Multi-Modal DeepFake Detection using Spatial and Frequency Domain Analysis

Thesis Submitted in partial fulfillment of the requirements of

BITS F421T: Thesis

By

Sivaramakrishnan KN

2017A7PS0153H

Under the supervision of

Dr. Abhinav Dhall

and

Dr. Manik Gupta

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(May 2021)

Birla Institute of Technology & Science, Pilani
Hyderabad Campus

CERTIFICATE

This is to certify that the thesis entitled **Multi-Modal DeepFake Detection using Spatial and Frequency Domain Analysis** submitted by **Mr. Sivaramakrishnan KN** (ID No. 2017A7PS0153H) in partial fulfillment of the requirements of the course BITS F421T Thesis, embodies the work done by him under my supervision and guidance.

Signature of the Supervisor
Dr. Abhinav Dhall
Designation
Date:

Signature of the Co-supervisor
Dr. Manik Gupta
Designation
Date:

Contents

Acknowledgement

I express my immense gratitude to Dr. Abhinav Dhall, Department of Human Centered Computing, Monash University, for accepting me under his tutelage to work on this thesis. His support and guidance have helped me put together this entire project. I would like to thank him for including me in the weekly meets which made me feel connected to the rest of the on-campus team despite not being able to be present there myself. His feedback was immensely valuable in bringing this work to shape.

I would like to thank Dr. Manik Gupta, Department of Computer Science, BITS Pilani Hyderabad Campus, for providing me this opportunity to work with such esteemed researchers and for continuously motivating me to work harder and push myself.

I would also like to thank Dr. Kalin Stefanov, Department of Human Centered Computing, Monash University, for his relentless support throughout my entire thesis duration. He has coached me at each and every step of the way and I have learned to be a more structured researcher because of him. I deeply appreciate his timely responses to all my queries and his patience in explaining various concepts and brainstorming different ideas over the past few months.

I express my gratitude to Monash University for providing me with the exceptional computational facilities even though the thesis was done remotely. This work would not have been possible without hardware capabilities of such high caliber.

Abstract

There is a battle raging between those who want to generate the highest quality deepfakes, indistinguishable to the naked eye to manipulate unsuspecting victims, and those who wish to help sift through the barrage of the endless content online and determine what is fake and what is real. In this thesis, we aim to improve the accuracy of Deepfake Detection mechanisms, but without unintentionally aiding DeepFake generators, as these generators can be trained on the latest detectors to improve their own quality of deepfakes. DeepFakes are generally created with the help of GANs or Auto-encoders, and these often leave traces in the high-level frequency domain during the up sampling operations. In this thesis we explore the features present in the frequency domain and utilize the high-level artifacts created by the DeepFake generators to aid in the detection of deepfakes. We also utilize transfer learning to extract the latent features from a lip-reading network to be used in combination with features extracted from high pass filtered images to detect deepfakes.

Introduction

Digitally manipulated videos and images have evolved from being manually photoshopped to being generated by deep learning techniques. This allows for a lot of fake content to be easily published on the internet.
Manipulated content can be created in a variety of forms. The most common manipulations done are the following-

- Entire Face Synthesis – this is a case where a completely new face which has never existed before with the use of GANs. This is used to create fake profiles and bait unsuspecting people.
- Identity Swap – This is the most controversial form of manipulation, where the identity of a person from a video is swapped out with the identity of another person. This is highly dangerous as it results in fake news being circulated on social media and the defamation of personalities by making people believe they have done things they haven't actually done.
- Attribute Manipulation – in this form of manipulation, only certain features from a person's face are changed, for instance hair and skin color, addition of glasses and ageing of the face. This is quite useful for consumers who wish to try out their new looks before they buy a product.
- Expression Swap – in this method, the facial expressions from one person is emulated onto the face of another person. This is again quite harmful as it can be used to create highly realistic fake propaganda.

Thus, we can see that there is a dire need to create accurate detection mechanisms which can keep a check on the content floating around on social media to ensure that people are falling for traps or being manipulated by fake propaganda. In this thesis we explore the artifacts generated by models in the frequency domain and utilize them to improve the accuracy of deepfake detection.

Related Work

There are a multitude of approaches for DeepFake detection, these are a few paradigms used for the same-

1. CNN-Based Methods

The XceptionNet (Chollet 2017) is a convolutional neural network which performs residual learning, i.e., it can learn and classify fake images generated by new methods without losing out accuracy on previously learnt methods (avoids catastrophic forgetting) by using depth wise separable convolutions (Rebuffi, et al. 2017). This method has been tested on most major deepfake datasets and has shown consistently high accuracies on all of them.
The VGG-Net is another popular convolutional neural network-based approach which is used in face recognition systems. The pre-trained VGGFace (Parkhi, Vedaldi and Zisserman 2015) network was taken, and the final fully connected layer was replaced with a binary SoftMax activation for classification of real and fake images (Do Nhu, et al. 2018). The advantages of using a VGG-Net were that it was faster than previous training methods such as the Alex-Net and it supported a greater depth of hidden layers.
Other approaches for classification using CNNs include using transfer learning from other tasks to reduce the training time for deepfake classification. (Ding, et al. 2020) used the weights from the ResNet-18 architecture which was previously trained on the ImageNet (Deng, et al. 2009) dataset and fine-tuned it to classify deepfakes.

