Contrastive Learning for DeepFake Classification and Localization via Multi-Label Ranking

Cheng-Yao Hong

Yen-Chi Hsu

Tyng-Luh Liu

Institute of Information Science, Academia Sinica, Taiwan

CVPR 2024

Outline

1 Introduction

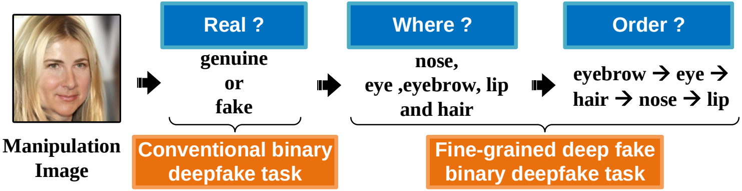
2 Method

3 Experiments

4 Conclusion

From convertional to fine-grained deepfake detection

- The task of conventional deepfake classification is usually formulated as a binary classification problem. Nevertheless, owing to the impressive development of generative networks, deep forgery is no longer limited to face-to-face interchange.
- A unified approach to **simultaneously address conventional binary deepfake detection and the fine-grained deepfake task**, where forged images created via deepfake mechanisms may be locally manipulated in one or more facial components or attributes.



Real ?

genuine
or
fake

Where ?

nose,
eye ,eyebrow, lip
and hair

Order ?

eyebrow → eye →
hair → nose → lip

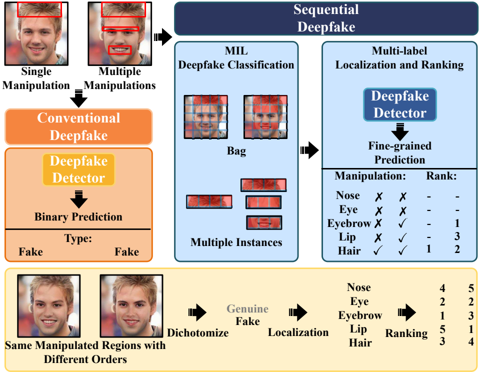
Manipulation
Image

Conventional binary
deepfake task

Fine-grained deep fake
binary deepfake task

Unified approach to deepfake classification and localization

- We introduce a multi-label ranking approach to tackling the “fine-grained” deepfake task (i.e., to localize the modified facial components and to identify the order of manipulations) and develop a contrastive multi-instance learning(MIL) framework to solve the binary classification.
- It is noteworthy to mention that manipulating the same regions in different orders could result in distinct manipulated images.



Manipulation:			Rank:	
Nose	✗	✗	-	-
Eye	✗	✗	-	-
Eyebrow	✗	✓	-	1
Lip	✗	✓	-	3
Hair	✓	✓	1	2

Nose	4	5
Eye	2	2
Eyebrow	1	3
Lip	5	1
Hair	3	4

Outline

1 Introduction

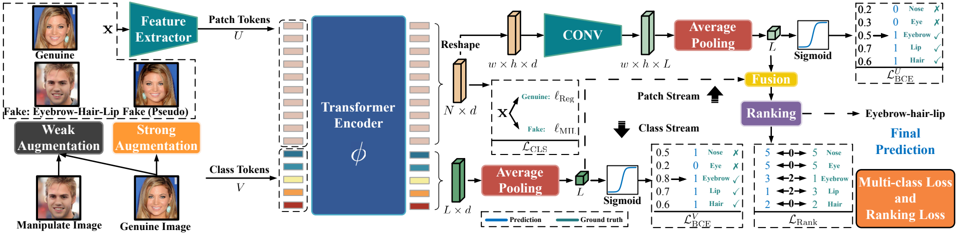
2 Method

3 Experiments

4 Conclusion

Overview of the model architecture

- **Deepfake classification:** For binary deepfake detection, the methodology adopts a MIL perspective, utilizing a sorted list of similarity scores between patch tokens.
- **Deepfake localization:** For fine-grained deepfake detection, we adopt a multi-label perspective and employ distinct loss functions tailored to enhance multi-label outcome predictions.
- **Manipulation order:** For manipulation ranking, we introduce a specialized loss term, designed to address the nuances of multi-label ranking challenges.



