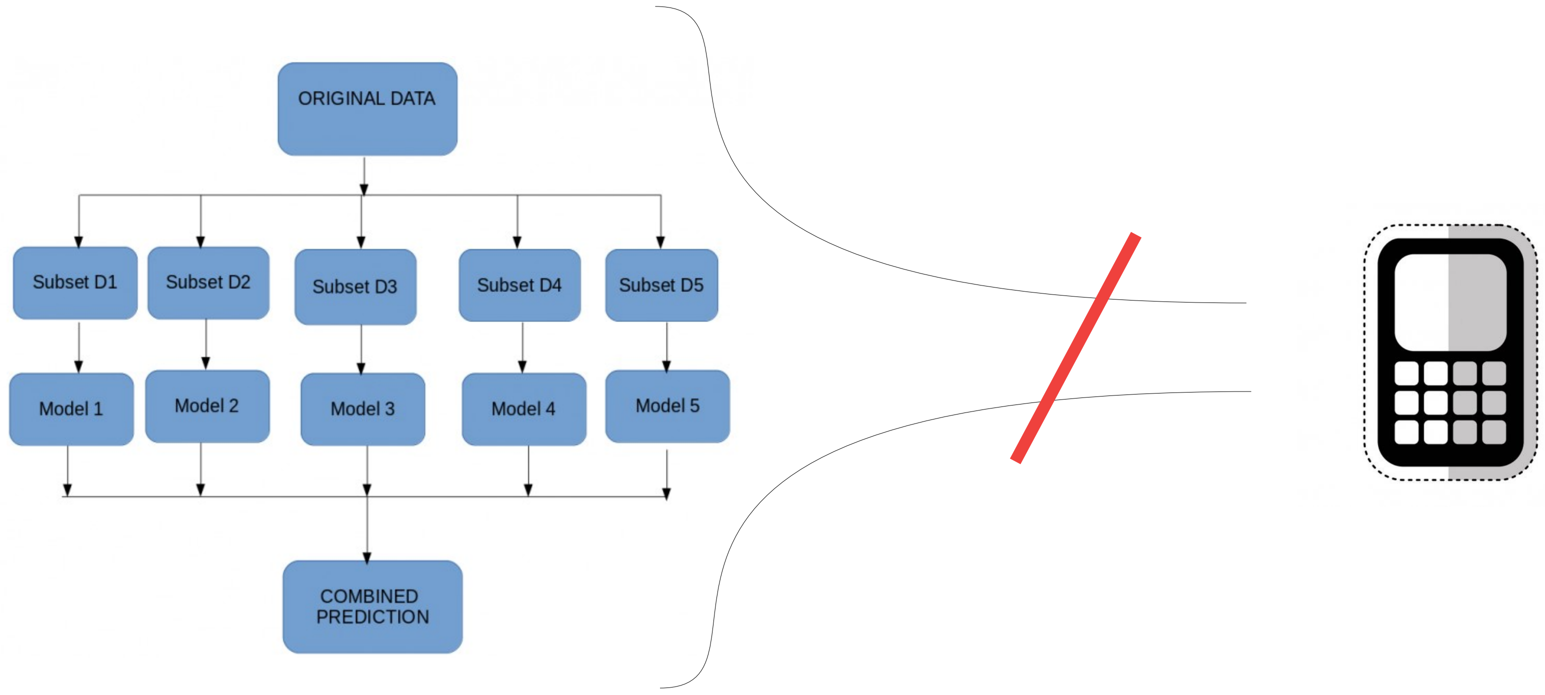
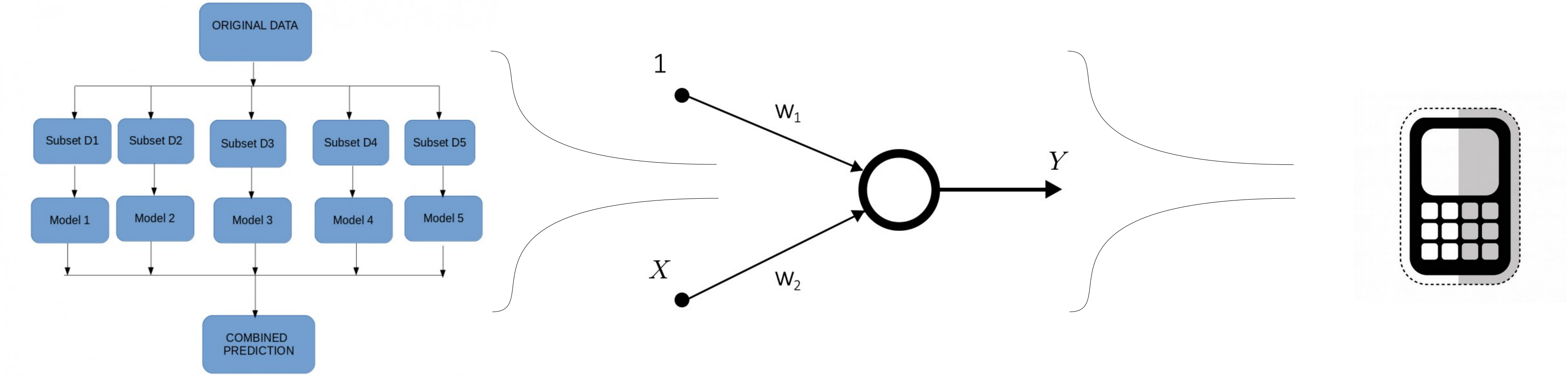
Distilling The Knowledge

Кузнецов Дмитрий
БПМИ171

Мотивация


ORIGINAL DATA
Subset D1
Subset D2
Subset D3
Subset D4
Subset D5
Model 1
Model 2
Model 3
Model 4
Model 5
COMBINED PREDICTION

Мотивация



ORIGINAL DATA
Subset D1
Subset D2
Subset D3
Subset D4
Subset D5
Model 1
Model 2
Model 3
Model 4
Model 5
COMBINED PREDICTION
1
X
w_1
w_2
Y

Перенос знаний

Тяжелая (Teacher)	**Легкая (Student)**
$\mu_H : \mathbf{X} \times \mathbf{Y} \to \mathbf{F}_H$	$\mu_L : \mathbf{X} \times \mathbf{Y} \to \mathbf{F}_L$

Перенос знаний

Тяжелая (Teacher)	**Легкая (Student)**
$\mu_H : \mathbf{X} \times \mathbf{Y} \to \mathbf{F}_H$	$\mu_L : \mathbf{X} \times \mathbf{Y} \to \mathbf{F}_L$

<u>*soft targets*</u>

<u>*hard targets*</u>

$$\mu'_L : \mathbf{X} \times \mathbf{F}_H \to \mathbf{F}_L$$

Например, можно усреднить распределения моделей в ансамбле для получения новых ц.п.

