УДК 004.056

РАЗРАБОТКА АЛГОРИТМА ЗАЩИТЫ НЕЙРОННОЙ СЕТИ ОТ АТАКИ ИНВЕРСИИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

Растворцева А.Е. (ИТМО)
Научный руководитель – ассистент Есипов Д.А. (ИТМО)

Введение. В настоящее время нейронные сети стали неотъемлемой частью многих сфер нашей жизни. Они применяются в различных областях, начиная от медицины, заканчивая технологиями и бизнесом. С учетом всеобщего распространения нейронных сетей важно гарантировать, что эти модели не раскроют конфиденциальную информацию о данных, на которых они были обучены. При использовании атаки с инверсией модели злоумышленники получают доступ к модели машинного обучения с целью извлечения конфиденциальных данных о конкретных лицах. Так как многие имеющиеся решения имеют свои недостатки, то существует потребность в разработке алгоритма, который позволит увеличить ошибку инверсии модели.

Основная часть. В рамках данной работы рассматривается атака на выходные данные модели машинного обучения, а именно атака инверсии модели. Атака инверсии относится к классу атак, в которых с помощью запросов к целевой модели, злоумышленник получает доступ к конфиденциальной информации. Целью использования атаки инверсии модели является восстановление злоумышленником набора данных, который используется для обучения нейронной сети [1].

В связи с ограниченным количеством проведенных исследований, относящихся именно к атакам инверсии модели, были проанализированы различные методы предотвращения для атак на выходные данные модели машинного обучения, с выделением наиболее подходящих в контексте инверсии модели [2]. Поскольку такой широко использующийся метод как дифференциальная приватность может приводить к значительной потере точности классификации, требуется разработать более универсальный алгоритм противодействия атаке.

Выводы. Были проанализированы результаты исследований и на основе полученных данных проведено сравнение с другими ранними работами. Представлено более эффективное с точки зрения сохранения точности классификации решение. Реализация данного алгоритма позволит увеличить ошибку инверсии модели и не потерять точность классификации.

Список использованных источников:

1. S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, “Model Inversion Attacks for Prediction Systems : Without Knowledge of Non-Sensitive Attributes,” in 2017 15th Annual Conference on Privacy, Security and Trust (PST), 2017.
2. Z Yang, B Shao, B Xuan, EC Chang, F Zhang, “Defending model inversion and membership inference attacks via prediction purification.” arXiv preprint arXiv:2005.03915, 2020. 76. 2020.