【解牛集】— 刊於〈信報〉，2018 年 5 月 8 日

打開 AlphaGo 橫掃棋壇技術面紗

楊毅
科大商學院資訊、商業統計及營運系助理教授

東方海外航運(OOCL)與微軟亞洲研究院（MSRA）近日展開一項研發合作計劃，透過應用人工智慧研究，以改善航運網絡營運，提升效率，預計每年可節省 1000 萬美元營運成本，可見人工智慧（Artificial Intelligence・AI）在商業領域的應用，將方興未艾，欲罷不能。

人工智慧無疑是目前計算機科學的前沿研究。去年 5 月，人工智慧 AlphaGo 三度擊敗中國圍棋高手柯潔，令內嵌到 AlphaGo 的「深度增強學習」（Deep Reinforcement Study）能力，進一步成為人工智慧域「最火」的詞彙。筆者於本文透過淺白語言，讓讀者瞭解一下 AI 技術的發展，探摸一下其令人目眩的發展脈搏。

深度增強學習能力其實並不神秘。顧名思義，就是把深度學習與增強學習結合起來，從而實現由感知（Perception）到動作（Action）的「端對端」學習之計算機新演算法，針對決策與控制的問題。

「智慧體」預測力躍進

事實上，人工智慧通過對以往的歷史數據進行分析，能夠協助人作出預測，譬如預測股票市場價格變化、公司未來的發展前景、明天天氣情況、甚至來預測一名消費者能否成為我方公司的潛在顧客等。

除了預測，人工智慧技術還用來進行決策。在人工智慧領域，習慣以 agent（代理）來表示一個具備行為能力的物體，比如機器人、無人車，甚至在廣義上也可以說人類自身的行為等。中國科學界趨向把 agent 翻譯為「智慧體」。深度增強學習考慮的，就是「智慧體」和「環境」（environment）之間交互的任務，當中蘊含「智能體」因應外部環境變化預測而作出的決策。

筆者試從人類生活的角度，來作說明。很顯然，我們每天都需要作出不少決策。譬如，今天是否早點起床，早起，可望（預測）完成某些工作任務；抑或懶床，多睡片刻，滿足甜睡樂趣，但某項工作便無法完成了。起床或是懶床，當中便牽

涉到決策選擇。

狀態觀察與行動

在決策過程有三種考慮。一，觀察當前的環境狀況（Observation）；二，以此基礎採取行動（action）；三，「智慧體」採取行動後，外部變化所產生好或壞的結果，常以「回饋值」（reward）來表示，簡單地可用「收益好壞」來理解。以上文的起床為例，早起，我可以完成某項工作；懶床不起來，完成不了工作，但可享受多睡片刻的樂趣。怎麼選擇好？顯然是以收益多寡與結果好壞作取捨圭臬。

人工智慧決策也是基於上述的邏輯，通過數據，幫助「智慧體」去觀察理解，在什麼情況下，採取什麼樣的行動，可以獲得最大的收益。譬如，在商業物流上的應用，如何把貨物從 A 點運送到 B 點，物流路線那一條可以最短，成本最低而取得更高的利潤。人工智慧的智慧決策都可派上用場，文章開首時提及東方航運與微軟的合作，利用 AI 節約成本，從中可見其一斑。

AlphaGo 也是據此來完成與柯潔在圍棋上對弈的決策任務，下文再作一些技術解釋。

預測與決策結合一體

再舉一個智慧體決策的例子。譬如，我參觀故宮博物館，我當然希望能在最短或有限時間內，把所有名畫都悉數瀏覽，無一錯過。如何安排觀賞路線，從而能夠達至理想目標？對我們來說，在當下情況，採取某項行動，會考慮兩方面因素，一，當前的收益；二，未來的潛在收益。

在故宮，我當下所在的位置，我看到一幅名畫，很開心，這是當前的收益，但我想前往另一處，參看另一幅名畫，但這幅畫所在的位置卻比較偏遠，我需要走一段長路，浪費了較多時間，有可能延宕了我接著參觀的腳程，預測代價較大，使得未來所得樂趣的收益減少，於是我考慮是否改往二樓，這裡正舉行一個畫展，名畫很多，我預測可能得到很多樂趣，在這一刻的時點，我需要作出決策。決策的目的，當然是希望看到更多名畫所帶來的樂趣，結果，或選擇爬上二樓。

這個例子的決策過程，牽涉把預測和決策能力結合，並以當下與未來收益作出綜合考量，以採取行動。把預測能力和決策能力結合起來，無疑是深度增強學習的核心。

多變環境的明智決策

看深一層，傳統決策過程，無疑也包含了一，觀察當前環境狀態；二，以此基礎採取行動；三，成本與收益衡量。只有瞭解三種因素後，才能作出決策行為。但傳統的決策方法，面對數量巨大的狀態和決策選擇，因為無法精准預測當前狀態，無法精准預測行動收益，往往難以勝任。

