1. 摘要

近年人工智慧的崛起對各領域皆有極為深厚的影響，不僅帶動科技的進步，同時也改善了人類的生活品質、娛樂等等，因此人工智慧的應用在近幾年成為炙手可熱的話題。

COVID-19 疫情的衝擊，使現今的人必須在保持社交距離的情況下進行遠端辦公等工作，人與人的交流更依賴網路及社群媒體。與現實互動的不同，網路及社群媒體的社交缺少副語言(Paralinguistics)[2]和非言語交際(Nonverbal Communication)，儘管有顏文字、Emoji、表情符號等發明的出現，但在傳遞上仍舊不夠精確。

因此本計畫期望以Discord社群軟體為例，捕捉使用者輸入之文字，透過改良Waifu Diffusion Model此AI 繪圖模型，依據使用者先前所設置之形象結合捕捉到的情緒語句，於Model中產生圖片，並以結果之圖片為基底生成多張相似的圖片，藉由Siamese Network[3]比對圖片之相似度進行過濾及排序，結合生成動態圖(Graphics Interchange Format, GIF），創造出不同於以往使用靜態貼圖的效果，有更為生動的情緒表達。

1. 研究動機與研究問題
2. 研究背景

當今COVID-19疫情的衝擊使人們必須在保持社交距離的情況下進行遠端辦公等工作，人與人的交流更為依賴網路及社群媒體。而這段時間虛擬主播的崛起，透過虛擬形象、不放自身在現實的樣子使大家排除現實上的顧慮進行交流，衍生出來的還有自製自身虛擬圖片等等。因此我們發現去除掉現實中一些因素，虛擬形象能使大家更坦開心胸進行交流。

人工智慧的興起，產生出許多訓練模型並運用在許多方面，像是預測未來績效、回應問題內容、識別物體形狀及功能。近幾年來，更有人將AI運用在繪圖上面，例如：Stable Diffusion[1]、Midjourney、DALL-E 2[4]等，讓使用者能夠透過文字的輸入，產生出對應輸入文字的圖畫，這些圖畫的精緻程度不亞於真人畫師所繪製的作品，引起世界各地的人們一陣討論。

1. 研究動機

溝通的時候，副語言(Paralinguistics)[2]是幫助我們解讀文意很重要的一環。在一般情況下的溝通中，副語言包含音量大小、語調、語氣停頓、聲音表情。除此之外，臉部表情或是肢體動作這樣的非言語交際也能幫助我們解讀語意。然而，我們在網路上使用文字進行交流時，就缺少了Paralinguistics和Nonverbal Communication來進行輔助，儘管有顏文字、Emoji、表情符號等發明的出現，但在傳遞上仍舊不夠精確。

雖然現階段Text to Image的模型能夠畫出精美的圖片，卻是靜態的圖片，而非動態圖。根據雙碼理論，大部分的人第一眼看到文字時，若為具體事物，腦中想像的會是一幅動態的場景，例如：一個金色長髮、綠色眼睛的女生在房間邊打遊戲邊喝可樂。

因此若是能夠將模型訓練成生成出連續且相似的圖片最後合併成短影片，符合想像的影片在傳達意境的效果上必定更為優良，若能將其用於教育、線上聊天室等場域，想必將大幅增進人際間的交流。

1. 研究問題

本計畫將研究如何改良Waifu Diffusion Model此圖像生成擴散模型，生成出連續且相似的圖片，並合併成 GIF 。

需要研究的問題如下：

1. 如何生成出類似的圖片？
2. 如何生成出具有細微變化的連續圖片？
3. 如何篩選生成出的圖片品質？
4. 如何將多張連續且相關之圖片合併成短影片？
5. 文獻回顧與探討
6. Generative Model

Generative Model[5]的概念大致為從一個已知的分布中採樣，將樣本放到Generative Model生成目標資料，最後要讓這些生成資料的分布與真實資料的分布越接近越好。

