Raising awareness about the climate crisis; an exploration of Machine Learning-based image generation

CP194 - Capstone Directed Study II

Carlos Rafael Garduno Acolt

Spring 2022

Table of contents

Executive Summary

Raising awareness about the climate crisis; an exploration of Machine Learning-based image generation

Carlos Rafael Garduño Acolt *Minerva University; Class of 2022*

Background and Justification: A global crisis

Humanity's exploitation and extensive disturbance of the Earth's biosphere threaten to shift the state of the global system towards a drastically different one. A state of the ecological system which will not be capable of sustaining biodiversity at the scale humanity just begins to register and understand (Vitousek, Mooney, Lubchenco, & Melillo, 1997; see also Barnosky et al., 2012).

Our generations have the challenge and responsibility of dealing with the climate crisis before it is too late. Humanity's efforts must be radical and sustained over the next several decades to restore the ecological balance of the Earth (Herring, & Lindsey, 2021).

Methods: Computer-generated art as a tool for science communication

Art has the potential to engage and nudge people to learn more about a topic by appealing to the affective domain of learning (i.e., engagement, attitude, or emotion) instead of the cognitive domain (i.e., understanding, comprehension, or application), usually targeted in traditional learning. (Lesen, Rogan, & Blum, 2016).

Technology has shaped human cognition throughout history. Industrialization and digitalization have not only physically separated humans from nature, but also intellectually (Kesebir, & Kesebir, 2017). Current societies conceive other living systems as objects of exploitation rather than beings deserving of respect as a consequence of colonialism (Flores, 2019).

Computer-generated art is becoming increasingly popular because of the philosophical implications it brings in regards to the definition and purpose of art as well as because of its signature aesthetic (Nugent, 2018).

DSTRTD: the project

Machine Learning methods mimic human cognitive processes including creativity. The program behind the artworks "learns" concepts (patterns in data) and creates novel representations of said concepts with a similar structure to the way humans learn and produce art.

Portraying nature through the lens of a computer is a metaphor for the way in which our perception of nature has been distorted by our growing dependency on digital technology while experiencing reality.

Each artwork is motivated by a different consequence of climate change, and is meant to exploit the cognitive dissonance of their signature "distorted realism" style, nudging viewers to engage and learn more about the piece, its purpose and context.

References

Barnosky, A. D., Hadly, E., Bascompte, J., Berlow, E. L., Brown, J. H., et al. (2012) Approaching a state shift in Earth's biosphere. Nature 486: 52–58

Flores, C. (2019). The disconnection between humans and nature. Center for humans and nature. Retrieved from: https://www.humansandnature.org/what-happens-when-we-see-ourselves-as-separate-from-or-as-a-part-of-nature-the-disconnection-between-humans-and-nature

Herring, D., & Lindsey, R. (2021). Can we slow or even reverse global warming? National Oceanic and Atmospheric Administration. Retrieved from: https://www.climate.gov/news-features/climate-qa/can-we-slow-or-even-reverse-global-warming

Kesebir, S., and Kesebir P. (2017). How Modern Life Became Disconnected from Nature. Greater Good Magazine. Retrieved from: https://greatergood.berkeley.edu/article/item/how_modern_life_became_disconnected_from_nature

Nugent, C. (2018). The Painter Behind These Artworks Is an AI Program. Do They Still Count as Art? Retrieved from: https://time.com/5357221/obvious-artificial-intelligence-art/

Lesen, A. E., Rogan, A., & Blum, M. J. (2016). Science Communication Through Art: Objectives, Challenges, and Outcomes. Trends in Ecology & Evolution, 31(9), 657–660. doi.org/10.1016/j.tree.2016.06.004

Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. (1997). Human domination of Earth’s ecosystems. Science 277, 494–499

Introduction and Background

Avoiding hysteresis at the global scale

Biological systems across scales have the potential to change from one state to a radically different one when critical thresholds are reached. The ultimate effect of a threshold-induced state shift is unidirectional. Once a system has experienced a critical transition from one state to another, it is extremely difficult or impossible for the system to return to its original state (Scheffer et al., 2009). Such unidirectionality of a state shift is often referred to as hysteresis in environmental science. A system is said to have undergone hysteresis when the reverse path from one state of the system to another is not the same as the forward path, often requiring disproportionately more energy and resources for the reverse shift to occur. In other words, the "work" that must be done to return to an original state after a disturbance is much more than the "work" done by the original perturbation itself (Beisner, 2012). The critical thresholds mentioned before are often reached by the cumulative effects of disturbances acting on the system.

Humanity's exploitation and extensive disturbance of the Earth's biosphere threaten to shift the state of the global system towards a drastically different one. A state of the system which will not be capable of sustaining biodiversity at the scale humanity just begins to register and understand (Vitousek, Mooney, Lubchenco, & Melillo, 1997; see also Barnosky et al., 2012). Such a state transition implies hysteresis over the global system. Once the cumulative effects of human disturbances on the Earth's ecosystems reach the system's threshold, it will be extremely difficult if not impossible to shift back to the original state.

The Earth's biosphere is no stranger to planetary-scale critical transitions and state shifts, which means such events can be triggered again in the future. The most recent and one of the

fastest planetary state shifts was the transition from the last glacial to the current interglacial condition:

> The critical transition was a rapid warm–cold–warm fluctuation in climate between 14,300 and 11,000 yr ago, and the most pronounced biotic changes occurred between 12,900 and 11,300 yr ago.
>
> The major biotic changes were the extinction of about half of the species of large-bodied mammals, several species of large birds and reptiles, and a few species of small animals; a significant decrease in local and regional biodiversity as geographic ranges shifted individualistically, which also resulted in novel species assemblages; and a global increase in human biomass and spread of humans to all continents.
>
> The pre-transition global state was a glacial stage that lasted about 100,000 yr and the post-transition global state is an interglacial that Earth has been in for approximately 11,000 yr. The global forcings were orbitally induced, cyclic variations in solar insolation that caused rapid global warming. Direct and indirect effects of humans probably contributed to extinctions of megafauna and subsequent ecological restructuring. (Barnosky et al., 2012. p. 53)

Other critical transitions on longer timescales include four of the ‘Big Five’ mass extinctions, each happening over the course of tens of thousands to a couple of million years. The global forcings of those extinctions were unusual climate changes and shifts in ocean and atmospheric chemistry, particularly due to extreme concentrations of carbon dioxide and hydrogen sulfide.

In spite of the diversity of timescales among the mentioned transitions, they all occurred very quickly relative to their bracketing states (less than 5% of the time the previous state lasted).

Furthermore, all of the transitions were connected to global-scale changes in the atmosphere, oceans, and climate (Barnosky et al., 2012). While forecasting biological change at the global scale, it is crucial to know if the magnitudes of present global forcings are sufficient to trigger a new critical transition. According to Steffen et al. (2011), current global-scale forcings are human population growth and consequent increase in resource consumption, habitat transformation and fragmentation, energy production and consumption, and climate change.

