Дистилляция знаний в нейронных сетях

Докладчик:
Цеховой Алексей,
БПМИ181

Мотивация

1. Меньшие затраты памяти.
2. Меньшее время вычислений.
3. Меньшее время на обучение.
4. Сохранение эффективности до какого-то порога.
5. Разработка продуктов для широкого пользования:
   на слабых и/или мобильных устройствах.

ImageNet – лучшие по top-1 результаты

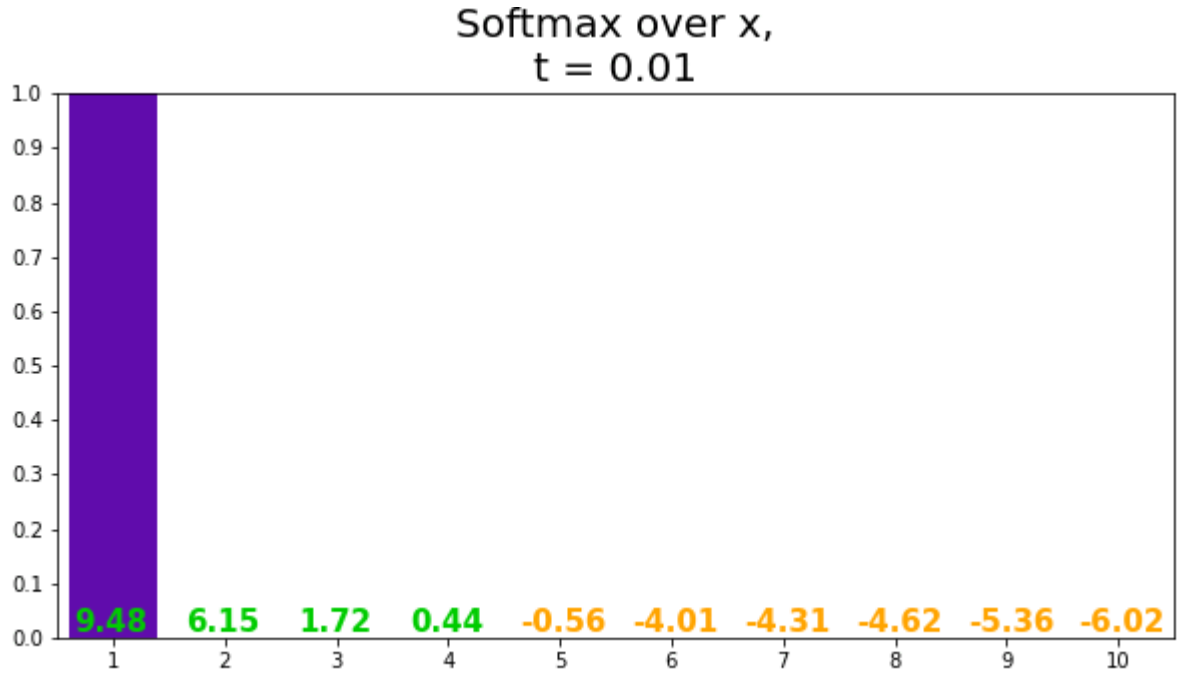
Rank	Model	Top-1 ACC	Top-5 ACC	# of params	Year
1	ViT-H/14	88.55%		632M	2020
2	FixEfficientNet-L2	88.50%	98.7%	480M	2020
3	NoisyStudent (EfficientNet-L2)	88.40%	98.7%	480M	2019
4	ViT-L/16	87.76%		307M	2020

Структура. Выход классификатора

Пусть t – температура, x – вектор логитов для классов, тогда $smax_t(x) = smax_t([x_1, x_2, \dots, x_k]) = [p_{1,t}, p_{2,t}, \dots, p_{k,t}]$, вычисляется как $p_{m,t} = \frac{e^{x_m/t}}{\sum_i e^{x_i/t}}$.

