HW - AI 資訊偵測與倫理作業報告

深度偽造 (deepfake)技術持續進步，偽造系統在部署時常對訓練未見的種類表現出很弱的檢測能力。然而，視覺語言模型（如 CLIP）可用少量額外參數就可以做到通用深偽檢測。本研究採用 CLIP 以少量可調參數進行 PEFT （Parameter-Efficient Fine-Tuning），展現其跨類型辨識 deepfake 能力。

1. Methodology

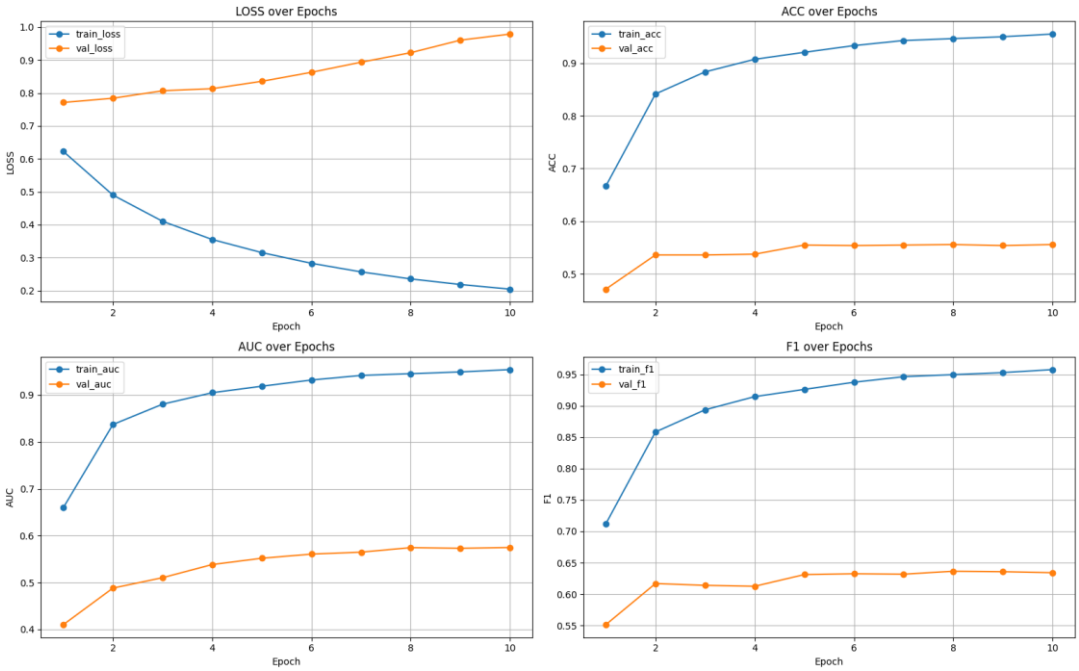
- ✧ Model: openai/clip-vit-base-patch32
- ✧ 訓練方法: LoRA
  - ■ PEFT 策略：在 Transformer 自注意力模組（q_proj、k_proj）嵌入 LoRA 層，不對骨幹權重除增量可訓練參數外進行微調。
  - ■ 凍結參數比例：

```
trainable params: 294,912 || all params: 63,460,864 || trainable%: 0.4647
```

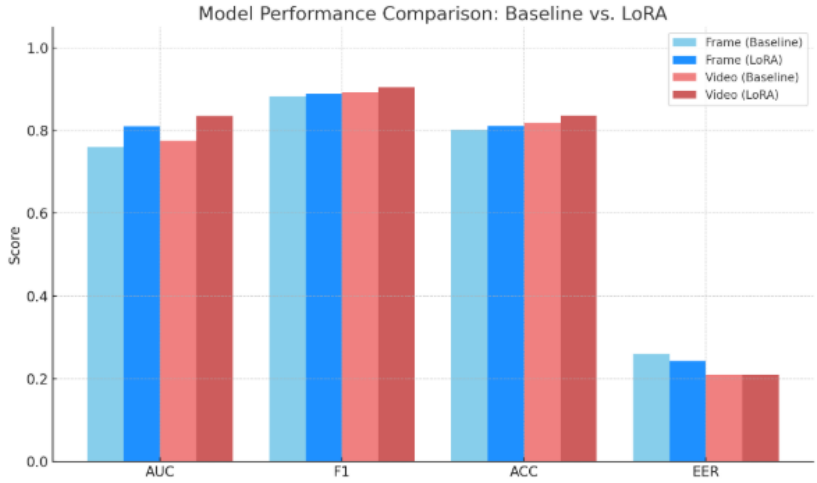
    Trainable ratio = 0.4647%
- ✧ Dataset 切分:
  - ■ Train: Real_youtube(80%) + FaceSwap(90%)
  - ■ Val: Real_youtube(10%) + FaceSwap(10%)
  - ■ Test: Real_youtube(10%) + NeuralTextures(100%)
- ✧ 固定隨機種子(seed = 42)
- ✧ 程式碼架構:
  - ■ model.py: 模型架構
  - ■ dataset.py: 建構資料集、處理資料集切分
  - ■ train.py: 主訓練程式
  - ■ test.py: 測試程式
  - ■ utils.py: 放工具函式

