**Student ID: W10174670**

**Name: Anu Baluguri**

**DEEPFAKE DETECTION PROJECT**

**Introduction**

Deepfake detection has emerged as a critical topic of research due to the possible misuse of manipulated media. This semester, I expanded my earlier work on image-based Deepfake detection to incorporate video analysis. My key goals were to improve preprocessing approaches, update the XceptionNet model implementation, and investigate a CNN+LSTM hybrid approach for more accurate temporal analysis. This report covers the approaches, lessons, and issues encountered throughout the semester.

**Key Learnings and Methods**

1. **Revised Preprocessing Steps**

Initially, I intended to reuse the preparation pipeline from Report 1, which included grayscale conversion, Sobel edge detection, and Gaussian blurring. However, I discovered that these strategies were insufficient for video data. I changed the preprocessing processes, as follows:

* **Frame Extraction:** Extracted frames from video files using OpenCV.
* **Resizing and Normalization:** Resized frames to 224x224 pixels and normalized pixel values for consistency.
* **Data Augmentation:** Applied rotation, flipping, and scaling to improve model robustness.
* **Batch Structuring:** Preprocessed frames were organized into structured batches, allowing a clear difference between Deepfake and real data.

This approach ensured compatibility with video datasets and improved the training data quality.

1. **XceptionNet Model Updates**

I used simple setups to construct XceptionNet in Reports 2 and 3. But the most recent coding file brought about some major enhancements:

* **Understanding Depthwise Separable Convolutions:**
* Depthwise separable convolutions, which XceptionNet uses, lower computing costs without sacrificing feature extraction effectiveness. I discovered how this structure enhances the ability to identify complex spatial aspects in pictures.
* **Transfer Learning Refinements:**
  + Used a pre-trained XceptionNet model (ImageNet) as a base.
  + By fine-tuning individual layers rather than freezing them entirely, the model was better able to adjust to the Deepfake dataset.
* **Improved Training Strategy:** 
  + Increased the number of epochs to 20, balancing overfitting with model performance.
  + Adjusted learning rates dynamically using a scheduler for better convergence.
* **Results:**
  + Achieved a peak validation accuracy of **91.67%** and a test accuracy of **96.67%**.
  + A blue squares with white text

    Description automatically generatedThe confusion matrix revealed high precision and recall for both Deepfake and Original classes.

Figure 1: Confusion matrix for XceptionNet

* The model's ability to accurately detect Deepfake and Original frames with few false positives or negatives is demonstrated by the confusion matrix above. This demonstrates how well XceptionNet extracts spatial data and produces accurate predictions.

1. **CNN+LSTM Hybrid Model**

To examine both spatial and temporal data in videos, I also investigated merging Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks this semester. Key insights include:

* **Model Design:**
  + CNN (based on XceptionNet) extracted spatial features from video frames.
  + LSTM layers captured temporal dependencies between consecutive frames, identifying anomalies in motion and expression continuity.
* **Efforts to Improve Accuracy:**
  + I attempted several techniques to improve the accuracy of the CNN+LSTM model, including adjusting the learning rate, introducing dropout layers, and using a ReduceLROnPlateau scheduler to optimize training. Despite these efforts, the model achieved a test accuracy of only 50%.
  + I learned from this experience how difficult it is to successfully integrate spatial and temporal variables, and it also made clear the necessity of bigger datasets or more reliable temporal feature representations.
* **Key Learnings:**
  + I gained a deeper understanding of Bidirectional LSTMs, which process sequences in both directions to capture contextual dependencies.
  + Combining XceptionNet with LSTMs reinforced my appreciation for the challenges of balancing spatial and temporal feature extraction.
  + I learned how temporal inconsistencies, such as unnatural motion transitions, are difficult to detect without significant dataset diversity.
* **Why Lower Accuracy for CNN+LSTM?**
  + Through my experiments, I realized that while XceptionNet excels in extracting spatial features, adding LSTM to capture temporal dependencies introduces more complexity. This made the hybrid model prone to underperformance, especially with a dataset that was not large or diverse enough to fully represent temporal patterns.
  + A graph showing a comparison of a number of classes

    Description automatically generated with medium confidenceDespite extensive efforts, including fine-tuning learning rates, adding dropout layers, and using a scheduler, the CNN+LSTM model achieved only 50.00% accuracy. I learned that effectively combining spatial and temporal features requires not just architectural adjustments but also a more robust dataset with greater temporal variety.
  + The confusion matrix below underscores the limitations of this hybrid model, with significant misclassifications highlighting the need for improved temporal data representations. However, this experience helped me understand the limitations of integrating complex models and the importance of dataset optimization.

1. **Results and Evaluation: Comparison of Models:**

* **XceptionNet:**
  + Focused on frame-by-frame analysis.
  + Demonstrated strong accuracy (96.67%) and reliability in detecting spatial inconsistencies.
* **CNN+LSTM Hybrid:**
  + Addressed temporal dependencies but faced challenges with dataset generalization.
  + Highlighted areas for future research, such as augmenting temporal features or increasing dataset diversity.
* **Key Metrics:**
  + Accuracy, precision, recall, and F1-score were used for evaluation. XceptionNet outperformed the CNN+LSTM model significantly across all metrics.

**Conclusion**

This semester, I deepened my understanding of Deepfake detection by transitioning from image-based to video-based approaches. The XceptionNet model proved effective for spatial analysis, while the CNN+LSTM hybrid offered insights into temporal challenges. Despite the limited success of the hybrid model, the experience provided valuable lessons in model integration and the complexities of video analysis. Future work will focus on improving temporal feature extraction and scaling models for larger datasets.

**References:**

1. Tsang, S. (2019, March 20). Review: XCeption — with depthwise separable convolution, better than Inception-V3 (Image classification). *Medium*. <https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568>
2. *Deepfake detection using XceptionNet*. (2023, November 8). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/10363477>
3. A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features. (2022, July 18). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/9892905>
4. Tipper, S., Atlam, H. F., & Lallie, H. S. (2024). An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection. *Applied Sciences*, *14*(21), 9754. <https://doi.org/10.3390/app14219754>
5. *CNN+LSTM for video classification*. (2023, July 29). PyTorch Forums. <https://discuss.pytorch.org/t/cnn-lstm-for-video-classification/185303>
6. <https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf>