國立清華大學 電機工程學系

實作專題研究成果摘要

Context-Aware Meme Suggestion Using Large Language Models

使用 LLM 進行合適梗圖回復建議

專題領域：資工領域

組　　別：A515

指導教授：孫民

組員姓名：陳宥宇

研究期間：113 年 2 月 19 日至 114 年 5 月 2 日止，共 13 個月

Abstract

This project aims to develop a context-aware meme recommendation system using Large Language Models (LLMs). The system consists of two main stages: the first stage analyzes a large set of meme images and automatically generates three representative tags for each image, describing its possible mood, theme, or situation. The second stage responds to user input by combining the image's filename, caption, and previously generated tags to iteratively filter and select the three most suitable memes for reply.

To address the issue of inference degradation when the model is overloaded with information, a grouping and elimination strategy was introduced. In each round, only five meme entries are input into the model, from which the top three are selected. These are then regrouped for the next round of selection, until only the final three remain. This approach significantly reduces cognitive load on the model and improves recommendation accuracy.

To evaluate the impact of different meme information formats on recommendation quality, four versions of meme metadata were tested: 1) tags generated by LLaVA-13B, 2) tags generated by GPT-4o, 3) a descriptive sentence generated by LLaVA-13B, and 4) a basic format containing only the filename and caption. A user survey was conducted to collect preference feedback on each set of recommendations. Results showed that the first two tag-based formats performed significantly better, especially in accurately matching tone and intent. The third, sentence-based format performed the worst, likely due to vague or imprecise information that misled the model.

In addition to demonstrating performance differences across data formats, the highest scoring configuration achieved excellent feedback in the survey, validating the feasibility and potential of the system's design. It is expected that in the future, this system can be integrated into popular messaging platforms to offer users a fast and intuitive meme recommendation experience.

摘要

本專題旨在設計一套利用大型語言模型（Large Language Models, LLMs）進行情境對應梗圖推薦的系統。系統主要分為兩大區塊：第一階段會針對大量梗圖進行圖像分析，自動生成每張圖的三個代表性標籤（tags），描述其可能代表的心情、主題或情境；第二階段則會根據使用者所輸入的文字訊息，結合每張梗圖的檔名、台詞及前述標籤資訊，進行多輪篩選，推薦出三張最適合作為回覆的梗圖。

考量到模型在處理過多圖像資訊時容易造成推論誤差，本專題設計了一套資訊分組與淘汰機制，每次只讓模型接收五張梗圖的資料，選出三張後再重新分組進行下一輪篩選，最終得到三張最為推薦的圖片，減緩了模型處理資訊的壓力。

為驗證不同梗圖資訊格式對模型推薦品質的影響，實驗設計四種資訊版本：1）LLaVA-13B 產生的圖像標籤；2）GPT-4o 產生的圖像標籤；3）LLaVA-13B 生成的一句圖像敘述；4）僅包含檔名與台詞的基本資訊。藉由問卷調查，收集使用者對各組推薦結果的偏好，結果顯示前兩種帶有標籤的格式顯著優於其他，尤其在提升推薦語氣與意境的準確度上更具幫助。第三種敘述型資訊效果最差，推測因敘述可能引入模糊或不精確內容，反而影響模型判斷。

實驗結果除了顯示各資料型態間的差別外，最高分項在問卷中展示了優異的表現，證實此系統的設計理念是可行且有發展性的，期望未來能將此系統結合各類通訊軟體，為使用者提供快速且便捷的圖片推薦服務。

章節目錄

1. Introduction

研究動機

隨著網路通訊技術的迅速發展，使用貼圖、表情符號與梗圖已成為現代人日常溝通的重要元素。特別是在年輕族群中，透過趣味圖片進行交流不僅增添對話的趣味性，也成為表達情緒與態度的方式，我個人以及我的交友圈亦深受此文化影響。

在進行專題研究期間，我深入探索電腦視覺相關領域，發現人工智慧技術在圖像分析與生成方面的進展非常快速。大型語言模型（Large Language Models, LLMs）已能夠有效地理解並生成各種型態的內容，為圖像與語言的結合應用開啟了新的可能性。

然而，觀察現有的通訊軟體，雖然普遍具備貼圖與表情符號的自動推薦功能，卻鮮少針對梗圖提供相應的推薦機制。考量到梗圖在現代溝通中的重要性，因此我希望能結合所學，開發一套能夠理解使用者語境，並自動推薦適合梗圖的系統，以提升溝通的趣味性與效率。