1. Diffusion Model

Diffusion Model 是一種Likelihood-based的模型，相比生成對抗網路(Generative Adversarial Network, GAN) 取得更好的生成效果。其核心精神是學習一個逐步降噪過程，在訓練模型時，逐步地將原圖加入雜訊。具體上來說，使用一個高斯分布一次又一次地在原圖上打上很小的雜訊，然後讓網路來學習如何反向這個雜訊。

它的主要作法是模擬馬可夫鏈 (Markov chain) 中的一系列高斯雜訊分佈來逐步地對原始訊號X\_0 (影像、音訊)添加高斯雜訊，進而生成一個服從高斯分佈的訊號 X\_T，接著訓練一個學習模型能夠將 X\_T 逐步地還原成 X\_0，步驟如下：

1. 從訊號逐步地添加高斯雜訊 (X\_0 ➞ … ➞ X\_t-1 ➞ X\_t ➞ … ➞ X\_T)的擴散過程 (Diffusion Process): q(X)
2. 從雜訊逐步地還原成訊號 (X\_T ➞ … ➞ X\_t ➞ X\_t-1 ➞ … ➞ X\_0)的逆擴散過程 (Reverse Process): p\_θ(X)
3. Waifu Diffusion Model

此計畫將要改良的 Waifu Diffusion Model 是微調 Stable Diffuison Model 架構，將資料集改為 Booru 網站中 680K 張的動漫圖片進行訓練，使其模型的輸入文字有條件生成為動漫人物類型的圖片。而 Stable Diffusiom Model 便是由更早之前所發表的 Latent Diffusion Model 下去進行優化。

Latent Diffusion Model 是一種自回歸模型，需要反復反覆運算計算，因而訓練和推理都十分昂貴。此提出一種擴散的過程改為在潛在空間上做的方法，從而大大減少計算複雜度，同時也能達到十分不錯的生成效果。

1. Text-to-Image Generation

DALL-E[6]作為開創高品質文字生成圖像的模型，是將文字生成圖像的問題視為序列與另一序列之間的轉換問題，使用Vector-Quantized Variational AutoEncoder(VQ-VAE)將文字標記轉換成圖像嵌入作為解決方案。往後更出現CogView2這種基於分層變壓(hierarchical transformer) 與局部平行自我回歸（Local Parallel Autoregressive）的文字生成圖像模型。

Aditya Ramesh 等作者提出之 DALL-E 2 模型[4]與 DALL-E 相異點在於此模型改良自GLIDE模型，使得模型能夠生成出與輸入文字有相同語意卻不完全一樣的圖片。模型架構分成兩個階段，一是訓練 Contrastive Language Image Pre-training (CLIP)[7]，將文字進行編碼，另一個訓練採用效率較高的擴散先驗(Diffusion prior)，將文字編碼轉換成圖像編碼，最後解碼，產生與輸入文字高度對應的高精緻度圖片。

Denoising Diffusion Probabilistic Models(DDPMs)[8]，其生成結果是根據Diffusion Probabilistic Models與使用朗之萬方程式(Langevin Dynamics)部分概念與去噪分數匹配(Denoising Score Matching)兩者結合而成。在西元 2022 年由 Robin Rombach 等作者提出的Latent diffusion[1]這種將 Diffusion Model 放到 潛在空間中訓練，在保有生成高品質圖片的特性下，解決訓練過程中計算複雜程度容易過高的情形，並且在模型架構中再另外加入交叉注意力層(cross-attention layer)，使其在無條件圖像生成（Unconditional Image Generation）、超解析（Super-resolution）、以及圖像修復（Inpainting）上皆有良好的表現。

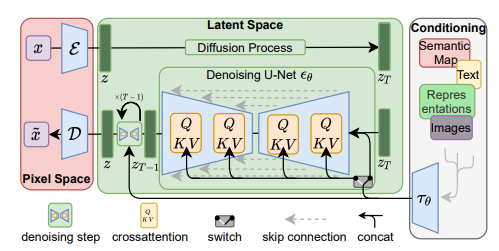


圖1. Latent Diffusion Model 架構圖[1]

