國立台北大學資訊工程學系專題報告

利用小模型交叉比較提升圖像語義通信方法

Enhanced Image Semantic Communication through Cross-Comparison of Compact Models

專題組員:曾浚睿、莊倚捷、侯加恩、王福莘

專題編號: PRJ-NTPUCSIE-113-001
執行期間:113 年 09 月 至 114 年 05 月

一、 簡介

語義通信是一種新興的通信方式，比起傳統通信更加專注於傳輸數據的語義內容，也就是數據所承載的意義或訊息，而非單純的文字序列。語義通信的目標是高效且準確地傳輸訊息的本質，而不是追求數據的無差錯傳輸，所以其允許在不影響理解的情況下丟棄非主要資訊的數據。

語義通信有以下幾個核心概念：

- 語義層面：語義通信會嘗試理解數據的上下文、目的和重要性，並優先傳輸句子的意思，而不是逐字逐句的無腦傳送。
- 高效性：透過聚焦於訊息的語義來減少冗餘數據的傳輸，從而節省頻寬和計算資源。
- 智能性：它通常依賴 AI 如機器學習或 NLP 模型來提取、編碼和解碼語義訊息。
- 容錯性：即使是在傳輸的過程中丟失了部分數據，只要語義能夠被正確還原的話，仍可視為傳輸成功。

而在執行語義通信過程中，若使用付費商用模型來提取語義資訊的話，因當前主流的付費方式皆為流量計費，在持續性使用下經常需付出大筆流量費用。為此我們嘗試尋找其他方法來免掉費用的同時又不失模型效能，並提出了利用多個免費開源小模型來替代付費模型，透過小模型之間交叉比較，互相判斷對方的回答是否正確並提供修改，以此來提升整體模型的效能。

二、 專題進行方式

I. 系統架構

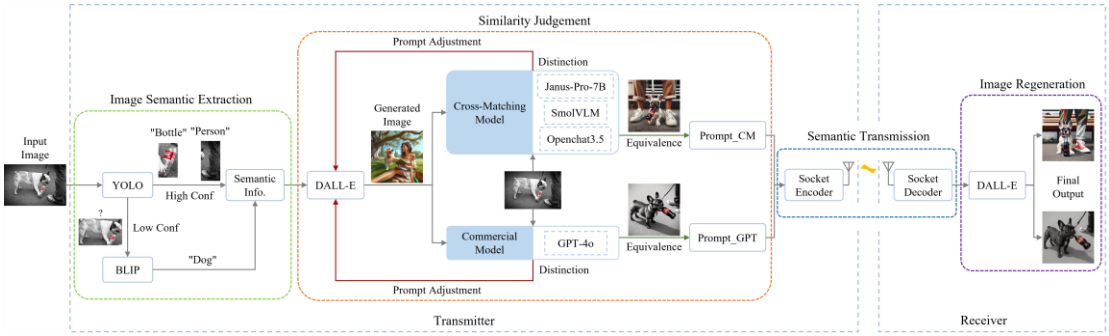
我們的語義通信系統可以大致分成四個部份(圖一)：

A. 圖像語義提取

我們使用了兩種物件辨識模型 YOLOv10 和 InstructBLIP 來提取圖片中的語義資訊，其中 YOLOv10 為主要辨識者，而 InstructBLIP 則是輔助辨識，當 YOLOv10 遇到分佈外 (Out-of-Distribution) 情況[1]時，再將辨識不出來的東西交由 InstructBLIP 來嘗試，可以想成 YOLOv10 知曉的領域不多，但都很精通，可以做出精準的辨識；而 InstructBLIP 則是知曉的領域廣，但並沒有那麼清楚，辨識結果會較為籠統。

