知識天地

音樂自動作曲的新展望——對抗式生成網路

楊理迦研究助理、楊奕軒副研究員（資訊科技創新研究中心）

摘要

在深度類神經網路（Deep neural network，通稱深度學習）崛起的時代，對抗式生成網路（Generative adversarial network, GANs）[1]的成功在過去三年獲得研究領域大量的注視。一對抗式生成網路由一判別模型（Discriminator）與一生成模型（Generator）所組成。判別模型借助類神經網日趨成成熟的模式識別（Pattern recognition）之能力，引導生成模型掌握真實資料的機率分佈，並加以生成模擬該分布的樣本。

本文將簡單依序介紹生成模型相關研究的重要性、對抗式生成網路以及其優勢、近期學界使用對抗式生成網路在影像領域的應用、以及我們在本院資創中心「音樂與音訊運算實驗室」使用對抗式生成網路對於音樂自動作曲的應用。

人類對生成模型（Generative modeling）的期盼

如果機器能夠自動產生照片、聲音甚至是帶有美感含義的畫作、音樂，將代表「人工智慧」達成了人們心中對於「創造」這項能力的期待。除此之外，在科學的角度上，生成模型的重要性代包含了相當重要的物理意義以及廣泛的應用價值。

物理意義上，真實世界中的事物可能擁有極高的維度。舉例來說，加速度方程式能夠被簡化為“F=ma”，但如果要定義「和諧的聲音」，其維度以及參數定義可能是超出傳統科學的認知的。而當一個生成模型能成功模擬目標事物間接代表了該模型能夠成功的掌握目標的機率分佈。能夠掌握高維度的機率分佈在應用數學以及工程領域將會是一大重要突破。

另一方面，在2012年的顯示卡加速技術突破後，資料科學以及機器學習的領域開始能夠乘載更大的資料量以及參數量，深度類神經網路中更大量的資料乘載量開啟了能夠訓練出更強大模型的可能性。而如果生成模型能夠產生大量的高度擬真樣本，將能再次突破現今深度學習領域中的對於資料量的瓶頸。除此之外，生成模型可以相當輕易的與半監督學習（Semi-supervised learning）結合，當資料的標記（label）量不完整時，生成模型在訓練掌握真實資料所擁有的歸納能力往往可以扮演成功的半監督學習模型的角色。

對抗式生成網路概念、優勢以及困難

由簡單的minimax博弈理論為出發點，一個基本的對抗式生成網路（GANs）包含了一個負責判斷樣本真偽的判別模型以及一個以產生可以騙過判別模型的「假樣本」的生成模型。這對模型組合在訓練中相互拉扯，生成模型從判別模型的結果中獲得回饋，並且依照判別模型給予的回饋漸漸逼近真實數據的分佈；另一方面，判別模型也因生成模型的進步，不斷更新達到更強大的判別能力。

