台 南 應 用 科 技 大 學

資訊管理系 日四技

專題製作

論文題目:

結合 AI 技術與電腦視覺的即時蔬果辨識系統

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 114 年 4 月

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蔬果辨識系統

組長：施坤政

組員：林聖祐　陳家翔

蔡景州　謝咏倫

指導老師：_____老師

主　　任：_____主任

中華民國 114 年 4 月

致謝辭

回顧本論文的撰寫歷程，深感收穫良多。首先，誠摯感謝指導老師張至豪老師在研究過程中所提供的細心指導與寶貴建議，幫助我們不斷修正思路、深入探討問題，並提升專題報告與論文的品質。張老師的專業素養與耐心指導，使我們在研究過程中克服重重困難，順利完成本論文。也感謝在研究過程中提供支持與協助的同儕與成員，讓我們能夠專注於研究與撰寫，順利完成此項研究成果。再次向所有在此過程中提供幫助與鼓勵的人士，致上最誠摯的感謝。

摘要

本研究開發一套結合 YOLOv7 與 GPT-3.5 的視覺語言互動模型（Vision-Language Interaction Model），旨在透過即時攝影畫面辨識蔬果，並依據使用者提問提供語意回應。研究動機源於現有多數人工智慧系統偏重文字輸入，缺乏與實體影像的整合能力，導致其無法有效應用於食材辨識、購物輔助與知識學習等日常場景的實際需求。為彌補此視覺與語言整合的缺口，本研究提出一套「能看會說」的 AI 解決方案，讓使用者可透過手機拍攝蔬果畫面，系統即時辨識其種類，並回應相關問題，以此提升 AI 系統對實際場景的應用價值。

在研究方法上，系統採用 YOLOv7 進行目標偵測，該模型以其優異的辨識速度與準確性被選為核心組件，並輔以 GPT-3.5 API 實現自然語言理解與生成，所有模型均以人工標註之蔬果影像資料進行訓練。本系統的目標是讓使用者透過手機畫面即時拍攝蔬果，系統能立即辨識其種類，並根據使用者提問給予相關資訊回答。整合視覺與語言能力的互動流程，將提升 AI 系統對實際場景的應用價值。

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

Abstract

Image recognition is a key topic in artificial intelligence development, especially

critical in real-time applications. YOLOv7 (You Only Look Once v7), released in 2022, offers high speed and accuracy, outperforming most object detection models. In the same year, OpenAI launched ChatGPT, a powerful language model demonstrating strong semantic understanding and generation capabilities, driving the adoption of large language models (LLMs) across various fields.

This study develops an integrated Vision-Language Model (VLM) system combining YOLOv7 and GPT-3.5. The system can recognize fruits and vegetables through a camera and respond to user queries in real time. It supports mobile operation, enabling instant recognition and immediate interaction, enhancing both efficiency and usability.

Keywords: Real-time visual recognition, Natural language processing, Artificial intelligence applications, GPT-3.5, YOLOv7

目錄

圖目錄

表目錄

第一章 緒論

一、研究動機

隨著生成式 AI 技術快速發展，ChatGPT 等語言模型逐漸普及，顯示出 AI 在理解與回應語言上的潛力。然而，現有系統多聚焦於文字輸入，缺乏與現實世界畫面結合的能力。在實際生活中，辨識食材、協助選購、學習蔬果知識等需求，需要搭配能夠即時處理大量影像資訊的辨識系統才能達到即時讓語言模型知曉我們對於影像需求的提問，並準確回答我們的問題。因此，我們希望設計一套系統，能即時辨識攝影機畫面中的物體，並透過語言模型提供有意義的回應，彌補現況來說 AI 缺乏視覺與語言整合的不足，提升 AI 與使用者互動的實用性與智慧程度。

二、研究目的

本研究旨在解決現有 AI 系統無法同時即時辨識影像並語意回答使用者問題的缺口，針對日常生活中如蔬果辨識與資訊查詢等需求，建立一套「能看、能說」的即時互動系統。

~~本系統的目標是讓使用者透過手機畫面即時拍攝蔬果，系統能立即辨識其種類，並根據使用者提問給予相關資訊回答。整合視覺與語言能力的互動流程，將提升 AI 系統對實際場景的應用價值。(放至摘要與結論)~~

三、本文架構

本論文共分為五章。

第二章為文獻探討，整理近年視覺語言模型（VLM）、目標檢測與自然語言處理的相關研究，並說明本研究的理論基礎與參考模型。

第三章為研究方法，介紹系統整體架構與流程，包含 YOLOv7 與 GPT-3.5 的應用方式、資料來源與標註方式，以及系統整合與手機端應用設計。

第四章為研究結果，呈現系統實作成果與各項指標評估，包含模型準確度、遮擋處理效果與使用者互動成果。

第五章為結論與建議，總結本研究的成果與限制，並提出未來可改進方向與應用發展建議。

第二章 文獻探討

近年來，全球科技及人工智慧領域的學術界與組織，對於視覺語言模型（Vision-Language Models, VLM）的討論與研究呈現蓬勃發展的態勢。顧名思義，VLM 便是結合了視覺辨識模型與自然語言模型，如同本研究採用的 YOLOv7 與 Chat GPT。自從自然語言模型嶄露頭角以來，各大研究機構如同，Microsoft 與 GOOGLE 便積極探索能與之結合應用的程式工具，以提升其實用性與便利性，而 VLM 正是其中備受矚目的關鍵技術之一。VLM 能夠整合不同的視覺模型，理解圖像中豐富的特徵內容，並根據不同的指令執行多樣化的視覺與語言相關任務，例如物體識別、特定區域分割、回答圖像相關問題，甚至能透過少量範例進行學習。以圖 1 中的蘋果辨識為例，模型不僅能精確定位圖像中的蘋果，更能根據輸入的描述（如「亮紅色的蘋果」）進行細緻的分割，回答關於蘋果種類的疑問（如「這是什麼種類的蘋果？」），並能基於少量的範例（如關於富士蘋果特徵的描述）辨識出蘋果的種類。**考量到 YOLOv7 在物體識別速度上的卓越表現，非常適合作為本研究即時辨識視訊系統的核心組件。**

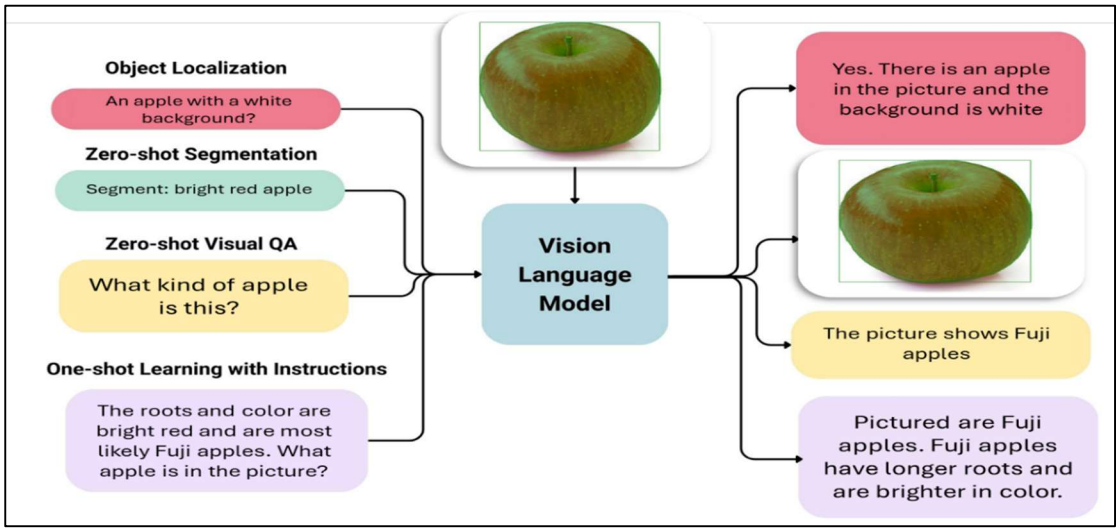

圖 1　VLM 模型的處理流程

為了實現 VLM 快速識別物體的目標，選用了在即時識別方面表現優異的 YOLOv7 作為視覺模型；語言模型則採用了易於連接且詞彙量豐富的 GPT。

一、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）[1]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。

