Methods for Detecting Deepfake Images

•••

MIT2019014 | Siddhanta Biswas

Under the guidance of
Dr. Vrijendra Singh, Associate Professor, HOD(IT)
IIIT Allahabad

Problem Statement

- Deepfake is a deep learning based technique which aims to replace the face of a targeted person by the face of someone else in digital media.
- Deepfake detection techniques in recent years have improved largely while still having relatively lower accuracy for low quality images.
- In this work, I am aiming to significantly increase the accuracy of manipulation detection in low quality images.

Present Work: Datasets Used

- Deepfake Database dataset (by the authors of the MesoNet paper)
- All videos are compressed using the H.264 codec but with different compression levels
- 80% of the Training set was used for training
- 20% of the Training set was used for model validation

Set	Fake Class	Real Class
Deepfake training	5111	7250
Deepfake testing	2889	4259

Table 1: Cardinality of each class in the Deepfake Database dataset

Present Work: Classification Setup

- Implemented the Meso4 architecture as shown in MesoNet paper in an effort to reproduce the results obtained by the authors, thereby having a baseline reading
- Input Shape = 200 x 200 x 3
- Testing set was divided into 80:20 ratio for training and model validation purposes simultaneously.
- Epochs = 200
- Batch Size = 75
- Learning Rate = 0.001 (fixed)
- Optimizer: ADAM with default parameters (β1 = 0.9 and β2 = 0.999)

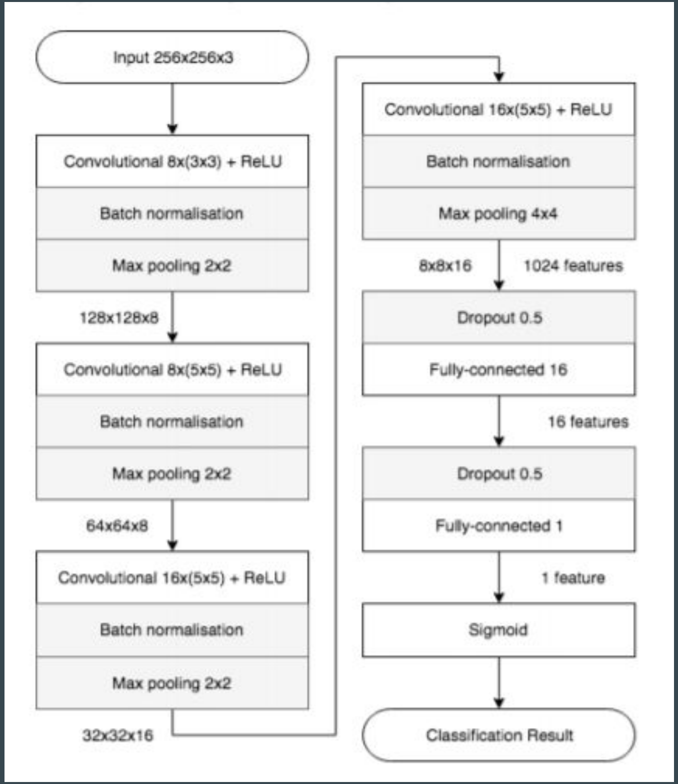

Figure 1: The network architecture of Meso4

Present Work: Results

- Achieved successful detection of deepfake images from real images, albeit with a lower accuracy than the authors
- Positive is Fake class
- Negative is Real class

Fake Images	96%
Real Images	84%
Total Accuracy	89%

Table 2: Classification scores of my Meso4 implementation on Deepfake dataset

True Positive	2734
False Positive	111
False Negative	664
True Negative	3595

Table 3: Confusion Matrix showing classification scores of Meso4 implementation on deepfake dataset.

Future Work: Plan of Action

- Improve present detection rates, especially for Real Images which are being detected with significantly lower accuracy.
- Testing the model on a different dataset with higher compression levels.
- Adding a bunch of Conv Layers in the 3rd and 4th blocks of the current architecture before Pooling, as shown in figure (similar to what VGG-16 does)
- If the number of trainable params increase substantially, resulting in huge training times, 1x1 conv layers may be used to decrease the number of feature maps while still retaining their salient features

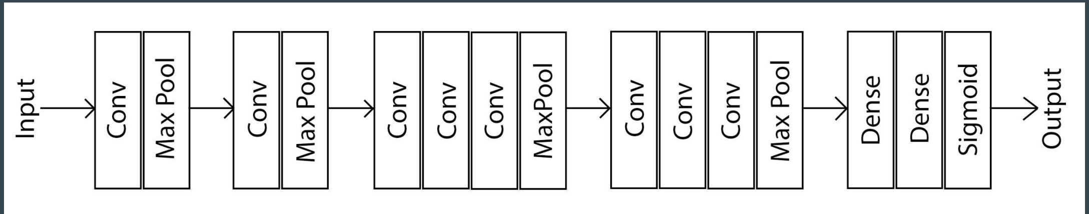

Figure 2: Architecture of my proposed model for deepfake classification, based on the Meso4 network, with several convolutional blocks added (inspired by VGG-16)

Thank You