Distilling the Knowledge in a Neural Network

Geoffrey Hinton ,Oriol Vinyals, Jeff Dean

Google Inc.

NIPS 2015

2020.04.23

임진혁

0 Index

관심있는 분야 :

실제 현실에 적용되고 활용될 수 있는 신경망

→ 학습데이터가 적거나 없는 상황에서도 일정한 성능이 가능한 신경망

관련 키워드:
Cold Start, Few shot Learning , Meta Learning , Transfer Learning , Explainable AI
Causality Training

그렇다면
KD(Knowledge Distillation)을 선택한 이유는?

관심있는 분야 :

실제 현실에 적용되고 활용될 수 있는 신경망

→ 학습데이터가 적거나 없는 상황에서도 일정한 성능이 가능한 신경망

관련 키워드:
Cold Start, Few shot Learning , Meta Learning , Transfer Learning , Explainable AI
Causality Training

실제 세상에서 활용할려면 Model의 Cost를 생각하지 않을 수 없다.

애초에 우리 곁에서 활용될려면 mobil이나 웹에서 돌아가야 한다.

또한, knowledge를 전달한다는 개념 자체가 매우 궁금하였다.
어떤 방식으로 전달한다는 걸까??

해당 주제를 선택한 이유

Geoffrey Hinton , Oriol Vinyals, Jeff Dean

제프리 힌턴 (Geoffrey Hinton)

컴퓨터 과학자

제프리 에버레스트 힌턴은 인공지능 분야를 개척한 영국 출신의 인지 심리학자이자 컴퓨터 과학자이다. 구글과 토론토 대학교에 재직 중이다.오류 역전파법과 딥 러닝 연구에 기여했다. . 힌턴 다이어그램을 발명했다. 위키백과

출생: 1947년 12월 6일 (72세), 영국 런던 윔블던

논문: Relaxation and its role in vision (1977)

학력: 에든버러 대학 (1972년–1975년), 더보기

유명한 제자: 얀 르쿤, 일리야 서츠케버, 루스 살라쿠디노프, 더보기

저서: Neural Network Architectures for Artificial Intelligence

수상: 튜링상, Rumelhart Prize, IJCAI 우수 연구 상, 더보기

Distilling the knowledge in a neural network

G Hinton, O Vinyals, J Dean - arXiv preprint arXiv:1503.02531, 2015 - arxiv.org

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. Unfortunately, making predictions using a whole ensemble of models is cumbersome and ...

☆ 99 3235회 인용 관련 학술자료 전체 20개의 버전 »

제프 딘(Jeff dean)에 대해서 알려주세요!!

검색해봐도 제프딘의 진실?그런 거 밖에 안나오던데 그거 빼고 업적이나 한 말? 같은거좀 알려주세요

Ξ 전기, 전자 공학

비공개 · 2018.05.26 · 조회수 731

1개 채택순 전체보기 최적 원문

NooN 님 답변

수호신 · 채택답변수 21,168 · 받은감사수 42 · 영어문법60위, 영어작문10위, 영어 공부, ···

질문자채택

Jeff Dean은 구글의 전설적인 프로그래머로, BigTable, MapReduce등 구글의 핵심기술을 만들었습니다. 그의 위대함을 칭송하기 위해 사람들은 '척 노리스의 진실'을 패러디한 '제프 딘의 진실'을 만들었습니다.

컴파일러는 제프 딘에게 경고하지 않는다. 그가 컴파일러에게 경고한다.
2000년 후반에 제프 딘의 코드 작성 속도가 40배로 빨라졌는데, 그가 키보드를 USB 2.0으로 업그레이드 했기 때문이다.
제프 딘은 커밋하기 전에 코드를 빌드해보는데, 컴파일러와 링커에 버그가 있는지 확인하기 위해서일

Deep neural networks for **youtube** recommendations

P Covington, J Adams, E Sargin - ... conference on **recommender systems**, 2016 - dl.acm.org

YouTube represents one of the largest scale and most sophisticated industrial recommendation **systems** in existence. In this paper, we describe the system at a high level and focus on the dramatic performance improvements brought by deep learning. The paper ...

☆ 99 833회 인용 관련 학술자료 전체 22개의 버전

Distilling Knowledge란?

Distillation(증류)가 뭡니까?

- 불순물이 섞여 있는 혼합물에서 특정 성분을 분리시키는 방법

두산백과

증류

[distillation , 蒸溜]

요약 어떤 용질이 녹아 있는 용액을 가열하여 얻고자 하는 액체의 끓는점에 도달하면 기체상태의 물질이 생긴다. 이를 다시 냉각시켜 액체상태로 만들고 이를 모으면 순수한 액체를 얻어낼 수 있는데, 이러한 과정을 증류라 한다.

NN에서의 Distillation이란?

불순물(불피요한 많은 파라미터)에서 특정 파라미터(일반화 능력)을 분리해내는 것

왜 이런 것을 해야 하는가?

Distilling Knowledge란?

모델 압축(compression)의 필요성

최근 SOTA 모델들은 굉장히 무거움(학습해야 할 파라미터 개수)

→ 효율이 떨어지고 , Test Inference도 매우 느림 (GoogleNet 등)

→ 컴퓨팅 파워가 떨어지는 장비에서는 돌릴 수도 없음
→ 모델이 무거울수록 Cost가 높아짐 (유지 비용 등)

Distilling Knowledge는 모델 압축의 한 방법

무거운 모델의 능력을 작고 얇은 모델에게 옮기는 것 → 에엥? 이게 가능하다고?

원하는 것
1. 무거운 모델과 비슷한 성능
2. 훨씬 가벼운 구조

Q: Transfer Learning과의 차이점은?

모델 압축(compression)의 필요성

최근 SOTA 모델들은 굉장히 무거움(학습해야 할 파라미터 개수)

→ 효율이 떨어지고 , Test Inference도 매우 느림 (GoogleNet 등)

→ 컴퓨팅 파워가 떨어지는 장비에서는 돌릴 수도 없음
→ 모델이 무거울수록 Cost가 높아짐 (유지 비용 등)

Distilling Knowledge는 모델 압축의 한 방법

무거운 모델의 능력을 작고 얇은 모델에게 옮기는 것 → 에엥? 이게 가능하다고?

