The @artbhot Text-To-Image Twitter Bot

Amy Smith[1] and **Simon Colton**[1,2]

[1] School of Electronic Engineering and Computer Science, Queen Mary University London, UK
[2] SensiLab, Faculty of Information Technology, Monash University, Australia

amy.smith@qmul.ac.uk s.colton@qmul.ac.uk

Abstract

@artbhot is a Twitter bot that brings the generative capabilities of CLIP-guided GAN image generation to the public domain by transforming user-given text prompts into novel artistic imagery. Until recently, access to such image synthesis techniques has been largely restricted to Google Colab notebooks, which require some technical knowledge to use, and limited services which require access. @artbhot increases access to text-to-image technology, as Twitter users already have the platform knowledge needed to interact with the model. We discuss here some of the technical challenges of implementing such a system, and provide some illustrative examples of its usage. We further discuss what this mounting of generative technology amongst social media could mean for autonomous computationally creative agents.

Introduction

Recent developments with generative deep learning technologies have enabled text-to-image computational models to produce artistic images and video content, given only text prompts from users. Colton et. al (2021) explored the possibilities for this, within the context of *generative search engines*, where images are generated rather than retrieved as per Google image search. Such approaches in the field of text-to-image synthesis (Agnese et al. 2019), allow the user to encode text in such a way as to drive a search for a latent vector input to a pre-trained image generation neural model. This technology has an impressive ability to innovate novel visual content from text, producing high quality and diverse imagery which reflects the prompt well, with images that are often surprisingly innovative. Examples of the kind of artwork that can be produced are given in (Smith and Colton 2021), and we describe the CLIP-Guided VQGAN text-to-image system in the background section below.

Interaction with such systems has been largely limited to Google Colab notebooks (Bisong 2019), but this has barriers to entry due to the the technical knowledge required to run the notebooks, and user interaction is limited to an image retrieval service. Other recent text-to-image generators (mentioned below) have invitation-only limited access for a small number of artists and researchers. To address this lack of access, we have built the @artbhot twitter-bot (Veale and Cook 2018), which embeds CLIP-guided VQGAN in the Twitter social media platform experience. As described and illustrated with examples below, people can tweet their text prompt with appropriate annotations, and expect an image to be returned in due course. This greatly increases accessibility to the public, as Twitter has over 200 million active users. Due to it's popularity and reach, and both the data and interaction available through its API, Twitter also provides an ideal platform for @artbhot to take on more creative autonomy. In particular, we plan to challenge the assumption that text-to-image users should be served only imagery which purely reflects their prompt. Instead, as described in the final section below, we aim for @artbhot to use prompts as springboards for creative ideation and visualisation and for it to enter into a dialogue with users in a fashion akin to discussions with artists on social media.

Background

In early 2021, Ryan Murdock combined OpenAI's Contrastive Learning Image Pretraining model (CLIP) (Radford et al. 2021) with the BigGAN generative adversarial network (Brock, Donahue, and Simonyan 2019) into a text-to-image generation process. He made the system available via a Colab notebook called *The Big Sleep*. In overview (with further details in (Colton et al. 2021)), the process involves first encoding a user-given text prompt into the CLIP latent space as vector v_1. Then the system performs a search for a latent vector input to BigGAN, v_2, which produces an image that, when encoded into the CLIP latent space as v_3, has optimally low cosine distance between v_1 and v_3. The search is performed using gradient descent to minimise a loss function based on this cosine distance. Given that related images and text are encoded by CLIP to similar places in the latent space, this approach tends to produce images which somehow reflect the given text prompt.

