Alberto Testolin, Marco Zorzi

L'approccio moderno all'intelligenza artificiale e la rivoluzione del deep learning

(doi: 10.1421/102685)

Giornale italiano di psicologia (ISSN 0390-5349)
Fascicolo 2, giugno 2021

Ente di afferenza:
Università di Padova (unipd)

INTERVENTI

L'APPROCCIO MODERNO ALL'INTELLIGENZA ARTIFICIALE E LA RIVOLUZIONE DEL *DEEP LEARNING*

ALBERTO TESTOLIN[1] E MARCO ZORZI[1,2]

[1] *Università di Padova,* [2] *IRCCS Ospedale San Camillo*

Riassunto. Nell'ultimo decennio le ricerche nel campo dell'intelligenza artificiale sono state segnate da una serie impressionante di successi, in gran parte dovuti all'introduzione di efficaci algoritmi di apprendimento automatico. Uno degli approcci più promettenti è costituito dal *deep learning*, che consente di creare reti neurali artificiali gerarchiche in grado di estrarre autonomamente conoscenza da enormi basi di dati. In questa rassegna discuteremo i principali progressi teorici e tecnologici alla base di questi modelli, sottolineandone anche la rilevanza per la psicologia e le neuroscienze cognitive. Evidenzieremo anche i principali limiti di questo approccio e le prospettive di ricerca per superarli.

1. INTRODUZIONE

Uno degli scopi delle scienze cognitive è di caratterizzare i fenomeni mentali da un punto di vista computazionale. Questo approccio è parzialmente condiviso dai ricercatori che operano nel campo dell'Intelligenza Artificiale (IA), il cui obiettivo è di implementare processi cognitivi all'interno di macchine, riproducendo (almeno in parte) la complessità del pensiero umano per renderle in qualche modo «intelligenti». Agli albori dell'invenzione dei calcolatori digitali, il matematico inglese Alan Turing ha proposto di adottare una definizione pragmatica del concetto di intelligenza concependo un gioco di imitazione noto come Test di Turing (Turing, 1950). Secondo tale definizione un sistema artificiale può essere considerato intelligente se il suo comportamento risulta indistinguibile da quello di un essere umano; sebbene rifletta una visione antropocentrica dell'intelligenza, questo test viene spesso considerato un metodo oggettivo per misurare i progressi nel campo dell'IA (Russell e Norvig, 2020).

Inizialmente l'approccio prevalente per costruire modelli di IA era basato sul paradigma del *cognitivismo*, che descrive il funzionamento della mente attraverso diagrammi di flusso e regole sintattiche (Chomsky, 1957; Pinker, 1999). La mente veniva concepita come un'entità indipendente dal sostrato fisico che la implementa, ipotizzando che i processi cognitivi potessero essere riprodotti in calcolatori universali

in grado di manipolare rappresentazioni simboliche (McCarthy e Hayes, 1968; Newell e Simon, 1961). Questo progetto si è rivelato più complesso del previsto a causa del problema del *symbol grounding*: non è chiaro come una macchina programmata per eseguire operazioni sintattiche possa attribuire un significato a tali rappresentazioni (Harnad, 1990; Searle, 1980).

Un approccio alternativo è costituito dal *connessionismo*, che trae origine dalle teorie cibernetiche e concepisce il pensiero come fenomeno emergente, intimamente collegato alle proprietà fisiche del sistema che lo supporta (Rumelhart e McClelland, 1986). Traendo ispirazione da alcune proprietà del sistema nervoso (McCulloch e Pitts, 1943), i modelli basati su reti neurali artificiali mirano quindi a sviluppare capacità cognitive attraverso meccanismi di apprendimento ed auto-organizzazione, che consentono di interagire con l'ambiente esterno mediante processi di *feedback*. Sfortunatamente, nonostante l'entusiasmo iniziale anche la strada del connessionismo si è rivelata più tortuosa del previsto, al punto da essere quasi abbandonata (tranne per la ricerca teorica nelle scienze cognitive) a seguito della comparsa di altre tecniche di apprendimento automatico (*machine learning*).

Negli ultimi anni stiamo tuttavia assistendo ad una rinascita delle reti neurali artificiali, resa possibile dalla rivoluzione del *deep learning* (LeCun, Bengio e Hinton, 2015). Come discuteremo in questa rassegna, questi modelli superano il Test di Turing in domini considerati da sempre appannaggio esclusivo dell'intelligenza umana, rendendo quindi sempre meno marcato il divario cognitivo tra uomo e computer ed alimentando aspettative – spesso esagerate – sulla possibile creazione di autentiche intelligenze artificiali. Discuteremo inoltre la rilevanza di questa nuova generazione di reti neurali artificiali per la modellizzazione computazionale nell'ambito delle scienze psicologiche e delle neuroscienze cognitive.

2. RETI NEURALI ARTIFICIALI E METODI DI APPRENDIMENTO AUTOMATICO

Un sistema di apprendimento automatico consiste in un algoritmo adattivo, che migliora la propria prestazione in un certo dominio in base all'esperienza ed è in grado di generalizzare la conoscenza appresa in situazioni nuove (Mitchell, 1997). L'aggettivo «adattivo» indica che il comportamento dell'algoritmo non è prestabilito, ma viene progressivamente rifinito a seguito dell'interazione con l'ambiente. È inoltre fondamentale che l'algoritmo sia in grado di «generalizzare», ovvero di estrarre conoscenza utile non solo per operare sui dati di apprendimento (*training set*) ma anche su dati mai incontrati prima

(*test set*). Generalmente si distinguono tre principali paradigmi di apprendimento automatico, noti rispettivamente come *supervisionato*, *non supervisionato*, e *per rinforzo*, in base alla modalità attraverso la quale viene fornito il segnale di *feedback* che guida l'apprendimento stesso (per una introduzione all'apprendimento nelle reti neurali dal punto di vista della psicologia si rimanda a Zorzi, 2016).

Nell'*apprendimento supervisionato* i dati di apprendimento sono etichettati, ovvero per ciascuno stimolo in input viene fornita anche la risposta di output desiderata. Uno dei primi esempi di modello di apprendimento supervisionato è stato il *perceptron* (Rosenblatt, 1958): come illustrato in figura 1A, si tratta di un dispositivo che riceve in input un insieme ordinato di valori, lo elabora secondo una precisa funzione matematica e restituisce in output un valore numerico che rappresenta l'attivazione del neurone. In un contesto di visione artificiale l'input potrebbe essere costituito dai valori di luminosità dei pixel di un'immagine, mentre in contesti di riconoscimento del linguaggio parlato potrebbe codificare le frequenze sonore di una parola; in entrambi i casi la risposta di output identificherebbe la categoria a cui appartiene lo stimolo sensoriale. L'obiettivo dell'apprendimento è di *minimizzare l'errore di classificazione*, ovvero ridurre la discrepanza tra la categoria predetta dal modello e la categoria corretta. Per ottenere questo risultato, l'algoritmo di apprendimento modifica progressivamente i pesi delle connessioni sinaptiche corrispondenti a ciascuna variabile di input, in modo da allineare l'output prodotto dal *perceptron* con la risposta desiderata. Sebbene questo tipo di modello sia effettivamente in grado di imparare a risolvere semplici problemi di classificazione, fallisce nel caso di problemi non lineari (Minsky e Papert, 1969). Tuttavia, il *perceptron* può essere utilizzato come unità di base per costruire reti neurali più complesse, costituite da uno strato aggiuntivo di neuroni «nascosti» (fig. 1B). In questo caso i pesi delle connessioni vengono modificati utilizzando l'algoritmo di *error backpropagation* (Rumelhart, Hinton e Williams, 1986).