2. Specific Artifact-Based Methods

Other novel methods of detecting deepfakes utilize artifacts left behind while generating them. (Yuezun Li 2019) detect fakes by comparing the modified regions and their nearby areas by exploiting the resolution inconsistency which results when the faces are warped onto a video. Similarly, the facial landmarks from a video are plotted and the head pose is estimated from both the central region of the face and the identified landmarks. They show that the difference between the two estimated head poses is higher in the case of deepfake since the central region is synthesized from another face and use this difference to train their classifier (Yang, Li and Lyu 2019).
There is also another interesting approach which uses biological signals (Ciftci, Demir and Yin 2020) to detect deepfakes. This method is independent of the model used to generate deepfakes and hence showed great results on a variety of in-the-wild datasets. They use the fact that a signal must coherent over both an individual image as well as over a period of frames to train a convolutional network which can accurately classify deepfakes.

3. Temporal Inconsistencies

Most of the methods discussed so far have tried to classify deepfakes by analyzing individual frames, but it has been observed that the temporal inconsistencies could also be used to enhance the quality of deepfake detection. (Sabir, et al. 2019) use a Recurrent Neural Network on top of a baseline CNN, the ResNet in this case, to aggregate the inputs over a period of time. They found out that a bidirectional RNN performed the best after comparing various networks.

4. GAN / Auto-Encoder Based Methods

(Guarnera, Giudice and Battiato 2020), again, exploit the fact that the correlation of local pixels is independent of the model used to generate it and train a GAN based network to estimate a kernel which would be similar to the GAN which generated the deepfake. The pitfall of this network is that it can only detect fakes from GANs it has been trained on previously and fails to distinguish new ones accurately, but it can identify which GAN was used to generate a given deepfake.

Another approach which creates a more generalized model which can detect unseen manipulations uses an Autoencoder trained on fake and real images, to represent the data in a latent space, and a classifier is then trained on features from this latent space to separate the real and fake images (Cozzolino, et al. 2018).

5. Frequency Domain Analysis

The problems that arise with using deep learning-based detection methods is that these can easily be learnt by another GAN and generate better quality fakes by training the discriminator on the detector. So, a way to counter this is to use features from the frequency domain spectrum obtained from an image as a feature.

The Discrete Cosine Transform (DCT) was used as the frequency domain transformation method and a local frequency analysis of this feature was performed upon it (Qian, et al. 2020). The XceptionNet was used as a backbone network to learn from the frequency domain features. This model was shown to be more robust to detecting fakes in both low- and high-quality images as compared to the previous state-of-the-art.

In (Durall, Keuper, et al., Unmasking DeepFakes with simple Features 2020), a Discrete Fourier Transform is applied on an image to convert it to the frequency spectrum. An Azimuthal Average of the 2-dimensional output from the previous step is done to compute a 1-dimensional feature upon which a simple classifier is trained to detect fakes. This method has been used as a part of this experiment for creating the baseline model.

6. Transfer Learning

Work done in the field of visual speech recognition is of great value to us since the same features recognized by lip reading networks could be used to successfully identify deepfakes. It has previously been show in (faceforensics) that the XceptionNet (xception) which was trained on the ImageNet database was able to produce great results in the task of deepfake detection by retraining the model after replacing the last fully connected layer with 2 outputs instead of 1000. This shows that networks which perform the task of image classification well can also generalize to perform the task of deepfake detection well.

Data

A good dataset is required for training any network, and lately there have been great advancements in the creation of dataset which has a mix of videos which have been created by various deepfake generation methods and videos of varying quality. This ensures that models trained on these do not overfit to identifying only specific types of fakes and are more generalizable. The following are two of the datasets upon which the following experiment has been conducted.

FaceForensics++

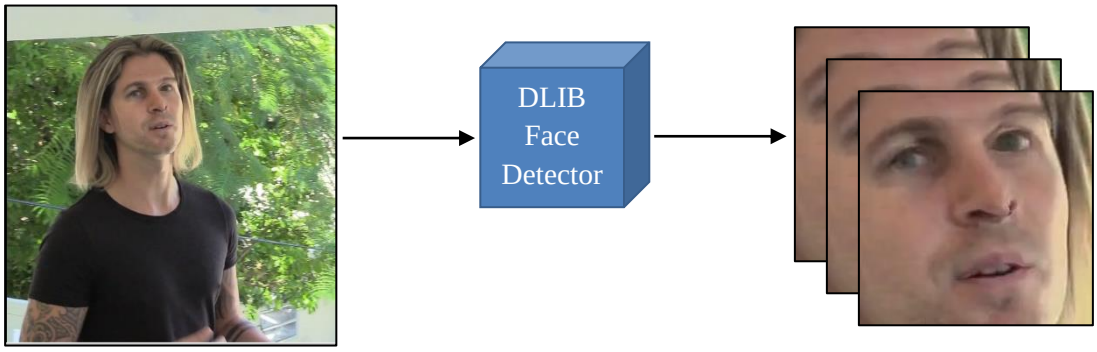
The FaceForensics++ dataset was released by Google & JigSaw in 2019 .
The dataset comprises of manipulations done using two computer graphics-based methods, Face2Face (Thies, Zollhofer, et al. 2016) and FaceSwap, along with two deep learning based methods, DeepFakes and NeuralTextures.
A 1000 original videos, with 977 actors, of quality varying from 480p, 720p or 1080p were taken from the internet and the above 4 algorithms were used to create manipulated videos from them. The videos were compressed using H.264 codec, which is the standard method used in most social media sites, to generate both Low Quality and High Quality videos to create a more realistic dataset. A preprocessed subset of this dataset taken from (prepro_deepFake.7z - Google Drive) was used in this experiment for evaluating the performance of the 1D Power spectrum analysis.

DFDC

The DeepFake Detection Challenge Dataset was released by the Facebook AI group (Dolhansky, et al. 2020) as part of their initiative to enhance the state of the art in deepfake detection.
The manipulations in this dataset were done using DFAE (DeepFake Autoencoders), Morphable-Mask face swap, Neural Talking Heads (NTH), FSGAN and StyleGAN. These methods are a combination of both low-effort, less-realistic deepfakes as well as state of the art methods to better represent the DeepFake content floating around the internet.
It contains over 100,000 unique fake videos made by 960 actors, with over 10 million frames in total to train from. They claim to be a part of the third generation of DeepFake datasets, with the first generation containing less than a thousand videos (or less than 1 million frames) and the second generation containing between 1000 and 10000 videos (or 1-10 million frames). The videos in this dataset were taken in varying lighting conditions which helps generalize the performance of models trained on it. As a result, the videos in this dataset are of varying resolution, with the manipulated region being either a 128x128 or a 256x256 face crop. A subset of DFDC, containing 400 training samples and 400 test samples was used in this experiment.

Methodology

Facial Feature Extraction

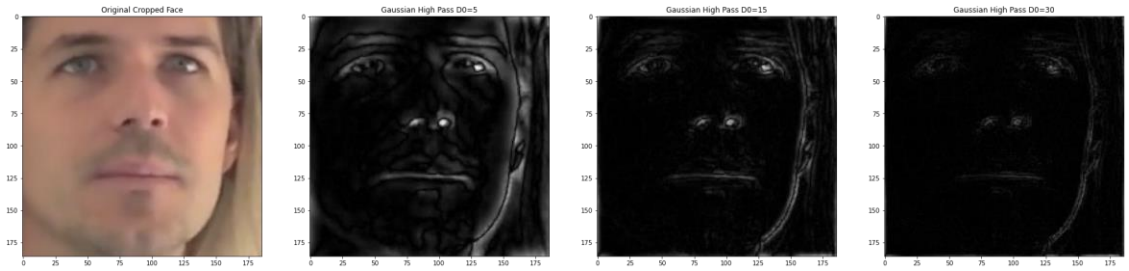


Original Uncropped Video *Chunks of extracted Faces*

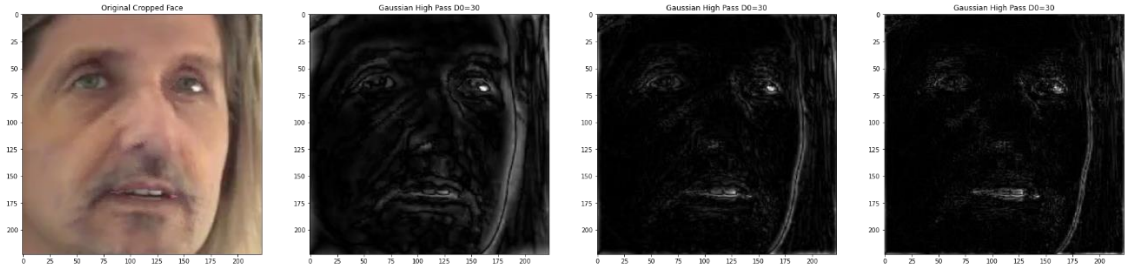
The HoG (Histogram of Oriented Gradients) Face Detector from the Dlib library was used to detect the bounding box of the faces present within the frame. This is one of the most widely used libraries for face recognition and has been trained on the ImageNet dataset and has an average detection rate of 92.68% (dlibhog). The frames were cropped to different sizes depending on which network was being used to extract the features.

A. Lip Reading Network - Each frame was cropped and resized to a uniform size of 160x160 and were stitched into 1 second clips of 30 frames each to fit the VGG Face Model.
B. Frequency Domain Analysis – Each frame was similarly cropped, run through a high pass filter as explained in the next section and resized to a uniform size of 224x224 to fit the Resnet18 architecture.

