Выбор структуры модели глубокого обучения

Бахтеев Олег

МФТИ

13.02.2019

Графовое представление модели глубокого обучения

Определение

Задан граф (V, E). Для каждого ребра $(j, k) \in E$ определен вектор базовых функций мощности $K^{j,k}$:

$$\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$$

. Пусть для каждой вершины $v \in V$ определена функция агрегации $\mathbf{agg}_v$. Граф (V, E) в совокупности со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и множеством функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ называется *параметрическим семейством моделей* $\mathfrak{F}$, если функция, задаваемая как

$$\mathbf{f}_k(\mathbf{x}) = \mathbf{agg}_k\left(\{\langle\gamma^{j,k}, \mathbf{g}^{j,k}\rangle(\mathbf{f}_j(\mathbf{x}))|j \in \mathrm{Adj}(v_k)\}\right), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \tag{1}$$

является моделью при любых значениях векторов, $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$.

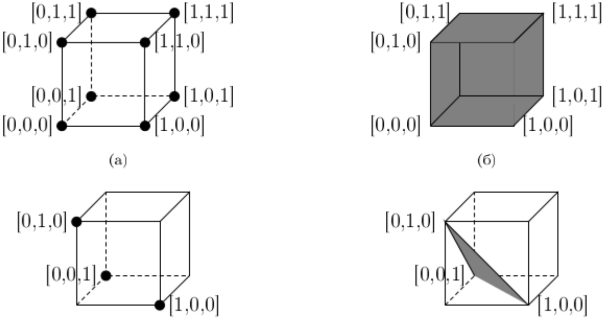
Пример: двуслойная нейросеть

$$\begin{tikzcd}[column sep=huge] \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left, "\mathbf{g}_0^{0,1}(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{w}_0^{0,1}\mathbf{x})"] \arrow[r, bend right, "\mathbf{g}_1^{0,1}(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{w}_1^{0,1}\mathbf{x})"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\mathbf{g}_0^{1,2}(\mathbf{x}) = \mathbf{softmax}(\mathbf{w}_0^{1,2}\mathbf{x})"] & \mathbf{f}_2(\mathbf{x}) \end{tikzcd}$$

Пример: сверточная сеть

$$
\begin{tikzcd}
\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left, "\mathbf{g}_0^{0,1}(\mathbf{x}) = \mathbf{Conv}(\mathbf{x}, c_0, c_1)"] \arrow[r, bend right, "\mathbf{g}_1^{0,1}(\mathbf{x}) = \mathbf{Conv}(\mathbf{x}, 1, c_1) \circ \mathbf{Conv}(\mathbf{x}, c_0, c_2)"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\mathbf{g}_0^{1,2}(\mathbf{x}) = \mathbf{softmax}(\mathbf{w}^{2,1}\mathbf{x})"] & \mathbf{f}_2(\mathbf{x})
\end{tikzcd}
$$

Ограничения на структурные параметры



(а) (б) (в) (г)

Статистические критерии качества модели

Параметрическая сложность — наименьшая дивергенция между априорным распределением параметров и апостериорным распределением параметров:

$$C_{\text{param}} = \min_{\mathbf{h}} \mathrm{D}_{\mathrm{KL}}\big(p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X})||p(\mathbf{W}, \mathbf{\Gamma}||\mathbf{h})\big).$$

Структурная сложность модели — энтропия апостериорного распределения структуры модели:

$$C_{\text{struct}} = -\mathsf{E}_p \log p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}).$$

Выбор оптимальной модели

Основные проблемы выбора оптимальной модели

- Интеграл правдоподобия $p(\mathbf{y}|\mathbf{X}, \mathbf{h})$ невычислим аналитически.
- Задача его оптимизации многоэкстремальна и невыпукла.

Требуется
Предложить метод поиска субоптимального решения задачи оптимизации, обобщающего различные алгоритмы оптимизации:

- Оптимизация правдоподобия.
- Последовательное увеличение и снижение сложности модели.
- Полный перебор вариантов структуры модели.

