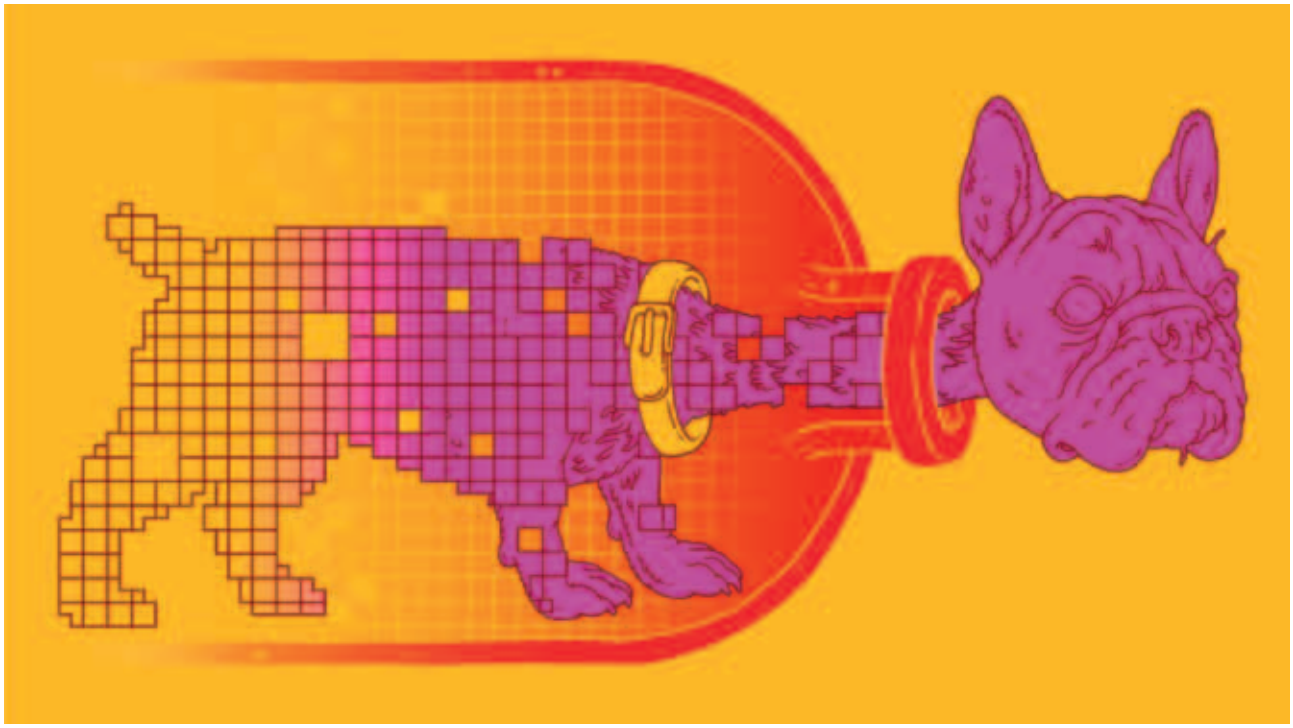
學習的奧祕是遺忘

揭開深度學習神秘黑箱的新理論

作者：**渥秋華 Natalie Wolchover**　譯者：**紀露結**

渥秋華是線上科普雜誌 *Quanta Magazine* 的資深主筆。她是塔夫斯大學物理學士，曾於加州大學柏克萊分校攻讀研究所。2016 年，她在該雜誌的報導為她贏得美國科學作家協會頒給年輕科學記者的 Evert Clark/Seth Payne 獎。



