Дистилляция на многодоменных выборках

К. М. Баязитов

Московский физико-технический институт

25 апреля 2022 г.

Слайд об исследованиях

Исследуется проблема переноса информации с более сложной модели *учителя* на модель *ученика* с меньшим числом параметров.

Цель исследования —

Понижение сложности модели машинного обучения при переходе к данным меньшей мощности.

Предложенные методы

1) дистилляция моделей,
2) доменная адаптация.

Решение

Предлагается при обучении модели ученика использовать помимо меток учителя также и связь между выборками.

Постановка задачи дистилляции для многодоменной выборки

Заданы

1) исходный и целевой наборы данных:

$$\mathfrak{D}_{\mathsf{s}} = (\mathbf{X}_{\mathsf{s}}, \mathbf{Y}_{\mathsf{s}}), \quad \mathbf{X}_{\mathsf{s}} \in \mathbb{X}_{\mathsf{s}}, \quad \mathbf{Y}_{\mathsf{s}} \in \mathbb{Y},$$

$$\mathfrak{D}_{\mathsf{t}} = (\mathbf{X}_{\mathsf{t}}, \mathbf{Y}_{\mathsf{t}}), \quad \mathbf{X}_{\mathsf{t}} \in \mathbb{X}_{\mathsf{t}}, \quad \mathbf{Y}_{\mathsf{t}} \in \mathbb{Y},$$

2) модель учителя на выборке большей мощности:

$$\mathbf{f} : \mathbb{X}_{\mathsf{s}} \to \mathbb{Y},$$

3) связь между исходной и целевой выборками:

$$\varphi : \mathbb{X}_{\mathsf{t}} \to \mathbb{X}_{\mathsf{s}}.$$

Требуется получить модель ученика для малоресурсной выборки:

$$\mathbf{g} : \mathbb{X}_{\mathsf{t}} \to \mathbb{Y}.$$

Предложенный метод

Предлагается при обучении модели ученика использовать

1) ответы модели учителя

$$\mathbf{f} : \mathbb{X}_\mathsf{s} \to \mathbb{Y},$$

2) связь между выборками

$$\varphi : \mathbb{X}_\mathsf{t} \to \mathbb{X}_\mathsf{s}.$$

Функция ошибки, учитывающая метки учителя и связь между выборками

$$\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi) = -\lambda \sum_{i=1}^{m} \sum_{r=1}^{R} \mathsf{I}[y_i = r] \log g^r(\mathbf{x}_i, \mathbf{w}) - (1-\lambda) \sum_{i=1}^{m} \sum_{r=1}^{R} (f \circ \varphi)^r(\mathbf{x}_i) \log g^r(\mathbf{x}_i, \mathbf{w}),$$

где λ — метапараметр, задающий вес дистилляции, I — индикаторная функция.
Оптимизационная задача:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi).$$

