HAML

Heterogenous and Accelerated Computing for Machine Learning

Advisor : Dr. Philip Jones
Client : JR Spidell

Justin Wenzel, Jonathan Tan, Kai Heng Gan, Josh Czarniak, Santiago Campoverde

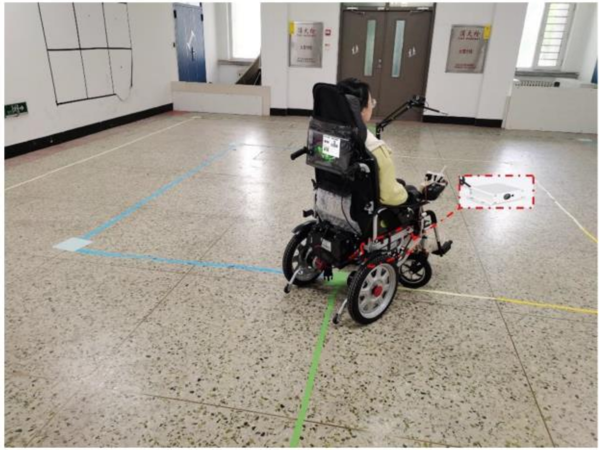
IOWA STATE UNIVERSITY

Problem Statement

- Client wants to create a system to help people with disabilities complete day-to-day activities by tracking eye movement.
  - Using pupil movement to control mouse cursor
  - Prediction of the user's state (predicting seizures, stress, fatigue, etc.)
- Three different algorithms crucial for the success of this goal:
  - Blink Detection
  - Pupil Tracking
  - Image Preprocessing
- Our project is a subcomponent of a larger ML powered wheelchair system.

Market Survey

- Obstacles Detection for Electric Wheelchair with Computer Vision [1]
  - Wheelchair system that uses computer vision to detect obstacle in a house environment.
  - Proof of concept, did not build custom hardware for the research.
- Eye Gaze Controlled Wheelchair Based on Deep Learning [2]
  - Trained GazeNet using 135k annotated images.
  - High complexity, poor reliability, and not totally realistic.
- Eyeware Webcam Eye Tracker [3]
  - Eye movement tracking software.

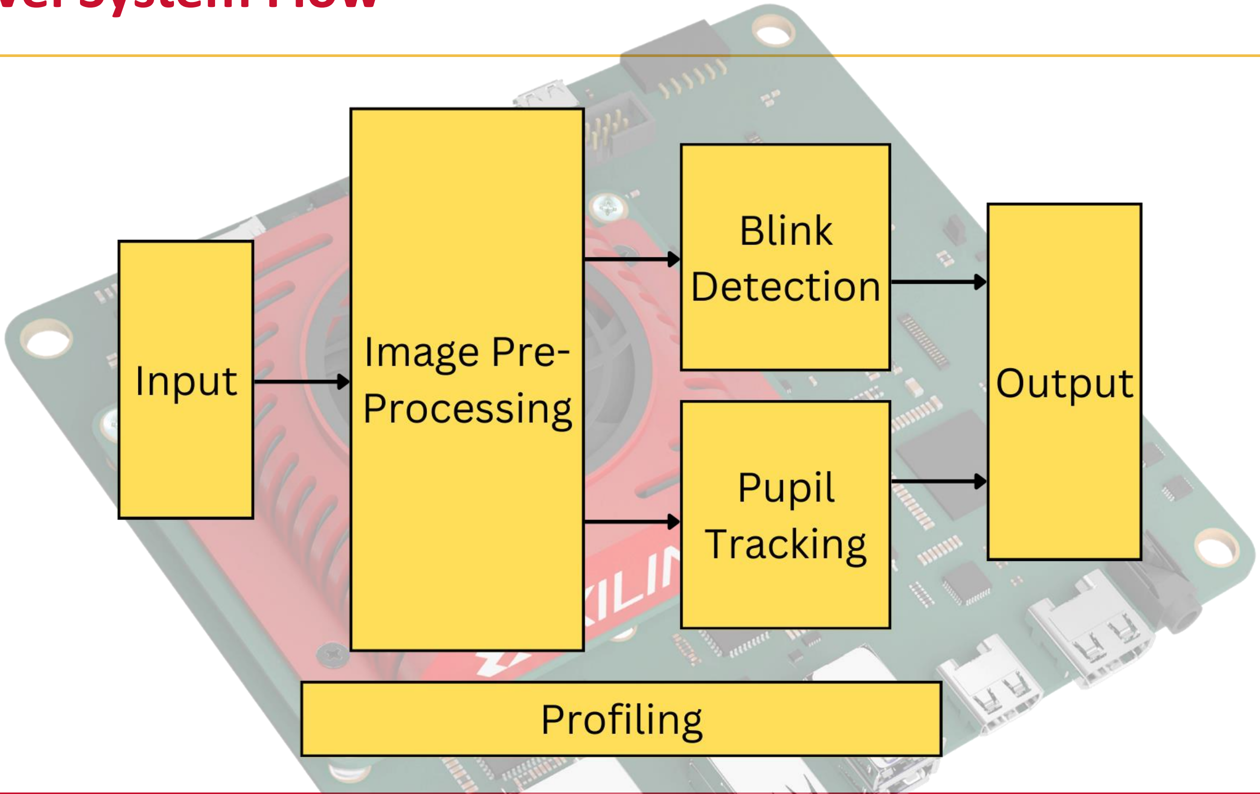


[1] "Obstacles Detection for Electric Wheelchair With Computer Vision." *IEEE Conference Publication | IEEE Xplore*, 26 Jan. 2022, ieeexplore.ieee.org/document/9729083.

