POLITECNICO DI TORINO
Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

Progettazione e sviluppo di una soluzione per interpretare i modelli black-box.
Caso di studio: classificazione di immagini e documenti testuali

Relatore

Prof.ssa Tania Cerquitelli

Candidato

Francesco Giacalone

ANNO ACCADEMICO 2017-2018

Ringraziamenti

Un sentito ringraziamento va alla professoressa Tania Cerquitelli per la sua competenza e disponibilità e a Francesco Ventura per il prezioso supporto.

Grazie a tutti i compagni di avventura con cui ho condiviso questo faticoso ma appassionante percorso accademico e di crescita. Non vi nominerò uno ad uno per la mia solita pigrizia, ma sappiate che dico proprio a voi.

Grazie infinite alla mia famiglia, lontana ma sempre vicinissima, per il sostegno, la pazienza, e l'affetto incondizionato.

Indice

Capitolo 1

Introduzione

Le tecniche di machine learning e deep learning forniscono degli strumenti estremamente utili per una vasta gamma di applicazioni di uso comune. Esse, tuttavia, potrebbero produrre dei comportamenti discriminatori e potenzialmente dannosi per gli utenti. Per ottenere quindi dei servizi basati su tali modelli black-box che si possano dire pienamente affidabili risulta fondamentale attestare l'importanza del principio di trasparenza algoritmica.

In questo lavoro di tesi si presentano due distinte soluzioni di trasparenza algoritmica, rispettivamente per classificatori di immagini e di documenti testuali, basati su reti neurali convoluzionali. Le soluzioni proposte permettono di produrre delle spiegazioni locali per le predizioni di un modello black-box mediante un processo di perturbazione di feature interpretabili. Il framework sviluppato sfrutta lo strumento delle ipercolonne, nel caso di immagini, oppure tecniche di information retrieval, nel caso di documenti testuali, per identificare un set interpretabile di feature dai dati in input. Mediante una serie di perturbazioni applicate a queste feature risulta possibile misurare l'influenza di ogni singola feature sulla predizione operata dalla rete neurale. Le soluzioni proposte sono state testate su un insieme eterogeneo di dati con buoni risultati che evidenziano l'efficacia di un approccio basato sull'estrazione e perturbazione di feature interpretabili.

Capitolo 2

Machine learning

Nel corso degli ultimi anni le discipline di intelligenza artificiale legate alla branca del *machine learning* stanno assumendo un ruolo di sempre maggiore rilievo in ambito tecnologico e la crescente disponibilità di dati forniti dalle moderne applicazioni ICT prospetta un futuro fortemente caratterizzato dalle tecnologie di apprendimento automatico data-driven. Lo scopo di questo capitolo è fornire una panoramica generale su queste tecnologie e sulle loro applicazioni con particolare attenzione a quelle utilizzate nel presente lavoro di tesi.

2.1 – Introduzione al machine learning

La definizione più citata di *machine learning* o apprendimento automatico è quella fornita da Tom M. Mitchell: "Si dice che un programma apprende dall'esperienza E con riferimento ad alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E".

L'idea alla base del concetto di apprendimento automatico è quindi quella di fornire ai computer l'abilità di imparare e replicare alcune operazioni senza essere stati esplicitamente programmati; tali operazioni sono in genere di tipo predittivo o decisionale e basate sui dati che si hanno a disposizione. [1][2]

2.1.1 I paradigmi dell'apprendimento automatico

I compiti dell'apprendimento automatico vengono solitamente classificati in tre categorie, o paradigmi:

- *Apprendimento supervisionato:* si forniscono esempi degli input e dei rispettivi output desiderati con l'obiettivo di estrarre una regola generale che associ l'input all'output corrispondente.

- *Apprendimento non supervisionato:* gli input forniti non hanno né una struttura definita né output associati. Lo scopo del calcolatore è quindi quello di identificare dei pattern negli input al fine di riprodurli o prevederli.
- *Apprendimento per rinforzo:* prevede l'interazione del calcolatore con un ambiente dinamico nel quale si cerca di raggiungere un obiettivo (per esempio il superamento di un livello in un videogioco).

2.1.2 Classificatori

Una delle applicazioni più frequenti per le tecniche di machine learning è quella dei *classificatori*. Il concetto di classificazione prevede l'individuazione delle caratteristiche di un'entità da classificare e la conseguente associazione di una classe di appartenenza e tale risultato può essere raggiunto tramite tecniche statistiche o di apprendimento supervisionato.

2.2 – Deep learning

Il *deep learning* o apprendimento profondo è un insieme di metodi riconducibili alla famiglia del machine learning che sono in grado di fornire dei modelli ad alto livello di astrazione per una vasta gamma di fenomeni non lineari. Tali tecniche hanno portato al raggiungimento di importanti progressi in varie discipline quali computer vision, natural language processing, riconoscimento facciale e vocale, e analisi di segnali in genere.

Il deep learning si basa su diversi modelli per rappresentare degli oggetti. Un'immagine, per esempio, può essere processata come un semplice vettore di campioni numerici oppure con altri tipi di rappresentazioni. Nello specifico la si potrebbe descrivere a partire da:

- l'intensità dei pixel
- i bordi degli elementi che la compongono
- le sue diverse regioni, con forme particolari

L'uso della giusta rappresentazione rende il compito di apprendimento più efficiente. La ricerca in quest'area quindi si sforza di costruire modelli della realtà quanto più efficienti con l'obiettivo di estrapolare le migliori rappresentazioni da vaste collezioni di dati non strutturati. Numerose tecniche di deep learning sono espressamente influenzate dalla neuroscienza e si ispirano ai modelli di elaborazione dell'informazione e di comunicazione del sistema nervoso, con particolare attenzione al modo in cui si stabiliscono le connessioni tra neuroni in base ai messaggi ricevuti, alle risposte neuronali e alle caratteristiche delle connessioni stesse.

Un'altra peculiarità delle tecniche di deep learning consiste nella sostituzione di alcuni artefatti particolarmente complessi con modelli algoritmici di apprendimento supervisionato o non supervisionato attraverso tecniche di estrazione gerarchica delle caratteristiche. Le tecniche di apprendimento profondo utilizzano infatti molteplici strati (layer) di unità di elaborazione non lineari per l'estrazione e la trasformazione di feature. Ogni layer prende in input l'output del precedente. Questa natura spiccatamente stratificata permette di operare con l'apprendimento su diversi livelli di dettaglio e rappresentazione dei dati. È quindi possibile passare dall'utilizzo di parametri di basso livello a parametri di alto livello, dove i diversi livelli corrispondono a diversi livelli di astrazione dei dati. In questo modo diviene possibile avvicinarsi maggiormente al significato semantico dei dati e dare loro la forma di immagini, suoni o testi. [3][4]

2.2.1 Il processo di apprendimento

Gli algoritmi di apprendimento possono essere supervisionati o non supervisionati e le loro applicazioni includono il riconoscimento di pattern e la classificazione statistica. La fase di apprendimento avviene nei seguenti passaggi:

1. **Pre-processamento dei dati**: i dati designati per l'apprendimento vengono trasformati e mappati in adeguate matrici numeriche così da poter essere forniti in input al modello. Nel caso di classificatori il dataset viene inoltre suddiviso negli insiemi di *training* e *test*. Per ogni istanza di questi dati è già presente la rispettiva etichetta di classificazione, indispensabile per il processo di training.
2. **Inizializzazione del modello**: per inizializzare il modello si assegnano dei valori ai parametri del modello. Questi valori possono essere assegnati randomicamente oppure ereditati da altri modelli di deep learning.
3. ***Training***: si ripartisce ulteriormente il set di training in batch. Ogni dato di un batch viene fornito in ingresso al modello, quest'ultimo calcola una *score function* che assegna al dato dei punteggi (ad esempio valori di appartenenza alle categorie di un classificatore). In seguito a questa predizione si calcola una *funzione di costo* che valuta la differenza tra i valori predetti e quelli reali. Lo scopo della rete neurale è quello di minimizzare il valore della funzione di costo mediante l'aggiornamento progressivo dei propri parametri (*back-propagation*).
4. ***Test***: in quest'ultima fase la rete processa i dati di test ed esegue una stima della qualità del modello.

2.2.2 Reti neurali artificiali

Una *rete neurale artificiale* (*Artificial Neural Network*, ANN) è un sistema di deep learning ispirato alle reti neurali biologiche. Si tratta di un modello matematico di calcolo basato su *percettroni*, neuroni artificiali in grado di apprendere, ovvero accumulare esperienza. L'elemento costituente più rilevante delle reti neurali è costituito dalle interconnessioni di informazioni che implementano un approccio connessionistico dinamico, quindi in grado di modificare la propria struttura in base ai flussi di dati che determinano l'apprendimento. Una rete neurale è organizzata in strati di neuroni, riceve le informazioni d'ingresso attraverso un *input layer* e restituisce i risultati dell'elaborazione per mezzo di un *output layer*, tra questi strati possono esserci uno o più *hidden layer*. [3][4]

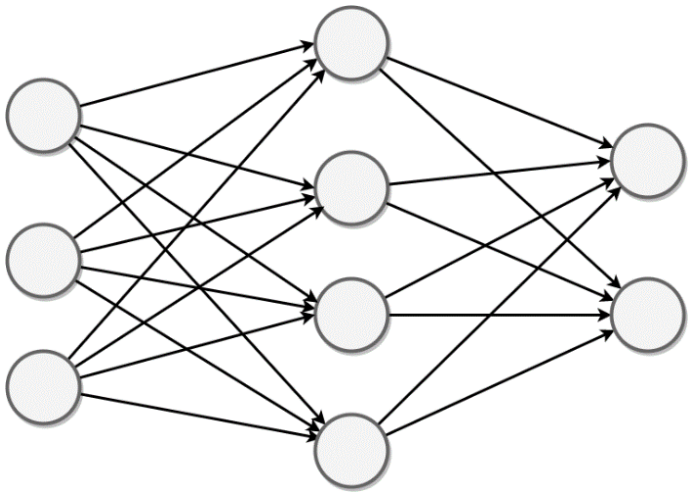


Figura 2.1: *Schema di una rete neurale artificiale. Da sinistra verso destra si hanno layer di input, un hidden layer e layer di output.*

Tra le categorie più diffuse di reti neurali si hanno:

- Reti neurali *convoluzionali*
- Reti neurali *ricorsive*
- Reti neurali *ricorrenti*
- Reti *deep belief*

2.3 – Reti neurali convoluzionali

In machine learning, una *rete neurale convoluzionale* (*CNN* o *ConvNet*) è una rete aciclica (*feed-forward*) di neuroni artificiali il cui il modello di connessione tra i

neuroni è ispirato alla corteccia visiva animale. I neuroni di questa regione del cervello sono disposti in modo tale da corrispondere a regioni sovrapposte del campo visivo. Le ConvNet consistono di una pila multistrato di percettroni, il cui scopo è quello di processare piccole quantità di informazione. Le reti neurali convoluzionali si stanno affermando come stato dell'arte per una vasta gamma di applicazioni nei campi della computer vision e del natural language processing. [3][5]

2.3.1 Architettura e funzionamento generale

Le reti neurali convoluzionali consistono di due tipi di neuroni che processano l'informazione:

- I neuroni di elaborazione, che elaborano una porzione limitata dell'immagine (chiamata *campo ricettivo*) attraverso una funzione di convoluzione
- I neuroni di aggregamento o di *pooling*

L'insieme di output di un layer consente di ricostruire un'*immagine* intermedia, che servirà da base per il livello successivo.

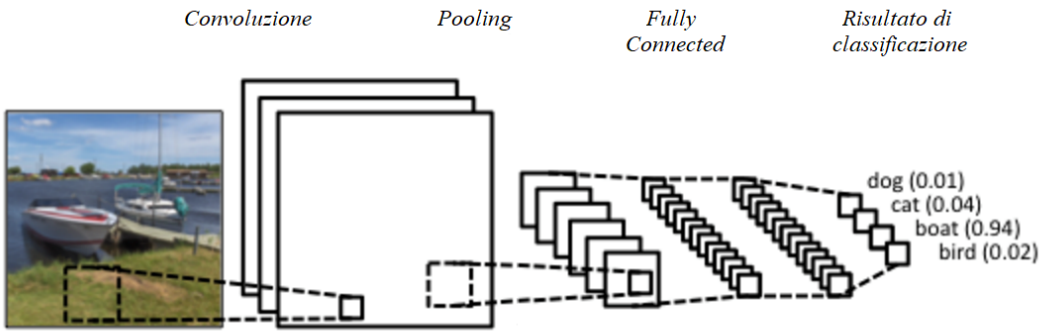


Figura 2.2: *Schema ridotto di una rete neurale convoluzionale con in evidenza i moduli di convoluzione e pooling e i layer fully-connected finali. Fonte www.wildml.com*

2.3.2 Convoluzione

L'operatore *convoluzione* rappresenta la componente principale delle ConvNet e il suo compito consiste nell'estrazione di caratteristiche dai dati in input.

Si consideri ad esempio l'elaborazione di un'immagine. Nei moduli di convoluzione i dati vengono processati con il meccanismo della *sliding window*, dove una piccola matrice chiamata *kernel di convoluzione* - i cui valori caratterizzano l'operatore - viene opportunamente traslata seguendo la struttura dell'immagine. La convoluzione è un operatore locale in quanto prende in input una piccola matrice di valori e fornisce

in output un solo valore in corrispondenza del centro della matrice e viene applicata in modo da mantenere la relazione spaziale tra i pixel. I kernel di convoluzione contengono i pesi utilizzati per calcolare la *score function* e sono associati a un *bias* che viene sommato al risultato della convoluzione.

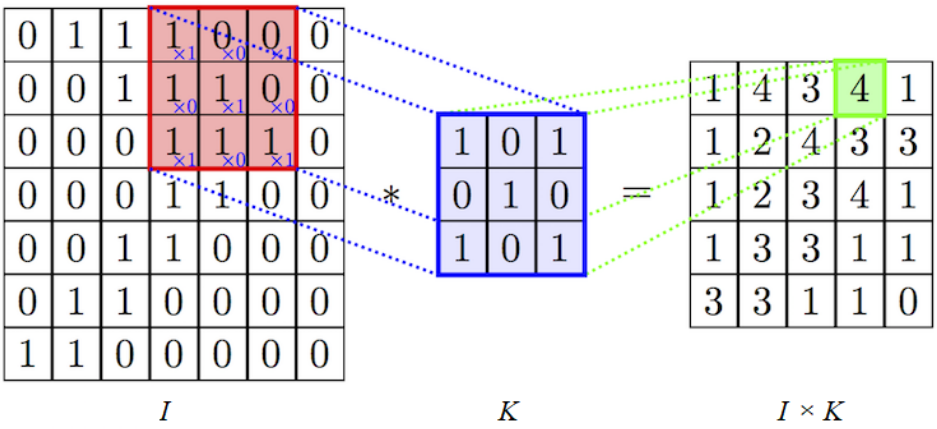


Figura 2.3: *Rappresentazione matriciale dell'operazione di convoluzione dove I indica la sliding window e K il kernel di convoluzione.*

2.3.3 Attivazione

La *funzione di attivazione* o di trasferimento viene solitamente applicata in seguito all'operazione di convoluzione e determina il comportamento in output di un neurone in funzione del suo livello di eccitazione. Le funzioni di attivazione sono solitamente non-lineari ma continue e differenziabili comprese nell'intervallo [-1, 1] come ad esempio la *standard logistic function* (*sigmoide*) o la *tangente iperbolica*.

2.3.4 Pooling

Il modulo di *pooling* (o aggregazione) si occupa di aggregare l'input e ridurne il volume per mezzo di un sotto-campionamento così da snellire l'elaborazione per i layer successivi. Il pooling, così come avviene per la convoluzione, agisce con l'ausilio di un kernel traslato lungo tutta l'immagine. Una tecnica diffusa di sotto-campionamento è il MaxPooling che ad ogni spostamento del kernel restituisce soltanto il valore massimo presente.

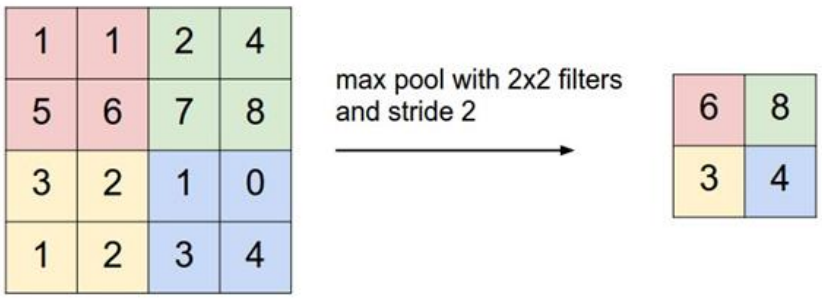


Figura 2.4: *Rappresentazione matriciale dell'operatore di MaxPooling. Per ciascuna porzione 2×2 della matrice si prende solo il valore massimo.*

2.3.5 Layer fully-connected

A valle della rete si trovano uno o più strati di neuroni completamente connessi che applicano al volume le trasformazioni finali. Mediante un set di pesi e un altro di bias si esegue la *funzione di classificazione* e si restituisce il corrispondente vettore di punteggi.

Capitolo 3

Machine learning e trasparenza algoritmica

I sistemi di machine learning stanno ottenendo un ruolo di sempre maggior rilevo nei moderni apparati decisionali. Tale ruolo è destinato a crescere ulteriormente in numerosi settori e per gli scopi più disparati, dall'industria alla sanità, dai web service alla finanza. Il fattore determinante di questa profonda trasformazione è l'inedita disponibilità di ampi volumi di dati provenienti dalle moderne applicazioni ICT.

Come esposto nel capitolo precedente, la disciplina del machine learning si propone di ricavare modelli analitici complessi con l'ausilio di tecniche di data mining allo scopo di assumere decisioni autonomamente o effettuare predizioni. Tali modelli, e di conseguenza i processi decisionali che si basano su essi, sono tuttavia dei modelli *black-box*, cioè offrono una visione estremamente limitata del loro comportamento interno. L'opacità dei processi di apprendimento automatico pone seri quesiti sull'affidabilità dei modelli e delle predizioni; non è possibile infatti motivare in alcun modo le decisioni adottate dagli algoritmi.

Le emergenti soluzioni di trasparenza applicate al machine learning mirano al raggiungimento di un maggior livello di interpretabilità dei modelli al fine di individuare comportamenti errati o discriminatori. Settori strettamente legati al benessere della collettività quali sanità, sicurezza, impiego, finanza necessitano infatti di algoritmi decisionali equi e interpretabili (*social good decision-making algorithms*). Questo capitolo offrirà una breve panoramica sugli aspetti sociali legati all'impiego di sistemi decisionali *data-driven*.

3.1 L'esigenza sociale di trasparenza algoritmica

I moderni approcci decisionali basati su big data e machine learning aprono a innumerevoli scenari di ricerca e prospettano profondi cambiamenti negli ambiti più disparati della società. In questo contesto risulta evidente l'importanza di comprendere l'impatto sociale di queste nuove metodologie e minimizzarne le eventuali criticità. Molte di queste criticità sono state identificate negli ultimi anni da

un numero sempre maggiore di osservatori ed esperti e costituiscono ciò che è stato denominato il *lato oscuro* dei sistemi decisionali *data-driven*. [6]

In questa sezione si analizzeranno quindi i potenziali aspetti negativi legati all'impiego massivo di sistemi decisionali che sfruttano tecniche di apprendimento automatico. Si evidenzieranno inoltre alcuni requisiti fondamentali per garantire che questo imminente cambiamento di paradigma risulti socialmente accettabile, con una particolare attenzione al concetto di *trasparenza algoritmica*.

3.1.1 Criticità dei sistemi di machine learning

L'opacità delle tecniche di apprendimento automatico limita significativamente il grado di controllo che si tiene sui dati. Ciò pone seri interrogativi in termini di affidabilità in quanto non si ha alcuna protezione contro fenomeni fortemente nocivi quali asimmetria dei dati, discriminazioni, esclusione sociale o violazioni della privacy. [6]

- **Asimmetria dei dati e mancanza di trasparenza:** con *asimmetria dei dati* si intende una condizione in cui l'accesso ai dati e le conoscenze per adoperarli sono elargiti solo a certi gruppi di persone (governi, aziende oppure organizzazioni in genere). Questi gruppi potrebbero facilmente usare gli strumenti a propria disposizione per operare a proprio vantaggio e a discapito degli altri. Il fattore determinante di questa asimmetria è la grave mancanza di conoscenza e consapevolezza negli individui che non hanno accesso ai dati. Per di più, questo fenomeno è spesso aggravato dalle scarse competenze dei soggetti danneggiati in ambito informatico e di data science;
- **Esclusione sociale e discriminazione:** le collezioni di dati su cui si basano i meccanismi decisionali potrebbero, anche involontariamente, nascondere comportamenti potenzialmente discriminatori. Questi comportamenti sono risultato del perpetrarsi di decisioni discriminanti prese nel passato o semplicemente riproducono tendenze già presenti e largamente diffuse nella società. Una mancata identificazione di queste tendenze finirebbe quindi per mantenere o addirittura esacerbare le iniquità;
- **Violazioni della privacy:** l'esplosione di discipline legate ai big data potrebbe permettere di risalire ai *dati personali* di specifici individui. Infatti non è raro che i moderni dataset contengano informazioni più o meno facilmente riconducibili a dati identificativi come nomi, date di nascita o indirizzi IP con serissime ripercussioni sulla privacy, e di conseguenza sulla libertà individuale, di singoli soggetti.

3.1.2 Requisiti chiave per sistemi equi di machine learning

Le tecniche decisionali basate su data mining e machine learning vedranno nei prossimi anni una crescita esponenziale. Risulta quindi di vitale importanza combattere le problematiche esposte nel paragrafo precedente per garantire sistemi decisionali equi. Si presentano di seguito alcuni punti chiave in quest'ottica. [6]

- **Trasparenza algoritmica:** i recenti modelli basati su machine learning agiscono come black-box e il loro impiego nei sistemi decisionali si traduce in una sostanziale mancanza di motivazioni a supporto delle decisioni adottate e di conseguenza in una pesante messa in discussione del livello di fiducia riposta nei risultati ottenuti. Per questo motivo gli sforzi della comunità scientifica legata al mondo del machine learning si stanno progressivamente orientando sulla tematica della trasparenza algoritmica. Una maggiore comprensione delle dinamiche che si attuano nei sistemi di apprendimento automatico permetterebbe di attenuare o eliminare del tutto alcune delle criticità individuate nel paragrafo precedente. Ad esempio, nel contesto di sistemi decisionali che presentano tendenze discriminatorie, molti studi si focalizzano sull'individuazione di tali tendenze e sulla comprensione delle le cause.
- **Gestione dei dati centrata sull'utente:** allo scopo di fornire ai singoli individui un maggior grado di controllo sui propri dati si sono proposti dei modelli centrati sull'utente (*user-centric*) per la gestione personale dei dati. Tali soluzioni cercano di garantire all'utente l'accesso ai propri dati personali altrimenti precluso. Approcci di questo tipo ambiscono quindi a sfruttare quanto più possibile il valore dei dati comportamentali umani e al contempo garantire privacy agli utenti coinvolti.

Capitolo 4

Stato dell'arte

Come esposto nella sezione precedente, il tema della trasparenza algoritmica si sta sempre più affermando come importante oggetto di studio nella comunità scientifica. In questa sezione si introdurranno alcuni dei risultati più rilevanti in questo ambito che hanno costituito il punto di partenza per lo sviluppo di questo lavoro di tesi. Come conseguenza del largo impiego di classificatori basati su machine learning, gran parte della ricerca si focalizza principalmente sullo studio della trasparenza algoritmica applicata ai classificatori.

4.1 Quantitative Input Influence

Nell'ambito della trasparenza algoritmica applicata ai classificatori si segnala lo studio di Datta, Sen e Zick sulla famiglia di misure *QII* (*Quantitative Input Influence*) [7]. Le misure QII permettono di generare dei report di trasparenza tramite la valutazione dell'influenza degli input sugli output in un sistema algoritmico che opera come *black-box*. Nel caso specifico di un classificatore si vuole misurare l'impatto di singole feature o gruppi di esse, appartenenti a un dataset strutturato, sul risultato di una classificazione applicata ad uno o più oggetti. Per misurare l'influenza sulla decisione finale di una feature si provvede a perturbarne opportunamente il valore così da rendere possibile l'analisi delle variazioni del comportamento del classificatore. Come esempio si voglia considerare un classificatore aziendale che opera sulle assunzioni di un insieme di candidati (dataset *adult* avente le feature: *Age*, *Gender*, *Weight*, *Marital Status* ed *Education*); si può decidere di operare una perturbazione sulla feature *Gender* (ad esempio impostando il medesimo valore della feature per l'intero insieme di candidati) per riuscire a stabilire se il classificatore presenta un'inclinazione a discriminare in base al sesso del candidato. Allo scopo di cogliere diversi aspetti della correlazione tra input e output sono stati definiti i seguenti tipi di misure:

- ***Quantità di interesse:*** rappresenta una proprietà statistica di un sistema algoritmico e fornisce una prima base per la stesura di report di trasparenza.

Un esempio di quantità di interesse è rappresentato dalla probabilità per uno o più elementi di essere assegnato ad una determinata classe in base alla configurazione degli input.

- ***Unary QII**:* analizza il rapporto di causalità tra i singoli input e il risultato di una predizione. La misura di Unary QII rappresenta la variazione di una quantità di interesse tra una distribuzione di input reale ed una distribuzione ipotetica opportunamente costruita in maniera specifica al fine di evidenziare eventuali input correlati. Rompendo le correlazioni tra i diversi input diviene possibile valutarne singolarmente l'impatto sulla decisione finale.
- ***Set QII*** e ***Marginal QII**:* nel caso di sistemi dove l'impatto delle singole feature è troppo contenuto per fornire informazioni rilevanti si vuole misurare l'influenza congiunta (*joint influence*) di un set di input sulla decisione finale. La misura di Set QII rappresenta quindi la generalizzazione dell'Unary QII al caso di un insieme di distribuzioni di input mentre la Marginal QII intende valutare il peso dei singoli input all'interno di tale insieme.

Questo approccio è applicabile solo su dataset strutturati, ovvero dataset dove le feature sono univocamente definite per ogni entità che vi appartiene. Di conseguenza lo studio sulle misure QII non considera tipologie di dati non strutturati quali immagini, suoni o documenti testuali.

4.2 Local Interpretable Model-agnostic Explanations

Il framework *LIME* (*Local Interpretable Model-agnostic Explanations*) sviluppato da Ribeiro, Singh e Guestrin [8] punta a determinare un modello interpretabile del classificatore al fine di stabilirne il grado di affidabilità. Tale modello deve fornire spiegazioni comprensibili per gli esseri umani e quindi composte da un insieme limitato di caratteristiche facilmente identificabili e semanticamente significative. Per un sistema di spiegazione delle predizioni di questo genere sono state inoltre definite le proprietà fondamentali di fedeltà locale e indipendenza dal modello:

- ***Indipendenza dal modello***: un sistema di spiegazione delle predizioni deve essere *model agnostic* ovvero indipendente dal modello. Deve quindi essere in grado di funzionare a prescindere dal tipo di algoritmo alla base del processo decisionale o predittivo.
- ***Fedeltà locale***: dal momento che le risposte fornite da un sistema di spiegazione delle predizioni riguardo l'esito di una certa istanza non potranno mai rispecchiare universalmente il comportamento di un classificatore, si mira

all'ottenimento di spiegazioni localmente valide che sappiano approssimare il comportamento del classificatore nelle vicinanze dell'istanza predetta.

Per una comprensione quanto più ampia di un sistema predittivo si sceglie quindi un opportuno insieme di istanze individuali da sottomettere al sistema e si cerca di fornire una spiegazione ai risultati ottenuti. Attraverso l'analisi del comportamento nelle località delle istanze sottoposte al sistema predittivo, è possibile approssimare le caratteristiche del modello a livello globale.

Capitolo 5

Architettura generale e stima delle influenze

In questo capitolo si introdurrà la metodologia sviluppata per le soluzioni di trasparenza algoritmica in classificatori basati su machine learning. Dopo una breve panoramica sull'architettura generale del framework si presenteranno le diverse metriche sviluppate per la stima dell'influenza delle feature sul comportamento del classificatore.

Per riuscire a comprendere il comportamento di un modello black-box è necessario valutare le sue risposte alle variazioni dei dati di input [7]. Nel caso di dataset strutturati, ovvero tabelle con campi ben definiti, si provvede ad operare una perturbazione su uno o più campi (feature) dell'input e si misurano le eventuali variazioni nell'output. Ciò risulta difficilmente applicabile nel caso di dati non strutturati come immagini o testi dove le feature sono delle entità proporzionalmente molto piccole, come pixel o singole parole, e quindi con influenza estremamente limitata, se non del tutto nulla, sugli output. Per poter applicare la perturbazione degli input anche su dati non strutturati diventa quindi necessario aggregare i dati in un numero limitato di macro-feature. Requisito fondamentale per le feature estratte è che siano valide semanticamente, o in altre parole, che risultino essere facilmente interpretabili per un osservatore umano. Una volta estratto un set limitato e interpretabile di feature si può eseguire il task di classificazione operando delle modifiche sugli input e misurando l'impatto di queste perturbazioni sui risultati del modello. Raccogliendo le misure ottenute è possibile generare dei report di trasparenza sulle classificazioni del modello e quindi raccogliere utili informazioni sul comportamento locale del modello.

Nella soluzione proposta si possono quindi individuare tre passaggi principali:

1. **Estrazione di feature interpretabili**: riduzione di un oggetto (immagine o testo) ad un set limitato di caratteristiche il cui significato semantico risulti comprensibile per un osservatore umano. Una feature interpretabile può essere una porzione d'immagine che individua un elemento ben distinto oppure un gruppo di parole vicine per significato.

2. **Perturbazione degli input**: per ognuna delle feature individuate nel passaggio precedente si opera un processo di mascheramento o rimozione al fine di minimizzare l'impatto della feature stessa sul risultato del modello. Si ottiene quindi un insieme di input perturbati, oggetti del tutto simili all'originale ma ciascuno privato di una sua feature.
3. **Predizioni e generazione del report di trasparenza**: si sottomettono al modello black-box l'istanza originale dell'oggetto e i relativi input perturbati. Eseguendo il task di classificazione si possono confrontare le variazioni dell'output del modello e valutare l'influenza di ogni singola feature sulla predizione del classificatore. I risultati ottenuti contribuiscono alla creazione di un report di trasparenza.

