http://synergy.ece.gatech.edu

MAESTRO: A Data-Centric Approach for Hardware and Mapping Explorations for Deep Learning Accelerators

Tutorial @ MICRO 2020
October 17 2020

http://maestro.ece.gatech.edu
https://tinyurl.com/maestro-tutorial-micro2020

Presenters

Tushar Krishna
Assistant Professor,
School of ECE,
Georgia Tech

tushar@ece.gatech.edu

Michael Pellauer
Sr. Research Scientist,
NVIDIA

mpellauer@nvidia.com

Geonhwa Jeong
PhD Student
School of CS
Georgia Tech

geonhwa.jeong@gatech.edu

Prasanth Chatarasi
Recent PhD
Georgia Tech
(joining IBM)

cprasanth@gatech.edu

Felix Kao
PhD Student
School of ECE
Georgia Tech

felix@gatech.edu

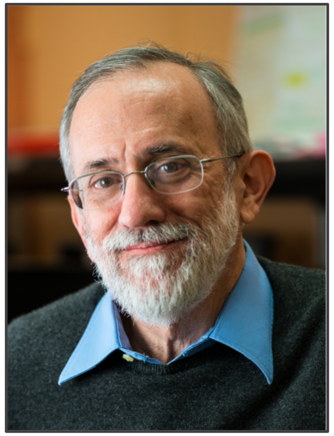
Acknowledgments

Hyoukjun Kwon
Recent PhD Grad
Georgia Tech
(Now at Facebook Reality Labs)

MAESTRO Core Developer

Angshuman Parashar
Sr. Research Scientist,
NVIDIA

Vivek Sarkar
Professor
Chair, School of CS,
Georgia Tech



Joel Emer
Sr. Distinguished
Research Scientist
NVIDIA

Professor
EECS, MIT

Schedule (EDT)

Time	Agenda	Presenter
10:00 - 10:10	**Welcome + Intro to DNNs**	Tushar
10:10 - 10:45	**DNN accelerators and dataflows**	Michael
	MAESTRO Cost Model	
10:45 - 11:15	Data-centric Mapping Directives	Hyoukjun^
11:15 - 11:30	Analytical Cost Model	Hyoukjun^
11:30 - 12:00	Compiling and Running MAESTRO	Geonhwa
12:00 - 12:10	**Break**	
	Automated Design-space Exploration using MAESTRO	
12:10 - 12:25	Marvel: Mapping Space Exploration via Heuristics	Prasanth
12:25 - 12:40	GAMMA: Mapping Space Exploration via Optimization	Felix
12:40 - 12:55	ConfuciuX: Hardware Design-space Exploration via RL and Optimization	Felix
12:55 - 13:00	**Wrap Up**	Tushar

^Pre-recorded Video

Brief Q/A at the end of each talk.

Please use Zoom's Q/A feature. You can upvote other questions. We will answer live or respond in the Q/A box.

Slides + Videos will be available on the MAESTRO tutorial website

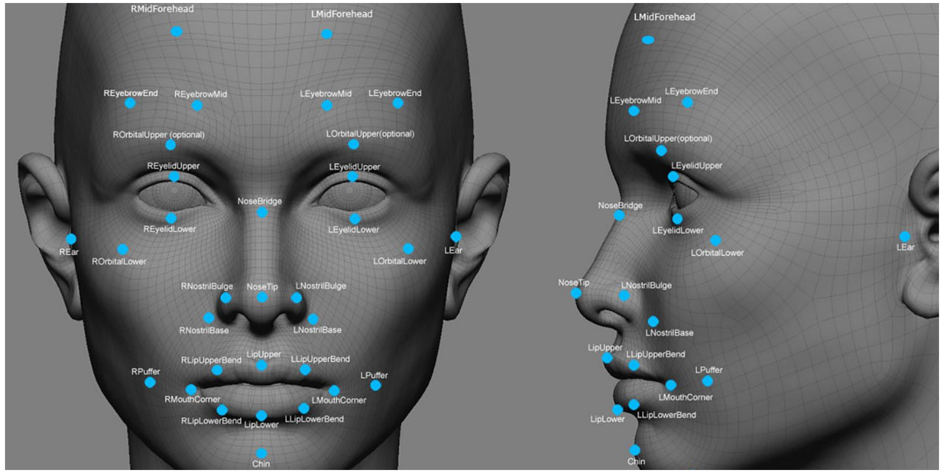
https://tinyurl.com/maestro-tutorial-micro2020

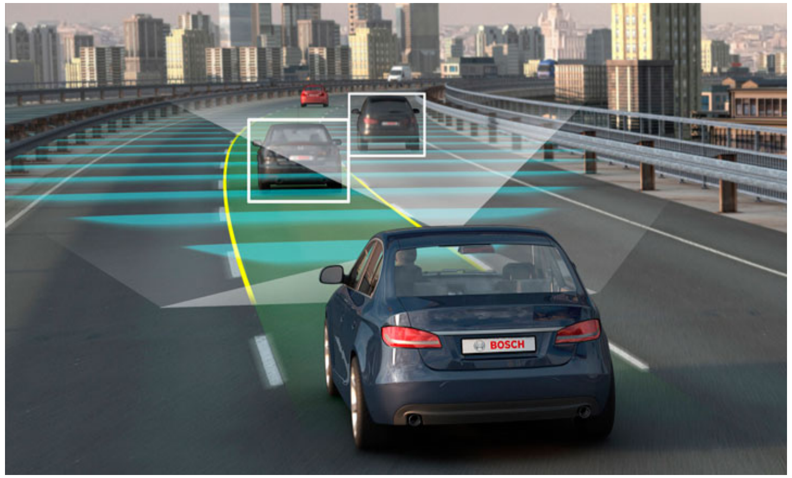
Outline

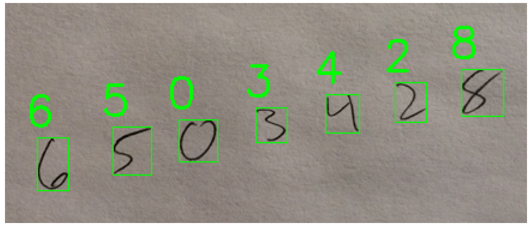
- **What this tutorial is about**
  - Role within the Deep Learning landscape
- **Resources**
  - MAESTRO
  - Synthesis Lectures
- **Relevant Background**
  - Deep Neural Networks

