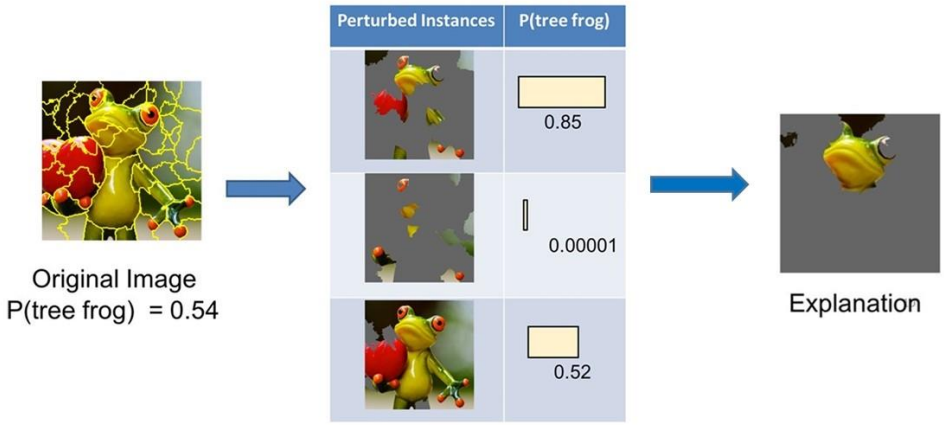
標題：掌握 AI 的思維
關鍵字：人工智慧、黑盒子、神經網路

人工智慧在現代社會中的地位日趨重要，利用類神經網路發展出來的深度學習(Deep Learning)功不可沒，然而人工神經網路的複雜程度如同人腦，科學家們從外部無法精準解釋它在下決定時的依據，就好比黑盒子般令人難以看透；雖然這代表人工智慧的技術大幅提升，但其思維的不透明性也帶來了不確定性與不安全感。Mark Riedl 曾說：「如果我們不問他們做事的原因並得到合理的回覆，人們便會將他們退貨。」，因此，詮釋人工智慧的想法成了一門新的課題，科學家們紛紛打造探索黑盒子的工具，種種方法發展成一種科學的研究，也就是所謂「AI 神經科學(AI neuroscience)」。

探索黑盒子—擾亂輸入，整理輸出

一位西雅圖華盛頓大學研究生 Marco Ribeiro 試圖打開神秘的黑盒子，透過改變輸入 AI 的文字或圖像來觀察輸出的辨識結果變化，開發出 LIME(Local Interpretable Model-Agnostic Explanations)作為解釋的工具，可對影像辨識做出「大量含有某些特徵的圖就是某個東西的影像」的解釋，也可以對文字分類做出「大量含有某些關鍵詞的文章就屬於某個分類」的解釋。



(圖一、利用 LIME 解讀相片中的哪些特徵會使 AI 判斷該圖片含有樹蛙)

假設我們想利用 LIME 解釋用來辨別圖片中是否含有樹蛙的分類器。首先我們會先將圖片分解成具解讀性的組件(如圖一左)，接著會將組件以不同的組合關閉產生不同的擾亂實例(perturbed instance，如圖一中間)，而對於每一個擾亂實例，我們都可以找出該實例可能包含樹蛙的機率(如圖一中的 P 行)；最後，我們會推舉出表現最好的實例作為解釋，而在這個例子中，樹蛙頭被作為解釋(圖一右)，也就是說當其他圖片中有出現類似的樹蛙頭時，該圖片有高機率會被分類器判斷含有樹蛙。

