Вариационная оптимизация моделей глубокого обучения с контролем сложности модели

О. С. Гребенькова, О. Ю. Бахтеев, В. В. Стрижов
grebenkova.os@phystech.edu; bakhteev@phystech.edu; strijov@ccas.ru

В работе исследуется задача построения модели глубокого обучения с возможностью задания ее сложности. Под сложностью модели понимается минимальная длина описания, минимальное количество информации, которое требуется для передачи информации о модели и о выборке. Предлагается метод оптимизации модели, основанный на представлении модели глубокого обучения в виде гиперсети с использованием байесовского подхода, где под гиперсетью понимается сеть, которая генерирует параметры другой сети. Вводятся вероятностные предположения о параметрах моделей глубокого обучения. Предлагается максимизировать вариационную нижнюю оценку байесовской обоснованности модели. Вариационная оценка рассматривается как условная величина, зависящая от задаваемой требуемой сложности модели. Для анализа качества представленного алгоритма проводятся эксперименты на выборках MNIST и CIFAR.

Ключевые слова: *вариационная оптимизация моделей; гиперсети; глубокое обучение; нейронные сети; байесовский подход; заданная сложность модели*

1 Введение

В данной работе рассматривается задача оптимизации модели глубокого обучения с заранее заданной сложностью модели. Построение модели заданной сложности является одной из фундаментальных проблем глубокого обучения, так как по построению данный класс моделей имеет избытычное число параметров [1].

Предлагаемый метод заключается в представлении модели глубокого обучения в виде гиперсети, с использованием байесовского подхода. Гиперсеть — сеть, которая задаёт параметры другой сети. В работе [2] рассмотрены статистические и динамические гиперсети для генерации весов сверточных и рекурентных сетей соответственно.

Вводятся вероятностные предположения о параметрах моделей глубокого обучения. В работе [1] предлагается использовать в качестве сетевой функции ошибки минимальную длину описания, т.е. минимальное количество информации, которое требуется для передачи информации о модели и о выборке, для оптимизации параметров модели глубокого обучения. Также в работе [1] получены виды функций потерь ошибки и потерь сложности для дельта и гауссова распределений апроксимации апостериорной вероятности.

Смежной задачей к построению модели заданной сложности выступает задача порождения и выбора оптимальной структуры сетей глубокого обучения. В работе [3] рассматривается возможность порождения широкого класса свёрточных сетей как подмоделей обощенной сети, которая называется «фабрикой»(англ. fabric). Данные структуры позволяют обойти процесс оптимизации параметров и проверки качества одиночных сетевых архитектур. В работах [4], [5] представлены сетевые архитектуры для решения задачи выбора структуры нейронной сети с использованием дифференцируемых алгоритмов - стохастическая (англ. Stochastic Neural Architecture Search — SNAS) и дифференцируемая (англ. Differentiable Neural Architecture Search — DNAS) нейронные архитектуры. Ососбенностью работы [5] является решение задачи выбора архитектуры модели, удовлетворяющей эксплутационным требованиям.

Проверка и анализ метода проводятся на выборках MNIST [6] и CIFAR-10 [7].

2 Постановка задачи

Задана выборка:

$$\mathfrak{D} = \{\boldsymbol{x_i}, y_i, \} \quad i = 1, \ldots, N,$$

где $\boldsymbol{x_i} \in \mathbb{R}^m$, $y_i \in \{1, \ldots Y\}$, Y — число классов. Рассмотрим модель $f(\boldsymbol{x}, \boldsymbol{w}) : \mathbb{R}^m \times \times \mathbb{R}^n \longrightarrow \{1, \ldots, Y\}$, где $\boldsymbol{w} \in \mathbb{R}^n$ — пространство параметров модели,

$$f(\boldsymbol{x}, \boldsymbol{w}) = softmax(f_1(f_2(\cdots(f_l(\boldsymbol{x}, \boldsymbol{w}))))),$$

где $f_k(\boldsymbol{x}, \boldsymbol{w}) = tanh(\boldsymbol{w}^\mathsf{T}\boldsymbol{x})$, $k \in \{1, \ldots, l\}; l$ — число слоёв нейронной сети. Параметр w_j модели f называется активным, если $w_j \neq 0$. Множество индексов активных параметров обозначим $\mathcal{A} \subset \mathcal{J} = \{1, \ldots, n\}$. Задано пространство парметров модели

$$\mathbb{W}_\mathcal{A} = \{\boldsymbol{w} \in \mathbb{R}^n | w_j \neq 0, \quad j \in \mathcal{A}\}.$$

Для модели f с множеством индексов активных параметров $\mathcal{A}$ и соответствующего ей вектора параметров $\boldsymbol{w} \in \mathbb{W}_\mathcal{A}$ определим логарфмическую функцию правдобподобия выборки:

$$\mathcal{L}_\mathfrak{D}(\mathfrak{D}, \mathcal{A}, \boldsymbol{w}) = \log p(\mathfrak{D}|\mathcal{A}, \boldsymbol{w}), \tag{1}$$

где $p(\mathfrak{D}|\mathcal{A}, \boldsymbol{w})$ - апостериорная вероятность выборки $\mathfrak{D}$ при заданных $\boldsymbol{w}, \mathcal{A}$. Оптимальные значения $\boldsymbol{w}, \mathcal{A}$ находятся из минимизации — $\mathcal{L}_\mathcal{A}(\mathfrak{D}, \mathcal{A})$ — логарифма правдоподобия модели:

$$\mathcal{L}_\mathcal{A}(\mathfrak{D}, \mathcal{A}) = \log p(\mathfrak{D}|\mathcal{A}) = \log \int\limits_{\boldsymbol{w} \in \mathbb{W}_\mathcal{J}} p(\mathfrak{D}|\boldsymbol{w}) p(\boldsymbol{w}|\mathcal{A}) d\boldsymbol{w}, \tag{2}$$

где $p(\boldsymbol{w}, \mathcal{A})$ — априорная вероятность вектора параметров в пространстве $\mathbb{W}_\mathcal{J}$.

