IDS 705 Project Proposal

Classification of Fake and Real Faces

And Its Implications for Dating Apps

Team Members:
Kashaf Ali
Beibei Du
Eric Rios Soderman
[in alphabetical order by last name]

Team Number: 3

Introduction

In the world of social media, fake profiles run absolutely rampant. A social media profile, in general, is a template that houses information and pictures indicative of the user's taste. Often, they convey some type of the users' information. On the other hand, fake profiles are ones that are generated by either disinterested or shy people on one extreme of the spectrum and people with nefarious intentions on the other. These false depictions can sometimes come accompanied by fake pictures and information. In the current age, one of the biggest and new advancements is the rise of AI-generated photos. Currently, and even more so in the future, fake images can be generated by certain individuals, who will use them as tools in phishing scams, catfishing, and other types of crimes [11].

The problem that we are trying to answer is strictly academic, as we have no plans on making a commercial tool from this. We will implement tools to train on real and fake image data, in order to test real and false images, which are proxies of the ones that would normally be present in a dating profile. We want to evaluate the viability and the implications of these models, which can potentially identify fake users as well as potentially harm or aid real users. In addition, we also want to get a sense of the limitations of these models, as well.

Goal/Objective

The Main Goal: The problem is understanding the implications of discriminating between real and fake image data for the types of pictures seen on dating apps. The types of pictures seen on dating apps are predominantly headshots or selfies and then pictures with friends.

Main Questions:

1. Will the model that we train have a racial bias, gender bias, and issues with pictures that do not have enough facial proximity or focus?

2. How well will this model perform when the fake images are good quality, and what is its long-term viability in the application space of dating apps?

3. What are the limitations of both the model selection and/or the application space?

4. What are the insights that we can glean to foresee the future of dating apps and the rise of elaborate scams?

Reach or Optional Goals:

1. To increase classification performance, we are considering stacking multiple model architectures, composed of CNNs, Vision Transformers, and GANs. We also aim to compare the performance of these models to understand which one performs better.

2. We are also considering an unsupervised learning model to learn what makes a real image as a form of feature extraction, and then use that information for classification.

3. Making fake images and assessing fake image quality by fooling discriminative models is also an extension of this study.

Realistically speaking, the thought process of designing the optional extensions of this study is rewarding enough. There exists a strong definitive possibility that we will lack both the time and computational resources to successfully reach a model stacking phase.

Background

In the current world of rapid progress in Artificial Intelligence (AI), the automated generation of poems, lyrics, music, and images has become increasingly popular in recent years. While newly developed and advanced technologies can bring convenience to human lives, there are many drawbacks, especially when ethical concerns come into play. Many researchers have delved into the topic of deep fakes, which we are also interested in, to differentiate between AI-generated faces and actual human faces. This is particularly important in a social context, as many people use fake faces as their profile pictures on dating apps, and fraud can sometimes occur.

Rosa et al. studied fake faces in the context of Tinder, where they stated that "judgments of moral character" may be more crucial than appearance in dating apps [1]. Specifically, they argued that physical attractiveness is a cue for rapid judgments in the swiping decision-making process. Thus, the profile picture and personal photos posted on the Tinder profile are decisive, and from a social psychology perspective, they tend to correlate with the cognition of morals. The emergence of fake faces might distort the original intention of dating apps, resulting in harmful social interactions.

This leads us to the question: How can we differentiate between real and fake facial images in order to prevent any potential detrimental influence on dating apps and online social interactions, or even the illegal theft of facial privacy? Luckily, there are many scientists who have tried different algorithms to predict whether an image depicts a real face or not in order to seek ways to prevent ethical violations of fake face usage online.

For this project, we aim to combine various approaches from different papers. We will refer to multiple papers in the references section to construct our classification algorithms.

Data

We plan on using the following datasets for the training, validation, and testing stages of the project.

1. **Flickr FacesHQ Dataset** - (linked here)

This dataset contains 70,000 high-quality PNG images of human faces at 1024 x 1024 resolution with considerable variation in terms of age, ethnicity, and image background. It was collected by NVIDIA from Flickr and thus, inherits all the biases of that website. The dataset also includes images with multiple accessories like glasses, hats, etc. The images were aligned and cropped using dlib, automatic

filters were used to prune the data, and Amazon Mechanical Turk was used to exclude paintings and statues from these images to make them ready for use as a training set.

This dataset is relevant to our project objectives as our aim is to design a model that is able to effectively detect fake images on dating apps and social media profiles, and most of those profiles have similar images of human faces with accessories. Moreover, the dataset is large enough to be used as our training and validation set.

2. **1 million fake faces** - (linked here)

This dataset contains 1 million images of human faces that have been artificially generated using a variety of Generative Adversarial Network models and it was created by the V7 Labs. Moreover, each image is of size 1024 x 1024 pixels, and is in the JPG format. The images also include a diverse range of races, ages, and genders.

We will be merging these two datasets for our model. However, we plan on exploring the following as an initial step:

- Do both datasets contain similar kinds of images, i.e is the division of gender, accessories, age, etc similar across both datasets?
- Will merging these datasets skew our balance of fake vs real images (ideally we would want approximately the same number of real and fake images in the final dataset)?

Datasets for the application space

In order to contextualize our tool in our application space, we hope to understand how our model results will perform when provided with new data. In this case, we will test our model on new data to drive qualitative insights on why our model performs better or worse on a new type of data. We will be using the following datasets for this purpose:

1. **Human Faces** - (linked here)

The Human Faces dataset on Kaggle contains 10,177 images of human faces, each with a resolution of 224 x 224 pixels. The images include a diverse range of ages, races, and genders, and the individuals depicted in the images are from various backgrounds and regions around the world.

2. **People Image Dataset** - (linked here)

This dataset on Kaggle is a collection of 2000 images from a variety of categories such as animals, landscapes, people, and objects, with resolutions varying from 128 x 128 pixels to 1024 x 1024 pixels.

Methods

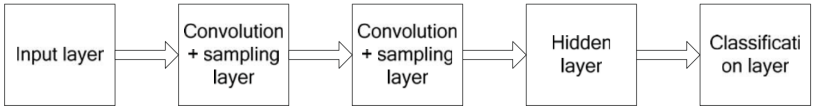
We have decided to consider using one of three potential machine-learning algorithms for our classification task at the first level. If time allows, we plan to compare the performance of all three algorithms or use ensemble learning techniques.

1. **GAN (Generative Adversarial Network)**

a. The formula is listed as: L(G) = αLGAN(cross-entropy loss) +βLadv +γLpert [6]
b. The generator and the discriminators are trained in a competition, with the generator trying to fool the discriminator and the discriminator trying to correctly identify the generated data. Over time, the generator gets better at creating realistic samples and the discriminator gets better at identifying generated data
c. GANs can be used to generate additional training data, which can improve the performance of image classifiers. For example, if there are few images of a particular class in the training set, a GAN can generate additional images from that class, which can help the classifier learn to recognize that class more accurately. Besides, it can generate adversarial examples, which are images that are designed to fool an image classifier, which can become more robust to these types of attacks.
d. GANs can be employed to produce synthetic images that closely resemble the desired target domain, and subsequently, these generated images can be utilized to pre-train an image classifier. To illustrate, suppose the objective is to classify deep fake face images; in that case, a GAN can be leveraged to generate artificial fake face images that mimic the actual ones, thereby enabling pre-training of a classifier on the synthesized images before fine-tuning it on the real dataset.

2. CNN (Convolutional Neural network)

a. $$h_{W,b}(x) = f\left(W^T x\right) = f\left(\sum_{i-1}^{3} W_i x_i + b\right)$$



[7]

b. CNNs are an effective tool for detecting deep fake images as they can learn to identify the subtle patterns that are unique to fake faces. These patterns can include blurred edges, discrepancies in lighting and shadows, and other unnatural facial expressions. The CNN model is trained on a dataset consisting of both real and fake faces, allowing it to learn the features and patterns that are indicative of a fake image.
c. Once the CNN has been trained, it can be used to classify new images as real or fake by analyzing the features present in the image. This process involves extracting the unique features of the image and comparing them against the learned patterns from the training data. If the image contains features that match those of the fake faces in the training set, CNN will predict whether the facial image is fake or not under the dating app setting.

3. Vision Transformer (ViT)

a. The biggest benefit compared to CNN: is less computational power needed for better

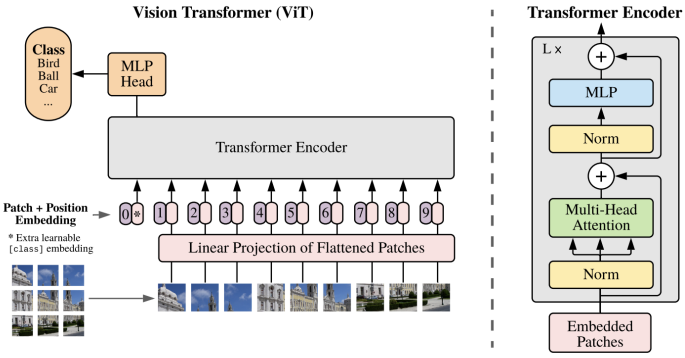


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

performance [8]

b. ViT processes image data similarly to how we typically process text data in NLP, by taking input data as sequences of tokens to capture details that may help classify whether a face is fake or real. This approach has shown better performance than CNNs in deep fake face classification.
c. In deep fake face classification, ViT splits images into a certain number of patches and feeds them into a traditional Transformer encoder for processing and classification. To improve the model's performance, we can fine-tune it by adjusting the number of layers, the hidden size, and the MLP size. By making these changes, we can optimize the ViT model's ability to identify the subtle patterns and features that are indicative of fake faces and achieve better accuracy in classification.

Experiments

Our study has two, certain experimental components. The first is preparatory in nature, where we tune our chosen models' hyperparameters with a validation set. The second component transpires after the model is tested on the test set. The model will be fully tested on proxy data for the type of pictures present in dating apps. This second component is the true experiment in this pipeline.

- Outcomes:
  - Test and validation accuracies during the training process
  - Overall accuracy on the application's proxy dataset.

- Performance Metrics
  - Validation and test accuracy (referenced across studies)
  - Classification accuracy on application's proxy data

- Reference Studies :

- Classification of Real and Fake Human Faces Using Deep Learning
  - **Authors:** Fatima Maher Salman and Samy S. Abu-Naser
  - **Study:** They wanted to discriminate real from fake faces generated by GAN models [5]
  - **Models:**
    - VGG16, ResNet50, MobileNet, and InceptionV3
    - Different flavors of the models were used to assess validation accuracy and test accuracy.
  - **Our takeaway**: We learned what model flavors we can utilize

- Using a GAN to Generate Adversarial Examples of Facial Image Recognition
  - **Authors:** Andrew Merrigan and Alan F. Smeaton from Dublin City University
  - **Study:** They evaluate the strength of the GANs in generating data to fool the discriminator [6].
  - **Model Selection:**
    - The introduction of Adversarial Transformation Networks
      - GANs optimized to fool the discriminators
  - **Our Takeaway:**
    - Gave us an idea of what GANs to consider if we desired to create our own synthetic data.

- Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection
  - **Vision Transformers**
    - The images are split into patches and are fed through the Transformer architecture.
    - The patches are 16x16 pixels, and the embeddings are vectors that can be learned [8].
  - **Study**: Although shown to be less effective than the EfficientNet variant of the CNN, a Vision Transformer was shown to be an alternate model to help with image classification [10].
  - **Model Selection**:
    - We are still evaluating which model flavor to use, as the architecture is a little over two years old.
  - **Our Takeaway**:
    - We found an additional alternative of a model that can be used to classify real and fake faces.

Roles

Here is a tentative overview of how we will divide specific roles and responsibilities:

- Data Preparation
  - Picture Rotations
- Computational Resource Search

    - Google Colab
    - Physical Laptops
    - Training Time Assessment
- Model Training
    - Model Selection
    - Model Training
        - Resource and time estimation
- Hyperparameter Tuning (Validation)
- Testing and Predictions
- Report Writing

We plan on ensuring sufficient code review throughout this process to ensure that we are offering ample support to each other in finishing all of these required steps. So far, no particular member has been assigned to any given process, since we are still very new to the space and have to research the implementation of such models.

