Технический отчет “Последовательное порождение моделей глубокого обучения оптимальной сложности”

О. Ю. Бахтеев

bakhteev@phystech.edu

Московский физико-технический институт, ФУПМ, Кафедра интеллектуальных систем

Работа посвящена поиску оптимальной структуры сети глубокого обучения и оптимизации ее параметров. Приняты статистические гипотезы о распределении зависимой переменной и параметров модели. Рассматриваемые модели декомпозируются на порождающую и разделяющую. На основании статистических предположений принимается оптимизируемая функция ошибки для каждой подмодели. Предлагаются методы оптимизации полученных функций ошибок. В качестве рассматриваемой порождающей модели используется вариационный автокодировщик. В качестве рассматриваемой разделяющей модели используется нейронная сеть с одним скрытым слоем. По предложенному алгоритму поиска структуры сети был проведен эксперимент на выборке изображений рукописных цифр MNIST.

Ключевые слова: *Поиск оптимальной структуры сети; глубокое обучение; байесовский вывод; вариационный вывод*

1 Основная часть

Пусть задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, N, \tag{1}$$

состоящая из множества пар «объект - класс», $\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n$. Каждый объект $\mathbf{x}$ принадлежит одному из Z классов с меткой $y_i \in \mathbf{Y}$.

Моделью классификации или сетью глубокого обучения $\mathbf{f}$ назовем суперпозицию функций:

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \boldsymbol{\mu}_1(\boldsymbol{\mu}_2(\dots \boldsymbol{\mu}_K(\mathbf{x}))) : \mathbb{R}^n \to [0, 1]^Z, \tag{2}$$

где $\boldsymbol{\mu}_k, k \in \{1, \dots, K\}$, — модели, параметрическое семейство вектор-функции; $\mathbf{w}$ — вектор параметров моделей; r-ю компоненту вектора $\mathbf{f}(\mathbf{x}, \mathbf{w})$ будем интерпретировать как вероятность отнесения объекта $\mathbf{x}_i$ к классу с меткой r.

Множество всех рассматриваемых моделей обозначим за $\mathfrak{F}$. Будем полагать, что для каждой модели $\mathbf{f} \in \mathbf{F}$ задано априорное распределение параметров $p(\mathbf{W}|\mathbf{f})$.

1.1 Определение

Модель классификации $\mathbf{f}$ является оптимальной среди моделей $\mathfrak{F}$, если достигается максимум интеграла:

$$p(\mathbf{X}, \mathbf{Y}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}, \mathbf{Y}|\mathbf{X}, \mathbf{w}) p(\mathbf{W}|\mathbf{f}) d\mathbf{W}. \tag{3}$$

Требуется найти оптимальную модель $\mathbf{f}$, а также значения ее параметров $\mathbf{W}$, доставляющие максимум интегралу:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{Y}, \mathbf{f}) \sim p(\mathbf{X}, \mathbf{Y}|\mathbf{W}, \mathbf{f}) p(\mathbf{W}|\mathbf{f}).$$

Декомпозируем модель $\mathbf{f}$ на порождающую $\mathbf{f}_G$ и разделяющую $\mathbf{f}_D$. Будем полагать, что выборка $\mathbf{X}$ была порождена некоторой случайной величиной $\mathbf{z}$. Положим решающее правило классификации:

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}_D(\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})),$$

где p — условная вероятность реализации случайной величины $\mathbf{z}$ при условии наблюдения $\mathbf{x}$ и порождающей функции f_G, f_D — разделяющая функция.

Поиск наилучшей модели будем осуществлять последовательно на некотором подмножестве моделей.

Определение 1. *Порождающая модель $\mathbf{f}_G$ является субоптимальной на множестве порождающих моделей $\mathfrak{F}_G$, если модель доставляет максимум нижней границе интеграла:*

$$log \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})p(\mathbf{W}|\mathbf{f})d\mathbf{W}. \tag{4}$$

В качестве множества моделей $\mathfrak{F}_G$ будем рассматривать модели вариационных автокодировщиков.

Определение 2. *Разделяющая модель $\mathbf{f}_D$ является субоптимальной на множестве разделяющих моделей $\mathfrak{F}_D$ при условии модели $\mathbf{f}_G$ и ее параметров $\mathbf{W}_G$ если модель доставляет максимум нижней границе интеграла:*

$$log \int_{\mathbf{w}} p(\mathbf{Y}|\hat{\mathbf{Z}})p(\mathbf{W}|\mathbf{f})d\mathbf{W}, \tag{5}$$

где $\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X})$.

Определение 3. *Модель классификации $\mathbf{f}$ является субоптимальной, если получена композицией субоптимальных моделей порождения и разделения.*

Таким образом, задача нахождения субоптимальной модели классификации $\mathbf{f}$ сводится к следующим подзадачам:
- нахождение субоптимальной модели порождения $\mathbf{f}_G$;
- нахождение оптимальных параметров $\mathbf{W}_G$ модели порождения $\mathbf{f}_G$;
- нахождение субоптимальной модели разделения $\mathbf{f}_D$;
- нахождение оптимальных параметров $\mathbf{W}_D$ модели порождения $\mathbf{f}_D$.