[2] Xu, Jingwei, et al. "Eye-Gaze Controlled Wheelchair Based on Deep Learning." Sensors, vol. 23, no. 13, July 2023, p. 6239. https://doi.org/10.3390/s23136239.

[3] Eyeware Beam. "Beam Eye Tracker - Turn Your Webcam Into an Eye Tracker." Beam Eye Tracker, 15 Apr. 2024, beam.eyeware.tech.

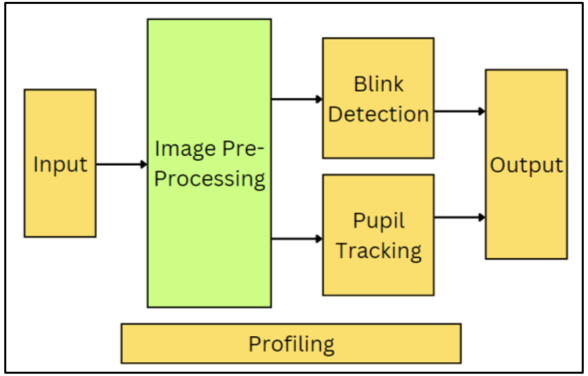
High Level System Flow



Functional Requirement and Constrains

- For some of the goal to work correctly (like emotion tracking), the system must achieve:
  - Throughput: 200 Frames Per Second
- Constraints:
  - Client provides ML models
  - Client wants it implemented on the Kria 260 eval board
  - Client wants it in multithreaded system

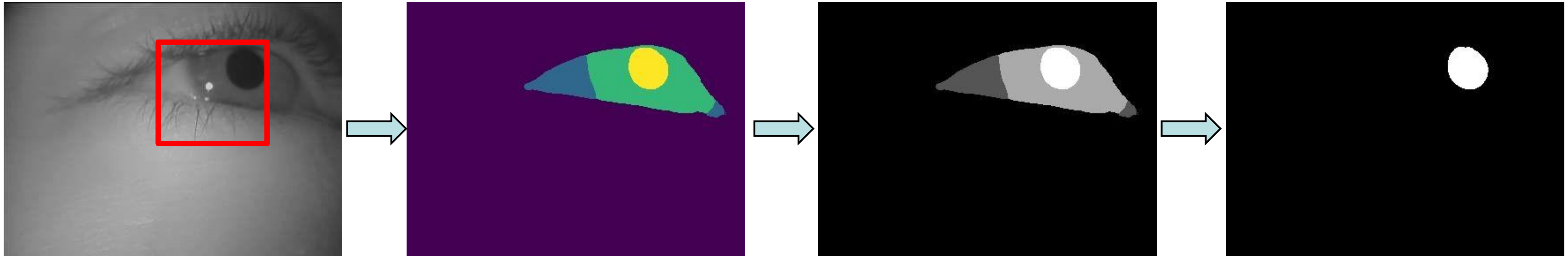
Image Pre-processing



Technique: Semantic Segmentation

Input: 1-channel image (grayscale)

Output: 1-channel image (black and white segmented image)



Blink Detection

Input:

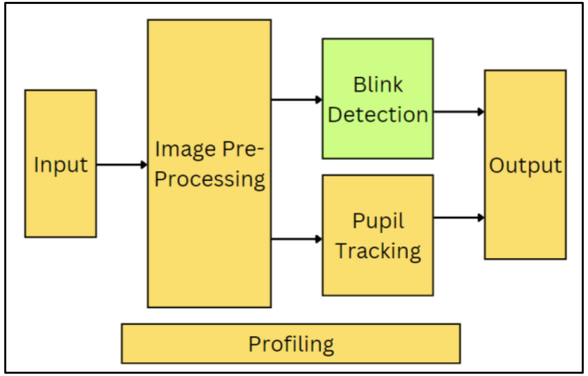
- Frames extracted from a given video

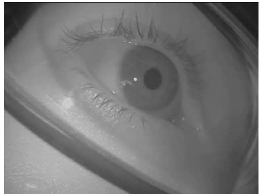
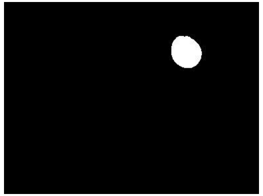
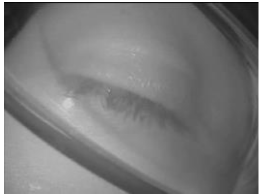
Model Type:

- Classification model

Output:

- Two classes of classification
  - Blink: "Frames contains a blink"
  - No Blink: "Frame does not contain a blink"
- Neural Network outputs the probability that a video frame is a blink or no blink



	Original	Preprocessed
No blink		
Blink		

Pupil Tracking

Input:

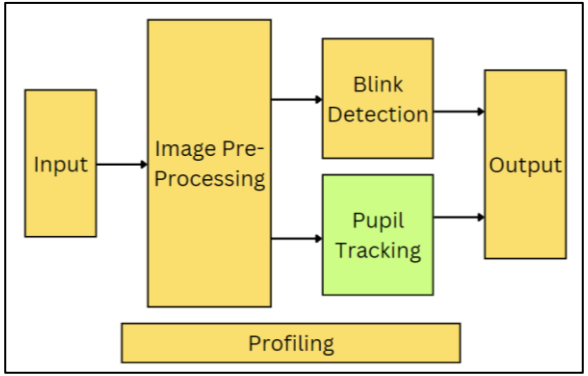
- Frames extracted from a given video

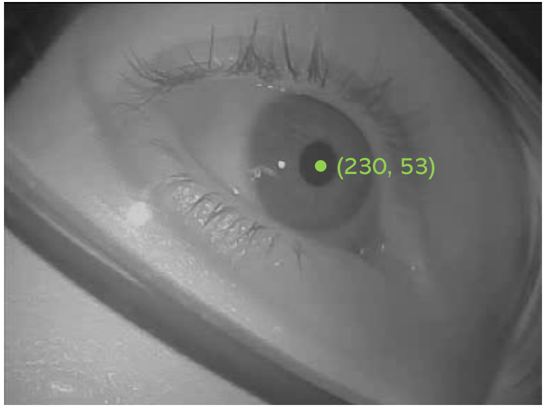
Model Type :

- Regression model

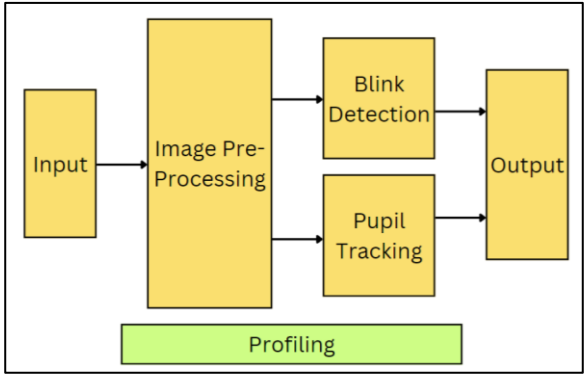
Outputs:

- Returns the location of the pupil within the image frame
- Given in X and Y coordinates using pixel measurements
- Slower run time than blink algorithm





Vitis AI Profiling



Input:

- A data tracer (VAI Trace) is run alongside the model and performance details are gathered
    - CPU/DPU compute units and AI inference throughput data

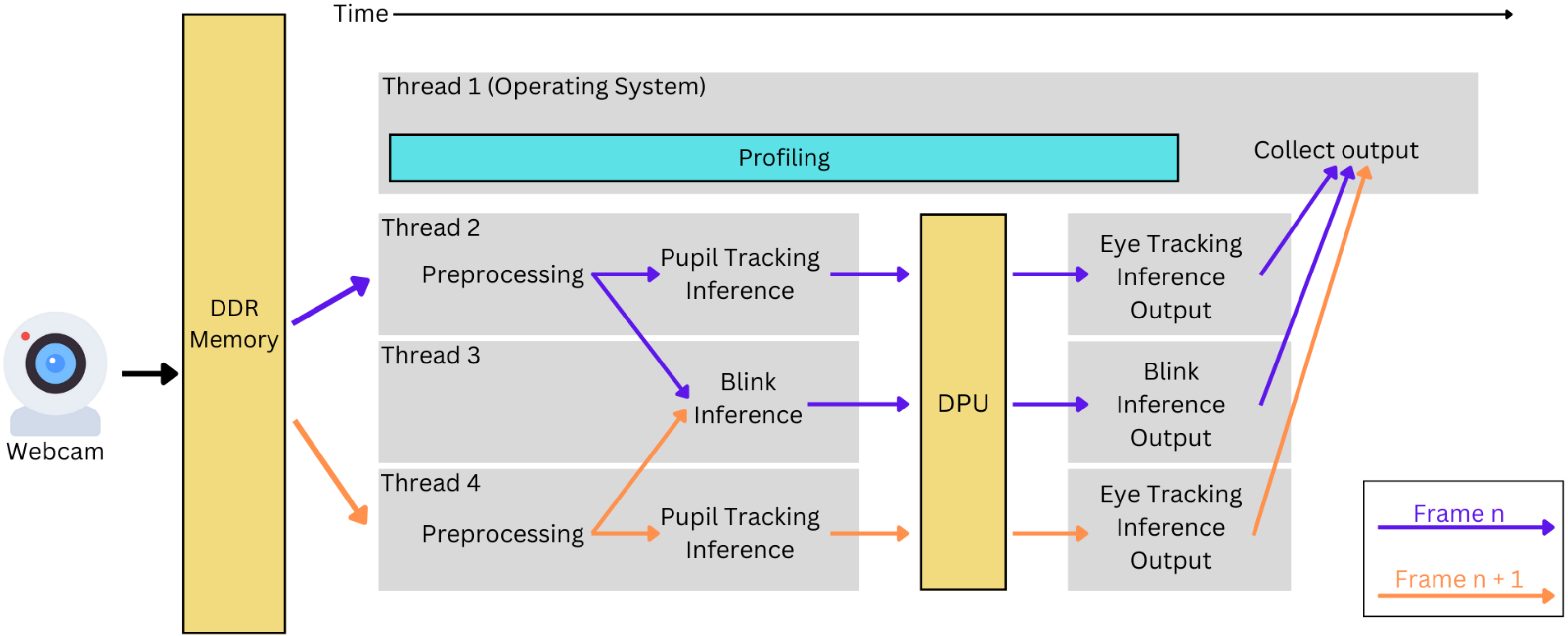
Output:

- A summary file for the data to graphically represent it on Vitis Analyzer

Goals:

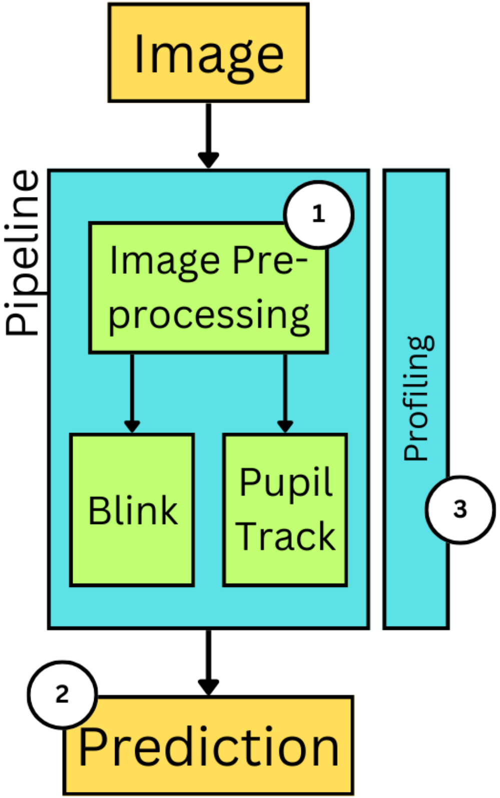
- Help team identify opportunities for performance optimizations
- Help current and future teams to compare performances for new implementations

Putting Everything Together - Multithreaded Conceptual Sketch



Test Plan

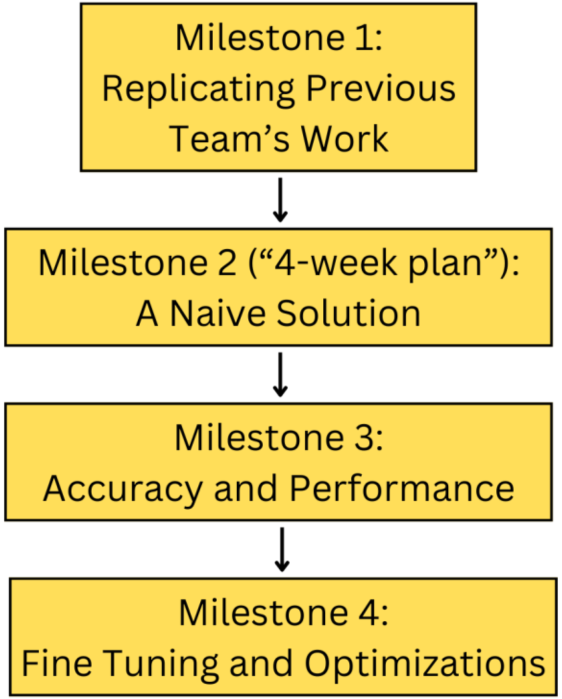
1. Pre-processing Accuracy
   - Metric: Mean intersection over union (IOU=0.9)
2. Blink and Pupil Tracking Accuracy
   - Metric:
     - Blink: Confusion Matrix
     - Pupil Tracking: Root Mean Squared Error
   - Visual Representation:
     - Compare prediction and ground truth
     - Colored dot overlay technique
3. Throughput
   - Ensure throughput of 200 Frames per Second



Potential Risk & Mitigations

Risks	Mitigations
Teammate unavailability due to personal circumstances	Regular information sharing and comprehensive documentation
Difficulty in debugging and testing	Incremental testing
Asynchronization among threads	Development of a thread synchronization control mechanism
Model doesn’t achieve accuracy and runtime expectation	Continuous evaluation of algorithmic options and collaborative partnerships