B. 相似度判斷

其中又可分成圖像生成部分和循



▲圖一，語義通信系統整體架構

環判斷部分。

在圖像生成方面，我們使用了 DALL-E-3，作為 Chat GPT 也在使用的圖像生成模型，只要有準確的 prompt 就能生成高品質的圖片出來。

雖然該模型在目前使用上需付費使用 API 流量，但我們在多次嘗試各種免費小模型後，為了能夠有效地辨別出我們的交叉比較模型和付費商用模型的差異，仍選擇使用高效能的圖像生成模型來更好的反映出兩者產生的提示詞之間的優劣。

而在循環比對方面，又可再分成交叉比較模型和付費商用模型兩個部份，在交叉比較模型中我們使用了 Janus-Pro-7B 、SmolVLM-Instruct 和 OpenChat3.5，其中 Janus-Pro-7B 和 SmolVLM-Instruct 負責最主要的小模型交叉比較，而 Openchat3.5 負責整合這兩個小模型生成的提示詞，這部分在後續會有更加詳細的說明。

在付費商用模型中，我們使用了現在的付費模型中數一數二強大的 GPT-4o 來實作，付費模型在這套系統中的功用是作為交叉比較模型的對照組，用來協助判斷選用我們的模型是否會捨棄掉大量的模型效能。

C. 語義傳輸

我們選用了 Socket 擔任傳輸橋梁，Socket 作為網路標準通信介面，幾乎所有作業系統和編程語言都支援，能在不同平台和網路環境下通訊，可以直接通過網路傳送原始資料封包，且不限制網路範圍，不論是局域網或跨網際網路都能連線。

D. 圖像再生成

這部分為接收端的系統，為減少操作變因，我們同樣選用 DALL-E-3 作為圖像再生成的模型。

II. 系統運作

我們的系統操作如下：首先選擇一張想要傳輸的圖片，丟入我們的語義通信系統中，接著系統會利用 YOLOv10 與 InstructBLIP 辨識圖片中的物件，也就是語義資訊，系統會先使用 YOLOv10 來辨識物件，並為每個物件加上一個信度值，若信度值高於我們設定的信度閥值的話，便會視為正確辨識，若低於信度閥值的話則會視為錯誤辨識，並將低於信度閥值的物件整理起來丟入 InstructBLIP 中做二次辨識，並利用貝葉森優化來輔助調整輸出的離散機率，使結果更加準確。

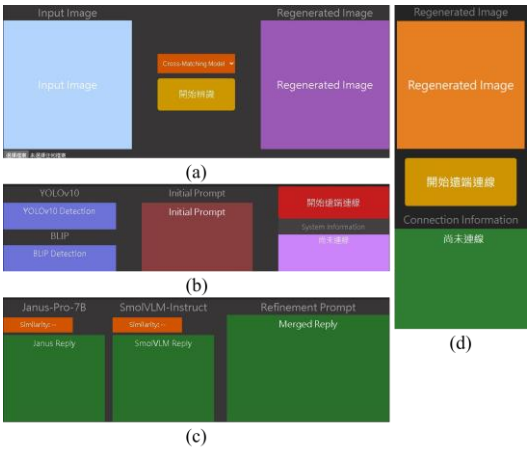
隨後將所有辨識出來的物件名稱合起來作為一串提示詞，丟入圖片生成模型中生成首張圖片，之後再依據

使用者的選擇來決定是利用我們設計的交叉比較模型，也就是 Janus-Pro-7B 和 SmolVLM-Instruct，或是利用付費商用模型來做相似度循環比較，這邊先使用我們的交叉比較模型來做解說。