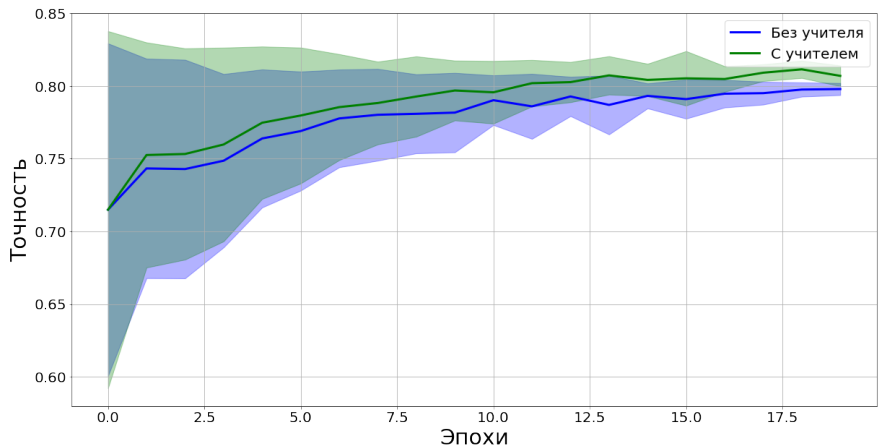
Экспериментальные данные

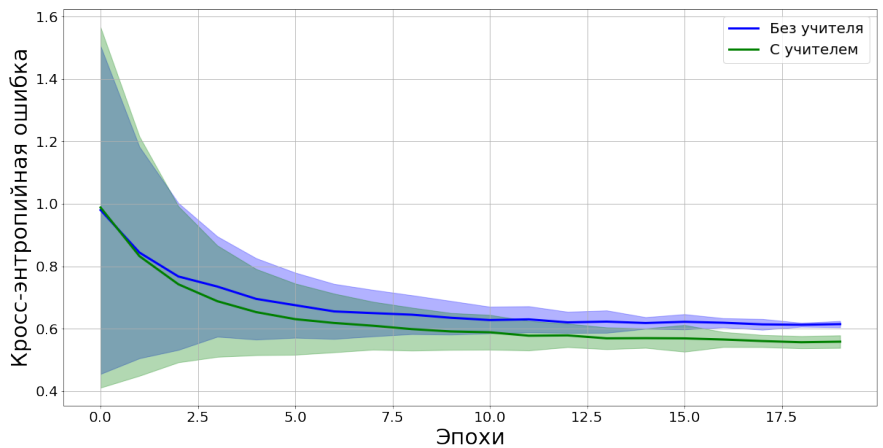
Эксперимент проводится для выборок FashionMNIST — набора изображений предметов одежды и MNIST — набора изображений рукописных цифр.
Каждая из выборок состоит из двух частей — обучающая и тестовая части. Обучающая часть разделяется на многоресурсную и малоресурсную части. Обучающая часть содержит 60000 объектов, тестовая часть содержит 10000 объектов, многоресурсная часть содержит 59000 объектов, малоресурсная часть содежит 1000 объектов.

Выборка	Пояснение	Размер выборки
FashionMNIST-Train	Обучающая часть	60000
FashionMNIST-Big	Многоресурсная часть	59000
FashionMNIST-Small	Малоресурсная часть	1000
FashionMNIST-Test	Тестовая часть	10000
MNIST-Train	Обучающая часть	60000
MNIST-Big	Многоресурсная часть	59000
MNIST-Small	Малоресурсная часть	1000
MNIST-Test	Тестовая часть	10000

Анализ дистилляции на малоресурсной части

Модель учителя обучается на многоресурсной части FashionMNIST-Big, а модель ученика обучается на малоресурсной части FashionMNIST-Small.
На графиках показаны метрики точности и кросс-энтропийной ошибки модели ученика.



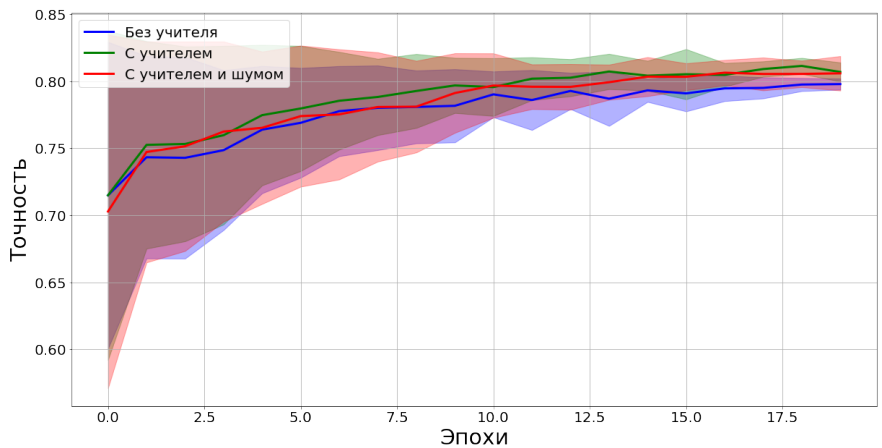


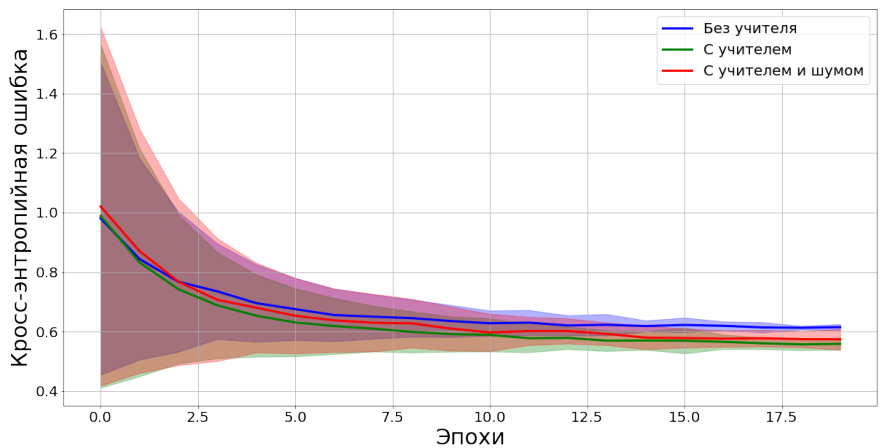
Модель, использующая метки учителя, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Анализ дистилляции с нормальным шумом

