



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

# Reti neurali: dall'algoritmo di backpropagation alle reti profonde

**Relatore:** Prof. Alberto Dennunzio

**Co-relatore:** Dott. Luca Manzoni

**Relazione della prova finale di:**

Alessandro Sassi

Matricola 807001

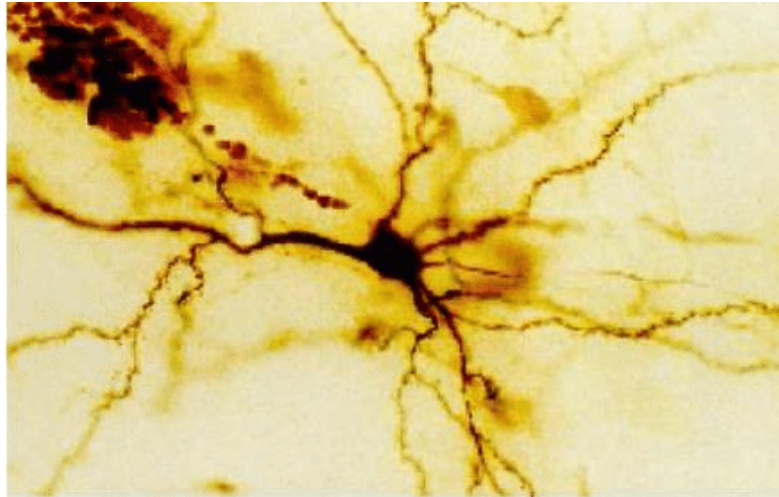
**Anno Accademico 2017-2018**



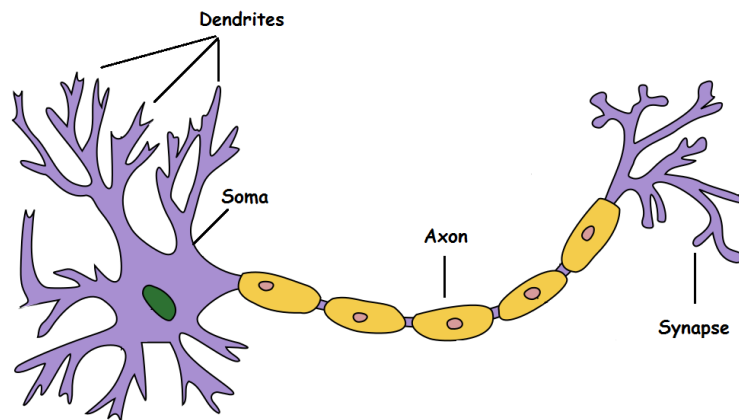
# Introduzione

## L'ispirazione biologica

Il cervello umano è composto da circa 10 miliardi di neuroni. Nella figura riportata in bassa attorno al nucleo di colore verde si estende il corpo cellulare delle cellule neurali , dette *soma* e i canali di input ed output che la circondano la collegano a circa altri 10000 (dieci mila) neuroni. Ogni neurone riceve degli stimoli elettrochimici dai neuroni circostanti attraverso i proprio dendriti. Se la somma degli input elettrici è potente abbastanza il neurone trasmette a sua volta un segnale elettrochimico lungo l'assone , che viene diffuso a tutti quei neuroni i cui dendriti sono connessi alle terminazione del neurone in oggetto. Per quel che vedremo in seguito è importante sottolineare come il neurone attivi il segnale in uscita solo se il segnale in ingresso totale raggiunge un certo livello, a quel punto il neurone attiva il proprio impulso, secondo un unico livello di segnale, ovvero, non ci sono sfumature di attivazione, o il segnale in uscita viene propagato o non viene propagato, non esistono segnali di potenza variabile da distribuire verso i neuroni connessi. Il nostro intero cervello è quindi composto interamente da questa fitta rete di cellule, i neuroni, che comunicano fra loro attraverso segnali elettrochimici. E' sorprendente vedere come la struttura alla base molto semplice, costituita da una cellula che sommando dei segnali in ingresso si attiva , o meno, comunicando con altre cellule attraverso segnali in uscita, riesca nel suo complesso reticolo di interazioni a svolgere funzioni complicate come quelle del cervello nella sua totalità. Lo studio del cervello e la sua comprensione ha potuto ispirare i modelli scientifici e matematici che oggi danno vita a quelle che sono dette oggi *reti neurali artificiali (ANN)*.



(a) *An impression of a real neuron. This image is a light microscope photograph by John Anderson, Institute for Neuroinformatics, Zurich, Switzerland.*



(b) *A biological neuron model*

## La storia delle reti neurali

La capacità dei computer di andare oltre la programmazione verso delle primordiali forme di apprendimento si sviluppò inizialmente come il tentativo di due scienziati di meglio comprendere il funzionamento dei neuroni nel cervello. Era il 1943 e il neurofisiologo Warrent McCulloch e il matematico Walter Pitts svilupparono con dei circuiti elettrici una semplice rete neurale che modellava i neuroni biologici. Nel 1949 Donald Hebb scrisse *L'organizzazione del Comportamento* in cui affermava di come i percorsi neurologici tendessero a rinforzarsi e a diventare più solidi mano a mano che questi venivano allenati e utilizzati. Un lavoro che troverà conferme in studi successivi e documentati nel saggio di Nicholas Carr del 2010, in cui si fa riferimento alla plasticità del nostro cervello e alla capacità dei nostri neuroni di rinforzare i loro collegamenti man mano che vengono "allenati". (PENSARE A DOVE METTERE LA CITAZIONE DI CARR CHE TROVO NELLA MAIL A CABITZA) Nel frattempo i calcolatori diventavano più potenti e sul finire degli anni '50 Bernard Widrow e Marcian Hoff della Stanford University svilupparono "ADALINE" and "MADALINE". La prima era in grado di riconoscere dei

