Explainable Visualizations as a Method of Improving Understanding of Artificial Intelligence Systems

Olivia Seow
Massachusetts Institute of Technology, olivias@mit.edu
Dave Ludgin
Massachusetts Institute of Technology, dludgin@mit.edu
John Liu
Massachusetts Institute of Technology, johncliu@mit.edu

Abstract

A primary barrier to using artificial intelligence is a proper understanding of how the system works, which enables trusts. To confront this, explainable visualizations provide a scaffolding for providing transparency to users. In order to provide this equitably, explainable visualizations should balance completeness and interpretability.

Keywords: Explainable AI, HCI, Visualization, GANs, ML

1 Introduction

Artificial Intelligence (AI) & Machine Learning (ML) are widely regarded as core technologies that will impact every corner of the global economy. In 2020, Sundar Pichai, CEO of Alphabet, stated at the World Economic Forum, "Artificial Intelligence will be more profound for humanity than fire and electricity."

While promising, the effectiveness of these AI systems is limited by computers' perceived inability to explain their decisions and actions to human users. The black-box perception of deep neural networks and their inability to adapt to change challenge its use in a wide array of applications, raising ethical and judicial concerns and inducing a lack of trust. In addition to business and academic stakeholders seeking proper understanding of advanced AI systems, policy makers are taking early steps to implement regulations that allow for the ethical deployment of ML, leading to an explosion of interest in interpreting the representations and decisions of "black-box models". [3]

There has been a considerable amount of research regarding the dangers of AI and ML, particularly in the context of biased systems. However, it is also important to highlight the massive potential of these systems. AI and ML can aid humans in improving decisions and outcomes by providing context and analysis we are simply incapable of doing ourselves. The key issue is that humans are hesitant to utilize these systems without trust. In the medical field, research has demonstrated that physicians do not utilize AI/ML systems that have a higher accuracy in predicting cancer because of the black box problem. Health professionals find that systems that are opaque, "lack quality assurance, fail to elicit trust, and restrict physician-patient dialogue."[6]

These missed opportunities for human-AI collaboration extend to all parts of society. Daniel Kahneman, a leading behavioral expert, has demonstrated that any system that relies on human interpretation contains "noise", which is the variability in judgements based on the same information. Additionally, Kahneman argues that the best way to remedy these problems is not to impose systems, but to ask the individuals within the system to come up with a solution. A broader understanding of AI/ML systems would allow these individuals to appropriately adopt technology to create more equitable systems.

As a result, Explainable Artificial Intelligence has emerged as a critical field of research that is essential for greater adoption & diffusion of Machine Learning applications across industries. "The ever increasing number of scientific articles, conferences, and symposia for this field has led to the development of domain-dependent and context specific methods for dealing with the interpretation of machine learning (ML) models and the formation of explanations for humans."

The purpose of the project described in this paper is to demonstrate the use of interactive data visualization techniques in explaining the power of generative neural networks to the average individual. The ML application that was developed for this paper can easily be designed for use cases in the context of social media, augmented reality, and creative technology tools. We conclude by considering the ethical implications of this technology when utilizing data sets with embedded bias or when generating fake media that result in the degradation of public trust.

2 Related Works

The research community has worked on explanations of intelligent systems since the 1970s, but recent successes of AI and machine learning in highly visible applications have sparked a new wave of interest in understanding these systems. Despite this interest, much work in AI and ML communities tends to suffer from a lack of usability, practical interpretability, and efficacy on real users and in turn, prevents trust and adoption of AI applications. As stated by Shneiderman et al. [158], the need for interfaces that allow users "to better understand underlying computational processes" and give users "the potential to better control their (the algorithms') actions" is one of the grand challenges for HCI researchers. [1]

3 Definitions & Concepts

To better address the effectiveness of Explainable AI, we first begin by defining two key concepts that influence explainability - "interpretability" and "completeness" [4].

The goal of interpretability is to describe the internals of a system in a way that is understandable to humans. The success of this goal is tied to the cognition, knowledge, and biases of the user: for a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user.