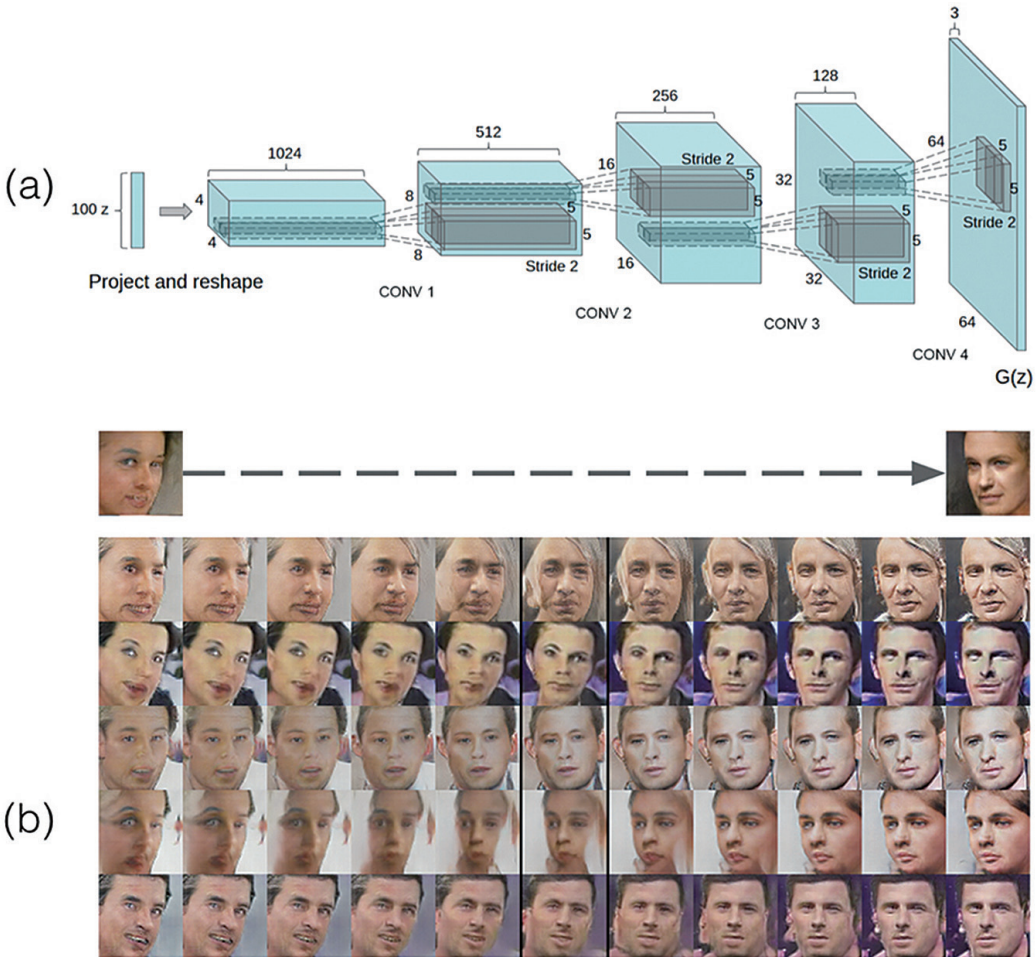
具體實作中，生成模型僅僅以一個雜訊z向量為其輸入，並且經由層層的網路設計，輸出一個與真實資料同樣規格的樣本。而判別模型則是分別將生成模型的輸出樣本（假樣本）以及真實資料為輸入，輸出「真」或「假」的判斷結果。然而生成模型經由以最大化判別模型的失誤為目標，判別模型反之的框架下，兩者將會達到所謂的奈許平衡（Nash equilibrium），這時生成模型所產生的樣本也與判別模型達到訓練中的最高相似度。

有鑒於這項設計，生成式對抗網路擁有不需給予明確參數定義（建模）的優勢，僅僅依賴一個自定義長度的雜訊向量，生成式對抗網路擁有相當高的自由度。舉例來說，相比於傳統上的變分自動編碼器（Variational autoencoder, VAE）生成模型框架，生成式對抗網路不需要單一具體的真實資料來作為媒介，即可產生高度變化的樣本。

然而，擁有高度自由的生成式網路同時背負了龐大的風險。基於兩模型互相拉扯的框架，生成式對抗網路的訓練過程往往難以控制，並且相當不穩定。舉例來說，當判別者無法有效判斷真假或反之，當生成者遲遲無法騙過片別者時，訓練將陷入無意義的更新迴圈。因此，如何藉由不同的訓練技巧以及模型變化增強並且穩定生成式對抗網路成為訓練成敗的關鍵。

影像生成的重量級突破

論及生成式對抗網路在影像生成上的重要里程碑，就必須一提2016年問世的DCGAN[2]架構。承襲了積卷式神經網（Convolutional neural networks, CNNs）在影像辨識中的強大能力，以及生成式對抗網路的概念，DCGAN成功地在影像生成上達到了驚人的表現。DCGAN中的判別模型是由一個常見的CNN分類架構所構成，而生成模型則是利用了反積卷（Transpose convolution）的設計，將輸入之雜訊向量逐層放大為目標樣本的影像尺寸。圖一分別為:（a）DCGAN的生成模型架構圖，以及（b）該模型在影像（人臉）生成上的成果。



圖一：(a)DCGAN的生成模型架構圖(b)該模型在影像（人臉）生成上的成果

在圖一（b）的實驗中，研究者將生成樣本中分別取了一張「向左看」以及「向右看」的臉部圖，並且將兩者個輸入雜訊取了若干個等距離的值。在此實驗中，發現雜訊向量中的「方向」可以表現出生成模型在生成樣本上的變化，而這些變化，是可以被賦予實質意義的。以此例則是臉部的方向。這項特性除了除了表現出DCGAN對於真實影像的模擬能力外，也代表了其高度自由度背後的應用淺力。

音樂自動作曲上的驚人表現

音樂，是一個同時具有了大量的理論基礎、卻又擁有無際自由的美學藝術。自動作曲在歷史上一直都是音樂相關研究領域最熱門也最具爭議性的話題。自生成式對抗網路的概念推出後，本實驗室不斷試圖從中獲得音樂創造與自動作曲全新詮釋的靈感，並且在近期得到了令人期待的正面結果。

