Knowledge Distillation

Jinyang Guo（郭晋阳）
jinyangguo@buaa.edu.cn

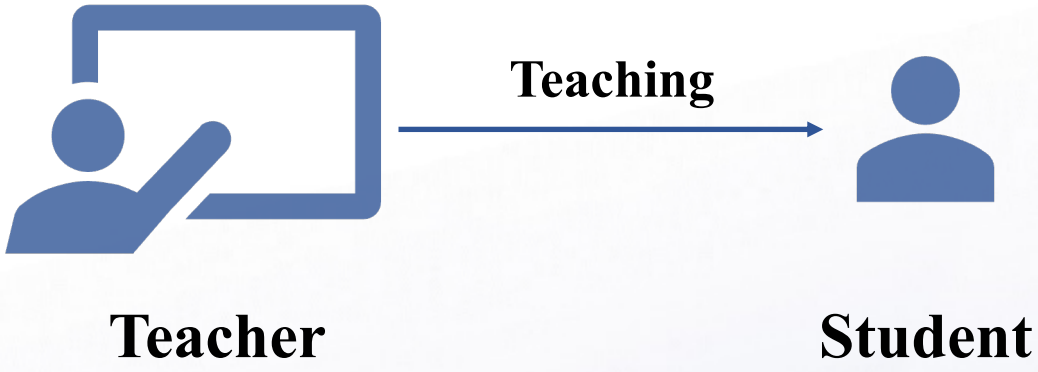
Last lecture

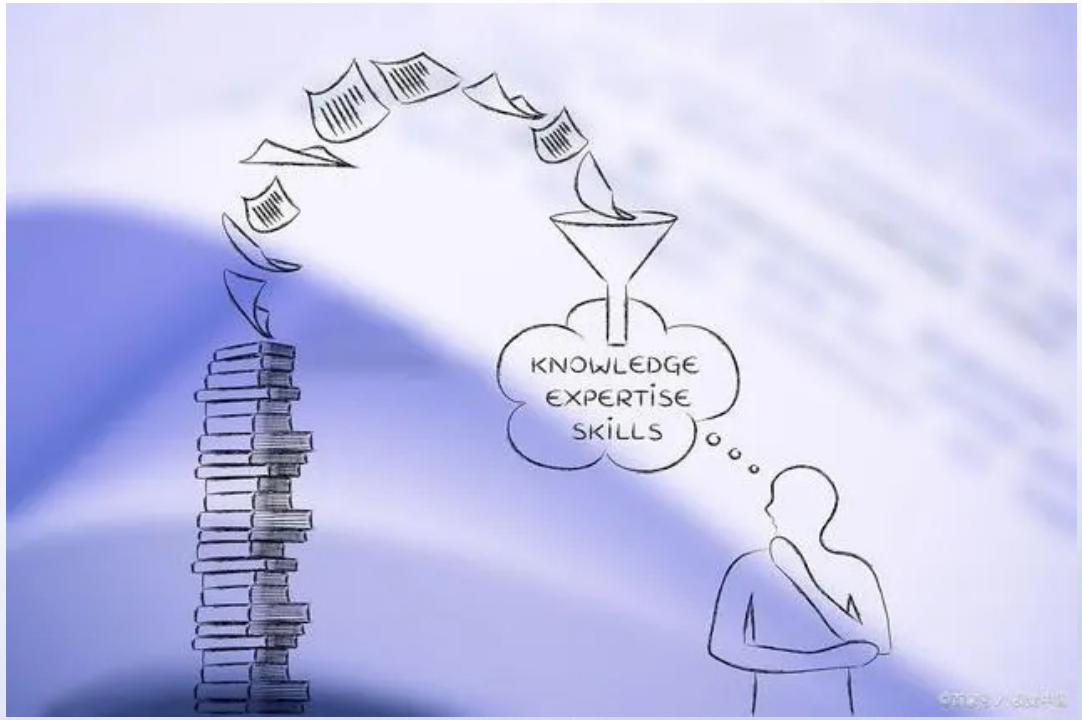
- Network pruning
- Make network sparse

目录

Introduction

- **Human knowledge can be transferred from teachers to students through teaching.**

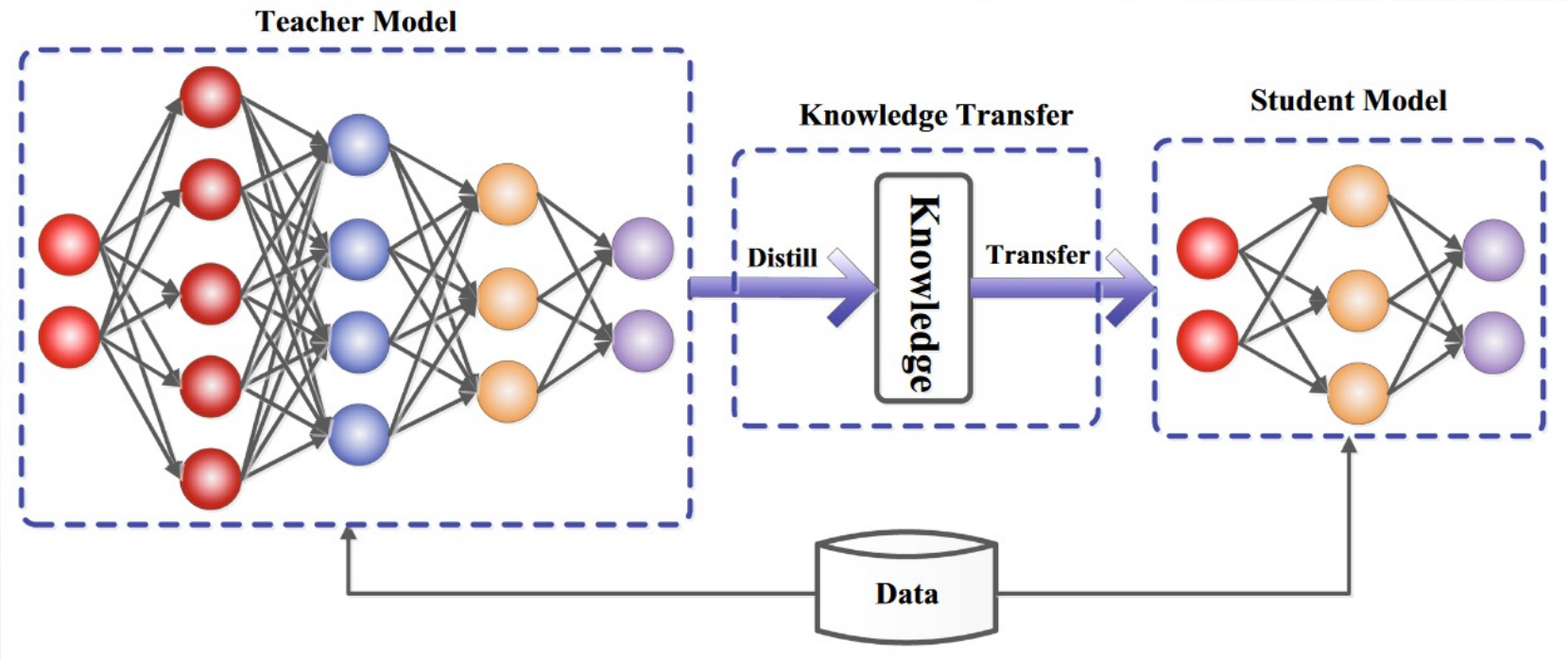




Introduction

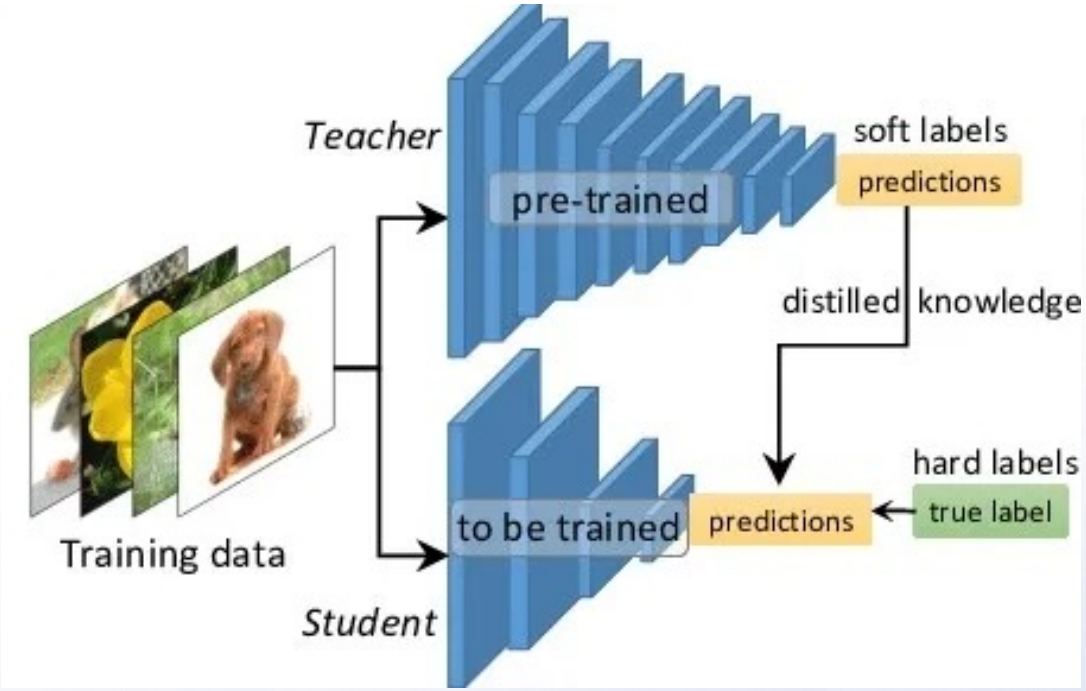
➢ Knowledge Distillation

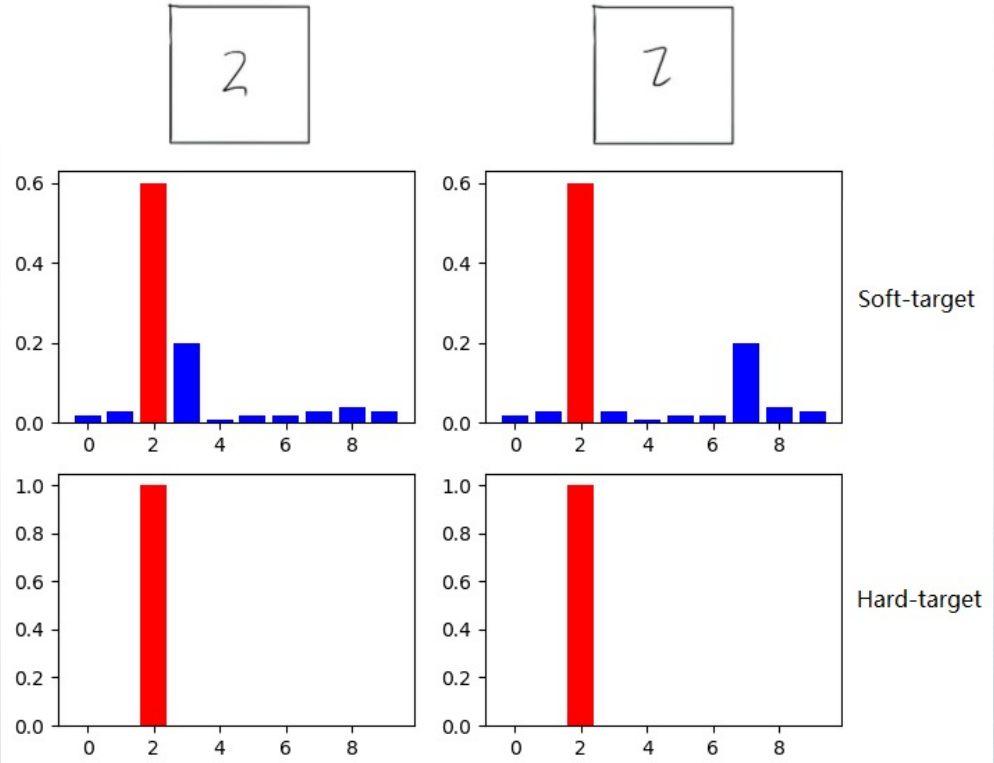
- ➢ transferring the knowledge from a large model (teacher) to a smaller model (student)
- ➢ a form of model compression



Introduction

➢**Soft target in knowledge distillation**

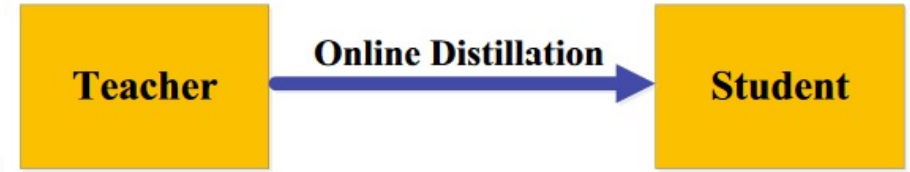




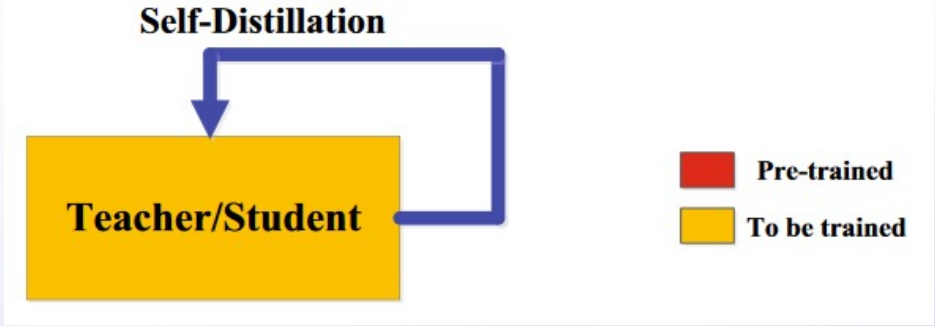
Introduction

➢ **Three types of knowledge distillation**

Offline Distillation



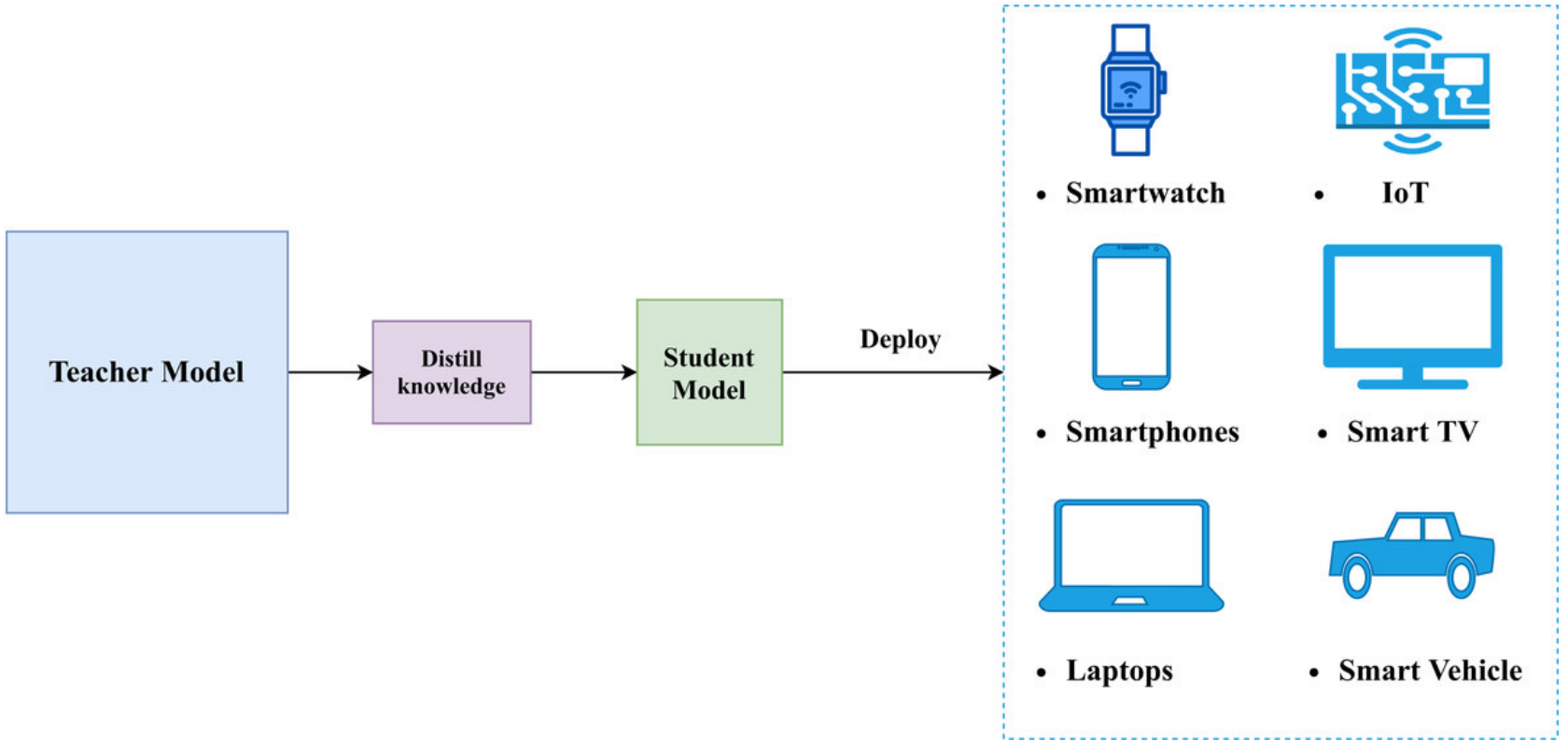
Online Distillation



Self-Distillation

Introduction

➢ Applications of Knowledge Distillation

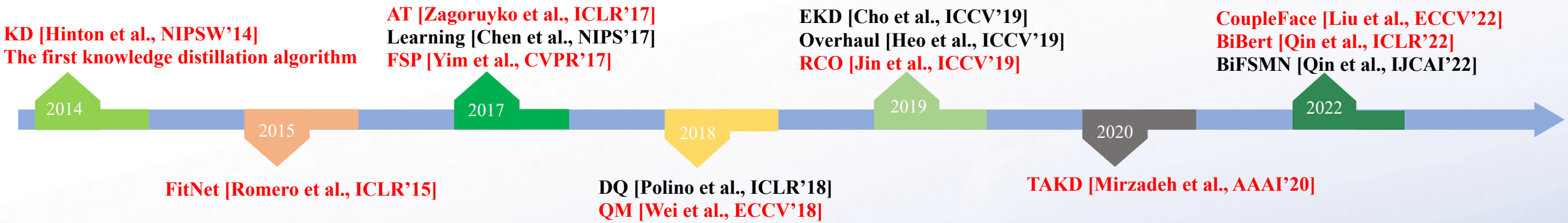


Use cases for knowledge distillation to deploy deep learning models on small devices with limited resources

目录

Knowledge Distillation: History

Knowledge distillation has gradually gained attention since 2015.



