A Survey of Interpretable Generative ML Tools & Techniques

Chris Williams
Clemson University
Clemson, SC
cwill47@g.clemson.edu

Abstract

This final project paper surveys state-of-the-art tools and techniques for investigating the interpretability of neural network models, explicitly focusing on generative models including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion models. Through experimentation with both GUI-based and API-based tools, we evaluate their effectiveness in interpreting model behaviors and decisions.

1. Introduction

Understanding the mechanisms behind generative machine learning is crucial for encouraging trust and acceptance of model outputs, understanding and limiting biases in trained models, and enabling improved performance, usability, and user interaction.

This project report surveys recent tools, techniques, and case studies that aim to capture and improve the **interpretability** of generative machine learning models, specifically related to image generation.

1.1. Background

Interpretability refers to how a model works, including its inner workings and how each component of input to the model contributes to the model's output. It is closely related to another term commonly used when describing how humans might understand machine learning models: explainability. While explainability focuses on why a model makes a prediction, interpretability emphasizes how a model makes a prediction.

1.1.1 Importance of interpretability

Interpretability is an important area of study for several reasons:

- Trust and acceptance: users and application developers need to trust the models that they are using, especially in applications that use sensitive data
- Model improvement: Understanding how a model decides and generates its outputs can help model developers identify flaws in the model architecture and possible deficiencies in the training data.
- Ethical considerations: understanding the decisions and outputs of a model can help identify potential biases and harmful outputs
- Improved usability: when it is clear how a model processes data, interprets prompts, and generates output (such as an image in the case of a generative model), users can craft more effective inputs for the model in question

1.1.2 Previous focus on non-generative models

Earlier work in the development of toolsets for model interpretability have been more focused on non-generative models for several reasons:

- Generative models involve more complex computations; non-generative models have a more understandable relationship between input and outputs
- Earlier deep learning successes were in the areas of classification and object detection, therefore tools and frameworks related to interpretability tended to be directed to those areas
- Interpretability for non-generative models is typically more straightforward since model correctness is often easier to explain for those use cases. Generative models for images tend to have subjective definitions regarding quality and correctness.

1.2. Research Objectives

The widespread adoption of various image generation models in the past few years, such as DALL-E[13] and Stable Diffusion[14], has made these models increasingly important in our daily lives, as they are an integral part of critical applications in society today.

For this reason, I chose this research area for my final project. There are unique challenges regarding the interpretation of generative models, and this is a rapidly evolving and fascinating field of study that will continue to gain importance.

The research questions I intend to address concern the current methods for the interpretability of generative models such as GANs, VAEs, and Diffusion Models. Additionally, I plan to examine related areas of work and identify potential future research for myself and others

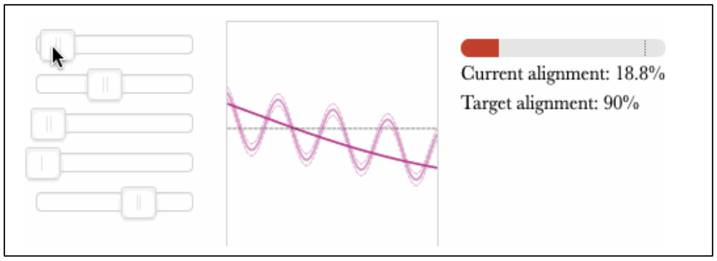


Figure 1: Interface for users to interactively manipulate features that affect a model.

2. Related Work

2.1. Interactive Reconstruction

One method of research involving the interpretability of models includes ***qualitative analysis***, which involves examining actual human perceptions of a model to gauge their understanding of the underlying model, thereby providing a more comprehensive understanding of the relationship between model characteristics and the perceived interpretability of the model. One such study [15], introduced the concept of interactive reconstruction, whereby users changed settings in a tool interface that manipulated the internal workings of a model to reconstruct a targeted model output as shown in Figure 1. Compared to quantitative analysis of the models collected via methods not requiring Human-Computer Interaction (HCI), their study confirmed that the qualitative and quantitative results agreed.

2.2. Mechanistic Interpretability

Mechanistic interpretability is a growing field of AI study that, like other types of interpretability research, relates to exploring and understanding the internals of neural networks. Rather than focus on the relationships between the models' inputs and outputs, mechanistic interpretability focuses on a more low-level understanding related to breaking down the computations involved into understandable components within the networks.