Nell'*apprendimento non supervisionato*, i dati di apprendimento non sono etichettati: viene fornito solo lo stimolo di input. In questa categoria rientrano i modelli di *clustering* e riduzione di dimensionalità. Lo scopo dell'apprendimento in questo caso è di estrarre un insieme di fattori latenti («rappresentazione interna») che consentano di descrivere i dati in maniera compatta. Per esempio, un *autoencoder* (fig. 1C) riceve in input uno stimolo (pixel di un'immagine) e cerca di riprodurre in output esattamente la stessa configurazione. L'obiettivo dell'apprendimento è di *minimizzare l'errore di ricostruzione*, ovvero ridurre la discrepanza tra l'input originale e la ricostruzione prodotta dal modello. Un'ulteriore estensione consiste nel dotare la rete neurale di connessioni ricorrenti (Elman, 1990), che consentono di ela-

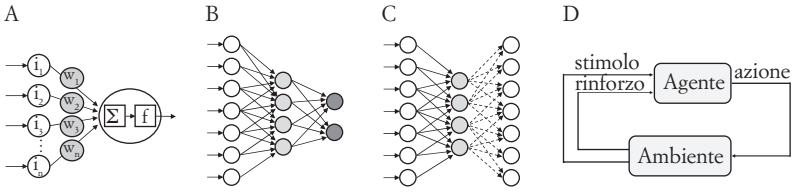

FIG. 1. (A) Rappresentazione grafica del *perceptron*. I segnali in input vengono inizialmente sommati, pesando il contributo di ciascuna dimensione *i* per il relativo peso sinaptico *w*. Viene poi applicata una funzione di attivazione *f* (solitamente non lineare, per esempio la funzione logistica) che stabilisce il livello di attivazione da trasmettere in output. (B) Rete neurale supervisionata, dotata di uno strato di unità di input (bianco), uno strato di unità nascoste (grigio chiaro) ed uno strato di unità di output (grigio scuro). (C) Rete neurale non supervisionata *(autoencoder)*, che ricostruisce i dati in input attraverso connessioni di *feedback* (frecce tratteggiate). (D) Modello di apprendimento con rinforzo: l'agente seleziona un'azione da compiere, alla quale viene associato un nuovo stato dell'ambiente ed eventualmente un valore di ricompensa o punizione.

borare stimoli sequenziali attraverso meccanismi di codifica predittiva (Clark, 2013). Questa classe di modelli risulta più plausibile dal punto di vista cognitivo rispetto alla controparte supervisionata, in quanto l'apprendimento consente di costruire un modello interno dei dati sensoriali attraverso mera osservazione (*statistical learning*) senza richiedere supervisione esplicita (Berkes, Orbán, Lengyel e Fiser, 2011; Perruchet e Pacton, 2006).

Nell'*apprendimento per rinforzo* i dati di apprendimento non sono etichettati ma l'agente è in grado di interagire con l'ambiente producendo una serie di azioni che hanno effetti sui successivi input sensoriali (fig. 1D). In questo caso l'ambiente fornisce un *feedback* indiretto (punizione o ricompensa) a seconda dell'azione scelta: lo scopo dell'apprendimento è di *massimizzare il rinforzo cumulativo* nell'arco dell'esistenza dell'agente e può essere ricondotto al paradigma di condizionamento operante.

Queste tre modalità di apprendimento non sono da considerarsi mutualmente esclusive, ma piuttosto complementari: la maggior parte dei ricercatori concorda sul fatto che la creazione di macchine intelligenti richieda la combinazione di tutti gli approcci (si vedano, per esempio, i modelli di apprendimento semi-supervisionato; Chapelle, Schölkopf e Zien, 2006).

3. LA RIVOLUZIONE DEL *DEEP LEARNING*

Il quadro teorico dei moderni modelli di reti neurali artificiali è rimasto sostanzialmente invariato rispetto al paradigma connessioni-

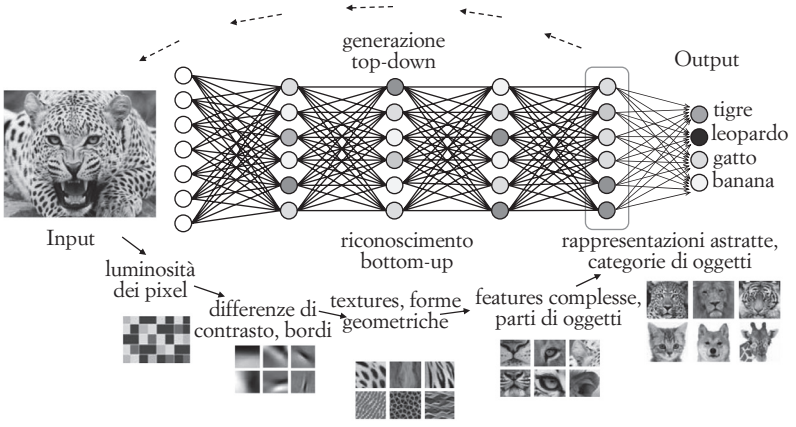

FIG. 2. Architettura di un sistema di *deep learning* per l'elaborazione visiva. La rete neurale riceve in input un'immagine digitale, codificata come matrice contenente i valori di luminosità di ciascun pixel. L'elaborazione avviene in maniera gerarchica: i neuroni dei primi strati identificano semplici caratteristiche (*features*) visive, che vengono combinate per creare rappresentazioni progressivamente più astratte dell'input. Nelle architetture supervisionate lo scopo è di riconoscere la classe di appartenenza dello stimolo, propagando il segnale attraverso le connessioni *bottom-up* fino ad attivare il neurone corrispondente nello strato di output. Nelle architetture non supervisionate non è invece presente lo strato di output, in quanto lo scopo è di ricostruire accuratamente lo stimolo a partire dalle rappresentazioni astratte attraverso le connessioni generative *top-down*.

sta sviluppato negli anni '80. Tuttavia, una serie di progressi ha consentito la creazione di *reti neurali profonde* dotate di molti strati di neuroni nascosti organizzati gerarchicamente (da cui il termine *deep networks* e *deep learning*). A differenza delle tradizionali architetture costituite da un unico strato nascosto (fig. 1B-C) ed in analogia con l'organizzazione gerarchica della corteccia cerebrale (Felleman e Van Essen, 1991), questo tipo di modelli elabora l'informazione sensoriale in maniera più efficace sfruttando molteplici livelli di rappresentazione (fig. 2). Per esempio, in un compito di riconoscimento visivo i neuroni che implementano i primi stadi di elaborazione imparano ad estrarre caratteristiche basilari delle immagini, come differenze di contrasto, gradienti di luminosità e barre orientate in diverse angolazioni; i neuroni degli strati successivi combinano queste informazioni per apprendere caratteristiche più complesse, codificando per esempio giunzioni tra barre, *texture* e forme geometriche; infine, negli strati più profondi della gerarchia vengono create rappresentazioni complesse come parti di oggetti, volti, categorie di oggetti ed intere scene visive.

