Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea Magistrale

Riconoscimento di dispositivi di protezione individuale in ambito industriale tramite infrastruttura cloud

Relatore:
Prof. Luca Ardito

Candidato:
Rei Zoto

Dicembre 2024

Indice

1 Introduzione **4**

2 Background **7**

2.1 Infortuni, Sicurezza Industriale e DPI 7

2.2 Computer Vision e Sicurezza sul Lavoro 12

2.3 Cloud Computing nell'Industria 16

2.4 Amazon Rekognition 17

2.5 Lavori Simili 19

3 Tecnologie **21**

3.1 Docker 21

3.2 MQTT 24

3.3 RTSP 28

3.4 Amazon Web Services 28

3.4.1 Infrastructure as Code 28

3.4.2 Serverless 28

3.4.3 Iam Policy 28

3.4.4 Kinesis Video Streams 28

3.4.5 AWS Iot Core e Greengrass 28

3.5 Apache Flink 28

3.6 GStreamer 28

4 Implementazione del sistema **29**

A Codice **31**

B Matematica Reti Convoluzionali **32**

Bibliografia **36**

Elenco delle figure

2.1 Infortuni sul lavoro accertati positivi per genere e modalità di accadimento nell'anno 2022. 7
2.2 infortuni in occasione di lavoro accertati positivi per settore di attività nell'anno 2022 8
2.3 onere economico complessivo stimato (approccio bottom up) 9
2.4 stima dei costi complessivi approccio top down 10
2.5 Modello del neurone artificiale sulla base del funzionamento di un neurone biologico. 13
2.6 Gatto di Hubel e Wiesel. Questi studi hanno permesso di dare una definizione di recettore visivo, scoprire l'organizzazione gerarchica della corteccia visiva nei mammiferi e formalizzare il concetto di retinotopia. 14
2.7 Lenet-5(1998). Primo modello ad aver dimostrato l'efficacia delle reti convoluzionali (CNN) nella comprensione delle immagini e ha aperto la strada a molte delle architetture moderne di deep learning. 16
2.8 Rilevamento tramite Rekognition dei dispositivi di sicurezza individuali. . . 18

3.1 Confronto tra macchine virtuali tradizionali e containers. Essi vengono eseguiti direttamente sul sistema operativo dell'host, senza la necessità di chiamate di sistema multiple. Vengono impacchettate solo le librerie necessarie nell'immagine di base fornita al container. 21
3.2 Esempio di condivisione di layers in Docker. Le due immagini risultanti hanno la stesso sistema di base, condividendo lo stesso ambiente (Debian) e runtime (Python). 23
3.3 Schema architettura publish-subscriber. 25
3.4 Flusso pacchetti di controllo per QoS2. I livelli più semplici funzionano in maniera simile, con un numero di pacchetti inferiore. Si può notare come il disacoppiamento fornito dal broker faciliti lo scambio dei messaggi. . . . 26

Capitolo 1

Introduzione

La sicurezza sul lavoro rappresenta un elemento fondamentale all'interno dell' industria manifatturiera, dove l'interazione tra macchinari complessi e operai espone a numerosi rischi. Come noto, gli infortuni sul lavoro nel settore manifatturiero sono tra i più frequenti e gravi, con conseguenze significative sia per i lavoratori che per le aziende. Garantire un ambiente di lavoro sicuro non solo tutela la salute e il benessere dei dipendenti, ma contribuisce anche a migliorare la produttività e a ridurre i costi associati agli incidenti. Essi infatti possono comportare gravi conseguenze per i lavoratori, inclusi infortuni permanenti, invalidità e, in casi estremi, decessi. Tali incidenti non solo influiscono sulla qualità della vita dei dipendenti e delle loro famiglie, ma comportano anche ripercussioni economiche rilevanti per le aziende. I costi diretti includono spese mediche e indennità di infortunio, mentre i costi indiretti comprendono la perdita di produttività, la necessità di sostituzione del personale e i danni alla reputazione aziendale. Oltre alle conseguenze dirette sugli individui, gli incidenti sul lavoro hanno un impatto economico significativo sulle aziende e sulla società nel suo complesso. Le aziende devono affrontare spese legali, aumenti dei premi assicurativi e potenziali sanzioni normative in caso di inadempienza alle leggi sulla sicurezza. Inoltre, la perdita di fiducia dei consumatori e dei partner commerciali può influenzare negativamente le performance finanziarie e la competitività dell'azienda sul mercato. Sul piano sociale, gli incidenti sul lavoro contribuiscono a un aumento dei costi sanitari e riducono la produttività nazionale. La società nel suo complesso subisce un impatto economico derivante dalla perdita di forza lavoro qualificata e dall'aumento delle richieste di assistenza sociale. Pertanto, investire nella sicurezza sul lavoro rappresenta non solo un obbligo etico e legale, ma anche una strategia economica vantaggiosa a lungo termine.

I Dispositivi di Protezione Individuale (DPI) sono strumenti essenziali per prevenire gli incidenti sul lavoro e ridurre l'esposizione dei lavoratori a rischi specifici. DPI comuni includono caschi, guanti, occhiali protettivi, maschere respiratorie e indumenti resistenti

agli agenti chimici. L'uso corretto e costante dei DPI è fondamentale per garantire la sicurezza dei lavoratori, ma la loro efficacia dipende dalla conformità e dalla corretta applicazione delle normative da parte dei dipendenti. Inoltre, monitorare l'uso dei DPI in ambienti industriali può risultare complesso, soprattutto in contesti ad alta dinamicità e con elevati volumi di produzione. Tradizionalmente, questo monitoraggio è stato effettuato attraverso ispezioni manuali, che possono essere dispendiose in termini di tempo e risorse e soggette a errori umani. Pertanto, vi è una crescente necessità di soluzioni automatizzate e tecnologicamente avanzate per garantire un controllo efficace e continuo dell'utilizzo dei DPI. L'innovazione tecnologica ha aperto nuove prospettive per migliorare la sicurezza sul lavoro nell' industria manifatturiera. In particolare, la computer vision e il cloud computing emergono come strumenti potenti per automatizzare il rilevamento dei DPI e monitorare in tempo reale le condizioni di sicurezza.

La **computer vision** permette alle macchine di interpretare e analizzare immagini e video, identificando automaticamente la presenza e l'uso corretto dei DPI. Attraverso algoritmi di deep learning, i sistemi di computer vision possono riconoscere oggetti specifici, come caschi e guanti, e verificare la loro corretta indossatura da parte dei lavoratori. Questo approccio non solo aumenta l'efficienza del monitoraggio, ma riduce anche la dipendenza da interventi manuali, minimizzando gli errori e garantendo una supervisione costante e accurata. Il **cloud computing**, d'altra parte, fornisce l'infrastruttura necessaria per gestire e analizzare grandi quantità di dati provenienti dai sistemi di computer vision. Attraverso piattaforme cloud, è possibile archiviare, elaborare e accedere ai dati in modo scalabile e flessibile, permettendo una gestione centralizzata e accessibile delle informazioni sulla sicurezza. Inoltre, il cloud computing facilita l'integrazione con altri sistemi aziendali, consentendo una visione completa delle operazioni e una risposta tempestiva agli incidenti rilevati. L'integrazione di computer vision e cloud computing rappresenta quindi una svolta nel campo della sicurezza industriale, offrendo soluzioni avanzate per il monitoraggio dei DPI e la prevenzione degli incidenti.

In questo contesto, la presente tesi si propone di sviluppare un sistema basato su Amazon Rekognition, un servizio di computer vision offerto da Amazon Web Services (AWS), per il rilevamento automatico dei DPI nell' industria manifatturiera. L'obiettivo principale è quello di generare una infrastruttura scalabile per l'analisi di dati semistrutturati e non strutturati all'interno di una fabbrica. In particolare, dato un insieme di macchinari, come ad esempio bracci robotici, si vuole ottenere il controllo dell'effettivo indossamento dei dispositivi di sicurezza da parte degli operatori(operai, manutentori) all'interno di uno stabilimento, in modo tale da garantirne loro la sicurezza sul posto di lavoro.

Capitolo 2

Background

2.1 Infortuni, Sicurezza Industriale e DPI

In questa sezione si vedranno delle statistiche relative agli infortuni sul lavoro, quale sia la risposta normativa al problema della sicurezza industriale dal punto di vista degli attori coinvolti, integrata con la definizione di dispositivi di sicurezza. Questo tema è di primaria importanza per garantire non solo la salute e il benessere dei lavoratori, ma anche l'efficienza operativa e la sostenibilità economica delle aziende. Secondo i dati forniti dall'Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro (INAIL), nel 2022 il settore manifatturiero ha registrato un tasso di infortuni del 13,9% sul totale [1].

