Оптимизация метапараметров в задаче дистилляции знаний

Горпинич М., Бахтеев О. Ю., Стрижов В. В.

Московский физико-технический институт (государственный университет)

2021

Дистилляция знаний

Цель

Предложить метод оптимизации метапараметров для задачи дистилляции. Метапараметры — это параметры оптимизационной задачи.

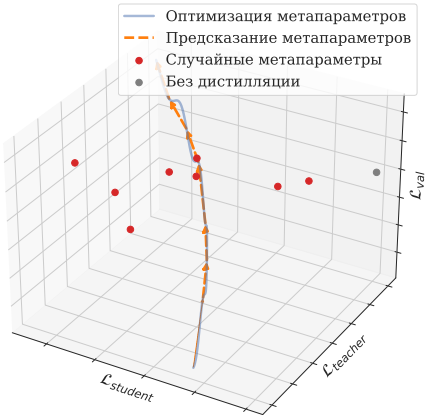
Проблема

Задача подбора метапараметров является вычислительно затратной. Однако, правильное назначение метапараметров значительно повышает качество модели.

Решение

Рассмотрим двухуровневую задачу оптимизации. Данная задача решается градиентными методами. Для уменьшения вычислительной сложности задачи значения метапараметров предсказываются с помощью линейных моделей.

Ключевая идея метода



Метапараметры задают значение функции потерь для рассмотренной модели:

$$\mathcal{L}_{\text{train}} = \lambda_1 \mathcal{L}_{\text{student}} + (1 - \lambda_1)\mathcal{L}_{\text{teacher}}.$$

Вместо непосредственной оптимизации значений метапараметров анализируется поведение траектории оптимизации, которая предсказывается с помощью линейных моделей.

Постановка задачи

Метапараметрами $\boldsymbol{\lambda}$ в задаче дистилляции являются коэффициенты слагаемых в функции потерь и температура:

$$\boldsymbol{\lambda} = [\lambda_1, T].$$

Температура является множителем логитов моделей в функции softmax.
Дистилляция знаний является задачей оптимизации параметров модели. Она учитывает:

1. информацию исходной выборки;
2. информацию, содержащуюся в модели-учителе.

Модель-учитель имеет более сложную структуру. Она обучается на исходной выборке. Модель-ученик имеет более простую структуру. Она оптимизируется путем переноса знаний модели-учителя.

Постановка задачи дистилляции

Дана выборка

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

$\mathbf{f}$ — фиксированная модель-учитель, $\mathbf{g}$ — модель-ученик.

Определение 1. Пусть функция $D : \mathbb{R}^s \to \mathbb{R}_+$ определяет расстояние между моделью-учеником $\mathbf{g}$ и моделью-учителем $\mathbf{f}$. Назовем D-дистилляцией модели-ученика такую задачу оптимизации параметров модели ученика, которая минимизирует функцию D.

Утверждение 1. Если $\lambda_1 = 0$, то минимизируется функция потерь, являющаяся D-дистилляцией с $D = D_{KL}\left(\sigma\left(\mathbf{f}/T\right), \sigma\left(\mathbf{g}/T\right)\right)$, где σ — функция softmax.

Функции потерь

Функция потерь на обучении

$$\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}) = -\lambda_1 \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\textbf{train}}} \underbrace{\sum_{k=1}^{K} y^k \log \frac{e^{\mathbf{g}(\mathbf{x},\mathbf{w})_k}}{\sum\limits_{j=1}^{K} e^{\mathbf{g}(\mathbf{x},\mathbf{w})_j}}}_{\text{исходная функция потерь}} - (1-\lambda_1) \sum_{(\mathbf{x},y)\in\mathfrak{D}_{\textbf{train}}} \underbrace{\sum_{k=1}^{K} \frac{e^{\mathbf{f}(\mathbf{x})_k/T}}{\sum\limits_{j=1}^{K} e^{\mathbf{f}(\mathbf{x})_j/T}} \log \frac{e^{\mathbf{g}(\mathbf{x},\mathbf{w})_k/T}}{\sum\limits_{j=1}^{K} e^{\mathbf{g}(\mathbf{x},\mathbf{w})_j/T}}}_{\text{слагаемое дистилляции}},$$