(這句話是否為抄襲)YOLOv7 (截止自 2022 年 12 月)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。

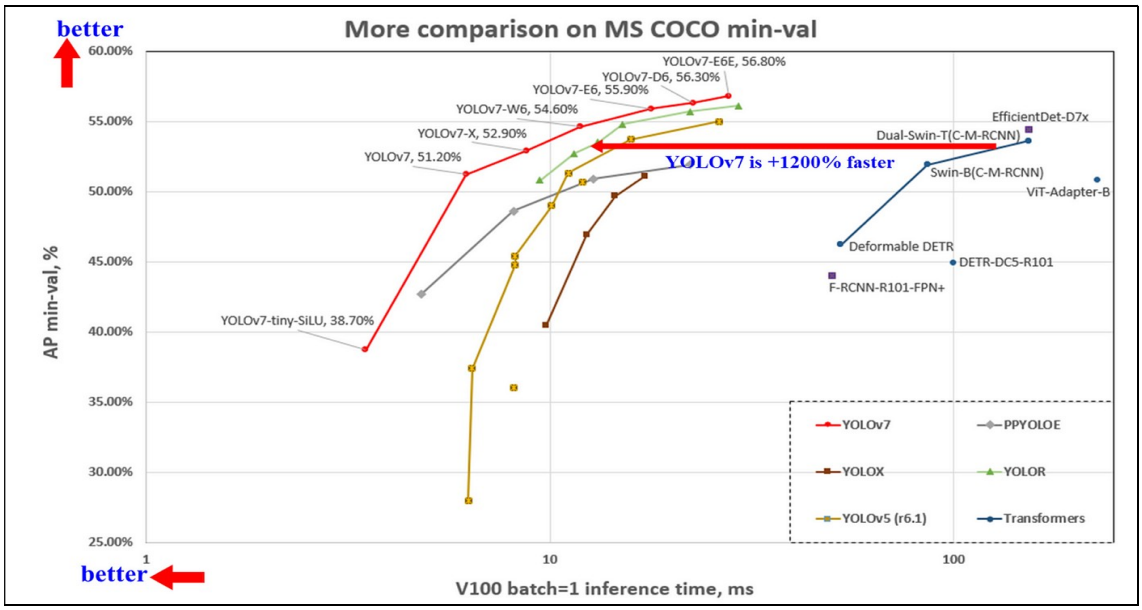

圖 2　YOLOv7 測試比較圖

二、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。

(一) 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[2]。

(二) 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。

思考如何做出程式雛形時，我們參考幾個類似的論文，可用於學習及效仿，例如 Uzair SHAH 等人的實驗與我們做的實驗相關性高，論文主要在視訊方面如何完整提供詳細資料給 GPT 做了詳細解說；

Uzair SHAH 等人[3] 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。

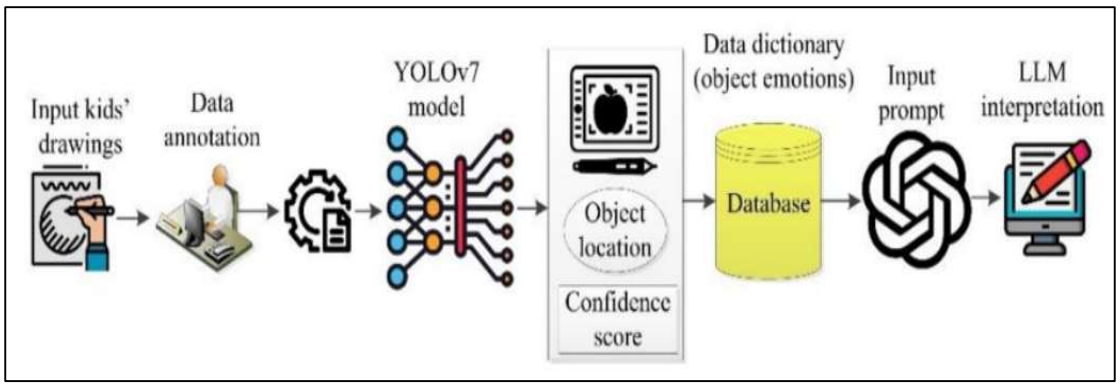

圖 3　情緒分析實驗架構圖

李等人[4] 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。

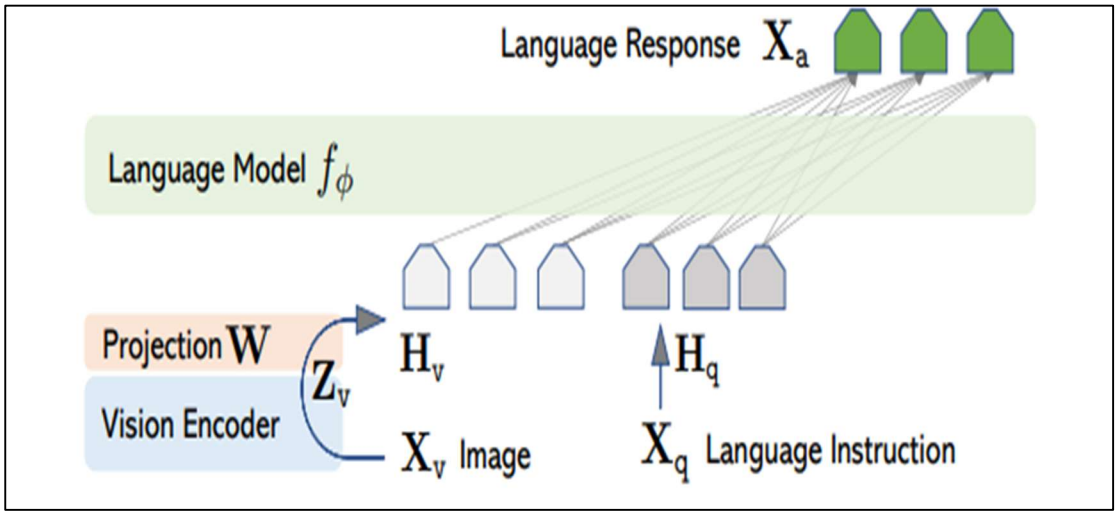

圖 4　LLaVA 架構圖

蘇等人[5]採用 Transformer 模型作為 VL-BERT（Visual-Linguistic 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（Masked Language Model）的方式，透過預測隨機屏蔽的單字或感興趣區域來提高模型對視覺-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。