pattern binari nel linee telefoniche e predire i successivi bit in ingresso; la seconda fu la prima rete neurale ad essere applicata in un problema del mondo reale, ovvero era in grado , grazie a dei filtri adattivi (COSA SONO000??) di eliminare gli echi nel chiamate al telefono. Questo dispositivo , per quanto datato e primitivo funziona così bene da avere ancora oggi una sua validità commerciale. Nel corso degli anni , nonostante i primi buoni risultati delle reti neurali l'architettura di von Neumann si impose sulla scena dello sviluppo dei calcolatori, nonostante von Neumann stesso suggerisse di apprezzare l'approccio dell'imitazione delle funzionali neurali. Inoltre all'epoca furono pubblicati degli articoli scientifici, basati però su un presupposto erroneo, ovvero la non derivabilità delle funzioni di apprendimento (CERCARE DI CAPIRE MEGLIO), che raffreddarono gli entusiasmi per questa tecnologia e con questi gli investimenti per ulteriori sviluppi. Inoltre i primordiali successi condussero ad esagerare potenziale ed aspettative nei confronti delle reti neurali, generando in un primo momento clamore e diffidenza, paura anche rispetto a quello che avrebbe potuto costituirsi come rapporto uomo-macchina e in seguito in delusione verso le aspettative non ripagate. Arrivando agli anni '70: nel 1972 furono sviluppate indipendentemente da Kohonen e Anderson delle reti molto simili che si basavano su una matematica matriciale ; del 1975 è invece la prima rete multistrato non supervisionata.

L'interesse verso le reti neurali artificiali si rinnovò nel 1982 quando John Hopfield della Caltech (USA) presentò un paper all'Accademia Nazionale delle Scienze. In questo documento presentava l'opportunità di rendere le macchine più utili ed efficienti usando reti bidirezionali rompendo con la tradizionale idea di reti in cui i segnali si propagassero in un'unica direzione. Nello stesso anno altri due ricercatori : Reilly e Cooper usarono reti "ibride" che applicavano differenti strategie di risoluzione per ognuno dei diversi strati della loro rete. Ancora nel 1982 una spinta ai finanziamenti in questo settore arrivò da una conferenza congiunta tra studiosi statunitensi e giapponesi sul tema. Vedendo i nipponici in vantaggio sullo sviluppo di queste tecnologie gli americani decisero di aumentare le risorse a disposizione della ricerca.

Nel 1986 arriviamo ad una svolta che diventa centrale nel corso di questa relazione, ovvero, lo sviluppo dell'idea del backpropagation error. Tre gruppi indipendenti di sviluppatori, tra i quali compariva David Rumelhart, un ex professore di Stanford del dipartimento di psicologia, arrivarono sempre indipendentemente all'idea di distribuire gli errori prodotti dalla computazione della rete su tutta la rete stessa in modo da riprogrammarne i parametri e farla apprendere dai propri errori. Quelle che prima abbiamo chiamato reti "ibride" avevano solo due layer mentre queste nuove reti a retropropagazione dell'errore presentavano molti livelli e , dato lo stato dell'arte degli hardware del tempo, erano ritenute molto lente; e così hanno proseguito ad essere, tanto che negli anni 2000, le computazioni potevano richiedere settimane intere. Così è stato finché i recenti sviluppi degli ultimi anni hanno messo a disposizione delle reti hardware performanti e notevoli quantità di dati da cui le reti possano trarre informazioni ed imparare [1].

## I diversi tipi rete neurale

Al giorno d'oggi le reti neurali sono tante e diverse tra loro, vengono riconosciute con l'utilizzo di sigle sempre più lunghe e in questa sezione vedremo alcune delle principali. Ad alcune accenneremo solamente mentre altre non verranno nemmeno citate per via del

loro numero elevato. Le reti neurali biologiche hanno ispirato le reti neurali artificiali che vengono utilizzate per approssimare delle funzioni che non sono conosciute a priori [6] e questo costituisce, semplicisticamente, un ribaltamento del classico paradigma con cui usiamo i computer: la programmazione infatti usa funzioni impostate dal programmatore per computare dei dati in output mentre le reti neurali cercano di creare una funzione a partire da dati forniti. e questo costituisce, semplicisticamente, un ribaltamento del classico paradigma con cui usiamo i computer: la programmazione infatti usa funzioni impostate dal programmatore per computare dei dati in output mentre le reti neurali cercano di creare una funzione a partire da dati forniti. Le reti feedforward furono le prime reti ad essere pensate e sono anche le più semplici. Sono costituite da diversi strati (*layers*), che possono essere di input per i dati di apprendimento, nascosti o di output. Il perceptrone può essere visto come modello di un neurone biologico e può approssimare nella sua semplicità dei circuiti logici. Gli strati delle FF sono tutti collegati fra loro e il flusso dei dati viaggia in una sola direzione, senza mai formare cicli, dallo strato di input, attraverso gli strati nascosti fino allo strato di output. Questa rete viene definita profonda all'aumentare degli *hidden layers* che si nascondono fra input ed output e prendono il nome di *deep feedforward neural networks* (SPECIFICARE SE ESISTE UNA DIFFERENZA SOSTANZIALE, DEL TIPO : UNA VOLTA C'ERANO SOLO QUELLE NON PROFONDE O PERKE HANNO CREATE QUELLE PROFONDE, QUALE ERA L'ESIGENZA CHE LE DISTINGUE).

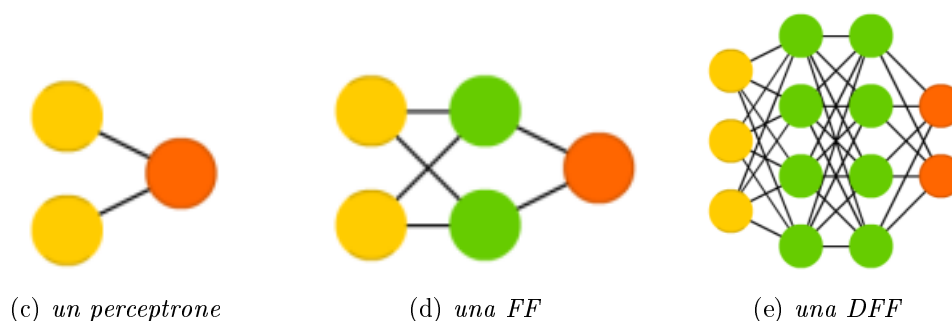


Figura 1: Confronto tra tipologie diverse di algoritmi

Esistono poi anche le *radial basis network* (RBF) che altro non sono che reti feedforward (e che possiamo schematizzare proprio come la FF in figura sopra) ma con una funzione di attivazione radiale di base. Non tutte le reti prendono il nome dalla propria funzione di attivazione ma questa sembrava avere particolari speranze nel 1988, quando uscì il documento che la presentava poiché la funzione che stava alla sua base era in grado di interpolare in spazi di molte dimensioni. (PERCHÈ QUESTA SI ALLORA ????)

