UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia

Corso di laurea in Scienze Psicologiche Cognitive e Psicobiologiche

Elaborato finale

Il riconoscimento visivo di oggetti nell'uomo e nelle reti neurali profonde

Visual object recognition in humans and in deep neural networks

Relatore
Prof. Zorzi Marco

Laureando: Marinello Matteo
Matricola:1222712

Anno Accademico 2022/2023

Indice

Abstract

La percezione visiva è una funzione mentale specifica che si basa sulla capacità di riuscire a discriminare forme, dimensioni, e altri stimoli visivi. Questa funzione è alla base del riconoscimento visivo d'oggetti che si basa sulla classificazione o denominazione di un oggetto dopo l'osservazione.
L'abilità di riconoscimento non appartenente unicamente all'uomo e agli animali ma è anche implementabile all'interno delle reti neurali profonde (DNN), la cui struttura è ispirata al funzionamento delle reti neuronali del cervello, le quali nei compiti di riconoscimento dimostrano di avere delle prestazioni equiparabili, se non superiori a quelle dell'uomo.
Ciò è stato permesso in particolar modo grazie all'utilizzo delle reti convoluzionali (CNN) che sfruttano dei campi recettivi locali per scansionare l'immagine presentatagli.

1. RICONOSCIMENTO VISIVO DI OGGETTI NELL'UOMO

La Percezione Visiva è "La Funzione mentale implicata nel distinguere forma, dimensione, colore e altri stimoli oculari"(Stucki et al., 2008) mentre il "riconoscimento d'oggetti" è la capacità dell'uomo di riuscire a riconoscere un oggetto il più rapidamente possibile (DiCarlo et al., 2012).

Questa abilità se pur apparentemente semplice, siccome intrinseca in molte attività quotidiane (es. lettura, selezione di uno strumento, …), nasconde la sua grandezza computazionale; infatti, vi è una continua rilevazione e classificazione di stimoli, (Biederman, 1987) in pochi istanti senza alcuno sforzo.

La complessità di tale abilità è confermata anche dalla visione evoluzionistica, in quanto nei primati non umani la metà della loro neocorteccia è implicata nell'elaborazione visiva (Felleman & Van Essen, 1991) essenziale per la sopravvivenza, che dipende dalla velocità e l'accuratezza con la quale riescono a determinare la natura dell'oggetto. Tale capacità però non è unicamente relegata a un'elaborazione visiva, ma presenta una natura multisensoriale (Amedi et al., 2005) perché le informazioni vengono elaborate in parallelo e successivamente integrate per fornire così una percezione coerente, standardizzandola a stimoli più familiare.

1.1. STRUTTURE NEURALI ALLA BASE DEL RICONOSCIMENTO

Studi su primati hanno mostrato un ruolo centrale, ai fini del riconoscimento d'oggetti, è svolto dal flusso di elaborazione visiva ventrale (Figura 1), costituito dalle vie visive e dai lobi temporali e occipitali (Gross, 1994), evidenziato anche da studi neuropsicologici, in quanto casi di soggetti o primati con compromissioni di queste aree presentano: cecità completa, se ad essere danneggiato è il flusso ventrale posteriore (Stoerig & Cowey, 1997), oppure vi sarà una scarsa o assente capacità di categorizzare gli oggetti se ad essere lesionate sono aree occipitali o temporali (Horel, 1996).

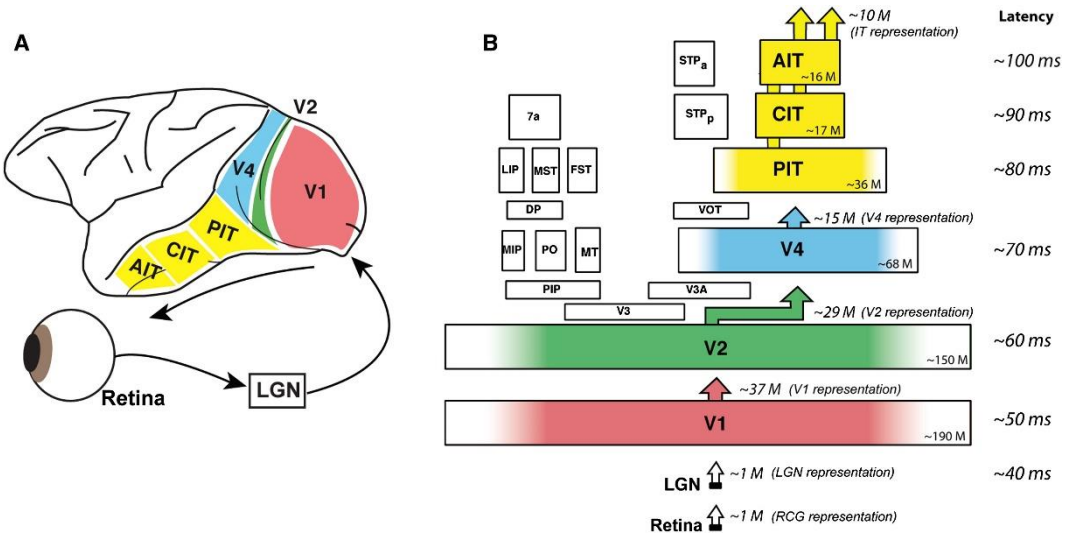

Figura 1 Il percorso visivo ventrale (DiCarlo et al., 2012)

(A) vengono mostrate le posizioni dell'area corticale del flusso ventrale nel cervello della scimmia macaco e il flusso di informazioni visive dalla retina.

