Порождение моделей заданной сложности с использованием байесовских гиперсетей

О. С. Гребенькова

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель: Бахтеев Олег Юрьевич

Отчёт по НИР
Осень 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 оптимизация и выбор модели с наперед заданной сложностью является вычислительно сложной задачей.

Метод решения

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

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 $\mathbb{R}^n$:

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

   где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации априорного распределения;

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

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

   Здесь $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации. 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 правдоподобия выборки:

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

Логарифм обоснованности модели:

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

При оценке интеграла получаем:

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

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

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

Максимизация функционала

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

Построение гиперсети для контроля сложности модели

Гиперсеть

Параметрическое отображение из множества Λ во множество параметров модели $\mathbb{R}^n$:

$$\mathbf{G} : \Lambda \times \mathbb{R}^u \to \mathbb{R}^n,$$

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

Реализация с отображением во множество матриц низкого ранга

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

Реализация с линейной аппроксимацией

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

Дополнительно была испробована локальная репараметризация параметров. Были рассмотрены другие виды гиперсетей, пока линейная аппроксимация показывает себя лучше всего.

Цель

Исследовать поведение обобщенной функции обоснованности модели. Сравнить методы построения разных моделей. Сравнить с теоретическими результатами.

Проведено сравнение следующих моделей:

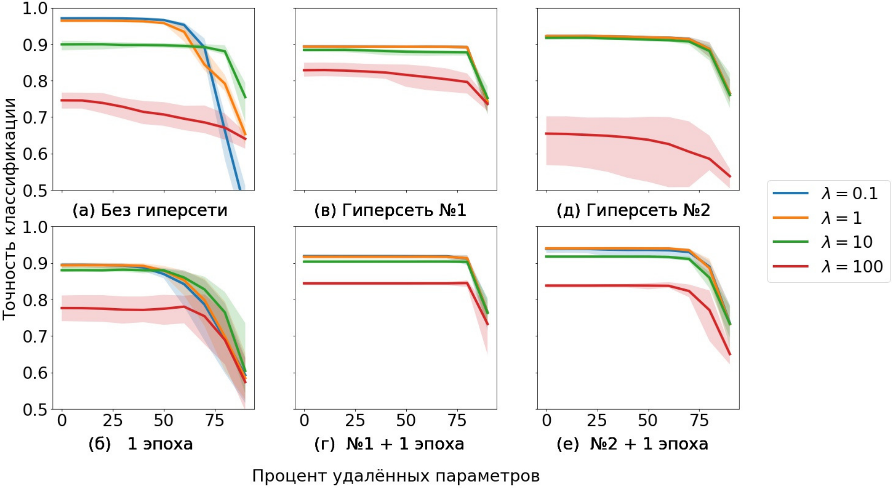
(а) построения модели напрямую без использования гиперсети;
(б) построения модели напрямую без использования гиперсети с оптимизацией за одну эпоху;
(в) построение с использованием гиперсети;
(г) построение с использованием гиперсети с дообучением итоговой модели за одну эпоху;
(д) построение с использованием гиперсети;
(е) построение с использованием гиперсети с дообучением итоговой модели за одну эпоху.
(ж) построение модели без вариационного вывода
(з) построение модели без вариационного вывода с использованием гиперсетей

Критерий удаления параметров — относительная плотность модели:

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

Критерии качества модели:

1. Точность классификации — Accuracy $= 1 - \frac{1}{m}\sum_{i=1}^{m}[f(\mathbf{x}_i, \mathbf{w}) \neq y_i]$.
2. Количество обновлений параметров модели.
3. Обобщенная обоснованность модели.
4. Стабильность модели S.


Точность классификации
1.0
0.9
0.8
0.7
0.6
0.5
(а) Без гиперсети
(в) Гиперсеть №1
(д) Гиперсеть №2
(б) 1 эпоха
(г) №1 + 1 эпоха
(е) №2 + 1 эпоха
0 25 50 75
Процент удалённых параметров
$\lambda = 0.1$
$\lambda = 1$
$\lambda = 10$
$\lambda = 100$

1. Вариационный метод позволяет удалить $\approx 60\%$ параметров при всех λ 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 исследовать свойства вариационных гиперсетей и их применение к задачи выбора модели с контролем сложности.
5. Сейчас ведется работа на базовым экспериментом для сравнения и теоретической частью диплома

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.

Jonathan Lorraine and David Duvenaud
Stochastic Hyperparameter Optimization through Hypernetworks // CoRR, vol. abs/1802.09419, 2018.