Knowledge Distillation

Definition of knowledge

- Distilling the Knowledge in a Neural Network, NIPSW'14
- Fitnets: Hints for Thin Deep Nets, ICLR'15
- Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer, ICLR'17
- A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning, CVPR'17

Transfer of knowledge

- Knowledge Distillation via Route Constrained Optimization, ICCV'19
- Improved Knowledge Distillation via Teacher Assistant, AAAI'20

Application of KD

- Quantization Mimic: Towards Very Tiny CNN for Object Detection, ECCV'18
- CoupleFace: Relation Matters for Face Recognition Distillation, ECCV'22
- BiBERT: Accurate Fully Binarized BERT, ICLR'22

目录

Definition of knowledge

Distilling the Knowledge in a Neural Network

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

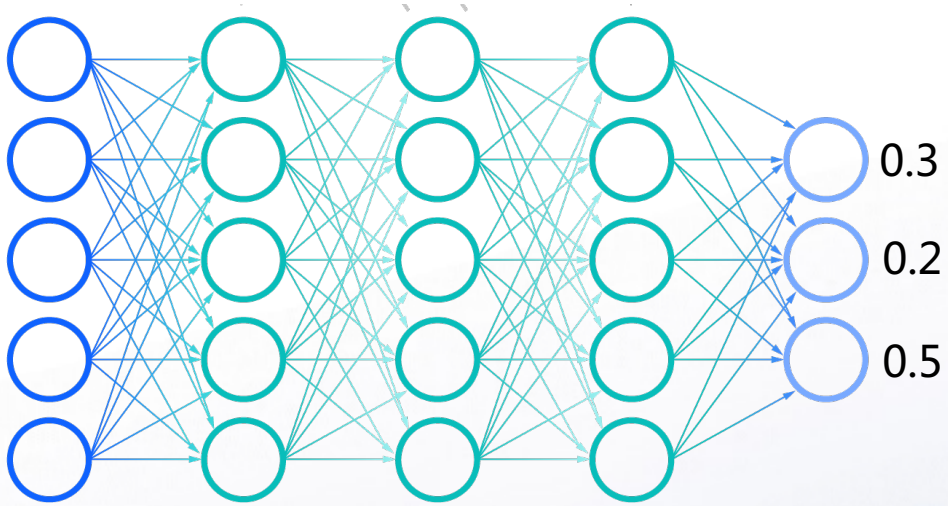
Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

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

Definition of knowledge

Problem: information in incorrect categories is ignored, during the training with cross entropy

Method: define the outputs of large model as knowledge, imitating the teacher's output distribution



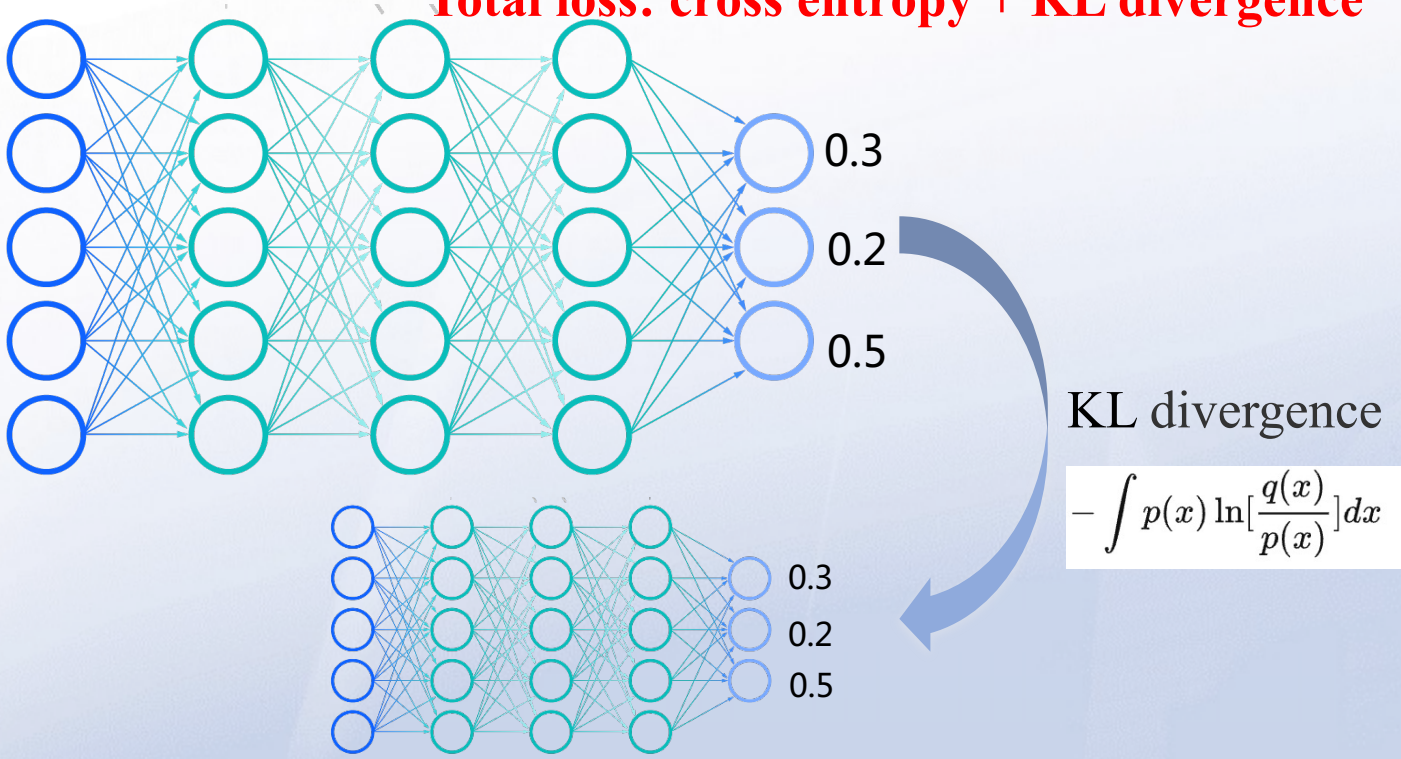
Cross entropy loss：

-(0xlog0.3 + 0xlog0.2 + 1xlog0.5)

Missing incorrect category prediction information

Add the KL divergence between the outputs of teacher and student

Total loss: cross entropy + KL divergence



Definition of knowledge

Result: 3x parameter compression on CIFAR10, 1.81% accuracy drop

Teacher Model	Student Model	Teacher Accuracy	directly training student	Knowledge Distillation
WRN-16-2, 0.7M	WRN-16-1, 0.2M	93.69%	91.23%	92.59%
WRN-40-1, 0.6M	WRN-16-1, 0.2M	93.42%	91.23%	91.61%
WRN-40-2, 2.2M	WRN-16-2, 0.7M	94.77%	93.69%	93.92%

Definition of knowledge

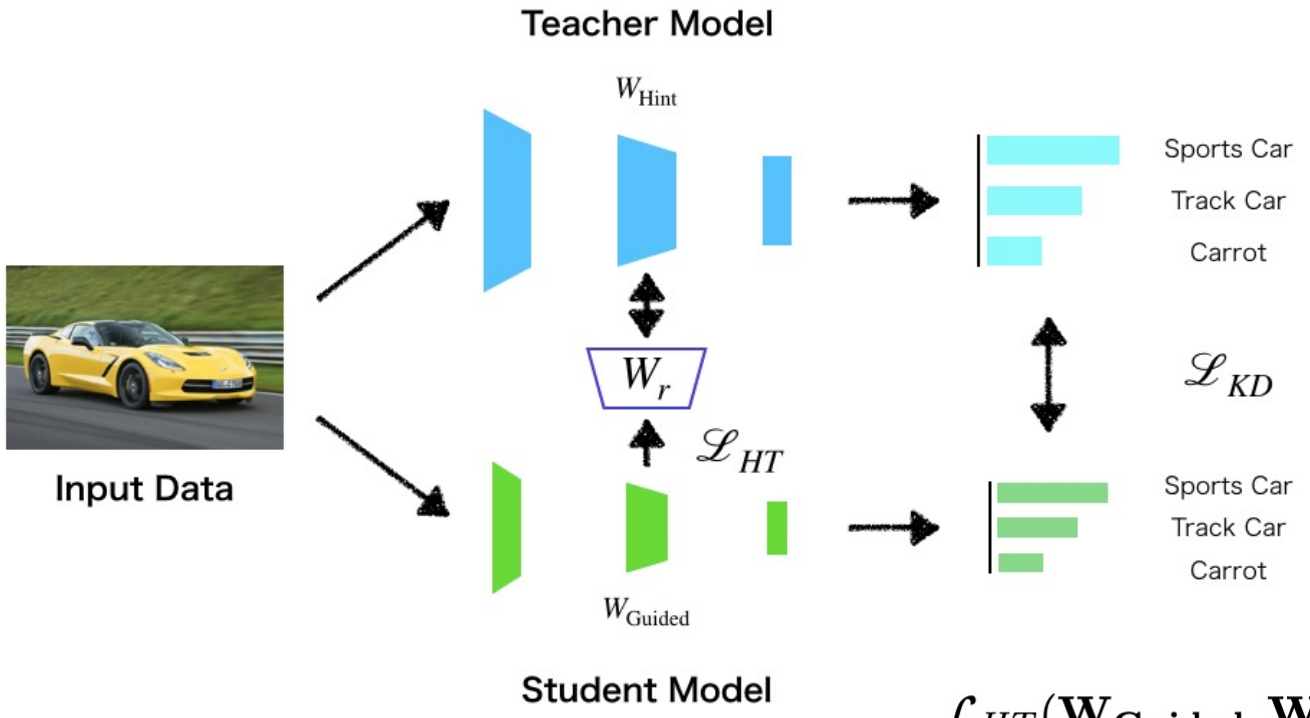
FITNETS: HINTS FOR THIN DEEP NETS

Adriana Romero[1], Nicolas Ballas[2], Samira Ebrahimi Kahou[3], Antoine Chassang[2], Carlo Gatta[4] & Yoshua Bengio[2]†
[1]Universitat de Barcelona, Barcelona, Spain.
[2]Université de Montréal, Montréal, Québec, Canada. †CIFAR Senior Fellow.
[3]École Polytechnique de Montréal, Montréal, Québec, Canada.
[4]Centre de Visió per Computador, Bellaterra, Spain.

Definition of knowledge

Problem: the KD method ignores the information of the middle layer in the network

Method: define the features of middle layer as knowledge



Total loss: cross entropy + the L2 distance of middle features + (other knowledge loss)

$$\mathcal{L}_{HT}(\mathbf{W_{Guided}}, \mathbf{W_r}) = \frac{1}{2}||u_h(\mathbf{x}; \mathbf{W_{Hint}}) - r(v_g(\mathbf{x}; \mathbf{W_{Guided}}); \mathbf{W_r})||^2,$$

Fitnets: Hints for Thin Deep Nets

Definition of knowledge

Result: improvements on multiple datasets

Network	# layers	# params	# mult	Acc	Speed-up	Compression rate
Teacher	5	~9M	~725M	90.18%	1	1
FitNet 1	11	~250K	~30M	89.01%	**13.36**	**36**
FitNet 2	11	~862K	~108M	91.06%	4.64	10.44
FitNet 3	13	~1.6M	~392M	91.10%	1.37	5.62
FitNet 4	19	~2.5M	~382M	**91.61%**	1.52	3.60

Table 5: Accuracy/Speed Trade-off on CIFAR-10.

Algorithm	# params	Accuracy
Compression		
FitNet	~2.5M	**91.61%**
Teacher	~9M	90.18%
Mimic single	~54M	84.6%
Mimic single	~70M	84.9%
Mimic ensemble	~70M	85.8%
State-of-the-art methods		
Maxout		90.65%
Network in Network		91.2%
Deeply-Supervised Networks		**91.78%**
Deeply-Supervised Networks (19)		88.2%

Table 1: Accuracy on CIFAR-10

Algorithm	# params	Accuracy
Compression		
FitNet	~2.5M	**64.96%**
Teacher	~9M	63.54%
State-of-the-art methods		
Maxout		61.43%
Network in Network		64.32%
Deeply-Supervised Networks		**65.43%**

Table 2: Accuracy on CIFAR-100

4x computation reduction,
0.12% accuracy drop

Definition of knowledge

PAYING MORE ATTENTION TO ATTENTION: IMPROVING THE PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORKS VIA ATTENTION TRANSFER

Sergey Zagoruyko, Nikos Komodakis
Université Paris-Est, École des Ponts ParisTech
Paris, France
{sergey.zagoruyko,nikos.komodakis}@enpc.fr

Definition of knowledge

Problem: the different parts in a picture have different importance

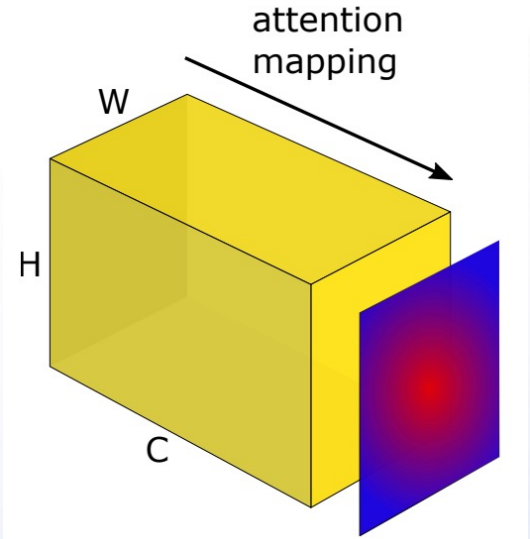
Method: define the attention as knowledge

attention map

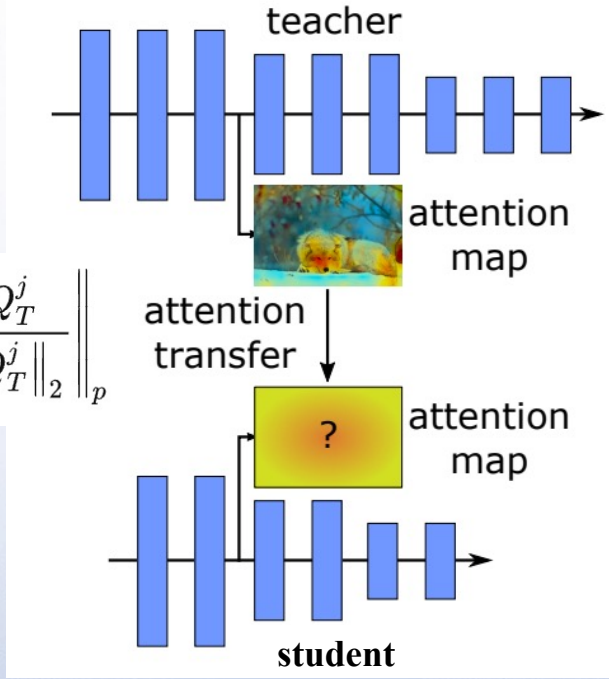
The central part of this picture has more information.

The network needs to pay more attention to the main part.