（Eric Nyquist/Quanta Magazine 繪製）

「資訊瓶頸」的新概念有助於解釋當今人工智慧演算法令人目眩神馳的成就，或許也能揭露人腦學習的機制。

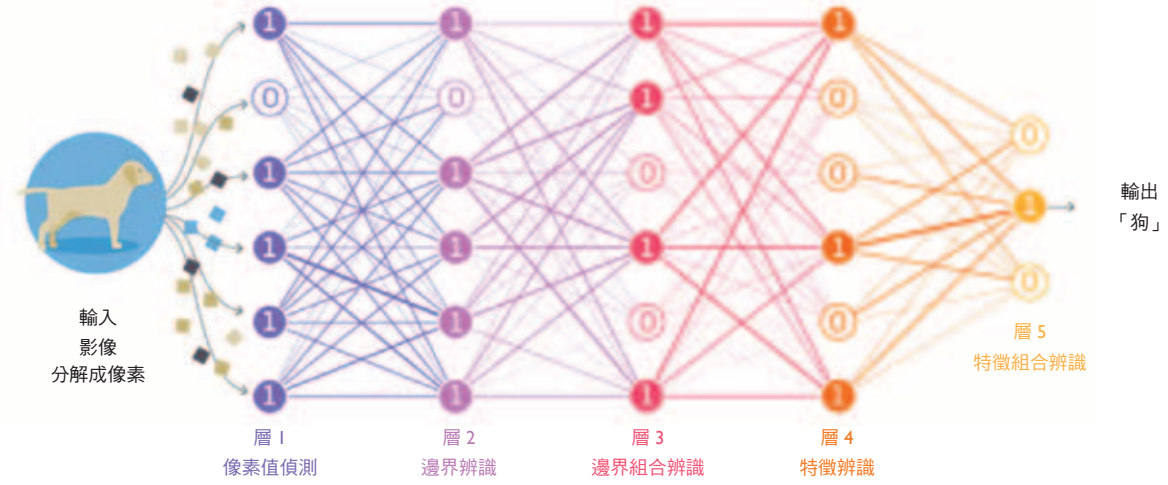
如今許多稱為「深度神經網絡」（deep neural network）的機器已經學會交談、開車、擊敗電玩與圍棋世界冠軍、做夢、繪畫，以及輔助科學發現，這讓打造出這些機器的人類大感意外，他們根本想不到所謂的「深度學習」（deep-learning）演算法會運作得如此成功，因為除了源自人腦架構的模糊靈感，這些學習系統的設計並沒有別的基本指導原則，而且其實也沒有人真正弄懂人腦究竟如何運作。

深度神經網絡就如人腦，有一層又一層的神經元（由電腦記憶體構成的人工物件），神經元受激發時，會傳遞訊號到上一層的連接神經元。系統進行深度學習時，網絡中每一條連結會根據需求增強或減弱，好讓系統把輸入資料的訊號（例如一張狗的照片像素），逐層傳遞到與正確高層次概念（譬如「狗」）相關聯的神經元。深度神經網絡「學習」

Quanta Magazine 是西蒙斯基金會（Simons Foundation）出版但編輯獨立之網路科普雜誌（http://www.quantamagazine.org/），希望能提高數學、物理與生命科學前沿研究進展的公眾能見度。本文譯自：
https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-20170921/
本刊感謝 *Quanta Magazine* 與主編 Thomas Lin 同意翻譯轉載，翻譯之文責由本刊自負。

從經驗中學習

深度神經網絡的學習是透過調整網絡連結的強度，將輸入訊號更有效的透過多層神經元傳至表示正確一般概念的神經元。



資料輸入網絡後，每個受激發的人工神經元（標記為「1」）會將訊號傳到下一層的特定神經元，這些神經元接收多個訊號後可能會被激發再往上傳遞。這個過程會過濾雜訊，只保留最為相關的特徵。（Lucy Reading-Ikkanda 繪製）

了幾千張狗的樣本照片後，它就能像人一樣準確認出新照片中的狗。深度神經網絡的威力，正是來自學習過程中從特例轉化成一般概念的神奇跳躍，就像人的推理、創意與其他統稱為「智慧」的能力背後的原因一樣。專家很好奇深度學習為何能促成一般化，也想知道到哪種程度，大腦也是以同樣的方式理解現實世界。

上個月在人工智慧研究社群廣為流傳的一支 YouTube 影片提供了可能的解答 ❶，影片內容是一場在柏林舉辦的會議演講。在演講中，耶路撒冷希伯來大學的電腦科學兼神經科學家泰斯比（Naftali Tishby）為一個解釋深度學習的新理論提出支持證據。泰斯比認為，深度神經網絡的「學習」是依據名為「資訊瓶頸」（information bottleneck）的程序；泰斯比與兩位合作者在 1999 年首次以純理論描繪這項程序。資訊瓶頸的概念可以簡述如下：網絡會讓資訊像擠過又窄又細的瓶頸一樣，把不相干細節這類雜訊輸入去除，只保留那些與一般概念最相關的特徵。泰斯比和學生許瓦茲齊夫（Ravid Shwartz-Ziv）以一些引人矚目的新電腦實驗，顯示這種擠壓的擷取程序如何出現在深度學習的過程中（至少就他們研究的案例是如此）。

