On the detection of GAN-generated facial imagery

Harshal Agrawal Ricky Parada Colin Sullivan

1. Introduction

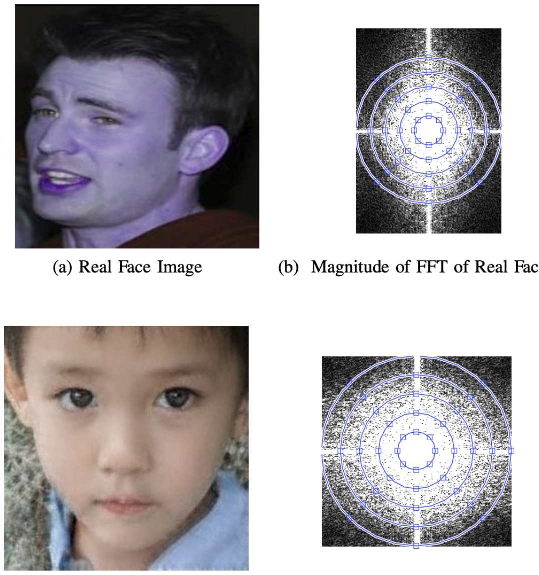
1.1. Problem

The AI revolution was (and is still) expected to threaten blue-collar work, but its largest impact thus far has hit an unlikely target — creators [6]. Artists, journalists, photographers, and even videographers have begun to face significant competition from generative AI models such as Midjourney, ChatGPT, and Sora which can generate human-level image, text, and video content, respectively. These creators have shot back by suing the developers of these models for their use of the creators' work in training with varying degrees of success (e.g., NYTimes [5]). With policymakers scrambling to regulate this new technology, concern about the impact of misinformation (e.g., Twitter bots, deep fakes) on the upcoming election, and the alarming spread of AI-generated internet content, a crucial question arises: can we differentiate real from synthetic content?

There exist many modes for which AI image content can be generated (e.g., cartoons, scenic photographs, paintings, facial imagery) and many potential methods across each of these modes (e.g., Generative Adversarial Networks (GANs), diffusion models). We restrict our focus in this project to facial imagery because it appears most consequential, considering the potential impact of impersonation on the upcoming election and already-realized effect on people's lives (synthesized "revenge porn" [3], identity theft, etc).

1.2. Related Work

Image quality assessment based fake face detection [7] a novel method to detect forged faces is proposed that trains a Random Forest (RF) classifier on Image Quality Assessment (IQA) based features. The approach is based on the hypothesis that the appearance of real and fake images is quite similar, so most of the discriminative information is available in the frequency and spatial domain of these images. As seen in Figure 1, visualizations of the difference in magnitude of the frequency domains between real and fake images appear to validate this idea. Further, the most utilized feature (highest SHAP value) is based on the frequency domain of input images. The approach achieves a reported 99% accuracy on a varied combined dataset of real images from CASIA (celebrities) + VGGFace2 (regular people of various professions), and fakes images from iFakeFaceDB (pulled from the This Person Does not Exist (TPDNE) dataset and modified by a GAN finger print remover). Performance on wild data suggests that the approach generalizes well.



(a) Real Face Image (b) Magnitude of FFT of Real Face

(c) Fake Face Image (d) Magnitude of FFT of Fake Face

Figure 1. We observe a significant difference between real and fake images in the magnitude of their respective frequency domains

GAN is a friend or foe?: a framework to detect various fake face images [8] As part of their neural-network (NN) based proposed framework, FakeFaceDetect, for detecting adversarially constructed fake facial imagery either on its own or as part of a larger photo, the authors try several classifier models, include three different shallow convolutional neural network (CNN) architectures and several fine-tuned deeper models. The authors note that, surprisingly, the shallow CNN classifiers, ShallowNetV1/2/3, trained from scratch perform significantly better than deeper fine-tuned approaches such as XceptionNet and VGG19. These larger models perform notably worse on the lower-fidelity 64p images than the proposed ShallowNet models. The data used in training and evaluation is pulled from the CelebA (celebrity images) and PGGAN (PG-GAN generated celebrity images trained on the CelebA dataset) datasets.

