Exploring Generative AI: What Makes An Image Steerable?

Paige Blum
Colgate University
13 Oak Drive
Hamilton, NY 13346
pblum@colgate.edu

Abstract

This study investigates the predictability of image steerability using convolutional neural networks and linear regression models. We based our approach on the study "ArtWhisperer: A Dataset for Characterizing Human-AI Interactions in Artistic Creations" (1), using their dataset as a starting point for our work. We integrated feature extraction and regression techniques while utilizing R-squared and mean-squared error metrics for evaluation. Our results show that the models have similar performance, with both identifying images being either real or generated by AI as being most influential in predicting steerability, giving us insights into areas where text-to-image model's performance could be improved.

1 Introduction

What makes an image steerable? Image steerability refers to the ability to guide or direct the generation process of a text-to-image model in creating specific visual outputs based on detailed prompts. Improving the steering process of AI can drastically reduce the time and resources used in creating custom visual content. Within the context of advertising and media, generative AI has the potential to allow clients to create personalized campaigns within minutes. Additionally, generative text-to-image models have the potential to open up a new world for creatives- allowing individuals to manipulate and produce artistic creations. Enhancing our understanding of steerability is crucial for advancing AI's interactive capabilities and the user experience of these tools. If you have ever played around with DALL-E you may have noticed that it can execute certain prompts better than others. A high steerability score is achieved when the AI can accurately follow detailed and specific instructions, producing an image that matches the user's input in a short number of prompts. Whereas a low score results from outputs that are unexpected by the user and a large number of prompts to achieve the desired image.

Can convolutional neural networks or linear regression models learn to predict steerability? To better understand the effects of steerability, we investigated how different modeling approaches can predict the steerability of AI-generated images. With a growing need for AI systems to produce highly customizable content efficiently, we hope to uncover features that enhance or reduce steerability. We obtained the dataset used in ArtWhisperer from the author's Hugging Face and added a steerability score column by running scripts from the author's Github repository. We assume that the data represents typical scenarios where AI models are applied and that these models were deployed in stable computer environments. The ArtWhisperer dataset was used to train a CNN and linear regression model. Each model processed target images with corresponding steerability scores to learn how to predict; the CNN directly analyzes a raw image to find features, while the linear regression model utilizes pre-defined discrete features such as whether the image depicts nature or has people in it. This approach allowed us to evaluate the performance of each model in predicting how well an image's characteristics inform the algorithms' steering capabilities.

2 Background

Three papers were quite informative in guiding the project. The first paper that inspired this study directly is "ArtWhisperer: A Dataset for Characterizing Human-AI Interactions in Artistic Creations" (1). In this study, developers created an interactive game called ArtWhisperer, which had players prompt a GAN to generate images similar to a given target. They were able to collect over 51,000 interactions documenting how participants were able to steer AI. They discovered that players made small changes to their prompts. Additionally, they defined a new model of measuring called steerability that is calculated using the stopping time of an empirical Markov model. We utilized the public dataset this study generated as a basis for our work. We also ran their scripts to generate steerability scores to train our models.

Another related work is "Steerable AI-powered Art-making Tools", a dissertation that explores AI's role in art-making (2). The study outlines existing tools and presents three new art-making tools called Artinter, TaleBrush, and Promptpaint. The findings indicate that multimodal interactions enhance the usability of AI. Multimodal interactions include integrated data from different sources or sensory inputs to perform tasks. This informed our research in using two different models to predict steerability.

Lastly, we drew off of the conclusions of "On The “Steerability” of Generative Adversarial Networks" to inform our research (3). This paper provides an in-depth overview of the evolution of deep generative models, specifically Generative Adversarial Networks. The authors demonstrated that discrepancies in GAN's ability to generate expected outputs are due to variability present in training data. The paper suggests that data augmentation and optimizing generator weights would bolster the performance of GAN steerability. This research influenced the conclusions of our study as well as the discussion section of this paper. We can better understand the results of this study by using their findings as a benchmark for future research.

