National College of Ireland

Data Mining and Machine Learning 2
Group B

MSc in Data Analytics
National College of Ireland

Name: Akshen Doke
ID: x18191592

Repeat Terminal Assignment Based Assessment

Lecturer: Prof. Anu Sahni

Case Study: Detection of DeepFake Images

Assumptions and Literature Review

For this case study we need to make use of a dataset on which we can make our report. The most easily available and one of the most popular dataset is that of celebrity images which we can make use of in our report by downloading it from here https://megapixels.cc/msceleb/ . We make use of Python as our programming language to code and its packages like keras, matplotlib and NumPy for further processing, visualization and model creation.

When the DeepFake false face creation technology was introduced, a lot of work was done on both sides i.e. to improve the quality of disguising fake face as real and it's detection. Most papers published in this category were related to fake face manipulation in videos but as they work on each frame per second, we can study the same methods applied on it for detection and use it to detect single images as well. For most researchers who've worked on fake face image detection, Convolutional Neural Network (CNN hereafter) based model has always been a preferred choice. Four models different based on CNN are majorly trained from the beginning namely VGG16, ResNet152, ResNet101 and ResNet50. (Tolosana, et al., 2020) in their research made use of Generative Adversarial Networks(GANs hereafter) architecture like StyleGAN and ProGAN to generate fake images of four categories like identity swap, expression swap, entire face synthesis and attribute manipulation. Each technique as name says implied a way of generating fake images using real image as a source. The most difficult amongst these to spot is Entire Face Synthesis where a face of person is used to generate an entirely random person which most of the times can be totally indistinguishable from the source image. (Rossler, et al., 2019) discussed about how synthetic image generation and manipulation is getting better and ways to tackle it. Also, they spoke about Deepfake and its implementation in particular where a target region on face is replaced by another face that the manipulator intended too. They made use of faceswap and implementation of deepfakes available on GitHub which we can also make use of in the project for creating fake image. Models like Alex Net, VGG19, ResNet50 Xception and Inceptionv3 were used to detect deepfakes based images (Khodabakhsh, et al., 2018) It was found that CNN based on AlexNet performed exceptional well with an accuracy rate of 98%.

Exploratory Data Analytics / Data Cleaning

This is the initial step in any data science project where we go through our available dataset to check for any missing values, inconsistency etc. First, we load the dataset from the source and check if any of the datapoint is corrupt in anyway like format and image size also we have to remove redundant images so an easy way for this is to give each image file a checksum and then keep only unique ones, another issue with the images which cannot be solved by assigning a checksum is that even if a single pixel is different than the checksum value would be different in this case we have to segregate the image based on who's it is and then in the next step of feature extraction, we would get the desired result. To maintain consistency throughout we will resize all the data to a single size of 480 X 320 and keep it in a single format i.e. jpeg. This way we make sure that all the images given as input to the model to train on are similar and through this consistency, even our model will be trained properly.

Feature Selection

First and for most important step is to select the face region from the image (Akhtar & Dasgupta, 2019) as face manipulation is the most crucial factor for identity forgery. Then for any face be it male or female, young or old its primary features are eyes, nose, lips and ears. In the initial part of our model, we will include layers which will help us identify these set of features by drawing a boundary over it. This model is pretrained to detect the mentioned features with better accuracy. In our CNN based model, we can make use of a machine learning technique called transfer learning where in a pretrained model is used to retrain and adjust it weights as per our dataset, this will help us achieve a better accuracy and save us time instead of doing everything from scratch.

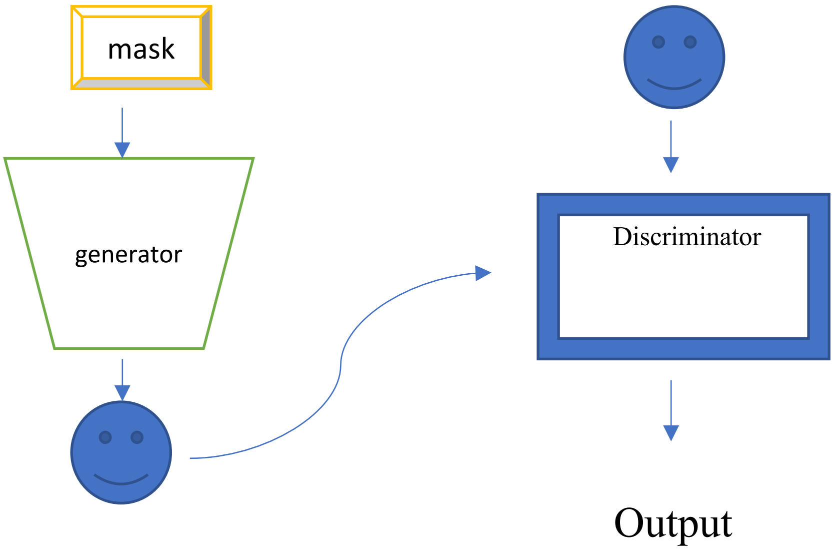
Feature Engineering / Feature Extraction