필요조건
1. 무거운 모델과 비슷한 성능
2. 훨씬 가벼운 구조

도메인 A -> B

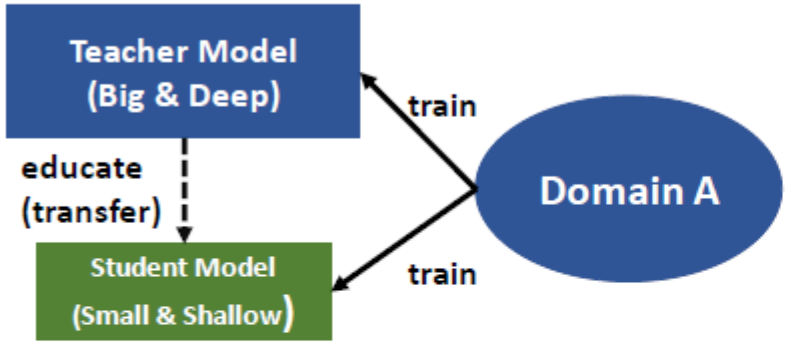
도메인 A -> A

Distilling Knowledge란?

도메인 A -> B

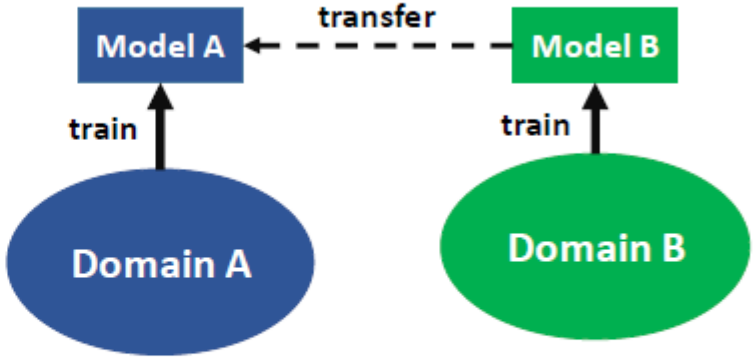
도메인 A -> A

Knowledge Distillation (Transfer)

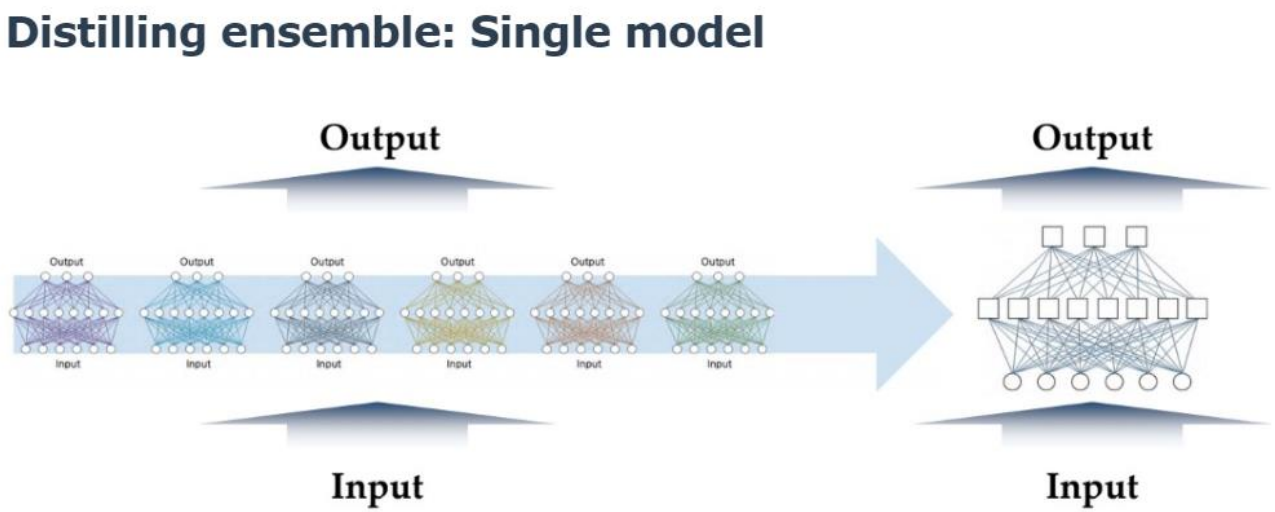


- For model compression
- To improve performance of student over teacher

Transfer Learning



- When data is not sufficient.
- When label for a problem is not presented.
- E.g., pretrained-model on ImageNet

