

Аннотация

Исследуется проблема понижения сложности аппроксимирующей модели при переходе к данным домена меньшей мощности. Вводятся понятия учителя, ученика, слабого и сильного доменов. Признаковые описания моделей ученика и учителя принадлежат разным доменам. Мощность одного домена больше мощности другого. Рассматриваются методы, основанные на дистилляции моделей машинного обучения. Вводится предположение, что решение оптимизационной задачи от параметров обеих моделей и доменов повышает качество модели ученика.

Ключевые слова: адаптация доменов, дистилляция, байесовский выбор модели, байесовская дистилляция

Содержание

1	Введение	4
2	Постановка задачи	6
2.1	Базовая постановка задачи дистилляции, предложенной Дж.Хинтоном	6
2.2	Постановка задачи дистилляции для многодоменной выборки	7
3	Вычислительный эксперимент	8
3.1	Анализ дистилляции Хинтона	9
4	Заключение	12

1 Введение

Сбор и обработка наборов данных для каждой новой задачи и области являются чрезвычайно дорогими и трудоемкими процессами, и не всегда могут быть доступны достаточные данные для обучения. Цель данной работы заключается в понижении сложности модели машинного обучения при переходе к домену меньшей мощности. Для этого предлагается использовать два основных метода - дистилляция моделей и доменная адаптация.

Дистилляция моделей машинного обучения использует метки модели с большим числом параметров для обучения модели с меньшим числом параметров. В [1] рассматривается метод дистилляции, предложенной Дж.Хинтоном, с учетом меток учителя при помощи функции softmax с параметром температуры, а в [2] рассматривается объединение методов дистилляции, предложенной Дж.Хинтоном, и привилегированной информации, предложенной В.Вапником, в обобщенную дистилляцию. Дистилляция моделей используется в широком классе задач. В [4] рассматривается метод дистилляции моделей для задачи распознавания речи.

Часто выборки могут состоять из объектов, которые можно разделить на домены. К примеру, можно составить отображение из множества реальных фотографий малой мощности во множество сгенерированных движком изображений, мощность которого естественно больше. Для задачи дистилляции, предложенной Дж.Хинтоном, исходный и целевой домены равны. Различные постановки задач доменной адаптации описываются в [5], встречаются постановки с частично размеченным целевым доменом и неразмеченным вовсе. Таким образом, доменная адаптация использует размеченные данные нескольких исходных доменов для выполнения новых задач в целевом домене.

Типичной задачей дистилляции моделей на многодоменных выборках является задача машинного перевода текстов, описанная в [3]. В качестве экспериментальных данных используются реальные данные и синтетическая выборка. В качестве реальных данных рассматривается выборка FashionMnist [6], состоящая из изображений одеж-

ды, для которой требуется решить задачу классификации на 10 типов одежды.

2 Постановка задачи

2.1 Базовая постановка задачи дистилляции, предложенной Дж.Хинтоном

Задана выборка $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$, где $\mathbf{X} \in \mathbb{X}, \mathbf{Y} \in \mathbb{Y}$. Множество $\mathbb{Y} = \{1, \dots, R\}$ для задачи классификации, где R - число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регрессии.

В качестве модели ученика \mathbf{g} рассматривается функция из множества:

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{z}(\mathbf{x})/T), \mathbf{z} : \mathbb{R}^n \rightarrow \mathbb{R}^R\}$$

В качестве модели учителя \mathbf{f} рассматривается функция из множества:

$$\mathfrak{F} = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^R\}$$

\mathbf{v}, \mathbf{z} - дифференцируемые параметрические функции заданной структуры, T - параметр температуры со свойствами:

- 1) при $T \rightarrow 0$ один из классов имеет единичную вероятность;
- 2) при $T \rightarrow \infty$ все классы равновероятны.

Функция потерь \mathcal{L} , учитывающая модель учителя \mathbf{f} при выборе модели ученика \mathbf{g} , имеет вид:

$$\mathcal{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{f}) = - \sum_{i=1}^m \sum_{r=1}^R y_i^r \log g^r(x_i)|_{T=1} - \sum_{i=1}^m \sum_{r=1}^R f^r(x_i)|_{T=T_0} \log g^r(x_i)|_{T=T_0},$$

где $\cdot|_{T=t}$ означает, что параметр температуры T в предыдущей функции равен t .

