生成式人工智慧-第11組
動物影像分析

國立成功大學工程科學系

楊皓宇	楊孟洵	陳稟翰	謝承翰	方威仁
N96121139	N96131176	N98121012	N96124624	N96131207

Outline
大綱>>

01/ 動機

02/ 資料集介紹、分割

03/ CNN、Transformer

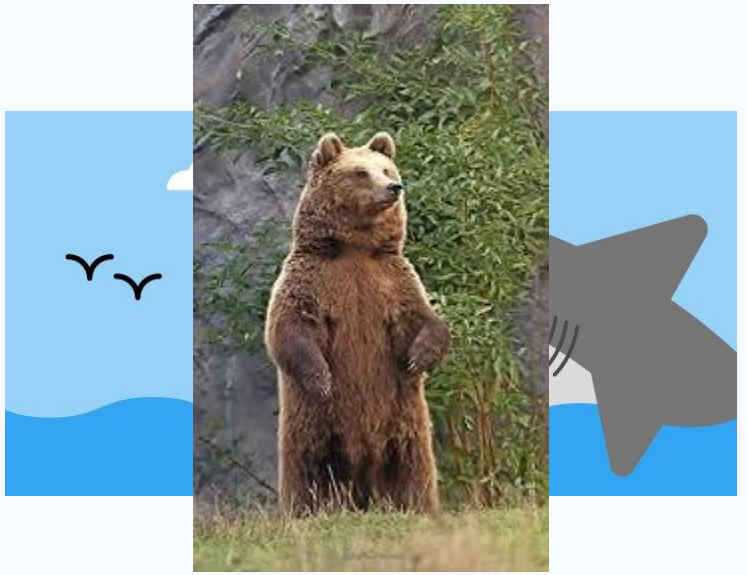
04/ LLM Fine-tune

05/ 總結

01／動機

- 從Kaggle上找了一些資料集
- 聚焦在分類任務上，使用了以下架構
  - CNN
  - Transformer
  - LLM Fine-tune
- 想做一些比較並分析各個架構模型之間做分類任務的差異

02／資料集介紹-動物照片



- 選用資料集為動物照片
- 照片總數：5400 張 (各個種類60張)
- 動物總類：90 種 (包括昆蟲)
- 資料集來源為 Kaggle 上整理而成，作者從Google圖片搜尋下載，所以資料集內的圖片解析度不一，在這之中也會有童話類型的照片，與AI生成或合成的動物照片

02／資料集分割-動物照片

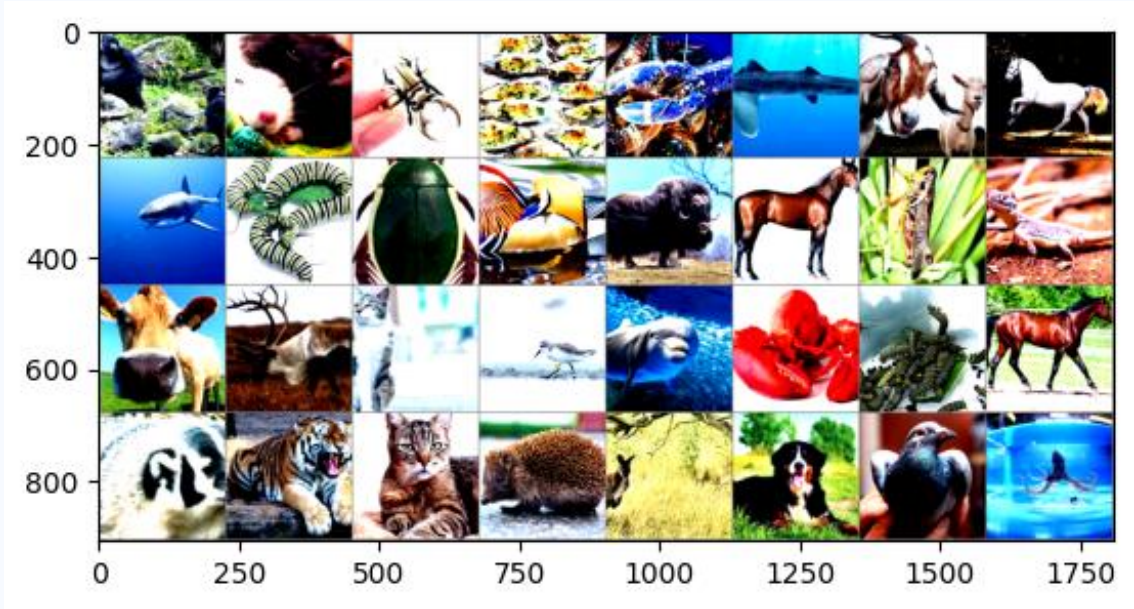
- 將資料集切分成訓練集、驗證集、測試集
  - 比例 : 70%、15%、15%
  - 張數： 3780張、810張、810張

