Аннотация

Исследуется проблема понижения сложности аппроксимирующей модели при переносе на новые данные меньшей мощности. Вводятся понятия учителя, ученика для разных наборов данных. При этом мощность одного набора данных больше мощности другого. Рассматриваются методы, основанные на дистилляции моделей машинного обучения. Вводится предположение, что решение оптимизационной задачи от параметров обеих моделей и доменов повышает качество модели ученика. Проводится вычислительный эксперимент на реальных и синтетических данных.

Ключевые слова: адаптация доменов, дистилляция, нейронные сети, обучение с учителем

Содержание

1 Введение

Актуальность темы. Традиционно алгоритмы машинного обучения разрабатываются для отдельной задачи, аппроксимирующей заданную выборку. Сбор и обработка наборов данных для каждой новой задачи и области являются чрезвычайно дорогими и трудоемкими процессами, и не всегда могут быть доступны достаточные данные для обучения. К тому же для каждой новой задачи сложную модель необходимо перестраивать с нуля. В этом случае можно использовать перенос информации с более сложной модели *учителя* на модель *ученика* с меньшим числом параметров. С другой стороны снижение сложности модели — приоритетная задача, необходимая для повышения интерпретируемости моделей.

Цель работы. Одним из способов повышения качества алгоритма машинного обучения является использование модели с большим числом параметров, ответы которой можно использовать при обучении модели с меньшим числом параметров. Модели с меньшим числом параметров являются более интерпретируемыми. Цель данной работы заключается в снижении сложности модели машинного обучения, а также переходе к данным меньшей мощности. Для этого предлагается использовать два основных метода — дистилляция моделей и доменная адаптация.

Новизна. Предложен метод для случая, когда модели учителя и ученика заданы на выборках разной мощности из разных, но схожих генеральных совокупностей. При чем задано отображение с выборки меньшей мощности в выборку большей мощности.

1.1 Обзор предметной области

Определение 1.1. *Учитель — модель с большим числом параметров, ответы которой используются при обучении модели ученика.*

Определение 1.2. *Ученик — модель с меньшим числом параметров, при обучении которой используются ответы модели учителя.*

Определение 1.3. *Дистилляция модели — выбор модели с меньшим числом параметров из параметрического семейства функций согласно решению оптимизационной задачи с учетом ответов модели с большим числом параметров.*

Определение 1.4. *Привилегированные признаки — набор признаков, доступных только на этапе обучения модели.*

Дистилляция моделей машинного обучения использует метки модели с большим числом параметров для обучения модели с меньшим числом параметров. В [1] рассматривается метод дистилляции, предложенной Джеффри Хинтоном, с учетом меток учителя при помощи функции softmax с параметром температуры, а в [2] рассматривается объединение методов дистилляци, предложенной Джеффри Хинтоном, и привилегированной информации [2], предложенной Владимиром Наумовичем Вапником, в обобщенную дистилляцию. В вероятностной дистилляции вводится гипотеза порождения выборки совместно с ответами учителя. В работе [25] рассмотрена байесовская дистилляция моделей глубокого обучения, в рамках которой вместе с ответами учителя используется апостериорное распределение параметров модели учителя. На основе этого апостериорного распределения задается априорное распределение модели ученика. Дистиляцция моделей используется в широком классе задач. В [4] рассматривается метод дистилляции моделей для задачи распознавания речи. В [18] рассматривается метод дистилляции моделей для задачи распознавания объектов с использованием взвешенной кросс-энтропии для улучшения качества на выборках с неслабансированными классами. В [20] рассматривается метод дистилляции моделей для задачи семантической сегментации с использованием GAN. В [19] предлагается усовершенствованный метод дистилляции с использованием помимо модели учителя также и модели помощника — сети среднего размера между размерами учителя и ученика.

В задаче доменнной адаптации используются наборы данных, схожих между собой. В общем случае выборки состоят из объектов, которые можно разделить на домены из близких генеральных совокупностей [9, 13]. К примеру, можно составить отображение из множества реальных фотографий малой мощности во множество сгенерированных движком изображений, мощность которого естественно больше [10, 12]. Так, в [9, 22] рассматриваются отображения, изменяющие стиль изображений. Одним из примеров генерации новых изображений является работа модели вариационного автокодировщика [8], способного для одного и того же объекта строить вероятностное распределение, на основе которого можно получить целое семейство новых объектов. Другим примером может служить работа генеративной состязательной сети GAN [23], в которой одновременно обучаются две модели. Модель генератора учится порождать новые объекты из шума, а модель дискриминатора учится отличать их от реальных объектов. Данные методы позволяют получить синтетически сгенерированную выборку из близкой генеральной совокупности к исходной выборке.

Различные постановки задач доменной адаптации описываются в [5], встречаются постановки с частично размеченным целевым доменом и неразмеченным вовсе. Задачи с неразмеченным целевым доменом направлены на то, чтобы модели, обученные на синтетических данных, адаптировались к реальным данным [12]. Таким образом, доменная адаптация использует размеченные данные нескольких исходных доменов для выполнения новых задач в целевом домене. В [4] рассматривается метод доменной адаптации для задачи распознавания речи. В [21] рассматривается метод доменной адаптации для обучения акустических моделей на основе дистилляции моделей.

Одной из постановок задач доменной адаптации является перенос стиля изображений [9, 22]. Так, в [9] предлагается использовать селфи-изображения в качестве исходного домена и перевести их в изображения желаемого художественного стиля на основе выборки из требуемого стиля. Таким образом можно использовать синтетически сгенерированные данные для обучения.

В рассмотренных выше методах дистилляции [1, 2, 25] рассмат-

ривается случай, когда модели учителя и ученика аппроксимируют выборки из разных генеральных совокупностей. Для задачи дистилляции, предложенной Джеффри Хинтоном [1], исходный и целевой наборы данных совпадают. Типичной задачей дистилляции моделей на многодоменных выборках является задача машинного перевода текстов, описанная в [3].

1.2 Предложенный метод

Предлагается при обучении модели ученика использовать помимо меток учителя на одном из доменов также и связь между доменами. При этом в качестве доменов должны служить близкие генеральные совокупности.

Определение 1.5. *Генеральная совокупность объектов B называется близкой к совокупности A, если существует инъективное отображение $\varphi : A \to B$*

Таким образом в качестве доменов разной мощности могут служить настоящие и сгенерированные изображения. В качестве отображений между изображениями рассматриваются нормальный шум, сверточные преобразования и генерация изображений с помощью модели вариацинного автокодировщика [8]. При этом исследуются отображения, для которых существования обратных не рассматриваются. Ожидается, что качество полученных моделей на одном домене будет превышать качество моделей, в обучении которых не использовались метки учителя на другом домене.

В качестве экспериментальных данных используются реальные данные и синтетическая выборка. В качестве реальных данных рассматривается выборка Fashion-MNIST [6], состоящая из изображений одежды, и выборка MNIST [7], состоящая из изображений рукописных цифр.

2 Постановка задачи

2.1 Базовая постановка задачи дистилляции

Задана выборка

$$\mathfrak{D} = (\mathbf{X}, \mathbf{Y}), \quad \mathbf{X} \in \mathbb{X}, \quad \mathbf{Y} \in \{1, ..., R\},$$

где R — число классов в задаче классификации.