Alarmingly, the magnitude and rate of current forcings exceed that of the triggers for the last global critical transition. For instance, human settlements and activities have transformed over 43% of the Earth's surface for agricultural and urban use (not taking road systems into account), which exceeds the 30% of the surface that went from ice-corevered to ice-free during the last critical transition. Quantifying land use can be utilized while anticipating planetay state shift (See Figure 1).

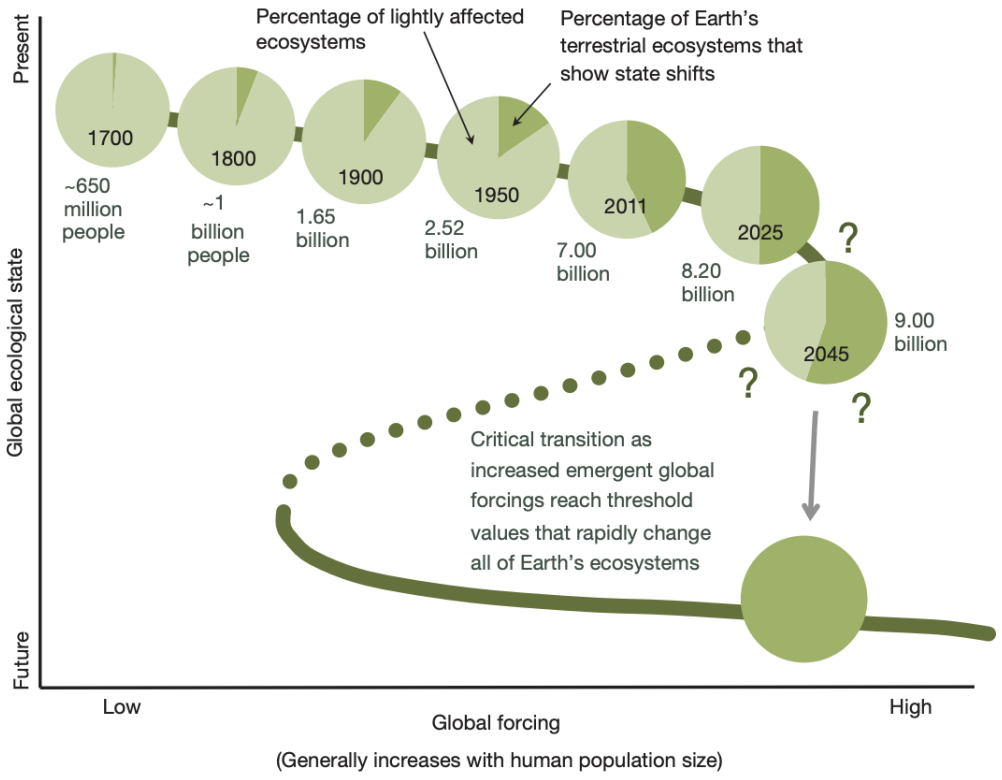


Figure 1. Adapted from Barnosky et al. (2012). The trajectory of the green line represents a fold bifurcation with hysteresis. At each time point, light green represents the fraction of Earth's land that probably has dynamics within the limits characteristic of the past 11,000 yr. Dark green indicates the fraction of terrestrial ecosystems that have unarguably undergone drastic state changes; these are minimum values because they count only agricultural and urban lands. Landscape-scale studies and theory suggest that the critical threshold may lie between 50 and 90% of land disturbed.

It is practically impossible to predict an exact tipping point because of the complexity of the plsnetary system. Nevertheless, evidence from smaller-scale ecological systems indicate that current global forcings may be close to, if not already have started to, induce a critical state shift at the global scale. According to Barnosky et al. (2012), landscape-scale studies and theory suggest a disturbance of around 50 to 90% of the Earth's surface will induce a critical transition, range which we will reach within the next decade if transformation rates remain similar.

The effects of climate change have become increasingly evident over the last couple of decades. Such effects include; rising maximum and minimum temperatures worldwide; more frequent and severe weather events; the shrinking of glaciers and thawing of permafrost; water, air, and soil pollution; and the rise of sea levels and ocean temperatures. Such effects directly impact Earth's ability to host life (including humans) which becomes apparent when we consider the increasing rates of wildlife extinction across all climates and biomes (Denchak, & Turrentine, 2021). A specific example of observable biotic responses (symptoms of a critical transition) to large-scale disturbance of the biosphere are the vast "dead zones" in near-shore marine ecosystems, areas of water bodies where aquatic life cannot survive because of low oxygen levels (US Environmental Protection Agency, 2022). Another is the drastic and rapid loss of biodiveristy in more than 40% of the Earth's landscapes, replaced with landscapes inhabited by few crop plant species, domestic animals, and humans.[1]

[1] #disturbances see Appendix A

Science communication; a driver of political action

The environmental crisis is not a politically unknown issue on a global scale. In fact, the World Economic Forum categorized the failure to mitigate and adapt to climate change as the "most impactful" risk facing communities worldwide in its Global Risks Report 2021 (World Economic Forum, 2021). Nevertheless, humanity's efforts to slow the rate and limit the amount of global warming must be radical and sustained over the next several decades (Herring, & Lindsey, 2021).

A situation of such complexity requires a multidisciplinary solution. Scientific debates in the modern world have and must continue to blur the lines between the pursuit of knowledge and the political, moral and legal implications of discovery. There have been many instances in the past when scientists successfully worked as political advocates and advisors. Such as the case of Albert Einstein's communication to US president Franklin D. Roosevelt in 1939 about the urgency of accelerating academic research of nuclear chain reactions, leading to the Manhattan Project. Or the case of Nobel prize winner Richard Smalley in the mid-1990s, who openly lobbied the US congress to establish a fund for the National Nanotechnology Initiative, leading to a multibillion-dollar investment by the US government for research in the area (Scheufele, 2014). Nevertheless, such overlaps between natural and political science can only happen when there is a clear and easily tractable path of action, and when the scale and complexity of the instance are not an obstacle. Scientists and scientific discovery are not always enough to drive necessary political action, especially for highly-complex issues like climate change, with several politically and economically powerful stakeholders and infinite room for conflicts of interest. Solving climate change is not as easy as changing the mind of one or a group of politicians. An effective, long-term solution requires local, regional, and global action, which can only be

achieved if people in all communities worldwide are aware of the causes, effects, and steps needed to slow down and stop human-caused global warming.

