Определение параметров нейросети, подлежащих оптимизации

Кирилл Федорович Муравьев

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 594, весна 2018

Цель исследования

Цели исследования

Получить метод определения параметров, не требующих дальнейшей оптимизации.

Проблемы

- Обучение глубоких нейросетей требует больших вычислительных ресурсов.
- Разные параметры нейросети сходятся с разной скоростью.

Было предложено

- Определять параметры, дальнейшая оптимизация которых не принесет результата, и удалять их из множества оптимизируемых параметров.
- Определять параметры, влияние которых на предсказания сети мало, и удалять их из множества весов сети.

Литература

- Alex Graves: Practical Variational Inference for Neural Networks - вероятностный подход к оптимизации нейросетей
- Max Welling, Yee Whye Teh: Bayesian Learning via Stochastic Gradient Langevin Dynamics - оптимизация с помощью стохастической динамики
- Chunyan Li, Changoyou Chen, David Carlson, Lawrence Carin: Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. 2015 - SGLD с матрицей коэффициентов вместо константного шага

Постановка задачи

- $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ - выборка; $\mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{Y}$.
- $f_{\mathbf{w}_0} : \mathbb{R}^n \to \mathbb{Y}$ - модель.
- T - градиентный оператор: $T(\mathbf{w}) = \mathbf{w} - \gamma \nabla L(\mathbf{w}, \mathfrak{D}')$.
- $\boldsymbol{\alpha}$ - вектор оптимизируемых параметров, $\boldsymbol{\beta}$ - вектор параметров, включенных в нейросеть.
- $T|_{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{w} - \gamma \boldsymbol{\alpha} \odot \nabla L(\mathbf{w}, \mathfrak{D})$ - градиентный оператор для вектора α параметров, включенных в оптимизацию.
- Ищем оптимальные вектора $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$:

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

Базовый алгоритм

Подлежат оптимизации: параметры с максимальной суммой абсолютных значений градиента функции потерь за последние t шагов:

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

Остаются в сети: параметры с наибольшими абсолютными значениями весов, веса остальных занулим:

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

Выбор параметров по ковариационной матрице

$\mathbf{C}$ - ковариационная матрица стохастического градиента:

$$\nabla L(\mathbf{w}, \mathfrak{D}') \sim \mathcal{N}(\mathbf{g}, \mathbf{C})$$

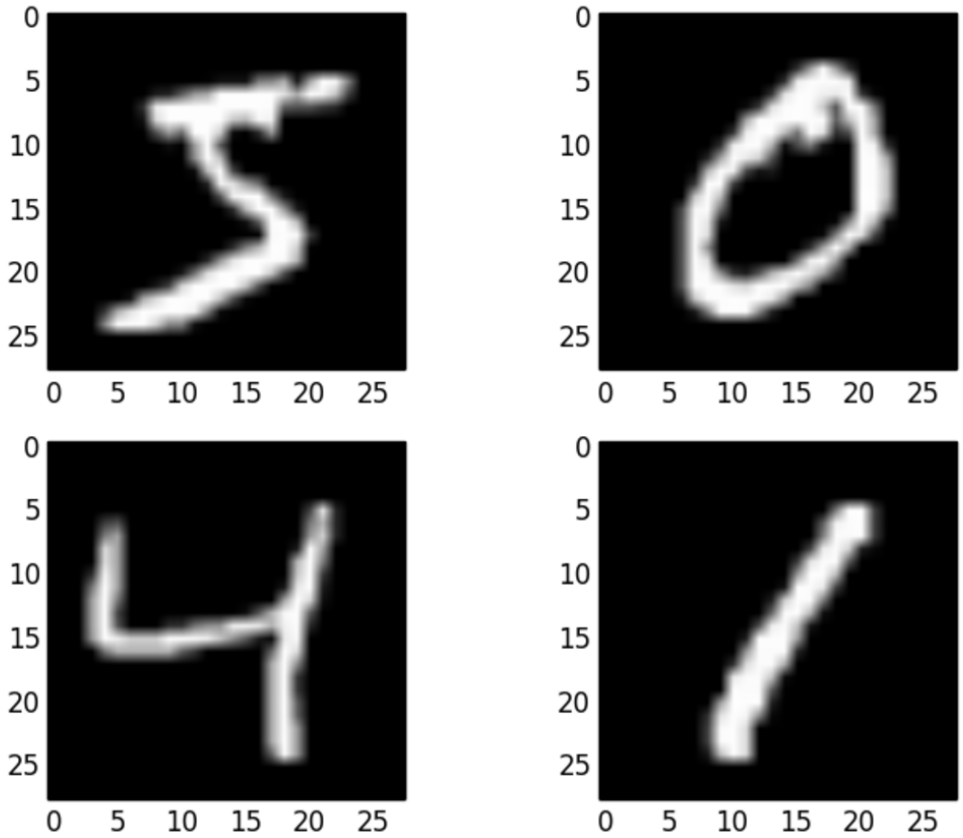
$\mathbf{H}$ - *precondition-матрица*:

$$\mathbf{H} = \frac{2S}{N\mathbf{C}};\ T(\mathbf{w}) = \mathbf{w} - \mathbf{H}\nabla L(\mathbf{w}, \mathfrak{D}')$$