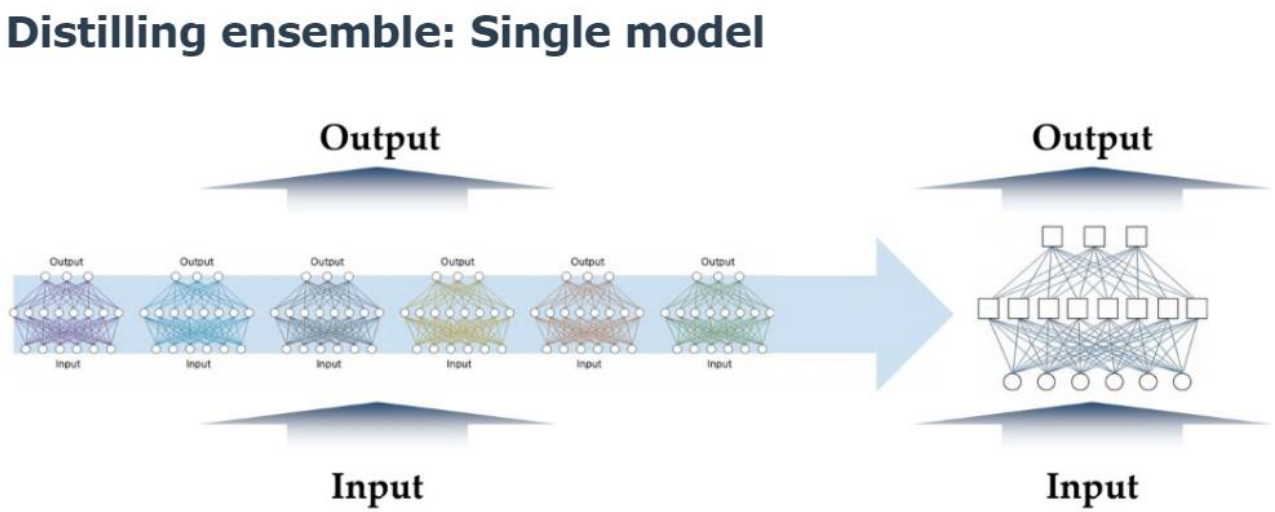


해당 논문에서의 Distilling Knowledge

무거운 모델(앙상블 모델)의 능력(일반화 능력)을 가벼운 모델(single model)로 옮기기

Hard Label -> Soft Label [*Knowledge Discovery and Data Mining(2006)]*

Q: Logit이란? (저번 시간에 나왔던 개념)_



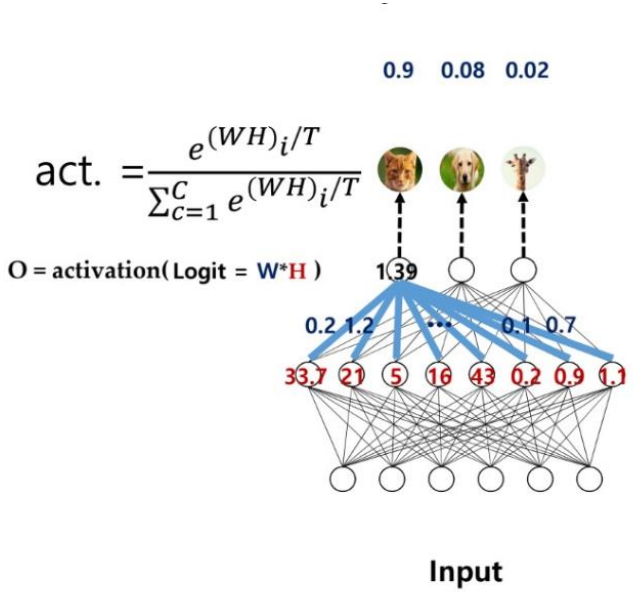
해당 논문에서의 Distilling Knowledge

무거운 모델(앙상블 모델)의 능력(일반화 능력)을 가벼운 모델(single model)로 옮기기

Hard Label -> Soft Label [*Knowledge Discovery and Data Mining(2006)]*

Q: Logit이란? (저번 시간에 나왔던 개념)_

Soft Label과 Temperature T



모델의 출력물에는 그 모델이 학습한 지식이 담겨있다

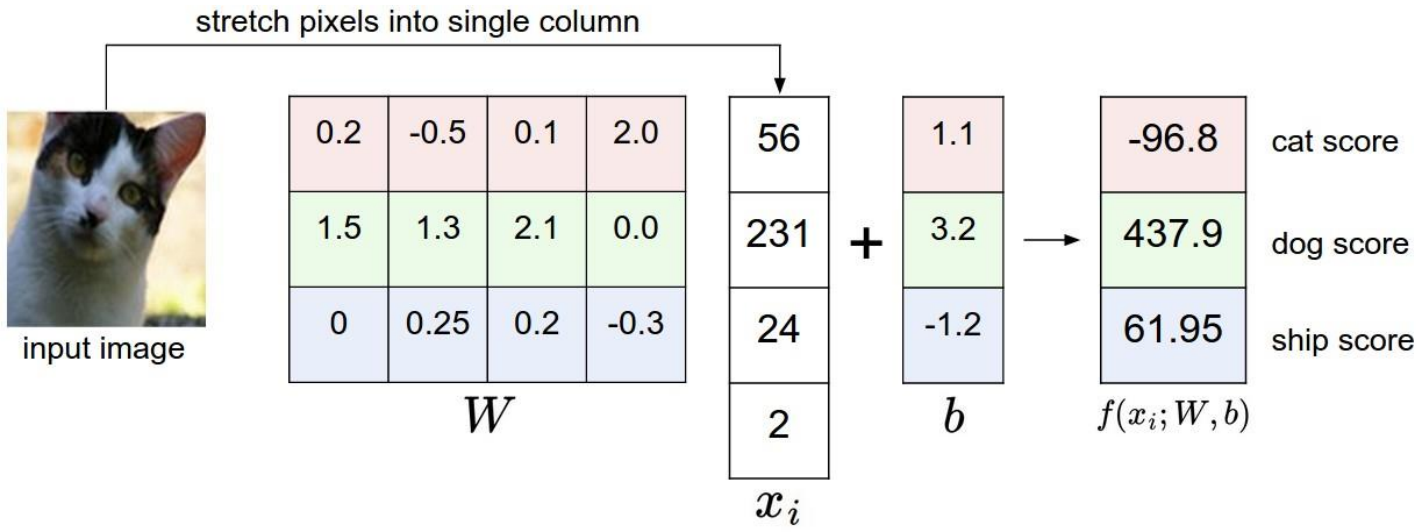
(사실상 Approach The End)

그럼 하고자 하는 것은?

그 지식(정보량)이 최대한 많이 들어가도록 해주자

Hard Target(argmax) -> output (soft max) -> Soft Target (soft max with T)

Hard Target(argmax) -> output (soft max) -> Soft Target (soft max with T)

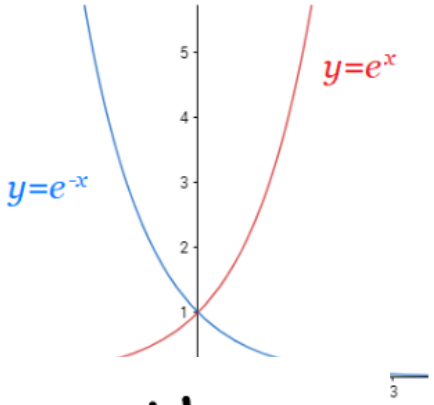


고양이 3 0.7
개 1 0.2
양 -2 0.1

Soft Label과 Temperature T

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

Softmax with temperature T



T가 커질수록 Smoothing

An example of hard and soft targets

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

Softened outputs reveal the dark knowledge in the ensemble.

