機器考托福：關於深度學習

文・圖／李宏毅

隨著深度學習技術的發展，機器的能力愈來越強，深度學習的應用除了眾所周知的AlphaGo以外，當你上傳一張照片到Facebook，Facebook可以自動把人臉標示出來，這也是深度學習的功勞。除了影像辨認外，深度學習也被用於讓機器裡解人類語言，例如：機器可以自動從文章中擷取關鍵詞彙[1]；給機器聽一段新聞或上課錄音，機器可以自動產生摘要[2]（註：語音文件摘要的發展已有很長一段時間，但過去機器只能從語音檔中選取重要片段產生摘要，但今日機器已可以在理解新聞內容後用自己的文句撰寫摘要）；Apple的Siri、微軟的Cortana、亞馬遜的Alexa等語音助理也都需要深度學習的技術。深度學習是機器學習的一種，本文不詳述深度學習技術，有興趣的讀者可以參考筆者為數理人文雜誌撰寫的科普文章[3]，簡單說來，深度學習模型以類神經網路模擬人類大腦神經元的結構，使得機器有學習能力。

語音內容: *Many people have been fascinated about Venus for centuries because of its thick cloud cover ..... Well, what's under those clouds? First of all, let me talk about how we have been able to get past those clouds ..... Radar can get through the clouds. So what have we learned? ..... The level of sulfur dioxide gas above Venus's clouds shows large and very frequent fluctuations. It is quite possible that these fluctuations, the huge increase and decrease of sulfur dioxide, happening again and again. It's quite possible that this is due to volcanic eruptions, because volcanic eruptions often emit gases. If that's the case , volcanism could very well be the root cause of Venus's thick cloud cover .....*

問題: *According to the professor, what is a possible origin of Venus' clouds?*

選項:

A. Gases released as a result of volcanic activity
B. Chemical reactions caused by high surface temperatures
C. Bursts of radio energy from the plane's surface
D. Strong winds that blow dust into the atmosphere

圖1：語音內容的部分為人工聽寫所得到的文字，全文長達800多字，此處只節錄和答題有關片段。本題正確答案是 A，你答對了嗎？這個考題機器也答對了，如果你沒有答對，那麼你的英文聽力就沒有機器強了。

令人好奇的是，有了深度學習技術後，機器可以聽懂人類的語言到什麼程度呢？如果學生要出國留學，通常要考英文能力檢定，以檢定結果證明自己的英文能力，因此我們也讓機器去做托福聽力測驗的試題，藉此了解機器可以聽懂多少人類語言。在托福聽力測驗中，受試者先聽一段聲音，內容可能是學校生活對話（例如：某位學生和宿舍管理員討論問題）或是教授講課片段，接下來根據考題，從4個選項中選出正確答案，圖1為真實的考題。我們希望透過深度學習技術來完成上述任務，人類只負責提供教材（大量英文文章和托福聽力測驗的考古題），不需要撰寫程式告訴機器如何答題，機器可以自己根據教材學習[4][5]。本研究和曾柏翔、方為、徐瑞陽、沈昇勳等同學以及李琳山院士共同完成。

機器如何根據語音內容回答問題呢？

機器要理解的對象是語音，故須先以語音辨識技術將聲音訊號轉成文字儲存下來，以作為答題之用。語音辨識已是目前手機必備功能，讀者一定都不陌生，現在語音辨識已全面採用深度學習技術，只要給機器準備大量的語音訊號及其對應文字，就可以自動學會語音辨識。