- 由於想比較各個成果的差異性，在做資料集切割的時候我們統一選定隨機種子皆為42

```
# 定義資料集分割比例
train_ratio = 0.7
val_ratio = 0.15
```

03／ CNN、Transformer-設定

- 訓練方式 : 訓練到正確率沒有明顯進展即停止
- 模型選用 : 使用 Pytorch 套件提供的各個預訓練模型下去執行
- Data Augmentation
  - Resize至 224* 224
  - Mixup + CutMix
  - 水平翻轉，亮度對比調整(max 0.4)
- Optimizer : AdamW
- Batch size : 32
- 訓練用設備 :
  - CPU :I5-12400F
  - RAM:48GB
  - GPU:4080S+3060TI

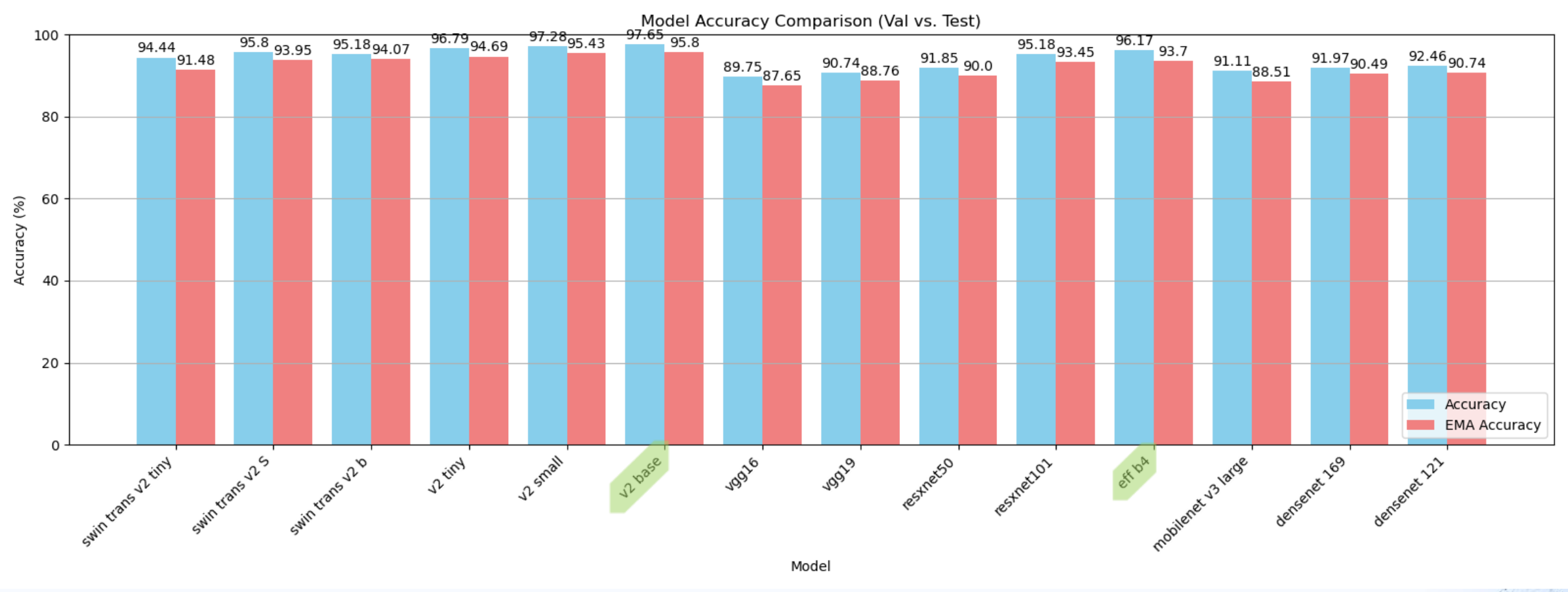


03／CNN、Transformer-結果