(B) Le dimensioni di ogni area sono proporzionali alla sua superficie corticale(Felleman & Van Essen, 1991). Il numero totale di neuroni è mostrato nell'angolo di ciascuna area (M=milioni). La dimensionalità di ciascuna rappresentazione è mostrata sopra ogni area, in base alla densità neuronale(Collins et al., 2010), alla frazione neuronale dello strato 2/3 (O'Kusky & Colonnier, 1982) e alla porzione dedicato all'elaborazione dei 10 gradi centrali del campo visivo(Brewer et al., 2002).

Il ruolo del lobo temporale è stato dimostrato da studi sui roditori (Balderas et al., 2008) alla quale viene iniettata l'anisomicina, un inibitore della sintesi proteica; dove a livello delle cortecce perinali e insulari si verifica un'incapacità di riconoscimento di oggetti nel lungo periodo (24h), ma non nel breve termine.

Mentre a livello ippocampale porta a un'assenza di memoria di riconoscimento basato sul contesto a lungo termine.

Tali evidenze denotano la presenza di due funzioni distinte delle aree temporali, una di consolidamento di oggetti familiari ad opera dalle cortecce perinali e insulari e un'altra di consolidamento delle informazioni contestuali fatto dall'ippocampo.

Nel lobo occipitale sono presenti le cortecce visive (V1, V2, V3, V4, V5). Nella corteccia visiva primaria (V1) sono presenti neuroni con un proprio campo recettivo stimolo-specifico, mostrando come ogni neurone tenda ad avere un'attivazione maggiore nel momento in cui lo stimolo presentato è simile allo stimolo per la quale si sono specializzati; in particolare in questo livello si elaborano stimoli semplici e ripetitivi.

La corteccia visiva secondaria (V2) riceve connessioni dalla V1 e le invia a V3, V4, V5 ma al contempo presenta anche connessioni di feedback con V1. Quest'area ha la capacità di elaborare stimoli più complessi come la discriminazione tra l'oggetto e il contesto e la formazione di contorni illusori.

La corteccia visiva terziaria (V3) presenta oltre a connessioni con V2 delle deboli connessioni con V1 ed è deputato al riconoscere gli oggetti in movimento. Le ultime due cortecce invece sono adibite alla codifica dei colori (V4) e all'elaborazione dei movimenti (V5) (Courtney & Ungerleider, 1997).

Oltre alle aree visive nel lobo occipitale è presente il complesso occipitale laterale (LOC) (Malach et al., 1995), che come evidenziato dalle tecniche di neuroimaging, è particolarmente attivato nel momento in cui si osserva un oggetto di uso quotidiano rispetto a oggetti la cui natura non è definita e vi è richiesta un'interpretazione per identificarlo, da sottolineare come l'area

tenda a rispondere anche ad oggetti non familiari sulla quale è necessario eseguire delle interpretazioni.

1.2. COME AVVIENE IL RICONOSCIMENTO?

Una prima modalità di spiegazione sul come i neuroni cooperino per il riconoscimento di oggetti è stata proposta da Hubel e Visel nel 1962, i quali per la prima volta, con studi sui gatti, riuscirono a identificare un'organizzazione gerarchica.

Identificarono due tipi di cellule, le prime definite cellule semplici le quali possedevano dei piccoli campi recettivi ed erano fortemente dipendenti dalla fase, cioè con distinti sottocampi inibitori e eccitatori, mentre le seconde furono definite le cellule complesse, in quanto possedevano dei campi recettivi ampi, i quali ricevevano informazioni dalle cellule semplici, e non erano dipendenti dalla fase (Hubel & Wiesel, 1962).

Un modello più aggiornato, basato sulla vista, è stato proposto da Riesenhuber & Poggio nel 2000 (Figura 2).

Alla base sono presenti le cellule comprese tra la corteccia visiva primaria (V1) e corteccia infero temporale posteriore (PIT) dove gli oggetti sono elaborati secondo una gerarchia di frammenti (Ullman, 2007), i quali sono stati appresi mediante la semplice osservazione, perché elementi distintivi dell'oggetti e forniscono un'elevata quantità di informazioni su di esso.

Nella seconda fase le cellule attivate si trovano nella corteccia infero temporale anteriore dove vi è una manipolazione delle informazioni ricevute, sfruttando rotazioni e modificazioni dell'illuminazione usufruendo l'invarianza delle trasformazioni, ottenendo in questo modo una rappresentazione più appropriata dell'oggetto. In fine come ultima fase vi è la categorizzazione o l'identificazione dell'oggetto, che può portare a una risposta motoria.

