



UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea triennale in *Ingegneria Informatica e dell'Automazione*

Sviluppo di un modello GAN per la generazione di immagini e relativa segmentazione

Develop of a GAN model for image generation with related segmentations

Relatore:

Prof. Adriano Mancini

Laureando:

Massimiliano Biancucci

Anno Accademico 2022/2023

Prefazione

Il mio percorso nel campo dell'intelligenza artificiale è iniziato diversi anni fa, alle superiori per l'esattezza, dove sentii per la prima volta parlare di reti neurali, ad un corso pomeridiano voluto dal prof. Roberto Lulli il quale mi ha mostrato per primo questo affascinante campo di ricerca.

Ho svolto durante il mio percorso di studi diversi progetti incentrati su questa tematica, partendo da semplici reti neurali, e confrontandomi con progetti sempre più complessi fino ad arrivare ai modelli generativi basati sull'architettura GAN (Generative Adversarial Network), del quale in questa tesi proporò una variante.

Lo scopo di questa tesi è quello di investigare la fattibilità di una potenziale soluzione ad uno dei grandi problemi che affligge oggi le aziende che si occupano di addestrare modelli neurali per la segmentazione di immagini, ovvero la difficoltà nel reperire immagini annotate, le quali hanno elevatissimi costi di realizzazione.

Tale scelta è stata naturale, in quanto ho dovuto confrontarmi in prima persona con questo problema nell'ultimo anno, come sviluppatore presso l'azienda Cloe.ai. In questa esperienza ho gestito per quasi un anno la realizzazione di un complesso dataset per l'addestramento di un modello di segmentazione di difetti, tale dataset aveva dei requisiti molto alti, e al contempo le risorse per realizzarlo erano limitate, per tale ragione ho potuto comprendere affondo le problematiche legate a questo tipo di progetto, facendo i conti io stesso con le spese e i progressi ottenuti.

La realizzazione di un dataset su larga scala è un'operazione molto complessa, che richiede una elevata coordinazione tra annotatori, revisori, sviluppatori, e un'accurata documentazione che in base al problema può richiedere anche diversi mesi per poter essere redatta efficacemente. Tutto ciò mi ha fornito la motivazione per cercare una soluzione per accorciare questo lungo e tedioso processo e dunque attenuare gli ingenti costi che un'azienda deve sostenere per realizzare un dataset di questo tipo.

Indice

Prefazione	ii
Indice	iii
1 Introduzione	1
1.1 Motivazione	1
1.1.1 Modelli neurali per la segmentazione nel controllo qualità	1
1.1.2 Il dataset, requisiti e problematiche di realizzazione	2
1.1.3 Approccio al problema	3
1.2 La nascita del deep learning	6
1.2.1 Dal machine learning al deep learning	6
1.2.2 Le reti neurali biologiche	7
1.3 Le reti neurali feed forward	7
1.3.1 Il neurone artificiale	8
1.3.2 Il Back-propagation	10
1.3.3 Teorema di approssimazione universale	12
1.4 Le reti neurali convoluzionali	13
1.4.1 Storia delle CNN	13
1.4.2 La convoluzione	15
1.4.3 I parametri della convoluzione	18
1.4.4 Il pooling	20
1.4.5 La convoluzione multichannel	22
2 Stato dell'arte	23
2.1 Il primo modello gan	23
3 Strumenti e metodi	24
3.1 Il Dataset: Severstal steel defect detection	24

3.2	Lama:	24
3.3	Stylegan2:	24
4	Sviluppo del progetto	25
5	Risultati	26
	Elenco delle figure	27
	Bibliografia	29

Capitolo 1

Introduzione

1.1 Motivazione

1.1.1 Modelli neurali per la segmentazione nel controllo qualità

Oggi le reti neurali trovano un vasto impiego in moltissimi campi, dall'industria alla medicina, fino alla vita di tutti i giorni. Il grande vantaggio che ci portano è la capacità di apprendere da un set di dati, e di generalizzare su di uno nuovo, permettendoci di risolvere problemi che altrimenti sarebbero matematicamente troppo complessi da risolvere con un algoritmo. Ci sono vari esempi in cui i modelli neurali raggiungono risultati superiori a quelli ottenuti dall'uomo, in determinati task, o almeno se non lo superano in termini di accuratezza, lo fanno in termini di velocità, scalabilità, costi e prestazioni.

Un task in cui le reti neurali eccellono è la segmentazione di immagini, ovvero la classificazione pixel per pixel di un'immagine, questo tipo di task è utilizzato ad esempio nel campo medico per la segmentazione di organi, tumori, o in campo industriale per la segmentazione di difetti, per la verifica automatica della qualità di un prodotto o di un semilavorato.

Nel caso specifico, per la segmentazione dei difetti l'utilizzo di questo tipo di modelli è molto diffuso, in quanto risolve un grave problema che affligge i reparti controllo qualità delle aziende, ovvero il calo della concentrazione al quale un operatore è soggetto dopo un certo numero di ore di lavoro. Infatti una persona per quanto allenata e preparata, dopo un certo numero di ore di lavoro, è soggetta a stanchezza e con essa la sua accuratezza nel riconoscere un difetto diminuisce, mentre un modello neurale adeguatamente addestrato, in condizioni ambientali stabili, come ad esempio una adeguata illuminazione, una videocamera ad alta risoluzione e un'adeguata distanza dal soggetto, sarà in grado di mantenere un'accuratezza costante, senza necessità di fermarsi per riposare. Questo si traduce in un risparmio di tempo e di denaro per l'azienda, In quanto il controllo manuale richiede più tempo ed è più soggetto ad errori, i quali spesso si trasformano in ritardi nella consegna dei prodotti, spese di trasporto aggiuntive per il ritorno o la sostituzione del prodotto, o addirittura la perdita di un cliente.

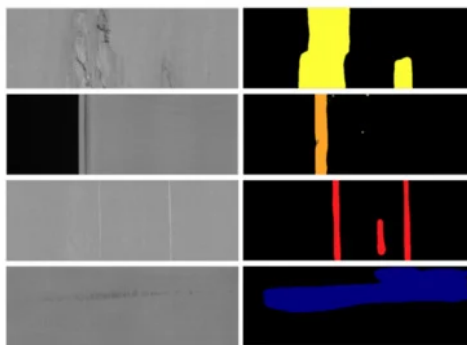


Figura 1.1: An example of segmented defects from the "Severstal steel defect dataset". credits: Neven Robby and Goedemé Toon, 2021, A Multi-Branch U-Net for Steel Surface Defect Type and Severity Segmentation. <https://www.mdpi.com/2075-4701/11/6/870>

1.1.2 Il dataset, requisiti e problematiche di realizzazione

La problematica di avere un modello con elevata accuratezza per task di segmentazione è relativa alla quantità e qualità dei dati necessari, i quali raramente sono disponibili opensource o per l'acquisto, rendendo necessaria la creazione di un dataset apposito. Molti task richiedono un grande quantità di dati per essere generalizzati correttamente, e ogni singolo esempio richiede molta concentrazione da parte dell'annotatore in quanto non sempre i difetti sono ben visibili.

In generale la creazione di un dataset è un'operazione molto complessa, non solo dal punto di vista dell'annotatore, ma in oltre da un punto di vista organizzativo e logistico. Infatti ci sono diversi step che si devono seguire:

- **Acquisizione delle immagini:** Le immagini devono essere acquisite in modo da avere una buona qualità, o almeno sufficiente ai fini dell'apprendimento. Se possibile in oltre dovrebbero avere una adeguata uniformità di condizioni (luce, distanza, ...) per garantire le migliori prestazioni da parte del modello, ovviamente solo se poi è possibile garantire le stesse condizioni anche nell'utilizzo finale del modello, altrimenti una grande varietà delle condizioni è preferibile.
- **Definizione delle classi:** Nel caso di dataset multi-classe, uno step molto importante è quello di scegliere accuratamente le classi e definire in maniera univoca l'associazione tra una classe e una particolare tipologia di difetto. Questo passaggio potrebbe sembrare banale ma in realtà nasconde delle grandi insidie, infatti una classificazione non adeguata andrà a causare confusione nel modello, diminuendo la sua accuratezza e/o rendendo il lavoro più difficile per gli annotatori andando a rallentare il processo di annotazione o comunque a ridurne la qualità. Questo tipo di problematiche purtroppo si manifestano chiaramente soltanto in uno stato avanzato del progetto, rendendo necessarie revisioni della documentazione, modifica di tutti gli esempi già annotati, con conseguente perdita di tempo e denaro.
- **Definizione della documentazione:** Questo passaggio è un'estensione del precedente, e consiste nella definizione di una documentazione che specifichi senza ambiguità, ad un nuovo annotatore come riconoscere senza dubbio un difetto e classificarlo nella giusta

classe. Questa fase spesso non termina prima dell'inizio dell'annotazione, ma si protrae per tutta la durata del progetto, in quanto spesso nuovi casi non previsti si presentano durante l'annotazione, e la documentazione deve essere aggiornata in tempo reale.

- **Annotazione:** Questo è il passaggio più lungo e costoso, in quanto richiede una squadra di persone, che devono essere formate per lo specifico task, e che devono essere costantemente seguite per garantire la qualità del lavoro.
- **Revisione:** Assieme all'annotazione questo è un passaggio chiave, in quanto permette di verificare che l'annotazione sia stata fatta correttamente, e che non ci siano errori nell'annotazione. Spesso infatti gli annotatori acquisiscono dei bias errati nei confronti di una certa classe, o di un certo tipo di difetto, che deve essere identificato e reso noto all'annotatore per correggerlo, ed evitare che questo errore si ripeta in futuro. Per evitare che ciò accada oltre al primo annotatore lo stesso esempio viene solitamente rivisto da 2 o 4 persone diverse. Si noti che gli errori degli annotatori che non vengono identificati verranno appresi dal modello finale come una corretta classificazione, ciò giustifica un tale dispendio di risorse in questa fase.

1.1.3 Approccio al problema

La creazione di un dataset come precedentemente illustrato è un processo complesso e dispendioso, che richiede molte risorse umane e finanziarie, dunque l'intento in questo progetto è quello di proporre un approccio alternativo che sia in grado di ridurre per quanto possibile la durata e il costo di questo lavoro. Partendo dal presupposto che almeno in parte il dataset deve essere realizzato manualmente, la proposta è quella di realizzare una certa quantità di campioni manualmente seguendo lo schema già visto, per poi addestrare un modello neurale per generare ulteriori esempi sintetici, raggiungendo un numero di esempi totali che permetta di addestrare un modello con buone prestazioni, ad un costo ridotto rispetto al caso in cui tutti i dati fossero stati realizzati manualmente.

