Enhanced Deepfake Detection Using Transfer Learning and Attention Mechanisms

Ashraf-Ul-Alam, Sudipta Progga Islam
ashrafamit9227@gmail.com, proggasudipta0@gmail.com
Computer Science & Engineering, Rajshahi University of Engineering and Technology

INTRODUCTION

- Deepfakes are AI-generated, highly realistic images or videos, making detection challenging.
- Once used for entertainment, they now pose threats to misinformation, privacy, and security.
- As deepfake quality advances, traditional detection methods struggle to keep up.

Fig 1: Examples of realistic-looking deepfake images [1].

OBJECTIVES

- Develop a robust model to accurately detect AI-generated deepfakes.
- Incorporate mechanisms to effectively handle the high volume of features extracted by the deep neural network, ensuring efficient model training.

PROPOSED METHODOLOGY

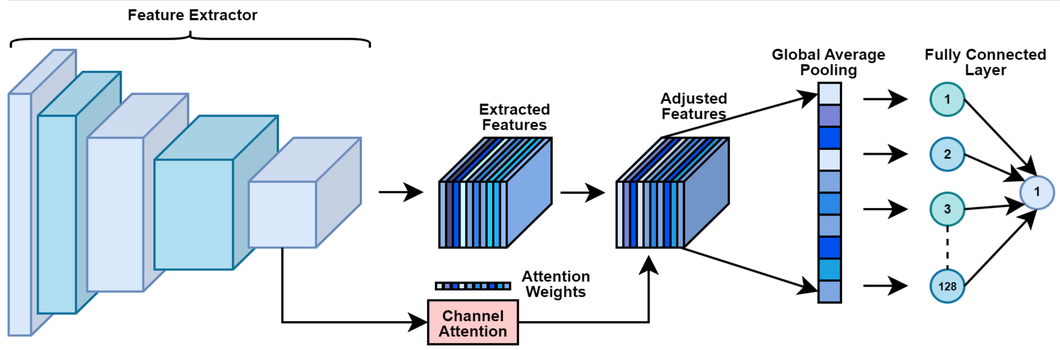

Fig 2: Proposed architecture.

- Pretrained models like ResNet-50[2] and VGG16[3] were used for feature extraction.
- VGG16 demonstrated better performance compared to ResNet50.
- To further enhance feature extraction, an attention mechanism was integrated into the model.

RESULTS

Table 1: Performance comparison(%)

Model	Accuracy	Precision	Recall
ResNet-50	96.95	99.87	93.42
VGG-16	99.64	99.92	99.35
ResNet-50 with attention	98.41	99.70	97.11
VGG16 with attention	**99.80**	**99.92**	**99.69**

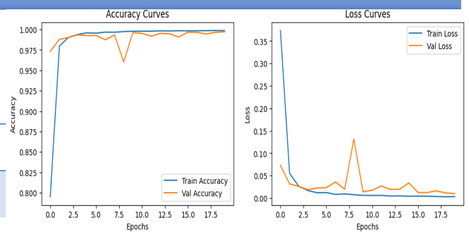

Fig 3: Accuracy and Loss Curves

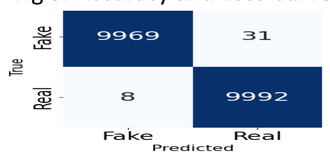

Fig 4: Confusion Matrix

DISCUSSION

- VGG16 outperformed ResNet-50 in detecting subtle manipulations.
- Attention mechanisms improved the model's focus on key features, enhancing detection accuracy.
- The improved performance and curves shows weight adjustment strategy leads towards optimized the weight adjustments while backpropagation.
- The approach is effective for high-quality, sophisticated deepfakes.

CONCLUSION

- VGG16 combined with attention mechanisms significantly improves deepfake detection.
- Future work could explore lightweight architectures and techniques for even better detection of real-world cases, including handcrafted deepfakes that are purposely modified in certain regions.

References:

[1] Xhlulu. (2020, February 10). *140k real and fake faces* [Dataset]. Kaggle. Retrieved September 4, 2024, from https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces/data

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.