Добавим к многоресурсной части FashionMNIST-Big нормальный шум $\mathcal{N}\left(0, \frac{1}{10}\right)$ и обучим на нем модель учителя. Модель ученика обучается на малоресурсной части FashionMNIST-Small без шума. В качестве отображения φ используется нормальный шум.

На графиках показаны метрики точности и кросс-энтропийной ошибки модели ученика.

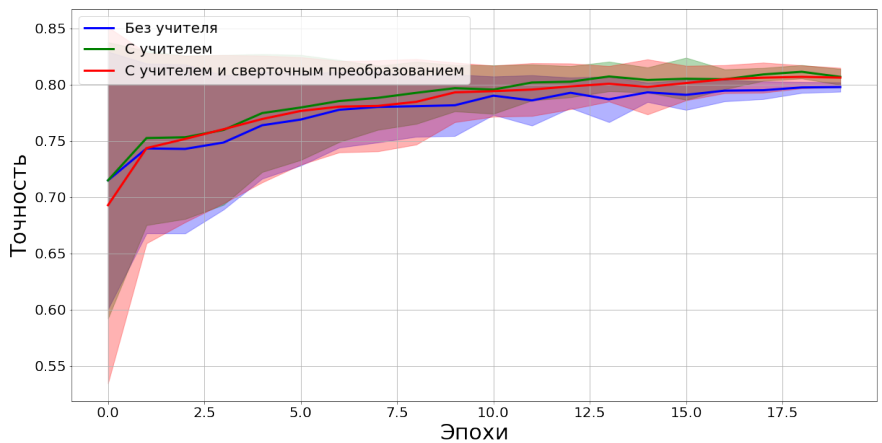


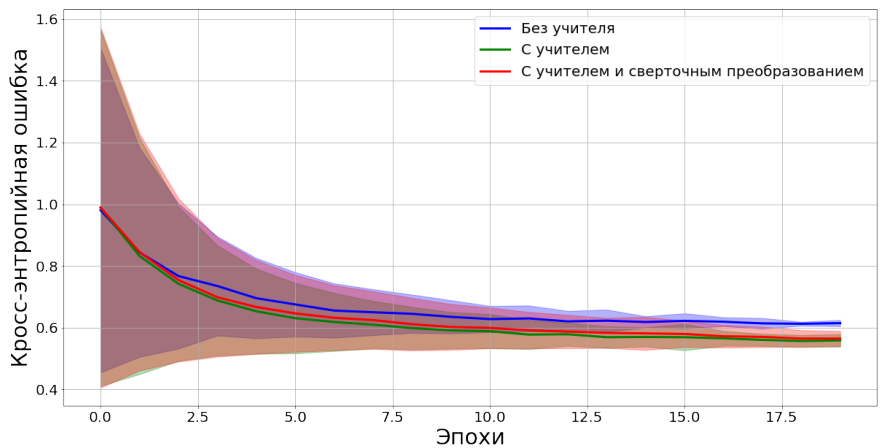


Модель, использующая метки учителя с применением шума, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Анализ дистилляции со сверточным преобразованием

Применим к многоресурсной части FashionMNIST-Big сверточное преобразование с параметром dilation $= 2$ и обучим на нем модель учителя. Модель ученика обучается на малоресурсной части FashionMNIST-Small. В качестве отображения φ используется сверточное преобразование.
На графиках показаны метрики точности и кросс-энтропийной ошибки модели ученика.