Limitations and Challenges

The limitations span multiple areas of this study. Beginning with the pictures themselves, we are concerned with the fake and real images overly adhering to a certain pattern. For example, maybe 40% of fake images have people tilting to the left of the picture. That imbalance will harm the model's generalization performance, as well as have the model learn the wrong thing. Another issue here is the dataset itself. Referencing the findings from Maher Salman and Abu-Naser (2022), they caution against fake images coming from different sources, as these images are manufactured with different types of noise and compression, as well as coming from multiple and different flavors of GANs.

Another area of limitation is the model selection. We may lack the time and expertise to implement a stacked, model architecture. Therefore, we have to currently focus on one model. Other goals or expansion of this project may be out of reach, such as applying or training other models for the problem. In addition, the most obvious hurdle is a computational constraint. CPUs and GPUs are fast, but the data is large, and the models take a long time to train. Training time, by far, is the strongest limitation, in addition to paying for the cloud computing required for the operations.

The last and final limitation is the proxy data for the application space, as well as the nuances of that space. We cannot legally obtain dating profile pictures, so we are using data that is reminiscent or proxy to what one could find on a dating app. Also, our model choice may generalize well to the test data, but it may perform horribly with the proxy data. For example, how will a model train on headshots to classify a picture of a person sitting on a couch with a parakeet? However, even failing at this task provokes great interest in us, as it will inform the limitations of our model choices, as well as the features our models extracted or learned. Furthermore, there are inherent challenges present in the application. Many scams involve criminals misrepresenting themselves by using pictures from real people, so our model could never account for something of that nature. In fact, an extensive study of this nature is what could be recommended to find ways to further isolate fake pictures from the real ones.

References

[1] Olivera-La Rosa A, Arango-Tobón OE, Ingram GPD. Swiping right: face perception in the age of Tinder. Heliyon. 2019 Dec 2;5(12):e02949. doi: 10.1016/j.heliyon.2019.e02949. PMID: 31872122; PMCID: PMC6909076.

[2] Taeb, Maryam, and Hongmei Chi. 2022. "Comparison of Deepfake Detection Techniques through Deep Learning" Journal of Cybersecurity and Privacy 2, no. 1: 89-106. https://doi.org/10.3390/jcp2010007

[3] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[4] Shahzad HF, Rustam F, Flores ES, Luís Vidal Mazón J, de la Torre Diez I, Ashraf I. A Review of Image Processing Techniques for Deepfakes. Sensors (Basel). 2022 Jun 16;22(12):4556. doi: 10.3390/s22124556. PMID: 35746333; PMCID: PMC9230855.

[5] Maher Salman F, Abu-Naser S. Classification of Real and Fake Faces Using Deep Learning. 2022. International Journal of Academic and Engineering Research. Vol. 6, pp. 1-14. 2022 March. ISSN: 2643-9085. https://philarchive.org/archive/SALCOR-3

[6] Merrigan A, Smeaton A. Using a GAN to Generate Adversarial Examples to Facial Image Recognition. 2021 November 30. arXiv:2111.15213v. https://arxiv.org/pdf/2111.15213.pdf

[7] Wang, J., & Li, Z. (2018). Research on face recognition based on CNN. IOP Conference Series: Earth and Environmental Science, 170, 032110. https://doi.org/10.1088/1755-1315/170/3/032110

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, June 3). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv.org. Retrieved March 10, 2023, from https://arxiv.org/abs/2010.11929

[9] M. Coşkun, A. Uçar, Ö. Yildirim and Y. Demir, "Face recognition based on convolutional neural network," 2017 International Conference on Modern Electrical and Energy Systems (MEES), Kremenchuk, Ukraine, 2017, pp. 376-379, doi: 10.1109/MEES.2017.8248937.

[10] Coccomini, D. A., Caldelli, R., Falchi, F., Gennaro, C., & Amato, G. (2022, June 28). Cross-forgery analysis of vision transformers and CNNS for deepfake image detection. arXiv.org. Retrieved March 10, 2023, from https://arxiv.org/abs/2206.13829

[11] Aura. The Unexpected Dangers of Online Dating [11 Scams to Know]. Retrieved March 10, 2023 from https://www.aura.com/learn/dangers-of-online-dating