

國立宜蘭大學資訊工程學系

碩士論文

Department of Computer Science and Information

Engineering

National Ilan University

Master Thesis

基於生成對抗網路的兒童英語繪本系統

Generative Adversarial Network-based Children's English Picture

Book System

研究生：藍珮瑄

Graduate Student：Pei-Hsuan Lan

指導教授：卓信宏 博士

Advisor：Hsin-Hung Cho Ph. D.

中華民國 111 年 7 月

July 2022

國立宜蘭大學碩士學位論文
指導教授推薦函

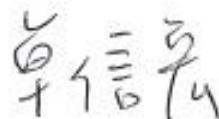
資訊工程學系碩士班 藍珮瑄君所提之

論文（題目）： 基於生成對抗網路的兒童英語繪本
系統

Generative Adversarial Network-based
Children's English Picture Book System

係由本人指導撰述，同意提付審查。

指導教授



（簽章）

系所主管



（簽章）

中 華 民 國 1 1 1 年 0 4 月 2 5 日

國立宜蘭大學碩士學位論文

口試委員會審定書

資訊工程學系碩士班藍珮瑄君所提之

論文（題目）：

基於生成對抗網路的兒童英語繪本系統

Generative Adversarial Network-based Children's English Picture Book System

經本委員會審議，認定符合碩士資格標準。

學位考試委員

曾繁勛

游家欽

卓信辰

蔡邦維

陳顯元

指導教授

卓信辰

中華民國111年7月9日

摘要

隨著英語能力在國際上的高度重要性，我國兒童學習英語的年齡逐漸降低，在兒童階段就開始接觸英語已成為趨勢，並隨著科技的日新月異，學習環境也從傳統學習發展出了數位學習，學習英語的方法變得越來越豐富，目前已有許多使用資訊技術方法來輔助英語學習的系統。但並非所有的學習方式都適用於兒童，由於許多兒童對於陌生的英語詞彙大多不感興趣，故兒童願不願意學習英語成了重要的課題。目前已有許多的英語學習系統，不過效果皆有限，甚至大部分的系統主要適合已有一定能力或是年紀較長的學習者。由於兒童的英語認知能力與成人不同，因此對於兒童的英語學習而言，應該設計適合他們的學習模式。對於兒童而言，他們的世界充滿了想像以及文字與圖像的各種結合，我們也發現大部分兒童對於漫畫風格圖像會比文字更感興趣，若是將這些英語詞彙轉換成漫畫風格圖像，則能藉此讓兒童從這些生澀的英語中產生興趣，並能讓兒童從圖像中學習英語。然而聘請漫畫家替英語繪本製作內容並不是件容易的事，除了須花費的成本過高，生產力也相當低落。因此本論文提出一個基於生成對抗網路的兒童英語繪本系統，並提出了在圖像生成後提取圖像特徵的方法以生成具連貫性圖像，並以行動學習的方式供兒童學習英語。透過由教師端自行輸入繪本故事內容，即能自動生成對應的漫畫圖像，並自製成繪本儲存於資料庫中，學生端則能透過行動裝置選取欲閱讀的繪本，藉由圖像學習英語以提高兒童的學習興趣。本系統分為 Client 端和 Server 端，Client 端是以行動裝置作為人機介面，可再細分為教師端和學生端，Server 端則為本系統之演算法、GAN 模型、網頁伺服器以及資料庫伺服器。在實驗中，為了求得最佳的輸出結果，我們系統性地調整了模型的超參數，例如平衡目標函數中兩個損失的超參數以及 Epoch 大小，使本系統能夠根據輸入的故事內容生成品質更高的連貫性圖像。

關鍵字：生成對抗網路、兒童英語學習、數位學習、兒童英語繪本

Abstract

With the increasing importance of English proficiency in the world, the age at which children in Taiwan begin to learn English is gradually decreasing, and it has become a trend to start contacting English at a young age. Due to the advancement of technology, the learning environment has also developed digital learning from traditional learning, and the methods of learning English have become more and more abundant. There are many systems that use information technology methods to assist English learning. However, not all learning styles are suitable for children. Since many children are mostly not interested in unfamiliar English vocabulary, whether children are willing to learn English has become an important issue. At present, there are many English learning systems, but the effectiveness is limited, and even most of the systems are more suitable for learners who already have a certain level of English or are older. Since children's English comprehension ability is different from that of adults, it is necessary to design a learning mode suitable for children's English learning. For children, their world is a world of whimsy and various combinations of words and images. We also found that most children are more interested in comic-style images than text. If these English words can be converted into comic-style images, children can be interested in the incurious English leaning, and children can learn English from images to increase learning efficiency. However, hiring cartoonists to draw content for English picture books is not an easy task. This is because in addition to the high cost, productivity is also quite low. Therefore, this thesis proposes a children's English picture book system based on generative adversarial network (GAN). This thesis also proposes a method to extract image features after image generation to generate coherent images, and these images can be used in mobile learning (M-Learning) for children to learn English. The teacher can input the story content of the picture book and then the corresponding comic image can be automatically generated. Finally, the picture book can be created and stored in

the database. Students can select the picture books they want to read through their mobile devices and learn English through images to enhance children's interest in learning. The system is divided into Client side and Server side. The Client side uses the mobile device as the human-machine interface, which can be subdivided into the teacher side and the student side. The Server side is the algorithm, GAN model, web server and database server of the system. In the experiment, in order to obtain the best output results, we systematically adjusted the hyperparameters of the model, such as the hyperparameters of the two losses in the balance objective function and the size of the Epoch, so that the system can generate quality according to the input story content higher corresponding images.

Keywords : Generative Adversarial Network, Children's English Learning, Electronic Learning, Children's English Picture Book

誌謝

首先，我要感謝我的指導教授卓信宏老師，在研究的過程中總是給我許多建議與想法，讓我的研究可以更加順利，當我有問題想找老師討論時，老師也總是不厭其煩的給予指導，除了在研究方面的指導，老師也時常帶實驗室的學生去吃甜點，讓我們可以緩解一些在研究上的壓力，由衷感謝老師的指導與對學生的用心。其次，我要感謝我的口試委員游家牧教授、蔡邦維教授、曾繁勛教授以及陳麒元教授，感謝口委們給予的建議與回饋，使我的論文可以更加完整。還要特別感謝陳麒元教授以及曾國鈞教授在每次的實驗室聯合 Meeting 時給予的指導。

接著，我要感謝實驗室的學長與學弟們，感謝旻諺學長教我寫演算法以及解決我程式上的問題，感謝江毅平時的幫忙，讓我能夠專心的完成論文，還要感謝陳麒元教授實驗室的助理明嫻，協助我解決行政上的事項。感謝我的男朋友鈺智，在我研究上遇到困難時，總會跟我一起討論，並且給予我各方面的幫助。

最後，我要感謝我的家人們，感謝爺爺和奶奶對我在宜蘭求學時的照顧，感謝爸爸讓我能無後顧之憂地完成學業，感謝媽媽一直以來的支持與鼓勵，由衷感謝你們。

目錄

摘要	I
Abstract	II
誌謝	IV
目錄	V
表目錄	VII
圖目錄	VIII
第一章 緒論	1
1.1 簡介	1
第二章 相關背景與文獻	4
2.1 英語學習	4
2.2 數位學習	7
2.3 人工智慧於數位學習之應用	9
2.4 Text-to-Image 生成系統	11
2.5 生成對抗網路	13
2.5.1 基本架構	14
2.5.2 訓練過程	15
2.5.3 條件生成對抗網路	18
2.6 生成對抗網路 Text-to-Image	19
2.7 文獻總結	24
第三章 研究方法	25
3.1 系統架構	25
3.2 生成對抗網路模型	27
3.2.1 文字處理	27
3.2.2 注意力生成網路	29
3.2.3 深度注意力多模態相似性模型	32
3.2.4 繪本圖像生成	35

3.3 介面設計	37
3.3.1 教師端	37
3.3.2 學生端	39
3.4 系統資料庫	41
3.5 系統流程	43
第四章 實驗	44
4.1 實驗資料	44
4.2 實驗環境	46
4.3 評估指標	47
4.4 實驗	49
4.4.1 模型超參數調整	53
4.4.2 實驗結果	57
第五章 結論與未來展望	71
參考文獻	72

表目錄

表 2-1 四種條件下聆聽簡短故事[9].....	6
表 2-2 E-Learning 學習環境比較表	8
表 2-3 Text-to-Image 圖像生成方法比較表.....	12
表 2-4 GAN 符號表[39]	15
表 2-5 GAN Text-to-Image 比較表	22
表 3-1 Attentional Generative Network 符號表[54]	29
表 3-2 DAMSM 符號表[54].....	32
表 4-1 實驗環境設置表.....	46
表 4-2 GAN 參數設置表	49
表 4-3 第一階段生成網路模型結構	49
表 4-4 第二階段生成網路模型結構	50
表 4-5 第三階段生成網路模型結構	50
表 4-6 第一階段鑑別網路模型結構	51
表 4-7 第二階段鑑別網路模型結構	51
表 4-8 第三階段鑑別網路模型結構	52

圖目錄

圖 1-1 全球語言使用排名	1
圖 2-1 年齡與第二語言學習表現關係圖[6].....	5
圖 2-2 歷年 GAN 相關文獻發表數量	14
圖 2-3 GAN 架構圖	15
圖 2-4 GAN 訓練過程	17
圖 2-5 CGAN 架構圖	18
圖 3-1 系統架構圖	26
圖 3-2 AttnGAN 架構圖	27
圖 3-3 BiLSTM 示意圖	28
圖 3-4 DAMSM 示意圖	32
圖 3-5 Inception-v3 架構圖	33
圖 3-6 繪本圖像生成過程	36
圖 3-7 教師端介面 (a)輸入英語故事 (b)選取重點單字	38
圖 3-8 教師端介面 (a)生成的圖像 (b)儲存於資料庫的繪本	39
圖 3-9 學生端介面 (a)儲存於資料庫的繪本 (b)選取的繪本內容	40
圖 3-10 SQLAlchemy ORM 框架	41
圖 3-11 資料表的屬性	42
圖 3-12 系統流程圖	43
圖 4-1 Irasutoya 網站中的圖像	44
圖 4-2 實驗圖像與對應的五句句子描述	45
圖 4-3 添加不同擾動等級的高斯雜訊對 FID 值的影響	48
圖 4-4 $\lambda=50$ 損失值變化	54
圖 4-5 $\lambda=20$ 損失值變化	54
圖 4-6 $\lambda=5$ 損失值變化	55
圖 4-7 $\lambda=1$ 損失值變化	55
圖 4-8 $\lambda=0.1$ 損失值變化	56

圖 4-9 不同 Epoch 之 FID 值	57
圖 4-10 不同 Epoch 之 R-precision 值	57
圖 4-11 <i>A boy's day</i> (a)輸入英語故事 (b)選取重點單字	59
圖 4-12 <i>A boy's day</i> (a)第一頁 (b)第二頁	60
圖 4-13 <i>A boy's day</i> (a)第三頁 (b)第四頁	61
圖 4-14 <i>A boy's day</i> 第五頁	62
圖 4-15 <i>The bathing boy</i> (a)輸入英語故事 (b)選取重點單字	63
圖 4-16 <i>The bathing boy</i> (a)第一頁 (b)第二頁	64
圖 4-17 <i>The bathing boy</i> (a)第三頁 (b)第四頁	65
圖 4-18 <i>The bathing boy</i> 第五頁	66
圖 4-19 <i>I don't want to walk home</i> (a)輸入英語故事 (b)選取重點單字	67
圖 4-20 <i>I don't want to walk home</i> (a)第一頁 (b)第二頁	68
圖 4-21 <i>I don't want to walk home</i> (a)第三頁 (b)第四頁	69
圖 4-22 <i>I don't want to walk home</i> (a)第五頁 (b)第六頁	70

第一章 緒論

1.1 簡介

現今我們處於一個全球化的時代，無論在世界何處的任何發明或生產的任何東西，它都需要一個共同語言與世界各地的所有人交流，根據[1]統計，2015 年英語使用者已超過 20 億人口，其中以英語為母語的人口約為 4 億，儘管世界各地說中文和西班牙語言的人口比說英語的人口多，如圖 1-1 所示，但由於英語目前為工程、學術以及貿易等眾多領域的首選語言，因此可以明顯得知英語是目前世界上最廣泛使用的語言，隨著英語能力在國際上的高度重要性，我國兒童學習英語的年齡逐漸降低，在兒童階段就開始接觸英語已成為趨勢，一方面可以自幼就開始培養國際觀，另一方面也可以避免學習英語年齡太晚而造成學習效率不彰的情形。

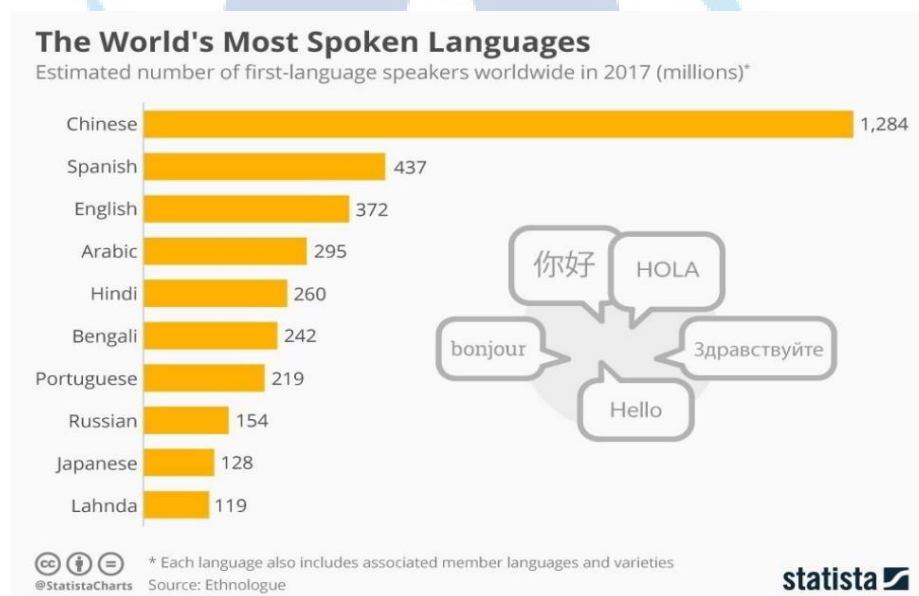


圖 1-1 全球語言使用排名

隨著科技的日新月異，學習環境也從傳統學習發展出了數位學習 (Electronic Learning, E-Learning)，E-Learning 是指經由數位媒介來進行學習，無論是在學習的地點或時間上都變得更有彈性，並且可以根據自身的學習

狀況重複學習，基於各種優勢 E-Learning 也被廣泛地應用在英語學習上。學習英語的方法變得越來越豐富，目前已有許多使用資訊技術方法來輔助英語學習的系統，常見的系統包括智能輔導系統(Intelligent Tutoring System, ITS) [2][3][4]和自適應學習系統(Adaptive Learning System, ALS) [5][6][7]。ITS 能以智能的方式給予多種類型的幫助和反饋，可在沒有教師干預的情況下為學生提供即時且適當的指導，例如[2]提出之系統可提供學生有關被動語態的學習問題，而系統輔導主要關注的是學生的錯誤診斷過程，當學生鍵入練習的答案時，系統會檢查答案的正確性，如果學生的答案是錯誤的，系統會診斷錯誤的原因，為了提供個性化的幫助，系統會為每個學生保存一個檔案，學生的進步和常見錯誤會被記錄到這個長期學生模型中，這些資訊可用於學生在後續課程中的錯誤分析。ALS 是使用電腦演算法來調整與學習者的互動，並提供客製化的學習資源來解決每位學習者的需求，例如[5]提出了個性化的上下文感知行動學習系統以輔助學生學習英語，基於上下文感知為不同的學習者提供適應性的內容，在他們的模型中，上下文包括位置、時間、管道以及學習者的知識，透過自適應的方式，學習者將獲得滿足其需求的自我調整內容。由上述可知，資訊的發展已為學習環境和學習模式帶來了巨大的衝擊及改革。

但並非所有的學習方式都適用於兒童，由於許多兒童對於陌生的英語詞彙大多不感興趣，故兒童願不願意學習英語成了重要的課題，[8]透過對幼兒英語教學現狀的分析，作者發現了一些問題，例如難以激發幼兒學習興趣、缺乏教學情境以及學習效率低等問題。目前在坊間已有許多的英語學習系統與媒體，不過效果皆有限，甚至大部分的系統主要適合已有一定能力或是年紀較長的學習者，例如[9]提出之個性化行動英語詞彙學習系統，可以根據學習者個人的詞彙能力和記憶週期推薦適合的英語詞彙進行學習。[10] 提出之英語學習系統可根據學習者的熟練程度來個性化教材內容。但這類的系統通常需要具備一定的英語認知能力，由於兒童並不具備一定的英語認知能力，因此這類的系統並不適用於兒童的學習。[11]提出了一個結合擴增實境(Augmented Reality, AR)和無所不在學習(Ubiquitous Learning, U-Learning)的英語學習系統，以提高真實情境下英語學習的效果，並利用

他們提出的系統研究了學習策略和認知風格是否會影響使用此系統的學習表現，實驗結果表明，學習策略和學習者的認知風格是會影響學習表現的，具有特定領域知識的學習者會比其他學習者更適合使用此學習系統。由上述研究可知，認知能力是會影響學習效果的，因此對於兒童的英語學習而言，應該設計適合他們的學習模式。

對於兒童而言，他們的世界充滿了想像以及文字與圖像的各種結合，也就是兒童對視覺圖像擁有更豐富且靈巧的理解[12]，我們也發現大部分兒童對於漫畫風格圖像會比文字更感興趣，若是將這些英語詞彙轉換成漫畫風格圖像，則能藉此讓兒童從這些生澀的英語中產生興趣，並能讓兒童從圖像中學習英語，讓英語的學習事半功倍。然而聘請漫畫家替英語繪本製作內容並不是件容易的事，除了須花費的成本過高，生產力也相當低落，因此本論文提出的兒童英語繪本系統結合人工智慧(Artificial intelligence, AI)的方式，利用低成本即可提供一套自動化的兒童英語學習繪本系統，並以行動學習(Mobile Learning, M-Learning)的方式供兒童學習英語。由於繪本是以圖像為主體並結合文字共同描述一個完整的故事，而繪本故事中通常會有幾個角色，這些角色在每頁故事的圖像中應保有相同的外觀以符合故事的描述，因此圖像之間須有連貫性，我們提出一個基於生成對抗網路(Generative Adversarial Network, GAN)的系統並提出了在圖像生成後提取圖像特徵的方法以產生具有連貫性的圖像。透過由教師端自行輸入繪本故事內容，即能自動生成對應的漫畫風格圖像，並能自製成繪本儲存於資料庫中，學生端則能透過行動裝置選取欲閱讀的繪本，藉由圖像學習英語以提高兒童的學習興趣，讓英語的學習事半功倍。

本論文分為五大章節，其組織架構如下：第一章緒論，說明研究背景與動機及論文架構，第二章相關背景與文獻，說明本研究的相關背景知識以及生成對抗網路 Text-to-Image 的研究近況，最後進行文獻總結，第三章研究方法，詳細說明本系統的架構與流程，第四章實驗，介紹本實驗所使用之資料集、評估指標、實驗設計以及展示實驗結果，第五章結論與未來展望，將對本論文進行總結並說明未來展望。

第二章 相關背景與文獻

隨著英語重要性的提高，擁有基本的英語能力已是必備技能，當前透過 E-Learning 的方式來學習已越來越普及，比起傳統的學習方式，透過 E-Learning 能夠提高學習效率，並且能隨時隨地的學習。由於 AI 的迅速發展，它也對教育領域產生了影響，現今已有許多加入 AI 的 E-Learning 應用，但較少有針對兒童設計的系統，且有許多學者提出圖像能夠增強兒童的記憶以幫助學習，目前已提出許多 Text-to-Image 系統。因此本論文欲透過 GAN 技術結合 E-Learning 來改善兒童的英語學習效果並輔助兒童學習英語。本章節將介紹英語學習、數位學習、人工智慧於數位學習之應用、Text-to-Image 生成系統、生成對抗網路的相關背景及概念以及本論文中使用的生成對抗網路 Text-to-Image 方法的相關研究，最後進行文獻總結。

2.1 英語學習

在我國以英語為第二語言是目前普遍的趨勢，英語也是世界上最多人使用的第二語言，第二語言通常是作為輔助語言或是通用語，而學習第二語言可以增強在國際化職場上的競爭優勢。許多兒童在幼兒時期就已接觸英語，但也有許多兒童是上小學後才真正開始學習英語，因此也造成許多父母的疑問，到底從幾歲開始讓兒童學習英語比較合適。[13]研究發現，0~7 歲是人類學習語言的關鍵期，7 歲以前的兒童完全有能力可以同時學習兩種語言，而當兒童 7 歲以後，學習語言的能力會急速下降，到了 17 歲之後，基本上將會失去第二語言的學習天賦，如圖 2-1 所示。但作者也表示並不是指過了青春期的就無法再學習第二語言，而是過了青春後，大腦的學習機制會發生變化，也就無法再像 0~7 歲的兒童能內隱學習(implicit learning)，只需要在新語言的環境中接受大量的新語言輸入，就能學習該語言，反之，過了青春後，人類變成需要外顯學習(explicit learning)，也就是需要明確的語言規則或指令來學習新的語言。

