Априорное распределение параметров в задачах выбора моделей глубокого обучения

А. В. Грабовой

Диссертация на соискание ученой степени
кандидата физико-математических наук
05.13.17 — Теоретические основы информатики
Научный руководитель: д.ф.-м.н. В. В. Стрижов

17 февраля 2022 г.

Априорное распределение параметров моделей

Исследуется проблема выбора моделей глубокого обучения. Для выбора моделей используется связный байесовский вывод. Для снижения размерности пространства параметров при выборе модели используется информация об их априорном и апостериорном распределениях.

Цель исследования —

предложить метод задания априорного распределения параметров модели глубокого обучения с учетом накопленной информации о решаемой задаче.

Требуется предложить

1) метод выбора модели, основанный на использовании привилегированной и накопленной информации,
2) метод выравнивания структур параметрических моделей,
3) метод снижения размерности пространства параметров моделей.

Решение

Для назначения априорного распределения параметров при выборе моделей глубокого обучения используются апостериорные распределения параметров ранее полученных моделей.

Привилегированное обучение В. Н. Вапника[1] и дистилляция Дж. Хинтона[2]

Заданы

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

Модели учителя $\mathbf{f} : \mathbb{R}^{n^*} \to \mathbb{Y}'$ и ученика $\mathbf{g} : \mathbb{R}^n \to \mathbb{Y}'$ — пространство оценок.
Ответы $\mathbf{s}_i = \mathbf{f}\left(\mathbf{x}_i^*\right)$ функции $\mathbf{f}$ для объектов $\mathbf{x}_i^*$.

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

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

Оптимизационная задача:

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

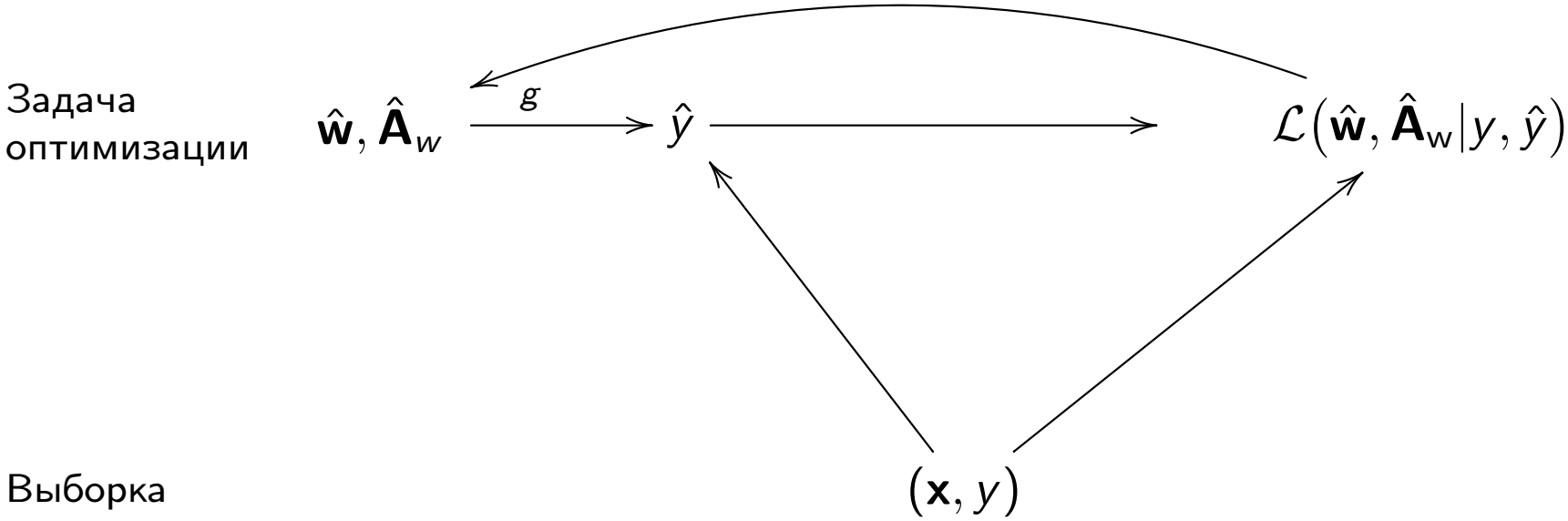
где $\mathcal{L}$ — функция ошибки.

---

[1] *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying distillation and privileged information // ICLR, 2016.

[2] *Hinton G., Vinyals O., Dean J.* Distilling the knowledge in a neural network // NIPS, 2015.

