Cross-Manipulation Deepfake Detection with Vision-Language Foundation Models

312831030 高士淵

Professor：許志仲

1.1 Model Backbone

We adopt the open-source CLIP (ViT-B/32) as our vision-language backbone.
The model weights are frozen throughout training.

1.2 Adaptation Strategy

We employ prompt tuning:

- Introduce learnable prompt tokens prepended to the text input for each class ("real", "fake").
- Only the prompt embeddings and a linear classifier head are updated during training; all CLIP backbone weights remain frozen.

2.1 Dataset and Split

	FaceSwap	NeuralTextures	Real_youtube
Train	100%	0%	90%
Test	0%	100%	10%

2.2 Training hyperparameters

Hyperparameter	Value
batch size	32
epochs	10
learning rate	1e-3
Optimizer	Adam

2.3 Evaluation

Metric	Value
AUC	0.9641
EER	0.1139
F1	0.9163
Accuracy	0.8589

3. Discussion

3.1 Limitations

- Some misclassifications occur on challenging videos with subtle artifacts or low-quality frames.
- The model may still be sensitive to distribution shifts not covered by the prompt tokens.

3.2 Future Work

- Explore other PEFT strategies (e.g., LoRA, adapters).
- Incorporate temporal reasoning or facial component guidance.
- Test on additional manipulation types and real-world datasets.

4. References

- CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection
- Standing on the Shoulders of Giants: Reprogramming VLM for General Deepfake Detection
- Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics
- Towards More General Video-based Deepfake Detection through Facial Component Guided Adaptation for Foundation Model
- C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection
- Unlocking the Hidden Potential of CLIP in Generalizable Deepfake Detection

THANKS!