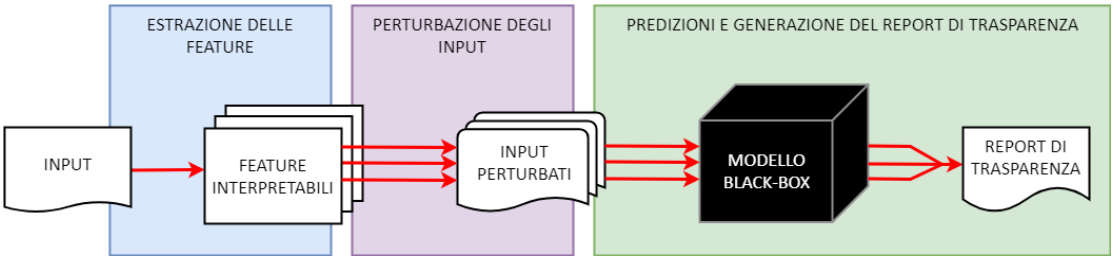


Figura 5.1: *Schema riassuntivo dell'architettura sviluppata. Nella figura si evidenziano i tre passaggi principali del processo.*

5.1 – Stima dell'influenza delle feature

Una volta eseguiti i passaggi estrazione e perturbazione delle feature diventa possibile procedere alla valutazione dell'impatto delle feature stesse sul risultato di classificazione fornito dal modello.

Si alimenta quindi la rete neurale con l'istanza originale dell'oggetto da classificare e si memorizza il risultato R. Per un classificatore a n classi, R sarà un vettore di lunghezza n contenente dei punteggi $r_1, r_2, r_3, \dots, r_{n-1}, r_n$ che indicano la probabilità dell'immagine di appartenere a ciascuna delle classi.

$$R = [r_1, r_2, r_3, \dots, r_{n-1}, r_n] \tag{5.1}$$

Si prosegue iterando la stessa operazione per ciascuna perturbazione i corrispondente ad una feature. Il risultato sarà un altro vettore R^i, naturalmente composto da valori più o meno differenti per via della perturbazione.

$$R^i = [r^i_1, r^i_2, r^i_3, \dots, r^i_{n-1}, r^i_n] \tag{5.2}$$

Dal confronto di R^i con R e attraverso una serie di metriche è possibile ricavare indicazioni di vario genere riguardo l'impatto della feature i sul risultato R. Le metriche proposte nel presente lavoro di tesi si dividono in due famiglie:

- Metriche relative ad una classe di interesse: misurano l'influenza locale delle feature in input con riferimento ad una particolare classe.
- Metriche inter-classe: misurano l'influenza delle feature sulla totalità delle classi.

5.1.1 Metriche relative ad una classe di interesse

Il primo obiettivo delle soluzioni di trasparenza sviluppate consiste nell'analisi dell'influenza delle feature su una determinata *classe di interesse*. Questa classe potrebbe essere scelta algoritmicamente oppure arbitrariamente da un utilizzatore umano. Determinata quindi una classe di interesse, si analizza in che quantità le feature impattano sul risultato.

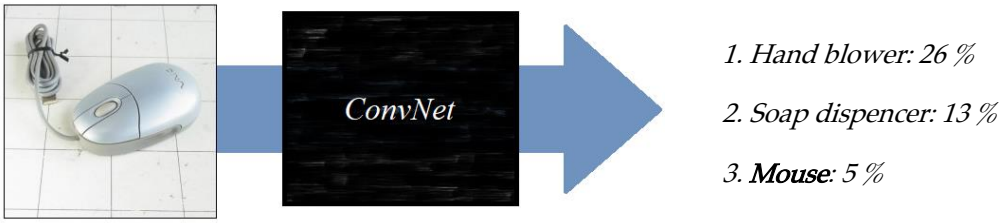


Figura 5.1: *Nell'esempio proposto una rete neurale convoluzionale non identifica correttamente la classe dell'immagine in input. Potrebbe quindi essere utile formulare delle query di trasparenza relative alla classe di interesse '**mouse**':*

- *Quali feature influiscono positivamente sulla classe '**mouse**'?*
- *Quali feature influiscono negativamente sulla classe '**mouse**'?*

Si considerino i valori r_c, pari al risultato di classificazione dell'immagine originale relativo alla classe c, e r^i_c derivato dalla perturbazione della feature i. Si definiscono le seguenti metriche per la stima locale dell'influenza di una feature rispetto a una classe di influenza:

- ***DI:*** l'indice *DI* (*Delta Impact*) viene calcolato come la semplice differenza tra i due valori e rappresenta un primo efficace indicatore dell'influenza positiva ($DI > 0$) o negativa ($DI < 0$) di una feature. Questa misura fornisce

un'informazione prevalentemente legata all'ampiezza di una variazione di predizione.

$$DI = r_c - r_c^i \tag{5.3}$$

- ***IR:*** l'indice *IR* (*Influence Relation*), calcolato come il rapporto tra i due valori di predizione, fornisce altre informazioni utili sull'influenza di una feature. Se il valore che si ottiene è compreso nell'intervallo [0, 1) allora indica un'influenza negativa da parte della feature sulla classe di interesse; se, al contrario, è compreso in (1, +∞) allora indica un'influenza positiva. Più il valore è vicino a 1, e minore è l'impatto della feature sul risultato di classificazione.

$$IR = \frac{r_c}{r_c^i} \tag{5.4}$$

- ***nIRI:*** l'indice *IRI* (*normalized Influence Relation Index*) unisce e normalizza le informazioni provenienti dal calcolo dell'indice *DI* rispetto a quelle dell'indice *IR*. Compreso nell'intervallo [-1, 1], indica un'influenza tanto più forte quanto più il valore si allontana dallo 0. La formula per il calcolo dell'indicatore IR è la seguente:

$$nIRI = softsign\left(d * \left(IR + \frac{1}{IR}\right)\right) \tag{5.5}$$

  Dove la presenza del delta d introduce una stima dell'ampiezza della variazione e fornisce inoltre il segno, meno (-) per le influenze negative, più (+) per quelle positive. Il rapporto *IR* assume valori generalmente grandi per feature con influenza positiva, ma trascurabili altrimenti e per questo motivo viene sommato al suo reciproco *1/IR* che al contrario misura bene le influenze negative. La funzione softsign normalizza il tutto nell'intervallo [-1, 1].

$$softsign(x) = \frac{x}{1 + |x|} \tag{5.6}$$

5.1.2 Metriche inter-classe

Le misure legate ad una classe di interesse offrono già delle precise indicazioni sul comportamento del classificatore. Risulta tuttavia interessante valutare l'influenza di una feature sul modello relativamente alla totalità delle classi o parte di esse.

- ***AIR:*** l'indice *AIR* (*Average Influence Relation*) fornisce il dato di influenza media di una feature sulla totalità delle classi. Viene calcolato come media degli *IR* su tutte le classi.

$$AIR = \frac{\sum_n IR_c}{n} \tag{5.7}$$

- ***WIR:*** come nel caso precedente l'indice *WIR* (*Weighted average Influence Relation*) è una media, ma pesata con i valori r_c del risultato di classificazione. In questo modo si ottiene una stima generale dell'influenza della feature per un set ristretto di classi individuate dal modello.

$$WIR = \frac{\sum_n r_c * IR_c}{\sum_n r_c} \tag{5.8}$$

- ***IRP:*** l'indice *IRP* (*Influence Relation Precision*) indica la tendenza di una feature ad influenzare positivamente e univocamente una determinata classe di interesse. Per valori di IRP minori di uno la feature in esame non ha un impatto solo sulla classe di interesse, ma anche su altre classi. Al contrario, se IRP presenta un valore molto maggiore di 1 si può determinare che la feature non solo ha una certa influenza sulla classe di interesse, ma contiene anche degli elementi univocamente caratterizzanti per la classe di interesse. Si possono invece considerare non caratterizzanti eventuali input perturbati con valori di IRP prossimi a 1. L'indice IRP si calcola come rapporto tra l'indice IR di una certa classe e l'indice medio WIR.

$$IRP = \frac{IR_c}{WIR} \tag{5.9}$$

Capitolo 6

Estrazione e perturbazione di feature – Immagini

In questo capitolo si analizzerà dettagliatamente l'approccio sviluppato per l'estrazione di feature da immagini e la loro perturbazione nel caso di una rete neurale convoluzionale che esegue task di classificazione.

La metodologia prevede che il compito di estrazione delle feature sia svolto con l'ausilio stesso modello di machine learning del quale si vuole motivare il comportamento.

6.1 – Ipercolonne

Nelle neuroscienze un'*ipercolonna* o colonna corticale è un insieme di neuroni della corteccia cerebrale disposti perpendicolarmente alla superficie corticale dell'encefalo. I neuroni appartenenti ad una stessa ipercolonna hanno campi ricettivi quasi identici, ovvero elaborano le stesse porzioni di informazione. [9]

Considerato che le reti convoluzionali imitano la struttura della corteccia visiva animale, è possibile estendere il concetto di ipercolonna anche nel campo del deep learning: un'ipercolonna è un vettore che raccoglie i valori di attivazione di ogni strato di una rete convoluzionale relativi ad un determinato pixel.

Tipicamente le applicazioni basate su ConvNet utilizzano soltanto le informazioni contenute nell'ultimo strato, generalmente un fully-connected layer. Per applicazioni che operano a diversi livelli di granularità tuttavia si dimostra utile ragionare su multipli livelli di astrazione e grandezza e quindi sfruttare anche le informazioni distribuite negli altri livelli della rete; ciò diventa possibile con l'uso di ipercolonne. Esse infatti consentono di sfruttare anche le feature contenute nei livelli intermedi della rete (hidden layers), caratterizzate da un maggior livello di dettaglio a fronte di una minore sensibilità semantica.

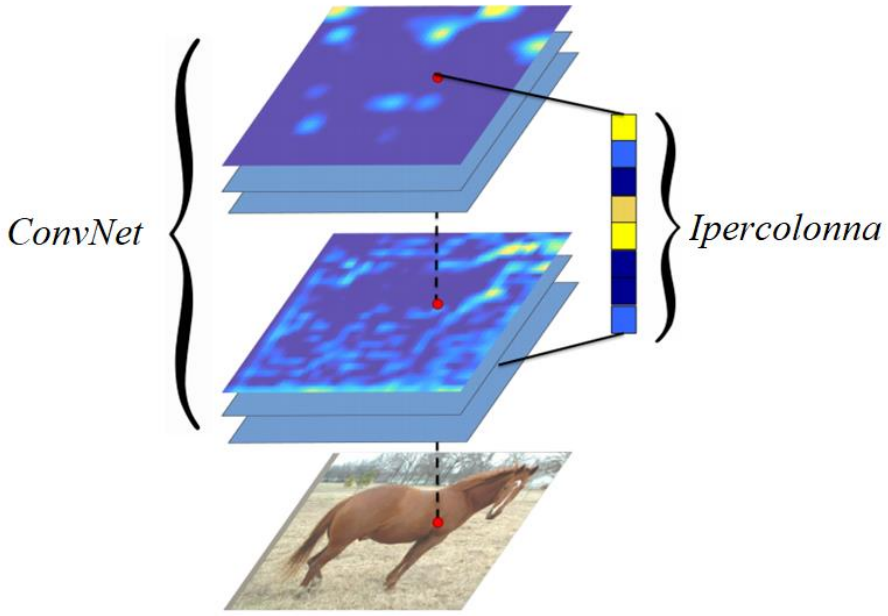


Figura 6.1: *Rappresentazione di un'ipercolonna. L'immagine in basso è l'input mentre le altre rappresentano le feature map degli strati intermedi della rete neurale convoluzionale. L'ipercolonna di un pixel è il vettore delle attivazioni di tutte le unità che corrispondono al pixel. Fonte: [9]*

Lo strumento delle ipercolonne è oggi utilizzato in molti task di computer vision basati su reti convoluzionali quali segmentazione o predizione di punti chiave di un'immagine, introducendo un sensibile miglioramento nella qualità dei risultati.

6.2 – Estrazione di feature intrepretabili con ipercolonne

Il processo di estrazione delle feature nel dominio delle immagini mira ad ottenere una segmentazione semanticamente rilevante. In altre parole, le feature devono essere in qualche modo riconducibili alle caratteristiche dell'immagine, e quindi agli oggetti presenti oppure a gruppi o porzioni di essi. Nel presente lavoro di tesi, al fine di ottenere feature esplicitamente interpretabili e semanticamente valide si è deciso di avvalersi dello strumento delle ipercolonne [10] applicato alla medesima rete neurale convoluzionale che opera il task di classificazione e della quale si vuole appurare l'affidabilità delle predizioni.

Figura 6.2: *Esempio di estrazione di feature per mezzo di ipercolonne. Dall'immagine di partenza (in alto a sinistra) si sono ottenute tre feature chiaramente riconducibili a oggetti distinti.*

6.2.1 Processo di estrazione delle feature

Data un'immagine di $m \times n$ pixel e un modello basato su ConvNet adeguatamente "allenato", il processo di estrazione delle feature si sviluppa nei seguenti passaggi:

1. Si alimenta la rete neurale convoluzionale con un'immagine. In corrispondenza di ogni pixel dell'immagine i neuroni della rete saranno "eccitati" in maniera diversa in base alla posizione dei layer corrispondenti all'interno modello. Raccogliendo i valori di attivazione di uno strato si compone una cosiddetta *feature map*. Si ottiene pertanto una serie di feature map il cui numero è pari al numero di layer della rete (altezza del modello).

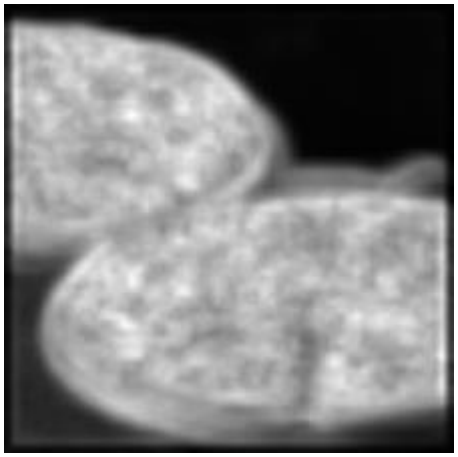


Figura 6.3: *Visualizzazione di una feature map. La luminosità dei pixel indica il grado di eccitazione del neurone corrispondente.*

2. Per ogni pixel dell'immagine si provvede a estrarre la relativa ipercolonna concatenando i valori di attivazione di ogni feature map nella posizione corrispondente al pixel di partenza. Si ottiene una matrice di $m \times n \times h$ valori, dove h rappresenta l'altezza delle ipercolonne ed è un parametro legato all'altezza della rete.
3. Dal volume ottenuto si scartano i valori corrispondenti ai livelli più bassi della rete. I primi layer di una ConvNet infatti sono caratterizzati da feature map ad alto livello di dettaglio ma prive di un sostanziale valore semantico, e sono quindi portati ad individuare feature poco significative per i nostri fini, come ad esempio i bordi degli oggetti. Al contrario gli ultimi layer, a fronte di una scarsa risoluzione a causa dei vari strati di pooling, offrono feature semanticamente rilevanti. Il volume si riduce quindi ad una nuova matrice di dimensione $m \times n \times h'$ dove h' corrisponde a circa un terzo dell'altezza originale delle ipercolonne.

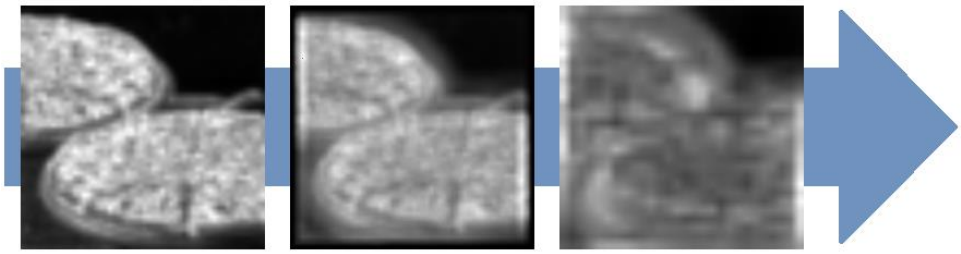


Figura 6.4: *Selezione di tre feature map corrispondenti a diversi livelli di una ConvNet. Si noti come con l'aumentare del livello si abbia un'evidente perdita di risoluzione delle immagini a fronte di un sostanziale guadagno in termini semantici.*

4. Infine si applica un algoritmo di *clustering* (*kmeans* in questo caso) sulle $m \times n$ ipercolonne ottenute. In questo modo si ottiene un raggruppamento delle ipercolonne in base al loro stato generale di eccitazione. Si ricavano quindi k

cluster di ipercolonne che, proiettati sull'immagine, restituiscono un partizionamento interpretabile della stessa.

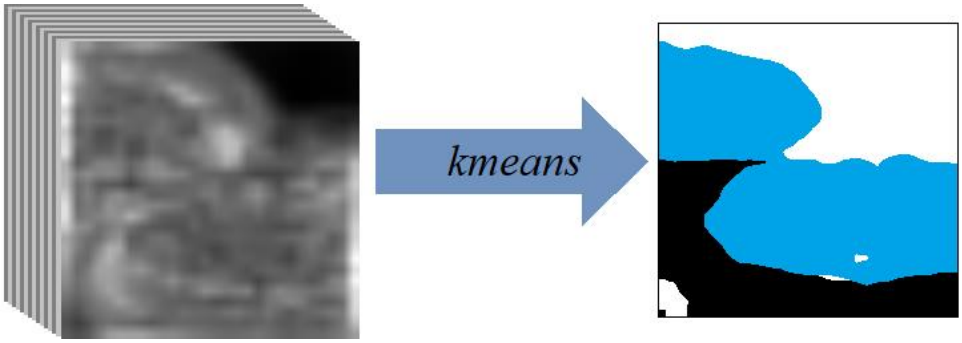


Figura 6.5: *Feature map risultante dall'applicazione dell'algoritmo kmeans (k = 3) sulla matrice di ipercolonne.*

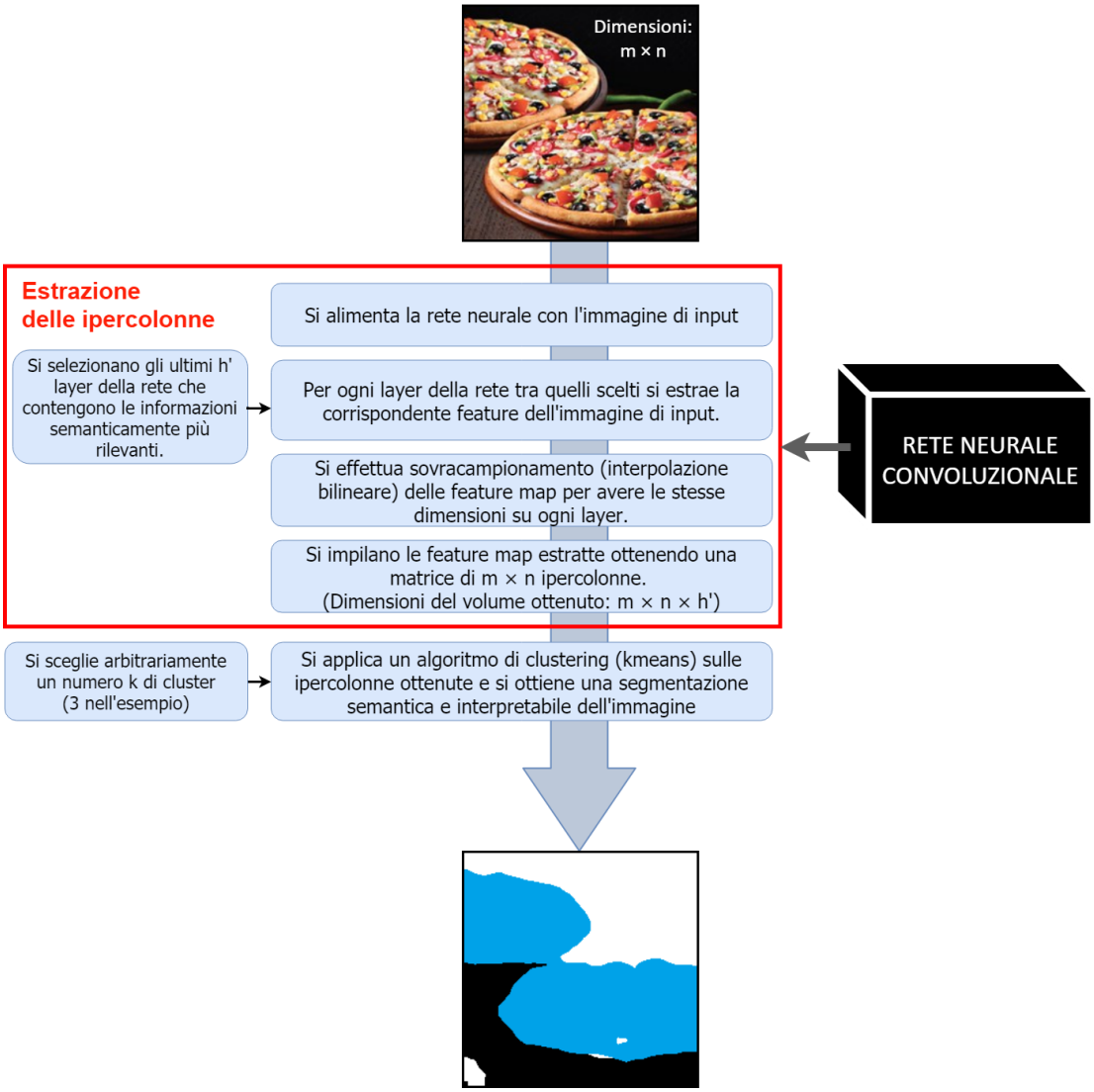


Figura 6.6: *Schema riassuntivo del processo di estrazione delle feature mediante ipercolonne.*

6.2.2 Contributo innovativo

Le feature estratte con il sistema proposto presentano i seguenti vantaggi:

- Rispecchiano il modo in cui la ConvNet "vede" l'immagine proposta e "viene stimolata" dai suoi pixel. Le porzioni d'immagine appartenenti alla stessa feature racchiudono quindi tutti i pixel che sono percepiti in maniera simile dagli strati della rete scelti.
- Le ipercolonne raccolgono il contributo di più feature map insieme, ciò permette di ottenere un buon compromesso tra risoluzione spaziale e consistenza semantica.
- Le feature estratte sono anche facilmente riconducibili agli oggetti presenti nell'immagine e di conseguenza risultano interpretabili per un osservatore umano.

6.3 – Perturbazione di immagini

Una volta completata l'estrazione delle feature, si passa alla fase di perturbazione degli input.

Per ogni area dell'immagine individuata da una feature si provvede ad applicare un'elaborazione grafica al fine di annullare o minimizzare quanto più possibile l'apporto della feature al risultato finale della classificazione. Questa elaborazione grafica si rende necessaria in quanto non è possibile rimuovere concretamente l'informazione da un'immagine senza pregiudicarne l'integrità della struttura matriciale $m \times n$.

6.3.1 Tecniche di perturbazione e criticità

Il tipo di elaborazione che si applica alle immagini è un aspetto critico del processo perché prevede la sostituzione di parte dei dati dell'immagine originale con nuove informazioni che potrebbero pesantemente inficiare i risultati. Si ricordi infatti che le immagini perturbate dovranno essere successivamente proposte al classificatore al fine di ricavare una stima dell'impatto delle singole feature sul risultato di classificazione.

Si propongono adesso le differenti tecniche di elaborazione analizzate nel presente lavoro di tesi:

- **Cancellazione**: si sostituiscono le aree interessate con un colore distribuito uniformemente. La presenza di queste aree uniformi tuttavia è un significativo elemento di disturbo per la ConvNet che rischia di essere "ingannata" e stravolgere le proprie predizioni anche a fronte di aree perturbate molto piccole. Questo tipo di perturbazione si è quindi rivelato inadeguato ai fini della tesi.

$$pix[i,j] = (255, 255, 255)_{rgb} \tag{6.1}$$

- **Rumore**: applicazione di un operatore randomico ai pixel delle aree interessate. Anche questa tipologia di perturbazione si è dimostrata sostanzialmente inapplicabile, a causa di un fattore di disturbo molto pesante del rumore sul modello di classificazione.

$$pix[i,j] = (rand, rand, rand)_{rgb} \tag{6.2}$$

oppure

$$pix[i,j] = pix[i + rand, j + rand] \tag{6.3}$$

- **Blur**: la tecnica del *blur* (o sfocatura) sostituisce il colore dei pixel con il valore medio dei pixel vicini, ovvero compresi entro un certo raggio *r*. In questo modo si ha un'attenuazione della feature e, di conseguenza, del suo impatto sul risultato di classificazione, senza introdurre sostanziali elementi di disturbo per la rete convoluzionale.

$$pix[i,j] = \frac{\sum_{\substack{i-r \le m \le i+r \\ i-r<n<i+r}} pix[m,n]}{r^2} \tag{6.4}$$

Sulla base di queste analisi, nel presente lavoro di tesi si è scelto di adoperare la tecnica del blur come metodologia di perturbazione degli input. Bisogna tuttavia considerare che, vista la grande variabilità delle caratteristiche delle immagini che si possono sottoporre ad un classificatore, non è possibile individuare un metodo di perturbazione che sia universalmente il migliore per ogni tipo di immagine.


(a)


(b)


(c)

Figura 6.7: *Perturbazione di una feature con i tre metodi analizzati:*

(a) *Cancellazione. In questo caso si è sostituita una feature con un'area di colore bianco. Si elimina l'apporto della feature originale ma la regione uniforme introdotta costituisce un significativo elemento di disturbo per il classificatore.*

(b) Applicazione dell'operatore rumore. Anche questo caso l'elemento di disturbo per il classificatore è molto forte.

(c) Blur. In questo caso si ha una buona attenuazione della feature in questione a fronte di un disturbo trascurabile per il modello.

Figura 6.8: *Risultato della perturbazione delle tre feature individuate in figura 6.2 con la tecnica del blur. Nella colonna di sinistra si hanno le immagini originali con le singole feature in evidenza, nella colonna di destra le versioni perturbate (sfocate).*

Capitolo 7

Estrazione e perturbazione di feature – Testi

In questo capitolo si analizzerà la soluzione sviluppata per l'estrazione di feature e la loro perturbazione nel caso di documenti testuali.

Come nel caso delle immagini, si operano delle perturbazioni su feature interpretabili. Tuttavia, in questo caso il processo di estrazione delle feature non sfrutta informazioni ricavate con l'ausilio della rete neurale usata per la classificazione, bensì prevede un'elaborazione preliminare sui dati di training del classificatore con l'uso di tecniche di *information retrieval*.

7.1 – Strumenti di Information Retrieval

La disciplina dell'*information retrieval* (*IR*) si occupa della rappresentazione semantica ed elaborazione di informazioni contenute in dati non strutturati di varia natura, solitamente documenti testuali o pagine web [11]. In questa sezione si presenteranno le tecniche di information retrieval adoperate nel processo di estrazione semantica delle feature interpretabili, e i concetti matematici che vi stanno alla base.

7.1.1 Pesatura Tf-idf

In ambito information retrieval un documento viene generalmente rappresentato come una *bag of words* (borsa di parole), ovvero come l'insieme dei termini (*token*) in esso presenti. Le tecniche di pesatura dei token sono largamente utilizzate all'interno dei processi di analisi semantica di dati testuali. Queste tecniche puntano a fornire una stima dell'importanza dei singoli termini all'interno di un documento o di una collezione. Il peso assegnato ad un token può essere infatti locale, quindi relativo ad un singolo documento, oppure globale, relativo all'intera collezione (detta *corpus*).

La pesatura Tf-idf (term frequency – inverse document frequency) è una delle tecniche più adoperate in tale ambito e combina gli effetti di una pesatura locale ed una globale.

- *Term frequency:* peso locale che misura il numero di occorrenze di un token all'interno di un testo. Il valore normalizzato di *tf* del termine *i* nel documento *j* viene calcolato come il numero di occorrenze n_{ij} diviso per il numero di parole del documento $|d_j|$.

$$tf_{ij} = \frac{n_{ij}}{|d_j|} \tag{7.1}$$

- *Inverse Document Frequency:* peso globale inversamente proporzionale al numero di occorrenze di un termine in una collezione di documenti. Si ricava come logaritmo del rapporto tra il numero totale di documenti nella collezione $|D|$ e il numero di documenti che contengono il termine *i*.

$$idf_i = log \frac{|D|}{|\{d : i \in d\}|} \tag{7.2}$$

Il peso tf-idf misura quindi l'importanza di un termine *i* appartenente ad un documento *j* all'interno di una collezione di documenti aumentando proporzionalmente al numero di occorrenze del termine all'interno del documento e in maniera inversamente proporzionale alla frequenza del termine nella collezione.

$$tfidf_{ij} = tf_{ij} \times idf_i \tag{7.3}$$

7.1.2 Latent Semantic Indexing

La tecnica del *Latent Semantic Indexing* (nota anche come *Latent Semantic Analysis*) viene utilizzata in ambito *natural language processing* per l'estrazione dei *concetti* o *componenti principali* che formano un documento. LSA infatti sfrutta la cosiddetta *ipotesi distribuzionale*, e assume che parole semanticamente vicine appariranno in documenti tra loro affini.

L'applicazione di LSA richiede che il corpus venga inizialmente rappresentato come una matrice *termini-documenti*, tipicamente molto sparsa. Questa matrice viene fattorizzata con la tecnica algebrica della *decomposizione ai valori singolari* (*SVD*) e di conseguenza significativamente ridotta nel numero di righe (corrispondenti alle parole) senza però perdere le relazioni di affinità tra le colonne (relative ai singoli documenti). Si voglia considerare una matrice termini-documenti C a valori reali e di dimensione $m \times n$. La decomposizione di C con SVD è la seguente:

$$C = U\Sigma V^T \tag{7.4}$$

dove U è detta la *matrice di similarità termini-concetti* ed è una matrice unitaria di dimensioni $m \times m$, Σ è detta *matrice dei concetti* e consiste in una matrice diagonale

rettangolare di dimensioni $m \times n$ e V^T è la *matrice di similarità documenti-concetti* e risulta essere la trasposta coniugata di una matrice unitaria di dimensioni $n \times n$. Si ha che:

- Ciascuna delle m colonne di U è detta *vettore singolare sinistro* di C, i vettori singolari sinistri sono gli autovettori di CC^T.
- Ciascuna delle n colonne di V è detta *vettore singolare destro* di C, i vettori singolari destri sono gli autovettori di C^TC.
- Gli elementi disposti lungo la diagonale di Σ sono detti *valori singolari* di C e corrispondono alle radici quadrate degli autovalori non nulli di CC^T e C^TC.

Questa fattorizzazione prende il nome di *SVD completa*. Nella pratica tuttavia ad essere solitamente adoperata è la cosiddetta *SVD ridotta* o *troncata*. La SVD troncata prevede che la matrice C termini-documenti sia approssimata con una matrice di rango k.

$$C_k \approx U_k \Sigma_k {V^T}_k \tag{7.5}$$

Dove il rango k rappresenta il numero di concetti al quale si vuole ridurre la matrice C. In questo modo si ottiene quindi una significativa riduzione della dimensionalità del corpus e la trasformazione dei vettori dei documenti in vettori di concetti.

