Introduction to Hardware Platform of Machine Learning

Ziliang Zhao

Shenzhen University

*Abstract*—Machine learning and Artificial Intelligence is quite a hot theme mentioned in sight of modern people. Begin with 2016, many high tech firms like Google, Amazon, Facebook, IBM, Baidu claim they are transforming to “AI Company”. Recently, more and more hardware vendors join the competition, for instance, Qualcomm. As we know DNN (Deep Neural Network) training always be performed in the cloud, which requires high performance GPU to accomplish. And in this case, scientists and engineers begin to find more efficient hardware platform to support their model.

I. INTRODUCTION

We would better have a little knowledge basic on machine learning before we discussing hardware platforms and requirements of it. What is exactly machine learning or artificial intelligence? Machine learning is a form of artificial intelligence that can execute tasks without too much hand-crafted programming or featuring.[1] And it will learn from previous example of you task in the process of training. Then your task in running on new data through inference. Generally, your task should go through feature extraction and classification to generate the result you want. Most AI services today reside in the cloud, where massive amounts of computational power can be harnessed to train the neural network and to handle the large volume of queries (called inference or scoring) in a cloud such as Microsoft, Facebook, Apple, or Google. Beyond the cloud, intelligent machines interact with people and objects through a device at the edge of the network. [2] And some applications need more resource than other, for instance, video compression always consuming mostly computational resource from GPU. That is why we gradually realize the boundary of our AI cloud server and urgently need hardware solutions.

II. APPLICATIONS

With the trend of machine learning and artificial intelligence, hardware applications begin to apply this new technology and many of them have been developed and produced. Personal VR Gaming could be the first one hit our head when we are discussing intelligent hardware. And in other industries, artificial intelligence is widely applied, such as autonomous cars and trucks for transportation, delivery drones and warehouse robots for commercial uses and surgical robots for healthcare. Such applications above are all tangible, there are also like Ad targeting and recommending system, traffic analytics and clinical recommendations which are based on complicated artificial intelligence algorithms.

III.HARDWARE

Basically we perform DNN on GPU like we mentioned above because we need precise and fast floating point computing ability. And cloud server usually do all the computational work, nevertheless, scientists assign batch of work to other components like CPU. For example, Facebook finish consumer image classification only by processor of smartphones- which always be Qualcomm- or desktop CPU, which provide sufficient processing capability yet. [1] Also edge devices sometimes share computing from cloud server to relieve resource pressure- witness Amazon Echo. The hardware requirements in each application are determined by the nature of the data that feeds into the system, the response time (latency) required for action, and the amount of synthesis or additional processing required. [3] So when you are processing massive computational work like compressing 64 frame 4K videos through DNN you still need a dictated accelerator and the additional work should be handled by it. Mostly the DNN needs GPU to spend up, however we still have FPGA, DSP or ASIC to provide hardware support. What is more, CPU is recommended for accelerating because it is friendly to programming which make itself able to handle more tasks deployed in program lifecycle. FPGA vendor Xilinx and IBM offer specific SoCs with integrated CPU cores, programmable components and sensors to support developer community and production, in addition, Qualcomm also publish its platform for developing flexible interfaces on Snapdragon 835. As industry forerunner- NVIDIA and AMD- also address platform-level tech support and production aimed to GPU processing. For example, NVIDIA Tegra X1 targets SoC and Intel Arria 10 SoC FPGA aims hybrid SoC. GPU performs excellently in the cloud, mostly the reason in accurate floating-point computing on massive batches of data and GPU is widely applied by cloud service vendors including IBM Softlayer and Amazon AWS.

Drawbacks of GPU in the cloud are probably high electricity consumption and so on. As we mentioned before, FPGA is another way burden the computing, however, FPGA is kind of rigid which means it is quite unfriendly to programming compared to CPU. The last way we discuss to computing aside the cloud is to compute locally or as we named it- hybrid environment. Hybrid computing with smartphone processor is relatively low cost and easy to program, unfortunately, their performance and latency is limited, which means longer training period.

IV.CONCLUSIONS

Applications demand varying computational capacity depending on where they reside: in the cloud, edge devices, or in a hybrid environment. The computational requirements vary with the type of data being analyzed and with the factors dictated by the AI environment where it operates. And when we decide to develop applications with machine learning we should not assign all the computing work to GPU, properly-clocked CPU [4], binding with hybrid machine learning environment devices could also be your target inferences.

REFERENCES

* [1] Moor Insights & Strategy, “A Machine Learning Application Landscape,” xilinx.com, March 2017.
* [2] Vivienne Sze, Yu-Hsin Chen, “Hardware for Machine Learning: Challenges and Opportunities,” MIT Invited Paper, October 2017.
* [3] Rob Farber, “Technology Requirements for Deep and Machine Learning,” nextplatform.com, July 2017.
* [4] Tim Dettmers, “A Full Hardware Guide to Deep Learning”, timdettmers.com, March, 2015.