Определение параметров нейронной сети, подлежащих оптимизации

К.Ф.Муравьев, О.Ю.Бахтеев, В.В.Стрижов
kirill.mouraviev@yandex.ru; bakhteev@phystech.edu; strijov@phystech.edu
Московский физико-технический институт

В работе рассматривается задача ускорения оптимизации параметров нейронной сети. Предлагается рассматривать градиентную оптимизацию параметров как стохастический процесс, позяоляющий получить оценки на апостериорное распределение параметров и отклонение оптимизируемых параметров от точки экстремума. Экспериментальный анализ качества алгоритма проводится на выборке рукописных цифр MNIST.
Ключевые слова: нейронные сети, методы оптимизации параметров, стохастический градиентный спуск, стохастическая динамика, precondition-матрица.

Введение

Оптимизация глубоких нейронных сетей является задачей высокой вычислительной сложности и требует больших временных затрат и вычислительных мощностей [6]. При этом оптимизация сходится по большинству параметров сети уже после небольшого числа итераций. Своевременное определение сходимости параметров позволит существенно снизить вычислительные затраты на обучение нейросетей.

Основные современные методы обучения нейронных сетей [3] работают довольно быстро и достигают хорошего качества по сравнению с обычным стохастическим градиентным спуском, но все они проводят одинаковое число шагов оптимизации для всех параметров. В работе [5] описывается байесовский подход к обучению нейронных сетей с использованием стохастической динамики Ланжевена (SGLD). В работе [4] описывается использование метода SGLD с precondition-матрицей вместо константного шага, которое ускоряет время оптимизации глубоких нейросетей до четырех раз.

В данной работе предлагается, анализируя precondition-матрицу градиентного спуска, определять параметры, для которых метод оптимизации сошелся, и удалять их из множества параметров, подлежащих дальнейшей оптимизации.

Эксперимент проводится на датасете MNIST для нескольких типов полносвязных и сверточных нейронных сетей. Проводится сравнение метода с другими методами оптимизации параметров.

Постановка задачи

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$$

Здесь $\mathbf{x}_i \in \mathbb{R}^n$ - вектор признаков для i-го объекта, $y_i \in \mathbb{Y}$ - ответ для i-го объекта.
Также задана модель $f_{\mathbf{w}_0} : \mathbb{R}^n \to \mathbb{Y}$ - нейросеть с весами $\mathbf{w}_0$.
Также даны натуральные числа η_0 - количество шагов оптимизации всех параметров, η_1 - количество шагов, на протяжении которых будет оптимизироваться выбранная часть параметров; k - количество параметров, оптимизация которых будет продолжена после η_0 шагов и l - количество параметров, веса которых не будут зануляться.
Пусть также задана функция потерь $\mathcal{L}(\mathbf{w}, \mathfrak{D})$.
Введем градиентный оператор T, описывающий один шаг градиентного спуска:

$$T(\mathbf{w}) = \mathbf{w} - \gamma \nabla L(\mathbf{w}, \mathfrak{D}'),$$

где $\mathfrak{D}'$ - случайное подмножество объектов выборки $\mathfrak{D}$.
Также для вектора $\boldsymbol{\alpha} \in \{0,1\}^N$ введем оператор

$$T|_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{w} - \gamma \boldsymbol{\alpha} \odot \nabla L(\mathbf{w}, \mathfrak{D})$$

Здесь знак $\odot$ означает поэлементное умножение векторов.
Требуется найти оптимальное множество, оптимизация которых будет продолжаться после η_0 шагов, а также оптимальное множество параметров, которые будут удалены из сети:

$$\arg \min_{\boldsymbol{\alpha} \in \{0,1\}^N, ||\boldsymbol{\alpha}||_1 = k} \min_{\boldsymbol{\beta} \in \{0,1\}^N, ||\boldsymbol{\beta}||_1 = l} \mathcal{L}(\boldsymbol{\beta} \odot (T|_{\boldsymbol{\alpha}})^{\eta_1}(T^{\eta_0}(\mathbf{w}_0)), \mathfrak{D})$$

Здесь вектор $\boldsymbol{\alpha}$ отвечает за включение/выключение параметров в оптимизацию после η_0 шагов, вектор $\boldsymbol{\beta}$ отвечает за обнуление весов нейросети.