Transformer	Val Accuracy
Swim trans. v2 tiny	94.44%
Swim trans. v2 S	95.80%
Swim trans. v2 b	95.18%
Vmamba V2 Tiny	96.79%
Vmamba V2 small	97.28%
Vmamba V2 **base**	**97.65%**

CNN	Val Accuracy
VGG16	89.75%
VGG19	90.74%
Resxnet50	91.85%
Resxnet101	95.18%
Efficientnet b4	**96.17%**
Mobilenet v3 large	91.11%
Densenet169	91.98%
Densenet121	92.47%

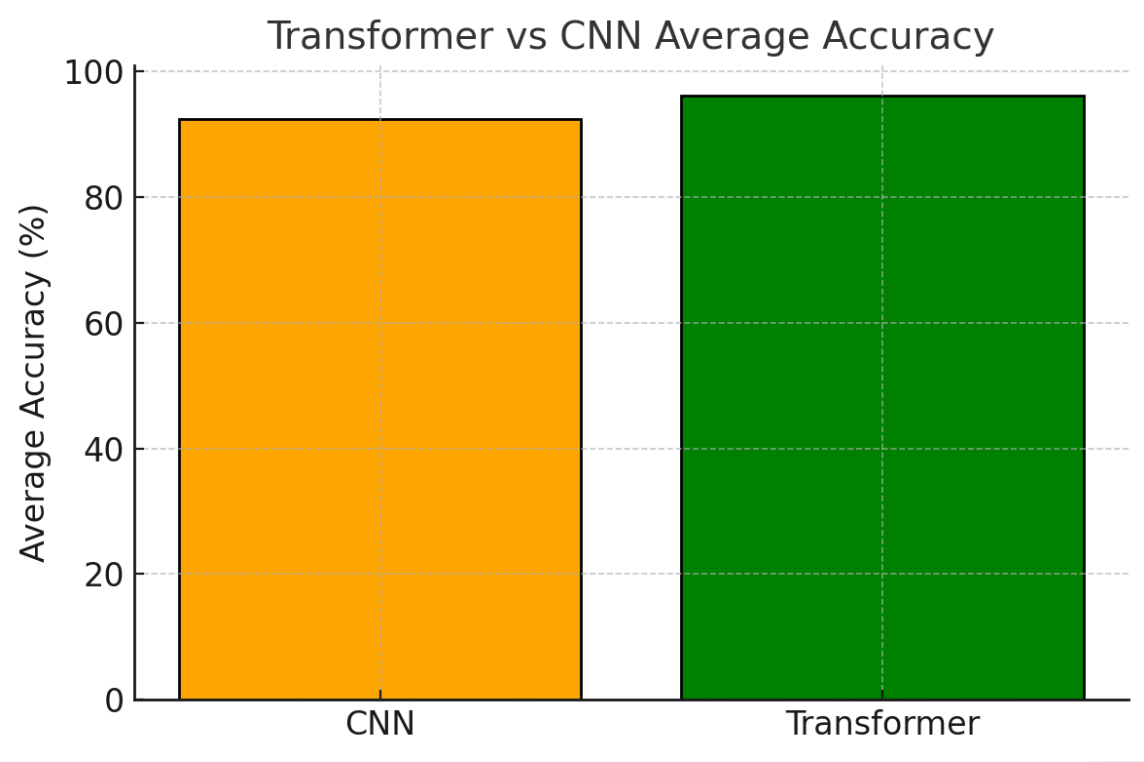
03／CNN、Transformer- Accuracy (Val vs. Test)