7.1.3 Confronto fra termini con LSI e similarità coseno

L'uso del Latent Semantic Indexing e della SVD ridotta permette di effettuare un confronto tra termini appartenenti allo stesso corpus. Ciò è possibile attraverso l'uso della metrica similarità coseno che misura la similitudine tra due vettori o matrici attraverso il calcolo del coseno tra i due.

$$cosine\ similarity = \frac{A * B}{\|A\|\|B\|} \tag{7.6}$$

Per confrontare quindi un determinato termine con tutti gli altri che appartengono allo stesso corpus e ottenere quindi i rispettivi valori di correlazione tra termini si procede nel modo seguente.

1. Dal corpus si ricava la matrice SVD troncata C_k di dimensione $m \times k$ dove m corrisponde al numero di termini nel corpus e k rappresenta il numero dei concetti, scelto arbitrariamente, al quale si è ridotta la matrice.
2. Dato un termine i si estrae la colonna corrispondente $C_k[i]$ dalla matrice SVD ridotta.

3. Si calcola la similarità coseno tra il vettore $C_K[i]$ e la matrice C_K ottenendo come risultato l'elenco di tutti i termini del corpus corredato dai valori di correlazione semantica con il termine *i*.

$$cosine\ similarity(C_k[i], C_k) \tag{7.7}$$

Le informazioni derivanti da questo processo possono essere utilizzate per costruire cluster semantici di termini in cui le parole vengono raggruppate in base a quanto spesso co-occorrono.

7.2 – Estrazione di feature intrepretabili con IR

Il processo di estrazione di feature nel dominio dei testi mira ad estrapolare dei gruppi di parole semanticamente correlate e, come nel caso delle immagini, le feature devono essere riconducibili a concetti distinti e interpretabili del testo. Nel presente lavoro di tesi, al fine di ottenere feature interpretabili e semanticamente valide si è deciso di avvalersi degli strumenti di information retrieval presentati nella sezione precedente. Il processo di estrazione prevede delle operazioni preliminari effettuate sull'intero corpus usato anche per il training del classificatore.

```
WASHINGTON — Inflight connectivity provider Gogo is
causing sparks to fly over its claim that leasing capacity
is superior to owning satellites and that proponents of
pure satellite ownership are only promoting ownership
models to protect their business. Gogo Chief Executive
Michael Small defended the Chicago company's Ku-band
leasing strategy during a May 4 company earnings call,
describing leasing capacity as "far more cost-effective"
than owning satellites.


Features:
['protect' 'executive' 'ownership' 'owning' 'business']
['cost' 'inflight' 'satellite' 'fly']
```

Figura 7.1: *Esempio di estrazione di feature. Dal testo si sono ricavati 2 diversi concetti, ciascuno identificato da un gruppo di termini.*

7.2.1 Operazioni preliminari per l'estrazione di feature

Nel presente lavoro di tesi, il processo di estrazione delle feature richiede una serie di elaborazioni preliminari sul corpus che viene adoperato per il training del classificatore. Queste operazioni preliminari si possono riassumere nei seguenti passaggi:

1. **Rappresentazione matriciale del corpus**: dopo le opportune operazioni di pulizia dei dati (normalizzazione, eliminazione stopwords e lemmatizzazione) si applica la tokenizzazione che restituisce la matrice termini-documenti C relativa all'intero corpus.
2. **Identificazione delle keyword con tf-idf**: sulla matrice C si opera la pesatura tf-idf ottenendo un'altra matrice $C_{tf\text{-}idf}$ che contiene i pesi relativi a tutti i termini. Si scorre quindi la matrice $C_{tf\text{-}idf}$ e selezionando i termini in base ai valori della pesatura si ottengono dei gruppi di keyword (contraddistinte da pesi superiori ad una soglia definita arbitrariamente). Queste keyword andranno a formare l'*insieme di partenza* delle feature.
3. **Estrazione dei concetti con SVD troncata**: per estrarre i concetti dal corpus si applica la fattorizzazione SVD troncata al k-esimo termine sulla matrice $C_{tf\text{-}idf}$ pesata con tf-idf nel passaggio precedente. La matrice ottenuta C_{svd} è una matrice temini-concetti relativa a k concetti e viene memorizzata per essere usata successivamente nel processo vero e proprio di estrazione delle feature.

7.2.2 Processo di estrazione delle feature

Si introduce adesso il processo di estrazione delle feature operato a runtime su un documento testuale *d*.

```
WASHINGTON – Inflight connectivity provider Gogo is causing
sparks to fly over its claim that leasing capacity is superior
to owning satellites and that proponents of pure satellite
ownership are only promoting ownership models to protect their
business. Gogo Chief Executive Michael Small defended the
Chicago company's Ku-band leasing strategy during a May 4
company earnings call, describing leasing capacity as "far
more cost-effective" than owning satellites.
```

Figura 7.2: *Esempio di un documento testuale sul quale verrà applicato il processo di estrazione delle feature.*

Il processo di estrazione delle feature si sviluppa nei seguenti passaggi:

1. **Preparazione del documento**: si applicano le tecniche di normalizzazione dei termini, eliminazione delle stopwords e lemmatizzazione sul documento. Infine con la tokenizzazione si ottiene la lista (*bag of words*) di termini normalizzati sulla quale si opera l'estrazione delle feature.

```
washington inflight connectivity provider gogo causing
spark fly claim leasing capacity superior owning satellite
proponent pure satellite ownership promoting ownership
model protect business gogo chief executive michael small
defended chicago company ku band leasing strategy may 4
company earnings call describing leasing capacity far cost
effective owning satellite
```

Figura 7.3: *Il documento di figura 7.2 dopo le operazioni di pulizia del testo*

2. **Ricerca delle keyword**: per ogni token del testo si controlla se appartiene alla lista delle keyword ricavata in precedenza. Tutte le keyword individuate nel documento costituiranno l'insieme delle feature di partenza.

```
'satellite'
'executive'
'inflight'
'ownership'
```

Figura 7.4: *Nel testo mostrato in figura 7.3 si identificano le seguenti feature di partenza*

3. **Aggregazione dei concetti**: una volta individuate le feature di partenza, da queste si vogliono ottenere dei cluster di termini semanticamente significativi. A questo scopo si sfrutta la matrice C_{svd} che contiene le informazioni sui concetti del corpus usato per il training del classificatore. Per ogni termine del documento è possibile trovare una serie di parole semanticamente correlate con l'applicazione della similarità coseno. In base alle parole correlate individuate per ogni termine del documento, si costruiscono i gruppi di parole che costituiranno le feature definitive. In caso di parole comuni a due o più gruppi di termini si controllano i valori di correlazione di tali parole con i gruppi e si mantengono i termini solo nei gruppi con maggiore correlazione.

In questo modo ci si assicura che i cluster saranno disgiunti, ossia con intersezione sempre nulla.

```
['ownership' 'protect' 'owning' 'business']

['inflight' 'fly']

['satellite' 'connectivity']

['executive' 'chief' 'company']
```

Figura 7.5: *Risultato finale del processo di estrazione. Si sono individuate quattro feature semanticamente coerenti.*

7.3 – Perturbazione di testi

Una volta completata l'estrazione delle feature, si passa alla fase di perturbazione degli input. Per ogni gruppo di termini definito da feature si provvede a rimuovere tutte le occorrenze dei termini nel testo al fine di annullare l'influenza della feature sul risultato finale della classificazione.

```
Feature rimossa: 'ownership' 'protect' 'owning' 'business'

WASHINGTON — Inflight connectivity provider Gogo is causing
sparks to fly over its claim that leasing capacity is
superior to owning satellites and that proponents of pure
satellite ownership are only promoting ownership models to
protect their business. Gogo Chief Executive Michael Small
defended the Chicago company's Ku-band leasing strategy
during a May 4 company earnings call, describing leasing
capacity as "far more cost-effective" than owning
satellites.
```

Figura 7.6: *Esempio di perturbazione di una feature. Tutte le occorrenze dei termini appartenenti alla feature, evidenziate in rosso, vengono rimosse dal testo.*

Capitolo 8

Analisi sperimentale

Nel seguente capitolo si presenteranno i risultati sperimentali ottenuti dall'applicazione delle soluzioni di trasparenza sviluppate per immagini e documenti testuali. L'analisi sperimentale è stata realizzata con l'ausilio di 3 diversi modelli di classificazione basati su reti neurali convoluzionali, ciascun modello sarà introdotto da una breve panoramica sulla struttura della rete neurale e sugli eventuali dataset adoperati. Per ogni modello si proporranno degli esempi di classificazione di immagini o testi e si forniranno delle spiegazioni alle predizioni con l'ausilio di elaborazioni grafiche e tabelle riassuntive. La prima sezione del capitolo sarà dedicata dedicata ai tool di sviluppo impiegati nella realizzazione del framework.

8.1 – Strumenti di sviluppo e framework

In questa sezione si introdurranno i framework di machine learning adoperati nel presente lavoro di tesi. Per i modelli di deep learning usati nel presente lavoro di tesi si è deciso di scegliere il framework *Keras* con la libreria *Tensorflow* come backend.

8.1.1 Tensorflow

Tensorflow è una libreria software open source sviluppata da Google Brain e rilasciata sotto licenza Apache 2.0 per applicazioni di machine learning e, in particolar modo, modelli di deep learning come reti neurali. Definita come una "libreria di machine learning di seconda generazione", Tensorflow si basa su Python ma mette a disposizione varie API native di alto e basso livello in anche linguaggio C/C++, Java e R. Alla base del framework Tensorflow vi è il concetto di *flow graph*. Un flow graph è un grafo diretto aciclico che rappresenta una computazione numerica. I nodi del grafo rappresentano delle operazioni matematiche, gli archi tra i nodi invece sono dei cosiddetti *tensori*, ovvero vettori multidimensionali di dati. Questo approccio offre la

possibilità di operare ad alti livelli di astrazione per l'esecuzione di task numerici di basso livello. [12]

Le principali qualità di Tensorflow sono rappresentate dalle elevate performance, dalla possibilità di monitorare la fase di training del modello secondo un'ampia gamma di metriche e dall'ottimo supporto fornito da una community sempre più vasta. Infatti, Tensorflow è oggi una delle librerie di machine learning più diffuse ed è impiegata in numerosi servizi di Google, come Google Maps, Gmail e Ricerca.

8.1.2 Keras

Per quanto Tensorflow risulti essere un'ottimo framework per applicazioni di deep learning, la creazione di modelli rimane un processo complesso, in particolar modo per utenti poco esperti. Il framework open source *Keras* risponde all'esigenza di definire modelli di deep learning in manera intuitiva, modulare e con un numero contenuto di righe di codice. La principale peculiarità di questa libreria è la capacità di funzionare come interfaccia di librerie di più basso livello come Tensorflow o Theano. Keras fornisce tutte le più comuni implementazioni dei modelli di deep learning con una perdita di prestazioni contenuta rispetto a Tensorflow. A partire dal 2017, il team di Tensorflow ha deciso di aggiungere dei moduli per il supporto di Keras nella propria libreria. [13]

8.2 – Classificatori e dataset utilizzati

In questa sezione verranno presentati i diversi classificatori, basati su deep learning, utilizzati per l'analisi sperimentale delle soluzioni di trasparenza proposte nei capitoli precedenti. Per il dominio delle immagini si è deciso di utilizzare il modello pre-allenato VGG-16, basato su ConvNet e con 1000 diverse classi di oggetti. Per la prova delle tecniche sviluppate nel dominio dei documenti testuali invece si è provveduto alla progettazione e al training di altre due reti neurali convoluzionali utilizzando i dataset *20-newsgroup* e *Large Movie Review*.

8.2.1 Modello VGG-16

Per l'applicazione delle soluzioni di trasparenza sviluppate nell'ambito di un classificatore di immagini si è scelto di utilizzare il modello *VGG-16*. Si tratta di un modello *pre-trained*, ovvero un modello già funzionante che ha precedentemente affrontato la frase di training su un dataset e contiene pesi e bias che rappresentano le

feature del dataset adoperato. L'utilizzo di un modello pre-trained si è dimostrato di grande utilità nell'ottica di risparmiare tempo e risorse di calcolo. [14][15]

Il modello VGG-16 è stato sviluppato e allenato dal team VGG (Visual Geometry Group) dell'Università di Oxford in occasione della competizione di image recognition ILSVRC-2014 (*ImageNet Large Scale Visual Recognition Challenge 2014*). L'obiettivo della competizione era quello di allenare una rete neurale che potesse classificare correttamente un'immagine di input in 1.000 classi separate di oggetti. Il training è avvenuto su un dataset comprendente circa 1.2 milioni di immagini, più altre 100.000 immagini per la fase di testing.

Il modello è basato su ConvNet e comprende 16 layer di elaborazione. La struttura si distingue per la sua semplicità, adoperando solo layer convoluzionali bidimensionali di kernel 3 × 3 sovrapposti con profondità crescente. La riduzione delle dimensioni del volume è gestita dall'operatore Maxpooling. Alla fine della rete due fully-connected layer, ciascuno da 4096 nodi, sono seguiti da una funzione *softmax* di classificazione.

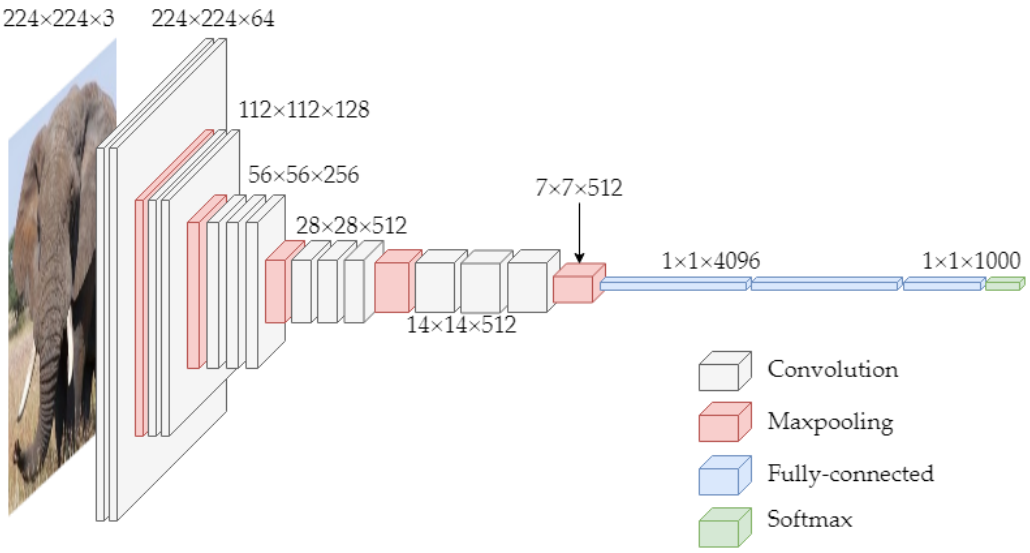


Figura 8.1: *Rappresentazione della struttura del modello VGG-16.*

La struttura del modello VGG-16 è quindi la seguente:

1. **Input**: immagine RGB (224 × 224)
2. **Layer di convoluzione 2D**: profondità = 64, dim. kernel = 3
3. **Layer di convoluzione 2D**: profondità = 64, dim. kernel = 3
4. **Maxpool 2D**: sottocampionamento di fattore 2
5. **Layer di convoluzione 2D**: profondità = 128, dim. kernel = 3
6. **Layer di convoluzione 2D**: profondità = 128, dim. kernel = 3
7. **Maxpool 2D**: sottocampionamento di fattore 2
8. **Layer di convoluzione 2D**: profondità = 256, dim. kernel = 3
9. **Layer di convoluzione 2D**: profondità = 256, dim. kernel = 3
10. **Layer di convoluzione 2D**: profondità = 256, dim. kernel = 3
11. **Maxpool 2D**: sottocampionamento di fattore 2
12. **Layer di convoluzione 2D**: profondità = 512, dim. kernel = 3
13. **Layer di convoluzione 2D**: profondità = 512, dim. kernel = 3
14. **Layer di convoluzione 2D**: profondità = 512, dim. kernel = 3
15. **Maxpool 2D**: sottocampionamento di fattore 2
16. **Layer di convoluzione 2D**: profondità = 512, dim. kernel = 3
17. **Layer di convoluzione 2D**: profondità = 512, dim. kernel = 3
18. **Layer di convoluzione 2D**: profondità = 512, dim. kernel = 3
19. **Maxpool 2D**: sottocampionamento di fattore 2
20. **Fully-Connected**: 4096
21. **Fully-Connected**: 4096
22. **Fully-Connected**: 1000
23. **Soft-max**

Il modello VGG-16 viene molto utilizzato in ambito di ricerca per task di classificazione e object detection in quanto risulta essere particolarmente accurato nonostante l'enorme quantità di classi:

- Accuratezza del primo risultato: 70.5 %
- Accuratezza dei primi 5 risultati: 90.0 %

8.2.2 Modello per la classificazione di documenti testuali

Per l'applicazione delle tecniche di trasparenza nel dominio dei testi si è reso necessario procedere alla progettazione di una rete neurale.

Il modello sviluppato si basa su ConvNet ed è composto da 4 layer di elaborazione. La rete riceve in input una sequenza di 600 termini mappati come valori numerici interi attraverso l'uso dei vettori pre-trained GloVe (Global Vectors for Word

Representation) [16]. L'input viene processato attraverso tra layer convoluzionali monodimensionali con kernel di dimensione 5, ciascuno seguito da un layer di pooling. Il modello comprende anche due layer di *dropout* che settano randomicamente a 0 il 25% dei dati in input così da evitare il fenomeno dell'*overfitting* durante la fase di training. In cima al modello si ha un layer fully-connected di 128 neuroni con la funzione di attivazione softmax che si occupa della classificazione.

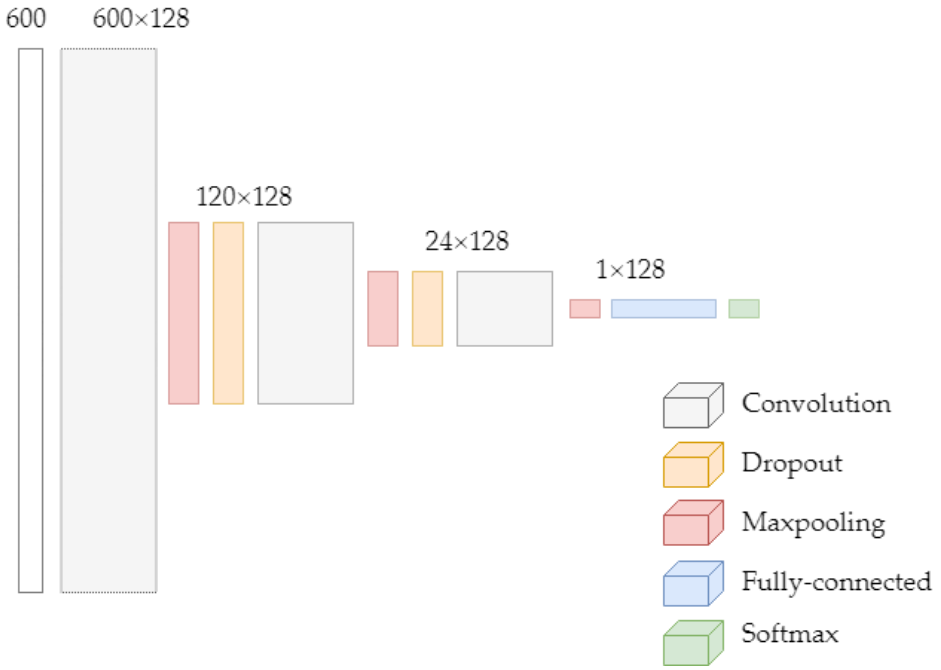


Figura 8.2: *Rappresentazione della struttura del modello per la classificazione dei testi.*

La struttura della rete è la seguente:

1. **Input**: sequenza di termini (lunghezza: 600)
2. **Layer di convoluzione 1D**: profondità = 128, dim. kernel = 5
3. **Maxpool 1D**: sottocampionamento di fattore 2, dim.kernel = 5
4. **Dropout**: rate = 0.25
5. **Layer di convoluzione 1D**: profondità = 128, dim. kernel = 5
6. **Maxpool 1D**: sottocampionamento di fattore 2, dim.kernel = 5
7. **Dropout**: rate = 0.25
8. **Layer di convoluzione 1D**: profondità = 128, dim. kernel = 5
9. **Maxpool globale 1D**
10. **Fully-connected:** 128
11. **Soft-max**

La fase di training del modello è avvenuta con due dataset testuali, molto diversi per caratteristiche e impostazione.

Con il primo dataset, *Large Movie Review*, composto da 50.000 recensioni etichettate come positive o negative, si è ottenuta un'accuratezza del 87,5%.

Con il secondo dataset *20-newsgroup*, comprendente circa 20.000 documenti appartenenti a 20 diverse classi, si è invece raggiunta un'accuratezza del 76,2%.

8.2.3 Large Movie Review Dataset

Il *Large Movie Review Dataset* [17] viene comunemente utilizzato per task di *sentiment analysis*, ovvero classificazione binaria con una classe positiva e una negativa. Il dataset contiene 50.000 recensioni di film provenienti dal popolare aggregatore IMDb. La distribuzione delle etichette è bilanciata (25.000 recensioni positive e 25.000 negative) e non si hanno più di 30 recensioni per un film perché le recensioni per lo stesso film tendono ad avere voti correlati. Sono considerate recensioni negative quelle con un voto minore o uguale a 4 e positive quelle con voto maggiore o uguale a 7.

8.2.4 Dataset 20-newsgroup

Il dataset *20-newsgroup* [18] è una raccolta di circa 20.000 documenti appartenenti a 20 diversi newsgroup (gruppi di discussione) corrispondenti a topic differenti. Alcuni dei newsgroup del dataset sono accomunabili per la vicinanza degli argomenti proposti (ad esempio *comp.sys.ibm.pc.hardware* e *comp.sys.mac.hardware*).

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast
rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	talk.religion.misc alt.atheism soc.religion.christian

Tabella 8.3: *Lista delle 20 classi del dataset. I newsgroup sono stati raggruppati in base alla vicinanza dei reativi topic.*

Il dataset 20-newsgroup è ad oggi una delle collezioni più utilizzate dalla comunità scientifica per esperimenti di machine learning in ambito *natural language processing* basati su task di classificazione o clustering.

8.3 – Risultati sperimentali – VGG-16

In questa sezione verranno riportati i risultati sperimentali ottenuti dall'applicazione delle soluzioni di trasparenza algoritmica sviluppate nel dominio delle immagini. Si presenteranno una serie di esempi relativi a immagini che sono state proposte al classificatore VGG-16 basato su reti convoluzionali. Per ciascuna immagine si proporranno le feature estratte con con lo strumento delle ipercolonne e per ciascuna di queste feature si valuterà l'impatto sulla predizione con l'ausilio di report di trasparenza relativi ad una o più classi di interesse. I report di trasparenza comprendono delle rappresentazioni grafiche delle influenze e delle tabelle dove l'impatto delle feature principali è misurato con le metriche mostrate nel capitolo 5. Le rappresentazioni grafiche si sono ottenute colorando le aree dell'immagine individuate dalle feature in base ai valori dell'indice nIRI. Le feature con influenza negativa presentano quindi una colorazione rossa più o meno accentuata in base all'influenza della feature mentre, al contrario, le feature con influenza positiva presentano una colorazione verde. Per le feature che presentano valori di nIRI vicini allo zero si è invece applicata una colorazione gialla per indicare l'impatto nullo o trascurabile sulla predizione. In chiusura del capitolo si presenterà l'analisi locale del comportamento del modello VGG-16 relativamente alle classi 'jellyfish', 'pizza', 'goose' e 'hotdog'. Verrà proposta infine una valutazione dell'efficacia del framework sviluppato relativamente al modello VGG-16 per quanto riguarda l'individuazione e la valutazione delle influenze positive e negative delle feature.

Le 10 immagini proposte per l'analisi di trasparenza sono le seguenti:

- Mouse (classificata come 'hand blower')
- Elefante africano (correttamente classificata come 'african elephant')
- Pizza (correttamente classificata come 'pizza')
- Limone (correttamente classificata come 'lemon')
- Jack o' Lantern (correttamente classificata come 'jack o' lantern')
- Medusa (correttamente classificata come 'jellyfish')
- Segnale stradale (classificata come 'boathouse')
- Kimono (classificata come 'vase')
- Bicchiere di birra (classificata come 'beer bottle')
- Gondola (classificata come 'dock')

8.3.1 Mouse

Primi 5 risultati:
hand_blower: 0.2620838
washbasin: 0.13937736
soap_dispenser: 0.11906022
toilet_seat: 0.087728485
toilet_tissue: 0.073539287

In questo primo esempio si sottopone al modello un'immagine contenente un comune mouse per computer. La classe 'mouse' tuttavia non compare tra le prime 5 predizioni del modello, che invece identifica prevalentemente classi relative ad articoli da bagno. Per motivare l'errata classificazione dell'immagine formuleremo una query di trasparenza relativa alla classe 'mouse' insieme ad altre query relative alle classi erroneamente predette da 'hand blower' a 'toilet tissue'.

Figura 8.4: *Risultato dell'estrazione di 10 feature mediante clustering di ipercolonne. Le aree di colore diverso corrispondono alle diverse feature. Ciascuna di queste feature sarà opportunamente perturbata al fine di misurarne l'influenza.*

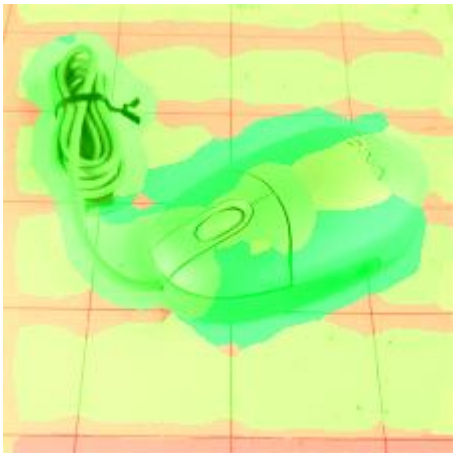


Figura 8.5: *Effettuando le predizioni sugli input perturbati e analizzando la classe di interesse 'mouse' si ottiene una rappresentazione grafica dove le feature che hanno un'influenza positiva sulla classe di interesse sono colorate in verde mentre le feature negativamente influenti sono colorate in rosso.*

Dall'osservazione della figura 8.5 si può notare come per questa immagine l'influenza negativa sulla classe 'mouse' provenga dalle feature che identificano il disegno a quadri della superficie sulla quale è poggiato il mouse. Le feature che compongono il mouse invece contribuiscono positivamente alla predizione corretta della classe 'mouse'. È possibile quindi registrare il comportamento errato del modello che in questo caso si affida maggiormente a feature marginali come lo sfondo che non all'oggetto al centro dell'immagine.

Osservando la tabella 8.6 si possono ottenere delle infomazioni più dettagliate sulle feature più rilevanti per quanto riguarda l'impatto sulla classe predetta 'mouse'. Le feature A, B e C sono le tre feature caratterizzate dalla maggiore influenza positiva sulla classe. In particolare, la feature A presenta il valore dell'indice nIRI pari a circa 0,18 e il valore di IRP pari a 2,96 a indicare la tendenza della feature a rappresentare univocamente la classe 'mouse'. Le feature D, E e F presentano invece valori del nIRI negativi, sono quindi negativamente influenti sulla classe in esame.

Effettuando l'analisi delle feature sulle prime 5 classi erroneamente individuate dal modello e ricavando le relative elaborazioni grafiche qualitative (figura 8.7) è possibile notare un'influenza generalmente positiva da parte delle feature che identificano il disegno a quadri della superficie d'appoggio sulle rispettive classificazioni (con la parziale eccezione della classe 'toilet seat'). Con solo un'osservazione prettamente qualitativa risulta quindi possibile attribuire proprio a queste feature le principali motivazioni dell'errata classificazione. Considerando che le classi predette risultano essere articoli legati all'igiene personale si può dedurre che il modello scambi lo sfondo dell'immagine per delle piastrelle da bagno venendone conseguentemente fuorviato.