Total loss: cross entropy + L2 distance of attention maps + (other knowledge loss)

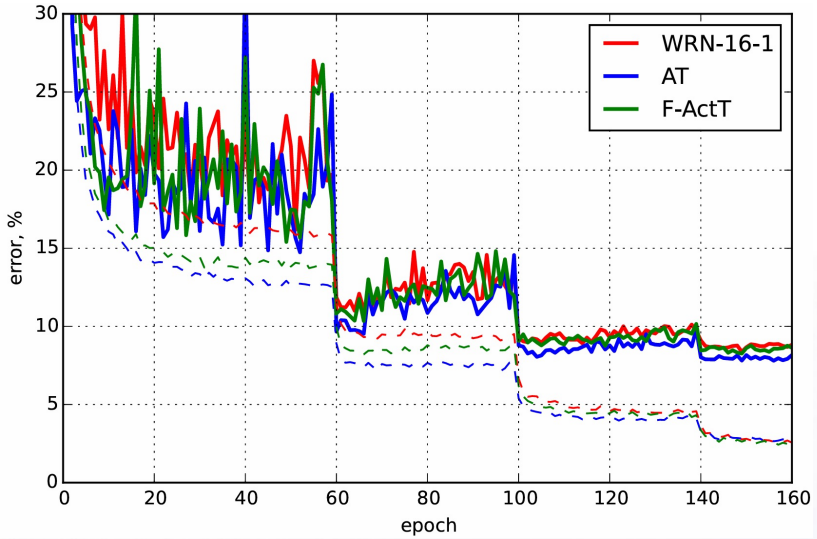


$$\frac{\beta}{2}\sum_{j\in\mathcal{I}}\left\|\frac{Q_S^j}{\left\|Q_S^j\right\|_2}-\frac{Q_T^j}{\left\|Q_T^j\right\|_2}\right\|_p$$



Definition of knowledge

Result: 3x parameter compression, 1.67% accuracy drop



accelerate the convergence speed and improve final accuracy

Teacher	Student	Teacher	directly training student	AT	AT+KD
WRN-16-2, 0.7M	WRN-16-1, 0.2M	93.69%	91.23%	92.07%	94.49%
WRN-40-1, 0.6M	WRN-16-1, 0.2M	93.42%	91.23%	91.75%	91.99%
WRN-40-2, 2.2M	WRN-16-2, 0.7M	94.77%	93.69%	94.15%	94.29%

Definition of knowledge

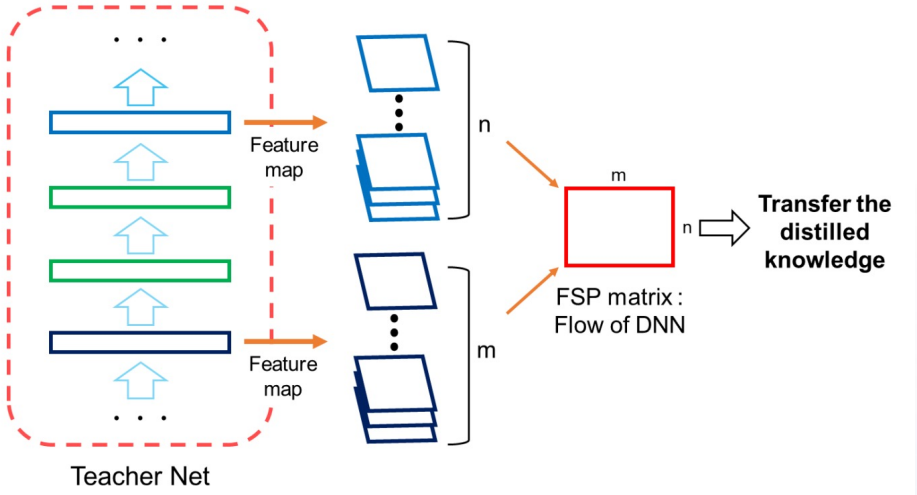
A Gift from Knowledge Distillation:
Fast Optimization, Network Minimization and Transfer Learning

Junho Yim[1] Donggyu Joo[1] Jihoon Bae[2] Junmo Kim[1]
[1]School of Electrical Engineering, KAIST, South Korea
[2]Electronics and Telecommunications Research Institute
{junho.yim, jdg105, junmo.kim}@kaist.ac.kr
{baejh}@etri.re.kr

Definition of knowledge

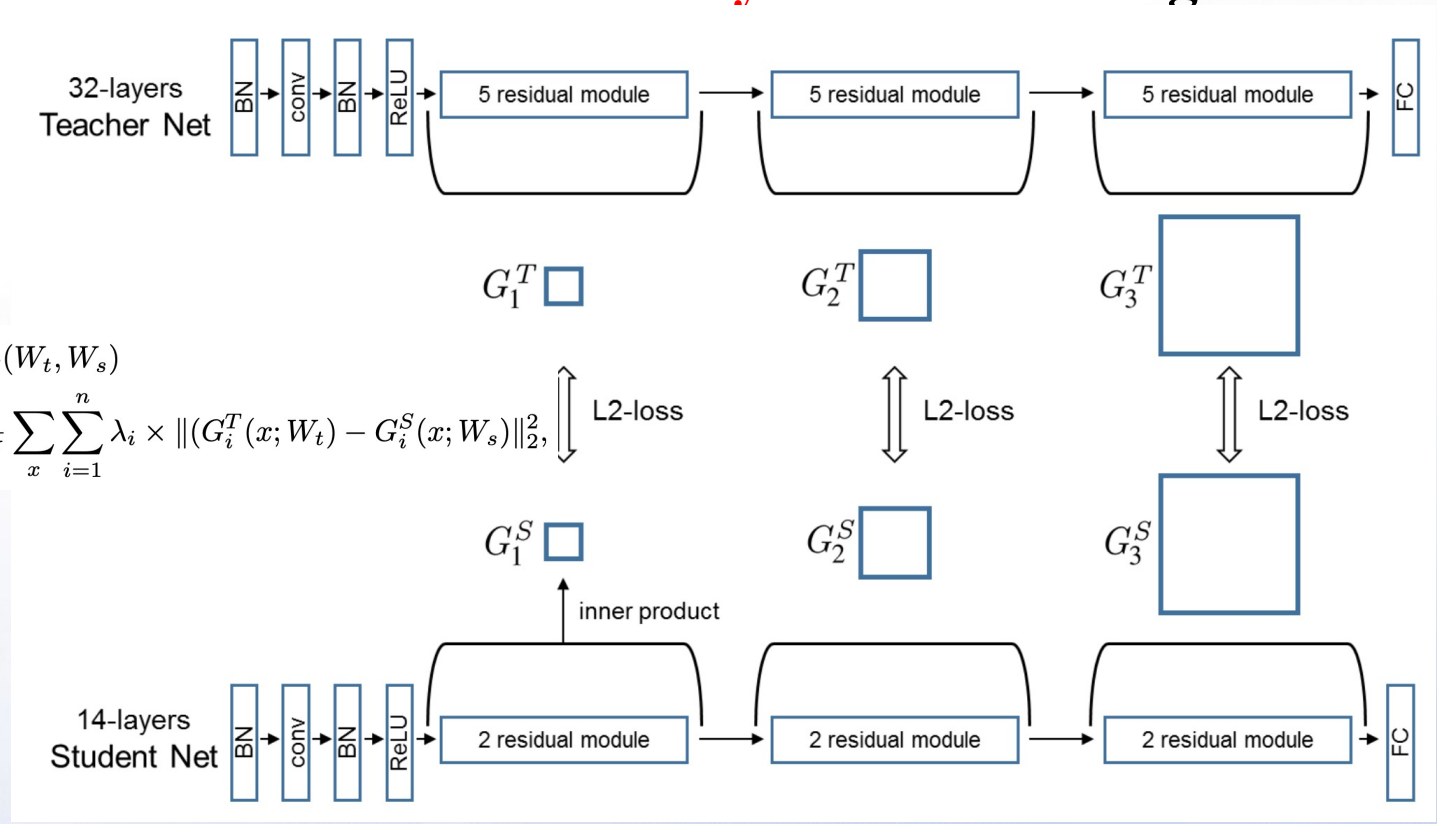
Problem: previous methods ignore the relationship between network layers

Method: define the relationship between features from two layers as knowledge



Inner product of features (FSP matrix)

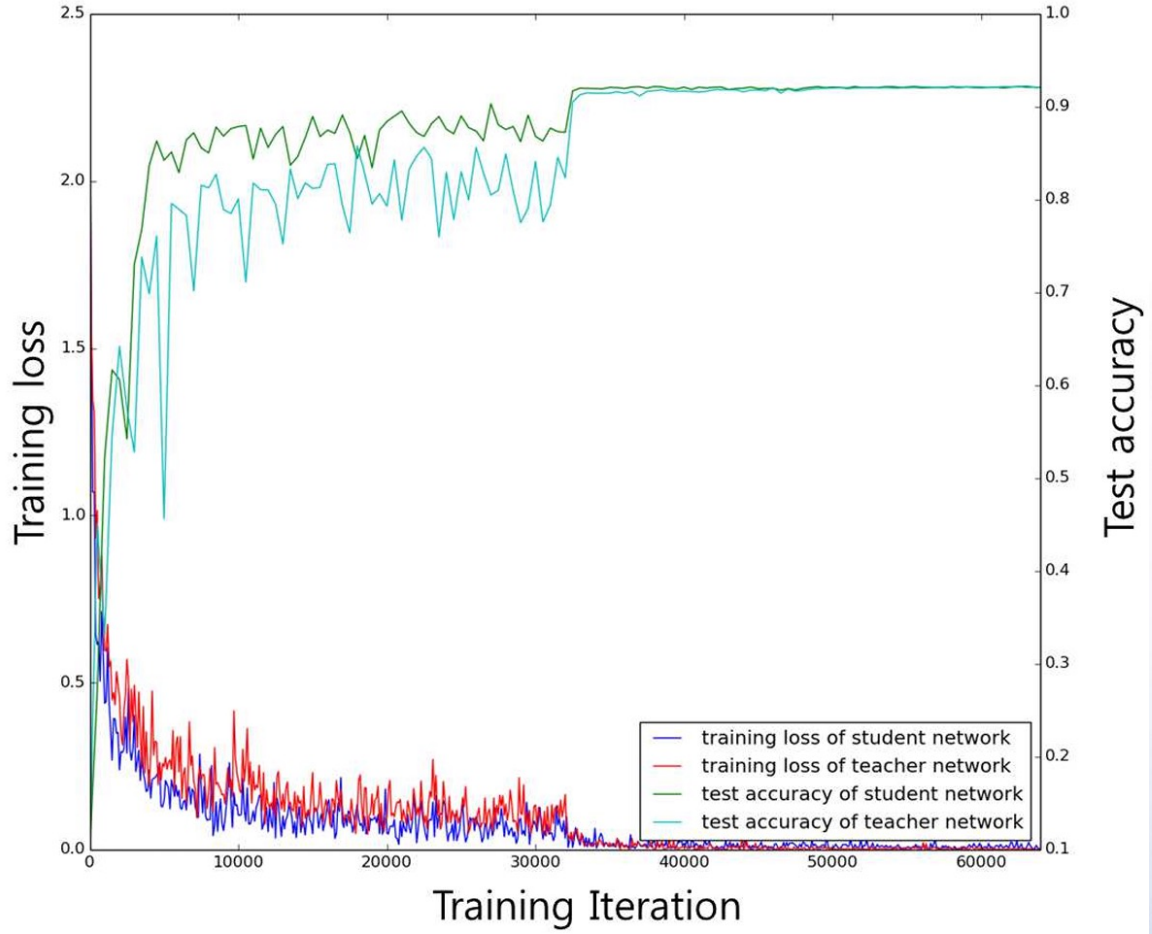
$$L_{FSP}(W_t, W_s) = \frac{1}{N}\sum_{x}\sum_{i=1}^{n}\lambda_i \times \|(G_i^T(x;W_t) - G_i^S(x;W_s)\|_2^2,$$



Total loss: cross entropy + L2 distance of FSP + (other knowledge loss)

Definition of knowledge

Result: 3.21% accuracy drop from ResNet-26 to ResNet-8 on CIFAR10



	Accuracy
Teacher-original	91.91
Student-original	87.91
FitNet [20]	88.57
Proposed Method	88.70

	Net 1	Net 2	Net 3	*Avg*	*Ensemble*	*#Iter*
Teacher	91.61	91.56	92.09	91.75	93.48	192k
Teacher *	90.47	90.83	90.62	90.64	92.6	63k
Teacher ‡	91.84	92.26	92.01	92.04	92.71	63k
1 loss FitNet [20]*	91.69	91.85	91.64	91.72	92.98	98k
3 loss FitNet [20]*	88.90	89.35	89.02	89.09	89.92	98k
Student *	92.28	92.08	92.07	92.14	93.26	84k
Student *†	92.28	91.89	92.08	92.08	93.67	126k

Definition of knowledge

SCOTT: Self-Consistent Chain-of-Thought Distillation

Peifeng Wang[1]*, **Zhengyang Wang**[2], **Zheng Li**[2], **Yifan Gao**[2], **Bing Yin**[2], **Xiang Ren**[1]
[1]Department of Computer Science, University of Southern California, [2]Amazon.com Inc
{peifengw,xiangren}@usc.edu,
{zhengywa,amzzhe,yifangao,alexbyin}@amazon.com

Definition of knowledge

Problem: chain-of-thought (CoT) prompting makes LMs to explain their multi-step reasoning, but the performance gains are typically only observed with sufficiently large language models

Method: rationales generated from a large language model (teacher) that are more grounded by the gold answers, and trains the student language model using a counterfactual reasoning objective

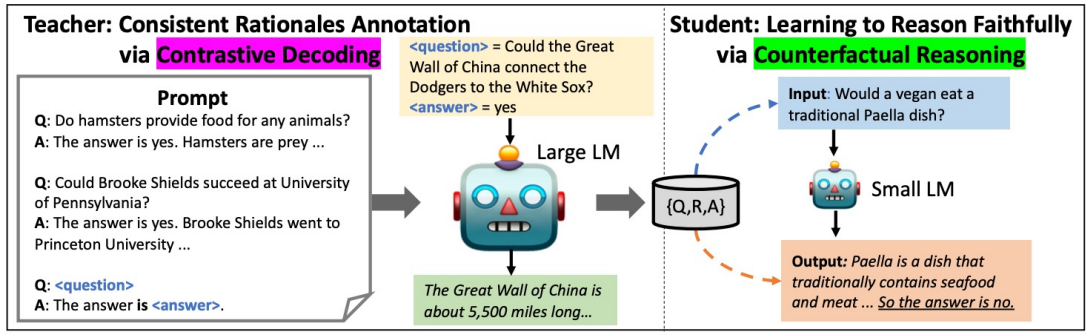
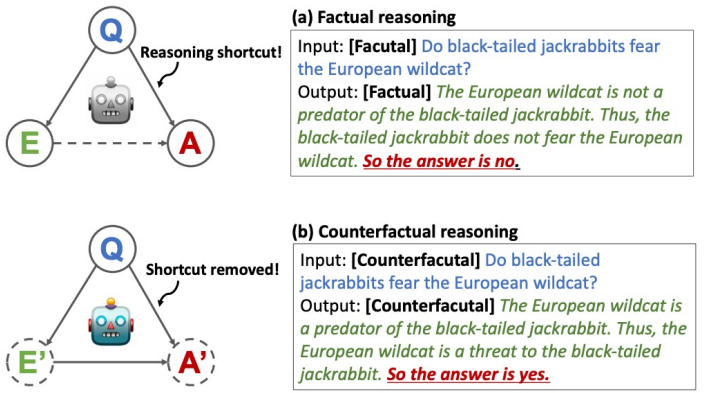


Figure 2: **Overview of our knowledge distillation framework for faithful reasoning.** (a) Teacher: A large LM prompted to generate a consistent rationale given a question and the gold answer in the training set via contrastive decoding. (b) Student: A small LM fine-tuned to generate a rationale and then answer via counterfactual reasoning.

