Preserving Fairness Generalization in Deepfake Detection (CVPR’24)

1. Lin Li, Xinan He, Yan Ju, Xin Wang, Feng Ding, Shu Hu
1Purdue University (lin1785, hu968@purdue.edu)
2Nanchang University (shuhu_fengding@ncu.edu.cn)
3University at Buffalo, State University of New York yanju@buffalo.edu
4University at Albany, State University of New York xwang@albany.edu

Motivation and Introduction

1. Deepfakes, created through advanced AI techniques, are highly realistic media that can pose serious threats like misinformation and political manipulations.
2. Unfairness in Deepfakes: Detection models have unfair performance disparities among demographic groups, such as race and gender.
3. Fairness Generalization can guarantee fair detection under intra-domain and cross-domain scenarios (detect deepfakes generated by unknown forgers). Enhance trustworthiness of AI systems and strengthen AI security.

Methodology

Objective: Train a fair deepfake detector using $S$ can then generalize to an unseen dataset while maintaining high detection accuracy.

Disentanglement Learning

\[ L_{dis} = \frac{1}{n} \sum_{i=1}^{n} [L_{cls} + \rho_1 L_{con} + \rho_2 L_{rec}], \]

\[ \rho_1, \rho_2 \text{ are trade-off hyperparameters.} \]

Classification Loss $L_{cls}$: disentangle demographic, forgery features.

Contrastive Loss $L_{con}$: enhance the encoder’s representation capabilities.

Reconstruction Loss $L_{rec}$: ensure the reconstructed image and the original images are consistent at the pixel level.

Fair Learning

\[ L_{fair} = \min_{\theta \in \Theta} \bar{\eta} + \frac{1}{|J|} \sum_{j=1}^{|J|} |L_j - \eta|_1, \]

$J$ represents user-defined subgroups e.g., Male-Asian or Female-Black. $|J|$ represents the size of set $J$, $L_j$ is the subgroup loss. $\alpha \in (0,1)$ is user-defined hyperparameter.

Optimization

 Flatten loss landscape with sharpness-aware minimization method.

\[ \min_{\theta} (L_{dis} + \lambda L_{fair})(\theta + \epsilon), \]

$\theta$ is model weights, $\epsilon$ is obtained through perturbing $\theta$ and calculating the gradient of $L$. The intuition is that the perturbation along the gradient norm direction increases the loss value significantly and then makes the model more generalizable in terms of fairness.

Results

Performance of Fairness Generalization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Fairness Metric (%)</th>
<th>Detection Metric (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF++</td>
<td>On1929</td>
<td>82.3</td>
<td>67.6</td>
</tr>
<tr>
<td>DFDC</td>
<td>On1929</td>
<td>57.2</td>
<td>44.0</td>
</tr>
<tr>
<td>CelebDF</td>
<td>On1929</td>
<td>75.7</td>
<td>64.0</td>
</tr>
</tbody>
</table>

3. Fairness Generalization can guarantee fair detection under intra-domain and cross-domain scenarios (detect deepfakes generated by unknown forgers). Enhance trustworthiness of AI systems and strengthen AI security.

Visualization of Loss Landscape

- With flattening, the landscape becomes smoother, suggesting an easier optimization path, potentially leading to better training and generalization.

Visualization of the Saliency Map

- Our method achieves the best performance on all datasets with different settings of backbones.

Conclusion & Future Work

- We propose the first method to improve fairness generalization in deepfake detection by addressing features, loss, and optimization.
- Our method outperforms state-of-the-art approaches in preserving fairness generalization.

In the future, we aim to preserve fairness generalization while detecting images generated by diffusion models or GANs.

Acknowledgements. This work is accepted by CVPR 2024. Shu Hu is supported by the Purdue University start-up grant.