Problem formulation

Consider now a deepfake dataset $\mathcal{D} = \{(\mathbf{x}, Y)\}$ with totally L facial components to which photorealistic manipulations, where $\mathbf{x}$ is an image and $Y = \{l_j\}_{j=1}^{k}$ with $k \leq L$ is an ordered subset of $\{1, 2, \ldots, L\}$, indicating that the jth $(j \leq k)$ deepfake modification has been performed on the l_jth facial component. When Y is an empty set, it implies that $\mathbf{x}$ is a genuine facial image.

From Y two L-dimensional vectors $\mathbf{y} = (y_i)$ and $\mathbf{r} = (r_i)$

$$y_i = \begin{cases} 1, & \text{if } i = l_j \in Y; \\ 0, & \text{otherwise,} \end{cases}$$

$$r_i = \begin{cases} j, & \text{if } i = l_j \in Y; \\ L, & \text{otherwise,} \end{cases}$$

where $\mathbf{y}$ *is the standard multi-label binary vector and* $\mathbf{r}$ *is the corresponding rank vector. We realize the above definitions with a hands-on example.*

Feature extraction

For each training sample $(\mathbf{x}, Y)$, the CNN+FPN module transforms $\mathbf{x}$ into feature maps of size $\mathbb{R}^{w \times h \times d}$. We then form two vectors of tokens, including the patch tokens $U \in \mathbb{R}^{N \times d}$ and the learnable class tokens, $V \in \mathbb{R}^{L \times d}$.

Dual-granularity of vectors of tokens

$$U \xrightarrow{\phi} \widetilde{U} \in \mathbb{R}^{N \times d}, \quad V \xrightarrow{\phi} \widetilde{V} \in \mathbb{R}^{L \times d}.$$

The two sets of tokens are passed through the transformer encoder ϕ, which performs self-attention to correlate their features

Similarity values of each patch token to all other tokens

$$S = \max(\widetilde{U}\widetilde{U}^\top, 0) \in \mathbb{R}^{N \times N}, \quad \mathbf{u} = (u_1, u_2, \ldots, u_n),$$

where S is rectified into a nonnegative matrix such that all of its elements are in $[0, 1]$. It suffices to focus on the upper triangular part of S, excluding those on the diagonal. We arrange these entries of interest in ascending order of similarity value

MIL deepfake classification

With the sorted list $\mathbf{u}$ of similarity responses between patch tokens, we can consider the task of deepfake detection from the multiple instance learning (MIL) viewpoint:

- A face image $\mathbf{x}$ as a bag and the positive label 1 indicates that $\mathbf{x}$ is indeed fabricated as a deepfake one. In terms of the elements in $\mathbf{u}$, if $\mathbf{x}$ is a deepfake image, we expect to uncover that there exists at least one u_i (starting from the front end of $\mathbf{u}$) with a small value close to 0.
- On the other hand, a negative bag (*i.e.*, $\mathbf{x}$ is not a deepfake image) implies all u_i are close to 1.
- To incorporate the above observations into the model learning process, we introduce a *contrastive* formulation to realize the MIL concept for deepfake detection.

Contrastive MIL loss for deepfake classification

Assume that a deepfake image $\mathbf{x}$ results in the k smallest similarity responses on the front end of the sorted list $\mathbf{u}$. We propose to compute its probability of being deepfake by contrasting the average responses from the positive and negative distributions:

Probability of being deepfake

$$P(\mathbf{x}; k) = 2 \times \frac{\exp(u^{+}(k)/\tau)}{\exp(u^{+}(k)/\tau) + \exp(u^{-}(k)/\tau)} - 1, \quad P(\mathbf{x}) = \max_{1 \le k \le n} P(\mathbf{x}; k).$$

Contrastive MIL loss

$$\ell_{\mathrm{MIL}}(\mathbf{x}) = -J(Y) \log P(\mathbf{x}) - (1 - J(Y)) \log(1 - P(\mathbf{x})), \quad \ell_{\mathrm{Reg}}(\mathbf{x}) = \sum_{i=1}^{n} \|1 - u_i\|_2.$$

where $J(Y) = 1$ if a sample $(\mathbf{x}, Y)$ is a deepfake image, and 0, otherwise. In addition, for an authentic image $\mathbf{x}$, it is reasonable to expect that all the similarity responses u_i should be close to 1.

