PRESENTATION ON RECURRENT CONVOLUTIONAL STRATEGIES FOR FACE MANIPULATION DETECTION IN VIDEOS

Pratikkumar.Prajapati@sjsu.edu

Apr/26/2020
OVERVIEW

- The original paper at [1]
- Features: Image + Temporal Information
- Model: CNN + RNN
- Dataset: FaceForensics++ (FF++) [2]
  - 1000 videos = 720 training + 140 validation + 140 test
- Accuracy:
  - AUC of 96.9% DeepFake, LQ
  - AUC of 96.3% FaceSwap, LQ
OVERALL ARCHITECTURE

Figure 1: The overall pipeline is a two step process. The first step detects, crops and aligns faces on a sequence of frames. The second step is manipulation detection with our recurrent convolutional model.
PREPROCESSING

- Uses masks provided by [2] to crop the face region

- Face alignment
  1. Landmark-based alignment
     - seven sparse points of the face are used.
     - corners of the eyes, the tip of the nose, and corners of the mouth.
  2. Spatial Transformer Network (STN)
     - performs spatial alignment of data with learnable affine transformation parameters.
     - a differentiable module which can be inserted into other CNN models to learn about features and landmarks.
MANIPULATION DETECTION

- Inputs are sequence of frames from the target video
- CNN
  - To learn about features of the image (frame(s) of video)
  - ResNet and DenseNet as backbone
- RNN
  - Uses GRU to exploit temporal discrepancies across frames.
  - Temporal discrepancies are expected to occur in images, since manipulations are performed on a frame-by-frame basis.
- The backbone n/w is first trained on FF++ training split minimizing cross-entropy loss to train to detect real from synthesized faces.
- The backbone is then extended with RNN and re-trained end-to-end.
- Adam optimizer
RNN TRAINING STRATEGIES

1. A single RNN on top of final features learnt from backbone n/w.
2. Multiple RNN at different levels of hierarchy of the backbone net.
   - To utilize micro, meso and macroscopic features
## RESULTS OF EXPERIMENTS (1/2)

<table>
<thead>
<tr>
<th>Manipulation</th>
<th>Frames</th>
<th>FF++  [34]</th>
<th>ResNet50</th>
<th>DenseNet</th>
<th>ResNet50 + Alignment</th>
<th>DenseNet + Alignment</th>
<th>ResNet50 + Alignment + BiDir</th>
<th>DenseNet + Alignment + BiDir</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deepfake</td>
<td>1</td>
<td>93.46</td>
<td>94.8</td>
<td>94.5</td>
<td>96.1</td>
<td>96.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-</td>
<td>94.6</td>
<td>94.7</td>
<td>96.0</td>
<td>96.7</td>
<td>94.9</td>
<td>96.9</td>
</tr>
<tr>
<td>Face2Face</td>
<td>1</td>
<td>89.8</td>
<td>90.25</td>
<td>90.65</td>
<td>89.31</td>
<td>87.18</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-</td>
<td>90.25</td>
<td>89.8</td>
<td>92.4</td>
<td>93.21</td>
<td>93.05</td>
<td>94.35</td>
</tr>
<tr>
<td>FaceSwap</td>
<td>1</td>
<td>92.72</td>
<td>91.34</td>
<td>91.04</td>
<td>93.85</td>
<td>96.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-</td>
<td>90.95</td>
<td>93.11</td>
<td>95.07</td>
<td>95.8</td>
<td>95.4</td>
<td>96.3</td>
</tr>
</tbody>
</table>

**Table 1:** Accuracy for manipulation detection across all manipulation types. DenseNet with alignment and bidirectional recurrent network is found to perform best. FF++ [34] is the baseline in these experiments.
RESULTS OF EXPERIMENTS (2/2)

<table>
<thead>
<tr>
<th>Manipulation</th>
<th>Base</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Spatial Transformer</td>
</tr>
<tr>
<td>Deepfake</td>
<td>96.9</td>
<td>91.7</td>
</tr>
<tr>
<td>Face2Face</td>
<td>94.35</td>
<td>87.46</td>
</tr>
<tr>
<td>FaceSwap</td>
<td>96.3</td>
<td>93.2</td>
</tr>
</tbody>
</table>

Table 2: Results on using variations to the recurrent convolutional architecture. Both spatial-transformer networks and multi-recurrent networks exhibit a decline in performance.
REFERENCES
