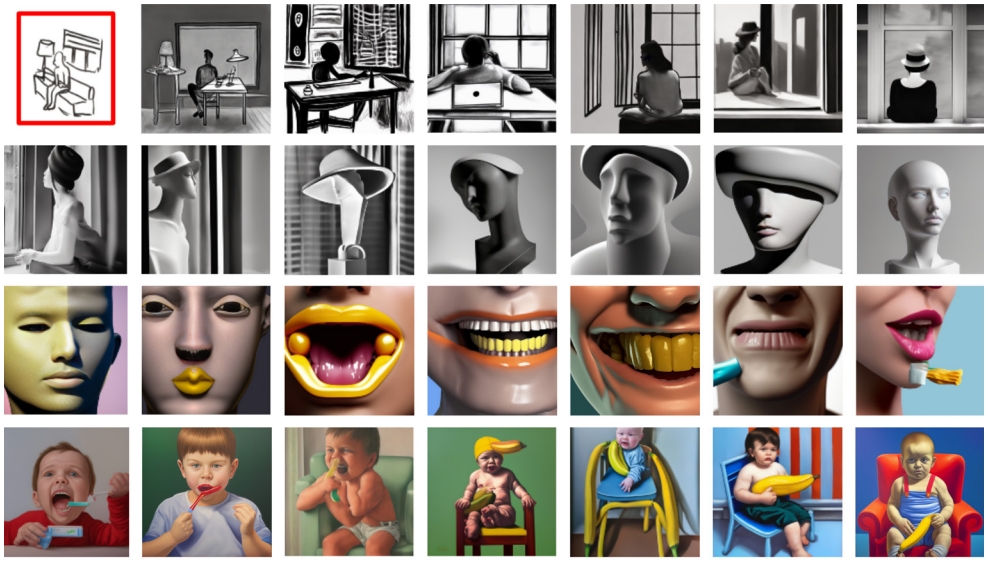
xCoAx 2023 11th Conference on Computation, Communication, Aesthetics & X
2023.xCoAx.org
Weimar, Germany

Visual Dialogues: Doodles that Spark Conversations between Deep Learning Networks



Theodoros Papatheodorou
theodoros@ust.hk
Computational Media and Arts,
Hong Kong University of Science and Technology (Guangzhou), China

Jack DiLaura
mail@jackdilaura.com
Independent computational artist,
Metro Detroit Area, USA

DOI 10.34626/xcoax.2023.11th.349

Visual dialogues is an interactive installation that explores and captures visually the dialogue between two complimentary, but functionally opposite deep learning networks: a text-to-image and an image-to-text model. The user submits a hand drawing to kickstart the process and a prompt describing that hand drawing is generated. This is then sent to a text-to-image model to generate an image, the result of which is submitted back to the image-to-text model and so on. This loop between image and prompt generation is extended to a few generations and we observe the networks slowly drift away from the original subject and style in the participant's original sketch, passing through interesting milestones in prompts and in images. This imperfect dialogue creates an appealing visual trajectory and gives viewers a more intuitive, visual understanding of the workings of these deep learning models.

Keywords: CLIP Network, Stable Diffusion, Generative Art, Interactive Installation, Machine Learning, Text-to-Image Model.

Introduction

Visual dialogues is an experimental, emerging media arts research project exploring the inner workings of deep learning networks and the new possibilities in creating visuals from diffusion models and image prompts. We propose a circuit between a drawing generated by a participant, a text-to-image and an image-to-text component that form an interactive installation that facilitates the dialogue between complimentary but functionally opposite systems.

The work was inspired by the sound-based installation *I Am Sitting in a Room* (Lucier 1969) by American composer and artist Alvin Lucier. The installation consisted of a recorded voice, which is played back into a room and re-recorded multiple times. With each iteration, the sound is gradually degraded, until the original words become unintelligible and are replaced by the resonant frequencies of the room itself. The result is a mesmerising soundscape that is unique to each room in which the piece is performed. Similarly, the sound recording *Disintegration Loops* (Basinski 2002), consists of a single, evolving loop of sound. The compositions are characterised by their slow, gradual evolution and their ethereal textures. The loops are meant to be played continuously, allowing the listener to experience the gradual deterioration of the sound over time.

Our work creates a similar loop, but focuses on generating visuals instead of sound, using state-of-the-art diffusion and image-to-text models to loop back and forth between prompt and image generation.

