CSE161-Final-Project

Kimi Holsapple

April 2022

Abstract

False images have become a growing concern over the years. This paper will discuss a simple strategy to detect altered images of people's faces using Convoluted Neural Networks. Meso-4 was the model of choice and two visualizations were implemented in python in order to perceive the important features for the model. These include feature mapping and Grad-CAM which aim to expose the activation mappings of the model and the influence of areas for the task of classification respectively. With these results, it is concluded that a there are future applications with such technology to automatically parse through false images online allowing news agencies and social media platforms to automatically detect them.

1 Introduction

Fake and altered images are everywhere from social media filters to funny photo shopped meme posts. While ubiquitous, it doesn't necessarily mean its all good. The phrase has also been connected to fake news, advertisements, destroying reputation, and even cases of blackmail and revenge.

Although it may be easy for some to identify fake images, this is not the case for everyone. According to this study[3], those who do not have sufficient internet experience, photo shop ability, and infrequently used social media had the hardest time telling real from fake. As deep fakes become more advanced they will become harder and harder to detect and social media users will have to become more vigilant with what they believe. According to Pew Research [6], 72% of Americans engage in some form of social media, making the issue of false images a relevant issue for many.

While some may argue it's effects are limited to the online realm, image alteration is serious since it has the ability to defame and cause emotional distress to people especially when used for inappropriate purposes. Studies have also shown even if we know the images were fake, it can alter our memories, and even subconsciously influence decision making down the road. Not only this, but people rely heavily on the authenticity of photo-journalism with more than eight in ten Americans getting their news from digital devices [10]. Altering images and presenting them as truth is unethical in cases where people can be misled away from reality. In cases, were the public finds out, it degrades trust in society and we have seen a large dip in American trust in news media outlets with Although defamation and blackmail are serious, the minor alterations on social media showing perfect people have also been shown to cause negative emotions in some viewers as it creates unrealistic standards [6].

2 Background

There are a variety of methods to be able to create false images. The most common machine learning algorithm to produce deep fakes is GAN or Generative Adversarial Networks. GAN is a deep-learning model that works by having two adversarial algorithms, a generator and a discriminator, compete against one another in order to produce realistic fakes. The generator tries to approach real images as close as possible while the discriminator tries to label the generators images as fake. When the discriminator begins to fail the output image is expected to look convincing. The technique is robust and well known results are captured in the website thispersondoesnotexist.com, which creates images of non-existent people shown in figure 1. This is technique called face synthesis where typically a version of GAN called styleGAN uses a generator architecture to learn facial features such as noses, eyes, and lips and then pass them through convolution layers in order to create non-existent images of people.

Figure 1: All of these images are fake and were created on thispersondoesnotexist.com

Another method is faceswap and is the most popular deep fake type as seen in figure 2. This is the technique focused on being detected in this paper. This approach works by taking two auto encoders with a shared encoder to recreate training images of target and source faces. The "swap" occurs in the decoding stage as the decoders are switched for the target and source images. [4]

Figure 2: A faceswap between Paul Rudd (Real) and Jimmy Fallon (Fake) has become a popularly shared deepfake

Although these methods sound complex, it has never been easier to create deep fakes as there exists a plethora of deep fake creation tools. These include DeepFakeLab which is product that allows easy faceswaps. It is an open-source project that is responsible for an estimated 95% of the created deep fake images in circulation today. It's powerful, convincing, and allows user to de-age, move and sync lips, as well as change faces with little understanding about deep-learning required from the user [5].

Deepfake and faceswap creation works by using GAN as described above. It is easy to achieve since it requires only a small amount of face photos in order to produce the face-swap. Since, most videos will be compressed when they are shared on social media, the work of digital forensics becomes harder as traditional signal processing tools have difficulty analyzing the compressed deep fakes. Thus there is still real and relevant work needed to be able to handle detailed videos and sort through degraded image features due to compression.