泰斯比的發現令人工智慧社群沸騰不已。任職於 Google Research 的阿勒米（Alex Alemi）說：「我相信『資訊瓶頸』會是未來深度神經網絡研究很重要的概念。」他表示，資訊瓶頸「不僅可作為理論工具解釋我們的神經網絡為何表現如此優異，也能成為建立嶄新標的與網絡架構的工具。」阿勒米目前已經發展出一些新的近似方法，將資訊瓶頸分析運用於大型深度神經網絡。

有些研究者仍然懷疑泰斯比提出的理論無法完全解釋深度學習的成功，但紐約大學的粒子物理學家克蘭默（Kyle Cranmer）表示，若當作普遍的學習原理，這項理論「感覺起來就是對的」。克蘭默現在正利用機器學習，分析大型強子對撞機（LHC）中的粒子碰撞。

在 Google 與多倫多大學工作的深度學習先驅辛騰（Geoffrey Hinton）看了泰斯比在柏林的演講影片後，電郵給泰斯比。他寫道：「太有趣了！我得重看一萬遍才能真的搞懂，如今已經很少聽到這種

演講，其中具有真正的原創性，而且可能解決真正的大問題。」

泰斯比把資訊瓶頸視為學習背後的基礎原理，無論是演算法、蒼蠅、有意識的生物，或是關於突現行為的物理計算，這個大家寄望已久的答案就是「學習時最重要的一環其實是遺忘。」

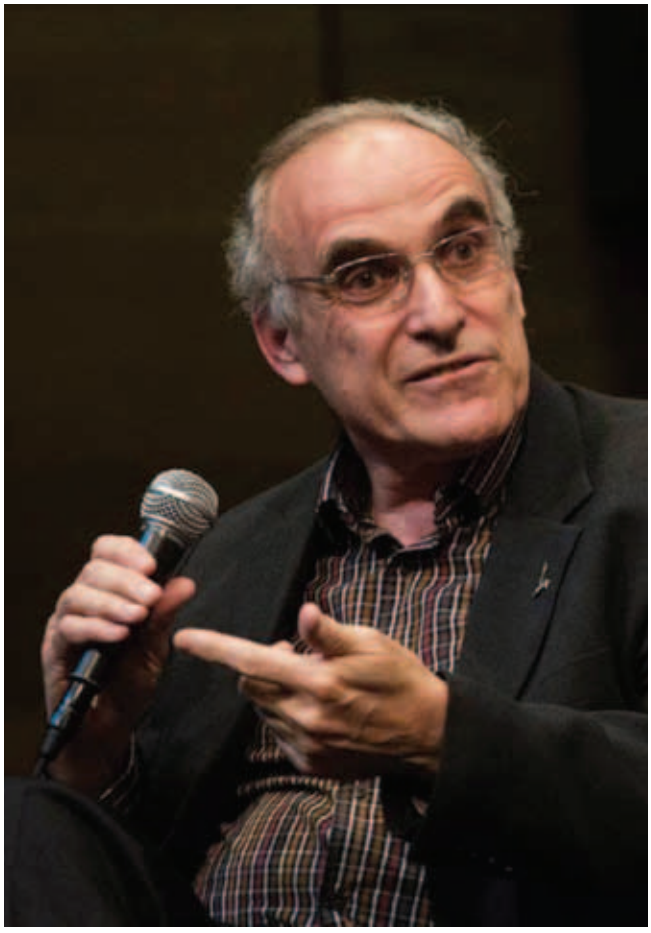
瓶頸

1980 年代，當其他研究者正開始探索深度神經網絡之際，泰斯比也開始推敲資訊瓶頸的想法，儘管這兩個概念當時都還沒有命名。在那個年代，語音辨識是人工智慧的一大挑戰，而泰斯比思考的是人類究竟有多擅長辨識語音。他意識到問題的核心是相關性：與口語最相關的特徵是什麼？要如何從腔調、聲調、含糊的嘟噥等伴隨的變數中擷取出這些特徵？一般來說，面對現實中浩瀚如海的數據時，我們應該保留哪些訊號？