1. Siamese Network

Siamese Network[3]為曾經用於簽字認證識別的神經網路，原理是先以神經網路得到特徵向量，而後利用兩個圖片的特徵向量判斷相似度，類似於尺度不變特徵轉換(Scale-invariant feature transform, SIFT)，只是是利用CNN進行提取特徵，並且用特徵向量構造損失函數進行訓練。

1. 副語言(Paralinguistics)

副語言[2]包括口音、音高、音量、語速和流利度，也包括某些非發聲現象，如:面部表情、眼球運動、手勢等，都能幫助人們進行解讀。而近幾十年來通過電子郵件、簡訊和社交媒體進行非面對面交流時，也興起如:表情符號、 Emoji 、顏文字等發明來代替副語言。

1. 雙碼理論 (Dual Coding Theory, DCT)

Allan Paivio於1971 年提出了「雙碼理論」[9]，該理論認為大腦的資訊尤以下兩種編碼組成：「圖像編碼」和「語言編碼」。「圖像編碼」為當人類想到狗的時候，想到的是一張狗的圖片；「語言編碼」則是想到「狗」這個詞。而當在思考抽象詞彙時，與思考貓、狗等具體詞彙會有差別。

當人類想到抽象的單詞時，就更容易用語言編碼來理解它們—找到定義或描述它們的單詞；對於具體的單詞，則使用圖像編碼並在腦海中呈現一張人或椅子的圖片，這往往比使用與之相關或描述性的單詞更容易被採用，而圖像編碼於腦海中呈現的圖像便是「心像」。

雙碼理論之圖像編碼，使人們在接收文字敘述後於腦中產生對應圖像，本計畫期望能根據使用者之形象設定，依據當前情緒或行為產生對應 GIF ，使原本的圖像編碼具現化。

1. 自然語言處理(Natural Language Processing)

自然語言處理(Natural Language Processing)[10]是計算機通過可計算的方法對自然語言的各級語言單位所進行的轉換、傳輸、儲存、分析等加工處理。

自然語言處理方法大致分為四類：符號法、統計法、連接法和混合法。符號法為透過該語言中通俗的解釋和相關演算法進行深度分析；統計法為使用大型文本語料庫，根據文本語料庫提供的這些現象的實際例子和數學運算建立語言現象(Language Phenomenon)的近似泛化模型；連接法與統計法類似，區別在於連接法的模型結合統計法與表徵理論。

1. 研究方法與步驟

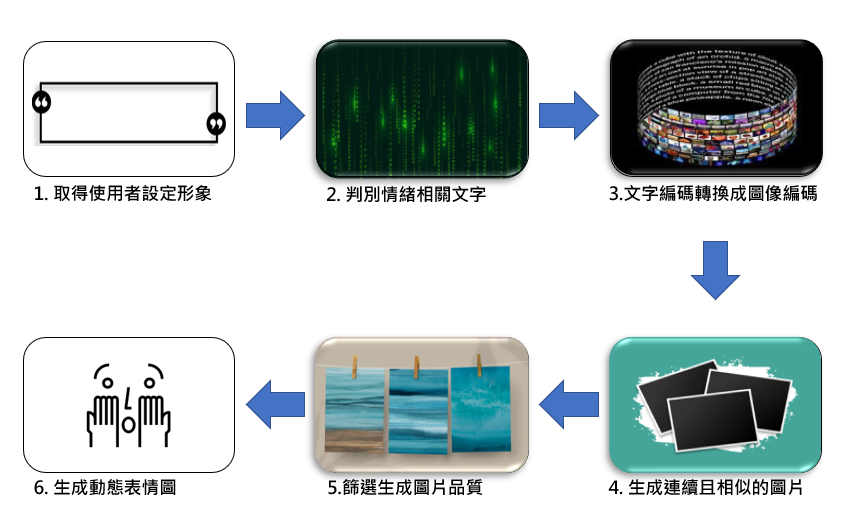


圖2. 研究計畫流程圖

本計畫預計設計一個介面讓使用者能夠輸入所想要生成動態圖的對應文字。將使用社群軟體，例如：Discord 所開發之機器人套件，並於聊天室內加入此機器人即時讀取使用者輸入之文字，以利後續行為判斷。

