# Knowledge Distillation

Еленик Константин

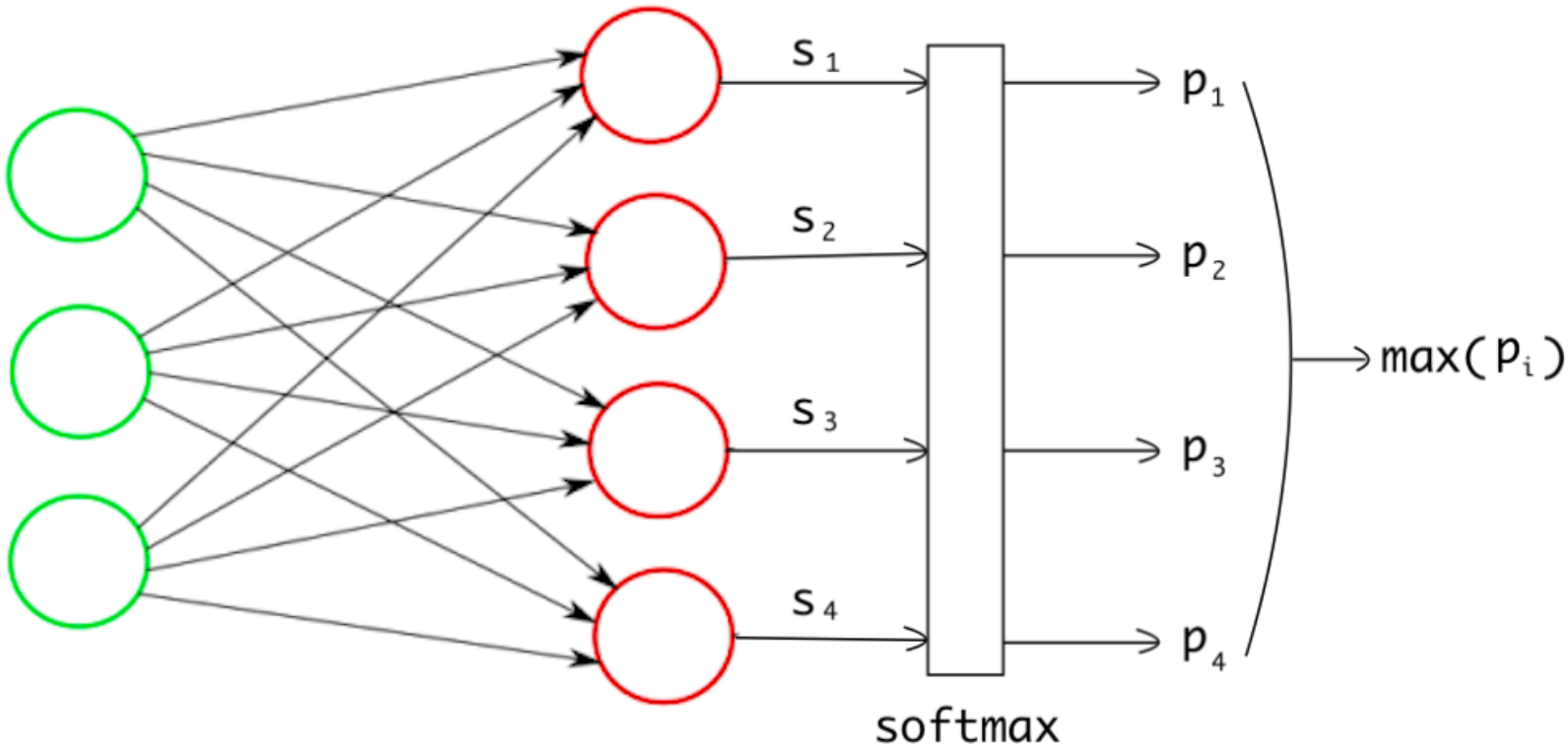
Национальный исследовательский университет
«Высшая школа экономики»

3 ноября 2020 г.

# Мотивация

- Модели имеющие наилучшее качество слишком громоздкие и медленные
- На этапе обучения и на тесте задачи модели различаются
- Хотелось бы получить с помощью большой модели, модель легче и быстрее, не сильно теряя в точности