Распределение на структуре

Пусть для каждого ребра (j,k) задан нормированный положительный вектор $\boldsymbol{\gamma}_{\boldsymbol{j},k} \in \mathbb{R}_{+}^{|K_{\boldsymbol{j},\boldsymbol{k}}|}$, определяющий веса базовых функций из $\mathbf{g}(j,k)$. Перечислим основные свойства, которыми должно обладать распределение такого вектора:

1. $p(\gamma^{j,k})$ является непрерывным на симплексе $\Delta^{K^{\boldsymbol{j},\boldsymbol{k}}-1}$.
2. При устремлении температуры к бесконечности распределение сходится к равномерному: $\lim_{c_{\mathbf{temp}}\to\infty} = p(\gamma^{j,k})|\mathrm{c}_{\mathrm{temp}}) = \mathcal{U}(\Delta^{K^{\boldsymbol{j},\boldsymbol{k}}-1})$.
3. При устремлении температуры к нулю распределение сходится к сингулярному распределению следующего вида:

$$\lim_{c_{\mathbf{temp}}\to 0} p(\gamma_k^{j,k}) = m_k^{j,k},$$

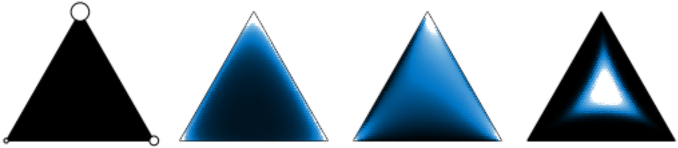
где $\mathbf{m}^{j,k}$ — параметр распределения.

Варианты распределений

1. Дирихле;
2. Гумбель-Софтмакс:

$$\hat{\gamma}_h = \exp\left(\log\left(m_h + \mathrm{Gum}_h\right) c_{\mathrm{temp}}^{-1}\right) \sum_{l=1}^{K_{\boldsymbol{j},k}} \exp\left(\log\left(m_l + \mathrm{Gum}_l\right) c_{\mathrm{temp}}^{-1}\right),$$

где $\mathrm{Gum} \sim -\log(-\log \mathcal{U}(0,1))$.



(a) $\lambda = 0$ (b) $\lambda = 1/2$ (c) $\lambda = 1$ (d) $\lambda = 2$

Maddison et al., 2017.

Оптимизация параметров вариационного распределения

Параметры вариационного распределения $q(\mathbf{W}, \mathbf{\Gamma}) = q_{\mathbf{W}}(\mathbf{W})q_{\mathbf{\Gamma}}(\mathbf{\Gamma})$ оптимизируем:

$$L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{temp}}) -$$

$$-c_{\text{reg}} \mathsf{D}_{KL}\left(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}), q(\mathbf{\Gamma})\right) \to \max_{\mathbf{A}_q, \boldsymbol{\mu}_q, \mathbf{m}_q}.$$

Теорема

Пусть $c_{\text{reg}} > 0$. Тогда $\frac{1}{m}L(c_{\text{reg}})$ сходится п.н. к той же функции, что и $\frac{c_{\text{reg}}}{m_0}L(c_{\text{reg}} = 1)$.

Интерпретация: для достаточно большого m и $c_{\text{reg}} \neq 1$ оптимизация параметров и гиперпараметров эквивалентна оптимизации ELBO для выборки другой мощности.

Теорема [Бахтеев, 2018].

Пусть $\mathbf{\Gamma}_1$ и $\mathbf{\Gamma}_2$ — реализации $\mathbf{\Gamma}$, такие что:

- $\mathbf{\Gamma}_1 \in \bar{\Delta}(\mathbf{\Gamma})$.
- $\mathbf{\Gamma}_2 \notin \bar{\Delta}(\mathbf{\Gamma})$.

Тогда для любых положительно определенных матриц $\mathbf{A}_1$ и $\mathbf{A}_2$ и векторов $\mathbf{m}_1, \mathbf{m}_2, \min(\mathbf{m}_1) > 0$ справедлива следующее отношение апостериорных вероятностей:

$$\lim_{c_{\mathbf{temp}} \to 0} \frac{p(\mathbf{\Gamma}_2, \mathbf{W}_2 | \mathbf{y}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}_2, c_{\text{temp}})}{p(\mathbf{\Gamma}_1, \mathbf{W} | \mathbf{y}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}_1, c_{\text{temp}})} = \infty.$$

Оптимизация параметров априорного распределения

Гиперпараметры $\mathbf{A}, \mathbf{m}$ оптимизируем:

$$Q = c_{\text{train}}\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}}\mathsf{D}_{KL}(p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}, \mathbf{\Gamma}))-$$

$$-c_{\text{comb}} \sum_{p' \in \mathsf{P}} \mathsf{D}_{KL}(\mathbf{\Gamma}|p') \to \max,$$

где P — множество (возможно пустое) распределений на структуре модели.

- c_{train} — коэффициент правдоподобия выборки;
- c_{prior} — коэффициент регуляризации модели;
- c_{comb} — коэффициент перебора структуры.