3 Methods

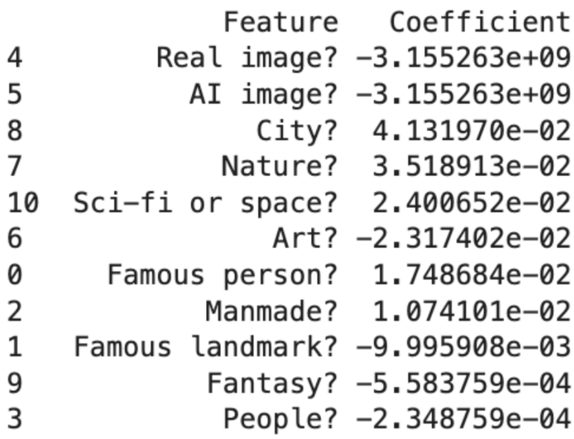
This study employs the ArtWhisperer dataset, which contains images with associated steerability. This dataset was crucial for training and testing the models on their ability to predict image steerability. We utilized a linear regression model to assess how well predefined discrete features extracted from the images can predict steerability. We additionally used a convolutional neural network to see how the results would change from analyzing raw images to find relevant features rather than using discrete features. We evaluated both of the models through feature extraction and using statistical measures to quantify how well each model predicts the steerability of images. A link to the code repository can be found here: $https://github.com/ekociubes/Generative - AI - Art - and - Steerability/$.

3.1 Datasets

We used the ArtWhisperer dataset. For more information regarding the creation of the dataset refer to the Background section of this paper. After obtaining the original dataset, we ran a steerability.py script to generate steerability scores for each target image. We then added the steerability scores as a new column to the original dataset and made sure there was only one entry for each target id. Depending on the model, we used specific columns of the dataset to train the model.

3.2 Models

We trained a linear regression model on discrete features of an image and a given steerability score to predict steerability. We also trained a convolutional neural network on target images and steerability scores to predict scores from raw images. The images were transformed to 128 by 128 pixels and normalized. The model was trained on 10 epochs with an Adam optimizer and a learning rate of 0.001.



```
              Feature   Coefficient
4         Real image? -3.155263e+09
5           AI image? -3.155263e+09
8               City?  4.131970e-02
7             Nature?  3.518913e-02
10  Sci-fi or space?  2.400652e-02
6                Art? -2.317402e-02
0      Famous person?  1.748684e-02
2           Manmade?  1.074101e-02
1   Famous landmark? -9.995908e-03
9            Fantasy? -5.583759e-04
3             People? -2.348759e-04
```

Figure 1: Linear Regression: Feature Importance

3.3 Evaluation

Mean squared error (MSE) was used to evaluate both models. We also utilized R-squared score to evaluate our linear regression model. The lower the score, the more accurate the model is performing. Furthermore, we implemented feature extraction by using the weights the linear regression model learned as a measure of feature importance. We used the absolute value of the weight to handle negative values. As well we performed error analysis by feature for both model types by taking the mean squared error of all predictions with ones under the specified feature. Figure 2 and Figure 3 show the mean squared error of all the images that contain a given individual feature.

4 Results

Both the linear regression and convolutional neural network models performed well. The linear regression model had a MSE of 0.002 and R-squared of 0.14. The convolutional neural network had a MSE of under 0.0001. Based off of these findings, we can conclude that both models performed very well at predicting steerability. After training a linear regression model to predict the steerability of each image using discsrete boolean features, we extracted the weights learned to measure feature importance. The model weighted real and AI images being features negatively and most strongly correlated in predicting a target image's steerability. You can refer to figure 1 for more information on the ranking of feature importance. In regards to error analysis by feature, the two models had very similar distributions across the subsets of images with each feature. We found that the highest error was in predicting the steerability of real target images. Features such as art, nature, fantasy, and AI images had higher accuracy in predictions.

