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 **4**
1.1 Обзор предметной области 4
1.2 Предложенный метод 6

2 Постановка задачи **7**
2.1 Базовая постановка задачи дистилляции 7
2.2 Постановка задачи дистилляции для многодоменной выборки 8

3 Вычислительный эксперимент **10**
3.1 Анализ базовой дистилляции 11
3.2 Вариационный автокодировщик 16
3.3 Анализ качества модели, предложенной на основе вариационного автокодировщика 19
3.4 Анализ качества модели на расширенной синтетически сгенерированной выборке 20

4 Заключение **22**

1 Введение

Актуальность темы. Традиционно алгоритмы машинного обучения разрабатываются для решения конкретной задачи. Поэтому, в зависимости от новой постановки задачи, модель необходимо перестраивать с нуля. 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. В этом случае можно использовать перенос знаний с большой модели на модель 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, более интерпретируемой. Цель данной работы заключается в понижении сложности модели машинного обучения при переходе к данным меньшей мощности. Для этого предлагается использовать два основных метода — дистилляция моделей и доменная адаптация.

Новизна. Предложен подход 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

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. Дистиляция 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 устранения дисбаланса классов. 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.

Выборки могут состоять из объектов, которые можно разделить на домены [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. Для задачи дистилляции, предложенной Джеффри Хинтоном [1], исходный и целевой наборы данных совпадают.

Различные постановки задач доменной адаптации описываются в [5], встречаются постановки с частично размеченным целевым доменом и неразмеченным вовсе. 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.

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 также и связь между доменами. 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 с помощью вариацинного автокодировщика [8]. При этом исследуются отображения, для которых существования обратных не рассматриваются. Ожидается, что качество полученных моделей будет превышать качество моделей, в обучении которых не использовались метки учителя.

В качестве экспериментальных данных используются реальные данные и синтетическая выборка. В качестве реальных данных рассматривается выборка FashionMnist [6], состоящая из изображений одежды, для которой требуется решить задачу классификации на 10 типов одежды.

2 Постановка задачи

2.1 Базовая постановка задачи дистилляции

Задана выборка

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

где множество $\mathbb{Y} = \{1, ..., R\}$ для задачи классификации, где R — число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регресии.

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:

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

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:

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

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

где $\mathfrak{D}_\text{s}, \mathfrak{D}_\text{t}$ — исходный и целевой наборы данных. Для задачи дистилляции, предложенной Джеффри Хинтоном, $\mathfrak{D}_\text{s} = \mathfrak{D}_\text{t}$.

Предполагается, что число объектов в выборках не совпадают:

$$|\mathbb{X}_\text{s}| \gg |\mathbb{X}_\text{t}|$$

Пусть при этом задана модель учителя на выборке большей мощности:

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

где $\mathbf{f}$ — модель учителя.

Задана связь между исходной и целевой выборками:

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

где φ — отображение, для которого существование обратного не рассматривается.

Требуется получить модель ученика для малоресурсной выборки:

$$\mathbf{g} : \mathbb{X}_\text{t} \to \mathbb{Y},$$

где $\mathbf{g}$ — модель ученика.

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

1. для задачи регрессии:

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

2. для задачи классификации:

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

где λ — метапараметр, задающий вес дистилляции.

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 путем дистилляции модели учителя в модель ученика, проводится вычислительный эксперимент для задачи классификации.

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 — ReLu. Для решения оптимизационной задачи используется градиентный метод оптимизации Adam [11].

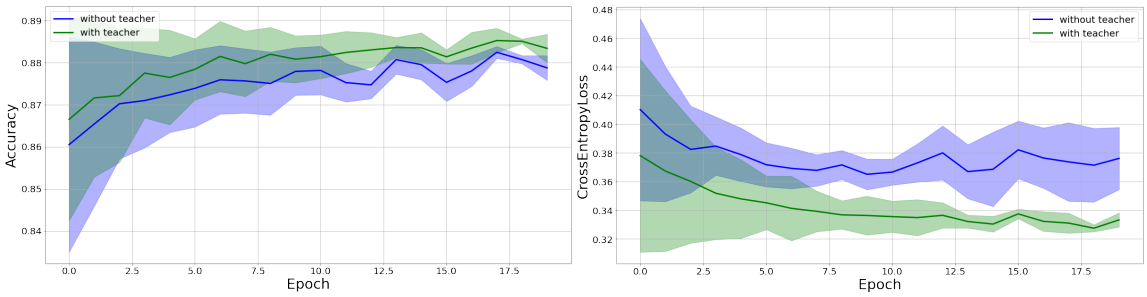
Выборки разделяются на 4 части: обучающая, многоресурсная, малоресурсная, а также тестовая часть. Обучающая часть содержит 60000 объектов, многоресурсная часть содержит 59000 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 10000 объектов.

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

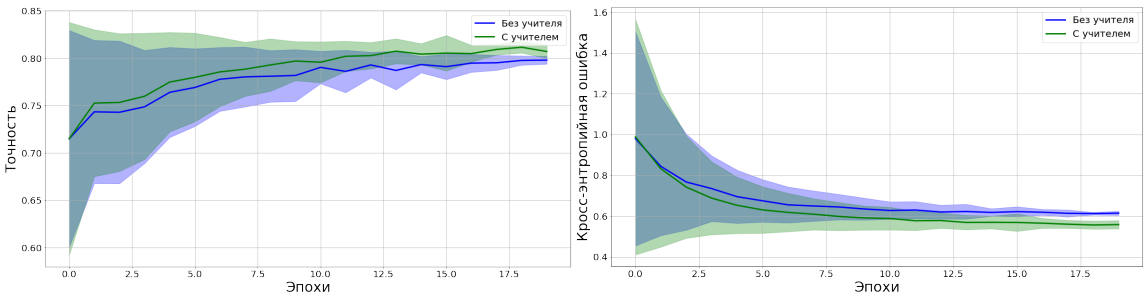
На рис.1а показан график зависимости метрики точности на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.1б показан график зависимости кросс-энтропийной ошибки на тестовой выборке между истинными метками объектов и ве-

роятностями, предсказанными моделью ученика.

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

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 тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.2б показан график зависимости кросс-энтропийной ошибки на 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.

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.

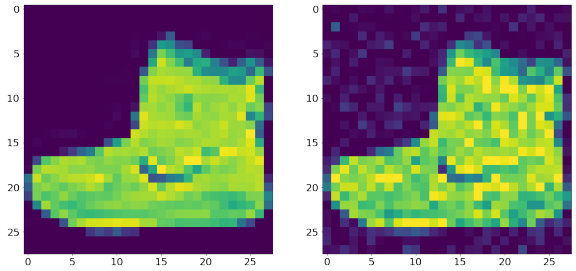
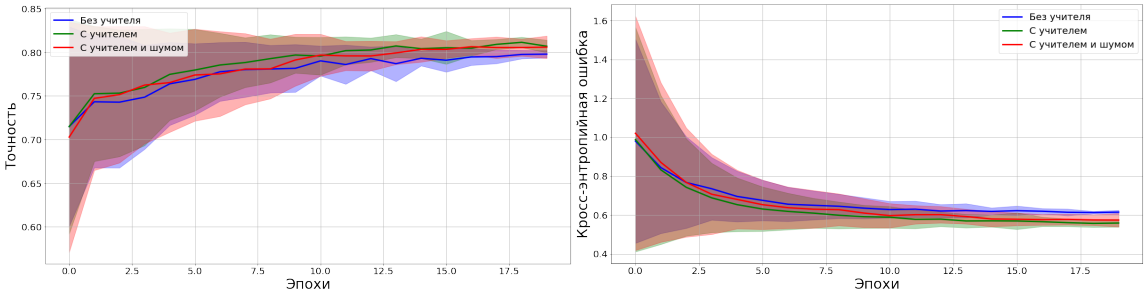


Рис. 3: Сравнение объекта выборки до и после добавления шума



(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 тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.4б показан график зависимости кросс-энтропийной ошибки на 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.

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.

