Оптимизация метапараметров в задаче дистилляции знаний

Горпинич М.

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель: к.ф.-м.н. Бахтеев Олег Юрьевич

Весна 2022 г.

Дистилляция знаний

Цель

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

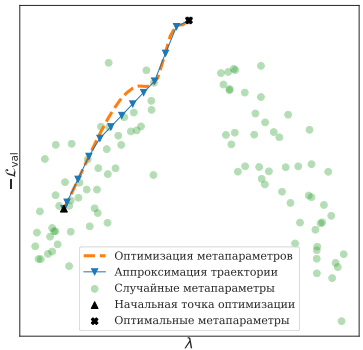
Проблема

Задача подбора метапараметров является вычислительно затратной. Правильное назначение метапараметров значительно повышает качество модели.

Решение

Задача метаоптимизации рассматривается как двухуровневая задача. Решение задачи предлагается проводить 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 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 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 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.

Корректность аппроксимации линейной моделью

Теорема 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

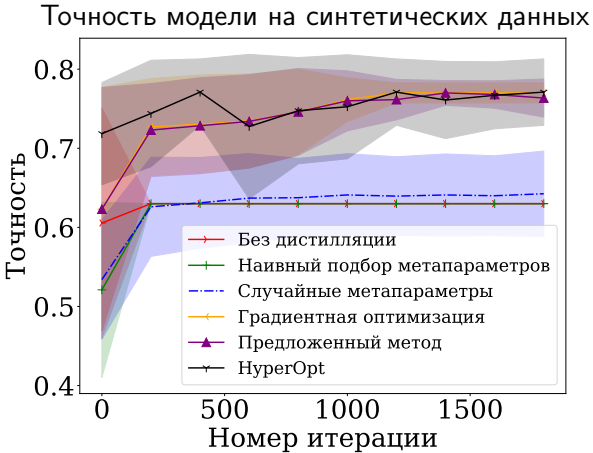
Синтетическая выборка, CIFAR-10 (вся выборка и уменьшенная выборка), Fashion-MNIST

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

Сравение подходов к оптимизации



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

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.

Результаты

Публикации ВАК по теме:

1. *Горпинич М., Бахтеев О.Ю., Стрижов В.В.* Градиентные методы оптимизации метапараметров в задаче дистилляции знаний // Автоматика и телемеханика (на рецензировании).

Выступление с докладом:

1. Metaparameter optimization in knowledge distillation // Maths & AI: MIPT-UGA young researchers workshop, 2021.
2. Градиентные методы оптимизации метапараметров в задаче дистилляции знаний // 64-я Всероссийская научная конференция МФТИ, 2021.
3. Градиентные методы оптимизации метапараметров в задаче дистилляции знаний // Математические методы распознавания образов (ММРО-2021), 2021.