ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.31

СРАВНЕНИЕ АППАРАТНЫХ УСКОРИТЕЛЕЙ ДЛЯ НЕЙРОСЕТЕЙ: CPU, NPU И IGPU НА БАЗЕ RYZEN 7 8700

Фадеев Вячеслав Олегович,
Студент группы ИУК5-71Б
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
fadeevvo@student.bmstu.ru

Ермаков Ярослав Владиславович,
Студент группы ИУК5-71Б
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
ermakovyav@student.bmstu.ru

Вершинин Евгений Владимирович,
Кандидат физико-математических наук, доцент и заведующий кафедрой «Системы обработки информации»
Калужский филиал Московского государственного технического университета имени Н.Э. Баумана
vershinin@bmstu.ru

Аннотация

В данной работе проводится сравнение аппаратных ускорителей для нейросетей: CPU, NPU и iGPU на базе Ryzen 7 8700, с использованием модели YOLOv8 в формате ONNX и квантованием. Анализируются производительность, энергоэффективность и применимость каждого ускорителя в различных вычислительных задачах. Работа включает теоретический обзор архитектурных особенностей ускорителей, экспериментальное тестирование и сравнение их производительности на практических примерах.

Ключевые слова: аппаратные ускорители, нейросети, CPU, NPU, iGPU, глубокое обучение, YOLOv8, квантование, энергоэффективность, производительность.

COMPARISON OF GAS PEDALS FOR NEURAL NETWORKS: CPU, NPU AND IGPU BASED ON RYZEN 7 8700

Vyacheslav O. Fadeev,
Student of group IUK5-71B
Bauman Moscow State Technical University (Kaluga Branch
fadeevvo@student.bmstu.ru

Yaroslav V. Ermakov,
Student of group IUK5-71B
Bauman Moscow State Technical University (Kaluga Branch
ermakovyav@student.bmstu.ru

Evgeny V. Vershinin,
Ph.D, Associate Professor and Head of the department "Information Processing Systems"
Bauman Moscow State Technical University (Kaluga Branch
vershinin@bmstu.ru

ABSTRACT

This paper compares hardware accelerators for neural networks: CPU, NPU, and iGPU based on Ryzen 7 8700, utilizing the YOLOv8 model in ONNX format with quantization. It analyzes the performance, energy efficiency, and applicability of each accelerator in various computational tasks. The work includes a theoretical overview of the architectural features of the accelerators, experimental testing, and comparison of their performance on practical examples.

Keywords: Hardware accelerators, neural networks, CPU, NPU, iGPU, deep learning, YOLOv8, quantization, energy efficiency, performance.

Введение

Современные нейронные сети требуют значительных вычислительных ресурсов, что привело к активному развитию аппаратных ускорителей. Среди таких решений выделяются центральные процессоры (CPU), нейронные процессоры (NPU) и интегрированные графические процессоры (iGPU). Каждая из этих архитектур предлагает свои подходы к обработке задач глубокого обучения, таких как детекция объектов с использованием моделей, например, YOLOv8.

Цель данной работы — сравнить производительность CPU, NPU и iGPU на примере модели YOLOv8 в формате ONNX с применением квантования. Исследование включает теоретический анализ архитектур, экспериментальные измерения и выводы о преимуществах и недостатках каждого решения.

Теоретический обзор: CPU, NPU и iGPU

Аппаратные ускорители стали ключевыми элементами для выполнения задач глубокого обучения (DL), таких как обработка изображений, классификация данных и другие ресурсоемкие операции. CPU, NPU и iGPU обладают уникальными архитектурными особенностями и технологиями, которые влияют на их производительность, энергоэффективность и применимость в различных задачах [1].

Центральный процессор (CPU)

CPU (Central Processing Unit) — это универсальное устройство для обработки данных, являющееся ядром любой компьютерной системы. CPU используется для выполнения практически всех вычислительных задач, включая задачи глубокого обучения. Архитектура CPU включает ограниченное количество высокопроизводительных ядер, которые способны эффективно обрабатывать последовательные и многопоточные вычисления [1].