Подлежат оптимизации: параметры с максимальными произведениями precondition и градиента:

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

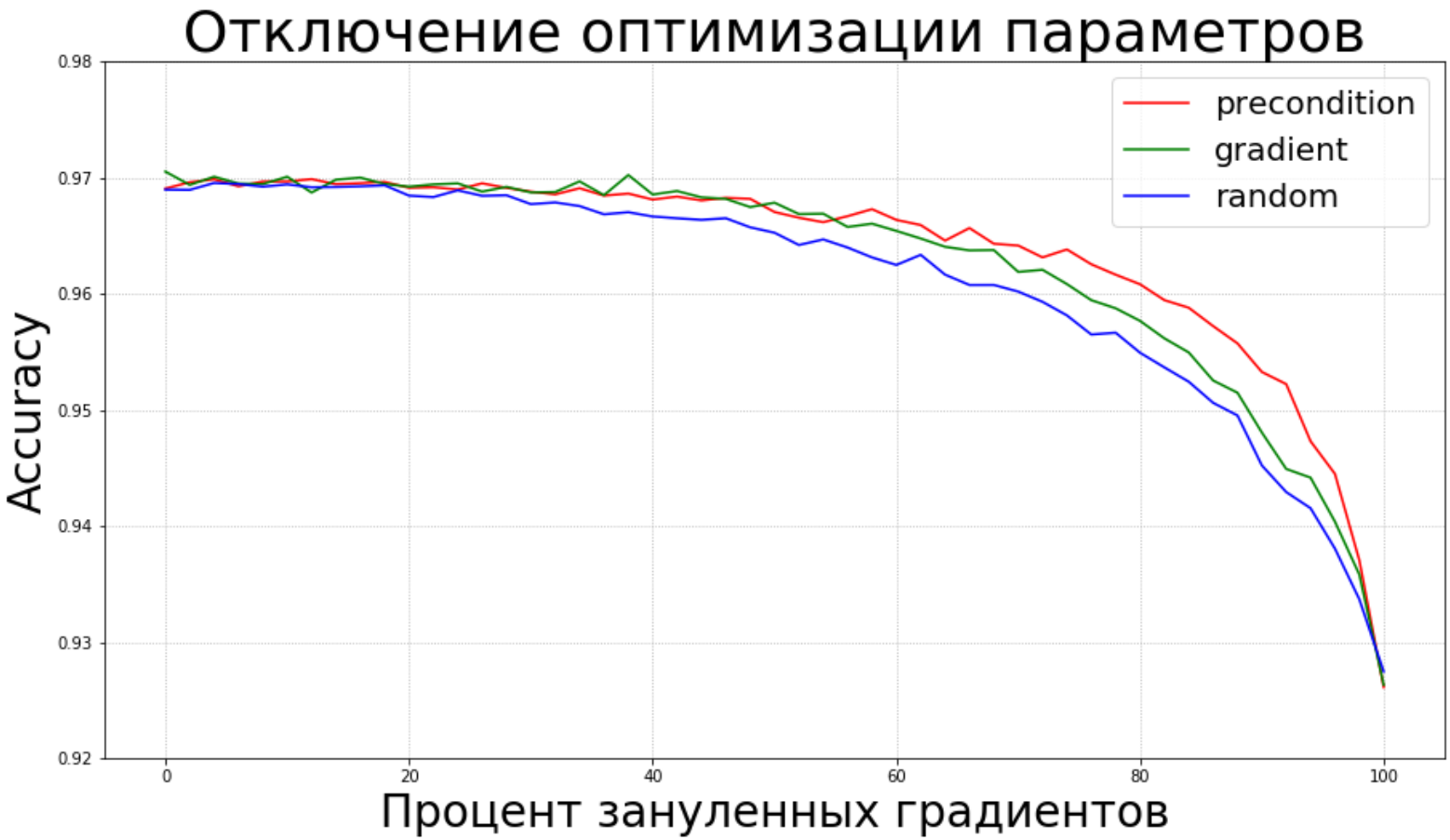
Вычислительный эксперимент: выборка MNIST