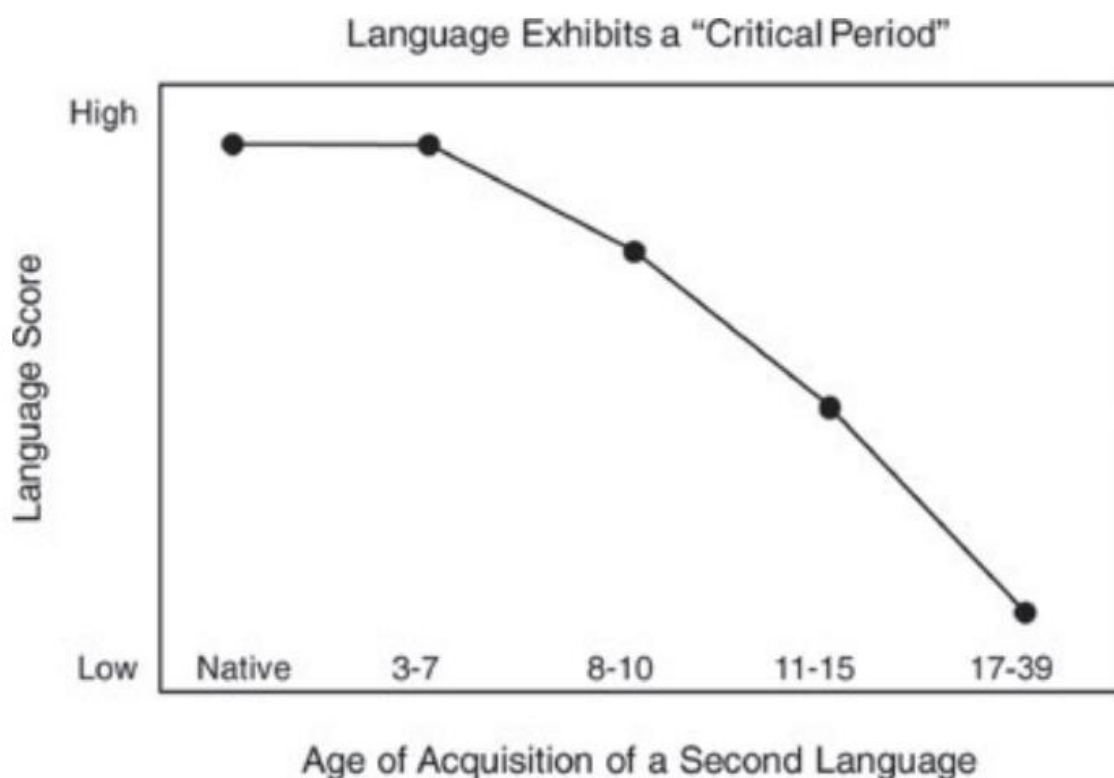





圖 2-1 年齡與第二語言學習表現關係圖[6]

由上述可知，從小就讓兒童接觸英語是具有優勢的，但由於兒童與成人的認知能力不同，因此需要選擇適合他們的學習方式，[14]提到閱讀是語言學習的重要部份，從小透過故事激發兒童學習英語的興趣，可以使兒童開闊視野並增強兒童對英語使用的意識，[15]表明人腦處理圖像的速度比處理文字的速度快，且圖像比文字更有可能保留在長期記憶中，因此使用圖像來輔助語言學習能夠提升學習效率，[16]透過眼動追蹤實驗，針對 41 名平均年齡為 64 個月的兒童，研究了在四種條件下聽了簡短的故事的實驗，分別是(1)口頭敘述與敘述一致的圖像；(2)口頭敘述與不一致的圖像；(3)只有圖像但沒有口頭敘述，以及 (4)只有口頭敘述，如表 2-1 所示，研究結果顯示，他們發現一致的圖像對兒童故事的複述有很大的貢獻，眼動追蹤數據顯示，兒童以探索圖像的方式可以有效地整合敘述和圖像。

表 2-1 四種條件下聆聽簡短故事[9]

	On screen	Oral narration
Congruent condition	 <p>Willem Wout kijkt nog even naar de reuzen poppen in de etalage. Hij had ook graag nog even gekeken naar het bad in het poppenhuis en het konijn in de poppenwagen. Buiten zit een hondje. Willem Wout wil hem graag aaien maar mama wil naar huis. Ze sjouwt de tas vol met nieuwe kleren. Nu nog gauw een ijsje eten en dan hollen naar de bus.</p> <p><i>[Willem Wout is looking at the giant dolls in the in the window.etcetera]</i></p>	<p>Willem Wout is looking at the giant dolls in the window. He also would like to have another look at the bath in the dollhouse and the rabbit in the doll wagon. Outside is a dog sitting. Willem Wout wants to pet him but mom would like to go home. She is carrying a bag full with new clothes. Let's eat an ice cream quickly and then run to the bus.</p>
Incongruent condition	 <p>Willem Wout kijkt nog even naar de reuzen poppen in de etalage. Hij had ook graag nog even gekeken naar het bad in het poppenhuis en het konijn in de poppenwagen. Buiten zit een hondje. Willem Wout wil hem graag aaien maar mama wil naar huis. Ze sjouwt de tas vol met nieuwe kleren. Nu nog gauw een ijsje eten en dan hollen naar de bus.</p> <p><i>[Willem Wout is looking at the giant dolls in the in the window.etcetera]</i></p>	<p>Willem Wout is going to the dentist with mom. Bear Baboen is coming along, 'Will the doctor hurt me', asks Willem Wout. 'Well, no', says mom, 'You have been brushing your teeth every evening. And you didn't eat candies.' 'Does Bear Baboen have a hole?' 'Perhaps he does', mom says, 'He often forgets to brush his teeth and really likes licorice.'</p>
Picture control condition	 <p>Willem Wout kijkt nog even naar de reuzen poppen in de etalage. Hij had ook graag nog even gekeken naar het bad in het poppenhuis en het konijn in de poppenwagen. Buiten zit een hondje. Willem Wout wil hem graag aaien maar mama wil naar huis. Ze sjouwt de tas vol met nieuwe kleren. Nu nog gauw een ijsje eten en dan hollen naar de bus.</p> <p><i>[Willem Wout is looking at the giant dolls in the in the window.etcetera]</i></p>	<p>No oral narration</p>
Written text condition	<p>Willem Wout kijkt nog even naar de reuzen poppen in de etalage. Hij had ook graag nog even gekeken naar het bad in het poppenhuis en het konijn in de poppenwagen. Buiten zit een hondje. Willem Wout wil hem graag aaien maar mama wil naar huis. Ze sjouwt de tas vol met nieuwe kleren. Nu nog gauw een ijsje eten en dan hollen naar de bus.</p> <p><i>[Willem Wout is looking at the giant dolls in the in the window.etcetera]</i></p>	<p>Willem Wout is looking at the giant dolls in the in the window. He also would like to have another look at the bath in the dollhouse and the rabbit in the doll wagon. Outside is a dog sitting. Willem Wout wants to pet him but mom would like to go home. She is carrying a bag full with new clothes. Let's eat an ice cream quickly and then run to the bus.</p>

由上述的研究可以發現，透過結合故事的圖像式學習能夠激發兒童的學習興趣並增加兒童語言學習的能力，因此本論文欲設計一套英語繪本系統來輔助兒童的英語學習。

2.2 數位學習

E-Learning 是指透過各種資訊科技工具來學習的方式，比起傳統學習，E-Learning 在學習時間上可以更有彈性，並且透過融入資訊技術能夠達到更高的學習效率，在成本效益方面，E-Learning 會隨著時間遞減成本但傳統教學不會，在 COVID-19 大流行的期間，E-Learning 更是發揮了重要的作用[17]，藉由遠距教學的方式減少人與人的接觸，讓學生也能達到預期的學習成效。

E-Learning 以學習模式區分可分為三種形式，分別為同步學習、非同步學習以及混合式學習[18]，同步學習是指教學者與學習者在指定時間內進行學習，而非同步學習在時間上則較有彈性，學習者可依照自己的需求進行學習，混合式學習則兼具同步和非同步學習之特性，透過多樣化的學習模式強化學習效果，採用何種學習模式則受使用對象及課程類型而定，但無論採用何種學習模式，若學習的過程缺乏互動性，將無法適用於學生及教師上。雖然 E-Learning 為我們帶來了不同的學習方式，但重點還是在於學習[19]，要如何運用資訊科技來提升教學者的教學品質和學習者的學習成效才是重點，所以單純把教材或影音內容放到網路上並不算是一種好的 E-Learning 方法，而是必須創造互動學習情境，讓學習者可以不斷地在學習環境中學習。

E-Learning 的發展在早期為使用電腦輔助教學(Computer-Assisted Instruction, CAI)利用與電腦交談式或互動式的方法來學習，到後來將教材燒錄至光碟中的數位教學內容，但教材與媒體整合性仍較低，隨著現今網際網路的發達發展出了線上學習(Online Learning)，讓學習者可以隨時上網學習提高了學習彈性，教學者也可以隨時回答學生的問題，提供了更多元的互動模式，透過網際網路即可上傳教材提高了教材與媒體的整合性，而隨著行動裝置的普及更發展出了 M-Learning，它是使用行動載具，經由無線網路與學習平台進行連結，讓學習者在學習上不必受到時間與空間的限制，在學習彈性方面又比 Online Learning 更高，各種 E-Learning 學習環境的比較如表 2-2 所示。

表 2-2 E-Learning 學習環境比較表

特性 學習環境	教學媒介	媒體形式	教材媒體 整合性	學習模式	學習彈性
傳統學習	黑板 投影片 影片	非數位化的 傳統學習	弱	同步式	低
CAI	電腦化教 學軟體	影音光碟	弱	非同步式	低
線上學習	個人電腦 電腦網路 數位教材	多媒體	強	混合式	中
行動學習	行動載具 無線網路	多媒體	強	混合式	高

隨著學習者為中心的觀念出現發展出了 E-Learning 2.0，學習者不再只是單向的接受與被動的參與而是自發性地學習，透過協同合作與討論達到由下而上的學習方式，學習者可以安排自己的學習進度和學習內容，也可以不斷反覆地聆聽，選擇性地加強複習內容。近年來，我國也不斷推廣自我調整學習(Self-Regulated Learning)，採取自我調整學習的目的是強調學習者能夠根據自身的認知及外在的動機調整學習，並能夠有效且適當的調節自我的學習方式[20]。學生在課業學習的過程中會遇到許多不同的外在影響因子，進而影響學生自我的學習效率，學習壓力、同儕關係以及身心狀況等都會成為學生學習困擾原因之一，這表示學生自我調整學習與學生的學習效率是息息相關的，自我調整學習策略可以幫助學生達到自我調整學習歷程預設目標。然而也是存在一些限制，主要是因為每位學習者的學習模式是不同的，如何讓每位學習者能擁有好的學習環境是非常重要的[21]。基於數位學習的各種優勢，本論文欲結合 E-Learning 中的 M-Learning 來輔助兒童的英語學習。

2.3 人工智慧於數位學習之應用

由於 AI 的興起，它也對教育領域產生了影響[22]，目前發展出了許多 AI 於 E-Learning 的應用，包括以下技術：智能輔導系統、自適應學習系統以及評估 E-Learning 環境之系統，大多數的研究都集中在智能輔導系統的開發和應用上，智能輔導系統可在不需要教師干預的情況下，為學生提供即時的反饋與指導。自適應學習是利用電腦演算法來調整與學習者的互動，學習教材會針對學習者的需求進行客製化設計，並適當調整教學環境以改進學習過程，有研究提出自適應學習對學習者的重要性[23]，若是將自適應學習加入 E-Learning[24]，則可以讓學習者更容易實現他們的學習目標，透過 AI 與 E-Learning 的結合，不但可以解決互動性的問題，還可以提供自適應學習。[25]主要是為了解決 E-Learning 在課程創建上花費大量的時間問題，作者提出了一種可以方便並且有效的課程創建管理系統，所提出的系統利用自然語言處理(Natural Language Processing, NLP)[26]和深度學習技術自動化生成 E-Learning 內容，首先，使用 NLP 和深度學習技術自動總結相關檔案之摘要，接著在生成的摘要中檢測關鍵字，再從摘要中刪除關鍵字，最後重新排列輸出結果，即可生成教學教材並將填充到 E-Learning 系統中。[27]提到在 E-Learning 中，學生的情緒常常會被忽視，因此作者使用卷積神經網路(Convolutional Neural Network, CNN)的深度學習方法來檢測臉部表情的情緒，此系統可分為情緒檢測模塊和教學策略調節模塊，在教師的教學過程中，情緒檢測模塊會檢測學生的情緒狀態，而檢測到的情緒將作為反饋回饋給教師，教師則能根據大多數學生的情緒來調整自己的教學策略和教學內容。[28]提出了一種基於深度學習的個性化學習模型來尋找合適的學習方法，他們利用深度學習和機器學習算法來分析學生資料，以找出資料之間的相關性，他們考慮了適應性學習、個性化學習、差異化學習和基於能力的學習等個性化學習因素。

由上述可知，採取資訊系統輔助學習是目前的教學趨勢，有許多研究都分析了科技的接受度，科技接受模型(Technology Acceptance Model, TAM)是以理性行動理論為基礎發展而來[29]，透過 TAM 可以解釋當使用者在電

腦科技中接觸新的資訊系統的行為，並分析出對於使用者有影響力的各項因素，更發展出預測使用者的行為模式及解釋使用者的接受度。TAM 有五個主要變數，分別為(1)認知有用 (perceived usefulness)；(2)認知易用 (perceived ease of use)；(3)使用者態度 (attitude toward using)；(4)行為意圖 (behavioral intention to use)；(5)外部變數 (external variables)。透過現有關於 TAM 研究可以發現 TAM 有以下幾項特徵：(1)透過人們使用電腦的行為，可以從行為的意圖進行推測及預測；(2)主要決定因素源於使用電腦行為意圖的認知有用性；(3)使用電腦行為意圖的認知易用則是次要的決定因素。總而言之，先預測及解釋資訊科技的接受狀況，透過去控制對應的方法來達到控制外在因子改變使用者自身的認知與想法進而提升使用者的接受度，這就是 TAM 的目的。TAM 除了用來探討使用者對於新科技決定因素的用途外，在 AI 系統盛行的現代，TAM 也常應用於 AI 相關應用系統的使用評鑑研究上。針對現階段有關 TAM 的研究現況，[30]表明，TAM 是一種有效且穩固的模型，此模型已被廣泛使用，這代表本論文使用 AI 協助兒童學習英語的方向是正確的。

2.4 Text-to-Image 生成系統

應用多媒體系統來輔助學習者學習已使目前趨勢，已有許多研究提出搭配圖像與文字來表達的效果高於僅使用文字表達[31]，在教育領域的研究者也支持這個論點，認為文字搭配圖像可以提高學習者的認知[32]。由於本系統為英語繪本，因此故事中的圖像須為卡通風格的圖像且圖像之間須具有連貫性，現今已有模型可根據文字生成圖像，但生成圖像的風格較為真實且生成的圖像為獨立的圖像不具連貫性，不適合應用在我們的系統，因此我們提出了在圖像生成後提取圖像特徵的方法，使得下一張圖像保有前一張圖像的特徵以生成具連貫性的圖像來解決此問題，目前已有許多從文字生成圖像的方法，以圖像生成方法區分可分為網頁搜索生成圖像、多媒體資料庫搜索生成圖像以及深度學習方法生成圖像。

以網頁搜索方法例如[33]提出的系統利用 NLP、電腦視覺以及機器學習，透過辨識文字，接著根據文字自動在網頁上搜索最相關的圖像，最後根據文字和圖像優化圖像佈局來組合這些物件。[34]利用網頁圖像集合開發了多媒體應用系統 Word2Image，他們的系統採用了包括相關性分析、語義和視覺聚類的各種技術，以生成多樣且有代表性的圖像集。[35]開發了一個系統 VizStory，它由三個步驟組成，首先，作者調查故事的敘述結構來分割整個故事，接著他們為每個分割細節選擇有代表性的關鍵詞，最後透過網頁圖像搜索，找到合適的圖像來構成文字的可視化。以上的系統皆是從網頁中去搜索匹配文字的圖像，以去表達文字描述，雖然可獲得具有多樣性的圖像，但生成的圖像皆是已存在的圖像，並且獲得的圖像可能風格不一致或不適合教學應用，在搜索適合的圖像上花費的運算及時間成本較高，較難搜索到適合的圖像。

以多媒體資料庫搜索方法例如[36]提出了一個基於概念圖匹配的多媒體文字到圖像移動學習系統，作者根據多媒體存儲庫中帶有圖像的文字以及用戶輸入的文字構建概念圖，基於兩個概念圖的匹配分數，對匹配的圖像進行相對排名。雖然以多媒體資料庫搜索可生成特定風格的圖像，搜索

成本也比以網頁搜索來的低，但由於圖像皆是從資料庫提取的，導致生成的圖像較缺乏多樣性及變化性。

隨著 AI 的快速發展，結合深度學習的方法也可達到從文字描述生成圖像，並且生成的圖像可以針對特定目標生成風格一致的圖像，不像以網頁搜索會導致生成圖像風格不一致，且藉由網路訓練的方式，可生成不同於訓練樣本的圖像，使生成圖像更具多樣性。GAN 是近年來研究中熱門的方法，目前已提出許多利用 GAN 的生成系統例如[37]提出的系統使用深度卷積生成對抗網絡(Deep Convolutional Generative Adversarial Network, DCGAN)生成新圖像，使用光學字元識別(Optical Character Recognition, OCR)引擎從兒童讀物中選取圖像和文字，並使用語素分析器對文字進行分類，最後將分類後的詞類與圖像的潛在向量進行匹配，DCGAN 即能生成與文字相關的圖像，[38]提出一個兒童故事可視化的系統，他們的系統包括常見動作和複雜動作之間的區別，這種區別有助於以一致和連貫的方式可視化場景中的對象及其關係，他們利用 GAN 為常見動作和複雜動作生成圖像和圖像序列。Text-to-Image 圖像生成方法比較如表 2-3 所示，基於 GAN 生成圖像的各種優勢，本論文欲使用 GAN 的方法生成圖像。

表 2-3 Text-to-Image 圖像生成方法比較表

論文編號	圖像生成方法	風格一致性	搜尋成本	圖像多樣性
[33]	網頁搜索	不一致	高	高
[34]				
[35]				
[36]	多媒體資料庫搜索	一致	低	低
[37]	GAN	一致	低	最高
[38]				

2.5 生成對抗網路

AI 的概念是在 1950 年代被提出的，它是能夠模仿人類智慧的技術，應用非常地的廣泛，到了 1980 年代出現了機器學習(Machine Learning, ML)，它可以透過以往的經驗和資料來學習並找出其中的規則，最後達成 AI 的方法，在近十年則發展出了深度學習(Deep Learning, DL)，它是機器學習的分支，是使用類神經網路為模組組成多層的類神經網路堆疊，而 GAN 便是深度學習的一種方法。GAN 是在 2014 年由 Ian Goodfellow 提出的[39]，主要由生成網路(Generator Network)與鑑別網路(Discriminator Network)組成，GAN 可視為一個生成模型也可當成分類模型，近年來發展非常迅速，是熱門的深度學習模型之一，如圖 2-2 所示，隨著 GAN 的快速發展，GAN 的應用也越來越廣，GAN 較常被應用在生成方面，如圖像或影像的生成、修復、融合、合成等，例如[40]可以生成出化妝風格的圖像，依照給定的化妝風格圖像即可轉換到非化妝臉部圖像並且同時保留臉部身份，以達到化妝轉化的效果。[41]可以針對圖像中的雨進行修復，以恢復成清晰細節的圖像。[42]將 GAN 和基於梯度的方法結合，提高了圖像融合的技術，可將兩個圖像融合在一起，由於 GAN 可從特定分佈生成自然圖像，但在捕捉紋理和邊緣等細節方面較弱，而基於梯度的方法在生成具有局部一致性的圖像方面效果較好，但生成的圖像較不自然，因此將 GAN 和基於梯度的方法結合可克服兩種方法的缺點，以提高融合圖像的逼真程度。而進階一點的應用則是輸入文字描述即能生成相應的圖像，將會在 2.6 小節詳細介紹。除了圖像生成外，GAN 也可應用於時間序列資料的生成，例如[43]提出了時間序列生成框架，它將非監督式 GAN 與監督式自回歸模型對條件時間動態的控制之特性相結合，可用來生成真實的時間序列資料。GAN 還可應用於物件辨識，例如[44]結合物件分割和物件生成技術來辨識物件，作者將圖像和可見區域作為輸入，可以生成整個遮擋物件的遮罩，生成網路和鑑別網路都以對抗的管道進行訓練，以生成遮擋區域的物件圖像。GAN 也可應用於分類問題上[45]，只需把鑑別網路的輸出層替換成 Softmax 分類器，假設訓練樣本有 n 類，在訓練模型時可把生成網路生成出

