

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

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.

Цели исследования

Получить метод определения параметров, не требующих дальнейшей оптимизации.

- Alex Graves: Practical Variational Inference for Neural Networks
- Chunyan Li, Changyou Chen, David Carlson, Lawrence Carin: Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. 2015
- Max Welling, Yee Whye Teh: Bayesian Learning via Stochastic Gradient Langevin Dynamics

- Дана выборка $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ и нейросеть $f_{\mathbf{w}_0}$.
- T - градиентный оператор: $T(\mathbf{w}) = \mathbf{w} - \gamma \nabla L(\mathbf{w}, \mathcal{D}')$.
- α - вектор оптимизируемых параметров, β - вектор параметров, включенных в нейросеть.
- $T|_{\alpha}(\mathbf{w}) = \mathbf{w} - \gamma \alpha \odot \nabla L(\mathbf{w}, \mathcal{D})$ - градиентный оператор для вектора α параметров, включенных в оптимизацию.
- Ищем оптимальные вектора α и β :

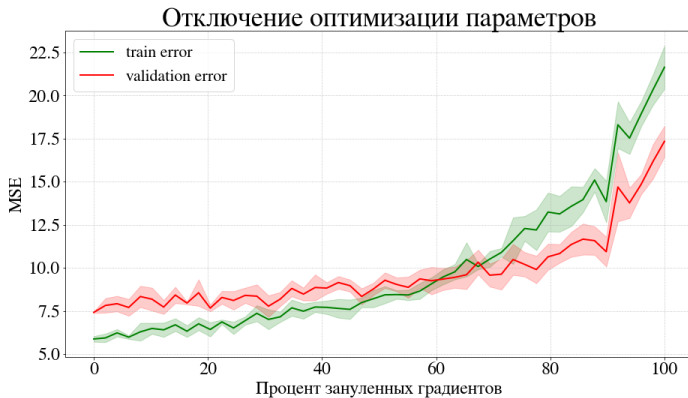
$$\arg \min_{\alpha \in \{0,1\}^N, \|\alpha\|_1=k} \min_{\beta \in \{0,1\}^N, \|\beta\|_1=l} \mathcal{L}(\beta \odot (T|_{\alpha})^{\eta_1}(T^{\eta_0}(\mathbf{w}_0)), \mathcal{D})$$

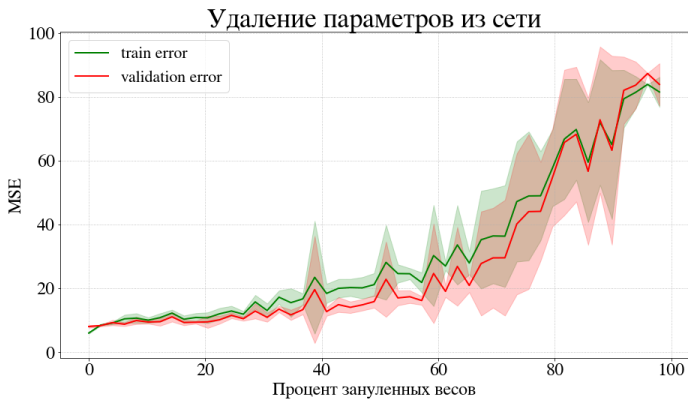
Будем считать подлежащими дальнейшей оптимизации параметры с максимальной суммой абсолютных значений градиента функции потерь за последние t шагов:

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

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

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





- Представлена формальная постановка задачи для выборочной оптимизации глубоких нейросетей.
- Построена экспериментальная зависимость качества предсказания от доли оптимизируемых параметров и от степени прореживания нейросети.
- Планируется проведение экспериментов с использованием precondition-матрицы.