Вариационная оптимизация модели глубокого обучения с контролем сложности

Гребенькова Ольга Сергеевна

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Консультант к.ф.-м.н. О. Ю. Бахтеев
Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва
2020 г.

Задача построения модели глубокого обучения

Цель

Предложить метод оптимизации модели глубокого обучения с контролем сложности модели.

Исследуемая проблема

По построению семейство моделей глубокого обучения имеет избыточное число параметров. Поэтому использование неоптимизированных моделей является вычислительно сложной задачей.

Метод решения

Предлагаемый метод заключается в представлении модели глубокого обучения в виде гиперсети, с использованием байесовского подхода. Гиперсеть — сеть, которая генерирует параметры для оптимальной модели.

- Поведения обобщенной функции обоснованности модели
- Влияния априорного распределения на сложность модели

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mathbf{w}_2^\top \mathbf{softmax}(\mathbf{w}_1^\top \mathbf{x}_1 + \mathbf{b}_1) + \mathbf{b}_2.$$

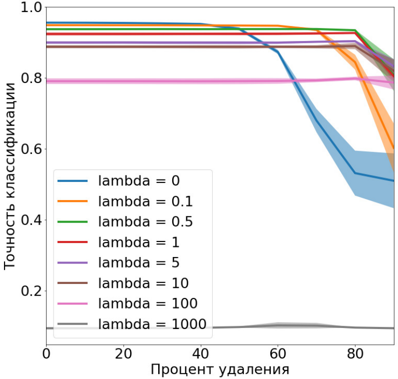


График зависимости точности классификации от процента удалённых параметров

Основная литература

Alex Graves
Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain

David Ha and Andrew M. Dai and Quoc V. Le
HyperNetworks // CoRR, vol. abs/1609.09106, 2018.

Tom Veniat and Ludovic Denoyer
Learning Time/Memory-Efficient Deep Architectures With Budgeted Super Networks // CVPR, 2018, Pp. 3492–3500.

1. выборка

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \ldots, m, \quad \mathbf{x}_i \in \mathbb{R}^m, \quad y_i \in \{1, \ldots, Y\},$$

2. модель

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}^Y,$$

   где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели,

3. априорное распределение вектора параметров в пространстве $\mathbb{W}$:

$$p(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}),$$

4. распределение, аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D})$:

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}).$$

   Предполагается, что:

$$q(\mathbf{w}) \approx p(\mathbf{w}|\mathfrak{D}) = \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D})}$$

Логарифмическая функция правдободобия выборки:

$$\mathcal{L}_{\mathfrak{D}}(\mathbf{w}|\mathfrak{D}) \propto \log p(\mathfrak{D}|\mathbf{w}).$$

Логарифм обоснованности модели:

$$\log p(\mathfrak{D}) = \log \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

При оценке интеграла получаем:

$$\mathcal{L}(\mathfrak{D}) = \log p(\mathfrak{D}) \geq \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log \frac{p(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} + \int\limits_{\mathbf{w}\in\mathbb{W}} q(\mathbf{w}) \log p(\mathfrak{D}|\mathbf{w}) d\mathbf{w} =$$

$$= \mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathbf{w}) + \mathcal{L}_E(\mathfrak{D}).$$

Обобщенная функция обоснованности: $\lambda\mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathbf{w}) - \mathcal{L}_E(\mathfrak{D})$

Максимизация функционала

$$\mathfrak{G}(\lambda) = \arg\max_{w\in\mathbb{W}}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{KL}(q(\mathbf{w})||p(\mathbf{w}))).$$

Построение гиперсети для контроля сложности модели

Гиперсеть

Параметрическое отображение из множества Λ во множество параметров модели $\mathbb{W}$:

$$\mathbf{G} : \Lambda \times \mathbb{U} \to \mathbb{W},$$

где $\mathbb{U}$ — множество допустимых параметров гиперсети, Λ — множество параметров, контролирующих сложность модели.