È legittimo chiedersi perché la gestazione delle architetture gerarchiche sia durata così a lungo. Possiamo innanzitutto identificare problemi di natura pratica. Gli algoritmi di apprendimento automatico basati su reti neurali richiedono un numero molto elevato di esempi di apprendimento e la richiesta di dati cresce all'aumentare delle dimensioni del modello. Inoltre, simulazioni su larga scala richiedono l'utilizzo di architetture di calcolo ad alte prestazioni. Questi problemi sono stati parzialmente risolti con l'avvento dei *big data*, ovvero database digitali di grandi dimensioni dai quali estrarre esempi per l'apprendimento, e con l'adozione di dispositivi di calcolo che parallelizzano le operazioni su migliaia di processori grafici (*Graphic Processing Units*, GPU) (Owens e Houston, 2008; Testolin, Stoianov, De Filippo De Grazia e Zorzi, 2013). La creazione di architetture gerarchiche comporta tuttavia anche problemi di natura teorica. Modelli con molti strati di neuroni soffrono del problema della scomparsa del gradiente (*vanishing gradient*; Bengio, Simard e Frasconi, 1994): essendo la modifica dei pesi sinaptici basata sul calcolo del gradiente della funzione di errore, utilizzando la classica funzione di attivazione logistica la derivata tende a diventare sempre più piccola man mano che ci si allontana dallo strato di output, rendendo irrisorie le modifiche dei pesi negli strati più vicini all'input sensoriale. Questo problema è stato mitigato introducendo funzioni di attivazione che non «saturano», come la *rectified linear unit* (Nair e Hinton, 2010) ed evitando di inizializzare i pesi delle connessioni in modo totalmente casuale (Glorot e Bengio, 2010; Sutskever, Martens, Dahl e Hinton, 2013). Sono stati inoltre perfezionati gli algoritmi di ottimizzazione utilizzati per minimizzare la funzione di errore, per esempio introducendo tassi di apprendimento adattivi (Kingma e Ba, 2015). Infine, la creazione di modelli su larga scala comporta un incremento sostanziale del numero di «parametri liberi» del modello. Dal punto di vista statistico, avere un modello con molti parametri (in questo caso, *milioni* di connessioni sinaptiche) incrementa notevolmente il rischio di *overfitting*, che viene contrastato con l'introduzione di efficaci regolarizzatori, come il *dropout* (Srivastava, Hinton, Krizhevsky, Sutskever e Salakhutdinov, 2014). Altri problemi sono specifici per particolari modalità di apprendimento: tratteremo le principali strategie adottate per risolverli nei paragrafi seguenti.

3.1. *Deep learning supervisionato*

I successi del *deep learning* si possono largamente attribuire alla creazione di enormi database digitali contenenti dati accuratamente etichettati. Per esempio, nel 2009 viene rilasciato il database Ima-

geNet (Deng *et al.*, 2009), costituito da milioni di immagini ad alta risoluzione annotate in base al contenuto semantico ed organizzate seguendo la struttura ontologica del database WordNet[1]. Creare un database di queste dimensioni ha richiesto un enorme investimento di risorse ed il reclutamento di migliaia di annotatori umani attraverso piattaforme online di *crowdsourcing*. L'investimento è stato però ampiamente ripagato: nel 2012 una rete neurale gerarchica addestrata in modo supervisionato sulle immagini di ImageNet ha ottenuto una prestazione di gran lunga superiore a qualsiasi algoritmo di visione artificiale esistente (Krizhevsky, Sutskever e Hinton, 2012) e nel giro di qualche anno il perfezionamento di questi modelli ha consentito di raggiungere accuratezze di classificazione superiori persino a quella umana (He, Zhang, Ren e Sun, 2016). Risultati simili sono stati ottenuti in compiti di elaborazione del linguaggio: grazie all'impiego di database contenenti la trascrizione di milioni di parole, il *deep learning* rappresenta lo stato dell'arte nei sistemi di riconoscimento vocale (Hinton *et al.*, 2012) e raggiunge notevoli prestazioni anche in difficili compiti di traduzione automatica (Sutskever, Vinyals e Le, 2014). L'utilizzo del *deep learning* supervisionato si è presto diffuso ad altri ambiti dove erano disponibili grandi quantità di dati annotati, come nel caso dell'analisi di composti chimici e la creazione di farmaci (Ma *et al.*, 2015). Si può quindi affermare che, nel caso dell'apprendimento supervisionato, i principali catalizzatori del progresso siano stati la creazione di grossi database annotati e la disponibilità di potenti calcolatori paralleli, in grado di supportare l'implementazione di modelli su larga scala.

3.2. *Deep learning non supervisionato*

Nonostante i progressi più eclatanti vengano solitamente attribuiti allo sviluppo del *deep learning* supervisionato, i primi modelli di reti neurali multistrato sono stati creati sfruttando progressi nell'apprendimento non supervisionato. In particolare, nel 2006 è stato dimostrato per la prima volta come fosse possibile apprendere molteplici livelli di rappresentazione creando un modello generativo gerarchico (Hinton e Salakhutdinov, 2006), in cui i neuroni degli strati interni rappresentano le variabili latenti (in una interpretazione Bayesiana, le ipotesi) che spiegano i dati osservati (ad esempio la distribuzione di pixel nell'immagine). Durante l'apprendimento i pesi della rete neurale vengono modificati per minimizzare la discrepanza tra l'informazione

[1] http://www.image-net.org.

sensoriale in arrivo e la «predizione» generata dalla rete stessa attraverso connessioni *top-down*. Tutte le connessioni tra strati sono infatti bidirezionali e l'attivazione dello strato più profondo può propagarsi all'indietro fino allo strato di input. Il fatto che in queste reti neurali artificiali le connessioni rientranti (*top-down*) siano necessarie per l'apprendimento di un modello interno fornisce una possibile spiegazione del perché i circuiti neurali nel cervello siano caratterizzati da una massiccia presenza di questo tipo di connessioni (per approfondimenti sui modelli generativi si vedano Zorzi, 2006a; Zorzi, Testolin e Stoianov, 2013).

