УДК 004.254

ПРИМЕНЕНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ В БЛОКАХ ПРЕДВЫБОРКИ ДАННЫХ В ПРОЦЕССОРАХ

Прошкин Н.А. (Университет ИТМО)
Научный руководитель - к.т.н. Кустарев П.В. (Университет ИТМО)

Введение. Производительность современных процессоров критически зависит от эффективности механизмов предвыборки данных (data prefetching), нужных чтобы компенсировать разницу между скоростью вычислений и скоростью доступа к оперативной памяти - явление, известное как “стена памяти” (Memory Wall). Классические эвристические алгоритмы предвыборки (Stride, Best Offset, Signature Path Prefetcher) демонстрируют высокую точность на программах с регулярными паттернами доступа, однако их эффективность падает при обработке современных нагрузок с нерегулярной структурой данных, таких как обход графов и pointer-chasing.

Исследования (например, работы Voyager [1], Traced [2]) показали, что блоки предвыборки на основе нейронных сетей способны выявлять сложные зависимости в потоке обращений к памяти, достигая покрытия (coverage) до 60% и точности (accuracy) до 90% на нерегулярных нагрузках. Однако применение этих решений в аппаратуре в текущем виде проблематично: задержка вывода (inference, инференс) составляет несколько сотен тактов при допустимом бюджете в десятки тактов для кэша последнего уровня, а необходимый объем памяти достигает нескольких МБ при доступных единицах-десятках КБ. Соответственно, требуется провести оптимизацию отдельных блоков и в целом префетчера, чтобы удовлетворять указанным (и иным подобным) ограничениям.

Однако, без понимания структуры, функций и устройства отдельных блоков префетчера и взаимосвязей между ними, невозможно определить эффективные направления оптимизации, а также оценку возможности интеграции таких блоков в конвейер современного процессора.

Основная часть. В рамках доклада предлагается декомпозиция процесса нейросетевой предвыборки данных на четыре последовательных этапа: предобработка входных данных (preprocessing), формирование представления (representation), инференс нейронной сети и генерация адреса (prediction generation). Для каждого этапа выполняется анализ вариантов реализации с оценкой их влияния на латентность, точность предсказания и требования к аппаратным ресурсам.

1. **Этап предобработки** определяет способ извлечения и структурирования признаков из сырого потока обращений к памяти. В некоторых подходах используют необработанные последовательности адресов и признаков (программный счетчик, разложение на страницу и смещение), но есть и более сложные варианты, например, группировка запросов по схожести признаков (кластеризация).
2. **Этап формирования представления** определяет структуру данных, подаваемых на вход нейронной сети. Здесь активно используют embedding-слои (что при миллионах уникальных адресов приводит к значительным затратам памяти), а также можно выделить интересный подход с графовыми моделями, позволяющими закодировать информацию о том, как регионы памяти логически связаны алгоритмом программы (например, пространственно-временные графы в Traced [2]).
3. **Этап инференса** нейросети является центральным с точки зрения точности предсказания и аппаратной сложности. В большинстве работ применяются рекуррентные сети (LSTM, GRU), также делаются попытки применения

графовых сетей, как например в Traced [2], где используется сочетание графовых сверток (GCN) и рекуррентных блоков.

4. **Этап генерации адреса** применяет специфические решения, продиктованные проблемой взрывного роста размерности пространства возможных исходов (Class Explosion Problem): прямое предсказание 64-битного адреса невозможно из-за огромного числа классов. Существующие подходы включают декомпозицию задачи на независимые предсказания компонентов адреса (например, раздельное определение страницы и смещения с использованием нескольких выходных слоев), либо сведение задачи к классификации среди ограниченного набора наиболее вероятных смещений (дельт) относительно базового адреса.

Выводы. Цель доклада - представить обзор одной из возможных декомпозиций конвейера блока предвыборки данных на базе нейронной сети и существующих решений для каждой ступени этого конвейера. Результаты в дальнейшем предполагается использовать для разработки методов оптимизации нейросетевых предсказателей и их аппаратных реализаций, обеспечивающих минимизацию задержки инференса и занимаемой площади.

Список использованных источников.

1. Shi Z., Jain A., Swersky K., Hashemi M., Ranganathan P., Lin C. A hierarchical neural model of data prefetching // Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. – 2021. – С. 861-873.
2. Jiang H., Fu L., Liu D., Ren Z., Chen Y., Qiao L. TRACED: A Temporal Graph Neural Networks-based // ACM Transactions on Architecture and Code Optimization. – 2025. – Т. 22. – №. 3. – С. 1-25.