



UNIVERSITA' DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

Corso di Laurea Magistrale in Informatica
Data Science & Machine Learning

Tesi di Laurea

**Una metodologia di Explainable AI per la
diagnosi di malattie da Tomografia Ottica
Computerizzata (OCT)**

Relatori

Ch.mo Prof. Deufemia Vincenzo
Dott. Stefano Cirillo
Dott. Gaetano Cimino

Candidato

Pierluigi Liguori

Matricola

0522500908

ANNO ACCADEMICO 2021/2022

Indice

1	Introduzione	6
2	Stato dell' arte	9
2.1	Tecniche di Medical Image Classification	9
2.2	XAI nella Medical Image Analysis	11
2.3	Studi relativi all'oftalmologia	12
2.4	Multi-task Learning	14
3	Medicina 4.0: L'Intelligenza Artificiale nella Sanità	16
3.1	Introduzione	16
3.2	L'utilizzo dell'AI nella diagnostica	18
3.3	Clinical Decision Support System	20
3.4	Malattie rilevabili dal fondo oculare	21
3.5	Intelligenza artificiale e oftalmologia	23
4	Identificazione di malattie da immagini di Tomografia Ottica Computerizzata (OCT)	26
4.1	Dominio di applicazione	27
4.1.1	Malattie studiate	28
4.2	Reti convoluzionali per Image Classification	30
4.2.1	Rete Neurale	30
4.2.2	Rete Neurale Convoluzionale	31
4.2.3	Transfer Learning	33
4.3	Image Captioning	35
4.3.1	Natural Language Processing	35
4.3.2	Architettura Encoder-Decoder	38
4.4	Transformer	40
4.5	Visual Explainability	42
4.5.1	Class Activation Maps	42

4.5.2	Grad-CAM	45
4.6	Learning congiunto	46
5	Estrazione e lavorazione dei dati	48
5.1	Panoramica	48
5.2	Dataset per l' Image Classification	49
5.2.1	Reperimento immagini	49
5.2.2	Preprocessing delle immagini	49
5.3	Data Preparation per il Captioning	52
5.3.1	Biomarkers	52
5.3.2	Creazione del trainset	54
5.3.3	Data augmentation su NLP	55
5.4	Dataset per GPT-2	56
6	Soluzione proposta	58
6.1	Modelli utilizzati	59
6.1.1	DenseNet	59
6.1.2	VGG	60
6.1.3	GPT-2 e GPT-Neo	60
6.2	Sviluppi e criticità affrontate	61
6.3	OCT Report Tool	64
6.3.1	Architettura soluzione finale	64
6.3.2	Implementazione modelli	66
6.3.3	Realizzazione della Web Application	66
6.3.4	Caso d' uso del Tool	67
7	Valutazione Sperimentale	68
7.1	Metriche per l' Image classification	68
7.2	Metriche per l' Image Captioning	70
7.3	Valutazione delle soluzioni proposte	73
7.3.1	Modelli con Visual Explainability	73
7.3.2	Modelli per Image Captioning	77
7.3.3	Modello Multi-Tasking	78
7.3.4	Generazione Report	80
8	Conclusioni e sviluppi futuri	81

Elenco delle figure

2.1	Analisi delle zone tumorali evidenziate da Grad-CAM	13
2.2	Architettura multi-task per classificazione e captioning.	15
3.1	Concettualizzazione dell'eHealth nei i suoi punti cardine	18
3.2	Modelli di intelligenza artificiale per la diagnostica	19
3.3	Architettura di un Clinical Decision Support System.	21
3.4	Retinopatia diabetica su fondo oculare ed OCT.	23
3.5	Screenshot del Software RetMarker.	25
4.1	Differenti tipologie di scan in OCT.	28
4.2	Sezioni della macula.	29
4.3	OCT Scan-B con i pattern tipici di drusen, neovascolarizzazione coroideale e edema maculare diabetico.	29
4.4	Rete neurale feed forward.	31
4.5	Filtro convoluzionale.	32
4.6	Architettura di una Rete Neurale Convolutzionale.	34
4.7	Architettura di una Rete Neurale Convolutzionale.	34
4.8	Language Modeling basato su Rete feed-forward.	37
4.9	Retroazione di una Rete Neurale Ricorrente.	38
4.10	Struttura di una cella LSTM	39
4.11	Architettura CNN-LSTM per il Task di Captioning	40
4.12	Architettura di un Trasformer	41
4.13	Confronto visivo delle heatmap per i diversi algoritmi di Visual XAI per la detection di una coccinella.	43
4.14	Esempio di Class Activation Maps	44
4.15	Architettura Grad-CAM	46
4.16	Condivisione dei layer tra diversi compiti in un' architettura Multi-Task.	47
5.1	Rappresentazione visiva della tecnica PCA	51

5.2	Da sinistra verso destra: reticular pseudodrusen, soft drusen, hard drusen.	53
5.3	Patterns del Diabetic Macular Edema.	54
5.4	Dataframe per l' allenamento congiunto.	55
5.5	Esempio di dataset per training GPT-2	57
6.1	Architettura di una DenseNet	59
6.2	Layers di una VGG19	60
6.3	Diverse implementazioni di GPT-2	61
6.4	Architettura della soluzione finale.	65
6.5	Output del modello proposto su un OCT con DRUSEN.	67
7.1	Esempio matrice di confusione	69
7.2	Curva ROC	70
7.3	Matrice di confusione e curva ROC per Densenet-161	73
7.4	Matrice di confusione e curva ROC per Densenet-121	74
7.5	Matrice di confusione e curva ROC per VGG19	74
7.6	Gradcam a confronto per DRUSEN	75
7.7	Gradcam a confronto per DME	75
7.8	Gradcam a confronto per CNV	76
7.9	Grad-CAM calcolato sulla classe DME (dx) e sulla classe CNV (sx)	76
7.10	Matrici di confusione per modello Multi-Task	78
7.11	Curva ROC per modello Multi-Task	79
7.12	Confronto Report generato da GPT2 e GPT-Neo	80

Abstract

Sempre più specialisti del settore sanitario fanno uso di sistemi di Intelligenza Artificiale (AI) per un supporto nel processo diagnostico: recenti modelli di Deep Learning sono in grado di classificare immagini biomediche con un'alta accuratezza. Tuttavia, la maggior parte di essi sono utilizzati con un livello di astrazione molto elevato, ossia senza effettivamente conoscere il criterio con il quale si è arrivati ad una determinata predizione. L'esigenza di una maggior comprensione ha portato alla diffusione di nuovi modelli di Explainable AI (XAI) che sono in grado di dare indizi che giustificano la decisione presa da un classificatore grazie all'ausilio di elementi testuali e visivi. Questo lavoro di tesi implementa tali meccanismi nel settore dell'oftalmologia. In particolare viene affrontato lo studio delle patologie più comuni che colpiscono la retina tramite l'analisi di immagini OCT (Tomografia ottica computerizzata) su malattie dovute alla degenerazione maculare senile, come DRUSEN e neovascolarizzazione coroideale (CNV) e l'edema maculare diabetico (DME). Questa ricerca si concretizza nello sviluppo di un sistema intelligente rivolto sia al supporto medico e sia al paziente, il quale sarà capace di ricevere una prima indicazione sulle proprie condizioni di salute. L'implementazione fa uso di reti convoluzionali per il task di classificazione; Densenet-161 raggiunge un accuracy del 92.75% su immagini di dimensione ridotta; Densenet-121 e VGG-19, invece, sono state addestrate su immagini OCT con maggior dettaglio per fornire oltre alla classificazione anche una Explainability visiva che consiste in una mappa di calore sovrapposta all'immagine al fine di marcare le aree di maggior criticità; le due reti hanno ottenuto rispettivamente un'accuracy del 91.67% e del 93.06%. Inoltre, si presenta un'architettura CNN-LSTM che genera una breve descrizione della scansioni OCT data in input; quest'ultima viene indirizzata ai Transformers GPT-2 e GPT-Neo che generano del contesto per un documento di reporting diagnostico. Tutte le componenti del sistema sono state riadattate in un unico modello che risolve tutti i task congiuntamente raggiungendo il 98% di accuracy, evidenziando come la correlazione tra compiti aumenti le performance.

Capitolo 1

Introduzione

L’Intelligenza Artificiale (AI) sta trasformando il mondo della sanità grazie all’adozione sempre più frequente di modelli di Machine Learning, Computer Vision e di Natural Language Processing (NLP). L’AI non solo offre un supporto nella diagnosi tramite Clinical Decision Support System (CDSS) ma automatizza anche compiti amministrativi del sistema sanitario che coinvolgono i sistemi informativi, come l’aggiornamento delle cartelle cliniche elettroniche, gestione del workflow dell’assistenza infermieristica e prenotazione di esami clinici in funzione dello stato di salute dei pazienti. Tuttavia, delegare il processo diagnostico solamente a sistemi intelligenti senza la consultazione di medici specialisti è un errore.

In passato, le prime soluzioni di AI si limitavano ad elaborare fonti di dati da cui trarre modelli per ottenere predizioni: su un dataset di immagini di scansioni al torace tramite tomografia computerizzata si poteva fare una stima sulle probabilità di avere un cancro ai polmoni, la quale impatta drasticamente sull’operato dei medici; purtroppo algoritmi troppo complessi hanno un grande svantaggio, ossia non forniscono le motivazioni che hanno portato ad una determinata valutazione, mancando di trasparenza. Negli ultimi anni sono state affinate nuove tecniche di Explainable AI (XAI) in grado di aiutare la comprensione dei comportamenti dei modelli di Deep Learning, requisito fondamentale per l’uso in ambito healthcare.

La XAI non solo ne aumenta la fiducia e l’adozione, ma le spiegazioni possono essere usate per migliorare lo sviluppo di nuovi modelli. Tale innovatività conduce la sanità verso un nuovo tipo di assistenza che verte sulla risposta immediata grazie a diagnosi supportate dall’AI e allo sviluppo di nuove tecnologie Cloud e dispositivi dell’Internet of Things (IoT). Questo progresso è già in corso in diversi Paesi, ma la pressione sui servizi sanitari è ancora molto alta, la quale

porta al rinvio di esami, interventi e terapie. Questa incapacità nell'assistere il paziente è riscontrata anche dai Paesi più sviluppati, non necessariamente per il mancato personale medico e strutture specializzate, ma anche per il tenore di vita dell'uomo moderno; la vita media si è decisamente allungata nell' ultimo secolo, per questo motivo l'incidenza di patologie è drasticamente aumentata, portando ad un maggior numero di casi da trattare.

Per quanto concerne l'ambito oftalmico (su cui si concentra questo elaborato), intorno ai quarant'anni è comune l'insorgere della presbiopia; dai sessanta anni in poi i controlli periodici dovrebbero avere una cadenza più frequente poiché si tende ad una degenerazione maculare dovuta all'età. Oltre la longevità, la tendenza degli ultimi tempi a consumare un alto quantitativo aumenta il rischio del diabete di tipo 2 che, oltre all'obesità e malattie cardiovascolari, può portare alla retinopatia diabetica che è tra le prime cause di cecità. Queste patologie, assieme al glaucoma, rientrano nello spettro di malattie individuabili con un esame OCT (tomografia ottica a radiazione coerente).

Dall'imaging OCT è possibile individuare particolari marker biologici a livello retinico non rilevabili con un classico esame del fondo oculare; per tale ragione è in crescita lo sviluppo di modelli di Deep Learning per questo tipo di scansioni oculari. Tra questi, solo una esigua percentuale di studi si esprime sulla spiegabilità dei risultati fornendo cenni di una didascalia generata che accompagna l'immagine ma che non fa le veci di un Clinical Decision Support System; inoltre la maggior parte dei lavori di visual XAI in oftalmologia non sono modelli addestrati su immagini OCT ed i pochi elaborati non trattano congiuntamente lo spettro delle malattie retiniche più comuni appena discusse.

Questo studio mira allo sviluppo di un sistema capace di diagnosticare le malattie della retina da una scansione OCT, evidenziando nell'immagine le aree che hanno portato ad una precisa classificazione; la spiegabilità viene ulteriormente ampliata grazie all'autogenerazione di un report che può essere letto sia dal medico, dando un supporto clinico, oppure dal paziente, in modo da ricevere un primo feedback istantaneo. L'intero modello è reso usabile grazie ad una Web Application che rende il facile accesso al sistema con il semplice caricamento dell'immagine tramite un' interfaccia grafica.

Questo capitolo è seguito da una panoramica dei lavori simili dello stato dell'arte (Capitolo 2), si discute successivamente di Medicina 4.0 e di come ha avuto un riscontro nella società, dei sistemi di supporto e delle malattie della retina; il Capitolo 4 formalizza le problematiche da affrontare, dopo un introduzione dal dominio di applicazione, si descrivono tutte le componenti che hanno contribuito

alla soluzione finale. Il Capitolo 5 è dedicato all'estrazione ed al preprocessing dei dati che costituiscono i dataset per le fasi di addestramento. Nel Capitolo 6 sono mostrate le implementazioni effettive dei modelli e la loro interconnessione; si evidenziano le criticità riscontrate e viene descritta l'applicazione Web. Le valutazioni dei modelli tramite le metriche più comuni sono riportate nel Capitolo 7, vengono inoltre confrontate diverse soluzioni al fine di trovare un giusto trade-off per i compiti stabiliti. Infine, il Capitolo 8 chiude l'elaborato con conclusioni e sviluppi futuri.

Capitolo 2

Stato dell' arte

Nel seguente capitolo si fornisce una panoramica dei lavori esistenti in letteratura sulle attuali tecniche di Image Classification ed Explainable AI in campo medico. L'attenzione viene successivamente ristretta all'oftalmologia, evidenziando le necessità che hanno portato allo sviluppo di questo elaborato. Nella sezione 2.4, inoltre, viene fatto cenno ai moderni approcci del multi-task learning, su cui si basa l'implementazione proposta.

2.1 Tecniche di Medical Image Classification

La moltitudine di immagini mediche è generata da dispositivi biomedici per mezzo di tecniche di imaging, ciascuna tratta delle aree specifiche del corpo umano poiché ogni struttura e tessuto organico risponde in maniera diversa a determinati stimoli. Tra queste ritroviamo la tomografia computerizzata (CT), imaging a risonanza magnetica (RM) e tecnologie ad ultrasuoni. Dopo una fase di acquisizione delle fonti, entra in gioco la Medical Image Analysis che, secondo [29], viene descritta come un susseguirsi di tre task: Feature Extraction, Feature Selection e Image (o più genericamente "Feature") Classification. L'estrazione delle caratteristiche può essere di tipo statistico a livello di pixel (SPL) o in funzione di determinate forme confinate nell' immagine (Shape Feature), soffermandosi su circolarità, compatezza.

La Feature Selection è un altro step importante per selezionare le features più appropriate per il task di classificazione e se combinato con tecniche di riduzione della dimensionalità e clustering è possibile migliorare le performance trovando dei pattern nei dati di training. Sempre in [29], vengono illustrate le principali metodologie per la Medical Image Classification rimarcando per ciascuna i pro e i contro, le tipologie di imaging associate e i risultati ottenuti in letteratura.

Tra questi ritroviamo di nuovo metodi statistici: si suddividono in approcci di tipo non supervisionato, dove vengono raggruppati i dati in base alla loro separazione nello spazio delle features, come il K-means e Fuzzy Clustering; nei metodi supervisionati rientrano metodi probabilistci come il K-Nearest Neighbors (k-NN) e classificatori Bayesiani, i quali hanno bisogno di dati di test e di training appositamente etichettati con una label.

In [37] viene proposto un classificatore KNN su immagini CT polmonari in grado di raggiungere un accuracy alta soltanto su un dataset limitato di immagini; in [42], un' implementazione basata su Fuzzy C-means non raggiunge la soglia dell' 80% di sensitivity su immagini di risonanza magnetica di tipo cerebrale. Si precisa come i modelli statistici si applicano maggiormente in questi due tecniche di imaging (Tomografia computerizzata e RM). Più versatilità è garantita invece su classificazioni di tipo Support Vector Machine (SVM), potenzialmente in grado di classificare tutte le modalità di imaging. L'obiettivo di SVM è trovare un iperpiano che separa nel miglior modo possibile le istanze di classe diverse per poi fare previsioni future su nuovi dati.

Lo studio presente in [28] è un esempio di applicazione in SVM per l' identificazione di processi di angiogenesi tumorale ed è stato testato su immagini provenienti da normale una fotocamera digitale attaccata al microscopio ottenendo un F-score dello 0.90; anche in [33] viene implementato SVM per immagini ad ultrasuoni per la classificazione di lesioni al fegato ottenendo una buona accuratezza ma allenando diversi classificatori soltanto su due delle N classi previste.

Un' ulteriore metodologia versatile è l'uso di reti neurali che statisticamente portano a migliori performance per l' Image Classification rispetto ai modelli visti in precedenza, soprattutto se si ha a disposizione una maggiore quantità di dati. Come spiegato in [6], il Deep Learning (DL) ha avuto un' adozione esponenzialmente crescente negli ultimi anni nel campo medico dovuto all' incremento del numero di immagini prodotte dai sistemi moderni.

Alcuni esempi sono: lo studio condotto da [7] per la classificazione del cancro al seno tramite l'uso di reti convoluzionali (CNN) su un dataset di 8000 immagini raggiungendo il 97.5% di accuratezza; in [1] le reti AlexNet and GoogLeNet con il 99% riescono ad identificare la tubercolosi nelle radiografie toraciche; gli ultimi elaborati prodotti per quanto concerne la diagnosi del SARS-CoV-2 hanno raggiunto il 96% tramite grazie alle reti EfficientNetB1, NasNetMobile e MobileNetV2.

È un dato di fatto come il Deep Learning abbia cambiato le carte in tavola nelle modalità di fare ricerca, d'altro canto sarebbe un errore condurre un

uso massivo di tal paradigma in ogni problema. Esiste un enorme numero di articoli che dimostrano come algoritmi basati su SVM e Naive Bayes risultano più efficienti, questo perché i modelli DL hanno una miriade di pesi che devono essere "calibrati" con i dati: se il numero di pesi è maggiore o quasi uguale al numero di samples di allenamento c'è il rischio di cadere in overfitting. Inoltre, la potenza computazionale richiesta per alcune reti non sempre è sufficiente per l' allenamento ed è preferibile optare per un apprendimento tradizionale.

2.2 XAI nella Medical Image Analysis

I sistemi deep learning-based citati in precedenza a causa della loro struttura complessa e non lineare vengono utilizzati come una "black box", ossia non sono fornite informazioni sul processo decisionale che ha portato ad una determinata predizione. Come si può ben intuire, la mancata trasparenza di questi modelli è un notevole svantaggio, specialmente in ambito medico, dove una decisione non può essere condotta da un algoritmo di intelligenza artificiale senza spiegarne il motivo (un approfondimento sul processo decisionale nell' eHealth è presente nel Capitolo 3).

In [39] si presentano le principali tecniche di Explainable AI (XAI) utilizzate nell' analisi di immagini mediche che vengono categorizzate in 3 tipi: visive, testuali e basate su samples; questo elaborato si focalizza sulle prime due. Nelle spiegazioni visuali, vengono definite delle "Saliency map" che mostrano le parti più importanti nell' immagine che sono state determinanti per una predizione e usano approcci backpropagation-based o perturbation-based.

Nei primi ritroviamo Class Activation Mapping (CAM) introdotta in [46], dove viene spiegato come rimpiazzando i layers pienamente connessi alla fine di un blocco convoluzionale con un layer di tipo Global Average Pooling (GAP) si ottiene una mappa di attivazione dalla somma dei pesi che fanno riferimento a pattern visivi in precise posizioni spaziali (Capitolo 4). Un esempio di applicazione è presente in [22], gli autori hanno ricavato CAMs dall' output di un ensamble di VGG-16, Resnet-50, Inception-V3 e Inception-Resnet-V2 ai fini di rilevare emorragia cranica acuta.

Una generalizzazione di CAM è Gradient-weighted Class Activation Mapping (Grad-CAM) che va bene per qualsiasi CNN e dà la possibilità di ottenere diverse mappe delle features per ogni strato convoluzionale della rete; uno studio in [44] mostra come con Grad-CAM si riesce a spiegare come il classificatore individua i tumori al cervello evidenziando le zone critiche (Figura 2.1). Il secondo approccio

visivo è il perturbation-based, il quale permette di esplorare una rete neurale "perturbando" gli input osservando i cambiamenti dell' output. L' occlusion sensitivity è un esempio di tal approccio, dove viene effettuata un' occlusione parziale di un'immagine di input e successivamente la verifica della classe di appartenenza, se il classificatore cambia la classe vuol dire che l' area appena occlusa è incisiva per la predizione.

Un altro metodo comune che rientra in questa categoria è la Local Interpretable Model-agnostic Explanations (LIME), spesso usato per effettuare diagnosi nell' eHealth. LIME sostituisce un modello complesso con uno più semplice (ad esempio una CNN viene approssimata ad un modello lineare) che viene usato per mappare i dati di input perturbati e il relativo output variato, la somiglianza tra l' input perturbato con quello originale viene usato come un peso per quantificare l' explainability della predizione.

Un' applicazione la si può trovare in [26], in cui LIME è usato per spiegare nelle immagini endoscopiche quali aree rappresentavano regioni contenenti sangue. Un' altra forma di XAI è quella testuale, facendo riferimento a didascalie per le immagini (Image Captioning) fino a referti medici. Un' architettura ampiamente usata per la textual explanation è la concatenazione di un encoder convoluzionale (CNN) che codifica l' immagine e restituisce caratteristiche visive, quest' ultime vengono date come input ad un decoder, rappresentato da Rete Neurale Ricorrente (RNN) o da una Long Short-Term Memory (LSTM) che genera le parole che compongono il testo (Capitolo 4); un uso CNN-LSTM per l' identificazione di lesioni nelle mammografie è descritto in [38]. Testo generato più assimilabile ad un vero e proprio report è Medical-VLBERT [24], uno strumento che identifica anomalie dovute al COVID-19 da scansioni polmonari, generando un report basato sulle regioni danneggiate raggiungendo ottimi risultati.

2.3 Studi relativi all'oftalmologia

Per quanto concerne il settore dell' oftalmologia, l' esplosione del Deep Learning ha dato un contributo significativo nella diagnosi di diverse malattie oculari, merito anche della condivisione di enormi dataset di immagini oculari da parte di specialisti. I formati di imaging più utilizzati sono la scansione del fondo oculare (EFI), l' optical coherence tomography (OCT) e, di recente, Optical Coherence Tomography-Angiography (OCTA) in grado di visualizzare anche le reti vascolari all' interno della retina. Sebbene in questo campo esistano diversi studi in letteratura riguardanti l' Image Classification, Captioning e XAI, la

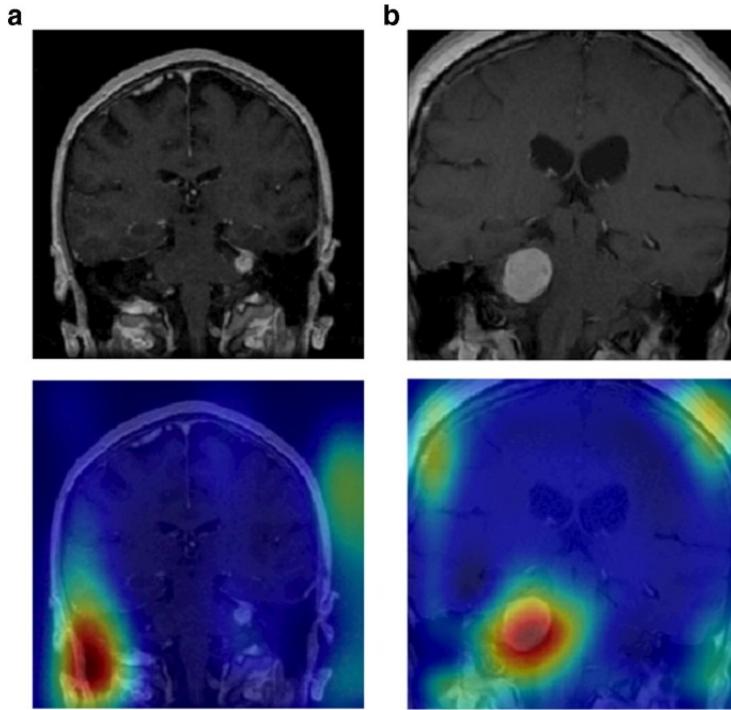


Figura 2.1: Analisi delle zone tumorali evidenziate da Grad-CAM

maggioranza sono svolti nelle immagini EFI. Un esempio è il lavoro svolto in [5], dove viene presentato un classificatore basato su reti neurali allenato su immagini EFI di glaucoma e retinopatia diabetica che raggiunge il 99,89% di accuracy anche se limitato a 2 classi.

Come si è già accennato in precedenza, anche nell' oftalmologia riscontriamo ottime performance con modelli che non rientrano nel DL: [30] apre le porte per un sistema di screening automatizzato per la detection di retinopatia dal fondo oculare, la ROC curve raggiunge 0.9968 con accuratezza del 97,4% utilizzando un SVM. Traguardi ugualmente eccelsi in contesto OCT sono riportati in [19]: 4 classificatori CNN binari, sono rispettivamente allenati per riconoscere DRUSEN-OTHER, NORMAL-OTHER, CNV-OTHER, DME-OTHER, sono sulla soglia dello 0.987 di accuracy, 0.987 sensitivity, e 0.996 specificity. Se ci si sposta in contesto XAI, sono presenti alcune proposte in letteratura che vertono su explainability testuale o visiva ma è scarna di studi in cui vengono trattate entrambe sullo stesso problema.

