

ALMA MATER STUDIORUM  
UNIVERSITY OF BOLOGNA

SCUOLA DI INGEGNERIA E ARCHITETTURA  
*Corso di Laurea Magistrale in Ingegneria Informatica*

MASTER THESIS

---

# Compression of Convolutional Neural Networks with tensor decomposition

---

*Author:*  
Ali Alessio Salman

*Thesis Supervisor:*  
Prof. Stefano MATTOCCIA

*Thesis Advisor*  
Ph.D Matteo Poggi

DISI  
Dipartimento di Informatica - Scienza e Ingegneria

III Session  
2017/2018



## *Abstract*

Quest'elaborato verte sullo studio delle reti neurali artificiali, e si prefigge principalmente due scopi: il primo consiste nel comprendere il funzionamento che sta alla base delle reti neurali ed il loro apprendimento; il secondo sta nell'affrontare studio ed implementazione dei modelli più all'avanguardia delle Convolutional Neural Networks e saggiare la loro nota efficacia nei compiti di visione artificiale, in particolare nella classificazione.

Nel *Capitolo 1* viene fornita un'introduzione alle reti neurali, che si conclude con la proposta di un problema da risolvere con un perceptron multi-strato; nel *Capitolo 2* si percorre passo passo l'implementazione da zero di quest'ultimo, fino a portare a termine l'effettivo addestramento; il *Capitolo 3* introduce le Convolutional Neural Networks, ne spiega l'architettura ed i loro ultimi successi; nel *Capitolo 4* vengono implementate queste reti e testate su un due dataset diversi; il *Capitolo 5* presenta un'architettura allo stato dell'arte che viene confrontata con quella del capitolo precedente; nel *Capitolo 6* viene analizzato un caso d'uso industriale della classificazione, affrontato sfruttando il transfer-learning; il *Capitolo 7*, infine, espone le riflessioni sul lavoro svolto e conclude l'elaborato.



## *Acknowledgements*

The acknowledgments and the people to thank go here, don't forget to include your project advisor...



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 . . . . .	1
1.2 Multi-layer Perceptron . . . . .	2
1.2.1 Strati Nascosti . . . . .	2
1.3 Caso di studio: prevedere il profitto di un ristorante . . . . .	3
<b>2 Convolutional Neural Networks</b>	<b>5</b>
2.1 . . . . .	5
2.2 Forward Propagation . . . . .	7
2.3 Backpropagation . . . . .	8
2.4 Verifica numerica del gradiente . . . . .	12
2.5 Addestramento . . . . .	14
2.5.1 Apprendimento supervisionato . . . . .	14
2.5.2 Discesa del gradiente . . . . .	15
2.6 Ottimizzazione: diverse tecniche . . . . .	15
2.7 Overfitting . . . . .	19
2.7.1 Rilevare l'overfitting . . . . .	19
2.7.2 Contromisure . . . . .	21
2.8 Risultati . . . . .	24
<b>3 Reti Neurali Convoluzionali</b>	<b>25</b>
3.1 Breve introduzione . . . . .	25
3.2 Architettura . . . . .	25
3.2.1 Strato di Convoluzione . . . . .	26
3.2.2 Strato di ReLU . . . . .	28
3.2.3 Strato di Pooling . . . . .	29
3.2.4 Strato completamente connesso (FC) . . . . .	30
3.3 Applicazioni e risultati . . . . .	31
3.3.1 Confronto con l'uomo . . . . .	31
<b>4 4 Tensor Decomposition</b>	<b>33</b>
4.1 Background . . . . .	33
4.1.1 Tensor rank . . . . .	33
4.1.2 Singular value decomposition . . . . .	34
4.2 Tensor mathematical tools . . . . .	35
4.2.1 Basic operations . . . . .	35
4.2.2 Tucker Decomposition . . . . .	37
HO-SVD . . . . .	37
4.2.3 Canonical Polyadic Decomposition . . . . .	38

4.3	Application of tensor decompositon on CNN . . . . .	38
4.3.1	Convolutional layer as 4-mode tensors . . . . .	38
4.3.2	CPD . . . . .	38
4.3.3	Tucker . . . . .	38
4.4	MNIST: addestramento . . . . .	39
4.5	CIFAR: preprocessing . . . . .	39
4.6	CIFAR: addestramento . . . . .	39
<b>5</b>	<b>Addestrare un modello allo stato dell'arte</b>	<b>43</b>
5.1	Deep Residual Network . . . . .	43
5.2	Implementazione di ResNet di Facebook . . . . .	45
5.3	Addestramento su CIFAR . . . . .	46
5.4	LeNet vs ResNet: confronto . . . . .	47
<b>6</b>	<b>Caso d'uso: fine-tuning su dataset arbitrario</b>	<b>49</b>
6.1	Il problema . . . . .	49
6.2	Transfer Learning . . . . .	49
6.2.1	Fine-tuning . . . . .	50
6.3	Dataset . . . . .	51
6.4	Fine-tuning su Resnet . . . . .	52
6.4.1	Training . . . . .	52
6.4.2	Testing . . . . .	55
<b>7</b>	<b>Conclusioni</b>	<b>59</b>
<b>A</b>	<b>MLP: Codice addizionale</b>	<b>61</b>
A.1	Classi in Lua . . . . .	61
A.2	La classe Neural_Network . . . . .	62
A.3	Metodi getter e setter . . . . .	63
<b>B</b>	<b>Il framework Torch</b>	<b>65</b>
B.1	Introduzione . . . . .	65
B.2	Utilizzo base per reti neurali . . . . .	66
B.2.1	Supporto CUDA . . . . .	67
B.3	ResNet . . . . .	67

# List of Figures

1.1	Modello di calcolo di un neurone (a sinistra) e schema del neurone artificiale (a destra) . . . . .	1
1.2	Struttura di un percettrone multistrato con un solo strato nascosto . . . . .	4
2.1	Architettura del MLP per la previsione dei profitti . . . . .	6
2.2	Cercare il minimo di una fz. seguendo la discesa del gradiente . . . . .	9
2.3	La funzione sigmoide e la sua derivata . . . . .	10
2.4	Apprendimento supervisionato: schema generale . . . . .	14
2.5	Confronto tra il momentum classico e NAG . . . . .	16
2.6	Confronto dei metodi di ottimizzazione durante il training . . . . .	18
2.7	L'output della rete $\hat{y}$ è vicino all'output desiderato $y$ . . . . .	18
2.8	La funzione polinomiale ha un errore nullo sul dataset laddove la funzione lineare invece lo ha del 100%. Tuttavia, la curva è eccessivamente complessa ed affetta da rumore; avrà quindi cattive capacità di generalizzazione. La retta, al contrario, approssima molto meglio i punti della distribuzione sottostante. Se definiamo questi punti come il test set, allora la retta avrà un'accuratezza maggiore. . . . .	19
2.9	La rete mostra overfitting. Dopo l'iterazione 100 le performance sul test set iniziano ad essere sbagliate. Attorno alla 250 il modello diventa troppo complesso e le predizioni sono pessime . . . . .	21
2.10	Dopo aver applicato la L2 regularization, la rete non è più affetta da overfitting. . . . .	23
3.1	Architettura di una CNN che classifica segnali stradali: si evidenzia la divisione tra gli strati che fungono da feature extractor ed il classificatore finale . . . . .	26
3.2	I diversi strati tipici di una CNN . . . . .	26
3.3	Convoluzione con un kernel: primi due step . . . . .	27
3.4	Ogni neurone è connesso a solo 1 regione locale dell'input ma a tutta la profondità (i.e. canali colori). La depth dell'output è data dal numero K di filtri, in questo caso 5 . . . . .	28
3.5	Max pooling: in uscita l'immagine avrà 1/4 dei pixel di partenza. . . . .	29
3.6	Architettura di una CNN . . . . .	30
3.7	Tipica CNN in un task di classificazione; la classe vincente è quella con la probabilità più alta, indicata alla fine . . . . .	30
4.1	Example of k-mode multiplication on 3-dimensional tensor. . . . .	40
4.2	Percentuali di accuracy a seconda della funzione d'attivazione. La ReLU produce indubbiamente risultati migliori. . . . .	41
4.3	Fibers and slices of a tensor: fibers is an equivalent term for a tensor mode. . . . .	41

4.4	Tucker-2 Decompositions for speeding-up a generalized convolution. Each box corresponds to a 3-way tensor $X, Z, Z'^{andY}$ in equation (??-??). Arrows represent linear mappings and illustrate each scalar value on the right is computed. Red tube, green cube and blue tube correspond to 1x1, dxd and 1x1 convolution respectively. . . . .	42
4.5	Tensor Decompositions for speeding up a generalized convolution. Each box correspond to a feature map stack within a CNN, (frontal sides are spatial dimensions). Arrows show linear mappings and demonstrate how scalar values on the right are computed. Initial full convolution (A) computes each element of the target tensor as a linear combination of the elements of a 3D subtensor that spans a spatial d d window over all input maps. Jaderberg et al. (B) approximate the initial convolution as a composition of two linear mappings in which the intermediate mpa stack has R maps, being R the rank of the decomposition. Each of the two-components computes each target value with a convolution based on a spatial window of size dx1 or 1xd in all input maps. Finally, CP-decomposition (C) by Lebedev et al. approximates the convolution as a composition of four smaller convolutions: the first and the last components compute a standard 1x1 convolution that spans all input maps while the middle ones compute a 1D grouped convolution <b>only on one</b> input map. . . . .	42
5.1	Un "Residual Block", all'input viene aggiunto $F(x)$ che è il residual . . . . .	43
5.2	Confronto architetture: VGG-Net la più innovativa della competizione ILSVRC 2014, rete classica a 34 strati (centro), Residual Network a 34 strati (sinistra) . . . . .	45
5.3	Top-1 e Top-5 training accuracy di ResNet sul dataset CIFAR10 . . . . .	47
5.4	Top-1 e Top-5 validation accuracy di ResNet sul dataset CIFAR10 . . . . .	48
6.1	Alcuni esempi delle immagini delle api del dataset . . . . .	51
6.2	Alcuni esempi delle immagini delle formiche del dataset . . . . .	52
6.3	Architettura della rete: il transfer learning avviene utilizzando i CNN codes provenienti da ResNet per addestrare il classificatore SoftMax . . . . .	53
6.4	Curve di apprendimento sul nuovo dataset: il modello generalizza in maniera ottimale e non vi sono segni di overfitting . . . . .	56
6.5	Testing su 153 esempi ancora non visti. Accuracy = 95.4% . . . . .	57
B.1	Architettura del framework Torch . . . . .	66

# List of Tables

2.1 Results after fine-tuning the . . . . .	7
2.2 Variabili usate nel testo e nel codice . . . . .	7



# List of Symbols

$a$	distance	m
$P$	power	W ( $\text{J s}^{-1}$ )
$\omega$	angular frequency	rad



*For/Dedicated to/To my...*



## Chapter 1

# Introduction

### 1.1

Una rete neurale artificiale – chiamata normalmente solo rete neurale (in inglese *Neural Network*) – è un modello di calcolo adattivo, ispirato ai principi di funzionamento del sistema nervoso degli organismi evoluti che secondo l'approccio connessionista [**WConnessionismo**] possiede una complessità non descrivibile con i metodi simbolici. La caratteristica fondamentale di una rete neurale è che essa è capace di acquisire conoscenza modificando la propria struttura in base alle informazioni esterne (i dati in ingresso) e interne (le connessioni) durante il processo di apprendimento. Le informazioni vengono immagazzinate nei parametri della rete, in particolare, nei pesi associati alle connessioni. Sono strutture non lineari in grado di simulare relazioni complesse tra ingressi e uscite che altre funzioni analitiche non sarebbero in grado di fare.

L'unità base di questa rete è il neurone artificiale introdotto per la prima volta da McCulloch e Pitts nel 1943 (fig. 1.1).

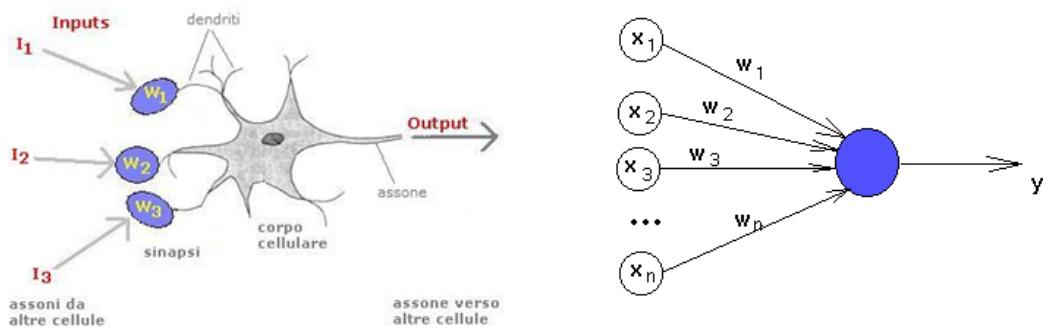


Figura 1.1: Modello di calcolo di un neurone (a sinistra) e schema del neurone artificiale (a destra)

Si tratta di un'unità di calcolo a  $N$  ingressi e 1 uscita. Come si può vedere dall'immagine a sinistra gli ingressi rappresentano le terminazioni sinaptiche, quindi sono le uscite di altrettanti neuroni artificiali. A ogni ingresso corrisponde un peso sinaptico  $w$ , che stabilisce quanto quel collegamento sinaptico influisca sull'uscita del neurone. Si determina quindi il potenziale del neurone facendo una somma degli ingressi, pesata secondo i pesi  $w$ .

A questa viene applicata una funzione di trasferimento non lineare:

$$f(x) = H(\sum_i (w_i x_i)) \quad (1.1)$$

ove  $H$  è la funzione gradino di Heaviside [**WHeaviside**]. Vi sono, come vedremo, diverse altre funzioni non lineari tipicamente utilizzate come funzioni di attivazioni dei neuroni. Nel '58 Rosenblatt propone il modello di *Percettrone* rifiinando il modello di neurone a soglia, aggiungendo un termine di *bias* e un algoritmo di apprendimento basato sulla minimizzazione dell'errore, cosiddetto *error back-propagation* [1].

$$f(x) = H\left(\sum_i(w_i x_i) + b\right), \quad \text{ove } b = \text{bias} \quad (1.2)$$

$$w_i(t+1) = w_i(t) + \eta \delta x_i(t) \quad (1.3)$$

dove  $\eta$  è una costante di apprendimento strettamente positiva che regola la velocità di apprendimento, detta *learning rate* e  $\delta$  è la discrepanza tra l'output desiderato e l'effettivo output della rete.

Il percettrone però era in grado di imparare solo funzioni linearmente separabili. Una maniera per oltrepassare questo limite è di combinare insieme le risposte di più percetroni, secondo architetture multistrato.

## 1.2 Multi-layer Perceptron

Il Multi-layer Perceptron (*MLP*) o percettrone multi-strato è un tipo di rete feed-forward che mappa un set di input ad un set di output. È la naturale estensione del percettrone singolo e permette di distinguere dati non linearmente separabili.

Il *mlp* possiede le seguenti caratteristiche:

- Ogni neurone è un percettrone come quello descritto nella sezione 1.1. Ogni unità possiede quindi una propria funzione d'attivazione non lineare.
- A ogni connessione tra due neuroni corrisponde un peso sinaptico  $w$ .
- È formato da 3 o più strati. In 1.2 è mostrato un MLP con uno strato di input, un solo strato nascosto (o *hidden layer*) ed uno di output.
- L'uscita di ogni neurone dello strato precedente è l'ingresso per ogni neurone dello strato successivo. È quindi una rete *completamente connessa*. Tuttavia, si possono disconnettere selettivamente settando il peso sinaptico  $w$  a 0.
- La dimensione dell'input e la dimensione dell'output dipendono dal numero di neuroni di questi due strati. Il numero di neuroni dello strato nascosto è invece indipendente, anche se influenza di molto le capacità di apprendimento della rete.

Se ogni neurone utilizzasse una funzione lineare allora si potrebbe ridurre l'intera rete ad una composizione di funzioni lineari. Per questo - come detto prima - ogni neurone possiede una funzione di attivazione non lineare.

### 1.2.1 Strati Nascosti

I cosiddetti *hidden layers* sono una parte molto interessante della rete. Per il teorema di approssimazione universale [2], una rete con un singolo strato nascosto e un numero finito di neuroni, può essere addestrata per approssimare una qualsiasi funzione continua su uno spazio compatto di  $\mathbb{R}^n$ . In altre parole, un singolo strato nascosto è abbastanza potente da imparare un ampio numero di funzioni. Precisamente, una rete a 3 strati è in grado di separare regioni convesse con un numero di lati  $\leq$  numero neuroni nascosti.

Reti con un numero di strati nascosti maggiore di 3 vengono chiamate reti neurali profonde o *deep neural network*; esse sono in grado di separare regioni qualsiasi, quindi di approssimare praticamente qualsiasi funzione. Il primo e l'ultimo strato devono avere un numero di neuroni pari alla dimensione dello spazio di ingresso e quello di uscita. Queste sono le terminazioni della "black box" che rappresenta la funzione che vogliamo approssimare.

L'aggiunta di ulteriori strati non cambia *formalmente* il numero di funzioni che si possono approssimare; tuttavia vedremo che nella pratica un numero elevato di strati migliora di gran lunga le performance della rete su determinati task, essendo gli hidden layers gli strati dove la rete memorizza la propria rappresentazione astratta dei dati in ingresso. Nel capitolo 4 vedremo un'architettura all'avanguardia con addirittura 152 strati.

### 1.3 Caso di studio: prevedere il profitto di un ristorante

Prendendo spunto dalla traccia d'esame di Sistemi Intelligenti M del 2 Aprile 2009:

*"Loris è figlio della titolare di una famoso spaccio di piadine nel Riminese e sta tornando in Italia dopo aver frequentato con successo un prestigioso Master in Business Administration ad Harvard, a cui si è iscritto inseguendo il sogno di esportare in tutto il mondo la piadina romagnola. Nel lungo viaggio in prima classe, medita su come presentare alla mamma, che sa essere un tantino restia alle innovazioni, il progetto di aprire un ristorantino a New York City."*

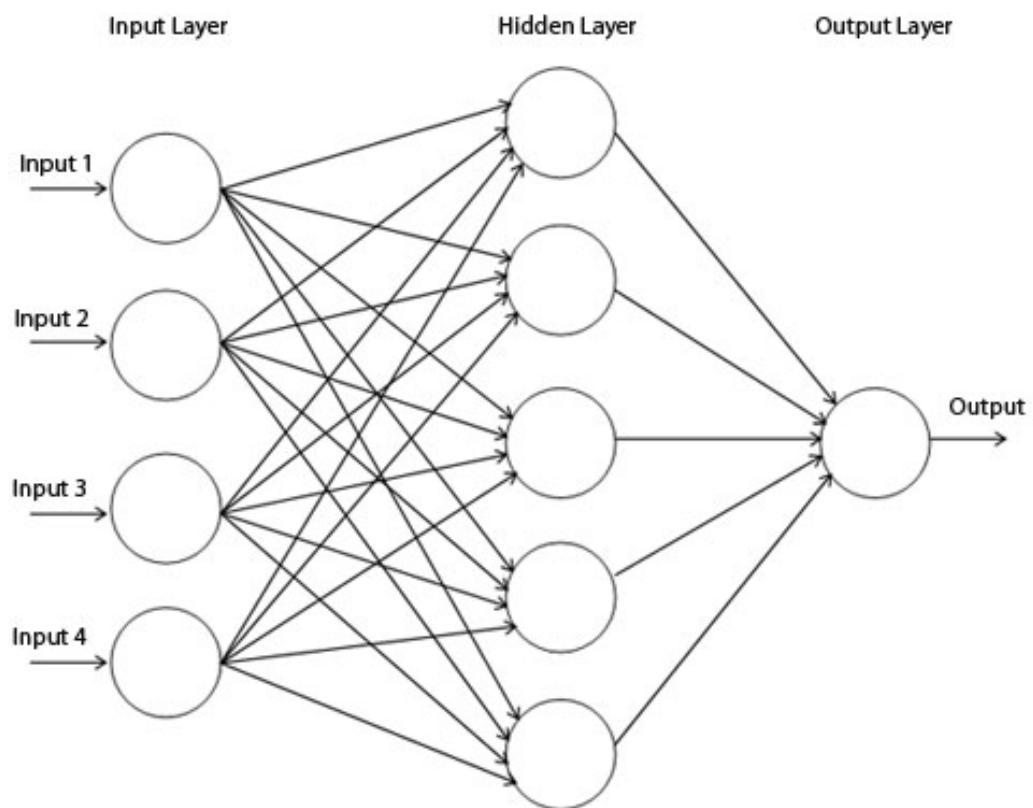
Loris ha esportato con successo la piadina a NY, (si veda [3]) ma col passare degli anni ha notato alcuni problemi e vuole utilizzare di nuovo le sue brillanti capacità analitiche per migliorare il profitto del suo ambizioso ristorante.

I problemi sono 2:

1. il ristorante è conosciuto ormai - si sa che tutti vogliono mangiare italiano - ma il numero dei coperti è rimasto a 22, come quelli iniziali;
2. gli orari di apertura sono troppo lunghi e vi sono alcune zone morte dove il costo di mantenere aperto il ristorante è maggiore rispetto al ricavo dei pochi clienti che si siedono a mangiare durante quelle ore;

Secondo la National Restaurant Association [4] [5], il profitto medio lordo annuo di un ristorante negli Stati Uniti varia dal 2 al 6%. Così Loris ha collezionato alcuni dati riguardo agli ultimi anni e - attratto da tutta quest'entusiasmo attorno alle reti neurali - decide di provare ad utilizzarle per trovare il trade-off ottimale di coperti e di orari di apertura settimanali per massimizzare il profitto del suo ristorante.

Dati questi presupposti, si vedrà nel capitolo ?? come implementare da zero un multi-layer perceptron e addestrarlo per sudetto scopo.



*Figura 1.2:* Struttura di un percettrone multistrato con un solo strato nascosto

## Chapter 2

# Convolutional Neural Networks

Per il caso di studio introdotto nel Capitolo ?? si è implementato da zero un percepitrone multistrato a 3 livelli come quello illustrato in sezione 1.2. Vediamo qui di seguito le varie parti da implementare passo passo per costruire un MLP, addestrarlo e verificare che l’addestramento sia stato eseguito in maniera corretta. Il progetto è realizzato in Lua, per utilizzare il framework per il *Machine Learning Torch*(si veda l’appendice B) e mantenere la consistenza con i capitoli successivi, nei quali si userà nuovamente Torch per addestrare reti neurali molto più complesse.

### 2.1

Nei vari anni, Loris ha cambiato le due variabili in gioco annotando di volta in volta i risultati. Siccome i coperti erano troppo pochi e le file d’attesa erano troppo lunghe, il ristorante perdeva alcuni clienti. Quindi Loris ha portato i coperti a 25 diminuendo le ore settimanali a 38, ed i profitti sono aumentati. Tuttavia, non era raro che ancora qualche cliente dovesse aspettare in piedi per troppo tempo (si sa la vita a NY è frenetica), finendo poi per scegliere un ristorante adiacente. Inoltre, aveva diminuito troppo drasticamente le ore; nel weekend i clienti arrivavano fino a tardi, quindi scegliere di rimanere aperti un’ora in più sarebbe stato lungimirante. Così, dopo l’allargamento della sala principale, ha aggiunto altri coperti ed aumentato le ore settimanali a 40, segnando un record personale di 4.4% di profitti annui.

Quindi, i dati in ingresso ed in uscita, in  $X$  e  $Y$  rispettivamente sono:

$$X = \begin{pmatrix} 22 & 42 \\ 25 & 38 \\ 30 & 40 \end{pmatrix} Y = \begin{pmatrix} 2.8 \\ 3.4 \\ 4.4 \end{pmatrix}$$

Osservando le dimensioni dei dati si nota che la rete deve avere 2 input e dare in uscita 1 output, che chiameremo  $\hat{y}$ , in contrapposizione a  $y$  che è l’uscita desiderata. Per quanto detto nella sezione 1.2, il MLP deve avere 2 neuroni nello strato di ingresso ed 1 solo in uscita. Inoltre, avrà uno strato nascosto con 3 neuroni. La dimensione di ogni strato fa parte di un insieme di parametri che viene deciso “a mano” sperimentando, i cosiddetti *hyperparameters*. Questi parametri non vengono aggiornati durante l’addestramento - come i pesi della rete - ma vengono decisi a priori.