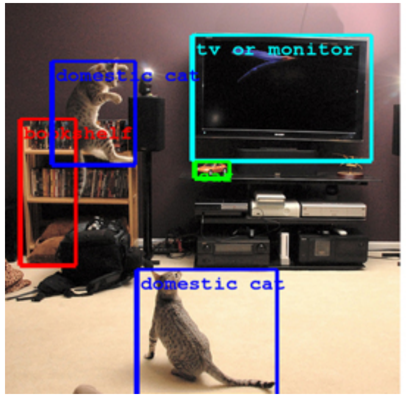
What this tutorial is about

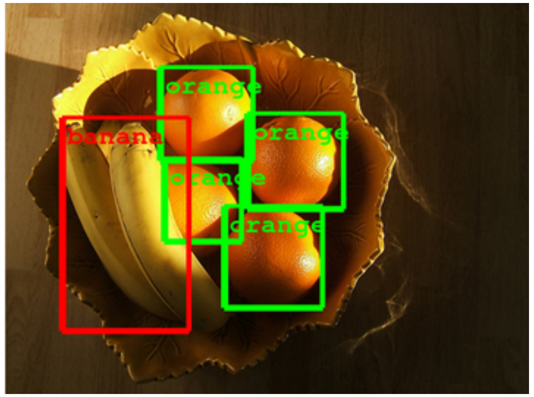
Deep Learning Applications

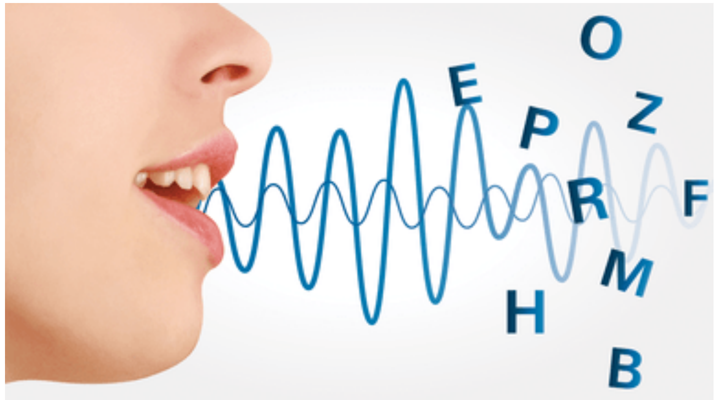




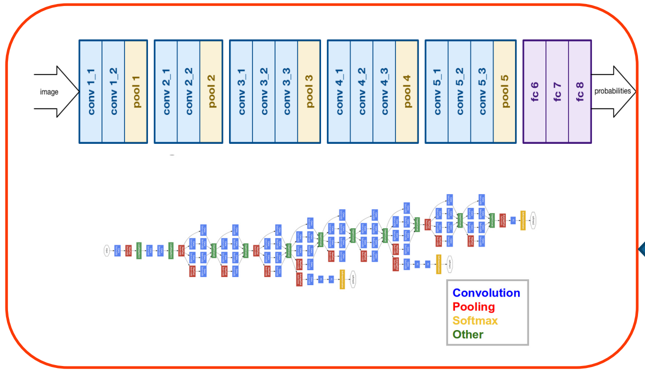








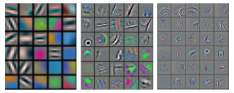
Deep Learning Landscape



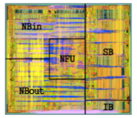
Model Creation

IMAGENET

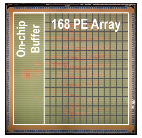
Training



Inference



ShiDIanNao



Eyeriss

Apple Neural Engine

NVDLA

ARM Trillum

CambriconX

TensorFlow

Caffe

cuDNN

intel Software

MLSL

NVIDIA

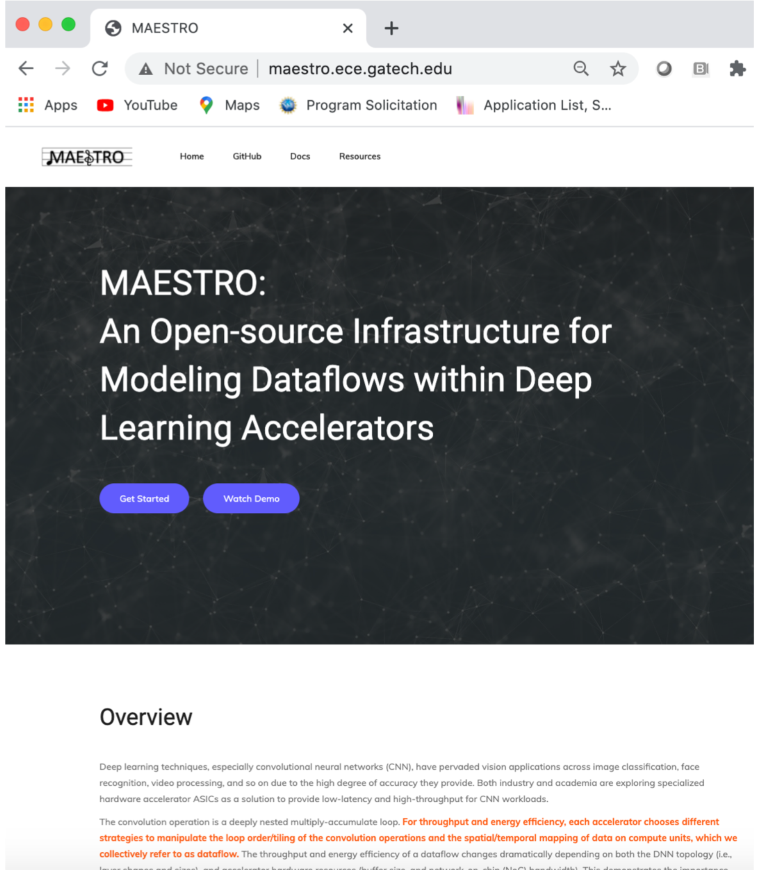
TensorRT

