Аннотация

Исследуется проблема понижения сложности аппроксимирующей модели при переходе к данным домена меньшей мощности. Вводятся понятия учителя, ученика, слабого и сильного доменов. Признаковые описания моделей ученика и учителя принадлежат разным доменам. 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.

Ключевые слова: адаптация доменов, дистилляция, байесовский выбор модели, байесовская дистилляция

Содержание

1 Введение

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 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.

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 дистилляци, предложенной Дж.Хинтоном, и привилегированной информации, предложенной В.Вапником, в обобщенную дистилляцию. Дистиляцция моделей используется в широком классе задач. В [4] рассматривается метод дистилляции моделей для задачи распознавания речи.

Часто выборки могут состоять из объектов, которые можно разделить на домены. К примеру, можно составить отображение из множества реальных фотографий малой мощности во множество сгенерированных движком изображений, мощность которого естественно больше. Одним из самых простых примеров генерации новых изображений является работа модели вариационного автокодировщика, способного для одного и того же объекта строить вероятностное распределение, на основе которого можно получить целое семейство новых объектов. Для задачи дистилляции, предложенной Дж.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 выбор-

ках является задача машинного перевода текстов, описанная в [3]. 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, предложенной Дж.Хинтоном

Задана выборка $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$, где $\mathbf{X} \in \mathbb{X}, \mathbf{Y} \in \mathbb{Y}$. Множество $\mathbb{Y} = \{1, ..., R\}$ для задачи классификации, где R - число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регресии.
В качестве модели ученика $\mathbf{g}$ рассматривается функция из множества:

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

В качестве модели учителя $\mathbf{f}$ рассматривается функция из множества:

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

$\mathbf{v}$, $\mathbf{z}$ - 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:

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

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 многодоменной выборки

Заданы два домена:

$$\mathbb{D}_{\text{s}}, \mathbb{D}_{\text{t}}$$

- исходный и целевой наборы данных. Для задачи дистилляции, предложенной Дж.Хинтоном, $\mathbb{D}_{\text{s}} = \mathbb{D}_{\text{t}}$. Предполагается, что числа объектов в доменах не совпадают:

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

$\mathbb{Y}$ - множество целевых переменных.
Пусть при этом задана модель учителя

$$\mathbf{f} : \mathbb{X}_{\text{s}} \to \mathbb{Y}, \text{ где } \mathbf{f} - \text{модель учителя}$$

и связь между исходным и целевым доменами:

$$\varphi : \mathbb{X}_{\text{t}} \to \mathbb{X}_{\text{s}}, \text{ где } \varphi - \text{необратимое отображение}$$

Требуется получить отображение

$$\mathbf{g} : \mathbb{X}_{\text{t}} \to \mathbb{Y}, \text{ где } \mathbf{g} - \text{модель ученика}$$

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:

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

Получаем оптимизационную задачу:

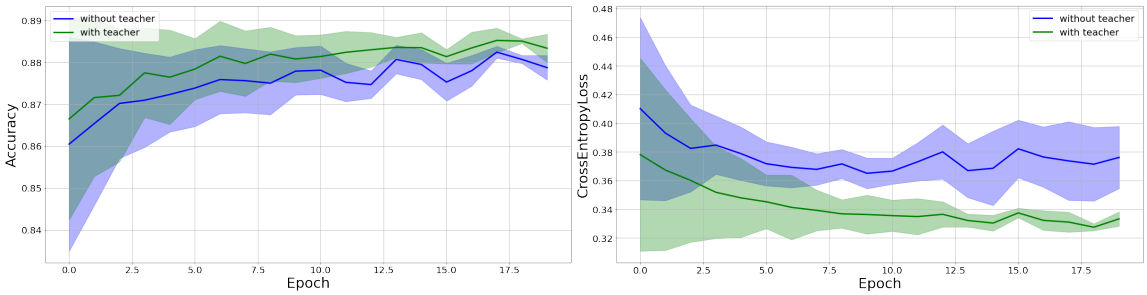
$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{f}, \varphi).$$

