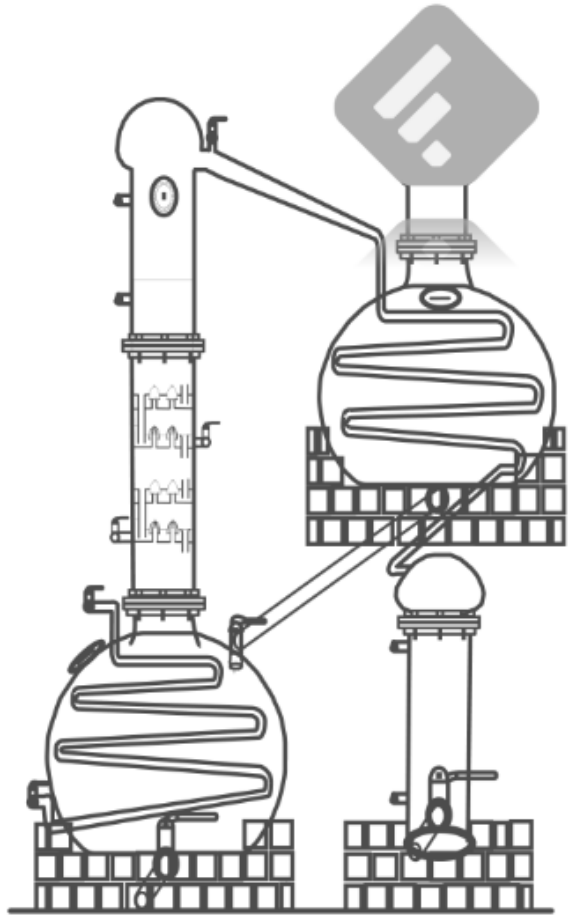
Knowledge Distillation

Градобоев Дмитрий, БПМИ171



Дистилляция знаний

Сложная задача

↓

