

Байесовское мультимоделирование: метаоптимизация

Московский Физико-Технический Институт

2021

Метапараметры

Wikipedia

A parameter that controls the value of one or more others.

Определение

Метапараметрами λ назовем параметры оптимизационной задачи.

Чаще всего метапараметры назначаются экспертно и не подлежат оптимизации в ходе решения задачи выбора модели.

Что можно считать метапараметрами:

- параметры оператора оптимизации;
- параметры задачи оптимизации;
- структуру модели;
- функции активации слоев сети;
- вид априорного распределения и функции правдоподобия.

A neural network that embeds its own meta-levels

Предлагается разделить подмодели внутри модели сети по назначениям:

- “Normal” model: обучение и вывод.
- Evaluation model: оценка качества Q .
- Analyzing model: анализ параметров модели.
- Modifiyng model: модификация параметров.

Представлен градиентный алгоритм оптимизации нейронной сети.

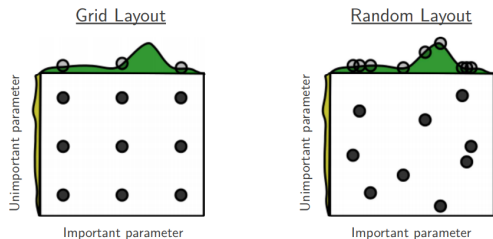
Базовые методы оптимизации гиперпараметров

Варианты:

- Поиск по решетке;
- Случайный поиск.

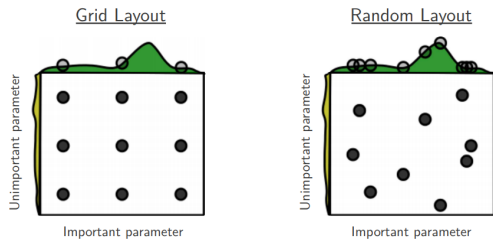
Оба метода страдают от проклятия размерности.

Случайный поиск может быть более эффективным, если пространство гиперпараметров вырождено.



Bergstra et al., 2012

Базовые методы оптимизации гиперпараметров



Bergstra et al., 2012

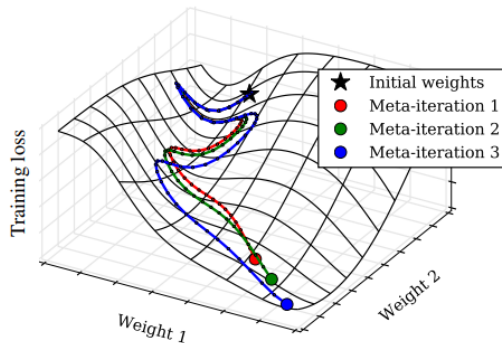
Что можно считать метапараметрами:

- параметры оператора оптимизации;
- параметры задачи оптимизации;
- структуру модели;
- функции активации слоев сети;
- вид априорного распределения и функции правдоподобия.

RMAD, Maclaurin et. al, 2015

- 1 Провести η шагов оптимизации с моментом γ : $\theta = T(\theta_0, \mathbf{h})$.
- 2 Положим $\hat{\nabla} \mathbf{h} = \nabla_{\mathbf{h}} Q(\theta, \mathbf{h})$.
- 3 Положим $d\mathbf{v} = 0$.
- 4 Для $\tau = \eta \dots 1$ повторить:
 - 5 Вычислить $\theta^{\tau-1}$.
 - 6 Вычислить градиент на шаге $\tau - 1$, используя RMD.

Алгоритм RMAD основывается на Reverse-mode differentiation.



Learning to learn by gradient descent by gradient descent

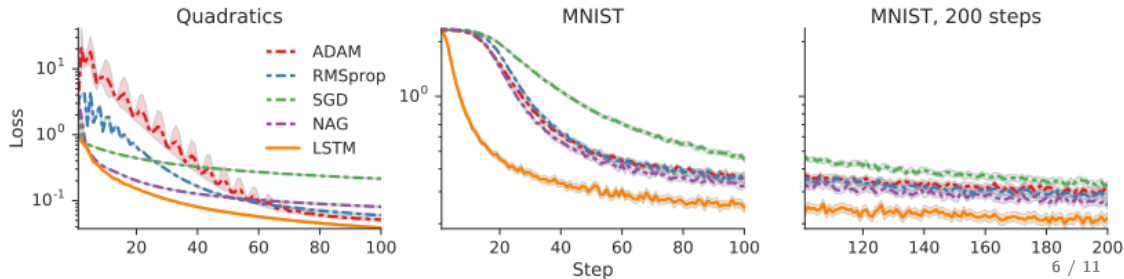
Идея: рассматривать оператор оптимизации T как дифференцируемую функцию:

$$T(\theta) = \text{LSTM}(\theta).$$

Оптимизационная задача:

$$\sum_{t=t_0}^{t_\eta} L(T^t(\theta_{t_0})) \rightarrow \max.$$

LSTM имеет небольшое число параметров и делит параметры между всеми метаметрами оператора.