CPU чаще всего используется для подготовки данных, обучения небольших моделей, выполнения задач на устройствах общего назначения и в случаях, когда не требуется специализированное оборудование.

Архитектурные особенности

1. Универсальность: CPU имеет гибкую архитектуру, поддерживающую широкий спектр инструкций, что позволяет ему выполнять самые разные вычислительные задачи [2].

2. Кэш-память: Современные процессоры оснащены многоуровневой кэш-памятью (L1, L2, L3), что обеспечивает высокую скорость доступа к данным.

3. Многопоточность: Поддержка технологий, таких как Hyper-Threading, позволяет запускать несколько потоков одновременно.

4. Поддержка SIMD: CPU может выполнять векторные операции с помощью SIMD (Single Instruction, Multiple Data), что улучшает производительность в обработке данных.

Преимущества CPU

- Гибкость и универсальность: CPU поддерживает широкий спектр приложений, от обработки текстов до сложных научных вычислений [3].

- Широкая совместимость: Поддержка всех основных операционных систем и программных платформ делает CPU универсальным инструментом.

- Последовательные вычисления: оптимально подходит для задач, где требуется строго последовательное выполнение операций.

Недостатки CPU

- Низкая энергоэффективность: для задач глубокого обучения и массивно-параллельных вычислений CPU потребляет значительно больше энергии по сравнению с NPU или GPU [3].

- Ограниченная производительность в параллельных вычислениях: хотя CPU может выполнять несколько потоков одновременно, его архитектура менее эффективна в параллельных вычислениях по сравнению с GPU или NPU.

Нейронный процессор (NPU)

NPU (Neural Processing Unit) — это специализированный процессор, разработанный исключительно для ускорения задач машинного обучения (ML) и глубокого обучения (DL). NPU оптимизирован для выполнения операций, связанных с обработкой больших массивов данных, таких как свертки, матричные умножения и другие линейные операции, часто используемые в нейронных сетях [4].

NPU находит применение в мобильных устройствах, встраиваемых системах и облачных платформах, где требуется высокая энергоэффективность и производительность в задачах DL.

Архитектурные особенности

1. Массивная параллельность: NPU содержит множество специализированных ядер, которые обрабатывают матрицы и векторы с высокой скоростью [3].

2. Оптимизация под тензорные операции: Архитектура NPUs позволяет эффективно выполнять операции с тензорами, такие как умножение матриц или свертки.

3. Поддержка INT8 и FP16: Использование низко разрядной арифметики (INT8, FP16) для выполнения операций позволяет добиться высокой производительности с минимальными затратами энергии [4].

4. Энергоэффективность: Аппаратная оптимизация под нейронные операции минимизирует энергопотребление по сравнению с CPU и GPU.

Преимущества NPU

- Высокая энергоэффективность: NPU обеспечивает оптимальную производительность при низком энергопотреблении, что особенно важно для мобильных и встроенных устройств.

- Ускорение операций нейросетей: Специализация на тензорных вычислениях делает NPU идеальным решением для задач DL.

- Минимальная задержка: NPU эффективно выполняет задачи в режиме реального времени, такие как детекция объектов.

Недостатки NPU

- Ограниченный набор задач: NPU оптимизирован исключительно под нейронные вычисления и плохо подходит для универсальных вычислительных задач [5].

- Сложности интеграции: В отличие от CPU, использование NPU требует специализированного ПО и драйверов, что может осложнить разработку и внедрение.

Интегрированный графический процессор (iGPU)

iGPU (Integrated Graphics Processing Unit) — это графический процессор, встроенный в центральный процессор. Первоначально предназначенный для обработки графики, iGPU также подходит для выполнения параллельных вычислений, что делает его эффективным для задач машинного обучения.

iGPU используется в бюджетных системах, ноутбуках и встраиваемых устройствах для ускорения нейросетевых вычислений в тех случаях, когда дискретный GPU или NPU нецелесообразны [5].

