Дистилляция моделей глубокого обучения

Грабовой Андрей Валериевич

Московский физико-технический институт

03.04.01 Прикладные математика и физика

Научный руководитель д.ф.-м.н. В. В. Стрижов

МФТИ, г. Долгопрудный
2021 г.

Байесовский вывод в задаче дистилляции

Исследуемая проблема

Снижения числа обучаемых параметров моделей машинного обучения для получения интерпретируемых моделей машинного обучения.

Задачи

1. Поставить вероятностную задачу дистилляции для задач классификации и регрессии.
2. Предложить метод байесовской дистилляции нейросетевых моделей.
3. Провести теоретический анализ полученных результатов.

Метод

Предлагается байесовская постановка задачи дистилляции моделей глубокого обучения. Предлагается метод дистилляции моделей в котором априорное распределение параметров модели *ученика* является функцией параметров *учителя*. Предлагается метод приведения структуры модели учителя к модели ученика, что позволяется использовать апостериорное распределения параметров учителя в качестве априорного распределения параметров ученика.

Оптимизация парам. ученика на основе парам. учителя

Заданы:

1) признаки $\mathbf{x}_i \in \mathbb{R}^n$;
2) привилегированные признаки $\mathbf{x}_i^* \in \mathbb{R}^{n^*}$;
3) целевая переменная $y_i \in \mathbb{Y}$;
4) индексы объектов, для которых известна привилегированная информация $\mathcal{I}$, а для которых она не известна $\bar{\mathcal{I}}$.

Функции учителя $\mathbf{f} : \mathbb{R}^{n^*} \to \mathbb{Y}'$ и ученика $\mathbf{g} : \mathbb{R}^n \to \mathbb{Y}'$ — пространство оценок. Ответ $\mathbf{s}_i = \mathbf{f}\left(\mathbf{x}_i^*\right)$ функции $\mathbf{f}$ для объекта $\mathbf{x}_i^*$ используется при решении оптимизационной задачи.

Требуется выбрать модель ученика $\mathbf{g}$ из множества

$$\mathfrak{G} = \left\{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}'\right\}.$$

Оптимизационная задача:

$$\mathbf{g} = \arg\min_{\mathbf{g} \in \mathfrak{G}} \mathcal{L}\left(\mathbf{g}, \mathbf{f}, \mathbf{X}, \mathbf{X}^*, \mathbf{y}\right),$$

где $\mathcal{L}$ функция ошибки.

Постановка задачи: дистилляция Хинтона[1]; Вапника[2]

Заданы:

1) $\mathbf{x}_i^* = \mathbf{x}_i$; $\mathbf{x}_i^* \neq \mathbf{x}_i$ для всех $i \in \{1, 2, \cdots, m\}$;
2) $y_i \in \mathbb{Y} = \{1, \cdots, K\}, \quad \mathbb{Y}' = \mathbb{R}^K$.

Параметрические семейства учителя и ученика:

$$\mathfrak{F}^*{}_{\text{cl}} = \left\{\mathbf{f} | \mathbf{f} = \text{softmax}\left(\mathbf{v}^*\left(\mathbf{x}^*\right)/T\right), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}^K\right\},$$

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

где $\mathbf{z}, \mathbf{v}^*$ — это дифференцируемые параметрические функции заданной структуры, T — параметр температуры.

Функция ошибки

$$\mathcal{L}_{\text{st}}\left(\mathbf{g}\right) = -\sum_{i=1}^{m} \underbrace{\sum_{k=1}^{K} y_i^k \log \mathbf{g}\left(\mathbf{x}_i\right)\big|_{T=1}}_{\text{исходная функция потерь}} - \sum_{i=1}^{m} \underbrace{\sum_{k=1}^{K} \mathbf{f}\left(\mathbf{x}^*{}_i\right)\big|_{T=T_0} \log \mathbf{g}\left(\mathbf{x}_i\right)\big|_{T=T_0}}_{\text{слагаемое дистилляция}},$$

где $\cdot\big|_{T=t}$ параметр температуры T равняется t.

