AI Art as a Tool to Explore Bias in Text-to-Image Generation from Natural Language

Valentine Goddard, lawyer, mediator, interarts curator and visual artist
Daniel Harris, PhD Physics, AI Research Director
AI Impact Alliance, Your Impact on the World
January 2021

Introduction

The purpose of this document is to explain how artists can use Text to Image Generation to explore ethical implications in AI such as gender bias. With our current partners, we have raised funds to build an interactive platform which can host the resulting visual art works and publications pertaining to the research findings. In this document, we present a brief overview of some of the potential research questions arising from our work so far, potential applications of this research of AI in industry, leading to the improvement of innovative technologies (Text-to-image Image Generation) and make scientific content accessible to a wider audience.

The Project

The Art + AI platform project includes 1) a platform (content holder and communication tool), and 2) an initial research project "AI Art as a Tool to Explore Bias in Text-to-Image Generation from Natural Language" that will include an art residency, lead to content creation and content publication (visual, sound, research publications, recording of panel discussions, conferences, etc). The research project "*AI Art as a Tool to Explore Bias in Text-to-Image Generation from Natural Language*" is best described as a collaboration leading to the improvement of innovative technologies (Text-to-image Image Generation) and make scientific content accessible to a wider audience.

We have agreed with our partners in France who benefit from a large network of women in the Global South to label with their perspective a short sentence (TBD) in order to compare the image generated. The new data set would be used to create a visual *fresque*, engaging the public and students in understanding through the art piece how data labeling works, what are the ethical issues involved (particularly gender bias) and so on.

Although this would need to be confirmed, our very brief research shows that the current state of the art at this time would be achieved with the CPGAN architecture (Liang et al., 2020). Two potentially interesting issues were identified : **1) Qualifiers** (adjectives) are problematic. How can a machine learn about cultural context, opinions and how could a visual art project assist in advancing this critical question in AI, while engaging a broader public on these questions. **2)** Prominent AI researchers are currently investigating the role of **attention in decision-making** process. How and why does the "attention" select its direction in an image, or in this case, why did the words women, beauty and imperfect result in a pear. Both could result in an engaging and interactive visual/sound art piece. Further multidisciplinary research collaborations could certainly be explored.
As of today, a research proposal focused on the issues identified above is being finalized, the AI Art Residency week-by-week program is in the making, and the interactive platform which would host

visual art work and publications has the funding to get started this month. Our current partners are the Canada Council for the Arts, the Montreal Art Council, Microsoft, Royal Bank of Canada, Isahit (responsible data-labeling based in France operating globally), FEF (Foundation behind Congreso Futuro/Dialogos in Chile) and AI Impact Alliance, an independent non-profit organization I founded in 2017, is the organizer/producer (Professor Yoshua Bengio is the Honorary President of our Advisory Board).

Current State of the Art (SotA)*

While understand there are other techniques[1] that can allow text to image generation, but using GAN based methods prove to be the most promising. Indeed, text can be used to condition an otherwise random image from a Generative Adversarial Network (GAN) in what is known as a class of GANs called Conditional GANs or cGANs for short. Arguably one of the most influential architectures achieving this is the AttnGAN created by Tao Xu et al. (2017). This architecture has become a popular backbone for a number of Text to Image generators since its first publishing. Below we give a quick update on the current state of the art, followed by a brief overview, and finally a discussion highlighting a few interesting questions to pursue.

The current state of the art at this time is achieved with the CPGAN architecture (Liang et al., 2020). Interestingly, the same study also introduces AttnGAN+ which is an improved version of the original AttnGAN architecture that vastly improves results. The results are gauged by the inception score and R-precision metrics where in both cases, higher is better. It is also worth noting that CPGAN achieves this better performance with only about ⅓ of the parameters that AttGan requires. The code for CPGAN using PyTorch has been shared in https://github.com/dongdongdong666/CPGAN and can be compared to the code for AttnGAN which was shared by the original authors in https://github.com/taoxugit/AttnGAN and has since then been widely implemented with over twelve implementations in either PyTorch and TensorFlow (https://paperswithcode.com/paper/attngan-fine-grained-text-to-image-generation).
Interestingly, Liang et al. (2020) also evaluated various models using 100 humans to test the quality of the synthesized images.