Архитектурные особенности

1. Массивно-параллельная обработка: iGPU состоит из множества ядер, которые способны выполнять вычисления независимо друг от друга.

2. Встроенность в CPU: Использование общей с CPU памяти снижает задержки при передаче данных.

3. Оптимизация для массовых вычислений: iGPU оптимально обрабатывает массивные параллельные операции, такие как матричные вычисления.

4. Экономичность: Совмещение CPU и GPU на одном чипе снижает стоимость и энергопотребление системы.

Преимущества iGPU

- Универсальность: iGPU подходит как для графических задач, так и для ускорения нейронных вычислений.

- Доступность: Встроенные в процессор iGPU обходятся дешевле дискретных GPU.

- Энергоэффективность: Совмещение с CPU позволяет сократить энергопотребление.

Недостатки iGPU

- Ограниченные ресурсы памяти: iGPU использует оперативную память системы, что может ограничить его производительность в больших задачах [3].

- Низкая производительность: По сравнению с дискретными GPU, iGPU имеет более низкую производительность в задачах глубокого обучения.

Экспериментальная часть

Для тестирования была выбрана модель YOLOv8 в формате ONNX, квантованная для повышения производительности. Модель запускалась на трех аппаратных ускорителях: CPU, NPU и iGPU. Тестирование проводилось на ПК с следующими характеристиками:

- Процессор: AMD Ryzen 7 8700 (включая NPU и iGPU).

- Оперативная память: 128 ГБ DDR5.

- Операционная система: Windows 11 Pro.

- Модель: YOLOv8 с квантованием для ускорения вычислений.

Методика эксперимента

1. Загрузка модели YOLOv8 в формате ONNX: Модель оптимизирована для использования на разных типах ускорителей.

2. Измерение производительности: Мониторинг производительности включал:
- Использование процессора и памяти.
- Частоту кадров (FPS).

Результаты экспериментов

Производительность на NPU. График представлен на рисунке 1.

