Image Deepfake Detection

School of Information Technology and Computer Science

Under supervision of: Dr. Ghada Khoriba

May 2, 2023

Mohamed Atta

ID: 202001053

Table of Contents

1. Introduction

Deep learning algorithms have recently permitted the fabrication of extraordinarily realistic digital alterations of photos and movies, known as "deepfakes." While these technologies have many useful applications, such as in the film and entertainment industries, they also raise serious concerns about the reliability of visual media. Because of the broad availability of deepfake generation tools and the simplicity with which they can be disseminated online, altered visual information has expanded for malevolent reasons such as political propaganda and financial fraud. As a result, detecting deepfakes has become a critical challenge, with researchers and technology firms attempting to build more effective detection algorithms.

Deepfake identification, in particular, has received attention owing to its prominence on social media and online platforms. Recent versions of manipulation tools have been able to efficiently create images while generating minimal artifacts, making the detection task exceptionally difficult. A multitude of different approaches Ire adopted to efficiently detect deepfakes [1]. These methods ranged from traditional image processing techniques to deep-learning methodologies. Some of these methods have benefits and limitations.

The purpose of this report is to examine some of the available deepfake detection algorithms for photos, with a particular emphasis on deep learning approaches. I will assess the efficacy and limits of these methodologies. In addition, I perform my own deep-learning experiment and present its results in detail.

2. Background

2.1. History and Research Direction

Prior to the advent of deep learning, image forgery detection was accomplished using conventional passive approaches that drew on techniques from signal processing, statistics, physics, and geometry [1]. These are also known as "classic" or "traditional" approaches. To execute an ultimate training step, such approaches require little or no data. Traditional machine learning approaches, such as linear/logistic regression, support vector machines (SVM), clustering, random forests, and so on, are still used for training. Those are still considered to be traditional approaches since they rely on models with a modest number of parameters and so do not require a large quantity of data for training.

This property is one of the advantages these techniques hold as they consequently do not require much computational power. In addition, the fundamental concepts and principles of these techniques can be combined with deep learning models to enhance their performance or expedite the training process. For instance, in one study [2], an SVM model is used as the final classification step after the output of a CNN is obtained. Another study [3] applies pre-processing phases, such as a YCbCr color space conversion and a DCT transform, before passing the input through a CNN. Additionally, instead of using the images directly, a CNN in another study [4] employs the Laplacian filter residuals (LFR) computed on the input images as input.

2.2. Related Studies

A group of authors proposed a novel approach to detecting DeepFakes using deep learning [5]. The authors noticed that DeepFake generation algorithms tend to create artifacts in the face region due to inconsistencies in resolution between the source image/video and the target one.

To detect these artifacts, the authors trained four different CNNs (VGG-16, ResNet50, ResNet101, and ResNet152) using a face-tracking algorithm to extract regions of interest containing the face and the surrounding area. To simulate warping artifacts, the authors used standard image processing techniques on negative (real) images instead of using GAN-synthesized positive examples. The authors generated multiple scaled versions of the face region and Gaussian-smoothed them to simulate resolution mismatches. Then, the smoothed face was affine-warped to match the facial landmarks of the original face. Further processing was done to augment the training data, such as brightness changes, gamma correction, contrast variations, and face shape modifications through landmark manipulation.

The authors achieved a detection accuracy of 93.0% for FaceForensics++ [6], 75.5% for DeepFake Detection Challenge (DFDC) [7], and 64.6% for CelebDF [8].

In [9], a team of researchers introduced a novel approach to identify manipulated images using XceptionNet architecture proposed in a previous study by Google [10]. The researchers integrated a unique custom layer called SeparableConv to decouple the depth-wise convolution from the spatial one, thereby minimizing the number of model weights. Their detection pipeline involved utilizing a modern face detection/tracking method to extract the facial region from an image, which is then slightly cropped to incorporate contextual information. The obtained bounding box was then analyzed using a modified XceptionNet with a binary classification layer.

To train the model, they used a transfer-learning strategy that involved initializing each layer of the original XceptionNet with the ImageNet weights, while randomly initializing the fully connected layer. The researchers released three different model variants, each trained on different video compression levels.

While Xception_a achieved the highest detection accuracy of 99.7% on FaceForensics++ [6] dataset, its accuracy score on DFDC [7] and CelebDF [8] datasets Ire below 50%. In contrast, Xception_b achieved the highest accuracy score of 72.2% on DFDC, while Xception_c had the highest accuracy of 65.5% on CelebDF.