Implementation

To kickstart the loop, we invite participants to sketch a basic doodle. This doodle is then submitted to a image-to-text neural network to extract an image description from it. To get this description we used CLIP interrogator, which is a prompt engineering tool. Prompt engineering involves transforming one or more tasks into a prompt-based dataset and training a language model through a process referred to as “prompt-based learning”. In 2022, the public was introduced to machine learning models such as DALL-E 2, Stable Diffusion, and Midjourney. These models are designed to accept text prompts as input and generate images, thereby creating and popularising a new category of prompt engineering focused on text-to-image generation. CLIP interrogator combines OpenAI’s CLIP and Salesforce’s BLIP to generate text prompts corresponding to a submitted image. CLIP was a milestone in AI as it was incredibly good at predicting the most relevant text description for an image, without optimising for a particular task. BLIP is similar in architecture, but trained to perform a slightly different task. CLIP focuses more on keywords, often single words, while BLIP is specifically trained to generate captions of about 4-8 words in length. These are then com-

bined to a final prompt that works well for the text-to-image model used in the next step. An example of the output of this process is seen in Figure 1.

Figure 1: Prompt generated by the CLIP interrogator: "A statue of a man standing in a room, a statue, by Michelangelo, pinterest, nubile body, pale grey skin, shows a leg, in style of davey adesida, from wikipedia, rambo, pale skin!" (Image ©Joerg Bittner CC BY-SA 3.0 Wikimedia Commons).

Figure 2: Image generated by Stable Diffusion with the prompt: "a photograph of an astronaut riding a horse" (Image ©Asanagi CC0 Wikimedia Commons)

The text prompt generated was then submitted to a text-to-image model. For this part we used Stable Diffusion which is a diffusion-based, deep learning, text-to-image model released in 2022. Text-to-image generation is primarily employed to produce intricate images based on textual descriptions, however, it can also be adapted to accomplish other tasks such as filling in missing parts of an image, extending an image beyond its original boundaries, and creating transformations on images guided by text prompts. An example of an image generated by Stable Diffusion is seen in Figure 2. The code and weights for the model have been released publicly and it can run on consumer hardware efficiently, which makes it ideal for interactive installations.

Once Stable Diffusion generated an image from the prompt this was submitted to the CLIP interrogator again and the output from that was used to generate a new image and so on. This process created an echo between the two neural networks that drifted slowly away from the original subject in the participant's sketch, passing through interesting milestones both in prompts as well as images. The circuit that makes up the installation can be seen in Figure 3.

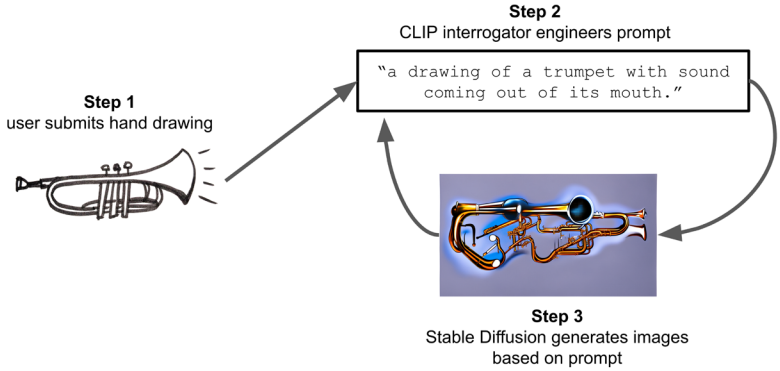


Figure 3: Outline of the installation circuit.

Participant's Experience and Results

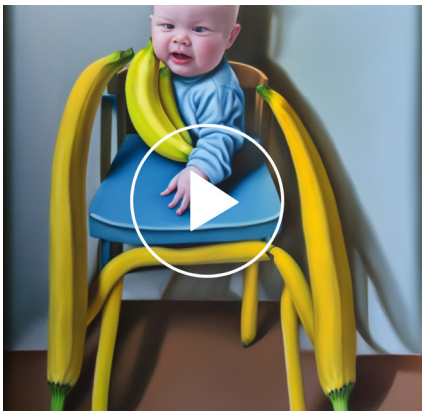


Figure 4: The following videos showcase different directions taken by the two networks during their dialogue: https://vimeo.com/798825496 and https://vimeo.com/798827525

Participants submit drawings online using a custom touch-based drawing web app. Then the loop between image-to-text and text-to-image is initiated. The results of the images generated through this process along with the prompts can be seen in the videos linked to in Figure 4. Due to the space limitations we only show a selection of outputs without the prompts in Figure 5. The first image, in a red frame, is the participant's hand drawing that was used as input. Harking back to Lucier's *I am sitting in a room* recording, the participant drew a person sitting in a room.