Teaching communities about climate change is not only necessary for making more people take action in their daily lives and within their communities, but also while developing strong and resilient political measures across scales in order to yield significant rapid change. An effective long-term solution requires a deep-rooted cultural shift. A shift that sets the stage for building an environmentally aware, sustainable future, and making sure the next generations do not commit the mistakes of the past.[2] [3]

Educating through art

Art is a powerful tool for science education and communication. An artpiece can have the potential to engage and nudge people to learn more about a topic by appealing to the affective domain of learning (i.e., engagement, attitude, or emotion) instead of the cognitive domain (i.e., understanding, comprehension, or application), which is usually targeted in traditional learning. (Lesen, Rogan, & Blum, 2016). In terms of the acclaimed book *Thinking, fast and slow*, art appeals to system I of thinking, intuition, rather than system II, reasoning. As Kahneman (2011) argues, we systematically overestimate the role of rationality for decision making. People take decisions impulsively more often than we are aware of, which is why appealing to system I is so powerful while persuading and nudging individuals into taking action. While pursuing a cultural shift as the one necessary to solve the climate crisis, it will not be sufficient for people to know the effects of climate change, but they need to believe there is a structural change that needs to be made.

[2] #socialframes see Appendix A
[3] #rightproblem see Appendix A

About the project

Relating to nature

One of the negative consequences of modernity is the growing disconnection between humans and nature. The trend of urbanization has not only physically separated us from it, but also intellectually. As Kesebir, and Kesebir (2017)'s research suggests, humans have become more and more distanced from nature ever since the 1950s. The team analyzed cultural products and assessed how frequently artists referenced nature in their works. More specifically, they created a list of nature-related words and registered the frequency of these words in works of popular culture over the last century. The data set included millions of English fiction books written between 1901 and 2000, thousands of songs listed as the top 100 between 1950 and 2011, and hundreds of thousands of storylines of movies and documentaries made between 1930 and 2014. According to their results, nature terms are significantly less frequent in popular culture products today than they were before the 1950s. Nature-related words represented around 1% of all words in the works before 1950, and only 0.5% of all words among more recent works (after 1980).

Besides, the disconnection between humans and nature certainly started way before the 1950s. Current western generations might consider themselves as separate from nature as a consequence of Christianity and colonization. Major Christian religions in the west developed the concept of the Earth belonging to humankind rather than humans being part of it, understanding which still permeates western cultures today. As a consequence of capitalism and globalization, these ideas have spread across the globe separated from religion, making this a worldwide phenomenon. The conceptual division between the natural world and humans is cemented in globalized culture, and even present in our language. Modern humans tend to

objectify ecosystems and other living beings. Conversations about nature across sciences tend to understand it as something far-removed from humans and miss the fact that for the scale of the Earth, we are just another species coexisting on the planet, rather than the planet's owners or masters (Flores, 2019). Nature, in many cases, has become something we possess and exploit rather than respect and explore.

The concept of private property in terms of land, ecosystem services, and living beings is not something that was practiced worldwide before colonization. For instance, in pre-colonial America, many of the native peoples did not have the concept of private property. In many cultures, one could not own the natural world the way we do today, and other living beings (e.g. animals, plants) were seen as equals. Humans relied on other living beings and ecosystems services for survival, but rituals involving the attainment of natural resources tended to personify rather than objectify nature. An example of this is hunting and the development of settlements, societies developed rituals where people would ask for permission to the spirit of the animal or place, or a deity related to it, before killing and using natural resources. The personification of nature allowed people to relate to the Earth as one does to a human being. Consequently, this led in many cases to the development of sustainable living practices, where societies existed and developed respecting, celebrating, and taking care of the ecosystems they were part of (Flores, 2019).

Certainly, modern societies do not need to personify nature the way pre-colonial societies did in order to empathize with it. We do not need the creation and spread of systems of belief rooted in animism to understand the urgent necessity to improve the ways we interact with nature. Nevertheless, the cultural shift needed does require societies to understand our existence and survival as dependent on the state of the Earth's biosphere.

The disconnection between people and nature might make it difficult for them to relate to the environmental crisis and realize how impactful its consequences will be to the way humanity experiences life over the upcoming decades (if not sufficient action is taken). Such observation sparks the purpose of the project, which is to appeal to the audience's emotions and nudge them into questioning their connection to nature and how it may be distorted by their postmodern lifestyle and dependency on digital technologies. Hoping that questioning triggers their curiosity and further interest in learning about nature and how to improve their relationship with it.[4]

[4] #purpose see Appendix A

Target audience

According to Lesen, Rogan, and Blum (2016), the intended target audience for most arts-based science communication projects in the US is the general public, followed by a professional audience, and university students. Such a trend makes sense when we consider our current understanding of the importance of social cues while making decisions. As psychologist Solomon Asch (1956) demonstrated through his seminal experiments on conformity, an individual's social environment has a powerful influence on their judgments and opinions. In the experiments, the researcher asked participants to judge the relative length of a line compared to other three lines, one of which had the same length. The participants were set to take that decision in front of other people who pretended to be participants and unanimously agreed that the line of equal length was one of the incorrect ones. Asch found that participants were less likely to pick the obvious right answer when experiencing the pressure of their answer going against the majority's. Participants tended to conform to other people's judgements due to the discomfort associated with opposing the majority's opinion. This is another example of how appealing to people's system 1 of thinking is so effective for persuasion, which directly informs our choice of a target audience. In order to trigger action at the scale needed, it is fundamental to motivate as many people as possible. The more people interested in fighting climate change, the greater the social pressure onto the rest of the population, including crucial political actors. Moreover, many people may already be interested but may not take action nor raise their voice because of conforming to a majority of the local population who do not think climate change is an urgent issue.

Consequently, the intended audience of this project is the general public, since we mean to reach as many people as possible. The final product is meant to appeal to the curiosity of

anyone who interacts with the pieces. Given that we are tailoring the deliverable to a general audience, we will assume the viewer has little to no knowledge about computer-generated art. The viewer has heard about AI, but most likely does not understand how a computer can generate a realistic image. In regards to the ecological subjects, we will assume the audience is able to distinguish the images presented (the audience knows how a forest or a coral reef look like), but is not able to describe the ecological importance of the subject.[5]

[5] #audience see Appendix A

Presenting the Artwork

The final product will most be displayed both through physical and online instances in order to reach the biggest audience possible.

Online

The artworks can be displayed in an Instagram account created specifically for the project. Given that the final deliverable will include several artworks, they could be organized into several posts. Each post will include the image, a short reflection related to the piece, and the context of how the image was generated. The account could also include posts describing the GAN process in more detail, as well as links to other pieces of climate change-related science communication for people who wish to learn more. Setting up an Instagram account will also allow for data collection regarding the number and kinds of interactions the audience has with the works. Once the account is set up, the content could stay there indefinitely.

Physical

The artworks can be printed into posters which can then be taped on walls as street art. Note that the placement of the artworks will follow the location's cultural norms and practices. More specifically, it will be fundamental to make sure the placements of the artworks are not unauthorized locations. As a rule of thumb, the posters will only be located on walls where other pieces of street art/posters are already placed. If possible, some of the posters will be placed in/near green spaces. We expect people to engage with the poster, read the online reflection and learn about the project while being

surrounded/close to nature, which will hopefully further nudge them to reflect on their perception and relationship with it.