Модель, использующая метки учителя со сверточным преобразованием, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Вариационный автокодировщик

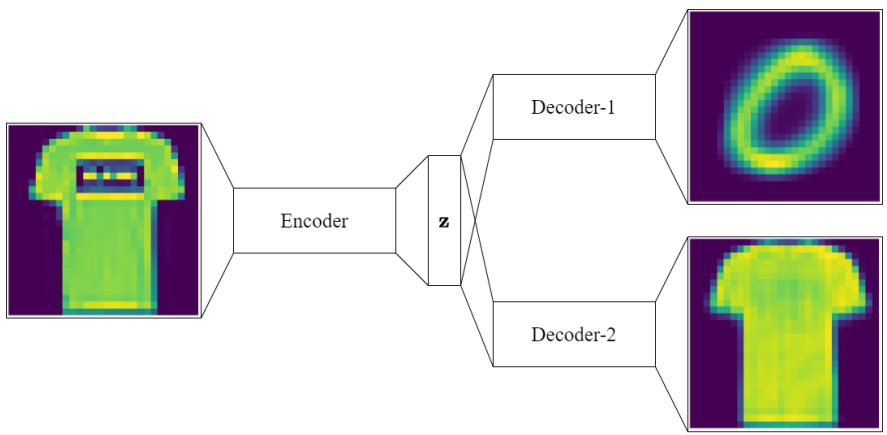
В качестве преобразования элементов выборки FashionMNIST в элементы выборки MNIST используем модель вариационного автокодировщика, аппроксимирующую отображение φ.

Функция ошибки:

$$\mathcal{L}_{\text{VAE}}(\alpha, \beta) = \sum_{i=1}^{l} \mathsf{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}_i, \alpha)} \log p(\mathbf{x}_i|\mathbf{z}, \beta) d\mathbf{z} - \mathsf{KL}(q(\mathbf{z}|\mathbf{x}_i, \alpha)||p(\mathbf{z})),$$

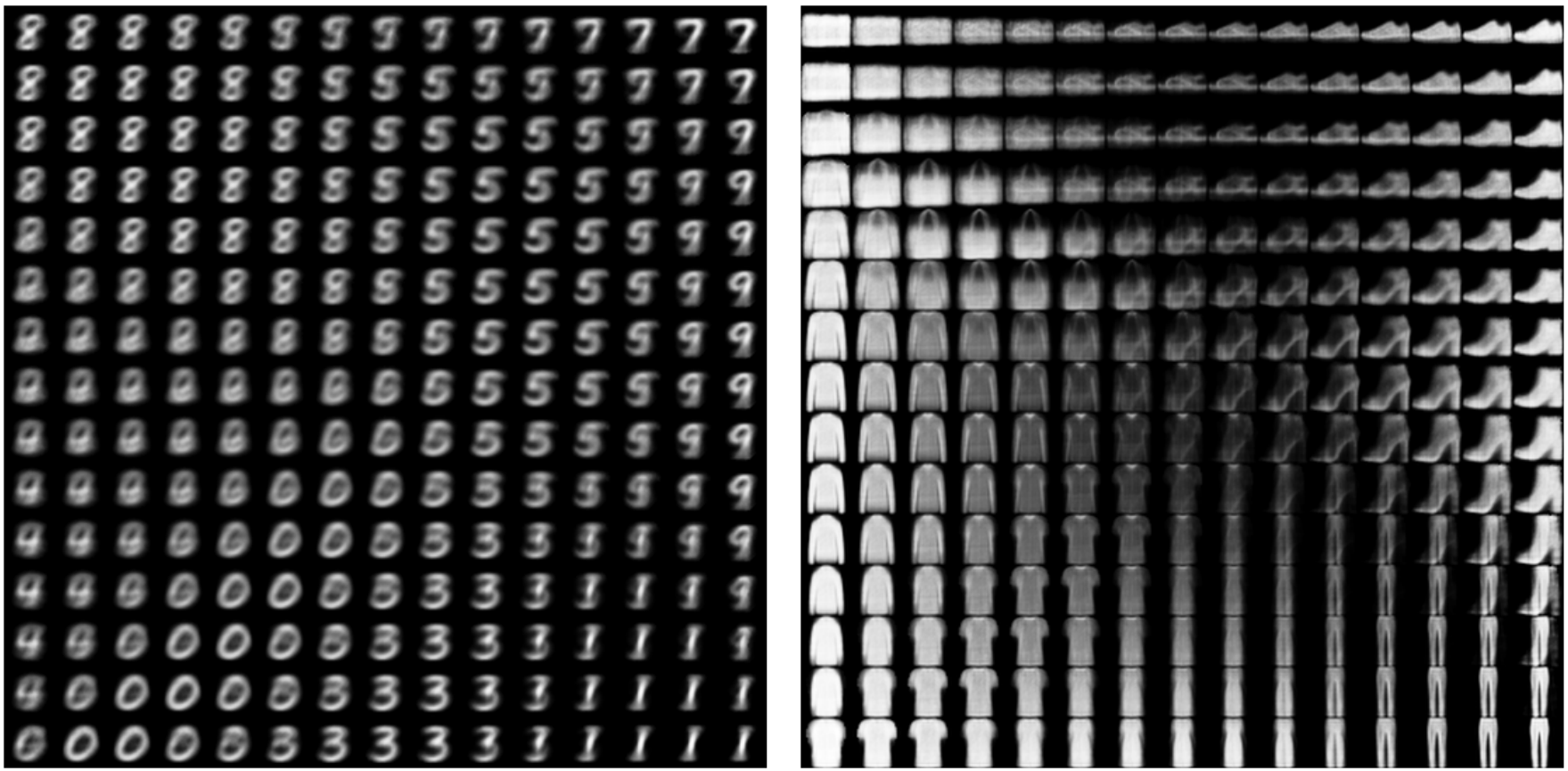
где $p(\mathbf{z}) \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ — априорное распределение, $q(\mathbf{z}|\mathbf{x}, \alpha)$ — вероятностный кодировщик, $p(\hat{\mathbf{x}}|\mathbf{z}, \beta)$ — вероятностный декодировщик.