In figura 2.1 è mostrata l’architettura generale della nostra rete.



Figura 2.1: Architettura del MLP per la previsione dei profitti

Di seguito gli snippet di codice per la definizione dei dati e dell'architettura della rete secondo lo schema appena presentato.

```

1 ----- Part 1 -----
2 th = require 'torch'
3 bestProfit = 6.0
4 -- X = (num coperti, ore di apertura settimanali), y = profitto lordo
   -- annuo in percentuale
5 torch.setdefaulttensortype('torch.DoubleTensor')
6 X = th.Tensor({{22,42}, {25,38}, {30,40}})
7 y = th.Tensor({{2.8},{3.4},{4.4}})
8
9 --normalize
10 normalizeTensorAlongCols(X)
11 y = y/bestProfit

1 ----- Part 2 -----
2 --creating the NN class in Lua, using a nice class utility
3 class=require 'class'
4 local Neural_Network = class('Neural_Network')
5
6 function Neural_Network:__init(inputs, hiddens, outputs)
7     self.inputLayerSize = inputs

```

Table 2.1: Results after fine-tuning the

Parafac Decomposition					
Layer	Decomposed	Fine-tuned	Params	Compres.	Impr
CONV4	Acc: 79% Loss : 0.193	Acc: 77% Loss: 0.023	36928	2878	<b>13x</b>
CONV (4+3)	Acc: 77% Loss: 0.082	Acc: 77% Loss: 0.018	18496	2206	<b>8.5x</b>
CONV (4+3+2)	Acc: 76% Loss: 0.300	Acc: 77% Loss: 0.033	9248	732	<b>13x</b>
CONV (4+3+2+1)	Acc: 77% Loss: 0.081	Acc: 77% Loss: 0.033	896	442	<b>2x</b>
CONV2FC1 Rank=170	Loss: 0.738	Acc: 80% Loss: 0.195	1180160	100472	<b>12x</b>
CONV2FC1 Rank=50	<b>Loss: 0.981</b>	<b>Acc: 80%</b> <b>Loss: 0.321</b>	<b>1180160</b>	<b>29912</b>	<b>40x</b>
Overall	<b>Acc: 79%</b>	<b>Acc: 80%</b>	<b>1252480</b>	<b>42922</b>	<b>29x</b>

```

8     self.hiddenLayerSize = hiddens
9     self.outputLayerSize = outputs
10    self.W1 = th.randn(net.inputLayerSize, self.hiddenLayerSize)
11    self.W2 = th.randn(net.hiddenLayerSize, self.outputLayerSize)
12 end

```

## 2.2 Forward Propagation

Table 2.2: Variabili usate nel testo e nel codice

Variabili			
S. Codice	S. Matematico	Definizione	Dimensione
X	$X$	Esempi, 1 per riga	(numEsempi, inputLayerSize)
y	$y$	uscita desiderata	(numEsempi, outputLayerSize)
W1	$W^{(1)}$	Pesi layer 1	(inputLayerSize, hiddenLayerSize)
W2	$W^{(2)}$	Pesi layer 2	(hiddenLayerSize, outputLayerSize)
z2	$z^{(2)}$	Input layer 2	(numEsempi, hiddenLayerSize)
a2	$a^{(2)}$	Uscita layer 2	(numEsempi, hiddenLayerSize)
z3	$z^{(3)}$	Input layer 3	(numEsempi, outputLayerSize)

Nella tabella 2.2 sono elencate le variabili della rete. Gli input dei layer indicati con  $z$  possono anche essere chiamati "attività dei layer" (indicando l'attività sulle loro sinapsi); e  $a^{(2)}$  indica l'uscita del neurone dopo aver applicato la sommatoria e la funzione di attivazione sulle attività provenienti dal layer precedente.

Per muovere i dati in parallelo attraverso la rete si usa la moltiplicazione fra matrici, per questo è molto comodo usare framework che supportano operazioni fra matrici come *Torch*, *Numpy* o *Matlab*. Per prima cosa, gli input del tensore  $X$  devono essere moltiplicati e sommati con i pesi del primo layer  $W^{(1)}$ , ottenendo l'ingresso per l'hidden layer:

$$z^{(2)} = XW^{(1)} \quad (1)$$

Si noti che  $z^{(2)}$  è di dimensione 3x3, essendo  $X$  e  $W^{(1)}$  di dimensione 3x2 e 2x3 rispettivamente.

Ora bisogna applicare la funzione di attivazione a  $z^{(2)}$ . Vi sono diverse funzioni di attivazione utilizzate per le reti neurali. Una delle prime a diventare popolare fu la funzione *sigmoide* [6], utilizzata per questa rete. Vedremo nei capitoli successivi funzioni più efficaci.

$$a^{(2)} = f(z^{(2)}), \quad \text{ove } f = \text{sigmoide} \quad (2)$$

Per completare la *forward propagation*, bisogna seguire lo stesso procedimento per lo strato di output: sommare i contributi provenienti dall'hidden layer ed applicare la sigmoide:

$$z^{(3)} = a^{(2)}W^{(2)} \quad (3)$$

$$\hat{y} = f(z^{(3)}) \quad (4)$$

Essendo  $a^{(2)}$  di dimensione 3x3 e  $W^{(2)}$  3x1 l'output  $\hat{y}$  sarà anch'esso di dimensione 3x1, risultando quindi in una previsione per ogni esempio in ingresso.

Si noti come la moltiplicazione fra matrici renda tutto esprimibile in poche righe di codice.

```

1 --Note: I didn't implement manually the sigmoid function as Torch has one
2   built-in.
3 --define a forward method
4 function Neural_Network:forward(X)
5   --Propagate inputs through network
6   self.z2 = th.mm(X, self.W1) --matrix multiplication
7   self.a2 = th.sigmoid(self.z2)
8   self.z3 = th.mm(self.a2, self.W2)
9   yHat = th.sigmoid(self.z3)
10  return yHat
11 end

```

## 2.3 Backpropagation

Addestrare una rete multistrato con diversi neuroni per strato, ognuno dei quali con uscita non lineare, non è semplice. Fortunatamente, Rumelhart-Hinton-Williams nel 1985 idearono l'algoritmo che è tutt'ora alla base dell'apprendimento delle reti neurali, la *backpropagation of errors*. Non si può introdurre l'algoritmo di "*backprop*" senza prima spiegare il concetto di funzione di costo (o *loss function*). Nelle reti neurali (e più specificatamente nell'apprendimento supervisionato [7], si veda anche sezione

2.5.1), la funzione di costo misura la discrepanza tra l'uscita desiderata e l'effettivo output della rete. È quindi una misura dell'errore della rete, per cui l'obiettivo dell'apprendimento è trovare il minimo di questa funzione (modificando la struttura interna della rete, ovvero i pesi sinaptici). Come per la funzione di attivazione, anche in questo caso ci sono ampie possibilità di scelta [8] a seconda del task su cui la rete viene addestrata.

E di nuovo, come per la funzione di attivazione, si è scelta una delle funzione di costo più popolari: l'errore quadratico medio.

$$J = \sum_{j=1}^n \frac{1}{2}(y - \hat{y})^2 \quad (5)$$

Da cui, il codice in Lua:

```

1 function Neural_Network:costFunction(X, y)
2     --Compute the cost for given X,y, use weights already stored in class
3     self.yHat = self:forward(X)
4     J = 0.5 * th.sum(th.pow((y-yHat), 2))
5     return J
6 end

```

La *backprop* ha alcuni requisiti:

- Reti stratificate;
- Ingressi a valori reali  $\in [0, 1]$ ;
- Neuroni non lineari con funzione di uscita sigmoidale (o altra fz. di attivazione derivabile).

Sotto queste condizioni l'algoritmo sfrutta la regola della catena [9] per la derivazione di funzione composite, per calcolare il gradiente della *funzione di costo*. I pesi della rete vengono quindi aggiornati secondo la *discesa del gradiente* (figura 2.2); ovvero variano in maniera tale da minimizzare la funzione di costo  $J$ .

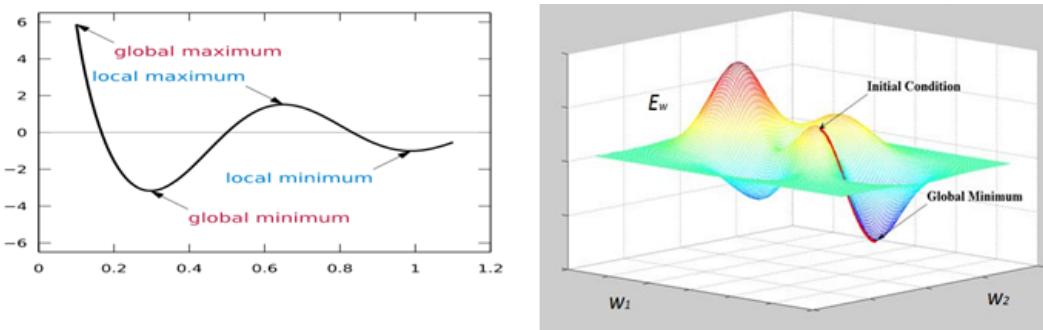


Figura 2.2: Cercare il minimo di una fz. seguendo la discesa del gradiente

Viene chiamato *backward propagation of errors* poiché l'errore calcolato a partire dall'output della rete viene distribuito in maniera proporzionale all'indietro, su tutti i neuroni della rete. È importante quindi, spezzare il calcolo del gradiente dell'errore in derivate parziali, dall'ultimo strato fino al primo, e poi combinarle insieme.

Si noti che le equazioni (1-5) formano una un'unica equazione che lega  $J$  a  $X, y, W^{(1)}, W^{(2)}$ . Tenendo questo in mente, si applica la regola della catena.

Partendo dal fattore riguardante lo strato di output si ha:

$$\frac{\partial J}{\partial W^{(2)}} = \sum \frac{\partial \frac{1}{2}(y - \hat{y})^2}{\partial W^{(2)}}$$

Sviluppando i calcoli si ottiene:

$$\frac{\partial J}{\partial W^{(2)}} = -(y - \hat{y}) \frac{\partial \hat{y}}{\partial W^{(2)}}$$

L'equazione (4) indica che  $\hat{y}$  è la funzione di attivazione di  $z^{(3)}$ . Da cui:

$$\frac{\partial J}{\partial W^{(2)}} = -(y - \hat{y}) \frac{\partial \hat{y}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial W^{(2)}}$$

Il 2 membro dell'equazione è semplicemente la derivata della funzione di attivazione sigmoide (fig. 2.3):

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$f'(z) = \frac{e^{-z}}{(1 + e^{-z})^2}$$

Definiamola, quindi, nel codice:

```

1 function Neural_Network:d_Sigmoid(z)
2     --Derivative of sigmoid function
3     return th.exp(-z):cdiv( (th.pow( (1+th.exp(-z)), 2) ) )
4 end

```

Figures/C2/sigmoidPrime.png

*Figura 2.3:* La funzione sigmoide e la sua derivata

L'equazione così ottenuta è:

$$\frac{\partial J}{\partial W^{(2)}} = -(y - \hat{y}) f'(z^{(3)}) \frac{\partial z^{(3)}}{\partial W^{(2)}}$$

Infine, dobbiamo trovare  $\frac{\partial z^{(3)}}{\partial W^{(2)}}$ , che rappresenta la variazione dell'attività del terzo layer rispetto ai pesi del secondo layer. Richiamando l'equazione (3):

$$z^{(3)} = a^{(2)}W^{(2)}$$

Tralasciando per un attimo la somma tra i vari neuroni, si nota una semplice relazione lineare fra i termini, con  $a^{(2)}$  che rappresenta la pendenza. Indi:

$$\frac{\partial z^{(3)}}{\partial W^{(2)}} = a^{(2)}$$

Indicando con  $\delta^{(3)}$ , l'errore sullo strato di uscita, si ha:

$$\delta^{(3)} = -(y - \hat{y})f'(z^{(3)})$$

Ora bisogna moltiplicare l'errore con  $a^{(2)}$ . Come indicato nelle ottime dispense di CS231 di Stanford [10]: guardare alle dimensioni delle matrici può essere utile in questo caso. Infatti per fare combaciare le dimensioni, c'è solo una maniera di calcolare la derivata qui, ed è facendo la trasposta di  $a^{(2)}$ :

$$\frac{\partial J}{\partial W^{(2)}} = (a^{(2)})^T \delta^{(3)}$$

Si noti che la sommatoria che abbiamo tralasciato all'inizio del calcolo viene inclusa "automaticamente" dalle somme delle moltiplicazione fra matrici.

L'altro termine da calcolare è  $\frac{\partial J}{\partial W^{(1)}}$ . Il calcolo è inizialmente simile a quello precedente, iniziando sempre dalla derivata sull'ultimo strato ed utilizzando i risultati trovati in precedenza:

$$\begin{aligned} \frac{\partial J}{\partial W^{(1)}} &= (y - \hat{y}) \frac{\partial \hat{y}}{\partial W^{(1)}} \\ \frac{\partial J}{\partial W^{(1)}} &= (y - \hat{y}) \frac{\partial \hat{y}}{\partial z^{(3)}} \frac{\partial z^{(3)}}{\partial W^{(1)}} \\ \frac{\partial J}{\partial W^{(1)}} &= -(y - \hat{y}) f'(z^{(3)}) \frac{\partial z^{(3)}}{\partial W^{(1)}} \\ \frac{\partial J}{\partial W^{(1)}} &= \delta^{(3)} \frac{\partial z^{(3)}}{\partial W^{(1)}} \end{aligned}$$

Ora rimane l'ultimo termine da calcolare, anch'esso da scomporre in diversi fattori andando a ritroso nella rete:

$$\frac{\partial z^{(3)}}{\partial W^{(1)}} = \frac{\partial z^{(3)}}{\partial a^{(2)}} \frac{\partial a^{(2)}}{\partial W^{(1)}}$$

Come prima, c'è una relazione lineare tra le sinapsi, ma stavolta la pendenza è data da  $W^{(2)}$ ; anche in questo caso da trasporre.

$$\frac{\partial J}{\partial W^{(1)}} = \delta^{(3)} (W^{(2)})^T \frac{\partial a^{(2)}}{\partial W^{(1)}}$$

$$\frac{\partial J}{\partial W^{(1)}} = \delta^{(3)}(W^{(2)})^T \frac{\partial a^{(2)}}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial W^{(1)}}$$

$\frac{\partial a^{(2)}}{\partial z^{(2)}}$  è di nuovo la derivata della  $f$  di attivazione. Il termine finale del calcolo  $\frac{\partial z^{(2)}}{\partial W^{(1)}}$ , rappresenta quanto varia l'uscita del primo strato al variare dei pesi. Richiamando l'equazione (1) si nota subito che questo valore è dato dal vettore di input  $X$  - come prima - traposto:

$$\frac{\partial J}{\partial W^{(1)}} = X^T \delta^{(3)}(W^{(2)})^T f'(z^{(2)})$$

Chiamando  $\delta^{(2)} = \delta^{(3)}(W^{(2)})^T f'(z^{(2)})$  diventa:

$$\frac{\partial J}{\partial W^{(1)}} = X^T \delta^{(2)}$$

Facendo un sommario:

$$\boxed{\frac{\partial J}{\partial W^{(2)}} = (a^{(2)})^T \delta^{(3)}} \quad (6)$$

$$\boxed{\frac{\partial J}{\partial W^{(1)}} = X^T \delta^{(2)}} \quad (7)$$

$$\boxed{\delta^{(2)} = \delta^{(3)}(W^{(2)})^T f'(z^{(2)})} \quad (8)$$

$$\boxed{\delta^{(3)} = -(y - \hat{y}) f'(z^{(3)})} \quad (9)$$

Implementando le equazioni sovrascritte in Lua, la classe `Neural_Network` è quindi completa (per i dettagli si veda l'appendice A).

```

1 function Neural_Network:d_CostFunction(X, y)
2   --Compute derivative wrt to W1 and W2 for a given X and y
3   self.yHat = self:forward(X)
4   delta3 = th.cmul(-(y-self.yHat), self:d_Sigmoid(self.z3))
5   dJdW2 = th.mm(self.a2:t(), delta3)
6
7   delta2 = th.mm(delta3, self.W2:t()):cmul(self:d_Sigmoid(self.z2))
8   dJdW1 = th.mm(X:t(), delta2)
9
10  return dJdW1, dJdW2
11 end

```

## 2.4 Verifica numerica del gradiente

Siccome la backprop è notoriamente difficile da debuggare una volta che la si usa per l'addestramento di una rete, bisogna controllare se l'implementazione della sezione precedente è corretta prima di proseguire nel progetto. A questo scopo, è stata scritta una funzione per il calcolo *numerico* del gradiente che andrà poi confrontata con il calcolo computato dalla Backprop.

L'algoritmo [11] è basato sulla seguente definizione di derivata:

$$\frac{d}{d\theta} J(\theta)) = \lim_{\epsilon \rightarrow 0} \frac{J(\theta + \epsilon) - J(\theta - \epsilon)}{2 * \epsilon}$$

Il gradiente che bisogna controllare è formato dai 2 vettori che contengono le derivate dei pesi di tutta la rete:  $\frac{\partial J}{\partial W^{(1)}} \quad e \quad \frac{\partial J}{\partial W^{(2)}}$ .

Approssimando  $\epsilon$  con un valore molto piccolo (*i.e.*  $10^{-4}$ ) è possibile perturbare singolarmente i pesi della rete ( contenuti nei vettori  $W^{(1)}$  e  $W^{(2)}$  ) e calcolare così il gradiente in maniera numerica.

Per poterlo fare, servono delle funzioni ausiliare (metodi getter e setter) per prendere e settare i pesi ed il gradiente della rete come singolo vettore "flat" di parametri (appendice A). Dopodiché basta un loop ed un array per memorizzare le derivate dei singoli pesi così calcolati:

```

1 function computeNumericalGradient(NN, X, y)
2     paramsInitial = NN:getParams()
3     numgrad = th.zeros(paramsInitial:size())
4     perturb = th.zeros(paramsInitial:size())
5     e = 1e-4
6
7     for p=1,paramsInitial:nElement() do
8         --Set perturbation vector
9         perturb[p] = e
10        NN:setParams(paramsInitial + perturb)
11        loss2 = NN:costFunction(X, y)
12
13        NN:setParams(paramsInitial - perturb)
14        loss1 = NN:costFunction(X, y)
15
16        --Compute Numerical Gradient
17        numgrad[p] = (loss2 - loss1) / (2*e)
18
19        --Return the value we changed to zero:
20        perturb[p] = 0
21    end
22
23    --Return Params to original value:
24    NN:setParams(paramsInitial)
25    return numgrad
26 end

```

Si può ora inizializzare una rete, eseguire la backpropagation e confrontare i valori con quelli calcolati numericamente:

```

1 --test if we actually make the calculations correctly
2 NN = Neural_Network(2,3,1)
3
4 print('Gradient checking...')
5 numgrad = computeNumericalGradient(NN, X, y)
6 grad = NN:computeGradients(X, y)
7 --[[
8 In order to make an accurate comparison of the 2 vectors
9 we can calculate the difference as the ratio of:
10 numerator --> the norm of the difference
11 denominator--> the norm of the sum
12 Should be in the order of 10^-8 or less
13 --]]
14 diff = th.norm(grad-numgrad)/th.norm(grad+numgrad)
15 print(string.format('The difference is %e',diff))

```

Per calcolare *quanto* siano effettivamente uguali i due gradienti si può usare un rapporto basato sulla norme della somma e della differenza dei gradienti (si veda il codice sopra). Se la backprop è stata implementata correttamente questa differenza dovrebbe essere nell'ordine di  $10^{-8}$  o inferiore. Difatti, quando si esegue lo script si ottiene:

```

1 $ th 4_gradCheck.lua
2 Gradient checking...

```

```
3 The difference is 2.123898e-10
```

## 2.5 Addestramento

Una volta accertati che l'implementazione della Backprop è corretta si può procedere ad addestrare la rete. Quello che si vuole ottenere è una rete che guardando ai dati accumulati negli anni, riesca a prevedere l'andamento del profitto del ristorante. Nel nostro dataset, ogni esempio è formato da una coppia  $\langle n. \ coperti, n. \ ore \ settimanali \rangle$  a cui è associata un'uscita *desiderata*. La rete cercherà di modificare la sua struttura interna (i.e. i pesi sinaptici) "creando" una funzione - la rete stessa rappresenta questa funzione - per riprodurre in maniera più precisa possibile quest'associazione. L'apprendimento sarà quindi di tipo *supervisionato*.

### 2.5.1 Apprendimento supervisionato

Con questo termine s'intende l'allenamento di un sistema tramite una serie di esempi ideali; l'insieme di questi esempi è chiamato *training set*. Il sistema impara quindi ad approssimare una funzione non nota a priori a partire da una serie di copie ingresso-uscita: per ogni input in ingresso gli si comunica l'output desiderato. L'apprendimento consiste nella capacità del sistema – tramite la funzione generata con l'allenamento – di generalizzare a nuovi esempi: avendo in ingresso dati non noti deve poter predire in modo corretto l'output desiderato. Matematicamente parlando, questi esempi non noti sono punti del dominio che non fanno parte dell'insieme degli esempi di training. Esistono diversi algoritmi di apprendimento supervisionato ma tutti condividono una caratteristica: l'addestramento viene eseguito mediante la minimizzazione di una funzione di costo (si veda la sezione 2.3). Nella *Learning Rule* in

## Fase di addestramento

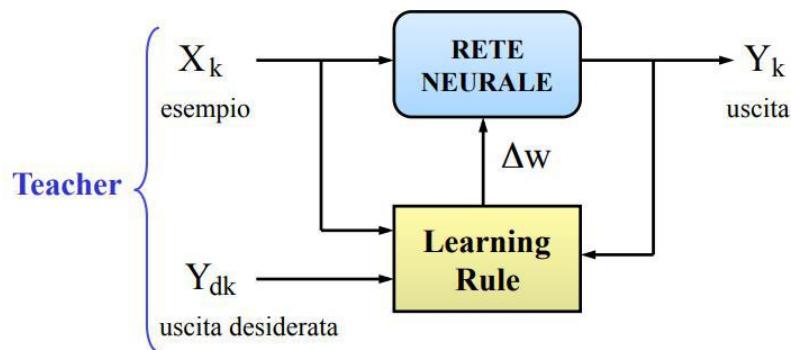


Figura 2.4: Apprendimento supervisionato: schema generale

figura 2.4 sono compresi 2 elementi:

1. Calcolo della discrepanza tra l'uscita della rete e l'uscita desiderata e backpropagation;

2. La strategia di aggiornamento dei parametri.

Riguardo all'ultimo punto ci sono diverse possibilità.

### 2.5.2 Discesa del gradiente

La loss function è una funzione che ha un numero di variabili pari al numero dei pesi della rete. Data la complessità - soprattutto in reti profonde molto più complesse di quella trattata in questo progetto - la ricerca del minimo non è semplice; si corre il rischio di rimanere bloccati in plateau o minimi locali. Per questo, si applicano metodi a raffinamento iterativo: si parte da una soluzione iniziale e si cerca di migliorarla ad ogni ciclo. Questo metodo è conosciuto come *Stochastic Gradient Descent* [12]. Calcolando il gradiente  $\nabla J$  si conosce la direzione di massima variazione quindi ci si sposta - lungo l'opposto di questa direzione, l'antigradiente - di una quantità pari a  $\eta$ . Questo parametro si chiama *learning rate* e regola appunto la velocità dell'apprendimento.

Tornando alla strategia di aggiornamento dei parametri, la più intuitiva è chiamata "*Vanilla*":

$$W_{t+1} = W_t - \eta \nabla J(W_t) \quad (2.1)$$

Questa regola di apprendimento però soffre di alcuni problemi:

- Effettua uno spostamento pari a  $\eta$  sia per features frequenti che non. Questo problema è noto come "*sparsità delle features*".
- $\eta$  è una costante e non è detto che garantisca la convergenza. Si potrebbe "saltare" da un lato all'altro del punto di minimo senza mai trovarlo.
- Come detto sopra, i punti di sella e plateau causano problemi. In questo caso il gradiente è nullo e quindi l'aggiornamento dei pesi si azzera, fermando l'apprendimento.

