Привилегированная информация и дистилляция моделей

Грабовой Андрей Валериевич

Московский физико-технический институт

МФТИ, г. Долгопрудный

Вероятностная интерпретация дистилляции моделей

Цель: предложить вероятностную постановку задачи дистилляции моделей глубокого обучения на основе существующих методов дистилляции и привилегированного обучения.

Задачи

1. Поставить вероятностную задачу дистилляции для задачи классификации и регрессии.
2. Провести теоретический анализ предложенной вероятностной постановки задачи для линейных моделей.

Исследуемая проблема: снижение размерности пространства параметров моделей глубокого обучения.

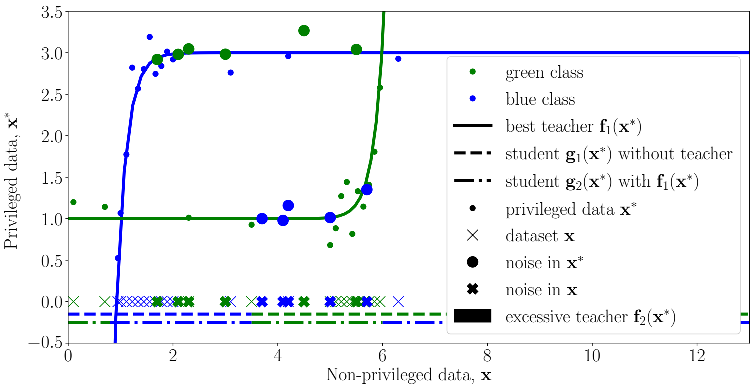
Метод решения

Предлагается поставить вероятностную постановку задачи дистилляции моделей глубокого обучения. В качестве базовой дистилляции предлагается использовать методы предложенные Дж. Хинтоном (2015г) и В. Вапником (2016г).

Список литературы

1. *Грабовой А.В., Стрижов В.В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и телемеханика, 2021 (текущая работа, на рецензировании).
2. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // In International Conference on Learning Representations. Puerto Rico, 2016.
3. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. 2015.
4. *Madala H., Ivakhnenko A.* Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc., 1994.

Привилегированное обучение и дистилляция



1. Учитель $\mathbf{f}$ влияет на выбор ученика $\mathbf{g}$ в пространстве x.
2. Учитель $\mathbf{f}_1$ корректирует шумные данные в x.
3. Модель учителя $\mathbf{f}_2$ более сложная, поэтому она аппроксимирует также и шум.

Постановка задачи обучения с учителем

Признаки $\mathbf{x}_i = \varphi(\omega_i)$, где $\mathbf{x}_i \in \mathbb{R}^n$.
Привилегированные признаки $\mathbf{x}_i^* \in \mathbb{R}^{n^*}$.
Целевая переменная $y_i \in \mathbb{Y}$.
Индексы объектов, для которых известна привилегированная информация обозначим $\mathcal{I}$, а $\bar{\mathcal{I}}$ для которых не известная.

Функции учителя $\mathbf{f} : \mathbb{R}^{n^*} \to \mathbb{Y}^*$ и ученика $\mathbf{g} : \mathbb{R}^n \to \mathbb{Y}^*$ — пространство оценок.
Ответ функции $\mathbf{f}$ для объекта $\mathbf{x}_i^*$ обозначим $\mathbf{s}_i = \mathbf{f}\bigl(\mathbf{x}_i^*\bigr)$.

Требуется выбрать модель ученика $\mathbf{g}$ из множества

$$\mathfrak{G} = \left\{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}^*\right\}.$$

Оптимизационная задача:

$$\mathbf{g} = \arg\min_{\mathbf{g} \in \mathfrak{G}} \mathcal{L}\bigl(\mathbf{g}, \mathbf{f}, \mathbf{X}, \mathbf{X}^*, \mathbf{y}\bigr),$$

где $\mathcal{L}$ функция ошибки.

Постановка задачи: дистилляция Хинтона

1) $\mathcal{I} = \{1, 2, \cdots, m\}$;
2) для всех $i \in \mathcal{I}$ выполняется $\mathbf{x}_i = \mathbf{x}_i^*$;
3) выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y} = \{1, \cdots, K\}$.