Gli autori di [17] si soffermano sul metodo CAM su un ensamble di Inception-V3, ResNet-152, e Inception-ResNet-V2 per distinguere EFI di soggetti sani da quelli affetti da retinopatia diabetica grave o moderata localizzando visivamente le regioni di interesse; in [10] su immagini OCT si implementa Grad-CAM+ sull'individuazione della miopia, difetto visivo meno trattato nonstante sia il

più diffuso. È da sottolineare come le immagini prodotte Tomografia ottica computerizzata hanno enormi potenzialità, l' oftalmologo tramite questo imaging può rilevare dei minimi cambiamenti della retina che con il semplice fondo oculare passerebbero inosservati. Secondo questo elaborato, condurre più studi su OCT colmerebbe anche il gap rispetto al numero di lavori sui immagini EFI ed è la miglior alternativa di imaging su cui fare ricerca: in alternativa, OCTA ha ancora un numero limitato di dataset per ricercare soluzioni accurate. Questo studio, tenendo conto di tali esigenze, mira alla realizzazione di un sistema di diagnosi su OCT che oltre alla classificazione di malattie retiniche sappia dare una forte e trasparente spiegazione sia visiva che testuale dei risultati raggiunti. Sebbene tentativi di Image Captioning OCT sono presenti in [40] con l' uso di LSTM, le descrizioni generate generalizzano sommariamente elementi delle immagini di test, inoltre, sia questo task che la classificazione (che raggiunge accuracy del 94%) sono svolti su 1750 immagini di training e solamente 8 di test. Questo studio si differenzia da [40] poiché oltre alla semplice caption, viene effettuato un vero e proprio report dalla retina che può essere letto sia dallo specialista e anche dal paziente in esame, similmente al lavoro su scansioni polmonari in [40] e dando anche una componente visiva delle aree di interesse come in [10] ma su malattie come la degenerazione maculare all' età e l' edema maculare diabetico.

2.4 Multi-task Learning

Si parla di Multi-task Learning (o anche Joint learning) quando due modelli vengono allenati su differenti task in modo simultaneo, ottimizzando più di una funzione loss contemporaneamente (Capitolo 6). La risoluzione può portare ad un sistema che sappia generalizzare meglio rispetto ad un addestramento separato dei singoli task: il modello congiunto tende a trovare una soluzione comune per tutti i compiti, più ne sono in allenamento, minore è il rischio di overfitting sui singoli task; ad esempio, un sistema di face detection sarebbe più accurato se dovesse in simultanea discriminare anche per genere. Tale paradigma risulta utile soprattutto quando bisogna allenare un Image Classifier ed un Image Captioner, soprattutto per minimizzare il problema del disagreement tra predizioni discordanti tra loro dei due modelli su input uguali.

In letteratura, il paper di riferimento è [45], il quale offre una soluzione elegante introducendo degli *embeddings multimodali* che sono la rappresentazione generalizzata di caratteristiche sia visive che testuali di un oggetto. Qui, le features linguistiche di un Decoder RNN e le features visive di un Encoder CNN

sono proiettate in uno spazio multimodale su cui il modello può lavorare su una semantica più latente. Questa concettualizzazione è stata applicata per descrivere e distinguere tipologie di fiori da specie di uccelli (Fig. 2.2), dando anche un explainability visiva CAM-based.

Per quanto concerne l' oculistica, [8] applica il Multi-task Learning per 3 compiti su immagini EFI: classificazione tra retinopatia diabetica, degenerazione maculare senile, glaucoma e melanoma; una seconda classificazione più ampia su 320 sub-categorie; generazione di una descrizione testuale.

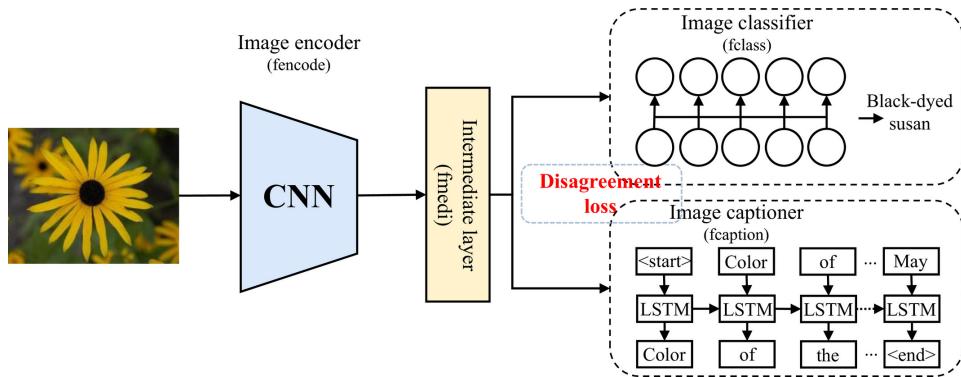


Figura 2.2: Architettura multi-task per classificazione e captioning.

Capitolo 3

Medicina 4.0: L’Intelligenza Artificiale nella Sanità

3.1 Introduzione

La Medicina 4.0 nasce dal turbine di cambiamento che ha portato alla quarta rivoluzione industriale. Punto chiave dell’Industria 4.0 è l’uso di tecnologie innovative che mirano alla collaborazione di tutti gli elementi della produzione, creando un’operazione tra operatore e macchina. Il sistema sanitario, che può essere visto come un’azienda, non ha potuto esimersi da questi nuovi modelli di business. Per questo motivo, nel 2001 viene coniato il termine ”eHealth” da Gunther Eysenbach nel suo paper [12], in cui si cercava di unire industria e sanità con Internet mentre il web 2.0 era ancora in uno stato embrionale. Da quel momento, il termine fu più volte ripreso negli anni, definendolo come un paradigma che unisce trasversalmente informatica, sanità pubblica e imprese e che porta ad un cambiamento nel modo di approcciare l’assistenza sanitaria. Quest’ultima dovrebbe avere un carattere proattivo, mirando alla prevenzione e al continuo monitoraggio della salute del singolo anziché reagire solo dopo l’insorgere della malattia. Una prima trasformazione della sanità è iniziata con la comparsa dei primi siti internet inerenti alla salute pubblica ed al trattamento delle malattie. Successivamente, l’introduzione delle cartelle cliniche elettroniche (EHR) assieme a tecnologie cloud e mobile hanno portato ad una maggiore integrazione dei dati creando una rete sanitaria intelligente.

Altre tecnologie abilitanti della Medicina 4.0 sono riassunte in [2]: l’Internet of Things (IoT) permette la connessione di dispositivi medici in una rete come ad esempio sensori per misurare la frequenza cardiaca, la temperatura corporea, il comportamento del sonno e la pressione sanguigna oppure macchine per raggi

X, dialisi e di diagnostica; per far riferimento a questi sistemi è solito usare anche il termine IoHT, ossia "Internet of Healthcare Things". La maggiore efficienza nello scambio di dati di tali dispositivi è dovuta dalla tecnologia di rete 5G che fornisce funzionalità avanzate come la comunicazione veloce e a bassa latenza, da questo punto di vista essa può essere considerata un altro fattore abilitante dell' eHealth.

Nuovi sistemi cibernetici (detti CPS) hanno facilitato le interazioni tra il mondo software e quello fisico, fornendo un monitoraggio continuo da una parte ed un servizio di assistenza costante dall'altra. I CPS medici utilizzano controlli di feedback incorporati per monitorare e reagire a specifici impulsi. Un esempio sono i dispositivi medici impiantabili, come simulatori cerebrali utilizzati per trattare l'epilessia, pacemaker cardiaci utilizzati per regolare la frequenza cardiaca e altri strumenti utilizzati per gestire i segnali biologici, compresi dispositivi indossabili. Questi sensori di nuova generazione, conducono ad una visione più lungimirante che è riscontrata in [27] in merito all' Healthcare 5.0, uno step successivo che prevede un' assistenza del paziente più pervasiva con l' uso di dispositivi CPS, robotici e nanotecnologie impiantante nel corpo umano per un intervento tempestivo ed immediato sulle patologie. La fattibilità di queste soluzioni sono ancora limitate da fattori economici, tecnologici, etici, di scalabilità e dai rischi per la salute imposti dall' uso di nanosensori e biosensori.

L'Healthcare intelligente del presente, invece, è adottata in diversi paesi già dall' ultimo decennio e si fonda su alcuni pilastri principali. Il primo principio restituisce al paziente il pieno controllo dei suoi dati. Ha il diritto di visualizzare tutti i suoi dati clinici e decidere quali informazioni rendere inaccessibili o meno da enti terzi. Questa filosofia genera fiducia e trasparenza nel sistema e ne facilita l'adozione. Le informazioni condivise, invece, sono utili per i medici che dovranno assistere il malato; per esempio, se un paziente viene trasferito in un altro ospedale, un nuovo medico potrà conoscere tutta la cronologia di tutti i suoi report passati immediatamente. Un altro punto fondamentale è la cura customizzata. I continui monitoraggi delle condizioni di salute generano un'enorme quantità di dati i quali potranno essere analizzati per permettere delle diagnosi ad hoc, senza divagare in assunzioni sommarie. In questo modo è possibile determinare il probabile andamento di salute di un individuo indirizzando ad un chiaro processo decisionale clinico grazie anche all' aiuto di sistemi intelligenti di supporto, argomento che viene approfondito nella Sezione 3.3. Questo capitolo si direzionerà successivamente verso l'ambito oftalmologico, dove verranno descritte le principali malattie rilevate dal fondo oculare e di come l' intelligenza

artificiale ha impattato nella prevenzione con metodi innovativi di Deep Learning e NLP.

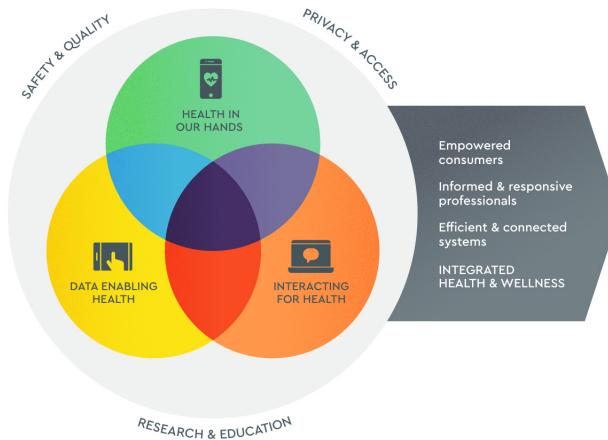


Figura 3.1: Concettualizzazione dell'eHealth nei suoi punti cardine

3.2 L'utilizzo dell'AI nella diagnostica

Nell'eHealth, vi sono alcune attività che richiedono una gran quantità di risorse umane. Grazie all'intelligenza artificiale, la quale è in grado di simulare il ragionamento e l'apprendimento umano, è possibile delegare dei compiti di tipo amministrativo o clinico. Automatizzare queste operazioni significa dare più spazio agli operatori sanitari che possono dedicare più tempo alla cura dei propri pazienti, migliorandone qualità della vita. Questi modelli sono stati raffinati grazie alla crescita del numero di cartelle cliniche digitali, dispositivi di monitoraggio e immagini derivate da risonanze magnetiche, mammografia, tomografia ed altri tipi di imaging. I compiti che riesce ad adempiere l'AI spaziano dalla diagnosi delle malattie, l'identificazione di fattori di rischio alla scoperta di nuovi farmaci. In letteratura, diversi modelli hanno ricoperto un ruolo cruciale per malattie come il cancro, diabete, malattie cardiache, Alzheimer e malattie del fegato. Tra questi, i più utilizzati sono la regressione logistica, logica Fuzzy, SVM ma soprattutto modelli di Deep Learning; una visione di tale contributo è riportata in Figura 3.2.

In [20] vengono raccolti casi di utilizzo di AI per la diagnostica. La malattia più trattata in letteratura nell'ultimo decennio è Alzheimer; diversi studi analizzano l'evoluzione della demenza con l'osservazione delle immagini dei neuroni tramite risonanza magnetica; metodi alternativi diagnosticano la malattia

tia analizzando dati vocali, estraendo caratteristiche dallo spettogramma dell' audio tramite sistemi di Machine Learning [25]. Questo è stato possibile grazie a dispositivi IoT che hanno raccolto dati vocali di 23 persone anziane, etichettati come sani o affetti di demenza per la fase di training. L'AI in questo caso è riuscita a rilevare sottili cambiamenti sonori non percepibili dall'uomo ma che sono campanelli d'allarme per l'Alzheimer. Al secondo posto tra le malattie più analizzate vi è il diabete. Lo studio in [3] utilizza sensori basati su Bluetooth per il monitoraggio di dati in tempo reale da soggetti affetti da diabete, come pressione sanguigna, frequenza cardiaca, peso e glicemia. I sensori erano collegati con il proprio smartphone personale ed inviavano costantemente informazioni ad un server. I vari test rivelano come le Long Short-Term Memory (LSTM) sono in grado di prevedere il futuro andamento dei livelli di glucosio nel sangue; le previsioni potrebbero essere combinate con suggerimenti personalizzati di diete e attività fisica al fine di migliorare la qualità della salute dei pazienti ed evitare condizioni critiche in futuro.

Sebbene i traguardi raggiunti grazie all'AI hanno già migliorato la vita di molte persone, ci sono ancora criticità da considerare negli sviluppi futuri. Un primo ostacolo è la dimensionalità dei dati: così come un modello non performa al meglio su un dataset di piccole dimensioni, anche avere troppe features può essere un problema, come nello studio del cancro, dove spesso non si riesce applicare una feature selection adeguata. Un'altra problematica sono i lunghi tempi di attesa prima che un modello sperimentale possa essere convalidato per l'effettiva adozione in contesti clinici. Il processo di validazione deve essere seguito con cautela, poiché vanno a scontrarsi con questioni etiche e legali. Qualora ci fosse un errore nella diagnosi o nella somministrazione di farmaci, a chi andrebbe la colpa? Per tal motivo, è necessario ampliare la ricerca verso soluzioni explainable che siano più chiare possibili.

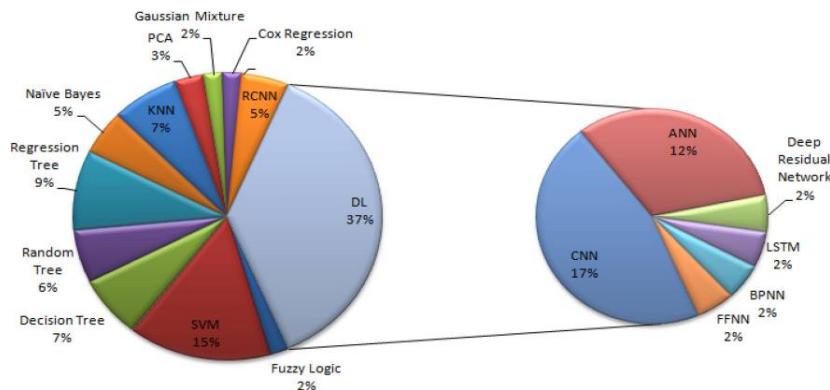


Figura 3.2: Modelli di intelligenza artificiale per la diagnostica

3.3 Clinical Decision Support System

Con Clinical Decision Support System (CDSS) si fa riferimento a un tipo di software che aiuta un medico nel processo decisionale analizzando i dati del paziente. Un CDSS, quindi, funge da supporto ai professionisti che possono ricevere nuove informazioni per restringere la diagnosi ed avere indicazioni per condurre dei nuovi test. Pensare che un CDSS possa rimpiazzare uno specialista è un errore comune, esso si limita a guidare il medico in modo che possa unire le sue competenze grazie a dei suggerimenti.

Un support System è composto da una architettura a tre layer (Figura 3.3). Vi è un livello che si occupa della gestione dei dati che ha il compito di mergiare un database contenente le informazioni su malattie, diagnosi e risultati con i dati del paziente, in più viene memorizzato una Knowledge Base o un modello di Machine Learning. Le informazioni processate sono inoltrate ad motore inferenziale che applica algoritmi sui dati del paziente e restuisce l' output ad un' interfaccia utente, la quale può trattarsi di una Web App, un' applicazione mobile o una dashboard.

Esistono due tipi di CDSS e si differenziano sul basarsi o meno su una base di conoscenza. Nel primo caso, i dati sono strutturati sotto forma di regole IF-THEN e seguono un processo inferenziale combinando le assunzioni memorizzate e la storia medica del paziente con le sue attuali condizioni di salute. I sistemi non knowledge-based non si basano su regole precise ma apprendono da esperienze passate ottenute dai dati storici e fanno uso di reti neurali e algoritmi genetici. Quest' ultimi tagliano di molto le spese sanitarie e sono in grado di ridurre la pressione sui medici. Lo svantaggio, oltre a richiedere potenza di calcolo e un'enorme quantità di dataset per l' addestramento, non sono in grado di spiegare le decisioni che generano e sono usate come black box.

Per questo motivo, la maggior parte di questi sistemi predilige sistemi knowledge-based. Oltre al supporto diagnostico, i CDSS hanno applicazione anche nella selezione di farmaci. Molti errori di prescrizione di farmaci provocano diversi morti all' anno, questa attività può essere automatizzata dai CDSS utilizzando il peso, l'età e le allergie del paziente. Anche la gestione delle cliniche può essere condotta da questi sistemi intelligenti che possono suggerire delle regole ospedaliere, come ad esempio guidare gli infermieri a seguire dei piani specifici con determinati pazienti. I sistemi CDSS contribuiscono a migliorare l'assistenza sanitaria, ma non sono privi di rischi e svantaggi. Inanzitutto, un CDSS può generare una moltitudine di consigli che possono confondere lo specialista che

non è più in grado di considerarli per grado di importanza. Oltre ai costi dell'adozione, un altro problema è quello dell'integrazione. Non tutti questi sistemi sono compatibili con un sistema informatico clinico pre-esistente e potrebbe essere necessario ristrutturare il formato dei dati.

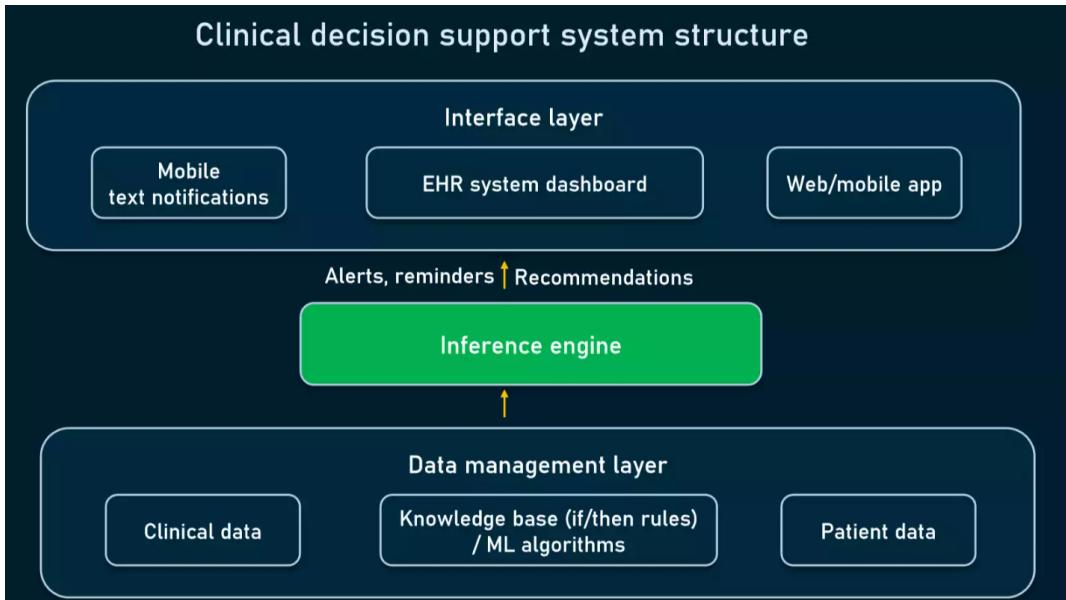


Figura 3.3: Architettura di un Clinical Decision Support System.

3.4 Malattie rilevabili dal fondo oculare

Il rapido invecchiamento della popolazione degli ultimi anni nei paesi sviluppati ha contribuito all'esplosione di un numero di casi di persone con difetti visivi. Oltre la longevità, lo stile di vita scorretto ha portato ad un aumento di casi affetti da diabete di tipo 1. Per tal motivo, le pressioni sul sistema sanitario hanno prolungato i tempi di attesa per essere visitati da un oftalmologo e questo può demotivare il paziente nel fare prevenzione.

Principalmente, gli esami più comuni di una visita oculistica sono due: l'esame del fondo oculare e la Tomografia a Coerenza Ottica (OCT), quest'ultima verrà più approfondita nella sezione 4.1.1. L'esame del "fundus" consiste nell'osservazione delle strutture posteriori del bulbo oculare, come la cavità vitreale, il nervo ottico e la retina. Varie malattie possono insorgere colpendo alcuni di questi elementi e necessitano di un trattamento immediato. Tra le malattie rilevabili dal fondo oculare ritroviamo il distacco della retina, la retinopatia e la degenerazione maculare legata all'età. L'OCT, invece, risulta più indicata

per malattie retiniche e per l' individuazione di segni precoci del glaucoma. Di seguito una descrizione più dettagliata delle malattie appena citate:

Distacco della retina può verificarsi quando i tessuti della membrana più interna dell' occhio cedono. Questo può portare alla visione di corpuscoli mobili che fluttuano nel vitreo e ad una riduzione della vista. In tali casi, è necessario un intervento immediato affinché si riattivino le funzionalità delle cellule retiniche evitando una cecità parziale o totale dell' occhio.

Retinopatia è un termine che genericamente include tutte le patologie che colpiscono la retina. La più comune è la retinopatia diabetica, la quale è una complicanza vascolare del diabete e può essere di due tipi, Una forma precoce (o non proliferativa) dove vi è un danneggiamento della retina, senza la crescita di nuovi vasi sanguigni. L' ostruzione dei vasi, però, può portare ad un edema maculare (DME). Nella forma proliferativa, nuovi vasi anomali possono disperdere sangue nel vitreo.

La degenerazione maculare legata all'età (AMD) è una delle tra le maggiori cause dell' acutezza visiva nei paesi occidentali. L' AMD è una patologia che comprende molti fattori, tra i quali la predisposizione genetica e diversi fattori ambientali. Oltre all' età, che è considerato il principale fattore di rischio, altre cause sono il fumo, l' ipertensione arteriosa, una dieta ricca di grassi, l' obesità addominale e ridotta attività fisica. L' Age-Related Eye Disease Study Group (AREDS) ha classificato la degenerazione maculare in due grandi categorie: AMD neovascolare (forma umidità o essudativa), caratterizzata dalla presenza di una neovascularizzazione coroideale (CNV). Questi vasi di origine coroideale, penetrano attraverso la membrana di Brunch e si estendono o sotto l' epitelio pigmentato retinico (CNV di tipo 1) o sopra l'epitelio pigmentato stesso (CNV di tipo 2). Questi vasi anomali possono indurre emorragie, essudazione di fluido e fibrosi, causando in un danno dei fotorecettori e conseguente perdita della funzione visiva; AMD non neovascolare (forma secca o atrofica), dove compiano accumuli di scorie cellulari di scorie e possono seguire alterazioni atrofiche della retina, cioè perdita di parte del tessuto retinico con conseguente calo della vista. La forma neovascolare e la forma d' atrofia geografica costituiscono uno stadio avanzato dell' AMD.

glaucoma è una malattia che compromette il nervo ottico e induce ad una progressiva perdita del campo visivo. In genere è caratterizzato da un'alta pressione intraoculare che può essere influenzata anche dallo spessore della cornea.

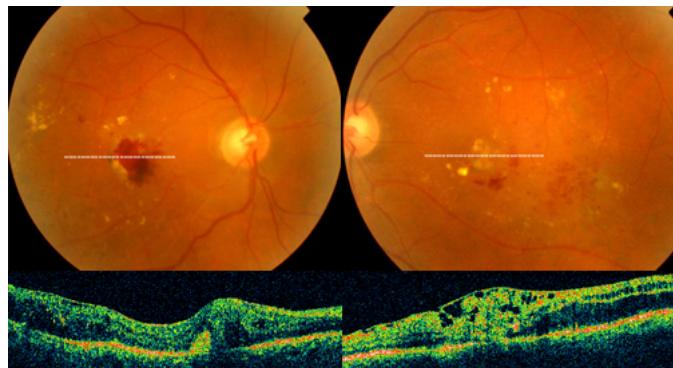


Figura 3.4: Retinopatia diabetica su fondo oculare ed OCT.

3.5 Intelligenza artificiale e oftalmologia

Uno studio in Inghilterra ha stimato che un ritardo di circa 22 settimane nella cura di una patologia oculare può comportare un deterioramento irreversibile nell'acutezza visiva e la perdita del campo visivo in alcuni pazienti che avrebbero potuto evitare tali danni con interventi precedenti [13]. Questo fa riflettere su quanto sia importante l'immediatezza nel trattamento di tali malattie e di come il sistema sanitario attuale non riesce a sostenere ritmi così accelerati.

In Scozia, per far fronte a ciò, nasce il progetto *Scottish Eyecare Integration*, dove un gruppo di oftalmologi specializzati lavorano in modo coordinato affinché si possano seguire più pazienti possibili tramite una precisa politica di smistamento di quest'ultimi in grado di ottimizzare le risorse a disposizione e minimizzare le attese. Non solo il numero di visite è bilanciato in modo equo tra gli specialisti, ma lavorano anche in modo interoperabile tra loro per il trattamento di soggetti che richiedono particolari attenzioni.

Il meccanismo scozzese ha ricevuto un'enorme soddisfazione dai pazienti ma di certo le code di attesa non sono totalmente scomparse. Tutt'ora la scalabilità di queste operazioni è limitata alle strutture e agli specialisti perciò ci si sta spostando su approcci automatizzati con l'AI di alcuni processi di diagnosi. Solo negli ultimi cinque anni sono stati pubblicati oltre ventimila articoli sull'intelligenza artificiale, più di mille solo relativi all'oftalmologia.

La promessa dell'applicazione dell'AI in oculistica è quella di potenziare l'accessibilità, la disponibilità e la produttività delle risorse esistenti dei servizi

di assistenza. La ricerca in questo settore negli ultimi anni ha saputo fornire risultati concreti che hanno eguagliato o superato le prestazioni umane grazie all'uso del Deep Learning.

Questi traguardi, soprattutto per quanto concerne la classificazione di fotografie del fondo oculare e scansioni di tomografia a coerenza ottica, sono stati legittimati come "clinicamente accettabili". Un esempio già tangibile lo si vede su diversi sistemi per lo screening della retinopatia diabetica (DR): la crescente proliferazione di nuovi casi di DR corre ad un passo troppo veloce rispetto al lungo percorso di formazione di nuovi professionisti, creando una divaricazione enorme nell'assistenza. Fortunatamente, i nuovi sistemi di AI sono stati capaci di automatizzare la diagnosi e segnalare con un alto grado di affidabilità solo i casi positivi, i quali sono trattati immediatamente dagli specialisti. L'AI ha avuto un impatto dirompente non solo sull'automatizzazione delle diagnosi ma anche su altri compiti; ad esempio sono già in uso dei software che prevedono la progressione della degenerazione maculare legata all'età tramite l'analisi della scansione OCT.