Soft Label과 Temperature T

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

Softmax with temperature T

T가 커질수록 Smoothing

An example of hard and soft targets

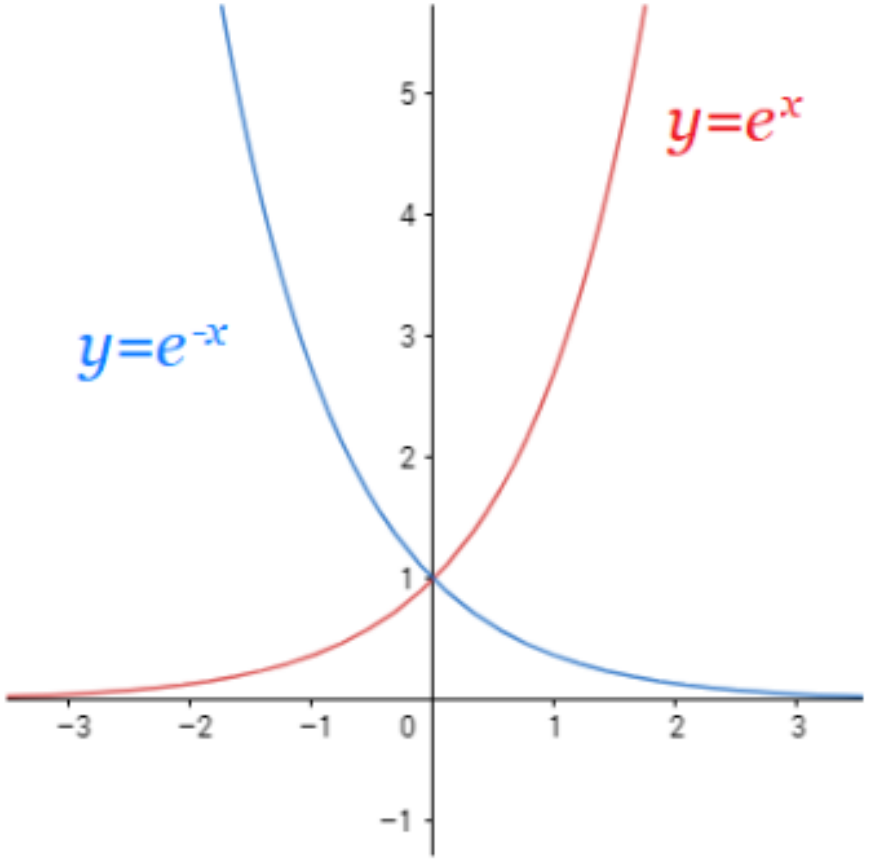
cow	dog	cat	car	
0	1	0	0	original hard targets
cow	dog	cat	car	
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble
cow	dog	cat	car	
.05	.3	.2	.005	softened output of ensemble

Softened outputs reveal the dark knowledge in the ensemble.

왜?

T가 커질수록 입력값들이

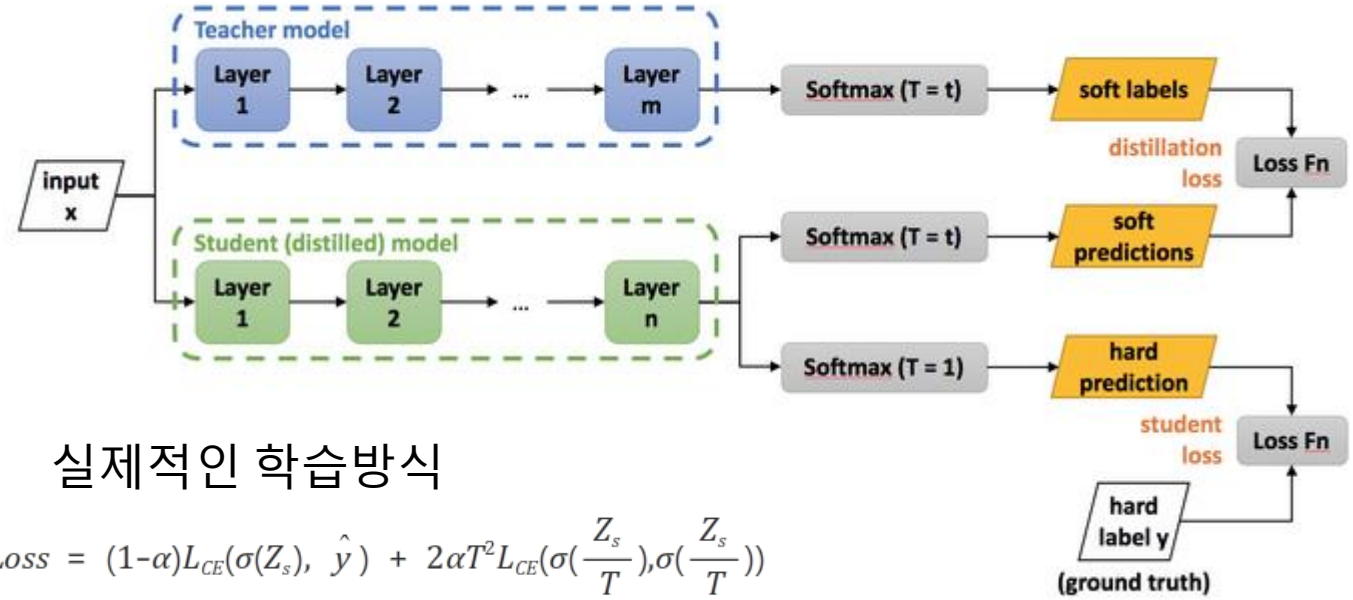
작아진다!!!

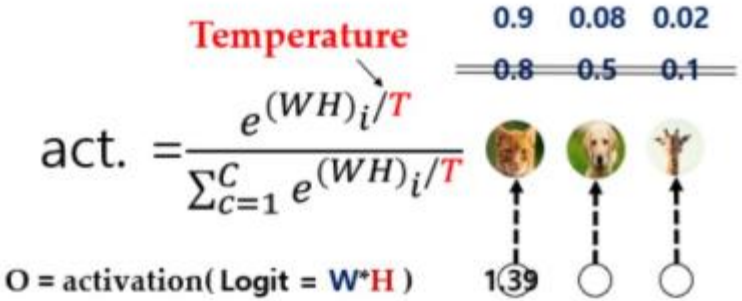


Smoothing -> More Soft

Hard target -> Soft Target

Soft Label과 Temperature T





실제적인 학습방식

$$Total\ Loss = (1-\alpha)L_{CE}(\sigma(Z_s),\ \hat{y}) + 2\alpha T^2 L_{CE}(\sigma(\frac{Z_s}{T}),\sigma(\frac{Z_s}{T}))$$

$L_{CE}()$: *Cross entropy loss*
$\sigma()$: *Softmax*
Z_s: *Output logits of Student network*
Z_t: *Output logits of Teacher network*
$\hat{y}$: *Ground truth(one-hot)*
α: *Balancing parameter*
T: *Temperature hyperparameter*

Q: 왜 loss에 2 * T가 곱해지는가?

