Implementazione di una Rete Convoluzionale in CUDA

Michele Valsesia

Nicholas Aspes

Anno accademico 2018/2019

Introduzione

Obiettivi

- Descrivere brevemente l'architettura ed il funzionamento di una*Rete Neurale*

Introduzione

Obiettivi

- Descrivere brevemente l'architettura ed il funzionamento di una*Rete Neurale*
- Motivare le differenti scelte implementative adottate durante lo svolgimento del progetto

Introduzione

Obiettivi

- Descrivere brevemente l'architettura ed il funzionamento di una*Rete Neurale*
- Motivare le differenti scelte implementative adottate durante lo svolgimento del progetto
- Valutare l'accuratezza e lo speed-up della rete rispetto ad una sua implementazione sequenziale

Reti Neurali

Reti Neurali

Significato Biologico

- Una *Rete Neurale* ha come scopo quello di modellare una rete neurale biologica

Reti Neurali

Significato Biologico

- Una *Rete Neurale* ha come scopo quello di modellare una rete neurale biologica
- Una rete neurale biologica si compone di unità cellulari di base: i *neuroni*

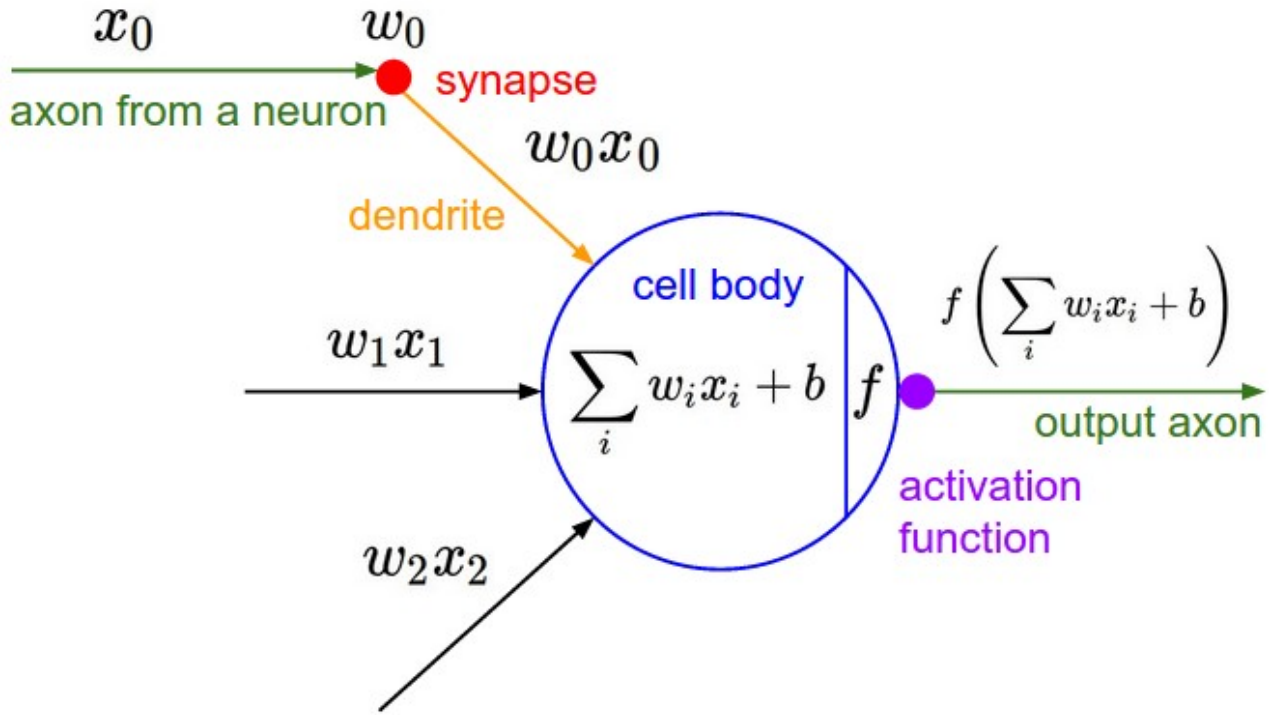
Reti Neurali

Significato Biologico

- Una *Rete Neurale* ha come scopo quello di modellare una rete neurale biologica
- Una rete neurale biologica si compone di unità cellulari di base: i *neuroni*
- I neuroni sono collegati tra loro per mezzo di specifiche giunture chiamate *sinapsi*