Per ovviare a questo ed altri problemi, sono stati studiati numerosi metodi. La prossima sezione ne elenca qualcuno utilizzato per questo progetto.

## 2.6 Ottimizzazione: diverse tecniche

L'ottimizzazione dell'apprendimento delle reti neurali è un argomento vasto ed imperioso, ma molto importante. Con addestramenti che possono durare mesi a seconda del tipo di apprendimento, sono nate, in relativamente breve tempo, moltissimi metodi di ottimizzazione. Qui si faranno solo dei cenni ai metodi utilizzati in questo progetto.

Prima di elencare tecniche più evolute dell'aggiornamento Vanilla, occorre precisare alcuni punti dello Stochastic Gradient Descent, essendo quest'ultimo il punto di partenza di ogni altro metodo.

- SGD: lo *Stochastic Gradient Descent* è, come lo descrive il nome stesso, una versione stocastica della discesa del gradiente. La discesa del gradiente standard non è scalabile, poiché il gradiente da calcolare tiene conto dell'errore quadratico calcolato *su ogni singolo esempio del dataset*. Come detto poco fa, la loss function ha già di per sé un numero di variabili proporzionale alla complessità della rete; quando il dataset è molto largo, ed è tipico per problemi reali di machine learning, questo calcolo diventa inefficiente. L'SGD risolve questo problema approssimando l'operazione "*gradiente → aggiornamento pesi*" da tutto il dataset

al singolo esempio. Questo rende l'approssimazione molto inaccurata, ma molto veloce da elaborare. Nonostante le oscillazioni dovute all'inaccuratezza, questo metodo consente nella pratica, dopo molte iterazioni, di trovare il minimo globale. Questo è soprattutto vero per problemi su larga scala.

Un compromesso tra il metodo standard e quello completamente stocastico è l'utilizzo di "*mini-batch*", cioè piccoli sottoinsiemi del dataset su cui viene eseguita l'iterazione di apprendimento. Se il dataset è abbastanza eterogeneo ed è ordinato in maniera randomica, il mini-batch approssima abbastanza bene l'intero dataset; di conseguenza, l'approssimazione sarà più verosimile al calcolo dell'intero gradiente, aumentando quindi l'accuracy senza intaccare la velocità del calcolo.

- **MOMENTUM & NAG:** Il "*Momentum*" controlla la quantità d'inerzia nella modifica dei pesi sinaptici, memorizzando nell'equazione la variazione  $\Delta W$  precedente:

$$W_t = W_{t-1} - \eta \nabla J(W_t) - \underbrace{\mu \nabla J(W_{t-1})}_{\text{momentum step}}$$

In questo modo si riducono le oscillazioni nella ricerca della soluzione permettendo di usare learning rate più alti. È ispirato dal momento nella Fisica, considerando il vettore dei pesi come una particella che viaggia in uno spazio parametrico e acquisisce velocità nella discesa.

Il *Nesterov Accelerated Gradient* è una versione più raffinata del Momentum update, che elabora una correzione della traiettoria calcolando il gradiente *dopo* aver fatto la somma con il gradiente accumulato precedentemente. Per aiutare a capire la differenza tra i due, si faccia riferimento alla figura 2.5.

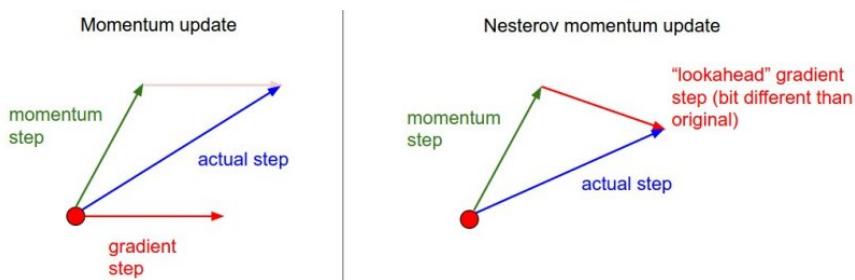


Figura 2.5: Confronto tra il momentum classico e NAG

- **BFGS:** l'algoritmo *Broyden-Fletcher-Goldfarb-Shanno* fa parte della famiglia "Quasi-Newton" [13]. Questo metodo supera i limiti della discesa del gradiente classica facendo una stima della matrice Hessiana, ovvero della curvatura della superficie della funzione di costo  $J$ . Utilizzando questa stima compie degli spostamenti più informati verso la discesa. Nella pratica si usa una versione efficiente che consuma meno memoria, detta Limited-BFGS [14].
- **ADAM:** l'*Adaptive Moment Estimation* cerca di ovviare ai problemi dell'aggiornamento vanilla modificando il learning rate in maniera diversa per ogni parametro e a seconda dello stadio dell'apprendimento. Fa parte quindi della famiglia dei metodi *adattivi*. In particolare, l'algoritmo calcola la media con decadimento esponenziale del gradiente e del quadrato del gradiente; i parametri  $\beta_1$  e  $\beta_2$  controllano il decadimento di queste medie mobili. Gli autori forniscono i valori consigliati per questi parametri, che sono infatti i valori di default anche in ogni framework che supporta Adam [15].

Per riuscire effettivamente ad addestrare la rete, si è utilizzato un package di ottimizzazione di Torch: `optim`. Quest'ultimo fornisce il supporto a tutte (e più) le tecniche di ottimizzazione viste prima. Si può quindi procedere all'addestramento della rete. Utilizzando un *logger* per mantenere tutti i dati sul training, si possono successivamente plottare le diverse curve di apprendimento per confrontare i diversi metodi. Definita quindi una classe `Trainer`, si definisce un metodo per addestrare la rete che astrae dalla tecnica di ottimizzazione utilizzata.

```

1 Trainer = class('Trainer')
2 function Trainer:__init__(NN)
3     --Make Local reference to network:
4     self.N = NN
5 end
6
7 --Let's train!
8 function Trainer:train(X, y)
9     --variables to keep track of the training
10    local neval = 0
11    --get initial params
12    params0 = self.N:getParams()
13    -- create closure to evaluate f(X) and df/dX
14    -- this is requested by the API of the optim package
15    local feval = function(params0)
16        local f = self.N:costFunction(X, y)
17        print(f)
18        local df_dx = self.N:computeGradients(X, y)
19        neval = neval + 1
20        logger:add{neval, f} --,timer:time().real}
21        return f, df_dx
22    end
23    if optimMethod == optim.cg then
24        newparams, _, _ = optimMethod(feval, params0, optimState)
25    else
26        for i=1,opt.maxIter do
27            newparams, _, _ = optimMethod(feval, params0, optimState)
28            self.N:setParams(newparams)
29        end
30    end
31 end

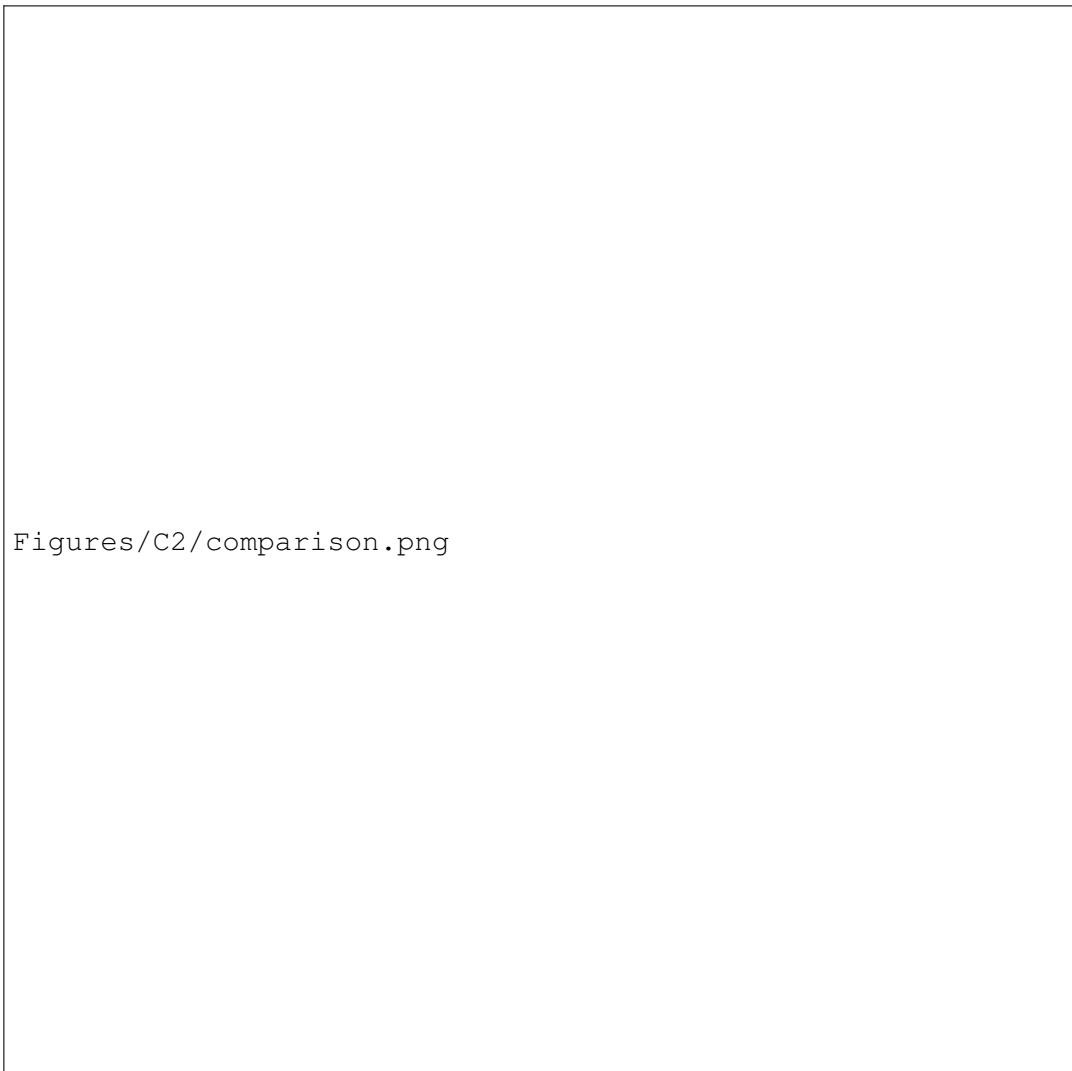
```

Procedendo iterativamente con le diverse tecniche di ottimizzazione si ottiene il risultato in figura 2.6.

Dal grafico si evidenzia come le premesse teoriche dei vari metodi siano state pienamente rispettate. Nonostante il problema sia molto semplice se comparato ai reali problemi di *deep learning*, già su un dataset ed un numero di iterazioni così limitato si nota la superiorità di un metodo verso il precedente. In particolare, nel corso "CS231" di deep learning applicato alla visione artificiale di Stanford [16], sviluppato da Andrej Karpathy et. al, si raccomanda Adam come metodo di default per applicazioni di deep learning:

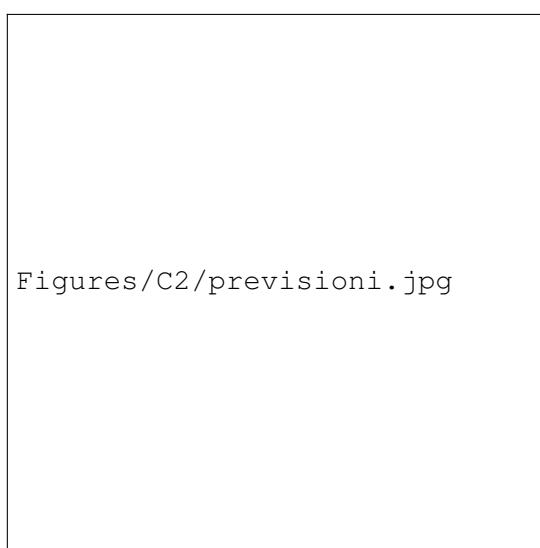
*In practice Adam is currently recommended as the default algorithm to use. However, it is often also worth trying SGD+Nesterov Momentum as an alternative.*

La rete è ora addestrata: dando in ingresso la matrice  $X$  di input si avranno in output previsioni molto più accurate, come mostrato in figura 2.7.



Figures/C2/comparison.png

*Figura 2.6:* Confronto dei metodi di ottimizzazione durante il training



Figures/C2/previsioni.jpg

*Figura 2.7:* L'output della rete  $\hat{y}$  è vicino all'output desiderato  $y$

## 2.7 Overfitting

Uno dei problemi più comuni dell'apprendimento automatico è *l'Overfitting*. Esso si verifica laddove il modello creato per fare predizioni risulta troppo complesso e troppo "legato" al solo training set, di cui apprende anche i rumori: la rete è quindi incapace di *generalizzare* ad esempi ancora non visti e, nonostante l'accuratezza estremamente alta sul training set, darà delle pessime predizioni. Questo può essere dovuto a diversi fattori, di cui i più classici sono:

- il modello ha troppi parametri rispetto al numero di osservazioni;
- il dataset è costituito da troppi pochi esempi;
- l'addestramento è stato fatto troppo a lungo.

In figura 2.8 è mostrato un esempio di 2 modelli statistici con complessità molto diverse. In un primo istante la curva polinomiale avrà risultati decisamente migliori sul training set, salvo poi veder capovolgere la situazione sul test set.



*Figura 2.8:* La funzione polinomiale ha un errore nullo sul dataset laddove la funzione lineare invece lo ha del 100%. Tuttavia, la curva è eccessivamente complessa ed affetta da rumore; avrà quindi cattive capacità di generalizzazione. La retta, al contrario, approssima molto meglio i punti della distribuzione sottostante. Se definiamo questi punti come il test set, allora la retta avrà un'accuratezza maggiore.

### 2.7.1 Rilevare l'overfitting

Per rilevare se è avvenuto o meno overfitting durante l'apprendimento; si possono dare in ingresso al percettrone esempi ancora non visti e controllare "ad occhio" se le predizioni hanno senso. Per averne la certezza però, bisogna dividere il dataset in un *training set* ed un *test set* su cui testare la rete durante l'apprendimento. Sugli esempi del *test set* non si farà backpropagation e non avverrà quindi apprendimento; si utilizzeranno solo per sapere quanto è corretto l'addestramento.

Testando la rete *durante* l'apprendimento si può capire il preciso momento in cui avviene l'overfitting. Per farlo, occorre plottare sullo stesso grafico l'errore sul training set e sul test set.

Dopo aver definito arbitrariamente alcuni esempi di testing, si modifica l'algoritmo di training:

```

1 --Need to modify trainer class a bit to check testing error during
2   training:
3 function Trainer:train(trainX, trainY, testX, testY)
4   --variables to keep track of the training
5   local neval = 0
6
7   params0 = self.N:getParams()
8   -- create closure to evaluate f(X) and df/dX
9   local feval = function(params0)
10      local f = self.N:costFunction(trainX, trainY)
11      local test = self.N:costFunction(testX, testY)
12      --printing training and testing error
13      print(f..' '..test)
14      local df_dx = self.N:computeGradients(trainX, trainY)
15      neval = neval + 1
16      --logging both training and testing data
17      logger:add{neval, f} --,timer:time().real}
18      testLogger:add{neval, test}
19
20      return f, df_dx
21  end
22
23  if optimMethod == optim.cg then
24    newparams,_,_ = optimMethod(feval, params0, optimState)
25  else
26    for i=1,opt.maxIter do
27      newparams,_,_ = optimMethod(feval, params0, optimState)
28      self.N:setParams(newparams)
29    end
30  end
31 end

```

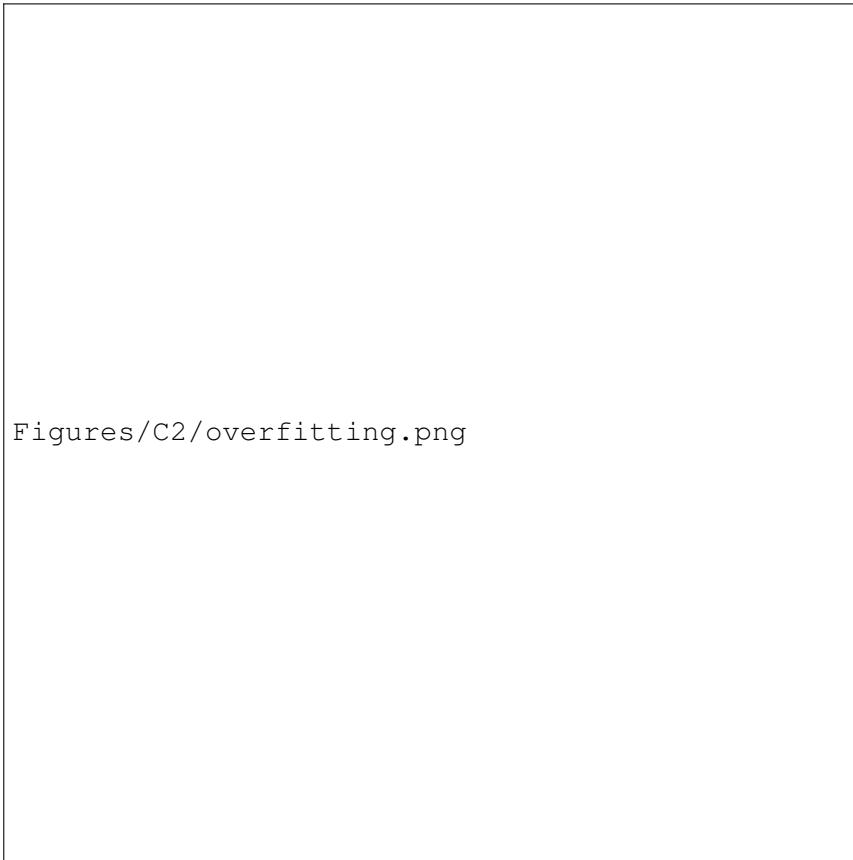
Si addestra la rete e si loggano i valori di training e di testing:

```

1 --now let's train and check where exactly the net is Overfitting
2 nn = Neural_Network(2,3,1)
3
4 init_params = nn:getParams()
5 logtrain = 'train.log'
6 logtest = 'test.log'
7 logger = optim.Logger(logtrain)
8 testLogger = optim.Logger(logtest)
9
10 trainer = Trainer(nn)
11 trainer:train(trainX, trainY, testX, testY)

```

Andando infine a plottare i dati così ottenuti, si osserva la presenza di overfitting in figura 2.9.



Figures/C2/overfitting.png

*Figura 2.9:* La rete mostra overfitting. Dopo l'iterazione 100 le performance sul test set iniziano ad essere sbagliate. Attorno alla 250 il modello diventa troppo complesso e le predizioni sono pessime

### 2.7.2 Contromisure

Vi sono diverse contromisure contro l'overfitting, che seguono le cause più comuni:

- *Dataset più ampio:* una regola empirica consiste nell'avere circa 10 esempi per ogni parametro della rete. Nel caso del MLP di questo progetto i pesi sono 9, quindi secondo questa regola sono necessari 90 esempi. È ovviamente impossibile, poiché Loris ha collezionato solamente un limitato numero di esempi.
- *K-fold cross-validation:* si divide il dataset in K set di eguale misura, ad ogni iterazione uno dei set viene utilizzato come test-set ed il resto per formare il training e validation set (si veda lez.1 del corso di Machine Learning per PhD di M. LippiWLippi). Anche questa procedura non è applicabile in questo caso.
- *Dropout:* il dropout è una tecnica di regolarizzazione maggiormente utilizzata nelle reti profonde, nella quale ad ogni iterazione si "spengono" alcuni neuroni (settando i pesi che li collegano a 0). Questo fa sì che neuroni vicini non si specializzino tutti a riconoscere le stesse caratteristiche, risultando in un minore overfitting. Una spiegazione approfondita esula dallo scopo di questo elaborato, per cui si suggerisce la consultazione del paper di Srivastava, Hinton et. al [17].
- *Arresto anticipato:* una volta rilevato quando avviene precisamente l'overfitting, si arresta l'apprendimento *prima* di quel momento. Osservando la figura 2.9 si potrebbe pensare di fermare le iterazioni alla n.100 (o comunque prima della

n. 250) e funzionerebbe. Così facendo, si ha una maniera piuttosto semplice di risolvere il problema. C'è uno svantaggio però: se la rete non è ancora ben addestrata non c'è possibilità di migliorarla ulteriormente, poiché ogni nuova iterazione causerebbe overfitting. Per questo, è stato usato l'ultimo metodo qui presentato.

- *Regolarizzazione*: consiste nell'aggiungere un termine alla funzione di costo  $J$  che penalizza modelli troppo complessi [18]. Questo termine,  $\lambda R(W)$  è regolato da un parametro  $\lambda$  che definisce l'intensità della regolarizzazione.

Un comune metodo di regolarizzazione è la *L2 Regularization*: si implementa aggiungendo alla funzione di costo il quadrato dei tensori dei pesi della rete. In questa maniera si tengono i valori di tutti i pesi piuttosto bassi evitando che si specializzino troppo sui dati del training set. Di seguito il codice dei metodi della rete che devono essere modificati:

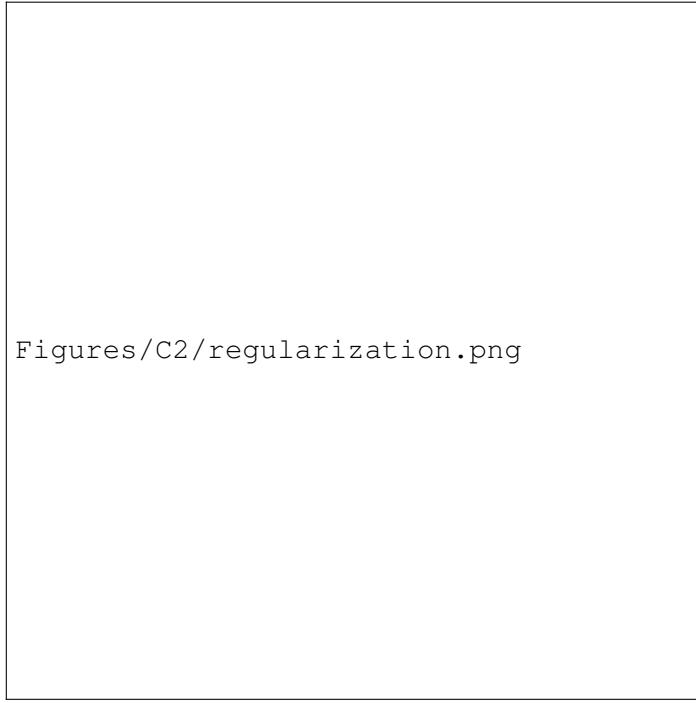
```

1 --[ [
2 ## Introducing a Regularization term to mitigate overfitting ##
3 Lambda will allow us to tune the relative cost:
4 higher values of Lambda --> bigger penalties for high model complexity
5 --]]
6
7 --so, the new Neural_Network class now becomes:
8 Neural_Network = class(function(net, inputs, hiddens, outputs, lambda)
9     net.inputLayerSize = inputs
10    net.hiddenLayerSize = hiddens
11    net.outputLayerSize = outputs
12    net.W1 = th.randn(net.inputLayerSize, net.hiddenLayerSize)
13    net.W2 = th.randn(net.hiddenLayerSize, net.outputLayerSize)
14
15    --regularization parameter
16    net.lambda = lambda
17 end)
18
19 function Neural_Network:costFunction(X, y)
20     --Compute the cost for given X,y, use weights already stored in class
21     self.yHat = self:forward(X)
22     J = 0.5 * th.sum(th.pow((y-self.yHat),2))/X:size()[1] +
23         (self.lambda/2) * (th.sum(th.pow(self.W1,2)) + th.sum(th.pow(
24             self.W2, 2)))
25
26     return J
27 end
28
29 function Neural_Network:d_CostFunction(X, y)
30     --Compute derivative wrt to W and W2 for a given X and y
31     self.yHat = self:forward(X)
32     delta3 = th.cmul(-(y-self.yHat), self:d_Sigmoid(self.z3))
33     --Add gradient of regularization term:
34     dJdW2 = th.mm(self.a2:t(), delta3)/X:size()[1] + self.lambda*self.W2
35
36     delta2 = th.mm(delta3, self.W2:t()):cmul(self:d_Sigmoid(self.z2))
37     --Add gradient of regularization term:
38     dJdW1 = th.mm(X:t(), delta2)/X:size()[1] + self.lambda*self.W1
39
40     return dJdW1, dJdW2
41 end

```

Eseguendo ora l'addestramento si ottengono risultati in figura 2.10. Si può osservare come la regolarizzazione sia stata efficace e le 2 curve siano pressoché identiche. Si noti anche che, data la modesta complessità del problema, dopo circa la 130esima

iterazione si è già raggiunto un minimo ed il processo di apprendimento non migliora ulteriormente.