Оптимизационная задача:

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathfrak{G}_{\text{cl}}} \mathcal{L}_{\text{st}}\left(\mathbf{g}\right).$$

---

[1] *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS, 2015.

[2] *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information //ICLR, 2016.

Вероятностная постановка задачи дистилляции

Гипотеза порождения данных:

1) задано распределение целевой переменной $p(y_i|\mathbf{x}_i, \mathbf{g})$;
2) задано совместное распределение $p(y_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g})$;
3) для всех $i \in \mathcal{I}$ элементы y_i и $\mathbf{s}_i$ являются зависимыми величинами;
4) если $|\mathcal{I}| = 0$ то решение равно решению максимума правдоподобия.

Совместное правдоподобие истинных меток и меток учителя:

$$p(\mathbf{y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i \notin \mathcal{I}} p(y_i|\mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(y_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}).$$

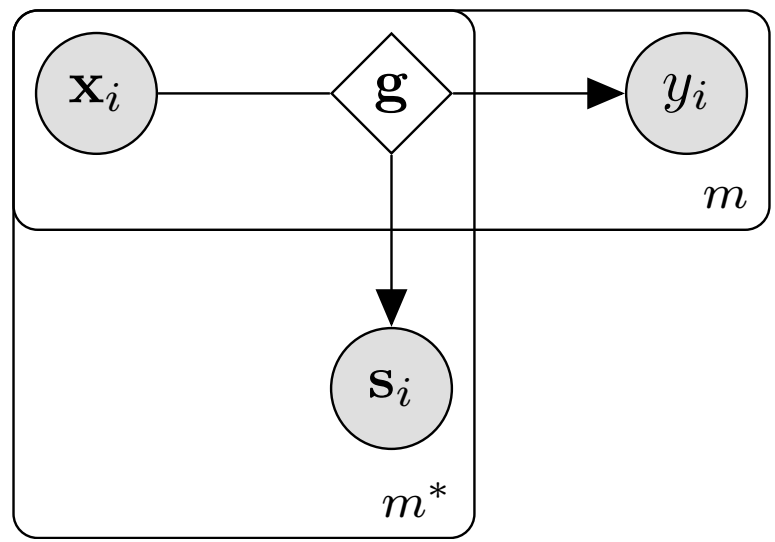
Задача оптимизации:

$$\mathbf{g} = \arg\max_{\mathbf{g} \in \mathfrak{G}} p(\mathbf{y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}),$$

имеет вид:

$$\sum_{i \notin \mathcal{I}} \log p(y_i|\mathbf{x}_i, \mathbf{g}) + (1 - \lambda) \sum_{i \in \mathcal{I}} \log p(y_i|\mathbf{x}_i, \mathbf{g}) + \lambda \sum_{i \in \mathcal{I}} \log p(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}),$$

где $\lambda \in [0, 1]$ — метапараметр.



Байесовская постановка задачи дистилляции

Задана модель учителя с фиксированными параметрами в виде суперпозиции отображений:

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

где $\mathbf{U}$ матрицы линейны отображений, $\boldsymbol{\sigma}$ нелинейность. Вектор параметров модели учителя:

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

На основе выборки $\{\mathbf{x}_i, y_i\}_{i=1}^m$ и учителя f требуется выбрать модель ученика:

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

где $\mathbf{W}$, $\boldsymbol{\sigma}$, $\mathbf{w}$ вводятся аналогично учителю. Выбор модели g эквивалентный оптимизации вектора параметров $\mathbf{w}$. Параметры $\mathbf{w}$ оптимизируются на основе вариационного вывода:

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

Априорное распределение $p\big(\mathbf{w}|\mathbf{A}\big)$ задается как функция от апостериорного распределения $p\big(\mathbf{u}|\mathbf{X}, \mathbf{y}\big)$. Апостериорное распределение параметров модели учителя задается нормальным распределением:

$$p\big(\mathbf{u}|\mathbf{X}, \mathbf{y}\big) = \mathcal{N}\big(\mathbf{u}_0, \boldsymbol{\Sigma}_0\big).$$

Проблема: пространства параметров учителя и ученика не совпадает.