Получаем оптимизационную задачу:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{f}).$$

2.2 Постановка задачи дистилляции для много-доменной выборки

Заданы два домена:

$$\mathbb{D}_s, \mathbb{D}_t$$

- исходный и целевой наборы данных. Для задачи дистилляции, предложенной Дж.Хинтоном, $\mathbb{D}_s = \mathbb{D}_t$. Предполагается, что число объектов в доменах не совпадают:

$$|\mathbb{X}_s| \gg |\mathbb{X}_d|$$

\mathbb{Y} - множество целевых переменных.

Пусть при этом задана модель учителя

$$\mathbf{f} : \mathbb{X}_s \rightarrow \mathbb{Y}, \text{ где } \mathbf{f} - \text{модель учителя}$$

и связь между исходным и целевым доменами:

$$\varphi : \mathbb{X}_t \rightarrow \mathbb{X}_s, \text{ где } \varphi - \text{необратимое отображение}$$

Требуется получить отображение

$$\mathbf{g} : \mathbb{X}_t \rightarrow \mathbb{Y}, \text{ где } \mathbf{g} - \text{модель ученика}$$

Функция потерь, учитывающая метки учителя и связь между доменами

1) для задачи регрессии:

$$\mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{y}, \mathbf{f}, \varphi) = \lambda \|\mathbf{y} - \mathbf{g}(\mathbf{X}, \mathbf{w})\|_2^2 + (1 - \lambda) \|\mathbf{g}(\mathbf{X}, \mathbf{w}) - (\mathbf{f} \circ \varphi)(\mathbf{X})\|_2^2$$

2) для задачи классификации:

Получаем оптимизационную задачу:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{X}, \mathbf{y}, \mathbf{f}, \varphi).$$

3 Вычислительный эксперимент

Для анализа моделей, полученных путем дистилляции модели учителя в модель ученика, проводится вычислительный эксперимент для задачи классификации.

Эксперимент проводится для выборки FashionMNIST [6] - набора изображений предметов одежды. В качестве моделей учителя \mathbf{f} и ученика \mathbf{g} рассматриваются четырёхслойная и однослойная нейронные сети соответственно. Для решения оптимизационной задачи используется Adam, функция активации - ReLu.

Выборка разделяется на 3 части: две для обучения многоресурсного и малоресурсного доменов, а также тестовая часть выборки. Многоресурсная часть содержит 59000 объектов, малоресурсная часть содержит 1000 объектов, а тестовая часть содержит 10000 объектов.

3.1 Анализ дистилляции Хинтона

Обучение на обоих доменах

На рис.1а показан график зависимости метрики ассурасу на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.1б показан график зависимости кросс-энтропии на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что модель, использующая метки учителя, показывает лучшее значение ассурасу, при этом наблюдается незначительное повышение ошибки.

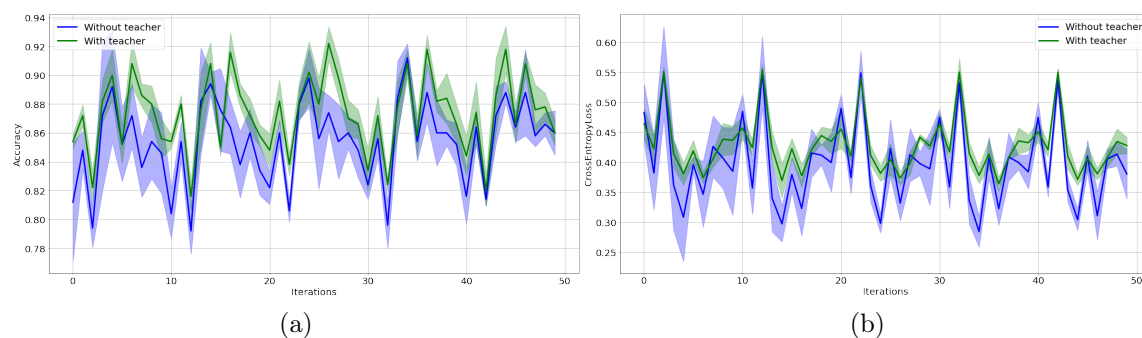


Рис. 1: Качество аппроксимации на тестовой выборке а) ассурасу; б) CrossEntropyLoss между истинными и предсказанными учеником метками

Обучение на многоресурсном домене На рис.1а показан график зависимости метрики ассигасу на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.1б показан график зависимости кросс-энтропии на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что модель, использующая метки учителя, показывает лучшее значение ассигасу, при этом наблюдается незначительное повышение ошибки.

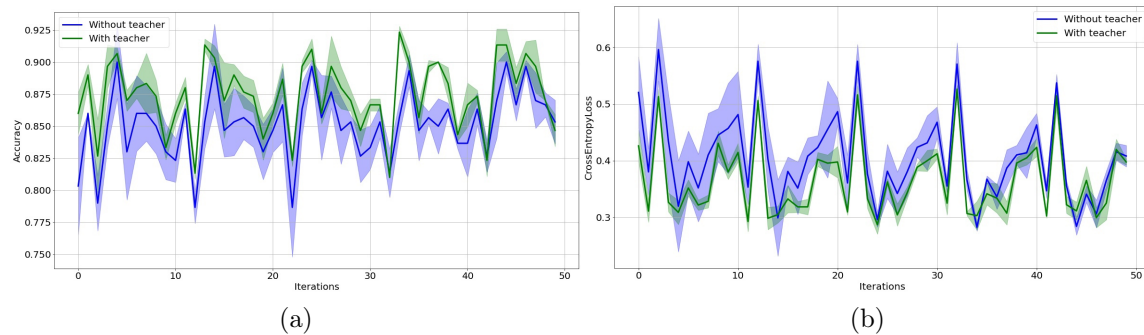


Рис. 2: Качество аппроксимации на тестовой выборке а) ассигасу; б) CrossEntropyLoss между истинными и предсказанными учеником метками

Обучение на малоресурсном домене На рис.1а показан график зависимости метрики ассигасу на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На рис.1б показан график зависимости кросс-энтропии на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

На графиках видно, что модель, использующая метки учителя, показывает лучшее значение ассигасу, при этом наблюдается незначительное повышение ошибки.

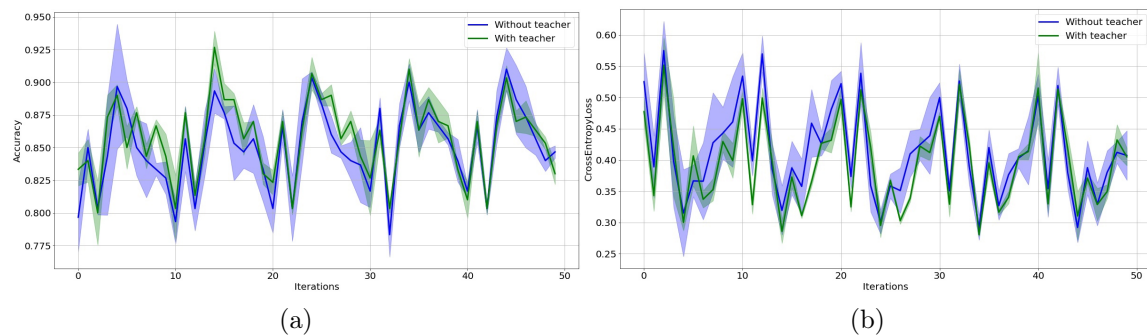


Рис. 3: Качество аппроксимации на тестовой выборке а) ассигасу; б) CrossEntropyLoss между истинными и предсказанными учеником метками

4 Заключение

Список литературы

- [1] *Hinton G., Vinyals O., Dean J* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. — 2015.
- [2] *D. Lopez-Paz, L. Bottou, B. Schölkopf, V. Vapnik* Unifying distillation and privileged information // ICLR. — 2016.
- [3] *Yoon Kim, Alexander M. Rush* Sequence-Level Knowledge Distillation. — 2016.
- [4] *H.Kim, M. Lee, H.Lee, T.Kang, J.Lee, E.Yang, S.Hwang* Multi-domain Knowledge Distillation via Uncertainty-Matching for End-to-End ASR Models. — 2021.
- [5] *Mei Wang, Weihong Deng* Deep Visual Domain Adaptation: A Survey. — 2018.
- [6] *Xiao H., Rasul K., Vollgraf R.* Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. — 2017. <https://arxiv.org/abs/1708.07747>.