Explaining Deepfake Detection by Analysing Image Matching

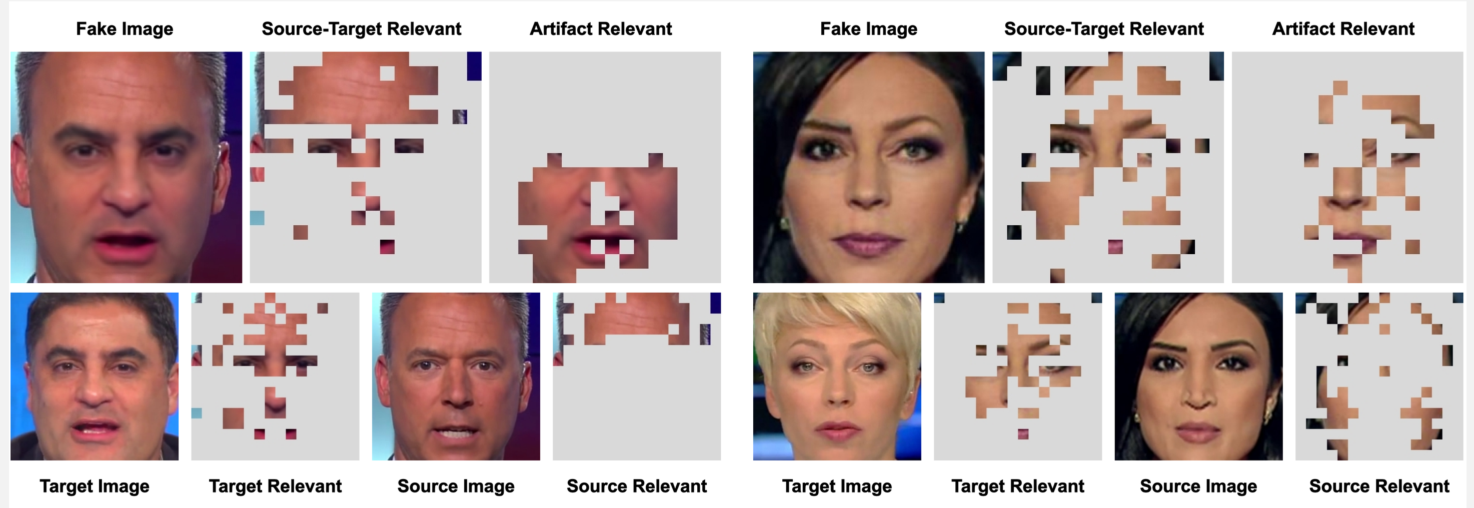
Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, Renhe Ji

ECCV TEL AVIV 2022

Introduction

- Considering it as a binary classification task, deepfake detecion models have achieved great success in detecting various manipulated media.
- In this paper, we focus on understanding how these models learn artifact features of images when just supervised by binary labels (real/fake) from the novel perspective of ***image matching***.

Image matching



- ➢ The face of the source image is manipulated with representations of the target image to generate the corrseponding fake image.
- ➢ The above **f**ake, **s**ource and **t**arget image are considered as matching images, termed as the FST-Matching.

Algorithm

From the perspective of FST-Matching, we propose three hypotheses and design several metrics to verify them.

Artifact representations for deepfake detection models

> ***Hypothesis 1***: *Deepfake detection models indicate real/fake images based on visual concepts that are neither source-relevant nor target-relevant, that is, considering such visual concepts as artifact-relevant.*

- ➢ We train a deepfake detection encoder $v_d(\cdot)$, a source encoder $v_s(\cdot)$ and a target encoder $v_t(\cdot)$ to indicate the artifact, source, and target relevant visual concepts.
- ➢ The source/target encoder $v_s(\cdot)/v_t(\cdot)$ learns to classify each fake image and the corresponding source/target image as the same category.
- ➢ We use the Shapley value [1] to evaluate regional contributions $\varphi_{v_d}/\varphi_{v_s}/\varphi_{v_t}$ of visual concepts to the prediction of each encoder.

[1]Shapley, L.S.: A value for n-person games, contributions to the theory of games, 2, 307–317 (1953)

- ➢ To verify the hypothesis, we design a metric to evaluate the intensities of the intersections between these visual concepts.

$$Q_\tau = \frac{(1-M_\tau)\cdot\varphi_{v_d}}{\sum[1-M_\tau]} - \frac{M_\tau\cdot\varphi_{v_d}}{\sum M_\tau}$$

where $M_\tau = I(max(\varphi_{v_s}, \varphi_{v_t}) > \tau)$ denotes the most source/target relevant visual concepts.

- ➢ $Q_\tau > 0$ represents that artifact-relevant visual concepts are more related to source/target-irrelevant visual concepts and vice versa.