Валидационная функция потерь

$$\mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\lambda}) = -\sum_{(\mathbf{x},y)\in\mathfrak{D}_{\textbf{val}}} \sum_{k=1}^{K} y^k \log \frac{e^{\mathbf{g}(\mathbf{x},\mathbf{w})_k/T_{\textbf{val}}}}{\sum\limits_{j=1}^{K} e^{\mathbf{g}(\mathbf{x},\mathbf{w})_j/T_{\textbf{val}}}}$$

Множество метапараметров:

$$\boldsymbol{\lambda} = [\lambda_1, T]$$

Задача оптимизации:

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

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

Градиентная оптимизация

Определение 2. Назовем *оператором оптимизации* алгоритм U выбора вектора параметров $\mathbf{w}'$ с использованием параметров на предыдущем шаге $\mathbf{w}$:

$$\mathbf{w}' = U(\mathbf{w}).$$

Оптимизируем параметры $\mathbf{w}$ используя η шагов оптимизации:

$$\hat{\mathbf{w}} = U \circ U \circ \cdots \circ U(\mathbf{w}_0, \boldsymbol{\lambda}) = U^{\eta}(\mathbf{w}_0, \boldsymbol{\lambda}),$$

где $\mathbf{w}_0$ — начальное значение вектора параметров $\mathbf{w}$.
Переопределим задачу оптимизации используя определение оператора U:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^3} \mathcal{L}_{\text{val}}\big(U^{\eta}(\mathbf{w}_0, \boldsymbol{\lambda})\big), \qquad U(\mathbf{w}, \boldsymbol{\lambda}) = \mathbf{w} - \gamma \nabla \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}).$$

Будем обновлять метапараметры последовательно по 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: в случае градиентной оптимизации траектория оптимизации может быть предсказана локально линейными моделями:

$$\boldsymbol{\lambda}' = \boldsymbol{\lambda} + \mathbf{c}^{\top} \begin{pmatrix} z \\ 1 \end{pmatrix},$$

где $\mathbf{c}$ — вектор параметров линейной модели.

Итоговый алгоритм

Algorithm 1 Оптимизация метапараметров

Require: число e_1 итераций с использованием градиентной оптимизации

Require: число e_2 итераций с предсказанием $\boldsymbol{\lambda}$ линейными моделями

1: **while** нет сходимости **do**

2: Оптимизация $\boldsymbol{\lambda}$ и $\mathbf{w}$ на протяжении e_1 итераций, решая двухуровневую задачу

3: **traj** =траектория ($\nabla\boldsymbol{\lambda}$) изменяется во время оптимизации;

4: Положим $\mathbf{z} = [1, \dots, e_1]^\mathsf{T}$

5: Оптимизация $\mathbf{c}$ с помощью МНК:

$$\hat{\mathbf{c}} = \arg \min_{\mathbf{c} \in \mathbb{R}^2} ||\mathbf{traj} - \mathbf{z} \cdot c_1 + c_2||_2^2$$

6: Оптимизация $\mathbf{w}$ и предсказание $\boldsymbol{\lambda}$ на протяжении e_2 итераций с помощью линейной модели с параметрами $\mathbf{c}$.

7: **end while**

Алгоритм для предложенного метода.

Корректность аппроксимации линейной моделью

Теорема 1. Если функция $\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda})$ является гладкой и выпуклой, и ее Гессиан $\mathbf{H} = \nabla^2_{\mathbf{w}} \mathcal{L}_{\text{train}}$ обратим и является единичной матрицей, $\mathbf{H} = \mathbf{I}$, а также если параметры $\mathbf{w}$ равны $\mathbf{w}^*$, где $\mathbf{w}^*$ — точка локального минимума для текущего значения $\boldsymbol{\lambda}$, тогда жадный алгоритм находит оптимальное решение двухуровневой задачи. Если существует область $\mathcal{D} \in \mathbb{R}^2$ в пространстве метапараметров, такая что градиент метапараметров может быть аппроксимирован константой, то оптимизация является линейной по метапараметрам.