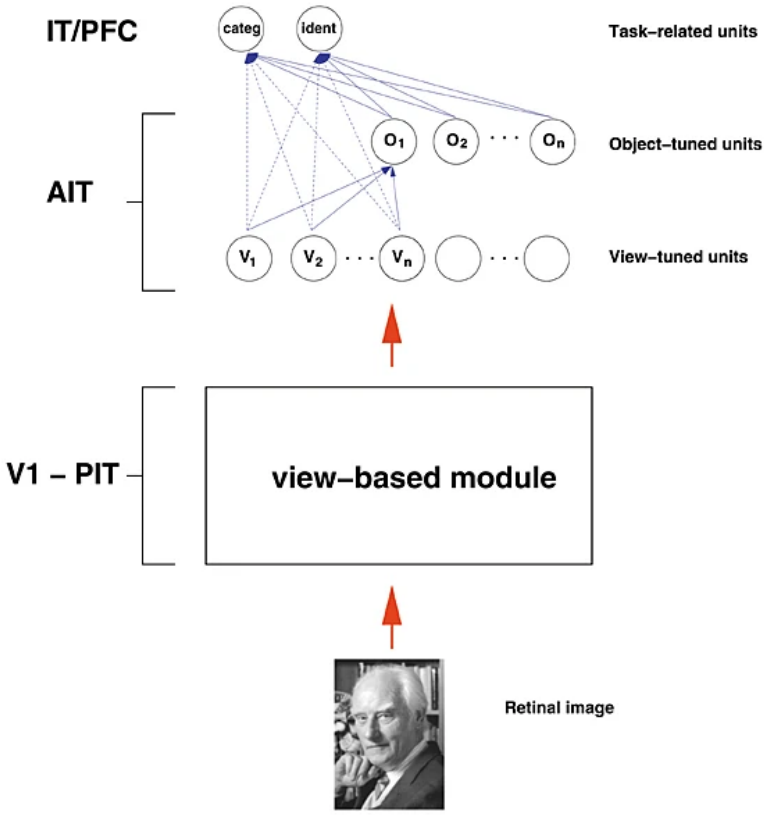

Figura 4 Modello gerarchico del riconoscimento oggetti (Riesenhuber & Poggio, 2000)

Nella parte superiore di un modello basato sulla vista, le unità ottimizzate per la vista (Vn) mostrano una stretta ottimizzazione della rotazione in profondità (e dell'illuminazione e di altre trasformazioni dipendenti dall'oggetto) ma sono tolleranti al ridimensionamento e alla traslazione della vista dell'oggetto preferita. Si noti che le celle qui etichettate come unità ottimizzate per la visualizzazione potrebbero essere ottimizzate per visualizzazioni complete o parziali, ovvero collegate solo ad alcune delle unità di funzionalità attivate dalla visualizzazione dell'oggetto. Tutte le unità nel modello rappresentano singole cellule modellate come neuroni semplificati con sinapsi modificabili. L'invarianza, ad esempio alla rotazione in profondità, può quindi essere aumentata combinando in un modulo di apprendimento diverse unità che sono sintonizzate su diverse viste dello stesso oggetto, creando unità invarianti vista-oggetto (On). Queste, così come le unità organizzate dalla vista, possono quindi servire come input per moduli di attività che eseguono compiti visivi come l'identificazione/discriminazione di oggetti o la categorizzazione (Riesenhuber & Poggio, 1999).

2. RETI NEURALI PROFONDE

Le reti neurali profonde sono sistemi informatici di apprendimento automatico, la cui funzione è quella di percepire le informazioni ambientali producendo di conseguenza delle azioni propense a massimizzare la probabilità di ottenere un successo.

Il loro nome deriva dalla loro architettura, in quanto emula l'organizzazione dei neuroni all'interno del cervello, in quanto al loro interno sono presenti neuroni artificiali interconnessi tra loro.

Oltre alla presenza dei neuroni e delle connessioni vi è un'organizzazioni in strati o gruppi:

- Strato di Input: strato il quale riceve le informazioni provenienti dall'ambiente.
- Stato di Output: strato che produce risposte dopo un'elaborazione delle informazioni avvenato negli strati precedenti.
- Strato nascosto: strato connesso con l'input il quale ha il compito di elaborare le informazioni ambientali per successivamente fornire la risposta più opportuna per ottenere un output corretto. Tale strato può presentarsi, a differenza degli strati di input e output, ripetuto e quindi esserci due o più strati nascosti all'interno della rete e ciò permette alla rete di prendere il nome di rete profonda (deep network), in quanto si aumenta la complessità dei dati che possono essere elaborati.

