Artificial Facial Detection, Separating Real Faces from the Artificially Generated: Project Midterm Report

Ante Tonkovic-Capin & Udhbhav Gupta

Introduction:

Artificially generated image technology has become increasingly sophisticated, sometimes termed as 'Deep Fakes', allowing the creation of realistic videos and images that can deceive viewers. In recent years this has become even more acute, with increasingly advanced techniques for creating artificial images and detecting them leading to a sort of arms race in the field. This is particularly dangerous when it comes to false or misleading representations of individuals. Detecting fake or manipulated images in the case where individuals are artificially generated is crucial to prevent their malicious use in spreading misinformation, identity theft, and other harmful activities. Our project aims to explore the potential of implementing a AI generated detection system for facial images using modern computer vision techniques.

Progress Made:

With generative modeling being a recent widespread success, a lot of prior work in this area of image forensics has focused on manual forgery (via Photoshop) detection. Techniques that analyze file format signatures, camera artifacts (like noise) and photography artifacts (like lighting and aberration) have been used with varying degrees of success. However, once such techniques are exposed, forgers, including generative models, can easily exploit them and produce artificial images that account for and comply with such image features. In the present environment of generative modeling, machine learning techniques stand a better chance as being somewhat of a black box, models can learn fake image detection without the specifics getting exposed to forgers.

Most current generative image models are built on the Generative Adversarial Network (GAN) architecture which comprises a Generator which generates artificial images by modeling the probability distributions of the underlying input image data, and a Discriminator which acts like a critic and tries to identify the artificial images generated from the real input ones.

We started our exploration into this project by studying types of machine learning models that have been successfully used for image classification. These include Convolutional Neural Networks (CNNs) and Vision Transformers. CNNs represent the fundamental deep learning architecture for image recognition and analysis tasks which process images in a grid-like fashion and use sliding filters to extract image features and information. Since the Discriminator of a GAN is built like a CNN, the CNN architecture makes for a promising candidate for this project. Vision Transformers represent a newer architecture for image analysis that divides images into patches, converts them into vector embeddings and processes embeddings using a transformer encoder to learn dependencies and relationships between parts of the images. For general image classification tasks, Vision Transformers have shown to provide better performance and accuracy results than CNNs.

We noted that previous image classification solutions with these model types have largely been based on classification of images based on their content eg: animals, landscapes, people etc. To test their performance on classifying images as real or fake, we first searched for datasets of real and fake human faces and came across the following on Kaggle:

- 140k Real and Fake Faces: This dataset comprises 70,000 real faces from Flickr and 70,000 fake faces generated from the StyleGAN model, originally collected and provided by Nvidia. The images

are sized to 256x256, split into training, testing and validation sets and collectively amount to ~4GB. (https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces)

- Real and Fake Face Detection: This dataset comprises approximately 1000 images each of real faces and fake faces generated by mixing of features from multiple different faces using Photoshop. This dataset was collected by Computational Intelligence and Photography Lab, Department of Computer Science, Yonsei University. The images are sized to 600x600 and collectively amount to ~250MB. (https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection)

Both datasets are comprehensive in terms of faces representing multiple ethnicities, genders, age and orientations. Following dataset discovery, we explored the need for and implemented some preprocessing techniques for the input images, discussed further in section 3.

As the next step, we implemented and trained models for a Convolutional Neural Network and a Vision Transformer, which take as input the training subset of images labeled as real or fake, learn the model parameters, and test accuracy on the test subset of image data. We trained and observed model performance for both the datasets mentioned above. Implementation details have been discussed in section 3 and results in section 4.

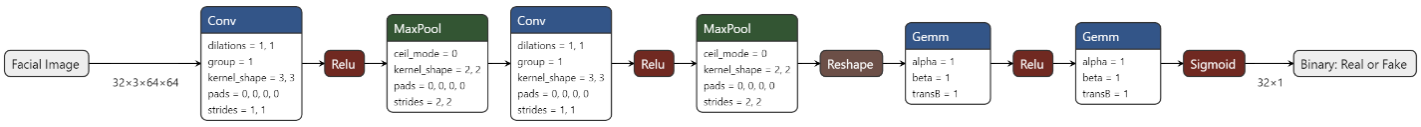
Methodology:

The goal of this project was to be able to detect artificially generated human faces from real ones. To accomplish this, our initial methodology focused on training a convolutional model on an available dataset of fake and real images to train the model and allow for accurate detection of input facial images as either real or fake. The first challenge we ran into was the dataset. While there is no shortage of image datasets out there, there is a wide range in terms of image size, mean and standard deviation of pixel intensity, quality of artificial manipulation or generation, varying formats and different focuses on facial features. We believe we finally found a quality dataset containing the right images of sufficient quality for our purpose, the dataset specified in section 2

To reduce the initial model training time and to move quickly on exploring various model architectures, we used randomized samples in the 5,000 to 10,000 range. We also applied some preprocessing to the input images, normalizing them using the datasets mean and standard deviation across all channels and applying a scaled resizing to reduce the image sizes from (128, 128, 3) to (64, 64, 3).

We initially tried training a base Vision Transformer model with patch size 16 in pytorch, however the model did not achieve satisfactory accuracy on either the Photoshop dataset or the StyleGAN dataset. Increasing the model complexity to the large Vision Transformer with patch size 32 also produced similar results (discussed in section 4). The Vision Transformer models also took a fairly long time to train on a single A40 GPU (~3 hours for the base model and ~5 hours for the large model).

