Video Deepfake Detection Model 1

1. Introduction

This report outlines the development and evaluation of a deepfake video classification system. The system aims to identify whether a video is real or fake by analyzing its frames through a machine-learning model. By leveraging deep learning techniques and pre-trained models, the system processes video data, extracts meaningful features, and builds a classifier to determine video authenticity.

2. Methodology

2.1 Data Preparation

- **Video Splitting**: Videos were initially collected and organized in a single directory. To ensure a representative training and testing dataset, an 80-20 split was used. This division allows for robust training of the model while retaining a significant portion of data for unbiased testing.
- **Metadata Handling**: Metadata associated with the videos was utilized to analyze the distribution of labels (real vs. fake). This metadata also facilitated the sampling of videos for visual inspection, ensuring the dataset's quality and diversity.

2.2 Feature Extraction

- **Pre-trained Model**: A pre-trained InceptionV3 model was employed to extract features from the video frames. This model, trained on the ImageNet dataset, is known for its efficacy in image classification tasks. The pre-trained weights were used to obtain meaningful representations of the video frames.
- **Frame Processing**: Each video frame was processed to ensure consistency in size and format. Frames were cropped to a central square region and resized to match the input dimensions required by the feature extractor. Additionally, frames were converted from BGR to RGB color space to align with the model's input expectations.
- **Inputs**: The model receives two types of inputs: frame features, which are the representations of the video frames, and frame masks, which indicate the presence of valid data. This setup allows the model to handle varying lengths of video sequences effectively.

3. Model Architecture

The deepfake detection system employs a sophisticated deep learning model with the following architecture:

- **Input Layer 2**: Accepts video frame sequences with a shape of `(None, 20, 2048)`. This layer is designed to handle the input data of video frames, each represented as a vector of size 2048.
- **Input Layer 3**: Handles additional input features with a shape of `(None, 20)`, providing supplementary temporal or contextual information.
- **GRU Layer**: A Gated Recurrent Unit (GRU) layer processes the sequence data with output shape `(None, 20, 16)`. This layer captures temporal dependencies in the video frames, with a total of 99,168 parameters.
- **GRU_1 Layer**: Another GRU layer reduces the sequence to a vector of shape `(None, 8)`, which distills the temporal features further, and contains 624 parameters.
- **Dropout Layer**: Applied with a dropout rate to the GRU_1 output, it prevents overfitting by randomly setting some of the layer's output units to zero during training.
- **Dense Layer**: A fully connected layer with an output shape of `(None, 8)`, which further processes the features extracted by the GRU layers, adding 72 parameters.
- **Dense_1 Layer**: The final output layer is a Dense layer with a single unit `(None, 1)` to produce the binary classification result, with 9 parameters.

In total, the model comprises **99,873 parameters**. All parameters are trainable, contributing to a model size of **390.13 KB**. This architecture is designed to effectively capture and classify deepfake video content by leveraging advanced sequence modeling and feature extraction techniques.

- **Output**: The model outputs a binary classification indicating whether a video is real or fake. A sigmoid activation function is used in the final layer to produce a probability score for classification.

2.4 Training

- **Training Setup**: The model was trained on the prepared dataset with a batch size of 8 and for a total of 50 epochs. A checkpoint mechanism was employed to save the best model weights based on validation performance, ensuring that the most effective model configuration was preserved.

2.5 Evaluation

- **Model Performance**: The model's performance was evaluated using accuracy as the primary metric. Accuracy measures the proportion of correctly classified videos out of the total number of videos in the test set.
- **Predictions**: The model was tested on new, unseen videos to validate its practical effectiveness. Predictions were made to classify these videos as either real or fake, demonstrating the model's ability to generalize to new data.

4. Model Prediction

4.1 Prediction Process

4.1.1 Input Data Preparation

- **Video Input**: New videos are preprocessed in the same manner as the training data (frame extraction, resizing, normalization).

4.1.2 Model Inference

- **Feature Extraction**: Features are extracted from video frames using the trained CNN model.
- **Sequence Classification**: The extracted features are processed by the RNN/LSTM to predict the authenticity of the video.

4.2 Performance Evaluation

4.2.1 Results

Accuracy and Loss:

- **Accuracy**: The model achieved an accuracy of **83.44%** on the validation set.
- **Loss**: The loss value during validation was approximately **0.4489**.

Dataset Used: **dfdc_train_part_49**