Пример работы вариационного автокодировщика, генерирующего семейство новых объектов — изображений цифры и изображений одежды для одного и того же изображения одежды.



Анализ работы вариационного автокодировщика

Проанализируем изменение выхода модели при изменении случайного вектора в скрытом представлении. Для визуализации рассмотрим скрытое представление размерности 2:

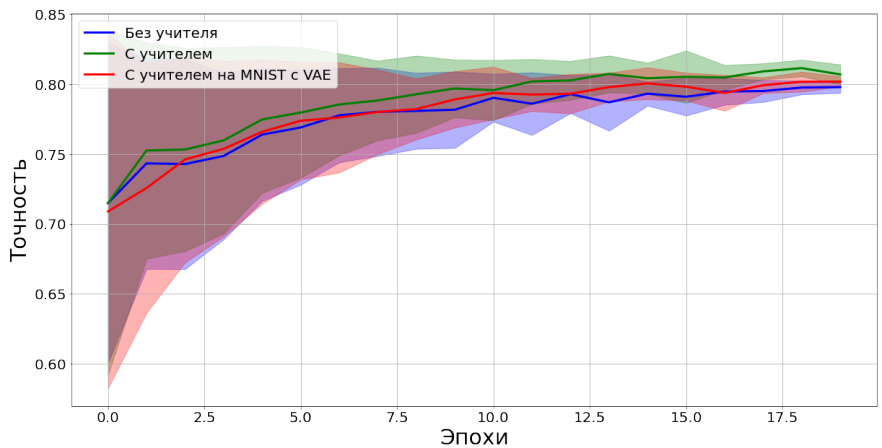


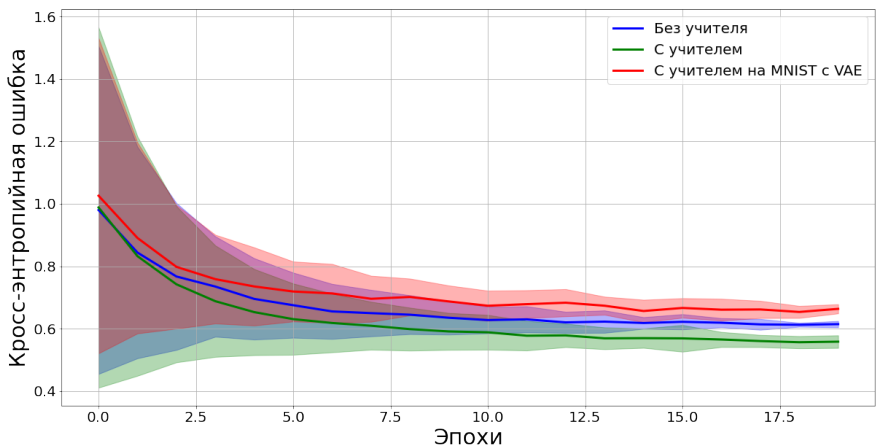
Видно, что классы одежды и цифр соответствуют друг другу.

Анализ дистилляции на основе вариационного автокодировщика

Модель учителя обучается на выборке MNIST-Big, а модель ученика на выборке FashionMNIST-Small. В качестве отображения φ используется выход вариационного автокодировщика, переводящего изображения одежды в изображения цифр.

На графиках показаны метрики точности и кросс-энтропийной ошибки модели ученика.