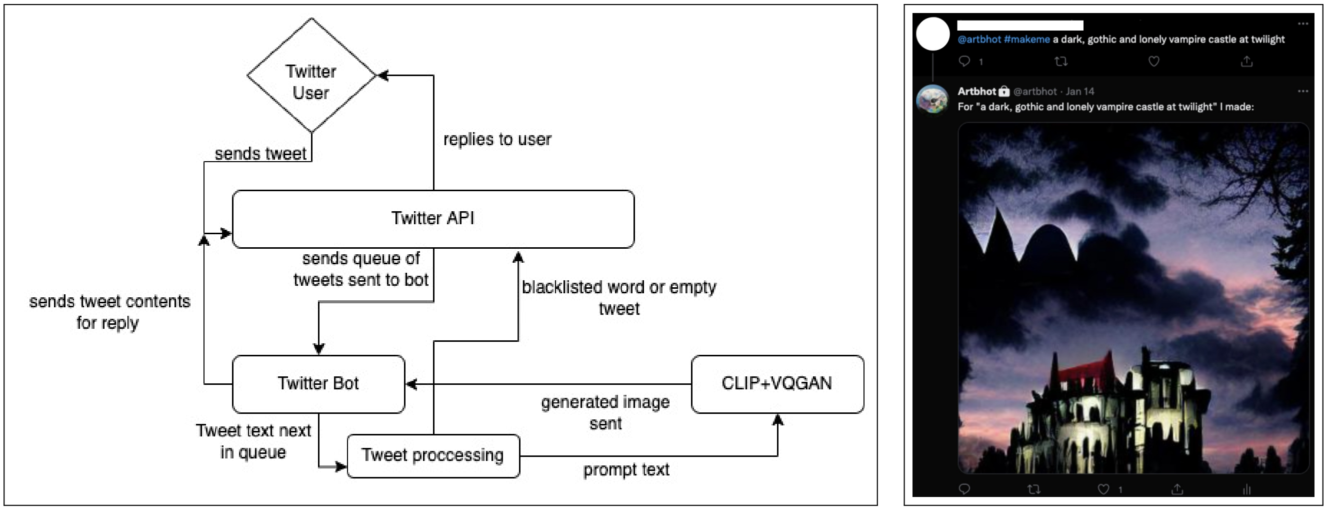
In the interim, many CLIP-guided text-to-image generators have been made available, with steadily improved quality and fidelity (with respect to the prompt) of the images produced. The most recent, and impressive examples of this generative technology are *@midjourney*[1], *Disco Diffusion*[2], DALL-E [3] from OpenAI and Imagen[4] from Google. DALL-

[1] midjourney.co

[2] tinyurl.com/yckn4h7

[3] openai.com/dall-e-

[4] imagen.research.google/



[redacted]
@artbhot #makeme a dark, gothic and lonely vampire castle at twilight

Artbhot @artbhot · Jan 14
For "a dark, gothic and lonely vampire castle at twilight" I made:

Figure 1: (a) Processing of a tweet by @artbhot (b) Example user interaction on Twitter.

E is particularly impressive as it employs a one-shot process, with an encoded text prompt fed-forward through a model to produce images near-instantaneously. However, the trained model is so large that access is limited, with the expectation that OpenAI will provide a subscription service for it soon. Currently, *Disco diffusion* is available as a Google Colab notebook, and @*midjourney* is only available to selected users. *Wombo Dream*[5] however is an app that is available for free from the app store, and appears to have been very popular. In addition to users being able to enter a prompt and receive an image based on this text, they can also select from several art styles that can influence the aesthetic of their generated image. These styles include 'Dark Fantasy', 'Mystical' and 'Salvador Dali'. There is also now *DALL.E mini* [6] which is available to the public and free of charge. It is a smaller version of the model mentioned above and is hosted on Hugging Face[7].

In a similar process to that of the Big Sleep approach, CLIP-guided VQGAN harnesses the perceptual power of CLIP and the image generation capabilities of the Vector Quantized Generative Adversarial Network (VQGAN) (Esser, Rombach, and Ommer 2021). This GAN architecture combines two approaches to interpreting meaning, using both discrete and continuous representations of content (Cartuyvels, Spinks, and Moens 2021). Discrete representations model a more human way of interpreting meaning aside from a pixel based approach, which is traditionally how computers have processed images. In particular it considers the image as a whole and interprets the relationships between the different compositional elements of the contents, i.e., relationships between different parts of an image (such as the sky and the ground in a landscape image).

VQGAN models these discrete representations as long range dependencies, meaning it can interpret the *relationships* between compositional elements, and not just the elements themselves, as described in (Esser, Rombach, and Ommer 2021). VQGAN models image elements, and the local relationships within visual parts of an image, using continuous representations (such as the RGB channels in a pixel). It also interprets discrete representations within image content using a transformer (Vaswani et al. 2017), but before a feature map can be passed to this, the model learns an intermediary representation of this image data using a *codebook*, as described at tinyurl.com/2vm3t9r8. This is a fixed size table of embedding vectors that is learned by the model. This intermediary stage is necessary, as transformers scale the length of an input sequence quadratically, making even a 224 x 224 pixel image above the processing capacity of most GPUs. CLIP-guided VQGAN is described in (Crowson et al. 2022), and various notebook for CLIP-guided VQGAN have been implemented, with a list of ten given here: ljvmiranda921.github.io/notebook/2021/08/11/vqgan-list/

@artbhot Implementation and Deployment

Twitter bots are usually small, autonomous programs running on a server, which regularly produce and tweet outputs composed of texts, images, animations and/or music/audio compositions, as described in (Veale and Cook 2018). More advanced bots can respond to replies on Twitter and/or tweets if they are hashtagged appropriately. Our Twitter bot, @artbhot, is currently only reactive, in that it is used as a service: people tweet text prompt requests at it, and it responds with a reply comprising an image that (hopefully) reflects the prompt, and a repetition of the prompt.

