WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning with AI Chips

Chetan Dayma, Dr. Kamal Raj R

Department of Computer Science and IT, Jain Deemend to be University
Chetandayma75@gmail.com, drkamalrajr@jainuniversity.ac.in
DOI: https://doi.org/10.55248/gengpi.5.0324.0656

A B S T R A C T

this study has been undertaken to investigate the determinants of latest Lenovo laptops Which Are using of AI chips with are train Machine Learning to learn the usage style of laptop according to user and adjust the performance for user according to the use. In the present Study we Figure out how we can train more efficiently, this chip to get more benefits of AI in upcoming Time of computing.

Keywords: Machine Learning, GPU, CPU, TDP, AI-engine, Tuning, hardware acceleration.

1. Introduction

We are surrounded by the AI chips in 21th century, now the companies are doing their best to get more output with this AI chips and pushing potential of this AI chips, Best example been taken as, one of the market leader manufacture of laptop Lenovo introduced the AI chips technology, which can be Adjust itself by knowing the usage of user, and try to get the best optimize performance to the user, given AI chips of Lenovo studies the user activity, whether system using more graphical GPU or more CPU Consumption, if system using more CPU power then it shifts more power to CPU, optimize system to use CPU more efficiently vice-versa with GPU graphical power, the advantage of given chip is more power efficiency and performance to the end user who is using it.

2. Literature review

According to current market situation and competitors for the particular brand is tougher, AI chips technology in laptops is new concept and with boom of AI many more companies acquire this technology, and the end user gets the good benefit with this technology. In present study we looking for, how we can train this chips to get more benefits and ease for usage. Currently with the latest event of Lenovo which is organize in Bengaluru, Leela Palace, where Lenovo is more looking about the complex demanding games and video editing software, so that Given AI chip train itself according to usage of user. With particular software which is given by Lenovo we can modify it with our usage type. With a short survey with users we easily find out what more we add with this chips so that, it work more efficiently and helping in different scenarios.

With the reference study of another papers based on hardware acceleration, we find a gap that, which is not using of AI, But with Help of this technology the gap has to be filled. Let's have a look at some previous Studies.

Hardware advancements in computers are largely responsible for many of the most recent improvements in AI [1], [2]. Specifically, advancements in contemporary computing have enabled many neural networks and other computationally intensive machine learning algorithms. Although machine-learning methods, such neural networks, have a long theoretical history [3], recent computing advancements have made it possible to apply

These algorithms in practice by supplying the processing capacity required to train and interpret large amounts of data. Even though the last ten years have seen many innovations in the computing space, more embedded and mobile applications that demand small, lightweight, and power (SWaP) systems will require capabilities beyond what the conventional architectures of central processing units (CPUs) and GPUs Can offer.

Low power Chips

Research chips from universities and businesses have been the majority of chips in the very low power regime. Still, a few suppliers have made announcements or are making items available in this market.

• The Vivienne Sze group at MIT CSAIL developed the research chip known as the MIT Eyeris chip, Eyerissi [4], [5], and [6]. Creating the most energy-efficient inference chip feasible was their aim. The AlexNet result was obtained without specifying the batch size.

• From the IBM Almaden research lab, the TrueNorth [7], [8] is a digital neuromorphic research chip. It was created in the Synapse program with support from DARPA tosi show the effectiveness of digitally spiked neural network chips, sometimes known as neuromorphic chips.

It should be noted that the graph shows points for both the chip, which consumes only 275 mW, and the system, which consumes 44 W of power.

• An embedded video processor with a neural engine for object detection and video processing is the Intel MovidiusX processor, MovidiusXi [9]. Google unveiled the HPUEdgei [10] TPU Edge processor for embedded inference applications at the beginning of 2019. TensorFlow Lite, which encodes the neural network model with low precision parameters for inference, is used by the TPU Edge.

GPU based Accelerators

On the chart, two AMD/ATI cards and four NVIDIA cards are presented in that order:

The Pascal architecture P100hP100i [11], [12], the Volta architecture V100, and the Maxwell architecture K80 hK80i .

TU106 Turing, MI6 hMI6i , hV100i and MI60 hMI60i. While the TU106 Turing GPU is targeted towards the gaming and graphics sector to incorporate inference processing within the graphics processing, the K80, P100, V100, MI6, and MI60 GPUs are pure compute cards meant for both inference and training.

3. AI Chip Working

Smooth gaming performance is paramount for gamers — one of the biggest pet peeves is a drop in frame rates, pulling them out of the immersion. All of them are powered by the world's first dedicated Lenovo Artificial Intelligence (LA) AI chip and Lenovo AI Engine+

Software to ensure an optimal gaming experience. The Lenovo AI Engine+ deploys a software machine learning scenario-detection algorithm that monitors in-game frame rates and fine-tunes system performance by allocating wattages between the CPU and GPU. This results in a 15 percent higher thermal design power (TDP), which optimizes gameplay seamlessly without the need for manual adjustment to the settings. Quieter fan noise, faster performance, and most of all, increased frames per seconds (FPS) can be expected. What's even better is that users will have full control to customize and adjust their new AI- powered machine through

Figure:3.1: lenovo.com

4. Improvement and Methodology

4.1 Improvement

Given AI chips are more optimize for Gaming and editing, with this kind of efficiency we train particular chip for all kind of process which we do in laptop for ex: movie watching to edit a document, this chip will able to optimize itself what kind of process is been executing on particular laptop, the best thing we can do we train it by ourselves and get the performance what we want, and we get the full potential of hardware. As current technology of the chip it works with CPU and GPU instructions.

