https://doi.org/10.26160/2474-5901-2026-55-65-71

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ЭНЕРГОЭФФЕКТИВНОСТИ МОДЕЛЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ НА ВСТРАИВАЕМОМ НЕЙРОСЕТЕВОМ УСКОРИТЕЛЕ: ПЛАТФОРМА ALLWINNER A733

Абдрахманов Д.С.
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, Россия

Ключевые слова: энергоэффективность, нейросетевой ускоритель, квантование нейронных сетей, встраиваемые системы, компьютерное зрение, ваттметр, тепловыделение, одноплатный компьютер, Allwinner A733.
Аннотация. В статье представлена экспериментальная оценка энергетической эффективности инференса моделей компьютерного зрения на встраиваемом нейросетевом ускорителе (NPU) Allwinner A733 в сравнении с центральным процессором (CPU). Исследован инференс пяти архитектур (YOLOv8, YOLOv11, YOLOv26, Fast-SCNN, DeepLabV3+) в формате FP32 на CPU и в трех форматах квантования (UINT8, PCQ, INT16) на NPU. Измерения выполнялись внешним USB-ваттметром UM34C методом вычитания мощности холостого хода. Установлено, что чистая мощность инференса на NPU (0,65-1,03 Вт) в 5,3-8,6 раза ниже мощности на CPU (5,10-5,77 Вт), а прирост температуры активного блока NPU на порядок меньше аналогичного показателя CPU.

EXPERIMENTAL EVALUATION OF THE ENERGY EFFICIENCY OF COMPUTER VISION MODELS ON AN EMBEDDED NEURAL NETWORK ACCELERATOR: THE ALLWINNER A733 PLATFORM

Abdrakhmanov D.S.
Saint-Petersburg Electrotechnical University “LETI”, Saint-Petersburg, Russia

Keywords: energy efficiency, neural processing unit, neural network quantization, embedded systems, computer vision, power meter, heat dissipation, single-board computer, Allwinner A733.
Abstract. This paper presents an experimental evaluation of the energy efficiency of computer vision model inference on the Allwinner A733 embedded NPU compared to the CPU. Five architectures (YOLOv8, YOLOv11, YOLOv26, Fast-SCNN, DeepLabV3+) were evaluated in FP32 on CPU and in three quantization formats (UINT8, PCQ, INT16) on NPU. Power consumption was measured with an external USB power meter (UM34C) by subtracting idle power. Net NPU inference power (0.65-1.03 W) was found to be 5.3-8.6 times lower than FP32 CPU inference power (5.10-5.77 W). The results confirm that NPUs provide a highly energy-efficient alternative for computer vision deployment on embedded platforms.

Введение

Ресурсоёмкость современных моделей детекции и сегментации ограничивает их применение во встраиваемых системах реального времени. Облачный инференс увеличивает сетевую задержку, поэтому для автономных устройств предпочтительны периферийные вычисления, где энергопотребление определяет время работы, тепловой режим и охлаждение [1]. Благодаря аппаратным целочисленным MAC-операциям NPU могут снижать энергозатраты относительно CPU в 7-125 раз [1]. Квантование

уменьшает требования к ресурсам, но его эффект зависит от поддержки операторов в SDK [2]. Продолжая исследование [2], настоящая работа оценивает мощность и тепловой режим NPU Allwinner A733 относительно CPU при инференсе моделей компьютерного зрения и определяет влияние формата квантования на энергопотребление и тепловыделение.

Материалы и методы

Аппаратная платформа

Эксперименты проведены на одноплатном компьютере Radxa Qubit A7A с SoC Allwinner A733 (CPU, NPU и GPU), 8 ГБ оперативной памяти и Debian (Linux arm64). В отличие от μNPU-плат с жёсткими ограничениями RAM/SRAM [3], 8 ГБ памяти позволяют исследовать лёгкие и тяжёлые архитектуры.

Для NPU модели квантовали и компилировали в Acuity Toolkit; на CPU использовали ONNX Runtime и FP32. GPU не рассматривался, поскольку используемые драйверы Linux arm64 не обеспечивают выполнение требуемых вычислений.

