National University of Sciences and Technology

School of Electrical Engineering and Computer Sciences

Deploying Vision Transformers on Embedded Devices

Problem Statement

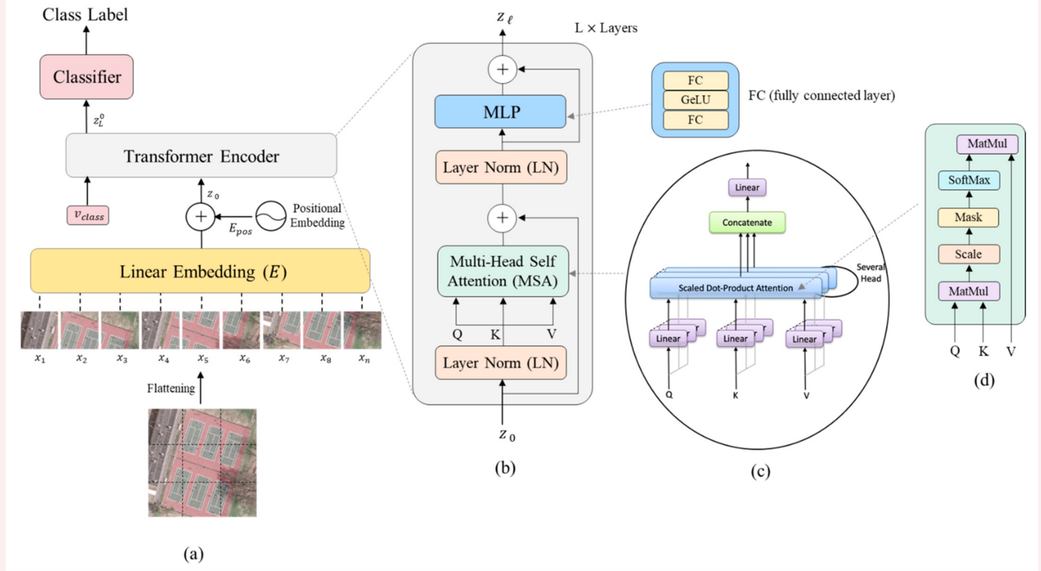
The advent of Vision Transformers (ViTs) has advanced accuracy in computer vision tasks. However, their intricate architecture, substantial computational and storage demands, large number of parameters, complex design (mainly attention module), and notable low throughput frequently lead to significantly slower performance when compared with lightweight convolutional networks. This highlights a compelling requirement for pioneering hardware accelerator design methodologies for Vision Transformer that can achieve high throughput while maintaining accuracy.

Methodology

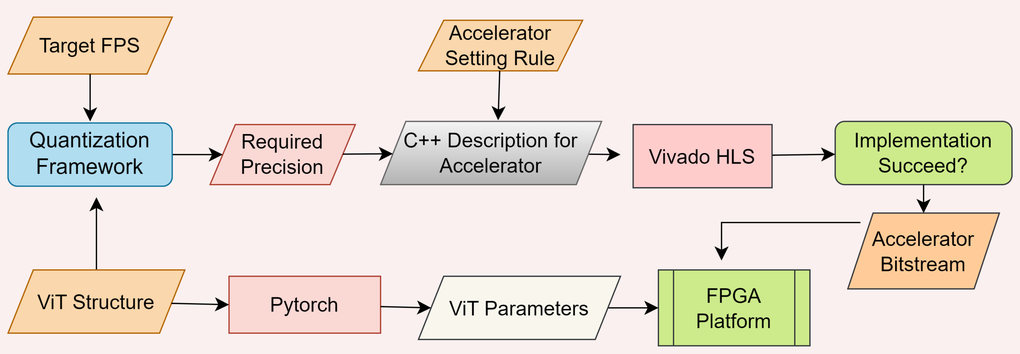
Our proposed solution is a FPGA-aware automatic acceleration framework for Vision Transformers (ViTs). We utilize a mixed-scheme quantization approach and High-Level Synthesis (HLS) to enhance the deployment of ViT models on FPGA. This involves reducing model size through quantization while preserving accuracy and converting the quantized model into FPGA-compatible C/C++ code using the Vitis HLS tool. The result is an efficient, high-throughput ViT accelerator suitable for real-time applications on resource-constrained edge devices.

Tools and Acknowledgments

ViT Architecture



Work Flow Structure



Challenges

1. Implementation of multiple modules not available in Brevitas and Finn HLS.
2. Model compression while maintaining the best accuracy and finding optimum bit-width for each layer.
3. Finding alternatives for some operations that are predominantly sequential, which can impede pipeline efficiency

Conclusion

Vision transformers on FPGA will revolutionize computer vision by enabling real-time, energy-efficient image analysis, transforming industries like autonomous vehicles and healthcare, and fostering innovation and breakthroughs.

FYP Supervisor: Dr Faisal Shafait Co-Advisor: Dr Adnan-ul-Hassan

Team Members: Talha Israr, Rakhmeen Gul, Muhammad Hamza Javaid