Обучающая выборка: 60000x28x28

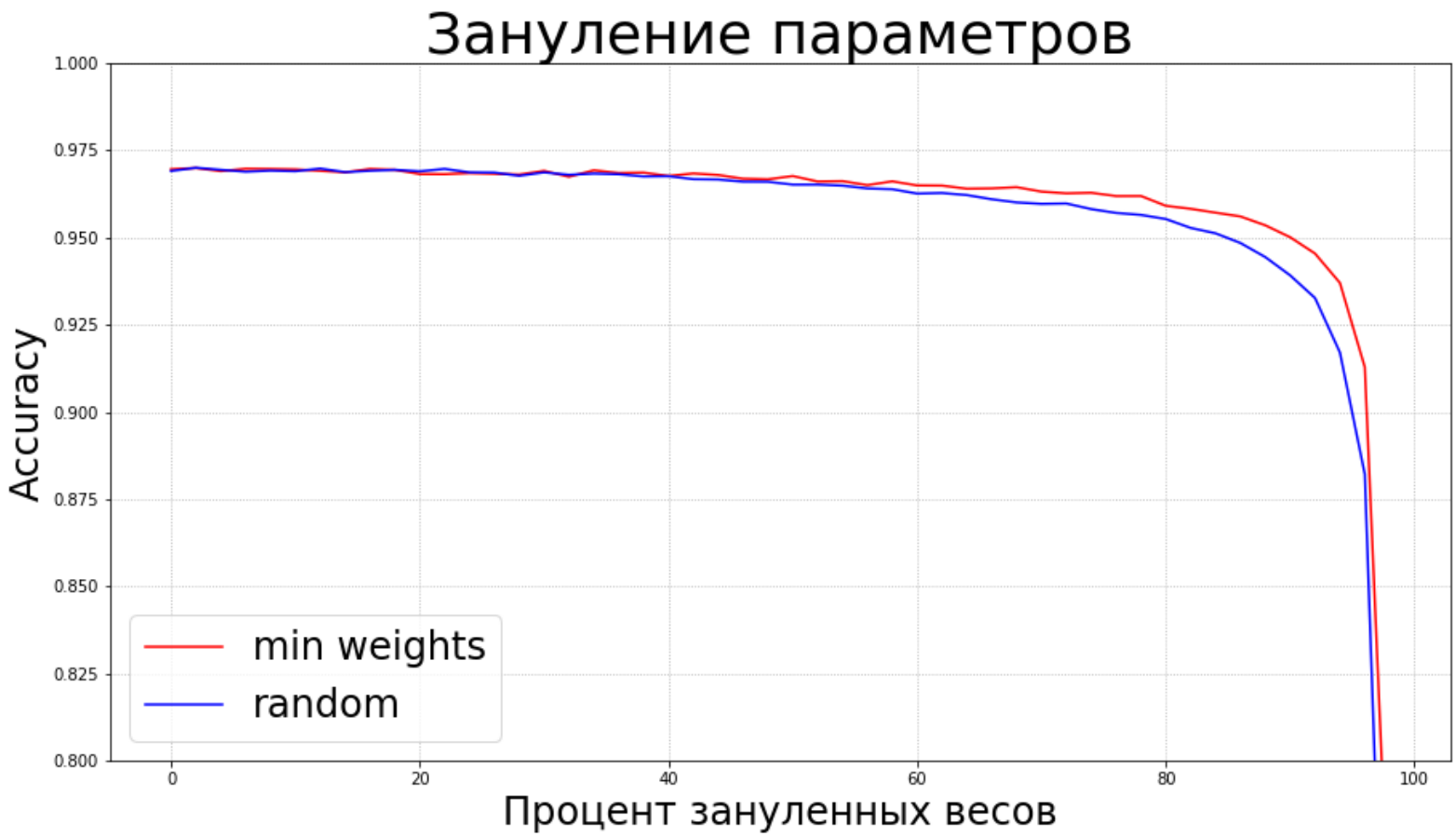
Тестовая выборка: 10000x28x28

Вычислительный эксперимент



Зависимость качества от доли параметров с отключенной оптимизацией

Вычислительный эксперимент



Зависимость качества от доли зануленных весов

Выводы

- Представлена формальная постановка задачи определения параметров сети, подлежащих оптимизации.
- Построена экспериментальная зависимость качества предсказания от доли оптимизируемых параметров и от степени прореживания нейросети.
- Проведено сравнение разных методов оптимизации.