때, 각 6면이 나올 확률을 1/6

$$H(x) = -\left(\frac{1}{6} log \frac{1}{6} + \frac{1}{6} log \frac{1}{6} + \frac{1}{6} log \frac{1}{6} + \frac{1}{6} log \frac{1}{6} + \frac{1}{6} log \frac{1}{6} + \frac{1}{6} log \frac{1}{6}\right) \approx 1.79$$

출처: **엔트로피(Entropy)와 크로스 엔트로피(Cross-Entropy)의 쉬운 개념 설명,** http://melonicedlatte.com/machinelearning/2019/12/20/204900.html

Cross-Entropy

Entropy vs Cross Entropy

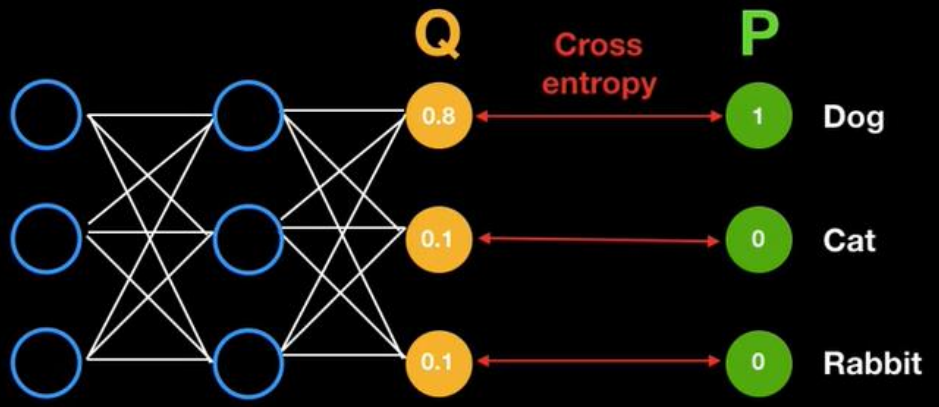
Amount of info

Probability distribution

$$H(X) = \sum_{i=1}^{n} \log_2 \frac{1}{p_i} * p_i$$

$$H(P, Q) = \sum_{i=1}^{n} \log_2 \frac{1}{q_i} * p_i$$

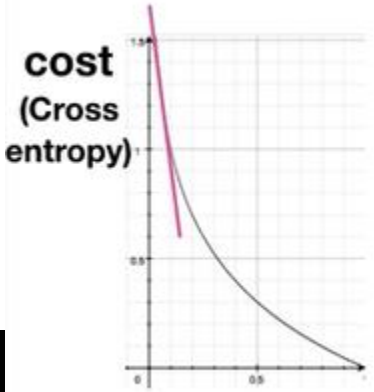
actual prediction info based on Q which is the output of the model