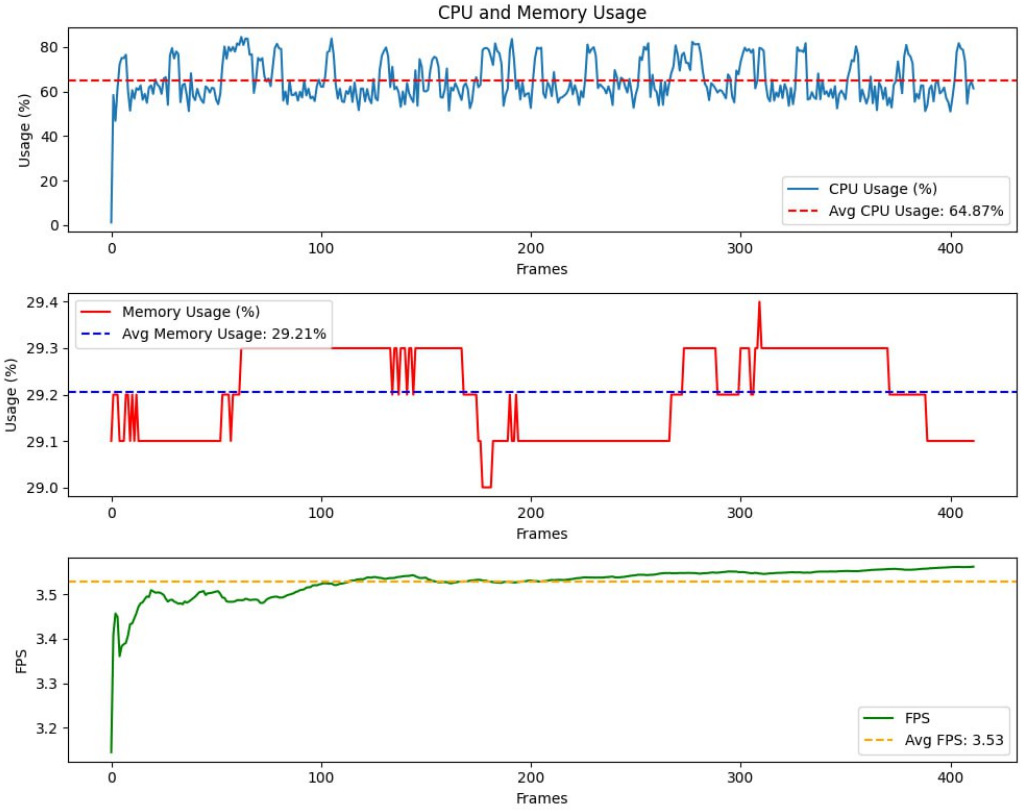

Рисунок 1 - График использования процессора, памяти и FPS для NPU

2. Производительность на CPU. График представлен на рисунке 2.

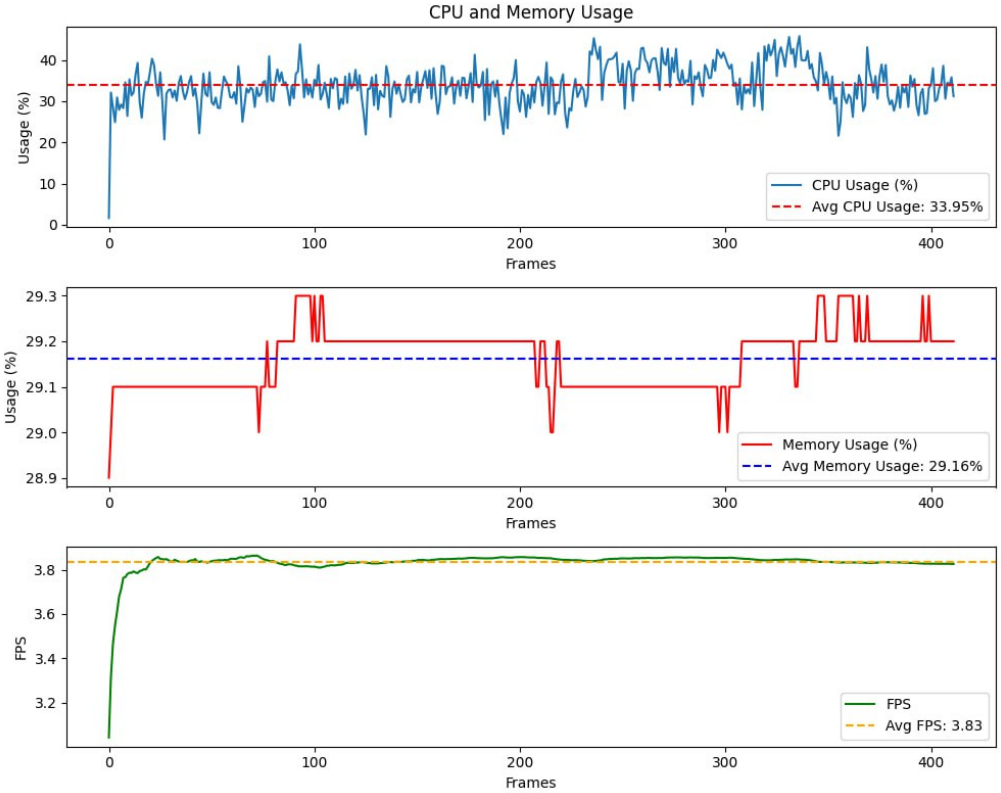

Рисунок 2 - График использования процессора, памяти и FPS для CPU

3. Производительность на iGPU. График представлен на рисунке 3.

