ANYA HURLBERT - TOMASO POGGIO

RENDERE LE MACCHINE (E L'INTELLIGENZA ARTIFICIALE) IN GRADO DI VEDERE

La visione è qualcosa di più di una capacità sensoriale, è un'intelligenza. Immaginiamo di guardare la televisione: da una luce tremola su di uno schermo a due dimensioni creiamo un mondo tridimensionale di persone, luoghi e cose. I processi mentali che vanno dal *pattern* di luce sulla nostra retina ad un'immagine interna del mondo sono altrettanto «intelligenti» quanto le analisi e interpretazioni che conducono il medico dai sintomi alla diagnosi. Eppure quando ammiriamo il cervello umano, siamo più pronti a lodare i poteri deduttivi di un logico che la capacità di una persona qualunque di riconoscere un viso.

Le capacità visive degli esseri umani sono molto sviluppate. Quasi il cinquanta per cento della neocorteccia — la parte del cervello che è arrivata per ultima nell'evoluzione ed è quindi più tipica dei primati — è dedicata alla visione. La visione gioca un ruolo così importante nella nostra comprensione del mondo che il termine «vedo» significa non solo decodificare segnali luminosi, ma anche cogliere il punto essenziale di un discorso. Perché allora abbiamo tentennato a parlare di intelligenza della visione?

Questo interrogativo ha una particolare rilevanza per i ricercatori in intelligenza artificiale (IA). Lo scopo dichiarato della ricerca in IA è di riprodurre nelle macchine l'intelligenza e allo stesso tempo di comprendere cos'è l'intelligenza. Storicamente, la ricerca in IA si è occupata di capacità avanzate quale il ragionamento, il *problem solving*, il linguaggio. Tutt'ora un'assunzione indiscussa sottostante la ricerca in IA è che la funzione dell'intelligenza è di arricchire l'interazione con il mondo esterno e di promuovere il controllo su di esso. Un'intelligenza isolata, per quanto intelligente nel risolvere i problemi di matematica, non potrebbe raggiungere quello scopo se fosse incapace di percepire o di influenzare il mondo. La robotica, ovvero lo studio di come unire la percezione alle azioni, è quindi una parte cruciale dell'IA.

Viene naturale dividere la ricerca in robotica in due settori principali: la visione artificiale («machine vision») e la robotica («robot movement»). Lo scopo della ricerca sulla visione artificiale è di costruire macchine che vedano e allo stesso tempo di comprendere la visione stessa. Analogamente, la ricerca sui robot cerca non solo di costruire robot che sappiano manipolare oggetti e muoversi nell'ambiente ma anche di capire che cos'è il controllo motorio. La tentazione ovvia in IA è di dare alla visione artificiale il compito di fornire l'input ad una macchina intelligente e di assegnare al robot il compito di eseguire il suo output. Agli inizi la ricerca in IA ha fatto proprio questo e così facendo ha escluso sia la visione che il controllo motorio dall'ambito delle attività intelligenti.

La ragione del precoce esilio della visione artificiale era in parte la stessa ragione per cui diamo per scontata la visione e ne sottovalutiamo i poteri quotidiani: vedere sembra facile e immediato. Ma i progressi della visione artificiale degli ultimi venti anni hanno fatto cadere questa illusione e hanno rivelato un dato che suona ironico. La visione è non solo intelligente ma è più difficile da capire o riprodurre che il più sofisticato dei ragionamenti matematici. Infatti la visione pone problemi così difficili che l'IA oggi è molto più preparata a sviluppare sistemi che potrebbero svolgere le funzioni di medici o avvocati, piuttosto che a costruire robot che potrebbero sostituire i giardinieri o i cuochi.

1. Le promesse della visione

Oggi la visione è parte integrante degli sforzi compiuti dall'IA non solo per la sua complessità e per la sua ovvia utilità, ma anche per le interessanti indicazioni generali contenute nel suo approccio allo studio dell'intelligenza. Recentemente, l'IA tradizionale (usiamo il termine ricerca in IA escludendo la visione artificiale e la robotica) è stata messa sotto accusa da una vecchia critica che ha assunto un nuovo nome, «connessionismo». L'accusa proviene da moderni gestaltisti che ritengono che la dissezione logica dell'intelligenza operata dall'IA tradizionale non può rivelare la vera struttura o capacità dei poteri più profondi della mente.

La critica ha preso di mira quella tradizione di ricerca che ha accettato senza riserve l'ipotesi di Newell e Simon secondo cui una struttura intelligente è un sistema di simboli. L'ipotesi afferma che: «un sistema di simboli ha i mezzi necessari e sufficienti per svolgere un'azione intelligente» [1]. Un sistema di simboli è una «macchina»

[1] Per una più completa definizione di simboli, strutture simboliche, designazione, si veda il capitolo di Newell e Simon (1981).

nata o costruita nel mondo fisico che utilizza simboli per descrivere oggetti nel mondo. Il sistema deve essere in grado di generare nuovi simboli a partire da simboli già esistenti e alla fine il sistema deve essere capace di influenzare oggetti reali per mezzo di simboli che li rappresentino. L'ipotesi giustifica la ricerca che si occupa di stabilire ciò che i computer sono in grado di fare: i computer sono sistemi di simboli e perciò essi sono intelligenti. Secondo l'IA tradizionale studiare ciò che i computer possono fare significa programmarli perché possano fare delle cose. La rivolta contro l'IA tradizionale è in parte contro il modo in cui si programma. L'approccio dell'IA è stato quello di scomporre i problemi in sottoproblemi, affrontandoli uno alla volta e seguendo regole esplicite di potatura dell'albero di soluzioni che si ramificano. I sistemi esperti, che funzionano con regole *ad hoc* generando inferenze da basi di dati organizzati *ad hoc*, sono i prodotti commerciali nati da questo approccio. MYCIN, un sistema basato su conoscenze per le diagnosi mediche, va dal sintomo ad una diagnosi differenziata attraverso l'uso di regole del tipo «se-allora» immagazzinate nella propria base di conoscenze. Guidato dai risultati dei test di laboratorio nella ricerca tra i propri dati di tutti i possibili sintomi patogeni, MYCIN ricaverà l'istruzione di volta in volta da una corretta affermazione del tipo «se-allora»: «Se trovi un certo batterio allora usa una certa medicina». Benché la maggior parte del lavoro non sia ancora giunto a piena maturazione — i medici non hanno ancora utilizzato alcun sistema medico esperto se non come una biblioteca o una bibliografia computerizzata — i sistemi esperti in alcune aree specializzate, quali la configurazione dei computer o la riparazione delle linee di telecomunicazione, si sono rivelati utili.

D'altro canto l'ipotesi di Newell e Simon sostiene anche che poiché gli esseri umani sono intelligenti, essi sono sistemi di simboli. Parte della nuova ondata di critiche verso l'IA è contraria a definire l'uomo un sistema di simboli. Il problema è la parola *simbolo*. In senso stretto, un simbolo potrebbe essere un testo su uno schermo, un insieme di interruttori elettronici, una rete di neuroni, o una corrente elettrica che attraversa la membrana di una cellula. Ma la tradizionale ricerca in IA ha assunto una definizione ancora più stretta: i simboli sono termini astratti che denotano oggetti conosciuti e obbediscono ad una grammatica di poche regole. Come i simboli logici in matematica o le parole nel linguaggio latviano, i simboli sono tenuti a comportarsi secondo regole formali. Newell e Simon hanno indicato il linguaggio di programmazione LISP come

un esempio di un sistema di simboli, anche se l'IA tradizionale sembra averlo considerato l'unico. Secondo l'IA tradizionale, fare ingegneria dell'intelligenza significa manipolare simboli di un linguaggio come il LISP in cui ogni rotella gira secondo le regole della logica.

L'ipotesi che l'intelligenza può essere rappresentata come un sistema di simboli è stata considerata come una scoperta di cruciale importanza per l'informatica. Ma se tutto ciò che i sistemi di simboli fisici possono fare è generare istruzioni per il calcolatore, allora alcuni considerano la scoperta come una sorta di maledizione per l'intelligenza umana. Il General Problem Solver (GPS), uno dei primi programmi che ha cercato di trovare meccanismi generali per risolvere i semplici problemi era costruito in base ai dati ottenuti da persone mentre risolvevano problemi ad alta voce. Ma ci si può chiedere come anche il migliore di questi GPS potrebbe guidare una automobile. Un esperto guidatore non applica coscientemente regole quali: «quando si viaggia ad una velocità superiore ai quaranta km all'ora, inserire la terza marcia» oppure «quando si scala di marcia, premere il pedale della frizione» oppure «quando ci si avvicina ad un casello, sollevare il piede dall'acceleratore». Il guidatore semplicemente guida, eseguendo le azioni necessarie automaticamente. Oppure si consideri un meccanico che cerca una pinza per allentare un bullone. Egli sceglie l'arnese appropriato al compito grazie ad una veloce intuizione, non attraverso una ricerca sistematica tra gli strumenti disponibili. Gli oppositori dell'IA tradizionale argomentano che nessun sistema esperto potrebbe manipolare i pezzi di una scacchiera in maniera altrettanto intelligente e veloce quanto un maestro di scacchi, a meno che il sistema non sia infinitamente grande e veloce.

