© 2021 г. А.В. ГРАБОВОЙ
(Московский физико-технический институт, Москва)
В.В. СТРИЖОВ, д-р физ-мат. наук
(Вычислительный центр имени 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. Вводится понятия учителя и ученика. Предполагается, что модель ученика имеет меньшее число параметров, чем модель учителя. Предлагается байесовский подход к выбору модели ученика. Авторами 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 сопоставления этих пространств путем изменения струткуры учителя. Проводится теоретический анализ предложенного механизма сопоставления. Вычислительный эксперимент проводится 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 таких моделей, с избыточным число параметров, являются AlexNet [5], VGGNet [6], ResNet [12], BERT [8, 7], mT5 [10], GPT3[9] и другие. Табл. 1

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

описывает глубокие модели машинного обучения. Число параметров моделей машинного обучения с годами растет. Это влечет снижение интерпретируемости моделей. Данная проблема рассматривается в специальном классе задач по состязательным атакам (англ. adversarial attack) [4]. Большое число параметров требует больших вычислительных ресурсов. Из-за этого данные модели не могут быть использованы в

[1]Работа выполнена при поддержке ... (грант №...).

мобильных устройствах. Для снижения числа параметров предложен метод дистилляции модели [14, 26, 27]. Дистиллируемая модель с большим числом параметров называется *учитель*, а модель получаемая путем дистилляции называется *ученик*. 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 в работах Дж. Е. Хинтона и В. Н. Вапником [14, 26, 27]. В этих работах предлагается 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 [15] показал применимость метода для дистилляции избыточно сложной нейросетевой модели в нейросетевую модель меньшей сложности. Эксперимент по распознаванию речи, в котором модель строилась путем дистилляции ансамбля моделей. Также в работе [14] был проведен эксперимент по обучению экспертных моделей на основе одной модели с большим числом параметров при помощи предложенного метода дистилляции на ответах учителя.

В работе [3] предложен метод передачи селективности нейронов (англ. neuron selectivity transfer) основаный на минимизации специальной функции потерь основаной на максимальном среднем отклонении (англ. 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 [1] и ImageNet [2].

В данной работе предлагается метод основанный на байесовском выводе. В качестве априорного распределения параметров модели ученика предлагается использовать апостериорное распределение параметров модели учителя. Решается задача сопоставления пространства параметров модели учителя и модели ученика. Авторы предлагают подход, основаный на последовательном сопоставлении пространств параметров модели ученика и учителя.

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

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

В результате сопоставления, параметры модели учителя и модели ученика лежат в одном пространстве. Как следствие в качестве априорного распределения параметров модели ученика выбирается апостериорное распределение параметров модели учителя. В данной работе 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 сопоставления параметрических моделей, не оговорен выбор порядка на множестве параметров модели учителя. Для этого предлагается упорядочивать параметры модели учителя на основе их значимости. Первый нейрон является наиболее значимым, а последний нейрон наименее значимым. Поря-

док задается на основе отношения плотности распределения параметра к плотности распределения данного параметра в нуле [24] или на основе метода Белсли [25] и т.д. В рамках данной работы порядок на параметрах в рамках одного слоя задается случайный образом.

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

2. Постановка задачи дистилляции

Задана выборка

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

где $\mathbf{x}_i, y_i$ — признаковое описание и целевая переменная i-го объекта, число объектов в обучающей выборке обозначается m. Размер признакового описания объектов обозначается n. Множество $\mathbb{Y} = \{1, \cdots, K\}$ 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:

$$f = \boldsymbol{\sigma} \circ \mathbf{U}_T \boldsymbol{\sigma} \circ \mathbf{U}_{T-1} \circ \cdots \mathbf{U}_2 \boldsymbol{\sigma} \circ \mathbf{U}_1,$$

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:

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

где vec операция векторизация последовательности матриц в вектор параметров. 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 ν учителя f:

$$\nu = \sum_{t=1}^{T} n_t n_{t-1}. \tag{2}$$

Для построения вектора параметров $\mathbf{u}$ задается полный порядок на элементов 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:

$$f\big(\mathbf{x}\big) = \boldsymbol{\sigma} \circ \mathbf{U}_3 \boldsymbol{\sigma} \circ \mathbf{U}_2 \boldsymbol{\sigma} \circ \mathbf{U}_1 \mathbf{x}, \tag{3}$$

вектор параметров $\mathbf{u}$ принимает вид:

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

Пусть для вектора параметров 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:

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

где L число слоев модели ученика. Число параметров υ модели ученика g вычисляется аналогично выражению (2). Вектор параметров модели ученика $\mathbf{w}$ строится аналогичным образом (1). Модель g задается своим вектором параметров $\mathbf{w}$. Следовательно, задача выбора модели g эквивалентна задаче оптимизации вектора параметров $\mathbf{w} \in \mathbb{R}^{\upsilon}$.

