© 2021 г. А.В. ГРАБОВОЙ (grabovoy.av@phystech.edu)
(Московский физико-технический институт),
В.В. СТРИЖОВ, д-р физ.-мат. наук (strijov@phystech.edu)
(Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН)

БАЙЕСОВСКАЯ ДИСТИЛЛЯЦИЯ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ[1]

Исследуется проблема понижения сложности аппроксимирующих моделей. Рассматриваются методы, основанные на дистилляции моделей глубокого обучения. Вводятся понятия учителя и ученика. Предполагается, что модель ученика имеет меньшее число параметров, чем модель учителя. Предлагается байесовский подход к выбору модели ученика. Предложен метод назначения априорного распределения параметров ученика на основе апостериорного распределения параметров модели учителя. Так как пространства параметров учителя и ученика не совпадают, предлагается механизм приведения пространства параметров модели учителя к пространству параметров модели ученика путем изменения структуры модели учителя. Проводится теоретический анализ предложенного механизма приведения. Вычислительный эксперимент проводился на синтетических и реальных данных. В качестве реальных данных рассматривается выборка FashionMNIST.

Ключевые слова: выбор модели; байесовский вывод; дистилляция модели; локальные преобразования; преобразования вероятностных пространств.

1. Введение

Исследуется проблема снижения числа обучаемых параметров моделей машинного обучения. Примерами моделей с избыточным числом параметров являются AlexNet [1], VGGNet [2], ResNet [3], BERT [4, 5], mT5 [6], GPT3 [7] и др. В табл. 1 приводится число параметров моделей глубокого обучения, которое с годами растет. Это влечет снижение интерпретируемости моделей. Данная проблема рассматривается в специальном классе задач по состязательным атакам (adversarial attack) [8]. Большое число параметров требует значительных вычислительных ресурсов. Из-за этого данные модели не могут быть использованы в мобильных устройствах. Для

[1]Настоящая статья содержит результаты проекта Математические методы интеллектуального анализа больших данных, выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы “Центр хранения и анализа больших данных”, поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ им. М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 №13/1251/2018. Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты №19-07-01155, №19-07-00875).

Таблица 1: Число параметров в моделях машинного обучения

Название	AlexNet	VGGNet	ResNet	BERT	mT5	GPT3
Год	2012	2014	2015	2018	2020	2020
Тип данных	изображение	изображение	изображение	текст	текст	текст
Число параметров, млрд	0,06	0,13	0,06	0,34	13	175

снижения числа параметров предложен метод дистилляции модели [9, 10, 11]. Дистиллируемая модель с большим числом параметров называется *учителем*, а модель, получаемая путем дистилляции, называется *учеником*. При оптимизации параметров модели ученика используется модель учителя с фиксированными параметрами.

Определение 1. Дистилляция модели — снижение сложности модели путем выбора модели в множестве более простых моделей на основе параметров и ответов более сложной фиксированной модели.

Идея дистилляции предложена Дж.Е. Хинтоном и В.Н. Вапником [9, 10, 11]. В их публикациях предлагалось использовать ответы учителя в качестве целевой переменной для обучения модели ученика. Поставлен ряд экспериментов, в которых проводилась дистилляция моделей для задачи классификации машинного обучения. Базовый эксперимент на выборке MNIST [12] показал результативную дистилляцию избыточно сложной нейросетевой модели в нейросетевую модель меньшей сложности. Проводился эксперимент по дистилляции ансамбля моделей в одну модель для решения задачи распознания речи. В [9] проведен эксперимент по обучению экспертных моделей на основе одной модели с большим числом параметров при помощи предложенного метода дистилляции на ответах учителя.

В [13] предложен метод передачи селективности нейронов (neuron selectivity transfer), основанный на минимизации специальной функции потерь. Метод основан на вычислении функции максимального среднего отклонения (maximum mean discrepancy) между выходами всех слоев модели учителя и ученика. Вычислительный эксперимент показал эффективность данного метода для задачи классификации изображений на примере выборок CIFAR [14] и ImageNet [15].

В данной статье предлагаются методы, основанные на байесовском выводе. В качестве априорного распределения параметров модели ученика предлагается использовать апостериорное распределение параметров модели учителя. Решается задача приведения пространства параметров модели учителя к пространству параметров модели ученика. Авторы предлагают подход, основанный на последовательном приведении пространства параметров модели учителя.

Определение 2. Структура модели — множество структурных параметров модели, которые задают вид суперпозиции.

Определение 3. Приведение параметрических моделей — изменение структуры модели (одной или нескольких моделей), в результате которого векторы параметров различных моделей лежат в одном пространстве.

В результате приведения параметры модели учителя и модели ученика лежат в одном пространстве. Как следствие, в качестве априорного распределения параметров модели ученика выбирается апостериорное распределение параметров моде-

ли учителя. В данной статье в качестве параметрических моделей рассматривается полносвязная нейронная сеть. В качестве структурных параметров модели выбраны число слоев, а также размер каждого скрытого слоя.

В рамках предложенного метода приведения параметрических моделей не оговорен выбор порядка на множестве параметров модели учителя. Для этого предлагается упорядочивать параметры модели учителя на основе их значимости [16]. Первый нейрон является наиболее значимым, а последний — нейрон наименее значимым. Порядок задается на основе отношения плотности распределения упорядочиваемого параметра к плотности распределения параметра в нуле [17] или на основе метода Белсли [18]. В рамках данной статьи порядок на параметрах задается случайный образом.