3 Experiments

NET A

NET B

NET C

Result 1: MNIST image data

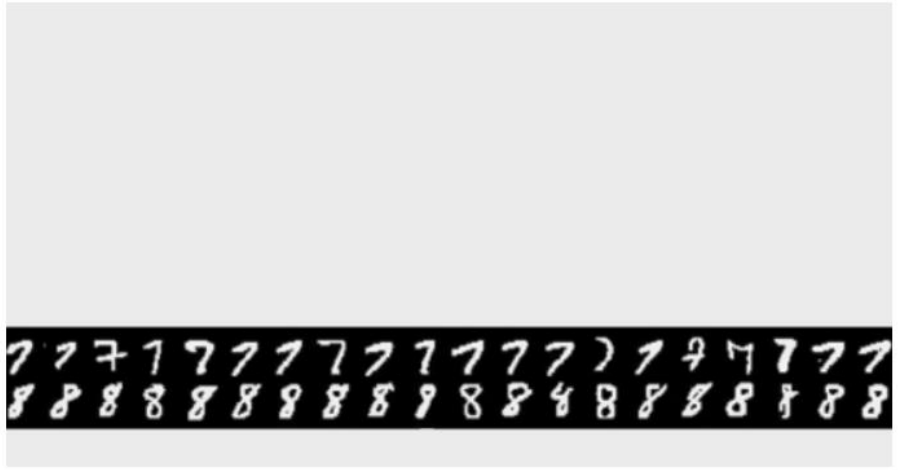
Two hidden layer + ReLU:	146 test errors
Two hidden layer + ReLU + DropOut:	67 test errors
Two hidden layer + ReLU + Soft Target:	74 test errors

Results 2: speech recognition

System	Test Frame Accuracy
Baseline	58.9%
10xEnsemble	61.1%
Distilled Single model	60.8%

3 Experiments

MNIST without “3”



MNIST with only “7” and “8”

학습데이터에 없는 “클래스”에 대해 큰 예측율을 보임

→ COLD START , FEW SHOT 러닝에 큰 영향

결론 및 정리

성능뿐만 아니라 효율을 원한다.

복잡한 모델 = 앙상블 모델 : Teacher

간단하고 빠른 모델 : Student

Teacher의 지식을 Student에게 전달(transfer)하고 싶다.

Soft max Out put은 해당 모델의 지식을 포함한다.

해당 지식의 정보량(함축정도)를 높이기 위해 Smoothing을 해준다. -> Soft Label

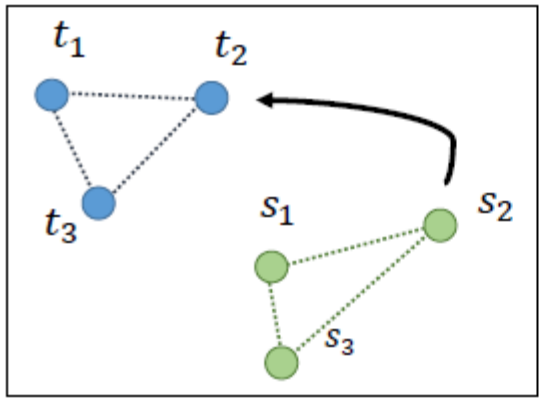
Soft Label로 학습!!

Relational Knowledge Distillation

$$\mathcal{L}_{\text{RKD-D}} = \sum_{(x_i, x_j) \in \mathcal{X}^2} l_\delta\left(\psi_{\text{D}}(t_i, t_j), \psi_{\text{D}}(s_i, s_j)\right)$$

	CIFAR-10	CIFAR-100
Baseline	92.47	71.26
Hinton *et al.*	92.84	74.26
RKD-D	92.64	72.27
RKD-DA	93.02	72.97
RKD-DA + Hinton *et al.*	**93.11**	**74.66**
Teacher	95.01	77.76

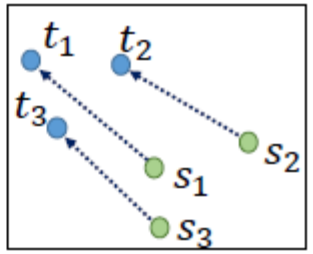
(a) Accuracy (%) on CIFAR-10 and CIFAR-100.



Structure to Structure

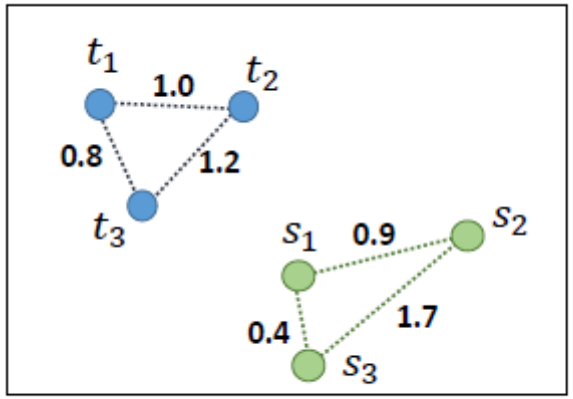
Relational KD

vs.



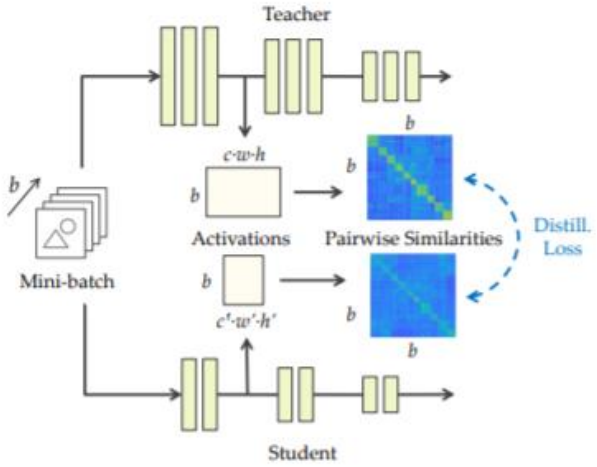
Point to Point

Individual KD



Embedding Space

Similarity-Preserving Knowledge Distillation (ICCV 2019)



$G_S^{(l')}$ WideResNet-16-1 (0.2M params)

$G_T^{(l)}$ WideResNet-40-2 (2.2M params)