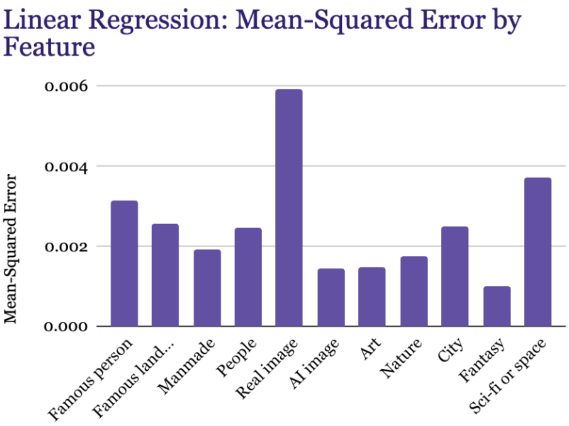


Figure 2: Linear Regression: Mean-Squared Error by Feature

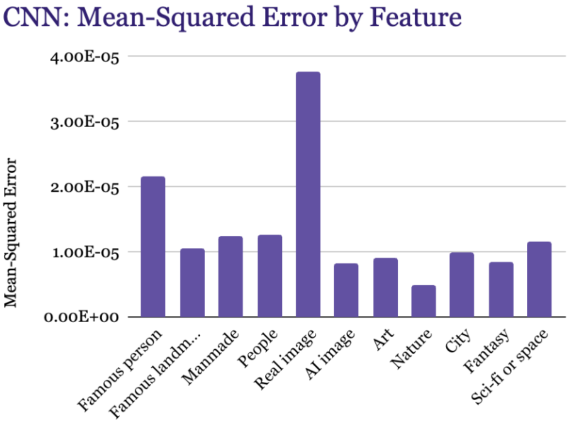


Figure 3: CNN: Mean-Squared Error by Feature

5 Discussion

Both of the models we created performed very well in predicting the steerability of a text-to-image model. We found that the convolutional neural network performed better than the linear regression model at predicting steerability. The CNN has the ability to discover new features that may not be explicitly encoded in the linear regression model, speaking to its lower mean squared error. We most notably found that an image being distinguished as real or created by artificial intelligence to have the largest importance in predicting steerability. It is important to note that these features negatively impacted steerability and that the error analysis demonstrated high error within these features. The negative instances in feature importance give us hints as to where the image-to-text model utilized in ArtWhisperer can be improved.

We came to a few conclusion that related to our background research. First off, our results directly match the results of ArtWhisperer. They discovered that images of cities and the natural world were most steerable than fantasy and artistic images. Our results indicated the same: the nature feature has a weight of .035, the city feature has a weight of .041, the art feature have a negative weight of -.023 and the fantasy feature has a weight of -.056. Second, we can also hypothesize that features that impacted low steerability scores could be due to variability in training data branching off of the results found in "On The “Steerability” of Generative Adversarial Networks"(3).

Some limitations on our work include that while convolutional neural networks are good at finding patterns and complex features in data, we did not have the time or resources to determine why certain features affect steerability or the presence of any new ones that had a significant impact. Our next steps for this study would be to start to determine if there are other features that exist that affect steerability and what those features are. We could start this by consulting with a domain expert. Furthermore we would implement feature extraction within our convolutional neural network using techniques like random forest models. Another interesting approach we were considering is having human subjects interact with the images so we could gauge their perceptions in real time. It could also be interesting to see what areas of the brain are active during this decision making process.

In conclusion, this study demonstrates the power of machine learning in predicting the steerability of images in a text-to-image generation context. Both models displayed robust performance. Our findings not only align with prior research, but also give us clues as to where these generative models can grow highlighting the specific features that influence steerability. Despite these promising results, there is still room for a lot of improvement within this study and a need to uncover additional information that significantly affects steerability. This work lays the groundwork for improving the functionality of generative AI tools within the context of image generation, which is increasingly becoming used by the public.

Acknowledgements

The stylistic parts of this template were taken from ACL 2023 Latex template.

References

[1] Anonymous authors. Artwhisperer: A dataset for characterizing human-ai interactions in artistic creations, 2024. Under review as a conference paper at ICLR 2024.

[2] J. Chung. *Steerable AI-powered art-making tools*. Dissertation, Michigan University, 2023.

[3] A. Jahanian, L. Chai, and P. Isola. On the “steerability” of generative adversarial networks. Technical report, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, n.d.