Поиск согласованных нейросетевых моделей в задаче мультидоменного обучения

К. Д. Яковлев[1] О. Ю. Бахтеев[1,2] В. В. Стрижов[1,2]

{iakovlev.kd, bakhteev, strijov}@phystech.edu

[1]Москва, Московский физико-технический институт

[2]Москва, Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

2023

Цель исследования

Цель

Предложить градиентный метод оптимизации гиперпараметров с ассимптотически несмещенным гиперградиентом.

Проблема

Существующие методы не гарантируют ассимптотически несмещенный градиент без увеличения вычислительных затрат.

Метод решения

Предлагаемый метод основан на аггрегации жадных гиперградиентов без дополнительных вычислительных затрат.

Постановка задачи оптимизации гиперпараметров

- Пусть задан вектор параметров модели $\mathbf{w} \in \mathbb{R}^p$ и вектор гиперпараметров $\boldsymbol{\alpha} \in \mathbb{R}^h$. задача оптимизации:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathcal{L}_2(\mathbf{w}^*, \boldsymbol{\alpha}),$$
$$\text{s.t.} \quad \mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}_1(\mathbf{w}, \boldsymbol{\alpha}).$$

- Пусть внутренняя задача решается с помощью оптимизатора $\boldsymbol{\Phi}(.,.)$:

$$\mathbf{w}_{t+1}(\boldsymbol{\alpha}) = \boldsymbol{\Phi}(\mathbf{w}_t, \boldsymbol{\alpha}), \quad t = \overline{1, T}.$$

- Гиперградиент запишется как:

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L}_2(\mathbf{w}_T(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \mathcal{L}_2(\mathbf{w}_T(\boldsymbol{\alpha}), \boldsymbol{\alpha}) + \sum_{t=1}^{T} \mathbf{B}_t \mathbf{A}_{t+1} \ldots \mathbf{A}_T \frac{\partial \mathcal{L}_2(\mathbf{w}_T(\boldsymbol{\alpha}), \boldsymbol{\alpha})}{\partial \mathbf{w}},$$
$$\mathbf{B}_t = \frac{\partial \boldsymbol{\Phi}(\mathbf{w}_{t-1}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}, \quad \mathbf{A}_t = \frac{\partial \boldsymbol{\Phi}(\mathbf{w}_{t-1}, \boldsymbol{\alpha})}{\partial \mathbf{w}_{t-1}}.$$

Аппроксимация гиперградиента

- Пусть задано $\gamma \in (0, 1)$. Тогда аппроксимация гиперградиента запишется как:

$$\hat{\nabla}_{\boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{\alpha}} \mathcal{L}_2(\mathbf{w}_T(\boldsymbol{\alpha}), \boldsymbol{\alpha}) + \sum_{t=1}^{T} \mathbf{B}_t \frac{\partial \mathcal{L}_2(\mathbf{w}_t, \boldsymbol{\alpha})}{\partial \mathbf{w}_t} \gamma^{T-t}.$$

- предлположения:
  1. $\mathcal{L}_1(., \boldsymbol{\alpha})$, $\mathcal{L}_2(., \boldsymbol{\alpha})$ являются L-гладкими и μ-сильно выпуклыми.
  2. $\frac{\partial^2 \mathcal{L}_1(., \boldsymbol{\alpha})}{\partial \mathbf{w} \partial \mathbf{w}^\top}$ является H_w-липшицева.
  3. $1 - \eta L \leq \gamma 1 - \eta \mu$
  4. $\|\frac{\partial \mathcal{L}_1(\mathbf{w}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \mathbf{w}^\top}\| \leq B$.
  5. $\frac{\partial^2 \mathcal{L}_1(., \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \mathbf{w}^\top}$ является M_b-липшицевой.
  6. $(\frac{\partial^2 \mathcal{L}_1(., \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \mathbf{w}^\top})^\top (\frac{\partial^2 \mathcal{L}_1(., \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \mathbf{w}^\top}) \succeq \kappa \mathbf{I}$.

Ассимптотическая несмещенность гиперградиента

Теорема (Яковлев, 2023)

Пусть выполнены предположения (1-6). Тогда:

$$\|\hat{\nabla}_{\boldsymbol{\alpha}} - \nabla_{\boldsymbol{\alpha}}\|_2 \leq \frac{2LB\|\mathbf{w}_0 - \mathbf{w}_*\|\sqrt{1-\eta\mu}^T}{\sqrt{1-\eta\mu}^{-1} - 1} + B\|\frac{\partial \mathcal{L}_2(\mathbf{w}_T, \boldsymbol{\alpha})}{\partial \mathbf{w}}\| \cdot$$

$$\left[\frac{1}{\eta}\left(\frac{1}{\mu} - \frac{1}{L} + \frac{1}{L}(1-\eta\mu)^T\right) + 2\eta H_w\left((T-1)\sqrt{1-\eta\mu}^T - \frac{\sqrt{1-\eta\mu}^{T-1} - (1-\eta\mu)^T}{\sqrt{1-\eta\mu}^{-1} - 1}\right)\right].$$

Теорема (Яковлев, 2023)

Предложенный градиентный метод оптимизации гиперпараметров предоставляет ассимптотически точный гиперградиент с ростом номера эпохи внутренней оптимизации.

Постановка вычислительного эксперимента

- Цель – сравнение качества предложенного подхода с существующими методами подсчета гиперградиента.
- Эксперимент проводится на задаче очистки обучающей выборки. Приводится точность предсказания на отложенной выборке.
- Сравниваются следующие методы: DrMAD, IFT, Truncated Backpropagation.

Method	**Valid. Acc.**	**#JVPs**
Truncated backpropagation (Lukethina)	72.5	1 (1)
DrMAD	69.8	99 $(2T-1)$
IFT(9, 5)	70.3	50 $((N+1)K)$
IFT(4, 10)	70.7	50 $((N+1)K)$
Proposed ($\gamma = 0.99$)	**73.5***	50 (T)

Из таблицы видно, что предложенный метод превосходит существующие методы оптимизации гиперпараметров в терминах точности предсказаний на отложенной выборке, имея сопоставимые вычислительные затраты.

Заключение

- Рассмотрена задача оптимизации гиперпараметров.
- Предложен метод оптимизации гиперпараметров с ассимптотически несмещенным гиперградиентом.
- Продемонстрирована работоспособность предлагаемого решения.