*Figura 2.10:* Dopo aver applicato la L2 regularization, la rete non è più affetta da overfitting.

## 2.8 Risultati

In questo capitolo si è visto come costruire da zero l’architettura di un percettrone multi-strato con paradigma OOP; come implementare la backpropagation e controllare numericamente che funzioni in maniera corretta. Successivamente si sono visti i principali algoritmi per ottimizzare l’apprendimento supervisionato di una rete neurale. Si è poi affrontato il problema dell’overfitting: come rilevarlo in maniera precisa ed un’introduzione alle tecniche più comuni per risolverlo.

Infine, si è ottenuto una rete neurale capace di predire in maniera corretta il profitto di un ristorante in base al numero di coperti ed il numero di ore di apertura settimanali.

## Chapter 3

# Reti Neurali Convoluzionali

In questo capitolo si introduce una panoramica generale sulle reti neurali convoluzionali. Essendo un argomento vasto, una trattazione teorica approfondita sarebbe materia di una tesi di laurea, ragion per cui gli argomenti sono introdotti con lo scopo di avere un'infarinatura per comprendere le applicazioni sviluppate nei capitoli successivi.

### 3.1 Breve introduzione

Le reti neurali convoluzionali, alle quali ci riferiremo con l'abbreviazione *CNN* - dall'inglese *Convolutional Neural Network*, sono un'evoluzione delle normali reti artificiali profonde caratterizzate da una particolare architettura estremamente vantaggiosa per compiti visivi (e non), che le ha rese negli anni molto efficaci e popolari. Sono state ispirate dalle ricerche biologiche di Hubel e Wiesel i quali, studiando il cervello dei gatti, avevano scoperto che la loro corteccia visiva conteneva una complessa struttura di cellule. Quest'ultime erano sensibili a piccole parti locali del campo visivo, detti campi recettivi (*receptive fields*). Agivano quindi da filtri locali perfetti per comprendere la correlazione locale degli oggetti in un'immagine. Essendo questi sistemi i più efficienti in natura per la comprensione delle immagini, i ricercatori hanno tentato di simularli.

### 3.2 Architettura

Le CNN sono reti neurali profonde costituite da diversi strati che fungono da estrattori delle features ed una rete completamente connessa alla fine, che funge da classificatore, come raffigurato in figura 3.1.

Questi strati in cui si estraggono le caratteristiche delle immagini sono detti strati di convoluzione e sono generalmente seguiti da una funzione non lineare e un passo di *pooling*. Vi possono poi essere degli strati di elaborazione dell'immagine, come quello di normalizzazione del contrasto, si veda la figura 3.2.

Convoluzione e pooling hanno come scopo quello di estrarre le caratteristiche, mentre l'unità non lineare serve a rafforzare le caratteristiche più forti e indebolire quelle meno importanti, ovvero quelle che hanno stimolato meno i neuroni (si dice che fa da “squashing”). Sempre dalla figura 3.1, possiamo inoltre notare che, per ogni immagine in input, corrispondono nei vari strati, diversi gruppi di immagini, che vengono chiamate *feature maps*. Le feature maps sono il risultato dell'operazione di convoluzione svolta tramite un banco di filtri, chiamati anche *kernel*, che altro non sono che delle matrici con dei valori utili a ricercare determinate caratteristiche nelle immagini.

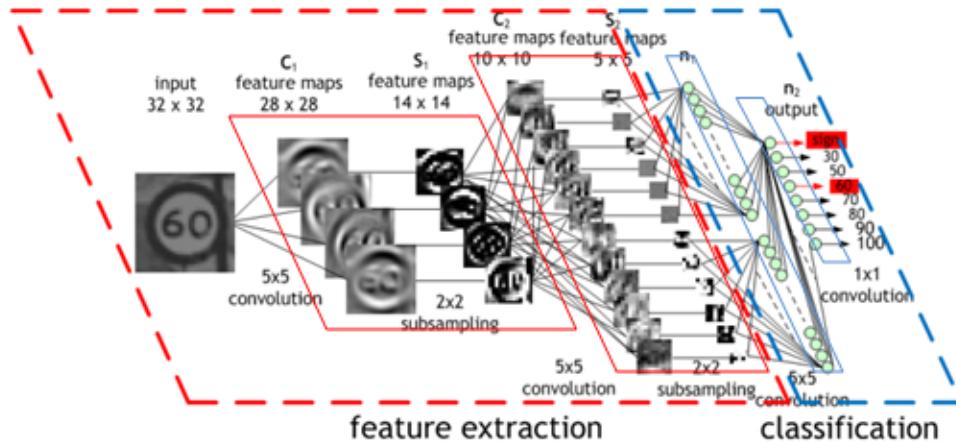


Figura 3.1: Architettura di una CNN che classifica segnali stradali: si evidenzia la divisione tra gli strati che fungono da feature extractor ed il classificatore finale

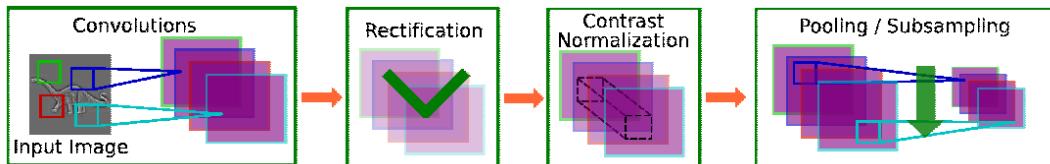


Figura 3.2: I diversi strati tipici di una CNN

Infine, terminati i convolutional layers, le feature maps vengono “srotolate” in vettori e affidate ad una rete neurale "classica" che esegue la classificazione finale.

Il numero di strati di convoluzione è arbitrario. Inizialmente, quando le CNN divennero famose grazie a Y. LeCun, che addestrò una CNN chiamata "*LeNet5*" al riconoscimento dei numeri [19], questo numero era compreso tra 2-5. Nel 2012, Alex Krizhevsky et al [20] addestrarono una rete costituita da 5 strati di convoluzione, 60 milioni di parametri e 650 mila neuroni. Ottennero la migliore percentuale d'errore al mondo sul dataset ImageNet ILSVRC-2010, contenente 1,2 milioni di immagini divise in 1000 categorie.

Da allora le cose si sono evolute con una velocità disaralente, e l'ImageNet challenge del 2015, è stata vinta da una rete con 152 strati [21]. Nel capitolo 5 si farà un confronto tra quest'ultima rete, soprannominata "*ResNet*" e la capostipite LeNet5 su un task di classificazione.

### 3.2.1 Strato di Convoluzione

Per comprendere appieno quello che avviene in una CNN, occorre introdurre il concetto di convoluzione fra matrici, e capire come questo sia importante per applicare dei filtri ad un'immagine digitale.

Un'immagine digitale può essere considerata come una matrice A di dimensione MxN valori reali o discreti. Ogni valore della matrice prende il nome di pixel e i suoi indici sono anche chiamati coordinate: ogni pixel  $A(m, n)$  rappresenta l'intensità nella posizione indicata dagli indici.

Si definisce “filtro” o “kernel” una trasformazione applicata ad un'immagine. Come detto prima, questi filtri sono a loro volta delle matrici; la trasformazione quindi si

effettua appunto tramite un'operazione di convoluzione tra l'immagine in ingresso ed il filtro. La convoluzione, discreta nel caso di immagini digitali, si può definire come:

$$y[m, n] = x[m, n] \otimes h[m, n] = \sum_{i=0}^m \sum_{j=0}^n x[i, j] \times h[m - i, n - j]$$

Ogni pixel di  $y[m, n]$  è così il risultato di una somma pesata tramite  $h[m, n]$  della sottoregione che ha centro nel pixel indicato dalle coordinate m,n. Un esempio di convoluzione è rappresentato in figura 3.3.

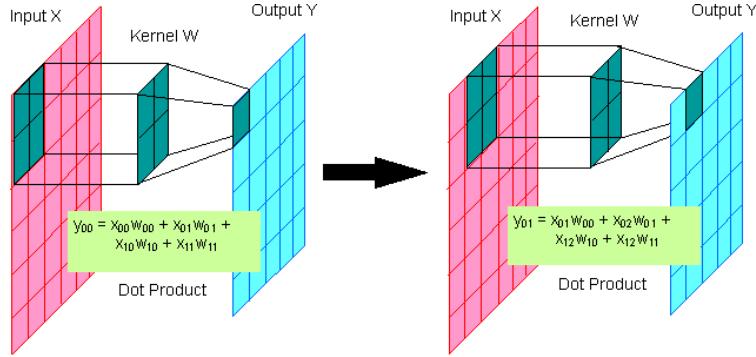


Figura 3.3: Convoluzione con un kernel: primi due step

Nei convolutional layers viene quindi fatta un'operazione di convoluzione tra l'immagine/i in ingresso e un numero arbitrario K di filtri. Questi filtri hanno valori tali da ottenere in uscita un riconoscimento di determinate caratteristiche.

I valori dei filtri sono all'inizio scelti casualmente, e vengono poi migliorati ad ogni iterazione mediante l'algoritmo di backpropagation, visto nel Capitolo ???. Così facendo, la rete addestra i suoi filtri ad estrarre le features più importanti degli esempi del training set; per cui cambiando training set i valori dei filtri saranno diversi. Ad esempio, i valori dei filtri di una rete allenata con immagini di pali verticali saranno diversi da quella allenata con immagini di palloni da calcio; nel primo caso i valori saranno calibrati per riconoscere lunghi orientamenti verticali, mentre nel secondo per riconoscere oggetti sferici.

Nelle reti convoluzionali quindi, la backpropagation migliora i valori dei filtri della rete, è lì quindi che si accumula l'apprendimento. I neuroni, in queste reti, devono intendersi come i singoli filtri.

Vi sono diversi *hyperparameters* da settare manualmente negli strati di convoluzione:

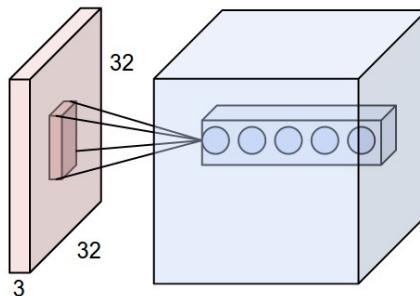
1. la misura del filtro  $F$ : chiamato anche *receptive field*. Ogni filtro cerca una determinata caratteristica in un'area locale dell'immagine, la sua misura quindi è il campo recettivo del singolo neurone. Tipicamente sono 3x3, 5x5 o 7x7.
2. Il numero  $K$  di filtri: per ogni strato, questo valore definisce la profondità dell'output dell'operazione di convoluzione. Infatti, mettendo una sopra l'altra le feature maps, si ottiene un cubo in cui ogni "fetta" è il risultato dell'operazione tra l'immagine in ingresso ed il corrispettivo filtro. La profondità di questo cubo dipende appunto dal numero dei filtri.
3. Lo "*Stride*" S: definisce di quanti pixel si muove il filtro della convoluzione ad ogni passo. Se lo stride è settato a 2, il filtro salterà 2 pixel alla volta, producendo quindi un output più piccolo.

4. Il "padding"  $P$ : definisce la misura con la quale si vuole aggiungere degli "0" all'input per preservare la dimensione in output. In generale, quando lo stride  $S=1$ , un valore di  $P = (F-1)/2$  garantisce che l'output avrà le stesse dimensioni dell'input.

Quando si elaborano delle immagini con le CNN si hanno generalmente in ingresso degli input tridimensionali, caratterizzati dall'altezza  $H_1$ , l'ampiezza  $W_1$  e il numero di canali di colore  $D_1$ . Conoscendo i parametri sopra specificati si può calcolare la dimensione dell'output di un layer di convoluzione:

$$\begin{aligned} H_2 &= (H_1 - F + 2P)/S + 1 \\ W_2 &= (W_1 - F + 2P)/S + 1 \\ D_2 &= K \end{aligned}$$

A questo proposito si osservi un esempio di "volume di neuroni" del primo strato di convoluzione in figura 3.4. Ogni neurone è collegato spazialmente solo ad 1 regione locale dell'input ma per tutta la profondità (i.e. i 3 canali del colore). Si noti che ci sono 5 neuroni lungo la profondità e tutti guardano alla stessa regione dell'input.



*Figura 3.4:* Ogni neurone è connesso a solo 1 regione locale dell'input ma a tutta la profondità (i.e. canali colori). La depth dell'output è data dal numero  $K$  di filtri, in questo caso 5

Gli strati di convoluzione mostrano molte proprietà interessanti. In primo luogo, se l'immagine in input viene traslata, l'output della feature map sarà traslato della stessa quantità ma rimarrà invariato altrove. Questa proprietà è alla base della robustezza rispetto alle traslazioni e alle distorsioni dell'immagine in ingresso; in secondo luogo, mettendo in fila diversi strati di convoluzione si ottiene una rete capace di avere una comprensione più "astratta" dell'immagine in ingresso. Il primo strato di convoluzione si occupa di estrarre features direttamente dai pixel grezzi dell'immagine e li memorizza nelle feature maps. Questo output diviene poi l'input di un successivo livello di convoluzione, il quale andrà a fare una seconda estrazione delle caratteristiche, combinando le informazioni dello strato precedente. Da questa astrazione a più livelli deriva una maggior comprensione delle features.

### 3.2.2 Strato di ReLU

Nel capitolo ?? si è detto che la funzione sigmoide non era la più efficace. Difatti, negli anni si è stabilita con sicurezza la *Rectified Linear Unit* (ReLU). La ReLU è più verosimile alla modalità di attivazione biologica dei nostri neuroni[22], ed è definita come:

$$f(x) = \max(0, x)$$

Y. LeCun ha dichiarato che la ReLU è inaspettatamente “*l’elemento singolo più importante di tutta l’architettura per un sistema di riconoscimento*”. Questo può essere dovuto principalmente a 2 motivi:

1. la polarità delle caratteristiche è molto spesso irrilevante per riconoscere gli oggetti;
2. la ReLU evita che quando si esegue pooling (sezione 3.2.3) due caratteristiche entrambe importanti ma con polarità opposte si cancellino fra loro.

### 3.2.3 Strato di Pooling

Un’altra proprietà che si vuole ottenere per migliorare i risultati sulla visione artificiale è il riconoscimento delle features indipendentemente dalla posizione nell’immagine, perché l’obiettivo è quello di rafforzare l’efficacia contro le traslazioni e le distorsioni. Questo si può ottenere diminuendo la risoluzione spaziale dell’immagine, il che favorisce una maggiore velocità di computazione ed è al contempo una contromisura contro l’overfitting, dato che diminuisce il numero di parametri.

Lo strato di pooling ottiene in ingresso N immagini di una risoluzione e restituisce in uscita lo stesso numero di immagini, ma con una risoluzione ridotta in una certa misura, solitamente del 75%. Infatti, la forma più comune di pooling layer utilizza dei filtri 2x2, che dividono l’immagine in zone di 4 pixel non sovrapposte e per ogni zona scelgono un solo pixel.

I criteri con cui scegliere il pixel vincente sono diversi:

- average pooling: si calcola il valore medio sui pixel del pool;
- median pooling: si calcola la mediana dei valori dei pixel del pool;
- LP-pooling: si calcola la p-norma della matrice dei pixel;
- max pooling: si sceglie il pixel col valore più alto.

Di questi, quello che si è dimostrato più efficace è il *max pooling*, figura 3.5.

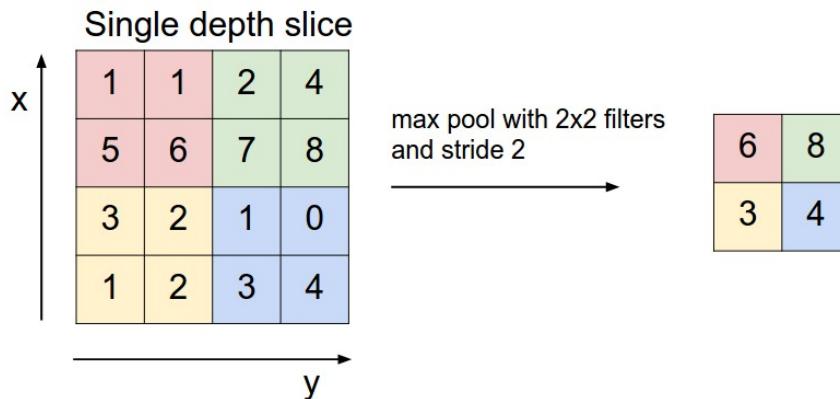


Figura 3.5: Max pooling: in uscita l’immagine avrà 1/4 dei pixel di partenza.

Attraversando la rete si avrà gradualmente un numero di feature maps più alto e quindi della ricchezza della rappresentazione delle features; ed una diminuzione della risoluzione dell’input. Questi fattori combinati insieme donano un forte grado di invarianza alle trasformazioni geometriche dell’input.

### 3.2.4 Strato completamente connesso (FC)

Nello strato completamente connesso, tutti gli input provenienti dai layer di convoluzione vengono dati in ingresso ad una normale rete neurale completamente connessa che fungerà da classificatore. I calcoli in questa parte finale sono quindi uguali alle moltiplicazioni tra matrici visti nel Capitolo ??.

Ultimamente però, si è notato che, eccetto per la modalità di connessione, questi neuroni sono funzionalmente identici a quelli dei layer di convoluzione (entrambi computano moltiplicazioni fra matrici). Quindi si possono sostituire con degli strati di convoluzione che hanno un receptive field pari alla risoluzione delle immagini in input [23].

In figura 3.6 è rappresentata l'architettura completa di una CNN. Si noti come la risoluzione dell'immagine si riduca ad ogni strato di pooling (chiamato anche sub-sampling) e come ogni pixel delle feature maps derivi dal campo recettivo sull'insieme di tutte le feature maps del livello precedente.

In figura 3.7 invece, si può osservare una CNN nell'atto di classificazione di un'auto. Sono visualizzati i filtri della rete durante tutti i vari livelli di elaborazione dell'input, per poi terminare in uno strato completamente connesso che da in output una probabilità. Questa probabilità è poi tradotta in uno score, da cui si sceglie la classe vincente.

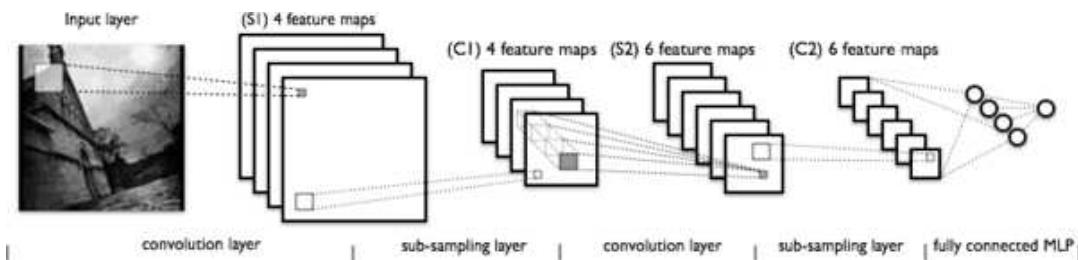


Figura 3.6: Architettura di una CNN

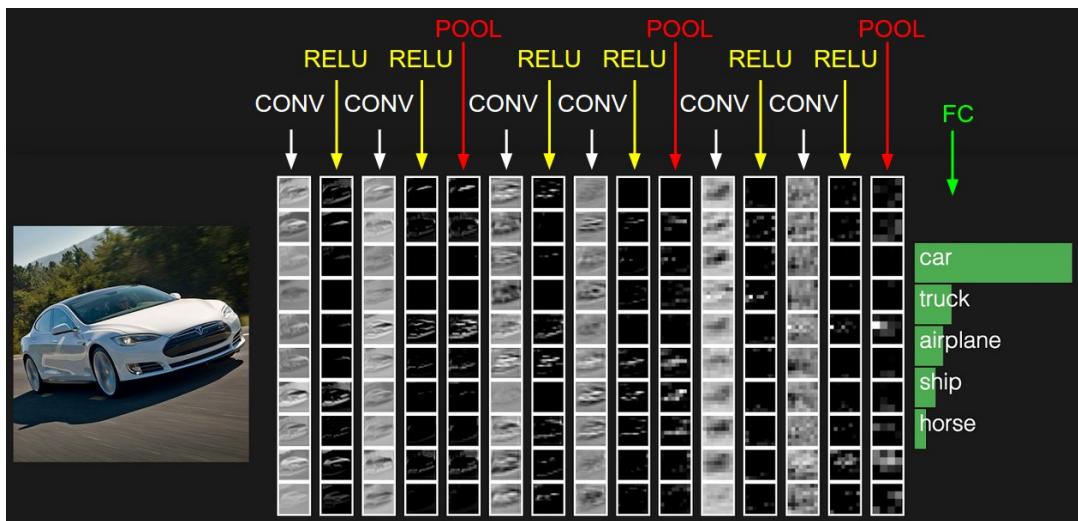


Figura 3.7: Tipica CNN in un task di classificazione; la classe vincente è quella con la probabilità più alta, indicata alla fine

### 3.3 Applicazioni e risultati

L'alta efficacia, la peculiare, vantaggiosa architettura insieme con l'enorme progresso tecnologico dell'hardware, hanno reso le CNN il sistema più promettente per compiti di visivo, con i più svariati ambiti applicativi come: riconoscimento e tagging facciale (si pensi a Facebook), ricerca intelligente di immagini (si pensi a Google Photos), automobili autonome, smartphone, robot, droni, (video)giochi ed altro. Le CNN hanno avuto eccellenti risultati anche nell'elaborazione naturale del linguaggio; nella scoperta di farmaci, poiché predicendo le interazioni tra determinate molecole e proteine hanno contribuito a scoprire potenziali biomolecole per il trattamento dell'ebola [24], tanto per citarne altri.

Già 3 anni fa, in un articolo pubblicato dal dipartimento di visione artificiale del KTH [25], si era analizzato l'utilizzo di "OverFeat", una CNN allenata per l'ImageNet Challenge del 2013. L'articolo sottolinea come abbiano utilizzato questa CNN "off-the-shelf" ovvero già pronta e, senza allenarla ulteriormente, testandola contro altri metodi allo stato dell'arte finemente perfezionati sviluppati fino ad allora. Come test hanno scelto attività gradualmente sempre più lontane dall'originario compito per cui OverFeat è stata addestrata e, con enorme stupore, hanno verificato che OverFeat surclassa i suddetti metodi su qualsiasi dataset (si rimanda all'articolo per i dettagli) nonostante sia stata allenata solo mediante l'ImageNet. L'articolo si chiude con una frase che qui cito:

*"Thus, it can be concluded that from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task."*

#### 3.3.1 Confronto con l'uomo

Nel 2011, le CNN hanno la prima volta battuto l'uomo raggiungendo un errore di 0.56% contro l'1.16% degli umani sul riconoscimento dei segnali stradali nella competizione "German Traffic Sign competition run by IJCNN 2011".

Due anni fa, nell'annuale competizione di ILSVRC, considerata ormai dalla comunità come le "Olimpiadi della Visione Artificiale", Microsoft Research Asia ha presentato ResNet [21]: una CNN a 152 strati che ha abbassato l'errore sulla classificazione (1000 classi) al *solo 3.6%*. Il risultato è impressionante, dato che un essere umano più o meno abile nel riconoscere tutte le classi ha un errore di circa 5-10%, si veda [26].

Un altro compito storicamente difficile per la visione artificiale era il riconoscimento di visi parzialmente occlusi, capovolti o spostati di diverse angolazioni. Tuttavia, nel 2015 un team del Yahoo Labs, è riuscito a far apprendere anche questo compito ad una CNN [27].