Так как вычисление интеграла (2) является вычислительно сложной задачей, рассмотрим вариационный подход для решения этой задачи. Пусть задано распределение:

$$q(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{A}_{ps}^{-1}).$$

Здесь $\boldsymbol{m}, \boldsymbol{A}_{ps}^{-1}$ — вектор средних и матрица ковариации, аппроксимирующеее неизвестное апостериорное распределение $p(\boldsymbol{w}|\mathfrak{D}, \mathcal{A})$, полученное при априорном предположении

$$p(\boldsymbol{w}|\mathcal{A}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A}_{pr}^{-1}),$$

где $\boldsymbol{\mu}, \boldsymbol{A}_{pr}^{-1}$ — вектор средних и матрица ковариации априорного распределения.

Приблизим интеграл (2):

$$\mathcal{L}_\mathcal{A}(\mathfrak{D}, \mathcal{A}) = \log p(\mathfrak{D}|\mathcal{A}) = \int\limits_{\boldsymbol{w} \in \mathbb{W}_\mathcal{J}} q(\boldsymbol{w}) \log \frac{p(\mathfrak{D}, \boldsymbol{w}|\mathcal{A})}{q(\boldsymbol{w})} d\boldsymbol{w} - \int\limits_{\boldsymbol{w} \in \mathbb{W}_\mathcal{J}} q(\boldsymbol{w}) \log \frac{p(\boldsymbol{w}|\mathfrak{D}, \mathcal{A})}{q(\boldsymbol{w})} d\boldsymbol{w} \approx$$

$$\approx \int\limits_{\boldsymbol{w} \in \mathbb{W}_\mathcal{J}} q(\boldsymbol{w}) \log \frac{p(\mathfrak{D}, \boldsymbol{w}|\mathcal{A})}{q(\boldsymbol{w})} d\boldsymbol{w} = \int\limits_{\boldsymbol{w} \in \mathbb{W}_\mathcal{J}} q(\boldsymbol{w}) \log \frac{p(\boldsymbol{w}|\mathcal{A})}{q(\boldsymbol{w})} d\boldsymbol{w} + \int\limits_{\boldsymbol{w} \in \mathbb{W}_\mathcal{J}} q(\boldsymbol{w}) \log p(\mathfrak{D}|\mathcal{A}, \boldsymbol{w}) d\boldsymbol{w} =$$

$$= \mathcal{L}_{\boldsymbol{w}}(\mathfrak{D}, \mathcal{A}, \boldsymbol{w}) + \mathcal{L}_E(\mathfrak{D}, \mathcal{A}) \tag{3}$$

Первое слагаемое формулы (3) — это сложность модели. Оно определяется расстоянием Кульбака– Лейблера:

$$\mathcal{L}_w(\mathfrak{D}, \mathcal{A}, \boldsymbol{w}) = -D_{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\mathcal{A})).$$

Второе слагаемое формулы (3) представляет собой математическое ожидание правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \boldsymbol{w})$. В данной работе оно является функцией ошибки:

$$\mathcal{L}_E(\mathfrak{D}, \mathcal{A}) = E_{\boldsymbol{w}\sim q}\mathcal{L}_{\mathfrak{D}}(\boldsymbol{y}, \mathfrak{D}, \mathcal{A}, \boldsymbol{w}).$$

Требуется найти параметры, доставляющие минимум суммарному функционалу потерь $\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \boldsymbol{w})$ из (3):

$$\hat{\boldsymbol{w}} = \underset{\mathcal{A}\subset\mathcal{J}, \boldsymbol{w}\in\mathbb{W}_{\mathcal{A}}}{\arg\min} -\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \boldsymbol{w}) = \underset{\mathcal{A}\subset\mathcal{J}, \boldsymbol{w}\in\mathbb{W}_{\mathcal{A}}}{\arg\min} D_{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\mathcal{A})) - \mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \boldsymbol{w}). \qquad (4)$$

3 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

[1] *Graves Alex.* Practical variational inference for neural networks // Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain / Eds. John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, Kilian Q. Weinberger. — 2011. P. 2348–2356. URL: `http://papers.nips.cc/book/advances-in-neural-information-processing-systems-24-2011`.

[2] *Ha David, Dai Andrew M., Le Quoc V.* Hypernetworks // CoRR, 2016. Vol. abs/1609.09106. URL: `http://arxiv.org/abs/1609.09106`.

[3] *Saxena Shreyas, Verbeek Jakob.* Convolutional neural fabrics // CoRR, 2016. Vol. abs/1606.02492. URL: `http://arxiv.org/abs/1606.02492`.

[4] *Xie Sirui, Zheng Hehui, Liu Chunxiao, Lin Liang.* Snas: Stochastic neural architecture search // CoRR, 2018. Vol. abs/1812.09926. URL: `http://arxiv.org/abs/1812.09926`.

[5] *Wu Bichen, Dai Xiaoliang, Zhang Peizhao, Wang Yanghan, Sun Fei et al.* Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search // CoRR, 2018. Vol. abs/1812.03443. URL: `http://arxiv.org/abs/1812.03443`.

[6] *LeCun Yann, Cortes Corinna.* MNIST handwritten digit database, 2010. URL: `http://yann.lecun.com/exdb/mnist/`.

[7] *Krizhevsky Alex, Nair Vinod, Hinton Geoffrey.* Cifar-10 (canadian institute for advanced research). URL: `http://www.cs.toronto.edu/~kriz/cifar.html`.

Received February 25, 2020