的樣本歸類為 $n + 1$ 類，在 Softmax 分類器上同時增加一個輸出神經元，用來表示輸入至鑑別網路為假樣本的機率，即可視為一個分類模型。

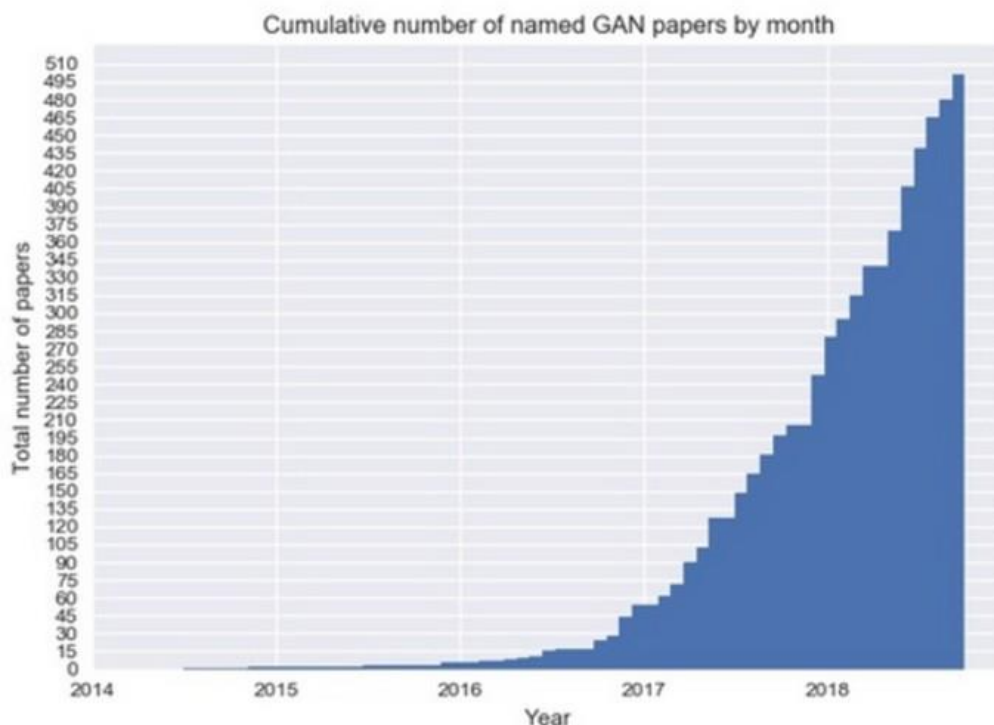


圖 2-2 歷年 GAN 相關文獻發表數量

2.5.1 基本架構

GAN 主要是由兩個神經網路組成，分別是生成網路與鑑別網路，如圖 2-3 所示，生成網路的輸入為隨機噪聲(Random Noise)，它的目標是了解真實樣本的特徵分佈並生成假樣本去欺騙鑑別網路，而鑑別網路的輸入為真實樣本和生成樣本，它的目標是分辨輸入樣本的真假。透過兩個網路互相對抗和學習，生成網路會根據鑑別網路的判斷結果進行改良，以生成出越來越接近真實分佈的樣本，而鑑別網路會藉由比較其判別結果與實際答案的差異來不斷提升自己的判別能力，以區分樣本的真假，直到最終鑑別網路無法區分生成網路所生成的樣本就可以說是一個完整的訓練。

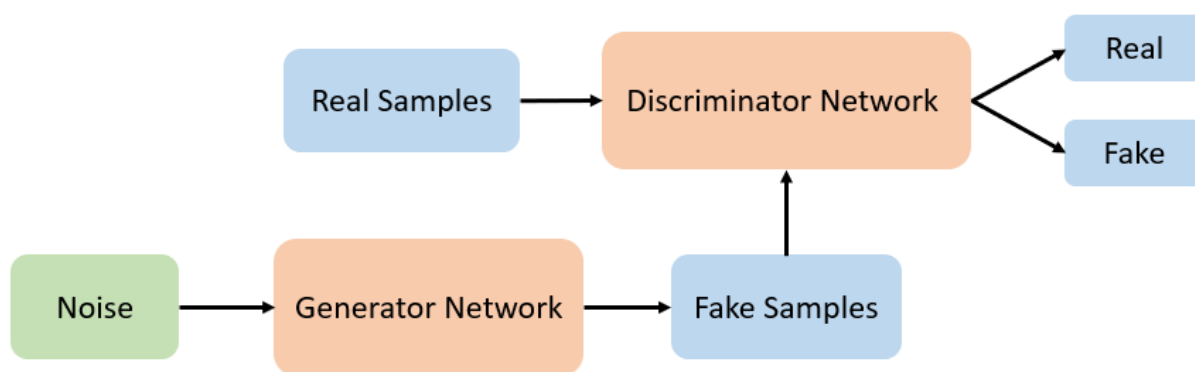


圖 2-3 GAN 架構圖

2.5.2 訓練過程

表 2-4 GAN 符號表[39]

G	生成網路
D	鑑別網路
r	真實樣本
$p_{data}(r)$	真實樣本機率分佈
z	隨機噪聲
$p_z(z)$	噪聲機率分佈
\mathbb{E}	期望值
∇	梯度運算子
θ_d	鑑別網路參數
θ_g	生成網路參數

GAN 的訓練過程可以視為 min-max game，生成網路和鑑別網路的目標是相反的，生成網路的目標是最求最小值，而鑑別網路的目標是求最大值，在訓練過程中兩個網路互相對抗進而不斷地優化直到達到平衡，GAN 的目標函數為公式(2.1)：

$$\min_G \max_D V(D, G) = \mathbb{E}_{r \sim p_{data}(r)} [\log D(r)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2.1)$$

其中， $D(r)$ 為鑑別網路判別真實樣本是否真實的機率，由於 r 就是真實樣本，因此理想情況下， $D(r)$ 越接近 1 越好，而 $\log D(r)$ 則會接近於 0。 $G(z)$ 為噪聲輸入生成網路後生成的樣本， $D(G(z))$ 則是鑑別網路判別生成樣本是否真實的機率，對於生成網路來說，它的目標是生成接近真實分佈的樣本，因此 $D(G(z))$ 越接近 1 越好，而 $\log(1 - D(G(z)))$ 則會越小，由此可知，生成網路的目標是最小化 V ，以達到以假亂真的目的，而鑑別網路的目標是最大化 V ，以得到判別能力更好的鑑別網路。

首先，初始化生成網路和鑑別網路的參數，在每一次的訓練迭代中，會固定其中一個網路的參數，只更新另一個網路的參數，之所以要限制生成網路和鑑別網路在訓練時只能調整自己的參數是為了避免其他參數造成的影響，這樣可以確保生成網路和鑑別網路都能得到明確的回饋資訊，而做出對應的調整。由於一開始鑑別網路的性能不佳，即便生成了樣本也無法判別樣本的真假，因此通常會先訓練鑑別網路，這時會固定生成網路，透過小批量隨機梯度上升法來更新鑑別網路參數，如公式(2.2)：

$$\nabla_{\theta_d} \frac{1}{n} \sum_{i=1}^n \left[\log D(r^{(i)}) + \log(1 - D(G(z^{(i)}))) \right] \quad (2.2)$$

從真實樣本隨機抽樣 n 個 r ，再取 n 個 z 生成假樣本 $G(z^{(i)})$ ，將 r 與 $G(z^{(i)})$ 放入鑑別網路，接著計算 $D(r^{(i)})$ 與 $D(G(z^{(i)}))$ 的分類損失，再利用反向傳播法，根據總誤差來調整鑑別網路的參數，最終的目標是要最大化公式(2.1)，因此前項 $\log D(r^{(i)})$ 表示要判別真實樣本越接近 1 越好，而後項 $\log(1 - D(G(z^{(i)})))$ 表示要判別假樣本越接近 0 越好。

接著訓練生成網路，這時會固定鑑別網路的參數，透過小批量隨機梯度下降法來更新生成網路參數，如公式(2.3)：

$$\nabla_{\theta_g} \frac{1}{n} \sum_{i=1}^n \log(1 - D(G(z^{(i)}))) \quad (2.3)$$

取 n 個 z 生成假樣本 $G(z^{(i)})$ ，接著計算 $D(G(z^{(i)}))$ 的分類損失，再利用反向傳播法，根據總誤差來調整生成網路的參數，生成網路的目標是要想辦法讓生成樣本騙過鑑別網路，也就是讓鑑別網路判別假樣本越接近 1 越好，最終的目標是要最小化公式(2.1)。

訓練完生成網路後，會再重新訓練鑑別網路，透過不斷重複訓練兩個網路，最終生成網路與鑑別網路都會得到最佳的結果，如圖 2-4 所示，在最理想的情況下，生成網路足以生成以假亂真的樣本 $G(z)$ ，而鑑別網路難以判別生成樣本的真假，此時 $D(G(z)) = 0.5$ 。

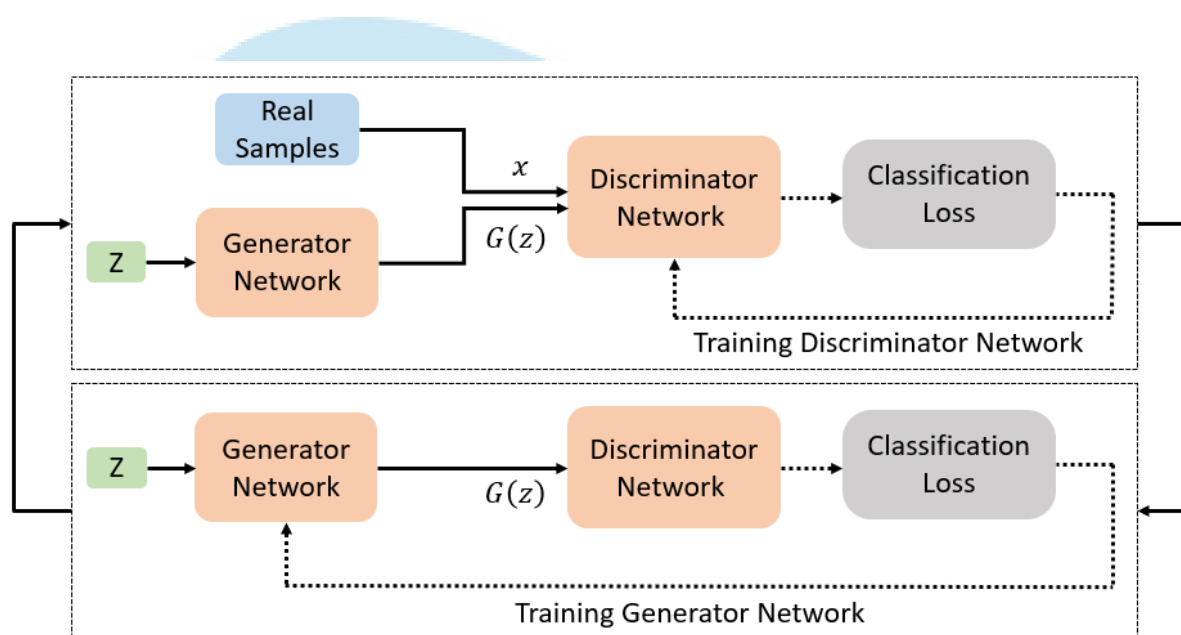


圖 2-4 GAN 訓練過程

GAN 的對抗訓練方式無需再要求一個假設的樣本分佈，而是使用一種分佈直接進行採樣，從而達到逼近真實樣本的分佈，然而這種不需預先建模方法的缺點是太過自由，對於較大的圖像和較多像素的情況下，基本 GAN 方法的結果就較不令人滿意，因此發展出了條件生成對抗網路 (Conditional Generative Adversarial Network, CGAN)。

2.5.3 條件生成對抗網路

由於基本 GAN 的輸出僅取決於隨機噪聲，沒有機制去控制它要生成的內容，因此[46]提出 CGAN 將 GAN 擴展為條件模型，它將額外條件添加到生成網路與鑑別網路中，額外條件可以是任何類型的輔助資訊，例如類別標籤或其它模態的數據，如圖 2-5 所示。

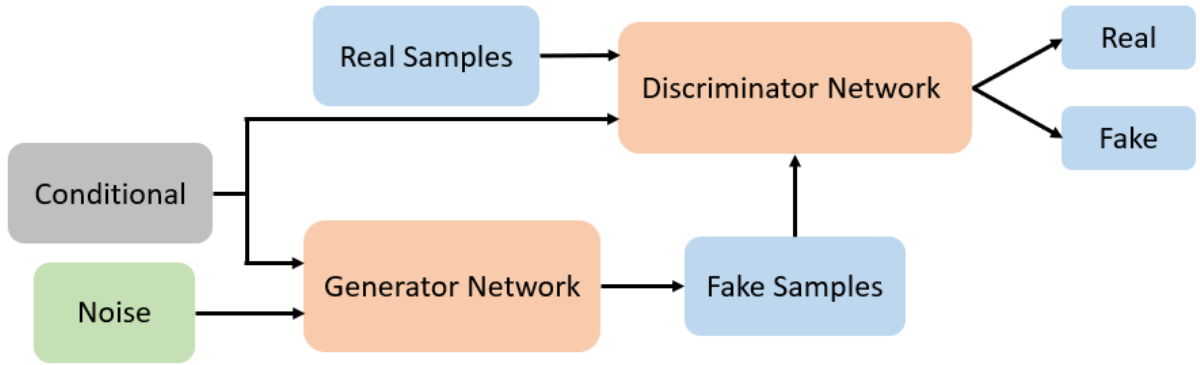


圖 2-5 CGAN 架構圖

CGAN 的目標函數[46]為公式(2.4)，與原始 GAN 類似，不同之處在於生成過程添加了額外條件 c 做控制，與原始 GAN 相比，CGAN 的目標函數轉變為條件概率，加入的條件不同，對應的目標函數也會不同，在加入條件變量的情況下，訓練方式從非監督式學習轉變為半監督式學習，提高了 GAN 的穩定性。

$$\min_G \max_D V(D, G) = \mathbb{E}_{r \sim p_{data}(r)} [\log D(r|c)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z|c)))] \quad (2.4)$$

由於本論文欲透過 GAN 從英語文字描述生成漫畫風格圖像，對於從文字生成圖像的 GAN 應用，是將文字描述作為條件變量加入至 GAN 模型中，將隨機噪聲和文字描述一起作為輸入，其中文字描述是用來確定主要內容，隨機噪聲負責變化，GAN Text-to-Image 將在 2.6 小節詳細介紹。

2.6 生成對抗網路 Text-to-Image

隨著 AI 技術發展越來越成熟，GAN 也延伸出許多擴展，現今已提出許多基於 GAN 的改良方法可以由文字生成對應的圖像 (Text-to-Image)，應用的範圍也越來越廣泛，GAN Text-to-Image 的應用例如[47]開發了一個多媒體資訊檢索系統，透過利用 GAN Text-to-Image 和 Image-to-Text 模型，開發的系統能夠讓用戶以語音作為輸入並且以高準確度輸出內容。[48]由於 GAN 的不可預測性，作者認為 GAN 在設計等需要創造力的領域具有很大的潛力，作者利用 GAN Text-to-Image 讓客戶只需要描述他們的需求，系統就可以生成出對應的室內設計圖。

在 GAN Text-to-Image 領域中無論使用何種 GAN 架構，都須先對文字進行預處理以得到文字特徵，進而從文字特徵生成後續的圖像，根據不同的文字訓練形式可分為以文字描述生成圖像和以場景佈局生成圖像。以文字描述生成圖像的方法是早期 GAN Text-to-Image 領域中最常見的方法，文字描述中通常包含不同物件之間的關係，最早期提出的論文為[49]，在當時 AI 技術應用於 Text-to-Image 還未發成熟，然而當時已經開發出通用的循環神經網絡(Recurrent neural network, RNN)[50]架構來學習和判別文字特徵，同時 DCGAN[51]已能生成特定類別的圖像，因此作者開發了一種深度架構 GAN，以促進文字和圖像建模的進步，他們在輸入中加入文字特徵作為生成網路和鑑別網路的約束，最終生成 64×64 的圖像，但生成圖像品質較低。隨著 GAN 的改良和進步，[52]使用兩個 GAN 來生成圖像，他們提出了一個兩階段的 GAN 架構 StackGAN，第一階段是根據給定的文字描述來繪製場景的原始形狀和顏色，從而生成低解析度圖像，第二階段會將第一階段的結果和文字描述作為輸入，並生成更多細節的高解析度圖像，並加入了條件增強(Conditioning Augmentation)，可緩解潛在空間中資料的不連續性，以提高資料的平滑性，但模型較難訓練且訓練時較不穩定。[53]提出了一種多階段 GAN 架構 StackGAN++，是由多個生成網路和多個鑑別網路組成，它們以樹狀結構排列，從樹的不同分支生成對應於同一場景的多個尺寸圖像，逐步生成更高品質的圖像，透過聯合多個分佈生成，StackGAN++比

StackGAN 有更穩定的訓練結果，但生成圖像較缺乏單字資訊。[54]提出之 AttnGAN 增加了注意力機制(Attention Mechanism)，不僅提取文字的句子特徵，同時還提取了單字特徵，透過描述中的相關單字來合成圖像中不同區域的細節，讓生成的圖像可以更突出文字中的細節，此外作者還提出了深度注意力多模態相似性模型(Deep Attentional Multimodal Similarity Model, DAMSM)來計算圖像和文字的匹配損失，讓生成的效果更好，但較難在多個複雜場景上生成圖像。[55]提出了 DM-GAN，它首先生成一個形狀和顏色粗糙的初始圖像，然後將初始圖像優化為高解析度圖像，當初始圖像不能很好地生成時，則利用動態記憶模塊進(Dynamic Memory Module)行圖像內容模糊的優化。他們還加入記憶寫入閥(Memory Writing Gate)可以根據初始圖像內容選擇重要的文字資訊，這使得它能夠從文字描述中準確的生成圖像，最後還利用回應閥(Response Gate)從記憶中和圖像特徵中融合讀取的資訊，但生成圖像的最終結果嚴重依賴於初始圖像的佈局。

雖然從文字描述生成圖像方面取得了不錯的進展，這些方法在生成簡單物件的圖像時效果較好，但很難在具有多個目標物件和關係的複雜句子上實現，由於資料的高維度，基於文字描述的生成在文字和圖像之間很難找到適合的映射關係，導致影響了訓練的穩定性，因此發展出了基於場景佈局的方法，透過提供額外的條件，例如物件的邊界框(Bounding Boxes)或物件之間的位置關係來控制圖像的生成。[56]提出了一種從場景圖(Scene Graphs)生成圖像的方法，能夠明確地推理物件之間的關係，他們的模型透過預測目標物件的邊界框和分割遮罩(Segmentation Masks)來計算場景佈局(Scene Layout)，並將 Scene Layout 加到 GAN 中生成圖像，該網絡還會針對鑑別網路進行對抗訓練，以確保圖像輸出的真實性，但生成圖像的品質較差且僅能生成 64×64 的低解析度圖像。[57] 提出了一種基於佈局的圖像生成方法稱為 Layout2image，只需給定大略的空間佈局，例如邊界框和物件類別，即可生成逼真的圖像，並且圖像的物件皆生成在正確的位置，每個物件分為確定部分(類別)和不確定部分(外觀)，首先使用詞嵌入對類別進行編碼，並將外觀提取為從正態分佈採樣的低維向量，接著使用卷積長短期記憶(convolutional Long Short-Term Memory, cLSTM)將各個物件表示組合

在一起，以獲得完整佈局的編碼，接著解碼為圖像，但生成圖像的解析度較低。相關的 GAN Text-to-Image 論文比較如表 2-5 所示。

由上述可知，無論是以文字描述或場景佈局為文字訓練方式，都有各自的優缺點，以文字描述形式訓練對於簡單描述或單個物件時生成圖像效果較好，而以場景佈局形式訓練可在具有多個物件的描述上生成圖像，但利用場景佈局形式生成圖像的訓練資料集需要做額外的處理，例如表 2-5 中利用場景佈局之論文皆使用 COCO 資料集[58]做訓練，COCO 資料集除了有圖像的文字描述外，還提供了圖像中每個物件的邊界框和分割遮罩，為圖像生成提供了額外的條件。

本論文是以生成漫畫風格圖像為目標，由於目前還未有這類的公開資料集，本論文是使用自製的資料集，對於圖像標注只提供了文字描述，且由於運算成本問題，利用場景佈局之訓練方式目前僅能生成 64×64 的圖像，對於本論文之繪本系統， 64×64 的圖像解析度無法清晰地顯示於系統上，因此選用的文字訓練形式為以文字描述的方式。在選擇 GAN 模型方面，考慮到多階段的訓練方式可以增加訓練的穩定性，因此選擇利用三個階段生成圖像的 AttnGAN，並且由於 AttnGAN 加入了注意力機制，能更突出文字中的細節，透過多階段的訓練方法，可以逐步生成更清晰的圖像，且加入 DAMSM 模型讓訓練有額外的監督，讓生成的圖像能更匹配於文字。雖然以文字描述生成圖像的方法較難生成多個複雜場景的圖像，但由於本論文的目標是製作兒童英語繪本系統，對於兒童繪本而言，繪本中的圖像應以簡單好理解為主，較少會出現多個複雜場景，基於上述綜合的考量，選用的訓練模型為 AttnGAN。

