台 南 應 用 科 技 大 學

資訊管理系 日四技

專題製作

結合 AI 技術與電腦視覺的即時物體辨識與回答系統-

以蔬果為例

組長：施坤政

組員：林聖祐　陳家翔

蔡景州　謝咏倫

指導老師：張至豪 老師

主　　任：黃玉枝 主任

中華民國 112 年 11 月

摘要

物體辨識是人工智慧領域中的重要課題，特別是在需要即時處理的應用場景中，準確且快速地辨識物體顯得至關重要。為了滿足這些需求，未來的影像辨識模型需要更高的精度與效率。YOLOv7（You Only Look Once v7）於 2022 年 7 月推出，憑藉其優異的辨識速度和高準確度，成為當前物體偵測的前沿技術之一。隨著人工智慧的進步，LLM（大型語言模型）逐漸接近人類語言的表達方式，並廣泛應用於各種場景，如客服、虛擬助手及內容生成工具。2022 年 11 月，OpenAI 推出了具備強大自然語言處理能力的 ChatGPT，能夠在多個學科領域生成高質量的文字回答。

本研究旨在開發一個結合視覺辨識與問答系統，將 GPT-3.5 語言模型與 YOLOv7 物體檢測技術融合，實現即時的物體識別與智能回答功能，形成視覺語言模型 (Vision-Language Model, VLM)。本系統能夠通過攝影機辨識物體，並根據用戶需求即時生成相關回答，實現了跨越影像與語言的互動。最終目標是在手機等便攜設備上實現高效、即時的辨識和問答應用，提升日常生活中的互動體驗。

關鍵字：即時視覺識別、自然語言處理、人工智慧應用、GPT-3.5、YOLOv7

目錄

摘要.....I

目錄.....II

圖目錄.....III

表目錄.....IV

第一章緒論.....1

1.1研究動機.....1

1.2研究目的.....2

第二章文獻探討.....3

第三章研究方法.....6

3.1研究架構.....6

3.2系統流程.....7

3.3YOLOv7.....8

3.3.1數據集.....9

3.3.2遮擋問題.....9

3.4GPT 3.5語言模型.....10

3.4.1成本花費.....10

3.4.2GPT提取成本.....11

3.5系統整合.....11

3.5.1手機連結.....11

第四章研究結果.....13

4.1評估指標.....13

4.1.1混淆矩陣.....14

4.1.2分析訓練結果.....15

4.1.3物件遮擋訓練.....17
第五章結論與建議.....20
5.1結論.....20
5.2未來建議.....20
參考文獻.....21
專題團隊工作分配一覽.....23

圖目錄

圖 1. VLM模型的處理流程.....3
圖 2. PICa 架構圖.....4
圖 3. LLaVA架構圖.....5
圖 4. VL-BERT架構圖.....6
圖 5. 研究架構圖.....7
圖 6. 系統流程圖.....8
圖 7. YOLOv7測試比較圖.....9
圖 8. 邊界框示例圖.....14
圖 9. 連結流程示意圖.....15
圖 10. 手機連結視覺辨識系統.....16
圖 11. 混淆矩陣500次訓練.....16
圖 12. 500次訓練-Percision.....17
圖 13. 500次訓練-Recall.....18
圖 14. 500次訓練-MAP0.5.....18
圖 15. 預測結果 17
圖 16. 500次訓練-Percision.....17
圖 17. 500次訓練-Recall.....18

圖 18. 500次訓練-MAP0.5.....18
圖 19. 有無針對遮擋訓練之差異.....19

表目錄

表 1. 訓練結果.....15
表 2. 遮擋物件訓練結果.....18

第一章 緒論

1.1 研究動機

2023 年被視為 AI 元年，人工智慧技術的應用在全球範圍內迅速擴展，生成式 AI（如 OpenAI 開發的 ChatGPT）成為其中的焦點之一。LLM（大型語言模型）應用範圍廣泛，不僅能進行自然對話，還能回答涉及各個學科領域的問題。基於此趨勢，我們提出了一個結合即時影像辨識與語言模型的應用構想，以蔬果識別為例進行研究。

過去，卷積神經網絡（Convolutional Neural Network, CNN）常被用於圖像分析[1]，而長短期記憶網絡（Long Short-Term Memory, LSTM）則被用於生成圖像描述[2]。然而，LSTM 在生成流暢且多樣的語言描述上存在局限性，且 CNN 在即時性與準確性方面的表現還有待提升。因此，我們計畫運用不同影像識別模型，以確定其能否滿足即時物體辨識和語言描述的需求。

如果目標檢測模型能在保持準確度的同時實現更快速的結果生成，並且語言模型能夠流暢且精確地描述圖像內容，這將大大提高影像識別問答系統的實用性。

1.2 研究目的

目標檢測結合圖像描述一直是近幾年討論的問題，考慮到目標檢測需要的即時與準確性，需要慎重挑選檢測模型。已有前人研究出 CNN 變種模型包括 (Faster Region with Convolution Neural Network，Fast R-CNN)[3]、(Single Shot MultiBox Detector，SSD)[4] 針對前兩者，我們使用本次實作的 YOLOv7 進行實驗，並使用混淆矩陣(Confusion Matrix)、準確率(Accuracy)、精確率(Precision)、召回率(Recall)、F1 Score、真陽率(True Positive Rate)、假陽率(False Positive Rate)效能指標進行評估。對語言模型，使用本研究專用的 GPT-

3.5 model 進行實驗。

在最終階段，確保正確識別食材並完整描述需求至關重要。這不僅需要在每個段落進行評估，還需要避免語言和目標檢測模型出現過擬合(overfitting)和樣本偏差（sample bias）等問題[5]，這些都是培養優質 VLM 模型的關鍵。因此，我們將著重於避免模型在訓練過程中對特定樣本過度適應，並尋找更具代表性的數據來訓練模型，以確保其性能在現實應用中的可靠性和穩定性。

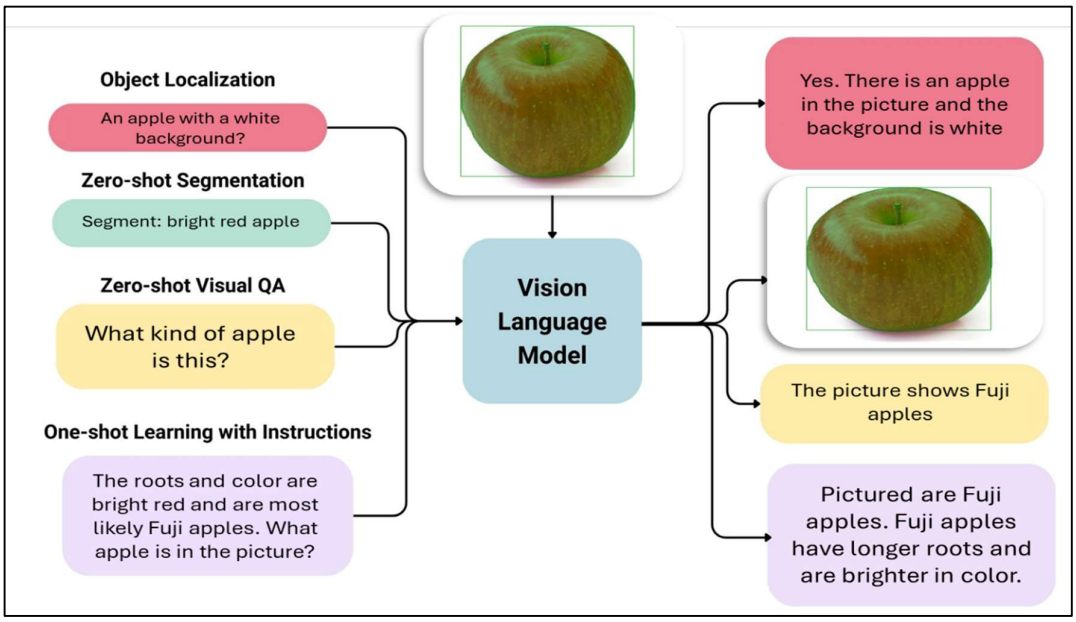

圖 1　VLM 模型的處理流程

圖 1 展示了 VLM 模型的架構及其應用流程，該模型結合了目標檢測和圖像描述功能。具體來說，VLM 能夠在物體定位、圖像分割、視覺問答及基於指令的學習等多個任務中實現端到端的處理。主要圖 1 強調了模型如何有效應對不同場景下的任務需求，並展示了模型在多任務學習中的靈活性與應用潛力。

