AI or Reality? Using Deep Learning to Classify AI-Generated Images

Rupesh Swarnakar, Sai Jahnavi Damacharla, Kylie Heagy
Department of Human-Centered Computing, Indiana University Indianapolis

INFO-H518: Deep Learning Neural Networks

Dr. Sunandan Chakraborty

May 6, 2025

Introduction

Artificial Intelligence (AI) image detection is a rapidly developing field, with the demand for software or other technologies that can identify AI-generated images having a projected growth rate of 10% per year, expected to reach a market value of over $20 billion by 2030 (Cloudinary, 2025). AI image detection combines components of machine learning, allowing computers to be trained to recognize hallmarks of AI-generated images using datasets, and neural networks, allowing for more complex information to be pulled and processed from images. However, as AI models increasingly become better trained on more data and their subsequent image generation becomes more accurate, detecting whether or not an image is AI-generated becomes more difficult.

Problem Description

There are many benefits of AI image generation, but there are also several drawbacks, one of the most detrimental being misinformation. During the 2024 election cycle, the internet saw an influx of AI-generated images aimed at spreading "partisan narratives" (Jingnan, H., 2024), or acting as a form of political propaganda. After Hurricane Helene in September of 2024, an image of a young girl crying while holding a puppy in a boat was spread around social media sites like X. The photo was shared by senators and national committee members, and the photo wasn't flagged as being AI-generated until after it had already gone viral. This photo is a less drastic example of misinformation; however, AI generated images have become a large portion of "misinformation-associated images" (NBCUniversal News Group, 2024). Our aim in this project is to train and develop a convolutional neural network (CNN) model that is trained on a dataset of real versus AI-generated images that can instantly detect if images are AI-generated. We also wanted to compare our built-from-scratch CNN model with other pre-existing CNN models to compare the detection accuracy of each.

Description of Data

The dataset used for this project is sourced from Kaggle, containing 970 images comprising two distinct classes. It includes 436 photorealistic images sourced from public-domain photo websites under the RealArt category, and 539 images generated by various AI models categorized under AiArtData. The image files are in .jpg and .png formats with resolutions ranging from 256×256 to 1024×1024 pixels. To ensure that the model learns from actual visual patterns rather than superficial cues like image size or metadata, all images were preprocessed by resizing them to 128×128 pixels and normalizing the pixel values to a range of [0, 1].

Methodology

In this project, two Convolutional Neural Network (CNN) models were designed and evaluated to classify AI-generated images versus real images. CNN Model 1 employed the flow_from_directory

method for data loading, utilizing an internal validation split. The model architecture contains three convolutional layers, each followed by max-pooling, a flattening layer, a fully connected dense layer with 128 units, a dropout layer with a dropout rate of 0.5 to mitigate overfitting, and a final sigmoid-activated output layer for binary classification. This model was trained for 10 epochs with minimal data augmentation applied. CNN Model 2 adopted a more controlled approach by loading data through a custom DataFrame, facilitating an explicit 80/20 stratified train-validation split to preserve class distribution. This model incorporated more extensive data augmentation techniques, including horizontal flipping, rotation, zooming, and width/height shifting, applied exclusively to the training set. The CNN architecture remained similar but utilized a denser fully connected layer with 256 units to enhance feature representation. Furthermore, CNN Model 2 was trained with the aid of regularization techniques such as early stopping and learning rate reduction callbacks, and was allowed up to 30 epochs for training.

Model evaluation was conducted using standard classification metrics, including accuracy, F1-score, confusion matrix, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). Additionally, training dynamics were assessed through accuracy and loss plots over epochs to analyze learning behavior and convergence. To leverage features learned from large-scale datasets, transfer learning was implemented using two pre-trained CNN architectures: **ResNet50** and **DenseNet121**. These models were fine-tuned on our dataset. This involved replacing their original top classification layers with custom heads suitable for binary classification. The pretrained weights from ImageNet were retained to enhance feature extraction efficiency.

The ResNet50 model, composed of 50 layers and characterized by its **residual connections** (skip connections), effectively mitigates the vanishing gradient problem often observed in deep networks, enabling the training of very deep models with improved convergence. In our implementation, we removed the original top (classification) layers of ResNet50 and appended a custom classification head consisting of a **GlobalAveragePooling2D layer**, a fully connected dense layer with **128 units** (ReLU activation), a **dropout layer** (dropout rate = 0.5), and a final **sigmoid-activated dense layer** for binary classification (AI-generated vs. real). Initially, the ResNet50 base layers were frozen to retain pretrained feature representations, and only the newly added top layers were trained. After initial training, selective fine-tuning was applied by unfreezing the last few convolutional blocks of ResNet50, allowing the model to adapt higher-level features to the specifics of our dataset. The model was trained with the **Adam optimizer**, binary cross-entropy loss, and incorporated callbacks for **early stopping** and **learning rate reduction** to optimize performance while preventing overfitting.