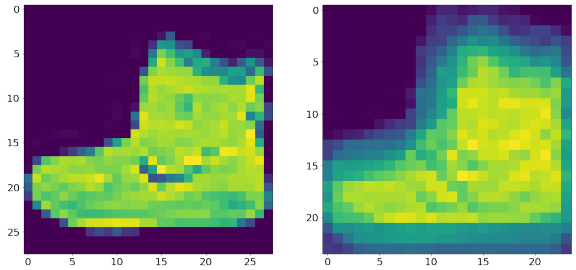
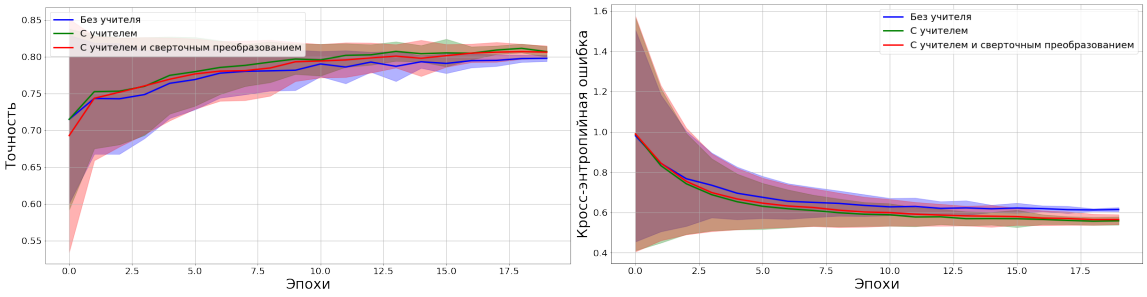


Рис. 5: Сравнение объекта выборки до и после преобразования



(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.6a показан график зависимости метрики точности на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.6б показан график зависимости кросс-энтропийной ошибки на 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.

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 вероятностоное 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. $q_\alpha(\mathbf{z}|\mathbf{x})$ — вероятностный кодировщик;
2. $p_\beta(\hat{\mathbf{x}}|\mathbf{z})$ — вероятностный декодировщик;
3. функция потерь:

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

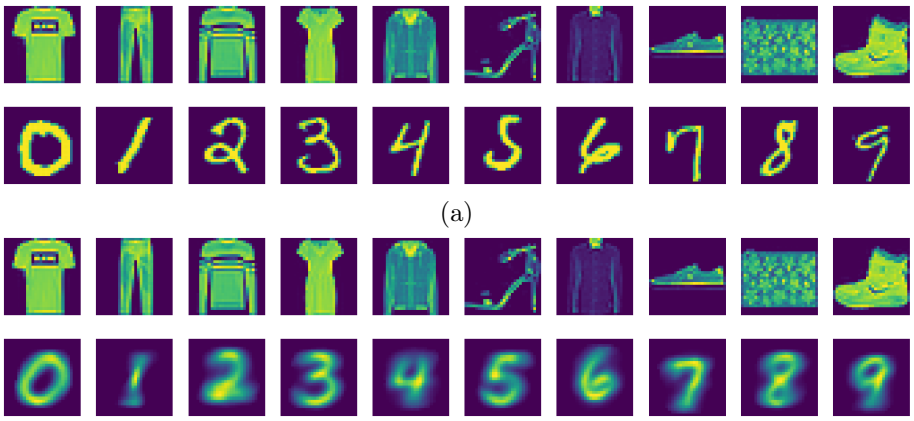
где $p(\mathbf{z}) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ — априорное распределение.

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].

Создадим 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 [8], 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 [8], 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.

