Автоматическое построение нейронной сети оптимальной сложности

Товкес А. А.[1]***, Бахтеев О. Ю.***[1]***, Стрижов В. В.***[2]
tovkes.aa@phystech.edu; bakhteev@phystech.edu; strijov@phystech.edu
Московский физико-технический институт[1];
Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН[1,2]

Работа посвящена задаче выбора оптимальной по сложности модели нейросети. Нейросеть представляется в виде вычислительного графа, где ребрам соответствуют базовые функции, а вершинам — промежуточные представления выборки под действием этих функций. Параметры сети разделяются на непосредственно параметры модели, которые определяют итоговое качество классификации; гиперпараметры, определяющие процесс обучения и предотвращение переобучения; структурные параметры, определяющие непосредственно структуру модели Для решения задачи оптимизации предлагается проводить релаксацию структуры нейросети. Рассмотрено изменение характеристик нейросети при возмущении структурных параметров. Для анализа качества представленного алгоритма проводятся эксперименты на выборках Boston, MNIST и CIFAR-10.

Ключевые слова: *нейронные сети, оптимизация гиперпараметров, оптимальная структура нейронной сети.*

1 Введение

В данной работе рассматривается задача построения оптимальной нейронной сети. В силу большого количества оптимизируемых параметров модели, задача выбора модели глубокого обучения является вычислительно сложной. Поэтому задача выбора структуры модели глубокого обучения включает в себя выбор стратегии параметризации структуры и построения модели, эффективной по вычислительным ресурсам.

Под оптимальной моделью понимается та модель, которая которая дает приемлемое качество при небольшом числе параметров модели. Нейросеть представляется в виде вычислительного графа, где ребрам соответствуют базовые функции, а вершинам — промежуточные представления выборки под действием этих функций. Сами же структурные параметры — это вектор весов каждой базовой функции, определяющий их вклад в итоговую модель.

Существует ряд подходов к построению нейронной сети. В работах [1, 2] предлагается использовать метод прореживания модели. Он заключается в построении переусложненной модели, с последующем удалением параметров, не влияющих на качество. В [3] используется предсказание структуры модели другой нейросетью. Кроме того в [4] используется метаобучение, которое по некторой входной выборке, возвращает оптимальные гиперпараметры.

В данной работе исследуется метод построения модели глубокого обучения, позволяющий производить оптимизацию параметров, гиперпараметров и структурных параметров в единой процедуре. В основе разработанного метода лежит алгоритм DARTS, предложенный в работе [5]. Для выбора оптимальных значений гиперпараметров и структурных параметров предлагается параметризовать структуру модели некотором действительным вектором, таким образом переходя от дискретного множетсва значений к непрерыному.

Проверка и анализ метода проводится на выборке Boston Housing [6], MNIST [7],CIFAR-10 [8] и синтетических данных. Результат сравнивается с моделью полученной при помощи

базовых алгоритмов. Критериями качества рассматриваемых алгоритмов выступают качество и полученных моделей и их устойчивость к возмущениям параметров, а также вычилсительная сложность методов.

2 Постановка задачи

Пусть заданы обучающая и вылидационная выборки:

$$\mathfrak{D}^{\text{train}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \ldots, m^{\text{train}},$$

$$\mathfrak{D}^{\text{valid}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \ldots, m^{\text{valid}},$$

состоящие из множеств пар объект-метка,

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^{\text{n}}, \quad y_i \in \mathbf{Y} \subset \mathbb{R}.$$

$\mathbf{Y} = \{1, \ldots, Z\}$, где Z - количество классов.

Модель задаётся ориентированным графом $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, где для каждого ребра (i, j) задан вектор базовых функций $\mathbf{g}^{i,j}$, с мощностью $|\mathbf{g}^{i,j}| = K^{i,j}$ и весами $\boldsymbol{\gamma}^{i,j}$. Требуется построить такую модель $\mathbf{f}$ с параметрами $\mathbf{W} \in \mathbb{R}^{\text{n}}$:

$$\mathbf{f}(\mathbf{x}, \mathbf{W}) = \{\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i)\}_{i=1}^{|\mathbf{V}|}$$

где $\mathbf{f_i}(\mathbf{x}, \mathbf{w_i})$ - подмодель с параметрами $\mathbf{w}_i$ задаётся через графовое представление как:

$$\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i) \ = \sum_{k \in adj(i)} \left\langle \boldsymbol{\gamma}^{i,k}, \mathbf{g}^{i,k} \right\rangle \mathbf{f}_k(\mathbf{x}, \mathbf{w}_k)$$

.

Тогда параметры модели — конкатенация всех параметров каждой подмодели: $\mathbf{W} = = [\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathbf{V}|}]$, а структура модели $\boldsymbol{\Gamma}$ задаётся вектором $\{\boldsymbol{\gamma}^{i,j}\}_{\mathbf{E}}$.