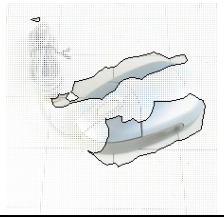
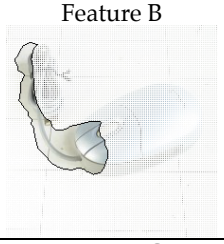
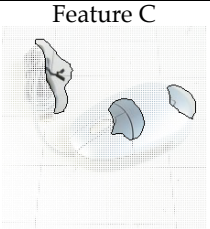
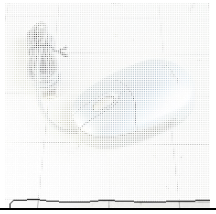
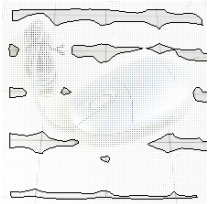
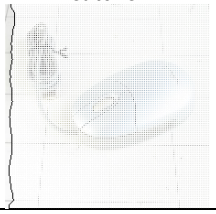
Classe: mouse, computer mouse	Prediction	DI	IR	nIRI	IRP
	5.102232				
Feature più influenti positivamente					
Feature A 	0.999437	4.102795	5.10511	**0.178637**	2.959124
Feature B 	1.772029	3.330202	2.87932	**0.097027**	2.004675
Feature C 	2.111192	2.99104	2.41675	**0.078054**	1.839328
Feature più influenti negativamente					
Feature D 	11.48096	-6.37873	0.444408	**-0.14667**	0.398884
Feature E 	7.494534	-2.3923	0.680794	**-0.04891**	0.361174
Feature F 	6.896386	-1.79415	0.739841	**-0.03617**	0.72312

Tabella 8.6: *Influenze principali sulla classe di interesse 'mouse'*

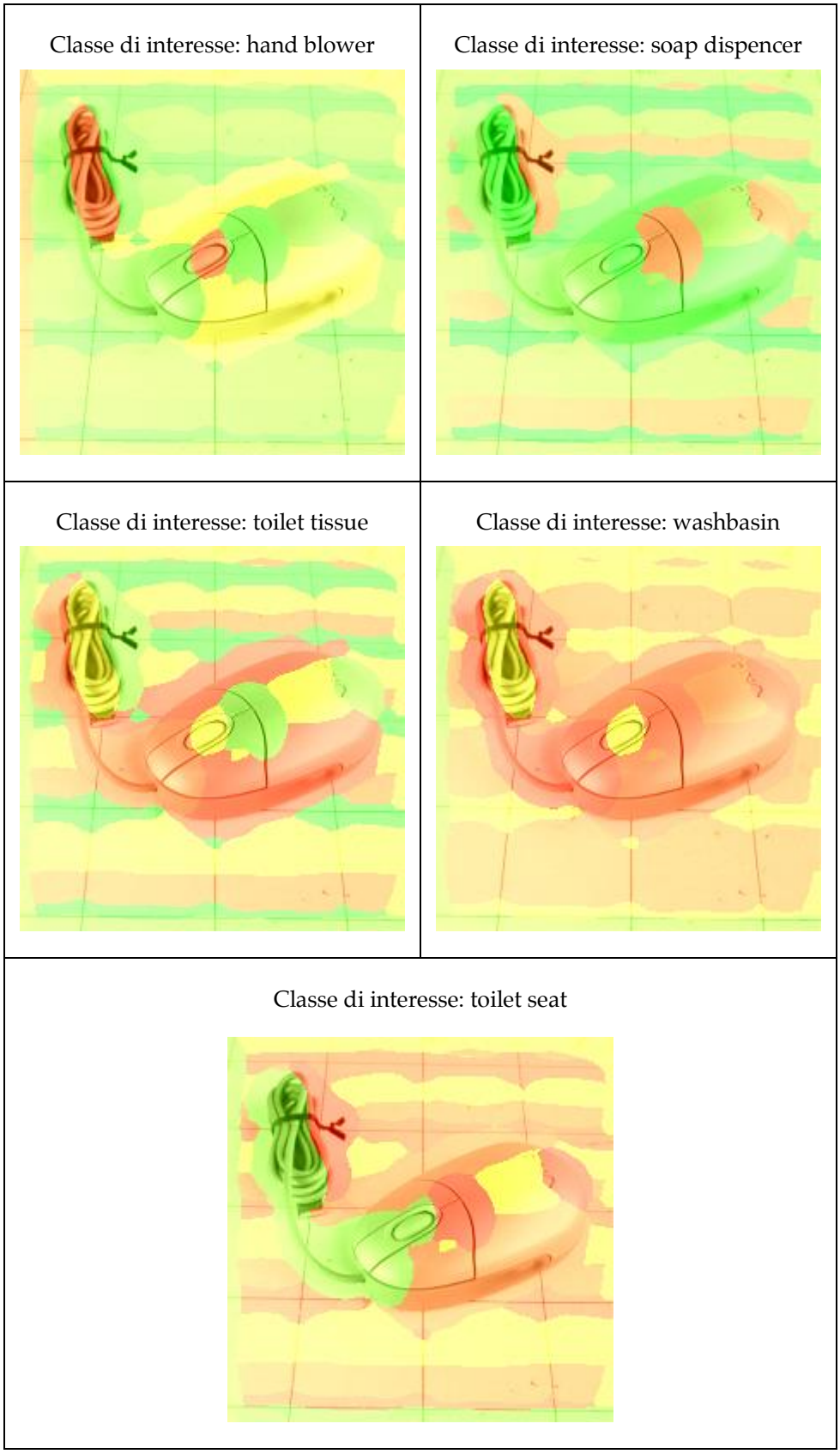


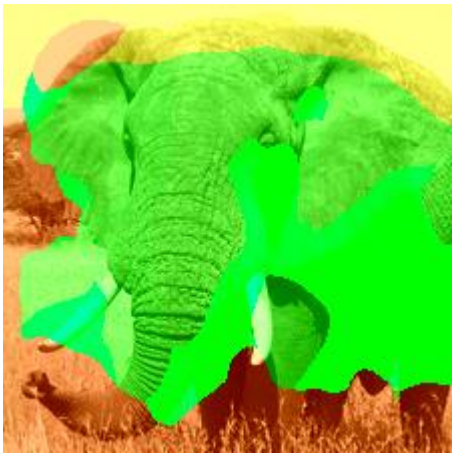
Figura 8.7: *Elaborazioni grafiche per le prime 5 classi predette dal classificatore*

8.3.2 Elefante africano

Primi 5 risultati:
African_elephant: 0.60234541
Bighorn: 0.17285174
Tusker: 0.10890929
Ram: 0.07381846
Indian_elephant: 0.016448317

La classe di questa immagine viene individuata correttamente dal classificatore e anche le altre classi che seguono appaiono come delle previsioni sensate; sono infatti quasi tutte specie diverse di elefanti. In questo esempio si è voluto indagare il perché della predizione della classe 'African elephant' (60%) a discapito della classe 'Indian elephant' (1%).

Figura 8.8: *Segmentazione dell'immagine in 10 feature interpretabili*


(a)

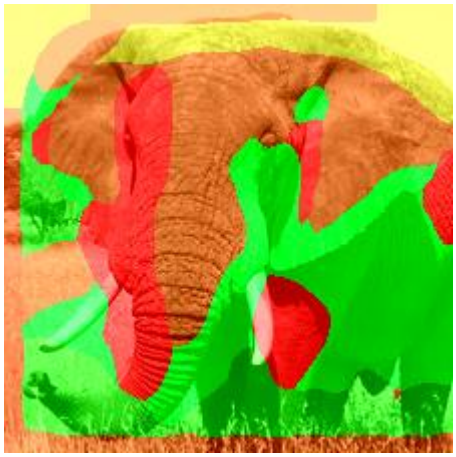

(b)

Figura 8.9: *Rappresentazioni grafiche qualitative dell'influenza delle feature sulle classi di interesse 'African elephant' (a) e 'Indian elephant' (b).*

Si noti come nella figura 8.9 le feature della testa, delle orecchie e di una delle due zanne risultino avere un impatto positivo sulla classe 'African elephant' e negativo su 'Indian elephant'. Ciò indica un corretto funzionamento da parte del modello in quanto l'elefante indiano è caratterizzato da una diversa forma della testa, orecchie più piccole e assenza di zanne negli esemplari femminili.

Dall'analisi delle tabelle 8.10 e 8.11 è possibile ottenere una stima più dettagliata dell'impatto delle feature che racchiudono testa, orecchie e zanne sulle due classi in esame. Per le due feature in esame A e B si hanno valori di nIRI moderatamente alti per la classe 'African elephant' e, al contrario, negativi per la classe 'Indian elephant'. Le due feature in esame, malgrado valori significativi dell'indice nIRI, non risultano essere univocamente caratterizzanti per la classe 'African elephant' in quanto presentano valori di IRP inferiori o prossimi a 1.

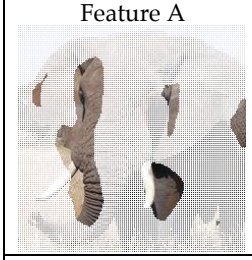
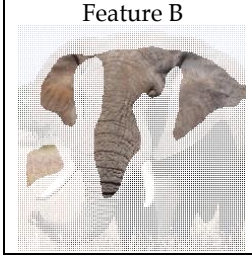
Classe: African elephant	Prediction	DI	IR	nIRI	IRP
	60.23454				
Feature A 	23.74696	36.48758	2.53652	**0.516759**	1.046909
Feature B 	26.14151	34.09303	2.30417	**0.48281**	0.448241

Tabella 8.10: *Influenza delle feature A e B sulla classe di interesse 'African elephant'*

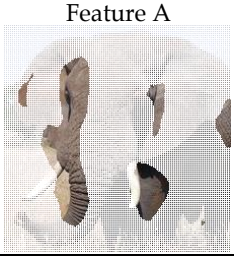
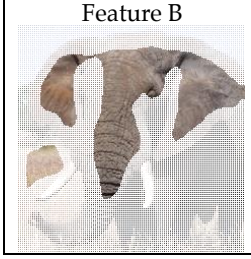
Classe: Indian elephant	Prediction	DI	IR	nIRI	IRP
	1.644832				
Feature A 	4.648682	-3.00385	0.353828	**-0.0872**	0.146037
Feature B 	2.162493	-0.51766	0.760618	**-0.01063**	0.147967

Tabella 8.11: *Influenza delle feature A e B sulla classe di interesse 'Indian elephant'*

8.3.3 Pizza

Primi 5 risultati:
pizza: 0.25180733
frying_pan: 0.079138257
sombrero: 0.052266803
toilet_seat: 0.035252444
wooden_spoon: 0.031031754

In questa immagine la classe 'pizza' viene correttamente identificata. Si è comunque condotta un'analisi di trasparenza sulla classe predetta per valutare l'influenza delle singole feature.

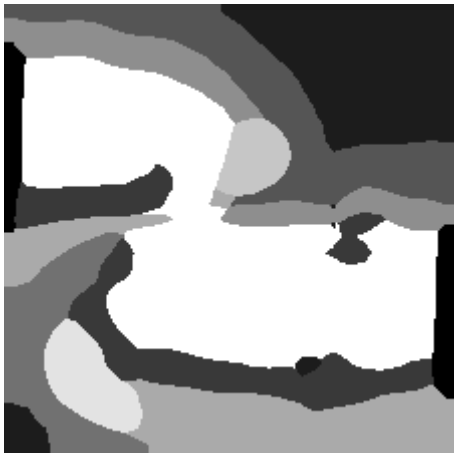


Figura 8.12: *segmentazione dell'immagine in 10 feature interpretabili*

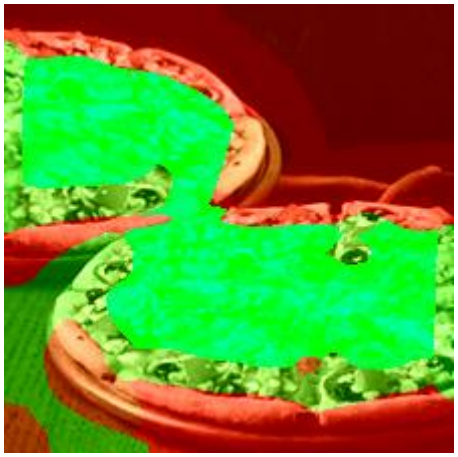


Figura 8.13: *report grafico qualitativo - classe di interesse: pizza*

Dal report grafico in figura 8.13 e dalla tabella 8.14 si osserva la netta influenza positiva della feature A, contenente la parte centrale delle due pizze. Osservando il valore dell'indice nIRI relativo alla feature A (0,98) si deduce che la feature in questione è assolutamente determinante nella predizione 'pizza' operata dal classificatore. Molto meno influenti, ma comunque non trascurabili, le feature B e C. In particolare, si noti come la feature C, che comprende la porzione dell'immagine occupata da una superficie in legno e che quindi ci si aspetterebbe come estranea alla classe 'pizza', abbia una significativa influenza positiva con valore di nIRI pari a 0,19. A giudicare dai risultati si può quindi ipotizzare che il classificatore sia particolarmente abile a riconoscere la classe 'pizza' a partire dalle feature che ne individuano il condimento.

Un altro aspetto da considerare è l'influenza negativa dei bordi della pizza (feature D ed E) che presentano dei valori dell'indice nIRI particolarmente negativi pari a -0,42 e -0,41.

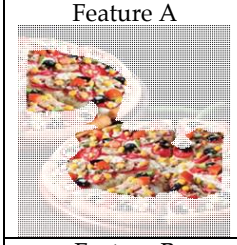
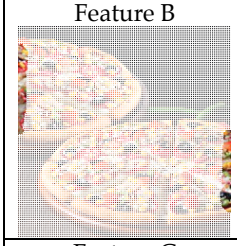
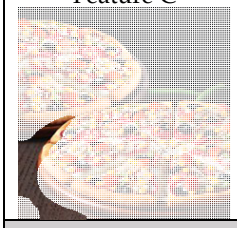
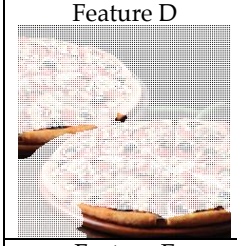
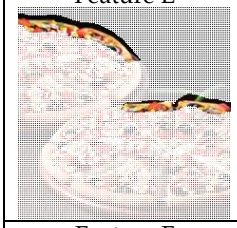
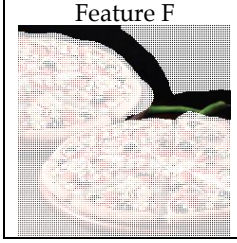
pizza, pizza pie	Prediction	DI	IR	nIRI	IRP
	25.18073				
Feature più influenti positivamente					
Feature A 	0.141403	25.03933	178.078	**0.978066**	1.562815398
Feature B 	13.13241	12.04832	1.91745	**0.227116**	1.437266829
Feature C 	14.67279	10.50794	1.71615	**0.194563**	1.356120624
Feature più influenti negativamente					
Feature D 	53.13112	-27.9504	0.473936	**-0.41935**	0.203286004
Feature E 	52.63395	-27.4532	0.478412	**-0.41355**	0.209505669
Feature F 	44.40693	-19.2262	0.567045	**-0.30943**	0.18340549

Tabella 8.14: *Influenze principali sulla classe di interesse 'pizza'*

8.3.4 Limone

Primi 5 risultati:
lemon: 0.15553676
bow_tie: 0.098650157
bathing_cap: 0.091984682
cowboy_hat: 0.089851722
mask: 0.084251054

In questo esempio volutamente ingannevole, il classificatore identifica correttamente la classe 'lemon'. Per evidenziare il il modo in cui il modello reagisce alle informazioni contrastanti fornite dall'immagine si è condotta l'analisi di trasparenza per classi 'lemon' e 'mask'.

Figura 8.15: *Segmentazione dell'immagine in 10 feature interpretabili*

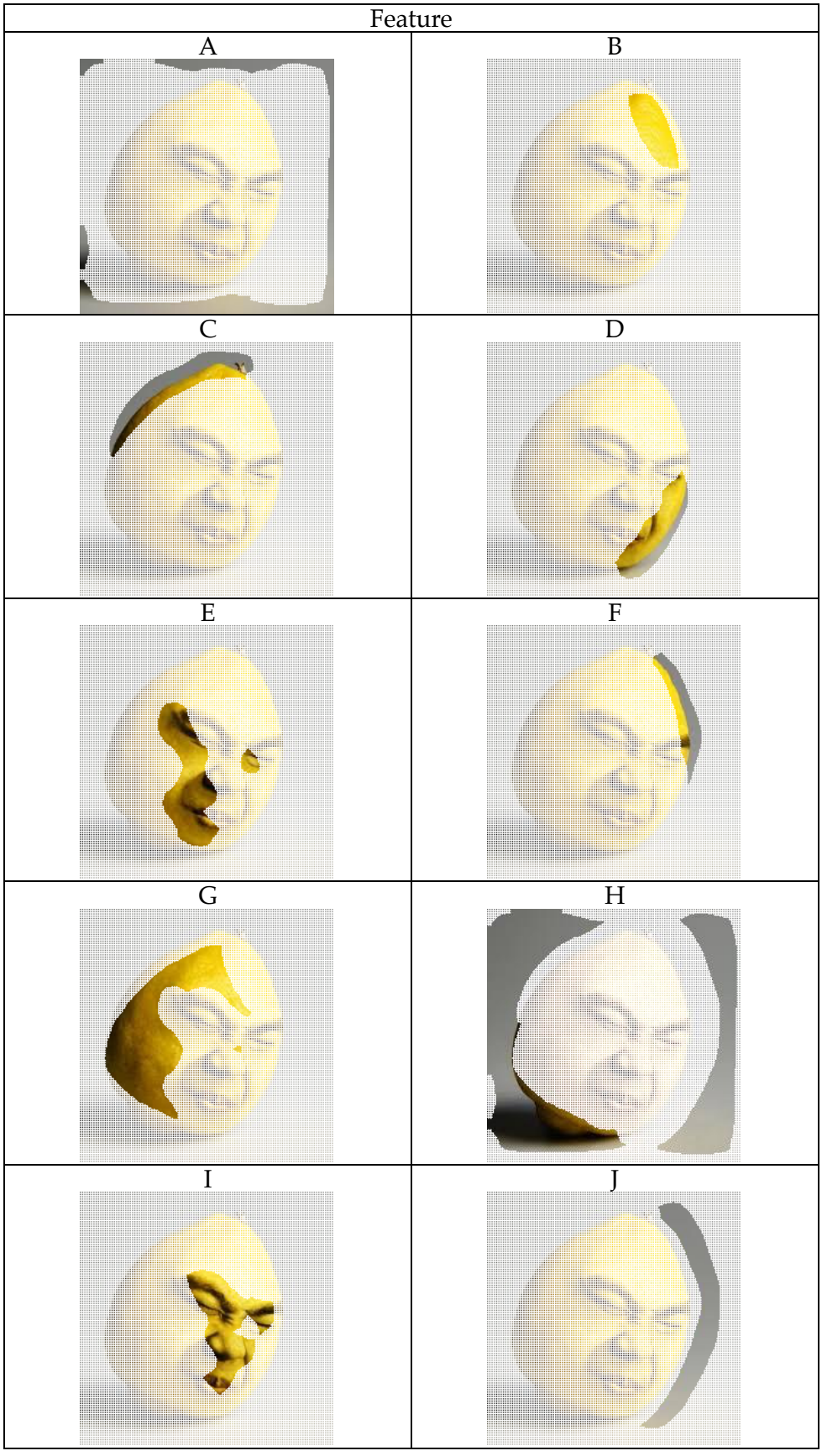


Figura 8.16: *Feature estratte dell'immagine*

Figura 8.17: *Report grafico qualitativo - classe di interesse: lemon*

Dal report grafico in figura 8.17 e dalla tabella 8.19 si osserva che il classificatore individua la classe 'lemon' prevalentemente grazie all'influenza delle feature G e H. La feature G in particolare, che racchiude un'ampia area della buccia, si dimostra essere molto influente e anche univocamente caratterizzante per la classe di interesse con valori dell'indice nIRI pari a 0,43 e IRP pari a 2,96. Si può quindi ipotizzare che il classificatore sia particolarmente sensibile e abile nell'individuazione della caratteristica superficie della scorza di limone. Le feature che invece influenzano negativamente la classe 'lemon' invece sono, come atteso, quelle relative al volto ovvero le feature I ed E con valori di nIRI pari a -0,66 e -0.42. Se l'influenza negativa delle feature I ed E era in qualche modo prevedibile molto meno lo è quella della feature C che delimita un'area dell'immagine comprendente la sommità del limone. La feature C presenta l'indice nIRI pari a -0,32.

Figura 8.18: *Report grafico qualitativo - classe di interesse: mask*

Analizzando la classe di interesse 'mask' si hanno risultati sostanzialmente speculari a quelli relativi alla classe 'lemon'. Le classi H, G ed A che avevano influenza positiva sulla classe 'lemon' presentano invece influenze leggermente negative nel caso della classe 'mask'. Le feature C, I ed E invece adesso influenzano positivamente la classe

'mask'. In questo esperimento si è quindi registrato un comportamento sostanzialmente corretto da parte del classificatore che, in presenza di informazioni intenzionalmente contrastanti e fuorvianti, è riuscito a restituire un risultato coerente soppesando in maniera adeguata le influenze provenienti dalle feature.

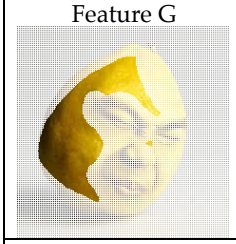
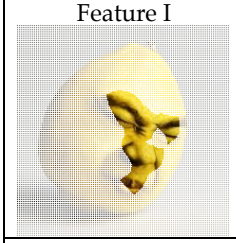
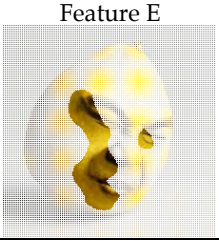
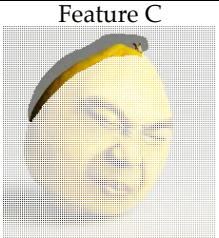
Classe: lemon	Prediction	DI	IR	nIRI	IRP
	15.55368				
Feature più influenti positivamente					
Feature G 	2.745513	12.80816	5.66513	**0.427985**	2.966326
Feature H 	7.576373	7.977302	2.05292	**0.168486**	1.500378
Feature più influenti negativamente					
Feature I 	61.54772	-45.994	0.252709	**-0.65943**	0.009141
Feature E 	40.45008	-24.8964	0.384515	**-0.42634**	0.12113
Feature C 	33.75359	-18.1999	0.460801	**-0.32379**	0.274651

Tabella 8.19: *Influenze principali sulla classe di interesse 'lemon'*

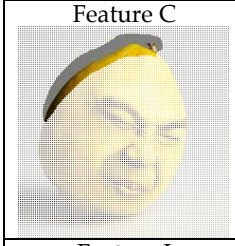
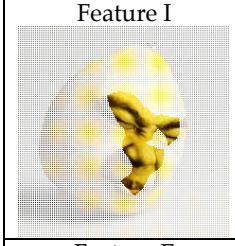
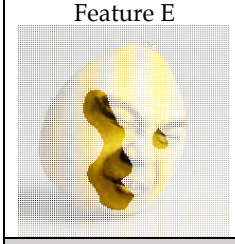
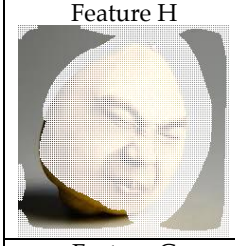
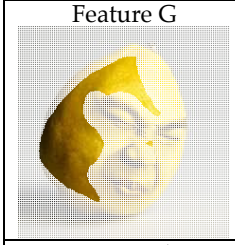
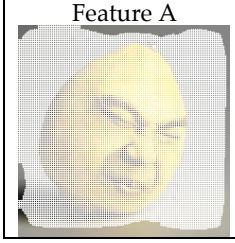
Classe: mask	Prediction	DI	IR	nIRI	IRP
	8.425105				
Feature più influenti positivamente					
Feature C 	2.161809	6.263296	3.89725	**0.206455**	2.322877
Feature I 	2.410627	6.014478	3.49498	**0.185279**	0.126414
Feature E 	2.711227	5.713878	3.10749	**0.163842**	0.978919
Feature più influenti negativamente					
Feature H 	10.31947	-1.89437	0.816428	**-0.03723**	0.596687
Feature G 	9.6787	-1.25359	0.870479	**-0.02469**	0.455793
Feature A 	9.395838	-0.97073	0.896685	**-0.01916**	0.870573

Tabella 8.20: *Influenze principali sulla classe di interesse 'mask'*

8.3.5 Jack o' Lantern

Primi 5 risultati:
jack-o'-lantern: 0.46362135
bakery: 0.086345926
candle: 0.049824145
toilet_seat: 0.020093719
tractor: 0.017202629

In questo caso il classificatore identifica correttamente la classe 'jack o lantern'. Nell'analisi di trasparenza si analizzerà l'influenza delle feature sulle prime due classi individuate dal modello: 'jack o lantern' e 'bakery'.

Figura 8.21: *Segmentazione dell'immagine in 10 feature interpretabili*

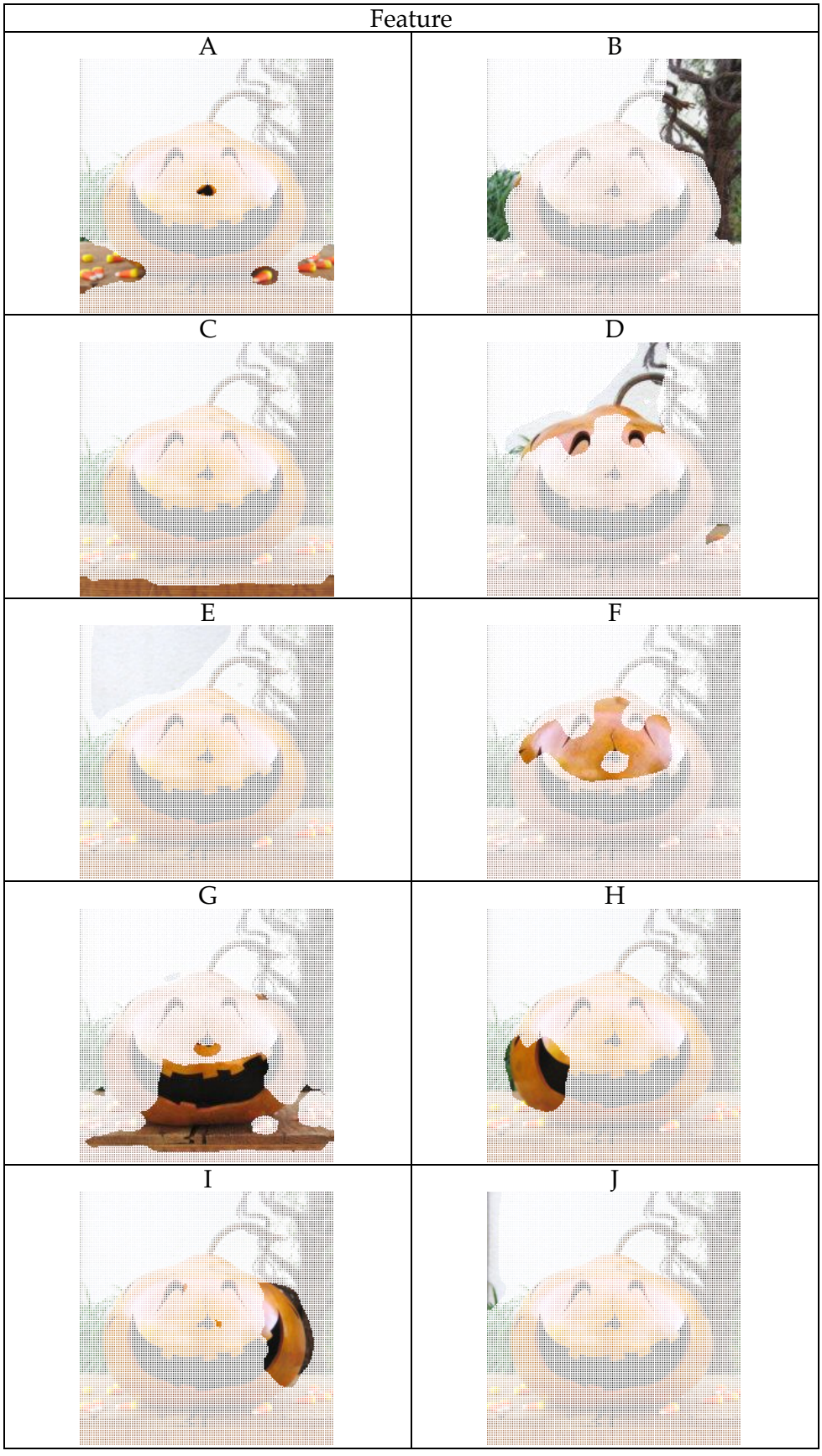


Figura 8.22: *Feature estratte dell'immagine*

Figura 8.23: *Report grafico qualitativo - classe di interesse: jack o' lantern*

Come si può osservare dal report grafico in figura 8.23 e dalla tabella 8.25 si sono individuate due feature (D e F) che presentano un impatto estremamente positivo sulla predizione della classe 'jack o lantern' con valori di nIRI pari a 0,88 e 0,80. Risulta interessante porre attenzione su come la feature B, che ricopre un'area relativa allo sfondo dell'immagine, abbia anche essa un'influenza positiva non indifferente sulla classe (nIRI = 0,22). Si sono individuate inoltre due feature che presentano un impatto leggermente negativo sulla classe in esame. Si tratta delle feature C e A relative alla porzione dell'immagine occupata dalla superficie sulla quale è poggiata la zucca.

Figura 8.24: *Report grafico qualitativo - classe di interesse: bakery*

Eseguendo l'analisi di trasparenza sulla classe 'bakery' si ottengono i risultati esposti nella figura 8.24 e nella tabella 8.26. Le feature G e A con valori di nIRI rispettivamente di 0,23 e 0,21 e IRP di 2,06 e 2,37 sono positivamente influenti e univocamente caratterizzanti per la classe in esame. La porzione di immagine coperta da queste due feature comprende la superficie di appoggio con la parte inferiore della zucca e alcune caramelle.

Le feature dall'impatto negativo sulla classe 'bakery' sono invece I, H e D.

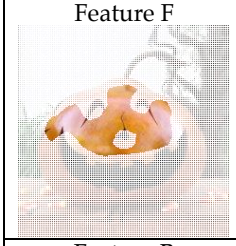
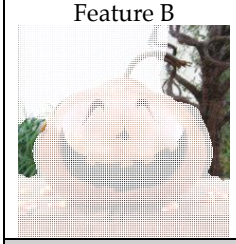
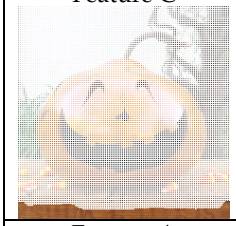
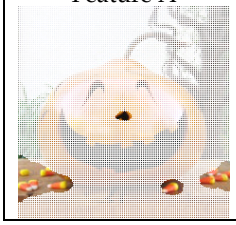
Classe: jack-o'-lantern	Prediction	DI	IR	nIRI	IRP
	46.36213				
Classi più influenti positivamente					
Feature D 	2.81339	43.54874	16.4791	**0.878091**	1.981226
Feature F 	4.859868	41.50227	9.53979	**0.800109**	1.773123
Feature B 	32.96434	13.3978	1.40643	**0.220997**	1.076831
Classi più influenti negativamente					
Feature C 	50.06802	-3.70589	0.925983	**-0.06919**	0.845685
Feature A 	48.10982	-1.74769	0.963673	**-0.0338**	0.564393

Tabella 8.25: *Influenze principali sulla classe di interesse 'jack o' lantern'*

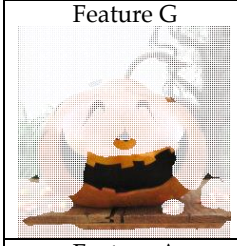
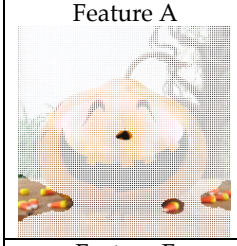
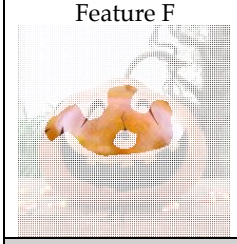
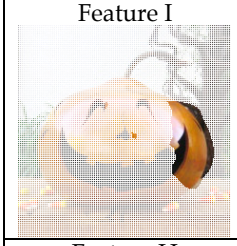
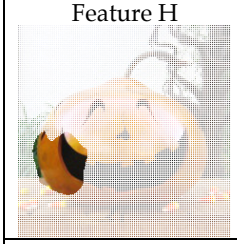
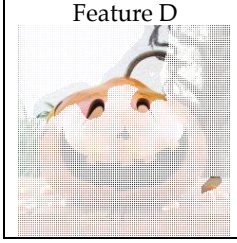
Classe: bakery	Prediction	DI	IR	nIRI	IRP
	8.634593				
Feature più influenti positivamente					
Feature G 	2.004083	6.63051	4.3085	**0.231399**	2.05576
Feature A 	2.133325	6.501268	4.04748	**0.218261**	2.370483
Feature F 	5.877264	2.757329	1.46915	**0.05596**	0.273065
Feature più influenti negativamente					
Feature I 	14.34479	-5.7102	0.601932	**-0.11445**	0.493748
Feature H 	13.51209	-4.87749	0.639027	**-0.09706**	0.511105
Feature D 	12.13863	-3.50404	0.711332	**-0.06906**	0.085521

Tabella 8.26: *Influenze principali sulla classe di interesse 'bakery'*

8.3.6 Medusa

Primi 5 risultati:
jellyfish: 0.32341242
sea_slug: 0.24017769
sea_anemone: 0.075401768
lionfish: 0.059361756
cardoon: 0.051633615

In questo caso il modello individua correttamente la classe 'jellyfish' con un punteggio del 32%. La classe 'jellyfish' è quindi la classe di interesse nell'analisi di trasparenza svolta. In aggiunta si sono presentano delle elaborazioni grafiche relative alle classi 'sea slug' e 'sea anemone'.

