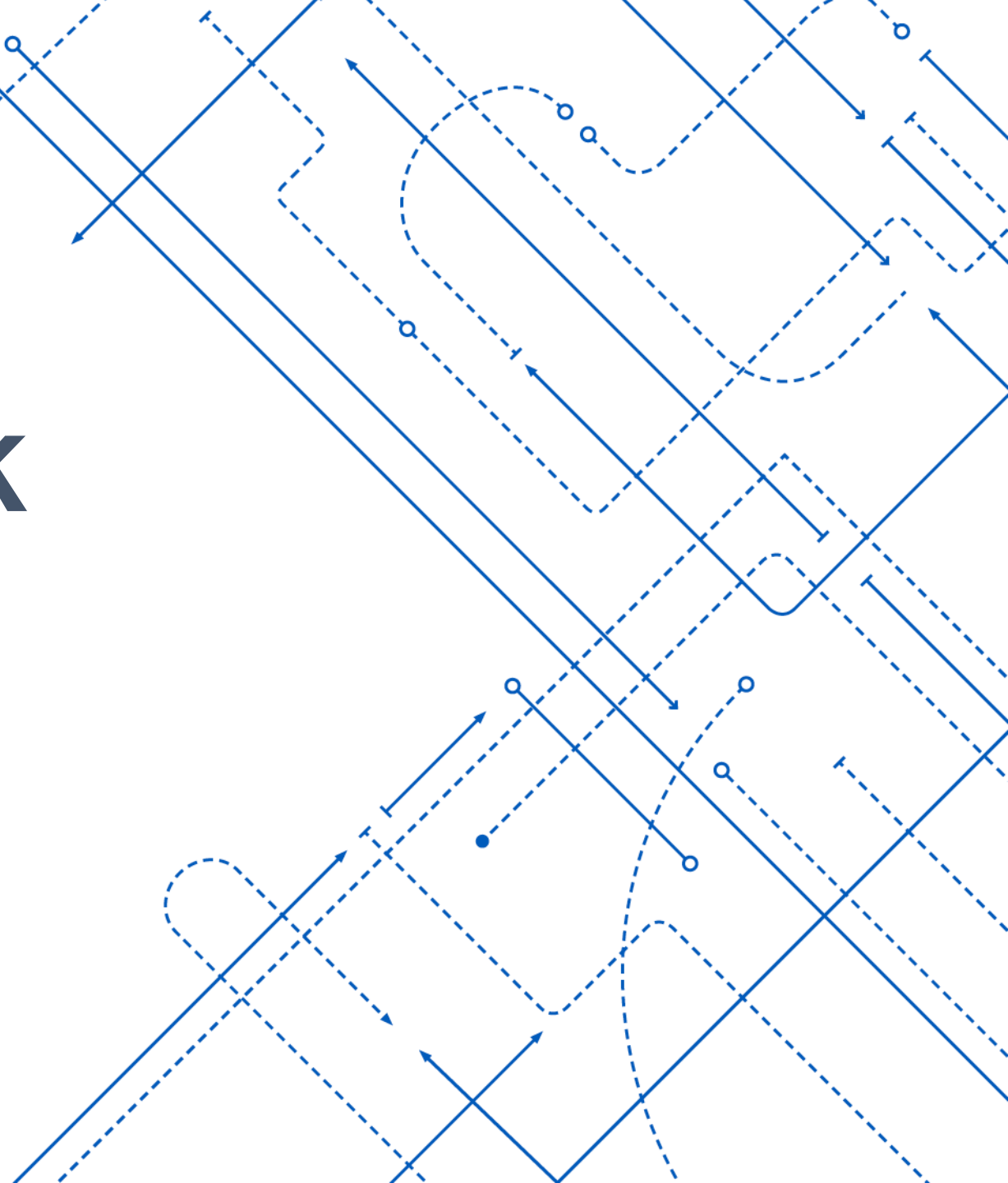
CONVOLUTION NEURAL NETWORK IN CUDA

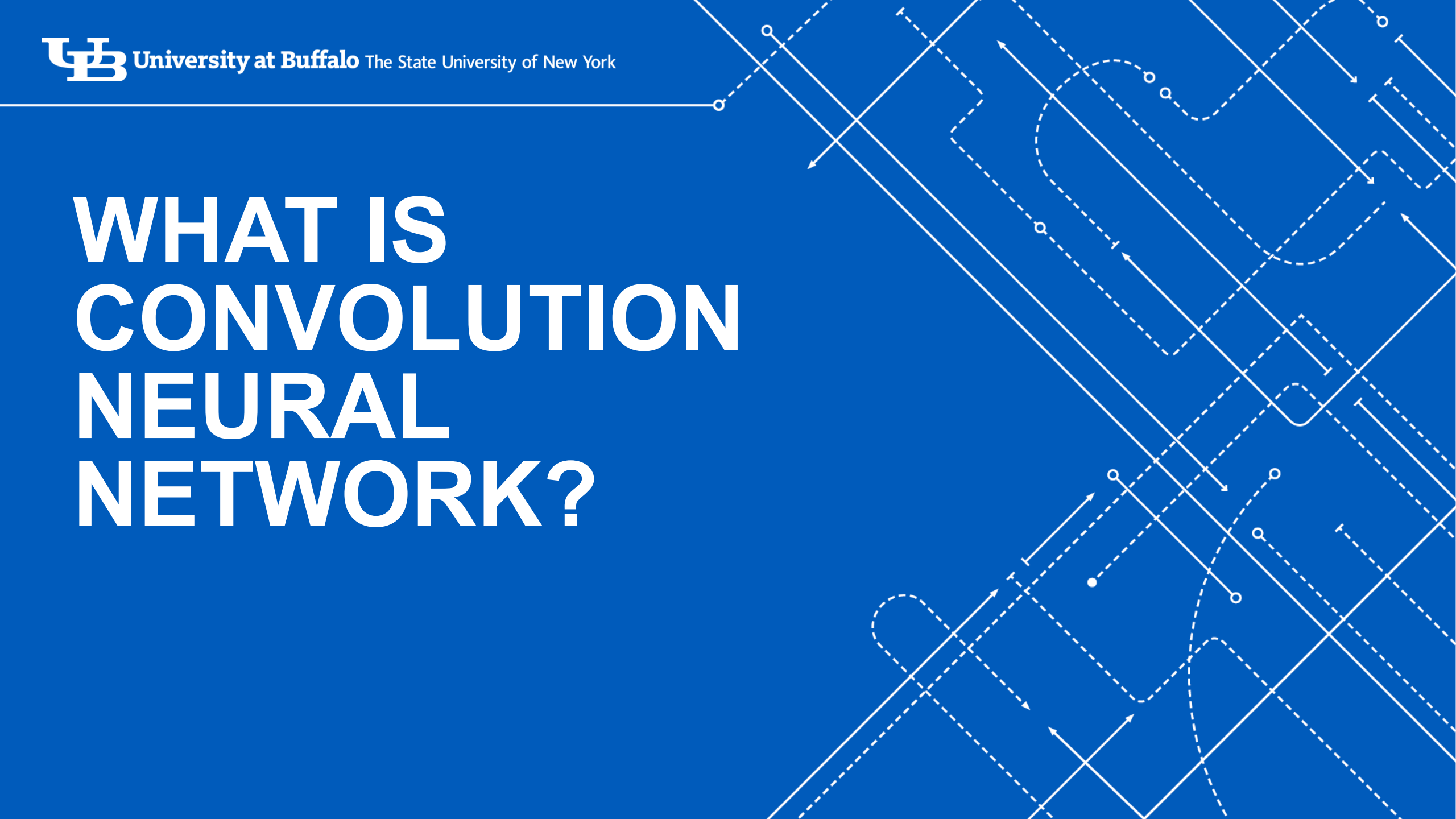
Jean Vigroux



University at Buffalo The