Una panoramica sulle applicazioni di NLP è illustrata in [9]; sono già in uso procedure di estrazione del testo attraverso algoritmi da cartelle cliniche elettroniche (EHR) per l'ottenimento di informazioni su diagnosi di cataratta e glaucoma, nome dei farmaci, procedure oftalmiche e complicazioni chirurgiche, operazioni che in passato venivano fatte manualmente anche per tracciare delle statistiche sulle varie patologie; la nuova conoscenza estratta può essere sfruttata anche per migliorare le cure cliniche.

Sempre dalle EHR, [4] ha condotto uno studio per ricavare dati strutturati sull'acutezza visiva. In generale, queste informazioni di solito sono presenti in testi non strutturati, ma grazie all'algoritmo TOVA ne si automatizza l'estrazione e la standardizzazione in valori numerici. Sebbene l'adozione di nuovi metodi innovativi stia crescendo in modo esponenziale, le varie implementazioni devono far fronte a sfide ancora aperte: in primis, i vari modelli di AI ancora non hanno raggiunto l'ottimalità, specialmente nel NLP; è necessario fornire i pazienti di dispositivi di uso comune per l'autoscansione del fondo oculare e di conseguenza migliorare l'operabilità notificando al sistema sanitario di eventuali malattie; migliorare sistemi di Question-Answering, in modo che il paziente possa in automatico ricevere le riposte di cui ha bisogno; estendere la fruizione di connessione internet in più località, dove non solo si dà la possibilità al singolo paziente di interagire con il sistema sanitario autonomamente, ma anche per poter fondare centri e piattaforme di screening (anche mobili) che fanno uso

di macchine intelligenti. Ovviamente, questo va a scontrarsi con i costi per la creazione di Hardware e Software specifici e costi per il mantenimento e aggiornamento di essi.

In commercio esistono già alcune soluzioni. IDx-DR è il primo software approvato per lo screening della retinopatia diabetica basato su cloud e raggiunge una sensitivity del 96.8%. Anche Retmarker fa screening per DR ed esamina il fondo oculare in "normale" o "anomalo" ed è in grado di fornire comparazioni nel tempo dell' andamento della malattia (vedi Figura 3.5). Un' ulteriore sfida è proiettata verso il massivo labelling di immagini oculari per migliorare le soluzioni di intelligenza di artificiale. Sempre più enti stanno iniziando a etichettare scansioni grazie all' aiuto di specialisti in diverse parti del mondo.

ODIR [16] è un set di dati pubblico cinese e utilizza immagini di fotocamere retiniche Canon, Zeiss e Kowa; comprende 8.000 immagini retiniche, classificate come normali o presenza di retinopatia diabetica, glaucoma, cataratta, AMD, ipertensione, miopia e altre condizioni; nelle etichette ODIR c'è una descrizione dell'età del paziente, ma nessuna descrizione del sesso, degli aspetti socioeconomici o dell'etnia. In Brasile, vengono forniti i dataset DR1 e DR2 [31] di pazienti con DR, composti da 1.597 immagini ottenuti da fotocamera D90 Nikon, purtanto non danno dettagli sulla stadio della malattia. IDRiD [32] è un dataset indiano che comprende 516 immagini e le scansioni provengono da una fotocamera digitale alfa Kowa VX-10.

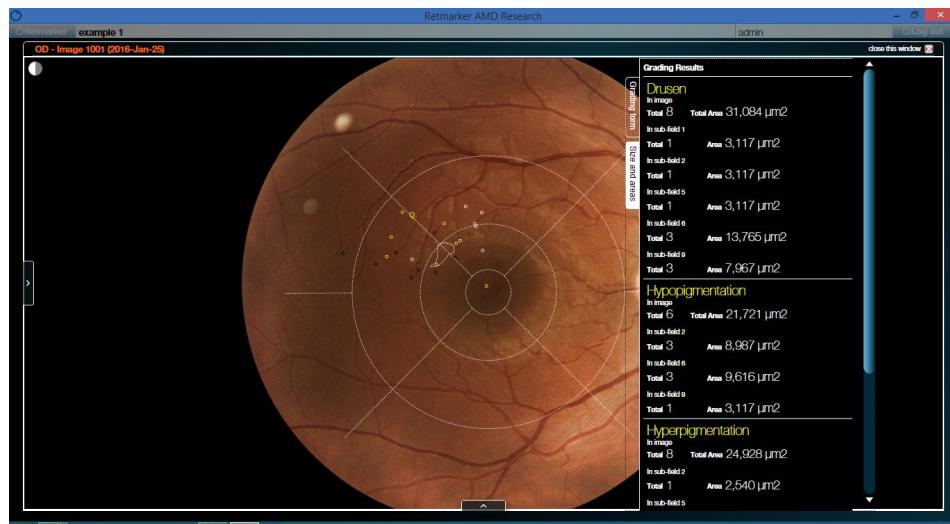


Figura 3.5: Screenshot del Software RetMarker.

Capitolo 4

Identificazione di malattie da immagini di Tomografia Ottica Computerizzata (OCT)

In questo studio si è cercato di ritagliare un campo d'applicazione nell' oftalmologia ancora austero e, con riferimento allo stato dell' arte di proposte in ambito eHealth e modelli all' avanguardia, si è scelto un set di componenti per l' implementazione di un sistema diagnostico intelligente. Tutti i vari tasselli sono composti prevalentemente da tipologie di Reti Neurali, ognuna dedita alla risoluzione di un problema. In questo capitolo si presentano a livello teorico tutti gli elementi che entrano in gioco per la soluzione presentata, descrivendone il funzionamento e l' utilità. Le sfide che si sono affrontate trattano problemi di classificazione di immagini OCT, ampiamente superate con l' uso di Reti Convoluzionali; Explainability testuale, dove vi è necessità di generare testo, questione che si rifà ai moderni sviluppi e traguardi raggiunti nel campo dell' NLP, in particolare nella Natural Language Generation. Un Modello di rilievo è rappresentati dalla Long Short-Term Memory (LSTM), il quale è utilizzato in un architettura Encoder-Decoder per la generazione della didascalia di diagnosi, e dai moderni Transformer che sono stati utili nella generazione di contesto grazie ai moduli di Self-Attention; un altro obiettivo prioritario è poter rilevare le aree di interesse che localizzano una patologia dalla scansione fondo oculare, a tal proposito strumenti di Visual XAI come CAM e Grad-CAM sono ampiamente discussi nelle sezioni successive. Tutta l' analisi ha avuto origine dallo studio del dominio oftalmologico, individuando pattern patologici ricorrenti e studiando la natura del tipo di imaging da trattare (OCT), fino all' introduzione del Multi-Task Learning, il quale ha raffinato il Tool presentato.

4.1 Dominio di applicazione

La tomografia ottica a luce coerente (dall'inglese "Optical Coherence Tomography") è uno strumento non invasivo, che non prevede l' uso di coloranti, di veloce esecuzione e permette di visualizzare alcune caratteristiche morfologiche come l' accumulo di fluidi e fenomeni di ispessimento della retina. La tecnologia OCT sfrutta la riflessione di onde luminose di strutture biologiche: quando un fascio di luce va a contatto con del materiale biologico, vengono indotte oscillazioni su di esso con frequenza pari a quella dei fotoni incidenti. Tali oscillazioni generano a loro volta emissioni di nuovi fotoni che vengono intercettati da un interferometro che processa il segnale generando l'immagine OCT. In questo modo è possibile catturare immagini ad alta risoluzione di sezioni trasversali della retina; queste ultime vengono interfogliate e possono dare un'immagine tridimensionale. Ogni sezione è una scansione rappresentante un'immagine diversa. I tracciati A-Scan, che sono monodimensionali, sono utilizzati per misurare la lunghezza dell' occhio e sono usate raramente. B-Scan rappresenta un' immagine bidimensionale ottenuta da 1600 A-Scan una in fila all'altra ed è la scansione più usata in oculistica, per tal motivo sono il tipo di imaging usato da questo studio. Su una B-Scan è possibile valutare la struttura interna dei tessuti. I tracciati C-Scan, invece, sono tridimensionali e sono generati dall' affiancamento di 256 immagini B-Scan; qui è possibile osservare la retina nella sua profondità in modo da valutarne i rigonfiamenti e deformazioni della superficie.

L'OCT ha maggior impiego sulle patologie retiniche poiché è in grado di rilevare trazioni di vitreo e di conseguenza della retina, identificare membrane epiretiniche maculari, osservare in modo dettagliato i vari strati retinici. Con OCT è possibile analizzare anche lo spessore e la curvatura della cornea (per individuare malattie come il cheratocono) e valutare la morfologia del nervo ottico che può essere alterato dal glaucoma. La tomografia a coerenza ottica ha sicuramente rivoluzionato il modo di fare un' inadgine oculistica. Tuttavia, non è in grado di distinguere e delimitare in modo preciso il tessuto neovascolare dal tessuto fibrotico e i tessuti circostanti.

Negli ultimi anni, un nuovo strumento diagnostico non invasivo è stato introdotto nella pratica clinica: l'angiografia OCT (OCT-A). L' OCT-A, utilizzando la decorrelazione del segnale tra diverse scansioni OCT ripetute, identifica il movimento degli eritrociti e ricostruisce da esso il volume di vasi retinici e coroideali. Anche OCT-A ha dei limiti: la presenza di artefatti di movimento e di proiezione può invalidare il risultato dell' esame; grazie all' introduzione di nuovi sistemi di

eye-tracking vengono risolte parzialmente queste problematiche che rimuovono tali artefatti.

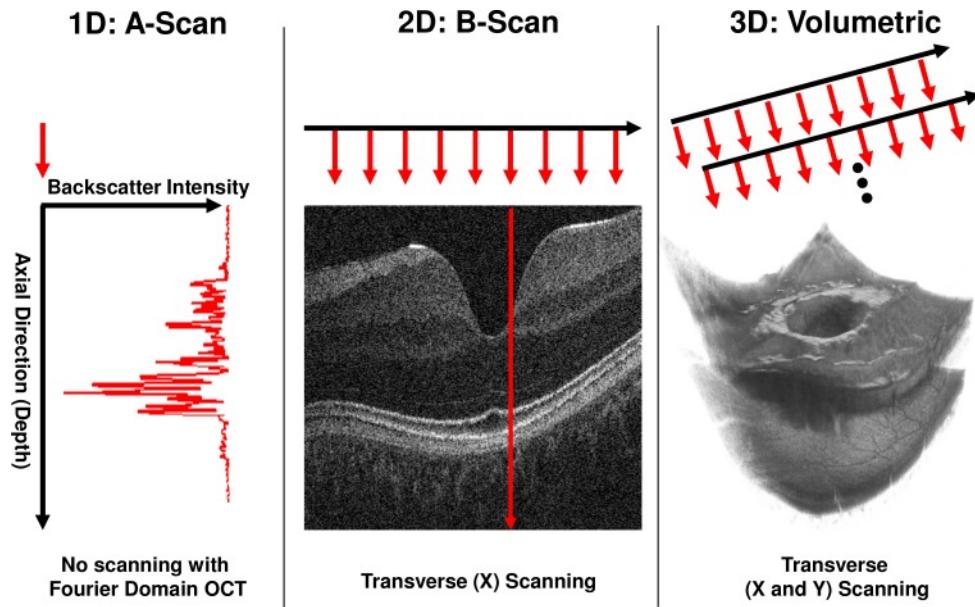


Figura 4.1: Differenti tipologie di scan in OCT.

4.1.1 Malattie studiate

Questo elaborato tratta diverse manifestazioni della degenerazione maculare legata all' età (patologia introdotta nella Sezione 3.4) e l' edema maculare diabetico, complicazione dovuta al diabete, individuabili tramite OCT Scan-B. Una macula sana (parte della retina che permette la visione centrale), se osservata trasversalmente, è formata da una stratificazione di diversi elementi (Figura 4.2). Dall' alto verso il basso: vi è la depressione foveale, la zona di migliore definizione visiva; uno strato di fotorecettori, che convertono lo stimolo luminoso in segnale elettrochimico; l' epitelio pigmentato retinico (RPE); la membrana di Bruch, costituita da diversi strati collagenosi e più esternamente la coroide che è un tessuto vascolarizzato. Normalmente quest' ultima rilascia attraverso nutrienti che vengono veicolati dai fotorecettori tramite l' attività dell' RPE. L' epitelio pigmentato retinico smaltisce poi tutti i prodotti di scarto che fagocita e che provengono dalla degradazione dei fotorecettori durante la visione. Il malfunzionamento dell' RPE e questi prodotti di degradazione tendono ad accumularsi sotto l' epitelio sottoforma di piccoli depositi giallastri chiamati **drusen**, considerata una forma precoce dell' AMD. Le drusen sono costituite da materiale proteico e glicemico e ostruiscono il passaggio durante lo scambio di nutrienti con

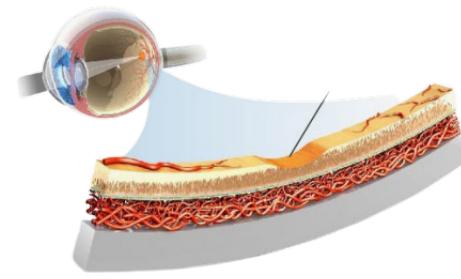


Figura 4.2: Sezioni della macula.

la coroide. L' RPE sovrastante, per questo motivo, va incontro ad un processo di distrofia finché non viene danneggiato. La forma precode di AMD, nei casi più gravi può evolversi nella forma "umida". A seguito della presenza delle drusen, si può avere un ridotto contributo di ossigeno nel ciclo vitale della retina. Questo può stimolare una produzione vascolare anomala che porta alla proliferazione di nuovi vasi sanguigni. Questi neovasi possono far uscire sangue ed altri fluidi creando un distacco (PED) dell' RPE dalla membrana di Bruch. Il distacco può aumentare finché i vasi non attraversano l'epitelio ed è qui che paziente presenta sintomi. Questa patologia viene chiamato **neovascolarizzazione coroideale** (CNV). Un'altra categoria, come già accennato è l'edema maculare diabetico. L' iperglicemia può portare alterazioni al microcircolazione retinica provocando l' ispessimento della retina, la chiusura di vasi fino a lesioni, caratteristiche della retinopatia diabetica. La rottura di un vaso può sfociare in emorragie che confluiscono nella macula causando l' edema maculare diabetico (**DME**). Concludendo, questo studio si sofferma su drusen, neovascolarizzazione coroideale e edema maculare diabetico rispettivamente assegnate le labels DRUSEN, CNV, DME 4.3. Oltre alle tre malattie retiniche, è stata necessaria aggiungere la categoria NORMAL in modo da creare modelli che sappiano riconoscere anche una macula senza patologie.

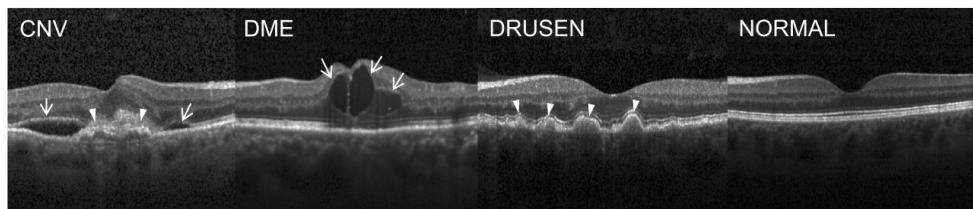


Figura 4.3: OCT Scan-B con i pattern tipici di drusen, neovascolarizzazione coroideale e edema maculare diabetico.

4.2 Reti convoluzionali per Image Classification

In questa sezione vengono introdotte le Reti Neurali Feed-Forward e il relativo processo di apprendimento. Successivamente si descrivono le Reti Neurali Convoluzionali e la tecnica del Transfer Learning.

4.2.1 Rete Neurale

Il modello designato da questo studio per la classificazione è la rete neurale. Ispiratosi in grandi linee ad una rete neurale biologica, questi modelli sono frequentemente utilizzati per risolvere problemi di qualsiasi genere. In particolare, l’ analogia con le neuroscienze deriva dalla legge di Hebb che afferma: se due neuroni sono attivi all’incirca nello stesso momento, le loro connessioni vengono rafforzate; allo stesso modo, i nodi del modello si comportano nella stessa maniera. Le reti neurali si differenziano da soluzioni algoritmiche classiche in quanto chi costruisce il modello non deve dichiarare a priori un set di passi ad hoc fortemente dipendenti dai dati di input: è una struttura che è la generalizzazione universale estesa di qualunque altro modello matematico. L’efficacia deriva da fasi di training da cui apprendere conoscenza dai dati. Generalmente, una rete di tipo Feed-Forward (Figura 4.4) ha un’ architettura partizionata in diversi layers consecutivi che a loro volta sono composti da più nodi (neuroni); ogni neurone del layer è collegato a ciascun neurone del layer successivo. Ogni input viene fatto passare per ciascuno strato; la particolare caratteristica di un neurone è che esso propaga l’informazione in avanti soltanto se supera una specifica *soglia di attivazione*, perciò ciascun nodo ha degli ingressi provenienti dalle elaborazioni di tutti i neuroni dei layers precedenti. Gli archi di collegamento sono pesati e il valore di un nodo è quindi la somma degli ingressi con i pesi più un valore chiamato bias (determina se e in quale misura il neurone debba attivarsi o meno). L’intero processo di apprendimento si completa iterando un insieme di passi per diverse epoche. Inizialmente i pesi vengono inizializzati in modo casuale; nella fase di Forward Propagation l’input viene fatto passare su tutta la rete dal primo all’ ultimo layer, il quale darà una predizione. L’ output risultante viene confrontato con la label reale ottenendo uno scarto (loss) che ne quantizza l’errore. Affinché si minimizzi l’ errore bisogna trovare il minimo globale di una funzione di costo rispetto tutti i pesi della rete neurale e, per capire in che direzione si trovi, è necessario calcolare il gradiente. Quest’ ultimo è un vettore n-dimensionale le cui componenti sono le derivate parziali della funzione costo rispetto a ciascun peso della rete. Nella fase di backpropagation

vengono calcolati i nuovi pesi grazie al gradiente che suggerisce come aggiornarli per minimizzare la loss.

$$w^{t+1} = w^t - \eta \frac{d}{dw} L(f(x; w), y) \quad (4.1)$$

Nell'equazione 4.1, il valore di un nuovo peso w^{t+1} viene aggiornato rispetto ad una funzione di loss $L(f(x; w), y)$ ed al Learning Rate η . Quest'ultimo è un iperparametro, ossia un parametro regolabile che consente di regolare il processo di training del modello; η stabilisce con quanta intensità aggiornare i pesi della nostra rete rispetto alla loss. Un tasso di apprendimento troppo grande può far convergere il modello troppo rapidamente verso una soluzione non ottimale, mentre un valore troppo piccolo può causare il blocco dell'apprendimento. Un altro iperparametro fondamentale è il BATCH-SIZE: è il numero di campioni alla volta che verranno propagati attraverso la rete per epoca. In genere, usare un BATCH-SIZE minore del numero di campioni richiede meno memoria, poiché la procedura di training è meno onerosa. D'altro canto, un BATCH-SIZE troppo piccolo diminuirà la stima del gradiente; un mini-batch tende a non far stabilizzare la rete e porta a risultati non soddisfacenti.

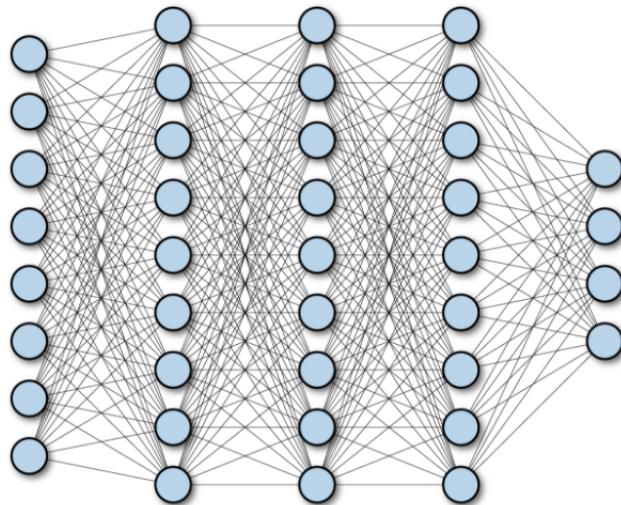


Figura 4.4: Rete neurale feed forward.

4.2.2 Rete Neurale Convoluzionale

Nell'ambito della Computer Vision, le **reti convoluzionali (CNN)** sono particolari reti neurali ampiamente utilizzate per il riconoscimento di oggetti in un'immagine. Un rete CNN funziona esattamente come una rete Feed Forward ma

si differenzia con la presenza di livelli di convoluzione. Il loro compito è quello di estrarre delle features dall' immagine che provengono dall' applicazione di diversi filtri. Ogni filtro è in grado di identificare elementi come linee verticali, orizzontali, diagonali e contorni delle figure; per questo motivo, la rete tenderà a concentrarsi su queste informazioni per la classificazione. Potenzialmente, una rete Feed Forward è in grado di adempiere allo stesso compito ma sarebbe particolarmente oneroso considerando il numero elevato di nodi negli strati nascosti fully connected. La CNN, invece, è in grado di trattare informazioni più specifiche ed essere di conseguenza più efficiente. Una rete convoluzionale è composta da un susseguirsi di blocchi di elementi diversi che vengono ripetuti a catena:

Livello di input: è composto da una serie di nodi che ricevono in input l' immagine. Quest' ultima dovrà essere prima ridimensionata e trasformata in tensore che rappresenterà i pixel. Qualora si trattasse di un'immagine a colori, per ogni pixel si avrebbero 3 valori rappresentanti i livelli di Rosso, verde e blu del formato RGB.

Livello Convoluzionale: è il layer cruciale di una CNN, poiché è in grado di riconoscere schemi visivi via via sempre più complessi tramite l' applicazione di filtri (kernel). Questi, possono essere visti come delle maschere di dimensioni ridotte che vengono fatte scorrere sull' immagine (vedi Figura 4.5); ad ogni spostamento viene generato il prodotto scalare tra la maschera e la relativa posizione che copre. Ogni pattern trovato dal kernel con il suo passaggio viene riportato in una features map risultante.

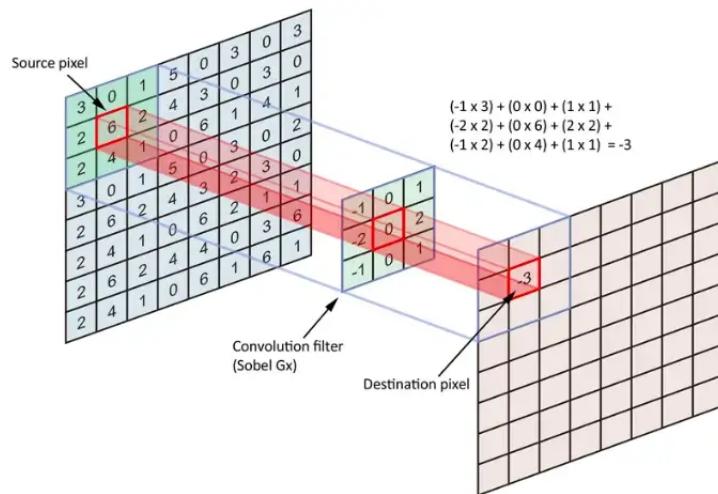


Figura 4.5: Filtro convoluzionale.

Livello di attivazione: determinano il passaggio o meno di alcuni valori provenienti dai livelli precedenti. Questo è possibile con l' uso di funzioni di attivazione come ReLU (Rectified Linear Units) o la sigmoide che determinano delle soglie che se non superate non propagano l' informazione in avanti; questo significa scartare delle caratteristiche che non sono state in grado di contribuire per la predizione.

Livello di Pooling: è un layer che aggrega le informazioni ricevute e genera una feature map di dimensione inferiore a seconda di un criterio. Nel caso del Max-Pooling, la dimensione dell' immagine viene ridotta suddividendola in vari blocchi (specificandone la taglia a priori) e per ognuno di essi solo il valore più alto sarà rappresentato nella nuova mappa delle caratteristiche. L' operazione di Pooling tende a conservare solo le aree con maggiore attivazione, riducendo così un eventuale overfitting.

Layer pienamente connesso: è lo strato denso di nodi che ha il compito di dare in output la predizione finale.

Esistono diverse architetture in cui variano Il numero di layer e la loro distribuzione. Più la catena è lunga, più la CNN sarà in grado di riconoscere dei pattern visivi. Una tipica catena è: *Input → Conv → ReLU → Pool → Conv → Relu → Pool → Relu → Conv → Relu → Pool → FullyConnected*. Alcuni strati, possono essere anche omessi, come nel caso della rete LeNet. Il modello ha solamente cinque strati di convoluzione seguiti da due strati completamente connessi; l' architettura (concepita ormai più di venti anni fa), anche se semplice, è tutt' ora utilizzata per varie attività come il riconoscimento di cifre scritte a mano, il riconoscimento dei segnali stradali e il rilevamento dei volti.