One of the biggest challenges with mechanistic interpretability is that with the continual growth and complexity of neural networks that generate images, the ability to reverse-engineer the complex computations within the networks is becoming increasingly difficult.

One group of researchers [2] have proposed a method for mechanistic interpretability for identifying layers within text-to-image models. Their research includes developing a tool that allows them to "evaluate the efficacy of closed-form model editing across a range of text-to-image models." This work is emerging and ongoing and promises to overcome some of the difficulties inherent in previous studies related to mechanistic interpretability.

3. Research Approach

The methodology for this survey focused on reviewing and analyzing recent developments in generative model interpretability. The focus was primarily on GANs (Generative Adversarial Networks), VAEs (Variational Autoencoders), and Diffusion Models, with particular emphasis on generation and manipulation of images.

Primary sources of data were research papers, with most available via arXiv.org. Focus was primarily on research papers published in the last 3 to 4 years since this is a rapidly evolving field. References to some of the underlying technologies were also consulted, and those have been included in the References section of this report.

In addition to the research papers, this survey included reviewing tools and frameworks, primarily open-source projects, that provided interpretability analysis for generative models. The tools and techniques were assessed based on:

- Support for generative models
- Maturity of product
- Ease of implementation
- Quality of analysis and output

Several well-designed tools were reviewed initially and not included in the final reviews because they were designed to be used with non-generative models. The tools selected to be reviewed in-depth will be detailed in the Experiments section. The goal of these reviews will be to explore the tools in the context of reviewing the model interpretability while reviewing and potentially improving the performance of the model in question.

Due to time constraints, research efforts will be focused on experimentation with a Variational Autoencoder (VAE). VAEs are particularly useful in this type of research because they learn meaningful latent representations and enable controlled generation through latent space manipulation, which allows systematic testing by isolating and modifying individual factors. This control helps map the relationship between latent representations and model outputs, making it easier for the interpretability tools being studied.

4. Experiments

This research involved testing API-based tools that required software development and GUI-based solutions that we applicable for visually exploring models. The goal was to use the guidelines set forth in the Research Approach and assess how effective the tools were at aiding in the task of model interpretability for generative models.

As this API-based vs. GUI-based solution led to different experiences, a key question was which approaches provide the most valuable insights into the internals of the analyzed models.

For the sake of brevity, the Experiments section of this document only details the tools our research found to be the most promising. A complete list of the tools investigated will be detailed on the project website at https://chriswil.github.io/interpret-ml-tools/, as well as more in-depth details about the tool findings.

4.1. GAN Paint (GUI)

The GANPaint image editing tool (https://ganpaint-demo.vizhub.ai/), shown in Figure 2, was developed as part of the research for the GAN Dissection paper [8]. This tool provides an interactive interface for interpreting and manipulating the internal representations learned by GANs.

Using a web-based interface, users of GANPaint can directly manipulate neuron activations within the GAN while working with several pre-loaded images. Users can add and remove objects and scene features such as trees, doors, and clouds. Using the research paper as a guide, we experimented with this tool to understand how GANs learn to recognize and manipulate features and encode spatial relationships between objects (for instance, it was understandably challenging to draw a door in the sky).

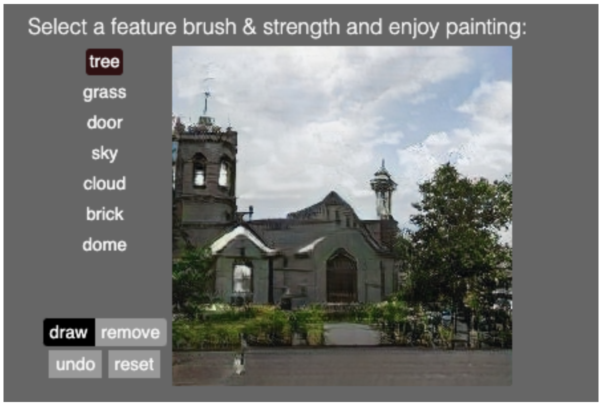
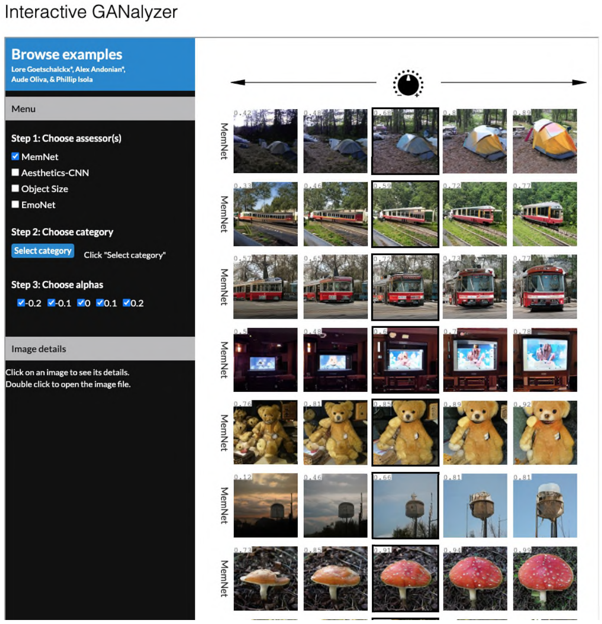

