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:

Содержание

1 Введение

Доменная адаптация использует размеченные данные нескольких доменов для выполнения новых задач в целевом домене.
Исходный и целевой домены могут содержать изображения, тогда расхождение признаковых описаний может быть вызвано разными сенсорными устройствами и разными стилями изображений (рисунки и фотографии).
Дистилляция моделей машинного обучения использует метки модели с большим числом параметров для обучения модели с меньшим числом параметров.

2 Анализ литературы

В [1] рассматривается метод дистилляции с учетом меток учителя при помощи функции softmax с параметром температуры.
В [2] рассматривается объединение методов дистилляции Хинтона и привилегированной информации Вапника в обобщенную дистилляцию.
В [3] рассматривается метод дистилляции моделей для задачи перевода текстов.
В [4] рассматривается метод дистилляиии моделей для задачи распознавания речи.
В [5] рассматривается задача машинного обучения при наличии исходного и целевых доменов.
В [6] приводится описание выборки, на которой проводятся все эксперименты.

3 Постановка задачи

3.1 Базовая постановка задачи дистилляции Хинтона

Задано множество объектов Ω и множество целевых переменных $\mathbb{Y}$. 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:

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

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

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

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

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

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

3.2 Постановка задачи дистилляции для многодоменной выборки

Заданы два домена: $\mathbb{D}^s, \mathbb{D}^t$, - исходный и целевой датасеты. (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. $\mathbb{Y} = \{1, ..., R\}$ для задачи классификации, где R - число классов, $\mathbb{Y} = \mathbb{R}$ для задачи регресии. Пусть при этом заданы модель ученика и связь между исходным и целевым доменами:

$$\mathbf{f} : \mathbb{X}^s \to \mathbb{Y}$$

$$\varphi : \mathbb{X}^t \to \mathbb{X}^s$$

Требуется получить модель ученика

$$\mathbf{g} : \mathbb{X}^t \to \mathbb{Y}$$

Функция потерь, учитывающая метки учителя и связь между доменами:

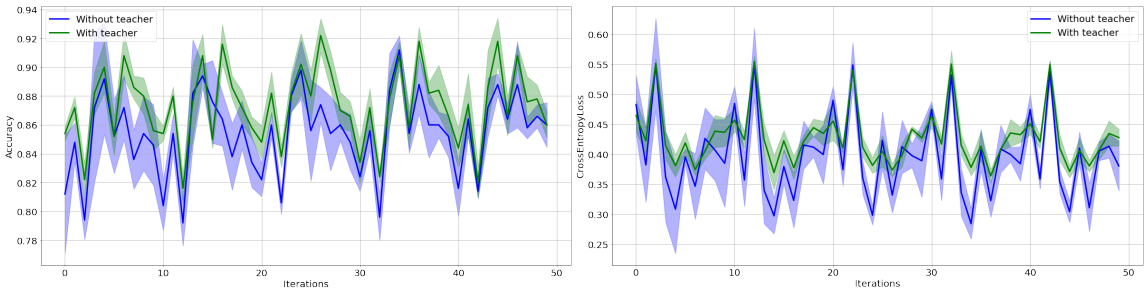
$$\mathcal{L} = \lambda \|\overline{y} - g(X, w)\|_2^2 + (1 - \lambda)\|g(X, w) - f \times \varphi(X)\|_2^2$$

4 Вычислительный эксперимент

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

Выборка FashionMNIST. 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.

На рисунках 1, 2 показаны графики зависимостей accuracy и 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 Accuracy, при этом наблюдается незначительное повышение ошибки.



(a) (b)

Рис. 1: Зависимость a) Accuracy; b) CrossEntropyLoss между истинными и предсказанными учеником метками от числа итераций на тестовой выборке

4.1 Базовый эксперимент

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.