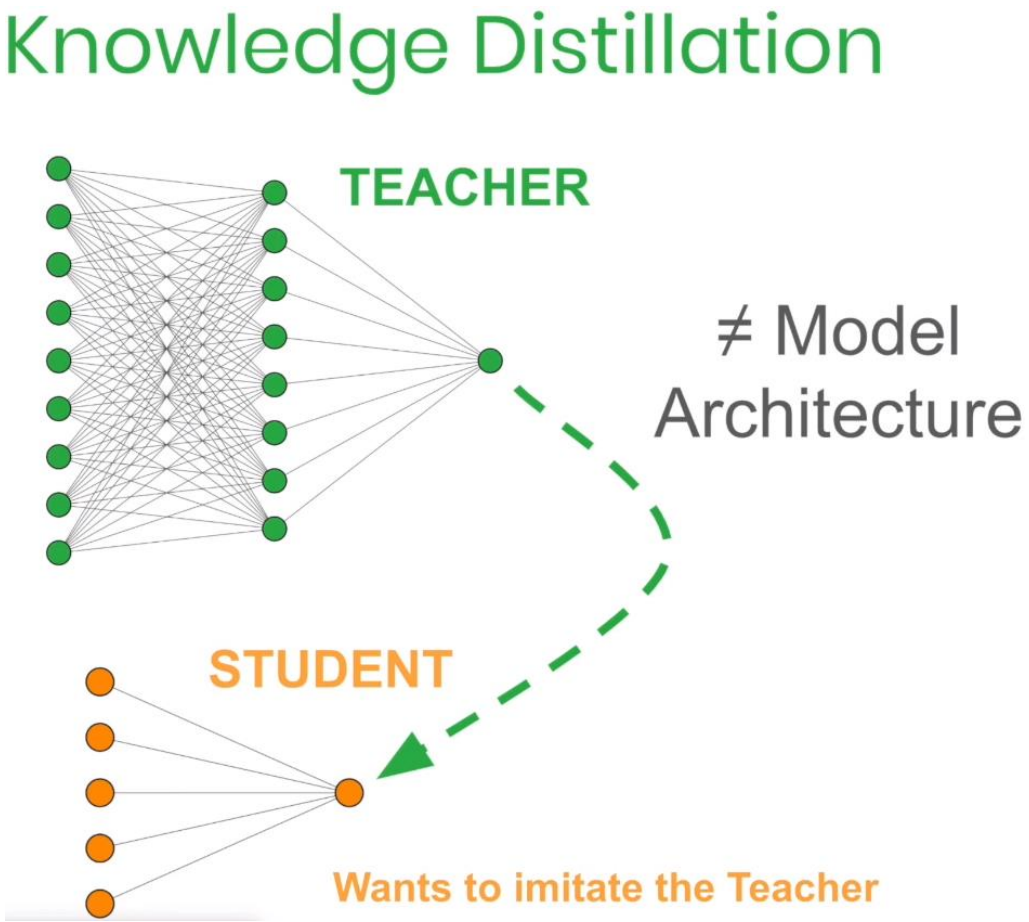
Большая сеть

↓

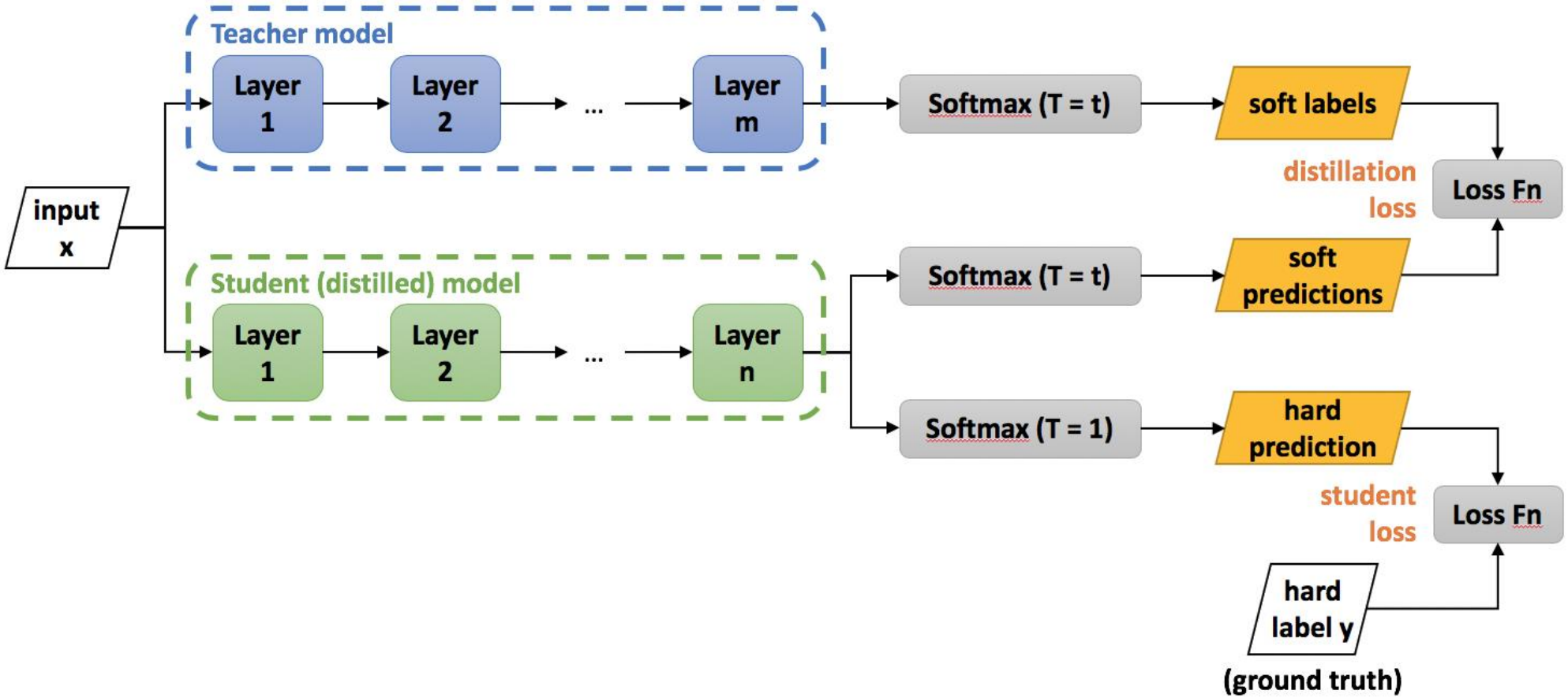
Большие вычислительные возможности и долгое время работы