The project will take advantage of the diversity of cities where Minervans are located in order to maximize the works' reach, making it possible to distribute posters in (at least) London, Berlin, San Francisco, and Seoul.

The posters will only present the artwork and a QR-code leading to the Instagram account previously described, this in order to create a sense of mystery around the pieces which nudges people into scanning the code.

Locating the works in public spaces (online and offline) will allow us to reach something closer to a truly general audience than the audience we could reach by presenting the works in a gallery or any other private space. Although there can be an immersive component to the experience (when the person reads the reflection and context of the artwork in/close to a natural space), this is intended to be a passive experience for the audience. Nevertheless, it will be interesting and relevant to explore ways to make this and future similar experiences interactive (outside of Minerva), as current research supports multiple benefits of interactive over passive art experiences for science communication (Lesen, Rogan, & Blum, 2016).

Interpretation of the pieces

The images are meant to present the natural world through an explicit digital distortion as a metaphor for the way in which our perception of nature has changed due to industrialization and digitalization. We expect the audience to be able to infer the works are images of nature but to also be able to tell they are distorted, this is also meant to capture the attention of bypassers, hoping they interact with the pieces. Once a viewer scans the QR-code, they will read the context of the artwork and a reflection about human perception and relationship to nature. By being a computer-generated representation, the viewer is prompted to question the ways technology shapes and informs their perception and interactions with the natural world.

Learning Journal: ML-based image generation

Computer-generated art; an overview

Since the rise and popularization of Machine Learning methods, their applications have extended across practically all fields. Art and design have not been the exception, both scientists and artists have been exploring computers' ability to generate creative products just like humans do. In an era where humans are every-time more closely in contact with technology, it is imperative to ask: How might we exploit Machine Learning frameworks and methods while working on creative processes such as the creation of a piece of art? Such an issue is too broad for the scope of this project. Therefore, we will narrow it down from the creation of art in general to the creation of 2D pieces of visual art, a theme that has become quite common in art galleries around the world over the last few years.

Making art is far from easy, it involves the artist using their knowledge, past experiences, and novel ideas to design a new piece. Therefore, it is practical to approach the creative process from a human-based perspective. In other words, we can develop computational processes that learn from the way humans make art and then recreate that on their own. In other words, a Machine Learning pipeline able to 'create' art, will first have to go through a training 'learning' process through which the system learns to recognize patterns (data distributions) 'concepts,' just like a human artist would do (Garduno, 2020).

While navigating the intersection between Machine Learning and visual art, one begins to wonder how computer-generated works compare to human-made ones. Elgammal et al. (2017) shed light on this regard with a study in which hundreds of pictures (some of them human-made and other computer-generated) were shown to a group of people. Each participant was prompted to analyze each artwork and rank the pictures on aspects such as 'novelty' and 'complexity.' The

research team assumed human-made works would rank higher in such categories. To the team's surprise, computer-generated images often scored higher than human-made ones. Furthermore, participants were asked to tell whether they thought each one of the works was human or computer-made. Participants concluded that humans had created most of the computer-generated works. It is clear, then, that Machine Learning tools have great potential while expanding the reach of human creation and creativity. A great example of this is the work of a group of French artists and computer scientists (under the name Obvious) who in 2018 became the first to sell a computer-generated work in an art auction. The artwork titled *Portrait of Edmond Bellamy* was sold for 432,500 USD in New York City (Ornes, 2020).

First implementation

This section presents the results and analysis of the initial research and experimentation with simple GANs for image generation.

Generative Adversarial Networks

Generative Adversarial Networks (GANs) are Machine Learning systems that are meant to learn and mimic patterns from a given data distribution. The structure of a GAN is quite simple. It is a neural network composed of other two neural networks 'subnetworks' that are in constant interaction with each other. The first network is often called the 'discriminator,' a classification tool that discerns between 'real' pictures from the training set, and 'fake' pictures generated by the second network. The second network is the generator, and as its name indicates, its job is to generate new images following the patterns identified in the training set without actually seeing the inputs (this 'game' or relationship between the subnetworks is the reason why these are called 'adversarial'). The way it works is that the generator starts creating random images (outputs only noise), which are then evaluated by the discriminator, who communicates how real or fake it finds them with the generator. The generator then 'learns', iteration after iteration, what the most prominent patterns and features of the training distribution are, in order to mimic them (this by changing its weights, fine-tuning the generator). Ideally, the process should iterate until equilibrium is achieved. Here, equilibrium refers to the instance when the discriminator cannot tell the difference between the real (training) images, and the fake (generated) versions (Candido, 2020). It is important to mention that this process is not creative per se. The generator has no means nor motivation to build a cohesive and completely novel piece of art just as a human artist would do. Instead, the generative network will just output

images that mimic the existent data from training so well that they cannot be distinguished (by an algorithm) from the validation set (Elgammal et al., 2017; see also Garduno, 2020).

Developing a basic GAN

Regarding my implementation, I developed a GAN using Keras' frameworks and methods for building neural networks. The full annotated code and descriptions can be found in this notebook.

About the dataset: CIFAR-100

Generative Adversarial Networks require large training and validation data sets in order to yield high accuracy. The artworks are meant to nudge the audience into thinking about the ways they perceive and relate to nature amidst the digital era. This is the reason why I decided to use pictures of trees for generating the artworks. The data set used comes from the readily available CIFAR-100 dataset. The set includes pictures of 100 categories. The subset used included 600 pictures of oak trees, 500 for training and 100 for validation. The pictures have low resolution (32x32). The structure of the set is a 10000x3072 numpy array of uint8, where the first 1024 entries contain the red channel values, the next 1024 the green, and the final 1024 the blue.

Analysis of outputs

These are some examples of the generated outputs using images of Oak Trees as input:

The model is successful at generating new images that resemble the shape of a tree (See examples of generates pictures in Appendix B), and after 50 epochs the model measures an accuracy score of 96.8% and 98.4% on real and generated pictures respectively, and a binary cross-entropy score of 0.0637. A cross-entropy score so close to 0 means that the generated

images at system equilibrium are almost indistinguishable from the validation set. Note that the validation set is completely different from the training set in order to minimize overfitting. The pipeline is highly efficient for the presented dataset, which makes sense given the nature of the input. All pictures in the training set are relatively small, and the shapes of trees are quite similar throughout pictures (one does not require too much detail to distinguish the picture of a tree).

The resolution of the generated images, just like the images in training and validation sets, is significantly low. This is practical for the exercise taking into account computational and time constraints (using higher-dimensional representations would require more computational power and running time). The generated images are very similar to the ones in the training and validation sets. Nonetheless, it could be possible to yield better results by increasing the number of epochs of our iteration, which would again increase the required computational power and running time.