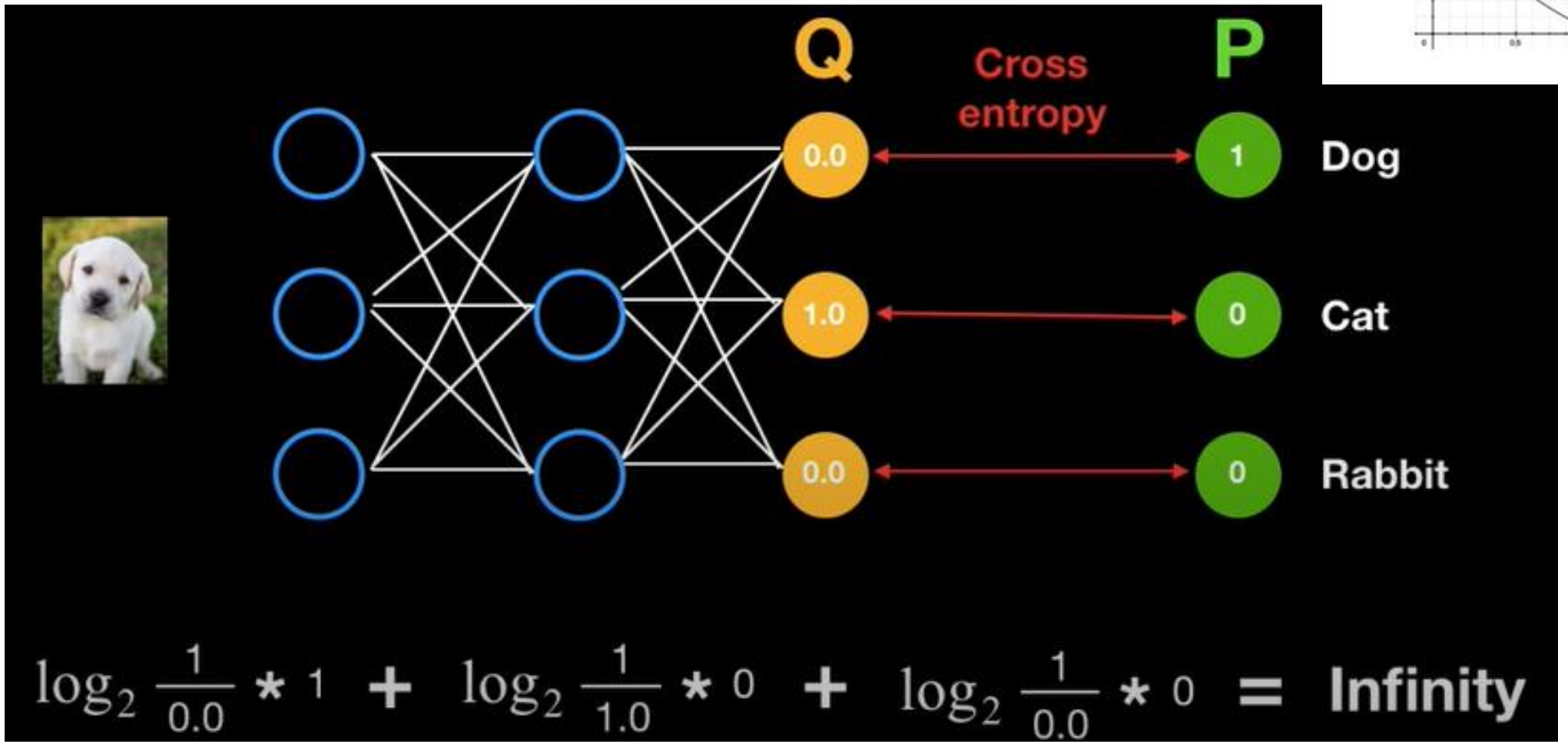
$$H(P, Q) = \sum_{i=1}^{n} \log_2 \frac{1}{q_i} * p_i$$