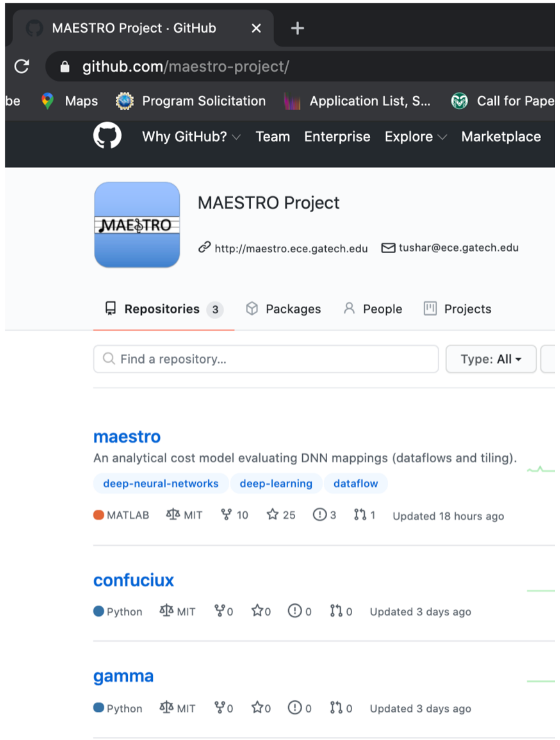
This Tutorial

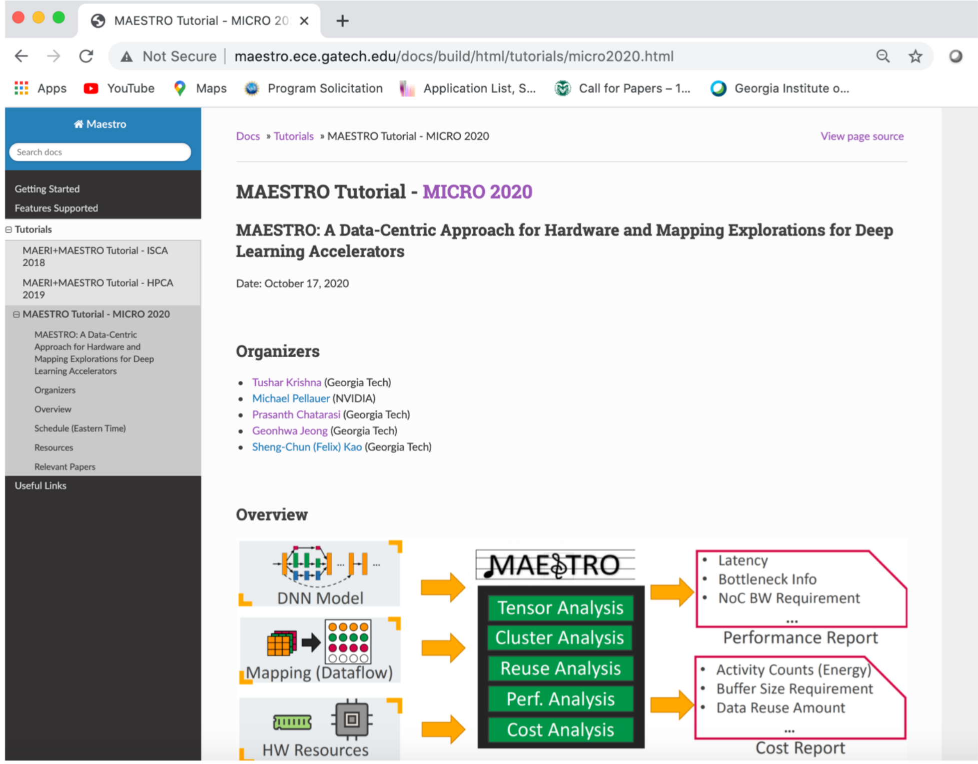
Design Tools

Resources

MAESTRO Webpage, Github and Tutorial







http://maestro.ece.gatech.edu

→ *Links to papers, wiki, …*

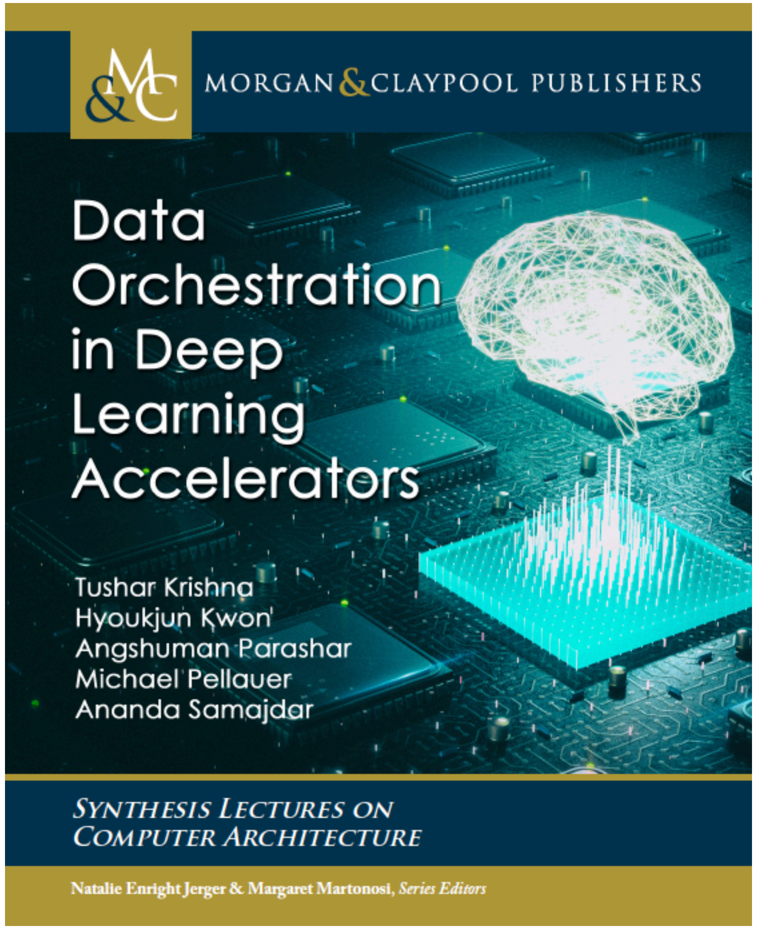
http://github.com/maestro-project

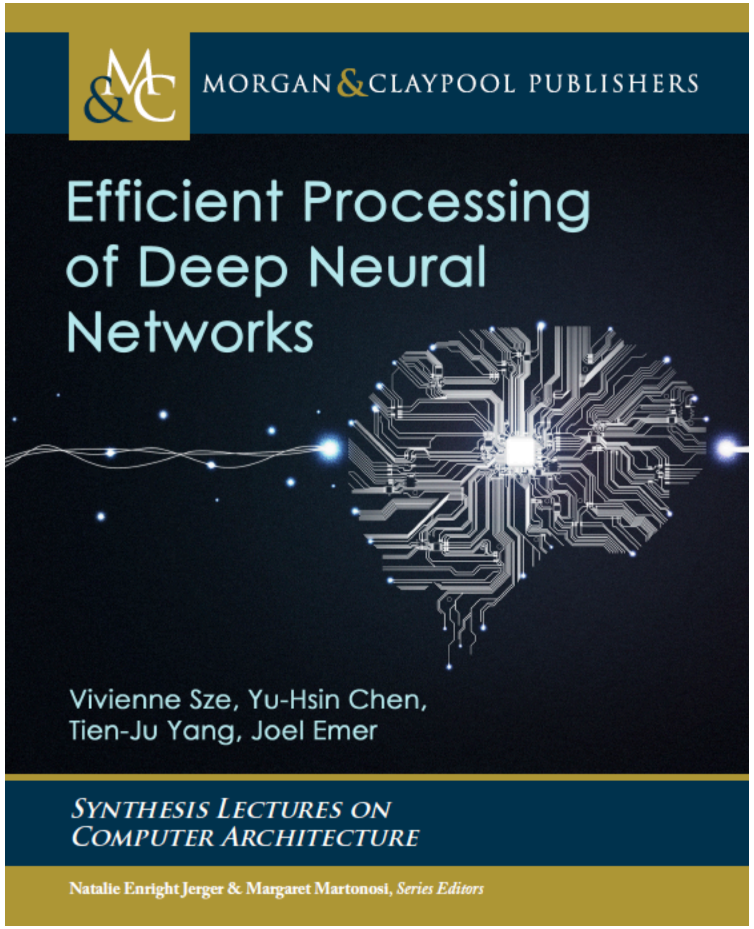
→ *MAESTRO cost model and associated tools: Marvel, GAMMA, ConfuciuX, ..*

https://tinyurl.com/maestro-tutorial-micro2020

→ *Slides + Videos + Docker Image*

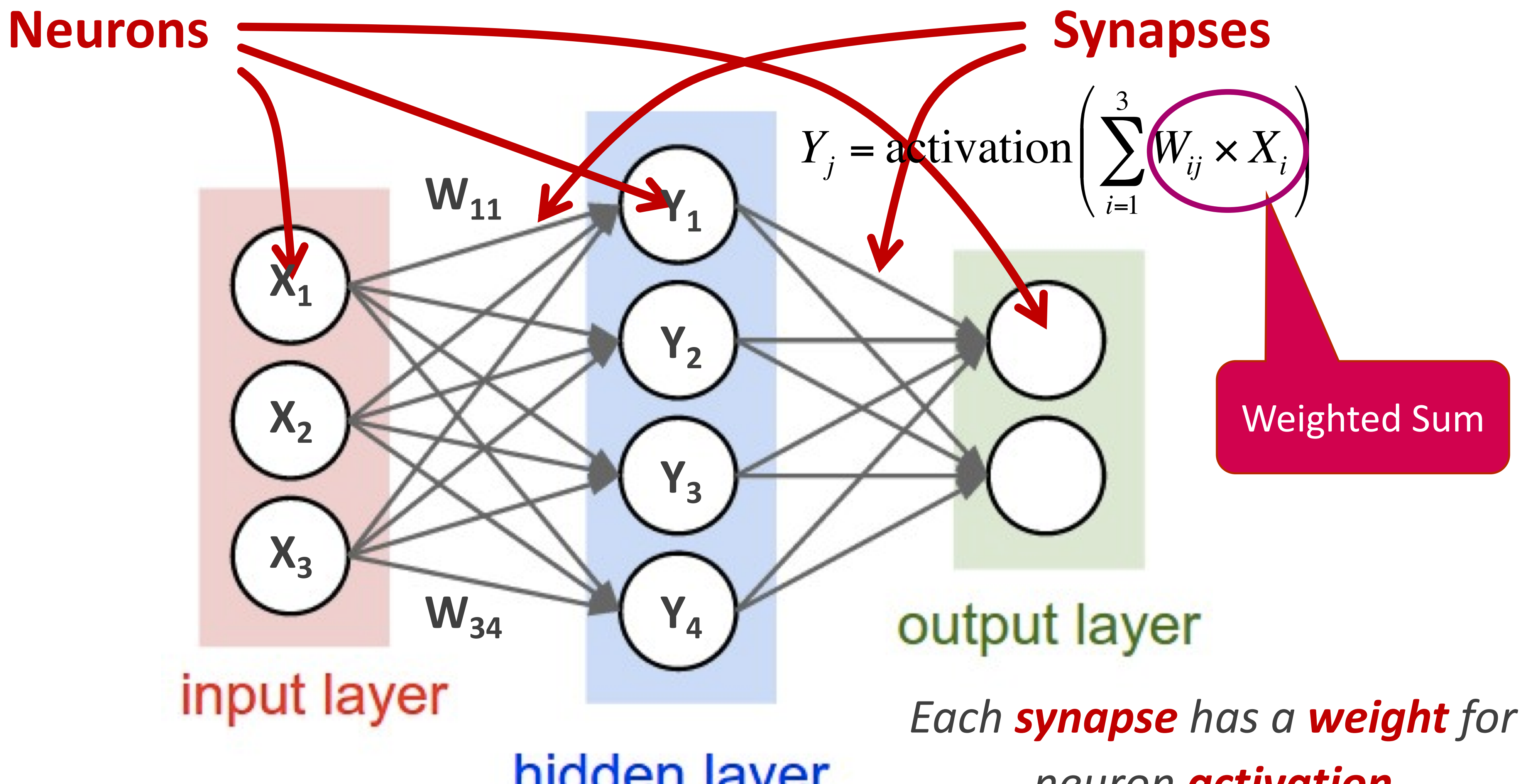
Synthesis Lectures





Relevant Background

What is a Deep Neural Network?