Предполагается, что задана обученная модель с большим числом параметров — модель учителя. Модель учителя $\mathbf{f}$ принадлежит параметрическому семейству функций:

$$\mathfrak{F} = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^R\}.$$

Требуется обучить модель ученика с меньшим числом параметров с учетом ответов учителя. Модель ученика $\mathbf{g}$ принадлежит параметрическому семейству функций:

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{z}(\mathbf{x})/T), \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^R\},$$

где $\mathbf{v}, \mathbf{z}$ — дифференцируемые параметрические функции заданной структуры, T — параметр температуры со свойствами:

1. при $T \to 0$ один из классов имеет единичную вероятность;
2. при $T \to \infty$ все классы равновероятны.

Функция потерь $\mathcal{L}$, учитывающая модель учителя $\mathbf{f}$ при выборе модели ученика $\mathbf{g}$, имеет вид:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}) = &- \sum_{i=1}^{m} \sum_{r=1}^{R} y_i^r \log g^r(x_i)\big|_{T=1} \\ &- \sum_{i=1}^{m} \sum_{r=1}^{R} f^r(x_i)\big|_{T=T_0} \log g^r(x_i)\big|_{T=T_0},\end{aligned}$$

где $\cdot\big|_{T=t}$ означает, что параметр температуры T в предыдущей функции равен t.

Получаем оптимизационную задачу:

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

2.2 Постановка задачи дистилляции для многодоменной выборки

Заданы две выборки:

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

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

где $\mathfrak{D}_\mathrm{s}, \mathfrak{D}_\mathrm{t}$ — исходный и целевой наборы данных. В базовой постановке задачи дистилляции предполагается, что $\mathfrak{D}_\mathrm{t} \subset \mathfrak{D}_\mathrm{s}, \mathbb{X}_\mathrm{t} = \mathbb{X}_\mathrm{s}$.

Предполагается, что число объектов в выборках не совпадают:

$$|\mathbf{X}_\mathrm{s}| \gg |\mathbf{X}_\mathrm{t}|$$

Пусть при этом задана модель учителя на выборке большей мощности:

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

где $\mathbf{f}$ — модель учителя, $\mathbb{Y}'$ — пространство оценок.

Задана связь между исходной и целевой выборками:

$$\varphi : \mathbb{X}_\mathrm{t} \to \mathbb{X}_\mathrm{s},$$

где φ — инъективное отображение.

Требуется получить модель ученика для малоресурсной выборки:

$$\mathbf{g} : \mathbb{X}_\mathrm{t} \to \mathbb{Y}',$$

где $\mathbf{g}$ — модель ученика.

В работе рассматривается функция потерь, учитывающая метки учителя и связь между доменами:

1. для задачи регрессии:

$$\begin{aligned}\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{Y}, \mathbf{f}, \varphi) =& \lambda \|\mathbf{y} - \mathbf{g}(\mathbf{x}, \mathbf{w})\|_2^2 \\ &+ (1-\lambda)\|\mathbf{g}(\mathbf{x}, \mathbf{w}) - (\mathbf{f} \circ \varphi)(\mathbf{x})\|_2^2;\end{aligned}$$

2. для задачи классификации:

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

   где λ — метапараметр, задающий вес дистилляции, $\mathbb{I}$ — индикаторная функция.

Получаем оптимизационную задачу:

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

3 Вычислительный эксперимент

Для анализа моделей, полученных при помощи дистилляции с модели учителя в модель ученика, проводится вычислительный эксперимент для задачи классификаци и регрессии.

Эксперимент проводится для выборок FashionMNIST [6] — набора изображений предметов одежды и MNIST [7] — набора изображений рукописных цифр. В качестве модели учителя $\mathbf{f}$ и модели ученика $\mathbf{g}$ рассматривается многослойный перцептрон с четырьми и одним скрытыми слоями соответственно:

Таблица 1: Описание моделей

	Учитель	Ученик
Структура	[784,256,128,64,64,10]	[784,64,10]
Число параметров	246400	50816

Функция активации после каждого скрытого слоя — ReLu. Для решения оптимизационной задачи используется градиентный метод оптимизации Adam [11].

Каждая из выборок состоит из обучающей и тестовой части, при этом обучающая часть разделяется на многоресурсную и малоресурсную части. Обучающая часть содержит 60000 объектов, многоресурсная часть содержит 59000 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 10000 объектов.

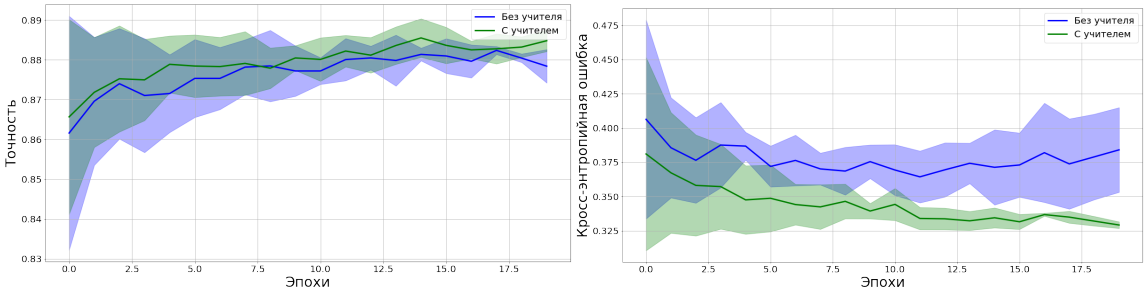
Для анализа качества дистилляции в работе [24] предложен интегральный критерий качества.

Таблица 2: Выборки

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

3.1 Анализ базовой дистилляции

Обучение на всей выборке. Модели учителя и ученика обучаются на обучающей части FashionMNIST-Train.



(a) (b)

Рис. 1: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

На рис.1а показан график зависимости метрики точности на отложенной тестовой выборке между истинными метками объектов и метками предсказанными моделью ученика.

На рис.1б показан график зависимости кросс-энтропийной ошибки на отложенной тестовой выборке между истинными метками объ-

ектов и вероятностями, предсказанными моделью ученика.

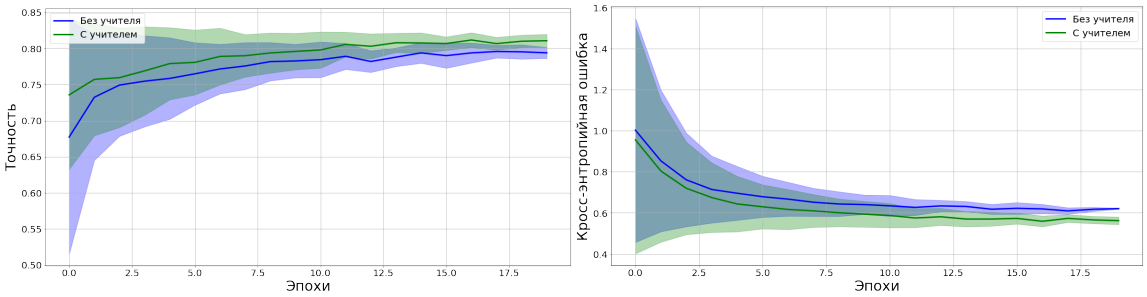
На графиках видно, что модель, использующая метки учителя, показывает лучшее значение точности, при этом наблюдается значительное снижение кросс-энтропийной ошибки.