Our model will be trained on real images and then using GANs we can try to create a fake image. When the GAN based model will create fake images based on the target image supplied this will be supplied to the CNN which will be labelled as fake images. Since this being a computer vision problem, features will be specific structures in the image like nose, facial edges, eyes etc. We will make use of a pretrained model present in keras library called VGG16. This model has 16 layers, which was trained on more than a million images with 1000 classes.

Choice of modelling technique

At first, we train a GAN on our dataset to create fake images or we can also make use of openly available deepfakes implementation called faceswap (Rossler, et al., 2019). In the GAN based model for creations of fake images, we first mask the critical face region with a masking image then we train the generator of the model to generate the fake image, depending on the quality and similarity of the generated image with that of the original/target image, the discriminator calculated the loss of the difference and gives feedback to the generator to improve its performance. Ones the generator gives a reasonable, acceptable output then we can process to train our CNN based detection model to detect fake from genuine based on the output from our GAN model.

The CNN model will have dedicated layers as mentioned in the above sections where required features will first be selected then extracted and examined to see if any manipulations have been made, in case of a totally random image, we can check for pixel density for blurriness, unnatural skin tone, teeth area and double chin. The model will be trained to classify based on if the image is fake or genuine which falls under the category of binary classification under supervised learning. The model will learn to identify the difference between the two types of images by updating its weights based on the features of both the input images.



Basic GAN model Layout

It is worth noting that we can use the discriminator itself as a separate model for detection of fake images as it is a collection of neural layers itself and does the job which we require, however for the sake of model with single image source as input, using similar principles as that of the discriminator we will build a CNN based detector model.

Steps for CNN based Model

1. Specify the batch size and number of epochs, generally a low batch gives proper accuracy the only drawback it has is that it takes time to be trained.
2. Make a layer for the data to go from 2D image into 1D image.
3. Normalize the pixel values.
4. Add the transfer learning model i.e. VGG16 in between, note that we've added the 2D to 1D in the first layer also the weights in the VGG16 would be frozen as we don't want to retrain them now, instead use it on our data.
5. Add a softmax layer at the end which would act as a classifier to set fake and real image aside.
6. Compile the model and save it in HDF5 format.

While the model is being trained, we can use the callback function available with the keras package to get a view on the internal statistics and states of the model during training. Also, there is another feature present in keras called Model Checkpoint which saves the model with its weights at a certain point in the training. It is a healthy practice to save the model weights only when an improvement is observed.

We are choosing an approach wherein we have to create an fake image first using the original image and later train a model and make it understand the difference between both, but if we follow an approach where in a model is trained to directly check an image whether it is fake on not based on the features, the probability of it making a low resolution picture as fake remains high.

Hyperparameter Optimization

This is a very important step when it comes to building an accurate model as proper hyperparameter tuning can highly increase the performance of the model and give the results we desire to achieve. (Julie, 2020) talks about using a package called ‘keras-tuner’ since we will be using keras as our transfer learning model provider, we would be using this package to fine tune our hyperparameters.

Model Evaluation

Initially we will have to split our dataset into two parts, one would be used for train the model and the other one would be used to test, assuming the split to be 70% for training and remaining 30% for testing. To evaluate the performance of our CNN based classification model, we will make use of confusion matrix where we can check how the model has classified true positives and true negatives values from the testing part of the dataset. In (Rossler, et al., 2019) it was observed that with lower quality images or videos it was easy to detect.

Scalability Issues

The Model will work fine on the dataset which we are using and other images of similar resolutions, but problems will arise if images with finer resolution appear also seeing the model, same model can’t be used in fake face detection in a video even though it will work on each frame. Also, as time passes by the efficiency of deepfake creation model will improve so this detection model won’t be efficient once that happens.

