HALO: High Autonomous Low-SWaP Operations

Team Members:

- Sloan Hatter shatter2022@my.fit.edu
- Blake Gisclair bgisclair2022@my.fit.edu

Faculty Advisor: Dr. Ryan T White rwhite@fit.edu

Client: Dr. Ryan T White rwhite@fit.edu

Progress Matrix of Milestone 3

Task	Completion %	Sloan	To Do
Hardware Swap	100%	Switch out the Raspberry Pi 5 AI HAT+ for the Jetson AGX Orin	None
Post Processing Scripts for Raspberry Pi	NULL	NULL	NULL
Post Processing Scripts for Jetson	90%	Modify	Modify available post processing scripts if needed.
8-bit Representation	50%	Implement	Implement the INT8 quantization methods.

Discussion of Accomplished Tasks for Milestone 3:

- **Hardware Swap**
  - Originally, HALO was being run on a Raspberry Pi 5 AI HAT+. I foresaw that this computer might not be able to support the level of computation required for lower quantization implementations of the neural network. Working through Milestone 3, I came to the realization that I would have to swap out the hardware sooner than I originally anticipated. The Raspberry Pi simply did not have the proper hardware capabilities to support the lower quantization implementations. Working with the Pi proved more difficult than it was worth, so the decision was made to switch to the Jetson.
- **Post-Processing Scripts for Raspberry Pi 5 AT HAT+**
  - Since I switched hardware from the Raspberry Pi to the Jetson, there was no point in continuing to work on the post processing scripts for the Raspberry Pi, as they would not transfer to the Jetson.
- **Post-Processing Scripts for Jetson AGX Orin**
  - The NVIDIA JetPack SDK has available resources that include post processing scripts for specific tasks. I will most likely use the DeepStream SDK library, as it provides a framework for video streaming analysis and includes pre-built

components for bounding box filtering, tracking, and metadata handling. I will only need to write my own scripts, or modify the available scripts, if I find that the frameworks from DeepStream SDK do not provide the proper analysis I require or if I want to optimize HALO's performance.

- **8-bit Representation**
    - Achieving 8-bit representation should not prove to be too difficult as the NVIDIA JetPack SDK package has ample support and available libraries for 8-bit representation through the NVIDIA TensorRT. The package offers two primary methods for implementing 8-bit quantization, including Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). I will be using the PTQ method, as I will have the network trained in 32-bit first and then will convert it to INT8. Although, QAT typically yields higher accuracy, I will not be keeping the 8-bit model.

Discussion of Contribution to Milestone 3:

- **Sloan Hatter:** Tasks contributed to this milestone include switching out the hardware from using a Raspberry Pi AT HAT+ to using a Jetson AGX Orin.

Task Matrix for Milestone 4:

Task	**Sloan**
4-bit Representation	100%
Tune 4-bit Model	100%
Research algorithms for binary quantization	100%

Discussion of Planned Tasks for Milestone 4:

- 4-bit Representation
    - The package that I am using for running HALO on the Jetson AGX Orin is the NVIDIA JetPack SDK. The primary software package for enabling 4-bit neural network quantization on NVIDIA Jetson devices is the NVIDIA TensorRT. This library is the core component for high performance neural network inference on NVIDIA GPUs. It should not be too difficult to represent HALO in 4-bits, being that there are available packages ready for fast deployment from NVIDIA. Representing HALO in 4-bits will get me closer to a 1-bit representation and will allow me to quantize the weights and activations at a closer scale.
- Tune 4-bit Model
    - After I implement the 4-bit model through the NVIDIA TensorRT package, I will have to implement techniques given from the package to tune the model for optimal performance and accuracy. The NVIDIA TensorRT Model Optimizer provides the following tuning techniques: Post-Training Quantization (PTQ), Quantization-Aware Training (OAT), Weight-Only Quantization (WoQ), and Automatic Kernel Selection and Optimization. The Quantized-Aware Training method will most likely be chosen as OAT provides the best balance of performance and accuracy for lower precisions.
- Research algorithms for binary quantization

    - Binary quantization converts each dimension of a vector into a single bit; this is the main goal of HALO: to represent the neural network through 1-bit quantization. The main disadvantage of 1-bit representation, and the main trouble I will encounter, is the loss of accuracy that comes with binary quantization. There are algorithms out there that aim to improve accuracy through incorporating methods such as Quantization Aware Training (QAT), Knowledge Distillation (KD), the use of scaling factors, and specialized algorithms like rescoring and mixed precision. Once I get the model down to 4-bit and tune it, I will be able to test different binary quantization algorithms and pick the one that fits the model the best.

Date of Meetings:

- 11/05/25
- 11/19/25

Client Feedback on Milestone 3:

See Faculty Advisor Feedback below.

Faculty Advisor Feedback on Milestone 3:

- Hardware Swap:
- Post Processing Scripts for Jetson:
- 8-bit Representation:

Faculty Advisor Signature: _____ Date: 24 Nov 2025