Deepfake Image Classification Analysis

Abstract
Investigating deep fake images is vital due to their societal impact—spreading misinformation, manipulating perception, and breaching privacy. Developing detection methods is crucial for mitigating these harms and addressing ethical concerns in AI development, requiring vigilance amid evolving technology in computer vision and artificial intelligence, focusing on image analysis and pattern recognition. How can we identify specific visual features or deep learning model activations, such as texture irregularities, color patterns, or high-frequency noise, that serve as the most reliable indicators for detecting AI-generated fake images within large-scale datasets? Our dataset, sourced from Kaggle, comprises facial images uniformly sized and categorized into real images and expert-labeled deep fake images with varying difficulty levels—easy, mid, and hard. These curated facial images offer a diverse range of high-quality, human-expertly manipulated content, distinguishing themselves from computationally generated images. This dataset is specifically tailored for training classifiers to discern between real and sophisticated expert-created fake facial images, crucial in distinguishing nuances missed by models trained solely on computer-generated content. Our study encompassed diverse models—SVC and Multi-Layer Perceptron—evaluated across a dataset featuring combined deep fake images categorized by difficulty levels (easy, mid, hard) and real images. Notably, we incorporated grayscale and PCA techniques across all models. Across the entirety of the dataset, SVC demonstrated an accuracy of 0.531. Meanwhile, Multi-Layer Perceptron, utilizing PCA, achieved an accuracy of 0.557. When specifically examining easy deep fake images against real ones, SVC produced an accuracy of 0.819 and MLP produced an accuracy of 0.791. This comprehensive analysis highlights the varying performance of SVC and Multi-Layer Perceptron concerning different difficulty levels of deep fake images. While both models excelled in distinguishing easy and hard scenarios, their performance remained relatively consistent across various difficulty levels, indicating robustness in their predictive capabilities. It emphasizes responsible AI development and ethical issues in the search for accuracy by navigating the continuing conflict between AI and reality. Identifying deep false pictures is a careful balance between fighting fake news, protecting people's rights and privacy, and making sure AI is developed in a way that is ethically sound as technology is always evolving and becoming more complex. Understanding deep fake image classification has enhanced our understanding of the process of image classification and building a model for one. For future direction we can understand the fundamentals when it comes to implementing the process of deep fake detection. We can utilize more sophisticated models so we can understand more of the decision-making for the predictions.

Introduction
In a modern context, deepfakes are images generated by computer software, made to look like they were real. Highly sophisticated deepfakes often have very few distinguishable features that can be used to detect whether the image is real or fake. It is essential to investigate deep fake images that can harm society by proliferating misinformation, manipulating the public's perception, and violating individual privacy. By exploring and developing methods to detect deep fake images, we can learn how to mitigate the negative impact of deep fake technology. Understanding how to detect and address deep fakes is also powerful knowledge for navigating ethical questions. Since technology is continually evolving, it is now more essential than ever to stay ahead and informed for research to develop actions accordingly. Notably, grasping deep fake technology and its ramifications provides for responsible development and deployment of AI systems and guarantees they are used with society's best interest in mind. This inquiry of classifying deep fake images falls under computer vision in artificial intelligence. It involves image analysis and pattern recognition. Our project is an exploration into a critical area of computer vision and artificial intelligence, posing the research question: "How can we identify specific deep learning model activations, such as texture irregularities, color patterns, or high-frequency noise, that serve as the most reliable indicators for detecting AI-generated fake images within datasets?"

Related Work
Recent years have witnessed a remarkable surge in deep fake investigations, driven by the application of Machine Learning and Deep Learning techniques that offer automated detection solutions. Diverse approaches have been explored, leveraging MLP, SVM, CNN, and hybrid frameworks for deep fake detection. While each approach has tackled specific challenges, they have encountered limitations concerning accuracy, robustness, and computational complexity. Despite notable achievements in accuracy, many models have grappled with challenges related to generalization, computational costs, recognition of compressed images, and handling low-quality videos. These performance issues underscore the complexities inherent in deep fake detection. Moreover, current Deep Learning algorithms, while promising, present challenges such as computational expenses, susceptibility to overfitting, and

substantial resource demands. Consequently, there is an urgent need to develop more robust, efficient, and accessible methods for deep fake detection and classification. Among the diverse approaches, some innovative techniques have emerged. For instance, a Hybrid Multi-Task Framework has employed the Firehawk Optimizer, significantly enhancing accuracy in detecting Arabic fake news. Additionally, novel architectures like the Convolution Vision Transformer (CVT) and MesoInception-4 have integrated attention mechanisms, CNN, and capsule networks, showcasing innovative strides in combating deep fakes. These advancements signal potential pathways for future research and development in this critical area.

Data

The dataset, sourced from a CILAB at Yonsei University on Kaggle (Yonsei, 2018), presents meticulously crafted composite images. These images are composed of distinct facial elements—eyes, nose, mouth, or entire faces—painstakingly assembled by human experts. The primary objective behind this dataset is to facilitate the training of a classifier with the ability to discern between authentic and expertly forged facial images. Structured within subdirectories, the dataset segregates content into 'training_real,' housing genuine facial photos, and 'training_fake,' which contains expertly manipulated fake face images. The latter category is further subjectively divided into three groups: 'easy,' 'mid,' and 'hard,' representing varying levels of complexity in the manipulation of these images. This organization aims to provide a diverse range of forged facial images for comprehensive classifier training and assessment. Initially, we had chosen a different dataset from HuggingFace (AI CONNECT, 2023), containing 28.3 gigabytes of real and fake images. However, this dataset contained many deepfakes, ranging from faces to objects or landscapes. These images were also of varying resolutions. Due to these obstacles, we were unable to conduct principal component analysis on the images and could not find an optimal resolution to resize the images without obscuring their details. We also determined that the models would likely face difficulty in finding patterns within the data due to the large amount of noise. Lastly, this dataset did not necessarily serve the ethical reasons behind this project, as deepfakes of human faces were more likely to cause harm to society than those of objects or landscapes. To narrow the focus of our project and ensure more consistency, we chose the Kaggle face dataset instead.

Approach

- **Pre-processing data:** In order to prepare the data for fitting, the images were resized, converted to grayscale, and transformed by PCA, with a number of components of 200.
- **Subsetting Data:** Each model was fit 4 times with different datasets. The first dataset contained all images in the dataset, and the other 3 were split into ‘easy’, ‘mid’ and ‘hard’ to represent how difficult it would be for the human eye to detect a deepfake. These classifications were provided within the Kaggle dataset. The data was stratified to account for class imbalances, and a 80-20 train test split was used to fit the models.
- **Model selection:** The models chosen were Support Vector Classifier (SVC) and Multi-Layer Perceptron (MLP) from the sci-kitlearn package.
- **Metric Selection:** As the models were solving a classification problem, accuracy score from the sklearn.metric package was chosen to measure the accuracy of the models.
- **Hyperparameter Tuning:** GridSearchCV was used to determine the best estimators for both models.
- **Other experiments:** Eigenfaces were generated in order to gain insight into which features varied the most throughout the images.

Experiments

- **Data Preparation**: Before fitting the model, we used resizing, grayscale reduction, and PCA in order to reduce the dimensionality of the data. While the initial images had a 600x600 resolution, we resized them to 100x100 to reduce the number of features. The images were then converted to grayscale by utilizing numpy. Lastly, through PCA, we generated a cumulative explained variance graph to pick the optimal number of PCA components. A value of 200 was chosen, as this coincided with approximately 95% cumulative explained variance. The data was then transformed through PCA.

- **Subsetting Data:**. To subset the data, we utilized the file names and imread in the sklearn.io package to classify the files into easy, medium and hard. There were 240 fake images in the easy subset, 480 fake images in the medium subset, 240 fake images in the hard subset. Each subset included all the real images provided. The combined dataset had approximately 1000 real images and 1000 fake images.
- **Splitting and Stratification:** The images were split using an 80-20 train-test split, and stratified to deal with the class imbalances.
- **Eigenface Analysis:** To gain more insight into the data, an analysis of the images through eigenfaces was conducted. The goal of this experiment was to determine what features of the images varied the most. At a lower number of components, most of the images had varying background color or skin tone. At a higher number of components, the most important features were complex facial features in the center of the face, such as the eyes, nose, and mouth.
- **Model Selection:** The models chosen were Support Vector Classifier and Multi-Layer Perceptron. The reason why we chose SVC was due to its ability to handle high dimensional spaces and also handle non-linear data such as pixels. We also decided to use SVC due to its ability to utilize small amounts of data very efficiently. The second model we chose was a Multi-Layer Perceptron model. The main reason we chose to use it is the simplicity in being able to interpret the machine learning process which allows us to have a clear understanding of how and why the model is making decisions. Although one downside of a Multi-Layer perceptron approach is typically these models are more data hungry as it is a neural network which could pose a limitation due to our dataset size.
- **Metrics:** As mentioned earlier, accuracy score was chosen as the key metric. We determined both the training and testing accuracy through running the model three times and calculating the mean accuracy.
- **Hyperparameter Tuning:** For each model, GridSearchCV was used to pick the hyperparameters that served as the best estimator of the data. GridSearchCV was carried out on each iteration of each model separately.

