Адверсариальный подход к дистилляции моделей глубинного обучения при помощи их весов

Колесов Александр, Бахтеев Олег

Московский физико-технический институт(ГУ)

Москва
2021 май

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.

Формальная постановка задачи

.

1. Задано параметрическое распределение $\mathbf{q}_t$, моделирующеее выходы промежуточных слоев или их скрытые представления "модели учителя включая логиты данной сети

2. Задано параметрическое распределение $\mathbf{q}_s$, моделирующее выходы промежуточных слоев или их скрытые представления"модели студента включая логиты данной сети

3. Задана параметрическая модель классификации D_θ, где θ - это элемент пространства параметров дискриминатора, призванного разделять скрытые представления двух заданных неоднородных моделей (в частности, логиты этих сетей).

Двухуровневая задача оптмизации

Ставится задача двухуровневая задача оптмизации:

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

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

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

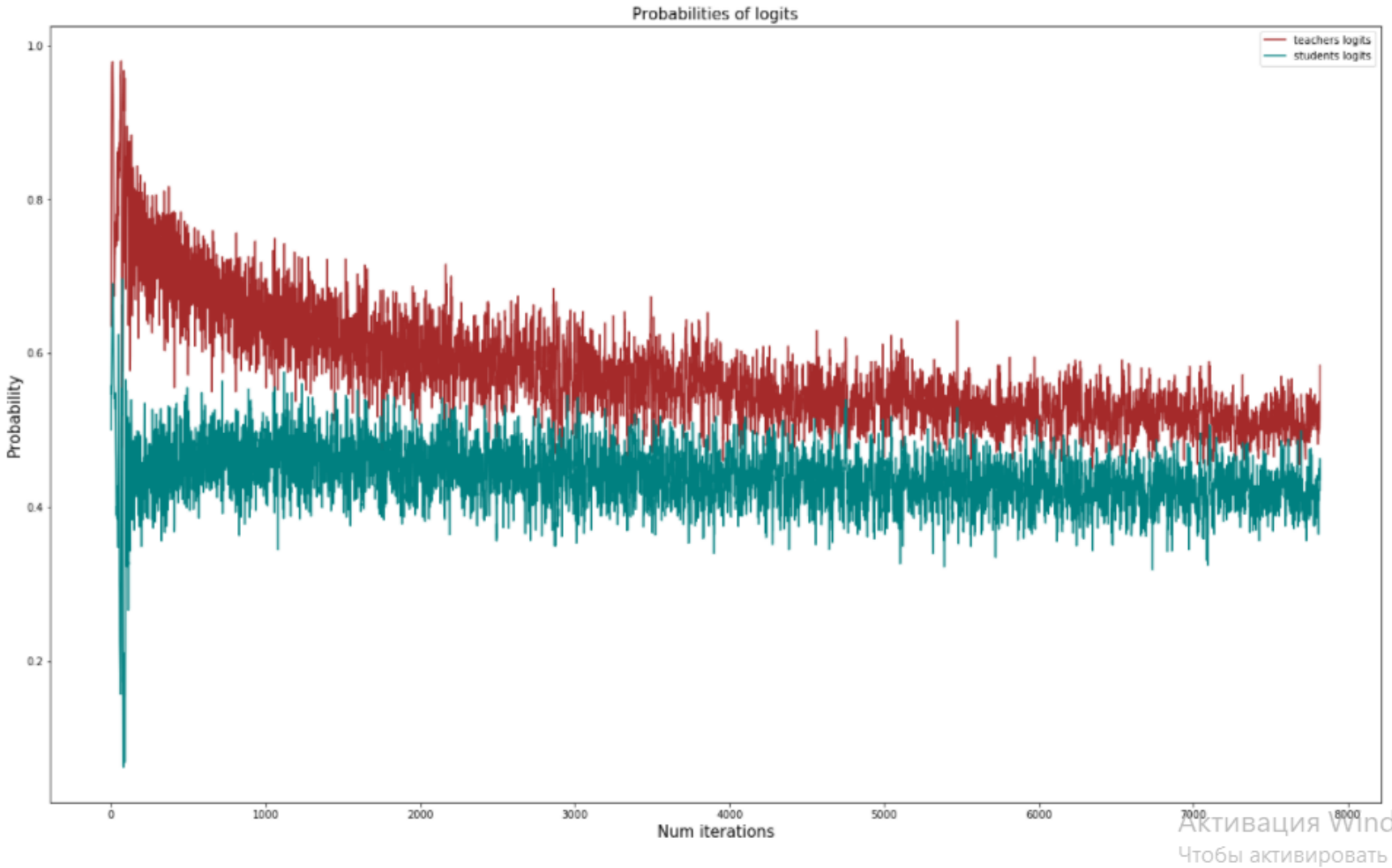

Рис.: Вероятность логитов "студента"и "учителя"при адверсариальном обучении

Результаты экспериментов на CIFAR-10

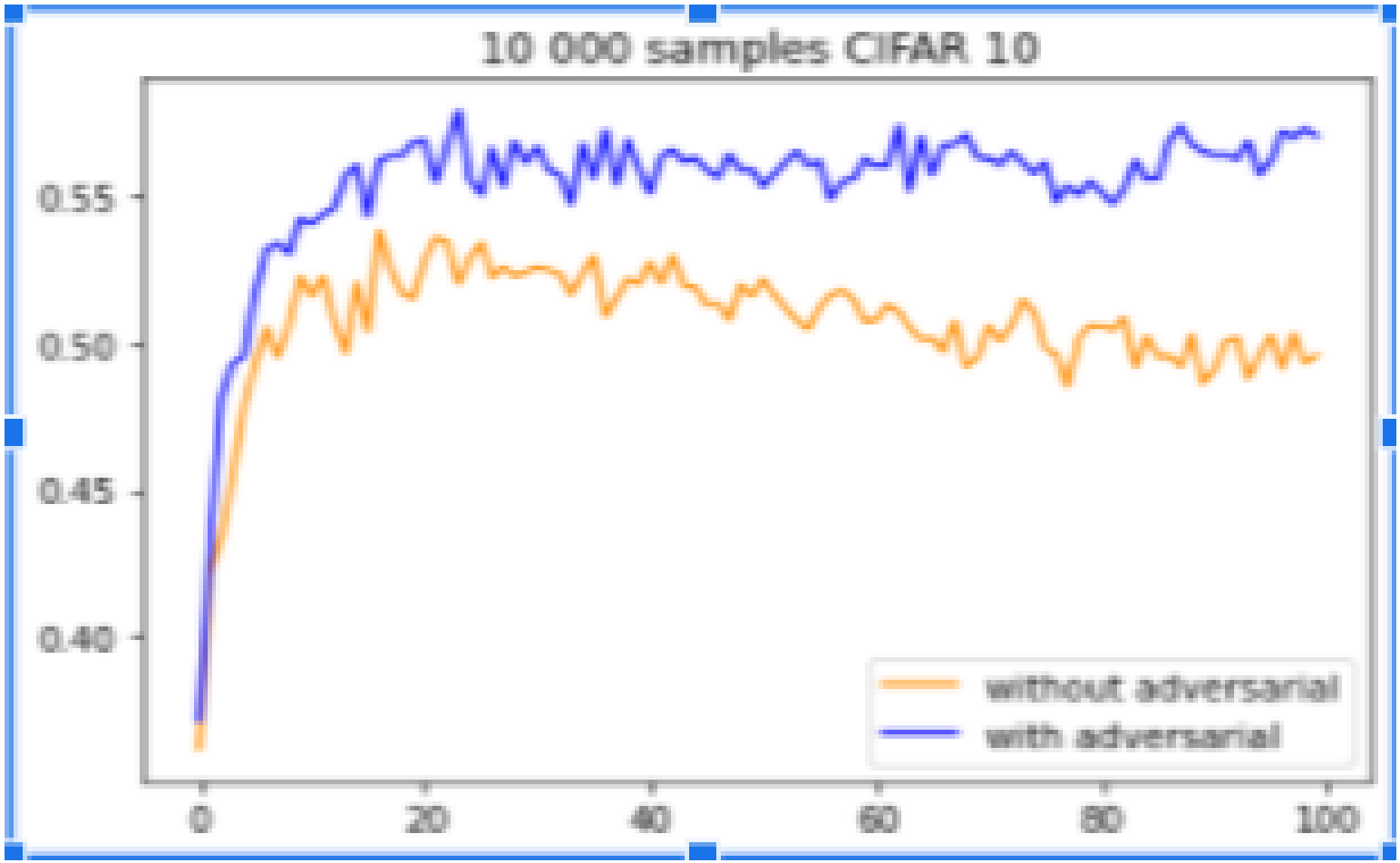

Рис.: Точность студента на валидации, обученного при помощи нашего метода и при помощи кросс-энтропии