3 Related Work

As already explained, the salience of deep fake detection necessitates their detection. Due to growing pressure large tech companies such as Facebook, Google, and Adobe have all released their own versions of fake image detection software recently however. This includes products like Google Jigsaw detect altered images in general while Adobe's version is specifically made at the time of this paper to detect photo shopped images created through splicing, cloning and object removal. The main focus for these applications is digital forensics to aid law enforcement.

Not only this but many researchers are interested on the topic with many papers detailing the topic to develop better methods in deep fake video detection. As it is a major topic in computer vision, there is interest in the application of neural networks to allow autonomous vehicles and robots the ability to detect various objects like roads and people.

This project aims to overlay the feature maps produced by CNN onto faked images to produce visualizations to help humans understand what makes the images fake.

In terms of visualization techniques, there exist multiple. The oldest are saliency maps which measures spatial attributes to explain classification decisions. Related and newer are CAM or class activation mappings which try to show the extent certain areas similarly explain features. While these two methods are essentially the same, CAM is computed in a variety of ways with the main way we explore being Grad-CAM which uses gradients from layers to calculate contribution. Feature maps are another technique that exists that is commonly used to display the activation of certain features for layers in the model.

4 Technical Detail

4.1 Data Collection

Two data sets were collected. The first involves a *DeepFake* and a *Real* data set provided by the Meso-4 authors and a youtube video explaining their model [1][8]. These images were taken from Deepfake videos, and processed to be 256x256 still images. Real images were taken from movie clips and videos and processed in the same manner.

Data was also collected from a Kaggle data set with over 2,041 entries at https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection. This will serve as the training data to allow the model to identify false images by classification. It is split into two groups: a altered image group and a real person image group. The altered group will be called the "Fake Image" group while the unaltered images will be called "Pristine". All the Fake Images in this data set fall into either easy, medium, or hard levels depending on the quality of alteration. All images in the dataset are jpeg files.

The dataset comes from deepfake stills and were processed by using OpenCV python package to detect faces. Faces were then cropped to size 256x256. Focus was placed on faces because the deep fakes used altered only faces. This was to reduce noise for the model and to allow it to narrow down feature detection to differences in real and fake frames.

For the model the images were left as compressed jpeg files in folder directories labeled either *Pristine* and *Fake*. The directories were read through using flow from directory in python. Although a NumPy array could be created to house all the images, resources available were limited, and it was thought impractical to load an entire NumPy array for every deep fake frame from disk. Compression from JPEG allows us to save space.

4.2 Model

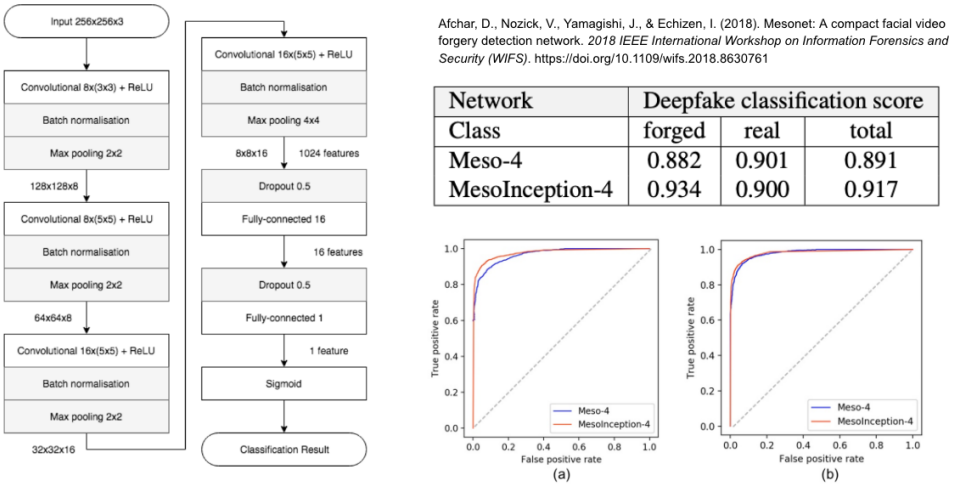
Convoluted Neural Networks were chosen for this task because they are among some of the most effective techniques to classify fake images. Furthermore, there already exists wide support for fake image detection tasks using Convoluted Neural Networks especially in the Python libray Keras. In this application, We will use the Meso-4 which was developed in 2018 and is perhaps one of the most well known DeepFake detection models produced today [1]. This model allows us to classify deep fake from their real counterparts and then extend this by taking the feature maps of the layers to produce interesting visualisations on deep fake image discrepancy. It was chosen because it was easy to use, has support in terms of tutorial information [8], and has good performance.