研究目的與應用方向

本研究旨在開發一套結合大型語言模型與電腦視覺技術的梗圖推薦系統，能夠根據使用者輸入的文字內容，自動從使用者建立的梗圖資料庫中，推薦最適合的梗圖作為回覆。未來，此系統可應用於各大通訊平台，如 LINE、Messenger、Discord 等，或作為瀏覽器的擴充套件，讓使用者在不同平台上皆能快速應用這項便利的功能。

研究架構與實驗方法

本推薦系統主要分為兩大模組：

1. **梗圖分析模組**：利用 LLaVA-13B 模型[1]對使用者提供的梗圖進行圖像分析，為每張梗圖生成三個代表性標籤（tags），描述其可能代表的心情、主題或情境。
2. **推薦模組**：採用 LLaMA3-8B 模型[2]，根據使用者輸入的文字訊息，結合每張梗圖的檔名、台詞及前述標籤資訊，進行多輪篩選，推薦出三張最適合作為回覆的梗圖。

為了讓使用者能在本地端使用相對應的模型功能，實驗過程必須控制所挑選的語言模型大小，考量到小型模型在處理大量圖像資訊時可能出現推論誤差，我為系統設計了一套資訊分組與淘汰機制，每次僅讓模型接收五張梗圖的資料，選出三張後再重新分組進行下一輪篩選，最終得到最佳三張推薦圖。

考量到系統推薦品質可能會受到提供給模型的梗圖資訊格式影響，我設計了一項實驗，針對以下四種資料格式進行推薦效果的比較：

1. **LLaVA tags**：圖像由 LLaVA 產生的情緒/主題/情境標籤
2. **GPT-4o tags**：改由 ChatGPT[3]生成的標籤版本
3. **LLaVA 敘述**：LLaVA 為每張圖像生成的一段敘述文字
4. **台詞 Only**：僅包含圖片原始檔名與台詞，不提供額外資訊

我設計了十組對話情境，並邀請 20 位使用者參與問卷評選，每位受測者針對每題的多張推薦圖片，可以自由選擇他們認為對他們有所幫助的推薦結果。

實驗結果

實驗結果顯示，前兩種格式（含標籤）在推薦準確性上表現最佳。僅台詞版本雖然沒有語意標記，但表現尚可。敘述文字格式得分最低（105分），推測是因為語意過於模糊，反而干擾模型判斷。

這項結果不僅驗證了圖像標籤在梗圖推薦系統中的重要性，同時也說明本推薦系統在正確資訊條件下具備良好的實用性與潛力。

這次實驗我使用的是由 Meta 設計的開源模型，對於圖片的分析是使用以圖片為輸入的 LLaVA(13B 版本)模型，對於輸入語意與資訊分析則是使用 LLaMA3(8B 版本)模型，並且我皆使用開源工具 Ollama[4]在本地端運行這些模型，除了模型本身為開源模型外，系統中所使用的 prompt 設計、資訊分組策略等皆由我自行設計與實作。

2. Research Methodology

本章將介紹本研究所開發之梗圖推薦系統的設計方法與實作流程。系統主要分為兩個階段：梗圖分析階段與回覆推薦階段，分別負責提取梗圖的語意特徵與根據對話語境進行適當圖片推薦，大致結構如圖2-1所示。實作中結合了兩種大型語言模型（LLaVA-13B 與 LLaMA3-8B）以及自訂之 prompt 設計與分組推薦策略，以提升整體系統的推薦品質與運行效率。以下將依照各階段詳細說明模型使用方式、資訊處理格式與演算法流程。

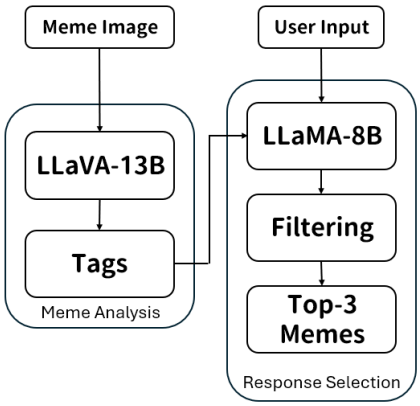

Fig. 2-1 實驗架構圖

梗圖分析階段

我首先使用 LLaVA-13B 模型分析輸入的梗圖圖像，為每張圖自動生成三個標籤（tags），描述其可能的情緒、主題或情境。我輸入的 prompt 如下：

「You are an AI that analyzes meme images. Look at the image and consider its visual content. Based on this, extract 3 English keywords (single words) that describe the mood, theme, or situation shown in the image. Only respond in this exact JSON

format: '{"tags": ["…, …, …"]}」

此段 prompt 主要包含兩個部分，第一是告訴模型其任務是分析圖片後生成三個關於它的標籤，第二部分則是限制模型的輸出格式，此方式能有效減少模型生成不必要文字的情況，提升推薦準確性與穩定度。

