Quy trình thực thi một model bằng C

* Step 1: Code các hàm chính của model

Trong quá trình chuyển đổi các model như Mobilenet, SiamRPN, Yolov8 thì các hàm sau đã được triển khai: convolution (cv.c), depth wise (dw.c), point wise (pw.c), add (add.c), average pool (avg.c), max pool (maxpool.c), scale for concatenation (scale.c).

* Step 2: Code một model cụ thể

Mỗi một layer x của model sẽ được thực thi trong một file opx.c. Đối với từng file này, sẽ có một file định nghĩa cấu hình của layer tương ứng (height, width, depth, flt\_off, inp\_off, out\_off, filter, bias,…). Toàn bộ các file định nghĩa này được lưu trong folder layer-cfgs. Gọi hàm thực thi chính của layer này (các hàm định nghĩa ở step 1) với tham số đầu vào lấy từ file cấu hình. Chú ý đối với định dạng shape của input, output có cần padding hay không. Padding ở đây là padding vào chiều channel để có thể truy cập được dữ liệu vào hai chiều cuối cùng lúc. Muốn xem shape chính xác cần như thế nào thì đọc trong các hàm chính ở step 1 để điền vào file cấu hình cho đúng.

* Step 3: Kiểm tra kết quả của từng layer

Trong file main.c có hàm CHECK để kiểm tra từng output của từng tensor đã đúng chưa. Ref\_output được lưu sẵn tại các file định nghĩa cấu hình (đề cập ở step 2).

* Step 4 : Đưa code chạy trên FPGA

Đối với bước này, cần chú ý tới MEM tối đa cho phép. Ví dụ với con zcu102 thì cho phép tối đa 2MB cho Mem VM, 1MB cho Mem WM. Mem VM là Mem sử dụng để tính toán, lưu trữ tensor, Mem WM là Mem để lưu trữ weight, bias. Đối với Mem VM thì tự tối ưu sao cho sử dụng Mem không vượt quá 2MB, còn đối với Mem WM thì sử dụng kĩ thuật DMA để đưa weight, bias từ Mem ngoài DDR vào trong Mem nội WM. Đối với DMA, sử dụng kĩ thuật ping-pong để đưa dữ liệu vào. Load data = ½ dung lượng tối đa cho phép của WM vào trước để tính toán, trong thời gian tính toán các layer đã có weight thì load tiếp data vào phần Mem còn lại để phục vụ tính toán cho các layer tiếp theo. Sau khi đã code xong DMA check lại kết quả đúng chưa rồi chuyển code cho đội phần cứng nạp và chạy trên FPGA.

**=>** Nếu cần cắt tensor để lưu vào mem thì ưu tiên cắt theo chiều C.