What’s wrong with this video?
Comparing Explainers for Deepfake Detection

Samuele Pino
samuele.pino@mail.polimi.it
CSE Track
Goals

Automatic classifiers can already detect if a video is real or fake.

But can we understand the reason why a video is detected as fake?

In this work we develop, extend and compare explanation techniques for deepfake detection.
Overview
Overview

• Introduction

Original showing Alison Brie

Deepfake showing Jim Carrey instead of Brie

[Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, Tim C. Kietzmann, Deepfakes: Trick or treat?, 2020]
Overview

• Introduction

• Approach
Overview

• Introduction

• Approach

• Experiments
Overview

- Introduction
- Approach
- Experiments
- Results
Introduction
• Replacing faces in videos

[Deepfake Detection Challenge, 2019]
Deepfakes

- Replacing faces in videos
- Deep learning technique

https://mc.ai/auto-encoder-in-biology/
Deepfakes

• Replacing faces in videos
• Deep learning technique
• Initially to generate adult contents

r/deepfakes has been banned from Reddit
Deepfakes

- Replacing faces in videos
- Deep learning technique
- Initially to generate adult contents
- No official implementation
Deepfakes

Why is it important to detect them?
Deepfakes

Why is it important to detect them?

• disinformation

[https://leidenlawblog.nl/articles/wash-your-hands-often-and-your-newsfeed-even-more-disinformation-in-covid-19-times]
Deepfakes

Why is it important to detect them?

- disinformation
- online abuse

Deepfakes

Why is it important to detect them?

• disinformation
• online abuse
• financial fraud
Deepfakes

Why is it important to detect them?

- disinformation
- online abuse
- financial fraud
- law enforcement

[http://www.forensicsciencesimplified.org/av/how.html]
Deepfakes

Let’s build our deepfake!
Deepfakes

Deepfakes

Explainability problem

• Also detectors use deep learning

• Should we trust them?
• «correct prediction for the correct reason»

• Complexity-interpretability trade off
Explainability problem

- Why do we need it:
  - law enforcement
  - journalists
  - dispute resolution in social media
Approach
Approach: overview

- **preprocessing**

- **detection models**
  - EfficientNet B4
    - Fake score
  - EfficientNet B4 LTPA
    - Fake score
  - EfficientNet Bonetini
    - Fake score

- **explainers**
  - GradCAM
    - Activation map 9x9
  - Kernel SHAP
    - SHAP values map (100 segments)
  - Attention map 12x12
  - Attention map 28x28

- **output explanations**
Approach: detection model

- EfficientNet as a backbone CNN
- Powerful and lightweight
- Winner’s solution in Deepfake Detection Challenge

[Ecientnet: Rethinking model scaling for convolutional neural networks, Tan and Le, 2019]
Approach: explainers

- Black-box (SHAP)

[A unified approach to interpreting model predictions, Lundberg and Lee, 2017]
Approach: explainers

- Black-box (SHAP)
- White-box (GradCAM)

[Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Ramprasaath et al., 2019]
Approach: explainers

- Black-box (SHAP)
- White-box (GradCAM)
- Self-attention (LTPA, Bonettini)

[Learn to pay attention, Jetley et al., 2018]
[Video face manipulation detection through ensemble of cnns, Bonettini et al., 2020]
Approach: SHAP

• Model-agnostic

• Kernel SHAP for image classification
  • Segmentation
  • SHAP values assignment

• Extension: 3D segmentation for video classification
Approach: GradCAM

• Class Activation Mapping

• Neural network gradients

• Binary classification extension
Approach: self-attention

- Learn To Pay Attention (LTPA)
  - 3 attention maps
Approach: self-attention

- Learn To Pay Attention (LTPA)
  - 3 attention maps
- Bonettini
  - Single attention map
Experiments
Experiments: setup

• Dataset (DFDC)