Задача оптимизации параметров модели

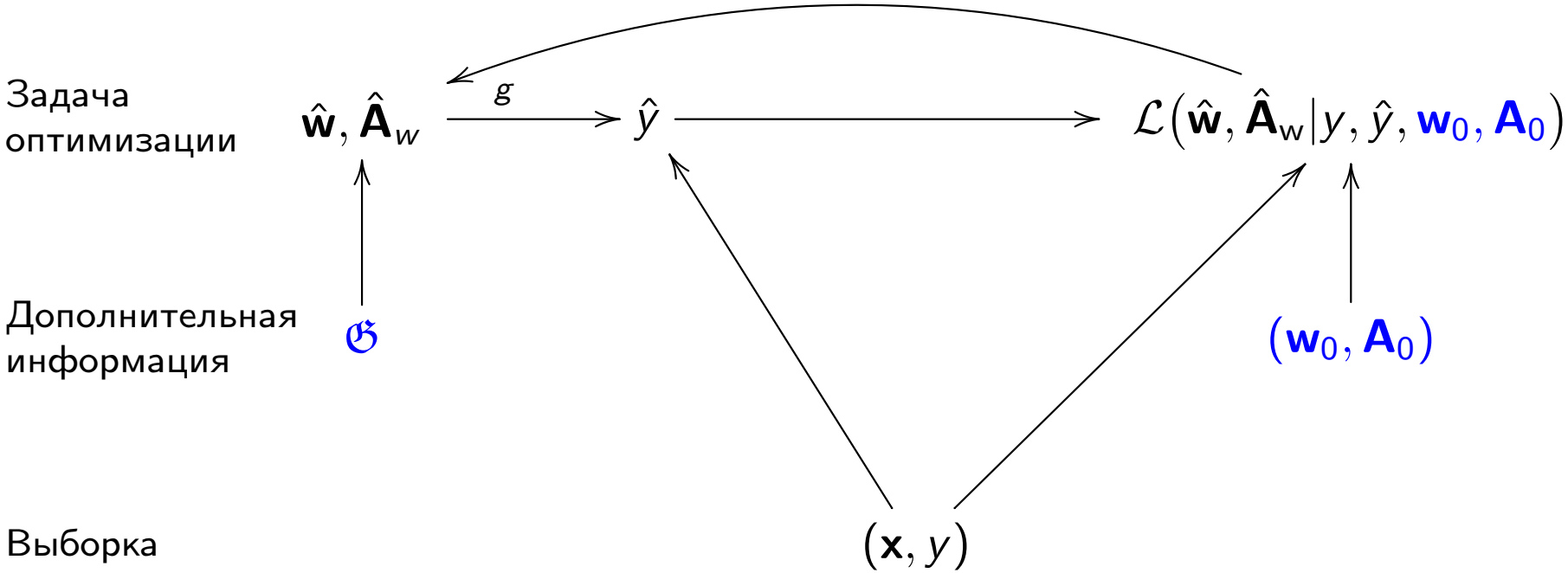


Выборка $\mathfrak{D} \ni (\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^n, y \in \mathbb{Y}$
Модели $\mathfrak{G} \ni \mathbf{g} : \mathbb{R}^n \times \mathbb{R}^{n_w} \to \mathbb{Y}'$
Прогноз $\hat{y} = g(\mathbf{x}, \hat{\mathbf{w}})$

Оптимизационная задача

$$\hat{\mathbf{w}}, \hat{\mathbf{A}}_w = \underset{\mathbf{w} \in \mathbb{R}^{n_w}, \mathbf{A} \in \mathbb{R}^{n_w \times n_w}}{\arg\min} \mathcal{L}(\mathbf{w}, \mathbf{A}_w | \mathfrak{D}, \mathbf{g})$$

Байесовский выбор модели



Выборка $\mathfrak{D} \ni (\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^n, y \in \mathbb{Y}$

Модели $\mathfrak{G} \ni \mathbf{g} : \mathbb{R}^n \times \mathbb{R}^{n_w} \to \mathbb{Y}'$