Ethical Implications

The main purpose of detecting deepfake is that it can be used for malicious purposes like political manipulation, evidence tampering, fake indecent video creation but if a genuine image is categorized as fake then this might lead to mistrust in technology or the company creating it. Also, clever hacker might just manipulate a bit of a genuine image so as to let detection algorithm classify it as fake and thus not let the user post his or her genuine image. Also (Akhtar & Dasgupta, 2019) spoke about how image recognition technology is used at border control in many countries and some even grant an e-pass online just based on merge image and a few other documents, if the image is tampered in between it may lead to a huge disaster if entry is granted to unlawful agents who pose a threat to the security of that country.

References

Akhtar, Z. & Dasgupta, D., 2019. A Comparative Evaluation of Local Feature Descriptors for DeepFakes Detection. *IEEE,* Volume 19, pp. 5-6.

Tolosana, R. et al., 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 22 June, pp. 1-18.
Rossler, A. et al., 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *IEEE,* pp. 1-14.

Khodabakhsh, A. et al., 2018. *Fake Face Detection Methods: Can They Be Generalized.* Gjovik, Research Gate.

Stewart, M., 2019. *Simple Guide to Hyperparameter Tuning in Neural Networks.* [Online]
Available at: https://towardsdatascience.com/simple-guide-to-hyperparameter-tuning-in-neural-networks-3fe03dad8594
[Accessed 1 August 2020].

Julie, 2020. *SICARA.* [Online]
Available at: https://www.sicara.ai/blog/hyperparameter-tuning-keras-tuner
[Accessed 1 August 2020].

Paper Review

DMGAN: Discriminative Metric-based Generative Adversarial Networks

Structure and Title

Yes, the title does describe the article as the paper talks about a novel way for Generative Adversarial Networks (GAN), it is very specific about its novelty which is "Discriminative Metrics".

Abstract

The abstract of the paper gives justice to the required criteria of briefly summarizing the content of the paper. Readers can get a gist of what the article is about only thing missing is future scope which if added could have covered the article completely.

Introduction

This part properly describes what the authors of the paper are trying to do initially by explaining what GAN is, where it is used and comparing the methods and approaches of design. However, it exceeds the notion of being expressed in one to two paragraphs long as the authors have stretched it up a bit.

Graphical abstracts and/or highlights

The authors have given a diagram of architecture for the proposed DMGAN which helps the reader visualize how the proposed GAN would have the data flowing through the pipeline.

Methodology

In this section, proposed methodology for the novel method with its improved method for strategy of weight adaption is explained. Both Models are explained with its respective pseudo code and equations. For the regular model, the authors have given a framework for the architecture of the model. Also, the authors are specific with how the model will be implemented, in the next model that is waDMGAN, a weight adaptive strategy is proposed instead of the fixed weight which was used in the initial one. Also, implementation of these models over datasets is explained under Experiments section were the models were trained on 3 datasets respectively, these datasets were collected for the following reasons, first is that they were all labeled which met their requirements, secondly most of the models based on GANs conduct their research using same databases. Exact hardware and software library specification is also given which will help in reproducibility of the experiment.

Results

No specific header is dedicated to result but we can derive it from the outcome of the experiments performed by the authors as a proof of their proposed model. Initially authors made use of human annotators as evaluators as quantitatively estimating models would have been quite a challenging task. But after realizing the issues related to human annotators, they substituted it with Inception Score (IS) and Frechet Inception Distance (FID) as both these approaches gave equivalent results which are correlated with human judgement and understanding. Also, the authors showed results of their experiments in a tabular format comparing performance of their each iterations with hyper-optimizing parameters as well.

Conclusion

Since the authors proposed a methodology and an enhancement to the same, the results were met however at the end they did mention its limitations. They failed to mention it with any relation to previous research work in this section. Authors didn't mention directly how their proposal will move the body of scientific knowledge ahead however they do have mentioned the gap of the methodology which can be filled in future.

Language

Apart from some minuet grammatical errors, the article was well written with concepts being explained as simple as possible although if the authors would have focused more on conciseness, it would have been easy for new researchers to get started with GANs and its successors.

Previous Research

The model proposed in the article is based on some previous work related to GANs which was mentioned in the introduction part and the authors have mentioned about it appropriately where-ever required. All references were done properly.