第二章 文獻探討

近幾年全球的科技及人工智慧學家與組織對於 VLM Model 的討論及研究可說是非常活躍，因為 LLM Model 華麗登場於世界舞台，在這之後各大研究機構也紛紛找尋能與 LLM 結合並用的程式工具，並提高實用性與便利性。這之中最有價值的工具之一就包含 VLM Model。以下解釋幾個有關的 VLM Model 文獻:

Uzair SHAH 等人[6] 的研究提出了一個創新的框架，結合了 YOLOv7 進行物件偵測，並使用 GPT-3.5 Turbo 進行語義分析，以評估兒童在藝術表現中的情緒。具體來說，他們利用 YOLOv7 來辨識和檢測兒童繪畫中的各類物件，如人物、動物、物品等，這些物件的識別結果隨後被傳遞給 GPT-3.5 Turbo，用於生成基於這些物件的情緒報告。GPT-3.5 通過分析這些圖像中物件的語義關聯性，推斷出畫作中隱含的情緒，並生成報表，為父母和治療師提供有關兒童心理狀態的寶貴見解。這樣的結合使得情感推斷從單一的圖像辨識提升到更高層次的語義解讀。他們的方法展示了如何有效運用 YOLOv7 進行高精度、即時的圖像物件辨識，並通過 GPT 模型將圖像資料轉化為有意義的語義資訊。他們的研究特別強調了在影像識別過程中的準確性和即時性，並探討了在實際應用中如何通過影像增強等技術來優化這些方面的表現。這項研究對我們的專題有很大啟發。我們同樣採用了 YOLOv7 進行物件偵測，主要應用於蔬果識別任務。Shah 等人的做法為我們提供了寶貴的參考，尤其是在如何處理複雜場景下的影像增強和保持高辨識準確率方面。此外，他們使用 GPT-3.5 Turbo 生成情緒報告的方式，也啟發了我們在研究中應用類似的語言生成技術。我們的做法是使用 YOLOv7 進行目標檢測，然後將檢測結果與物體名稱進行關聯，並使用 GPT 進行語義化描述，這不僅提升了我們系統的可讀性，也加強了識別結果的語義表達能力。

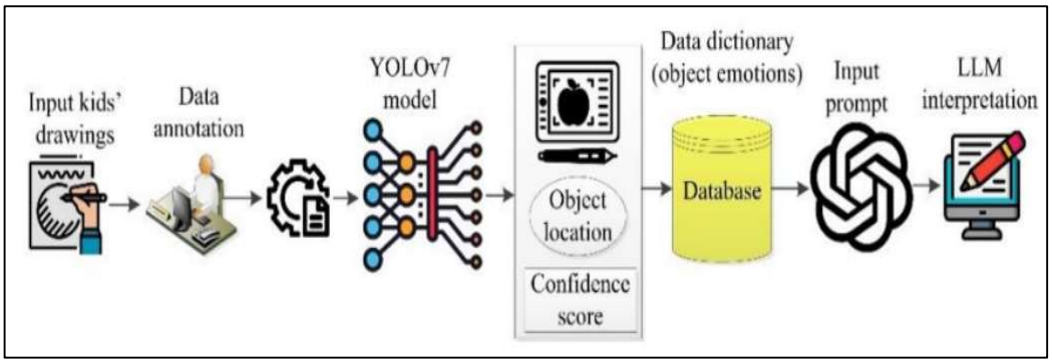

圖 2　情緒分析實驗架構圖

李等人[7] 開發出了一個名為 LLaVA（Large Language and Vision Assistant）的多模態模型的發展。 LLaVA 是一個端到端訓練的大型多模態模型，它能夠連接視覺編碼器和語言模型，用於通用的視覺和語言理解任務。 在論文中，作者將視覺編碼器選用了 CLIP（Contrastive Language-Image Pre-training），這是一種用於影像理解的視覺模型。 透過連接 CLIP 的視覺編碼器和語言模型，LLaVA 能夠同時處理影像和語言輸入，從而實現多模態理解任務。 此外，論文也探討了多模態指示跟隨中的挑戰，並提出了一些解決方案。其中，一種方法是使用僅基於語言的模型（如 GPT-4）產生多模態語言-圖像指示跟隨數據，並透過這些數據對大型多模態模型進行微調。 此外，論文還介紹了一些基準測試、實驗結果，並宣布了模型和相關資源的開源。

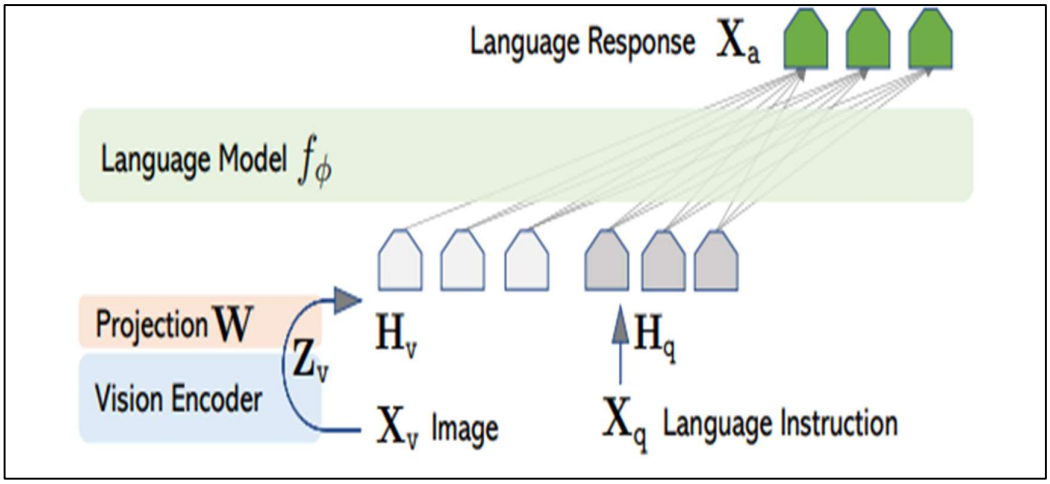

圖 3　LLaVA 架構圖

蘇等人[8]採用 Transformer 模型作為 VL-BERT 的基礎架構，Transformer 是一種強大的神經網路模型，特別適用於處理序列資料。 VL-BERT 將 Transformer 模型擴展到視覺-語言任務中，同時考慮了圖像和文字輸入。

在預訓練階段，VL-BERT 使用了大規模的 Conceptual Captions 資料集和純文字資料集。 在視覺-語言資料集上的預訓練採用了一種掩碼語言模型（MLM）的方式，透過預測隨機屏蔽的單字或感興趣區域（RoI）來提高模型對視覺-語言線索的聚合和對齊能力 。 這樣的預訓練方式有助於 VL-BERT 更能理解和處理影像和文字之間的關聯。

透過在大規模資料集上進行預訓練，VL-BERT 能夠學習到豐富的視覺和語言表示，從而在各種視覺-語言任務中取得顯著的效能提升。 這項研究為跨模態資訊融合提供了新的思路和方法。