Таблица 3: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка	Интегральный критерий
FashionMNIST-Train	—	—	$0{,}878 \pm 0{,}004$	$0{,}384 \pm 0{,}031$	$7{,}151 \pm 0{,}459$
FashionMNIST-Train	FashionMNIST-Train	—	$\mathbf{0{,}885 \pm 0{,}003}$	$\mathbf{0{,}329 \pm 0{,}002}$	$\mathbf{6{,}520 \pm 0{,}303}$

В таблице 3 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

Обучение на малоресурсной части. Модель учителя обучается на многоресурсной части FashionMNIST-Big, а модель ученика обучается на малоресурсной части FashionMNIST-Small.



(a) (b)

Рис. 2: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

На рис.2а показан график зависимости метрики точности на отложенной тестовой выборке между истинными метками объектов и метками, предсказанными моделью ученика.

На рис.2б показан график зависимости кросс-энтропийной ошибки на отложенной тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что модель, использующая метки учителя, показывает лучшее значение точности, при этом наблюдается снижение кросс-энтропийной ошибки.

Таблица 4: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка	Интегральный критерий
FashionMNIST-Train	—	—	$0{,}878 \pm 0{,}004$	$0{,}384 \pm 0{,}031$	$7{,}151 \pm 0{,}459$
FashionMNIST-Train	FashionMNIST-Train	—	$\mathbf{0{,}885 \pm 0{,}003}$	$\mathbf{0{,}329 \pm 0{,}002}$	$\mathbf{6{,}520 \pm 0{,}303}$
FashionMNIST-Small	—	—	$0{,}794 \pm 0{,}008$	$0{,}621 \pm 0{,}002$	$12{,}728 \pm 1{,}743$
FashionMNIST-Small	FashionMNIST-Big	—	$\mathbf{0{,}811 \pm 0{,}009}$	$\mathbf{0{,}562 \pm 0{,}018}$	$\mathbf{11{,}803 \pm 1{,}885}$

В таблице 4 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

Обучение на выборке с 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.

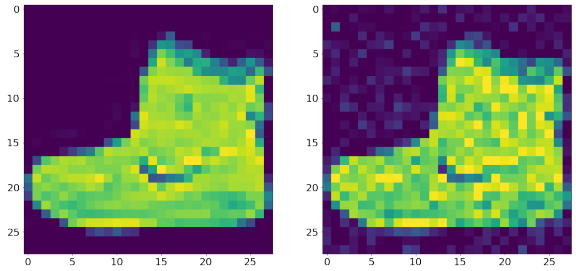
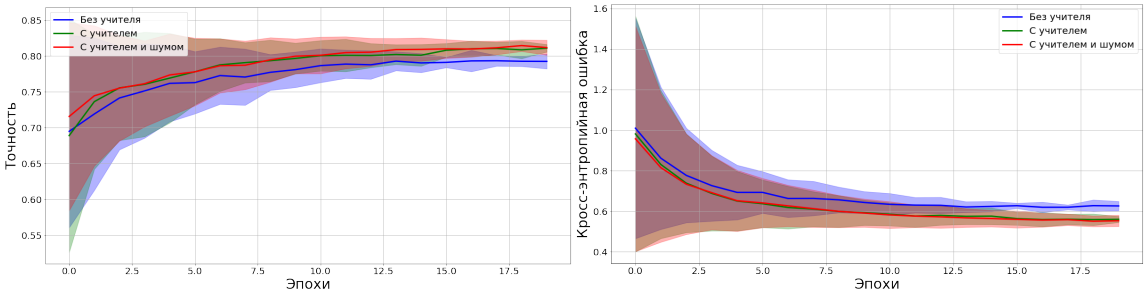


Рис. 3: Сравнение объекта выборки до и после добавления шума



(a) (b)

Рис. 4: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

На рис.4а показан график зависимости метрики точности на отложенной тестовой выборке между истинными метками объектов и метками, предсказанными моделью ученика.

На рис.4б показан график зависимости кросс-энтропийной ошибки на отложенной тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что значения точности и кросс-энтропийной ошибки модели, использующей метки учителя на выборке с шумом,

лежат между соответствующими значениями для модели без учителя и для модели, использующей метки учителя на выборке без шума.

Таблица 5: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка	Интегральный критерий
FashionMNIST-Train	—	—	$0{,}878 \pm 0{,}004$	$0{,}384 \pm 0{,}031$	$7{,}151 \pm 0{,}459$
FashionMNIST-Train	FashionMNIST-Train	—	$\mathbf{0{,}885 \pm 0{,}003}$	$\mathbf{0{,}329 \pm 0{,}002}$	$\mathbf{6{,}520 \pm 0{,}303}$
FashionMNIST-Small	—	—	$0{,}794 \pm 0{,}008$	$0{,}621 \pm 0{,}002$	$12{,}728 \pm 1{,}743$
FashionMNIST-Small	FashionMNIST-Big	—	$\mathbf{0{,}811 \pm 0{,}009}$	$\mathbf{0{,}562 \pm 0{,}018}$	$\mathbf{11{,}803 \pm 1{,}885}$
FashionMNIST-Small	FashionMNIST-Big	Noise	$0{,}812 \pm 0{,}010$	$0{,}553 \pm 0{,}028$	$11{,}800 \pm 2{,}098$
FashionMNIST-Small	FashionMNIST-Big	Dilation	$0{,}808 \pm 0{,}006$	$0{,}564 \pm 0{,}020$	$11{,}921 \pm 1{,}973$
FashionMNIST-Small	MNIST-Big	VAE	$\mathbf{0{,}803 \pm 0{,}012}$	$\mathbf{0{,}631 \pm 0{,}022}$	$\mathbf{13{,}123 \pm 2{,}063}$
FashionMNIST-Small	MNIST-Big	—	$0{,}457 \pm 0{,}017$	$1{,}227 \pm 0{,}030$	$24{,}184 \pm 1{,}940$
FashionMNIST-Small	GeneratedMNIST-Big	VAE	$0{,}806 \pm 0{,}006$	$0{,}588 \pm 0{,}014$	$12{,}393 \pm 1{,}684$

В таблице 5 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

Получаем, что шум в выборке не влияет на качество.

Обучение на выборке с dilation. 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.

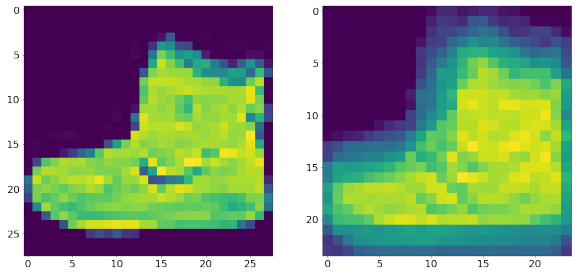
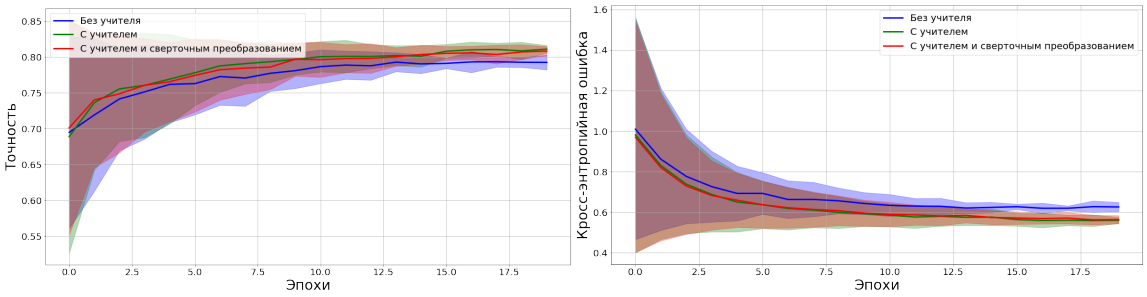


Рис. 5: Сравнение объекта выборки до и после преобразования

На рис.6а показан график зависимости метрики точности на отложенной тестовой выборке между истинными метками объектов и метками, предсказанными моделью ученика.