Исследуемые архитектуры

Исследованы детекторы YOLOv8, YOLOv11 и тяжёлая YOLOv26 с расширенными feature pyramid-блоками, а также сегментационные Fast-SCNN и DeepLabV3+ с Atrous Spatial Pyramid Pooling. Для каждой архитектуры испытаны FP32 на CPU и UINT8, PCQ, INT16 на NPU – 20 конфигураций.

Методика измерения энергопотребления

Из-за отсутствия штатного мониторинга на многих встраиваемых устройствах [4] и возможной дополнительной нагрузки программных средств [5] мощность измеряли внешним USB-ваттметром UM34C в цепи питания. Аналогичный подход применялся для встраиваемых ускорителей, Jetson Nano и Raspberry Pi [4, 6].

Чистую мощность вычисляли как разность мощности при инференсе и в простое [7]. Внешний прибор не зависит от программной поддержки и обеспечивает более частую регистрацию, чем большинство встроенных средств [5].

Каждую конфигурацию трижды запускали примерно на 60 с; рассчитывали среднее и стандартное отклонение. Повторы соответствуют практике повышения надёжности энергетических оценок [8].

Температурный мониторинг

Показания встроенных термозон регистрировали примерно раз в секунду как независимый индикатор энергетической нагрузки [1].

Дизайн эксперимента

В каждом прогоне пять моделей последовательно выполнялись на CPU/FP32, затем на NPU в UINT8, PCQ и INT16 по 60 с на конфигурацию. Простой составлял 30 с до начала, 90 с между блоками CPU и NPU и 30 с после

измерений. Получено 20 конфигураций × 3 повтора = 60 измерений чистой мощности и температурного отклика.

Результаты

На рисунке 1 представлена общая временная развёртка мощности и температуры NPU для всех трёх прогонов эксперимента. Видно, что фазы инференса FP32 на CPU (мощность 8,5-10,5 Вт) чётко отделены по уровню мощности от фаз инференса на NPU (мощность 3,5-5 Вт), а кривые трёх прогонов практически совпадают на всём протяжении эксперимента, что свидетельствует о высокой повторяемости измерений.

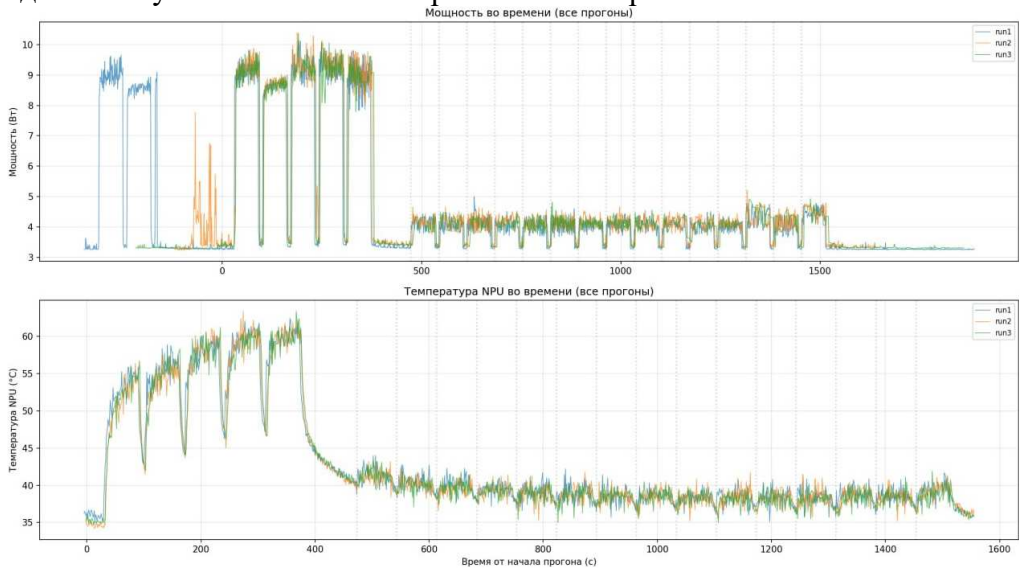

Рис. 1. Мощность и температура NPU во времени для трёх прогонов эксперимента

В таблице 1 представлены результаты мощности инференса.