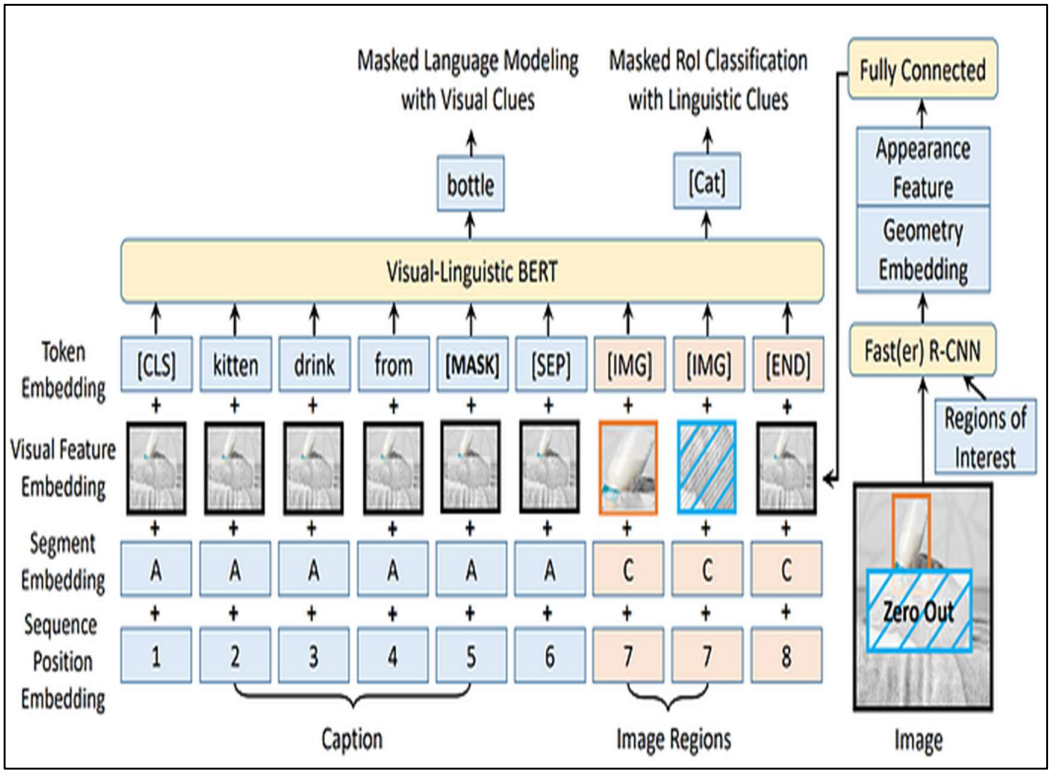

圖 4　VL-BERT 架構圖

第三章 研究方法

3.1 研究架構

研究強調了在目標圖像檢測中實現速度和準確性的重要性。該研究分為三個主要步驟。

[1] 第一步是收集相關食材圖片並對它們進行標籤，以便進行訓練。在訓練完成後，使用四項效能評估方法，如 Precision 等，來評估模型的優缺點。

[2] 第二步是啟用 GPT-3.5 模型，並將 YOLOv7 的結果集成到其中。這些結果暫時存儲在 GPT-3.5 模型的輸入層中並轉換為數字形式。

[3] 最後一步是，根據給予 GPT-3.5 指定的提示，GPT-3.5 將這些數字化的結果解釋為對象代碼，直到用戶輸入問題促使它們被傳送到 GPT-3.5 。

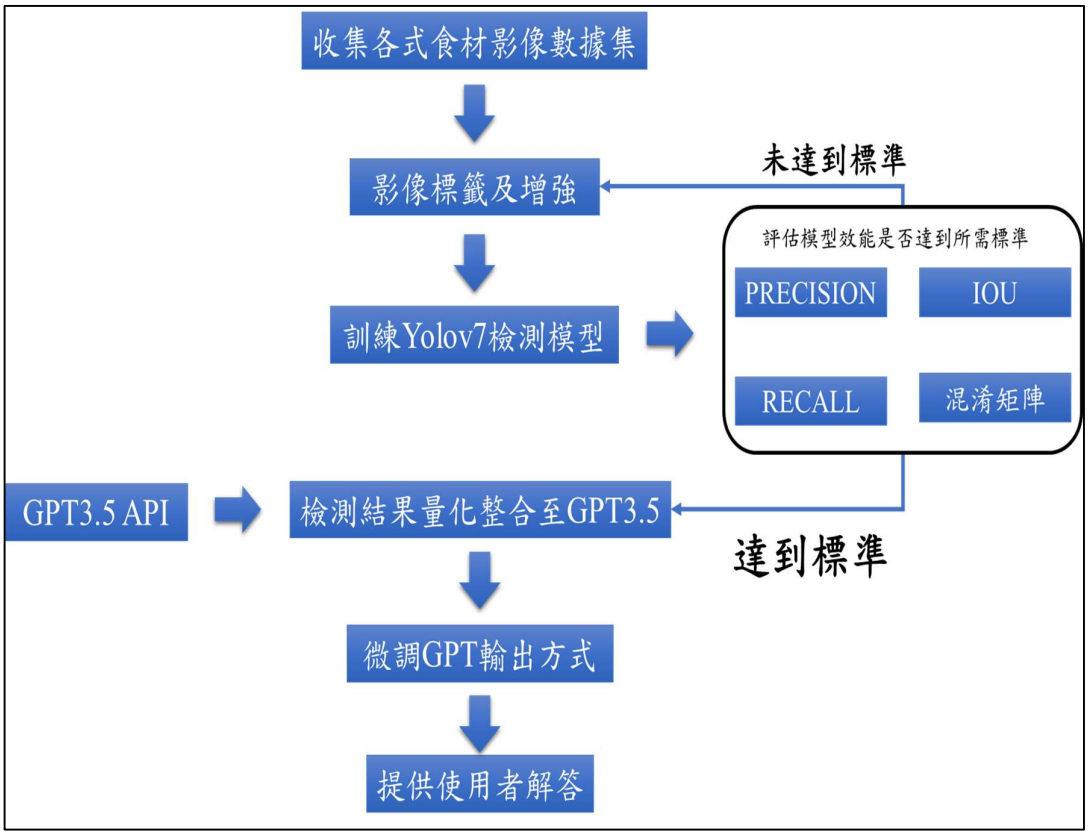

圖 5 研究架構

3.2 系統流程

本系統採用 Python 套件結合 YOLOv7 及 GPT-3.5 API。使用者透過手機連接至電腦後台，透過 Web 頁面提供問題輸入介面，同時顯示系統接收到的畫面。GPT-3.5 回答的結果會在提交問題後即時呈現，並支援持續對話功能。

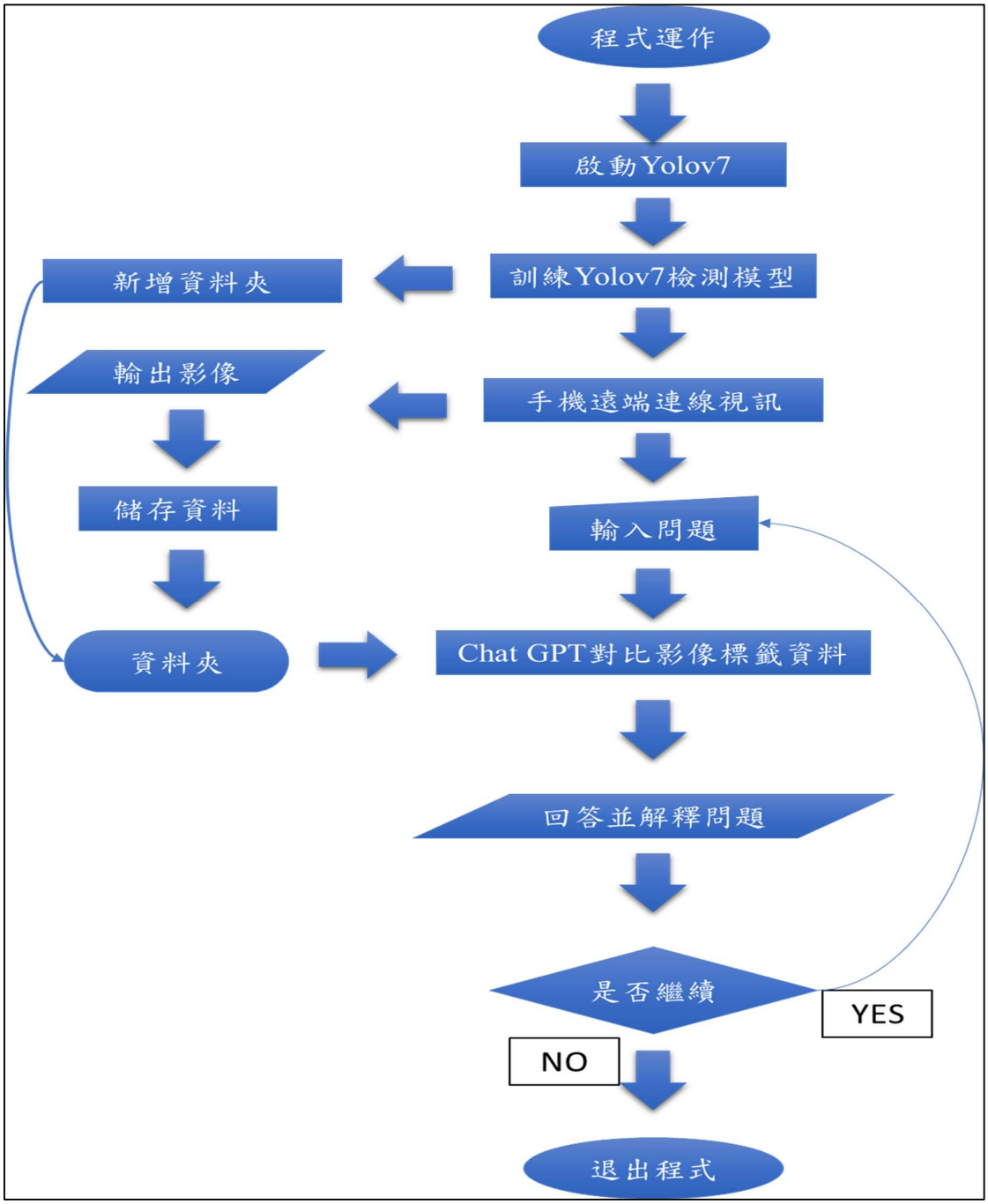

圖 6 系統流程圖

3.3 YOLOv7（You Only Look Once v7）

要能夠迅速準確識別影像目標並且要在好開發的前提下，本研究採用YOLOv7 物體檢測模型，YOLOv7（You Only Look Once v7）[9]是一種物體檢測模型，於 2022 年 7 月問世。它在物件辨識速度和高準確度上展現出卓越的表現。

