Model to Detect Deep Fakes

Devadharshini Ayyappan, Kriti Singh, Rashmi Datta
Advisor: Dr. Hamid Krim

Department of Electrical and Computer Engineering

Introduction

- Deepfake technology has rapidly advanced, making it increasingly difficult to distinguish real from fake videos. The potential for deepfakes to be used in disinformation, propaganda, and other malicious activities has led to a growing need for effective deepfake detection methods.
- In this project, we propose to use a combination of Multitask Cascaded CNNs (MTCNN) facenet and metric learning to detect deepfakes. In this work, we analyze several deep learning approaches in the context of deepfakes classification in high compression scenario and demonstrate that a proposed approach based on metric learning can be very effective in performing such a classification.
- Using less number of frames per video to assess its realism, the metric learning approach using a triplet network architecture proves to be fruitful. It learns to enhance the feature space distance between the cluster of real and fake videos embedding vectors. We validated our approaches on two datasets to analyze the behavior in different environments.

Datasets and Preprocessing

- We analyzed our video classification approaches using the datasets Celeb-DF and Celeb-DF-v2. This dataset comprises 52 celebrities whose interviews are available on YouTube. They considered various factors such as gender, age and ethnic group bias to make the dataset more challenging.
- Our training data consisted of approximately 30,000 images with equal distribution of manipulated and original images from both Celeb DF and Celeb DF v2 datasets.
- Our labels for the training data were categorical in nature with '0' representing the image is real and '1' representing that the image is manipulated. We have used semi-hard triplet loss.
- With 25 frames per video, we took 30,000 embedding into consideration. We generated 512 face embedding vectors using facenet.
- Then we trained a network with triplet loss function to get a better segregation of features and once we got the features we used it as a training dataset for different classifiers like KNN, SGD and Random Forest.

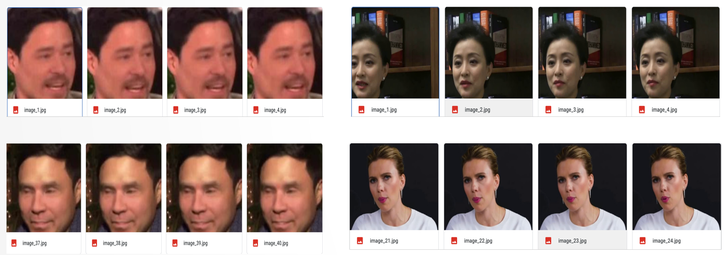


Fig. Above, the first set (left) depicts the deep-fake frames from the CelebDF-v2 dataset. The next set (right) are examples of original sequences from the same dataset.

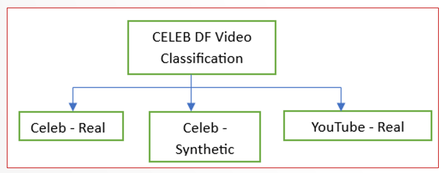


Fig.Structure of CELEB DF dataset

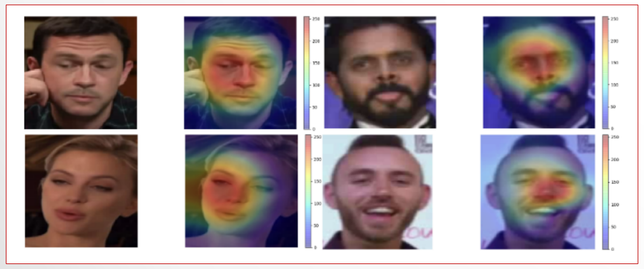


Fig. Heat map focusing on features that were learned by the network

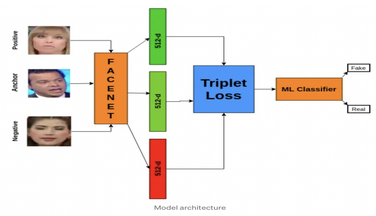
Methodology

Xception Net: Xception architecture is a linear stack of depthwise separable convolution layers with residual connections.We used Xception architecture to learn the crucial feature about real and fake faces. Xception net based on Inception V3 uses Inception module, with modification of the spatial convolutions to depthwise separable convolutions.After separating each channel, 1x1 depthwise convolutions helps network to capture the cross-channel correlations. Compared to Inception architecture convolutions, depthwise separable convolution differs in two ways: 1) Xception modules performs channel wise convolutions first, then, 1x1 convolution, compared to Inception where the 1x1 is performed earlier, and, 2) There's no non-linearity after depthwise separable convolutions.