Функция потерь на обучении L и функция потерь на валидации Q задаются как:

$$L(\mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A}) = \log p(\mathbf{Y}^{\text{train}} | \mathbf{X}^{\text{train}}, \mathbf{W}, \boldsymbol{\Gamma}) + \boldsymbol{e}^{\mathbf{A}} ||\mathbf{W}||^2,$$

$$Q(\mathbf{W}, \boldsymbol{\Gamma}) = \log p(\mathbf{Y}^{\text{valid}} | \mathbf{X}^{\text{valid}}, \mathbf{W}, \boldsymbol{\Gamma}),$$

В итоге получаем задачу двухуровневой оптимизации:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} L(\mathfrak{D}^{train}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A})$$

$$\boldsymbol{\Gamma}^*, \mathbf{A}^* = \arg\min_{\boldsymbol{\Gamma}, \mathbf{A}} Q(\mathfrak{D}^{valid}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A}).$$

Релаксация модели

Для того чтобы более эффективно решать задачу поиска оптимальной структуры нейросети, переходим от дискретной задачи поиска оптимальной базовой функции к непрерывной,

производя релаксацию структуры модели используя softmax:

$$\overline{g}^{(i,j)}(x) = \sum_{g \in \mathbb{K}} \frac{exp(\gamma_g^{(i,j)})}{\sum\limits_{\overline{g} \in \mathbb{K}} exp(\gamma_{\overline{g}}^{(i,j)})} g(x),$$

где $\gamma^{(i,j)}$ — вектор размерности $|\mathbb{K}|$, где $\mathbb{K}$ — мощность множества кандидатов на роль базовой функции. Этот вектор параметризует комбинацию базовых функций. Таким образом мы перешли к задаче поиска базовой функции, подбирая непрерывные параметры γ. В конце поиска, каждая комбинация базовых функций $\overline{g}^{(i,j)}(x)$ меняется на $g^{(i,j)} = \arg\max\limits_{g \in \mathbb{K}} \gamma_g^{(i,j)}$.

Регуляризация структуры модели

Регуляризация структуры проводится добавлением к фунции потерь Q слагаемого $\lambda P(\Gamma)$, где $P(\Gamma)$ есть произведение всех вероятностей возникновения веса $\gamma^{(i,j)}$.
Таким образом функция потерь принимает вид:

$$Q = \log \mathrm{p}(\mathbf{Y}^{valid} | \mathbf{X}^{valid}, \mathbf{W}, \mathbf{\Gamma}) + \lambda P(\Gamma)$$

В качестве вероятности для структуры можно использовать Gumble-Softmax или распределение Дирихле.

Оптимизация гиперпараметров и структурных параметров модели

Потери на валидационной и обучающей выборке обусловлены как структурой модели Γ так и параметрами модели W. Цель поиска архитектуры найти модель Γ^*, которая минимизирует ошибку на валидационной выборке $Q(w^*, \Gamma^*)$. При этом w^* находится из условия минимизации функции потерь $L(w, \Gamma^*)$.
Таким образом получается задача двухуровневой оптимизации:

$$\min_{\Gamma} \quad L(w^*(\Gamma), \Gamma)$$

$$s.t. \quad w^*(\Gamma) = \arg\min_{w} Q(w, \Gamma)$$

.
Чтобы решить эту задачу мы используем итеративную оптимизационную процедуру, в которой Γ и w оптимизируются по очереди с помощью градиентного спуска. На k-м шаге, имея структуру модели Γ_{k-1}, полуачем w_k изменяя w_{k-1} в сторону минимизации $L(w_{k-1}, \Gamma_{k-1})$. Далее, фиксируя w_k, находим Γ_k, минимизируя $Q(w_{k-1} - \xi \nabla_w L(w_{k-1}, \Gamma_{k-1}), \Gamma_{k-1})$, где ξ шаг градиентного спуска.

Литература

[1] John S. Denker Yann Le Cun and Sara A. Solla. Optimal brain damage. 1989.

[2] Alex Graves. Practical variational inference for neural networks.

[3] Weinan Zhang Yong Yu Jun Wang. Han Cai, Tianyao Chen. Efficient architecture search by network transformation. 2017.

[4] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. 2015.

[5] Yang .Y Hanxiao L., Simonyan K. Darts: Differentiable architecture search. 2018.

[6] Daniel L. Harrison Jr., Rubinfeld D. Hedonic housing prices and the demand for clean air. 1978.

[7] Christopher J.C. Burges Yann LeCun, Corinna Cortes. The mnist database of handwritten digits. 1998.

[8] G. Hinton A. Krizhevsky, V. Nair. The cifar-10 dataset. 2009.