State University of New York

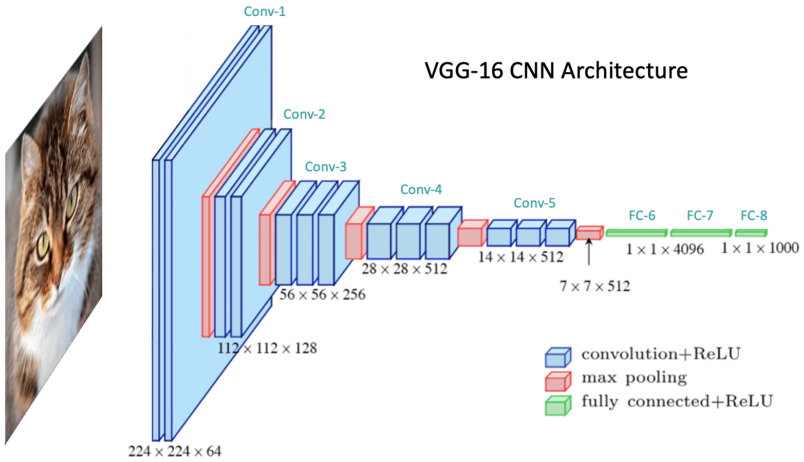
University at Buffalo The State University of New York

WHAT IS CONVOLUTION NEURAL NETWORK?