Le intercessioni tra i vari strati della rete ne determinano la tipo:

- Reti feed-forward: la rete presenta connessioni unidirezionali tra gli strati con l'assenza di connessioni interstrato (Bottom-up) nella quale possono essere presenti anche strati nascosti.
- Reti ricorrenti: la rete presenta delle connessioni bidirezionali (Top-down), capaci di inviare messaggi di feedback allo strato precedente, ma sono assenti connessioni intra-strato.
- Reti Interamente ricorrenti: le quali presentano connessioni bidirezionali e intra-strato.

Le reti artificiali però, come nell'uomo, prima di poter eseguire un qualsiasi compito richiedono una fase di addestramento che permette di apprendere delle rappresentazioni, garantendogli in questo modo di riuscire a eseguire compiti di riconoscimento partendo da una serie di dati grezzi (LeCun et al., 2015).

Al momento attuale, con l'utilizzo massiccio delle deep networks, i metodi di deep learning sono i più usati, perché metodi di rappresentazioni apprendimento, in quanto le rappresentazioni apprese della rete sono distribuite tra gli strati nascosti.

Ad esempio, quando si presenta un'immagine come array, valori di pixel, si osserva:

- Primo livello: si estrae la presenza o l'assenza di bordi all'interno dell'immagine e il loro orientamento.
- Secondo livello: rileva orientamenti particolarmente informativi e in che modo sono disposti i bordi nell'immagine.
- Terzo strato: combina i bordi identificati in modo da ottenere rappresentazioni più complesse
- Strati successivi: partendo dalle 'informazioni provenienti dallo stato precedente produrranno una rappresentazione volta a massimizzare le probabilità di ottenere un output corretto.

L'elemento vincente di tale metodo è che l'estrazione delle caratteristiche (featurs) non arbitrario, non è programmato da un operatore esterno, ma è la stessa rete ad apprenderlo mediante un apprendimento di tipo gerarchico.

La forma più semplice di deep learning è l'apprendimento supervisionato, nella quale non è possibile mostrare una semplice immagine grezza alla rete ma si necessita che l'immagine si etichettata.

Durante l'addestramento si mostra un'immagine alla rete, la quale produce degli output di vettori numerici, uno per ogni categoria, selezionando quello con valore maggiore.

Nei primi tentativi si nota come l'output fornito e il target (output desiderato) non coincidano, in quanto molto probabilmente il valore della categoria

target non sarà il maggiore. Viene in questo modo calcolato una funzione di errore che misura la distanza tra l'output ottenuto e il target, modificando in questo modo i parametri, detti pesi delle connessioni, con l'obiettivo minimizzare l'errore.

Ai fini di una corretta modifica dei pesi si calcola un vettore del gradiente, su ogni peso, in modo da indicare come varierebbe l'errore se il peso aumentasse di una piccola quantità, causando così una regolarizzazione del peso in direzione opposta al vettore gradiente.

Al momento attuale però non si calcola la funzione d'errore per ogni singola esposizione alla rete, ma si usa tecnica definita discesa stocastica del gradiente, in quanto si calcola la funzione d'errore dopo un numero predefinito di esposizioni ottenendo così un gradiente medio che porterà di conseguenza una regolarizzazione dei pesi delle connessioni, riuscendo in questo modo a ottenere una buona serie di pesi rapidamente (Bousquet, O., et al., 2007).

Dopo la fase di addestramento vi è una fase di testing della rete, nella quale viene mostrato un set di immagini mai mostratogli prima, in modo da comprendere le capacità di generalizzazione, osservando se fornisce degli output sensati nonostante non gli si presenti gli input di addestramento(LeCun et al., 2015).

2.1. IL RICONOSCIMENTO OGGETTI NELLE DNN

Reti convoluzionali (CNN)

Le reti convoluzionali sono delle reti profonde nella quale è presente almeno uno strato convoluzionale, uno strato non interamente connesso con lo strato precedente.

Nello strato convoluzionale i neuroni per estrarre le caratteristiche dall'immagine dallo strato precedente utilizzano dei campi recettivi locali, uno per ogni neurone, che tende a risponde featurs specifica (detta kernel o filtro); quindi l'immagine, presentata alla rete, come una matrice, viene scansionata interamente da ogni neurone che, come risultato, produrrà l'immagine di partenza ma nella quale sarà evidenziata un'unica caratteristica, ottenedo così una featurs map (Figura 3)