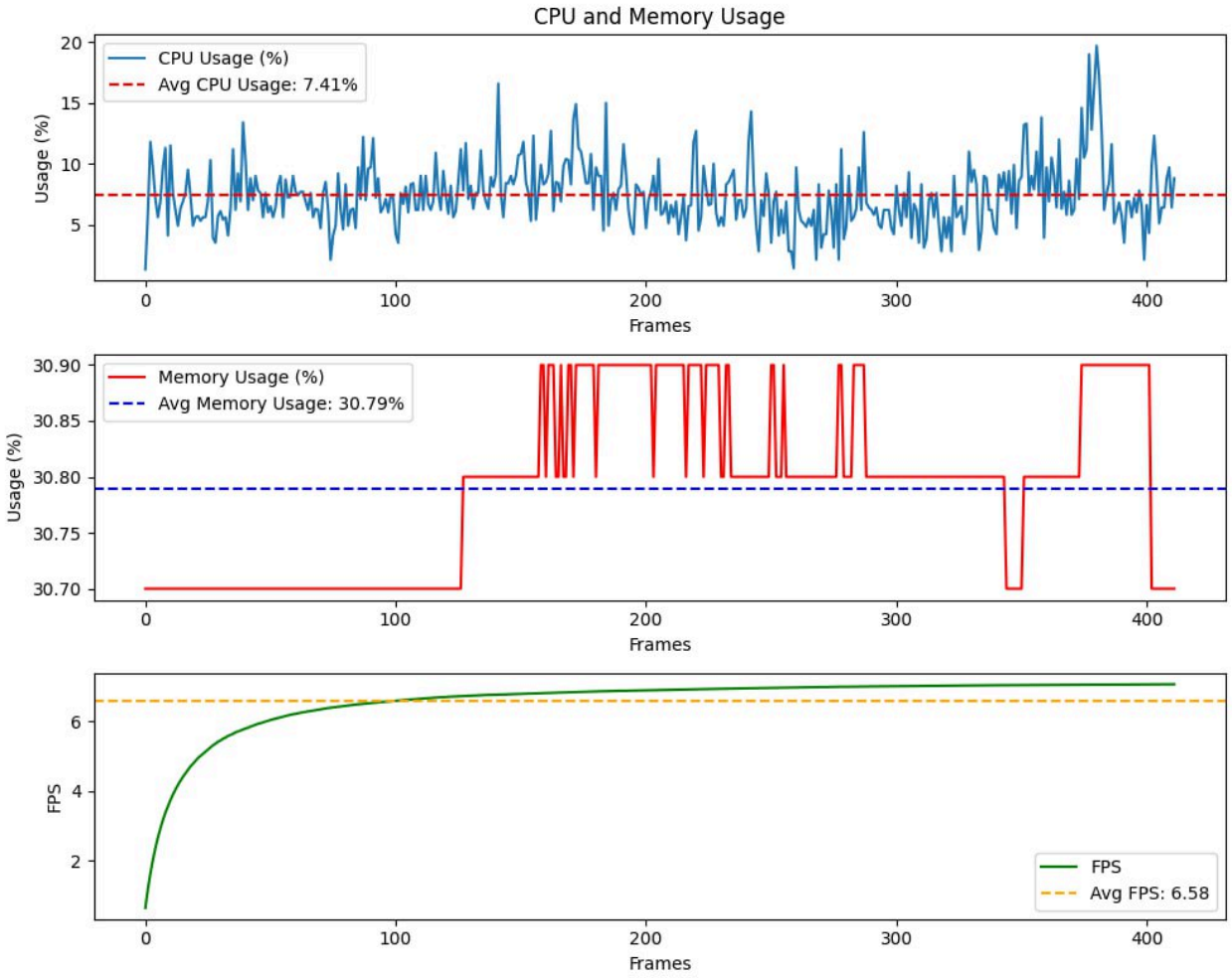

Рисунок 3 - График использования процессора, памяти и FPS для iCPU

В Таблице 1 представлены результаты сравнительного анализа производительности трех аппаратных ускорителей: CPU, NPU и iGPU, на примере модели YOLOv8 в формате ONNX с применением квантования.