L'intuizione iniziale di Hinton e colleghi è stata quella di utilizzare come punto di partenza una rete con un singolo strato di unità nascoste addestrata come *autoencoder* (fig. 1C) e poi utilizzare le attivazioni dei neuroni nascosti (che costituiscono le «rappresentazioni interne» del modello) come input per un successivo *autoencoder*, che impara quindi a rappresentare i dati estraendo correlazioni di ordine superiore (Hinton, 2007). Questo approccio ha dimostrato come fosse possibile estrarre caratteristiche astratte dai dati sensoriali «grezzi»: per esempio, un modello generativo gerarchico addestrato su un database contenente milioni di immagini non annotate ha sviluppato neuroni con profili di risposta estremamente sofisticati che fungevano da «rilevatori di volti» (Le *et al.*, 2012). Successivi sviluppi hanno portato all'introduzione degli *autoencoder variazionali* (Kingma e Welling, 2013) e delle *generative adversarial networks* (Goodfellow *et al.*, 2014), che apprendono in maniera più efficace i fattori latenti della distribuzione dei dati di input e possono quindi generare in modo *top-down* dati sensoriali estremamente realistici. Per esempio, questi modelli sono in grado di generare immagini sintetiche di volti alterandone a piacimento caratteristiche come l'espressione, il colore ed il taglio dei capelli, la forma degli occhi e del naso (Karras, Aila, Laine e Lehtinen, 2018).

Altri sviluppi teorici hanno migliorato i modelli ricorrenti basati su *Long-Short Term Memory* dotandoli di una maggior capacità di «memoria di lavoro» (intesa come capacità di mantenere attiva nel tempo l'informazione utile per il compito) (Lipton, Berkowitz e Elkan, 2015) e di meccanismi di attenzione selettiva (Vaswani *et al.*, 2017), consentendo di elaborare più efficacemente dati sequenziali. Questi progressi hanno portato alla creazione di reti neurali in grado di apprendere in maniera molto accurata la struttura del linguaggio umano a partire da corpora linguistici e grossi archivi testuali, che si riflette nella capacità di produrre dialoghi estremamente realistici (Devlin, Chang, Lee e Toutanova, 2018). I modelli più recenti, dotati di *centinaia di miliardi* di connessioni sinaptiche (Brown *et al.*, 2020), sono in grado di generare notizie inventate, poesie in rima, sonetti, brevi storie e parodie

in maniera totalmente autonoma, esibendo un livello di creatività così impressionante da spingere gli autori stessi a renderne privato il codice per evitare che possa essere utilizzato per scopi malevoli[2].

3.3. *Deep learning con rinforzo*

Negli anni più recenti il *deep learning* ha raggiunto l'apice della popolarità grazie alla combinazione con algoritmi di apprendimento per rinforzo, come il *Q-learning* (Watkins e Dayan, 1992). Questi algoritmi venivano applicati di rado in contesti pratici a causa della loro complessità computazionale, che rendeva l'apprendimento lento ed inefficace. A partire dal 2015, una serie di studi ha dimostrato come fosse possibile utilizzare reti neurali multistrato per approssimare in maniera efficace la stima delle ricompense che l'agente riceverà nel tempo, riducendo quindi in modo significativo la complessità computazionale dell'apprendimento. Un primo risultato è stata la creazione di un sistema in grado di imparare a giocare a videogiochi interattivi, decidendo come agire sui controlli di gioco in base solamente all'informazione visiva presente sul monitor ed al punteggio ottenuto, ottenendo prestazioni comparabili a quelle di giocatori umani (Mnih *et al.*, 2015). L'anno successivo lo stesso approccio è stato utilizzato per creare AlphaGo (Silver *et al.*, 2016), un sistema che ha imparato a giocare all'antico gioco cinese Go, da sempre considerato fuori portata per le macchine a causa della complessità dello spazio delle soluzioni (Müller, 2002). In una serie di partite seguite online da migliaia di telespettatori, AlphaGo ha ripetutamente sconfitto il campione mondiale in carica Lee Sedol, dimostrando come l'approccio connessionista sia in grado di produrre macchine in grado di rivaleggiare con gli aspetti più sofisticati dell'intelligenza umana, come la pianificazione e l'apprendimento di strategie. Una delle frontiere della ricerca consiste nell'estendere questi modelli a contesti multi-agente, dove l'interazione con l'ambiente è resa più complessa dalla presenza di dinamiche di cooperazione e competizione tra agenti. Risultati incoraggianti continuano ad arrivare dal dominio dei giochi: lo scorso anno è stato presentato AlphaStar (Vinyals *et al.*, 2019), un sistema in grado di imparare a giocare a StarCraft, uno dei più difficili giochi di strategia online multi-giocatore, raggiungendo il livello di «Grandmaster» ed ottenendo un punteggio superiore al 99.8% dei 90.000 giocatori presenti nella classifica ufficiale della lega.

[2] https://openai.com/blog/better-language-models/.

Gli impressionanti successi del *deep learning* nelle applicazioni di IA hanno riacceso l'entusiasmo che si era creato a seguito dell'invenzione dell'algoritmo di *error back-propagation* (Crick, 1989). Tuttavia, oggi come allora ci si chiede se questi sistemi possano essere effettivamente considerati un buon modello per spiegare come i processi cognitivi possano emergere dalle complesse dinamiche del cervello umano. In molti rispondono affermativamente (Hassabis, Kumaran, Summerfield e Botvinick, 2017; Richards *et al.*, 2019), sebbene il consenso generale sia che per descrivere la complessità del cervello sarà necessario sviluppare un approccio integrato nel quale coesisteranno molteplici classi di modelli a diversi livelli di astrazione, dal singolo neurone all'organizzazione su larga scala del sistema nervoso (Bassett e Gazzaniga, 2011; Gerstner, Sprekeler e Deco, 2012). In questa prospettiva i modelli di *deep learning* potrebbero occupare una posizione privilegiata, in quanto consentirebbero di creare un indispensabile collegamento tra modelli biofisici, in grado di simulare in modo dettagliato le dinamiche neuronali, e modelli cognitivi in grado di catturare le dinamiche del comportamento (Testolin e Zorzi, 2016). Nei paragrafi successivi descriviamo brevemente alcuni studi particolarmente rilevanti per la comprensione dei meccanismi alla base della percezione e della cognizione umana (per una introduzione metodologica alla modellistica computazionale in psicologia e nelle neuroscienze cognitive si rimanda a Zorzi, 2006b, 2017).