YOLOv7 在速度和準確度方面超越了所有已知的物體檢測器，其速度範圍從每秒 5 幀到每秒 160 幀，並且在 GPU V100 上以每秒 30 幀或更高的速度達到了最高的準確率，即 56.8% AP。YOLOv7-E6 物體檢測器（V100，每秒 56 幀，準確率 55.9% AP）在速度上比基於 transformer 的 SWINL Cascade-Mask R-CNN（A100，每秒 9.2 幀，準確率 53.9% AP）快了 509%，而準確度提高了 2%；比基於卷積的 ConvNeXt-XL Cascade-Mask R-CNN（A100，每秒 8.6 幀，準確率 55.2% AP）快了 551%，而準確度提高了 0.7% AP。此外，YOLOv7 在速度和準確度方面也超越了 YOLOR、YOLOX、Scaled-YOLOv4、YOLOv5、DETR、Deformable DETR、DINO-5scale-R50、ViT-Adapter-B 等許多其他物體檢測器。值得注意的是，團隊僅使用 MS COCO 數據集從頭訓練了 YOLOv7，沒有使用任何其他數據集或預訓練權重。

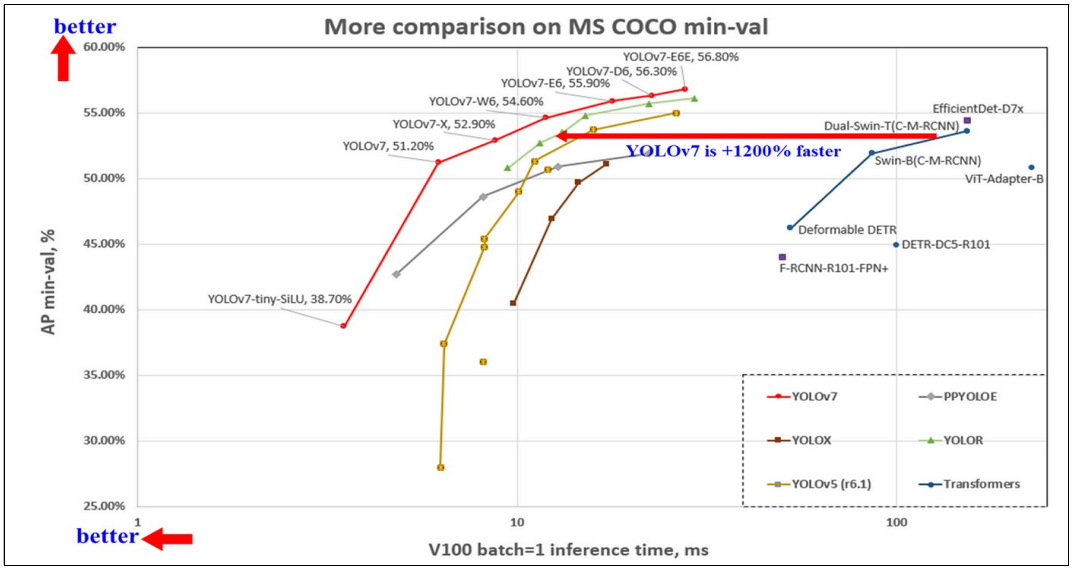

圖 7　YOLOv7 測試比較圖

3.3.1 數據集

圖片標註(Image Annotation Types)可以使用多種不同的技術和方法來標記圖像中的物體和特徵。這些方法包括邊界框(Bounding boxes)、多邊形分割(Polygonal Segmentation)、語義分割(Semantic Segmentation)、 3D 長方體標註(3D cuboids)、關鍵點和界標標註(Key-Point and Landmark)、直線和樣條標註(Lines and Splines)[10]。而邊界框(Bounding boxes)，通常用於物體檢測和識別，這也是我們 YOLOv7 所使用的方式，並且在識別後會以方框標記目標物，如圖 8。

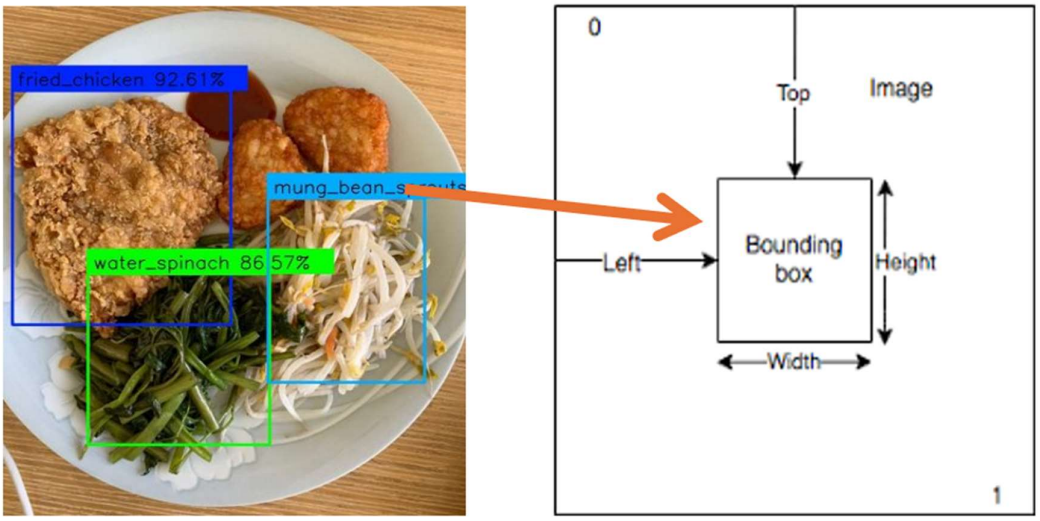

圖 8　邊界框示例圖

本研究總共採用 25 種蔬果作為檢測對象，每個種類的蔬果都以三十張為準則。包括:

[胡蘿蔔、洋蔥、高麗菜、萵苣、香菇、茄子、番茄、蘋果、芒果、梨、哈密瓜、南瓜、香蕉、馬鈴薯、花椰菜、白蘿蔔、蘆筍、葡萄、西瓜、青蘋果、筊白筍、芭蕉、紫甘藍、白葡萄、印度芒果]。

3.3.2 遮擋問題

在我們的 YOLOv7 實作中，引入了許多新方法，特別是在非極大值抑制（NMS）[11]階段針對袋裝食材的遮擋問題進行處理。 這項設計想法非常有效率，因為在模型設計和訓練階段解決袋裝食材遮擋問題需要涉及大量的研究，而且不一定能夠完全解決問題。 透過利用更好的 NMS 演算法，我們能夠顯著

提高袋裝食材的偵測效果，尤其是在夜間光照不足、目標被遮蔽導致資訊缺失以及行人目標多尺度的情況下。

在使用我們的 YOLOv7 進行 NEU-DET 實驗時，我們注意到最終結果中存在重複檢測的問題。 因此，如何在解決袋裝食材遮擋問題的同時減少重複檢測，即保留更少的框，也成為我認為 NMS 優化演算法中值得研究的方向。

我們的 YOLOv7 在設計上借鑒了 Yolov4 中的一些最佳化方法，如採用了改進的 NMS 演算法，例如 DIOU_nms[12]。 透過這種方式，我們能夠更好地處理被遮蔽的袋裝食材，進一步提高了檢測的準確性。

3.4 GPT 3.5 語言模型

GPT-3.5 是一種基於深度學習的語言模型，由 OpenAI 開發。它建立在前一代模型 GPT-3 的基礎上，擁有更多的參數和更強大的能力。GPT-3.5 通過預訓練和微調的方式，能夠在多個自然語言處理任務上實現卓越表現。GPT-3.5 採用了深度轉換器架構，其中包含 2000 億個參數。它由多個編碼器-解碼器堆疊而成，每個編碼器-解碼器都是一個自注意力機制的 Transformer 模塊。編碼器負責將輸入文本（例如問題或提示）轉換為一系列特徵向量。這些特徵向量捕捉了輸入文本的語義信息。解碼器接收編碼器生成的特徵向量，並根據這些特徵生成輸出文本。解碼器使用自注意力機制來關注輸入的不同部分，以生成流暢自然的文本。這種架構使得 GPT-3.5 能夠捕捉長距離的語義依賴關係，並生成流暢自然的文本。