Таблица 1. Производительность аппаратных ускорителей

Параметр	NPU	CPU	iGPU
Средняя загрузка CPU, %	64.87%	33.95%	7.41%
Среднее использование памяти, %	29.21%	29.16%	30.79%
Среднее значение FPS	3.53	3.83	6.58

Заключение

На основе проведённого анализа и тестирования можно сделать следующие выводы:

По нагрузке на CPU: iGPU демонстрирует самую низкую загрузку процессора (7.41%), что делает его наиболее подходящим выбором для задач, где требуется минимальное использование CPU. NPU показывает наибольшую нагрузку (64.87%), что может быть связано с архитектурными особенностями специализированных вычислений. CPU занимает среднее положение с загрузкой 33.95%.

Использование памяти: Все ускорители демонстрируют схожий уровень использования памяти. Однако, iGPU показывает немного более высокое значение (30.79%), что может быть связано с использованием общей оперативной памяти системы. Это незначительное увеличение компенсируется другими преимуществами iGPU.

Среднее значение FPS (кадров в секунду): iGPU демонстрирует наибольшее количество кадров в секунду (6.58 FPS), опережая как NPU (3.53 FPS), так и CPU (3.83 FPS). Это делает iGPU наиболее производительным решением для тестируемой задачи, особенно при использовании квантованных моделей.

В целом можно сделать следующие выводы: iGPU на базе Ryzen 7 8700 демонстрирует лучший баланс между производительностью, энергоэффективностью и загрузкой процессора, что делает его идеальным выбором для сценариев, где важны универсальность и доступность. CPU остаётся хорошим решением для задач с ограниченной параллельностью, где требуется универсальность и предсказуемость. NPU, несмотря на свою специализированность, в данном тесте уступает iGPU и CPU. Это может быть связано с недостаточной оптимизацией модели под архитектуру NPU или спецификой реализации аппаратной платформы. Таким образом, iGPU оказывается наиболее перспективным выбором для ускорения нейросетевых вычислений на бюджетных или энергоэффективных системах без GPU, особенно с использованием квантованных моделей.

Список литературы:

1. Ardavan Pedram & Kunle Olukotun, "Hardware Accelerators for Training Deep Neural Networks", [Электронный ресурс] URL: https://web.stanford.edu/~perdavan/DNNTrain/ (дата обращения 01.10.2024)
2. MIT, "Tutorial on Hardware Accelerators for Deep Neural Networks", [Электронный ресурс] URL: https://eyeriss.mit.edu (дата обращения 01.10.2024)
3. Lukas Baischer, Matthias Wess, "Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors", [Электронный ресурс] URL: https://ar5iv.labs.arxiv.org/html/2104.09252 (дата обращения 01.10.2024)
4. S M Mojahidul Ahsan et al., "Hardware Accelerators for Artificial Intelligence", [Электронный ресурс] URL: https://ar5iv.org/pdf/2411.13717 (дата обращения 01.10.2024)
5. Liangzhen Lai, "Recent Advances in Efficient Computation of Deep Convolutional Neural Networks", [Электронный ресурс] URL: https://ar5iv.labs.arxiv.org/html/1802.00939 (дата обращения 01.10.2024)

References:

1. Ardavan Pedram & Kunle Olukotun, "Hardware Accelerators for Training Deep Neural Networks", [Online Resource] URL: https://web.stanford.edu/~perdavan/DNNTrain/ (accessed on 01.10.2024)
2. MIT, "Tutorial on Hardware Accelerators for Deep Neural Networks", [Online Resource] URL: https://eyeriss.mit.edu (accessed on 01.10.2024)
3. Lukas Baischer, Matthias Wess, "Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors", [Online Resource] URL: https://ar5iv.labs.arxiv.org/html/2104.09252 (accessed on 01.10.2024)
4. S M Mojahidul Ahsan et al., "Hardware Accelerators for Artificial Intelligence", [Online Resource] URL: https://ar5iv.org/pdf/2411.13717 (accessed on 01.10.2024)

5. Liangzhen Lai et al., "Recent Advances in Efficient Computation of Deep Convolutional Neural Networks", [Online Resource] URL: https://ar5iv.labs.arxiv.org/html/1802.00939 (accessed on 01.10.2024).