Convoluted Neural Network

Convoluted Neural Network in general are deep learning algorithms that are based off the linkages similar to the neural pathways. CNN for short, this is what the model will be referred to for the rest of the paper.

Meso-4

Meso-4 exploits mesoscopic features of images in order to detect if they are real or fake. Proposed by Afchar *et al*, Meso-4 is built by using four convolutions blocks followed by a fully connected hidden layer [1]. It uses Adam optimization for gradient descent and the overall architecture is displayed in Figure 3.



Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: A compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. https://doi.org/10.1109/wifs.2018.8630761

Network	Deepfake classification score		
Class	forged	real	total
Meso-4	0.882	0.901	0.891
MesoInception-4	0.934	0.900	0.917

Figure 3: From Afchar et al, this shows the model confusion matrix, ROC curve and structure. As is seen it preforms relatively well on the DeepFake and FaceSwap datasets.

This model was selected as it is optimized for use on images. It utilized gradient descent to identify most prominent features in the image. These are assigned to kernel weights, provided in the file called "MESO4_DF", that will be filtered through layers allowing for the task of classification.

Another filter layer can then be applied to identify particular areas of images classified. By identifying these areas it will be possible to highlight wrong parts and provide a visualization component. This can be done by overlaying the feature maps of CNN over the images. Accuracy of the model can also be visualized as statistical graphs and maps depending on the level of difficulty of each image.

The activation function used at the end of the model was Sigmoid. Sigmoid works by keeping input values between 0 and 1. In this application, Leaky ReLU will be the most relevant as this is the activation function that feeds the convolution layer inputs which our visualizations rely on.

4.3 Visualization

The visualizations were produced using a Python script called "fake_img_det.py". This script utilized TensorFlow/Keras and numpy for image processing as well as mathplotlib library to generate the visualizations.

Feature Maps In order to create a visualization of identified features, the activation output of each convolution layer was displayed using viridis color mapping. Another color map was created for the final feature activation. Feature mappings allow us to understand and visualize where activation occurs, what features are being recognized, and when certain features die out. As the model goes deeper there is resolution loss making resulting images more pixelated as demonstrated below.

The Figure 4 image was in the pristine data set folder to test the model to distinguish between real and fake images. Activation is color encoded with strong activation of the ReLU function corresponding to high yellow values. Values of no activation are encoded dark blue, and this comes in from the fact reLU maps input between 0 and above defined as below:

$$LReLU(x) = \left\{ \begin{array}{ll} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \leq 0 \end{array} \right\} \tag{1}$$

After undergoing reLU, the result was then normalized by subtracting by the mean value and dividing by the standard deviation. It was then mapped to the the rbg color scale and mathplotlib was used to create the final feature maps.

