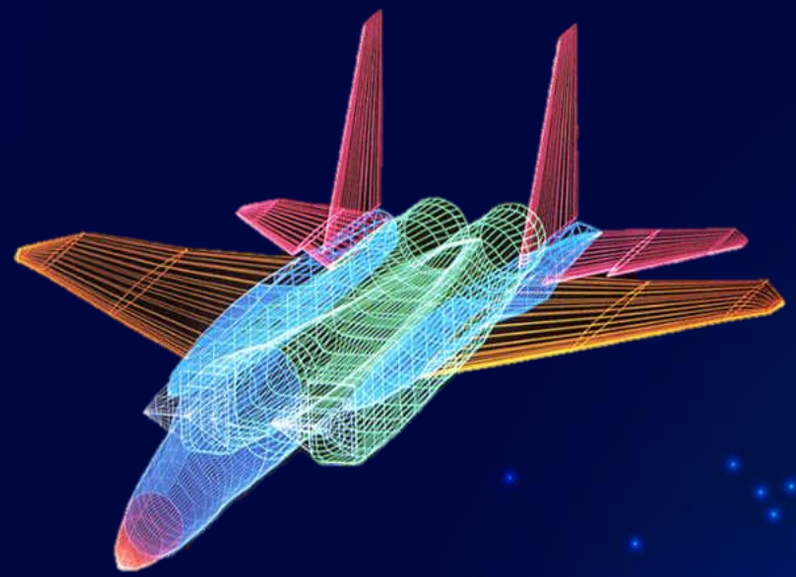
DL Deployment (Inference) Syllabus and Expectations

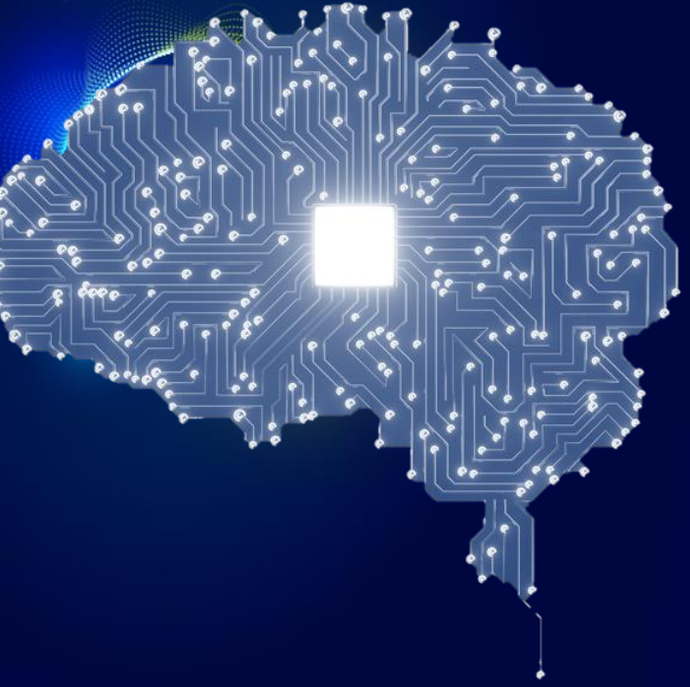
Oron Guterman



נעים להכיר ☺

https://www.linkedin.com/in/oron-guterman-30352879/

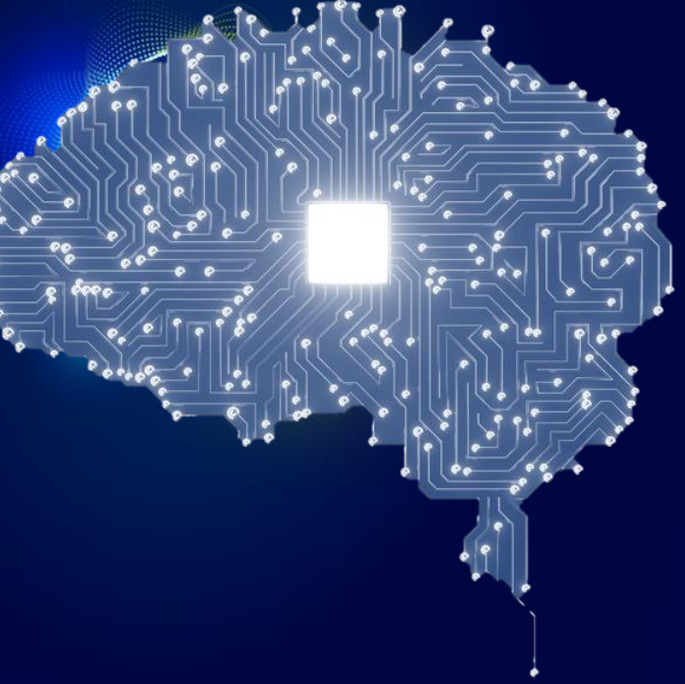
- Since 2017: HPC programmer
  Working in the fields of High Performance Computing (HPC) with focus on classic algorithms and Deep Learning (DL) trained models deployment (inference) on Heterogeneous architectures
- 2011-2017: SW team leader
  Avionics upgrades Smart Displays and Mission Computers
  Graphic SW solutions for civilian & military avionics
- 2005-2011: SW engineer
  Graphics & Displays domain, HMD
- 2004-2005: Student of SW engineering
  Graphics & Displays domain, Helmet Mounted Display (HMD) team