Параметрическое семейство учителя и ученика:

$$\mathfrak{F}_{\mathrm{cl}} = \left\{\mathbf{f} | \mathbf{f} = \mathrm{softmax}\big(\mathbf{v}(\mathbf{x})/T\big), \quad \mathbf{v}: \mathbb{R}^n \to \mathbb{R}^K\right\},$$
$$\mathfrak{G}_{\mathrm{cl}} = \left\{\mathbf{g} | \mathbf{g} = \mathrm{softmax}\big(\mathbf{z}(\mathbf{x})/T\big), \quad \mathbf{z}: \mathbb{R}^n \to \mathbb{R}^K\right\},$$

где $\mathbf{z}, \mathbf{v}$ — это дифференцируемые параметрические функции заданной структуры, T — параметр температуры.

Функция ошибки $\mathcal{L}_{\mathrm{st}}$:

$$\mathcal{L}_{\mathrm{st}}(\mathbf{g}) = -\sum_{i=1}^{m} \underbrace{\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\Big|_{T=1}}_{\text{исходная функция потерь}} - \sum_{i=1}^{m} \underbrace{\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i)\Big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\Big|_{T=T_0}}_{\text{слагаемое дистилляция}},$$

где $\cdot\Big|_{T=t}$ обозначает, что параметр температуры T в предыдущей функции равняется t.

Оптимизационная задача:

$$\hat{\mathbf{g}} = \arg \min_{\mathbf{g} \in \mathfrak{G}_{\mathrm{cl}}} \mathcal{L}_{\mathrm{st}}(\mathbf{g}).$$

Постановка задачи: дистилляция Вапника

Задано:

1) $\mathcal{I} = \{1, 2, \cdots, m\}$;
2) для всех $i \in \mathcal{I}$ выполняется $\mathbf{x}_i \neq \mathbf{x}_i^*$;
3) выборка $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^{m}, \mathbf{x}_i \in \mathbb{R}^n, \mathbf{x}_i^* \in \mathbb{R}^{n^*}, y_i \in \{1, \cdots, K\}$.

Параметрическое семейство учителя:

$$\mathfrak{F}_{\mathsf{cl}}^* = \left\{\mathbf{f} \middle| \mathbf{f} = \mathrm{softmax}\bigl(\mathbf{v}^*(\mathbf{x}^*)/T\bigr), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}^K\right\},$$

где $\mathbf{v}^*$ — это дифференцируемые параметрические функции заданной структуры, T — параметр температуры.

Функция ошибки:

$$\mathcal{L}_{\mathsf{st}}(\mathbf{g}) = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\bigr|_{T=1} - \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i^*)\bigr|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\bigr|_{T=T_0},$$

где $\cdot\bigr|_{T=t}$ обозначает, что параметр температуры T в предыдущей функции равняется t.

Оптимизационная задача:

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathfrak{G}_{\mathsf{cl}}} \mathcal{L}_{\mathsf{st}}(\mathbf{g}).$$

Вероятностная постановка

Гипотеза порождения данных:

1) задано распределение целевой переменной $p\bigl(y_i|\mathbf{x}_i, \mathbf{g}\bigr)$;
2) задано совместное распределение $p\bigl(y_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr)$;
3) для всех $i \in \mathcal{I}$ элементы y_i и $\mathbf{s}_i$ являются зависимыми величинами;
4) если $|\mathcal{I}| = 0$ то решение равно решению максимума правдоподобия.

Совместное правдоподобие истинных меток и меток учителя:

$$p\bigl(\mathbf{y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\bigr) = \prod_{i \not\in \mathcal{I}} p\bigl(y_i|\mathbf{x}_i, \mathbf{g}\bigr) \prod_{i \in \mathcal{I}} p\bigl(y_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr).$$

Задача оптимизации:

$$\mathbf{g} = \arg\max_{\mathbf{g} \in \mathfrak{G}} p\bigl(\mathbf{y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\bigr),$$

данная задача оптимизации переписывается в следующем виде:

$$\begin{aligned}\hat{\mathbf{g}} = \arg\max_{\mathbf{g} \in \mathfrak{G}} \sum_{i \not\in \mathcal{I}} \log p\bigl(y_i|\mathbf{x}_i, \mathbf{g}\bigr) + (1 - \lambda) \sum_{i \in \mathcal{I}} \log p\bigl(y_i|\mathbf{x}_i, \mathbf{g}\bigr) \\ + \lambda \sum_{i \in \mathcal{I}} \log p\bigl(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\bigr),\end{aligned}$$

где $\lambda \in [0, 1]$ — мета-параметр для взвешивания слагаемых отвечающих модели учителя относительно истинных меток.