表 2-5 GAN Text-to-Image 比較表

論文編號	GAN 模型	文字訓練形式	輸出圖像解析度	特性	限制	資料集
[49]	GAN-INT-CLS	文字描述	64×64	最早提出的 GAN Text-to-image 論文。	生成圖像品質較差。	CUB Oxford-102 MS COCO
[52]	StackGAN		256×256	利用兩個 GAN 生成圖像，並添加條件增強可緩解潛在空間中資料的不連續性。	訓練較不穩定。	CUB Oxford-102 MS COCO
[53]	StackGAN++		256×256	利用三個階段生成圖像，可有更穩定的訓練結果。	生成圖像缺乏單字資訊。	CUB Oxford-102 COCO
[54]	AttnGAN		256×256	利用三個階段生成圖像，並添加注意機制可以更突出文字的細節。	較難生成多個複雜的場景。	CUB COCO

[55]	DM-GAN		256×256	利用兩個階段生成圖像，並加入動態記憶模塊進行初始圖像優化。	最終結果嚴重依賴於初始圖像的佈局。	CUB COCO
[56]	Sg2im	場景 佈局	64×64	利用場景圖生成圖像，能夠推理物件之間的關係。	生成圖像品質較差且解析度較低。	COCO Visual Genome
[57]	Layout2image		64×64	給定邊界框和物件類別即可生成圖像。	生成圖像解析度較低。	COCO Visual Genome

2.7 文獻總結

由上述的章節可以得知從小就讓兒童學習英語是具有優勢的，但由於兒童與成人的認知能力不同，因此須選擇適合兒童的學習方式，而結合故事的圖像式學習不但能激發兒童的學習興趣還能增加兒童語言學習的能力，因此我們欲設計一套兒童英語繪本系統。我們在文獻中比較了不同的 Text-to-Image 生成系統，由於 GAN 生成圖像的多樣性及風格一致性，我們最終選擇利用 GAN 生成圖像，並且已有許多研究提出利用科技技術輔助學習是具有優勢的，為了讓兒童能隨時隨地學習，我們欲結合行動學習的方式製作本系統。在選擇 GAN 模型方面，考慮到多階段的訓練方式能提高模型訓練的穩定性，因此選擇 AttnGAN。

由於目前神經網路模型訓練的限制，最終生成圖像的解析度為 256×256 ，雖然生成圖像目前只局限於 256×256 ，但能確保訓練的穩定性且生成較佳的圖像，若是只追求圖像解析度而盲目地上採樣圖像則會導致訓練不穩定，並生成不符合文字描述的圖像。在實務操作上，對於較大型的行動裝置可能會導致解析度不足，但本系統主要是針對繪本的概念而設計，因此會著重於圖像的生成品質及圖像中的連貫性去研究，解析度不足的部分將在未來中持續研究。

第三章 研究方法

本論文提出的兒童英語學習系統是以繪本為主要設計，繪本是以圖像為主體並結合文字共同描述一個完整的故事，因此一張圖像可以對應一句或多句句子，透過多張圖像連貫成一個故事。本論文提出一個基於GAN的英語繪本系統，可以根據輸入的英語故事內容生成對應的連貫性圖像，並自製成繪本，本章節將描述系統架構、生成對抗網路模型、介面設計、系統資料庫以及系統流程。

3.1 系統架構

本論文提出之兒童英語繪本系統架構有 Client 端和 Server 端，Client 端是以行動裝置作為人機介面，可再細分為教師端和學生端，Server 端則為本系統之演算法、GAN 模型、網頁伺服器以及資料庫伺服器。教師端是由教師輸入自行設計的英語故事內容的詞句，輸入完成後會送出一個網頁的請求，Server 端的網路伺服器則會接收請求讀取輸入文字，接著利用自然語言處理工具包(Natural Language Tool Kit, NLTK) [59]對輸入文字進行前處理，處理完後的文字會分別回傳至教師端及進入 Text Encoder，教師可選擇每句句子的重點單字，再回傳至 Server 端，Server 端會利用爬蟲去爬取有道字典中的內容，而進入 Text Encoder 的文句會進行訓練以得到 Word features 和 Sentence features，接著輸入至 GAN 模型透過訓練好的 GAN 模型生成與文句對應的漫畫風格圖像，最後英語教學的文字內容、重點單字以及漫畫圖像會自製成英語繪本並回傳至教師端讓教師確認上傳，若選擇上傳後資料會儲存至資料庫中，學生端則能透過人機介面選擇欲閱讀的繪本，經由網頁請求，網頁會向資料庫請求資料並接收資料庫的回應，最後回傳至介面，學生則能透過繪本學習英語，如圖 3-1 所示。

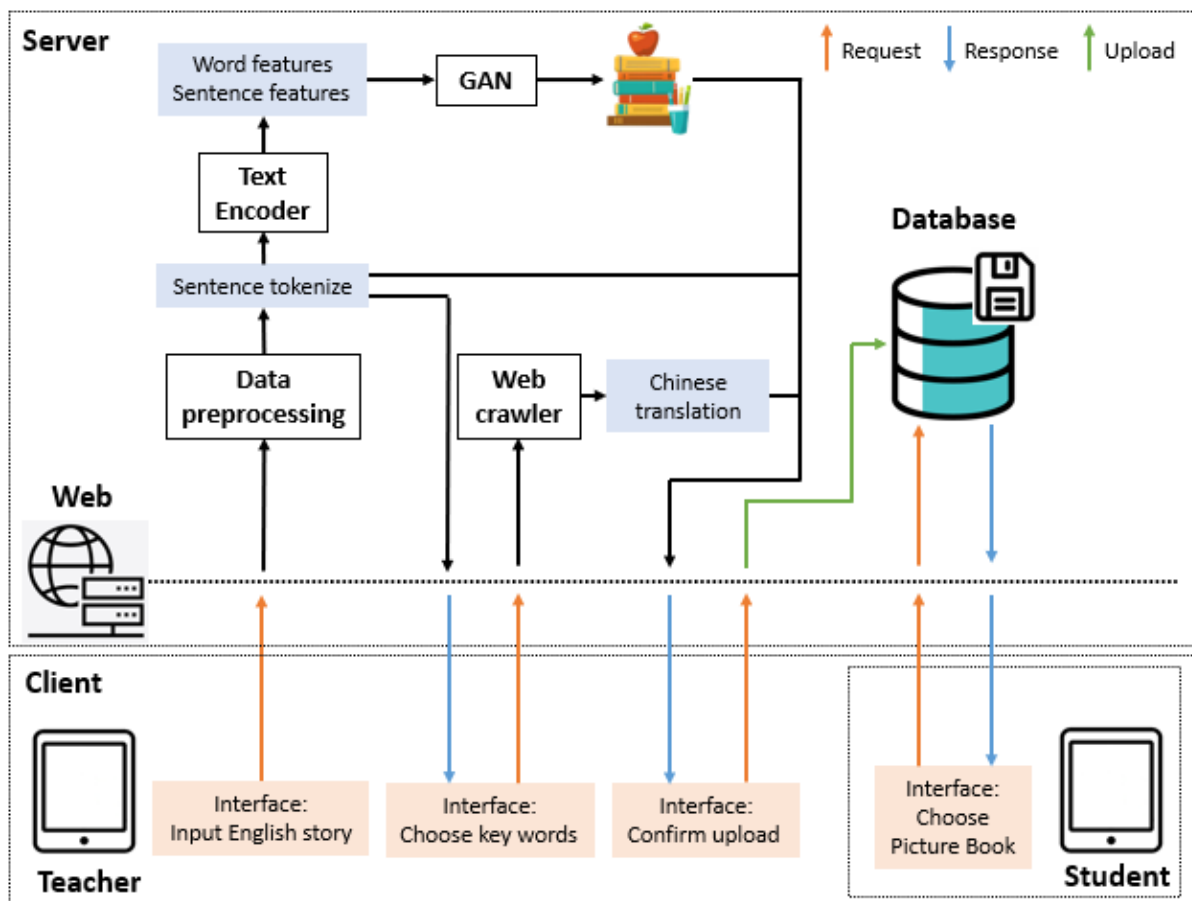


圖 3-1 系統架構圖

3.2 生成對抗網路模型

本論文使用的 GAN 模型是以 AttnGAN[54]做改良，AttnGAN 主要有兩個模組，分別為注意力生成網路(Attentional Generative Network)和 DAMSM，藉由注意力生成網路，AttnGAN 可以透過關注輸入句子中的相關單字來生成圖像中不同子區域的細節，以三階段的方式優化生成的圖像，而 DAMSM 訓練了兩個神經網路，分別為 Text Encoder 和 Image Encoder，以計算圖像和文字的相似度，如圖 3-2 所示，接著將分別介紹模型中的文字處理、注意力生成網路、DAMSM 以及繪本圖像生成。

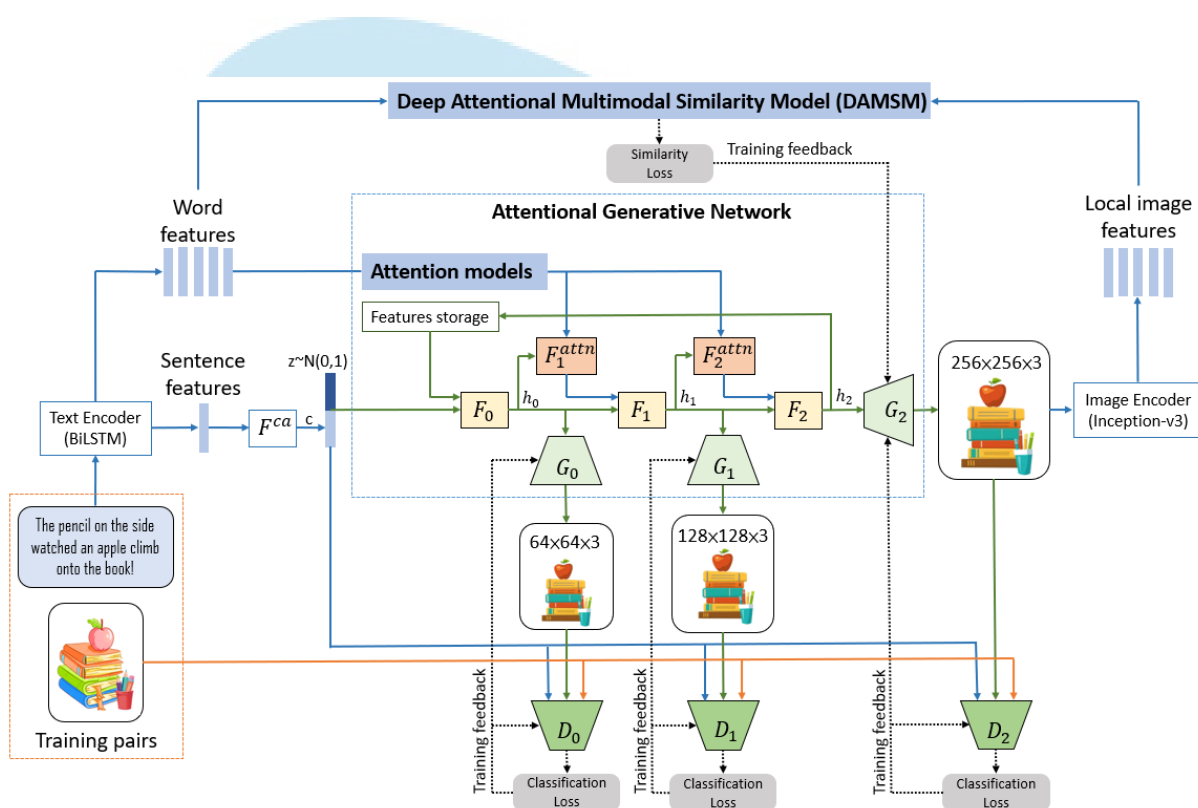


圖 3-2 AttnGAN 架構圖

3.2.1 文字處理

基於繪本為一個連貫故事的特性，教師端的輸入為整篇完整的故事內容，由於兒童在學習時需有段落性，需以一段句子搭配一張圖像，並透過多張圖像連貫成一篇故事，因此須將整篇故事分解為單一句子，接著利用句子中的主詞、動詞及受詞之間的關係去生成對應的圖像。首先，使用

NLTK 對輸入的英語故事進行分句，以將整篇故事分解成句子，再將分解後的句子輸入至 Text Encoder，Text Encoder 是使用雙向長短期記憶(Bi-directional Long Short-Term Memory, BiLSTM)[60]，由於 LSTM 只能依照之前的狀態來預測下一個輸出，但當前時刻的輸出不僅和之前的狀態有關，還可能與未來的狀態有關，因此使用 BiLSTM 不但可以根據前文來判斷，還可以考慮後面的內容，達到基於前後文的判斷。Text Encoder 會逐字讀取輸入句子將單字編碼為隱藏狀態，每個單字會對應兩個隱藏狀態，分別在 Backward Layer 和 Forward Layer 各一個，接著連接兩個隱藏狀態來表示一個單字特徵(Word features)，而 BiLSTM 的最後隱藏狀態會連接為句子特徵(Sentence features)，如圖 3-3 所示。

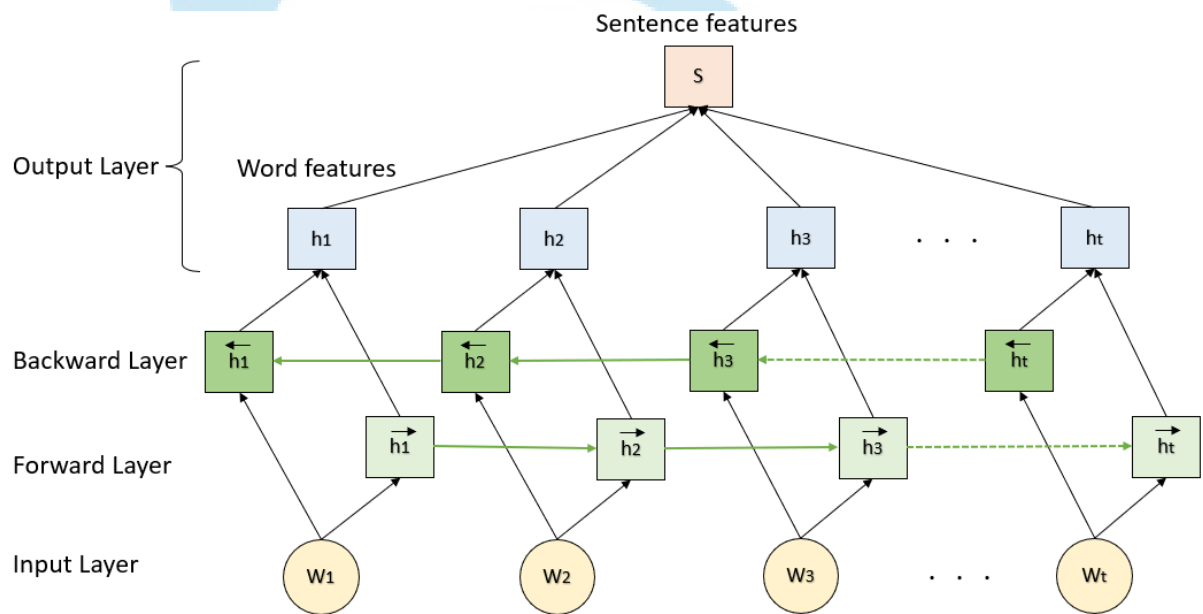


圖 3-3 BiLSTM 示意圖

3.2.2 注意力生成網路

表 3-1 Attentional Generative Network 符號表[54]

G	生成網路
D	鑑別網路
h	圖像特徵
\hat{x}	生成圖像
x	真實圖像
z	隨機噪聲
\bar{e}	句子特徵向量
e	單字特徵向量
m	生成網路的數量
F^{ca}	條件增強網路
F	神經網路
F^{attn}	注意力模型
N	子區域的數量
T	單字的數量
c_j	第 j 個子區域的 word-context
$\beta_{j,i}$	第 i 個單字對於第 j 個子區域的重要程度
p_G	生成圖像分佈
p_{data}	真實圖像分佈

注意力生成網路共使用了三個生成網路，生成網路以圖像特徵之隱藏狀態(hidden states)作為輸入，並生成不同解析度的圖像如公式(3.1)，每個生成網路都有對應的鑑別網路用來計算分類損失，並透過反向傳播訓練生成網路與鑑別網路。在第一階段會將從標準正態分佈中採樣的噪聲 z 與經過條件增強(Conditioning Augmentation)[52]的句子特徵 $F^{ca}(\bar{e})$ 結合如公式(3.2)，其中 F_0 表示的是神經網路，經過多次的上採樣產生圖像特徵 h_0 和生

成 64×64 的低解析度圖像，第二階段的輸入為第一階段的圖像特徵 h_{i-1} 和注意力模型 $F_i^{attn}(e, h_{i-1})$ 如公式(3.3)，其中 m 為生成網路的數量，透過注意力機制來生成更詳細的圖像細節並生成 128×128 的圖像，第三階段的輸入同樣為上一階段的圖像特徵和注意力模型，最終生成 256×256 的圖像，如圖 3-2 所示。

$$\hat{x}_i = G_i(h_i) \quad (3.1)$$

$$h_0 = F_0(z, F^{ca}(\bar{e})) \quad (3.2)$$

$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, \dots, m-1 \quad (3.3)$$

注意力模型有兩個輸入，分別為單字特徵 e 和上一階段的圖像特徵 h 如公式(3.4)，其中 N 表示子區域的數量，注意力模型是用來計算圖像中每個子區域的 word-context，用句子中所有的單字向量來進行表示，相關的單字有較大的權重，而不相關的單字則權重較小，每個子區域計算單字向量加權和的結果即為 word-context。

$$F^{attn}(e, h) = (c_0, c_1, \dots, c_{N-1}) \quad (3.4)$$

對於第 j 個子區域，它的 word-context 可以表示為 c_j ，其計算公式為 (3.5)，在計算 word-context 前會先將單字特徵 e 轉換為與圖像特徵同樣維度的 e' ，其中 $\beta_{j,i}$ 為第 i 個單字對於圖像特徵中第 j 個子區域的重要程度，透過第 j 個子區域向量與第 i 個單字向量的內積，除以第 j 個子區域向量與句子中所有單字向量內積之和，可以得到第 i 個單字的權重係數，而第 j 個子區域上的 word-context 為所有單字向量的加權和。最後將上一階段的圖像特徵和相應的 word-context 結合，即可在下一階段生成更詳細的圖像。

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \text{ where } \beta_{j,i} = \frac{\exp(h_j^T e'_i)}{\sum_{k=0}^{T-1} \exp(h_j^T e'_k)} \quad (3.5)$$

注意力生成網路的最終目標函數定義為公式(3.6)：

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i} \quad (3.6)$$

其中 λ 為超參數，用於平衡方程式的兩項，第一項為聯合條件和無條件分佈的 GAN 損失，在第 i 個階段，生成網路 G_i 有一個對應的鑑別網路 D_i ， G_i 的對抗損失定義為公式(3.7)：

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log D_i(\hat{x}_i)]}_{\text{unconditional loss}} - \underbrace{\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log D_i(\hat{x}_i, \bar{e})]}_{\text{conditional loss}} \quad (3.7)$$

其中前項為無條件損失，用來確認圖像的真假，而後項為條件損失，用來確認圖像和句子是否匹配。 D_i 會和 G_i 交替訓練， D_i 透過最小化定義的損失如公式(3.8)將輸入分類為真假類別。

$$\begin{aligned} \mathcal{L}_{D_i} = & \underbrace{-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i)] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i))]}_{\text{unconditional loss}} \\ & + \underbrace{-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i, \bar{e})] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(1 - D_i(\hat{x}_i, \bar{e}))]}_{\text{conditional loss}} \end{aligned} \quad (3.8)$$

其中前項為無條件損失，後項為條件損失， x_i 為來自第 i 個解析度的真實圖像， p_{data_i} 為第 i 個解析度的真實圖像分佈， \hat{x}_i 為來自相同解析度的生成圖像， p_{G_i} 為相同解析度的生成圖像分佈，。每個 D_i 都是獨立的，分別用來鑑別不同解析度的圖像，因此他們可以並行訓練。公式(3.6)的第二項 \mathcal{L}_{DAMSM} 是用來計算圖像和文字的相似度損失，將在 3.2.3 節詳細說明。

3.2.3 深度注意力多模態相似性模型

表 3-2 DAMSM 符號表[54]

f	局部圖像特徵向量
\bar{f}	全局圖像特徵向量
v	添加感知器的局部圖像特徵向量
\bar{v}	添加感知器的全局圖像特徵向量
$s_{i,j}$	第 i 個單字和第 j 個子區域的相似性
c_i	第 i 個單字的 region-context
Q	整個圖像
D	句子描述

