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.

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 [4]. Своевременное определение сходимости параметров позволит существенно снизить вычислительные затраты на обучение нейросетей.

Основные современные методы оптимизации нейронных сетей [3] позволяют значительно ускорить оптимизацию по сравнению со 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 метод SGLD с использованием матрицы коэффициентов вместо константного шага градиентного спуска, позволяющий ускорить время оптимизации глубоких нейросетей до четырех раз.

В данной работе предлагается использовать *precondition-матрицу* - диагональную матрицу коэффициентов, определеяющую размер шага вдоль антиградиента для параметров нейросети, вычисляемую на основе ковариационной матрицы стохастического градиента [1]. Предлагается выбирать параметры, которым соответствуют наименьшие значения в precondition-матрице, и удалять их из множества оптимизируемых параметров.

Эксперимент проводится на выборке рукописных цифр MNIST с использованием двухслойной полносвязной нейронной сети. 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 алгоритма

При оптимизации в качестве шага градиентного спуска используется диагональная precondition-матрица H:

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

Значения матрицы H вычисляются на основе ковариационной матрицы стохастического градиента C [1]:

$$H = \frac{2S}{NC},$$

где S - размер батча.
Значения ковариационной матрицы C инициализируются случайно и вычисляются динамически во время обучения по следующей рекуррентной формуле:

$$C_t = (1 - k_t)C_{t-1} + k_t(\widehat{g}_{1,t} - \widehat{g}_{S,t})(\widehat{g}_{1,t} - \widehat{g}_{S,t})^T,$$

где $\widehat{g}_{1,t}$ - это значение стохастического градиента на одном элементе выборки, а $\widehat{g}_{S,t}$ - значение стохастического градиента на батче.

В данной работе сравниваются три метода выборов параметров для отключения оптимизации: случайный выбор параметров, метод наибольших значений и выбор на основе precondition-матрицы.

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 после η_0 шагов градиентного спуска отключить оптимизацию для всех параметров, кроме k параметров с наибольшими абсолютными значениями частной производной функции потерь на шаге η_0:

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

где под $|\cdot|$ понимается поэлементная операция взятия модуля

Выбор параметров на основе ковариационной матрицы

Предлагается после η_0 шагов градиентного спуска отключать оптимизацию для всех параметров, кроме k с наибольшими по модулю произведениями коэффициента в precondition-матрице и градиента на шаге η_0:

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

Вычислительный эксперимент

Эксперимент проводится на выборке MNIST. В датасете представлены черно-белые изображения рукописных цифр размером 28 на 28 пикселей. Выборка делится на обучающую, размера 60000, и тестовую, размера 10000.

Модель представляет собой полносвязную нейросеть с одним скрытым слоем из 200 нейронов. Число ее параметров N равно 159010. В качестве функции потерь выбрана категориальная кроссэнтропия. Метрикой качества является точность классификации. Модель оптимизируется с помощью стохастического градиентного спуска с precondition-матрицей.

В ходе эксперимента изучается влияние отключения оптимизации параметров и зануления весов на качество предсказания модели.

Отключение оптимизации части параметров
В данной части эксперимента исследуется зависимость качества модели от доли p параметров, для которых отключается оптимизация. Сначала проводится 100 шагов оптимизации всех параметров, затем отключается оптимизация pN параметров и проводится еще 500 шагов оптимизации. После этого измеряется точность предсказания модели. Проводится сравнение трех методов выбора параметров: случайного, на основе precondition-матрицы и на основе градиента.

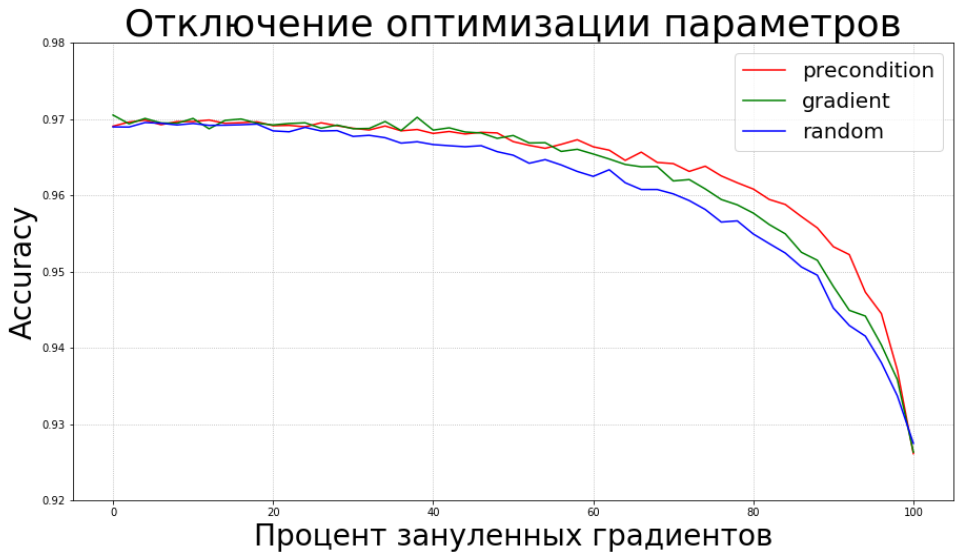

Рис. 1. Зависимость качества предсказания на отложенной выборке от доли зануленных градиентов

Зануление части весов

Исследуется зависимость качества предсказания модели от доли p зануляемых параметров. Сначала проводится 100 шагов оптимизации всех весов нейросети, затем pN весов зануляются и проводится еще 500 шагов оптимизации остальных весов. После этого измеряется точность предсказания модели. Проводится сравнение метода выбора наименьших по модулю весов со случайным выбором. Результаты представлены на рис. 2.

