Байесовский дистилляция разнородных моделей глубокого обучения

Декабрь 2020

1 Формальная постановка задачи

Задача: Рассматривается задача многоклассовой классификации.

1. **Моделью глубокого обучения** будем называть дифференцируемую по параметрам w функцию $f(w, x)$ из множества признаковых описаний объекта во множество меток:

$$f : \mathbb{W} \times \mathbb{X} \to \mathbb{Y}$$

Где $\mathbb{W}$ - это пространство параметров функции f

2. Модели глубокого обучения f и g называются **неоднородными** , если число скрытых слоев этих моделей или число нейронов в них не эквивалентны друг другу.

Имеются две неоднородные модели глубокого обучения $f(w_t, x)$ и $g(w_s, x)$, именуемые "модель учитель пространство параметров которой является заранее оптимизированным, и "модель студент"с неоптимальным элементом $_s$ из пространства собственных параметров соответсвенно.

3. Задано параметрическое распределение q_t , моделирующеее выходы промежуточных слоев или их скрытые представления "модели учителя включая логиты данной сети

4. Задано параметрическое распределение q_s, моделирующее выходы промежуточных слоев или их скрытые представления"модели студента включая логиты данной сети

5. Задана параметрическая модель классификации $\mathbb{D}_\theta$, где θ - это элемент пространства параметров дискриминатора, призванного разделять скрытые представления двух заданных неоднородных моделей (в частности, логиты этих сетей).

Ставится задача двухуровневая задача оптмизации:

$$\theta^* = \arg\min_\theta (\mathbb{E}_{t\sim q_t(x,w_t)} \log \mathbb{D}_\theta(t(x, w_t)) + \mathbb{E}_{s\sim q_s(x,w_s)} \log(1 - \mathbb{D}_\theta(s(x, w_s))))$$

$$w_s^* = \arg\max_{w_s} \mathbb{E}_{s\sim q_s(x,w_s)} \mathbb{D}_{\theta^*}(s(x, w_s)) - \sum_k \mathcal{L}(y_k|g(w_s, x_k))$$

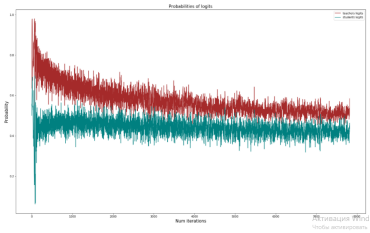

Рис. 1: Вероятность логитов "студента"и "учителя"при адверсариальном обучении

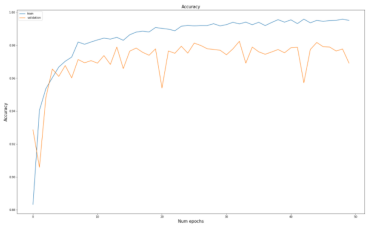

Рис. 2: Точность "студента"предсказаний на трейне и валидации

Где $\mathcal{L}(y_k|g(w_s, x_k))$ - это ошибка модели на объекте x_k.

Цель: Обучить "модель студент"для задачи многоклассовой классификации при помощи минимизации кросс-энтропии между предсказанными метками $\hat{\mathbb{Y}}$ "моделью студентом"и действительными метками $\mathbb{Y}$ на тех же объектах $\curvearrowleft$ и обучения дискриминатора $\mathbb{D}_\theta$, который должен быть способен различать семплы из распределения q_t и q_s, либо их скрытые представления z_t и z_s соответственно.

"Модель студент"представляет собой генератор, поскольку именно она генерирует распределения q_s в каждом слое, которые дискриминатор $\mathbb{D}_\theta$ должен определять как "фейковые".

Процесс обучения:

1. Берем батч объектов $x_{batch} = \{x_1, ..., x_B\}$, где B размер батча 2. Получаем выход из функции "модели учителя"$f(\psi, x_{batch})$ как логиты $\mathbb{L}(x_{batch}, \psi)$ 3. Получаем выход из функции "модели студента"$g(\phi, x_{batch})$ как логиты $\hat{\mathbb{L}}(x_{batch}, \phi)$ 4.Минимизируя следующую функцию потерь по парметрам дискриминатора, находим оптимальные:

$$\theta^* = \arg\min_\theta(\mathbb{E}_{t\sim q_t(x,w_t)} \log \mathbb{D}_\theta(t(x, w_t)) + \mathbb{E}_{s\sim q_s(x,w_s)} \log(1 - \mathbb{D}_\theta(s(x, w_s))))$$

5. На втором шаге мы минимизируем кросс энтропию между параметрическими распределениями промежуточных слоев "модели студента"и "модели учителя"по параметрам первого, чтобы найти оптимальные:

$$w_s^* = \arg\max_{w_s} \mathbb{E}_{s\sim q_s(x,w_s)} \mathbb{D}_{\theta^*}(s(x, w_s)) - \sum_k \mathcal{L}(y_k|g(w_s, x_k))$$

И так далее

2 Численные эксперименты на MNIST

Описание эксперимента появится чуть позже. Результаты приведенны на изображениях 1 и 2.