Frequency Domain High Pass Filtering



Real Face *D0=5* *D0=15* *D0=30*



Fake Face *D0=5* *D0=15* *D0=30*

To filter an image in the frequency domain the following steps were followed –

1. Frequency Domain Transformation - The extracted frame is transformed to the frequency domain by computing the Discrete Fourier Transform (DFT) of the input image. This is done by applying the Fast Fourier Transform (FFT) function on each channel of the input image (dftwolfram).

$$F_n \equiv \sum_{k=0}^{N-1} f_k \, e^{-2\pi i n k/N}.$$

2. High Pass Filter – An Ideal high pass filter cuts off all frequencies below a particular threshold (D0). A Gaussian High Pass filter was used in this experiment to allow only the high frequencies and cut off the lower frequencies in a decreasing gradient. The DFT is multiplied with this filter function *H(u,v)* to eliminate the lower frequencies. Three different threshold values, D0 = 5, 15, 30, were used and the resulting filtered images were saved.

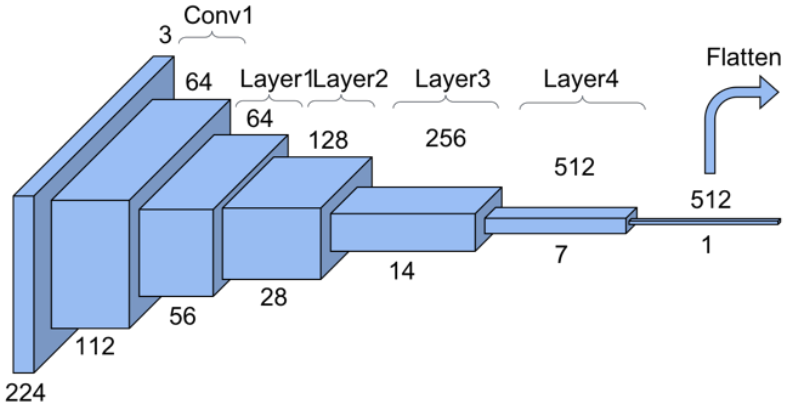
$$H(u,v) = 1 - e^{-D^2(u,v)/2D_0^2}$$

3. Inverse Fourier Transform – after the High Pass filter is applied, the image is brought back to the spatial domain by applying an inverse Fourier Transform function.

$$f_k = \frac{1}{N} \sum_{n=0}^{N-1} F_n \, e^{2\pi i k n/N}.$$

Convolutional Feature Extraction

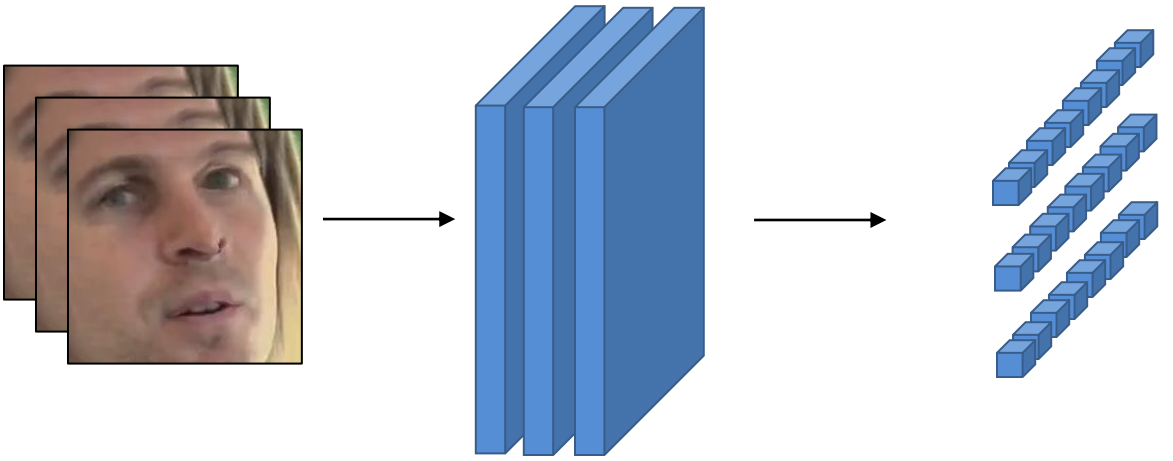
The ResNet18 architecture with pretrained weights on the ImageNet dataset was used to extract the features from the High Pass filtered images. Each of the three high pass filtered images was passed through the network and 3 sets of a 512-dimensional feature vector was extracted from the penultimate max-pooling layer.



ResNet18 Architecture

Lip Reading Features

The front-end of the Oxford lip-reading module (deeplipreading) was considered as a baseline for extracting facial features to be trained on a classifier. The intuition behind this approach was that the temporal consistency across video chunks would be captured by the features generated by the 3D convolution layer in the lip-reading network. These same features could be used to discriminate the inconsistencies observed in artificially generated images.
The visual frontend of this network comprises of a 3D convolutional network, which has a temporal width of 5 frames. The output is then passed through a ResNet18 which reduces the spatial resolution and yields a 512-dimensional feature vector as an output for each input frame. This network was pretrained on the LRW dataset (lrw) where every temporal chunk was trained to be classified as a word label.