3.4.1 成本花費

在本研究中，使用 GPT-3.5 模型產生的成本，主要是費用方面：

API 使用費用：

GPT-3.5 定價：

每 100 萬個輸出 token：6.00 美元

每 100 萬個輸入 token：3.00 美元

平均而言，一個中文字大約相當於 2 個字元或約 0.49 個 token。這意味著，相較於中文文本，英文文本在 token 使用方面相對更有效率[13]。

3.4.2 GPT 提取成本

OpenAI 的生成式預訓練語言模式（GPT-3.5）在自然語言處理任務中表現優異。然而，頻繁使用這些模型會帶來顯著的 Token 消耗和費用。本節旨在探討並提出降低使用成本的有效策略。目前提出的有效方法如下:

1. 透過簡化問題描述、刪除冗餘詞彙和使用縮寫來減少輸入中的 Token 數量。
2. 透過設定 max_tokens 參數限制每次請求的最大 Token 數量，避免不必要的運算資源浪費。
3. 控制模型生成回答的長度，減少不必要的 Token 消耗。
4. 對頻繁查詢的結果進行緩存，並將多個問題合併成一次請求以減少單次請求的 Token 數量。

3.5 系統整合

使用 HTML 設計網頁介面，包括輸入文字描述的文本框和上傳圖像的按鈕。通過使用者的 Google 瀏覽器訪問本機伺服器，用戶可以輸入文字描述，通過 OpenCV 庫的函數將影像編碼為 JPEG 格式，返回一個元組，其中包含編碼結果的布爾值和編碼後的影像資料(buffer)。點擊按鈕上傳目前影像擷取的圖像。當用戶提交表單時，表單中的文字描述和上傳的圖像將被後端 Python 應用程式接收。後端使用 Python 套件（Flask）處理用戶輸入的數據。將接收到的圖像傳送到 YOLOv7 模型進行物體檢測。YOLOv7 模型會識別圖像中的物體並返回其位置和類別。接著將 YOLOv7 的檢測結果傳送到 GPT-3.5 模型進行解釋，生成對象代碼或其他自然語言描述。GPT-3.5 模型會根據收到的圖像檢測結果生成相應的語言描述。系統支持持續對話，後端可以追蹤用戶的歷史輸入並基於上下文提供連續的對話回應。系統會將處理後的結果即時返回給前端，用戶可以在瀏覽器中看到系統的回答和解釋。

3.5.1 手機連結

本系統支持用戶通過手機與電腦後台連接，具體流程如下：

1. **連接方式：**
   用戶可以使用手機通過 Wi-Fi 或 USB 連接至本地伺服器，然後使用 Web 瀏覽器進入系統網頁介面。當使用者掃描 QR Code 進入伺服器頁面後，網頁將首先請求授權以持續啟用手機相機（若使用者允許），相機將自動啟動，並開始由 YOLOv7 模型進行即時物體檢測。

2. **界面交互：**

   在網頁介面中，用戶可輸入問題或上傳圖片，並選擇直接提交。系統會即時顯示手機攝像頭捕捉的畫面，並允許用戶上傳圖片進行進一步分析。

3. **系統處理：**

   用戶輸入的文字和上傳的圖片會被後端的 Python 應用程序處理。YOLOv7 模型負責進行目標檢測，GPT-3.5 模型則根據檢測結果生成語言描述並回答用戶的問題。

4. **結果展示：**

   隨後，網頁將重新載入並即時顯示處理結果（包括檢測到的物體和描述），用戶可以在手機瀏覽器中查看。系統還支援持續對話，根據用戶的歷史輸入提供連貫的回應。

這樣的設計使得用戶能夠輕鬆通過手機與系統進行即時互動，並獲得準確的反饋和資訊。

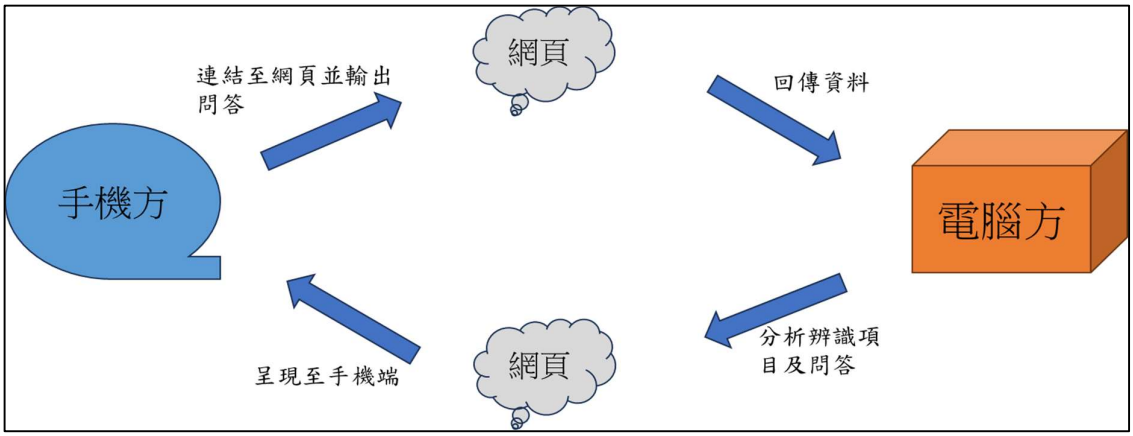

圖 9 連結流程示意圖

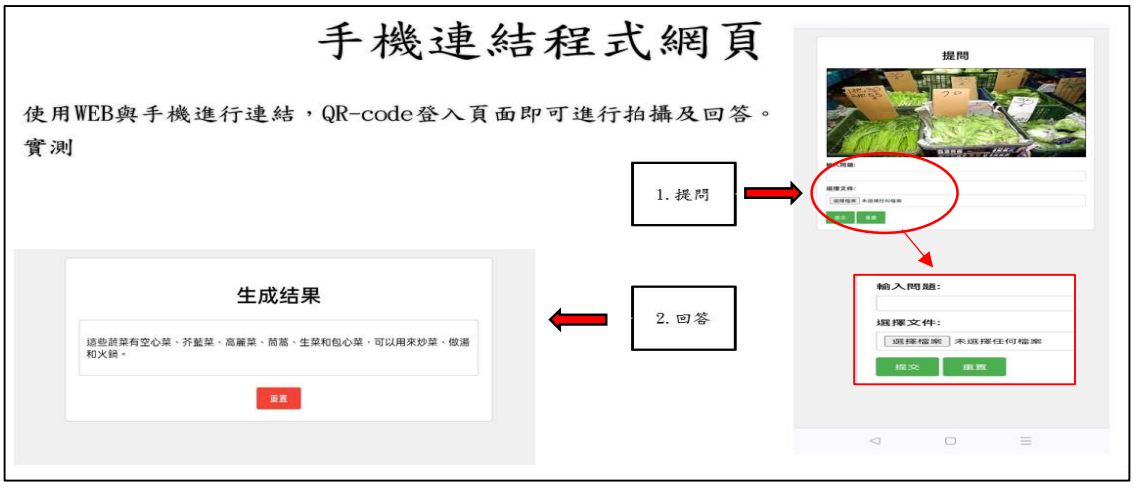

圖 10 手機連結視覺辨識系統

第四章 研究結果

4.1 評估指標

本研究旨在實現在目標檢測模型上套用語言模型，也就是所謂的 Visual Language Model（VLM）模型，效能評估主要採取構建訓練數據集並依照訓練次數儲存相對的精度及召回率以及損失值等等。其次以混淆矩陣圖示例目標檢測模型訓練結果。

關於評估模型的方法，使用精度 (Precision)、召回率 (Recall) 及 F1-分數 (F1-score)，精度的公式如下:

$$\text{Precision}=\frac{True\ Positives}{True\ Positives+Fa\quad Positives} \tag{1}$$

True Positives（TP）是模型正確預測為正例的樣本數。

False Positives（FP）是模型錯誤地將負例預測為正例的樣本數。

精確率的範圍在 0 到 1 之間，值越高表示模型正確地預測正例的能力越強。

$$\text{Recall}=\frac{True\ Positives}{True\ Positives+Fa\quad Negatives} \tag{2}$$