Learning the artifact representations

> ***Hypothesis 2***: *Besides the supervision of binary labels, deepfake detection models implicitly learn artifact-relevant visual concepts through the FST-Matching in the training set.*

- ➢ To verify the hypothesis, we train two models with paired and unpaired training set, which are downsampled from original dataset.
- ➢ In the paired training set, the real images are only the corresponding source images and target images of fake images.
- ➢ In the unpaired images, the real images do not correspond to any fake images but are of the same number as the paired training set.

Vulnerability of artifact representations to video compression

> ***Hypothesis 3***: Implicitly learned artifact visual concepts through the FST-Matching in the raw training set are vulnerable to the video compression.

- ➢ To verify the hypothesis, we design the stability metric of implicitly learned artifact visual concepts to the video compression.

$$\delta_{v_d} = E_{cmp\in\{c23,c40\}}[cos(\varphi_{v_d}^{cmp}, \varphi_{v_d}^{raw})]$$

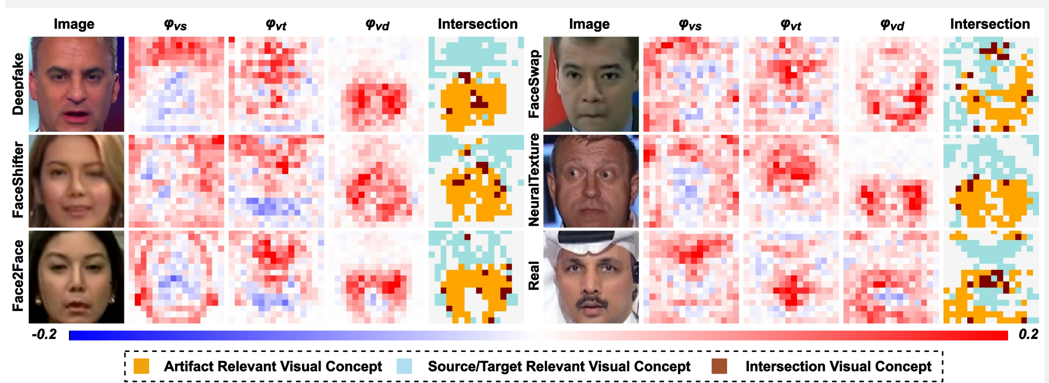
where $\varphi_{v_d}^{cmp}/\varphi_{v_d}^{raw}$ represents the regional contributions to the predictions of the v_d when tested on the compressed/raw images.

- ➢ We also evaluate the stability of the learned source/target visual concepts for v_s/v_t on compressed videos for more comparisons.

Experiments

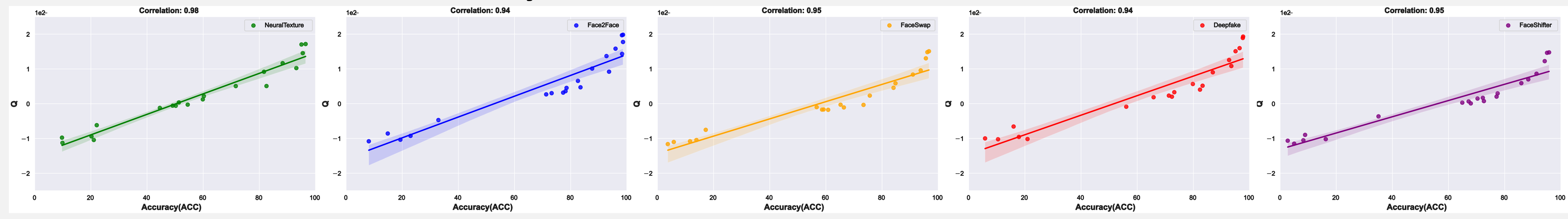
Verification of hypothesis 1

- ➢ Qualitative analysis



Artifact-relevant visual concepts barely have intersections with source/target-relevant visual concepts.

- ➢ Quantitative analysis



Values of Q and accuracy of models are positively correlated.

Verification of hypothesis 2

- ➢ Deepfake detection performance analysis

Models	Forgery Methods	Baseline		Pair		Unpair	
		ACC	*AUC*	*ACC*	*AUC*	*ACC*	*AUC*
ResNet-18 [18]	FaceSwap [21]	98.93	100	97.50	99.91	53.93	75.41
	Face2Face [40]	96.79	99.43	97.14	99.27	64.29	85.74
	FaceShifter [23]	99.29	99.99	97.14	99.82	81.07	93.03
	Deepfake [11]	98.21	100	97.50	99.87	69.64	86.51
	NeuralTexture [39]	90.71	98.89	95.71	98.73	60.00	76.60
Efficient-b3 [38]	FaceSwap [21]	100	100	99.64	100	77.50	87.51
	Face2Face [40]	99.29	99.77	99.29	99.72	81.79	93.36
	FaceShifter [23]	99.29	99.93	99.29	99.96	84.29	96.10
	Deepfake [11]	100	100	100	100	85.36	97.81
	NeuralTexture [39]	99.29	99.85	98.93	99.56	82.86	92.30

Models trained on the paired/unpaired training set achieved similar/worse performance to the models on the full dataset.