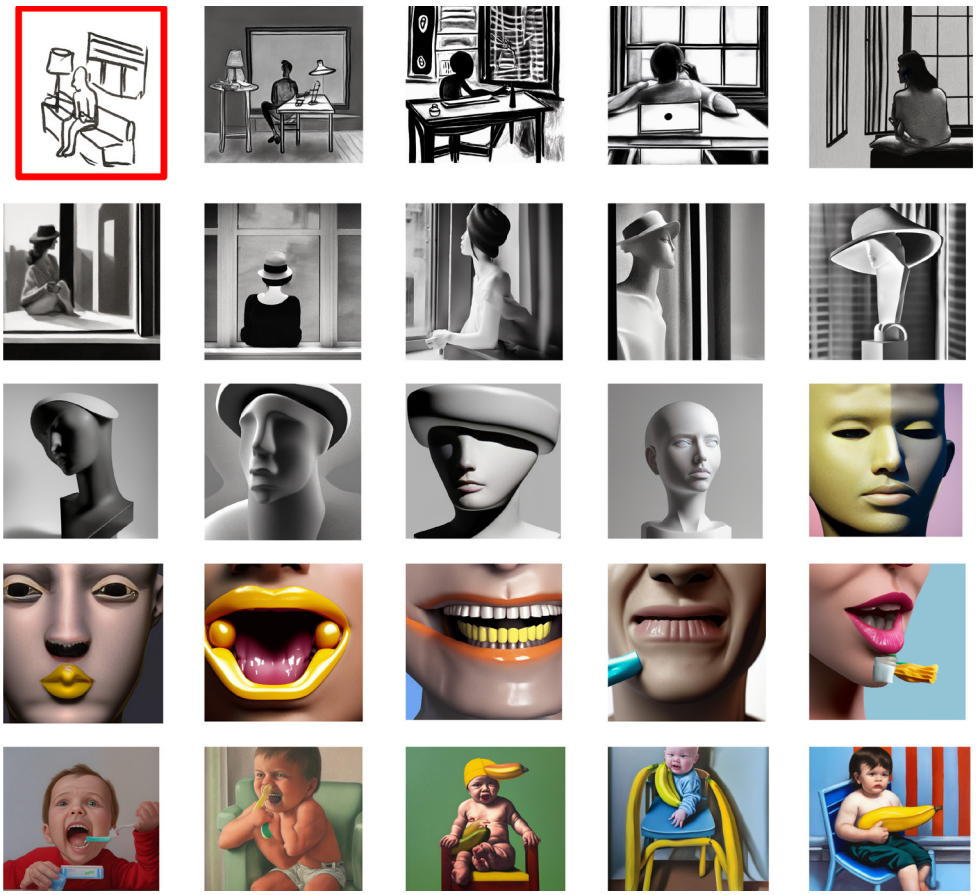


Figure 5: A sample of outputs from the dialogue, in sequence from left to right. The first image, in a red frame, is the participant's hand-drawing that kickstarted the circuit. Harking back to Lucier's I am sitting in a room recording, the participant drew a person sitting in a room.

Discussion

What makes the installation interesting is the theme of the generation loss concept in the fidelity of the image. Like the degrading recordings, not all information is captured or transmitted between generations. This degradation resulting from the dialogue between the image-to-text and text-to-image loop gives us a visual understanding of the underlying mechanisms governing these algorithms. We can observe with greater clarity the things that each model considered important, what it ignored and occasionally around which themes it circled for a long time without being able to escape.

This imperfect dialogue between the two models opens creative possibilities, as it allows us to explore the image space freely to find image styles that we like. Once these are located, we can intercept

the process by injecting more words in the prompt to calibrate the style or content we want. The interactive component, allowing the participant to input the seed that kick-starts the process, introduces a level of unpredictability and variety to the output of this dialogue that might not be possible if using a fixed set of input data. However, abstracted it may become, it is the participant's original input that is echoed across generations of images. Observing them one can always see links from generation to generation which slowly fade away as more loops have taken place.

References

Lucier, Alvin. 1969. *I Am Sitting in a Room*. Retrieved February 14, 2023 from https://en.wikipedia.org/wiki/I_Am_Sitting_in_a_Room

Basinski, William. 2002. *The Disintegration Loops*. Retrieved February 14, 2023 from https://en.wikipedia.org/wiki/The_Disintegration_Loops