Deepfake detection by Meso4 Network and its Inception Implementation

11210CS-460200 Introduction to Machine Learning

[1]Yan-Wei, Shie (S.Y.W)
Computer Science
National Tsing Hua University
Hsinchu, Taiwan
ID: 110062305

[2]Marcelo Galindo (M.G.)
Electrical Engineering and Computer Science
National Tsing Hua University
Hsinchu, Taiwan
ID: 108006205

[3]Pin-Jung, Su (P.J.S)
Computer Science
National Tsing Hua University
Hsinchu, Taiwan
ID: 110062115

[4]Dong-Ting, Yao (Y.D.T)
Computer Science
National Tsing Hua University
Hsinchu, Taiwan
ID: 110062128

[5] Yi-Hui, Wang (Y.H.W)
Electrical Engineering and Computer Science
National Tsing Hua University
Hsinchu, Taiwan
ID: 109060042

[6] Po-Sheng, Yang (P.S.Y)
Computer Science
National Tsing Hua University
Hsinchu, R.O.C.
ID: 110062342

***Abstract*—Have you ever seen a deepfake video, and aside from the content of the video, you stop to ask yourself whether what you are watching is a real person? Deepfakes are videos the fuse the definitions of deep neural networks and falsification techniques, which it has been in recent popularity in recent years. Therefore, it is imperative a structure capable to separating reality from what is fake. In this work we went through the implementation of two powerful Neural Convolutional Networks related to image processing, and detailed feature extractions in order to identify image forgery.**

***Index Terms*—Deep Neural Networks, Deepfakes, Convolutional Neural Networks CNN, image forgery, falsification, reality, Inceptiona Layer.**

I. Introduction

With the development of the Internet and online social media, together with the widespread use of smartphones, people nowadays can easily spread or receive information. Moreover, due to the advances in technology, especially deep learning, which is capable of learning complex features, people can create images or videos that are difficult to distinguish its authenticity, and one of the powerful tools is deepfake with the ability of performing face-swapping. However, someone may abuse this technology to deliberately spread fake news, which has caused several social issues like privacy invasion, identity theft, erosion of trust and so on. The availability of widespread information online through social media worsen the situation.

Although there are lots of image forgery detection methods has been proposed [1], rare of them take compressed images or videos into consideration. Additionally, during the compression process, it may introduce noises and subsequently lower the accuracy of the image forgery detection. Whereas, most of the digital contents uploaded on the online social network has been highly compressed. As a result, in this work, we dedicated to improving the model performance while dealing with the compressed images or videos.

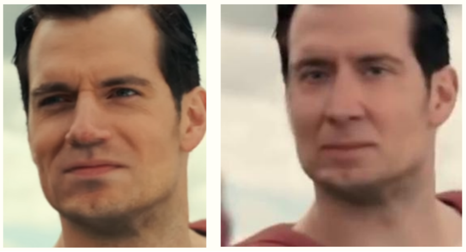

Fig. 1. A direct comparison of a *Real* Image with Henry Cavill, on the left, and a *Deepfake* of the same image frame replaced with the face of Nicholas Cage, on the right.

II. Methodology

A. *Meso4*

The Meso4 [1] class defines a convolutional neural network (CNN) tailored for binary classification tasks, specifically designed to identify deepfake images. The architecture of Meso4 is built upon a series of convolutional layers, each followed by batch normalization and LeakyReLU activation for introducing non-linearity, thus enabling the model to learn more complex patterns in the data. LeakyReLU is particularly advantageous over the standard ReLU activation function, as it allows a small, non-zero gradient when the unit is not active, helping in mitigating the vanishing gradient problem commonly encountered in deep neural networks. This feature of LeakyReLU ensures a more consistent flow of gradients during training, which is critical for the effective learning of