# Среднестатистическая нейронная сеть




$p_i$ могут дать информацию о похожести классов

# Softmax

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

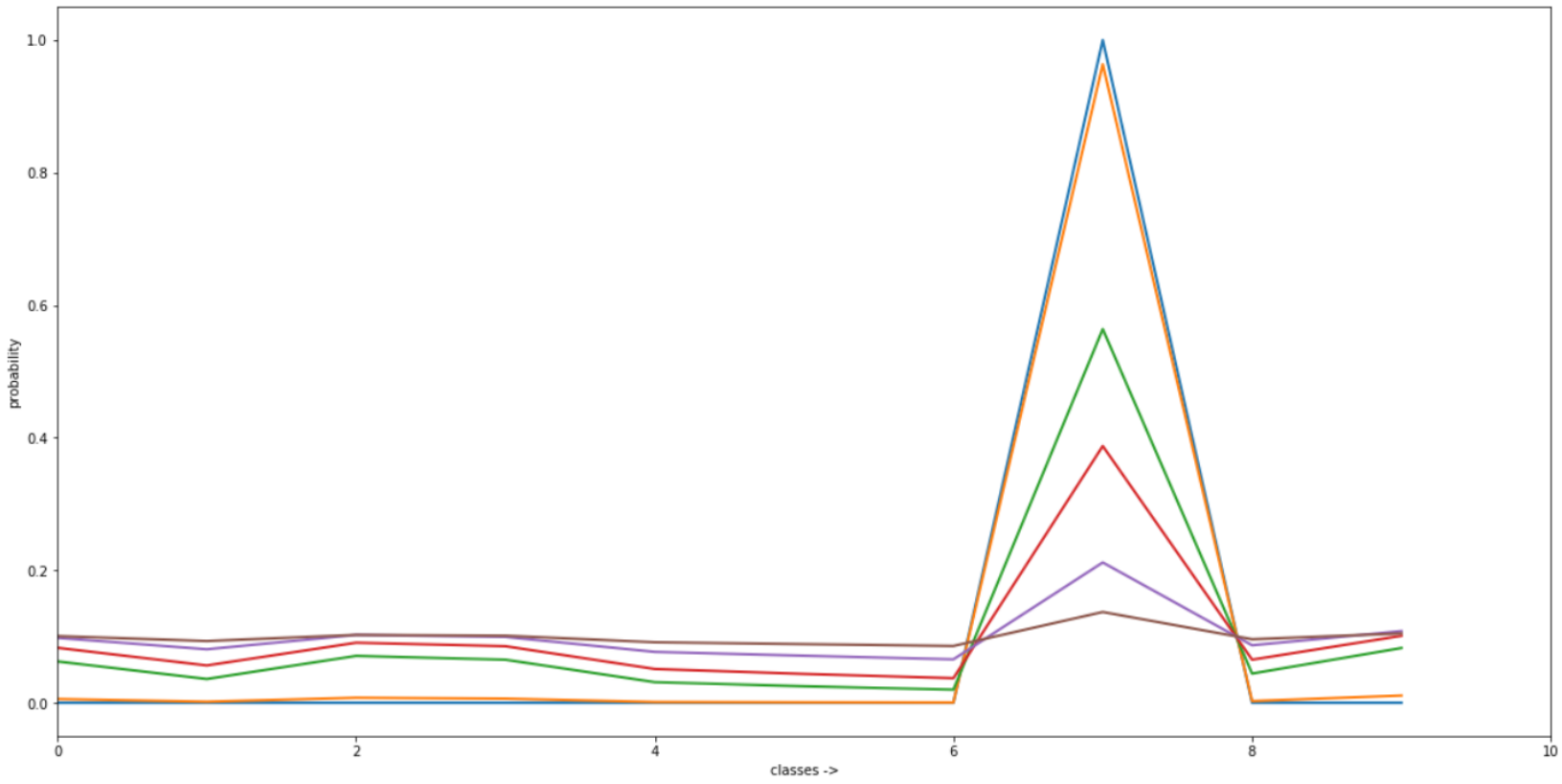
где

- $T$ – температура