Un altro caso speciale o un'evoluzione delle FF sono le *reti convoluzionali profonde* (*deep convolutional network*, DCN) il cui utilizzo primario è il processing delle immagini e in alcuni casi anche degli input audio. Si ispirano a quello che è il funzionamento del sistema visivo degli uomini. Le reti convoluzionali sono in grado di riconoscere soggetti nelle immagini sapendo quindi fare una clas-

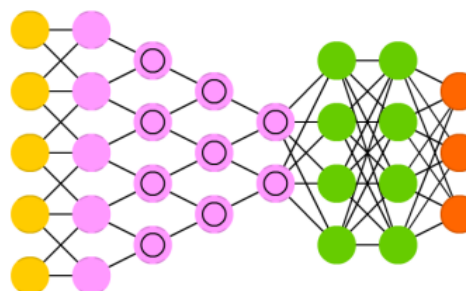


Figura 2: Un perceptrone (P).

sificazione in un insieme di immagini che gli vengono sottoposte. Le DCN scansionano le immagini partendo da piccoli blocchi di pixel vicini tra loro e individuano le caratteristiche delle immagini porzione dell'immagine per porzione, fino a classificare l'immagine nella sua totalità. Diversamente delle tradizionali FF i layer di mezzo non sono collegati nodo per nodo fra loro, piuttosto i nodi vicini sono collegati ad un intorno dei nodi vicini nello strato successivo. Questo vuole ispirarsi al modo in cui il nostro campo visivo funziona e reagisce agli stimoli esterni. Man mano che il segnale fluisce lungo la rete attraversa strati costituiti da sempre meno nodi fino ad arrivare alla terminazione della rete in cui viene "attaccata" una piccola rete FF che consente di processare ulteriormente i dati e fare nuove astrazioni.

Figure 2 is a wrapped figure.

Un recente sviluppo, del 2014, delle reti che sembra essere promettente è la combinazione di due reti insieme. Di solito vengono associate tra loro FF e reti convoluzionali e si impostano in modo tale che esse si contrappongano nello svolgere dei compiti. Immaginiamo un pittore che dipinge su una tela un gatto, in sua compagnia un giudice verifica la qualità del lavoro confrontandolo che un insieme di fotografie di gatti reali; il pittore si allena a ritrarre sempre più fedelmente i gatti e il giudice viene sfidato nel giudicare via via se il gatto sottopostogli dall'amico sia una fedele riproduzione artistica o una foto reale. Ecco nelle reti *GAN*, come sono abbreviate, una delle due fa un lavoro di generazione mentre l'altra fa un lavoro di giudizio, in pratica una genera contenuti che vengono giudicati dall'altra e si allenano vicendevolmente nel giudicare una l'operato dell'altra, spingendo la seconda a produrre contenuti sempre più difficili da predire mentre la prima viene allenata dalla seconda, che producendo via via lavori sempre più approssimabili a quelli dell'insieme di confronto.

Abbiamo parlato fin'ora solamente di reti in cui i segnali partendo da uno strato di input si propagano unidirezionalmente verso la fine della rete. E' giusto annoverare brevemente anche quelle reti che invece formano dei cammini ciclici fra i loro nodi.





# L'algoritmo di Backpropagation

## L'idea dell'apprendimento

Cosa significa apprendere? Una domanda non così scontata per un'azione che a diversi livelli compiamo lungo tutta la nostra vita. Ci troviamo ad imparare memorizzando concetti, dati, stimoli sensoriali dai cinque sensi e rielaboriamo tutto questo nel nostro cervello producendo idee nuove che derivano dagli elementi di partenza conservati nella nostra memoria.

Un esempio banale è il contatto con un oggetto caldo, da una singola esperienza generalizzeremo l'idea che il dato oggetto con cui abbiamo avuto a che fare procuri la spiacevole sensazione di bruciare. Ma cosa accade se l'esperienza è più complessa? Ad esempio potremmo toccare un ferro da stiro bollente e farci l'idea che quell'oggetto dalla data forma del ferro sia intrinsecamente caldo, ma toccandolo in un momento in cui è spento non ci accadrebbe nulla, anzi percepiremmo il freddo della piastra metallica. La moltitudine di esperienze per un contesto simile può fornirci un'idea più chiara del perché di determinati avvenimenti. Avvicinandoci ad un ferro da stiro collegato alla corrente elettrica con la spina avremmo un ulteriore elemento di ragionamento in base al quale fare nuove ipotesi sullo stato dell'oggetto: conoscendo come funzionano gli elettrodomestici potremmo pensare che se il ferro è attaccato alla corrente potrebbe essere acceso e quindi caldo, potremmo pensare che sia rimasto attaccato alla corrente perché nessuno si è preoccupato di staccarlo e che quindi passato del tempo sia ormai freddo e che possa essere riposto senza bruciarsi. Tutto lineare e chiaro, sono intuizioni semplici per un adulto che ha vissuto una simile scena tante volte.

Ma se fossimo un neonato? Da adulti penseremmo che questo neonato non ha di certo dei genitori responsabili se si trova a giocare con un ferro da stiro, ma così è e da neonati in cui siamo sarà probabilmente la prima volta che vedremo questo oggetto cuneiforme e non avremo esperienze pregresse per deliberare quale tipo di interazione sia più intelligente avere con esso. Un bambino di un anno infatti non può sapere a cosa serve quell'oggetto, non può sapere che appartiene ad una classe degli oggetti della vita che può trovarsi nel duplice stato di acceso/spento, non può sapere che questi oggetti hanno un filo che li collega al muro e non può sapere che da quel filo passa l'energia a renderlo funzionante, non sanno tante cose. Ecco il punto del discorso, le esperienze ci insegnano a comprendere, inferire idee a partire dalla somma di altre. Il bimbo crescendo imparerà a capire che il ferro lo utilizza la mamma in determinate situazioni, imparerà a capire che solo una parte dell'oggetto può bruciare e capirà che dopo diverso tempo dopo che la mamma ha terminato di adoperarlo non brucerà più. Ma questa conoscenza da cosa sarà data? Sarà data con la serie di evidenze avute dal bambino nel toccare il ferro: ci saranno state volte in cui, vispo di curiosità, lo avrà avvicinato per vederlo meglio e sfiorandolo si sarà bruciato; altre volte la mamma, certa della sua esperienza gli avrà