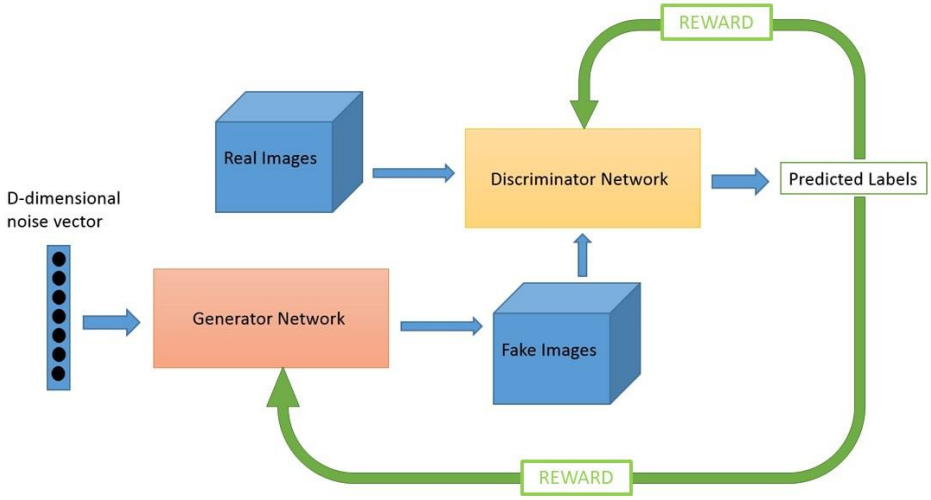
控制黑盒子—掌握已知，將曖昧透明化

面對黑箱子所帶來的困擾時，不一定要選擇解釋 AI 的想法，也可以試著降低神經網路的不確定性。一位「透明盒子(Glass Box)」專案的領袖 Maya Gupt 曾經提出單調插值查表(Monotonic interpolated lookup tables)，其中記錄著不同參數間的單調關係(monotonic relationship，即正相關或負相關)，例如購買商品的總價會隨著商品數量增加而上升，或者尿意會隨著體內的缺水程度上升而減弱；Gupta 將它提供給神經網路，使得在訓練神經網路時不需要再重新學習那些已知的訊息，而當特定資料已經先預定好，人們便更能掌控神經網路的學習走向，也有效縮小黑盒子的範圍。

機器解釋機器—模型間的合作

在舊金山 Uber 總部中，Jason Yosinski 設計一款影像的分類器套用在自駕系統上，用以辨識路上的各種物件，例如：斑馬線、紅綠燈、消防栓等等，他讓神

經系統記憶大量已標記的影像，訓練它的辨識能力，好讓自駕系統能精準掌握路況。但是，分類器的神經網路如同前文所言，像個黑盒子般神秘，而當面對這種狀況，Yosinski 選擇用另一個人工智慧幫他解釋原本的神經系統。
首先，Yosinski 的團隊重新調整分類器的工作，將「辨識影像」改為「產生影像」，再對模型投入彩色雜訊(colored statics)，並設定輸出影像的分類(例如：火山)；他們期待這模型可以將雜訊捏成類似火山形狀的影像，而結果就某程度來說它確實完成工作，產生了一張能被分類到「火山」的圖片，但就人類肉眼所見，輸出的影像還是很像雜訊，這也意味著「機器與人類看到的不一樣」。



(圖二、GAN 中生成器與辨別器的運作)

接著，Yosinski 的團隊使用了生成式對抗網路(GAN)，其中包含兩個互相競爭的神經網路—「產生器」(Generator)與「辨別器」(Discriminator)：產生器會先從訓練圖片集中合成指定的影像，企圖欺騙辨別器—該圖片是真實的而非合成產生的，而辨別器也會不斷學習以增強自己的辨識能力，對抗產生器的欺騙。當合成的影像被辨別器識破時，產生器就會改進自己的技術，產出更逼近真實的影像；相反的，當合成影像蒙騙成功，辨別器就會修正判斷的標準，增強自己的識別能力。(即圖二的 reward)

利用 GAN 可以彌補在訓練神經系統時真實資料的不足，也減少工程師標記資料分類的時間，而且也解決了黑盒子的問題，有助於讓我們了解模型學習到的是哪些特徵；但事實上，GAN 還是會產出一些錯誤輸出，而 Yosinski 解釋這可能是跟訓練資料或神經系統的問題有關，反覆類似的實驗就可以發現問題的根源並改正，「這些提示將會是 AI 神經科學上重要的方向」，他說，「這只是個開始，就如同一面空白地圖上的一小點。」

參考資料：

1. Paul Voosen, “How AI detectives are cracking open the black box of deep learning”, Science, 7 July 2017

2. Marco Tulio Ribeiro and Sameer Singh and Carlos Guestrin, “Introduction to Local Interpretable Model-Agnostic Explanations (LIME)”, O’Reilly, 12 August 2016

3. Jon Bruner and Adit Deshpande, “Generative Adversarial Networks for beginners”, O’reilly, 7 June 2017