Табл. 1. Чистая мощность инференса по моделям и форматам, Вт

Модель	**FP32 / CPU**	**UINT8 / NPU**	**PCQ / NPU**	**INT16 / NPU**
YOLOv8	5,770±0,111	0,696±0,077	0,703±0,011	0,732±0,045
YOLOv11	5,762±0,102	0,708±0,049	0,663±0,049	0,697±0,020
YOLOv26	5,531±0,171	0,683±0,025	0,691±0,031	0,682±0,057
Fast-SCNN	5,099±0,024	0,706±0,004	0,645±0,031	0,648±0,017
DeepLabV3+	5,428±0,055	0,951±0,062	0,706±0,010	1,034±0,073

Таким образом, чистая мощность инференса на NPU в 5,3-8,6 раза ниже мощности инференса FP32 на CPU в зависимости от модели и формата.

В таблице 2 приведён прирост температуры активного вычислительного блока: для CPU – относительно температуры простоя непосредственно перед началом нагрузки; для NPU – относительно той же исходной точки отсчёта.

Прирост температуры активного блока при инференсе FP32 на CPU составил от 27,9°C (YOLOv26) до 36,1°C (YOLOv8); при инференсе на NPU во

всех форматах квантования – от 3,0°C до 6,1°C. В рамках NPU-блока температура монотонно снижалась во всех трёх прогонах.

Табл. 2. Прирост температуры активного блока по моделям и форматам

Модель	FP32 / CPU, °C	UINT8 / NPU, °C	PCQ / NPU, °C	INT16 / NPU, °C
YOLOv8	36,088±1,221	6,069±0,471	4,987±0,489	4,479±0,329
YOLOv11	34,764±0,192	4,048±0,549	3,675±0,469	3,576±0,554
YOLOv26	27,851±0,429	3,270±0,658	3,262±0,579	3,049±0,639
Fast-SCNN	29,916±0,536	3,030±0,664	3,091±0,634	3,030±0,667
DeepLabV3+	35,732±0,758	3,296±0,628	3,197±0,584	3,786±0,646

Сопоставление латентности тех же конфигураций, полученной методом Single-Frame Profiler в предшествующей работе автора [2], показывает, что для моделей переход с UINT8 на INT16 увеличивает латентность практически в два раза, тогда как соответствующий рост чистой мощности для тех же конфигураций составляет 5,2% и –1,6%, то есть существенно меньше прироста латентности.

Для архитектуры YOLOv26 и формата PCQ, у которых в работе [2] зафиксирован аномальный рост латентности относительно UINT8/INT16, чистая мощность не выходит за пределы общего диапазона значений, наблюдаемых для изучаемых форматов.