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 соответсвенно. Для решения оптимизационной задачи используется Adam, функция активации - ReLu.
Выборка разделяется на 3 части: две для обучения многоресурного 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 дистилляции Хинтона

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



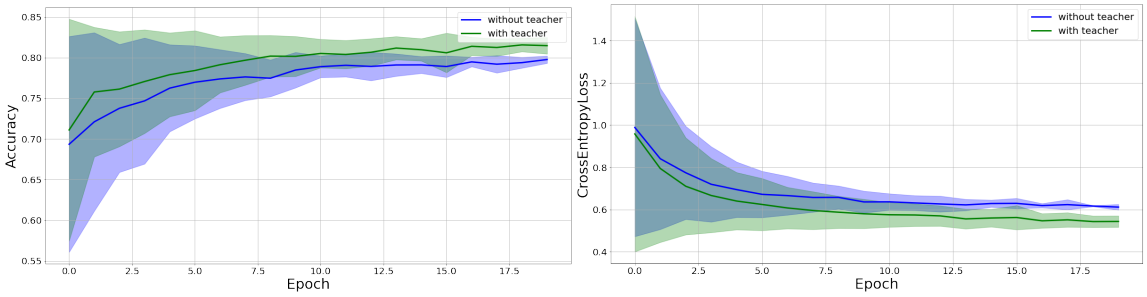
(a) (b)

Рис. 1: Качество аппроксимации на тестовой выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

На рис.1a 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 accuracy, при этом наблюдается значительное снижение ошибки.

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



(a) (b)

Рис. 2: Качество аппроксимации на тестовой выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

На рис.2а показан график зависимости метрики accuracy 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 accuracy, при этом наблюдается снижение ошибки.

Обучение на выборке с шумом. Добавим к многоресурсному домену нормальный шум $\mathcal{N}(0, 0.1)$ и обучим на нем модель учителя. Модель ученика обучается на малоресурсном домене.

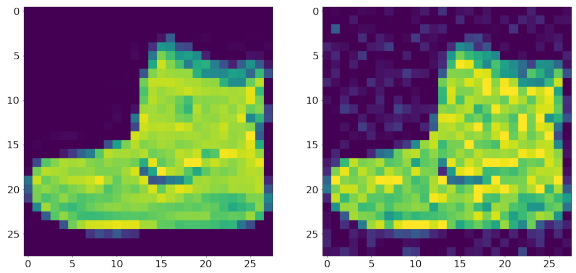
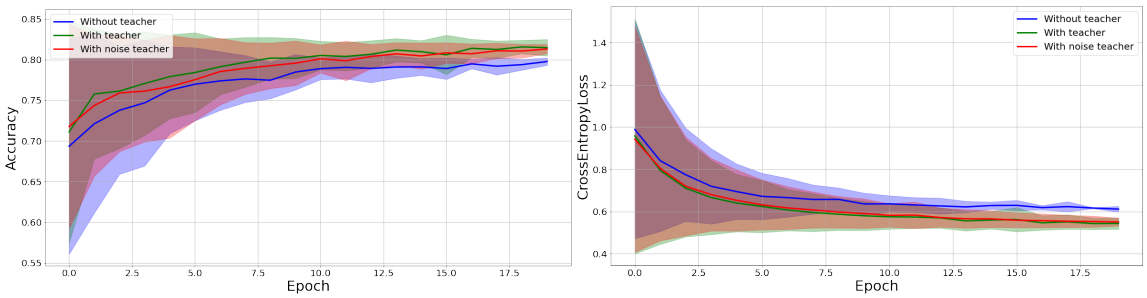


Рис. 3: Сравнение объекта выборки до и после добавления шума



(a) (b)

Рис. 4: Качество аппроксимации на тестовой выборке a) accuracy; b) 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.

Обучение на выборке с dilation. Применим к многоресурсному домену сверточное преобразование с параметром dilation $= 2$ и обучим на нем модель учителя. Модель ученика обучается на малоресурсном домене.