4.2.3 Transfer Learning

Il Transfer Learning [15] è una tecnica di Machine Learning in cui un modello pre-addestrato viene riutilizzato per un nuovo task. Se il modello di partenza svolge un compito simile al nuovo task, allora la conoscenza appresa dal primo può essere generalizzata per la risoluzione del secondo. Questo significa che il training sul nuovo problema non deve partire da zero e non ha bisogno di una grande quantità di dati e tante ore di addestramento e confluire a performance ideali sulla base dei pesi del task primario. Per questo motivo questo approccio sta diventando sempre più utilizzato; i modelli di AI sono costosi, basti pensare

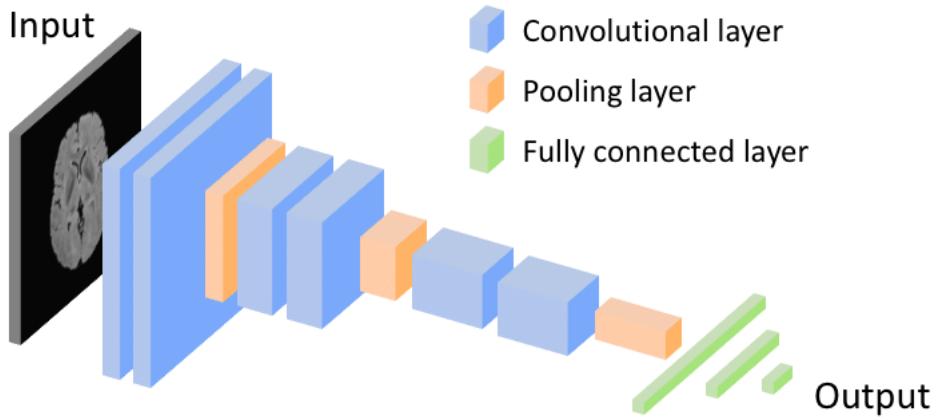


Figura 4.6: Architettura di una Rete Neurale Convoluzionale.

che GPT-3 è costato milioni di dollari per l' addestramento, BERT sui 12000 ed è open source; chiunque può iniziare a fare ricerca a partire da questi modelli e ritagliarli per task customizzati. Più formalmente, su un modello pre-addestrato vengono rimossi i layers densi specifici della classificazione del task A e il resto viene conservato meramente come estrattore delle features. Quest' ultimo viene esteso con un layer denso che servirà a ridirigere l' output verso il riconoscimento delle nuove classi del task B ed i pesi dell' estrattore rimarranno congelati (si parla di Freezing). Si parla di Fine-Tuning quando invece viene riaddestrato l' intero modello, in genere con un learning rate più basso, in modo da adattare in modo incrementale le features pre-allenate con i nuovi dati.

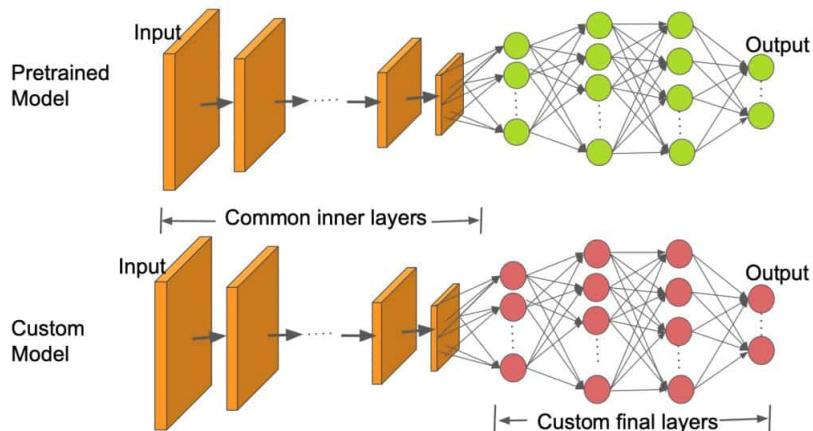


Figura 4.7: Architettura di una Rete Neurale Convoluzionale.

4.3 Image Captioning

L' Image Caption Generator è il processo di riconoscimento del contesto in un' immagine con la generazione di didascalie testuali pertinenti utilizzando il Deep Learning e la Computer Vision. In particolare fa uso di reti convoluzionali e altri modelli per il Natural Language Processing (NLP) i quali sono discussi nelle sezioni successive.

4.3.1 Natural Language Processing

Il Natural Language Processing (NLP) è una disciplina dell' Intelligenza artificiale che comprende algoritmi per la comprensione, rappresentazione ed analisi del linguaggio umano. La formalizzazione di quest' ultimo è uno step necessario per la realizzazione di sistemi in grado di rispondere e comprendere l' uomo, tuttavia la miriade di aspetti che entrano in gioco come la semantica, la fonetica, la sintassi, rendono il compito particolarmente arduo. Esistono diversi approcci per formalizzare in termini matematici il significato delle parole; una tecnica usata per la comprensione di una frase è la rappresentazione tramite vettori (vector semantics). Si basa sull' idea che due parole, più sono presenti in contesti simili e più possiamo considerarle sinonimi e quindi con molta probabilità si troveranno vicine in uno spazio semantico che si va a definire. Questo spazio N-dimensionale contiene tutte le parole, le quali sono raggruppate in vettori, chiamati *Embeddings*.

L'NLP trova molte applicazioni nella vita quotidiana. Vi sono sistemi di AI che effettuano Machine Translation, ossia traduzione di testo da una lingua con ormai con alta affidabilità; tecniche di Information Extraction sono già utilizzate in molti servizi di posta elettronica in cui vengono estratte indicazioni su data, ora e luogo in appuntamento per l' autogenerazione di un evento sul proprio calendario personale. Soluzioni di sentiment analysis sono in grado di capire il tasso di apprezzamento di un trend, classificando i commenti in modo negativo o positivo. Questi task appena elencati si comportano già in modo accettabile dalla comunità anche se possono essere ulteriormente raffinati. Esistono invece delle problematiche già risolte dall'NLP come la detection di email spam o la name entity recognition, ossia capire un termine a quale entità fa riferimento (una persona, una località, un'organizzazione). La Part-of-speech (POS) tagging è anch' esso considerata risolta: data una frase, il sistema deve associare per ogni parola la corretta parte variabili del discorso (ossia individuare se trattasi di un avverbio, nome, aggettivo, etc).

Tuttavia, sfide ancora in corso riguardano la Question Answering, quindi riuscire a interpretare una domanda e dare una risposta, o compiti di Summarization, in cui, dato un documento, l'obiettivo è sintetizzarlo in modo breve e conciso. Nemmeno i Chatbot hanno raggiunto una verosomiglianza di dialogo pari a quella dell'uomo poiché hanno difficoltà nel mantenere un filo sensato del contesto man mano che si prosegue la conversazione.

Sono ora aperte questioni su come un modello deve approcciarsi alla comprensione di un testo; una persona comune leggerebbe un testo concettualizzando idee che ha del mondo reale, mettendo in relazione oggetti, desideri e credenze ed è da chiedersi se un sistema di intelligenza artificiale dovrebbe simulare in modo così astratto tale processo o meno. Tuttavia, citando Kevin Gimpel, ricercatore specializzato in NLP, finché non si tende al pensiero umano, i modelli non saranno altro che dei meccanismi che eseguono matching di pattern senza mostrare una comprensione "reale".

Natural Language Generation si occupa della generazione automatica di testo in linguaggio naturale sulla base di contesti forniti come input. Gli usi più diffusi sono per la generazione di risposte per chatbot e assistenti vocali, creazione di storie e poesie, creazione di didascalie, modelli di sintesi di testi, reporting verbale di dati strutturati e molto altro ancora. In [34] viene data una visione degli approcci NLG utilizzati in passato fino a quelli più recenti. Tradizionalmente, l'architettura di questi sistemi era composta da una pipeline di sei operazioni per generare testo:

- **Analisi del Contenuto:** i dati vengono filtrati in modo da ottenere solo i contenuti ritenuti più importanti.
- **Comprensione:** tramite tecniche di Machine Learning vengono identificati dei pattern con cui generare contesti.
- **Strutturazione del documento:** viene scelta una precisa struttura narrativa per il testo da generare.
- **Aggregazione:** vengono pescati e uniti pezzi di vari frasi per crearne ex novo.
- **Strutturazione Grammaticale:** si analizza la struttura sintattica del testo per correggerlo e renderlo più naturale.
- **Presentazione:** formattazione e visualizzazione dell'output.

Dopo gli anni 2000, si è passati ad approcci più statistici per ridurre lo sforzo di dover scrivere a mano regole grammaticali. Tuttavia, un principale svantaggio di tali metodi è l' alto carico computazionale, in quanto generano un sacco di possibili frasi che di solito sono filtrate con l' ausilio di particolari classificatori. Una svolta si è avuta con l' avvento di nuovi Reti Neurali profonde, dove l' introduzione di embeddings spostò il focus sulla semantica anziché sulla sintassi. Il primo Language Model (modello probabilistico in grado di prevedere la parola successiva data una sequenza di parole precedenti) di stampo Neurale fu una Rete Feed-Forward. Questa implementazione richiede un numero fissato N di parole da dare in input per predire la parola seguente; poiché un testo fornito può avere lunghezza maggiore, viene fatta passare una "finestra scorrevole" di N parole sull' input. Ad esempio, in Fig. 4.8 si vuole predire quale sarà la parola w_t ; per ogni parola nella finestra corrente viene calcolato l' embedding nel Projection Layer, il quale sarà propagato attraverso i layer nascosti fino all' Output Layer che indicherà una nuova parola w_t tra tutte quelle presenti nel vocabolario V . In conclusione, con la "sliding windows" stiamo assumendo che $P(w_t | w_1^{t-1}) \approx P(w_t | w_{t-N+1}^{t-1})$. Lo svantaggio principale di una rete Feed-

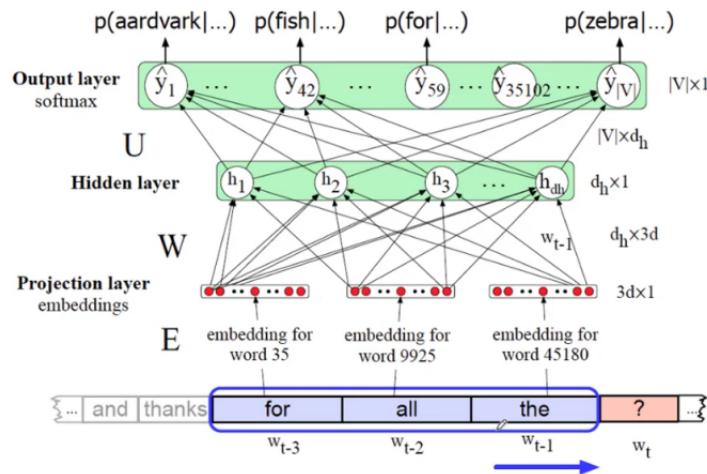


Figura 4.8: Language Modeling basato su Rete feed-forward.

Forward è la persistenza temporale della frase limitata alle parole nella finestra; tutte le parole che ne sono fuori vengono dimenticate. Un' alternativa più robusta che ha considerazione anche della cronologia passata sono le Reti Neurali Ricorrenti (RNN). In questo caso, per il calcolo della predizione può essere considerata anche la prima parola del testo. In un RNN l' output di un nodo viene dato in input al livello precedente (Fig. 4.9): così facendo, l' output O nell'

istante t dipende dall' input X e da V che per effetto della retroazione sarebbe O stesso ma dell' istante precedente.

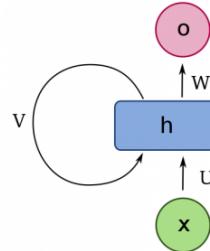


Figura 4.9: Retroazione di una Rete Neurale Ricorrente.

Purtroppo, tale architettura va ad accentuare due problemi comuni che erano già presente nelle Feed-Forward. Durante l' apprendimento, in particolare durante il calcolo delle derivate, le funzioni di attivazioni possono azzerare o aumentare il valore del gradiente. La funzione Sigmoid tende a mantenere i valori dei gradienti nell' intervallo $[0, 1]$, perciò più prodotti a catena tra gradienti porta ad avere un valore sempre più basso, il cosiddetto porta ad un Vanishing Gradient. Al contrario, la ReLu ha valori dei gradienti superiori a 1 e la loro moltiplicazione può via via raggiungere valori enormi (Exploding Gradient). Un ulteriore problematica del gradiente è che i pesi della parte iniziale della sequenza sono molto vicino a zero, quindi vi è una tendenza a dimenticare le ultime parole e limitare il contesto solo alle ultime. Per minimizzare queste eventualità vengono introdotte le LSTM (Long Short-term Memory) [43] che verranno descritte nel paragrafo successivo in merito all' architettura per il Captioning. Lo stato dell' arte dell' NLG dei tempi moderni sono i Transformers (paragrafo 4.4), utilizzati anch' essi per questo elaborato.

4.3.2 Architettura Encoder-Decoder

Il task di Image Captioning permette di fornire un spiegazione testuale di un' immagine data in input. L' architettura che realizza tal compito è proposta in [36] ed è definita come una struttura Encoder-Decoder. L' Encoder rappresenta un modello in grado di estrarre delle features da pattern visivi. Generalmente è implementato da una rete neurale convoluzionale che codifica l' immagine in input. Le features ricavate vengono successivamente propagate su un layer latente detto "Latent Space Vector", il quale fa da ponte per il modello di Decoder. Le caratteristiche mappate su tal vettore sono decodificate da una LSTM.

Long Short-term Memory sono un tipo di Reti Neurali Ricorrenti che esentano dai problemi sui gradienti accennati in precedenza grazie all' introduzione di gate che suggeriscono quale informazioni del contesto mantenere e quali ignorare. Si introducono delle porte (gate) che mantengono le informazioni da portare avanti e quella da ignorare. La cella LSTM (Fig.4.10) ha come input x_t , lo stato hidden precedente e un vettore c che mantiene il contesto ereditato dall' istante $t - 1$ precedente, mentre in output viene dato c_t per l' istante successivo. Il vettore c tiene traccia dell' informazione a lungo termine. Un nodo è strutturato con una matrice dei pesi W che viene decomposta in i, f, o, g che tengono traccia di diversi tipi di informazioni:

- **Input Gate (i)**: quanto del contesto attuale si vuole mantenere.
- **forget Gate (f)**: quanto si vuole cancellare dal contesto precedente.
- **Output Gate (o)**: quante informazioni sono utilizzate per l' h_t corrente.
- **New Memory Cell (g)**: cosa si vuole scrivere.

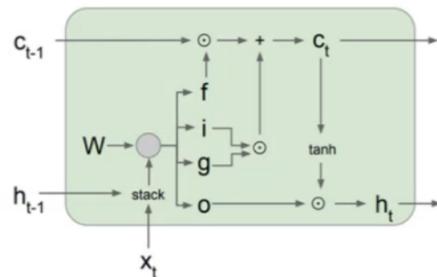


Figura 4.10: Struttura di una cella LSTM

Più nello specifico, la sezione convoluzionale dell' Encoder è seguita da un layer pienamente connesso che riduce le features visive alla dimensione dell' *embed size*, ossia alla dimensione degli embeddings. Il decoder, invece, deve processare queste caratteristiche in uscita trasformandole in parole. Prima di far partire l' addestramento, bisogna eseguire un pre-processing sulle caption a disposizione, ossia bisogna trasformarle con un processo di *Tokenizzazione*: ossia creare un dizionario da tutte le parole tratte dalle didascalie di training rendendole dei "token" che devono rappresentare una word in modo univoco, ogni token deve essere rappresentato da un intero. Inoltre, ciascuna frase di training tokenizzata sarà estesa con $< start >$ e $< end >$, rispettivamente messi all' inizio e alla fine della caption. Durante l' addestramento, ogni caption sotto forma di un insieme

token deve passare sequenzialmente su tutte le celle della LSTM. Prima di ciò, un modulo dovrà convertire ciascuna parola (quindi ciascun indice univoco) nel suo embedding corrispondente di dimensione *embed size*. A questo punto, avendo in input una coppia immagine-didascalia, Encoder e Decoder vengono allenamenti congiuntamente; l' immagine viene ridotta ad un vettore latente che ha la stessa taglia dell' embedding corrispondente alla prima parola della didascalia che entrerà nella prima cella della Long Short-term Memory. L' allenamento andrà avanti finché non verranno associate caratteristiche visive con la semantica delle parole creando un sistema di Image Captioning.

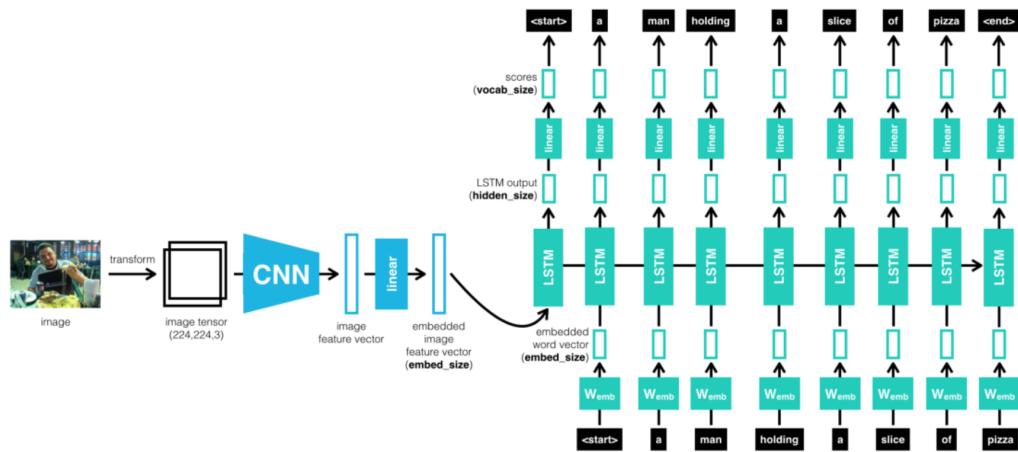


Figura 4.11: Architettura CNN-LSTM per il Task di Captioning

4.4 Transformer

Uno svantaggio delle reti ricorrenti è la difficoltà che hanno nel catturare informazioni tra i vari moduli. Inoltre, per questi modelli non sono strutturati per parallelizzare la computazione; i Transformers, invece, sono capaci di superare a queste problematiche. Il cuore di questa architettura sono i moduli *Self-Attention*, i quali catturano le informazioni contestuali tra una sequenza di dati. Questo modulo prende in input N vettori x_1, \dots, x_n e lo trasforma in una sequenza di vettori della stessa lunghezza, facendo una sorta di mapping. Più nello specifico, un Self-Attention layers è formato da una serie di nodi dove il loro input è dato dall' elemento attuale più tutti quelli precedenti. Si differisce da LSTM poiché non bisogna propagare un contesto dagli hidden layer, qui tutte le parole vengono prese in input e tutti i nodi possono anche lavorare in parallelo. Un blocco Transformers (Fig. 4.12) include un Self-Attention, dei layer feed-forward, delle connessioni residuali e dei layer di normalizzazione. L 'input x_1, \dots, x_n

viene dato in input al Self-Attention, il livello di normalizzazione riceve l' output e anche l' input del Self-Attention sommandoli; questo passaggio viene definito come residual connection, ossia passare un dato dai livelli più bassi ai livelli più alti e serve per semplificare l' addestramento. Il risultato dopo la normalizzazione viene propagato al feed-forward dove è presente anche qui una connessione residual tra l'output precedente e l'output del feed-forward che è spedito ad un altro layer di normalizzazione. il prodotto finale è una nuova sequenza y_1, y_n . Il modulo Self-Attention cattura le possibili relazioni tra le parole, qualora si volessero intercettare aspetti più profondi è possibile usare più moduli (Layer MultiHead). Una caratteristica importante sono i Positional Embeddings: se prendiamo una frase e ne cambiamo l' ordine delle parole, i risultati ottenuti sono simili a quelli sull' ordine di partenza; questo perché, a differenza delle RNN, si tiene in considerazione la posizione in cui appare la parola e viene concatenato all' embedding anche un Positional Embedding (Embedding ad ogni posizione della parola). Gli usi dei Trasnformers sono molteplici. Quello più comune è usare il modello come Autoregressive Language Model, ossia generare la prossima parola da una frase di partenza; compiti di Text Summarization, ossia ottenere riassunti da un documento; la Contextual Generation, un' attività che prevede il completamento delle frasi in funzione di un contesto fissato ed è questo il task che hanno seguito i Trasformer addestrati per questo studio.

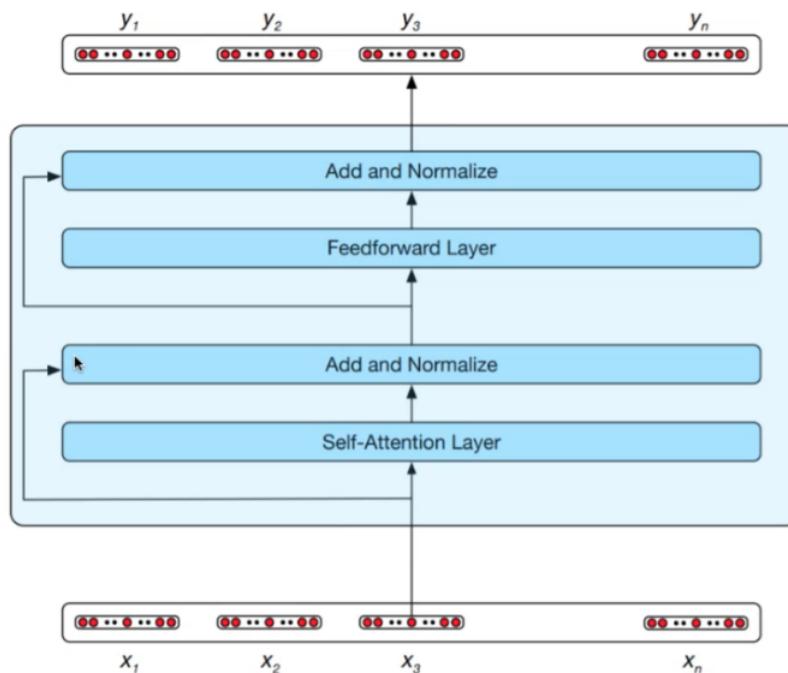


Figura 4.12: Architettura di un Trasformer

4.5 Visual Explainability

Per un algoritmo di classificazione di DL in Computer Vision, dove l' input è rappresentato da un' immagine, il processo di discriminazione è racchiuso in una "black box". In tal caso, un tipo di spiegabilità può essere rappresentato da delle heatmap, le quali hanno il compito di evidenziare le regioni più importanti per la predizione. Questo permette all'utente di capire visivamente quali pixel sono correlati con l'etichetta prevista e di decidere se il modello di Deep Learning si è concentrato su una zona ragionevole dell'immagine di input. In un heatmap (o mappe di calore), il rosso è associato a valori alti ed il blu a valori bassi. Quando un heatmap viene sovrapposta ad un'immagine, i pixel con colori caldi stanno ad indicare un alto contributo per la label fornita, per i colori freddi il contributo è prossimo a 0. Ovviamente, la mappa relativa alla classe corretta tende a concentrarsi sull'elemento in questione, mentre se si sovrappongono mappe di classi sbagliate il calore verrà distribuito con confusione sullo sfondo e zone adiacenti all'elemento. Tra i vari metodi di visual XAI, le heatmap di Vanilla Gradient e LRP forniscono delle indicazioni su zone ampie e discontinue che si espandono anche sui dintorni, stesso comportamento anche sulle classi sbagliate (vedi Fig. 4.13); la mappa relativa a Guided Backpropagation è leggermente più precisa delle precedenti ma non rileva una differenza tangibile con le etichette diverse dalla classe assegnata; LIME e Grad-CAM sono in grado di dare invece una spiegazione molto più umana e precisa, inoltre si evidenziano anche i tentativi di trovare dei pattern visivi sbagliati per le label non corrette; per RISE, le spiegazioni sono più scarne e le regioni che considera importanti sono sempre le stesse, a prescindere dall'etichetta. Tenendo conto di queste considerazioni, si è scelto di implementare Gradient-weighted Class Activation Mapping (Grad-CAM) per la classificazione di OCT. Nella sezione seguente, viene introdotto il concetto di Class Activation Maps fondamentale per la comprensione della versione più generale Gradient-weighted.

4.5.1 Class Activation Maps

Le Class Activation Maps (abbreviato in CAM) [46] sono una tecnica in grado di indicare le regioni che usa una CNN per identificare una classe nell' immagine. Sostanzialmente, l' applicazione di CAM produce delle heatmap che sono calcolate senza conoscere la predizione finale del classificatore. Rispetto ai metodi di occlusione (come ad esempio LIME), le Activation maps evidenziano una semantica più profonda che risiede nei blocchi convoluzionali. Durante la fase di

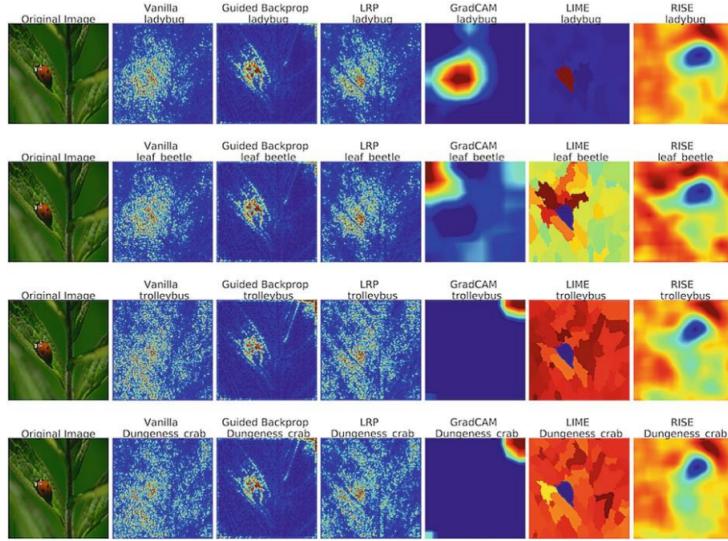


Figura 4.13: Confronto visivo delle heatmap per i diversi algoritmi di Visual XAI per la detection di una coccinella.

training di un modello, un' immagine ad una risoluzione più o meno elevata viene spinta attraverso i blocchi convoluzionali della rete; il superamento di un blocco produrrà una mappa delle caratteristiche (o più features maps, dipende dall' architettura della rete) e per questo motivo la risoluzione risultante sarà via via minore. L' ultima feature map dell' ultimo blocco verrà propagata presumibilmente verso layers pienamente connessi che servono per la classificazione. L' idea alla base di CAM è il rimpiazzo di tali layers con uno di Global Average Pooling (GAP); in questo modo, quando verrà propagato l' input, la feature map verrà "collassata" in un singolo valore scalare (o un array di scalari). Il nuovo layer dopo GAP (indicato in Fig. ??) contiene le features che in combinazione lineare con i pesi w_1, w_2, \dots, w_n portano al risultato della classificazione. Questi pesi, che verranno calibrati con l' allenamento, se moltiplicati con le corrispondenti features maps e sommati tra loro genereranno una Class Activation Map. Più formalmente, le features maps possono essere viste come delle funzioni $f_k(x, y)$ che prendono in input x e y, rappresentazione delle coordinate spaziali dei pixel, e restituiscono un valore. Il risultato dell' applicazione del GAP è definito come in 4.2.

$$F_k = \sum_{x,y} f_k(x, y) \quad (4.2)$$

Successivamente, viene fatta una somma delle k features maps rispetto i pixel, eventualmente la somma può essere divisa per il numero di pixel per fornire una media dei valori. La predizione per una determinata classe c è una combinazione

lineare tra i pesi w_k^c (che quindi dipendono dalla classe) e F_k , come mostrato in 4.3.

$$S_c = \sum_k w_k^c F_k \quad (4.3)$$

Qualora si volesse ottenere la probabilità P_c della classe c , basterebbe applicare la funzione di attivazione Softmax, come in 4.4.

$$P_c = \frac{\exp(S_c)}{\sum_{c'} \exp(s_{c'})} \quad (4.4)$$

La predizione S_c può essere riformulata sostituendo 4.2 in 4.3 per ottenere 4.5. Si faccia presente come nella nuova equazione ogni operazione è lineare, quindi possiamo spostare w_k^c nella sommatoria più interna.