Reti Neurali

Neurone



Modello matematico di un neurone

Reti Neurali

Funzionamento Neurone

- Attraverso un meccanismo di eccitazione ed inibizione i pesi sinaptici controllano quanto un neurone venga influenzato dagli altri

Reti Neurali

Funzionamento Neurone

- Attraverso un meccanismo di eccitazione ed inibizione i pesi sinaptici controllano quanto un neurone venga influenzato dagli altri

- I segnali pesati dalle differenti sinapsi vengono trasportati dai dendriti all'interno del neurone e sommati tra loro

Reti Neurali

Funzionamento Neurone

- Attraverso un meccanismo di eccitazione ed inibizione i pesi sinaptici controllano quanto un neurone venga influenzato dagli altri
- I segnali pesati dalle differenti sinapsi vengono trasportati dai dendriti all'interno del neurone e sommati tra loro
- Se la somma supera una certa soglia, il neurone *spara* un segnale lungo l'assone

Reti Neurali

Funzionamento Neurone

- Attraverso un meccanismo di eccitazione ed inibizione i pesi sinaptici controllano quanto un neurone venga influenzato dagli altri
- I segnali pesati dalle differenti sinapsi vengono trasportati dai dendriti all'interno del neurone e sommati tra loro
- Se la somma supera una certa soglia, il neurone *spara* un segnale lungo l'assone
- La *frequenza di sparo* del neurone viene modellata con una funzione di attivazione f

Reti Neurali

Funzioni di Attivazione

Definizione

Una *funzione di attivazione* è una funzione matematica non lineare usata per calcolare l'output di un neurone. Riceve come input la somma pesata dei segnali in ingresso al neurone

Reti Neurali

Funzioni di Attivazione

Definizione

Una *funzione di attivazione* è una funzione matematica non lineare usata per calcolare l'output di un neurone. Riceve come input la somma pesata dei segnali in ingresso al neurone

- *Sigmoide*

Reti Neurali

Funzioni di Attivazione

Definizione

Una *funzione di attivazione* è una funzione matematica non lineare usata per calcolare l'output di un neurone. Riceve come input la somma pesata dei segnali in ingresso al neurone

- *Sigmoide*
- *Tangente Iperbolica*

Reti Neurali

Funzioni di Attivazione

Definizione

Una *funzione di attivazione* è una funzione matematica non lineare usata per calcolare l'output di un neurone. Riceve come input la somma pesata dei segnali in ingresso al neurone

- *Sigmoide*
- *Tangente Iperbolica*
- *Softplus*

Reti Neurali

Sigmoide

Definizione

La *Sigmoide* $\sigma : \mathbb{R} \to [0, 1]$ è definita come $\sigma(x) = \frac{1}{(1+e^{-x})}$

Reti Neurali

Sigmoide

Definizione

La *Sigmoide* $\sigma : \mathbb{R} \to [0, 1]$ è definita come $\sigma(x) = \frac{1}{(1+e^{-x})}$

- Per elevati valori negativi di input la sigmoide restituisce 0: il neurone non spara affatto

Reti Neurali

Sigmoide

Definizione

La *Sigmoide* $\sigma : \mathbb{R} \to [0, 1]$ è definita come $\sigma(x) = \frac{1}{(1+e^{-x})}$

- Per elevati valori negativi di input la sigmoide restituisce 0: il neurone non spara affatto
- Per elevati valori positivi di input la sigmoide restituisce 1: il neurone satura e spara con una frequenza di sparo pari a 1