MTCNN: Crops out images using Proposal, Refine and Output Net. Proposal network detect faces across multiple resolutions, then, refine net suppress the overlapping boxes using nonmax suppression. Finally, output network gives the bounded face using five landmarks.

FaceNet: Based on learning a Euclidean embedding per image using a deep convolutional network. The network is trained such that the squared L2 distances in the embedding space directly correspond to face similarity: faces of the same person have small distances and faces of distinct people have large distances.

The embedding is represented by f(x) ∈ R d . It embeds an image x into a d-dimensional Euclidean space. Additionally, we constrain this embedding to live on the d-dimensional hypersphere, i.e. kf(x)k2 = 1.We want to ensure that an image x _a (anchor) of a specific person is closer to all other images x_p(positive) of the same person than it is to any image x_n(negative) of any other person.



Triplet Loss:

The constraints for embeddings in the D-dimensional hyperspace is:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \qquad \forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \mathcal{T}$$

Here, x_i^a = Anchor image and f (x_i^a)= embeddings of the Anchor image

x_i^p = Positive image and f (x_i^p)= embeddings of the Positive image

x_i^n = Negative image and f (x_i^n)= embeddings of the Negative image

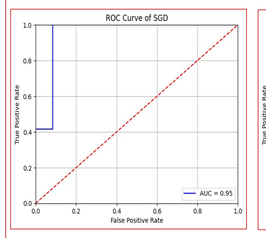
The loss to be minimized is L = $\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]$

Experiments Performed and the selected Hyperparameters

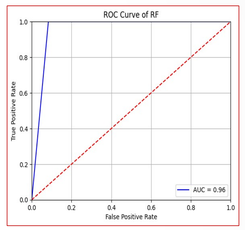
Architectures Used	Dataset	Best Learning Rate	Number of Epochs	Batch SIze	Accuracy
XceptionNet(Frames only)	CELEB DF	0.05	30	64	50%
LSTM+Frames	CELEB DF	0.01	30	64	52%
XceptionNet(30k frames)	CELEB DF+CELEB DF v2	0.05	30	64	70%
Facenet+Metric Learning	CELEB DF+CELEB DF v2	0.0003	50	32	87.5%

Results

Scatter plots comparison of data before and after training the model with Triplet Network



ROC curve of the SGD Classifier depicting False vs True positive rate



ROC curve of the RF Classifier depicting False vs True positive rate

- Our classifiers were able to classify real and fake faces with a high degree of accuracy. We achieved a testing accuracy of 87% on the dataset we used.
- We gave videos for testing where for each video we extracted frames and made predictions on the features of the 25 frames and then took the mean of the predictions to come up with a final classification output.

Measures	Triplets+ SGD	Triplets+ RF
AUC Score	0.95138	0.95833
Accuracy	0.875	0.875
Precision	0.8	0.8
Recall	1.0	1.0
F1 Score	0.888889	0.888889

Conclusion

- We were able to extract faces from frames using the Triplet network and classify real and fake faces in videos using the Facenet architecture.
- We have addressed the issue of spatio-temporal learning and less accuracy in case of convolution based xception network.Our model is innovative and uses metric learning to learn crucial features through the triplet network architecture.

Future Direction

- We have demonstrated performance close to the one shown in our base literature and we aspire to improve our model further to outdo the current achievements.
- The major limitation of the approaches is their generalizability across different datasets.
- In future, our aim is to use unsupervised domain adaptation to adapt the feature space from source dataset to target dataset, to make our model robust and label independent.

References

[1].FaceForensics++: Learning to detect manipulated facial images https://arxiv.org/abs/1901.08971#:~:text=https%3A//doi.org/10.48550/arXiv.1901.08971

[2] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and S. Li. 2021. Face Forgery Detection by 3D Decomposition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2928–2938

[3] Detecting deepfakes with metric learning https://arxiv.org/abs/2003.08645#:~:text=https%3A//doi.org/10.48550/arXiv.2003.08645

[4] Shreyan Ganguly, Aditya Ganguly, Sk Mohiuddin, Samir Malakar, Ram Sarkar/ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection.

[5] Xception: Deep Learning with Depthwise Separable Convolutions https://doi.org/10.48550/arXiv.1610.02357

[6] FaceNet: A Unified Embedding for Face Recognition and Clustering https://arxiv.org/abs/1503.03832