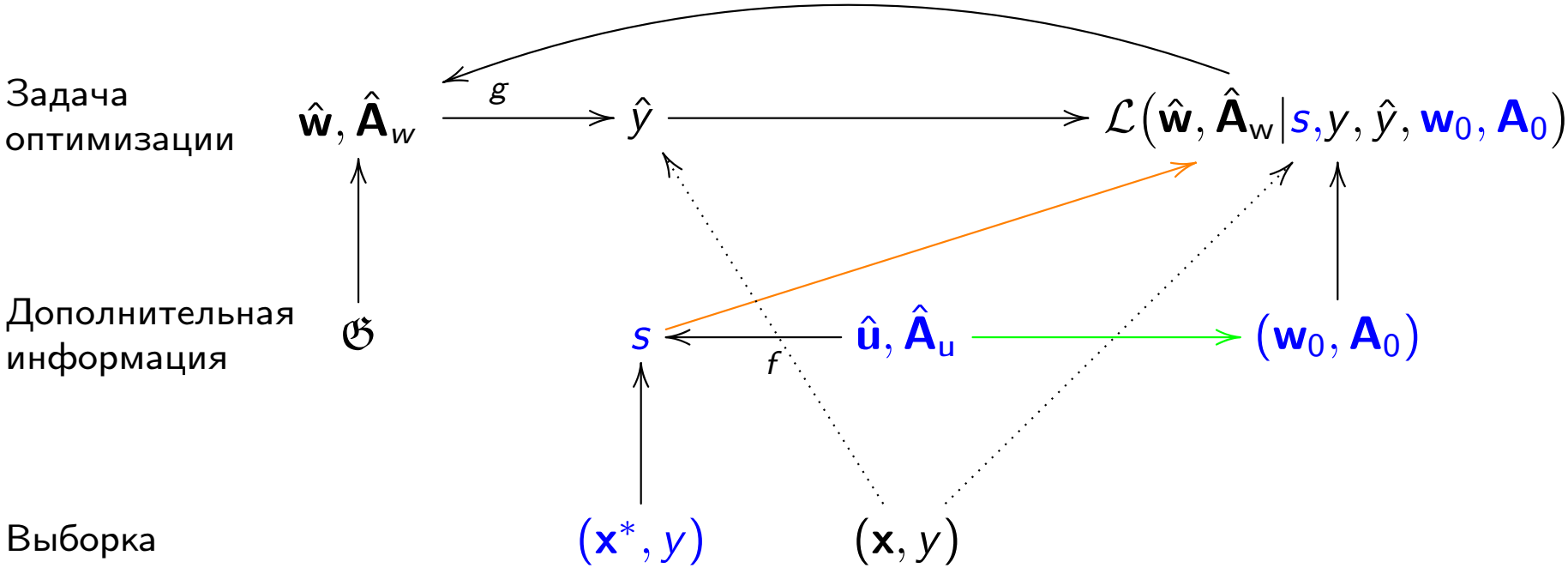
Прогноз $\hat{y} = g(\mathbf{x}, \hat{\mathbf{w}})$

Априорное распределение $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \mathbf{A}_0)$

Оптимизационная задача

$$\hat{\mathbf{w}}, \hat{\mathbf{A}}_w = \underset{\mathbf{w} \in \mathbb{R}^{n_w}, \mathbf{A} \in \mathbb{R}^{n_w \times n_w}}{\arg\min} \mathcal{L}(\mathbf{w}, \mathbf{A}_w | \mathfrak{D}, \mathbf{w}_0, \mathbf{A}_0, \mathbf{g})$$

Байесовская дистилляция модели



Выборка $\mathfrak{D} \ni (\mathbf{x}, \mathbf{x}^*, y)$, $\mathbf{x} \in \mathbb{R}^n, y \in \mathbb{Y}$, дополнительно $\mathbf{x}^* \in \mathbb{R}^{n^*}$

Модели $\mathfrak{G} \ni \mathbf{g} : \mathbb{R}^n \times \mathbb{R}^{n_w} \to \mathbb{Y}'$ и учитель $\mathbf{f} : \mathbb{R}^{n^*} \times \mathbb{R}^{n^*_w} \to \mathbb{Y}'$

Прогноз $\hat{y} = g(\mathbf{x}, \hat{\mathbf{w}})$ и $s = f(\mathbf{x}^*, \hat{\mathbf{u}})$

Априорное распределение $\mathbf{w} \sim \mathcal{N}\left(\mathbf{w}_0(\hat{\mathbf{u}}, \hat{\mathbf{A}}_u), \mathbf{A}_0(\hat{\mathbf{u}}, \hat{\mathbf{A}}_u)\right)$.

Оптимизационная задача

$$\hat{\mathbf{w}}, \hat{\mathbf{A}}_w = \underset{\mathbf{w} \in \mathbb{R}^{n_w}, \mathbf{A} \in \mathbb{R}^{n_w \times n_w}}{\arg\min} \mathcal{L}\left(\mathbf{w}, \mathbf{A}_w | \mathbf{s}, \mathfrak{D}, \mathbf{w}_0, \mathbf{A}_0, \mathbf{g}, \mathbf{f}\right)$$

Оптимизация модели ученика на основе учителя и привилегированных признаков

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 $\mathbf{z}, \mathbf{v}^*$ — дифференцируемые по параметрам функции заданной структуры, T — параметр температуры.