Reti Neurali

Sigmoide

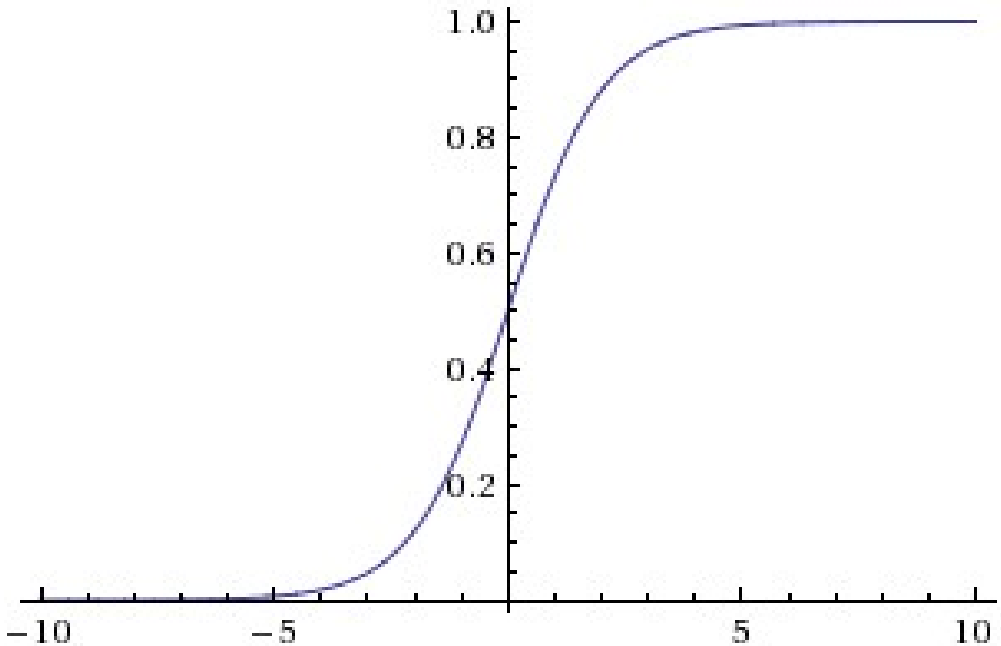
Definizione

La *Sigmoide* $\sigma : \mathbb{R} \to [0, 1]$ è definita come $\sigma(x) = \frac{1}{(1+e^{-x})}$

- Per elevati valori negativi di input la sigmoide restituisce 0: il neurone non spara affatto
- Per elevati valori positivi di input la sigmoide restituisce 1: il neurone satura e spara con una frequenza di sparo pari a 1
- La sua derivata è uguale a $\sigma'(x) = 1 - \sigma(x)$

Reti Neurali

Sigmoide



Rappresentazione grafica Sigmoide

Reti Neurali

Tangente Iperbolica

Definizione

La *Tangente Iperbolica* $\tanh : \mathbb{R} \to [-1, 1]$ è definita come $\tanh(x) = 2\sigma(2x) - 1$

Reti Neurali

Tangente Iperbolica

Definizione

La *Tangente Iperbolica* $\tanh : \mathbb{R} \to [-1, 1]$ è definita come $\tanh(x) = 2\sigma(2x) - 1$

- La tangente iperbolica è una sigmoide scalata

Reti Neurali

Tangente Iperbolica

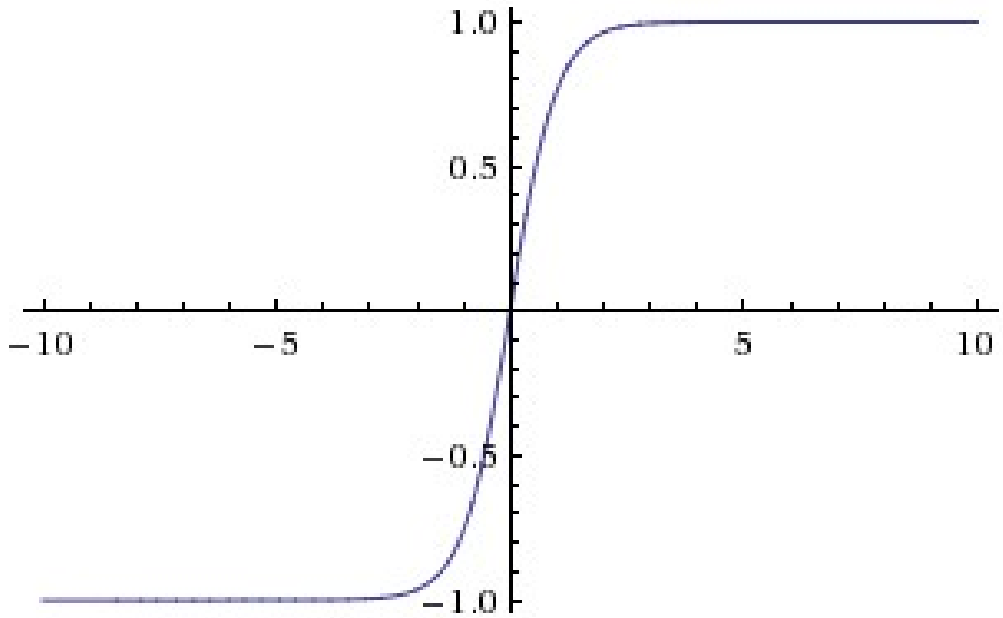
Definizione

La *Tangente Iperbolica* $\tanh : \mathbb{R} \to [-1, 1]$ è definita come $\tanh(x) = 2\sigma(2x) - 1$