[Image Source: Stanford]

*Each **synapse** has a **weight** for neuron **activation***

Modern Deep Learning Landscape

Model Creation ⇄ Training → **Inference**

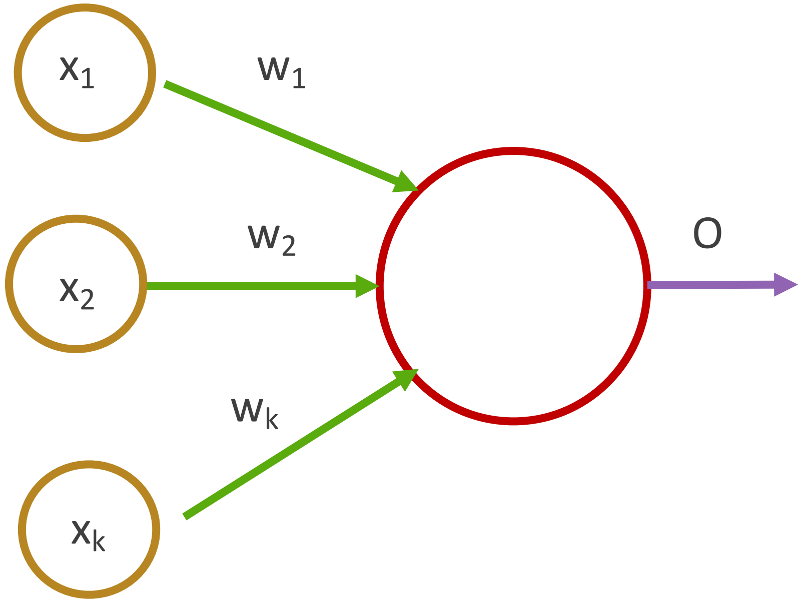
Convolutional Layers (Feature Extraction)

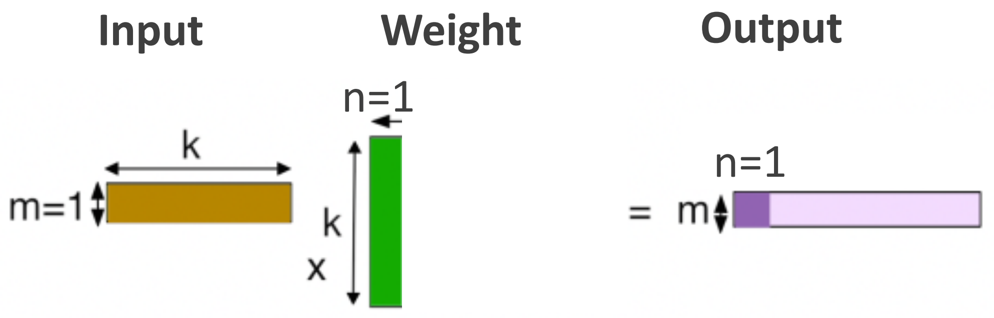
Summarize features

→ Conv. Layer → Conv. Layer … Conv. Layer → Pool. Layer → FC Layer → **"Georgia Institute of Technology, Atlanta"**