(a) (b)

Рис. 6: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

На рис.6б показан график зависимости кросс-энтропийной ошибки на отложенной тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что значения точности и кросс-энтропийной ошибки модели, использующей метки учителя на выборке с преобразованием, лежат между соответствующими значениями для модели без учителя и для модели, использующей метки учителя на выборке без преобразования.

Таблица 6: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка	Интегральный критерий
FashionMNIST-Train	—	—	$0{,}878 \pm 0{,}004$	$0{,}384 \pm 0{,}031$	$7{,}151 \pm 0{,}459$
FashionMNIST-Train	FashionMNIST-Train	—	$\mathbf{0{,}885 \pm 0{,}003}$	$\mathbf{0{,}329 \pm 0{,}002}$	$\mathbf{6{,}520 \pm 0{,}303}$
FashionMNIST-Small	—	—	$0{,}794 \pm 0{,}008$	$0{,}621 \pm 0{,}002$	$12{,}728 \pm 1{,}743$
FashionMNIST-Small	FashionMNIST-Big	—	$\mathbf{0{,}811 \pm 0{,}009}$	$\mathbf{0{,}562 \pm 0{,}018}$	$\mathbf{11{,}803 \pm 1{,}885}$
FashionMNIST-Small	FashionMNIST-Big	Noise	$0{,}812 \pm 0{,}010$	$0{,}553 \pm 0{,}028$	$11{,}800 \pm 2{,}098$
FashionMNIST-Small	FashionMNIST-Big	Dilation	$0{,}808 \pm 0{,}006$	$0{,}564 \pm 0{,}020$	$11{,}921 \pm 1{,}973$

В таблице 6 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

3.2 Вариационный автокодировщик

В качестве преобразования элементов выборки FashionMNIST [6] в элементы выборки MNIST [7] используем модель вариационного автокодировщика [8], аппроксимирующую отображение φ.

Базовая модель автокодировщика. Данная модель состоит из двух частей. Сначала строится вероятностное распределение в скрытом пространстве, которое позволяет генерировать скрытые представления для одного объекта. Далее с помощью декодировщика строится вероятностное распределение, позволяющее генерировать реконструкции исходного объекта.

1. $q(\mathbf{z}|\mathbf{x}, \alpha)$ — вероятностный кодировщик, где α — параметры кодировщика;

2. $p(\hat{\mathbf{x}}|\mathbf{z}, \beta)$ — вероятностный декодировщик, где β — параметры декодировщика;

3. функция потерь:

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

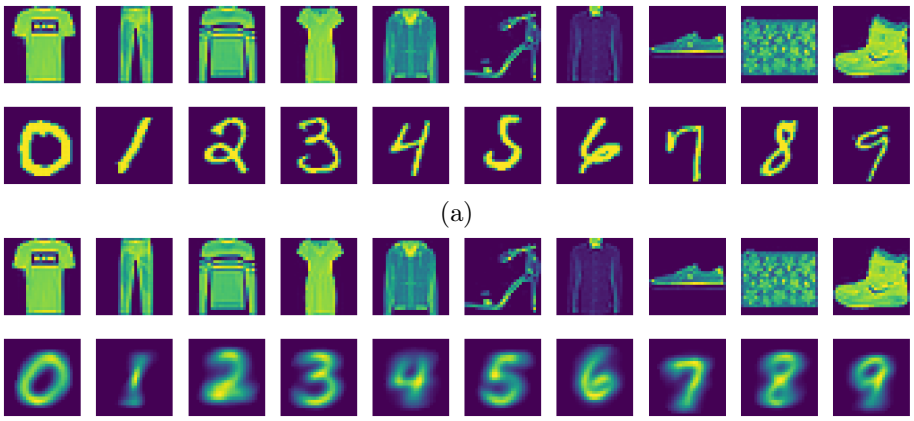
   где $p(\mathbf{z}) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ — априорное распределение.

Получаем оптимизационную задачу:

$$\hat{\alpha}, \hat{\beta} = \arg\max_{\alpha,\beta} \mathcal{L}(\alpha, \beta).$$

Генерация отображения из FashionMNIST в MNIST. Воспользуемся моделью вариационного автокодировщика [8] для преобразования изображений одежды из выборки FashionMNIST [6] в изображения цифр на основе выборки MNIST [7].

Сгенерируем синтетическую выборку, где каждому изображению одежды выборки FashionMNIST-Train будет соответстовать случайное изображение цифры из выборки MNIST-Train из того же класса.



(a)

(b)

Рис. 7: a) Объекты синтетический выборки; b) Объекты исходной выборки до и после работы автокодировщика

Далее воспользуемся моделью вариационного автокодировщика, состоящего из одного кодировщика и двух декодировщиков, соответствующих генерации объектов цифр и одежды соответственно. Используем модель с размером скрытого представления, равным 64.

На основе полученной выборки обучим модель вариационного автокодировщика, минимизируя ошибку между выходом модели и исходным значением — изображением одежды и ошибку между выходом модели и целевым значением — изображением цифр, соответствующего исходному объекту.

Полученная модель генеририрует семейство новых объектов — изображений цифры и изображений одежды для одного и того же изображения одежды.

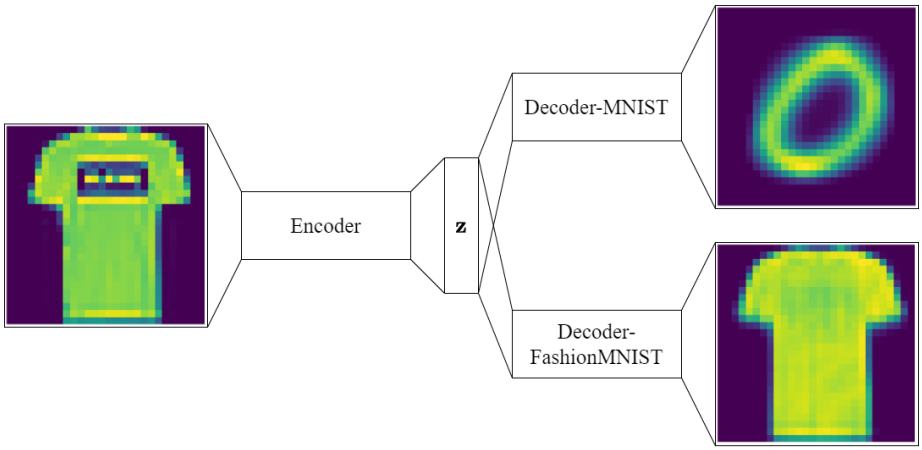
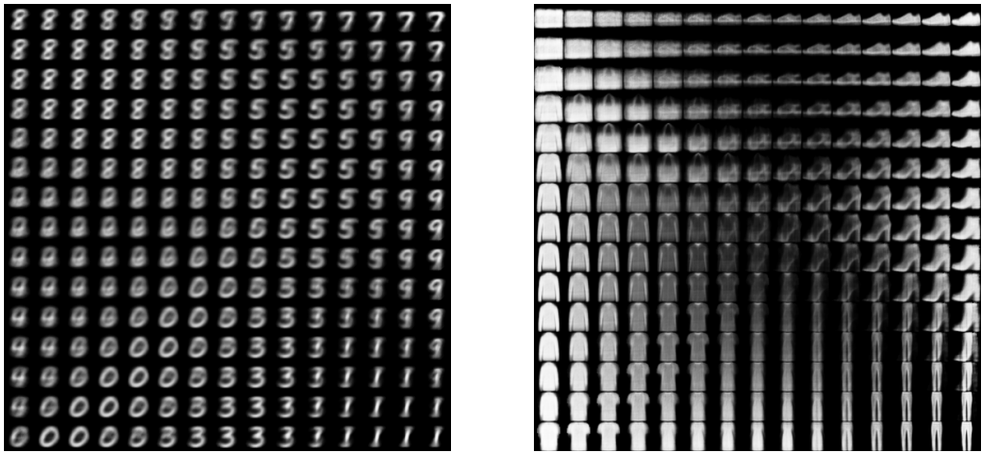


