
IsSwap?
Deep Fake Detection using Deep Learning

Jaypee Institute of Information Technology

IS Swap?

Supervised By:
Dr. Rashmi Kushwah

•••

Aakriti Aggarwal
Siddhant Wadhwa
Pallav Gupta
Nishit Anand

OVERVIEW

In the current world scenario where technology has taken over every aspect of human life and has become an essential part of our daily lives. We use technology to complete routine tasks, this has actually made life simpler in more ways than we can imagine. The sheer amount of information we consume from digital media is enormous.

As our intake of information through digital media has increased, new problems have arisen. How do we know that the same technology we love and trust is not being used against us? Over time with technological advancements people have found ways to use technology to falsify information and spread it to achieve personal vendetta. One of such practices is called ‘Deep-Fake’. It is a photo or video of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information.

POTENTIAL AND PROPOSED THREATS

1. Celebrities
2. Politicians
3. Businessman
4. Spokesperson
5. Fashion Industry

UNDERSTANDING THE PROBLEMS

01 Information through digital media has grown exponentially and it has provided people with personal motives to spread falsified information specially during elections to create political unrest among the masses or simply to spread a rumor.

02 The explosive growth in deep fake video and its undetected use is a major threat to democracy, justice, and public trust. Due to this there is a increased the demand for fake video analysis, detection and intervention

03 Deep Fakes are raising a set of challenging policy, technology, and legal issues.

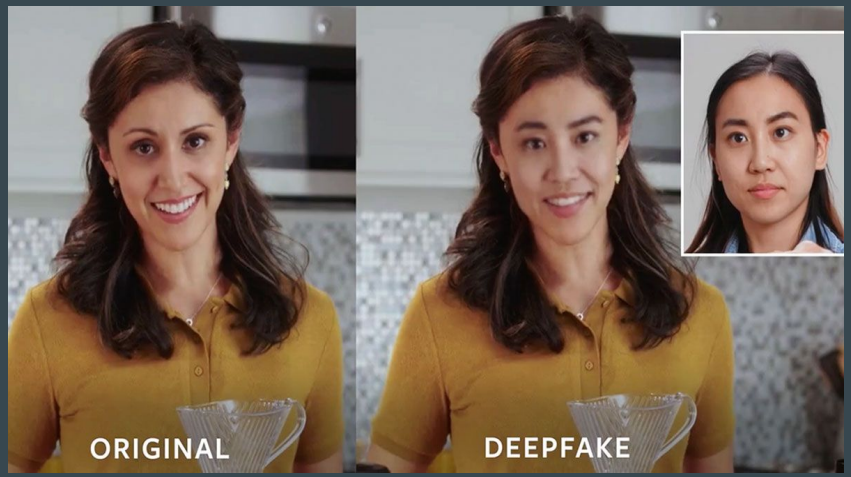
PROJECT OBJECTIVE

Users upload more than 500 hours of fresh video content per minute which roughly translate to 7.2 lakhs of new content uploaded every day and this is just on one platform, namely YouTube.

isSwap? attempts to protect people from believing in false information which is being spread through the said deep-fakes by identifying such videos. It is a web application which uses deep learning techniques to achieve its goal which is to help the user determine the authenticity of a given video.

HOW DEEP FAKES ARE CREATED

For creation of deep fakes, GANs or General Adversarial Networks are used. What is done is that a source image is taken and then another person's image whose face we want to be in there and whose face we want to be faked is used. The model is given both the original video and 2nd person photos as input and it is trained. The new person's face is either mixed and morphed or sometimes overlaid on the face of the first person. Thus, the resulting face we get is either a mixture of both the faces or only the face of the second person and we get a deep fake as the 2nd person was not actually there in the video and the video is fake.



STATE OF THE ART

XceptionNet-

- It is a deepfake detection method using convolutional neural network (CNN). Xception is a convolutional neural network pre trained on more than a million images from the ImageNet database.
- The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images.
- CNN (Convolutional Neural Network) is employed to extract frame-level features. A fully connected network is finally used for calculating probabilities of the frame sequence belonging to either authentic or deepfake class and thus classifying doctored videos from real ones.