Select a feature brush & strength and enjoy painting:

Figure 2: GAN Paint tool

4.2. GANalyzer (GUI)

The Interactive GANalyzer tool (http://ganalyze.csail.mit.edu/), shown in Figure 3, was developed as part of the GANalyze paper [9]. This paper introduces a framework that uses GANs "to study cognitive properties like memorability, aesthetics, and emotional valence." The authors leverage pre-existing CNN-based models to assess image memorability and aesthetics and develop a custom model by fine-tuning a ResNet50 model on the Cornell Emotion6 Image Database.


Interactive GANalyzer

Browse examples
Lore Goetschalckx*, Alex Andonian*, Aude Oliva, & Phillip Isola

Menu

Step 1: Choose assessor(s)
☑ MemNet
□ Aesthetics-CNN
□ Object Size
□ EmoNet

Step 2: Choose category
Select category Click "Select category"

Step 3: Choose alphas
☑-0.2 ☑-0.1 ☑0 ☑0.1 ☑0.2

Image details

Click on an image to see its details.
Double click to open the image file.

Figure 3: Interactive GANalyzer tool

The authors acknowledge that memorability, aesthetics, and emotional valence are complex concepts that are difficult to describe. Rather than defining these properties definitively, they proposed exploring these characteristics with the GANalyzer tool.

Our experiments with this tool tended to confirm the findings that the authors noted in their research. For instance, larger objects and images with centered subjects were often associated with higher memorability. Our experiments indicated that this tool doesn't provide any interpretability functions directly. Still, its ability to predict complex and subjective properties gives some insight into a model's decision-making process. That insight can increase our understanding of how these models work and what visual features are important.

4.3. AI Explainability 360 (API)

AI Explainability 360 (AIX360)[18] is an open-source Python library developed by IBM that provides algorithms for interpretability and explainability. As with most of the other tools examined during our research, AIX360 was initially designed for traditional, non-generative machine learning models.

In order to explore our VAE, we implemented the following algorithms:

- ProtoDash explainer, which identifies prototype examples that are most representative of the learned features. This provides insights into what patterns the model has captured.
- Contrastive Explanations Method (CEM), which helps identify minimal changes required in the latent space to transform one type of maritime image into another.

4.4. Alibi (API)

Alibi[17] is an open-source Python library developed by Seldon that provides a set of algorithms for analyzing machine learning model interpretability. The library offers a unified API that allows developers to work with different functions within the library in a consistent manner. While Alibi has been primarily built for supporting non-generative model types, the support for generative models, including models related to image generation, is growing.

We experimented with Alibi using a basic VAE that we implemented to work with custom training and test datasets. The Maritime-related dataset was generated using a Stable Diffusion v2.1 pipeline. Using a prompt of 'photo realistic x' where x = {boat, lighthouse}.

Figure 4: Samples from our custom maritime dataset

We experimented with Alibi's API and its included analysis features:

- Integrated Gradients explainer, to analyze feature importance to identify parts of the input images that have the most influence in the reconstruction process
- Anchor Image explainer, to understand decision rules which can show patterns in the images that result in better or worse reconstructions
- Accumulated Local Effects (ALE) to understand feature interactions which can show how different image features interact to influence the reconstruction process.

Our results indicated Alibi's strengths in well-defined interpretability and explanation algorithms and general ease of use as a software framework. Our research confirms the consensus in the literature that Alibi is one of the most production-ready frameworks.