Per la definizione della pipeline di generazione dei dati, si è partiti dal concetto di *generative adversarial network* (GAN), che è una tecnica di machine learning che permette di generare dati sintetici utilizzando come base dati reali, tali dati sintetici possono essere utilizzati per addestrare un modello neurale. Tale tecnica ha trovato riscontri positivi in molte ricerche pubblicate in ambito di computer vision [4], in cui i modelli GAN vengono utilizzati per espandere il numero di immagini presenti in un dataset e migliorare la generalizzazione di un modello di classificazione. Ovviamente gli aumenti di accuratezza, precisione e recall dipendono dal numero di esempi presenti nel dataset e dalla complessità del problema. Tale tecnica potrebbe essere considerata una versione più sofisticata di data augmentation, in quanto permette di generare dati sintetici molto più complessi e realistici di quelli che si possono ottenere con semplici trasformazioni geometriche o matematiche. Per generare dati utilizzabili per addestrare un modello di segmentazione però è necessario risolvere un'ulteriore problema, infatti un normale modello GAN, fedele alla sua definizione originale [8], è in grado di generare intere immagini, che possono essere utilizzate per addestrare un modello di classificazione, ma non sono utilizzabili per addestrare un modello di segmentazione, in quanto per l'addestramento

di tale architettura è necessario che gli oggetti di interesse abbiano una maschera che specifichi la loro posizione. Ci sono vari approcci di augmentation per la segmentazione che risultano molto più semplici di addestrare un modello GAN, come il caso del metodo "copy paste" [7], il quale propone come augmentation per i dataset di segmentazione la copia di un oggetto presente in un'immagine, ritagliandolo attraverso la sua maschera, e incollandolo su di un nuovo background potenzialmente in una nuova posizione, tale metodo risulta estremamente efficace per oggetti indipendenti dal contesto con contorni ben definiti, ma risulta inutile nel momento in cui l'oggetto che vogliamo generare ha una interdipendenza forte con l'area immediatamente circostante, pensiamo ad esempio un difetto su di un'auto, un graffio o una bozza, non potrà essere copiato da un'auto e incollato su di un'altra in quanto subentreranno una serie di di artifacts, come la variazione netta di colore tra l'auto e il difetto, rischiando di introdurre un bias nel modello, il quale finirebbe per cercare la variazione netta di colore e non più le features del difetto. Per risolvere questo problema con questa particolare categoria di dataset ci sono 2 principali strade illustrate di seguito.

Generatore con architettura a solo decoder

Questo approccio prevede un'architettura a solo decoder, ovvero un modello che prende in ingresso un tensore di determinate dimensioni e che attraverso una serie di operazioni di upsampling o dilated convolution ad esempio, effettua un'espansione di tale tensore portandolo alle dimensioni finali. Generalmente si mette in ingresso un vettore casuale di dimensione definita, ottenendo in uscita un tensore delle dimensioni di un'immagine con i canali rgb ed eventualmente altri n canali per la maschere che identificano le classi desiderate. Un'esempio di tale architettura è illustrata di seguito.

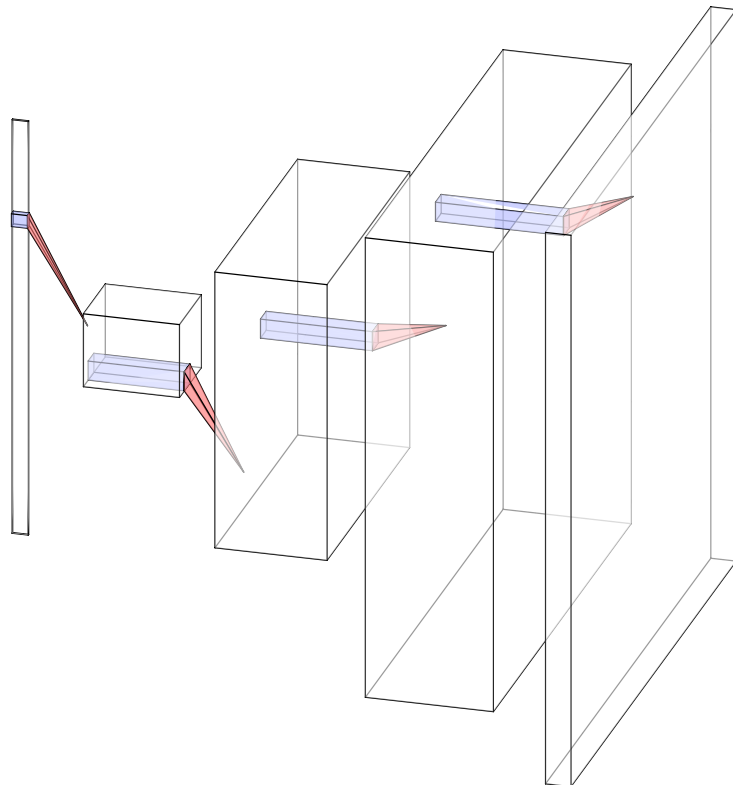


Figura 1.2: Esempio di architettura a solo decoder.

Questo approccio risulta più semplice da implementare, lasciando però al modello il compito di imparare a generare correttamente le immagini e delle maschere coerenti, compito non facile, che a seconda del task può necessitare di un elevato numero di esempi. Questa architettura dovrà imparare oltre alla struttura degli oggetti target, a posizionarli nell'immagine e a generare lo sfondo. Un'altra problematica di questo approccio è il controllo, infatti l'unico modo di interagire con tale modello è modificando il valore del vettore z dato in input, il quale permette di spostarsi nello spazio latente, al quale il modello associa diverse caratteristiche dell'immagine di output in maniera altamente non lineare, rendendo un eventuale controllo dell'output del modello molto difficile. La difficoltà di controllare il modello rende dunque difficoltoso o impossibile controllare, qualora fosse necessario, la posizione, l'intensità, la dimensione o la forma degli oggetti generati.

Generatore con architettura a encoder-decoder

Quest'ultimo è l'approccio scelto in questo progetto, in quanto permette di avere un maggiore controllo sull'output del modello, anche se prevede una training pipeline più complessa da gestire. Al contrario del caso precedente il modello con struttura encoder-decoder permette di passare in ingresso un'immagine base e una o più maschere che identificano le aree dove determinati oggetti devono essere generati, trasformando il task di generazione puro in un task di inpainting. I vantaggi principali di questa tecnica stanno nel fatto che il modello non deve più apprendere la distribuzione degli oggetti nello spazio dell'immagine, nè deve apprendere in maniera troppo approfondita i background, ma si può focalizzare maggiormente sulla struttura degli oggetti da generare.

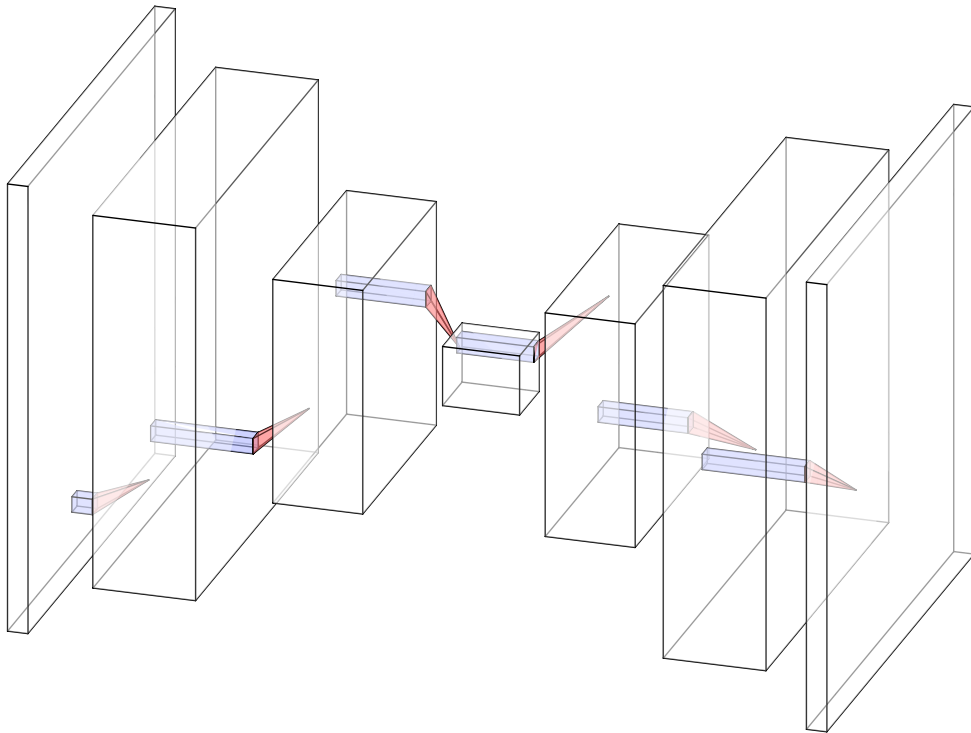


Figura 1.3: Esempio di architettura con encoder e decoder.

1.2 La nascita del deep learning

In questa sezione si darà un'idea generale di cos'è il deep learning e di come questo campo di ricerca sia nato dallo studio della neuroscienza e del machine learning.

1.2.1 Dal machine learning al deep learning

L'intelligenza artificiale è un campo di ricerca con l'obiettivo di risolvere una grande varietà di problemi, che per essere risolti attraverso la programmazione classica avrebbero bisogno di una grande quantità di conoscenze non disponibili, o semplicemente di troppo lavoro.

I programmi basati sul paradigma dell'intelligenza artificiale si propongono di superare questi ostacoli acquisendo direttamente queste conoscenze dai dati grezzi, tale capacità è nota come machine learning. Sotto questa grande famiglia di algoritmi si trovano altri sottogruppi quali il representation learning e all'interno di quest'ultimo il deep learning.

Il deep learning rispetto ai metodi più classici, tipicamente in grado di riconoscere soltanto relazioni lineari (come ad esempio l'SVM o support vector machine), si propone come alternativa per l'apprendimento di funzioni non lineari anche molto complesse. Il termine "deep learning" deriva proprio dalla capacità di riuscire a cogliere queste relazioni molto "profonde" tra i dati di ingresso e uscita.