True Positives（TP）是模型正確預測為正例的樣本數。

False Negatives（FN）是模型錯誤地將正例預測為負例的樣本數。

召回率的範圍同樣在 0 到 1 之間，值越高表示模型能夠更全面地捕捉到正例。

$$\text{F1-score}=\frac{2*Precision*Recall}{Precision+R} \tag{3}$$

F1-分數將精準率和召回率結合在一起，通常用於平衡模型的準確性和全面性。

它的取值範圍也在 0 到 1 之間，值越高表示模型的整體性能越好。

$$\text{mAP}=\frac{1}{N}\sum_{i=1}^{N} APi \tag{4}$$

N 是類別的數量，APi第 i 個類別的平均精確度（AP）。

mAP 是根據每個類別的偵測結果計算的，具體步驟包括：

1. 對每個類別的檢測結果依照置信度排序。
2. 計算每個類別的精確率-召回率曲線，並根據 11-point 插值計算曲線下的面積。
3. 將曲線下的面積作為該類別的平均精度（AP）。
4. 將所有類別的 AP 求平均，得到 mAP。

4.1.1 混淆矩陣

混淆矩陣是一種用於評估二元分類模型性能的表格，以四個不同的組合來描述模型的預測結果和實際情況。混淆矩陣包含以下四個元素：

- True Positives (TP)：模型正確預測為正例的樣本數。
- False Positives (FP)：模型錯誤地將負例預測為正例的樣本數。
- True Negatives (TN)：模型正確預測為負例的樣本數。
- False Negatives (FN)：模型錯誤地將正例預測為負例的樣本數。

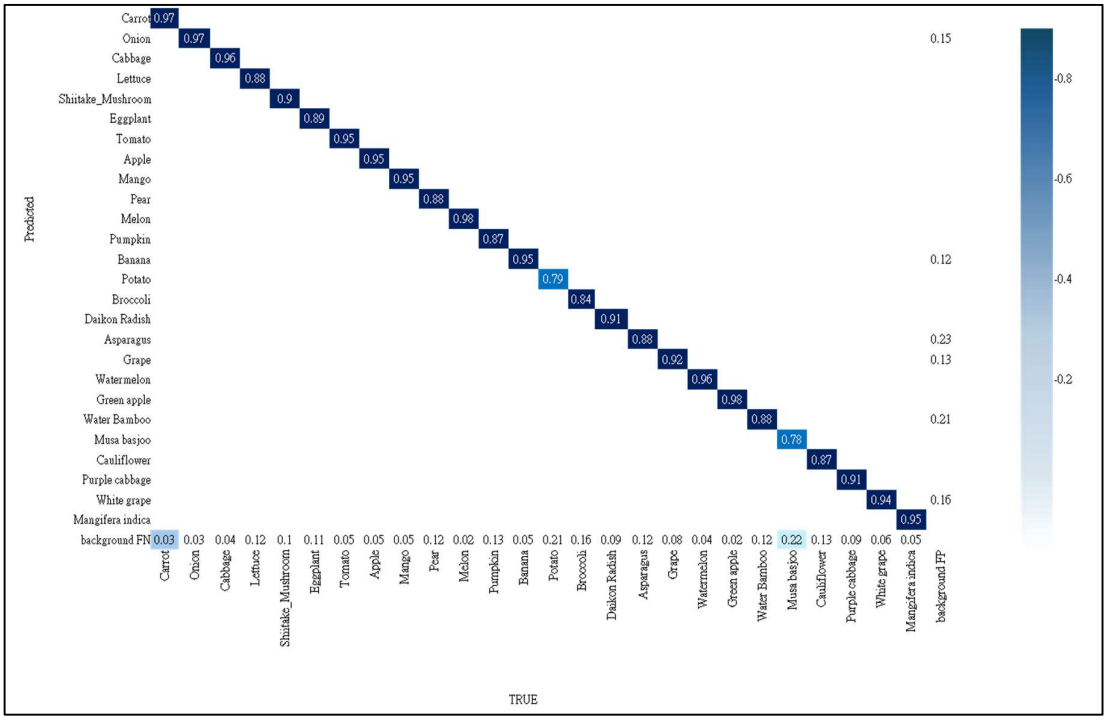

圖 11　混淆矩陣

本研究只為檢測準確為主要，並未應用 TN 元素。下方以 Carrot 為例:

$$\text{Precision}=\frac{0.96}{0.96+0.12}=0.88\approx 88\%$$

$$\text{Recall}=\frac{0.96}{0.96+0.04}=0.96\approx 96\%$$

$$\text{F1-score}=\frac{2*0.88*0.96}{0.96+0.88}=0.92\approx 92\%$$

運用矩陣可以算出 Precision 在 Epoch 為 500 次時是 88%，Recall 則為 96%，最終兩個值的 F1-score 則為 92%。

4.1.2 分析訓練結果

在每個訓練時期所產生的數值依序對比可以觀察到在第 400 次訓練時期產生的 MAP@0.5 數值最高，為 0.9741。連帶 precision、recall、F1-score 所觀察對比的，25 樣蔬菜種類在訓練時應該保持 400 次左右就能差不多，不過增加種類會影響多少的訓練次數這還有待往後訓練才能得知。

訓練次數	precision	recall	F1-score	MAP0.5
100/499	0.9133	0.894	0.90354	0.9391
200/499	0.9361	0.9546	0.94525	0.9672
300/499	0.9215	0.9694	0.94484	0.9707
400/499	0.9414	0.9709	0.95074	0.9741
499/499	0.9385	0.9702	0.95408	0.9727