5. Conclusions & Findings

5.1. Preferred Tool

After research and testing, our research indicates that Alibi is currently our best choice tool for generative model interpretability analysis, particularly for more mature, production environments. It has a comprehensive and growing feature set, unified API design for ease of use across the available algorithms, and strong community support.

5.2. API vs. GUI

One goal of this research project was a comparative look at the GUI tools for examining model interpretability versus products that provide an API and require additional software development. Our research uncovered advantages for both GUI and API-based approaches. While GUI tools provide immediate feedback and a lower technical barrier to entry, the API-based tools offered customizable analysis as well as more detailed analysis.

5.3. Future Research

As the field of generative models continues to mature, it will be important to continue monitoring this area of interpretability tools for those types of models. The tools chosen in our research for additional software development are being actively maintained and generative AI appears to be a focus in their efforts.

The downsides of steeper learning curve and limited interactive capabilities with the API-based solutions can potentially be aided by future research into hybrid solutions, combining the flexibility of using an API, while adding the improved accessibility of a GUI.

References

[1] Alber, M., S. Lapuschkin, P. Seegerer, M. Hagele, K. T. Schutt, G. Montavon, W. Samek, and K.-R. Muller, "iNNvestigate neural networks!" arXiv:1808.04260, 2018.

[2] Arya, V., R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models," in Journal of Machine Learning Research 21, 2020.

[3] Basu, S., K. Rezaei, P. Kattakinda, R. Rossi, C. Zhao, V. Morariu, V. Manjunatha, and S. Feizi, "On Mechanistic Knowledge Localization in Text-to-Image Generative Models." arXiv:2405.01008, 2024.

[4] Bau, D., J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba, "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks," in Proceedings of the National Academy of Sciences, Sep 2020. arXiv:1811.10597, 2020.

[5] Bereska, L. and E. Gavves, "Mechanistic Interpretability for AI Safety -- A Review," Under review as submission to TMLR. arXiv:2404.14082, 2024.

[6] Bergmann, D. and C. Stryker, "What is a variational autoencoder?" IBM Think, Jun. 2024.

[7] Dandolo, D., M. Chiaraa, M. Carlettib, D. D. Pezzeb, and G. A. Susto, "AcME - Accelerated Model-agnostic Explanations: Fast Whitening of the Machine-Learning Black Box." arXiv:2112.12635, 2021.

[8] Fel, T., L. Hervier, D. Vigouroux, A. Poche, J. Plakoo, R. Cadene, M. Chalvidal, J. Colin, T. Boissin, L. Bethune, A. Picard, C. Nicodeme, L. Gardes, G. Flandin, and T. Serre, "Xplique: A Deep Learning Explainability Toolbox," in CVPR 2022 Workshop on Explainable Artificial Intelligence for Computer Vision. arXiv:2206.04394, 2022.

[9] Goetschalckx, L., A. Andonian, A. Oliva, and P. Isola, "GANalyze: Toward Visual Definitions of Cognitive Image Properties." arXiv:1906.10112, 2019.

[10] Kitouni, O., N. Nolte, V. S. Pérez-Díaz, S. Trifinopoulos, and M. Williams, "From Neurons to Neutrons: A Case Study in Mechanistic Interpretability," in Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. arXiv:2012.01264, 2024.

[11] Klaise, J., A. Van Looveren, G. Vacanti, and A. Coca, "Alibi Explain: Algorithms for Explaining Machine Learning Models," in Journal of Machine Learning Research 22, 2021.

[12] Müller, R., M. Abdelaal, and D. Stjelja, "Open-Source Drift Detection Tools in Action: Insights from Two Use Cases," in ACM Conference'17, Washington, DC, USA. arXiv:2404.18673, 2024.

[13] Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation (DALL-E)." arXiv:2102.12092, 2021.

[14] Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion)," in CVPR 2022. arXiv:2112.10752, 2022.

[15] Ross, A. S., N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez, "Evaluating the Interpretability of Generative Models by Interactive Reconstruction," in CHI '21, Yokohama, Japan. arXiv:2102.01264, 2021.

[16] Seldon Technologies Ltd., "Alibi Explain Online Documentation." 2019.

[17] Sharma, R., N. Reddy, V. Kamakshi, N. C. Krishnan, and S. Jain, "MAIRE - A Model-Agnostic Interpretable Rule Extraction Procedure for Explaining Classifiers," in Artificial Intelligence Journal. arXiv:2011.01506, 2020.