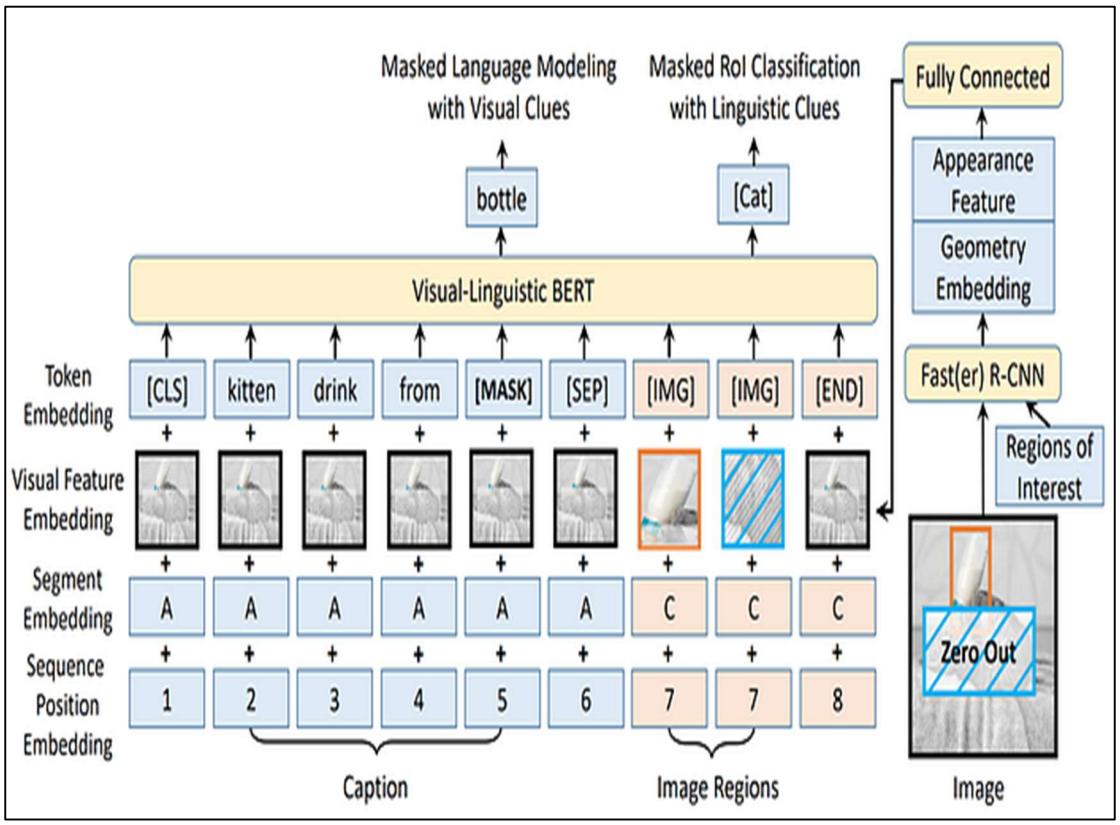

圖 5　VL-BERT 架構圖

第三章 研究方法

一、研究架構

本研究旨在建構一套結合即時目標檢測與自然語言回應的視覺語言模型（VLM）系統，著重於即時性與準確率的平衡。系統整合 YOLOv7 進行蔬果辨識，並結合 GPT-3.5 回答使用者問題，整體架構如圖 6 所示，涵蓋前端影像擷取、後端推論處理與語意生成三大模組。

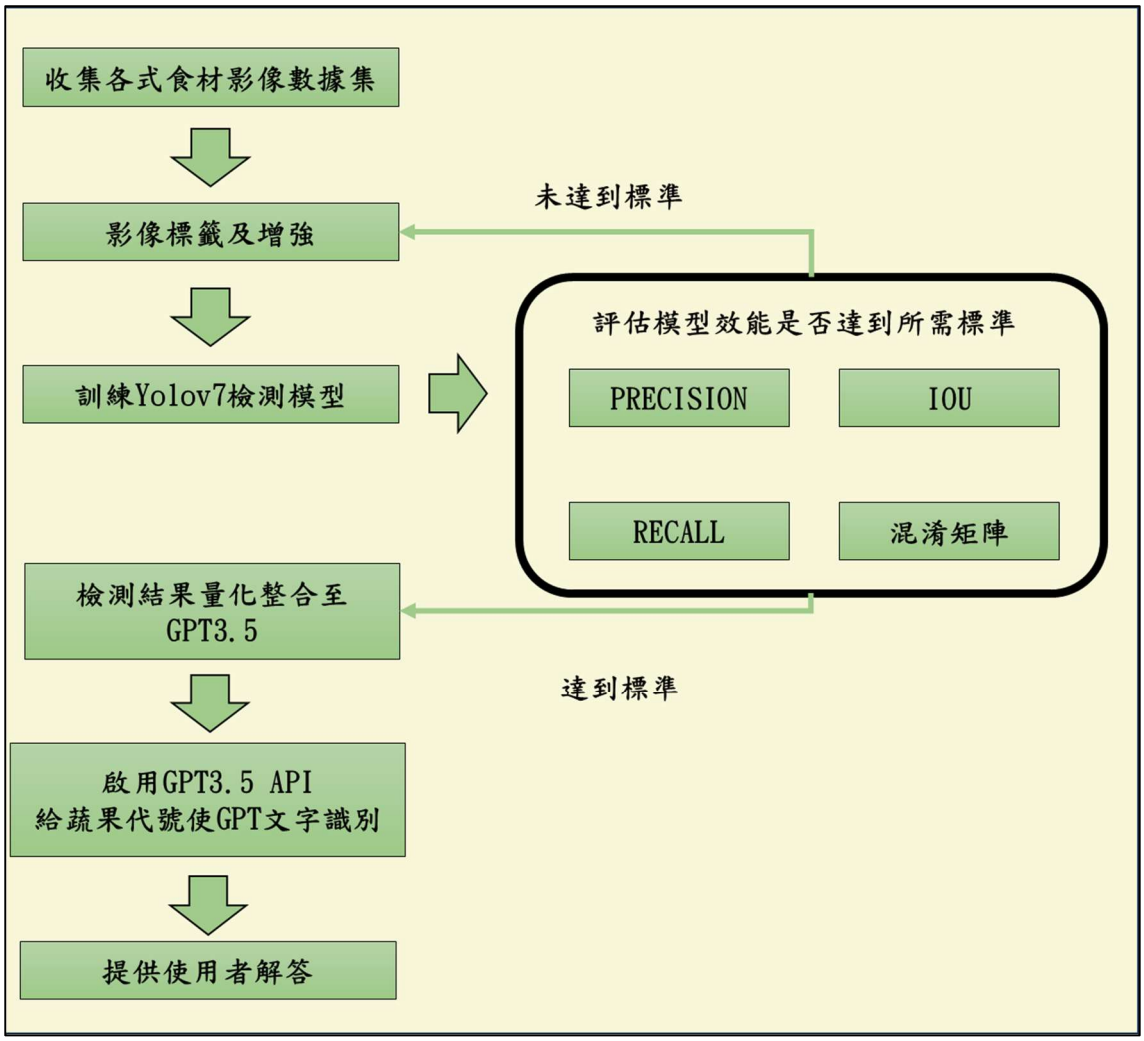

圖 6 研究整體流程架構圖

研究流程三大步驟

本研究使用 Labelimg 工具對蒐集到的蔬果影像資料進行標註 。Labelimg 是一款圖形化影像標註工具，支援多種標註類型，其中本研究針對 YOLOv7 模型採用了邊界框（Bounding boxes）的標註方式 。邊界框標註通常用於物體檢測和識別任務 。透過 Labelimg，研究團隊在每張蔬果圖片上繪製矩形框，精確框選出蔬果的位置，並為每個框指定對應的類別名稱（如「蘋果」、「香蕉」等） 。這些標註資訊會被儲存為 YOLOv7 模型所需的特定格式檔案（通常是 TXT 格式），其中包含類別編號、物體中心點的相對座標以及邊界框的寬高比例 。訓練完成後，透過 Precision、Recall、F1-score 等指標評估模型偵測效能 。圖 7 顯示了 YOLOv7 的訓練流程，從原始圖片開始，經過圖片標

籤、訓練，最終進行效能評估 。

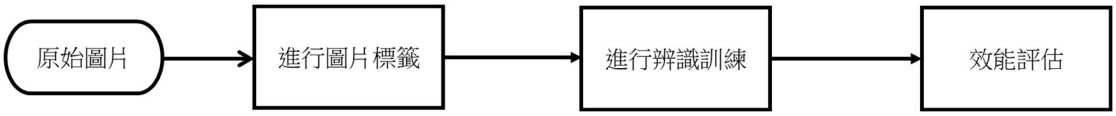

圖 7　YOLOv7 訓練流程

在 OpenAI 平台申請並啟用了 GPT-3.5 的應用程式介面 (API)。為了讓 GPT-3.5 能夠精確理解 YOLOv7 辨識出的蔬果種類，並提供相關的語意回應，模型需要進行微調 (Fine-tuning)。微調的目的是讓 GPT-3.5 學習蔬果名稱與其對應數字代碼之間的映射關係，例如將「蘋果」與其在 YOLOv7 中定義的特定數字 ID（如 0）進行關聯。透過提供大量的範例，讓 GPT-3.5 了解當接收到特定數字代碼時，應該如何將其解讀為對應的蔬果類別，並準備生成相關的語意內容。