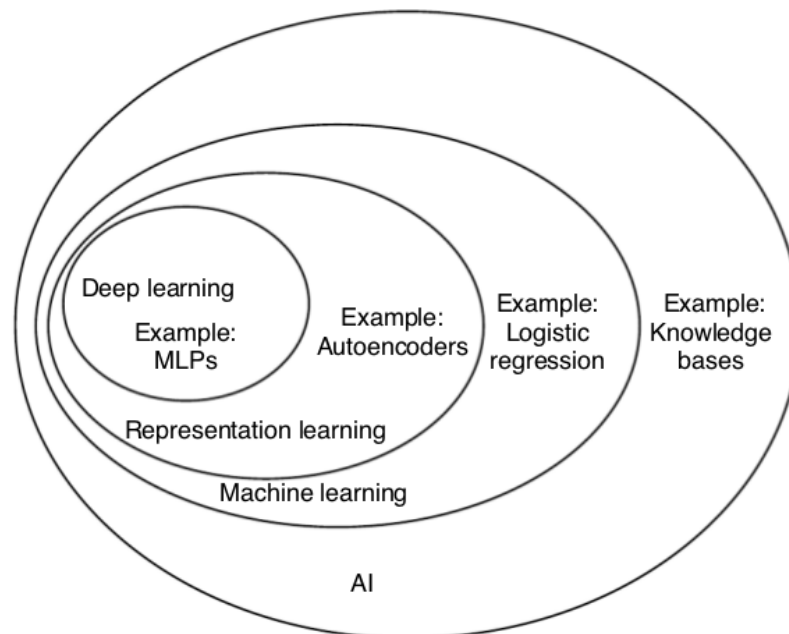


Figura 1.4: Un diagramma di ven che illustra le relazioni tra i diversi sottogruppi dell'intelligenza artificiale, vediamo infatti come il deep learning sia un sottogruppo del representation learning, che a sua volta è un sottogruppo del machine learning.
credits: Yoshua Bengio, Ian J. Goodfellow, Aaron Courville 2015, From the book "Deep Learning"

1.2.2 Le reti neurali biologiche

Le reti neurali come concetto matematico fecero la loro prima comparsa in un articolo del 1957, pubblicato da Warren McCulloch e Walter Pitts "A logical calculus of the ideas immanent in nervous activity", articolo che gettò le basi per la costruzione di reti neurali artificiali come le conosciamo oggi partendo proprio dal sistema nervoso. Infatti le reti neurali artificiali sono ispirate alle reti neurali biologiche, le quali hanno un comportamento più complesso della controparte artificiale, in quanto le reti neurali artificiali devono fare i conti con la complessità computazionale che deve essere ridotta per garantire una elaborazione efficiente nei calcolatori. Il neurone biologico componente principale del cervello e del sistema nervoso, è costituito da un corpo cellulare o soma, dai dendriti, dall'assone e dalle sinapsi.

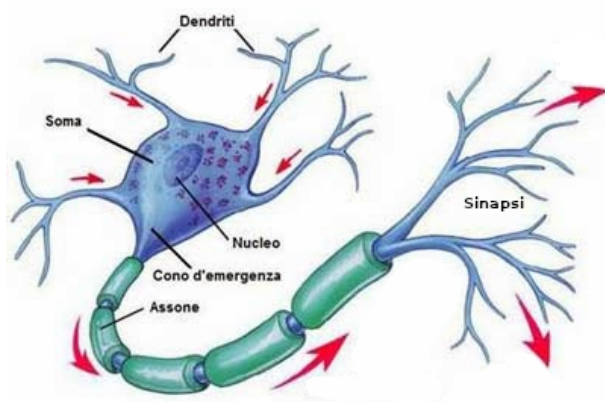


Figura 1.5: Schema di un neurone biologico.

Il neurone riceve degli impulsi da altri neuroni o da altri apparati sensoriali attraverso i dendriti, che sono delle strutture ramificate che si estendono dalla cellula, questi impulsi vengono accumulati all'interno del soma, e se la somma supera un certo valore di soglia si innesca la propagazione di un'impulso, che viene trasmesso attraverso l'assone verso altri neuroni o verso altri organi. L'assone è una struttura che si estende dalla cellula, a seconda della tipologia di neurone può estendersi da pochi micrometri fino anche ad un metro, e presenta all'estremità opposta del soma le sinapsi, le quali consentono la propagazione dell'impulso dall'assone ad altri neuroni. La struttura dell'assone è rivestita dalla guaina mielinica che ne facilita la conduzione degli impulsi, maggiore è lo spessore della guaina minore è la resistenza al passaggio dell'impulso, e dunque maggiore sarà l'ampiezza del segnale in uscita a parità di quello di ingresso. Tale meccanismo è utilizzato per accumulare informazione nella struttura della rete neurale biologica.

1.3 Le reti neurali feed forward

Uno dei primi modelli ad essere proposti e utilizzati nella pratica è stato quello delle reti neurali feedforward (o multi layer perceptron MLP), in cui i neuroni sono disposti in strati, e l'output di ogni neurone di uno strato è connesso con l'input di tutti i neuroni dello strato successivo, attraverso delle connessioni che conservano un peso, la configurazione di tali pesi determina il comportamento della rete. Tali reti come dice il nome propagano l'informazione dallo strato di

input a quello di output attraverso i layer intermedi, in modo lineare, senza retropropagazioni intermedie dell'informazione.

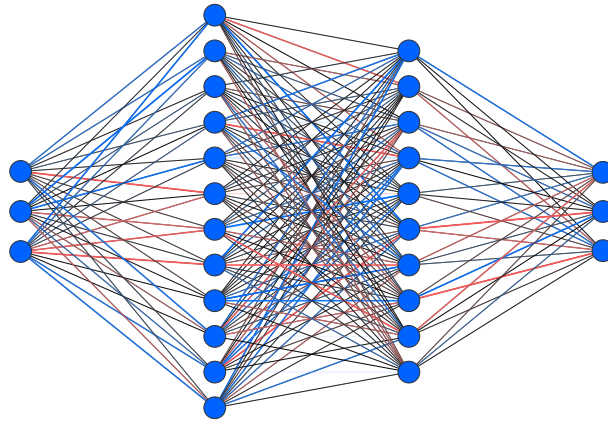


Figura 1.6: Esempio di rete neurale feedforward. In tale rete è possibile vedere i neuroni rappresentati dai nodi del grafo, e le interconnessioni tra di essi che definiscono il peso di ogni relazione, con in rosso un peso positivo e in blu un peso negativo, l'intensità del colore indica la forza della relazione.

1.3.1 Il neurone artificiale

Il neurone artificiale emula il comportamento del neurone biologico, semplificandone notevolmente la complessità, una delle più importanti semplificazioni è che il neurone artificiale opera in un regime temporale discreto e non continuo come la controparte. Il neurone artificiale inoltre non utilizza un meccanismo di accumulazione e spike, ma restituisce un output per ogni input ricevuto, ciò che varia è l'intensità di questo output, che dipende dall'intensità degli input ricevuti, e dai pesi delle connessioni con tali input. Il neurone artificiale inoltre è provvisto di una funzione di trasferimento che mappa la somma pesata degli input ricevuti con l'uscita. Vediamo dunque l'espressione che caratterizza il comportamento di un neurone artificiale:

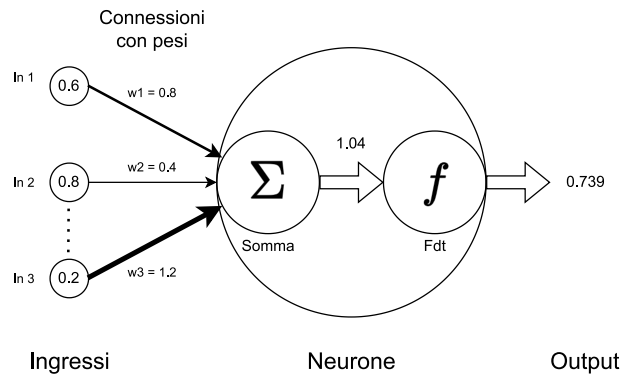
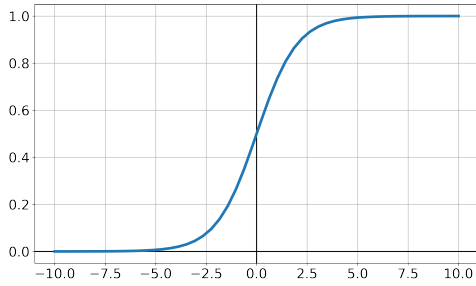


Figura 1.7: Esempio di Neurone artificiale.

$$y = f(P) = f(\vec{w} \cdot \vec{x} + b) = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (1.1)$$

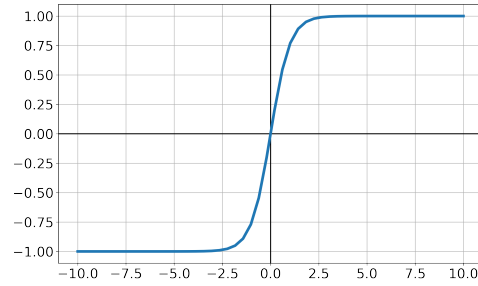
Considerando y l'output del neurone, \vec{w} il vettore dei pesi delle connessioni in ingresso, \vec{x} il vettore degli input, b il bias ovvero un valore aggiunto alla somma pesata degli input dipendente dal neurone e f la funzione di trasferimento. La funzione di attivazione o funzione di trasferimento del neurone è la componente che conferisce alla rete la capacità di generare degli output che hanno una relazione non lineare rispetto agli input ricevuti, e dunque che gli permette di apprendere funzioni non lineari. La funzione di attivazione solitamente deve essere derivabile, o almeno derivabile a tratti per poter essere utilizzata in un contesto di apprendimento, in

quanto la derivata della funzione di attivazione viene utilizzata per calcolare il gradiente della funzione di errore dall'algoritmo Backpropagation, e in seguito per aggiornare i pesi da parte della funzione di ottimizzazione (es. SGD). Di seguito sono mostrate alcune delle più comuni funzioni di attivazione:



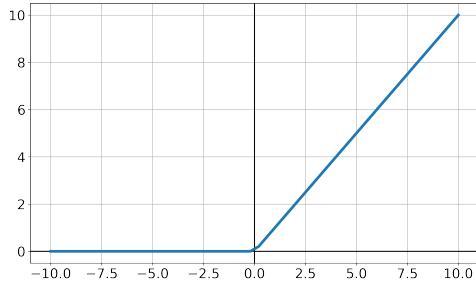
(a) Sigmoide

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



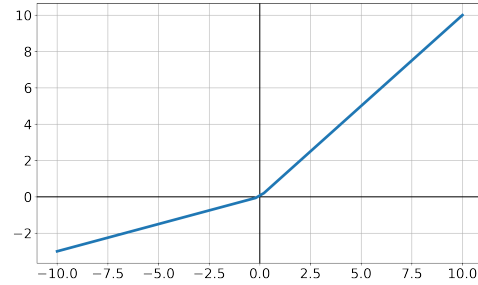
(b) Tangente iperbolica

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



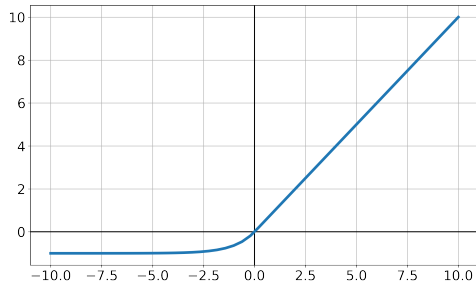
(e) Rectified linear unit (ReLU)

$$f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$$



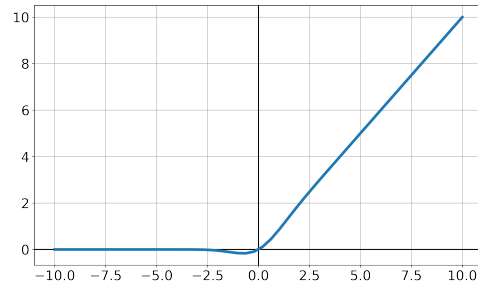
(f) Leaky ReLU

$$f(x) = \begin{cases} \alpha x & x < 0 \\ x & x \geq 0 \end{cases}$$



(i) Exponential linear unit (ELU)