Le critiche rivolte da molti all'IA tradizionale si cristallizzano nel connessionismo. I connessionisti sostengono che i tratti caratteristici e più importanti dell'intelligenza umana sono, tra gli altri, il pensiero associativo e la capacità di generalizzare dagli esempi. Essi sostengono che queste caratteristiche non sono rispecchiate dalle procedure di ricerca seriale e dalle strutture ad albero dei sistemi esperti dell'IA. Al contrario, essi sostengono, l'intelligenza emergerà solo da un hardware speciale che riproduca il parallelismo del cervello umano in cui un enorme numero di cellule interconnesse affrontano parti diverse dello stesso compito nello stesso tempo.

Prima di dar ragione alle argomentazioni sollevate contro l'approccio tradizionale in IA, dobbiamo avere ben chiaro dove e

perché esso fallisce. Proprio quelle cose che l'IA all'inizio ha con più decisione chiamato intelligenti — il ragionamento matematico, la comprensione del linguaggio, la logica astratta — sono le cose che i sistemi esperti fanno meglio. In effetti le cose che consideriamo mentalmente più difficili sono semplicemente le cose di cui siamo più coscienti, e ne siamo più coscienti in quanto sono le cose che abbiamo imparato più tardi nella evoluzione e conseguentemente facciamo meno bene. I sistemi esperti hanno una buona probabilità di superarci in questi compiti coscienti e difficili e non dovrebbero essere prematuramente incolpati a causa della lentezza del progresso tecnologico. La debolezza intrinseca dei sistemi esperti medici non risiede nella loro presente incapacità di abbracciare l'enorme dominio di conoscenze mediche — col tempo probabilmente essi ce la faranno — ma nella loro incapacità di riprodurre l'arte della medicina. Nonostante che un odierno sistema esperto medico possa essere più bravo di un internista (che non abbia dormito a sufficienza) nel ricordare un intero elenco di test di laboratorio, l'internista ha sempre un vantaggio sul computer nel percepire una situazione di infelicità personale come la causa della perdita di appetito del paziente.

Da un lato i lampi di intuizione e le folgorazioni istantanee e dall'altro ordinarie capacità percettive quali il riconoscimento del linguaggio parlato: queste sono le capacità della mente che l'IA tradizionale ha difficoltà a modellare. Esse sono le attività mentali di cui siamo meno consci e più capaci. L'evoluzione ha impiegato millenni per perfezionare tali talenti inconsci, ed è più che logico pensare che per riprodurre questi talenti sia necessario impiegare tattiche che non siano quelle usate esclusivamente (e in modo relativamente inefficiente) dalla attività cerebrale più cosciente.

La visione è probabilmente il più intelligente dei meccanismi mentali sottratti alla coscienza. Eppure le metodologie per studiare l'intelligenza sviluppate dalla visione artificiale non sono né recondite né magiche né prendono le distanze dalla logica o dall'intuizione. Nella visione artificiale, il meglio della tradizione dell'IA si incontra con il meglio del connessionismo per costruire una scienza che si differenzia da entrambe.

La visione ha derivato la sua filosofia dalla stessa fonte usata dalla IA tradizionale. Alla base dell'ipotesi dell'intelligenza come sistema di simboli è l'idea che i simboli sono oggetti arbitrari indipendenti dalle macchine sottostanti e senza significato fino a che non gliene viene assegnato uno. La visione artificiale ha trasformato questa idea in quello che definiamo il suo dogma centrale: l'intelli-

genza può essere studiata come un sistema astratto di information-processing, indipendentemente dalla macchina su cui gira. La visione artificiale ha sostenuto il dogma in un'unica direzione, l'approccio computazionale. L'approccio computazionale descrive esattamente quali informazioni un sistema riceve e quali informazioni produce in uscita e cerca di trovare un algoritmo che trasformi l'input nell'output. Per i sistemi della visione, naturali o artificiali, un tale computo è obbligatoriamente vincolato dalle proprietà dell'ambiente, dell'occhio e della luce che viaggia tra di loro. La strada che porta alla giusta computazione è scoprire e rispettare questi vincoli. La visione artificiale ha trasformato la ricerca dei vincoli in una scienza del mondo naturale.

2. Il problema della visione

Per comprendere la forza dell'approccio della visione artificiale, è necessario innanzitutto rendersi conto della difficoltà dei problemi che esso affronta. Un fatto divertente e un po' apocrifo appartenente alla biografia di Marvin Minsky, uno dei padri fondatori dell'IA, illustra quanto questa presa di coscienza sia difficile. Circa venti anni fa, egli assegnò ad uno studente un problema, da affrontare durante l'estate: collegare una cinepresa ad un computer e fare in modo che il computer descrivesse ciò che vedeva. Questo progetto estivo si è espanso in una vera e propria disciplina di ricerca e nonostante gli enormi progressi della visione artificiale, il problema non è stato ancora risolto.

Gran parte dei progressi sono derivati dalla faticosa ricerca su come formulare in modo corretto il problema: che cosa fa la visione? La risposta più semplice è che la visione trasforma i segnali di luce in rappresentazioni interne di quegli oggetti che li trasmettono. La visione umana inizia da un *pattern* di luce (un'immagine) bidimensionale su ciascuna retina e termina con una descrizione di oggetti tridimensionali in termini della loro forma, colore, tessitura, grandezza, distanza e movimento. Il primo ostacolo è la stessa immagine retinica: essa contiene un enorme, quasi inimmaginabile quantità di informazione. Sulla retina sono disposti più di 100 milioni di fotorecettori. La lente dell'occhio mette a fuoco la luce sulla retina in modo tale che il mondo tridimensionale viene appiattito e messo in diretta corrispondenza con il mosaico dei fotorecettori, dove ogni fotorecettore corrisponde ad un particolare punto nel campo visivo. La quantità di luce che colpisce un singolo fotorecettore è determinata dalla quantità di luce riflessa da un qualsiasi

oggetto che occupa il punto corrispondente nel campo visivo. A sua volta la quantità di luce che un oggetto riflette dipende, tra le altre cose, dalla quantità di luce che lo colpisce (che dipende per esempio da quanto esso è vicino ad una lampada, o se esso giace all'ombra di un altro oggetto) e dal materiale di cui esso è fatto (metallo riflettente, velluto nero, garza trasparente, pelle vegetale).

Se, per ogni fotone colpito, ogni fotorecettore depositasse un granello scuro di argento su di un negativo dietro il nostro occhio, potremmo scoprire immagini del mondo dalla macchina fotografica Polaroid costruita dentro di noi. Ma la retina non è così ubbidiente e l'immagine che manda al cervello è più astrusa. Ad ogni istante, l'immagine è un insieme ordinato di segnali elettrochimici, dove la grandezza di ogni segnale è proporzionale alla quantità di luce che colpisce il fotorecettore che la trasmette. In ogni secondo il cervello deve elaborare circa un centinaio di queste immagini, poiché l'occhio vaga in un mondo che muta continuamente. Così invece di registrare fotografie statiche, la retina trasmette un flusso di informazioni visive dinamiche.

Nella visione artificiale, l'immagine retinica è tradotta in una configurazione di punti bidimensionale. Ogni punto rappresenta una minuscola suddivisione dell'immagine e contiene un numero che rappresenta la grandezza del segnale trasmesso da un singolo sensore di luce (oppure in altre parole, l'intensità di luce che colpisce il sensore). Il compito della macchina è di eseguire operazioni matematiche su questa matrice di numeri per trasformarlo in una matrice più ricca di significato: per esempio, la matrice che codifica esplicitamente le distanze degli oggetti dalla cinepresa, o la matrice che assegna un colore a ciascun materiale.

Un'immagine tipica della visione artificiale può essere composta da un milione di punti. Ogni punto è codificato da un numero a otto cifre (*bit*). La quantità totale di informazione, sebbene molto minore di quella contenuta in un'immagine retinica umana, è pur tuttavia ben otto milioni di *bits*. Si moltiplichi questo numero per il numero di immagini al secondo che una telecamera deve inviare per poter imitare l'occhio umano e la velocità di trasmissione dell'informazione sale ad almeno centinaia di milioni di *bits* al secondo. Così perfino l'operazione matematica più semplice che la macchina esegue sul flusso di immagini richiede miliardi di moltiplicazioni e addizioni al secondo. Questa operazione potrebbe essere eseguita da più di un milione di personal computers in funzione tutti allo stesso tempo.