@artbhot is comprised of two parts: the generative process, which is provided by CLIP-guided VQGAN; and code which enables it to interact with the Twitter API. The implementation is hosted on a remote server which runs 24 hours a day, so users can access image generation capabilities on demand. Users can read instructions on how to use the bot from a document linked in the bio section of the @artbhot's

[5] wombo.art

[6] tinyurl.com/4eyr5yjv

[7] huggingface.co

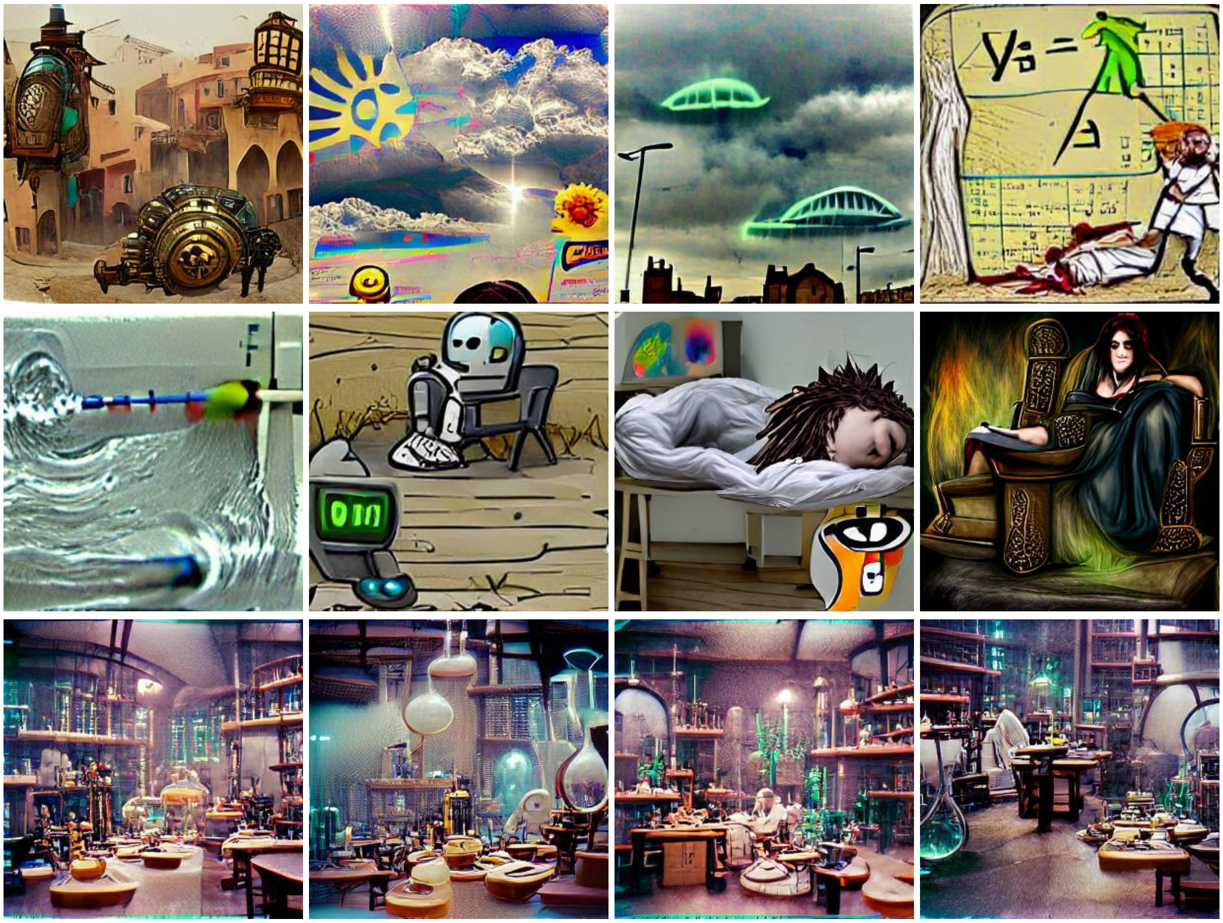


Figure 2: Generated images for prompts. **Top row:** “Steampunk morocco, concept art”; “🌞☀🌤🌦🔆🌥”; “Aliens invading Newcastle Upon Tyne”; “Pythagoras killing his student because the square root of 2 is irrational”. **Middle row:** “A positive lateral flow test”; “Waiting for the bot”; “Wake up @artbhot”; “The Scribe, sitting in her throne. Deviant art character illustration”. **Bottom row (all):** “A 35mm analog film photo of an alchemists lab in the distant future”.