03／ CNN Vs. Transformer

各個模型跑出來的結果，做平均後比較CNN架構與Transformer架構。

我們認為在這個資料集中做分類任務使用Transformer架構會比CNN更好，除了正確率的結果更高之外，另一部分認為是資料集本身圖片品質繁雜，導致卷積網路在特徵提取沒有比注意力機制取得優勢，並且在做訓練時所花費的時間也差距不大。



04／LLM Fine-Tune - 模型介紹

- 使用模型：Llama 3.2 Vision Model
- 特色：
  - 支援圖像辨識任務
  - 採用 4-bit 量化技術降低記憶體需求
  - 使用 Unsloth 框架加速訓練
  - 支援梯度檢查點 (gradient checkpointing) 處理長序列


```
model, tokenizer = FastVisionModel.from_pretrained(
    "unsloth/Llama-3.2-11B-Vision-Instruct",
    load_in_4bit = True,  # 使用4-bit量化降低記憶體使用
    use_gradient_checkpointing = "unsloth"  # 長序列處理
)
```


圖片來源：https://github.com/unslothai/unsloth

04 LLM Fine-Tune - 訓練說明

- 訓練平台：
  - Colab
  - GPU：T4
- 資料集：
  - 動物圖片資料集：90 種動物類別
  - 訓練資料：3,780 張圖片
  - 驗證資料：810 張圖片
- 訓練設定：
  - 使用 LoRA 技術進行模型調整
  - 批次大小 (Batch size)：2
  - 梯度累積步數：4
  - 學習率：2e-4
  - 採用 8-bit Adam 優化器

```
model = FastVisionModel.get_peft_model(
    model,
    finetune_vision_layers = True,      # 微調視覺層
    finetune_language_layers = True,    # 微調語言層
    r = 16,                             # LoRA rank
    lora_alpha = 16,                    # LoRA alpha值
    lora_dropout = 0,                   # LoRA dropout率
)

trainer = SFTTrainer(
    model = model,
    args = SFTConfig(
        per_device_train_batch_size = 2,
        gradient_accumulation_steps = 4,
        learning_rate = 2e-4,
        max_steps = 30,
        optim = "adamw_8bit"
    )
)
```

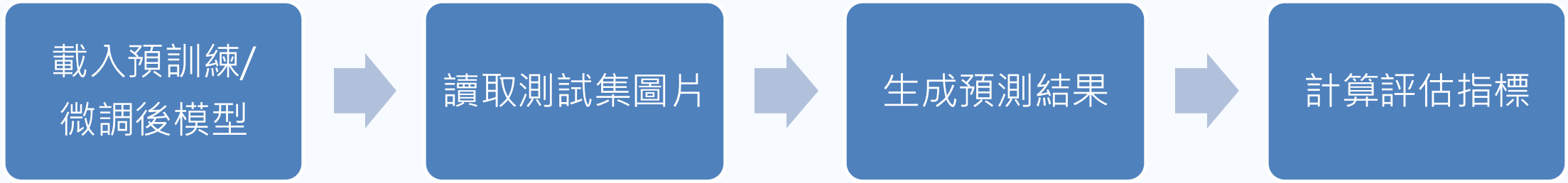
04／LLM Fine-Tune - 訓練說明

- 分批訓練：
  - 因 Colab 的 **CPU RAM 用量限制**，因此將訓練集拆分為三份，分批進行三次訓練。
- 指令設計：
  - 提示模型其為專業的分類器，要求其在指定的動物類別中選擇一個作為正確答案，並且只能回傳該動物的名稱，不得包含任何額外的訊息

```
# 模型指令模板
instruction_template = (
    "You are a professional image classification model. "  # 定義模型角色
    "Your task is to classify the given image into one of the following categories: "
    f"{animal_list} "  # 動態插入90種動物類別清單
    "Respond with the category name only, "  # 回應格式要求
    "and do not include any additional information."  # 額外限制說明
)
```

04／ LLM Fine-Tune - 測試方法

- 測試流程：



04 LLM Fine-Tune - 測試方法

- 評估方法：
  - #A 輸出完全符合則預測**成功**
  - #B 輸出非正解動物預測**失敗**
  - #C 輸出包含動物名稱以外的資訊則預測**失敗**
- 評估指標：
  - 總體準確率 (考慮#A、#B和#C)
  - 過濾後準確率 (考慮#A和#B)
  - 召回率
  - F1分數

動物類別(烏龜)

#A → 烏龜 ○

#B → 鱷魚 ✕

#C → 正在游泳的烏龜 ✕

```
# 評估指標計算
from sklearn.metrics import accuracy_score, classification_report
accuracy = accuracy_score(filtered_true_labels, filtered_predictions)
report = classification_report(true_labels, predictions, target_names=animal_list)
```

04／ LLM Fine-Tune - 結果討論

- 整體表現提升：
  - 總體準確率：35% → 92%
  - 過濾後準確率：93.69% → 95.16%
  - 召回率：0.34 →0.91
  - 加權平均F1分數：0.47 → 0.92

效能提升：

	Before	After
Accuracy	35%	92%
F.Acc	93.69%	95.16%
Precision	0.90	0.95
Recall	0.34	0.91
F1-score	0.47	0.92

04／ LLM Fine-Tune - 結果討論

- 顯著改進：
  - 完全正確識別的類別大幅增加
  - 原本無法識別的類別 (如Shark、Zebra) 結果變好
- 待改進項目：
  - Penguin 類別準確率低
  - 部分類別 (如Squirrel、Reindeer) 的召回率低

Classification Report:

antelope	1.00	1.00	1.00	9
badger	1.00	0.78	0.88	9
bat	1.00	0.89	0.94	9
bear	1.00	1.00	1.00	9
bee	1.00	1.00	1.00	9
beetle	1.00	1.00	1.00	9
bison	1.00	0.89	0.94	9
boar	0.90	1.00	0.95	9
butterfly	1.00	0.22	0.36	9
cat	0.64	1.00	0.78	9
caterpillar	1.00	1.00	1.00	9
chimpanzee	1.00	1.00	1.00	9
cockroach	1.00	1.00	1.00	9
cow	1.00	1.00	1.00	9
coyote	0.75	1.00	0.86	9
crab	1.00	0.67	0.80	9
crow	1.00	1.00	1.00	9
deer	1.00	1.00	1.00	9
dog	0.90	1.00	0.95	9
dolphin	1.00	0.78	0.88	9
donkey	1.00	1.00	1.00	9
dragonfly	1.00	1.00	1.00	9
duck	1.00	1.00	1.00	9
eagle	1.00	0.89	0.94	9
elephant	1.00	1.00	1.00	9
flamingo	1.00	1.00	1.00	9
fly	1.00	1.00	1.00	9
fox	1.00	1.00	1.00	9
goat	1.00	1.00	1.00	9
goldfish	0.90	1.00	0.95	9

Classification Report:

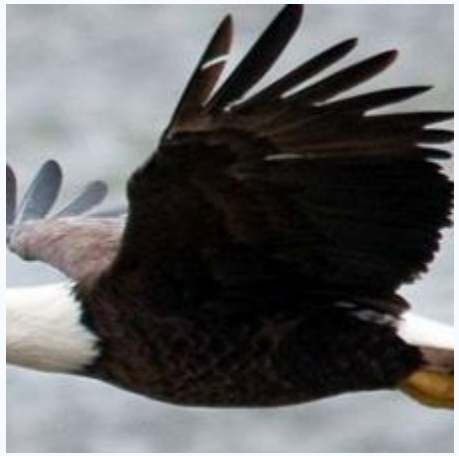
antelope	0.67	0.22	0.33	9
badger	0.50	0.11	0.18	9
bat	1.00	0.22	0.36	9
bear	1.00	0.11	0.20	9
bee	0.71	0.56	0.62	9
beetle	1.00	0.56	0.71	9
bison	0.67	0.22	0.33	9
boar	1.00	0.44	0.62	9
butterfly	1.00	0.11	0.20	9
cat	0.33	0.11	0.17	9
caterpillar	1.00	0.33	0.50	9
chimpanzee	1.00	0.11	0.20	9
cockroach	1.00	0.22	0.36	9
cow	1.00	0.67	0.80	9
coyote	0.89	0.89	0.89	9
crab	1.00	0.33	0.50	9
crow	0.75	0.33	0.46	9
deer	1.00	0.22	0.36	9
dog	1.00	0.56	0.71	9
dolphin	1.00	0.89	0.94	9
donkey	1.00	0.11	0.20	9
dragonfly	1.00	0.33	0.50	9
duck	1.00	0.22	0.36	9
eagle	1.00	0.44	0.62	9
elephant	1.00	0.67	0.80	9
flamingo	1.00	0.11	0.20	9
fly	1.00	0.56	0.71	9
fox	1.00	0.33	0.50	9
goat	1.00	0.67	0.80	9
goldfish	1.00	0.56	0.71	9

04／ LLM Fine-Tune - 結論

- 微調成功提升模型表現
- 大多數動物類別識別準確度顯著提高
- 輸出為動物列表中動物機率有明顯改進
- 模型預測穩定性增強
- 仍有特定類別需要改進

05／總結 - 錯誤資料分析

- Vmamba V2 Base(Transformer) 與 Efficient net b4(CNN) 容易出錯的圖片
  - **部分圖像或遮擋(可能沒偵測到特徵)**
  - 複數動物或背景複雜(辨識到其他特徵)
  - 顏色相近(Vmamba V2 Base)






05／總結 - 錯誤資料分析

- Vmamba V2 Base(Transformer) 與 Efficient net b4(CNN) 容易出錯的圖片
  - 部分圖像或遮擋(可能沒偵測到特徵)
  - **複數動物或背景複雜(辨識到其他特徵)**
  - 顏色相近(Vmamba V2 Base)






05／ 總結 - 錯誤資料分析

- Vmamba V2 Base(Transformer) 與 Efficient net b4(CNN) 容易出錯的圖片
  - 部分圖像或遮擋(可能沒偵測到特徵)
  - 複數動物或背景複雜(辨識到其他特徵)
  - **顏色、紋理相近(Vmamba V2 Base)**