這些標籤將與圖檔名稱、原始台詞一同構成梗圖的資訊描述格式，存入 JSON 檔中，供後續階段使用。針對台詞資料，由於模型給出的台詞翻譯結果一直不甚理想，可能受到其他圖片資訊影響而參雜許多錯誤資訊，因此我是透過手動翻譯後，透過程式統一將台詞寫入完整資訊中。

回覆推薦階段

完成對圖片的分析後，LLaMA3-8B 模型會對使用者輸入的訊息進行語意理解，並與梗圖資料庫進行比對，從中推薦出最適合的三張圖片。我輸入的 prompt 如下：

「You are a meme recommender AI. The user said: "{user_input}". You are given a list of memes, each with a filename, description and caption. Pick the 3 memes that best responds the user's message in terms of tone, topic, or mood. Output ONLY a JSON list of filenames. Example: ["001.jpg", "005.jpg", "009.jpg"]. Do NOT explain anything. Meme list: {my dataset}」

此段 prompt 的設計與上一階段相同，同樣是在說明任務後，限制模型的輸出格式，但在研究過程中，我發現一次將所有資訊同時輸入時，LLaMA 模型容易產生語意雜訊，因此我設計出一種分組淘汰制的推薦策略，具體流程如下：

1. 隨機將所有梗圖分成數個五張為一組的小群組。
2. 針對每組，模型根據使用者輸入訊息，選出三張最適合回覆的梗圖。
3. 所有勝出的梗圖進入下一輪分組。
4. 重複上述過程，直到剩下最後三張圖片，即為本次推薦結果。

資料格式與模型設定

這次實驗總共使用了 50 張以海綿寶寶為主題的梗圖，我從網路上的開放來源[5]挑選下載後，作為輸入放進模型中分析，每筆梗圖資訊包含以下欄位：

1. **filename**: 檔案名稱，例如 "001.jpg"
2. **tags**: 圖像分析生成的三個關鍵詞，例如 ["shock", "surprise", "funny"]
3. **caption**: 原始梗圖中的台詞，例如 "I can't believe this!"

系統預設使用兩個模型，分別為：

1. 圖像分析模型：llava，13b 版本
2. 對話理解與推薦模型：llama3，8b 版本，所有模型皆透過 Ollama 呼叫執行。

3. Experimental Results

下圖 3-1 展示了此推薦系統的輸出結果，系統在接收到使用者的輸入內容後，進入上述的分析步驟，並輸出三張圖片的檔名，使用者即可依照檔名下載對應的圖片，或自行設計圖片的獲取管道，來將梗圖使用在想使用的地方。

Fig. 3-1 推薦系統輸出結果

Ablation Study

為了最佳化這個推薦系統，我共實驗了四種要輸入給模型的梗圖資訊格式，分別是：1）LLaVA-13B產生的圖像標籤；2）GPT-4o產生的圖像標籤；3）LLaVA-13B生成的一句圖像敘述；4）僅包含檔名與台詞的基本資訊。為評估推薦系統在不同「梗圖資訊格式」下的推薦效果，我設計了一份問卷，模擬實際使用情境，讓使用者根據推薦結果選擇他們認為符合標準的的回覆圖像。藉由分析各種資訊型態在問卷中的得分表現，來判斷哪一種資料呈現方式能有效提升推薦品質。

問卷設計

本問卷共設計 10 題，每題提供一段日常對話輸入，例如：”I'm hungry.”、”I love you!”、”Let's go have fun!”等，針對每一題，我的推薦系統會根據四種不同格式的梗圖資訊進行推薦，總共產生 4 組，每組 3 張圖片，即每題共提供 12 張推薦圖片。問卷採用多選題形式，測驗者可針對每題自由選擇認為最適合的梗圖，無數量限制。