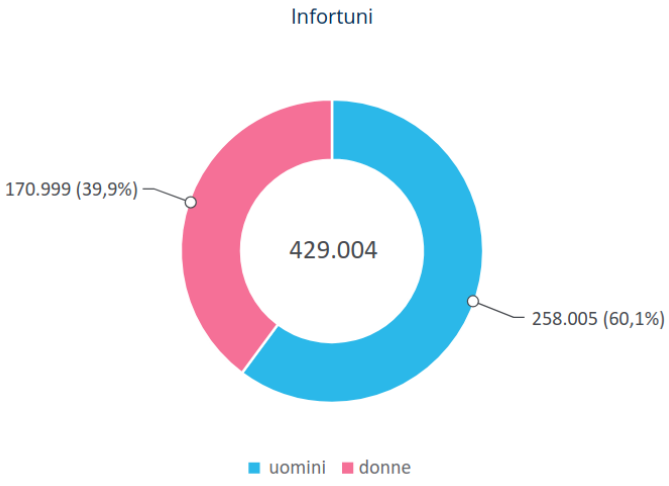

Figura 2.1: Infortuni sul lavoro accertati positivi per genere e modalità di accadimento nell'anno 2022.

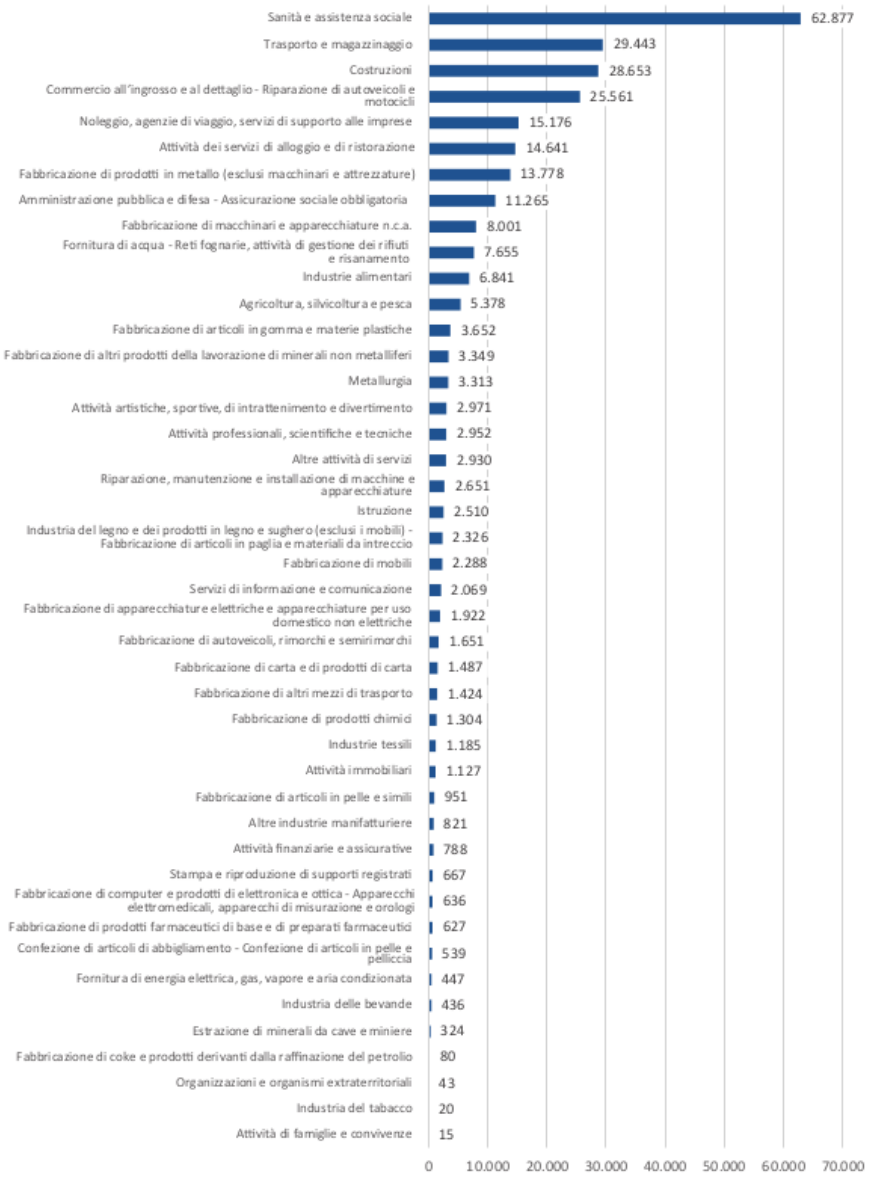

Figura 2.2: infortuni in occasione di lavoro accertati positivi per settore di attività nell'anno 2022

Essi comportano gravi conseguenze per i dipendenti, inclusi infortuni permanenti, invalidità e, nei casi più gravi, decessi. Oltre al costo umano, gli incidenti sul lavoro hanno un impatto significativo sull'economia delle aziende, generando costi diretti come spese mediche e indennità di infortunio, e costi indiretti come perdita di produttività, danni reputazionali e aumento dei premi assicurativi. L'EU-Occupational Safety and Health Administration (EU-OSHA) a questo proposito ha stimato in due diversi approcci l'impatto degli incidenti sul lavoro all'interno dell'Unione Europea[2]. Nell'indagine sono stati presi in esame i dati relativi a 5 Paesi, poiché più completi e accessibili (tra cui figura anche l'Italia) e sono stati mostrati i risultati seguendo due diversi approcci: uno bottom-up, perché prende i valori dei costi per ciascun infortunio e li valuta globalmente; l'altro top-down, in quanto stima l'impatto dell'infortunio sulla vita del lavoratore e da valori macroeconomici come il PIL pro-capite valuta il costo effettivo dell'infortunio sul singolo. In termini pratici, nel primo caso si tiene conto dei costi diretti, indiretti e immateriali (effetti sulla vita e sulla salute) mentre nel secondo del valore monetario espresso in DALY, cioè il costo in termini di anni di vita persi a causa di un infortunio o di una malattia.

Paese		Finlandia	Germania	Paesi Bassi	Italia	Polonia
Numero di casi		131 867	2 262 031	323 544	1 907 504	1 156 394
Costi diretti	In Mio EUR	484	10 914	2 137	8 491	1 882
Costi diretti, % rispetto al totale		8	10	9	8	4
Costi indiretti	In Mio EUR	4 362	70 658	6 468	58 961	19 588
Costi indiretti, % rispetto al totale		72	66	69	56	45
Costi immateriali	In Mio EUR	1 196	25 557	5 147	37 392	22 311
Costi immateriali, % rispetto al totale		20	24	22	36	51
Onere economico complessivo	In Mio EUR	6 042	107 129	23 751	104 844	43 781
Percentuale rispetto al PIL		2,9	3,5	3,5	6,3	10,2

Figura 2.3: onere economico complessivo stimato (approccio bottom up)

Il risultato di queste analisi ha mostrato che per l'Italia il costo di un infortunio o malattia causata dal posto di lavoro aveva un impatto percentuale sul PIL del 6,3% nel primo caso, mentre nell'approccio top down, riferendosi alla metodologia VSLY - considerata più coerente con i risultati dell'approccio bottom up - il valore medio era del 7,7% rispetto alla produzione interna. I valori ottenuti, indipendentemente dall'approccio utilizzato, non si discostano troppo l'uno dall'altro, confermando l'attendibilità dell'analisi. Si può dedurre perciò quanto questo problema sia concreto e impatti sulla società e sull'economia dell'Italia, dove il posto di lavoro è in gran parte costituito dall'industria.

	Germania	Finlandia	Italia	Paesi Bassi	Polonia
DALY					
Totale dei DALY professionali	1 236 855	64 516	853 817	248 464	507 068
Percentuale rispetto ai DALY totali	4,9	4,2	5,1	5,7	4,0
DALY professionali per ogni 10.000 persone occupate	308	265	380	299	315

	Mio EUR	% rispetto al PIL	Mio EUR	% rispetto al PIL	Mio EUR	% rispetto al PIL	Mio EUR	% rispetto al PIL	Mio EUR	% rispetto al PIL
COSTI										
Approccio basato sul capitale umano										
Valore minimo	24 597	0,8	1 419	0,7	13 530	0,8	5 290	0,8	2 692	0,6
Media	55 429	1,8	3 106	1,5	31 475	1,9	11 879	1,7	6 929	1,6
Mediana	39 712	1,3	2 291	1,1	23 865	1,4	8 708	1,3	4 656	1,1
Massimo	138 404	4,5	7 393	3,5	69 671	4,2	30 114	4,4	17 037	4,0
Approccio WTP										
Valore minimo	32 324	1,1	1 637	0,8	20 929	1,3	3 276	0,5	5 118	1,2
Media	66 251	2,2	5 814	2,8	42 895	2,6	14 613	2,1	9 676	2,3
Mediana (a)	66 251	2,2	4 335	2,1	42 895	2,6	13 953	2,0	8 863	2,1
Massimo	100 177	3,3	17 453	8,3	64 861	3,9	30 767	4,5	15 861	3,7
Approccio VSLY/VOLY										
Valore minimo	60 609	2,0	4 214	2,0	52 304	3,2	9 649	1,4	12 790	3,0
Media	191 939	6,3	9 345	4,5	133 789	8,1	38 016	5,6	43 836	10,2
Mediana	166 943	5,5	8 633	4,1	126 876	7,7	33 248	4,9	31 026	7,2
Massimo	420 489	13,8	19 425	9,3	256 120	15,5	77 016	11,3	119 149	27,7

(a) Nel caso della Germania e dell'Italia i valori mediani e medi dell'approccio WTP coincidono perché, per questi due paesi, abbiamo potuto inserire solo due valori centrali europei di riferimento (i valori minimi e massimi riportati nella tabella).

Figura 2.4: stima dei costi complessivi approccio top down

L'utilizzo corretto dei Dispositivi di Protezione Individuale (DPI) è fondamentale per prevenire tali incidenti. Secondo la legislazione italiana, per DPI si intende *qualsiasi attrezzatura destinata ad essere indossata e tenuta dal lavoratore allo scopo di proteggerlo contro uno o più rischi suscettibili di minacciarne la sicurezza o la salute durante il lavoro, nonché ogni complemento o accessorio destinato a tale scopo*[3].

La normativa in materia di sicurezza sul lavoro è un sistema complesso e articolato, volto a tutelare la salute e la sicurezza dei lavoratori in ogni settore produttivo. Il fulcro di questo sistema è rappresentato dal **Decreto Legislativo 81/2008**, conosciuto come *Testo Unico sulla Salute e Sicurezza sul Lavoro*. Questo decreto introduce una serie di obblighi inderogabili per i datori di lavoro, al fine di garantire un ambiente salubre e sicuro. Tra i principi cardine si evidenziano:

- **Valutazione dei rischi**: il datore di lavoro, con l'ausilio di un responsabile di sicurezza ed un medico esperto, è tenuto ad effettuare un'attenta e completa valutazione di tutti i rischi presenti sul luogo di lavoro, compresi anche per gruppi di lavoratori specifici. A questo scopo deve redarre un documento dove vengono presi in considerazione tutti i criteri utilizzati nella valutazione dei rischi.
- **Programmazione della prevenzione**: sulla base della valutazione dei rischi, il datore di lavoro, nello stesso documento, deve individuare i dispositivi di sicurezza

necessari nelle attività lavorative ed elaborare un piano di prevenzione, nell'ottica di eliminare o ridurre al minimo i rischi individuati. Questo piano deve essere integrato con le condizioni tecniche, ambientali e produttive dell'azienda, garantendo la sua effettiva applicabilità e sostenibilità.

- **Informazione e formazione dei lavoratori**: i lavoratori devono essere informati in modo chiaro e completo sui rischi generali dell'azienda e su quelli specifici a cui sono esposti durante lo svolgimento delle loro mansioni, su come effettuare un primo soccorso e a chi rivolgersi nell'ottica di prevenzione dei rischi. Devono inoltre ricevere una formazione adeguata sulla loro prevenzione, adottare comportamenti sicuri e utilizzare correttamente macchinari e dispositivi di protezione individuale. L'informazione e la formazione devono essere fornite prima dell'inizio dell'attività lavorativa e devono essere ripetute periodicamente, garantendo l'aggiornamento costante dei lavoratori, nel caso ad esempio vengano cambiate le mansioni, oppure siano introdotte nuove attrezzature e tecnologie.
- **Sorveglianza sanitaria**: questa misura è fondamentale per monitorare lo stato di salute degli operatori in relazione ai rischi cui sono esposti, prevenire l'insorgenza di malattie professionali e garantire l'idoneità alla mansione. La sorveglianza sanitaria è effettuata da un medico competente, che ha il compito di visitare i lavoratori, effettuare gli accertamenti sanitari necessari e rilasciare il giudizio di idoneità.

I DPI rappresentano l'ultima barriera di protezione, quando le misure tecniche e organizzative non sono sufficienti a eliminare o ridurre i rischi. Pertanto, la loro scelta, il loro utilizzo e la loro manutenzione devono essere effettuati con la massima attenzione e responsabilità. Vengono suddivisi nelle seguenti categorie in base alla loro funzione:

- **Protezione della testa**: caschi di protezione per l'industria.
- **Protezione dell'udito**: cuffie antirumore, tappi auricolari.
- **Protezione degli occhi e del viso**: occhiali protettivi, visiere.
- **Protezione delle vie respiratorie**: mascherine antipolvere e respiratorie.
- **Protezione degli arti superiori e inferiori**: guanti di protezione, scarpe antinfortunistiche, ginocchiere.
- **Indumenti di protezione**: tute, grembiuli, indumenti ad alta visibilità.

Gli standard, giocano un ruolo fondamentale nel definire tecnicamente i criteri di produzione, utilizzo e manutenzione dei DPI, garantendo un elevato livello di protezione per gli

utenti. La legge europea, come evidenziato nel **Regolamento (UE) 2016/425** stabilisce i requisiti che i DPI devono soddisfare nel mercato unico[4], tra cui:

- **Ergonomia**: i DPI devono essere progettati e fabbricati in modo da essere comodi da indossare e non limitare la libertà di movimento del lavoratore, evitando di interferire con lo svolgimento delle sue attività, garatendone allo stesso tempo la sicurezza.
- **Livelli e classi di protezione**: i DPI devono fornire un livello di protezione adeguato al rischio specifico da cui proteggono. La classificazione dei DPI in base al livello di protezione consente di scegliere il dispositivo più idoneo in relazione al rischio da prevenire.
- **Marcatura**: i DPI devono essere marcati con il simbolo **CE**, a indicare la loro conformità ai requisiti di sicurezza dell'Unione Europea. La marcatura CE deve essere apposta in modo visibile, leggibile e indelebile sul DPI o sulla sua confezione.
- **Istruzioni e informazioni del fabbricante**: i DPI devono essere accompagnati da istruzioni chiare e complete (e.g. rischi coperti, prestazioni, classi di protezione, accessori, pezzi di ricambio etc.) per l'utilizzatore, che indichino in modo dettagliato come utilizzare, conservare, pulire e manutenere correttamente il dispositivo. Le istruzioni devono essere redatte in una lingua comprensibile nello Stato membro in cui il DPI è commercializzato. I dispositivi fabbricati devono avere una sorgente (il produttore e il suo indirizzo) ed essere identificati dal lotto messo in commercio.

2.2 Computer Vision e Sicurezza sul Lavoro

La computer vision è un campo dell'informatica incentrata sulla comprensione del contenuto di immagini o video per mezzo di un calcolatore. I task che si possono svolgere sono di diversi tipologie, tra cui la classificazione, l'object detection, la segmentazione, il riconoscimento di volti, l'encoding e l'applicazione di filtri per la modifica delle immagini originali. La ricerca sulle reti neurali nell'ambito della computer vision è stata tra le prime a mostrare le potenzialità di questa tecnologia nella risoluzione di problemi nel mondo reale. Storicamente l'insieme di diversi sviluppi nelle discipline di neuroscienza, deep learning e matematica ha permesso il raggiungimento di questo traguardo. Le scoperte relative al neurone biologico, la modellazione dei primi neuroni artificiali e la successiva estensione a più strati, l'utilizzo del calcolo differenziale per l'aggiornamento dei pesi, l'introduzione di funzioni di attivazione e di perdita sempre più complesse, ed infine la formulazione del teorema di approssimazione universale sono sicuramente gli elementi fondamentali di

questo successo. Alla fine degli anni '50 è stato modellato il primo neurone artificiale, prendendo ispirazione dal neurone biologico, composto dalla combinazione lineare di input e pesi in ingresso ad una funzione di attivazione.

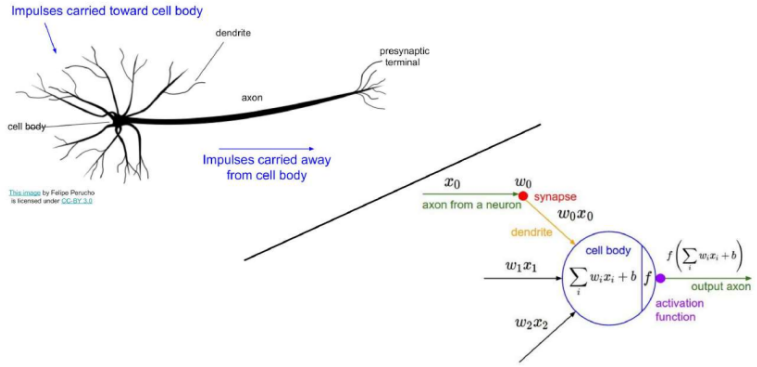

Figura 2.5: Modello del neurone artificiale sulla base del funzionamento di un neurone biologico.

Questo semplice meccanismo era in grado di mimarne grossolanamente il comportamento, generando una risposta a dei dati in ingresso, in modo tale che, superata una certa soglia, producesse o meno un valore in uscita. La funzione di attivazione era una semplice funzione gradino (al tempo non era scontato generare funzioni non lineari e continue), ma comunque questo oggetto era in grado di risolvere problemi di classificazione binaria. Il limite principale di questo modello consisteva nell'aggiornamento dei pesi in caso di predizioni sbagliate, basato su una delta di valori discreti. La funzione di aggiornamento dei pesi forniva in maniera euristica una direzione verso l'insieme ottimale delle variabili interne al modello, per ottenere la predizione il più possibile corretta ad ogni nuovo input. Per risolvere questo limite, venne definita una funzione di attivazione continua, trasformando il problema da uno di classificazione ad uno di regressione. Questa nuova costruzione permetteva di introdurre una funzione di perdita, nell'ottica di minimizzare l'errore nelle predizioni attraverso un approccio più rigoroso. Dalla teoria delle regressioni lineari infatti si poteva utilizzare il metodo dei minimi quadrati, che in termini pratici significava ridurre il più possibile l'errore nella rappresentazione della funzione che si voleva apprendere dai dati. Fino alla fine degli anni '60 si sperimentò l'utilizzo di questi modelli, di cui gli esempi più famosi sono Adaline e Madaline, costituiti da semplici reti di neuroni artificiali, rispettivamente ad uno e due strati. Esse non riuscivano a rappresentare correttamente le non linearità all'interno della distribuzione dei dati, ma si trattava solo di un limite tecnico e non teorico, poiché non erano ancora state introdotte funzioni di attivazione non lineari continue come la sigmoide e non era ancora stato compreso come propagare l'aggiornamento dei pesi negli strati nascosti.

Nella seconda ondata di ricerca sulle reti neurali, iniziata negli '80, è stato dimostrato che è teoricamente possibile approssimare qualsiasi distribuzione dei dati attraverso l'apprendimento automatico di reti neurali con almeno uno strato di neuroni artificiali, aventi delle funzioni di attivazione non lineari. Questo teorema prende il nome di teorema di approssimazione universale. Esso si applica a tutte le tipologie più comuni di problemi risolti nel machine learning, quindi problemi discriminativi come la classificazione e la regressione e problemi generativi, come ad esempio l'encoding di immagini, la generazione di testo etc. Le implicazioni di questa dimostrazione hanno avuto un forte impatto solo in tempi più recenti, ma per comprenderne appieno le cause bisogna ancora revisionare alcuni elementi fondamentali in questa storia. Dalla neuroscienza infatti, non si è soltanto preso ispirazione per la modellazione del perceptron, tant'è che a partire dagli anni '50 è stato studiato il funzionamento della corteccia visiva nel cervello di alcuni mammiferi. Fondamentalmente con questi studi è stato dimostrato che i neuroni all'interno di questa zona sono organizzati gerarchicamente e nel livello più semplice rispondono a stimoli visivi con caratteristiche specifiche, come l'orientamento e le traslazioni.

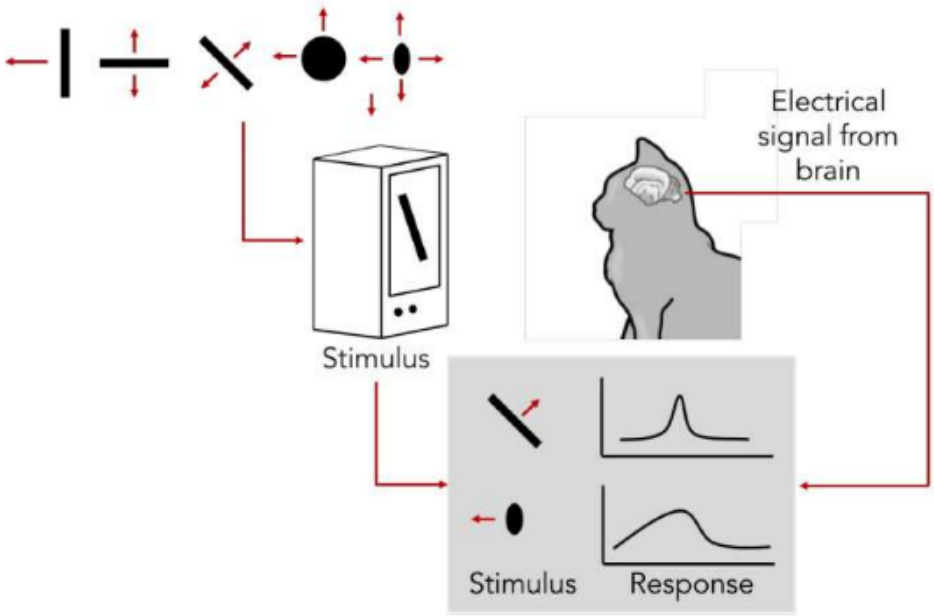

Figura 2.6: Gatto di Hubel e Wiesel. Questi studi hanno permesso di dare una definizione di recettore visivo, scoprire l'organizzazione gerarchica della corteccia visiva nei mammiferi e formalizzare il concetto di retinotopia.

Nel 1980 venne proposto il Neocognitron, un antenato delle moderne reti convoluzionali. In questo modello sono stati trasposti i precedenti principi, in quanto, oltre ad implementare una architettura gerarchica con strati di neuroni, è stato definito matematicamente come modellare dei campi recettivi, cioè in che modo identificare delle forme semplici con

diverse orientazioni dall'immagine di input, come succede per i recettori delle immagini provenienti dal campo visivo oculare. La definizione è stata presa della teoria dei segnali usando la formula della convoluzione:

$$y[n] = (x * h)[n] = \sum_{k=-\infty}^{+\infty} x[k] \cdot h[n-k] \tag{2.1}$$

Classicamente, questa espressione permette la generazione di diversi filtri, in modo tale da modulare o isolare solo parti del segnale di interesse, eliminandone altre che possono non essere utili a successive trasformazioni o semplicemente perché fonti di rumore. Nel dominio dell'image processing si voleva sfruttare esattamente questa proprietà: applicare la funzione di convoluzione in modo da isolare le caratteristiche desiderate all'interno di una figura. Applicando questa formula nel dominio spaziale e defininendo dei filtri bidimensionali, l'espressione assume la seguente forma:

$$y[i,j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} h[m,n] \cdot x[i-m, j-n] \tag{2.2}$$

Così non solo era possibile emulare il comportamento dei recettori visivi, ma allo stesso tempo si implementava il concetto di retinotopia. Il prodotto scalare di un filtro in una singola sezione dell'immagine genera la stessa formula di un neurone artificiale (vedi Appendice B per dettagli), per cui ogni attivazione all'interno di ciascuna feature map (il risultato di una intera convoluzione) simula esattamente il modello del perceptron. La retinotopia definisce una relazione locale tra elementi vicini del campo visivo e neuroni vicini all'interno della corteccia visiva. Allo stesso modo attivazioni vicini nella feature map corrispondono ad elaborazioni di elementi vicini nell'immagine. In questo modo si potevano identificare forme di interesse in diverse angolazioni. Per ottenere invece lo stesso effetto di invarianza dalla posizione delle forme nell'immagine - in altri termini l'identificazione di queste indipendentemente da traslazioni nell'immagine - sono stati definiti degli strati di pooling. Questo modello presentava principalmente un grosso limite: il metodo di allenamento non era supervisionato e non si basava su una funzione di perdita globale, infatti l'utilizzo della backpropagation non era ancora stato formalizzato. Gli strati più interni della rete non permettevano la rappresentazione di forme più complesse e più coerenti con l'oggetto da classificare. Inoltre la rete era addestrata per il pattern recognition, ma non aveva una utilità pratica rispetto ai problemi più comuni nella computer vision. L'introduzione di uno strato fully connected, l'utilizzo di una funzione di perdita globale per la classificazione e della backpropagation portarono all'architettura di Lenet, nel 1998. L'allenamento di questa rete era specifico per la classificazione e tutti i neuroni dell'architettura partecipavano al training, quindi anche quelli degli strati convoluzionali. In altre parole si trattava di una rete end-to-end. AlexNet, la rete che segna

una netta linea di demarcazione nel deep learning, mantiene la stessa architettura, con una principale differenza: le funzioni di attivazione all'interno della rete permettono la propagazione del gradiente senza saturazioni negli strati più interni.

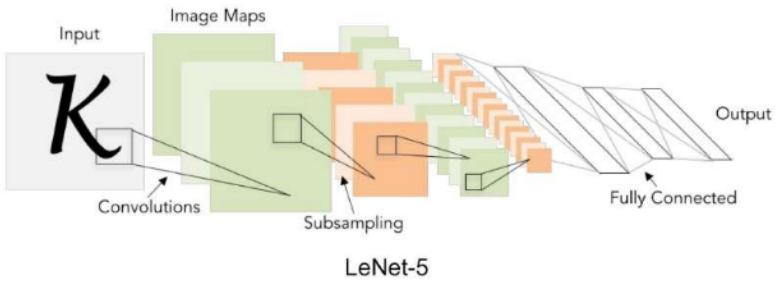

Figura 2.7: Lenet-5(1998). Primo modello ad aver dimostrato l'efficacia delle reti convoluzionali (CNN) nella comprensione delle immagini e ha aperto la strada a molte delle architetture moderne di deep learning.

La riduzione dell'errore nei problemi di classificazione nei problemi di vision artificiale è stata solo una naturale conseguenza: l'architettura ormai era chiara e funzionante, si trattava solo di aumentare il numero di neuroni e strati all'interno della rete, grazie ad una potenza di calcolo che ai tempi di Lenet non era disponibile.

2.3 Cloud Computing nell'Industria

Il cloud computing rappresenta una delle innovazioni più rilevanti degli ultimi decenni nel settore IT, trasformando il modo in cui le aziende gestiscono le proprie risorse informatiche e processi produttivi. Questo nuovo paradigma, basato sull'erogazione di servizi tramite Internet, consente di accedere a risorse come server, storage, database e applicazioni software in modo scalabile e on-demand, senza dover effettuare investimenti iniziali significativi in infrastrutture hardware. Le implicazioni di questa trasformazione sono profonde, in quanto ridefiniscono i modelli di gestione IT e le strategie aziendali, favorendo un approccio più veloce e flessibile nell'implementazione di nuove soluzioni. La principale innovazione apportata dal cloud computing risiede nella possibilità di adattare rapidamente le risorse informatiche alle necessità aziendali, garantendo una scalabilità notevolmente superiore rispetto alle infrastrutture tradizionali. In passato, le aziende che desideravano espandere i propri sistemi erano costrette a effettuare investimenti consistenti in hardware e a sostenere costi elevati per la relativa gestione e manutenzione. Inoltre, la diversa geolocalizzazione dei datacenter comporta vantaggi in termini di accessibilità, permette di risolvere problemi di latenza e di personalizzare i servizi in base alla regione in cui l'applicazione eseguita sul cloud viene deployata.

Oltre a facilitare la gestione delle risorse e ridurre i costi infrastrutturali, il cloud computing è considerato una tecnologia abilitante nell'implementazione dell'Industria 4.0. Ogni era industriale è stata segnata da una svolta tecnologica: nella prima l'introduzione della macchina a vapore, nella seconda l'elettrificazione delle macchine e la conseguente implementazione della catena di montaggio. La terza rivoluzione è stata possibile grazie all'invenzione del transistor e la successiva democratizzazione dei calcolatori. Questa nuova ondata invece è incentrata sui dati: ad esempio, nel 2015 è stato stimato che solo l'1% delle informazioni generate dai sensori all'interno di una fabbrica veniva effettivamente elaborata[5]. L'adozione di tecnologie quali l'Internet of Things (IoT), gli sviluppi moderni nell'intelligenza artificiale e l'analisi di big data sono gli elementi che concorrono a questa nuova rivoluzione. Il primo di questi fattori è fondamentale per la generazione e l'ingestion, mentre gli altri due per il processamento: indipendemente dalle loro funzioni, i dati restano il fulcro delle operazioni. In questo contesto, il cloud fornisce l'infrastruttura e i servizi necessari per l'integrazione di questi elementi, rendendo possibile l'implementazione delle smart factories.

La capacità del cloud di raccogliere, archiviare ed elaborare grandi quantità di dati in tempo reale è cruciale per sfruttare appieno il potenziale dell'Industria 4.0. Le aziende che operano in settori industriali tradizionali, come la manifattura, possono trasformare le loro linee di produzione in sistemi autonomi e ottimizzati, capaci di adattarsi alle esigenze del mercato e di ridurre significativamente gli sprechi. La connettività fornita dal cloud consente invece di collegare dispositivi, sensori e macchinari all'interno della fabbrica, creando un ecosistema in cui ogni componente è in grado di comunicare e condividere le proprie informazioni, rendendo più semplice il monitoraggio dei processi produttivi. Inoltre con gli avanzamenti nella ricerca sul deep learning, diventato sempre più consistente negli anni, i relativi modelli sono stati adottati per migliorare il processo decisionale nelle aziende. Un esempio concreto è la manutenzione predittiva, che sfrutta i dati provenienti dai sensori per rilevare anomalie e prevedere i guasti delle macchine. E' così possibile ridurre i tempi di inattività, prolungare la vita utile delle apparecchiature e migliorare la loro efficienza complessiva. Sempre nello stesso contesto, un'azienda potrebbe utilizzare il cloud per raccogliere e analizzare dati provenienti dalle linee di assemblaggio, applicando modelli predittivi per migliorare la qualità dei componenti e ridurre i difetti di produzione.

2.4 Amazon Rekognition

Esistono diversi providers di servizi cloud, di cui uno dei più diffusi è Amazon Web Services (AWS). Fin dalla sua nascita è stato sempre considerato uno dei principali innovatori in questo dominio, non solo perché il primo ad avere introdotto il concetto di cloud nel

2006. Amazon è da sempre all'avanguardia nel fornire nuove funzionalità: è passata dalle risorse di calcolo on-demand e dei servizi gestiti come storage e database, all'introduzione del paradigma serverless, fino all'integrazione di servizi per l'IoT e per il machine learning. AWS infatti possiede un ricco ecosistema per la generazione di soluzioni basate sull'apprendimento automatico come SageMaker, Bedrock e Rekognition. Ciascuno risponde a esigenze specifiche: SageMaker funge da piattaforma di base per lo sviluppo e l'addestramento di modelli personalizzati, mentre Bedrock offre accesso a diversi foundation models, semplificando l'utilizzo di questa nuova tecnologia. Infine, Rekognition si posiziona come una soluzione specializzata per l'analisi di immagini, video e streaming, dove le aziende possono implementare le relative funzionalità senza dover sviluppare o addestrare modelli. Serve solo invocare la API di interesse. Rekognition è quindi particolarmente utile per le aziende che necessitano di integrazioni rapide e affidabili nell'ambito della visione artificiale all'interno dei loro processi. I casi d'uso spaziano su numerosi domini: ad esempio, può essere utilizzato per estrarre metadati da un testo scritto a mano, oppure per la moderazione dei contenuti nelle piattaforme social.

Figura 2.8: Rilevamento tramite Rekognition dei dispositivi di sicurezza individuali.

Nell'ambito di questo scritto, Rekognition viene usato per l'identificazione dei dispositivi di sicurezza e per la valutazione del loro corretto utilizzo. Dalla documentazione Amazon viene mostrato un esempio di questo servizio in azione. La API utilizzata è DetecProtectiveEquipment, che in questo caso è in grado di rilevare casco, maschera e guanti da lavoro. Più nel dettaglio, la risposta di questa richiesta sarà una struttura dati contenente: le persone all'interno dell'immagine, le parti del corpo che indossano i dispositivi di sicurezza (differenziando ad esempio quale mano indossa i guanti), il tipo di dispositivi rilevati e l'associazione tra parti del corpo e dispositivi, in modo tale da verificare quali siano correttamente indossati. La chiamata può essere configurata per rilevare tutti i DPI

più comuni oppure solo un sottoinsieme di essi in base alla necessità. Le API di Rekognition si differenziano in due possibili sottocategorie: storage e non-storage. La distinzione consiste nel fatto che il servizio può salvare o meno le informazioni relative all'analisi dell'immagine o del video. Per questioni di privacy, non viene tracciatp alcun individuo e non c'è alcuna correlazione tra gli id restituiti da ciascun processamento. L'operazione è focalizzata solo sull'utilizzo regolare dei DPI e la documentazione è molto chiara in questo punto. Altre chiamate di Rekognition, come quelle basate sul riconoscimento facciale, hanno invece bisogno di salvare queste informazioni altrimenti non potrebbero funzionare.