L'ultima pietra migliare in ordine cronologico della sfida "Uomo vs. Macchina" è senza dubbio quella di ALPHAGo [28]. AlphaGo è il primo programma a riuscire a battere un giocatore umano professionista (Lee Sedol, 18 volte campione del mondo) all'antico gioco cinese di Go.

Go è conosciuto per essere computazionalmente estremamente complesso: vi sono  $10^{170}$  possibili combinazioni della scacchiera, un numero più alto degli atomi dell'Universo conosciuto. È quindi inaffrontabile con un approccio a "forza bruta".

AlphaGo si basa su una combinazione di deep learning + tree search. In particolare, utilizza 3 CNN: 2 "policy network" per scegliere la strategia più vincente ed 1 "value

"network" come funzione euristica per valutare la bontà di un'ipotetica mossa. In più, l'output di queste reti è combinato con una "Montecarlo Tree Search" per avere una risposta finale sulla prossima mossa da giocare. Maggiori dettagli si trovano sull'esaustivo paper pubblicato da DeepMind [29].

Questi risultati bastano per comprendere la potenzialità delle convolutional neural networks.

## Chapter 4

# 4 Tensor Decomposition

In this chapter the main mathematical tools for the manipulation of tensors are introduced. Following, we will see how to apply these operators to decompose a convolutional layer through two different techniques, effectively exploring the state-of-the-art methods of low-rank approximation presented in Chapter 2.

## 4.1 Background

[A better introduction to tensors is needed. !!]

A tensor is a geometric object that generalizes all the structures usually defined in linear algebra to the  $n$ -dimensional space. As such, they can be defined as  $n$ -dimensional arrays.

Given a reference basis of vectors - i.e. a set of linearly independent vectors with which we can represent every other vector in the correspondent vector space - (cite needed) a tensor can be represented as a multidimensional array of numeric values. We define *rank* (or *order*) of a tensor the dimensionality of the array needed to represent it with respect to this basis, or the number of indices needed to label a component of that array. Thus, an  $k$ -th order tensor in an  $n$ -dimensional space is a mathematical object that has  $n$  indices and  $n^k$  components; each index ranges over the number of dimensions of the space.

A third-order tensor is showed in ??.

### 4.1.1 Tensor rank

Intuitively, a scalar would be a tensor of *order 0*; a vector of *order 1*; a matrix of *order 2* and so on.

The intuitive definition can also be used using "rank" instead of order, but that may be somewhat misleading since there's a subtle difference. To avoid confusion, the following definitions are introduced:

- a tensor of *rank-1* or a decomposable tensor tensor-hackbush2009 is a tensor that can be written as a product of a product of tensors of the form:

$$T = a \circ b \circ \dots \circ d \quad (4.1)$$

- The *rank* of a tensor  $T$  is the minimum number of rank-1 tensor that sum to  $T$ .

The product used in 4.1 is an outer product for vectors and will be defined in detail in the coming section.

Tensors are extensively used in many applications [30] to modelize multi-dimensional data. In the CNN scenario, a CONV layer with  $\mathcal{W}$  weights is defined through a tensor of size:

$$\dim(\mathcal{W}) = [T \times S \times D \times D] \quad (4.2)$$

where,

- $T$  is the number of output filters
- $D$  is the size of the kernel of the convolution
- $S$  is the number of input filters

In pure mathematical terms, there exist a lot of different methods to decompose a tensor. In section 4.2 the formal tools to wield tensors are given and in section 4.3 the application of these methods to convolutional layers are presented.

#### 4.1.2 Singular value decomposition

In order to understand better how a tensor decomposition work and its properties, it is necessary to introduce the *singular value decomposition* (SVD) for matrices.

Let  $M$  be a matrix  $\in \mathbf{F}$  of size  $m \times n$ , the SVD is given by:

$$M = U\Sigma V^* \quad (4.3)$$

Where  $U$  and  $V$  are an  $m \times m$  and  $n \times n$  unitary matrix respectively. In the case of  $\mathbf{F} = \mathbf{R}$ ,  $U$  and  $V$  are also orthogonal matrices.  $V$  is the conjugate transpose of  $V$ .  $\Sigma$  is a *diagonal* matrix with non-negative real numbers, which holds the singular values of  $M$  in its diagonal.

A thorough explanation of the above terms is required, so the following definitions must be kept in mind:

- A *diagonal* matrix is a matrix whose values are all zero except for those on the diagonal. A special case of diagonal matrix is an *identity* matrix, where all these diagonal elements are equal to 1.
- Given a matrix  $M$ , the *transpose* is an operator that flips  $M$  over its diagonal producing another matrix  $M^T$  as a result, whose column and rows indices are therefore also switched. Hence, the rows of  $M$  becomes the columns of  $M^T$  and viceversa.
- Given a matrix  $M \in \mathbf{C}$ , its *complex conjugate*,  $\bar{M}$ , is the conversion of each element  $m_{i,j}$  to its conjugate i.e., the real part are the same while the imaginary part have opposite sign and same magnitude.
- A *conjugate transpose* is a matrix  $M^*$  who's been obtained by first transposing  $M$  and then taking the *complex conjugate* of each entry. Also, the following properties holds:

$$M^* = (\bar{M})^T = M^{\top} T$$

- A quadratic matrix  $M$  is said to be *unitary* if  $MM^* = M^*M = I$ , where  $I$  is the identity matrix. In the case  $M \in \mathbf{R}$  the matrix is called *orthogonal* and it satisfies the equivalence  $MM^T = M^TM = I$ .
- An *orthogonal* matrix has rows and columns that are unitary or orthogonal between each other, respectively.
- A non-negative real number  $\sigma$  is a singular value for  $M$  of a space  $F^{m \times n}$  if and only if there exist unit-length vectors  $u \in \mathbf{F}^m$ ,  $v \in \mathbf{F}^n$  such that  $Mv = \sigma u$  and  $M^*u = \sigma v$ . The vectors  $u$  and  $v$  are called left-singular and right-singular vectors for  $\sigma$  respectively.

Recalling the SVD definition ??, the  $\times n$  rectangular matrix  $\Sigma$  holds the *singular values* (the square roots of the non-zero *eigen-values*)  $\sigma_i$   $i = 1, \dots, k$  of  $M$  on its diagonal. The first  $k = \min(m, n)$  columns of  $U$  and  $V$  are, respectively, left-singular vectors and right-singular vectors for the corresponding singular values. Consequently, the SVD theorem implies that:

- An  $m \times n$  matrix  $M$  has at most  $k$  distinct singular values;
- It is always possible to find a unitary basis  $U$  for  $\mathbf{F}^m$  with a subset of basis vectors spanning the left-singular vectors of each singular value of  $M$ ;
- It is always possible to find a unitary basis  $V$  for  $\mathbf{F}^n$  with a subset of basis vectors spanning the right-singular vectors of each singular value of  $M$ .

### 4.1.3 SVD Applications

The SVD factorization is useful in many fields of research. It can be used to solve the *linear least-squares* and the *total least-squares* problems; it is widely used in statistics where it is related to *principal component analysis* (PCA); it's successfully used in signal processing and pattern recognition. SVD is also fundamental in *recommender systems* to predict people's item ratings [recsys1] [recsys2]; for instance, Netflix has launched a global competition to find the best implementation of SVD for clusters to improve its collaborative filtering technique [recsys3-netflix]. The main area of interest of SVD for convolutional neural networks is that of *low-rank matrix approximation* and image processing.

In fact, SVD can be thought of as decomposing a matrix into a *weighted, ordered* sum of separable matrices. Separable here means exactly the same concept introduced for tensors in section ?? i.e., a matrix  $A$  can be written as an outer product of two vectors  $A = u \circ v$ . More precisely, the matrix can be factorized as:

$$M = \sum_i A_i = \sum_i \sigma_i U_i \circ V_i^T \quad (4.4)$$

this turns out to be enable a fast convolution computation:

$$F = \sum_i^R \sigma_i (I * U_i) * V_i \quad (4.5)$$

where  $I$  is the input image,  $F$  is the resultant feature map and  $U_i$  and  $V_i$  are  $i$ -th column of  $U$  and  $V$  respectively.

Note that  $\sigma_i$  are the  $R$  largest singular values (the other singular values are replaced by zero). Since the number of non-zero  $\sigma_i$  correspond exactly to the rank of a matrix, the approximated matrix is thus of rank  $R$ .

For this purpose, given a matrix  $M$ , the best method to find a truncated matrix  $\tilde{M}$  of rank  $R$  that approximates  $M$ , is to find the solution that minimizes the *frobenius norm* of their difference:

$$\|M - \hat{M}\|, \quad \text{with } \hat{M} = \sum_i U_i \circ V_i \quad (4.6)$$

The Eckart-Young theorem [Weckart] states that the best solution to the problem is given by its SVD decomposition.

## SVD on Convnets

$$\begin{aligned}
 & (U_{n \times t} \Sigma_{t \times t} V_{m \times t}^T) x + b \\
 = & U_{n \times t} (\Sigma_{t \times t} V_{m \times t}^T x) + b
 \end{aligned}$$

## 4.2 Tensor mathematical tools

### 4.2.1 Basic operations

In multi-linear algebra it actually does not exist a method that satisfy all SVD properties for *m-way arrays*, i.e. tensors. However, by taking a closer look at SVD we can formulate two main requirements that a tensor decomposition algorithm should satisfy to be a feasible alternative to SVD in a tensor-world:

1. a Rank-R decomposition
2. the orthonormal row/column matrices.

SVD computes both of them simultaneously. Regarding tensors, these properties can be captured separately by two different family of decompositions.

The first property is extended to the multi-linear world by a class of decompositions that fall under the name of *CP decomposition* (named after the two most popular variants, CANDECOMP and PARAFAC). The latter is provided by the Tucker methods (and many other names). For each axes of a tensor, these methods compute the associated orthonormal space. Therefore, Tucker methods are also used in multilinear principal component analysis (PCA).

Historically, much of the interest in higher-order SVDs was driven by the need to analize empirical data, especially in psychometrics, chemometrics and neuroscience (CIT NEEDED). As such, these techniques have been rediscovered many times with different names leading to a confused literature. Thus, these methods are often presented in a more practical goal-driven way, rather than through rigourous abstract general theorems, which are in fact rare.

Before diving into tensor decomposition algorithms, we need to introduce some fundamental tensor operations:

- **Tensor times matrix: *k*-mode product:** The *k*-mode product of a tensor  $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$  with a matrix  $M \in \mathbb{R}^{J \times I_k}$  is written as:

$$\mathbf{Y} = \mathbf{X} \times_k M \tag{4.7}$$

The resulting tensor  $\mathbf{Y}$  is of size  $I_1 \times \dots \times I_{k-1} \times J \times I_{k+1} \times \dots \times I_N$ , and contains the elements:

$$y_{i_1 \dots i_{k-1} j i_{k+1} \dots i_N} = \sum_{i_k=1}^{I_k} x_{i_1 i_2 \dots i_N} a_{j i_k}.$$

It can be hard to visualize this operation at first, but effectively it boils down to multiply each mode-*k* fiber of  $\mathbf{X}$  by the matrix  $M$ . Looking at ??, we can also represent the same operation with  $\mathbf{Y}_{(i)} = \mathbf{X}_{(i)} \cdot M$ , being  $\mathbf{X}_{(i)}$  the mode-*i* unfolding of the tensor  $\mathbf{X}$ .

To simplify things, let  $\mathbf{X}$  be a 3-mode tensor  $\in F^{I \times J \times K}$  and  $M$  a matrix  $\in F^{N \times J}$ , the k-mode product on axis 1 of X and M is:

$$\mathbf{Y} = \mathbf{X} \times_1 M, Y \in F^{I \times J \times K} \quad (4.8)$$

So each element  $(n, j, k)$  of Y is obtained by:

$$y_{n,j,k} = \sum_i x_{i,j,k} \cdot b_{n,j} \quad (4.9)$$

A visual example is depicted in ??.

Few interesting properties of the k-mode product are:

- $X \times_m A \times_n B = S \times_n B \times_m A$  if  $n \neq m$
- $X \times_n A \times_n B = X \times_n (BA) \neq X \times_n B \times_n A$ .

- **Tensor times vector:** Given the same matrix X, the tensor-vector multiplication on the  $i$ -axis is defined as:

$$\mathbf{Y} = \mathbf{X} \times_1 v, \mathbf{Y} \in \quad (4.10)$$

with each element  $y_{j,k}$ :

$$y_{j,k} = \sum_i x_{i,j,k} \cdot a_i \quad (4.11)$$

An example is illustrated in ??.

- **Matrix Kronecker product:** A Kronecker product of two matrices  $A \in \mathbf{R}^{M \times N}$  and  $B \in \mathbf{R}^{P \times Q}$  is defined as:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \quad (4.12)$$

and more explicitly:

*MATRICIONE*

- **Outer product:** If we take the *Kronecker product for matrices* definition and apply it to vectors, we obtain the outer product:

$$\mathbf{a} \circ \mathbf{b} = \mathbf{ab}^T = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} = \begin{bmatrix} a_1b_1 & a_1b_2 & a_1b_3 \\ a_2b_1 & a_2b_2 & a_2b_3 \\ a_3b_1 & a_3b_2 & a_3b_3 \\ a_4b_1 & a_4b_2 & a_4b_3 \end{bmatrix}. \quad (4.13)$$

Let  $a \in \mathbf{R}^I$ ,  $b \in \mathbf{R}^J$   $c \in \mathbf{R}^K$  be three vectors. Computing the outer product  $(a \circ b)$  of two of them will result in a matrix, as showed above. Proceeding in this way is easy to show that an outer product of 3-vectors will result in a 3-dimensional tensor, as illustrated in ??.

This comes in handy the other way around: a rank-1 tensor can be decomposed into 3 vectors. As we will see in the following sections, this operation is fundamental for tensor decomposition.

Another interesting way to look at it is that the outer product operation “ $\otimes$ ” is a way of combining a tensor of  $d_1$ -order and a tensor of  $d_2$ -order to obtain a tensor of order- $(d_1 + d_2)$ .

- **Matrix Khatri-Rao product:** given two matrices  $A \in \mathbf{R}^{M \times N}$  and  $B \in \mathbf{R}^{P \times R}$  is defined as:

$$A \odot B = [a_1 \otimes b_1, a_2 \otimes b_2, \dots, a_N \otimes b_R] \in \mathbf{R}^{MN \times R} \quad (4.14)$$

Note that the Kronecker matrix operation returns the same number of elements of the Khatri-Rao product, but while the former produce a matrix, the latter is shaped into a vector.

### 4.2.2 Tucker Decomposition

#### HO-SVD

As the *Higher-order singular value decomposition* was studied in many scientific fields, it is historically referred in different ways: multilinear singular value decomposition, m-mode SVD, or cube SVD, and it is often incorrectly identified with a Tucker decomposition.

HOSVD is actually a specific orthogonal version of the Tucker decomposition. To put it in other words, it is a specialized algorithm to compute the Tucker decomposition. HOSVD involves solving each  $k$ -mode matricized form of the specific tensor [31], relying on the following equivalence:

$$\begin{aligned} Y &= X \times_1 A^{(1)} \times_2 A^{(2)} \times_3 \dots \times_N A^{(N)} \\ \Leftrightarrow Y_{(k)} &= A^{(k)} X_{(k)} \left( A^{(N)} \otimes \dots \otimes A^{(k+1)} \otimes A^{(k-1)} \otimes \dots \otimes A^{(1)} \right)^T. \end{aligned}$$

The algorithm steps follow:

```

for  $k = 1, 2, \dots, N$  do
     $A^{(k)} \leftarrow$  left orthogonal matrix of SVD of  $X_{(k)}$ 
end for
 $G \leftarrow X \times_1 (A^{(1)})^T \times_2 (A^{(2)})^T \times_3 \dots \times_N (A^{(N)})^T$ 
```

This approach may be regarded as one generalization of the matrix SVD, because:

- each matrix  $A^k$  is an orthogonal matrix
- Two subtensors of the core tensor  $\mathbf{G}$  are orthogonal, i.e.  $\langle \mathbf{G}_p, \mathbf{G}_q \rangle \quad if \quad p \neq q$
- the subtensors in the core tensor  $\mathbf{G}$  are ordered according to their Frobenius norm, i.e.  $\|\mathbf{G}_1\| \geq \|\mathbf{G}_2\| \geq \dots \geq \|\mathbf{G}_n\| \quad for \ n=1,\dots,N$

The tensor  $\mathbf{G}$  is said to be “*ordered*” and “*all-orthogonal*”. For further explanation see [32].

$$\mathbf{X} = \mathbf{G} \times_1 A^{(1)} \times_2 A^{(2)} \times_3 \dots \times_N A^{(N)} \quad (4.15)$$

### 4.2.3 Canonical Polyadic Decomposition

The Polyadic Decomposition (PD) [33] approximates a tensor with a sum of  $R$  rank-one tensors. If the number of rank-one terms  $R$  is minimal, then  $R$  is called the rank of the tensor and the decomposition is called minimal or *canonical* (CPD).

For any other arbitrary rank- $r$ , the decomposition is often referred to as CANDECOMP/PARAFAC (CP). As we will discover later, selecting the perfect rank is an *NP-Hard* problem. Hence, from now on we will refer to this decomposition as CP.

Recall the outer product between vectors introduced in section 4.2. Let  $\vec{a}$ ,  $\vec{b}$  and  $\vec{c}$  be nonzero vectors in  $\mathbf{R}^n$ , then  $\vec{a} \circ \vec{b} \equiv \vec{a} \cdot \vec{b}$  is a rank-one matrix and  $\vec{a} \circ \vec{b} \circ \vec{c}$  is defined to be a rank-one tensor. Let  $T$  be a tensor of dimensions  $I_1 \times I_2 \times \dots \times I_N$ , and let  $U^{(n)}$  be matrices of size  $I_n \times R$  and  $\vec{u}_r^{(n)}$  the  $r$ -th column of  $U^{(n)}$ , then:

$$T \approx \sum_{r=1}^R \vec{u}_r^{(1)} \circ \vec{u}_r^{(2)} \circ \dots \circ \vec{u}_r^{(N)}. \quad (4.16)$$

A visual representation of this decomposition in the third-order case is shown in ??

It's interesting to notice that CPD can be regarded as a special case of a Tucker Decomposition in which the core tensor  $\mathbf{G}$  is constrained to be a super-identity  $\mathbf{I}$ , which is an extension of the identity matrix and has all one's on its superdiagonal and all zero's off the superdiagonal.

## 4.3 Application of tensor decompositon on CNN

### 4.3.1 Convolutional layer as 4-mode tensors

### 4.3.2 CPD

$$\sum_{r=1}^R K_r^x(i) K_r^y(j) K_r^s(s) K_r^t(t). \quad (4.17)$$

$$\begin{aligned} V(x, y, t) &= \sum_i \sum_j \sum_s K(x - i, y - j, s, t) X(i, j, s) \\ &= \sum_r \sum_i \sum_j \sum_s K_r^x(x - i) K_r^y(y - i) K_r^s(s) K_r^t(t) X(i, j, s) \\ &= \sum_r K_r^t(t) \sum_i \sum_j K_r^x(x - i) K_r^y(y - i) \sum_s K_r^s(s) X(i, j, s) \end{aligned} \quad (4)$$

### 4.3.3 Tucker

The Tucker Decomposition, also known as the Higher Order SVD *HOSVD* is a generalization of SVD for tensors:

$$K(i, j, s, t) = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} \sum_{r_4=1}^{R_4} G_{r_1, r_2, r_3, r_4} * K_{r_1}^x(i) K_{r_2}^y(j) K_{r_3}^s(s) K_{r_4}^t(t)$$

The Tucker Decomposition has a property that's useful for our purposes: it doesn't have to be applied along all modes (axis) of the tensors. We have seen how CPD

decomposes the kernel also spatial-wise; this acts pretty aggressively on the number of parameters and since the convolutional kernels are small in most of the recent implementations ( $3 \times 3$  or  $5 \times 5$ ) it arguably doesn't save a lot of computation. Hence, we can skip this decomposition with Tucker, going for what is known as a Tucker-2 Decomposition:

$$K(i, j, s, t) = \sum_{r_3=1}^{R_3} \sum_{r_4=1}^{R_4} \mathbf{G}_{i, j, r_3, r_4}(j) K_{r_3}^s(s) K_{r_4}^t(t) \quad (4.18)$$

Using equation 4.15 and plugging into the formula for the convolutional forward pass, we obtain the new equation for the Tucker convolutional forward pass:

$$V(x, y, t) = \sum_i \sum_j \sum_s K(x - i, y - j, s, t) X(i, j, s) \quad (4.19)$$

$$V(x, y, t) = \sum_i \sum_j \sum_s \sum_{r_3=1}^{R_3} \sum_{r_4=1}^{R_4} \mathbf{G}(x - i)(y - j) r_3 r_4 K_{r_3}^s(s) K_{r_4}^t(t) X(i, j, s) \quad (4.20)$$

$$V(x, y, t) = \sum_i \sum_j \sum_{r_4=1}^{R_4} \sum_{r_3=1}^{R_3} K_{r_4}^t(t) \mathbf{G}(x - i)(y - j) r_3 r_4 \sum_s K_{r_3}^s(s) X(i, j, s) \quad (4.21)$$

## 4.4 MNIST: addestramento

## 4.5 CIFAR: preprocessing

## 4.6 CIFAR: addestramento

```

1 ----- preprocess/normalize train/test sets -----
2 print '<trainer> preprocessing data (color space + normalization)'
3
4 -- preprocess trainSet
5 normalization = nn.SpatialContrastiveNormalization(1, image.gaussian1D(7))
6 for i = 1,trainData:size() do
7   --rgb -> yuv
8   local rgb = trainData.data[i]
9   local yuv = image.rgb2yuv(rgb)
10
11  -- normalize Y locally:
12  yuv[1] = normalization(yuv[{1}])
13  trainData.data[i] = yuv
14 end
15 -- normalize U globally:
16 mean_u = trainData.data[{ {},2,{},{} }]:mean()
17 std_u = trainData.data[{ {},2,{},{} }]:std()
18 trainData.data[{ {},2,{},{} }]:add(-mean_u)
19 trainData.data[{ {},2,{},{} }]:div(-std_u)
20
21 -- normalize V globally:

```

```
22 mean_v = trainData.data[{: {},3,{},{} }]:mean()
23 std_v = trainData.data[{: {},3,{},{} }]:std()
24 trainData.data[{: {},3,{},{} }]:add(-mean_v)
25 trainData.data[{: {},3,{},{} }]:div(-std_v)
26
27 --same applies to test set...
```

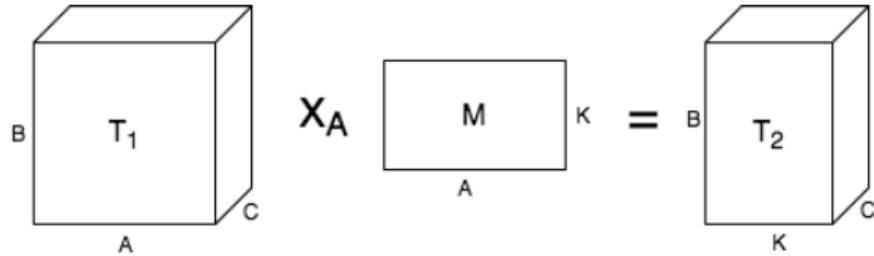
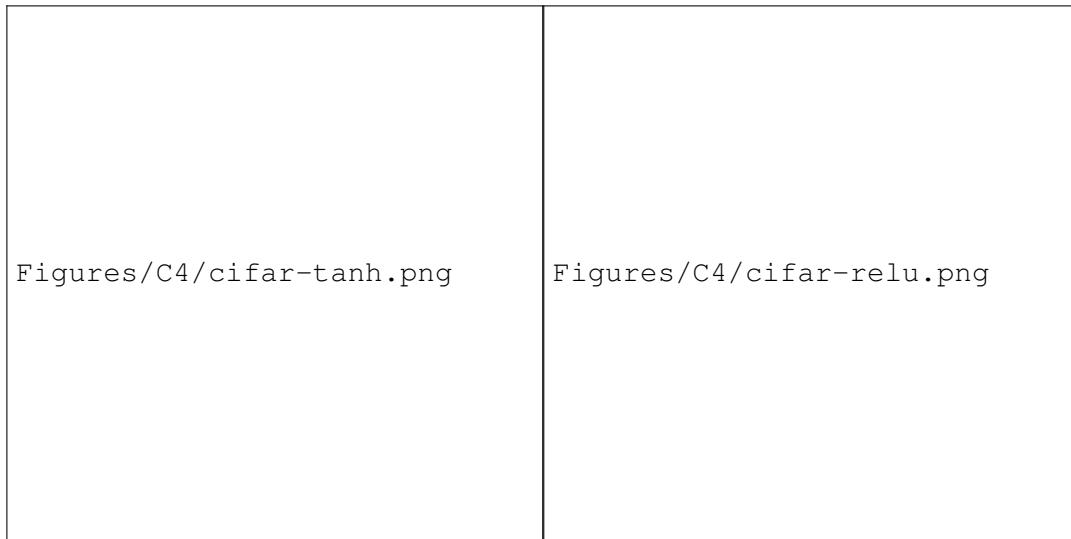
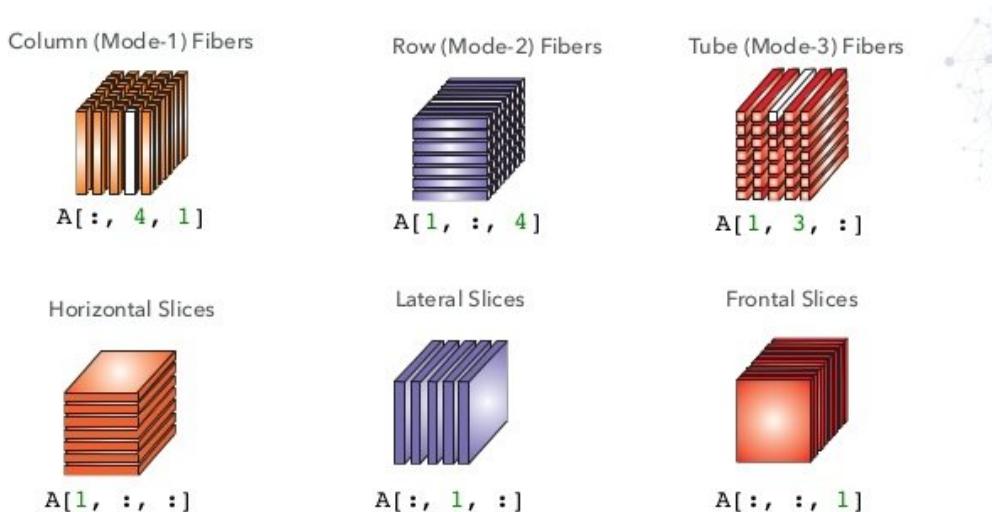


Figura 4.1: Example of k-mode multiplication on 3-dimensional tensor.