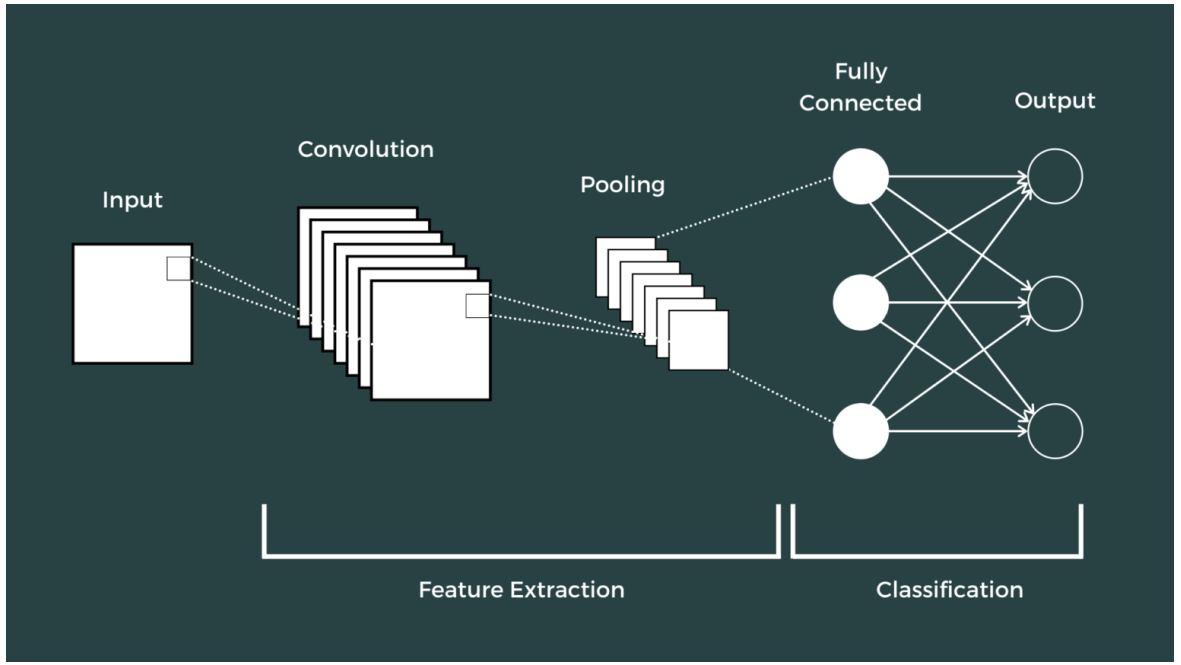
Convolution Neural Network

- Type of deep learning neural network
- Typically used for classification using images
- Computational demanding
  - Training
  - Large Labeled data



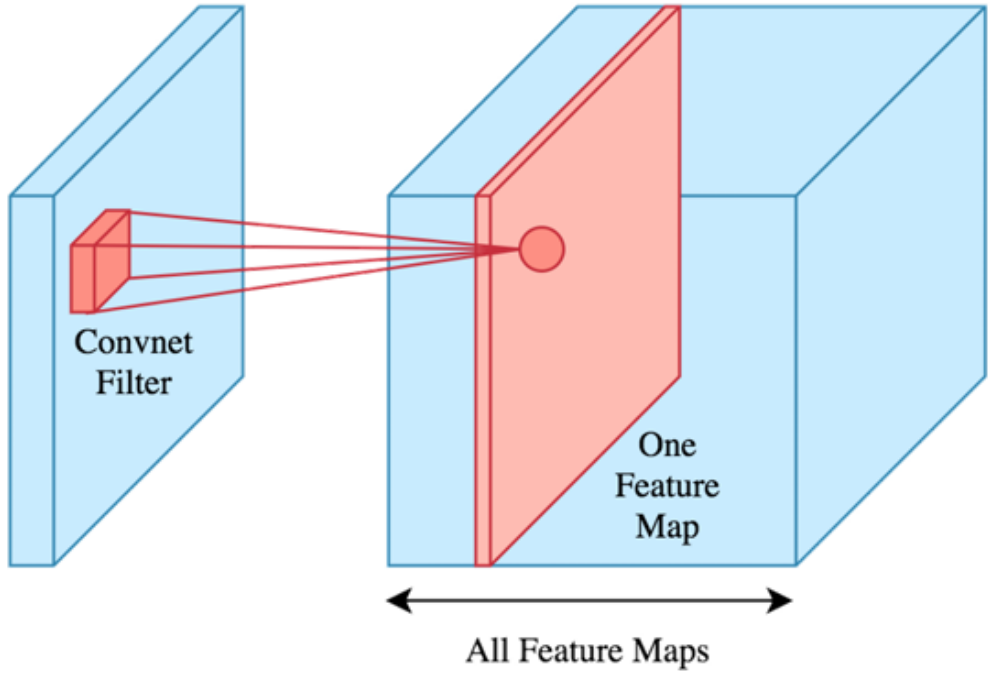
CNN Structure

- **Feature Extraction**
  - Applies current kernel/s to Input
  - Activation function is applied
  - Use pooling layer
- **Classification**
  - Fully Connected layer is used to create output



CNN Structure

- **Feature Extraction**
  - **Applies current kernel/s to Input**
  - Activation function is applied
  - Use pooling layer