В рамках вычислительного эксперимента проводится теоретический анализ. Предложенный метод дистилляции анализируется на примере синтетической выборки, а также на реальной выборке FashionMnist [19].

2. Постановка задачи дистилляции

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}, \qquad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y}, \tag{1}$$

где $\mathbf{x}_i, y_i$ — признаковое описание и целевая переменная i-го объекта, число объектов в обучающей выборке обозначается m. Матрица признаковых описаний обозначается $\mathbf{X} = \left[\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_m^\mathsf{T}\right]^\mathsf{T}$, а вектор целевых переменных обозначается $\mathbf{y} = [y_1, \ldots, y_m]$. Размер признакового описания объектов обозначается n. Множество $\mathbb{Y} = \{1, \ldots, K\}$ для задачи классификации, где K число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регрессии.

Задана модель учителя в виде суперпозиций линейных и нелинейных преобразований:

$$f\big(\mathbf{x}\big) = \boldsymbol{\sigma} \circ \mathbf{U}_T \boldsymbol{\sigma} \circ \mathbf{U}_{T-1} \circ \ldots \mathbf{U}_2 \boldsymbol{\sigma} \circ \mathbf{U}_1 \mathbf{x},$$

где T — число слоев модели учителя, $\boldsymbol{\sigma}$ — функция активации, а $\mathbf{U}_t$ обозначает матрицу линейного преобразования. Матрицы $\mathbf{U}$ соединяются в вектор параметров $\mathbf{u}$ модели учителя f:

$$\mathbf{u} = \mathrm{vec}\big(\mathbf{U}_T, \mathbf{U}_{T-1}, \ldots, \mathbf{U}_1\big), \tag{2}$$

где vec — операция векторизации соединенных матриц. Каждая матрица $\mathbf{U}_t$ имеет размер $n_t \times n_{t-1}$, где $n_0 = n$, а $n_T = 1$ для задачи регрессии и $n_T = K$ для задачи классификации на K классов. Число параметров N_{tr} учителя f

$$N_{\mathrm{tr}} = \sum_{t=1}^{T} n_t n_{t-1}. \tag{3}$$

Для построения вектора параметров $\mathbf{u}$ задается полный порядок на элементах матриц $\mathbf{U}_t$. Для полносвязнной нейронной сети вводится естественный порядок, индуцированный номером слоя t, номером нейрона и номером элемента вектора параметров нейрона: выбирается матрица $\mathbf{U}_t$, строка этой матрицы и элемент строки.

Например, для модели учителя в задаче регрессии:

$$f\big(\mathbf{x}\big) = \boldsymbol{\sigma} \circ \mathbf{U}_3\boldsymbol{\sigma} \circ \mathbf{U}_2\boldsymbol{\sigma} \circ \mathbf{U}_1\mathbf{x} \tag{4}$$

вектор параметров $\mathbf{u}$ принимает вид

$$\mathbf{u} = \big[u_1^{1,1}, \ldots, u_1^{1,n}, \ldots, u_1^{n_1,1}, \ldots, u_1^{n_1,n}, u_2^{1,1}, \ldots, u_2^{1,n_1}, \ldots, u_2^{n_2,1}, \ldots, u_2^{n_2,n_1}, u_3^{1,1}, \ldots, u_3^{1,n_2}\big].$$

Пусть для вектора параметров $\mathbf{u}$ учителя f известно апостериорное распределение параметров $p\big(\mathbf{u}|\mathfrak{D}\big)$.

На основе выборки $\mathfrak{D}$ и апостериорного распределения параметров учителя f требуется выбрать модель ученика из параметрического семейства функций:

$$g\big(\mathbf{x}\big) = \boldsymbol{\sigma} \circ \mathbf{W}_L\boldsymbol{\sigma} \circ \ldots \circ \mathbf{W}_1\mathbf{x}, \quad \mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l-1}},$$

где L — число слоев модели ученика. Число параметров N_{st} модели ученика g вычисляется аналогично выражению (3). Вектор параметров модели ученика $\mathbf{w}$ строится аналогичным образом (2). Модель g задается своим вектором параметров $\mathbf{w}$. Следовательно, задача выбора модели g эквивалентна задаче оптимизации вектора параметров $\mathbf{w} \in \mathbb{R}^{N_{\text{st}}}$.

Параметры $\hat{\mathbf{w}} \in \mathbb{R}^{N_{\text{st}}}$ оптимизируются при помощи вариационного вывода на основе совместного правдоподобия модели и данных:

$$\mathcal{L}\big(\mathfrak{D}, \mathbf{A}\big) = \log p\big(\mathfrak{D}|\mathbf{A}\big) = \log \int\limits_{\mathbf{w}\in\mathbb{R}^{N_{\text{st}}}} p\big(\mathfrak{D}|\mathbf{w}\big)p\big(\mathbf{w}|\mathbf{A}\big)d\mathbf{w}, \tag{5}$$