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y) \quad (4.5)$$

In questo modo è possibile definire M_c come la mappa di attivazione ("Class Activation Map" in Fig. 4.14) per la classe c come la sommatoria:

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \quad (4.6)$$

La class Activation Map risultante appare con una risoluzione più bassa rispetto l'immagine di partenza, per questo motivo bisogna aumentare la risoluzione applicando un interpolazione bi-lineare in modo da poter poi overlapparla sull'immagine di input. Solo in questo modo si potranno chiaramente le chiazze di calore con cui la CNN determina la sua predizione. Ovviamente, la sostituzione con il layer GAP porta ad un cambiamento dell'architettura della rete ed quindi è necessario riallenarla per visualizzare una CAM.

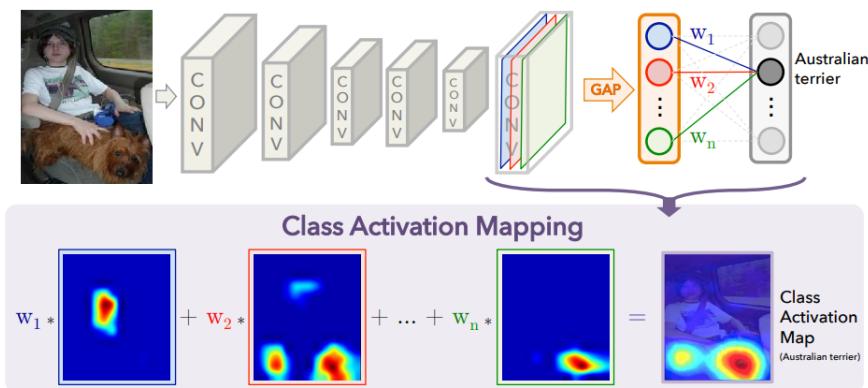


Figura 4.14: Esempio di Class Activation Maps

4.5.2 Grad-CAM

Con CAM, è stato necessario cambiare l' architettura con un Global Average Pooling per poter svolgere i calcoli in modo lineare. Con le mappe di attivazione Gradient-weighted è possibile mantenere gli strati densi della classificazione senza la necessità di linearizzare il problema. Anche in questo caso, l' input passa attraverso blocchi convoluzionali con cui ottenere delle features map. Alla fine del blocco, le features possono essere eventualmente indirizzate a diversi task, come la classificazione, l'Image Captioning o Question Answering. Nel caso di un Image Classifier, alcuni nodi degli strati pienamente connessi fino ad indicare la classe target. Il primo step da affrontare è il calcolo del gradiente della classe predetta c rispetto l' ultima feature map A^k , ossia $\frac{\partial y^c}{\partial A_{ij}^k}$, dove k è il numero di features map; questo poiché si vuole calcolare quanto linearmente ogni singolo pixel dell' immagine di input contribuisce rispetto A^k . Il calcolo porta a dover risolvere più gradienti per la classe c dei quali si farà la media (si applica un un Global Average Pooling) sulle dimensioni U e V per ottenere dei pesi α_k^c (4.7).

$$\alpha_k^c = \frac{1}{UV} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (4.7)$$

Tutti i pesi α_k^c ottenuti dovranno essere combinati linearmente con le features map A^k . Infine, viene applicato al risultato la funzione di attivazione ReLU. Questa funzione reasta a 0 per i valori negativi, mentre cresce linearmente per i valori maggiori di 0, in questo modo selezioniamo solo i contributi positivi per la classificazione finale. L' heatmap generata da Grad-CAM risultante per la classe c è:

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (4.8)$$

Anche in questo caso, la risoluzione dell' heatmap è inferiore, quindi bisogna portarla alla risoluzione corretta per stenderla sull' immagine di partenza. Come accennato, Grad-CAM può essere vista come una generalizzazione di CAM. La predizione per una classe classe c può essere scritta come:

$$S^c = \sum_k \alpha_k^c \frac{1}{UV} \sum_i \sum_j A_{ij}^k \quad (4.9)$$

Si suppone $\alpha_k^c = w_k^c$ e $A_{ij}^k = f_k(x, y)$; anche in questo caso, avendo un' espressione totalmente lineare, è possibile spostare le sommatorie in qualsiasi ordine, ottenendo:

$$S^c = \frac{1}{UV} \sum_x \sum_y \sum_k w_k^c f_k(x, y) \quad (4.10)$$

Dove $\sum_k w_k^c f_k(x, y)$ equivale a $M_c(x, y)$, ossia la Class Activation Map (equazione 4.6).

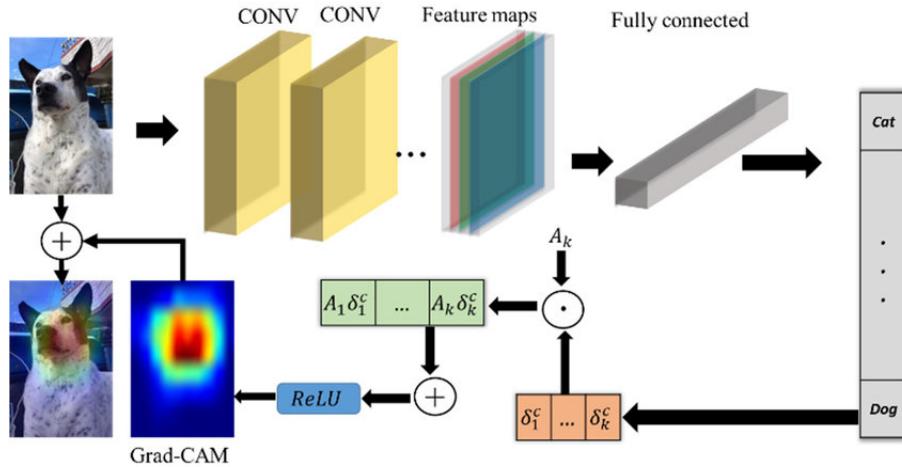


Figura 4.15: Architettura Grad-CAM

4.6 Learning congiunto

Il Multi-Task Learning (MTL) è un tipologia di apprendimento in cui le features tra task correlati vengono condivise, consentendo una migliore generalizzazione di entrambi i compiti. Il MTL è utilizzato nel Machine Learning, nel Natural language processing e nella Computer Vision ottenendo ottimi risultati; le applicazioni spaziano dalle previsioni finanziarie/economiche, dove bisogna prevedere il valore di diversi indicatori correlati, fino alla bioinformatica, cercando di prevedere simultaneamente i sintomi di più malattie. MTL si ispira all’ apprendimento naturale umano; lo sviluppo del pensiero e delle capacità nel saper discernere dipendono dall’ attuazione delle conoscenze acquisite imparando da compiti correlati; ad esempio, un neonato che interagisce con l’ ambiente impara a riconoscere i volti delle persone, la stessa abilità viene poi applicata anche nel riconoscere i singoli oggetti. È possibile vedere il multi-Task Learning come una sorta di apprendimento induttivo, il quale migliora il modello tramite un cosiddetto ”pregiudizio induttivo” tramite un modello che fa un compito ausiliario, aumentandone la generalizzazione. L’ apprendimento multi-task, spesso intercambiato in letteratura con il termine ”Joint-Learning” (allenamento congiunto) è classificato in due categorie, la differenza sta nella condivisione dei parametri dei layer nascosti. Un MTL con condivisione *Hard*, gli hidden layer sono condivisi tra le varie attività, differenziandosi solo sugli output specifici di

ciascun task. Un modello "hard" riduce notevolmente il rischio di overfitting; in effetti, più compiti stiamo imparando contemporaneamente, più il nostro modello deve trovare una rappresentazione che vada bene per tutti i compiti. Nella condivisione *Soft* dei parametri, ogni task ha il suo modello con i propri parametri; qui viene semplicemente introdotta una metrica di distanza dai parametri che vengono regolarizzati in modo da apprendere influenzandosi l' uno con l'altro. Il MLT risulta quasi sempre un vantaggio per diversi motivi. In primis, avere compiti correlati significa avere molteplici dataset, magari di natura diversa ma intrinsecamente affini, quindi vi è una sorta di Data-Augmentation implicita per i singoli task che attingono dai dati dell' altro. Il modello deve ottimizzarsi su un set di dati unificato che è molto più "rumoroso" e per un compito significa focalizzarsi sulle caratteristiche che contano davvero. Un altro vantaggio è che se l' apprendimento sul compito A risulta difficile, può comunque imparare da un altro compito B a trovare una soluzione ottimale da un' altra prospettiva catturata dall' addestramento su B (fenomeno di *intercettazione*). Inoltre, il joint learning porta ad un alto livello di generalizzazione per un modello con cui possono essere trattati efficientemente anche altri nuovi task futuri purché provengano dallo stesso ambiente.

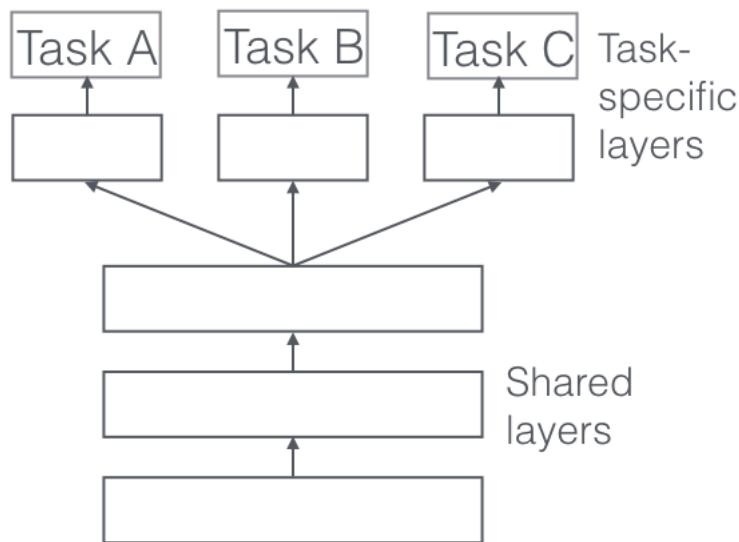


Figura 4.16: Condivisione dei layer tra diversi compiti in un' architettura Multi-Task.

Capitolo 5

Estrazione e lavorazione dei dati

5.1 Panoramica

Acquisire, pulire e organizzare i dati è uno step fondamentale per condurre una buona ricerca. Quasi mai il dato grezzo a disposizione risulta adeguato, bensì bisogna applicare trasformazioni che standardizzano l' input e che siano comprensibili per il modello. Oltre ciò, bisogna far fronte anche alle limitazioni dovute agli strumenti hardware e tarare di conseguenza le metodologie per il trattamento dei dati di addestramento. Il capitolo descrive inizialmente il dataset su scansioni OCT ed illustra la pipeline di processing sulle immagini. Tecniche riduzione di dimensionalità e undersampling sono applicate per costituire tre dataset per l' Image Classification; ognuno pensato per computare soluzioni con diversi gradi di explainability. La seconda parte verte sulla fase di Preparation dei dati per l' Image Captioner. Si discutono le assunzioni per la costruzione delle didascalie di allenamento cercando di raggruppare un insieme di marker che individuino delle caratteristiche riconoscibili e significative nell'immagine per la diagnosi testuale. Un ulteriore menzione è rivolta all' allenamento multi-Tasking che necessita di un Dataset che è un ibrido tra quello della classificazione e dell' Image Captioning, poiché deve fornire una tripla di item alla volta per condurre un allenamento congiunto. Successivamente verrà descritta la creazione del Dataset per le didascalie e di come poterlo ampliare. Nell'ultimo paragrafo viene mostrato il procedimento di estrazione e formattazione delle fonti per l' allenamento di GPT-2 verticalizzato al dominio dell' oftalmologia.

5.2 Dataset per l' Image Classification

5.2.1 Reperimento immagini

Il dataset di immagini OCT con cui sono stati condotti gli studi in [18], scaricabile anche da <https://www.kaggle.com/>. Contiene al suo interno 84.495 immagini provenienti da retine di soggetti adulti acquisite da ospedali, università e istituti di ricerca nel periodo 2013 - 2017. Alcune scansioni appartengono a stessi pazienti che hanno rifatto l'esame nei mesi successivi, in tal modo è stato possibile anche catturare l' andamento degli stadi di una malattia sia in negativo e sia in positivo tramite l' aiuto di terapie specifiche. L' intera raccolta ha seguito un processo di labelling stratificato a 3 livelli: una prima cernita delle scansioni è stata effettuata dagli studenti di medicina che hanno passato l' esame di diagnosi su OCT, il loro compito è consistito nell' escludere scansioni invalide o di scarsa qualità. Le immagini che hanno superato il primo controllo sono state analizzate da 4 oftalmologi (rappresentano il secondo livello) che hanno etichettato ogni OCT con una classe tra NORMAL, DME, DRUSEN, CNV in modo indipendente. Il terzo era composto da due specialisti con molti anni di esperienza sulle malattie retiniche che hanno validato tutte le immagini del livello precedente, finché non sono state rese disponibili al pubblico. Il dataset è già diviso in una cartella TRAIN ed una cartella TEST, ognuna ripartita a sua volta in altre 4 sottocartelle divisa nelle quattro classi.

5.2.2 Preprocessing delle immagini

Prima di applicare qualsiasi algoritmo di machine learning è buona norma pre-elaborare i dati grezzi in maniera che il training non risulti troppo ostico, oneroso e che non porti a pessimi risultati. Anche per l' elaborazioni delle immagini si è usato *PyTorch*, in particolare il package *Torchvision* che mette a disposizione comuni funzioni di Computer Vision per la trasformazione delle immagini. Un primo punto cruciale su cui soffermarsi è la dimensione delle immagini di training: non avrebbe senso dare in input OCT altamente definite, l' informazione necessaria per distinguere una malattia da un' altra potrebbe risiedere in una manciata di pochi pixel in una regione particolare. Inoltre, sovraccaricare il passaggio di un input del genere comporterebbe un sovraccarico della memoria in quanto le reti convoluzionali dovrebbero far scorrere delle *sliding windows* di filtri su una superficie troppo ampia. In conclusione, bisogna renderle in un formato ai limiti delle proprie capacità computazionali (RAM o GPU); da precisare che

il carico da computare non solo dipende da ciò, ma anche dalla dimensione dei modelli di addestramento e dal numero di immagini da fornire in input. A tal proposito, è stato definito il datasetA che è riservato solo al task di classificazione con immagini di dimensione 100x100 e il datasetB con immagini 224x224 per modelli di classificazione con explainability visiva. Un terzo dataset (datasetC) 224x224 è stato riservato per le immagini per l' architettura congiunta, le quali sono anche accompagnate da didascalie. Ovviamente, tutte le immagini devono essere standardizzate con un unico processo comune, devono avere la stessa dimensione, la stessa trasformazione e convertite in tensori per permettere l' elaborazione da parte delle CNN. Quest' ultime sono allenate su precise tipologie di immagini, ad esempio, le Densenet hanno raffinato i propri pesi su immagini a colori dal dataset ImageNet e, affinché si possa applicare del transfer learning sul problema in questione, bisogna alimentare il nuovo modello sperimentale con immagini OCT a 3 canali (RGB).

Le scansioni di [18] sono fornite su scala di grigi, quindi su un solo canale, ma grazie al modulo Image della libreria Pillow viene duplicata l' immagine su canali RGB. Una trasformazione comune che è stata applicata è la *Normalizzazione* che generalmente porta a velocizzare l'addestramento poiché i gradienti agiscono in modo uniforme per ciascun canale: per ogni immagine viene sottratta la media dei valori di ciascun canale divisa la sua deviazione standard; questo è possibile con la funzione *Normalize* di *Torchvision*.

Per ridurre la dimensionalità delle immagini OCT, si è scelto di utilizzare la Principal Component Analysis, la quale permette di riassumere l' informazione delle variabili di un dataset, quindi ridurre la dimensionalità dei dati. PCA rende il modello meno espressivo e riducendo il numero di features può in molti casi ridurre l' overfitting (ma non necessariamente). In sostanza, si applica una trasformazione lineare che proietta le variabili correlate in un nuovo sistema cartesiano (vedi Fig. 5.1) in cui le nuove variabili non sono correlate, quest' ultime sono dette *componenti principali*. Dal set delle componenti ricavate si può ricostruire tutta l' informazione di partenza; è possibile conservare solo poche componenti che detengono la maggior parte dell' informazione, alleggerendo lo spazio occupato dai dati. La libreria *decomposition* di *scikit-learn* implementa PCA e permette di specificare il numero di componenti da mantenere. Effettuando vari test in modo iterativo, un buon compromesso si aggira intorno alle prime 25 componenti affinché l' immagine OCT non perda le informazioni chiave per la classificazione ed è stata applicata per l' intero datasetA.

Prima di effettuare un allenamento sui dati è necessario che il numero di

samples per ogni classe sia bilanciato. In caso contrario, il modello rischierebbe di focalizzarsi troppo sulla classe prevalente e ignorare le altre, portando sempre alla stessa predizione. Esistono vari meccanismi per fornire una distribuzione più equilibrata. Tra questi i più comuni sono l' oversampling che provvede all' aumento di campioni per la classe con meno elementi con alcune tecniche, come il metodo *SMOTE (Synthetic Minority Over-sampling TEchnique)* che sintetizza nuove istanze perturbando le osservazioni della classe minoritaria; al contrario, l' undersampling ridimensiona la classe maggioritaria prendendo solo un sottoinsieme di campioni. In questo caso, i modelli di questa ricerca hanno già abbastanza scansioni OCT da elaborare se si tiene conto dei limiti hardware a disposizione, si è fatto quindi undersampling generale dal dataset di partenza portando tutte le classi allo stesso numero di immagini, questo valore per motivi pratici è stato definito *IMG4CLASS*. Il datasetB, riservato per la explainability visiva ha IMG4CLASS pari a 1000; il datasetA, usato solamente per la classificazione e avendo immagini di dimensione ridotta e alleggerite con PCA ha invece 6000 immagini per classe; il datasetC, dedito all' explainability visiva e testuale ma per un' architettura disposta per l' allenamento congiunto, ha solamente un IMG4CLASS di 200, limitato dalle caption a disposizione. Si riporta la tabella riassuntiva (Tab. 5.1) dei 3 dataset di training.

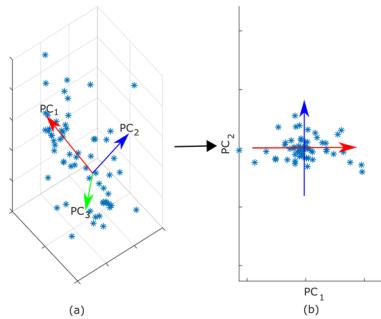


Figura 5.1: Rappresentazione visiva della tecnica PCA

Name	IMG4CLASS	WxH	PCA	Vis. XAI	Text. XAI	Joint
DatasetA	6000	100x100	Sì	No	No	No
DatasetB	1000	224x224	No	Sì	No	No
DatasetC	200	224x224	No	Sì	Sì	Sì

Tabella 5.1: Panoramica dei 3 Training Set per l'addestramento.

5.3 Data Preparation per il Captioning

Il processo di Captioning richiede la preparazione di un dataset a sé: ogni input per il training del modello CNN-LSTM deve essere una coppia composta da un' immagine ed una didascalia. Quest' ultima dovrebbe fornire una descrizione testuale di feature visive dell' immagine ed essere concorde con la definizione della classe appartenenza. In sostanza, è necessario descrivere le immagini per l' allenamento in termini di particolari marker biologici che identificano univocamente la malattia.

5.3.1 Biomarkers

Il numero di bio-marker deve essere non troppo ristretto (rendendo le caption banali) e nè troppo elevato; l' inclusione di troppi markers genererebbe caption troppo lunghe e complicherebbe l' associazione di elementi visivi con le parole. Circa una decina di markers e descrizioni che oscillano tra le 15 e 20 parole di lunghezza massima sono un ottimo trade-off, da raffinare anche in funzione di quanto si voglia bilanciare il carico informativo rispetto all' accuratezza sintattica e semantica che si vuole ottenere. Per ogni malattia trattata si è fatto riferimento a lavori presenti in letteratura per capire per ognuna quali sono i pattern visivi più ricorrenti.

Markers per DRUSEN: in [35] si dà una panoramica su tutte le tipologie di drusen, definite in termini di forma, contenuto, livello di integrità e riflettività nell' imaging. C'è da precisare che le proprietà di una drusen non sono verificabili in termini assoluti: ad esempio, la riflettività dei tessuti della retina è dovuta a particolari condizioni di luce nel momento della scansione o taratura dello strumento, così come non è possibile risalire alla grandezza in termini di micrometri avendo solamente un' immagine. La scelta più sensata è stata ricadere sulla forma e sulla loro posizione rispetto il livello dell' RPE, sia perché può risultare più facile individuarne la differenze e anche per la possibilità di dare un ulteriore sub-categorizzazione. Più nello specifico, si parla di *Soft Drusen* (**SOFT**) quando sono presenti depositi sopra l' epitelio pigmentato retinico e hanno una forma tondeggiante, in genere sono più grandi (definite anche come *Large Drusen*) e si differenziano dalle *Hard Drusen* (**HARD**) che hanno un aspetto cosiddetto a "dente di sega" e sono più appuntite, si associano anche alla definizione di *Cuticular Drusen* (**CUTICULAR**) (Figura 5.2) . In [11] si evidenzia anche la presenza di depositi drusenoidi che compaiono sopra l' RPE

ma sotto lo strato subretinico, per tal motivo vengono definiti anche come *Reticular Pseudodrusen* (**RPD**). Un ulteriore marker è il distacco dell' epitelio pigmentato (**PED**) dalla membrana di Brunch.

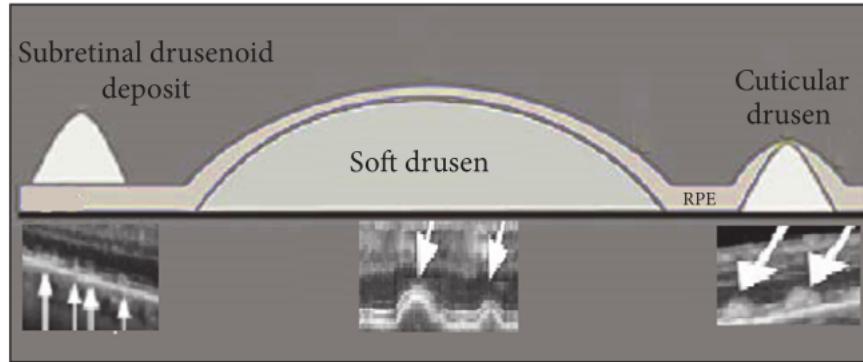


Figura 5.2: Da sinistra verso destra: reticular pseudodrusen, soft drusen, hard drusen.

Markers per CNV: la neovascolarizzazione coroideale, essendo una progressione dell' AMD secca, può contenere tutti i marker drusenoidi appena descritti. CNV, nella sua forma occulta del fluido può elevare l' RPE fino ad assumere una forma segmentata con bordi irregolari formando un distacco fibrovascolare dell' epitelio pigmentato retinico (**FVPED**) e può essere accompagnato da fluido sotto-retinico (**SRF**). Nella sua forma classica, i vasi proliferano attraverso l' RPE provocando una lesione vascolare con la fuoriuscita di materiale iperriflettente (Subretinal Hyperreflective Material, in breve **SHRM**), come descritto anche in [23]. Non si escludono la presenza di liquido intraretinico (**IRF**) e a forma di cisti (**IRC**).

Markers per DME: l' edema maculare diabetico non presenta anomalie a livello di RPE ma ha complicazioni a livello retinico. Per la selezione di pattern tipici (Figura 5.3) si è fatto riferimento a [41], dove si illustra un sistema di segmentazione delle tipologie di DME; Cystoid Macular Edema (CME) se sono presenti cisti intraretiniche (**IRC**), distacco sieroso retinico se è presente **SRF** e *Diffuse Retinal Thickening* se c'è un ampio accumulo di fluidi retinico (**IRF**), specialmente negli strati inferiori.

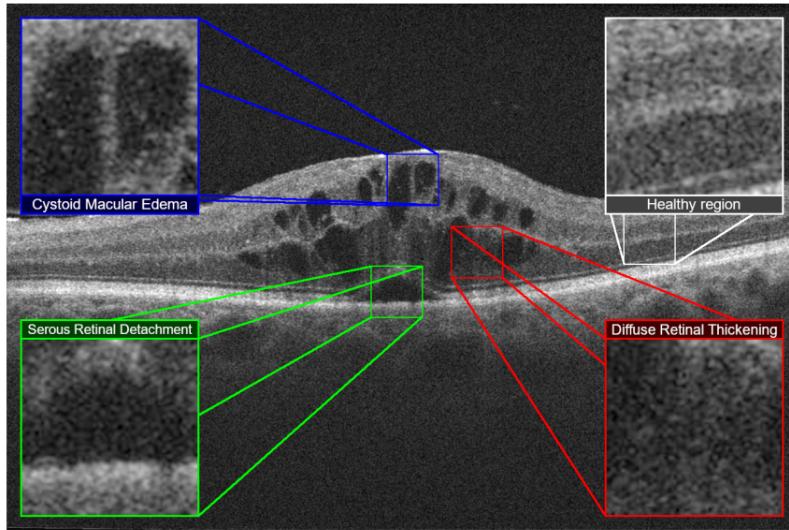


Figura 5.3: Patterns del Diabetic Macular Edema.

