Fake Video Detection Using Multiple Instance Learning

Avnish Kumar
ak2626@cornell.edu

Pratyush Sharma
ps984@cornell.edu

Sarath Chandra
src257@cornell.edu

Ravi Theja Reddy
rs2538@cornell.edu

Ananya Shivaditya
aks298@cornell.edu

Prashant Jain
pj263@cornell.edu

Abstract

With the advent of deep fake technology, authenticity of an image or a video has been severely compromised, leading to repeated manipulation of public sentiment. There exist methods of detecting the authencity of a video, but these are targeted for specific kinds of fakes and can't always be generalised. With the goal of a generalisable method to reliably detect authenticity of a video, in this paper, we explore the use of Multiple Instance Learning for this task and compare accuracy, scalability and generalisability with existing methodologies. By classifying bags of frames from the video, we achieve high accuracies for a lower compute cost. These bags containing individual frames are hypothesised to require a relatively less complex model for classification than for the entire video. This approach could be especially advantageous if a positive (fake) video doesn't have manipulated content in all frames.

1 Introduction

Rapid innovation and development of most recent techniques not only in the field of computer vision and deep learning but also in the ubiquitous image software such as Adobe Photoshop, it has now become very easy to generate synthetic realistic multimedia content even in real time. Manipulation of digital multimedia content on social media and the internet is a burgeoning issue that can potentially cause huge problems in the future if misused or left unregulated. It is in fact increasingly becoming more challenging for humans to distinguish original videos from their fake counterparts.

Using the novel Multiple Instance Learning (MIL) technique, we aim to detect the authenticity of a given image or a video. MIL allows to divide a corpus of multimedia content into bags. MIL algorithm accepts a set of labeled bags instead of a set of instances which are individually labeled, such that each bag contains many instances. MIL labels a bag as negative only if all the instances in the bag are negative. Once the model labels all the bags, MIL tries to label individual elements correctly. In this paper, we describe a stochastic-sampling methodology based MIL technique that facilitates the detection of fake images or videos without the need to process the complete file.

2 Related Work

There is significant literature in the topics involving image/video forgery and their identification. Before we dive deeper into detection it makes sense to know what methods are used to generate these videos. Also MIL and its applications has to be investigated. In the subsection we would discuss:

1. Methods for video generation

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2. Methods for detection of fake videos and images
3. MIL methods and applications

2.1 Fake Video Generation Techniques

An image-based approach called Video Rewrite to automatically create a new video of a person with generated mouth movements has been presented by Bregler et al[3]. With Video Face Replacement [20], Dale et al. presented one of the first automatic face swap methods. Using single-camera videos, they reconstruct a 3D model of both faces and exploit the corresponding 3D geometry to warp the source face to the target face. Garrido et al. [29] presents method to use single-camera video to construct the 3D models of both the faces and then use the 3d geometry to warp the source face to the target face while preserving the original expressions. Face2Face, proposed by Thies et al. [11], is an advanced real-time facial reenactment system, capable of altering facial movements in commodity video streams, e.g., videos from the internet.

Generative adversarial networks (GANs) are used to apply Face Aging [2], to generate new viewpoints, or to alter face attributes like skin color. Deep Feature Interpolation [12] shows impressive results on altering face attributes like age, mustache, smiling etc. Most of these deep learning based image synthesis techniques suffer from low image resolutions. Recently, Karras et al. [7] have improved the image quality using progressive growing of GANs, producing high-quality synthesis of faces. Instead of a pure image-to-image translation network, NeuralTextures optimizes a neural texture in conjunction with a rendering network to compute the reenactment result.