此階段將微調後的 GPT-3.5 API 整合到 YOLOv7 的程式架構中。這意味著在 YOLOv7 完成目標檢測並輸出蔬果類別及置信度後，這些資訊會立即作為輸入傳遞給 GPT-3.5 API。這種整合確保了視覺辨識與自然語言理解與生成之間的同步運行，實現了即時的視覺語言互動。圖 8 所示的流程圖描述了此整合過程，從獲取 GPT-3.5 API 到數位化蔬果代表編號，最終將 GPT 程式碼嵌入 YOLOv7 程式碼內。

圖 8 GPT 微調方法

YOLOv7 完成影像中的蔬果辨識後，會輸出每個辨識到的物體的類別名稱（例如：「蘋果」、「香蕉」）以及其相對應的數字編號（例如：「蘋果」可能對應編號 0，「香蕉」對應編號 1）。這些類別名稱和編號會作為結構化的資料傳遞給 GPT-3.5。GPT-3.5 接收到這些資訊後，會進行解析，將數字編號轉換回對應的蔬果名稱，並理解這些辨識結果所代表的語意內容。根據 YOLOv7 辨識出的蔬果種類（例如：「蘋果」），GPT-3.5 會結合其強大的自

然語言生成能力，產生與該蔬果相關的自然語言回答。這些回答不僅限於簡單的名稱確認，還可以根據預先設定的指令或其內建知識，提供更多詳細資訊，例如：蔬果的營養價值、食用方法、產地、保存方式等。例如，如果 YOLOv7 辨識出「蘋果」，GPT-3.5 可能會生成「您看到的是蘋果，它富含維生素 C 和膳食纖維，有助於消化。」這樣的回答，並以文字串的形式輸出，即時回饋給使用者。整個過程旨在實現「能看會說」的互動體驗，讓使用者透過視覺輸入獲得語意豐富的回應

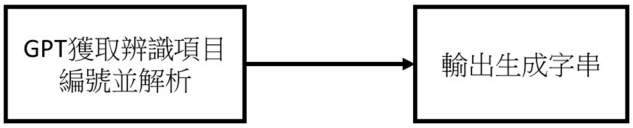

圖 9 GPT 輸出結果

二、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。

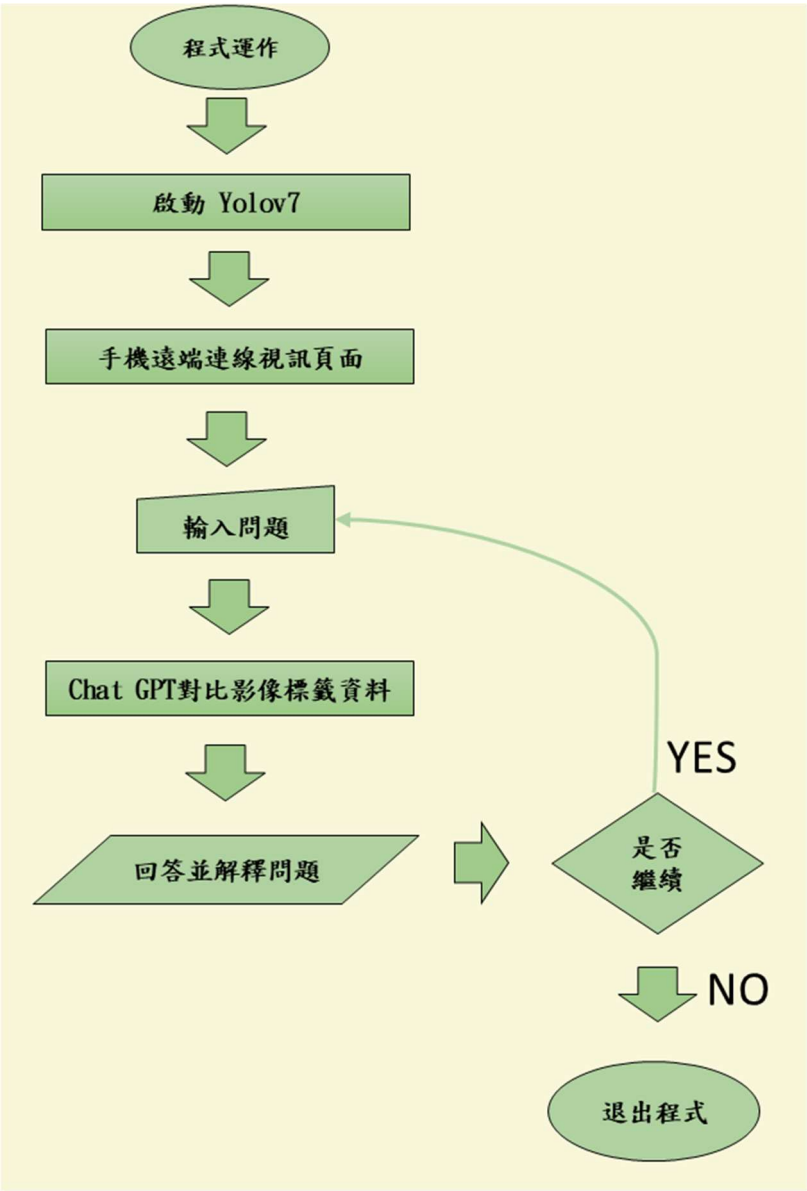

圖 10 系統流程圖(系統流程改為 3 章的末端)

(一) 數據集製作

訓練集使用 Labelimg 這個軟體進行圖片標註，Labelimg 可進行多樣標註，其中，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)[6]。在本研究 YOLOv7 所採用的是邊界框(Bounding boxes)如圖 X，通常用於物體檢測和識別。

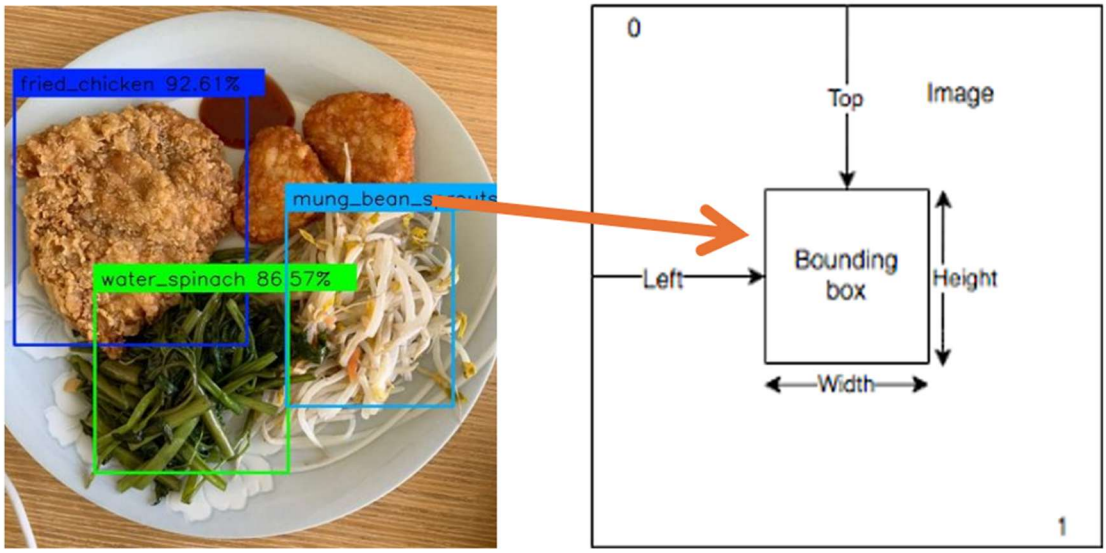

圖 11　邊界框示例圖

本研究總共採用 26 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、紅棗、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]。