1. **取得使用者設定形象**

Discord Bot 於進入伺服器內時，使用者以對應指令設定形象，例如：/set blond\_long\_hair\_with\_green\_eyes\_wearing\_uniform\_and\_wearing\_a\_brac-elet，此外使用者所設定之形象將會綁定使用者自己的user id。

由於 Discord Bot 指令後方參數之設定，使用者輸入形象語句時空格需以底線替代，以確保存進同個參數內。此形象語句於之後偵側到情緒相關詞彙時會根據使用者所定義之形象產生。

1. **判別情緒相關文字**

於系統內建立情緒詞彙相關資料庫，例如：笑死、哈哈、哭了等網路流行詞彙或Emoji，並將各詞彙對應至情緒之形容詞。

情緒類別為該情緒的分類，本計畫將其簡單分為喜、怒、哀、樂，分數為該情緒所表現的張力，範例如下：

表1. 情緒詞彙與情緒類別對照表

|  |  |  |  |
| --- | --- | --- | --- |
| 網路常用情緒詞彙 | 情緒對應形容詞/動詞 | 情緒類別 | 情緒分數 |
| 笑死 | laugh | 喜 | 50 |
| 哈哈 | smile | 樂 | 30 |
| 哭了 | sob | 哀 | 50 |
| 難過 | sad | 哀 | 60 |
| 真的很不爽 | very angry | 怒 | 50 |

以Discord Bot即時捕捉使用者輸入之文字內容，若語句內包含資料庫所設置之詞彙，將資料庫中詞彙與原先設置之形象結合，例如：

使用者輸入：笑死 到底什麼東西

轉換使用者輸入語句：blond long hair with green eyes wearing uniform and wearing a bracelet laugh.

形成可以作為模型輸入的具體語句，並產生對應 GIF 。

若一個語句包含兩個以上情緒詞彙，檢查是否造成衝突，若為相同情緒類別，取高情緒分數者；若非相同之情緒類別，將語句增加連結詞。

範例如下：

1. 相同情緒類別

使用者輸入：笑死到底什麼東西啦哈哈

轉換使用者輸入語句：blond long hair with green eyes wearing uniform and wearing a bracelet laugh.

1. 非相同情緒類別

使用者輸入：哭了到底什麼東西啦哈哈

轉換使用者輸入語句：blond long hair with green eyes wearing uniform and wearing a bracelet sob and laugh.

1. **文字編碼轉換成圖像編碼**

本計畫將訓練 Contrastive Language-Image Pretraining(CLIP) 此文字與圖像預訓練模型的text encoder與 image encoder 。

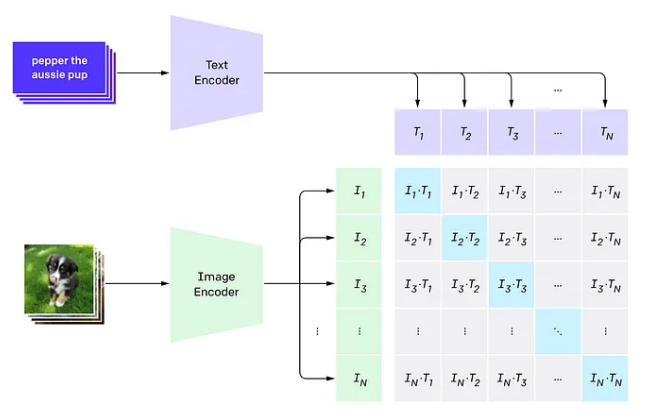
****

圖3. CLIP架構圖 (圖片來源：https://openai.com/blog/clip/)