Больше информации, градиент между точками имеет меньшую «дисперсию»

Формально

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

z_i- логит iго класса

T- параметр температуры. По умолчанию равен 1

Больше T, более сглаженным получается выходное распределение

Формально

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

z_i- логит iго класса

T- параметр температуры. По умолчанию равен 1

Больше T, более сглаженным получается выходное распределение

$z^s,\ z^t$ - вектор логитов студента и учителя, соответственно

$q^{s,T},\ q^{t,T}$ - вектор softmax студента и учителя, соответственно, при температуре T

y - истинные лейблы (hard targets)

$$\mathcal{L}_H(x) = \mathcal{H}(q^{s,1}(x),\ y(x)) = \mathcal{H}(softmax(z^s(x)),\ y(x))$$

$$\mathcal{L}_S(x) = T^2\mathcal{H}(q^{t,T}(x),\ q^{s,T}(x)) = -T^2 \sum_k q_i^{t,T}(x) \log q_i^{s,T}(x)$$

Улучшение

$z_s,\ z_t$ - вектор логитов студента и учителя, соответственно

$q_s^T,\ q_t^T$ - вектор softmax студента и учителя, соответственно, при температуре T

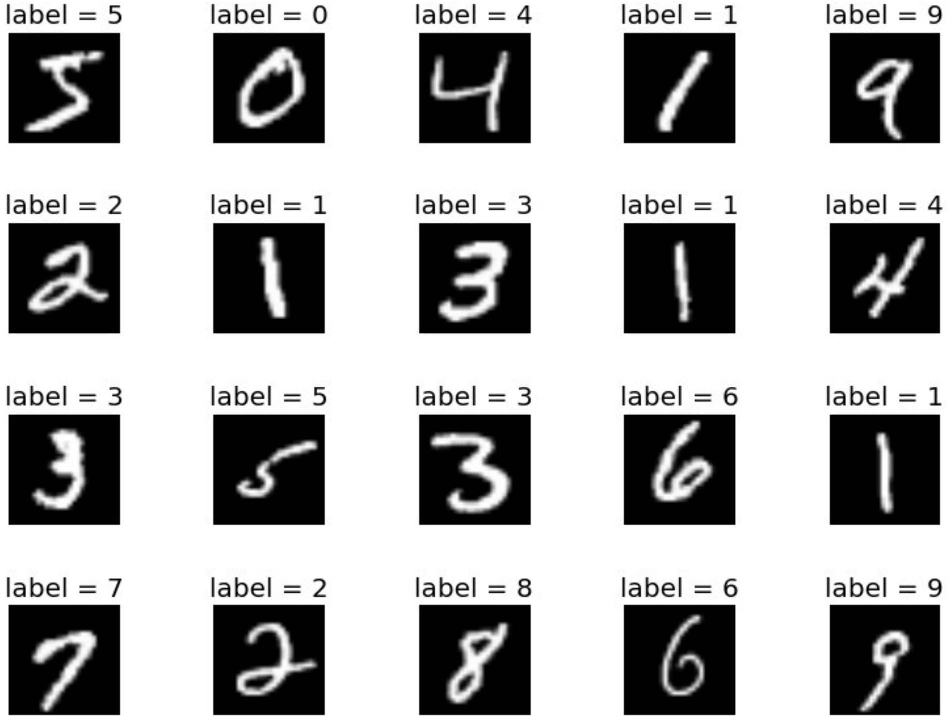
y - истинные лейблы (hard targets)

$$\mathcal{L}_H(x) = \mathcal{H}(softmax(z^s(x)),\ y(x))$$

$$\mathcal{L}_S(x) = T^2\mathcal{H}(q^{t,T}(x),\ q^{s,T}(x))$$

$$\mathcal{L}_{student} = \lambda\mathcal{L}_H + (1-\lambda)\mathcal{L}_S$$

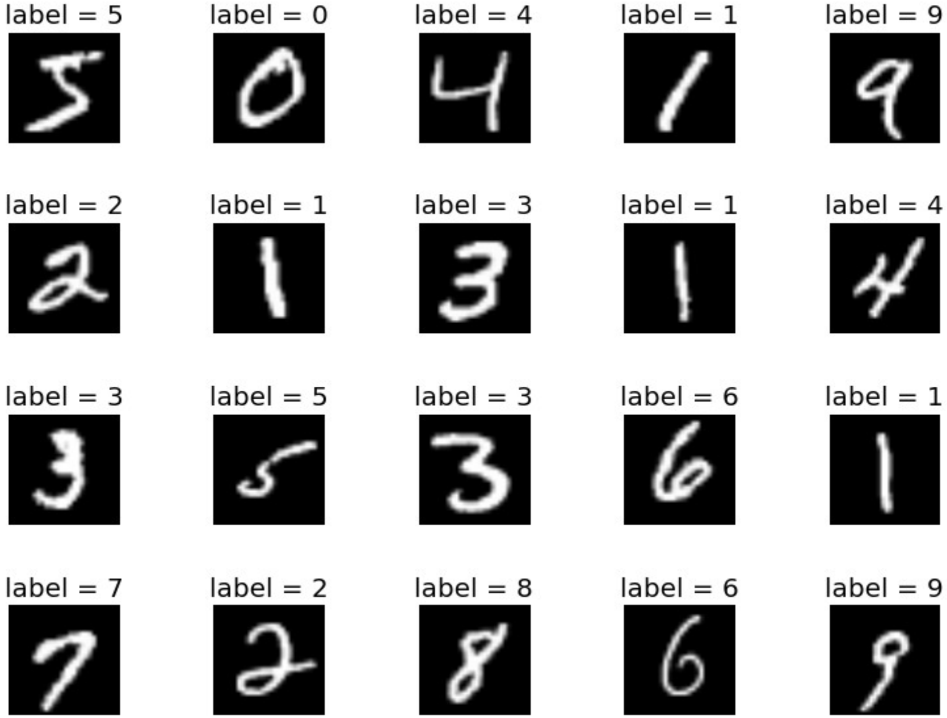
Поясняющий пример



Передаем модели больше информации.

По факту, показываем не просто класс, а дополнительную информацию о похожести тех или иных цифр на объекты других классов.