Рис. 8: Пример работы вариационного автокодировщика

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



(a) (b)

Рис. 9: Зависимость выхода модели от изменения вектора в скрытом представлении для генерации a) цифр; b) одежды

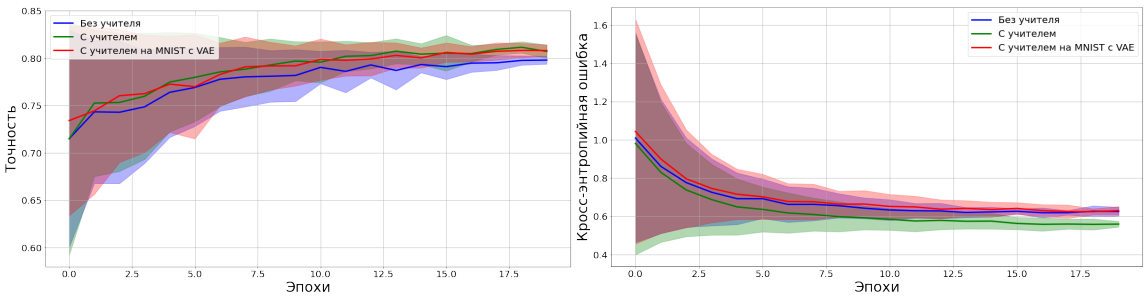
Видно, что при варьировании вектора в скрытом состоянии меняется также и выход автокодировщика. При этом всегда наблюдается равенство классов полученных цифр и предметов одежды, как представлено на рис.7a. То есть рис.9 не противоречит рис.7

Согласно определению 1.5 существует отображение из выборки FashionMNIST в выборку MNIST. Значит данные выборки являются близкими генеральными совокупностями и их можно использовать для задачи дистилляции для многодоменной выборки.

3.3 Анализ качества модели, предложенной на основе вариационного автокодировщика

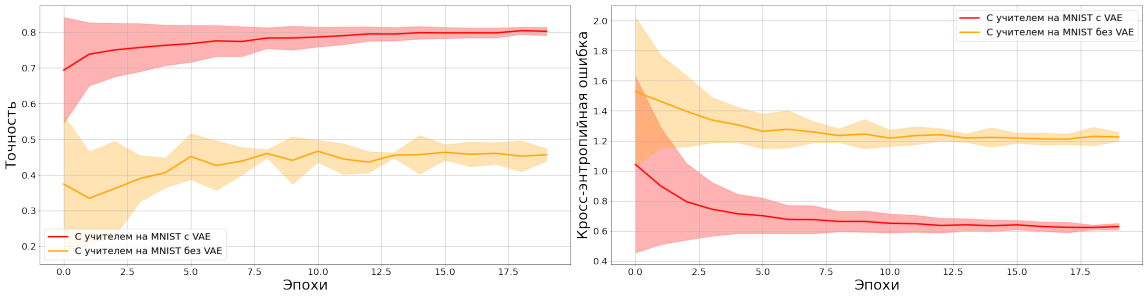
Модель учителя обучается на выборке MNIST-Big, а модель ученика на выборке FashionMNIST-Small. При этом обученная модель ученика использует метки учителя, на вход которого подается выход вариационного автокодировщика [8], переводящего изображения одежды в изображения цифр.

Сравнивается качество аппроксимации без использования вариационного автокодировщика: модель ученика обучается на выборке FashionMNIST-Small, модель учителя обучается на выборке MNIST-Big и используется при обучении ученика, получая на вход изображения одежды без преобразования вариационным автокодировщиком.



(a) (b)

Рис. 10: Качество аппроксимации при использовании VAE на малодоменной выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками



(a) (b)

Рис. 11: Сравнение качества аппроксимации в зависимости от использования VAE на малодоменной выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

На графиках видно, что без использования отображения φ модель становится более шумной с явным понижением качества аппроксимации.

Таблица 7: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка	Интегральный критерий
FashionMNIST-Train	—	—	$0{,}878 \pm 0{,}004$	$0{,}384 \pm 0{,}031$	$7{,}151 \pm 0{,}459$
FashionMNIST-Train	FashionMNIST-Train	—	$\mathbf{0{,}885 \pm 0{,}003}$	$\mathbf{0{,}329 \pm 0{,}002}$	$\mathbf{6{,}520 \pm 0{,}303}$
FashionMNIST-Small	—	—	$0{,}794 \pm 0{,}008$	$0{,}621 \pm 0{,}002$	$12{,}728 \pm 1{,}743$
FashionMNIST-Small	FashionMNIST-Big	—	$\mathbf{0{,}811 \pm 0{,}009}$	$\mathbf{0{,}562 \pm 0{,}018}$	$\mathbf{11{,}803 \pm 1{,}885}$
FashionMNIST-Small	FashionMNIST-Big	Noise	$0{,}812 \pm 0{,}010$	$0{,}553 \pm 0{,}028$	$11{,}800 \pm 2{,}098$
FashionMNIST-Small	FashionMNIST-Big	Dilation	$0{,}808 \pm 0{,}006$	$0{,}564 \pm 0{,}020$	$11{,}921 \pm 1{,}973$
FashionMNIST-Small	MNIST-Big	VAE	$\mathbf{0{,}803 \pm 0{,}012}$	$\mathbf{0{,}631 \pm 0{,}022}$	$\mathbf{13{,}123 \pm 2{,}063}$
FashionMNIST-Small	MNIST-Big	—	$0{,}457 \pm 0{,}017$	$1{,}227 \pm 0{,}030$	$24{,}184 \pm 1{,}940$

В таблице 7 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

3.4 Анализ качества модели на расширенной синтетически сгенерированной выборке

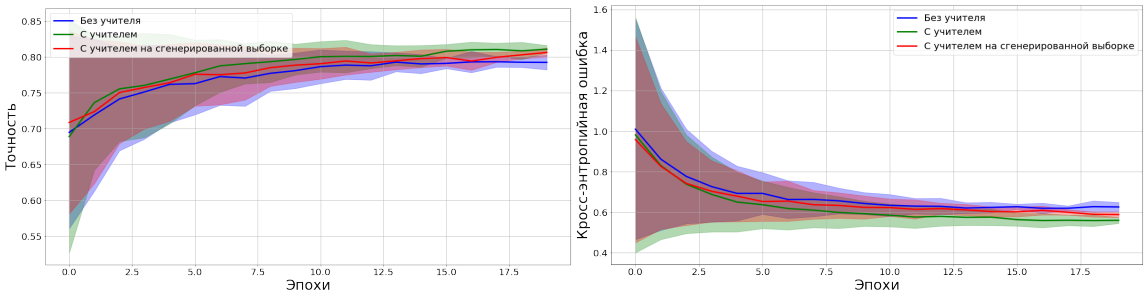
На основе малоресурсной части выборки FashionMNIST-Small сформируем новую выборку, сгенерировав для каждого объекта одежды 70 изображений цифр с помощью модели вариационного автокодировщика [8]. Полученная выборка состоит из обучающей и тестовой части, при этом обучающая часть разделяется на многоресурсную и малоресурсную части. Обучающая часть содержит 60000 объектов, многоресурсная часть содержит 59000 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 10000 объектов.

Таблица 8: Расширенная сгенерированная выборка

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

Модель ученика обучается на малоресурсной части FashionMNIST-Small, модель учителя на многоресурсной части GeneratedMNIST-Big сгенерированной расширенной выборки и используется при обучении ученика.

На графиках видно, что значения точности и кросс-энтропийной ошибки модели, использующей метки учителя, обученного на сгенерированной расширенной выборке, лежат между соответствующими значениями для модели без учителя и для модели, использующей метки учителя, обученного на многоресурсной части выборки.



(a) (b)

Рис. 12: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

3.5 Анализ дистилляци на основе преобразования стиля изображений

Используем подвыборку ImageNet — набора изображений, для которого нужно решить задачу классификации на 10 классов. Выборка состоит из обучающей и тестовой части, при этом обучающая часть разделяется на многоресурсную и малоресурсную части. Обучающая часть содержит 9469 объектов, многоресурсная часть содержит 8469 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 3925 объектов.

Таблица 9: Выборка ImageNet

Выборка	Пояснение	Размер выборки
ImageNet-Train	Обучающая часть	9469
ImageNet-Big	Многоресурсная часть	8469
ImageNet-Small	Малоресурсная часть	1000
ImageNet-Test	Тестовая часть	3925

В качестве модели учителя $\mathbf{f}$ рассматривается нейронная сеть с пятью сверточными слоями и тремя полносвязными слоями, в качестве модели ученика рассматривается нейронная сеть с двумя свер-

точными слоями и двумя полносвязными слоями. Функция активации после каждого скрытого слоя — ReLU.

Таблица 10: Структура учителя

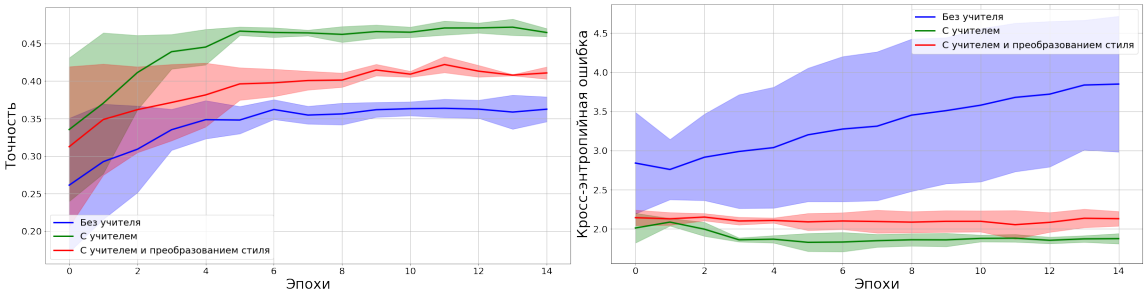
Слой	Размер входного вектора	Число параметров
Входной слой	(3, 200, 200)	0
CONV1 (kernel size=5)	(24, 196, 196)	1800
POOL1	(24, 98, 98)	0
CONV2 (kernel size = 5)	(48, 94, 94)	28800
POOL2	(48, 47, 47)	0
CONV3 (kernel size = 8)	(96, 40, 40)	294912
POOL3	(96, 20, 20)	0
CONV4 (kernel size = 5)	(192, 16, 16)	460800
POOL4	(192, 8, 8)	0
CONV5 (kernel size = 7)	(384, 2, 2)	3612672
POOL5	(384, 1, 1)	0
Полносвязный слой	(384)	0
Полносвязный слой	(120)	46080
Полносвязный слой	(84)	10080
Полносвязный слой	(10)	840
		$\sum = 4455984$

Таблица 11: Структура ученика

Слой	Размер входного вектора	Число параметров
Входной слой	(3, 200, 200)	0
CONV1 (kernel size=5)	(24, 196, 196)	1800
POOL1	(24, 98, 98)	0
CONV2 (kernel size = 5)	(48, 94, 94)	28800
POOL2	(48, 47, 47)	0
Полносвязный слой	(106032)	0
Полносвязный слой	(120)	12723840
Полносвязный слой	(10)	1200
		$\sum = 12755640$

Применим к многоресурсной части ImageNet-Big преобразование стиля на основе сверточной нейронной сети VGG-19 и обучим на ней модель учителя. Модель ученика обучается на малоресурсной части ImageNet-Small без преобразования.

Рис. 13: Сравнение объекта выборки до и после преобразования



(a) (b)

Рис. 14: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 3 запускам. a) точность; b) кросс-энтропийная ошибка между истинными и предсказанными учеником метками

На рис.6а показан график зависимости метрики точности на отложенной тестовой выборке между истинными метками объектов и метками, предсказанными моделью ученика.

На рис.6б показан график зависимости кросс-энтропийной ошибки на отложенной тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что значения точности и кросс-энтропийной ошибки модели, использующей метки учителя на выборке с преобразованием, лежат между соответствующими значениями для модели без учителя и для модели, использующей метки учителя на выборке без преобразования.

Таблица 12: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка/ Среднеквадратичная ошибка	Интегральный критерий
ImageNet-Small	—	—	$0{,}363 \pm 0{,}017$	$3{,}849 \pm 0{,}866$	$46{,}615 \pm 11{,}498$
ImageNet-Small	ImageNet-Big	—	$\mathbf{0{,}465 \pm 0{,}005}$	$\mathbf{1{,}876 \pm 0{,}066}$	$\mathbf{26{,}488 \pm 0{,}996}$
ImageNet-Small	ImageNet-Big	StyleTransfer	$0{,}411 \pm 0{,}008$	$2{,}131 \pm 0{,}093$	$29{,}476 \pm 1{,}495$

В таблице 12 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

3.6 Анализ дистилляции для задачи регрессии

Сгенерируем синтетическую выборку из нормального распределения и разделим ее на обучающую и тестовую часть. При этом обучающая часть разделяется на многоресурсную и малоресурсную части. Обучающая часть содержит 9000 объектов, многоресурсная часть содежит 8700 объектов, малоресурсная часть содержит 300 объектов, а тестовая часть содержит 1000 объектов.

Таблица 13: Выборки

Выборка	Пояснение	Размер выборки
Reg-Train	Обучающая часть	9000
Reg-Big	Многоресурсная часть	8700
Reg-Small	Малоресурсная часть	300
Reg-Test	Тестовая часть	1000

Обучение на всей выборке. Модели учителя и ученика обучаются на обучающей части Reg-Train.