The output is a great starting point for the project. Nevertheless, it does not fit the desired quality of output. The resolution is very low, and the size is too small, this can be resolved by curating a new data set instead of using the one found in CIFAR-100. Although this could be challenging, such exploration will be necessary in future iterations. Furthermore, it would be interesting to curate other kinds of training and validation sets. For instance, a mix between tree/nature related pictures and technological devices/modern architecture.

Research Stage

After analyzing the outputs of the first implementation, it was clear that fundamental research on GANs was missing for the project's completion. This section presents a summary of the research stage, which informed the direction of the project into the second implementation.

Training a Generative Adversarial Network

GANs are difficult to develop and train because of their adversarial nature. The efficacies of the generator and discriminator exist in a zero-sum game, where improving the performance of one neural network worsens the performance of the other. More specifically, the generator is meant to have a consistent output that assimilates the training distributions of the discriminator. On the other hand, the discriminator is meant to identify fake and real data. The more consistent and similar to training data the generated images are, the worse the performance of the discriminator will be. The less consistent and different from training data the generated images are, the better the performance of the discriminator.

Therefore, when describing the optimal performance of a GAN one should focus on finding a point of equilibrium where the performance of the larger system is optimized and individual global maxima are disregarded. In other words, finding convergence. Unusual convergence is referred to as failure modes. Once the larger model is defined, the error calculated by the discriminator is used to train the weights of the generator (Brownlee, 2019).

Measures of performance

Generator

One can assess the performance of the generator by keeping track of the values of a loss function across iterations during training. Stable GANs tend to measure a training generator loss between 0.5 and 2.0.

Qualitative measures:

Nearest Neighbors: Used to detect overfitting. Samples are shown next to their nearest neighbors in the training set and their similarities are analyzed with Euclidean distance. A concern is that Euclidean distance is very sensitive to minor perceptual disturbances. Therefore, two images can be nearly identical and have large distances between them.

Rating and Preference Judgment: The most used for generator qualitative measurements. Human subjects are asked to classify and compare real and fake images from the GAN's output and training set. Judges are presented with one of each and asked to say which one is more realistic. A disadvantage is that subjects can learn over time to identify the fake images, so the metric does not stay fixed over time (Brownlee, 2019).

Quantitative measures:

Average Log-likelihood: Kernel density estimation. It estimates the density function of a distribution from samples. It measures the likelihood of the true data under the generated distribution from samples of the output. It is very intuitive, if a model has maximum likelihood, the generated images must be perfect reproductions of the training set. A

central drawback is that it requires large sample sizes of generated images. The required sample size increases with the number of dimensions, making it an unfeasible measure for complex datasets (e.g. dealing with colors).

Inception Score: The most common quantitative measure for GANs. It uses an external neural network to learn the desirable properties for generated samples. Namely, a highly classifiable and diverse with respect of class labels set. It predicts the probability of the generated image to belong in each of the network's training classes. Nevertheless, the measure is unable to detect overfitting, it does not hold out a validation set, it is affected by image resolution, and fails to detect if the model is stuck in a bad mode. Moreover, the fact that it the external network is trained in a specific object set leads to the measure being a better indicator of the image portraying an object rather than the desired distribution.

Frechet Inception Distance: An improvement on the inception score. It applies the Frechet distance to measure the distance between two distributions. Each distribution is a multivariate Gaussian which summarizes the salient features of the real and generated images. It performs well in terms of discriminability, robustness and computational efficiency. However, it assumes the data follows the characteristics of a multivariate Gaussian distribution, which in many cases is not true (Borji, 2018).

Discriminator

Similar to the generator, we can assess its performance accuracy with a loss function for each real and fake images. Stable GANs tend to measure a training discriminator loss of 0.5 and up to 0.8, and an accuracy of around 0.7 or 0.8.

GAN

The most reliable way to evaluate the performance of the larger model is to subjectively assess the output. At the end of the day, the purpose of a GAN is to generate a novel data distribution that is indistinguishable from a family of distributions (used for the discriminator's training). Given that this process "simulates" mechanisms of human cognition, the assessment of the final output will favor from visual examination. Nevertheless, visual evaluation can be expensive, biased, difficult to reproduce, and might not fully reflect the performance of the system (Borji, 2018).[6]

[6] #modelmetrix see Appendix A

Choices for training

During the training of the GAN, it is up to the developer to set the number of epochs to run, as well as the bach size of real and fake images to use. The bach size will in turn determine the number of batches (model updates), depending on the total number of images available.

The quality of generated images is not correlated with the number of iterations in the training. The quality of images will reach its maximum when the larger model reaches stability. Further training, once stability is reached, can lead to problems like overfitting which will decrease the quality of the output.

Therefore, at the beginning of training, we expect the quality of output to increase after each epoch. Nevertheless, once the system reaches equilibrium, more training epochs could decrease the quality of output. Besides, we expect generated quality to vary while the system is converging, making visual examination or other metrics during this stage fundamental (Brownlee, 2021).

Improving the model:

For my GAN implementation, I followed the training structure of basic GAN examples I found while learning how to use Keras and TensorFlow. There was no rationale behind the number of epochs and batches other than minimizing running time given that I ran the experiments locally. After learning more about training alternatives and measures of performance, I see how the implementation is missing an analysis of model performance through training. My implementation only considers the discriminator's loss function value and accuracy at the end of the last epoch. Instead, a revised implementation should keep track of both the generator and discriminator's loss function values and accuracy across epochs in order to make sure model convergence is reached, as well as while informing the sampling of generated images for display.

Failure Modes

Mode collapse

When the generator only leads to one or a small number of different output distributions. This can be identified when the generated images show little diversity, where the same output distribution (or very similar ones) repeats multiple times. It can also be inferred by analyzing the discriminator, and especially the generator loss function through time. A system may experience mode collapse when the loss functions show great oscillations. In other words, it is very good at discerning between real and fake data for some instances and very bad for others.

A system can experience mode collapse when we overly restrict the size of plausible outputs to generate. This will directly impair the performance of the generator, and improve the performance of the discriminator (Brownlee, 2021).

Convergence failure

The most common failure when training a GAN. It occurs when the larger model fails to find performance equilibrium between the competing neural networks. This is identifiable when the loss function of the discriminator stays close or equal to 0, or when the loss of the generator keeps increasing. Convergence failure often occurs when the generated distributions are very different from each other in regards to the features that should be similar, making it very easy for the discriminator to tell which distributions are real and which are fake. Some unstable models will experience convergence failure for

some batches or epochs, then converge and then fail again. There are many possible causes for a convergence failure, including insufficient capacity, and too large or too small kernel sizes (Brownlee, 2021).

Exploring other GAN architectures:

CycleGAN: Mostly used for the transformation between images with different styles. The model is able to learn how to map the features of one image or data distribution into another. More specifically, the system learns the mapping function between two distributions and the inverse using an Adversarial loss. For example, transforming images of horses into zebras. It is the architecture behind FaceApp.

