# Верифкация ИНС на основе тензорного параллелизма

***Воробьев Сергей Юрьевич***

*Студент*

*Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия*

*E-mail:* sergei.vorobyov01@yandex.ru

***Научный руководитель*** *— Ильюшин Евгений Альбинович*

Формальная верификация нейронных сетей — задача математического доказательства того, что сеть удовлетворяет заданным спецификациям (например, робастность к состязательным возмущениям входных данных) для всех допустимых входов. Для сетей с функцией активации ReLU задача является NP-полной [1], что обуславливает экспоненциальный рост вычислительных затрат с увеличением размера модели. Ведущим инструментом верификации является Alpha-Beta-CROWN — победитель международного соревнования VNN-COMP в 2021–2024 годах [2]. Инструмент использует метод распространения линейных границ (Linear Relaxation based Perturbation Analysis, LiRPA) с оптимизируемыми параметрами релаксации $\alpha$ и двойственными множителями Лагранжа $\beta$, реализованный в виде операций над тензорами на GPU в библиотеке `auto_LiRPA` [3].

С ростом масштабов верифицируемых моделей (Transformers, большие языковые модели) возникает проблема ограниченной памяти и вычислительных ресурсов одного GPU. В области обучения нейронных сетей аналогичная проблема решается методами параллелизма, в частности Tensor Parallelism (TP), впервые предложенным в работе [4]. Цель настоящего исследования — адаптация TP к задаче верификации нейронных сетей методом распространения границ.

Основная идея состоит в распределении весовых матриц линейных слоёв между несколькими GPU с параллельным выполнением обратного распространения границ (CROWN backward pass [5]). Реализованы два типа параллельных слоёв: Column Parallel, в котором матрица весов разделяется по выходной размерности, а результаты агрегируются операцией AllReduce; и Row Parallel, в котором матрица разделяется по входной размерности, и коммуникация при обратном проходе не требуется. Чередование этих слоёв образует схему, минимизирующую объём межпроцессных коммуникаций.

Реализация выполнена в виде расширений auto_LiRPA для распределённых вычислений через коммуникации между GPU. При использовании $N$ GPU потребление памяти весов и матриц границ на

каждом ускорителе снижается в $N$ раз: для MLP-блока с размерностью $d = 4096$ это соответствует сокращению с 512 до 256 МБ при $N = 2$. Замеры на конфигурации из двух NVIDIA A100, соединённых NVLink, показали ускорение вычисления границ CROWN в $\sim$1,97 раза при накладных расходах на коммуникации AllReduce менее 2% от общего времени. Полученные результаты демонстрируют принципиальную возможность масштабирования формальной верификации нейронных сетей на несколько GPU, что открывает путь к верификации более крупных архитектур.

**Литература**

1. Robustness Verification in Neural Networks // arXiv preprint arXiv:2403.13441. 2024.
2. The Fifth International Verification of Neural Networks Competition (VNN-COMP 2024): Summary and Results // arXiv preprint arXiv:2412.19985. 2024.
3. Xu K., Shi Z., Zhang H. et al. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond // Advances in Neural Information Processing Systems 33 (NeurIPS). 2020.
4. Shoeybi M., Patwary M., Puri R. et al. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism // arXiv preprint arXiv:1909.08053. 2020.
5. Zhang H., Weng T.-W., Chen P.-Y. et al. Efficient Neural Network Robustness Certification with General Activation Functions // Advances in Neural Information Processing Systems 31 (NeurIPS). 2018.