在今日大數據時代，環境狀態特別多，決策選擇也多，譬如，在故宮參觀，於路線決策一刻，你向前走，還是向後走？向左走或向右走？是向上走或是往下走？除了決策選擇眾多，收益也難以估計。我走這一步，收益有多少，走另一步，收益又有多少？

總言之，環境狀態多、決策選擇也多、收益難以準確衡量下，如何處理？很顯然，決策時刻，對於環境變化的預測能力便起關鍵的作用。深度學習精准的預測能力在此處便體現了巨大的優勢。 將深度學習與增強學習兩者的優勢結合，這也是「深度增強學習」過人之處。

無人車暢行其道

深度學習和增強學習結合的濫觴，成功開端，可以說由 DeepMind 團隊，在 2013 年「神經資訊處理系統大會（Conference and Workshop on Neural Information Processing Systems・NIPS）上，發表「Playing Atari with Deep Reinforcement Learning」一文為標誌。該文其中位研究團隊成員 David Silver——如今是谷歌（Google）旗下 DeepMind 團隊最出色的成員之一，也是 AlphaGo 的開發者，他強調，人工智慧未來的發展大趨勢，就是深度學習與增強學習的結合。亦即 AI= DL+RL（Deep Learning+ Reinforcement Learning）。

預測能力可以通過深度學習得以加強。雖然傳統的決策也可以根據觀察、行動和收益計算來進行，但在環境狀態多、決策選擇多和收益無法可靠衡量下，便難以有效勝任。然而，深度增強學習所取得對環境變化的精準預測，不僅提升了商業行為決策的能力，更重要是「智慧體」，如 AlphaGo、無人車以至機器人等，能夠發揮強大的預測和決策能力。譬如，無人車，透過攝像鏡頭來感知路面交通環境，從而根據每一次的觀察，作出決策，是停車、左轉、或右轉……。深度增強學習使人工智慧發展，邁向一個更具應用潛力的新階段。

AlphaGo 如何擊敗柯潔

技術的簡單表述是，「智慧體」按照當前的環境觀察，以確定下一步的行動（action）。每一次的環境觀察，是為「智慧體」所處的狀態（State）。因此，狀態和動作存

在「映射關係」。簡單來說，就是一個「狀態」可以對應一個「行動」，或者對應不同動作的概率；而概率最高者，往往就是最值得執行的動作。用估值函數公式來表達，可以寫成——Max Q（s, a）=current reward（當前收益） + future reward（未來收益）。

估值函數中，s 為狀態（state），a 為行動（action）。換言之，由觀察狀態到採取行動的過程，也就是策略選擇的決策。智慧體會選擇採取行動 a，使得在當前狀態 s 下的收益最大化。很顯然「深度增強學習」，就是如何在多變環境中，取得最好的決策收益結果。

AlphaGo 在圍棋對弈中，棋逢敵手。關鍵是，在當前棋盤的狀態下，如何走下一步棋，蘊含對每一步棋的落子預測，對每一步棋落子的收益估值，AlphaGo 通過計算機數據演算法，對環境變化〈棋局〉作出精準預測，往往取得更高勝算，令柯潔三度飲恨，含淚甘拜下風。

商業應用前景遼濶

當前，谷歌更運用了 DeepMind 的深度增強學習技術，毋須改變機房設計，只需靠軟體動態調整，就使得冷卻設備整體耗電減少 40%的成效。

在谷歌的資料中心，最大的耗電量，是進行冷卻降溫，因為谷歌一個資料中心，動輒有上萬台伺服器，產生大量的熱能，為了讓伺服器持續正常運作，必須靠冷卻裝置來降溫。然而，這些冷卻設備多是大型的機電設備，像是抽風機、冷水機和冷卻塔等。但如此複雜和高度變動環境中，很難準確地操作這些機電設備來降溫。DeepMind 的深度增強學習，精準找出設備與機房環境狀態的對應關係，令耗電量明顯下降。

可以看到，深度增強學習除了讓「智慧體」，包括機器人、無人車等能夠因應外部狀態變化而作出智能行動外，在商業領域的應用，無論是應用範圍和應用前景，都非常廣泛。

當然，人工智慧也有其局限性，例如，需要大量數據支援，更重要一點是，人工智慧的技術，雖然在預測能力上近年有所精進，但目前在對「反饋值」，亦即收益的估值中並不是那麼精確。因此，即使人工智慧在當前取得飛躍進步，但始終不能視之為解決所有問題的「萬靈藥」，而我們應以「實是求是」的態度，來直視和促進這項技術的未來發展，為人類社會服務，也為商業領域作出更具效率和收益的決策行為。

〔本文由科大商學院傳訊部筆錄，楊毅教授口述及整理定稿〕