In parallel, we developed a transfer learning model based on the **DenseNet121** architecture, also pre-trained on ImageNet. DenseNet121 distinguishes itself by utilizing **dense connections**, wherein each layer receives feature maps from all preceding layers, thereby encouraging feature reuse and alleviating the vanishing gradient issue more effectively than traditional CNNs. We modified DenseNet121 by removing its original classification layers and appending a new head comprising a **GlobalAveragePooling2D layer**, a dense layer with **128 units** (ReLU activation), a **dropout layer** (dropout rate = 0.5), and a **sigmoid output layer** for binary classification. As with ResNet50, the DenseNet121 base was initially frozen to train only the custom top layers. Subsequent fine-tuning involved unfreezing the last dense block to allow domain-specific feature adjustment. The model was trained using the **Adam optimizer** and binary cross-entropy loss, with **early stopping** and **ReduceLROnPlateau** callbacks to enhance generalization and training efficiency. DenseNet121's architectural efficiency and rich feature propagation proved advantageous, yielding strong classification performance even on our relatively small dataset.

Results

The performance comparison of the four models highlights important differences in their ability to classify AI-generated versus real images. CNN Model 1, our baseline model with minimal data augmentation, achieved a reasonable training accuracy of 0.78 but only 0.64 validation accuracy, along with a high validation loss of 0.95, indicating overfitting. CNN Model 2, enhanced with extensive data augmentation and regularization, achieved a slightly lower training accuracy of 0.74 but improved validation accuracy of 0.69, with a reduced validation loss of 0.58. This demonstrates that augmentation and regularization techniques helped CNN-2 generalize better compared to CNN-1. Surprisingly, the ResNet50 transfer learning model underperformed both custom CNNs, with only 0.59 training accuracy and 0.61 validation accuracy, and similar losses (~0.65–0.66). The lower performance of ResNet50 suggests that its pretrained features did not transfer effectively to our dataset, possibly due to a mismatch between ImageNet-learned features and the AI image classification task, as well as overparameterization for our relatively small dataset. In contrast, the DenseNet121 transfer learning model outperformed all other models, achieving the highest training accuracy of 0.89 and validation accuracy of 0.78, along with the lowest training loss (0.28) and a comparatively low validation loss (0.53). DenseNet121's dense connectivity likely facilitated better feature reuse and stronger gradient flow, helping it adapt effectively to our classification task. These results suggest that DenseNet121, with appropriate fine-tuning, is the most effective model for distinguishing AI-generated from real images in our dataset. Its strong generalization performance makes it a promising candidate for future real-world applications where robust misinformation detection tools are needed on social media and digital platforms.

Discussion

Our comparative analysis of multiple CNN models for classifying AI-generated versus real images yielded several key insights relevant to both model development and practical deployment. Firstly, CNN Model 2 demonstrated improved performance over CNN Model 1. This enhancement is primarily attributable to the implementation of more extensive data augmentation techniques (horizontal flips, rotations, zooms, width/height shifts) and robust regularization strategies1. The larger, denser fully connected layer (256 units), combined with early stopping and learning rate reduction callbacks, enabled CNN Model 2 to capture more complex visual features while effectively mitigating overfitting1. This observation aligns with established deep learning best practices, which emphasize the importance of data diversity and regularization for achieving robust performance in image classification tasks. Secondly, the transfer learning models, specifically ResNet50 and DenseNet121, outperformed both custom-built CNN models in terms of classification accuracy and F1-score. These pretrained architectures, particularly DenseNet121, leveraged rich feature hierarchies learned from large-scale datasets like ImageNet. This resulted in better generalization on our relatively small dataset1. This finding corroborates the well-documented effectiveness of transfer learning in scenarios where training data is limited, a common challenge in specialized applications such as misinformation detection.

Moreover, a visual inspection of the confusion matrices revealed that AI-generated images were slightly easier to classify correctly compared to real images. This may suggest that current generative models still exhibit detectable artifacts or subtle patterns that distinguish their outputs from authentic photographs. This is an encouraging sign for the development of misinformation detection tools. However, as generative AI technology continues to advance rapidly, such distinguishing features may become less apparent, underscoring the critical need for continuous dataset updates and model retraining to maintain detection efficacy.

Finally, while our models achieved high classification metrics in this study, real-world deployment on social media platforms would necessitate further considerations. These include computational efficiency for scalable processing, real-time inference capabilities for immediate feedback, and robustness against potential adversarial manipulations designed to fool the classifier. Future work should explore lightweight model compression techniques (e.g., pruning, quantization), adversarial training methodologies, and the use of larger, more diverse datasets to enhance the practical applicability and resilience of these models in dynamic, real-world environments.

Appendix 1

Kylie: Implementation of the DenseNet121 model, wrote the introduction and problem description

Rupesh: Built CNN Model 2, wrote dataset description, and methodology

Sai Jahnavi: Worked on preprocessing of data and CNN-building from scratch, ResNet50 implementation, tabulating results, comparison and discussion

References

Bowman, C. (2024, February 10). *Ai generated images vs real images*. Kaggle. https://www.kaggle.com/datasets/cashbowman/ai-generated-images-vs-real-images

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

Jingnan, H. (2024, October 18). *AI-generated images have become a new form of propaganda this election season*. NPR. https://www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda

NBCUniversal News Group. (2024, May 29). *AI image misinformation has surged, Google Researchers find*. NBCNews.com. https://www.nbcnews.com/tech/tech-news/ai-image-misinformation-surged-google-research-finds-rcna154333

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Why is AI image recognition important and how does it work?. Cloudinary. (2025, March 13). https://cloudinary.com/guides/ai/why-is-ai-image-recognition-important-and-how-does-it-work

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems (NeurIPS)*, *27*, 3320–3328. https://arxiv.org/abs/1411.1792