The goal of completeness is to describe the operation of a system in an accurate way. An explanation is more complete when it allows the behavior of the system to be anticipated in more situations. When explaining a self-contained computer program such as a deep neural network, a perfectly complete explanation can always be given by revealing all the mathematical operations and parameters in the system. The challenge facing Explainable AI is in creating explanations that are both complete and interpretable: it is difficult to achieve interpretability and completeness simultaneously. The most accurate explanations are not easily interpretable to people; and conversely the most interpretable descriptions often do not provide predictive power. Additionally, both interpretability and completeness are significant factors impacting trust in AI/ML systems. Figure 1 illustrates this tradeoff that a designer faces in creating visualizations for this field.

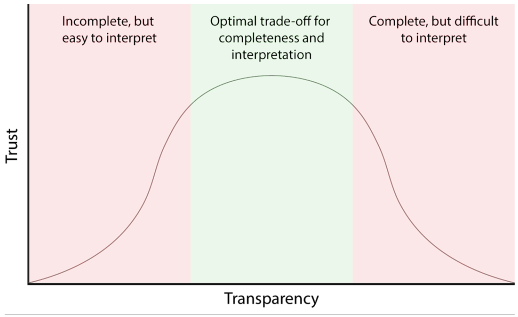

Figure 1: Graph depicting the relative trade-off between interpretability and completeness, resulting in different levels of trust.

4 Method

Our objective in this visualization project was to help individuals with a basic understanding of Artificial Intelligence to comprehend the mechanics of machine learning and generative neural networks. To achieve this objective, our team focused on simplicity and ease of understanding as measures of visualization success. When considering the trade-off between interpretability and completeness, we focused more on interpretability and nuanced higher-level understanding for our end users.

5 Data Set

The dataset used for this visualization project was originally created for a published paper by Tero Karras on "Training Generative Adversarial Networks with Limited Data". The dataset consists of 1,336 high-quality portrait PNG images at 1024x1024 resolution downloaded from the Metropolitan Museum of Art Collection API and automatically aligned and cropped using dlib (see Figure 2). The original dataset included portrait titles and artist names. During exploratory data analysis, we expanded this dataset to include additional attributes such as museum department, accession year, medium, dimensions, and search tags.

Figure 2: Sampling of images from original dataset

6 Narrative, Data Processing, Design Rationale

6.1 "What are Portrait?"

Our visualization uses the scrollytelling interactive format to tell a narrative from a computer's perspective. The user is introduced to the content through an opening section titled "What are portrait?" which is meant to establish a conversational tone for the computer narrator. The subtitle "A machine's learnings from 1,336 portraits at the Met" is meant to personify the AI and this dialogue between the user is carried forward in each section. We believe that the personification of the AI will make it easier for users to empathize with the system and increase the impact of analogies in the visualization. The backdrop of vibrant portraits is meant to introduce the user to the dataset that will be explored throughout the narrative. For increased visual interest, this backdrop is reshuffled every time the page is refreshed.

The transition from the opening section to this second section of the visualization is extremely important as the animated portraits moving into place on the scatterplot convey much of the meaning described below without words by transforming the portraits into their dominant color pixels. While the user cannot trace each and every portrait in the animation, they do get a sense that these images are moving to their appropriate position on this scatter plot, which sets up the next part of the narrative. We chose not to give a deep explanation of how the RGB values are created due to our focus on simplicity, which is why this animation is critical. It conveys the key concept without complexity.

On the backend, this transformation was achieved by running k-means on the pixels and returning the centroid of the largest cluster. To speed up processing, the 1024 × 1024 images were first downsampled to 25 × 25. A 3D plot (shown in Figure 3 below) was created in python to investigate the result and determine interesting views that fit into the central narrative.

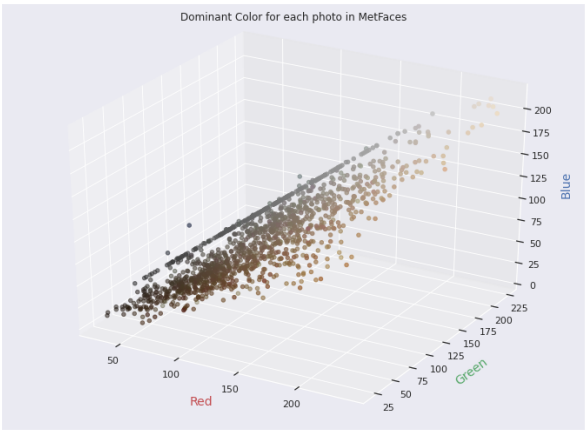

Figure 3: 3D scatter plot of dominant colors in dataset

6.2 "Pixels, Pixels, Pixels"

This section of the scrollytelling was created to support two purposes - explanation and exploration. The scrolling explanatory text boxes provide the viewer with an understanding for how a computer sees portraits and these dominant color pixels as defined colored values. We found the scatter plot to be most appropriate and effective in conveying the varying degrees of dominant color pixels in each portrait and highlighting early clusters of these RGB values. Exploration is facilitated by the tooltip, which is introduced early on so the user may discover specific portraits and artist names. Furthermore, the visual encoding of the average color in a portrait for each point bolsters the concept without complex explanations.

6.3 "Next, Same Data, New Perspective"

The next animation is initiated by the user scrolling down and changing the y-axis color value from Green to Blue. The subtle dispersion of data points is meant to provide visual guidance as the text explains the variability of dominant colors within this dataset. We continue to offer opportunities for exploration using the same hover tooltip interaction. While the first section may have sufficed for some users to understand how computers see images, the additional visualization serves as a check and reinforcement. The check is critical because the narrative builds and missed understanding impacts later interpretability.

6.4 “Moving up from Pixel values”

As the user moves through the first third of the narrative, a more dramatic transition takes place with the visualization hiding the scatter plot and reconfiguring data points into a web floating in white space. The change in form is meant to showcase new image clusters based on features and shapes rather than only a single color metric. Additionally, the dramatic animation serves as a pivot point in the story. We are taking a step up in understanding here by moving from seeing to interpreting that information. This creates a moment of friction causing a user to slow down and be more thoughtful about what they are viewing.

On the backend, this is accomplished by performing a Uniform Manifold Approximation and Projection (UMAP) projection onto a two-dimensional space, illustrated in FIgure 4.

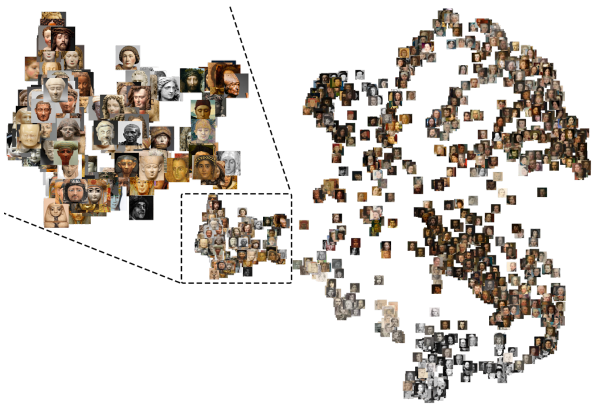

Figure 4: UMAP encoding of dataset, with one cluster magnified to show details

While this method does not directly feed into the Generative Adversarial Network (GAN) process described in the later sections of the scrollytelling, it is a helpful reference for viewers to contemplate how a machine learning model reduces high dimensional data in order to perform predictions. To demonstrate the arbitrary nature of the visual space in this encoding, we also rearrange the data into a grid view in the following section.

6.5 “Clustering by Likeness”

At this point in the narrative, the user has become familiar with the layout of the scrolly-telling format. In place of the scatter plot and cluster web, we use a grid for ease of viewing. Through the text, we convey that this view is meant to aid human understanding, and is not critical for the computer. In this section, the visualization highlights 10 cluster representative images and upon hovering over the images, the user can see the associated cluster within the new portrait grid. The visualization is highly responsive and intuitive. We chose this interaction to draw attention to the critical link between the clusters and the representative images while offering a critical point that machines cannot tell how the images are alike in the real world. An exploration of these clusters shows how AI systems sometimes cluster in ways that are familiar to our users, but also in ways that are not intuitive.