2.2 Fake Video Detection Techniques

Researchers have proposed forensics methods to detect a variety of face manipulations. Zhouet al. [8] and Roessler et al. [9] propose neural network models to detect face swapping and face reenactment, i.e. manipulations where one face is wholly replaced with another (perhaps taken from the same subject) after splicing, color matching, and blending. FaceForensics++ [9] presents an in depth analysis of the various detection methods and their applications. Deep networks which can detect subtle inconsistencies arising from low-level and/or high level features have also been proposed. RNN based method proposed by a simple convolutional LSTM structure can accurately predict if a video has been subject to manipulation or not with as few as 2 seconds of video data. A few more methods are listed below:

1. Xception
2. MesoNet
3. Sentinel
4. Inception Resnet V1
5. RNN Based Detection

2.3 Multiple Instance Learning (MIL) Application

But a common assumption in all the methods mentioned take in one frame of a video and do the analysis to check the authenticity of the video. This approach may not be adequate when we try to do large scale analysis of videos. To put things in context large scale implies these three things:

1. Forged part of the video is a small interval in a long video
2. Forged video is a part of a huge set of video (like Youtube or Vimeo)
3. A combination of both

Large quantities of data necessitate a growing labeling effort. Weakly supervised methods, such as MIL, can alleviate this burden since weak supervision is generally obtained more efficiently. In our case a video can be labeled morphed but the exact intervals may be problematic.

MIL was first introduced for the problem of drug activity prediction [4], where axis-parallel hyper-rectangles (APR) were used to design three variants of enclosure algorithms. The APR algorithms tried to surround at least one positive instance from each positive bag while eliminating any negative instances inside it. A test bag was positive if it had at least one instance within the APR. Conversely if there is no instance in the APR the bag is classified as negative.

The technique MIL deals with training data arranged in sets, called bags. Supervision is provided only for entire sets, and the individual label of the instances contained in the bags are not provided. This problem formulation has attracted much attention[13], where the amount of data needed to address large problems has increased exponentially.

A video is condensed into a bag of snaps sampled according to some distribution and then used for MIL. The main advantage of using this approach is that the drastically reduces the computational complexity. The three common ways through which we utilize MIL are 1. MIL pooling 2. MIL with neural networks 3. MIL and attention.

MIL pooling and MIL with neural networks have few disadvantages. They are already predefined and non-trainable since the max operator could be a good choice in instance based approach but it might be inappropriate or the embedding based approach. Similarly mean operator is definitely a bad MIL pooling to aggregate the instance scores. Therefore a flexible and adaptive MIL pooling could potentially achieve better results by adjusting task to data. MIL based attention can assist in greatly dealing with the adaptability

The attention mechanism is widely used in deep learning for image captioning or text analysis. In the context of the MIL problem it has rarely been used and only in a very limited form. Pappas and Popescu-Belis proposed an attention-based MIL was proposed but attention weights were trained as parameters of an auxiliary linear regression model. The idea was further expanded from linear regression to a neural network. The only recent work on attention based MIL, the attention was calculated using the dot product

3 Data

1. **Kaggle Dataset**

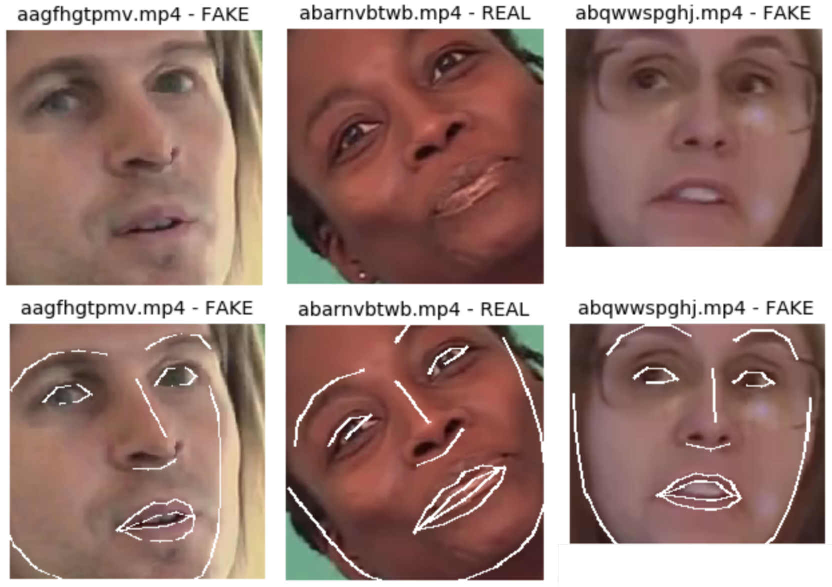
   The kaggle deep fake challenge has just over 470 GB volume of videos split over 50 files each 10 GB. Each video is compressed in mp4 format and consists of an individual speaking. Each original video in the dataset has a replica of the same video but with morphed content of the individual.