2. Experiments

✧ 訓練結果紀錄：



✧ 指標計算與結果：



Metric	Baseline (Frozen CLIP)	LoRA 微調	差異 Δ
Frame AUC	0.7602	**0.8100**	+0.0498
Frame F1	0.8823	**0.8879**	+0.0056
Frame ACC	0.8019	**0.8106**	+0.0087
Frame EER	0.2606	**0.2426**	↓0.0180
Video AUC	0.7750	**0.8350**	+0.0600
Video F1	0.8925	**0.9043**	+0.0118
Video ACC	0.8182	**0.8364**	+0.0182
Video EER	**0.2100**	**0.2100**	-

✧ 與 Baseline (Frozen CLIP + Linear Probe)比較：

- baseline 採用 frozen CLIP 的 image embedding，僅訓練線性分類器，泛化性高但對 deepfake 細節不足。
- LoRA 微調部分參數後，模型能更貼近 deepfake 特徵(如邊緣 artifacts、臉部不一致等)。
- LoRA 微調後顯著提升 AUC (Frame: +5%、Video: +6%)，顯示微調能有效改善模型對偽造內容的辨識能力。
- Frame-level 與 Video-level 全部指標皆略有提升，無指標下降。
- EER 持平或略降，表示 LoRA 微調並未犧牲辨識平衡性。

3. Discussion

✧ PEFT 效益：僅凍結 CLIP 骨幹仍具基本檢測能力，加入 LoRA 後 AUC 提升 ≈8%，驗證稀疏微調能提高泛化能力。

✧ 錯誤案例分析：

- 查看 FP/FN，列出錯誤率影片最高 top5

```
== False Positives ==
{'video_id': '218', 'score': 0.5023550391197205, 'label': 0}
{'video_id': '218', 'score': 0.5462505221366882, 'label': 0}
{'video_id': '218', 'score': 0.5037362575531006, 'label': 0}
{'video_id': '218', 'score': 0.5198004841804504, 'label': 0}
{'video_id': '218', 'score': 0.5626910924911499, 'label': 0}

== False Negatives ==
{'video_id': '611', 'score': 0.3318840563297272, 'label': 1}
{'video_id': '611', 'score': 0.4071859121322632, 'label': 1}
{'video_id': '611', 'score': 0.3317265808582306, 'label': 1}
{'video_id': '611', 'score': 0.30216893553733826, 'label': 1}
{'video_id': '611', 'score': 0.3920201063156128, 'label': 1}

== Top Videos by Error Rate ==
Video 578: error rate = 100.00%
Video 791: error rate = 100.00%
Video 484: error rate = 100.00%
Video 965: error rate = 100.00%
Video 746: error rate = 100.00%
```

模型的 False Positive 主要集中於 score 接近 threshold（0.5）區域，顯示模型對這些 frame 信心不足；而 False Negative 影片如 611 則可能因偽造特徵不明顯而難以辨識。此外，部分影片（如 578, 791）出現 100% 錯誤，顯示模型對特定資料分布外樣本無法泛化，可能需重新檢查標註品質或加入更多樣本以強化模型對異質影片的魯棒性。

✧ 未來改進方向：

- 可結合 prompt learning 或視覺重編程(masking)
- 加入更多 deepfake 類型訓練或評估 zero-shot 能力

4. References

https://arxiv.org/abs/2103.00020

https://huggingface.co/openai/clip-vit-base-patch32

https://arxiv.org/abs/2402.12927