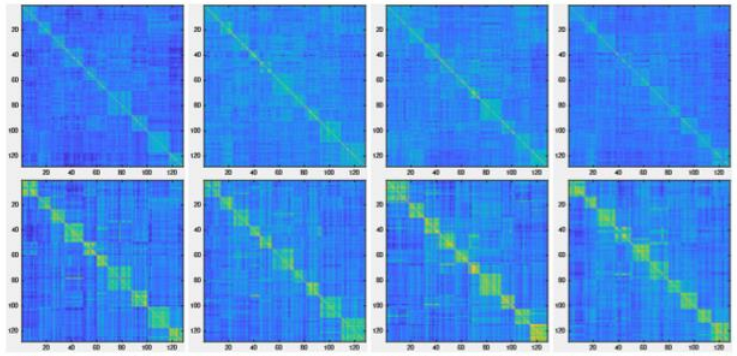


Figure 3. Activation similarity matrices G (Eq. 2) produced by trained WideResNet-16-1 and WideResNet-40-2 networks on sample CIFAR-10 test batches. Each column shows a single batch with inputs grouped by ground truth class along each axis (batch size = 128). Brighter colors indicate higher similarity values. The blockwise patterns indicate that the elicited activations are mostly similar for inputs of the same class, and different for inputs across different classes. Our distillation loss (Eq. 4) encourages the student network to produce G matrices closer to those produced by the teacher network.

아예 Activation map 자체를 Target으로 학습

비슷한 사진은 비슷한 activation map을 가질 것이다.

Activation similiarity를 통해 학습확인

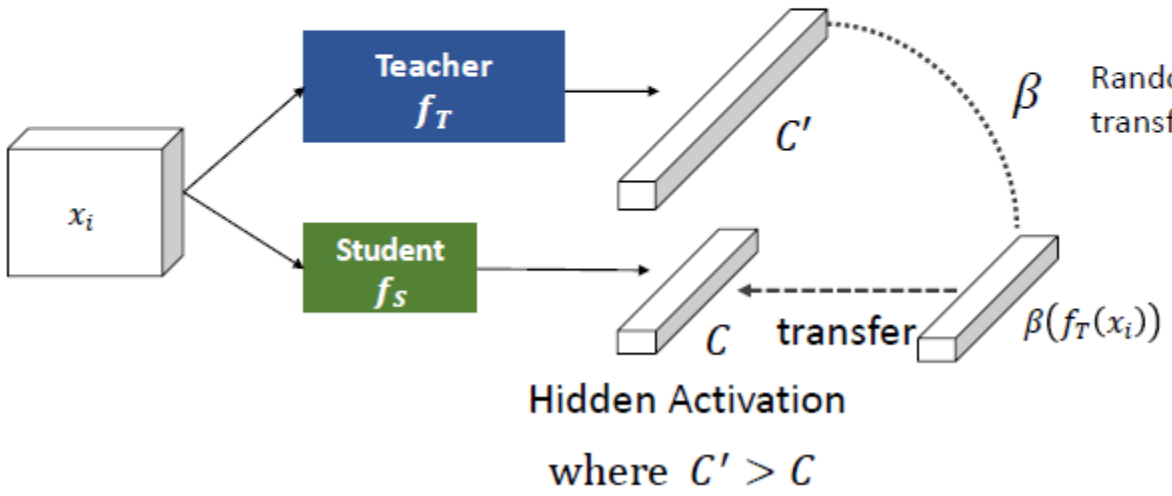
$$L = L_{CE}(y, \sigma(z_S)) + \gamma L_{SP}(G_T, G_S)$$

γ: *Balancing parameter*

Deep한 네트워크에서는 최적화가 잘안되는 문제가 발생

- **FitNets: Hints for Thin Deep Nets**
  Romero *et al*. In *ICLR*, 2015.

Objective: $\sum_{x_i \in \mathcal{X}} \left\| f_T(x_i) - \beta\left(f_S(x_i)\right) \right\|_2^2$



