Realizzazione di un layer di max pooling per una Convolutional Neural Network

Una Convolutional Neural Network (CNN) è un tipo di rete neurale tipicamente utilizzata per applicazioni di riconoscimento vocale, (ad esempio Keywork spotting system) riconoscimento di oggetti etc.

Un layer di max pooling viene introdotto per ridurre progressivamente le dimensioni delle matrici su cui vengono effettuate le operazioni di una CNN, riducendo il numero di parametri, il numero di operazioni e la memoria necessari.

In questo caso, il layer di max pooling è costituito da:

- N canali d'ingresso C_{in} di dimensione 10x10.
- N canali d'uscita C_{out}.

Ad ogni canale d'ingresso $C_{in}(i)$ è applicato un "filtro" di max pooling di dimensioni 2x2 che produce un elemento del corrispondente $C_{out}(i)$. Il filtro di max pooling seleziona l'elemento massimo tra quelli considerati come mostrato nell'equazione (1):

$$\max\_pool = \max\{a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}\} \quad (1)$$

Dove $a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}$ sono i valori del canale d'ingresso $C_{in}(i)$ i considerati ad un certo istante.

Un esempio di max pooling applicato ad un singolo canale è mostrato in Figura 1. Partendo dalla situazione iniziale indicata in figura 1.a, per calcolare gli elementi successivi si sposta il filtro verso destra fino a raggiungere la colonna finale della matrice d'ingresso come mostrato in figura 1.b. Per calcolare gli elementi successivi è necessario spostare il filtro nella riga successiva come evidenziato in figura 1.c. Procedere in questo modo fino a scorrere tutti gli elementi della matrice di ingresso e raggiungere la posizione evidenziata in figura 1.d.

La stessa operazione deve essere svolta in maniera indipendente sugli N canali d'ingresso C_{in}.

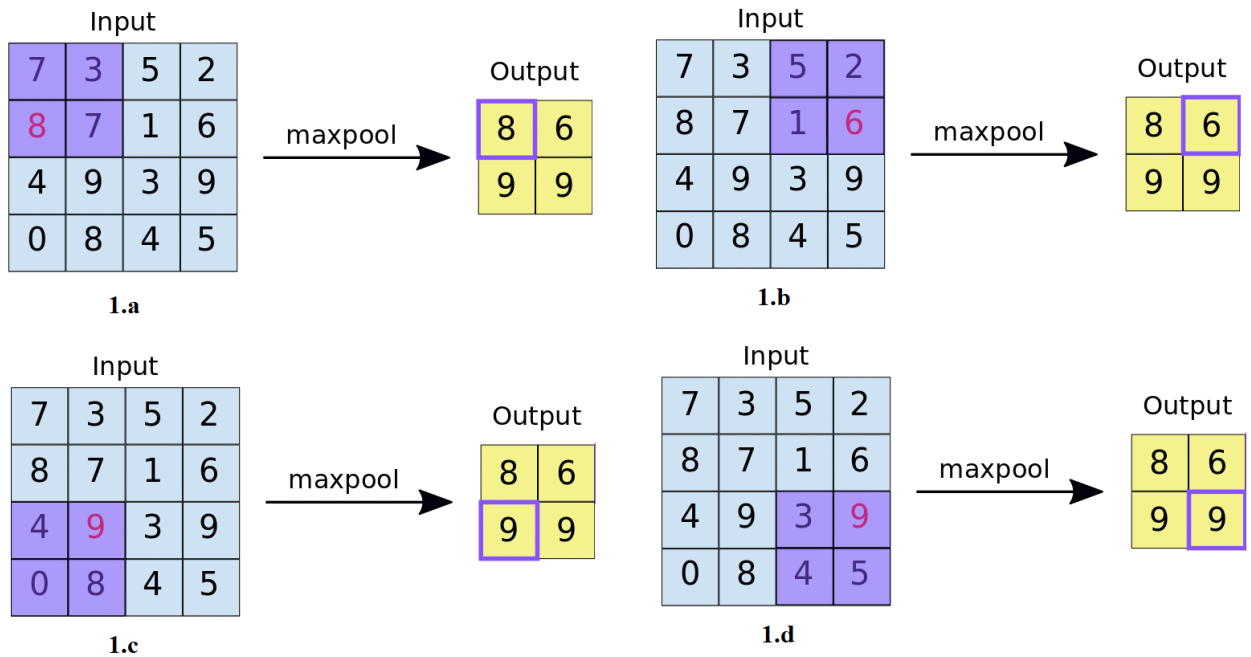

Figura 1

Nel caso specifico, realizzare un layer max pooling in cui viene calcolato un elemento per ogni canale $C_{out}(i)$ per ciclo di clock (N elementi per ciclo di clock). Salvare i dati parziali in una memoria.

Utilizzare parametri generici là dove è possibile

La relazione finale del progetto deve contenere:

- Introduzione (descrizione algoritmo, possibili applicazioni, possibili architetture, etc.)
- Descrizione dell'architettura selezionata per la realizzazione (diagramma a blocchi, ingressi/uscite, etc.)
- Codice VHDL (con commenti dettagliati)
- Test-plan e relativi Testbench per la verifica
- Risultati della sintesi logica automatica su piattaforma Xilinx FPGA Zynq: risorse utilizzate (slice, LUT, etc.), massima frequenza di funzionamento, cammino critico, etc. commentando eventuali messaggi di warnings.
- Conclusioni