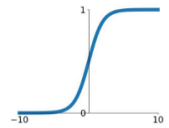
CNN Structure

- **Feature Extraction**
  - Applies current kernel/s to Input
  - **Activation function is applied**
  - Use pooling layer

Activation Functions

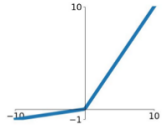
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



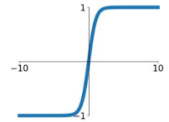
Leaky ReLU

$\max(0.1x, x)$



tanh

$\tanh(x)$

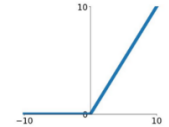


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

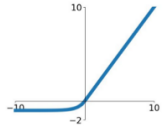
ReLU

$\max(0, x)$



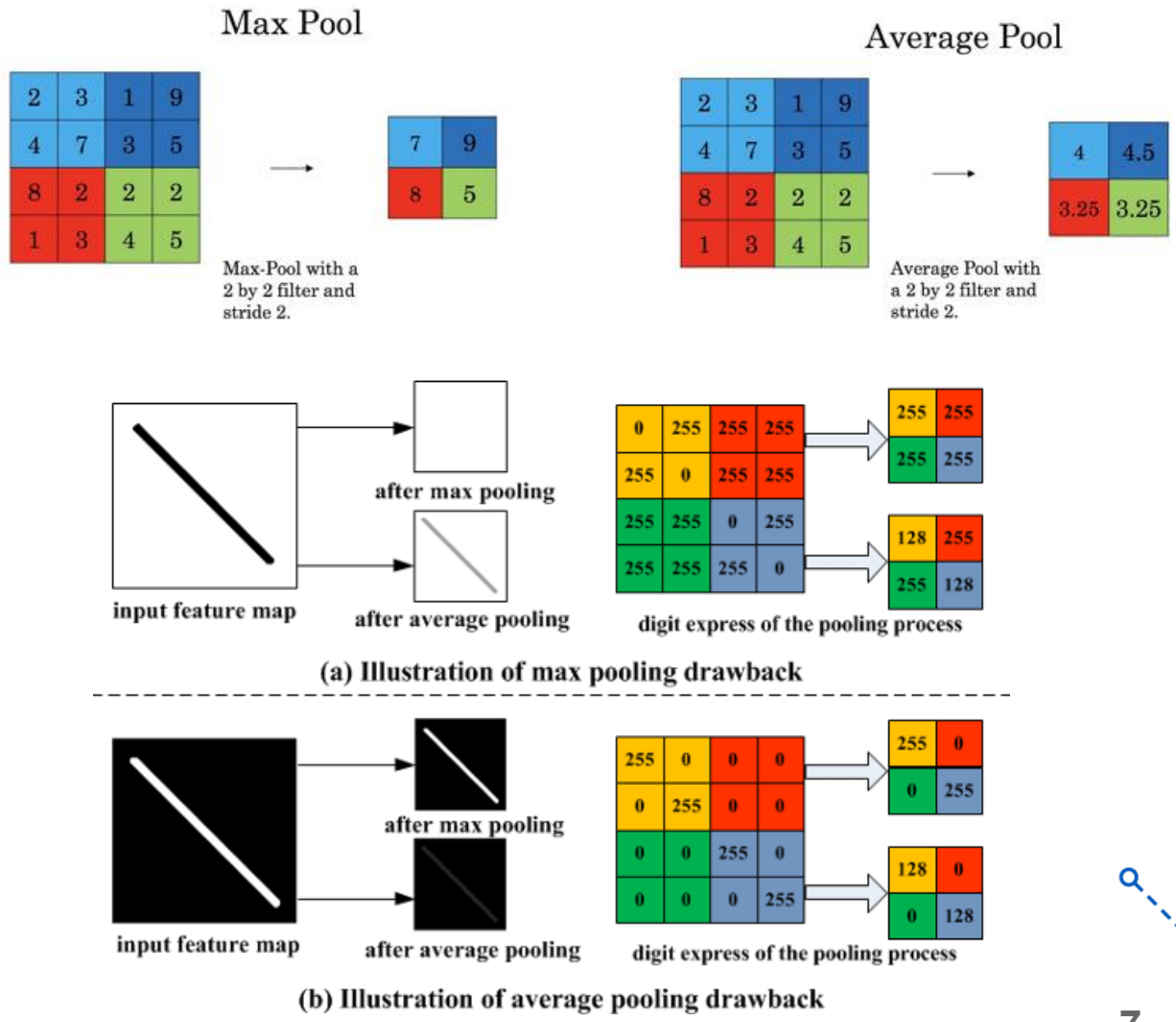
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



