A Novel Approach To DeepFake Detection

Pritam Mishra[1] , Ignatius Ezeani[2] , Rob Young[2]
Department of Computing And Communications,
Lancaster University

Abstract

- *Manipulated images and videos have become increasingly realistic due to the tremendous progress of deep convolutional neural networks (CNNs), such progress raises a number of social concerns related to the advent and spread of fake information.*
- *In this project we investigated some of the state of the art detection methodologies including pre-trained CNNs (VGG16, ResNet50, Xception, InceptionResNetV2), MESONET, Baseline CNN from scratch, CNN-LSTM architecture (Several variations - sequential, Integrated approach, altering pre-trained models) , Domain Adapted GAN (Discriminator), 3DCNN (A variation of C3D architecture) to evaluate and compare their performance on Kaggle deep-fake detection challenge data-set.*
- *The findings indicates that performance of these proposed models rely on not only more computing resources, longer training time, accurate face detection, video compression rate, image noise, video resolution but also huge number of training examples to form a comprehensive training data distribution to discern between fake and authentic videos.*

Motivation

The social implications of deepfake videos that motivated this project are as follows,

- In January 2019, Fox affiliate KCPQ aired a deepfake of Trump during his Oval Office address, mocking his appearance and skin color[5].
- In June 2019, the United States House Intelligence Committee held hearings on the potential malicious use of deepfakes to sway elections[3].
- In June 2019, a downloadable Windows and Linux application called DeepNude was released which used generative adversarial networks to remove clothing from images of women[2].

DeepFake Generation

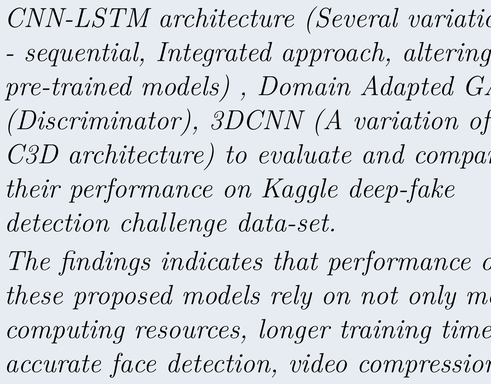

Figure 1:DeepFake Generation with Two Autoencoders(shared Encoder)[4]

Deepfake is a deep learning methodology which aims to replace the face of a targeted person by the face of someone else in a video with the help of generation techniques such as Autoencoders or DcGAN [1].

Vulnerabilities

- Pixel level artifacts observed in faces of deepfake videos
- Temporal inconsistency among image frames of a tampered video.

Proposed DeepFake Detection Architecture

The highlights of proposed DeepFake Detection system architecture can be summarized as follows,

- Keyframe extraction from videos with maximum entropy from every image sequence.
- Comparison between face detection models HaarCascades, DLIB, MTCNN and RetinaFace to obtain optimal face extraction model with highest accuracy.
- Implementation of pre-processing techniques (viz. Standardisation, Normalization, Normalization-Standardisation and Standardisation-Normalization) on extracted face images for better loss convergence and improved accuracy of deepfake models.
- Finally comparison of proposed deepfake detection models (Baseline CNN(3 variations), MESONET Transfer Learning (Vgg16, ResNet50, Xception, InceptionResNetV2), CNN-LSTM (sequential and integrated model), 3DCNN and domain adapted DcGAN(Discriminator) in terms of performance (average weighted precision, recall, f1, accuracy, logloss).

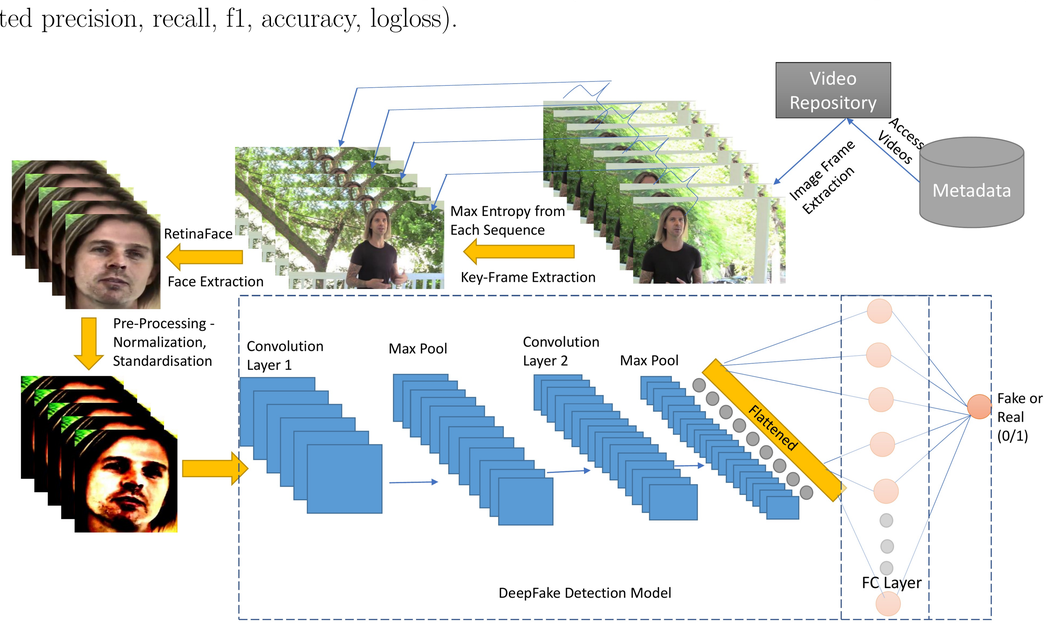

Figure 2:Proposed DeepFake Detection Architecture

Evaluation of proposed methodologies

The evaluation of top 2 models during training-validation out 15 proposed deep-fake detection models has been demonstrated as follows,

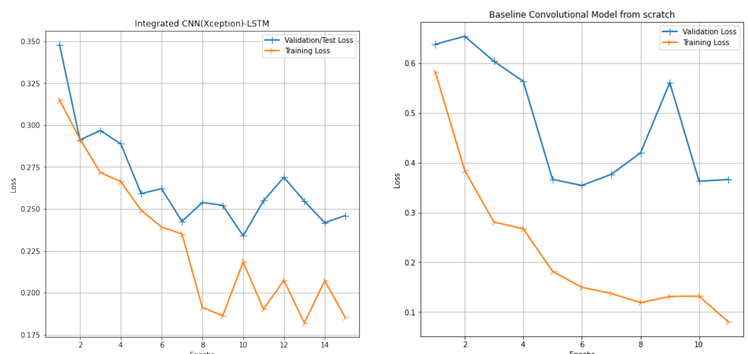

Figure 3:Evaluation of Proposed Xception-LSTM and Pritam's Baseline 16-32 Model

From the loss charts of Figure 3, it can be observed the minimum loss of baseline model and Xception-LSTM has been found to be 0.35 and 0.23 respectively on the validation set which is very impresive, it could further be evaluated on the testset.
The performance of top 5 proposed models out of 15 on kaggle deepfake detection testset is shown below,

Table 1:Evaluation of Proposed DeepFake Detection Models(Only Top 5 shown out of 15 proposed models

Proposed Model Name	Accuracy	Precision	Recall	F1	Log Loss
CNN from scratch (Pritam's Baseline var 1)	65	62	65	58	0.67
CNN from scratch (Pritam's Baseline var 2)	59	54	59	54	0.79
Integrated CNN(Xception)-LSTM	64	41	64	50	0.65
Integrated CNN(ResNet50)-LSTM(Fine Tuned)	63	55	63	53	12.77
Transfer Learning with ResNet50(Fine Tuned)	64	59	64	53	12.43

Table 2:Performance of top 5 wining solution vs top proposed methodology

Team/Model Name	Overall Log Loss
Selim Seferbekov	0.4279
WM	0.4284
NTechLab	0.4345
Eighteen Years Old	0.4347
The Medics	0.4371
Proposed Integrated CNN(Xception)-LSTM	0.65
CNN from Scratch (Pritam's Baseline Model var 1)	0.67

Evaluation of Pre-processing techniques on Model Performance

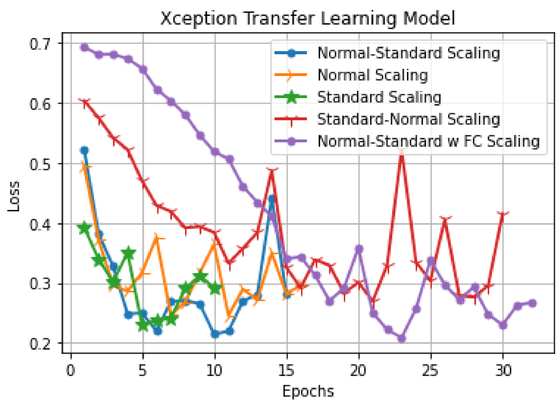

Figure 4:Effect of pre-processing Techniques on Loss Convergence

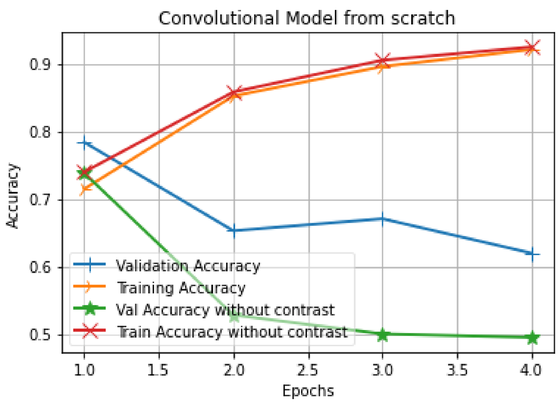

Figure 5:Effect of Image Contrast on DeepFake Model Performance

Discussion

- Even though the top proposed methodologies could not outperform the top 5 solutions from kaggle, it came close considering the limitations in this project.
- The performance of Pritam's baseline model variation 1 surprisingly performed very well with a significant low model complexity compared any of the top 5 winning solutions from Kaggle.

Conclusion And Future Work

- All of the proposed methodologies could have performed really well if enough computing resources and time were available.
- If all of the proposed 15 deepfake models could have been ensembled into one model the performance could improve as a result.

References

[1] Nozick et al. Afchar D. "MesoNet: a compact facial video forgery detection network". In: *IEEE International Workshop* (2018).

[2] Koebler Jason Cole Samantha Maiberg Emanuel. "This Horrifying App Undresses a Photo of Any Woman with a Single Click". In: *Vice* (June 2019).

[3] O'Sullivan Donie. "Congress to investigate deepfakes as doctored Pelosi video causes stir". In: *CNN* (Nov. 2019).

[4] D. Guera and E. J Delp. "Deepfake video detection using recurrent neural networks". In: *15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6). IEEE* (2018).

[5] Swenson Kyle. "A Seattle TV station aired doctored footage of Trump's Oval Office speech. The employee has been fired". In: *The Washington Post* (Jan. 2019).