

## Аннотация

Исследуется проблема понижения сложности аппроксимирующей модели при переносе модели к новым данным меньшей мощности. Вводятся понятия учителя, ученика для сильного и слабого доменов соответственно. Предполагается, что признаковые описания моделей ученика и учителя принадлежат разным доменам. При этом мощность одного домена 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

<b>1</b>	<b>Введение</b>	<b>4</b>
1.1	Обзор предметной области . . . . .	4
1.2	Предложенный метод . . . . .	5
<b>2</b>	<b>Постановка задачи</b>	<b>6</b>
2.1	Базовая постановка задачи дистилляции . . . . .	6
2.2	Постановка задачи дистилляции для многодоменной выборки . . . . .	7
<b>3</b>	<b>Вычислительный эксперимент</b>	<b>9</b>
3.1	Анализ базовой дистилляции . . . . .	9
3.2	Вариационный автокодировщик . . . . .	13
3.3	Анализ качества модели, предложенной на основе ва- риационного автокодировщика . . . . .	15
3.4	Анализ качества модели на расширенной синтетически сгенерированной выборке . . . . .	17
<b>4</b>	<b>Заключение</b>	<b>18</b>

# 1 Введение

**Актуальность темы.** Одним из самых простых способов повышения качества алгоритма машинного обучения является использование ансамбля моделей. Однако 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.

**Цель работы.** Вместо ансамбля моделей можно использовать модель с большим числом параметров, ответы которой можно использовать при обучении модели с меньшим числом параметров, более подходящей для развертывания. Цель данной работы заключается в понижении сложности модели машинного обучения при переходе к домену 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. При этом задана связь между выборками.

## 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 Дж.Хинтоном, с учетом меток учителя при помощи функции softmax с параметром температуры, а в [2] рассматривается объединение методов дистилляции, предложенной Дж.Хинтоном, и привилегированной информации [2], предложенной В.Валником, в обобщенную дистилляцию. Дистилляция моделей используется в широком классе задач. В [4] рассматривается метод дистилляции моделей для задачи распознавания речи.

Часто выборки могут состоять из объектов, которые можно разде-

лить на домены [9]. К примеру, можно составить отображение из множества реальных фотографий малой мощности во множество сгенерированных движком изображений, мощность которого естественно больше [10]. Одним из самых простых 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 Джефффри Хинтоном, исходный и целевой домены равны. 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.

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 с возможностью выбора веса дистилляции. Таким образом в качестве доменов разной мощности могут служить настоящие и сгенерированные изображения.

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

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

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

Предполагается, что есть обученная модель с большим числом параметров — модель учителя. В качестве модели учителя  $\mathbf{f}$  рассматривается функция из множества:

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

Требуется обучить модель ученика с меньшим числом параметров при использовании ответов учителя. В качестве модели ученика  $\mathbf{g}$  рассматривается функция из множества:

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

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

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

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

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

где  $\cdot|_{T=t}$  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 много-доменной выборки

Заданы две выборки:

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

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

— исходный и целевой наборы данных. Для задачи дистилляции, предложенной Дж.Хинтоном,  $\mathfrak{D}_s = \mathfrak{D}_t$ . Предполагается, что число объектов в выборках не совпадают:

$$|\mathbb{X}_s| \gg |\mathbb{X}_t|$$

$\mathbb{Y}$  — множество целевых переменных.

Пусть при этом задана модель учителя на выборке большей мощности

$$\mathbf{f} : \mathbb{X}_s \rightarrow \mathbb{Y}, \text{ где } \mathbf{f} \text{ — модель учителя}$$

Задана связь между исходной и целевой выборками:

$$\varphi : \mathbb{X}_t \rightarrow \mathbb{X}_s, \text{ где } \varphi \text{ — необратимое отображение}$$

Требуется получить модель ученика для малоресурсной выборки

$$\mathbf{g} : \mathbb{X}_t \rightarrow \mathbb{Y}, \text{ где } \mathbf{g} \text{ — модель ученика}$$

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:

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

где  $\lambda$  — метопараметр, 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 выборки FashionMNIST [6] - набора изображений предметов одежды. В качестве моделей учителя  $\mathbf{f}$  и ученика  $\mathbf{g}$  рассматриваются четырёхслойная и однослойная нейронные сети соответственно. Функция активации — ReLu. Для решения оптимизационной задачи используется градиентный метод Adam [11]. Выборка разделяется на 3 части: две для обучения многоресурсного и малоресурсного доменов, а также тестовая часть выборки. Многоресурсная часть содержит 59000 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 10000 объектов.