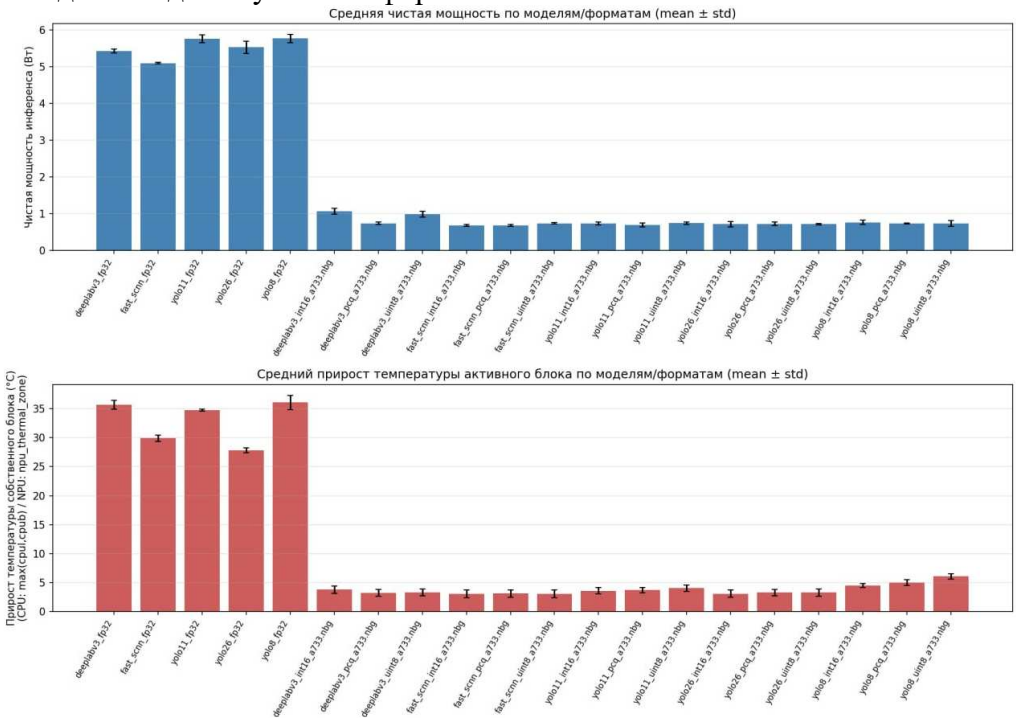

Рис. 2. Средняя мощность инференса (сверху) и средний прирост температуры (снизу)

Обсуждение

Энергетическое превосходство NPU

Снижение чистой мощности на NPU в 5,3-8,6 раза согласуется с выигрышем 7-125 раз для других встраиваемых NPU [1]. Причина – выполнение целочисленных MAC-операций специализированными блоками при более низком напряжении, чем FP-операции универсальных ядер CPU [1, 3].

Разрядность формата квантования и энергопотребление

Переход UINT8→INT16 почти удваивает латентность совместимых детекторов, но изменяет чистую мощность лишь на ~5% для YOLOv8 и почти не влияет на YOLOv11. Следовательно, связь разрядности с латентностью и энергопотреблением нелинейна; аналогичный вывод получен для квантованных языковых моделей [9].

Разделение эффектов: латентностные аномалии и мгновенная мощность

Сопоставление с латентностью [2] разделяет вычислительную интенсивность и накладные расходы software fallback. Для YOLOv26/PCQ латентность достигает 750 мс, но мощность остаётся в общем диапазоне. Сходно, per-group-квантование крупных моделей увеличивало латентность в 8,1-10,7 раза из-за разбиения операций и суммирования в FP [10]. Вероятно, штраф YOLOv26/PCQ обусловлен координацией и передачей данных, а не ростом мгновенной мощности.

Температура NPU: ограничение методики измерения

Прирост температуры NPU 3,0-6,1°C (табл. 2) оценён после 90 с простоя за блоком CPU/FP32.

На рисунке 3 показано наложение температурных кривых CPU и NPU по трём прогонам с разметкой фаз выполнения моделей.

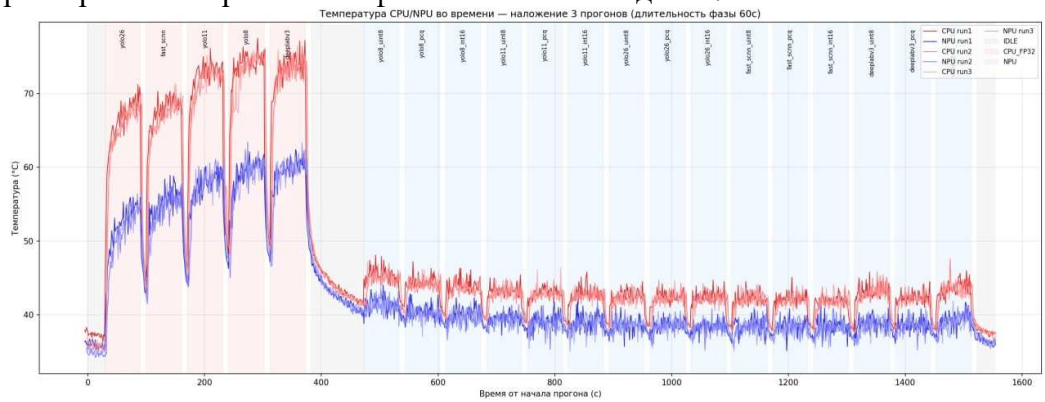

Рис. 3. Температура CPU/NPU во времени, наложение трёх прогонов

Во всех прогонах зона NPU остывала в течение NPU-блока примерно от 41 до 38°C: 90 с недостаточно для теплового равновесия. Значения 3,0-6,1°C поэтому являются верхней границей с остаточным нагревом CPU, но даже они

на порядок ниже прироста при CPU/FP32. В дальнейшем интервал между блоками следует увеличить до теплового равновесия.