CNN Structure

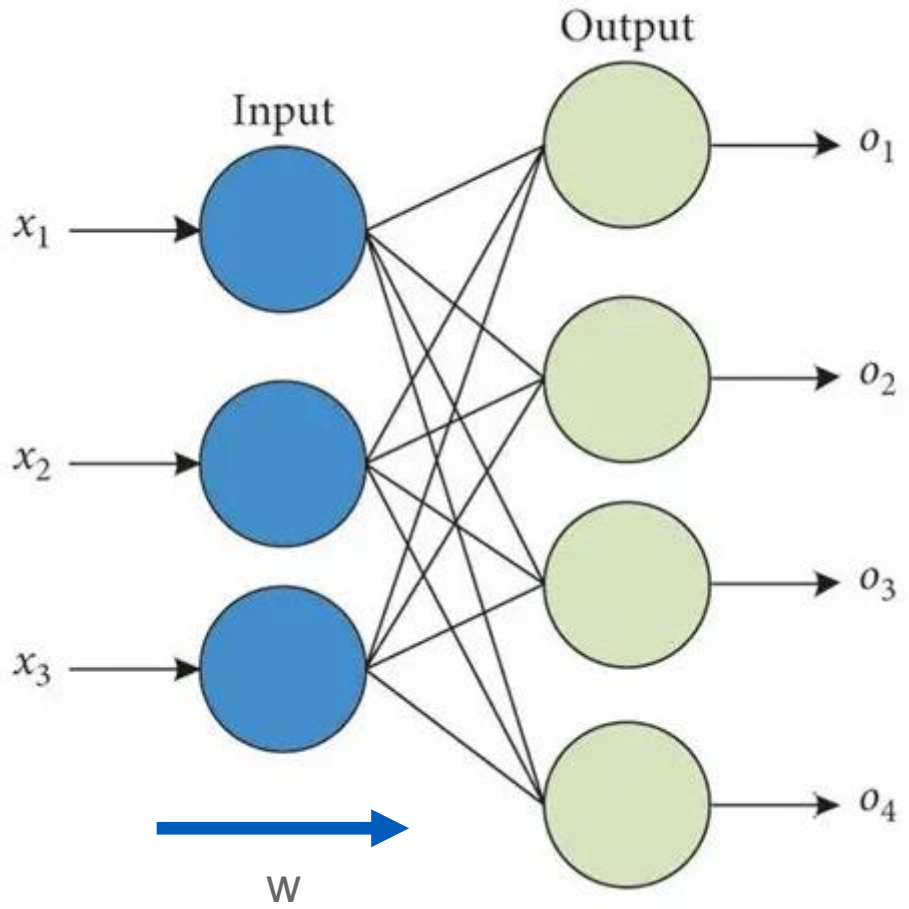
- **Feature Extraction**
  - Applies current kernel/s to Input
  - Activation function is applied
  - **Use pooling layer**



CNN Structure

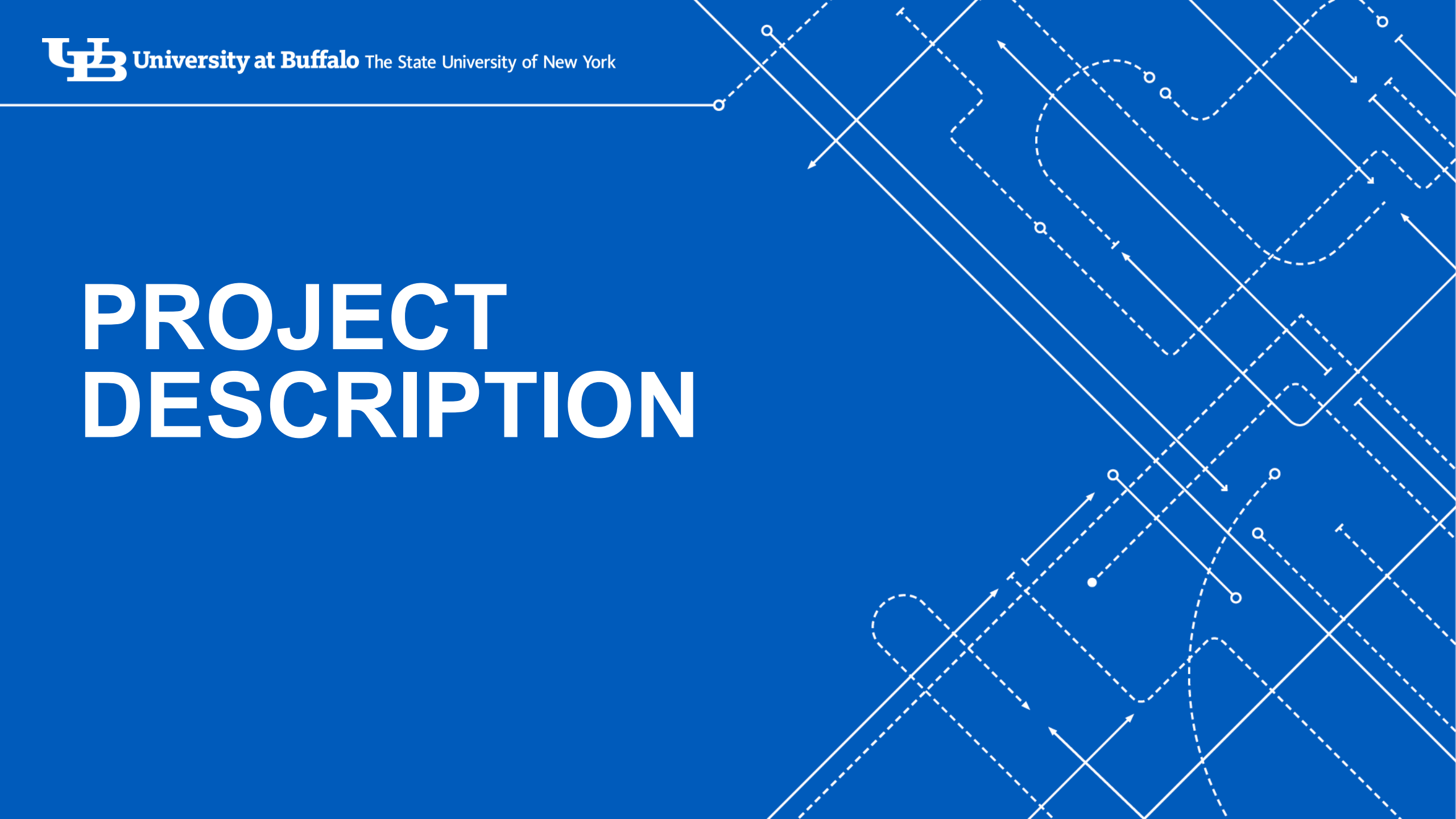
- **Classification**
  - **Fully Connected layer is used to create output**

