Introduction

~~AI content detectors and AI generative models are in a cold war/arms race and stakeholders must adapt by choosing the best models to accommodate the new~~

~~The advent of Deep Neural Networks, along with increased computing and Big Data, gave rise to powerful generative models that can generate media that is near or as indistinguishable from human made media. These synthetic media are major achievements AI engineers have been studying for decades and have come into fruition only in the last decade due to the availability of data and the increased computing power.~~

~~Yet, with the advancements came the downsides of having such access~~

# REDO

Generative AI

Generative technology in recent years has come under scrutiny due to the number of cases where the misuse of the technology has become rampant in cases of fraud, deception, and misinformation. Due to this, countermeasures in detecting AI-generated contents have been deployed but have returned with mixed results in detecting newer models of these technologies, requiring frequent updates to their countermeasures. Akin to an arms race, development in one area will require the other to adapt.

CNNs or Convolutional Neural Networks are the leading countermeasures against generative images, but as stated before, suffer when predicting data from more recent models. This is due to Neural Networks being dependent on the data they are trained on. The

The countermeasures for detecting generative content are mostly in the form of Convolutional Neural Networks (CNN)

An arms race has developed in the area, where an improvement in the other will force the other to also improve, akin to a game of cat and mouse or the nuclear arms race. This in turn has required stakeholders to update their detection technologies to keep up with the rapid advancements of generative technologies frequently.

Common technologies used to generate such content include Generative Adversarial Networks and Diffusion models, with popular examples being StyleGAN and DALLE-3 respectively. Though different in how they generate images, both have the same goal, to generate an image indistinguishable from a real one.

Earlier generative image models suffered from distortions, low resolution and prevalence of artifacts in the image, which in turn made it very easy to recognize that it’s an artificial image, but with the rapid advancements came more impressive results with less to no distortions and high resolution. With this, came the need for systems that can detect such content as they become more indistinguishable from real images.

Technologies that are most promising are a counterpart of generative models, instead of creating an image output, takes in an image instead. These are Convolutional Neural Networks (CNN). Neural Networks that operate by passing an image through a filter to create a map that detects features on an image.

This study hinges on the idea that generated content, images in this case, have features only present or has a pattern commonly found in generated content that can be exploited by Neural Networks, specifically Convolutional Neural Networks.

A common technology of these generative technologies are adversarial networks in the form of General Adversarial Networks (GAN) which pit a Generator against a Discriminator in a game of where the generator must produce samples that will fool the discriminator, in the case where the generator fools the discriminator, the discriminator must update its parameters to match up, and in the case where the generator does not fool the discriminator, the generator must then update its parameters to match up. A sufficiently trained GAN would produce samples that are nearly indistinguishable from a real image but these GANs are limited by their complexity in training, and due to this, most generated images are generated from diffusion models.

Another technology, though more complex, is Diffusion models, with popular examples such as Stable Diffusion and DALLE-3 from OpenAI. These work in tandem with Transformers to turn words into latent representations that the model then understands.

The diffusion, of which it got its name from, is a decoder which generates noise at first then passes through a series of denoisers that “diffuse” the noise into an output

**Statement of the Problem**

The purpose of this study is **to evaluate the performance of common CNN architectures in discriminating against real images from fakes** through training and analysis.

Specifically, this research attempts to give an answer to the following:

1. What are the metrics of performance in detection given the same dataset? In terms of:
   1. Time to Train
   2. Accuracy
   3. Precision
   4. Recall
   5. F1 Score
   6. Confusion Matrix
2. Among the CNN architecture, which has the best performance considering all the metrics?
3. Is there a significant difference in performance on using a Pretrained model over training a model from scratch?

**Hypothesis** H0:

1. Each model will display the same performance over all the measured metrics.
2. Inconclusive in determining the best model among the set.
3. No discernable difference in performance.

**Objectives of the Study**

1. To identify architectural characteristics in distinguishing images.
2. To identify the best architecture in handling the task of being a detector of fake and real images.
3. To determine if the pretrained weights already present in the model affect the performance of the model.

**Scope and Delimitations**

This study focuses on comparing common Convolutional Neural Networks architectures in their ability to distinguish between generative images and real images, in which is a binary classification task. The study will utilize pre-existing datasets sourced on popular repositories for machine learning. This is done to constrain the scope to just training and analyzing the performance of the models. And to further focus on analyzing the performance, its set of image genres are constrained to only one genre, specifically portraits of people. The number of datasets used is adjusted accordingly to accommodate the available computational resources the researcher has.

**Significance of the Study**

The outcomes of this study would be beneficial for the following entities:

To **Social Media Platforms**, the outcomes of this study would be helpful in choosing a model in detecting deceptive generative images.

To **E-commerce**, the result would help in detecting falsified identifications that would result in fraud.

To **Researchers**, this study could be a steppingstone for future research regarding a related subject.

~~Review of Related Literature and Studies~~

~~Related Literature and studies that give context, support and validate the study was reviewed and analyzed in depth in this section. Studies that provide context for the motivations for why this study was undertaken are presented first, then will shift to reviewing literatures regarding attempts in assessing the efficacy of CNNs in detecting AI-generated images.~~

**Review of Related Literature and Studies**

This section provides an in-depth review of related literature and studies that contextualize, support and validate the present research. Beginning with studies establishing the motivations for this work are discussed, followed by explorations of preceding studies on assessing the effectiveness of Convolutional Neural Networks (CNNs) in detecting AI-generated images.

**Generative AI in the Real World**

Generative, multimodal Artificial Intelligence (GenAI; Marchal et al., 2024) has become prevalent in recent years. As high-performance computing hardware becomes consumer-grade, the barrier to AI misuse has lowered significantly.

Marchal et al. (2024) concluded that manipulation of the human likeness –including impersonation, sockpuppeting, and falsification—as the most frequent misuse. Impersonation, the study’s top category, commonly uses synthesized audio and video to replicate a person’s likeness. However, artificially generated images (e.g., profile pictures) emerged as the dominant GenAI misuse overall.

Though majority of misuses are unsophisticated, this presents a precedent of GenAI offering relatively simpler means of conducting such abuses of the tool. This then becomes a problem of identifying such uses, particularly images.

A report by Dash et at. (2024) reported that uses of AI among research circles have increased and expected that at least 34% of research in 2024 has used AI in some ways. Though not a problem on its own, its misuse may cause potential for fraud. AI-generated images have found their way on research journals dating back to the 2020s. This has led to scrutiny of the usage of AI in research.

On a paper on the ethical implications of Image Synthesis. Bendel (2024) reported that the concept of images used for forgeries and spreading false information is nothing new. Examples of image manipulation has existed since the 19th century. Moreso in the 20th and 21st centuries. But with AI-based image generators execute manipulations in mere seconds.

**Detection Studies**

On a literature review by Rana et al. (2022) reviewed detection solutions, in detecting deepfakes, used throughout the industry and categorize it into 4 methods, Machine Learning, Deep Learning, Statistical Measurements and Blockchain. Deep Learning, the most prevalent among the methods, achieved high accuracy as compared to its Machine Learning and Statistical counterparts. Leveraging deep neural networks like CNNs, achieved accuracy of over 90%.

Barni et al. (2020) explored the use of a custom CNN by “Cross-Band Co-occurrences Analysis” by exploiting the fact that GANs, though able to generate images indistinguishable from real images, always create artifacts that are detectable by comparing channels.

Another study utilizing custom CNNs by Zhong et al. (2023) proposed a detection method leveraging the fact that AI-generated images produce distinct artifacts that are not uniform all over the image. It decomposes the image into smaller subsets and rather than passing it over a pooling layer first, it extracts features over a specialized set of kernels.

**Technologies**

Synthesis

Theoretical Framework

Conceptual Framework

Methodologies

~~The study is wholly quantitative in nature due to the nature of Neural Networks in terms of development, training and analysis. And, as such, the study will be structured in the development and training of a model in a set of discrete stages of which being data gathering, data preprocessing, architecture development, training and analysis~~.

The study is quantitative in nature as it analyzes the performance of CNNs on their ability to detect AI-generated images of which is a binary classification task. We analyze the performance of the models by comparing them using frequently used metrics that exist in the space of machine learning, of which are: Accuracy, Precision, Recall, F1 Test, and the Confusion matrix. (geeksforgeeks.com) These tests will take in the number of True positives and negatives and False negatives and positives to produce a digestible result.

**Dataset**

The dataset is procured at Kaggle, which is a database of public datasets for machine learning. The dataset is composed of fake and real human portraits based on the OpenForensics project, which is a dataset of deepfake faces and real faces. Sample size of the training set is 140K, with the validation set at 40K, and the test set at 10K. In actual training, the number of training samples will be downsized to 20K to accommodate the computing constraints.

Each sample is a 3 channel, 256 x 256 image, and will be passed through a preprocessor that will resize the image to 224 x 224, then augment the image by transforming the image with a random flip and rotation before being converted into a tensor. It is then normalized then passed into the model for training or evaluation.

**Hardware and Libraries**

A computer with an AMD Ryzen Zen 5 7600 and an Intel Arc B580 12GB will be used to train and evaluate the models on a Windows 11 OS. The library the model will be built on is PyTorch, due to its rich environment in the Deep Learning space and the availability of pretrained models.

Additional Libraries include:

* Scikit-learn for its k-fold cross validation to prevent overfitting
* Seaborn for visualization of metrics

**Training**

To achieve a fair comparison of performance for all the models, they will be trained on the same dataset, use the same optimizer and use the same batch size. The number of epochs will also be the same.

# EXPAND

**Evaluation and Analysis**

Metrics such as the confusion matrix

References:

Marchal, N., X., R., Elzayn, R., *et al.* (2024). *Generative AI misuse: A taxonomy of tactics and insights from real-world data*. arXiv. <https://doi.org/10.48550/arXiv.2406.13843>

Dash, K. S., Mehta, V., & Kharat, P. (2024). We are entering a new era of problems: AI-generated images in research manuscripts. *Oral Oncology Reports, 10*, 100289. <https://doi.org/10.1016/j.oor.2024.100289>

Remya Revi, K., Vidya, K.R., Wilscy, M. (2021). Detection of Deepfake Images Created Using Generative Adversarial Networks: A Review. In: Palesi, M., Trajkovic, L., Jayakumari, J., Jose, J. (eds) Second International Conference on Networks and Advances in Computational Technologies. Transactions on Computational Science and Computational Intelligence. Springer, Cham. <https://doi.org/10.1007/978-3-030-49500-8_3>

Bendel, Oliver. (2023). Image synthesis from an ethical perspective. AI & SOCIETY. 40. 437-446. 10.1007/s00146-023-01780-4.

Rana, Md & Nobi, Mohammad & Murali, Beddhu & Sung, Andrew. (2022). Deepfake Detection: A Systematic Literature Review. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3154404.

M. Barni, K. Kallas, E. Nowroozi and B. Tondi, "CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis," *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, <https://doi.org/10.48550/arXiv.2007.12909>

Nan Zhong, Yiran Xu, Zhenxing Qian, Xinpeng Zhang, *Rich and Poor Texture Contrast: A Simple yet Effective Approach for AI-generated Image Detection*, 2023, <https://doi.org/10.48550/arXiv.2311.12397>

DATASET REFERENCE  
https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images/data