泰斯比在上個月受訪時提到：「歷史上曾多次論及相關資訊（relevant information）的概念，但從未正確表述出來。多年來大家都認為資訊理論不是思考相關性的正確工具，誤解的起源甚至可直接上溯到夏農（Claude Shannon）本人。」

夏農是資訊理論的創立者，就某種意義上，他算是解放了始於 1940 年代的資訊研究，因為他將資訊提升至抽象思考層面——變成只有數學意義的諸多 1 與 0。泰斯比指出，夏農認為「資訊無關語意。」❷ 但是泰斯比不認同夏農的觀點，他意識到，利用資訊理論，「我們可以精確定義『相關』的意思。」

假設 X 是一個複雜的數據集，例如一張狗照片的所有像素，而 Y 是由這些數據表示的較簡變數，例如「狗」這個字。我們可以在不喪失預測 Y 的能力下盡力壓縮 X，藉此獲得 X 中所有與 Y 相關的資訊。1999 年，泰斯比與佩雷拉（Fernando Pereira，目前任職 Google）、巴亞萊克（William Bialek，目前在普林斯頓大學任教）合寫的論文中，他們將這個問題表述成一個數學的最佳化問題。這是一個基礎想法，並沒有發展成殺手級應用（killer



泰斯比是耶路撒冷希伯來大學的電腦科學教授。（Miriam Alster 攝於 ELSC Art and Brain Week 2016）

❶ 編註：這支 2017 年 8 月放上 YouTube 的影片見 https://www.youtube.com/watch?v=bLqJHjXihK8 會議是 2017 年 6 月底在柏林舉行的 *Deep Learning: Theory, Algorithms, and Applications*。

❷ 譯註：夏農在 1948 年的論文〈通訊的數學理論〉（A Mathematical Theory of Communication）為通訊與資訊理論奠定基礎。在引言中，夏農表示「通訊中的語意層面與工程問題無關。」（These semantic aspects of communication are irrelevant to the engineering problem.）

application）。

泰斯比說：「30 年來，我一直在各種脈絡中思考這些課題，幸運的是深度神經網絡恰好變得如此重要。」

泰斯比的兩位研究生：札斯拉夫斯基（左）與許瓦茲齊夫（右），兩人協助泰斯比發展深度學習的資訊瓶頸理論。（照片提供：Ravid Shwartz-Ziv）

風景的人的臉的眼

雖然深度神經網絡背後的概念已經歷數十年的談論，但一直要到 2010 年代初期，基於網絡訓練機制的改善與更強大的電腦處理器，深度神經網絡在語音辨識與影像辨識的實務表現才真正起飛。2014 年，泰斯比在讀到物理學家施瓦布（David Schwab）與梅塔（Pankaj Mehta）令人意外的研究論文之後，才看出這些結果與資訊瓶頸原理的潛在關聯。

這兩位物理學家發現，在特定的情況下，辛騰發明的深度學習演算法「深度信念網」（Deep Belief Net），正好和物理學裡的「重整化」（renormalization）程序相同。所謂的重整化，就像把對準物理系統的鏡頭拉遠，將系統細節粗粒化（coarse-graining）並據此計算出整體狀態。施瓦布與梅塔把深度信念網演算法應用在達到「臨界點」的磁體模型[3]，這時系統會呈現碎形（fractal）或全尺度自相似的狀態。結果他們發現深度網絡會自動使用與重整化相似的程序去找出模型的狀態。正如生物物理學家奈門曼（Ilya Nemenman）當時所說的，這項令人目瞪口呆的發現顯示：「在統計物理脈絡中擷取相關特徵與在深度學習脈絡中擷取相關特徵的方法，不僅僅是相似而已，根本就一模一樣！」

唯一的問題是，一般的真實世界並非碎形。克蘭默說：「在大自然裡，耳朵不是長在耳朵的耳朵的耳朵上，而是風景中的人的臉上的眼珠。所以我不認為深度學習辨識影像的出色能力可歸功於重整化的程序。」然而，當時正在接受胰臟癌化療的泰斯比意識到，無論是深度學習或粗粒化程序，都可以納入一個更寬廣的架構。他說：「思考科學以及自己昔日想法的意義，是我痊癒與康復過程中很重要的一環。」