**Our project includes research time allocated to a 1) literature review to ensure this is indeed SotA, 2) ensure that the code is available via GitHub to evaluate feasibility within time and budget constraints, 3) work with AI ethics experts, curator and artists to enable knowledge sharing that results in a meaningful art piece. The visual art result should be accessible to a broad public, breaking down a complex issue into engaging interactive visual fresque.*

Overview

AttnGAN

A number of traits of the AttnGAN architecture are important to understand. First, the GAN is initially conditioned on a sentence feature using the process of Conditioning Augmentation explained in an earlier work

[1] *Very recently an Open AI blog has been has been published showing exciting results, but few detail or comparisons to AttnGAN or CPGAN are mentioned, other than this is not a GAN approach but rather a Variational Autoencoder (VAE) approach. https://openai.com/blog/dall-e/?fbclid=IwAR1hspd8_ok_AmaKg2OKfxXJM2JhYKXyZSvFL6cdNfxDNH9J1JVjWxAA4Dk

by some of the authors (Zhang et al, 2017) that converts a sentence vector to a conditioning vector. This generates a low-resolution image conditioned on the sentence as a whole.

The AttnGAN architecture then uses individual word encodings (vectors) to create subsequently higher resolution images from the previous lower-resolution images. Each attention model automatically retrieves the conditions (i.e., the most relevant word vectors) for generating different sub-regions of the image. The architecture is shown in Figure 1.

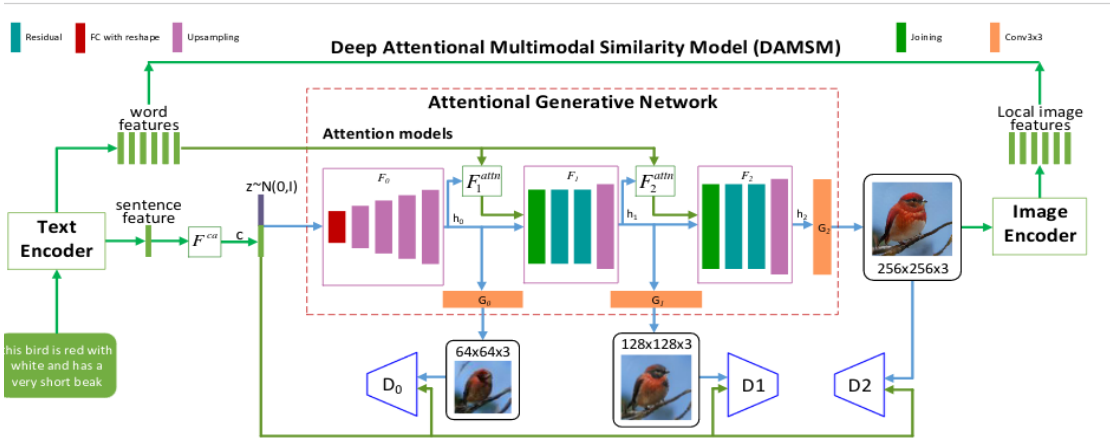


Figure 1. Architecture schematic from original attnGAN paper (Xu et al., 2017) showing how sentence features and individual word features are used.

CPGAN

The architecture of the CPGAN (Liang et al, 2020) is similar to that of the AttnGAN and shown in Figure 2. Both architectures have a word feature to image feature mapping called the Deep Attentional Multimodal Similarity Model (DAMSM) in the AttnGAN and the Text-Image Semantic Consistency Loss (TISCL) in the CPGAN.

While the architectures look similar, the key difference lies in the difference between the TISCL and DAMSM, which in turn lies in the encoding mechanisms for both input text (TextEnc) and the synthesized image (ImageEnc). CPGAN's proposed Memory-Attended Text Encoder and Object-Aware Image Encoder focuses on 1) distilling the underlying semantic information contained in text and image, and 2) capturing the semantic correspondence between them and these are the key differences that give CPGANs superior performance when compared to AttnGANs.

Being more recent, the CPGAN paper also contains a thorough review of other methods and publications as well.