4.1. *Percezione e riconoscimento*

Nell'ambito della modellazione dei processi sensoriali e percettivi, recenti studi hanno mostrato come unità con profili di risposta simili a quelli dei neuroni della corteccia visiva primaria possano emergere in modelli di *deep learning* che apprendono una codifica efficiente di immagini naturali (Güçlü e van Gerven, 2014; Testolin, Stoianov e Zorzi, 2017; Xiong, Rodríguez-Sánchez, Szedmak e Piater, 2015). Aggiungendo ulteriori strati di neuroni emergono unità con profili di risposta più sofisticati, simili a quelli dei neuroni della corteccia visiva secondaria (Le *et al.*, 2008), che si possono specializzare per rappresentare stimoli più complessi se il modello viene successivamente esposto ad un set di immagini di lettere stampate (Testolin *et al.*, 2017). Infine, nei livelli più profondi di modelli addestrati in compiti di riconoscimento di oggetti emergono rappresentazioni simili a quelle osservate nel cervello negli stadi più avanzati

della via visiva ventrale, in termini di codifica distribuita (Güçlü e van Gerven, 2015; Kriegeskorte, 2015; Yamins e DiCarlo, 2016) ma anche nei profili di risposta di singoli neuroni (Bashivan, Kar e DiCarlo, 2019). Risultati analoghi sono stati ottenuti con modelli che apprendono abilità visuo-spaziali, nei quali emergono rappresentazioni simili a quelle osservate nella via visiva dorsale (Stoianov e Zorzi, 2012; Testolin, De Filippo De Grazia e Zorzi, 2017; Zorzi e Testolin, 2018). Oltre a riflettere un'ampia gamma di proprietà neurofisiologiche, questi modelli possono riprodurre fedelmente anche il comportamento umano in esperimenti di psicofisica, simulando curve psicometriche (Testolin, Dolfi, Rochus e Zorzi, 2020), giudizi di similarità (Kubilius, Bracci e Op de Beeck, 2016) e matrici di confusione (Testolin, Stoianov *et al.*, 2017).

4.2. *Apprendimento e memoria*

Ponendo enfasi sui meccanismi di apprendimento, i modelli connessionisti si prestano particolarmente alla simulazione dell'acquisizione di capacità cognitive (Elman *et al.*, 1996). Recentemente, modelli di *deep learning* sono stati utilizzati con successo per simulare le traiettorie di sviluppo del senso del numero nel bambino (Testolin, Zou e McClelland, 2020) e per spiegare il processo graduale di differenziazione di categorie semantiche e prototipi concettuali (Saxe, McClelland e Ganguli, 2019). Alcune moderne tecniche impiegate in sistemi di IA hanno anche fornito preziose intuizioni sul possibile ruolo dell'ippocampo nei processi di memorizzazione ed integrazione dell'informazione (Kumaran, Hassabis e McClelland, 2016). In particolare, è stata proposta un'analogia tra il ruolo della memoria episodica ed il meccanismo di *memory replay* utilizzato per stabilizzare gli algoritmi di *deep learning* con rinforzo. Questa tecnica consente di incorporare nuova informazione nella rete neurale senza interferire con la conoscenza precedentemente acquisita, migliorando le capacità di generalizzazione dell'agente e supportando processi di pianificazione volti a massimizzare i guadagni futuri. Se addestrati in compiti di navigazione spaziale, questi modelli sviluppano neuroni con profili di risposta simili a quelli delle *grid cells* osservate nella corteccia entorinale dei ratti (Banino *et al.*, 2018). Sebbene la plausibilità biologica dell'algoritmo di *backpropagation* rimanga oggetto di dibattito, sono state inoltre avanzate alcune proposte riguardo a come i circuiti corticali potrebbero effettivamente implementare questo tipo di operazioni (Lillicrap, Santoro, Marris e Akerman, 2020).

4.3. *Comprensione del linguaggio e cognizione di alto livello*

Modelli basati su *deep learning* sono stati utilizzati per simulare l'apprendimento di strutture ortografiche e la generazione di pseudoparole (Testolin, Stoianov, Sperduti e Zorzi, 2016) e, più recentemente, l'apprendimento di strutture sintattiche (Futrell *et al.*, 2019) e regole composizionali (Baroni, 2020). Per misurare le capacità linguistiche di questi sistemi sono stati creati test specifici, simili alle prove di comprensione del testo utilizzate nelle scuole primarie e secondarie (Wang *et al.*, 2019). In molti casi, i modelli di *deep learning* hanno ottenuto punteggi comparabili a quelli umani (Devlin *et al.*, 2018; Liu, He, Chen e Gao, 2020). L'applicazione di questi modelli si sta estendendo a scenari sempre più complessi, per esempio simulando come rudimentali abilità di comunicazione possano emergere dalle necessità di cooperazione e comunicazione tra agenti (Mordatch e Abbeel, 2018). Questi successi hanno riacceso importanti dibattiti tra gli esperti di linguistica e psicolinguistica, suggerendo che l'acquisizione del linguaggio non sia necessariamente legata alla presenza di strutture cerebrali innate (Linzen e Baroni, 2020).

Le frontiere della ricerca sul *deep learning* si stanno ora spostando verso la simulazione di abilità intellettive ancora più sofisticate, come il ragionamento astratto, l'inferenza logica e l'apprendimento di strutture causali (Bottou, 2014). La maggior parte di questi domini rimane fuori portata per le reti neurali artificiali, che spesso falliscono persino in compiti relativamente semplici usati per testare questo tipo di abilità intellettive nell'umano (Barrett, Hill, Santoro, Morcos e Lillicrap, 2018). Questi limiti potrebbero apparire inspiegabili, considerate le prestazioni ottenute in situazioni che presuppongono straordinarie capacità di ragionamento, come il gioco del Go; tuttavia, la manipolazione di rappresentazioni simboliche è considerata il Sacro Graal dei modelli connessionisti, ed alcuni autori affermano che potrebbe servire un cambio di paradigma per poter implementare questo tipo di abilità nelle macchine (Marcus, 2018). Un caso di studio particolarmente interessante è costituito dall'apprendimento di concetti matematici: nonostante le reti neurali artificiali inizino ad esibire abilità notevoli, per esempio imparando ad integrare funzioni (Lample e Charton, 2019), simulare l'apprendimento della semantica dei simboli numerici rappresenta ancora una sfida aperta (Testolin, 2020).

5. CONCLUSIONI E PROSPETTIVE

La rivoluzione del *deep learning* ha riportato in auge i modelli connessionisti di intelligenza artificiale: attraverso meccanismi di appren-

dimento dominio-generale, le moderne reti neurali imparano a classificare stimoli sensoriali, riconoscere *pattern* e regolarità statistiche, creare rappresentazioni astratte di concetti, elaborare il linguaggio umano e persino scoprire strategie per pianificare azioni in ambienti complessi. Oltre a generare grande entusiasmo tra i ricercatori di IA, questi modelli hanno catturato l'attenzione dell'accademia e dell'industria, grazie alla loro potenziale applicazione nei più svariati ambiti scientifici e tecnologici. L'entusiasmo si è rapidamente diffuso anche tra gli investitori ed il grande pubblico, dando spesso credito a speculazioni futuristiche ed alimentando aspettative poco realistiche. Infatti, sebbene i modelli di *deep learning* siano in grado di superare il Test di Turing in specifici domini, siamo ancora lontani dal poter affermare di aver riprodotto la complessità dell'intelligenza umana nelle macchine (Lake, Ullman, Tenenbaum e Gershman, 2017).