Частный случай: задача классификации

1) функции учителя $\mathbf{f} \in \mathfrak{F}^*_{\mathsf{cl}}$ и ученика $\mathbf{g} \in \mathfrak{G}_{\mathsf{cl}}$;
2) распределение истинных меток $p\bigl(y|\mathbf{x}, \mathbf{g}\bigr) = \mathsf{Cat}\bigl(\mathbf{g}(\mathbf{x})\bigr)$;
3) распределение меток учителя $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) = C\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}$.

Оптимизационная задача:

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathfrak{G}} \sum_{i\not\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1} + (1-\lambda)\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1}$$

$$+\lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} s_{i,k}\log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0} + \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\left(\log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0} + \log\log\frac{1}{g_k\bigl(\mathbf{x}_i\bigr)}\bigr|_{T=T_0}\right)$$

Теорема (Грабовой, 2020)

Пусть вероятность каждого класса отделима от нуля и единицы, то есть для всех k выполняется $1 > 1-\varepsilon > g_k\bigl(\mathbf{x}\bigr) > \varepsilon > 0$, тогда при

$$C = (-1)^K \frac{K^{K/2}}{2^{K(K-1)/2}}\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)\log g_k\bigl(\mathbf{x}\bigr)$$

функция $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) = C\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}$ является плотностью распределения.

Частный случай: задача регрессии

1) учитель $\mathbf{f} \in \mathfrak{F}^*_{\text{rg}} = \left\{\mathbf{f} | \mathbf{f} = \mathbf{v}^*\left(\mathbf{x}^*\right), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}\right\}$;
2) ученик $\mathbf{g} \in \mathfrak{G}_{\text{rg}} = \left\{\mathbf{g} | \mathbf{g} = \mathbf{z}\left(\mathbf{x}\right), \quad \mathbf{z} : \mathbb{R}^{n} \to \mathbb{R}\right\}$;
3) распределение истинных меток $p\left(y | \mathbf{x}, \mathbf{g}\right) = \mathcal{N}\left(y | \mathbf{g}\left(\mathbf{x}\right), \sigma\right)$;
4) распределения меток учителя $p\left(s | \mathbf{x}, \mathbf{g}\right) = \mathcal{N}\left(s | \mathbf{g}\left(\mathbf{x}\right), \sigma_s\right)$.

Оптимизационная задача:

$$\hat{g} = \arg\min_{g \in \mathfrak{G}} \sum_{i \not\in \mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\left(\mathbf{x}_i\right)\right)^2 + \left(1 - \lambda\right) \sum_{i \in \mathcal{I}} \sigma^2 \left(y_i - \mathbf{g}\left(\mathbf{x}_i\right)\right)^2 + \lambda \sum_{i \in \mathcal{I}} \sigma_s^2 \left(s_i - \mathbf{g}\left(\mathbf{x}_i\right)\right)^2.$$

Теорема (Грабовой, 2020)

Пусть $\mathfrak{G}_{rg}$ *— класс линейных функций* $\mathbf{g}\left(\mathbf{x}\right) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$. *Тогда решение оптимизационной задачи эквивалентно решению задачи линейной регрессии* $\mathbf{y}'' = \mathbf{X}\mathbf{w} + \varepsilon,\ \varepsilon \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ *, где* $\boldsymbol{\Sigma}^{-1} = diag\left(\boldsymbol{\sigma'}\right)$ *и* $\mathbf{y}''$ *имеют следующий вид:*

$$\sigma'_i = \begin{cases} \sigma^2, \text{ если } i \not\in \mathcal{I} \\ \left(1 - \lambda\right) \sigma^2 + \lambda \sigma_s^2, \text{ иначе,} \end{cases} \qquad \mathbf{y}'' = \boldsymbol{\Sigma}\mathbf{y}', \quad y'_i = \begin{cases} \sigma^2 y_i, \text{ если } i \not\in \mathcal{I} \\ \left(1 - \lambda\right) \sigma^2 y_i + \lambda \sigma_s^2 s_i, \text{ иначе.} \end{cases}$$