Общая задача оптимизации

Общая задача оптимизации — двухуровневая:

$$\hat{\mathbf{A}}, \hat{\mathbf{m}} = \arg\max_{\mathbf{A},\mathbf{m}} Q =$$

$$= c_{\text{train}} \mathsf{E}_{\hat{q}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}} \mathsf{D}_{KL}(p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||\hat{q}(\mathbf{W}, \mathbf{\Gamma}))-$$

$$-c_{\text{comb}} \sum_{p' \in \mathbf{P}} \mathsf{D}_{KL}(\mathbf{\Gamma}|p'),$$

где

$$\hat{q} = \arg\max_{q} L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{temp}}) - c_{\text{reg}} \mathsf{D}_{KL}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}), q(\mathbf{\Gamma}))$$

Параметрическая сложность

Обозначим за $F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, c_{\text{comb}}, \mathbf{P}, c_{\text{temp}})$ множество экстремумов функции L при решении задачи двухуровневой оптимизации.

Утверждение

Пусть $\mathbf{f} \in F(1, 1, c_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$. При устремлении c_{prior} к бесконечности параметрическая сложность модели $\mathbf{f}$ устремляется к нулю.

$$\lim_{c_{\text{prior}} \to \infty} C_{\text{param}}(\mathbf{f}) = 0.$$

Параметрическая сложность

Обозначим за $F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, c_{\text{comb}}, \mathbf{P}, c_{\text{temp}})$ множество экстремумов функции L при решении задачи двухуровневой оптимизации.

Утверждение

Пусть $\mathbf{f}_1 \in F(1, 1, c^1_{\text{prior}}, 0, \varnothing, c_{\text{temp}}), \mathbf{h}_2 \in F(1, 1, c^2_{\text{prior}}, 0, \varnothing, c_{\text{temp}}),\ c^1_{\text{prior}} < c^2_{\text{prior}}$. Пусть вариационные параметры моделей $\mathbf{f}_1$ и $\mathbf{f}_2$ лежат в области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда модель $\mathbf{f}_1$ имеет параметрическую сложность, не меньшую чем у $\mathbf{f}_2$.

$$C_{\text{param}}(\mathbf{f}_1) \geq C_{\text{param}}(\mathbf{f}_2).$$

Структурная сложность

Утверждение

Пусть для каждого ребра (i,j) семейства моделей $\mathfrak{F}$ априорное распределение

$$p(\gamma_{i,j}) = \lim_{c_{\textbf{temp}} \to 0} \mathcal{GS}(c_{\text{temp}}).$$

Пусть $c_{\text{reg}} > 0, c_{\text{train}} > 0, c_{\text{prior}} > 0$. Пусть $\mathbf{f} \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$. Тогда структурная сложность модели $\mathbf{f}$ равняется нулю.

$$C_{\text{struct}}(\mathbf{f}) = 0.$$

Структурная сложность

Гипотеза

Пусть $\mathbf{f}_1 \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c^1_{\text{temp}}), \mathbf{h}_2 \in \lim_{c^2_{\text{temp}} \to \infty} F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c^2_{\text{temp}})$. Пусть вариационные параметры моделей f_1 и f_2 лежат в области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда разница структурных сложностей моделей ограничена выражением:

$$C_{\text{struct}}(\mathbf{f}_1) - C_{\text{struct}}(\mathbf{f}_2) \leq \mathsf{E}^1_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c^1_{\text{temp}}) - \mathsf{E}^2_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}^{-1}).$$

Полный перебор

Пусть для каждого ребра (i, j) семейства моделей $\mathfrak{F}$ априорное распределение

$$p(\gamma_{i,j}) = lim_{c_{\mathbf{temp}} \to 0} \mathcal{GS}(c_{\mathbf{temp}}).$$

Рассмотрим последовательность $\mathbf{P}$, состоящую из $N = \prod_{(j,k) \in E} K_{j,k}$ моделей, полученных в ходе оптимизаций вида:

$$f_1 \in F(c_{\mathbf{reg}}, 0, 0, \varnothing, c_{\mathbf{comb}}, c_{\mathbf{temp}}),$$

$$f_2 \in F(c_{\mathbf{reg}}, 0, 0, \{q_1(\mathbf{\Gamma})\}, c_{\mathbf{comb}}, c_{\mathbf{temp}}),$$

$$f_3 \in F(c_{\mathbf{reg}}, 0, 0, \{q_1(\mathbf{\Gamma}), q_2(\mathbf{\Gamma})\}, c_{\mathbf{comb}}, c_{\mathbf{temp}}),$$

где $C_{\mathbf{reg}} > 0, c_{\mathbf{comb}} > 0$.

Гипотеза

Вариационные распределения $q_{\mathbf{\Gamma}}$ структур последовательности $\mathbf{P}$ вырождаются в распределения вида $\delta(\hat{\mathbf{m}})$, где $\hat{\mathbf{m}}$ — точка на декартовом произведении вершин симплексов структуры модели.
Последовательность соответствует полному перебору структуры $\mathbf{\Gamma}$.