2015 年，泰斯比與學生札斯拉夫斯基（Noga Zaslavsky）推測深度學習是一種資訊瓶頸程序，會盡可能壓縮有雜訊的數據，同時保留這些數據所表現的相關資訊。泰斯比與許瓦茲齊夫對深度神經網絡的新實驗，顯示了資訊瓶頸程序的實際發生過程。在其中一個實驗裡，他們使用了含有 282 個神經連結的小型網絡，訓練替輸入資料標上 1 或 0（想成「有狗」或「沒狗」），其中連結的初始強度是隨機設定的。接著，他們就追蹤該網絡深度學習 3000 個樣本輸入資料集的狀況。

大多數深度學習程序使用的基本演算法稱為「隨機梯度下降法」（stochastic gradient descent），它會隨著資料的變動修改神經元連結的強度：當每次訓練資料輸入網絡，就會激發一連串訊號由下而上傳過各層人工神經元。當訊號抵達最上層，系統會將最終激發的模式與圖片正確的標籤（如 1 或 0；

「有狗」或「沒狗」）做比較。任兩者間有任何不一致，都會「反向傳遞」（back-propagated）下傳網絡各層，就像老師批閱考卷，演算法會增強或減弱每個連結，讓網絡層更能產生正確的輸出訊號。在訓練過程中，訓練資料的共同模式逐漸反映在網絡連結的強度上，最後網絡成為能正確標記這類資料的專家，辨識出照片中的狗、特定的字、符號 1 等等。

泰斯比與許瓦茲齊夫在他們的實驗中，記錄了深度神經網絡每一層留存多少與輸入資料有關、多少與輸出標籤有關的資訊量。這兩位科學家發現，網絡會逐層收斂到資訊瓶頸的理論界限：這是 1999 年泰斯比、佩雷拉及巴亞萊克三人原初論文所導出的理論極限值，代表系統在擷取相關性資訊方面所能做到的絕對極致。當網絡達到界限值時，就表明系統在未犧牲準確預測標籤能力的條件下，已經盡可能壓縮輸入的資料。

在實驗中，泰斯比與許瓦茲齊夫也獲得奇妙的發現：深度學習過程分為兩個階段（phase，相）：時間比較短的適配階段（fitting）與時間長得多的壓縮階段（compression）。在適配階段，網絡學習如何標記訓練資料；在壓縮階段，系統會精熟如何一般化，並以標記新測試資料的表現來衡量。

當深度神經網絡採用隨機梯度下降法調節連結強度時，剛開始與輸入資料相關的資訊儲存量會大致維持固定或稍微增加。此時連結強度開始調整，讓網絡可以編碼輸入資料的模式，更能與正確的標籤適配。有些專家將適配階段與學習的記憶階段相提並論。

接著，深度網絡的學習轉換到壓縮階段。網絡開始去除輸入資料中不相干的資訊，只保留與輸出標籤最為相關的特徵。這是因為在隨機梯度下降法的每一次迭代中，訓練資料中強弱不一的相關性會對網絡產生不同的效應，使得網絡神經元連結的強度跟著上下起伏宛如隨機漫步（random walk），這些隨機化在效果上大致與系統如何正確呈現輸入資料的壓縮過程相近。譬如有些狗照片的背景是房子，有些不是。當網絡反覆面對這些訓練照片時，可能會「忘記」狗與房子之間的相關性，因為背景沒有房子的照片會抵消這種關連。泰斯比與許瓦茲齊夫認為，正是因為忘記這些細節，才能讓系統形成一般性的概念。確實，他們的實驗顯示深度神經網絡會在壓縮階段提升一般化能力的表現，更正確的標記測試資料。例如受過訓練從照片辨識狗的深度神經網絡，可以用來測試是否有狗在新照片裡面。