GPUs: Graphics Processing Units

Deep learning frequently uses matrix operations, which GPUs excel at handling because of their high level of parallelization. We used NVIDIA GeForce RTX 3090 GPUs for our experiments due to their high computational power and wide support in the machine learning community.

TPUs, or tensor processing units:

TPUs, developed by Google, are designed specifically for deep learning tasks. We utilized Google Cloud TPUs to benchmark their performance in comparison to GPUs, particularly for large-scale neural network training.

FPGAs (Field-Programmable Gate Arrays):

Hardware circuits called FPGAs can be reconfigured to perform particular functions. To investigate the possible advantages of hardware modification in machine learning applications, we used Xilinx Alveo U280 FPGAs.

ASICs: Application-Specific Integrated Circuits

ASICs are semiconductors that have been specially created and optimized for a certain use. Although we didn't use ASICs directly, we looked at how they could speed up inference in the setting of edge AI devices.

Figure:4.1: lenovo.com

4.2 Methodology

ML technique and providing a particular data to chip with a short survey so, that how many Lenovo users using the laptop by which type and which software they are using mostly, with bigger dataset, So that AI chip trained according to user, with the current scenario the particular chip is coded, so that it can be learn by itself and give the full potential performance to the user. Like Lenovo provided mux switch to shift CPU to GPU according to use. If possible Lenovo might be provides the developer option to this AI tuned Chip so Enthusiast user can do more with it and train it by himself and get the best performance .

The Lenovo Vantage software, including overclocking, AI-performance tuning, and system monitoring through a real- time performance dashboard.

5. Models Come With AI CHIPS

Also boasting the new LA AI chip, the new 16-inch Lenovo Legion Pro 5 and 5i laptops are touted by the company as having “esports styling” that “hints at the gaming powerhouse at its heart”.

They’re available with either a 13th Gen Intel Core or AMD Ryzen 7000 Series processor and up to an Nvidia GeForce RTX 4070 Laptop GPU. Internals are kept cool thanks to Lenovo Legion’s Cold Front 5.0, which uses massive exhaust and intake systems, a turbo-charged dual fan system, phase-change thermal compound, and advanced hybrid heat pipes to move more air out. The Lenovo LA1 chip settings can be messed with through Lenovo Vantage, which will give you full control over features like fans and overclocking controls.

The new chip will come preinstalled on the new Lenovo Legion Pro 7 and 7i laptops and the Lenovo Legion Pro 5 and 5i laptops. Here’s what the company had to add:

Lenovo AI Engine+, powered by the Lenovo LA AI chip, deploys a software machine learning algorithm to optimally tune system performance. The chip uses software machine learning, deployed through Lenovo Vantage, to help monitor in-game FPS and dynamically adjust for the highest performance output. Offering up to 15% higher TDP, this chip and machine learning software combo allows Legion Pro Series laptops to deliver higher performance compared to previous generations. Lenovo Legion Pro 7i and Lenovo Legion Pro 5 and 5i laptops also come with Tobii Horizon, providing gearless head tracking that gives players an extra level of immersion when playing their favourite games, as well as Tobii Aware.

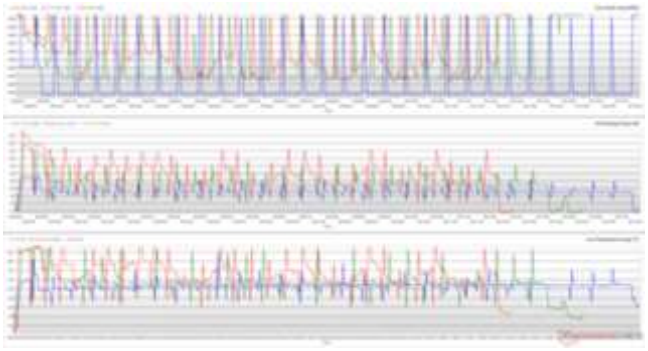

Figure:5.1 Reddit.com

5.1 L2 chip:

As of January 26, 2022, a new firm named Dingdao Zhixin has been officially registered as a commercial entity in Shanghai, China. Its new listings indicate that it would deal with integrated circuit design and sales as well as PC hardware, software, and some retail equipment.

It would be intriguing enough to see a new company in this industry emerge, and it becomes even more intriguing in light of rumors that Lenovo China actually owns the entire company. Consequently, this subsidiary may be an early indication that the OEM hopes to match other businesses like Apple with its own line of in-house processors, the M-series.

Lenovo has begun to design parts of its own silicon, as evidenced by the Yoga Pro 14s that was just released.

