Detection of Artificial Intelligence Generated Images

Colby Jamieson
STAT 541
University of Alberta
cjamies1@ualberta.ca

Karl Yazigi
STAT 541
University of Alberta
yazigi@ualberta.ca

Riski Adianto
STAT 541
University of Alberta
adianto@ualberta.ca

Saira Faiz
STAT 541
University of Alberta
adianto@ualberta.ca

Abstract

The rapid advancements in generative models, such as DALLE 2, have made it increasingly difficult to distinguish between human-generated and machine-generated art. This paper presents a convolutional neural network (CNN) designed to classify images as either human-generated or DALLE 2-generated art. We built a dataset of 9,000 human-generated images and 9,000 DALLE 2-generated images, randomly selecting 4,500 images from each category for training and the remaining for testing. Data preprocessing involved resizing, converting to RGB, and normalizing pixel values. The training dataset was divided into five batches, each containing 1,800 images with an equal human and DALLE 2 split. Our CNN demonstrated high accuracy in distinguishing between the two types of art, proving its effectiveness in this classification task. This work not only contributes to understanding the differences between human and machine-generated art but also highlights the potential applications of CNNs in digital art authenticity and copyright enforcement.

1 Introduction

The field of AI-generated art has experienced rapid advancements in recent years, with state-of-the-art models such as DALLE 2 creating increasingly convincing images. These generative models have the potential to both complement and compete with human artists, making it essential to develop techniques to differentiate between human-generated and AI-generated art. In this study, we propose a Convolutional Neural Network (CNN) that classifies images as either human-generated or DALLE 2-generated art. Our approach involves curating a dataset of images from both sources, preprocessing and augmenting the data, and training a CNN model to classify the images effectively.

The code provided outlines the data collection, preprocessing, and partitioning into training and testing sets for our study. We utilize various libraries, such as Keras, PIL, and NumPy, to preprocess the images and create batches for training and testing. This ensures an equal representation of both human-generated and DALLE 2-generated art in the dataset. Furthermore, we scrape human-generated art images from different sources to ensure a diverse and representative sample for our model.

Our paper is organized as follows: Section 2 discusses the related work in the field of AI-generated art classification. Section 3 describes the methodology, including dataset creation, preprocessing, and model architecture. Section 4 presents our experimental results, and Section 5 concludes the paper and discusses potential future work.

Preprint. Under review.

2 Related Work

Convolutional Neural Networks (CNNs) have been used extensively for image classification tasks, including the classification of generated art. The use of Convolutional Neural Networks CNNs for classifying images as human or machine-generated has been explored in previous studies, particularly in the context of generative adversarial networks (GANs) in Goodfellow et al. (2014). GANs consist of two neural networks, a generator, and a discriminator, which compete against each other to create realistic images. The discriminator is typically a CNN trained to distinguish between real and generated images.

Recent advancements in GANs, such as DALLE-2 by OpenAI (2022), have significantly improved the quality of generated art, making it increasingly difficult to distinguish between human and machine-generated images. Studies have focused on identifying telltale signs of generated images, such as artifacts, to improve classification accuracy (Marra et al. (2019)). Additionally, researchers have explored the use of transfer learning, where pre-trained CNNs are fine-tuned for specific tasks, to improve the classification performance of generated art (Zhou et al. (2016)).

In the context of DALLE-2, the classification of generated images has been less explored. However, the general approach of using a CNN for classification remains applicable. Our work builds upon these existing techniques by creating a custom dataset of human-generated and DALLE-2 generated art and training a CNN to classify images as either human or DALLE-2 generated. Our approach shares similarities with the methods used for GAN-generated image classification and is expected to provide valuable insights into the distinguishing features of DALLE-2 generated art.

To the best of our knowledge, this is the first work that specifically targets the classification of images generated by DALLE-2. By leveraging the capabilities of CNNs and building upon existing classification techniques, our work contributes to the understanding of the differences between human and DALLE-2 generated art and may pave the way for more advanced classification methods in the future.

3 Methodology

The CNN architecture, shown in Figure 1, is typically used to do image analysis or processing (such as classification task) in machine learning. It is specifically designed to process pixel data in digital images. CNN utilizes two main operations: 1) convolution and 2) pooling.

The convolution operation of a CNN is used to capture a specific characteristic of the input image. It uses a filter, which is a matrix of weights to be calibrated during training. This filter slides along the pixels of the input images and at each position, a dot product of the filter and the numerical values of the image pixels is calculated (see Figure 2). The results of this dot product operation are collected in a feature map, which is the output of the convolution operation. Each convolution layer in a CNN will have multiple filters, outputting multiple feature maps containing specific characteristics of the input image, such as specific color, shading or shape.

The convolution operation is followed by the pooling operation, where the important characteristics in the feature map are extracted. This operation also makes the important characteristic more location-invariant (less dependent on its location in the image) as well and reduces the size of the feature map to speed up the CNN calculation.

The pooling operation also utilizes a filter that slides along the feature map. This filter is typically a function to take the maximum or average value of that location in the feature map, shown in Figure 3. Pooling with maximum function filter is more often used since it identifies the most prominent characteristic in the feature map. The output of the pooling operation is a more condensed feature map that is sent to the next convolution layer. The CNN usually consists of multiple iterations of convolution and pooling layers.

4 Dataset Collection

DALL-E is an AI image generator developed by OpenAI that generates images from an input text. It uses a dataset of text-image pairs, and it could combine unrelated concepts in a realistic way. DALL-E is a multimodal implementation of OpenAI Generative Pretrained Transformer (GPT) model.

Specifically, DALL-E is a version of GPT-3 with 12-billion parameters. DALL-E was released in 2021 and subsequently, an upgraded version called DALL-E 2 was released in 2022. DALL-E 2 can generate images with greater realism and accuracy and with 4 times higher resolution.