DAMSM 是用來計算圖像和文字的相似度，它訓練了兩個神經網路，分別為 Text Encoder 和 Image Encoder，並從整個句子和句子中的每個單字來計算相似度損失。由於 DAMSM 是在真實資料集上預訓練的，在輸入生成的圖像和對應的句子描述時，它會迫使注意力生成網路生成更逼真且與對應句子相關的圖像，如圖 3-4 所示。

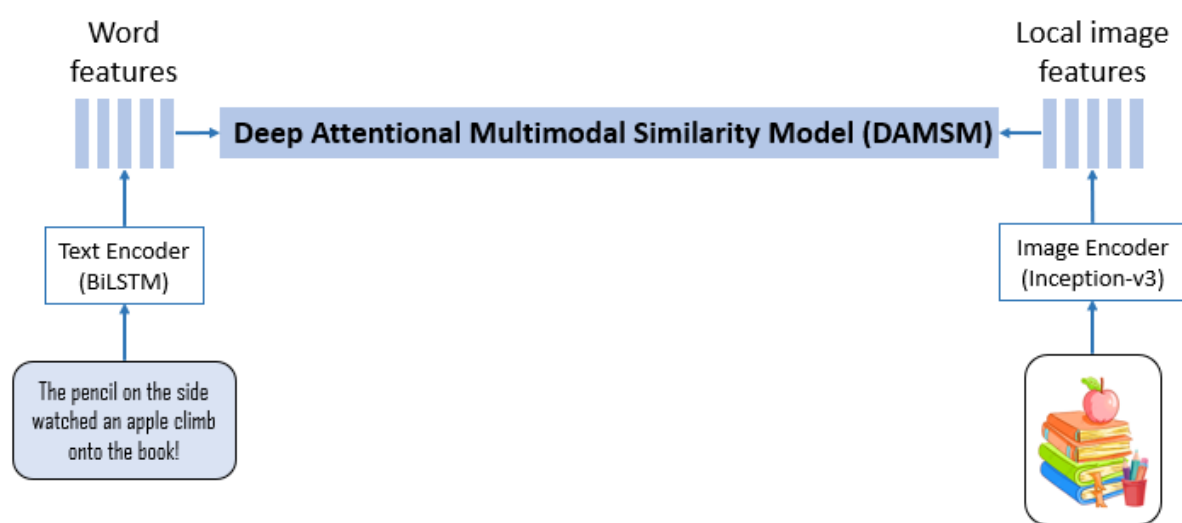


圖 3-4 DAMSM 示意圖

Text Encoder 使用 BiLSTM，訓練過程與文字處理的部份相同，會輸出單字特徵與句子特徵，所有單字的特徵矩陣由 $e \in \mathbb{R}^{D \times T}$ 表示，其中 D 為單字向量的維度， T 為單字的數量， e_i 代表第 i 個單字的特徵向量，句子向量則由 $\bar{e} \in \mathbb{R}^D$ 表示。

Image Encoder 使用 CNN，是在 ImageNet[61] 上預訓練的 Inception-v3 模型[62]建構的，CNN 的中間層會學習圖像中不同子區域的局部特徵，而後面的層會學習圖像的全局特徵。從 Inception-v3 的 Mixed_6e 層中提取局部特徵向量 $f \in \mathbb{R}^{768 \times 289}$ ，它是從 $768 \times 17 \times 17$ 重塑的，其中 768 代表局部特徵向量的維度，289 代表圖像中子區域的數量，而全局特徵向量 $\bar{f} \in \mathbb{R}^{2048}$ 是從 Inception-v3 的最後一個平均池化層(Average Pooling Layer)中提取的，如圖 3-5 所示，為了計算圖像與文字的相似度，文字和圖像的維度須相同，因此透過添加感知器層將圖像特徵與文字特徵保持一致如公式 (3.9)，其中 $v \in \mathbb{R}^{D \times 289}$ ，其中 D 為特徵空間的維度， v_i 代表圖像中第 i 個子區域的視覺特徵向量，而 $\bar{v} \in \mathbb{R}^D$ 代表整個圖像的全局圖像特徵向量。

$$v = W f, \quad \bar{v} = \bar{W} \bar{f} \quad (3.9)$$

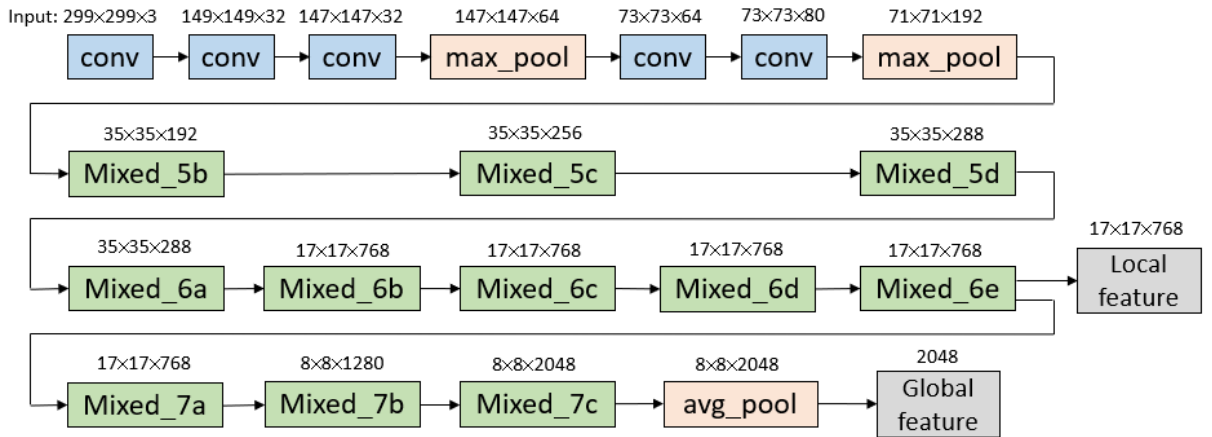


圖 3-5 Inception-v3 架構圖

首先透過公式(3.10)計算句子中所有單字和圖像中子區域的相似性矩陣，其中 $s \in \mathbb{R}^{T \times 289}$ ，其中 T 為單字的數量， $s_{i,j}$ 代表句子中第 i 個單字和圖像中第 j 個子區域的相似性，接著對相似性矩陣進行歸一化如公式(3.11)。

$$s = e^T v \quad (3.10)$$

$$\bar{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})} \quad (3.11)$$

接下來利用注意力模型，針對每一個單字計算所有子區域視覺向量的加權和以得到 region-context，對於第 i 個單字，它的 region-context 可以表示為 c_i ，其計算公式如(3.12)，其中 γ_1 表示對於相關的子區域的關注程度，相似性越大則關注程度越高。

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \text{ where } \alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_1 \bar{s}_{i,k})} \quad (3.12)$$

最後，使用 c_i 和 e_i 之間的餘弦相似性來定義第 i 個單字與圖像之間的相似性，即 $R(c_i, e_i) = (c_i^T e_i) / (\|c_i\| \|e_i\|)$ ，整個圖像 Q 和句子描述 D 之間的匹配分數定義為公式(3.13)，其中 γ_2 是為了突出最相關的 word-to-region-context，它用來調節重要程度。

$$R(Q, D) = \log \left(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i)) \right)^{\frac{1}{\gamma_2}} \quad (3.13)$$

DAMSM 損失的監督標籤是整個圖像與整句句子是否匹配，對於一批圖像和句子 $\{(Q_i, D_i)\}_{i=1}^M$ ，用圖像 Q_i 去匹配句子 D_i 的事後機率 (posterior probability) 為公式(3.14)，其中 γ_3 是透過實驗確定的平滑因數，在這一批句子中，只有 D_i 與圖像 Q_i 匹配，而將其他 $M - 1$ 的句子視為不匹配的描述，

將損失函數定義為圖像與其對應的句子匹配的負對數事後概率為公式(3.15)，其中 w 代表單字。

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^M \exp(\gamma_3 R(Q_i, D_j))} \quad (3.14)$$

$$\mathcal{L}_1^w = - \sum_{i=1}^M \log P(D_i|Q_i) \quad (3.15)$$

同樣地，在以句子 D_i 去匹配圖像 Q_i 的情況下，損失函數定義為公式(3.16)。

$$\mathcal{L}_2^w = - \sum_{i=1}^M \log P(Q_i|D_i) \quad (3.16)$$

如果用 $R(Q, D) = (\bar{v}^T \bar{e}) / (\|\bar{v}\| \|\bar{e}\|)$ 重新定義公式(3.13)，並將其代入公式(3.14)(3.15)(3.16)，則可以利用句子向量 \bar{e} 和全域圖像向量 \bar{v} 獲得損失函數 \mathcal{L}_1^s 和 \mathcal{L}_2^s ，其中 s 代表句子。

最後 DAMSM 損失可以定義為公式(3.17)。

$$\mathcal{L}_{DAMSM} = \mathcal{L}_1^w + \mathcal{L}_2^w + \mathcal{L}_1^s + \mathcal{L}_2^s \quad (3.17)$$

3.2.4 繪本圖像生成

本研究之兒童英語繪本系統的輸入為一整篇英語故事，經過文字處理將整篇故事段句後，每句句子會依序輸入至 GAN 模型中生成圖像，由於繪本是以圖像為主體並結合文字共同描述一個完整的故事，而繪本故事中通常會有幾個角色，這些角色在每頁故事的圖像中應保有相同的外觀以符合故事的描述，因此圖像之間須具有連貫性，我們提出了在圖像生成後提取圖像特徵的方法，使得生成圖像之間具有連貫性，當第一句句子輸入至

GAN 模型後，會利用 Inception-v3 提取第三階段生成的最終圖像之局部圖像特徵，並儲存於特徵暫存器內，當下一句句子輸入至 GAN 模型時，會將暫存器內的圖像特徵與此句的文字特徵結合，以生成保有上一句局部圖像特徵的連貫性圖像，如圖 3-6 所示。

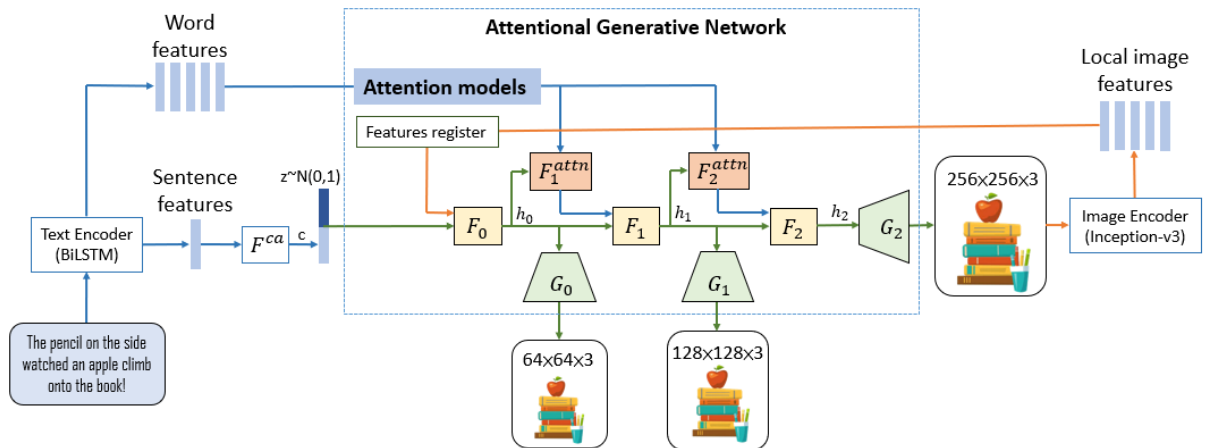


圖 3-6 繪本圖像生成過程

3.3 介面設計

使用者介面是以網站作為開發，使用的框架為 Flask[63]，它是使用 Python 語言編寫的 Web 應用框架，是基於 Jinja2 模板引擎和 Werkzeug WSGI 套件，它提供了許多架設網站需求的基本工具，包括網頁模板 (Templates)、路由(Routes)、權限(Authorization)等。為了導入行動學習的模式，因此利用 Android Studio 設計成 app 即可於行動裝置上使用。

3.3.1 教師端

在首頁的頁面中，由教師輸入繪本名稱和英語故事內容的詞句，按下確認後，輸入的內容會傳送至 Sever 端利用 NLTK 進行分句，以將整篇故事分成一段一段的句子，接著再回傳至介面中，教師可選取每句句子中的重點單字，按下生成圖像後，Sever 端會進行 GAN 模型的運算，如圖 3-7 所示。

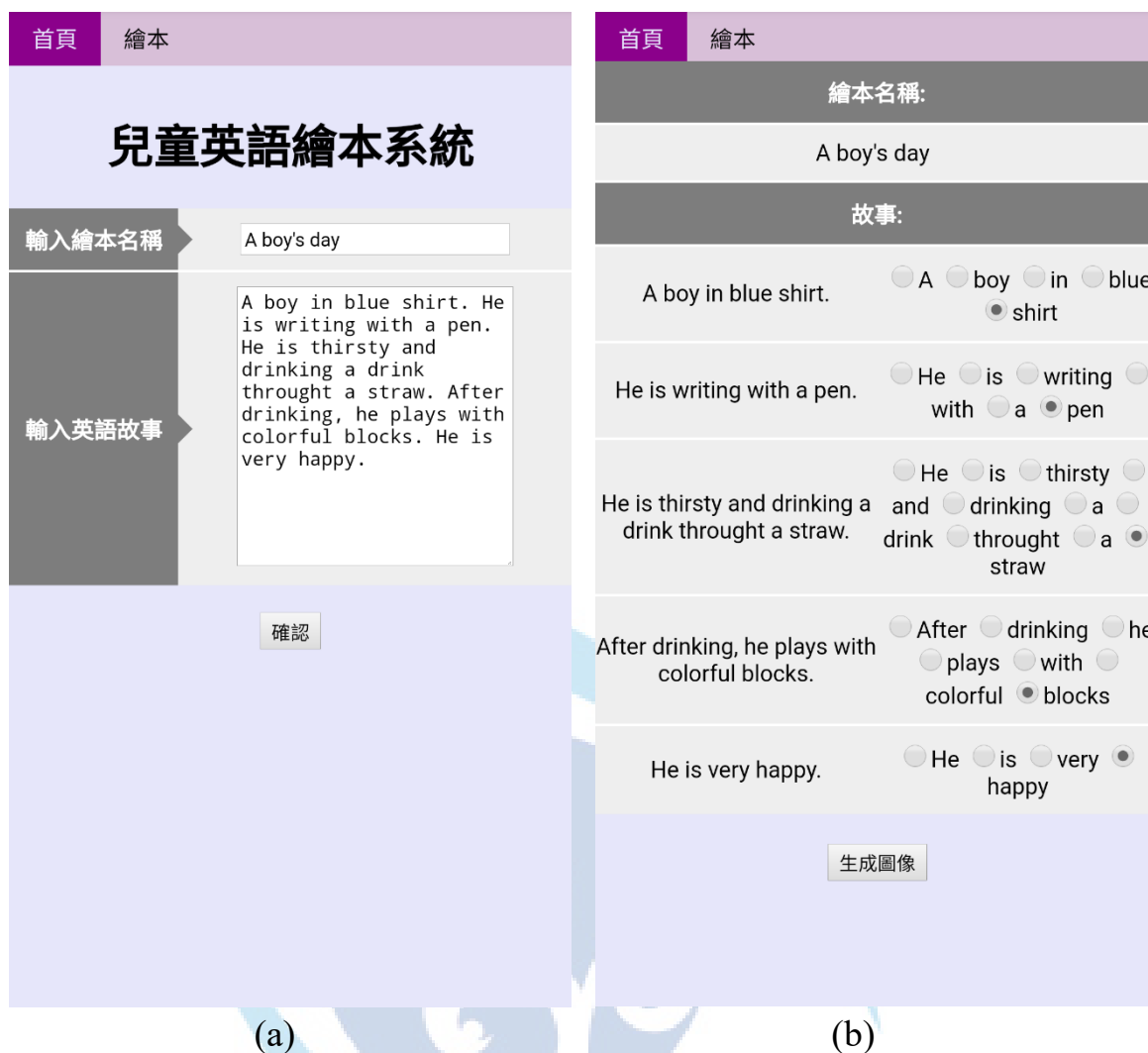


圖 3-7 教師端介面 (a)輸入英語故事 (b)選取重點單字

GAN 模型運算完後，每段句子會生成一張對應的漫畫圖像，選取的單字則會顯示單字的中文翻譯，它是利用爬蟲去爬取有道字典中的內容，接著有兩個按鍵可以選取，若生成的結果不滿意可選取重新生成，頁面會導向首頁，可重新輸入內容，若是確定上傳則可選取上傳繪本，繪本的故事內容、生成圖像以及重點單字內容會儲存至資料庫，接著頁面會導向繪本，即可查看所有繪本的內容，如圖 3-8 所示。



圖 3-8 教師端介面 (a)生成的圖像 (b)儲存於資料庫的繪本

3.3.2 學生端

學生可查看所有儲存於資料庫的繪本，在所有繪本的頁面中，每本繪本皆會顯示繪本的首張圖像和繪本名稱，選取欲閱讀的繪本，則會顯示繪本的故事內容，點選向右箭頭則可進行翻頁，如圖 3-9 所示。



(a)



(b)

圖 3-9 學生端介面 (a)儲存於資料庫的繪本 (b)選取的繪本內容

3.4 系統資料庫

本論文使用 SQLAlchemy 來操作資料庫，SQLAlchemy 是以 Python 語言撰寫的物件關係對映(Object Relational Mapping, ORM)框架，ORM 框架是在關聯式資料庫 (Database) 和物件 (Application) 之間做映射，可直接使用 Python 語法操作資料庫，無需再編寫複雜的結構化查詢語言 (Structured Query Language, SQL)語法處理資料，ORM 會自動將 Python 代碼轉換成對應的 SQL 語法，再對資料庫進行操作，開發者只需要操作物件的屬性與方法即可達到編寫 SQL 語法的效果，如圖 3-10 所示。

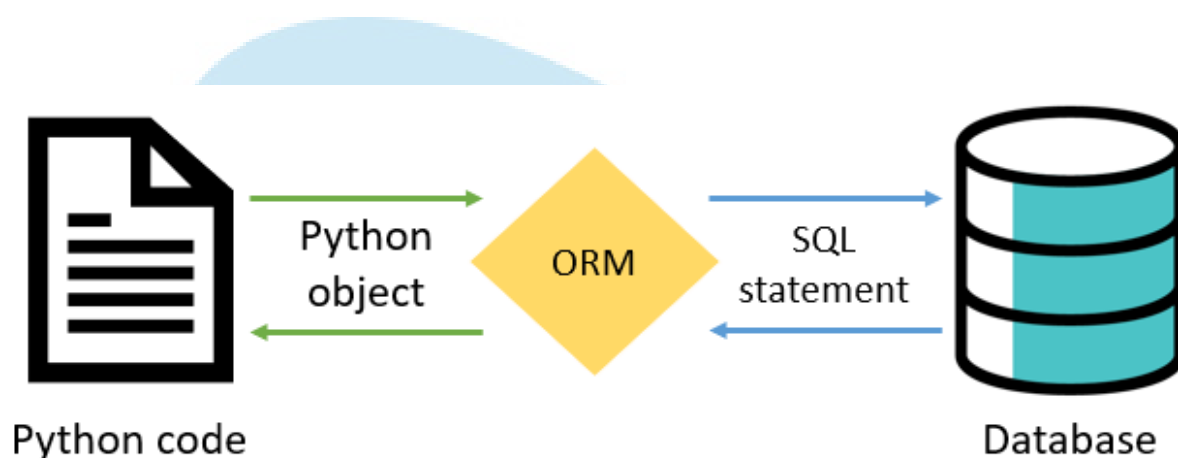


圖 3-10 SQLAlchemy ORM 框架

SQLAlchemy 可以連線大多數常見的資料庫，例如 MySQL、SQLite、PostgreSQL 等，本論文連接的資料庫為 MySQL，MySQL 為關聯式資料庫管理系統(Relational Database Management System, RDMS)，也就是資料庫是由多個資料表(Table)組成的，其中資料表包含紀錄(Record)、欄位(Field)及資料(Data)，Data 表示資料儲存在資料庫的形式，而一筆一筆橫向的資料代表 Record，直向代表 Field 是一筆資料的不同屬性。每筆資料都有互相關聯的特性，並且可以透過關聯各個資料表去連結多個資料表之間的關係。

本論文將每一篇上傳的繪本設計成一個資料表，資料表內的每一筆 Record 代表一頁的繪本故事，Field 的屬性有 id、text、word、image、example，如圖 3-11 所示，其中 text 代表斷句後的英語句子、word 代表選

取的重點單字、image 代表生成的圖像以及 example 代表重點單字的中文翻譯。

```
mysql> show columns from book;
```

Field	Type	Null	Key	Default	Extra
id	int	NO	PRI	NULL	auto_increment
text	varchar(100)	YES		NULL	
word	varchar(100)	YES		NULL	
image	longblob	YES		NULL	
example	varchar(500)	YES		NULL	