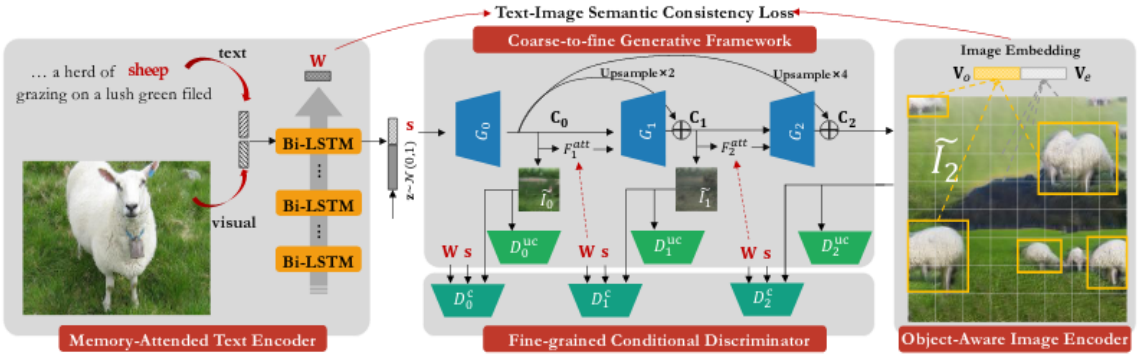


Figure 2. CPGAN Architecture.

Discussion

There are many interesting and highly relevant research questions that require further analysis and exploration. Two initial discussion points are presented below that could lead to further research. The first concerns Opinion Adjectives and arguably constitute a limitation to image generation from text that may require management. The second concerns further attention analysis.

Language - Opinion Adjectives

The first interesting point to note is that language consists of nouns and qualifiers (adjectives) that are not opinion based as machines would have a problem being trained on the subjective nature of opinions. To be clear, consider the accepted ordering of qualifiers in the English language as illustrated in Figure 3. Of the adjective types listed, "Opinion" stands out as being problematic for tying text to images. Consider the statements "A *beautiful* welded joint" with "A *beautiful* red bird". Perhaps extremely large amounts of data could be used to overcome this challenge. However, two people may have completely different opinions on a "*Beautiful* dress" reflecting diversity of taste causing the exactly same image of a dress to be labeled "*Beautiful*" by one person and even "*Ugly*" by another. This diversity of opinions would arguably make it impossible for a machine to learn anything about opinions. These questions could be very interesting to explore from an artistic perspective but we'd need to confirm there are viable solutions to explore from an NL/AI point of view.

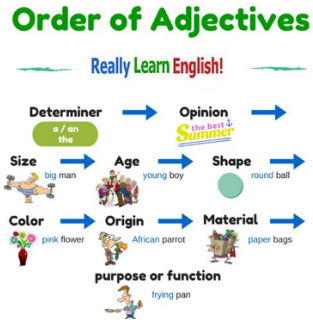


Figure 3. English language proper ordering of words.

Attention Analysis

The second discussion point concerns the appearance that as one uses AttnGAN, one observes the frequent generation of abstract images. It then becomes interesting, knowing how the AttnGAN works, to map areas of attention in the image to specific words in the phrase generating the image. Figure 4 illustrates the idea, although it is troubling that the "very short beak" qualifying statement seems to focus largely on the feet rather than the beak. Some experimentation would be required to determine the robustness of such a visualization. It is also worth noting that the CPGAN method aims to address these artifacts by parsing the content of both the input text and the synthesized image thoroughly and thereby modeling the semantic correspondence between them. On the side of text modality, they design a memory mechanism to parse the textual content by capturing the various visual context information across relevant images in the training data for each word in the vocabulary. On the side of image modality, they propose to encode the generated image in an object-aware manner to extract the visual semantics. The obtained text embeddings and the image embeddings are then utilized to measure the text-image consistency in the semantic space.

Figure 4. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0,G_1 and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively.

Coming back to the original observation of the case where an image is generated that is more abstract as in Ex 2 below, it would be interesting to see, visually, where the attention of the words overlaps with the pear, for instance. Or perhaps the pear is even generated by the combination of all words since one must recall that the first low-resolution image is generated by the entire phrase.

Development and Improvement of Innovative Scientific Technology or Processes

One of the most common and challenging problems in Natural Language Processing and Computer Vision is that of image captioning: given an image, a text description of the image must be produced. Text to image synthesis is the reverse problem: given a text description, an image which matches that description must be generated. In short, the forward problem (image generation from text) helps solve the reverse problem (text captioning).

Scientific Content Accessible to a Wide Audience with AI Art

First, the potential applications for real world problems and industry are broad and show important potential for creative industries.

- Visual Question Answering is used to ask a question/query in natural language and receive desired information from an image which is related to creating a caption based on an image (visual caption generation) to facilitate search through written commands.
- Interactive entertainment.
- Accessible AI-Assisted design (video games, fashion, animation, home decor, etc) where Image Editing using Natural Language commands to change qualifiers. Ex: change short red pants for long green pants.
- Improves document accessibility where, for instance, visually impaired benefit from image captioning which is the reverse of text to image generation however benefits from the latter.