Поясняющий пример

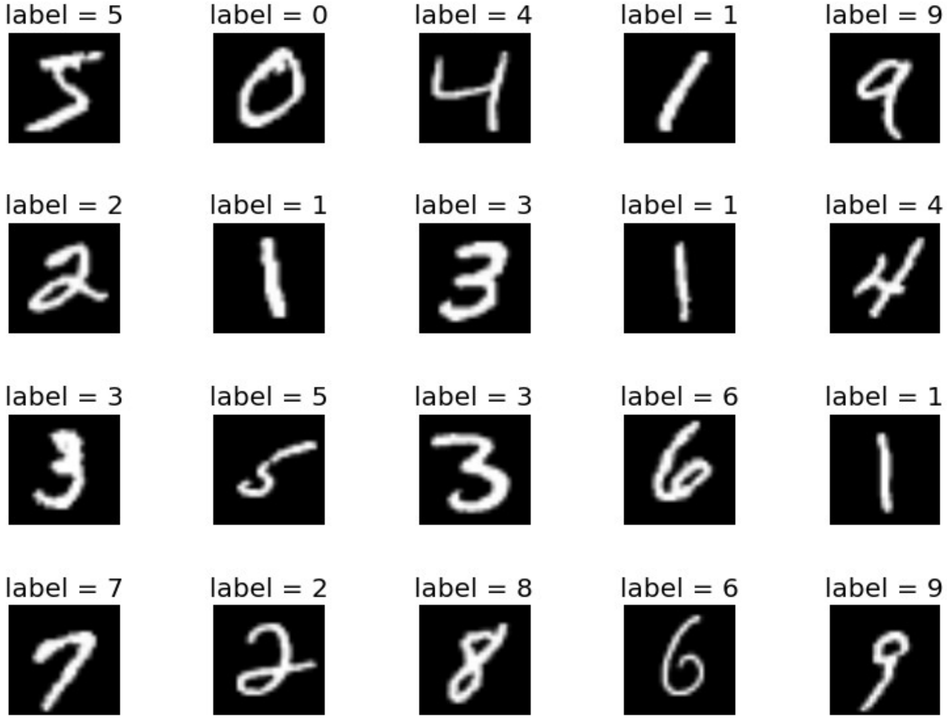


Передаем модели больше информации.

По факту, показываем не просто класс, а дополнительную информацию о похожести тех или иных цифр на объекты других классов.

Но если мы даем в качестве ц.п. просто вероятности, то эта дополнительная информация практически не вносит вклад в cross-entropy, т. к. обычно вероятности очень близки к нулю.

Поясняющий пример



Чтобы решить эту проблему, можно приближать логиты вместо вероятностей.

Соотнесение логитов

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} = \frac{1}{T}\left(q_i^{s,T} - q_i^{t,T}\right)$$

Соотнесение логитов

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} = \frac{1}{T}\left(q_i^{s,T} - q_i^{t,T}\right) = \frac{1}{T}\left(\frac{\exp^{z_i^s/T}}{\sum_j \exp^{z_j^s/T}} - \frac{\exp^{z_i^t/T}}{\sum_j \exp^{z_j^t/T}}\right)$$

Соотнесение логитов

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} = \frac{1}{T}(q_i^{s,T} - q_i^{t,T}) = \frac{1}{T}\left(\frac{\exp^{z_i^s/T}}{\sum_j \exp^{z_j^s/T}} - \frac{\exp^{z_i^t/T}}{\sum_j \exp^{z_j^t/T}}\right)$$

При достаточно большом T, можем разложить экспоненты в ряд Тейлора:

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} \simeq \frac{1}{T}\left(\frac{1 + z_i^s/T}{N + \sum_j z_j^s/T} - \frac{1 + z_i^t/T}{N + \sum_j z_j^t/T}\right)$$

Соотнесение логитов

При достаточно большом T, можем разложить экспоненты в ряд Тейлора:

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} \simeq \frac{1}{T}\left(\frac{1+z_i^s/T}{N+\sum_j z_j^s/T} - \frac{1+z_i^t/T}{N+\sum_j z_j^t/T}\right)$$

Из предположения о том, что: $\sum\limits_j z_j^s = \sum\limits_j z_j^t = 0$

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} \simeq \frac{1}{NT^2}(z_i^s - z_i^t)$$

Соотнесение логитов

При достаточно большом T, можем разложить экспоненты в ряд Тейлора:

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} \simeq \frac{1}{T}\left(\frac{1 + z_i^s/T}{N + \sum_j z_j^s/T} - \frac{1 + z_i^t/T}{N + \sum_j z_j^t/T}\right)$$

Из предположения о том, что: $\sum\limits_j z_j^s = \sum\limits_j z_j^t = 0$

$$\frac{\partial \mathcal{L}_S}{\partial z_i^s} \simeq \frac{1}{NT^2}(z_i^s - z_i^t)$$

Значит, при достаточно больших T, задача эквивалентна:

$$\frac{1}{2}(z_i^s - z_i^t)^2 \to \min$$

Соотнесение логитов

Это приводит нас к выводу о том, что нам не требуется приближать логиты. Достаточно обучаться при высокой температуре.

Также, при достаточно низких температурах при дистилляции уделяется значительно меньше внимания логитам, которые сильно меньше среднего (« 0).

Можем использовать в лучшую сторону. Убирает шумовые куски распределения, после дистилляции увеличиваем потенциал обобщающей способности.

Результаты (MNIST)

Модель	Кол-во ошибок на тесте
single	67
single(s, hard)	146
single(s, soft)	74

Single: 2 скрытых слоя по 1200 ReLU
+ dropout
+ weight-constraints

Single(s): 2 скрытых слоя по 800 ReLU
- без какой-либо регуляризации

Результаты (MNIST)

Модель	Кол-во ошибок на тесте
single	67
single(s, hard)	146
single(s, soft)	74

Single: 2 скрытых слоя по 1200 ReLU
+ dropout
+ weight-constraints

Single(s): 2 скрытых слоя по 800 ReLU
- без какой-либо регуляризации

Кол-во нейронов на слое (small)	Оптимальное значение температуры
300	8 (и выше)
30	2.5 - 4

Подтверждает необходимость низких температур при высокой разнице архитектур студента и учителя

Результаты (ASR)

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

Baseline: 8 скрытых слоев по 2560 ReLU
- 14000 классов (с softmax)

Data: 2000 часов разговорного Английского (источник не уточняется)

Результаты (ASR)

Система и обучающее множество	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% training set)	63.4%	58.9%
Baseline (3% training set)	67.3%	44.5%
Soft targets (3% training set)	65.4%	57.0%