- $z_i$ – логит
- $q_i$ – вероятность

Температура помогает регулировать гладкость распределения

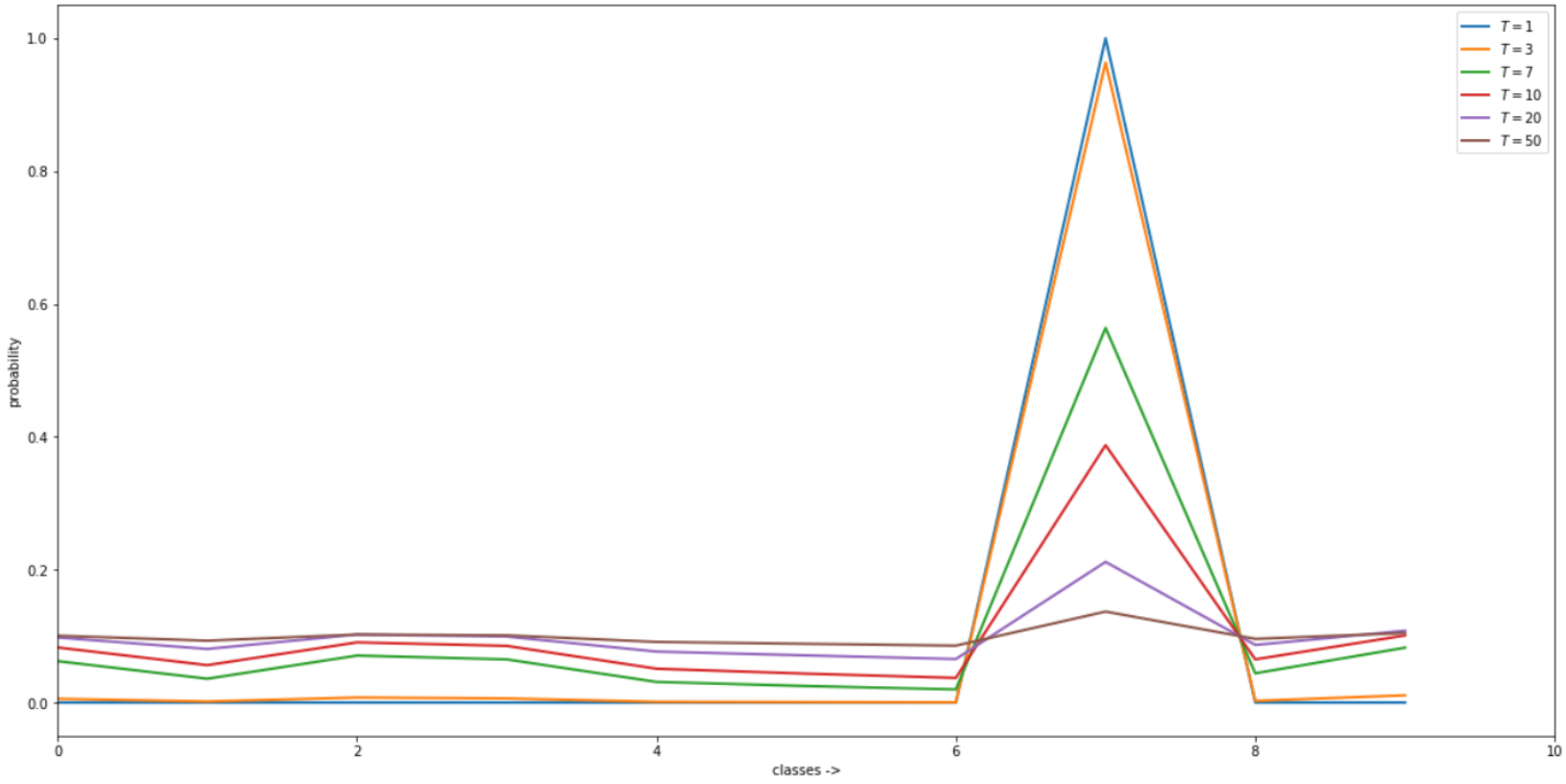
# Распределение вероятностей в зависимости от температуры в softmax