To enhance the experience, users are encouraged to click on a representative image, which will animate the grid axis and reveal more details about that cluster. The associated portraits are magnified and re-arranged on the screen. When users select another representative image, they are taken to the new cluster within the grid. A key effect of the zooming across and in and out of the grid is to give the user a true feeling of exploration within the dataset. Figure 5 shows a mid-transition screenshot of images falling to place. When the user navigates the cursor away from the representative images, the visualization resets. While some users might want to explore deeper, this is part of scaling complexity. We intend for the user to have an understanding of the clustering as a whole before continuing on.

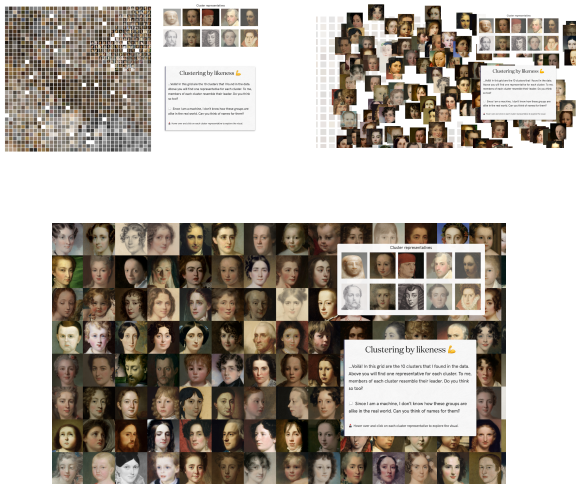

Figure 5: Mid-transition screenshots of images from a specific cluster flying into place, and other clusters fading out

5.6 “Latent Space”

As the narrative moves to its climax, the user is introduced to the idea of latent space encoding as a means to generate entirely new portraits. The images displayed look oddly real but upon closer inspection via a tooltip, the user will notice titles are inappropriately matched with the subject within the painting. This mismatch was an intentional design choice to signal to the user the short-comings of the AI system and the bias embedded within the dataset. We then introduce the idea of latent space arithmetic which explains at a high-level how these new images are created by seeking the in-between imaginary images between two portraits, as demonstrated in Figure 6.

Figure 6: Gridview illustrating latent space arithmetic - the bottom rows are midpoints between the top row and the left image

The portrait titles were created using a text-generating neural net trained on the titles in the original dataset. The network also learns the structure of the titles, such as letter casing and the inclusion of years in some samples. Examples titles generated include “Madame Head of a Lady" and “The Grandre Colley (1762 - 1791)”.

6.7 Have I Met You

This final section serves as a reinforcement around the idea of latent space. We had some stretch goals for this section we were not able to complete - we wanted to enhance the “martini glass” navigation and allow users to explore the latent space with new input by uploading a photo or turning on their web camera. Instead, we provide a video of the transformations on four celebrities so that users can imagine how we can perform projections into the latent space of generative neural networks, and how the outcome is limited by biases in its training data. This video is also interactive in that viewers can drag the image to scrub frames back and forth.

The visuals in sections 3.2.6 to 3.2.7 were created by a StyleGAN2 model trained on the MetFaces dataset. To reduce overfitting, an adaptive discriminator augmentation (ADA) mechanism was used to stabilize training on this relatively small (1000+ images) dataset.

7 Results

Upon testing the data visualization with a small group of 5 users with varying degrees of AI knowledge, two areas of improvement related to the completeness of the Explainable AI visualization stood out. The first relates to the leap from an AI analyzing dominant color pixels within an image to the AI analyzing more complex shapes and patterns that result in representative clusters. The second relates to the use of latent space encoding to produce new portraits and images.

While we do not have quantitative measurements for the design effectiveness of our Explainable AI, we can qualitatively assess our visualization against the Guidelines for Human-AI Interaction that were presented by researchers at Microsoft and University of Washington at CHI 2019. [5] Their paper laid out 18 generally applicable design guidelines based on AI-infused products allocated across 4 scenarios for when they apply to interactions with users.