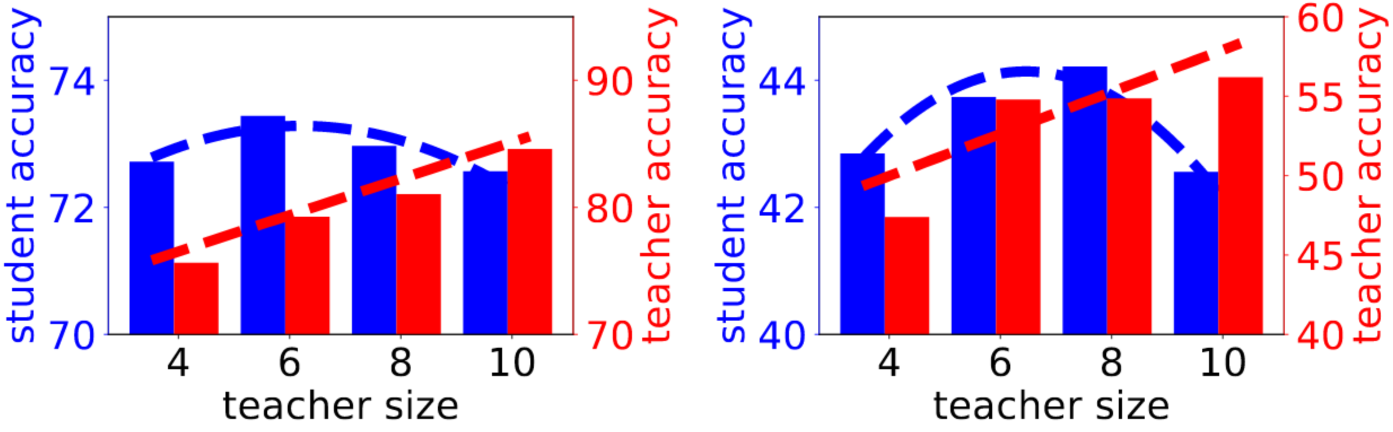
При исользовании soft targets мы снимажем риск переобучения, необходимое кол-ва данных для обучения.

Пропасть между моделями

Важный вопрос, который стоит задать: *что происходит, если обобщающие способности (сложность) архитектур учителя и студент сильно разнятся?*

Пропасть между моделями

Важный вопрос, который стоит задать: *что происходит, если обобщающие способности (сложность) архитектур учителя и студент сильно разнятся?*

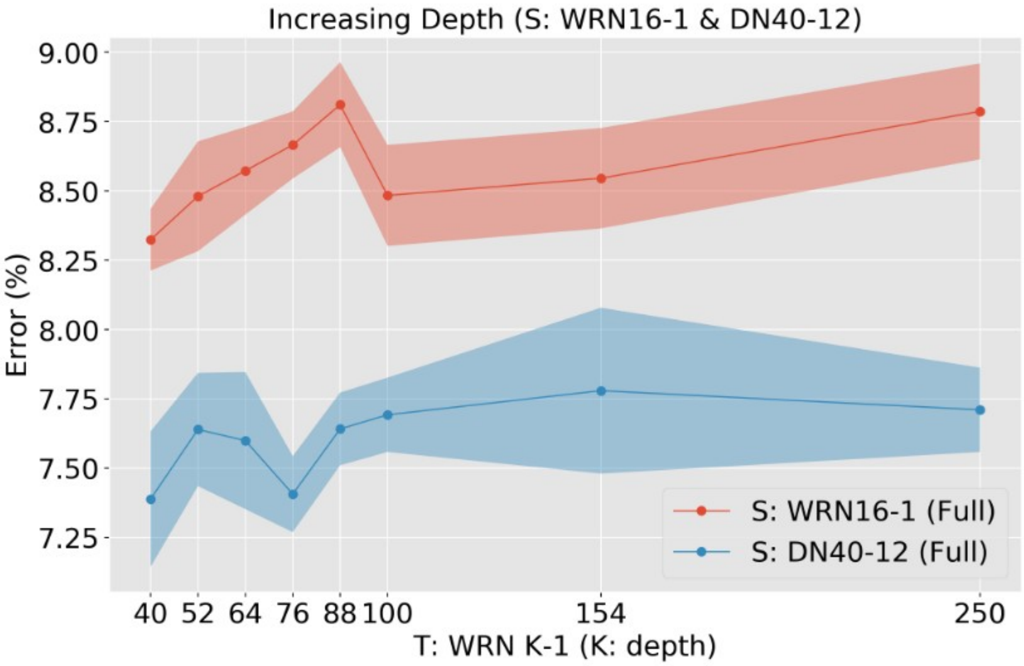


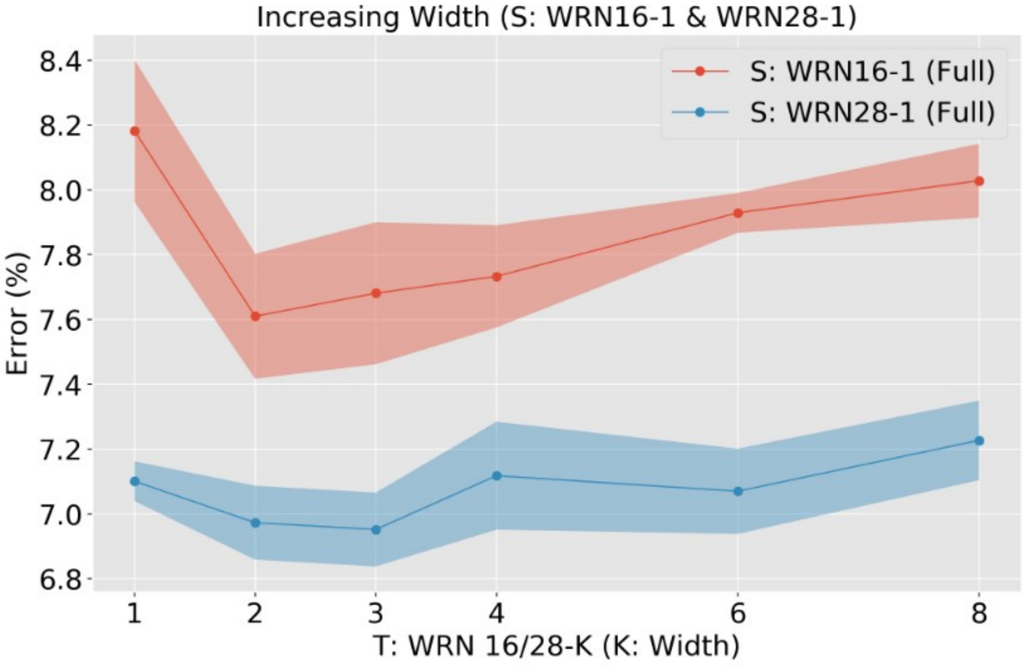
a) CIFAR-10 b) CIFAR-100

Student: *2 layers CNN*
Teacher: *Size layers CNN*

+ max pool
+ FC

Пропасть между моделями





Dataset: CIFAR-10

Может быть дело в метрике?

