DeepFake Video Detection Using EfficientNet

Styliani Kalaitzaki
aid24009

September 2024

1 Introduction

With the increasing spread of deepfake content, distinguishing real videos from fake ones is a crucial task. This project focuses on detecting deepfakes by processing video frames and leveraging a deep learning model (EfficientNet) to classify videos as real or fake based on facial features.

1.1 DeepFakes

Deepfakes are a form of synthetic media where artificial intelligence (AI) is used to manipulate or generate visual and audio content, typically videos, to make it appear as though someone said or did something they never actually did. These techniques often involve replacing a person's face in a video with someone else's, mimicking facial expressions, voice, and gestures with a high level of realism. Deepfakes are created using advanced AI models, including deep learning and generative adversarial networks (GANs).

1.2 Problem Statement

The goal is to build an efficient model to detect deepfake videos by processing video frames, extracting faces, and training a CNN-based model using EfficientNet for classification.

1.3 Objective

To classify videos as real or fake using a deep learning model trained on facial features extracted from video frames.

2 Data Preprocessing

2.1 Video Metadata and Dataset

The dataset consists of 400 videos labeled as "REAL" or "FAKE". Metadata from the `metadata.json` file was processed to obtain labels for the videos.

2.2 Face Extraction from Video Frames

Videos were processed frame by frame. Face detection was performed using the MTCNN face detector, and the faces were saved as individual images.

Face Detection The MTCNN library was used to detect faces within frames. Confidence scores ensured that low-confidence detections were skipped, and bounding boxes were calculated to crop the face regions.

2.3 Dataset Splitting

The dataset was split into training, validation, and testing sets using an 80:10:10 ratio. Downsampling of fake faces was applied to ensure that real and fake classes were balanced in the training dataset.

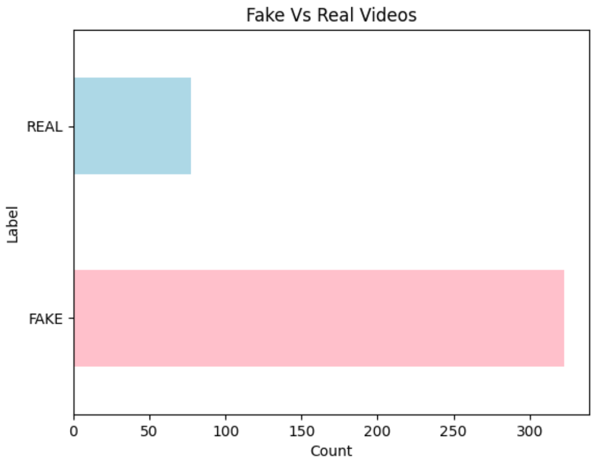

Figure 1: Dataset Balance

3 Model Architecture

3.1 EfficientNet Model

EfficientNetB0, pre-trained on ImageNet, was selected as the base model for feature extraction due to its efficiency and performance. The model was modified by adding fully connected layers, dropout for regularization, and a final sigmoid activation for binary classification.

Model Layers

- EfficientNet as the base model (without the top layer)
- Global Max Pooling layer
- Dense layers with ReLU activations
- Dropout layer (to avoid overfitting)
- Sigmoid activation for binary classification

4 Model Training

4.1 Training Setup

The dataset was augmented using `ImageDataGenerator` to improve generalization by applying transformations such as rotations, zoom, and flips. Early stopping and checkpointing were used to monitor validation performance and save the best-performing model.

4.2 Training Results

The model was trained for 20 epochs, and accuracy, validation accuracy, loss, and validation loss were plotted to visualize the training progress.