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

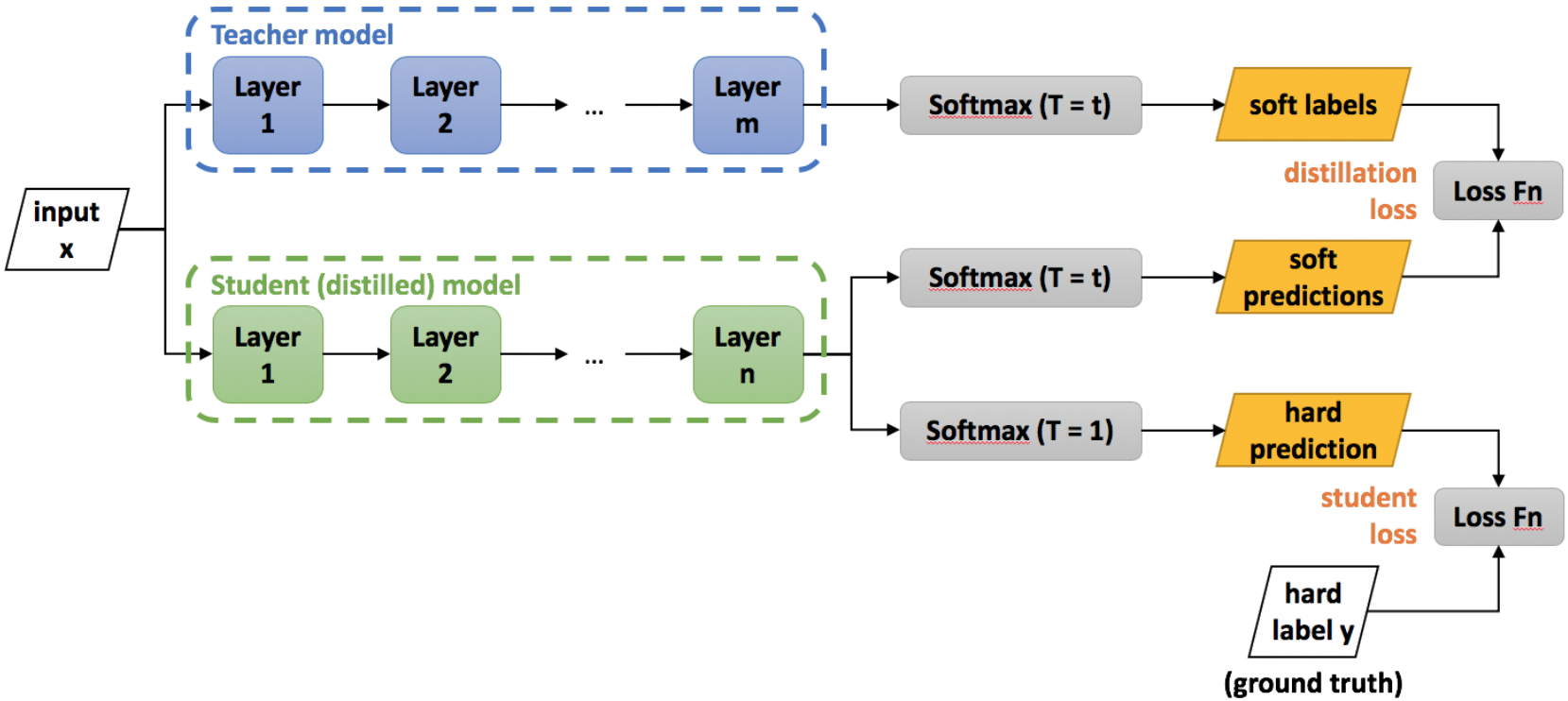


# Распределение вероятностей в зависимости от температуры в softmax

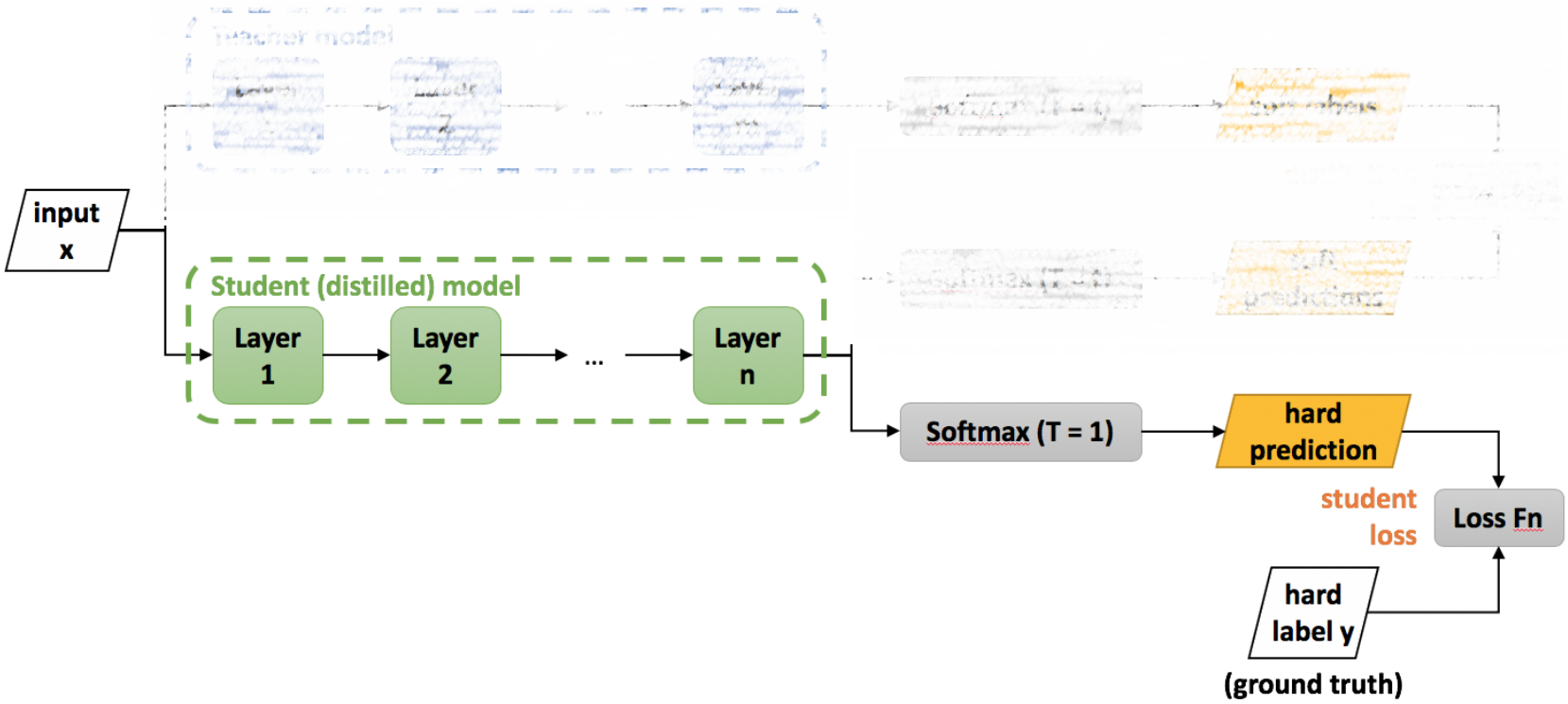
$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$



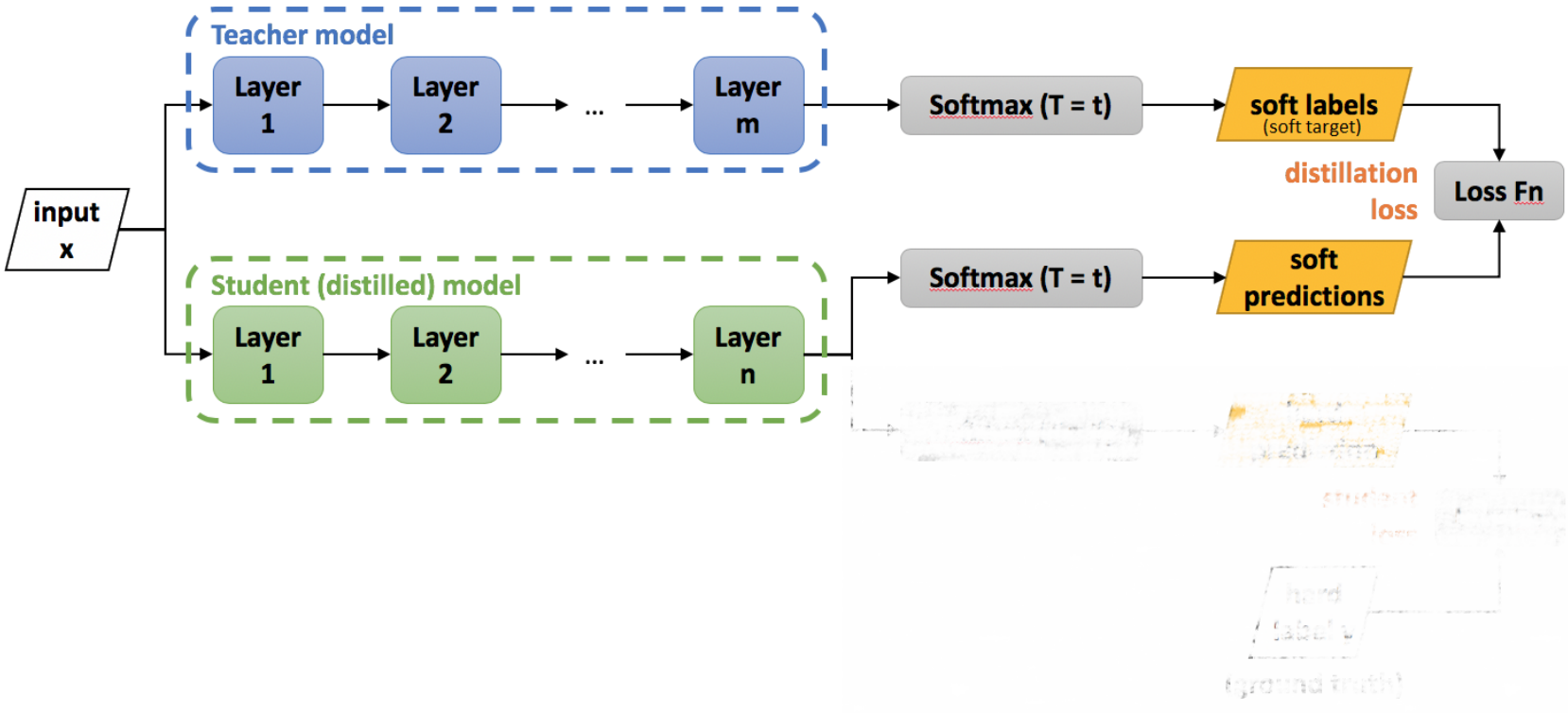
# Дистилляция



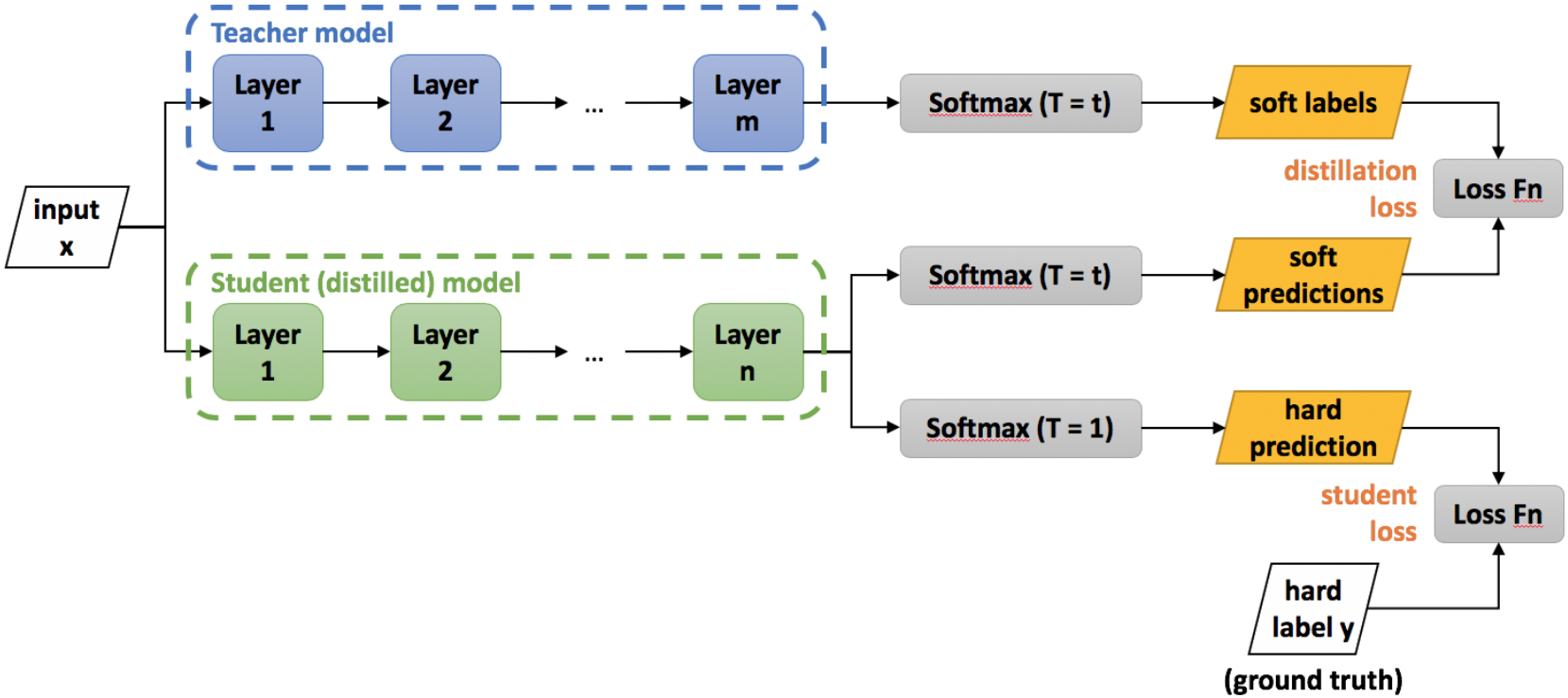
# Дистилляция



# Дистилляция

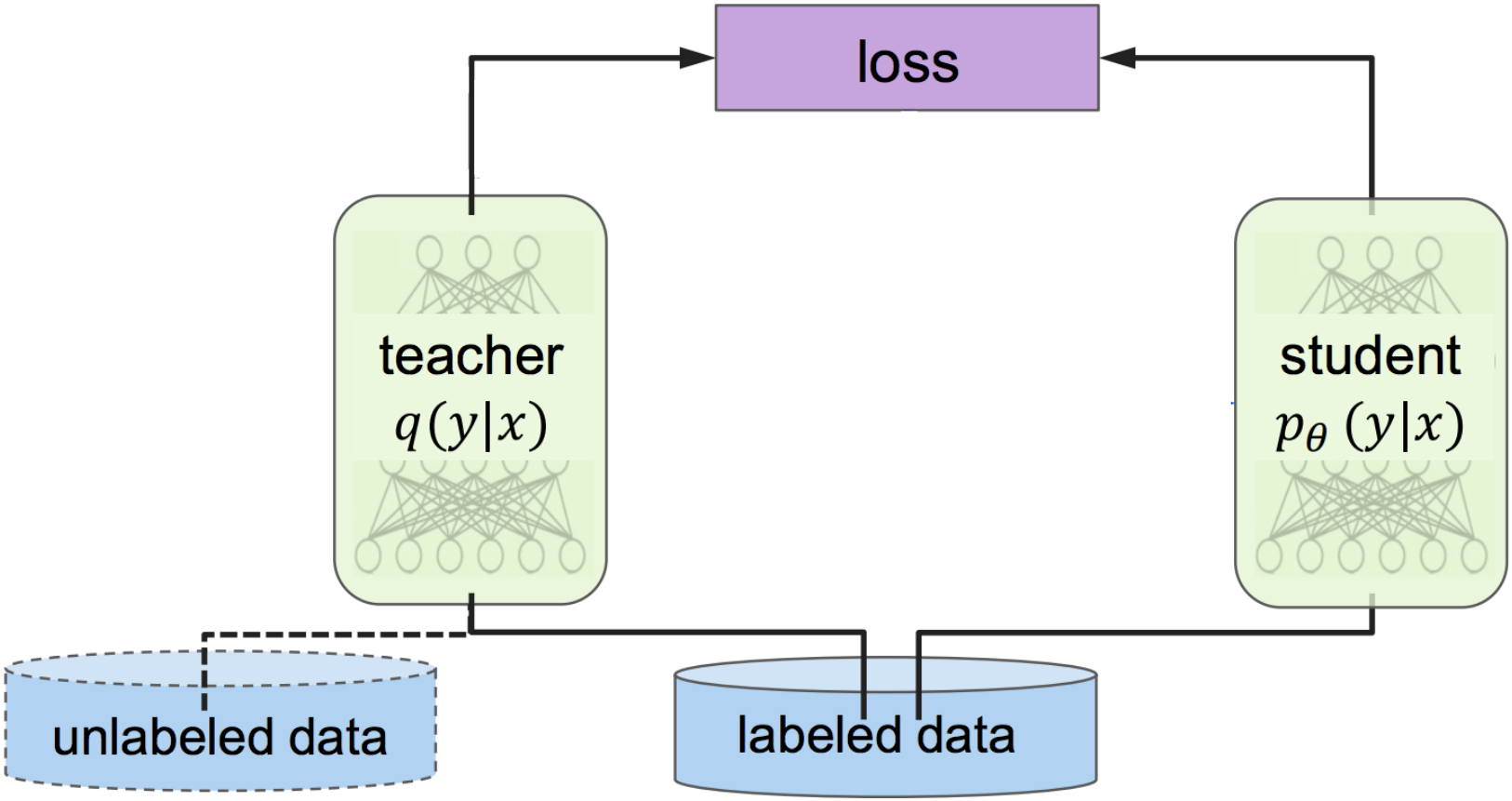


# Дистилляция



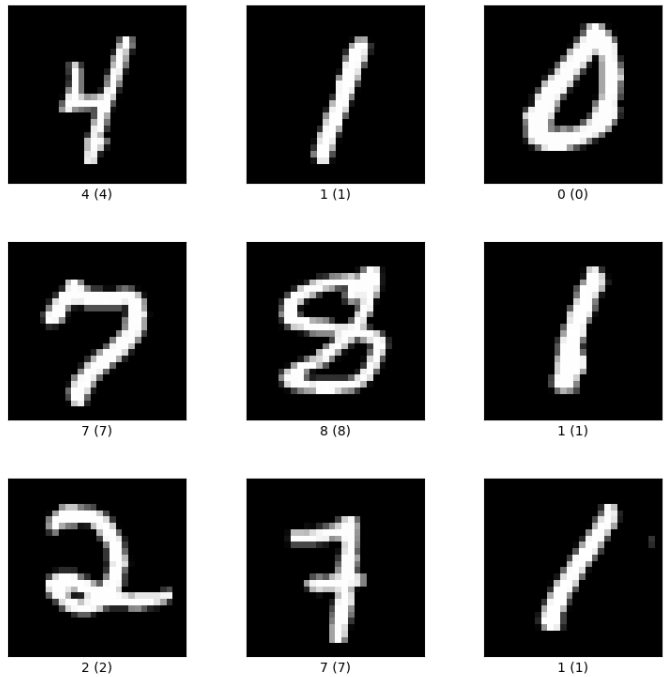

$$\mathcal{L}_{total} = (1 - \lambda) \cdot \mathcal{L}_{student} + \lambda \cdot \mathcal{L}_{distilation}$$

# Дистилляция



# Результаты на MNIST

| | | |
|---|---|---|
| Teacher | 2x1200 units | 67 test errors |
| Scratch | 2x800 units | 146 test errors |
| Student | 2x800 units | 74 test errors |
| Student (without 3) | 2x800 units | 109 test errors (14 on 3s) |



