

Дистилляция моделей на многодоменных выборках

К. М. Баязитов

Выпускная квалификационная работа
03.03.01 — Прикладные математика и физика
Научный руководитель: д.ф.-м.н. В. А. Семенов
Научный консультант: к.ф.-м.н. А. В. Грабовой

25 апреля 2022 г.

Слайд об исследованиях

Исследуется задача построения моделей глубокого обучения на основе предобученных моделей на выборках из близких генеральных совокупностей.

Цель исследования —

Адаптация моделей машинного обучения при переходе к данным из близких генеральных совокупностей.

Предложенный метод —

Дистилляции моделей в случае когда выборки учителя и ученика из разных генеральных совокупностей.

Решение

Предлагается при обучении модели ученика использовать не только метки учителя и истинные метки, а также связь между двумя выборками.

Базовая постановка задачи дистилляции

Заданы

1) выборка:

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

2) параметрическое семейство функций:

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

Выбирается оптимальная модель учителя $\hat{\mathbf{f}} \in \mathfrak{F}$.

Требуется выбрать модель ученика \mathbf{g} из параметрического семейства функций:

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

Оптимизационная задача:

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

где \mathcal{L} — функция ошибки.

Постановка задачи дистилляции для многодоменной выборки

Определение

Генеральная совокупность объектов B называется близкой к совокупности A , если существует инъективное отображение $\varphi : A \rightarrow B$

Заданы

- 1) исходный и целевой наборы данных из близких генеральных совокупностей:

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

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

- 2) модель учителя на выборке большей мощности:

$$\hat{\mathbf{f}} : \mathbb{X}_s \rightarrow \mathbb{Y}', \quad \mathbb{Y}' - \text{пространство оценок}$$

- 3) связь между исходной и целевой выборками:

$$\varphi : \mathbb{X}_t \rightarrow \mathbb{X}_s.$$

Требуется получить модель ученика для малоресурсной выборки:

$$\mathbf{g} : \mathbb{X}_t \rightarrow \mathbb{Y}'.$$

Предложенный метод

Предлагается при обучении модели ученика использовать

1) ответы модели учителя

$$\hat{\mathbf{f}} : \mathbb{X}_s \rightarrow \mathbb{Y}',$$

2) связь между выборками

$$\varphi : \mathbb{X}_t \rightarrow \mathbb{X}_s.$$

Функция ошибки, учитывающая метки учителя и связь между выборками

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

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

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

Экспериментальные данные

Заданы выборки:

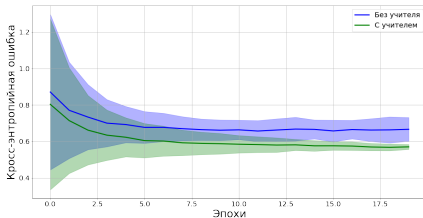
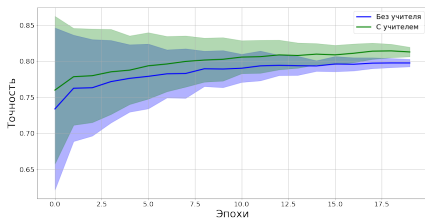
- 1) FashionMNIST — набор изображений предметов одежды,
- 2) MNIST — набор изображений рукописных цифр.

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

Анализ дистилляции на малоресурсной части

- 1) Модель учителя обучается на FashionMNIST-Big,
- 2) Модель ученика обучается на FashionMNIST-Small, тестируется на FashionMNIST-Test.

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



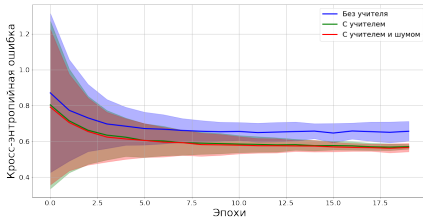
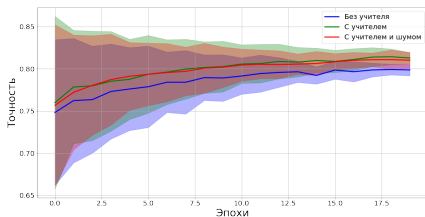
Модель, использующая метки учителя, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Анализ дистилляции с нормальным шумом

- 1) Модель учителя обучается на FashionMNIST-Big с шумом $\mathcal{N}(0, \frac{1}{10})$,
- 2) Модель ученика обучается на FashionMNIST-Small, тестируется на FashionMNIST-Test.