#### 3.1 Анализ базовой дистилляции

**Обучение на обоих доменах.** Модели учителя и ученика обучаются на обоих доменах.

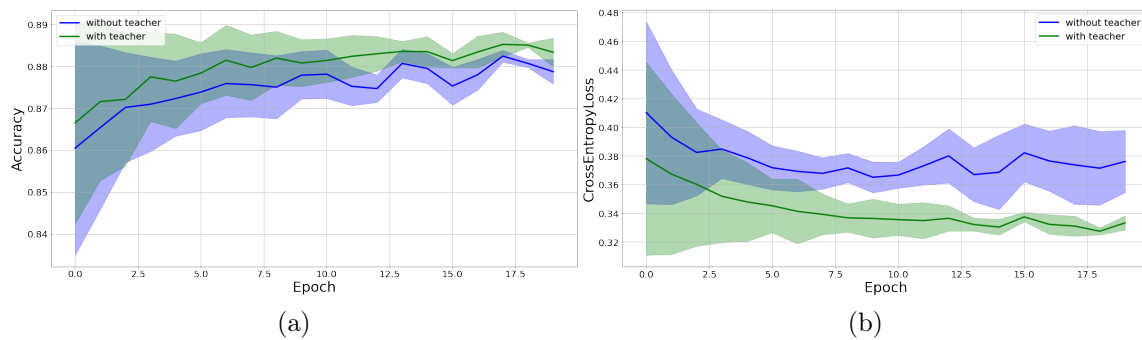


Рис. 1: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. а) accuracy; б) CrossEntropyLoss между истинными и предсказанными учеником метками

На рис.1а показан график зависимости метрики accuracy на тестовой выборке между истинными метками объектов и вероятностями,



предсказанными моделью ученика.

На рис.1б показан график зависимости кросс-энтропии на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

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

**Обучение на малоресурсном домене.** Модель учителя обучается на многоресурсном домене, а модель ученика обучается на малоресурсном домене.

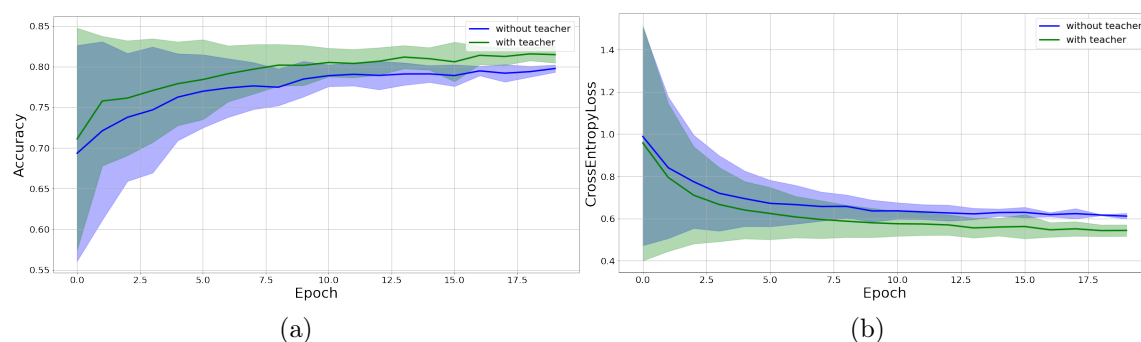


Рис. 2: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. а) accuracy; б) CrossEntropyLoss между истинными и предсказанными учеником метками

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

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

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

**Обучение на выборке с шумом.** Добавим к многоресурсному домену нормальный шум  $\mathcal{N}(0, \frac{1}{10})$  и обучим на нем модель учителя. Модель ученика обучается на малоресурсном домене без шума.

