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

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. Так, например в [1] рассматривается метод дистилляции Хинтона с учетом меток учителя при помощи функции softmax с параметром температуры, а в [2] рассматривается объединение методов дистилляции Хинтона и привилегированной информации Вапника в обобщенную дистилляцию. Дистиляцция моделей используется в широком классе задач, так в [4] рассматривается метод дистилляции моделей для задачи распознавания речи.

Большое количество информации можно разделить на домены. Так например, можно составить отображение из множества реальных фотографий малой мощности во множество сгенерированных движком изображений, мощность которого естественно больше. Для традиционной задачи машинного обучения 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 одежды, для которой требуется решить задачу классификации на 10 типов одежды.

2 Постановка задачи

2.1 Базовая постановка задачи дистилляции Хинтона

Задана выборка $\mathcal{D} = (X, Y)$, где $X \in \mathbb{X}, Y \in \mathbb{Y}$. Множество $\mathbb{Y} = \{1, ..., R\}$ для задачи классификации, где R - число классов, множество $\mathbb{Y} = \mathbb{R}$ для задачи регресии.
В качестве модели ученика $\mathbf{g}$ рассматривается функция из множества:

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

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

$$\mho = \{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^R\}$$

$\mathbf{v}$, $\mathbf{z}$ - дифференцируемые параметрические функции заданной структуры, T - параметр температуры со свойствами:
1) при $T \to 0$ один из классов имеет единичную вероятность;
2) при $T \to \infty$ все классы равновероятны.
Функция потерь $\mathcal{L}$ учитывает перенос информации от модели учителя $\mathbf{f}$ к 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 многодоменной выборки

Заданы два домена:

$$\mathbb{D}^s, \mathbb{D}^t$$

- исходный и целевой наборы данных. Для традиционной задачи машинного обучения $\mathbb{D}^s = \mathbb{D}^t$. Предполагается, что признаковые описания доменов не совпадают, а именно

$$|\mathbb{X}^s| \gg |\mathbb{X}^d|$$

$\mathbb{Y}$ - множество целевых переменных. Пусть при этом заданы модель учителя и связь между исходным и целевым доменами:

$$\mathbf{f} : \mathbb{X}^s \to \mathbb{Y}, \text{ где } \mathbf{f} - \text{модель учителя}$$

$$\varphi : \mathbb{X}^t \to \mathbb{X}^s, \text{ где } \varphi - \text{необратимое отображение}$$

Требуется получить отображение

$$\mathbf{g} : \mathbb{X}^t \to \mathbb{Y}, \text{ где } \mathbf{g} - \text{модель ученика}$$

Функция потерь, учитывающая метки учителя и связь между доменами:

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

Получаем оптимизационную задачу:

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

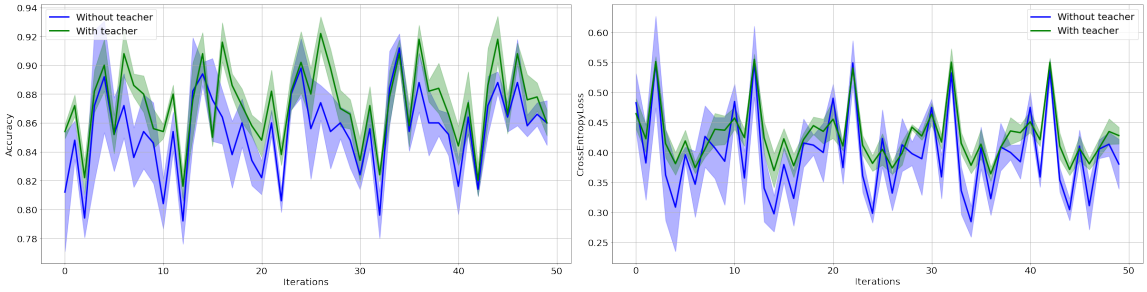
3 Вычислительный эксперимент

Для анализа моделей, полученных путем дистилляции модели учителя в модель ученика, был проведен вычислительный эксперимент для задачи классификации.
Эксперимент проводился для выборки FashionMNIST [6] - набора изображений предметов одежды. В качестве моделей учителя $\mathbf{f}$ и ученика $\mathbf{g}$ рассматриваются четырёхслойная и однослойная нейронные сети соответсвенно, в качестве функции активации рассматривается ReLu. Градиентный метод оптимизации - Adam.
Выборка разделяется на 3 части: две для обучения многоресурного 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

На рисунках a, b показаны графики зависимостей accuracy и кросс-энтропии на тестовой выборке между истинными метками объектов и вероятностями, предсказанными моделью ученика.

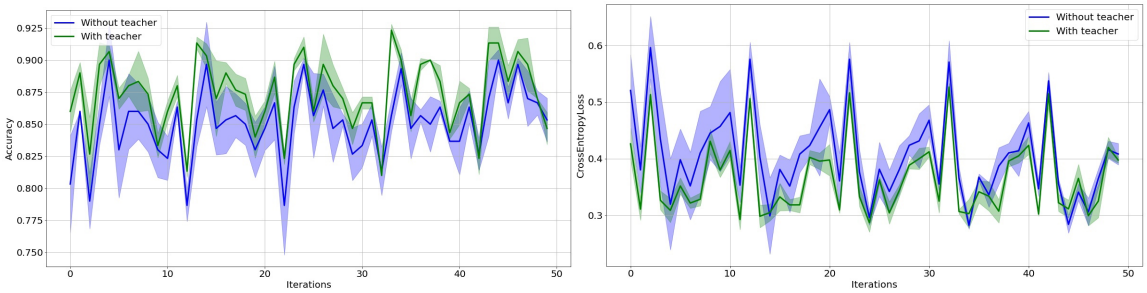
На всех данных На графиках видно, что модель, использующая метки учителя, показывает лучшее значение accuracy, при этом наблюдается незначительное повышение ошибки.



(a) (b)

Рис. 1: Качество аппроксимации на тестовой выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

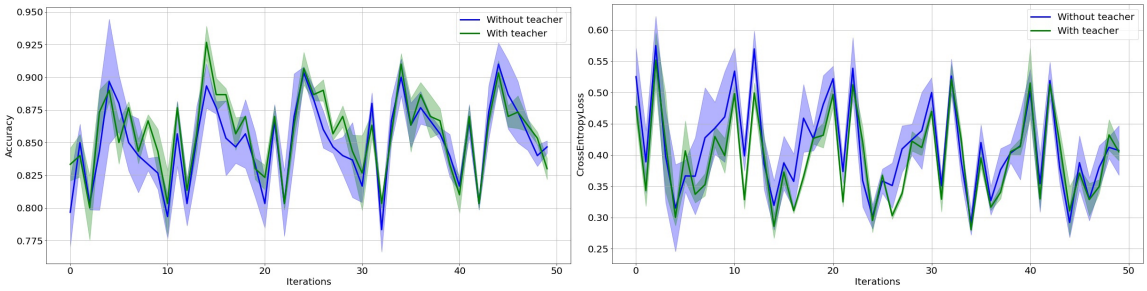
На многоресурсном домене На графиках видно, что качество модели почти не изменилось.



(a) (b)

Рис. 2: Качество аппроксимации на тестовой выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

На малоресурсном домене На графиках видно, что качество модели понизилось.



(a) (b)

Рис. 3: Качество аппроксимации на тестовой выборке a) accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками

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.