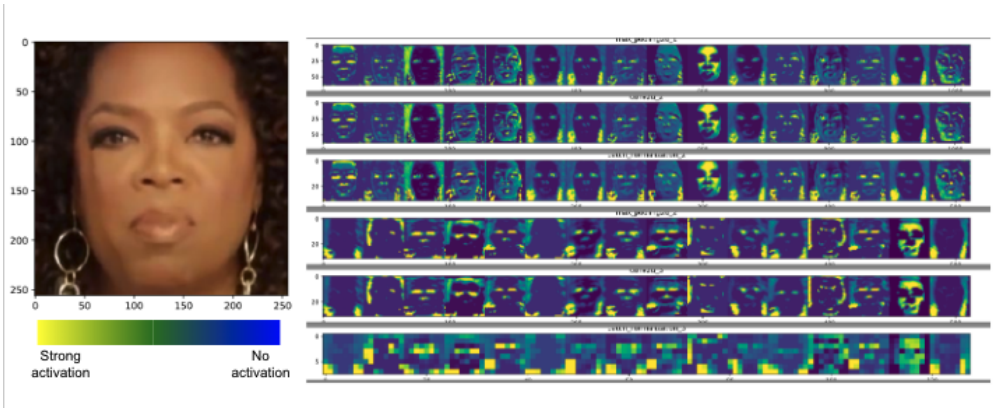


Figure 4: Produced visualization of feature map for real image showing the loss of resolution and the coloring legend.

Class Activation Maps These mappings allow us to see what regions of activation were used in classification. A color mapping is mapped over the image in order to help create a better understanding of what factors and areas aided in the understanding of the final output. More red colored hues indicate strong influence to the end classification and bluer areas show little regard to helping the end classification.

Grad-Cam was used for this application and stands for Gradient-weighted Class Activation Mapping. This works by taking the targets gradients in order to create a coarse location mapping that highlights important regions. Since Meso-4 has 4 convolution layers, the fourth layer was taken in order to produce these visualizations. Results vary by task and model, but still remain useful and effective methods of understanding model decisions. Grad-Cam is unsuitable for instances where an image contains multiple occurrences of the same class. That is why images of single people are focused on instead of groups of people. Furthermore, it is possible where localisation may not correspond to the full object. This mean highlights will occur on bits and pieces of the object. However, if this occurred in the close up images, it would not be an issue. This is because we already know that images are classified as true or false, and if the entire object was highlighted then the entire image would end up red and this would not be very useful for visualization.

5 Results

These resulting visualizations are saved in "viz_output.pdf". This document includes the predictions score for the image, whether that prediction was correct (True) or incorrect (False), the feature maps for all layers, and the Grad-CAM map for that image. It can be see in Figure 5 that there may exist perceivable differences in real and fake image activation mappings. As the authors of Meso-4 have already shown in a visualization they produced of the aggregate output of their data set there seems to exist some common activation for real images in the eye and mouth regions, while fake images tend to have the background be areas of interest.

Meanwhile Figure 6 shows the image regions used by the model to discriminate between real and fake. It seems that a common trend is the coloring of eye and neck region. Furthermore it seems that in the figure with hair, that this may have played a role in the final prediction. From the perspective of people, eye regions make sense as DeepFakes often struggle to create convincing eyes. Furthermore what is surprising is that the mouth region do not particularly contribute to the final outcome. Usually DeepFakes are created to have users say or do things they normally would never say. Perhaps this is simply due to these three images not having this be a common theme or a specific functionality of Meso-4.

