Valutazione dell'Impatto della Riduzione della Precisione dei Pesi in una Rete Neurale nei Confronti dell'Accuratezza di Classificazione

Descrizione:
La quantità di memoria utilizzata per memorizzare i parametri di una rete neurale è dominata dalla memorizzazione dei pesi. Ridurre il numero di bit per rappresentare i pesi ha quindi un impatto positivo sull'utilizzo sulla quantità di risorse richieste. La riduzione del numero di bit offre inoltre la possibilità di utilizzare circuiti aritmetici ridotti con una conseguente riduzione di area, potenza e possibilmente una riduzione del percorso critico e quindi un aumento della frequenza di clock. La riduzione del numero di bit utilizzati per rappresentare i pesi ha sicuramente un impatto sull'accuratezza della rete neurale. Si vuole valutare tale impatto.

Flusso generale per l'analisi:
1) Allenare una rete neurale nella sua configurazione originale
2) Valutarne l'accuratezza della rete (Aorig)
3) Modificare i pesi riducendo i bit di rappresentazione
4) Valutare l'accuratezza della rete (Amod)
5) Riallenare la rete in cui i pesi sono rappresentati con un numero ridotto di bit
6) Valutare l'accuratezza della rete (Amod2)
7) Confrontare Aorig, Amod e Amod2

Note:
Riguardo al punto 3) in prima battuta è possibile semplicemente modificare il peso troncando per ridurne l'accuratezza. Sarebbe però più interessante utilizzare una libreria per la rappresentazione dei float a dimensione variabile ed utilizzare, quindi, un numero ridotto di bit. Per esempio, date un'occhiata a https://github.com/Ghost047/Fap (io non l'ho ancora provata)

Come implementazione di rete neurale utilizzate una che preferite. Una molto utilizzata è tinyDNN.

Vi allego un paio di articoli rappresentativi in questo contesto che, se volete, potete sfogliare per ispirarvi.

Per qualsiasi problema non esitate a contattarmi.

M. Palesi