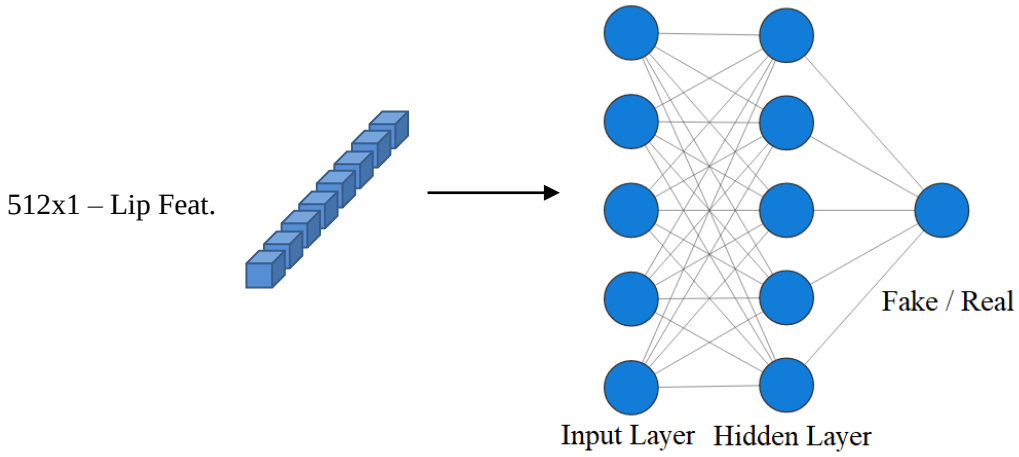


Input Video Chunk (30x160x160) *Lip Reading Frontend Output* *Feature Vector (30x512)*

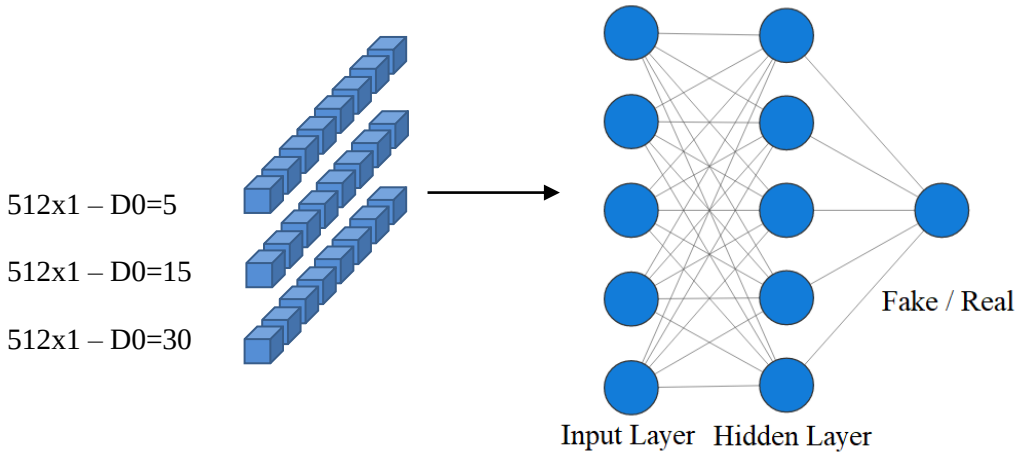
Fully Connected Layer Classification

Three individual fully connected networks were comprising of a hidden layer containing 256 and a fully connected layer with 2 output was initialized.

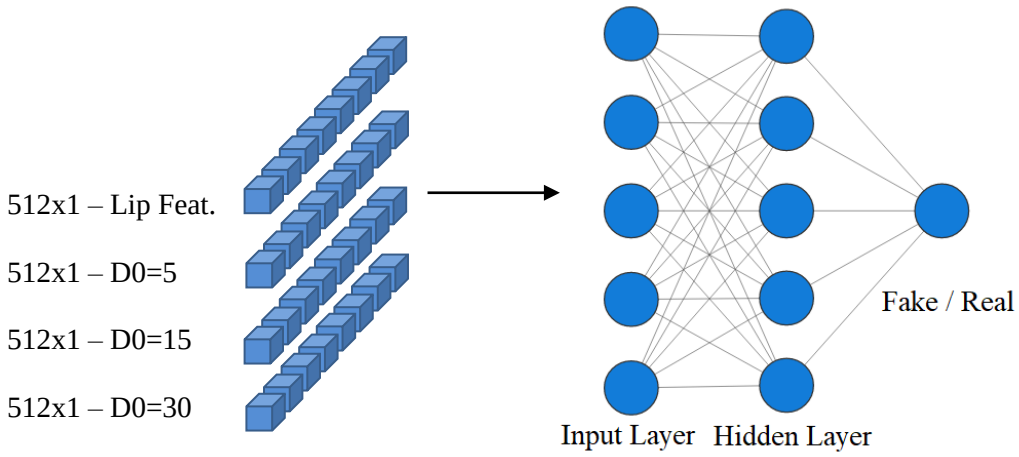
1. The Lip Reading network took a 512 dimensional feature vector as input and passed it through a fully connected (FC) layer with a 256-dimensional hidden layer and a 2-dimensional output layer which was then trained for 50 epochs with a batch size of 32, dropout rate of 0.1 and learning rate of 0.01.
2. The Frequency Domain network concatenated the weights from the high pass filtered images and took a 1536-dimensional feature vector as input and passed it through a fully connected (FC) layer with a 512-dimensional hidden layer and a 2-dimensional output layer which was then trained for 50 epochs with a batch size of 64, dropout rate of 0.1 and learning rate of 0.01.
3. The Multimodal network concatenated all 4 sets of feature vectors and took a 2048-dimensional feature vector as input and passed it through a fully connected (FC) layer with a 512-dimensional hidden layer and a 2-dimensional output layer which was then trained for 50 epochs with a batch size of 64, dropout rate of 0.1 and learning rate of 0.01.