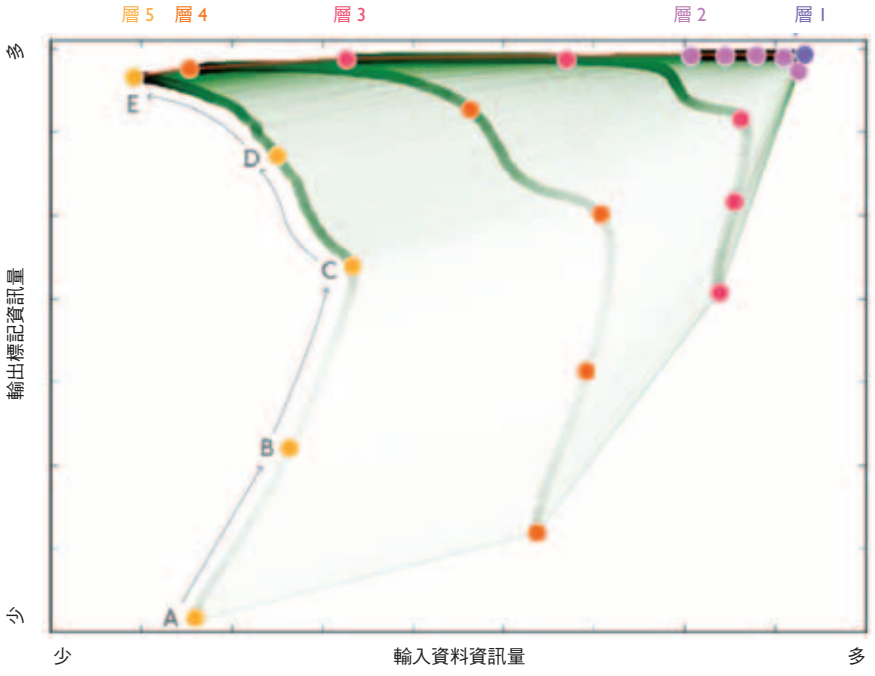
尚待觀察的是，資訊瓶頸是否掌握了所有深度學習的機制，還是在壓縮之外，另有有其他途徑可以達到一般化。有些人工智慧專家認為，泰斯比的想法只是近期關於深度學習的許多重要理論洞見之一。例如哈佛大學的人工智慧兼理論神經科學家薩克歇（Andrew Saxe）指出，有些非常大型的深度神經網絡似乎不需要曠日持久的壓縮階段就很善於一般化，研究者在程式中編寫所謂的「提早停止」（early stopping）步驟，縮短訓練過程，防止網絡在初期編碼太多的相關性。

---

❸ 譯註：即易辛模型（Ising model）。施瓦布與梅塔的研究可見 *Quanta* 另一篇更仔細的報導 "A Common Logic to Seeing Cats and Cosmos"（看見貓與宇宙的共同邏輯）。

深度學習的階段

新實驗揭示了深度神經網絡在學習過程中的演變方式。



A 初始狀態：第一層神經元將輸入資料全部編碼，包括與標籤有關的所有資訊。在此階段，最上層神經元的狀態近乎隨機，與輸入資料或標籤都幾乎沒有關係。

B 適配階段（相）：深度學習開始後，上方各層神經元獲得與輸入資料有關的資訊，標記資料的表現越來越好。

C 相變：各層突然「換檔」，開始「遺忘」輸入資料的資訊。

D 壓縮階段（相）：上方各層壓縮輸入資料的表現，只留下與輸出標記最相關的資訊，預測標記能力提升。

E 最終狀態：最上層達到準確度與壓縮的最佳平衡，僅保留預測標記所需資訊。

（Lucy Reading-Ikkanda 繪製；改繪自 arXiv:1703.00810 [cs.LG]）

泰斯比認為薩克歇及其同事分析的網絡模型不同於標準的深度神經網絡架構，儘管如此，資訊瓶頸理論界限還是比其他方法更能清晰界定這些網絡的一般化表現。至於資訊瓶頸理論是否適用於更大型網絡的問題，泰斯比與許瓦茲齊夫在最新的實驗嘗試給出回答（這些在他們的初步論文並未提及）。他們訓練規模大得多、具有 33 萬連結的深度神經網絡去學習辨識手寫數字，訓練素材來自 MINST 數據庫（Modified National Institute of Standards and Technology database）多達六萬張的手寫數字圖片。此數據庫用於檢測深度學習演算法的成效，是個廣為人知的基準。他們再度目睹網絡收斂到資訊瓶頸的理論界限，也觀察到深度學習的兩個截然不同的階段，而且比起小型網絡，這兩個階段（相）在大型網路中的的相變更明顯。泰斯比說：「我如今完全相信這是一個普遍現象。」