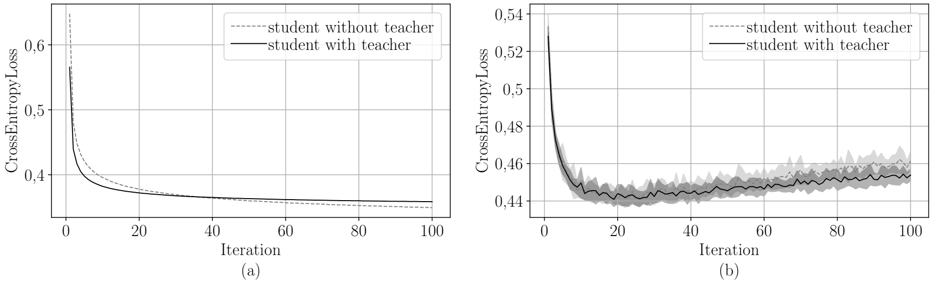
Вычислительный эксперимент

Выборка FashionMNIST:
Изображения размера 28×28. Решается задача классификации с $K = 10$ классами.
Объем выборки $m_{\text{train}} = 60000$ и $m_{\text{test}} = 10000$ объектов.

Синтетическая выборка:
$$\mathbf{X} = \left[\mathcal{N}\big(x_{ij}|0,1\big)\right]_{m \times n}, \qquad \mathbf{W} = \left[\mathcal{N}\big(w_{jk}|0,1\big)\right]_{n \times K},$$
$$\mathbf{S} = \text{softmax}\left(\mathbf{XW}\right), \qquad \mathbf{y} = \left[\text{Cat}\big(y_i|\mathbf{s}_i\big)\right],$$
где функция softmax берется построчно.
Число признаков $n = 10$, число классов $K = 3$, объем выборки $m_{\text{train}} = 1000$ и $m_{\text{test}} = 100$ объектов.

Выборка Twitter Sentiment Analysis:
Англоязычные твиты пользователей. Решается задача бинарной классификации текстовых сообщений.
Объем выборки $m_{\text{train}} = 1{,}18$млн и $m_{\text{test}} = 0{,}35$млн объектов.

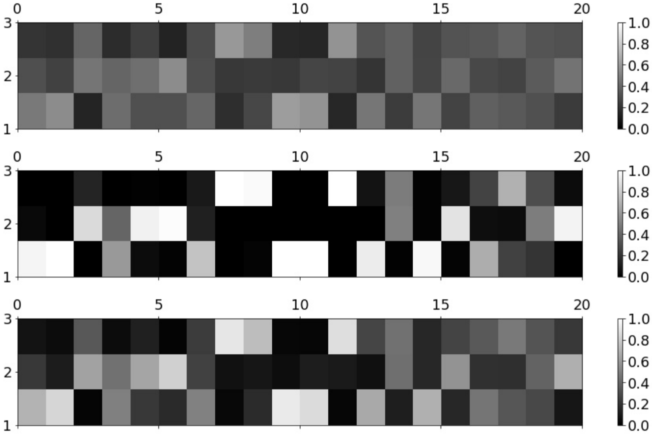
Выборка FashionMNIST



Кросс-энтропийная функция ошибки модели ученика: a) на обучающей выборке; b) на тестовой выборке.

Обе модели начинают переобучатся после 30-й итерации. Модель, которая получена путем дистилляции переобучается не так быстро: ошибка на тестовой выборке растет медленней, а на обучающей выборке падает медленней.

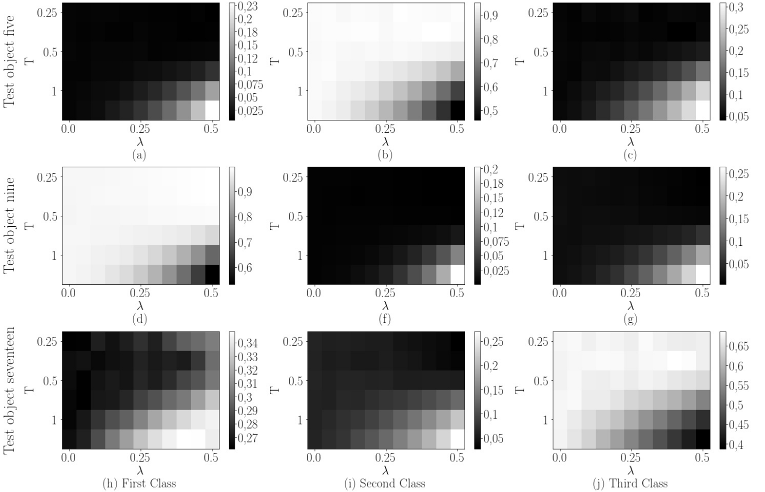
Синтетический эксперимент: распределение классов



Столбец — вероятность класса, строка — объекты. Сверху вниз: истинное распределение; без учителя; с учителем.