В качестве отображения φ используется нормальный шум $\mathcal{N}(0, \frac{1}{10})$.

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



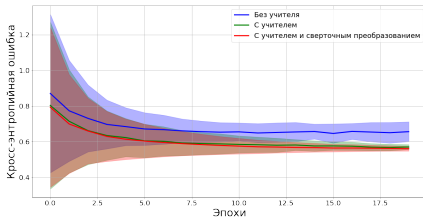
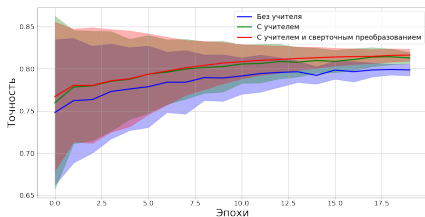
Модель, использующая метки учителя с применением шума, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Анализ дистилляции со сверточным преобразованием

- 1) Модель учителя обучается на FashionMNIST-Big со сверточным преобразованием с параметром $dilation = 2$,
- 2) Модель ученика обучается на FashionMNIST-Small, тестируется на FashionMNIST-Test.

В качестве отображения φ используется сверточное преобразование с $dilation = 2$.

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



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

Вариационный автокодировщик

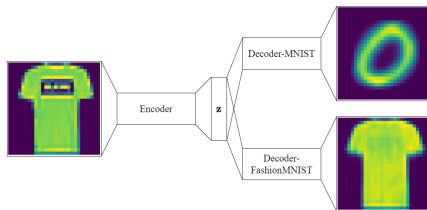
Отображение φ аппроксимируется моделью автокодировщика.

Функция ошибки для обучения автокодировщика:

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

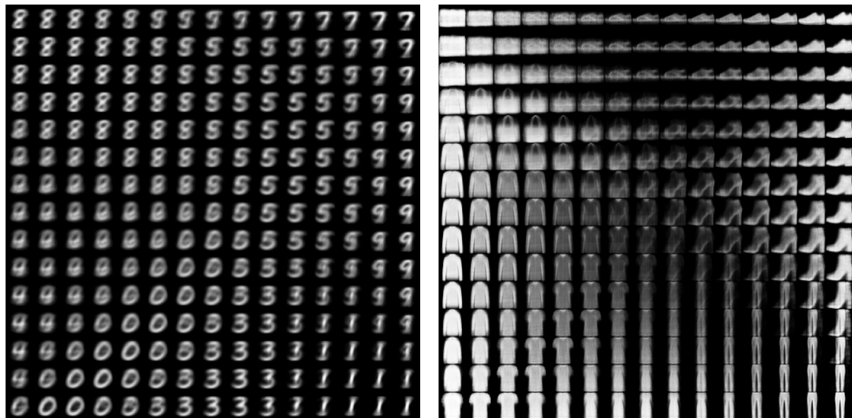
где $p(\mathbf{z}) \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ — априорное распределение, $q(\mathbf{z}|\mathbf{x}, \alpha)$ — вероятностный кодировщик, $p(\hat{\mathbf{x}}|\mathbf{z}, \beta)$ — вероятностный декодировщик.

Вариационный автокодировщик генерирует новые объекты — изображения цифр и одежды для одного изображения одежды.



Визуализация сгенерированных изображений автокодировщиком

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

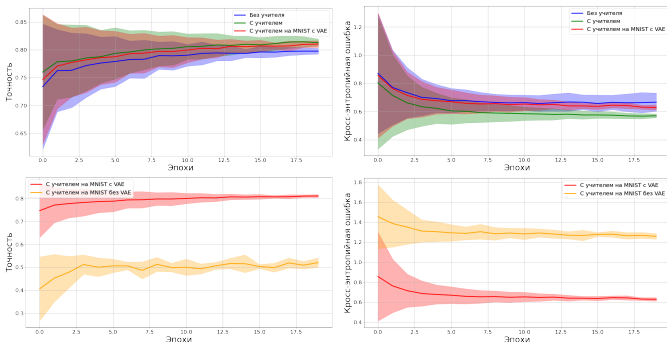


Видно, что классы одежды и цифр соответствуют друг другу.