Only search rationale tokens that are more plausible for the gold answer by contrastive decoding

$$G(t_i|a^*) = \log \frac{P(t_i|p, q, a^*, t_{<i})}{P(t_i|p, q, a', t_{<i})}.$$

appending the keyword [Factual] or [Counterfactual]



Sum the loss and train jointly

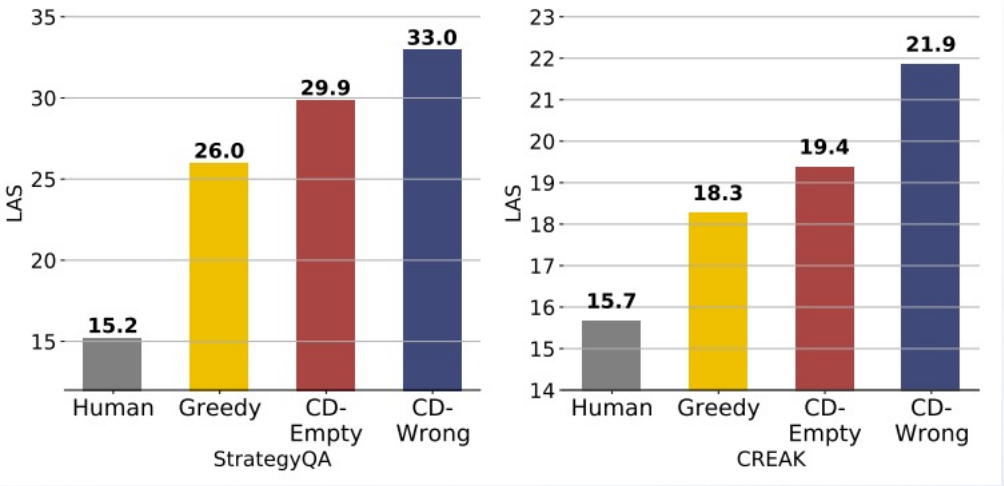
$$\mathcal{L}_{factual} = -\sum_i \log P(t_i|q, t_{<i}),$$

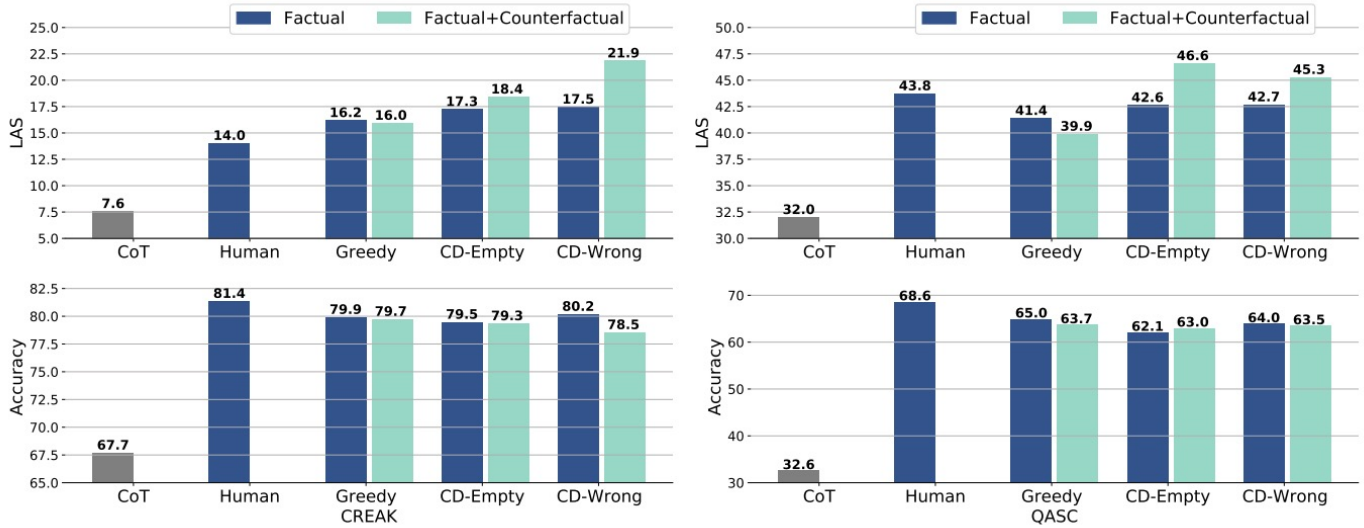
$$\mathcal{L}_{counterfactual} = -\sum_i \log P(t_i|q, r', t_{<i}).$$

Definition of knowledge

Result: generate CoT rationales that are more faithful

Teacher Model	Grammaticality	New Info	Supports Answer
Greedy	**0.99**	0.65	0.48
Contrast.-Empty	0.97	0.77	0.58
Contrast.-Wrong	0.97	**0.82**	**0.63**





{Greedy, CD-Empty, CD-Wrong} refer respectively to using greedy decoding, contrastive decoding with empty/wrong answer to obtain rationale tokens from the teacher.

Definition of knowledge

ON-POLICY DISTILLATION OF LANGUAGE MODELS: LEARNING FROM SELF-GENERATED MISTAKES

Rishabh Agarwal[12*] **Nino Vieillard**[1*] **Yongchao Zhou**[13]

Piotr Stanczyk[1†] **Sabela Ramos**[1†] **Matthieu Geist**[1] **Olivier Bachem**[1]

[1]**Google DeepMind** [2]**Mila** [3]**University of Toronto**

Definition of knowledge

Problem: current KD methods for auto-regressive sequence models suffer from distribution mismatch

Method: introduce **on-policy Knowledge Distillation**，which trains the student on its self-generated output sequences

$$L_{OD}(\theta) := \mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim p_S(\cdot|x)}\left[\mathcal{D}_{KL}\left(p_T \,\|\, p_S^{\theta}\right)(y \mid x)\right]\right]$$

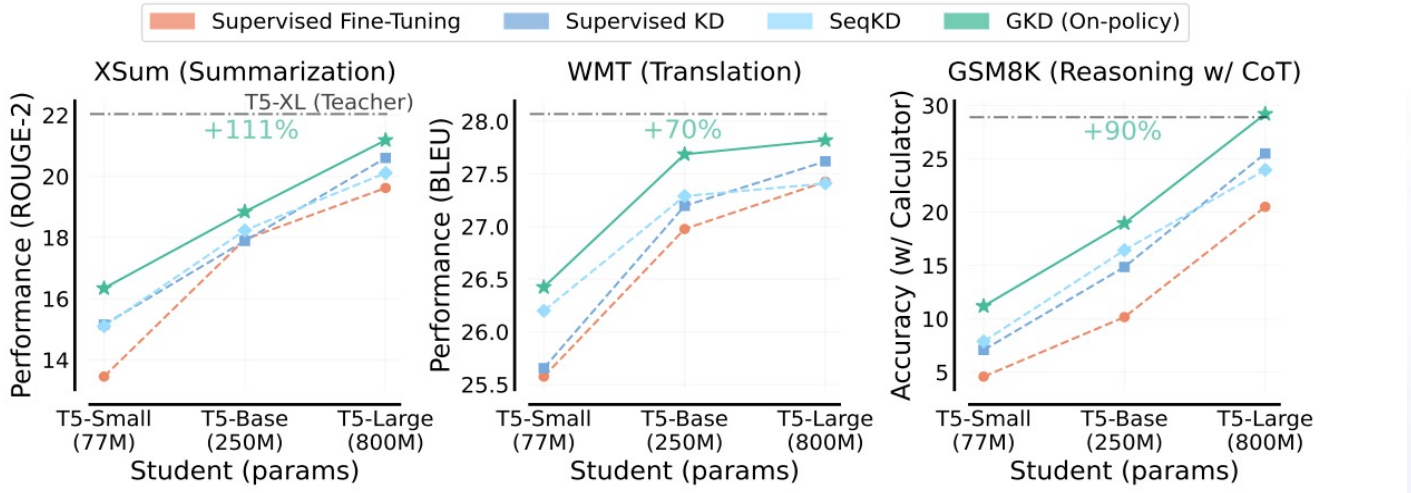
student receives token-specific feedback from the teacher's logits

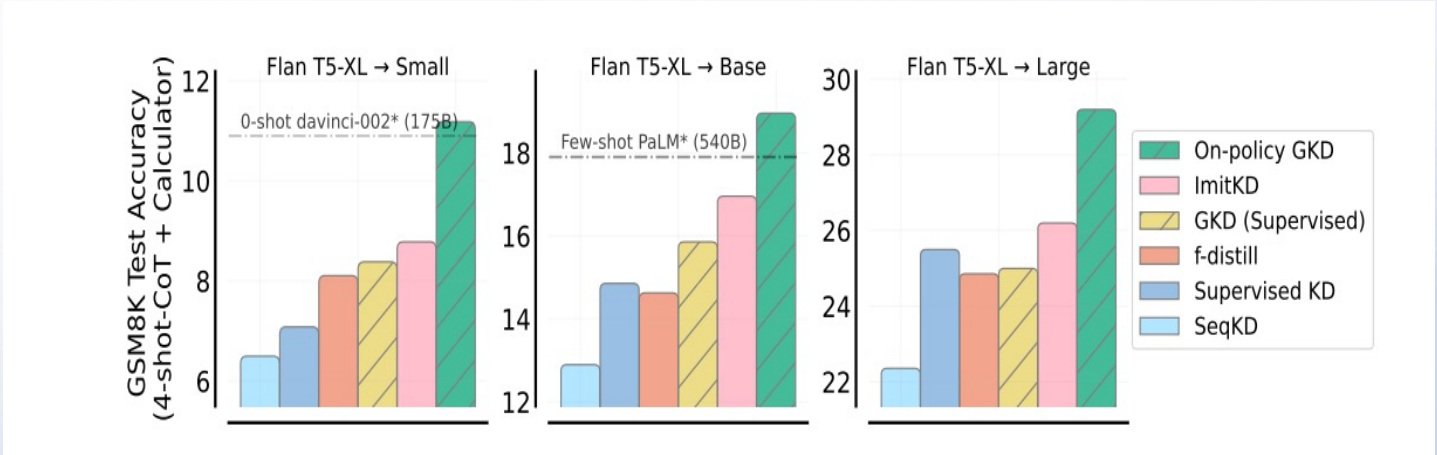
$$L_{GKD}(\theta) := (1-\lambda)\mathbb{E}_{(x,y)\sim(X,Y)}\left[\mathcal{D}\left(p_T \,\|\, p_S^{\theta}\right)(y \mid x)\right] + \lambda\mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim p_S(\cdot|x)}\left[\mathcal{D}\left(p_T \,\|\, p_S^{\theta}\right)(y \mid x)\right]\right]$$

use a mixture of teacher-generated and on-policy student-generated sequences for training

Definition of knowledge

Result: **on-policy GKD outperforms commonly-used KD approaches**





目录

Transfer of knowledge

Knowledge Distillation via Route Constrained Optimization

Xiao Jin[1*], Baoyun Peng[2*], Yichao Wu[1], Yu Liu[3], Jiaheng Liu[4],
Ding Liang[1], Junjie Yan[1], Xiaolin Hu[5]

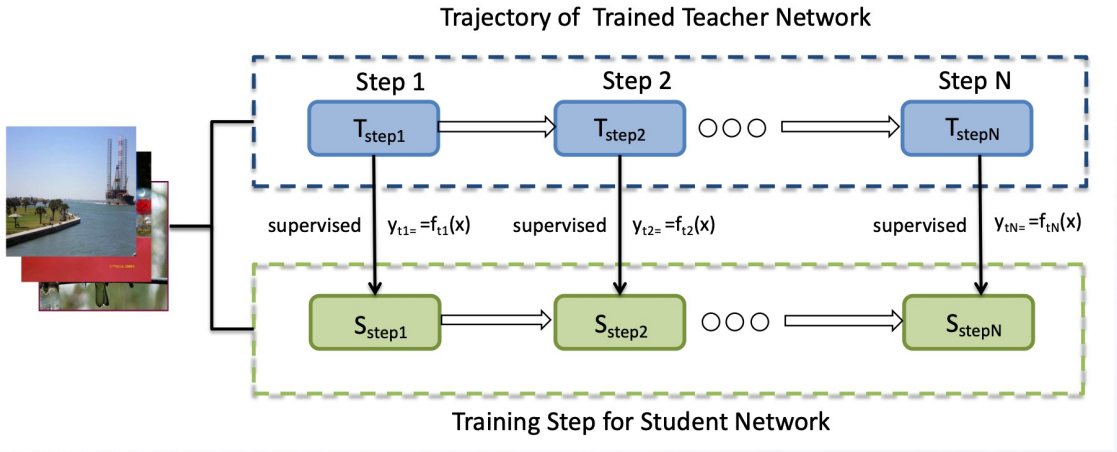
[1] SenseTime Group Limited [2] National University of Defense Technology
[3] Chinese University of Hong Kong [4] Beihang University [5] Tsinghua University

jinxiaocuhk@gmail.com, pengbaoyun13@nudt.edu.cn, yuliu@ee.cuhk.edu.hk, liujiaheng@buaa.edu.cn
{wuyichao,liangding,yanjunjie}@sensetime.com, xlhu@mail.tsinghua.edu.cn

Transfer of knowledge

Problem: there is **a significant performance gap** between the pre-trained teacher moel and the student model

Method: simultaneous learning between teachers and students



Algorithm 2 Greedy Search

Require: Student network with parameter W_s after mimicking former i_{th} anchor point C_i, where $i \in \{1, 2, ..., N\}$, relaxation factor δ.
compute KL divergence $\mathcal{H}_i$
$j = i + 1$
while $j < N$ **do**
  compute $\mathcal{H}_j$ on validation set
  compute $r_{ij} = \frac{\mathcal{H}_j - \mathcal{H}_i}{\mathcal{H}_i}$
  if $r_{ij} > \delta$ **then**
    Return j-1;
  end if
  $j = j + 1$
end while
Return N;

imitate the teacher network's checkpoint step by step

avoid significant gap between teacher and student

greedy search to find the optimal checkpoint

Transfer of knowledge

Result: 68.21% accuracy from ResNet-50 to MobileNetV2 on ImageNet

Method	Network	top-1	top-5
Teacher-Softmax	ResNet-50	75.49	92.48
Student-Softmax	MobileNetV2	64.2	85.4
Student-KD	MobileNetV2	66.75	87.3
Student-RCO	MobileNetV2	**68.21**	**88.04**

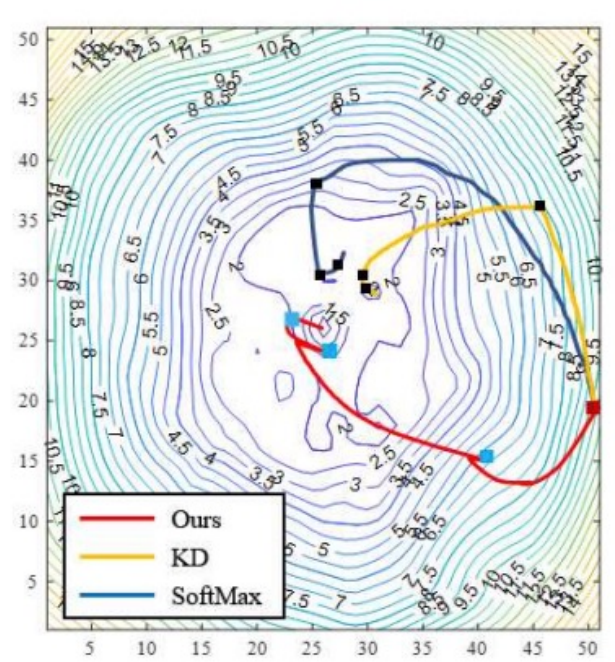
Table 4: Results on ImageNet

	top-1 @ distractor size					
Method	e^1	e^2	e^3	e^4	e^5	e^6
Teacher	99.78	99.67	99.38	98.86	97.70	94.83
Softmax	99.20	96.37	91.49	84.45	75.60	65.91
FitNet	99.62	98.80	96.83	93.53	88.28	81.02
RCO	**99.69**	**99.01**	**97.52**	**94.84**	**90.55**	**84.3**

Table 5: Results on MegaFace

68.21% accuracy, 13x computation reduction on ImageNet

higher performance than FitNet on MegaFace



better training trajectories

Transfer of knowledge

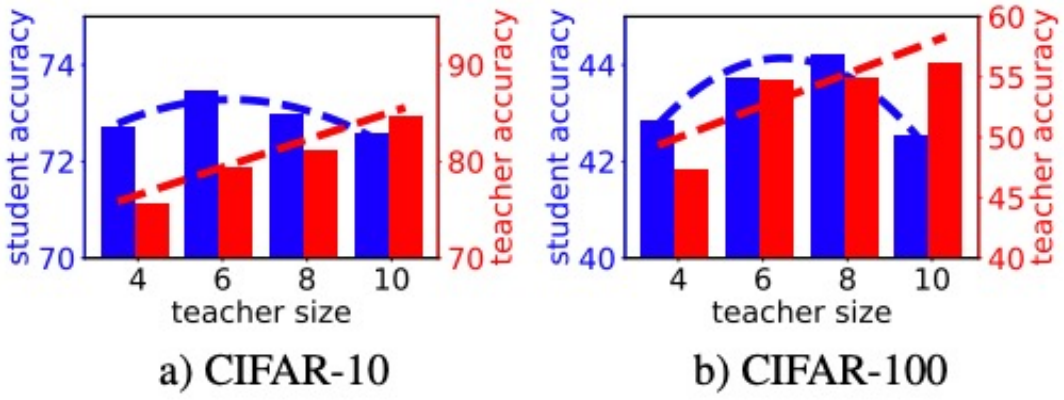
Improved Knowledge Distillation via Teacher Assistant

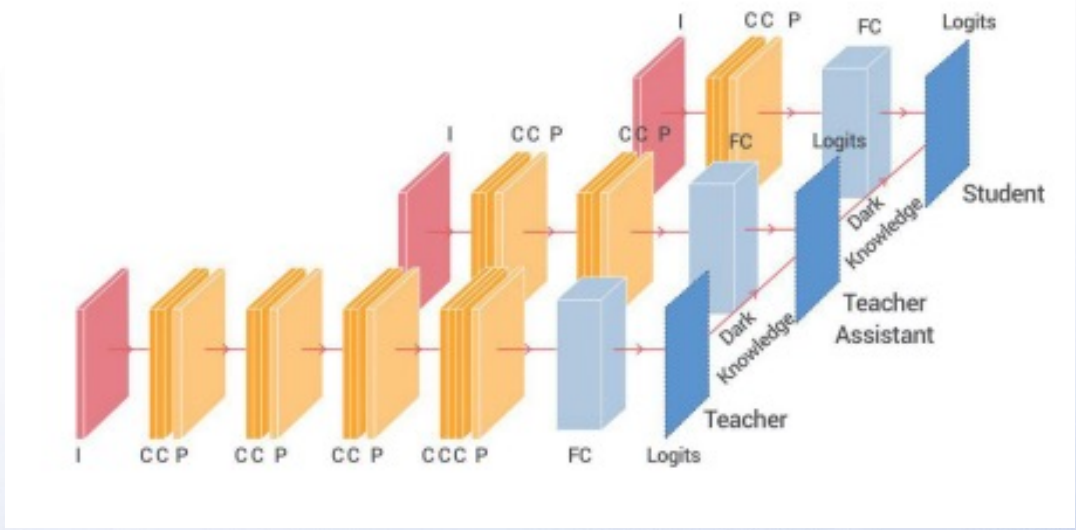
Seyed Iman Mirzadeh,[*1] Mehrdad Farajtabar,[*2] Ang Li,[2]
Nir Levine,[2] Akihiro Matsukawa,[†3] Hassan Ghasemzadeh[1]
[1]Washington State University, WA, USA
[2]DeepMind, CA, USA
[3]D.E. Shaw, NY, USA
[1]{seyediman.mirzadeh, hassan.ghasemzadeh}@wsu.edu
[2]{farajtabar, anglili, nirlevine}@google.com
[3]akihiro.matsukawa@gmail.com

Transfer of knowledge

Problem: the **significant difference** between teacher and student leads to a decrease in student's performance

Method: introduce **teacher assistant** network, transferring knowledge **gradually** to student





When the gap between teacher and student is too large, the performance of student will decrease.

introduce **teacher assistant** network to assist **knowledge transfer**

Transfer of knowledge

Result: higher performance on multiple datasets

How to determine the size of teacher assistant model?