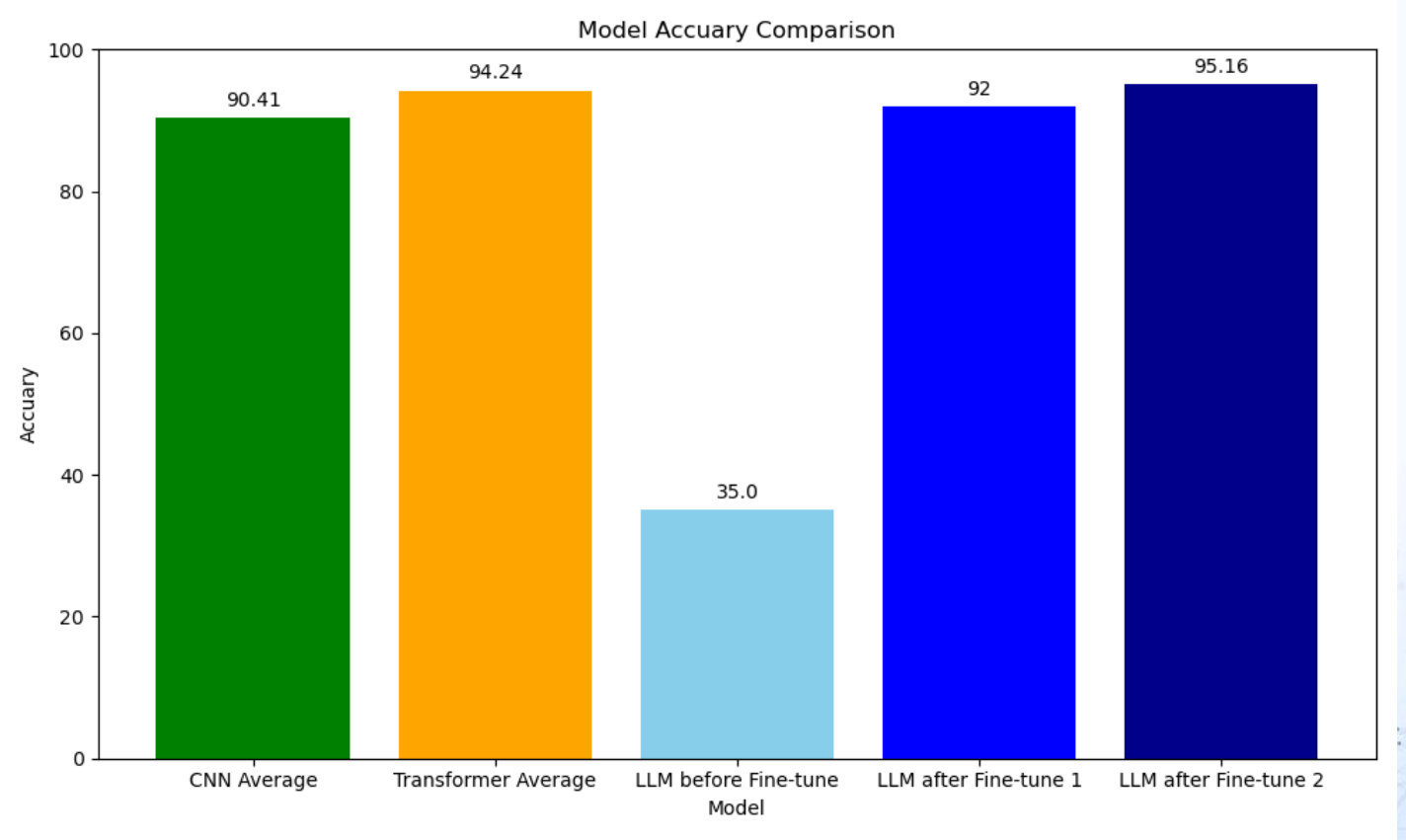
05／ 總結 - 比較

LLM Fine-tune 相較於 Transformer 跟 CNN 在某些動物表現較好（如水母）

皆使用測試集做測試結果

LLM after Fine-tune 1
為總體正確率

LLM after Fine-tune 2
為過濾正確率




Q
&
A

06／分工表

	分工
陳稟翰(組長)	視覺模型實驗(CNN、Transformer)、模型實驗統籌、報告者
方威仁	簡報製作(資料集介紹、CNN、Transformer)、CNN(VGG16)、簡報彙整、報告者
楊孟洵	簡報製作(LLM)、LLM Fine-Tune、報告者
楊皓宇	簡報製作(LLM)、LLM Fine-Tune、LLM 實驗統籌、報告者
謝承翰	簡報製作(總結)、總結部分(錯誤部分分析)

07 补充

Transformer	Machine
Swim trans. v2 tiny	3060ti
Swim trans. v2 S	3060 12G
Swim trans. v2 b	4080S
Vmamba V2 Tiny	3060ti
Vmamba V2 small	3060 12G
Vmamba V2 base	4080S

CNN	Machine
VGG16	3060ti
VGG19	3060 12G
Resxnet50	3060 12G
Resxnet101	4080S
Efficientnet b4	4080S
Mobilenet v3 large	3060ti
Densenet169	4080S
Densenet121	4080S

07 補充

DataSet&圖片來源：
https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals