

Байесовский выбор субоптимальной структуры модели глубокого обучения

О. Ю. Бахтеев

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

Московский физико-технический институт
6 февраля 2020 г.

Выбор структуры модели глубокого обучения

Цель: предложить метод выбора структуры модели глубокого обучения.

Задачи

- 1 Предложить критерии оптимальной и субоптимальной сложности модели глубокого обучения.
- 2 Предложить алгоритм построения модели субоптимальной сложности и оптимизации параметров.

Исследуемые проблемы

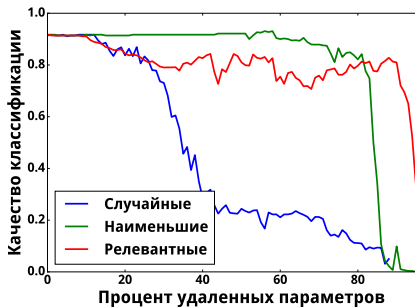
- 1 Большое число параметров и гиперпараметров модели, высокая вычислительная сложность оптимизации.
- 2 Многоэкстремальность и невыпуклость задачи оптимизации.

Методы исследования

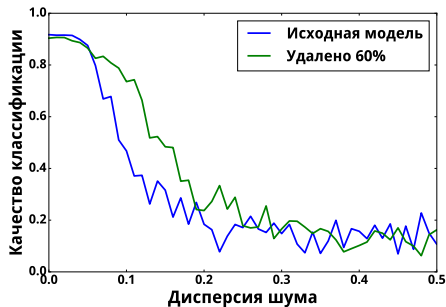
Рассматривается графовое представление нейронной сети. Используются методы вариационного байесовского вывода. Для получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров с использованием градиентных методов оптимизации гиперпараметров и структурных параметров модели.

Проблема выбора оптимальной структуры

Правдоподобие моделей с избыточным числом параметров значимо не меняется при их удалении.



Избыточность параметров модели



Устойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

Модель глубокого обучения

Определение

Моделью $\mathbf{f}(\mathbf{w}, \mathbf{x})$ назовем дифференцируемую по параметрам \mathbf{w} функцию из множества признаков описаний объекта во множество меток:

$$\mathbf{f} : \mathbb{X} \times \mathbb{W} \rightarrow \mathbb{Y},$$

где \mathbb{W} — пространство параметров функции \mathbf{f} .

Особенность задачи выбора модели *глубокого обучения* — значительное число параметров моделей приводит к неприменимости ряда методов оптимизации и выбора структуры модели (AIC, BIC, кросс-валидация).

Модель определяется параметрами \mathbf{W} и структурой Γ .

Структура задает набор суперпозиций, входящих в модель и выбирается согласно статистическим критериям сложности модели.

Эмпирические оценки статистической сложности модели:

- ① число параметров;
- ② число суперпозиций, из которых состоит модель.

Выбор структуры: двуслойная нейросеть

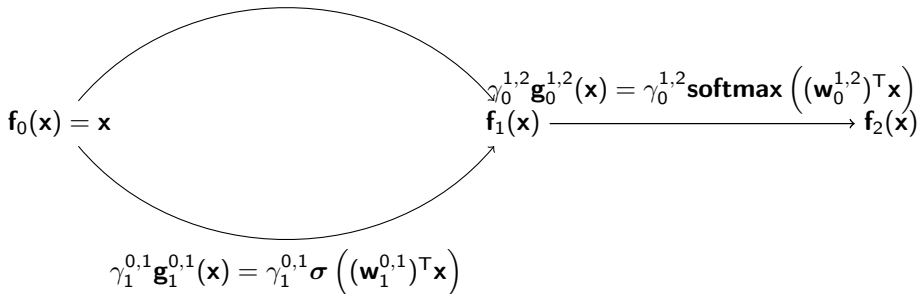
Модель \mathbf{f} задана структурой $\Gamma = [\gamma^{0,1}, \gamma^{1,2}]$.

Модель: $\mathbf{f}(\mathbf{x}) = \text{softmax} \left((\mathbf{w}_0^{1,2})^\top \mathbf{f}_1(\mathbf{x}) \right)$, $\mathbf{f}(\mathbf{x}) : \mathbb{R}^n \rightarrow [0, 1]^{|Y|}$, $\mathbf{x} \in \mathbb{R}^n$.

$$\mathbf{f}_1(\mathbf{x}) = \gamma_0^{0,1} \mathbf{g}_0^{0,1}(\mathbf{x}) + \gamma_1^{0,1} \mathbf{g}_1^{0,1}(\mathbf{x}),$$

где $\mathbf{w} = [\mathbf{w}_0^{0,1}, \mathbf{w}_1^{0,1}, \mathbf{w}_0^{1,2}]^\top$ — матрицы параметров, $\{\mathbf{g}_0^{0,1}, \mathbf{g}_1^{0,1}, \mathbf{g}_0^{1,2}\}$ — обобщенно-линейные функции скрытых слоев нейросети.

$$\gamma_0^{0,1} \mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1} \sigma \left((\mathbf{w}_0^{0,1})^\top \mathbf{x} \right)$$



Графовое представление модели глубокого обучения

Заданы:

- ① ациклический граф (V, E) ;
- ② для каждого ребра $(j, k) \in E$: вектор базовых дифференцируемых функций $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$ мощности $K^{j,k}$;
- ③ для каждой вершины $v \in V$: дифференцируемая функция агрегации agg_v .
- ④ Функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

$$\mathbf{f}_v(\mathbf{w}, \mathbf{x}) = \text{agg}_v \left(\{ \langle \gamma^{j,k}, \mathbf{g}^{j,k} \rangle \circ \mathbf{f}_j(\mathbf{x}) \mid j \in \text{Adj}(v_k) \} \right), v \in \{1, \dots, |V|-1\}, \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \quad (1)$$

и являющаяся функцией из признакового пространства \mathbb{X} в пространство меток \mathbb{Y} при значениях векторов, $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$.

Определение

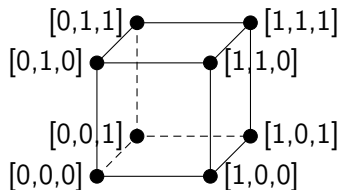
Граф (V, E) со множеством векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и функций агрегаций $\{\text{agg}_v, v \in V\}$ назовем *параметрическим семейством моделей* \mathfrak{F} .

Утверждение

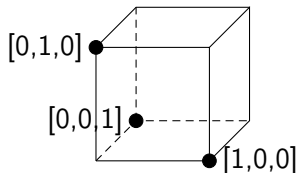
Для любого значения $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Ограничения на структурные параметры

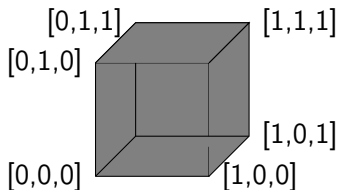
Примеры ограничений для одного структурного параметра γ , $|\gamma| = 3$.



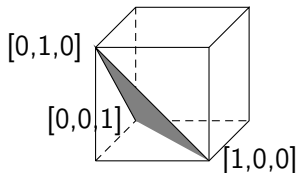
На вершинах куба



На вершинах симплекса



Внутри куба

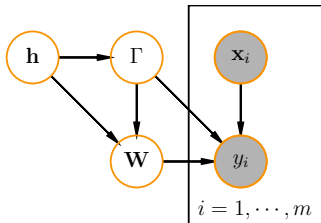


Внутри симплекса

Априорное распределение параметров

Определение

Априорным распределением параметров \mathbf{w} и структуры Γ модели \mathbf{f} назовем вероятностное распределение $p(\mathbf{W}, \Gamma | \mathbf{h}, \lambda) : \mathbb{W} \times \mathbb{\Gamma} \times \mathbb{H} \rightarrow \mathbb{R}^+$, где \mathbb{W} — множество значений параметров модели, $\mathbb{\Gamma}$ — множество значений структуры модели.



Определение

Гиперпараметрами $\mathbf{h} \in \mathbb{H}$ модели назовем параметры распределения $p(\mathbf{w}, \Gamma | \mathbf{h}, \mathbf{f})$ (параметры распределения параметров модели \mathbf{f}).

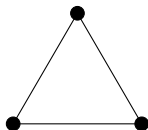
Модель \mathbf{f} задается следующими величинами:

- **Параметры** $\mathbf{w} \in \mathbb{W}$ задают суперпозиции \mathbf{f}_v , из которых состоит модель \mathbf{f} .
- **Структурные параметры** $\Gamma = \{\gamma^{j,k}\}_{(j,k) \in E} \in \mathbb{\Gamma}$ задают вклад суперпозиций \mathbf{f}_v в модель \mathbf{f} .
- **Гиперпараметры** $\mathbf{h} \in \mathbb{H}$ задают распределение параметров и структурных параметров модели.
- **Метапараметры** $\lambda \in \mathbb{A}$ задают вид оптимизации модели.

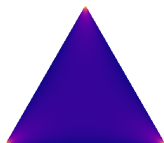
Априорное распределение на структуре модели

Каждая точка на симплексе задает модель.

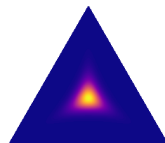
Распределение Гумбель-софтмакс: $\Gamma \sim \text{GS}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \rightarrow 0$

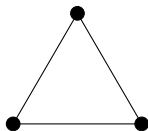


$\lambda_{\text{temp}} = 0.995$

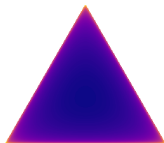


$\lambda_{\text{temp}} = 5.0$

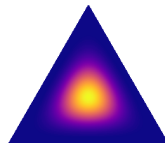
Распределение Дирихле: $\Gamma \sim \text{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \rightarrow 0$



$\lambda_{\text{temp}} = 0.995$

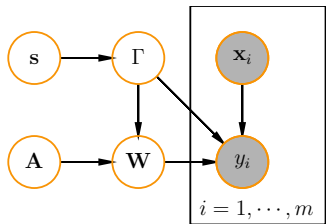


$\lambda_{\text{temp}} = 5.0$

Байесовский выбор модели

Базовая модель:

- параметры модели
 $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1})$,
- гиперпараметры модели $\mathbf{h} = [\alpha]$.



Предлагаемая модель:

- параметры модели
 $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, (\gamma_r^{j,k})^2 (\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ —
диагональная матрица параметров,
соответствующих базовых функций
 $\mathbf{g}_r^{j,k}$,
 $(\mathbf{A}_r^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$,
- структурные параметры модели
 $\Gamma = \{\gamma^{j,k}, (j, k) \in E\}$,
 $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- гиперпараметры модели
 $\mathbf{h} = [\text{diag}(\mathbf{A}), s]$,
- метапараметры $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Обоснованность как статистическая сложность

Статистическая сложность модели f :

$$\text{MDL}(\mathbf{y}, f) = -\log p(\mathbf{h}|\mathbf{f}) - \log p(\hat{\mathbf{w}}|\mathbf{h}, \mathbf{f}) - \log (p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}, \mathbf{f})\delta\mathfrak{D}),$$

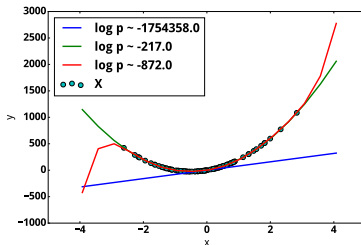
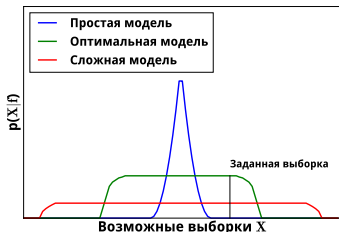
где $\delta\mathfrak{D}$ — допустимая точность передачи информации о выборке \mathfrak{D} .

Оптимизация параметров \mathbf{w} производится согласно апостериорному распределению параметров:

$$L = \log p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{h}, \boldsymbol{\lambda}) \propto \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}) + \log p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda}).$$

Оптимизация гиперпараметров производится в согласно апостериорному распределению гиперпараметров:

$$Q = \log p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \propto \log p(\mathbf{h}|\boldsymbol{\lambda}) + \log \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\lambda}) p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda}) d\mathbf{w}.$$



Вариационная нижняя оценка обоснованности

Интеграл обоснованности невычислим аналитически.