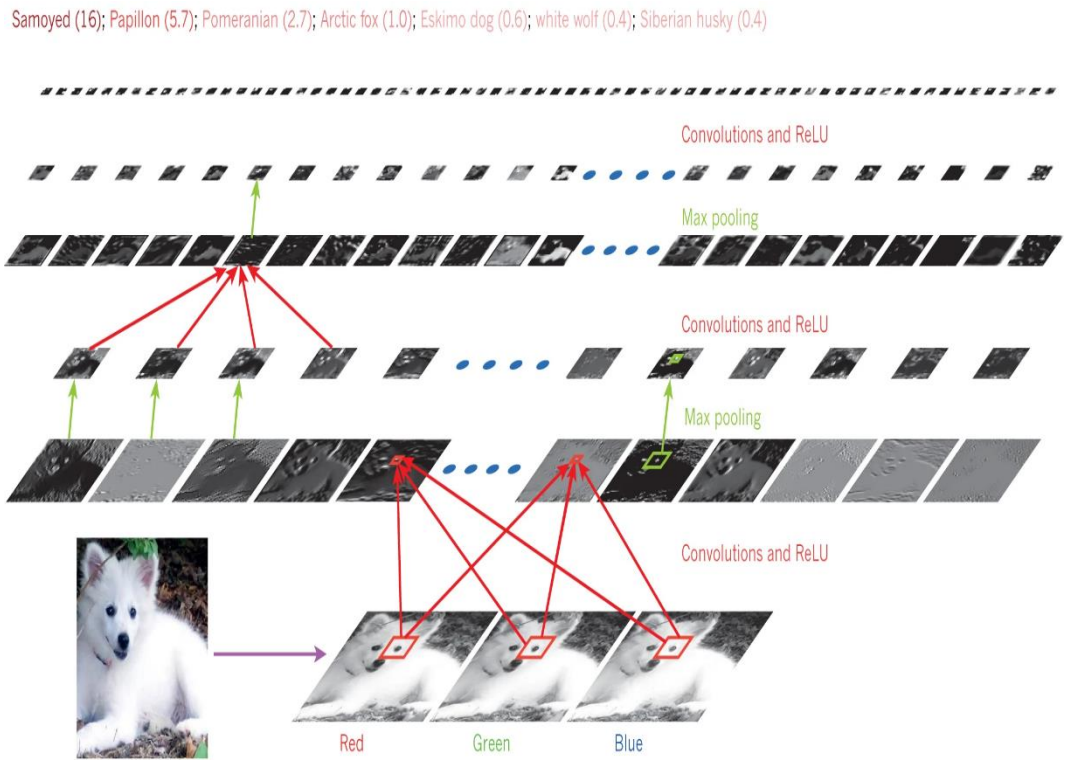

Figura 3 Funzionamento di una rete convoluzionale (LeCun et al., 2015)

A seguito di uno strato convoluzionato vi è sempre uno strato di pooling, il quale ha tre obiettivi:

- Ridurre il numero di parametri
- Controllare l'overfitting
- Promuovere l'invarianza

Per ridurre il numero di parametri, evidenziando la featurs estratta nello strato precedente, viene applicata una operazione non lineare, definita max pooling, la quale suddivide la matrice proveniente dallo strato precedente in blocchi e seleziona l'elemento con valore maggiore sfruttando il fenomeno del "winner takes all", che si compie tra le cellule nervose quando sono in competizione tra loro, dando come risultato una nuova matrice dove si può identificare una caratteristica dominante su tutte le altre.

Gli ultimi strati sono interamente connessi fra loro e potrebbero essere connessi uno strato di output, addestrato con apprendimento supervisionato, avente il compito di classifica le informazioni elaborate negli strati sottostanti; in particolare l'ultimo strato, per eseguire una classificazione il più precisa possibile, utilizza la funzione soft-max, la quale può assumere valori compresi fra 0 e 1. Tale funzione quando viene presentata l'immagine alla rete fornisce la medesima probabilità a tutte le categorie della rete, successivamente però tenderà ad attribuire una maggiore probabilità alla categoria che presenta maggiori affinità con le caratteristiche estratte dalla rete riducendo al contempo la probabilità di selezione delle altre categorie.

Le rete neurali profonde sfruttano la presenza di più strati nascosti per riuscire a costruire delle rappresentazioni più complesse, unendo negli strati più elevati le caratteristiche estratte nei livelli inferiori; quindi da un'immagine vengono estratte inizialmente dei bordi, i quali unendosi formano dei motivi che assemblati a loro volta formano degli oggetti (LeCun et al., 2015).

2.2. Confronto tra CNN e Uomo

In uno studio di Cadieu et al., 2014 vengono messe a confronto le capacità di riconoscimento oggetti tra i primati (macachi) con le loro reti neurali profonde, che a differenza delle reti bio-ispirate costituite da 3 o 4 strati nascosti, quest'ultime possedevano dai 7 fino ai 9 strati nascosti e il loro addestramento era avvenuto per addestramento supervisionato su milioni di immagini etichettate, detti anche big data.

All'interno dell'esperimento bisognava riuscire a riconoscere l'immagine presentata, tra un set di 1960 immagini ottenendo in questo modo una grande variazione allo stimolo come, la sua grandezza, l'angolazione o lo sfondo, eliminando così ogni fenomeno di dipendenza presenti nel mondo reale, introducendo la possibilità di controllare la loro variabilità e difficoltà (Figura5).