Paradossalmente, il vero problema della visione è che tutta

l'informazione contenuta in un'immagine non è mai sufficiente. Nella proiezione del mondo tridimensionale sulla superficie bidimensionale, viene perduta troppa informazione, cosicché il valore di ciascun *pixel* di una grande configurazione è altamente ambiguo. Si consideri il tipico errore di un fotografo alle prime armi: quello di situare il proprio soggetto davanti ad un palo del telefono per ritrarre gli alberi e il verde del paesaggio su entrambi i lati. Nel proiettare la scena tridimensionale su una pellicola bidimensionale si è perduta l'informazione cruciale circa la profondità. La grandezza del palo telefonico è un indice che può essere interpretato in almeno due modi: potrebbe essere un palo telefonico grosso ma distante oppure una sottile asta di legno che sporge dalla testa del soggetto. L'ambiguità relativa alla profondità è la più ovvia; un'altra ambiguità è nell'interpretazione della luminosità o oscurità. Se i valori di intensità di un gruppo di punti sono molto più grandi di quelli in un gruppo vicino, la disparità tra essi può essere in diversi modi. Può darsi che un'ombra attraversi un unico pezzo di carta, dando così l'illusione di un confine tra un foglio di carta chiaro e uno scuro, oppure può darsi che un foglio di carta bianca giaccia accanto ad un foglio nero. A questo riguardo, i soli numeri non sono indicativi.

La visione artificiale formula il problema nel seguente modo: data una matrice ordinata di valori di intensità, trovare il corrispondente insieme tridimensionale di oggetti e superfici cha l'ha prodotto. Definito così, il problema appare insolubile. Le caratteristiche che devono essere recuperate — i colori, i gradienti, le tessiture, le relazioni spaziali tra gli oggetti, la posizione e il colore della sorgente di luce — sono mescolati in una matrice di numeri che scoraggia qualsiasi tentativo di interpretazione. Eppure la visione artificiale ha fatto grandi progressi da quando lo studente di Minsky venne alle prese con il problema. Ciò che è emerso negli ultimi quindici anni è una concezione della struttura della visione che ci permette di scomporre il problema in parti che sono trattabili e indipendenti tra loro. Innanzitutto la visione può essere chiaramente divisa in due stadi: la *visione di basso livello* (che determina dove sono gli oggetti) e la *visione ad alto livello* (che determina che cosa sono gli oggetti). In secondo luogo, la visione di basso livello può essere studiata come un insieme di moduli visivi separati, dove ognuno estrae dall'immagine un tipo specifico di informazione visiva. Trovando la giusta prospettiva, la visione artificiale è andata al di là dei limiti dell'IA tradizionale e ha costruito una sua propria scienza: la scienza dell'ottica inversa («inverse optics»).

3. Uno sguardo alla visione artificiale

Un'idea chiara sulla struttura della visione non è emersa immediatamente. In effetti, i primi tentativi degli scienziati che si occupavano della visione, tentavano di risolvere comunque il problema, in un modo o nell'altro: veniva utilizzata qualsiasi strategia che potesse risolvere l'ambiguità nell'immagine, senza tener conto di quanto fosse limitato e circoscritto il compito di interesse. Questo approccio ha prodotto *sistemi esperti* nella visione: programmi per l'interpretazione delle immagini che si basavano su una riserva di conoscenze specifiche, fatte «su misura», per alimentare l'applicazione di regole fatte «su misura». Un esempio estremo di un programma del genere intraprende il compito di localizzare un telefono su una scrivania. Tra le assunzioni che il programma fa per vincolare la ricerca, ci sono quelle che il telefono è nero, e che si trova ad un'altezza e una distanza fissata dalla macchina che sta facendo la fotografia. La ricerca eseguita dal programma è relativamente semplice: esamina la giusta fila di *pixels*, tenendo presente che un raggruppamento di valori bassi dei *pixels* significano la presenza di un oggetto scuro, poi controlla che la grandezza del raggruppamento corrisponda alla grandezza che un telefono dovrebbe avere quando osservato da una distanza specificata. Se l'assunzione è vera, il programma funziona bene, ma fallisce irrimediabilmente se il telefono è bianco o se è stato spostato dalla scrivania sul pavimento.

Le soluzioni *ad hoc* prodotte da questo approccio lasciano poco spazio allo sviluppo o applicazione di principi scientifici generali. Invece di delineare chiaramente i passi che il sistema visivo deve fare mentre passa dall'immagine alla rappresentazione degli oggetti, i primi programmi di visione mescolavano casualmente i livelli, sfruttando decisioni ad alto livello per analizzare informazioni a basso livello registrate da insiemi ordinati di valori di *pixel*. Come nel caso di nuovi sistemi esperti in altri domini, i primi sistemi di visione artificiale potevano funzionare solo in ambienti circoscritti e artificiali. Le loro tecniche per affrontare circoscritti micromondi non potevano essere generalizzate nel caso di circostanze imprevedibili in cui gli esseri umani e gli animali più primitivi sfruttano la visione in modo tanto efficace.

3.1. *La visione di basso livello*

L'insuccesso ottenuto dai primi tentativi di affrontare il problema, ha favorito un crescente interesse per la visione di basso livello.

I pionieri (gli artefici) di questo cambiamento si sono resi conto che, senza una teoria sulla descrizione dell'immagine abbastanza generale da guidare l'interpretazione di una immagine qualsiasi, la visione artificiale sarebbe condannata ad una infinta ripetizione di espedienti, dove ciascuno è concepito più intelligentemente del precedente, ma nessuno di essi sarebbe capace di raggiungere la flessibilità del sistema visivo umano. La strada per questa teoria generale passa attraverso una scienza sul mondo, non una scienza della mente, cioè una analisi completa della fisica e dell'interazione tra la luce, l'occhio e l'oggetto. Un impegno tenace negli ultimi quindici anni ha prodotto uno schema per la visione di basso livello il cui scopo principale è di costruire una mappa della scena che registri, per ogni superficie nell'immagine, la sua distanza e orientamento rispetto all'osservatore. Cioè, la visione di basso livello tenta di trasformare l'iniziale insieme ordinato di valori di *pixel* in un altro insieme di numeri che esplicitamente raggruppi parti di un'immagine che appartengono tutte ad uno stesso oggetto e dica dove ogni oggetto si trovi, relativamente all'osservatore. Gli stessi «oggetti» della visione a basso livello sono solo parti di oggetti, di facciate visibili e di lati più grandi appartenenti a intere configurazioni ancora non riconosciute. Questa mappa, chiamata da David Marr rappresentazione a due dimensioni e mezzo («2 e 1/2D *sketch*»), fornisce la base da cui partire per perseguire gli scopi successivi della visione a basso livello: assegnare ad ogni oggetto di un'immagine una forma, un colore, una tessitura, una velocità, e una direzione del moto. Si potrebbero poi tracciare diverse mappe dove ognuna registri un diverso tipo di informazione visiva, e ciascuna potrebbe essere sovrapposta nel *register* con la rappresentazione a due dimensioni e mezzo. Allineando in questo modo la mappa del colore con il «2 e 1/2 D *sketch*», un ipotetico supervisore potrebbe per esempio, leggere la distanza, l'orientamento e il colore di ogni singola superficie nell'immagine.

La visione di basso livello deve compiere due funzioni durante il computo delle proprietà visive di ogni superficie dell'immagine: 1) ridurre la grande quantità di informazioni dell'immagine solamente alle caratteristiche importanti, e 2) ridurre il grado di ambiguità delle informazioni stesse.

3.2. *Estrazione dei contorni*

Il passo iniziale per fare la prima mappa è di delineare i confini tra regioni distinte nell'immagine. Il processo di *estrazione* dei

contorni, che è il più studiato dalla visione a basso livello, ha questo compito. Esso viene svolto tracciando i confini che separano raggruppamenti di valori di *pixel* significativamente diversi. È facile accettare il fatto che i rilevatori dei confini esistano al livello più basso del sistema visivo umano data la tendenza dei nostri organi sensoriali a preferire i cambiamenti agli stati stabili. Infatti il nostro sistema visivo sembra essere organizzato soprattutto per rilevare i mutamenti, piuttosto che i valori assoluti, dei segnali di luce. Immaginate cosa vedreste di uno spazio uniformemente bianco che attraversa il vostro campo di vista: certamente non vedreste molto. L'immagine è noiosa perché non ci sono cambiamenti dell'intensità del segnale che attraversa lo spazio. Ma il nostro sistema visivo non cerca solo di rilevare cambiamenti spaziali dell'intensità. Se la percezione della scena complessa che vedete sollevando gli occhi da questa pagina, dovesse rimanere perfettamente immobile sulla vostra retina, essa perderebbe molto presto i suoi tratti distintivi e diventerebbe come lo spazio bianco. Sparirebbe perché non ci sarebbero cambiamenti *temporali* del segnale di luce registrato da ogni fotorecettore.

