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: к.ф.-м.н. Бахтеев Олег Юрьевич

Весна 2021 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 избытычное 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.

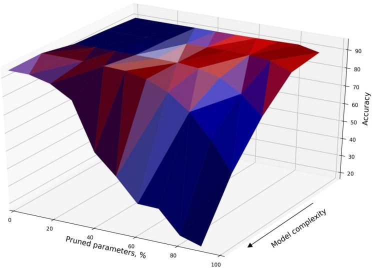


График зависимости точности классификации от процента удалённых параметров и параметра сложности модели

Заданы

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 правдобподобия 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:

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

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

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathbf{w}) = -D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w})\big).$$

$$\mathcal{L}_E = \mathbb{E}_{q(\mathbf{w})}\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}|\mathbf{w})$$

Обобщенная функция обоснованности:

$$\mathfrak{L}(\lambda) = \lambda\mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathbf{w}) + \mathcal{L}_E(\mathfrak{D}); \tag{1}$$

$$\mathfrak{L}(\lambda) = \mathcal{L}_{\mathbf{w}}(\mathfrak{D}, p(\lambda\mathbf{w}), q(\mathbf{w})) + \mathcal{L}_E(\mathfrak{D}); \tag{2}$$

$$\mathfrak{L}(\lambda) = \lambda\|\mathbf{w}\|^2 + \mathcal{L}_E(\mathfrak{D}). \tag{3}$$

Максимизация функционала

$$\mathfrak{G}(\lambda) = \arg\max_{\mathbf{w}\in\mathbb{R}^n}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{KL}(q(\mathbf{w})||p(\mathbf{w}))), \tag{4}$$

$$\mathfrak{G}(\lambda) = \arg\max_{\mathbf{w}\in\mathbb{R}^n}(\log p(\mathfrak{D}|\mathbf{w}) - D_{KL}(q(\mathbf{w})||p(\lambda\mathbf{w}))), \tag{5}$$

$$\mathfrak{G}(\lambda) = \arg\max_{\mathbf{w}\in\mathbb{R}^n}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda\|\mathbf{w}\|^2). \tag{6}$$

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 линейной аппроксимацией

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

Для аппроксимации оптимизационной задачи (4) предлагается оптимизировать параметры гиперсети $\mathbf{U} \in \mathbb{R}^u$ по случайно порожденным значениям параметра сложности $\lambda \in \Lambda$:

$$\mathbb{E}_{\lambda \sim P(\lambda)}(\log p(\mathfrak{D}|\mathbf{w}) - \lambda D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w})) \to \max_{\mathbf{U} \in \mathbb{R}^u}.$$

Прореживание параметров

Критерий удаления параметров — относительная плотность модели:

$$\rho(w_i) \propto \exp\frac{\mu_i^2}{2\sigma_i^2},$$

$$\rho(w_i) \propto \exp\left(-w_i^2\right).$$

Теорема

Пусть выполнены следующие условия:

1. существует компакт $\mathbb{U} \in \mathbb{R}^n$, который содержит единственный минимум $\mathbf{f}(\mathbf{w}^*(\lambda))$ для каждого $\lambda \in \Lambda$;
2. существует последовательность моделей $\mathbf{f}(\mathbf{w}_n(\lambda))$ такая, что $\mathbb{E}\mathfrak{L}(\mathbf{f}(\mathbf{w}_n(\lambda))) \underset{n\to\infty}{\to} \max$.

Тогда $g(\mathbf{f}(\mathbf{w}_n(\lambda))) \xrightarrow{L_1} g(\mathbf{f}(\mathbf{w}^*(\lambda)))$, где g это непрерывный критерий для удаления параметров.

Цель

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

Проведено сравнение следующих моделей:

(а) вариационная сеть;

(б) сеть с репараметризацией;

(в) базовая нейросеть с регуляризатором;

(г) вариационная сеть с линейной гиперсетью;

(д) сеть с репапраметризацией и линейной гиперсетью ;

(е) базовая нейросеть с регуляризаторс и линейной гиперсетью.

Вид используемой нейросети для эксперимента на MNIST:

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

где $\mathbf{w}_1, \mathbf{b}_1$ — параметры первого слоя, $\mathbf{w_2}, \mathbf{b}_2$ — параметры второго слоя,

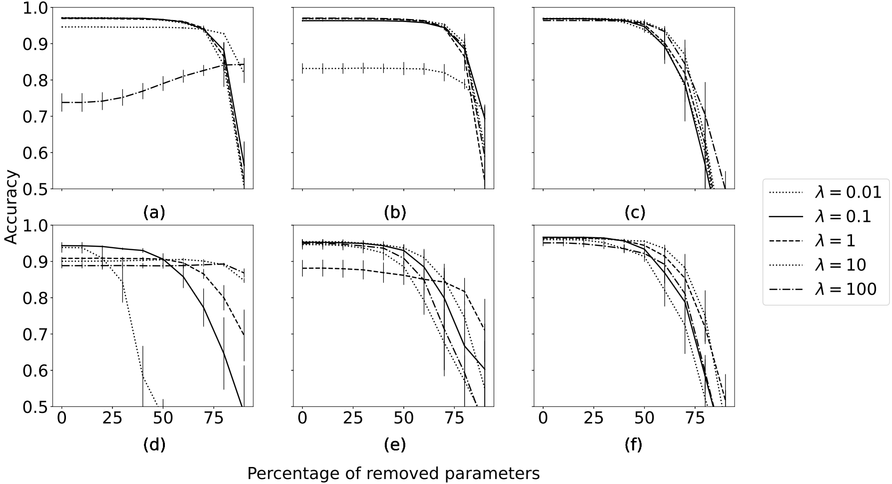
$$\mathbf{softmax}(\mathbf{x})_i = \frac{\exp{(x_i)}}{\sum_{j=1}^k \exp{(x_j)}} \quad i = 1, \ldots, k,$$

$$\mathbf{ReLU}(\mathbf{x}) = \mathbf{max}(0, \mathbf{x}).$$

Критерий качества модели — точность классификации

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

.


1.0
0.9
0.8
0.7
0.6
0.5
Accuracy
(a)
(b)
(c)
(d)
(e)
(f)
0
25
50
75
Percentage of removed parameters
$\lambda = 0.01$
$\lambda = 0.1$
$\lambda = 1$
$\lambda = 10$
$\lambda = 100$

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. Вариационная сеть запущена на CIFAR, идет настройка эксперимента. Планируется провести полноценное сравнение на CIFAR и добавить описание эксперимента в дипломную работу.

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.