Figura 8.27: *Segmentazione dell'immagine in 10 feature interpretabili*

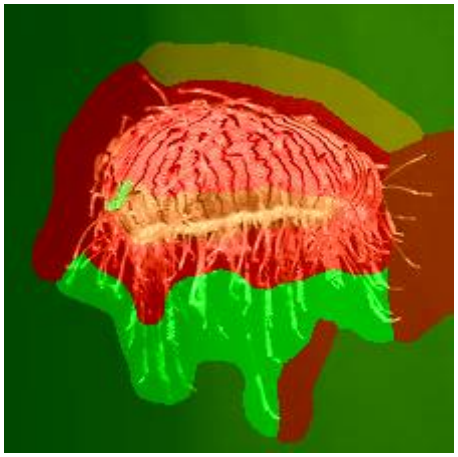
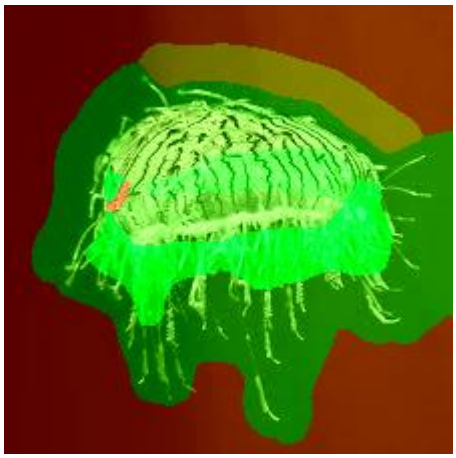


Figura 8.28: *Report grafico qualitativo - classe di interesse: jellyfish*

La feature che contribuisce in maniera determinante alla corretta predizione è la feature A relativa ai tentacoli della medusa che presenta un nIRI di 0,48 e un IRP di 1,88. Si distingue per la sua influenza positiva anche la feature B che copre buona gran parte dello sfondo dell'immagine e ha un indice nIRI di 0,11.

Le feature F, E, e D sono invece caratterizzate da un impatto estremamente negativo sulla classe 'jellyfish'. Come osservabile dalle rappresentazioni grafiche in figura 8.29 infatti queste feature conducono il risultato del modello verso altre classi come 'sea slug' o 'sea anemone'.

(a)



(b)

Figura 8.29: *Rappresentazioni grafiche qualitative dell'influenza delle feature sulle classi di interesse 'sea slug' (a) e 'sea anemone' (b).*

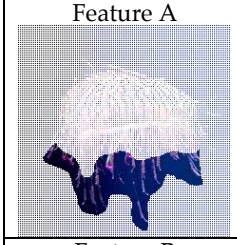
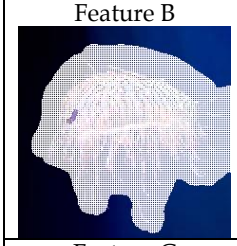
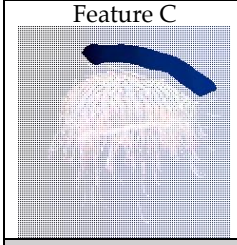
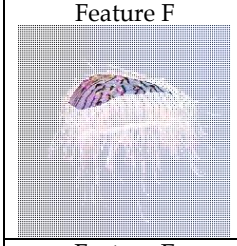
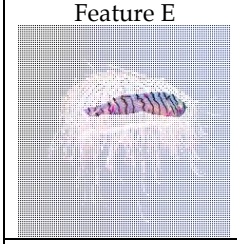
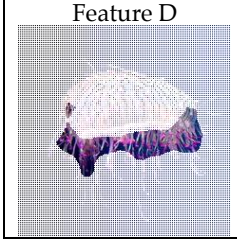
Classe: jellyfish	Prediction	DI	IR	nIRI	IRP
	32.34124				
Feature più influenti positivamente					
Feature A 	8.784293	23.55695	3.68171	**0.482209**	1.883774
Feature B 	25.90675	6.434494	1.24837	**0.116506**	1.178451
Feature C 	31.51028	0.830963	1.02637	**0.016353**	1.023195
Feature più influenti negativamente					
Feature F 	78.64544	-46.3042	0.411228	**-0.5683**	0.050326
Feature E 	77.41876	-45.0775	0.417744	**-0.55896**	0.157443
Feature D 	71.31584	-38.9746	0.453493	**-0.50888**	0.15251

Tabella 8.30: *Influenze principali sulla classe di interesse 'jellyfish'*

8.3.7 Segnale stradale

Primi 5 risultati:
boathouse: 0.2066157
street_sign: 0.12135011
alp: 0.069113672
lakeside: 0.059708256
birdhouse: 0.057770401

In questo esempio il classificatore individua la classe dell'immagine 'street sign', ma solo al secondo posto con una percentuale del 12,1%. Il primo risultato è infatti 'boathouse' con un punteggio superiore al 20%. L'analisi di trasparenza svolta ha preso in esame le prime due classi predette dal modello 'boathouse' e 'street sign' così da poter motivare le ragioni di questa inesattezza nei risultati.

Figura 8.31: *Segmentazione dell'immagine in 10 feature interpretabili*

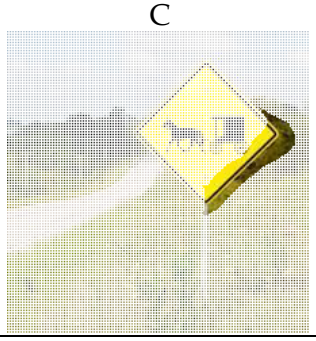
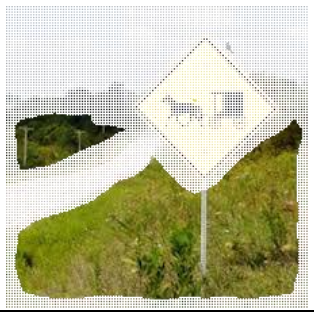
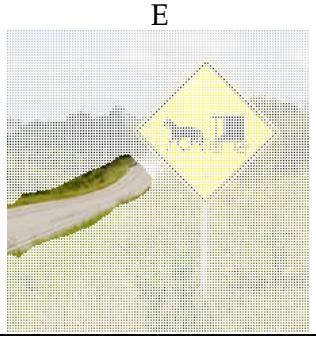
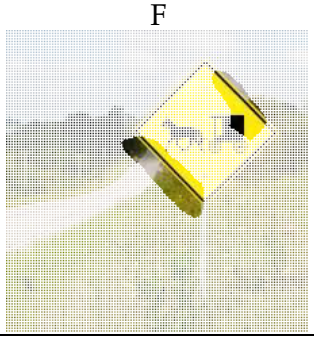
Feature	
A 	B 
C 	D 
E 	F 
G 	H 
I 	J 

Tabella 8.32: *Feature estratte dell'immagine*

Figura 8.33: *Report grafico qualitativo - classe di interesse: boathouse*

Dall'osservazione del report grafico in figura 8.33 e dalla tabella delle influenze in figura 8.35 è possibile individuare i segmenti che contribuiscono alla predizione indesiderata della classe 'boathouse' da parte del modello. Le tre feature (E, F, D) che si contraddistinguono per l'alta influenza sulla predizione della classe al tratto di strada presente nell'immagine che presenta un valore molto alto dell'indice nIRI pari a circa 0,71 e tende anche ad essere univocamente caratterizzante per la classe in quanto presenta un valore di IRP pari a 2,07. Le feature che presentano un impatto negativo sulla classe sono invece G e J (nIRI rispettivamente di -0,19 e -0,04) che ricoprono l'area dell'immagine occupata dal segnale stradale.

Figura 8.34: *Report grafico qualitativo - classe di interesse: street sign*

I risultati relativi alla classe desiderata 'street sign' risultano essere sostanzialmente speculari a quelli della classe 'boathouse'. Le feature G, C e J come previsto presentano un impatto significativamente positivo univocamente sulla classe di interesse. La feature E invece è in maniera inequivocabile il segmento dell'immagine che causa l'errata predizione da parte del classificatore. Dalla sua perturbazione si ricava infatti un indice nIRI nettamente negativo pari a -0,73.

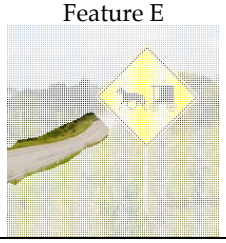
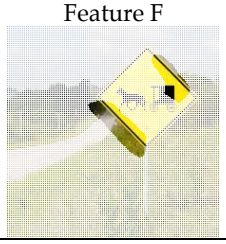
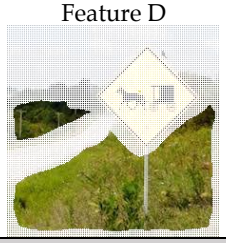
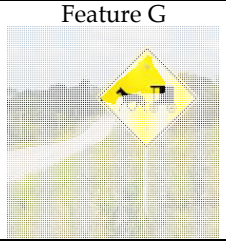
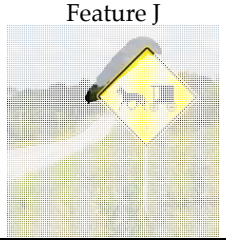
Classe: boathouse	Prediction	DI	IR	nIRI	IRP
	20.66157				
Feature più influenti positivamente					
Feature E 	1.651817	19.00975	12.5084	**0.705277**	2.074602
Feature F 	3.029595	17.63198	6.81991	**0.551235**	1.156676
Feature D 	3.854244	16.80733	5.36073	**0.482495**	2.072487
Feature più influenti negativamente					
Feature G 	31.33114	-10.6696	0.659458	**-0.18841**	0.416943
Feature J 	22.46984	-1.80827	0.919525	**-0.03502**	0.762566

Tabella 8.35: *Influenze principali sulla classe di interesse 'boathouse'*

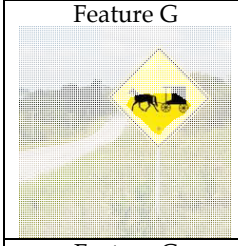
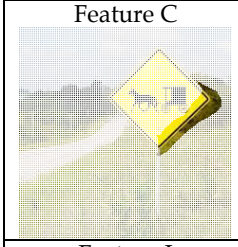
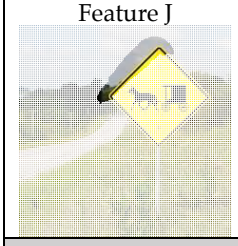
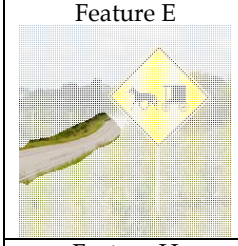
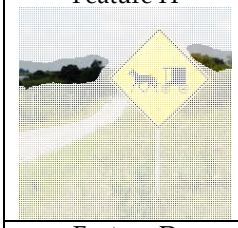
Classe: street sign	Prediction	DI	IR	nIRI	IRP
	12.13501				
Feature più influenti positivamente					
Feature G 	2.685032	9.449979	4.5195	**0.309393**	2.186905
Feature C 	3.754026	8.380985	3.23253	**0.228898**	2.055227
Feature J 	5.301936	6.833075	2.28879	**0.157007**	1.898102
Feature più influenti negativamente					
Feature E 	61.86723	-49.7322	0.196146	**-0.72475**	0.032532
Feature H 	15.88981	-3.7548	0.763698	**-0.07222**	0.557496
Feature D 	14.1476	-2.01259	0.857743	**-0.03913**	0.331608

Tabella 8.36: *Influenze principali sulla classe di interesse 'street sign'*

8.3.8 Kimono

Primi 5 risultati:
vase: 0.12474733
lampshade: 0.080761641
candle: 0.07888414
birdhouse: 0.073079087
kimono: 0.072604351

In questo esempio la classe desiderata 'kimono' è solo il quinto risultato restituito dal classificatore black-box con una percentuale di solo il 7% circa. L'analisi di trasparenza si è quindi concentrata sulla classe 'kimono' e su 'vase' che risulta essere la prima predizione del classificatore.

Figura 8.37: *Segmentazione dell'immagine in 10 feature interpretabili*

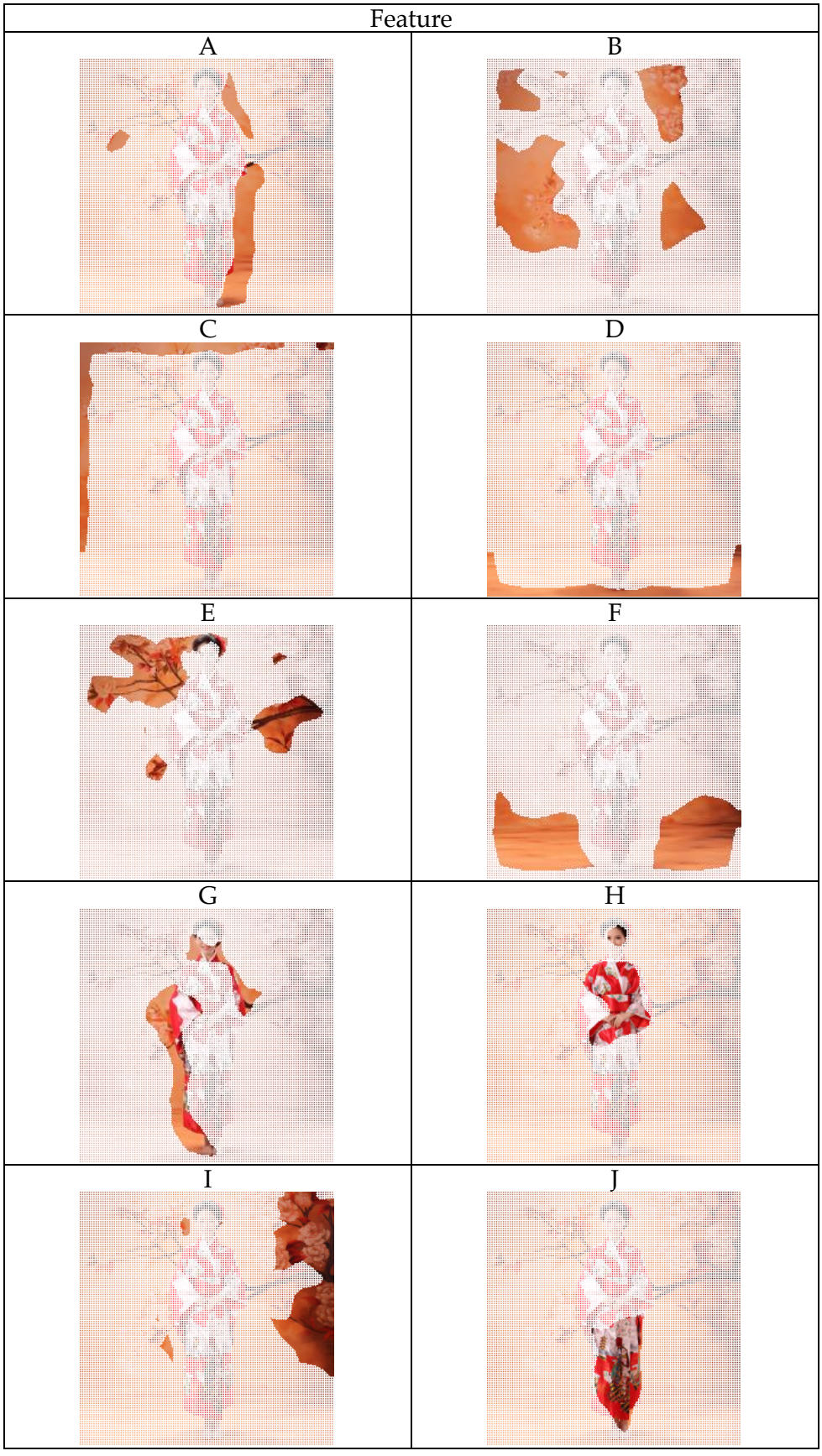


Figura 8.38: *Feature estratte dell'immagine*

Figura 8.39: *Report grafico qualitativo - classe di interesse: vase*

Per la classe di interesse 'vase' si individuano le influenze positive da parte delle feature E e F con valori di nIRI rispettivamente di 0,47 e 0,23. La feature E racchiude un'area dell'immagine con dei motivi floreali, risulta quindi assolutamente chiaro il motivo della influenza positiva sulla classe 'vase'. La feature J presenta invece un impatto fortemente negativo sulla classe in esame con in indice nIRI pari a circa -0,33.

Figura 8.40: *Report grafico qualitativo - classe di interesse: kimono*

Ripetendo l'analisi di trasparenza con la classe di interesse 'kimono' e osservando i risultati prodotti nel report grafico in figura 8.40 e nella tabella 8.42 si può valutare l'impatto delle feature sulla predizione. Spicca l'influenza positiva della classe J che per la classe 'vase' presentava un'influenza negativa e qui invece risulta la feature più positiva con un valore di nIRI pari a 0,47. La feature J risulta essere univocamente caratterizzante per la classe in modo netto con un IRP significativamente alto pari a circa 4.

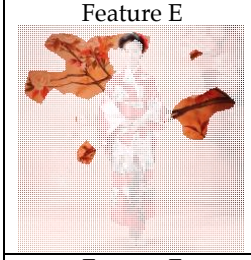
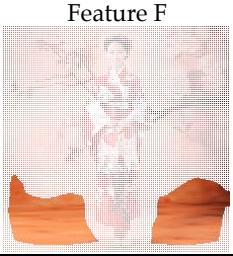
Classe: vase	Prediction	DI	IR	nIRI	IRP
	12.47473				
Feature più influenti positivamente					
Feature E 	1.546317	10.92842	8.06738	**0.472347**	2.278489
Feature F 	3.949062	8.525671	3.15891	**0.228578**	2.213477
Feature più influenti negativamente					
Feature J 	29.89348	-17.4187	0.417306	**-0.3289**	0.123472

Tabella 8.41: *Influenze principali sulla classe di interesse 'vase'*

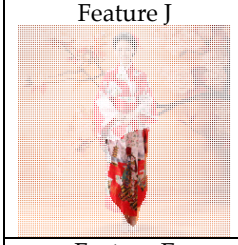
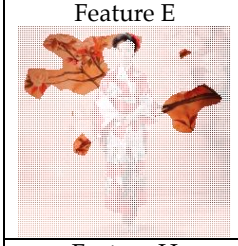
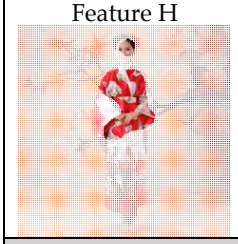
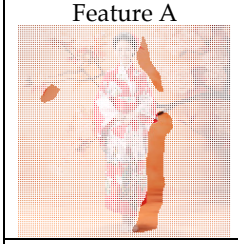
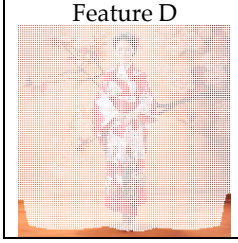
Classe: kimono	Prediction	DI	IR	nIRI	IRP
	7.260435				
Feature più influenti positivamente					
Feature J 	0.550985	6.70945	13.1772	**0.470678**	3.898849
Feature E 	2.306836	4.953599	3.14736	**0.1465**	0.888915
Feature H 	2.322692	4.937743	3.12587	**0.145404**	1.597405
Feature più influenti negativamente					
Feature I 	13.35907	-6.09864	0.543483	**-0.12691**	0.35036
Feature A 	11.54297	-4.28253	0.628992	**-0.08678**	0.491328
Feature D 	8.268887	-1.00845	0.878043	**-0.01993**	0.862955

Tabella 8.42: *Influenze principali sulla classe di interesse 'kimono'*

8.3.9 Bicchiere di birra

Primi 5 risultati:
beer_bottle: 0.6838693
pill_bottle: 0.069878176
beer_glass: 0.045961853
wine_bottle: 0.037885152
pop_bottle: 0.033832375

In questa immagine la classe corretta è presente tra le prime 5 ma viene dopo le classi 'beer bottle' e 'pill bottle' e con un puntaggio pari solo al 6%. L'analisi di trasparenza in questo caso si focalizzerà quindi sulle prime tre classi individuate così da capire le motivazioni dei risultati.

Figura 8.43: *Segmentazione dell'immagine in 10 feature interpretabili*

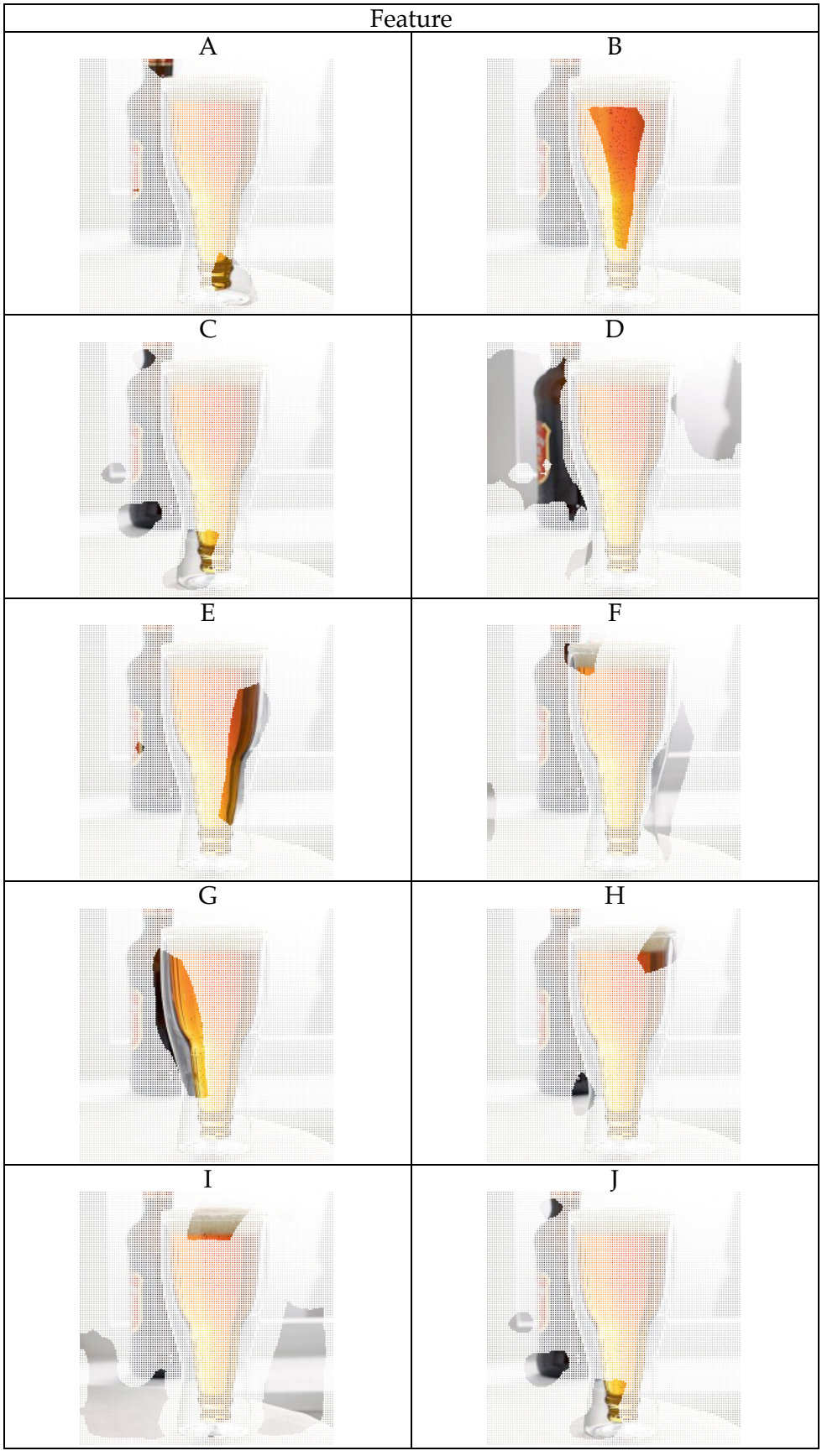


Figura 8.44: *Feature estratte dell'immagine*

Figura 8.45: *report grafico qualitativo - classe di interesse: beer bottle*

In questo caso il classificatore è parzialmente tratto in inganno dalla insolita forma del bicchiere, il cui fondo richiama il collo di una bottiglia capovolta. Non a caso, come mostrato nella tabella 8.46 relativa alla classe di interesse 'beer bottle', tra le feature con la maggiore influenza positiva se ne hanno due (A e C) che racchiudono proprio il particolare del fondo del bicchiere. Queste due feature sono entrambe corredate da valori significativi dell'indice nIRI (0,34 e 0,36). Tra le feature con un'influenza negativa sulla classe 'beer bottle' si individua la feature D (nIRI = -0.06) relativa alla bottiglia sullo sfondo dell'immagine.

Figura 8.46: *Report grafico qualitativo – classe di interesse: beer glass*

In questo caso, come atteso, nella figura 8.46 il fondo del bicchiere presenta una colorazione tendente al rosso a indicare una leggera influenza negativa sulla classe di interesse 'beer glass' da parte delle feature A e C. Le feature che invece influenzano positivamente la classe di interesse sono F, B e G (tabella 8.48).

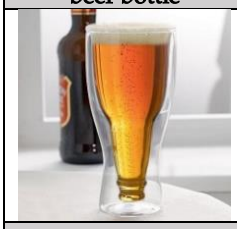
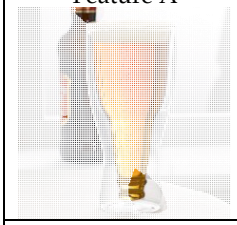
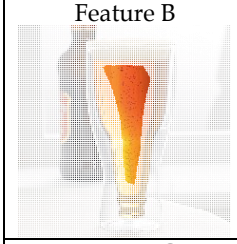
Classe: beer bottle	Prediction	DI	IR	nIRI	IRP
	68.38693				
Feature più influenti positivamente					
Feature A 	44.3387	24.04823	1.54238	**0.345049**	1.194653
Feature B 	46.57094	21.81599	1.46845	**0.319228**	1.170846
Feature C 	47.78249	20.60444	1.43121	**0.305005**	1.16109
Feature più influenti negativamente					
Feature D 	74.57187	-6.18494	0.917061	**-0.11045**	0.863047
Feature E 	69.30945	-0.92252	0.98669	**-0.01812**	0.931206
Feature F 	68.68246	-0.29553	0.995697	**-0.00588**	0.958943

Tabella 8.47: *Influenze principali sulla classe di interesse 'beer bottle'*

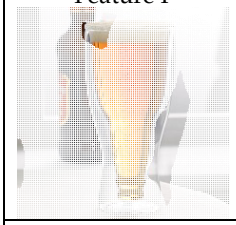
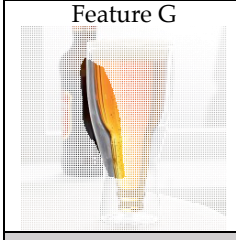
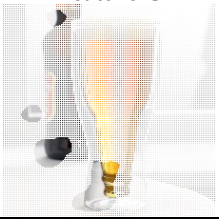
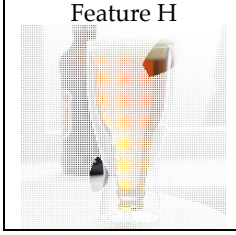
Classe: beer glass	Prediction	DI	IR	nIRI	IRP
	4.596185				
Feature più influenti positivamente					
Feature F 	2.40087	2.195315	1.91438	**0.050778**	1.843718
Feature B 	2.758864	1.837321	1.66597	**0.039973**	1.328339
Feature G 	3.422211	1.173974	1.34305	**0.023922**	1.26969
Feature più influenti negativamente					
Feature C 	6.751858	-2.15567	0.680729	**-0.04429**	0.55225
Feature A 	6.707486	-2.1113	0.685232	**-0.04332**	0.53075
Feature H 	5.206463	-0.61028	0.882784	**-0.01215**	0.846996

Tabella 8.48: *Influenze principali sulla classe di interesse 'beer glass'*

Figura 8.49: *Report grafico qualitativo – classe di interesse: pill bottle*

Per quanto riguarda invece la classe di interesse 'pill bottle', dal report grafico in figura 8.49 si possono facilmente individuare le due aree (feature D e E) che presentano un'influenza positiva. La feature E, in particolare, viene probabilmente scambiata per la bottiglia di pillole individuata dal classificatore.

Classe: pill bottle	Prediction	DI	IR	nIRI	IRP
	6.987818				
Feature D 	5.444403	1.543414	1.28349	**0.030853**	1.20789
Feature E 	6.094153	0.893665	1.14664	**0.017721**	1.082164

Tabella 8.50: *Influenze principali sulla classe di interesse 'pill bottle'*

8.3.10 Gondola

Primi 5 risultati:
dock: 0.22443144
pirate: 0.21815692
container_ship: 0.14292516
drilling_platform: 0.14217149
boathouse: 0.11376126

In questa immagine la classe attesa 'gondola' non è presente nei primi 5 risultati. La prima classe predetta 'dock' tuttavia non si può considerare sbagliata in quanto buona parte raffigura effettivamente un porto. L'analisi di trasparenza in questo caso si focalizzerà quindi sulle classi di interesse 'dock' e 'gondola'.

