農地作物現況調查影像辨識競賽 - 春季賽：AI作物影像判釋

結果報告

內容

壹、環境

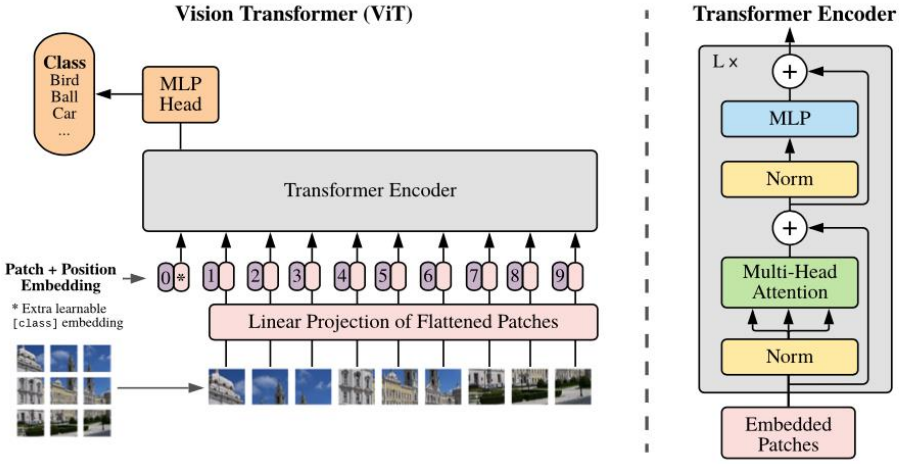
1. 作業系統：Windows 11
2. 語言：Python
3. 套件(函式庫)：
    1. matplotlib==3.5.1
    2. numpy==1.22.3
    3. Pillow==9.1.0
    4. pycm==3.4
    5. scikit-learn==1.0.2
    6. tensorboard==2.8.0
    7. torch==1.11.0
    8. torchmetrics==0.8.0
    9. torchvision==0.12.0
    10. tqdm==4.64.0
    11. transformers==4.18.0
    12. fire
4. 預訓練模型：google/vit-base-patch16-224-in21k
   (https://huggingface.co/google/vit-base-patch16-224-in21k)

貳、演算方法與模型架構

模型架構：Vision Transformer

Vision Transformer (ViT) [1]自提出之後，在各大影像辨識資料集如 ImageNet、CiFar100，相比於傳統 ConvNet 網路都獲得更好的表現。引用了 Transformer 內的 self-attention 自注意力機制，讓全域資訊能更好傳遞至深層網路。同時由於缺少 inductive bias，使模型的表現能隨著資料量有更好的加成。實際拿農業資料集訓練比對後，ViT 在驗證集表現遠 ResNet-50 約莫 10%，因此選擇 ViT 作為本次比賽的主要模型。

本次選擇模型之參數：86.791M，20.627G flops



(圖 1：Vision Transformer 架構)

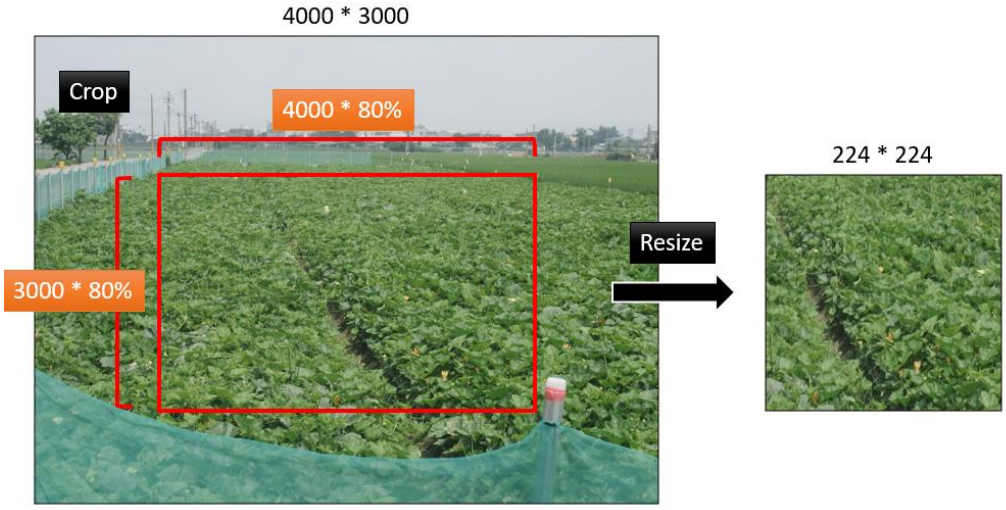
(來源：https://github.com/google-research/vision_transformer)

參、資料處理

資料刪減：

由於硬體效能限制，無法容納所有圖片資料(共計 150GB 左右)，且將原始圖片直接訓練會消耗大量時間，因此決定將所有圖片先經過圖像處理，使每張圖片由原先最小的解析度 1280*720，全部統一縮放至解析度 224*224，總資料容量也大幅下降 98%，僅剩下 2.5GB。

依照對資料集的觀察，可以發現目標(農作物)幾乎位於圖片的中央偏下。依上述觀察，圖像處理時，我們僅保留由中間點算起，往上下左右 4 方向各取 40%範圍內的圖像，再以內插法將圖片縮小至 224*224。



(圖 2：資料刪減示意圖)

相關程式碼請見 Github ./data/dataset/preprocessing.py

資料增強：

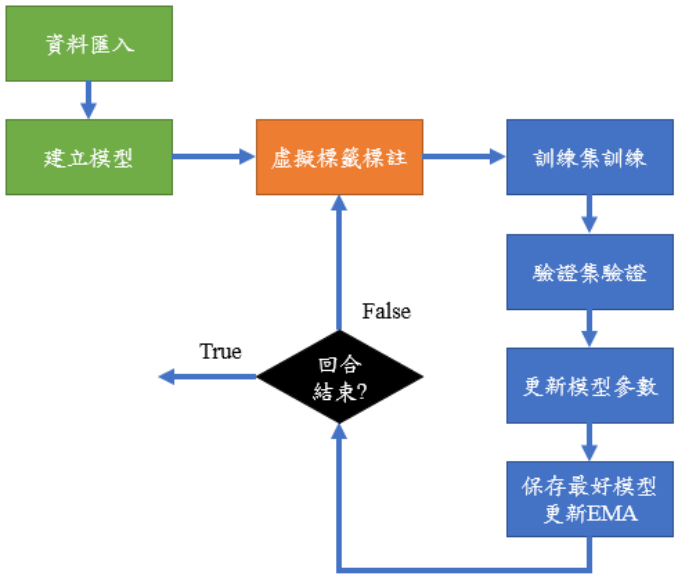
1. RandomResizedCrop()
2. ColorJitter()
3. RandomHorizontalFlip()

肆、訓練方式

- Optimizer：Adam
- Batch size：32
- Epoch：15
- Learning rate：學習率初始 0.00002，並使用 linear decay 每 2 epoch 減少 90%的學習率
- Criterion：Cross-Entropy，並對各 class 施加不同權重(根據該類別擁有的訓練集比例而定，避免不平衡訓練集對模型的負面影響)
- Regularization：Label smoothing
- 其他：
  1. 加入 EMA (Exponential moving average)
  2. 對驗證集進行 Self-supervised 自監督訓練，當模型對驗證集內圖片信心水準夠高時，為該圖片建立標籤，納入訓練集內。(每次 Epoch 都會重新進行 pseudo label 虛擬標籤的標註)

訓練流程：

1. 匯入資料並建立模型後，開始進行訓練
2. 進行 pseudo label 虛擬標籤標註，試圖增加訓練集規模
3. 常規模型訓練過程，訓練集、驗證集參與後，更新參數
4. 更新 EMA，並保存當前表現最好之模型，若尚未結束，則回到步驟 2

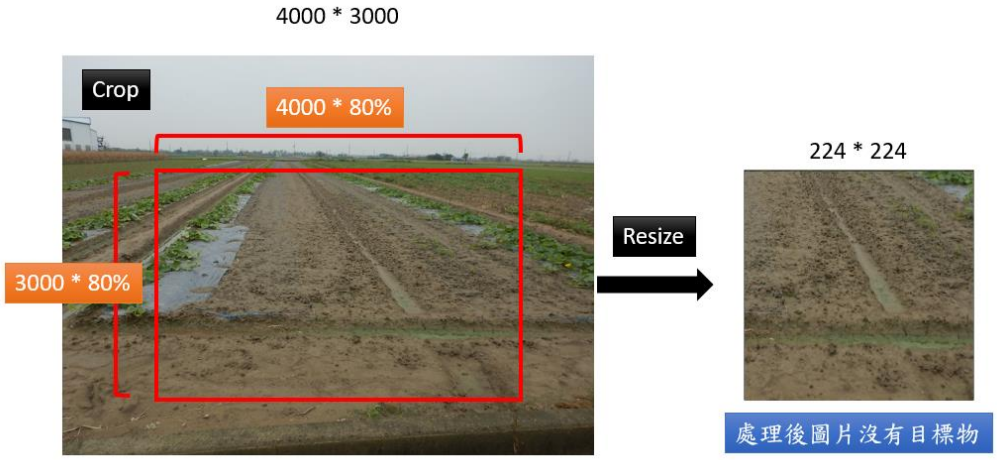


(圖 3：模型訓練流程圖)

伍、分析與結論

資料處理

第三章所提到的資料刪減方式，雖然有效降低了資料量，但相對也讓輸入圖片的資料含量降低，同時也可能出現如圖 2 沒有切割到目標的圖片。



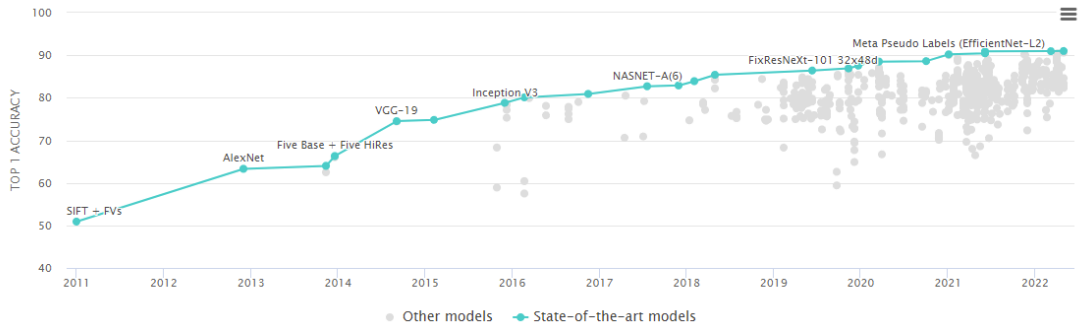
(圖 4：資料刪減中有問題之圖片示意圖)

後續應保留全圖，最大化資料含量，避免資料處理出現問題，導致模型學習不穩定。

模型選擇

Vision Transformer (ViT)現在已經有相當多的演化，如 Swin Transformer[2]等等，都對最一開始提出的基礎 ViT 架構再進行不同方面的改進；CNN 架構方面，重新統整的網路如 ConvNeXt[3]也表現出不輸給 ViT 的表現。

後續應嘗試更多不同網路，來比較不同網路在農業資料集的預測結果。



(圖 5：影像辨識 ImageNet Benchmark)

陸、程式碼

- Github：https://github.com/jimmylin0979/AICUP-2022-CropsClassifier

柒、使用的外部資源與參考文獻

[1]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. 2017.

[2]. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021.

[3]. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie. A ConvNet for the 2020s. 2022.

附件

作者聯絡資料表

- **隊伍**

隊伍名稱	Private Leaderboard 成績	Private Leaderboard 名次
TEAM_1514	0.9802984	48/151

- **隊員(隊長請填第一位)**

姓名 (中英皆需填寫)	學校名稱 (中英文皆需填寫)	系所 (中英皆需填寫)	電話	E-mail
林哲豪 (Jhe-Hao Lin)	清華大學	工業工程與工程管理學系	0979268400	jimmylin0979@gmail.com

註：E-mail 請填寫常用信箱，得獎後將以此信箱作為聯繫窗口。