Автоматизация в промышленности 2024. №09
www.avtprom.ru

DOI: 10.25728/avtprom.2024.09.05

А.А. Жиленков (СПбГМТУ)

Анализ современного состояния проблемы проектирования программно-аппаратных ускорителей вывода моделей глубокого обучения

Описываются перспективы применения программно-аппаратных ускорителей вывода моделей глубокого обучения, а также результаты сравнения аппаратных архитектур ускорителей вывода моделей глубокого обучения. Рассматриваются варианты оптимизации моделей глубокого обучения, функционирующих на различных программно-аппаратных архитектурах, по критериям увеличения пиковой производительности, увеличения эффективности, снижения рабочей нагрузки. Показаны преимущества применения ПЛИС как платформы для развёртывания моделей глубокого обучения.

Ключевые слова: проектирование, программно-аппаратные ускорители, модели глубокого обучения, ПЛИС.

***Жиленков Антон Александрович** – канд. тухн. наук, доцент, директор Института робототехники и интеллектуальных систем, зав. кафедрой киберфизических систем, СПбГМТУ.*

Список литературы

1. Popov A.V., Sayarkin K.S., Zhilenkov A.A. The scalable spiking neural network automatic generation in MATLAB focused on the hardware implementation // Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018. – Pp. 962-965.
2. Marozzo F., Orsino A., Talia D. and Trunfio P. Edge Computing Solutions for Distributed Machine Learning // 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Falerna, Italy. 2022. Pp. 1-8.
3. Ning Z. et al. Dynamic Computation Offloading and Server Deployment for UAV-Enabled Multi-Access Edge Computing // IEEE Transactions on Mobile Computing. 2023. Vol. 22. N 5. Pp. 2628-2644.
4. Muralidharan S., O'Brien K. and Lalanne C. A Semi-Automated Tool Flow for Roofline Anaylsis of OpenCL Kernels on Accelerators. High-performance Reconfigurable Computing (H2RC). 2015. – 8 p.
5. Zhou S., Ni Z., Zhou X., Wen H., Wu Y. and Zou Y. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients // Computing Research Repository (CoRR). – 2016. – 13 p.
6. Krizhevsky A., Sutskever I. and Hinton G. E. Imagenet classification with deep convolutional neural networks // International Conference on Neural Information Processing Systems (NIPS). 2012. Vol. 25. Pp. 1097–1105.
7. Жиленков А.А., Черный С. Г. Извлечение информации из BigData с помощью нейросетевых архитектур как сетей ассоциаций информационных гранул // Тр. Института системного анализа РАН. 2022. Т. 72. № 3. С. 81-90.

Zhilenkov A.A. The analysis of the state of the art of the design of hard-/software accelerators for deep learning model inference

The paper analyzes the application outlook for hard-/software accelerators of deep learning model inference as well as the methodology and design criteria for such accelerators and compares their hardware architectures. It discusses three basic approaches to the optimization of deep learning models for their effective deployment on various hard-/software architectures: peak capacity increase, efficiency improvement, workload reduction. The advantages of PLD as a platform for deep learning model deployment are shown.

Keywords: design, hard-/software accelerators, deep learning models, PLD.