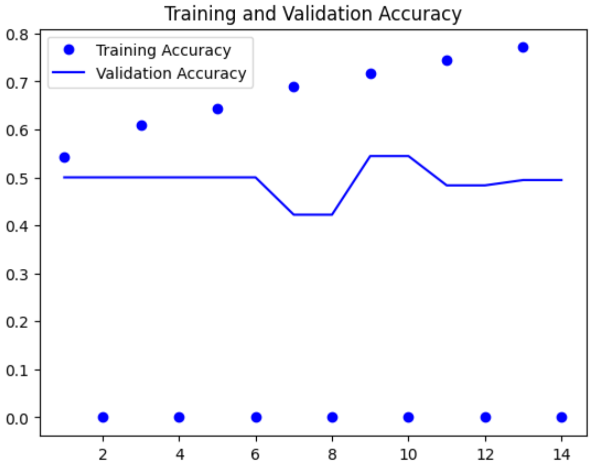

Figure 2

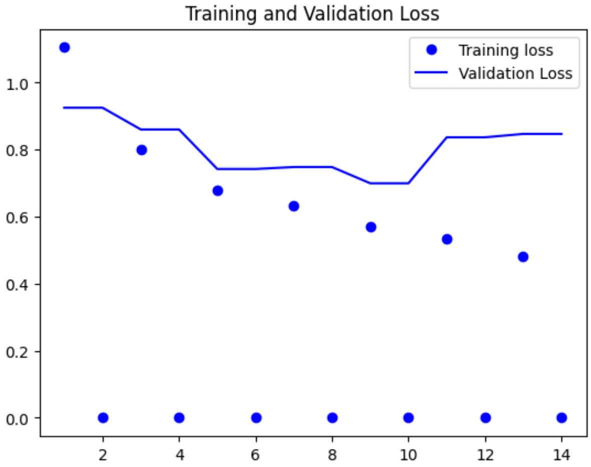

Figure 3

5 Model Evaluation

5.1 Prediction on Test Set

After training, the model was used to predict the labels on the test dataset. To improve video-level predictions, three different frame aggregation methods were explored: mean prediction, max prediction, and majority voting.

5.2 Performance Metrics

The following metrics were used to evaluate the model:

- **Accuracy**: Percentage of correct predictions.
- **Precision**: The proportion of predicted fake videos that were truly fake.
- **Recall**: The proportion of actual fake videos correctly identified.
- **F1 Score**: The harmonic mean of precision and recall.

```
Mean Prediction Without Downsampling:
Accuracy: 0.23
Precision: 1
```

```
Recall: 0.02
F1: 0.03

Mean Prediction With Downsampling:
Accuracy: 0.48
Precision: 0.55
Recall: 0.64
F1: 0.59

Max Prediction:
Accuracy: 0.48
Precision: 0.55
Recall: 0.66
F1: 0.6

Majority Voting:
Accuracy: 0.48
Precision: 0.55
Recall: 0.56
F1: 0.56
```

5.3 Confusion Matrix

A confusion matrix was generated to visualize the model's performance on both real and fake videos.