Results

Combined Dataset	
SVC Training Accuracy: 1.0 SVC Testing Accuracy: 0.531	MLP Training Accuracy: 0.885 MLP Testing Accuracy: 0.557
Easy Dataset	
SVC Training Accuracy: 1.0 SVC Testing Accuracy: 0.819	MLP Training Accuracy: 0.811 MLP Testing Accuracy: 0.791
Medium Dataset	
SVC Training Accuracy: 1.0 SVC Testing Accuracy: 0.693	MLP Training Accuracy: 0.691 MLP Testing Accuracy: 0.651
Hard Dataset	
SVC Training Accuracy: 1.0 SVC Testing Accuracy: 0.819	MLP Training Accuracy: 0.807 MLP Testing Accuracy: 0.790

Discussion

Some findings we derived from this project are first within the Multilayer Perceptron model as it actually performed best within the combined dataset while it underperformed in the easy, medium and hard dataset. This result leads us

to conclude that because of the nature of MLP models which is a type of neural network and typically data hungry. If our dataset had been expanded the MLP model may have performed better in the long run, but this is yet to be fully proven which if we had more time we may have been able to include more images/generate data. Alongside this we also were able to identify that the SVC model actually performed better in the restricted cases(easy,medium and hard) while it underperformed in the combined dataset which may be indicative of noise being introduced. These two takeaways are important as it gives us a deeper understanding of what models may perform better when doing image recognition and future deepfake detection which in this case out finding point towards collecting more data and trying other types of neural network models.

Limitations and Future Work

Since deepfake detection is still a novel research topic, we faced many challenges throughout the project. The most notable limitation was the lack of a larger dataset. Neural networks like Multi-Layer Perceptron are more suited for image classification, with one downside being that they need a large amount of data to result in more accuracy. The size of the dataset likely had an impact on the accuracy. One area where this is visible is in the SVC model for the easy and hard dataset: each had only 240 images and resulted in the same accuracy. Another notable limitation is model selection. Through preliminary research, we found that models suited for computer vision such as convolutional neural networks would provide better fits, but this was out of the scope of the sci-kit learn package. Conducting the same experiments using a more sophisticated model through PyTorch or TensorFlow could potentially yield more insightful results. Lastly, processing images is computationally expensive, and we had to resize our images down from 600x600 to 100x100 to run the code for our project. This limitation may have obscured some of the inconsistencies in the images and made it more difficult for the models to find patterns in the pixels. Overall, our goals for future work on this project are expanding our datasets to include more images, fitting more sophisticated models, and employing image processing techniques that are less computationally intensive would greatly contribute to improving our project.

Ethical Considerations

Some ethical considerations we have for this project is due to the nature of the data. It may be subject to privacy concerns as using people's face images can be an invasion of privacy and thus make it more difficult to collect and utilize real faces when training models at large scales. Alongside this there is also the possibility of potential bias in how certain models determine if someone’s face is a deepfake for example if someone has a distorted face/large facial injury because most machine learning detection models use distortions it may misclassify people with disabilities and facial injuries disproportionately. Finally a major ethical concern which may arise is bias within image selection itself as making sure there is an equal distribution of race and culture within the dataset is imperative as not being equitable within that distribution may lead to worse detection of deepfakes for underrepresented and marginalized communities at risk of misinformation.

Conclusion

The investigation uncovered several pivotal insights in the realm of deep fake detection. It shed light on challenges concerning data imbalance, model sensitivity to different complexity levels, and the impact of noise interference in amalgamated datasets. These findings emphasize the critical hurdles that need addressing in the pursuit of robust and accurate detection systems. Moreover, the study employed a diverse array of approaches, leveraging Support Vector Classifier (SVC) and Multi-Layer Perceptron (MLP) techniques, offering a comprehensive approach to image classification. This varied methodology showcased the importance of employing multiple techniques for a more thorough and nuanced analysis of deep fake images. Notably, significant enhancements were implemented in detection systems to fortify their resilience against evolving deep fake technologies. These improvements represent a proactive step towards developing more resilient and adaptable detection mechanisms, crucial in the ongoing battle against the proliferation of deceptive content.

References

AI CONNECT. (2023). *Mncai/Fake_or_Real_Competition_Dataset · Datasets at Hugging Face*. Retrieved December 8, 2023, from https://huggingface.co/datasets/mncai/Fake_or_Real_Competition_Dataset

Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., & Alshehri, A. H. (2023). Deep fake detection and classification using error-level analysis and deep learning. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-34629-3

Yonsei University. (2018). *Real and Fake Face Detection*. Retrieved December 8, 2023, from https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detectionb