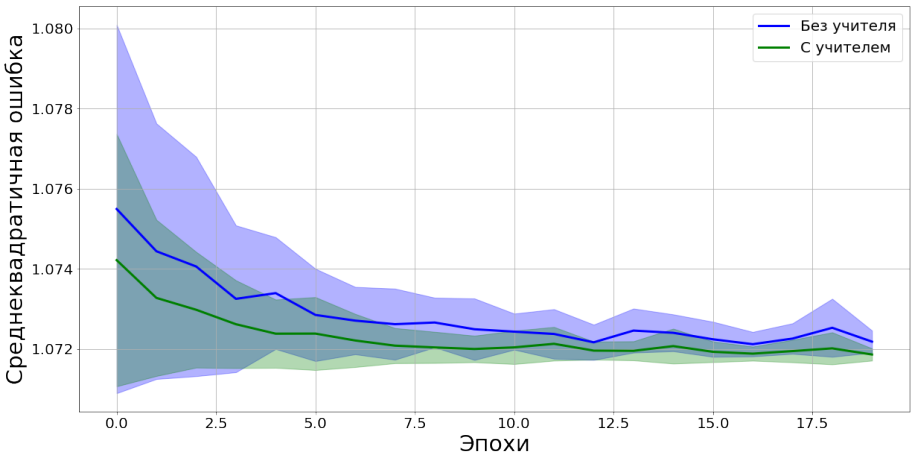


Рис. 15: Среднеквадратичная ошибка между истинными и предсказанными учеником значениями на тестовой выборке. Все результаты усреднены по 5 запускам.

На рис.13 показан график зависимости среднеквадратичной ошибки на отложенной тестовой выборке между истинными значениями объектов и значениями, предсказанными моделью ученика.

На графике видно, что модель, использующая ответы учителя, показывает лучшее значение среднеквадратичной ошибки.

Таблица 14: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка/ Среднеквадратичная ошибка	Интегральный критерий
Reg-Train	—	—	—	$1{,}0647 \pm 0{,}0001$	—
Reg-Train	Reg-Train	—	—	$\mathbf{1{,}0645 \pm 0{,}0001}$	—

В таблице 14 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

Обучение на малоресурсной части. Модель учителя обучается на многоресурсной части Reg-Big, а модель ученика обучается на малоресурсной части Reg-Small.

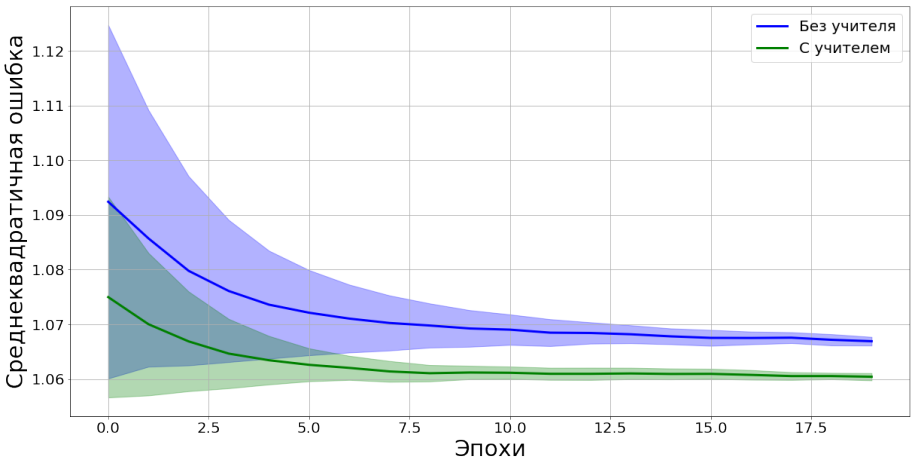


Рис. 16: Среднеквадратичная ошибка между истинными и предсказанными учеником значениями на тестовой выборке. Все результаты усреднены по 5 запускам.

На рис.14 показан график зависимости среднеквадратичной ошибки на отложенной тестовой выборке между истинными значениями

объектов и значениями, предсказанными моделью ученика.

На графике видно, что модель, использующая ответы учителя, показывает лучшее значение среднеквадратичной ошибки.

Таблица 15: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка/ Среднеквадратичная ошибка	Интегральный критерий
Reg-Train	—	—	—	$1{,}0647 \pm 0{,}0001$	—
Reg-Train	Reg-Train	—	—	$\mathbf{1{,}0645 \pm 0{,}0001}$	—
Reg-Small	—	—	—	$1{,}0804 \pm 0{,}0004$	—
Reg-Small	Reg-Big	—	—	$\mathbf{1{,}0755 \pm 0{,}0004}$	—

В таблице 15 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

Обучение на выборке с преобразованием Добавим к многоресурсной части Reg-Big преобразование $\sin x$ и обучим на ней модель учителя. Модель ученика обучается на малоресурсной части Reg-Small без преобразования.

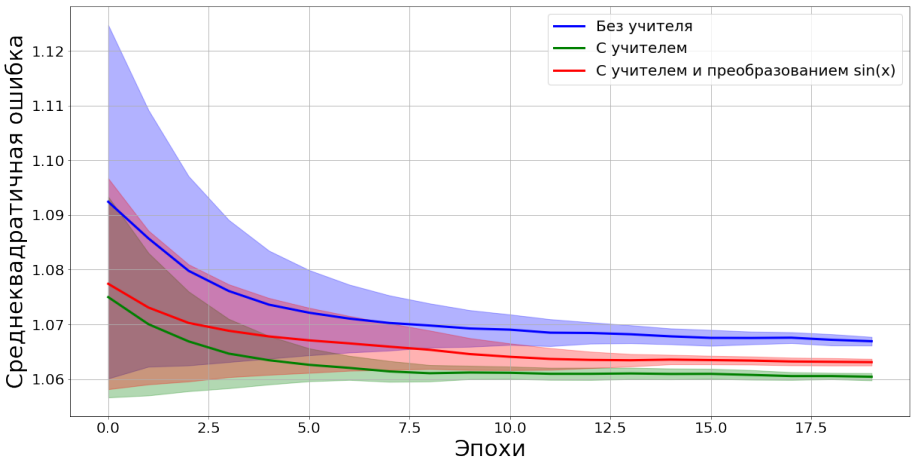


Рис. 17: Среднеквадратичная ошибка между истинными и предсказанными учеником значениями на тестовой выборке. Все результаты усреднены по 5 запускам.

На рис.15 показан график зависимости среднеквадратичной ошибки на отложенной тестовой выборке между истинными значениями объектов и значениями, предсказанными моделью ученика.

На графике видно, что модель, использующая ответы учителя, показывает лучшее значение среднеквадратичной ошибки.

Таблица 16: Качество моделей

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка/ Среднеквадратичная ошибка	Интегральный критерий
Reg-Train	—	—	—	$1{,}0647 \pm 0{,}0001$	—
Reg-Train	Reg-Train	—	—	$\mathbf{1{,}0645 \pm 0{,}0001}$	—
Reg-Small	—	—	—	$1{,}0804 \pm 0{,}0004$	—
Reg-Small	Reg-Big	—	—	$\mathbf{1{,}0755 \pm 0{,}0004}$	—
Reg-Small	Reg-Big	Sin	—	$1{,}0782 \pm 0{,}0003$	—

В таблице 16 представлены результаты сравнения моделей ученика, полученных с использованием и без использования дистилляции.