我共邀請 20 位受測者填答，每位受測者可於每題中選擇任意組圖片，每選擇一組推薦圖片，即代表該資訊型態獲得一分。每種資訊型態在一份問卷中的最高得分為 10 分（假設受測者每個該資訊型態下的圖片皆合適），共有 20 位受測者，故每種資訊型

態理論上的總得分上限為 200 分。

實驗結果

Table 3-1
問卷調查之結果

序號	資訊類別	實際得分(分)	百分比(%)
Type 1	LLaVA-13B tags	183	91.5%
Type 2	PT-4o tags	181	90.5%
Type 3	LlaVA-13B 敘述	105	50.25%
Type 4	原始檔名與台詞	147	73.5%

從得分分布可發現：

1. 包含圖像tags的資訊型態明顯優於其他格式，Type 1與 Type 2 的表現接近，顯示不論是由 LlaVA 或 GPT-4o 所產生的關鍵詞，都對模型推薦有顯著幫助。
2. Type 4（無任何額外資訊）雖然缺乏語意標記，但仍優於 Type 3，顯示只靠原始台詞也能提供一定的判斷依據。
3. Type 3（敘述性文字）效果最差，推測可能因文字冗長、語意不精準，反而造成語言模型判斷偏誤。
4. 本系統整體在最佳資訊型態下(Type 1)達到 183 分,接近理論上限的 200 分，代表系統能根據上下文語境精準推薦適合的梗圖，證明系統架構與推薦演算法的實用性與可行性。

4. Conclusion

本研究成功實作出一套基於大型語言模型的梗圖推薦系統,具備根據使用者輸入訊息，自動推薦最適合梗圖回覆的功能。系統架構中，我透過 LLaVA-13B 模型對圖片進行語意標記,並利用 LLaMA3-8B 模型搭配多輪分組淘汰機制進行推薦,不僅克服了模型輸入限制的問題，也有效提升了推薦的精確度。

從實驗結果來看，本推薦系統表現符合預期，整體推薦準確度高，特別是在搭配含有圖像語意標籤（tags）的資訊格式下，獲得測試者明顯的偏好支持。雖然在每輪推薦中，由於必須選出三張圖片，有時候會出現部分圖像與語境略有出入的情況，但整體系統的運作邏輯與結果仍具備高度實用性與穩定性。考慮到目前的圖片庫僅包含 50 張梗圖，隨著資料量逐步擴充，未來推薦的精準度預期將有更顯著的提升。

此外，透過四種不同資料格式的實驗比較，也進一步驗證了圖像語意資訊在提升此推薦表現中的關鍵作用。這個發現證實了本系統的設計理念是有效的。

本系統目前已具備基礎的圖片分析與推薦能力，未來仍有多項發展與改進方向，包括但不限於：

1. 梗圖語言自動翻譯功能：針對圖片上的台詞自動生成翻譯，進一步提升整體的

自動化體驗。

2. 與通訊平台整合：可考慮將系統開發成通訊軟體的擴充套件（如 LINE BOT、Discord 插件等），進一步實現在真實聊天情境中的快速推薦功能。
3. 使用者個人化調整：未來可加入使用者偏好學習機制，針對不同用戶的使用習慣或回饋進行推薦優化。
4. 引入圖片生成技術：結合圖像生成模型（如 Stable Diffusion）在推薦不足時自動生成梗圖，擴展系統的應用彈性。

5. Reference

[1] Liu H, Li C, Wu Q et al. Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems . Red Hook, NY: Curran Associates, 2023, 34892 – 916.

[2] Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/

[3] ChatGPT. https://chatgpt.com/

[4] Ollama. https://ollama.com/

[5]【情報】有點高畫質的"海綿寶寶梗圖". https://forum.gamer.com.tw/C.php?bsn=60076&snA=5491441

6. Reflection and Thoughts

這是我第一次進行大型的研究專題，包括從前期的資料找查、訂定方向等，再到後期的所有實驗，讓我明白要專研一個領域並不是一件容易的事。找尋研究方向是對我來說最困難的一環，雖然實驗室提供我足夠的時間與資源去研究，但我對如何做好一項研究仍沒有太多的想法，感謝我的指導老師與帶領我的實驗室的學長，讓我最後得以沒有煩惱的進行所有研究與實驗，期望未來的我能將這段期間所學到的所有專業知識，以及對研究的態度，都化為成長的養分。