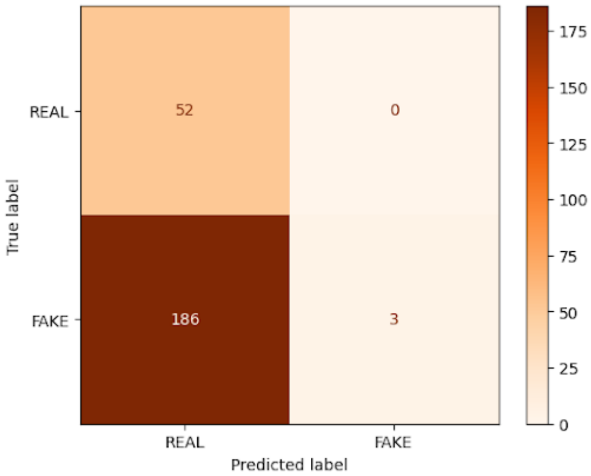

Figure 4: Confusion Matrix of Mean Prediction Without Downsampling

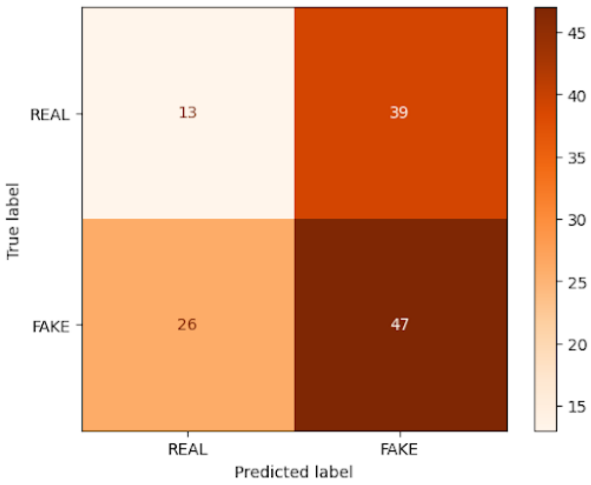

Figure 5: Confusion Matrix of Mean Prediction With Downsampling

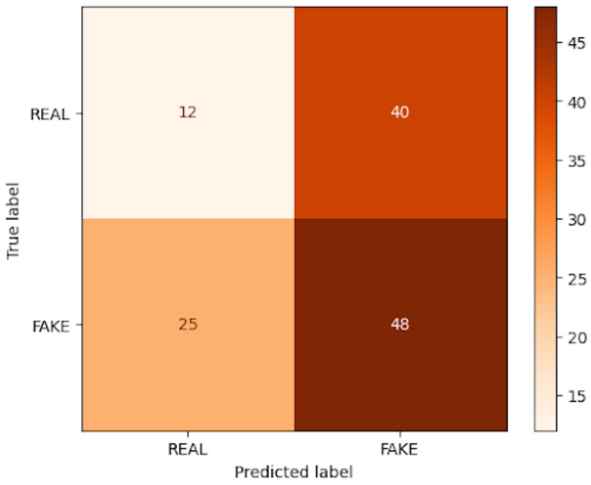

Figure 6: Confusion Matrix of Max Prediction

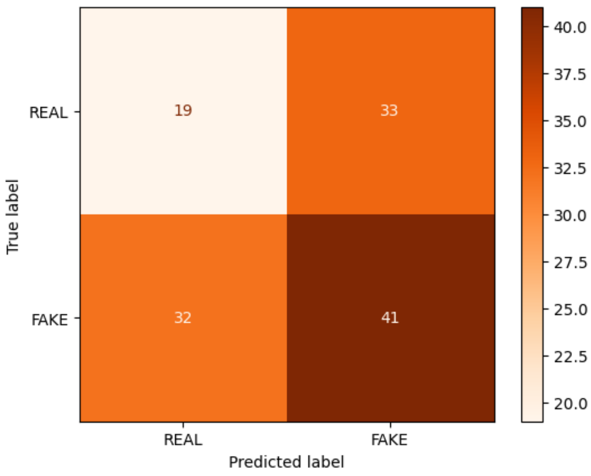

Figure 7: Confusion Matrix of Majority Voting

6 Conclusion

6.1 Downsampling Vs No Downsampling

The evaluation of the deepfake detection model revealed significant differences in performance based on whether downsampling was applied to the dataset.

With downsampling, the model achieved an accuracy of 48%, precision of 55%, recall of 64%, and an F1 score of 59%. It correctly identified 47 fake videos but missed 26, and misclassified 39 real videos as fake.

Without downsampling, accuracy dropped to 23%, with precision at 100%, recall at 2%, and an F1 score of 3%. The model identified only 3 fake videos correctly while missing 186, and did not misclassify any real videos as fake.

Downsampling improved the model's balance between precision and recall, but challenges remain in achieving high accuracy and handling class imbalance.

6.2 Comparison of Frame Aggregation Methods

Using mean prediction, where the average of all frame-level predictions was taken, the model achieved an accuracy of 48%, with a precision of 0.55, recall of 0.64, and an F1 score of 0.59. In this approach, the model correctly classified 47 fake videos (TP) but incorrectly flagged 39 real videos as fake (FP).

With the max prediction method, where the most confident frame-level prediction was chosen as the final video classification, the accuracy remained the same at 48%, but recall slightly improved to 0.66, and the F1 score increased to 0.60. This method led to 48 correct fake video classifications (TP) and 40 incorrect classifications of real videos (FP).

Finally, majority voting was employed, where the prediction chosen by the majority of frame-level outputs determined the video label. This method resulted in an accuracy of 48%, a precision of 0.55, and a recall of 0.56, with an F1 score of 0.56. It had a slightly higher number of true negatives (19 TN) but more false negatives (32 FN), indicating that this method was more conservative, correctly identifying fewer fake videos.

Overall, the max prediction method slightly improved recall and the F1 score compared to the other methods, though the changes in the confusion matrix were minimal, shifting only by one or two values.

6.3 Challenges

This project encountered a number of challenges during development. One significant hurdle was the high computational requirements, particularly when processing high-resolution videos, which demanded substantial memory and time for both training and inference. Additionally, the model exhibited sensitivity to video quality—low-resolution, noisy, or heavily compressed videos posed difficulties for the model in detecting subtle visual manipulations. Another key challenge was class imbalance within the dataset, as authentic videos were underrepresented compared to manipulated ones, which led to a bias in model performance. Although techniques like data augmentation and resampling were employed, achieving balanced accuracy remained difficult. Finally, the dataset included audio manipulation in some videos, which the model was not equipped to detect, further limiting its ability to identify cases involving solely audio tampering. These challenges underscore the need for more advanced and optimized approaches in future iterations of the model.

6.4 Future Work

Future improvements could focus on reducing the model's computational demands through more efficient architectures and better handling of low-quality videos via pre-processing techniques. Addressing class imbalance with synthetic data or advanced resampling methods is also crucial. Additionally, incorporating audio analysis would allow the model to detect manipulations in both visual and auditory domains, enhancing its overall effectiveness. Exploring multi-modal approaches and ensemble techniques could further improve detection accuracy across a wider range of manipulation types.