Анализ дистилляции на основе вариационного автокодировщика

- 1) Модель учителя обучается на MNIST-Big,
- 2) Модель ученика обучается на FashionMNIST-Small, тестируется на FashionMNIST-Test.

На графиках показаны сравнения метрик точности и кросс-энтропийных ошибок модели ученика в зависимости от использования автокодировщика, аппроксимирующего отображение φ .



Без использования отображения φ модель становится более шумной.

Анализ дистилляции на расширенной синтетически сгенерированной выборке

На основе выборки FashionMNIST-Small с помощью модели вариационного автокодировщика генерируется новая выборка GeneratedMNIST объектов цифр.

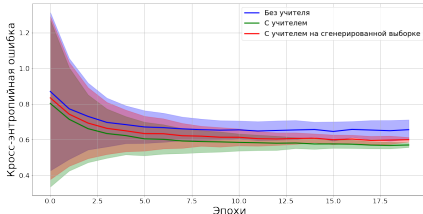
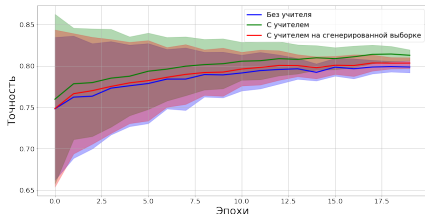
Выборка	Пояснение	Размер выборки
GeneratedMNIST-Train	Обучающая часть	60000
GeneratedMNIST-Big	Многоресурсная часть	59000
GeneratedMNIST-Small	Малоресурсная часть	1000
GeneratedMNIST-Test	Тестовая часть	10000

Анализ дистилляции на расширенной синтетически сгенерированной выборке

- 1) Модель учителя обучается на GeneratedMNIST-Big,
- 2) Модель ученика обучается на FashionMNIST-Small, тестируется на FashionMNIST-Test.

В качестве отображения φ используется выход вариационного автокодировщика, переводящего изображения одежды в изображения цифр.

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



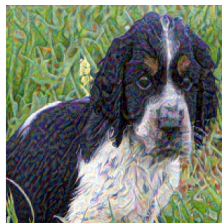
Модель, использующая метки учителя с применением вариационного автокодировщика, показывает лучшее значение точности и кросс-энтропийной ошибки, чем модель без учителя.

Анализ дистилляции на основе преобразования стиля изображений

Используем подвыборку ImageNet для задачи классификации на 10 классов.

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

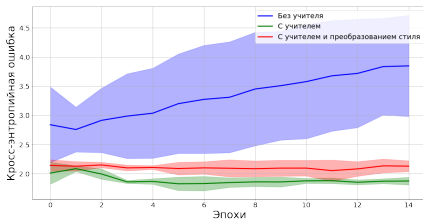
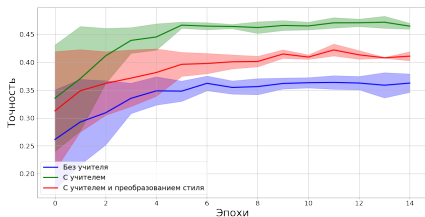
Применим к многоресурсной части выборки преобразование стиля на основе сверточной нейронной сети VGG-19:



Анализ дистилляции на основе преобразования стиля изображений

- 1) Модель учителя обучается на ImageNet-Big с преобразованием стиля,
- 2) Модель ученика обучается на ImageNet-Small, тестируется на ImageNet-Test.

В качестве отображения φ используется преобразование стиля на основе VGG-19. На графиках показаны метрики точности и кросс-энтропийной ошибки модели ученика.



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

Выводы

1. Рассмотрена задача снижения сложности модели при ее переносе к новым данным меньшей мощности из близкой генеральной совокупности.
2. Рассмотрен подход на основе дистилляции моделей глубокого обучения.
3. Предложен подход для случая, когда модели учителя и ученика заданы на выборках разной мощности с известным отображением между выборками.
4. Проведен вычислительный эксперимент по анализу качества предложенного метода.
5. Предложен метод генерации выборки из близкой генеральной совокупности на основе вариационного автокодировщика.
6. Проведен вычислительный эксперимент на реальных данных ImageNet с использованием отображения, изменяющего стиль изображений