mostrato l'oggetto per farglielo conoscere e avvicinandolo non si sarà fatto male. Nella rete intricata di esperienze, sensazioni e fatti tangibili il cervello del bimbo imparerà a dare maggior peso ad alcuni elementi cognitivi piuttosto che ad altri realizzando dentro di lui una consapevolezza che si alimenta ed arricchisce nel tempo. Il bimbo imparerà a capire che lo stato caldo/freddo del ferro dipenderà in larga parte dalla sua posizione nello spazio: vedendolo infatti riposto sullo scaffale dello sgabuzzino saprà che con probabilità non è stato usato di recente e che quindi non deve bruciare ; mentre darà poco peso, sempre nel determinare lo stato dell'oggetto, alla sua disposizione nel tempo: poco importerà che il ferro si trovi in salotto la mattina o la sera , sua mamma è una eccellente casalinga e può stirare a qualunque ora. Di conseguenza non potrà determinare dentro di lui l'idea di come il ferro sia, potrebbe averlo toccato una mattina ed essersi scottato così come potrebbe averlo toccato la sera successiva ed essersi scottato nuovamente... allora il bruciare del ferro non dipenderà dal momento della giornata.

Ecco , se questa metafora è stata ben digerita possiamo di certo avvicinarci più facilmente con la spiegazione di come agisca l'algoritmo di backpropagation, l'algoritmo che aiuta le reti neurali a dare il giusto peso alle caratteristiche degli elementi che le reti analizzano.

## La matematica dietro a backpropagation

Il nostro algoritmo viene applicato in modo da riprogrammare ad ogni iterazione i parametri della rete , apprendendo via via la loro configurazione migliore per approssimare la funzione che risolve il nostro problema. Nell'apprendimento supervisionato delle reti FF di cui ci occupiamo alleniamo la rete fornendole un paragone con i risultati aspettati del set di training. Nel set di training sono quindi presenti i vettori di soluzione per ogni problema. Indichiamo con  $y(x) = (y_1, y_2, \dots, y_n)^T$  il vettore di  $n$  dimensioni di output che ci aspetteremmo in uscita dalla nostra rete;  $y(x)$  è la funzione che la nostra rete vogliamo approssimi, la forma di questa funzione non è data e non si conosce a priori, conosciamo soltanto i suoi valori. BP valuta ad ogni iterazione ( o EPOCH ?) quanto lontano dal valore atteso è il risultato della rete, computa per cui una funzione dei costi

$$C(w, b) \equiv \frac{1}{2n} \sum_x \| y(x) - a \|^2 \quad (1)$$

che chiamiamo *errore quadratico medio (MSE)*;  $a$  invece è a sua volta una funzione dei pesi  $w$  e dei bias  $b$  e rappresenta il vettore output della rete: in questo modo abbiamo quindi una funzione che mette in relazione il risultato atteso con quello effettivo e che diventa tanto più piccola nel suo valore quanto sono simili quando  $a$  approssima  $y(x)$ . Quindi, se l'intento della rete è produrre un output che approssimi al meglio le soluzioni del set di apprendimento, e  $C$  ne è la misura, dobbiamo cercare di capire per quali valori questa funzione si minimizza, ovvero trovare per quali valori delle sue  $w$  e  $b$  variabili l'errore è minimo.

Per far questo abbiamo bisogno di un metodo. Potremmo pensare di trattare questa minimizzazione con un approccio analitico ma considerando che le variabili in gioco nelle reti reali possono essere anche miliardi questo non è fattibile. Facciamo allora ricorso ad un algoritmo, che in diversi passi successivi ci aiuta a trovare il punto di minimo che cerchiamo. Questo algoritmo è il *metodo della discesa del gradiente*, spieghiamo in

primis cosa è il gradiente e poi come funziona l'algoritmo. Data una funzione di due o più variabili, il suo gradiente in un punto  $x = (x_1, x_2, \dots, x_n)$  del dominio è il vettore delle sue derivate parziali rispetto a tutte le variabili:  $\nabla C \equiv (\frac{\partial C}{\partial x_1}, \frac{\partial C}{\partial x_2}, \dots, \frac{\partial C}{\partial x_n})^T$ . Questa è solo una formula matematica e non ci dà bene idea di cosa stiamo parlando. Dobbiamo allora rifarci ad uno scenario reale per capire meglio cosa sia il gradiente; la realtà è fatta di tre dimensioni quindi rimaniamo in uno spazio tale per spiegarci meglio: se fossimo in una valle e avessimo una palla perfettamente tonda, il gradiente è il vettore della direzione lungo la quale la funzione cresce maggiormente e, di contrario,  $-\nabla C$  è la direzione in cui la palla si muoverebbe se, appoggiata a terra, venisse lasciata libera di rotolare lungo il pendio della valle. In una tale ipotesi le variabili libere non sono  $n$  come abbiamo generalizzato, ma sono soltanto due  $v_1, v_2$  (usiamo le  $v$  e non le  $x$  per separare la spiegazione dalla sua metafora). Possiamo vedere una figura che rappresenta quanto detto.