где $p\big(\mathbf{w}|\mathbf{A}\big)$ — априорное распределение вектора параметров модели ученика, $\mathbf{A}$ обозначает гиперпараметры априорного распределения. Взятие интеграла (5) является вычислительно сложной задачей. В качестве приближенного решения используется вариационный подход [17, 18]. Для этого задается вариационное распределение параметров модели ученика $q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$, которое аппроксимирует неизвестное апостериорное распределение $p\big(\mathbf{w}|\mathfrak{D}\big)$

$$q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big) \approx p\big(\mathbf{w}|\mathfrak{D}\big),$$

где оптимальные гиперпараметры распределения $\hat{\boldsymbol{\mu}}$ и $\hat{\boldsymbol{\Sigma}}$ требуется найти вместе с оптимальными параметрами $\hat{\mathbf{w}}$, решив оптимизационную задачу:

$$\hat{\mathbf{w}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{w}} \mathrm{D}_{\mathrm{KL}}\big(q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)||p\big(\mathbf{w}|\mathbf{A}\big)\big) - \sum_{i=1}^{m} \log p\big(y_i|\mathbf{x}_i, \mathbf{w}\big), \tag{6}$$

где D_{KL} обозначает расстояние Кульбака–Лейблера между вариационным распределением $q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$ и априорным распределением $p\big(\mathbf{w}|\mathbf{A}\big)$. Второе слагаемое формулы (6) является логарифмом правдоподобия $\log p\big(y_i|\mathbf{x}_i, \mathbf{w}\big)$ объекта $(\mathbf{x}_i, y_i) \in \mathfrak{D}$ выборки (1). Выражение (6) не учитывает параметры учителя f. Для использования информации о распределении параметров учителя предлагается рассмотреть параметры априорного распределения $p\big(\mathbf{w}|\mathbf{A}\big)$ как функцию от апостериорного распределения учителя $p\big(\mathbf{u}|\mathfrak{D}\big)$.

3. Построение априорного распределения ученика

Апостериорное распределение параметров модели учителя предполагается нормальным:

$$p\big(\mathbf{u}|\mathfrak{D}\big) = \mathcal{N}\big(\mathbf{m}, \mathbf{\Sigma}\big), \tag{7}$$

где $\mathbf{m}$ и $\mathbf{\Sigma}$ — гиперпараметры этого распределения. На основе гиперпараметров $\mathbf{m}$ и $\mathbf{\Sigma}$ требуется задать параметры $\mathbf{A}$ априорного распределения $p\big(\mathbf{w}|\mathbf{A}\big)$. Когда структура моделей учителя и ученика задаются числом слоев и размером этих слоев, возможны следующие варианты: 1) число слоев и размер каждого слоя совпадают; 2) число слоев совпадает, а размеры слоев различаются; 3) не совпадает число слоев.

3.1. Учитель и ученик имеют одну структуру

Рассмотрим следующие условия:

1) число слоев модели ученика равняется числу слоев модели учителя $L = T$;

2) размеры соответствующих слоев совпадают, другими словами, для всех t, l, таких что $t = l$, выполняется $n_l = n_t$, где n_t обозначает размер t-го слоя учителя, а n_l — размер l-го слоя ученика.

В случае выполнения этих условий априорное распределение параметров модели ученика приравнивается к апостериорному распределению параметров учителя $p\big(\mathbf{w}|\mathbf{A}\big) = p\big(\mathbf{u}|\mathfrak{D}\big)$.

3.2. Удаление нейрона в слое учителя

Приведем структуру модели учителя к структуре модели ученика согласно определению 3 при помощи последовательных преобразований вектора параметров $\mathbf{u}$. Рассмотрим преобразование

$$\phi\big(t, \mathbf{u}\big) : \mathbb{R}^{N_{\text{tr}}} \to \mathbb{R}^{N_{\text{tr}} - 2n_t}$$

вектора $\mathbf{u}$, которое описывает удаление одного нейрона из t-го слоя учителя. Обозначим новый вектор параметров $\boldsymbol{\upsilon} = \phi\big(t, \mathbf{u}\big)$, а элементы вектора, которые были удалены, — через $\bar{\boldsymbol{\upsilon}}$. Заметим, что векторы $\boldsymbol{\upsilon}$ и $\bar{\boldsymbol{\upsilon}}$ являются случайными величинами.

Теорема 1. Пусть выполняются следующие условия:

1) *апостериорное распределение $p\big(\mathbf{u}|\mathfrak{D}\big)$ параметров модели учителя является нормальным распределением* (7)*;*

2) *число слоев модели учителя равняется числу слоев модели ученика $T = L$;*

3) *размеры соответствующих слоев не совпадают, другими словами, для всех t, l, таких что $t = l$, выполняется $n_t \leqslant n_l$.*

Тогда апостериорное распределение параметров модели учителя $p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big)$ также является нормальным.

Доказательство. Не уменьшая общности, пусть $\phi\big(t, \mathbf{u}\big)$ удаляет j-й нейрон в t-м слое, что является удалением j-й строки матрицы $\mathbf{U}_t$. Заметим, что удаление j-й строки матрицы $\mathbf{U}_t$ влечет удаление j-й компоненты вектора z_{t+1}, где

$$\mathbf{z}_t = \boldsymbol{\sigma} \circ \mathbf{U}_{t-1}\boldsymbol{\sigma} \circ \ldots \mathbf{U}_2\boldsymbol{\sigma} \circ \mathbf{U}_1\mathbf{x}.$$

Удаление j-й компоненты вектора $\mathbf{z}_{t+1}$ эквивалентно занулению j-го столбца матрицы $\mathbf{U}_{t+1}$. Заметим, что тогда предсказание модели не зависит от параметров j-й строки матрицы $\mathbf{U}_t$, а поэтому данными параметрами также можно пренебречь.