我們找的蔬果訓練圖數據來源有從網路下載，也有自己去菜市場拍，每個蔬果個別有 20 張照片，蔬果照比例如下表 1。

資料來源	**張數**
網路下載	364
實體市場拍攝	156
合計	520

表 1 YOLOv7 訓練集影像來源

為提高模型訓練效果並強化模型的泛化能力，本研究對原始影像資料進行了資料擴增（Data Augmentation）處理，包含：

- 隨機裁切（Random Cropping）
- 水平翻轉（Horizontal Flipping）
- 亮度調整（Brightness Adjustment）
- 輕微旋轉（Rotation）

經由擴增後，總圖像數量增至原始圖數之兩倍，共計 1040 張。

(二) YOLOv7 訓練

本研究採用 YOLOv7 作為目標檢測模型，因其具備優異的辨識速度與準確性，適合應用於即時蔬果識別系統中。

使用 Labelimg 工具對蔬果圖像進行邊界框標註生成文字檔，內含類別編號與相對座標（中心點與寬高比例）。

依照 YOLOv7 所需結構建立資料夾階層，使用 train 腳本啟動模型，結構設定參考 yolov7.yaml 原始檔，訓練參數如表 2 所列，使用 500 次訓練週期（Epochs），Batch size 為 16，圖片尺寸 640×640，並使用 GPU 進行加速訓練。

參數項目	設定值
Epochs	**500**
Batch Size	**16**
圖像尺寸	**640 x 640**
模型結構檔案	**yolov7.yaml**
儲存路徑	**runs/train/exp/weights**
最終輸出模型	**best.pt**

表 2 訓練參數

(三) 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）[7]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，採用了改進的 NMS 演算法。 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。

(四) 視訊傳輸至 GPT 做法(放到 3 章的尾端，跟流程一起)

經過目標辨識與語意轉換，最終回傳自然語言回答給使用者。各階段模組功能與技術對應如表 3 所示。

1.前端影像擷取

使用 HTML5 與 JavaScript 建置網頁介面，透過裝置相機擷取即時畫面，並將影像資料以 JPEG 格式傳輸至後端伺服器。

2.目標偵測

後端系統接收影像後，利用 YOLOv7 模型結合 OpenCV 進行即時目標辨識，提取圖像中各物體的類別名稱及邊界框座標資訊。

3.資訊轉換

將 YOLOv7 輸出的偵測結果以 Python 程式處理，轉換為適合輸入 GPT-3.5 的自然語言提示（Prompt），例如看到蘋果則輸出「蘋果」至 GPT。

4.回應產生

將生成的 Prompt 傳送至 GPT-3.5 API，經語意分析後，輸出自然語言回答，提供使用者更具體且具理解性的回應。

模組階段	處理項目	技術工具	資料格式
前端影像擷取	相機擷取與畫面顯示	HTML5 / JavaScript	JPEG / Base64
目標偵測	即時物件辨識	YOLOv7 / OpenCV	類別名稱、框座標
資訊轉換	偵測結果轉換成敘述文字	Python	Prompt (自然語言)
回應產生	回傳對象語意回答	GPT-3.5 API	自然語言輸出

表 3 影像傳輸方法

系統流程

視訊傳輸至 GPT 做法

三、系統整合(放到研究結果)

本研究整合前端使用介面、後端伺服邏輯及深度學習模型，建置一套即時互動的視覺語言辨識系統。整體架構分為三大模組：前端介面、後端服務與 AI 模型模組，如**圖 12** 所示。

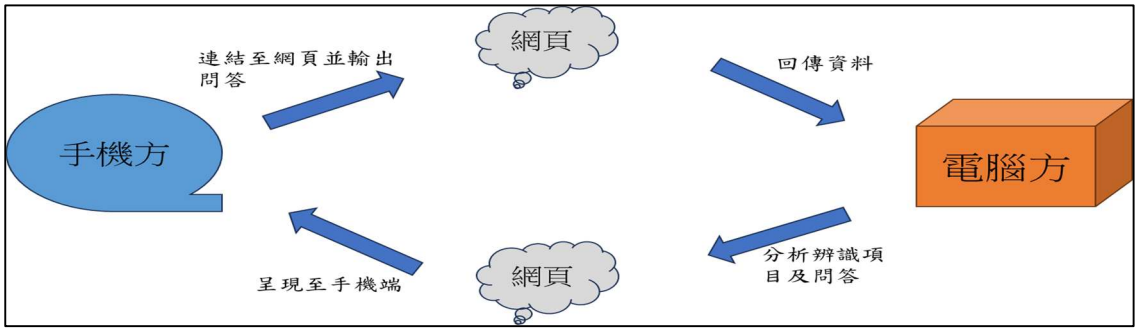

圖 12 連結流程示意圖

1.前端介面

以 HTML5 和 JavaScript 建構網頁平台，提供影像擷取、問題輸入及回應顯示功能，支援桌機與手機瀏覽器操作。

2.後端服務

使用 Python Flask 框架作為伺服核心，負責接收前端影像與輸入資料，並將資訊傳遞給 YOLOv7 與 GPT-3.5 模型進行處理。

3.AI 模型模組

整合 YOLOv7 進行目標偵測，並以 GPT-3.5 進行語意生成，實現影像辨識與自然語言回答的串接功能。

4.跨網區支援

為實現跨網區（Cross-Network）連接與行動裝置即時互動，本系統採用 ngrok 工具建立安全的 HTTP 隧道，公開網址轉換為 QR Code，授權啟用手機相機後，畫面即時傳輸至後端，使行動裝置能透過外網直接連接

至本地端伺服器。

第四章 研究結果

本研究旨在於目標檢測模型中結合語言模型，形成視覺語言模型（Visual Language Model, VLM），並針對物體辨識任務進行效能評估。研究過程中，我們以 YOLOv7 模型為基礎進行改良與訓練，並記錄不同訓練階段下的精度、召回率、F1-score 以及損失值指標。為進一步了解模型預測表現，亦繪製混淆矩陣作為輔助分析工具。

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:

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

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。

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

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。

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

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）。

AP 是根據每個類別的偵測結果計算的，具體步驟包括：

對每個類別的檢測結果依照置信度排序。

計算每個類別的精確率-召回率曲線，並根據 11-point 插值或其他方法計算曲線下的面積。

將曲線下的面積作為該類別的平均精度（AP）。

將所有類別的 AP 求平均，得到 mAP。

一、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。

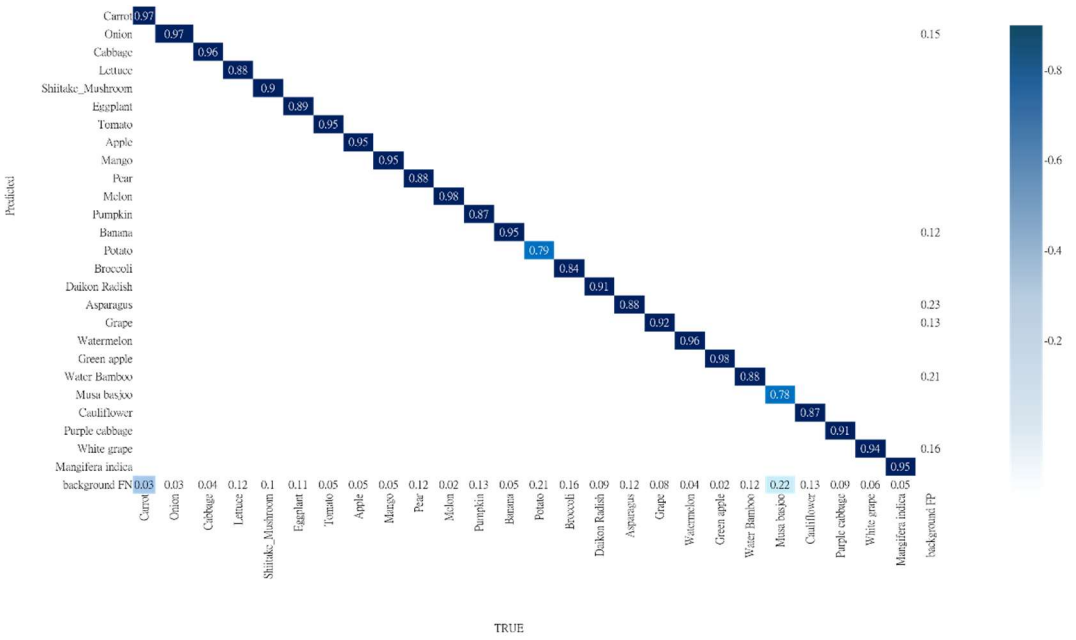

圖 13　混淆矩陣圖

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:

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

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

$$\text{F1-score}=\frac{2*0.88*0.96}{0.96+\ .88}=0.92\approx 92\%$$

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%

二、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，26 樣蔬果種類在訓練時應該保持 400 次左右就能差不多。

訓練次數	precision	recall	F1-score	MAP0.5

100/499	0.9133	0.894	0.903547	0.9391
200/499	0.9361	0.9546	0.945259	0.9672
300/499	0.9215	0.9694	0.944843	0.9707
400/499	0.9314	0.9709	0.95074	0.9741
499/499	0.9385	0.9702	0.954087	0.9727

表 4　訓練結果

下圖顯示模型於前 100 次訓練過程中的變化，特別聚焦於指標波動較大的區段。

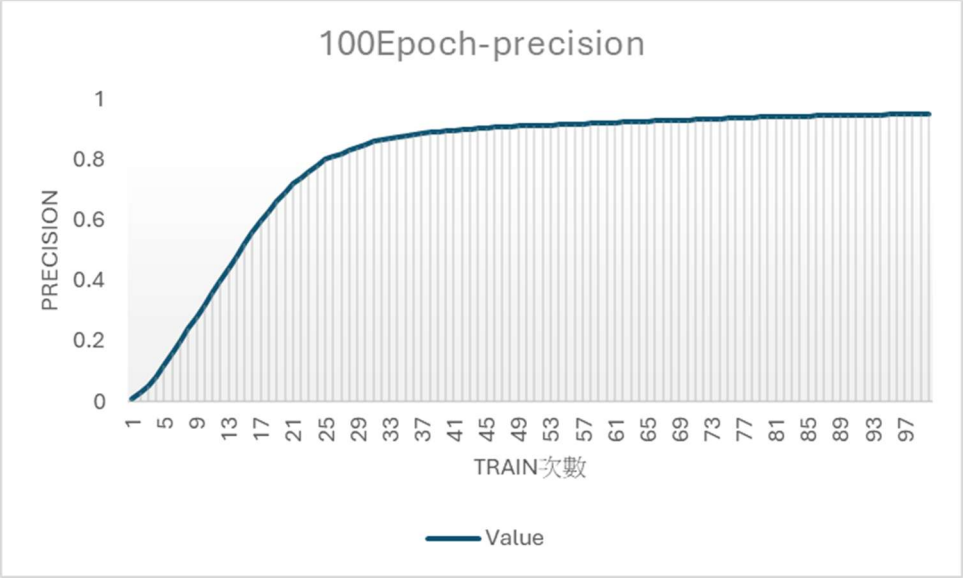

圖 14　100 次訓練-Percision 圖

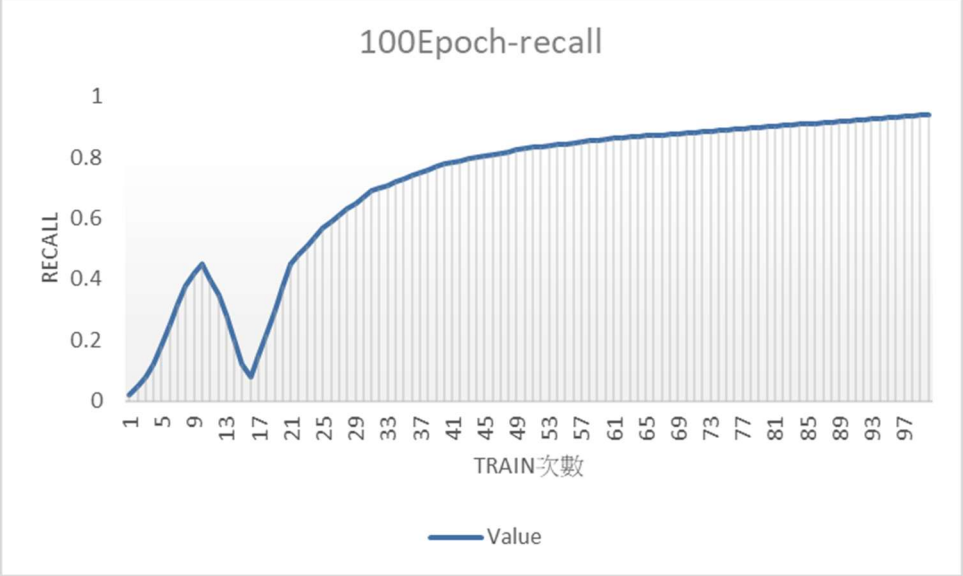

圖 15　100 次訓練-Recall 圖

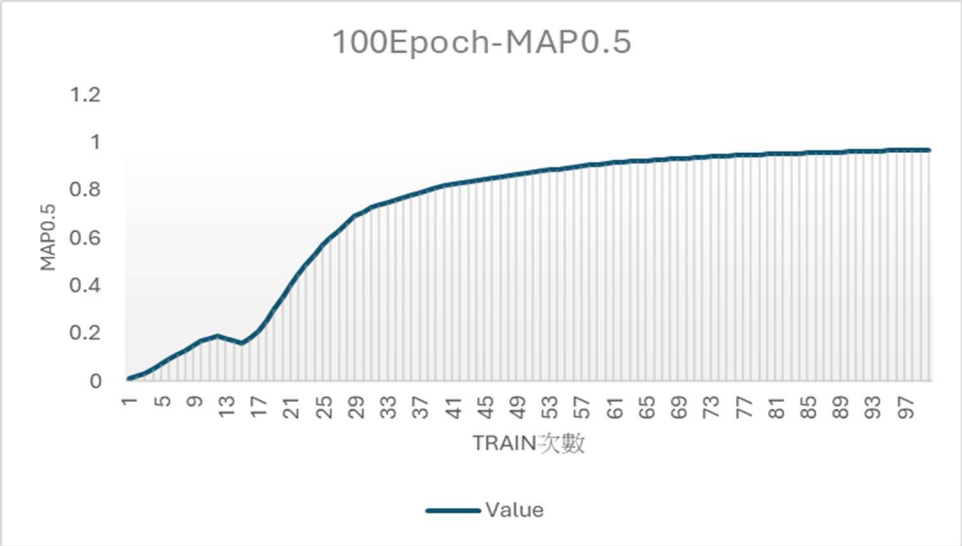

圖 16　100 次訓練-MAP0.5 圖

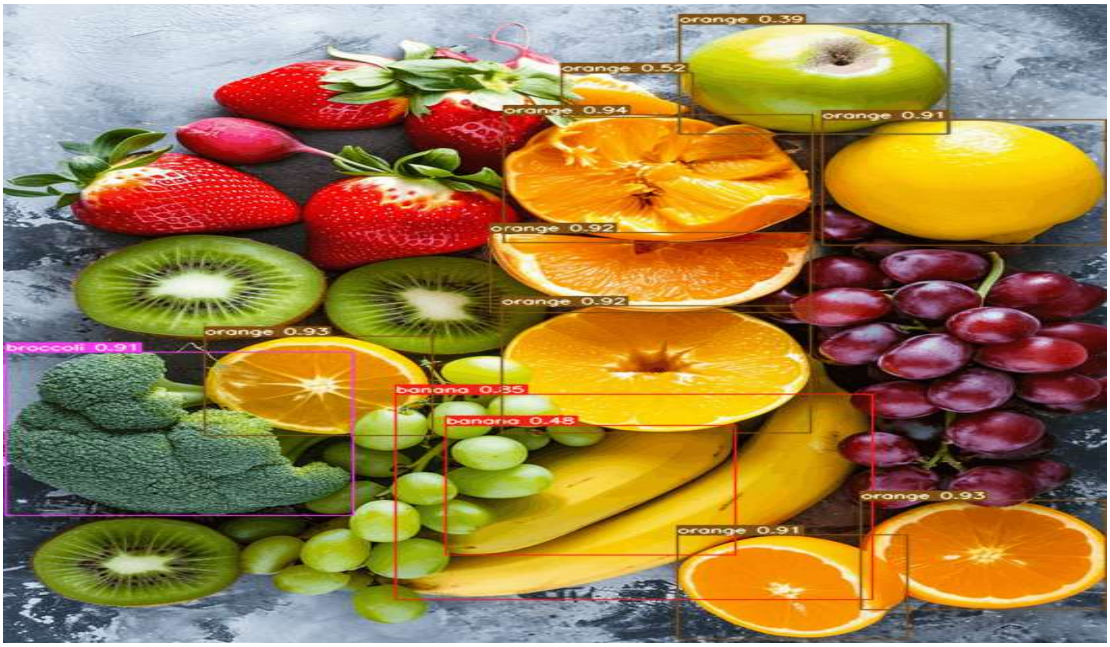

圖 17　預測結果圖

- 圖 17 可以看到已成功識別了幾種水果，橙子、香蕉和花椰菜。
- 圖片中可看到草莓未被識別，主要是因為未天加入至訓練當中，檸檬則是被誤判為橘子。
- 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（香蕉的邊界框）的置信度分數較低，這表明需要進一步訓練提高迭代次數或是調整訓練參數。

三、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 5　遮擋物件訓練結果

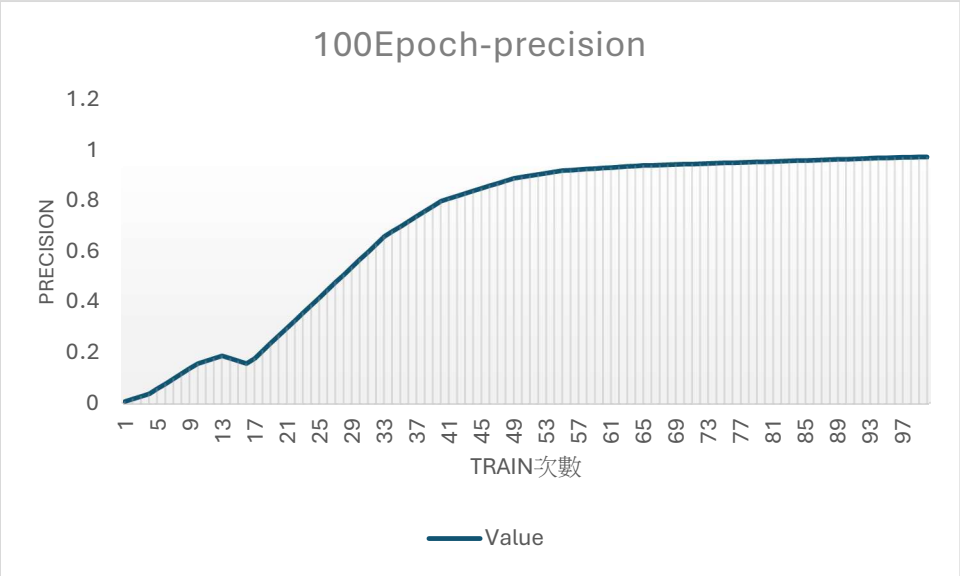

圖 18　100 次遮擋物件訓練-Percision 圖

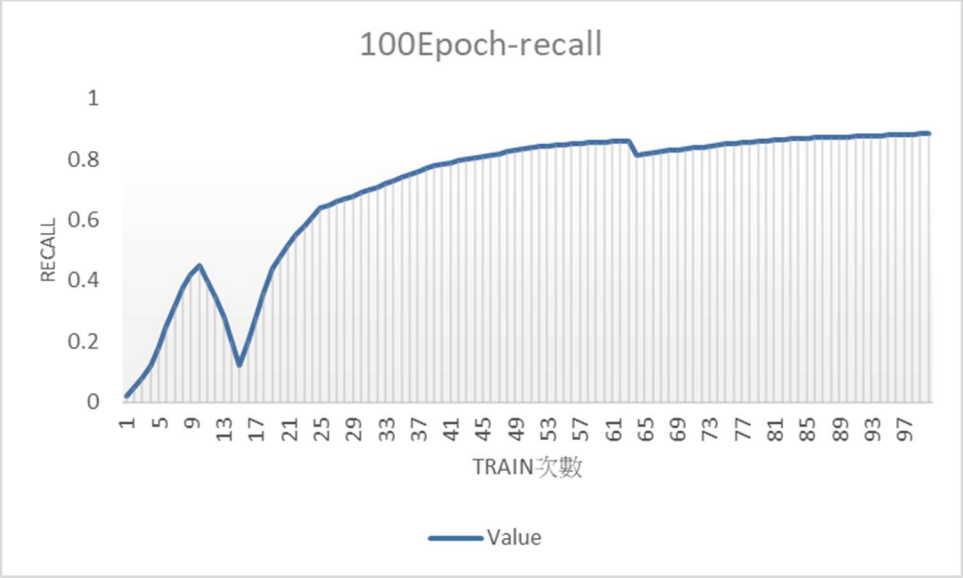

圖 19　100 次遮擋物件訓練-Recall 圖

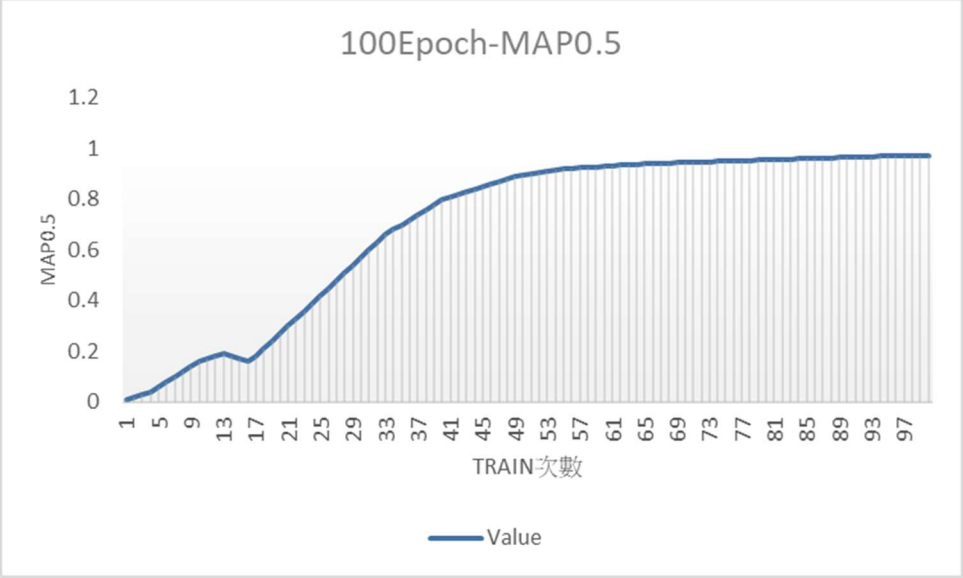

圖 20　100 次遮擋物件訓練-MAP0.5 圖

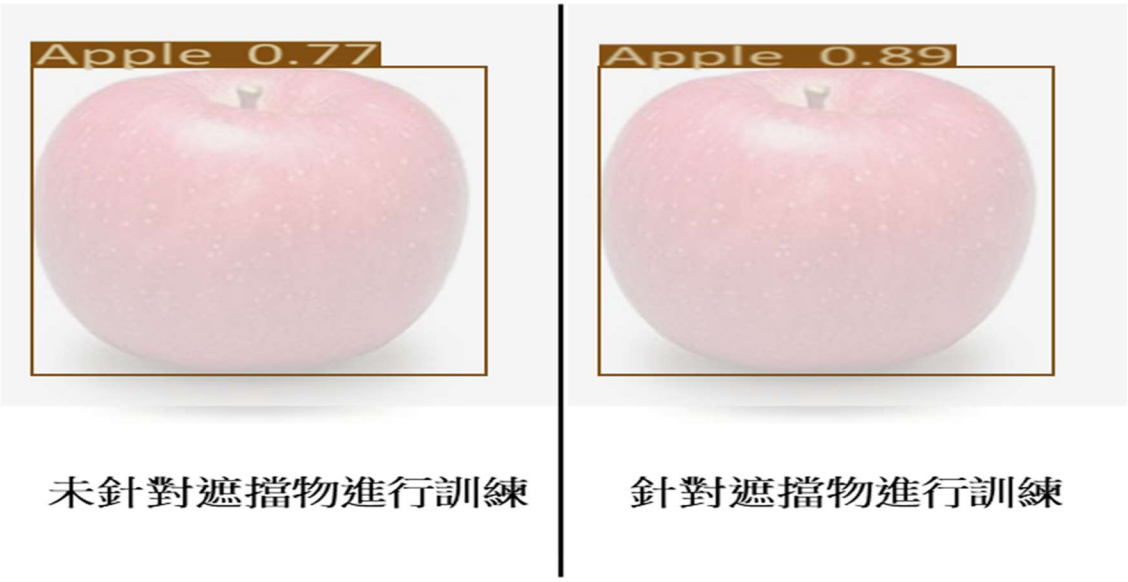

圖 21 有無針對遮擋訓練之差異圖

圖 21 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。

雲端技術(3 章的)

四、用戶端操作結果

本系統在實際操作中展現出穩健的性能。在物件偵測（Object Detection）方面，系統能準確識別大部分蔬果，即使目標部分存在部分遮擋（Partial Occlusion）的情況，仍可透過分析其輪廓特徵（Contour Features）與局部紋理（Local Textures）有效判別蔬果類型及顏色，這得益於 YOLOv7 模型在複雜場景下的強健性。

圖 23 可清楚識別圖中物件並準確回答圖

圖 24 進行遮擋訓練後也可以進行識別

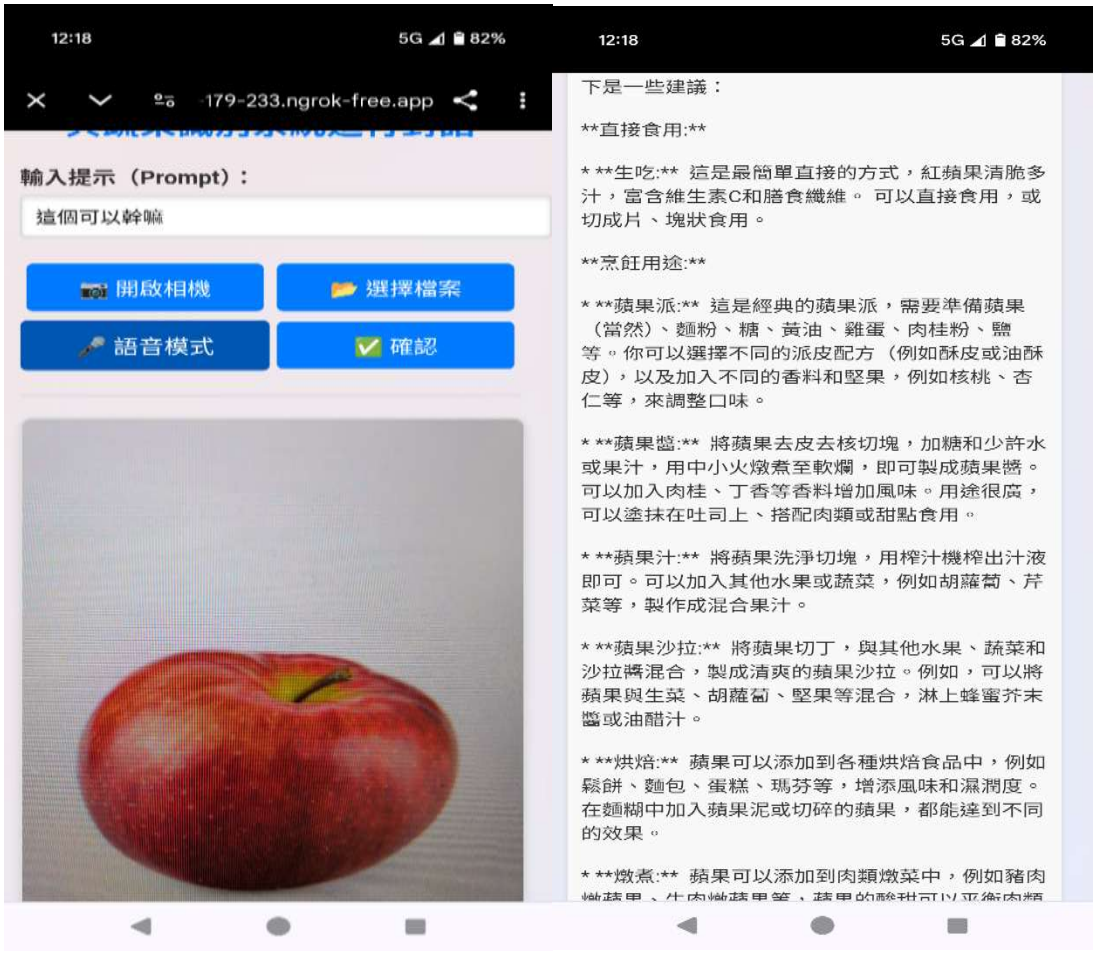

圖 25 手機版進行操作圖

第五章 結論與建議

二、結論(先討論在結論再發展，討論寫多點)

本研究成功整合 YOLOv7 目標檢測與 GPT-3.5 語言模型，實現了即時蔬果辨識與自然語言回應。系統能即時辨識手機畫面中的蔬果，並根據使用者提問給予語意說明，提升了 AI 與使用者互動的實用性。這成果呼應了本研究「建立能看、能說」即時互動系統的目標。本系統的目標是讓使用者透過手機畫面即時拍攝蔬果，系統能立即辨識其種類，並根據使用者提問給予相關資訊回答。整合視覺與語言能力的互動流程，將提升 AI 系統對實際場景的應用價值。(放至摘要與結論)

一、討論(為什麼用這個那個)

本系統證明視覺與語言整合的可行性，但也有下列限制：