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

Student: ResNet18
Dataset: ImageNet

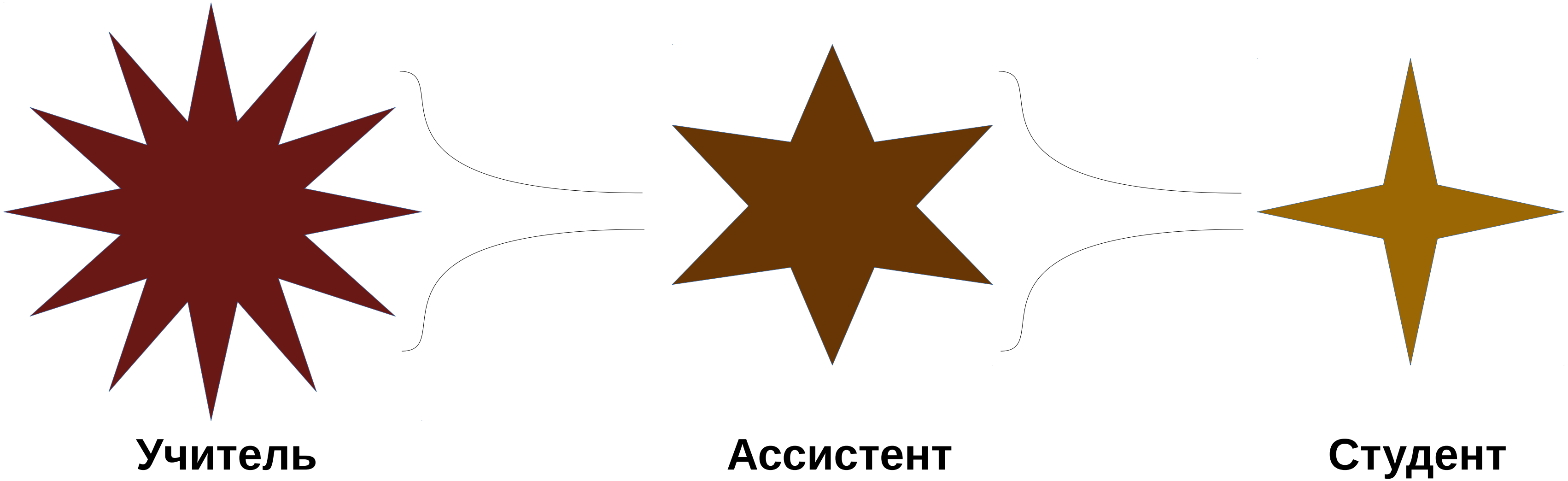
Первая строка — остутствие дистилляции

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

Dataset: CIFAR-10

Teacher Assistant Knowledge Distillation (TAKD)

Заметим, что мы не можем менять размер ни учителя, ни студента. Иначе мы ограничиваем потенциал применимости дистилляции.



Здесь схематично изображена сложность моделей. Больше вершин — больше параметров.

TAKD Experiments

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

NOKD: малая модель без дистилляции

BLKD: классическая дистилляция

TAKD: дистилляция с ассистентом

Метрика — Accuracy

Конфигурация:
CIFAR CNN: S=2, TA=4, T=10
CIFAR ResNet: S=8, TA=20, T=110
ImageNet ResNet: S=14, TA=20, T=50