Conclusioni sui markers esistono ulteriori elementi che possono comparire nelle scansioni OCT ma non sono caratteristiche che discriminano le classi trattate, come hyperreflective foci, geographic atrophy e disturbi del vitreo. Un paper con cui compararsi è [21]; gli autori propongono una segmentazione automatizzata di alcuni biomarkers prefissati, tra questi (SRF, IRF, Drusen, RPD, IRC, FPED) vengono identificati come segni dell' AMD. Sempre in [21], si implementa un Decision tree che classifica CNV, DME e DRUSEN con la presenza o meno di quest' ultimi.

5.3.2 Creazione del trainset

Una volta raccolto tutti gli elementi, si passa alla stesura delle didascalie per il training. Ogni immagine deve essere descritta con frasi diverse che semanticamente devono concordare, magari ognuna con qualche sfaccettatura del contenuto con un grado di generalizzazione più o meno ampio rispetto la frase di partenza. Questo è possibile anche con l' aiuto di sinonimi dei tessuti della retina e delle patologie che estendono ulteriormente il vocabolario, ad esempio in DME la presenza di **SRF** denota anche un *Neurosensory detachment* oppure l' assenza di **IRC** ma con liquido intraretinico viene definito come *Diffuse non-cystoid Edema*. Altre diversificazioni sulle frasi sono frutto di tecniche descritte nella Sezione 6.3.3. Per le classi DRUSEN, CNV e DME sono state "markerate" 200 immagini per il training e per ognuna di esse è stato effettuato un processo di generazione di 4 caption diverse; le NORMAL non seguono lo stesso iter in quanto non

esistono dei marker precisi, nonostante ciò si sono prodotte anche qui 4 caption per le 200 immagini di pazienti sani. Il dataset di training risultante è composto da 800 immagini, ciascuna replicata 4 volte per ogni sua caption poiché l' architettura CNN-LSTM prende in input una coppia immagine-didascalia alla volta, per un totale di 3200 istanze. Per quanto concerne l' allenamento congiunto dei Task, non solo bisogna fornire al DataLoader immagine e description ma anche la relativa classe. Per questo motivo, si è stato svolto un lavoro di matching in modo da estendere il CaptionSet di partenza in un dataset dove ogni stanza restituisce la tripletta [Classe, Immagine, Caption] (Figura 5.4).

CLASS	filename	description
DRUSEN	DRUSEN-1083159-3.jpeg	This is soft drusen also pointed deposits tend to protrude into the RPE layer
DRUSEN	DRUSEN-1083159-3.jpeg	This is soft drusen and sharped debris causing irregular contour of the RPE layer
DRUSEN	DRUSEN-1083159-3.jpeg	soft drusen also pointed tiny accumulations leading to elevations at the RPE level ascertained
DRUSEN	DRUSEN-1083159-3.jpeg	rounded light-colored debris under RPE layer and then sharped light-colored debris causing RPE elevations observed
DRUSEN	DRUSEN-1071961-8.jpeg	here looks hard drusen and dome-shaped lumps of accumulations causing elevations at the RPE level
DRUSEN	DRUSEN-1071961-8.jpeg	The picture highlight hard drusen and dome-shaped light-colored deposits causing elevations of the RPE layer
DRUSEN	DRUSEN-1071961-8.jpeg	hard drusen also soft drusen noticed
DRUSEN	DRUSEN-1071961-8.jpeg	soft drusen and cuticular drusen recognized
DME	DME-778975-75.jpeg	The picture shows intraretinal cystoid fluid and serous retinal detachment denote the presence of DME mixed Type
DME	DME-778975-75.jpeg	The image shows intraretinal cystoid fluid also detachment in the subretinal space denote the presence of DME mixed Type
DME	DME-778975-75.jpeg	subretinal fluid and then intraretinal cysts distinguished so it may be DME mixed Type
DME	DME-778975-75.jpeg	subretinal fluid and then formation of cysts in the central part of the retina shown reminding DME mixed Type
DME	DME-778975-73.jpeg	Presence of intraretinal cysts also accumulations fluid between the sensory retina and the RPE typical of DME
DME	DME-778975-73.jpeg	There is fluid collects in the subretinal space and intraretinal cystoid fluid so the patient has DME mixed Type
DME	DME-778975-73.jpeg	detachment under the retina and then intraretinal cystoid fluid ascertained so it could be DME mixed Type
DME	DME-778975-73.jpeg	fluid collects between the sensory retina and the RPE also intraretinal cysts distinguished as it presents DME

Figura 5.4: Dataframe per l' allenamento congiunto.

5.3.3 Data augmentation su NLP

La Data Augmentation comprende un insieme di metodologie per la generazione di nuovi campioni tramite trasformazioni applicate a quelli di partenza in modo da poter lavorare su un dataset più fornito con l' intento di migliorare le prestazioni. Questo studio non esegue Data Augmentation sulle immagini OCT, è possibile però attuare una procedura simile in campo NLP. Questo task è utile per aiutare la generazione di nuove didascalie per un Image Captioner. Le tecniche più conosciute sono:

- **Easy Data Augmentation (EDA) operations:** la tecnica più tradizionale e consiste in diverse operazioni che prevengono l' overfitting e rendono i modelli più robusti. Esse sono il *rimpiazzamento di sinonimi*, dove si scelgono casualmente N parole dalla caption (purché non siano termini che chiudono la frase) e si sostituiscono casualmente da un set di sinonimi; *L' inserimento casuale*, dove per N volte viene selezionato un sinonimo a caso di una parola nella didascalia e viene inserito al posto di una parola in mo-

do randomico; in *Random Swap*, si scelgono randomicamente due parole nella frase e vengono swappate di posto, lo si fa N volte; *Random deletion*, rimuove casualmente una parola se un valore generato p è inferiore ad una soglia predefinita per quella parola.

- **Back translation:** viene selezionata una frase di partenza in una lingua A e tradotta in un lingua B che fa "ponte". Successivamente si effettua un ulteriore traduzione nella lingua di partenza della frase appena tradotta. In questo modo si ottiene una nuova frase aumentata nel linguaggio A.
- **NLP Albulmentation** il termine fa riferimento ad un Augmentation analogo a quello che si fa nella Computer vision e consiste nel prendere un insieme di frasi da un testo e mischiarle tra loro, creando un nuovo periodo da cui estrarre nuove frasi, dopo un' operazione di rimozione di espressioni duplicate.

Tra le metodologie appena elencate, si è scelto di implementare le operazioni di rimpiazzamento di sinonimi e l'inserimento casuale della tecnica EDA per il Captionset (DatasetC).

5.4 Dataset per GPT-2

È possibile riallenare GPT-2 per uno specifico dominio di interesse, fornendo qualsiasi forma di testo come romanzi, testi delle canzoni, poesie, notizie e altro purché siano in inglese. La preparazione dell' input è definita in funzione del compito specifico con cui viene regolato GPT; questo studio segue un task di *Knowledge Feeding*, ossia di aggiunta di conoscenza al modello. La prima cosa da fare è reperire tutte le fonti che riguardano l' oftalmologia, verticalizzate alle malattie della retina. La selezione e segmentazione del testo non deve essere casuale, GPT è più efficace se le informazioni sono divise con spezzoni di frasi che raggruppano un concetto completo e finito nel modo più atomico possibile, delimitati da un tag speciale, ovvero $<\text{endoftext}>$. L' estrazione dei testi è stata fatta fatta su documenti, antologie e libri di diagnostica, una buona parte delle fonti sono estrapolate da:

- **Retina and Vitreous:** American Academy of Ophthalmology, Section 12, 2014-2015
- **Ophthalmology:** 3rd Edition, di Myron Yanoff e Jay S. Duker

- **DME EyeWiki:** "<https://eyewiki.aao.org/DiabeticMacularEdema>"

Tutte le informazioni contenute in periodi e/o capitoli che si distaccassero troppo dalla degenerazione Maculare e dalla retinopatia diabetica non sono state prese in considerazione. Il testo è stato preprocessato automaticamente con la condensazione di caratteri speciali, annotazioni e riferimenti non ammissibili per l'allenamento, inoltre, le frasi sono state divise nel modo più semanticamente atomico possibile.

```
<|endoftext|>

Choroidal neovascularization (CNV) is defined as an ingrowth of vessels and associated tissue
that usually occurs in the macular region. CNV is associated with a number of disorders, but
the most important one is age-related macular degeneration (AMD)

<|endoftext|>

Several other conditions associated with CNV include intraocular inflammation, angiod
streaks, choroidal rupture, pathologic myopia, chorioretinal scars, or chorioretinal
dystrophy. Yet, the techniques for diagnosis, and in many cases treatment, are common to any
form of CNV

<|endoftext|>

CNV related to AMD is by far the most common form of CNV, and many of the treatment strategies
and studies were developed for CNV secondary to AMD

<|endoftext|>
```

Figura 5.5: Esempio di dataset per training GPT-2

Capitolo 6

Soluzione proposta

Questo capitolo fornisce delle implementazioni sulla base dei modelli teorici introdotti nel capitolo 4 fino alla costruzione di un Tool sulla diagnosi e spiegabilità di malattie rilevabili da OCT. L’ obiettivo da raggiungere si dirama nella risoluzione di molteplici compiti che spaziano dalla classificazione, Textual XAI, Visual XAI, Reporting, rendendo particolarmente ardua la ricerca di soluzione che sia un trade-off per tutti i tasks. Per questo motivo, seguendo il principio del *Divide et Impera*, tutti i compiti sono stati trattati singolarmente in modo atomico cercando di ottenere le massime prestazioni per ciascuno. Come già anticipato nel Capitolo 5, sono stati realizzati tre dataset, ognuno con proprietà differenti e pensate per la risoluzione di un singolo compito (vedi Tab. 5.1). Con queste premesse, da un punto di vista applicativo, tutti i modelli possono essere già raggruppati insieme e resi fruibili per mezzo di un software che risponde istantaneamente alle esigenze del paziente. Tuttavia, si fa notare nel Capitolo 4 come di fronte a problemi dello stesso dominio è possibile andare oltre, l’ interconnessione tra compiti diversi aiuta a catturare degli aspetti che elementi a ”camera stagna” non sono in grado di rilevare. Questo capitolo presenta inizialmente i modelli utilizzati; alcuni di essi sono stati presenti sin dalle prime fasi sperimentali fino all’ implementazione finale. Successivamente, vengono elencate le criticità riscontrate nel trovare una soluzione ai vari compiti; infine viene descritta l’ architettura generale che risolve in toto e parallelamente i task con l’ unione di tutti i moduli. Le ultime sezioni sono improntate sulla descrizione dell’ Applicazione Web sviluppata per dare accesso al modello agli utenti finali.

6.1 Modelli utilizzati

6.1.1 DenseNet

Le Densenet (Dense Convolutional Network) [14] sono modelli che possono fare Image Classification e sono stati sviluppati per trattare il problema del vanishing gradient tramite l' utilizzo di link diretti tra i layer interni, aumentando notevolmente l' accuratezza. Nella versione 121 (definita così per il suo numero di layer), l' input è indirizzato verso layer convoluzionali di dimensione 7x7 con stride pari a 2; viene seguito da uno strato di Pooling 3x3 e da un DenseBlock. Quest' ultimo è l' elemento centrale dell' architettura: ogni blocco è composto da 6 layer di convoluzione in cui ciascuno è collegato con tutti i suoi successivi; questo significa che ogni layer riceve le feature dai livelli precedenti; queste feature accumulate vengono messe in stack. Le features map tendono a dimensioni sempre minori man mano che ci si addentra nei layers in avanti (downsampling) fino a ridursi a poche feature map essenziali. L' output del denseblock confluisce in un transition Layer: è un livello che racchiude un ulteriore livello convoluzione ed un Average Pooling. Questa catena è ripetuta più volte in funzione dell' implementazione di Densenet, dove si susseguono DenseBlock alternati da Transition Layer. Gli ultimi strati prima di ottenere l' output sono un livello GAP ed un pienamente connesso che restituisce la predizione. Un vantaggio della densenet è il rafforzamento della propagazione delle features, proprio poiché un layer L_i ha un link diretto a $L_{i+1}, L_{i+2}, \dots, L_n$ e, per il medesimo motivo si incoraggia al riuso delle features perché la conoscenza che ha imparato uno strato è tramandata immediatamente anche ad un altro. D' altro canto, i molteplici collegamenti portano ad uno spreco di memoria per mantenere features replicate più volte.

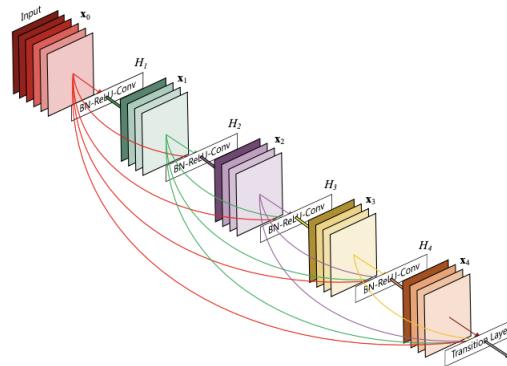


Figura 6.1: Architettura di una DenseNet

6.1.2 VGG

Una VGG (visual Geometry Group) è una rete convoluzionale ideata per l' image recognition per una vasta quantità di dati; il modello è stato addestrato su milioni di immagini provenienti dal dataset ImageNet. Le due implementazioni di spicco sono VGG-16 e VGG-19, le cifre fanno riferimento al numero di layer presenti in ciascuna. Infatti, la versione VGG-16 presenta 13 layer convoluzionali e 3 layer pienamente connessi (per un totale di 16) ed è una tra le architetture più popolari. Prende in input immagini di dimensione 224x224 che vengono trattati con filtri convoluzionali 3x3. L'unicità di VGG sta nel non puntare su un gran numero di iperparametri ma sull' avere strati di convoluzione con kernel piccoli che usano sempre la stessa dimensione di padding e del layer di Max Pooling.

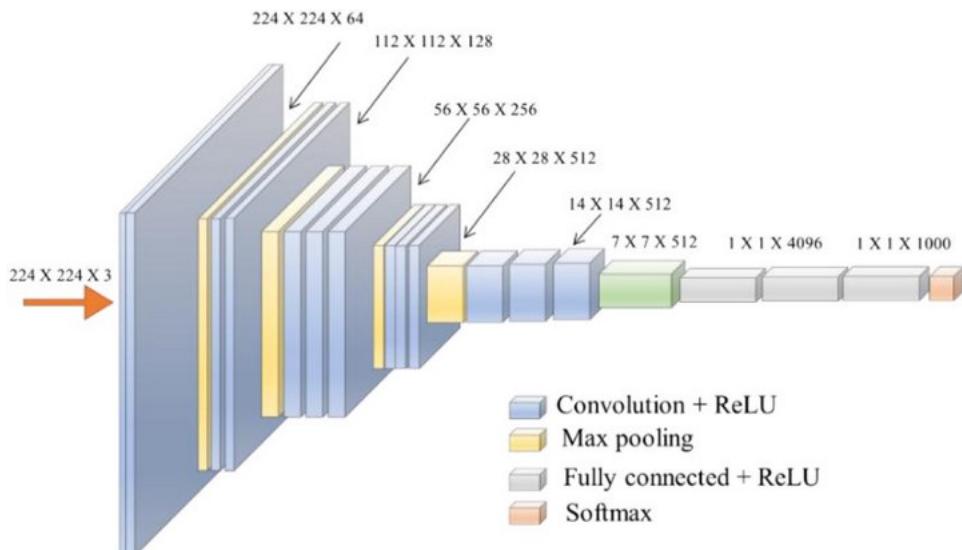


Figura 6.2: Layers di una VGG19

6.1.3 GPT-2 e GPT-Neo

Con GPT (Generative Pre-trained Transformer) si fa riferimento ad una categoria di Transformer che possono fungere da language model e che sono rilasciati da OpenAI. Questi modelli si basano sui meccanismi di attenzione descritti nel Capitolo 4 e possono essere riallenati per diversi scopi. GPT-2 è la versione rilasciata nel 2019 ed è strutturato su 1,5 miliardi di parametri nella sua versione più pesante (vedi Figura 6.3); l'architettura prevede 12 blocchi di tipo decoder messi in stack, ciascuno formato da 3 strati: un Masked Self-Attention, un Self-Attention e da una rete Feed-Forward. Un Self-Attention di tipo Masked, si differenzia dal Self-Attention classico poiché, scorrendo l'input, da una data

posizione evita di considerare anche i token successivi. GPT-2 nel 2020 è stato superato da GPT-3 con ben 175 miliardi di parametri ma non è stato ancora reso open source e il suo utilizzo solo tramite API. Tuttavia, viene rilasciato poco dopo GPT-Neo (anch'esso utilizzato da questo lavoro di tesi) ossia una replica open source di GPT-3 e ha 2,7 miliardi di parametri ed è attualmente è la versione più leggera pur essendo la più vicina a GPT-3.

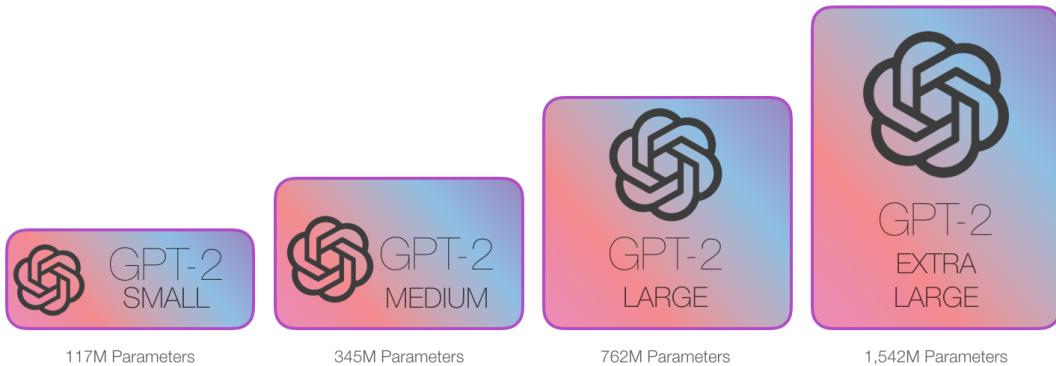


Figura 6.3: Diverse implementazioni di GPT-2

6.2 Sviluppi e criticità affrontate

Le reti appena descritte sono coinvolte in più sperimentazioni con configurazioni diverse su più dataset, principalmente limitate nella loro esecuzione da potenza computazionale non sufficiente. Sono stati effettuati più perfezionamenti incrementali degli iperparametri, questo elaborato tiene traccia solo del tuning che ha portato a migliori performance. Il numero di epoche adeguato è stato individuato visionando i grafici che riportavano l'andamento delle loss durante l'allenamento, il quale è stato arrestato in vista di nessun margine di miglioramento.

DenseNet-161 per Image Classification Una prima soluzione nel classificare le malattie retiniche è data dalla Densenet-161. Un modello che a causa della sua lunga catena di DenseBlock e delle replicazioni delle Feature Maps perpetuati ai livelli successivi rendono l'architettura potente ma con un notevole spreco di RAM (e/o GPU). Le macchine a disposizione, durante il runtime, non sono in grado di addestrare un modello così grande su un cospicuo numero di immagini a risoluzione 224x224. Per questo motivo, per affrontare la classificazione di OCT, Densenet-161 è stata allenata sul DatasetA, sfruttando i pesi preaddestrati di default e riaddestrando dei nuovi layer densi aggiuntivi per la predizione delle

classi OCT. Questo dataset contiene 24000 immagini (IMG4CLASS = 6000) con risoluzione 100x100 e sono ulteriormente alleggerite dalla tecnica PCA, affinché DenseNet-161 possa essere eseguito senza problemi.

- **Nome rete:** Densenet-161
- **Dataset:** DatasetA
- **Epoche:** 20
- **Learning Rate:** 0.001
- **Batch Size:** 32

VGG19 con Visual XAI L’ elaborato presentato non si limita al compito di classificazione. La configurazione precedente agisce in maniera ”black box” senza dare spiegazione dei risultati ottenuti. Nel capitolo 4 è stato introdotto l’ algoritmo Grad-CAM, esso è stato applicato ad una rete VGG-19 per fare predizioni e preservando anche le heatamp che forniscono mappe di calore per evidenziare aree critiche dell’ immagine data in input. Anche se trattasi di una rete pesante, riesce comunque a girare al limite delle capacità Hardware; si tenga in considerazione che l’ addestramento è stato fatto sul DatasetB con 4000 immagini a 224x224. La risoluzione elevata ha quasi saturato la memoria ma è stato necessario affinché le heatmaps prodotte da Grad-CAM fossero significative. Inoltre ci sono altre due considerazioni da fare sulle mappe di attivazione Gradient-weighted: questa volta non è possibile effettuare un transfer learning come nella soluzione precedente; freezare il blocco convoluzionale limita l’ azione di Grad-CAM, perciò è opportuno fare un fine-tuning su tutti i layers di VGG-19, riallenandola sulle nuove immagini con un learning rate più basso. Una seconda considerazione sta nel non usare la funzione di *noGrad()* di PyTorch che comporta la disattivazione dei gradienti. In alcune fasi dell’ addestramento è possibile usare questa funzione per risparmiare memoria (usata anche con Densenet-161) ma questa procedura non permette la generazione delle heatmaps.

- **Nome rete:** VGG-19
- **Dataset:** DatasetB
- **Epoche:** 30
- **Learning Rate:** 0.001
- **Batch Size:** 64

DenseNet-121 con Visual XAI Anche se DenseNet-161 è computazionalmente onerosa, le DenseNet restano comunque una categoria di ottimi classificatori. Una versione depotenziata della versione 161 è la DenseNet-121. Quest'ultima riesce a reggere un task di Image Classification su 4000 immagini 224x224 ottenendo ottimi risultati. Per quanto concerne la spiegabilità visiva, l'implementazione della Densenet in PyTorch non permette facilmente di accedere ai singoli layer per aggiungere della logica per catturare il gradiente in una data posizione: è stato necessario disassemblare i blocchi nidificati della rete e prendere l'ultimo layer di normalizzazione dall'architettura (VGG, permette invece facile accesso al blocco convoluzionale e ai layer per la classificazione). Dopo aver scomposto la rete, è stata riassemblata aggiungendo anche in livello di Global Average Pooling.

- **Nome rete:** Densenet-121
- **Dataset:** DatasetB
- **Epoche:** 30
- **Learning Rate:** 0.001
- **Batch Size:** 64

VGG-19/DenseNet-121 per l'Image Captioning Per generare caption si è implementata l'architettura Encoder-Decoder descritta nel Capitolo 4. Entrambe le reti VGG-19 e DenseNet-121 sono ideali come estrattori di features. I Decoder sono allenati congiuntamente sul DatasetC (che contiene solamente 800 immagini ma che sono fornite di 4 didascalie ciascuna) con il Decoder, rappresentato da un LSTM. Le due componenti comunicano tramite un vettore ponte di dimensione 256 (al quale si adegua la CNN), che corrisponde alla taglia degli embeddings usati dal decoder. L'addestramento è stato effettuato sempre con gli stessi iperparametri, ma ha richiesto un numero di epoche maggiore.

- **Dataset:** DatasetC
- **Epoche:** 55
- **Learning Rate:** 0.001
- **Batch Size:** 64

DenseNet-121 su tutti i task congiuntamente Il modello Multi-Task è formato da 3 moduli che lavorano in sincronia. Oltre all' Encoder ed il Decoder, si introduce un modulo che si occupa solamente di classificazione. Questo è formato solamente da un layer denso di taglia (256, 4), in cui 256 è la dimensione dell' output dell' Encoder (256 è anche la dimensione dell' input dell' LSTM mentre 4 è il numero delle classi finali). L' allenamento in modo congiunto è stato il più oneroso e ha visto margini di miglioramento fino alla 150° epoca circa.

- **Encoder:** Densenet-121
- **Dataset:** DatasetC
- **Epoche:** 150
- **Learning Rate:** 0.001
- **Batch Size:** 64

GPT-2 e GPT-Neo anche in questo caso, a causa delle limitazioni hardware, il re-training di entrambe le versioni in ambito oftalmico è stato effettuato sulla versione Medium che comprende 345 milioni di parametri. I modelli sono stati addestrati anche con gli stessi iperparametri:

- **Learning Rate:** 0.001
- **Num Step per Epoch (= $NumIstanze/BatchSize$):** 10000

6.3 OCT Report Tool

6.3.1 Architettura soluzione finale

Il fine di questo studio si concretizza con la realizzazione di un Software di auto-diagnosi intelligente. Alla base, vi sono i modelli sperimentali che operano in un'unica architettura mostrata in Figura 6.4. Questa soluzione, è stata progettata al fine di computare contemporaneamente la classificazione e la generazione delle caption. Ogni istanza del trainset è rappresentata da un' immagine, una label ed una didascalia che rispettivamente confluiscono nel blocco convoluzionale, nel classificatore e nell' LSTM. Durante l' addestramento, ogni singola immagine viene propagata nella CNN che estrae delle features che sono propagate su un layer denso di dimensione 256. Questa struttura che è comune a tutti i task usa le stesse feature per alimentare sia il decoder (LSTM) e sia un

classificatore. Sulla base della predizione della classe e dell' output del decoder vengono calcolate due loss differenti, L_{cap} ed $L_{classif}$, specifiche per task; trattasi di due loss di tipo cross entropy. Supponendo che l' istanza sia composta da (*images, captions, classImg*), le equazioni di seguito descrivono come le feature dell' encoder (rappresentato dalla funzione F_{enc}) sono necessarie sia per il decoder (F_{dec}) e sia per il classificatore (F_{class}) nel calcolo delle loss.

$$features = F_{enc}(images)$$

$$L_{cap} = L_{cross}(F_{dec}(features, captions), targetCaption)$$

$$L_{classif} = L_{cross}(F_{class}(features), classImg)$$

Viene successivamente introdotta una terza loss L_{MULTI} che sarà l' unica da minimizzare e sarà composta dalla somma delle due precedenti, quindi $L_{MULTI} = L_{cap} + L_{classif}$; in questo modo, il modello cercherà di trovare delle soluzioni che allo stesso tempo saranno pesate dal contributo che hanno nella risoluzione di entrambi i compiti simultaneamente. Quando l' allenamento sarà terminato, un' immagine di test che è data in input ripercorre tutti gli elementi appena citati finché le features non arrivano sia al classificatore e sia nell' LSTM. Nel primo caso, viene calcolato il gradiente rispetto la label predetta, in questo modo si riesce a calcolare una Heatmap Grad-CAM e quindi fornire una visual explainability; nel secondo caso, viene generata una caption di una dozzina di parole che vengono passate ai modelli GPT per generare un Report.