Twitter page. These instructions include how to communicate with the bot using the following tweet format:

```
@artbhot #makeme prompt text
```

(e.g. @artbhot #makeme an oil painting of a burger).

Every 15 seconds, the bot code checks for new tweets in this format from any user, using the python Twitter API. Once found, the prompt text is extracted, processed and either used as input for a CLIP-guided VQGAN process, or rejected for containing any prohibited words. This cross-referencing of the prompt against a list of prohibited words aims to keep the experience of using the bot as friendly as possible. If a prohibited word is found, a textual reply is automatically generated and sent to the user as a reply to their tweet, asking them to try again. The processing performed by @artbhot for a given input tweet is portrayed in fig. 1(a).

If an image is generated, it is then sent to the user via the Twitter API as a reply to their initial tweet, with a reminder of the prompt they used (this is to ensure that the prompt text follows the generated image in the case where a bot reply is shared on Twitter without the original tweet from the user to provide context). An example user interaction on Twitter with @artbhot is given in figure 1(b). The first iteration of @artbhot incorporated CLIP guided BigGAN for image generation, as this model was one of the best CLIP guided GANs available to the public. This was a local version of the code released in the Big Sleep colab notebook, installed on our server. Later, an implementation of CLIP-guided VQGAN was released (github.com/nerdyrodent/VQGAN-CLIP). On experimenting with this text-to-image generator, we found that the output from the newer model showed improvements in multiple ways. Firstly, almost no images were outright failures from VQGAN in the way that BigGAN regularly generated blank or highly noisy/textured uninterpretable images. Also, the fidelity of the image to the prompt was usually much better and there was much less visual indeterminancy (Hertzmann 2020), making the images more coherent from VQGAN than from BigGAN. For these reasons, we replaced BigGAN in @artbhot with VQGAN. The top two rows of figure 2 show 8 example images generated in response to tweets sent to it, which we refer to in the next subsection.

A Preliminary Evaluation

We plan to make @artbhot open to the public in 2022, after some additional implementation described in future work below. Before this, we have made it available to a user group of 16 people. It has been running for 5 months and has processed over 600 tweets, taking, on average, around 2 minutes for a user to receive an image in response to their tweet. While there have been no outright failures where images don't reflect the prompt at all, after an informal evaluation (by ourselves) of the most recent 100 replies to Twitter prompts, we found 16% of the images were not visually coherent enough to reflect the prompt satisfactorily. Two examples of this can be seen on the left of row two in figure 2, with neither properly reflecting the prompt "A positive lateral flow test" or "Waiting for the bot". Generally, the images that are less successful have a high degree of visual indeterminacy (Hertzmann 2020), making it difficult to interpret the content of the image and how it may be associated with the tweet text. Other factors for relative failure include content that is off topic, inaccurate colours for the subject matter, or image content that is too small and/or off-centre. We do acknowledge however that this is a subjective evaluation and that other opinions may differ regarding interpretations of image content.

We found that @artbhot was able to handle unexpected prompts, for instance ones containing emojis. As per the second image in the first row of figure 2, CLIP-guided VQ-GAN interpreted the weather emojis correctly and produced an image with sun and clouds. Diversity was also a concern, as users would expect a variety of images for similar prompts. We asked four users to each use the prompt "a 35mm analog film photo of an alchemists lab in the distant future", with the resulting images portrayed in the bottom row of figure 2. We see that there is some diversity, but perhaps not enough to be satisfying, and this is something we hope to improve upon, probably with automated augmentation/alteration of prompts.

Overall, the interactions users have had with @artbhot have been playful and casual, with people feeling free to try out all manner of interesting and unusual prompts, often trying to stretch the bot past its limitations. The qualitative responses we've gathered have been largely positive, with people reporting they have used it for amusement, entertainment and conversation, but wish it would return images faster, as attention can wane. We noticed some trends in the kinds of prompts users sent, including: referring to the bot itself (see middle row of figure 2); setting moods or styles such as *steampunk* (first image of top row); setting up imaginary or historical scenes such as aliens over cityscapes or pythagorean murders (top row, right); and asking for design inspiration (final image on the middle row). One user wanted longer interactions with @artbhot, in particular to ask it to enhance images and to combine their prompts/images with those from friends.