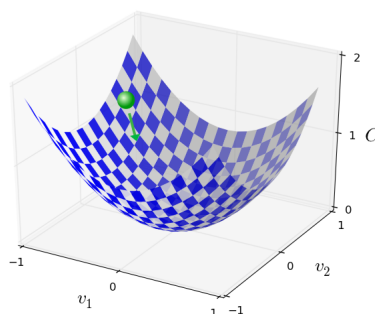


Figura 3: Il grafico della funzione  $C(v_1, v_2)$  con la pallina verde che rappresenta il punto in analisi e il vettore gradiente spiccato da esso, ad indicare in quale direzione moverebbe nella discesa

Se ora abbiamo un'idea più chiara di cosa sia il vettore gradiente spieghiamo in cosa consiste il metodo della discesa che lo utilizza. Ipotizzando la stessa metafora della valle vogliamo sfruttare il fatto che la discesa della pallina si arresterà nella parte più bassa dalla depressione dipinta a scacchi della Baviera. Il nostro metodo computa una discesa simile ma diversa nel fatto che il percorso che farà la nostra pallina non approssimerà la discesa *fisica* che avverrebbe nella realtà. Per cui il nostro metodo non si rifà alle equazioni della dinamica di Newton, è un pò diverso. Per ora sappiamo che, dato un punto sulla superficie di una valle, la direzione opposta a quella del gradiente è quella in cui la palla si muoverebbe se fosse libera di scendere; nel nostro metodo seguiamo un percorso di discesa fatto di tanti piccoli passi che prevedono, partendo da un punto, di muoversi verso  $-\nabla C$  per un piccolo tratto. Mossi nel nuovo punto, che chiamiamo  $P_2$ , ricalcoleremo il gradiente  $\nabla C(P_2)$  e faremo altri piccoli passi in direzione  $-\nabla C(P_2)$ . Si capisce che questo metodo va iterato a lungo per scendere lungo la valle fino ad arrivare nel suo punto più basso. Tra un iterazione  $k$  e la successiva  $k + 1$ , ci saremo spostati dal punto  $P_k$  al punto  $P_{k+1}$  per cui  $\Delta P = P_{k+1} - P_k = -\eta \nabla C$  dove  $\eta$  è un piccolo parametro positivo detto il *learning rate* che fornisce un'indicazione di quanto grande sia la distanza percorsa fra un iterazione e l'altra nell'algoritmo; va da sé che se il  $\eta$  cresce l'algoritmo procede più speditamente, ma questo potrebbe portare anche a degli errori. Ritorniamo dal parallelo con la valle, la palla e le variabili  $v$  alla nostra funzione dei costi  $C(w, b)$  e

vediamo come il metodo del gradiente funzioni con più di due variabili, anche se di questa situazione non potremo avere una visualizzazione. Così come ci spostiamo da un punto all'altro nella valle, aggiornando le due variabili della nostra posizione, ora aggiorniamo le variabili della funzione dei costi, ovvero  $w$  i pesi e  $b$  i bias. La regola per aggiornare queste componenti sarà:

$$w_k \rightarrow w'_k = w_k - \eta \frac{\partial C}{\partial w_k} \quad (2)$$

$$b_l \rightarrow b'_l = b_l - \eta \frac{\partial C}{\partial b_l} \quad (3)$$

dove  $k$  è il (FARSI AIUTARE A CAPIRE BENE  $k$  ED  $L$ ). Ripetendo questa regola arriveremo a trovare il minimo della funzione. (SPIEGARE IL MINIBATCH E IL FATTO CHE È STOCASTICO IL GRADIENTE)

Ora che abbiamo spiegato come funziona il gradiente dobbiamo andare più a fondo a spiegare come si rapportano fra loro i neuroni collegati, come si attivano e come propagano il proprio segnale. Ogni neurone  $j^{th}$  è collegato (stiamo sempre parlando del caso delle reti feed forward) ai neuroni dello strato precedente dall'arco di peso  $C$  (cioè dal neurone  $k$  in  $l-1$  al neurone  $j$  in  $l$ ) e la sua attivazione viene determinata da una computazione fatta sui segnali in ingresso, più precisamente è la somma delle attivazioni dei neuroni dello strato precedente a lui collegati, moltiplicati per il peso del loro arco verso  $j^{th}$  più una quantità  $b_j$  specifica del neurone, questa somma prende il nome di  $z_j^l$ .

$$a_j^l = \sigma(z_j^l) = \sigma\left(\sum_k w_{kj}^l a_k^{l-1} + b_j^l\right) \quad (4)$$

Spieghiamo brevemente cosa è  $\sigma$ . Ogni neurone riceve input dagli strati precedenti e internamente deve decidere se propagare a sua volta lo stimolo o meno. In biologia abbiamo già anticipato che i neuroni, una volta eccitati, non hanno una via di mezzo nel rispondere al segnale per cui o propagano a loro volta lo stimolo nella rete "accendendosi" oppure rimangono in quiete senza reagire. Nelle reti neurali la funzione di attivazione di solito non emette un segnale binario, piuttosto propaga vari livelli d'intensità del segnale.  $\sigma$  rappresenta quella classe di funzioni sigmoidali (che hanno una caratteristica curva a  $S$ ) che ben si adattano a rappresentare l'attivazione dei neuroni e che hanno lo scopo di normalizzare gli output nel range da 0 a 1. Quindi, se un neurone reale può veder rappresentata la propria funzione di attivazione come una funzione a gradino una sigmoide è la versione addolcita della funzione gradino, come mostriamo qui sotto con il plot della funzione sigmoide:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

Esiste una quantità  $a_j^l$  per ogni neurone  $j$  nello strato  $l$ , possiamo annotare come un vettore questi  $a$  output per tutti gli  $n$  neuroni di quello strato  $a^l = [a_1^l, a_2^l, \dots, a_j^l, \dots, a_n^l]$ . Vediamo ora come di strato in strato i neuroni si attivano fra loro e per questo adottiamo una notazione matriciale di quanto abbiamo appena visto: denotiamo allora una matrice  $w^l$  dei pesi per ogni layer della rete, dove le sue entry sono  $w_{kj}^l$  i vari pesi che legano i neuroni fra lo strato  $l-1$  ed  $l$  e denotiamo, allo stesso modo di  $a^l$  anche il vettore dei bias  $b^l$  per il livello  $l$  e riscriviamo la formula 3 come:

$$a^l = \sigma(w^l a^{l-1} + b^l) \quad (6)$$

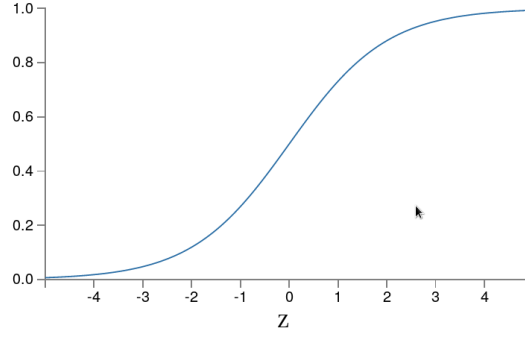


Figura 4: Il grafico della funzione sigmoide  $\sigma(z)$

La propagazione del calcolo degli output dei singoli livelli procede verso il livello finale, l'*output layer* e una volta calcolato l'errore nella funzione dei costi, che abbiamo chiamata  $C$  questo errore viene propagato in senso opposto fino al layer degli input. Spieghiamo meglio quale è il senso di questo "propagarsi".  $C(w, b)$  è stata definita in (1) e viene calcolata sull'*output layer*  $a$  della rete, questo che possiamo indicare per chiarezza come  $a^{out}$  per esplicitare a quale livello sia riferito è calcolato a sua volta come indicato nella (5) ed è quindi funzione dei pesi e dei bias del livello che lo precedono:  $a^{out}(w^l, b^l, a^l)$  dove  $l$  è l'ultimo *hidden layer*, quello prima di *out*; ma compare anche  $a^l$ , questo in catena è una funzione di pesi, attivazioni e bias di  $l - 1$ . Vediamo quindi come in verità tutta la funzione dei costi non sia altro che una lunghissima e imponente funzione composta, che richiama i parametri dei livelli sottostanti. Compreso questo possiamo dire che se la  $C$  è una funzione composta dei pesi e dei bias di tutti i suoi sottolivelli e quindi della rete in generale; possiamo cercare di capire in che modo questi parametri, lungo la rete, contribuiscano alla definizione di  $C$  e in ultima analisi di come contribuiscano alla quantità dell'errore espresso da  $C$  e della rete nel suo complesso; in questo senso l'errore viene retropropagato lungo la rete. Per vedere che peso ha complessivamente un peso  $w_k^l$  si calcola la sua derivata parziale  $\frac{\partial C}{\partial w_k^l}$  ma per poter far questo bisogna procedere livello per livello. Abbiamo detto che  $C$  è una grande funzione composta ed è necessario partire dal fronte di output e andare indietro per calcolare le derivate dei pesi più in profondità applicando la *chain rule* delle funzioni composte  $(f \circ g)' = (f' \circ g) \cdot g'$ .

$$C(a^{out}) = C(a^{out}(w^l, b^l, a^l)) = C(a^{out}(w^l, b^l, a^l(w^{l-1}, b^{l-1}, a^{l-1}))) = \dots \quad (7)$$

La formula sopra dovrebbe rendere l'idea di come si procede a comporre la funzione costi, innestando via via i pesi e i bias. Per trovare la derivata  $\frac{\partial C}{\partial w_k^m}$  di un peso di un neurone nel livello  $m$  dovremo quindi calcolare prima di lui tutte le derivate delle funzioni che lo contengono:

$$\frac{\partial C}{\partial w_k^m} = \frac{\partial C}{\partial a^{out}} \cdot \frac{\partial a^{out}}{\partial a^{l-1}} \cdot \frac{\partial a^{l-1}}{\partial a^{l-2}} \cdot \dots \cdot \frac{\partial a^m}{\partial w_k^m} \quad (8)$$

In questo senso la procedura avanza verso la base della rete, calcolando tutte le derivate parziali rispetto ai  $w$  e ai  $b$  della rete (per ogni neurone di ogni layer); il che corrisponde ad avere calcolato il gradiente della funzione dei costi:

$$\nabla C = \left[ \frac{\partial C}{\partial w_1^1}, \frac{\partial C}{\partial w_2^1}, \dots, \frac{\partial C}{\partial w_n^1}, \frac{\partial C}{\partial b_1^2}, \dots, \frac{\partial C}{\partial b_2^2}, \dots \right] \quad (9)$$

Una volta ottenuto il gradiente di  $C$  possiamo procedere all'aggiornamento dei pesi, l'apprendimento della rete si basa infatti su questo, sulla modifica dei suoi parametri interni in modo che dal loro cambiamento scaturisca un comportamento complessivamente diverso (migliore) nei confronti degli input che le vengono passati e con diversi training riesca a darci i risultati che speriamo. La regola di aggiornamento di un generico peso è questa:

$$w_j^{l+} = w_j^l + \eta \frac{\partial C}{\partial w_j^l} \quad (10)$$

che altro non è che la procedura della discesa del gradiente spiegata in precedenza.

## I miglioramenti

### Migliorare backpropagation attraverso diversi approcci sulla discesa del gradiente

L'algoritmo di backpropagation si basa sull'ottimizzazione di una funzione d'errore tra l'output della rete e il risultato desiderato. La funzione in questione ha come variabili i pesi associati ai collegamenti tra i vari neuroni della rete e i loro bias. Nella sua versione più semplice la funzione cerca il proprio minimo seguendo la direzione del proprio gradiente. Venendo alle variazioni su questo metodo; furono proposti il *metodo dei minimi quadrati* (in inglese OLS: Ordinary Least Squares) e il *metodo detto di quasi-newton* (che è una variazione sul canonico metodo di Newton) che risultano essere però troppo lenti da computare, in special modo nelle reti molto ampie e ad entrare nello specifico il problema per i metodi di quasi-newton è che lo spazio in memoria richiesto per salvare una approssimazione dell'inversa dell'Hessiana cresce quadraticamente rispetto al numero dei pesi su cui viene effettuato il calcolo [3]. Altre tecniche come *BFGS* o il *gradiente coniugato* possono essere in alcuni casi delle valide alternative. Più genericamente il problema del migliorare l'ottimizzazione dell'errore trova soluzione quando si migliora la capacità di computare la più utile lunghezza di passo possibile (step-length) ovvero quanto distante una iterazione deve essere da quella che la precede avendo seguito la direzione del gradiente. Esistono per questo algoritmi del secondo ordine e del primo ordine, che possiamo vedere a confronto nelle due immagini in fig.1.1.

Si può vedere come l'andamento della traiettoria conduca al minimo in numero inferiori di passi utilizzando algoritmi del secondo ordine.

Per velocizzare BP è stato anche introdotto il momento, che ha lo scopo di ridurre le oscillazioni della traiettoria dovute ad una non proficua scelta della lunghezza del passo per le iterazioni e fu proprio nel 1988 che Robert A. Jacobs, propose quattro euristiche per il miglioramento del tasso di convergenza del BP.

