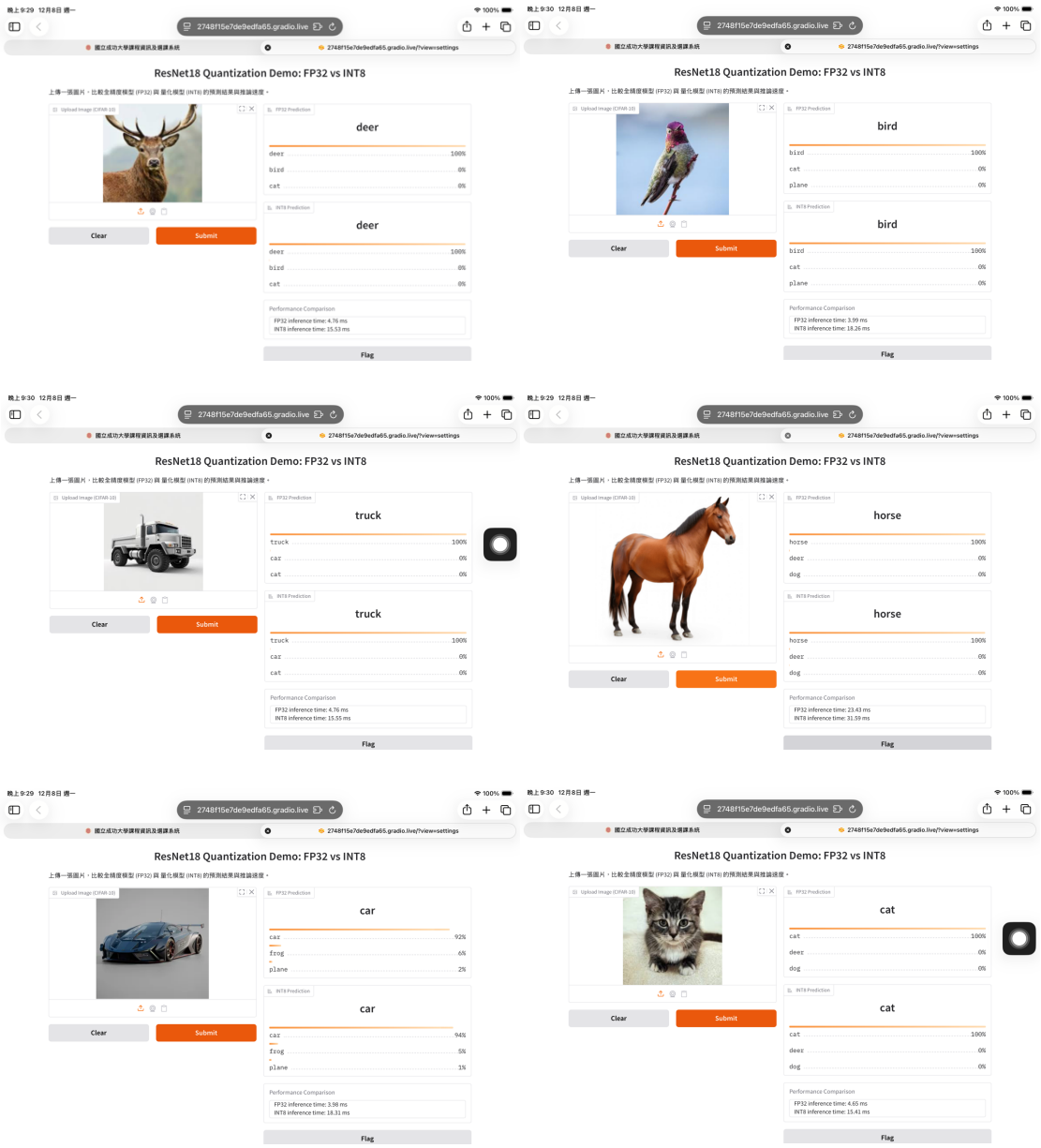
EAI-Lab5

學號：NM6131051　　姓名：張芷晴

一、在公開的 Gradio 介面網址上，上傳圖片之結果：

共測試六種圖片，分別為：deer、bird、truck、horse、truck、cat

圖片來源皆來自 google 搜尋，大小選擇圖示。



二、說明使用 compare_fp32_int8 function 中，你觀察到 FP32 與 INT 8 有什麼差別，需附上相關證據。

1. 效能差異：

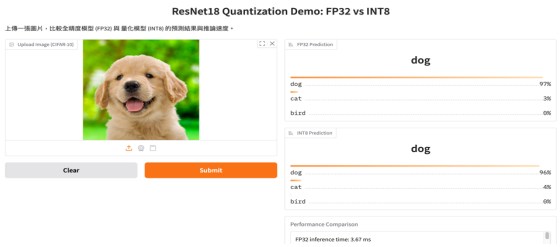
根據第一部分所測試的六張圖片，發現 INT8 的模型比 FP32 的模型慢大約 3~4 倍，但理論上 INT8 應該要比 FP32 快？有在討論區看到其他同學也有相同的疑問，看了助教的回答後，我嘗試換用 colab 去跑，發現同樣的 car 變成 INT18(14.33ms)比 FP32(26.09)快。所以這部分的確有可能是我自己所使用的硬體不支援 INT8 加速。

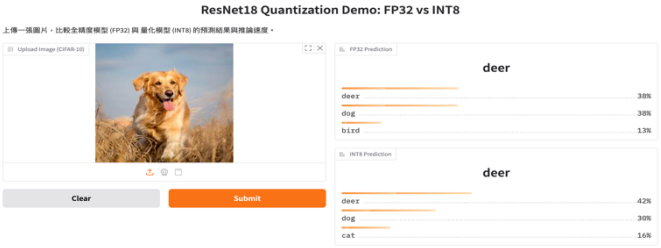
下表為 FP32 和 INT8 效能差異：

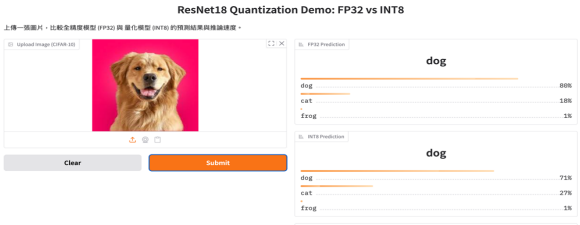
	FP32	Int8
Deer	4.76	15.53
Bird	3.99	18.26
Truck	4.76	15.55
Horse	23.43	31.59
Car	3.98	18.31
Cat	4.65	15.41

2. 準確度差異：

這次我用三張不同的狗狗照片去測試，發現 FP32 的準確率較高，INT8 較低，這部分滿合理的，因為他是量化過後的模型，自然沒那麼精確。







三、對於本次 lab 的心得與建議

我覺得這次的 Lab 滿有趣的，第一次使用 Gradio，感覺之後也可以把它用在其他地方，或是期末專案。還有討論區也是一個很不錯的設置，常常遇到問題的時候，會先在上面看有沒有其他同學有相同的疑問，可以省去滿多自己盲目找尋解答的時間。