5.2 Benchmarks:

Figure:5.2 Reddit.com

1440p I play most games at 60 frames per second, maxed out, and I'm loving it. I mostly play on an external 1440p display. I often run between 80 and 100 frames per second, and to aid with temperature, I've added an undervolt and a small overclock. Higher performance and better temperatures than stock (GPU stays at 78°) (timespy score of 10800).

Even the loudest fans are not too horrible, but you can modify them to your liking. You may underclock and lower the fan speed to get approximately 5 to 10% less performance, and the fan will be extremely quiet.

With the given graph we can see how the Ai chips are giving more performance compare to another laptops, which are working with only the cpu and gpu.

6. Future Work

With a game Changing technology this AI chips were more being intelligent with the time, for future user won't have to tune laptop for particular application or software, the laptop tune automatically itself by just observing the user activity how he/she doing what kind of work on the laptop, which gives more performance to laptop and work will easy for its user.

Integration and Customization of AI Chips:

Examine the potential integration of AI chips into a range of products and platforms, including IoT devices and driverless cars. AI chips that are specifically designed for a given application may perform significantly better.

Sustainability and effectiveness:

Examine strategies to make AI processors more eco-friendly and energy-efficient. Create AI processors that use less energy, and look into environmentally friendly alternatives to materials and designs.

Co-design of hardware and software:

Encourage research into methods for hardware-software co- design to enhance machine learning models for particular AI

chip architectures. This might result in more specialized and effective AI solutions.

Explicit AI on the Cutting Edge

Look into the use of explainable AI methods in edge AI devices. It is essential to create AI processors that can deliver comprehensible results in real-time applications for vital industries like healthcare and autonomous vehicles.

Accelerating quantum machine learning:

Investigate the pairing of AI chips with quantum computing technology. Understanding how quantum computers interact with AI chips is a fascinating direction to pursue since they have the ability to exponentially accelerate some machine learning operations.

Improved Security and Privacy:

Identify and create AI chips with built-in security and privacy protection. Address data privacy and security issues in machine learning models from the outset rather than after the fact.

7. Conclusion

So, we had done good research on the particular chip and we get to know about the chip about its potential, how we can make it better for upcoming models, given paper gives you sufficient Knowledge about AI chips using by companies to extract more performance from the machines. With some xtra add on features or customizable developer mode gives more freedom to user to train according to himself, so it would be more controllable and useful.

In summary we benchmarked some laptops with ai chips and some laptops with non ai chips here Lenovo ai chips are Cleary win with low power consumption and better performance.

Utilizing the Lenovo LA AI processor, the Lenovo AI Engine+ is a software machine learning technique that enhances system performance. To improve gaming, the chip makes use of sensors and machine learning algorithms. Additionally, it tracks frame-rate while playing games and makes dynamic adjustments to maximize performance.

During gaming, the Lenovo AI Engine+ also dynamically adjusts power and controls thermal performance. Superior Rapid Charge, which provides quicker charging and longer battery life, is supported by all Lenovo LOQ laptops.

In current IT industry Role of AI in every single corner is there, this AI chips are the best example for it, which is used by you on daily basis.

8. Reference

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Neural Information Processing Systems, vol. 25, 2012.

[2] N. P. Jouppi, C. Young, N. Patil, and D. Patterson, “A Domain-Specific Architecture for Deep Neural Networks,” Communications of the ACM, vol. 61, no. 9, pp. 50–59, aug 2018. [Online]. Available: http://doi.acm.org/10.1145/3154484

[3] M. L. Minsky, Computation: Finite and Infinite Machines. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1967.

[4] V. Sze, Y. Chen, T. Yang, and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” Proceedings of the IEEE, vol. 105, no. 12, pp. 2295–2329, dec 2017.

[5] Y. Chen, J. Emer, and V. Sze, “Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks,” IEEE Micro, p. 1, 2018

[6] Y. Chen, T. Krishna, J. S. Emer, and V. Sze, “Eyeriss: An EnergyEfficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” IEEE Journal of Solid-State Circuits, vol. 52, no. 1, pp. 127–138, jan 2017.

[7] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, “TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 10, pp. 1537–1557, oct 2015.

[8] M. Feldman, “IBM Finds Killer App for TrueNorth Neuromorphic Chip,” sep 2016. [Online]. Available: https://www.top500.org/news/ ibm-finds-killer-app-for-truenorth-neuromorphic-chip

[9] J. Hruska, “New Movidius Myriad X VPU Packs a Custom Neural Compute Engine,” aug 2017.

[10] “Edge TPU,” 2019. [Online]. Available https://cloud.google.com/edge-tpu/

[11] “NVIDIA Tesla P100.” [Online]. Available: https://www.nvidia.com/ en-us/data-center/tesla-p100/

[12] R. Smith, “NVIDIA Announces Tesla P100 Accelerator - Pascal GP100 Power for HPC,” apr 2016. [Online]. Available: https://www.anandtech.com/show/10222/ nvidia-announces-tesla-p100-accelerator-pascal-power-fo