當圖片輸入進交叉比較模型後，Janus-Pro-7B 和 SmolVLM-Instruct 會先各自判斷該輸入圖片和原始圖片之間的相似度，若相似度未達到我們設定的相似度閥值的話，便會生成一串用於修改輸入圖片，使其能更加接近原始圖片的提示詞。接著他們會將提示詞互相交給對方做檢查，比較對方的提示詞是否能夠正確的修改原始圖片，若可以的話便會將該提示詞作為對方的最終回應提供給 Openchat3.5，若不行的話則會修改提示詞內容，使其能更加準確地提供修改建議，改完後便會將該提示詞再交還給對方，讓對方再次檢查該提示詞是否有誤，並重複此循環直到兩方都認為該提示詞能夠正確修改輸入圖片，再將其提供給 Openchat3.5。而在兩模型的最終回應都確定後，Openchat3.5 便會將兩個提示詞整合在一起，將相似的句子整合成一句，並剔除掉重複的句子，得出交叉比對模型的最終輸出，一串由兩個小模型共同提出的提示詞，並將該提示詞提供給 DALL-E-3 來再次生成一張圖片，之後便是重複生成圖片➔比較相似度➔提供提示詞➔生成圖片的循環，直至某次循環中，Janus-Pro-7B 和 SmolVLM-Instruct 皆認為輸入圖片和原始圖片足夠相似，有正確包含原始圖片的語義資訊後，即可將生成出該圖片的提示詞傳送給接收端，也就是我們的 Socket 傳輸僅會傳送一串文字給接收端而已，達到語義通信所要的節省頻寬效果。

而在接收端收到提示詞後，系統會自動利用這串提示詞生成出帶有原始圖片語義資訊的圖片，以達到僅傳送文字亦能傳達圖片資訊的效果。

III. 系統介面



▲圖二，系統 UI 介面

我們的系統全程在 Pycharm 中進行，並且利用 Python 和 Vue 3 製作整個系統，其中 Python 用於製作後端系統，Vue 3 則是製作前端的使用者介面。

在發送端，使用者選好圖片及所使用的模型後，即可點擊「開始辨識」按鈕來擷取圖片語義資訊(圖二 a)，畫面上會顯示 YOLOv10 和 InstructBLIP 各自的擷取結果，以及一串整合起來的提示詞(圖二 b)，接著生成出來的圖片也會顯示在畫面上供使用者確認(圖二 a 紫色方框)，並且循環判斷的過程也會顯示在螢幕上，包括交叉比較模型的相似度分數、小模型提供的提示詞，以及整合後的提示詞(圖二 c)，而待模型確認正確的提示詞後，便可點擊「開始遠端連線」按鈕來與接收端做 Socket 連線(圖二 b 紅色按鈕)。

在接收端方面則只需和發送端一起點擊連線按鈕即可接收提示詞，並在接收到後自動生成圖片(圖二 d)。

IV. 專題製作過程與困難

在進行專題的過程中，我們四人各司其職，有人負責編寫程式，有人負責整理流程圖及計畫每周目標，雖然無法每天見面，但我們每周仍會找時間聚在一起討論當前進度與困難，例如我們曾遇到過：語義資訊辨識成果不佳，後來透過查找相關論文，找到了利用貝葉森優化來調整輸出結果的離散機率的方法[2]、Janus-Pro-7B [3]的輸出結果隨機性過高，無法正確輸出我們所需的回答格式，後來透過我們幾人不斷地微調參數，才找出一個可以穩定輸出正確格式，回答也不會太過死板的設置。

三、 主要成果與評估

我們的成品功能很完善，在學期初制定的預期功能皆有成功達成，雖然在製作過程中有過些許調整，例如在最初構想中的交叉比較模型是由 Gemini 1.5 Pro 和 SmolVLM-Instruct 來實作，後來才換成 Janus-Pro-7B，且原先設計上，整合提示詞的工作是交由 SmolVLM-Instruct 來執行，後面考慮到 SmolVLM-Instruct 專攻於圖像識別而非自然語言處理能力，才加入了 Openchat 3.5 來專門執行這項工序。儘管如此，我們最後還是成功製作出一套能夠無須付任何費用也能執行語義通信的系統出來，對此我們感到十分高興且驕傲。