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

$$\mathcal{L}\big(\mathfrak{D}, \mathbf{A}\big) = \log p\big(\mathfrak{D}|\mathbf{A}\big) = \log \int_{\mathbf{w}\in\mathbb{R}^{\upsilon}} p\big(\mathfrak{D}|\mathbf{w}\big) p\big(\mathbf{w}|\mathbf{A}\big) d\mathbf{w}, \tag{4}$$

где $p\big(\mathbf{w}|\mathbf{A}\big)$ — априорное распределение вектора параметров модели ученика. Так как взятие интеграла (4) является вычислительно сложной задачей, используем вариационный подход [24, 25]. Для этого зададим вариационное распределение параметров модели ученика $q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$, которое аппроксимирует неизвестное апостериорное распределение $p\big(\mathbf{w}|\mathfrak{D}\big)$:

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

Далее распределение $q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$ будем обозначать просто $q\big(\mathbf{w}\big)$. Оптимизация параметров $\mathbf{w}$ сводится к решению задачи:

$$\hat{\mathbf{w}} = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{w}} \mathrm{D}_{\mathrm{KL}}\big(q\big(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)||p\big(\mathbf{w}|\mathbf{A}\big)\big) - \log p\big(\mathbf{y}|\mathbf{X}, \mathbf{w}\big). \tag{5}$$

Выражение (5) не учитывает параметры учителя f. Для использования информации о распределении параметов 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

Зададим апостериорное распределение параметров модели учителя как нормальное распределение:

$$p\big(\mathbf{u}|\mathfrak{D}\big) = \mathcal{N}\big(\mathbf{m}, \boldsymbol{\Sigma}\big), \tag{6}$$

где $\mathbf{m}$ и $\boldsymbol{\Sigma}$ параметры этого распределения. На основе параметров $\mathbf{m}$ и $\boldsymbol{\Sigma}$ требуется задать параметры $\mathbf{A}$ априорного распределения $p\big(\mathbf{w}|\mathbf{A}\big)$. В случае, когда структура моделей учителя и ученика задаются числом слоев и размером этих слоев, то возможны следующие варианты: число слоев и размер каждого слоя совпадает; число слоев совпадает, а размеры различаются; не совпадает число слоев.

3.1. Учитель и ученик принадлежат одному семейству

Рассмотрим следующие условия:

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

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 пприравнивается к апостериорному распределения параметров учителя, то есть $p\big(\mathbf{w}|\mathbf{A}\big) = p\big(\mathbf{u}|\mathfrak{D}\big)$.

3.2. Удаление нейрона в слое учителя

Проведем согласование модели учителя и модели ученика, согласно определению 3 при помощи последовательных преобразований параметров $\mathbf{u}$. Рассмотрим преобразование:

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

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 как $\bar{\boldsymbol{\upsilon}}$. 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* (6)*;*

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

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

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

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:

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

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 следовательно донными параметрами также можно пренебречь.

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

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

Из свойств нормального распределения следует, что распределение $p\big(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big)$ 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 введены обозначения $\mathbf{m}_{\bar{\boldsymbol{\nu}}_1}, \mathbf{m}_{\boldsymbol{\nu}_1}$ соответствует 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 $p\big(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big)$ по параметрам $\boldsymbol{\nu}_2$. Используя свойства нормального распределения получаем распределения:

$$p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big) = \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{\upsilon}}, \boldsymbol{\Xi}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}}\big), \tag{7}$$

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

Теорема 1 задает апостериорное распределение параметров (7) 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 выполняется последовательно применения отображения $\phi\big(t, \mathbf{u}\big)$ для каждого t-го слоя.

3.3. Удаление слоя учителя

Проведем согласование модели учителя и модели ученика, согласно определения 3 при помощи последовательных преобразований вектора параметров $\mathbf{u}$. Рассмотрим преобразование:

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

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 как $\bar{\boldsymbol{\upsilon}}$.