Описание базового алгоритма

Случайный выбор параметров

Данный метод сначала случайно выбирает вектор $\boldsymbol{\alpha}$ среди всех векторов из $\{0,1\}^N$ с суммой k, затем случайно выбирает вектор $\boldsymbol{\beta}$ среди всех векторов из $\{0,1\}^N$ с суммой l.

Метод наибольших значений

Предлагается продолжить оптимизацию для k параметров с наибольшими абсолютными значениями частной производной функции потерь за последние t шагов:

$$\boldsymbol{\alpha} = \arg \max_{\boldsymbol{\alpha} \in \{0,1\}^N, ||\boldsymbol{\alpha}||_1 = k} \boldsymbol{\alpha} \cdot \sum_{i=0}^{t-1} |\nabla \mathcal{L}(T^{\eta_0 - i}(\mathbf{w}_0), \mathfrak{D})|,$$

где под $|\cdot|$ понимается поэлементная операция взятия модуля.
Далее, после $\eta_0 + \eta_1$ шагов оптимизации, предлагается оставить ненулевыми l весов с наибольшими абсолютными значениями:

$$\boldsymbol{\beta} = \arg \max_{\boldsymbol{\beta} \in \{0,1\}^N, ||\boldsymbol{\beta}||_1 = l} \boldsymbol{\beta} \cdot |(T|_{\boldsymbol{\alpha}})^{\eta_1}(T^{\eta_0}(\mathbf{w}_0))|,$$

где под $|\cdot|$ также понимается поэлементное взятие модуля.

Базовый вычислительный эксперимент

Датасет Boston Housing

Базовый эксперимент проводится на датасете Boston Housing. Датасет содержит описание 178 домов. Для каждого дома известны значения 13 признаков, описывающих дом, и его стоимость, являющаяся целевой переменной. Датасет разбивается на обучающую выборку размера 142 и отложенную выборку размера 36.

Модель

Модель представляет собой полносвязную нейросеть с одним скрытым слоем из 32 нейронов. Число ее параметров N равно 481. В качестве функции потерь выбрана среднеквадратичная ошибка. Метрикой качества является среднеквадратичное отклонение. Модель оптимизируется с помощью стохастического градиентного спуска с шагом 10^{-3}.

Оценка влияния доли параметров с отключенной оптимизацией

В данной части эксперимента исследуется зависимость качества модели от доли p параметров, для которых отключается оптимизация. Сначала проводится 500 шагов оптимизации всех параметров, затем отключается оптимизация pN параметров с наименьшими по модулю значениями градиента функции потерь (градиенты функции потерь по ним в дальнейшем приравниваются к нулю), и проводится еще 5000 шагов оптимизации. После этого измеряется среднеквадратичная ошибка предсказания модели.