Структура. Выход классификатора

Пусть t – температура, x – вектор логитов для классов, тогда $smax_t(x) = smax_t([x_1, x_2, \dots, x_k]) = [p_{1,t}, p_{2,t}, \dots, p_{k,t}]$, вычисляется как $p_{m,t} = \frac{e^{x_m/t}}{\sum_i e^{x_i/t}}$. На примере $x = [9.48,\ \ 6.15,\ \ 1.72,\ 0.44, -0.56, \dots]$:



Структура. Функция потерь на входе

$loss_{CLS}(p(in)) = CE(p_{true}, p) = -\sum_i p_{true,i}(in) \log p_i(in)$, где
in – вход, p – полученные вероятности, а p_{true}- целевые
В случае, когда целевое распределение шифрует лишь 1 класс, имеем:
$CE(p_{true}, p) = -\log p_{trg}(in)$, где trg – номер этого класса.

Дистилляция знаний. Основная идея

- На выходе обученная модель обычно даёт грубый вектор, из которого сложно извлечь что-то кроме класса, в котором она наиболее уверенна.
- Сглаженное температурой распределение гораздо более информативно и может быть применено при обучении меньшей модели.

Дистилляция знаний. Алгоритм

1. Обучить некоторый классификатор.
2. Установить температуру $t = t^* > 0$.
3. Обучить сеть с допустимыми размерами, используя функцию ошибки $loss_{KD}^{\sim t^*}(p(in)) = (t^*)^2 CE\left(p_M^{\sim t^*}(in),\ p^{\sim t^*}(in)\right)$, где $p_M^{\sim p^*}(in)$ – сглаженный температурой p^* результат основной модели. Если данные размечены – можно улучшить, взяв взвешенное среднее ошибок $loss\big(p(in)\big) = \alpha\ loss_{CE}\big(p(in)\big) + (1-\alpha) loss_{KD}^{\sim p^*}\big(p(in)\big)$, $\alpha \ll 1$, обычно в пределах 0.05 и 0.1.
4. Установить исходную температуру (1). Сеть училась давать сглаженные распределения и это вернёт их в норму.

Результат. MNIST

- Исходная сеть – 1200 ReLU нейронов на 2 скрытых слоях.
- Применение drop-out и ограничения весов.
- Данные могли получать лёгкое искажение в целях расширения обучающей выборки.
- Результат - 67 ошибок на 10,000 тестах.

- Сравнительная сеть на 800 ReLU по 2 слоям, без дополнительных регуляризаций.
- Результат – 146 ошибок.

- Та же сеть, но обученная с дистилляцией знаний исходной.
- Результат – 74 ошибки.

Результат. MNIST

- Тот же опыт. Все тройки исключены из обучающего набора.
- Результат – 206 ошибок, 133 из которых на тройках. Всего троек в тестирующем наборе 1010.

- Та же сеть, но с повышающим коэффициентом на уверенность в тройках.
- Результат - 109 ошибок, 14 из которых на тройках.

- Сеть распознала 98,6% троек, хотя ни разу не видела их в процессе обучения. Исходная сеть успешно передала знания о тройках через дистилляцию.

Результат. MNIST

- Тот же опыт. В процессе обучения сети показывают лишь семёрки и восьмёрки.
- Результат – 47,3% ошибок.

- Та же сеть, но с понижающими коэффициентами на уверенность в семёрках и восьмёрках.
- Результат – 13,2% ошибок на всём тестовом наборе. Ещё более сильное подтверждение информативности сглаженных распределений.