Lip Reading Network



Frequency Domain Network



Multimodal Network

Results

Lip Reading Network

Lip Reading Features	DFDC Preview	DFDC Full
Accuracy	0.7792	0.6100
AUC Score	0.70477	0.6070
F1 Score	0.5137	0.3371
Test/Validation Set Frames	19,284	629,760
Train Set Frames	77,136	3,242,268

Table 1

Table 1 contains the comparison of results from the classification done on the DFDC Preview dataset and the DFDC full dataset using only the features from the Lip Reading network. It is noted that even though the model performed decently well on the preview dataset, its performance was reduced on the larger dataset.

	Predicted Fake	Predicted Real
Actual Fake	12,778	2,630
Actual Real	1,627	2,249

Table 2

Table 2 represents the confusion matrix of the result of classification on the validation set split of the DFDC Preview dataset. The Lip Reading Network was trained on 38,711 fake frames and 38,425 real images. Epoch 29 resulted in the best accuracy on the validation set, with a dropout rate of 0.1 for the hidden layer.

	Predicted Fake	Predicted Real
Actual Fake	321,700	204,410
Actual Real	41,191	62,459

Table 3

Table 3 represents the confusion matrix of the classification on the test set split of the full DFDC dataset. The Lip Reading Network classifier was trained on 1621637 fake and 1620631 real frames with a 0.1 train and validation split. This was then tested on a set containing 3,62,891 fake and 2,66,869 real frames. Epoch 21 resulted in the best accuracy on the validation set, with a dropout rate of 0.1 for the hidden layer.

Frequency Domain Network

	Predicted Fake	Predicted Real
Actual Fake	15129	279
Actual Real	138	3738

Table 4

Table 4 represents the confusion matrix of the result of classification on the validation set split of the DFDC Preview dataset. The Frequency Domain Network was trained on 38,498 fake frames and 38,638 real images. Epoch 31 resulted in the best accuracy on the validation set, with a dropout rate of 0.1 for the hidden layer.

Multimodal Network

This section contains the results from the classification done using the combination of Lip Reading features and the Frequency Domain features.

	Predicted Fake	Predicted Real
Actual Fake	15277	131
Actual Real	74	3802

Table 5

Table 5 represents the confusion matrix of the result of classification on the validation set split of the DFDC Preview dataset. The Multimodal Network was trained on 38,461 fake frames and 38,675 real images. Epoch 45 resulted in the best accuracy on the validation set, with a dropout rate of 0.1 for the hidden layer.

Comparison

A comparison of the performance of the three models on the same split of the DFDC Preview dataset is shown in Table 6.

DFDC Preview	Accuracy	AUC Score	F1 Score
Lip Reading	0.77929	0.51376	0.70477
Frequency Domain	0.97843	0.97314	0.94717
Multimodal	**0.98939**	**0.98620**	**0.97375**

Table 6

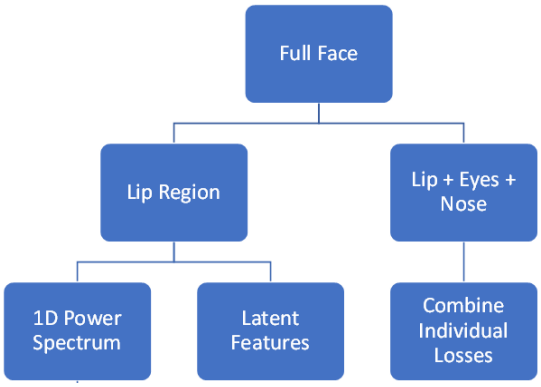
It can be seen that there is a great improvement in performance when the model was trained on both, spatial and frequency domain features.

Future Work

The performance of the Frequency Domain model and the Multimodal network must be tested upon the larger DFDC dataset to validate their robustness to different types of deepfakes. These are the ways in which the current model can be improved-

Combine Loss from Different Regions

Apart from improving the classifier trained only on the lip region, we can also consider the other regions of the face such as the eyes, hair, and nose where there is a high probability of there being a manipulation and combine the individual losses from each of these features to improve our classifier. The hypothesis here is that the combination of individual losses from each region of the face should improve the accuracy as compared to training the network over the entire face region.



Temporal

On top of this, a temporal loss can also be introduced before the final classification step by using a Recurrent Neural Network or an LSTM. It is possible to exploit the irregularities and artifacts present in fake images, when passed through a high pass filter, along the temporal dimension. The transitions between frames are smooth in real videos whereas the transition between frames have erratic jumps occasionally in fake videos. These are enhanced under a high pass filter as the artifacts are generally jagged edges which prop up in the frames which are of high frequencies.

