Using deep learning to detect deepfake videos

Atharva Shende[1], Shubham Paliwal[2], Tarun Kumar Mahay[3]

[1]Department Of Applied Mathematics, Delhi Technological University, (India)
[2]Department Of Applied Mathematics, Delhi Technological University, (India)
[3]Department Of Applied Mathematics, Delhi Technological University, (India)

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: In recent times, software based on deep learning, due to their ease of availability have made the modelling of real- looking face swapping in videos very easy that leave little signs of forgery. Such forged videos are termed deepfake(DF) videos. Manipulation of digital videos has been illustrated for many decades via an adequate usage of visual effects. Recent progress in the field of deep learning has caused an abrupt increase in the realism of forged content and they can be made very conveniently.These AI-produced means are also called by DF. Forming the DF with the artificial intelligence tools is an easy task. Developing an application to find out whether the given video is a deepfake isn'tan easy thing to do. This is mainly because any such algorithm's training will require lots of computations. To accomplish this daunting task, we decided to use convolutional neural networks and recurrent neural networks. We begin by extracting features at the frame level using CNN. We will train a recurrent neural network using those extracted features. Our RNN will then be able to identify if a video is fake or not and also check for any temporal variations between frames which were caused by the deepfake forming tools. We will compare the performance of our model with some results from a standard data set. We will keep on improvising our model till it becomes good enough to work with real world data.

Keywords: Artificial Intelligence, Convolutional Neural Network, Deep Fake Detection, Deep Learning, Long Short Term Memory

1. Introduction

1.1. What are deepfakes?

Deepfakes are forged audio or video recordings that resemble the real source. Once commonly produced by intelligence agencies, like the CIA, and professional special effects artists but nowadays anyone can download deepfake application like faceswap and create compelling forged audio or video.

Till now, only amateur hobbyists have majorly used deepfakesto make politicians say funny things. But someone could effortlessly create a DF of an emergency alert warning thatan attack is going to occur, or affect an election by creating a forged audio or video recording of someone from the political party just before the polls. [1]

1.2. Why are deepfakes dangerous?

The ill use of deepfakes can have a huge impact on the security and economy of the country. It can cause harm to individuals and democracy. Deepfakes will furthermore erode already decreasing trust in the media. Threats of deepfakescan be categorized into following categories:

Threat to Individuals: Initially vengeful use of DF was seenin pornography, causing emotional trauma, and in few cases, violence towards the person.

Threat to Society: Artificial media based upon AI may speed up the already diminishing trust in media. This erosion is leading towards a culture of factual relativism, wearing the already worn-out fabrics of civil society.

Threat to Democracy: DF of a politician can destroy his/her image and reputation and thus may strongly influence thecourse of an election.

Threat to Businesses: Deepfakes can be used to impersonate identities of business leaders and executives to carry fraud andmarket manipulation. [2]

1.3. How deepfakes are made?

DF can be generated by autoencoders. At the highest level, auto encoders work like this, when data is processed suchas image data gets compressed by an encoder. This is doneto suppress the effect of noise in the data and to reduce computational complexity. Then, the original image can be restored by passing a compressed version of the image througha decoder.

The concept of DF can be explained by following example, let's say we want to create a deep fake that blends Van Gogh'sStarry Night and De Vinci's Mona Lisa. To do so, we trainthe auto encoders for different datasets, we allow the encodersto share weights while keeping their decoders separate. In this way, an image of Mona Lisa can be compressed accordingto general logic considering things like illumination, position and expression of her face. But when it gets restored, it will be in accordance with the logic of Starry Night painting. It is analogous to a crime sketch. The descriptions from a witness (encoder) are features and are made use of by some compositesketch artist (decoder) to recreate an image of the suspicious person.

This is followed by training of encoders as well as decoders.This is done with help of backpropagation so that the input matches most nearly with the output. This process takes a lot of time.

After training, the video is processed frame wise to switch aperson's face with some other person. One person's face (let'ssay person A) is extracted out using face detection and passed on to the encoder. But inplace of passing it back to its original decoder, decoder of person B is used to rebuild the picture. That means, a person B is being created using the features ofA in the original video. After that this new face is merged intothe original image.

The encoder detects angles of the face, facial expression, skin tone, amount of lighting and some other information which is significant to recreate the person A. By using this second decoder to recreate the image, we're sketching thesecond person with the characteristics of the first one. [3]

2. Literature review

In recent times there has been an eruptive hike in creation and illegal use of DF videos. DeepFakes are a threat to society,democracy and businesses. Due to this increased threat, there is grave need for a deep fake video detection application. Below are three major methods used for detection of deepfakesusing deep learning.

Uncovering DF videos by detecting face warping artifacts: describes a method based on the idea that fake videos are generally created using warping techniques such as rotation, scaling and shearing to match the configuration of originalvideo. This lack of consistency between warped face area and surrounding space can be used to detect artifacts using CNN based models. [4]

The approach illustrated here is based on observation thatmost of the deep fakes that are generated today are of limitedresolution and have a clear distinction between warped facesand surrounding regions based on resolution and this contrastin resolution can be used to find if a video is fake or not. [5]

Uncovering AI-produced videos by detecting eye blinking [6] illustrates a method based on blinking of the human eyeto detect deep fake videos generated using neural network models. It has been seen in DF videos that blinking of eye, something which is a physiological signal, is not well produced in these videos. It has been researched and foundthat a biological eye has an interval of 2-10 seconds between consecutive blinks and the blink lasts for about 0.1-0.2 secondsbut in fake video this blinking rate is much lower because of asmaller number of images of people with closed eyes available.This unique method has shown promising results in detection of fake videos.

This method only used absence of blinking as a clue but it does not use other parameters for identification of deep fake videos like face's wrinkles etc. The model we are proposingis taking all these parameters into consideration.

Using capsule networks to detect forged images and videos [6] describes a way that takes help of a capsule network to identify that the given image or video is fake or not in various scenarios, like replay attack detection and computer-generated video detection.

In the method they used, they have included some noisein the training phase which probably wasn't the right way to go. That model might have done decently in the dataset they took but might have failed when it comes to real time data considering the training noise. We believe that our model will do well in both the scenarios.

3. Research work done

We started our learning through online courses on Deep Learning. Then we read the following research papers - Exposing DF Videos by Detecting Face Warping Artifacts [4],Exposing AI Created Fake Videos by Detecting Eye Blinking [6], Using capsule networks to detect forged images and videos [7]. Our model is inspired by these three research papers.

4. Problem statement

Research work that we have carried out has led us to our problem statement. Now since we have an authentic problem statement and resources to go for a solution. The problem statement helped us to be clear with our prerequisites,challenges and direction. The literature review and all theimplementation are a byproduct of this problem statement.

5. Methodology

There were some online courses like NPTEL IITM deep learning course that helped us getting through project.

Course like Deep Learning with PyTorch: Zero to GANs [8]was imperative in helping us write clean and efficient code forthe project.

5.1 DEEP LEARNING

Artificial Intelligence has a new branch called Deep learning. This is driven by data, different layers of machine learning models are stacked over each other. Complex inputs can be learnt and complex outputs can be generated by this model.

A neural network is used to create deep learning models. Similar to neurons in a biological brain, a neural network is a graph that contains nodes which are connected by edges. The nodes represent neurons in a brain. The connections between the neurons are depicted by the edges. As a first step, data is given through the input neurons to the neural network. After that, some mathematical operations are performed on each neuron’s data.

The resulting value of each neuron is passed on to its connected neuron. This process is repeated for all of the nodes in the hidden layer of the network. Same is done for the edges in the hidden layer.

In the final step, the result from the output neurons is used by our model to make a prediction. A neural network with more than one hidden layer is called a deep neural network. More complex patterns and functions can be modelled using a deep neural network with more hidden layers.

Let’s say we need to train a deep neural network to identify human faces. In the first step, a set of labeled images is fed to the input layer of this network. We do this to teach the network about the faces of each person along with their name.

