Регуляризация траектории оптимизации параметров модели глубокого обучения на основе дистилляции знаний

Мария Горпинич

Московский физико-технический институт

Курс: Автоматизация научных исследований
(практика, В. В. Стрижов)/Группа 874

Эксперт: В. В. Стрижов

Консультант: О. Ю. Бахтеев

2021

Задача дистилляции знаний

Цель

Предложить метод назначения метапараметров в задаче обучения с применением дистилляции знаний.

Исследуемая проблема

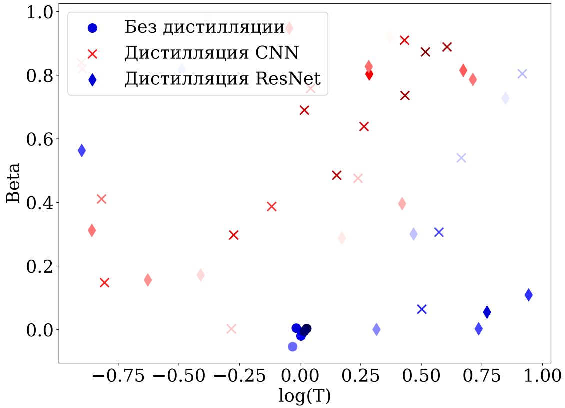
Назначение метапараметров задачи дистилляции является плохо исследуемой задачей.

Метод решения

Предлагается рассмотреть задачу как двухуровневую задачу оптимизации. Решение задачи оптимизации метапараметров производится градиентными методами. Для ускорения вычислительно затратной процедуры оптимизации метапараметров производится прогнозирование локально-линейными моделями.

Оптимизация параметров модели на основе дистилляции знаний

Назовем дистилляцией знаний задачу оптимизации параметров модели прогнозирования, при которой учитывается не только информация, содержащаяся в выборке, но также и информация, содержащаяся в сторонней модели (модели-учителе).



Основная литература

- Geoffrey E. Hinton, Oriol Vinyals и Jeffrey Dean. “Distilling the Knowledge in a Neural Network”. в: *CoRR* abs/1503.02531 (2015). URL: http://arxiv.org/abs/1503.02531.
- Jelena Luketina и др. “Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters”. в: *CoRR* abs/1511.06727 (2015). URL: http://arxiv.org/abs/1511.06727.
- Oleg Yu. Bakhteev и Vadim V. Strijov. “Comprehensive analysis of gradient-based hyperparameter optimization algorithms”. в: *Ann. Oper. Res* 289.1 (2020), с. 51—65.
- Marcin Andrychowicz и др. “Learning to learn by gradient descent by gradient descent”. в: *CoRR* abs/1606.04474 (2016). URL: http://arxiv.org/abs/1606.04474.

Постановка задачи

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}, \ \mathbf{x}_i \in \mathbb{R}^n, \ y_i \in \mathbb{Y} = \{1, \dots, K\}, \qquad \mathfrak{D} = \mathfrak{D}_{\text{train}} \sqcup \mathfrak{D}_{\text{val}}$$

Дистилляция

$$\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}) = -\beta_1 \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\textbf{train}}} \underbrace{\sum_{k=1}^{K} y^k \log \mathbf{g}(\mathbf{x}, \mathbf{w})\big|_{T=1}}_{\text{исходная функция потерь}} - \beta_2 \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\textbf{train}}} \underbrace{\sum_{k=1}^{K} \mathbf{f}(\mathbf{x})\big|_{T=T_0} \log \mathbf{g}(\mathbf{x}, \mathbf{w})\big|_{T=T_0}}_{\text{слагаемое дистилляции}}$$

где $\mathbf{f}$ — модель учителя, $\mathbf{g}$ — модели ученика, $\boldsymbol{\lambda} = [\beta_1, \beta_2, T]$ — множество метапараметров.

Постановка задачи

Итоговая оптимизационная задача:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^2} \mathcal{L}_{\text{val}}(\hat{\mathbf{w}}, \boldsymbol{\lambda})$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^s} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda})$$

Градиентные методы оптимизации

Оптимизационную задачу решает оператор градиентного спуска:

$$U(\mathbf{w}, \boldsymbol{\lambda}) = \mathbf{w} - \gamma \nabla \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}).$$

Будем обновлять метапраметры последовательно согласно следующему правилу:

$$\boldsymbol{\lambda}' = \boldsymbol{\lambda} - \gamma_{\boldsymbol{\lambda}} \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\text{val}}(U(\mathbf{w}, \boldsymbol{\lambda}), \boldsymbol{\lambda}) = \boldsymbol{\lambda} - \gamma_{\boldsymbol{\lambda}} \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\text{val}}(\mathbf{w} - \gamma \nabla \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}), \boldsymbol{\lambda}).$$