Intermediate features

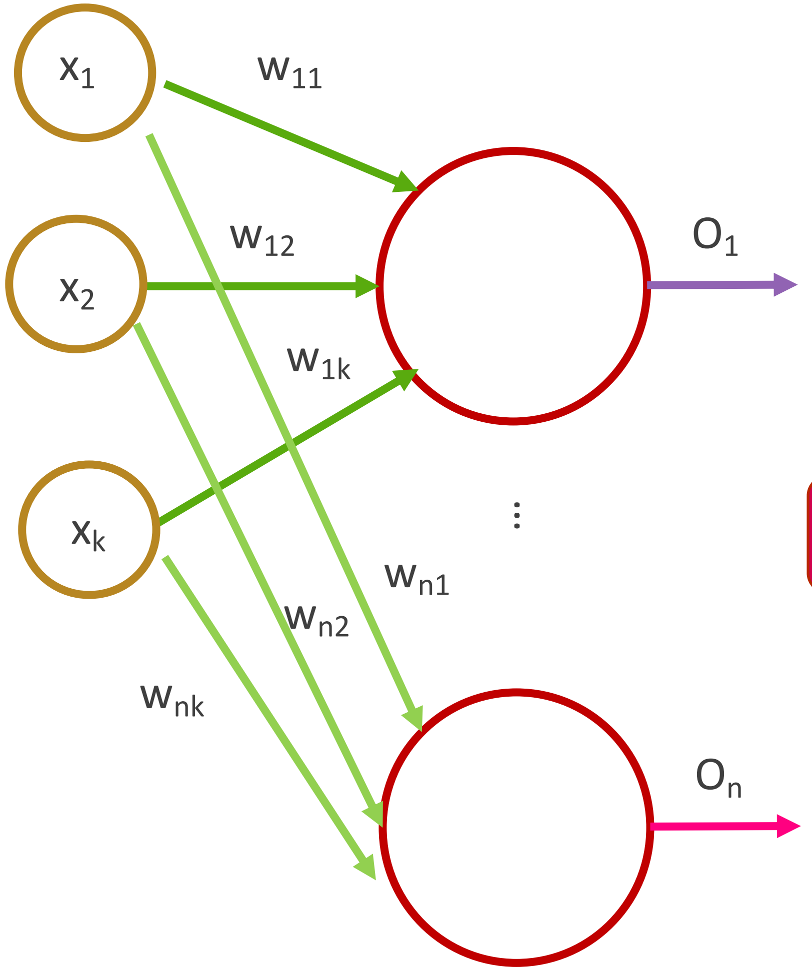
Computations in a DNN → Linear Algebra

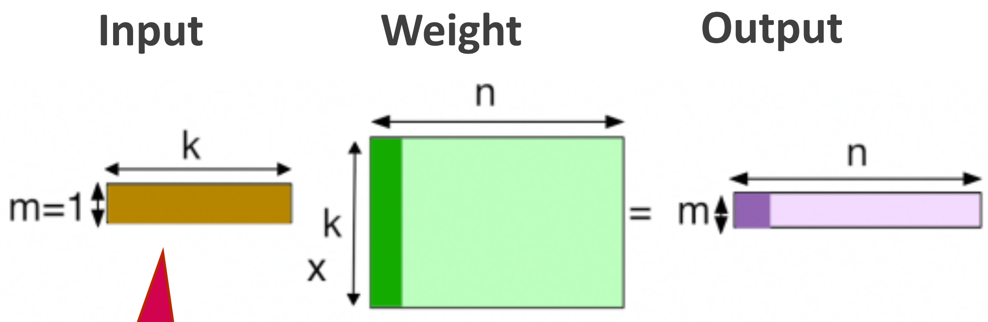




Neuron => Vector x Vector

Computations in a DNN → Linear Algebra

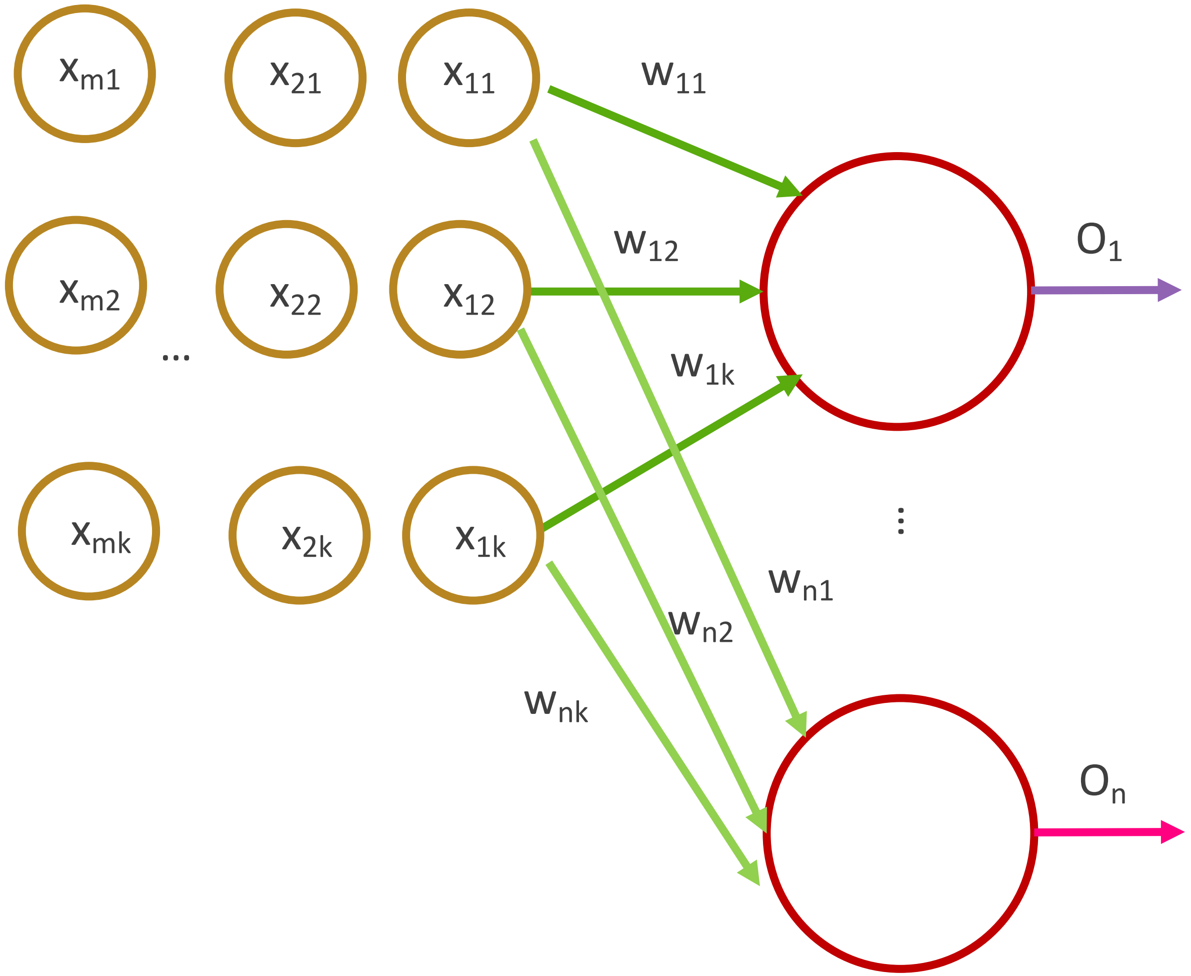


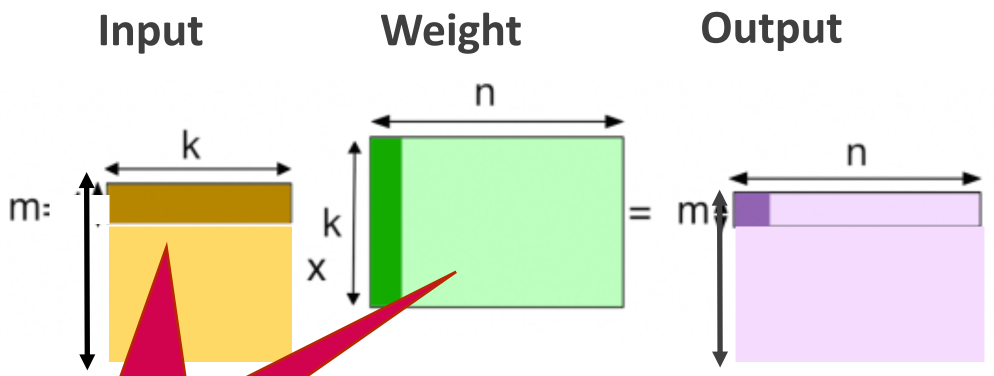


DNN Layer => Vector x Matrix

Data "Reuse"

Computations in a DNN → Linear Algebra



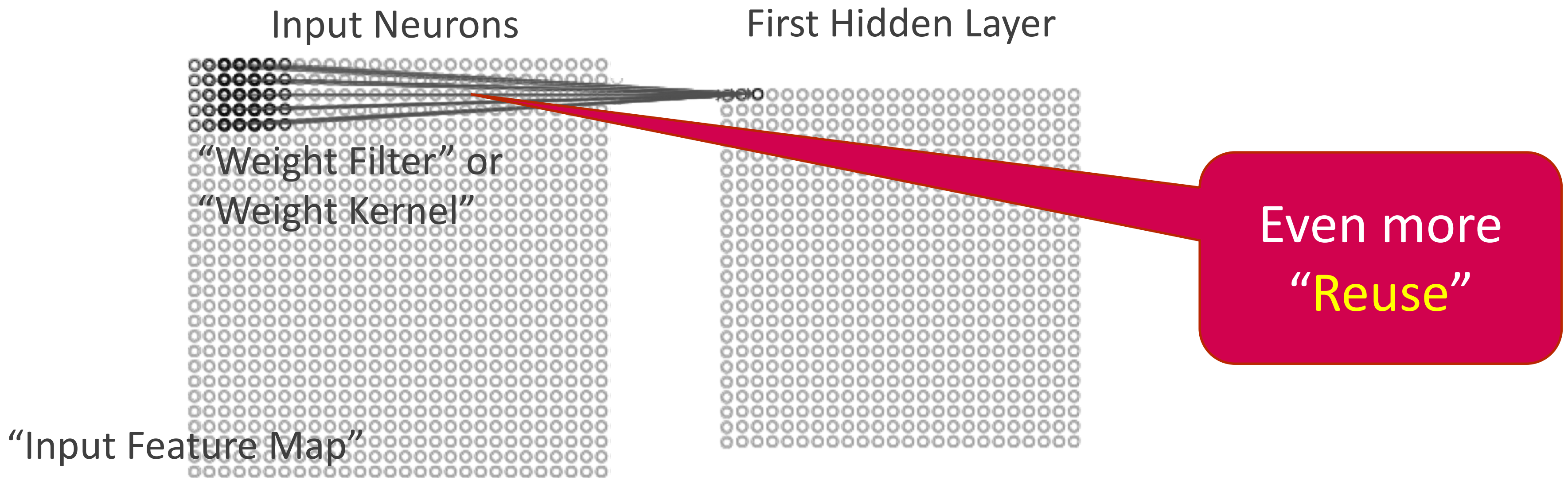


Batching => Matrix x Matrix

Data "Reuse"