5.1 Feature maps

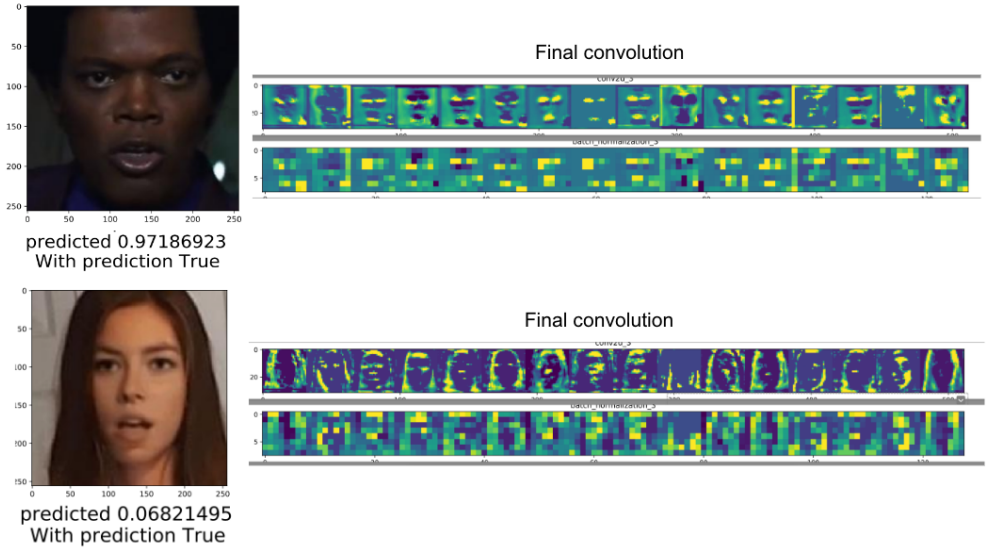


Figure 5: Feature map of the last two layers, the first being the final convolution layer. The first image is a real image and the bottom is a fake one both correctly classified. The final convolution layer was chosen to help reflect the most identifying features in the final output before flattening. As can be seen, it seems in this particular example that Meso-4 was very activated on the eye and mouth region of the first image. For the second image, which is fake, it seems more emphasis for features are placed in the background. These images were chosen for similar facial expressions, however there may be bias in regards to skin tone that are reflected in the mappings.

5.2 CAM

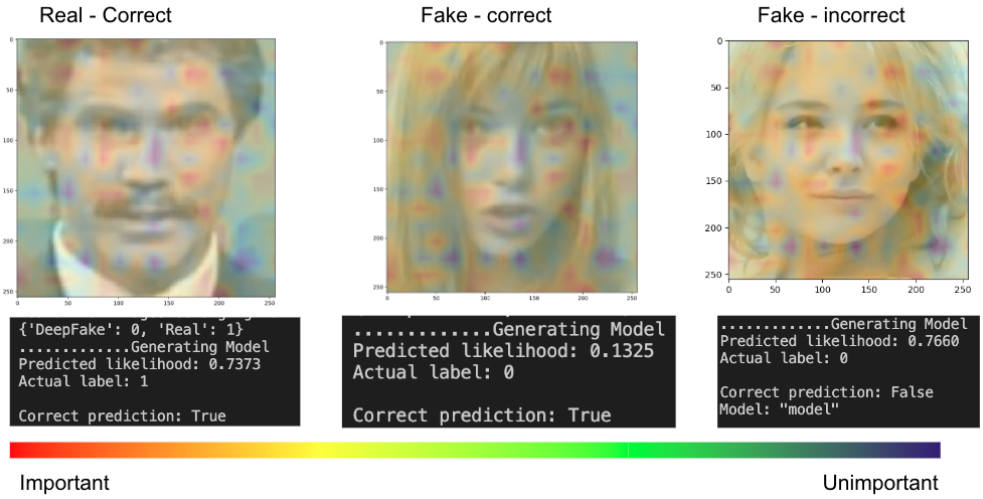


Figure 6: Result of CAM showing three outcomes a real image guessed correctly and two fake images, one guessed correctly and one incorrectly.

6 Limitations

Obviously there are limitations with our approach. The data set is limited in several ways. The first being the high prevalence of female photos in the Deep Fake data set that may skew results. Secondly, as can be seen with Figure 4, darker images and skin tones react differently to lightly ones as seen in the activation of the feature map.

Furthermore, it is difficult to interpret the Grad-CAM visualizations. As CNN remain a black box method, these visualizations help by allowing users to analyze model decisions. Some limitations already discussed for Grad-CAM includes the failure to identify local objects. Obviously for this task, users would like to identify what features contribute to the classification of "Real" or "DeepFake". Although the highlighted mappings expose the influence of certain features on the final classification, this is specific to the model and only helps in this regard. For users trying to differentiate Real and DeepFakes this visualization may not be as useful for general use, and instead, serves to help understand how Meso-4 perceives objects.