$$H_p(q) = -\sum_{i=1}^{n} q(x_i) log p(x_i)$$

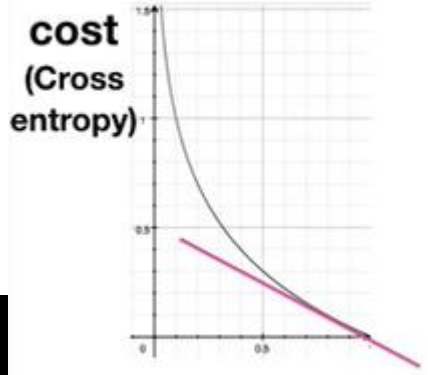
Cross-Entropy



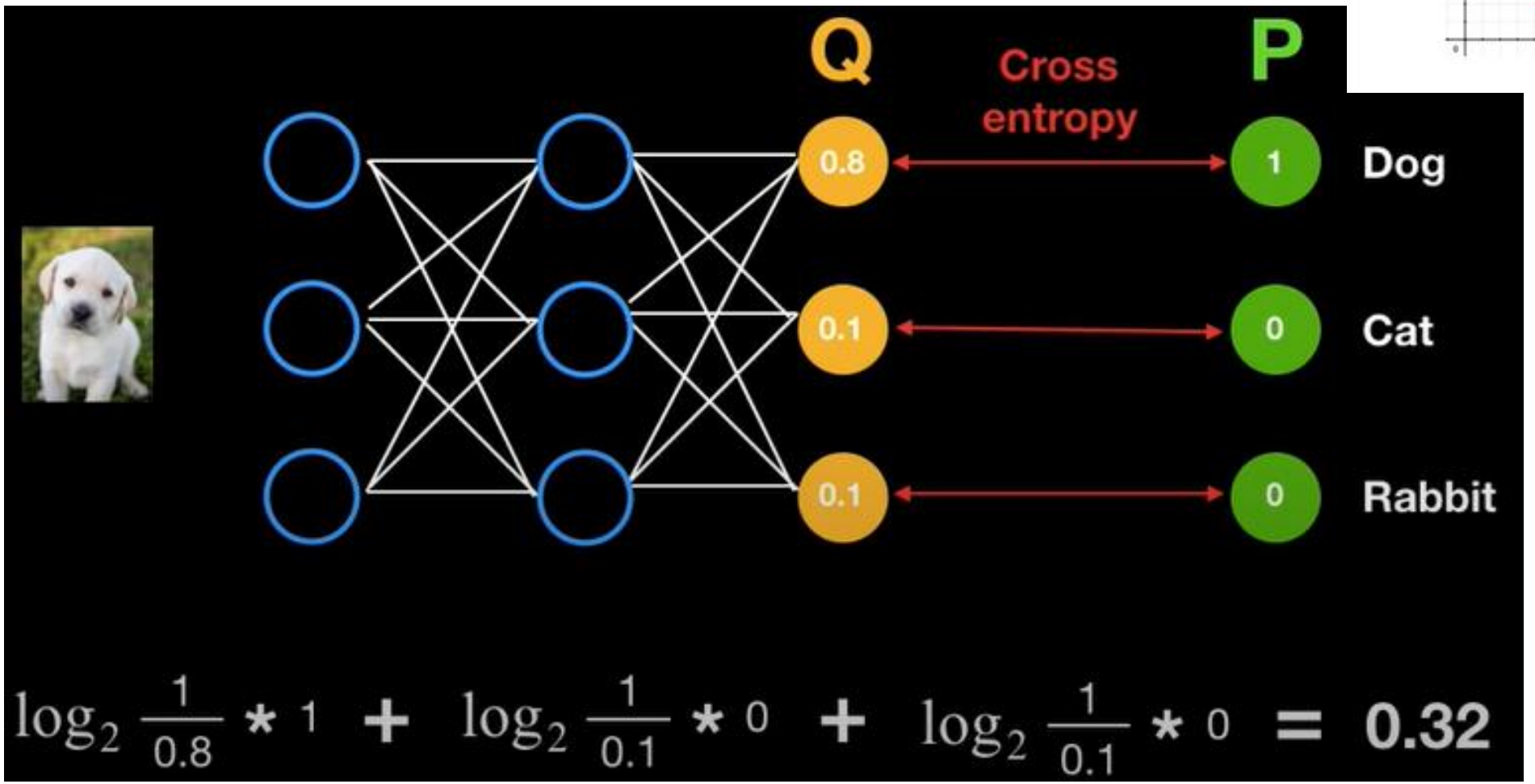
예측값이 실제값과 완전히 다를 경우



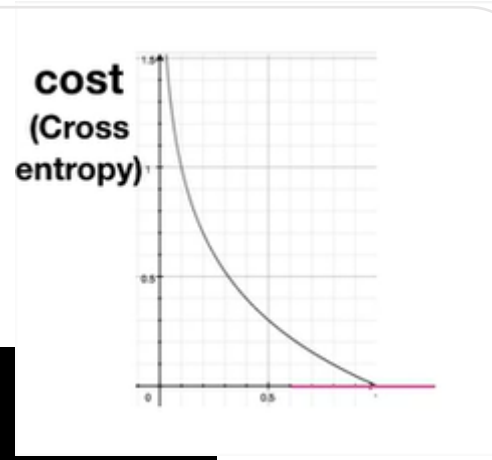
Cross-Entropy