StyleGAN: This architecture is able to generate high-resolution images. Its central feature is a stack of layers where the initial ones create low-resolution representations, which are then enhanced by following layers, gradually increasing resolution. The model remembers features of instances from the training set and adds noise to the output at each layer, increasing the final resolution.

PixelRNN: An auto-regressive generative model. It learns an explicit data distribution while a regular GAN learns from an implicit probability distribution. The model is able to predict the pixels of a full image from a portion of it. The model starts with an occluded image (half of it), after training it outputs plausible completions fot the original image (the entire picture).

Text-2-image: The model generates meaningful images based in explicit textual descriptions of them. The process starts with a textual representation (e.g. flower with small pink petals), which is transformed into text embedding, concatenated with a noise vector and given as input to the generator. The discriminator nor only distinguishes

between real and fake pictures but it also computes the likelihood of the generated image fitting the input text description (Shibsankar, 2021).

I chose to use the most basic GAN structure for the ease of implementation, given that there are several online resources on the topic, making troubleshooting faster. It would be interesting to experiment with architectures which allow for better resolution of the generated images. One of the problems with the current implementation is that the resolution of the output is very low, which does not allow make sense for the kind of work I intend to create. Looking more deeply into other architectures is a place to start while choosing an architecture able to generate the kinds of images I intend to.

Relevant existent projects:

Edmond de Belamy (2018)

Obvious (collective)

Training data: The algorithm was trained on a set of 15,000 portraits from the online art encyclopedia WikiArt. The training works spanned the 14th to the 19th centuries (Nugent, 2018).

GAN architecture: They used a Deep Convolutional Generative Adversarial Network (DCGAN)

Measure of performance: Nearest Neighbor through Euclidean distance (Barrat, 2018)

Context and reaction to the piece: The piece was printed and sold for $432,500 in an auction. The image looks clearly like a painted portrait. Nevertheless, there is clearly something "wrong" about it. The strokes seem to come from a perfect grid, which is not the way human-made paintings look like. The piece was involved in controversy because of the philosophical questions it raises:

> Despite some uncertainty around calling it art—including from Lloyd—the team has been surprised by the level of interest in their work. "One thing about our art is that nobody is indifferent to it," Fautrel says. "People either love it or really like it, or basically hate it. But nobody says whatever (Nugent, 2018).

Namely, whether it should be considered a piece of art or not, given that it was technically not entirely "created" by a human or a "conscious" being. Obvious (the collective behind the work) believes the image should be considered art, comparing the technique to early photographs:

> Caselles-Dupré [who is responsible for most of the ML implementation] insists they're not viewing AI as a mass-producing replacement for human artists. He compares today's experiments with AI to the dawn of photography in the mid 1800's, when miniature portrait artists lost their jobs. "Back then people were saying that photography is not real art and people who take pictures are like machines," he says. "And now we can all agree that photography has become a real branch of art (Nugent, 2018)."

What can I learn from it?

Another problem with my current implementation is where to find a high-quality, large-enough training and validation set. I will look into databases similar to WikiArt. Furthermore, the use of DCGAN seems to be popular for projects whose output presents the "style," I intend the

generated images to have. As well as using the Nearest Neighbor measure to assess the model's performance. Furthermore, the philosophical considerations the Obvious team raises will be fundamental to take into account while thinking about the ways to communicate the process behind the final images and the meaning that can come from such considerations.

Memories of Passersby I (2018)

Mario Klingemann

Training data: A set of thousands of portraits from the 17th to 19th century (Onkaos, 2018).

GAN architecture: DCGANs. However, the architecture is way more complicated, given that it has several GANs working at the same time and feeding each other's training sets.

Measure of performance: Unknown. Nevertheless, it was also based on the same paper as the previous case.

Context and reaction to the piece: This was the main inspiration for me to think of working with GANs for my CS156 final project, which later sparked my interest in this Capstone project. It has the uncanny, almost-realistic style I am looking for in my project. The artwork includes a computer that is constantly running a network of multiple GANs. The artwork is ever-evolving and displays completely new portraits every couple of seconds. According to Onkaos (2018), Klingemann (the artist) wants his work to "understand, question and subvert the inner workings of systems of any kind," particularly being interested in challenging human perception and aesthetic theory.

What can I learn from it?

Given that the code for the piece was also based on the same paper as the previous one, that must be some seminal research in this particular field. It will be very useful to analyze the methods in it, and see what I can apply to my project. It is very interesting to think about the fact that the artwork comes from a network of GANs. The creativity of the work comes not only from the training set, but from the GAN architecture itself. I would like to experiment with different GAN architectures to see the kinds of images I can generate. Furthermore, human perception is a common theme among the interpretation of pieces generated with Machine Learning, which is a fundamental phenomenon I intend my audience to challenge.

Second Implementation

The main problems with the output of the first implementation were that the images were too small and of low resolution, and that the kinds of images that could be produced heavily relied on the specific training sets used.

Instead of improving the simple GAN, I decided to explore other, more complicated, architectures used for image generation. While doing research about DCGANs, I learned about the following architecture, which does not have the first two problems of the first implementation. VQGAN can generate high-resolution images of any size with high diversity of ouput. Still, the possible output heavily relies on the training set. Fortunately, the researchers behind the development of this architecture have pre-trained models for generating landscape images and have made the model weights from their pre-training available in their repository.

VQGAN + Transformer

The discriminator of the first implementation was a convolutional neural network (CNN). Such feedforward architecture has been widely applied to analyze visual imagery. It is characterized by the ability to "take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate between classes" (Saha, 2018). CNNs contain a built-in inductive prior (bias) on the locality of interactions, which means they exploit prior knowledge about strong local (pixels close to each other) correlations within images, allowing them not to learn all relationships between the elements of a sequence, but only the most important for classification. Convolutional approaches exploit the two-dimensional structure of images (where local interactions are important) by restricting interactions between input variables to a local neighborhood defined by the kernel size of the

convolutional kernel. Being biased towards learning local correlations make CNNs efficient for low-resolution images (fewer pixels). Nevertheless, applying a kernel makes the complexity scale linearly with the overall sequence length (total number of pixels), and quadratically in the kernel size (which is usually fixed to a constant), making the method inefficient for high-quality images (Esser, Rombach, & Ommer, 2021).

Transformer networks, on the other hand, are deep learning models which learn about the correlations between elements within an input solely through attention (without inductive bias that prioritize local interactions like CNNs). Here, attention is defined as mimicking human cognitive attention, which enhances some parts of the input data, while diminishing others. Attention is, in turn, translated as the different weighting of relationships between the elements of a sequence (e.g. pixels of an image). Such a difference in learning style makes transformers good at learning long-range interactions (correlations between elements far apart from each other), which ultimately make them highly expressive. Here, we define expressivity as the ability of an architecture to fit a diverse array of functions. However, high expressivity means transformers must learn all possible relationships between the elements of the input sequence, making them highly inefficient for high-resolution image generation, given that sequence length scales quadratically with resolution (Esser, Rombach, & Ommer, 2021). Still, transformers are widely used for other tasks, like text generation (e.g. OpenAI's powerful GPT-3). Furthermore, transformers are able to learn on extremely large datasets when time and space complexities are

not constraints.