Функция ошибки

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

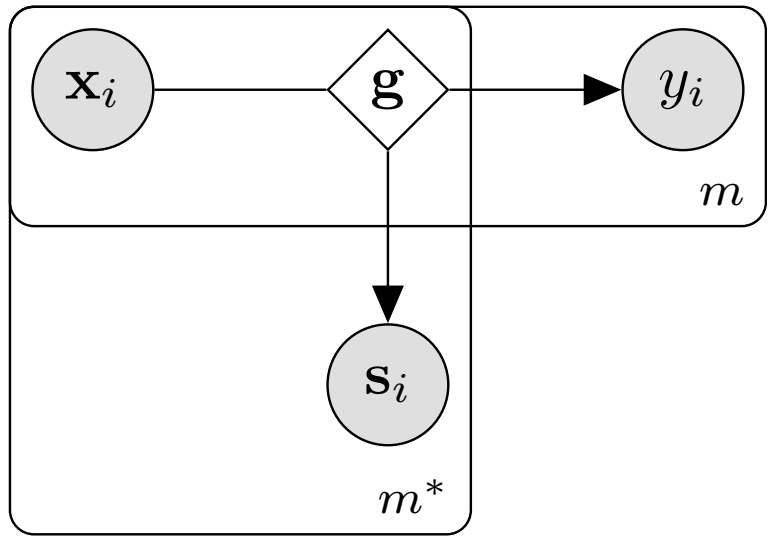
где $\cdot\big|_{T=t}$ фиксирует температуру T.

Оптимальная модель выбирается из класса, $\hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathfrak{G}_{\text{cl}}} \mathcal{L}\left(\mathbf{g}\right)$.

Вероятностная постановка задачи дистилляции

Гипотеза порождения данных:

1) распределение целевой переменной $p(y_i|\mathbf{x}_i, \mathbf{g})$,
2) 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

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

Вероятностная постановка задачи классификации

Заданы

1) учитель $\mathbf{f} \in \mathfrak{F}^*_{\text{cl}}$ и ученик $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$,
2) распределение истинных меток $p\big(y|\mathbf{x},\mathbf{g}\big) = \text{Cat}\big(\mathbf{g}(\mathbf{x})\big)$,
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.*

Вероятностная постановка задачи регрессии

Заданы

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 $p\left(y|\mathbf{x}, g\right) = \mathcal{N}\left(y|g\left(\mathbf{x}\right), \sigma\right)$,
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 $\mathbf{y}'' = \mathbf{X}\mathbf{w} + \varepsilon$, $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$, где $\boldsymbol{\Sigma}^{-1} = diag\left(\boldsymbol{\sigma}'\right)$ и $\mathbf{y}''$ имеют следующий вид:

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

Байесовская постановка задачи дистилляции

Задана модель учителя, суперпозиция

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

где $\mathbf{U}$ матрицы линейных отображений, $\boldsymbol{\sigma}$ монотонная вектор-функция. Параметры учителя фиксированы

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

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

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

где $\mathbf{W}$, $\boldsymbol{\sigma}$ вводятся как и отображения учителя. Задача выбора модели $\mathbf{g}$ состоит в оптимизации вектора $\mathbf{w}$. Решается вариационным выводом

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

Априорное распределение $p\big(\mathbf{w}|\mathbf{A}\big)$ задается как функция от апостериорного распределения параметров учителя $p\big(\mathbf{u}|\mathbf{X}, \mathbf{y}\big)$. Оно задано,

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

Проблема: пространства параметров учителя и ученика не совпадают.

Выравнивание структур моделей

Множество структурных параметров задает вид суперпозиции моделей $\mathbf{f}, \mathbf{g}$.

Определение

Структурные параметры — последовательность размерностей n_s скрытых представлений после каждого слоя нейросетевой модели.

Определение

Выравнивание структур параметрических моделей — изменение структуры одной или нескольких моделей в результате которого векторы параметров моделей различных структур лежат в одном пространстве.

Пространства параметров совпадают

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

тогда

$$p\big(\mathbf{w}|\mathbf{A}\big) = p\big(\mathbf{w}|\mathbf{X}, \mathbf{y}\big).$$

Пространства параметров не совпадают

- выполняется выравнивание моделей учителя $\mathbf{g}$ и ученика $\mathbf{f}$,
- апостериорное распределение учителя $p\big(\mathbf{u}|\mathbf{X}, \mathbf{y}\big)$ назначается априорным распределением ученика $p\big(\mathbf{w}|\mathbf{A}\big)$.

Размеры скрытых слоев учителя и ученика различны

Преобразования t-го слоя учителя:

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

описывает удаление одного нейрона из t-го слоя. Новый вектор параметров $\boldsymbol{\upsilon} = \phi\big(t, \mathbf{u}\big)$. Исходный вектор $\mathbf{u}$ условно делиться на подвекторы $\boldsymbol{\nu}_2, \boldsymbol{\nu}_1, \boldsymbol{\upsilon}$.