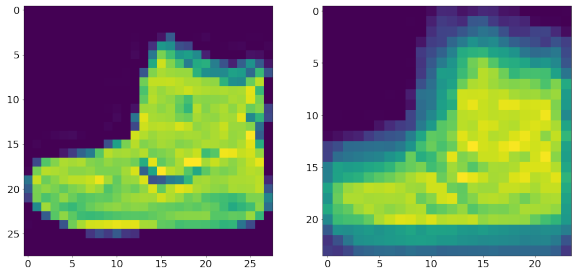
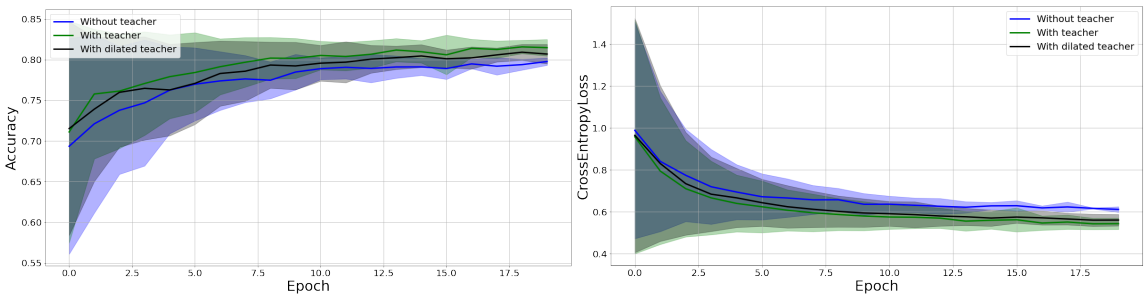


Рис. 5: Сравнение объекта выборки до и после преобразования



(a) (b)

Рис. 6: Качество аппроксимации на тестовой выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

На рис.6a 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 выборки FashionMNIST [6] будем использовать модель вариационного автокодировщика. Данная модель состоит из двух частей. Сначала строится вероятностоное распределение в скрытом пространстве, которое позволяет генерировать кодовые представления для одного объекта. Далее с помощью декодировщика строится вероятностное распределение, позволяющее генерировать реконструкци исходного объекта.

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

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

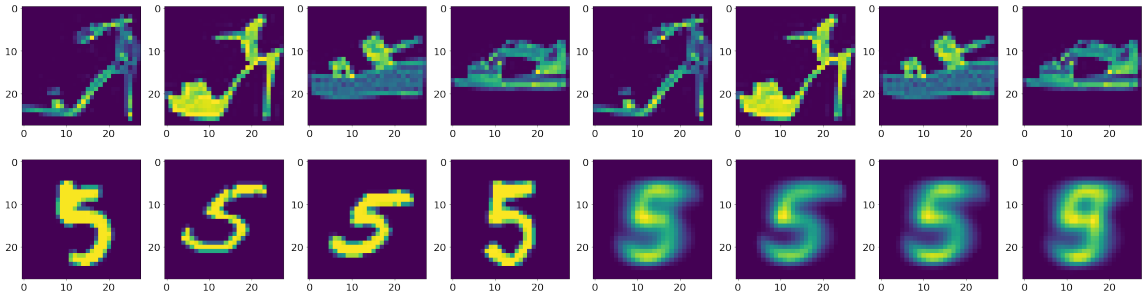
$\mathbf{p}(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \sigma^\mathbf{2}\mathbf{I})$ - априорное распределение

Получаем оптимизационную задачу:

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

3.3 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.



(a) (b)

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

Далее, на основе данной выборки обучим модель вариационного автокодировщика, минимизируя ошибку между выходом модели и целевым значением - изображением цифры, соответствующего исходному объекту.
Получили модель, генерирующую семейство новых объектов - изображений цифры для одного и того же изображения одежды.

