Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

LAIDLAW SCHOLARS

Investigating the Use of FPGAs for Object Detection in High-Speed Autonomous Navigation

Emran Yasser Moustafa, Supervisor: Dr. Shreejith Shanker
Department of Electronic & Electrical Engineering, Funded by the Laidlaw Foundation

Background

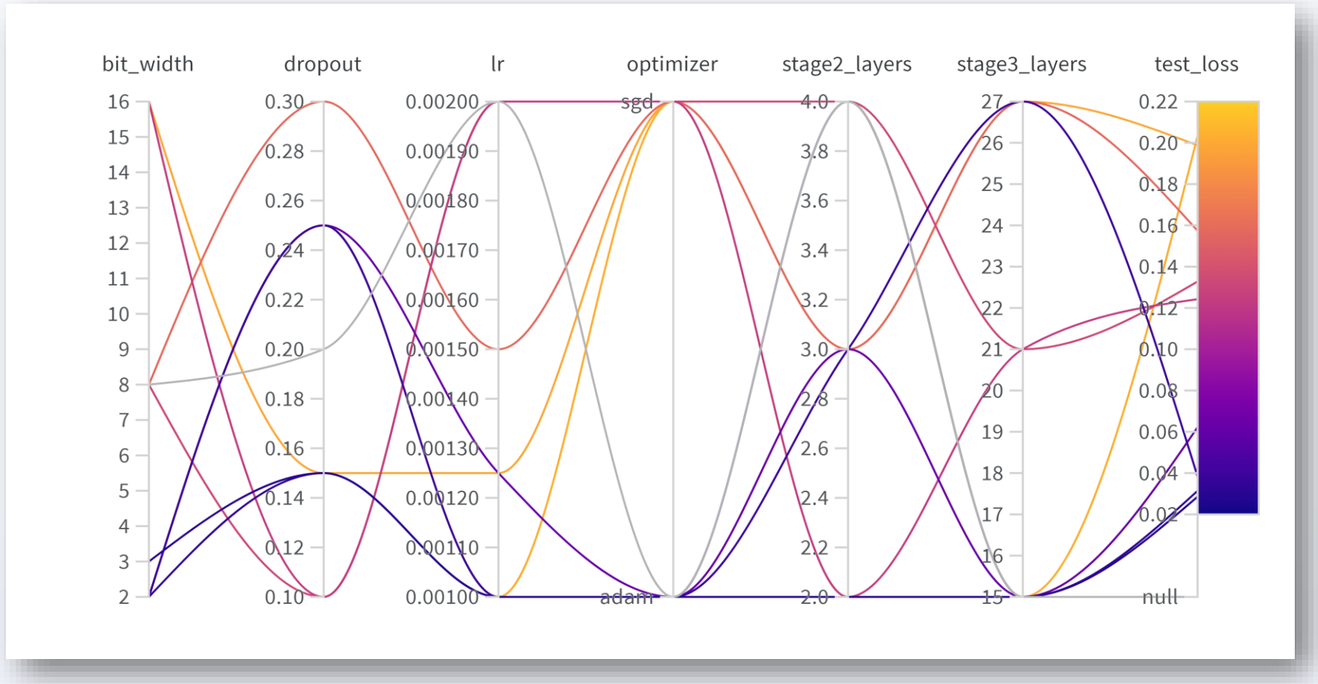
Perception is an essential component of many autonomous navigation systems. Observations of objects and obstacles are made using a range of sensors, from stereo cameras to LiDAR sensors. These observations are often made using an AI model known as a deep neural network!
However, one issue engineers often have to face is the high energy consumption of using these models. For quick computation, these models are often run on GPUs. This greatly limits the range and sustainability of many lightweight autonomous vehicles.

Neural Network

Segmentation is a method of object detection where each pixel of an image is assigned a “class”. This creates a mask of the image, locating the exact position of different objects in view of the camera. This is an essential method of computer vision for many autonomous vehicles.
For my project, I used the CGNet (1) architecture for my model. CGNet is a lightweight and high accuracy segmentation model, roughly 500-700 thousand parameters.

Deployment

Xilinx’s Kria SoM (KV260) was the FPGA I used for my project. The KV260 is built to be used for accelerated computer vision applications, making it perfect for this project.
VitisAI was also used to quantise and compile the model before deployment on the board. This tool came with some limitations, such as unsupported layers and kernel sizes.
The model architecture was changed to be made compatible with the VitisAI library.