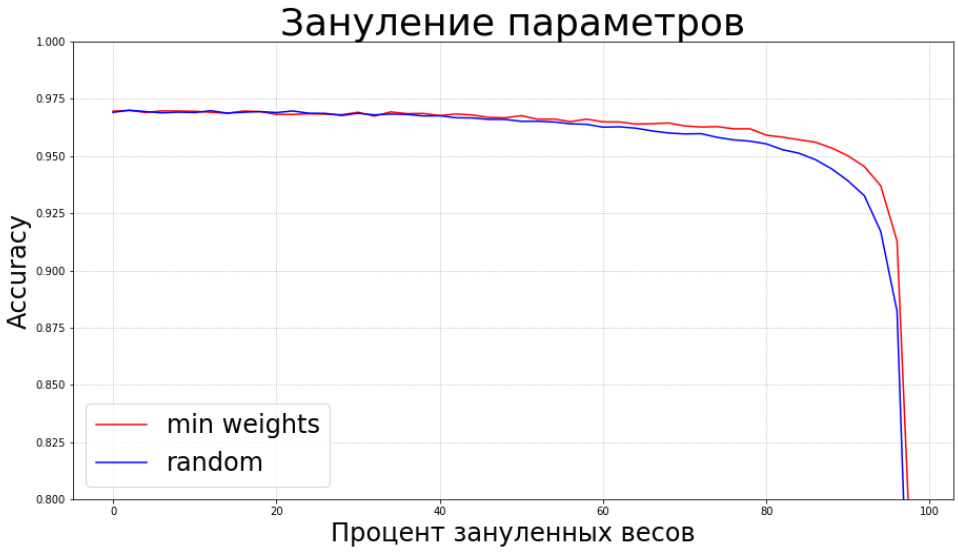

Рис. 2. Зависимость качества предсказания на отложенной выборке от доли зануленных параметров

Выводы

В ходе эксперимента было оценено качество трех методов выбора параметров, не подлежащих оптимизации: случайного выбора, выбора параметров с наи-

меньшими значениями градиента и выбора параметров на основе ковариационной матрицы. С помощью всех трех методов удалось получить качество классификации на датасете MNIST более 0.96 при отключении оптимиации 60% параметров после 100 шагов градиентного спуска. Наибольшую точность классификации в среднем показал метод, основанный на precondition-матрице.
Также было экспериментально оценено качество двух методов прореживания весов нейросети: со случайным выбором параметров и с выбором наименьших по модулю параметров. Оба метода показали на датасете MNIST точность классификации более 0.95 при 80% зануленных весов. Средняя точность классификации оказалась больше при занулении наименьших по модулю весов.

Заключение

В работе были предложены методы определения параметров нейросети, подлежащих оптимизации. Был рассмотрен случайный выбор параметров, выбор параметров с наименьшими значениями стохастического градиента и выбор параметров с наименьшими значениями градиента, помноженного на значения precondition-матрицы. Также были предложены методы прореживания нейросети: удаление случайно выбранных параметров и удаление параметров с наименьшими по модулю значениями. Для сравнения методов и оценки качества их работы был проведен вычислительный эксперимент. Наивысшее качество среди методов отключения оптимизации показал метод выбора на основе precondition-матрицы. Наивысшее качество среди методов зануления весов показал метод выбора весов с наименьшими абсолютными значениями. Реализация эксперимента на языке Python находится в свободном доступе [7].

Список литературы

[1] *Stephan Mandt, Matthew D. Hoffman, David M. Blei* Stochastic Gradient Descent as Approximate Bayesian Inference. 2017 // *Journal Of Machine Learning Research,* 2017, 18(134), pp.1-35.

[2] *Alex Graves* Practical Variational Inference for Neural Networks // *Advances in Neural Information Processing Systems 24 (NIPS 2011)*

[3] *Diederik P. Kingma, Jimmy Lei Ba* Adam: a Method for Stochastic Optimization // *arxiv.org,* https://arxiv.org/pdf/1412.6890.pdf

[4] *Chunyan Li, Changoyou Chen, David Carlson, Lawrence Carin* Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks // *Phoenix,* Arizona — February 12 - 17, 2016

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

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

[7] *Муравьев К.Ф.* экспериментальное сравнение методов определения параметров нейросети, подлежащих оптимизации // *github.com,* https://github.com/KirillMouraviev/science_publication/blob/master/code