表 1　訓練結果　　**單位:%**

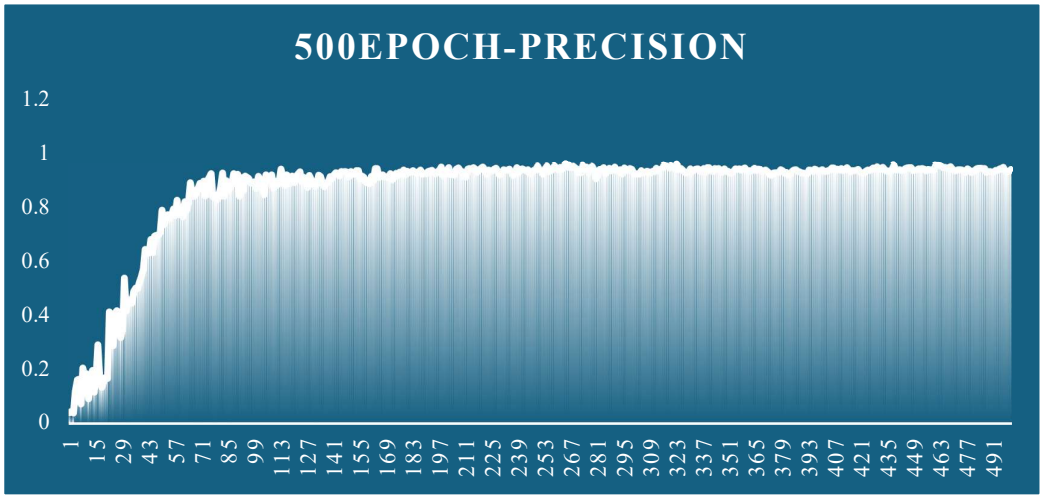

圖 12　500 次訓練-Precision

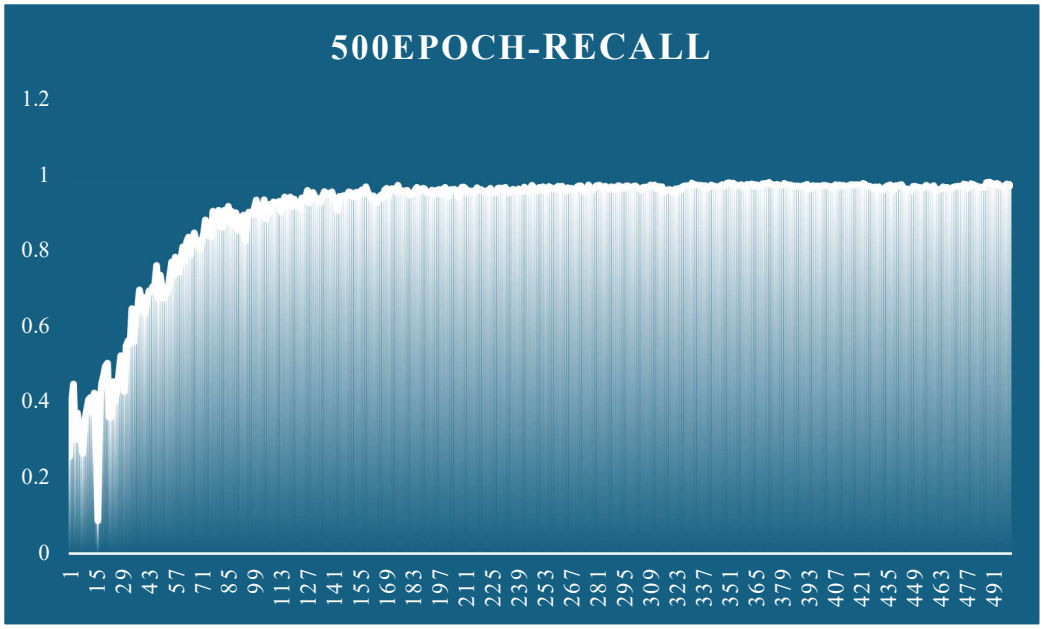

圖 13　500 次訓練-Recall

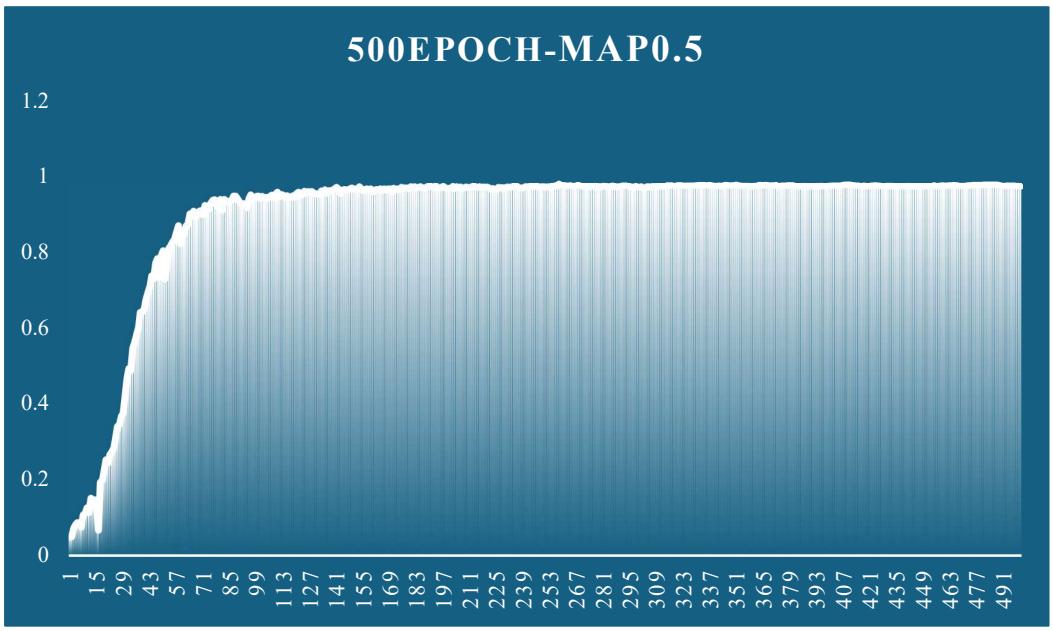

圖 14　500 次訓練-MAP0.5

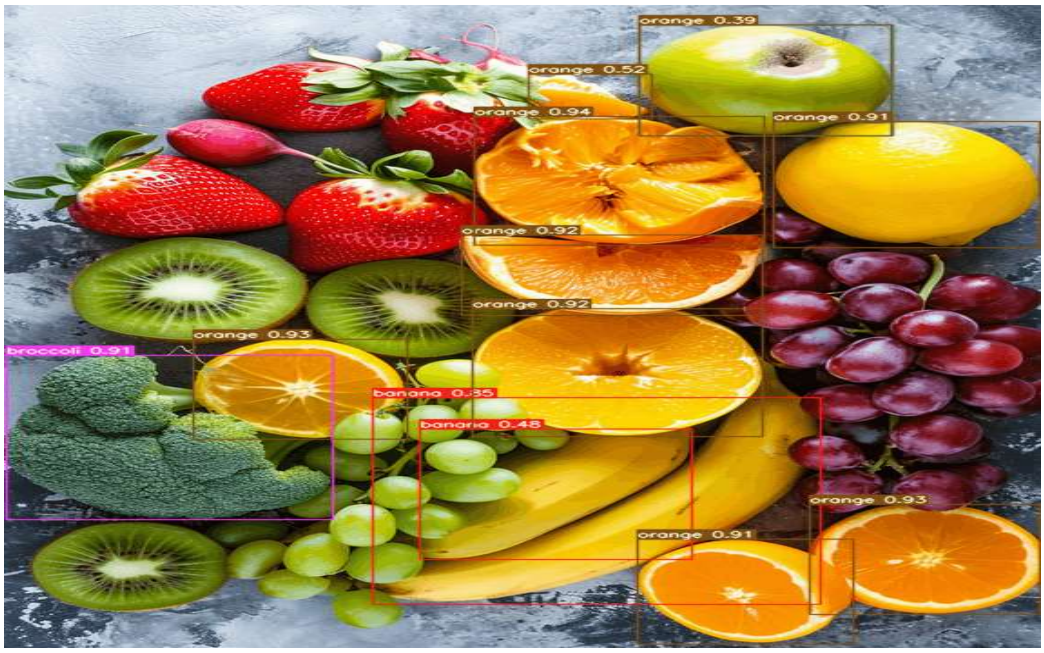

圖 15 預測結果

- 圖 9 可以看到已成功識別了幾種水果，橙子、香蕉和花椰菜。
- 某些物件（如小香蕉）的置信度較低，而其他物件（如蘋果、草莓、奇異果、葡萄）無法識別原因在於還未添加這些蔬果或是置信度過低而被忽略。
- 檢測模型對於有添加訓練過的物件似乎相當準確，但某些邊界框（例如香蕉的邊界框）的置信度分數較低，這可能表明需要進一步訓練提高迭代次數或是調整訓練參數。

4.1.3 物件遮擋訓練

把每個蔬果的訓練數據都套上類似塑膠袋裝這樣的遮擋效果再次進行訓練，在每個訓練時期所產生的數值依序對比可以觀察到在第 500 次訓練時期產生的 MAP@0.5 數值最高，為 0.9706。連帶 precision、recall、F1-score 所觀察對比的，19 樣蔬菜種類在訓練時依照現階段能夠判斷有遮擋的影像應該能夠增加更多訓練次數，好以判斷最終效果落在何處。

訓練次數	precision	recall	F1-score	MAP0.5
100/499	0.8233	0.8114	0.899099	0.9405
200/499	0.8461	0.852	0.941046	0.967
300/499	0.8315	0.8748	0.940751	0.9631
400/499	0.8485	0.7585	0.951846	0.9694
499/499	0.8414	0.8805	0.947652	0.9706