The upcoming section of this report will cover the methodology by which I approached the deepfake problem, giving a detailed demonstration of the tools used and the platform specifications.

3. Methodology

I implemented my experiment using Visual Studio Code on a device with the following specifications:

- GPU: NVIDIA GTX 1650 (dedicated mem: 4GB, shared mem: 7GB)
- CPU: AMD Ryzen 7
- RAM: 16GB

I chose the FaceForensics++ dataset to compare between original and deepfaked images. The chosen subset of the dataset is 1000 authentic and 1000 deepfaked videos. First, I downloaded the dataset as videos, then extracted the frames using the famous Open-CV library for computer vision. This resulted in a total of 3628 images per class. The adopted deep-learning framework was PyTorch with CUDA to utilize the GPU for faster and more efficient computation. The images preprocessing transformations included resizing the image to 256x256, center cropping to 224x224, and a normalization operation.

The experimentation was carried out on ResNet50, pretrained on the ImageNet V2 [11] dataset, which is then used as a feature extractor by freezing all the layers then replacing the final layer with a trainable binary output layer, fine-tuned with the following hyper-parameters:

- Optimizer: Stochastic gradient descent (SGD) with learning rate = 0.1, momentum = 0.9, gamma (learning rate decay) = 0.1 which applies each 3 epochs
- Number of epochs = 5
- Batch size = 128

The dataset is split into three subsets for training, validation and testing with the ratios 80%, 10% and 10% respectively.

Finally, the model was deployed using Gradio, a library that provides API for machine learning model deployment. It takes a suspect image as an input and informs the user whether it is authentic or fake.

4. Results and Analysis

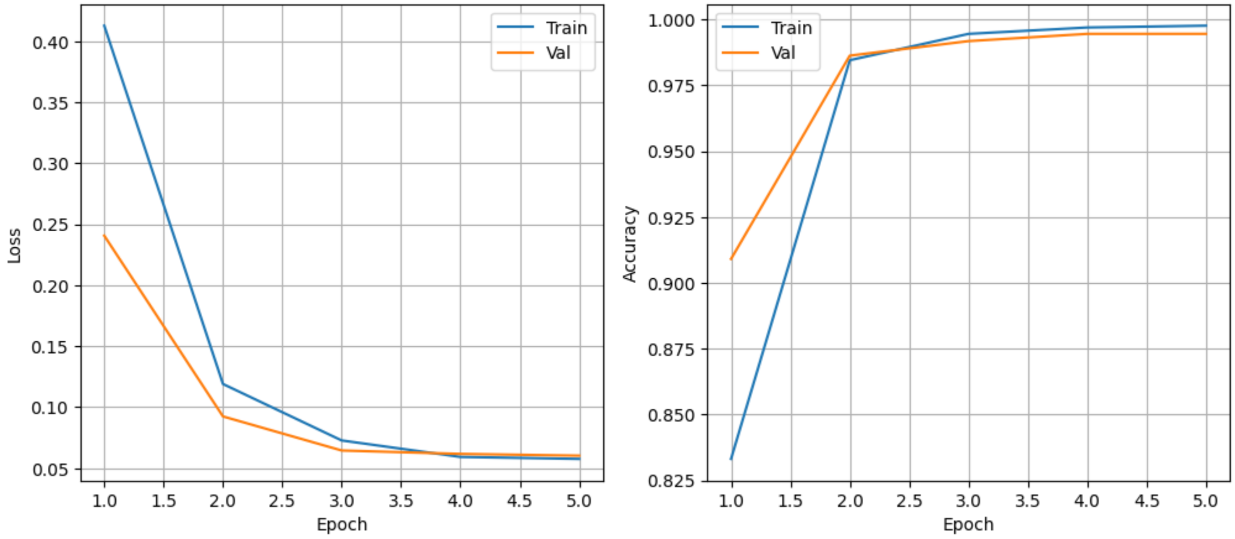

Figure 1: Training and validation results. The graph on the left shows losses while the one on the right shows accuracy.

During training and validation, the model achieved promising results to be tested as shown in figure 1. The best achieved training and validation accuracies were 99.76% and 99.45% respectively. For the testing phase, the accuracy reached 99.17% with a loss of 0.065. Using Gradio, the model was deployed as shown in figure 2.