Теорема (Грабовой, 2021)

Пусть задано распределения вектора параметров $p\big(\mathbf{u}\big)$. Тогда распределение вектора параметров $p\big(\boldsymbol{\upsilon}\big)$ представимо в виде:

$$p\big(\boldsymbol{\upsilon}\big) = \int\limits_{\boldsymbol{\nu}_2 \in \mathbb{R}^{n_t - 1}} p\big(\bar{\boldsymbol{\nu}}_1 | \mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big) d\boldsymbol{\nu}_2.$$

Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

1) *апостериорное распределение параметров $p\big(\mathbf{u}|\mathfrak{D}\big) = \mathcal{N}\big(\mathbf{m}, \boldsymbol{\Sigma}\big)$,*
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 $n_t \geq n_l$.*

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

Решение задачи выравнивания структур моделей

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

Параметры $\mathbf{u}$ модели $\mathbf{f}$ делятся на удаляемые $\boldsymbol{\nu}_2$, зануляемые $\boldsymbol{\nu}_1$, оставшиеся $\boldsymbol{\upsilon}$.
Суперпозиция слоев модели учителя $\mathbf{f}$ в окрестности t-го слоя:

$$\mathbf{f}(\mathbf{x}) = \cdots \circ \underbrace{\begin{pmatrix} u_{1,1} & \cdots & u_{1,j} & \cdots & u_{1,n_t} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{n_{t+1},1} & \cdots & u_{n_{t+1},j} & \cdots & u_{n_{t+1},n_t} \end{pmatrix}}_{\mathbf{U}_{t+1}} \boldsymbol{\sigma} \circ \underbrace{\begin{pmatrix} u_{1,1} & \cdots & u_{1,n_{t-1}} \\ \vdots & \ddots & \vdots \\ u_{j,1} & \cdots & u_{j,n_{t-1}} \\ \vdots & \ddots & \vdots \\ u_{n_t,1} & \cdots & u_{n_t,n_{t-1}} \end{pmatrix}}_{\mathbf{U}_t} \boldsymbol{\sigma} \circ \cdots \circ \mathbf{U}_1\mathbf{x}$$

Апостериорное распределение параметров $\boldsymbol{\upsilon}$ модели $\mathbf{f}$:

$$p(\boldsymbol{\upsilon}|\mathfrak{D}) = \int_{\boldsymbol{\nu}_2 \in \mathbb{R}^{n_{t-1}}} p(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0})\, d\boldsymbol{\nu}_2.$$

Из свойства распределения $p(\bar{\boldsymbol{\nu}}_1|\mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Xi})$, с параметрами $\boldsymbol{\mu}, \boldsymbol{\Xi}$:

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

$$\boldsymbol{\Xi} = \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\bar{\boldsymbol{\nu}}_1} - \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\boldsymbol{\nu}_1} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\nu}_1,\boldsymbol{\nu}_1} \boldsymbol{\Sigma}_{\bar{\boldsymbol{\nu}}_1,\boldsymbol{\nu}_1},$$

Маргинализация нормального распределения $p(\boldsymbol{\upsilon}|\mathfrak{D}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\upsilon}}, \boldsymbol{\Xi}_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}})$.

Число скрытых слоев учителя и ученика различны

Преобразования t-го слоя учителя:

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

описывает удаление t-го слоя. Новый новый вектор параметров $\boldsymbol{\upsilon} = \psi\left(t, \mathbf{u}\right)$, а элементы вектора, которые были удалены как $\bar{\boldsymbol{\upsilon}}$.

Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

1) *апостериорное распределение параметров* $p\left(\mathbf{u}|\mathfrak{D}\right) = \mathcal{N}\left(\mathbf{m}, \boldsymbol{\Sigma}\right)$,
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:

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

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

Обобщение для рекурентной сети RNN

Структура модели RNN задается размерностью скрытых слоев $n_1, n_2, \ldots, n_T$. Отображение $\phi_{\text{RNN}}\big(t, \mathbf{u}\big) : \mathbb{R}^{\mathsf{p_{tr}}} \to \mathbb{R}^{\mathsf{p_{tr}}-2n_t}$ описывает уменьшение размерности пространства после t-го слоя. Вектор параметров $\boldsymbol{\upsilon} = \phi\big(t, \mathbf{u}\big)$. Исходный вектор $\mathbf{u}$ условно делиться на подвекторы $\boldsymbol{\nu}_2, \boldsymbol{\nu}_1, \boldsymbol{\upsilon}$.

