Project Proposal - Deepfake Facial Imagery Detection

Alex Zhang, Apple Jin, Alex Xiang

1. Motivation

AI-generated and modified imagery technology has found extensive applications across various domains, from generating captivating fake photos for social media engagement to deploying forged imagery to discredit opponents during elections, fabricating academic papers, and even executing extortion schemas[1] [2]. Despite its vast potential, this technology harbors significant adverse consequences. Studies have shown that most Internet users place undue trust in online images and seldom question their authenticity[3]. Additionally, in academic publishing, fake scientific images frequently evade detection[4]. These scenarios underscore the urgent need to develop proficient fake image detection methodologies to match the rapid advancements in AI-driven image generation and modification, enabling timely identification and mitigation of challenges posed.

Given the significant role of applications of AI-generated facial images in a spectrum of applications, from crafting memes for daily amusement to spoofing facial recognition systems and facilitating fraudulent activities, this project specifically focuses on face-centric images. Our objective is to provide an effective method to detect AI-generated fake face images to meet the evolving challenges of digital authenticity.

2. Related works

Using deep learning to detect AI-generated images is not a new topic in the field. The paper by Barni, Mauro, et al. introduces a novel approach by considering cross-band co-occurrences, which could potentially reveal the spectral inconsistencies introduced by GANs in generating synthetic images. It attains a high accuracy of 99.7% and is robust against post-processed images (i.e., Gamma correction, average blurring, resizing, zooming, etc.). However, the method is developed and evaluated primarily based on the images generated by StyleGAN-2, thus its generalization to other image datasets or images generated by other

[1] Hamid, Y., Elyassami, S., Gulzar, Y., Balasaraswathi, V. R., Habuza, T., & Wani, S. (2023). An improvised CNN model for fake image detection. *International Journal of Information Technology*, 15(1), 5-15.

[2] Gu, J., Wang, X., Li, C., Zhao, J., Fu, W., Liang, G., & Qiu, J. (2022). AI-enabled image fraud in scientific publications. *Patterns*, 3(7).

[3] Kasra, M., Shen, C., & O'Brien, J. F. (2018, April). Seeing is believing: How people fail to identify fake images on the web. In Extended abstracts of the 2018 CHI conference on human factors in computing systems (pp. 1-6).

[4] Gu et al., 2022, *Patterns*, 3(7).

algorithms needs further exploration. Another drawback is that the model performance degrades when JPEG compression is applied to the images.[5]

In another informative work, 'An Improvised CNN Model for Fake Image Detection,' Yasir Hamid et al. embark on a detailed exploration of Convolutional Neural Network (CNN) models for face detection, utilizing the CelebA-HQ dataset enhanced with GAN-generated fake images to test five distinct CNN architectures: DenseNet, EfficientNet, MobileNet, ResNet, and VGG16. These models, augmented with layers like averaging pooling, flatten, dense, dropout, and softmax, are evaluated across metrics including accuracy, loss, precision, recall, and F-1 score, demonstrating that all CNNs significantly outperform traditional ML models like KNN, with ResNet achieving the highest accuracy. Notably, ResNet50 reached 100% accuracy on the validation set within 20 epochs through optimizations such as data augmentation and adaptive learning. Despite its high accuracy, the paper acknowledges limitations, such as its heavy reliance on the CelebA-HQ dataset, which may affect generalization, and the need for continual updates to combat evolving deepfake technologies, underscoring the dynamic challenge of maintaining effectiveness in fake image detection.[6]

3. Methodology

Our project will leverage the advanced capabilities of U-net, a ConvNet architecture that excels in image segmentation, and CAM(Class Activation Mapping) to enhance the interpretability of the model's predictions. U-net is particularly adept at detailing the nuances within images, crucial for our task of distinguishing AI-generated content. Its ability to accurately localize regions of interest aligns perfectly with our goal of not just classifying images but also identifying specific AI-generated features. CAM's ability to highlight decision-driving regions of an image allows us to understand and trust the model's reasoning, a vital aspect when distinguishing between AI-generated and human-captured content.

From the course, I will leverage the knowledge of deep learning architectures, focusing on the implementation and fine-tuning of pre-trained CNN models, such as VGG or ResNet, which have shown great success in image classification tasks. To implement and evaluate this methodology, we will use Tensorflow or PyTorch for their support in developing convolutional neural networks, along with OpenCV for preprocessing. These libraries also provide functionalities for data augmentation, which is crucial for ensuring the model generalizes well to unseen data.

Given the challenge of not having true labels for the areas highlighted by U-net, our evaluation strategy will primarily focus on the classification accuracy of the model alongside a

[5] Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2020, December). CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.
[6] Hamid et al., 2023, *International Journal of Information Technology*, 15(1), 5-15.

nuanced qualitative analysis. For the binary classification of images as AI-generated or human-captured, we will rely on standard metrics such as accuracy, precision, and recall. These metrics will provide a clear measure of the model's overall performance in distinguishing between the two categories of images. Qualitatively, the evaluation will shirt towards an in-depth analysis of the areas highlighted by U-Net as indicative of AI generating. This analysis involves comparing the U-Net identified areas with known characteristics and artifacts typical of AI-generated images, such as texture inconsistencies, unnatural edges, or color anomalies. By documenting and categorizing the types of features identified across a broad set of images, we can assess the model's effectiveness in localizing areas of interest.

4. Dataset

We plan to use the 140k Real and Fake Faces dataset found on Kaggle by XHLULU. This dataset includes 70k real faces from Flickr-Faces-HQ (FFHQ), a high-quality image dataset of human faces, and 70k GAN-generated fake faces. Since we are concerned about distinguishing real human face images and fake (either photoshopped or AI-generated images), the dataset helps combine two existing datasets for deep fake image detection purposes. This is a large dataset with 140k images, which provides a solid ground for training the model. All images are in size of 256 pixels x 256 pixels. Below are some sample face images from both real and fake classes.

Sample Fake Images

Sample Real Images

5. Work plan

Week	Task	Details
1-2	Project Planning and Dataset Collection	**Team Meeting:** Define project goals, roles, and milestones. **(Alex Z, Alex Xiang, Apple Jin)** **Dataset Collection:** Collect a balanced dataset of AI-generated and human-taken images. Ensure diversity in the types of images gathered to cover various scenarios and artifacts. **(Alex Z, Alex Xiang, Apple Jin)**
3-4	Data Preprocessing and Model Setup	**Data Preprocessing:** Clean and preprocess the data. This includes resizing images, normalizing pixel values, and augmenting the dataset to increase its diversity. **(Alex Z, Apple Jin)** **Model Setup:** Initialize the U-Net and CAM models using TensorFlow and Keras. Set up the training and validation framework. **(Alex Xiang, Apple Jin)**
5-6	Model Training and Initial Evaluation	**Model Training:** Begin training the models on the collected dataset. Monitor performance and adjust hyperparameters as necessary. **(Alex Z, Alex Xiang)** **Initial Evaluation:** Conduct an initial evaluation using accuracy, precision, and recall metrics for the classification task. Begin qualitative assessments with a subset of images to gauge the effectiveness of U-Net's highlighting.**(Alex Z, Alex Xiang, Apple Jin)**
7	Model Refinement and Extended Evaluation	**Model Refinement:** Refine the models based on initial evaluation feedback. This may involve further tuning of hyperparameters or adjustments to the model architecture. **(Alex Xiang, Apple Jin)** **Extended Evaluation:** Expand the qualitative evaluation to include more images and potentially involve more domain experts and user study participants to validate the effectiveness of highlighted areas.**(Alex Z, Alex Xiang, Apple Jin)**
8	Final Evaluation, Documentation, and Presentation Prep	**Final Evaluation:** Conduct a comprehensive final evaluation of the models, focusing on both quantitative metrics and qualitative feedback. **(Alex Z, Alex Xiang)** **Documentation:** Compile detailed documentation of the project, including the methodology, model architecture, evaluation results, and insights gained.**(Alex Xiang, Apple Jin)** **Presentation Preparation:** Prepare a presentation to showcase the project's objectives, methodology, findings, and implications.**(Alex Z, Apple Jin)**

6. Proposal Team Responsibilities:

- Motivation: Apple Jin
- Related work: Alex Xiang, Apple Jin
- Methodology: Alex Zhang
- Dataset: Alex Xiang
- Work plan: Alex Zhang