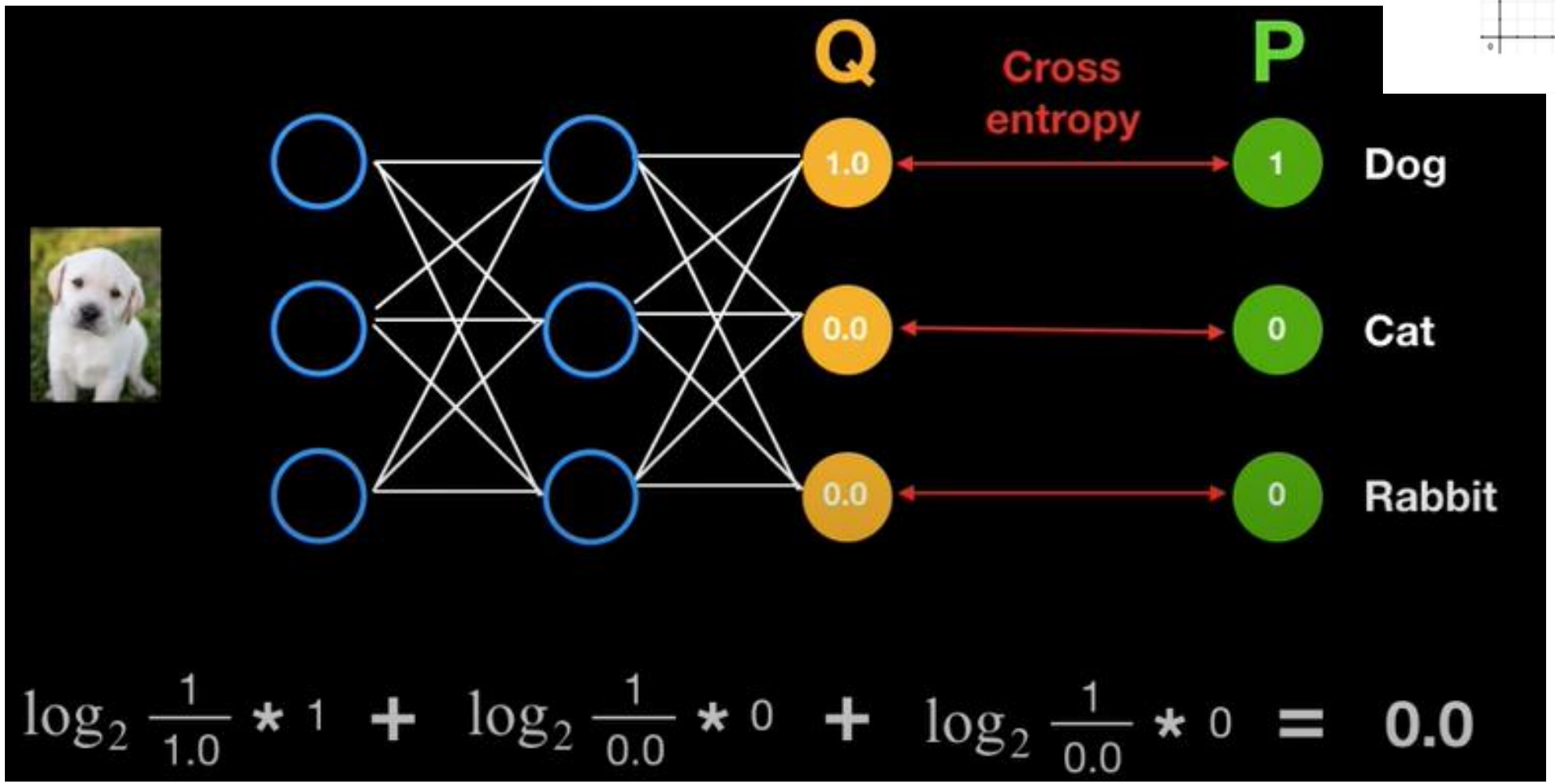
예측값이 실제값과 비슷한 경우



Cross-Entropy



예측값이 실제값과 동일한 경우



MNIST

MNIST dataset 는 손글씨 이미지 분류에 사용되는 data set으로 손으로 쓴 0~9 까지의 글자로 이루어져 있고, 학습 데이터 6만장과 테스트 데이터 1만장이 주어진 셋으로 이중 트레이닝 셋을 학습데이터로 사용하고 테스트 셋을 신경망을 검증하는 데에 사용한다.