Обоснованность модели:

$$p(y|\mathbf{X}, \mathbf{h}, \lambda) = \iint_{\mathbf{w}, \Gamma} p(y|\mathbf{X}, \mathbf{w}, \Gamma) p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda) d\mathbf{w} d\Gamma.$$

Определение

Вариационными параметрами модели $\theta \in \Theta$ назовем параметры распределения q , приближающие апостериорное распределение параметров и структуры $p(\mathbf{w}, \Gamma|\mathbf{X}, \mathbf{y}, \mathbf{h}, \lambda)$:

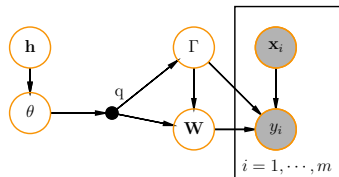
$$q \approx \frac{p(y|\mathbf{X}, \mathbf{w}, \Gamma) p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)}{\iint p(y|\mathbf{X}, \mathbf{w}', \Gamma') p(\mathbf{w}', \Gamma'|\mathbf{h}, \lambda) d\mathbf{w}' d\Gamma'}.$$

Получим нижнюю оценку $\log \hat{p}(y|\mathbf{X}, \mathbf{h}, \lambda)$ интеграла

$$\log p(y|\mathbf{X}, \mathbf{h}, \lambda) \geq E_q \log p(y|\mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) - D_{\text{KL}}(q(\mathbf{w}, \Gamma) \| p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)).$$

Она совпадает с интегралом обоснованности при

$$D_{\text{KL}}(q(\mathbf{w}, \Gamma) \| p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \lambda, \mathbf{h})) = 0.$$



Задача выбора модели

Зададим вариационное распределение $q = q_w q_\Gamma$ с параметрами θ , приближающие апостериорное распределение $p(\mathbf{w}, \Gamma | \mathbf{X}, \mathbf{y}, \mathbf{h}, \lambda)$ параметров и структуры.

Определение

Функцией потерь $L(\theta | \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$ назовем дифференцируемую функцию, качество модели на обучающей выборке при параметрах θ распределения q .

Функцией валидации $Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \theta, \lambda)$ назовем дифференцируемую функцию, качество модели при векторе θ , заданном неявно.

Задачей выбора модели \mathbf{f} назовем двухуровневую задачу оптимизации:

$$\mathbf{h}^* = \arg \max_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \theta^*, \lambda),$$

где θ^* — решение задачи оптимизации

$$\theta^* = \arg \max_{\theta \in \Theta} L(\theta | \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda).$$

Обобщающая задача

Задачу выбора модели \mathbf{h}^*, θ^* назовем обобщающей на множестве $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{A}$, если выполнены условия:

- 1 Область параметров, гиперпараметров и метапараметров не является пустым или точкой.
- 2 Для каждого $\mathbf{h} \in U_h$ и каждого $\lambda \in U_\lambda$ решение θ^* определено однозначно.
- 3 Критерий непрерывности: L, Q непрерывны по метапараметрам.
- 4 Критерий **перехода между структурами**: существует константа $K_3 > 0$, такая, что для произвольных локальных оптимумов $\mathbf{h}_1, \mathbf{h}_2$ задачи оптимизации Q , полученных при метапараметрах λ и удовлетворяющих неравенствам

$$D_{\text{KL}}(p(\Gamma|\mathbf{h}_1, \lambda)|p(\Gamma|\mathbf{h}_1, \lambda)) > K_3, D_{\text{KL}}(p(\Gamma|\mathbf{h}_1, \lambda)|p(\Gamma|\mathbf{h}_2, \lambda)) > K_3,$$

$$Q(\mathbf{h}_1|\lambda) > Q(\mathbf{h}_2|\lambda),$$

существует значение метапараметров $\lambda' \neq \lambda$, такое, что

- 1 соответствие между вариационными параметрами $\theta^*(\mathbf{h}_1), \theta^*(\mathbf{h}_2)$ сохраняется при λ' ,
- 2 выполняется неравенство $Q(\mathbf{h}_1|\lambda') < Q(\mathbf{h}_2|\lambda')$.

Обобщающая задача

Задачу выбора модели \mathbf{h}^*, θ^* назовем обобщающей на множестве $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{A}$, если выполнены условия:

- ⑤ **Критерий максимизации правдоподобия выборки:** существует $\lambda \in U_\lambda$ и $K_1 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h$, $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$: выполнено:
$$E_{q(\mathbf{w}, \Gamma | \theta^*(\mathbf{h}_1))} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) > E_{q(\mathbf{w}, \Gamma | \theta^*(\mathbf{h}_2))} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma).$$
- ⑥ **Критерий минимизации параметрической сложности модели:** существует $\lambda \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h$, $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, сложность первой модели меньше, чем второй.
- ⑦ **Критерий максимизации обоснованности модели:** существует значение гиперпараметров λ , такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели:
$$\mathbf{h}^* \propto \arg \max E_{q(\mathbf{w}, \Gamma | \theta)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)) + \log p(\mathbf{h} | \lambda),$$

$$\theta^* = \arg \min D_{\text{KL}}(q | p(\mathbf{w}, \Gamma | \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)).$$

Анализ задач выбора моделей

Теорема [Бахтеев, 2019]

Следующие задачи выбора модели не являются обобщающими:

- ① критерий максимума правдоподобия: $\max_{\theta} E_q \log p(\mathbf{y}|\mathbf{X}, \theta, \lambda_{\text{temp}}, \mathbf{f})$;
- ② критерий максимума апостериорной вероятности $\max_{\theta} E_q \log p(\mathbf{y}|\mathbf{X}, \theta, \mathbf{f}) p(\theta|\mathbf{h}, \lambda_{\text{temp}})$;
- ③ метод максимума вариационной оценки обоснованности модели $\max_{\mathbf{h}} \max_{\theta} E_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{f}) - D_{KL}(q(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}, \lambda_{\text{temp}})) + \log p(\mathbf{h}|\mathbf{f})$;
- ④ кросс-валидация $\max_{\mathbf{h}} E_q \log p(\mathbf{y}_{\text{valid}}|\mathbf{X}_{\text{valid}}, \theta^*, \lambda_{\text{temp}}, \mathbf{f})$,
 $\theta^* = \arg \max_{\theta} E_q \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}, \theta, \lambda_{\text{temp}}, \mathbf{f}) p(\theta|\mathbf{h})$.
- ⑤ AIC: $\max_{\theta} E_q \log p(\mathbf{y}|\mathbf{X}, \theta, \lambda_{\text{temp}}, \mathbf{f}) - |\theta_i : D_{KL}(q(w_i) || p(w_i|\mathbf{\Gamma}, \mathbf{h}, \lambda)) < \lambda|$;
- ⑥ BIC:
 $\max_{\theta} E_q \log p(\mathbf{y}|\mathbf{X}, \theta, \lambda_{\text{temp}}, \mathbf{f}) - \frac{1}{2} \log(|\mathbb{W}| |\theta_i : D_{KL}(q(w_i) || p(w_i|\mathbf{\Gamma}, \mathbf{h}, \lambda)) < \lambda|$;
- ⑦ перебор структуры модели:
 $\max_{\mathbf{\Gamma}'} \max_{\theta} E_q \log p(\mathbf{y}|\mathbf{X}, \theta, \lambda_{\text{temp}}, \mathbf{f}) \mathbb{I}(q(\mathbf{\Gamma}\mathbf{\Gamma} = \mathbf{p}'))$, где \mathbf{p}' — распределение на структуре (метапараметр).

Предлагаемая задача оптимизации

Теорема [Бахтеев, 2018]

Тогда следующая задача является обобщающей:

$$\begin{aligned} \mathbf{h}^* &= \arg \max_{\mathbf{h}} Q = \\ &= \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta^*)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) - \\ &\quad - \lambda_{\text{prior}}^Q D_{KL}(q(\mathbf{w}, \Gamma | \theta^*) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)) - \\ &\quad - \sum_{p' \in \mathfrak{P}, \lambda \in \lambda_Q^{\text{struct}}} \lambda D_{KL}(p(\Gamma | \mathbf{h}, \lambda) | p') + \log p(\mathbf{h} | \lambda), \end{aligned}$$

где

$$\begin{aligned} \theta^* &= \arg \max_{\theta} L = \mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) \\ &\quad - \lambda_{\text{prior}}^Q D_{KL}(q^*(\mathbf{w}, \Gamma) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)). \end{aligned}$$

Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия и обоснованности, последовательное увеличение и снижение сложности модели, полный перебор структуры.



$$\lambda_{\text{struct}}^Q = [0; 0; 0].$$



$$\lambda_{\text{struct}}^Q = [1; 0; 0].$$



$$\lambda_{\text{struct}}^Q = [1; 1; 0].$$

Адекватность задачи оптимизации

Теорема, [Бахтеев, 2018]

Пусть задано параметрическое множество вариационных распределений: $q(\theta)$.