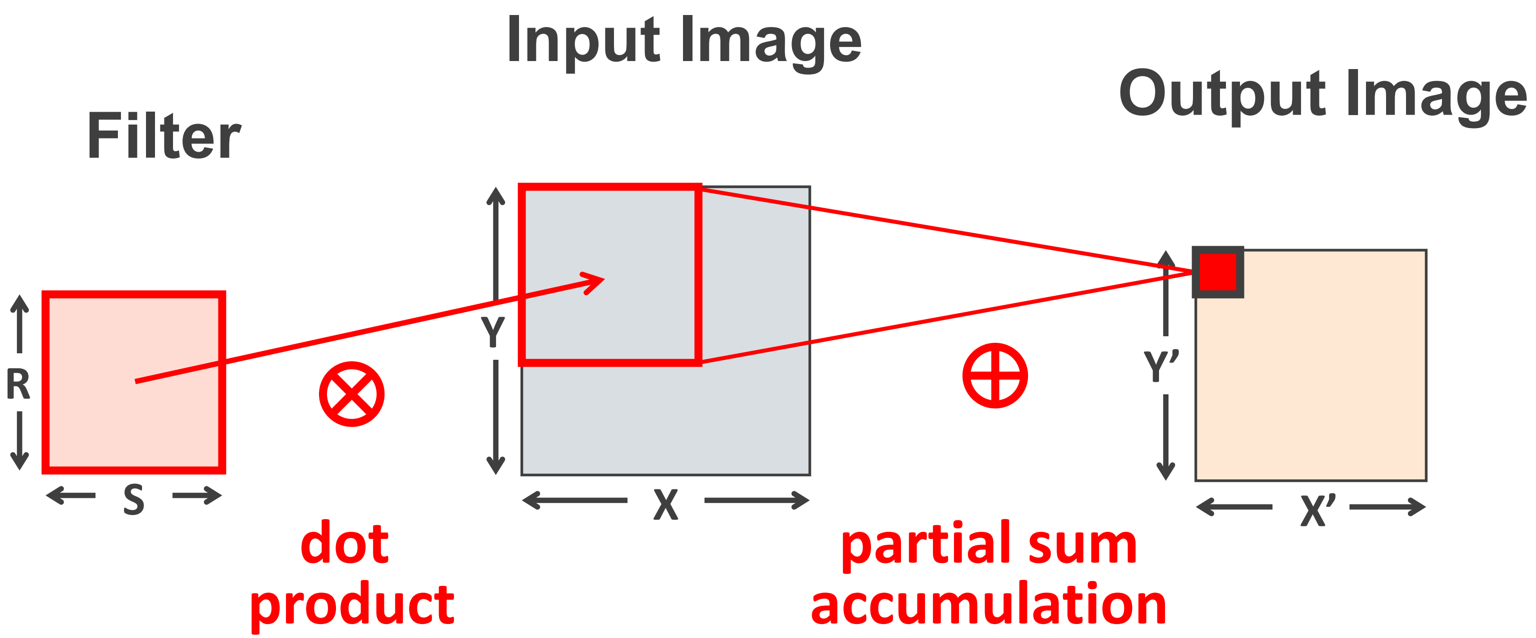
GEMM

Convolutional Neural Networks

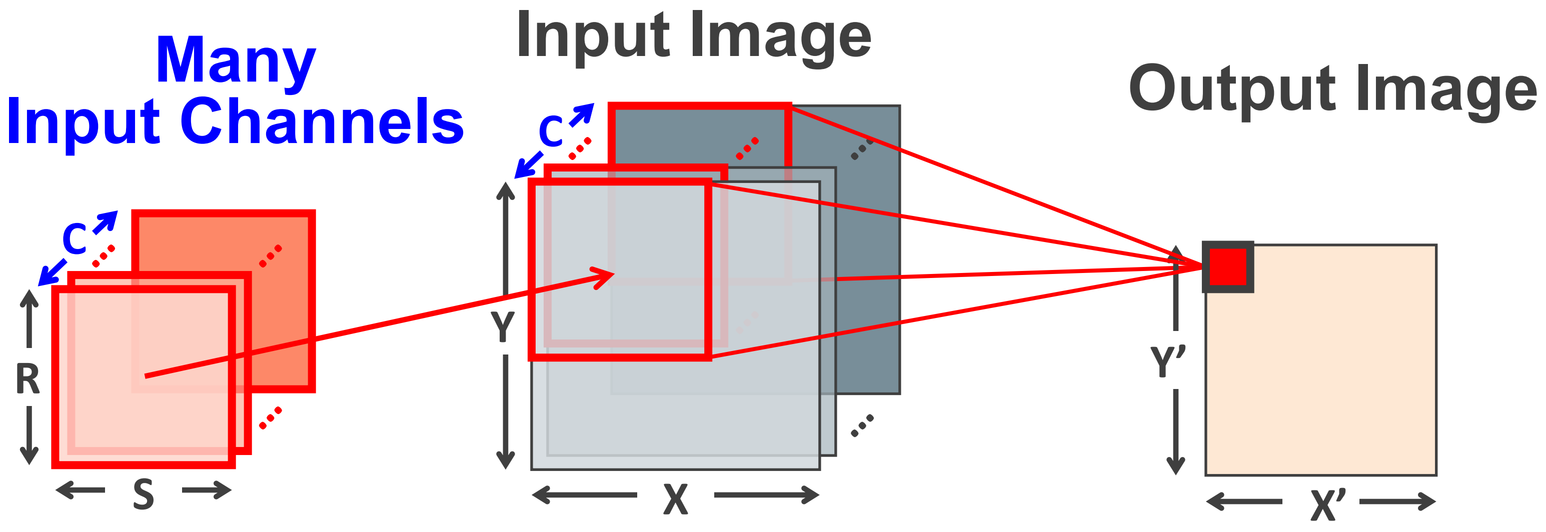


Shared Weights:
All neurons use the *same* filter weights

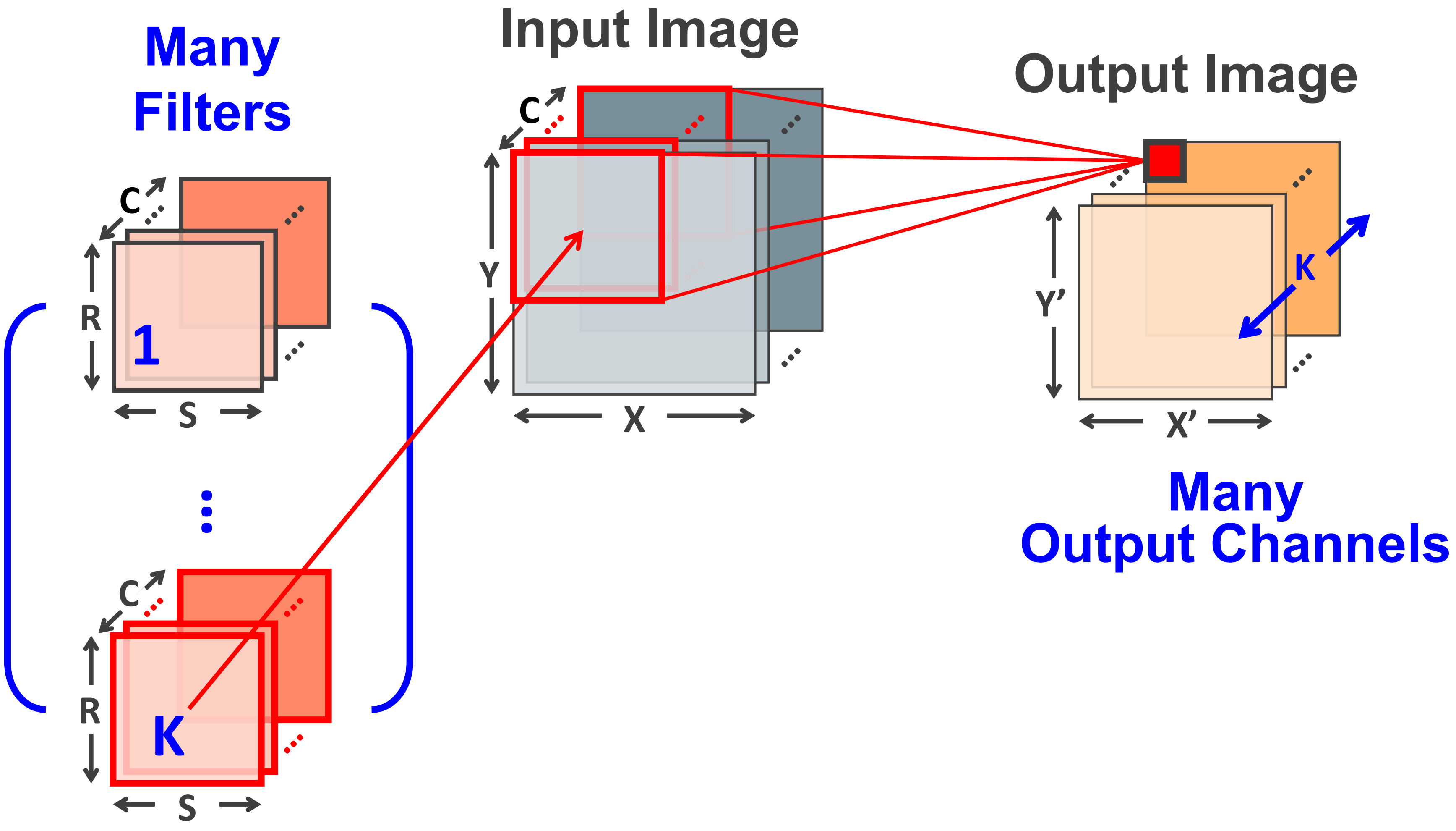
Convolution in CNN



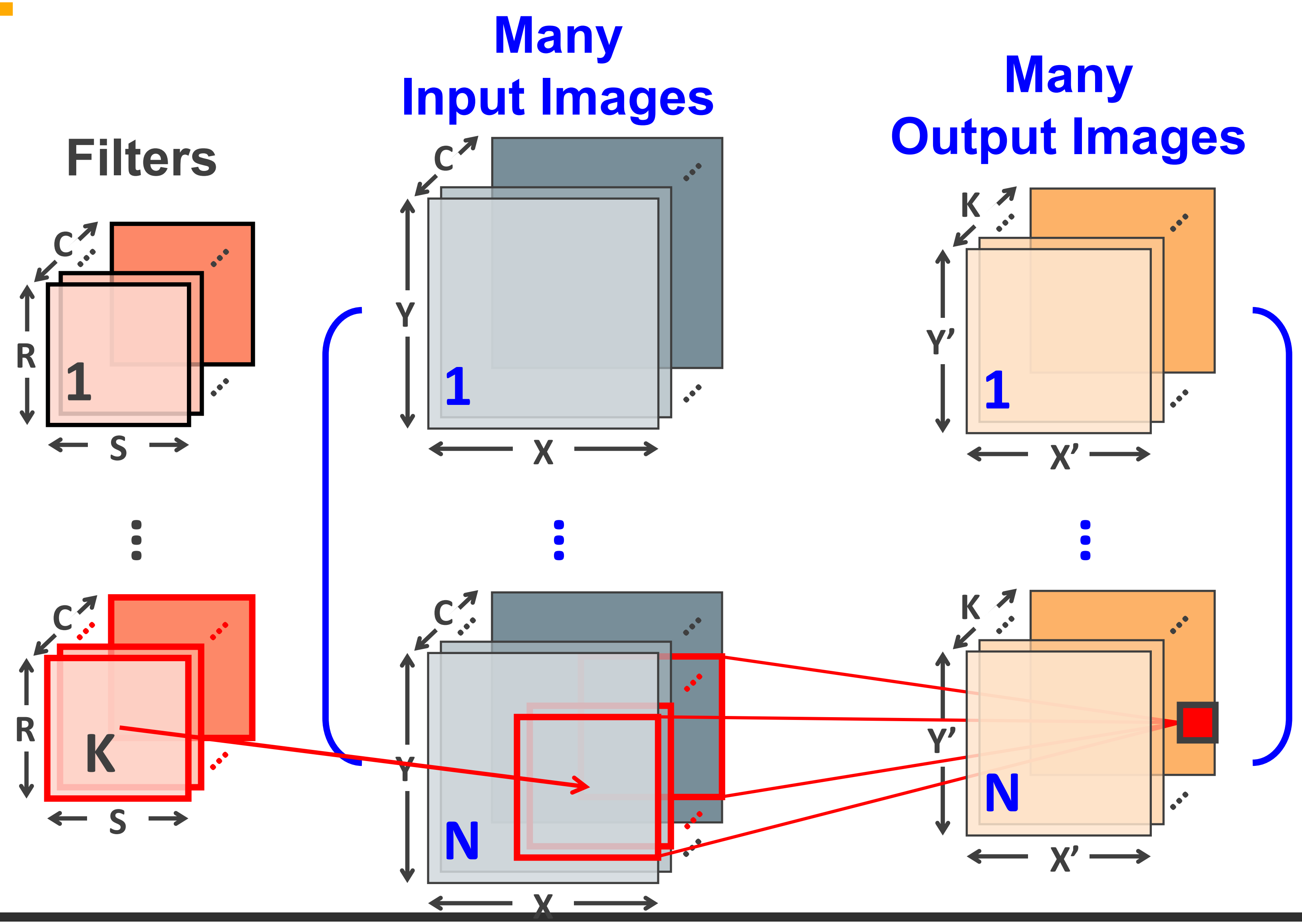
Convolution in CNN



Convolution in CNN



Convolution in CNN



7 loops

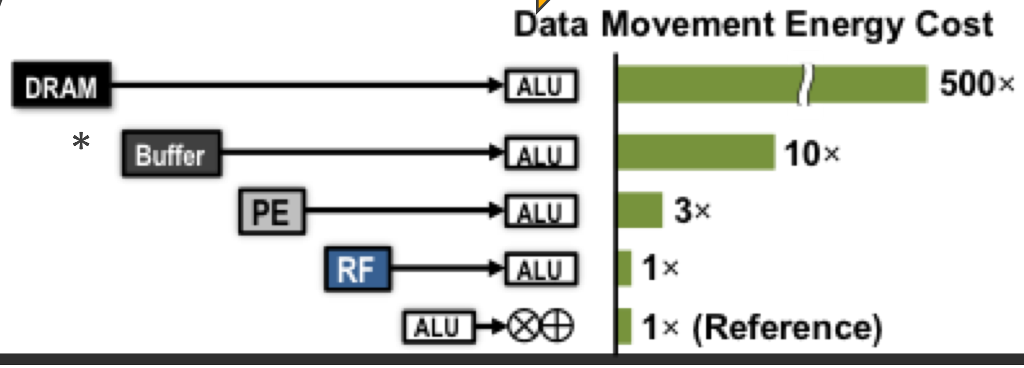
Challenges with DNN Computations

- **Millions of Parameters (i.e., weights)**