Здесь подразумевается кол-во сверточных слоев

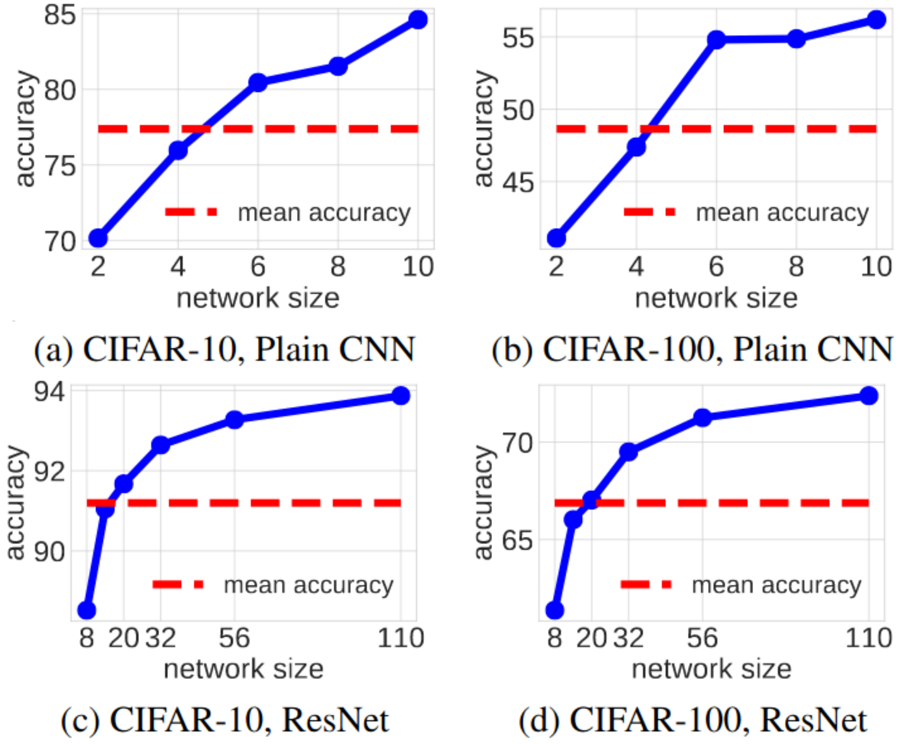
Как подобрать размер TA

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

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

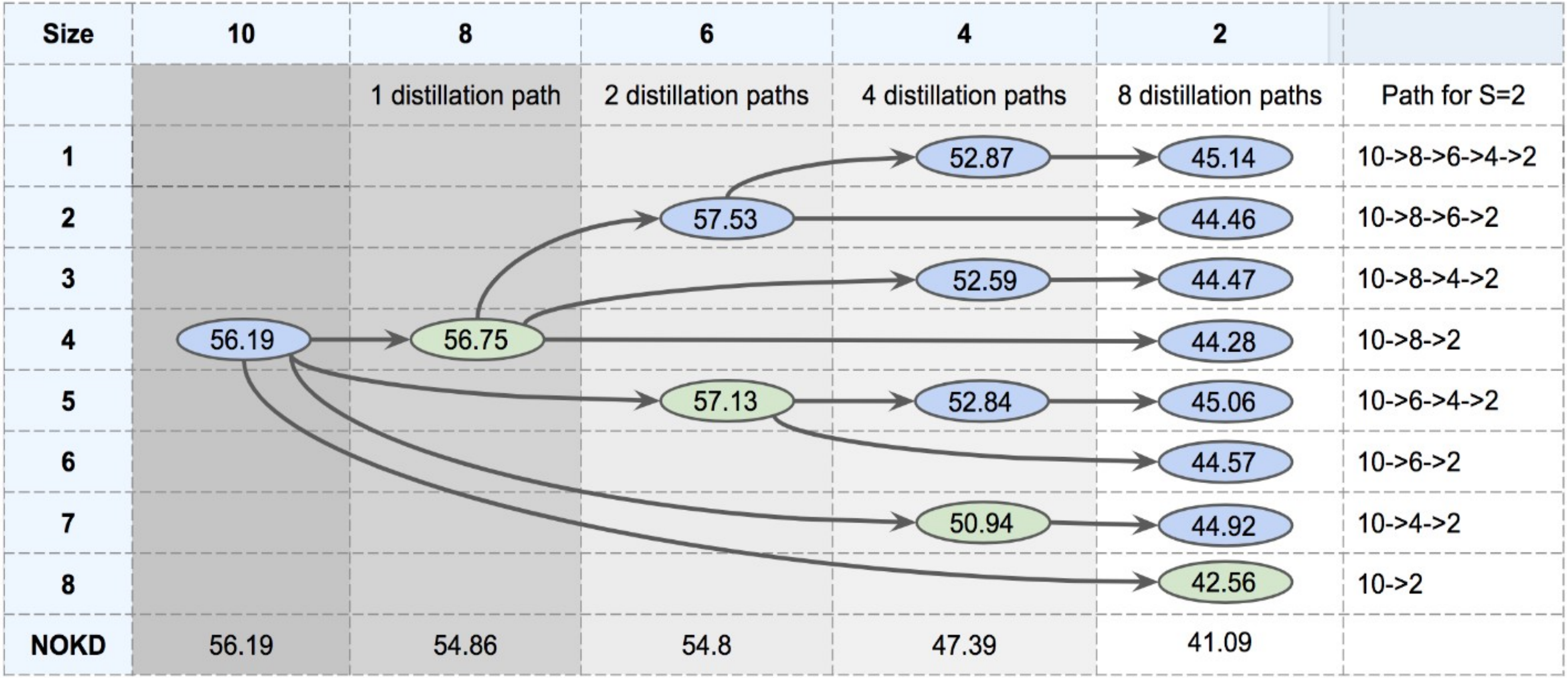


TA train accuracy

Лучшие показатели у TA, accuracy которого приближена к среднему accuracy учителя и студента.

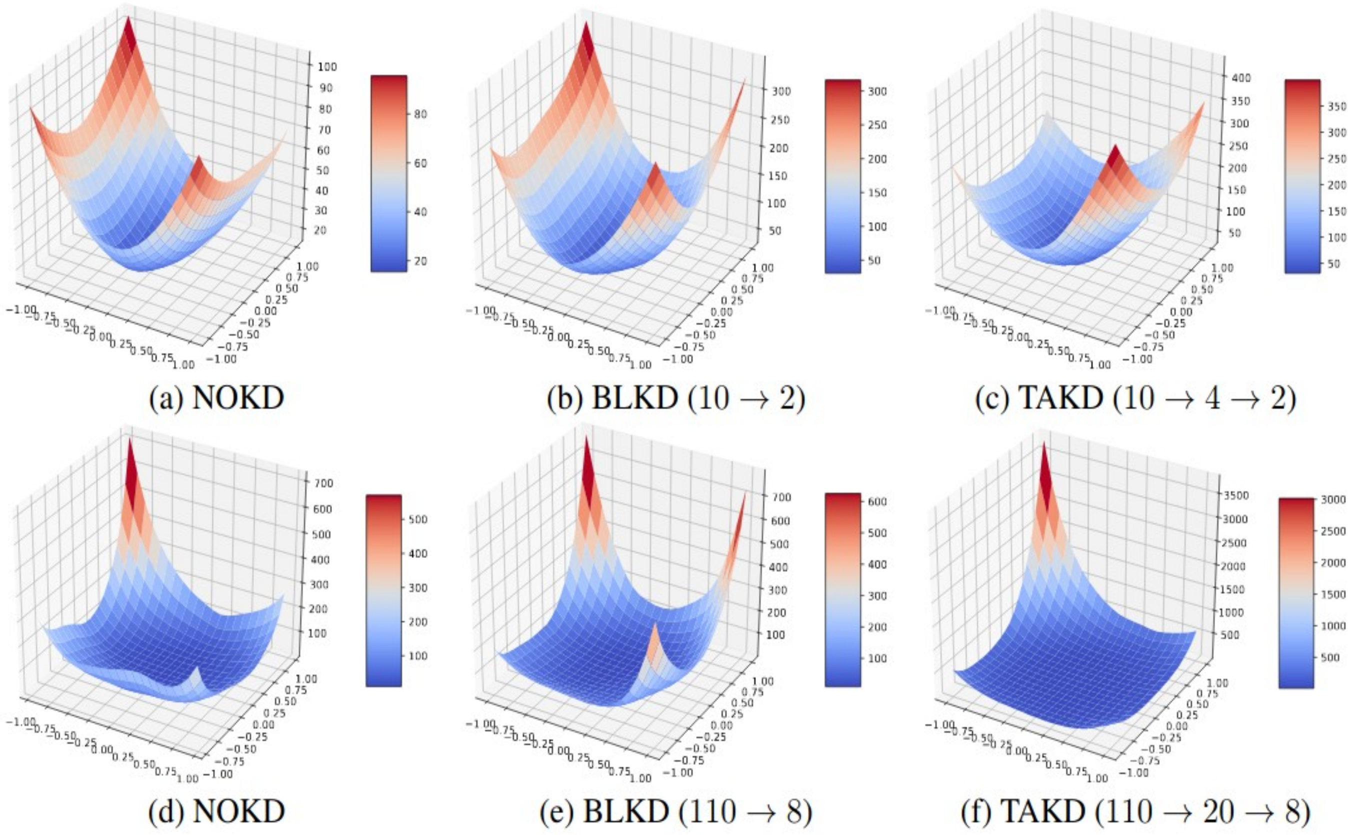
Почему только один ассистент?

T=10. CNN CIFAR-100



Все еще всякий путь лучше, чем классическая дистилляция без ассистентов в целом.