We found more promising results training our own CNN classifier in `pytorch`. The model, `FakeFaceDetector`, consists of four total hidden layers, two convolutional layers followed by two fully connected layers:



The first convolutional layer (`conv1`) takes an input with 3 channels, applies filters of size 3x3, and outputs 32 feature maps. The second convolutional layer (`conv2`) takes the 32 feature maps from `conv1` as input, applies filters of size 3x3, and outputs 64 feature maps. The ReLU activation function is applied after each convolutional layer and after that a 2x2 max pooling operation is applied to reduce the spatial dimensions of the feature maps. The outputs from these convolutional layers are then flattened and passed through two fully connected layers (`fc1` and `fc2`). The first fully connected layer (`fc1`) has 128 neurons, and the second one (`fc2`) has a single neuron, corresponding to the binary output of the classifier. The ReLU activation function is applied after the first fully connected layer where afterwards the final output is passed through a sigmoid function (`F.sigmoid`), which squashes the output values between 0 and 1, making it suitable for binary classification. We're still trying to identify which layer structure, kernel size, pooling and stride parameters are best for our ultimate goal, but we seem to be making progress with this model structure and methodology so we're hopeful we're on the right track.

One change from our initial proposal might be to introduce a feature extraction step after the initial input preprocessing and resizing to try to detect and extract the bounding box surrounding just the face of the image using the pre trained MTCNN model to isolate and identify the key features comprising the human face. We hope this might improve our current best accuracy, around 83%, by removing additional noise and isolating only the relevant facial features required for our CNN to learn the correct parameters pertaining to artificial generation of fake facial features and not erroneous or peripheral features unrelated to the primary objective.

Another potential change to our initial proposal that we're exploring is to potentially include a post-model inference analysis. There is a lot of work on detecting digital signatures from manipulated images, and as our goal remains to detect real facial images from fake ones, the inclusion of some form of digital signature analysis such as an EXIF analysis or another technique for detecting artificially generated images. If available in the digital image signature, may increase our ability to detect manipulated images even further in the event our model is unable to correctly classify the input. A third potential change to our initial proposed methodology is to use images from multiple datasets originating from varying types of GANs to ensure a generalizable model able to infer across multiple types of previously unseen artificially generated images.

The primary evaluation metric of our project remains the same as it was in the initial proposal, to be able to detect fake facial images from real ones. We believe that our trained model's accuracy on unseen data is the best way to assess that outcome. This will be calculated during the test phase of training and represented by the ratio of correct predictions to the total number of previously unseen samples. We hope this will give a measure of how well the model's predictions match additional unseen data in the real-world.

Initial Results:

Our initial results include a lot of exploration into what the right dataset was for our project, as well as trying out which type of model and in what type of architecture worked best for our task. The two datasets our earlier exploration focused around, as discussed in the earlier section, was a dataset containing manually forged or "photoshopped" images of human faces along with real ones, labeled in the table below as 'photoshop'. With the second dataset containing more convincing forgeries generated by StyleGAN, labeled in the table below as 'StyleGAN'. The vit prefixed models correspond to their appropriate pytorch VisionTransformer model with the corresponding model type and patch size. Whereas the models prefixed ffd represent our own CNN and the corresponding randomized sample size used for training and testing. Here is the summary of some of

our initial trials during the early phase of model selection, sorted from lowest to highest test accuracy achieved:

Model	Dataset	Epochs	Learning Rate	Test Accuracy
vit-l-32	StyleGAN	20	5e-3	50.18%
vit-b-16	Photoshop	20	1e-2	52.58%
ffd-1k	Photoshop	20	3e-4	54.23%
ffd-5k	StyleGAN	20	3e-4	77.90%
ffd-5k	StyleGAN	80	1e-4	78.70%
ffd-5k	StyleGAN	80	2e-4	79.30%
ffd-5k	StyleGAN	50	3e-4	80.20%
ffd-10k	StyleGAN	50	3e-4	83.30%

As illustrated by the initial results, our CNN trained with 10,000 randomized samples for 50 epochs achieved the highest test accuracy on previously unseen test data. Accordingly, this is where we're planning on furthering our project and investing the additional training time and resources to try and improve our results.

An outcome that we didn't expect was that all the models would perform relatively poorly on the obviously photoshopped dataset. These images are less convincing to the naked eye than the artificial ones generated by GAN models. We've faced a few challenges while gathering our initial results, perhaps the most relevant one is the long training times associated with training models using higher quality RGB images. In order to achieve improved results, we're encountering longer and longer training times and GPU requirements. Another challenge we've faced is being unsure how much resources and energy should be spent on further improvements to initial models and architectures without knowing how well those future improvements would fair against previously unseen data that may have been generated by a different GAN or manipulated in a way completely foreign to our model and therefore lead to poor inference and classification.

Planned Project Activities

- Implement face isolation and extraction via MTCNN as an additional preprocessing step and explore impact on accuracy *(3 days)*
- Test increasing model depth and tuning training parameters to improve accuracy *(3 days)*
- Train model on entire 140,000 images instead of subset. We might need to adapt the model to a Distributed Data Parallel training method to be able to do this in a reasonable time frame *(7 days)*
- Explore and implement an additional detection technique/post-processing step involving digital signatures or image artifacts to combine with detection model *(7 days)*
- Evaluation and testing of model, collection of final performance metrics possibly including an additional dataset of images generated from a different type of GAN provided it can be obtained *(3 days)*
- Project Presentation, Webpage *(7 days)*

Contributions:

All group members, Ante Tonkovic-Capin and Udhbhav Gupta, contributed equally to various components of this project.

Signed 22 March 2024