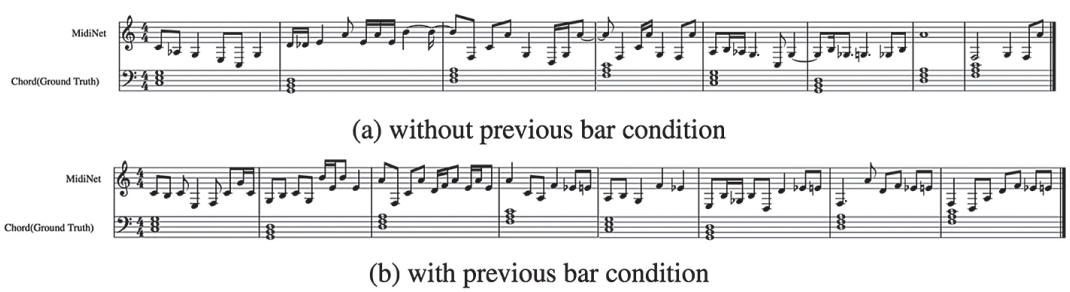
回歸人本：創造音樂是一件經由大量的自身經驗以及學習才能擁有的能力。以此為出發點，我們將生成式對抗網路中的生成模型想像成一位剛接觸音樂的孩子，並且以如何教導孩子學音樂為概念來設計我們的自動作曲模型。

首先，一個初學音樂的孩子往往都會先接觸一些簡單的音樂旋律，例如筆者兒時接觸鋼琴印象最深刻的

旋律就是上下課的鐘聲，看似零散的片段卻能是建構音樂的基礎。接觸一定數量的旋律後，我們試圖將範圍稍微拉大，配合著樂理的基礎，開始介紹和弦以及段落小節的概念，並且開始訓練孩子如何透過和弦的組合配合旋律產生悅耳的音樂。

透過這樣的概念，我們將生成式對抗網路接上了！同樣的，判別模型扮演著給予生成模型「好」或「壞」標準的老師角色，但除此之外，我們加入了和弦、小節以及段落的資訊。在具體應用上，我們將資料中每小節的旋律對應上了該小節的和弦標記，並且以最基本的八小節為段落來分割歌曲資料。當這樣的資料設計應用到生成式對抗網路時，我們將和弦的標記資訊加入到了生成模型以及判別模型中。接下來，我們設計了另一組類神經網，這組類神經網負責學習每個段落中小節與前一個小節間的關聯性，並且將學到的資訊也加入了生成模型中。換句話說，結合以上設計的框架下，生成模型能夠同時學習「不同和弦下的適當旋律」以及「與先前內容連貫的旋律」。

在圖二中，我們簡單展示兩段由生成模型創作的八小節的片段，圖二（a）為同時給予「和弦資訊」以及設定成「考慮先前小節旋律」的結果，而圖二（b）為只給予「和弦資訊」所產生的結果。由圖二中我們可以除了可以看出生成模型能夠掌握不同小節下符合樂理的音符組成外，還能明顯在設定成考慮先前小節內容時會產生較有連貫性的旋律。



圖二：音樂自動作曲生成範例

我們將此計畫命名為MidiNet，此論文已於西元2017年十月發表於International Society for Music Information Retrieval Conference [3]。我們另外在以下網址中提供了現階段的實驗成果音擋可供讀者直接線上聆聽：

https://richardyang40148.github.io/TheBlog/midinet_arxiv_demo.html

MidiNet的延伸版，MuseGAN，可以更進一步產生五個音軌的流行音樂段落，包含鋼琴、吉他、貝斯、鼓、以及其他弦樂。此論文甫於西元2018年二月以口頭論文方式發表於人工智慧領域頂尖會議Association for the Advancement of Artificial Intelligence [4]。可至下列網址聆聽MuseGAN所產生之音樂：

https://salu133445.github.io/musegan/

未來展望-AI音樂家

有鑒於我們在目前的生成式對抗網路設計上保有了相當大的資料適應能力與可擴充性，未來我們將加入更完全的音樂訓練資料，例如：樂風、情緒等等，如同繼續教導著一個學音樂的孩子般，希望有朝一日能夠成為更貼近人性的AI音樂家的角色。除此之外，我們更希望能夠藉由這項研究帶給音樂教育、音樂創作界的更多啟發以及幫助。

參考文獻

[1] I. J. Goodfellow, “NIPS 2016 tutorial: Generative adversarial networks,” arXiv preprint arXiv:1701.00160, 2017.

[2] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep con- volutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[3] L.-C. Yang, S.-Y. Chou, Y.-H. Yang, “MidiNet: A convolutional generative adversarial network for symbolic-domain music generation,” in Proc. Int. Society of Music Information Retrieval Conf., pp. 324-331, 2017.

[4] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, “MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment,” in Proc. AAAI Conf. Artificial Intelligence, 2018.