Deepfake Caricatures: Human-guided Motion Magnification Improves Deepfake Detection by Humans and Machines

Camilo Fosco*, Emilie Josephs*, Alex Andonian, Allen Lee, Xi Wang, Aude Oliva
MIT CSAIL

Problem

Deepfake videos threaten the integrity of our digital ecosystem.

- Mitigating the spread of deepfakes requires 1) identifying them and 2) alerting the human user to their presence.
- Current detection methods consist of black-box computer vision models, which do not consider how best to alert humans.

Can we develop computer-based deepfake detection methods that also effectively reveals fake videos to human users?

We introduce CariNet, a deepfake detection model that relies on selective motion magnification to improve both human and machine detection.

CariNet has 2 main components:
- A deepfake detection module that learns to attend to artifacts through human and machine supervision.
- A caricature module which amplifies artifacts on fake videos to facilitate human detection, creating deepfake caricatures.

Model Details

- **Human Artifact Maps**
  - 1015 fake videos from DFDC
  - Human annotations of regions with artifacts
  - 10 annotations per video, 1572 distinct workers
  - Artifact maps obtained by averaging annotations across participants

- **Model Results**
  - **Human Artifact Maps**
    - Annotation interface
    - Annotation examples
    - Artifact maps for training

- **Do Caricatures improve detectability of fake videos by humans?**
  - **Method**
    - 600 videos: 200 each real, standard fake, caricature
    - Unspeeded 2AFC response: Real or Fake?
    - 6 presentation times between 300ms and 5000ms
    - 10 participants per video per timepoint
    - 178 participants performed 223 HITs on Amazon Mechanical Turk
    - Included “engagement probe” trials to gauge online engagement
  - **Results**
    - Caricatures led to better correct detection than standard fakes overall (main effect of condition p<0.001, interaction with time F(2,2429)=5.13, p<0.001)
    - This difference is significant at all timepoints tested (a priori, Bonferroni corr.)
    - The Caricature advantage is 17% at fastest timepoint, and increases with longer exposure, up to 42%
    - Caricature led to better detection even for participants demonstrating low attentiveness to the tasks.

- **Model frameworks that jointly consider to detect and communicate the authenticity of a video may have larger societal impact.**

- **Comparison against other detectors**
  - 1015 fake videos from DFDC
  - Human annotations of regions with artifacts
  - 10 annotations per video, 1721 distinct workers
  - Artifact maps obtained by averaging annotations across participants

- **Behavioral Results**
  - **Do CariNet improve performance for low-engagement participants?**
  - **Results**
    - Caricature led to better detection even for participants demonstrating low attentiveness to the tasks.

Contact: camilolu@mit.edu, ejpaoths@mit.edu