Figura 8.51: *Segmentazione dell'immagine in 10 feature interpretabili*

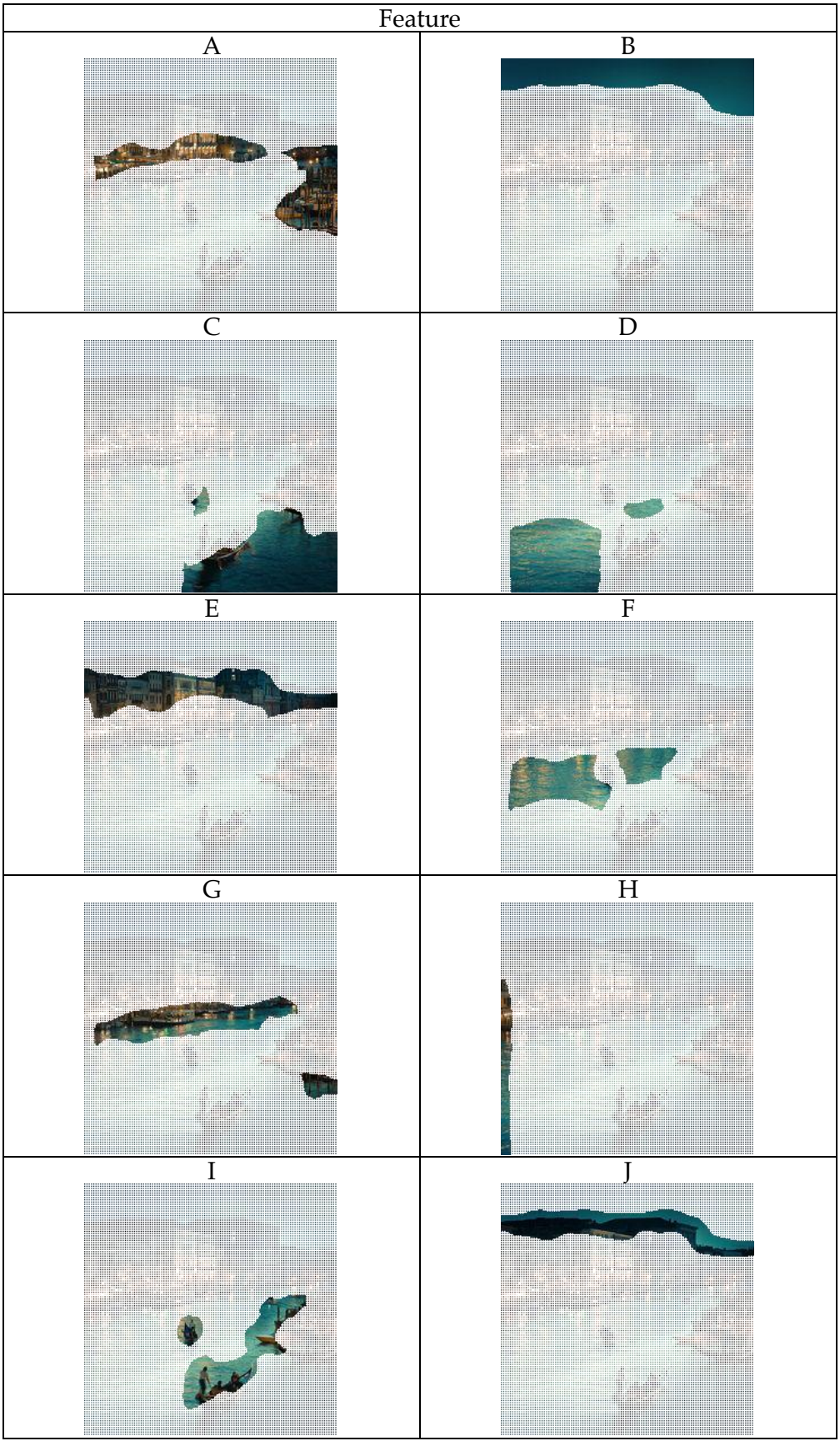


Figura 8.52: *Feature estratte dell'immagine*

Figura 8.53: *Report grafico qualitativo – classe di interesse: dock*

In questo caso il classificatore identifica la classe 'dock' grazie prevalentemente all'influenza delle feature A e G che effettivamente coprono l'area dell'immagine relativa al porto.

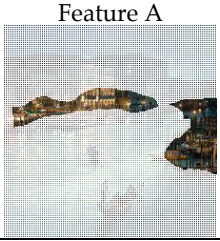
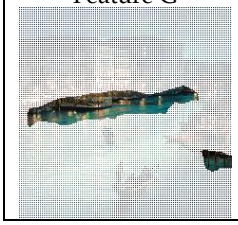
Classe: dock	Prediction	DI	IR	nIRI	IRP
	22.44314				
Feature più influenti positivamente					
Feature A 	13.23656	9.206584	1.69554	**0.173827**	0.803427
Feature G 	16.12109	6.322055	1.39216	**0.117718**	0.82808

Tabella 8.55: *Influenze principali sulla classe di interesse 'dock'*

Figura 8.54: *Report grafico qualitativo – classe di interesse: gondola*

Ripetendo l'analisi per la classe d'interesse 'gondola' e analizzando le feature più influenti (A, E) si può notare che esse individuano prevalentemente le caratteristiche rive del canale veneziano e gli edifici. Questi risultati suggeriscono che il contesto veneziano influisce positivamente sulla classe 'gondola' molto di più di quanto non faccia la feature I che contiene la piccola gondola in evidenza nell'immagine. Sono particolarmente degni di nota gli indici relativi alla feature A che presenta un valore di nIRI non molto alto pari a 0,08 ma anche un indice IRP pari a 6,84 che delinea la feature come univocamente caratterizzante per la classe di interesse.

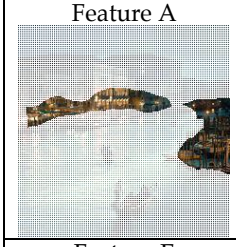
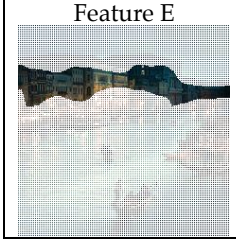
Classe: gondola	Prediction	DI	IR	nIRI	IRP
	0.666045				
Feature più influenti positivamente					
Feature A 	0.04608	0.619965	14.454	**0.082601**	6.84897
Feature E 	0.115799	0.550246	5.75174	**0.031576**	2.484972

Tabella 8.56: *Influenze principali sulla classe di interesse 'gondola'*

8.3.11 Altri esempi notevoli

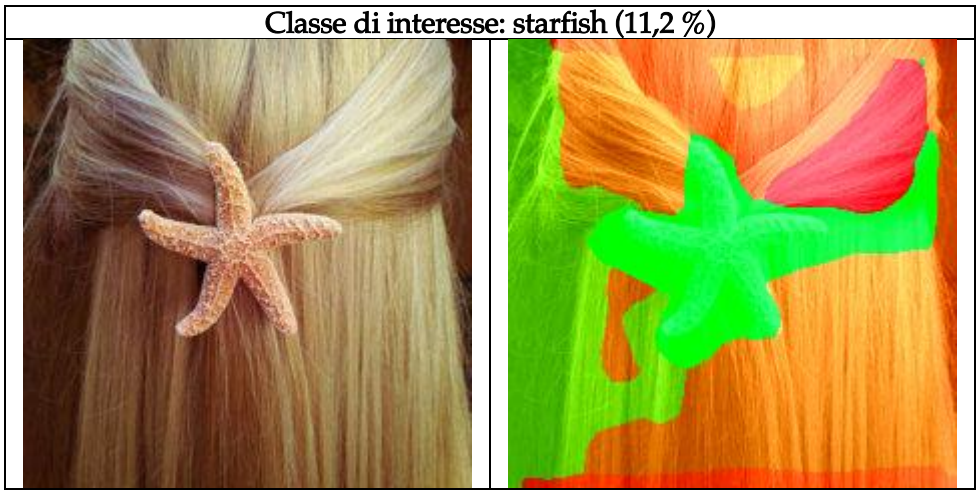


Figura 8.57: *Report grafico qualitativo per la classe di interesse 'starfish'*

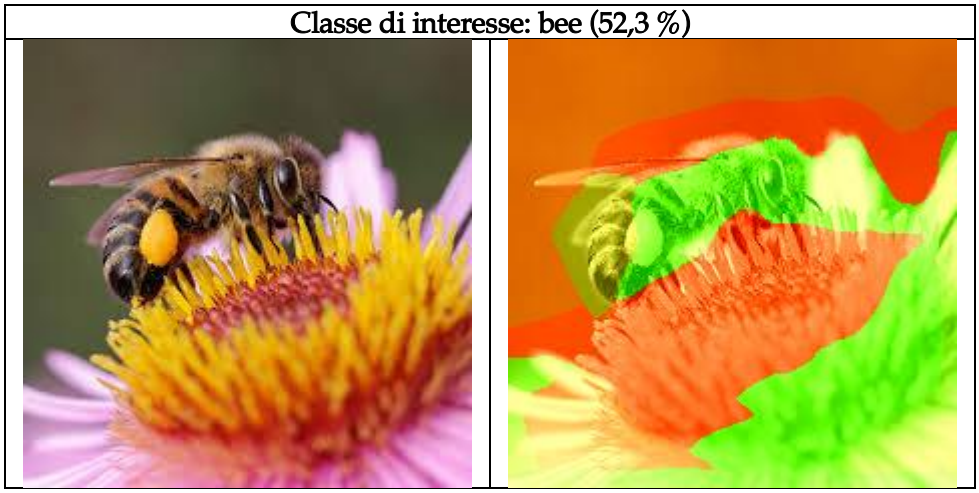


Figura 8.58: *Report grafico qualitativo per la classe di interesse 'bee'*

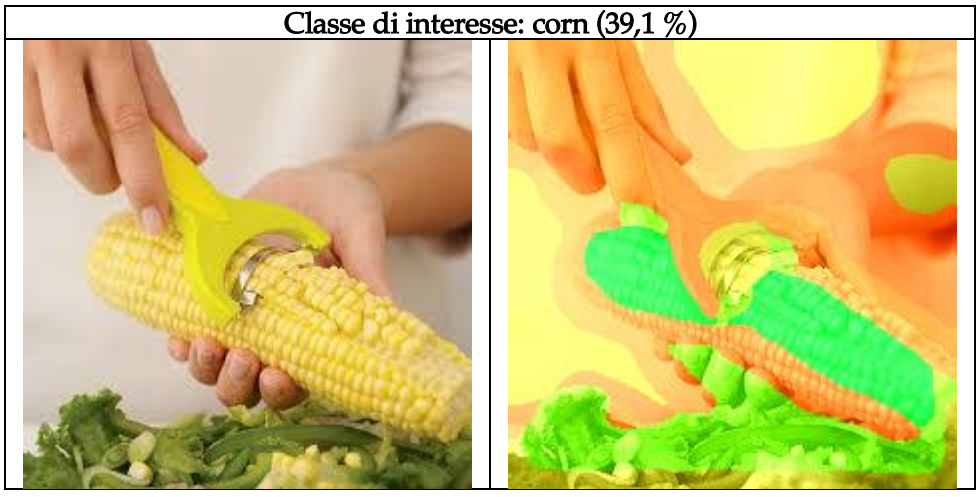


Figura 8.59: *Report grafico qualitativo per la classe di interesse 'corn'*

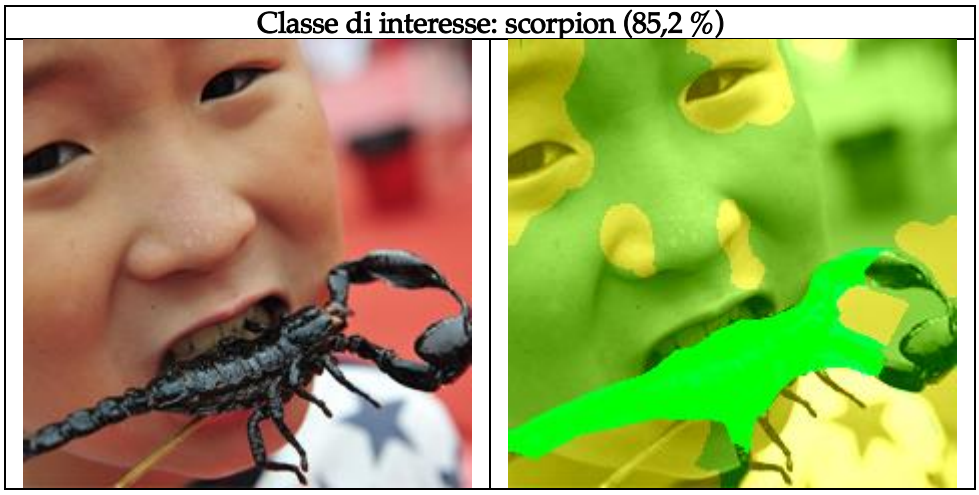


Figura 8.60: *Report grafico qualitativo per la classe di interesse 'scorpion'*

Figura 8.61: *Report grafico qualitativo per la classe di interesse 'carbonara'*

Figura 8.62: *Report grafico qualitativo per la classe di interesse 'table_lamp'*

8.3.12 Analisi locale della classe 'jellyfish'

Dopo lo studio di trasparenza per singole immagini si è passati ad analizzare molteplici oggetti appartenenti a determinate classi così da ottenere una comprensione più profonda del comportamento locale del modello VGG-16. La prima classe analizzata è 'jellyfish'. Si sono sottoposte al classificatore 12 immagini raffiguranti delle meduse e per ciascuna immagine si è individuata la feature più influente sulla classe per capire quali sono le caratteristiche che condizionano maggiormente il modello nella predizione della classe 'jellyfish'.

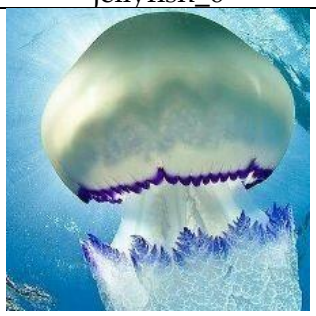
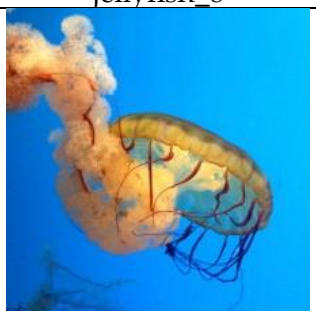
 jellyfish_0	 jellyfish_1	 jellyfish_2
 jellyfish_3	 jellyfish_4	 jellyfish_5
 jellyfish_6	 jellyfish_7	 jellyfish_8
 jellyfish_9	 jellyfish_10	 jellyfish_11

Tabella 8.63: *Le 12 immagini con meduse date in input al modello*

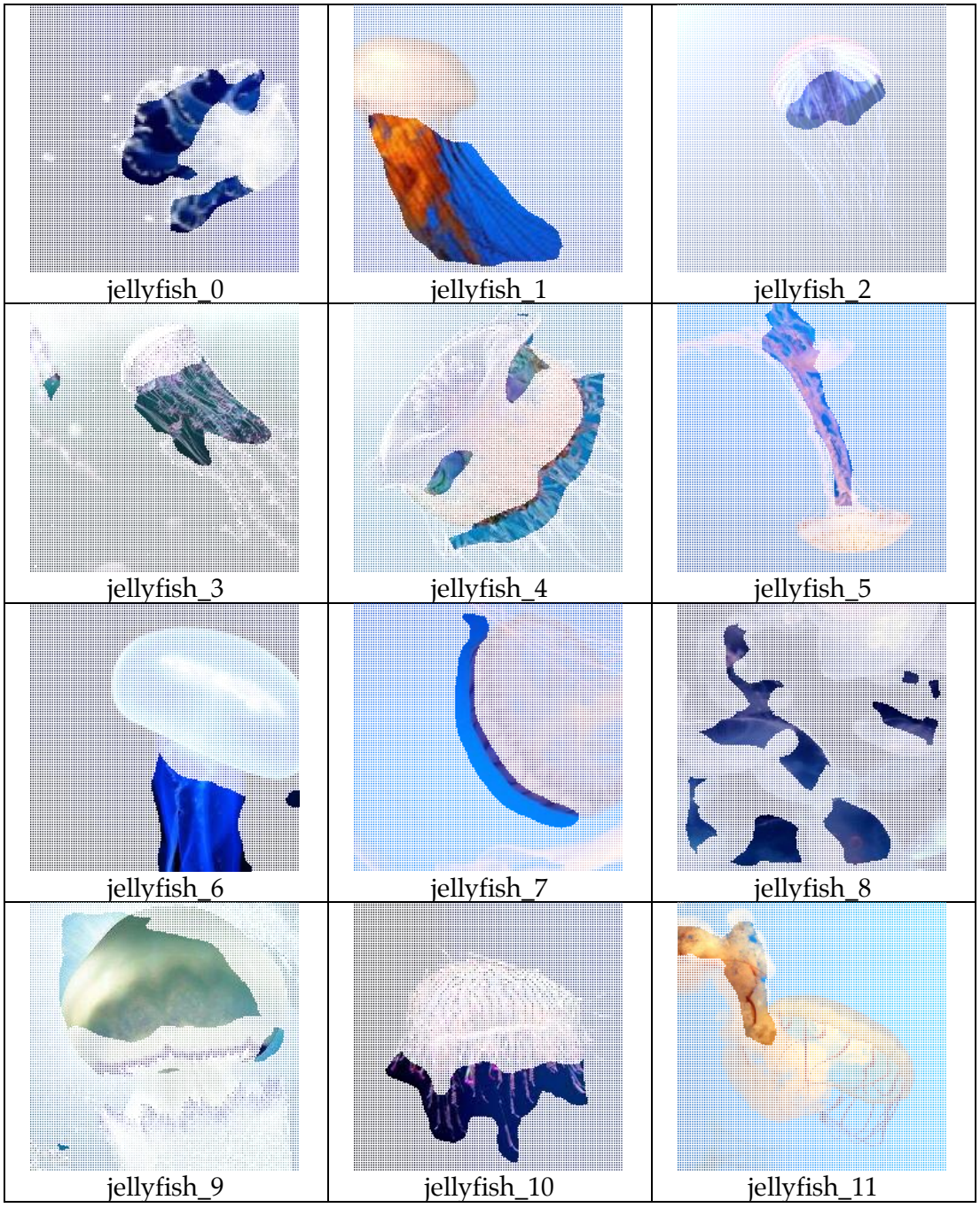


Tabella 8.64: *Le 12 feature più influenti sulla classe 'jellyfish'*

I risultati mostrati in figura 8.64 suggeriscono che il modello è estremamente influenzato dalle feature che comprendono i tentacoli delle meduse. Per ogni immagine la feature individuata è quella relativa alla coda della medusa con l'eccezione delle immagini 7, 8 e 9 dove infatti i tentacoli non sono ben visibili o presentano una forma inusuale.

8.3.13 Analisi locale della classe 'pizza'

 pizza_0	 pizza_1	 pizza_2
 pizza_3	 pizza_4	 pizza_5
 pizza_6	 pizza_7	 pizza_8
 pizza_9	 pizza_10	 pizza_11

Tabella 8.65: *Le 12 immagini con pizze date in input al modello*

Identificando e raccogliendo le feature più influenti sulla predizione 'pizza' per le 12 immagini proposte si ottengono dei risultati alquanto netti. Le feature che contribuiscono maggiormente, come mostrato in figura 8.66, sono infatti quelle relative alla parte centrale delle pizze dove si trova il condimento. Fanno eccezione le immagini 6 e 7 dove, probabilmente a causa del particolare modo in cui sono state tagliate, le feature più influenti sono costituite dai bordi.

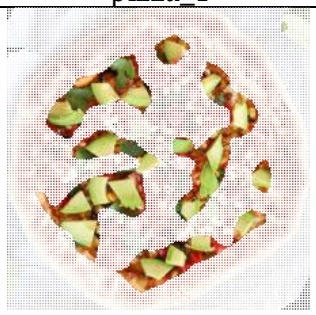
 pizza_0	 pizza_1	 pizza_2
 pizza_3	 pizza_4	 pizza_5
 pizza_6	 pizza_7	 pizza_8
 pizza_9	 pizza_10	 pizza_11

Tabella 8.66: *Le 12 feature più influenti sulla classe 'pizza'*

8.3.14 Analisi locale della classe 'goose'

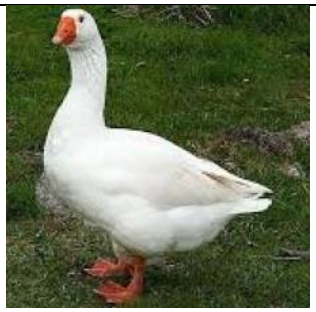
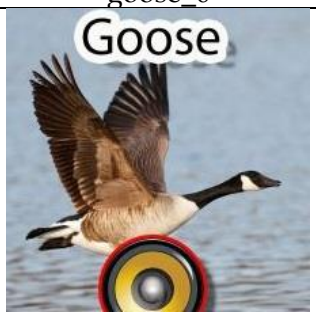
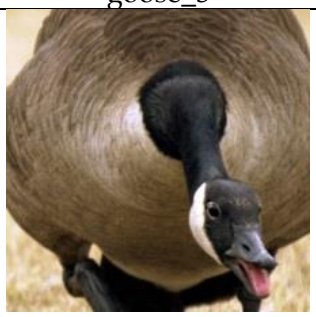
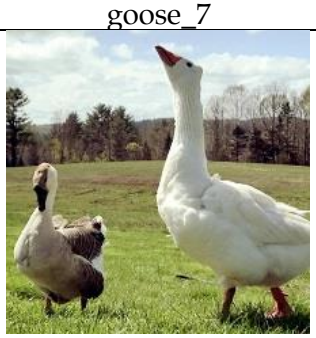
 goose_0	 goose_1	 goose_2
 goose_3	 goose_4	 goose_5
 goose_6	 goose_7	 goose_8
 goose_9	 goose_10	 goose_11

Tabella 8.67: *Le 12 immagini con oche date in input al modello*

Analizzando la classe 'goose' l'identificazione di un'unica caratteristica delle immagini che impatti sulla classe è meno scontata, anche a causa della varietà delle immagini in input. La caratteristica più influente sulla classe sembrebbe essere la testa e, in particolare, il dettaglio dell'occhio, specialmente se in primo piano, come osservabile in tabella 8.68 nelle immagini 0, 4, 6 e 9. Nelle immagini 7 e 8 invece la feature più influente consiste nel caratteristico piumaggio grigiastro delle ali e della coda.

Tabella 8.68: *Le 12 feature più influenti sulla classe 'goose'*

8.3.15 Analisi locale della classe 'hotdog'

Figura 8.69: *Le 12 immagini con hotdog date in input al modello*

L'analisi delle feature più influenti sulla classe 'hotdog', mostrate in figura 8.70, evidenzia che il modello è maggiormente condizionato nella predizione dai dettagli relativi al pane (immagini 1, 2, 7, 9 e 11) e dalle estremità dell'hotdog (immagini 3, 4, 6).

Figura 8.70: *Le 12 feature più influenti sulla classe 'hotdog'*

8.3.16 Efficacia delle analisi di trasparenza

Nei paragrafi precedenti si sono proposti i risultati relativi all'analisi di trasparenza di singole immagini o intere classi. Si vuole adesso valutare l'efficacia della soluzione di trasparenza proposta nel presente lavoro di tesi. Per fare ciò si sono raccolti e rappresentati graficamente i risultati dall'analisi delle 48 immagini sottoposte al classificatore VGG-16 per l'analisi locale delle classi svolta nei paragrafi precedenti. Le immagini sono state scelte per essere equamente distribuite nelle 4 classi 'jellyfish', 'pizza', 'goose' e 'hotdog'.
Nei grafici rappresentati nelle figure 8.74, 8.77, 8.80 e 8.83 si riassumono i risultati dell'approccio sviluppato per 12 immagini relative ad una classe di interesse. La serie in nero indica la predizione della classe in questione da parte del modello VGG-16 per ciascuna immagine, i valori sono compresi nell'intervallo [0, 1]. Le due serie in verde e rosso chiaro indicano i valori massimi e minimi dell'indice nIRI relativo alle feature di ciascuna immagine, le serie in verde e rosso scuro indicano invece i valori del 25° e il 75° percentile dell'indice nIRI per le immagini in questione. Si ha quindi che le aree verdi esprimono la capacità della soluzione proposta di individuare feature positivamente influenti mentre le aree rosse al contrario indicano la capacità di individuare le feature influenti negativamente. Lungo l'asse delle ascisse si hanno le 12 immagini, ordinate in modo decrescente in base al valore di predizione del modello per la classe di interesse.

La tendenza che si osserva dai grafici suggerisce che la soluzione di trasparenza proposta funziona al meglio per le immagini che non presentano delle predizioni troppo nette. Per valori di predizione vicini a 1 solitamente non si riesce a individuare delle feature influenti, o al limite, si riesce a individuarne solo alcune che influiscono positivamente ma in maniera generalmente debole. Per valori di predizione minori si cominciano a individuare delle influenze prevalentemente positive e avvicinandosi allo 0 si trovano sempre più feature dall'influenza negativa. Per valori molto vicini allo zero di nuovo risulta molto difficile individuare influenze positive o negative.

predizione ≅ 1	predizione ∈ (0.5, 1)	predizione ∈ (0, 0.5)	predizione ≅ 0
Influenze positive basse o nulle	Influenze positive medio-alte	Influenze positive medio-basse	Influenze positive nulle
Influenze negative nulle	Influenze negative medio-basse	Influenze negative medio-alte	Influenze negative basse o nulle

Figura 8.71: *Schematizzazione dei risultati ottenuti con la soluzione di trasparenza proposta*

Figura 8.72: *Le 12 immagini con meduse date in input al modello*

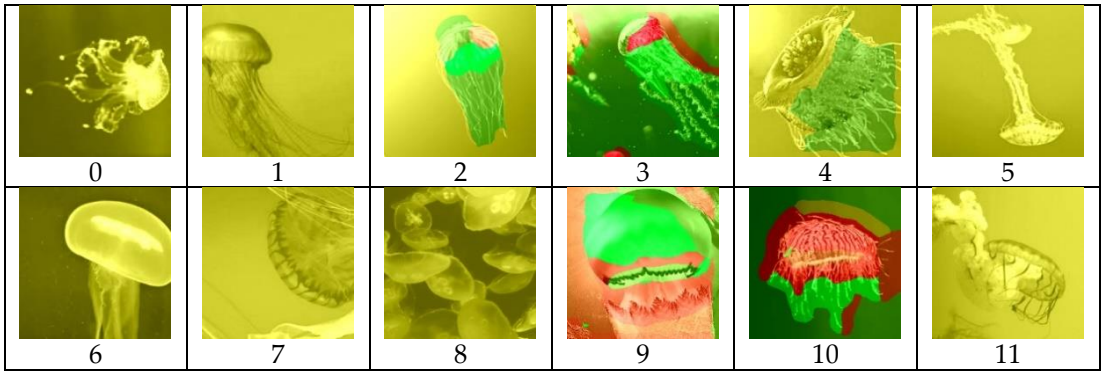


Figura 8.73: *I 12 report grafici qualitativi delle immagini in figura 8.72. Si notino le molte immagini completamente gialle che indicano una bassa efficacia da parte del framework sviluppato nell'individuazione di feature influenti.*

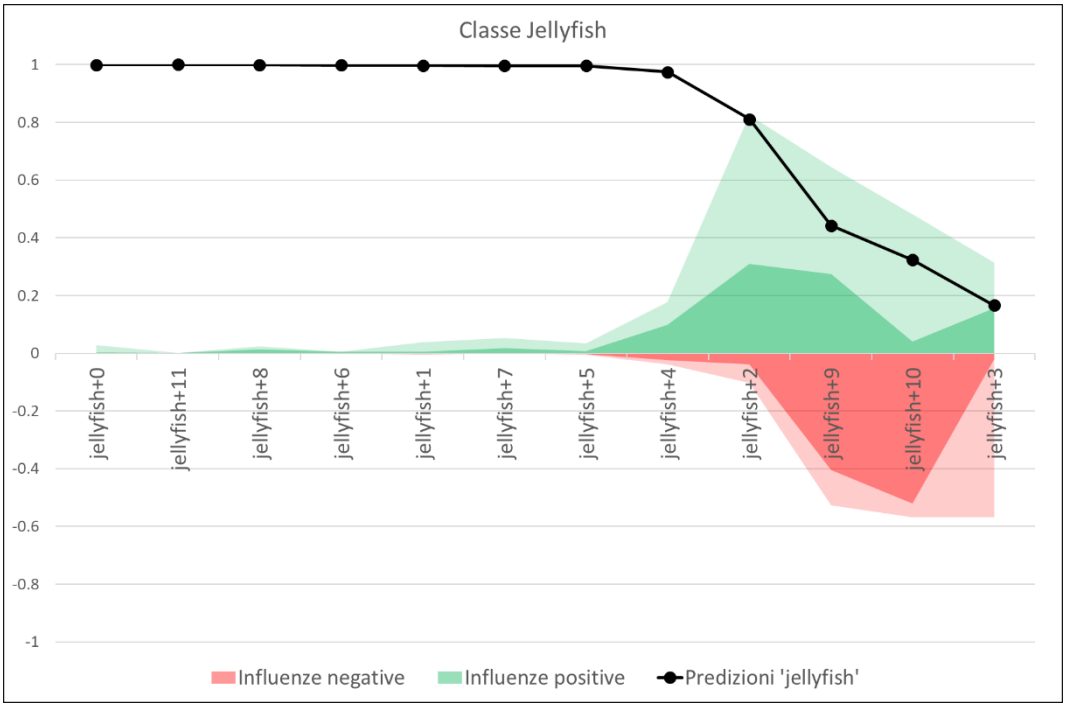


Figura 8.74: *Influenze positive e negative individuate per la classe 'jellyfish'*