Le cellule retinali a cui i fotorecettori mandano i segnali, sono particolarmente organizzate per rilevare i cambiamenti della luce attraverso lo spazio e il tempo. Le cellule paragonano continuamente gli attuali valori di intensità con i valori di intensità registrati l'istante precedente e mandano risposte transitorie che codificano i cambiamenti dei valori di intensità invece che gli stessi valori di intensità. Se i valori d'intensità non cambiano, la risposta delle cellule scende a zero e l'immagine sfuma. I cambiamenti di intensità nello spazio e nel tempo sono interconnessi; mentre l'occhio si gira senza sosta nella sua cavità, i contorni nello spazio si spostano attraverso i fotorecettori e vengono convertiti in contorni nel tempo.

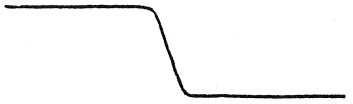
Nella visione artificiale il rilevatore di contorni cerca di trovare i cambiamenti di intensità, in quanto essi sono i tratti fondamentali più importanti e in questo modo il primo compito viene eseguito riducendo l'enorme quantità di informazione dell'immagine. La difficoltà nel rilevare i contorni, sta non nel trovare i cambiamenti di intensità ma nello scartare quelli non importanti. In un'immagine qualsiasi, piccoli cambiamenti di intensità tra raggruppamenti di *pixels* sono frequenti. Persino l'immagine di un foglio di carta bianco non avrebbe lo stesso valore per ogni *pixel*, a meno che il bianco della carta fosse assolutamente uniforme, la luce che illumina la carta fosse esattamente la stessa ad ogni punto, e i sensori che

fedelmente registrano l'immagine, registrassero la quantità di luce che li colpisce. In realtà queste condizioni non sono mai soddisfatte: il bianco del foglio è macchiato di impurità, la luce colpisce il foglio con una certa angolazione, gli fa ombra e i sensori, continuamente bombardati da fotoni casuali, trasmettono un segnale disturbato. Un buon rilevatore di contorni dovrebbe trascurare le casuali oscillazioni d'intensità del segnale proveniente dalla superficie del foglio e rilevarne solo i veri contorni. Teoricamente, la rilevazione dei contorni dovrebbe generare una linea che disegna la scena (traccia i confini della) cogliendo i contorni fisici e i confini degli oggetti e lasciando bianche le superfici all'interno degli stessi confini. Eppure progettare un rilevatore di contorni che faccia questo non è un compito semplice.

Si sono compiuti grandi sforzi per costruire efficienti rilevatori di contorni. Il problema nel progettarli è di controbilanciare tra l'appiattimento del disturbo dell'intensità del segnale da un lato e il cogliere tutti i contorni significativi. In due passaggi i buoni rilevatori di contorni imitano il modo in cui le cellule retiniche umane fanno entrambe le cose. Innanzitutto, il rilevatore sfuma l'immagine sommando ad ogni valore di *pixel* la media dei valori di *pixel* in un raggruppamento che lo circonda. In secondo luogo, calcola la differenza tra il nuovo valore di *pixel* e una media dei valori di *pixel* di un raggruppamento più grande che lo circonda e assegna il valore di quella differenza ad un *pixel* centrale [2]. Queste due operazioni, di *sfumatura* o *filtraggio* e di *differenziazione,* assicurano che i piccoli cambiamenti di intensità del segnale vengano trascurati (i piccoli cambiamenti diventano differenze di valore ancora più piccole dopo il filtraggio) e i grandi mutamenti vengono favoriti. L'effetto di queste azioni sulla seguente linea di scorrimento (*scan line*) su di una immagine

la trasformerebbe in un segnale più chiaro.

[2] A dire il vero, il modo in cui si determina la differenza è un po' più complicato, ma il risultato è simile. Per una descrizione più lunga dell'estrazione dei contorni si veda Berthold e Horn (1986).



Si può controllare in vari modi il valore al di sotto del quale i cambiamenti di intensità sono scartati e sopra il quale essi sono mantenuti: si modifica la grandezza del raggruppamento di *pixels* su cui è calcolata la media, si modificano i pesi che hanno in media i valori di *pixel*, e si conservano solo quei valori sfumati e differenziati che sono al di sopra di una soglia fissata in precedenza. Se il valore è troppo basso può accadere che vengano rilevati contorni che non rappresentano contorni fisici nella scena, ma se il valore è troppo alto, alcuni contorni reali possono essere scartati insieme al disturbo. Coloro che hanno lavorato per lo sviluppo dei rilevatori di contorni per la visione artificiale hanno lavorato molto nella progettazione di filtri e differenziatori che favoriscono più efficacemente i contorni e rendono lisce le superfici. Tuttavia, nessun rilevatore di contorni può essere perfetto. Le linee che vengono tracciate in questo stadio inizale della visione a basso livello contengono molte linee estranee.

3.3. *I vincoli naturali*

Lo scopo della visione di basso livello di ridurre l'ambiguità delle immagini, ha promosso l'idea di *vincoli naturali.* Fin dall'inizio era ovvio che i vincoli fossero necessari per ridurre il numero di possibili interpretazioni di un'immagine, ma non era ovvio quale avrebbe dovuto essere la fonte di questi vincoli. Quindi, *vincoli innaturali* come quelli contenuti nel programma per il riconoscimento del telefono sono stati considerati come buoni pari di altri possibili vincoli. La fisica delle immagini ha spinto gli scienziati della visione a rivolgersi al mondo fisico per trovare dei vincoli, cioè per formulare assunzioni che quasi sempre sarebbero vere circa le proprietà delle luci, delle superfici, e le geometrie: i vincoli naturali. I vincoli naturali riducono il numero delle interpretazioni di un'immagine escludendone alcune in quanto fisicamente impossibili.

3.4. *Un algoritmo stereoscopico*

Uno dei compiti più difficili della visione a basso livello è stato quello di trovare e formulare esattamente i giusti vincoli naturali. Per la soluzione del problema della percezione della profondità i vincoli naturali appaiono di primaria importanza.

Con un semplice esperimento potete rendervi conto del problema della percezione della profondità e di come gli esseri umani lo semplificano durante la visione. Ponete un dito a diversi centimetri dal vostro viso e guardatelo tenendo l'occhio destro chiuso. Ora aprite l'occhio destro e chiudete il sinistro. Il vostro dito sembra spostarsi a destra. Ora tenete il dito lontano dal vostro viso il più possibile e alternate gli sguardi degli occhi allo stesso modo. Il dito si sposterà leggermente a sinistra. Ecco la spiegazione: poiché i due occhi occupano una posizione leggermente diversa sul vostro viso, l'immagine del mondo colpisce l'occhio in un modo lievemente diverso.

La diversità nella posizione del dito nelle due immagini è dovuta alla disparità binoculare del dito e, come mostrato dal vostro esperimento, è in diretta relazione con la distanza tra il dito e il vostro viso. Più il dito è lontano, più piccola è la sua disparità. La visione stereoscopica si avvantaggia di questo dato usando la disparità per stimare le relative profondità delle superfici in una scena, ricreando così oggetti solidi tridimensionali. Per calcolare la disparità binoculare di un particolare punto della scena, il sistema visivo deve in primo luogo identificare quali *pixels* nelle due immagini corrispondono allo stesso punto. Questo problema della corrispondenza ha troppe soluzioni: ciascun *pixel* dell'immagine di sinistra potrebbe corrispondere ad ognuno dei tanti *pixels* nell'immagine destra con valori di *pixels* simili. È necessario introdurre vincoli naturali che eliminano tutte le corrispondenze tranne quelle fisicamente corrette.

Nel ricercare i vincoli che siano abbastanza potenti e generali da risolvere il problema, emergono due assunzioni circa la realtà quotidiana. Il vincolo della *unicità* incorpora il fatto che la maggior parte delle superfici non sono trasparenti; stabilisce la regola che ad ogni *pixel*, per esempio nell'immagine dell'occhio sinistro, possa venir assegnata una ed una sola profondità (questo vincolo non risulterebbe vero se la maggior parte delle scene fosse vista attraverso molti strati di vetro). Il vincolo della *continuità* esprime il dato che la maggior parte delle superfici sono lisce, così che se ad un punto in una immagine viene assegnata una certa distanza dall'oc-

chio, distanze simili verranno assegnate ai punti vicini. Solo se il mondo fosse composto da una marea di punti sconnessi a profondità casuali ci aspetteremmo di trovare disparità estremamente diverse tra i *pixels*.