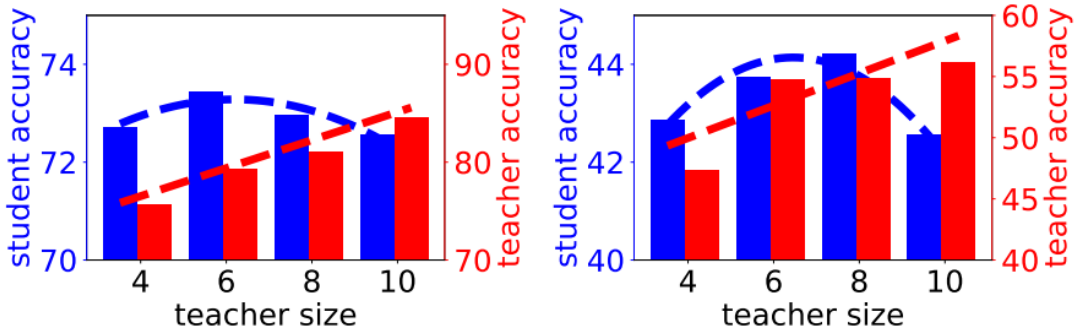
# Ансамбли

Распознавание речи:

| | |
|---|---|
| Baseline | 58.9% |
| 10xEnsemble | 61.1% |
| Distiled | 60.8% |

Дистилляция не всегда позволяет добиться лучших результатов.
Иногда точность сильно проседает, становясь хуже чем у модели обученной с нуля.

# Связь размера учителя с его преподавательскими способностями

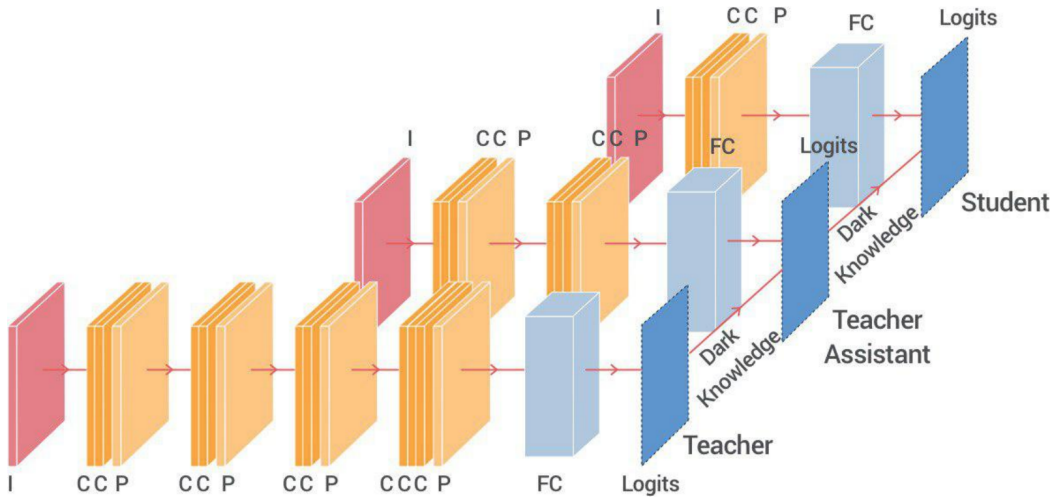


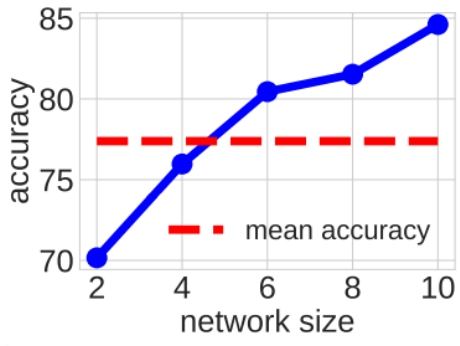

a) CIFAR-10 b) CIFAR-100

Возможные причины:

1. Недостаточная гладкость вероятностей
2. Функция потерь дистилляции не отражает нужную метрику
3. Найденное решение лежит вне пространства решения

# Ассистент




Лучший размер ассистента – тот на котором достигается средняя точность между учителем и студентом

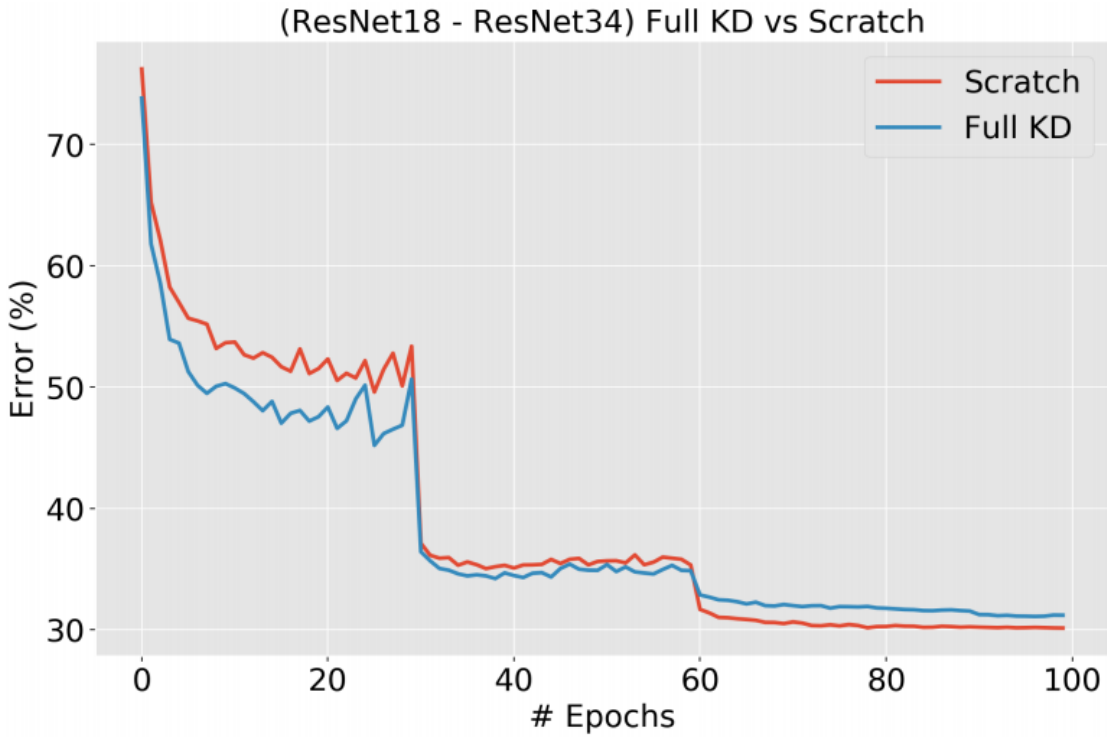
# Больше ассистентов!