Теорема 2. Пусть выполняются следующие условия:

1) *апостериорное распределение параметров* $p\big(\mathbf{u}|\mathfrak{D}\big)$ *модели учителя является нормальным распределением* (6)*;*

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

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

Тогда апостериорное распределения также описывается нормальным распределением со следующей плотностью распределения:

$$p\big(\boldsymbol{\upsilon}|\mathfrak{D}\big) = \mathcal{N}\big(\mathbf{m}_{\boldsymbol{\upsilon}} + \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}^{-1}_{\bar{\boldsymbol{\upsilon}},\bar{\boldsymbol{\upsilon}}}\left(\mathbf{i} - \bar{\boldsymbol{\upsilon}}\right), \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}^{-1}_{\bar{\boldsymbol{\upsilon}},\bar{\boldsymbol{\upsilon}}}\boldsymbol{\Sigma}_{\boldsymbol{\upsilon},\bar{\boldsymbol{\upsilon}}}\big), \tag{8}$$

где вектор $\mathbf{i}$ *задается следующим образом:*

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

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 теоремы. Заметим, что если заметить t-й слой нейронной сети с $T+1$ слоем, то он будет эквивалентным архитектуре с T слоями:

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

Получаем, что удаление t-го слоя нейросети эквивалентно приравнивают 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 распределения также является нормальным распределением с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

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

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

Теорема 2 задает апостериорное распределение параметров (8) 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 согласовывают пространства параметров учителя f и ученика g. После сопоставления параметрических моделей получаем, что параметры модели учителя и модели учителя принадлежат одному семейству 3.1.

4. Вычислительный эксперимент

Проводится вычислительный эксперимент для анализа предложенного метода дистилляции на основе апостериорного распределения параметров модели учителя.

4.1. Синтетические данные

Проанализируем модель на синтетической выборке. Выборка построенная следующим образом:

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

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 (3). 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 (5), 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:

$$g = \boldsymbol{\sigma} \circ \mathbf{W}_3\boldsymbol{\sigma} \circ \mathbf{W}_2\boldsymbol{\sigma} \circ \mathbf{W}_1, \tag{9}$$

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.

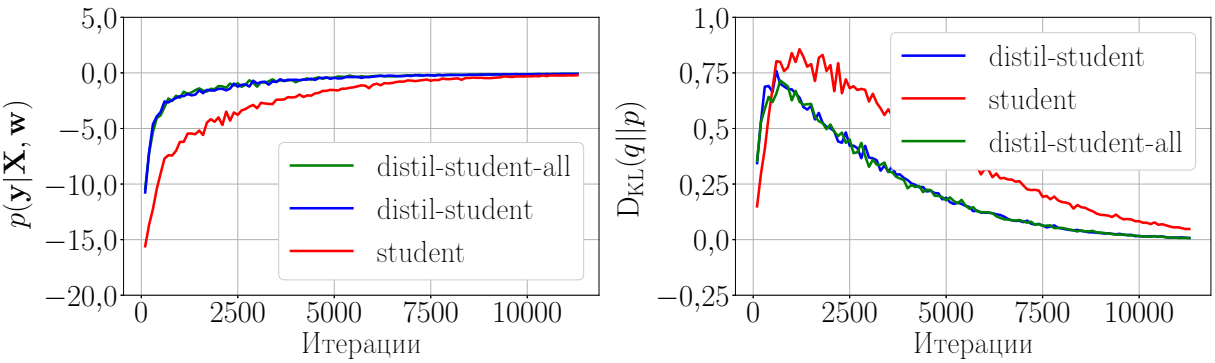

Рис. 1. Структура (9) модели ученика g. Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL–дивергенция между вариационным и априорным распределениями параметров модели.

Рис. 1 сравнивает модели ученика, со структурой (9). Представлено сравнение разных моделей: модель без дистилляции, где в качестве априорного распределения выбирается стандартное нормальное распределение (на легенде обозначается student); модель с частичной дистилляцией, где в качестве среднего значения параметров выбираются параметры согласно выражения (7), а ковариационная матрица была приравнена к единичной матрицы (на легенде обозначается distil-student); модель с полной дистилляцией согласно выражения (7) (на легенде обозначается distil-student-all). Видно, что модели ученика, где в качестве априорного распределения выбраны распределения основанные на апостериорном распределение учителя имеют больше правдоподобие, чем модель где в качестве априорного распределения выбрано стандартное нормальное. Также заметим, что использования параметра среднего из апостериорного распределения дает основной вклад при дистилляции, так как качество моделей distil-student и distil-student-all совпадает.