Empirical setting: when training from scratch, the teaching assistant model falls between the teacher and student.

Table 2: Student's accuracy given varied TA sizes for (S=2, T=10)

Model	Dataset	TA=8	TA=6	TA=4
CNN	CIFAR-10	72.75	73.15	73.51
	CIFAR-100	44.28	44.57	44.92

Table 3: Student's accuracy given varied TA sizes for (S=8, T=110)

Model	Dataset	TA=56	TA=32	TA=20	TA=14
ResNet	CIFAR-10	88.70	88.73	88.90	88.98
	CIFAR-100	61.47	61.55	61.82	61.5

Table 1: Comparison on evaluation accuracy between our method (TAKD) and baselines. For CIFAR, plain (S=2, TA=4, T=10) and for ResNet (S=8, TA=20, T=110) are used. For ImageNet, ResNet (S=14, TA=20, T=50) is used. Higher numbers are better.

Model	Dataset	NOKD	BLKD	TAKD
CNN	CIFAR-10	70.16	72.57	73.51
	CIFAR-100	41.09	44.57	44.92
ResNet	CIFAR-10	88.52	88.65	88.98
	CIFAR-100	61.37	61.41	61.82
ResNet	ImageNet	65.20	66.60	67.36

68.21% accuracy from ResNet-50 to ResNet-14 on ImageNet

Transfer of knowledge

Efficient Knowledge Distillation from Model Checkpoints

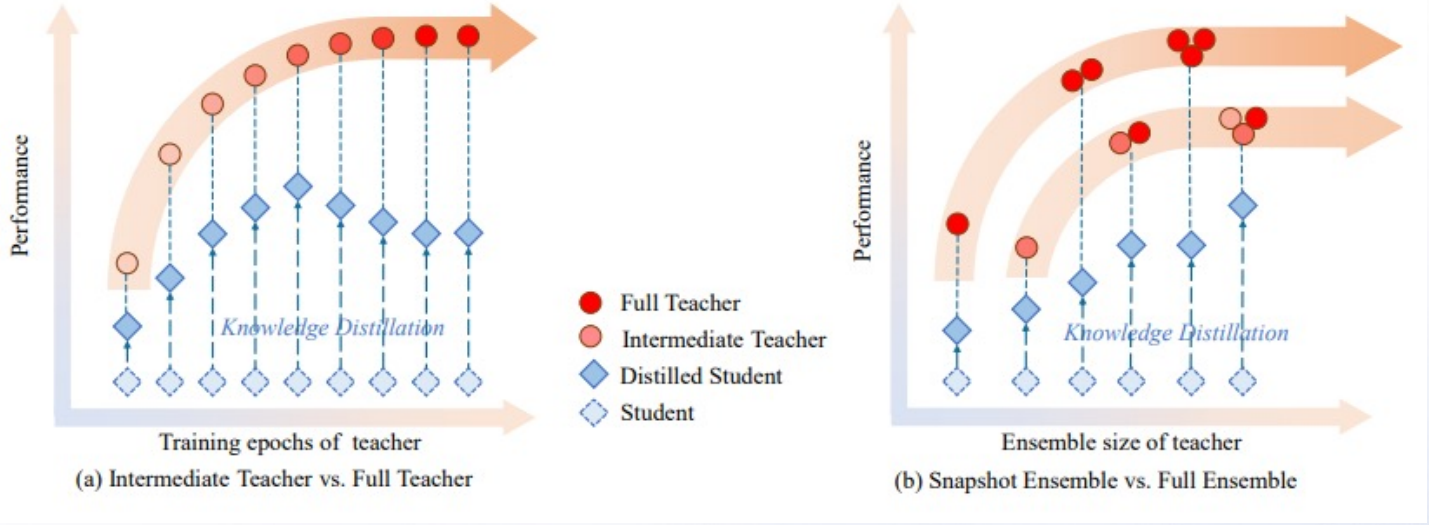
Chaofei Wang*, **Qisen Yang***, **Rui Huang**, **Shiji Song**, **Gao Huang**†
Department of Automation, Tsinghua University, China
wangcf18, yangqs19, hr20@mails.tsinghua.edu.cn
shijis, gaohuang@tsinghua.edu.cn

Transfer of knowledge

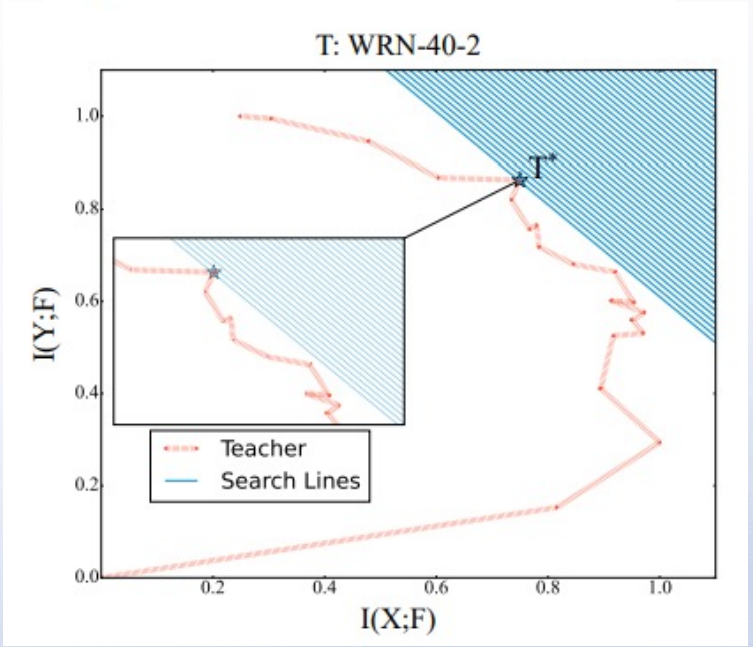
Problem: fully trained teachers do not always provide the best knowledge distillation

Method: select optimal intermediate teacher checkpoints by **maximizing mutual information**

$$\max_{F} \{I(X;F) + I(Y;F)\}$$



A weak **intermediate model** can serve as a better teacher than the strong fully converged model.



select the optimal intermediate teacher

Transfer of knowledge

Result: improved distillation performance by selecting intermediate teachers

Network structure		Accuracy of T&S		KD accuracy of different intermediate teachers				
T	S	T	S	$T^{0.3}$	$T^{0.5}$	$T^{0.7}$	T^{full}	T^{*}
WRN-40-2	WRN-40-1	76.53	70.38	72.34±0.10	72.76±0.24	73.08±0.05	72.68±0.10	**73.26**±0.03
	MobileNetV2		64.49	68.21±0.33	**68.99**±0.12	68.54±0.07	68.03±0.34	68.58±0.34
ResNet-110	ResNet-32	73.41	70.16	70.74±0.18	72.49±0.32	72.46±0.30	72.48±0.22	**72.63**±0.13
	MobileNetV2		64.49	67.84±0.26	68.79±0.17	**69.01**±0.20	68.63±0.35	68.99±0.33
Average		74.97	67.38	69.78	70.76	70.77	70.46	**70.87**

Transfer of knowledge

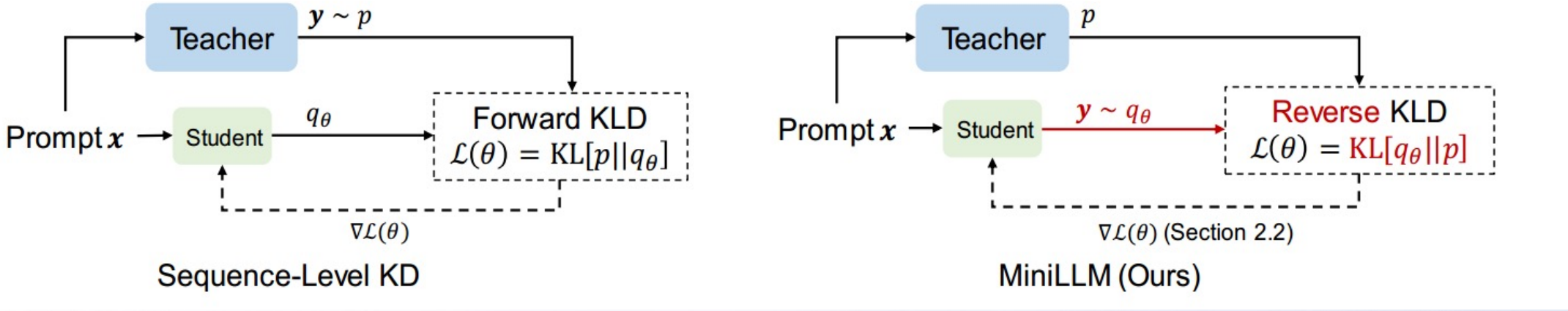
MiniLLM: Knowledge Distillation of Large Language Models

Yuxian Gu[1,2]*; **Li Dong**[2], **Furu Wei**[2], **Minlie Huang**[1†]
[1]The CoAI Group, Tsinghua University
[2]Microsoft Research
guyx21@mails.tsinghua.edu.cn {lidong1,fuwei}@microsoft.com
aihuang@tsinghua.edu.cn

Transfer of knowledge

Problem: In open-ended text generation tasks，student model overestimates the low-probability regions of the teacher distribution

Method: replace the standard KLD with **reverse KLD**



reverse KLD causes q_θ to seek the major modes of p, and assign low probabilities to p's void regions

$$\theta = \arg\min_\theta \mathcal{L}(\theta) = \arg\min_\theta \mathrm{KL}[q_\theta \,\|\, p]$$
$$= \arg\min_\theta \left[-\mathop{\mathbb{E}}_{x\sim p_x, y\sim q_\theta} \log \frac{p(y \mid x)}{q_\theta(y \mid x)} \right]$$

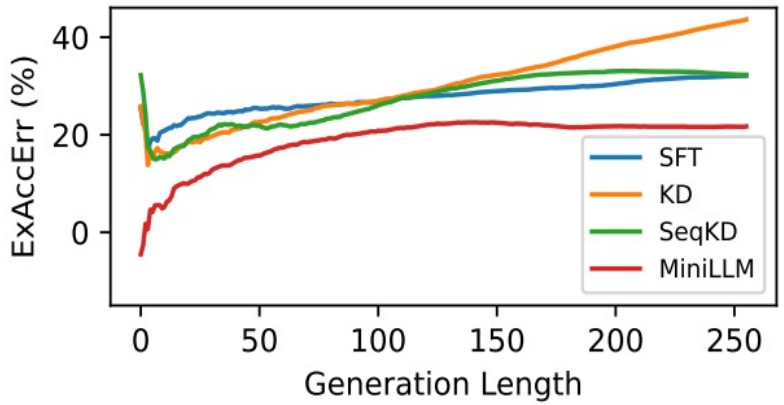
Transfer of knowledge

Result: In contrast to previous KD methods, MINILLM narrows the performance gap between the student and the teacher

Model	#Params	Method	DollyEval GPT4	DollyEval R-L	SelfInst GPT4	SelfInst R-L	VicunaEval GPT4	VicunaEval R-L	S-NI R-L	UnNI R-L
GPT-2	1.5B	Teacher	58.4	27.6	42.9	14.3	48.6	16.3	27.6	31.8
	120M	SFT w/o KD	38.6	23.3	26.3	10.0	32.8	14.7	16.3	18.5
		KD	40.3	22.8	27.8	10.8	31.9	13.4	19.7	22.0
		SeqKD	41.2	22.7	26.2	10.1	31.0	14.3	16.4	18.8
		MINILLM	**44.7**	**24.6**	**29.2**	**13.2**	**34.1**	**16.9***	**25.3**	**26.6**
	340M	SFT w/o KD	51.9	**25.5**	39.6	13.0	42.3	16.0	25.1	32.0
		KD	51.6	25.0	39.2	12.0	42.8	15.4	23.7	31.0
		SeqKD	50.5	25.3	39.0	12.6	**43.0**	16.9*	22.9	30.2
		MINILLM	**52.2**	25.4	**40.5**	**15.6**	42.6	**17.7***	**27.4**	**34.5**
	760M	SFT w/o KD	50.7	25.4	38.3	12.4	43.1	16.1	21.5	27.1
		KD	53.4	25.9	40.4	13.4	43.4	16.9*	25.3	31.7
		SeqKD	52.0	25.6	38.9	14.0	42.4	15.9	26.1	32.9
		MINILLM	**54.7**	**26.4**	**44.6***	**15.9**	**45.7**	**18.3***	**29.3***	**37.7***
OPT	13B	Teacher	70.3	29.2	56.1	18.4	58.0	17.8	30.4	36.1
	1.3B	SFT w/o KD	52.6	26.0	37.7	11.4	40.5	15.6	23.1	28.4
		KD	52.7	25.4	36.0	12.2	40.8	14.9	21.9	27.0
		SeqKD	51.0	26.1	36.6	12.7	42.6	16.6	21.4	28.2
		MINILLM	**60.7**	**26.7**	**47.0**	**14.8**	**50.6**	**17.9***	**28.6**	**33.4**
	2.7B	SFT w/o KD	55.4	27.1	38.9	13.9	44.8	16.6	24.9	32.3
		KD	60.5	25.9	48.6	13.8	51.3	16.7	26.3	30.2
		SeqKD	57.6	27.5	40.5	13.3	44.5	16.5	25.3	32.3
		MINILLM	**63.2**	**27.4**	**52.7**	**17.2**	**55.9**	**19.1***	**30.7***	**35.1**
	6.7B	SFT w/o KD	67.9	27.6	56.4	16.4	57.3	17.8	30.3	28.6
		KD	68.6	28.3	58.0	17.0	57.0	17.5	30.7*	26.7
		SeqKD	69.6	28.5	54.0	17.0	57.6	17.9*	30.4	28.2
		MINILLM	**70.8***	**29.0**	**58.5***	**17.5**	**60.1***	**18.7***	**32.5***	**36.7***
LLaMA	13B	Teacher	79.0	29.7	75.5	23.4	65.1	19.4	35.8	38.5
	7B	SFT w/o KD	73.0	26.3	69.2	20.8	61.6	17.5	32.4	35.8
		KD	73.7	27.4	70.5	20.2	62.7	18.4	33.7	37.9
		SeqKD	73.6	27.5	71.5	20.8	62.6	18.1	33.7	37.6
		MINILLM	**76.4**	**29.0**	**73.1**	**23.2**	**64.1**	**20.7***	**35.5**	**40.2***

	SST2 ECE	SST2 Acc.	BoolQ ECE	BoolQ Acc.
Teacher	0.025	93.0	0.356	74.5
KD	0.191	84.7	0.682	63.5
SeqKD	0.243	66.5	0.681	62.8
MINILLM	**0.099**	**89.7**	**0.502**	**67.8**

Higher scores on SST2 and BoolQ



Lower ExAccErr

Transfer of knowledge

DISTILLM: Towards Streamlined Distillation for Large Language Models

Jongwoo Ko [1] Sungnyun Kim [1] Tianyi Chen [2] Se-Young Yun [1]

https://github.com/jongwooko/distillm

Transfer of knowledge

Problem: current KD methods for auto-regressive sequence models suffer from missing a standardized objective function

Method: propose a method that consists of novel Skew KLD and Aaptive off-policy approach