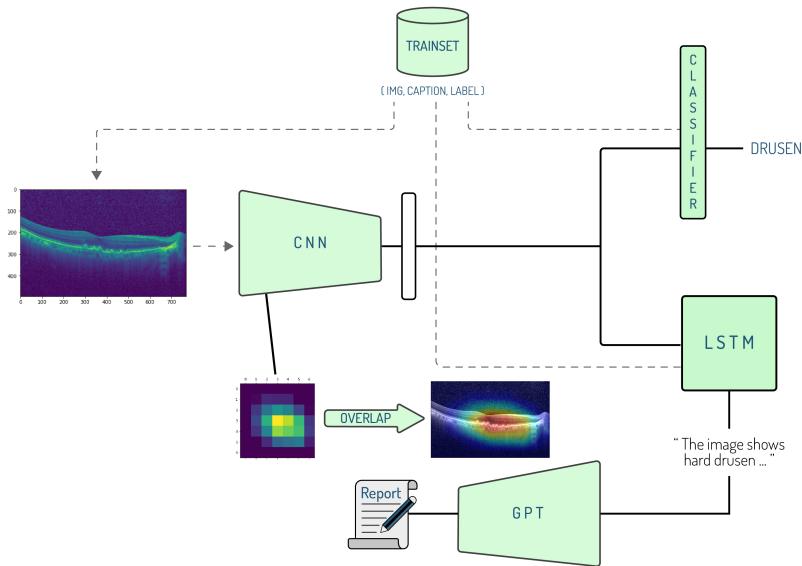


Figura 6.4: Architettura della soluzione finale.

6.3.2 Implementazione modelli

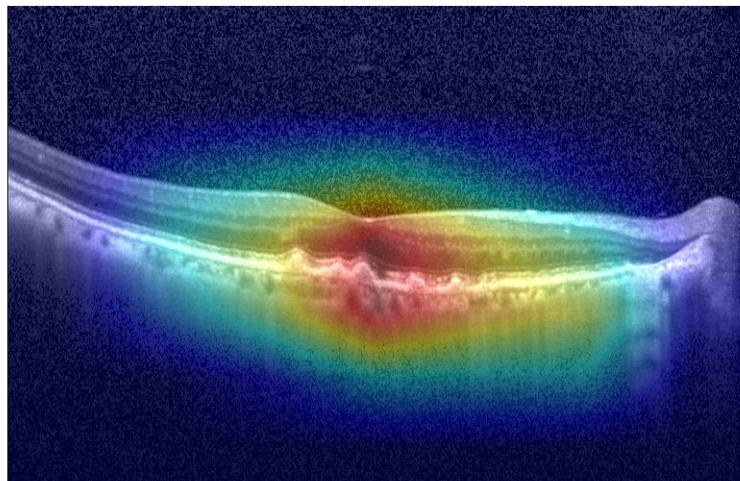
L' intero sviluppo, a causa delle limitazioni hardware delle macchine locali, è stato condotto sulla piattaforma cloud Google Colaboratory (Colab); nel suo piano base dispone di una GPU NVIDIA TESLA k80 e 16GB di RAM. Per l' implementazione si è deciso di scegliere il linguaggio Python per la facilità d' uso e per le ottime funzionalità fornite dalle librerie per la Data Science. Le librerie cruciali per l' elaborato sono state: *PyTorch* che fornisce un ampio supporto per l' intelligenza artificiale, dando la possibilità di creare modelli ex novo basati su Reti Neurali ed anche di riusare modelli pre-addestrati. Inoltre l' elaborazione dei Dataset con relative trasformazioni sulle immagini sono state rese possibili rispettivamente con package DataLoader e Torchvision; con *Numpy* è stata condotta la lavorazione di vettori, matrici e Tensori per standardizzare l' input; *Pandas* ha permesso la scrittura di tutti i Datasets e tutte le attività di pre-processing e Data Cleaning coinvolte; con *Matplotlib* ha permesso la visualizzazione dei dati e risultati.

6.3.3 Realizzazione della Web Application

Tutti i modelli realizzati necessitano di essere fruibili tramite un' interfaccia di facile accesso. Si è deciso di realizzare un' applicazione web in cui l' utente può caricare la propria scansione OCT ed eseguire i modelli di classificazione, captioning, visual XAi e report. L' applicazione è composta da una parte backend che gestisce le richieste e le immagini indicate, le elabora ed esegue i modelli su di esse; il risultato deve essere passato alla componente frontend che renderizza il contenuto sul client dell' utente. Poiché tutti modelli sono stati implementati in Python, è stato necessario utilizzare un framework web dello stesso linguaggio; tra le varie alternative, Django è sembrato il più adeguato. Trattasi di un framework altamente scalabile, sicuro e versatile ed è diventato molto popolare negli ultimi anni; in Django sono state implementate delle API REST che rispondono alle richieste da parte del client. Si precisa come i modelli eseguiti in real-time sono stati già allenati precedentemente su cloud con Colab, i pesi raffinati sono stati scaricati e richiamati dal server locale. L'applicazione lato client è sviluppata con ReactJS, una libreria basata su JavaScript che ha rivoluzionato il modo di concepire un App Web, grazie alla strutturazione delle entità che costituiscono Component atomici che possono comunicare tra loro tramite una struttura gerarchica in cui il flusso dei dati è passato dal componente padre al figlio tramite l' assegnazione di props, ossia le proprietà dell' oggetto.

6.3.4 Caso d' uso del Tool

L' OCT-Tool prevede possibili casi d' uso sia da parte del paziente che dall' oftalmologo. Facendo riferimento alle sezioni 3.2 e 3.3 sull' utilizzo dell' AI nella diagnostica e CDSS, un primo scenario vede lo specialista che usa l' App Web come un Clinical Decision Support System basato su Machine Learning con cui può interrarsi. L' oculista carica l' immagine della scansione OCT e riceve immediatamente la patologia più probabile riscontrata (o nessuna se il fondo oculare è sano). In allegato alla classificazione, viene visualizzata l' immagine appena caricata con la colorazione dei pattern sulla retina che discriminano la malattia rilevata con cui il medico può confrontarsi e fare le sue ipotesi; inoltre ha la possibilità di leggere una breve report che descrive il fondo oculare ed estende ulteriori considerazioni sulla base di testi di oftalmologia. In un secondo scenario l' utente è il paziente stesso che, dopo aver fatto l' esame OCT può caricare la propria scansione autonomamente sulla piattaforma con le quale può avere già una prima indicazione sullo stato di salute della retina e, in caso di positività ad una delle malattie, contattare immediatamente lo specialista; in tal senso, il tool riesce a far fronte alla ricezione immediata di una prima diagnosi senza necessariamente appesantire il sistema sanitario ritagliando già una lista di casi gravi che richiedono assistenza immediata. Si riporta in Figura 6.5 il risultato su un' immagine di test con nome della patologia, le aree di interesse in risalto e la descrizione.



Predicted disease: Drusen
Description: "The image shows hard drusen since pointed accumulations."

In a mild stage of the disease, visual acuity usually is good through the first five or six decades of life, although scotomata detected patients may have reduced acuity early in the course of the disease. Most patients retain good vision in this setting. The differential diagnosis of disease includes other night-blinding disorders, fundus flavimaculatus, and Bietti crystalline dystrophy.

Figura 6.5: Output del modello proposto su un OCT con DRUSEN.

Capitolo 7

Valutazione Sperimentale

Tutti i modelli presentati fino ad ora sono valutati con le metriche introdotte in questo capitolo. Per ogni soluzione si discutono i risultati dei test, cercando di decretare il trade-off ottimale per ogni task.

7.1 Metriche per l' Image classification

Il lavoro è stato valutato con gli indicatori prestazionali più comuni nel machine learning, come *accuracy*, *precision*, *recall (sensitivity)* e la *Receiver Operating Characteristic (ROC)*, usata per la classificazione binaria. Tutte le valutazioni si basano sul concetto di:

- **True Positive (TP):**

Il classificatore predice correttamente un caso positivo.

- **True Negative (TN):**

Il classificatore predice correttamente un caso negativo.

- **False Negative (FN):**

Il classificatore predice erroneamente un caso negativo come positivo.

- **False Positive (FP):**

Il classificatore predice erroneamente un caso positivo come negativo.

Questi valori posso essere combinati in una matrice di confusione (confusion matrix), ossia una tabella dove in ogni cella si indica un aspetto prestazionale del modello (Fig.7.1). La confusion matrix può essere applicata anche per classificazioni multiclasse; sulle righe sono riportate tutte le label reali, mentre sulle colonne i valori predetti dal modello; le caselle corrispondenti alla riga e colonna i -esima fanno riferimento ad item predetti con la giusta classe.

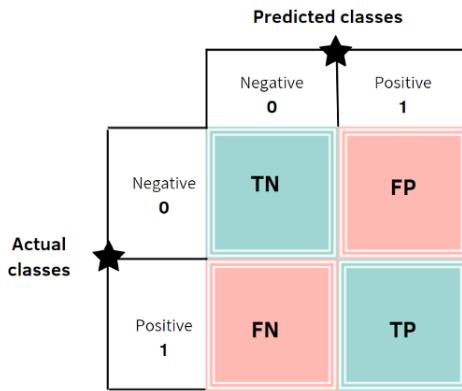


Figura 7.1: Esempio matrice di confusione

A seguire, le metriche più comuni:

- **Accuracy** È la frazione dell’insieme di dati di test su di cui il modello fornisce una previsione corretta, ha valore compreso tra 0 e 1 ed è la prima metrica da prendere in considerazione. L’ accuratezza non distingue, però, tra falsi positivi e falsi negativi, dove a volte è fondamentale farne una distinzione.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** È la frazione di casi identificati come positivi che sono correttamente positivi. Peggiora se ci sono tanti falsi positivi.

$$\text{precision} = \frac{TP}{TP + FP}$$

- **Recall (sensitivity)** È la frazione di casi identificati come positivi che il modello classifica come positivi. Questo significa che peggiora se vi sono tanti falsi negativi. È di particolare importanza se un modello è dedito alla diagnosi di una malattia, come in tal caso.

$$\text{recall} = \frac{TP}{TP + FN}$$

- **Specificity (True Negative Rate)** Considera il rapporto d’ istanze negative correttamente identificate. Il valore più alto di specificità sta a indicare una maggior presenza di True Negative e meno falsi positivi.

$$\text{specificity} = \frac{TN}{TN + FP}$$

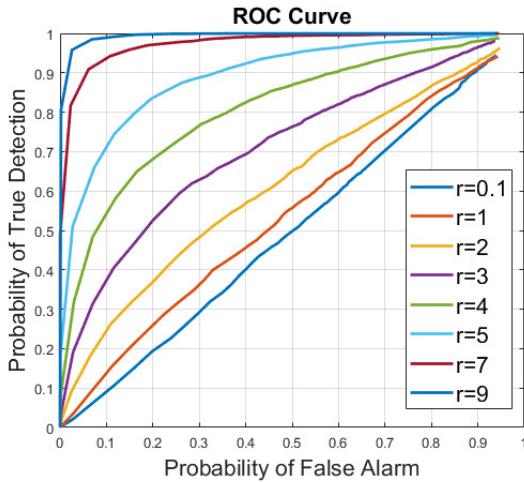


Figura 7.2: Curva ROC

- **Curva ROC** Traccia su grafico il *true positive rate*, altro modo di vedere la recall, con il *false positive rate*, $FPR = \frac{FP}{FP+TN}$. La bontà della misura è proporzionale a quanto è ampia l'area che sottende la curva che si ottiene.

In questo contesto, è importante evitare il più possibile dei falsi negativi. La mancata individuazione di una patologia potrebbe ritardare eventuali follow-up, la quale potrebbe nel tempo sfociare in degenerazioni irreversibili di maggior gravità. Ad esempio, in merito alla metriche appena definite, tendere ad una sensitivity alta significa prediligere un sistema che sappia rilevare un numero di casi con patologie con maggior facilità pur rischiando di cadere più probabilmente in falsi positivi. Un altro aspetto da considerare per questo elaborato è l'uso della curva ROC: poiché le DRUSEN sono anomalie che non compromettono la vista, sono considerate di una minor gravità; le neovascolarizzazioni e l'edema maculare sono già casistiche che portano sintomi, per questo motivo si è scelto di usare la ROC curve per confrontare una macrocategoria di minor rischio di cecità composta da NORMAL e DRUSEN con un'altra di alto rischio che comprende DME e CNV.

7.2 Metriche per l' Image Captioning

La valutazione di un problema di classificazione è basata sul confronto della label reale con quella predetta, riuscendo a dare una chiara indicazione sulla bontà del modello. Per quanto concerne i modelli di NLG, risulta più difficile fornire una metrica che risponda alla correttezza della soluzione in termini assoluti, limitarsi su aspetti sintattici e tralasciare quelli semanticci (o viceversa) può sottostimare

il modello; la descrizione di un' immagine può essere formulata con strutture grammaticali e sinonimi differenti che esprimono lo stesso concetto in maniera differente. Per questo studio, si introduce la metrica **BLEU** (Bilingual Evaluation Understudy) che misura la vicinanza tra la frase generata dal modello e frasi definite dall'uomo con un grado di accettabilità in un range tra 0 e 1. Elementi fondamentali per BLEU sono gli **N-Grammi**: un N-gram è una sottosequenza di N parole che occorrono in una sequenza di riferimento; si possono avere unigram, bigram e trigram a seconda della lunghezza di N. Ad esempio, dalla frase "*Dopo aver fatto il test, saprai il risultato*", si estraggono i seguenti N-Gram:

- **1-gram**: "dopo", "aver", "fatto", "il", "test", "saprai", "il", "risultato"
- **2-gram**: "dopo aver", "aver fatto", "fatto il", "il test", "test saprai", "saprai il", "il risultato"
- **3-gram**: "dopo aver fatto", "aver fatto il", "fatto il test", "il test saprai", "test saprai il", "saprai il risultato"
- **4-gram**: "dopo aver fatto il", "aver fatto il test", "fatto il test saprai", "il test saprai il", "test saprai il risultato"

La metrica BLEU confronta l'n-gram delle frasi generate con l'n-gram della frase target contando il numero di matching indipendentemente dalla posizione. Si definisce con **precision** il numero di parole generate che occorrono in qualsiasi frase target diviso il numero totale di parole nella frase generata. Si consideri, però, il seguente caso:

- **Frase A generata**: *il il il il il il il*
- **Frase B generata**: *il metronomo è pianoforte il sopra*
- **Frase Target**: *il metronomo è sopra il pianoforte*

Per la frase A si ha precision 2/7 (28,5%), per la frase B è il 100%; anche se la precision è elevata in questo ultimo caso, non è una frase di senso compiuto. Per risolvere tal problema si fa utilizzo della Clipped precision, che viene calcolata grazie a dei $Count_{clip}$ per ogni n-gram. Questi ultimi si ottengono con i seguenti step:

1. si definisce come *Count* il numero massimo di volte in cui un n-gram occorre in una singola frase target.

2. per ogni frase target si conta il numero di volte in cui occorre un n-gram.
Ogni conteggio è definito come Ref_i per una frase i .
3. Si calcola *Max Ref Count*, ossia il numero più alto tra tutte le Ref_i
4. Un $Count_{clip}$ è dato da: $Count_{clip} = \min(Count, MaxRefCount)$.

La **Clipped precision** sarà definita come:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{ngrams \in \{C\}} Count_{clip}(ngrams)}{\sum_{C' \in \{Candidates\}} \sum_{ngrams' \in \{C\}} Count(ngrams')} \quad (7.1)$$

In sostanza, tutti i $Count_{clip}$ per ogni n-gram sono sommati e divisi per il numero totale degli n-grammi candidati. Prima di definire BLEU, c'è un altro aspetto da considerare. Qualora la frase generata fosse troppo corta, ad esempio composta dalle parole "sopra il", la precision sarebbe pari a 1, assegnando un punteggio ottimo ad una frase che non ha senso: con l' introduzione della Brevity Penalty (BP) si disincentiva in modo esponenziale la valutazione positiva per frasi al di sotto di una certa lunghezza, più nello specifico se la lunghezza della frase generata (c) è minore della lunghezza della frase target (r).

$$\text{Brevity Penalty (BP)} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (7.2)$$

Infine, si calcola l' indice BLEU che sarà dato dalla formula:

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (7.3)$$

La presenza del peso w_n permette di poter usare diverse varianti della metrica; di default w_n è fissato a 0.25 (con N=4).

La semplice formulazione di BLEU ha diversi vantaggi, come la velocità di calcolo, l'alta usabilità in più compiti (oltre l'Image Captioning, è applicabile nello speech recognition, nella text summarization e nelle traduzioni) e l'indipendenza dal linguaggio usato. Inoltre la sua frequente adozione in letteratura permette l'immediato confronto con tanti altri lavori. Tuttavia, non viene influenzato dall'ordine delle parole. Un altro svantaggio è che il suo basarsi su match esatti fa in modo che se un sinonimo della stessa parola non è presente tra le frasi target il risultato finale sarà penalizzato.

7.3 Valutazione delle soluzioni proposte

Di seguito i risultati dei test condotti per ogni modello; ciascuno è corredata da accuracy, matrice di confusione e curva ROC o indice BLEU qualora si trattasse di un Image Captioner; per brevità solo la soluzione finale Multi-Tasking è analizzata più a grana fine, dove si riporta per ogni malattia anche specificity e recall.

Densenet-161

- Accuracy: 92,75%

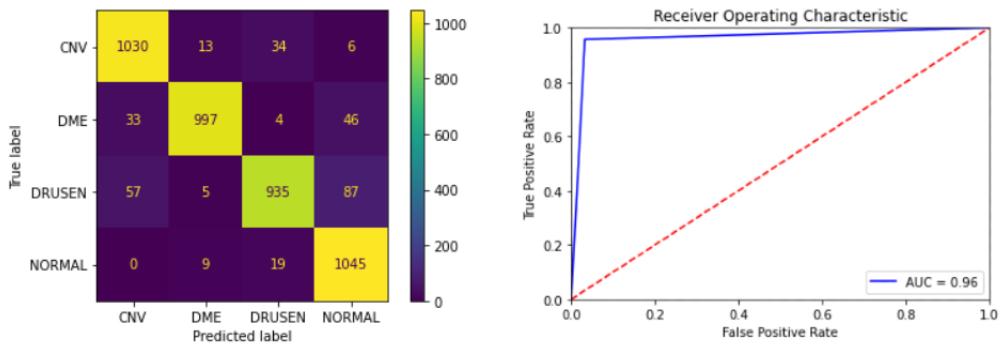


Figura 7.3: Matrice di confusione e curva ROC per Densenet-161

Il learning della Densenet-161 converge velocemente a plateau, sono necessarie solamente 20 epoche per raggiungere un' accuratezza soddisfacente. Le maggiori difficoltà della rete sono state riscontrate nel riconoscere le DRUSEN, confuse spesso con CNV o con scansioni di retine in salute, come mostrato nella matrice di confusione in Figura 7.3. Questa soluzione, alleggerita soprattutto da immagini ridotte, potrebbe essere migliorata ulteriormente con nuove immagini, ma non sarebbe comunque in grado di dare una Explainability visiva, il quale è un punto cruciale per questo studio, pertanto si è deciso di prediligere modelli alternativi. Tuttavia, Densenet-161 è un'ottima rete per massimizzare le performance per il compito di Image Classification; anche la curva ROC sottende un' ampia area sottostante con un valore AUC=0.96.

7.3.1 Modelli con Visual Explainability

In questo caso, la classificazione verte su immagini di risoluzione più alta. Grad-CAM è applicato su una versione di Densenet più leggera, raggiungendo un buon 91.67% di accuratezza (Figura 7.4).

Densenet-121

- Accuracy: 91.67%

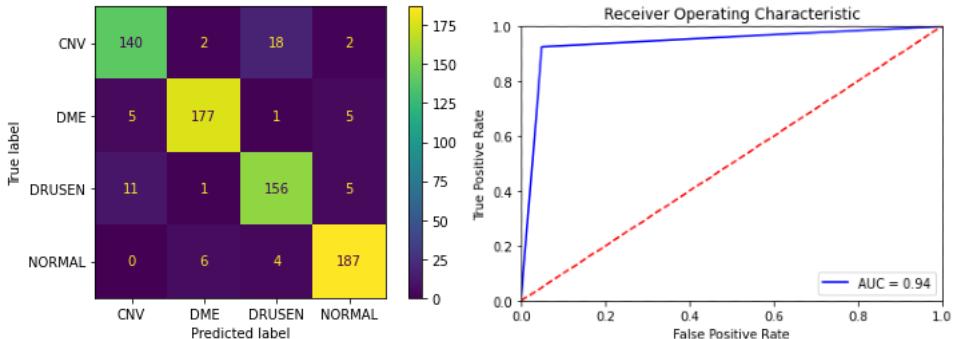


Figura 7.4: Matrice di confusione e curva ROC per Densenet-121

Anche qui la rete sbaglia più di frequente DRUSEN che sono etichettate come CNV e viceversa. Ciononostante, si faccia presente come lo sviluppo di una CNV si può manifestare con l'insorgenza di drusen: una OCT con neovascolarizzazione coroideale può comunemente contenere drusen di diverso tipo, quando l'anomalia si sposta sugli strati retinici sopra l'RPE, allora l'immagine deve essere classificata come una CNV.

VGG-19

- Accuracy: 93.06%

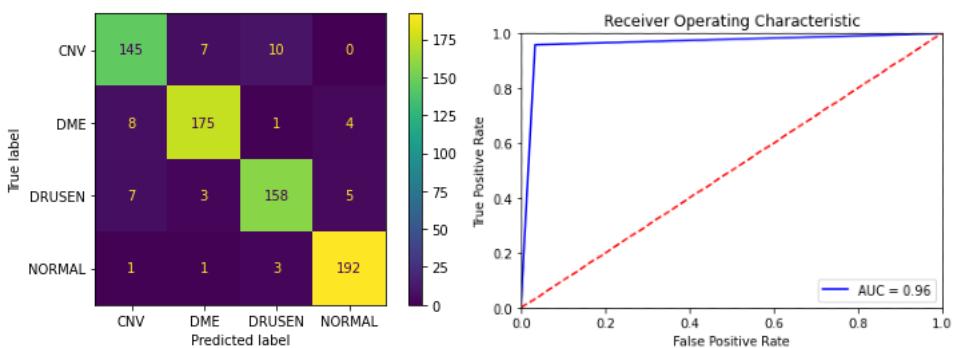


Figura 7.5: Matrice di confusione e curva ROC per VGG19

Con VGG-19 si riesce a limitare leggermente l'errore di discriminazione appena descritto (Figura 7.5) ed è fin ora il miglior classificatore in grado di dare una componente visiva. Le Figure 7.6-7.7-7.8, mostrano la mappa di calore applicata rispettivamente per una DRUSEN, DME e CNV. In ogni figura, l'immagine a

sinistra rappresenta l'applicazione di Grad-CAM su una VGG-19, Densenet-121 sulla destra, entrambe con la relativa heatmap a dimensione ridotta. Le DRU-SEN sono riconosciute per i loro tumuli che sollevano l'RPE; la DME è identificata da una ciste a livello intraretinico; il distacco fibrovascolare con bordi irregolari e liquido sottoretinico nella zona centrale (tipico di CNV) è correttamente individuato da una macchia di calore. È interessante notare come la mappa termica della densenet è più ampia, mentre quella della VGG è più localizzata; questo probabilmente è dovuto alla differente architettura degli ultimi strati convoluzionali.