2.5 Lavori Simili

Capitolo 3

Tecnologie

3.1 Docker

Docker è diventato negli anni il sistema più popolare per gestire la virtualizzazione basata su container. In realtà, la sua value proposition è legata alla possibilità di poter lanciare una applicazione ovunque, con la garanzia che questa funzioni sempre. L'analogia con i container nel mondo reale è diretta: i container utilizzati nel commercio, possono essere adattati indipendemente dall'ambiente di trasporto. Allo stesso modo l'applicazione che viene eseguita, funzionerà indipendemente dall'ambiente di esecuzione, qualsiasi sistema operativo esso sia, con l'unica limitazione sull'instruction set per cui è stata generata la build. Possono anche essere visti come macchine virtuali leggere, perché riescono a garantire le stesse caratteristiche di quelle tradizionali, ma con un overhead minore, sia in termini di efficienza nell'esecuzione, che di spazio occupato.

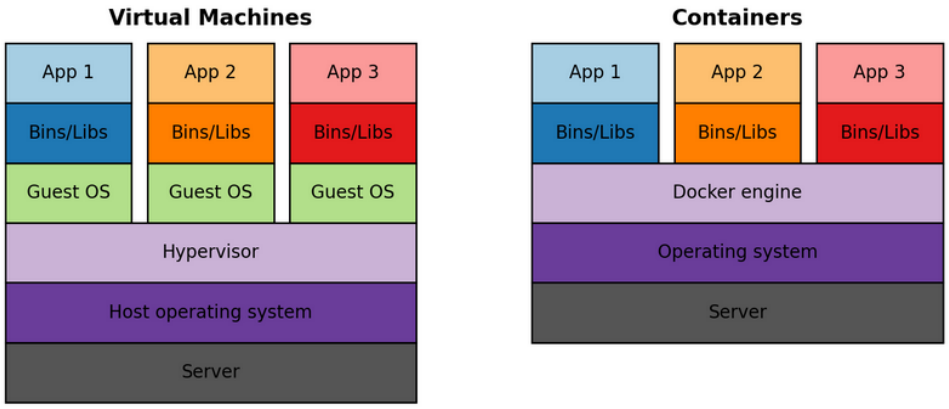

Figura 3.1: Confronto tra macchine virtuali tradizionali e containers. Essi vengono eseguiti direttamente sul sistema operativo dell'host, senza la necessità di chiamate di sistema multiple. Vengono impacchettate solo le librerie necessarie nell'immagine di base fornita al container.

Nella pratica esistono due definizioni secondo chi è l'utilizzatore di questa tecnologia: se si parla di gestione di una infrastruttura, allora l'attenzione è spostata verso la virtualizzazione, mentre se il focus è relativo alle applicazioni allora si può vedere Docker come una tecnologia orientata alla portabilità. In generale, le macchine virtuali devono garantire un serie di proprietà per poter essere definite tali, che sono:

- **consolidamento**: inizialmente ciascuna applicazione veniva eseguita su una macchina distinta, rendendo l'utilizzo delle risorse inefficiente. La virtualizzazione ha permesso di superare questa limitazione, con l'esecuzione di software differenti all'interno della stessa macchina.

- **isolamento**: questa proprietà consente ad ogni applicazione di vedere l'ambiente come se fosse una macchina a sé stante, nonostante il consolidamento. I sistemi operativi offrono funzionalità di isolamento base come l'astrazione dello spazio di indirizzamento di un processo, ma questo non è sufficiente per una applicazione self-contained. Diversi applicativi infatti possono interferire tra di loro già solo nell'utilizzo della memoria, nel caso di sovra-allocazioni, portando ad un degrado delle prestazioni o al crash(verifica). Lo stesso ragionamento si può applicare anche per la gestione della cpu, della rete e del filesystem che all'interno di un sistema operativo sono risorse completamente condivise tra processi.

- **flessibilità**: il controllo dei container è molto semplice in quanto docker prevede dei comandi standard per la loro gestione, come l'avvio, la pausa e lo stop. Gli orchestratori giocano un ruolo fondamentale per riallocare i container in nodi diversi dell'infrastruttura, per esempio in caso di manutenzione o sovraccarico. Questo processo avviene senza particolari overhead, in quanto i container, a differenza delle macchine virtuali tradizionali hanno dei tempi di attivazione molto bassi.

- **portabilità**: come già visto, i container possono essere avviati in qualsiasi piattaforma, indipendemente dal sistema operativo, dall'hardware, dalle librerie necessarie e dalle loro versioni. Questo comporta un grande vantaggio anche nell'uso di linguaggi di programmazione di basso livello quando l'efficienza è fondamentale. Linguaggi come Java, ad esempio, si servono di macchine virtuali per poter garantire la portabilità del codice, ma il tempo di esecuzione aumenta notevolmente.

L'implementazione dei container è stata possibile grazie a due elementi presenti all'interno dei sistemi operativi: namespaces e control groups(cgroups). Si tratta di features che inizialmente non erano state pensate per la virtualizzazione, ma solo nell'ottica di avere dei processi isolati. Non si è quindi arrivati subito alla definizione di container, ma ci sono state delle tappe risolutive di diversi problemi che gli sviluppatori avevano

all'inizio degli anni 2000. I namespaces permettevano la gestione della visibilità relativa alle risorse per ogni processo, mentre i cgroups servivano a limitare l'utilizzo di quanto allocato. Nel secondo caso, esistevano già dei meccanismi di questo tipo all'interno del kernel Linux (e.g. cpulimit, nice), ma non erano centralizzati e potevano essere applicati solo a singoli processi. L'evoluzione dei cgroups è fondamentalmente quella di estendere il tutto ad un gruppo di task in maniera modulare, con la possibilità di definire delle gerarchie nella gestione delle risorse. Le tecnologie basate su container come Docker hanno integrato questi meccanismi nel loro layer di virtualizzazione, in modo tale da fornire delle primitive di alto livello per la gestione dei container, che altrimenti con le precedenti astrazioni sarebbero accessibili solo ad utenti esperti, ed in ogni caso inclini ad errori a causa della complessità di utilizzo.

Docker presenta infine una particolarità a livello di filesystem: rispetto alle macchine virtuali tradizionali o container LXC, che sfruttano un filesystem in senso stretto, questa tecnologia utilizza il concetto di Union Filesystem. Si tratta di un sistema che combina diversi filesystem restituendo logicamente un'unica struttura. Nell'implementazione di Docker, lo Union Filesystem si presenta come una gerarchia di strati, tutti in sola lettura, a ne cui viene aggiunto uno scrivibile al lancio del container. Questo meccanismo permette la condivisione della stessa immagine di base, evitando allo stesso tempo di dover duplicare lo spazio di memoria e che modifiche da parte di una macchina possano influenzare ciò che vedono gli altri container basati sulla stessa istanza, attuando così un primo livello di isolamento (quello più generale con l'host viene sempre implementato attraverso i namespaces).

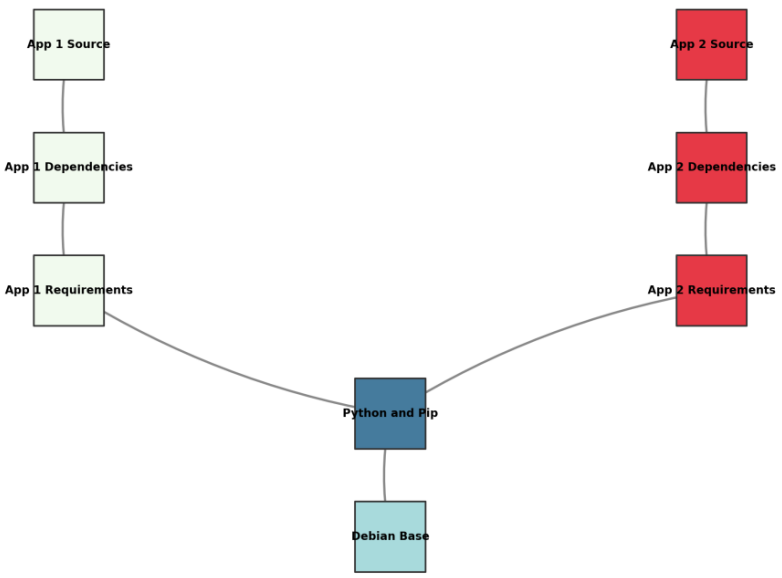

Figura 3.2: Esempio di condivisione di layers in Docker. Le due immagini risultanti hanno la stesso sistema di base, condividendo lo stesso ambiente (Debian) e runtime (Python).