$$Output = Input * Weight + bias$$



University at Buffalo The State University of New York

PROJECT DESCRIPTION



CNN in Parallel

- Convolutional Neural Network are ideal for parallel implementation
  - Matrix Multiplication
  - Large Data

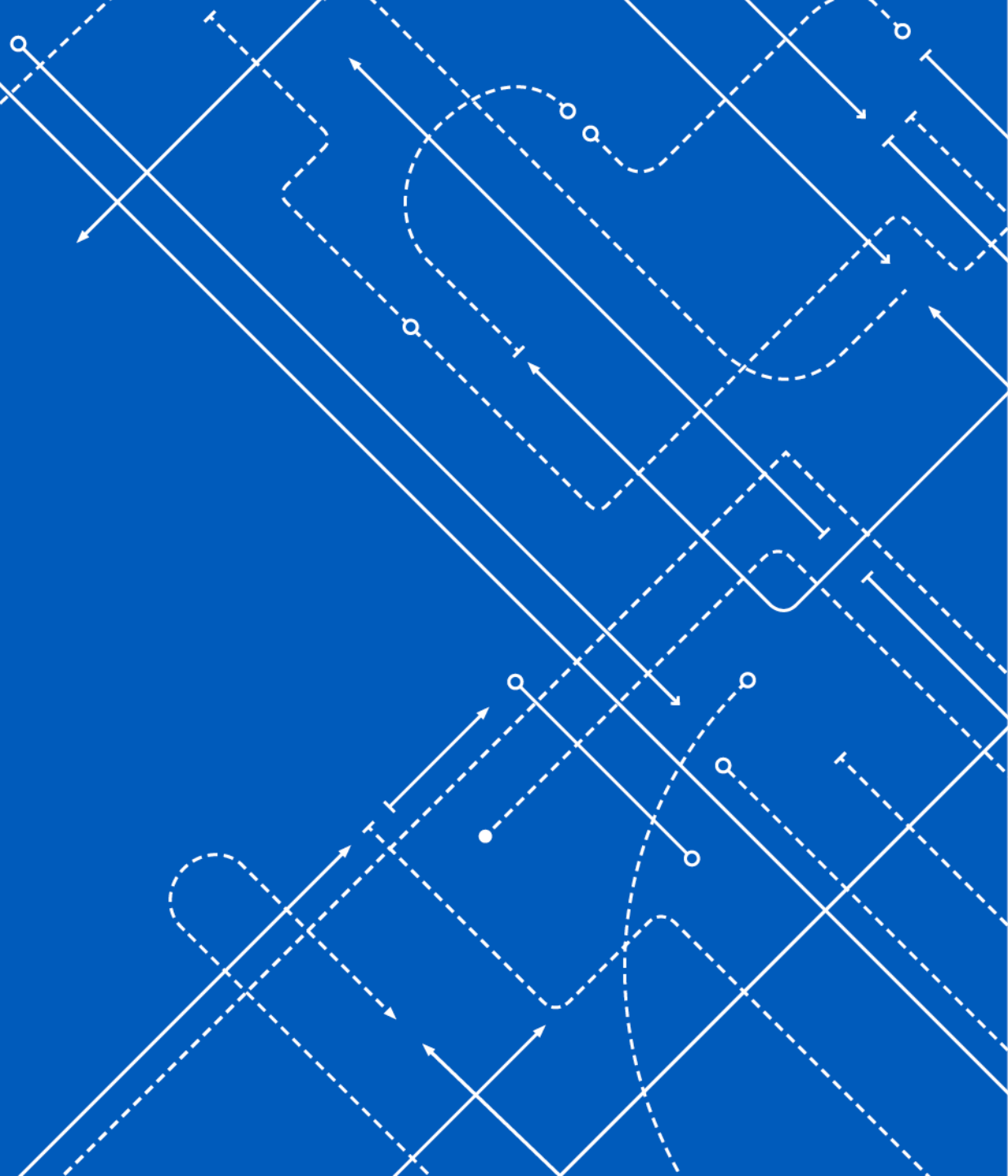
- Nvidia CUDA parallel computing platform
  - Allows utilization of Nvidia GPUs
  - GPU has the ability to run more threads
  - Shared memory within thread blocks