表 2　遮擋物件訓練結果　　**單位:%**

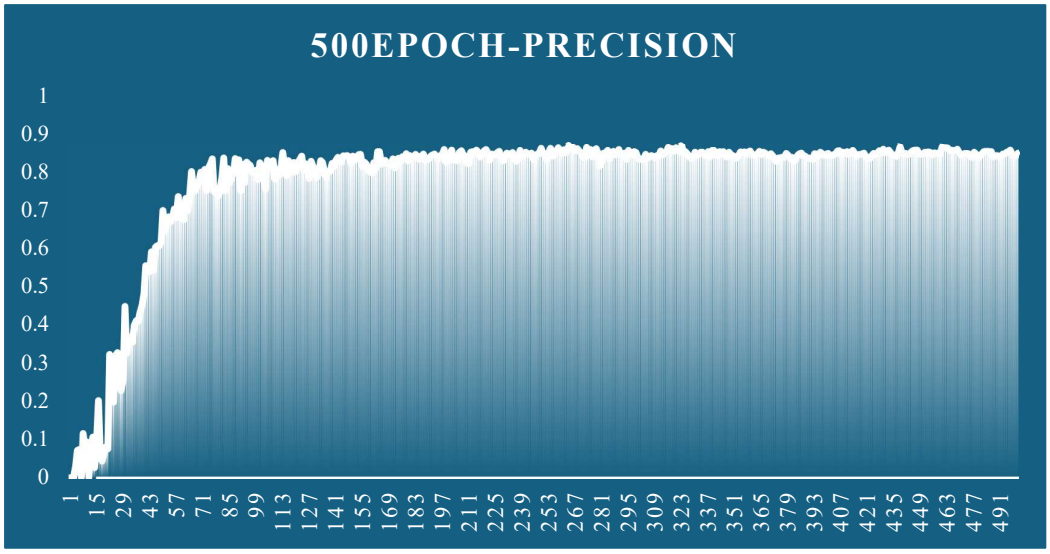

圖 16　500 次訓練-Precision

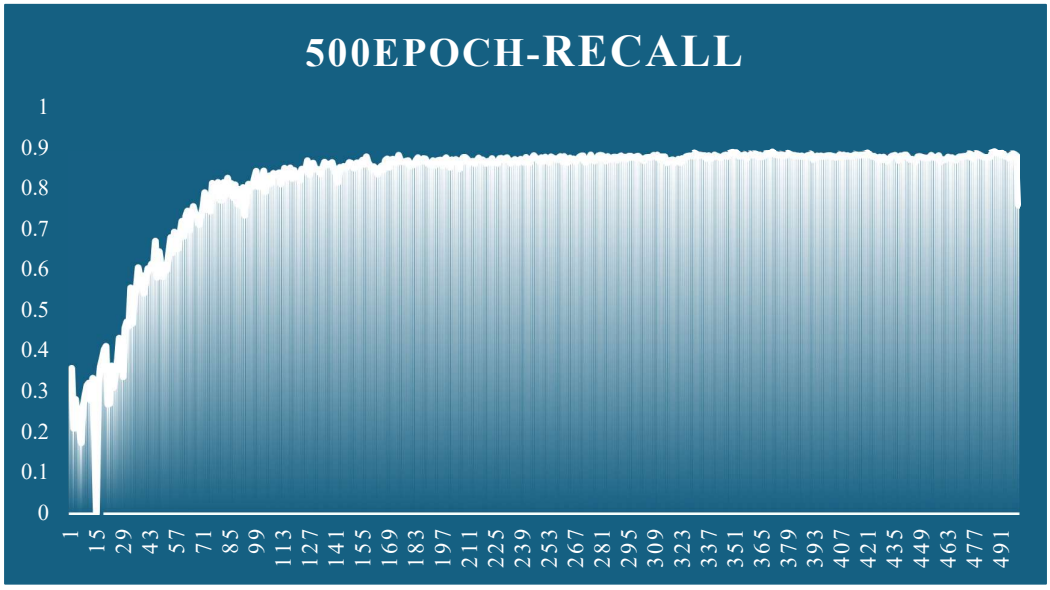

圖 17　500 次訓練-Recall

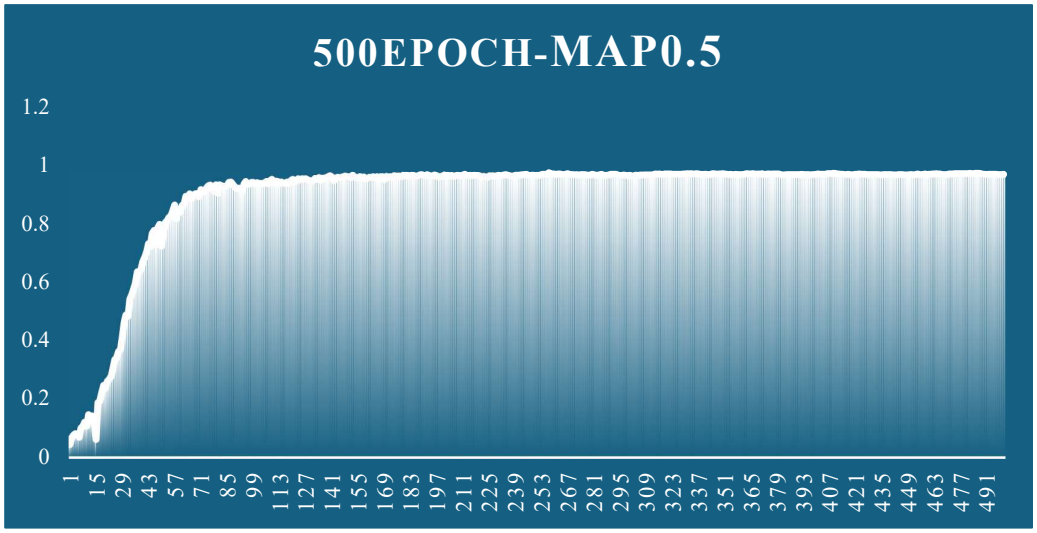

圖 18　500 次訓練-MAP 0.5

圖 19 有無針對遮擋訓練之差異

圖 19 中的左側和右側顯示了對蘋果進行目標檢測的兩種情境。

- 左圖標註為「未針對遮擋物進行訓練」，置信度為 0.77。在未經過遮擋訓練的情況下，對於偵測到的蘋果有相對較低的信心。
- 右圖標註為「針對遮擋物進行訓練」，置信度提升至 0.89。這表示模型在經過遮擋物的專門訓練後確實是有效的。

第五章 結論與建議

5.1 結論

本研究旨在實現在目標檢測模型上套用語言模型，並提出了一種稱為 VLM 的新方法。通過將 YOLOv7 模型與 GPT-3.5 模型整合，我們成功地將圖像檢測的結果轉換為自然語言描述或對象代碼，從而實現了圖像檢測和自然語言解釋的一體化。該系統不僅在圖像檢測的準確性上表現出色，而且能夠提供自然語言解釋，使用戶能夠更容易地理解和使用系統的輸出。此外，我們還實現了持續對話功能，使系統能夠根據用戶的歷史輸入提供連續的對話回應，進一步提高了系統的實用性和用戶體驗。

5.2 未來建議

未來的研究可以進一步改進本研究中使用的模型和方法，以提高系統的性能和效率。例如，可以探索更先進的目標檢測模型，如 YOLOv8 或其他基於深度學習的模型，以提高圖像檢測的準確性和速度。同時，可以考慮使用更大的語言模型或結合多個語言模型，以提高自然語言解釋的質量和多樣性。此外，還可以進一步優化系統的用戶界面和互動方式，以提高系統的易用性和可操作性，並擴展系統的應用範圍和場景。例如以下:

數據擴充： 在訓練目標檢測模型時，可以採取數據擴充技術來增加訓練數據的多樣性，從而提高模型的泛化能力和魯棒性。

多模態融合： 考慮引入多模態融合的方法，將圖像信息和文本信息有效地融合在一起，以提高系統的綜合理解能力和表現。

用戶反饋收集： 收集用戶的反饋意見和需求，並根據用戶的實際使用情況進行系統的持續改進和優化。

參考文獻

[1.] O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. arXiv:1511.08458 [cs.NE], November 26, 2015 (v1), revised December 2, 2015 (v2),4-5. Retrieved from https://doi.org/10.48550/arXiv.1511.08458.

[2.] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[3.] Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 1440-1448).

[4.] Here's the citation for the paper "SSD: Single Shot MultiBox Detector":Liu, W., et al. (2016). SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham.

.

[5.] Sabiri, B., EL Asri, B., & Rhanoui, M. (2023). Efficient Deep Neural Network Training Techniques for Overfitting Avoidance. In B. Sabiri, B. EL Asri, & M. Rhanoui (Eds.), Enterprise Information Systems (ICEIS 2022), Lecture Notes in Business Information Processing (Vol. 487, pp. 198–221).

[6.] Shah, U., Khan, S., Alzubaidi, M., Agus, M., & Househ, M. (2024). Unveiling the potential of ChatGPT and YOLOv7 for evaluating children's emotions using their artistic expressions. *Studies in Health Technology and Informatics*, *316*, 409–413. https://doi.org/10.3233/SHTI240434

[7.] Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV], October 5, 2023. https://doi.org/10.48550/arXiv.2310.03744

[8.] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

[9.] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable

bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696 [cs.CV], July 6, 2022.

[10.] Telus International. (2021, January 1). An Introduction to 5 Types of Image Annotation. Retrieved from https://www.telusinternational.com/insights/ai-data/article/an-introduction-to-5-types-of-image-annotation.

[11.] Neubeck, A., & Van Gool, L. (2006, August). Efficient non-maximum suppression. In 18th international conference on pattern recognition (ICPR'06) (Vol. 3, pp. 850-855). IEEE.

[12.] Wang, Y., Zou, X., Shi, J., & Liu, M. (2024). YOLOv5-Based Dense Small Target Detection Algorithm for Aerial Images Using DIOU-NMS. RADIOENGINEERING, 33(1), 12-22.

[13.] OpenAI. (n.d.). Tokenizer. OpenAI. Retrieved May 26, 2024, from https://platform.openai.com/tokenizer.

專題團隊工作分配一覽

工作分配

姓名 / 工作分配	施坤政	陳家翔	謝咏倫	蔡景州	林聖祐
收集相關資訊及文獻					
訓練集製作集修改					
Yolov7資料增強					
ChatGPT API 網頁連結手機					
論文及報告撰寫					