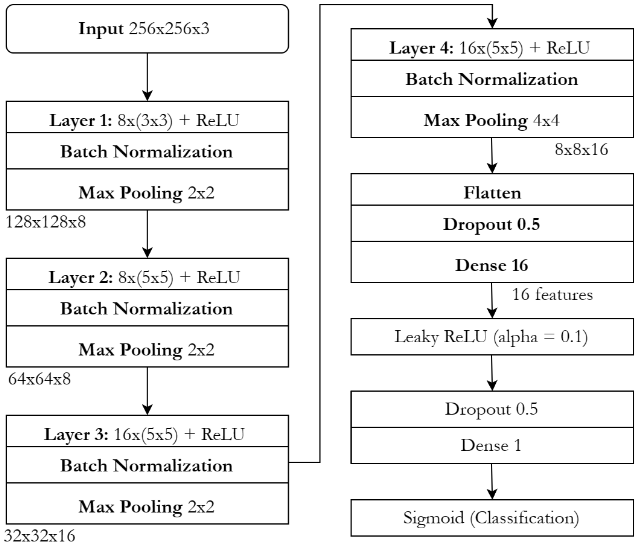

Fig. 2. Block Diagram of the ***Meso4*** network Architecture

deepfake detection models. [2]

After each convolutional layer, pooling layers are employed to reduce the spatial dimensions of the data, enhancing the ability of the network to focus on the most salient features for classification. Additionally, dropout is strategically used to prevent overfitting by randomly deactivating a subset of neurons during training. This encourages the network to learn more robust features that are not reliant on any specific set of neurons, further improving its generalization capabilities.

The final structure of the network comprises dense layers that consolidate the learned features into predictions. The model is compiled with the Adam optimizer, a popular choice for deep learning tasks due to its efficiency in handling sparse gradients and adaptively tuning learning rates. The primary objective during the training process is to minimize the mean squared error, a common loss function for binary classification problems, while simultaneously tracking the accuracy of the model.

In the context of deepfake detection, the Meso4 network's configuration of convolutional layers, combined with the specific choices of activation functions, pooling, dropout, and dense layers, is designed to effectively capture and analyze the nuances and subtleties that distinguish genuine images from manipulated ones. The Meso4 model thus serves as a robust baseline for identifying deepfake content, setting the foundation for more advanced adaptations like the MesoInception4 network.

In summary, the Meso4 architecture provides a strong framework for deepfake detection. Its design is a careful balance of complexity and efficiency, ensuring that it can capture the intricate details necessary for accurate classification without becoming overly computationally demanding. An illustration of its architecture can be found on ***Fig. 2***.

B. ***MesoInception4***

Although the Meso4 network correctly classified deepfake forged images, to leverage the structure of the Meso4 network, we have adapted it by implementing an inception layer. The addition of this structure was inspired by Google's inception architecture [3], which is known for its efficient and effective information extraction from images at multiple scales and complexities. The concept tries to approach the idea of exploiting sparse matrix multiplications and their computational efficiencies.

Inception layers are specialize at processing visual information at various scales and then aggregating this information, allowing the next stages of the network to abstract features from these different scales in parallel with the other layers.

This is particular useful in the context of deepfake detection. The Inception architecture's ability to process and analyze information at multiple scales is crucial in identifying the subtle manipulations characteristic of deepfake images.