ResNeXt-

- It is a CNN model which is also used for classifying images and videos and gives high accuracy when used for classifying tasks.

REVIEW OF THE DATASET

Files

- **train_sample_videos.zip** - A ZIP file containing a sample set of training videos and a metadata.json with labels. The full set of training videos is available through the links provided in report
- **sample_submission.csv** - A sample submission file in the correct format.
- **test_videos.zip** - A zip file containing a small set of videos to be used as a public validation set.

Metadata Columns

- **filename** - the filename of the video
- **label** - whether the video is REAL or FAKE
- **original** - in the case that a train set video is FAKE, the original video is listed here
- **split** - this is always equal to "train".

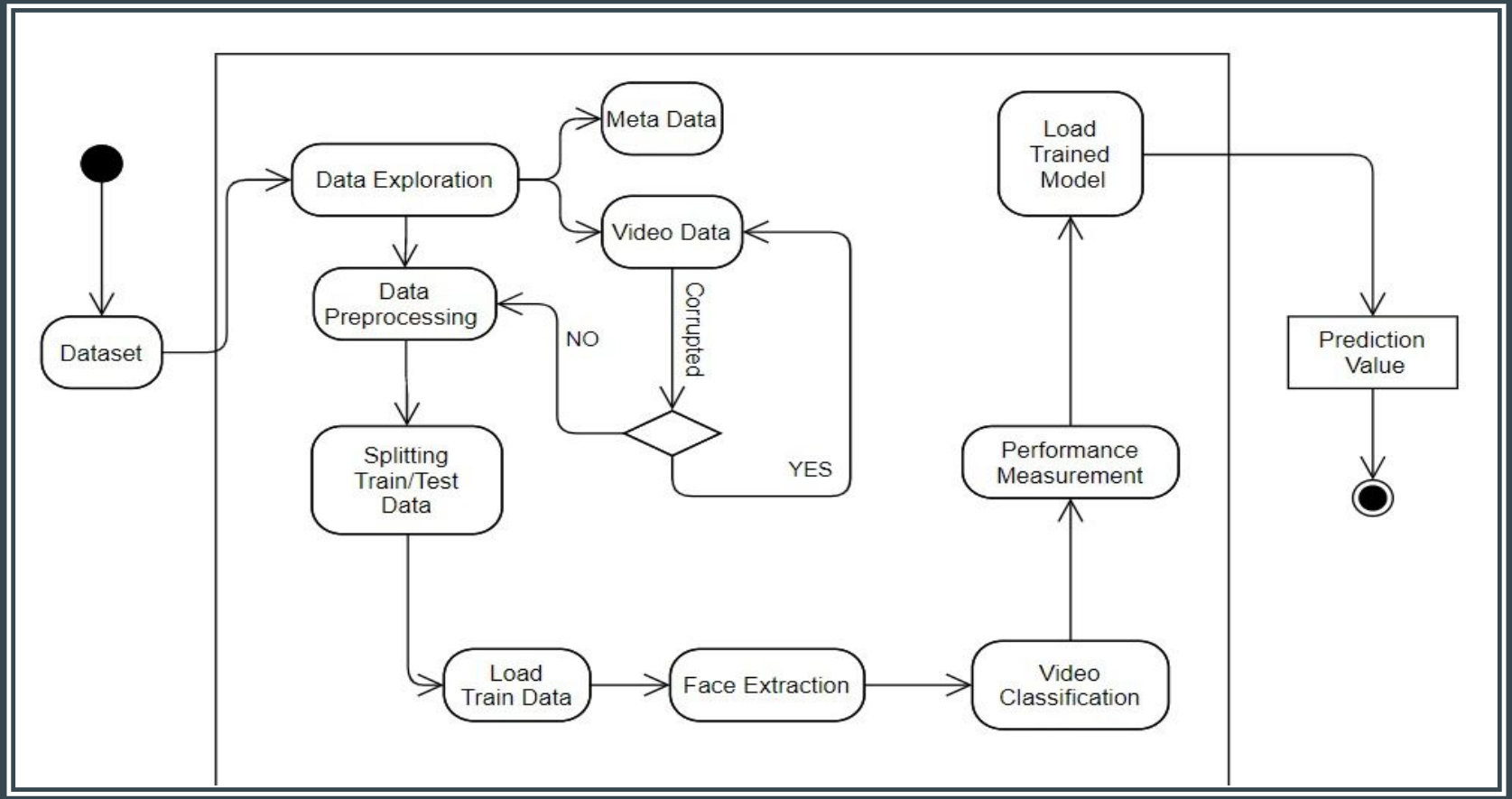
TOOLS USED FOR PROJECT

- Python
- Javascript
- HTML/CSS
- Pytorch
- Flask
- Kaggle Cloud
- Google Colab
- Jupyter notebook
- PyTorch

TECHNOLOGIES USED

- Frontend
- Backend
- Neural Networks
- Deep Learning
- CNN
- Matplotlib
- Fully Connected Layer
- Numpy
- Pandas

DETAILED DESIGN


Dataset
Data Exploration
Meta Data
Video Data
Data Preprocessing
NO
Corrupted
YES
Splitting Train/Test Data
Load Train Data
Face Extraction
Video Classification
Performance Measurement
Load Trained Model
Prediction Value

DATA PREPROCESSING

a. Split the video into the frame.

b. Validate the video to check if the video is corrupted or not. If it is then delete the video.

c. Blaze Face classifier is used to detect the faces in the frame.