Project Milestone & Schedules



Milestone 1: Replicating Previous Team’s Work

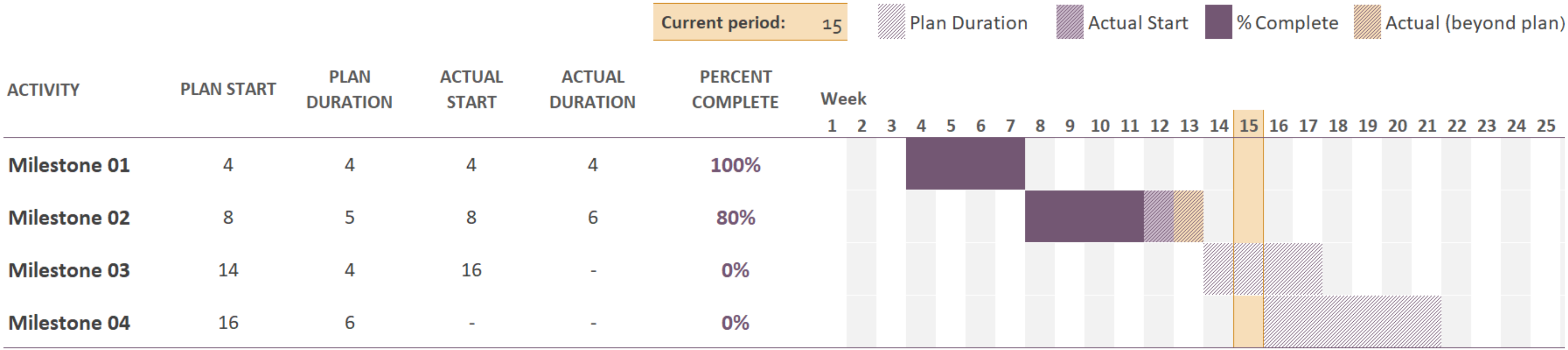
Milestone 2 (“4-week plan”): A Naive Solution

Milestone 3: Accuracy and Performance

Milestone 4: Fine Tuning and Optimizations

- Milestone 2:
  - Implementing a single threaded approach
  - Encourage team to identify issues and seek help
  - Identify bottlenecks and overheads through profiling
- Milestone 3:
  - Implement a multithreaded approach
  - Measure model accuracy
  - Profile hardware resources
- Milestone 4:
  - Optimization of implementation

Current Project Progress



Current period: 15

▨ Plan Duration ▨ Actual Start ■ % Complete ▨ Actual (beyond plan)

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Milestone 01	4	4	4	4	100%
Milestone 02	8	5	8	6	80%
Milestone 03	14	4	16	-	0%
Milestone 04	16	6	-	-	0%

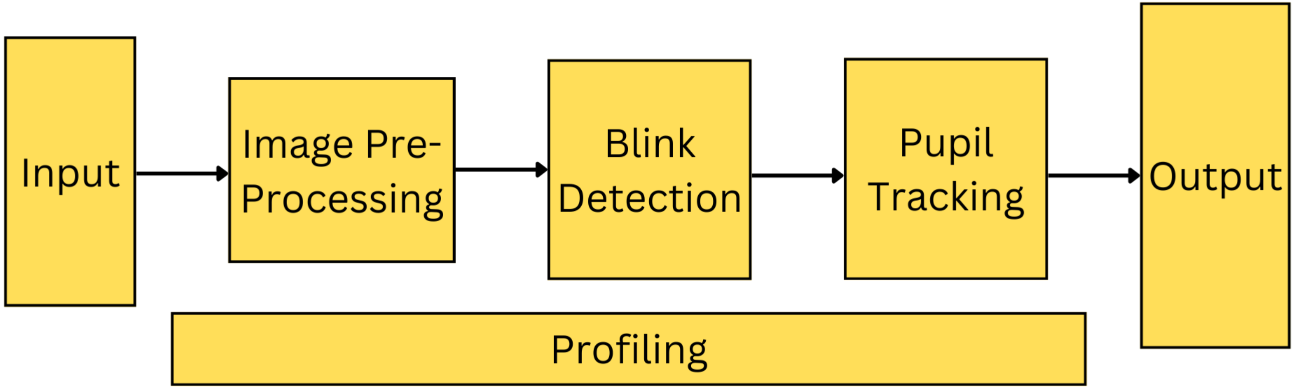
Agile – Flexibility and Adaptability

- Our project has a lot of different solutions, being flexible allows us to reevaluate our implementations/decisions.
- E.g., Unexpected delays in milestone 2, redefinition for milestone 3 needed.