Пусть $\lambda_{\text{likelihood}}^L = \lambda_{\text{prior}}^L = \lambda_{\text{prior}}^Q = 1, \lambda_{\text{struct}}^Q = 0$. Тогда:

- 1 Предлагаемая задача оптимизации доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности:
$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda) + \log p(\mathbf{h}|\lambda) \rightarrow \max_{\mathbf{h}}.$$
- 2 Вариационное распределение q приближает апостериорное распределение $p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda, \mathbf{f})$ наилучшим образом:
$$D_{\text{KL}}(q||p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)) \rightarrow \min_{\theta}.$$

Пусть также распределение q декомпозируется на два независимых распределения для параметров \mathbf{w} и структуры Γ модели \mathbf{f} :

$$q = q_{\mathbf{w}} q_{\Gamma}, q_{\Gamma} \approx p(\Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda), q_{\mathbf{w}} \approx p(\mathbf{w}|\Gamma, \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda).$$

Если существуют значения вариационных параметров, такие что $q(\mathbf{w}) = p(\mathbf{w}|\Gamma, \mathbf{h}, \lambda)$, $q(\Gamma) = p(\Gamma|\mathbf{h}, \lambda)$, то решение задачи оптимизации для функции L доставляет эти значения.

Оператор оптимизации

Определение

Назовем *оператором оптимизации* T выбор вектора параметров θ' по параметрам предыдущего шага θ .

Оператор стохастического градиентного спуска:

$$\begin{aligned}\hat{\theta} &= T \circ T \circ \dots \circ T(\theta_0, \mathbf{h}) = T^\eta(\theta_0, \mathbf{h}), \quad \text{где } T(\theta, \mathbf{h}) = \\ &= \theta - \lambda_{lr} \nabla (-L(\theta, \mathbf{h})|_{\hat{\mathcal{D}}}),\end{aligned}$$

λ_{lr} — длина шага градиентного спуска, θ_0 — начальное значение параметров θ , $\hat{\mathcal{D}}$ — случайная подвыборка исходной выборки \mathcal{D} .

Перепишем итоговую задачу оптимизации:

$$\mathbf{h}' = T^\eta(Q, \mathbf{h}, T^\eta(L, \theta_0, \mathbf{h})),$$

где θ_0 — начальное значение θ .

Теорема, [Бахтеев, 2019]

Пусть $\frac{\lambda_{\text{prior}}^Q}{\lambda_{\text{likelihood}}^Q} = \lambda_{\text{prior}}^L$. Тогда задача оптимизации представима в виде одноуровневой задачи.

Нижняя вариационная оценка обоснованности на основе мультистарта

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda) \geq E_{q(\mathbf{w})} \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h}, \lambda) - E_{q_{\mathbf{w}}}(-\log(q_{\mathbf{w}})).$$

Теорема [Бахтеев, 2016]

Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C .

Пусть $\theta = [\mathbf{w}^1, \dots, \mathbf{w}^k]$ — начальные приближения оптимизации модели, λ_{lr} — шаг градиентного спуска.

