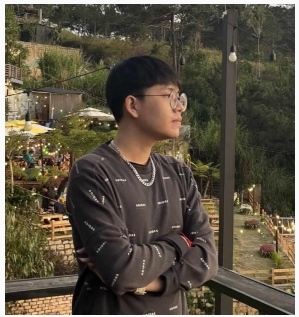
NGHIÊN CỨU VÀ THIẾT KẾ PHẦN CỨNG CHO CONVOLUTIONAL NEURAL NETWORK

Trần Quốc Lượng - 20521590
Võ Phan Hoàng Kha - 20521428

Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/KhaVo12/CS519.O11
- Link YouTube video: https://youtu.be/LvsbiyFaVxE
- Ảnh + Họ và Tên của các thành viên:

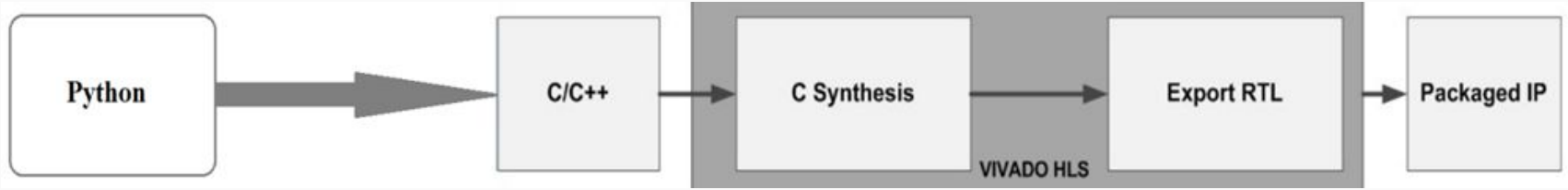


Trần Quốc Lượng

Võ Phan Hoàng Kha

Giới thiệu

- Cải thiện về năng lượng tiêu thụ và hiệu năng trong các hệ thống cần độ chính xác cao.
- Khả năng triển khai các mô hình lên những hệ thống nhúng có diện tích nhỏ nhưng vẫn đảm bảo về tốc độ xử lý và độ chính xác .



Mục tiêu

- Xây dựng thành công một mạng CNN với mô hình Lenet5 bằng ngôn ngữ Python và C/C++.
- Thiết kế được IP (Intellectual Property) của mô hình bằng công cụ tổng hợp cấp cao.
- Phát triển mô hình từ nhận diện chữ số viết tay lên nhận diện các vật thể khác.

Nội dung và Phương pháp

- Nội dung
  - Nghiên cứu về cách hoạt động của CNN và mô hình Lenet5, từ đó bắt đầu xây dựng trên ngôn ngữ Python.
  - Từ những nghiên cứu về CNN, bắt đầu mô phỏng lại mô hình bằng ngôn ngữ C/C++
  - Tiến hành chuyển đổi mô hình sang ngôn ngữ HDL với công cụ Vivado HLS, kèm theo các chỉ thị tối ưu để phù hợp khi chuyển sang phần cứng.

Nội dung và Phương pháp

- Phương pháp
  - Sử dụng các tập dữ liệu có sẵn và các mô hình phổ biến cho việc học của mô hình.
  - Đánh giá độ chính xác của từng mô hình cho từng tập dữ liệu tương ứng.
  - Cầu nối giữa phần mềm và phần cứng - Vivado HLS.

Kết quả dự kiến

- Mô hình CNN được triển khai bằng ngôn ngữ Python và C/C++ hoạt động chính xác như mong đợi, có khả năng dự đoán chính xác đầu vào là tập dữ liệu MNIST với các chữ số viết tay.
- Mô hình từ phần mềm hiện thực thành công sang phần cứng, cho ra hệ thống có khả năng và độ chính xác tương tự như phần mềm.

Tài liệu tham khảo

[1]. Yongming Shen, Michael Ferdman, Peter Milder: Maximizing CNN Accelerator Efficiency Through Resource Partitioning. ISCA ’17, 2017.

[2]. Jiantao Qiu, Jie Wang, Song Yao, Kaiyuan Guo, Boxun Li, Erjin Zhuo, Jincheng Yu, Tianqi Tang, Ningyi Xu, Sen Song, Yu Wang, Huazhong Yang: Going Deeper with Embedded FPGA Platform for Convolutional Neural Network. FPGA ’16, 2016.

[3]. Lei Shan, Minxuan Zhang, Lin Deng, Guohui Gong: A Dynamic Multi-precision Fixed-Point Data Quantization Strategy for Convolutional Neural Network. NCCET 2016, 2016.

[4]. Yufei Ma, Yu Cao, Sarma Vrudhula, Jae-sun Seo: An automatic RTL compiler for high-throughput FPGA implementation of diverse deep convolutional neural networks. FPL 2017, 2017.

[5]. Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan: Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. FPGA ’15, 2015.