Модель Хинтона

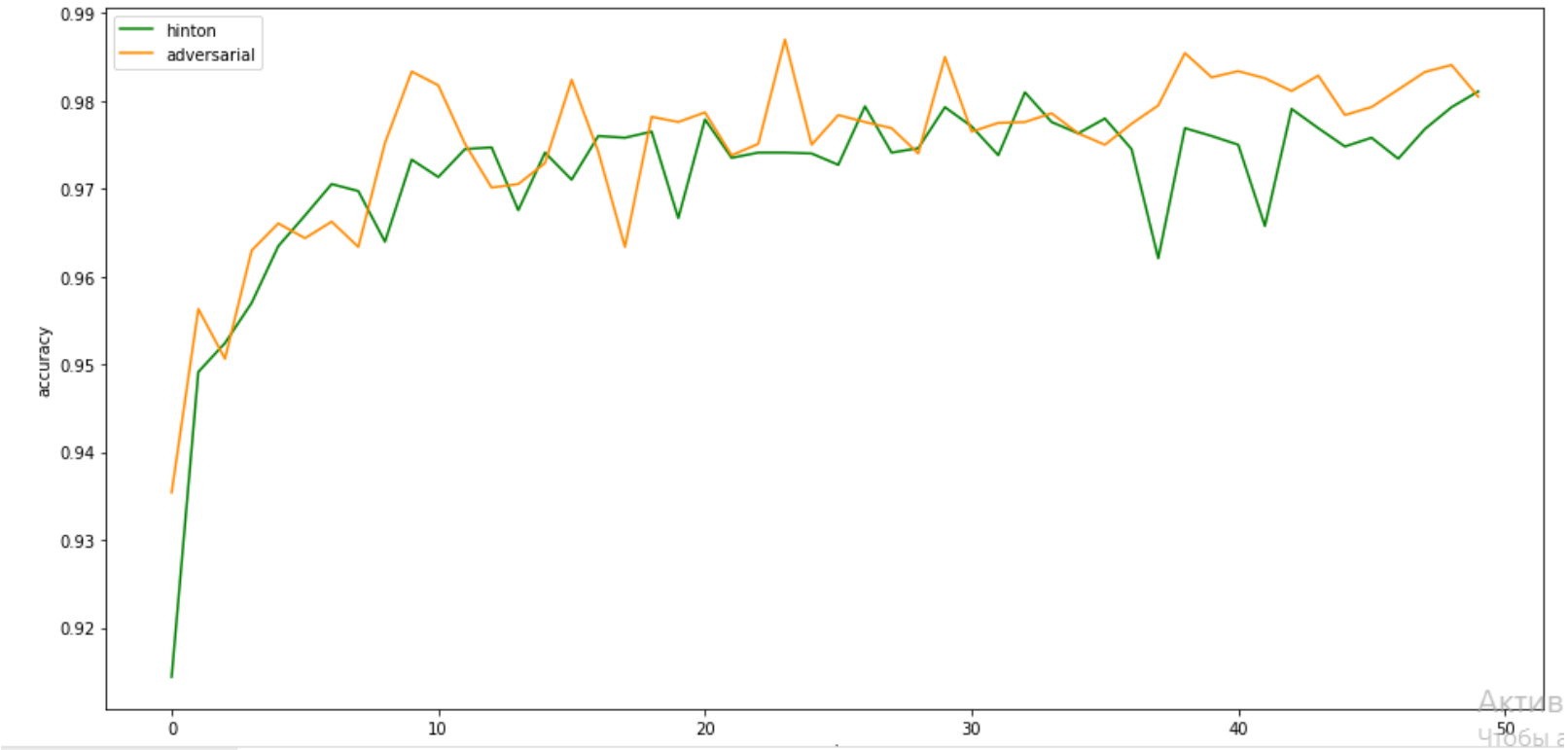

Рис.: Сравнение модели Хинтона и адверсариальной моделей