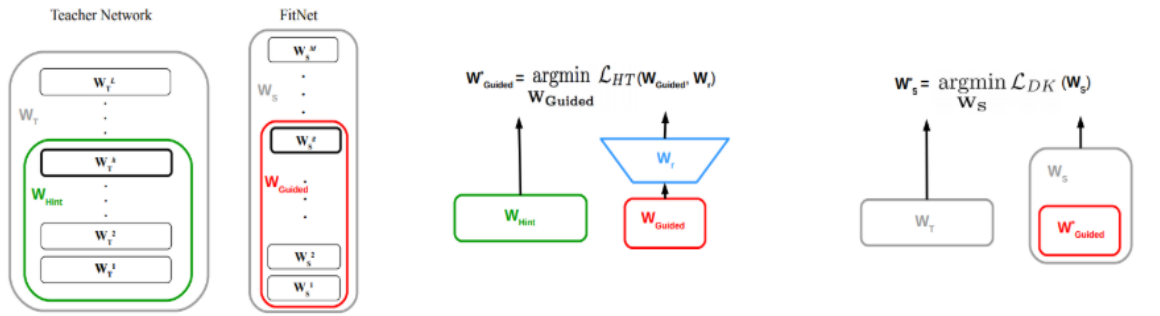
Student를 더 깊은 모델로 설정

아예 중간 레이어 자체를 Target으로 학습

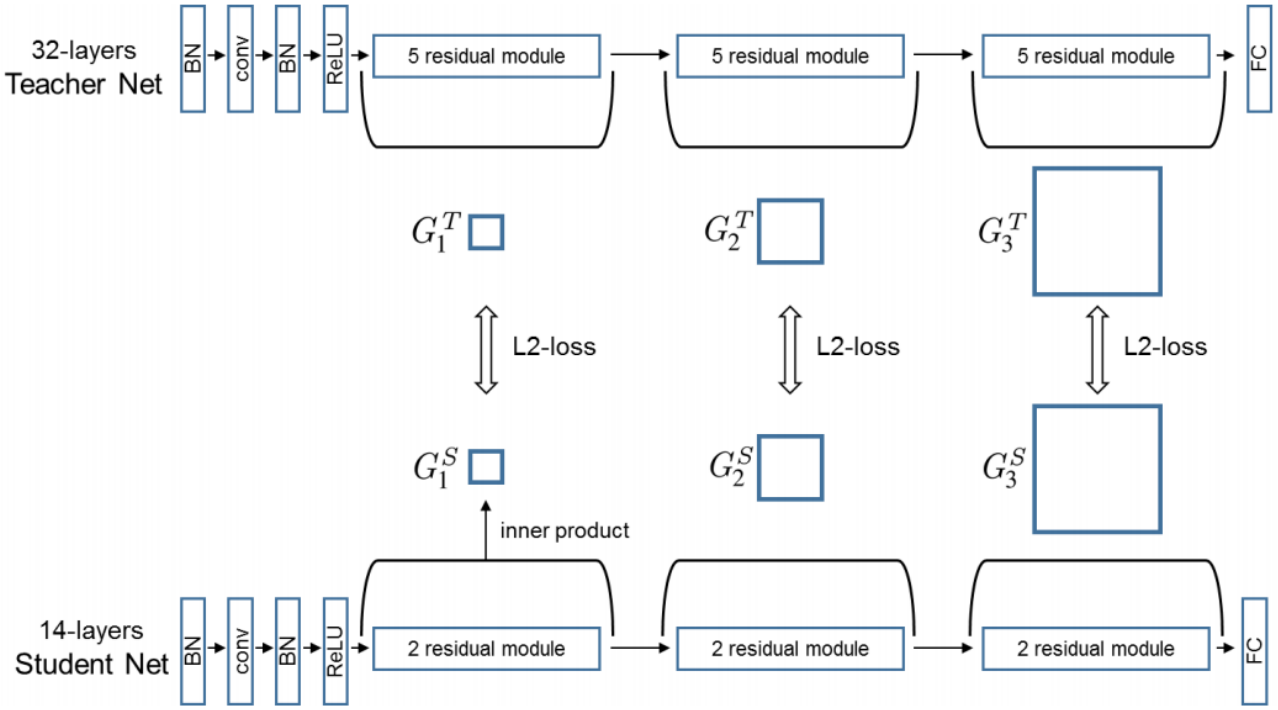
크기가 다른 점을 해결하기 위해

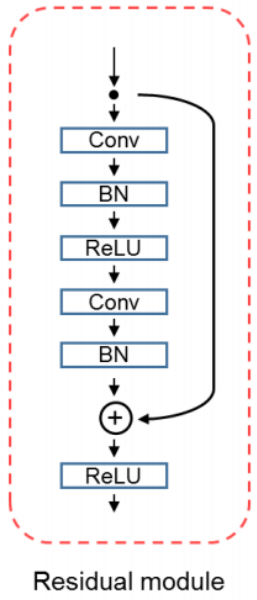
형태를 맞춰주는 레이어가 필요

중간 레이어 (Hint)를 학습하도록 함.



A Gift from Knowledge Distillation- Fast Optimization, Network Monimization and Tranfer Learning





Residual module

앞서 언급한 방법이 모델 학습의 자유도를 방해하다고 주장

KD는 단순히 답을 전달하는게 아니라 Knowledge가 문제를 푸는 방식(flow)에 있을 것이라고 생각

Flow는 각 레이어 사이의 features라고 정의

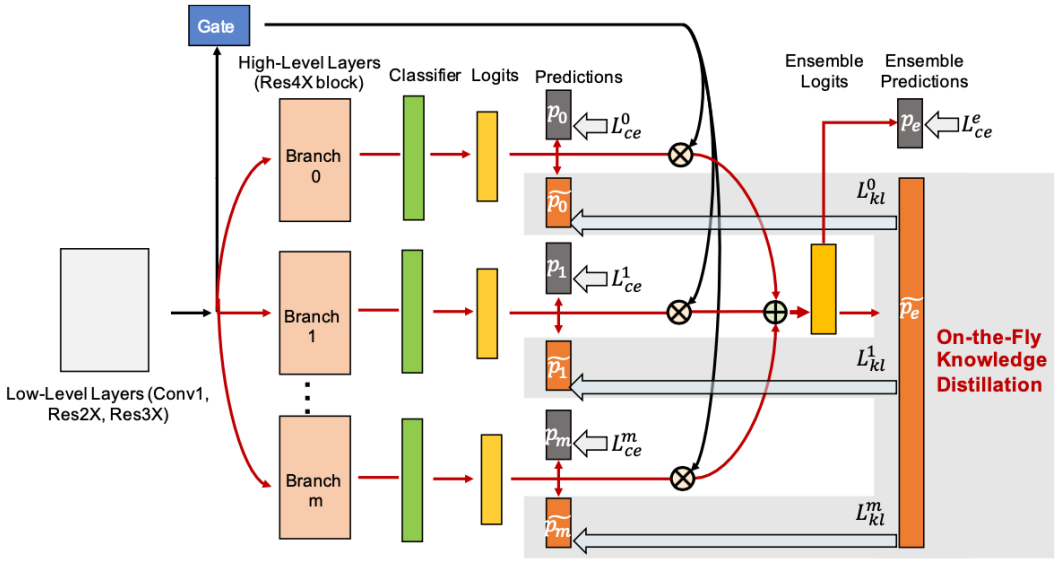
보다 빠르게 딥한 네트워크에 대한 최적화가 가능

다른 task에서 학습된 knowledge도 가져올 수 있음

Knowledge Distillation by On-the-Fly Native Ensemble (2018 NIPS)

지금까지의 KD 방법론들에 문제점이 있다고 주장
1. longer training process
2. extra computational cost and memory usage
3. complex multi-phase training procedure.

결국 복잡도와 비용 증가.
해당 논문은 이에 대해 간소화를 제안



각각의 (peer) Branch에서 얻은 최종 logit에 대한 가중치합이 Teacher(앙상블모델)의 logit을 대체할 수 있다.

$$\mathcal{L} = \sum_{i=0}^{m} \mathcal{L}_{\text{ce}}^{i} + \mathcal{L}_{\text{ce}}^{e} + T^2 * \mathcal{L}_{\text{kl}}$$

Loss를 총 3가지항으로 구성. 앙상블모델을 필요로 하지 않는다.

Born Again Neural Networks (Icml 2018)

Student -> Born Again Network(BAN)

Dense NET과 Res NET으로 실험하였다.
(똑같은 구조 / 같은 파리미터 수 다른 구조 / 아예 다른 모델)

KD Loss를 일종의 정규화항으로 생각하였다.

사람의 지능을 equence of teaching selves으로 생각하여

순차적으로(sequencially) knowledge transfer을 하였다.
모델 k는 k-1 모델을 teacer로 삼는다.

Student > Teacher : 청출어람의 성능

Table 1: Born Again DenseNet: test error on CIFAR100 for DenseNet of different depth and growth factor, the respective sequence of BAN-DenseNet, and the BAN-ensembles resulting from the sequence. Each BAN is trained from the label loss and cross-entropy with respect to the model at its left. We include the original teacher as a member of the ensemble for Ens*3 for 80-120 since we did not train a BAN-3 for this configuration.