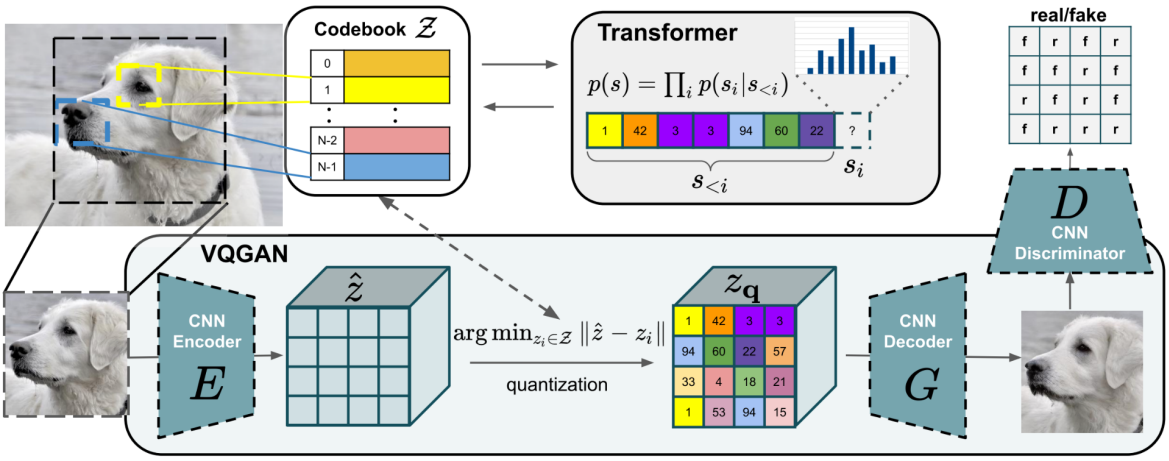


Figure 2. Adapted from Esser, Rombach, & Ommer (2021). Our approach uses a convolutional VQGAN to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

Vector Quantised Generative Adversarial Networks (VQGAN) combine the effectivenes of the inductive bias of CNNs with the high expressivity of transformer networks to generate high-resolution images (larger than 64 x 64 pixels, which is the highest resolution transformers are able to generate otherwise). As Esser, Rombach, & Ommer (2021) explain:

> Taken together, convolutional and transformer architectures can model the compositional nature of our visual world. We use a convolutional approach to efficiently learn a codebook of context-rich visual parts and, subsequently, learn a model of their global compositions. The long-range interactions within these compositions require an expressive transformer architecture to model distributions over their consituent visual parts.

In other words, the transformer network of VQGAN allows the model to understand the global composition of an image, leading to globally-consistent patterns, while the CNN learns the composition of details, making generated output locally realistic.

Moreover, VQGAN is based on the earlier VQVAE model (Vector Quantised Variational Autoencoder). When an image *x* is input into VQVAE, it first goes through an encoder *E* which maps the image onto a sequence of discrete latent variables, producing *E(x). E(x)* is then quantized (hence, Vector Quantised) based on its distance to the code vectors of a codebook (*E(x)* is replaced by the index of the nearest code vector in the codebook). Finally, the discrete codebook representation of *E(x)* goes through a decoder *G* which reconstructs and outputs a generated version of the image. The generated version is then fed into a discriminator network, which then assesses the image with a reconstruction loss score, similar to the simple GAN discussed previously (Thakur, 2021). Therefore, the model follows an adversarial training, where Generator (Encoder-Codebook-Decoder) and Discriminator are trained simultaneously, looking for convergence.

The main structural difference between the VQGAN architecture and the previously implemented GAN is the generation of a codebook, an efficient and rich representation of the images used for training. Building the codebook is the purpose of the first half the training VQGAN requires. At this point, the model is able to generate realistic images. Nevertheless, it runs into one of the same problems the first implementation ran into. The convolutional nature of the encoder and decoder do not allow to model long-range interactions. Namely, the receptive field of every convolutional layer is limited, and cannot generate large images (the system is biased to prioritize local patterns, making it hard to scale in size of the output).

Here is where the transformer is used. During its training, the transformer uses the codebook as a training set, learning to predict the distribution of possible next indices inside the codebook representation. Just like a regular autoregressive model, the transformer builds a regression equation which uses previous time steps as inputs to predict the values of future time steps. VQGAN is able to generate large, high-resolution images because the transformer will not learn from relationships between individual pixels or elements in a sequence (e.g. GPT-3 learns from individual words to generate text), but from a codebook, whose elements are perceptually-rich image constituents (from exploiting the locally-biased learning style of the CNN); extremely compressed data that can be read sequentially by the transformer.

Another advantage of this method over basic GANs is that the size of the images generated is not constrained by the size of the images used in training. When we generate an image (after the second half of training is done; training the transformer) we ask the transformer to generate a novel sequence of elements or “codewords” from the codebook. Given that the transformer predicts indices from previous ones, the size of the generated sequence can have, conceptually, any size. The novel sequence is then fed into the decoder, which turns the codewords into pixel data.

Image generation with VQGAN is often subjected to initial conditioning. Starting from some guide to what we want to generate and asking the model to fill-in the gaps. For example, half of an image, depth-maps, or segmentation maps, among others (see Figure 3).