Fake Face Detection Methods: Can They Be Generalized? [4] The authors provide a new dataset, Fake Face in the Wild (FFW), of 53k images from 150 videos, originating from multiple sources of digitally generated fakes, including CGI generation, and the commonly used Swap-Face application. The authors make various attempts to classify this data, using both pre-trained deep CNNs (AlexNet, VGG19, ResNet50, Xception, GoogLeNet/Inceptionv3) and an SVM with Local Binary Patterns (LBPs) as features. While not particularly impressive, Xception appears to generalize best to new "unknown" data, but the LBP-featured approach also performs surprisingly well. The authors suggest that future research be directed toward the generalizability of fake facial imagery detection methods, suggesting that many current CNN-based approaches may fail to succeed in practice despite impressive experimental results.

1.3. Problem Statement

Given a perfectly balanced dataset of real and fake images with binary labels, we perform binary classification with accuracy as our metric of success. The images are RGB and 256p. No additional image metadata is used to inform classification.

Baseline We train a support vector machine (SVM) directly on the flattened pixel data. This approach has two potential upsides:

- The approach will either (a) identify any significant, easy-to-spot, biases in our image data prior to other approaches, or (b) provide a nice sanity check assuring us that the problem is non-trivial.
- We gain a potentially useful and interpretable prototype of what real and fake facial images look like.

We use hinge loss and stochastic gradient descent (SGD) to avoid loading our approach into memory, decaying the learning rate exponentially over time.

2. Dataset

We make use of the Kaggle 140k Real and Fake Faces Kaggle dataset [10] which consists of 70k real faces from Nvidia's Flickr dataset and 70k fake faces generated by Nvidia's StyleGAN [2]. The latter, fake data, was generated and compiled by Bojan Tunguz as a Kaggle dataset of 1 million such samples at 1024p [9] before it was down-sampled to 256p and subsampled from to get the 70k images in the Kaggle dataset we use for this project. We note that this means our model will essentially act as a discriminator for the StyleGAN generator. This approach of training on both data generated by a model and the data the was trained on has been done [8] and reduces the potential for bias introduced in the image-gathering process, allowing our classifier to focus instead on the peculiarities of the fake image generation process. We also believe this makes the problem tractable for us for the short duration of our project and that some of the findings we discover while trying several different approaches may generalize to facial imagery generated by other models. To this end, we will additionally test our approaches on wild data at the end of our project: including, potentially, images from the CelebA, TPDNE, and PGGAN datasets.

3. Technical Approaches

Frequency and LBP-based Features Classifier This approach is based off of the finding from [7] that high-frequency image data is helpful in differentiating between real and fake images combined with the semi-successful use of LBP features in [4]. We concatenate the Fourier Transform of our input image with its LBP features and train a Random Forest (RF) to perform classification.

Fine-tuned Large Model Inspired by some promising results of the Xception model [1] in the previously cited studies [4] [8], we experiment with this deep pre-trained model ourselves. We freeze the final layer of this model, and fine-tune the final fully-connected layer weighted for our fake image classification task.

Shallow CNN We train a shallow CNN from scratch for our classification task, the idea being that even humans struggle to differentiate these real and fake images, so perhaps we need to look for non-intuitive features that would not have been picked up on by the previous larger models. For this model, we use the 2-strided convolutional layers in the architecture shown in Figure 2 with 2x2 max pool and batch norm layers inserted every other convolutional layer. We use BCELoss and optimize with the Adam optimizer at a learning rate of 10^{-4}.

4. Preliminary Results

Due to time and compute constraints (not able to secure cloud compute resources in time), we decided to use a smaller sample of our data to get preliminary results. We trained our model on 1000 images from our training data, evaluating our model after training on 200 images from our validation data for 10 epochs. After training, we evaluated the model on a small test set of 100 images (pulled from a different subset of validation data in order to preserve the

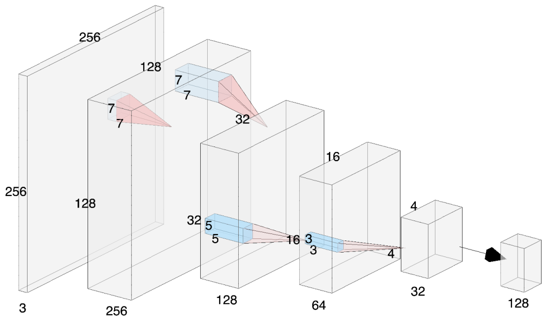


Figure 2. Shallow CNN Architecture

test dataset during full training). This gave us an initial read on the viability of our first proposed approach, the Shallow CNN, as well as our baseline method.