Recognizing the face can be broken down into various sub- problems. The first layer would recognize various primitive geometric shapes like small curves, little edges, diagonal lines etc. Similarly more complex features like mouths, noses, ears, and eyes, are learnt by the second hidden layer.

Continuing this idea, the entire face structure would belearnt by third hidden layer. Most complete depiction of a person would be captured by the Output layer, in our case the name of the person whose face is fed, is learnt by the model. Complexity of features that a layer learns is increased in each subsequent layer. As an effect of each additional layer, the data becomes more abstract.

5.2 RECURRENT NEURAL NETWORKS

Recurrent neural networks , also known as RNN are a class of artificial neural networks specialized in techniques that are helpful in handling sequential information .The main functionality of an RNN is the ability to retain the results of preceding computations and use that info in the present computation .Thus making our models fit to model context dependencies in inputs of random length so as to create a right configuration of the input which is the perfect fit for natural language processing and sequence data like videos applications . As we are feeding a sequence of words into the RNN, the state gets updated for each word being input as a result , the state essentially becomes a representation of all the words which have been processed so far and since the state gets updated in a sequential manner ,the state will also contain information about the order of the words as well as the words themselves .Taking an example sentence of ' deep learning is hard but fun ' , consider the states at each step as the RNN is processing this sentence when deep is fed into the RNN the state contains the representation of just the word deep , next when we feed learning into the RNN , it will update the state which had a representation of just deep to now contain a representation of deep learning as the RNN continues to get the words from the sequence . The final state contains the representation of deep learning is hard but fun .The final state of the RNN contains both semantic information of the words in the sentence as well as sequential information regarding the order of the words which is perfect to understand the sentence . Since it works just like our brain. The usage of recurrent neural networks isn't only restricted to generation of text , image captioning , machine translation and authorship detection even though these applications will not replace any humans , it's believable that with more training data in the larger model ,a neural network would be able to synthesize new reasonable patient abstracts .[9]

5.3 LONG SHORT-TERM MEMORY

Long-term memory (LSTM) is a modified RNN also called artificial RNN architecture. LSTM has feedback connections which makes it different from the normal feed forward neural network. In addition to single data points, an important feature of LSTM is that it can process whole sequences of data like videos which contain numerous frames. Time series data can be best predicted, processed and classified with the help of LSTM networks. This is because a time series can have delaysof unknown time periods between two events.

The most important advantage of LSTM over RNN is that the vanishing gradient problem was addressed here.

6. Implementation

Deep Fakes are a matter of grave concern and it is a needof hour to develop a tool for detection and analysis of deep fakes. Our model is easy to understand. If an application is made using this model, it would be helpful in limiting theflow of Deepfakes on the internet. This can also be used along with popular apps like Twitter and Instagram and enhance their security and allow the users to verify the videos before sharingthem. We aim to constantly evaluate our project on parameterslike user experience, precision, security and authenticity.

6.1 DATASET

We have used the Celeb-DF dataset. [10] We mixed datasetsin such a way that it contains equal proportions of fake and real videos. Thus, we had 70-80 percent of training data and 20-30 percent data was used for testing.

6.2 PREPROCESSING

We started with splitting of video into frames. Next step was detection of the region that contained the face. This was followed by cropping the frame. We found the average of our videos to maintain consistency in frames count. We ignored the frames which didn't contain faces.

All our videos had a frame rate of 30 fps, so in processinga 10 second video, we had to deal with 300 frames, which would have been heavy computationally. So for this model, we used only the initial 100 frames for the purpose of model's training.

6.3 MODEL

Our model consisted of resnext50 32x4d and a layer of LSTM. Data Loader loaded the preprocessed face croppedvideos. This was followed by splitting of videos into two sets- one for training and one for testing. After this, the frames from the processed videos were given as input to our model. Here, training and testing took place in mini batches.

6.4 ResNext CNN FOR FEATURE EXTRACTION

Extraction and learning of frame level features was done with the help of the ResNext CNN model. This was followed by optimization of our model by addition of extra necessary layers and selection of suitable learning rate so that gradient descent for the model would converge properly .After the final pooling layer ,the 2048D feature vectors were given as inputin sequential LSTM.

