Research

Deceptiscan: Comparative Analysis of Deepfake Detection Models

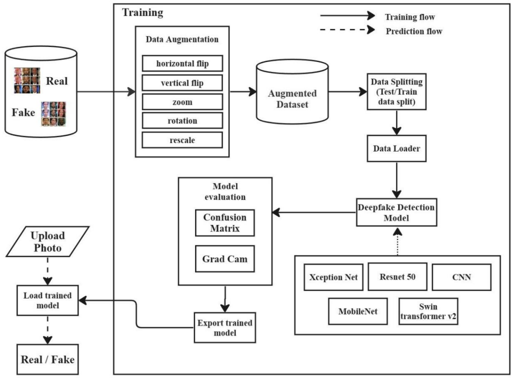
Om Jannu , Tushar Padhy , Vendra Sekar
Department of Information Technology, The Bombay Salesian Society's
Don Bosco Institute of Technology Mumbai - 400070

2023 - 2024

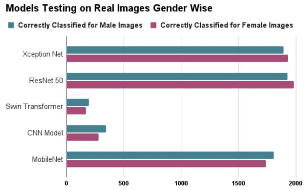
Abstract

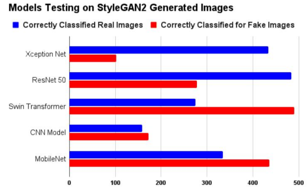
Deepfakes, which leverage advanced machine learning techniques such as generative adversarial networks (GANs), pose a significant threat to the integrity of visual content and raise concerns about misinformation and identity theft. This research provides a comparative analysis of various deepfake detection models, aiming to dissect their strengths and weaknesses. Notable architectures like Xception and ResNet50 exhibit high accuracy, precision, and recall with minimal gender bias. However, the Swin Transformer, while excelling in fake image detection, faces challenges with real images, suggesting potential bias. The CNN model demonstrates subpar performance, emphasizing limitations in classifying both fake and real images effectively. MobileNet shows moderate overall performance but maintains balanced precision and recall. Results of the ”Model Testing on Real Images Gender Wise” test case show that Xception achieved high accuracy, correctly classifying 94.65% of male images and 96.60% of female images. In the ”Model Testing on StyleGAN2 generated images” test case, ResNet50 also demonstrated strong performance, correctly classifying images with an accuracy of 96%. The study recommends an ensemble approach to combine model strengths and address individual weaknesses. Future work should focus on refining model architectures, exploring ensemble strategies, and mitigating biases in real image detection.

System Architecture



Results


Models Testing on Real Images Gender Wise


Models Testing on StyleGAN2 Generated Images

Model	Real Images	Fake Images	Overall Accuracy	Gender bias	Precision	Recall	F1-Score
Xception Net	81%	65%	73%	Low	0.95	0.95	0.95
Resnet 50	97%	60%	78%	Low	0.96	0.96	0.96
Swin Transformer v2	54%	98%	76%	High	0.61	0.60	0.61
Cnn Model	45%	55%	50%	High	0.61	0.60	0.61
MobileNet	65%	85%	75%	Moderate	0.78	0.78	0.78

Methodology

Data Collection

Our research utilized diverse datasets for training our models, ensuring comprehensive learning. These datasets included:

- **Deepfake and Real Images Dataset:** Obtained from Zenodo, sourced specifically from the OpenForensics Dataset.
- **This Person Does Not Exist Dataset:** Consists of 10,000 images of person faces generated by the "This Person Does Not Exist" website, offering styleGan2 samples.
- **140k Real and Fake Faces Dataset:** Gathered from Kaggle, contains 70,000 real faces from the Flickr dataset by Nvidia, alongside 70,000 fake faces sampled from the 1 Million FAKE faces generated by StyleGAN.

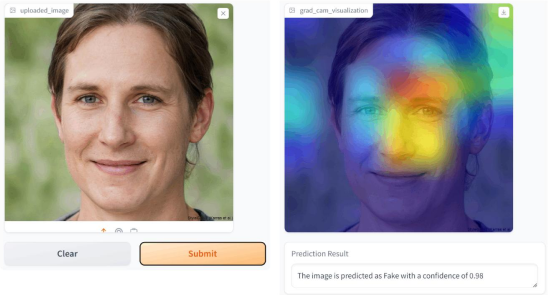
Testing Methodology

For evaluating the performance of our models, we conducted tests using real-world scenarios. Our testing strategy comprised two primary methods:

- **Deepfake Detection:** This method aimed to evaluate the models' proficiency in distinguishing between genuine images and those generated by StyleGAN2. We tested the models' effectiveness in identifying synthetic manipulations in images, crucial for digital content verification and security.
- **Gender-based Classification:** Through this method, we examined the accuracy of the models in classifying real images based on gender. This test measured the precision in gender identification, important for demographic analysis and

Analysis

Our trained model has effectively identified the image as a deepfake. To gain insight into the model's decision-making process, we utilize Grad-CAM visualization. This technique highlights significant regions in red, indicating facial irregularities, unnatural lighting, or background artifacts—typical indicators of manipulation. Additionally, green and yellow areas provide supplementary information, while blue regions denote insignificance. This helps us understand the model's thought process for deepfake detection. By iteratively analyzing these visualizations, we can potentially refine the model's ability to identify increasingly sophisticated deepfakes.



Prediction Result

The image is predicted as Fake with a confidence of 0.98

References

[1] V. Gupta, “vidhig/deepfake-image-detection,” GitHub, Apr. 21, 2022. [Online]. Available: https://github.com/vidhig/deepfake-image-detection

[2] C. Tan, Y. Zhao, S. Wei, et al., “Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection,” Accessed: Oct. 15, 2023. [Online]. Available:https://openaccess.thecvf.com/content/CVPR2023/papers/Tan_Learning_on_Gradients_Generalized_Artifacts_Representation_for_GAN-Generated_Images_Detection_CVPR_2023_paper.pdf

[3] J. Pu, N. Mangaokar, B. Wang, et al., “NoiseScope: Detecting Deep- fake Images in a Blind Setting,” Annual Computer Security Applications Conference, Dec. 2020, doi: https://doi.org/10.1145/3427228.3427285

[4] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” arXiv.org, 2018. Available: https://arxiv.org/abs/1812.04948

[5] H. Li, B. Li, S. Tan, et al., “Identification of deep network generated images using disparities in color components,” Signal Processing, vol. 174, p.107616, Sep. 2020, doi: https://doi.org/10.1016/j.sigpro.2020.107616

[6] U. Ojha, Y. Li, and Y. Lee, “Towards Universal Fake Image Detectors that Generalize Across Generative Models.” Accessed: Oct.15, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/papers/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.pdf

[7] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, “Deep Fake Image Detection Based on Pairwise Learning,” Applied Sciences, vol. 10, no. 1, p. 370, Jan. 2020, doi: https://doi.org/10.3390/app10010370

INNOVEX 2024