多模態 AI 模型微調於無基準網頁視覺缺陷監測

Fine-Tuning Multi-Modal AI Models for Baseline-Free Web Visual Defect Detection

指導教授：李信杰
專題成員：曾立綸、毛宸薇
開發工具：Python 3.10.12、Playwright、YOLOv11n、LLaMA 3.2 Vision Instruct
測試環境：Linux Ubuntu 22.04.3 LTS

一、簡介：

軟體測試領域中，網頁破版是個難以被偵測及定義的問題。我們使用自製的網頁破版圖片資料集，針對不同模型進行微調訓練，預期將網頁截圖輸入至模型就能偵測到網頁上是否有不對齊的元素，若有則指出其位置。

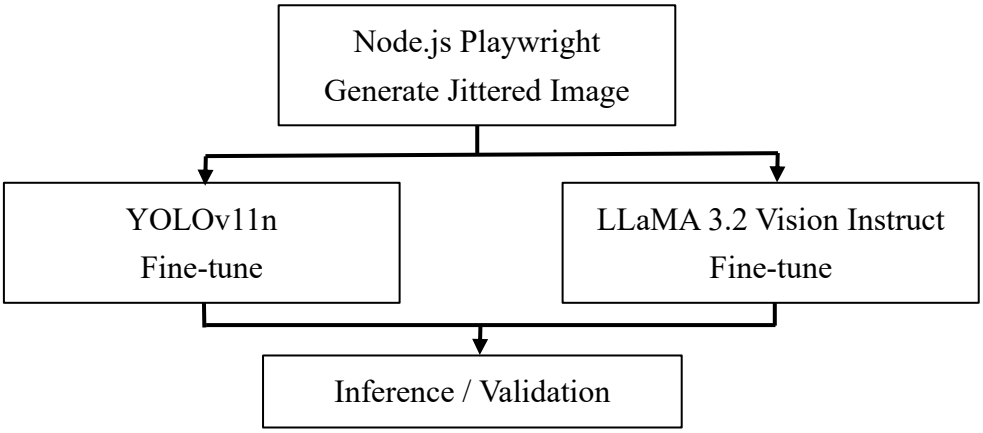
資料集部份我們透過爬蟲工具 Playwright 對網頁進行修改，找出網頁上有 flex 屬性的父元素，確認該元素至少包含兩子元素、視覺上可見、且長度不大於網頁的 1/4。則隨機取一子元素移動後截圖，並標註修改處的座標。

我們首先以 Ultralytics 推出的 YOLOv11n 模型作為基礎，訓練一套專門針對網頁截圖中元素錯位現象的物件偵測模型。經過訓練後，即可輸入網頁截圖並自動框出畫面中視覺不對齊區域。

我們進一步測試多模態模型的應用，使用 LLaMA 3.2 Vision Instruct 模型，使每筆圖片資料搭配自然語言輸入與預期回答進行微調訓練。訓練資料來源為標註過的 JSON 檔案，每筆資料包含截圖圖片路徑與一組「錯位區域」的左上與右下像素座標。若圖中無錯位情況，則以文字訊息表達無誤。

微調訓練完成後，允許進行圖片輸入與回答生成。推論時，將圖片與對應提示語包裝成對話格式，再以視覺語言模型共同編碼，產生最終預測回答。

以下為系統架構圖：



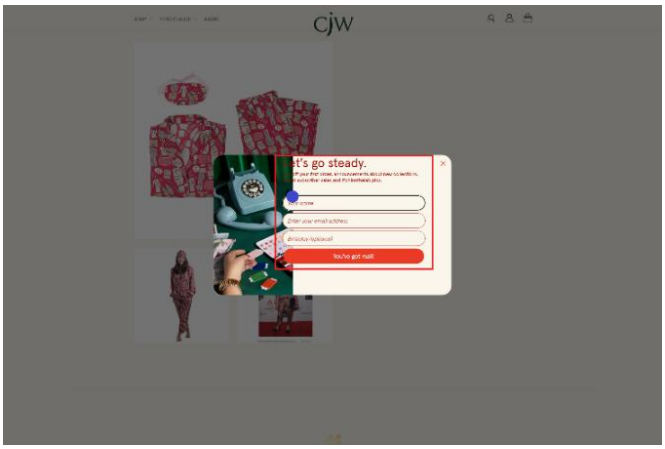
二、測試結果：

1. YOLO v11n 模型圖片偵測結果

圖一：網頁截圖偵測結果

圖二：網頁截圖偵測結果

2. LLaMA 3.2 Vision 模型圖片偵測結果



圖三：實際破版區域與偵測結果（手動繪製）