

# GenAI to explain CNN

Riccardo Maria Villaggio

## Introduzione

L'interpretabilità delle reti neurali convoluzionali (CNN) rappresenta una delle sfide più attuali e rilevanti nel campo del deep learning. In questo lavoro viene proposto un approccio che combina tecniche di AI generativa e strumenti di analisi semantica per esplorare e spiegare il funzionamento interno di una CNN.

In particolare, utilizziamo il modello generativo **Stable Diffusion 3.5 Large** per creare un dataset sintetico di immagini, a partire da lemmi selezionati tramite **Core WordNet**. Queste immagini vengono poi impiegate come input per la rete **ResNet-50**, di cui analizziamo le attivazioni interne tramite l'uso di *forward hook*. Questo consente di mettere in relazione i pattern di attivazione con concetti semanticamente ben definiti, rappresentati dai synset di WordNet.

A complemento di questa prima analisi, la vera sfida di questo lavoro è stata il dimostrare, e valutare, la presenza di bias nelle predizioni che la CNN produce. Per farlo, sono state scelte 5 classi particolarmente rappresentative in questo contesto, poiché particolarmente soggette a bias, grazie a esempi riportati su **Salient Imagenet**, e si è scelto di generare prompt strutturati in tre modalità distinte — “object-only”, “context-only” e “mixed”.

I risultati ottenuti mostrano che i layer più profondi della CNN sono in grado di discriminare efficacemente tra i concetti semanticamente proposti, confermando la capacità del modello di apprendere rappresentazioni complesse. L'analisi del bias, invece, rivela una significativa dipendenza della rete dal contesto visivo: la presenza di elementi di contesto nei prompt aumenta sensibilmente la confidenza delle predizioni.

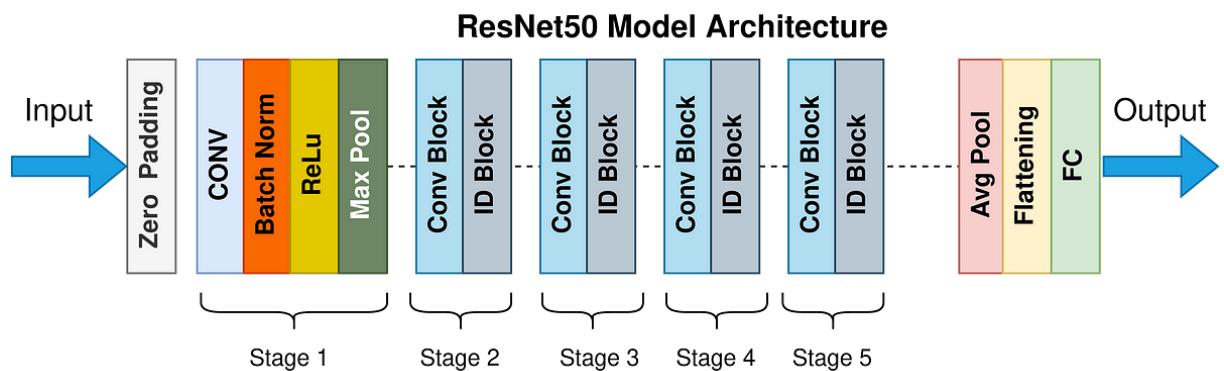


Figura 1: Architettura di *ResNet50*

## 1. Obiettivi e metodologia

### 1.1 Analisi interpretativa della CNN

Prima di affrontare il tema del bias, ci si è posti l’obiettivo di riuscire a esplorare in profondità il funzionamento interno della rete neurale convoluzionale ResNet-50, con l’obiettivo di comprendere come, e quanto, i suoi layer rappresentino e distinguano i concetti semanticici. A tal fine, viene adottata una metodologia che prevede la generazione di un dataset sintetico di immagini, costruito a partire da lemmi selezionati tramite Core WordNet e descritti tramite prompt testuali generici e pre-fabbricati. Queste immagini vengono poi utilizzate come input per ResNet-50, su cui vengono registrate le attivazioni interne dei principali layer tramite l’uso di forward hook. Per quantificare questi aspetti, sono state utilizzate le seguenti metriche

- **Attivazione massima per synset e layer:** per ogni layer della rete, viene calcolato il valore massimo di attivazione associato a ciascun concetto semantico, individuando così i neuroni più sensibili a determinati synset (su cosa si sono specializzati).
- **Correlazione neurone-synset:** per i neuroni più attivi di ciascun layer, vengono analizzate le statistiche delle attivazioni (media, deviazione standard, numero di campioni) rispetto ai diversi synset, sempre al fine di evidenziare eventuali specializzazioni.
- **Separabilità semantica:** viene misurata la distanza media tra i centroidi delle attivazioni dei synset nei vari layer, fornendo un indice quantitativo della capacità del layer di distinguere tra concetti diversi.

### 1.2 Analisi del bias

Concentrando adesso la nostra attenzione sul problema del bias all’interno delle CNN, quello che si vuole davvero cogliere è la bontà della loro capacità di generalizzare sulle classi da riconoscere: in particolare, si vuole valutare fino a che punto le predizioni di un modello, nel nostro caso **ResNet-50**, dipendano da elementi di contesto piuttosto che dall’oggetto di interesse stesso della classificazione. Partendo dai dati di **Salient ImageNet**, sono state scelte cinque classi — **bee, frog, plane, guitar, strawberry** — note per essere frequentemente associate a scenari riconoscibili (fiori per le api, stagni per le rane, cielo per gli aerei, palchi per le chitarre, piante per le fragole).