Una volta trovati i vincoli, come li utilizziamo? Marr e Poggio (1976) li hanno inglobati in un algoritmo stereoscopico che, nella sua ultima versione, inizia da un'immagine destra e una sinistra, ognuna delle quali è stata filtrata e differenziata da un rilevatore di contorni e produce un'immagine finale in cui ogni valore di *pixel* rappresenta una profondità. Per ogni *pixel* nell'immagine sinistra, l'algoritmo computa la disparità con ognuna delle sue possibili corrispondenze con l'immagine destra. Per ciascuna disparità di ciascun *pixel*, l'algoritmo allora conta il numero di corrispondenze dei *pixels* vicini cha producono la stessa disparità. La disparità con i numeri di «voti» più alti vince. Il vincolo della continuità assicura che il metodo della votazione sia giusto — se una disparità è sostenuta da molte corrispondenze vicine, è probabile che sia quella giusta per una superficie liscia — e il vincolo della unicità impone che ci sia solo un vincitore.

Il semplice metodo di implementare vincoli naturali produce risultati realistici per la maggior parte delle immagini. A differenza dei suoi predecessori ai tempi delle soluzioni *ad hoc*, l'algoritmo stereoscopico non fa affidamento su segnali di alto livello che indicano, per esempio, la punta del naso nell'immagine sinistra e poi risolvono il problema della corrispondenza riconoscendo la punta del naso nell'immagine destra. Gli esseri umani riescono a vedere la profondità negli stereogrammi di punti casuali[3], dimostrando il notevole fatto che noi, come l'algoritmo stereoscopico, non abbiamo bisogno di riconoscere gli oggetti prima di poterli vedere in profondità.

[3] Gli stereogrammi di punti casuali sono immagini tridimensionali generate mostrando all'occhio destro un'immagine di punti neri disposti casualmente su uno sfondo bianco e mostrando all'occhio sinistro un'altra immagine. I *patterns* di punti nelle duc immagini sono identici fatta eccezione per la parte centrale di punti nell'immagine sinistra che è leggermente spostata rispetto alla corrispondente posizione nell'immagine destra. Per l'occhio quel leggero spostamento equivale ad una differenza di profondità tra la parte centrale e il suo sfondo così che quest'ultima sembra rimanere al di sopra o al di sotto dei punti circostanti.

3.5. *L'ottica inversa*

Lo studio delle proprietà dell'ambiente che interagiscono con la visione e il soddisfacimento dei vincoli naturali, ha spinto la ricerca sulla visione di basso livello a sviluppare una nuova scienza: lo studio delle immagini speculari («inverse optics»). Così come l'ottica è la fisica delle formazioni delle immagini bidimensionali a partire da scene tridimensionali, allo stesso modo l'ottica inversa è la fisica del processo di ricostruzione di scene tridimensionali a partire da immagini bidimensionali. Così come gli artisti del Rinascimento seguendo le regole della prospettiva lineare, hanno riprodotto su un foglio bidimensionale confini e contorni tridimensionali, così i moderni scienziati della visione stanno imparando a decodificare immagini bidimensionali scoprendo le nuove regole dell'ottica inversa. L'ottica inversa è una scienza dei problemi impossibili. L'informazione fornita dai punti di un'immagine è insufficiente e la soluzione potrebbe non essere unica né ben definita. Può essere molto difficile per l'IA tradizionale definire precisamente un problema, ma una volta definito la soluzione si trova in modo unico e diretto (benché essa richieda molto tempo per essere realizzata). Nell'ottica inversa è vero il contrario: è facile formulare i problemi ma è molto difficile risolverli. Uno dei maggiori progressi compiuti dalla visione di basso livello è di aver compreso che questi problemi impossibili appartengono ad una classe di problemi, chiamati tecnicamente problemi malposti che è stata studiata a lungo in matematica. Poiché i matematici hanno già sviluppato tecniche adatte ad affrontare la forma generale dei problemi malposti, gli scienziati della visione possono utilizzare le stesse tecniche per risolvere problemi particolari di ottica inversa. Queste tecniche vanno sotto il nome di *teoria della regolarizzazione* (Poggio, Torre e Koch 1985).

La teoria della regolarizzazione fornisce uno schema in cui trovano posto, una volta scoperti, i vincoli naturali. Lo schema è lo stesso per tutti i problemi della visione di basso livello, e indica una soluzione generale che può essere adattata ad ogni problema. Quindi la difficoltà dei cosiddetti problemi «non proprio impossibili» della visione sta quasi esclusivamente nel trovare i giusti vincoli naturali poiché una volta identificati, esiste un metodo formale per renderli operativi.

Ci sono tre operazioni di carattere generale che contraddistinguono le strategie della visione di basso livello. La prima è di isolare il problema dagli altri problemi della visione; cioè se lo scopo è di recuperare la mappa dei colori delle superfici dell'immagine, non è

necessario calcolare la disparità binoculare (i nostri neuroni suddividono i compiti in modo simile: un neurone sensibile alla selezione del colore generalmente non si preoccupa molto della disparità). La seconda operazione identifica i vincoli naturali che governano il problema. L'ultimo compito è di utilizzare un algoritmo che funzioni per i vincoli naturali, possibilmente usando la teoria della regolarizzazione.

3.6. *La visione di alto livello*

Un «mobot» corre nel corridoio del laboratorio di IA al Mit, fermandosi quando ha il sospetto che qualcosa di solido sia troppo vicino per continuare indisturbato. Può seguire un muro o invertire il percorso, muovendosi come se potesse vedere. Ma i suoi sensori infrarossi trasmettono solo i messaggi più importanti: se c'è o meno qualcosa. Esso non è in grado di dire che cos'è quel qualcosa.

Sebbene molto più progredito delle immagini reali del mobot, la rappresentazione a due dimensioni e mezzo generata dalla visione di basso livello, è ben lontana dal dire che cosa sono gli oggetti. La sua funzione è quella di dire dove sono gli oggetti. La funzione della visione di alto livello è di determinare cosa sono gli oggetti. Se la rappresentazione a due dimensioni e mezzo isolasse ogni oggetto di un'immagine dagli altri, il compito di riconoscere gli oggetti dalle caratteristiche del colore, forma, trama, e così via, non sarebbe molto difficile. Ma è proprio tracciare i confini tra oggetti separati ciò che la visione di basso livello, almeno come implementata sulle macchine, non riesce ancora a fare. Gli oggetti di un'immagine che la visione di basso livello coglie sono al massimo parti di oggetti, superfici continue con un solo colore, piccoli gonfiori scuri in una superficie altrimenti uniforme, ma non gli oggetti stessi. La difficoltà non è nel trovare i confini tra le diverse regioni di una stessa immagine, ma nel determinare quali sono le regioni che servono a distinguere e classificare gli oggetti.

I compiti della segmentazione dell'immagine (ritagliare l'immagine in regioni che probabilmente corrispondono a differenti oggetti) e del riconoscimento dell'oggetto (far corrispondere queste regioni significative ad oggetti classificati nella memoria) sono stati al centro di un intenso lavoro nel campo della visione di alto livello. Nonostante questo anche i programmi più sofisticati per il riconoscimento degli oggetti pongono un eccesso di requisiti sull'immagine

che essi elaborano. I programmi richiedono che gli oggetti da riconoscere siano stati precedentemente localizzati con esattezza nell'immagine. Così una volta saputo dove guardare, il programma può far corrispondere ad ogni caratteristica di un oggetto localizzato un'immagine praticamente identica della sua memoria. Ma messo di fronte solo ad una rappresentazione di due dimensioni e mezzo, il programma non sa da dove cominciare. Il problema che ora gli scienziati della visione affrontano con interesse è come conciliare i due approcci alla visione, cioè la visione a basso ed alto livello. Come si passa dai contorni agli oggetti?

Benché quella domanda rimanga ancora senza risposta, la visione artificiale sin dalla sua nascita come scienza ha compiuto notevoli progressi. Cinque anni fa, ci volevano trenta minuti perché un calcolatore estraesse i contorni di un'immagine; oggi ci vuole una frazione di secondo e il programma genera un'immagine pulita che, tra gli altri, i costruttori di aerei, trovano di grande utilità. Gli algoritmi che recuperano il colore, la profondità, il movimento e la forma delle superfici a partire da immagini del mondo bidimensionali sono più veloci che mai. Essi assistono già i veicoli autonomi sperimentali dei militari nella navigazione sulla terra e assistono i robot industriali nell'ispezione dei manufatti. Il successivo ostacolo da superare è integrare algoritmi differenti in un sistema visivo che sia capace di vedere in tempi reali. La Macchina per la Visione del laboratorio di IA del Mit è la prima versione di un sistema del genere. La visione artificiale ha anche stabilito dei collegamenti tra la biologia e la psicologia e ha mostrato che l'esistenza di macchine in grado di vedere implica avere una miglior visione della mente umana.