- La tangente iperbolica è una sigmoide scalata
- La sua derivata è uguale a $\tanh'(x) = 1 - \tanh^2(x)$

Reti Neurali

Tangente Iperbolica



Rappresentazione grafica Tangente Iperbolica

Reti Neurali

Softplus

Definizione

La *Softplus* $s : \mathbb{R} \to [0, +\infty]$ è definita come $s(x) = \log(1 + e^x)$

Reti Neurali

Softplus

Definizione

La *Softplus* $s : \mathbb{R} \to [0, +\infty]$ è definita come $s(x) = \log(1 + e^x)$

- La softplus è un approssimazione della *Rectifier Linear Unit (ReLU)*

Reti Neurali

Softplus

Definizione

La *Softplus* $s : \mathbb{R} \to [0, +\infty]$ è definita come $s(x) = \log(1 + e^x)$

- La softplus è un approssimazione della *Rectifier Linear Unit (ReLU)*
- Viene usata per sostituire la ReLU che presenta un punto di discontinuità in 0

Reti Neurali

Softplus

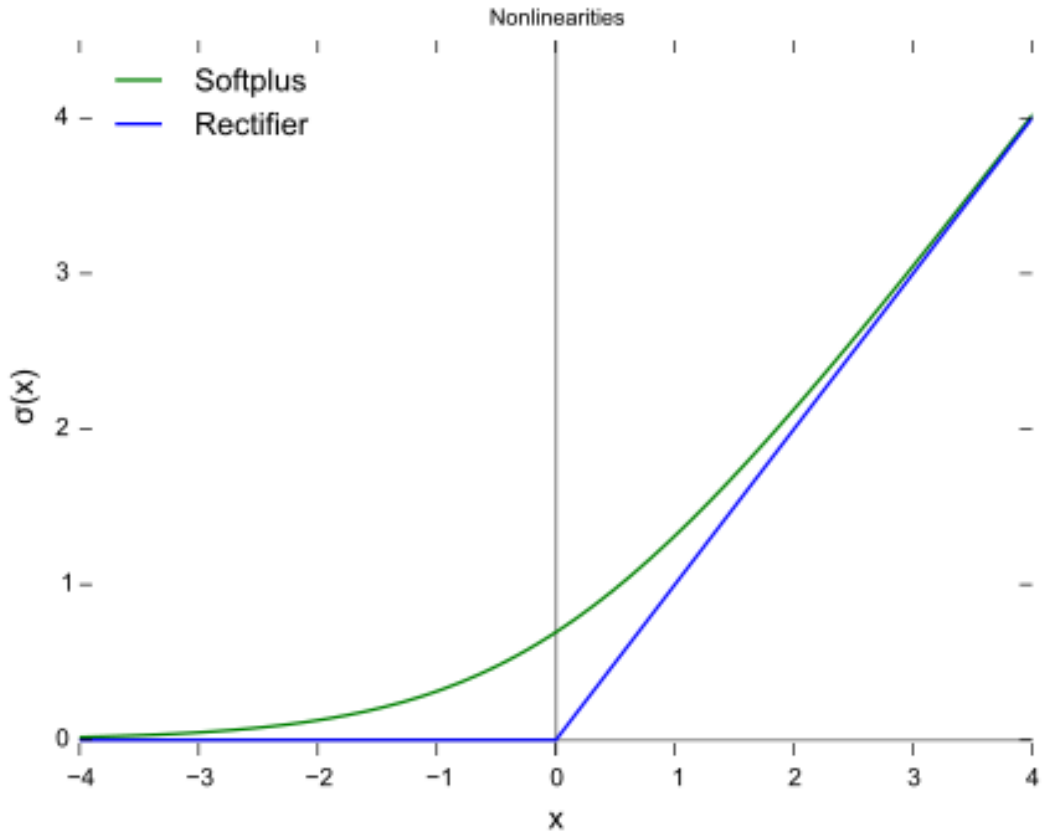
Definizione

La *Softplus* $s : \mathbb{R} \to [0, +\infty]$ è definita come $s(x) = \log(1 + e^x)$