2 Рассматриваемые модели

2.1 Softmax-сеть

Softmax-сеть $\boldsymbol{\mu}_{\text{SM}}$ представляет собой однослойную нейросеть вида:

$$\mathbf{a}(\mathbf{x}) = \mathbf{W}_2^{\mathsf{T}}\mathbf{tanh}(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}),$$

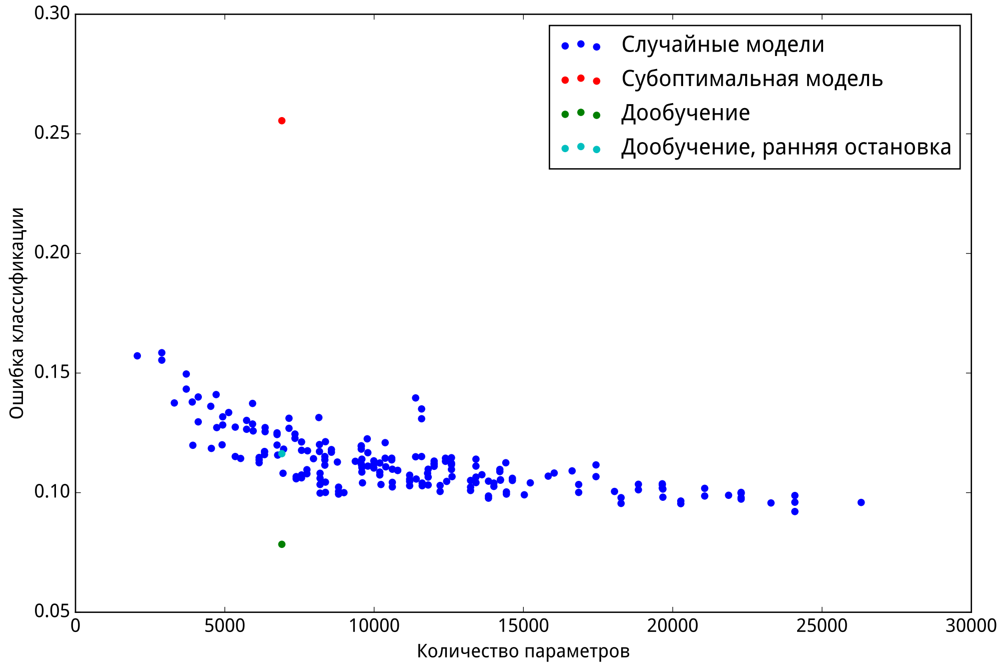
$$\boldsymbol{\mu}_{\text{SM}}(\mathbf{x}) = \frac{\mathbf{exp}\big(\mathbf{a}(\mathbf{x})\big)}{\sum_{j=1}^{Z}\exp\big(a_j(\mathbf{x})\big)}.$$

2.2 Вариационный автокодировщик

Будем полагать, что объекты $\mathbf{X}$ порождены случайно величиной $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{w}).$$

Рис. 1 График качества полученных моделей



Пусть $q_\varphi(\mathbf{z}|\mathbf{x})$ — распределение, аппроксимирующее неизвестное распределение $p(\mathbf{z}|\mathbf{x}, \mathbf{w})$. Тогда

$$\log p(\mathbf{x}|\mathbf{w}) \geqslant \mathbb{E}_{q_\varphi(\mathbf{z}|\mathbf{x})}\log p(\mathbf{x}|\mathbf{z}, \mathbf{w}) - \mathrm{D}_{KL}(q_\varphi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})),$$

где $q_\varphi(\mathbf{z}|\mathbf{x}), p(\mathbf{x}|\mathbf{z}, \mathbf{w})$ — распределения, параметризуемые нейросетью:

$$q_\varphi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{x} + \boldsymbol{\mu}_\varphi(\mathbf{x}), \boldsymbol{\sigma}^2_\varphi(\mathbf{x})),$$

$$p(\mathbf{x}|\mathbf{z}, \mathbf{w}) = \mathcal{N}(\mathbf{z} + \boldsymbol{\mu}_\mathbf{w}(\mathbf{z}), \boldsymbol{\sigma}^2_\mathbf{w}(\mathbf{z})).$$

3 Вычислительный эксперимент

Вычислительный эксперимент проводился на выборке изображений рукописных цифр MNIST. Предварительно выборка была домножена на случайную матрицу размера 784. В качестве сравнения предложенного подхода был обучен ряд случайных моделей той же структуры. Как видно из графика на рис. **??**, субоптимальная модель дает достаточно слабое качество классификации. Тем не менее, после дообучения сети, результат превосходит качество моделей с тем же количеством параметров, но полученных без предобучения. Был также рассмотрен метод предобучения с ранней остановкой оптимизации, где в качестве критерия остановки выступало снижение оценки интеграла (**??**). Кроме того, как видно из графика, предложеный подход позволяет выбирать структуру сети, имеющую достаточно небольшое количество параметров и качество, сравнимое с качеством более сложных моделей.

4 Заключение

В работе предложен метод выбора модели субоптимальной сложности. Метод основывается на статистических предположениях и эвристической декомпозиции модели. Предложенный метод позволяет оптимизировать структуру каждой из подмоделей в отдельности. Вычислительный эксперимент показал хорошее качество классификации на моделях, полученных данным методом.

Consequently generated deep learning models of optimal complexity

O. Y. Bakhteev
bakhteev@phystech.edu
Moscow Institute of Physics and Technology

Background: The paper investigates the problem of optimal model selection in deep learning. The one of the problems of deep learning model selection is its ambiguity: either if the classification quality of a model is rather good, the model can be unstable or overfitted.

Methods: The paper focuses on the statistical hypothesis about the distribution of data and parameter values. Based on this hypothesis we decompose the model into generative and discriminative part. For both of these parts the marginal likelihood maximization criterion is introduced. In order to evluate the marginal likelihood variational methods are considered.

Results: In model selection experiment we considered two types of models: softmax-neural network with one hidden layer and variational autoencoder. The experiments were conducted on MNIST dataset.

Keywords: *model selection; deep learning; variational inference; marginal likelihood; evidence*