Посмотрим также на изменение выхода модели при изменении случайного вектора в скрытом представлении:

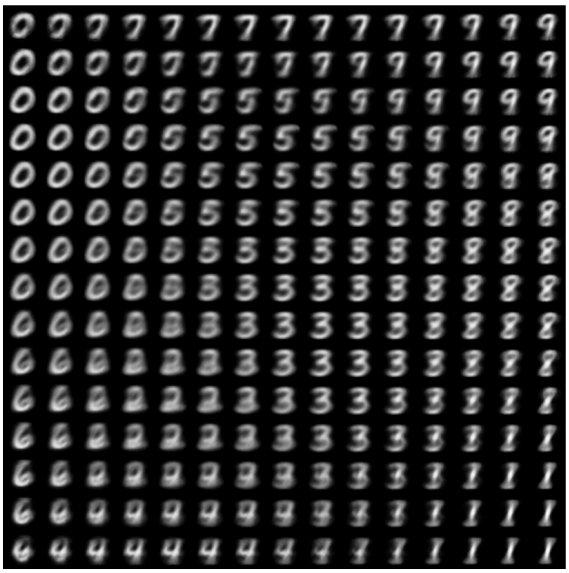
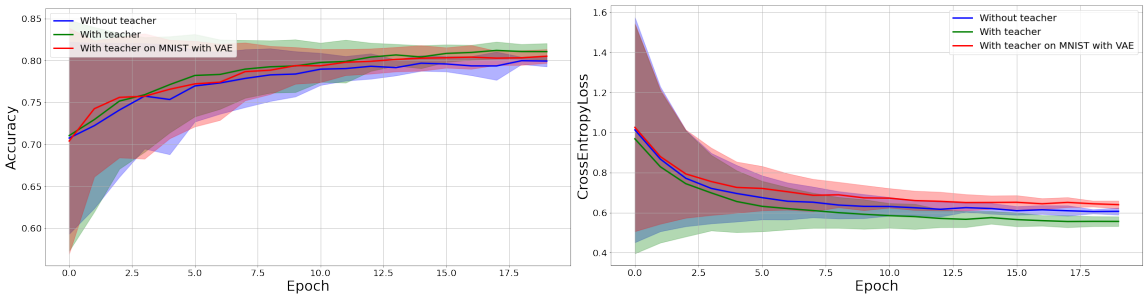


Рис. 8: Зависимость выхода модели от изменения вектора в скрытом представлении

3.4 Качество модели в зависимости от использования автокодировщика

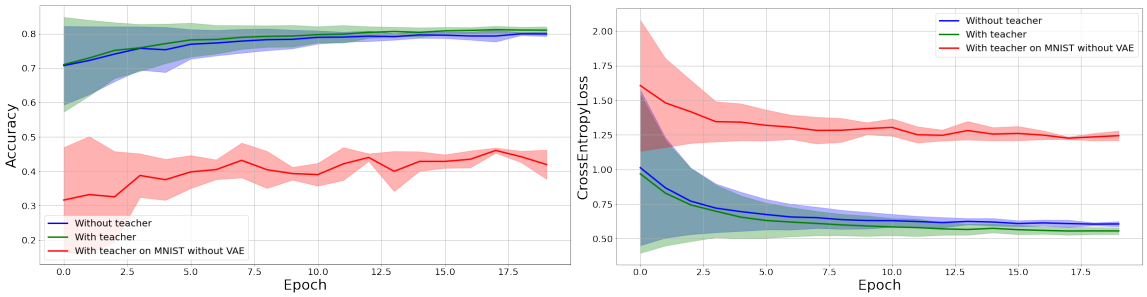
Будем обучать модель учителя на выборке MNIST, а модель ученика на выборке FashionMNIST. При этом при обучении модели ученика будем использовать метки учителя, подавая ему на вход выход вариационного автокодировщика, переводящего изображения одежды в изображения цифр.

Также для сравнения покажем качество аппроксимации без использования вариационного автокодировщика.



(a) (b)

Рис. 9: Качество аппроксимации при использовании VAE на малодоменной выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками



(a) (b)

Рис. 10: Качество аппроксимации без использования VAE на малодоменной выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

3.5 Качество модели на 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.