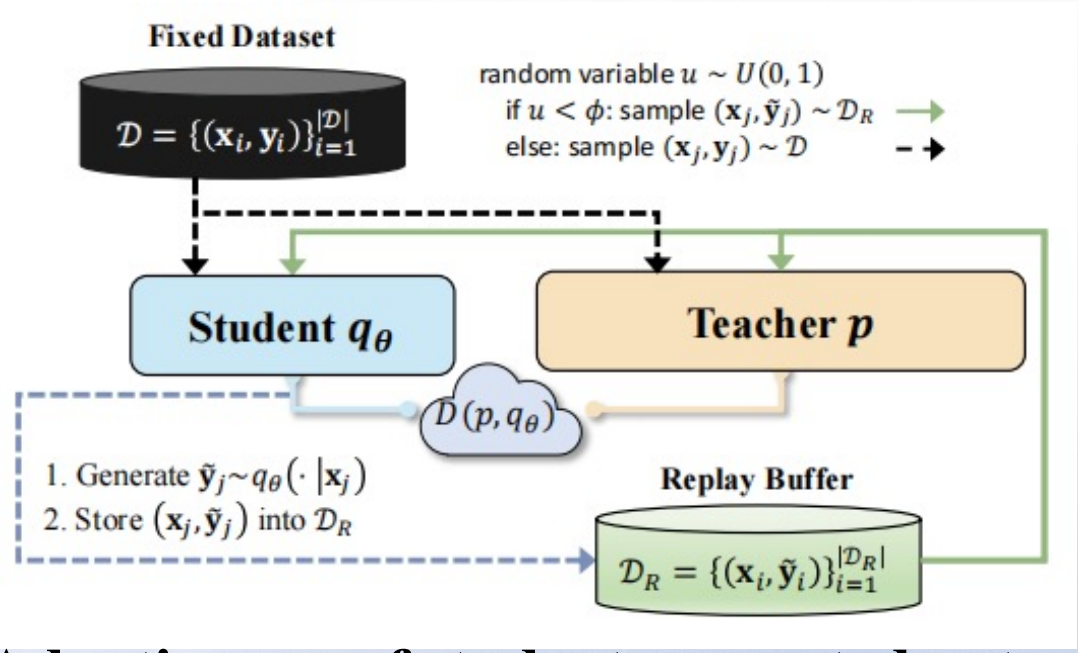
$$D_{\mathrm{KL}}(p, q_\theta) = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y} \sim p(\cdot|\mathbf{x})} \left[\log \frac{p(\mathbf{y} \mid \mathbf{x})}{q_\theta(\mathbf{y} \mid \mathbf{x})} \right]$$

mix two distributions to replace student

$$D_{\mathrm{SKL}}^{(\alpha)}(p, q_\theta) = D_{\mathrm{KL}}(p, \alpha p + (1-\alpha) q_\theta)$$

more stable gradient

smaller approximation error

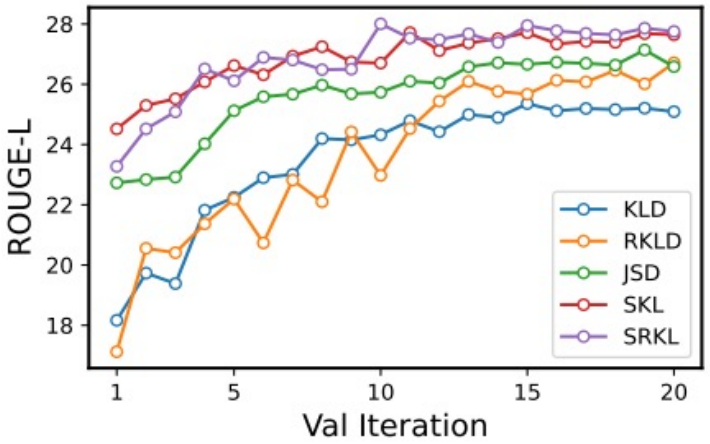


Adaptive use of student generated output

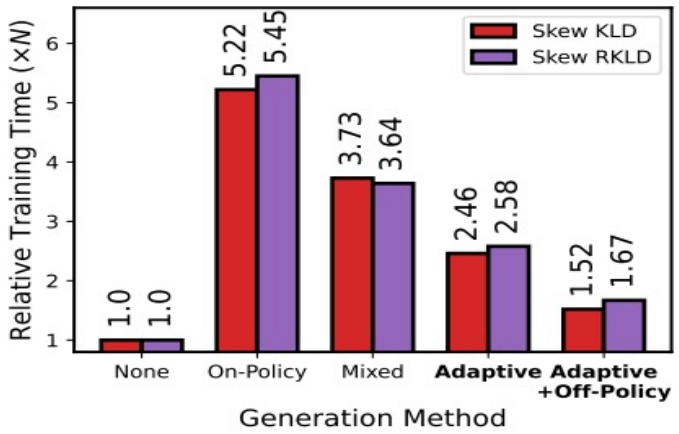
randomly draw samples from Replay Buffer

Transfer of knowledge

Result: **achieves significant training efficiency and performance improvement.**



Dataset	SAMSum		IWSLT 2017 En-De	
T5-XL →	***T5-Base***	***T5-Small***	***T5-Base***	***T5-Small***
KD (Hinton et al., 2015)	46.23	39.52	29.36	21.15
SeqKD (Kim & Rush, 2016)	46.89	40.24	29.07	21.42
ImitKD (Lin et al., 2020)	48.57	41.44	29.87	21.52
GKD (Agarwal et al., 2024)	48.49	41.92	30.24	22.04
DISTILLM (ours)	**49.11**	**42.37**	**30.32**	**22.53**



Dataset	Dolly Eval		Self-Instruct		Super-Natural	
Sampling	on-	off-	on-	off-	on-	off-
ImitKD (Lin et al., 2020)	21.63	20.62	10.85	10.09	19.94	18.04
GKD (Agarwal et al., 2024)	23.75	22.89	12.73	12.78	26.05	24.97
DISTILLM (ours)	26.37	26.12	13.14	13.16	28.24	28.20

目录

Application of KD

TinyBERT: Distilling BERT for Natural Language Understanding

Xiaoqi Jiao[1]*†, Yichun Yin[2]*‡, Lifeng Shang[2]‡, Xin Jiang[2]
Xiao Chen[2], Linlin Li[3], Fang Wang[1]‡ and Qun Liu[2]
[1]Key Laboratory of Information Storage System, Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics
[2]Huawei Noah's Ark Lab
[3]Huawei Technologies Co., Ltd.
{jiaoxiaoqi,wangfang}@hust.edu.cn
{yinyichun,shang.lifeng,jiang.xin}@huawei.com
{chen.xiao2,lynn.lilinlin,qun.liu}@huawei.com

Application of KD

Problem: pre-trained language models (PLMs) like BERT are computationally expensive and difficult to deploy on edge devices due to the large size

Method: propose a new Transformer distillation method and a two-stage learning framework called TinyBERT

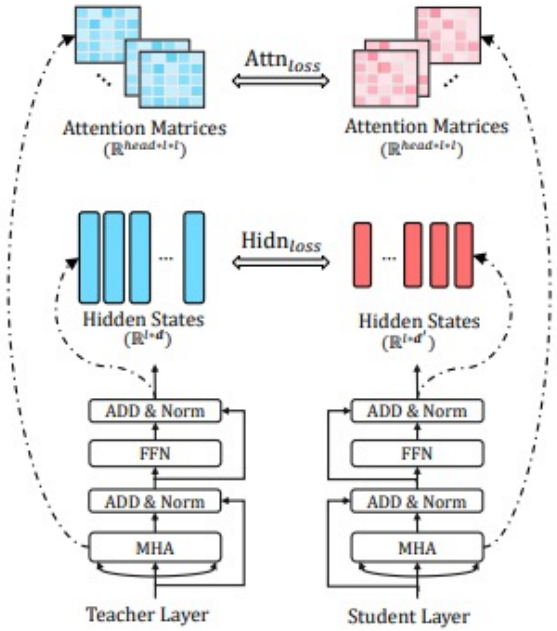


Figure 2: The details of Transformer-layer distillation consisting of Attn$_{loss}$(attention based distillation) and Hidn$_{loss}$(hidden states based distillation).

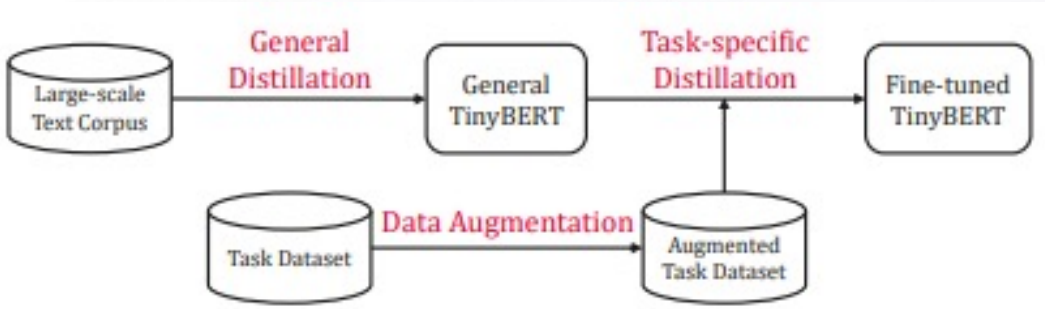


Figure 1: The illustration of TinyBERT learning.

distill knowledge from a large teacher BERT into a small student model through both general and task-specific distillations

Application of KD

Result: TinyBERT4 achieves more than 96.8% of BERT's performance while being 7.5x smaller and 9.4x faster, and TinyBERT6 performs on par with BERT

System	#Params	#FLOPs	Speedup	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg
$BERT_{BASE}$ (Teacher)	109M	22.5B	1.0x	83.9/83.4	71.1	90.9	93.4	52.8	85.2	87.5	67.0	79.5
$BERT_{TINY}$	14.5M	1.2B	9.4x	75.4/74.9	66.5	84.8	87.6	19.5	77.1	83.2	62.6	70.2
$BERT_{SMALL}$	29.2M	3.4B	5.7x	77.6/77.0	68.1	86.4	89.7	27.8	77.0	83.4	61.8	72.1
$BERT_4$-PKD	52.2M	7.6B	3.0x	79.9/79.3	70.2	85.1	89.4	24.8	79.8	82.6	62.3	72.6
$DistilBERT_4$	52.2M	7.6B	3.0x	78.9/78.0	68.5	85.2	91.4	32.8	76.1	82.4	54.1	71.9
$MobileBERT_{TINY}$†	15.1M	3.1B	-	81.5/81.6	68.9	**89.5**	91.7	**46.7**	80.1	**87.9**	65.1	**77.0**
$TinyBERT_4$ (ours)	14.5M	1.2B	9.4x	**82.5/81.8**	**71.3**	87.7	**92.6**	44.1	**80.4**	86.4	**66.6**	**77.0**
$BERT_6$-PKD	67.0M	11.3B	2.0x	81.5/81.0	70.7	89.0	92.0	-	-	85.0	65.5	-
PD	67.0M	11.3B	2.0x	82.8/82.2	70.4	88.9	91.8	-	-	86.8	65.3	-
$DistilBERT_6$	67.0M	11.3B	2.0x	82.6/81.3	70.1	88.9	92.5	49.0	81.3	86.9	58.4	76.8
$TinyBERT_6$ (ours)	67.0M	11.3B	2.0x	**84.6/83.2**	**71.6**	**90.4**	**93.1**	**51.1**	**83.7**	**87.3**	**70.0**	**79.4**

Evaluation on the GLUE benchmark

Application of KD

Model Compression with Two-stage Multi-teacher Knowledge Distillation for Web Question Answering System*

Ze Yang†
yaze@microsoft.com
STCA NLP Group, Microsoft

Linjun Shou†
lisho@microsoft.com
STCA NLP Group, Microsoft

Ming Gong
migon@microsoft.com
STCA NLP Group, Microsoft

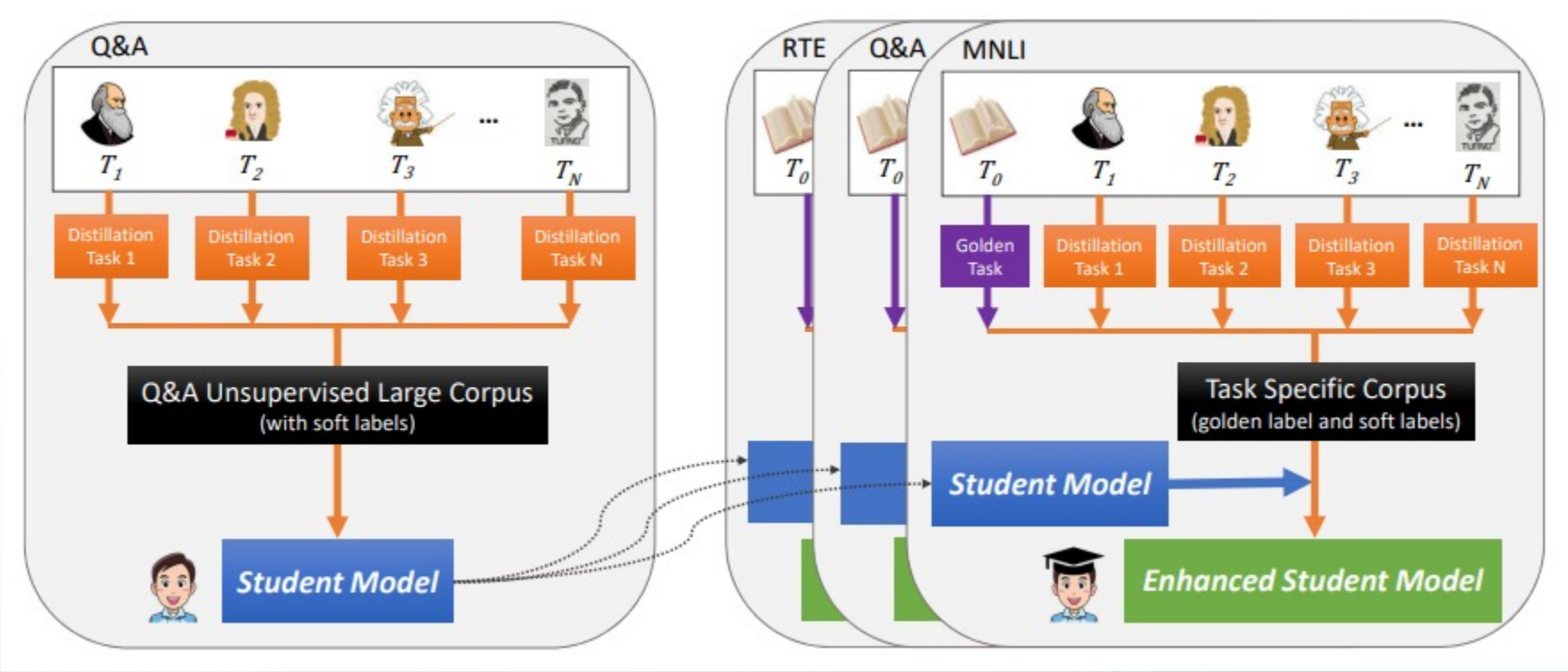
Wutao Lin
wutlin@microsoft.com
STCA NLP Group, Microsoft

Daxin Jiang
djiang@microsoft.com
STCA NLP Group, Microsoft

Application of KD

Problem: previous model compression methods for PLMs usually suffer from information loss, leading to inferior models compared with the original one

Method: propose a **Two-stage Multi-teacher Knowledge Distillation** (TMKD) method



general distillation task for student model pre-training in the first stage

multi-teacher distillation for fine-tuning on downstream tasks in the second stage

Application of KD

Result: **significantly outperforms baseline methods and even achieves comparable performance with the original teacher models, with substantial improvement in model inference speed**

Model		Performance (ACC)					Inference Speed(QPS)	Parameters (M)
		DeepQA	MNLI	SNLI	QNLI	RTE		
Original Model	BERT-3	75.78	70.77	77.75	78.51	57.42	207	50.44
	$BERT_{large}$	81.47	79.10	80.90	90.30	68.23	16	333.58
	$BERT_{large}$ ensemble	81.66	79.57	81.39	90.91	70.75	16/3	333.58*3
Traditional Distillation Model	Bi-LSTM (1-o-1)	71.69	59.39	69.59	69.12	56.31	207	50.44
	Bi-LSTM (1_{avg}-o-1)	71.93	59.60	70.04	69.53	57.35	207	50.44
	Bi-LSTM (m-o-m)	72.04	61.71	72.89	69.89	58.12	207/3	50.44*3
	BERT-3 (1-o-1)	77.35	71.07	78.62	77.65	55.23	217	45.69
	BERT-3 (1_{avg}-o-1)	77.63	70.63	78.64	78.20	58.12	217	45.69
	BERT-3 (m-o-m)	77.44	71.28	78.71	77.90	57.40	217/3	45.69*3
Our Distillation Model	Bi-LSTM ($TMKD_{base}$)	74.73	61.68	71.71	69.99	62.74	207	50.45
	*$TMKD_{base}$	79.93	71.29	78.35	83.53	66.64	217	45.70
	*$TMKD_{large}$	**80.43**	**73.93**	**79.48**	**86.44**	**67.50**	**217**	**45.70**

* These two models are BERT-3 based models.