Si riportano per chiarezza, esempi di feature spurie per tali classi di bias selezionate, ossia elementi del contesto che la CNN “guarda” per aiutarsi nella classificazione, ma che non sono reali feature proprie del soggetto, quanto elementi di contorno:



Figura 2: Esempio di feature spurie per *bee*



Figura 3: Esempio di feature spurie per *guitar*



Figura 4: Esempio di feature spurie per *strawberry*

### 1.3 Design dei Prompt

Per ciascuna classe, sono stati generati 8 prompt utilizzando (elencati in Tabella 1) e suddivisi in tre categorie:

- **Object-only**: descrizioni focalizzate unicamente sull'oggetto (es. “*A professional close-up photograph of a bee with detailed anatomical features.*”).
- **Context-only (spurious)**: descrizioni del solo contesto tipico dell'oggetto, senza citare l'oggetto stesso (es. “*A vibrant garden full of colorful flowers.*”).
- **Mixed**: combinazioni di oggetto e contesto per riprodurre scene realistiche (es. “*A bee in a beautiful garden setting with colorful flowers.*”).

Così facendo, possiamo valutare quando, e se, la rete "imbroglio" guardando a elementi non necessari, ma spesso strettamente legati all'oggetto stesso.

Si vuole porre l'attenzione su un particolare non di poco conto, connesso a quanto poc'anzi detto: **ResNet-50**, così come altre CNN, è stata addestrata su **ImageNet**, e tali feature spurie su cui lei basa la sua confidence nella predizione è dunque da ricondurre alla correlazione soggetto-contesto che lei ha, seppur erroneamente, imparato durante la fase di training su questo specifico dataset di immagini. Allora il nostro scopo è proprio qui che si colloca, vedere fino a che punto

possiamo spingere la generalizzazione della rete e se davvero utilizza il contesto per boostare le sue predizioni.

Tabella 1: Prompt utilizzati per ciascuna classe di bias e tipologia di prompt.

<b>Classe</b>	<b>Object-only</b>	<b>Context-only</b>	<b>Mixed</b>
bee	<ul style="list-style-type: none"> <li>– A professional close-up photograph of a bee with detailed anatomical features and natural coloring</li> <li>– A bee on a clean white background showing intricate wing and body structure</li> <li>– A simple illustration of a bee isolated against a plain background</li> </ul>	<ul style="list-style-type: none"> <li>– A vibrant garden full of colorful flowers and greenery</li> <li>– A close-up of pollen-covered blossoms in a sunlit garden</li> <li>– A field of daisies and sunflowers under bright daylight</li> </ul>	<ul style="list-style-type: none"> <li>– A bee in a beautiful garden setting with colorful flowers and natural lighting</li> <li>– A bee collecting nectar from blooming flowers in a sunny meadow</li> </ul>
frog	<ul style="list-style-type: none"> <li>– A professional wildlife photograph of a frog with detailed skin texture and natural markings</li> <li>– A frog on a clean white background showing intricate skin patterns</li> <li>– A simple illustration of a frog isolated against a plain background</li> </ul>	<ul style="list-style-type: none"> <li>– A tranquil pond scene with aquatic plants and lily pads</li> <li>– A marshland with crystal clear water and lush vegetation</li> <li>– A wetland ecosystem with reeds and water lilies</li> </ul>	<ul style="list-style-type: none"> <li>– A frog sitting near a crystal clear pond with lush marsh vegetation</li> <li>– A frog resting on lily pads in a serene wetland environment</li> </ul>
plane	<ul style="list-style-type: none"> <li>– A professional close-up photograph of a plane with detailed engineering design</li> <li>– A plane on a clean white background showing aerodynamic features</li> <li>– A simple illustration of a plane isolated against a plain background</li> </ul>	<ul style="list-style-type: none"> <li>– A clear blue sky filled with puffy white clouds</li> <li>– An airport runway with control tower in the distance</li> <li>– A vast blue sky with scattered cumulus clouds at sunset</li> </ul>	<ul style="list-style-type: none"> <li>– A plane flying through clear blue sky with puffy white clouds in perfect weather conditions</li> <li>– A plane approaching the runway at a busy international airport</li> </ul>

Classe	Object-only	Context-only	Mixed
guitar	<ul style="list-style-type: none"> <li>- A professional product photograph of a guitar with detailed craftsmanship and wood grain texture</li> <li>- A guitar on a clean white background showing its shape and strings</li> <li>- A simple illustration of a guitar isolated against a plain background</li> </ul>	<ul style="list-style-type: none"> <li>- A concert stage with dramatic stage lights and microphone setup</li> <li>- A music venue with amplifiers and performance equipment</li> <li>- A recording studio with professional audio equipment and lights</li> </ul>	<ul style="list-style-type: none"> <li>- A guitar on a concert stage with dramatic stage lights and microphone setup for performance</li> <li>- A guitar in a recording studio with professional lighting and audio equipment</li> </ul>
strawberry	<ul style="list-style-type: none"> <li>- A professional macro photograph of a strawberry with detailed surface texture and natural red coloring</li> <li>- A strawberry on a clean white background showing its seeds and surface</li> <li>- A simple illustration of a strawberry isolated against a plain background</li> </ul>	<ul style="list-style-type: none"> <li>- A farm field full of lush green plants and ripe berries</li> <li>- An agricultural setting with rows of leafy plants in rich soil</li> <li>- A garden bed with green foliage and growing vegetation</li> </ul>	<ul style="list-style-type: none"> <li>- A strawberry growing on a healthy green plant with leaves in a farm field setting</li> <li>- A strawberry ripening on its plant surrounded by lush green foliage</li> </ul>