6.5 SEQUENCE PROCESSING BY LSTM

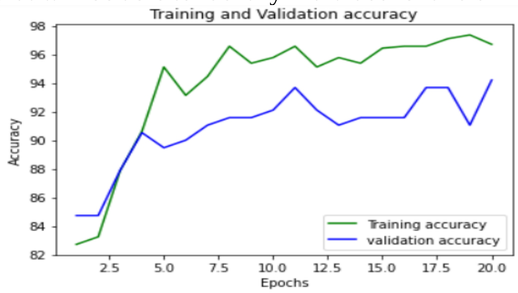
We can consider here the sequence of ResNext CNN feature vectors that are generated from input frames fed to the model as input and 2-node neural network can be assumed to be probability that sequence belongs to forged video or real video.The major challenge that we encountered was to design aclassifier that would recursively process the sequence in arelevant way. We think this problem can be addressed with the help of 2048 LSTM unit with a probability of dropout 0.4. Temporal analysis of the video can be done by comparison of its frame at 't' second with the frame of 't-n' seconds, 'n' hereis any number of frames before t seconds. This is how LSTM works to process frames in a sequential manner.

6.6 PREDICT

For prediction purposes, we passed a video to the model which had already been trained. We also preprocessed that video that brought it in the trained model's format. Then we split our video into frames and finally we did face cropping. We didn't store the video into our local storage. Instead of that, we directly passed the cropped frames to the model we trained. The detection took place there.

7. Results and analysis

We have designed a model that will be able to identify if a video is fake or not.

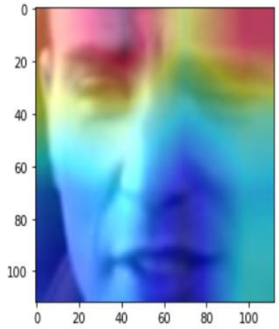


```
[[158   3]
 [  8  21]]
True positive =  158
False positive =  3
False negative =  8
True negative =  21
```

Our Model displays if the video is real or is it fake. Output also includes the confidence of that prediction. We have shown two such examples here.

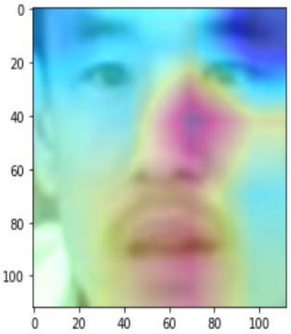
We used 190 of the CelebDF dataset as a test set. On which our model predicted 94.21% correctly.

confidence of prediction: 96.53757214546204



REAL

confidence of prediction: 99.99945163726807



FAKE

Conclusion

In an attempt to check if a video is fake or not, the challenge here was in pre-processing, because the dataset was huge and maintaining uniformity in videos was difficult.

In our entire process in this project, we didn't consider audio of the video as a factor. So it wasn't able to detect the video where the audio is fake. But we will find a way to implement that in the future.

References

A. N. N. NOW, "How deepfake videos are made." [Online]. Available: https://youtu.be/jlBNF1SucLo
B. S. Adee, "What are deepfakes and how are they created?" [Online]. Available: https://spectrum.ieee.org/tech-

talk/computing/software/what- are-deepfakes-how-are-they-created

C. J. Hui, “How deep learning fakes videos (deepfake) and how to detect it?” Medium Corporation, vol. 28, 2018.

D. Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” arXiv preprint arXiv:1811.00656, 2018.

E. D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018, pp. 1–6.

F. Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018, pp. 1–7.

G. H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2307–2311.

H. A. N. S, “Deeplearning with pytorch zero to gans.” [Online]. Available: https://jovian.ai/learn/deep-learning-with-pytorch-zero-to-gans

I. S. LABS, “Understanding deep learning: Dnn, rnn, lstm, cnn and r-cnn.” [Online]. Available: https://medium.com/@sprhlabs/understanding- deep-learning-dnn-rnn-lstm-cnn-and-r-cnn-6602ed94dbff

J. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in IEEE Conference on Computer Vision and Patten Recognition (CVPR), 2020.