Application of KD

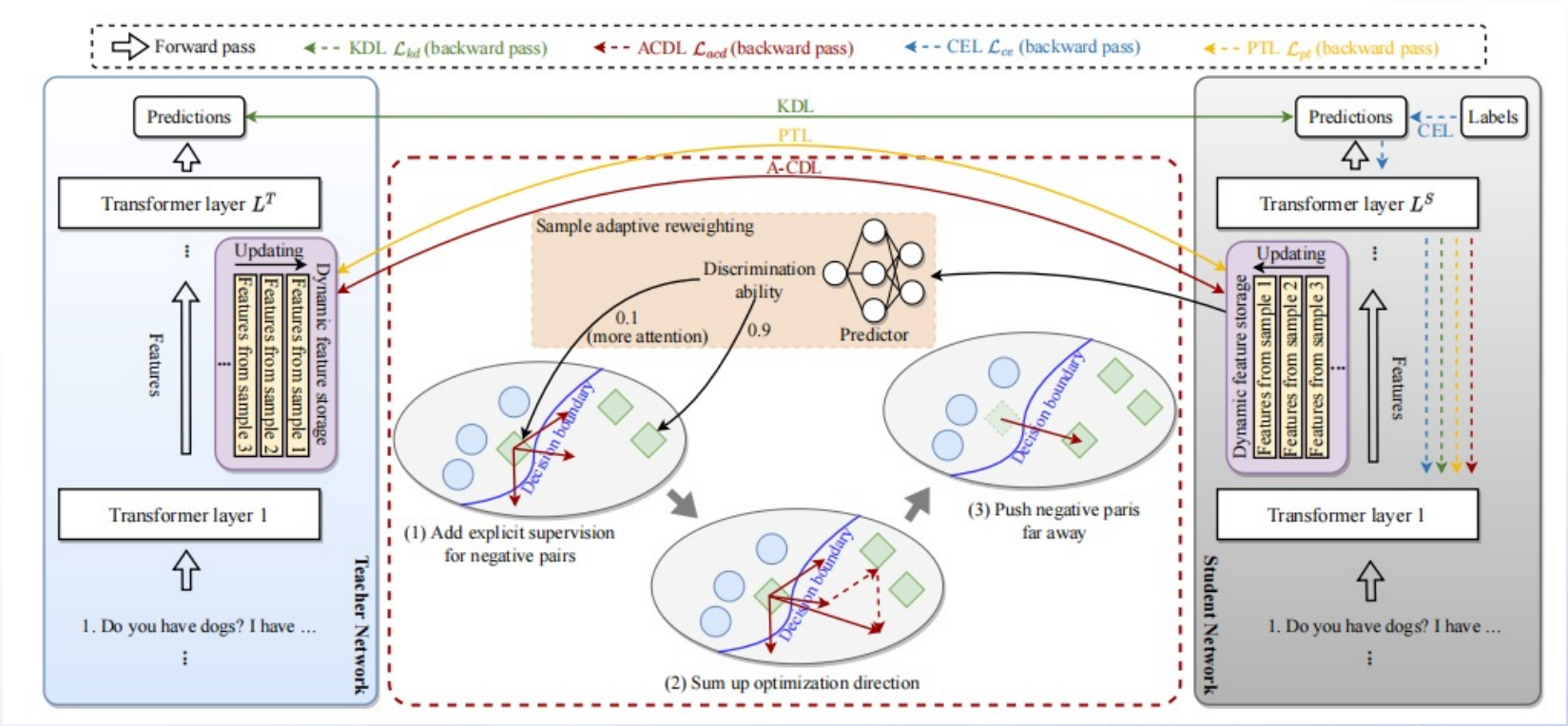
Adaptive Contrastive Knowledge Distillation for BERT Compression

Jinyang Guo[1,2]*, Jiaheng Liu[2]*, Zining Wang[2], Yuqing Ma[2],
Ruihao Gong[2,3], Ke Xu[2] and Xianglong Liu[2]†
[1]Institute of Artificial Intelligence, Beihang University
[2]State Key Lab of Software Development Environment, Beihang University
[3]SenseTime Group Limited
{jinyangguo,liujiaheng}@buaa.edu.cn, xlliu@buaa.edu.cn

Application of KD

Problem: it is hard to distinguish similar sentences with completely different meanings

Method: propose a novel **contrastive distillation** loss to introduce **explicit supervision** for learning discriminative student features

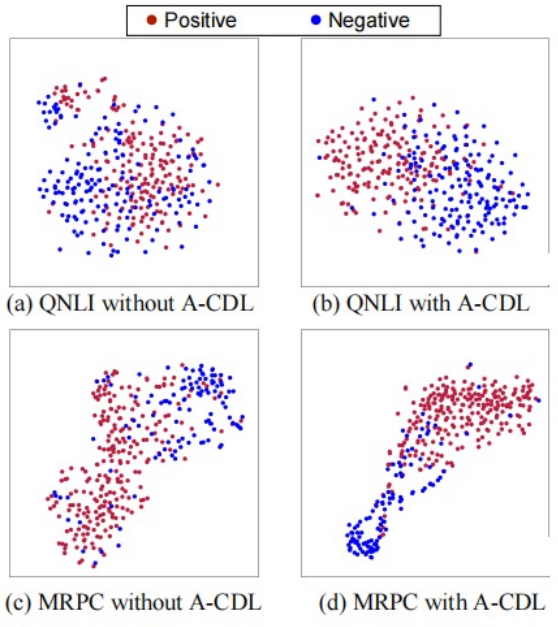


adaptively pay more attention to hard samples with fewer discrimination abilities

Application of KD

Result: performance improvements of language models on multiple datasets

Method	#Param	Speed-up	GLUE							
			CoLA (Matt.)	MNLI (Acc -m/-mm)	MRPC (F1/Acc)	QNLI (Acc)	QQP (F1/Acc)	RTE (Acc.)	SST-2 (Acc.)	STS-B (Pear./Spear.)
Teacher Network: BERT-Base										
BERT-Base (Devlin et al., 2018)	110M	1.0×	60.8	84.6/84.4	91.6/87.6	91.6	88.5/91.4	71.4	93.0	90.2/89.8
Student Network: $BERT_3$										
PKD (Sun et al., 2019)	46M	4.0×	39.8	75.9/76.6	84.1/75.0	84.3	85.3/89.2	62.8	87.4	86.3/86.1
RCO (Jin et al., 2019)	46M	4.0×	31.4	76.3/76.9	85.3/77.5	83.4	85.4/88.7	66.1	86.8	84.8/84.4
TAKD (Mirzadeh et al., 2020)	46M	4.0×	35.7	76.2/76.8	83.2/73.5	83.8	83.7/87.5	59.2	87.9	83.8/83.4
DistilBERT (Sanh et al., 2019)	46M	4.0×	34.0	77.0/77.0	83.2/73.0	83.8	85.1/88.9	62.8	86.9	86.6/86.2
TinyBERT (Jiao et al., 2019)	46M	4.0×	38.7	76.5/76.9	82.8/72.8	84.2	85.1/88.8	60.6	86.8	86.4/86.1
CRD (Tian et al., 2019)	46M	4.0×	38.6	76.1/76.8	85.2/77.5	84.6	83.9/88.0	65.7	87.6	86.1/85.6
SFTN (Park et al., 2021)	46M	4.0×	38.1	76.6/77.1	83.1/73.3	84.2	83.9/87.7	60.3	88.0	83.9/83.5
MetaDistill (Zhou et al., 2022)	46M	4.0×	39.3	75.9/76.4	82.0/71.1	83.8	83.7/88.1	62.1	88.0	86.6/86.4
Annealing KD (Jafari et al., 2021)	52M	3.0×	36.0	73.9/74.8	86.2/-	83.1	-/86.5	61.0	**89.4**	74.5/-
ACKD (ours)	46M	4.0×	**42.7**	**79.5/80.6**	**87.5/81.4**	**86.2**	**86.1/89.7**	**67.9**	88.5	**87.1/86.8**
Student Network: $BERT_6$										
PKD (Sun et al., 2019)	66M	2.0×	54.5	82.7/83.3	89.4/84.7	89.5	87.8/90.9	67.6	91.3	88.6/88.1
RCO (Jin et al., 2019)	66M	2.0×	53.6	82.4/82.9	89.5/85.1	89.7	87.4/90.6	67.6	91.4	88.7/88.3
TAKD (Mirzadeh et al., 2020)	66M	2.0×	53.8	82.5/83.0	89.6/85.0	89.6	87.5/90.7	68.5	91.4	88.2/88.0
DistilBERT (Sanh et al., 2019)	66M	2.0×	53.0	82.5/83.1	89.3/85.0	89.2	87.2/90.6	66.1	91.5	88.7/88.5
TinyBERT (Jiao et al., 2019)	66M	2.0×	52.4	83.6/83.8	90.5/86.5	89.8	87.6/90.6	67.7	91.9	89.2/88.7
CRD (Tian et al., 2019)	66M	2.0×	55.8	83.2/83.4	89.5/85.5	89.8	87.6/90.8	67.1	91.5	88.8/88.3
SFTN (Park et al., 2021)	66M	2.0×	53.6	82.4/82.9	89.8/85.3	89.5	87.5/90.4	68.5	91.5	88.4/88.5
MetaDistill (Zhou et al., 2022)	66M	2.0×	58.6	83.5/83.8	**91.1**/86.8	90.4	88.1/91.0	69.4	92.3	89.4/**89.1**
ALP-KD (Passban et al., 2021)	66M	2.0×	46.4	82.0/-	-/85.8	89.7	-/90.6	69.0	91.9	88.8/-
CoDIR (Sun et al., 2020)	66M	2.0×	56.4	**83.9**/-	87.9/-	**90.7**	-/91.2	66.3	**92.4**	-/-
ACKD (ours)	66M	2.0×	**59.7**	83.6/**83.9**	91.0/**87.0**	90.6	**88.5/91.3**	**69.7**	92.3	**89.5/89.1**



(a) QNLI without A-CDL (b) QNLI with A-CDL

(c) MRPC without A-CDL (d) MRPC with A-CDL

significant improvement in feature discrimination

Application of KD

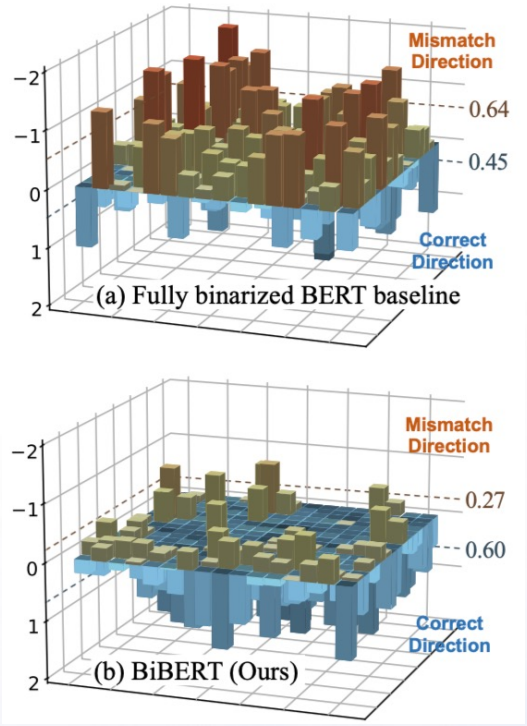
BiBERT: Accurate Fully Binarized BERT

Haotong Qin[*1,4], **Yifu Ding**[*1,4], **Mingyuan Zhang**[*2], **Qinghua Yan**[1], **Aishan Liu**[1], **Qingqing Dang**[3], **Ziwei Liu**[2], **Xianglong Liu**[✉ 1]
[1]State Key Lab of Software Development Environment, Beihang University [3]Baidu Inc.
[2]S-Lab, Nanyang Technological University [4]Shen Yuan Honors College, Beihang University
{qinhaotong,yifuding,yanqh,aishanliu,xlliu}@buaa.edu.cn
mingyuan001@e.ntu.edu.sg zwliu.hust@gmail.com dangqingqing@baidu.com

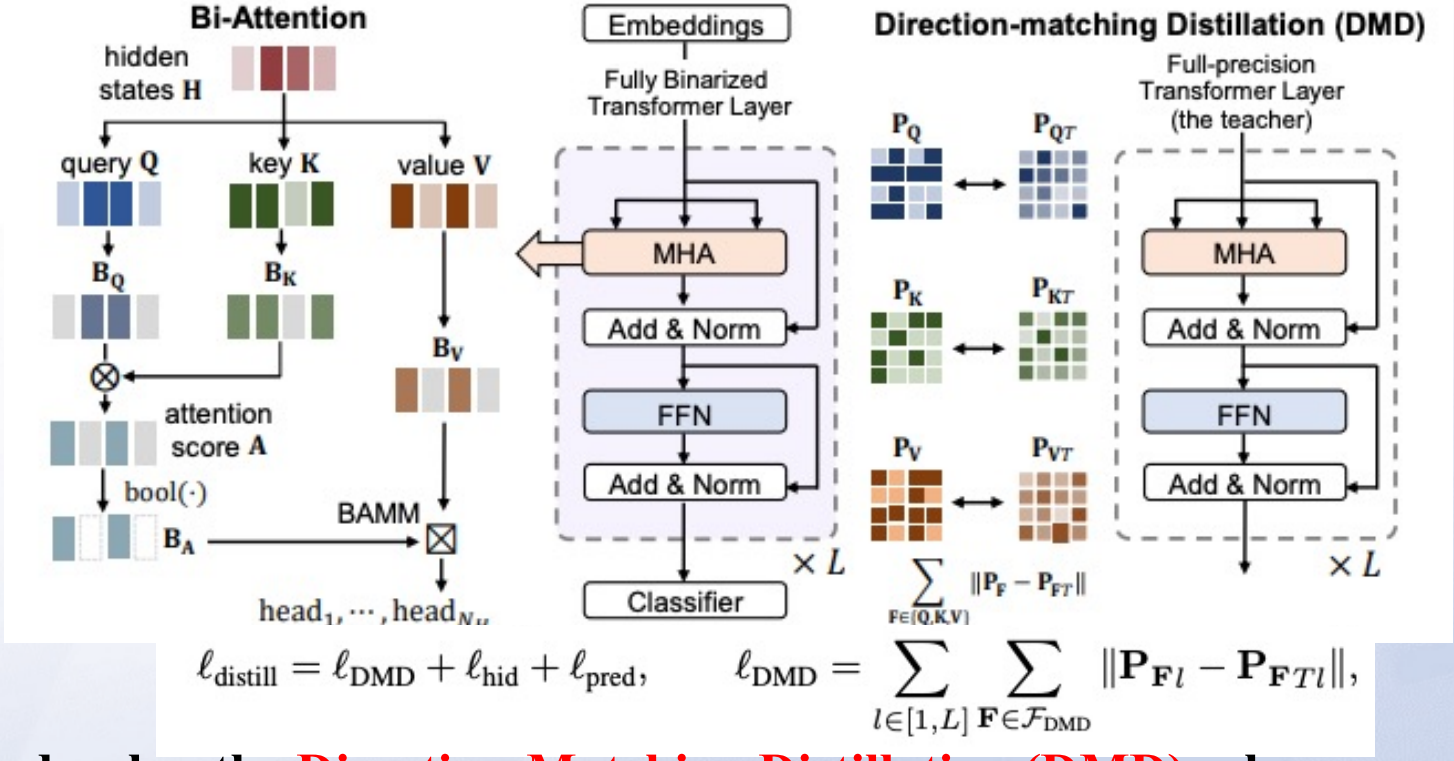
Application of KD

Problem: transfer knowledge from full-precision teacher network to binarized student network

Method: Direction-Matching Distillation



optimization direction mismatch



$$\ell_{\text{distill}} = \ell_{\text{DMD}} + \ell_{\text{hid}} + \ell_{\text{pred}}, \qquad \ell_{\text{DMD}} = \sum_{l\in[1,L]} \sum_{\mathbf{F}\in\mathcal{F}_{\text{DMD}}} \|\mathbf{P}_{\mathbf{F}l} - \mathbf{P}_{\mathbf{F}Tl}\|,$$

develop the Direction-Matching Distillation (DMD) scheme to eliminate the direction mismatch in distillation

Application of KD

Result

Table 2: Comparison of BERT quantization methods without data augmentation

Quant	#Bits (W-E-A)	Size (MB)	FLOPs (G)	MNLI -m/mm	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg.
Full Prec.	32-32-32	418	22.5	84.9/85.5	91.4	92.1	93.2	59.7	90.1	86.3	72.2	83.9
Baseline	1-1-1	13.4	0.4	45.8/47.0	73.2	66.4	77.6	11.7	7.6	70.2	54.1	50.4
BinaryBERT	1-1-1	16.5	0.4	35.6/35.3	63.1	51.5	53.2	0	6.1	68.3	52.7	40.6
BinaryBERT	1-1-2	16.5	0.8	35.4/35.2	63.1	52.6	82.5	14.6	6.5	68.3	52.7	45.7
TernaryBERT	2-2-1	28.0	0.8	32.7/33.0	74.1	59.3	53.1	0	7.1	68.3	53.4	42.3
TernaryBERT	2-2-2	28.0	1.5	40.3/40.0	63.1	50.0	80.7	0	12.4	68.3	54.5	45.5
Q2BERT	2-8-8	43.0	6.5	47.2/47.3	67.0	61.3	80.6	0	4.4	68.4	52.7	47.7
Q-BERT	2-8-8	43.0	6.5	**76.6/77.0**	–	–	84.6	–	–	68.3	52.7	–
BiBERT (ours)	**1-1-1**	**13.4**	**0.4**	59.3/60.0	**82.4**	**70.2**	**86.9**	**25.3**	**33.5**	**72.9**	**58.5**	**61.0**

Table 3: Comparison of BERT quantization methods with data augmentation

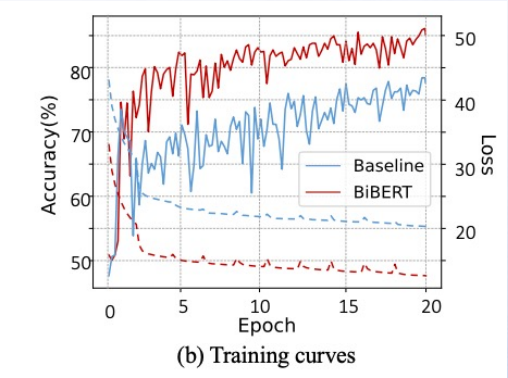
Quant	#Bits (W-E-A)	Size (MB)	FLOPs (G)	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg.
Full Prec.	32-32-32	418	22.5	92.1	93.2	59.7	90.1	86.3	72.2	82.3
Baseline	1-1-1	13.4	0.4	69.2	84.0	23.3	14.4	71.4	50.9	52.2
BinaryBERT	1-1-1	16.5	0.4	66.1	68.0	7.3	22.1	69.3	57.7	48.4
BinaryBERT	1-1-2	16.5	0.8	51.0	89.6	33.0	11.4	71.0	55.9	52.0
TernaryBERT	2-2-1	28.0	0.8	50.9	80.3	6.5	10.3	71.5	53.4	45.5
TernaryBERT	2-2-2	28.0	1.5	50.0	87.5	20.6	**72.5**	72.0	47.2	58.3
BiBERT (ours)	**1-1-1**	**13.4**	**0.4**	**76.0**	**90.9**	**37.8**	56.7	**78.8**	**61.0**	**67.0**

BiBERT outperforms existing binarized quantization methods of BERT, even higher bits quantization methods.