  - Billions of computations ⇨ **Need lots of parallel computations**

This makes CPUs inefficient

Neural Networks	Number of Weights
AlexNet (2012)	3.98M
VGGnet-16 (2014)	28.25M
GoogleNet (2015)	6.77M
Resnet-20 (2016)	0.27M
Resnet-110 (2016)	1.7M

  - Heavy data movement ⇨ **Need to reduce energy**



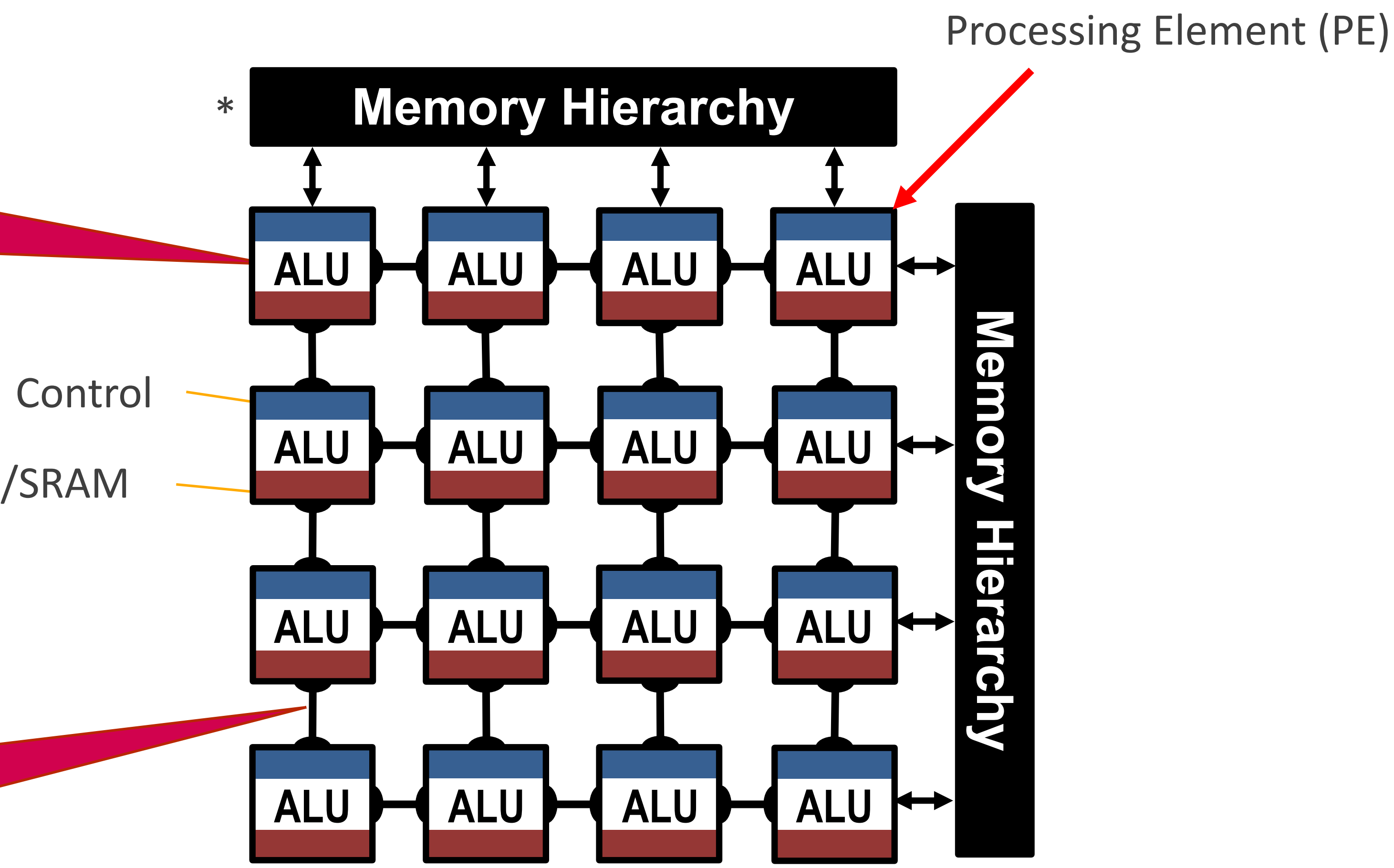
This makes GPUs inefficient

*Y. Chen et. al., "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," ISCA, 2016.