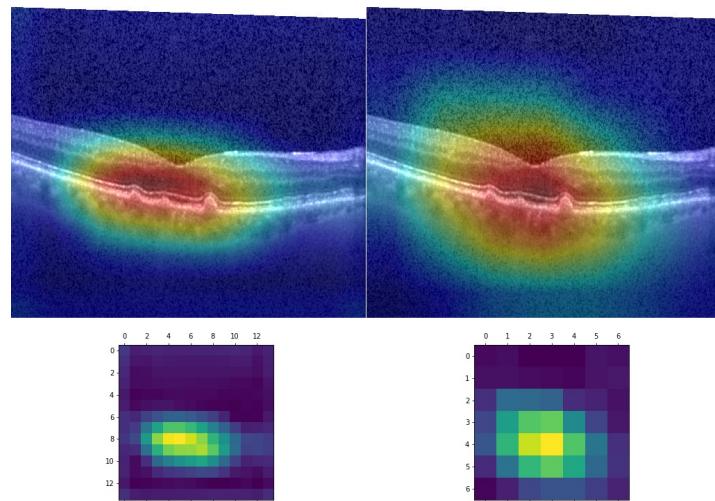


Figura 7.6: Gradcam a confronto per DRUSEN

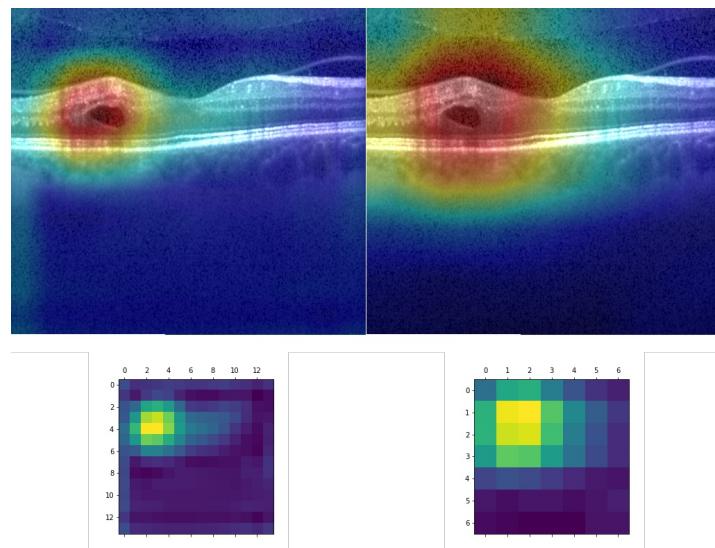


Figura 7.7: Gradcam a confronto per DME

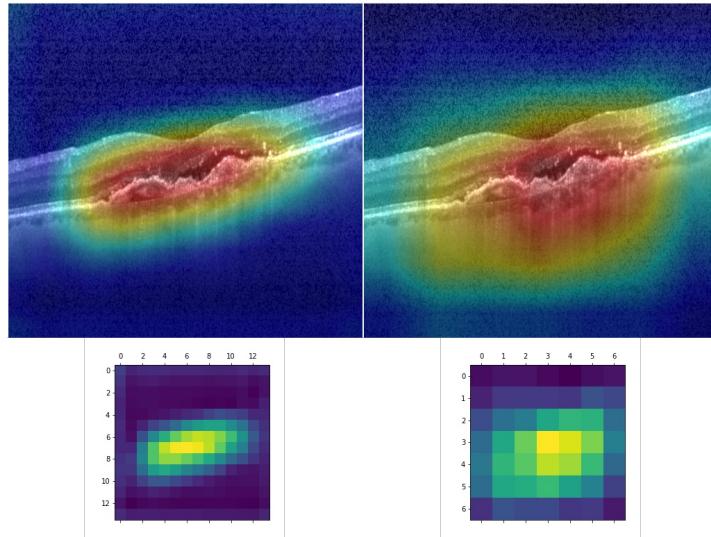


Figura 7.8: Gradcam a confronto per CNV

Un' ultima osservazione: bisogna ottenere Grad-CAM calcolando il gradiente in funzione della classe predetta con la classificazione; se questo non dovesse accadere, la mappa di calore si concentrerebbe nel trovare le caratteristiche della classe sbagliata. Ad esempio, la Figura 7.9 rappresenta una CNV; sull'immagine a destra il gradiente è stato calcolato davvero per CNV, invece sulla sinistra per l'edema maculare diabetico: pur sbagliando (poiché l'RPE non dovrebbe essere circondato da materiale iperriflettente) ha comunque spostato il focus su una ciste intraretinica, marker del DME.

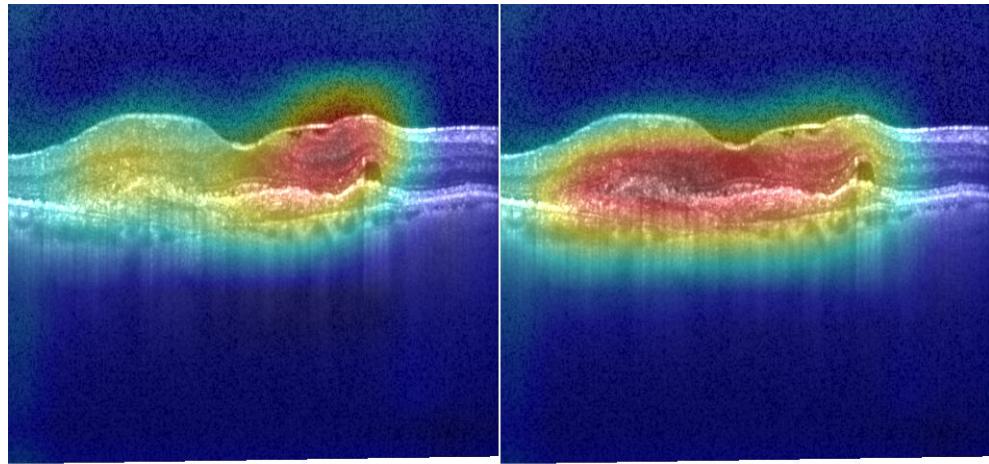


Figura 7.9: Grad-CAM calcolato sulla classe DME (dx) e sulla classe CNV (sx)

7.3.2 Modelli per Image Captioning

La valutazione della generazione di didascalie riguarda due architetture Encoder-Decoder che si differenziano nella CNN. Le reti DenseNet-121 e VGG-19 sono addestrate questa volta sul DatasetC, contenente un numero di ridotto di immagini con uno split 60%-40% per Training e Test.

Ovviamente questo modello fornisce solamente una caption e non una classe. Per questo motivo, creare un unico sistema che svolge più compiti implicherebbe inglobare modelli eterogenei che non necessariamente concordano con le loro predizioni; ad esempio l' Image Captioner potrebbe generare una didascalia per una malattia A che contraddice la predizione del classificatore che indica una malattia B. Dovendo quantificare questa discordanza, questo elaborato definisce una metrica di "Agreement" che dovrebbe dare una stima di quanto la caption generata si discosta dalla label predetta. L' Agreement Score è stato calcolato in questo modo: per ogni immagine OCT viene prodotta una didascalia dall' Image Captioner e predetta una classe dall' Image Classifier; si conta un' occorrenza qualora nella caption fosse presente almeno un termine che richiama inequivocabilmente solo la classe predetta, 0 altrimenti. Iterando questo processo per tutte le N immagini, vengono accumulate diverse occorrenze α che vengono sommate. Il tasso di Agreement sarà: $AgreementScore = \frac{\sum_i^N \alpha_i}{N}$. Ovviamente non si esclude che sia il Caption Generator che il classificatore possano concordare su una previsione sbagliata, contando comunque l'occorrenza in modo positivo. Tuttavia questa metrica non è stata definita in relazione alla correttezza delle predizioni, ma è usato solamente come indice per capire quanto i due modelli sono discordanti.

Per quanto concerne i risultati della valutazione BLEU, potrebbero sembrare al quanto bassi. Tuttavia, solitamente uno score tra lo 0,6 e 0,7 di questa metrica è già considerato ottimale. Questo perché un punteggio vicino all'1 non sarebbe realistico in quanto ci sarebbe il sospetto di un probabile overfitting del modello; sostanzialmente l' indice misura il grado di differenza tra traduzioni umane rispetto quelle automatiche, pertanto anche punteggi tra 0,2 e 0,4 possono essere considerati nella media. In ogni caso, per le reti proposte si prevede ulteriori margini di miglioramento qualora le caption di training fossero di un numero maggiore.

Si riporta nella Tab.7.1 i risultati degli indici BLEU per valori di N da 1 a 4 per DenseNet121-decoder e VGG19-decoder con Agreement Score rispetto ai relativi classificatori DenseNet-121 e VGG-19 (della sezione precedente) testate sulle stesse immagini.

Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Agreement
Densenet-121	0.438	0.189	0.0815	0.034	0.797
VGG-19	0.422	0.177	0.072	0.029	0.754

Tabella 7.1: Indice BLEU e Agreement Score per Densenet-121 e VGG-19

7.3.3 Modello Multi-Tasking

Nel task di classificazione sul datasetC, la Densenet-121 ha raggiunto un accuracy del 98,01%; osservando la matrice di confusione multiclasse presente in Figura 7.10 si nota come adesso gli errori commessi si sono ristretti nel distinguere DME che viene confusa con CNV. In tal caso sono riportati anche i valori di specificity e recall per ogni malattia, che tendono all' ottimalità nel confronto binario NORMAL-MALATTIA.

- **Accuracy:** 98.01%

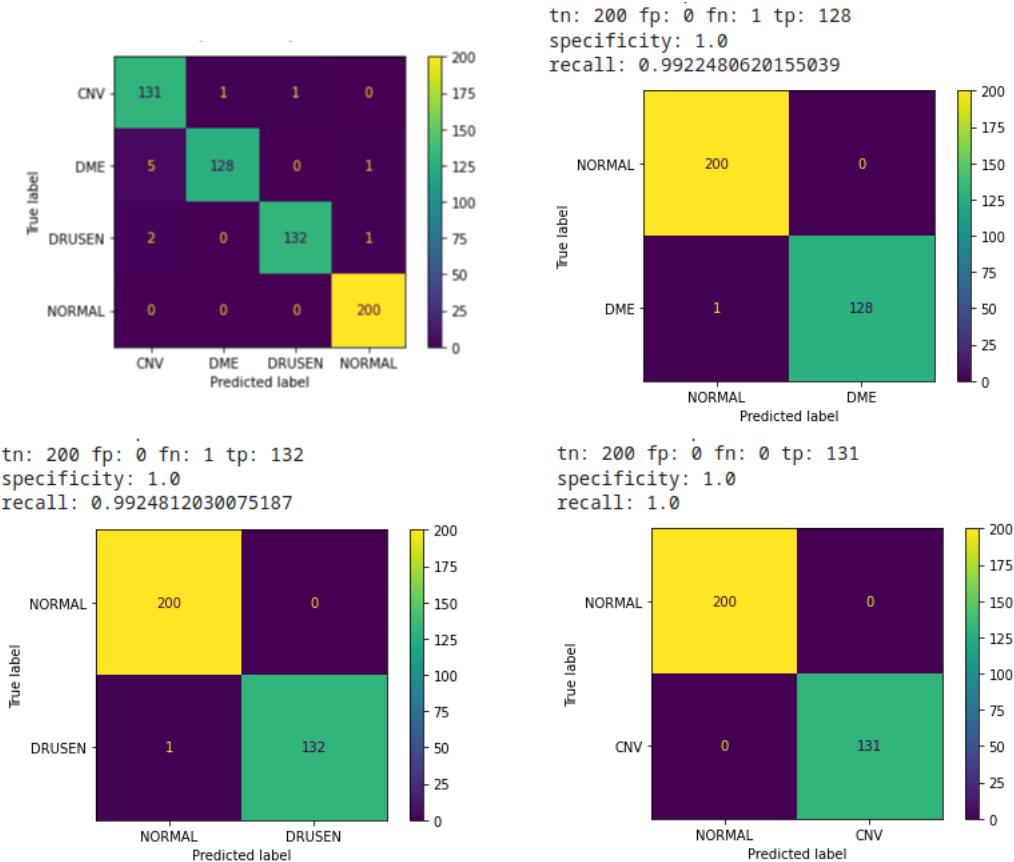


Figura 7.10: Matrici di confusione per modello Multi-Task

In particolare: il modello non è riuscito ad individuare un solo fondo oculare con drusen, per tal motivo la recall per DRUSEN è 0,99%; discorso analogo per DME, dove vi è solamente un falso negativo. L'ottimalità dei valori di specificity e recall è riscontrata invece nel riconoscimento di neovascolarizzazioni coroideali (CNV). La curva ROC in Figura 7.11 sottende quasi l'area più ampia possibile, questo significa che massimizza contemporaneamente sensibilità e specificità nel riconoscere patologie lievi (o assenti) da quelle gravi.

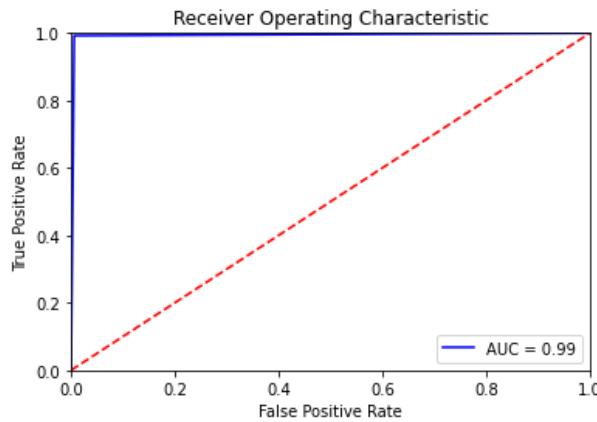


Figura 7.11: Curva ROC per modello Multi-Task

Il modello Multi-task, oltre alla predizione, genera contemporaneamente sia heatmap con Grad-CAM e sia una didascalia. Le caption sono valutate con gli indici BLEU e con l' Agreement Score; quest'ultimo raggiunge l'86% rispetto al 75% e 79% degli Image Caption Generator precedenti, probabilmente poiché tende a trovare una direzione comune alla risoluzione di entrambi i problemi, inoltre di questo meccanismo ne giova anche la valutazione BLEU che ha avuto un leggero miglioramento. Un ultimo tentativo sperimentale di questo elaborato è consistito in un Transfer Learning delle reti DenseNet-t121 e VGG-19, dedita alla classificazione sul datasetB, sui pesi dei rispettivi Encoder DenseNet-121 e VGG-19 nei sistemi di captioning visti in precedenza. Successivamente, si è verificato come il fine-tuning sulla nuova configurazione ha portato ad egualare gli indici BLEU del modello MultiTask, inoltre il tasso di Agrement risultante supera di poco quello dell'allenamento congiunto (Tab.7.2). Tuttavia non è da considerare un confronto alla pari con il sistema Multi-Task, quest'ultimo è stato allenamento su poche istanze composte da triple (*LABEL, Image, Caption*) rispetto alle precedenti reti allenate sul datasetB di sole immagini.

Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Agr.
MULTIDensenet-121	0.452	0.210	0.093	0.045	0.864
TRANSF-Densenet-121	0.459	0.209	0.095	0.048	0.870
TRANSF-VGG-19	0.455	0.204	0.081	0.037	0.885

Tabella 7.2: Indice BLEU e Agreement Score con strategia Multi-Task e Transfer Learning

7.3.4 Generazione Report

L’architettura finale del sistema diagnostico prevede che le caption siano passate al Transformer più performante tra GPT2 e GPT-NEO. Poiché le metriche da applicare su questi modelli sono ancora ambigue, l’unica valutazione possibile sarebbe con la consultazione con esperti del dominio. In Figura 7.12 la stessa caption viene data in input ad entrambi i modelli, i quali generano due report differenti. In questo caso sembrerebbe che GPT-NEO sia stato maggiormente in grado di dare un’indicazione all’oftalmologo.

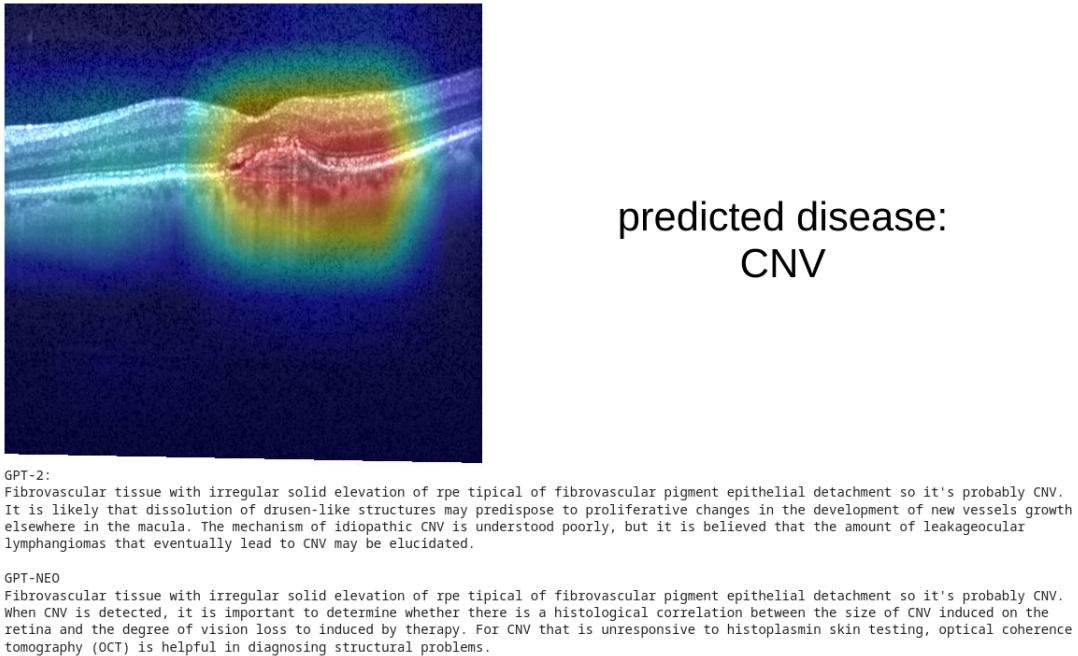


Figura 7.12: Confronto Report generato da GPT2 e GPT-Neo

Capitolo 8

Conclusioni e sviluppi futuri

Dai risultati sperimentali del Capitolo 7 si evince come il modello Multi-Tasking ha performance migliori rispetto a modelli che risolvono singoli task in modo autonomo; quest'ultimi, proprio per la mancata comunicabilità reciproca, soffrono maggiormente del problema della discordanza delle predizioni. D'altro canto, poiché l'allenamento del modello congiunto ha a disposizione un numero ridotto di istanze (dove ciascuna deve essere necessariamente correlata di Label, immagine e caption) non è possibile ottenere ulteriori margini di miglioramento, specialmente per le valutazioni sul testo generato. Se si dispone di un secondo Dataset molto più ampio per il task di classificazione, come in questo caso, applicare un Transfer Learning aumenta o migliora le performance dei sistemi di captioning. Come sviluppi futuri si prevede l'addestramento di modelli Multi-Tasking su hardware con più potenza computazionale (in modo da poter implementare reti più complesse) e su molte più immagini OCT corredate di descrizione, al fine di esplorare nuove soluzioni trasversali per tutti i task. Inoltre, si spera che il sistema di supporto diagnostico presentato sia raffinato verso la scrittura di un report sempre più clinicamente accettato tramite il sostegno di specialisti nella valutazione dei risultati generati.

Bibliografia

- [1] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. *Computers in biology and medicine*, 100:270–278, 2018.
- [2] Jameela Al-Jaroodi, Nader Mohamed, and Eman Abukhousa. Health 4.0: on the way to realizing the healthcare of the future. *Ieee Access*, 8:211189–211210, 2020.
- [3] Ganjar Alfian, Muhammad Syafrudin, Muhammad Fazal Ijaz, M Alex Syae-khoni, Norma Latif Fitriyani, and Jongtae Rhee. A personalized healthcare monitoring system for diabetic patients by utilizing ble-based sensors and real-time data processing. *Sensors*, 18(7):2183, 2018.
- [4] Douglas M Baughman, Grace L Su, Irena Tsui, Cecilia S Lee, and Aaron Y Lee. Validation of the total visual acuity extraction algorithm (tova) for automated extraction of visual acuity data from free text, unstructured clinical records. *Translational Vision Science & Technology*, 6(2):2–2, 2017.
- [5] Omar Bernabé, Elena Acevedo, Antonio Acevedo, Ricardo Carreño, and Sandra Gómez. Classification of eye diseases in fundus images. *IEEE Access*, 9:101267–101276, 2021.
- [6] Chandradeep Bhatt, Indrajeet Kumar, V Vijayakumar, Kamred Udhama Singh, and Abhishek Kumar. The state of the art of deep learning models in medical science and their challenges. *Multimedia Systems*, 27(4):599–613, 2021.
- [7] Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep learning and its applications in biomedicine. *Genomics, proteomics & bioinformatics*, 16(1):17–32, 2018.

- [8] Sahil Chelaramani, Manish Gupta, Vipul Agarwal, Prashant Gupta, and Ranya Habash. Multi-task learning for fine-grained eye disease prediction. In *Asian Conference on Pattern Recognition*, pages 734–749. Springer, 2019.
- [9] Jimmy S Chen and Sally L Baxter. Applications of natural language processing in ophthalmology: present and future. *Frontiers in Medicine*, 9, 2022.
- [10] Kyung Jun Choi, Jung Eun Choi, Hyeon Cheol Roh, Jun Soo Eun, Jong Min Kim, Yong Kyun Shin, Min Chae Kang, Joon Kyo Chung, Chaeyeon Lee, Dongyoung Lee, et al. Deep learning models for screening of high myopia using optical coherence tomography. *Scientific reports*, 11(1):1–11, 2021.
- [11] Mahdad Esmaeili, Alireza Mehri Dehnavi, and Hossein Rabbani. 3d curvelet-based segmentation and quantification of drusen in optical coherence tomography images. *Journal of Electrical and Computer Engineering*, 2017, 2017.
- [12] Gunther Eysenbach et al. What is e-health? *Journal of medical Internet research*, 3(2):e833, 2001.
- [13] Dinesh Visva Gunasekeran and Tien Yin Wong. Artificial intelligence in ophthalmology in 2020: a technology on the cusp for translation and implementation, 2020.
- [14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [15] Mahbub Hussain, Jordan J Bird, and Diego R Faria. A study on cnn transfer learning for image classification. In *UK Workshop on computational Intelligence*, pages 191–202. Springer, 2018.
- [16] Md Tariqul Islam, Sheikh Asif Imran, Asiful Arefeen, Mahmudul Hasan, and Celia Shahnaz. Source and camera independent ophthalmic disease recognition from fundus image using neural network. In *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pages 59–63. IEEE, 2019.

- [17] Hongyang Jiang, Kang Yang, Mengdi Gao, Dongdong Zhang, He Ma, and Wei Qian. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2045–2048. IEEE, 2019.
- [18] Kang Kermany, Daniel; Zhang. Large dataset of labeled optical coherence tomography (oct) and chest x-ray images. 2018.
- [19] Jongwoo Kim and Loc Tran. Retinal disease classification from oct images using deep learning algorithms. In *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–6. IEEE, 2021.
- [20] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–28, 2022.
- [21] Thomas Kurmann, Siqing Yu, Pablo Márquez-Neila, Andreas Ebneter, Martin Zinkernagel, Marion R Munk, Sebastian Wolf, and Raphael Sznitman. Expert-level automated biomarker identification in optical coherence tomography scans. *Scientific reports*, 9(1):1–9, 2019.
- [22] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature biomedical engineering*, 3(3):173–182, 2019.
- [23] Karis Little, Jacey H Ma, Nan Yang, Mei Chen, and Heping Xu. Myofibroblasts in macular fibrosis secondary to neovascular age-related macular degeneration—the potential sources and molecular cues for their recruitment and activation. *EBioMedicine*, 38:283–291, 2018.
- [24] Guangyi Liu, Yinghong Liao, Fuyu Wang, Bin Zhang, Lu Zhang, Xiaodan Liang, Xiang Wan, Shaolin Li, Zhen Li, Shuixing Zhang, et al. Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):3786–3797, 2021.

- [25] Lin Liu, Shenghui Zhao, Haibao Chen, and Aiguo Wang. A new machine learning method for identifying alzheimer’s disease. *Simulation Modelling Practice and Theory*, 99:102023, 2020.
- [26] Avleen Malhi, Timotheus Kampik, Husanbir Pannu, Manik Madhikermi, and Kary Främling. Explaining machine learning-based classifications of in-vivo gastral images. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2019.
- [27] Elliot Mbunge, Benhildah Muchemwa, John Batani, et al. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. *Global Health Journal*, 5(4):169–177, 2021.
- [28] Mutlu Mete, Leah Hennings, Horace J Spencer, and Umit Topaloglu. Automatic identification of angiogenesis in double stained images of liver tissue. In *BMC bioinformatics*, volume 10, pages 1–14. Springer, 2009.
- [29] Eka Miranda, Mediana Aryuni, and E Irwansyah. A survey of medical image classification techniques. In *2016 international conference on information management and technology (ICIMTech)*, pages 56–61. IEEE, 2016.
- [30] Meindert Niemeijer, Michael D Abramoff, and Bram van Ginneken. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Medical image analysis*, 10(6):888–898, 2006.
- [31] Ramon Pires, Herbert F Jelinek, Jacques Wainer, Eduardo Valle, and Anderson Rocha. Advancing bag-of-visual-words representations for lesion classification in retinal images. *PloS one*, 9(6):e96814, 2014.
- [32] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.
- [33] Anju Rani, Deepti Mittal, et al. Detection and classification of focal liver lesions using support vector machine classifiers. *Journal of Biomedical Engineering and Medical Imaging*, 3(1):21, 2016.
- [34] Sashank Santhanam and Samira Shaikh. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*, 2019.

- [35] Ferdinand Schlanitz, Bernhard Baumann, Stefan Sacu, Lukas Baumann, Michael Pircher, Christoph K Hitzenberger, and Ursula Margarethe Schmidt-Erfurth. Impact of drusen and drusenoid retinal pigment epithelium elevation size and structure on the integrity of the retinal pigment epithelium layer. *British Journal of Ophthalmology*, 103(2):227–232, 2019.
- [36] M Soh. Learning cnn-lstm architectures for image caption generation; dept. *Comput. Sci., Stanford Univ., Stanford, CA, USA*, 2016.
- [37] Yang Song, Weidong Cai, Heng Huang, Yun Zhou, Yue Wang, and David Dagan Feng. Locality-constrained subcluster representation ensemble for lung image classification. *Medical image analysis*, 22(1):102–113, 2015.
- [38] Li Sun, Weipeng Wang, Jiyun Li, and Jingsheng Lin. Study on medical image report generation based on improved encoding-decoding method. In *International Conference on Intelligent Computing*, pages 686–696. Springer, 2019.
- [39] Bas HM van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.
- [40] Sivamurugan Vellakani and Indumathi Pushbam. An enhanced oct image captioning system to assist ophthalmologists in detecting and classifying eye diseases. *Journal of X-Ray Science and Technology*, 28(5):975–988, 2020.
- [41] Plácido L Vidal, Joaquim de Moura, Macarena Díaz, Jorge Novo, and Marcos Ortega. Diabetic macular edema characterization and visualization using optical coherence tomography images. *Applied Sciences*, 10(21):7718, 2020.
- [42] Hesheng Wang and Baowei Fei. A modified fuzzy c-means classification method using a multiscale diffusion filtering scheme. *Medical image analysis*, 13(2):193–202, 2009.
- [43] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*, 2015.
- [44] Paul Windisch, Pascal Weber, Christoph Fürweger, Felix Ehret, Markus Kufeld, Daniel Zwahlen, and Alexander Muacevic. Implementation of model

- explainability for a basic brain tumor detection using convolutional neural networks on mri slices. *Neuroradiology*, 62(11):1515–1518, 2020.
- [45] Yeo Chan Yoon, So Young Park, Soo Myoung Park, and Heuiseok Lim. Image classification and captioning model considering a cam-based disagreement loss. *ETRI Journal*, 42(1):67–77, 2020.
- [46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.