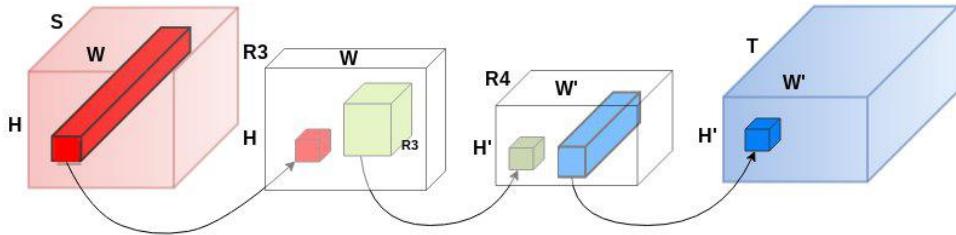
(a) Acc.= ~65% f. di attivazione = TanH      (b) Acc.= ~73% f. di attivazione = ReLU

*Figura 4.2:* Percentuali di accuracy a seconda della funzione d'attivazione. La ReLU produce indubbiamente risultati migliori.

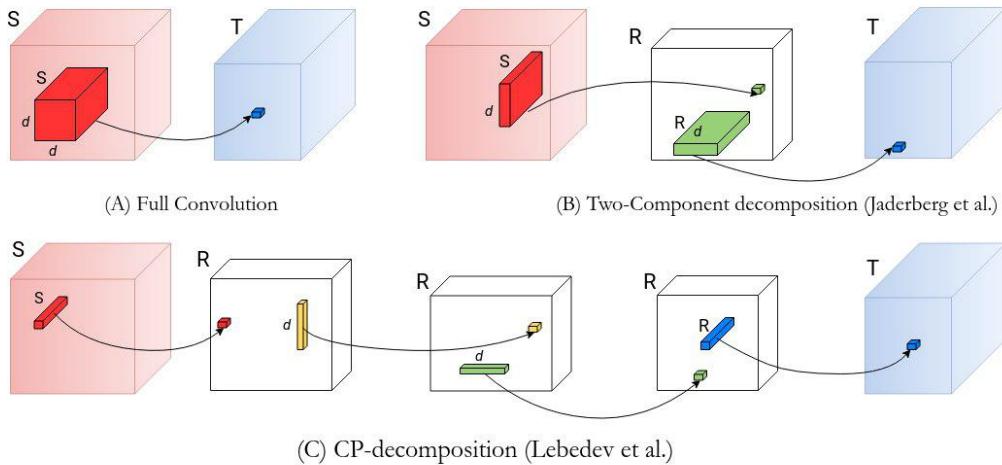


Cichocki et al. Nonnegative Matrix and Tensor Factorizations

*Figura 4.3:* Fibers and slices of a tensor: fibers is an equivalent term for a tensor mode.



*Figura 4.4:* Tucker-2 Decompositions for speeding-up a generalized convolution. Each box corresponds to a 3-way tensor  $X, Z, Z'$  and  $Y$  in equation (??-??). Arrows represent linear mappings and illustrate each scalar value on the right is computed. Red tube, green cube and blue tube correspond to  $1 \times 1$ ,  $d \times d$  and  $1 \times 1$  convolution respectively.



*Figura 4.5:* Tensor Decompositions for speeding up a generalized convolution. Each box corresponds to a feature map stack within a CNN, (frontal sides are spatial dimensions). Arrows show linear mappings and demonstrate how scalar values on the right are computed. Initial full convolution (A) computes each element of the target tensor as a linear combination of the elements of a 3D subtensor that spans a spatial  $d \times d$  window over all input maps. Jaderberg et al. (B) approximate the initial convolution as a composition of two linear mappings in which the intermediate map stack has  $R$  maps, being  $R$  the rank of the decomposition. Each of the two-components computes each target value with a convolution based on a spatial window of size  $dx1$  or  $1xd$  in all input maps. Finally, CP-decomposition (C) by Lebedev et al. approximates the convolution as a composition of four smaller convolutions: the first and the last components compute a standard  $1 \times 1$  convolution that spans all input maps while the middle ones compute a 1D grouped convolution **only on one** input map.



## Chapter 5

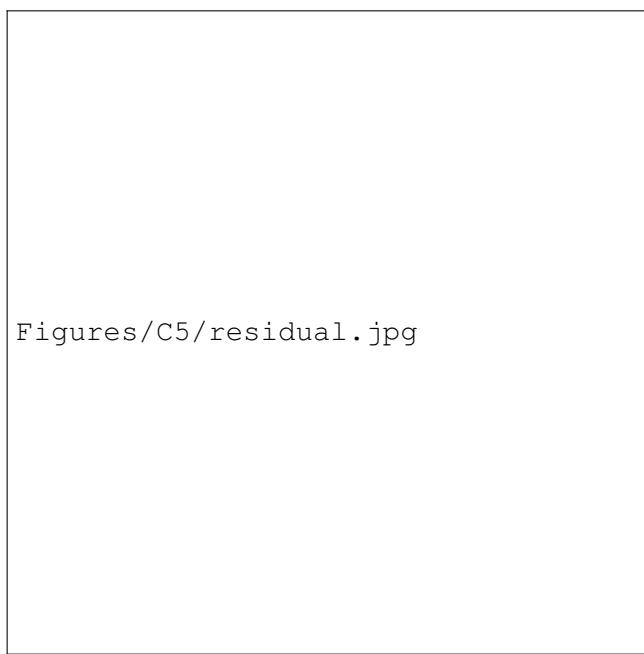
# Addestrare un modello allo stato dell'arte

## 5.1 Deep Residual Network

Menzionata nel Capitolo 3 a causa del clamore riscosso per l'incredibilmente bassa percentuale d'errore nella ImageNet Challenge del 2015, la *Deep Residual Network* - o ResNet - è una CNN *molto* profonda presentata da Microsoft Research Asia [21] che ha stravolto tutti i record in classification, detection e localization.

L'architettura è modulare: si basa su un blocco detto "*residual block*" (figura 5.1) che viene ripetuto N volte, prima dell'ultimo strato di classificazione. L'idea dietro a questo blocco viene chiamata *residual learning* e funziona così:

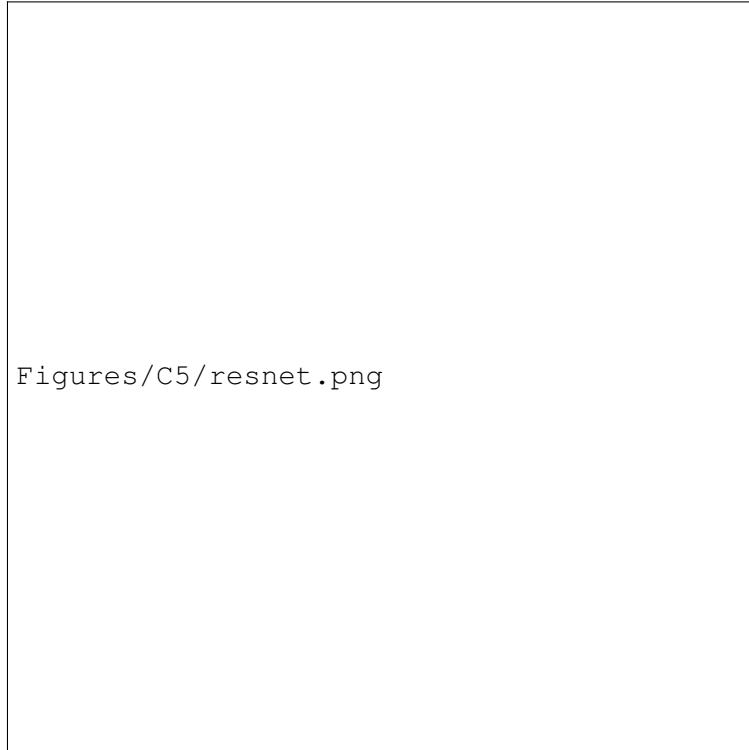
1. l'input  $x$  va attraverso una serie conv-relu-conv dando in output un certo  $F(x)$ ;
2. a questo viene aggiunto l'input originale:  $H(x) = F(x) + x$ ;
3.  $F(x)$  è chiamato residual learning.



*Figura 5.1:* Un "Residual Block", all'input viene aggiunto  $F(x)$  che è il residual

Nelle CNN normalmente si passa direttamente da  $x$  a  $F(x)$ , la quale è una rappresentazione completamente nuova che non mantiene l'informazione dell'input originale. La novità qui, invece, consiste nel mantenere queste informazioni lungo tutta la rete; ogni blocco esegue una sorta di fine-tuning delle informazioni apprese negli strati precedenti, e secondo gli autori è più facile ottimizzare un "residual mapping" anziché un "unreferenced mapping".

Un esempio della profondità di ResNet è rappresentato in figura 5.2, dove viene confrontata la versione da 34 strati con la VGG-Network, la rete più profonda che venne presentata 2 anni prima.



Figures/C5/resnet.png

Figura 5.2: Confronto architetture: VGG-Net la più innovativa della competizione ILSVRC 2014, rete classica a 34 strati (centro), Residual Network a 34 strati (sinistra)

Un'altra ragione per cui ResNet è stupefacente è che la backpropagation funziona meglio a causa delle somme di ogni layer che distribuiscono la derivata lungo la rete. Inoltre, avere queste scorciatoie che non passano attraverso conv-relu aiuta contro il cruciale problema dei *vanishing gradients*. Quando si usano molti layer uno sull'altro si corre il rischio di non riuscire ad ottimizzare l'apprendimento perché l'errore calcolato con la backprop si perde tra i vari strati e dopo alcuni layer arriva un gradiente nullo che non aiuta nell'apprendimento. Per questa ragione anche, il paper di ResNet è importante: ha dimostrato come poter costruire CNN molto profonde (numero di strati  $> 100$ ) riuscendo a completare l'apprendimento ed avere inoltre risultati allo stato dell'arte.

Tuttavia, altre analisi sul successo di ResNet [34] sostengono il contrario: anche in essa il problema dei vanishing gradients si manifesta dopo 20 moduli, e quindi la profondità da sola non può essere la caratteristica principale della rete. Suggeriscono invece, che la rete si comporti come un gruppo di "sotto-reti" più piccole che contribuiscono insieme a generare la risposta corretta.

È sicuramente un argomento recente e complesso; per ulteriori dettagli si rimanda alla bibliografia.

## 5.2 Implementazione di ResNet di Facebook

L'implementazione di ResNet è decisamente complessa ma fortunatamente nella community si rilasciano spesso le implementazioni dei modelli più popolari per stimolare i ricercatori a riprodurre gli esperimenti e migliorare sempre di più lo stato dell'arte. A questo proposito Facebook ha rilasciato il codice per diversi modelli di ResNet da 18-34-50-101-152-200 strati.

Nei seguenti snippet, vi è l'implementazione della scorciatoia identità e del residual block:

```

1 -- The shortcut layer is either identity or 1x1 convolution
2 local function shortcut(nInputPlane, nOutputPlane, stride)
3   local useConv = shortcutType == 'C' or
4     (shortcutType == 'B' and nInputPlane ~= nOutputPlane)
5   if useConv then
6     -- 1x1 convolution
7     return nn.Sequential()
8       :add(Convolution(nInputPlane, nOutputPlane, 1, 1, stride,
9         stride))
10      :add(SBatchNorm(nOutputPlane))
11    elseif nInputPlane ~= nOutputPlane then
12      -- Strided, zero-padded identity shortcut
13      return nn.Sequential()
14        :add(nn.SpatialAveragePooling(1, 1, stride, stride))
15        :add(nn.Concat(2)
16          :add(nn.Identity())
17          :add(nn.MulConstant(0))))
18    else
19      return nn.Identity()
20    end
21  end
22
23 -- The basic residual layer block for 18 and 34 layer network, and the
24 -- CIFAR networks
25 local function basicblock(n, stride)
26   local nInputPlane = iChannels
27   iChannels = n
28
29   local s = nn.Sequential()
30   s:add(Convolution(nInputPlane, n, 3, 3, stride, stride, 1, 1))
31   s:add(SBatchNorm(n))
32   s:add(ReLU(true))
33   s:add(Convolution(n, n, 3, 3, 1, 1, 1, 1))
34   s:add(SBatchNorm(n))
35
36   return nn.Sequential()
37     :add(nn.ConcatTable()
38       :add(s)
39       :add(shortcut(nInputPlane, n, stride)))
40       :add(nn.CAddTable(true))
41       :add(ReLU(true)))
42 end

```

Microsoft ha addestrato ResNet per circa 3 settimane su 8 GPU di ultima generazione. Data l'evidente mancanza di hardware opportuno per un'"impresa" del genere si è deciso di utilizzare il modello più piccolo, quello a 18 strati. In appendice B è mostrata una rappresentazione testuale del modello, ove se ne può apprezzare la complessità e lunghezza.

### 5.3 Addestramento su CIFAR

Si è addestrato il modello ResNet-18 su CIFAR10 per circa 160 epoch. I risultati sul training e validation set sono presentati rispettivamente in figura e figura. Sull'asse X vi sono le epoch di training, sull'asse Y la percentuale di errore.

La rete da in uscita le 10 classi più probabili; con top1 s'intende quando la classe corretta è la prima più probabile mentre con top5 quando la classe corretta è tra le

prime 5 più probabili.



*Figura 5.3:* Top-1 e Top-5 training accuracy di ResNet sul dataset CIFAR10

## 5.4 LeNet vs ResNet: confronto

Nel Capitolo ?? si è addestrato uno dei primi modelli di CNN sullo stesso dataset, CIFAR10. La superiorità di ResNet è evidente, raggiunge un'accuratezza di circa il 92-93% laddove LeNet arrivava appena ad un 73%.

Questo è dovuto a diversi elementi, ma sicuramente è una riprova del fatto che aumentando il numero di layer di convoluzione si aumenta la ricchezza della rappresentazione delle features delle immagini in ingresso e di conseguenza aumenta la capacità della rete di astrarre e "comprendere" molto meglio le informazioni nelle immagini.

Ogni anno le architetture delle reti subiscono variazioni e vengono introdotte novità dirompenti che generano nuove riflessioni che spingono quest'ambito di ricerca in avanti ad una velocità tale da mantenere il passo con l'industria delle GPU di cui hanno bisogno per essere realizzate.



Figures/C5/ResNet-Cifar10-Val.png

*Figura 5.4:* Top-1 e Top-5 validation accuracy di ResNet sul dataset  
CIFAR10

## Chapter 6

# Caso d'uso: fine-tuning su dataset arbitrario

Un tipico scenario d'uso industriale è quello di avere un piccolo dataset di immagini annotate a disposizione e, da questo, dover costruire un modello che riesca a generalizzare bene su esempi ancora mai visti. In questo capitolo si vedrà come affrontare questo problema utilizzando le convolutional neural networks ed il *transfer learning*.

### 6.1 Il problema

Occorre trovare una maniera di costruire un modello che riesca ad essere addestrato su un dataset discretamente piccolo ma al contempo sia successivamente capace di classificare correttamente esempi ancora non visti. Le CNN sono la norma quando si tratta di task di classificazione, tuttavia è difficile riuscire ad affrontare un problema del genere per i seguenti motivi:

1. un modello troppo semplice non sarebbe capace di estrarre tutte le caratteristiche necessarie per catalogare le immagini;
2. addestrare finemente un modello complesso richiede delle risorse computazionali non indifferenti, di cui non si dispone;
3. un modello troppo complesso con un dataset troppo piccolo avrebbe la certezza di incorrere in overfitting.

In pratica, in pochi addestrano interamente una CNN da zero proprio per i problemi elencati sopra. Quello che in genere si fa è usare modelli pre-addestrati per settimane su un dataset enorme come quello di *ImageNet* come estrattori di features da applicare a dataset più piccoli per i più svariati scopi.

### 6.2 Transfer Learning

Sfruttare l'apprendimento di grosse CNN per trasferirlo alla nostra rete è un processo che prende il nome di "*transfer learning*".

Gli scenari principali del transfer learning sono i seguenti:

- **CNN come estrattore di feature:** si prende una rete addestrata su ImageNet, si rimuove l'ultimo strato completamente connesso che faceva da classificatore per le 1000 classi di ImageNet, e si utilizza la restante rete come estrattore di features per il nuovo dataset. Si danno in ingresso alla rete le immagini del nuovo dataset e si ottengono in uscita le attivazioni dell'ultimo hidden layer prima del classificatore finale, chiamate anche *CNN codes*. Una volta ottenuti

questi CNN codes, gli si addestra sopra un classificatore lineare (una SVM o un SoftMax).

Uno schema di questo approccio si è visto nella figura 3.1 del Capitolo 3, dove il primo rettangolo rossa indica la CNN come estrattore di features.

- **Fine-tuning su una CNN:** la seconda strategia consiste non solo nel rimpiazzare e ri-addestrare il classificatore usando CNN sul nuovo dataset; ma anche nel "rifinire" (fine-tune) i pesi di tutta la CNN continuando la backpropagation. Si possono rifinire i pesi di tutti gli strati o solamente dei primi se si hanno preoccupazioni di overfitting. Questa scelta è motivata dal fatto che i primi strati della rete contengono feature più generiche (per esempio edge detectors) che possono essere utili a diversi task, mentre gli ultimi strati diventano sempre più specifici ai dettagli delle classi contenuti nel dataset su cui sono state addestrate. Ad esempio, una CNN addestrata sull'ImageNet challenge che contiene tantissime razze diverse di cane, può avere molto del suo potere di rappresentazione dedicato a distinguere tra le diverse razze.

Siccome le archittture moderne richiedono un addestramento di 2-3 settimane su una moltitudine di GPUs, nella comunità scientifica è consueto rilasciare la rete addestrata on-line, cosicché gli altri ne possano beneficiare. Ad esempio, il framework per il Deep Learning "*Caffe*" ha una repository online chiamato "Model Zoo" [35], dove appunto gli utenti caricano i pesi sinaptici dei loro modelli già addestrati. Grazie a delle librerie che traducono le rappresentazioni dei modelli da un framework ad un altro, è facile importare ed esportare questi modelli (in Torch, ad esempio).

### 6.2.1 Fine-tuning

Il fine-tuning è il processo con il quale si rifiniscono i pesi della rete alle specificità del dataset su cui la si sta addestrando. Come visto prima, è una strategia molto vantaggiosa poiché è un risparmio sia in termini di sviluppo di software ma soprattutto in termini di risorse computazionali.

Tuttavia, vi sono dei rischi nel fine-tuning che si possono arginare ricorrendo ad alcune buone pratiche che fanno riferimento soprattutto a 2 fattori: *la dimensione* del nuovo dataset e *la similarità* al dataset su cui la rete è stata pre-addestrata.

A questo proposito le combinazioni possibili sono 4:

1. *Il nuovo dataset è piccolo e simile al dataset originale:* poiché ci sono pochi dati, non è una buona idea fare il fine-tuning dell'intera rete a causa dell'overfitting. Siccome il dataset è simile all'originale ci si può aspettare che le anche le features degli ultimi livelli della CNN siano rilevanti per questi dati. Di conseguenza, la miglior idea è addestrare un classificatore lineare sui CNN Codes.
2. *Il nuovo dataset è grande e simile al dataset originale:* dal momento che si hanno molti dati a disposizione, si ha più confidenza che la rete non presenti overfitting se si esegue il fine-tuning su tutti gli strati.
3. *Il nuovo dataset è piccolo ma molto diverso dal dataset originale:* siccome i dati sono pochi, probabilmente la scelta migliore è addestrare un classificatore lineare. E siccome il dataset è molto diverso, non conviene addestrarlo sull'ultimo layer della CNN; è meglio addestrarlo sulle attivazioni di uno degli strati precedenti.
4. *Il nuovo dataset è grande e molto piccolo dal dataset originale:* dato che è il dataset è grande è possibile addestrare una CNN da zero. In pratica però, è

sempre vantaggioso inizializzare i pesi da un modello pre-addestrato rispetto ad una inizializzazione random. Dopodiché, ci si può aspettare che un fine-tuning totale della rete funzioni avendo a disposizione tanti dati.

Un ultimo punto a tenere a mente riguarda il *learning rate*: è caldamente consigliato usare learning rate più bassi del solito quando si esegue il fine-tuning. Questo perché si parte dal presupposto che i pesi siano già regolati relativamente bene e non si vuole distorcerli troppo o troppo velocemente, a maggior ragione se sopra questi avviene l'addestramento di un classificatore lineare i cui pesi invece vengono inizializzati in maniera casuale.

### 6.3 Dataset

Il dataset d'esempio contiene 377 immagini appartenenti a 2 classi: api e formiche. Il training set è costituito da 224 immagini, mentre 153 sono per il validation set. Si noti a questo punto, la differenza di 2 ordini di grandezza rispetto ai dataset usato nel Capitolo ??, che contenevano 60 mila immagini ciascuno; e la differenza di 4 ordini di grandezza con ImageNet su cui è stata pre-addestrata ResNet.

Alcuni esempi delle immagini nel dataset si possono osservare nelle figure 6.1 e 6.2.