Вторая конфигурация получается путем удаления слоя модели учителя:

$$g = \boldsymbol{\sigma} \circ \mathbf{W}_2 \boldsymbol{\sigma} \circ \mathbf{W}_1, \tag{10}$$

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

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

В качестве функции активации была выбрана функция активации ReLu.

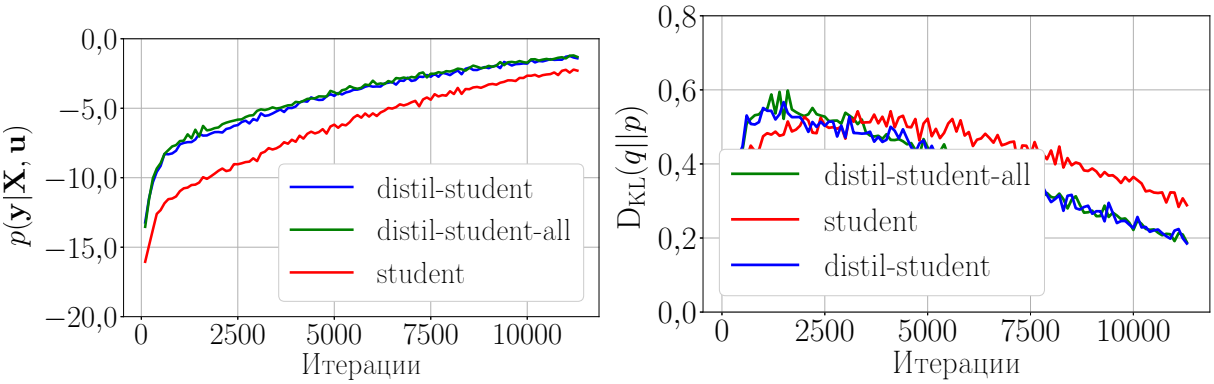

Рис. 2. Структура (10) модели ученика g. Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL–дивергенция между вариационным и априорным распределениями параметров модели.

Рис. 2 сравнивает модели ученика со структурой (10). Аналогично рис. 1 на рис. 2 представлено сравнение модели без дистилляции (student), модели с дистилляцией параметра среднего значение (distil-student), модели с полной дистилляцией (distil-student-all). 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 (3). 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 (5), где в качестве априорного распределения параметров выбрано стандартное нормальное распределение.

Таблица 2. Сводная таблица результатов вычислительного эксперимента.

	teacher	student	distil-student	distil-student-all
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	-	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

В качестве модели ученика были выбрана конфигурация с одним скрытым слоем (10), 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.

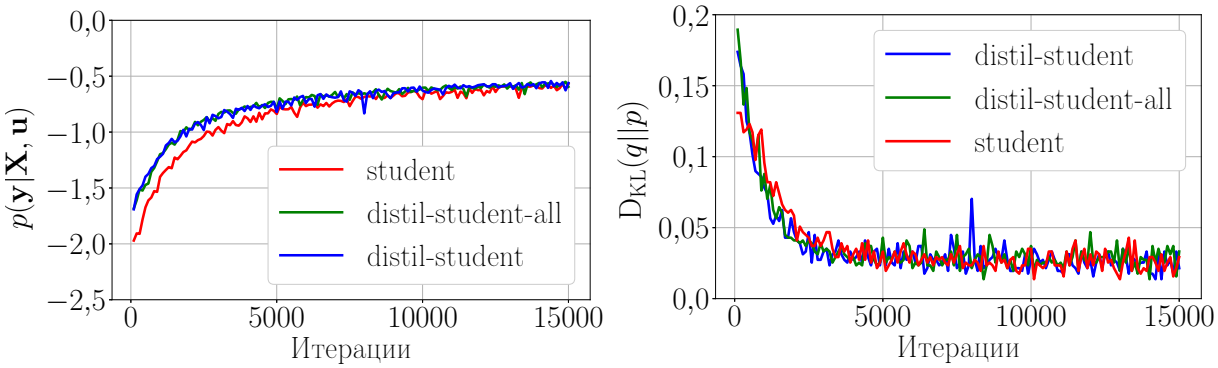

Рис. 3. Слева: правдоподобие выборки в зависимости от номера итерации при обучении. Справа: KL–дивергенция между вариационным и априорным распределениями параметров модели.

Рис. 3 сравнивает модели ученика с разными априорными распределениями на параметры. Аналогично синтетическому эксперименту сравнивался случай стандартного нормального распределения, использования априорного распределения с заданным средним значением параметров на основе апостериорного распределения, с определением полного априорного распределения на основе формулы (8). 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 student и моделями distil-student и distil-student-all соответсвенно:

$$S = \sum_{k}^{K} p\big(\mathbf{y}|\mathbf{X},\mathbf{u}^{\mathrm{s}}\big) - p\big(\mathbf{y}|\mathbf{X},\mathbf{u}^{\mathrm{ds}}\big),$$

где $\mathbf{u}^{\mathrm{s}}, \mathbf{u}^{\mathrm{ds}}$ обозначает параметры модели студента и модели дистиллированного студента соответсвенно. Заметим, что площадь S имеет знак: чем большее положительное число, тем дистиллированная модель лучше, чем модель построенная без учителя. В случае, 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, то есть модели ученика полученные при помощи дистилляции являются лучше чем модель ученика без дистилляции.

5. Заключение

В данной работе проанализирована байесовская дистиляция модели учителя в модель ученика на основе вариационного вывода. В рамках данной работы 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 сопоставления параметров модели учителя и ученика в случае, если число слоев совпадает, но размер слоев различается. Теорема 2 описывает механизм сопоставления параметров модели учителя и 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 сопоставления. Краткое описания эксперимента представлено в табл. 2.

СПИСОК ЛИТЕРАТУРЫ

1. *Alex Krizhevsky and Vinod Nair and Geoffrey Hinton* CIFAR-10 (Canadian Institute for Advanced Research) // `http://www.cs.toronto.edu/~kriz/cifar.html`

2. *Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.* Imagenet: A large-scale hierarchical image database // IEEE conference on computer vision and pattern recognition, 2009. P. 248–255.

3. *Huang, Zehao and Wang, Naiyan* Like What You Like: Knowledge Distill via Neuron Selectivity Transfer // arXiv e-prints, 2017.

4. *Kui Ren and Tianhang Zheng and Zhan Qin and Xue Liu* Adversarial Attacks and Defenses in Deep Learning // Engineering, 2020. P. 346–360.

5. *Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton* ImageNet Classification with Depp Convolutional Neural Networks // NIPS, 2012.

6. *Karen Simonyan and Andrew Zisserman* Very Deep Convolutional Networks for Large-Scale Image Recognition // NIPS, 2014.

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

8. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprinted, 2018.

9. *Tom B. Brown et al* GPT3: Language Models are Few-Shot Learners // arXiv preprinted, 2020.

10. *Linting Xue and Noah Constant and Adam Roberts and Mihir Kale and Rami Al-Rfou and Aditya Siddhant and Aditya Barua and Colin Raffel.* mT5: A massively multilingual pre-trained text-to-text transformer // arXiv preprinted, 2021.

11. *Yang, Ziqing and Cui, Yiming and Chen, Zhipeng and Che, Wanxiang and Liu, Ting and Wang, Shijin and Hu, Guoping* TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. P. 9–16.

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

13. *Бахтеев О. Ю., Стрижов В. В.* Выбор моделей глубокого обучения субоптимальной сложности // АиТ. 2018. № 8. С. 129–147.

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

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

16. *Vapnik V., Izmailov R.* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // Journal of Machine Learning Research. 2015. No 16. P. 2023–2049.

17. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // In International Conference on Learning Representations. Puerto Rico, 2016.

18. *Madala H., Ivakhnenko A.* Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc., 1994.

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

20. *Wilson T., Kozareva Z., Nakov P., Rosenthal S., Stoyanov V., Ritter A.* SemEval-2013 Task 2: Sentiment Analysis in Twitter // Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, 2013. P. 312–320.

21. *LeCun Y., Boser B., Denker J., Henderson D., Howard R., Hubbard W., Jackel L.* Backpropagation Applied to Handwritten Zip Code Recognition // Neural Computation. 1989. V. 1. No 4. P. 541–551.

22. *Hochreiter S., Schmidhuber J.* Long short-term memory // Neural Computation. 1997. V. 9. No 8. P. 1735–1780.

23. *Kingma D, Ba J.* Adam: A Method for Stochastic Optimization // arXiv preprint arXiv:1412.6980. 2014.

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

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

26. *Vapnik V., Izmailov R.* Learning Using Privileged Information: Similarity Control and Knowledge Transfer // Journal of Machine Learning Research. 2015. No 16. P. 2023–2049.

27. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // In International Conference on Learning Representations. Puerto Rico, 2016.