Анализ. Структура

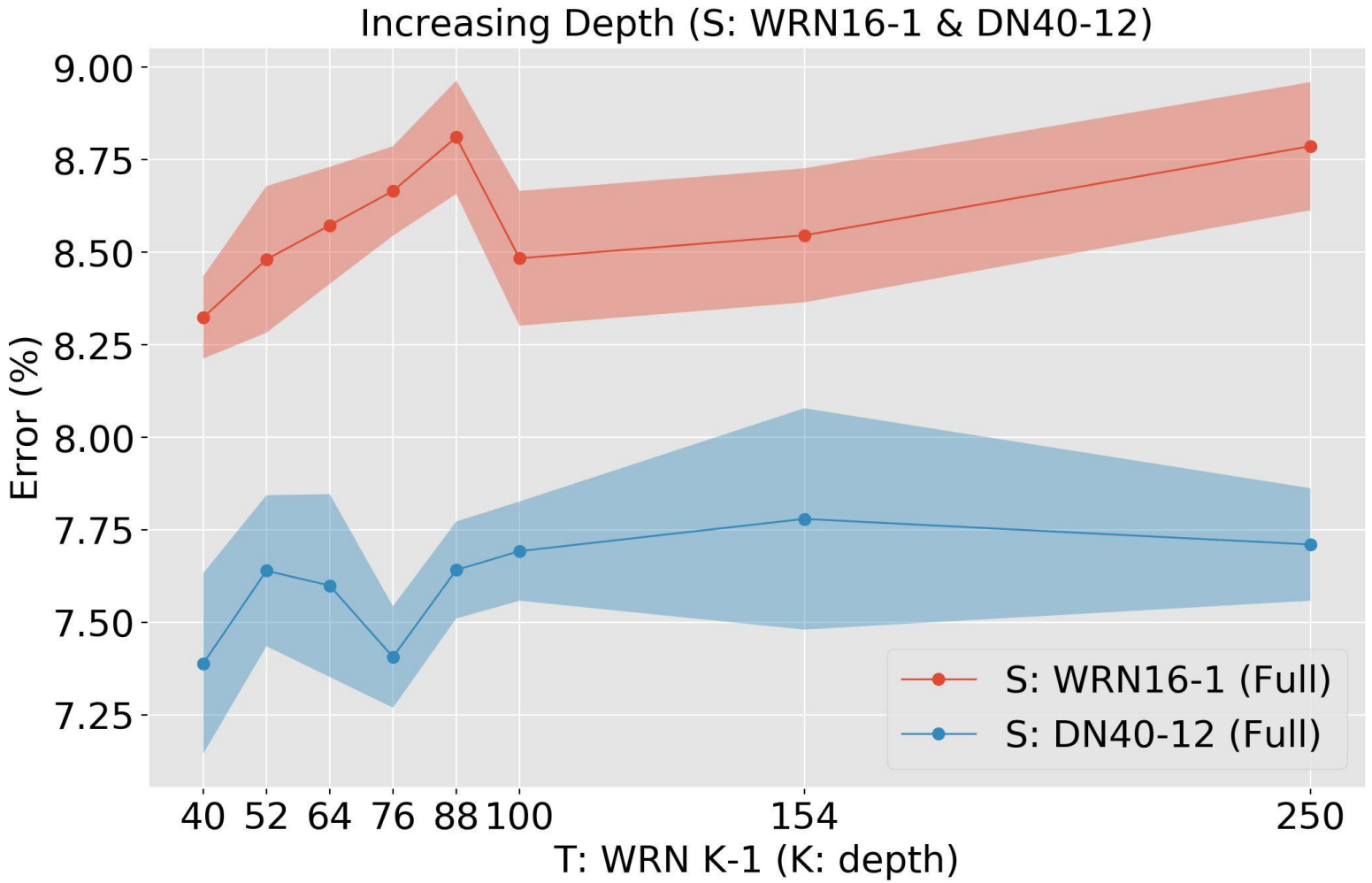
Опыты проводились на наборах данных CIFAR10 и ImageNet.

- ImageNet – постоянно пополняется, более 14M фотографий и 28K категорий.
- CIFAR10 – цветные фотографии 32x32px из 10 категорий, по 6000 фотографий на каждую. Позволяет быстро проводить различные опыты.

В качестве испытуемых использовались ResNet, WideResNet и DenseNet.