Network	Parameters	Baseline	BAN-1	BAN-2	BAN-3	Ens*2	Ens*3
DenseNetBC-112-33	6.3 M	18.25	17.61	17.22	**16.59**	15.77	15.68
DenseNetBC-90-60	16.1 M	17.69	16.62	**16.44**	16.72	15.39	15.74
DenseNetBC-80-80	22.4 M	17.3	16.26	16.30	**15.5**	15.46	15.14
DenseNetBC-80-120	50.4 M	16.87	**16.13**	16.13	/	**15.13**	**14.9**

$$\min_{\theta_k} L(y, f(x, \theta_k) + L(f(x, arg\min_{\theta_{k-1}} L(y, f(x, \theta_{k-1}))), f(x, \theta_k))$$

5 Bonus 발표자의 잡다한 생각들

T[1]. Knowledge 추출이라고도 생각할 수 있는 것 같은데 더 일반적인 지능에 관련된 Knowledge를 추출할려면 어떻게 해야할까? 여러 개의 모델에서 한 번에 Knowledge를 뽑는 방법은 없을까? 여러가지 수치 예측 테스크와 각각의 테스크에 최적화된 신경망 모델들에서 “수치 예측 " 이라는 성능에 대한 Knowledge를 뽑을 수는 없을까

T[2]. 해당 논문에서는 앙상블 모델의 일반화 성능을 Knowledge로 보고 그 것을 전이시켰다.
근데, 앙상블을 이루는 각각의 모델에 대해 single net으로 전이시키고 그 결과를 앙상블하면 어떨지도 궁금하다.
비교해보고 싶다. Mnist로 한번 해볼까?

T[3]. 어떻게보면 흔히들 사용하는 머신러닝에서의 스텍킹과 비슷한 것 같다. 스테킹도 모델의 결과 또는 결과의 확률값 형태를 Target으로 학습시키니까, 해당논문에서는 원래의 타켓 (Hard Target)과 Soft Target을 같이 학습시킴으로써 (Loss가 2개의 항) 큰 성능 향상을 보였다고 하는데 머신러닝의 스테킹에서도 스테킹할 값과 True label을 동시에 학습시킬 수 없을까? 머신러닝에서는 loss를 손 볼 수 없나?
만약 가능하다면 저번 빅콘의 이탈문제에서도 동일한 데이터로 두 개의 task가 있어서 모델도 2개가 필요했는데 그 2개를 하나로 통합시키는 것은 불가능했을까 궁금하다

5 Bonus 발표자의 잡다한 생각들

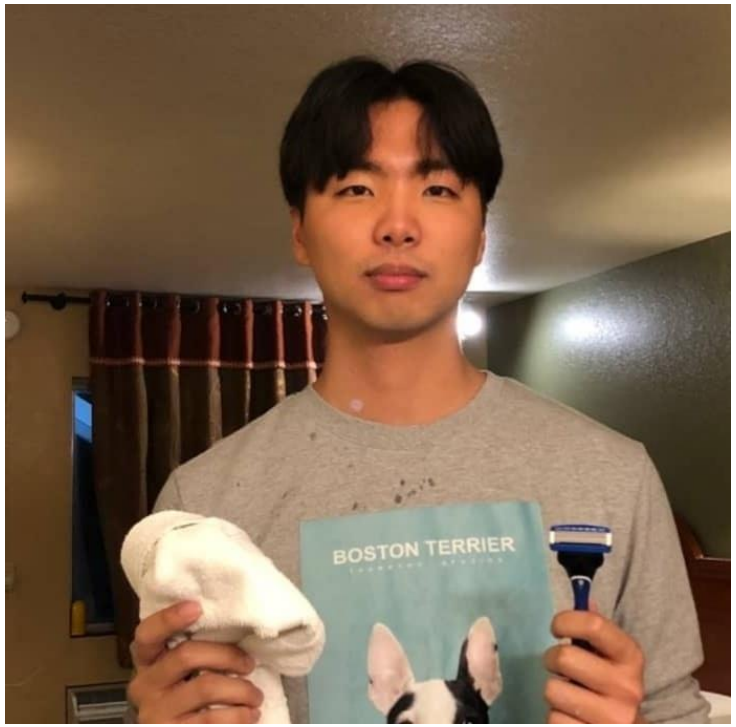
T[4]. VAE나 Gan도 파라미터가 매우 크던데 이에 대해서도 KD가 가능할지 궁금하다.

T[5]. 특정 모델의 결과값을 사용함으로써 해당 모델의 성능을 갖게 되는게
NN이 아닌 다른 방법론을 NN에 이식하는 방법이 될 수 있는지 궁금하다. 예를 들어 DT의
결과로 NN을 학습시켜서 DT와 비슷한 결과를 내게 만드는게 가능할까?

Recommender System에서도 가능할지 궁금하다.
SAR의 결과를 사용해서 NN이 비슷한 성능을 갖게끔 유도하는게 가능할까?

관련 자료를 찾아보니 이미 Ranking Distillation이라는 논문이 존재한다.
정확하게 어떤 방식인지는 추후 읽어야겠다.

T[6]. MNIST에 대해 돌려볼 수 있는 KD 코드를 구하던지 구현해보던지
다음 발표자 발표 때까지 가져오도록 하겠습니다!



Reference

https://www.youtube.com/watch?v=tOItokBZSfU
(youtube: Choung young jae [PR-009: Distilling the Knowledge in a Neural Network (Slide: English, Speaking: Korean)
논문리뷰 영상

https://tv.naver.com/v/5082046/list/150913
(네이버 테크 톡: **Relational Knowledge Distillation)**
(ppt:https://www.slideshare.net/NaverEngineering/relational-knowledge-distillation?from_action=save)

관련 연구들
https://kdst.tistory.com/
https://jayhey.github.io/deep%20learning/2018/01/27/BAN/
https://sysyn.tistory.com/24
https://seongkyun.github.io/papers/2019/04/03/a_gift_from_distillation/
https://creamnuts.github.io/paper/Fitnet/
https://m.blog.naver.com/PostView.nhn?blogId=hist0134&logNo=220916802890&proxyReferer=https:%2F%2Fwww.google.c
https://light-tree.tistory.com/195

Q&A