Задача классификации: вероятностная постановка

Заданы:

1) учитель $\mathbf{f} \in \mathfrak{F}^*_{\text{cl}}$ и ученик $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$;
2) распределение истинных меток $p\big(y|\mathbf{x},\mathbf{g}\big) = \text{Cat}\big(\mathbf{g}\big(\mathbf{x}\big)\big)$;
3) распределение ответов учителя $p\big(\mathbf{s}|\mathbf{x},\mathbf{g}\big) = C\prod_{k=1}^{K} g_k\big(\mathbf{x}\big)^{s^k}, \quad C < \infty.$

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathfrak{G}} \sum_{i\notin\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\big(\mathbf{x}_i\big)\big|_{T=1} + (1-\lambda)\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\big(\mathbf{x}_i\big)\big|_{T=1}$$

$$+\lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} s_{i,k} \log g_k\big(\mathbf{x}_i\big)\big|_{T=T_0} + \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\left(\log g_k\big(\mathbf{x}_i\big)\big|_{T=T_0} + \log\log\frac{1}{g_k\big(\mathbf{x}_i\big)}\big|_{T=T_0}\right)$$

Теорема (Грабовой, 2020)

Пусть всех k выполняется $1 > 1-\varepsilon > g_k\big(\mathbf{x}\big) > \varepsilon > 0$, *тогда при*

$$C = (-1)^K \frac{K^{K/2}}{2^{K(K-1)/2}} \prod_{k=1}^{K} g_k\big(\mathbf{x}\big)\log g_k\big(\mathbf{x}\big)$$

функция $p\big(\mathbf{s}|\mathbf{x},\mathbf{g}\big) = C\prod_{k=1}^{K} g_k\big(\mathbf{x}\big)^{s^k}$ *является плотностью распределения.*

Задача регрессии: вероятностная постановка

1) учитель $f \in \mathfrak{F}^*_{\text{rg}} = \left\{f | f = \mathbf{v}^*\left(\mathbf{x}^*\right), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}\right\}$;
2) ученик $g \in \mathfrak{G}_{\text{rg}} = \left\{g | g = \mathbf{z}\left(\mathbf{x}\right), \quad \mathbf{z} : \mathbb{R}^{n} \to \mathbb{R}\right\}$;
3) распределение истинных меток $p\left(y|\mathbf{x}, g\right) = \mathcal{N}\left(y|g\left(x\right), \sigma\right)$;
4) распределения меток учителя $p\left(s|\mathbf{x}, g\right) = \mathcal{N}\left(s|g\left(\mathbf{x}\right), \sigma_s\right)$.

Оптимизационная задача:

$$\hat{g} = \arg\min_{g \in \mathfrak{G}} \sum_{i \notin \mathcal{I}} \sigma^2 \left(y_i - g\left(\mathbf{x}_i\right)\right)^2 + (1-\lambda) \sum_{i \in \mathcal{I}} \sigma^2 \left(y_i - g\left(\mathbf{x}_i\right)\right)^2 + \lambda \sum_{i \in \mathcal{I}} \sigma_s^2 \left(s_i - g\left(\mathbf{x}_i\right)\right)^2.$$

Теорема (Грабовой, 2020)

Пусть $\mathfrak{G}_{rg}$ — класс линейных функций $g\left(\mathbf{x}\right) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$. Тогда решение оптимизационной задачи эквивалентно решению задачи линейной регрессии $\mathbf{y}'' = \mathbf{X}\mathbf{w} + \varepsilon,\ \varepsilon \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right)$, где $\mathbf{\Sigma}^{-1} = diag\left(\boldsymbol{\sigma}'\right)$ и $\mathbf{y}''$ имеют следующий вид:

$$\sigma'_i = \begin{cases} \sigma^2, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\,\sigma^2 + \lambda\sigma_s^2, \text{ иначе,} \end{cases} \quad \mathbf{y}'' = \mathbf{\Sigma}\mathbf{y}', \quad y'_i = \begin{cases} \sigma^2 y_i, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\,\sigma^2 y_i + \lambda\sigma_s^2 s_i, \text{ иначе.} \end{cases}$$

Байесовкая дистилляция: приведение парам. моделей

Определение

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

Определение

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

Пространства параметров совпадают

- число слоев совпадает $L = T$;
- размеры соответствующих слоев совпадают,

тогда $p\big(\mathbf{w}|\mathbf{A}\big) = p\big(\mathbf{w}|\mathfrak{D}\big)$.

