УДК 004.6, 004.8, 004.9

Приватность меток в вертикальном федеративном обучении

К. О. Шестакова, А. Н. Безносиков
Московский физико-технический институт (национальный исследовательский университет)

В настоящий момент масштабные модели машинного обучения стремительно набирают популярность. Для распределенной обработки данных и обучения модели используется такой подход, как федеративное обучение. Федеративное обучение, в свою очередь, разделено на два направления: вертикальное и горизонтальное. При вертикальном федеративном обучении различные участники процесса обучения располагают некоторым подмножеством свойств, характеризующим все семплы глобального датасета, в то время как при горизонтальном федеративном обучении различные участники имеют доступ ко всем свойствам, характеризующим некоторое подмножество семплов глобального датасета.

Одним из главных достоинств подхода федеративного обучения считается приватность данных, достигаемая за счет того, что стороны, принимающие участие в обучении, не делятся данными из датасета напрямую. Однако даже при таком подходе возможна утечка данных при передаче промежуточных результатов обучения, например, градиентов, активаций локальных моделей и промежуточных весов [1]. Если одна из сторон обучения является недобросовестной, то данные других сторон или же итоговая модель могут быть похищены и использованы злоумышленниками.

В данной работе рассматривается проблема приватности в распределенных моделях компьютерного зрения. Изучены различные способы распределения модели ResNet [2] между участниками федеративного обучения - двумя клиентами и одним сервером. Для каждого способа разделения изучена эффективность атаки KMeans и защиты от нее при помощи регуляризации. В результате удалось уменьшить качество атаки KMeans на 20% без потерь в качестве предсказания итоговой модели при помощи использования дополнительного блока ResidualBlock на конечном клиенте и добавления слагаемого регуляризации с весом 0.3 и шагом 10.

Рассмотрены способы разделения модели ResNet, состоящей из трех слоев, каждый из которых содержит от 0 до 3 блоков ResidualBlock, при которых начальный клиент обучает начальный слой, сервер обучает средний слой, конечный клиент обучает конечный слой, при разных количествах блоков в каждом из слоев. Результаты представлены на рис. 1. Наименьшая точность атаки Kmeans достигнута для случая, когда конечный клиент обучает глубокий слой, т.е. состоящий из 2-3 блоков. При регуляризации с параметрами веса 0.3 и шага 10 достигается точность 0.25-0.35, в то время как точность без регуляризации достигает 0.5.

Также рассмотрена постановка, когда начальный клиент обучает некоторое количество легковесных слоев, называемых предобработкой, сервер обучает три слоя, состоящих из блоков ResidualBlock, конечный клиент обучает некоторое количество легковесных слоев, называемых постобработкой. Данная постановка актуальна для практических ситуаций, в которых сервер обладает моделью с приватной архитектурой,

не доступной клиентам. На рис. 2 представлены результаты для разных способов постобработки. В частности, протестирована классическая постобработка ResNet с добавлением 1, 2, 4 дополнительных линейных слоев и 1 дополнительного блока ResidualBlock. Наименьшая точность атаки Kmeans достигнута при добавлении 1 дополнительного блока ResidualBlock и регуляризации с весом 0.3 и шагом 10. При данных параметрах достигнута точность 0.58 без потери качества предсказания итоговой модели, в то время как для других способов постобработки и параметров регуляризации точность достигает 0.8-0.9.