La realizzazione pratica di questa funzionalità è possibile grazie al concetto di copy-on-write, già utilizzato nella creazione dei processi. In questo caso il funzionamento è analogo, ma si applica a livello di memoria secondaria. Ogni modifica all'immagine di base, attraverso il layer in scrittura di più alto livello, è ciò che viene salvato. Non si apportano modifiche all'intera memoria: in pratica è un meccanismo basato sulle differenze. Una volta che si esegue il commit di un'immagine, gli strati vengono congelati e ne formano una nuova. Al livello dell'utilizzatore questo sistema si rivela molto utile perché si possono comporre nuove immagini impilando nuovi layer a quelli già esistenti, senza preoccuparsi dei livelli sottostanti, perché grazie alla portabilità il loro funzionamento è garantito. La composizione non è l'unica proprietà interessante che emerge da questa architettura, ma anche modularità, riutilizzo ed efficienza. Quest'ultima è importante sia in termini di spazio, perché come già visto i layer inferiori vengono condivisi, sia in termini di tempo. Quando si costruisce una nuova immagine, infatti, verranno aggiunti o scaricati solo gli strati di più alto livello per la generazione della nuova immagine, invece di dover ri-eseguire l'operazione ogni volta.

3.2 MQTT

Message Queue Telemetry Transport (MQTT) è un protocollo orientato ai messaggi, basato su TCP, il cui obiettivo è quello di fornire un protocollo leggero, per dispositivi con risorse limitate e connessioni instabili. Il formato dei dati infatti è semplice ed è pensato per essere affidabile e fault-tolerant. Proprio per le sue caratteristiche, con il tempo è diventanto uno standard nelle le applicazioni IoT, ad esempio monitoraggio remoto e raccolta di dati in industria, domotica e ambiente. MQTT si basa su un'architettura publish-subscribe. Ciò permette di disaccoppiare l'invio dei messaggi tra produttori e consumatori, attraverso un broker. Si tratta di un metodo di comunicazione indiretto, in quanto per chi pubblica non è necessario conoscere la destinazione, come invece avviene nelle architetture client-server. Il funzionamento è analogo a quello delle newsletter. Quando un utente è interessato ad un certo argomento, lascia il proprio indirizzo email, in modo tale da ricevere aggiornamenti. Analogamente in un sistema publish-subscribe le entità che devono ricevere determinati messaggi da un publisher, si iscrivono ad un canale.

Oltre al disaccoppiamento, questo tipo di pattern fornisce ulteriori vantaggi. E' possibile aggiungere nuovi publisher o subscriber senza dover modificare la logica esistente, il che rende l'architettura facilmente scalabile. Si tratta di uno dei motivi per cui il cloud riesce a gestire facilmente milioni di messaggi in un breve intervallo di tempo ed avere migliaia di publishers. Inoltre ogni parte del sistema non dipende dalle altre, favorendo

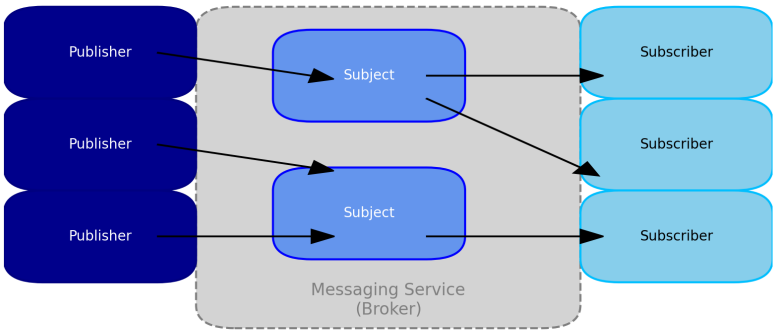

Figura 3.3: Schema architettura publish-subscriber.

la comunicazione asincrona, poiché nessun publisher deve aspettare per l'invio di nuovi messaggi prima che i consumatori finiscano. Allo stesso modo i consumatori non devono attendere che altri ricevano i messaggi per poter proseguire con l'elaborazione delle proprie informazioni. Questo tipo di pattern viene implementato in tanti sistemi come quelli delle notifiche push nelle applicazioni mobili, nei servizi gestiti del cloud ed infine nei sistemi IoT, attraverso il protocollo MQTT, focus dei paragrafi successivi. In questo contesto, le unità di interesse vengono descritte come topic. Essi sono univocamente definiti all'interno dei broker attraverso una struttura gerarchica che ricorda i path nelle URI. Ad esempio, un topic può essere dichiarato come: `factory/floor/sensor1`. Non esiste un meccanismo esplicito di creazione dei topic, ma vengono definiti contestualmente alla pubblicazione di un messaggio.

Il broker ha la responsabilità di connettere, se autorizzati, i diversi client della rete, di tenere traccia dello stato di connessioni precedenti e di filtrare i dati in arrivo dai publisher. Per il discorso di efficienza previsto da questo protocollo, infatti, devono essere distribuiti solo i pacchetti necessari al funzionamento del sistema, e quindi la capacità di filtro è un elemento fondamentale. In particolare, esistono diverse tipologie di controllo dei pacchetti in transito. Quello più importante è il topic filter, dove un messaggio deve contenere necessariamente un topic nell'header, e sarà compito del broker ritrasmetterlo o ignorarlo in base a come è stata implementata la sua logica. In generale i filtri vengono usati per riferirsi ad un insieme di topic, sfruttando due tipologie di wildcards. Si utilizza il segno "+" per indicare un intero livello nella gerarchia del path definito dal topic. Ad esempio `factory/+/sensor1` serve per tutti i sensor1 nei diversi piani di una fabbrica. Il secondo simbolo, cioè "#" si riferisce ad una intera sottogerarchia nel path. `factory/#` e `factory/first/#` filtrano rispettivamente tutti i messaggi provenienti dalla fabbrica e quelli provenienti dal primo piano di un complesso industriale.

MQTT è un protocollo progettato con l'affidabilità in mente: in base al contesto (hardware, banda, latenza) e al tipo di applicazione, fornisce diversi livelli di Quality of Service(QoS):

- **QoS0**: modalità di funzionamento default in MQTT, se non specificato. I messaggi vengono inviati con una logica fire-and-forget, senza la garanzia che il processo vada a buon fine. Il publisher o il broker inviano i messaggi senza alcun tipo di risposta da parte di un subscriber. Non ci sono ulteriori garanzie oltre a quelle fornite dal protocollo TCP.
- **QoS1**: garantisce che il messaggio venga inviato almeno una volta. Questo meccanismo si basa su messaggi di acknowledgment provenienti dal ricevente, per cui se non arriva nessuna risposta, il messaggio deve essere ritrasmesso.
- **QoS2**: si tratta del livello di qualità più alto. Il messaggio viene inviato una sola volta, grazie ad una serie di messaggi di controllo generati dalle entità in gioco. Può essere visto come un acknowledgement incrociato tra publisher e ricevente. Quello che avviene in più rispetto a QoS1 è l'invio dei pacchetti publish-realease per indicare il rilascio del messaggio dalla memoria del publisher, e publish-complete da parte del ricevente per chiudere la comunicazione. In questo modo alla fine del processo il messaggio è stato inviato e gli attori in gioco sono totalmente sincronizzati dal punto di vista della ricezione.