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(e^x - 1) & \text{if } x < 0 \end{cases}$$



(j) Gaussian error linear unit (GELU)

$$f(x) = x * \Phi(x)$$

La sigmoide e la tangente iperbolica, sono funzioni molto vecchie utilizzate sin dai primi anni '90, ma sono state sostituite da funzioni più moderne e semplici da calcolare, come le funzioni ReLU e Leaky ReLU. Vengono ancora utilizzate però in casi particolari come nello stadio finale di una rete neurale, quando si ha bisogno di un output che sia compreso tra 0 e 1. Le funzioni ELU e GELU sono invece funzioni di attivazione più recenti, che vengono utilizzate per migliorare l'apprendimento delle reti neurali in casi specifici, ma in generale ReLU e Leaky ReLU sono le funzioni più utilizzate, in quanto offrono un buon compromesso tra prestazioni e accuratezza.

1.3.2 Il Back-propagation

La prima volta che questo algoritmo fu proposto fu nel 1986 da David E. Rumelhart, Geoffrey E. Hinton e Ronald J. Williams, su *nature* con l'articolo *Learning representations by back-propagating errors*. Questo algoritmo ha segnato da quel momento una vera e propria rivoluzione nel campo dell'apprendimento automatico, rendendo possibile l'addestramento dei modelli neurali come li conosciamo oggi e plasmando lo scenario attuale dell'intelligenza artificiale.

Ciò che questo algoritmo fa effettivamente è calcolare il gradiente della funzione di errore nello spazio dei pesi di una rete neurale, partendo dal layer di uscita e andando a calcolare il gradiente per ogni layer precedente fino ad arrivare al layer di input, tale gradiente può essere poi utilizzato per aggiornare i pesi della rete attraverso una funzione di ottimizzazione, in modo da minimizzare o massimizzare la funzione di costo che si vuole ottimizzare.

Da un punto di vista matematico, data la precedente definizione di neurone e rete neurale, possiamo definire una rete neurale come una funzione $\mathbf{g}(\mathbf{x})$ come combinazione di composizione di funzioni e moltiplicazioni di matrici.

$$\tilde{\mathbf{y}} = \mathbf{g}(\tilde{\mathbf{x}}) = \mathbf{f}_L(\mathbf{W}^L \cdot \mathbf{f}^{L-1}(\mathbf{W}^{L-1} \cdot \mathbf{f}^{L-2}(\dots \mathbf{W}^3 \cdot \mathbf{f}^2(\mathbf{W}^2 \cdot \mathbf{f}^1(\mathbf{W}^1 \cdot \tilde{\mathbf{x}})) \dots))) \quad (1.2)$$

Dove \mathbf{f}_L è la funzione di attivazione del layer di uscita, \mathbf{f}_i è la funzione di attivazione del layer i e \mathbf{W}^i è la matrice dei pesi del layer i . Abbiamo inoltre che $\tilde{\mathbf{x}}$ e $\tilde{\mathbf{y}}$ sono rispettivamente il vettore di input e il vettore di output della rete neurale, se consideriamo $\hat{\mathbf{y}}$ il vettore di output desiderato, possiamo definire la funzione di costo. In questo caso utilizzeremo la loss MSE (Mean Squared Error), ma si può sostituire con qualsiasi funzione di costo.

$$\mathbf{E} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (1.3)$$

A questo punto data una determinata coppia di input e output $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, possiamo calcolare il gradiente della funzione di costo attraverso la *regola della catena*, che ci permette di calcolare il gradiente della funzione di costo per ogni peso della rete.

$$\frac{\partial \mathbf{E}}{\partial \mathbf{w}_{ij}^l} = \frac{\partial \mathbf{E}}{\partial \mathbf{a}^l} \cdot \frac{\partial \mathbf{a}^l}{\partial \mathbf{z}^l} \cdot \frac{\partial \mathbf{z}^l}{\partial \mathbf{w}_{ij}^l} \quad (1.4)$$

Otteniamo così il gradiente della funzione di costo per il peso w_{ij}^l appartenente al layer l , al neurone i e alla sinapsi j .

Per l'esecuzione dell'algoritmo back propagation è necessario effettuare il caching dei valori intermedi calcolati durante il passaggio in avanti, nello specifico degli input pesati dei neuroni prima della funzione di attivazione $\tilde{\mathbf{z}}^l$ e l'output dei neuroni dopo la funzione di attivazione $\tilde{\mathbf{a}}^l$ per ogni layer.

Consideriamo la derivata della funzione di errore rispetto all'input del modello:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{E}}{\partial \mathbf{a}^L} \right) \circ \left(\frac{\partial \mathbf{a}^L}{\partial \mathbf{z}^L} \cdot \frac{\partial \mathbf{z}^L}{\partial \mathbf{a}^{L-1}} \right) \circ \left(\frac{\partial \mathbf{a}^{L-1}}{\partial \mathbf{z}^{L-1}} \cdot \frac{\partial \mathbf{z}^{L-1}}{\partial \mathbf{a}^{L-2}} \right) \circ \dots \circ \left(\frac{\partial \mathbf{a}^2}{\partial \mathbf{z}^2} \cdot \frac{\partial \mathbf{z}^2}{\partial \mathbf{a}^1} \right) \circ \left(\frac{\partial \mathbf{a}^1}{\partial \mathbf{z}^1} \cdot \frac{\partial \mathbf{z}^1}{\partial \mathbf{x}} \right) \quad (1.5)$$

Dove \circ è il prodotto di Hadamard, un semplice prodotto element-wise tra matrici di dimensioni uguali, che moltiplica elemento per elemento le due matrici. Osservando questa formulazione della derivata possiamo notare che $\frac{\partial \mathbf{a}^L}{\partial \mathbf{z}^L} \cdot \frac{\partial \mathbf{z}^L}{\partial \mathbf{a}^{L-1}}$ è la derivata della funzione di attivazione moltiplicata per la matrice dei pesi del layer stesso. Inoltre possiamo riscrivere la derivata $\frac{\partial \mathbf{E}}{\partial \mathbf{a}^L}$ in termini di gradiente $\nabla_{\mathbf{a}} \mathbf{E}$, invertendo l'ordine dei prodotti e trasponendo le matrici:

$$\nabla_{\mathbf{x}} \mathbf{E} = (\mathbf{W}^1)^T \cdot (\mathbf{f}^1)' \circ \dots \circ (\mathbf{W}^{L-1})^T \cdot (\mathbf{f}^{L-1})' \circ (\mathbf{W}^L)^T \cdot (\mathbf{f}^L)' \circ \nabla_{\mathbf{a}^L} \mathbf{E} \quad (1.6)$$

A questo punto possiamo introdurre i prodotti parziali del gradiente per determinare il gradiente della funzione di errore ad un determinato layer l :

$$\delta^l = (\mathbf{f}^l)' \circ (\mathbf{W}^{l+1})^T \cdot (\mathbf{f}^{l+1})' \circ \dots \circ (\mathbf{W}^L)^T \cdot (\mathbf{f}^L)' \circ \nabla_{\mathbf{a}^L} \mathbf{E} \quad (1.7)$$

E notiamo che ogni prodotto parziale può essere definito come il prodotto tra il gradiente e la matrice trasposta dei pesi del layer successivo per la derivata della funzione di attivazione del layer stesso.

$$\delta^{l-1} = (\mathbf{f}^{l-1})' \circ (\mathbf{W}^l)^T \cdot \delta^l \quad (1.8)$$

Quindi:

$$\delta^L = (\mathbf{f}^L)' \circ \nabla_{\mathbf{a}^L} \mathbf{E}$$

$$\delta^{L-1} = (\mathbf{f}^{L-1})' \circ (\mathbf{W}^L)^T \cdot \delta^L = (\mathbf{f}^{L-1})' \circ (\mathbf{W}^L)^T \cdot (\mathbf{f}^L)' \circ \nabla_{\mathbf{a}^L} \mathbf{E}$$

...

$$\delta^2 = (\mathbf{f}^2)' \circ (\mathbf{W}^3)^T \cdot \delta^3 = (\mathbf{f}^2)' \circ (\mathbf{W}^3)^T \dots \circ (\mathbf{W}^L)^T \cdot (\mathbf{f}^L)' \circ \nabla_{\mathbf{a}^L} \mathbf{E}$$

$$\delta^1 = (\mathbf{f}^1)' \circ (\mathbf{W}^2)^T \cdot \delta^2 = (\mathbf{f}^1)' \circ (\mathbf{W}^2)^T \cdot (\mathbf{f}^2)' \circ (\mathbf{W}^3)^T \dots \circ (\mathbf{W}^L)^T \cdot (\mathbf{f}^L)' \circ \nabla_{\mathbf{a}^L} \mathbf{E}$$

In tal modo possiamo calcolare i prodotti parziali del gradiente per ogni layer, partendo dallo strato di output all'indietro, minimizzando così il numero di operazioni richieste per il calcolo del gradiente. Per ottenere i gradienti dei pesi è sufficiente moltiplicare il prodotto parziale del gradiente per l'output del layer precedente:

$$\nabla_{\mathbf{w}^l} \mathbf{E} = \delta^l \cdot (\mathbf{a}^{l-1})^T \quad (1.9)$$

L'oggetto $\nabla_{\mathbf{w}^l} \mathbf{E}$ rappresenta una matrice della stessa dimensione della matrice dei pesi del layer l , contenente i gradienti della funzione di errore di tali pesi, mentre \mathbf{a}^{l-1} è l'output di tutte le unità del layer precedente.