Найдем распределение вектора $\boldsymbol{\upsilon}$. Для поиска распределения вектора параметров после зануления j-го столбца матрицы $\mathbf{U}_{t+1}$ воспользуемся формулой условной вероятности $p\big(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big)$, а для удаления j-й строки матрицы $\mathbf{U}_t$ воспользуемся маргинализацией распределения $p\big(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big)$. Обозначим зануляемые параметры модели через $\boldsymbol{\nu}_1$, а удаляемые параметры — через $\boldsymbol{\nu}_2$. Также обозначим все параметры, которые не были занулены, через $\bar{\boldsymbol{\nu}}_1 = [\boldsymbol{\upsilon}^\mathsf{T}, \boldsymbol{\nu}_2^\mathsf{T}]$. Итоговое распределение параметров принимает вид:

$$p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big) = \int\limits_{\boldsymbol{\nu}_2} p\big(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big) d\boldsymbol{\nu}_2.$$

Из свойств нормального распределения следует, что распределение

$$p\big(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big) \tag{8}$$

также является нормальным распределением с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

$$\boldsymbol{\mu} = \mathbf{m}_{\bar{\nu}_1} + \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\boldsymbol{\nu}_1}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\nu}_1,\boldsymbol{\nu}_1}\left(\mathbf{0} - \mathbf{m}_{\boldsymbol{\nu}_1}\right),$$
$$\boldsymbol{\Xi} = \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\bar{\boldsymbol{\nu}}_1} - \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\boldsymbol{\nu}_1}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\nu}_1,\boldsymbol{\nu}_1}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\boldsymbol{\nu}_1},$$

где введенные обозначения $\mathbf{m}_{\bar{\nu}_1}, \mathbf{m}_{\boldsymbol{\nu}_1}$ соответствуют подвектору вектора $\mathbf{m}$, который относится к параметрам $\bar{\boldsymbol{\nu}}_1$ и $\boldsymbol{\nu}_1$ соответсвенно. Ковариационная матрица $\boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\boldsymbol{\nu}_1}$ обозначает подматрицу матрицы $\boldsymbol{\Sigma}$, которая соответсвует ковариационной матрицей между параметрами $\bar{\boldsymbol{\nu}}_1$ и $\boldsymbol{\nu}_1$.

Распределение $p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big)$ найдем при помощи маргинализации распределения (8) по параметрам $\boldsymbol{\nu}_2$. Используя свойства нормального распределения, получаем распределение

$$p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big) = \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{\upsilon}}, \boldsymbol{\Xi}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}}\big), \tag{9}$$

где $\boldsymbol{\mu}_{\boldsymbol{\upsilon}}$ обозначает подвектор вектора $\boldsymbol{\mu}$, который относится к вектору параметров $\boldsymbol{\upsilon}$, а матрица $\boldsymbol{\Xi}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}}$ является подматрицей матрицы $\boldsymbol{\Xi}$, которая относится к вектору параметров $\boldsymbol{\upsilon}$.

Теорема 1 задает апостериорное распределение параметров (9) после зануления нейронов в модели нейросети — учителя. Заметим, что аналогичным образом можно удалить сразу подмножество нейронов в рамках одного слоя. В случае если число нейронов отличается в нескольких слоях модели нейросети учителя, то выполняются последовательно применения отображения $\phi\big(t, \mathbf{u}\big)$ для каждого t-го слоя.

3.3. Удаление слоя учителя

Приведем структуру модели учителя к модели ученика согласно определению 3 при помощи последовательных преобразований вектора параметров $\mathbf{u}$. Рассмотрим преобразование

$$\psi\big(t, \mathbf{u}\big) : \mathbb{R}^{N_{\text{tr}}} \to \mathbb{R}^{N_{\text{tr}} - n_t n_{t-1}}$$

вектора $\mathbf{u}$, которое описывает удаление одного t-го слоя. Обозначим новый вектор параметров $\boldsymbol{\upsilon} = \psi\big(t, \mathbf{u}\big)$, а элементы вектора, которые были удалены, — через $\bar{\boldsymbol{\upsilon}}$.

Теорема 2. Пусть выполняются следующие условия:

1) *апостериорное распределение параметров* $p\big(\mathbf{u}|\mathfrak{D}\big)$ *модели учителя является нормальным распределением* (7)*;*

2) *соответствующие размеры слоев совпадают,* $n_t = n_{t-1}$*, т.е. матрица* $\mathbf{U}_t$ *является квадратной;*

3) *функция активации удовлетворяет свойству идемпотентности* $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$*.*

Тогда апостериорное распределение также описывается нормальным распределением с плотностью распределения

$$p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big) = \mathcal{N}\big(\mathbf{m}_{\boldsymbol{\upsilon}} + \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\upsilon}},\bar{\boldsymbol{\upsilon}}}^{-1}\left(\mathbf{i} - \bar{\boldsymbol{\upsilon}}\right), \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\upsilon}},\bar{\boldsymbol{\upsilon}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\big), \tag{10}$$