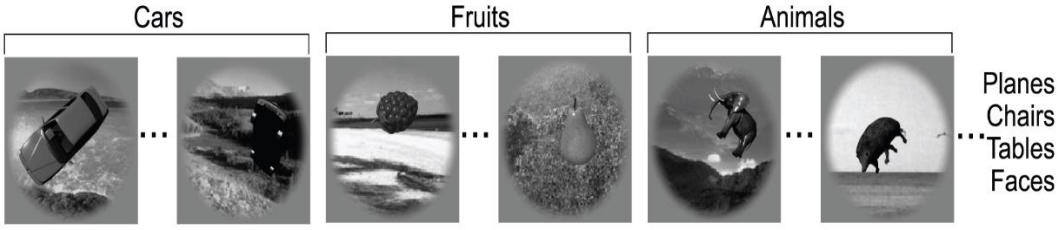

Figura 5 esempio di immagini dell'esperimento (Cadieu et al., 2014)

Come primo step è stato misurato la frequenza di attivazioni dell'area visiva V4 e della corteccia IT dei macachi mentre eseguivano una fissazione passiva, nell'intervallo che intercorre tra i 70 e 170 ms dopo l'esposizione, di ogni immagine delle 1960, per 100ms (Keysers et al., 2001), ripetuto per 47 volte .

Per valutare invece la prestazione della rete neurale si utilizza un'estensione dell'analisi del kernel dove si valuta l'efficacia della rappresentazione misurando come varia la precisione del problema della regressione all'aumentare della complessità della funzione di regressione (Montavon et al., 2011).

Intuitivamente le rappresentazioni più efficaci sono quelle che riescono a ignorare la variabilità irrilevante, (sfondo, dimensione e orientamento dell'oggetto) prestando unicamente attenzione all'oggetto target.
Come primo passaggio, prima di confrontare le reti artificiali con le capacità di riconoscimento dei primati, sono state osservate le prestazioni delle reti artificiali sul test set (Figura 6); confrontando le reti bio-ispirate (V2-like, V2-like e HMAX) con le DNN, reti neuronali profonde convoluzionate (HMO, Krizhevsky et al. 2012, Zeiler & Fergus 2013)

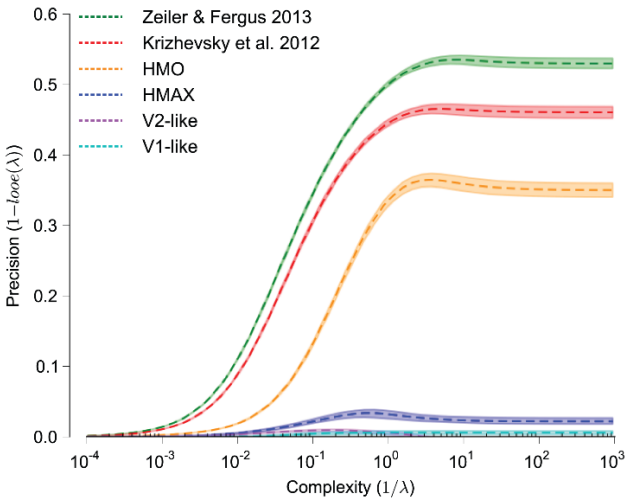

Figura 6 Curva di analisi del kernel delle prestazioni del modello (Cadieu et al., 2014)

Si osserva come le reti V1, V2 e HMAX, nonostante quest'ultima presenti delle lievi migliori prestazioni rispetto alle altre due, ottengono dei pessimi risultati nel compito di riconoscimento, dovuta, non tanto alla loro incapacità di riconoscimento dell'oggetto, ma a causa delle variazioni subite dall'oggetto (Pinto et al., 2008).
Al contrario si osserva come le CNN presentino oltre a una migliore capacità di riconoscimento dell'immagine, anche un aumento della precisione nel riconoscimento con l'aumentare della complessità dell'immagine.

Per riuscire a confrontare le prestazioni delle reti neurali e dei macachi, si necessita di rendere eque le rappresentazioni tra i due fissando un numero di campioni neurali e introducendo del rumore nelle rappresentazioni delle reti, ciò è dovuto a delle variabili impossibili da esulare nel momento della misurazione delle risposte dei macachi e dal rumore intrinseco dello strumento utilizzato per ottenere i dati (Cadieu et al., 2014).

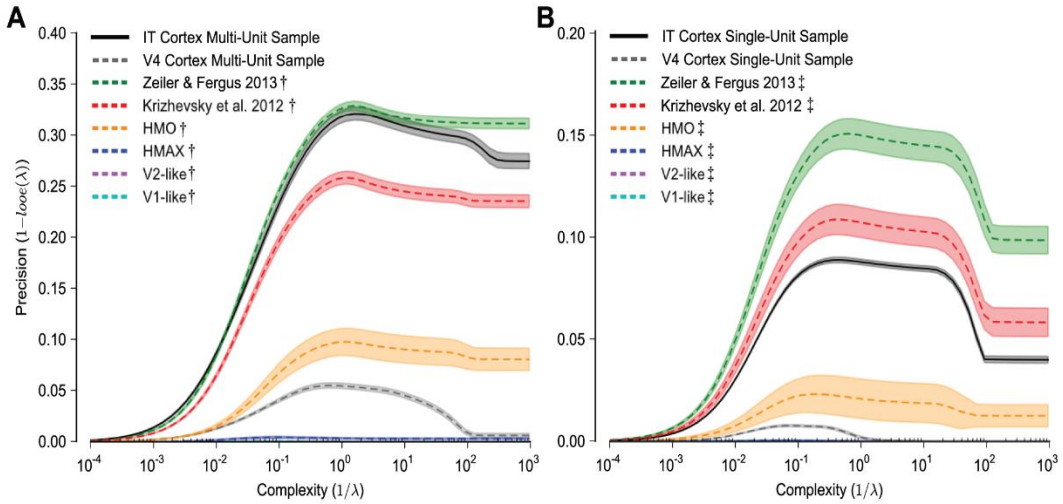

Figura 7 Curve di analisi del kernel delle rappresentazioni neuronali e della rete abbinate al campione e al rumore (Cadieu et al., 2014)

Dal confronto tra campioni multi-unità (Figura 7A), costituito da 80 campioni neuronali e 80 caratteristiche del modello, si può notare come le prestazioni della area visiva V4 siano significamenti inferiori alla corteccia IT (Rust & DiCarlo, 2010), le quali prestazioni sono soddisfacenti tal punto da essere equiparabili alla DNN Zeiler & Fergus 2013.

Se invece si fra campioni a singola unità (Figura 7B), costituito da 40 campioni neuronali e 40 caratteristiche del modello, a causa del numero minore di dati raccolti con questa condizione e la maggior pervasività del rumore si può rilevare come le prestazioni della corteccia IT siano superiori a HMAX e leggermente inferiori a Krizhevhevsky et al. 2012.

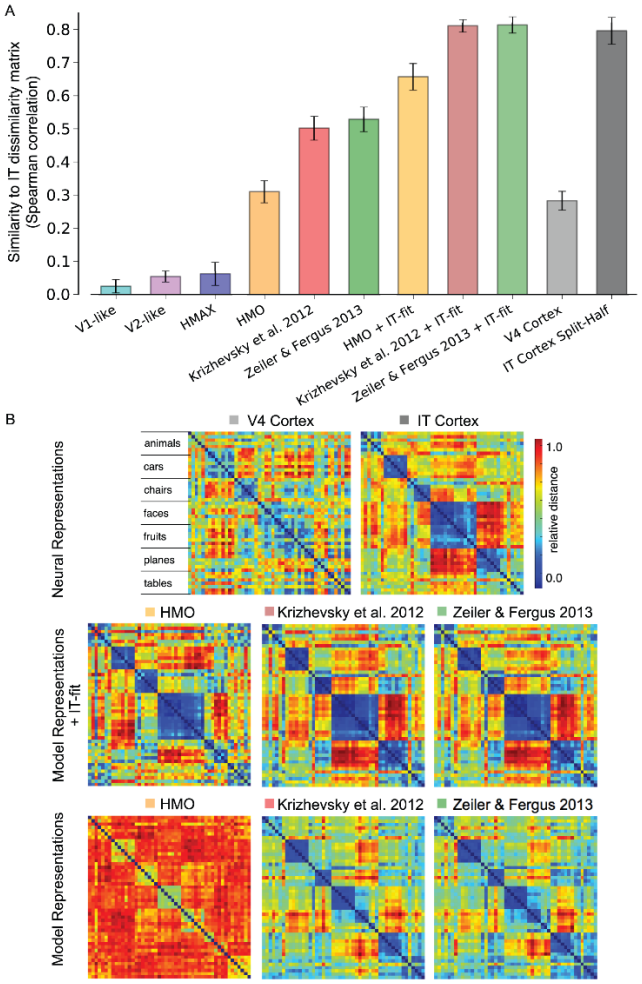

Figura 8 Analisi della somiglianza rappresentativa a livello di oggetto che confronta il modello e le rappresenta azioni neurali con la rappresentazione multi-unità IT (Cadieu et al., 2014)

In fine è stata osservata il grado di somiglianza tra le rappresentazioni.

Per fare ciò è stata costruita una matrice di dissimilarità a livello dell'oggetto (RDM) (di dimensioni 49x49, dato che all'interno delle immagini erano presenti 49 oggetti) (Figura 8B) e il coefficiente di correlazione per ranghi di Sperman tra le RDM delle reti e la RDM multi-unità IT (Figura 8A).

Per confronto è stata anche considerata la somiglianza tra le rappresentazioni formatesi nelle due cortecce IT.

Oltre a ciò, per seguire la mediologia proposta da (Yamins et al., 2014), la quale prevede di sfruttale le previsioni del sito multi-unità IT, da parte delle

reti, per costruire una nuova rappresentazioni (nella figura è aggiunto “+IT-fit”).

Confrontandoli si osserva come le rappresentazioni delle reti sono particolarmente differenti dalle rappresentazioni ottenute in IT e ciò non spiegabile dalla presenza del rumore ma unicamente da un divario a livello delle rappresentazioni tra le reti profonde e la corteccia IT (Cadieu et al., 2014).