Si consideri che è possibile scomporre un elemento della matrice $\nabla_{\mathbf{w}^1} \mathbf{E}$ nel seguente modo:

$$\nabla_{\mathbf{w}^1} \mathbf{E}_{k,k} = \frac{\partial \mathbf{E}}{\partial \mathbf{w}_{k,k}^1} \quad (1.10)$$

Tale gradiente a questo punto può essere utilizzato per aggiornare il peso $w_{k,k}^l$, applicando un semplice coefficiente di apprendimento η :

$$\mathbf{W}_{k,k}^1 = \mathbf{W}_{k,k}^1 + \Delta \mathbf{W}_{k,k}^1 \quad (1.11)$$

$$\Delta \mathbf{W}_{k,k}^1 = -\eta \cdot \nabla_{\mathbf{w}^1} \mathbf{E}_{k,k} \quad (1.12)$$

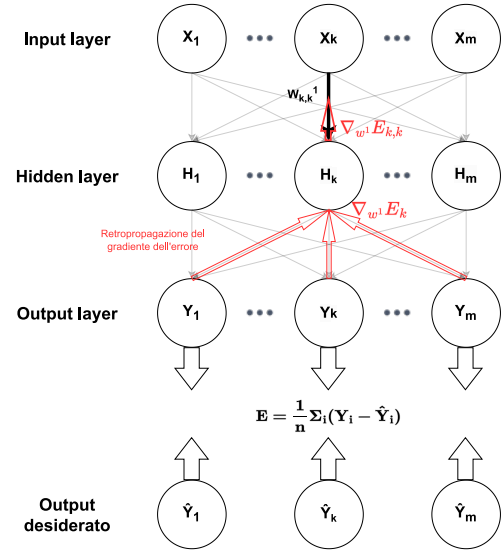


Figura 1.9: Esempio di retropropagazione su di una rete semplificata a 2 strati.

é importante notare che questo è un'approccio

molto basilare, che può essere migliorato attraverso l'uso di tecniche di ottimizzazione più evolute come ad esempio l'ottimizzazione SGD (Stochastic Gradient Descent) o l'ottimizzazione ADAM.

1.3.3 Teorema di approssimazione universale

Le reti neurali artificiali, non hanno dimostrato solo nella pratica con risultati sperimentali la loro efficacia nel risolvere i problemi, ma anche nella teoria, infatti in molti studi è stata provata la loro efficacia come approssimatori universali. Quando si parla di approssimatori universali ci si riferisce a delle funzioni che possono approssimare su un dato intervallo di valori qualunque altra funzione continua. Nel caso particolare delle reti neurali vi è una grande quantità di varianti, per le quali si ha una diversa dimostrazione per tale proprietà.

Tra i più importanti nel 1989 George Cybenko in *Approximation by superpositions of a sigmoidal function* [2], dimostrò che la sovrapposizione di un numero finito di istanze di una singola funzione univariata può approssimare qualsiasi funzione continua di n variabili con supporto nell'ipercubo unitario. Questo risultato permette di asserire che ogni funzione continua può essere approssimata da una rete neurale MLP (feedforward) avente uno strato nascosto con un numero finito di unità, con come funzione di attivazione una funzione sigmoide arbitraria. Di seguito il relativo teorema:

Teorema 1 *Sia sigma qualsiasi funzione sigmoide arbitraria. Le sommatorie finite della forma*

$$\mathbf{G}(\mathbf{x}) = \sum_{i=1}^N \alpha_i \cdot \sigma(\mathbf{y}_i^T \mathbf{x} + \Theta_i) \quad (1.13)$$

sono dense in $C(I_n)$. In altre parole, data qualunque $f \in C(I_n)$ e $\epsilon > 0$ esiste una somma $G(x)$ con la forma sopra descritta tale che:

$$|\mathbf{G}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| < \epsilon \quad \forall \mathbf{x} \in \mathbf{I}_n \quad (1.14)$$

Con questo teorema Cybenko dimostra che una rete neurale MLP con un solo strato nascosto può approssimare qualsiasi funzione continua ma non è in grado di stabilire quante unità deve avere, in base alla complessità della funzione da approssimare il numero potrebbe essere astronomico. Altre ricerche negli anni seguenti hanno studiato molteplici varianti di questo teorema, ad esempio per neuroni con funzione di attivazione RELU o il caso con profondità del modello arbitraria.

1.4 Le reti neurali convoluzionali

Le reti neurali convoluzionali sono modelli concepiti per analizzare dati che hanno una struttura spaziale, e nei quali le relazioni tra i dati hanno relazioni locali molto forti, come nel caso delle immagini. La loro struttura è simile a quella che si trovano nel nervo ottico degli organismi viventi. Uno dei maggiori vantaggi che ha portato l'avvento delle CNN è stato quello di ridurre drasticamente il numero di parametri da apprendere, i quali con le comuni MLP crescevano esponenzialmente all'aumentare della risoluzione dell'immagine, rendendo di fatto il problema intrattabile. Sono dette anche *Shift invariant artificial neural networks* (SIANN) in quanto sono resistenti alla traslazione dell'input.

1.4.1 Storia delle CNN

La prima pubblicazione alla quale si deve l'invenzione delle CNN è stato un lavoro di David Hunter Hubel e Torsten Wiesel "*Receptive fields of single neurones in the cat's striate cortex*", i quali nel 1959 hanno studiato la struttura della corteccia visiva del cervello di un gatto, scoprendo che tali neuroni avevano una struttura particolare, che li rendeva soggetti agli stimoli di una certa porzione di retina, detta *campo ricettivo*, tale area era regolare per tutti i neuroni e tutti avevano una certa sovrapposizione dei rispettivi campi ricettivi.

Il primo lavoro a sfruttare i concetti appresi da Hubel e Wiesel è stato quello di Kunihiko Fukushima il quale nel 1980 ha pubblicato il suo lavoro *Neocognitron* [5], un sistema di riconoscimento di immagini, utilizzato per la prima volta per il riconoscimento di numeri scritti a mano in giapponese. Questa implementazione di modello neurale è molto vicino al modello naturale, utilizza infatti due tipologie di neuroni:

- **S-cells:** sono i neuroni che si occupano di estrarre le features locali, come ad esempio linee o bordi, presentando le caratteristiche delle cellule della corteccia visiva primaria, come il campo ricettivo limitato ad un'area ristretta.
- **C-cells:** sono i neuroni che si occupano di estrarre le features globali, infatti questi ricevono in input le uscite dei neuroni S-cells, e dunque lavorano con pattern più complessi, come ad esempio semi-cerchi o quadrati ecc. Anche in questo caso c'è una somiglianza con il modello biologico in quanto questi neuroni hanno un campo ricettivo molto più ampio, e costituiscono la corteccia visiva secondaria.

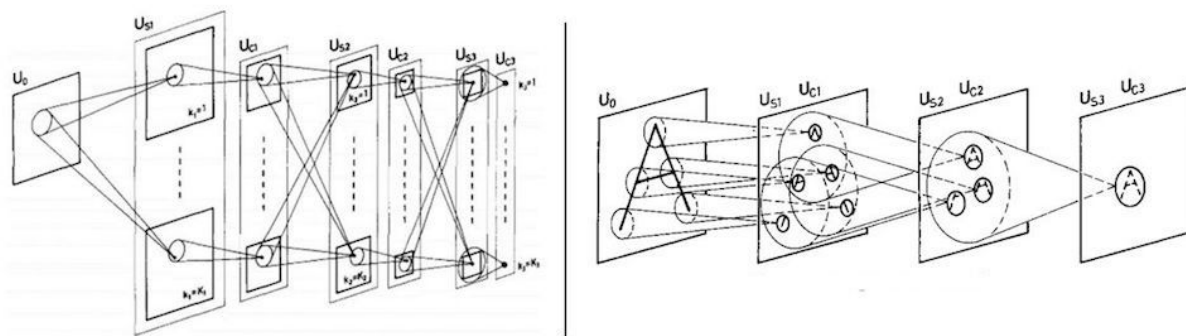


Figura 1.10: L'immagine raffigura schematicamente il funzionamento di Neocognitron.
credits: Kunihiko Fukushima 1980 [5].

Date queste caratteristiche, possiamo notare come il modello neurale di Fukushima sia molto simile a quello che si trova nel cervello ma ancora presenta delle problematiche, come ad esempio la mancanza di un algoritmo di addestramento end-to-end come il backpropagation e l'invarianza alla traslazione, in quanto ogni S-cell ha un campo ricettivo molto limitato e non condivide i suoi pattern con le altre, dunque un modello addestrato con questa struttura potrebbe non essere in grado di riconoscere un oggetto in ogni punto dell'immagine con la stessa precisione. Sempre a Fukushima si deve anche l'introduzione della funzione di trasferimento *RELU*, precedentemente discussa.

Seguì poi il lavoro di Alex Waibel "*Phoneme Recognition Using Time-Delay Neural Networks*" [18] nel 1989, che segnò un'ulteriore passo nella direzione delle CNN, introducendo per la prima volta l'invarianza alla traslazione, ma si parlava ancora di una convoluzione 1D. Questo lavoro però non era focalizzato sulle immagini ma sull'analisi del suono, in particolare sul riconoscimento di fonemi. Ma l'utilizzo di questo modello su 2 dimensioni, ovvero il tempo e la frequenza, lo rendeva concettualmente adatto anche all'analisi delle immagini. Un'altro fatto interessante è che in questo lavoro viene introdotto il concetto di *pooling*, introdotto proprio per ridurre la dimensione dei tensori durante la propagazione nel modello, e ridurre così il numero di parametri da apprendere, soluzione oggi molto utilizzata nelle CNN moderne, in diverse varianti che poi verranno discusse. Nell'immagine che segue, tratta proprio dall'articolo citato si può vedere come questo modello effettuava la convoluzione 1D nel tempo ed il pooling nel tempo e nella frequenza, tra i vari layer.

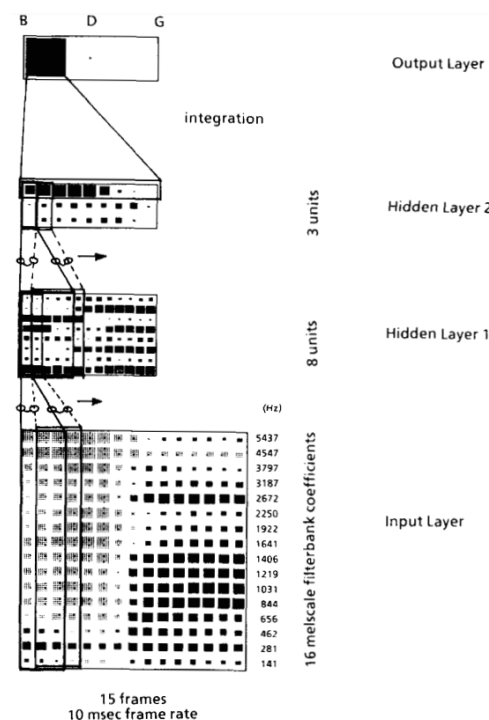


Figura 1.11: L'immagine tratta dallo stesso articolo illustra schematicamente l'architettura del modello TDNN.

credits: Alex Waibel et al. 1989 [18].