7 Possible Extension

Since videos frames are essentially images, it would be an interesting extension to extend this to deep fake videos. A bounding box could keep track of the fake objects in the frame. A further extension would be to apply the CAM visualization to the entire video. This could be utilized by social media companies such as Twitter and Instagram in order to further their "fake news" tags and included fake videos that contain DeepFakes.

Further analysis could be applied to the generated Grad-CAM images in order to determine if some relationship exists in classification among all images in the data set. Although the results could only be interpreted in the scope of the model, it could lead into insights on how to improve Meso-4 for broader application. This could look like taking the average activation mappings from feature maps from convolution layer 3 to as well as for the Grad-CAM images. However, due to the already resource intensive nature of this application, such visualizations were not created in this report and further modification to improve run-time would be necessary to average over thousands of images.

Since this project focuses solely on faces, it would also be interesting to see how it fairs with photo shopped objects and generated images that fill in missing parts of the image. This would involve reading in the image stream and applying the model to each frame. Another application would then be extending this to real time and allowing users to scan videos with their cameras in order to detect alteration.

8 Conclusion

The detection and understanding of DeepFakes continues to remain a major issue. As the complexity of DeepFake creation techniques improves it becomes easier than ever before to generate convincing fake images without intimate knowledge of the subject and expensive tools. This makes it a challenge for consumers to differentiate between the growing number of persuasive images and necessitates more robust technique to aid digital forensics.

This paper attempted to utilize the preexisting model and weights for Meso-4, a CNN made to classify DeepFake images. From this model, layers were pulled and processed in order to generate feature maps and Grad-CAM visualizations.

9 Reference

[1] Afchar, D., Nozick, V., Yamagishi, J., amp; Echizen, I. (2018). Mesonet: A compact facial video forgery detection network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS). https://doi.org/10.1109/wifs.2018.8630761

[2] Agrawal, S., Karandikar, A., Deshpande, V., Singh, S., amp; Nagbhidkar, S. (n.d.). Deepfake Video Detection Using Convolutional Neural Network. Retrieved June 6, 2022, from http://warse.org/IJATCSE/

[3] Auxier, Brooke, and Monica Anderson. "Social Media Use in 2021." Pew Research Center: Internet, Science Tech, 31 Jan. 2022, www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021.

[4] CS230 deep learning. CS230 Deep Learning. (n.d.). Retrieved June 6, 2022, from https://cs230.stanford.edu/

[5] DeepFaceLab. Softonic. (n.d.). Retrieved June 6, 2022, from https://deepfacelab.en.softonic.com/: :text=DeepFaceLab

[6] Gottfried, Jeffrey, and Jacob Liedke. "Partisan Divides in Media Trust Widen, Driven by a Decline among Republicans." Pew Research Center, 30 Aug. 2021, www.pewresearch.org/fact-tank/2021/08/30/partisan-divides-in-media-trust-widen-driven-by-a-decline-among-republicans.

[7] Herrman, John. "Fixation on Fake News Overshadows Waning Trust in Real Reporting." The New York Times, 19 Nov. 2016, www.nytimes.com/2016/11/19/business/media/exposing-fake-news-eroding-trust-in-real-reporting.html.

[8] Kite. (2020). How to detect DeepFakes with MesoNet — 20 Min. Python Tutorial. YouTube. YouTube. Retrieved June 6, 2022, from https://www.youtube.com/watch?v=kYeLBZMTLjk.

[9] Shearer, Elisa. "More than Eight-in-Ten Americans Get News from Digital Devices." Pew Research Center, 12 Jan. 2021, www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices.

[10] Matsa, Katerina Eva, et al. "Trust in America: Do Americans Trust the News Media?" Pew Research Center, 29 Mar. 2022, www.pewresearch.org/2022/01/05/trust-in-america-do-americans-trust-the-news-media.