Figura 6.1: Alcuni esempi delle immagini delle api del dataset

Utilizzando un modello pre-addestrato si estraggono i CNN Codes, ovvero le feature che si hanno in output dagli ultimi layer di convoluzione e si salvano in due file '`train_ants.t7`' e '`train_beans.t7`' (`t7` è il formato specifico usato da Torch). Dopodiché, si caricano i dati nei tensori, gli si associano le label corrette (1 e 2) e li si prepara per il training.

```

1 --Load features for training
2 t1 = torch.load(opt.data..'train_ants.t7').features:float()
3 t2 = torch.load(opt.data..'train_beans.t7').features:float()
4 Z = torch.cat(t1, t2, 1) --CNN Codes
5 dataDim = Z:size(2)
6
7 --1: ants

```



Figura 6.2: Alcuni esempi delle immagini delle formiche del dataset

```

8 --2: bees
9 classes = {1,2}
10 lab1 = torch.Tensor(t1:size(1)):fill(1)
11 lab2 = torch.Tensor(t2:size(1)):fill(2)
12 labels = torch.cat(lab1, lab2)
13
14 --Shuffling the whole given dataset
15 --before dividing it in training and val sets
16 torch.manualSeed(opt.manualSeed)
17 shuffle= torch.randperm(Z:size(1))
18
19 --Creating the datasets objects for the validation and training sets
20
21 validationset={}
22 -- we are going to use 30 % of the whole dataset as the validation set
23 -- the ratio can be changed using the validationSize option
24 function validationset:size() return opt.validationSize * Z:size(1) end
25 for i=1, validationset:size() do
26   validationset[i]={Z[shuffle[i]], labels[shuffle[i]]}
27 end
28
29 trainingset={}
30 function trainingset:size() return Z:size(1) - validationset:size() end
31 for i=1, trainingset:size() do
32   trainingset[i]={Z[shuffle[ validationset:size()+ i ]], labels[shuffle[
33 validationset:size()+i ]]}
```

## 6.4 Fine-tuning su Resnet

### 6.4.1 Training

Come menzionato nel Capitolo 5, Facebook ha reso pubbliche le sue implementazioni di ResNet, compresi i pesi dei modelli pre-addestrati sugli 1.2 milioni di esempi di ImageNet. È stato utilizzato il modello da 18 strati, sopra il quale si è addestrato un

classificatore SoftMax. La funzione SoftMax [36] generalizza la funzione logistica con una funzione esponenziale normalizzata:

$$\sigma(z) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad per \quad j = 1, \dots, K.$$



Figures/C6/ft-arch.png

*Figura 6.3:* Architettura della rete: il transfer learning avviene utilizzando i CNN codes provenienti da ResNet per addestrare il classificatore SoftMax

L'architettura finale è raffigurata in 6.3.

Per addestrare il classificatore lineare non è necessario ricorrere alla libreria di ottimizzazione `optim` di Torch usata per il training delle CNN. Basta utilizzare un modulo di training più semplice chiamato `StochasticGradient`, che appunto lo allenerà mediante una discesa del gradiente classica.

Di default, questo modulo stampa sul terminale solo l'errore corrente. Tuttavia, per riuscire a capire se sta avvenendo overfitting o meno, bisogna definire una funzione di valutazione che verrà chiamata poi durante il training dalla funzione di callback del modulo di training.

```
1 --Defining the evaluation function
2 function eval(model, dataset, validation)
3     local correct=-1
4     local r={}
5     for i=1, dataset:size() do
6         local example=dataset[i]
7         local img = example[1]
8         local label = example[2]
9         local prediction= model:forward(img) --this output the prob (class
10            \| image)
11         local confidences, indices = torch.sort(prediction, true) -- let's
12            sort the prob
13         r[i]=indices[1] -- Picking up the class with highest confidence
14         if validation then --If this is the validation set we can estimate
15            the accuracy
16             if r[i]==label then
17                 correct=correct+1
18             end
19         end
20     end
21     return r, correct
22 end
```

Una volta definita la funzione di valutazione, definiamo la callback function che avrà queste funzioni:

- chiamerà la funzione di valutazione sul training e sul validation set;
- stamperà le rispettive percentuali di accuratezza;
- farà logging dei dati per un'eventuale plotting off-line degli stessi.

```

1 local function evaluation_callback(trainer, iteration, currentError)
2   _, correct=eval(trainer.module, trainingset, true)
3   training_acc= correct / trainingset:size()
4   print("# test accuracy = " .. training_acc)
5
6   _, correct=eval(trainer.module, validationset, true)
7   acc = correct / validationset:size()
8   print("# validation accuracy = " .. acc)
9
10  --save values to be logged later on
11  if trainer.stats then
12    logger:add{iteration, training_acc, acc}
13    trainer.stats.tr[iteration]=training_acc
14    trainer.stats.val[iteration]=acc
15  end
16 end

```

A questo punto possiamo lanciare l'addestramento.

```

1 --define the trainer parameters
2 trainer = nn.StochasticGradient(model, criterion)
3 trainer.hookIteration=evaluation_callback --link callback function
4 trainer.stats={tr={},val={}} --we will use this table to save the stats
5 trainer.learningRate = opt.LR
6 trainer.maxIteration = opt.nEpochs -- epochs of training
7 trainer.verbose = false -- print stats in the callback
8 --let's train
9 trainer:train(trainingset)
10 --save model
11 torch.save(opt.save..'_tuned.t7', model)

```

In output si ottiene qualcosa di simile a:

```

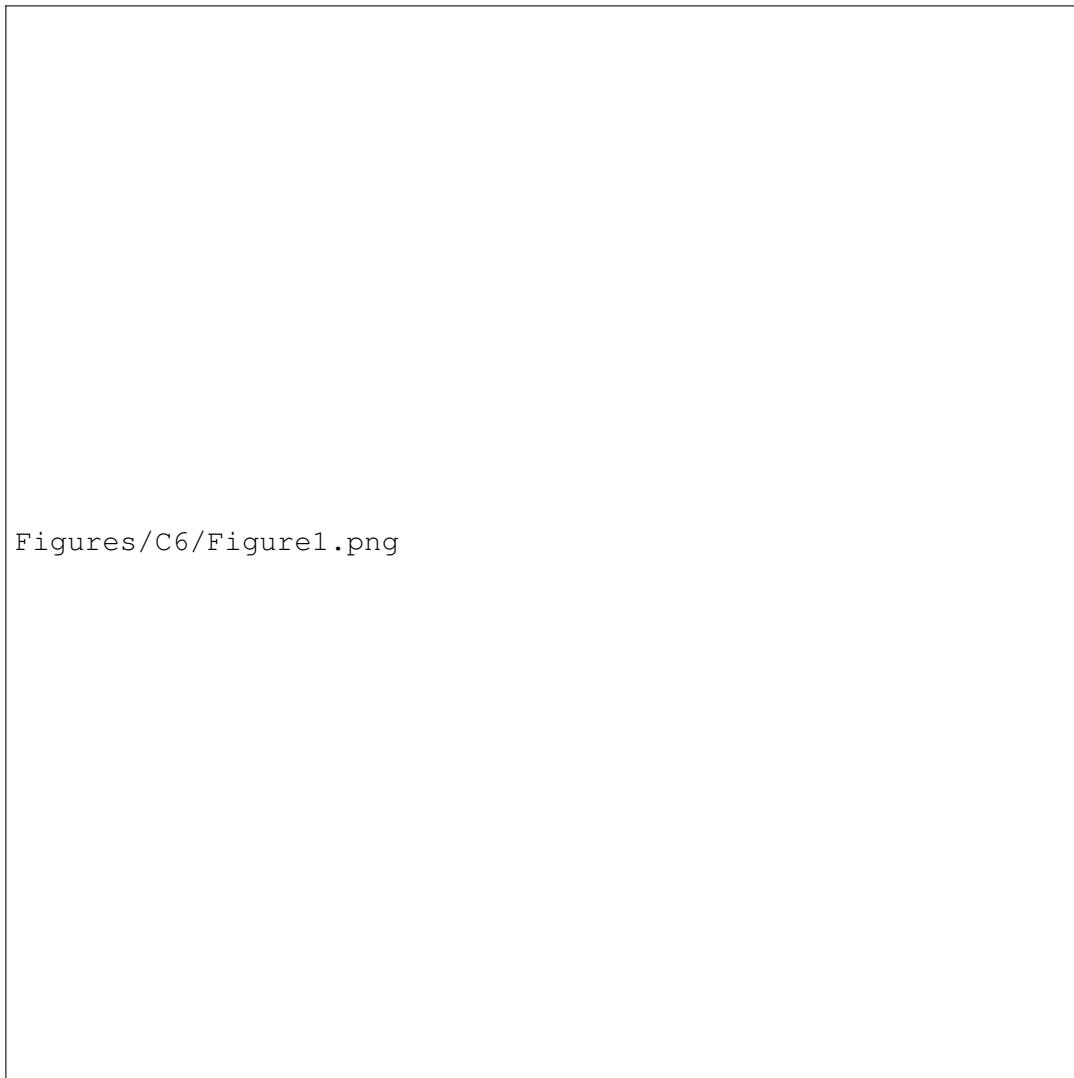
1 # StochasticGradient: training
2 # .....
3 # training accuracy = 0.84248771281362
4 # validation accuracy = 0.72571834721429
5 # current error = 1.89795655439887
6 # training accuracy = 0.84428662170892
7 # validation accuracy = 0.76952567752381
8 # .....

```

I risultati sono in figura 6.4. Come si nota, l'addestramento è andato a buon fine e la rete non ha dato problemi di overfitting, quindi il classificatore è effettivamente pronto per distinguere le api dalle formiche. Nella sezione 6.4.2 lo si mette alla prova.

### 6.4.2 Testing

È il momento di testare il classificatore su esempi ancora non visti. In precedenza si sono salvati su altri due file le immagini di testing; si caricano queste immagini nei tensori, gli si associa le label corrette e si utilizza la stessa funzione di valutazione



Figures/C6/Figure1.png

*Figura 6.4:* Curve di apprendimento sul nuovo dataset: il modello generalizza in maniera ottimale e non vi sono segni di overfitting

usata nel training per avere in uscita la categoria scelta dalla rete. Dopodiché si confronta questa risposta con la rispettiva label e si conta la percentuale di classificazioni corrette.

```

1 --load model and save prediction on unseen test data
2 model = torch.load(opt.save..'tuned.t7')
3 t1 = torch.load(opt.data..'test_ants.t7').features
4 t2 = torch.load(opt.data..'test_bees.t7').features
5 Z = torch.cat(t1,t2,1)
6
7 labels = torch.Tensor(t1:size(1)):fill(1)
8 lab2 = torch.Tensor(t2:size(1)):fill(2)
9 labels = torch.cat(labels, lab2)
10
11 dataset={}
12 function dataset:size() return Z:size(1) end
13
14 for i=1, dataset:size() do
15     dataset[i]={Z[i]:float()}
16 end
17

```

```
18 pred,_ = eval(model, dataset, false)
19
20 output={}
21 correct = 0
22 for i=1,#pred do
23     if labels[i] == pred[i] then
24         correct = correct + 1
25     end
26     output[i]={id=i,Label=pred[i]}
27 end
28 accuracy = correct / labels:size(1)
29 print('Testing on '..labels:size(1)..' unseen examples...')
30 print(string.format('#Test Accuracy: %.3f',accuracy))
31 --save results in a CSV file
32 csvigo.save('test.csv', output, true)
```

Eseguendo il test si ottiene l'output in figura 6.5.



Figura 6.5: Testing su 153 esempi ancora non visti. Accuracy = 95.4%

Si è quindi dimostrato come addestrare con risultati eccellenti un classificatore su un dataset arbitrario molto piccolo sfruttando una rete allo stato dell'arte.



## Chapter 7

# Conclusioni

Questa attività progettuale ha permesso sia di testare sul campo che di fornire un’opportunità di studio ed analisi ad ampio spettro delle reti neurali artificiali, che hanno fatto tanto parlare di sé negli ultimi anni. Dapprima, si sono esposte le basi teoriche delle reti neurali che gettano le loro radici a più di mezzo secolo fa; poi si è implementato da zero un percepitrone multi-strato cercando di "demistificarne" il funzionamento interno, essendo quest’ultimo poco intuitivo. Particolare enfasi è stata data alla spiegazione dell’algoritmo che è alla base dell’apprendimento delle reti, la **backpropagation of errors**. Dopodiché è stata fornita una panoramica sull’architettura delle **Convolutional Neural Networks**. Considerate lo stato dell’arte per la visione artificiale, hanno confermato le aspettative durante i test sui diversi dataset. Nel confronto fra le due architetture è emerso che avere un maggior numero di strati di convoluzione dona alla rete una maggiore potenza espressiva ed una rimarchevole capacità di astrazione e comprensione dei soggetti nelle immagini. Infine, il **transfer-learning** si è dimostrato una strategia di successo per ottenere classificatori ottimizzati per dataset arbitrari, anche piccoli.

Il framework **Torch** ha mantenuto le sue promesse sulla prototipazione flessibile e veloce. È un framework che può contare su un largo ecosistema di librerie guidate dalla comunità di ricercatori e l’astrazione che fornisce semplifica molto la progettazione e l’ottimizzazione delle reti, mantenendo al tempo stesso l’elasticità che i ricercatori necessitano per costruire modelli bleeding-edge. Tuttavia, non è facile eseguire il debugging di reti molto complesse ed, in un campo dove il trial-and-error è il pane quotidiano, sembra che ci possa essere ancora un margine di miglioramento.

Si noti, infine, che il progetto poteva essere realizzato anche con una delle molte altre librerie per il Deep Learning (TensorFlow, Keras, Caffe, DL4J) le quali presentano ciascuna le proprie peculiarità. Vale certamente la pena seguirne gli sviluppi e sperimentarne l’utilizzo, magari con applicazioni anche in ambiti diversi. Tutto ciò dimostra quante alternative il panorama del deep learning abbia da offrire e quanto questo campo sia in crescita esponenziale.



## Appendix A

# MLP: Codice addizionale

### A.1 Classi in Lua

In Lua manca il costrutto delle classi. Si può tuttavia crearle utilizzando tables e meta-tables. Per realizzare il multi-layer perceptron si è utilizzato una piccola libreria, di seguito riportata.

```

1 -- class.lua
2 -- Compatible with Lua 5.1 (not 5.0).
3 function class(base, init)
4     local c = {} -- a new class instance
5     if not init and type(base) == 'function' then
6         init = base
7         base = nil
8     elseif type(base) == 'table' then
9         -- our new class is a shallow copy of the base class!
10        for i,v in pairs(base) do
11            c[i] = v
12        end
13        c.__base = base
14    end
15    -- the class will be the metatable for all its objects,
16    -- and they will look up their methods in it.
17    c.__index = c
18
19    -- expose a constructor which can be called by <classname>(<args>)
20    local mt = {}
21    mt.__call = function(class_tbl, ...)
22        local obj = {}
23        setmetatable(obj,c)
24        if init then
25            init(obj,...)
26        else
27            -- make sure that any stuff from the base class is initialized!
28            if base and base.init then
29                base.init(obj, ...)
30            end
31        end
32        return obj
33    end
34    c.init = init
35    c.is_a = function(self, klass)
36        local m = getmetatable(self)
37        while m do
38            if m == klass then return true end
39            m = m.__base
40        end
41        return false
42    end
43    setmetatable(c, mt)

```

```

44     return c
45 end

```

## A.2 La classe Neural\_Network

Nel capitolo ?? si sono mostrati i vari snippet di codice man mano che si introducevano i concetti teorici che stanno alla base di questa implementazione. Di seguito è presentata l'intera classe Neural\_Network :

```

1 --creating class NN in Lua, using a nice class utility
2 class = require 'class'
3
4 Neural_Network = class('Neural_Network')
5
6 --init NN
7 function Neural_Network:__init(inputs, hiddens, outputs)
8     self.inputLayerSize = inputs
9     self.hiddenLayerSize = hiddens
10    self.outputLayerSize = outputs
11    self.W1 = th.randn(net.inputLayerSize, self.hiddenLayerSize)
12    self.W2 = th.randn(net.hiddenLayerSize, self.outputLayerSize)
13 end
14
15 --define a forward method
16 function Neural_Network:forward(X)
17     --Propagate inputs though network
18     self.z2 = th.mm(X, self.W1)
19     self.a2 = th.sigmoid(self.z2)
20     self.z3 = th.mm(self.a2, self.W2)
21     yHat = th.sigmoid(self.z3)
22     return yHat
23 end
24
25 function Neural_Network:d_Sigmoid(z)
26     --derivative of the sigmoid function
27     return th.exp(-z):cdiv( (th.pow( (1+th.exp(-z)), 2) ) )
28 end
29
30 function Neural_Network:costFunction(X, y)
31     --Compute the cost for given X,y, use weights already stored in class
32     self.yHat = self:forward(X)
33     --NB torch.sum() isn't equivalent to python sum() built-in method
34     --However, for 2D arrays whose one dimension is 1, it won't make any
35     --difference
36     J = 0.5 * th.sum(th.pow((y-yHat), 2))
37     return J
38 end
39
40 function Neural_Network:d_CostFunction(X, y)
41     --Compute derivative wrt to W1 and W2 for a given X and y
42     self.yHat = self:forward(X)
43     delta3 = th.cmul(-(y-self.yHat), self:d_Sigmoid(self.z3))
44     dJdW2 = th.mm(self.a2:t(), delta3)
45
46     delta2 = th.mm(delta3, self.W2:t()):cmul(self:d_Sigmoid(self.z2))
47     dJdW1 = th.mm(X:t(), delta2)
48
49     return dJdW1, dJdW2
50 end

```

### A.3 Metodi getter e setter

Nella sottosezione 2.4 del capitolo ?? si è dimostrato come calcolare numericamente il gradiente. Si è fatto cenno ai *getter e setter* per ottenere dei *flattened gradients*, ovvero "srotolare" i tensori dei gradienti in vettori monodimensionali. I metodi, qui mostrati, necessitano di una comprensione dei comandi di Torch. Data la maggiore popolarità di Python *Numpy*, nel caso il lettore fosse più familiare con quest'ultimo, nell'appendice B è mostrata anche una tabella di equivalenza dei metodi fra i due.

```

1 --Helper Functions for interacting with other classes:
2 function Neural_Network:getParams()
3     --Get W1 and W2 unrolled into a vector
4     params = th.cat((self.W1:view(self.W1:nElement()), (self.W2:view(
5         self.W2:nElement())))
6     return params
7 end
8
9 function Neural_Network:setParams(params)
10    --Set W1 and W2 using single parameter vector.
11    W1_start = 1 --index starts at 1 in Lua
12    W1_end = self.hiddenLayerSize * self.inputLayerSize
13    self.W1 = th.reshape(params[{ {W1_start, W1_end} }], (
14        self.inputLayerSize, self.hiddenLayerSize)
15    W2_end = W1_end + self.hiddenLayerSize*self.outputLayerSize
16    self.W2 = th.reshape(params[{ {W1_end+1, W2_end} }], (
17        self.hiddenLayerSize, self.outputLayerSize)
18 end
19
20 --this is like the getParameters(): method in the NN module of torch, i.e.
21 --compute the gradients and returns a flattened grads array
22 function Neural_Network:computeGradients(X, y)
23     dJdW1, dJdW2 = self:d_CostFunction(X, y)
24     return th.cat((dJdW1:view(dJdW1:nElement()), (dJdW2:view(dJdW2:
25         nElement())))
26 end

```



## Appendix B

# Il framework Torch

«««< HEAD

### B.1 Introduzione

Torch[37] è un framework per il calcolo numerico versatile che estende il linguaggio Lua. L’obiettivo è quello di fornire un ambiente flessibile per il progetto e addestramento di sistemi di machine learning anche su larga scala.

La flessibilità è ottenuta grazie a Lua stesso, un linguaggio di scripting estremamente leggero e efficiente. Le prestazioni sono garantite da backend compilati ed ottimizzati (C,C++,CUDA,OpenMP/SSE) per le routine di calcolo numerico di basso livello.

Gli obiettivi degli autori erano: (1) facilità di sviluppo di algoritmi numerici; (2) facilità di estensione, incluso il supporto ad altre librerie; (3) la velocità.

Gli obiettivi (2) e (3) sono stati soddisfatti tramite l’utilizzo di Lua poiché un linguaggio interpretato risulta conveniente per il testing rapido in modo interattivo; garantisce facilità di sviluppo e, grazie alle ottime C-API, unisce in modo eterogeneo le varie librerie, nascondendole sotto un’unica struttura di linguaggio di scripting. Infine, essendo ossessionati dalla velocità, è stato scelto Lua poiché è un veloce linguaggio di scripting e può inoltre contare su un efficiente compilatore JIT. Inoltre, Lua ha il grosso vantaggio di essere stato progettato per essere facilmente inserito nelle applicazioni scritte in C e consente quindi di “wrappare” le sottostanti implementazioni in C/C++ in maniera banale. Il binding C per il Lua è tra i più semplici e dona quindi grande estensibilità al progetto Torch.

Torch è un framework auto-contenuto e estremamente portabile su ogni piattaforma: iOS, Android, FPGA, processori DSP ecc. Gli script che vengono scritti per Torch riescono ad essere eseguiti su queste piattaforme senza nessuna modifica.

Per soddisfare il requisito (1) hanno invece ideato un oggetto chiamato ‘Tensor’, il quale altro non è che una “vista” geometrica di una particolare area di memoria, e permette una efficiente e semplice gestione di vettori a N dimensioni, tensori appunto. L’oggetto Tensor fornisce anche un’efficiente gestione della memoria: ogni operazione fatta su di esso non alloca nuova memoria, ma trasforma il tensore esistente o ritorna un nuovo tensore che referenzia la stessa area di memoria. Torch fornisce un ricco set di routine di calcolo numerico: le comuni routine di Matlab, algebra lineare, convoluzioni, FFT, ecc. Ci sono molti package per diversi ambiti: Machine Learning, Visione Artificiale, Image & Video Processing, Speech Recognition, ecc.

I package più importanti per il machine learning sono:

- **nn**: Neural Network, fornisce ogni sorta di modulo per la costruzione di reti neurali, reti neurali profonde (deep), regressione lineare, MLP, autoencoders ecc. È il package utilizzato nel progetto. Per topologie di rete bleeding-edge si suggerisce il package '**nmx**';
- **optim**: package per l'ottimizzazione della discesa del gradiente Fondamentale per avere buone performance nel training della rete;
- **unsup**: è un toolbox per l'apprendimento non supervisionato;
- **Image**: contiene tutte le funzioni atte all'image processing;
- **cunn**: package per utilizzare le reti neurali sfruttando la potenza di calcolo parallelo delle GPU, mediante l'architettura CUDA.

L'architettura del framework è raffigurata in figura B.1.

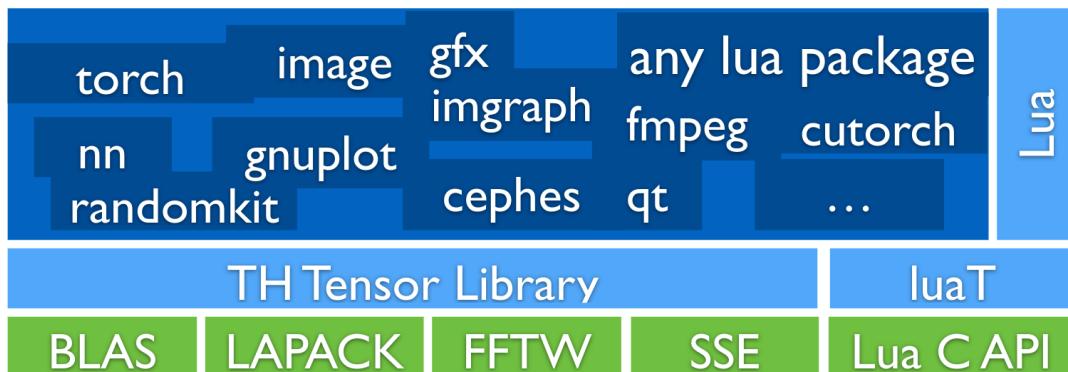


Figura B.1: Architettura del framework Torch

Torch è adottato da una fiorente comunità attiva di ricercatori in diverse università e importanti centri di ricerca come quelli di IBM; il dipartimento di IA di Facebook (FAIR); Google DeepMind prima di passare a TensorFlow nel 2016.