Тогда разность энтропий на смежных шагах оптимизации приближается следующим образом:

$$E_{q_{\mathbf{w}}^{\tau}}(-\log(q_{\mathbf{w}}^{\tau})) - E_{q_{\mathbf{w}}^{\tau-1}}(-\log(q_{\mathbf{w}}^{\tau-1})) \approx \frac{1}{k} \sum_{r=1}^k (\lambda_{lr} \text{Tr}[\mathbf{H}(\mathbf{w}^r)] - \lambda_{lr}^2 \text{Tr}[\mathbf{H}(\mathbf{w}^r)\mathbf{H}(\mathbf{w}^r)]),$$

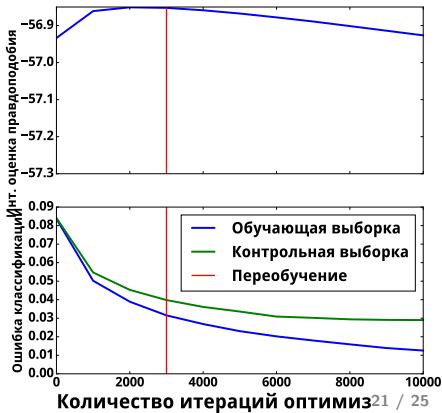
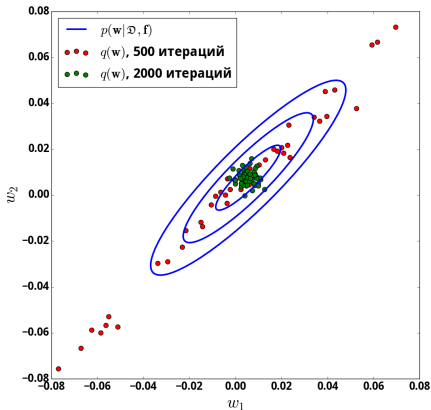
где \mathbf{H} — гессиан минус функции потерь $-L$, $q_{\mathbf{w}}^{\tau}$ — распределение $q_{\mathbf{w}}^{\tau}$ в момент оптимизации τ .

Градиентный спуск как вариационная оценка обоснованности модели

Эмпирическое распределение на точках старта оптимизации — вариационное распределение.

Градиентный спуск не оптимизирует оценку обоснованности.

Снижение вариационной оценки обоснованности — начало переобучения.



Анализ обобщающей задачи оптимизации

Теорема, [Бахтеев, 2018]

Пусть $\lambda_{\text{prior}}^L > 0$, $m \gg 0$, $\frac{m}{\lambda_{\text{prior}}^L} \in \mathbb{N}$. Тогда оптимизация функции

$$L = E_q \log p(y|\mathbf{X}, \mathbf{w}, \Gamma) - \lambda_{\text{prior}}^L D_{\text{KL}}(q||p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda))$$

эквивалентна минимизации $E_{\hat{\mathbf{X}}, \hat{\mathbf{y}} \sim p(\mathbf{x}, y)} D_{\text{KL}}(q||p(\mathbf{w}, \Gamma|\hat{\mathbf{X}}, \hat{\mathbf{y}}, \mathbf{h}, \lambda))$, где $\hat{\mathbf{X}}, \hat{\mathbf{y}}$ — случайные подвыборки мощностью $\frac{m}{\lambda_{\text{prior}}^L}$ из генеральной совокупности.

Определение

Параметрической сложностью модели назовем минимальную дивергенцию между априорным и вариационным распределением:

$$C_p = \min_{\mathbf{h}} D_{\text{KL}}(q||p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)).$$

Теорема, [Бахтеев, 2018]

Пусть $\lambda_{\text{struct}}^Q = 0$. Пусть $\theta_1, \theta_2, \mathbf{h}_1, \mathbf{h}_2$ — результаты оптимизации при разных значениях гиперпараметров $\lambda_{\text{prior}_1}^Q, \lambda_{\text{prior}_2}^Q, \lambda_{\text{prior}_1}^Q > \lambda_{\text{prior}_2}^Q$ на компакте U . Пусть функция $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \theta, \lambda)$ является вогнутой на U при $\lambda_{\text{prior}_2}^Q$. Тогда:

$$C_p(\theta_1|U_{\mathbf{h}}, \lambda_1) - C_p(\theta_2|U_{\mathbf{h}}, \lambda_2) < \frac{\lambda_{\text{prior}}^L}{\lambda_{\text{prior}_2}^Q} (\lambda_{\text{prior}_2}^Q - \lambda_{\text{prior}}^L) C,$$

где C — некоторая константа.