- La softplus è un approssimazione della *Rectifier Linear Unit (ReLU)*
- Viene usata per sostituire la ReLU che presenta un punto di discontinuità in 0
- La sua derivata è uguale a $s'(x) = \frac{1}{(1+e^{-x})}$

Reti Neurali

Softplus



Confronto grafico tra ReLU e Softplus

Reti Neurali

Rete Neurale

Definizione

Una *Rete Neurale* è composta da un insieme di neuroni connessi tra loro in un grafo aciclico

Reti Neurali

Rete Neurale

Definizione

Una *Rete Neurale* è composta da un insieme di neuroni connessi tra loro in un grafo aciclico

- I neuroni sono organizzati in insiemi distinti chiamati *livelli* o *layer*

Reti Neurali

Rete Neurale

Definizione

Una *Rete Neurale* è composta da un insieme di neuroni connessi tra loro in un grafo aciclico

- I neuroni sono organizzati in insiemi distinti chiamati *livelli* o *layer*
- I livelli vengono posti uno di seguito all'altro in modo da formare una sequenza

Reti Neurali

Rete Neurale

Definizione

Una *Rete Neurale* è composta da un insieme di neuroni connessi tra loro in un grafo aciclico

- I neuroni sono organizzati in insiemi distinti chiamati *livelli* o *layer*
- I livelli vengono posti uno di seguito all'altro in modo da formare una sequenza
- I livelli intermedi prendono il nome di *hidden*

Reti Neurali

Rete Neurale

Definizione

Una *Rete Neurale* è composta da un insieme di neuroni connessi tra loro in un grafo aciclico

- I neuroni sono organizzati in insiemi distinti chiamati *livelli* o *layer*
- I livelli vengono posti uno di seguito all'altro in modo da formare una sequenza
- I livelli intermedi prendono il nome di *hidden*
- L'output dei neuroni di un livello diventano l'input dei neuroni del livello successivo

Reti Neurali

Rete Neurale

- Quando si effettua il conteggio dei livelli di una rete non si considera il livello di input

Reti Neurali

Rete Neurale

- Quando si effettua il conteggio dei livelli di una rete non si considera il livello di input

- Una rete a *singolo livello* non presenta livelli hidden

Reti Neurali

Rete Neurale

- Quando si effettua il conteggio dei livelli di una rete non si considera il livello di input
- Una rete a *singolo livello* non presenta livelli hidden
- Per determinare la grandezza di una rete ci si concentra sul numero di neuroni e sui relativi pesi ad essi associati

Reti Neurali

Livello Fully-Connected

Definizione

Un livello è di tipo *Fully-Connected* quando i neuroni di due livelli adiacenti sono completamente connessi tra loro ed i neuroni che formano un livello non condividono nessuna connessione

Reti Neurali

Livello Fully-Connected

Definizione

Un livello è di tipo *Fully-Connected* quando i neuroni di due livelli adiacenti sono completamente connessi tra loro ed i neuroni che formano un livello non condividono nessuna connessione

- I pesi associati ai neuroni di un livello sono salvati in matrici

Reti Neurali

Livello Fully-Connected

Definizione

Un livello è di tipo *Fully-Connected* quando i neuroni di due livelli adiacenti sono completamente connessi tra loro ed i neuroni che formano un livello non condividono nessuna connessione

- I pesi associati ai neuroni di un livello sono salvati in matrici
- Le righe della matrice corrispondono ai neuroni del livello mentre le colonne ai pesi di ciascun neurone

Reti Neurali

Livello Fully-Connected

Definizione

Un livello è di tipo *Fully-Connected* quando i neuroni di due livelli adiacenti sono completamente connessi tra loro ed i neuroni che formano un livello non condividono nessuna connessione

- I pesi associati ai neuroni di un livello sono salvati in matrici
- Le righe della matrice corrispondono ai neuroni del livello mentre le colonne ai pesi di ciascun neurone
- Le reti neurali sono organizzate con una struttura a livelli perché risulta più facile ed efficiente fare operazioni matriciali

Reti Neurali

Rete Neurale Convoluzionale

Una *Rete Neurale Convoluzionale* si differenzia da una più classica in quanto assume che l'input della rete sia un'immagine

Implementazione della Rete

Analisi dei Risultati