| Size | 10 | 8 | 6 | 4 | 2 | |
|---|---|---|---|---|---|---|
| | | 1 distillation path | 2 distillation paths | 4 distillation paths | 8 distillation paths | Path for S=2 |
| 1 | | | | 52.87 | 45.14 | 10->8->6->4->2 |
| 2 | | | 57.53 | | 44.46 | 10->8->6->2 |
| 3 | | | | 52.59 | 44.47 | 10->8->4->2 |
| 4 | 56.19 | 56.75 | | | 44.28 | 10->8->2 |
| 5 | | | 57.13 | 52.84 | 45.06 | 10->6->4->2 |
| 6 | | | | | 44.57 | 10->6->2 |
| 7 | | | | 50.94 | 44.92 | 10->4->2 |
| 8 | | | | | 42.56 | 10->2 |
| NOKD | 56.19 | 54.86 | 54.8 | 47.39 | 41.09 | |

Figure 5: Distillation paths for plain CNN on CIFAR-100 with T=10

вывод: ограничивают только ресурсы

# Ранняя остановка студента



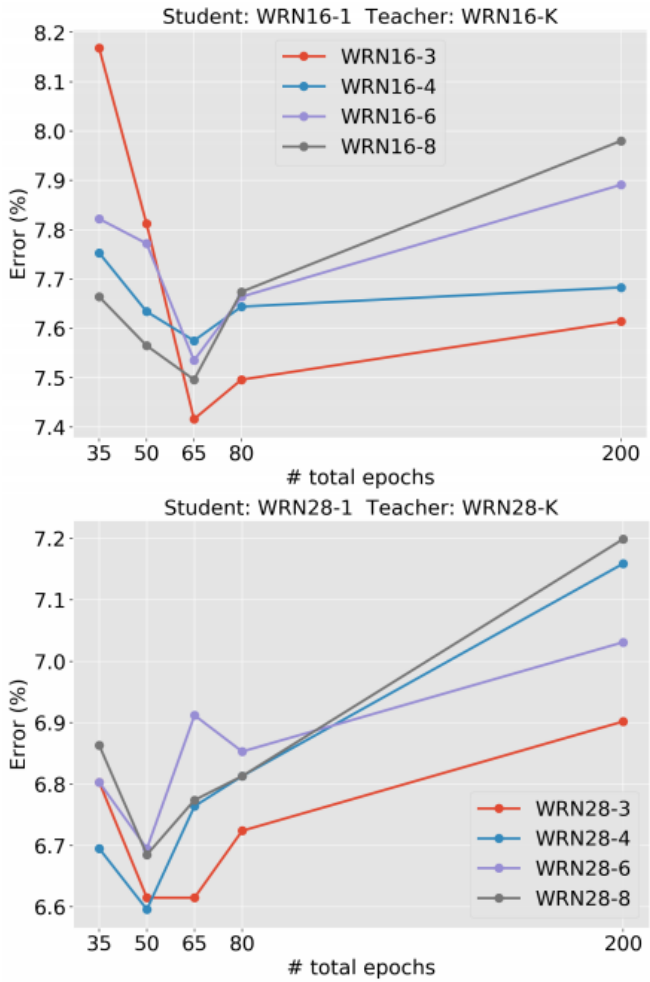

Проблема: В конце обучения у студента не достаточная вместимость для минимизации двух функций потерь.

Решение: Начиная с какой-то эпохи минимизируем только оригинальный loss

# Ранняя остановка учителя

Проблема: Найденное учителем решение лежит вне пространства решений студента.

Решение: Сделать так, чтобы учитель нашел решение, которое может достичь студент.




Считается, что если остановить обучение раньше, то поведение большой модели будет аналогично поведению меньшей, но возможность находить лучшие решения всё еще будет больше.

# Итог

С помощью больших моделей (а именно – их логитов) можно обучать меньшие, более пригодные для использования по размеру и скорости, при этом получая лучшие результаты, чем при обычном обучение с нуля. Этот подход называется дистилляция.

Для достижения лучшего результата существует некоторое количество эвристик, в том числе рассмотренные сегодня:

- дистилляция в несколько шагов (с учебным ассистентом)
- дистилляция в с ранней остановкой студента
- дистилляция на рано остановленном учителе

# Список источников:


- https://arxiv.org/pdf/1503.02531.pdf
- https://arxiv.org/pdf/1902.03393.pdf
- https://arxiv.org/pdf/1910.01348.pdf
- https://www.youtube.com/watch?v=-4PGNj1CqZc
- https://www.youtube.com/watch?v=veYYkjC3Yvc