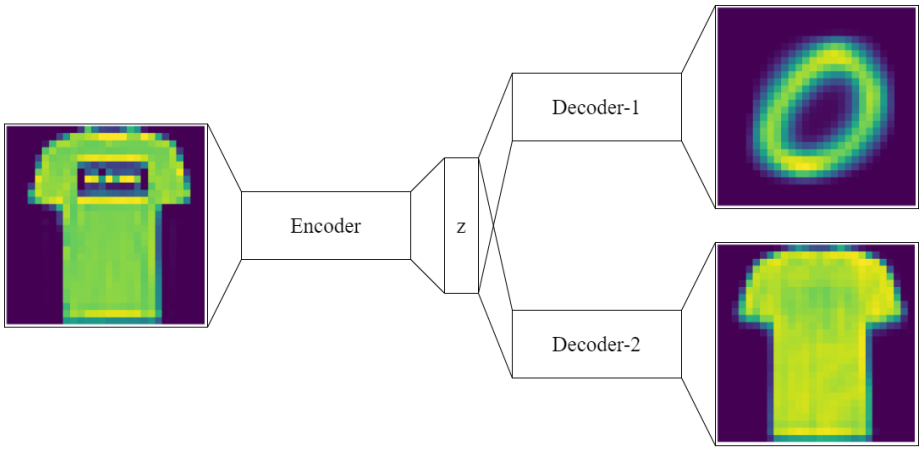
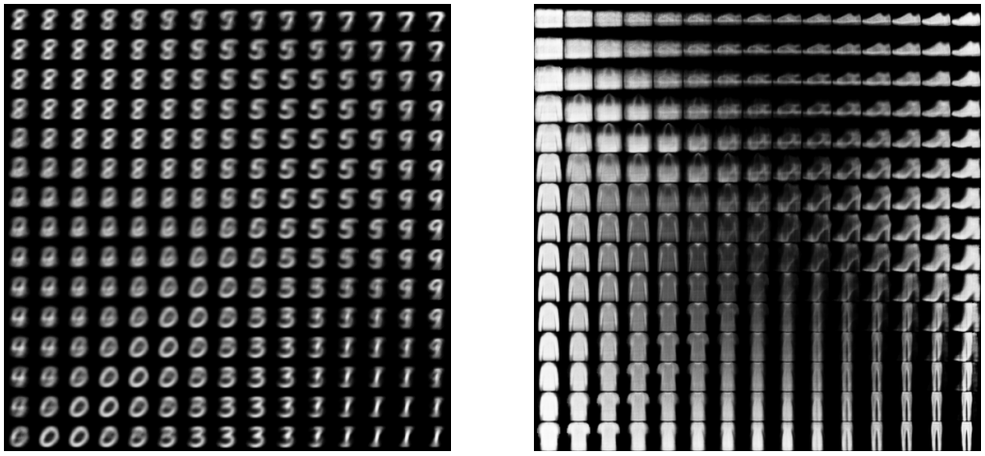


Рис. 8: Пример работы вариационного автокодировщика

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



(a) (b)

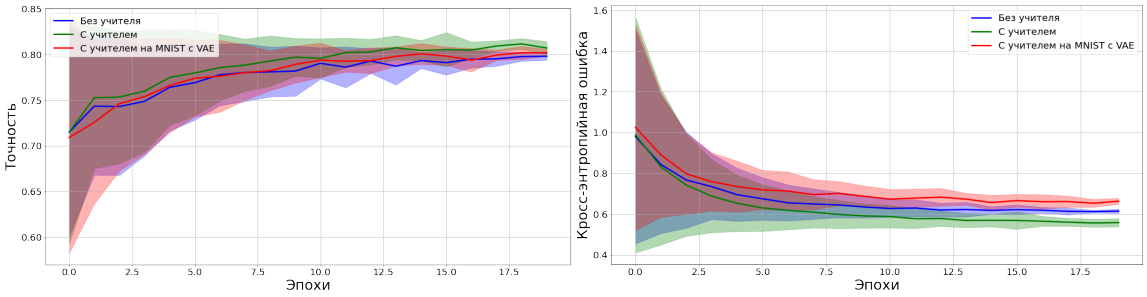
Рис. 9: Зависимость выхода модели от изменения вектора в скрытом представлении для генерации a) цифр; b) одежды

Видно, что классы соответствуют друг другу, как представлено на рис.7a. То есть рис.9 не противоречит рис.7.

3.3 Анализ качества модели, предложенной на основе вариационного автокодировщика

Обучается модель учителя на выборке MNIST-Big, а модель ученика на выборке FashionMNIST-Small. При этом при обучении модели ученика будем использовать метки учителя, подавая ему на вход выход вариационного автокодировщика [8], переводящего изображения одежды в изображения цифр.