Spatial (or Dataflow) Accelerators

- **Millions of Parameters (i.e., weights)**
  - Billions of computations

Spread computations across hundreds of ALUs



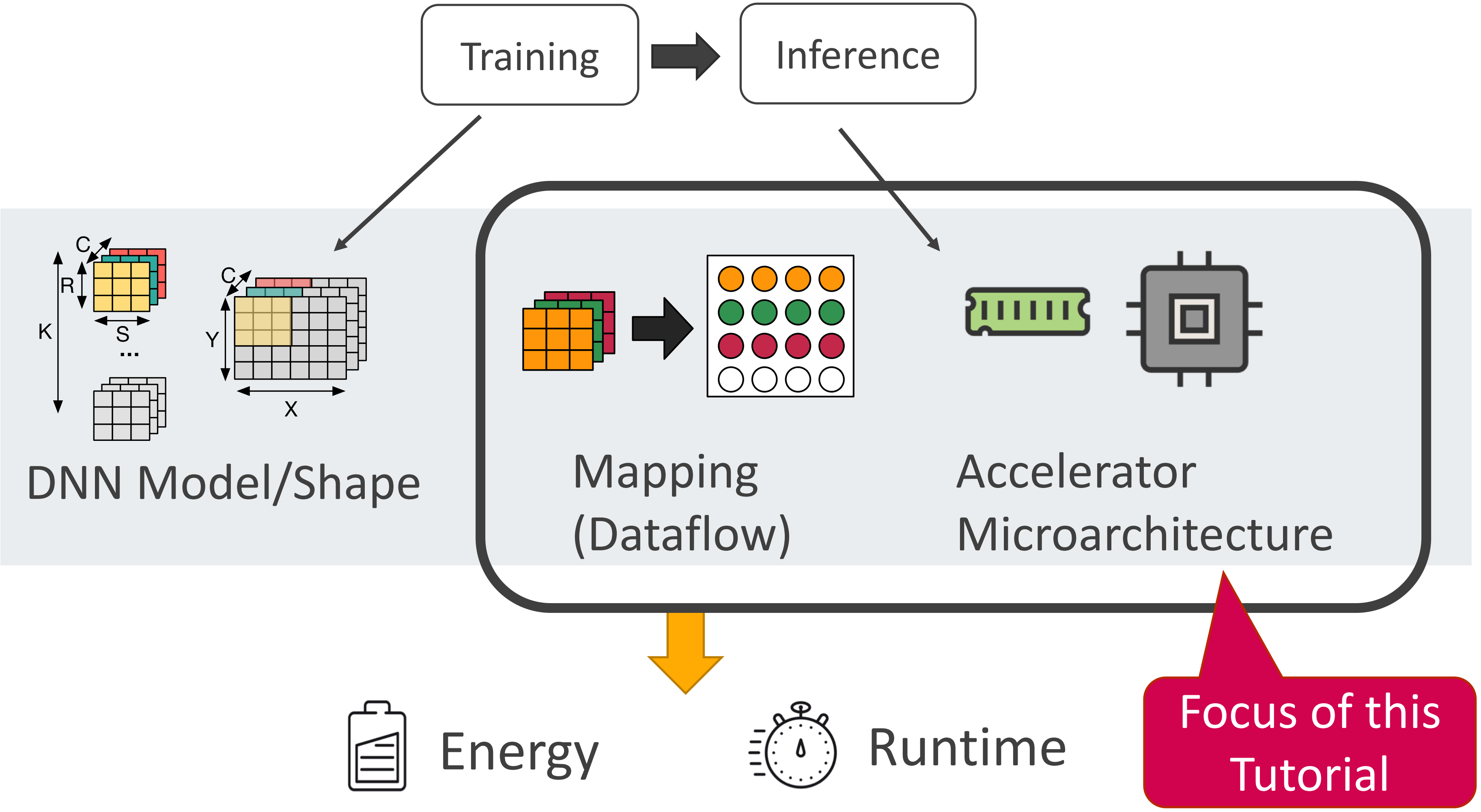
  - Heavy data movement

Reuse data within the array via local memories and direct communication

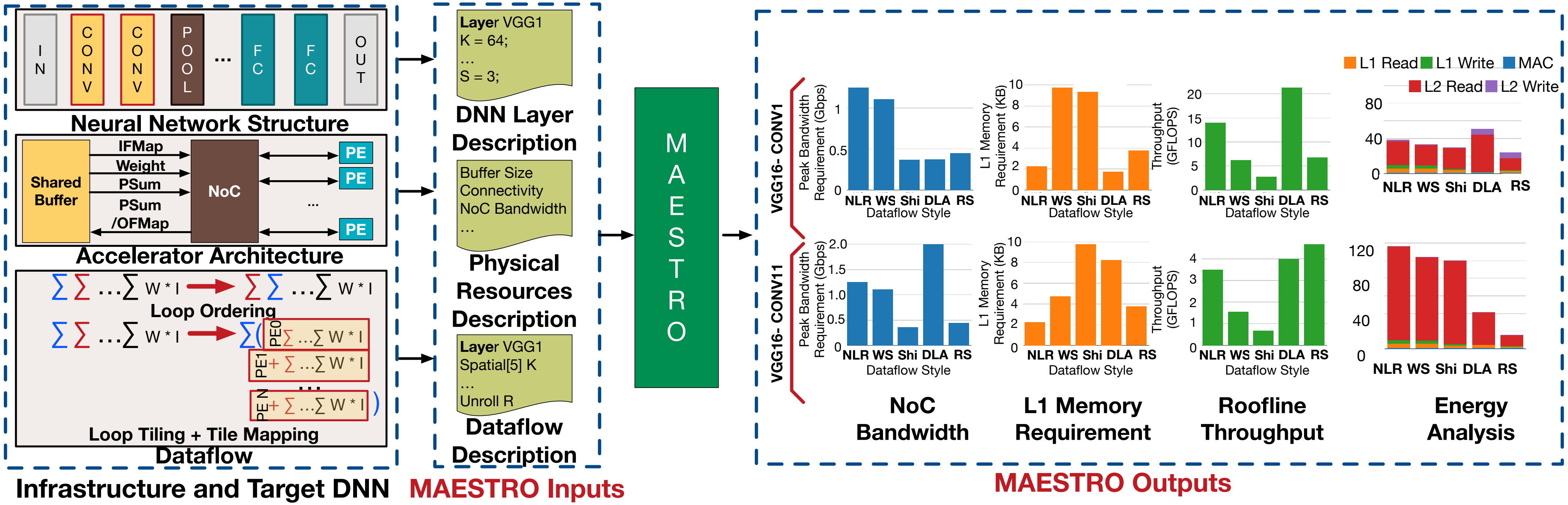
Examples: MIT Eyeriss, Google TPU, …

***Y. Chen et. al., "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," ISCA, 2016.**

Challenges in Design and Deployment



MAESTRO: Analytical cost model for DNN dataflows



Hyoukjun Kwon, Prasanth Chatarasi, Michael Pellauer, Angshuman Parashar, Vivek Sarkar and Tushar Krishna

"Understanding Reuse, Performance, and Hardware Cost of DNN Dataflows: A Data-Centric Approach *using MAESTRO*"

MICRO 2019, *IEEE Micro Top Picks 2020*

Welcome to the Tutorial!

Schedule (EDT)

Time	Agenda	Presenter
10:00 - 10:10	**Welcome + Intro to DNNs**	Tushar
10:10 - 10:45	**DNN accelerators and dataflows**	Michael
	MAESTRO Cost Model	
10:45 - 11:15	Data-centric Mapping Directives	Hyoukjun^
11:15 - 11:30	Analytical Cost Model	Hyoukjun^
11:30 - 12:00	Compiling and Running MAESTRO	Geonhwa
12:00 - 12:10	**Break**	
	Automated Design-space Exploration using MAESTRO	
12:10 - 12:25	Marvel: Mapping Space Exploration via Heuristics	Prasanth
12:25 - 12:40	GAMMA: Mapping Space Exploration via Optimization	Felix
12:40 - 12:55	ConfuciuX: Hardware Design-space Exploration via RL and Optimization	Felix
12:55 - 13:00	**Wrap Up**	Tushar

^Pre-recorded Video

https://tinyurl.com/maestro-tutorial-micro2020

Brief Q/A at the end of each talk.

Please use Zoom's Q/A feature. You can upvote other questions. We will answer live or respond in the Q/A box.

Slides + Videos will be available on the MAESTRO tutorial website