Milestone 2 – A Naïve Solution

A Naïve Solution where we run different algorithms in series (instead of in parallel)

- Throughput: ~16.1 Frames Per Second (without image pre-processing)



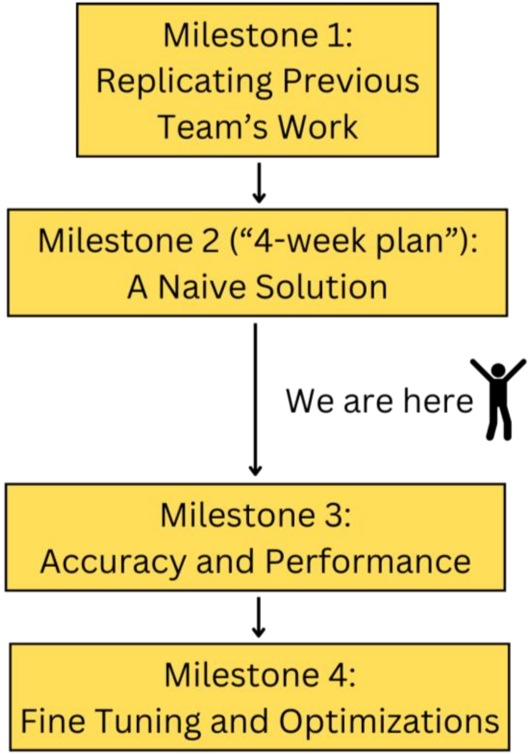
Conclusion

Achievements for Semester 1:

- Successfully implemented all three algorithms on the Kria board
- Successfully ran one inference through all algorithms

Moving forward:

- Redefining milestone 3's goals according what we learned in milestone 2


Milestone 1:
Replicating Previous Team's Work
Milestone 2 ("4-week plan"):
A Naive Solution
We are here
Milestone 3:
Accuracy and Performance
Milestone 4:
Fine Tuning and Optimizations

Questions?

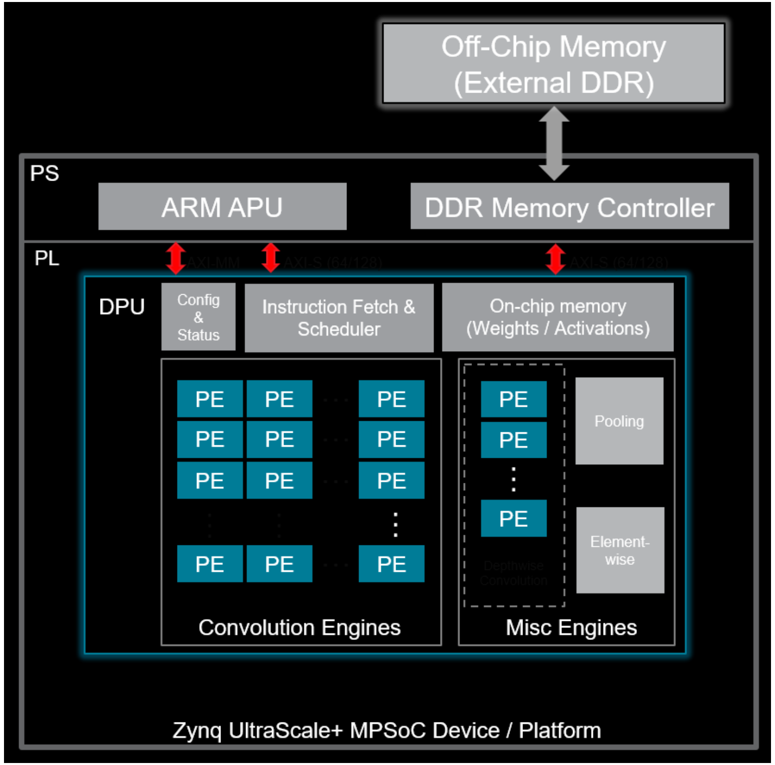
Questions

1. What is the DPU?
2. What is BRAM?
3. Project cyber security implications
4. What is the point of the blink model then?
5. What is mean intersection over union?
6. What is milestone 2?
7. What are some potential improvements?
8. How would multithreading help us hit the 200FPS target?
9. What is the latency target for the system?
10. Broader Context
11. RMSE
12. Confusion Matrix
13. Why 200FPS?

What is the DPU?

Deep-learning **P**rocessing **U**nit

- A programmable engine for convolutional neural network
- An IP block in Vivado, ours is the B4096 architecture