Table 1: Ablation study.

Quant	#Bits	DA	SST-2	MRPC	RTE	QQP
Full Precision	32-32-32	–	93.2	86.3	72.2	91.4
Baseline	1-1-1	✗	77.6	70.2	54.1	73.2
Bi-Attention	1-1-1	✗	82.1	70.5	55.6	74.9
DMD	1-1-1	✗	79.9	70.5	55.2	75.3
BiBERT (ours)	1-1-1	✗	**88.7**	**72.5**	**57.4**	**84.8**
Baseline	1-1-1	✓	84.0	71.4	50.9	-
Bi-Attention	1-1-1	✓	85.6	73.2	53.1	-
DMD	1-1-1	✓	85.3	72.5	56.3	-
BiBERT (ours)	1-1-1	✓	**90.9**	**78.8**	**61.0**	-

Higher model performance



Faster convergence rate

Application of KD

HuatuoGPT, Towards Taming Language Models To Be a Doctor

Hongbo Zhang[1,2†], **Junying Chen**[1,2†], **Feng Jiang**[1,2,3†], **Fei Yu**[1,2], **Zhihong Chen**[1,2], **Jianquan Li**[2], **Guiming Chen**[1,2], **Xiangbo Wu**[2], **Zhiyi Zhang**[2], **Qingying Xiao**[1], **Xiang Wan**[1,2], **Benyou Wang**[1,2 *], **Haizhou Li**[1,2]

[1]Shenzhen Research Institue of Big Data, [2]The Chinese University of Hong Kong, Shenzhen

[3]University of Science and Technology of China

hongboz183@gmail.com, junying.chen.cs@gmail.com, jeffreyjiang@cuhk.edu.cn

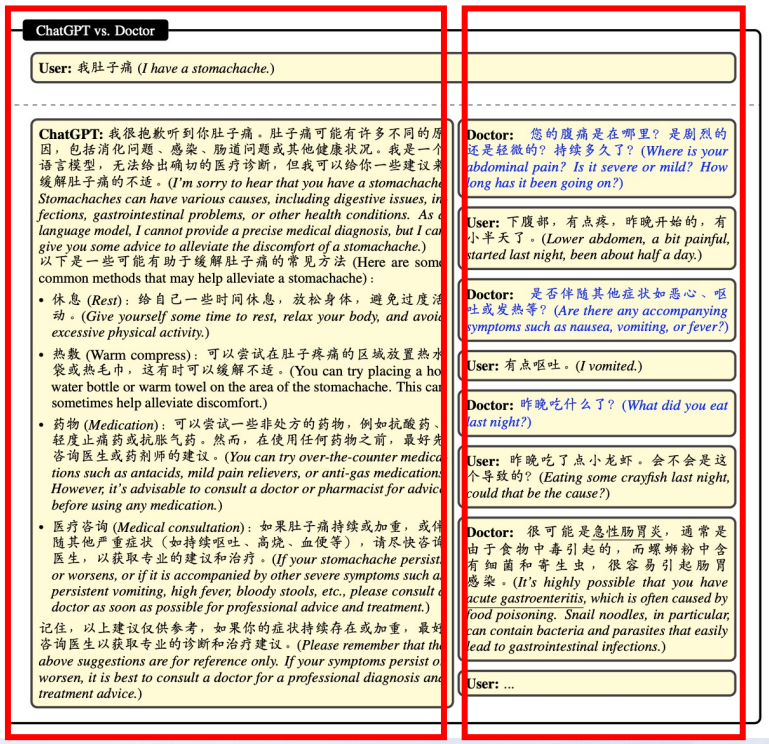
wangbenyou@cuhk.edu.cn

Application of KD

Problem: ChatGPT did not reach the expert level and lack of proficiency in medical knowledge that medical doctors have

Method: propose HuatuoGPT, combining the strengths of both distilled data (from ChatGPT) and real-world data (from Doctors) to tame the medical LLM

Distilled data from ChatGPT



Real world data from doctors

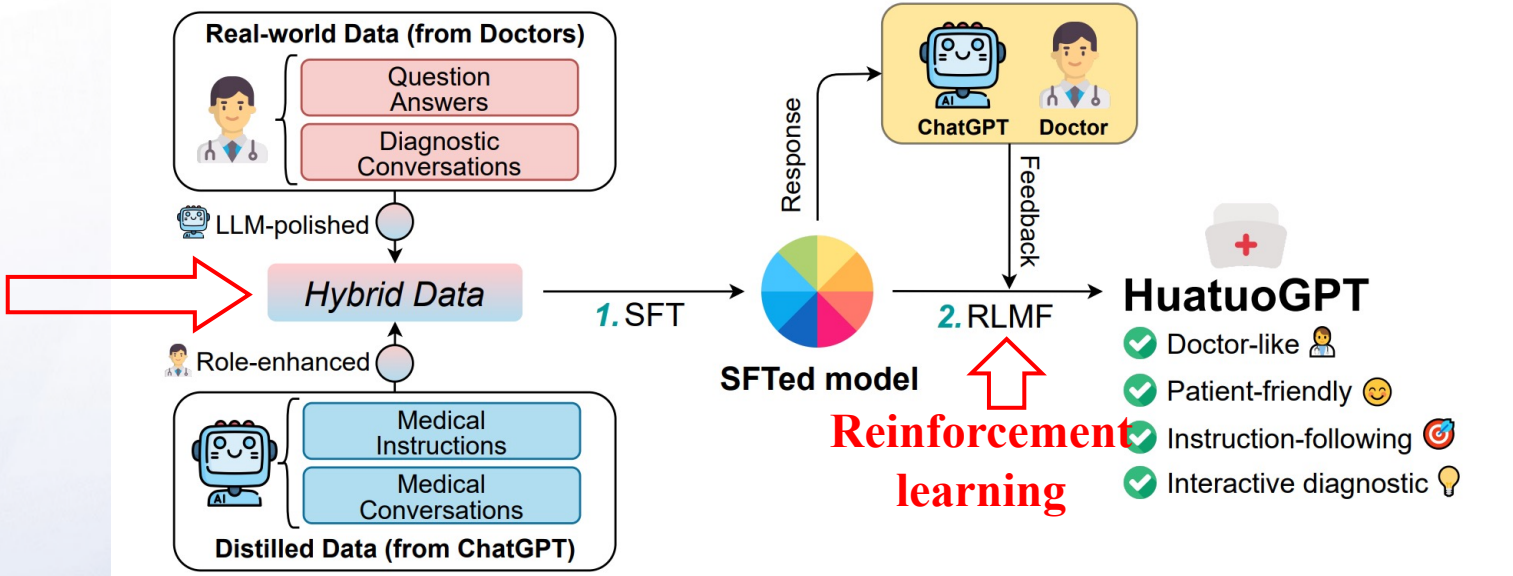
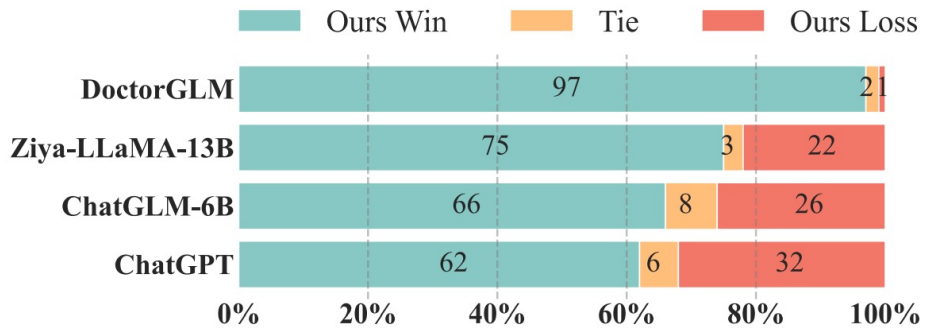


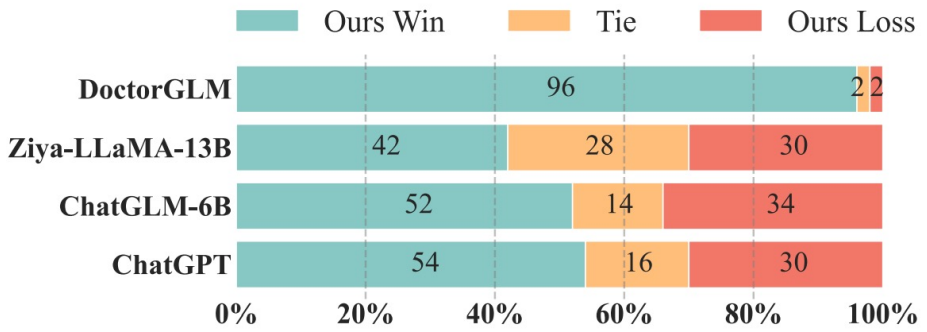
Figure 2: Schematic of HuatuoGPT.

Application of KD

Result: achieves state-of-the-art results in performing **medical consultation** among open-source LLMs, **outperforms its teacher model** (i.e., ChatGPT) in most cases.



(a) GPT-4 Evaluation



(b) Doctor Evaluation

User:
我最近拉肚子，有泡沫黄水，怀疑是喝牛奶喝的，是不是喝牛奶的原因。谢谢。
(I've been experiencing diarrhea recently, with my stool appearing as frothy yellow liquid. I suspect it might be due to the milk I've been drinking. Could this be the reason? Thank you.)

HuatuoGPT (w/ real data):
您好！根据你描述的症状，可能存在肠胃疾病的情况。此外，饮食习惯也会影响到这种情况发生。从中医角度来看，这主要与脾胃功能失调有关。因为脾胃相表里，所以需要通过健脾胃的方法进行治疗。同时，平常饮食也需注意清淡易消化。
(Hello! Based on the symptoms you described, there might be a potential issue with your gastrointestinal system. Moreover, dietary habits can also affect the occurrence of such symptoms. From a traditional Chinese medicine perspective, this is primarily related to a disorder of spleen and stomach functions, since they are closely linked. Therefore, treatment would involve methods to strengthen the spleen and stomach. Additionally, you should consider a light and easily digestible diet in your daily eating habits.)

Application of KD

VALLEY: VIDEO ASSISTANT WITH LARGE LANGUAGE MODEL ENHANCED ABILITY

Ruipu Luo[1,2*], Ziwang Zhao[1,3*], Min Yang[1*], Junwei Dong[1,4], Da Li[1],
Pengcheng Lu[1], Tao Wang[1], Linmei Hu[5], Minghui Qiu[1,†], Zhongyu Wei[2]
[1]ByteDance Inc. [2]Fudan University [4]Chongqing University
[3]Beijing University of Posts and Telecommunications
[5]Beijing Institute of Technology
{luoruipu,zhaoziwang,yangmin.priv,dongjunwei}@bytedance.com

Application of KD

Problem: jointing video and language understanding has not been widely explored

Method: adopt a pre-training-then-instructions-tuned pipeline to align visual and textual modalities and improve the instruction-following capability

Figure 1: Valley architecture.

Application of KD

Result: a highly effective video assistant that can understand complex video scenarios and provide descriptions in text.

Method	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	32.2	–	16.8	–	24.7	–
VideoChat	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA Adapter	54.9	3.1	43.8	2.7	34.2	2.7
Video LLaMA	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT	64.9	3.3	49.3	2.8	35.2	2.7
Valley-v1	**65.4**	**3.4**	45.7	2.5	42.9	3.0
Valley-v2	59.1	**3.4**	49.9	**3.0**	32.5	2.6
Valley-v3	60.5	3.3	**51.1**	2.9	**45.1**	**3.2**

Table 1: Zero-shot video QA result on MSVD-QA, MSRVTT-QA and ActivityNet-QA

Method	COR	DO	CU	TU	CON
VideoChat	2.23	2.50	2.53	1.94	2.24
LLaMA Adapter	2.03	2.32	2.30	1.98	2.15
Video LLaMA	1.96	2.18	2.16	1.82	1.79
Video-ChatGPT	2.40	**2.52**	2.62	1.98	2.37
Valley-v1	2.06	2.42	2.74	1.83	2.41
Valley-v2	2.35	2.13	2.85	1.99	2.17
Valley-v3	**2.43**	2.13	**2.86**	**2.04**	**2.45**

Table 2: Results on video-based text generation benchmark provided by Video-ChatGPT.

Method	Setup	BERT-F1
Flamingo	zero-shot	65.51
	zero-shot COT	58.23
MiniGPT4	zero-shot	65.81
	zero-shot COT	68.45
Valley	zero-shot	84.82
	zero-shot COT	**85.26**

Table 3: Results on Meme-Cap, the image metaphor understanding benchmark.

Leading performance in video Q&A

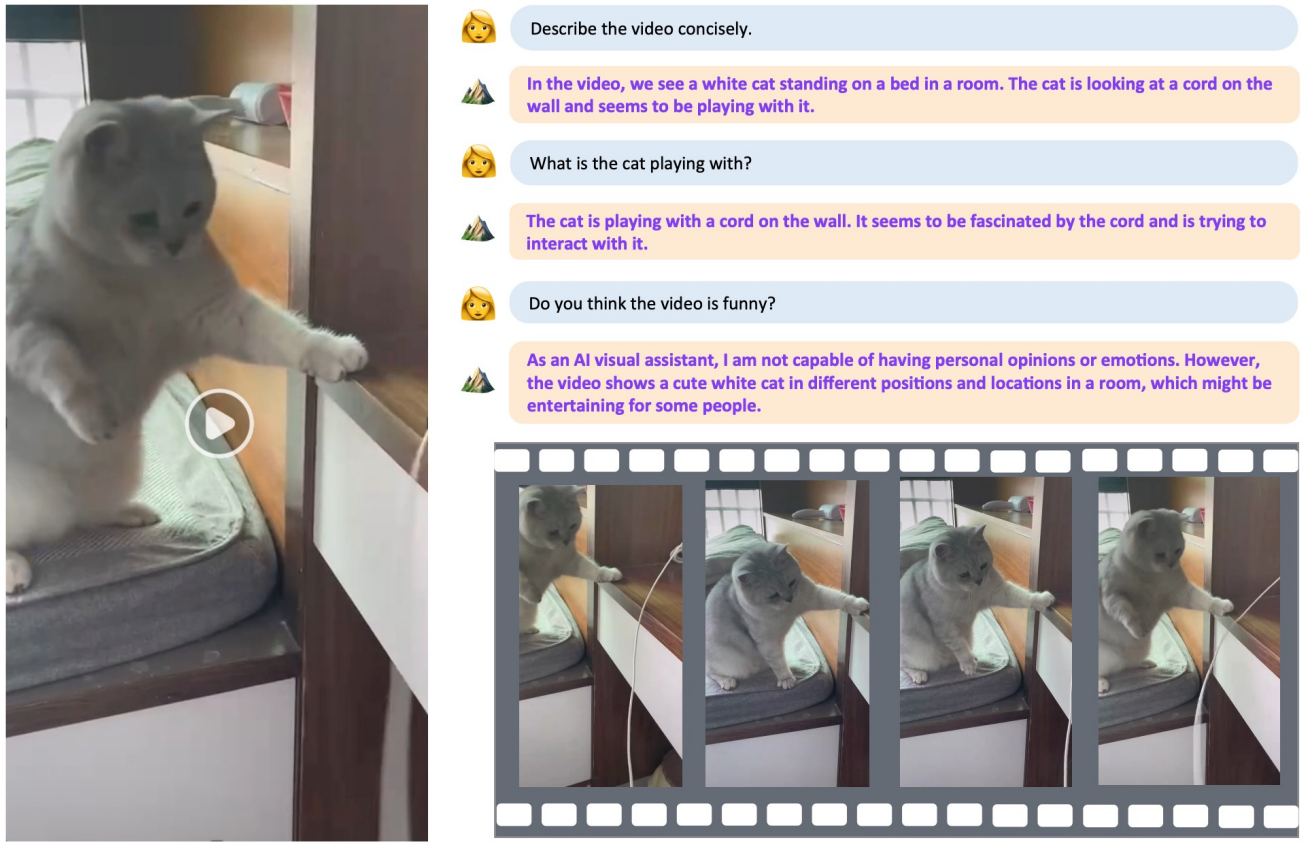


Figure 6: Video description, funny things identification, and recognition of interactions of the objects in the video.

Thanks!