然而僅僅把聲音訊號轉成文字，機器仍只是「聽而不聞」，並不知道文字所代表的意思。如何讓機器知道一個詞彙所代表的意思呢？透過大量閱讀，機器可以根據詞彙的上下文自動學到詞意，閱讀的文章越多，學習成果就越好，所以在考托福聽力測驗之前，機器就已經讀過大量英文文章，知道多數英文單字的意思，這裡我們所用的技術為Glove向量[6]。在臺大語音實驗室，我們嘗試讓機器閱讀大量PTT的文章後，機器真的可以學到部分鄉民用語，例如：當我們要機器輸出和「本魯」意思最相近詞彙時，機器的答案是「小弟」、「小妹」、「敝人」等詞彙（註：「本魯」是「本魯蛇」的簡稱，就算你本來不知道「本魯」是什麼意思，看了機器的答案後應該略知一二了）；如果我們問機器，「魯蛇」之於「loser」，等於「溫拿」之於什麼，機器的答案會是「winner」。機器不只可以知道單一詞彙的意思，還可以知道整個句子的意思，這裡需要用到較為複雜的類神經網路如結構遞歸神經網絡（Recursive Neural Network）

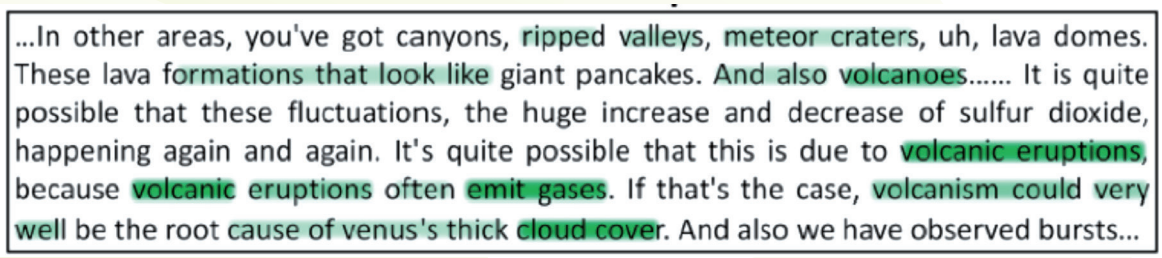

圖2：機器在回答圖1的問題時，在文章內容上所畫的「重點」。此處重點以綠色表示，顏色越深就代表機器覺得越重要。圖片來源[4]

等來頗析文句，有興趣的讀者可以參考文獻[5]。

有了上述的文句理解技術後，輸入一個問題，機器會根據問題內容再透過語音辨識得到的文章中尋找相關資訊，也就是在文章上畫「重點」，如何選取重點的位置也是透過類神經網路來完成，這個技術稱之為「專注」（Attention），因為該技術會讓機器學會只「專注」在和當前任務有關的資訊上，而忽略無關的資訊。該技術被廣泛的應用於深度學習，這個技術不只可以用在文字處理上也可以用在影像，它可以學會專注於影像中重要物件而不受背景干擾。接下來，機器會閱讀文章數次，不斷修改重點，然後根據畫好的重點產生答案，最後看答案和哪一個選項最相近，就選那個選項。圖2是機器回答圖1中的問題時所畫的重點（真實實驗結果）。

機器的測驗成績如何呢？

在使用深度學習技術前，我們先試一些較簡單的方法。第一個方法是，機器完全不看文章，直接比較問題和4個選項的相近程度，這個方法得到 24.6% 的正確率，因為托福聽力測驗有4個選項，所以亂猜答對的機率是 25%，所以這個方法跟亂猜沒兩樣。第二個方法是補習班的教法，機器從文章中取出和問題有最多同樣詞彙的段落，然後再看哪一個選項和該段落有最多重複的詞彙，這個方法的正確率是 31.2%，比隨機好一些。至於透過深度學習模型機器可以得到什麼樣的成績呢？在聽了717 題考古題後（考古題和實際測驗的題目是不同的），在施測時可以達到的正確率是 48.8%（註：類神經網路的訓練過程具有隨機性，故看同樣的教材每次訓練出來的結果都不同，此處是平均值，單次最高是55.7%）。目前的測驗成績大約是兩題答對一題，雖然這樣的英文程度大概申請不到美國名校，但是遠比隨機亂猜還要好。