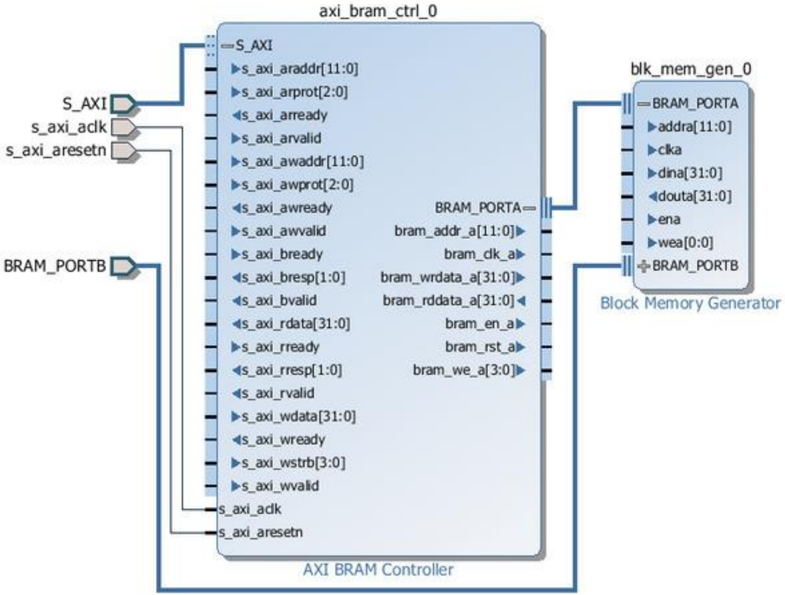


“DPU For Convolutional Neural Network.” AMD, www.xilinx.com/products/intellectual-property/dpu.html.

Return to Questions

What is BRAM and why is it important?

- **B**lock **R**andom **A**ccess **M**emory
- A type of memory used in FPGAs
- High-speed data access
- In our project, it is used by the DPU



AMD. support.xilinx.com/s/question/0D52E00006hpZsESAU/axi-bram-controller-unable-to-change-address-to-least-significant-bits?language=en_US.

Return to Questions

Cyber security implications?

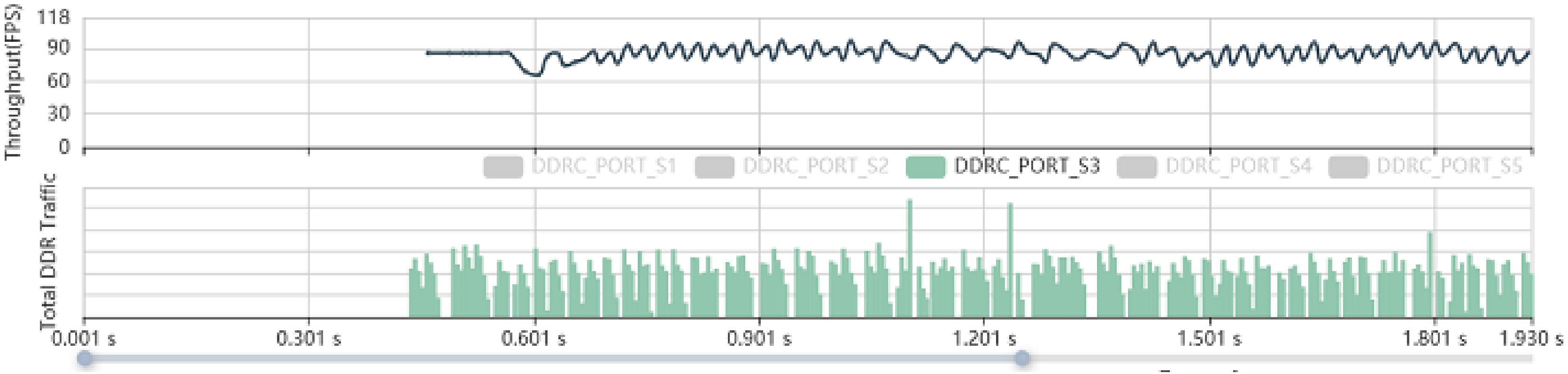
Although the system will mostly be used “off the grid,” some security implications are worth paying attention to:

- Intrusion during firmware update
- Install malware or ransomware on the board accidentally
- Physical security risks: Attacker installing malware physically

Return to Questions

How to identify issues through Vitis Analyzer?

- Data will be analyzed by looking at the trends represented by the graphs



From: https://docs.amd.com/r/1.2-English/ug1414-vitis-ai/What-Information-Can-Be-Obtained-from-This-Tool

Return to Questions

If input is reduced to black & white segments, what is the point of the blink model?

- Image preprocessing (semantic segmentation) isn't perfect
- Image preprocessing might take too long and other techniques might need to be used instead
- Blink model handed to us by previous teams uses unprocessed image as input

Return to Questions

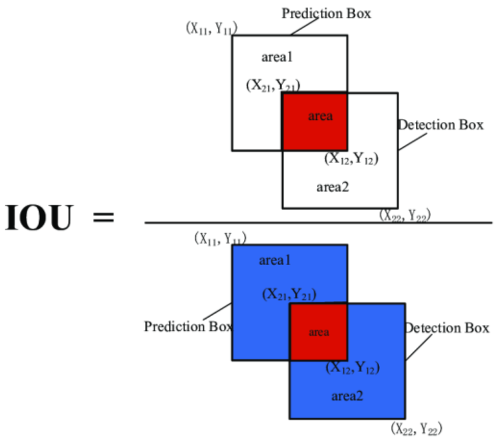
What is mean intersection over union (IoU)?