Figura 8.75: *Le 12 immagini con pizze date in input al modello*

Figura 8.76: *I 12 report grafici qualitativi delle immagini in figura 8.75*

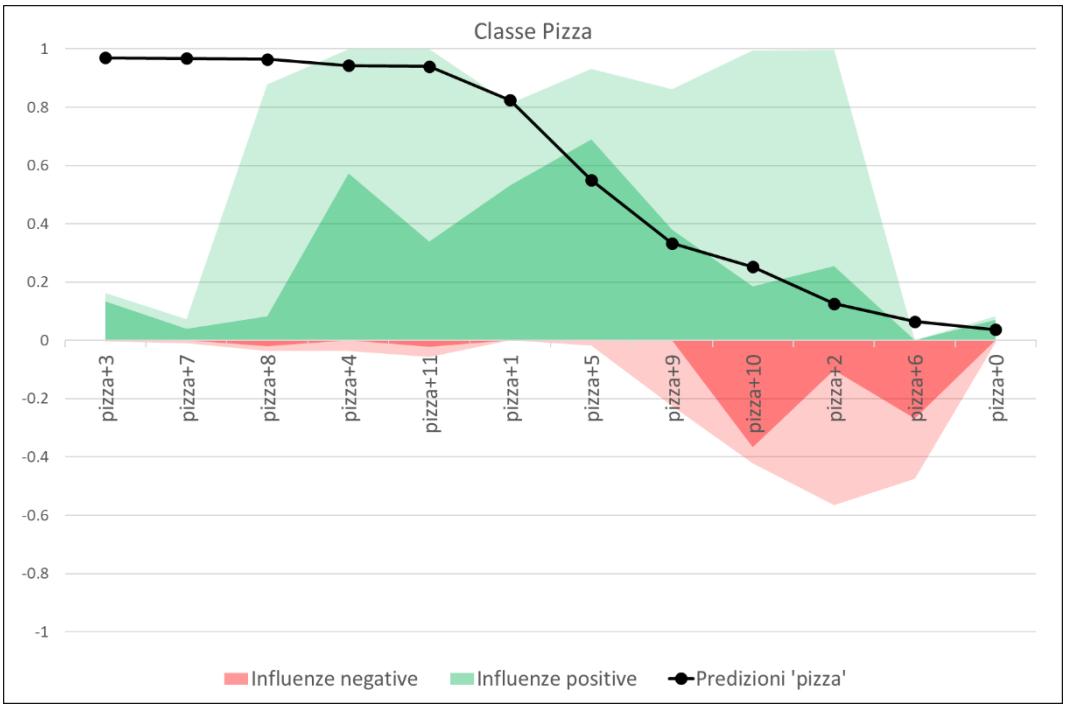


Figura 8.77: *Influenze positive e negative individuate per la classe 'pizze'*

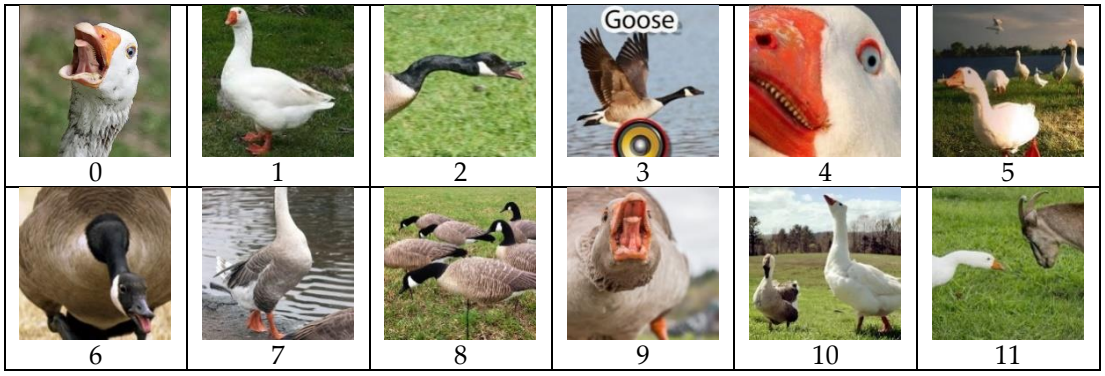


Figura 8.78: *Le 12 immagini con oche date in input al modello*

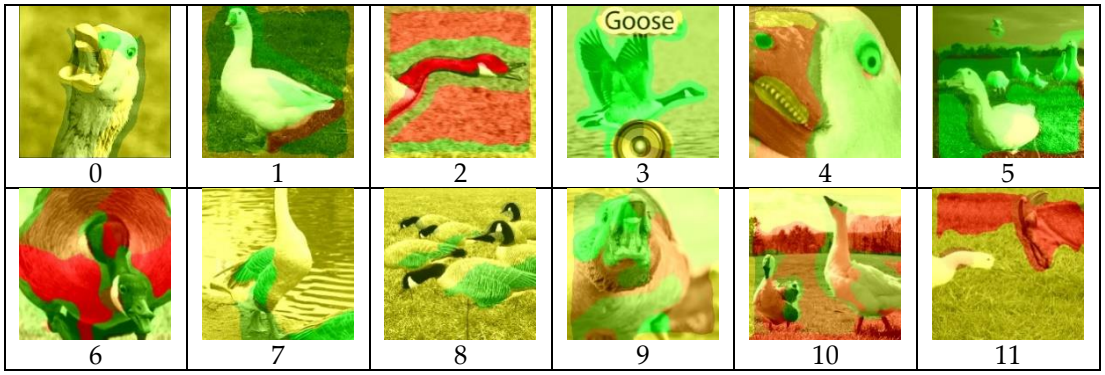


Figura 8.79: *I 12 report grafici qualitativi delle immagini in figura 8.78*

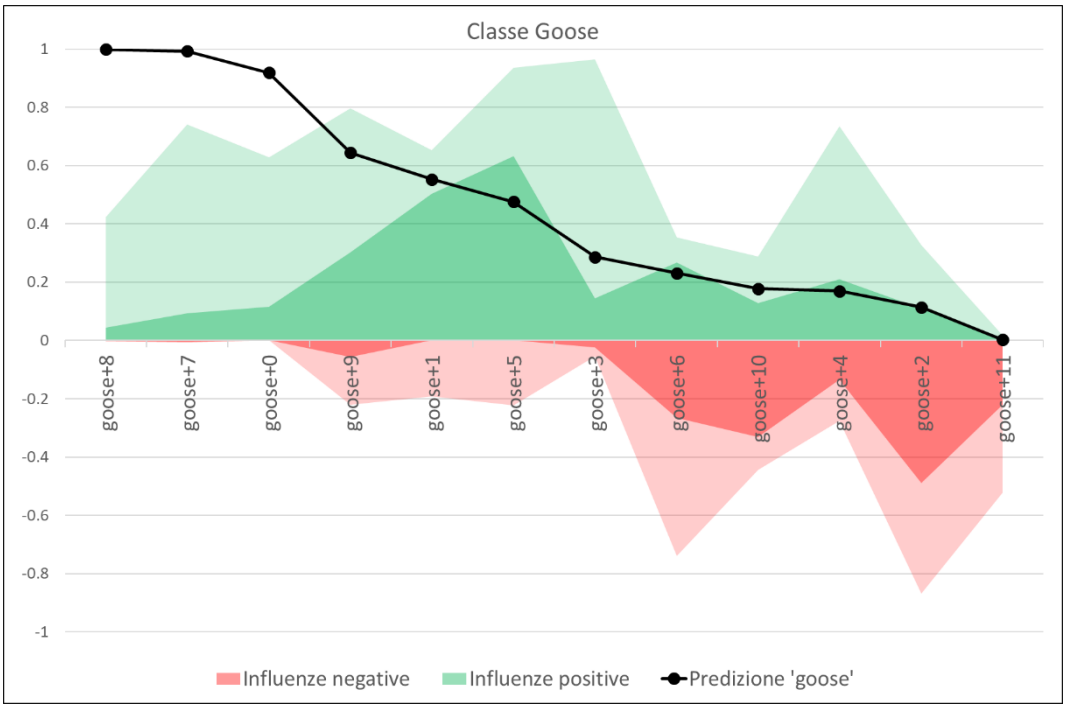


Figura 8.80: *Influenze positive e negative individuate per la classe 'goose'*

Figura 8.81: *Le 12 immagini con hotdog date in input al modello*

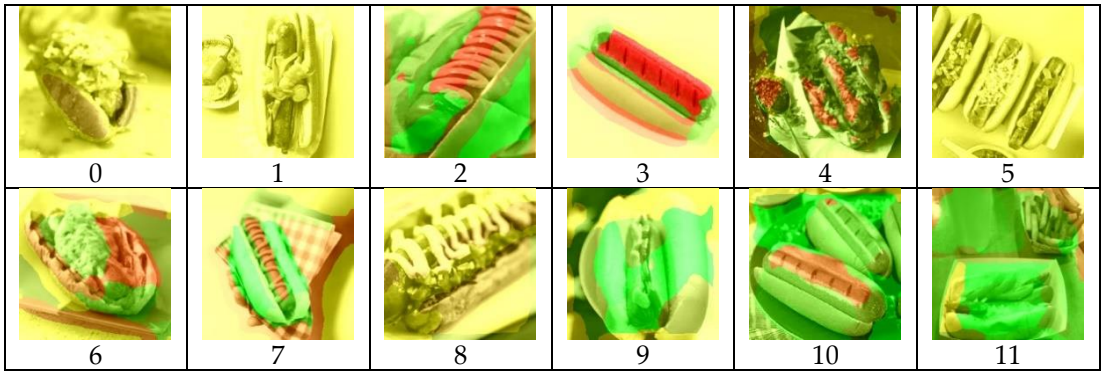


Figura 8.82: *I 12 report grafici qualitativi delle immagini in figura 8.81*

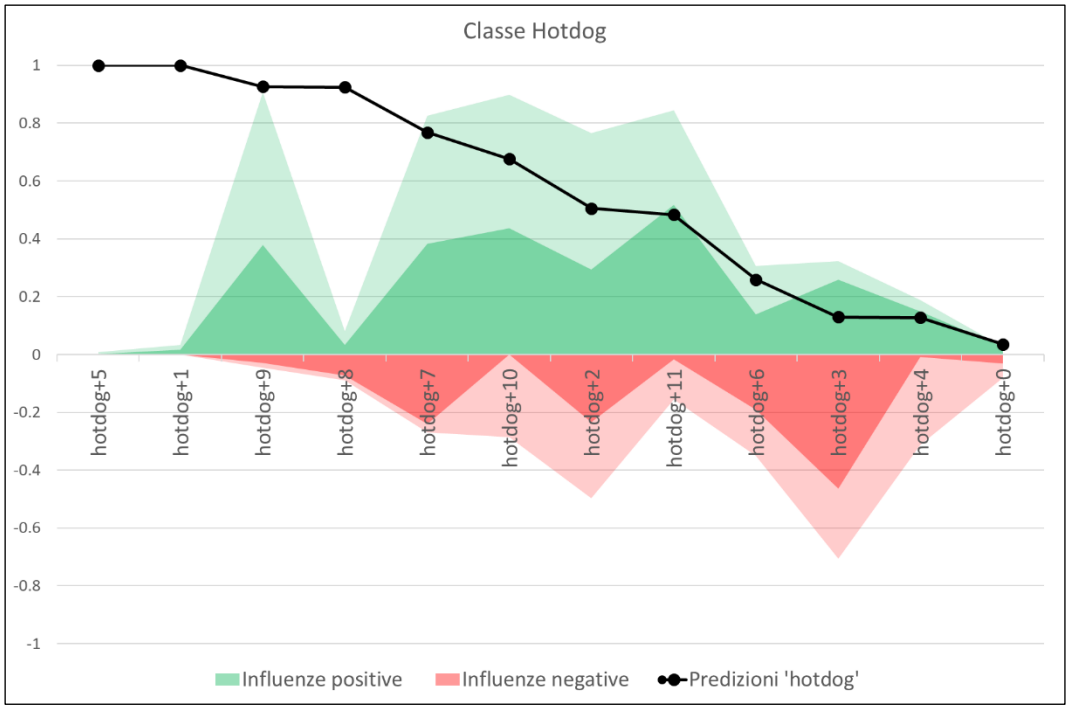


Figura 8.83: *Influenze positive e negative individuate per la classe 'hotdog'*

8.4 – Risultati sperimentali – Recensioni Imdb

In questa sezione verrà riportata la prima parte dei risultati sperimentali ottenuti dall'applicazione delle soluzioni di trasparenza algoritmica sviluppate nel dominio dei documenti testuali. Si presenteranno una serie di recensioni che sono state sottoposte al classificatore basato su ConvNet sviluppato e allenato con il Large Movie Review Dataset e i rispettivi report di trasparenza. Trattandosi di un classificatore che esegue task di sentiment analisys si ha a che fare con soltanto due classi, 'pos' e 'neg', tra loro complementari. Per questo motivo nei report di trasparenza si terrà conto della sola classe 'pos' e si considereranno positive le recensioni con una predizione superiore a 0,5 e negative le altre con predizione inferiore a 0,5. Inoltre, nei report di trasparenza non si terrà conto della misura IRP che con sole due classi perde di significato trattandosi di una metrica che valuta l'influenza inter-classe delle feature.

Le 10 recensioni proposte sono le seguenti:

- The Room (positiva, classificata come negativa)
- Black Panther (negativa, classificata come negativa)
- Dunkirk (negativa, classificata come positiva)
- Inception (negativa, classificata come negativa)
- Pulp Fiction (positiva, classificata come positiva)
- Red Sparrow (positiva, classificata come positiva)
- Lord of the Rings: The Return of the King (negativa, classificata come negativa)
- Three Billboards outside Ebbing Missouri (negativa, classificata come positiva)
- Fight Club (negativa, classificata come negativa)
- Get Out (negativa, classificata come negativa)

8.4.1 The Room

In questo primo esempio si sottopone al modello una recensione positiva ma relativa ad un film noto per essere talmente brutto da diventare divertente. La recensione è quindi ironica e viene identificata come negativa dal modello. Attraverso l'identificazione delle feature e la loro perturbazione è possibile valutare le influenze positive e negative sul risultato della predizione e quindi motivare l'errata predizione.

*Is it possible that tommy wiseau intended the film to be a "black comedy" all along, as the posters would lead one to believe? i don't think so - to have made a film this incompetent on almost every level is nothing short of a miracle. that's right, "the room" is a miracle, part of the holy trinity of bad movies that includes "troll 2" and "dangerous men" (another work of genius), and maybe even "the apple" depending on who you talk to (but then that wouldn't really be a trinity anymore, would it?). ***SPOILERS*** the film is full of many brilliant cinematic devices. the sight of tommy wiseau's naked torso causes men and women to scream in terror, as does "pleasantly plump" juliette danielle's throbbing neck. could it be that images like these were inspired by wiseau's deep affection for the cinema of David Cronenberg? wiseau also has a mastery of staging that is unheralded in modern cinema, as is evinced by the scene in which denny and lisa kneel on the ground with an empty chair visible in the foreground of the shot. i have a feeling that the empty chair is not insignificant, or as claudette would say, "that is not nothing!" on the subject of claudette, i feel that in addition to her "definite" case of breast cancer she may suffer from dementia as well, given that she and lisa seem to have nearly identical conversations every time they share the screen. she may want to get that checked out. and you may want to check out "the room." it will change your life and the lives of those around you forever. in the words of wiseau, "you may not like it, but you will learn something. that's what entertainment is: the process of learning."*
Risultato della classificazione pos: 21.58 % - neg: 78.41 %

Feature	Parole
A	*['modern', 'image', 'subject', 'mastery', 'foreground', 'cinema', 'film', 'process', 'cinematic']*
B	*['breast', 'plump', 'bad']*
C	*['deep', 'life', 'genius', 'right', 'inspired', 'brilliant', 'feeling', 'depending', 'insignificant', 'screen', 'change']*
D	*['level', 'scene', 'dementia', 'torso', 'given', 'sight', 'conversation', 'nearly', 'intended', 'short', 'identical', 'possible', 'room']*
E	*['anymore', 'really', 'think', 'cause', 'suffer', 'wouldn', 'maybe', 'say', 'talk', 'like']*
F	*['includes', 'juliette', 'share', 'affection', 'ha', 'addition']*
G	*['ground', 'black', 'lead', 'dangerous', 'case']*
H	*['word', 'believe', 'holy', 'don']*
I	*['tommy', 'apple']*
J	*['pleasantly', 'poster']*
K	*['shot', 'incompetent']*
L	*['lisa', 'denny']*
M	*['empty']*
N	*['miracle']*
O	*['movie', 'danielle', 'forever']*
P	*['scream', 'troll']*
Q	*['device']*
...	

Tabella 8.84: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.215894 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
F	includes juliette share affection ha addition	0.147539	0.068355	1.4633	**0.12796**
A	modern image subject mastery foreground cinema film process cinematic	0.151152	0.064743	1.42833	**0.121112**
J	pleasantly poster	0.177907	0.037987	1.21352	**0.071841**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
K	shot incompetent	0.584931	-0.36904	0.369094	**-0.53185**
B	breast plump bad	0.347979	-0.13209	0.620423	**-0.22771**
M	empty	0.309362	-0.09347	0.69787	**-0.16608**

Tabella 8.85: *Influenze principali sulla predizione*

Dai risultati esposti in figura 8.85 si registra la forte influenza negativa da parte delle feature K, B e M. La prima in particolare, composta dalle parole 'shot' e 'incompetent' presenta un indice nIRI pari a -0,53. L'influenza positiva da parte delle feature F, A e J invece è significativamente più contenuta.

8.4.2 Black Panther

In questo esempio si è proposta al classificatore una recensione negativa del film 'Black Panther' correttamente identificata dal modello.

After the excellent.....and hilarious.....Thor 3, I was rather looking forward to this. Positive reviews had also whetted the appetite. Perhaps it was the expectation that undermined the enjoyment? The genre is a little hard to quantify, but it reminded a lot of Wonder Woman. Except, instead of the WW2 setting, we had a kind of Bond-style thriller instead, complete with ridiculous gadgets and under-developed baddies. The film looks great (some dodgy CGI aside) - I liked the armour, weaponry and cultural aesthetics. I also felt a lot of the performances were quite compelling. Unfortunately, it is all rather dull. And doesn't feature anything we haven't seen before. What is most disappointing, is that it doesn't seem to tie into the Marvel continuity at all (apart from the unremarkable Everett Ross, who appeared from nowhere in Civil War and is a poor attempt to replace Agent Coulson). Like Wonder Woman, this is simply a "Black Panther tale" and you can take it or leave it when embarking on your next Marvel marathon. I also found the advanced technology to be completely over the top. Weapons and armour, I can understand. Miracle healing and Panther suits concealed within necklace teeth is VERY hard to swallow indeed. Vibranium was an indestructible Alien metal that could absorb sound. Fair enough. Great material for weapons and armour. But spaceships and miracle healing? Really? Couldn't we have toned it down just a little? I was bored. My friend was bored. After his wonderful debut in Civil War, this was a bit of a wasted opportunity.
Risultato della classificazione neg: 86.58 % - pos: 13.41 %

Feature	Parole
A	*['ridiculous', 'poor', 'continuity']*
B	*['necklace', 'aesthetic', 'cultural', 'style']*
C	*['opportunity', 'enjoyment', 'replace', 'kind', 'doesn', 'leave', 'attempt', 'marathon', 'feature', 'apart', 'complete', 'material', 'simply', 'aside', 'tie']*
D	*['advanced', 'haven']*
E	*['embarking', 'miracle', 'swallow']*
F	*['technology', 'spaceship']*
G	*['indestructible', 'genre', 'thriller']*
H	*['dull', 'just', 'bored', 'undermined', 'instead', 'wasted', 'bit', 'dodgy']*
I	*['gadget']*
J	*['teeth']*
K	*['baddie']*
L	*['reminded', 'developed']*
M	*['tale', 'absorb', 'wonderful', 'compelling', 'setting', 'healing', 'excellent', 'debut']*
N	*['felt', 'wa', 'unremarkable', 'disappointing', 'toned', 'understand', 'film', 'liked', 'forward', 'review', 'expectation', 'completely', 'appeared']*
O	*['sound', 'metal']*
P	*['appetite', 'hilarious', 'look', 'concealed', 'armour', 'looking', 'hard', 'suit']*
Q	*['weaponry', 'weapon']*

Tabella 8.86: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.134169 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
M	tale absorb wonderful compelling setting healing excellent debut	0.043268	0.090901	3.10089	**0.237334**
P	appetite hilarious look concealed armour looking hard suit	0.089168	0.045001	1.50468	**0.088938**
I	gadget	0.114037	0.020132	1.17654	**0.039199**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
H	dull just bored undermined instead wasted bit dodgy	0.500532	-0.36636	0.268053	**-0.59431**
A	ridiculous poor continuity	0.259284	-0.12512	0.517458	**-0.23461**
C	opportunity enjoyment replace kind doesn leave attempt marathon feature apart complete material simply aside tie	0.250709	-0.11654	0.535158	**-0.21883**

Tabella 8.87: *Influenze principali sulla predizione*

Nella tabella in figura 8.87 si identifica il significativo impatto positivo della feature M (nIRI = 0,24) e quello negativo molto più forte delle feature H (nIRI = -0,59), A e C.

8.4.3 Dunkirk

In questo esempio la recensione negativa del film 'Dunkirk' è stata classificata come positiva dal modello.

Christopher Nolan is a visual genius: this film is just stunning to look at. From the bullets flying near soldier's head to gorgeous explosions, it transports you right into the war zone within the first 10 minutes of the screening. Furthermore, it is an accurate description of the "Miracle of Dunkirk". This is where the movie nails it. However, apart from that, I've had a couple of issues with it. The film is told from three perspectives: Air, Mole and Sea. This is where the movie falls a bit short. The editing at some points feels lazy and not very consistent: it cuts from a dramatic scene or intense action scene very quickly. The film is told in a non-linear way: this makes us watch certain scenes twice through different perspectives. Although this could've been done in a very interesting way, it's very difficult to keep track of whose perspective we're watching at times. Even when certain semi-important (I'll get back to this later) soldiers die, it took me a while to realise this happened. This is where my third and final problem come into question: the characters lack depth. You don't care about the main character, nor any of the other soldiers that are dying. If I'm watching a film about war, I like to bond with the characters I'm seeing on screen. If none of them show any real emotion, the viewer won't as well. All in all, Dunkirk could've been amazing. I personally don't understand why it has such a high rating besides being directed by a very well-known director / starring famous actors (including infamous Harry Styles) / being a war biography. Disappointing.
Risultato della classificazione neg: 17.7 % - pos: 82.29 %

Feature	Parole
A	*['biography']*
B	*['transport', 'visual', 'flying']*
C	*['die', 'care']*
D	*['amazing', 'don', 'linear', 'stunning', 'dramatic', 'main', 'depth', 'character', 'consistent']*
E	*['known', 'intense', 'screen', 'near', 'head', 'starring', 'gorgeous', 'nail', 'directed', 'ha', 'track']*
F	*['bullet']*
G	*['soldier', 'war', 'zone']*
H	*['action', 'explosion']*
I	*['10', 'rating', 'movie', 'high']*
J	*['semi', 'apart', 'non', 'll', 'quickly', 'cut', 'couple', 'seeing', 'lack', 'editing', 'twice', 'like']*
K	*['bond']*
L	*['screening', 'director', 'short', 'film']*
M	*['famous', 'final', 'scene', 'including', 'time', 'infamous']*
N	*['later', 'dying', 'fall']*
O	*['difficult', 'make', 'description', 'issue', 'understand', 'told', 'accurate', 'genius', 'won', 'problem', 'right', 'way', 'viewer', 'different', 'emotion', 'perspective', 'question', 'point', 'happened', 'personally', 'certain', 'realise', 'important', 'took']*
P	*['lazy', 'just']*

Tabella 8.88: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.822958 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
O	difficult make description issue understand told accurate genius won problem right way viewer different emotion perspective question point happened personally certain realise important took	0.73291	0.090049	1.12286	**0.15348**
M	famous final scene including time infamous	0.735479	0.087479	1.11894	**0.149706**
D	amazing don linear stunning dramatic main depth character consistent	0.738684	0.084274	1.11409	**0.144958**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
J	semi apart non ll quickly cut couple seeing lack editing twice like	0.922775	-0.09982	0.891829	**-0.16732**
P	lazy just	0.888156	-0.0652	0.926592	**-0.11565**
I	10 rating movie high	0.841888	-0.01893	0.977515	**-0.03649**

Tabella 8.89: *Influenze principali sulla predizione*

8.4.4 Inception

In questo esempio viene data in input al modello una recensione negativa del film 'Inception' che viene correttamente classificata come tale.

Well, first of all, I have to absolutely clearly admit that I'm writing this review to lower the rating of this movie, since it was one of the biggest movie disappointments of my life. Currently 9.1 for this thing is unbelievably high. But I guess I get why the rating of this movie is so high. It makes the people feel good about themselves because they understood the "complicated" plot. Yes, thats it! Inception actually managed to plant the idea to the people that it was a good movie. At best its an average action movie with too long boring action scenes with the "added value" of a "smart" story. However the story is not smart at all. Its about people that somehow are able to get into people's dreams and do stuff there, or even go to dreams in dreams, and dreams in dreams in dreams, and limbo... Wow, you get it? You are really smart, and should rate this movie 10!!! But seriously, this story is really stupid. Why should I care that the Japanese guy wants to destroy the other guy's (the scarecrow from Batman) company? How did the architect girl become expert on dreams and psychology after two dream sessions? How are they even able to get into the dreams? I guess the acting overall is not bad, but who cares if the plot is so annoying. And finally, the twist at the end is really predictable. Yay, the spinning thing stops to spin: he might be dreaming! Really original. You might say that I did not get the message of the movie. Well, I did get it, but it was too stupid to care about it.
Risultato della classificazione neg: 96.74 % - pos: 3.25 %

Feature	Parole
A	*['destroy']*
B	*['scarecrow', 'high']*
C	*['plot', 'unbelievably', 'bad']*
D	*['good', 'movie', 'thing', 'stupid', 'guess', 'stuff', 'say', 'really', 'thats', 'acting']*
E	*['dreaming', 'story', 'life', 'girl', 'complicated', 'psychology', 'architect', 'finally', 'message', 'dream']*
F	*['10', 'rating', 'rate']*
G	*['guy', 'smart', 'care', 'people']*
H	*['company', 'limbo', 'expert', 'spin']*
I	*['admit', 'added', 'make', 'idea', 'boring', 'biggest', 'overall', 'predictable', 'value', 'writing', 'able', 'understood', 'wa', 'disappointment', 'annoying', 'managed', 'long', 'end', 'review', 'actually', 'clearly', 'average', 'absolutely', 'twist']*
J	*['seriously', 'lower', 'spinning', 'stop']*
K	*['plant']*
L	*['session']*

Tabella 8.90: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.032507 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
E	dreaming story life girl complicated psychology architect finally message dream	0.007329	0.025178	4.43561	**0.10503**
G	guy smart care people	0.018008	0.014499	1.80514	**0.033073**
B	scarecrow high	0.023452	0.009054	1.38607	**0.018725**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
D	good movie thing stupid guess stuff say really thats acting	0.120727	-0.08822	0.269256	**-0.26003**
I	admit added make idea boring biggest overall predictable value writing able understood wa disappointment annoying managed long end review actually clearly average absolutely twist	0.086848	-0.05434	0.374293	**-0.14202**
K	plant	0.035192	-0.00269	0.923696	**-0.00536**

Tabella 8.91: *Influenze principali sulla predizione*

8.4.5 Pulp Fiction

La recensione positiva del film 'Pulp Fiction' sottomessa al classificatore in questo esempio viene correttamente identificata come positiva dal modello, ma con una percentuale non troppo netta pari a circa il 61,2 %. Dall'analisi dell'influenza delle feature è possibile comprendere il motivo di questo risultato.

To put this in context, I am 34 years old and I have to say that this is the best film I have seen without doubt and I don't expect it will be beaten as far as I am concerned. Obviously times move on, and I acknowledge that due to its violence and one particularly uncomfortable scene this film is not for everyone, but I still remember watching it for the first time, and it blew me away. Anyone who watches it now has to remember that it actually changed the history of cinema. In context- it followed a decade or more of action films that always ended with a chase sequence where the hero saved the day - you could have written those films yourself. Pulp had you gripped and credited the audience with intelligence. There is not a line of wasted dialogue and the movie incorporates a number of complexities that are not immediately obvious. It also resurrected the career of Grease icon John Travolta and highlighted the acting talent of Samuel L Jackson. There are many films now that are edited out of sequence and have multiple plots etc but this is the one they all want to be, or all want to beat, but never will.
Risultato della classificazione neg: 38.79 % - pos: 61.2 %

Feature	Parole
A	*['number']*
B	*['beat', 'beaten', 'away']*
C	*['credited']*
D	*['particularly', 'cinema', 'history', 'context', 'audience']*
E	*['uncomfortable']*
F	*['written', 'expect', 'movie', 'saved', 'obvious', 'wasted', 'dialogue', 'don']*
G	*['icon', 'resurrected']*
H	*['remember', 'highlighted', 'immediately', 'concerned', 'blew', '34', 'ended', 'decade', 'followed', 'changed', 'far', 'doubt', 'time', 'year']*
I	*['intelligence', 'acknowledge']*
J	*['hero', 'sequence', 'violence', 'action']*
K	*['talent', 'best', 'incorporates', 'film', 'career', 'ha']*
L	*['multiple', 'plot']*
M	*['gripped', 'complexity']*
N	*['chase']*
O	*['edited', 'scene']*

Tabella 8.92: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.612059 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
K	talent best incorporates film career ha	0.375523	0.236535	1.62988	**0.346682**
H	remember highlighted immediately concerned blew 34 ended decade followed changed far doubt time year	0.4523	0.159759	1.35321	**0.250513**
M	gripped complexity	0.521381	0.090677	1.17392	**0.155185**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
F	written expect movie saved obvious wasted dialogue don	0.901319	-0.28926	0.67907	**-0.38363**
E	uncomfortable	0.663737	-0.05168	0.92214	**-0.09395**
O	edited scene	0.622136	-0.01008	0.983802	**-0.01976**

Tabella 8.93: *Influenze principali sulla predizione*

8.4.6 Red Sparrow

In questo esempio si è data in input al modello una recensione positiva del film 'Red Sparrow'. Il classificatore ha correttamente identificato la classe positiva con un punteggio pari a al 61,15 %.

Red Sparrow is not what it seems to be and that's actually a good thing. Reunited with director Francis Lawrence, Jennifer Lawrence brings the star power to this cold and rough thriller. This is not an action movie as it has barely any action sequences, but it's a clever and well thought out political drama. It's one of those few interesting cases when you can't quite read the protagonist, and Jennifer proves a wild card here. The chemistry just isn't there though, it's hard to buy Lawrence and Edgerton's romance. And it bothers a bit that all the russian folks are played by americans (it can't be that hard to find decent russian actors). But make no mistakes, this is a hell of ride, it's violent, it's brutal, and it's nasty bones will creep up on you.
Risultato della classificazione neg: 38.84 % - pos: 61.15 %

Feature	Parole
A	*['good', 'thriller']*
B	*['bit', 'ride']*
C	*['card', 'played']*
D	*['hard', 'movie', 'thing', 'make', 'barely', 'bother', 'buy', 'just', 'mistake']*
E	*['clever', 'prof', 'brings', 'chemistry', 'rough']*
F	*['american', 'russian']*
G	*['sequence', 'decent', 'nasty', 'bone', 'hell', 'ha', 'isn']*
H	*['wild', 'folk', 'romance', 'power']*
I	*['violent', 'director', 'protagonist', 'case']*
J	*['political']*
K	*['creep', 'actually']*
L	*['brutal', 'drama', 'cold']*

Tabella 8.94: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.611554 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
L	brutal drama cold	0.406657	0.204897	1.50386	**0.307663**
B	bit ride	0.423215	0.188339	1.44502	**0.286982**
E	clever prof brings chemistry rough	0.497044	0.11451	1.23038	**0.1896**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
D	hard movie thing make barely bother buy just mistake	0.824126	-0.21257	0.742064	**-0.30758**
G	sequence decent nasty bone hell ha isn	0.711563	-0.10001	0.859451	**-0.16827**
I	violent director protagonist case	0.659522	-0.04797	0.927269	**-0.08777**

Tabella 8.95: *Influenze principali sulla predizione*

8.4.7 Lord of the Rings: The Return of the King

Questa recensione del film 'Lord of the Rings' è stata correttamente identificata come negativa.