圖 3-11 資料表的屬性

3.5 系統流程

本論文提出之兒童英語繪本系統流程可分為教師端和學生端，教師端可輸入繪本名稱及英語故事內容，接著選擇每段句子的重點單字，選擇完後即可生成對應於每句句子的圖像及重點單字的中文翻譯，若生成的圖像不滿意可重新輸入繪本故事，若確認上傳即可上傳繪本至資料庫。學生端可選取儲存於資料庫內的所有繪本，接著即可透過繪本學習英語，如圖 3-12 所示。

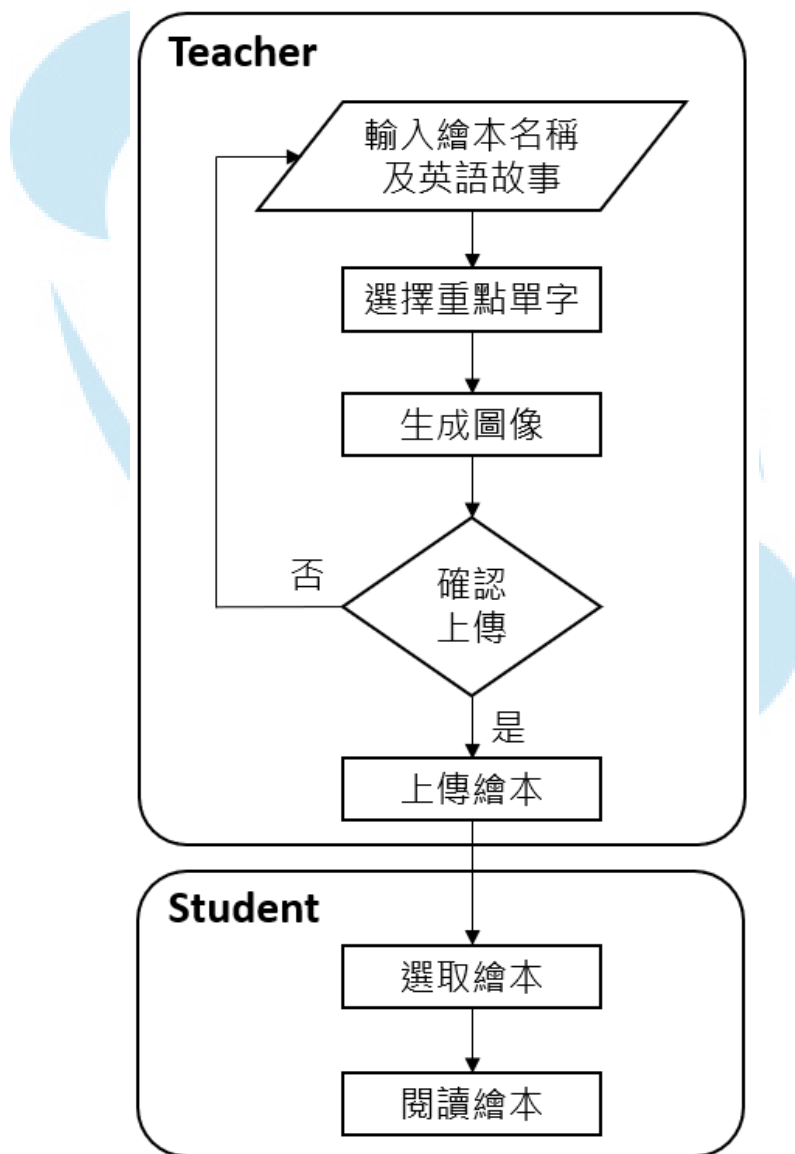


圖 3-12 系統流程圖

第四章 實驗

4.1 實驗資料

由於目前未有適合本論文的公開資料集，因此本論文訓練 GAN 模型所使用的資料集為自製資料集，基於生成漫畫圖像為目標，我們使用 Irasutoya[64]網站中的圖像，Irasutoya 是一個提供大量插圖素材的日本網站且每張圖像的繪畫風格相近，與本論文欲生成的圖像相符，我們從 Irasutoya 網站中挑選 20,000 張插圖，如圖 4-1 所示，再自行為每張插圖標上對應圖像的句子描述，每張圖像共有五句句子描述，如圖 4-2 所示。



圖 4-1 Irasutoya 網站中的圖像



Woman in pink dress is holding a water gun on hand
She is sprinkling water with the hose in her hand
The water gun on her hand squirted water
Woman in pink dress is holding a water pipe on hand
She was spraying water with the hose in her hand

圖 4-2 實驗圖像與對應的五句句子描述



4.2 實驗環境

本實驗的硬體設備為 Intel i7-11700KF 處理器和 16GB 記憶體之桌上型主機，軟體的設置為 Window10 以及 Python 語言，並使用 Anaconda 內的 Spyder 作為深度學習的開發環境，詳細的實驗環境設置如表 4-1 所示。

表 4-1 實驗環境設置表

項目	內容
處理器	Intel i7-11700KF
顯示卡	NVIDIA GeForce RTX 3060
記憶體	16GB DDR4
顯示卡記憶體	20GB
作業系統	Window10
開發環境	Spyder
程式語言	Python 3.6.13
深度學習框架	Pytorch 1.7.1

4.3 評估指標

為了評估生成圖像的品質，在定量評估(Quantitative evaluation)中使用 Fréchet Inception Distance(FID)[65]和 R-precision[54]進行評估。

FID 是用來計算真實圖像與生成圖像之間特徵分佈的距離，FID 能夠代表生成圖像的多樣性和品質，首先利用 Inception 網絡來提取特徵，其特徵來自 Inception 最後一個平均池化層(Average Pooling Layer)，接著使用平均值和協方差矩陣來計算兩個分佈之間的距離，FID 計算公式如(4.1)：

$$\text{FID}(x, g) = \|\mu_x - \mu_g\|^2 + \text{Tr}\left(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}\right) \quad (4.1)$$

其中 x 表示真實圖像、 g 表示生成圖像、 μ_x 表示真實圖像特徵的平均值、 μ_g 表示生成圖像特徵的平均值、 Σ_x 表示真實圖像特徵的協方差、 Σ_g 表示生成圖像特徵的協方差、 Tr 為一個線性代數，表示矩陣對角線上元素之和。FID 值越低則表示圖像多樣性和品質越好，[65]於實驗中評估了在圖像中添加不同擾動等級的高斯雜訊對 FID 值的影響，如圖 4-3 所示，FID 值隨著圖像失真程度提高而增加，代表圖像品質越來越低，由此能證明 FID 可以用來評估圖像的品質。[66][67][68][69]皆利用 FID 作為評估 GAN 生成卡通風格圖像或繪畫風格圖像的指標，可以從[66]的實驗結果發現，FID 在評估繪畫風格圖像時整體分數會比真實圖像來得高，根據不同的訓練資料集，FID 在評估真實圖像時，分數通常在 100 以下，而當 FID 用來評估卡通圖像或繪畫風格圖像時，分數則會在 100 以上，意指只要考慮所屬圖像是哪一種類別並選擇合適區間，FID 仍是有效的評估指標。

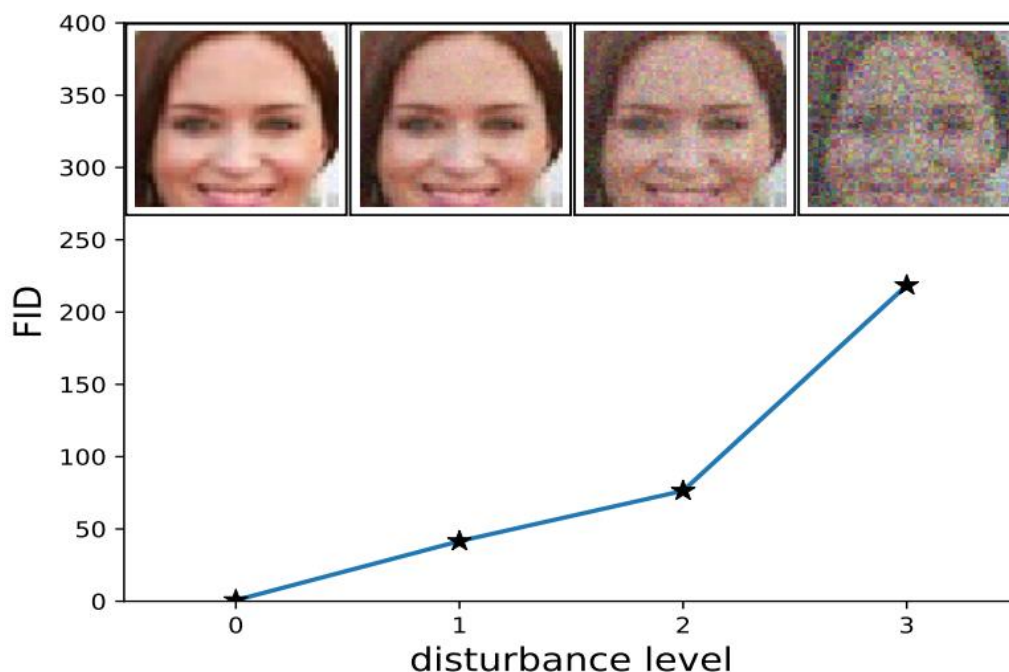


圖 4-3 添加不同擾動等級的高斯雜訊對 FID 值的影響

R-precision 是用來評估生成圖像是否匹配於給定的文字描述，它是使用生成圖像來檢索對應的文字描述，首先，使用預訓練 DAMSM 中的 Text Encoder 和 Image Encoder 來提取生成圖像的全局特徵和給定的文字描述，接著計算圖像向量和文字向量之間的餘弦相似度，最後按降序對每個圖像的候選文字描述進行排序，如果排名前 R 的檢索描述中的 r 個結果是相關的，則 R-precision 可以定義為 r/R ，R-precision 值越高則生成圖像越匹配於給定的文字描述。在本實驗中我們計算 $R = 1$ 的 R-precision，從文字描述中生成 20,000 張圖像，每個圖像的候選文字描述包刮 R 個基本事實(ground truth)和 $100 - R$ 個隨機的不匹配描述。

4.4 實驗

本論文提出之基於 GAN 的兒童英語繪本系統除了能夠從英語故事生成繪本外，還要能生成逼真且符合輸入故事文字的連貫性圖像，為了使生成的圖像達到最佳效果，我們系統性地調整模型的超參數，最後將使用調整後的最佳超參數展示實驗結果，實驗中的參數設置及相關設置如表 4-2 所示，三個階段的生成網路模型結構分別如表 4-3、表 4-4 以及表 4-5 所示，三個階段的鑑別網路模型結構分別如表 4-6、表 4-7 以及表 4-8 所示。

表 4-2 GAN 參數設置表

世代	140
批次大小	25
生成網路學習率	0.0002
鑑別網路學習率	0.0002
生成網路優化器	Adam
鑑別網路優化器	Adam

表 4-3 第一階段生成網路模型結構

Layer	Dimension
Input	sentence embedding=256 noise=100
Fully Connected Layer	16,384
Batch Normalization	16,384
Reshape	4×4×512
Transposed Convolution	8×8×256
Batch Normalization	8×8×256
Transposed Convolution	16×16×128
Batch Normalization	16×16×128

Transposed Convolution	$32 \times 32 \times 64$
Batch Normalization	$32 \times 32 \times 64$
Transposed Convolution	$64 \times 64 \times 32$
Batch Normalization	$64 \times 64 \times 32$
Convolution	$64 \times 64 \times 3$
Tanh	$64 \times 64 \times 3$

表 4-4 第二階段生成網路模型結構

Layer	Dimension
Input	$64 \times 64 \times 32$
Residual	$64 \times 64 \times 64$
Transposed Convolution	$128 \times 128 \times 32$
Batch Normalization	$128 \times 128 \times 32$
Convolution	$128 \times 128 \times 3$
Tanh	$128 \times 128 \times 3$

表 4-5 第三階段生成網路模型結構

Layer	Dimension
Input	$128 \times 128 \times 32$
Residual	$128 \times 128 \times 64$
Transposed Convolution	$256 \times 256 \times 32$
Batch Normalization	$256 \times 256 \times 32$
Convolution	$256 \times 256 \times 3$
Tanh	$256 \times 256 \times 3$

表 4-6 第一階段鑑別網路模型結構

Layer	Dimension
Input	64×64×3
Convolution	32×32×64
LeakyReLU	32×32×64
Convolution	16×16×128
Batch Normalization	16×16×128
LeakyReLU	16×16×128
Convolution	8×8×256
Batch Normalization	8×8×256
LeakyReLU	8×8×256
Convolution	4×4×512
Batch Normalization	4×4×512
LeakyReLU	4×4×512
Convolution	1
Sigmoid	1

表 4-7 第二階段鑑別網路模型結構

Layer	Dimension
Input	128×128×3
Convolution	64×64×64
LeakyReLU	64×64×64
Convolution	32×32×128
Batch Normalization	32×32×128
LeakyReLU	32×32×128
Convolution	16×16×256
Batch Normalization	16×16×256

LeakyReLU	16×16×256
Convolution	8×8×512
Batch Normalization	8×8×512
LeakyReLU	8×8×512
Convolution	4×4×1024
Batch Normalization	4×4×1024
LeakyReLU	4×4×1024
Convolution	4×4×512
Batch Normalization	4×4×512
LeakyReLU	4×4×512
Convolution	1
Sigmoid	1

表 4-8 第三階段鑑別網路模型結構

Layer	Dimension
Input	256×256×3
Convolution	128×128×64
LeakyReLU	128×128×64
Convolution	64×64×128
Batch Normalization	64×64×128
LeakyReLU	64×64×128
Convolution	32×32×256
Batch Normalization	32×32×256
LeakyReLU	32×32×256
Convolution	16×16×512
Batch Normalization	16×16×512
LeakyReLU	16×16×512
Convolution	8×8×1024

Batch Normalization	8×8×1024
LeakyReLU	8×8×1024
Convolution	4×4×2048
Batch Normalization	4×4×2048
LeakyReLU	4×4×2048
Convolution	4×4×1024
Batch Normalization	4×4×1024
LeakyReLU	4×4×1024
Convolution	4×4×512
Batch Normalization	4×4×512
LeakyReLU	4×4×512
Convolution	1
Sigmoid	1

4.4.1 模型超參數調整

本實驗調整的超參數為公式(3.6)中的 λ ，它是用於平衡 GAN 損失以及計算圖像和文字相似度的 DAMSM 損失，在一開始的訓練時，我們將 λ 設置為 50，並觀察第三階段的生成網路及鑑別網路的損失值在 160 個 Epoch 的變化，可以由圖 4-4 看到鑑別網路的損失值幾乎持續趨近於 0，這代表鑑別網路判別真假圖像的能力過強，以至於無法與生成網路達到對抗訓練，並導致訓練結果不好，而生成網路的損失值也無法收斂，因此，我們將 λ 依序調整為 20、5、1、0.1，並觀察它們的結果，由圖 4-5 可以看到，將 λ 調整為 20 時，鑑別網路的損失值依然持續趨近於 0，生成網路的損失值也還無法收斂，圖 4-6 為將 λ 調整為 5 的結果，可以看到鑑別網路的損失值開始有變化，生成網路也有逐漸收斂的趨勢，而圖 4-7 為將 λ 調整為 1 的結果，可以看到鑑別網路的損失值隨著生成網路的損失值升高而降低，代表兩個網路有達到有效的對抗訓練，而生成網路也逐漸收斂，若將 λ 調整為 0.1，如圖 4-8 所示，結果並無較好，因此最終設置 λ 為 1。