	Train	Validation	Test
Baseline (SVM)	0.917	0.650	0.490
Shallow CNN	0.925	0.729	0.706

Table 1. Final accuracy of our baseline and initial CNN Approach across all datasets.

From Table 1, we observe that the Shallow CNN performs marginally better than our baseline SVM across all datasets, especially the test set. The training loss and accuracy plots in Figure 3, 4 and 5 indicate that the baseline model underfits the data significantly whereas the Shallow CNN continues to learn at a steady rate. We also note that our model may benefit from further training.

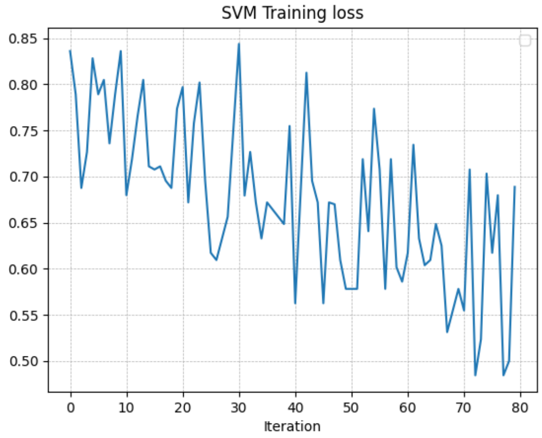


Figure 3. Training loss for baseline SVM.

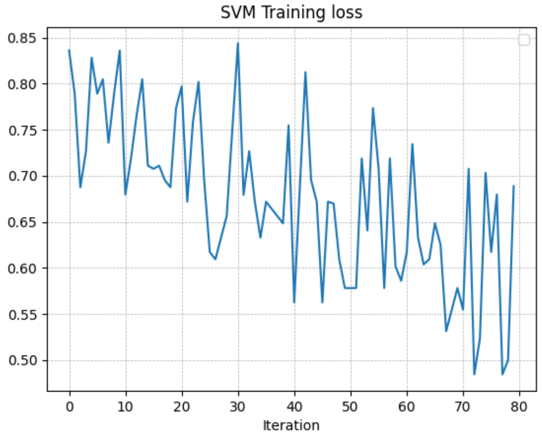


Figure 4. Training loss for shallow CNN.

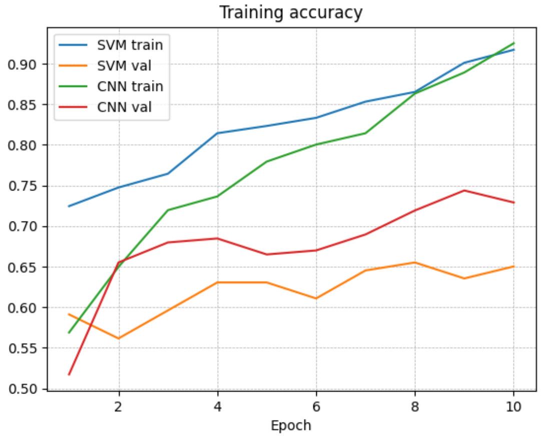


Figure 5. Training accuracy for shallow CNN and baseline SVM.

References

[1] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[2] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks, 2019.

[3] K. Kelleher. Revenge porn and deep fake technology: The latest iteration of online abuse. Boston University School of Law: Dome, 08 2023.

[4] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2018.

[5] R. M. Michael M. Grynbaum. The times sues openai and microsoft over a.i. use of copyrighted work. NYTimes, 12 2023.

[6] A. Roy. Blue-collar jobs may weather raging ai storm better: experts. The Economic Times, 01 2024.

[7] K. S. and V. Masilamani. Image quality assessment based fake face detection. *Multimedia Tools and Applications*, 82, 01 2022.

[8] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo. Gan is a friend or foe? a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 1296–1303, New York, NY, USA, 2019. Association for Computing Machinery.

[9] B. Tunguz. 1 million fake faces. Kaggle, 2019.

[10] xhlulu. 140k real and fake faces. Kaggle, 2020.