где вектор $\mathbf{i}$ *задается как*

$$\mathbf{i} = [\underbrace{1,0,\ldots,0}_{n_t},\underbrace{0,1,\ldots,0}_{n_t},\underbrace{0,0,1,\ldots,0}_{n_t},\underbrace{0,\ldots,1}_{n_t}]^{\mathsf{T}}.$$

Доказательство. Рассмотрим структуру нейронной сети с T слоями и $T+1$ слоем. Не уменьшая общности, для удаления рассматривается t-й слой, для которого выполняются условия этой теоремы. Заметим, что если t-й слой нейронной сети с $T+1$ слоем приравнять к единичной матрице, то он будет эквивалентным архитектуре с T слоями:

$$\begin{aligned} f &= \boldsymbol{\sigma} \circ \mathbf{U}_{T+1}\boldsymbol{\sigma} \circ \mathbf{U}_T \ldots \boldsymbol{\sigma} \circ \mathbf{U}_t\boldsymbol{\sigma} \circ \ldots \mathbf{U}_2\boldsymbol{\sigma} \circ \mathbf{U}_1 = \\ &= \boldsymbol{\sigma} \circ \mathbf{U}_{T+1}\boldsymbol{\sigma} \circ \mathbf{U}_T \ldots \boldsymbol{\sigma} \circ \mathbf{I}\boldsymbol{\sigma} \circ \ldots \mathbf{U}_2\boldsymbol{\sigma} \circ \mathbf{U}_1 = \\ &= \boldsymbol{\sigma} \circ \mathbf{U}_{T+1}\boldsymbol{\sigma} \circ \mathbf{U}_T \ldots \boldsymbol{\sigma} \circ \boldsymbol{\sigma} \circ \ldots \mathbf{U}_2\boldsymbol{\sigma} \circ \mathbf{U}_1 = \\ &= \boldsymbol{\sigma} \circ \mathbf{U}_{T+1}\boldsymbol{\sigma} \circ \mathbf{U}_T \ldots \boldsymbol{\sigma} \circ \ldots \mathbf{U}_2\boldsymbol{\sigma} \circ \mathbf{U}_1. \end{aligned}$$

Получаем, что удаление t-го слоя нейросети эквивалентно приравниванию матрицы параметров t-го слоя к единичной матрице. Распределение параметров после приравнивая к единичной матрице вычисляется при помощи условного распределения. В силу общих свойств нормального распределения условное распределение также является нормальным распределением с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{m}_{\boldsymbol{\upsilon}} + \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\upsilon}},\bar{\boldsymbol{\upsilon}}}^{-1}\left(\mathbf{i} - \bar{\boldsymbol{\upsilon}}\right), \\ \boldsymbol{\Xi} &= \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\upsilon}},\bar{\boldsymbol{\upsilon}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\big), \end{aligned}$$

где вектор $\mathbf{m}_{\boldsymbol{\upsilon}}$ является подвектором вектора $\mathbf{m}$ соответсвующей параметрам $\boldsymbol{\upsilon}$, а матрица $\boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}$ является подматрицей ковариационной матрицы $\boldsymbol{\Sigma}$, соответствующей векторам параметров $\boldsymbol{\upsilon}$ и $\bar{\boldsymbol{\upsilon}}$.

Теорема 2 задает апостериорное распределение (10) параметров после удаления слоя нейросети. Полученное распределение $p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big)$ является оценкой апостериорного распределения модели без одного слоя.

3.4. Выполнение последовательных преобразований

Преобразования ϕ, ψ приводят пространство параметров учителя f к пространству параметров ученика g. После приведения параметрических моделей получаем, что параметры модели учителя и модели ученика принадлежат одному семейству 3.1.

4. Вычислительный эксперимент

Вычислительный эксперимент анализирует предложенный метод дистилляции на основе апостериорного распределения параметров модели учителя.

4.1. Синтетические данные

Проанализируем модель на синтетической выборке. Выборка построена следующим образом:

$$\mathbf{w} = \big[w_j : w_j \sim \mathcal{N}\big(0,1\big)\big]_{n\times 1}, \quad \mathbf{X} = \big[x_{ij} : x_{ij} \sim \mathcal{N}\big(0,1\big)\big]_{m\times n},$$
$$\mathbf{y} = \big[y_i : y_i \sim \mathcal{N}\big(\mathbf{x}_i^{\mathsf{T}}\mathbf{w}, \beta\big)\big]_{m\times 1},$$

где $\beta = 0{,}1$ — уровень шума в данных. В эксперименте число признаков $n = 10$, для обучения и тестирования было сгенерировано $m_{\text{train}} = 900$ и $m_{\text{test}} = 124$ объекта.

В качестве модели учителя рассматривалась модель — многослойный перцептрон с двумя скрытыми слоями (4). Матрицы линейных преобразований имеют размер:

$$\mathbf{U}_1 \in \mathbb{R}^{100\times 10}, \quad \mathbf{U}_2 \in \mathbb{R}^{50\times 100}, \quad \mathbf{U}_3 \in \mathbb{R}^{1\times 50}.$$

В качестве функции активации была выбрана функция активации ReLu. Модель учителя предварительно обучена на основе вариационного вывода (6), где в качестве априорного распределения параметров выбрано стандартное нормальное распределение.