4. I livelli di comprensione

L'impegno della ricerca sulla visione artificiale di comprendere la visione a tutti i livelli nasce dalla natura e difficoltà dei problemi che essa affronta. La visione è un problema complesso e non servirebbe tentare di risolverlo costruendo dei programmi per il riconoscimento del telefono (*telephone-recognition*). I consolidati successi delle ricerche sulla visione artificiale originano dalla singolare applicazione del dogma centrale (che l'intelligenza può essere studiata come un sistema astratto di elaborazione dell'informazione, indipendente dalla macchina che lo sostiene), che è diventato una filosofia e una scienza a sé stante. La scienza è l'ottica inversa

che ha le sue fondamenta nella fisica ed è formalizzata in termini matematici. La filosofia alla base della visione artificiale (per esempio come è applicata ai laboratori di IA del Mit) è caratterizzata dall'assunzione dell'esistenza di livelli di comprensione ed analisi secondo cui i problemi di elaborazione dell'informazione sono affrontati a tre livelli: computazione, algoritmo e hardware. Il credo della visione artificiale è di stabilire per ogni problema delle buone fondamenta a livello computazionale.

L'approccio computazionale sostiene che i problemi della visione possono essere studiati come se si trattasse di problemi di matematica e fisica, vincolati dalle proprietà del mondo che viene osservato (*imaged*) e dall'occhio che costruisce le immagini. Le soluzioni devono essere completamente formulate indipendentemente dalla macchina che le implementerà; i vincoli naturali grazie a cui esiste un'unica soluzione sono gli stessi, siano i neuroni o i transistors a utilizzarli. Come ha scritto David Marr (1981): «Una volta che una teoria computazionale è stata definita per un problema particolare, essa non deve essere più riformulata». Essa diventa un pilastro dell'IA così come un teorema è un principio di base della matematica. La teoria computazionale sottostante l'estrazione dei contorni, che afferma che le caratteristiche più significative e primitive in un'immagine sono proprio i contorni, non è legata al modo in cui essi sono rilevati da un qualsivoglia pezzo di hardware: al contrario, la teoria fornisce dati empirici già comprovati circa le informazioni visive.

L'algoritmo è una procedura che esegue uno dopo l'altro i comandi della computazione: riguardo l'estrazione dei contorni, l'algoritmo è dato dall'insieme di istruzioni per calcolare la somma di un gruppo di valori di *pixels*, per dividere quella somma per il numero di *pixels* nel gruppo, aggiungere quel numero al valore medio del *pixel*, e così via. L'hardware è il marchingegno che implementa l'algoritmo: nel sistema visivo umano, le cellule retiniche sono estesamente interconnesse le une alle altre, permettendo così alle cellule vicine di inviare alla cellula centrale la somma della loro attività. A livello computazionale, la macchina per la visione determina ciò che vuole calcolare; a livelli di algoritmo e di hardware essa stabilisce come eseguire la computazione.

Inizialmente, quando Marr e Poggio si sono dichiarati favorevoli ad affrontare i problemi a livello computazionale essi ne sottolinearono l'indipendenza da altri livelli per liberarlo da una moltitudine di algoritmi dell'IA. Eppure in realtà i livelli interagiscono: l'algoritmo è determinato dal tipo di problema da risolvere ed è spesso

vincolato dalle proprietà e limiti dell'hardware. Alterare la procedura di un algoritmo, per esempio aumentare la velocità o migliorarne l'affidabilità, significa spesso modificare l'elaborazione che esso esegue. Cambiare anche solo dei dettagli di un algoritmo può portare all'invenzione di una nuova elaborazione o ad un'intuizione circa la reale natura del problema da risolvere. Analogamente, l'algoritmo potrebbe richiedere delle operazioni che esistenti hardware semplicemente non possono eseguire in modo efficiente, come per esempio moltiplicare grandi matrici di numeri, e a sua volta ciò potrebbe stimolare l'evoluzione o la scoperta di nuove macchine. Gli sforzi compiuti dalla ricerca sulla visione automatica degli ultimi quindici anni dimostrano da un lato un serio impegno ad affrontare la visione a tutti i livelli, e dall'altro l'affermarsi dell'opinione che i livelli sono così intrinsecamente connessi che un tale impegno è necessario se si vorrà mai comprendere la visione.

5. Aspetti contrapposti dell'intelligenza

Prima che diventasse ovvio che la visione era uno dei problemi più difficili che l'IA potesse affrontare, alcuni ricercatori in IA, all'esortazione di Marr di elaborare teorie computazionali robuste piuttosto che deboli algoritmi, hanno risposto sostenendo che tutto ciò andava molto bene per abilità semplici come la visione ma non per compiti più difficili caratteristici dell'intelligenza di alto livello. Oggi, l'impatto che la filosofia della visione artificiale potrebbe avere sulla tradizionale ricerca in IA è rafforzato dal conflitto tra l'IA tradizionale e il connessionismo. Eppure ci sembra che il connessionismo possa imparare qualcosa anche dalla visione artificiale.

L'IA tradizionale e il connessionismo sono due branche dello stesso campo di ricerca e possono venir considerate come i due poli dell'intelligenza.

La filosofia connessionista si ispira alle nostre capacità *associative* (messa in soggezione dal modo in cui avanziamo tra quelle sabbie mobili costituite dai vincoli multipli, dal parlare, dal canticchiare, dal guidare la macchina, dall'afferrare una tazza di caffè, dal riconoscere i visi tra la folla) mentre l'IA tradizionale si ispira alle nostre capacità deduttive (influenzata positivamente dalla logica, dalle dimostrazioni matematiche, dagli autorevoli dibattiti e dalla sistematica eliminazione di tutti i possibili *bugs* dai codici della macchina). Le loro diverse concezioni dell'intelligenza hanno con-

dotto l'IA tradizionale e il connessionismo a formulare progetti diversi per ricostruirla.

L'IA punta sugli algoritmi, il connessionismo insiste sull'hardware. Il connessionismo sostiene che gli algoritmi da soli non possono ricreare l'intelligenza e che l'enfasi posta dall'IA sugli algoritmi concede un ingiusto primato al processamento simbolico, che non potrà mai cogliere la «fluidità e adattabilità» dell'intelligenza umana (1986).

L'hardware è l'essenza dell'intelligenza, sostiene il connessionismo e non solo l'IA manca nel riconoscere questo fatto, ma usa l'hardware sbagliato. L'IA si è avvantaggiata del rapido sviluppo di computer seriali sempre più potenti, le «macchine di Von Neumann», che eseguono le istruzioni una dopo l'altra. I connessionisti sostengono che l'hardware dovrebbe eseguire le operazioni non in serie ma in parallelo, e che le quantità con cui opera dovrebbero essere numeri, non simboli, analogici, non digitali[4].

Inoltre il connessionismo ritiene che l'hardware più appropriato sia una rete di unità semplici fortemente interconnesse, in grado di elaborare contemporaneamente parti reciprocamente interagenti dello stesso problema. L'output del sistema è determinato dalla somma totale dei valori di attivazione di tutte le unità nella rete, e non semplicemente dal valore di un solo predicato che termina una serie di deduzioni logiche.

Il sogno ultimo dei connessionisti e pure dei ricercatori in IA è di costruire una macchina che sia in grado di imparare. I connessionisti predicono che il giusto hardware si organizzerà, come magicamente, in un sistema intelligente, non semplicemente grazie a ciò che gli è stato detto di fare, ma perché esso è in grado di imparare e generalizzare da esempi. L'hardware conterrà proprio quegli elementi della mente che una volta aggregati mostreranno proprietà emergenti quali l'intelligenza, proprio come le molecole d'acqua che si uniscono per formare dei fiocchi di neve. Si dia al giusto tipo di rete una lista di parole scritte insieme alla loro corretta pronuncia e quella rete stabilirà lo stato in cui dovrebbe venire a trovarsi per pronunciare parole che non sono nella sua lista iniziale (*training list*). Non è necessario analizzare il calcolo che la rete esegue per passare dal testo al parlato.

[4] Nei computer digitali i dati vengono rappresentati e utilizzati con stringhe di zero e uno (numeri binari). Nel computer analogico i dati sono rappresentati da quantità fisiche, quali il voltaggio, che possono assumere una gamma di valori continua.

Più che discutere di hardware in opposizione al software, di simboli oppure di numeri, o di operazioni parallele in contrapposizione a quelle seriali, il dibattito si riduce veramente ad una sola questione: qual è lo scopo finale dell'intero progetto? O in altri termini, a che scopo studiare l'intelligenza? Per costruire macchine intelligenti? Per comprendere come il cervello si sia costituito? O invece, per descrivere la struttura e i poteri dell'intelligenza, vista come un'entità autonoma che non è legata né al cervello né alla macchina?

Se supponiamo di aver di fronte un connessionista tipico e un sostenitore dell'IA e poniamo loro questo quesito, otterremo risposte profondamente diverse. Il connessionista risponderà che il suo scopo è di costruire un modello del cervello simulando la rete neuronale. Il modello dovrebbe cogliere in misura adeguata le capacità naturali del cervello in modo tale da essere commerciabile. D'altra parte il tipico connessionista eviterebbe teorie dell'intelligenza non legate al funzionamento del cervello, per evitare i livelli intermedi dei simboli che si introducono nella transazione dai dati alla soluzione. In realtà nella maggior parte delle reti connessioniste, le unità sono semplificate a tal punto da non assomigliare più ai veri neuroni, che sono componenti molto complessi da un punto di vista biofisico e computazionale. Un vero connessionista ammetterà che la sola vera rassomiglianza tra le reti artificiali e il cervello è da ricercarsi ad un livello astratto, ovvero nella presenza di molte connessioni e molte operazioni simultanee.

D'altra parte il sostenitore dell'IA potrebbe affermare di essere il primo a voler costruire una macchina intelligente, ma probabilmente esiterebbe circa la necessità di studiare a fondo la struttura del cervello. Ciò non implica che egli rifiuterebbe i chiarimenti offerti dagli scienziati che si occupano del cervello. Eppure, sebbene il sostenitore dell'IA dica di avere a cuore l'importante compito di comprendere l'intelligenza, vista unicamente come un sistema astratto di elaborazione delle informazioni, egli genera solo programmi che risolvono compiti molto specifici.

In termini di livelli di comprensione, l'IA sostiene di operare a livello computazionale quando in realtà è bloccata a livello di algoritmi[5]. Il connessionismo sostiene di ignorare il livello computazionale e di cercare unicamente di costruire hardware simile al cervello. Tuttavia l'hardware delle reti connessioniste è molto lontano da quello del cervello, e molte delle reti funzionano solo

[5] Daniel Dennet (1984) esprime un'opinione simile.

perché l'analisi computazionale è stata già fatta in precedenza[6]. Il messaggio della visione artificiale al connessionismo e all'IA è che nessuno dei suoi scopi può essere conseguito senza il simultaneo conseguimento degli altri.

6. La visione: una sintesi

In realtà i confini tra l'IA tradizionale e il connessionismo non sono così distinti. Sebbene le loro teorie e le loro tecniche appaiono essere diametricalmente opposte tanto quanto lo sono gli aspetti dell'intelligenza a cui esse si ispirano, essi convergono nell'approccio della visione. Se da un lato il tipico connessionista e il tipico sostenitore dell' IA si confrontano, dall'altro lo scienziato della visione artificiale segue una strada precisa che coincide parzialmente con entrambi gli approcci.

La facilità, l'immediatezza e l'impenetrabilità della visione la collocano nel polo associativo dell'intelligenza. Così è considerata la visione e in questo modo la questione è velocemente risolta. Eppure analizzata più dettagliatamente dalle teorie computazionali, la visione spesso funziona seguendo regole deduttive. Alternativamente, la visione abbraccia metodologie contrastanti che derivano dalla considerazione dei due poli dell'intelligenza: essa sviluppa degli algoritmi fortemente paralleli, utilizza numeri naturali, fa affidamento su teorie astratte di elaborazione delle informazioni, e assembla sistemi esperti per la visione.

Il campo della visione artificiale è oggi fortemente caratterizzato da un approccio quantitativo e parallelo. La funzione di molte procedure della visione è quella di trasformare un enorme insieme ordinato di numeri in un altro insieme ordinato di numeri, piuttosto che di valutare la verità logica di una singola affermazione. A causa delle dimensioni dell'insieme ordinato e poiché tutti i punti in esso contenuti devono essere trasformati allo stesso modo (per esempio un rilevatore di contorni esegue esattamente la stessa operazione indipendentemente dallo specifico raggruppamento di *pixels*), il modo più naturale per eseguire la trasformazione è quello di

[6] Per esempio, la rete di John Hopfield è semplicemente una macchina per la minimizzazione. Cioè prima di utilizzarla per risolvere il problema, bisogna esprimere il problema, se possibile, come una quantità matematica da essere minimizzata. L'analisi preliminare è a livello computazionale ed ha poco a che vedere con la stessa rete.

eseguirla simultaneamente e in parallelo per ogni *pixel*. Molti tra gli algoritmi per la visione sono stati formulati tenendo presente l'idea di elaborazione parallela, usando come modelli la retina e il cervello (considerati tipicamente organi «paralleli»)[7]. Molti di questi algoritmi sebbene inizialmente siano stati provati su computer digitali, possono immediatamente e in maniera più efficace essere implementati su macchine a parallelismo elevato con reti in cui i livelli di attivazione delle unità sono espresse in quantità analogiche; la teoria della regolarizzazione che unifica molti dei primi algoritmi della visione, mostra in modo semplice come ciò possa essere fatto[8]. La *Connection Machine*, un potente computer composto da molte migliaia di semplici processori fortemente interconnessi, è stato in parte concepito per la ricerca sulla visione.

La visione artificiale ha tenuto anche in considerazione la struttura e il funzionamento del cervello. I ricercatori della visione si sono resi conto che le macchine artificiali possono trovare dei vantaggi emulando quelle biologiche che sono straordinariamente efficienti nella visione così come nelle altre capacità sensoriali. Di conseguenza, i rilevatori dei contorni sono stati disegnati seguendo il modello delle cellule retiniche e sono state prese in considerazione altre indicazioni provenienti da ricerche sul cervello. Allo stesso tempo, i neurobiologi si sono rivolti al campo della visione artificiale per cercare suggerimenti circa le operazioni che i neuroni devono

[7] Il primo articolo di Marr e Poggio sulla visione stereoscopica intitolato «Calcolo della disparità stereoscopica» del 1976, inizia così: «Può darsi che una delle differenze più evidenti tra il cervello e i computer odierni sia nella quantità di collegamenti (*wiring*). Nel computer digitale, il rapporto tra le connessioni e i componenti è di circa 3 mentre nel caso della corteccia dei mammiferi, il rapporto oscilla tra 10 e 10.000. Benché questo dato segnali una chiara differenza strutturale tra i due, questa distinzione non è fondamentale riguardo al tipo di elaborazione delle informazioni che ciascun sistema realizza, piuttosto lo è circa i dettagli di come essi operano. In termini chomskiani, questa differenza riguarda teorie della *performance* non teorie della competenza, poiché il tipo di calcolo che viene eseguito da una macchina o da un sistema nervoso dipende solo dal problema da risolvere, non dall'hardware disponibile. Cionondimeno ci si potrebbe aspettare che un sistema nervoso e un computer digitale usino diversi tipi di algoritmo, anche nell'eseguire lo stesso compito. Algoritmi dalla struttura parallela che richiedono simultaneamente molte operazioni locali su grandi insiemi di dati, sono costosi per i computer odierni, ma probabilmente adatti ad un'organizzazione dei sistemi nervosi altamente interattiva...»

[8] Barthold Horn fu probabilmente il primo (nel 1974) a usare reti analogiche per risolvere un problema sulla visione, il calcolo della luminosità (*lightness*) (si veda Berthold e Horn 1986). Per il legame tra le reti analogiche e i primi algoritmi per la visione, si veda Poggio et al. (1985).

eseguire per risolvere i vari problemi percettivi. Eppure la visione artificiale non si limita ad implorare l'IA affinché essa si converta ai numeri e al parallelismo. Dopo tutto, l'IA non può semplicemente e immediatamente trasformare le basi di dati e le macchine inferenziali dei sistemi esperti in enormi insiemi ordinati di numeri[9]. Né la visione artificiale impone al connessionismo di essere più attenta alla struttura e al funzionamento del cervello. Sopra ogni altra cosa, la visione artificiale dice al connessionismo e all'IA di cercare di trovare soluzioni ad un livello computazionale.

Una soluzione di tipo connessionista al problema della visione stereoscopica potrebbe essere quella di presentare ad una rete artificale di neuroni degli insiemi di tre immagini, un'immagine di input per ogni occhio (che dà l'intensità della luce per ogni *pixel*) e un'immagine di output, la soluzione (che dà la distanza dell'osservatore da ciascun *pixel*). Per ciascun insieme, la rete genererebbe l'immagine di output dai due input, la paragonerebbe con la giusta soluzione, e aggiusterebbe i pesi delle connessioni tra le unità per far corrispondere le due immagini di output. Dopo che la rete ha ricevuto un numero sufficiente di insiemi di *training*, essa finalmente si stabilirà su di un *pattern* di pesi che genererà un'immagine di profondità accurata dopo una coppia di immagini di input completamente nuova. L'osservazione della rete può rivelare una maglia di connessioni eccitatorie e inibitorie molto simili a quelle che l'algoritmo di Marr e Poggio descrive per risolvere lo stesso problema. Eppure sebbene l'algoritmo di Marr e Poggio è il risultato di un'analisi computazionale sulla percezione della profondità, e funzioni quindi all'interno di un dominio che è stato completamente descritto, si possono solo fare delle supposizioni circa che cosa e quanto la rete connessionista potrebbe essere in grado di fare.