In fine una delle prime ricerche che si sono avvicinate di più all'implementazione moderna

delle CNN è stata quella di Yann LeCun et al. *"Backpropagation applied to handwritten zip code recognition"* [12] pubblicato nel 1989, in cui è stato utilizzato il backpropagation per l'addestramento di una rete convoluzionale per il riconoscimento di numeri scritti a mano.

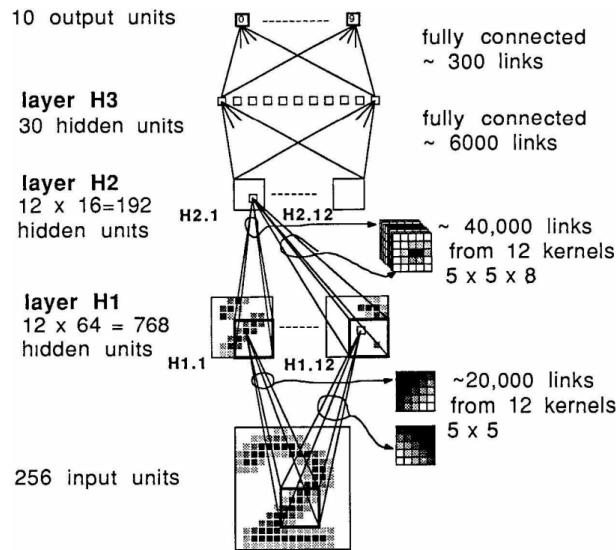


Figura 1.12: L'immagine tratta dal medesimo articolo illustra l'architettura proposta.
credits: Yann LeCun et al. 1989 [12].

La rete proposta da LeCun in questa pubblicazione aveva un'architettura molto semplice, basta su 2 strati convoluzionali 2D, e 2 strati MLP, l'utilizzo del backpropagation per l'addestramento dei kernel di questi due strati convoluzionali, per quanto oggi sembri banale era un'innovazione di grande impatto, in quanto fino a quel momento i kernel dei filtri convoluzionali venivano definiti manualmente. Un'altra caratteristica interessante fu quella dell'utilizzo della tecnica del "weights sharing", la caratteristica che consente alle CNN di ridurre drasticamente il numero di parametri da apprendere, rispetto alle MLP o a precedenti implementazioni come Neocognitron.

1.4.2 La convoluzione

In questa sezione andremo a vedere come funziona la convoluzione, e come viene implementata nelle CNN moderne. partendo dalla definizione formale di convoluzione, che è la seguente:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \quad (1.15)$$

dove f e g sono due funzioni, e $*$ è l'operatore di convoluzione. Questa definizione è valida per funzioni definite su un dominio continuo, ma spesso nella pratica si utilizza la convoluzione su un dominio discreto, come ad esempio le immagini, le quali sono definite da un numero finito di pixel. In questo caso la convoluzione si può definire come segue:

$$(f * g)(t) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(t - \tau) \quad (1.16)$$

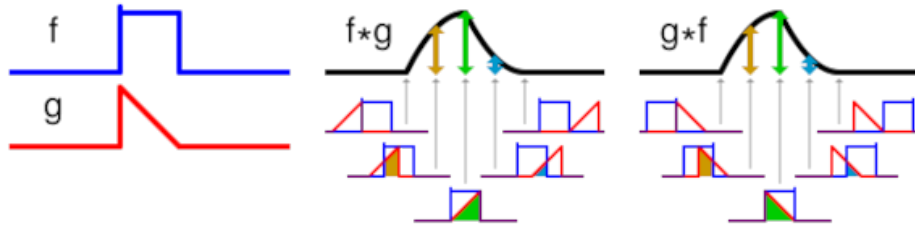


Figura 1.13: L'immagine illustra l'operazione di convoluzione.
credits: Wikipedia. <https://en.wikipedia.org/wiki/Convolution>

Per fare un esempio pratico la convoluzione equivale ad effettuare un prodotto tra due funzioni, ma con un'operazione di shift, che è il motivo per cui la convoluzione è anche detta operazione di correlazione. In generale si avrà come risultato una funzione che è più alta dove le due funzioni in ingresso sono più simili, e più bassa dove queste sono più diverse, o ancora negativa dove queste sono opposte. Come è possibile vedere nell'immagine 1.13, dove l'onda a dente di sega ha una sovrapposizione maggiore con l'onda quadra la loro convoluzione avrà un valore maggiore, inoltre è possibile notare come tale operazione sia commutativa.

La convoluzione discreta nelle immagini, rispetto al caso continuo tra funzioni, si avvale dell'uso di kernel, ovvero di matrici di dimensioni finite, che vengono spostate lungo l'immagine, moltiplicandole e sommando i valori in ogni posizione, in modo da produrre un'immagine di output che è la convoluzione dell'immagine di input con il kernel.

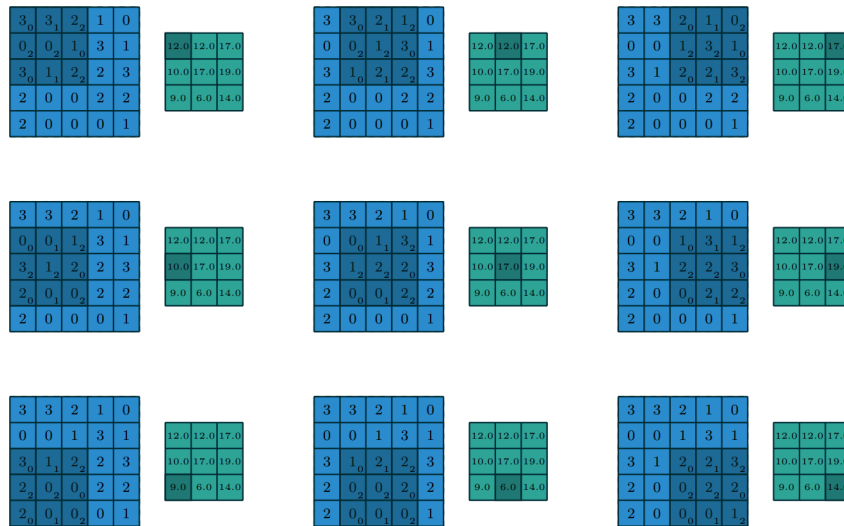


Figura 1.14: Operazione di convoluzione 2D discreta, in blu abbiamo una matrice 5x5 che rappresenta un'immagine, mentre in verde il risultato di una convoluzione, con un kernel 3x3. I valori del kernel sono raffigurati in basso a destra delle caselle interessate dalla convoluzione.
credits: Dumoulin et al. 2016 [3].

Vediamo un'esempio pratico, consideriamo due kernel 3x3, uno con un pattern verticale, e l'altro con un pattern orizzontale e un'immagine di input 5x5:

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 3 & 4 & 3 & 2 \\ 3 & 4 & 5 & 4 & 3 \\ 2 & 3 & 4 & 3 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix} \quad (1.17)$$

Vediamo la convoluzione di questi due kernel con la matrice di esempio 5x5:

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 3 & 4 & 3 & 2 \\ 3 & 4 & 5 & 4 & 3 \\ 2 & 3 & 4 & 3 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} -6.0 & 0.0 & 6.0 \\ -6.0 & 0.0 & 6.0 \\ -6.0 & 0.0 & 6.0 \end{bmatrix} \quad (1.18)$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 3 & 4 & 3 & 2 \\ 3 & 4 & 5 & 4 & 3 \\ 2 & 3 & 4 & 3 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} -6.0 & -6.0 & -6.0 \\ 0.0 & 0.0 & 0.0 \\ 6.0 & 6.0 & 6.0 \end{bmatrix} \quad (1.19)$$

In questo modo si ottiene un'immagine di output che ha valori più alti dove i pattern presenti nel kernel sono simili mentre si hanno valori negativi dove il pattern è opposto, nello specifico, possiamo vedere come nella convoluzione 1.18, ha un andamento decrescente verso destra, e infatti la convoluzione con l'immagine restituisce un valore elevato dove l'immagine ha un andamento decrescente verso destra. Vediamo nel dettaglio però quali operazioni vengono effettuate per ottenere l'immagine di output, considerando sempre la convoluzione 1.18, vediamo una lista di operazioni, una per ogni valore della matrice di output, che corrispondono alle moltiplicazioni e somme che vengono effettuate tra il kernel e l'immagine, in ogni posizione in cui il kernel si sovrappone all'immagine:

- $1 * 1 + 0 * 2 + -1 * 3 + 1 * 2 + 0 * 3 + -1 * 4 + 1 * 3 + 0 * 4 + -1 * 5 = -6.0$
- $1 * 2 + 0 * 3 + -1 * 4 + 1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 = 0.0$
- $1 * 1 + 0 * 2 + -1 * 3 + 1 * 2 + 0 * 3 + -1 * 4 + 1 * 3 + 0 * 4 + -1 * 5 = 6.0$
- $1 * 2 + 0 * 3 + -1 * 4 + 1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 = -6.0$
- $1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 + 1 * 1 + 0 * 2 + -1 * 3 = 0.0$
- $1 * 2 + 0 * 3 + -1 * 4 + 1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 = 6.0$
- $1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 + 1 * 1 + 0 * 2 + -1 * 3 = -6.0$
- $1 * 2 + 0 * 3 + -1 * 4 + 1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 = 0.0$
- $1 * 3 + 0 * 4 + -1 * 5 + 1 * 2 + 0 * 3 + -1 * 4 + 1 * 1 + 0 * 2 + -1 * 3 = 6.0$

Di seguito è riportata l'immagine 1.15 di Lena Forsén, un'attrice svedese la cui foto è usata frequentemente per esempi e test di algoritmi di computer vision, a cui sono stati applicati i due kernel visti prima, alla prima immagine è stato applicato il kernel 1.18 mentre alla seconda è stato applicato il kernel 1.19.

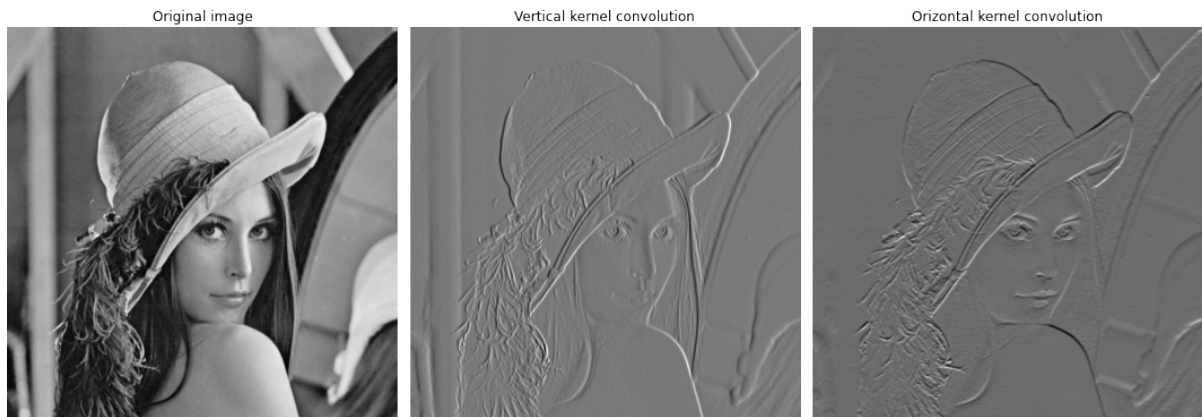


Figura 1.15: Immagine di Lena Forsén, prima e dopo la convoluzione con i due kernel, che amplificano i bordi verticali e orizzontali.