Ecco alcuni esempi delle 3 tipologie di immagini generate in base ai prompt adottati:



Figura 5: Esempio di immagine *object only*



Figura 6: Esempio di immagine *context only*



Figura 7: Esempio di immagine *mixed*

## 1.4 Fasi operative

Di seguito vengono descritte in dettaglio tutte le fasi operative che hanno caratterizzato l'intera analisi effettuata, articolata in due modalità principali: interpretabilità della CNN su classi generiche e analisi del bias su classi selezionate. Seppure con finalità diverse, entrambi i due aspetti sono perfettamente integrabili, come dimostrano i risultati ottenuti. Per maggiore chiarezza schematica, si faccia riferimento alla Figura 8.

1. **Selezione delle classi e generazione dei prompt:** Inizialmente vengono selezionate le classi di interesse. Per l'analisi interpretativa vengono scelti synset generici da Core WordNet, mentre per l'analisi del bias si utilizzano le cinque classi bias-aware da Salient ImageNet (**bee**, **frog**, **plane**, **guitar**, **strawberry**). Per ciascuna classe vengono generati prompt testuali: nella modalità interpretativa si usano prompt descrittivi e neutri, mentre nella modalità bias-aware si creano prompt strutturati in tre categorie (object-only, context-only, mixed) per isolare l'effetto del contesto.
2. **Generazione delle immagini:** Tutti i prompt vengono forniti a Stable Diffusion 3.5 Large, che produce le immagini coerenti con la descrizione testuale fornita.
3. **Analisi delle attivazioni:** Le immagini generate vengono processate dalla rete ResNet-50. Tramite l'utilizzo di forward hook (catturare gli output di ogni layer della rete) vengono estratte le attivazioni interne dei principali layer della CNN, permettendo di analizzare come la rete rappresenta i diversi concetti semanticici. Questo processo può essere applicato in entrambi gli ambiti di analisi di cui abbiamo discusso.
4. **Estrazione delle probabilità di classificazione:** Per ogni immagine viene registrata la probabilità assegnata da ResNet-50 alla classe target. Questo consente una valutazione quantitativa della capacità del modello di riconoscere sia le classi generiche, che le classi bias-aware, in funzione del tipo di prompt.
5. **Analisi quantitativa:** Le attivazioni e le probabilità ottenute vengono aggregate e confrontate. Per la modalità interpretativa si valutano metriche come l'attivazione massima, la separabilità semantica e la specializzazione dei layer. Per la modalità bias-aware, invece, si confrontano le probabilità ottenute per le tre categorie di prompt (object-only, context-only, mixed), quantificando l'impatto del contesto sulla predizione della rete e mettendo in evidenza eventuali bias semanticci per le singole classi.

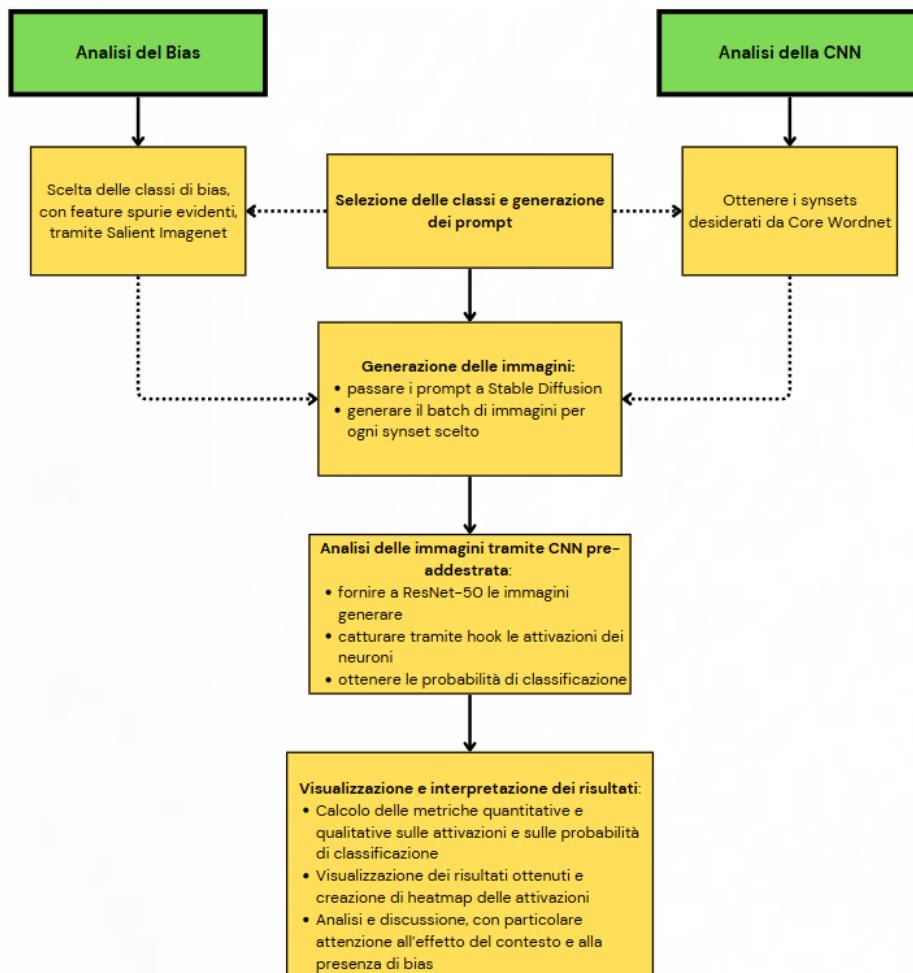


Figura 8: Schematizzazione delle fasi operative

## 2. Architettura del Codice

Il progetto è strutturato in cinque classi principali, ognuna con responsabilità ben definite:

1. **CoreWordNetManager**: si occupa del download e parsing del file Core WordNet, della selezione dei synset (distinguendo tra quelli bias-aware e non), della generazione dei prompt sia generici che mirati all'analisi del bias, e della gestione del lessico WordNet.
2. **StableDiffusionGenerator**: gestisce il caricamento del modello Stable Diffusion 3.5 Large e la generazione di batch di immagini a partire dai prompt.
3. **CNNActivationAnalyzer**: carica la rete ResNet-50, esegue il preprocessing delle immagini, estrae le attivazioni medie per layer, calcola metriche utili all'analisi e visualizza i pattern di attivazione.
4. **CNNInterpretabilityProject**: coordina l'intera pipeline di interpretabilità, collegando le attivazioni più significative ai concetti semanticci estratti da WordNet e producendo le metriche e visualizzazioni finali.
5. **ImageNetBiasAnalyzer**: gestisce le classi bias, effettua il mapping tra indice di ImageNet e label, identifica i prompt con maggiore probabilità predetta per ciascuna classe bias e produce le metriche specifiche per il report.

## 3. Funzionalità delle classi

### 3.1 Configurazione e Parametri (ProjectConfig)

Questa classe centralizza tutti i parametri di configurazione: percorsi delle cartelle, nomi dei modelli, dimensioni delle immagini, parametri di generazione e le classi di bias da analizzare.

### 3.2 CoreWordNetManager

- **Download**: la funzione `download_and_parse_from_url()` scarica e legge il file Core WordNet, estraendo un set di lemmi.
- **Caricamento synset**: `load_synsets_from_wordnet()` carica i synset specifici (ad esempio `dog.n.01`), escludendo le classi di bias.
- **Fallback**: se i synset trovati non sono sufficienti, viene utilizzata una lista di parole comuni (utile anche per debug).
- **Prompt CNN-pure**: genera prompt descrittivi utilizzando template predefiniti per ogni categoria semantica.
- **Prompt bias-aware**: genera gruppi di prompt per ciascuna classe bias (`bee`, `frog`, `plane`, `guitar`, `strawberry`), suddivisi in object-only, context-only e mixed.

### 3.3 StableDiffusionGenerator

- **Caricamento**: `load_model()` tenta prima il caricamento quantizzato a 4 bit, con fallback al modello standard se necessario.
- **Generazione**: `generate_image()` e `generate_batch()` generano immagini a partire dai prompt, salvano i risultati e ottimizzano la gestione della memoria GPU.

### 3.4 CNNActivationAnalyzer

- **Hook:** registra i forward hook sui layer principali di ResNet-50; questi hook catturano le attivazioni durante il forward pass.
- **Analisi:** calcola e stampa diverse metriche utili per l'interpretazione della rete, tra cui:
  - Attivazione massima per synset e layer: per ogni layer e synset, viene riportato il valore massimo di attivazione e il numero di immagini considerate.
  - Correlazioni neurone/synset: per ciascun layer, vengono mostrate le statistiche (media, deviazione standard, numero di campioni) delle attivazioni dei neuroni più attivi rispetto ai diversi synset.
  - Separabilità semantica: viene calcolata la distanza media tra i centroidi delle attivazioni dei synset nei vari layer, utile per valutare la capacità del layer di distinguere tra concetti diversi.

### 3.5 CNNInterpretabilityProject

Questa classe rappresenta il cuore del progetto e integra tutte le funzionalità principali:

- Gestione del mapping prompt-synset
- Analisi delle metriche prodotte tramite gli hook di ResNet
- Esecuzione centralizzata degli esperimenti, facilitando la riproducibilità
- Visualizzazione e salvataggio dei risultati

### 3.6 ImageNetBiasAnalyzer

La classe `ImageNetBiasAnalyzer` si occupa di:

- Selezionare e gestire le classi bias
- Eseguire la classificazione delle immagini generate per gruppi di prompt (object-only, context-only, mixed)
- Identificare e classificare i prompt in base alla probabilità predetta dalla CNN per ciascuna categoria
- Produrre e salvare le metriche e i report specifici per l'analisi del bias

## 4. Risultati ottenuti

Analizziamo adesso quanto ottenuto da un'esecuzione dove si integrano entrambe le due analisi.

### Separabilità tra classi nei layer profondi

La heatmap in Figura 9 mostra i valori di attivazione massima per ciascun synset sui migliori 8 layer con maggiore separabilità semantica, ovvero la capacità di distinguere tra i centroidi delle attivazioni delle diverse classi di synsets.