Come può l'IA trarre beneficio dall'approccio computazionale? Si consideri ad esempio il modo in cui una qualsiasi mosca vola. La mosca ha degli obiettivi semplici in conseguenza delle sue primitive capacità visive. La mosca può essere richiamata da un qualsiasi contrasto bianco-nero (una briciola di pane sulla tovaglia) o da un punto mobile nero (un'altra mosca, forse di sesso opposto). Un tipico sostenitore dell'IA sarà preso dalla tentazione di formulare regole esplicite che dicano alla mosca come seguire le tracce di un potenziale compagno: se un punto scuro vira a destra, girare a

[9] Alcuni connessionisti stanno cercando di fare proprio questo: testimonianza ne è la creazione di James Anderson di basi di dati medici codificate in reti.

destra. Se un punto scuro vola intorno punta ad esso. Lo schema che sta sotto le regole è quello di paragonare continuamente la direzione dell'obiettivo (*target*) con la direzione della mosca, e effettuare dei movimenti per far coincidere la mosca con l'obiettivo. L'approccio computazionale potrebbe sviluppare uno schema simile però senza fare una lista di regole che valgano solo per un insieme finito di spostamenti.

La teoria di Poggio e Reichardt (1976) su come le mosche controllano visivamente il loro volo è una teoria computazionale classica. Lascia da parte le regole descrivendone la struttura soggiacente. La teoria incorpora le regole in un'unica asserzione matematica che dimostra come un input visivo sia trasformato in un output motorio, vincolato dalla fisica del volo e dalla biologia della visione. L'equazione eguaglia la coppia esercitata dalle ali della mosca (che a sua volta controlla la sua posizione e velocità) con la differenza tra la posizione reale e desiderata dell'immagine dell'obiettivo sulla retina. La velocità con cui quella differenza muta determina la velocità con cui la coppia cambia. Un'unica equazione riassume il *pattern* di volo della mosca durante un inseguimento.

Benché un connessionista probabilmente consideri estremamente semplice il comportamento di una mosca, l'azione di riflesso che l'uomo compie nel mettere il piede sul freno quando una macchina di fronte si ferma ad un incrocio, potrebbe essere governata da un'equazione simile. Ma l'ostacolo maggiore nel programma del volo della mosca scritto dal sostenitore dell'IA è più illustrativo del connessionista che denigra la mosca: il programma tradizionale assume che il più difficile lavoro di individuare l'obiettivo sia già stato fatto. Le sue regole si applicano solo una volta ricevute le informazioni ad alto livello circa la posizione e velocità dei punti neri. Se fosse un programma veramente intelligente, partirebbe dall'immagine iniziale generata dall'occhio primitivo della mosca, ne troverebbe i punti salienti e ne seguirebbe in modo continuo le tracce. Il lavoro sta nel decifrare gli errori delle posizioni retiniche relative all'obiettivo e nello scoprire che la coppia generata dalle ali è il giusto output dell'ala. Allo stesso modo, la rete connessionista per la visione stereoscopica avrebbe difficoltà a stabilizzarsi in uno stato produttivo se ricevesse solo immagini pure e semplici. La quantità di informazioni in queste immagini risulterebbe eccessiva per una rete artificiale a meno che questa non fosse irrealisticamente vasta e complessa. La rete per la visione stereoscopica funzionerebbe molto meglio con immagini i cui contorni siano già stati rilevati e che è utilizzata dall'algoritmo di Marr e Poggio per ridurre

gradualmente la quantità di possibili corrispondenze tra i *pixels* delle due immagini. Il programma della mosca e la rete per la visione stereoscopica richiedono, per poter funzionare, l'appropriata rappresentazione degli inputs.

Una teoria computazionale costruisce una rappresentazione adeguata, indica come andare dall'input all'output trovando il giusto input e output e specifica sia l'informazione che il modo in cui debba essere elaborata. Nella ricerca degli elementi di base della conoscenza, l'IA ha cercato i passaggi deduttivi più semplici all'interno di regole più globali del pensiero, mentre il connessionismo ha cercato i legami associativi più reconditi. La visione artificiale ha rivolto il suo sguardo all'esterno: ha cercato generalizzazioni sul mondo che siano quasi sempre vere (gli oggetti sono rigidi, le superfici sono lisce, i contorni sono continui) e le ha tradotte in vincoli sugli elementi informativi di base.

La visione artificiale condivide il sogno di costruire una macchina che sia in grado di imparare. Tuttavia prima di tutto ci sono domande a cui è necessario rispondere: è possibile imparare un qualsiasi compito iniziando da una tabula rasa? Noi pensiamo di no. Per la maggior parte dei problemi, deve esistere un quadro di riferimento che guidi i dati verso una soluzione prima ancora che l'apprendimento possa influenzare la soluzione, concretizzandola e migliorandola. Alcune delle trasformazioni che dai dati conducono alla soluzione del problema non possono essere apprese in alcun modo se non attraverso un esame esaustivo di tutte le possibili soluzioni. Ma indipendentemente dalla risposta, questa domanda rappresenta un filone di indagine nella miniera della ricerca. È necessario portare alla luce le caratteristiche delle computazioni che possono essere apprese ed evidenziare quanto efficacemente e da quali classi generali di reti esse possono essere imparate [10]. Attualmente la ricerca è inadeguata. Tuttavia, alcuni scienziati della visione stanno esaminando le implicazioni della teoria della regolarizzazione, in cui si dimostra che, a certe condizioni, alcuni algoritmi della visione possono essere imparati da esempi.

[10] Ci sono diverse questioni di base che sorgono, ma non sono state risolte dall'approccio connessionista all'apprendimento. Le tecniche di apprendimento connessioniste (esemplificate dall'esempio della visione stereoscopica) funzionano solo per problemi di piccole dimensioni? Più importante è sapere che tipi di apprendimento funzionano per quali classi di problemi. Infine, gli algoritmi di apprendimento connessionisti sono significativamente diversi dalle tecniche classiche di regressione e classificazione? Ci azzardiamo a dire che essi sono probabilmente simili.

Saremmo soddisfatti semplicemente costruendo macchine che sono in grado di apprendere? Da un punto di vista pratico la risposta sarà probabilmente positiva: queste macchine saranno certamente molto utili. Tuttavia se lo scopo è di capire l'intelligenza, la risposta è negativa. La semplice riproduzione di una capacità non ne spiega la sue strategie di fondo. Gli esseri umani possono apprendere, eppure non sappiamo come ciò avvenga. La teoria evolutiva fornisce una descrizione completa e coerente di come la vita, e i cervelli, si sono evoluti. Ci informa su come costruire un sistema nervoso, sebbene la procedura sia sfortunatamente troppo dispendiosa in termini di tempo. Eppure, questa teoria sulla vita e sull'intelligenza, così come il punto di vista secondo cui sia sufficiente costruire una macchina intelligente, non è sufficiente per coloro che vogliono càpire che cos'è l'intelligenza. Allo stesso modo, anche se si fosse scoperta la rete magica che è in grado di imparare a risolvere qualsiasi problema, colui che sinceramente crede nei livelli di comprensione, insisterebbe ancora nel domandare: che cosa ha imparato la rete?

L'uomo non dovrebbe aversene per il fatto di essere usato come prova dell'esistenza delle macchine per l'elaborazione delle informazioni semplicemente perché egli rifugge dalla manipolazione simbolica dei tradizionali programmi in IA. Sebbene questo stile di programmazione sia ancora il più adatto per affrontare le tradizionali questioni relative al ragionamento, alla risoluzione dei problemi e alla logica, esso da solo non può bastare. Così come il pensiero ha due aspetti e l'intelligenza ha due poli, lo studio dell'intelligenza deve attingere a due filosofie. Azioni così profondamente intelligenti come la percezione, il riconoscimento del parlato, il controllo motorio necessitano un approccio in cui l'aspetto quantitativo, parallelo, analogico, siano più influenti. Non dovremmo dimenticare che coloro che aspirano a costruire macchine intelligenti hanno tutto il tempo per confutare la loro ipotesi iniziale: che l'intelligenza sia riproducibile da una macchina. Oggi l'intelligenza umana supera di molto le capacità dei sistemi esperti o delle reti connessioniste, ma in futuro macchine più sofisticate potrebbero risentirsi ad un'affermazione di questo genere. Quelle macchine potrebbero ricordarsi con affetto il tempo in cui la visione artificiale che unifica tutti i livelli di comprensione dell'intelligenza umana, congiunse i propri genitori.