Реализация с отображением во множество матриц низкого ранга

$$\mathbf{G}_{\text{lowrank}}(\lambda) = (\mathbf{f}(\lambda)\mathbf{U}_1)(\mathbf{f}(\lambda)\mathbf{U}_2)^\top + \mathbf{b}_1.$$

Реализация с линейной аппроксимацией

$$\mathbf{G}_{\text{linear}}(\lambda) = \lambda\mathbf{b}_2 + \mathbf{b}_3.$$

Метод — оптимизация параметров гиперсети по случайно порожденным значениям параметра сложности λ

$$\mathsf{E}_{\lambda\sim\mathcal{U}(\Lambda)}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{KL}(q(\mathbf{w})||p(\mathbf{w}))) \to \max_{\mathbf{U}\in\mathbb{U}},$$

где $\mathcal{U}$ — равномерное распределение.

Цель

Исследовать поведение обобщенной функции обоснованности модели. Сравнить методы построения моделей напрямую без использования гиперсетей и с их использованием.

Критерий качества модели — точность классификации

$$\text{Accuracy} = 1 - \frac{1}{m}\sum_{i=1}^{m}[\mathbf{f}(\mathbf{x}_i, \mathbf{w}) \neq y_i].$$

Критерий удаления параметров — относительная плотность модели

$$\rho(w_i) \approx \exp\frac{\mu_i^2}{2\sigma_i^2}.$$

Вычислительный эксперимент: модель без гиперсети

Была рассмотрена нейросеть состоящая из двух слоёв: 100 и 10 нейронов соответственно, где второй слой отвечает за softmax-функцию.

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mathbf{w}_2^\top \mathbf{softmax}(\mathbf{w}_1^\top \mathbf{x}_1 + \mathbf{b}_1) + \mathbf{b}_2.$$

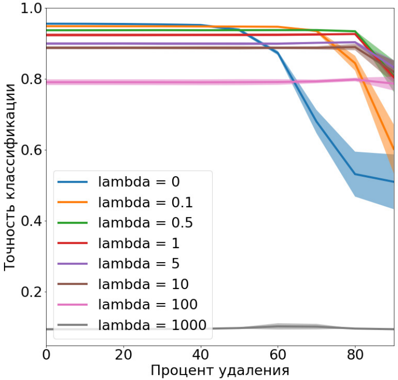
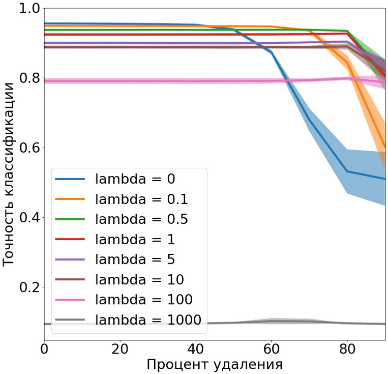


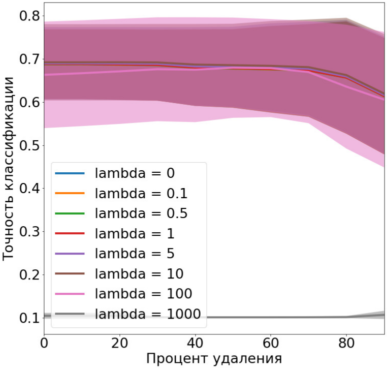
График зависимости точности классификации от процента удалённых параметров

Параметры эксперимента:

1. оптимизатор SGD;
2. 50 эпох;
3. параметр априорной дисперсии $= 0.1$;
4. для низкоранговой аппроксимации рассмотрен случай с 50 нейронами в скрытом слое и с функций активации ReLU;
5. для линейной аппроксимации каждый батч проходил процесс оптимизации с 5 разными значениями сэмплируемой λ.