В качестве модели ученика были выбраны две конфигурации. Первая конфигурация получается путем удаления нейронов в модели учителя:

$$g\big(\mathbf{x}\big) = \boldsymbol{\sigma} \circ \mathbf{W}_3\boldsymbol{\sigma} \circ \mathbf{W}_2\boldsymbol{\sigma} \circ \mathbf{W}_1\mathbf{x}, \tag{11}$$

где $\boldsymbol{\sigma}$ является нелинейной функцией активации, а матрицы линейных преобразований имеют размер:

$$\mathbf{W}_1 \in \mathbb{R}^{10\times 10}, \quad \mathbf{W}_2 \in \mathbb{R}^{10\times 10}, \quad \mathbf{W}_3 \in \mathbb{R}^{1\times 10}.$$

В качестве функции активации была выбрана функция активации ReLu.

На рис. 1 сравниваются модели ученика со структурой (11). Представлено сравнение разных моделей: модель без дистилляции (график 1), где в качестве априорного распределения выбирается стандартное нормальное распределение; модель с частичной дистилляцией (график 2), где в качестве среднего значения параметров выбираются параметры согласно (9), а ковариационная матрица была приравнена к единичной матрице; модель с полной дистилляцией (график 3) согласно (9). Видно, что модели ученика, где в качестве априорного распределения выбраны распределения, основанные на апостериорном распределении учителя имеют большее правдоподобие, чем модель, где в качестве априорного распределения выбрано стандартное нормальное. Также заметим, что использование параметра среднего из апостериорного распределения дает основной вклад при дистилляции, так как качество моделей без дистилляции и с полной дистилляцией совпадает.

Вторая конфигурация получается путем удаления слоя модели учителя:

$$g\big(\mathbf{x}\big) = \boldsymbol{\sigma} \circ \mathbf{W}_2\boldsymbol{\sigma} \circ \mathbf{W}_1\mathbf{x}, \tag{12}$$

где $\boldsymbol{\sigma}$ является нелинейной функцией активации, а матрицы линейных преобразований имеют размер:

$$\mathbf{W}_1 \in \mathbb{R}^{50\times 10}, \quad \mathbf{W}_2 \in \mathbb{R}^{1\times 50}.$$

В качестве функции активации была выбрана функция активации ReLu.

На рис. 2 сравниваются модели ученика со структурой (12). Аналогично рис. 1 на рис. 2 представлено сравнение модели без дистилляции (график 1), модели с дистилляцией параметра среднего значение (график 2) и модели с полной дистилляцией (график 3). В рамках данного эксперимента по дистилляции модели учителя в модель ученика с меньшим числом параметров получены результаты, которые подтверждают, что задание априорного распределения параметров ученика позволяет улучшить число итераций при выборе оптимальных параметров модели ученика.

4.2. Выборка FashionMnist

В рамках данного эксперимента проводился анализ байесовского подхода к дистилляции на реальных данных. В качестве реальных данных выбрана выборка FashionMnist [19], которая является задачей классификации изображений на 10 классов.

В качестве модели учителя рассматривалась модель многослойный перцептрон с двумя скрытыми слоями (4). Матрицы линейных преобразований имеют размер:

$$\mathbf{U}_1 \in \mathbb{R}^{800\times 784}, \quad \mathbf{U}_2 \in \mathbb{R}^{50\times 800}, \quad \mathbf{U}_3 \in \mathbb{R}^{10\times 50}.$$

В качестве функции активации была выбрана функция активации ReLu. Модель учителя предварительно обучена на основе вариационного вывода (6), где в качестве априорного распределения параметров выбрано стандартное нормальное распределение.

В качестве модели ученика были выбрана конфигурация с одним скрытым слоем (12), где матрицы линейных преобразований имеют размер:

$$\mathbf{W}_1 \in \mathbb{R}^{50\times 784}, \quad \mathbf{W}_2 \in \mathbb{R}^{50\times 10}.$$

В качестве функции активации была выбрана функция активации ReLu.

На рис. 3 сравниваются модели ученика с разными априорными распределениями параметров. Аналогично синтетическому эксперименту модель, где в качестве априорного распределения использовалось стандартное нормальное распределение, сравнивалась с моделью, где параметры распределения определялись на основе формулы (10). Видно, что у моделей с заданием априорного распределения на основе апостериорного распределения параметров учителя правдоподобие выборки выше, чем у модели, где в качестве априорного распределения выбрано стандартное нормальное распределение.

В табл. 2 представлен результат вычислительного эксперимента. Для численного сравнения качества моделей выбрана разность площадей графика $p\big(\mathbf{y}|\mathbf{X},\mathbf{u}\big)$ между моделью без дистилляции и моделями с частичной дистилляцией и полной дистилляцией соответственно:

$$S = \sum_s p\big(\mathbf{y}|\mathbf{X},\mathbf{u}_{\mathrm{s}}^s\big) - p\big(\mathbf{y}|\mathbf{X},\mathbf{u}_{\mathrm{ds}}^s\big), \tag{13}$$