1. **生成連續且相似的圖片**

使用Denoising Autoencoder(DAE)在圖像生成過程中加入輸入文字的類型。再結合Text to Image生成預想中的形象圖片後，使用 Image to Image的技術來產生大量連續且相似的圖片。

1. **篩選生成圖片品質**

將生成出來的圖片進行相似度比對，捨棄與其他圖片偏差過多的結果並將篩選出來的圖片進行相似度排序。

本計畫以 Siamese Neural Network 或 Convolutional Neural Networks[11] 比對多張生成圖片之間的相似程度，篩選掉差異性過大的圖片。

圖片排序：遞增再遞減。

例如，13542 （數字為圖片排序後的 Index）。

由於最終生成的結果為GIF 檔，將圖片以循環方式排序後去轉換便能減少其不順暢的部分，最大化減少連續幀數間圖片的落差，以達成連續動作之效果。

1. **生成動態表情圖**

使用Python的套件：Imageio，將篩選、排序後的圖片存進清單，以Imageio中的 Mimsave function將該清單儲存成GIF 並保存。

1. 預期結果
2. 預期結果

本計畫預期將依照實驗方法與下方進度表時程逐步完成輸入文字生成動態表情的系統開發，該系統會先讓使用者設定角色形象後，再判輸入文字描述中情緒字眼，最後將兩者結合後產生對應該輸入文字的動態表情貼圖，以達到比使用靜態貼圖更為生動的情緒表達，並由獨一無二的自行生成形象，使用者能有確實在不同人建立情感交流的效果。

以聊天視窗為例，以下圖片為我們的預期結果:

一張含有 文字, 白板 的圖片

自動產生的描述

圖4. 預期結果示意圖

本計畫期望透過自定義形象及語句情緒詞彙捕捉，以產生對應 GIF 替代副語言，使人與人之間互動、溝通時更能確切地傳遞資訊及縮短網路造成的距離感。

1. 一張含有 文字 的圖片

   自動產生的描述預期進度表(2023/07~2024/07)
2. 指導教授教導內容

本計畫包含訓練及改良Diffusion Model等內容。在資料蒐集的過程中雖能獲取相關作法與文獻，但由於尚未累積足夠程度的實務經驗，仍需指導教授的指導，且指導教授在訓練人工智慧的經驗豐富，故若以指導教授的經歷來協助本計畫尋求解決改進之方法，本計畫將會更為完善。

需要指導的內容摘要如下：

1. Diffusion Model 訓練
2. 篩選生成圖片之方法設計
3. Discord Bot 連結所訓練之Model 功能
4. 參考文獻

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10674-10685, 2021.

[2] B. Schuller *et al.*, "Paralinguistics in speech and language - State-of-the-art and the challenge," *Comput. Speech Lang.,* vol. 27, pp. 4-39, 2013.

[3] J. Bromley *et al.*, "Signature Verification Using A "Siamese" Time Delay Neural Network," in *International journal of pattern recognition and artificial intelligence*, 1993.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *ArXiv,* vol. abs/2204.06125, 2022.

[5] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM,* vol. 63, no. 11, pp. 139-144, 2020.

[6] A. Ramesh *et al.*, "Zero-Shot Text-to-Image Generation," *ArXiv,* vol. abs/2102.12092, 2021.

[7] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning*, 2021.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *ArXiv,* vol. abs/2006.11239, 2020.

[9] J. M. Clark and A. Paivio, "Dual coding theory and education," *Educational Psychology Review,* vol. 3, pp. 149-210, 1991.

[10] E. D. Liddy, *Natural Language Processing* (In Encyclopedia of Library and Information Science, 2nd Ed. ). NY. Marcel Decker, Inc., 2001.

[11] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4353-4361, 2015.