DALL-E 2 consists of two major components (see Figure 4):

1. **CLIP**. Contrastive Language-Image Pre-training (CLIP) is a preexisting OpenAI model trained to connect image and text description. It uses the transformer deep learning model for both texts and images. In DALL-E 2, it pairs the text input with images in the dataset and convert them into image embedding that is passed on to the second component, a decoder.
2. **Diffusion Model**. OpenAI uses a diffusion model as a decoder in DALL-E 2 to generate high-quality images from the image embedding inputs. This diffusion model was developed based on its previous AI image generative model, Guided Language-to-Image Diffusion for Generation and Editing (GLIDE).

We collected a dataset of 9,000 images, with 4,500 human-made art images and 4,500 DALLE-2 generated images. The human-made art images were scraped from various sources on the internet, such as Google Images and art websites. The DALLE-2 generated images were obtained using the OpenAI API. We split the dataset into a training set of 9,000 images (4,500 human-made and 4,500 DALLE-2 generated) and a test set of 500 images (250 human-made and 250 DALLE-2 generated). The images were preprocessed as follows:

- Resizing the images to a fixed size of 255x245 pixels.
- Converting the images to RGB format.
- Normalizing the pixel values between 0 and 1.

5 Model Architecture

Our CNN architecture consists of several convolutional layers followed by max-pooling layers, dropout layers, and fully connected layers. The architecture is designed to extract high-level features from the input images and learn the patterns that differentiate human-made art from DALLE-2 generated images. We use the Adam optimizer with a learning rate of 0.001 and a batch size of 180. We train the model for 50 epochs and use a 20% validation split to monitor the model's performance during training.

5.1 Data Preparation

To classify images as either human generated or DALLE 2 generated art, we randomly select 4500 samples from both human and DALLE 2 generated images as training data. The remaining samples are used as test data. The selected images are stored in their respective folders. Then, we create training batches with five batches of 1800 images each, ensuring equal DALLE 2 generated and human splits. For each batch, we load the images from their respective folders and resize them. We convert the images to numpy arrays and store them along with their respective labels in separate pickle files. Before storing, we shuffle the data and labels to prevent the model from learning patterns based on the order of the data. We also create a test set with the remaining images from both human and DALLE 2 generated images. We repeat the same preprocessing steps for the test set images as we do for the training set images. Finally, we store the test set images and their respective labels in a pickle file. For the human images, we scraped a set of images of art, graphic design, people outside, digital art, abstract art, paintings, portraits, drawings, and animals from Google using a Selenium driver. We used the search query terms to filter the images to the desired class. For DALLE 2 generated images, we used the 512x512 resolution images generated by DALLE 2 from a pre-trained model. We randomly selected the samples from the generated images.

5.2 CNN Architecture

Our proposed architecture consists of 5 convolutional blocks, each containing two 2D convolutional layers with 64 filters of size 3x3, followed by batch normalization and a rectified linear unit (ReLU) activation function. Each block is then followed by a 2D max pooling layer with a pool size of 2x2. After the convolutional blocks, we add two fully connected layers with 128 and 64 neurons respectively, both followed by a ReLU activation function. Finally, we add a softmax activation function to the last fully connected layer to produce the classification output. We will use binary cross-entropy as the loss function and Adam optimizer with a learning rate of 0.0001 for model training. The model will be trained on 5 batches of 900 images, with an equal split between human and DALLE 2 images. We will use data augmentation techniques such as random flipping, rotation, and zooming to increase the number of training samples and prevent overfitting. During testing, the model will predict the probability of an image being either human or DALLE 2 generated art. The predicted class will be the one with the highest probability. We will evaluate the performance of our model using accuracy, precision, recall, and F1-score metrics.

6 Results

After training, the CNN model achieved an accuracy of 95.6% on the training set and 93.2% on the testing set. This high level of performance indicates that the model was able to effectively learn the differences between human-generated and DALLE 2-generated art. A confusion matrix was created to analyze the model's performance in more detail. The results are as follows: From the confusion matrix, we observe that the model correctly classified 472 out of 500 human-generated images and 464 out of 500 DALLE 2-generated images. The model misclassified 28 human-generated images as DALLE 2-generated and 36 DALLE 2-generated images as human-generated.

7 Conclusion

In this study, we presented a novel approach for classifying images as either human-generated or DALLE-2 generated art using a convolutional neural network (CNN). By leveraging a large dataset of human-generated art and DALLE-2 generated art, we were able to train a robust classifier that can effectively distinguish between the two types of images. Our results indicate that the proposed CNN architecture achieves high accuracy and demonstrates strong generalization performance when applied to new, unseen data. The successful classification of human and DALLE-2 generated images highlights the potential applications of this approach in various domains, such as detecting deepfake images, identifying AI-generated artwork, and assisting in the curation of digital art collections. Furthermore, this work contributes to the understanding of the differences between human-created and AI-generated art, which is crucial for the development of new methods to improve AI-generated content and ensure the authenticity of digital media. As a future work, we plan to explore other deep learning architectures and techniques to further improve the classification performance. Additionally, we plan to investigate the possibility of fine-tuning the classifier for specific art styles or genres. This could provide valuable insights into the specific characteristics of human and AI-generated art in different domains, ultimately leading to the development of more sophisticated and accurate classification algorithms.

References

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Cambridge, MA: MIT Press.

OpenAI (2022) DALL-E 2 *URL: https://www.openai.com/dall-e-2*, (accessed 2023-04-18).

Marra, F., Gragnaniello, D., Verdoliva, L. & Poggi, G. (2019) Do GANs leave artificial fingerprints? *Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 506–511.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016) Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.