Пространства параметров не совпадают В данном случае дистиляция проходит в два этапа:

- выполняется приведение модели учителя к ученику;
- апостериорное распределение учителя назначается априорным распределением ученика.

Размеры скрытых слоев учителя и ученика отличаются

- число слоев совпадает $L = T$;
- размеры соответствующих слоев не совпадают: $n_l \leq n_t$, где n_t для всех $l \in \{1, \cdots, L\}, t \in \{1, \cdots, T\}$.

Преобразования t-го слоя учителя:

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

описывает удаление одного нейрона из t-го слоя. Новый вектор параметров обозначим $\mathbf{u}' = \phi\big(t, \mathbf{u}\big)$, а выброшенные параметры обозначим $\mathbf{u}''$

Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

- $p\big(\mathbf{u}|\mathbf{X}, \mathbf{y}\big) = \mathcal{N}\big(\mathbf{u}_0, \mathbf{\Sigma}_0\big)$;
- *число слоев совпадает* $L = T$;
- *для всех* t, l *таких, что* $t = l$ *выполняется* $n_s \leq n_l$,.

Тогда $p\big(\mathbf{u}'|\mathbf{X}, \mathbf{y}\big)$ *является нормальным распределением:*

$$p\big(\mathbf{u}'|\mathbf{X}, \mathbf{y}\big) = \mathcal{N}\big(\mathbf{u}_0' + \mathbf{\Sigma}_0^{'''} \mathbf{\Sigma}_0^{''-1} \left(\mathbf{0} - \mathbf{u}_0^{''}\right), \mathbf{\Sigma}_0^{'} - \mathbf{\Sigma}_0^{'''} \mathbf{\Sigma}_0^{''-1} \mathbf{\Sigma}_0^{'''}\big).$$

Число скрытых слоев учителя и ученика отличаются

- соответствующие размеры слоев совпадают, $n_t = n_{t-1}$;
- функция активации удовлетворяет свойству $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$.

Преобразования t-го слоя учителя:

$$\psi\big(t\big) : \mathbb{R}^{\mathsf{p_{tr}}} \to \mathbb{R}^{\mathsf{p_{tr}} - n_t n_{t-1}}$$

описывает удаление t-го слоя. Новый вектор параметров обозначим $\mathbf{u}' = \phi\big(t, \mathbf{u}\big)$, а выброшенные параметры обозначим $\mathbf{u}''$

Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

- $p\big(\mathbf{u}|\mathbf{X}, \mathbf{y}\big) = \mathcal{N}\big(\mathbf{u}_0, \boldsymbol{\Sigma}_0\big)$;
- *соответствующие размеры слоев совпадают,* $n_t = n_{t-1}$;
- *функция активации удовлетворяет свойству* $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$.

Тогда $p\big(\mathbf{u}'|\mathbf{X}, \mathbf{y}\big)$ *является нормальным распределением:*

$$p\big(\mathbf{u}'|\mathbf{X}, \mathbf{y}\big) = \mathcal{N}\Big(\mathbf{u}_0' + \boldsymbol{\Sigma}_0^{'}{}^{,''} \boldsymbol{\Sigma}_0^{''-1} \Big(\mathbf{i} - \mathbf{u}_0^{''}\Big), \boldsymbol{\Sigma}_0^{'} - \boldsymbol{\Sigma}_0^{'}{}^{,''} \boldsymbol{\Sigma}_0^{''-1} \boldsymbol{\Sigma}_0^{'}{}^{,''}\Big),$$

где $\mathbf{i} = \mathrm{vec}\big(\mathbf{I}\big)$

Порядок на множестве параметров

Задание порядка на множестве параметров:

- случайным образом (используется в рамках вычислительного эксперимента)
- на основе метода оптимального прореживания нейросети:

$$\xi = \arg\min_j h_{jj} \frac{u_j^2}{2},$$

где h_{jj} коэффициент при квадратичном члене в разложении Тейлора функции ошибки по параметрам модели.