Модель ученика, которая использует информацию учителя, более точно восстанавливает вероятности классов в пространстве оценок, чем модель ученика, которая не использует информацию учителя.

Синтетический эксперимент: анализ параметра λ и T



Зависимость распределения по классам при разных параметрах λ и T. Видно, что при увеличении температуры T распределение на классах становится более равномерным.

Выборка Twitter Sentiment Analysis

В твитах была выполнена следующая предобработка:

- все твиты были переведены в нижний регистр;
- все никнеймы вида “@andrey” были заменены на токен “name”;
- все цифры были заменены на токен “number”.

Описание моделей:

- модель учителя: модель на основе Bi-LSTM с ≈ 30 миллионов настраиваемых параметров;
- модель ученика: модель на основе предобученной модели BERT с 1538 настраиваемых параметров.

Model	CrossEntropyLoss	Accuracy	StudentSize
без учителя	$0{,}501 \pm 0{,}006$	$0{,}747 \pm 0{,}005$	1538
с учителем	$0{,}489 \pm 0{,}003$	$0{,}764 \pm 0{,}004$	1538

При использовании учителя качество модели ученика увеличивается на 2% по сравнению с аналогичной моделью без использования меток учителя.

Сводная таблица вычислительного эксперимента

Dataset	Model	CrossEntropyLoss	Accuracy	StudentSize
FashionMnist	without teacher	$0{,}461 \pm 0{,}005$	$0{,}841 \pm 0{,}002$	7850
	with teacher	$0{,}453 \pm 0{,}003$	$0{,}842 \pm 0{,}002$	7850
Synthetic	without teacher	$0{,}225 \pm 0{,}002$	$0{,}831 \pm 0{,}002$	33
	with teacher	$0{,}452 \pm 0{,}001$	$0{,}828 \pm 0{,}001$	33
Twitter	without teacher	$0{,}501 \pm 0{,}006$	$0{,}747 \pm 0{,}005$	1538
	with teacher	$0{,}489 \pm 0{,}003$	$0{,}764 \pm 0{,}004$	1538

В таблице показаны результаты вычислительного эксперимента для разных выборок. Точность аппроксимации выборки учеником улучшается при использовании модели учителя при обучении.

Заключение

1. Поставлена вероятностная задача дистилляции моделей глубокого обучения.
2. Проведен теоретический анализ предложенной вероятностной задачи.
3. Результат анализа сформулирован в виде теорем: для задачи классификации, а также для задачи линейной регрессии.
4. Теорема для задачи линейной регрессии показала, что обучение модели линейной регрессии с учителем сводится к задаче линейной регрессии со скорректированными ответами.
5. Проведен вычислительный эксперимент для анализа предложенной модели.

Планируется:

1. Обобщить предложенный метод на случай задачи регрессии более корректно.
2. Использовать байесовский подход выбора моделей машинного обучения для решения данной задачи.

Публикации по теме

1. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019, 13(2).
2. *Грабовой А.В., Бахтеев О. Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020, 14(2).
3. *A. Grabovoy, V. Strijov.* Quasi-periodic time series clustering for human. Lobachevskii Journal of Mathematics, 2020, 41(3).
4. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал Вычислительной математики и математической физики, 2021. 61(5).
5. *Грабовой А.В., Стрижов В.В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и телемеханика, 2021 (текущая работа, на рецензировании).
6. *T. Gadaev, A. Grabovoy, A. Motrenko, V. Strijov* Numerical methods of minimum sufficient sample size estimation for linear models // in progress.
7. *Базарова А.И., Грабовой А.В., Стрижов В.В.* Анализ свойств вероятностных моделей в задачах обучения с экспертом // в разработке.