As our visualization is not a full-fledged product, we focus on the general interaction experience rather than scenarios when interactions take place incorrectly or during overtime. The 6 most relevant guidelines are: Making clear what the system can do, Making clear how well the system can do what it can do, Time services based on context, Show contextually relevant information, Match relevant Social Norms, and Mitigate Social Bias. In future iterations of this visualization project, we would map enhancements of the design based on this criteria

8 Discussion

Our initial intended outcome is a more educated viewer about AI/ML. Ideally they will understand how the output of an AI/ML system (e.g. the generated faces) is a product of its analysis of a dataset and the dataset itself. By having a better understanding of how the AI/ML systems work and how the datasets are incorporated, viewers can ask appropriate questions when considering the use of such systems in the future. Further work could be done to help illuminate the pitfalls and concerns at each level.

A subsequent outcome we hope our visualization provides is increased trust as a result of greater transparency through understanding. Some qualitative discussions with users lead us to believe that this result was achieved, although we mention further aspirations for measuring this result in the future work section.

Both of these outcomes are supported by the overall design approach of using a simplified scenario and presenting concepts in different visual representations. One of the more challenging concepts to understand is that of ”latent space.” Through conversations with individuals that do not have experience with AI/ML, they found the “Moving up from pixel values” section to be helpful in imagining an abstract concept. This insight is worth further investigating as certain operations by AI/ML systems do not easily match with a familiar concept. By identifying these more abstract components of systems, developers can focus their effort on optimizing transparency.

Another key takeaway from the visualization is the varying outcomes when projecting four celebrity faces into the latent encoding. Viewers saw how the model performs better on Joe Biden and Taylor Swift, and less well on Kamala Harris and Dwayne Johnson (Figure 7), and are prompted to consider how the original data affected this outcome. This serves as an important point of consideration for anyone deliberating the use of machine learning, in that a model can only be as aware about the real world as the data it is given.

Figure 7: Outputs from projecting four celebrity faces into the StyleGan2 latent space

Lastly, our visualization has provided users with a better understanding of what computers can do for us. From the beginning, the visualization shows how computers can abstract and transform data into a format that we can use. More importantly, users get to see how generative systems can create portrait faces.

9 Future Work

From an industry and technological adoption perspective, human-centered AI developers must focus on establishing, nurturing, and maintaining trust in AI systems. The question that arises then is how does a design team measure and optimize this trust over time with explainable visualization tools?

If we were able to continue the project, we would like to quantitatively measure the impact of the design using validated measures. The first such measure is the General Attitudes towards Artificial Intelligence Scale [2]. To further understand mechanisms, we would like to also measure whether the participants are System 1 or System 2 dominant in their thinking. An MIT research group has a paper to be published in the next year that demonstrates that system 1 thinkers are inherently less trusting of AI/ML systems. Having this additional information can illuminate how these groups may be differently impacted by our project.

With regard to AI education, Explainable Visualizations also serve the purpose of bringing a more diverse community to AI research and emerging disciplines. Creating transparency and understanding is not simply a matter of trust, but it provides equitable use of technology. Requiring AI developers and companies to provide minimum levels of transparency of the output, AI/ML system, and the underlying datasets and transformations can greatly improve the ability of users to responsibly analyze and use the technology. However, this is a labor heavy endeavor for companies and identifying key opportunities and methods to create transparency can reduce this burden and improve the adoption of industry standards. This visualization project serves as a proof point for an exciting future for the positive proliferation of Artificial Intelligence in society.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable, and Intelligible Systems: An HCI Research Agenda. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3173574.3174156

[2] Astrid Schepman, Paul Rodway. 2020. Initial validation of the general attitudes towards Artificial Intelligence Scale. https://doi.org/10.1016/j.chbr.2020.100014

[3] Bryce Goodman, Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a “right to explanation”. presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY https://doi.org/10.1609/aimag.v38i3.2741.

[4] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, Lalana Kagal. 2019. Explaining Explanations: An Overview

of Interpretability of Machine Learning. In The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018).

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300233

[6] Thomas P Quinn, Stephan Jacobs, Manisha Senadeera, Vuong Le, Simon Coghlan. 2020. The Three Ghosts of Medical AI: Can the Balck-Box Present Deliver?