Uno dei principali limiti del *deep learning* riguarda l'efficienza di questi modelli. Addestrare reti neurali di grandi dimensioni è un processo estremamente dispendioso: per ottenere buone prestazioni servono enormi moli di dati – spesso annotati – e molta potenza di calcolo. In termini di risorse energetiche, è stato stimato che la costruzione di un modello per l'elaborazione del linguaggio abbia un impatto ambientale ben superiore a quello annuale di un'automobile (Strubell, Ganesh e McCallum, 2020). Si sta quindi investendo nella ricerca di algoritmi in grado di apprendere con pochi dati (Snell, Swersky e Zemel, 2017) e di generalizzare la conoscenza appresa in modo flessibile, esibendo capacità di «meta apprendimento» (Finn, Abbeel e Levine, 2017). Un'altra direzione promettente consiste nell'implementare questi modelli su hardware neuromorfo, che riduce i consumi riproducendo *in silico* alcune proprietà fisiche delle sinapsi biologiche (Prezioso *et al.*, 2015; Zidan, Strachan e Lu, 2018).

Un ulteriore aspetto critico riguarda l'interpretabilità dei modelli di *deep learning*, che vengono paragonati a «scatole nere» in quanto spesso risultano impenetrabili ad un osservatore esterno (Castelvecchi, 2016). Se impiegati in contesti sensibili, come la diagnosi medica o la guida autonoma, questi algoritmi non solo devono essere in grado di produrre decisioni accurate, ma anche fornire indicazioni precise sul perché una certa decisione è stata presa (Murdoch, Singh, Kumbier, Abbasi-Asl e Yu, 2019). Questo aspetto è reso ancora più critico dalla scoperta di alcune vulnerabilità di questi sistemi, che possono essere soggetti ad attacchi mirati che ne alterano il comportamento introducendo perturbazioni non rilevabili da un umano (Papernot *et al.*, 2016). Problemi etici nascono anche dal possibile uso improprio di questi sistemi, che possono essere utilizzati per generare *fake news* o alterare immagini e video a scopo diffamatorio (Rossler *et al.*, 2019).

In conclusione, non si può negare il forte impatto che le ricerche di intelligenza artificiale stanno producendo nelle società moderne. Questi progressi vanno accolti con entusiasmo, ma ci invitano anche a riflettere sui potenziali sviluppi di queste tecnologie e sulle questioni etiche che ne deriveranno. Un esempio emblematico è costituito dagli scenari che potrebbero richiedere ad una macchina di esprimere giudizi morali, la cui validità dipende dal contesto sociale e culturale di chi prende la decisione (Awad *et al.*, 2018). Tali questioni andranno affrontate con estrema cautela e responsabilità, unendo tutti i saperi e le competenze interdisciplinari che da sempre contribuiscono allo sviluppo delle scienze cognitive.

BIBLIOGRAFIA

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I. (2018). The Moral Machine experiment. *Nature*, *563*, 59-64, doi: https://doi.org/10.1038/s41586-018-0637-6.

Banino A., Barry C., Uria B., Blundell C., Lillicrap T., Mirowski P., Pritzel A., Chadwick M.J., Degris T., Modayil J., Wayne G., Soyer H., Viola F., Zhang B., Goroshin R., Rabinowitz N., Pascanu R., Beattie C., Petersen S., ... Kumaran D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*, 429-433, doi: https://doi.org/10.1038/s41586-018-0102-6.

Baroni M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375* (1791), doi: https://doi.org/10.1098/rstb.2019.0307.

Barrett D.G.T., Hill F., Santoro A., Morcos A.S., Lillicrap T. (2018). Measuring abstract reasoning in neural networks. *ArXiv*, doi: https://arxiv.org/abs/1807.04225.

Bashivan P., Kar K., DiCarlo J.J. (2019). Neural population control via deep image synthesis. *Science*, *364* (6439), doi: https://doi.org/10.1126/science.aav9436.

Bassett D.S., Gazzaniga M.S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences*, *15* (5), 200-209. doi: https://doi.org/10.1016/j.tics.2011.03.006.

Bengio Y., Simard P., Frasconi P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*, 157-166.

Berkes P., Orbán G., Lengyel M., Fiser J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331* (6013), 83-87, doi: https://doi.org/10.1126/science.1195870.

Bottou L. (2014). From machine learning to machine reasoning: An essay. *Machine Learning*, *94* (2), 133-149, doi: https://doi.org/10.1007/s10994-013-5335-x.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., ... Amodei D. (2020). Language models are few-shot learners, *ArXiv*, doi: arXiv:2005.14165v4.

Castelvecchi D. (2016). The black box of AI. *Nature*, *521* (7553), 452-459, doi: https://doi.org/10.1038/538020a.
Chapelle O., Schölkopf B., Zien A. (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.
Chomsky N. (1957). *Syntactic structures*. New York: Mouton De Gruyter.
Clark A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-204, doi: https://doi.org/10.1017/S0140525X12000477.
Crick F. (1989). The recent excitement about neural networks. *Nature*, *337* (12), 129-132.
Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. (2009). Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition*, 248-255.
Devlin J., Chang M.-W., Lee K., Toutanova K. (2018). BERT: Pre-training of Deep bidirectional transformers for language understanding. *arXiv preprint 1810.04805v2*.
Elman J.L. (1990). Finding structure in time. *Cognitive Science*, *14* (2), 179-211.
Elman J.L., Bates E., Johnson M., Karmiloff-Smith A., Parisi D., Plunkett K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
Felleman D.J., Van Essen D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1* (1), 1-47.
Finn C., Abbeel P., Levine S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *34th International Conference on Machine Learning, ICML 2017*, *3*, 1856-1868.
Futrell R., Wilcox E., Morita T., Qian P., Ballesteros M., Levy R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, *1*, 32-42, doi: https://doi.org/10.18653/v1/n19-1004.
Gerstner W., Sprekeler H., Deco G. (2012). Theory and simulation in neuroscience. *Science*, *338* (6103), 60-65, doi: https://doi.org/10.1126/science.1227356.
Glorot X., Bengio Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *9*, 249-256, doi: https://doi.org/10.1.1.207.2059.
Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672-2680.
Güçlü U., van Gerven M.A.J. (2014). Unsupervised feature learning improves prediction of human brain activity in response to natural images. *PLoS Computational Biology*, *10* (8), e1003724, doi: https://doi.org/10.1371/journal.pcbi.1003724.
Güçlü U., van Gerven M.A.J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35* (27), 10005-10014. doi: https://doi.org/10.1523/JNEUROSCI.5023-14.2015.
Harnad S. (1990). The symbol grounding problem. *Physica D: Nonlinar Phenomena*, *42* (1-3), 335-346, doi: https://doi.org/10.1016/0167-2789(90)90087-6.

