實務專題計畫成果摘要報告
物件辨識模型加速與輕量化

專題編號：111-CSIE-S010
執行期限：110 年第 1 學期至 111 年第 1 學期
指導教授：陳彥霖
專題參與人員： 108820019 陳昶宇
108820044 蕭岳

一、摘要

在很多的應用中，計算的資源像是 CPU、GPU 都是相對比較缺乏的，像是車用嵌入式物件辨識系統，為了能將龐大的模型放入規格有限的裝置中並維持一定的辨識需求，就需要對模型進行適當的調整。

我們採用的模型會符合現有的框架: PyTorch、TensorRT 等，並進行加速與模型輕量化，方法包括批次預測、模型量化，在保持相對準確的情況下辨識的速度越快越好。

專題的最終目標是在裝置可以正常進行辨識時，達到一定的準確率，速度達到多面實時，也就是基於特定嵌入式系統(型號為 eNVP-JTX-IV-V0008)完成一個多方向實時的影像物件辨識系統。

關鍵詞：嵌入式系統、物件辨識、批次預測(batch-inference)、模型輕量化、量化模型(quantization)

二、緣由與目的

大型車種駕駛需要注意的方向很多，死角也多，所以常常發生意外，因此之前政府提出大型車須強制加裝視野輔助系統，透過車內外的監視鏡頭，在螢幕上顯示周邊路況與車輛狀況的影像輔助駕駛人行駛。

我們希望在用以偵測物件的深度學習演算法方面，將 yolov4 進行優化，使模型更小、運算複雜度更低，最後將優化後的模型放入車用嵌入式 AI 物件辨識系統中。

三、進行方式

(一) 確認物件辨識模型
蒐集開源的物件辨識模型，並實際嘗試在特定的嵌入式系統上的辨識效果與實際占用空間符合預期，最後選擇最好的模型，也就是 YOLOv4

(二) 特定的嵌入式系統環境模擬
為了減少使用維修不便的嵌入式系統，參考嵌入式系統的環境，在實驗室的虛擬機上盡量的安裝相同配置的環境，後續實驗都先在虛擬機上進行並記錄結果

(三) 模型加速
將 YOLOv4 進行 TensorRT 模型加速

(四) 批次預測(batch-inference)
為了提升辨識速度，同時輸入多張影像讓模型更好的分配利用計算資源，讓 YOLOv4 模型進行批次預測

(五) 整合系統
將模型上到嵌入式系統，完善整體架構，測試整體穩定度

四、成果

- 將我們優化過後的 yolov4 模型上到車用嵌入式 AI 物件辨識系統
- 將影像送入到嵌入式系統中進行物

件偵測

- 辨識結果維持一定的 FPS 與辨識準確率

五、工具說明

(一) eNVP-JTX-IV-V0008: 嵌入式系統

(二) VMware Workstation Pro:
一款桌面平台虛擬化管理程序，我們用於遠端連結虛擬機

(三) Visual Studio Code: 程式編輯器

六、參考文獻

PyTorch 轉 TensorRT 模型部署 - Dynamic Shape (Batch Size)
https://zhuanlan.zhihu.com/p/387853124

Dynamic Quantization
https://pytorch.org/tutorials/recipes/recipes/dynamic_quantization.html