УДК 004.89

РАЗРАБОТКА МЕХАНИЗМА РЕВЕРС ИНЖИНИРИНГА ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ПО ДАННЫМ ТЕЛЕМЕТРИИ

Крисанов Р.В. (ИТМО)

Научный руководитель - кандидат технических наук, доцент Кугаевских А. В. (ИТМО)

Введение.

Глубокие нейронные сети активно применяются в критически важных системах (финансы, медицина, безопасность), являясь объектами интеллектуальной собственности. Их широкое распространение создает риски утечки информации об архитектуре. Это формирует потребность в исследовании методов защиты и аудита безопасности. Векторы атак включают анализ выходных данных модели, физический доступ к оборудованию и побочные каналы при выполнении на графических процессорах. Последний вектор представляет особый интерес. Существующие подходы к реверс-инжинирингу через побочные каналы используют анализ физических утечек - электромагнитного излучения и энергопотребления. В данной же работе рассматривается подход на основе телеметрии GPU. Актуальность обусловлена тем, что аппаратные счетчики производительности используются в задачах анализа безопасности и демонстрируют информативность. Например, в одном из исследований показано, что статистика промахов кэша позволяет предсказывать классы входных данных. Это указывает на потенциальную возможность использования телеметрии для восстановления архитектурных особенностей самих нейронных сетей. Новизна работы заключается в применении профилирования GPU на уровне отдельных вычислительных ядер для восстановления архитектуры через классификацию типов слоев.

Основная часть.

Работа направлена на разработку механизма восстановления типов слоев нейронных сетей по программно доступной телеметрии GPU. Достаточная детализация телеметрии достигается использованием инструмента профилирования Nsight Compute, который собирает данные на уровне отдельных вычислительных ядер CUDA. Собирается около 50 телеметрических параметров, включающих загрузку вычислительных блоков, активность специализированных модулей, паттерны доступа к памяти, статистику выполнения инструкций. Данная телеметрия формирует уникальный «цифровой отпечаток» каждого типа операции. Последовательность операций формирует нейросетевой слой. В рамках работы был создан генератор синтетических моделей и автоматизированная система сбора датасета. Для классификации типов слоев обучена мета-модель. Задача машинного обучения сформулирована как маркировка последовательности: по последовательности векторов телеметрических параметров вычислительных ядер предсказывается последовательность типов соответствующих слоев нейронной сети. На валидационной выборке достигнута общая точность классификации 88%, micro-average precision/recall/F1-score = 0.88, weighted-average precision = 0.89, weighted-average F1-score = 0.86.

Выводы.

Предложенный механизм восстановления типов слоев нейронных сетей по телеметрии GPU экспериментально подтверждает жизнеспособность подхода на основе побочных каналов. Практическое применение результатов включает аудит безопасности сторонних моделей и защиту собственных архитектур от утечек через побочные каналы. Дальнейшее развитие предполагает расширение обучающей выборки для повышения качества классификации, автоматическую разметку, восстановление топологии.

Список использованных источников:

1. Oh, S.J. Towards Reverse-Engineering Black-Box Neural Networks / S.J. Oh, M. Augustin, B. Schiele, M. Fritz // arXiv. – 2018. – arXiv:1711.01768.

2. Luo, Y. NNReArch: A Tensor Program Scheduling Framework Against Neural Network Architecture Reverse Engineering / Y. Luo, S. Duan, C. Gongye, Y. Fei, X. Xu // arXiv. – 2022. – arXiv:2203.12046.

3. Alam, M. How Secure are Deep Learning Algorithms from Side-Channel based Reverse Engineering? / M. Alam, D. Mukhopadhyay // arXiv. – 2018. – arXiv:1811.05259.

4. Smith, J.T.H. Reverse Engineering Recurrent Neural Networks with Jacobian Switching Linear Dynamical Systems / J.T.H. Smith, S.W. Linderman, D. Sussillo // arXiv. – 2021. – arXiv:2111.01256.

5. NVIDIA. Nsight Compute Profiling Guide [Электронный ресурс] // NVIDIA Documentation. – 2025. – URL: https://docs.nvidia.com/nsight-compute/.