	Latency (s)	Power Consumption (w)	Accuracy (% IoU)	Energy Consumption (J)
GPU (GTX 1080 Ti)	0.0144	48.0154	0.7084	0.6914
FPGA (KV260)	0.0741	9.3877	0.7194	0.6956

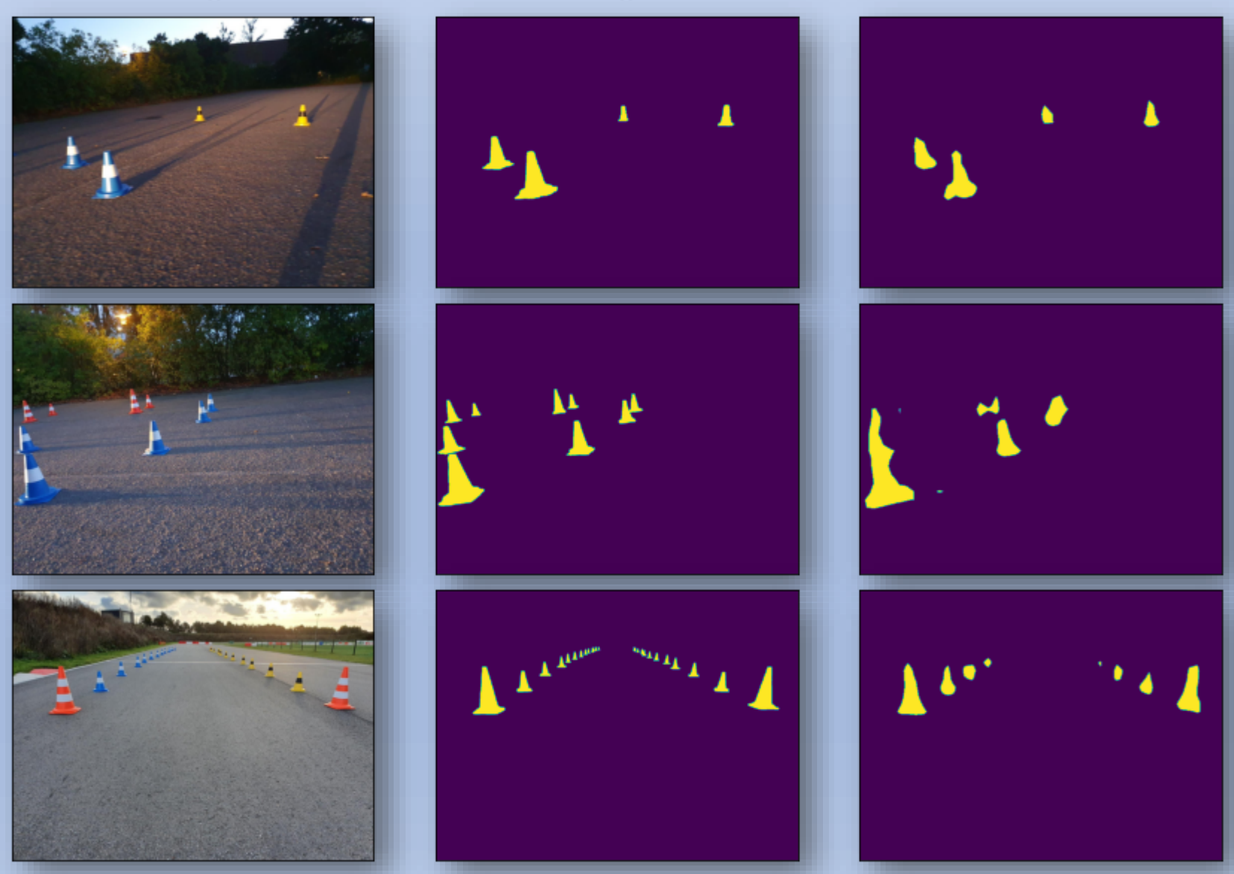
Results

As you can see above the FPGA had a similar power consumption to the GTX 1080 Ti. This may come as a surprise, however this issue can be remedied in a number of different ways.
If we look at the masks created we also gain an understanding of the model performance. Due to a resolution drop in the image in the initial stages of the model cones a great distance from the camera are undetected. Similarly, cones that are overlapping are joined together in some cases.

Where to go from here?

This project shows the viability of the use FPGAs in high-speed autonomous vehicles.
The latency of the model can be improved, which would in turn reduce the energy consumption per inference. As seen in the sweep diagram, a more conservative architecture can actually yield greater performance.
The issues of information being lost in the initial stage of the model can be rectified by replacing the pooling stages at the start of the network with a series of height-wise and width-wise convolutional layers (2).

Image | Target | Mask



References
(1) Tianyi Wu, Sheng Tang, Rui Zhang, Yongdong Zhang, 2018, CGNet: A Light-weight Context Guided Network for Semantic Segmentation, arXiv:1811.08201
(2) Lukasz Kaiser, Aidan N. Gomez, Francois Chollet, 2017, Depthwise Separable Convolutions for Neural Machine Translation, arXiv:1706.03059