不過我們仍未感到滿足，雖然最後我們的交叉比較模型能夠成功的生成能傳遞原始圖片中的語義資訊的提示詞，但若將其與付費商用模型的成果做比對，仍能看出兩者之間的差距，商用模型的成果不僅也能正確傳遞語義資訊，在物件之間的互動、位置、色調上都更加接近原始圖片，這在我們的交叉比較模型中是做不到的，因為一旦下了過多的指示，模型便會混淆我們的要求，導致輸出的結果雜亂且無邏輯，最大主因便是我們受到硬體限制，記憶體空間不足以應付我們的工作要求，因此我們必須將模型做量化處理，才能夠避免系統執行時間過長，但也因此導致模型效能降低，無法執行太過複雜的任務。

(a) (b)

▲圖三，(a)交叉比較模型認為與原始圖片相似的圖；(b)接收端再生成的圖

另外 DALL-E-3 雖然有著強大的圖像生成能力，但是仍帶有一定的隨機性，如同圖三，雖然使用相同模型和相同提示詞，但兩張圖仍然有所差異，若要解決這項問題就必須對模型參數做微調，但因為我們是使用呼叫 API 的方式來使用 DALL-E-3，因此我們並沒有辦法更改模型的參數。

在未來，待我們的硬體設備得到升級後，便不再需要對模型做量化處理，甚至能使用更多、更好的小模型來做交叉比較，或者能找到更加優秀、隨機性更低的圖像生成模型來取代現行的 DALL-E-3，相信屆時能夠得

到更加優秀的成果。

在這次長達近一年的專題製作課程中，我們不僅學到了許多語義通信和人工智慧的專業知識，還體認到了更勝以往的團隊合作能力，與過去的團體報告不同，專題製作的時間更長，難度也更高，若缺少了良好的團隊溝通和協作能力，想必會紛爭不斷，幸好我們團隊間不僅相處融洽，從未發生過爭執，在合作能力上更是默契絕佳，而這也多虧了我們良好的分工作業，各司其職的同時，也會向遇上困難的組員提供幫助，合作應付眼前的難題，不僅能夠減少爭吵的機會，更能加快專題製作的速度，若只將工作推給一個人做的話，不僅進度緩慢，更是極大佔用了此人的個人時間，想必會產生不少的爭端。對此我們所有人都很感謝彼此的體諒和信任，若缺少了任何一人，必定無法獲得這次的成果。

四、 結語與展望

在本專題中，我們製作出了一套使用了交叉比較模型的語義通信系統，該模型整合了多個免費開源小模型，並且會互相比較對方的輸出，不斷地修改調整，產生出模型們認為最佳的結果。透過這個設計，我們可以將其代替付費商用模型來做語義通信中的語義資訊提取和最佳化過程，達到完全免費的語義通信系統。

在未來不僅能夠擴充交叉比較模型中的小模型數量，還能隨意更換各種不同的小模型，並且操作簡單，即使是剛入門的人工智慧新手也能使用，不需大量的模型微調知識，只要將模型載入並替換變數名稱即可，對於和我們一樣的人工智慧新手十分友好。

五、 銘謝

在此感謝所有對本次專題製作有所幫助的人們，包括我們全組組員，以及我們的專題指導教授，若缺少了他的專業指導，我們不只有許多困難無法突破，更是連整體架構都會出錯。另外還要感謝實驗室的學長姐們，在我們遇到問題的第一時間往往會尋求他們的協助，在不需頻繁叨擾指導教授的同時又能迅速解決問題，十分的感謝它們的協助。

六、 參考文獻

[1] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, “Towards out-of-distribution generalization: A survey,” arXiv preprint arXiv: 2108.13624, 2021.

[2] Feifan Zhang, Yuyang Du Kexin Chen, Yulin Shao, Soung Chang Liew, “Addressing Out of Distribution Challenges in Image Semantic Communication Systems with Multi-modal Large Language Models", arXiv：2407.15335v1 [eess.SP] 22 Jul 2024.

[3] Chen, Xiaokang ,Wu, Zhiyu ,Liu, Xingchao ,Pan, Zizheng ,Liu, Wen ,Xie, Zhenda ,Yu, Xingkai , Ruan, Chong, “Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling”, arXiv preprint arXiv : 2501.17811, 2025.