- **蔬果辨識準確率受限於資料多樣性**：目前訓練資料以常見蔬果為主，對於罕見或外觀變異大的蔬果辨識效果較差，影響整體系統的實用性。
- **語言回應在複雜問題下有待加強**：當使用者提問較複雜或多步驟問題時，語言模型有時無法給出完整或精確的答案，偏離主題與方向。

三、未來發展(往好的地方寫)

未來從以下方向優化:

- **優化(可能要另外想)語言模型，增強回應多樣性**

  原因：優化模型或結合進階語言模型，提升回答的深度與靈活度。

  預期成果：系統能更完整、準確地回應使用者各類問題，增強互動智慧與用戶滿意度。

- **改善介面與互動方式，提升用戶體驗**

  原因：實際使用者數量有限，缺乏用戶回饋數據，無法全面掌握系統介面與互動流程的優化方向。

  預期成果：更友善的介面與多樣互動方式，將降低使用門檻，吸引更多用戶並提升系統使用率。

參考文獻

[1.] 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.

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

[3.] Shah, Uzair, et al. "Unveiling the Potential of ChatGPT and YOLOv7 for

Evaluating Children's Emotions Using Their Artistic Expressions." *Digital Health and Informatics Innovations for Sustainable Health Care Systems*. IOS Press, 2024. 409-413.

[4.] 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

[5.] 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.

[6.] 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.

[7.] 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.