3.7 Код вычислительного эксперимента

Весь код вычислительного эксперимента представлен в [26]. Также доступны письменный отчет и результаты экспериментов.

4 Заключение

Таблица 17: Результаты экспериментов

Ученик	Учитель	Отображение φ	Точность	Кросс-энтропийная ошибка/ Среднеквадратичная ошибка	Интегральный критерий
FashionMNIST-Train	—	—	$0{,}878 \pm 0{,}004$	$0{,}384 \pm 0{,}031$	$7{,}151 \pm 0{,}459$
FashionMNIST-Train	FashionMNIST-Train	—	$\mathbf{0{,}885 \pm 0{,}003}$	$\mathbf{0{,}329 \pm 0{,}002}$	$\mathbf{6{,}520 \pm 0{,}303}$
FashionMNIST-Small	—	—	$0{,}794 \pm 0{,}008$	$0{,}621 \pm 0{,}002$	$12{,}728 \pm 1{,}743$
FashionMNIST-Small	FashionMNIST-Big	—	$\mathbf{0{,}811 \pm 0{,}009}$	$\mathbf{0{,}562 \pm 0{,}018}$	$\mathbf{11{,}803 \pm 1{,}885}$
FashionMNIST-Small	FashionMNIST-Big	Noise	$0{,}812 \pm 0{,}010$	$0{,}553 \pm 0{,}028$	$11{,}800 \pm 2{,}098$
FashionMNIST-Small	FashionMNIST-Big	Dilation	$0{,}808 \pm 0{,}006$	$0{,}564 \pm 0{,}020$	$11{,}921 \pm 1{,}973$
FashionMNIST-Small	MNIST-Big	VAE	$\mathbf{0{,}803 \pm 0{,}012}$	$\mathbf{0{,}631 \pm 0{,}022}$	$\mathbf{13{,}123 \pm 2{,}063}$
FashionMNIST-Small	MNIST-Big	—	$0{,}457 \pm 0{,}017$	$1{,}227 \pm 0{,}030$	$24{,}184 \pm 1{,}940$
FashionMNIST-Small	GeneratedMNIST-Big	VAE	$0{,}806 \pm 0{,}006$	$0{,}588 \pm 0{,}014$	$12{,}393 \pm 1{,}684$
ImageNet-Small	—	—	$0{,}363 \pm 0{,}017$	$3{,}849 \pm 0{,}866$	$46{,}615 \pm 11{,}498$
ImageNet-Small	ImageNet-Big	—	$\mathbf{0{,}465 \pm 0{,}005}$	$\mathbf{1{,}876 \pm 0{,}066}$	$\mathbf{26{,}488 \pm 0{,}996}$
ImageNet-Small	ImageNet-Big	StyleTransfer	$0{,}411 \pm 0{,}008$	$2{,}131 \pm 0{,}093$	$29{,}476 \pm 1{,}495$
Reg-Train	—	—	—	$1{,}0647 \pm 0{,}0001$	—
Reg-Train	Reg-Train	—	—	$\mathbf{1{,}0645 \pm 0{,}0001}$	—
Reg-Small	—	—	—	$1{,}0804 \pm 0{,}0004$	—
Reg-Small	Reg-Big	—	—	$\mathbf{1{,}0755 \pm 0{,}0004}$	—
Reg-Small	Reg-Big	Sin	—	$1{,}0782 \pm 0{,}0003$	—

В работе рассмотрена проблема понижения сложности модели при ее переносе к новым данным меньшей мощности. Рассмотрены методы дистилляции моделей и доменной адаптации. Был предложен подход для случая, когда модели учителя и ученика заданы на выборках разной мощности с известной связью между выборками.

В ходе экспериментов, проведенных на реальных и синтетических данных, показано что предложенные методы хорошо работают для передачи знаний от большой модели к меньшей дистиллированной модели. Результаты экспериментов представлены в таблице 10.

Из таблицы видно, что качество модели зависит от размера выборки: модель ученика, обученная на всей обучающей выборке, имеет наилучшее качество. Также во всех экспериментах качество модели ученика повышается при использовании ответов учителя. Использование отображения между выборками также влияет на качество дистиллированой модели: точность модели с использованием вариационного автокодировщика почти в два раза больше точности модели без использования автокодировщика.

Список литературы

[1] *Hinton G., Vinyals O., Dean J* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. — 2015.

[2] *D.Lopez-Paz, L.Bottou, B.Schölkopf, V.Vapnik* Unifying distillation and privileged information // ICLR. — 2016.

[3] *Yoon Kim, Alexander M.Rush* Sequence-Level Knowledge Distillation. — 2016.

[4] *H.Kim, M. Lee, H.Lee, T.Kang, J.Lee, E.Yang, S.Hwang* Multi-domain Knowledge Distillation via Uncertainty-Matching for End-to-End ASR Models. — 2021.

[5] *Mei Wang, Weihong Deng* Deep Visual Domain Adaptation: A Survey. — 2018.

[6] *Xiao H., Rasul K., Vollgraf R.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017. https://arxiv.org/abs/1708.07747.

[7] *LeCun Y., Cortes C.* MNIST handwritten digit database. — 2010. http://yann.lecun.com/exdb/mnist/

[8] *Diederik P.Kingma, M. Welling* Auto-Encoding Variational Bayes. — 2014. https://arxiv.org/pdf/1312.6114.pdf

[9] *Y. Pang, J. Lin, T. Qin* Image-to-Image Translation: Methods and Applications. — 2021.

[10] *S. Sankaranarayanan, Y. Balaji, A. Jain* Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation. — 2018.

[11] *Kingma D., Ba J.* Adam: A Method for Stochastic Optimization // ICLR. — 2015.

[12] *Hongruixuan Chen, Chen Wu, Yonghao Xu, Bo Du* Unsupervised Domain Adaptation for Semantic Segmentation via Low-level Edge Information Transfer. — 2021.

[13] *Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai* Domain Adaptation via Prompt Learning. — 2022.

[14] *Zhiyuan Wu, Yu Jiang, Minghao Zhao, Chupeng Cui* Spirit Distillation: A Model Compression Method with Multi-domain Knowledge Transfer

[15] *Y.Ganin, V.Lempitsky* Unsupervised Domain Adaptation by Backpropagation. — 2015.

[16] *Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu* Multi-source Distilling Domain Adaptation. — 2020.

[17] *Brady Zhou, Nimit Kalra, Philipp Krahenbuhl* Domain Adaptation Through Task Distillation. — 2020.

[18] *Guobin Chen, Wongun Choi, Xiang Yu* Learning Efficient Object Detection Models with Knowledge Distillation. — 2017.

[19] *Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li* Improved Knowledge Distillation via Teacher Assistant. — 2020.

[20] *Yifan Liu, Ke Chen, Chris Liu* Structured Knowledge Distillation for Semantic Segmentation. — 2018.

[21] *T. Asami, R. Masumura, Y. Yamaguchi* Domain adaptation of DNN acoustic models using knowledge distillation. — 2017.

[22] *Srikanth Tammina* Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. — 2019.

[23] *Antonia Creswell, Tom White, Vincent Dumoulin* Generative Adversarial Networks: An Overview. — 2017.

[24] *Грабовой А.В., Стрижов В.В.* Вероятностная интерпретация задачи дистилляции // Автоматика и телемеханика, 2022.

[25] *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика, 2021.

[26] *Код эксперимента*
https://github.com/kbayazitov/distillation/blob/main/code/main.ipynb