1. Problem Statement

The goal of my project is to detect deepfake images of human faces. Using a machine learning model, it will identify the legitimacy of a picture, and give the probability that it is a deepfake.

2. Data Preprocessing

The dataset I am working with is this one: https://www.kaggle.com/dagnelies/deepfake-faces, containing more than 95k images. I will use this one instead of the previously chosen one (official dataset for the Deepfake Detection Challenge). I decided to change because of simplicity, since this dataset contains images of the first frame of every video from the previous dataset. The images are also centered on the face and resized to 224x224 pixels. This way, I can work directly with images instead of getting images from the huge video dataset, run a facial recognition model to identify faces, and then resize the input. To preprocess the data, I simply ran several convolution, activation, and pooling layers.

3. Machine learning model

The model used is an image classification CNN as discussed in the last deliverable. I built a Keras Sequential model with 3 convolution layers and 2 hidden layers. I did a 80% training set vs 20% testing set split, so that I had enough data for the testing set and making sure the model was not overfitting. However, due to time constrains, I did not get to set customized hyper-parameters and test around to see if I could improve the accuracy. After running 20 epochs, I observed that the model was wildly overfitting as my validation set accuracy was constantly dropping for increased training set accuracy. Again, due to time constraints and a last minute project change, I did not have time to change models.

4. Preliminary results

To see detailed a more analysis of the model, see the notebook. In summary, the validation accuracy and loss quickly plateaued and began to decrease after a few epochs. The model is extremely overfit, and we can see this model is clearly not appropriate for deepfake detection.

5. Next steps

For the next steps, I will try to implement a different model completely. Instead of a classic image classifier, I will use a pre-trained face recognition model to better evaluate the legitimacy of a detected face. This way, instead of choosing between binary classes “real” or “fake”, the model will give the probability that the input image is a deepfake. I will then compare the future results with the accuracy of the current model.