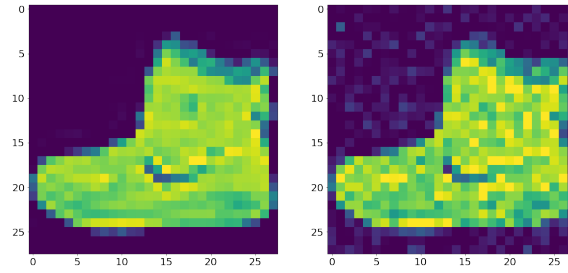


Рис. 3: Сравнение объекта выборки до и после добавления шума

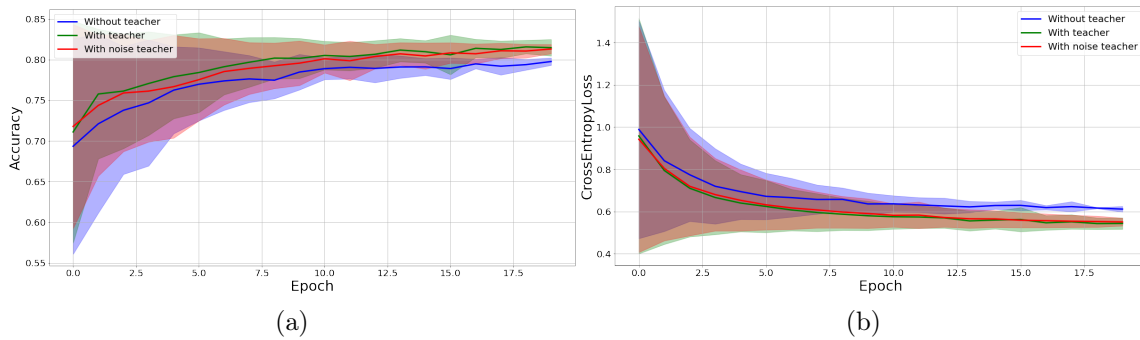


Рис. 4: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. а) accuracy; б) CrossEntropyLoss между истинными и предсказанными учеником метками

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

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

На графиках видно, что значения accuracy и CrossEntropyLoss 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 многоресурсному домену сверточное преобразование с параметром  $\text{dilation} = 2$  и обучим на нем модель учителя. Модель ученика обучается на малоресурсном домене.

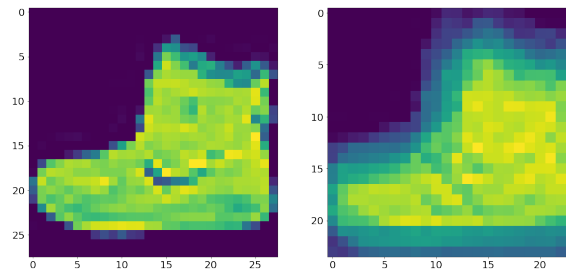


Рис. 5: Сравнение объекта выборки до и после преобразования

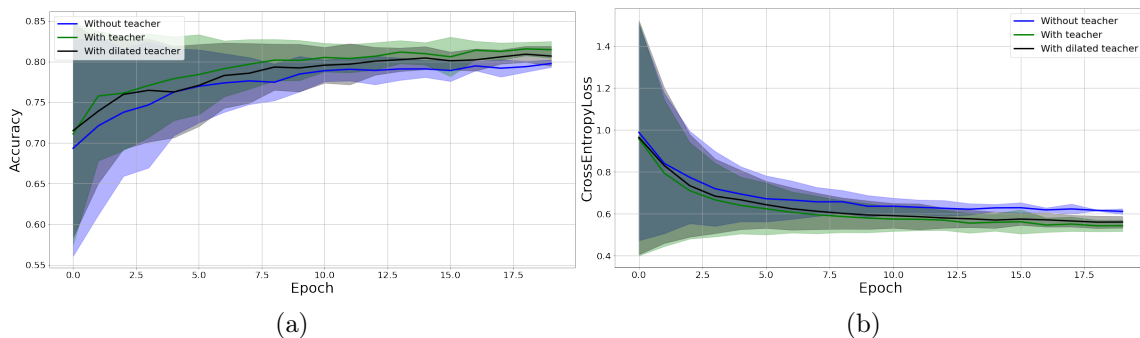


Рис. 6: Качество аппроксимации на тестовой выборке. Все результаты усреднены по 5 запускам. а) accuracy; б) CrossEntropyLoss между истинными и предсказанными учеником метками

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

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

На графиках видно, что значения accuracy и CrossEntropyLoss 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 используем модель вариационного автокодировщика, аппроксимирующую отображение  $\varphi$ .

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

$\mathbf{q}_\alpha(\mathbf{z}|\mathbf{x})$ -вероятностный кодировщик

$\mathbf{p}_\beta(\hat{\mathbf{x}}|\mathbf{z})$ -вероятностный декодировщик

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

$\mathbf{p}(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$  - 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 для преобразования изображений одежды из выборки FashionMNIST [6] в изображения цифр на основе выборки MNIST.

Создадим синтетическую выборку, где каждому изображению одежды будет соответствовать случайное изображение цифры из того же класса.

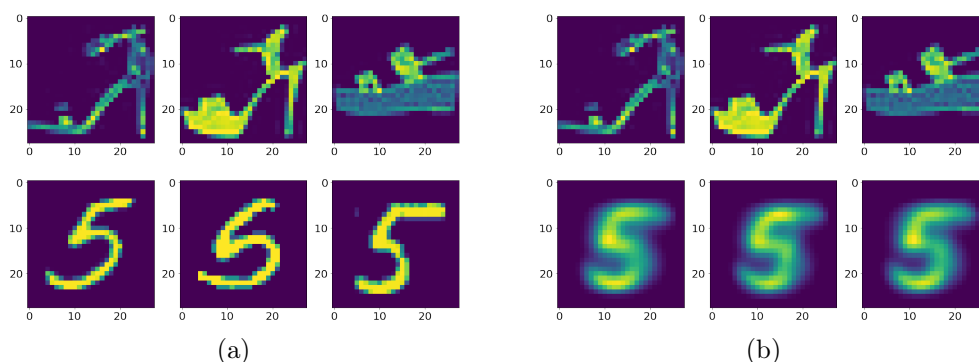


Рис. 7: а) Объекты синтетической выборки; б) Объекты исходной выборки до и после работы автокодировщика

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

На основе данной 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.

Получили модель, способную генерировать семейство новых объектов - изображений цифры или изображений одежды для одного и того же изображения одежды.

Посмотрим также на изменение выхода модели при изменении случайного вектора в скрытом представлении. Для визуализации построим скрытое представление размерности 2:

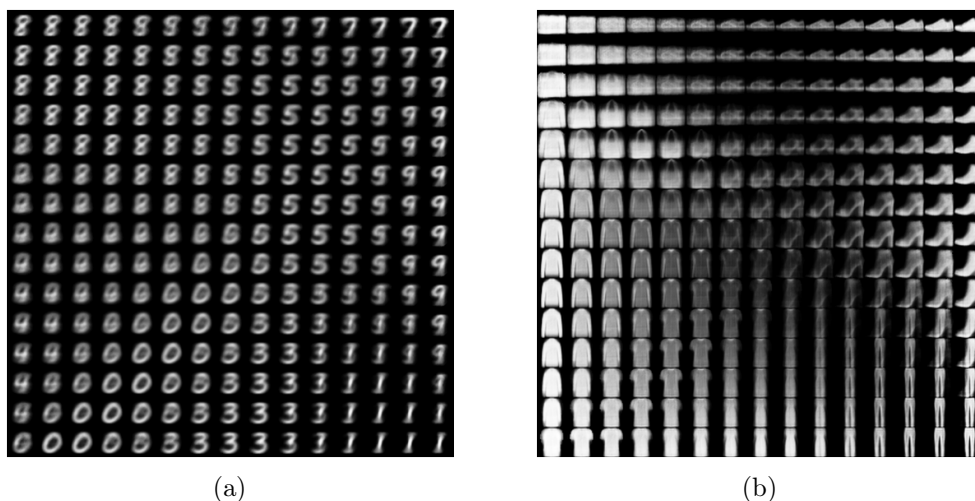


Рис. 8: Зависимость выхода модели от изменения вектора в скрытом представлении для генерации а) цифр; б) одежды

### 3.3 Анализ качества модели, предложенной на основе вариационного автокодировщика

Будем обучать модель учителя на выборке MNIST, а модель ученика на выборке FashionMNIST. При этом при обучении модели ученика будем использовать метки учителя, подавая ему на вход выход вариационного автокодировщика, переводящего изображения одежды в изображения цифр.