1.4.3 I parametri della convoluzione

In questa sezione andremo a vedere quali sono i principali parametri della convoluzione, e come questi la influenzano, vedremo in particolare: padding, stride, kernel size e numero di filtri.

Padding

Il padding è un parametro che viene utilizzato per aumentare la dimensione dell'immagine di ingresso prima di effettuare la convoluzione, e ottenere così una dimensione di output maggiore, questo è utile quando si vuole mantenere la dimensione dell'immagine di output uguale a quella di input, tale risultato si ottiene aggiungendo delle righe e colonne di valori intorno alla matrice di input originale. Ci sono due tipologie principali di padding:

- **Zero padding:** il padding viene aggiunto con dei valori pari a zero, questo è utile quando si vuole mantenere la dimensione dell'immagine di output uguale a quella di input.
- **Reflection padding:** il padding viene aggiunto con dei valori uguali ai valori dell'immagine di input, riflessi rispetto ai bordi.

Stride

Lo stride è un parametro che viene utilizzato per decidere di quanto si sposta il kernel durante la convoluzione, ad esempio con uno stride pari a 2, il kernel si sposterà di 2 pixel in orizzontale e/o in verticale, questo è utile quando si vuole ridurre la dimensione dell'immagine di output, con il giusto stride si può effettuare la convoluzione e ottenere ad esempio un'immagine di output di dimensione pari a $\frac{1}{2}$ dell'immagine di input.

Kernel size

La kernel size o dimensione del kernel è un parametro che viene utilizzato per decidere la dimensione del kernel con cui effettuare la convoluzione. tipicamente i kernel sono quadrati, ma è possibile utilizzare anche kernel rettangolari, ad esempio un kernel di dimensione 3x5 o 1x7. Per riprendere il collegamento con la controparte biologica delle reti neurali, il kernel è l'equivalente del campo visivo del neurone, aumentando la dimensione del kernel aumentiamo il campo visivo del neurone, e quindi aumentiamo la capacità di riconoscere pattern più complessi. Uno svantaggio importante di aumentare eccessivamente la dimensione del kernel è che aumenta anche il numero di parametri da apprendere, ma ancor di più aumenta considerevolmente il numero di operazioni da effettuare durante la convoluzione, per tale ragione è consuetudine utilizzare kernel di dimensione 1x1, 3x3 o 5x5, difficilmente si utilizzano kernel di dimensione maggiore.

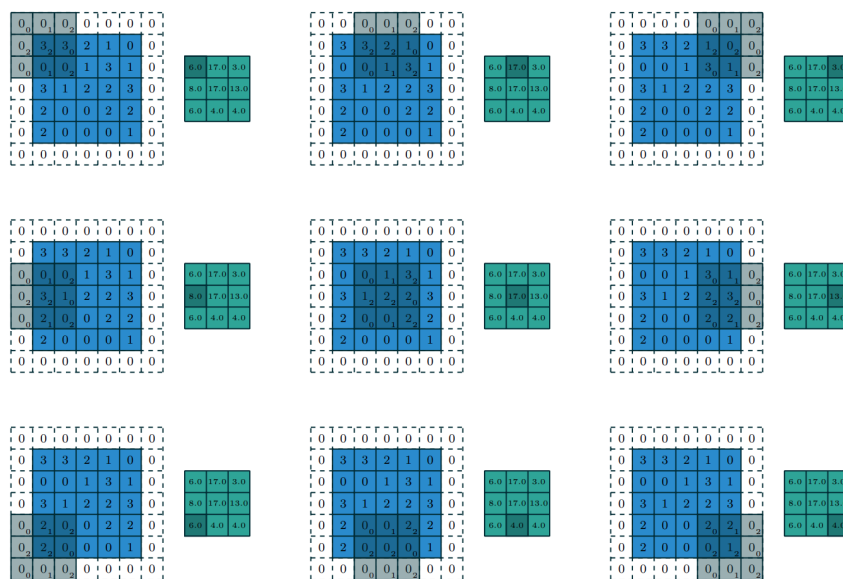


Figura 1.16: Esempio di convoluzione con un kernel 3x3, stride pari a 2, e padding (zero padding) pari a 1.

credits: Dumoulin et al. 2016 [3]

Numero di filtri

Il numero di filtri è un parametro che viene utilizzato per decidere il numero di filtri da utilizzare per effettuare la convoluzione, tipicamente equivale a decidere quanti kernel diversi si vogliono utilizzare per effettuare la convoluzione, ottenendo altrettanti canali nel tensore di output. Ad esempio se abbiamo un tensore 1x5x5 in input e decidiamo di utilizzare 3 filtri 3x3 con stride 1 e padding 0, otteniamo un tensore 3x3x3 in output, dove ogni canale è il prodotto della convoluzione tra il tensore di input e un kernel diverso.

La dimensione di output

Dopo aver visto i principali parametri della convoluzione, vediamo come questi influenzano la dimensione dell'immagine di output. Possiamo calcolare la dimensione dell'immagine di output in base ai seguenti parametri di input:

- W_{in}, H_{in} : dimensione del tensore di input.
- W_k, H_k : dimensione del kernel.
- P : padding.
- S : stride.
- W_{out}, H_{out} : dimensione del tensore di output.

Ipotizzando lo stride e il padding uguali lungo entrambe le direzioni di convoluzione, le dimensioni di output si calcolano come segue:

$$H_{out} = \frac{H_{in} - H_k + 2P}{S} + 1 \quad (1.20)$$

$$W_{out} = \frac{W_{in} - W_k + 2P}{S} + 1 \quad (1.21)$$

1.4.4 Il pooling

Il pooling è un'operazione che viene utilizzata per ridurre la dimensione di un tensore, tipicamente viene utilizzato dopo una o più convoluzioni, in modo da ridurre la quantità di dati da elaborare, rendendo le convoluzioni successive più veloci, e permettendo di lavorare con kernel più piccoli su pattern di più alto livello. Infatti riducendo la dimensione di un tensore diciamo della metà, si potrà lavorare con features più complesse, senza la necessità di utilizzare kernel più grandi dato che saranno le features stesse a venire ridotte in dimensione. Le tipologie principali di pooling sono il max pooling e l'average pooling, entrambe vengono applicate in maniera simile alla convoluzione, con lo scorrimento di una finestra, ma invece di effettuare una combinazione lineare dei valori del tensore con i pesi di un kernel, viene applicata una diversa funzione.

Anche per il pooling vale una regola simile alla convoluzione per il calcolo della dimensione dell'immagine di output, con l'unica differenza che nel pooling solitamente il padding non viene utilizzato, quindi la dimensione dell'immagine di output si calcola come segue:

$$H_{out} = \frac{H_{in} - H_k}{S} + 1 \quad (1.22)$$

$$W_{out} = \frac{W_{in} - W_k}{S} + 1 \quad (1.23)$$

Max pooling

Con il **max pooling** viene applicata la funzione di massimo, ovvero considerando una data finestra di scorrimento, il tensore di output avrà per ogni valore il valore massimo presente nella corrispondente finestra del tensore di input.

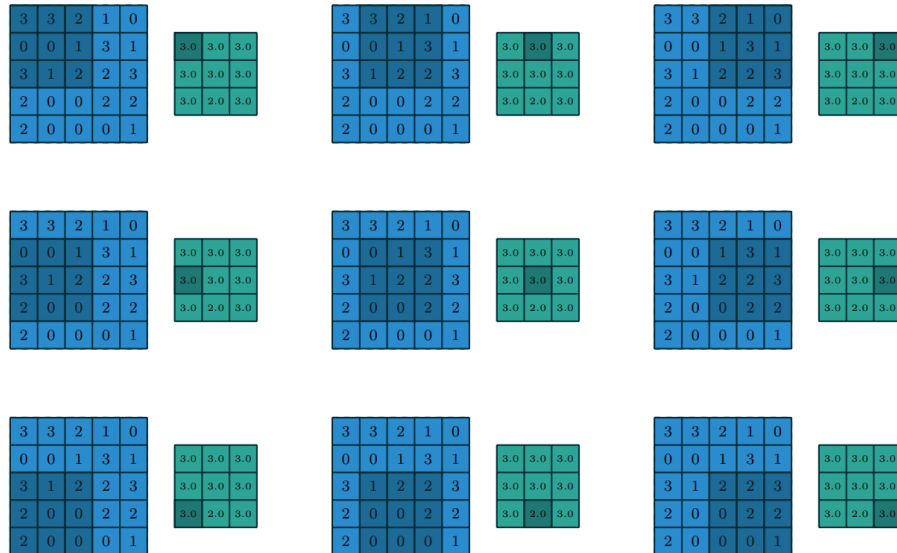


Figura 1.17: Esempio di max pooling con una finestra di dimensione 3x3, stride pari a 1 e padding 0.

credits: Dumoulin et al. 2016 [3]

Average pooling

Con l'**average pooling** viene applicata la funzione di media, ovvero considerando una data finestra di scorrimento, il tensore di output avrà per ogni valore la media dei valori presenti nella corrispondente finestra del tensore di input.