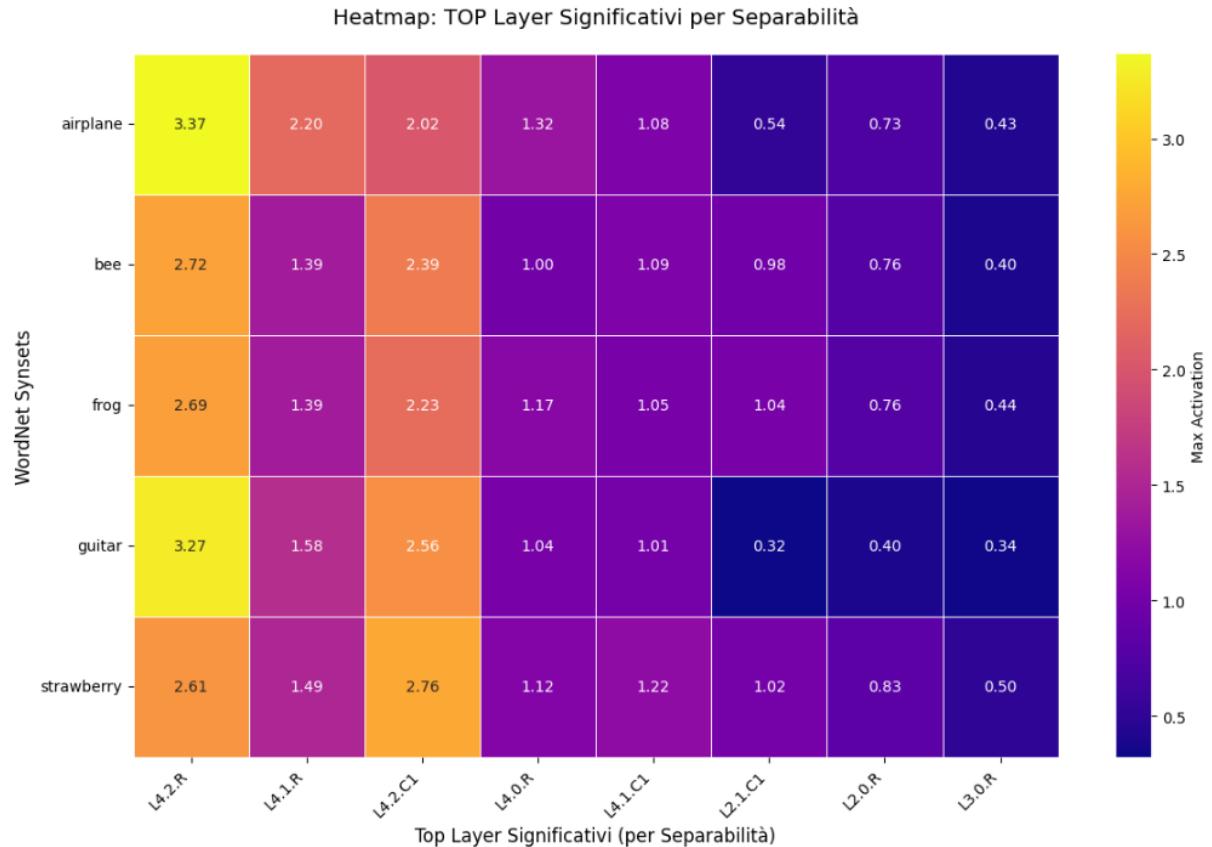


Figura 9: Heatmap per classi di bias sui layer più separabili

Osserviamo:

- **guitar** raggiunge 3.27 su **layer4.2.relu**.
- **airplane** ottiene 3.37 su **layer4.2.relu**, seguito da valori 2.20 e 2.02 per layer precedenti.
- **strawberry** si attesta 2.76 su **layer4.2.conv1** e 2.61 su **layer4.2.relu**.
- **bee** e **frog** presentano 2.72 e 2.69 nei layer top, ma scendono rapidamente.

## Risultati quantitativi

Analizzando i migliori 6 layer, si osserva che la separabilità media tra classi — calcolata come distanza media tra i centroidi delle attivazioni dei diversi synset — varia a seconda della profondità del layer: nei layer più profondi si raggiungono valori elevati (ad esempio, 16.828 in layer4.2.relu), a testimonianza della maggiore capacità di distinguere efficacemente i cinque synset testati. È importante sottolineare che la separabilità media globale (cioè tra tutti i centroidi dei synset) risulta sempre superiore rispetto alle attivazioni massime osservate per i singoli synset, poiché la metrica globale tiene conto di tutte le possibili coppie di classi. Ad esempio, la distanza media globale in layer4.2.relu (16.828) è nettamente superiore rispetto all'attivazione massima di synset come airplane o guitar nello stesso layer (rispettivamente 3.37 e 3.27).