Second, if we consider the general public to be part of our wider audience, the outcoming visual art work that will outcome from this particular project helps citizens become acquainted with artificial intelligence and its potential in an accessible, engaging, and playful, magical way.

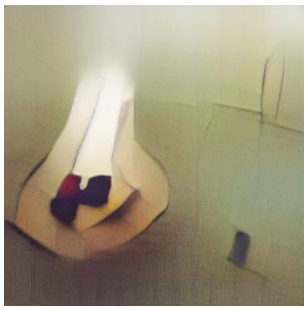
Brief History / Context

I'm a lawyer, mediator, interarts curator and a visual artist, and my expertise is in AI ethics. In 2019-early 2020, with the support of the Canada Council for the Arts, I co-designed the Art Impact AI and held workshops across the country engaging with artists on how AI would impact them, their community, their practice, and how they wanted to impact AI. I published a report with policy recommendations, presented those to the United Nations as part of an Expert Group on Digital Governance, resulting in public policies underscoring of the role of the Arts. I've been invited to speak at various conferences on the role of the Arts in AI Governance including UNESCO and ITU supported conferences. This link offers an overview.

Parallel to the policy work, I used RunwayML's AttGans model to convert the artists' voices per city visited into a unique AI Art piece. Below, for example, was a result of the artists, critical designers, students, professors, community leaders, voices at the last in-person workshop, Vancouver 2020.

That work led to further explorations on how I could use AI Art to teach data/AI ethics as I noticed sometimes that the intention was surprisingly well interpreted, and others the intention was completely missed. This research and art-science collaboration would improve the understanding of we can better control the outcomes, which touches on complex issues in Text-to-Image Generation (see last paragraph).

Ex 1. The text that generated this image was on the role of the art in shaping democratic laws and policies on AI governance. The text's intention was empowering, inspiring but the "attention" seems to have been put on the civil society and it being interpreted as civil unrest.



Ex 2. This one was a result of the words: women, beauty, imperfect resulting in an abstract pear-looking visual art piece.

The Project Overview

The Art + AI platform project includes 1) an interactive web-based platform (content holder and communication tool), and 2) an initial research project "AI Art as a Tool to Explore Bias in Text-to-Image Generation from Natural Language" that will include an art-science residency, lead to content creation and content publication (visual, sound, research publications, recording of panel discussions, conferences, etc). Depending on the pandemic situation, this will be showcased at the annual AI on a Social Mission Conference.

The research project **"*AI Art as a Tool to Explore Bias in Text-to-Image Generation from Natural Language*"** is best described as a collaboration leading to the improvement of innovative technologies (Text-to-image Image Generation) and make scientific content accessible to a wider audience. In short, we would work with our international partners who have a large network of women in the Global South to label with a women's perspective a short sentence in order to compare the image being generated using a short sentence. Although this would need to be confirmed, our very brief research shows that the current state of the art at this time would be achieved with the CPGAN architecture (Liang et al., 2020).

Two potentially interesting issues were identified : **1) Qualifiers (adjectives)** are problematic. How can a machine learn about cultural context, opinions and how could a visual art project assist in advancing this critical question in AI, while engaging a broader public on these questions. **2)** Prominent AI researchers are currently investigating the **role of attention in decision-making** processes. How and why does the "attention" select its direction in an image, or in this case, why did the words women, beauty and imperfect result in a pear. Both could result in an engaging and interactive visual/sound art piece, a process identified as socio-technical mediation, or cultural mediation, if we see it as the process of engaging viewers in understanding what the artists are expressing through AI Art. Further multidisciplinary research collaborations could certainly be explored.

References

Xu, T., Zhang, P., Huang, Q.,Zhang, H., Gan, Z., Huang, X., He, X., *AttnGAN: Fine-Grained Text to Image Generationwith Attentional Generative Adversarial Networks*, https://arxiv.org/pdf/1711.10485.pdf, 2017

Liang, J., Pei, W., Lu, F., CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-ImageSynthesis, https://arxiv.org/abs/1912.08562, 2020

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D., Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D., Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916, 2017

Bodnar, Cristian, under the supervision of Dr Jon Shapiro, Text to Image Synthesis Using Generative Adversial Networks, https://arxiv.org/pdf/1805.00676.pdf, 2018

Li, B., Qi, X., Lukasiewicz, T., Torr, P., Controllable Text-to-Image Generation, Oxford University, https://papers.nips.cc/paper/2019/file/1d72310edc006dadf2190caad5802983-Paper.pdf, 2019