   The competition specifies that the train data is provided as archived chunks. There are 119,146 mp4 files in total (training set: 118346 videos) and the ratio of real/fake in each chunk of data as approximately 1:5 . Within each chunk, we found that the real data has missing original videos, and the number of unique samples is therefore reduced. We explored a subset of data by displaying a selected image from a video, viewing real videos and their fake counterparts, and different sections of the same video comparing parts that were modified. Some of the alterations were even visible to the naked eye, on faces. Face (front and side profile) and eye detection was further useful to visualise alterations in fakes.

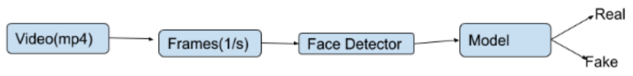
2. **FaceForensics++ Dataset**

   The FaceForensics++ dataset comprises 38 GB of original videos from youtube. Classical computer graphics-based methods Face2Face and FaceSwap as well as learning based approaches DeepFakes and NeuralTextures were utilized to generate a large-scale dataset of manipulations around 500 GB. These morphed videos were compressed lossless with H.264 and have a constant frame rate of 30 fps.

A few images along with their face detection features are shown below:



The Data Pipeline is shown below:



4 Data Preprocessing

Kaggle dataset of 470gb was prohibitively large to use, so decided to condense by extracting relevant features. Prior Work has shown, facial features can be used to identify fake image. We used dlib library to extract face from each second of video, leading to generation of 1mn photos which we used. DLib used CNN based face detector to delineate a rectangular window of the face. This reduced datasize from 470gb videos to 10gb 11mn images.

5 Method

5.1 Deepfake video detection using RNN

The Recurrent neural networks are a class of artificial neural networks where the connection between nodes form a directed graph along a temporal sequence. Unlike the feed-forward neural networks, the RNNs use their internal state memory for processing sequences. This behavior of RNNs is leveraged in image detection, video analysis, and other similar applications. To build our deepfake video detection model, we explored GRUs which is known as a Gated Recurrent Unit in an RNN architecture. GRU comprises the reset gate and the update gate in the basic RNN structure. Combining new input with the previous memory is determined by the reset gate, and the threshold of the previous memory used in the overall model is determined by the update gate.

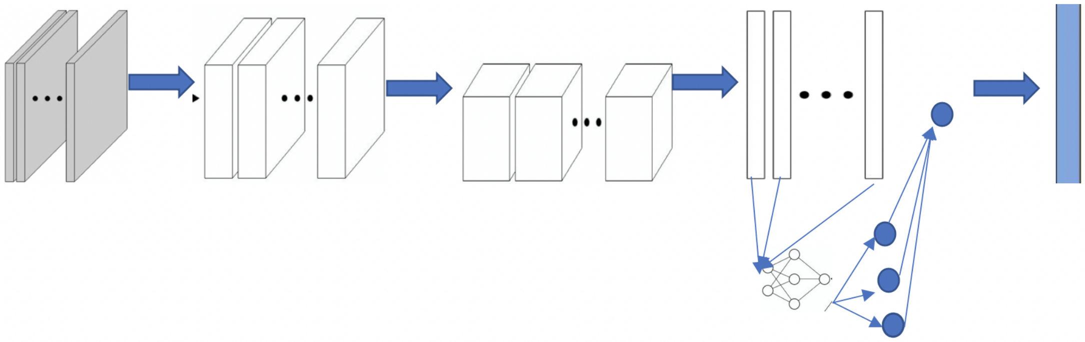
We followed the approach presented in “Deepfake Video Detection Using Recurrent Neural Networks” [5] research paper for deep fake video detection. Given the temporal nature of our dataset and its dependence on chronologically ordered actions, we have introduced a “TimeDistributed” layer in our model, which is immediately followed by the GRU layer (instead of LSTM due to limited memory availability at the expense of lesser efficiency). In summary, following high-level architecture is used for detecting the deep fake videos in our data set:

1. Building a convolutional neural network and introducing a TimeDistributed layer in the network to detect temporal features
2. The output from the TimeDistributed layer is fed into a Gated Recurrent Layer unit to treat the time-series like features of the data.
3. Finally, a DenseNet neural network layer is added to classify the data into real and fake videos, and take the decision based on a softmax layer. We only used the output of the final image.