Multi-label loss for deepfake localization

The Transformer encoder ϕ generates, for each sample $(\mathbf{x}, Y)$, two sets of features from the patch tokens, $U \in \mathbb{R}^{N \times d}$ and the class tokens, $V \in \mathbb{R}^{L \times d}$. Our network model applies convolutions to U and then carries out average pooling to obtain the patch-token logits $\mathbf{f}^U = (f_i^U) \in \mathbb{R}^L$. In a similar way, we have the class-token logits $\mathbf{f}^V = (f_i^V) \in \mathbb{R}^L$. By independently applying a sigmoid function σ to each logit, we obtain two sets of multi-label predictions as

Multi-label prediction and BCE loss

$$P_i^{\mathcal{X}}(\mathbf{x}) = \sigma(f_i^{\mathcal{X}}) \in [0,1], \quad i = 1, \ldots, L, \quad \mathcal{L}_{\mathrm{BCE}}^{\mathcal{X}} = \sum\nolimits_{(\mathbf{x},Y)\in\mathcal{D}} \mathbf{1} \cdot \ell^{\mathcal{X}}(\mathbf{x}),$$

$$\ell_i^{\mathcal{X}}(\mathbf{x}) = -y_i \log P_i^{\mathcal{X}}(\mathbf{x}) - (1 - y_i) \log(1 - P_i^{\mathcal{X}}(\mathbf{x})).$$

where $\mathcal{X}$ can be replaced by U or V to respectively imply that the predictions are based on the features from patch tokens or class tokens. "$\cdot$" denotes inner product, $\mathbf{1}$ is all-ones vector.

Multi-label ranking loss for manipulation order

To begin with, we average the patch-token and the class-token logits to obtain $\mathbf{f} = (f_i) = (\mathbf{f}^U + \mathbf{f}^V)/2$. The fusion between the two streams gives rise to multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^{L}$. The main idea behind our formulation is as follows: by constructing a rank-aware loss term, the learned network model is expected to output multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^{L}$ that respect the rank order $\mathbf{r} = (r_i)$, implied by the given sample $(\mathbf{x}, Y) \in \mathcal{D}$.

Loss term for tackling multi-label ranking

$$\mathcal{L}_{\mathrm{Rank}} = \sum\nolimits_{(\mathbf{x},Y)\in\mathcal{D}} \mathbf{w}(\mathbf{r}, \{P_i(\mathbf{x})\}) \cdot \ell(\mathbf{x})$$

$$w_i = \begin{cases} \alpha, & \text{if } i \notin Y \ \wedge \ r_i > |Y|; \\ \alpha \times |o_i - r_i|, & \text{otherwise,} \end{cases}$$

where α is a hyperparameter to our method.

Total loss

To train the proposed network model for simultaneously carrying out deepfake classification and localization, our method considers the following total loss:

Classification loss

$$\mathcal{L}_{\mathrm{CLS}} = \sum_{(\mathbf{x},Y)\in\mathcal{D}} \ell_{\mathrm{MIL}}(\mathbf{x}) + (1 - J(Y))\,\ell_{\mathrm{Reg}}(\mathbf{x})$$

Multi-label loss

$$\mathcal{L}_{\mathrm{BCE}} = \mathcal{L}_{\mathrm{BCE}}^{U} + \mathcal{L}_{\mathrm{BCE}}^{V}$$

Total loss

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{CLS}} + \lambda_1\,\mathcal{L}_{\mathrm{BCE}} + \lambda_2\,\mathcal{L}_{\mathrm{Rank}}$$