Теорема (Грабовой, 2021)

Пусть задано распределения вектора параметров $p\big(\mathbf{u}\big)$. Тогда распределение вектора параметров $p\big(\boldsymbol{\upsilon}\big)$ представимо в виде:

$$p\big(\boldsymbol{\upsilon}\big) = \int\limits_{\boldsymbol{\nu}_2 \in \mathbb{R}^{n_t - 1}} p\big(\bar{\boldsymbol{\nu}}_1 | \mathfrak{D}, \boldsymbol{\nu}_1 = \mathbf{0}\big)\, d\boldsymbol{\nu}_2.$$

Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

1) *апостериорное распределение параметров $p\big(\mathbf{u}|\mathfrak{D}\big) = \mathcal{N}\big(\mathbf{m}, \boldsymbol{\Sigma}\big)$,*
2) *число слоев модели учителя равняется числу слоев модели ученика $T = L$,*
3) *размеры соответствующих пространств не совпадают $n_t \geq n_t$.*

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

Обобщение для рекурентной сети RNN

Преобразования t-го слоя учителя:

$$\psi_{\text{RNN}}\left(t\right) : \mathbb{R}^{\mathsf{p}_{\mathbf{tr}}} \to \mathbb{R}^{\mathsf{p}_{\mathbf{tr}} - n_t n_{t-1}}$$

описывает удаление t-го слоя. Новый новый вектор параметров $\boldsymbol{\upsilon} = \psi\left(t, \mathbf{u}\right)$, а элементы вектора, которые были удалены как $\bar{\boldsymbol{\upsilon}}$.

Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

1) *апостериорное распределение параметров* $p\left(\mathbf{u}|\mathfrak{D}\right) = \mathcal{N}\left(\mathbf{m}, \boldsymbol{\Sigma}\right)$,
2) *соответствующие размеры пространств совпадают,* $n_t = n_{t+1}$.
3) *функция активации удовлетворяет свойству идемпотентности* $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$.

Тогда апостериорное распределение также описывается нормальным распределением с плотностью распределения:

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

Последовательность выравнивающих преобразований

Множество всех структур задается последовательностью натуральных чисел:

$$\mathfrak{H} = \{(n_1, n_2, \ldots, n_T), \quad n_i \in \mathbb{N}, T \in \mathbb{N}\}.$$

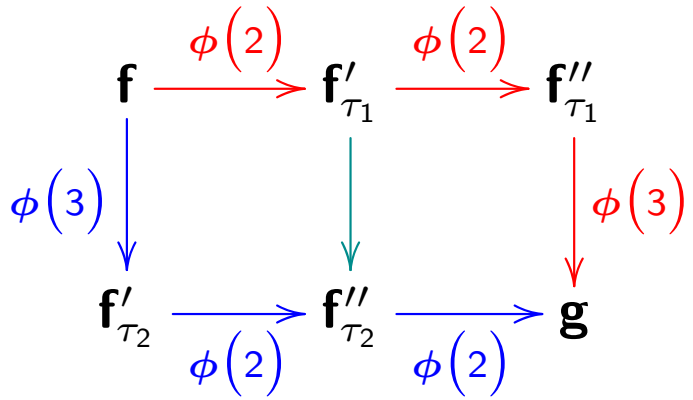
Множество структур, порождаемое структурой $(n'_1, n'_2, \ldots, n'_L)$ учителя $\mathbf{f}$:

$$\mathfrak{H}_{\mathbf{f}} = \{(n_1, n_2, \ldots, n_T); \quad n_i \in \mathbb{N}, L \in \mathbb{N}; \quad n_i \le n'_i,\ 3 \le T \le L; \quad n_{T-1} \le n'_i,\ i > T\}$$

описывает конечное подмножество структур в континуальном множестве.

Теорема (Грабовой, 2021)

Для произвольной структуры $s \in \mathfrak{H}_{\mathbf{f}}$ существует последовательность локальных выравнивающих преобразований $\tau = (\ldots, \phi, \ldots, \psi, \ldots)$ сохраняющее распределение параметров модели $\mathbf{f}$.



Пример: Учитель $\mathbf{f}$ имеет структуру $(10, 6, 7, 10)$, модель ученика $\mathbf{g}$ имеет структуру $(10, 4, 6, 10)$.

Последовательности выравнивающих преобразований f в g:

$\tau_1 = (\phi(2), \phi(2), \phi(3)), \tau_2 = (\phi(3), \phi(2), \phi(2)), \tau_3 = (\phi(2), \phi(3), \phi(2))$

Последовательность выравнивающих преобразований существует, но не единственно.

Введение отношения порядка на множестве параметров

Полный порядок при выборе моделей задается:

1) случайным образом (базовая гипотеза),
2) 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 функции ошибки $\mathcal{L}$ по параметрам модели $\mathbf{u}$,

3) на основе отношения плотности апостериорного распределения параметра в нуле к плотности апостериорного распределения параметра:

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

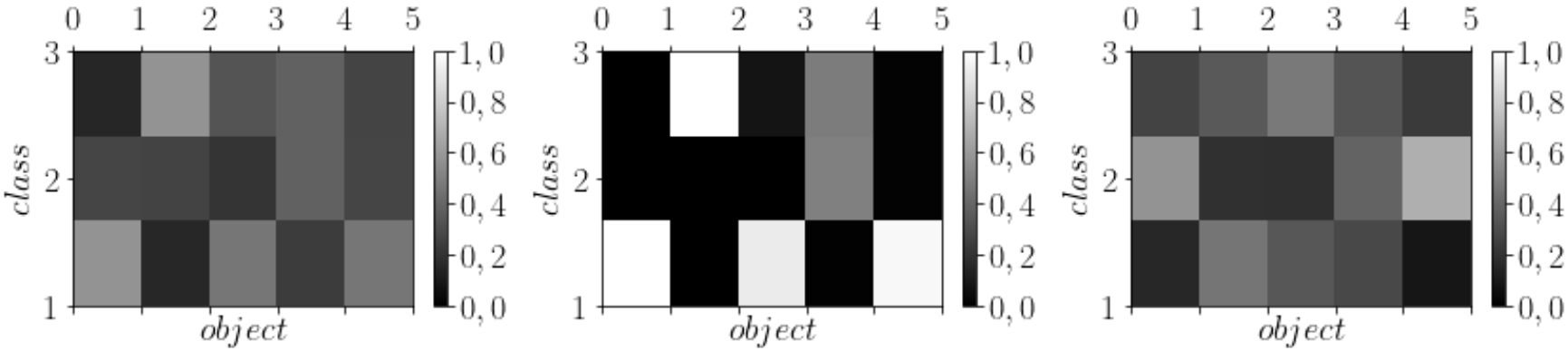
4) на основе анализа мультиколлинеарности параметров методом Белсли:

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

где λ являются сингулярными числами ковариационный матрицы параметров,

5) на основе ковариационной матрицы градиентов функции ошибки $\mathcal{L}$ по параметрам $\mathbf{u}$.

Анализ вероятностных свойств ответов модели ученика



Истинное распределение | Модель без учителя | Модель с учителем

Распределение классов объектов обучающей выборки восстанавливается точнее с учителем.

Выборка	Модель	Кросс-энтропийная ошибка	Кросс-энтропийная ошибка с реальными вероятностями	Вероятностная разница	Точность	Число Параметров
FashionMnist	с учителем	$0{,}453 \pm 0{,}003$	-	$0{,}84 \pm 0{,}13$	$0{,}842 \pm 0{,}002$	7850
	без учителя	$0{,}461 \pm 0{,}005$	-	$0{,}86 \pm 0{,}18$	$0{,}841 \pm 0{,}002$	7850
Systetic	с учителем	$0{,}618 \pm 0{,}001$	$\mathbf{1{,}17 \pm 0{,}05}$	$\mathbf{0{,}45 \pm 0{,}20}$	$0{,}828 \pm 0{,}002$	33
	без учителя	$0{,}422 \pm 0{,}002$	$2{,}64 \pm 0{,}02$	$0{,}75 \pm 0{,}22$	$0{,}831 \pm 0{,}001$	33
Twiter	с учителем	$0{,}489 \pm 0{,}003$	-	$0{,}79 \pm 0{,}17$	$0{,}764 \pm 0{,}005$	1538
	без учителя	$0{,}501 \pm 0{,}006$	-	$0{,}83 \pm 0{,}22$	$0{,}747 \pm 0{,}004$	1538

Точность предсказания модели ученика и учителя на одном уровне.

Модель с учителем аппроксимирует истинные вероятности классов.

Модель с учителем имеет меньшую разницу между вероятностями классов, то есть вероятность не концентрируются в одном классе.

Анализ правдоподобия выборки, выравнивание моделей с разным числом скрытых слоев

Вид суперпозиции модели учителя:

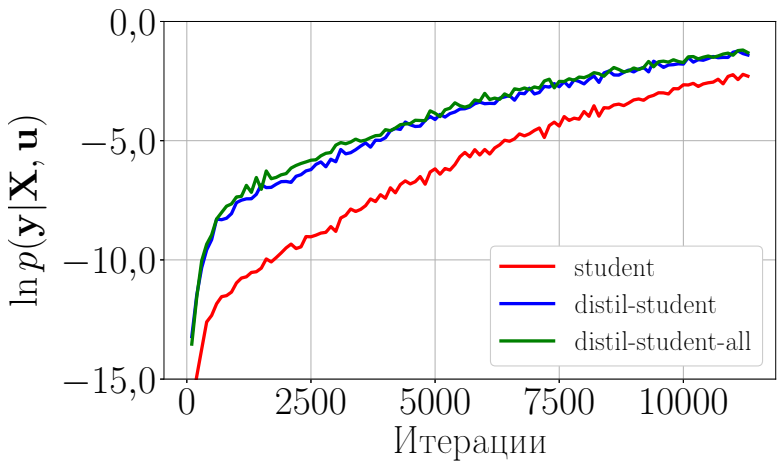
$$f(\mathbf{x}) = \boldsymbol{\sigma} \circ \mathbf{U}_3 \circ \boldsymbol{\sigma} \circ \mathbf{U}_2 \circ \boldsymbol{\sigma} \circ \mathbf{U}_1 \mathbf{x},$$

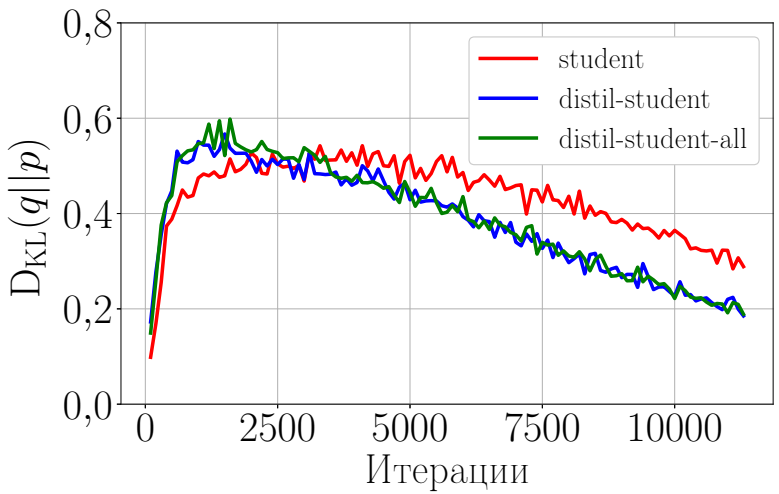
Вид суперпозиции модели ученика:

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

	Учитель	Ученик	Дистил.-ученик	Дистил.-ученик-все
Структура	$[10, 100, 50, 1]$	$[10, 50, 1]$	$[10, 50, 1]$	$[10, 50, 1]$
Число параметров	6050	550	550	550
Разность площадей	-	0	**23310**	**25506**

Анализ решения задачи: интегральная разность $\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ по итерациям.





Правдоподобие модели ученика с дистилляцией растет быстрее чем без нее.

Выносится на защиту

1. Предложен байесовский метод выбора моделей с использованием модели учителя с привилегированной и накопленной информацией.
2. Доказаны теоремы о свойствах дистилляции,
   - *теоремы об эквивалентности* для дистилляции моделей в случае задачи регрессии и классификации,
   - *теоремы о виде априорного распределения* параметров модели ученика в байесовской дистилляции.
3. Предложен метод выравнивания структур параметрических моделей. Предложен метод выбора априорного распределения параметров модели ученика с использованием апостериорного распределения параметров модели учителя для случаев
   - различных размерностей пространств параметров отдельных слоев,
   - различного в числа слоев нескольких моделей.
4. Предложены методы задания порядка на множестве параметров моделей
   - на основе корреляции параметров,
   - на основе оценки скорости сходимости параметров.
5. Предложена вероятностная интерпретации дистилляции моделей глубокого обучения. Исследованы свойства дистилляции моделей глубокого обучения.

Список работ автора по теме диссертации

Публикации в журналах ВАК

1. *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика, 2021.
2. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал вычислительной математики и математической физики, 2021.
3. *A. Grabovoy, V. Strijov.* Quasi-periodic time series clustering for human // Lobachevskii Journal of Mathematics, 2020.
4. *Грабовой А.В., Бахтеев О. Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020.
5. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019.
6. *Грабовой А.В., Стрижов В.В.* Вероятностная интерпретация задачи дистилляции // Автоматика и телемеханика, 2022.
7. *A. Grabovoy, T. Gadaev, A. Motrenko, V. Strijov* Numerical methods of minimum sufficient sample size estimation for linear models // Lobachevskii Journal of Mathematics, 2023.
8. *Bazarova A.I., Grabovoy A.V., Strijov V.V.* Analysis of the properties of probabilistic models in learning problems with an expert // Journal of Computational Mathematics (на рецензировании), 2021.

Выступления с докладом

1. Задача обучения с экспертом для построения интерпретируемых моделей машинного обучения, Международная конференция «Интеллектуализация обработки информации», 2020.
2. Привилегированная информация и дистилляция моделей, Всероссийская конференция «63-я научная конференция МФТИ», 2020.
3. Введение отношения порядка на множестве параметров нейронной сети, Всероссийская конференция «Математические методы распознавания образов ММРО», 2019.
4. Анализ априорных распределений в задаче смеси экспертов, Всероссийская конференция «62-я научная конференция МФТИ», 2019.
5. Поиск оптимальной модели при помощи алгоритмов прореживания, Всероссийская конференция «61-я научная конференция МФТИ», 2018.
6. Автоматическое определение релевантности параметров нейросети, Международная конференция «Интеллектуализация обработки информации», 2018.