Why Fake ? Unveiling the Semantic Vocabulary of Deepfake Detectors

Vazgken Vanian; Alexandros Doumanoglou; Dimitris Zarpalas
Information Technologies Institute (ITI) | Centre For Research and Technology HELLAS (CERTH)

CVPR JUNE 3-7, 2026 DENVER COLORADO

1. MOTIVATION

The Problem

Deepfake detectors usually say only:

✔ Real
✖ Fake

…but **not WHY.**

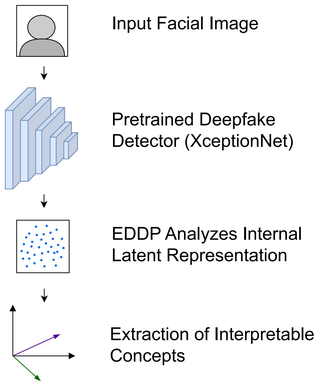
Our Goal

Create trustworthy explanations for deepfake detectors by uncovering:

1. which concepts, detectors learn
2. where these concepts appear
3. and how they affect decision

2. METHOD OVERVIEW

Post-hoc Concept Extraction with EDDP



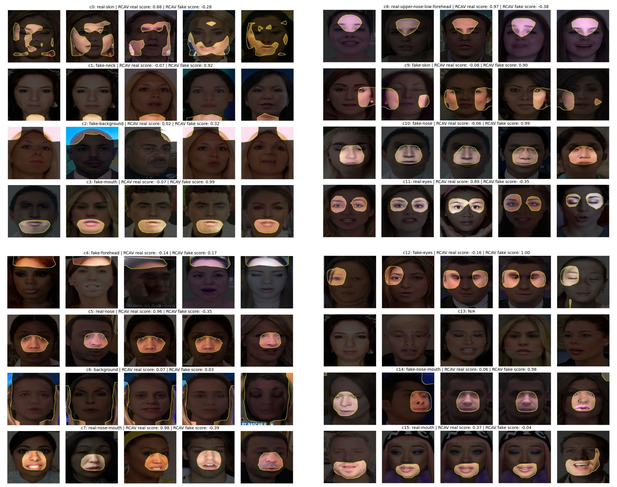
3. GLOBAL UNDERSTANDING

Concept Identification

To ground concepts into human semantics we perform

- RCAV analysis
- IoU analysis for face - concept overlap
- Dataset distribution analysis
- Manual visual inspection

Concept Examples



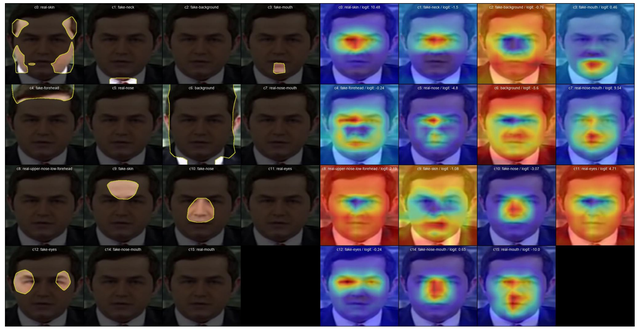
4. CONCEPT FAITHFULNESS ASSESSMENT

Concept-Transfer Accuracy	Correcting Misclassified Samples
87.34%	99.8%

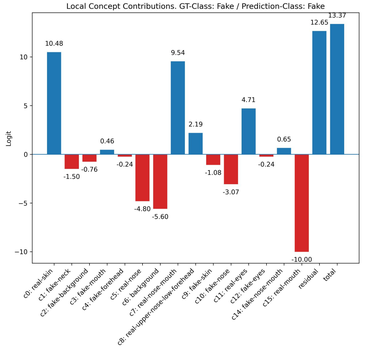
5. LOCAL EXPLANATIONS

Concept Presence Maps & Concept Contribution Maps

- Concept Presence Maps show where each concept is active on a sample image
- Concept Contribution Maps show pixel contributions to the output prediction



Logit Decomposition based on Concept Contribution Maps



6. COUNTERFACTUAL ANALYSIS

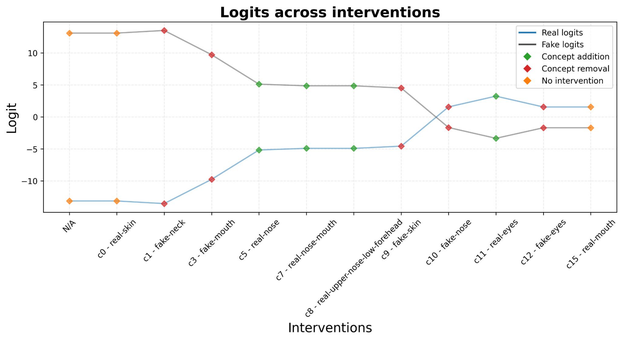
Intervention Methodology

We have access to the concept directions and therefore we can intervene on them based on desired target class. The intervention operations we can perform are

- Addition of concepts associated with the target class
- Removal of concepts associated with the opposing target class
- No intervention if concept is already present (or absent) based on target class

Single Sample Intervention Trace

We perform a series of interventions and successfully flip the prediction from fake to real



7. SUMMARY

In summary our contributions are the following:

- First post-hoc concept-based XAI application for deepfake detection
- Global concept explanations & understanding
- Faithfulness explanations
- Local concept based explanations
- Counterfactual analysis and intervention mechanism

Contact

Vazgken Vanian
Center for Research and Technology Hellas
Email: vvanian@iti.gr

PROJECT PAGE

CODE

This research has been supported by the European Commission funded program DETECTOR, under Horizon Europe Grant Agreement 101225942.