Модель, использующая метки учителя с применением вариационого автокодировщика, показывает лучшее значение точности, но большее значение кросс-энтропийной ошибки, чем модель без учителя.

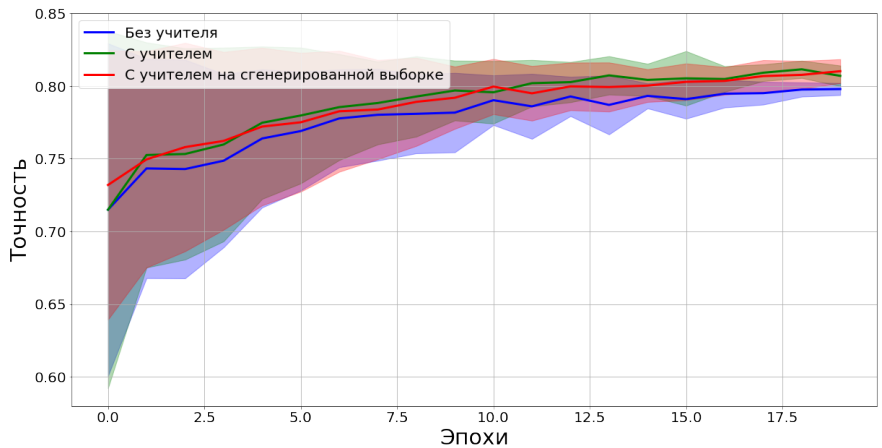
Анализ дистилляции на расширенной синтетически сгенерированной выборке

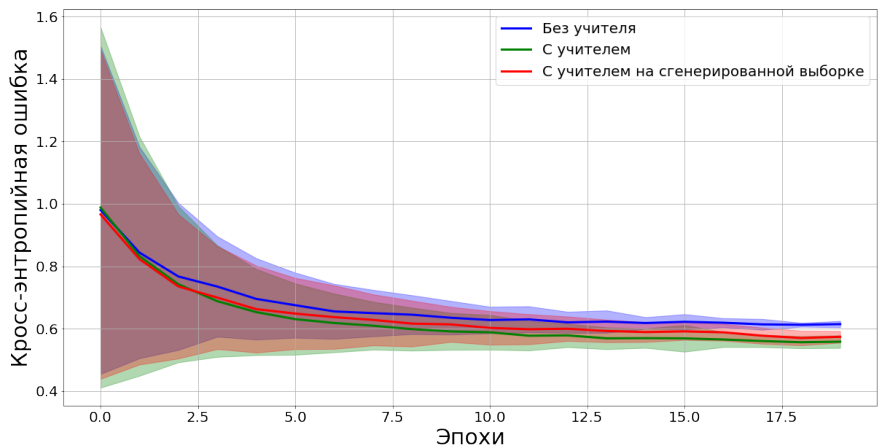
На основе малоресурсной части выборки FashionMNIST-Small сформируем новую выборку, сгенерировав для каждого объекта одежды 70 изображений цифр с помощью модели вариационного автокодировщика. Выборка разделяется на 2 части — обучающая и тестовая части. Обучающая часть разделяется на многоресурсную и малоресурсную части. Обучающая часть содержит 60000 объектов, тестовая часть содержит 10000 объектов, многоресурсная часть содержит 59000 объектов, малоресурсная часть содежит 1000 объектов.

Выборка	Пояснение	Размер выборки
GeneratedMNIST-Train	Обучающая часть	60000
GeneratedMNIST-Big	Многоресурсная часть	59000
GeneratedMNIST-Small	Малоресурсная часть	1000
GeneratedMNIST-Test	Тестовая часть	10000

Анализ дистилляции на расширенной синтетически сгенерированной выборке

Модель ученика обучается на малоресурсной части FashionMNIST-Small, модель учителя на многоресурсной части GeneratedMNIST-Big. В качестве отображения φ используется выход вариационного автокодировщика, переводящего изображения одежды в изображения цифр.
На графиках показаны метрики точности и кросс-энтропийной ошибки модели ученика.





Модель, использующая метки учителя с применением вариационого автокодировщика, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Выводы

1. Рассмотрена проблема понижения сложности модели при ее переносе к новым данным меньшей мощности.
2. Рассмотрены два основных подхода
   - дистилляция моделей,
   - доменная адаптация.
3. Предложен подход для случая, когда модели учителя и ученика заданы на выборках разной мощности с известным отображением между выборками.