Optimizing Feature Representation of Deep Neural Networks for Enhanced Deepfake Detection

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].

PREVIOUS FINDINGS

Table 1: Performance comparison with Previous Works(%)

Model	Accuracy	Precision	Recall
MiniNet[2]	95.18	-	-
CNN+ResNet-50+VGG16[3]	98.79	99.87	99.63

Due to the complexity and large volume of features extracted by hybrid models, the training failed to fully optimize the learning of relevant neurons, which can be addressed by adjusting the weights of the important feature vectors.

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 FRAMEWORK

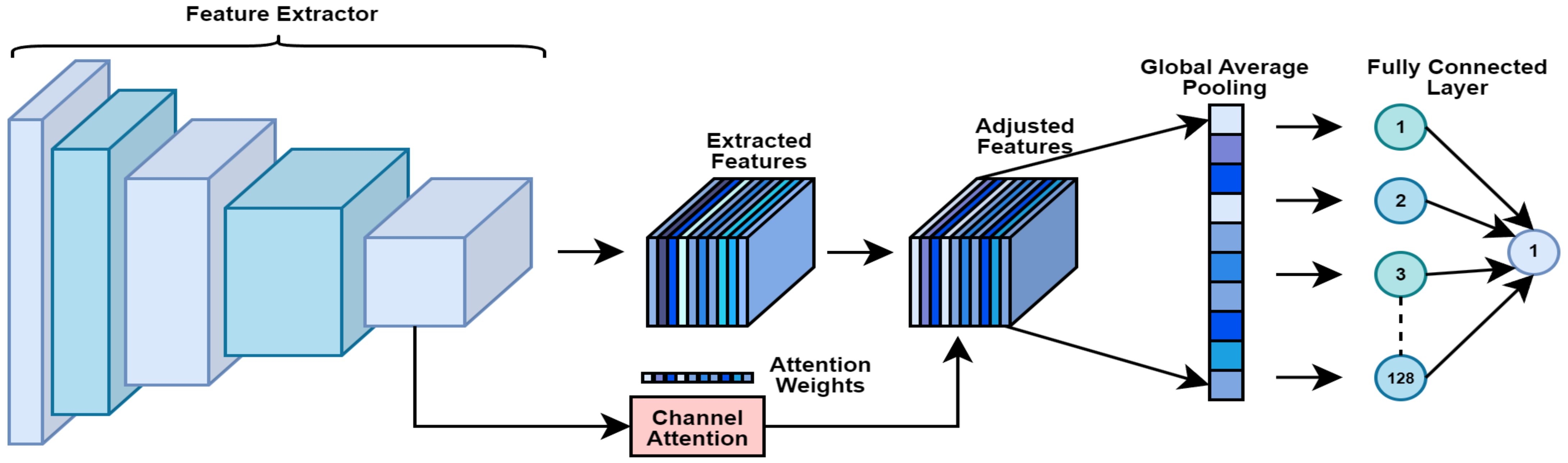

Fig 2: Proposed architecture.

- Pretrained models like ResNet50 and VGG16 were used for feature extraction.
- VGG16 demonstrated better performance compared to ResNet50.
- Channel Attention generates channel-wise weights using global pooling and fully connected layers to learn the importance of each feature channel. These weights are multiplied with the feature vectors to enhance focus on relevant channels.

RESULTS

Table 1: Performance comparison with different models (%)

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**

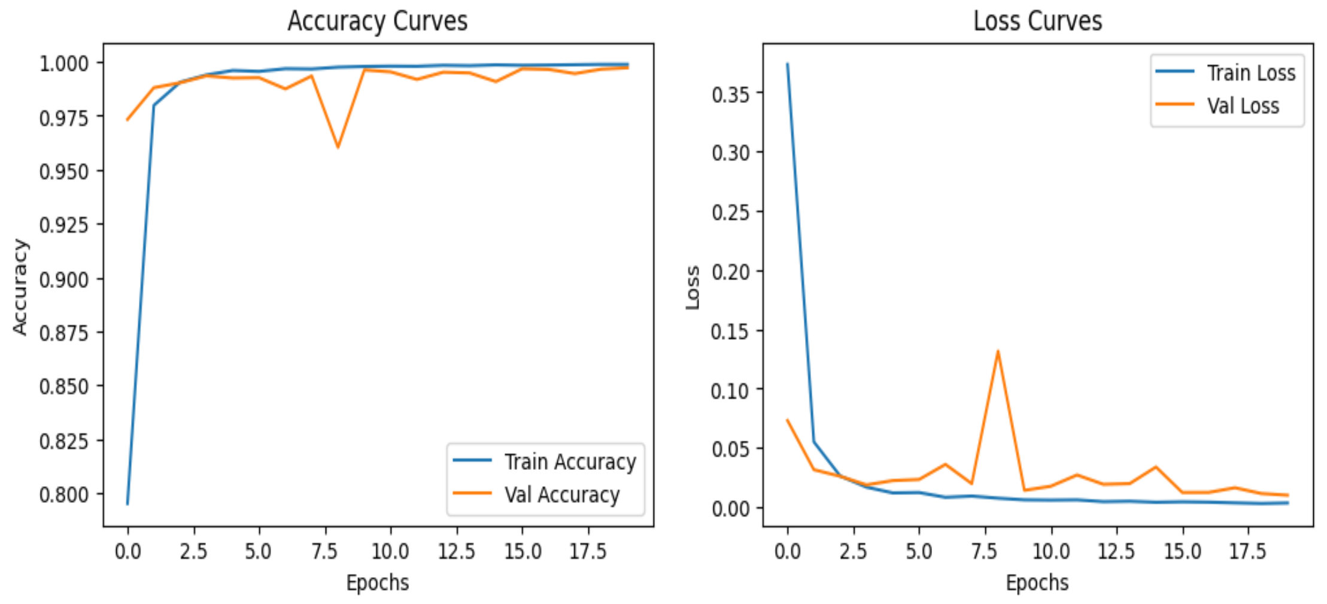

Fig 3: Accuracy and loss curves.

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.

LIMITATIONS

- Complexity of VGG16.
- May not be able to detect handcrafted fake images.

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] Tyagi, S., & Yadav, D. (2023). MiniNet: a concise CNN for image forgery detection. Evolving Systems, 14(3), 545-556.

[3] Sharma, J., Sharma, S., Kumar, V., Hussein, H. S., & Alshazly, H. (2022). Deepfakes Classification of Faces Using Convolutional Neural Networks. Traitement du Signal, 39(3).