Поверхности функции ошибок



(a) NOKD (b) BLKD (10 → 2) (c) TAKD (10 → 4 → 2)

(d) NOKD (e) BLKD (110 → 8) (f) TAKD (110 → 20 → 8)

Окрестность локального минимума

Сверху: CNN, S=2

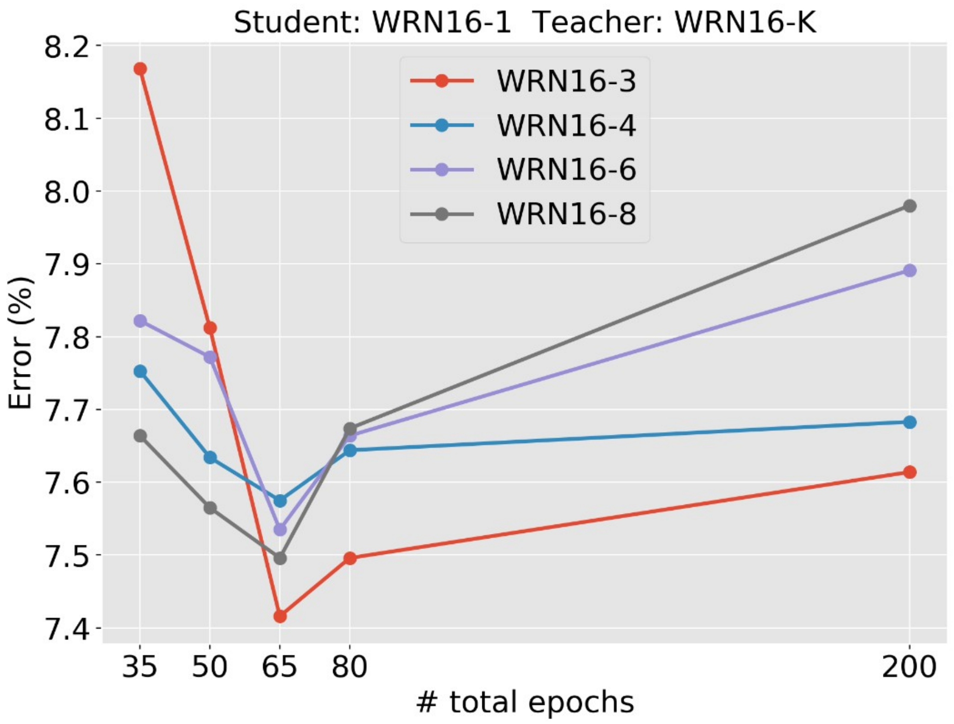
Снизу: ResNet, S=8

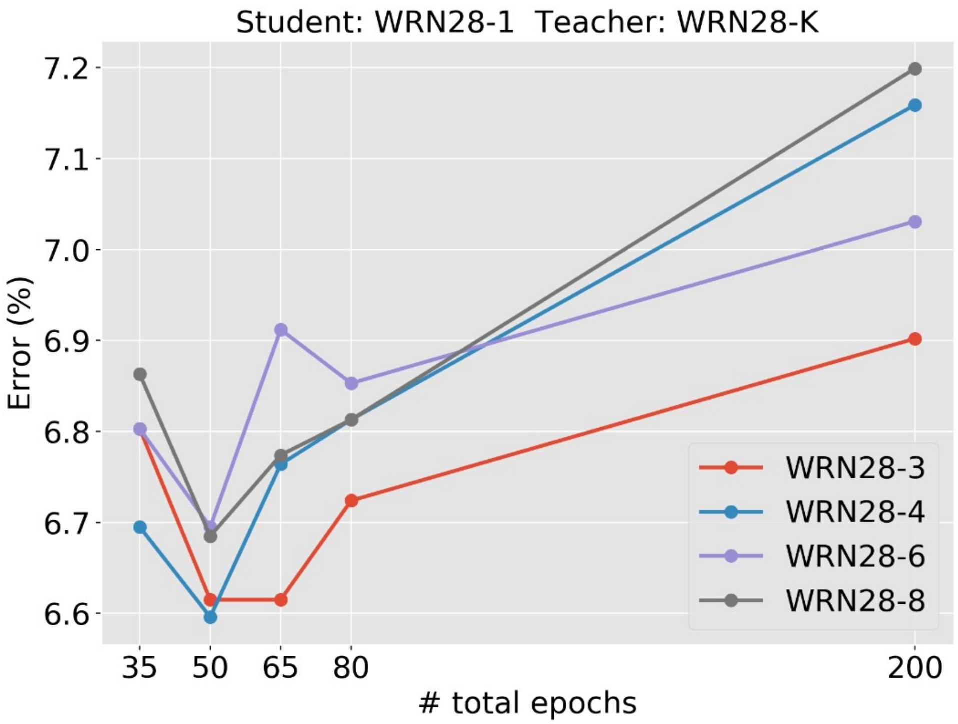
Сравнение с другими подходами

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**

FITNET — дистилляция на промежуточные слои (не последние)
AT — перенос активаций CNN
FSP — перенос с помощью FSP матриц
BSS — учит студента на состязательных примерах

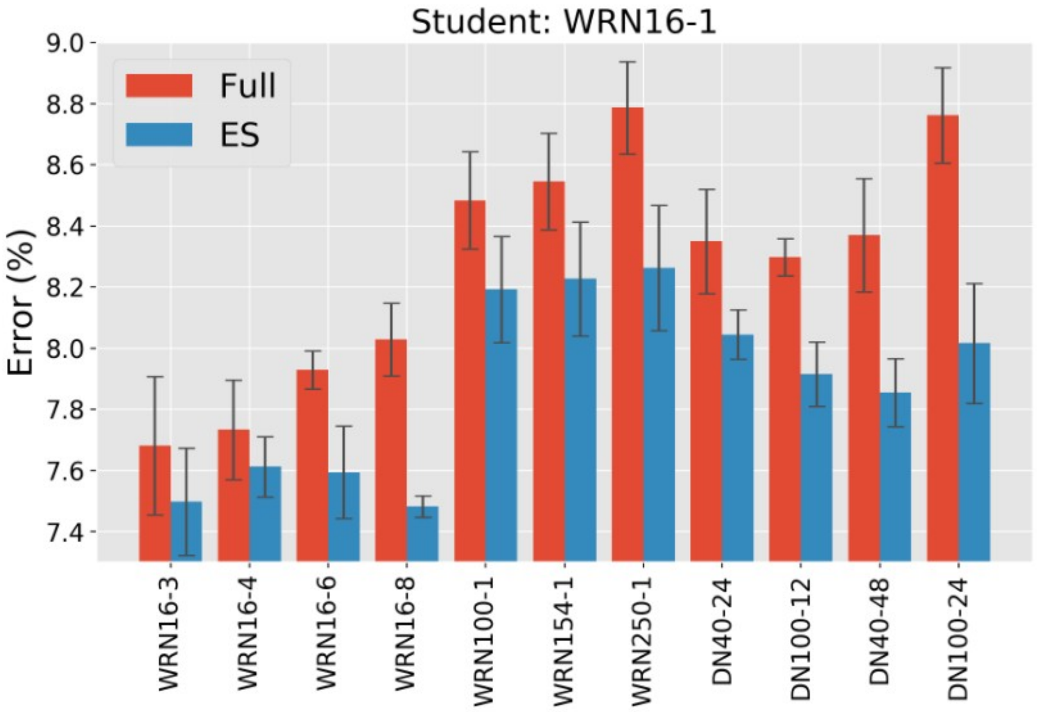
Ранний останов

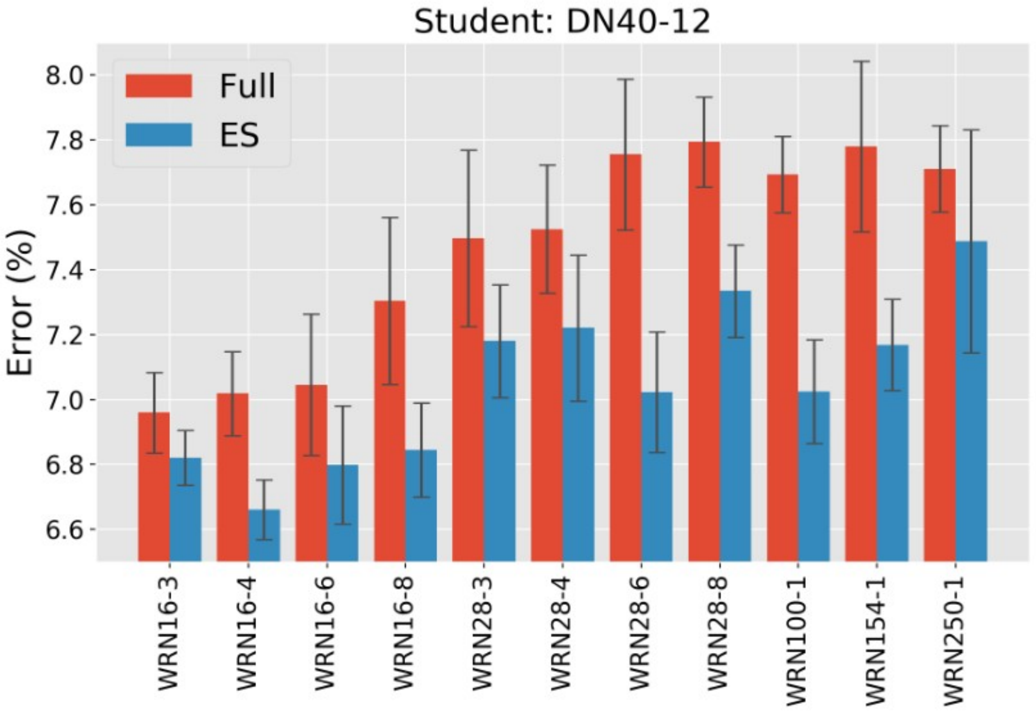




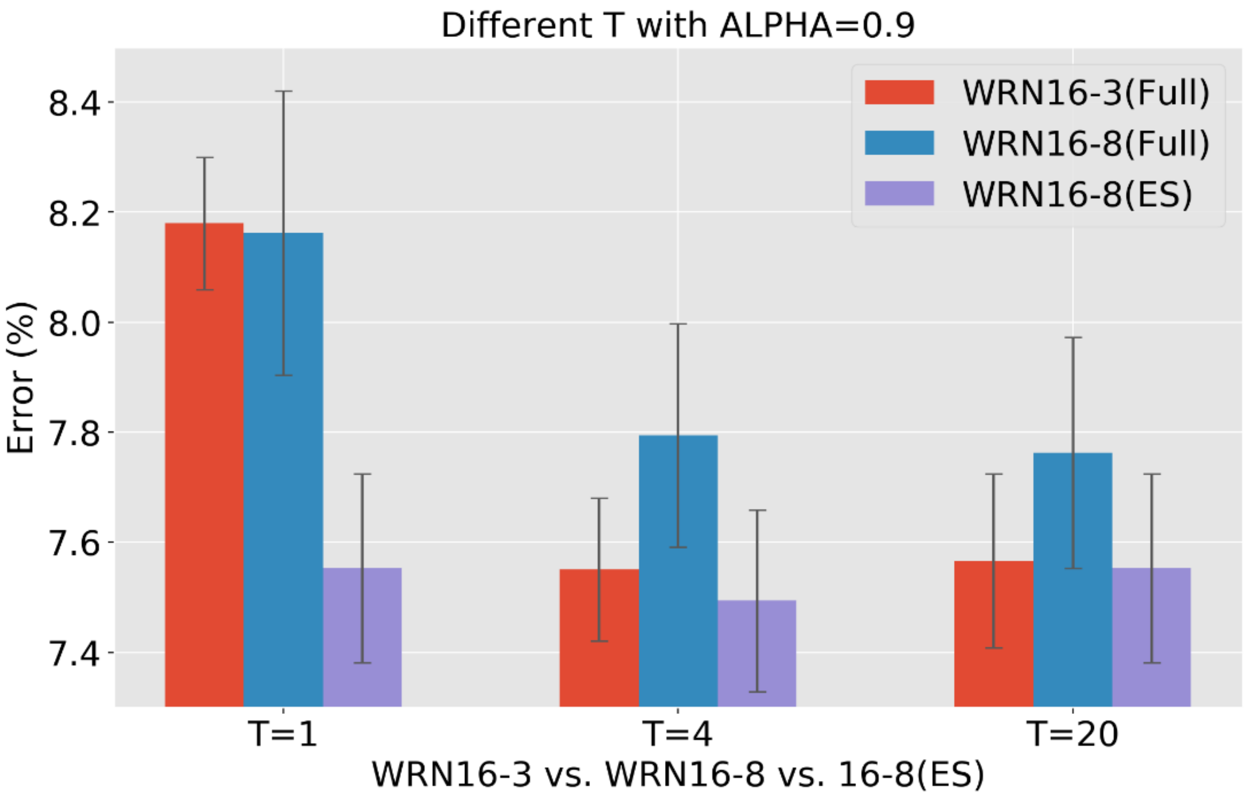
Dataset: CIFAR-10.
Top-1 Error rate(Accuracy)

Ранний останов





Еще немного про влияние температуры



Заметим, что даже при низкой температуре ранний останов улучшает качество, чем повышение температуры

Dataset: CIFAR-10

Источники

Distilling the Knowledge in a Neural Network
https://arxiv.org/pdf/1503.02531.pdf

Improved Knowledge Distillation via Teacher Assistant
https://arxiv.org/pdf/1902.03393.pdf

On the Efficacy of Knowledge Distillation
https://arxiv.org/pdf/1910.01348.pdf

Вопросы

Вопрос 1:
Что такое knowledge distillation?
Сформулируйте один из классических подходов реализации данной идеи.
Почему дистилляция может приводить к более хорошим результатам, чем обучение модели с нуля?

Вопрос 2:
Что такое TAKD?
Какую проблему дистилляции он решает?

Вопрос 3:
Что происходит с очень отрицательными логитами при классической дистилляции? Содержат ли они полезную информацию? Когда стоит их сгладить до нуля?
Если мы хотим сохранить информацию, содержащуюся в этих логитах, как мы можем улучшить процесс дистилляции?