Hardware

- **GPU: NVIDIA GeForce RTX 4070 Ti**
  - CUDA Cores: 7680
  - Streaming Multiprocessor: 60
  - Architecture: Ada Lovelace
  - VRAM: 12 GB GDDR6X
    - Data rate: 21 Gbps
    - Interface: 192-bit
    - Bandwidth: 504.05 GB/s
- **CPU: AMD Ryzen 7950x 16-Core Processor**
- **RAM: 64 GB DDR5**

University at Buffalo The State University of New York

PROJECT DESIGN & ALGORITHM



Design

Dataset
Large Dataset EX: 1.2m Images

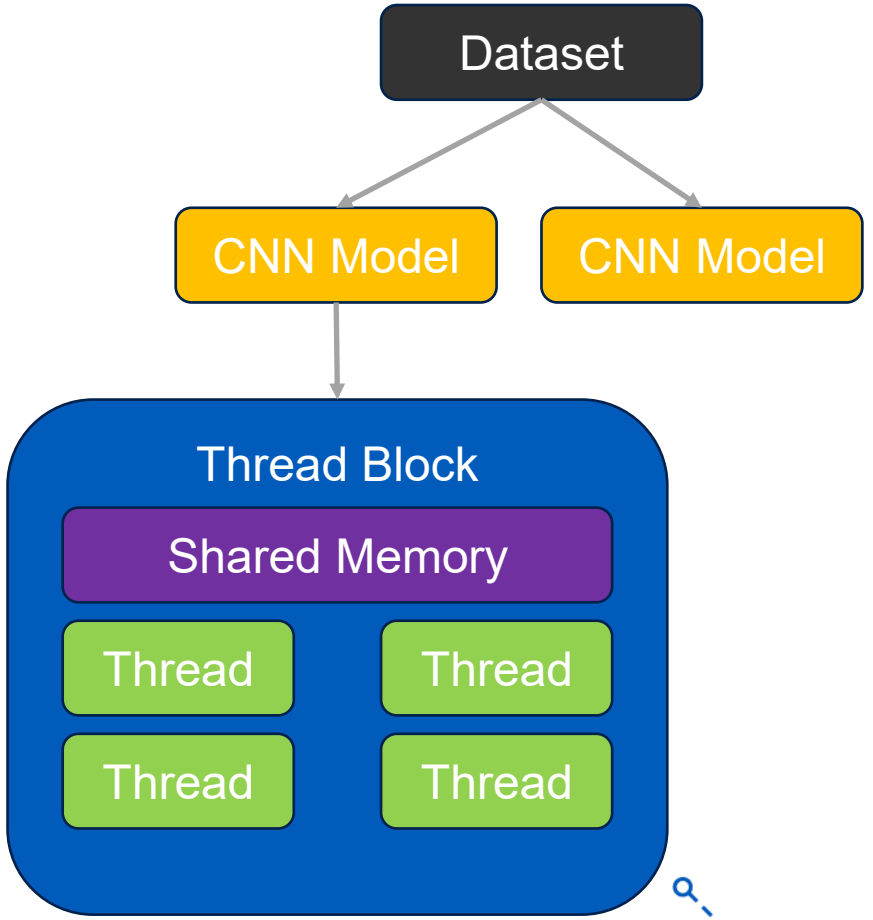
CNN Model
For each data entry, there will be an instant of the CNN model