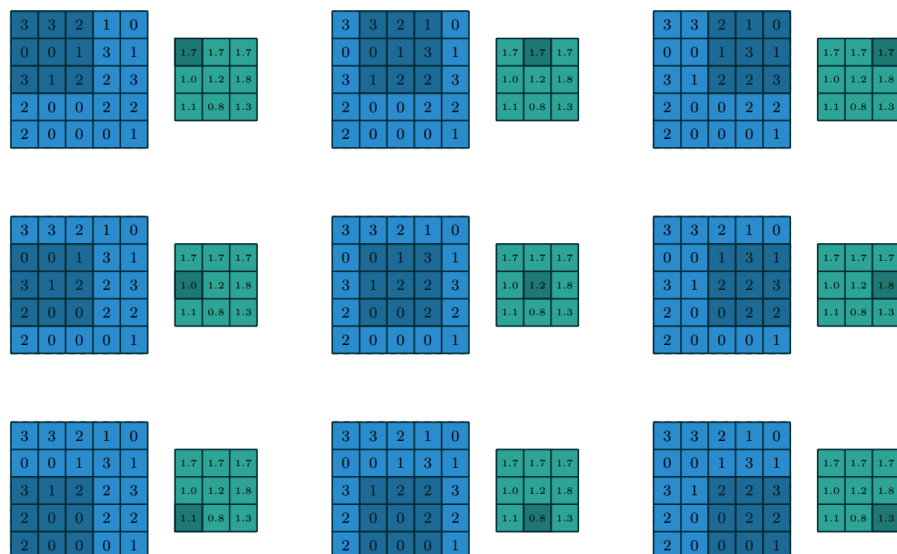


Figura 1.18: Esempio di average pooling con una finestra di dimensione 3x3, stride pari a 1 e padding 0.

credits: Dumoulin et al. 2016 [3]

1.4.5 La convoluzione multichannel

Diversamente da quanto visto nei precedenti esempi, le operazioni di convoluzione utilizzate nelle CNN sono applicate a tensori 3D, dunque con più canali, perciò si parla di convoluzione 2D su tensori con 3 dimensioni ($W \times H \times C$), dove W e H sono altezza e larghezza del tensore mentre C è il numero di canali. Le immagini tipicamente hanno questa struttura e si ha che i diversi canali enfatizzano diversi aspetti dell'immagine, per tale ragione non si effettua la convoluzione della stessa matrice di pesi su tutti i canali, ma per un dato filtro si ha una matrice di pesi per ogni canale, il risultato delle n convoluzioni viene poi sommato channel-wise per ottenere un canale del tensore di output.

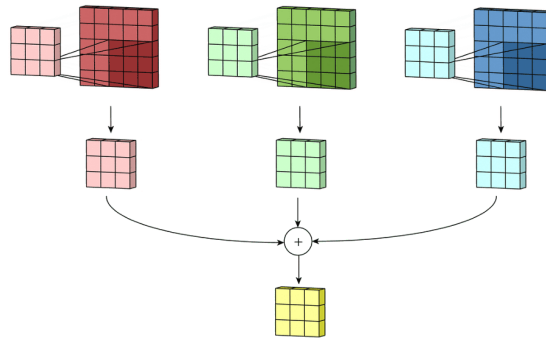


Figura 1.19: Esempio di convoluzione di un'immagine rgb.
credits: Irhum Shafkat [16]

Facciamo però un pò di chiarezza sulla differenza tra kernel e filtro, nella convoluzione 2D, con kernel si intende la matrice di pesi ($W_k \times H_k \times 1$) che viene applicata ad un canale di un tensore, mentre con filtro (Il tensore arancio in 1.20) si intende un tensore composto da C kernel, il quale ha una dimensione $W_k \times H_k \times C$, dove W_k e H_k sono altezza e larghezza del kernel, mentre C è il numero di canali del tensore di input. Per tale ragione in un dato layer convoluzionale che ha in input un tensore con n canali e in output un tensore con m canali si avranno m filtri, composti a loro volta da n kernel, perciò il numero di parametri di tale layer convoluzionale sarà pari a $m \times n \times W_k \times H_k$, dove W_k e H_k sono le dimensioni del kernel.

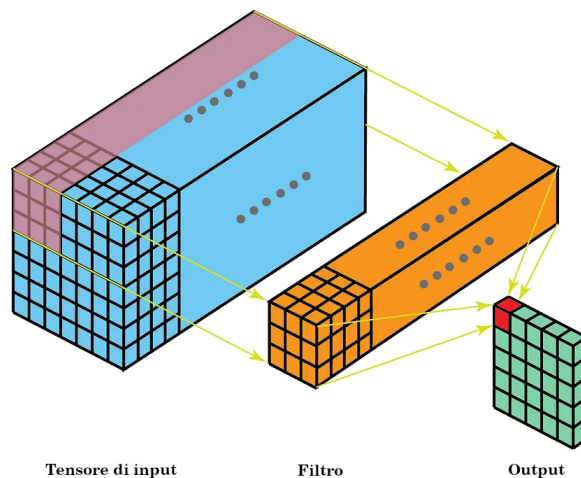


Figura 1.20: Esempio di convoluzione multichannel.
credits: Irhum Shafkat [16]

Capitolo 2

Stato dell'arte

In questa sezione verranno mostrati i principali modelli generativi basati su architettura GAN, che sono stati utilizzati negli ultimi anni, partendo dal primo modello proposto da Goodfellow et al. [8] capostipite di questa famiglia di modelli passando per le principali innovazioni proposte da altri autori, fino ad arrivare a modelli più recenti come StyleGAN2 [11] e LAMA [17].

2.1 Il primo modello gan

Nel 2014 Ian Goodfellow et al. [8] hanno proposto il primo modello generativo basato su architettura GAN (Generative Adversarial Network)., Questa pubblicazione ha segnato un punto di svolta nella ricerca dei modelli di deep learning generativi, i quali fino ad allora erano basati su architetture ad autoencoder. La differenza sostanziale tra i due tipi di modelli è che i modelli basati su autoencoder

Capitolo 3

Strumenti e metodi

3.1 Il Dataset: Severstal steel defect detection

L'acciaio è uno dei materiali più comunemente utilizzati in tutto il mondo, e la sua produzione è in continua crescita. La sua versatilità e la sua resistenza lo rendono un materiale molto utilizzato in diversi settori, come l'edilizia, l'automotive, l'industria elettronica, l'industria aerospaziale, ecc. Per produzioni su larga scala di acciaio come di altri materiali o prodotti, è necessario che il materiale sia di qualità, e che non contenga difetti, ma è difficile per gli operatori umani rilevare difetti come graffi, crepe, ecc. con elevata affidabilità, per tale ragione è necessario adottare sistemi automatizzati che siano in grado di rilevare difetti in modo affidabile e veloce. Severstal è una delle principali aziende produttrici di acciaio in Russia, e produce circa 10 milioni di tonnellate di acciaio all'anno, quest'ultima ha iniziato dunque ad utilizzare il machine learning per automatizzare il processo di rilevazione dei difetti, e ha messo a disposizione questo dataset nel 2019 per permettere a chiunque di partecipare alla challenge, e di testare le proprie idee, sperando di riuscire ad aumentare l'affidabilità del proprio sistema di rilevazione difetti.

Questo dataset è diviso in 2 parti, training e test set, ma per il test set non sono stati rilasciati i ground truth, quindi non è possibile testare il modello finale sul test set originale e verranno dunque utilizzati solo i dati del training set, i quali verranno suddivisi in training e test set. Il training set originale è composto da 12568 immagini, che verranno divise in 50% per il nuovo training set e 50% per il nuovo test set, quindi un totale di 6284 immagini per il training set e 6284 immagini per il test set.

il training set verrà utilizzato per il training del generatore di difetti sintetici, mentre il test set verrà utilizzato per testare il modello di segmentazione finale addestrato con il dataset sintetico.

3.2 Lama:

3.3 Stylegan2:

Capitolo 4

Sviluppo del progetto

Capitolo 5

Risultati

Elenco delle figure

1.1	An example of segmented defects from the "Severstal steel defect dataset". credits: Neven Robby and Goedemé Toon, 2021, A Multi-Branch U-Net for Steel Surface Defect Type and Severity Segmentation. https://www.mdpi.com/2075-4701/11/6/870	2
1.2	Esempio di architettura a solo decoder.	4
1.3	Esempio di architettura con encoder e decoder.	5
1.4	Un diagramma di ven che illustra le relazioni tra i diversi sottogruppi dell'intelligenza artificiale, vediamo infatti come il deep learning sia un sottogruppo del representation learning, che a sua volta è un sottogruppo del machine learning. credits: Yoshua Bengio, Ian J. Goodfellow, Aaron Courville 2015, From the book "Deep Learning"	6
1.5	Schema di un neurone biologico.	7
1.6	Esempio di rete neurale feedforward. In tale rete è possibile vedere i neuroni rappresentati dai nodi del grafo, e le interconnessioni tra di essi che definiscono il peso di ogni relazione, con in rosso un peso positivo e in blu un peso negativo, l'intensità del colore indica la forza della relazione.	8
1.7	Esempio di Neurone artificiale.	8
1.9	Esempio di retropropagazione su di una rete semplificata a 2 strati.	12
1.10	L'immagine raffigura schematicamente il funzionamento di Neocognitron. credits: Kunihiko Fukushima 1980 [5].	14
1.11	L'immagine tratta dallo stesso articolo illustra schematicamente l'architettura del modello TDNN. credits: Alex Waibel et al. 1989 [18].	14
1.12	L'immagine tratta dal medesimo articolo illustra l'architettura proposta. credits: Yann LeCun et al. 1989 [12].	15
1.13	L'immagine illustra l'operazione di convoluzione. credits: Wikipedia. https://en.wikipedia.org/wiki/Convolution	16

1.14	Operazione di convoluzione 2D discreta, in blu abbiamo una matrice 5x5 che rappresenta un'immagine, mentre in verde il risultato di una convoluzione, con un kernel 3x3. I valori del kernel sono raffigurati in basso a destra delle caselle interessate dalla convoluzione. credits: Dumoulin et al. 2016 [3].	16
1.15	Immagine di Lena Forsén, prima e dopo la convoluzione con i due kernel, che amplificano i bordi verticali e orizzontali.	18
1.16	Esempio di convoluzione con un kernel 3x3, stride pari a 2, e padding (zero padding) pari a 1. credits: Dumoulin et al. 2016 [3]	19
1.17	Esempio di max pooling con una finestra di dimensione 3x3, stride pari a 1 e padding 0. credits: Dumoulin et al. 2016 [3]	21
1.18	Esempio di average pooling con una finestra di dimensione 3x3, stride pari a 1 e padding 0. credits: Dumoulin et al. 2016 [3]	21
1.19	Esempio di convoluzione di un'immagine rgb. credits: Irhum Shafkat [16]	22
1.20	Esempio di convoluzione multichannel. credits: Irhum Shafkat [16]	22

Bibliografia

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.
- [2] George Cybenko. Approximation by superpositions of a sigmoidal function, 1989.
- [3] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning, 2016.
- [4] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, 2018.
- [5] Kunihiro Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, 1980.
- [6] Rinon Gal, Dana Cohen, Amit Bermano, and Daniel Cohen-Or. Swagan: A style-based wavelet-driven generative model, 2021.
- [7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation, 2021.
- [8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
- [9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.
- [10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.
- [11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.
- [12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition, 1989.
- [13] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge?, 2018.

- [14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.
- [15] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.
- [16] Irhum Shafkat. Intuitively understanding convolutions for deep learning, 2018.
- [17] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021.
- [18] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks, 1989.
- [19] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.