Постановка эксперимента

Выборки

Синтетическая выборка, CIFAR-10 (вся выборка и уменьшенная выборка), Fashion-MNIST

Методы оптимизации

1) оптимизация без дистилляции;
2) оптимизация со случайной инициализацией метапараметров. Метапараметры выбираются из равномерного распределения
$$\lambda_1 \sim \mathcal{U}(0;1), \quad T \sim \mathcal{U}(0.1, 10).$$
3) оптимизация с "наивным" назначением метапараметров:
$$\lambda_1 = 0.5, T = 1;$$
4) градиентная оптимизация;
5) hyperopt;
6) предложенный метод.

Внешний критерий качества:
$$\text{accuracy} = \frac{1}{m}\sum_{i=1}^{m}[\mathbf{g}(\mathbf{x}_i, \mathbf{w}) = y_i].$$

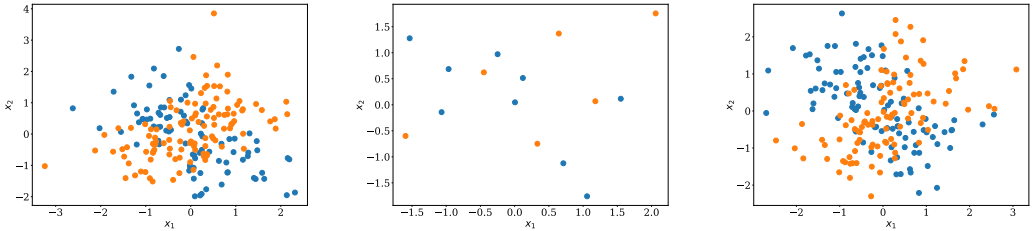
Эксперимент на синтетических данных

Выборка

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

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

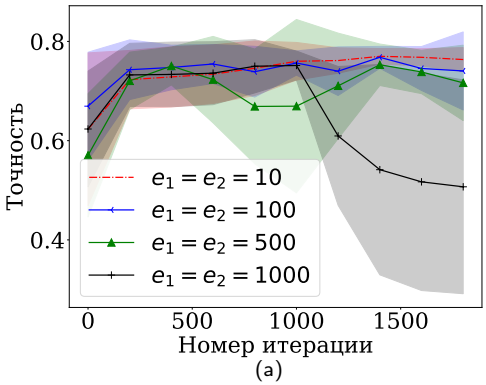
Размер выборки для модели-ученика существенно меньше размера выборки для модели-учителя.



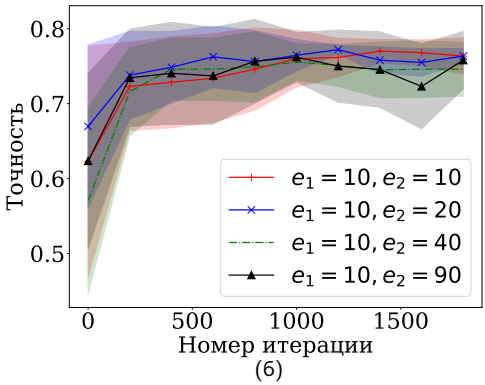
Визуализация выборки для а) модели-учителя; б) модели-ученика; в) тестовой выборки

Настройка параметров алгоритма

Точность модели со значениями e_1 и e_2: а) $e_1 = e_2$; б) подбор e_2 при $e_1 = 10$.



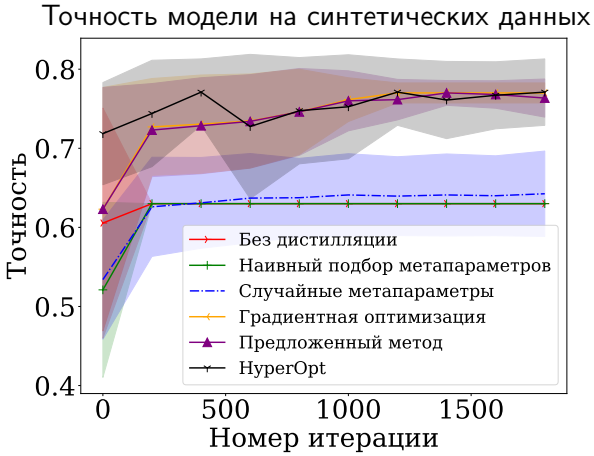
(а)



(б)

Лучшая точность получена при $e_1 = e_2 = 10$.

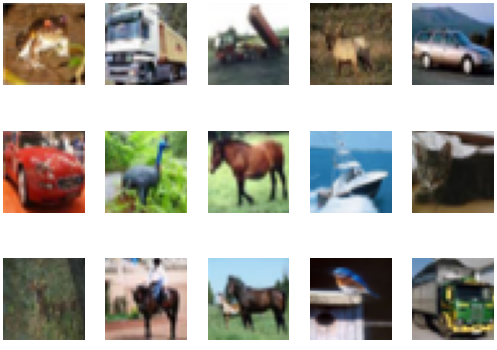
Сравение подходов к оптимизации



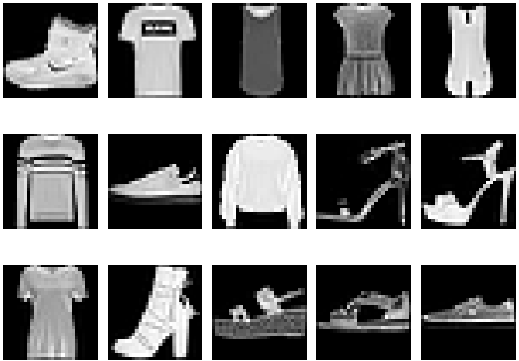
Результаты, полученные с помощью градиентной оптимизации близки к результатам, полученным с помощью аппроксимации линейными моделями.

Выборки CIFAR-10 и Fashion-MNIST

Метод оценивался на выборках Fashion-MNIST, CIFAR-10 и подмножестве CIFAR-10, которое составляет 10% от исходной выборки.

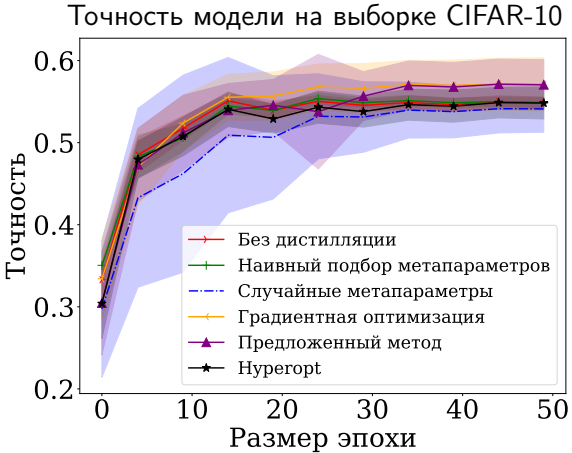


CIFAR-10



Fashion-MNIST

Результаты эксперимента на выборке CIFAR-10



Точность модели при обучении с дистилляцией значительно выше, чем без нее. Наибольшая точность получена при использовании предложенного метода.

Результаты эксперимента

Метод	Синтетическая выборка	Fashion-MNIST	Уменьшенный CIFAR-10	CIFAR-10
Без дистилляции	0.63 (0.63)	0.87 (0.88)	0.55 (0.56)	0.65 (0.66)
Наивный выбор	0.63 (0.63)	0.87 (0.88)	0.55 (0.56)	0.66 (0.67)
Случайные метапараметры	0.64 (0.72)	0.79 (0.88)	0.54 (0.57)	0.64 (0.67)
Градиентная оптимизация	**0.77** (0.78)	**0.88** (0.89)	**0.57** (0.61)	**0.70** (0.72)
Hyperopt	**0.77** (0.78)	0.87 (0.88)	0.55 (0.58)	-
Предложенный метод	0.76 (0.78)	**0.88** (0.89)	**0.57**	**0.70** (0.72)

Использование предложенного метода и градиентных методов дает похожий результат. Градиентные методы являются предпочтительными, так как они дают схожее качество, но требуют меньше вычислительных затрат.

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.

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.