AI Accelerators

Flash Memory, IoT and AI
- Bringing it All Together

EMBD 302B-1: PANEL DISCUSSION

August 8, 2019

Kiran Gunnam, Western Digital

Forward-Looking Statements

Safe Harbor | Disclaimers

This presentation contains certain forward-looking statements that involve risks and uncertainties, including, but not limited to, statements regarding machine learning technology, growth opportunities, market adoption, demand for digital storage and market trends. Forward-looking statements should not be read as a guarantee of future performance or results, and will not necessarily be accurate indications of the times at, or by, which such performance or results will be achieved, if at all. Forward-looking statements are subject to risks and uncertainties that could cause actual performance or results to differ materially from those expressed in or suggested by the forward-looking statements.

Key risks and uncertainties include volatility in global economic conditions, actions by competitors, business conditions, growth in our markets, product development, pricing trends and fluctuations in average selling prices, and other risks and uncertainties listed in our filings with the Securities and Exchange Commission (the "SEC") and available on the SEC's website at www.sec.gov, including our most recently filed periodic report, to which your attention is directed. We do not undertake any obligation to publicly update or revise any forward-looking statement, whether as a result of new information, future developments or otherwise, except as required by law.

Challenges, 1/2

- ***AI workloads are compute, storage and networking intensive and incur huge power dissipation.***
  - For a typical image classification inference workload such as ResNet50 using 299x299 image, need ~4 Billion INT8 MAC (multiply and accumulate) operations, ~25 MB of weight storage, ~10 MB for activations for each image. Typically multiple image sensors with higher resolutions in a given application and 25 to 50 frames per second.
  - Each new image that needs to be used in training, ~12 Billion FP32 MAC operations, weights, gradients, updates: ~300 MB of storage, activations, gradients: 80 MB. Training uses 1.2 million images and 100 epochs. Total requirement ~1.4x10^18 (exa) FP32 operations. Needs multiple GPUs, storage and networking cluster to be able to do training on daily/weekly basis.
  - Current computing architectures such as CPUs and GPUs are expensive and contain lot of processing resources that are not either used or overkill for AI workloads and draw huge power. In addition, they require expensive memories such as HBM and expensive networking to scale up. So it is currently difficult to deploy AI workloads at edge.

- ***IDC reports estimates 41.6 billion connected IoT devices, or "things," generating 79.4 zettabytes (ZB) of data in 2025. Compute, storage and networking can not be scaled up with current cloud centric architectures and CPUs/GPUs. Need a new approach.***

Sources : https://www.idc.com/getdoc.jsp?containerId=US45066919 and HotChips-2018 Proceedings.

Challenges, 2/2, Convolutional Neural Networks (CNNs)

Flash Memory Summit

- ❑ *Image classification accuracy is 62.5% (top-1 accuracy) for AlexNet on ImageNet benchmark.*

- ❑ *New models such as Inception-v4 and PNAS-Net improves the Top-1 accuracy to 82.5%.*

- ❖ *Increased depth of the models (i.e. deeper Nets)*
(increased number of layers. AlexNet-8 layers, ResNet-152-152 layers)

- ❖ *New constructs such as inception, residual connections, cells, blocks etc.*

- ❑ *However newer models have even larger computation requirements*
- ❖ *More computations, up to 32 GoPs per image frame (~ 20x for VGG16 vs AlexNet)*

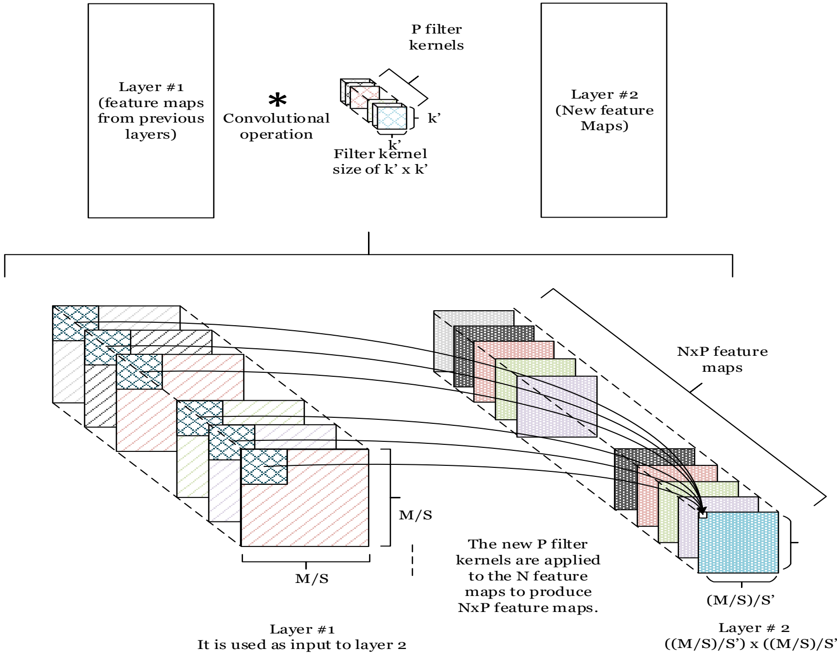
Sources: Venieris, Stylianos & Kouris, Alexandros & Bouganis, Christos. (2018). Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions. ACM Computing Surveys. 51. 10.1145/3186332.
And https://arxiv.org/pdf/1712.00559.pdf

AI Accelerators, 1/2

Flexibility	AI Accelerators	Efficiency
↑	CPUs (x86, RISC-V etc.)	↓
	CPUs with vector extensions (x86-Advanced Vector Extensions, RISC-V Vector Extension)	
	GPUs	
	Soft Cores on FPGAs	
	Programmable Systolic or tensor/matrix computation processors	
	Programmable ASICs targeted for specific ML application	

AI accelerators typically have hardware acceleration for Artificial Intelligence applications.

AI Accelerators, 2/2



AI Accelerators speed up Multiply-Accumulate dominated tensor computations with customized compute, interconnect and memory/storage architectures. Also they provide higher energy efficiency reducing the power dissipation by minimizing the traffic to DRAM/HBM with emphasis on data locality.

Image Source : K. Gunnam, IEEE Tutorial on Machine Learning, 2017

Use Case, Edge AI for Virtual Assistant

Devices with advanced edge AI computing can make Virtual Assistants more smarter.

Enables voice biometrics for user authentication.

- Prevents unauthorized users

  children talking to Alexa to add items to shopping

  burglars shouting to Alexa to disarm the home alarm
- Works offline
- Users privacy is protected.
- Zero latency vs 0.82 msec for every 100 miles data travels.
- Low power

Enables Robust face biometrics for user authentication.

- Prevents unauthorized users using 3D masks

No need to send the voice recordings (for voice search) or images at home to cloud

- Vs Cloud processing➔ on-going privacy concerns and inadvertent leaks or transmission to unauthorized third parties and labelers.
- faster image analysis at edge with privacy for social networking and search apps.

Advantages: privacy latency, reliability, low cost