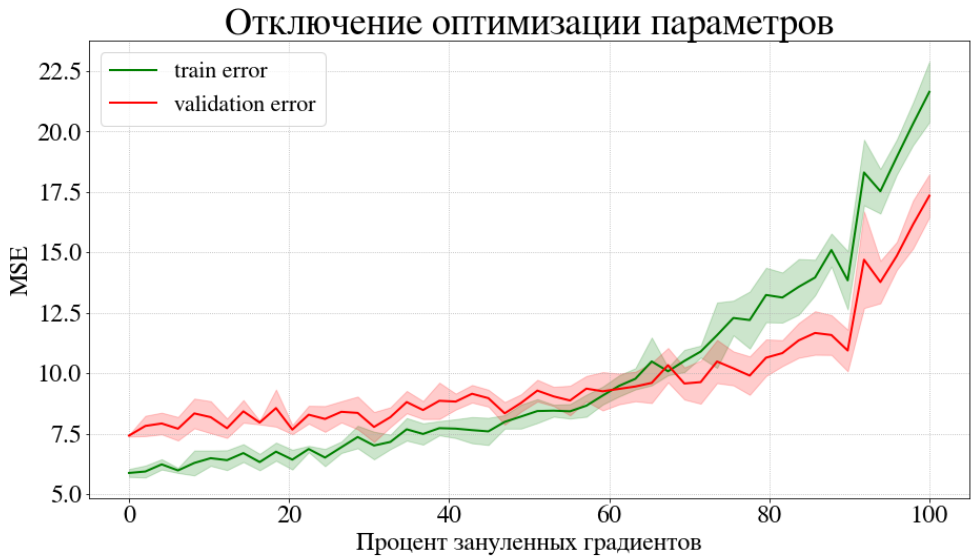

Рис. 1. Зависимость качества предсказания от доли зануленных градиентов

Видно, что при отключении оптимизации 30-40% параметров качество на отложенной выборке ненамного хуже, чем при полной оптимизации.

Оценка влияния доли зануленных весов

Исследуется зависимость качества предсказания модели от доли p зануляемых параметров. Сначала проводится 500 шагов оптимизации всех весов нейросети,

затем pN наименьших по модулю весов зануляются и проводится еще 5000 шагов оптимизации остальных весов. После этого измеряется среднеквадратичная ошибка предсказания модели на обучающей и на отложенной выборке.

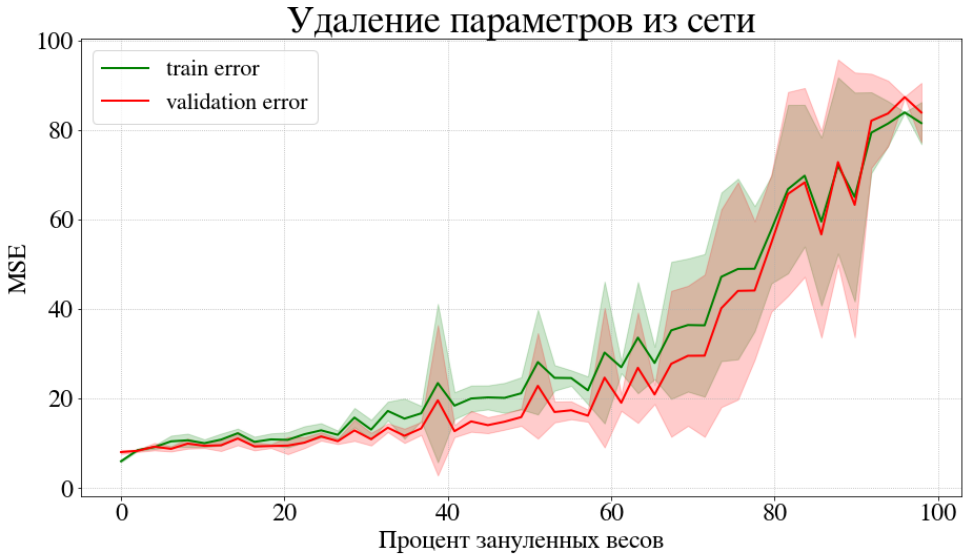

Рис. 2. Зависимость качества предсказания от доли зануленных параметров

Согласно графику, удаление из сети 20-30% параметров ухудшает качество предсказания незначительно.

Список литературы

[1] Stephan Mandt, Matthew D. Hoffman, David M. Blei: Stochastic Gradient Descent as Approximate Bayesian Inference. 2017
url: https://arxiv.org/pdf/1704.04289.pdf

[2] Alex Graves: Practical Variational Inference for Neural Networks
url: http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf

[3] Diederik P. Kingma, Jimmy Lei Ba: Adam: a Method for Stochastic Optimization. 2015
url: https://arxiv.org/pdf/1412.6980.pdf

[4] Chunyan Li, Changoyou Chen, David Carlson, Lawrence Carin: Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. 2015
url: https://arxiv.org/pdf/1512.07666.pdf

[5] Max Welling, Yee Whye Teh: Bayesian Learning via Stochastic Gradient Langevin Dynamics
url: https://arxiv.org/pdf/1702.05575.pdf

[6] Barret Zoph, Quoc V.Le: Neural Architecture Search with Reinforcement Learning. 2017
url: https://openreview.net/pdf?id=r1Ue8Hcxg