Рис. 1, Сравнение точности Kmeans для разных разбиений по слоям

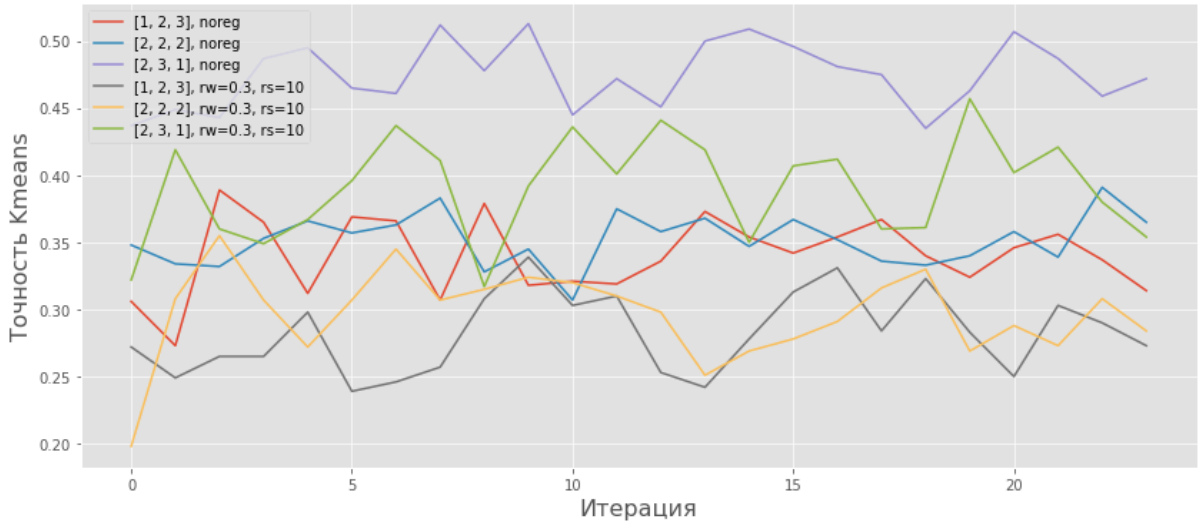
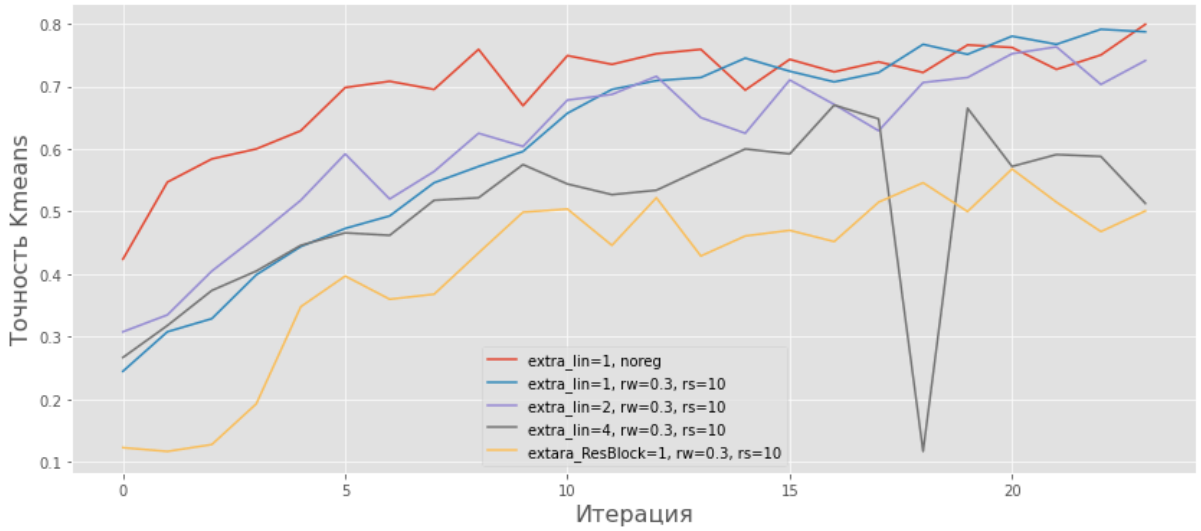


Рис. 2, Сравнение точности Kmeans для разных постобработок



Литература

[1] *Lei Yu, Meng Han, Yiming Li, Changting Lin, Yao Zhang, Mingyang Zhang, Haiqin Weng, Yuseok Jeon, Ka-Ho Chow, Stacy Patterson.* A Survey of Privacy Threats and Defense in Vertical Federated Learning: From Model Life Cycle Perspective. *arXiv:2402.03688v1*. 2024.
[2] *Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.* Deep Residual Learning for Image Recognition. *arXiv:1512.03385*. 2015.