- на основе отношения плотности апостериорного распределения параметра к плотности апостериорного распределения параметра к нулю:

$$\xi = \arg\max_j \frac{p\big(0|\mathbf{X},\mathbf{t}\big)}{p\big(u_j|\mathbf{X},\mathbf{t}\big)},$$

- выбор на основе анализа мультиколиниарности параметров методов Белсли:

$$\xi = \arg\max_j \frac{\lambda_{\max}}{\lambda_j},$$

где λ являются сингулярными числами ковариационный матрицы параметров.

Вычислительный эксперимент: вероятностная дистиляция

Выборка FashionMNIST:
Изображения размера 28×28. Решается задача классификации с $K = 10$ классами.
Объем выборки $m_{\text{train}} = 60000$ и $m_{\text{test}} = 10000$ объектов.

Синтетическая выборка:

$$\mathbf{X} = \left[\mathcal{N}\left(x_{ij}|0,1\right)\right]_{m\times n}, \qquad \mathbf{W} = \left[\mathcal{N}\left(w_{jk}|0,1\right)\right]_{n\times K},$$

$$\mathbf{S} = \text{softmax}\left(\mathbf{XW}\right), \qquad \mathbf{y} = \left[\text{Cat}\left(y_i|\mathbf{s}_i\right)\right],$$

где функция softmax берется построчно.
Число признаков $n = 10$, число классов $K = 3$, объем выборки $m_{\text{train}} = 1000$ и $m_{\text{test}} = 100$ объектов.

Выборка Twitter Sentiment Analysis:
Англоязычные твиты пользователей. Решается задача бинарной классификации текстовых сообщений.
Объем выборки $m_{\text{train}} = 1{,}18$млн и $m_{\text{test}} = 0{,}35$млн объектов.

Сводная таблица вычислительного эксперимента: вероятностная дистиляция

Dataset	Model	CrossEntropyLoss	Accuracy	StudentSize
FashionMnist	without teacher	$0{,}461 \pm 0{,}005$	$0{,}841 \pm 0{,}002$	7850
	with teacher	$\mathbf{0{,}453 \pm 0{,}003}$	$0{,}842 \pm 0{,}002$	7850
Synthetic	without teacher	$0{,}225 \pm 0{,}002$	$0{,}831 \pm 0{,}002$	33
	with teacher	$\mathbf{0{,}452 \pm 0{,}001}$	$0{,}828 \pm 0{,}001$	33
Twitter	without teacher	$0{,}501 \pm 0{,}006$	$0{,}747 \pm 0{,}005$	1538
	with teacher	$\mathbf{0{,}489 \pm 0{,}003}$	$\mathbf{0{,}764 \pm 0{,}004}$	1538

В таблице показаны результаты вычислительного эксперимента для разных выборок. Точность аппроксимации выборки учеником улучшается при использовании модели учителя при обучении.

Вычислительный эксперимент: байесовская дистилляция

Синтетическая выборка:

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

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

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

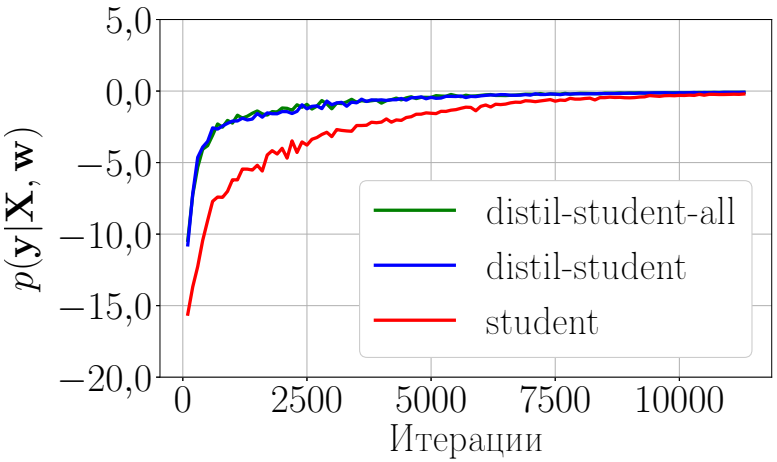
Первая конфигурация модели ученика:

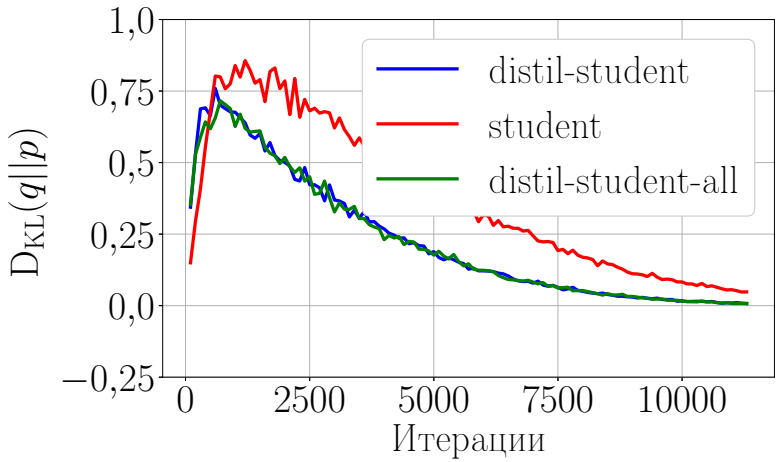
$$g = \boldsymbol{\sigma} \circ \mathbf{W}_3 \circ \boldsymbol{\sigma} \circ \mathbf{W}_2 \circ \boldsymbol{\sigma} \circ \mathbf{W}_1, \quad \mathbf{W}_1 \in \mathbb{R}^{10\times 10}, \mathbf{W}_2 \in \mathbb{R}^{10\times 10}, \mathbf{W}_3 \in \mathbb{R}^{1\times 10}.$$

Вторая конфигурация модели ученика:

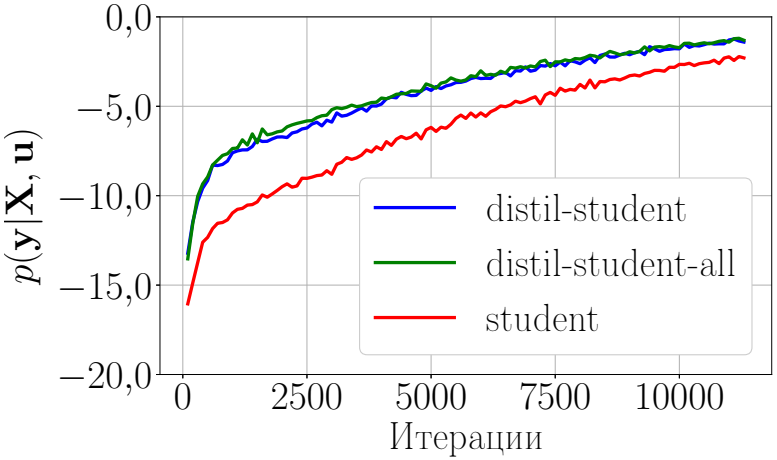
$$g = \boldsymbol{\sigma} \circ \mathbf{W}_2 \circ \boldsymbol{\sigma} \circ \mathbf{W}_1, \quad \mathbf{W}_1 \in \mathbb{R}^{1\times 50}, \mathbf{W}_2 \in \mathbb{R}^{50\times 10}.$$

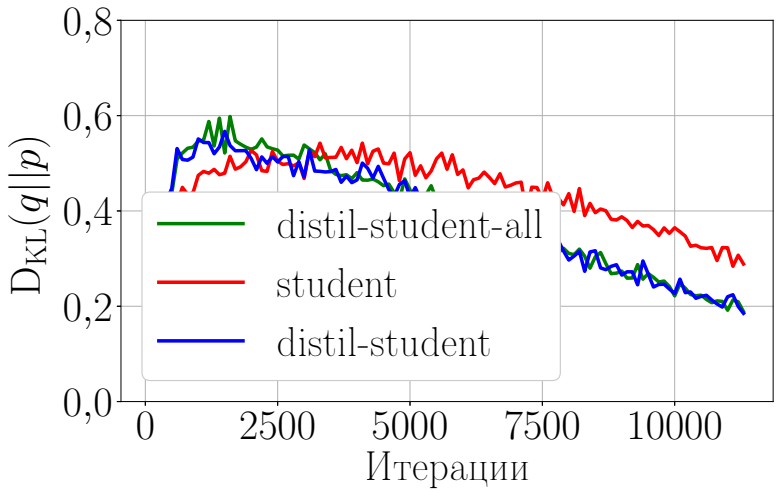
Результаты эксперимента: байесовская дистилляция





Первая конфигурация модели ученика. Дистиллированная модель имеет большее правдоподобие.





Вторая конфигурация модели ученика. Дистиллированная модель имеет большее правдоподобие.

Сводная таблица вычислительного эксперимента: байесовская дистилляция

	teacher	student	distil-student	distil-student-all
Эксперимент на синтетической выборке (удаление нейрона)				
Архитектура	$[10, 100, 50, 1]$	$[10, 10, 10, 1]$	$[10, 10, 10, 1]$	$[10, 10, 10, 1]$
Число параметров	6050	210	210	210
Разность площадей	-	0	**16559**	**16864**
Эксперимент на синтетической выборке (удаление слоя)				
Архитектура	$[10, 100, 50, 1]$	$[10, 50, 1]$	$[10, 50, 1]$	$[10, 50, 1]$
Число параметро	6050	550	550	550
Разность площадей	-	0	**23310**	**25506**
Эксперимент на выборке FashionMnist				
Архитектура	$[784, 800, 50, 10]$	$[784, 50, 10]$	$[784, 50, 10]$	$[784, 50, 10]$
Число параметро	667700	39700	39700	39700
Разность площадей	-	0	**1165**	**1145**

Для численного сравнения качества моделей выбрана разность площадей графика $p\big(\mathbf{y}|\mathbf{X}, \mathbf{w}\big)$. Чем больше значение тем лучше.

Выносится на защиту

1. Проведен вероятностный анализ задачи дистилляции.
2. Выполнена обобщение классического подхода введя вероятностные предположения о природе данных.
3. Теоретические анализы сформулированы в виде теорем для задачи классификации и регрессии.
4. Поставлена задача байесовской дистилляции моделей глубокого обучения.
5. Предложен метод задания априорного распределения параметров модели ученика на основе апостериорного распределения парамтеров учителя.
6. Доказаны теоремы, которые позволяют приводить структуры модели учителя к структуре модели ученика.
7. Проведен ряд вычислительных экспериментов, которые показывают применимость предложенных методов.

Публикации ВАК по теме

1. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019, 13(2).
2. *Грабовой А.В., Бахтеев О. Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020, 14(2).
3. *A. Grabovoy, V. Strijov.* Quasi-periodic time series clustering for human. Lobachevskii Journal of Mathematics, 2020, 41(3).
4. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал Вычислительной математики и математической физики, 2021. 61(5).
5. *Грабовой А.В., Стрижов В.В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и телемеханика, 2021 (на рецензировании).
6. *T. Gadaev, A. Grabovoy, A. Motrenko, V. Strijov* Numerical methods of minimum sufficient sample size estimation for linear models // подано.
7. *Базарова А.И., Грабовой А.В., Стрижов В.В.* Анализ свойств вероятностных моделей в задачах обучения с экспертом // подано.
8. *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика, 2021 (на рецензировании).