Nel 1988 la ricerca aveva già riconosciuto la bontà della procedura BP e si investigavano quegli algoritmi di ricerca dell'ottimo per la funzione di minimo che computassero la discesa del gradiente soltanto localmente (ovvero rimanendo in intorni molto vicini al punto di iterazione). Questo cosiddetto *locality constraint* veniva motivato dal fatto che costituiva una buona metafora tecnologica della controparte biologica, le reti neurali, a cui si ispirava e in secondo luogo dall'ipotesi che questi algoritmi *locali* fossero più adatti ad essere processati in parallelo. Fra le quattro euristiche di Jacobs vi era l'adattamento

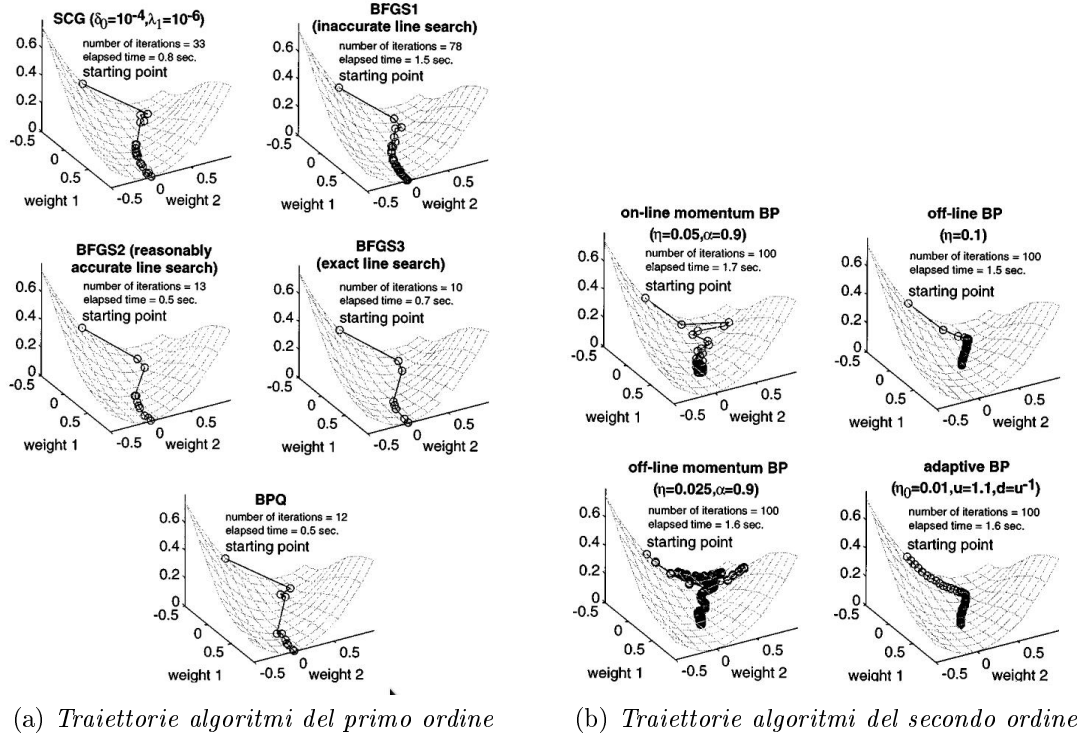


Figura 5: Confronto tra tipologie diverse di algoritmi

dei tassi d'apprendimento nel tempo e si spiega che ogni peso della rete dovrebbe avere il proprio tasso d'apprendimento specifico. Le implementazioni di queste euristiche venivano individuate nel *momento* e nella regola d'apprendimento *delta-bar-delta*.

## Low complexity NNs

Alcuni altri indirizzi di miglioramento dell'efficacia delle reti sfruttano il BP per modellare delle reti meno complesse (*low complexity*) che possano essere utili a scopi meno sofisticati. Ad esempio, nelle reti convoluzionali, dove il processo di riconoscimento degli oggetti ha luogo, gli algoritmi vengono eseguiti su costose GPU che dissipano grandi quantità di energia e un tale scenario non è adatto a scopi dove il livello di dettaglio nel riconoscere gli oggetti non è così elevato. Applicazioni più popolari e frequenti come il riconoscimento facciale nei dispositivi mobili deve per forza di cose girare in locale sui processori embedded che animano gli smartphone, per questo i ricercatori sfruttano varianti del BP per creare reti convoluzionali a bassa complessità. Queste modellano problemi molto meno *demanding* e possono quindi essere eseguite più velocemente anche sui dispositivi meno prestanti [5].

## Le performance tra CPU e GPU

Nella sezione riguardante la matematica di backpropagation abbiamo citato come i calcoli per le attivazioni dei neuroni non vengano eseguiti uno alla volta ma , livello per livello,

ma vengano raggruppati in vettori e matrici non solo per snellire la notazione usata ma anche perché a tutti gli effetti è quello che accade realmente. La quantità di dati coinvolta in una rete neurale può essere davvero notevole, con lo strato di input che può avere fino a migliaia di nodi, per questo le matrici che vengono coinvolte nei calcoli possono diventare molto grandi ed occupare notevoli quantità di spazio in memoria.

Tale utilizzo di memoria costituisce un limite difficile per la computazione da parte delle CPU, anche quelle multi-core. Le CPU ottimizzano la velocità di calcolo sui singoli processi seriali e sono un'alternativa ad un altro tipo di architettura per processori: le GPU. Le GPU furono inventate nel 1999 da Nvidia per la grafica dei PC e nel 2007 gli sviluppatori poterono sfruttare le potenzialità di calcolo parallelo anche per applicazioni più generiche grazie alla piattaforma di programmazione *CUDA*. Nel 2009 [INSERIRE NOTA DA QUI: <https://www.quora.com/Who-introduced-GPU-to-deep-learning>] un documento accademico di Stanford illustrava quanto più velocemente si potesse addestrare delle reti neurali utilizzando metodi con GPU, fino a 70 volte più veloci della controparte con CPU. Tale approccio consentì l'uso di dataset più ricchi e un aumento dei parametri a disposizione per settare la rete. Altri benefici collegati all'introduzione delle schede grafiche furono una democratizzazione della ricerca sull'AI. I costi derivati dall'attività di ricerca diminuirono, evitando l'utilizzo di cluster di CPU e molti più laboratori in giro per il mondo poterono contribuire all'avanzamento del settore.

