Project Proposal: Aritifcal Facial Detection, Separating Real Faces from the Artificially Generated

Udhbhav Gupta & Ante Tonkovic-Capin

Problem Statement

Artifically generated image technology has become increasingly sophisticated, sometimes termed as 'Deep Fakes', allowing the creation of realistic videos and images that can deceive viewers. 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.

Motivation

The proliferation of such artificially generated misrepresentations poses a significant threat to society. With the advent of ever more advanced media generation systems like OpenAI's SORA and DALL-E, artificially generated content is becoming more convincing by the day. There may soon come a time where it becomes increasingly challenging to distinguish between real and fake content. By exploring and building an accurate detection model to identify artificially generated facial images, we can explore potential options in how to combat against misleading and maliciously generated content.

Proposed Solution

Our project will follow these steps to research the problem:

1. **Data Collection and Preprocessing**:
    - Gather a diverse dataset containing both real and aritifically generated images of human faces.
    - Preprocess the data by removing noise, resizing images, and ensuring consistent formats.
    - Multiple datasets are publicly available containing sufficient numbers of labeled images, including:
        - https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection
        - https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
        - https://huggingface.co/datasets/InfImagine/FakeImageDataset
        - https://datagen.tech/blog/face-datasets/
2. **Model Selection and Training**:
    - Evaluate various machine learning models (e.g., CNNs, recurrent neural networks) for artificial facial image detection.
    - Train the selected model on the preprocessed dataset.
    - Use techniques like transfer learning to leverage pre-trained models.
3. **Loss Function and Optimization**:
    - Define an appropriate loss function that penalizes misclassifications.

    - Optimize the model using gradient descent or other optimization algorithms.

4. **Evaluation Metrics**:
    - Assess the model's performance using metrics such as accuracy, precision, recall, and F1-score.
    - Use cross-validation to validate the model's robustness.

5. **Post-processing and Confidence Threshold**:
    - Apply post-processing techniques to refine predictions.
    - Determine an optimal confidence threshold for classifying images as real or fake.

6. **Explore Feature Based Detection Approach**:
    - Analyze statistical properties like noise patterns and pixel distributions
    - Analyze texture details of image regions
    - Analyze lighting and shadow patterns through forensic analysis
    - Perform feature extraction using techniques such as Convolutional Neural Networks (CNNs) or pre-trained models like VGG16 or ResNet.
    - Train a classifier on extracted features

7. **Conclusion and Report**:
    - Analyze the results of our attempt at accurately detecting fake images
    - Summarize analyses and results into final submission report.

Timetable

Phase	Duration
Data Collection	2 weeks
Preprocessing	1 week
Model Selection & Training	3 weeks
Evaluation and Testing	1 weeks
Feature based approach	Time permitting
Documentation, Report, Webpage	1 week

Conclusion

By exploring solutions to detecting potential misleading AI generated images, our project aims to contribute to a safer digital environment. We believe that combining modern computer vision techniques with rigorous evaluation can lead to an effective solution to help combat the spread of misinformation and protect the integrity of visual content.

Open Questions

- For learning-based approach, which model and training method will be the most effective?

- For feature-based approach, which features and model should be used and will the approach be effective?
- Will biases in the available datasets affect results?