**Tabella 1:** Separabilità media per layer (Top 6)

<b>Layer</b>	<b>Avg Centroid Distance</b>
layer4.2.relu	16.828
layer4.1.relu	7.331
layer4.2.conv1	4.443
layer4.0.relu	4.329
layer4.1.conv1	2.952
layer2.1.conv1	2.270

Chiaramente i valori mostrati sono coerenti a quanto mostrato precedentemente in Figura 9

**Tabella 2:** Top 5 synset per i 6 layer con maggiore separabilità

<b>Layer</b>	<b>Synset</b>	<b>Max Activation</b>
layer4.2.relu	airplane	3.3669
	guitar	3.2708
	bee	2.7210
	frog	2.6922
	strawberry	2.6121
layer4.1.relu	airplane	2.2011
	guitar	1.5823
	strawberry	1.4935
	bee	1.3906
	frog	1.3889
layer4.2.conv1	strawberry	2.7643
	guitar	2.5606
	bee	2.3852
	frog	2.2322
	airplane	2.0180
layer4.0.relu	airplane	1.3186
	frog	1.1688
	strawberry	1.1203
	guitar	1.0430
	bee	0.9994
layer4.1.conv1	strawberry	1.2223
	bee	1.0875
	airplane	1.0846
	frog	1.0466
	guitar	1.0144
layer2.1.conv1	frog	1.0400
	strawberry	1.0224
	bee	0.9789
	airplane	0.5362
	guitar	0.3222

## Confronto con una Esecuzione Precedente (100 prompt)

Un'esecuzione precedente è stata effettuata su un diverso sottoinsieme semantico, con 10 immagini per ciascuno di 10 synset scelti. Di seguito si riportano i risultati nel dettaglio, che mostrano per alcuni layer le classifiche di separabilità per i 3 synsets su cui risultano meglio specializzati (in base alla massima attivazione media):

- **layer4.1.relu:**
  - tree.n.01: 2.5172
  - boat.n.01: 2.3055
  - flower.n.01: 2.1233
- **layer4.0.relu:**
  - flower.n.01: 1.6804
  - boat.n.01: 1.6279
  - computer.n.01: 1.3162
- **layer3.3.relu:**
  - computer.n.01: 0.4060
  - dog.n.01: 0.3881
  - flower.n.01: 0.3645
- **layer2.2.bn2:**
  - computer.n.01: 0.3557
  - car.n.01: 0.3469
  - cat.n.01: 0.3369
- **layer2.1.bn3:**
  - tree.n.01: 0.2985
  - flower.n.01: 0.2932
  - bicycle.n.01: 0.2736
- **layer1.0.bn2:**
  - computer.n.01: 0.5834
  - dog.n.01: 0.5657
  - chair.n.01: 0.5599
- **layer4.0.bn1:**
  - dog.n.01: 0.1840
  - cat.n.01: 0.1813
  - flower.n.01: 0.1747
- **layer4.2.bn1:**
  - bicycle.n.01: 0.2196
  - tree.n.01: 0.2030
  - chair.n.01: 0.1647
- **layer3.5.conv3:**
  - dog.n.01: 0.1595
  - bird.n.01: 0.1476
  - cat.n.01: 0.1155
- **layer4.2.conv2:**
  - chair.n.01: 0.0890
  - cat.n.01: 0.0845
  - dog.n.01: 0.0830

**Osservazioni:** Anche in questa run si conferma il trend: i layer più profondi (`layer4.*`) mostrano valori di separabilità semantica più elevati, evidenziando una maggiore capacità di discriminare tra concetti anche in presenza di prompt più vari. Le classi maggiormente separabili differiscono leggermente, ma includono concetti diversi come *dog*, *tree* e *flower*.

### Esempi di attivazione per singoli neuroni

Per approfondire meglio il comportamento interno della rete, sono state anche analizzate le attivazioni medie dei neuroni in risposta ai prompt generati. Di seguito si mostrano tre esempi significativi tratti da `layer4.2.relu`, poiché come visto è il layer più di alto livello e performante per il riconoscimento delle classi, scelti per la loro capacità di evidenziare pattern distintivi rispetto ai diversi synset. Insieme alle medie, si riportano anche le deviazioni standard, che indicano la variabilità dell'attivazione sulle 8 immagini di ciascuna classe.

**Tabella 3:** Attivazioni medie e deviazioni standard per i neuroni in `layer4.2.relu`

Neurone	Synset	Mean	Std
624	bee	1.569	0.683
	frog	2.616	0.868
	airplane	1.152	1.171
	guitar	2.276	0.710
	strawberry	1.369	0.690
589	bee	1.194	0.607
	frog	1.028	0.606
	airplane	1.487	1.465
	guitar	3.220	1.660
	strawberry	0.518	0.258
617	bee	2.721	1.374
	frog	1.629	1.274
	airplane	0.684	0.278
	guitar	0.487	0.328
	strawberry	1.732	1.211

In questi esempi, si osserva come ogni neurone risponda in modo differenziato alle classi testate. Ad esempio, il neurone 624 è particolarmente attivo per `frog`, mentre il neurone 617 mostra una ampia variabilità su `bee`. Tali pattern indicano la presenza di neuroni specializzati nella rilevazione di caratteristiche visive proprie dei vari synset.

## 5. Analisi del Bias

I test bias-aware evidenziano una chiara dipendenza del modello dalla presenza esplicita di un contesto visivo all'interno dell'immagine. In particolare, l'aggiunta di elementi riconoscibili e centrali porta sistematicamente a un incremento della confidenza nella classificazione. L'analisi dei prompt generati mostra il seguente comportamento:

- I prompt **object-only**, in cui l'oggetto target è presentato senza contesto, generano predizioni corrette con buona confidenza. I valori variano da circa 0.61 (**airplane**) a oltre 0.99 (**strawberry**), confermando che il modello riesce a riconoscere la classe anche in assenza di indizi ambientali. Nonostante ciò, è possibile osservare anche l'esatto opposto: per la classe **bee**, **airplane** e **strawberry**, dove alcuni prompt non sono stati correttamente classificati.
- I prompt **context-only**, che rimuovono completamente l'oggetto lasciando solo ambienti semanticamente coerenti, vengono generalmente ignorati: la confidenza è sempre inferiore all'1%, anche in casi fortemente associativi (come stagni per **frog** o piste d'atterraggio per **airplane**). Questo suggerisce che il modello non è affatto da bias nei confronti del contesto.
- I prompt **mixed**, che includono sia l'oggetto che un contesto plausibile, mostrano un comportamento variegato. Per classi come **bee** e **strawberry**, il contesto rafforza la predizione e porta a confidenze prossime al 100%. Al contrario, in classi come **guitar**, l'aggiunta del contesto può talvolta ridurre la confidenza, probabilmente perché l'ambiente può essere fuorviante per la classificazione.

Le seguenti tabelle riportano, per ciascuna classe target, i valori di confidenza ottenuti dal modello sui diversi prompt, suddivisi per categoria:

Tabella 2: Classe: **bee**

Prompt	Confidence
<i>Mixed</i>	
A bee in a beautiful garden setting with colorful flowers and lush greenery.	0.9950
A bee collecting nectar from blooming flowers in a sunny meadow.	0.9805
<i>Object-only</i>	
A bee on a clean white background showing intricate wing and body detail.	0.9735
A professional close-up photograph of a bee with detailed anatomy.	0.1645
A simple illustration of a bee isolated against a plain background.	0.0097
<i>Context-only</i>	
A close-up of pollen-covered blossoms in a sunlit garden.	0.2678
A field of daisies and sunflowers under bright daylight.	0.0047
A vibrant garden full of colorful flowers and greenery.	0.0044

Tabella 3: Classe: **frog**

Prompt	Confidence
<i>Mixed</i>	
A frog resting on lily pads in a serene wetland environment.	0.9020
A frog sitting near a crystal clear pond with lush marsh vegetation.	0.0698
<i>Object-only</i>	
A frog on a clean white background showing intricate skin patterns.	0.9348
A simple illustration of a frog isolated against a plain background.	0.6945
A professional wildlife photograph of a frog with detailed skin texture.	0.5284
<i>Context-only</i>	
A garden bed with green foliage and growing vegetation.	0.0005
A close-up of pollen-covered blossoms in a sunlit garden.	0.0005
A tranquil pond scene with aquatic plants and lily pads.	0.0004

Tabella 4: Classe: **airplane**

Prompt	Confidence
<i>Mixed</i>	
A airplane flying through clear blue sky with puffy white clouds.	0.7531
A airplane approaching the runway at a busy international airport.	0.9531
<i>Object-only</i>	
A airplane on a clean white background showing aerodynamic features.	0.6493
A professional close-up photograph of a airplane with detailed fuselage.	0.4126
A simple illustration of a airplane isolated against a plain background.	0.7717
<i>Context-only</i>	
An airport runway with control tower in the distance.	0.0555
A clear blue sky filled with puffy white clouds.	0.0005
A vast blue sky with scattered cumulus clouds at sunset.	0.0001

Tabella 5: Classe: **guitar**

Prompt	Confidence
<i>Mixed</i>	
A guitar in a recording studio with professional lighting and equipment.	0.9498
A guitar on a concert stage with dramatic stage lights and microphones.	0.0040
<i>Object-only</i>	
A simple illustration of a guitar isolated against a plain background.	0.9789
A guitar on a clean white background showing its shape and strings.	0.9227
A professional product photograph of a guitar with detailed wood grain.	0.0004
<i>Context-only</i>	
A music venue with amplifiers and performance equipment.	0.0090
A recording studio with professional audio equipment and lighting.	0.0030
A concert stage with dramatic stage lights and microphone setup.	0.0143

Tabella 6: Classe: **strawberry**

Prompt	Confidence
<i>Mixed</i>	
A strawberry ripening on its plant surrounded by lush green leaves.	0.9998
A strawberry growing on a healthy green plant with leaves in a garden.	0.9979
<i>Object-only</i>	
A strawberry on a clean white background showing its seeds and texture.	0.9988
A professional macro photograph of a strawberry with detailed surface.	0.9856
A simple illustration of a strawberry isolated against a plain background.	0.3255
<i>Context-only</i>	
A close-up of pollen-covered blossoms in a sunlit garden.	0.0072
A garden bed with green foliage and growing vegetation.	0.0007
A tranquil pond scene with aquatic plants and lily pads.	0.0006

Infine, si osserva che nella classifica dei top prompt per ciascuna classe, può capitare che i prompt context-only presenti siano in realtà prompt pensati per altre classi, ma, nonostante ciò, attivano maggiormente la classe target rispetto a quelli specificatamente progettati per essa. Alla luce di quanto esposto, questo fenomeno conferma che la CNN tende a basare la classificazione anche (e, in alcuni casi, soprattutto) sul contesto su cui è stata allenata a riconoscere la classe target.

## 5.1 Analisi per classe

Adesso vediamo, per ogni classe, cosa ricaviamo dai risultati sopra esposti; si faccia sempre riferimento alle tabelle riportate in precedenza.

**Bee:** Le immagini prodotte permettono di concludere che la classe bee è sicuramente affetta da elevato bias: i prompt mixed sono praticamente prossimi al 100% di confidence, anche se un particolare caso di object-only è stato correttamente classificato:



Figura 10: A bee on a clean white background showing intricate wing and body detail.

Risultano invece non riconosciuti gli altri due casi di prompt object-only:



Figura 11: A professional close-up photograph of a bee with detailed anatomy.



Figura 12: A simple illustration of a bee isolated against a plain background.

Il bias sulla classe è chiaramente visibile se si relazionano le confidence di questi due casi sopra mostrati con questa immagine context-only, la cui confidenza risulta circa 27%. Ciò che induce la rete a riconoscere l'ape, infine, sono proprio i fiori in primo piano, dato che immagini ritraenti giardini più vasti hanno confidence pressoché nulle.



Figura 13: A close-up of pollen-covered blossoms in a sunlit garden.

**Frog:** La classe frog, come per bee, riceve un boost nella confidence nel caso in cui siano presenti in primo piano foglie di ninfea riconoscibili. Infatti, nel caso in cui siano assenti e la rana sia circondata da uno stagno generico, la confidence scende al 6%: è chiaramente un esempio di bias.

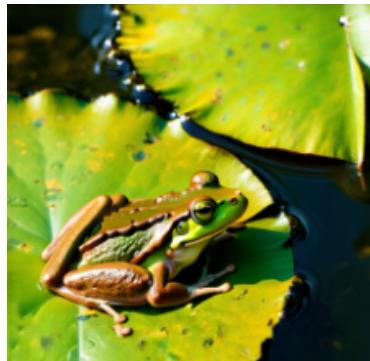


Figura 14: A frog resting on lily pads in a serene wetland environment.



Figura 15: A frog sitting near a crystal clear pond with lush marsh vegetation.

Nonostante ciò, a differenza di bee, i prompt object only hanno avuto ottimi risultati, dimostrando che isolando il soggetto la confidence rimane comunque accettabile e quindi potendo concludere che il contesto, al netto delle foglie di ninfea, confonde il modello invece che aiutarlo.

**Plane:** La classe plane risulta estremamente equilibrata, con medie percentuali ottime sia nel caso mixed che object only. Unico caso degno di nota è il prompt di una foto close up sulla parte frontale dell'aereo, dove la confidence è di circa il 41%. La rete probabilmente guarda a più dettagli per poter essere certa che si tratti di un aereo.



Figura 16: A professional close-up photograph of an airplane with detailed fuselage.

**Guitar:** Precisiamo che la label di ImageNet che stiamo analizzando riguarda le chitarre elettriche, e questo dettaglio ci servirà tenerlo a mente per un'analisi che faremo a breve. Le chitarre sono frequentemente associate a palchi, microfoni e luci da concerto. Anche in questo caso, la confidenza della rete cresce nei prompt mixed e context-only, anche se ci sono due casi particolari. Il primo è la seguente immagine mixed, che ha ottenuto percentuale pressoché nulla sicuramente perché il modello si confonde su quale dettaglio analizzare e tenere in considerazione per la classificazione.



Figura 17: A guitar on a concert stage with dramatic stage lights and microphones.

Il secondo è questa immagine object-only, che ha ottenuto anche in questo caso percentuale nulla. Ciò è dovuto sicuramente al fatto che ritrae una chitarra acustica e non elettriche (nostro caso d'interesse), quindi si tratta di un ottimo risultato per la rete, che non confonde le due classi anche rimuovendo il contesto,



Figura 18: A professional product photograph of a guitar with detailed wood grain.

**Strawberry:** Le fragole sono riconosciute con maggiore probabilità quando sono rappresentate su piante o in ambienti agricoli. La classificazione è stata praticamente perfetta, al netto del prompt riportato sotto, dimostrando che la classe presenta un bias molto basso.



Figura 19: A simple illustration of a strawberry isolated against a plain background.

## 6. Conclusioni

L'analisi delle attivazioni interne ha evidenziato, correttamente, come i layer più profondi della rete siano in grado di specializzarsi nella distinzione tra concetti semanticici anche molto diversi tra loro, confermando la capacità del modello di apprendere rappresentazioni complesse e strutturate.

Tuttavia, l'indagine mirata sulle classi di bias ha messo in luce una significativa dipendenza del modello dal contesto visivo: la presenza di elementi ambientali tipici, spesso appresi durante la fase di addestramento su ImageNet, può incrementare sensibilmente la confidenza della rete nella classificazione, anche in assenza dell'oggetto target: dunque, per questo motivo, il lavoro di ricerca effettuato è stato un successo.

Questi risultati sottolineano come le CNN, pur mostrando ottime capacità di generalizzazione, siano ancora vulnerabili a correlazioni spurie e bias semanticici introdotti dai dataset di training. La metodologia adottata in questo progetto si è rivelata efficace non solo per spiegare il funzionamento interno della rete, ma anche per individuare e quantificare tali bias, offrendo così uno strumento utile sia per la ricerca sull'interpretabilità che per lo sviluppo di modelli più robusti e affidabili.

## Riferimenti

- Salient ImageNet Explorer – University of Maryland:  
<https://salient-imagenet.cs.umd.edu/explore.html>
- Stable Diffusion 3.5 Large – Hugging Face (Stability AI):  
<https://huggingface.co/stabilityai/stable-diffusion-3.5-large>
- Core WordNet Synsets (Princeton University):  
<https://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt>
- “Core-ImageNet: Measuring Bias in Vision Models Using Core WordNet”, 2024 – arXiv preprint:  
<https://arxiv.org/html/2412.13079v1>