• Training

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (balanced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EfficientNet B4, 224×224</td>
<td>0.888</td>
</tr>
<tr>
<td>EfficientNet B4, 380×380</td>
<td>0.931</td>
</tr>
<tr>
<td>EfficientNet B7, 224×224</td>
<td>0.906</td>
</tr>
<tr>
<td>EfficientNet B7, 380×380</td>
<td>0.926</td>
</tr>
<tr>
<td>EfficientNet B4, LTPA, 224×224</td>
<td>0.879</td>
</tr>
<tr>
<td>EfficientNet B4, LTPA, 380×380</td>
<td>0.929</td>
</tr>
<tr>
<td>EfficientNet B7, LTPA, 224×224</td>
<td>0.893</td>
</tr>
<tr>
<td>EfficientNet B7, LTPA, 380×380</td>
<td>0.904</td>
</tr>
</tbody>
</table>
Experiments: explanations

(a) GradCAM
(b) LTPA lv. 2
(c) SHAP
(d) Bonettini
Experiments: evaluation

• Metrics

  • Variance
    \[ V = \text{avg}_{f \in \text{frames}}(\text{var}(f)) \]
  
  • Inter-frame consistency
    \[ \tau = \text{avg}_{f \in \text{frames}}(\text{PCC}_{f,f+1}) \]
  
  • Intra-frame consistency
    \[ \rho = \text{avg}_{f \in \text{frames}} \left( \frac{\text{avg}_{s \in S}(a_{0.1\cdot s}(f))}{a_{0,0}(f)} \right) \]
  
  • Centredness
    \[ \mu = \text{avg}_{f \in \text{frames}} \left( \frac{I(\text{inner 50\%})}{I(\text{full frame})} \right) \]
Experiments: evaluation

- User study
  - 20 real and 20 fake videos
  - 20 sections
  - 2 questions per section
Results
Results: metrics

• Average over 58 fake videos

• GradCAM performs best

<table>
<thead>
<tr>
<th></th>
<th>( V )</th>
<th>( \tau )</th>
<th>( \rho )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonettini</td>
<td>0.0951</td>
<td>0.7390</td>
<td>0.1262</td>
<td>0.5286</td>
</tr>
<tr>
<td>GradCAM</td>
<td>0.0135</td>
<td>0.8756</td>
<td>0.7489</td>
<td>0.8666</td>
</tr>
<tr>
<td>LTPA</td>
<td>0.0108</td>
<td>0.7991</td>
<td>0.3333</td>
<td>0.6386</td>
</tr>
<tr>
<td>SHAP</td>
<td>0.0302</td>
<td>0.4496</td>
<td>0.2326</td>
<td>0.7348</td>
</tr>
</tbody>
</table>
Results: user study

- Number of answers: 67
- Accuracy: 85%
- Preferred explainer: SHAP
- Statistical “sign test” for validation
- Preference dependent on video

<table>
<thead>
<tr>
<th>Number of answers</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen used</td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>43%</td>
</tr>
<tr>
<td>Small</td>
<td>57%</td>
</tr>
<tr>
<td>Are you familiar with deepfakes?</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>37%</td>
</tr>
<tr>
<td>Heard of it</td>
<td>33%</td>
</tr>
<tr>
<td>No</td>
<td>30%</td>
</tr>
<tr>
<td>Correct video identification</td>
<td>85%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explainer choices</th>
<th>GradCAM</th>
<th>SHAP</th>
<th>LTPA</th>
<th>Bonettini</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>165</td>
<td>221</td>
<td>137</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>23%</td>
<td>31%</td>
<td>19%</td>
<td>26%</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion

• We implemented and extended 4 explanation techniques

• We defined intrinsic and extrinsic metrics

• We empirically compared the explainers based on them

• We performed a user survey

• Human perception is not always aligned with objective metrics
Future work

- Captioning explanation maps
- Weighting explanations on classifier’s confidence
Thank you