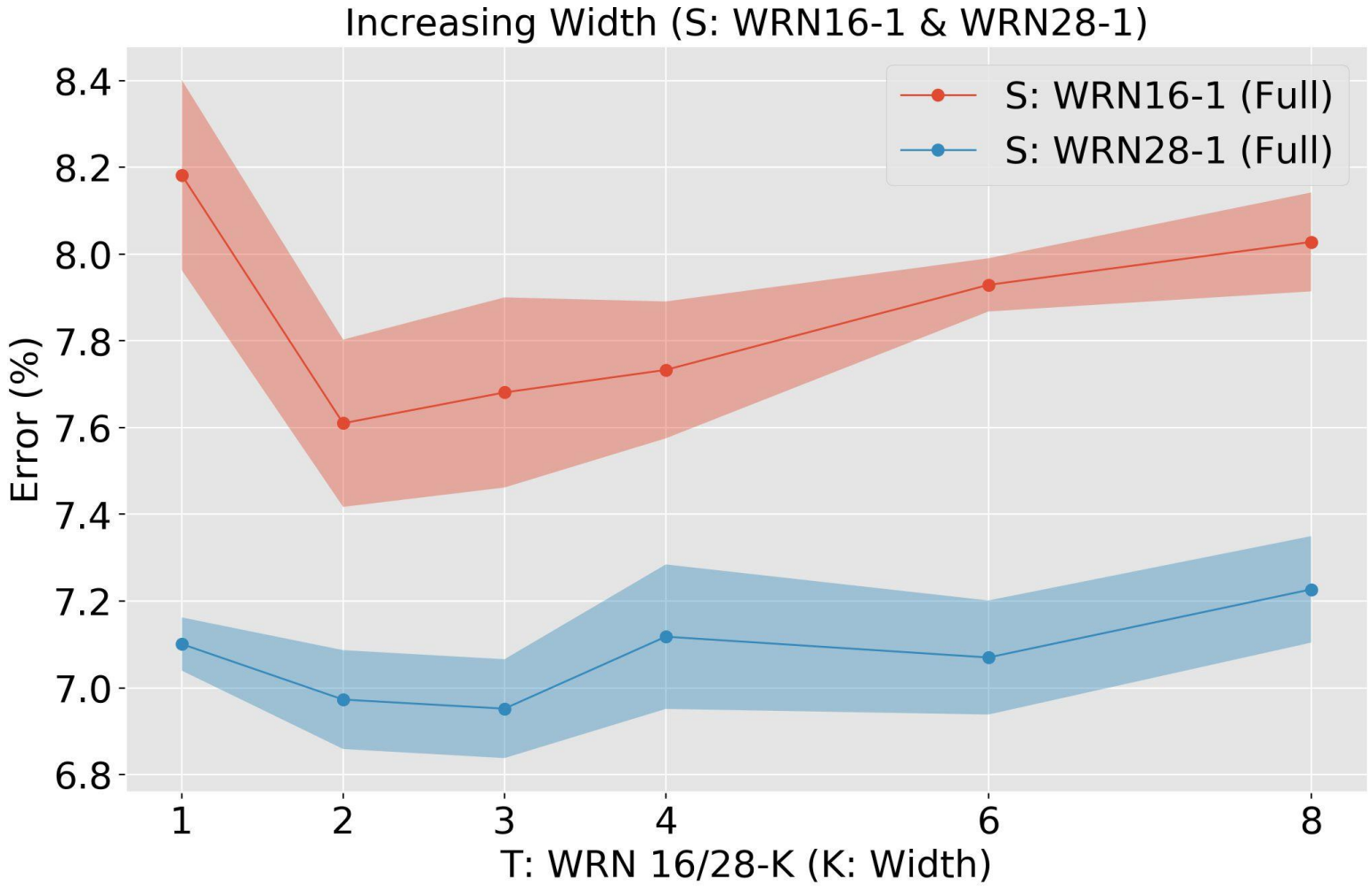
Проблемы. Сильнее учитель – слабее ученик

Ученики – WRN16-1 и DN40-12. Учителя – WRN K-1. CIFAR10.



Проблемы. Сильнее учитель – слабее ученик

Ученики – WRN16-1 и WRN28-1. Учителя – WRN16/28-K. CIFAR-10.



Предложение. Последовательная дистилляция

WRN16-3→WRN16-1 или WRN16-3→^{5}WRN16-3→WRN16-1. CIFAR10.

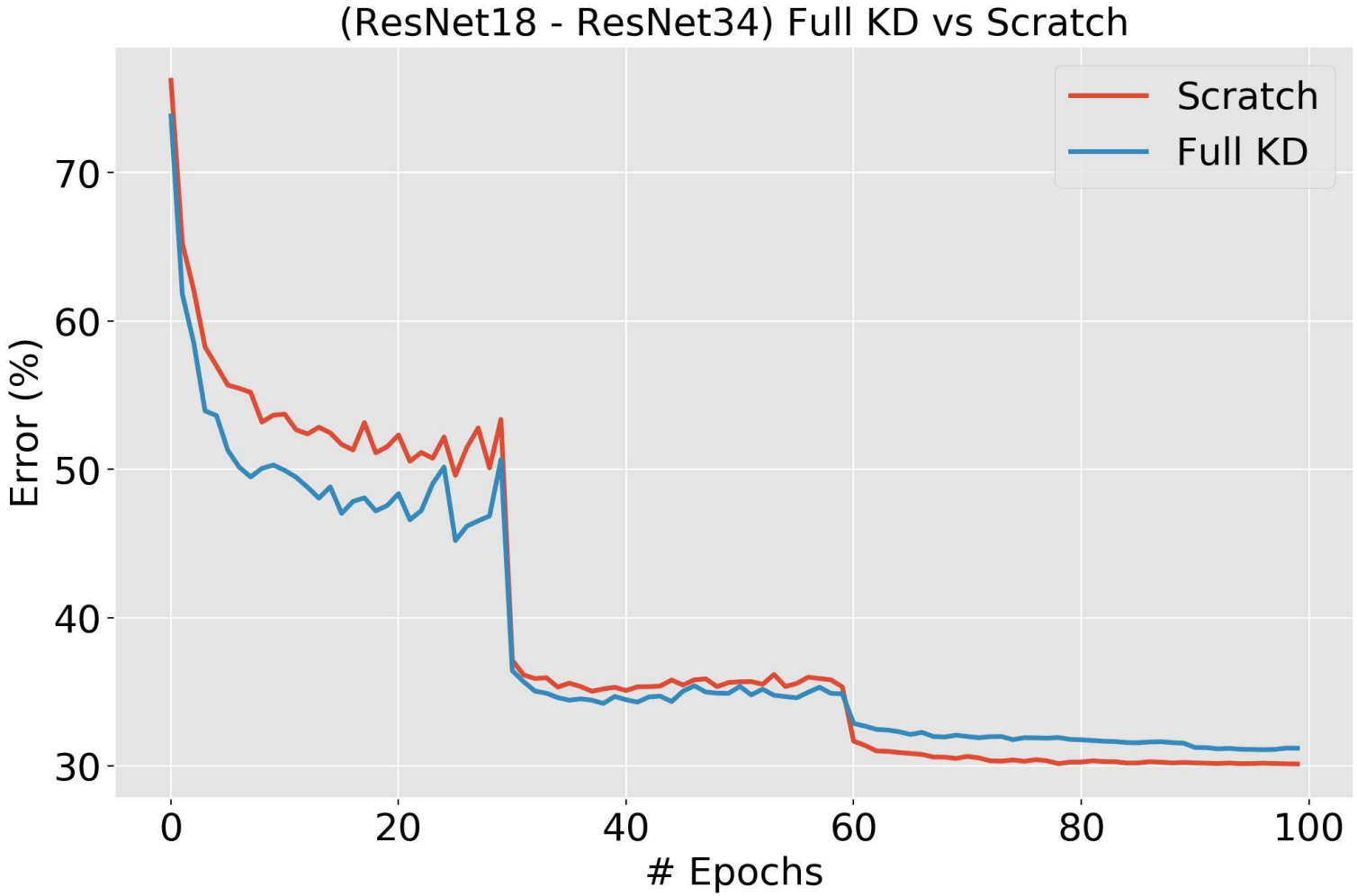
Учитель	Ошибка учителя (%)	Ошибка ученика (%)
Самостоятельно	**5.34**	**7.61 (7.68±0.259)**
5 дистилляций	4.89	7.79 (7.67±0.19)

WRN16-8 (L)→WRN16-3 (M)→WRN16-1 (S), WRN16-3→WRN16-1 или WRN16-8 →WRN16-1. CIFAR10.

Метод	Ошибка L (%)	Ошибка M (%)	Ошибка S (%)
L→M→S	4.41	4.80	8.04 (7.99±0.24)
M→S	-	**5.34**	**7.614 (7.68±0.26)**
L→S	4.41	-	7.98 (8.03±0.14)

Проблемы. Дистилляция хуже, чем с нуля

Ученик – ResNet18, учителя – самостоятельно или ResNet34. ImageNet.



Предложение. Остановка дистилляции и спуск

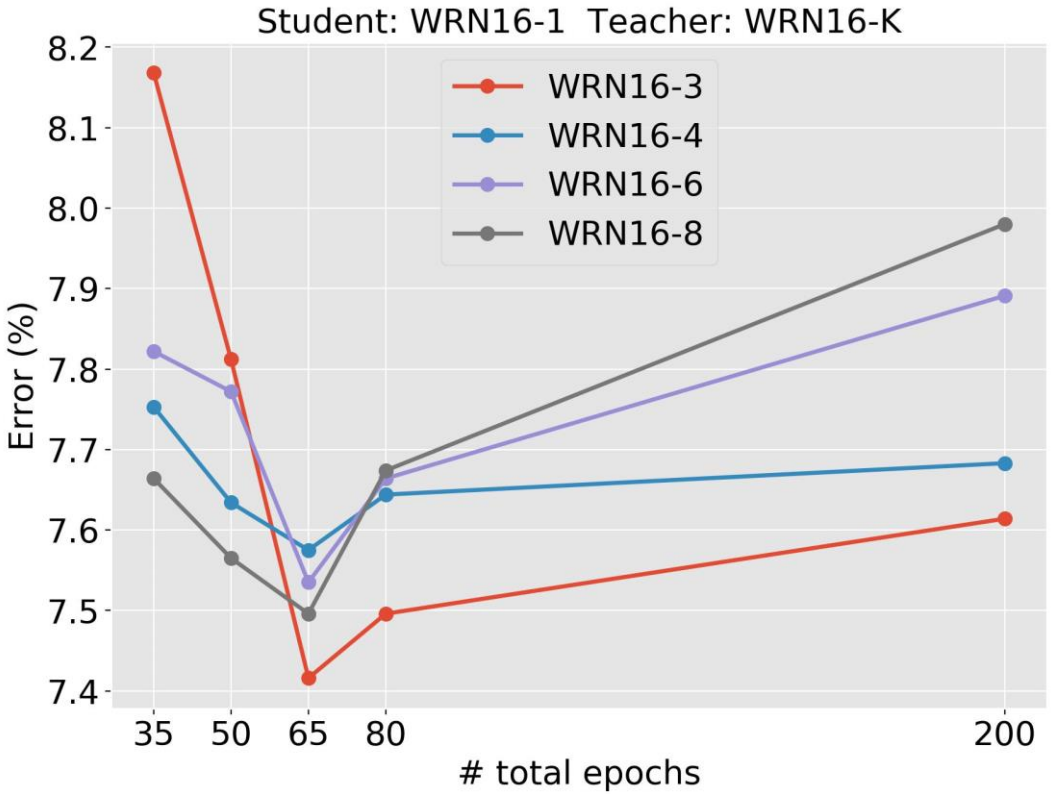
Ученик – ResNet18, учителя – самостоятельно, ResNet18, ResNet34 или ResNet50. Полная дистилляция (KD) или остановленная с обычным дообучением (ES KD). ImageNet.

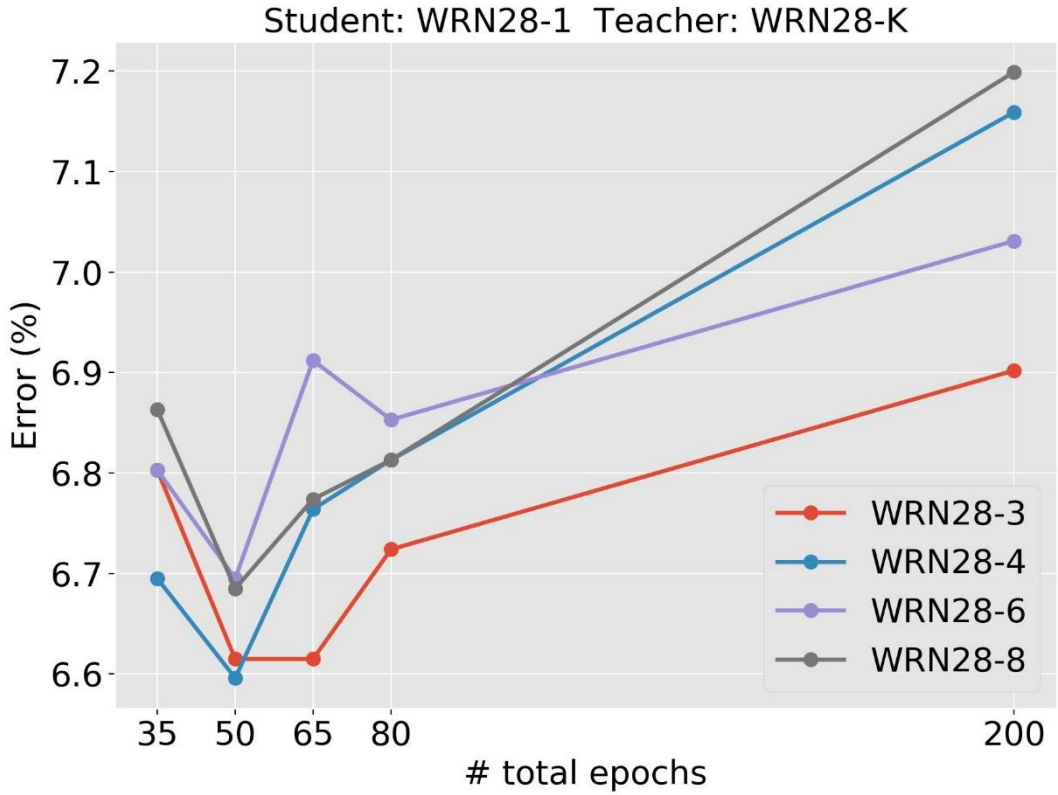
Учитель	Top-1 ACC (%)	Err CE на обучении	Err KD на обучении	Err KD на тесте
Самостоятельно	30.24	-	-	-
ResNet18 (KD)	30.57	0.146	2.916	3.358
ResNet18 (ES KD)	**29.01**	0.123	2.234	2.491
ResNet34 (KD)	30.79	0.145	1.357	1.503
ResNet34 (ES KD)	**29.16**	0.123	2.359	2.582
ResNet50 (KD)	30.95	0.146	1.553	1.721
ResNet50 (ES KD)	**29.35**	0.124	2.659	2.940

Проблемы. Чем лучше обучен учитель, тем хуже

Ученик - WRN16-1, учителя – WRN16-K разных эпох. CIFAR10.

Ученик – WRN28-1, учителя – WRN28-K разных эпох. CIFAR10.





Заключение

- Метод может быть очень эффективным.
- Однако, не является универсальным, особенно если разрыв между учителем и учеником слишком большой.
- Преждевременная остановка обучения учителя может улучшить результат.
- Преждевременная остановка дистилляции и последующее дообучение может улучшить результат.
- Цепная дистилляция не рекомендуется.

Список источников

- https://arxiv.org/abs/1503.02531
- https://arxiv.org/abs/1910.01348
- https://paperswithcode.com/sota/image-classification-on-imagenet