Ma perché le GPU? Tanta più memoria si necessita per sviluppare i calcoli di queste matrici tanto più l'utilizzo delle GPU migliora le prestazioni poiché queste ottimizzano i calcoli del processo su una banda più ampia verso la memoria, date le numerose unità di calcolo. A voler fare un paragone i calcoli richiesti sono numerosi come i pacchi postali da scambiare fra due località che vengono messe in collegamento da un capiente ma lento camion e da una rapida automobile che però ha una ridotta capacità, nel momento in cui i pacchi diventano molti allora tanto maggiori sono i vantaggi di trasportarli in unico viaggio, per quanto lento, con il camion [4].

Un esperimento presentato da alcuni studiosi nel 2016 comparava due reti neurali nello svolgere il medesimo compito, una implementata su GPU l'altra su CPU. L'esperimento consiste nel riconoscere cifre numeriche scritte a mano, presentate in un dataset chiamato *MNIST*, come quelle in figura. Presentiamo i risultati delle velocità di computazione nella tabella e nel grafico riportato sotto, dai quali possiamo evidenziare come aver allenato la rete neurale attraverso la GPU aumenti di un fattore molto grande le performance rispetto alla controparte in CPU e che la GPU sia da preferire quando nella rete sono presenti un grande numero di ATTRIBUTI (SPECIFICARE MEGLIO), altrimenti il costo dell'inizializzazione della scheda non verrebbe giustificato dall'incremento delle velocità [2].



Population	Epochs	CPU time in seconds	GPU time in seconds
10	100	5.8687	0.283682
20	1000	117.362351	2.674261
30	1000	177.518307	2.646332
40	1000	236.354287	2.677458
50	1000	295.3512	2.675260
60	10000	3552.425203	26.610284

Figura 6: Riassunto dei risultati dell'esperimento

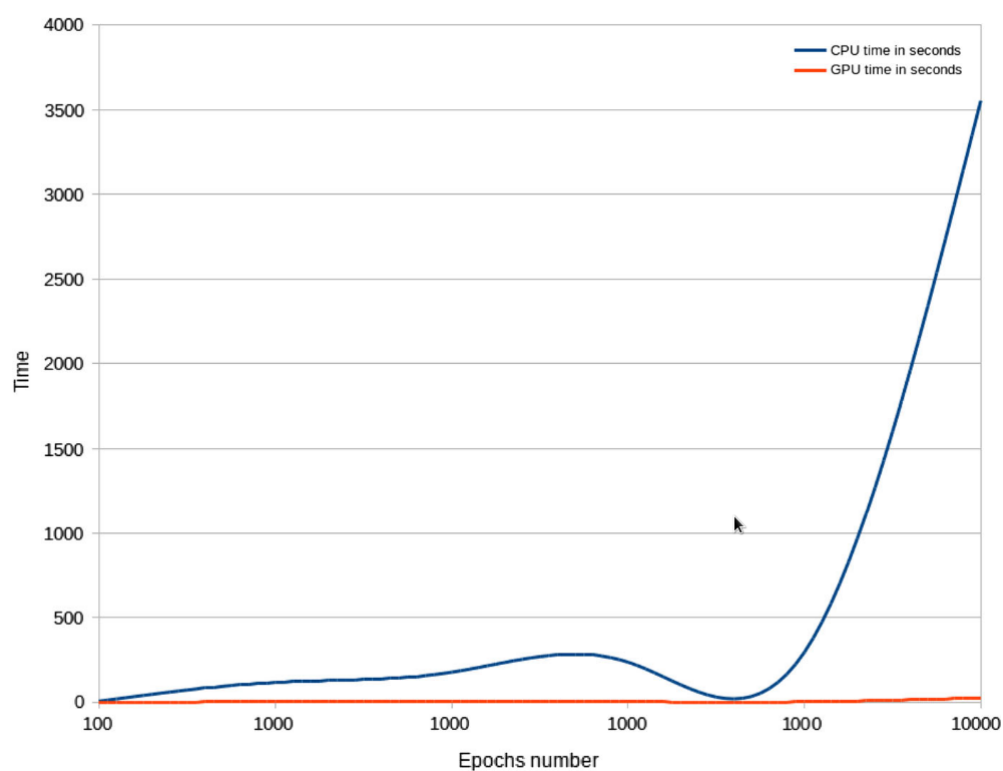


Figura 7: Il grafico che mostra le differenze tra le velocità di esecuzione tra CPU e GPU

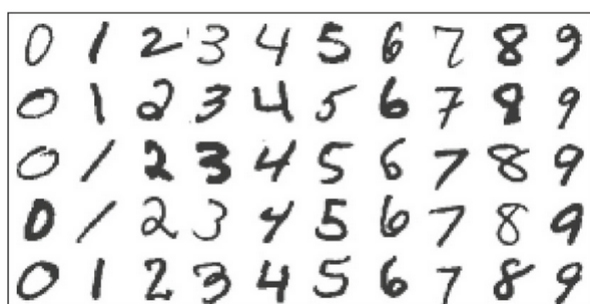


Figura 8: Un esempio di cifre da riconoscere da parte della reti nell'esperimento



# Bibliografia

- [1] . History of neural networks. <https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/index.html>, 2010. Online; accessed 28 Febraury 2019.
- [2] Ricardo Brito, Simon Fong, Kyungeun Cho, Wei Song, Raymond Wong, Sabah Mohammed, and Jinan Fiaidhi. Gpu-enabled back-propagation artificial neural network for digit recognition in parallel. *The Journal of Supercomputing*, 72(10):3868–3886, 2016.
- [3] Kazumi Saito and Ryohei Nakano. Partial bfgs update and efficient step-length calculation for three-layer neural networks. *Neural Computation*, 9(1):123–141, 1997.
- [4] Tim Dettmers. Answer to why gpu are well suited for deep learning?, 2016. [Online; accessed 10-March-2019].
- [5] Subarna Tripathi, Gokce Dane, Byeongkeun Kang, Vasudev Bhaskaran, and Truong Nguyen. Lcdet: Low-complexity fully-convolutional neural networks for object detection in embedded systems. 05 2017.
- [6] Wikipedia contributors. Types of artificial neural networks — Wikipedia, the free encyclopedia, 2019. [Online; accessed 1-March-2019].