While the first two films of the trilogy where very good in my opinion, the third one is a complete catastrophe, in every aspect one can think of. Its attitude towards the book (and without the book, no one would have heard of Peter Jackson, except few admirers) moves between insult to ignorance - almost no text from the original survives, and the characters and scenes are terribly distorted. Even the climax scene is changed from the genius of Tolkien to a bad Holiwoodic cliché. The most disastrous is what the movie does to the character of Frodo, which, from the original tragic hero, becomes a stupid and gullible fellow, which does nothing (expect stupid things). The "go home Sam" scene, should be candidate to one of the worse movie scenes ever. Other characters, such as Denethor, Faramir, Aragorn, Theoden or Elrond, are either inconsistent, irrational, stupid, badly acted, or all of the above. In addition, unlike the beautiful design of the middle-earth in general in the first movie, or of Rohan in second, there is nothing left from Gondor, except one picture of Minas-Tirith and few fires. It amazes me that not only this film is considered "excellent" by many viewers, but that it is considered the best of the three.
Risultato della classificazione neg: 92.08 % - pos: 7.91 %

Feature	Parole
A	*['trilogy']*
B	*['design', 'beautiful', 'inconsistent']*
C	*['second', 'catastrophe', 'earth', 'thing', 'amazes']*
D	*['book', 'text']*
E	*['survives', 'hero']*
F	*['ignorance']*
G	*['insult', 'badly', 'movie', 'stupid', 'bad', 'terribly', 'worse', 'gullible', 'disastrous', 'complete', 'irrational']*
H	*['good', 'excellent', 'candidate']*
I	*['picture', 'genius', 'considered', 'film', 'distorted', 'middle', 'climax', 'changed', 'opinion', 'expect', 'viewer', 'general', 'heard', 'aspect', 'left', 'fellow', 'admirer', 'tragic', 'acted', 'character', 'unlike', 'attitude', 'addition', 'home']*
J	*['fire']*

Tabella 8.96: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.079176 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
H	good excellent candidate	0.010978	0.068199	7.2126	**0.333931**
I	picture genius considered film distorted middle climax changed opinion expect viewer general heard aspect left fellow admirer tragic acted character unlike attitude addition home	0.059697	0.01948	1.32631	**0.038945**
J	fire	0.063803	0.015373	1.24095	**0.030506**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
G	insult badly movie stupid bad terribly worse gullible disastrous complete irrational	0.861902	-0.78273	0.091862	**-0.89575**
C	second catastrophe earth thing amazes	0.09509	-0.01591	0.832643	**-0.03135**
B	design beautiful inconsistent	0.090815	-0.01164	0.87184	**-0.02296**

Tabella 8.97: *Influenze principali sulla predizione*

L'influenza della feature G risulta essere la principale ragione della predizione negativa da parte del classificatore.

8.4.8 Three Billboards Outside Ebbing Missouri

Questa recensione negativa viene erroneamente classificata come positiva dal modello.

First, let's all just accept the premise that police beat up random black people for no reason whatsoever. Also, police are inept, because they can't find someone to arrest for the murder and rape of your daughter, even though there is zero forensic evidence. Next let's take a moment to reflect on the time you told your daughter "I hope you get raped on the way too". Now it's time to start lashing out at the world because you are angry. Start by committing 2 felonies against a dentist. It's also a good idea to go into a police station and tell them what horrible people they are. When your signs get set on fire, climb onto one of the burning signs and stand at the top of the flames. Next, fire-bomb a police station, because you are still mad, even though there is still no evidence or suspects. Finally, if you can't find the actual person who committed the crime, instead go murder a complete stranger on a hunch. This is all very profound and heart-wrenching...can't you tell by the music? The police chief isn't so bad though, because he has cancer and is spitting up blood. He thinks it's best to leave the hospital, because what do doctors know about dealing with cancer. He doesn't want his wife and kid's final memories of him to be slowly dying in a hospital, so instead their final memories will be of him blowing his own brains out in the barn. Thankfully, a streetwise black police chief came in to take over for the inept and racist white police officers. Now things will finally get done.
Risultato della classificazione neg: 29.71 % - pos: 70.28 %

Feature	Parole
A	*['mad', 'just', 'blood', 'burning', 'time']*
B	*['let', 'idea', 'brain', 'want', 'instead', 'isn', 'random', 'premise']*
C	*['black', 'white', 'racist', 'people']*
D	*['hospital', 'climb', 'wife', 'dying']*
E	*['blowing', 'came', 'beat', 'finally', 'actual', 'sign', 'set', 'hope', 'station', 'spitting', 'stand', 'memory']*
F	*['final', 'start', 'moment', 'way', 'stranger', 'leave', 'slowly', 'angry', 'heart', 'tell', 'streetwise']*
G	*['dentist']*
H	*['flame', 'hunch', 'doesn']*
I	*['committing', 'person', 'barn', 'told', 'accept', 'profound', 'committed', 'evidence', 'forensic', 'murder', 'dealing', 'reflect']*
J	*['bomb']*
K	*['doctor']*
L	*['wrenching', 'cancer', 'daughter']*
M	*['rape', 'raped', 'lashing']*
N	*['horrible', 'reason', 'zero', 'bad', 'complete', 'whatsoever', 'inept']*
O	*['chief', 'officer', 'crime', 'arrest', 'suspect', 'police', 'ha']*

Tabella 8.98: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.702891 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
O	chief officer crime arrest suspect police ha	0.449369	0.253522	1.56417	**0.358412**
I	committing person barn told accept profound committed evidence forensic murder dealing reflect	0.473542	0.229349	1.48433	**0.331077**
C	black white racist people	0.56105	0.14184	1.25281	**0.225357**
L	wrenching cancer daughter	0.56568	0.13721	1.24256	**0.21931**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
N	horrible reason zero bad complete whatsoever inept	0.914851	-0.21196	0.768312	**-0.30494**
E	blowing came beat finally actual sign set hope station spitting stand memory	0.729437	-0.02655	0.963607	**-0.05045**

Tabella 8.99: *Influenze principali sulla predizione*

8.4.9 Fight Club

In questo esempio una recensione negativa del film 'Fight Club' viene correttamente classificata con un punteggio del 59,63 %.

*How can we get Brad Pitt, Edward Norton and Meatloaf to beat the c**p out of each other? This movie solves that "problem," and gives the studio license to shoot a film veiled in psuedo-philosophies and violence. It's just dumb. Don't try looking beneath the surface, because nothing is there (maybe that's the point). It provides that men are imasculated and shackled by modern life... their spirits crushed by their jobs and their possessions. In order to escape, you fight. It's so simple! Give up the banalities of ordinary life and find your individuality in a black t-shirt fight gang. Makes sense. Teenage males will certainly think so, since it is filmed in the MTV style... bright color palette, fancy edits, blood... mesmerizing (and simple philosophies tend to go over well on that demographic). It actually starts our promisingly. I loved the first quarter. Only when it attempts to provide us with answers that are, sorry, far out of the grasp of the writers, does it fail. Film buffs may want to view it simply for one of the sloppiest, tacked-on "surprise" endings in cinema history. I could only laugh. Bottom line: Dumb.*
Risultato della classificazione neg: 59.63 % - pos: 40.36 %

Feature	Parole
A	*['just', 'beat', 'dumb', 'shackled', 'laugh', 'looking', 'male', 'sorry', 'shirt']*
B	*['palette', 'shoot', 'men', 'studio', 'gang', 'escape', 'try', 'start', 'black', 'fight']*
C	*['movie', 'quarter', 'tacked']*
D	*['sense', 'writer', 'fail', 'answer', 'violence', 'film', 'maybe', 'doe', 'attempt', 'demographic', 'license', 'surface', 'order', 'point', 'grasp', 'far', 'problem', 'edits', 'simply', 'individuality', 'job']*
E	*['cinema', 'life', 'philosophy', 'ordinary']*
F	*['blood', 'style', 'simple', 'modern', 'provides', 'surprise', 'provide', 'possession', 'mesmerizing', 'loved', 'fancy', 'view', 'filmed', 'spirit', 'buff', 'history', 'bright', 'color', 'tend', 'crushed', 'promisingly', 'beneath', 'ending', 'certainly']*

Tabella 8.100: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.403697 pos					
Feature influente positivamente		**Result**	**Delta**	**IR**	**nIRI**
F	blood style simple modern provides surprise provide possession mesmerizing loved fancy view filmed spirit buff history bright color tend crushed promisingly beneath ending certainly	0.238977	0.16472	1.68927	**0.273133**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
E	cinema life philosophy ordinary	0.413535	-0.00984	0.976208	**-0.0193**
C	movie quarter tacked	0.444378	-0.04068	0.908453	**-0.07556**
B	palette shoot men studio gang escape try start black fight	0.454744	-0.05105	0.887745	**-0.09323**
A	just beat dumb shackled laugh looking male sorry shirt	0.57987	-0.17617	0.696185	**-0.2731**
D	sense writer fail answer violence film maybe doe attempt demographic license surface order point grasp far problem edits simply individuality job	0.638818	-0.23512	0.631943	**-0.34238**

Tabella 8.101: *Influenze principali sulla predizione*

8.4.10 Get Out

In questo esempio la recensione negativa del film 'Get Out' è stata correttamente classificata dal modello.

First, let me say I loved Key and Peele's comedy sketches. They're refreshing and make light of racism in America. However, their movie is not as good as the reviews say it is. There isn't one good Caucasion in this movie. Every single one is racist, egocentric, deluded, deceitful, immoral, and the list goes on. Every single African-American are down to earth, authentic, humorous, and so on. During the movie, I was watching evil vs good or black vs white, which was clear as day. However, this is an anti-racism movie, isn't it? Then why are there so many stark contrasts? Abolishing racism isn't about creating two polar opposites, in fact, that is exactly what racism is! If you want to create something with any sense of seriousness, then make something that shows reality: show that doesn't matter what ethnicity you are, you are capable of good and bad. What was the point of this movie other than to show how horribly evil Caucasians are, oh an one Asian guy? Unfortunately, the comedy is lost within all this nuanced racist rhetoric. I'd rather describe this movie as a light torture-porn with subliminal racism. If the movie was described as that, then I could give it a 10/10. If this was based on reality then I could accept it as such but it's not(like portraying history). As Foucault and Morgan Freeman tried to express: it's time to get rid of the barriers and rather celebrate the unity. If you don't quite understand my point, imagine this movie but reverse the races. Imagine the hero is running away from African-American hypnotizing slavers because white people are, I dunno, in fashion? Wouldn't be so funny anymore, would it? Even if the context of the movie was satirically humorous, the theme alone would brand everyone involved a racist. If you like this movie, then good for you. However, see it as it really is - not an ant-racism movie. That, it certainly is not.
Risultato della classificazione neg: 71.74 % - pos: 28.25 %

Feature	Parole
A	*['hero', 'guy', 'slaver']*
B	*['refreshing', 'really', 'create', 'certainly', 'capable', 'clear', 'creating', 'portraying', 'let', 'light', 'running', 'exactly', 'fashion']*
C	*['reverse', 'subliminal', 'earth', 'polar', 'ant']*
D	*['brand']*
E	*['sense', 'wa', 'anymore', 'make', 'bad', 'understand', 'tried', 'isn', 'movie', 'review', 'oh', 'dunno', 'horribly', 'like']*
F	*['rid']*
G	*['sketch', 'comedy', 'funny']*
H	*['white', 'black', 'racist', 'ethnicity']*
I	*['good', 'evil']*
J	*['race']*
K	*['torture', 'don', 'egocentric']*
L	*['rhetoric', 'reality', 'deceitful', 'accept', 'people']*
M	*['humorous', 'involved', 'opposite', 'lost', 'list', 'based', 'away', 'matter', 'single', 'imagine', 'loved', 'described', 'time']*
N	*['porn']*
O	*['racism']*
P	*['deluded', 'contrast', 'point', 'seriousness', 'context', 'nuanced']*
Q	*['stark', 'authentic']*
R	*['satirically','theme','barrier','express','celebrate']*
S	*['immoral', 'fact', 'history', 'anti', 'unity']*

Tabella 8.102: *Elenco delle principali feature individuate nel testo*

Risultato del testo originale: 0.282547 pos					
Feature più influenti positivamente		**Result**	**Delta**	**IR**	**nIRI**
M	humorous involved opposite lost list based away matter single imagine loved described time	0.205729	0.076818	1.3734	**0.138996**
B	refreshing really create certainly capable clear creating portraying let light running exactly fashion	0.221372	0.061175	1.27634	**0.111909**
S	immoral fact history anti unity	0.223759	0.058788	1.26273	**0.107772**
Feature più influenti negativamente		**Result**	**Delta**	**IR**	**nIRI**
E	sense wa anymore make bad understand tried isn movie review oh dunno horribly like	0.478012	-0.19547	0.591087	**-0.30854**
O	racism	0.370623	-0.08808	0.762357	**-0.15446**
C	reverse subliminal earth polar ant	0.329571	-0.04702	0.857317	**-0.0869**

Tabella 8.103: *Influenze principali sulla predizione*

8.5 – Risultati sperimentali – 20-newsgroup

In quest'ultima sezione relativa ai risultati sperimentali si analizzeranno delle predizioni effettuate dal modello allenato con il dataset 20-newsgroup. Verranno proposti una serie di testi appartenenti al dataset 20-newsgroup, estrapolati prima della fase di training, e per ciascuno di essi si fornirà il relativo report di trasparenza sotto forma di tabella.

I documenti proposti sono 5 e appartengono ai newsgroup 'alt.atheism' e 'talk.politics.mideast':

- Alt.atheism A (classificato come alt.atheism)
- Alt.atheism B (classificato come talk.politics.mideast)
- Alt.atheism C (classificato come soc.religion.christian)
- Talk.politics.mideast A (classificato come talk.politics.mideast)
- Talk.politics.mideast B (classificato come talk.politics.guns)

8.5.1 alt.atheism A

Messaggio del newsgroup 'alt.atheism' classificato correttamente.

An Introduction to Atheism by mathew <mathew@mantis.co.uk>. This article attempts to provide a general introduction to atheism. Whilst I have tried to be as neutral as possible regarding contentious issues, you should always remember that this document represents only one viewpoint. I would encourage you to read widely and draw your own conclusions; some relevant books are listed in a companion article. To provide a sense of cohesion and progression, I have presented this article as an imaginary conversation between an atheist and a theist. All the questions asked by the imaginary theist are questions which have been cropped up repeatedly on alt.atheism since the newsgroup was created. [...]

Risultato della classificazione

alt.atheism: 87.56 %
comp.graphics: 0.0 %
comp.os.ms-windows.misc: 0.0 %
comp.sys.ibm.pc.hardware: 0.0 %
comp.sys.mac.hardware: 0.0 %
comp.windows.x: 0.0 %
misc.forsale: 0.0 %
rec.autos: 0.0 %
rec.motorcycles: 0.0 %
rec.sport.baseball: 0.0 %
rec.sport.hockey: 0.0 %
sci.crypt: 0.0 %
sci.electronics: 0.0 %
sci.med: 0.0 %
sci.space: 0.0 %
soc.religion.christian: 7.94 %
talk.politics.misc: 0.0 %
talk.politics.guns: 0.0 %
talk.politics.mideast: 0.0 %
talk.religion.misc: 4.48 %

Feature	Result	DI	IR	nIRI	IRP
Classe di interesse: alt.atheism - Risultato del testo originale: 87.5667					
alt discussion contentious	82.2661	5.3006	1.0644	**0.0960**	1.0388
atheism	53.2606	34.3061	1.6441	**0.4358**	1.1117
weak atheist agnosticism agnostic disbelieving	76.3746	11.1920	1.1465	**0.1842**	1.0716
Classe di interesse soc.religion.christian - Risultato del testo originale: 7.9433					
alt discussion contentious	13.2489	-5.3055	0.5995	**-0.1073**	0.5851
atheism	37.6596	-29.7163	0.2109	**-0.5953**	0.1426
weak atheist agnosticism agnostic disbelieving	16.0078	-8.0645	0.4962	**-0.1684**	0.4638
Classe di interesse talk.religion.misc - Risultato del testo originale: 4.4884					
alt discussion contentious	4.4830	0.0054	1.0012	**0.0001**	0.9771
atheism	9.0020	-4.5135	0.4986	**-0.1015**	0.3371
weak atheist agnosticism agnostic disbelieving	7.6118	-3.1233	0.5896	**-0.0666**	0.5511

Tabella 8.104: *Influenze principali sulla predizione*

Risulta particolarmente significativo il ruolo della feature composta dalla sola parola 'atheism'. La feature infatti presenta un buon impatto positivo sulla classe 'alt.atheism' (nIRI = 0,43) e negativo sulle classi 'soc.religion.christian' (nIRI = -0,59) e 'talk.religion.misc' (nIRI = -0,10).

8.5.2 alt.atheism B

Messaggio del newsgroup 'alt.atheism' classificato come 'talk.politics.mideast'.

In article <N4HY.93Apr5120934@harder.ccr-p.ida.org>, n4hy@harder.ccr-p.ida.org (Bob McGwier) writes:

|> [1] HOWEVER, I hate economic terrorism and political correctness worse than I hate this policy.

|> [2] A more effective approach is to stop donating to ANY organizating that directly or indirectly supports gay rights issues until they end the boycott on funding of scouts.

Can somebody reconcile the apparent contradiction between [1] and [2]?

Rob Strom, strom@watson.ibm.com, (914) 784-7641 IBM Research [...]

Risultato della classificazione

alt.atheism: 11.47 %
comp.graphics: 0.01 %
comp.os.ms-windows.misc: 0.0 %
comp.sys.ibm.pc.hardware: 0.0 %
comp.sys.mac.hardware: 0.0 %
comp.windows.x: 0.0 %
misc.forsale: 0.0 %
rec.autos: 0.0 %
rec.motorcycles: 0.0 %
rec.sport.baseball: 0.0 %
rec.sport.hockey: 0.0 %
sci.crypt: 0.85 %
sci.electronics: 0.0 %
sci.med: 0.21 %
sci.space: 0.04 %
soc.religion.christian: 0.8 %
talk.politics.misc: 4.59 %
talk.politics.guns: 18.79 %
talk.politics.mideast: 53.12 %
talk.religion.misc: 10.04 %

Feature	Result	DI	IR	nIRI	IRP
Classe di interesse: alt.atheism - Risultato del testo originale: 11.4781					
approach economic effective issue	20.3000	-8.8219	0.5654	**-0.1707**	0.4780
terrorism political	28.1318	-16.6537	0.4080	**-0.3225**	0.2819
apparent gay	30.2702	-18.7921	0.3792	**-0.3618**	0.2860
Classe di interesse: talk.politics.guns - Risultato del testo originale: 18.7985					
terrorism political	7.5156	11.2829	2.5013	**0.2466**	1.7279
approach economic effective issue	9.7383	9.0602	1.9304	**0.1816**	1.6319
Classe di interesse: talk.politics.mideast - Risultato del testo originale: 53.1209					
terrorism political	36.0397	17.0813	1.4740	**0.2688**	1.0182
approach economic effective issue	43.9100	9.2110	1.2098	**0.1579**	1.0227
funding	60.7210	-7.6000	0.8748	**-0.1330**	0.8397
Classe di interesse: talk.religion.misc - Risultato del testo originale: 10.0458					
watson ibm	6.4905	3.5554	1.5478	**0.0724**	1.4756
contradiction	6.7695	3.2763	1.4840	**0.0660**	1.3743
funding	10.1201	-0.0743	0.9927	**-0.0015**	0.9528
apparent gay	14.7153	-4.6695	0.6827	**-0.0911**	0.5150
approach economic effective issue	18.2938	-8.2480	0.5491	**-0.1635**	0.4642
terrorism political	20.8004	-10.7546	0.4830	**-0.2155**	0.3336

Tabella 8.105: *Influenze principali sulla predizione*

8.5.3 alt.atheism C

Messaggio del newsgroup 'alt.atheism' classificato come 'soc.religion.christian'.

>> Didn't you say Lucifer was created with a perfect nature?

>Yes.

Define perfect then.

>> I think you playing the usual game here, make sweeping statements like omni-, holy, or perfect, and don't note that they mean exactly what they say. that says that you must not use this terms when it leads to contradictions.

>I'm not trying to play games here. But I understand how it might seem that way especially when one is coming from a completely different point of view such as atheism [...]

Risultato della classificazione

alt.atheism: 19.74 %
comp.graphics: 0.0 %
comp.os.ms-windows.misc: 0.0 %
comp.sys.ibm.pc.hardware: 0.0 %
comp.sys.mac.hardware: 0.0 %
comp.windows.x: 0.0 %
misc.forsale: 0.0 %
rec.autos: 0.0 %
rec.motorcycles: 0.0 %
rec.sport.baseball: 0.0 %
rec.sport.hockey: 0.0 %
sci.crypt: 0.0 %
sci.electronics: 0.0 %
sci.med: 0.0 %
sci.space: 0.0 %
soc.religion.christian: 70.8 %
talk.politics.misc: 0.0 %
talk.politics.guns: 0.01 %
talk.politics.mideast: 0.08 %
talk.religion.misc: 9.33 %

Feature	Result	DI	IR	nIRI	IRP
Classe di interesse: alt.atheism - Risultato del testo originale: 19.7478					
religion atheism	13.4848	6.2629	1.4644	**0.1185**	1.4054
perfect truly knowing fall nature different presence known knowledge	25.9475	-6.1998	0.7611	**-0.1140**	0.7455
play playing game	26.5846	-6.8369	0.7428	**-0.1250**	0.7207
course clearly view ability exactly claim possibility imply fact argument true necessarily understand reason simply statement	33.6281	-13.8804	0.5872	**-0.2412**	0.5182
completely harm causing usual mouth said foot avoiding taking away	40.2833	-20.5355	0.4902	**-0.3419**	0.3976
evil happen	51.2620	-31.5142	0.3852	**-0.4844**	0.2462
allow especially alternative choice free choose greater self allowing doing	58.9924	-39.2446	0.3348	**-0.5659**	0.1721
Classe di interesse: soc.religion.christian - Risultato del testo originale: 70.8099					
allow especially alternative choice free choose greater self allowing doing	27.6649	43.1450	2.5596	**0.5600**	1.3157
evil happen	35.2424	35.5675	2.0092	**0.4714**	1.2839
completely harm causing usual mouth said foot avoiding taking away	47.0577	23.7522	1.5048	**0.3400**	1.2206
course clearly view ability exactly claim possibility imply fact argument true necessarily understand reason simply statement	52.6575	18.1524	1.3447	**0.2749**	1.1866
moral omniscient conscious	62.3283	8.4816	1.1361	**0.1460**	1.1014
play playing game	62.5142	8.2957	1.1327	**0.1433**	1.0990
perfect truly knowing fall nature different presence known knowledge	64.7629	6.0471	1.0934	**0.1083**	1.0711
religion atheism	78.9237	-8.1138	0.8972	**-0.1403**	0.8610

Tabella 8.106: *Influenze principali sulla predizione*

8.5.4 talk.politcs.mideast A

Messaggio del newsgroup 'talk.politics.mideast' classificato correttamente.

>> An Israeli soldier, stabbed in the neck, was lightly injured. Soldiers opened fire when a 19-year-old Gazan standing among visitors to the jail stabbed the soldier, who pushed the attacker away. Witnesses said up to eight soldiers fired at the man and he was hit repeatedly in the head.

> A soldier is attacked and his attacker is killed. What is the problem here? Would your opinion be any different if the soldiers wounds were more severe? How about if his attacker was only crippled for life, instead of being killed? I suspect that any Army would reprimand soldiers who shot to injure, rather than kill.

You are really sick, my friend. [...]

Risultato della classificazione

alt.atheism: 5.05 %
comp.graphics: 0.0 %
comp.os.ms-windows.misc: 0.0 %
comp.sys.ibm.pc.hardware: 0.0 %
comp.sys.mac.hardware: 0.0 %
comp.windows.x: 0.0 %
misc.forsale: 0.0 %
rec.autos: 0.0 %
rec.motorcycles: 0.0 %
rec.sport.baseball: 0.0 %
rec.sport.hockey: 0.0 %
sci.crypt: 0.0 %
sci.electronics: 0.0 %
sci.med: 0.0 %
sci.space: 0.0 %
soc.religion.christian: 0.03 %
talk.politics.misc: 4.79 %
talk.politics.guns: 17.74 %
talk.politics.mideast: 69.31 %
talk.religion.misc: 3.04 %

Feature	Result	DI	IR	nIRI	IRP
Classe di interesse: alt.atheism - Risultato del testo originale: 5.0529					
article discus statement merely example different talking wrong specific writes poster	1.6781	3.3748	3.0111	**0.1014**	2.6396
ethnic justifiable religion racist bigotry civil supporter	2.7270	2.3259	1.8529	**0.0527**	1.7265
Classe di interesse: talk.politics.guns - Risultato del testo originale: 10.4289					
article discus statement merely example different talking wrong specific writes poster	10.4289	7.3152	1.7014	**0.1434**	1.4915
ethnic justifiable religion racist bigotry civil supporter	11.7367	6.0074	1.5119	**0.1155**	1.4087
Classe di interesse: talk.politics.mideast - Risultato del testo originale: 69.3128					
article discus statement merely example different talking wrong specific writes poster	80.7188	-11.4060	0.8587	**-0.1875**	0.7528
ethnic justifiable religion racist bigotry civil supporter	79.8643	-10.5515	0.8679	**-0.1757**	0.8087
walking year age term man bound similar old	74.8648	-5.5519	0.9258	**-0.1002**	0.9089

Tabella 8.107: *Influenze principali sulla predizione*

8.5.5 talk.politcs.mideast B

Messaggio del newsgroup 'talk.politics.mideast' classificato come 'talk.politics.guns'.

>> How many Mutlus can dance on the head of a pin?

> That reminds me of the Armenian massacre of the Turks. Joel, I took out SCT, are we sure we want to invoke the name of he who greps for Mason Kibo's last name lest he include AFU in his daily rounds?

I dunno, Warren. Just the other day I heard a rumor that "Serdar Argic" (aka Hasan Mutlu and Ahmed Cosar and ZUMABOT) is not really a Turk at all, but in fact is an Armenian who is attempting to make any discussion of the massacres in Armenia of Turks so noise-laden as to make serious discussion impossible, thereby cloaking the historical record with a tremendous cloud of confusion.

Risultato della classificazione

alt.atheism: 5.67 %
comp.graphics: 0.0 %
comp.os.ms-windows.misc: 0.0 %
comp.sys.ibm.pc.hardware: 0.0 %
comp.sys.mac.hardware: 0.0 %
comp.windows.x: 0.0 %
misc.forsale: 0.0 %
rec.autos: 0.0 %
rec.motorcycles: 0.0 %
rec.sport.baseball: 0.0 %
rec.sport.hockey: 0.0 %
sci.crypt: 0.0 %
sci.electronics: 0.0 %
sci.med: 0.0 %
sci.space: 0.0 %
soc.religion.christian: 0.01 %
talk.politics.misc: 0.82 %
talk.politics.guns: 83.75 %
talk.politics.mideast: 9.29 %
talk.religion.misc: 0.41 %

Feature	Result	DI	IR	nIRI	IRP
Classe di interesse: alt.atheism - Risultato del testo originale: 5.6769					
israel nysernet	22.4703	-16.7934	0.2526	**-0.4142**	0.1607
Classe di interesse: talk.politics.guns - Risultato del testo originale: 83.7592					
israel nysernet	46.1085	37.6506	1.8166	**0.4712**	1.1557
cloaking jfurr laden historical massacre	66.1384	17.6208	1.2664	**0.2659**	1.0908
cloud	73.0643	10.6949	1.1464	**0.1776**	1.0762
noise	74.1986	9.5606	1.1289	**0.1615**	1.0721
Classe di interesse: talk.politics.mideast - Risultato del testo originale: 9.2941					
israel nysernet	27.8101	-18.5160	0.3342	**-0.3812**	0.2126
cloaking jfurr laden historical massacre	25.3038	-16.0097	0.3673	**-0.3310**	0.3164
cloud	18.4762	-9.1821	0.5030	**-0.1861**	0.4722
noise	17.2798	-7.9857	0.5379	**-0.1607**	0.5108
round record	16.4402	-7.1461	0.5653	**-0.1430**	0.5409

Tabella 8.108: *Influenze principali sulla predizione*

Capitolo 9

Conclusioni

L'obiettivo di questo capitolo è di ripercorrere quanto descritto nelle sezioni precedenti di questo elaborato mettendo in evidenza i passaggi salienti che hanno portato all'ottenimento di un framework che fornisce spiegazioni interpretabili alle predizioni di classificatori black-box. L'intuizione di estrarre un set limitato di feature da tipologie di dati non strutturati così da poter operare perturbazioni e misurare le variazioni nei risultati del modello si è rilevata funzionale allo scopo prefissato. Con la sola iterazione del task di classificazione sulla collezione di input perturbati relativi alle feature è possibile infatti ottenere una stima visiva ed efficace dei fattori che influiscono positivamente o negativamente sulla predizione, e da lì arrivare ad una vera e propria spiegazione della predizione.

Per quanto riguarda la metodologia sviluppata nel dominio delle immagini, il principale contributo innovativo proviene dall'uso dello strumento delle ipercolonne. Le feature estratte mediante il clustering di ipercolonne infatti rispecchiano il modo in cui la rete neurale convoluzionale percepisce le varie componenti di un'immagine. Per un osservatore umano, le feature risultano inoltre riconducibili facilmente alle caratteristiche semantiche di un'immagine.

Le analisi locali effettuate su alcune classi del modello VGG-16 si sono dimostrate soddisfacenti, permettendo di identificare una o più caratteristiche principali che il modello associa ad ogni classe. Infine, attraverso il confronto delle influenze positive e negative di immagini appartenenti alla stessa classe, si è verificato che la metodologia sviluppata tende a funzionare al meglio in presenza di risultati di predizione non troppo netti. Nel caso di predizioni estremamente sicure (quindi molto vicine a 1 o 0) risulta molto difficile individuare eventuali influenze sulla predizione e quindi fornire delle motivazioni al risultato del modello.

9.1 – Sviluppi futuri

La qualità delle spiegazioni ottenute dal framework sviluppato è direttamente proporzionale alla qualità delle feature individuate. Per eventuali sviluppi futuri di

questo lavoro di tesi sarebbe quindi opportuno affinare le tecniche di estrazione delle feature, in particolare per quanto riguarda il processo relativo ai documenti testuali, magari provando ad elaborare un approccio unificato per i due domini. Sarebbe inoltre interessante estendere la metodologia sviluppata anche su altre applicazioni di machine learning come *voice* o *face recognition*. Infine, sarebbe necessario continuare il lavoro svolto sulla definizione delle metriche puntando ad ottenere misure nuove e più raffinate, come, ad esempio, una misura che sappia valutare l'influenza congiunta di due più feature. In questo contesto, guardare alla disciplina della 'teoria dei giochi' sembrerebbe la scelta più naturale.

Bibliografia

[1] Tom M. Mitchell, *Machine Learning*, McGraw Hill, 1997

[2] it.wikipedia.org/wiki/Apprendimento_automatico

[3] deeplearning.net

[4] it.wikipedia.org/wiki/Apprendimento_profondo

[5] it.wikipedia.org/wiki/Rete_neurale_convoluzionale

[6] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, Nuria Oliver, *The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good*, 2016

[7] Anapum Datta, Shayak Sen, Yair Zick, *Algorithmic Transparency via Quantitative Input Influence*, 2016

[8] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, 2016

[9] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, Jitendra Malik, *Hypercolumns for Object Segmentation and Fine-grained Localization*, 2015

[10] blog.christianperone.com/2016/01/convolutional-hypercolumns-in-python

[11] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *An Introduction to Information Retrieval*, Cambrige University Press, 2008

[12] github.com/tensorflow/tensorflow

[13] keras.io

[14] Karen Simonyan, Andrew Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2015

[15] VGG16 for Keras, gist.github.com/baraldilorenzo/07d7802847aaad0a35d3

[16] Glove embeddings, nlp.stanford.edu/projects/glove/

[17] Large Movie Review Dataset, ai.stanford.edu/~amaas/data/sentiment/

[18] Dataset 20-newsgroup, qwone.com/~jason/20Newsgroups/