5.2 Multiple Instance Learning

We practiced the methodology presented in "Attention-based Deep Multiple Instance Learning"[1] for deep fake detection in videos. The motivation behind using MIL is that we sample the video into a bag of screenshots and use this bag to calculate the attention based on two layered neural network and using this attention to predict the class of image based on the attention by passing through a fully convoluted network. The below picture encapsulates the high level architecture of our MIL model

Architecture of our MIL based Attention Model:



Structure of MIL Attention Model:

On the bag,we have applied CNN on the first pass, then we have used one more CNN with Relu Activation and Max pooling. In the third pass we have used Relu to condense the image and then we have used MIL based attention on this condensed image to extract the feature wights which is then passed through a Fully Convoluted Network on which the final classification of detecting whether a video is fake or not is determined.

Model Parameters:

1. bag size = 8
2. number of training bags = 8
3. method = attention

We utilized the method of using Multiple Instance Learning to detect altered content in the video by sampling frames of the video following a certain distribution. We bootstrapped a sample of frames from a video based on certain distribution (Uniform, Beta, Gamma), made a bag out of these samples and trained a Convolutional Neural Network on this bag of samples to classify as (positive)fake.

Structure of Convolution Neural Network:

We experimented with the layers of the neural network architecture. Our baseline model emphasized on the creation of the bags using MIL and tested how it affects the accuracy. We extended[6] by modifying our pooling to a multi layered pooling. We observed whether max pooling/ min pooling / mean pooling for the bagged images first form a pooled image,then use a second layer of pooling to reduce the size of the final pre-processed image. This image is then sent to the the convoluted neural network. The pooling steps are described below:

First layer of pooling:
For each pixel of the image y = (argmax y_k) for all the images in the bag to capture a pooled image.

Second layer of pooling:
We use one more layer of max pooling on this pooled image and use this as input to the CNN model.

6 Evaluation

6.1 Compute

We used Colab Pro account with Tesla GPU. Data preprocessing was done in batches of 100gb's with each speed of 4 minutes/GB.

Table 1: Time per epoch

Model	Hour
Attention Based MIL	3hr
RNN + Inception	4hr
FineTuned Inception Resnet	1hr

6.2 Metrics

Primary evaluation metric is kaggle contest score which is logloss on test set in kaggle competition.[1]

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Where:

- n is the number of videos being predicted
- $\hat{y}$ i is the predicted probability of the video being fake
- y_i is 1 if the video is FAKE, 0 if real
- log() is the natural (base e) logarithm

Model was further tested on Faceforensics++ dataset.

7 Results

7.1 Testing on FaceForensics++

Table 2: Results on FaceForensics++ Dataset

Goal	TestAccuracy
Facenet	99%
MIL	96 %

We see that we are able to attain, decent accuracy by using Attention based MIL to detect facially morphed images from facenet++ challenge. Facenet++ was trained specifically to identify specific mutations of face.

7.2 Testing on Kaggle DeepFake Challenge

Table 3: Results on Kaggle Deep Fake Challenge

Goal	Dev NLL	TestNLL
Attention Based MIL	0.64	0.65
RNN + Inception[10]	0.64	0.69
FineTuned Resnet+Facenet++	0.63	0.63
Voting Ensemble(MIL+ Resnet)	0.61	0.61

The best leaderboard submission had a negative log loss of 0.43.

8 Error Analysis

We saw our models were predicting with high variance rather than having 2 large peaks at 0 and 1. This may have been due to improper loss function. Analysing the ROC curve may lead to better results.

