Detection of Artificial Intelligence Generated Images

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

Colby Jamieson
STAT 541
University of Alberta
cjamies1@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.

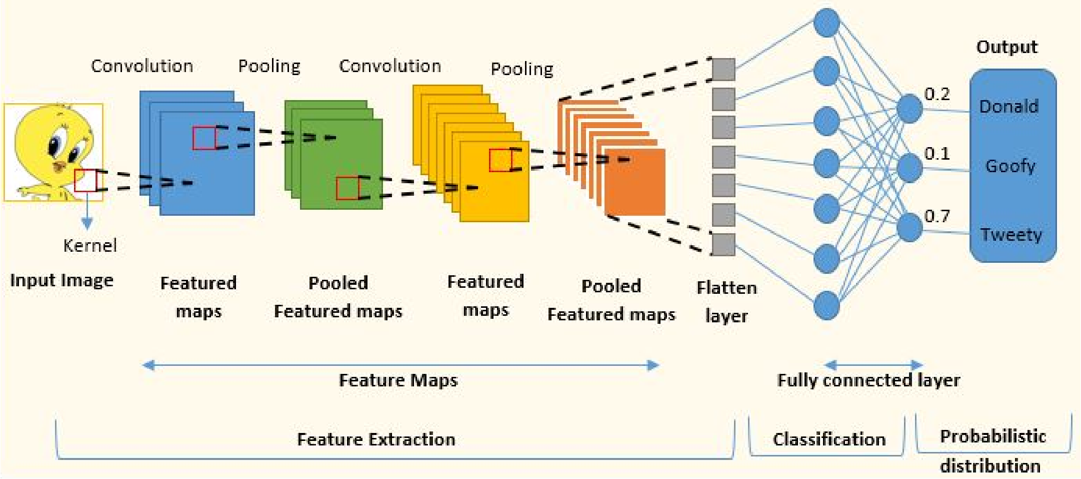

Figure 1: A typical architecture of a Convolutional Neural Network for Image Classification Task. Image by Saily Shah, March 15 2022. https://www.analyticsvidhya.com/blog/2022/01/convolutional-neural-network-an-overview

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 a specialized neural network architecture that is specifically designed to process pixel data in digital images for pattern recognition tasks(O'Shea & Nash (2015)). One unique characteristic of CNN is the organization of its layers to handle three dimensional inputs: height, length and depth of the image. The image depth refers to the color scheme of the input image (e.g. depth of 3: RGB color). CNN utilizes two main operations: 1) convolution and 2) pooling.

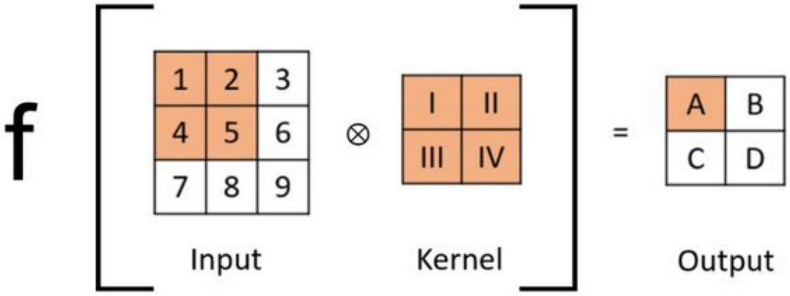

Figure 2: Convolution operation in a CNN. Image by Diego Unzueta, October 18 2022. https://builtin.com/machine-learning/fully-connected-layer

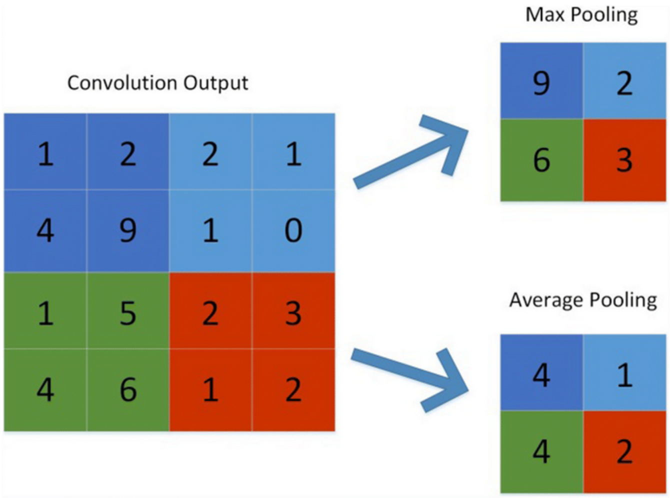

Figure 3: Pooling operation in a CNN. Image by Debasish Kalita, March 3 2022. https://www.analyticsvidhya.com/blog/2022/03/basics-of-cnn-in-deep-learning

The convolution operation of a CNN is used to capture a specific characteristic of the input image. It uses a kernel or filter, which is a matrix of weights to be calibrated during training. This kernel slides along the pixels of input images and at each position, a dot product of the kernel and numerical values of the image pixels is calculated (see Figure 2). An activation function (typically the rectified linear unit (ReLu) function) is then applied to the outputs of this dot product operation, and the results are then collected in a feature map. Feature maps are the outputs 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, which extracts the important characteristics in the feature map are reduces the size and dimensionality of the feature map, which will further reduce learnable parameters and computational complexity of the model. This operation also makes the important characteristic more location-invariant (less dependent on its location in the image).

The pooling operation also utilizes a kernel or filter that slides along the feature map. This kernel is typically a function to take the maximum (max-pooling) or average (average-pooling) value of that location in the feature map. Max-pooling (Figure 3) 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 typically consists of multiple iterations of convolution and pooling layers with fully-connected layers at the very end to produce prediction scores.

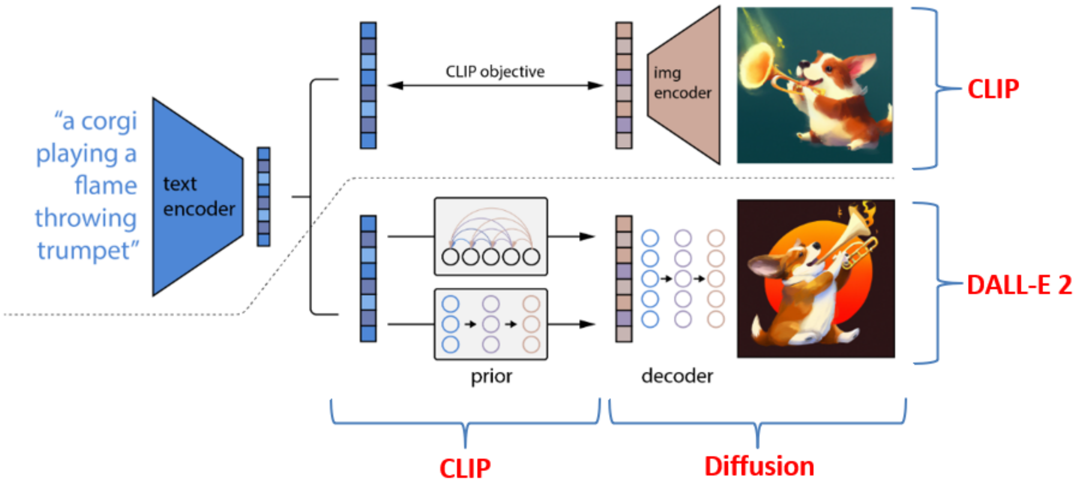

Figure 4: Diagram of OpenAI DALL-E 2 Model. The top diagram shows the training of CLIP to learn joint representation between text and images. The bottom diagram depicts the DALL-E 2 generative process that uses CLIP encoder to generate text-image embeddings that are then used to generate the desired final image using diffusion model. Figure from Ramesh et al. (2022).

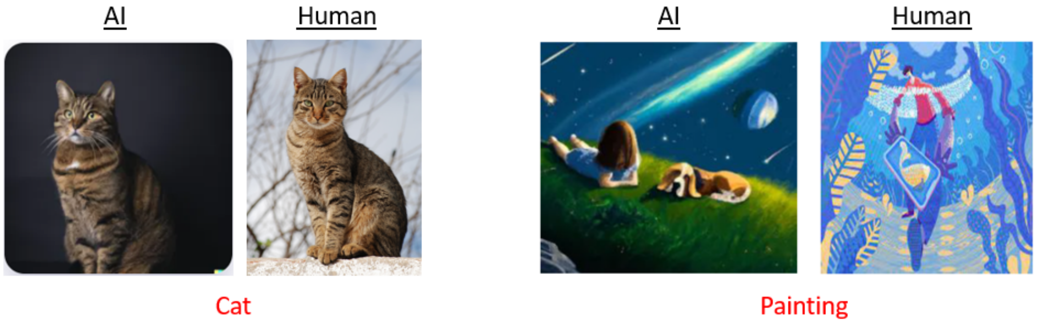

Figure 5: Examples of images used in the model development.

