ВикипедиЯ

# Процессор глубокого обучения

Материал из Википедии — свободной энциклопедии

**Процессор глубокого обучения** (**Deep learning processor**, **DLP**) или ускоритель глубокого обучения — это электронная схема, разработанная для алгоритмов глубокого обучения, обычно с отдельной памятью данных и специализированной архитектурой набора команд. Процессоры глубокого обучения варьируются от мобильных устройств, таких как блоки нейронной обработки (NPU) в мобильных телефонах Huawei[1], до серверов облачных вычислений, таких как Тензорный процессор Google (TPU) в Google Cloud Platform[2].

Цель создания специализированных устройств DLP — обеспечить более высокую эффективность и производительность для алгоритмов глубокого обучения, чем обычные центральные процессоры (CPU) и графические процессоры (GPU). Большинство DLP используют большое количество вычислительных компонентов для использования параллелизма на высоком уровне данных, относительно большие буфер / память на кристалле для использования шаблонов повторного использования данных и операторы ограниченной ширины данных для обеспечения устойчивости к ошибкам при глубоком обучении.

## Содержание



# История

### Использование центральных и графических процессоров

Первоначально для выполнения алгоритмов глубокого обучения были адаптированы процессоры общего назначения. Позже для целей глубокого обучения стали использоваться и графические процессоры. Например, в 2012 году Алекс Крижевский использовал два графических процессора для обучения сети глубокого обучения, названной AlexNet[3], которая стала победителем конкурса ISLVRC-2012. Поскольку интерес к алгоритмам глубокого обучения и DLP продолжил расти, производители графических процессоров начинают добавлять функции, связанные с глубоким обучением, как в аппаратное обеспечение (например, операторы INT8), так и в программное обеспечение (например, библиотеку cuDNN). Так, Nvidia выпустила ядро Turing Tensor Core — DLP — для ускорения обработки глубокого обучения.

### Первые DLP

Чтобы обеспечить более высокую эффективность в производительности и энергопотреблении, разработчики оборудования обращают внимание предметно-ориентированный дизайн устройств. В 2014 году команда исследователей под руководством Tianshi Chen предложила первый в мире DLP, DianNao (по-китайски «электрический мозг»)[4], специализированный для ускорения глубоких нейронных сетей. DianNao обеспечивает пиковую производительность 452 Gop / s (ключевых операций в глубоких нейронных сетях) при небольшой занимаемой площади 3,02 мм2 и потребляемой мощности 485 мВт. Следующие версии процессора (DaDianNao[5], ShiDianNao[6], PuDianNao[7]), образующие семейство микросхем DianNao были предложены той же группой разработчиков[8].

### Дальнейшее развитие

После появления семейства процессоров DianNao, аналогичные по идеологии разработки велись как в академических кругах, так и в промышленности. Только на ежегодной Международной конференция по компьютерной архитектуре ISCA 2016 три сессии, 15% (!) принятых докладов описывали проекты архитектуры процессоров глубокого обучения. В числе заслуживающих упоминания проектов можно назвать Eyeriss[9] (Массачусетский технологический институт), EIE[10] (Стэнфорд), Minerva[11] (Гарвард), Stripes[12] (Университет Торонто) - из числа академических работ и TPU[13] (Google), MLU[14] (Cambricon) - из числа промышленных разработок.

## Примечания


1. HUAWEI Reveals the Future of Mobile AI at IFA 2017 | HUAWEI Latest News | HUAWEI Global (https://consumer.huawei.com/en/press/news/2017/ifa2017-kirin970/). *consumer.huawei.com*. Дата обращения: 10 ноября 2021. Архивировано (https://web.archive.org/web/20211110131401/https://consumer.huawei.com/en/press/news/2017/ifa2017-kirin970/) 10 ноября 2021 года.
2. *Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal.* In-Datacenter Performance Analysis of a Tensor Processing Unit (https://dl.acm.org/doi/10.1145/3140659.3080246) (англ.) // ACM SIGARCH Computer Architecture News. — 2017-09-14. — Vol. 45, iss. 2. — P. 1–12. — ISSN 0163-5964 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0163-5964). — doi:10.1145/3140659.3080246 (https://dx.doi.org/10.1145%2F3140659.3080246). Архивировано (https://web.archive.org/web/20211120115739/https://dl.acm.org/doi/10.1145/3140659.3080246) 20 ноября 2021 года.
3. *Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton.* ImageNet classification with deep convolutional neural networks (https://dl.acm.org/doi/10.1145/3065386) (англ.) // Communications of the ACM. — 2017-05-24. — Vol. 60, iss. 6. — P. 84–90. — ISSN 1557-7317 0001-0782, 1557-7317 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0001-0782,). — doi:10.1145/3065386 (https://dx.doi.org/10.1145%2F3065386). Архивировано (https://web.archive.org/web/20211115034158/https://dl.acm.org/doi/10.1145/3065386) 15 ноября 2021 года.