Secondly we noticed, in cases wherein faces were visibly fake were also causing issues, but cases wherein glasses were missing were clearly identified to be fake. As we had filtered out the presence of non facial data and primarily focused on one face per image so images with multiple people, filtered out a lot of data.

9 Discussion Future Work

In interest of compute we filtered out a lot of frames of video, completely removed the temporal aspect of video for Attention based MIL. We also removed all audio based features, which could have had significant use for identifying fake videos.

Kaggle contests commonly use ensembled models of multiple classifier to produce better results.

Pretrained GAN discriminator[14] have also been used to identify fake images. Adding such models in our ensemble would identify deepfakes generated by the same technique.

Further we noticed that size of MIL model wasn't as large as the InceptionResnet[10], so ideally using Resnet Features in MIL models could have helped.

Acknowledgments

We would like to thank Professor Christopher Kanan for mentoring our team on devising a novel methodology for deepfake detection and suggesting us to look into multiple instance learning techniques and its applications. We are also thankful to Prof. Jin Sun for guiding us through the project requirements and helping us along the way in refining our approach towards testing and implementing a novel technique to a novel problem.

References

[1] Deepfake Detection Challenge, 2020.

[2] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[3] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997.

[4] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[5] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[6] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[9] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV 2019*, 2019.

[10] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[11] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[12] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

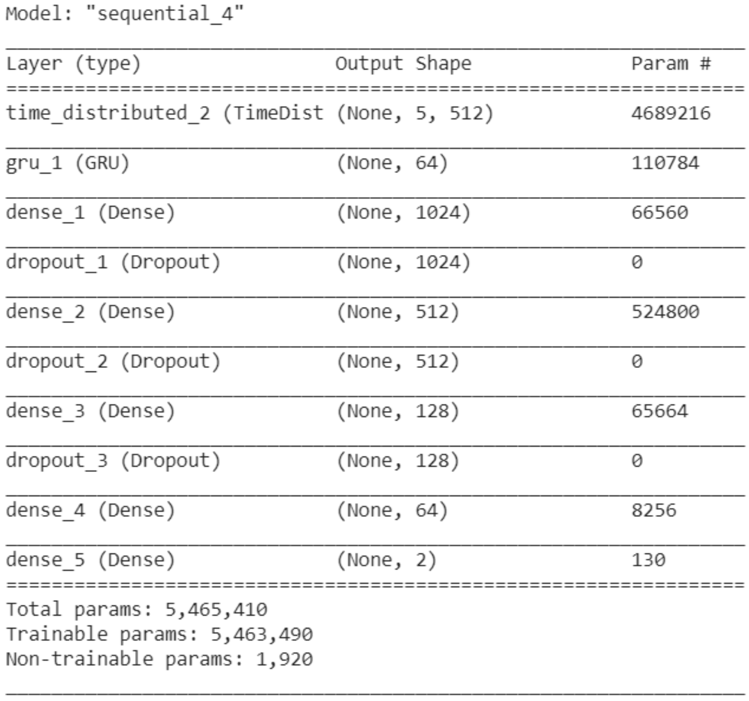
[14] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515*, 2019.

9.1 Appendix

Everyone is equally responsible for all parts of the project, the members listed below are for continuous drive and morale.

Table 4: Contributions

Contributor	Responsibility
Sarath	Experimentation with network structure
Avnish	Sampling for MIL
Ananya	Baseline development
Pratyush	Infrastructure for scaling model
Prashant	Data storage, preprocessing and management
Ravi	Evaluation over diverse datasets



```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
time_distributed_2 (TimeDist (None, 5, 512)            4689216   
_________________________________________________________________
gru_1 (GRU)                  (None, 64)                110784    
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              66560     
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0         
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800    
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_3 (Dense)              (None, 128)               65664     
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0         
_________________________________________________________________
dense_4 (Dense)              (None, 64)                8256      
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 130       
=================================================================
Total params: 5,465,410
Trainable params: 5,463,490
Non-trainable params: 1,920
_________________________________________________________________
```

Figure 1: The image describes the infrastructure of the Recurrent Neural Network used to classify the deep fake images