Анализ параметрической сложности

Определение

Относительной вариационной плотностью назовем отношение:

$$\rho(w|\Gamma, \theta_w, h, \lambda) = \frac{q_w(\text{mode } p(w|\Gamma, h, \lambda))}{q_w(\text{mode } q_w)}.$$

Теорема, [Бахтеев, 2018]

Пусть заданы компактные множества $U_h \subset \mathbb{H}$, $U_{\theta_w} \subset \Theta_w$, $U_{\theta_\Gamma} \subset \Theta_\Gamma$, вариационное и априорное распределение $q_w(w|\Gamma, \theta_w)$, $p(w|\Gamma, h, \lambda)$ являются абсолютно непрерывным и унимодальным на U_θ с совпадающей модой и матожиданием. Пусть мода и матожидание априорного распределения не зависят от гиперпараметров h и структуры Γ .

Пусть задана бесконечная последовательность векторов вариационных параметров $\theta[1], \theta[2], \dots, \theta[i], \dots \in U_\theta$, такая, что $\lim_{i \rightarrow \infty} C_p(\theta[i]|U_h, \lambda) = 0$. Тогда:

$$\lim_{i \rightarrow \infty} E_{q_\Gamma(\Gamma|\theta_\Gamma[i])} \rho(w|\Gamma, \theta_w[i], h[i], \lambda)^{-1} = 1, h[i] = \arg \min D_{KL}(q(w, \Gamma|\theta_i) || p(w, \Gamma|h, \lambda)).$$

Результаты, выносимые на защиту

- ① Предложен метод байесовского выбора субоптимальной структуры модели глубокого обучения с использованием автоматического определения релевантности параметров.
- ② Предложены критерии оптимальной и субоптимальной сложности модели глубокого обучения.
- ③ Предложен метод графового описания моделей глубокого обучения.
- ④ Предложено обобщение задачи оптимизации структуры модели, включающее ранее описанные методы выбора модели:
 - ▶ оптимизация обоснованности;
 - ▶ последовательное увеличение сложности модели;
 - ▶ последовательное снижение сложности модели;
 - ▶ полный перебор вариантов структуры модели.
- ⑤ Предложен метод оптимизации вариационной оценки обоснованности на основе мултистарта оптимизации модели.
- ⑥ Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
- ⑦ Исследованы свойства оптимизационной задачи при различных значениях метапараметров. Рассмотрены ее асимптотические свойства.

Список работ автора по теме диссертации

Публикации ВАК

- 1 Bakhteev O., Strijov V. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research. 2019.
- 2 Bakhteev, O., Kuznetsova, R., Romanov, A. and Khritankov, A. A monolingual approach to detection of text reuse in Russian-English collection // In 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) (pp. 3-10). IEEE. 2015.
- 3 Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации // Системы и средства информатики. 26:2 (2016), 4–22.
- 4 Romanov, A., Kuznetsova, R., Bakhteev, O. and Khritankov, A. Machine-Translated Text Detection in a Collection of Russian Scientific Papers. // Computational Linguistics and Intellectual Technologies. 2016.
- 5 Bakhteev, O. and Khazov, A., Author Masking using Sequence-to-Sequence Models // In CLEF (Working Notes). 2017.
- 6 Бахтеев О.Ю., Стрижов В.В. Выбор моделей глубокого обучения субоптимальной сложности // Автоматика и телемеханика. 2018, № 8, 129–147.
- 7 Огальцов А.В., Бахтеев О.Ю. Автоматическое извлечение метаданных из научных PDF-документов // Информатика и её применения. 12:2 (2018), 75–82.
- 8 Смердов А.Н., Бахтеев О.Ю., Стрижов В.В. Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза // Информатика и её применения. 12:4 (2018), 63–69.
- 9 Грабовой А.В., Бахтеев О.Ю., Стрижов В.В. Определение релевантности параметров нейросети // Информатика и её применения. 13:2 (2019), 62–70.

Выступления с докладом

- 1 “Восстановление панельной матрицы и ранжирующей модели в разнородных шкалах”, Всероссийская конференция «57-я научная конференция МФТИ», 2014.
- 2 “Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия”, Международная конференция «Интеллектуализация обработки информации», 2016.
- 3 “Градиентные методы оптимизации гиперпараметров моделей глубокого обучения”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
- 4 “Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
- 5 “Байесовский выбор наиболее правдоподобной структуры модели глубокого обучения”, Международная конференция «Интеллектуализация обработки информации», 2018.