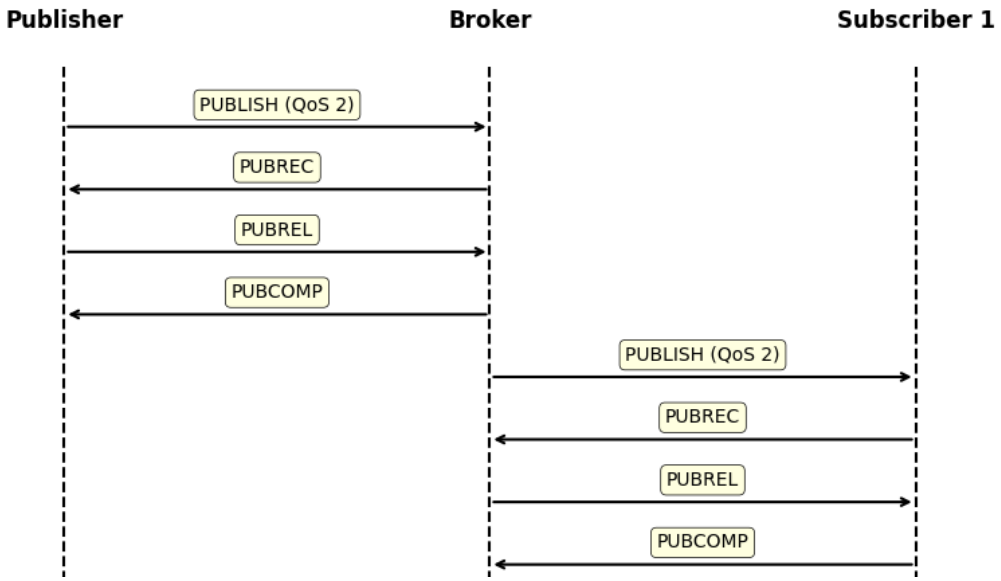

Figura 3.4: Flusso pacchetti di controllo per QoS2. I livelli più semplici funzionano in maniera simile, con un numero di pacchetti inferiore. Si può notare come il disacoppiamento fornito dal broker faciliti lo scambio dei messaggi.

Il QoS viene negoziato al momento della connessione con il broker, ed in base alle sue capacità può essere dinamicamente modificato in un secondo momento. Nell'affrontare il concetto di qualità di servizio, si è visto quindi il funzionamento della publish da parte di un client, ma esistono altri tipi di operazioni in questo pattern, cioè subscribe, unsubscribe e ping. In particolare quest'ultimo risulta fondamentale per l'implementazione del Last Will and Testament(LWT), un metodo per comunicare la disconnessione di un publisher all'interno della topologia. LWT è un messaggio contente il topic, il QoS e un payload da inoltrare a tutti i subscriber che fino a quel momento ricevevano le notifiche. Questo pacchetto viene fornito nel momento in cui un client si connette per la prima volta al broker, prima che inizi l'invio effettivo dei dati per gli altri client. Se attraverso delle operazioni di ping, impostate con un certo timer, una delle due parti non riceve risposta, allora si assume esserci stata una disconnessione. A quel punto, se è il client ad essere offline, il broker invierà il Will ai subscribers.

Un'ultima caratteristica importante per il funzionamento di MQTT, è la persistenza. Il broker mantiene le informazioni di sessione per i client disconnessi, in modo tale da riottenere facilmente le iscrizioni ai topic e i messaggi in coda che non avevano ricevuto (se si tratta di QoS1 e QoS2, pensati come già visto per la trasmissione affidabile). Inoltre, i publisher possono impostare un flag RETAINED nei messaggi, in modo tale che venga salvati in maniera indefinita nel broker. L'obiettivo di questa funzionalità è garantire l'invio di un determinato messaggio ad un client indipendetemente dal momento della sua iscrizione al topic.

3.3 RTSP

3.4 Amazon Web Services

3.4.1 Infrastructure as Code

3.4.2 Serverless

3.4.3 Iam Policy

3.4.4 Kinesis Video Streams

3.4.5 AWS Iot Core e Greengrass

3.5 Apache Flink

3.6 GStreamer

Capitolo 4

Implementazione del sistema

Appendice A

Codice

Appendice B

Matematica Reti Convoluzionali

Equivalenza algebrica tra perceptron e filtro convoluzionale su una regione dell'immagine

Un **perceptron** è un modello neurale che calcola un output y basato su un insieme di input $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, pesi associati $\mathbf{w} = [w_1, w_2, \ldots, w_n]$, e un bias b. La formula del perceptron è:

$$y = \phi\left(\sum_{j=1}^{n} w_j x_j + b\right)$$

Dove:

- ϕ è la funzione di attivazione (ad esempio, ReLU, Sigmoide, Step Function).
- $\sum_{j=1}^{n} w_j x_j + b$ è la somma pesata degli input più il bias.

Convoluzione su una Singola Regione dell'Immagine

Consideriamo una regione $\mathbf{R}_i$ dell'immagine di dimensioni $M \times N$ e un filtro $\mathbf{K}$ (o kernel) di dimensioni $M \times N$. L'operazione di **convoluzione** su questa regione è definita come:

$$S_i = \sum_{m=1}^{M} \sum_{n=1}^{N} K(m,n) \cdot R_i(m,n) + b$$

Dove:

- S_i è il risultato della convoluzione prima dell'applicazione della funzione di attivazione.
- $K(m,n)$ sono i pesi del filtro.
- $R_i(m,n)$ sono i pixel della regione $\mathbf{R}_i$ dell'immagine.

- b è il bias.

Espansione Completa della Sommatoria

Espandiamo la sommatoria per una specifica regione $\mathbf{R}_i$ di dimensioni 3×3:

$$\begin{aligned} S_i = K(1,1) \cdot R_i(1,1) + K(1,2) \cdot R_i(1,2) + K(1,3) \cdot R_i(1,3) \\ + K(2,1) \cdot R_i(2,1) + K(2,2) \cdot R_i(2,2) + K(2,3) \cdot R_i(2,3) \\ + K(3,1) \cdot R_i(3,1) + K(3,2) \cdot R_i(3,2) + K(3,3) \cdot R_i(3,3) + b \end{aligned}$$

Rappresentazione come Prodotto Scalare

Possiamo rappresentare questa operazione come un **prodotto scalare** tra due vettori appiattiti: uno che contiene i pixel della regione $\mathbf{R}_i$ e l'altro che contiene i pesi del filtro $\mathbf{K}$.

Definiamo i vettori appiattiti:

$$\mathbf{x}_i = \begin{bmatrix} R_i(1,1) \\ R_i(1,2) \\ R_i(1,3) \\ R_i(2,1) \\ R_i(2,2) \\ R_i(2,3) \\ R_i(3,1) \\ R_i(3,2) \\ R_i(3,3) \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} K(1,1) \\ K(1,2) \\ K(1,3) \\ K(2,1) \\ K(2,2) \\ K(2,3) \\ K(3,1) \\ K(3,2) \\ K(3,3) \end{bmatrix}$$

Il prodotto scalare tra $\mathbf{w}$ e $\mathbf{x}_i$ è:

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i = K(1,1)R_i(1,1) + K(1,2)R_i(1,2) + K(1,3)R_i(1,3) \\ + K(2,1)R_i(2,1) + K(2,2)R_i(2,2) + K(2,3)R_i(2,3) \\ + K(3,1)R_i(3,1) + K(3,2)R_i(3,2) + K(3,3)R_i(3,3) \end{aligned}$$

Quindi, possiamo riscrivere S_i come:

$$S_i = \mathbf{w} \cdot \mathbf{x}_i + b$$

Convoluzione con Funzione di Attivazione

Dopo aver calcolato S_i, applichiamo una **funzione di attivazione** ϕ per ottenere l'output y_i:

$$y_i = \phi(S_i) = \phi\left(\mathbf{w} \cdot \mathbf{x}_i + b\right)$$

Equivalenza con la Formula del Perceptron

La formula del **perceptron** è data da:

$$y = \phi\left(\sum_{j=1}^{n} w_j x_j + b\right)$$

Dove:

- $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ sono gli input.
- $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ sono i pesi.
- b è il bias.
- ϕ è la funzione di attivazione.

Confrontando le due formule, vediamo che:

$$y_i = \phi\left(\mathbf{w} \cdot \mathbf{x}_i + b\right) = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \phi\left(\sum_{j=1}^{n} w_j x_j + b\right)$$

Bibliografia

[1] I. N. per l'Assicurazione contro gli Infortuni sul Lavoro (INAIL), "Rapporto annuale inail 2023," 2023, accesso: 24 settembre 2024. [Online]. Available: https://www.inail.it/content/dam/inail-hub-site/documenti/2023/09/infografiche-relazione-annuale-inail-2022.pdf

[2] O. Safety and H. A. E. Union, "Il valore della sicurezza e della salute sul lavoro e i costi sociali degli infortuni e delle malattie professionali," 2019, accesso: 26 settembre 2024. [Online]. Available: https://osha.europa.eu/sites/default/files/Summary_Value_of_OSH_and_societal_costs_injuries_and_diseases_IT.pdf

[3] G. U. della Repubblica Italiana, "Decreto legislativo 81/2008," 2008, accesso: 18 ottobre 2024. [Online]. Available: https://www.gazzettaufficiale.it/eli/id/2008/04/30/008G0104/sg

[4] G. U. dell' Unione Europea, "Regolamento (ue) 2016/42," 2016, accesso: 22 ottobre 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016R0425

[5] J. Manyika, L. Woetzel, and R. Dobbs. (2015) Unlocking the potential of the internet of things. Accesso: 24 ottobre 2024. [Online]. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world