Гипотеза: траекторию градиентной оптимизации можно аппроксимировать локально-линейной моделью

Вычислительный эксперимент

Цель эксперимента

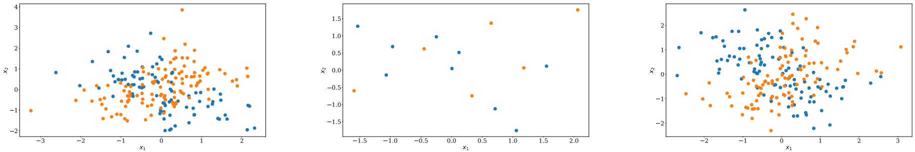
Анализ градиентной оптимизации и проверка гипотезы об аппроксимации траектории оптимизации локально-линейной моделью.

Выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}, \ x_{ij} \in \mathcal{N}(0,1), \ j = 1, 2, x_{i3} = [\mathrm{sign}(x_{i1}) + \mathrm{sign}(x_{i2}) > 0]$$

$$y_i = \mathrm{sign}(x_{i1} * x_{i2} + \delta) \in \mathbb{Y}$$

Размер выборки модели-ученика намного меньше размера выборки модели-учителя.



а) **б)** **в)**

Визуализация выборки а) учителя; б) ученика; в) тестовой

Вычислительный эксперимент

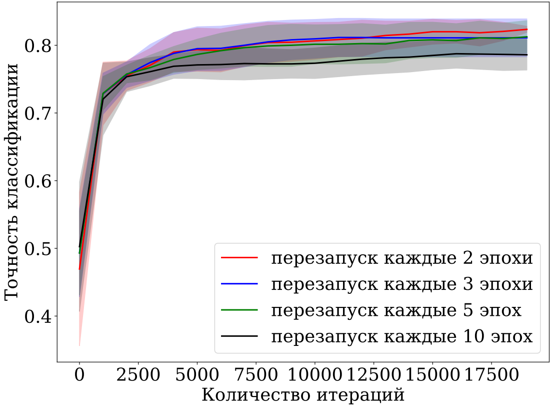
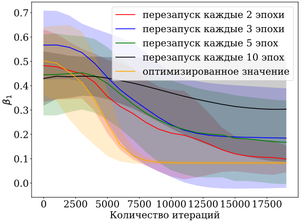


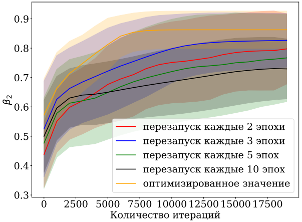
График зависимости точности классификации от номера итерации при различном количестве перезапусков

Вычислительный эксперимент

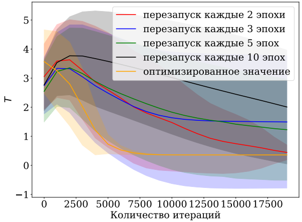
Графики зависимости значений метапараметров от номера итерации: а) β_1; б) β_2; в) температуры



а)



б)



в)

Вычислительный эксперимент

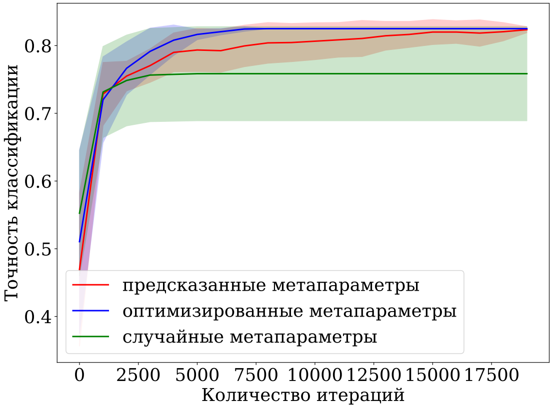


График зависимости точности классификации от номера итерации

Заключение

- исследовано применение градиентных методов оптимизации для метапараметров задачи дистилляции
- предложена и проверена гипотеза по аппроксимации траектории оптимизации метапараметров
- вычислительный эксперимент показал, что оптимизация метапараметров применима к задаче дистилляции;
- подтверждена возможность аппроксимации метапараметров локально-линейными моделями
- планируется дальнейшее исследование оптимизационной задачи и анализ качества аппроксимации траектории оптимизации метапараметров более сложными прогностическими моделями.