Также для сравнения покажем качество аппроксимации без использования вариационного автокодировщика.

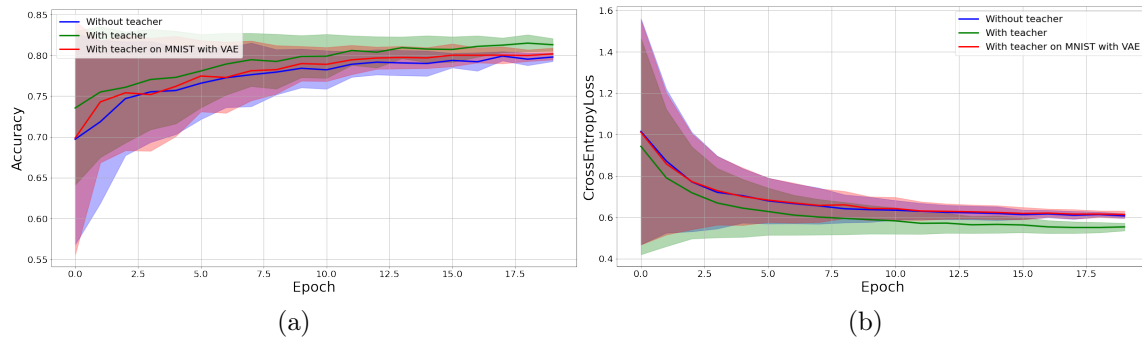


Рис. 9: Качество аппроксимации при использовании VAE на малодоменной выборке. Все результаты усреднены по 5 запускам. а) accuracy; б) CrossEntropyLoss между истинными и предсказанными учеником метками

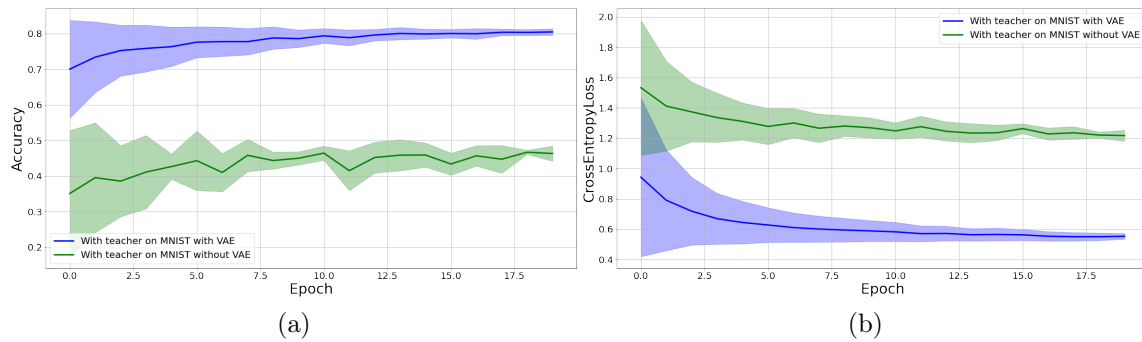


Рис. 10: Сравнение качества аппроксимации в зависимости от использования VAE на малодоменной выборке. Все результаты усреднены по 5 запускам. а) accuracy; б) CrossEntropyLoss между истинными и предсказанными учеником метками

На графике видно, что без использования отображения  $\varphi$  модель становится более шумной с явным снижением точности.

### **3.4 Анализ качества модели на расширенной синтетически сгенерированной выборке**

Для каждого объекта малодоменной выборки сгенерируем 70 изображений цифр с помощью модели вариационного автокодировщика и создадим новую выборку из 70000 объектов.

Модель ученика обучается на малодоменной выборке, модель учителя на новой выборке и используется при обучении ученика.



## 4 Заключение

## 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.