Таблица 2: Сводная таблица результатов вычислительного эксперимента

	учитель	модель ученика без дистилляции	модель ученика с частичной дистилляцией	модель ученика с полной дистилляцией
Эксперимент на синтетической выборке (удаление нейрона)				
Структура	$[10, 100, 50, 1]$	$[10, 10, 10, 1]$	$[10, 10, 10, 1]$	$[10, 10, 10, 1]$
Число параметров	6050	210	210	210
Разность площадей S	-	0	16559	16864
Эксперимент на синтетической выборке (удаление слоя)				
Структура	$[10, 100, 50, 1]$	$[10, 50, 1]$	$[10, 50, 1]$	$[10, 50, 1]$
Число параметров	6050	550	550	550
Разность площадей S	-	0	23310	25506
Эксперимент на выборке FashionMnist				
Структура	$[784, 800, 50, 10]$	$[784, 50, 10]$	$[784, 50, 10]$	$[784, 50, 10]$
Число параметров	667700	39700	39700	39700
Разность площадей S	-	0	1165	1145

где $\mathbf{u}_{\mathrm{s}}^{s}, \mathbf{u}_{\mathrm{ds}}^{s}$ обозначают параметры модели ученика и модели дистиллированного ученика после s-й итерации оптимизационного процесса. Заметим, что площадь S имеет знак: чем большее положительное число, тем дистиллированная модель лучше, чем модель, построенная без учителя. Если площадь S принимает отрицательное значение, то, значит, модель без дистилляции является лучше, чем модель с дистилляцией. В рамках вычислительного эксперимента видно, что площадь S под графиками принимает положительные значения, т.е. модели ученика полученные при помощи дистилляции являются лучше чем модель ученика без дистилляции.

Код вычислительного эксперимента доступен по ссылке `https://github.com/andriygav/BayesianDistilation`.

5. Заключение

В данной статье проанализирована байесовская дистиляция модели учителя в модель ученика на основе вариационного вывода. В рамках данной статьи дистилляция основывается на задании априорного распределения параметров модели ученика. Априорное распределение параметров модели ученика задается на основе апостериорного распределения параметров модели учителя. Механизм преобразования структуры модели учителя в структуру модели ученика представлен в теореме 1 и теореме 2.

Теорема 1 описывает механизм приведения пространства параметров модели учителя к пространству параметров модели ученика в случае, если число слоев совпадает, но размер слоев различается. Теорема 2 описывает механизм приведения пространства параметров модели учителя к пространству параметров модели ученика в случае, если число слоев различается.

В вычислительном эксперименте сравнивается модель ученика, которая обучена без использования распределения параметров учители и модель ученика, где в качестве априорного распределения параметров выбрано апостериорное распределение

параметров модели учителя после приведения. В табл. 2 показано, что модели ученика с заданием априорного распределения параметров на основе апостериорного распределения параметров учителя сходятся быстрее, что подтверждается положительным значением метрики (13), которая введена для численного сравнения модели без дистилляции с дистиллированной моделью.

СПИСОК ЛИТЕРАТУРЫ

1. *Krizhevsky A., Sutskever I., Hinton G.* ImageNet Classification with Depp Convolutional Neural Networks // Proc. 25th Int. Conf. on Neural Information Procescing Systems. 2012. V. 1. P. 1097–1105.

2. *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition // Int. Conf. on Learning Representations. San Diego. 2015.

3. *He K., Ren S., Sun J., Zhang X.* Deep Residual Learning for Image Recognition // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas. 2016. P. 770–778.

4. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minnesota. 2019. V. 1. P. 4171–4186.

5. *Vaswani A., Gomez A., Jones L., Kaiser L., Parmar N., Polosukhin I., Shazeer N., Uszkoreit J.* Attention Is All You Need // In Advances in Neural Information Processing Systems. 2017. V. 5. P. 6000–6010.

6. *Al-Rfou R., Barua A., Constant N., Kale M., Raffel C., Roberts A., Siddhant A., Xue L.* mT5: A massively multilingual pre-trained text-to-text transformer // Proc. 2021 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. P. 483–498.

7. *Brown T., et al* GPT3: Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020. V. 33. P. 1877–1901.

8. *Zheng T., Liu X., Qin Z., Ren K.* Adversarial Attacks and Defenses in Deep Learning // Engineering. 2020. V. 6. P. 346–360.

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

10. *Vapnik V., Izmailov R.* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // J. of Machine Learning Research. 2015. V. 16. P. 2023–2049.

11. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // Int. Conf. on Learning Representations. Puerto Rico. 2016.

12. *Burges C., Cortes C., LeCun Y.* The MNIST dataset of handwritten digits. 1998. http://yann.lecun.com/exdb/mnist/index.html.

13. *Huang Z., Naiyan W.* Like What You Like: Knowledge Distill via Neuron Selectivity Transfer // arXiv:1707.01219. 2017.

14. *Hinton G., Krizhevsky A., Nair V.* CIFAR-10 (Canadian Institute for Advanced Research) // `http://www.cs.toronto.edu/~kriz/cifar.html`

15. *Deng J., et al* Imagenet: A large-scale hierarchical image database // Proc. IEEE Conf. on computer vision and pattern recognition. Miami. 2009. P. 248–255.

16. *LeCun Y., Denker J., Solla S.* Optimal Brain Damage // Advances in Neural Information Processing Systems. 1989. V. 2. P. 598–605.

17. *Graves A.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems. 2011. V. 24. P. 2348–2356.

18. *Grabovoy A.V., Bakhteev O.Y., Strijov V.V.* Estimation of relevance for neural network parameters // Informatics and Applications. 2019. V. 13 No. 2. P. 62–70.