מטרה

Pre trained model deployment skills & resources

- ✓ הגדרת תפקיד איש HPC בכלל ובפרט בעולם התוכן של DL
- ✓ פרקטיקה יום יומית שלנו כאנשי DL Deployment בצוות HPC
- ✓ מה היכולות והכישורים הנדרשים
- ✓ מה הם המשאבים הנדרשים לשם השגת יכולות וכישורים אלו
- ✓ ממשקים
- ✓ אתגרים, מעצורים פוטנציאליים, דרכי התמודדות
- ✓ חדשנות



נושאים

Project:

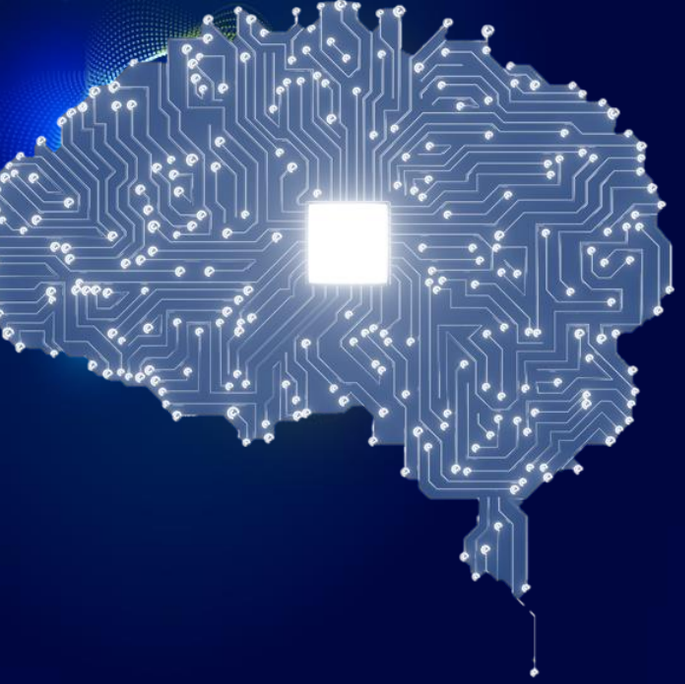
- Optimize & deploy pretrained model using TensorRT
- Filter image based on blur filter using Numba (Bonus)
- project-instructions.ipynb

Overview:

- SIGHT
- HPC
- Dleware

GPGPU & CUDA

- GPGPU_Intro
- CUDA-Basics
- Setup



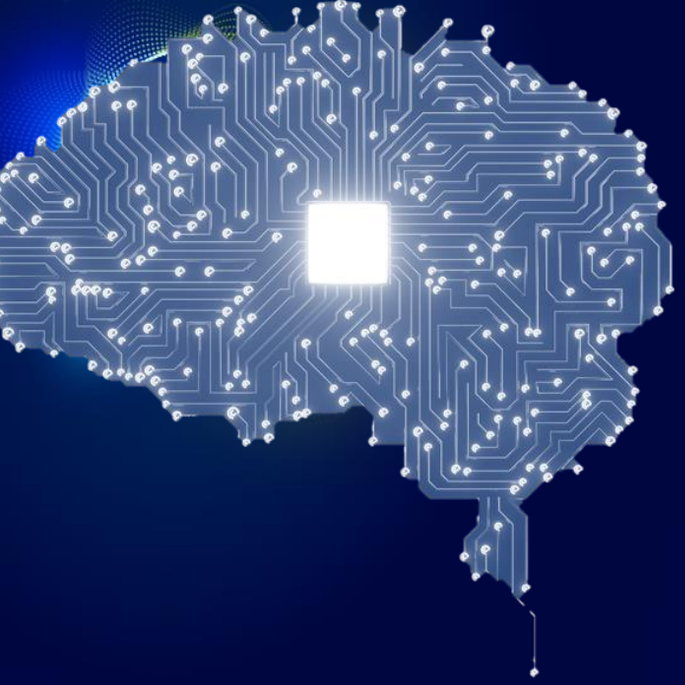
נושאים

CUDA (Numba Python based)

- numba-cuda-kernels.ipynb – Part#1 (Overview)
- numba-introduction.ipynb
    - Exe#1 – Device-Host performance comparison
    - Exe#2 – Host memory and device memory
    - Exe#3 – Vector addition fp32bits

- numba-cuda-kernels.ipynb – Part#2 (Practice)
    - Exe#1 – Device arrays
    - Exe#2 – Histogram
    - Exe#3 –
      Convert colored input image to grayscale
      numba_cuda_kernels-focus-on-Multi-dimensional-grids.ipynb

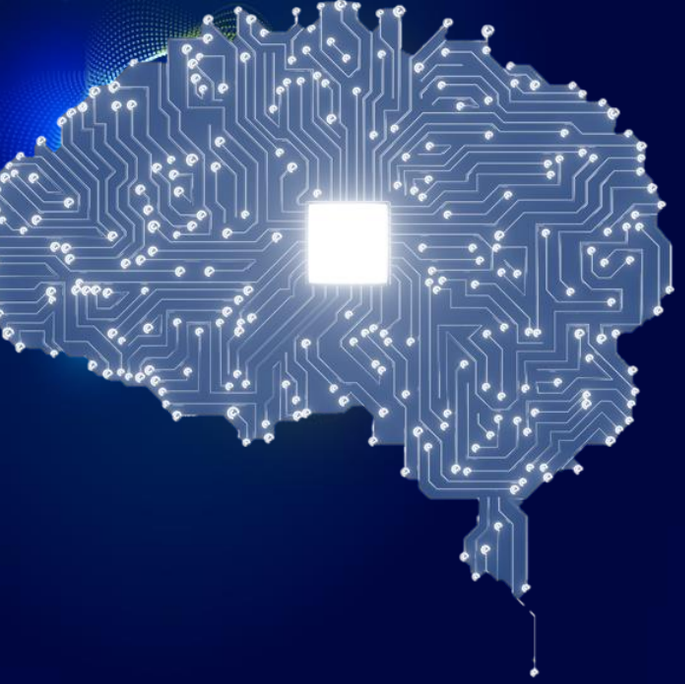
- Convolution exercise



נושאים

Deploy pre trained DL model (focus on CNN)

- C++:
  - VS (Cross platform)
  - Cmake
- Python - PyCharm IDE based
- PyTorch – Python\C++
  - https://pytorch.org/tutorials/advanced/super_resolution_with_onnxruntime.html
- Onnx – (Focused on PyTorch)
  - Get started –
    - https://github.com/onnx/onnx
    - https://onnx.ai/get-started.html
  - Deploy –
    - https://onnx.ai/supported-tools.html#deployModel
    - https://github.com/onnx/tutorials#converting-to-onnx-format
    - https://github.com/onnx/optimizer
    - https://github.com/onnx/tutorials/blob/main/tutorials/PytorchOnnxExport.ipynb
    - onnx-model-convertion.ipynb – Part#1



נושאים

Deploy pre trained DL model (focus on CNN)

- TensorRT – Pyhon\C++
  - NVIDIA_TensorRT
  - SDK
    - Overview
    - Samples (Python based)
      - onnx_resnet50.py
      - yolov3_onnx
        - yolov3_to_onnx.py
        - onnx_to_tensorrt.py
        - onnx_packnet (graph surgeon)
        - engine_refit_onnx_bidaf
  - Notebooks:
    - onnx-model-convertion.ipynb – Part#2
    - Exe#1 - exercise-tensorrt-onnx-conversion-solution.ipynb
  - OSS (Advanced)
    - Polygraphy
    - Trex

- OpenVINO (optional) – C++

מתחילים!