Вариационная оптимизация модели глубокого обучения с контролем сложности модели

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

Решаемая проблема

По построению семейство моделей глубокого обучения имеет избыточное число параметров.

Метод решения

Предлагаемый метод заключается в представлении модели глубокого обучения в виде гиперсети, с использованием байесовского подхода. Гиперсеть — модель, которая задаёт параметры модели.

Исследование

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

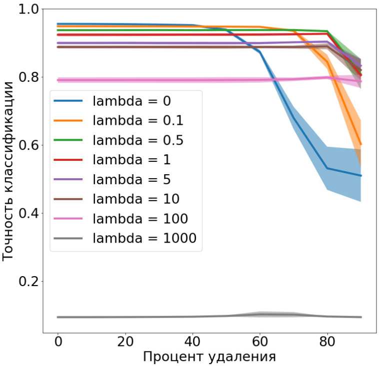


График зависимости точности классификации от процента удалённых параметров

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
$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, m, \quad \mathbf{x}_i \in \mathbb{R}^m, \quad y_i \in \{1, \dots, Y\},$$
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 $p(\mathbf{w}|\mathfrak{D})$:
$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}).$$

Постановка задачи оптимизации

Логарифмическая функция правдободобия выборки:

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

Логарифм обоснованности модели:

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

При оценке интеграла получаем:

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

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

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

Максимизация функционала

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

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.

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

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

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

Цель

Исследовать поведение обобщенной функции обоснованности модели.

Критерий качества модели — точность классификации

$$\text{Accuracy} = 1 - \frac{1}{m}\sum_{i=1}^{m}[f(x_i, w_i) \neq y_i].$$

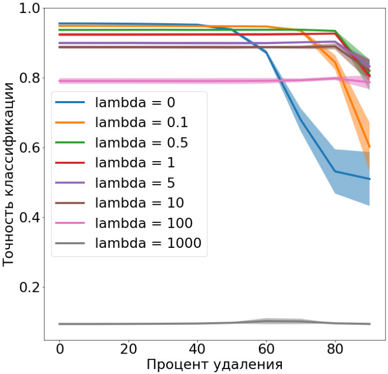
Критерий удаления параметров — относительная плотность модели

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

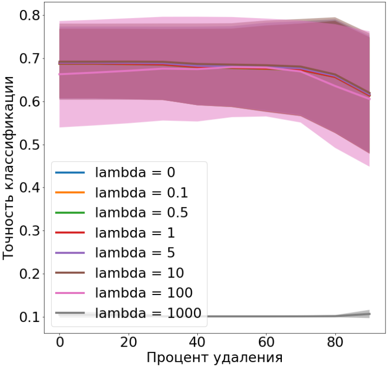
Была рассмотрена нейросеть состоящая из двух слоёв: 100 и 10 нейронов соответственно, где второй слой отвечает за softmax-функцию.

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

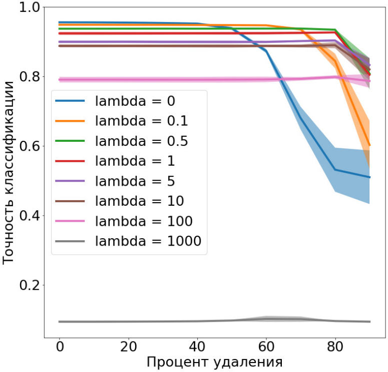
Сравнение базовой модели с низкоранговой гиперсетью



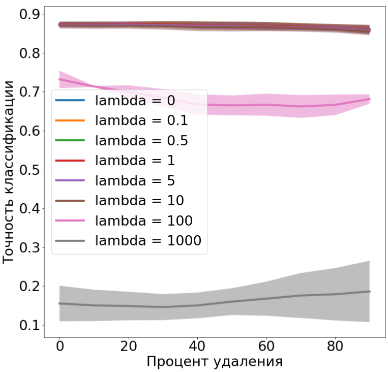
Базовая модель



Гиперсеть с отображением во множество матриц низкого ранга

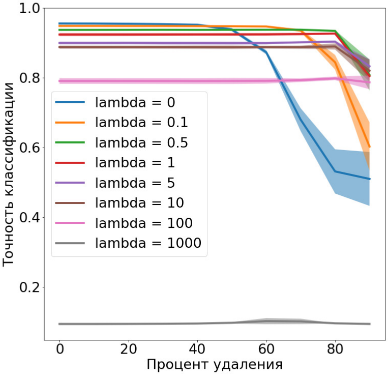


Базовая модель

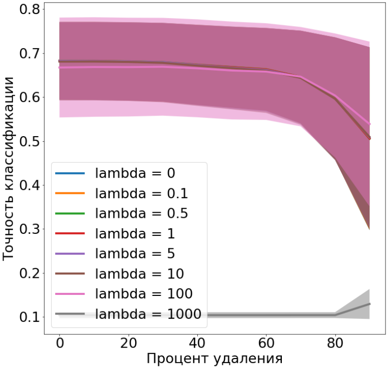


Гиперсеть с линейной аппроксимацией

Сравнение базовой модели с низкоранговой аппроксимацией и дообучением

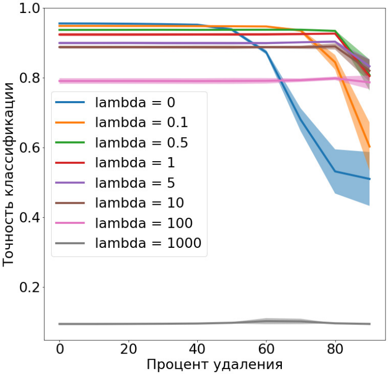


Базовая модель

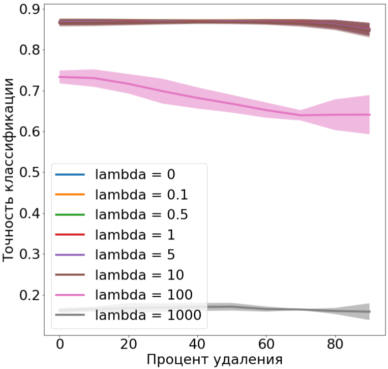


Модель с низкоранговой аппроксимацией с дообучением

Сравнение базовой модели с линейной аппроксимацией и дообучением



Базовая модель



Модель с линейной аппроксимацией с дообучением

Выносится на защиту

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.