人與機器

人類的大腦如何篩選來自感官的訊號，把這些訊號提升到意識可察覺的層次，這個謎團促使人工智慧的先行者對深度神經網絡產生興趣，他們希望能以逆向工程研究大腦的學習法則。但自那時之後，人工智慧的實踐者因為科技進步的愚蠢挫敗，大多放棄這條道路，毅然轉向可促進表現的花俏手段，漠視擬似生物性的初衷。捱過這段日子之後，如今他們的思維機器達成前所未見的偉大成就（甚至讓人擔憂人工智慧有朝一日可能威脅人類的生存），許多研究者希望藉由這些探索，能夠揭開學習與智慧的廣泛洞識。

紐約大學的心理學兼數據科學助理教授雷克（Brenden Lake）是研究人類與機器學習異同的學者，他認為泰斯比的發現是「揭開神經網絡神秘黑

箱重要的一步」。但他強調，人腦是一個更大更黑的黑箱。成人大腦的 860 億個神經元之間有幾百兆個連結，很有可能會運用一大籮筐提升一般化的手法，遠多於發生在嬰兒期的基本影像與聲音辨識學習程序，而且在許多方面可能與深度學習類似。

雷克用自己的研究舉例說明，幼兒學習寫字的過程中，似乎就沒有類似泰斯比所發現的適配與壓縮階段。幼兒不需要把同一個字看上千遍，再花上一大段時間壓縮成心靈表示，然後才能辨認出這個字的其他範例並寫出這個字。實際上，幼兒需要的只是單一範例。雷克及其同事的研究模型，顯示大腦會把新的字拆解成一系列的筆畫（既存的心智構造），並容許把新字的概念加入既存的知識體系。雷克解釋說：「與其〔像標準機器學習演算法那樣〕把字的圖像看成一個像素模式，再從中擷取特徵習得該字的概念，我更傾向為字建立簡單的因果模型（causal model）。」這是通往一般化的較短路徑。

如此聰明的想法或許會為人工智慧社群上了一課，促進人工智慧與大腦研究兩個領域之間的交流。泰斯比相信他的資訊瓶頸理論在這兩個領域最終都會證明是有用的，即使這個理論在人類學習領域的形式會比人工智慧更為一般化。譬如從這個理論可以得到一項立即的洞察，更能理解哪類問題能以大腦或人工神經網絡來解決。泰斯比表示：「它完整的刻畫什麼是可以學習的問題。〔這些都是〕能去除輸入中的雜訊，又不傷害分類能力的問題。自然的視覺問題、語音辨識都是這一類問題。這些也正是大腦能處理的問題。」

同時，每個細節都舉足輕重、些微差異就能翻盤的問題，用大腦或人工神經網絡處理都會失敗。例如，許多人無法快速心算兩個大數字的乘積。「我們有一大串這類問題，例如對單一變數改變很敏感的邏輯問題、離散分類問題、密碼學問題。我不認為深度學習有辦法幫我破解密碼。」

一般化（或許是藉由穿越資訊瓶頸達成的）意味著留下一些細節。這對於快速進行代數運算或許沒什麼好處，然而這並不是大腦的要務。人類需要的是在人潮中搜尋熟悉的臉孔，自吵雜紛擾的世界捕捉重要訊號，從混沌發現秩序。

本文出處

Quanta Magazine September 21, 2017。

譯者簡介

紀露結現就讀於臺灣大學數學研究所。

延伸閱讀

▶ Tishby, Naftali "18. Information Theory of Deep Learning. Naftali Tishby" YouTube 影片。片源：*Deep Learning: Theory, Algorithms, and Applications* 會議（2017/6）。

網址：https://www.youtube.com/watch?v=bLqJHjXihK8

▶ Tishby, Naftali & Pereira, Fernando & Bialek, William, "The information bottleneck method"

網址：https://arxiv.org/pdf/physics/0004057.pdf

▶ Wolchover, Natalie "A Common Logic to Seeing Cats and Cosmos", *Quanta magazine*（2014/12/04）

網址：goo.gl/ccT61i