d. Remove the face from the frame. (Cropping)

e. Resize the images so as to have fixed pixel size for better output.

- We have used ensemble of two models. The first model is ResNeXt and the second is Xception Net.
- ResNeXt model has convolutional layers followed by pooling layers. The output of pooling layers is then fed into the sequential layer.
- In the Xception Net, after importing Xception model what we have done is after the Xception model we put a few layers more layers to improve the accuracy of the model. Output of Xception Net is passed to an AdaptiveAvgPool2D layer. Then we have put a flatten layer to flatten the outputs.
- After that we use a linear layer followed by a dropout layer of 0.5 probability to stop the model from overfitting. Output of Dropout Layer is processed by another Linear Layer. Then we use batch normalization layer followed by a ReLU activation layer.

- After this we run both the models to make inference and both of them give us a score between 0 and 1 is 0 means it is real and 1 means it is fake. Then we take weighted average of scores achieved from both the models for any particular video.
- We found that we achieved the best results when we gave slightly less weightage to ResNeXt model as compared to Xception Net model. Then we find the final score. If the score for that video is greater than 0.5 then it means that our model has predicted that it is a fake video and if the score is less than 0.5 then it means our model has predicted that it was a real video.

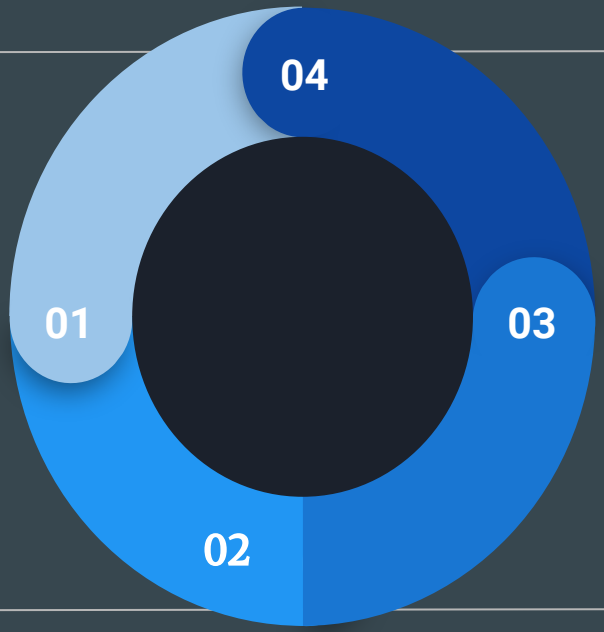
CYCLE DIAGRAM

Data exploration (Files)