## B.2 Utilizzo base per reti neurali

Costruire modelli di reti neurali è una procedura rapida con Torch, ecco alcuni esempi:

```

1
2 -- simple linear model: logistic regression
3
4 model:add(nn.Reshape(3*32*32))
5 model:add(nn.Linear(3*32*32, #classes))
6
7 -- classic 2-layer fully-connected MLP
8
9
10 model:add(nn.Reshape(3*32*32))
11 model:add(nn.Linear(3*32*32, 1*32*32))
12 model:add(nn.ReLU())
13 model:add(nn.Linear(1*32*32, #classes))
14
15 -- convolutional layer
16
17

```

```

18 --hyper-parameters
19 nfeats = 3 --3D input volume
20 nstates = {16, 64, 128} --output at each level
21 filtsize = 5 --filter size or kernel
22 poolsize = 2
23
24 --Here's only the first stage.
25 --The others look the same except for the nstates you're gonna use
26
27 -- filter bank -> squashing -> max pooling
28 model:add(nn.SpatialConvolutionMM(nfeats, nstates[1], filtsize, filtsize))
29 model:add(nn.ReLU())
30 model:add(nn.SpatialMaxPooling(poolsize, poolsize, poolsize, poolsize))

```

### B.2.1 Supporto CUDA

CUDA (Compute Unified Device Architecture) è l'architettura di elaborazione in parallelo di NVIDIA che permette netti aumenti delle prestazioni di computing grazie allo sfruttamento della potenza di calcolo delle GPU per operazioni "general purpose". Torch offre un package chiamato 'cunn' per usufruire di CUDA. Il package è basato su un tensore chiamato 'torch.CudaTensor()' che altro non è che un normale Tensor che risiede ed utilizza la memoria della DRAM della GPU; tutte le operazioni definite per l'oggetto Tensor sono definite normalmente anche per il CudaTensor, il quale astrae completamente dall'utilizzo della GPU, offrendo un'interfaccia semplice e permettendo di sfruttare gli stessi script che si usano per l'elaborazione CPU. L'unica modifica da apportare, quindi, è cambiare il tipo di tensore.

```

1 tf = torch.FloatTensor(4,100,100) -- CPU's DRAM
2 tc = tf:cuda() -- GPU's DRAM
3 tc:mul() -- run on GPU
4 res = tc:float() -- res is instantiated on CPU's DRAM
5
6 --similarly, after we've built our model
7 --we can move it to the GPU by doing
8 model:cuda()
9 --we also need to compute our loss on GPU
10 criterion:cuda()
11
12 --now we're set, we can train our model on the GPU
13 --just by following the standard training procedure seen in (Capitolo 4)

```

## B.3 ResNet

Una rappresentazione testuale del modello a 18 strati di Residual Network.

```

1 nn.Sequential {
2   [input -> (1) -> (2) -> (3) -> (4) -> (5) -> (6) -> (7) -> (8) -> (9) ->
3     output]
4   (1): cudnn.SpatialConvolution(3 -> 16, 3x3, 1,1, 1,1) without bias
5   (2): nn.SpatialBatchNormalization (4D) (16)
6   (3): cudnn.ReLU
7   (4): nn.Sequential {
8     [input -> (1) -> (2) -> (3) -> output]
9     (1): nn.Sequential {
10       [input -> (1) -> (2) -> (3) -> output]
11       (1): nn.ConcatTable {
12         input
13         | -> (1): nn.Sequential {
14           [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]

```

```

14         |      (1): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
15     without bias
16         |      (2): nn.SpatialBatchNormalization (4D) (16)
17         |      (3): cudnn.ReLU
18         |      (4): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
19     without bias
20         |      (5): nn.SpatialBatchNormalization (4D) (16)
21         |
22     }
23 (2): nn.CAddTable
24     (3): cudnn.ReLU
25   }
26 (2): nn.Sequential {
27   [input -> (1) -> (2) -> (3) -> output]
28   (1): nn.ConcatTable {
29     input
30     |`-> (1): nn.Sequential {
31       |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
32       |  (1): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
33     without bias
34       |  (2): nn.SpatialBatchNormalization (4D) (16)
35       |  (3): cudnn.ReLU
36       |  (4): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
37     without bias
38       |  (5): nn.SpatialBatchNormalization (4D) (16)
39       |
40     }
41   (2): nn.CAddTable
42   (3): cudnn.ReLU
43   }
44 (3): nn.Sequential {
45   [input -> (1) -> (2) -> (3) -> output]
46   (1): nn.ConcatTable {
47     input
48     |`-> (1): nn.Sequential {
49       |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
50       |  (1): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
51     without bias
52       |  (2): nn.SpatialBatchNormalization (4D) (16)
53       |  (3): cudnn.ReLU
54       |  (4): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
55     without bias
56       |  (5): nn.SpatialBatchNormalization (4D) (16)
57       |
58     }
59   (2): nn.CAddTable
60   (3): cudnn.ReLU
61   }
62 }
63 (5): nn.Sequential {
64   [input -> (1) -> (2) -> (3) -> output]
65   (1): nn.Sequential {
66     [input -> (1) -> (2) -> (3) -> output]
67     (1): nn.ConcatTable {
68       input
69       |`-> (1): nn.Sequential {
70         |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]

```

```

71      |      (1): cudnn.SpatialConvolution(16 -> 32, 3x3, 2,2, 1,1)
72      without bias
73      |      (2): nn.SpatialBatchNormalization (4D) (32)
74      |      (3): cudnn.ReLU
75      |      (4): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
76      without bias
77      |      (5): nn.SpatialBatchNormalization (4D) (32)
78      |
79      `-> (2): nn.Sequential {
80          [input -> (1) -> (2) -> output]
81          (1): nn.SpatialAveragePooling(1x1, 2,2)
82          (2): nn.Concat {
83              input
84              |`-> (1): nn.Identity
85              |`-> (2): nn.MulConstant
86              ... -> output
87          }
88      }
89      (2): nn.CAddTable
90      (3): cudnn.ReLU
91  }
92  (2): nn.Sequential {
93      [input -> (1) -> (2) -> (3) -> output]
94      (1): nn.ConcatTable {
95          input
96          |`-> (1): nn.Sequential {
97              [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
98              (1): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
99          without bias
100         |      (2): nn.SpatialBatchNormalization (4D) (32)
101         |      (3): cudnn.ReLU
102         |      (4): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
103         without bias
104         |      (5): nn.SpatialBatchNormalization (4D) (32)
105         |
106         `-> (2): nn.Identity
107         ... -> output
108     }
109     (2): nn.CAddTable
110     (3): cudnn.ReLU
111  }
112  (3): nn.Sequential {
113      [input -> (1) -> (2) -> (3) -> output]
114      (1): nn.ConcatTable {
115          input
116          |`-> (1): nn.Sequential {
117              [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
118              (1): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
119          without bias
120          |      (2): nn.SpatialBatchNormalization (4D) (32)
121          |      (3): cudnn.ReLU
122          |      (4): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
123          without bias
124          |      (5): nn.SpatialBatchNormalization (4D) (32)
125          |
126          `-> (2): nn.Identity
127          ... -> output
128      }
129      (2): nn.CAddTable
130      (3): cudnn.ReLU
131  }

```

```

128 }
129 (6): nn.Sequential {
130   [input -> (1) -> (2) -> (3) -> output]
131   (1): nn.Sequential {
132     [input -> (1) -> (2) -> (3) -> output]
133     (1): nn.ConcatTable {
134       input
135         |`-> (1): nn.Sequential {
136           |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
137           |  (1): cudnn.SpatialConvolution(32 -> 64, 3x3, 2,2, 1,1)
138       without bias
139         |  (2): nn.SpatialBatchNormalization (4D) (64)
140         |  (3): cudnn.ReLU
141         |  (4): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
142       without bias
143         |  (5): nn.SpatialBatchNormalization (4D) (64)
144         |
145       `-> (2): nn.Sequential {
146         [input -> (1) -> (2) -> output]
147         (1): nn.SpatialAveragePooling(1x1, 2,2)
148         (2): nn.Concat {
149           input
150             |`-> (1): nn.Identity
151             |`-> (2): nn.MulConstant
152             ... -> output
153           }
154         ...
155       }
156     (2): nn.CAddTable
157     (3): cudnn.ReLU
158   }
159   (2): nn.Sequential {
160     [input -> (1) -> (2) -> (3) -> output]
161     (1): nn.ConcatTable {
162       input
163         |`-> (1): nn.Sequential {
164           |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
165           |  (1): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
166       without bias
167         |  (2): nn.SpatialBatchNormalization (4D) (64)
168         |  (3): cudnn.ReLU
169         |  (4): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
170       without bias
171         |  (5): nn.SpatialBatchNormalization (4D) (64)
172         |
173       `-> (2): nn.Identity
174       ...
175     }
176   (2): nn.CAddTable
177   (3): cudnn.ReLU
178 }
179 (3): nn.Sequential {
180   [input -> (1) -> (2) -> (3) -> output]
181   (1): nn.ConcatTable {
182     input
183       |`-> (1): nn.Sequential {
184         |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
185         |  (1): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
186     without bias
187       |  (2): nn.SpatialBatchNormalization (4D) (64)
188       |  (3): cudnn.ReLU

```

```

185         |      (4): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
186     without bias
187     |      (5): nn.SpatialBatchNormalization (4D) (64)
188     |      }
189     `-> (2): nn.Identity
190     ... -> output
191   }
192   (2): nn.CAddTable
193   (3): cudnn.ReLU
194   }
195   (7): cudnn.SpatialAveragePooling(8x8, 1,1)
196   (8): nn.View(64)
197   (9): nn.Linear(64 -> 10)
198 }
199 =====
200 \section{Introduzione}
201 \parencite{WTorch}
202
203 \section{Utilizzo base per reti neurali}
204 \subsection{Supporto CUDA}
205
206 \section{ResNet}
207 Una rappresentazione testuale del modello a 18 strati di Residual Network.
208
209 \begin{lstlisting}[language={[5.2]Lua}]
210 nn.Sequential {
211   [input -> (1) -> (2) -> (3) -> (4) -> (5) -> (6) -> (7) -> (8) -> (9) ->
212     output]
213   (1): cudnn.SpatialConvolution(3 -> 16, 3x3, 1,1, 1,1) without bias
214   (2): nn.SpatialBatchNormalization (4D) (16)
215   (3): cudnn.ReLU
216   (4): nn.Sequential {
217     [input -> (1) -> (2) -> (3) -> output]
218     (1): nn.Sequential {
219       [input -> (1) -> (2) -> (3) -> output]
220       (1): nn.ConcatTable {
221         input
222         | `-> (1): nn.Sequential {
223           |      [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
224           |      (1): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
225         without bias
226           |      (2): nn.SpatialBatchNormalization (4D) (16)
227           |      (3): cudnn.ReLU
228           |      (4): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
229         without bias
230           |      (5): nn.SpatialBatchNormalization (4D) (16)
231           |      }
232           `-> (2): nn.Identity
233           ... -> output
234         }
235       (2): nn.CAddTable
236       (3): cudnn.ReLU
237     }
238     (2): nn.Sequential {
239       [input -> (1) -> (2) -> (3) -> output]
240       (1): nn.ConcatTable {
241         input
242         | `-> (1): nn.Sequential {
243           |      [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]

```

```

243           |      (1): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
244           without bias
245           |      (2): nn.SpatialBatchNormalization (4D) (16)
246           |      (3): cudnn.ReLU
247           |      (4): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
248           without bias
249           |      (5): nn.SpatialBatchNormalization (4D) (16)
250           |
251           ... -> output
252       }
253   (2): nn.CAddTable
254   (3): cudnn.ReLU
255 }
256 (3): nn.Sequential {
257   [input -> (1) -> (2) -> (3) -> output]
258   (1): nn.ConcatTable {
259     input
260     |`-> (1): nn.Sequential {
261       [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
262       |      (1): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
263       without bias
264       |      (2): nn.SpatialBatchNormalization (4D) (16)
265       |      (3): cudnn.ReLU
266       |      (4): cudnn.SpatialConvolution(16 -> 16, 3x3, 1,1, 1,1)
267       without bias
268       |      (5): nn.SpatialBatchNormalization (4D) (16)
269       |
270       ... -> output
271     }
272   (2): nn.CAddTable
273   (3): cudnn.ReLU
274 }
275 (5): nn.Sequential {
276   [input -> (1) -> (2) -> (3) -> output]
277   (1): nn.Sequential {
278     [input -> (1) -> (2) -> (3) -> output]
279     (1): nn.ConcatTable {
280       input
281       |`-> (1): nn.Sequential {
282         [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
283         |      (1): cudnn.SpatialConvolution(16 -> 32, 3x3, 2,2, 1,1)
284       without bias
285       |      (2): nn.SpatialBatchNormalization (4D) (32)
286       |      (3): cudnn.ReLU
287       |      (4): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
288       without bias
289       |      (5): nn.SpatialBatchNormalization (4D) (32)
290       |
291       ... -> (2): nn.Sequential {
292         [input -> (1) -> (2) -> output]
293         (1): nn.SpatialAveragePooling(1x1, 2,2)
294         (2): nn.Concat {
295           input
296             |`-> (1): nn.Identity
297             |`-> (2): nn.MulConstant
298             ... -> output
299           }
300         }
301       ... -> output
302     }
303   }
304 }
```

```

300     (2): nn.CAddTable
301     (3): cudnn.ReLU
302   }
303 (2): nn.Sequential {
304   [input -> (1) -> (2) -> (3) -> output]
305   (1): nn.ConcatTable {
306     input
307     |`-> (1): nn.Sequential {
308       |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
309       |  (1): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
310     without bias
311       |  (2): nn.SpatialBatchNormalization (4D) (32)
312       |  (3): cudnn.ReLU
313       |  (4): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
314     without bias
315       |  (5): nn.SpatialBatchNormalization (4D) (32)
316     ... -> output
317   }
318   (2): nn.CAddTable
319   (3): cudnn.ReLU
320 }
321 (3): nn.Sequential {
322   [input -> (1) -> (2) -> (3) -> output]
323   (1): nn.ConcatTable {
324     input
325     |`-> (1): nn.Sequential {
326       |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
327       |  (1): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
328     without bias
329       |  (2): nn.SpatialBatchNormalization (4D) (32)
330       |  (3): cudnn.ReLU
331       |  (4): cudnn.SpatialConvolution(32 -> 32, 3x3, 1,1, 1,1)
332     without bias
333       |  (5): nn.SpatialBatchNormalization (4D) (32)
334     ... -> output
335   }
336   (2): nn.CAddTable
337   (3): cudnn.ReLU
338 }
339 }
340 (6): nn.Sequential {
341   [input -> (1) -> (2) -> (3) -> output]
342   (1): nn.Sequential {
343     [input -> (1) -> (2) -> (3) -> output]
344     (1): nn.ConcatTable {
345       input
346       |`-> (1): nn.Sequential {
347         |  [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
348         |  (1): cudnn.SpatialConvolution(32 -> 64, 3x3, 2,2, 1,1)
349     without bias
350       |  (2): nn.SpatialBatchNormalization (4D) (64)
351       |  (3): cudnn.ReLU
352       |  (4): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
353     without bias
354       |  (5): nn.SpatialBatchNormalization (4D) (64)
355       |
356     ... -> (2): nn.Sequential {
357       [input -> (1) -> (2) -> output]
358       (1): nn.SpatialAveragePooling(1x1, 2,2)

```

```

357             (2): nn.Concat {
358                 input
359                     | `-> (1): nn.Identity
360                     `-> (2): nn.MulConstant
361                     ... -> output
362             }
363         }
364     ...
365   }
366   (2): nn.CAddTable
367   (3): cudnn.ReLU
368 }
369 (2): nn.Sequential {
370     [input -> (1) -> (2) -> (3) -> output]
371     (1): nn.ConcatTable {
372         input
373             | `-> (1): nn.Sequential {
374                 |     [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
375                 |     (1): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
376             without bias
377                 |     (2): nn.SpatialBatchNormalization (4D) (64)
378                 |     (3): cudnn.ReLU
379                 |     (4): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
380             without bias
381                 |     (5): nn.SpatialBatchNormalization (4D) (64)
382                 |
383             `-> (2): nn.Identity
384             ...
385             -> output
386         }
387         (2): nn.CAddTable
388         (3): cudnn.ReLU
389     }
390     (3): nn.Sequential {
391         [input -> (1) -> (2) -> (3) -> output]
392         (1): nn.ConcatTable {
393             input
394                 | `-> (1): nn.Sequential {
395                     |     [input -> (1) -> (2) -> (3) -> (4) -> (5) -> output]
396                     |     (1): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
397             without bias
398                 |     (2): nn.SpatialBatchNormalization (4D) (64)
399                 |     (3): cudnn.ReLU
400                 |     (4): cudnn.SpatialConvolution(64 -> 64, 3x3, 1,1, 1,1)
401             without bias
402                 |     (5): nn.SpatialBatchNormalization (4D) (64)
403                 |
404             `-> (2): nn.Identity
405             ...
406             -> output
407         }
408         (2): nn.CAddTable
409         (3): cudnn.ReLU
410     }
411 }
412 (7): cudnn.SpatialAveragePooling(8x8, 1,1)
413 (8): nn.View(64)
414 (9): nn.Linear(64 -> 10)
415 }
416 >>>>> ece01a6c88b2ea9153728cdbf63dcf2c83a18f6a

```

# Bibliography

- [1] Wikipedia. (2016). Percettrone, [Online]. Available: <http://it.wikipedia.org/w/index.php?title=Percettrone&oldid=88722444> (visited on 08/20/2017).
- [2] ——, (2016). Universal Approximation Theorem, [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Low-rank\\_approximation&oldid=822406390](https://en.wikipedia.org/w/index.php?title=Low-rank_approximation&oldid=822406390) (visited on 01/03/2018).
- [3] (2016). Gradisca new york, [Online]. Available: [www.gradiscanyyc.com](http://www.gradiscanyyc.com) (visited on 08/20/2017).
- [4] A. Central. (2016). The average profit margin for a restaurant, [Online]. Available: <http://yourbusiness.azcentral.com/average-profit-margin-restaurant-13113.html> (visited on 08/20/2017).
- [5] R. R. Group. (2016). The restaurant financial red flags, [Online]. Available: [http://rrgconsulting.com/ten\\_restaurant\\_financial\\_red\\_flags.htm](http://rrgconsulting.com/ten_restaurant_financial_red_flags.htm) (visited on 08/20/2017).
- [6] Wikipedia. (2016). The Sigmoid Function, [Online]. Available: [https://en.wikipedia.org/wiki/Sigmoid\\_function](https://en.wikipedia.org/wiki/Sigmoid_function) (visited on 08/20/2017).
- [7] ——, (2016). Supervised learning, [Online]. Available: [https://en.wikipedia.org/wiki/Supervised\\_learning](https://en.wikipedia.org/wiki/Supervised_learning) (visited on 08/20/2017).
- [8] StackExchange. (2016). List of cost function used in Neural Networks applications, [Online]. Available: <https://stats.stackexchange.com/questions/154879/a-list-of-cost-functions-used-in-neural-networks-alongside-applications> (visited on 08/20/2017).
- [9] Wikipedia. (2016). Regola della catena, [Online]. Available: [https://it.wikipedia.org/wiki/Regola\\_della\\_catena](https://it.wikipedia.org/wiki/Regola_della_catena) (visited on 08/20/2017).
- [10] CS231-Stanford. (2016). Vector, matrix, and tensor derivatives, [Online]. Available: <http://cs231n.stanford.edu/vecDerivs.pdf> (visited on 08/23/2017).
- [11] Stanford. (2016). Gradient checking and advanced optimization, [Online]. Available: [http://ufldl.stanford.edu/wiki/index.php/Gradient\\_checking\\_and\\_advanced\\_optimization](http://ufldl.stanford.edu/wiki/index.php/Gradient_checking_and_advanced_optimization) (visited on 08/20/2017).
- [12] Wikipedia. (2016). Stochastic Gradient Descent, [Online]. Available: [https://en.wikipedia.org/wiki/Stochastic\\_gradient\\_descent](https://en.wikipedia.org/wiki/Stochastic_gradient_descent) (visited on 08/20/2017).
- [13] ——, (2016). The Quasi-Newton method, [Online]. Available: [https://en.wikipedia.org/wiki/Quasi-Newton\\_method](https://en.wikipedia.org/wiki/Quasi-Newton_method) (visited on 08/20/2017).
- [14] ——, (2016). Limited-memory bfgs, [Online]. Available: [https://en.wikipedia.org/wiki/Limited-memory\\_BFGS](https://en.wikipedia.org/wiki/Limited-memory_BFGS) (visited on 08/20/2017).

- [15] M. Mastery. (2017). A gentle introduction to adam, [Online]. Available: <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/> (visited on 08/25/2017).
- [16] CS231-Stanford. (2017). Cnn for visual recognition, [Online]. Available: <http://cs231n.github.io/neural-networks-3/> (visited on 08/25/2017).
- [17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting”, vol. 15, pp. 1929–1958, Jun. 2014.
- [18] M. Lippi. (2017). Machine learning phd course, [Online]. Available: [http://lia.disi.unibo.it/Staff/MarcoLippi/teaching/ml\\_phd.html](http://lia.disi.unibo.it/Staff/MarcoLippi/teaching/ml_phd.html) (visited on 08/25/2017).
- [19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, in *Intelligent Signal Processing*, IEEE Press, 2001, pp. 306–351.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105, 2012. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>.
- [22] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks”, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudík, Eds., ser. Proceedings of Machine Learning Research, vol. 15, Fort Lauderdale, FL, USA: PMLR, Nov. 2011, pp. 315–323. [Online]. Available: <http://proceedings.mlr.press/v15/glorot11a.html>.
- [23] CS231-Stanford. (2016). Convert fc to convolution, [Online]. Available: <https://cs231n.github.io/convolutional-networks/#convert> (visited on 08/29/2017).
- [24] Wikipedia. (2016). "convolutional neural networks", [Online]. Available: [https://en.wikipedia.org/wiki/Convolutional\\_neural\\_network#Applications](https://en.wikipedia.org/wiki/Convolutional_neural_network#Applications) (visited on 08/20/2017).
- [25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition”, *CoRR*, vol. abs/1403.6382, 2014. [Online]. Available: <http://arxiv.org/abs/1403.6382>.
- [26] A. Karpathy. (2016). "what i learned from competing against a convnet on imagenet", [Online]. Available: <http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/> (visited on 08/30/2017).
- [27] M. T. Review. (2015). The face detection algorithm set to revolutionize image search, [Online]. Available: <https://www.technologyreview.com/s/535201/the-face-detection-algorithm-set-to-revolutionize-image-search/> (visited on 08/29/2017).

- [28] Wired. (2016). In a huge breakthrough, google's ai beats a top player at the game of go, [Online]. Available: <https://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/> (visited on 08/29/2017).
- [29] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search", *Nature*, vol. 529, no. 7587, pp. 484–489, jan 2016, issn: 0028-0836. doi: 10.1038/nature16961.
- [30] Wikipedia. (2018). <https://en.wikipedia.org/w/index.php?title=Tensoroldid=826559352>, [Online]. Available: <https://en.wikipedia.org/wiki/Tensor?oldformat=true> (visited on 01/03/2018).
- [31] alex gossmann. (2016). Understanding tucker decomposition, [Online]. Available: [http://www.alexejgossmann.com/tensor\\_decomposition\\_tucker/](http://www.alexejgossmann.com/tensor_decomposition_tucker/) (visited on 01/03/2018).
- [32] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition", *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.
- [33] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications", *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
- [34] A. Veit, M. J. Wilber, and S. J. Belongie, "Residual networks are exponential ensembles of relatively shallow networks", *CoRR*, vol. abs/1605.06431, 2016. [Online]. Available: <http://arxiv.org/abs/1605.06431>.
- [35] B. Vision and L. Center. (2016). "model zoo", [Online]. Available: <https://github.com/BVLC/caffe/wiki/Model-Zoo> (visited on 08/30/2017).
- [36] Wikipedia. (2016). "la funzione softmax", [Online]. Available: [https://it.wikipedia.org/wiki/Funzione\\_softmax](https://it.wikipedia.org/wiki/Funzione_softmax) (visited on 08/20/2017).
- [37] (2017). Torch: A scientific computing framework for luajit, [Online]. Available: [torch.ch](http://torch.ch) (visited on 08/20/2017).