Shared Memory
For each data entry, there will be an instant of the CNN model

Data
- Kernel
- Input
- Output

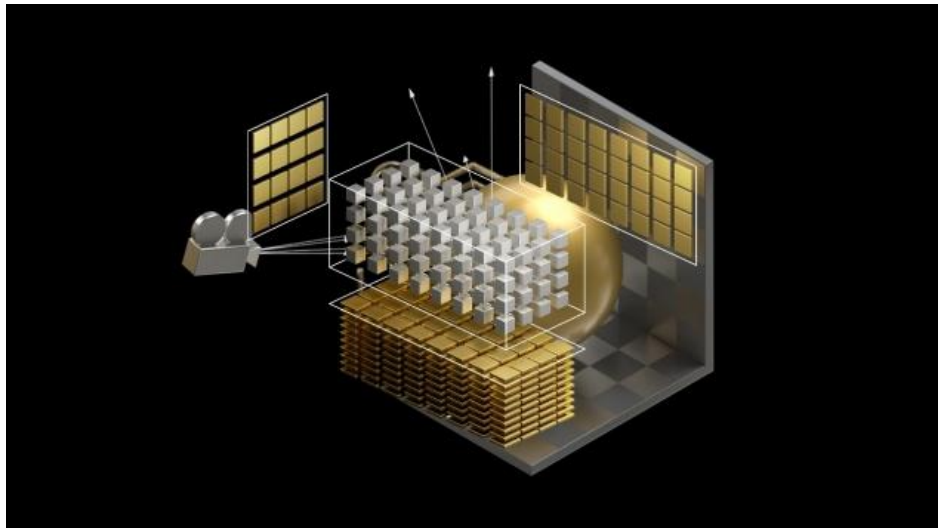
Thread Block
Uses number of threads equal to the size of the largest convolution layer

Thread
Each thread will calculate use the kernel to calculate an index for each layer



Limitations

- Ada Lovelace Architecture
  - Streaming Multiprocessor
    - Max Thread Blocks: 24
    - Max Threads: 1536 | Max Warp: 48
  - Thread Block
    - Max Threads: 1024
    - Shared Memory: 99 KB
- Shared Data < 48 KB
- Valid Data
- Activation function is always ReLU
- Static amount of layers



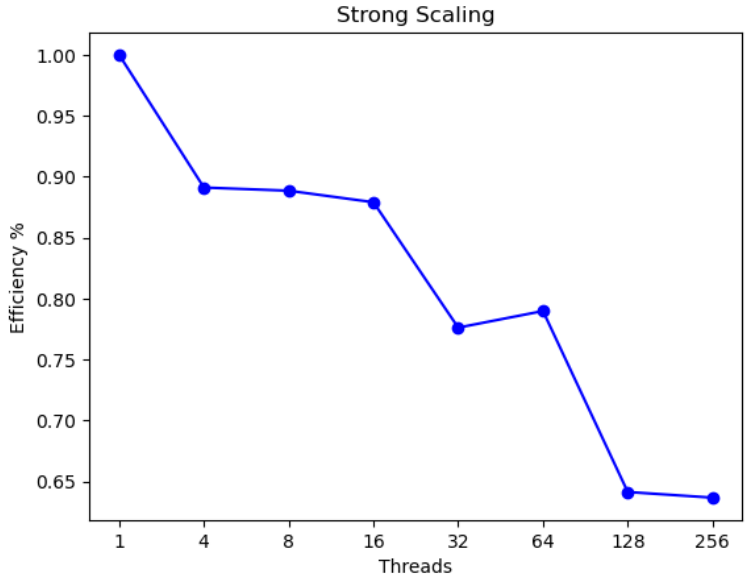
University at Buffalo The State University of New York

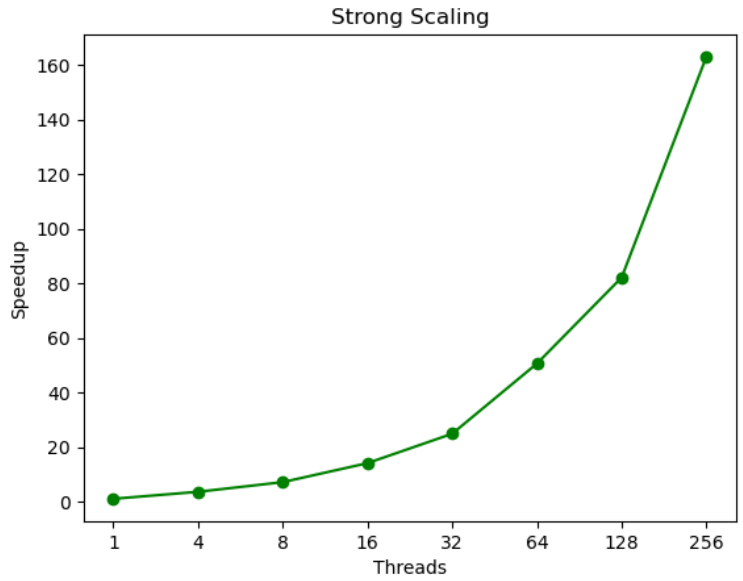
RESULTS

Results – Strong Scaling

- Input: 64 x 64
- Batch Size: 128

Threads	Speedup	Efficiency
1	1	1
4	3.56	0.89
8	7.10	0.88
16	14.06	0.87
32	24.83	0.77
64	50.54	0.78
128	82.07	0.64
256	162.96	0.63





University at Buffalo The State University of New York

THANK YOU!!!

Any questions?