4. *Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu.* DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning (https://dl.acm.org/doi/10.1145/2654822.2541967) (англ.) // ACM SIGARCH Computer Architecture News. — 2014-04-05. — Vol. 42, iss. 1. — P. 269–284. — ISSN 0163-5964 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0163-5964). — doi:10.1145/2654822.2541967 (https://dx.doi.org/10.1145%2F2654822.2541967). Архивировано (https://web.archive.org/web/20211110133118/https://dl.acm.org/doi/10.1145/2654822.2541967) 10 ноября 2021 года.
5. *Yunji Chen, Tao Luo, Shaoli Liu, Shijin Zhang, Liqiang He.* DaDianNao: A Machine-Learning Supercomputer (https://ieeexplore.ieee.org/document/7011421) // 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture. — 2014-12. — С. 609–622. — doi:10.1109/MICRO.2014.58 (https://dx.doi.org/10.1109%2FMICRO.2014.58). Архивировано (https://web.archive.org/web/20211201161659/https://ieeexplore.ieee.org/document/7011421/) 1 декабря 2021 года.
6. *Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Li.* ShiDianNao: shifting vision processing closer to the sensor (https://dl.acm.org/doi/10.1145/2872887.2750389) (англ.) // ACM SIGARCH Computer Architecture News. — 2016-01-04. — Vol. 43, iss. 3S. — P. 92–104. — ISSN 0163-5964 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0163-5964). — doi:10.1145/2872887.2750389 (https://dx.doi.org/10.1145%2F2872887.2750389). Архивировано (https://web.archive.org/web/20211110133120/https://dl.acm.org/doi/10.1145/2872887.2750389) 10 ноября 2021 года.
7. *Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou.* PuDianNao: A Polyvalent Machine Learning Accelerator (https://dl.acm.org/doi/10.1145/2786763.2694358) (англ.) // ACM SIGARCH Computer Architecture News. — 2015-05-29. — Vol. 43, iss. 1. — P. 369–381. — ISSN 0163-5964 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0163-5964). — doi:10.1145/2786763.2694358 (https://dx.doi.org/10.1145%2F2786763.2694358). Архивировано (https://web.archive.org/web/20211110133120/https://dl.acm.org/doi/10.1145/2786763.2694358) 10 ноября 2021 года.
8. *Yunji Chen, Tianshi Chen, Zhiwei Xu, Ninghui Sun, Olivier Temam.* DianNao family: energy-efficient hardware accelerators for machine learning (https://dl.acm.org/doi/10.1145/2996864) (англ.) // Communications of the ACM. — 2016-10-28. — Vol. 59, iss. 11. — P. 105–112. — ISSN 1557-7317 0001-0782, 1557-7317 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0001-0782,). — doi:10.1145/2996864 (https://dx.doi.org/10.1145%2F2996864). Архивировано (https://web.archive.org/web/20211110133118/https://dl.acm.org/doi/10.1145/2996864) 10 ноября 2021 года.
9. *Yu-Hsin Chen, Joel Emer, Vivienne Sze.* Using Dataflow to Optimize Energy Efficiency of Deep Neural Network Accelerators (http://ieeexplore.ieee.org/document/7948671/) // IEEE Micro. — 2017. — Т. 37, вып. 3. — С. 12–21. — ISSN 0272-1732 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0272-1732). — doi:10.1109/MM.2017.54 (https://dx.doi.org/10.1109%2FMM.2017.54). Архивировано (https://web.archive.org/web/20181122201940/https://ieeexplore.ieee.org/document/7948671/) 22 ноября 2018 года.
10. *Han, Song.* EIE: Efficient Inference Engine on Compressed Deep Neural Network / Song Han, Xingyu Liu, Huizi Mao … [и др.]. — 2016-02-03.
11. *Brandon Reagen, Paul Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee.* Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators (https://ieeexplore.ieee.org/document/7551399/) // 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA). — Seoul: IEEE, 2016-06. — С. 267–278. — ISBN 978-1-4673-8947-1. — doi:10.1109/ISCA.2016.32 (https://dx.doi.org/10.1109%2FISCA.2016.32). Архивировано (https://web.archive.org/web/20211201161613/https://ieeexplore.ieee.org/document/7551399/) 1 декабря 2021 года.

12. *Patrick Judd, Jorge Albericio, Andreas Moshovos.* Stripes: Bit-Serial Deep Neural Network Computing (http://ieeexplore.ieee.org/document/7529197/) // IEEE Computer Architecture Letters. — 2017-01-01. — Т. 16, вып. 1. — С. 80–83. — ISSN 1556-6056 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:1556-6056). — doi:10.1109/LCA.2016.2597140 (https://dx.doi.org/10.1109%2FLCA.2016.2597140). Архивировано (https://web.archive.org/web/20210308075215/https://ieeexplore.ieee.org/document/7529197/) 8 марта 2021 года.
13. *Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal.* In-Datacenter Performance Analysis of a Tensor Processing Unit (https://dl.acm.org/doi/10.1145/3079856.3080246) (англ.) // Proceedings of the 44th Annual International Symposium on Computer Architecture. — Toronto ON Canada: ACM, 2017-06-24. — P. 1–12. — ISBN 978-1-4503-4892-8. — doi:10.1145/3079856.3080246 (https://dx.doi.org/10.1145%2F3079856.3080246). Архивировано (https://web.archive.org/web/20211110134715/https://dl.acm.org/doi/10.1145/3079856.3080246) 10 ноября 2021 года.
14. 思元100智能加速卡 - 寒武纪 (https://www.cambricon.com/index.php?m=content&c=index&a=lists&catid=21). *www.cambricon.com*. Дата обращения: 10 ноября 2021. Архивировано (https://web.archive.org/web/20211110134723/https://www.cambricon.com/index.php?m=content&c=index&a=lists&catid=21) 10 ноября 2021 года.

---

Источник — https://ru.wikipedia.org/w/index.php?title=Процессор_глубокого_обучения&oldid=125780671

---