1. Load Data
2. Load Packages
3. Check File types

MetaData Exploration

1. Missing Values
2. Unwanted Data
3. Fixing structural data



Training Data

1. Splitting Datasets
2. Model Training

Video Data Exploration

1. Rescalling Images
2. Missing Real or corresponding fake videos and vice versa.

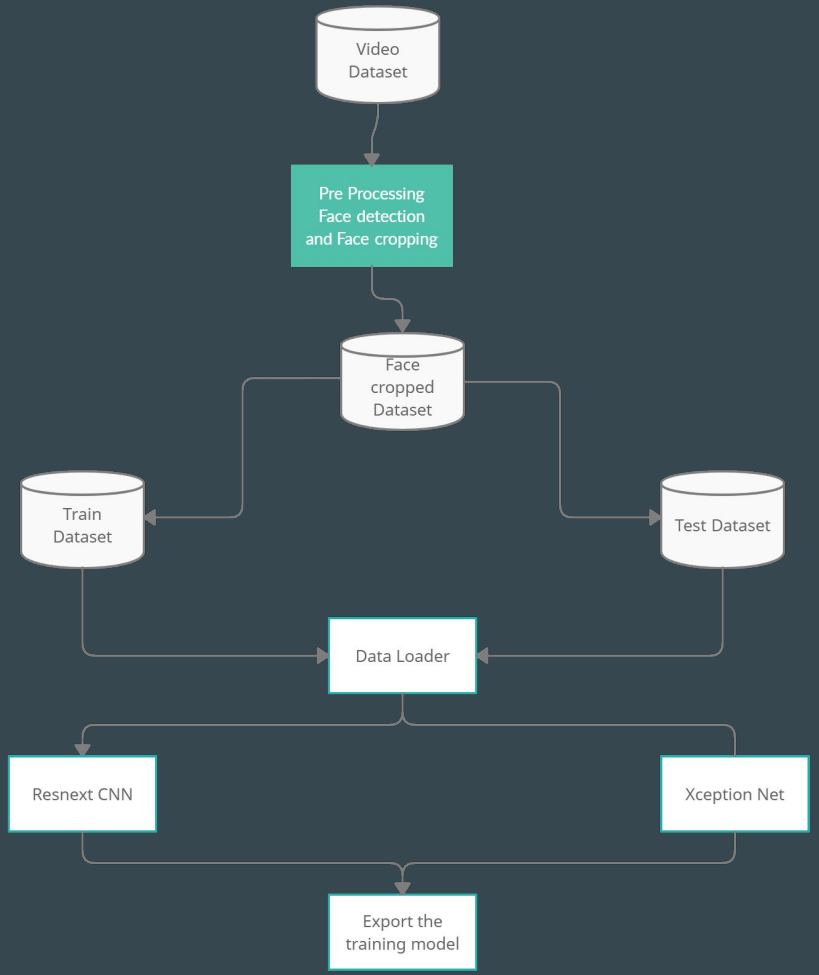
MODEL ARCHITECTURE

Following are the layers available in our model for training a deep neural network.

a. ResNeXt-50 32*4 dimension pre trained model for feature extraction. It consists of 50 layers with 32 nodes in each layer which is capable of learning a large number of parameters.
b. The output of ResNeXt is a pooling layer which gives us a feature vector which is then fed into a sequential layer.
c. The output of Xception Net is AdaptiveAvgPool2d layer.
d. Sequential layer fed the input into the AdaptiveAvgPool2d layer. We have used 1 linear layer with the chance of Dropout of 0.5.
e. The output of (d) is further processed by a linear layer.
f. Finally the softmax layer is implemented which gives the output whether the video is Real or Fake.
g. Train_epoch trains the model based on the given number of epochs. Epochs is a hyperparameter that defines the number of times that the learning algorithm will work through the entire training dataset.