Hassabis D., Kumaran D., Summerfield C., Botvinick M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95* (2), 245-258, doi: https://doi.org/10.1016/j.neuron.2017.06.011.
He K., Zhang X., Ren S., Sun J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778, doi: https://doi.org/10.1109/CVPR.2016.90.
Hinton G.E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11* (10), 428-434.
Hinton G.E., Deng L., Yu D., Dahl G.E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T.N., Kingsbury B. (2012). Deep neural networks for acoustic modeling in speech recognition. *Ieee Signal Processing Magazine*, *November*, 82-97, doi: https://doi.org/10.1109/MSP.2012.2205597.
Hinton G.E., Salakhutdinov R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313* (5786), 504-507, doi: https://doi.org/10.1126/science.1127647.
Karras T., Aila T., Laine S., Lehtinen J. (2018). Progressive growing of GANs for improved quality, stability, and variation. *International Conference on Learning Representations*, 1-26.
Kingma D.P., Ba J.L. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 1-15.
Kingma D.P., Welling M. (2013). Auto-encoding variational bayes. *ArXiv*.
Kriegeskorte N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1* (1), 417-446, doi: https://doi.org/10.1146/annurev-vision-082114-035447.
Krizhevsky A., Sutskever I., Hinton G.E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *24*, 609-616.
Kubilius J., Bracci S., Op de Beeck H.P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12* (4), 1-26, doi: https://doi.org/10.1371/journal.pcbi.1004896.
Kumaran D., Hassabis D., McClelland J.L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20* (7), 512-534, doi: https://doi.org/10.1016/j.tics.2016.05.004.
Lake B.M., Ullman T.D., Tenenbaum J.B., Gershman S.J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, 1-72, doi: https://doi.org/1511.09249v1.
Lample G., Charton F. (2019). Deep learning for symbolic mathematics. *arXiv preprint 1912.01412v1*.
Le Q.V, Ranzato M.A., Monga R., Devin M., Chen K., Corrado G.S., Dean J., Ng A.Y. (2012). Building high-level features using large scale unsupervised learning. *International Conference on Machine Learning*.
LeCun Y., Bengio Y., Hinton G.E. (2015). Deep learning. *Nature*, *521* (7553), 436-444, doi: https://doi.org/10.1038/nature14539.
Lee H., Ekanadham C., Ng A.Y. (2008). Sparse deep belief net models for visual area V2. *Advances in Neural Information Processing Systems*, *20*, 873-880.
Lillicrap T.P., Santoro A., Marris L., Akerman C. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21* (6), 335-346, doi: https://doi.org/10.1038/s41583-020-0277-3.

Linzen T., Baroni M. (2020). Syntactic Structure from Deep Learning. *Annual Reviews of Linguistics*, *7*, 195-212, doi: https://doi.org/10.1146/annurev-linguistics-032020-051035.
Lipton Z.C., Berkowitz J., Elkan C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv*, doi: arXiv:1506.00019.
Liu X., He P., Chen W., Gao J. (2020). Multi-task deep neural networks for natural language understanding. *57th Annual Meeting of the Association for Computational Linguistics*, 4487-4496.
Ma J., Sheridan R.P., Liaw A., Dahl G.E., Svetnik V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling*, *55* (2), 263-274, doi: https://doi.org/10.1021/ci500747n.
Marcus G. (2018). Deep Learning: A critical appraisal. *arXiv*, doi: arXiv:1801.00631.
McCarthy J., Hayes P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & Donald Michie (eds.), *Machine Intelligence 4*. Edinburgh: Edinburgh University Press, pp. 463-502.
McCulloch W.S., Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5* (4), 115-133, doi: https://doi.org/10.1007/BF02478259.
Minsky M., Papert S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.
Mitchell T.M. (1997). *Machine learning*. New York: McGraw Hill.
Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., Graves A., Riedmiller M., Fidjeland A.K., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wierstra D., Legg S., Hassabis D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518* (7540), 529-533, doi: https://doi.org/10.1038/nature14236.
Mordatch I., Abbeel P. (2018). Emergence of grounded compositional language in multi-agent populations. *32nd AAAI Conference on Artificial Intelligence*, 1495-1502.
Müller M. (2002). Computer Go. *Artificial Intelligence*, *134* (1-2), 145-179, doi: https://doi.org/10.1016/S0004-3702(01)00121-7.
Murdoch W.J., Singh C., Kumbier K., Abbasi-Asl R., Yu B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116* (44), 22071-22080, doi: https://doi.org/10.1073/pnas.1900654116.
Nair V., Hinton G.E. (2010). Rectified linear units improve restricted Boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*, *3*, 807-814, doi: https://doi.org/10.1.1.165.6419.
Newell A., Simon H. (1961). Computer simulation of human thinking. *Science*, *134* (3495), 2011-2017.
Owens J., Houston M. (2008). GPU computing. *Proceedings of the IEEE*, *96* (5), 879-899.
Papernot N., McDaniel P., Goodfellow I., Jha S., Celik Z.B., Swami A. (2016). *Practical Black-box attacks against machine learning*, doi: https://doi.org/10.1145/3052973.3053009.
Perruchet P., Pacton S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10* (5), 233-238, doi: https://doi.org/10.1016/j.tics.2006.03.006.

Pinker S. (1999). How the mind works. *Annals of the New York Academy of Sciences*, *882* (1), 119-127.
Prezioso M., Merrikh-Bayat F., Hoskins B.D., Adam G.C., Likharev K.K., Strukov D.B. (2015). Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, *521* (7550), 61-64, doi: https://doi.org/10.1038/nature14441.
Richards B.A., Lillicrap T.P., Beaudoin P., Bengio Y., Bogacz R., Christensen A., Clopath C., Costa R.P., de Berker A., Ganguli S., Gillon C.J., Hafner D., Kepecs A., Kriegeskorte N., Latham P., Lindsay G.W., Miller K.D., Naud R., Pack C. C., … Kording K.P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22* (11), 1761-1770, doi: https://doi.org/10.1038/s41593-019-0520-2.
Rosenblatt F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65* (6), 386-408.
Rossler A., Cozzolino D., Verdoliva L., Ries, C., Thies J., Niessner M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE International Conference on Computer Vision*, *2019 October*, 1-11, doi: https://doi.org/10.1109/ICCV.2019.00009.
Rumelhart D.E., Hinton G.E., Williams R.J. (1986). Learning representations by back-propagating errors. *Nature*, *323* (6088), 533-536.
Rumelhart D.E., McClelland J.L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.
Russell S.J., Norvig P. (2020). *Artificial Intelligence: A modern approach (4th Edition)*. New York: Prentice Hall.
Saxe A.M., McClelland J.L., Ganguli S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, doi: https://doi.org/10.1073/PNAS.1820226116.
Searle J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3* (3), 417-424.
Silver D., Huang A., Maddison C.J., Guez A., Sifre L., van den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., Dieleman S., Grewe D., Nham J., Kalchbrenner N., Sutskever I., Lillicrap T., Leach M., Kavukcuoglu K., Graepel T., Hassabis D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529* (7587), 484-489, doi: https://doi.org/10.1038/nature16961.
Snell J., Swersky K., Zemel R.S. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, doi: https://doi.org/10.12783/dtetr/mcee2017/15746.
Srivastava N., Hinton G.E., Krizhevsky A., Sutskever I., Salakhutdinov R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929-1958.
Stoianov I., Zorzi M. (2012). Emergence of a «visual number sense» in hierarchical generative models. *Nature Neuroscience*, *15* (2), 194-196, doi: https://doi.org/10.1038/nn.2996.
Strubell E., Ganesh A., McCallum A. (2020). Energy and policy considerations for deep learning in NLP. *ACL 2019 – 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *1*, 3645-3650, doi: https://doi.org/10.18653/v1/p19-1355.
Sutskever I., Martens J., Dahl G.E., Hinton G.E. (2013). On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on Machine Learning*, *28* (3), 1139-1147.