Conclusions and Future Work

Text-to-image colab notebooks are very popular, and initial responses to @artbhot suggest that it would also be very popular on twitter. Unfortunately, it is beyond our computational resources to provide GPU processing to anyone on twitter who tweets a prompt. Moreover, as predicted in (Colton et al. 2021), there seems little doubt that consumer text-to-image generation services will become available soon, and will likely find their way into products such as Adobe's Creative Suite eventually. For these reasons, we are interested in offering more than a service which fulfils image generation requests, as @artbhot currently does. Instead, we will open up @artbhot so that it can receive tweets from any member of the public (which it currently does not), and select a few tweets each day to reply to that have the highest potential for a meaningful, creative and thought-provoking interaction with the user. Once a user is selected, this longer interaction with @artbhot may take the form of a string of iterations on an image; as the user asks to 'evolvethis' image to repeatedly evolve the image with new prompts. This may also take the form of merging several tweets in to a prompt, that is then used to generate an image, using a 'mergethis' hashtag. In this way, the user will still feel in control of the process, but will receive innovative and surprising output as the bot takes on more autonomy.

On responding to the chosen prompts, we plan for @artbhot to apply a range of generative techniques and appeal to a number of computational creativity theories and practices. These include (on the text side) fictional ideation, humour, narrative generation, poetry, etc., and (on the imagery side) style transfer, animations, and visual stories. @artbhot will employ framing and explainable computational creativity techniques (Llano et al. 2020) to get users to look more closely at its ideas and creations. We further aim to enable @artbhot to learn from feedback, so as to be more interesting and engaging for users.

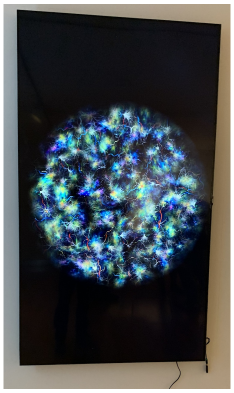


Figure 3: Exhibition piece: Pericellular Nests

We also aim to encourage conversation and collaboration with users, to ultimately generate pieces deemed to be artworks rather than just imagery reflecting text. To do this, we will need to utilise existing evaluation techniques from casual creators (Compton and Mateas 2015) and computational creativity in general, and to develop new ones specific to the project. We will also need to implement more advanced artistic image generation techniques. We have already taken first steps in this direction by writing software which takes animations from @artbhot and makes a large collaged animation (as per fig. 3) for an exhibition[8] at the Pablo Gargallo Museum in Zaragoza, Spain; celebrating the life and work of nobel laureate Santiago Ramon y Cajal.

[8] zaragoza.es/sede/servicio/cultura/evento/232731

Author Contributions

AS is lead author, SC is second author and contributed to writing, supervision, reviewing & editing. Both AS and SC contributed to the concept of @Artbhot, AS developed and evaluated @Artbhot. SC implemented initial interactions with CLIP + BigGAN while AS implemented initial interactions with CLIP + VQGAN.

Acknowledgments

We would like to thank Dr. Mike Cook for his support. We would also like to thank Ryan Murdock, Katherine Crowson and Nerdy Rodent for their ground breaking contributions in pioneering CLIP guided image generation. We would like to thank the anonymous reviewers for their helpful comments, and also thank the participants of the pilot study from the GSIG Discord group.

References

Agnese, J.; Herrera, J.; Tao, H.; and Zhu, X. 2019. A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis. arXiv: 1910.09399.

Bisong, E. 2019. Google colaborator. In *Building Machine Learning & Deep Learning Models on Google Cloud Platform*. Springer.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv: 1809.11096.

Cartuyvels, R.; Spinks, G.; and Moens, M.-F. 2021. Discrete and continuous representations and processing in deep learning: Looking forward. *AI Open* 2:143–159.

Colton, S.; Smith, A.; Berns, S.; Murdock, R.; and Cook, M. 2021. Generative search engines: Initial experiments. In *Proceedings of the International Conference on Computational Creativity*.

Compton, K., and Mateas, M. 2015. Casual creators. In *Proceedings of the International Conference on Computational Creativity*.

Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv:2204.08583*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. arXiv: 2012.09841.

Hertzmann, A. 2020. Visual Indeterminacy in GAN Art. *Leonardo* 53(4):424–428.

Llano, M. T.; dÄôInverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable Computational Creativity. In *Proceedings of the International Conference on Computational Creativity*.

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

Smith, A., and Colton, S. 2021. CLIP-Guided GAN Image Generation: An Artistic Exploration. In *Proceedings of the EvoMusArt conference*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, .; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in NeurIPS*,

Veale, T., and Cook, M. 2018. *Twitterbots: Making Machines that Make Meaning*. MIT Press.