4 Data Collection

DALL-E is an AI image generator developed by OpenAI that generates images from an input text (Ramesh et al. (2021)). 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 (Ramesh et al. (2022)).

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 to similar embeddings (Radford et al. (2021)). It uses a transformer deep learning model for both texts and images. In DALL-E 2, it was used as an encoder to pair the text input with images in the dataset and convert them into text-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 text-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) (Nichol et al. (2022)).

A dataset of 10,580 images were collected for the model development, of which 5200 were human-made images and 5380 were AI generated images. The human-made images were scraped from various sources on the internet using Google Image search. The AI generated images were obtained using from an online DALL-E 2 image database (DALL-E 2 Gallery (2022)). The types of image collected include artistic images (e.g. cartoons, paintings, Japanese animations, digital images) and photorealistic images (e.g. peoples, animals, sceneries) (see Figure 5 for examples).

The images were stored in an RGB format as a three-dimensional numerical array of $245 \times 255 \times 3$, representing the pixel height, length and color depth of the images. Values of the arrays were also normalized to be between 0 and 1.

5 Model Architecture

The CNN model 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. The model uses the Adam optimizer with a learning rate of 0.001 and a batch size of 1800. The model was trained 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, 4500 samples from both human and DALLE 2 generated images were used as training data, while the remaining samples are used as test data. Training batches were created with five batches of 1800 images each, ensuring equal DALLE 2 generated and human splits. For each batch, images are resized to $245 \times 255 \times 3$ pixels and converted to numpy arrays in RGB format.

The images were cropped to exclude the watermarks to avoid the model learning from irrelevant data. Additionally, data and labels are shuffled to prevent the model from learning patterns based on the order of the data. Human-generated images were scraped from Google using a Selenium driver and included art, graphic designs, digital art, abstract art, paintings, portraits and other drawings. DALLE 2 images were randomly selected from DALL-E 2 Gallery (2022).

5.2 CNN

The CNN model architecture consists of 5 convolutional blocks, each containing two 2D convolutional layers with 64 filters of size 3×3, followed by batch normalization and a ReLU 3 activation function. Each block is then followed by a 2D max pooling layer with a pool size of 2×2. After the convolutional blocks, two fully connected layers are added with 128 and 64 neurons respectively, both followed by a ReLU activation function. Finally, a softmax activation function is used in 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 0.00005 for model training. The model will be trained on 5 batches of 1800 images and tested on 880 DALLE 2 and 700 human-generated images. Data augmentation techniques such as random flipping, rotation, and zooming are used to increase the number of training samples and prevent overfitting. The model uses 7.5 million parameters during training. During testing, the model predicts the probability of an image being either human or DALLE 2 generated art. Prediction output will be between 0 and 1, with an output above 0.5 predicting the image to be DALEE 2 generated, and below 0.5 as human-generated. We will evaluate the performance of our model using accuracy, and a receiver operating characteristic curve (ROC).

6 Results

After training the model for a combined 50 epochs, the CNN model achieved an accuracy of 84% on the training set and 93% on the testing set. Training and testing accuracy are nearly identical until around 20 epochs. At this point the results diverge, with training accuracy continuing to increase with further training, and validation accuracy oscillating in the neighbourhood of 80% accuracy. The

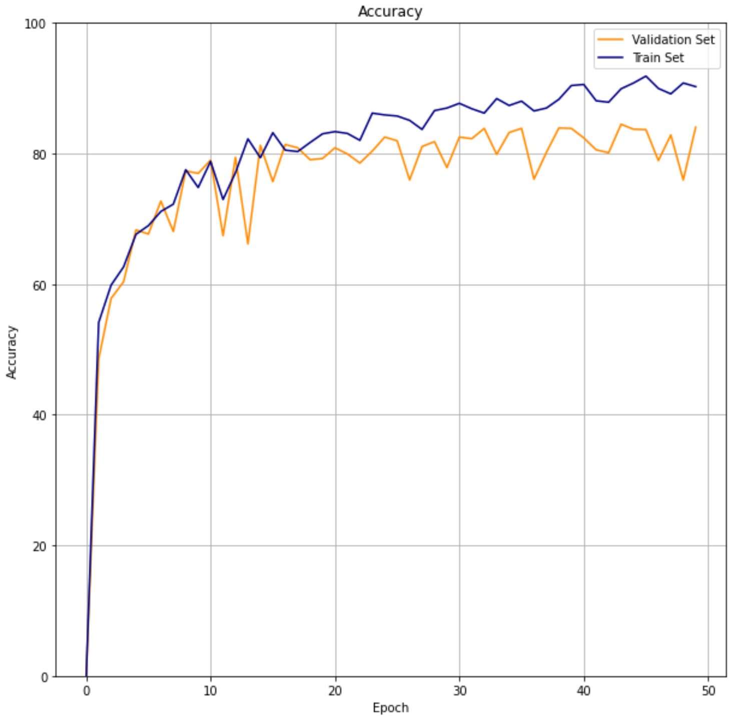

Figure 6: The CNN model reaches peak accuracy of around 80% on the testing data.

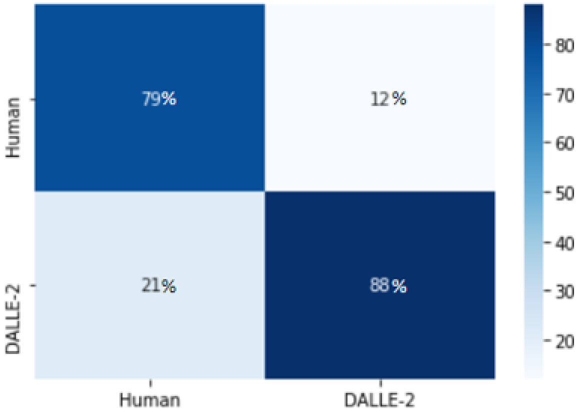

Figure 7: The confusion matrix indicates the percentage of images labelled correctly by image type.

divergence of these results provides evidence of over-fitting the training data. However, validation accuracy does not diminish as a result of this over-fitting so there does not seem to be a risk of accuracy loss from over-training the model.

Of the 880 DALLE-2 images tested, the model correctly labelled 780 (88%). Of the 700 human-generated images, 550 were correctly labelled (79%). This suggests that the CNN model is better at correctly classifying DALLE-2 generated images than those created by a human. One reason for this may be that the DALLE-2 images have specific and relatively consistent markers that make them easier to classify.

When analyzing the trade-off between model sensitivity and specificity, an analysis of the true positive and false positive rate was performed using an ROC curve. True positives are DALLE-2 images classified correctly, while false positives are human-generated images that were misclassified. The ROC curve plots the true positive rate on the y-axis and false positive rate on the x-axis. A one-to-

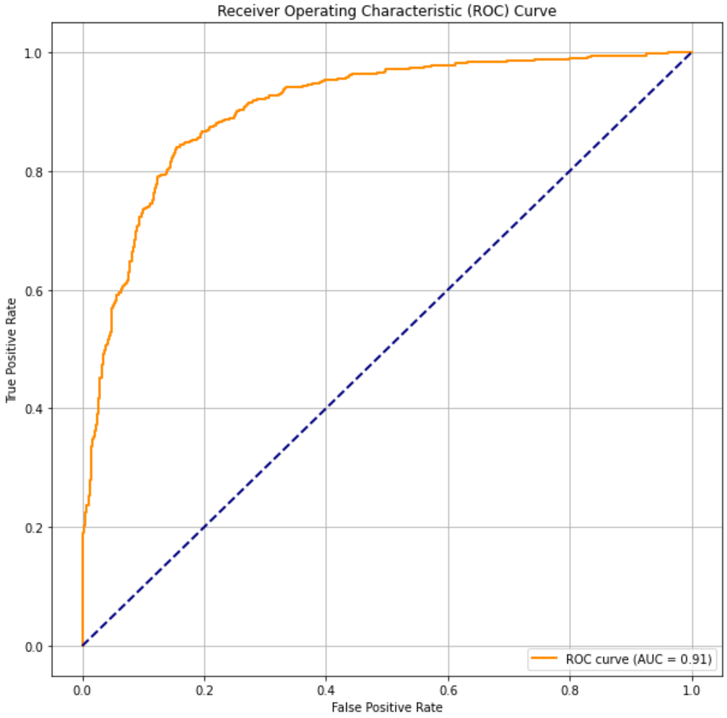

Figure 8: The CNN model scored an AUC of 0.91, indicating a nearly ideal trade-off between true and false positives.

one trade-off, shown in the graph below as a blue dashed line, indicates a random, and non-effective classifier. A concave ROC curve that takes up more area indicates a better performing classifier. This is because the true positive rate increases at a higher rate than false positives, indicating a high performing model. The area under the curve (AUC) score for the model was 0.91, with an AUC score of one being perfect, and 0.5 being a random classifier.