Гиперсети

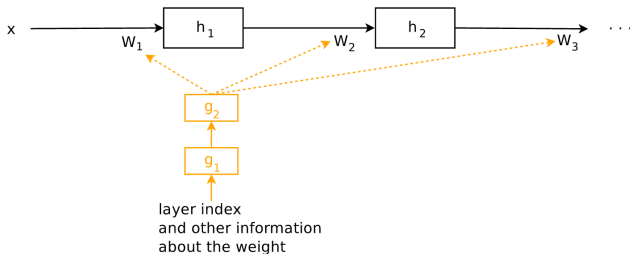
Определение

Пусть задано множество Λ .

Гиперсеть — это параметрическое отображение из множества Λ во множество параметров \mathbb{R}^n модели f :

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

где \mathbb{R}^u — множество параметров гиперсети.

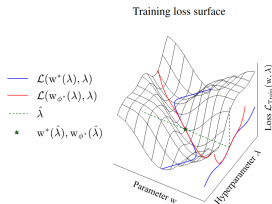


Stochastic Hyperparameter Optimization through Hypernetworks

$$E_{\lambda} \left(-\log p(\mathcal{D} | \mathbf{w}(\lambda)) + \lambda \|\mathbf{w}(\lambda)\|_2^2 \right) \rightarrow \min$$

Theorem

Sufficiently powerful hypernetworks can learn continuous best-response functions, which minimizes the expected loss for all hyperparameter distributions with convex support.



Lorraine et al., 2016

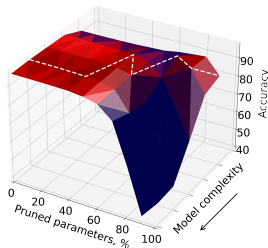
Deep learning model selection with parametric complexity control

Теорема

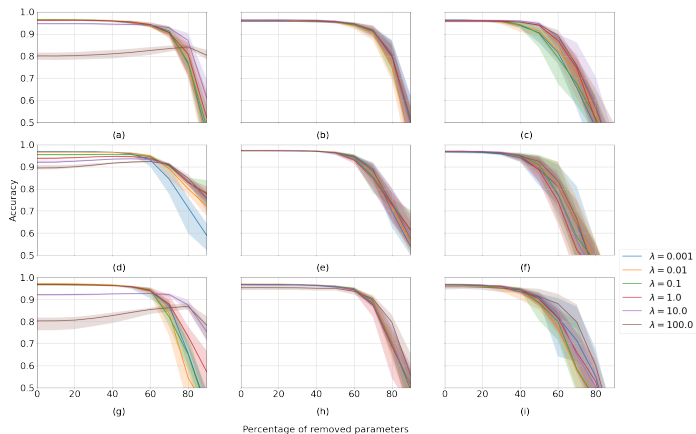
Гиперсеть позволяет не только достичь оптимальных значений аппроксимируемой модели, но также и повторить ее статистические свойства.

Следствие

Можем контролировать метапараметры, отвечающие за прунинг и сложность модели.



Deep learning model selection with parametric complexity control



Литература и прочие ресурсы

- Bergstra et al., Random Search for Hyper-Parameter Optimization, 2012
- Dougal Maclaurin et. al, Gradient-based Hyperparameter Optimization through Reversible Learning, 2015
- Andrychowicz M. et al. Learning to learn by gradient descent by gradient descent //Advances in neural information processing systems. – 2016. – С. 3981-3989.
- Bobak Shahriari et. al, Taking the Human Out of the Loop: A Review of Bayesian Optimization, 2016
- Chen Y. et al. Learning to learn without gradient descent by gradient descent //International Conference on Machine Learning. – PMLR, 2017. – С. 748-756.
- Ha D., Dai A., Le Q. V. Hypernetworks //arXiv preprint arXiv:1609.09106. – 2016.
- Lorraine J., Duvenaud D. Stochastic hyperparameter optimization through hypernetworks //arXiv preprint arXiv:1802.09419. – 2018.
- Гребенькова О. С., Бахтеев О. Ю., Стрижов В. В. Вариационная оптимизация модели глубокого обучения с контролем сложности //Информатика и её применения. – 2021. – Т. 15. – №. 1. – С. 42-49.