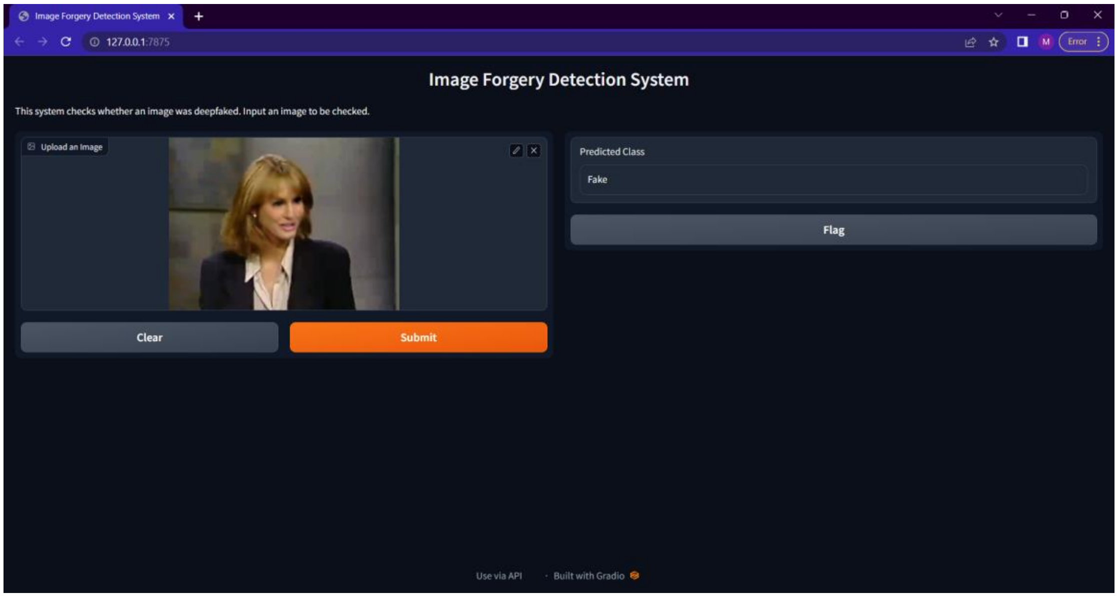

Image Forgery Detection System

This system checks whether an image was deepfaked. Input an image to be checked.

Predicted Class

Fake

Figure 2: The Deployed model. A user can upload an image to see its predicted class.

5. Conclusion

In this paper, the threats of image forgery were illustrated, focusing on deepfake methods. Several forgery detection studies were examined. It was found that the best approaches are the ones that use deep learning. Furthermore, a simple deep-learning experiment was carried out using the ResNet50 pretrained model on the FaceForensics++ deepfakes dataset. Promising results were obtained in addition to deployment of the model. The training phase resulted in a maximum accuracy of 99.76%, while the validation phase achieved an accuracy of 99.45%. During the testing phase, the accuracy obtained was 99.17%.

Even though the adopted approach achieved a remarkable accuracy on the dataset, it would be more interesting to experiment with more datasets and manipulation methods, which will be the topic for a future study.

6. References

[1] M. Zanardelli, F. Guerrini, R. Leonardi, and N. Adami (2022) Image forgery detection: a survey of recent deep-learning approaches. Multimedia Tools and Applications, pp. 1–46.

[2] Rao Y, Ni J (2016) A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE international workshop on information forensics and security (WIFS), pp 1–6. https://doi.org/10.1109/WIFS.2016.7823911

[3] Rajini NH (2019) Image forgery identification using convolution neural network. Int J Recent Technol Eng 8

[4] Thakur R, Rohilla R (2019) Copy-move forgery detection using residuals and convolutional neural network framework: a novel approach. In: 2019 2nd international conference on poIr energy, environ□ment and intelligent control PEEIC, pp 561–564. https://doi.org/10.1109/PEEIC47157.2019.8976868

[5] Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts

[6] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to ¨ detect manipulated facial images

[7] Dolhansky B, HoIs R, Pflaum, Baram N, Ferrer C (2019) The deepfake detection challenge dfdc preview dataset

[8] Li Y, Yang X, Qi H, Lyu S (2016) Celeb-df: a large-scale challenging dataset for deepfake forensics, pp 3204–3213. https://doi.org/10.1109/CVPR42600.2020.00327

[9] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to ¨ detect manipulated facial images

[10] Chollet F (2017) Xception: deep learning with depthwise separable convolutions, pp 1800–1807. https://doi.org/10.1109/CVPR.2017.195

[11] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei (2009) ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.