Although there is room for improvement, the level of performance from testing indicates that the model was able to effectively learn the differences between human-generated and DALLE 2-generated art.

7 Prediction Examples

Figure 9 shows prediction examples for various human and DALLE 2-generated art. The photos in the left column have output scores that indicate the model was relatively confident in both its correct and incorrect labeling. The photos in the right column show images that the model had more difficulty identifying. These example images may hold clues to what makes a either class of image distinctive from each other.

8 Conclusion

This study proposes a novel approach for classifying images as either human-generated or DALLE-2 generated art using a CNN. By leveraging a large dataset of human-generated art and DALLE-2 generated art, a robust classifier was trained that can effectively distinguish between the two types of images. Results indicate that the proposed CNN architecture achieves high accuracy and demonstrates strong generalization performance when applied to testing 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

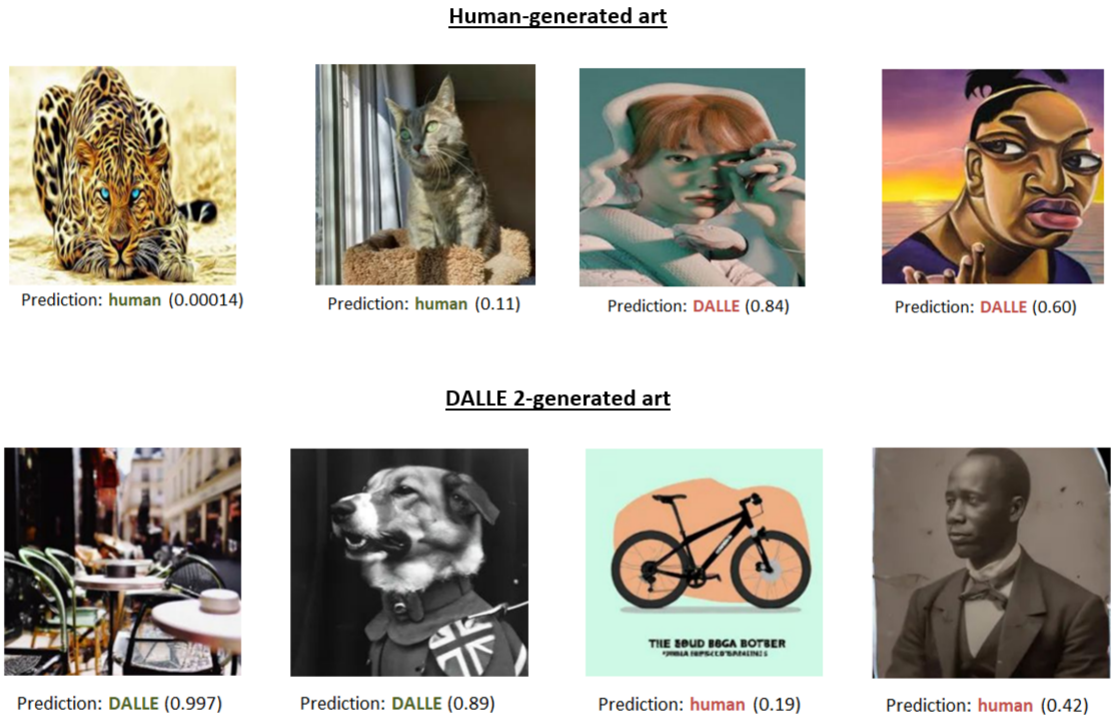

Figure 9: Model predictions for DALLE 2-generated art by relative confidence.

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. Future work will explore other deep learning architectures and techniques to further improve classification performance.

The model could also be fine-tuned 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.

O'shea, K. & Nash, R. (2015) An Introduction to Convolutional Neural Networks. *ArXiv:1511.08458*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021) Zero-Shot Text-to-Image Generation. *ArXiv:2102.12092v2*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022) Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv:2204.06125*.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021) Learning Transferable Visual Models From Natural Language Supervision. *ArXiv:2103.00020*.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022) GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *ArXiv:2112.10741*.

contact@dalle2.gallery (2022) DALL-E 2 Image Database *URL: https://dalle2.gallery/#search*, (accessed 2023-04-03).