Ограничения применённого измерительного оборудования

USB-ваттметр обеспечивает доступные воспроизводимые измерения, но регистрирует суммарную мощность платы с периферией и не изолирует потребление ядра NPU или CPU [8].

Заключение

Эксперимент с пятью архитектурами компьютерного зрения на Allwinner A733 показал:

1) мощность NPU составила 0,645-1,034 Вт – в 5,3-8,6 раза ниже CPU/FP32 (5,099-5,770 Вт); три повтора подтверждают пригодность внешнего USB-ваттметра с вычитанием мощности простоя;

2) аномальная латентность YOLOv26/PCQ без роста мощности указывает на накладные расходы передачи данных;

3) прирост температуры NPU (3,0-6,1°C, верхняя граница) на порядок ниже CPU/FP32 (27,9-36,1°C);

4) GPU A733 неприменим в использованном Linux arm64 из-за отсутствия совместимых драйверов.

NPU обеспечивает энергоэффективный инференс; формат квантования и поддержка операторов определяют латентность, но не обязательно пропорционально изменяют мгновенную мощность.

Список литературы / References

1. Fanariotis A., Orphanoudakis T., Fotopoulos V. Evaluating the Energy Efficiency of NPU-Accelerated Machine Learning Inference on Embedded Microcontrollers // arXiv. 2025. doi.org/10.48550/arXiv.2509.17533.
2. Абдрахманов Д.С. Экспериментальное исследование методов квантования нейросетей на встраиваемом нейросетевом ускорителе: платформа Allwinner A733 // Наукосфера. – 2026. – Т. 6, №1. – С. 246-252. – doi.org/10.5281/zenodo.20759171.
    2. Abdrakhmanov D.S. Experimental Study of Neural Network Quantization Methods on an Embedded Neural Network Accelerator: The Allwinner A733 Platform // Naukosfera. 2026, vol. 6, no. 1, pp. 246-252. doi.org/10.5281/zenodo.20759171.
3. Millar K., Huang X., Sethi G., Madhavapeddy A., Haddadi H. Benchmarking Ultra-Low-Power μNPUs // arXiv. 2025. doi.org/10.48550/arXiv.2503.22567.
4. Wang H., Li X., Zhou T., Lin M. Data-driven Software-based Power Estimation for Embedded Devices // arXiv. 2024. doi.org/10.48550/arXiv.2407.02764.
5. Gholami A., Kim S., Dong Z., Yao Z., Mahoney M.W., Keutzer K. A Survey of Quantization Methods for Efficient Neural Network Inference // arXiv. 2021. doi.org/10.48550/arXiv.2103.13630.
6. Mohammadi M., Smith H., Khan L., Zand R. Facial Expression Recognition at the Edge: CPU vs GPU vs VPU vs TPU // Proceedings of the Great Lakes Symposium on VLSI. 2023, pp. 243-248. doi.org/10.1145/3583781.3590245.
7. Fister Jr. I., Fister D., Podgorelec V., Fister I. Profiling the Carbon Footprint of Performance Bugs // Proceedings of the International Conference on Artificial

Intelligence, Computer, Data Sciences and Applications. 2024. doi.org/10.48550/arXiv.2401.01782.
8. Bartoli P., Veronesi C., Giudici A., Siorpaes D., Trojaniello D., Zappa F. Benchmarking Energy and Latency in TinyML: A Novel Method for Resource-Constrained AI // arXiv. 2025. doi.org/10.48550/arXiv.2505.15622.
9. Husom E.J., Goknil A., Astekin M., Shar L.K., Kæsen A., Sen S., Mithassel B.A., Soylu A. Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency // ACM Transactions on Internet of Things. 2025, vol. 6, iss. 4, p. 28. doi.org/10.1145/3767742.
10. Xu D., Zhang H., Yang L., Liu R., Huang G., Xu M., Liu X. Fast On-device LLM Inference with NPUs // Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. Rotterdam: ACM. 2025, vol. 1, 18 p. doi.org/10.1145/3669940.3707239.

Абдрахманов Данил Сергеевич – магистрант	**Danil Sergeevich Abdrakhmanov** – master's student
abdraxmanov_03@list.ru	

Received 11.07.2026