Conclusion

In this thesis, we have explored the usefulness of analyzing features from the frequency domain and used it in combination with spatio-temporal features to create a multimodal network to detect deepfake videos. We have found that this increases the accuracy of the classifier to detect deepfakes.

References

Adouani, Amal, Wiem Mimoun Ben Henia, and Zied Lachiri. 2019. "Comparison of Haar-like, HOG and LBP approaches for face detection in video sequences." *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD).* 266–271.

Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. 2018. "Deep lip reading: a comparison of models and an online application." *arXiv preprint arXiv:1806.06053.*

Amato, Giuseppe, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. 2018. "A Comparison of Face Verification with Facial Landmarks and Deep Features."

Chollet, François. 2016. "Xception: deep learning with depthwise separable convolutions (2016)." *arXiv preprint arXiv:1610.02357.*

—. 2017. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1251–1258.

Chung, Joon Son, and Andrew Zisserman. 2016. "Lip reading in the wild." *Asian Conference on Computer Vision.* 87–103.

Ciftci, Umur, Ilke Demir, and Lijun Yin. 2020. "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP: 1-1. doi:10.1109/TPAMI.2020.3009287.

Cozzolino, Davide, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. "Forensictransfer: Weakly-supervised domain adaptation for forgery detection." *arXiv preprint arXiv:1812.02510.*

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition.* 248–255.

Ding, Xinyi, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger, and Michael Hahsler. 2020. "Swapped face detection using deep learning and subjective assessment." *EURASIP Journal on Information Security* (Springer) 2020: 1–12.

Do Nhu, Tai, In Na, Hyung-Jeong Yang, Guee-Sang Lee, and S. H. Kim. 2018. "Forensics Face Detection From GANs Using Convolutional Neural Network."

Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. "The DeepFake Detection Challenge (DFDC) Dataset."

Durall, Ricard, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. 2020. "Unmasking DeepFakes with simple Features."

Durall, Ricard, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. 2020. "Unmasking DeepFakes with simple Features."

Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. 2020. "Deepfake detection by analyzing convolutional traces." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.* 666–667.

King, Davis E. 2009. "Dlib-ml: A Machine Learning Toolkit." *Journal of Machine Learning Research* 10: 1755-1758.

Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. 2015. "Deep Face Recognition." Edited by Xianghua Xie and Gary K. L. Tam. *Proceedings of the British Machine Vision Conference (BMVC).* BMVA Press. 41.1-41.12. doi:10.5244/C.29.41.

Qian, Yuyang, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. "Thinking in frequency: Face forgery detection by mining frequency-aware clues." *European Conference on Computer Vision.* 86–103.

Rebuffi, Sylvestre-Alvise, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. "icarl: Incremental classifier and representation learning." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 2001–2010.

Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. "Faceforensics++: Learning to detect manipulated facial images." *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1–11.

Sabir, Ekraam, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. "Recurrent convolutional strategies for face manipulation detection in videos." *Interfaces (GUI)* 3.

Thies, Justus, Michael Zollhöfer, and Matthias Nießner. 2019. "Deferred neural rendering: Image synthesis using neural textures." *ACM Transactions on Graphics (TOG)* (ACM New York, NY, USA) 38: 1–12.

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

Torfi, A., S. M. Iranmanesh, N. Nasrabadi, and J. Dawson. 2017. "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition." *IEEE Access* 5: 22081-22091. doi:10.1109/ACCESS.2017.2761539.

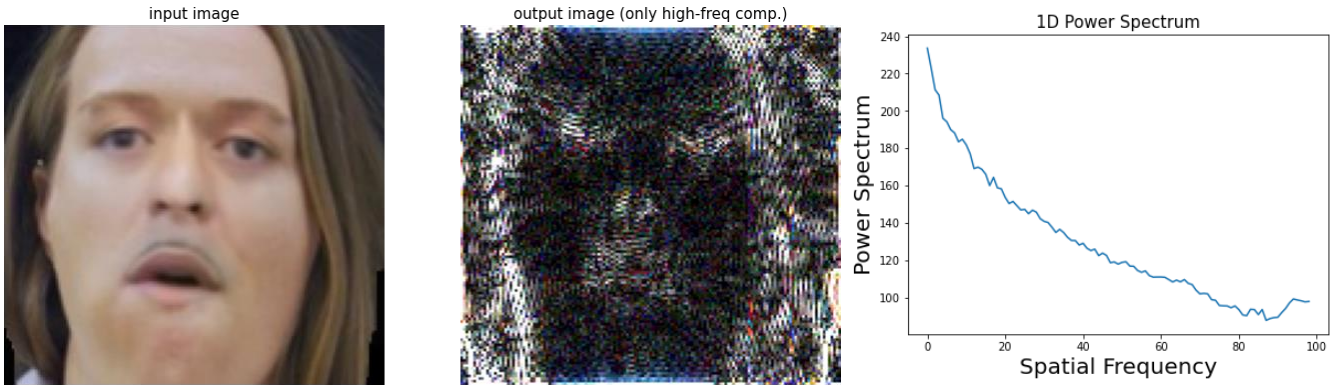
Weisstein, Eric W. 2002. "Discrete fourier transform." *https://mathworld. wolfram. com/* (Wolfram Research, Inc.).

Yang, X., Y. Li, and S. Lyu. 2019. "Exposing Deep Fakes Using Inconsistent Head Poses." *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 8261-8265. doi:10.1109/ICASSP.2019.8683164.

Yuezun Li, Siwei Lyu. 2019. "Exposing DeepFake Videos By Detecting Face Warping Artifacts."

Appendix

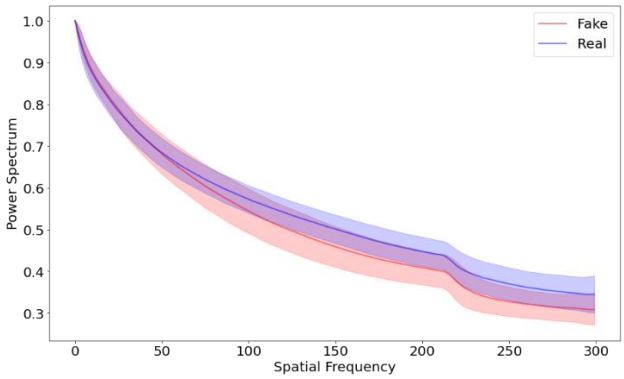
An initial study was conducted to draw a comparison between the results of the classifiers on the full-face region and the cropped lip and eye region. This was to find out how much information was contained in each of the individual regions vs. the entire face. The 68-point facial landmark detector from the Dlib library was then used to extract the lip and eye region from the identified face (Amato, et al. 2018). The frames were converted to the frequency domain by applying a DFT and a radial, Azimuthal average was calculated to obtain a 1D Power-spectrum for each video.



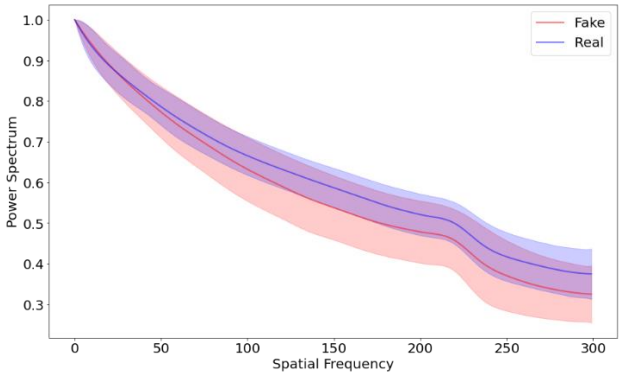
The above figure shows, form left to right, the face cropped input image, the frequency domain transformed image and the corresponding graph of the 1D power spectrum.
A standard SVM and Logistic Regression classifier were trained on the resultant 1D features to discriminate between the two classes. The SVM was trained by using both the RBF kernel and the Polynomial kernel. The following results were observed –

	FaceForensics++	DFDC Preview
SVM	0.8765	0.8699
SVM (RBF Kernel)	0.9187	0.8858
SVM (Polynomial)	0.9046	0.8699
Logistic Regression	0.8234	0.8752
No. of Fake Frames	1680	149122
No. of Real Frames	2468	21642

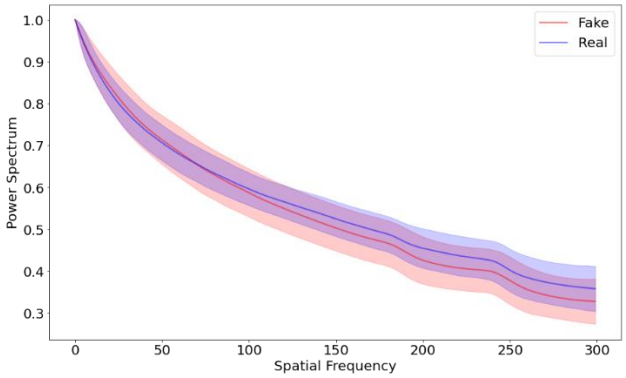
The classifier worked well on a small dataset, but the accuracy started dropping when trained on larger datasets. This showed that the 1D features weren't robust enough to handle a variety of deepfake manipulations.



Full Face Spectrum



Lip Region Spectrum



Eye Region Spectrum

FaceForensics++	Full Face	Lip Region	Eye Region
SVM	0.8765	0.775	0.742
SVM (RBF Kernel)	0.9187	0.829	0.812
SVM (Polynomial)	0.9046	0.790	0.756
Logistic Regression	0.8234	0.770	0.721

It can be noted that there is a noticeable decrease in accuracy when only considering the Lip region, this can also be observed from the frequency graphs. In Fig.1 the split between the real and fake graph is wider than in Fig.2, where the overlap between the blue and the red region is more, which indicates that simple classifiers like SVM and LR will be able to separate it more easily.

From this, we can conclude that some information is lost when taking into consideration only the lip region from the entire face. The loss in information is due to the fact that not all videos have manipulations occurring near the lip region, but there is a possibility that when this is combined with the information from other regions such as the eyes and nose the overall accuracy can increase.