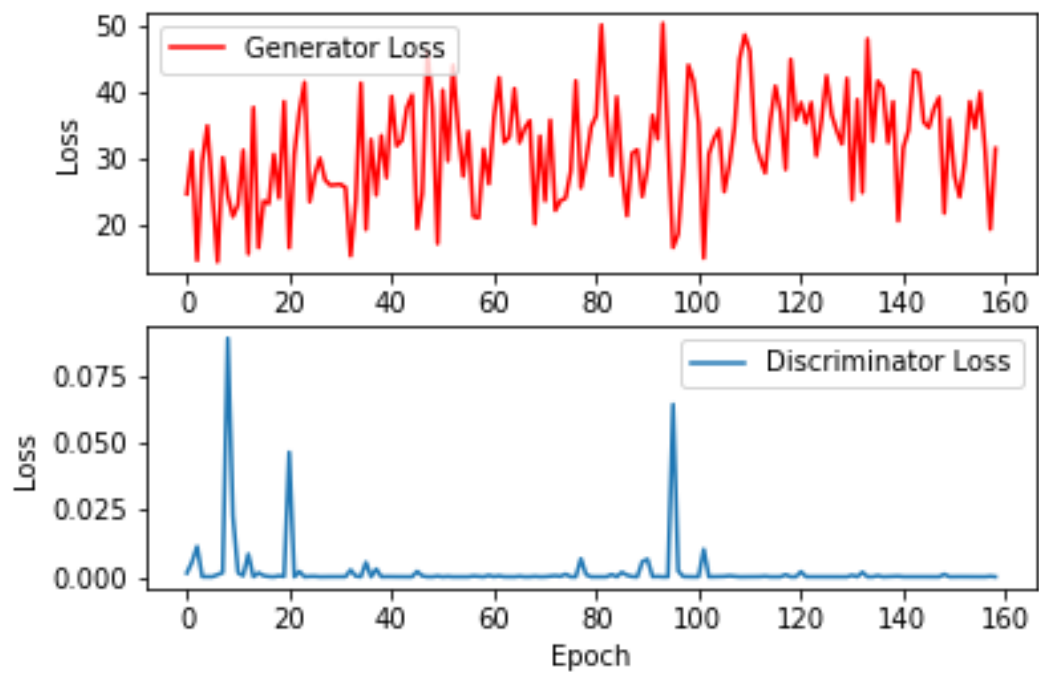


圖 4-4 $\lambda=50$ 損失值變化

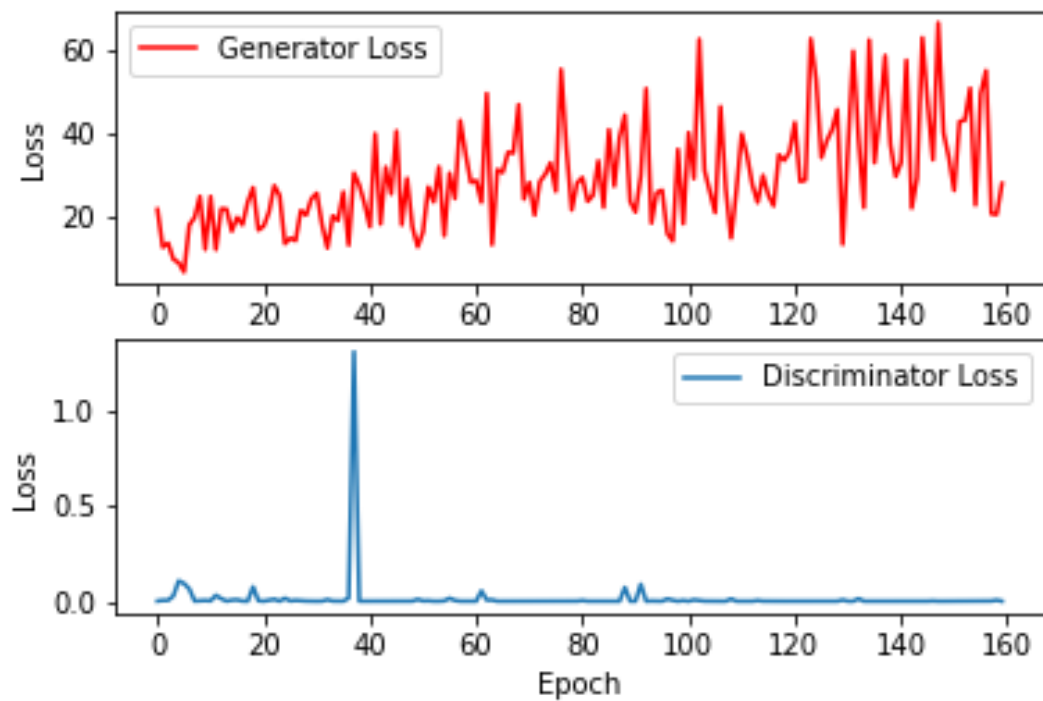


圖 4-5 $\lambda=20$ 損失值變化

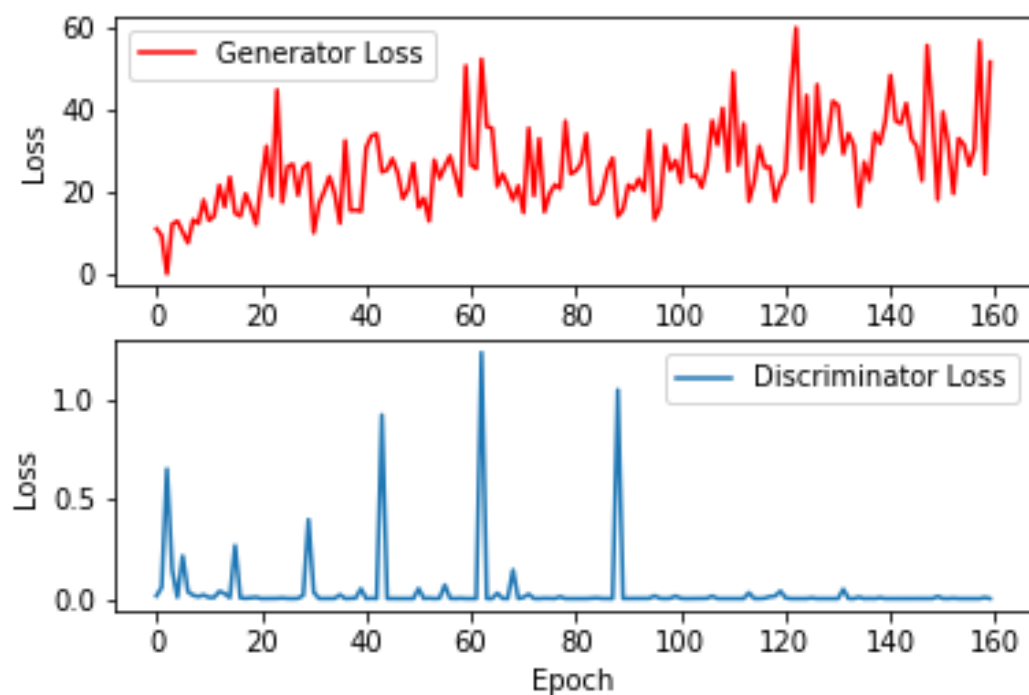


圖 4-6 $\lambda=5$ 損失值變化

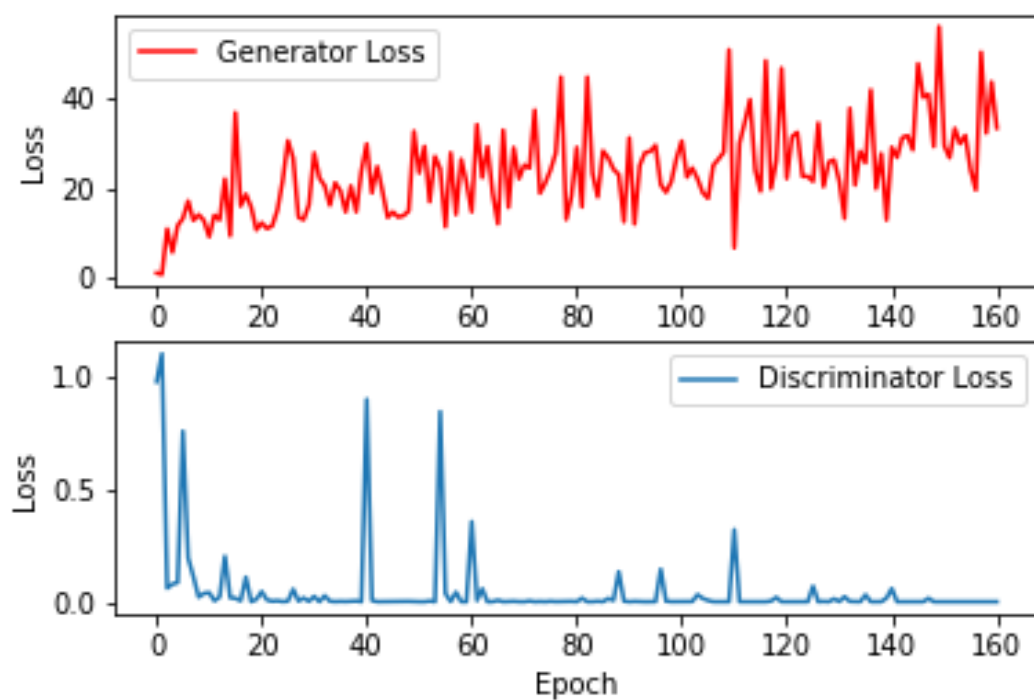


圖 4-7 $\lambda=1$ 損失值變化

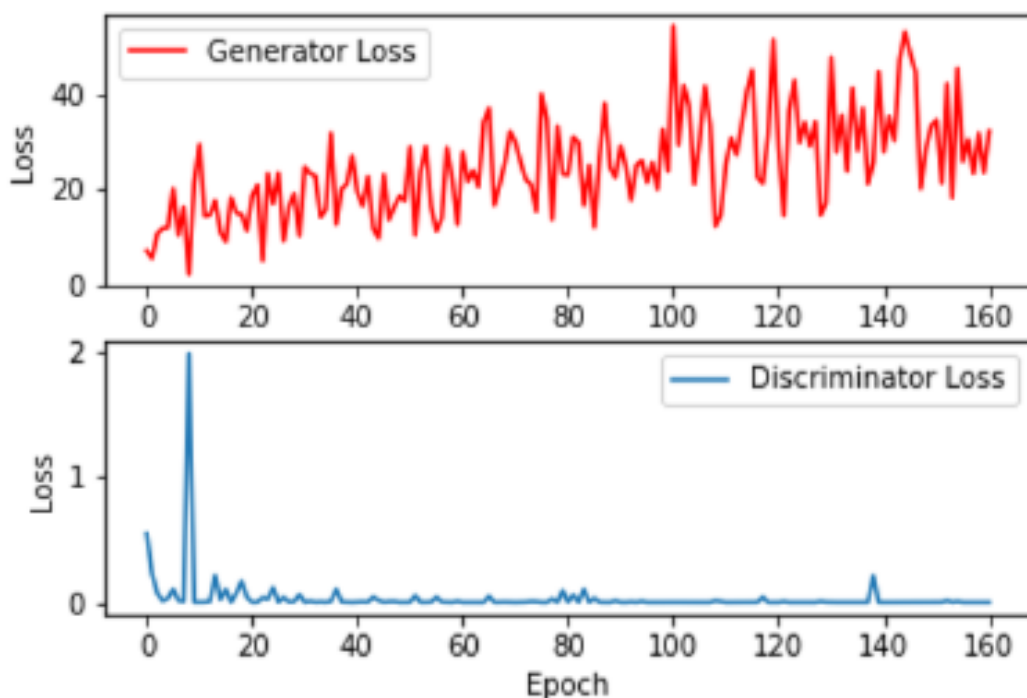


圖 4-8 $\lambda=0.1$ 損失值變化

在訓練模型時，Epoch 的大小也會影響訓練的結果，若執行的 Epoch 次數太少，會導致無法完整地學習訓練資料的特徵，若執行的 Epoch 次數太多，則會導致訓練資料過度學習，而造成過擬合現象，為了求得最佳的 Epoch，我們針對 Epoch 為 0 到 160 遞增去評估 FID 值和 R-precision，以得到最佳的 Epoch。圖 4-9 顯示了不同 Epoch 大小的 FID 值比較，越低的 FID 值代表圖像品質越好，可以看到在剛開始訓練時 FID 值非常的高，到了 Epoch 為 20 時，FID 值開始有下降的趨勢，當 Epoch 為 140 時，獲得最佳的 FID 值，而當 Epoch 為 160 時，FID 值開始上升且訓練時間相對增長。圖 4-10 為不同 Epoch 大小的 R-precision 值比較，R-precision 值可用來表示圖像與文字的相似性，R-precision 值越高代表相似性越高。我們從文字描述中生成 20,000 張圖像，每張生成圖像的候選文字描述包刮 1 個基本事實 (ground truth) 和 99 個隨機的不匹配描述，接著利用這 100 句句子的排序去計算 R-precision 值。可以看到當 Epoch 為 20 時，R-precision 值開始有明顯的提升，當 Epoch 為 140 時，獲得最佳的 R-precision 值，綜合上述的結果我們設定 Epoch 為 140。

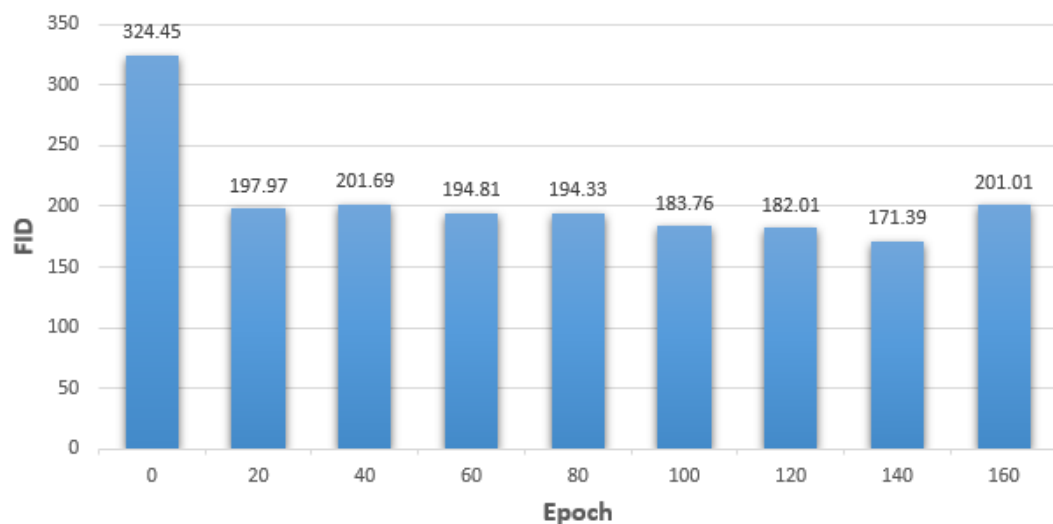


圖 4-9 不同 Epoch 之 FID 值

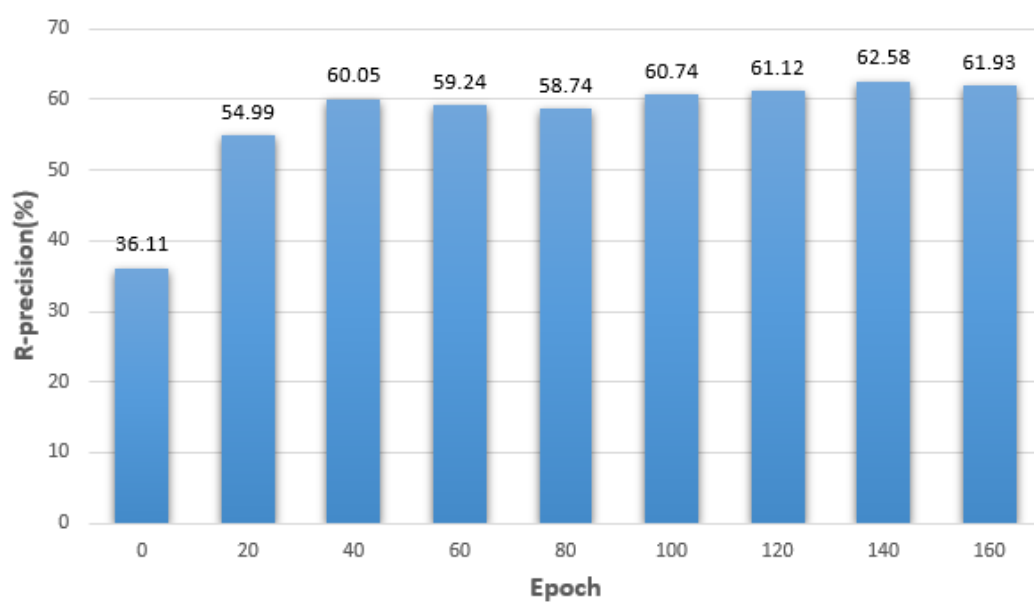


圖 4-10 不同 Epoch 之 R-precision 值

4.4.2 實驗結果

本實驗分別以自編故事以及英語網站中的故事展示實驗結果，首先透過自編故事驗證本系統的準確性，為了測試本系統能夠在不同的故事中產生具連貫性圖像，我們還使用了英語網站中的故事做實驗，英語網站為三毛英語季[70]，它是一個英語學習網站，包含英語聽力、口說和閱讀等教

材，在英語閱讀中提供了許多英語故事，我們從中挑選適合兒童閱讀的故事，選擇的英語故事名稱為 *The bathing boy* 和 *I don't want to walk home*。實驗共展示了三篇不同複雜度故事的生成結果，分別為一篇自編故事和兩篇英語網站故事，故事的複雜度為根據故事中的角色數量定義。

自編故事的故事名稱為 *A boy's day*，輸入的英語故事內容如圖 4-11(a) 所示，本故事的角色數量為 1 人，圖 4-11(b) 為故事斷句後的結果，可選擇每句句子的重點單字，生成繪本的結果如圖 4-12、圖 4-13 及圖 4-14 所示，可以看到生成的圖像皆符合輸入的文句，基本上人物的表情和動作都可以清楚地顯示，且能生成具連貫性的圖像，例如圖 4-13(a) 為描述男孩用吸管在喝飲料，可以看到圖像中包含了吸管與杯子，而圖 4-13(b) 為描述男孩在玩積木，可以看到圖像中包含了積木，且每頁中的角色固定為同個男孩，但在一些細節的部份，例如上衣細節及手部細節生成的結果較不理想，這是由於模型無法抓取到細部的圖像特徵。

英語網站中的故事 *The bathing boy* 之英語故事內容如圖 4-15(a) 所示，本故事的角色數量為 2 人，分別為一個男孩和一個男人，圖 4-15(b) 為故事斷句後的結果，可選擇每句句子的重點單字，生成繪本的結果如圖 4-16、圖 4-17 及圖 4-18 所示，與自編故事相比，*The bathing boy* 的文句較為複雜，文句中包含較多抽象的描述，例如：“He say: It is dangerous to bath in the river!”，由於模型在訓練時較少訓練到抽象的文字，導致生成的結果較不清晰。圖 4-16(a) 為描述男孩在河邊洗澡，可以看到圖像生成了河邊的背景，但男孩的生成結果較差，這是由於洗澡這個動作在訓練時較少出現，因此在人物的動作和表情上較難正確地抓取特徵，但生成的圖像依然能表達文字的內容，圖 4-17(a) 為描述第二個角色，可以看到圖像生成了不同於前頁的新角色，雖然在抽象的文句中生成較差，但依然能夠生成連貫性的圖像。

英語網站中的故事 *I don't want to walk home* 之英語故事內容如圖 4-19(a) 所示，本故事的角色數量為 3 人，分別為老男人、老女人與警察，圖 4-19(b) 為故事斷句後的結果，可選擇每句句子的重點單字，生成繪本的結果如圖 4-20、圖 4-21 及圖 4-22 所示，與上述兩篇故事相比，*I don't want to walk home* 的故事中包含較多的角色，當句子描述中包含一個角色時，例如：

“Tom is a very old man.”，生成的結果較佳，而當句子描述中包含多個角色時，例如：“The policeman tells Tom’s wife: The old man couldn’t find his way in the street, he asked me to take him in the car.”，由於句子過於複雜，句子中包含多個角色，以至於角色之間的特徵重疊，導致模型無法抓取到正確的特徵，因此較難生成清晰的圖像，但生成的圖像依然符合文字描述且具連貫性。

首頁

繪本

兒童英語繪本系統

輸入繪本名稱

A boy's day

輸入英語故事

A boy in blue shirt. He is writing with a pen. He is thirsty and drinking a drink through a straw. After drinking, he plays with colorful blocks. He is very happy.

確認

首頁

繪本

繪本名稱:

A boy's day

故事:

A boy in blue shirt.

☐ A
☐ boy
☐ in
☐ blue
☒ shirt

He is writing with a pen.

☐ He
☐ is
☐ writing
☐ with
☐ a
☒ pen

He is thirsty and drinking a drink through a straw.

☐ He
☐ is
☐ thirsty
☐ and
☐ drinking
☐ a
☐ drink
☐ through
☐ a
☒ straw

After drinking, he plays with colorful blocks.

☐ After
☐ drinking
☐ he
☐ plays
☐ with
☐ colorful
☒ blocks

He is very happy.

☐ He
☐ is
☐ very
☒ happy

生成圖像

(a)

(b)

圖 4-11 *A boy's day* (a)輸入英語故事 (b)選取重點單字

首頁 繪本

A boy in blue shirt.



shirt • n. 襯衫，恤衫；球衣；<英>運動隊隊員

1 2 3 ... 5 >

(a)

首頁 繪本

He is writing with a pen.



pen • n. 筆，鋼筆；（圈養動物用的）圍欄，圈；（計算機上使用的）電子筆；<北美>監獄；<英，非正式>處罰，（尤指足球的）點球；墨水；雌天鵝；（動）（槍鳥賊的）羽狀殼；（西印度羣島）農莊，大農場；（潛艇等戰艦的）隱蔽塢，掩藏塢；國際筆會（國際詩人，劇作家，編輯，散文家和小說家協會）（PEN）；祕魯新索爾（PEN） vt. 寫，撰寫；

(b)

圖 4-12 *A boy's day* (a)第一頁 (b)第二頁

首頁 繪本

He is thirsty and drinking a drink through a straw.



straw

- n. (乾燥的) 麥稈，稻草；(喝飲料用的) 吸管；稻草色，淡黃色；一文不值的東西adj. 稻草的；無價值的【名】 (Straw) 斯特勞 (人名)

< 1 2 3 4 5 >

(a)

首頁 繪本

After drinking, he plays with colorful blocks.



blocks

- n. [建]街區；積木；樓羣 (block 的複數)；黑金石；雙頭木魚v. 阻礙；封鎖；使成塊 (block 的三單形式)

< 1 ... 3 4 5 >

(b)

圖 4-13 *A boy's day* (a)第三頁 (b)第四頁

首頁

繪本

He is very happy.



happy

- adj. 快樂的；幸福的，使人高興的；滿意的；樂意的；幸運的；合適的comb. <非正式> 濫用.....的【名】 (Happy) (英、瑞典、喀) 哈皮 (人名)

<

1

...

4

5

圖 4-14 *A boy's day* 第五頁

首頁

繪本

輸入繪本名稱

The bathing boy

輸入英語故事

One day,A boy is bathing in a river. He doesn't swim well and will be drown, he calls out loud for help. A man is just passing by but he doesn't help the boy. He say: It is dangerous to bath in the river! The boy say: Please help me out and scold me afterward, counsel without help is unless.

確認

(a)

首頁

繪本

繪本名稱:

The bathing boy

故事:

One day,A boy is bathing in a river.

One

day

A

boy

is

bathing

in

a

river

He doesn't swim

He doesn't swim well and will be drown, he calls out loud for help.

He

doesn't

swim

well

and

will

be

drown

he

calls

out

loud

for

help

A man is just passing by but he doesn't help the boy.

A

man

is

just

passing

by

but

he

doesn't

help

the

boy

He say: It is dangerous to bath in the river!

He

say

It

is

dangerous

to

bath

in

the

river

The boy say: Please help me out and scold me afterward, counsel

The

boy

say

Please

help

me

out

and

scold

me

afterward,

counsel

without

help

is


unless.

(b)

圖 4-15 *The bathing boy* (a)輸入英語故事 (b)選取重點單字

首頁 繪本

One day, A boy is bathing in a river.



river • n. 河，江n. (River) 人名；(英) 裏弗

1 2 3 ... 5 >

(a)

首頁 繪本

He doesn't swim well and will be drown, he calls out loud for help.



swim • v. 游泳；橫渡；遊（某一泳姿）；（魚、禽等）遊過，遊動；浸，泡；漂浮；使漂浮，使渡過；（物體）彷彿旋轉；眩暈n. 游泳；（河流中）適合釣魚的靜止深水處，深潭；漂浮；眩暈adj. 游泳時穿戴的

< 1 2 3 4 5 >

(b)

圖 4-16 *The bathing boy* (a)第一頁 (b)第二頁

首頁 繪本

A man is just passing by but he doesn't help the boy.

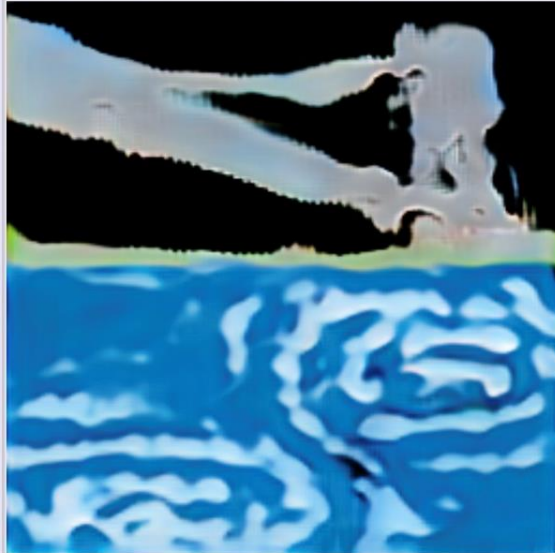


help	<ul style="list-style-type: none"> v. 幫助，援助；改善狀況，對.....有益；給（自己或某人）食物或飲料；擅自拿走，竊取；攙扶，帶領n. 幫助，協助；有助益的東西（如忠告、錢等）；有用；救助；有幫助的人（或事物）；傭人；（計算機程序提供使用說明的）求助程序int. 救命（呼救用語）【名】（Help）（芬）海爾普（人名）
------	--

(a)

首頁 繪本

He say: It is dangerous to bath in the river!



dangerous	<ul style="list-style-type: none"> adj. 危險的，有威脅的
-----------	---

< 1 ... 3 4 5 >

(b)

圖 4-17 *The bathing boy* (a)第三頁 (b)第四頁

The boy say: Please help me out and scold me afterward, Counsel without help is useless.