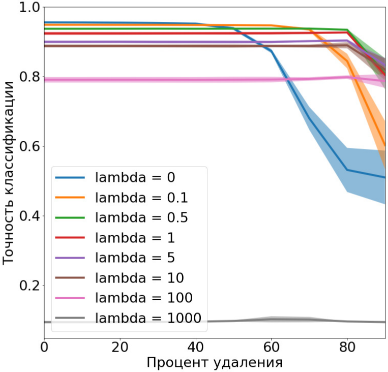


Базовая модель

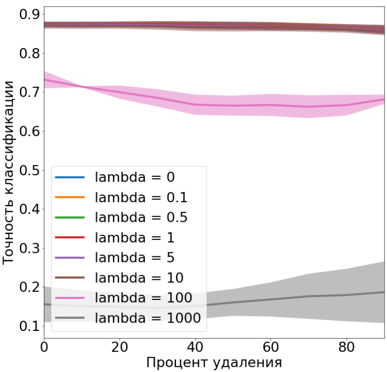


Гиперсеть с отображением во множество матриц низкого ранга

Сравнение базовой модели с гиперсетью с лин. аппроксимацией

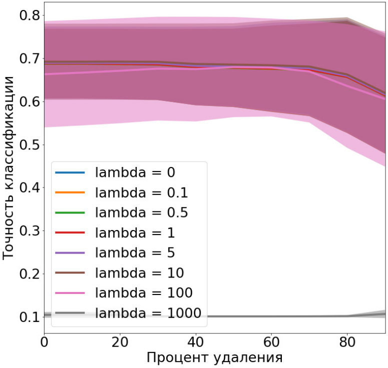


Базовая модель

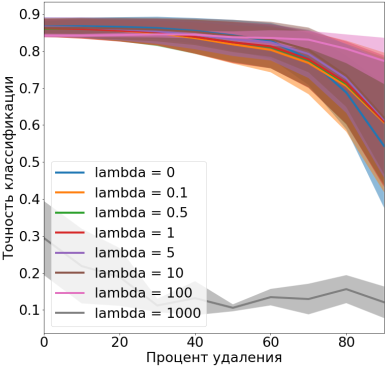


Гиперсеть с линейной аппроксимацией

Сравнение моделей после дообучения

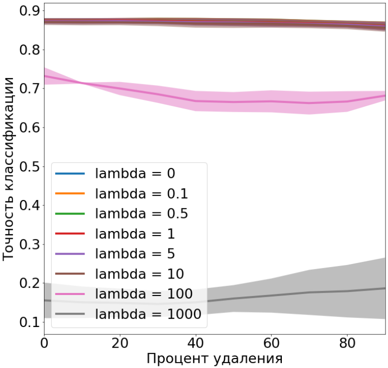


Гиперсеть с отображением во множество матриц низкого ранга

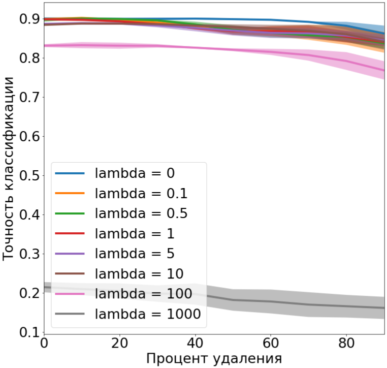


Модель с низкоранговой аппроксимацией с дообучением

Сравнение моделей после дообучения



Гиперсеть с линейной аппроксимацией



Модель с линейной аппроксимацией с дообучением

1. Вариационный метод позволяет удалить $\approx 80\%$ параметров при всех λ, кроме значений 0 и 0.1, без значительной потери точности классификации.
2. Несмотря на потерю в качестве, гиперсеть получает схожие результаты, что и обычные модели при меньших вычислительных затратах.
3. По графикам видно, что модель сохраняет схожие свойства (к примеру точность классификации) при прореживании.
4. Планируется провести эксперимент с большим количеством запусков для уточнения результатов.