19. *Rasul K., Vollgraf R., Xiao H.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms // arXiv preprint arXiv:1708.07747. 2017.

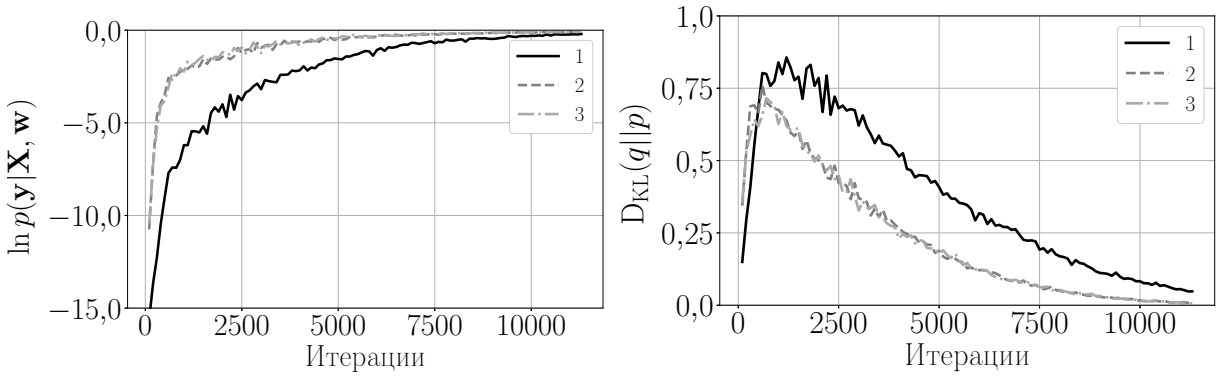

Рис. 1

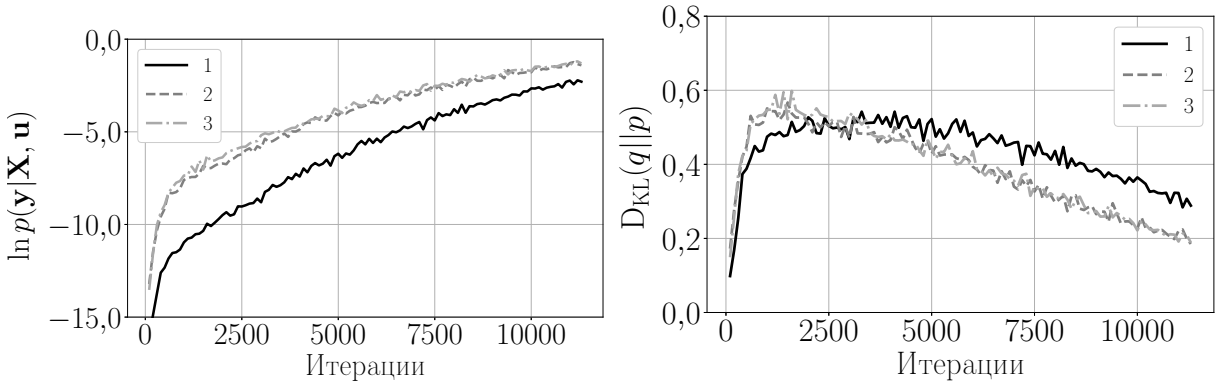

Рис. 2

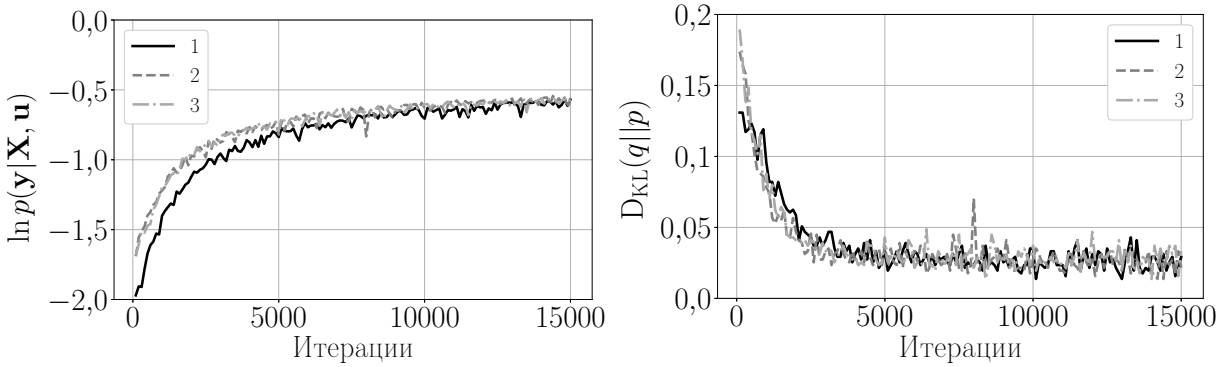

Рис. 3

Рис. 1. Структура (11) модели ученика g. Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: дивергенция Кульбака–Лейблера между вариационным и априорным распределениями параметров модели.

Рис. 2. Структура (12) модели ученика g. Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: дивергенция Кульбака–Лейблера между вариационным и априорным распределениями параметров модели.

Рис. 3. Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: дивергенция Кульбака–Лейблера между вариационным и априорным распределениями параметров модели.