Counsel • n. 忠告，建議；辯護律師；<古>諮詢，磋商
 v. 建議，勸告（做.....）；為.....提供諮詢，給.....提供建議；商討，提出忠告

圖 4-18 *The bathing boy* 第五頁

首頁

繪本

繪本名稱:

I don't want to walk home

故事:

Tom is a very old man.

☐ Tom
☐ is
☐ a
☐ very
☒ old
☐ man

After dinner, he likes walking in the street and he goes to bed at seven o'clock.

☐ After
☐ dinner
☐ he
☐ likes
☒ walking
☐ in
☐ the
☐ street
☐ and
☐ he
☐ goes
☐ to
☐ bed
☐ at
☐ seven
☐ o'clock

But tonight, a car stopped at his house and policeman helps him get out.

☐ But
☐ tonight
☐ a
☐ car
☐ stopped
☐ at
☐ his
☐ house
☐ and
☒ policeman
☐ helps
☐ him
☐ get
☐ out

The policeman tells Tom's wife,The old man couldn't find his way in the street, He asked me to take him in the car.

☐ The
☐ policeman
☐ tells
☐ Tom's
☐ wifeThe
☐ old
☐ man
☐ couldn't
☐ find
☐ his
☐ way
☐ in
☐ the
☒ street
☐ He
☐ asked
☐ me
☐ to
☐ take
☐ him
☐ in
☐ the

輸入繪本名稱

I don't want to walk home

輸入英語故事

Tom is a very old man.
After dinner, he likes walking in the street and he goes to bed at seven o'clock.
But tonight, a car stopped at his house and policeman helps him get out.
The policeman tells Tom's wife,The old man couldn't find his way in the street, He asked me to take him in the car.
After the policeman leaves there, his wife asks: Tom, you go to the street every night but tonight you can't find the way, what's the matter?
The old man smiles like a child and says: I couldn't find my way?I didn't want to walk home.

確認

(a)

(b)

圖 4-19 *I don't want to walk home* (a)輸入英語故事 (b)選取重點單字

首頁 繪本

Tom is a very old man.



old

- adj. (人)歲的, (事物) 存在.....久的; 年老的, 年紀大的; 衰老的; 古老的, 歷史悠久的; 陳舊的; 過去的, 從前的; 原來 (屬於自己的) 的; 結識久的; <非正式> (表示親暱) 老.....; (語言形式) 古的, 早期的; 老派的, 守舊的; 老一套的, 經歷多次的
- n. 老年人; 某個年齡段的人; 古時【名】
(Old) (英) 奧爾德 (人名)

1 2 3 ... 6 >

(a)

首頁 繪本

After dinner, he likes walking in the street and he goes to bed at seven o'clock.



walking

- n. 步行; 散步 v. 步行 (walk 的 ing 形式)
adj. 步行的

< 1 2 3 4 ... 6 >

(b)

圖 4-20 *I don't want to walk home* (a)第一頁 (b)第二頁

首頁 繪本

But tonight, a car stopped at his house and policeman helps him get out.



policeman

- n. 警察，警員；[分化] 澱帚（橡皮頭玻璃攪棒）

< 1 2 3 4 5 6 >

(a)

首頁 繪本

The policeman tells Tom's wife, The old man couldn't find his way in the street, He asked me to take him in the car.



street

- n. 街道adj. 街道的n. (Street) 人名；(英、葡) 斯特里特；(德) 施特雷特

< 1 ... 3 4 5 6 >

(b)

圖 4-21 *I don't want to walk home* (a)第三頁 (b)第四頁



(a)



(b)

圖 4-22 *I don't want to walk home* (a)第五頁 (b)第六頁

由上述三篇故事的實驗結果可以得知，本系統在生成簡單故事繪本的結果較好，隨著故事複雜度和角色數量的增加，生成圖像的品質則會隨之降低，雖然在較複雜的故事中生成的圖像結果較差，但本系統依然可根據輸入的故事生成對應的圖像且圖像之間具有連貫性。

第五章 結論與未來展望

隨著英語能力重要性的提高，兒童學習英語的年齡逐漸下降，並隨著科技的進步，學習環境從傳統學習發展出了數位學習，學習英語的方法變得越來越豐富，目前已有許多英語學習系統，但大部份的系統只適合具有一定能力的學習者，由於兒童與成人的認知能力不同，因此應設計適合他們的學習方式，我們發現兒童對於圖像會比文字更感興趣，若是將這些文字轉為圖像，則能讓兒童產生興趣，然而聘請漫畫家替英語繪本製作內容並不是件容易的事，因此本論文欲透過 GAN 解決這個問題。

本論文提出一個基於 GAN 的兒童英語繪本系統，透過 GAN 以及資訊技術來輔助兒童的英語學習，本系統分為 Client 端和 Server 端，Client 端可再分為教師端及學生端，教師端能夠輸入自行設計的英語故事，Server 端則會對輸入的英語詞句進行處理，再透過 GAN 生成對應的圖像，並自製成英語繪本儲存於資料庫中，學生端則能透過行動裝置選取欲閱讀的繪本來學習英語。

本實驗使用自製的資料集，從 Irasutoya 網站中挑選 20,000 張圖像，再自行為每張圖像標上對應的五句句子描述。為了求得訓練模型的最佳參數，我們系統性地調整模型的超參數，例如平衡目標函數中兩個損失的超參數以及 Epoch 大小，提高了鑑別網路與生成網路之間的訓練結果，並在實驗結果表明，本系統能夠根據輸入的故事，在簡單的故事及抽象的故事中生成對應的連貫性圖像。

目前由於神經網路訓練的限制，因此生成圖像僅侷限於 256×256 ，未來希望能透過提升神經網路的效能以生成更高解析度的圖像，使得此系統更符合實際的應用。

未來將嘗試生成不同漫畫風格的圖像，由於每個兒童都有各自喜歡的特定漫畫風格，若是能增加選擇圖像風格的選項以提高繪本風格的多樣性，則能提升本系統的學習效果，但由於目前較少有此類型的資料集，因此在圖像標記上需花費較多的時間。

參考文獻

- [1] P. S. Rao, "The importance of english in the modern era," *Asia. Journ. of Multidimensi. Resear. (AJMR)*, vol. 8, no. 1, p. 7, 2019, doi: 10.5958/2278-4853.2019.00001.6.
- [2] M. Virvou, D. Maras, and V. Tsiriga, "Student Modelling in an Intelligent Tutoring System for the Passive Voice of English Language," *Journal of Educational Technology & Society*, vol. 3, no. 4, pp. 139–150, 2000.
- [3] M. I. Alhabbash, A. O. Mahdi, and S. S. A. Naser, "An Intelligent Tutoring System for Teaching Grammar English Tenses," *European Academic Research*, vol. 4, no. 9, pp. 1–15, 2016.
- [4] M. J. Abu Ghali, A. Abu Ayyad, S. S. Abu-Naser, and M. Abu Laban, "An Intelligent Tutoring System for Teaching English Grammar," 2018, Accessed: Apr. 10, 2022. [Online]. Available: <http://dspace.alazhar.edu.ps/xmlui/handle/123456789/289>
- [5] V. A. Nguyen, V. C. Pham, and S. D. Ho, "A Context - Aware Mobile Learning Adaptive System for Supporting Foreigner Learning English," in *2010 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, Nov. 2010, pp. 1–6. doi: 10.1109/RIVF.2010.5632316.
- [6] V. Bradac and B. Walek, "A comprehensive adaptive system for e-learning of foreign languages," *Expert Systems with Applications*, vol. 90, pp. 414–426, Dec. 2017, doi: 10.1016/j.eswa.2017.08.019.
- [7] L. Zilio and C. Fairon, "Adaptive System for Language Learning," in *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, Jul. 2017, pp. 47–49. doi: 10.1109/ICALT.2017.46.
- [8] Y. Chen, D. Zhou, Y. Wang, and J. Yu, "Application of Augmented Reality for Early Childhood English Teaching," in *2017 International Symposium on Educational Technology (ISET)*, Jun. 2017, pp. 111–115. doi: 10.1109/ISET.2017.34.
- [9] C.-M. Chen and C.-J. Chung, "Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle," *Computers & Education*, vol. 51, no. 2, pp. 624–645, Sep. 2008, doi: 10.1016/j.compedu.2007.06.011.
- [10] S. Sfenrianto, Y. B. Hartarto, H. Akbar, M. Mukhtar, E. Efriadi, and M. Wahyudi, "An Adaptive Learning System based on Knowledge Level for English Learning," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 12, Art. no. 12, Dec. 2018, doi:

10.3991/ijet.v13i12.8004.

- [11] S.-C. Ho, S.-W. Hsieh, P.-C. Sun, and C.-M. Chen, "To Activate English Learning: Listen and Speak in Real Life Context with an AR Featured U-Learning System," *Journal of Educational Technology & Society*, vol. 20, no. 2, pp. 176–187, 2017.
- [12] D. Lewis, *Reading Contemporary Picturebooks: Picturing text*. London: Routledge, 2001. doi: 10.4324/9780203354889.
- [13] P. K. Kuhl, "Early Language Learning and Literacy: Neuroscience Implications for Education," *Mind, Brain, and Education*, vol. 5, no. 3, pp. 128–142, 2011, doi: 10.1111/j.1751-228X.2011.01121.x.
- [14] C. T. Mart, "Encouraging Young Learners to Learn English through Stories," *English Language Teaching*, vol. 5, no. 5, pp. 101–106, May 2012.
- [15] L. Baker, "How Many Words Is a Picture Worth? Integrating Visual Literacy in Language Learning with Photographs," *English Teaching Forum*, vol. 53, no. 4, pp. 2–13, 2015.
- [16] Z. K. Takacs and A. G. Bus, "How pictures in picture storybooks support young children's story comprehension: An eye-tracking experiment," *Journal of Experimental Child Psychology*, vol. 174, pp. 1–12, Oct. 2018, doi: 10.1016/j.jecp.2018.04.013.
- [17] A. Y. Alqahtani and A. A. Rajkhan, "E-Learning Critical Success Factors during the COVID-19 Pandemic: A Comprehensive Analysis of E-Learning Managerial Perspectives," *Education Sciences*, vol. 10, no. 9, Art. no. 9, Sep. 2020, doi: 10.3390/educsci10090216.
- [18] A. S. Yucel, "E-Learning Approach in Teacher Training," *Turkish Online Journal of Distance Education*, vol. 7, no. 4, Art. no. 4, Dec. 2006.
- [19] A. A. Hamid, "e-Learning: Is it the 'e' or the learning that matters?," *The Internet and Higher Education*, vol. 4, no. 3, pp. 311–316, Jan. 2001, doi: 10.1016/S1096-7516(01)00072-0.
- [20] B. J. Zimmerman, "Self-Regulated Learning and Academic Achievement: An Overview," *Educational Psychologist*, vol. 25, no. 1, pp. 3–17, Jan. 1990, doi: 10.1207/s15326985ep2501_2.
- [21] T. Takahashi, K. Asahi, H. Suzuki, M. Kawasumi, and Y. Kameya, "A Cloud Education Environment to Support Self-Learning at Home - Analysis of Self-Learning Styles from Log Data," in *2015 IIAI 4th International Congress on Advanced Applied Informatics*, Jul. 2015, pp. 437–440. doi: 10.1109/IIAI-AAI.2015.213.
- [22] M. Chassignol, A. Khoroshavin, A. Klimova, and A. Bilyatdinova, "Artificial Intelligence trends in education: a narrative overview," *Procedia Computer*

- Science*, vol. 136, pp. 16–24, Jan. 2018, doi: 10.1016/j.procs.2018.08.233.
- [23] M. Liu, E. McKelroy, S. B. Corliss, and J. Carrigan, “Investigating the effect of an adaptive learning intervention on students’ learning,” *Education Tech Research Dev*, vol. 65, no. 6, pp. 1605–1625, Dec. 2017, doi: 10.1007/s11423-017-9542-1.
- [24] S. Ennouamani and Z. Mahani, “An overview of adaptive e-learning systems,” in *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, Dec. 2017, pp. 342–347. doi: 10.1109/IN^{TEL}CIS.2017.8260060.
- [25] Y. Wang and K. Okamura, “Automatic Generation of E-Learning Contents Based on Deep Learning and Natural Language Processing Techniques,” in *Advances in Internet, Data and Web Technologies*, Cham, 2020, pp. 311–322. doi: 10.1007/978-3-030-39746-3_33.
- [26] K. R. Chowdhary, “Natural Language Processing,” in *Fundamentals of Artificial Intelligence*, K. R. Chowdhary, Ed. New Delhi: Springer India, 2020, pp. 603–649. doi: 10.1007/978-81-322-3972-7_19.
- [27] A. Sun, Y.-J. Li, Y.-M. Huang, and Q. Li, “Using Facial Expression to Detect Emotion in E-learning System: A Deep Learning Method,” in *Emerging Technologies for Education*, Cham, 2017, pp. 446–455. doi: 10.1007/978-3-319-71084-6_52.
- [28] A. B. Firdausiah Mansur, N. Yusof, and A. H. Basori, “Personalized Learning Model based on Deep Learning Algorithm for Student Behaviour Analytic,” *Procedia Computer Science*, vol. 163, pp. 125–133, Jan. 2019, doi: 10.1016/j.procs.2019.12.094.
- [29] Y. Lee, K. A. Kozar, and K. R. T. Larsen, “The Technology Acceptance Model: Past, Present, and Future,” *CAIS*, vol. 12, 2003, doi: 10.17705/1CAIS.01250.
- [30] J. Schepers and M. Wetzels, “A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects,” *Information & Management*, vol. 44, no. 1, pp. 90–103, Jan. 2007, doi: 10.1016/j.im.2006.10.007.
- [31] R. E. Mayer and R. Moreno, “Aids to computer-based multimedia learning,” *Learning and Instruction*, vol. 12, no. 1, pp. 107–119, Feb. 2002, doi: 10.1016/S0959-4752(01)00018-4.
- [32] R. N. Carney and J. R. Levin, “Pictorial Illustrations Still Improve Students’ Learning from Text,” *Educational Psychology Review*, vol. 14, no. 1, pp. 5–26, Mar. 2002, doi: 10.1023/A:1013176309260.
- [33] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, “A text-to-picture synthesis system for augmenting communication,” in *Proceedings of*

- the 22nd national conference on Artificial intelligence - Volume 2*, Vancouver, British Columbia, Canada, Jul. 2007, pp. 1590–1595.
- [34] H. Li, J. Tang, G. Li, and T.-S. Chua, “Word2Image: Towards visual interpreting of words,” 2008. Accessed: May 20, 2022. [Online]. Available: <https://scholarbank.nus.edu.sg/handle/10635/41084>
- [35] C. T. Li, C. J. Huang, and M. K. Shan, “Automatic generation of visual story for fairy tales with digital narrative,” *Web Intelligence and Agent Systems*, vol. 13, pp. 115–122, Jul. 2015.
- [36] A. G. Karkar, J. M. Alja’am, and A. Mahmood, “Illustrate It! An Arabic Multimedia Text-to-Picture m-Learning System,” *IEEE Access*, vol. 5, pp. 12777–12787, 2017, doi: 10.1109/ACCESS.2017.2710315.
- [37] J. Cho and N. Moon, “Design of Image Generation System for DCGAN Based Picture Book Text,” in *Advances in Computer Science and Ubiquitous Computing*, Singapore, 2020, pp. 265–270. doi: 10.1007/978-981-13-9341-9_46.
- [38] J. Zakraoui, M. Saleh, S. Al-Maadeed, J. M. Alja’am, and M. S. Abou El-Seoud, “Visualizing Children Stories with Generated Image Sequences,” in *Visions and Concepts for Education 4.0*, Cham, 2021, pp. 512–519. doi: 10.1007/978-3-030-67209-6_55.
- [39] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA, Dec. 2014, pp. 2672–2680.
- [40] T. Li *et al.*, “BeautyGAN: Instance-level Facial Makeup Transfer with Deep Generative Adversarial Network,” Oct. 2018, pp. 645–653. doi: 10.1145/3240508.3240618.
- [41] R. Li, L.-F. Cheong, and R. T. Tan, “Heavy Rain Image Restoration: Integrating Physics Model and Conditional Adversarial Learning,” 2019, pp. 1633–1642. Accessed: Apr. 11, 2022. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Li_Heavy_Rain_Image_Restoration_Integrating_Physics_Model_and_Conditional_Adversarial_CVPR_2019_paper.html
- [42] H. Wu, S. Zheng, J. Zhang, and K. Huang, “GP-GAN: Towards Realistic High-Resolution Image Blending,” in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, Oct. 2019, pp. 2487–2495. doi: 10.1145/3343031.3350944.
- [43] J. Yoon, D. Jarrett, and M. van der Schaar, “Time-series Generative Adversarial Networks,” in *Advances in Neural Information Processing Systems*, 2019, vol. 32.

- [44] K. Ehsani, R. Mottaghi, and A. Farhadi, “SeGAN: Segmenting and Generating the Invisible,” Jun. 2018, pp. 6144–6153. doi: 10.1109/CVPR.2018.00643.
- [45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2016, pp. 2234–2242.
- [46] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” *arXiv:1411.1784 [cs, stat]*, Nov. 2014, Accessed: May 06, 2022. [Online]. Available: <http://arxiv.org/abs/1411.1784>
- [47] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, “Voice-Input Multimedia Information Retrieval System Based on Text-to-image GAN,” in *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, Oct. 2019, pp. 943–944. doi: 10.1109/GCCE46687.2019.9015535.
- [48] Q.-V. Dang, T.-H. Pham, M.-L. Tran, M.-H. Dang, Q.-K. Nguyen, and A.-N. Nguyen, “Towards an automatic interior design system using GAN,” in *2021 4th International Conference on Data Science and Information Technology*, New York, NY, USA, Jul. 2021, pp. 170–172. doi: 10.1145/3478905.3478939.
- [49] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative Adversarial Text to Image Synthesis,” in *International Conference on Machine Learning*, Jun. 2016, pp. 1060–1069. Accessed: Dec. 06, 2021. [Online]. Available: <https://proceedings.mlr.press/v48/reed16.html>
- [50] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, “DRAW: A Recurrent Neural Network For Image Generation,” in *Proceedings of the 32nd International Conference on Machine Learning*, Jun. 2015, pp. 1462–1471. Accessed: May 06, 2022. [Online]. Available: <https://proceedings.mlr.press/v37/gregor15.html>
- [51] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv:1511.06434 [cs]*, Jan. 2016, Accessed: May 06, 2022. [Online]. Available: <http://arxiv.org/abs/1511.06434>
- [52] H. Zhang *et al.*, “StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks,” Oct. 2017, pp. 5908–5916. doi: 10.1109/ICCV.2017.629.
- [53] H. Zhang *et al.*, “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019, doi: 10.1109/TPAMI.2018.2856256.

- [54] T. Xu *et al.*, “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1316–1324. doi: 10.1109/CVPR.2018.00143.
- [55] M. Zhu, P. Pan, W. Chen, and Y. Yang, “DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5795–5803. doi: 10.1109/CVPR.2019.00595.
- [56] J. Johnson, A. Gupta, and L. Fei-Fei, “Image Generation from Scene Graphs,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1219–1228. doi: 10.1109/CVPR.2018.00133.
- [57] B. Zhao, W. Yin, L. Meng, and L. Sigal, “Layout2image: Image Generation from Layout,” *Int J Comput Vis*, vol. 128, no. 10, pp. 2418–2435, Nov. 2020, doi: 10.1007/s11263-020-01300-7.
- [58] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
- [59] E. Loper and S. Bird, “NLTK: the Natural Language Toolkit,” in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1*, USA, Jul. 2002, pp. 63–70. doi: 10.3115/1118108.1118117.
- [60] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.
- [61] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int J Comput Vis*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
- [62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.
- [63] P. S. Lokhande, F. Aslam, N. Hawa, J. Munir, and M. Gulamgaus, “Efficient way of web development using python and flask,” Mar. 2015, Accessed: Apr. 28, 2022. [Online]. Available: <http://localhost:8080/xmlui/handle/123456789/1367>
- [64] “かわいいフリー素材集 いらすとや.” <https://www.irasutoya.com/>
- [65] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter,

- “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2017, pp. 6629–6640.
- [66] X. Wang and J. Yu, “Learning to Cartoonize Using White-Box Cartoon Representations,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8087–8096. doi: 10.1109/CVPR42600.2020.00811.
- [67] Y. Shu *et al.*, “GAN-based Multi-Style Photo Cartoonization,” *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021, doi: 10.1109/TVCG.2021.3067201.
- [68] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, “Gated-GAN: Adversarial Gated Networks for Multi-Collection Style Transfer,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 546–560, Feb. 2019, doi: 10.1109/TIP.2018.2869695.
- [69] B. Proven-Bessel, Z. Zhao, and L. Chen, “ComicGAN: Text-to-Comic Generative Adversarial Network.” arXiv, Sep. 19, 2021. doi: 10.48550/arXiv.2109.09120.
- [70] “三毛英語季 - 英語口語，英語閱讀從此不是問題。”
<https://www.smyyj.com/>