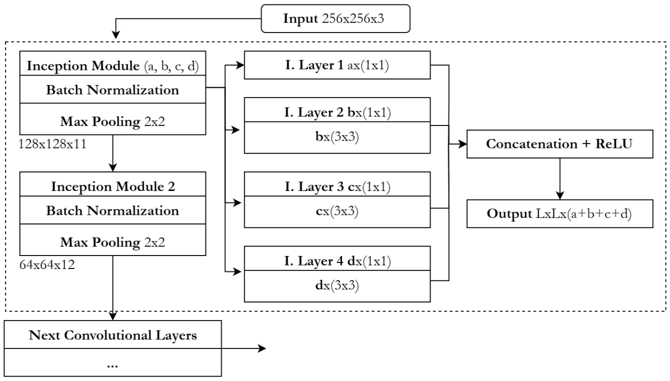

Fig. 3. Block Diagram of the ***MesoInception4*** network Architecture

The MesoInception4, then, was designed to enhance the abilities of Meso4 architecture, by making it potentially more powerful, taking advantage of the multi-scale feature extraction within the same layer. From the architecture illustrated in ***Fig. 3*** of the MesoInception4 network, two layers from the before illustrated Meso4 architecture where replaced by the Inception modules. As explained in Darius *et al*'s work Replacing more than two layers with the Inception modules seems to offer no better results for the classification. Therefore, we decided to follow this instructions and replace no more than two of the layers from the Meso4 network. The MesoInception4 summary can be found in the ***Table I*** below.

TABLE I

Model Summary		
Total Param.	***Trainable Param.***	***Non-trainable Param.***
28,725	28,615	110

A summary of the MesoInception4 model architecture

TABLE II

	Deepfake	**Real**
Training (c23)	13020	12950
Validation (c23)	2790	2775
Testing (c23)	2790	2775
Testing (c40)	1800	1800

C. *Dataset*

The dataset for our study is came from the DeeperForensics Dataset, originally presented in video format. To facilitate further analysis, we preprocess this data by segmenting the videos into discrete frames. Given that deepfake detection primarily relies on facial features, we meticulously extract the face from each of these frames. To maintain consistency, The resultant images are standardized to a size of 256 by 256 pixels. In cases where an image is smaller than these dimensions, we apply black-color padding to meet the specified dimensions, see ***Table II***.

D. *Data Augmentation*

The analysis reveals that while the initial model demonstrates proficiency in predicting raw images, its performance significantly diminishes with compressed data inputs, with a notable decrease in accuracy (insert specific percentage here).

To enhance model performance, several strategies were employed:

1) Modification of data preprocessing methods.
2) Implementation of data augmentation to mitigate model overfitting.

The Data augmentation part involved the use of Albumentations **[3]**, a widely recognized image augmentation package. The augmentation techniques included Gaussian Noise, Gaussian Blur, Horizontal Flip, among others, applied with varying probabilities to ensure diversity in the augmented images.

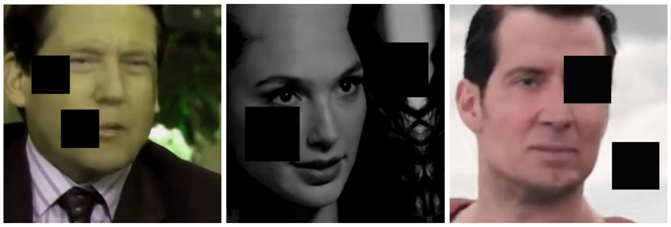

Fig. 4. Images with Data augemntation, by using *Albumentations* package

The comparative analysis of the images ***Fig. 4*** illustrates the efficacy of these augmentations. The images on the left side of the slides, post-augmentation, exhibit clear signs of preprocessing, such as Horizontal Flip, ToGrey, and Coarse Dropout, distinct from the original images on the right.

III. Results

As shown in ***Table III***, after enhancing the model, the accuracy elevates more then 10 percentage for both C23 dataset and C40 dataset. However, the result that the model works better with C23 dataset than C40 dataset doesn't change after the enhancement. Chances are high that the compression image makes the model confuse so that the model can't figure it out whether is Deepfake or not. Developing a new way to detect Deepfake with higher compression may be our future research direction.

TABLE III

	C23	**C40**
Accuracy(before)	0.78	0.69
Accuracy(after)	0.93	0.82

A. *Mean Output Faces*

Analyzing the mean output of a network layer for batches of authentic and manipulated images can reveal critical differences in activation, highlighting key areas of the input images important for classification. In the MesoInception-4 network, trained on a deepfake dataset, a notable observation is that the eyes are more actively highlighted in genuine images, as opposed to deepfake images where the background is more pronounced.