托福聽力測驗的問題類型根據官方分類有3種，圖3為不同結構的類神經網路在這3類問題上的答題正確率，在此我們不討論不同結構所造成的差異，而是討論不同結構在同類題型上的一致之處。我們發現機器普遍在第二類題型上表現較差，

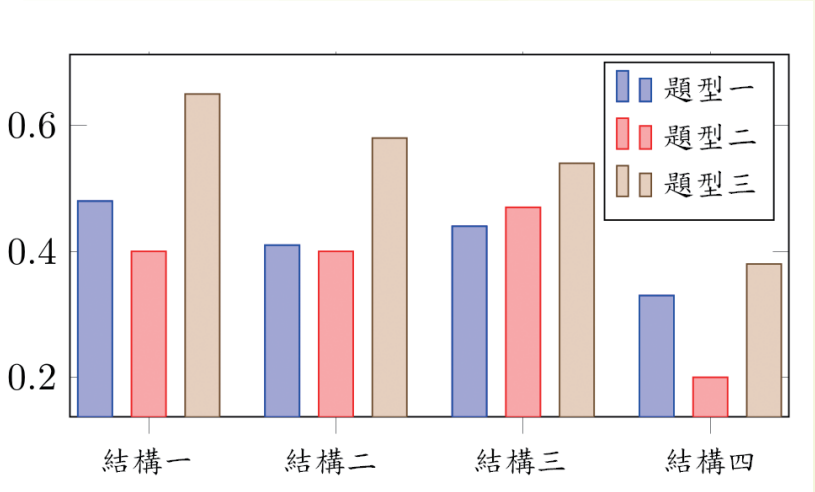

圖3：不同結構的類神經網路在托福聽力測驗的3種問題類型的答題正確率。圖片來源[5]

第二類題型是「語用」題，該類題目是要測驗考生可否了解語言中的「弦外之音」，而非語言本身的表面意義，要答對這種問題需要人類社會的基本常識，對只有看過考古題的機器來說，要答對這類問題是強「機器」所難。第三類題目考的是篇章結構、概念間的關係、推論等，機器較有可能從考古題中學會作答，故在這類題目上表現最好。

結語

機器沒有需要申請出國，所以讓機器去考托福聽力測驗似乎沒有什麼實際的用途，但是如果有一天機器可以在聽力測驗上拿高分，同樣的技術可以有很多應用，例如：給機器聽課程的錄音，它聽懂後，就可以擔任助教來回答學生的問題。（本專題策畫／電機系簡韶逸教授&生命科學系鄭貽生教授&醫技系方偉宏教授）

參考文獻：

[1] Sheng-syun Shen, Hung-Yi Lee, "Neural Attention Models for Sequence Classification: Analysis and Application to Key Term Extraction and Dialogue Act Detection", INTERSPEECH, 2016

[2] Lang-Chi Yu, Hung-Yi Lee, Lin-Shan Lee, "Abstractive Headline Generation for Spoken Content by Attentive Recurrent Neural Networks with ASR Error Modeling", SLT, 2016

[3] 李宏毅，"什麼是深度學習？"，數理人文第10期，P.20

[4] Bo-Hsiang Tseng, Sheng-syun Shen, Hung-Yi Lee, Lin-Shan Lee, "Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine", INTERSPEECH, 2016

[5] Wei Fang, Juei-Yang Hsu, Hung-Yi Lee, Lin-Shan Lee, "Hierarchical Attention Model for Improved Machine Comprehension of Spoken Content", SLT, 2016

[6] Jeffrey Pennington, Richard Socher, Christopher D. Manning, "Glove: Global Vectors for Word Representation", EMNLP, 2014

李宏毅小檔案

現任臺大電機系助理教授，研究方向是以機器學習技術讓機器辨識並理解語音訊號的內容。以深度學習技術為基石，致力於語音數位內容搜尋、語音數位內容之自動化組織以及從語音數位內容擷取關鍵資訊等前瞻性研究，這些技術有很多應用，例如：人機互動、問答系統、智慧型線上教學平台等等。