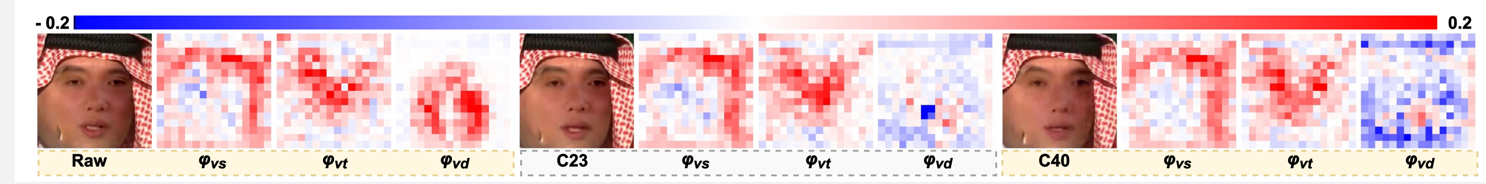
- ➢ Comparison of the proposed metric Q_τ

Models trained on the paired training set also have larger values of Q_τ than models trained on the unpaired training set.

Verification of hypothesis 3

- ➢ Comparison of the stability metric $\delta_{v_s}/\delta_{v_t}/\delta_{v_d}$

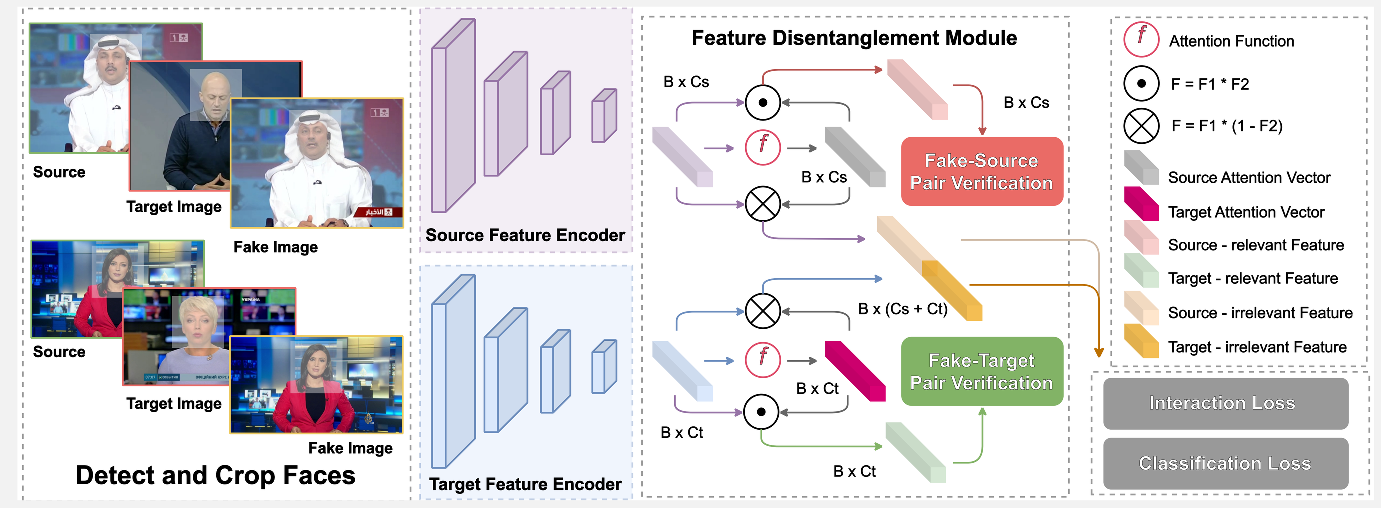
Visual Concept	Forgery Methods (δ)				
	FaceSwap	Face2Face	FaceShifter	Deepfake	NeuralTexture
Source	0.73	0.74	0.73	0.74	0.74
Target	0.73	0.76	0.71	0.75	0.76
Artifact (Baseline)	0.17	-0.02	0.14	-0.15	-0.14



Source/target visual concepts show better consistency than the implicitly learned artifact visual concepts to compression.

FST-Matching Deepfake Detection Model

Based on our analysis, we propose a novel method to boost the performance of deepfake detection on compressed videos.



Our method disentangles source/target-irrelevant representations from source/target visual concepts to indicate images, achieving great performance on compressed videos.