Outline

1 Introduction

2 Method

3 Experiments

4 Conclusion

Experiment: Sequential deepfake manipulation

Method	**Seq-FaceComp Acc.**		**Seq-FaceAttr Acc.**	
	Multi-label (%)	Ranking (%)	Multi-label (%)	Ranking (%)
Multi-Cls*	78.32	69.66	85.14	66.99
DETR*	-	69.87	-	67.93
SeqFakeFormer*	-	72.13	-	67.99
Ours*	**82.31** ↑ 3.99	**73.72** ↑ 4.06	**86.42** ↑ 1.28	**68.82** ↑ 1.83
Multi-Cls†	79.54	69.75	88.23	66.66
DRN†	-	66.06	-	64.42
DETR†	-	69.75	-	67.62
MA†	-	71.31	-	67.58
Two-Stream†	-	71.92	-	66.77
SeqFakeFormer†	-	72.65	-	68.86
MMNet†	-	73.93	-	69.27
Ours†	**84.12** ↑ 4.58	**74.54** ↑ 4.79	**90.45** ↑ 2.22	**69.58** ↑ 2.92
Ours‡	**84.36** ↑ 4.82	**74.97** ↑ 5.22	**90.74** ↑ 2.51	**70.02** ↑ 3.36

Experiment: Binary deepfake classification

Method	Intra-testing AUC		Cross-testing (Train on FF++ only) AUC			
	FF++ (%)	CDF (%)	CDF (%)	WDF (%)	DFDC (%)	DFD (%)
Xception	96.30	99.73	61.80	62.72	48.98	87.86
EifficientNet-B4	99.70	99.81	64.29	63.83	-	-
Multi-Att†	99.29	99.94	67.44	59.74	-	-
SPSL*	96.91	-	76.88	-	66.16	-
RECCE*	99.32	99.94	68.71	64.31	69.06	-
Face X-Ray*	99.17	-	80.58	-	**80.92**	95.40
LRL*	99.46	-	78.26	-	76.53	89.24
SBIs†	99.64	93.74	**93.18**	-	72.42	97.56
SBIs‡	99.72	95.68	89.12	70.56	71.08	97.34
Ours‡	**99.82** ↑ 3.52	**99.98** ↑ 0.25	91.56 ↑ 29.76	**73.41** ↑ 10.69	75.17 ↑ 26.19	**97.88** ↑ 10.02

Ablation studies

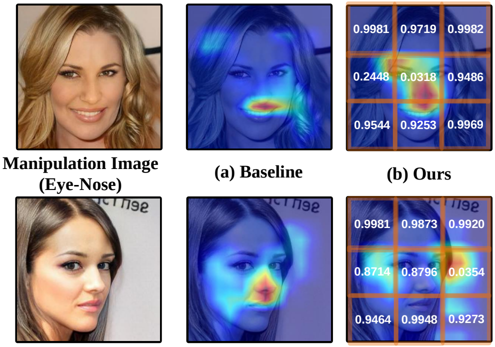
Model	$\mathcal{L}^{U}_{\mathrm{BCE}}$	$\mathcal{L}^{V}_{\mathrm{BCE}}$	$\mathcal{L}^{U}_{\mathrm{CLS}}$	$\mathcal{L}^{U}_{\mathrm{Rank}}$	**Seq-FaceComp Acc.** Ranking (%)
I	✔				52.43
II	✔		✔		54.21
III	✔			✔	72.52
IV	✔		✔	✔	73.43
V		✔	✔	✔	72.87
VI	✔	✔	✔	✔	**74.54**

(a) The proposed losses

Manipulation Components	**Nose**	**Eye**	**Eyebrow**	**Lip**	**Hair**
Baseline (Multi-Cls)	0.41	0.38	0.35	0.46	0.42
Ours	**0.72**	**0.61**	**0.66**	**0.75**	**0.74**

(b) The correlation

Qualitative visualization



Manipulation Image (Eye-Nose) **(a) Baseline** **(b) Ours**

Manipulation Image (Eyebrow) **(a) Baseline** **(b) Ours**

Outline

1 Introduction

2 Method

3 Experiments

4 Conclusion

Conclusion

This work aims to develop a unified framework that comprehensively addresses both sequential deepfake manipulations and binary deepfake classification.

- We propose to decompose the general deepfake problem into three parts: deepfake classification, deepfake localization, and manipulation order
- The proposed approach introduces novel contrastive MIL learning and explores multi-label ranking to elegantly tackle all three subtasks.
- The extended experimental results demonstrate the effectiveness and flexibility of the proposed formulation in dealing with the various deepfake application scenarios. The provided analyses are also reasonable to support the usefulness of our method.

Thanks for your attention!