3. CONCLUSIONE

Al giorno d'oggi la differenza nei compiti di riconoscimento d'oggetti dimostrano come le reti neurali artificiali presentino dei livelli di prestazione pari, se non superiori, a quelli dell'uomo.

È interessante però far notare come se pur differenti presentano una grande analogia fra esse, ovvero un'organizzazione gerarchica di elaborazione dell'informazioni, costituita nelle aree occipitali (V1, V2, V3, V4, V5) e temporali (in particolare la corteccia IT) nell'uomo e dalla disposizione stratificata all'interno delle reti (numero di strati nascosti).

Infatti, in entrambi i casi si può osservare come le informazioni più semplici, come linee e il loro orientamento, sia elaborato nei livelli più bassi della gerarchia, successivamente le informazioni vengono integrate in modo da ottenere rappresentazioni più complesse.

Le reti artificiali però presentano delle limitazioni relative al riconoscere un oggetto nel contesto, con ciò non si intende che non sanno usare lo sfondo per trarre informazioni per aggiungere dettagli relative all'immagine, come dimostrato da (LeCun et al., 2015) quando implemento una rete ricorrente a una CNN per descrivere delle immagini, ma sulla capacità di riconoscere un oggetto ambiguo se isolato dall'immagini (Es. denominare una ciocca di capelli come un animale).

In conclusione, le reti artificiali presentano data la loro somiglianza architettonica con la disposizione delle cellule celebrali, possono essere, usate per studiare nel campo della psicologia sperimentale, per spiegare il funzionamento del cervello riuscendo a esulare il rumore misurato dagli strumenti di neuroimaging dovuto allo stato di attivazioni di neuroni non ignorabili dalla registrazione, dovute alla loro vicinanza spaziale.

4. BIBLIOGRAFIA

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, *166*(3), 559–571. https://doi.org/10.1007/s00221-005-2396-5

Balderas, I., Rodriguez-Ortiz, C. J., Salgado-Tonda, P., Chavez-Hurtado, J., McGaugh, J. L., & Bermudez-Rattoni, F. (2008). The consolidation of object and context recognition memory involve different regions of the temporal lobe. *Learning & Memory*, *15*(9), 618–624. https://doi.org/10.1101/lm.1028008

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Brewer, A. A., Press, W. A., Logothetis, N. K., & Wandell, B. A. (2002). Visual Areas in Macaque Cortex Measured Using Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *22*(23), 10416–10426. https://doi.org/10.1523/JNEUROSCI.22-23-10416.2002

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Collins, C. E., Airey, D. C., Young, N. A., Leitch, D. B., & Kaas, J. H. (2010). Neuron densities vary across and within cortical areas in primates. *Proceedings of the National Academy of Sciences*, *107*(36), 15927–15932. https://doi.org/10.1073/pnas.1010356107

Courtney, S. M., & Ungerleider, L. G. (1997). What fMRI has taught us about human vision. *Current Opinion in Neurobiology*, *7*(4), 554–561. https://doi.org/10.1016/S0959-4388(97)80036-0

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10), 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Gross, C. G. (1994). How Inferior Temporal Cortex Became a Visual Area. *Cerebral Cortex*, *4*(5), 455–469. https://doi.org/10.1093/cercor/4.5.455

Horel, J. A. (1996). Perception, learning and identification studied with reversible suppression of cortical visual areas in monkeys. *Behavioural Brain Research*, *76*(1), 199–214. https://doi.org/10.1016/0166-4328(95)00196-4

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106-154.2.

Keysers, C., Xiao, D.-K., Földiák, P., & Perrett, D. I. (2001). The Speed of Sight. *Journal of Cognitive Neuroscience*, *13*(1), 90–101. https://doi.org/10.1162/089892901564199

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Art. 7553. https://doi.org/10.1038/nature14539

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139. https://doi.org/10.1073/pnas.92.18.8135

O'Kusky, J., & Colonnier, M. (1982). A laminar analysis of the number of neurons, glia, and synapses in the visual cortex (area 17) of adult macaque monkeys. *Journal of Comparative Neurology*, *210*(3), 278–290. https://doi.org/10.1002/cne.902100307

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is Real-World Visual Object Recognition Hard? *PLOS Computational Biology*, *4*(1), e27. https://doi.org/10.1371/journal.pcbi.0040027

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), Art. 11. https://doi.org/10.1038/14819

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*(11), Art. 11. https://doi.org/10.1038/81479

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Stoerig, P., & Cowey, A. (1997). Blindsight in man and monkey. *Brain*, *120*(3), 535–559. https://doi.org/10.1093/brain/120.3.535

Stucki, G., Kostanjsek, N., Ustün, B., & Cieza, A. (2008). ICF-based classification and measurement of functioning. *European Journal of Physical and Rehabilitation Medicine*, *44*(3), 315–328.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, *11*(2), 58–64. https://doi.org/10.1016/j.tics.2006.11.009

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111