Mean Intersection over Union (IoU):

- Average IoU across all classes in multi-class segmentation tasks.
- Reflects overall segmentation accuracy, normalized between 0 (no overlap) and 1 (perfect overlap).

Advantages of mIoU:

- Normalizes performance across imbalanced class sizes.
- Widely used metric for comparing model effectiveness in benchmarks.



Return to Questions

What are some potential improvements?

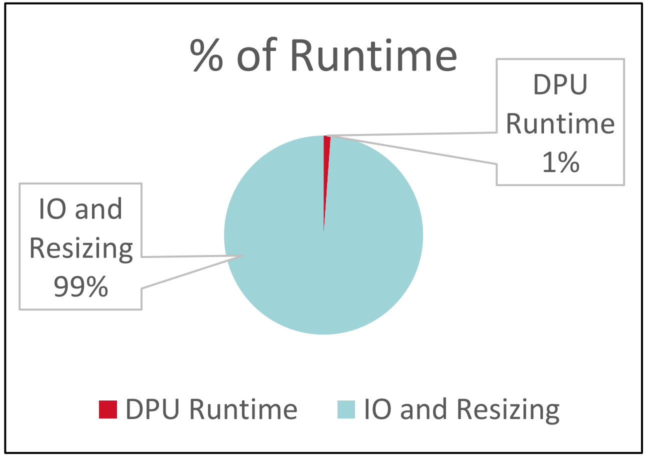
Throughput of Naïve solution: 16FPS

- Single threaded
- Redundant file read (reading same input twice)
- Inefficient image resizing technique
- Using Amdahl's Law, if we remove redundant file read:

$$Speedup = \frac{T_{org}}{\left((1-f)+\frac{f}{a}\right)\times T_{org}} = \frac{1}{(1-f)+\frac{f}{a}}$$

$$= \frac{1}{(1-0.99)+\frac{0.99}{2}} = 1.98$$

- Assumption: Both Blink and Pupil Tracking algorithm perform same IO read and resizing.



What are some potential improvements? (cont.)

Throughput of Naïve solution: 16FPS

- Resizing is inefficient, and not the focus of our project, looking into other techniques like:
    - Cropping, instead of resizing
    - Ensure correct resolution at other components (e.g., at the camera)

Return to Questions

How would multithreading help us hit the 200FPS target?

Throughput of Naïve solution: 16FPS

Making the basic assumption of running 2 frames in parallel = 2x improvements:

We can expect our application runtime to 2x.

What is the latency target for the system?

- Latency isn’t a concern for our project, throughput is.
- Latency target for the system is $< 50ms$.
- So, the idea is to not exceed $50ms$.

Return to Questions

Broader Context

Area	Description
Public health, safety, and welfare	• Bring an accessibility solution to people with disabilities.
Environmental	• ML application can be energy intensive. • While meeting functional target is important, we don't rule out optimizing resources used by our system.
Global	• Opportunities for a machine vision application is endless, allowing huge improvements to human's day-to-day life.
Economic	• The economic factors involve the cost of deployment, development, and the potential market demand which will influence future production costs and scale.

Return to Questions

RMSE

RMSE calculates the square root of the average squared differences between the predicted and actual values.

- Square to make sure all numbers are positive, and errors are bigger
- Add all values up and then divide by the number of predictions
- Square root the number to bring it back into the original measurement range

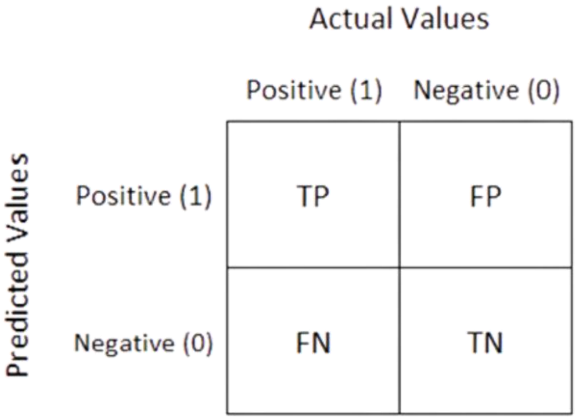
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- n = number of observations
- y_i = observed values
- $\hat{y}_i$ = predicted values

Return to Questions

Confusion Matrix

- A binary classification Confusion Matrix utilizes a 2x2 table
  - **True Positive (TP)**: The number of cases correctly predicted as positive.
  - **True Negative (TN)**: The number of cases correctly predicted as negative.
  - **False Positive (FP)**: The number of cases incorrectly predicted as positive (also known as Type I error).
  - **False Negative (FN)**: The number of cases incorrectly predicted as negative (also known as Type II error).



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- Accuracy equation: $\frac{(TP+TN)}{(TP+FP+FN+TN)}$

Return to Questions

Why 200 FPS?

- 200 FPS is required for capturing an individual’s state using the movement of the eye
- It is a requirement of our project

Return to Questions