↓

Невозможность запускать на слабых устройствах (например, на смартфонах)

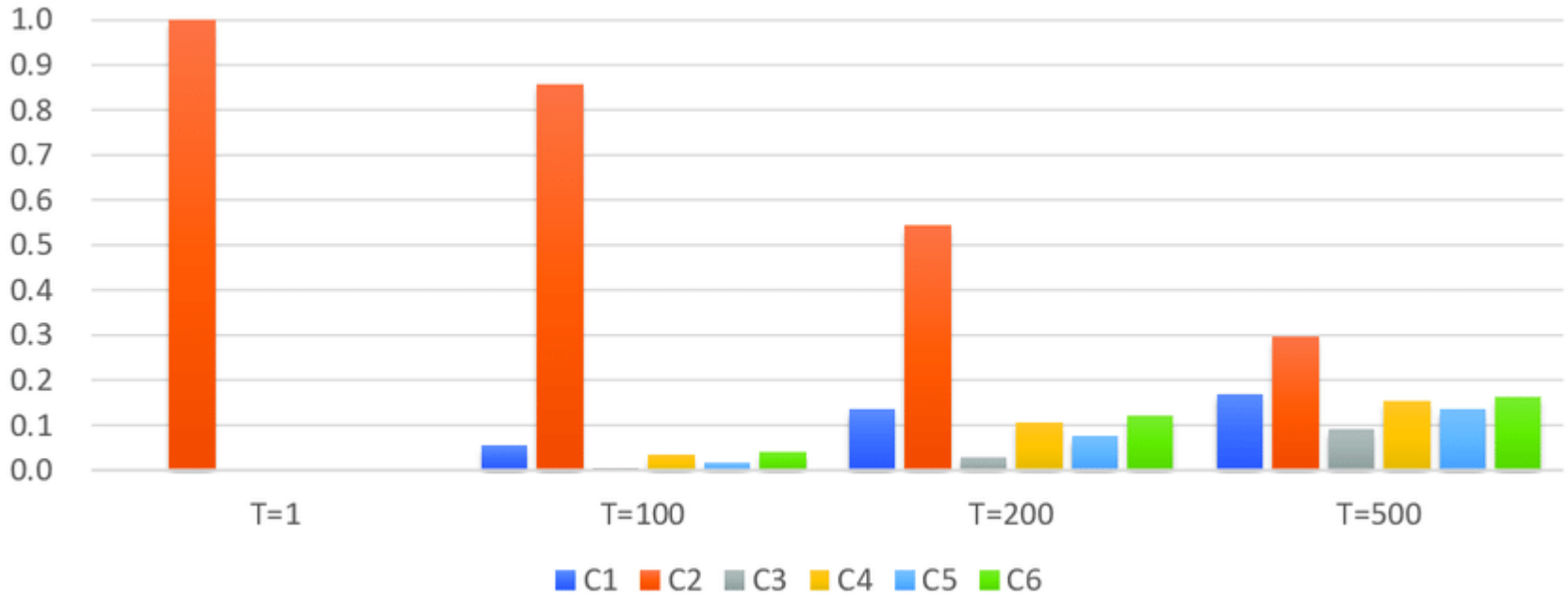


Алгоритм обучения



Softmax temperature

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}, \text{ где T - температура}$$



Обучение

$$\mathcal{L}(x; W) = \alpha \cdot \mathcal{H}(y, \sigma(Z_S|T = 1)) + \beta \cdot \mathcal{H}(\sigma(V_T|T = \tau), \sigma(Z_S|T = \tau))$$

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}(\frac{exp(z_i/T)}{\sum_j exp(z_j/T)} - \frac{exp(v_i/T)}{\sum_j exp(v_j/T)})$$

Если температура достаточна большая по сравнению с z и v, то:

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}(\frac{1+z_i/T}{N+\sum_j z_j/T} - \frac{1+v_i/T}{N+\sum_j v_j/T})$$

Если теперь предположить, что мы выравниваем логиты так, что $\sum_j z_j = \sum_j v_j = 0$, то:

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

MNIST

- Учитель: 2 слоя, каждый по 1200 юнитов, строгая регуляризация (dropout, weight constraints) – 67 ошибок
- Студент: 2 слоя, каждый по 800 юнитов, без регуляризации. – 146 ошибок
- Студент с дистилляцией: 2 слоя, каждый по 800 юнитов, без регуляризации. – 74 ошибки

Студент получает знания о данных, которые не видит при обучении, но необходимо скорректировать смещение.

Распознавание речи

Улучшение результатов, за счёт дистилляции от ансамбля сетей

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Дистилляция в качестве регуляризатора

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%

The JFT dataset

System	Conditional Test Accuracy	Test Accuracy
Baseline	43.1%	25.0%
+ 61 Specialist models	45.9%	26.1%

100 миллионов картинок, 15000 классов
Baseline – 6 месяцев обучения
Получаем прирост качества за счет дистилляции и ансамблирования

$$KL(p^g, q) + \sum_{m \in A_k} KL(p^m, q) \rightarrow_q min$$

On the Efficacy of Knowledge Distillation

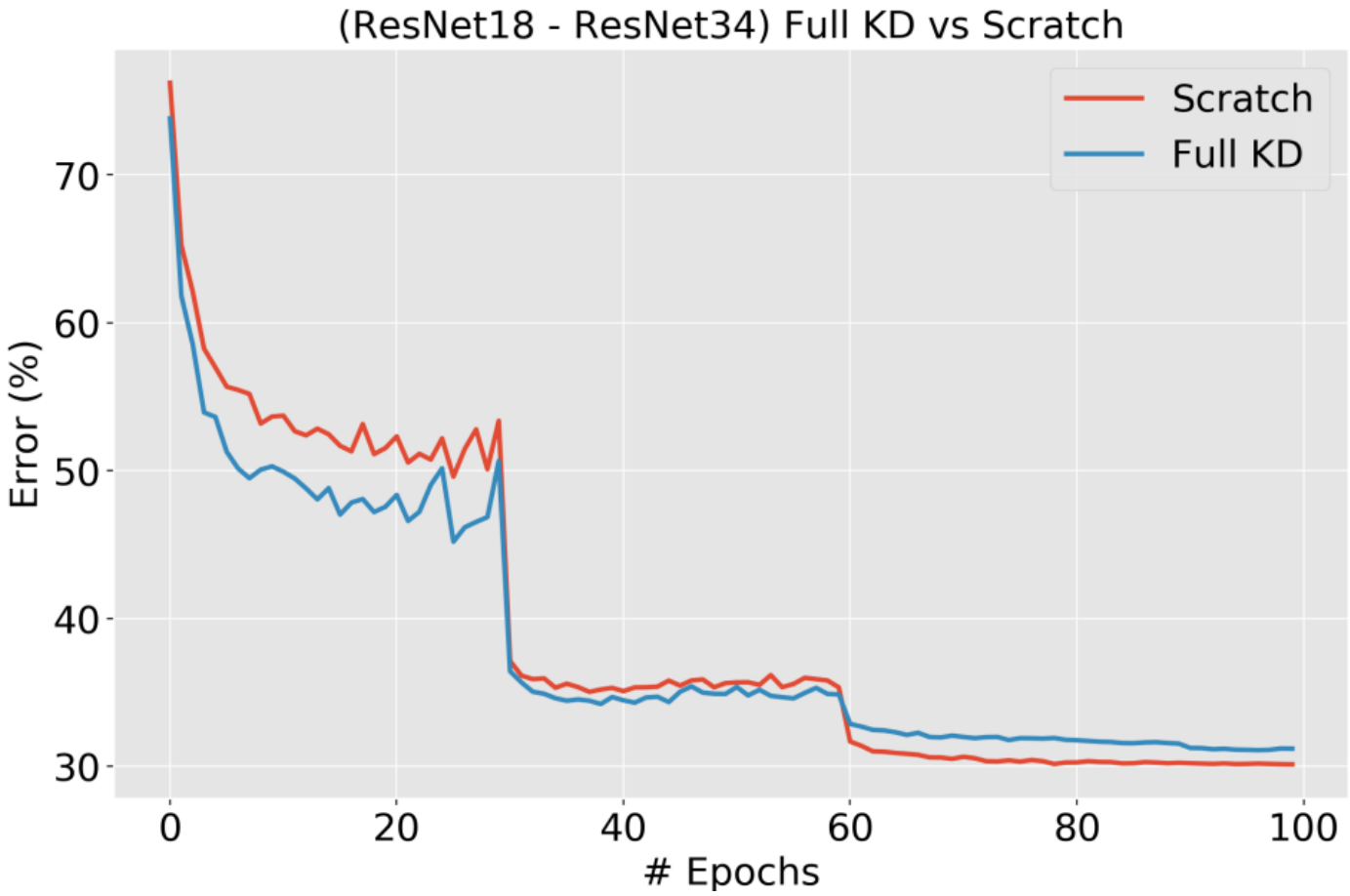
- Чем больше учитель, тем хуже качество студента

1. Учитель становится более уверенным и чётким в ответах.
2. Студент способен имитировать учителя, но качество не улучшается. (Проблемы с функцией потерь и метрикой)
3. Студент не может имитировать больших учителей.

Teacher	Teacher Error (%)	Student Error (%)
-	-	30.24
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

Student	Teacher	KD Error (%,Train)	KD Error (%,Test)
WRN28-1	WRN28-3	0.23	4.05
	WRN28-4	0.25	4.53
	WRN28-6	0.23	4.54
	WRN28-8	0.31	4.81
WRN16-1	WRN16-3	1.70	6.32
	WRN16-4	1.69	6.52
	WRN16-6	1.94	6.91
	WRN16-8	1.69	7.01

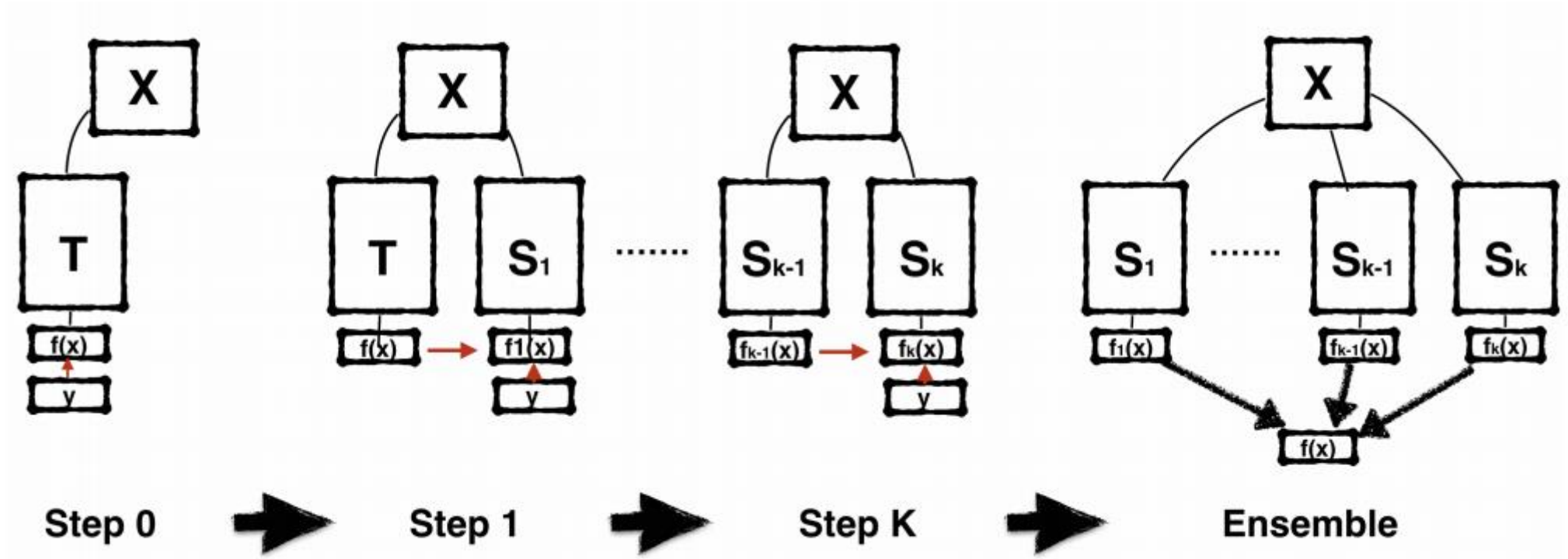
Лучший результат – отсутствие дистилляции



Early-stopped KD

Teacher	Top-1 Error (%, Test)	CE (Train)	KD (Train)	KD (Test)
ResNet18	30.57	0.146	2.916	3.358
ResNet18 (ES KD)	29.01	0.123	2.234	2.491
ResNet34	30.79	0.145	1.357	1.503
ResNet34 (ES KD)	29.16	0.123	2.359	2.582
ResNet50	30.95	0.146	1.553	1.721
ResNet50 (ES KD)	29.35	0.124	2.659	2.940

Последовательный KD



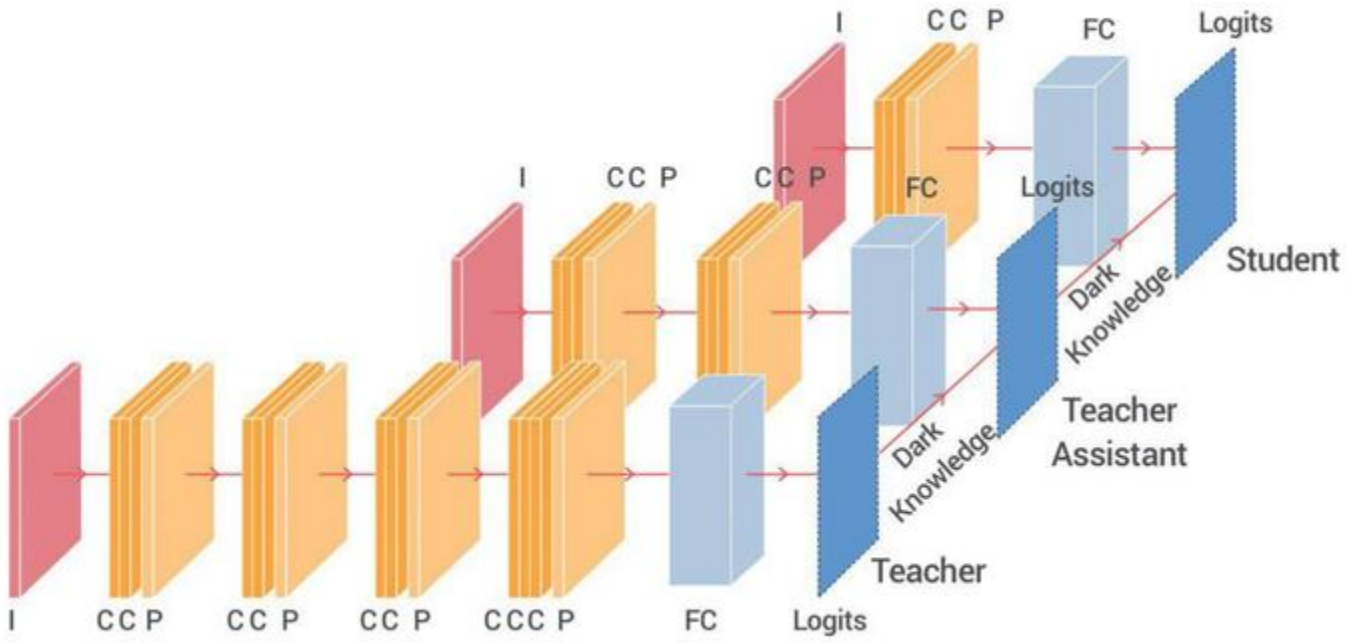
Model	# Params	Method	Last Gen. Err.	All Gen. Ensmeble Err.	Scratch Err.	Scratch Ensemble Err.
ResNet8	0.07M	AT+KD	13.469	12.786	12.569*	**10.176**
ResNet14	0.17M	AT+KD	9.226	8.653	9.078*	**6.675**
WRN16-2	0.69M	KD	6.101	5.181	6.428	**4.865**
WRN16-2	0.69M	AT+KD	5.696	5.310	6.418	**5.003**

Результаты экспериментов

Method	Teacher	Top-1 Error (%)
Scratch	-	30.24
Full KD [12]	ResNet18	30.57
Full KD [12]	ResNet34	30.79
Full KD [12]	ResNet50	30.95
Seq. Full KD [23]	3 Gen.	30.12*
Seq. Full KD [23]	6 Gen.	29.6*
KD+ONE [17]	3 Branches	29.45 ±0.23*
Full KD + AT [26]	ResNet34	30.94
Full KD + AT [26]	ResNet34	29.3*
ESKD	ResNet18	29.01
ESKD	ResNet34	29.16
ESKD	ResNet50	29.35
ESKD	ResNet152	29.45
ESKD	ResNet34 (50)	29.02
ESKD	ResNet50 (35)	29.05
ESKD	ResNet152 (35)	29.26
Seq. ESKD	L →S→S	29.41
Seq. ESKD	M→S→S	29.35
Seq. ESKD	S →S→S	29.15
ESKD + AT	ResNet34	28.84
ESKD + AT	ResNet34 (50)	**28.61**

Method	Teacher	Top-1 Error (%)	Top-5 Error (%)
Scratch	-	30.24	10.92
Full KD	ResNet18	30.75	11.11
Full KD	ResNet50	30.98	10.20
Full KD	ResNet152	31.27	11.59
ESKD	ResNet18	29.00	9.91
ESKD	ResNet50	29.00	9.76
ESKD	ResNet50 (35)	**28.89**	**9.76**

Teacher assistant KD



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

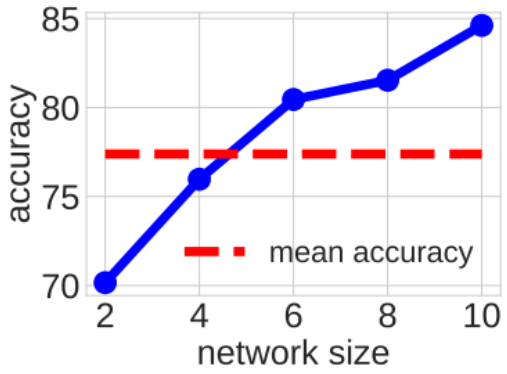
Model	Dataset	TA=56	TA=32	TA=20	TA=14
ResNet	CIFAR-10	88.70	88.73	88.90	88.98
	CIFAR-100	61.47	61.55	61.82	61.5

Подбор размера ассистента

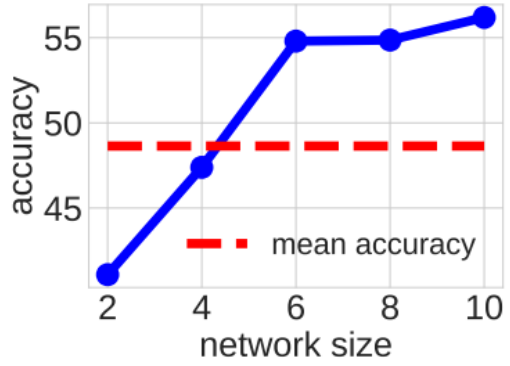
Берем среднее не по размеру сети, а по качеству.

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

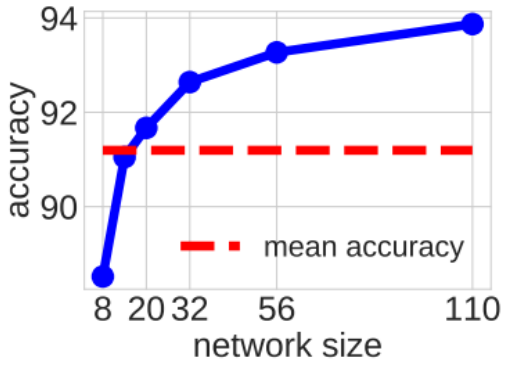
Model	Dataset	TA=56	TA=32	TA=20	TA=14
ResNet	CIFAR-10	88.70	88.73	88.90	88.98
	CIFAR-100	61.47	61.55	61.82	61.5



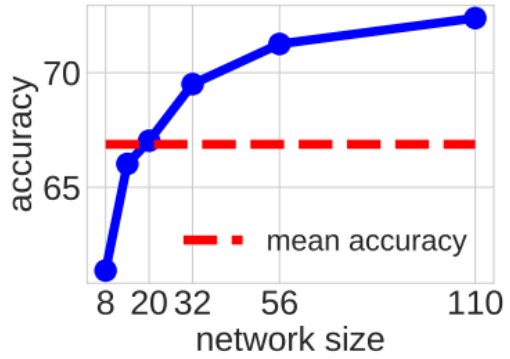
(a) CIFAR-10, Plain CNN



(b) CIFAR-100, Plain CNN



(c) CIFAR-10, ResNet



(d) CIFAR-100, ResNet

Итоговые результаты

Student	NOKD	BLKD	FITNET	AT	FSP	BSS	MUTUAL	TAKD
ResNet8	86.02	86.66	86.73	86.86	87.07	87.32	87.71	**88.01**
Resnet14	89.11	89.75	89.72	89.84	89.92	90.34	90.54	**91.23**

Теоретическое доказательство

$$R(f_s) - R(f_r) \leq O\left(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{sr}}}\right) + \epsilon_{sr}$$

Из теории Вапника-Червоненкиса (1998)

$$O\left(\frac{|\mathcal{F}_t|_C}{n^{\alpha_{tr}}} + \frac{|\mathcal{F}_a|_C}{n^{\alpha_{at}}} + \frac{|\mathcal{F}_s|_C}{n^{\alpha_{sa}}}\right) + \epsilon_{tr} + \epsilon_{at} + \epsilon_{sa}$$

$$\leq O\left(\frac{|\mathcal{F}_t|_C}{n^{\alpha_{tr}}} + \frac{|\mathcal{F}_s|_C}{n^{\alpha_{st}}}\right) + \epsilon_{tr} + \epsilon_{st}$$

$$\leq O\left(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{sr}}}\right) + \epsilon_{sr}.$$

Вопросы

1. В чём основная идея подхода дистиляции знаний, для чего это нужно? Нарисуйте схему обучения модели и поясните её.

2. Приведите два любых примера KD метода и для каждого поясните преимущества использования KD.

3. Как влияет размер архитектуры на качество модели при KD? Как это можно объяснить и можно ли исправить?

4. Как устроен метод TAKD и почему он работает лучше, чем BLKD и NOKD?

Источники

- https://arxiv.org/pdf/1503.02531.pdf
- https://arxiv.org/pdf/1902.03393.pdf
- https://arxiv.org/pdf/1910.01348.pdf