Sutskever I., Vinyals O., Le Q.V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 3104-3112, doi: https://doi.org/10.1007/s10107-014-0839-0.
Testolin A. (2020). The challenge of modeling the acquisition of mathematical concepts. *Frontiers in Human Neuroscience*, *14*, 1-9, doi: https://doi.org/10.3389/fnhum.2020.00100.
Testolin A., De Filippo De Grazia M., Zorzi M. (2017). The role of architectural and learning constraints in neural network models: A case study on visual space coding. *Frontiers in Computational Neuroscience*, *11*, 1-17, doi: https://doi.org/10.3389/fncom.2017.00013.
Testolin A., Dolfi S., Rochus M., Zorzi M. (2020). Visual sense of number vs. sense of magnitude in humans and machines. *Scientific Reports*, *10* (1), 1-13, doi: https://doi.org/10.1038/s41598-020-66838-5.
Testolin A., Stoianov I., De Filippo De Grazia M., Zorzi M. (2013). Deep unsupervised learning on a desktop PC: A primer for cognitive scientists. *Frontiers in Psychology*, *4*, 251, doi: https://doi.org/10.3389/fpsyg.2013.00251.
Testolin A., Stoianov I., Sperduti A., Zorzi M. (2016). Learning orthographic structure with sequential generative neural networks. *Cognitive Science*, *40* (3), 579-606, doi: https://doi.org/10.1111/cogs.12258.
Testolin A., Stoianov I., Zorzi M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, *1* (9), 657-664, doi: https://doi.org/10.1038/s41562-017-0186-2.
Testolin A., Zorzi M. (2016). Probabilistic models and generative neural networks: towards an unified framework for modeling normal and impaired neurocognitive functions. *Frontiers in Computational Neuroscience*, *10* (73), doi: https://doi.org/10.3389/fncom.2016.00073.
Testolin A., Zou W.Y., McClelland J.L. (2020). Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental Science*, doi: https://doi.org/10.1111/desc.12940.
Turing A.M. (1950). Computing machinery and intelligence. *Mind*, *59* (236), 433-460.
Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30*.
Vinyals O., Babuschkin I., Czarnecki W.M., Mathieu M., Georgiev P., Oh J., Horgan D., Kroiss M., Danihelka I., Huang A., Sifre L., Cai T., Agapiou J.P., Jaderberg M., Vezhnevets A.S., Leblond R., Pohlen T., Dalibard V., Budden D., … Apps C. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*, 350-354, doi: https://doi.org/10.1038/s41586-019-1724-z.
Wang A., Singh A., Michael J., Hill F., Levy O., Bowman S.R. (2019). GLUE: A Multi-Task benchmark and analysis platform for NLU. *International Conference on Learning Representations*, 1-20.
Watkins C.J.C.H., Dayan P. (1992). Q-learning. *Machine Learning*, *8* (3-4), 279-292, doi: https://doi.org/10.1007/BF00992698.
Xiong H., Rodríguez-Sánchez A.J., Szedmak S., Piater J. (2015). Diversity priors for learning early visual features. *Frontiers in Computational Neuroscience*, *9* (104), doi: https://doi.org/10.3389/fncom.2015.00104.
Yamins D.L.K., DiCarlo J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19* (3), 356-365, doi: https://doi.org/10.1038/nn.4244.

ZIDAN M.A., STRACHAN J.P., LU W.D. (2018). The future of electronics based on memristive systems. *Nature Electronics*, *1* (1), 22-29, doi: https://doi.org/10.1038/s41928-017-0006-8.
ZORZI M. (2006a). Dai neuroni al comportamento: la simulazione dei processi cognitivi con modelli generativi. *Sistemi Intelligenti*, *18* (1), 115-124.
ZORZI M. (2006b). L'approccio computazionale in psicologia cognitiva. *Giornale Italiano di Psicologia*, *33* (2), 229-252.
ZORZI M. (2016). Apprendimento e memoria nelle reti neurali. In V. Girotto, M. Zorzi (a cura di), *Manuale di psicologia generale*. Bologna: Il Mulino.
ZORZI M. (2017). Modelli computazionali e simulazione dei processi neurocognitivi. In P.S. Bisiacchi, A. Vallesi (a cura di), *Il cervello al lavoro. Nuove prospettive in neuropsicologia*. Bologna: Il Mulino.
ZORZI M., TESTOLIN A. (2018). An emergentist perspective on the origin of number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373* (1740), doi: https://doi.org/10.1098/rstb.2017.0043.
ZORZI M., TESTOLIN A., STOIANOV I. (2013). Modeling language and cognition with deep unsupervised learning: A tutorial overview. *Frontiers in Psychology*, *4* (515), doi: https://doi.org/10.3389/fpsyg.2013.00515.

The modern approach to artificial intelligence and the deep learning revolution

Summary. In the past decade artificial intelligence research has achieved impressive results, mostly due to the creation of efficient machine learning algorithms. One of the most promising approaches is constituted by *deep learning*, which allows to build multi-layer artificial neural networks that can autonomously extract knowledge from large-scale data sets. In this review we will discuss the main theoretical and technological progresses underlying these achievements, also focusing on their relevance for psychology and cognitive neuroscience. We will also highlight some of the limits of deep learning models and possible research directions to overcome them.

Keywords: Artificial intelligence, machine learning, deep learning, computational modelling, connectionism, artificial neural networks.

La corrispondenza va inviata a Alberto Testolin e Marco Zorzi, Dipartimento di Psicologia Generale, Università di Padova, Via Venezia 8, 35131 Padova. E-mail: alberto.testolin@unipd.it e marco.zorzi@unipd.it