Each epoch consists of a number of batches. In our function we are defining a range of parameters for each epoch. It will compute the best possible epoch value and batch size

TRAINING WORKFLOW



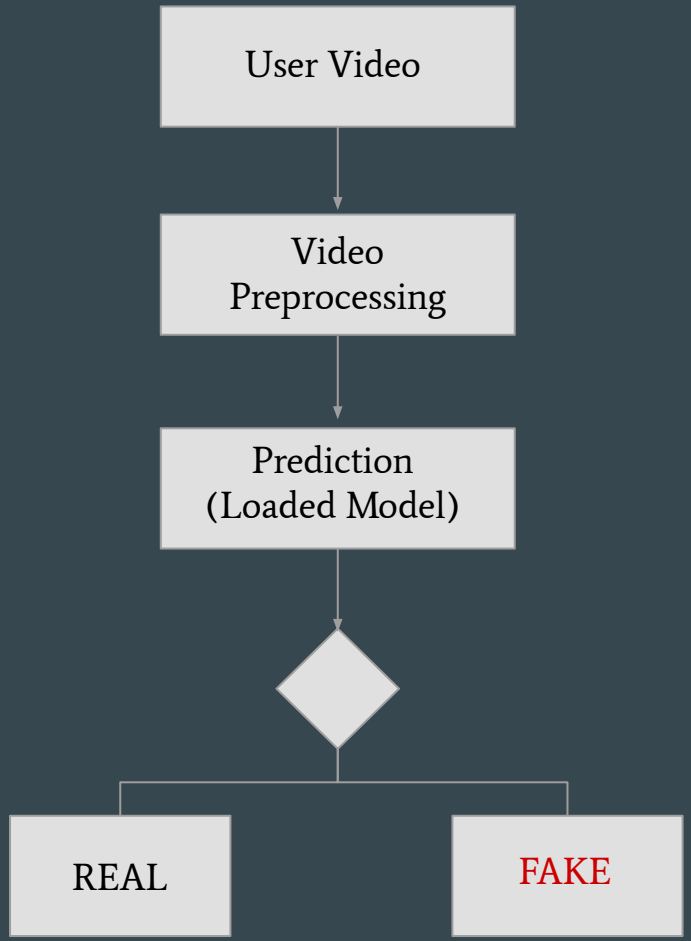
PREDICTION WORKFLOW

Prediction workflow starts as soon as the user uploads a video. Then the video enters to the Video Preprocessing function where face extraction, feature extraction and splitting into frames is done. It is then loaded to the trained model.

Based on the model it calculate the label corresponding to the video.

If label >= 0.5; Fake Video

Else; Real Video



RESULT

An important observation after training the final model is that as we are increasing the number of frames, i.e. as the sequence length is increasing, the accuracy also increases. We have checked the sequence length upto 100 frames. But by using Hyperparameter optimization we concluded that 64 frames gives the best accuracy.

We have trained the model using both the neural networks that are ResNet and Xception Net. The outcomes demonstrate two points:

1. ResNet independently performed less precisely as compared to the Xception Net network.
2. By applying the probability on each neural network's precision and performing the given below mathematical equation, produces a more desirable outcome to the model.

```
r1 = 0.46441
r2 = 0.52189
total = r1 + r2
r11 = r1/total
r22 = r2/total
```

```
submission_df["label"] =
r22*submission_df_resnext["label"] +
r11*submission_df_xception["label"]
```

LIMITATIONS

- This technology is not 100% accurate
- It requires access to internet.
- The file has to be uploaded manually, insertion through URL not available.
- Inability to produce reliable results if multiple faces are present in the video.
- While preprocessing 500+ videos we were facing the main challenge in using Haar cascades and blazeface, as they don't seem to be very convenient tools. Setting up the parameters to match various images seems to be nearly manual.
- Inaccurate results when the subject is farther than 5m.

CONCLUSION

IsSwap? is proposed to detect the fake face and general images generated by the state-of-the-art GANs.. Our results using a large collection of manipulated videos have shown that using a convolutional ResNET and Xception Net structure we can accurately predict if a video has been subject to manipulation or not within a few seconds. .The experimental results demonstrated that the proposed method outperformed other state-of-the-art methods in terms of precision and recall rate. Fake video detection is also an important issue, so in our future work, we will extend the proposed method to fake video detection, incorporating the object detection.

THANK YOU!