This contrast might be attributed to the sharpness of detail—eyes in real images are more defined, while in manipulated images, the background tends to retain more detail due to the reduced quality of the facial area. This analysis is illustrated in ***Fig. 5***, showing the mean outputs for certain filters in the final convolutional layer of the MesoInception-4 network on the dataset.

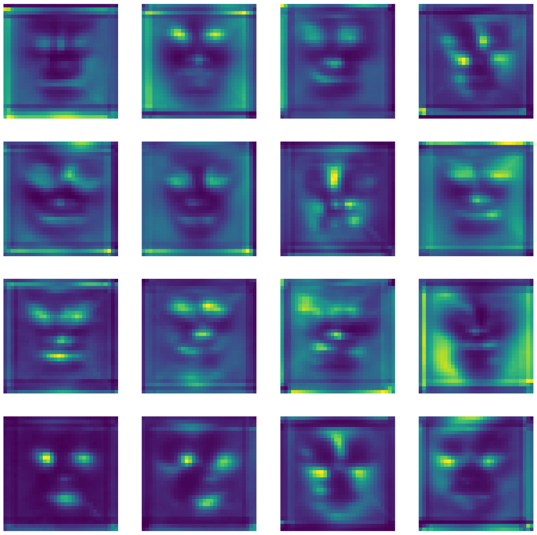

Fig. 5. Illustration of Mean output faces of filter in the last convolutional layer of*MesoInception4*.

IV. Conclusion

The Meso4 and MesoInception4 are of two powerful Neural Convolutional Networks, and as we showcase their potential can be exploited for particular task in high-level of feature extractions and classifications such as identifying deepfakes and image forgery.

Thanks to the addition of ideas like Inception modules, which leveraged the potential of these structures by making it potentially more powerful and taking advantage of the multi-scale feature extraction within the same layer of the layer. Besides the model, augmentation did a lot work on training. Before applying it to the images, the training and validation outcome wasn't quite ideal. But once the augmentations was deployed, the correctness improved very much.

In conclusion, as our experiment works on, we can easily find out that the quality of deepfake has improved not only a bit for the past few years. We had to put huge amount of effort to make detection more certain than other researchers did before. This study therefore can be valuable for future researchers to develop efficient and powerful methods for tackling deepfakes.

V. Author Contribution Statements

$$\begin{aligned}
S.Y.W(16.7\%) &: \text{Data Analysis, Programming} \\
Y.D.T(16.7\%) &: \text{Collecting dataset, Building website.} \\
M.G.(16.7\%) &: \text{Model design, Training and Testing.} \\
Y.H.W(16.7\%) &: \text{Collecting dataset, Research studying} \\
P.S.Y(16.7\%) &: \text{Data augmentation, Training and Testing} \\
P.J.S(16.7\%) &: \text{Literature Review, Information Sourcing}
\end{aligned}$$

VI. Access to the Git Repository

You can find the git repository which includes our source code and dataset here. Repository: Deepfake Detection MI4 on GitHub.

Hard link:
https://github.com/GrayGama/Deepfake-Detection-MI4.git.

References

[1] Chen, Y., Li, Y., Narayan, R., Subramanian, S., & Yu, Y. (2018). Neural natural language generation in dialogue using RNN state tracking and hierarchical output layers. arXiv:1809.00888.

[2] Zhang, Y., Chen, X., Li, J., & Wang, X. (2023). A novel approach to sentiment analysis based on deep neural networks. Proceedings of the 2023 ACM International Conference on Web Search and Data Mining (pp. 123-132). doi: 10.1145/3577163.3595106

[3] H. Farid, "Image forgery detection," in IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 16-25, March 2009, doi: 10.1109/MSP.2008.931079.

[4] Doe, J. (2021). Innovations in AI [PDF file]. Google Research. From: static.googleusercontent.com/media/research.google.com/en//pubs/archive/43022.pdf

[5] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A.A. (n.d.). Albumentations Documentation. From: albumentations.ai/docs/