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

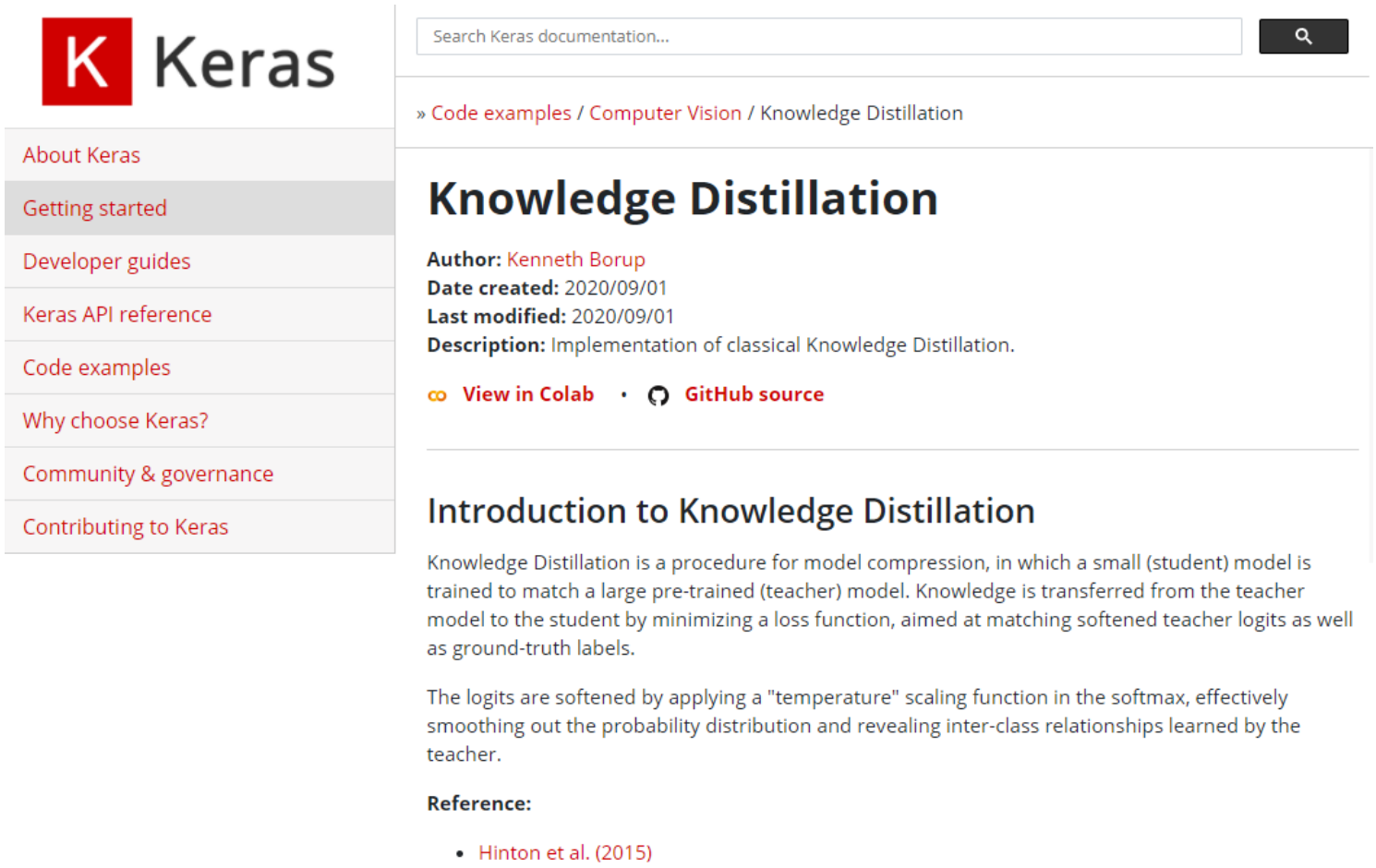
System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

- 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 $2 \le T \le 4$)
- Soft label을 통해 소실된 데이터를 유추
- Soft target을 사용하는 것은 overfitting을 방지 Regularizer

Application



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/**

추가

- **https://baeseongsu.github.io/posts/knowledge-distillation/**
- **엔트로피(Entropy)와 크로스 엔트로피(Cross-Entropy)의 쉬운 개념 설명**http://melonicedlatte.com/machinelearning/2019/12/20/204900.html
- 유튜브,**크로스 엔트로피, https://www.youtube.com/watch?v=Jt5BS71uVfI**
- 딥 러닝을 이용한 자연어 처리 입문 9)소프트 맥스 회귀 **(Softmax Regression) - 다중 클래스 분류,** https://wikidocs.net/35476
- "Similiarity preserving knowledge distillation", https://intellabs.github.io/distiller/knowledge_distillation.html
- https://towardsdatascience.com/distilling-knowledge-in-neural-network-d8991faa2cdc

Thank You