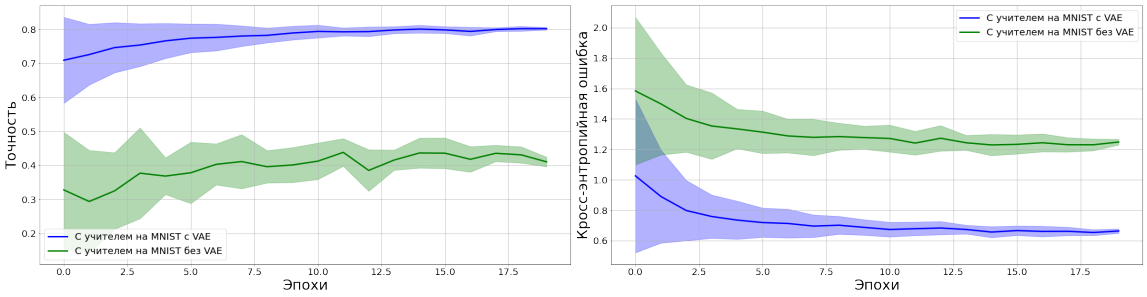
Сравнивается качество аппроксимации без использования вариационного автокодировщика [8]: модель ученика обучается на выборке FashionMNIST-Small, модель учителя обучается на выборке MNIST-Big и используется при обучении ученика, получая на вход изображения одежды без преобразования вариационным автокодировщиом.



(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

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



(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

3.4 Анализ качества модели на расширенной синтетически сгенерированной выборке

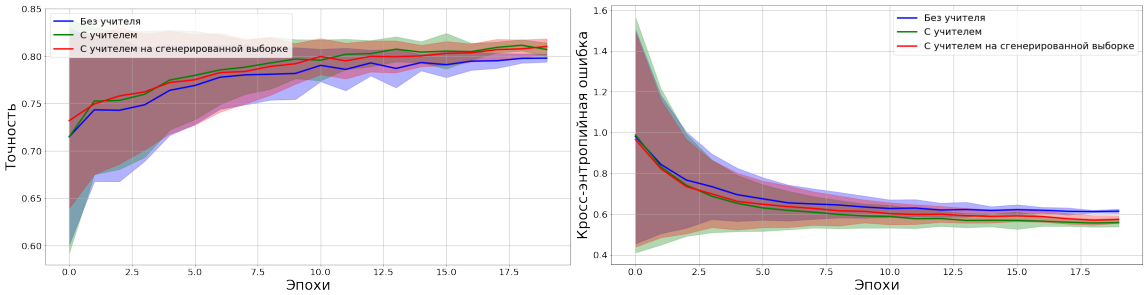
На основе малоресурсной части выборки FashionMNIST-Small сформируем новую выборку, сгенерировав для каждого объекта одежды 70 изображений цифр с помощью модели вариационного автокодировщика [8]. Далее разделим полученную выборку на 4 части: обучающая, многоресурсная, малоресурсная, а также тестовая часть. Обучающая часть содержит 60000 объектов, многоресурсная часть содержит 59000 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 10000 объектов.

Таблица 3: 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.



(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

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.

4 Заключение

Таблица 4: Результаты экспериментов

Ученик	Учитель	Связь φ	Точность	Кросс-энтропия
FashionMNIST-Train	—	—	0,879	0,376
FashionMNIST-Train	FashionMNIST	—	0,884	0.332
FashionMNIST-Small	—	—	0,798 ±0,004	0,615 ±0,010
FashionMNIST-Small	FashionMNIST-Big	—	0,807 ±0,007	0,559 ±0,020
FashionMNIST-Small	FashionMNIST-Big	Noise	0,806 ±0,013	0,574 ±0,035
FashionMNIST-Small	FashionMNIST-Big	Dilation	0,806 ±0,009	0,565 ±0,025
FashionMNIST-Small	MNIST-Big	VAE	0,802 ±0,004	0,664 ±0,015
FashionMNIST-Small	MNIST-Big	—	0,410 ±0,014	1,248 ±0,019
FashionMNIST-Small	GeneratedMNIST-Big	VAE	0,810 ±0,008	0,574 ±0,019

В работе исследована 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 4.

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] *Code on Github*
https://github.com/kbayazitov/distillation/blob/main/code/main.ipynb