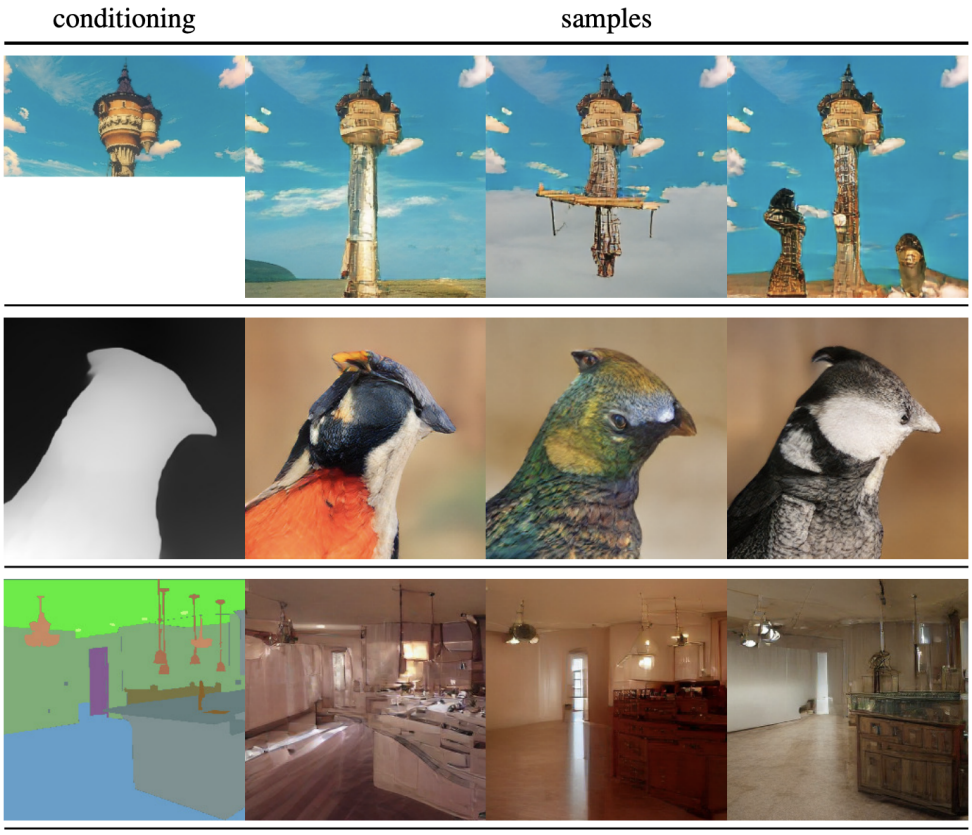
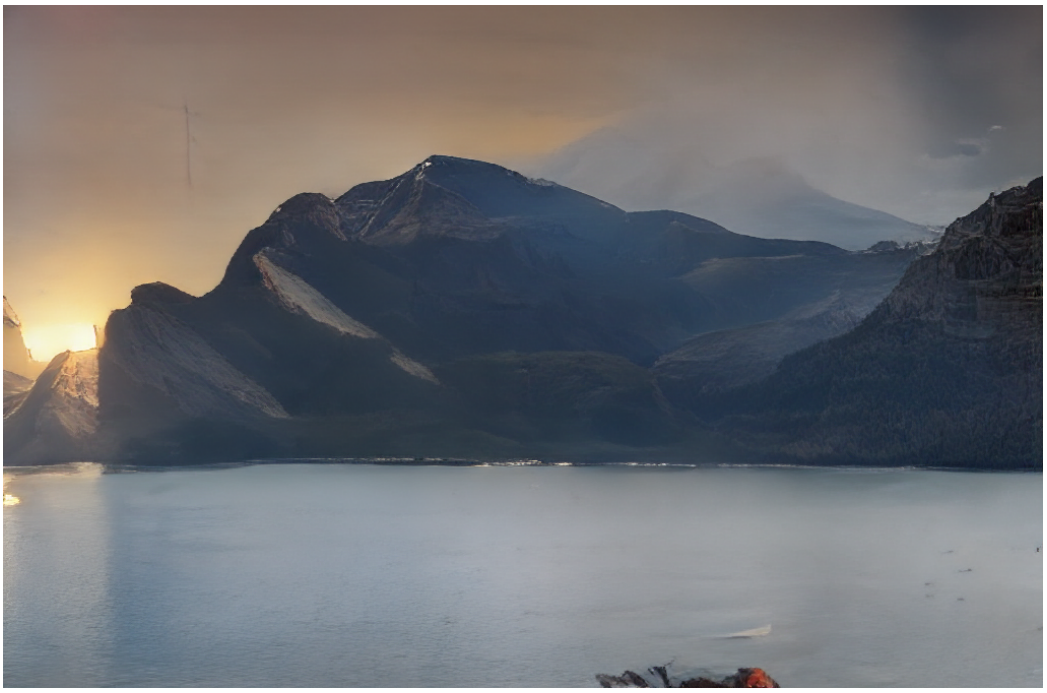


Figure 3. Adapted from Esser, Rombach, & Ommer (2021). Examples of image generation using VQGAN and Transformer method. The novel distribution the transformer builds is often guided with a starting point which can be (from top to bottom) half of an image, depth-maps, or segmentation maps, among others.

These are three example images of landscapes generated with the VQGAN and transformer architecture, applying the methods included in Esser, Rombach, & Ommer (2021).



The images are large, realistic, and high-resolution. Furthermore, their generation was guided with the same initial segmentation map; the resemblance is clear:

Analysis of output

As previously mentioned, the outputs of the second implementation do not have some of the problems of the outputs of the first. The images are larger, have high resolution, and there is greater diversity of outputs (the first example generated more similar outputs; slightly-different trees). Starting from a certain segmentation map, VQGAN is able to generate different landscapes, varying on the details (different lighting, vegetation, textures, etc.). Nevertheless, the images might be too realistic for the purpose of the project. The pieces are supposed to look clearly distorted (see Relevant Existent Projects section for examples of such style), which is the whole metaphor connecting to its meaning. These outputs could be perceived as real photographs or paintings, especially if the viewer does not look closely. Furthermore, although there is great diversity of the kinds of landscapes to generate, the project would beneficiate from having an even more diverse array of outputs (e.g. images of living beings, more abstract works). We can generate images of living beings and things with VQGAN, it will only require more training and other datasets.

Third Implementation

Another kind of ML architecture I was interested in learning about was that of text-to-image methods (see Exploring other GAN architectures section). These methods usually combine a GAN with other algorithms that allow for the generated images to match an input text.

CLIP

Contrastive Language–Image Pre-training (CLIP), proposed by OpenAI, is a model that uses transformers to learn about the patterns between images and associated texts. The model was trained by the team on a set of 400 million pairs of image and textual descriptions, and was meant to be used to "determine which caption from a set of captions best fits with a given image" (Steinbrück, 2021). The model is similar to a discriminator, with the difference that it is not constrained to a few classes, but it has built millions of associations to discriminate from. The revolutionary feature of this model is that it is capable of zero-shot learning, which means that it performs exceptionally well on unseen datasets (because of its unique training).

VQGAN + CLIP method

Combining CLIP with VQGAN has shown to lead to complex, high-resolution and diverse outputs. CLIP (Perceptor) is able to guide the training of VQGAN (Generator) to generate an image that matches a given text. In other words "we can use CLIP to guide a search through VQGAN's latent space to find images that match a text prompt very well according to CLIP" (Steinbrück, 2021). The model starts with a noise vector, which is fed into VQGAN to generate an initial image. The output and input text (parameter input by user) is fed into CLIP, which processes the data pair and computes the similarity between the text and image with

cosine similarity relevance as a loss function. CLIP computes a cumulative loss that includes how well the text matches the image, and how well the image matches the text (these distances are different). The result is then used to update the weights of VQGAN through gradient back-propagation. The new initial vector for a second iteration of VQGAN generation includes such feedback from CLIP, matching text and image after several iterations.

I adapted an implementation of the VQGAN + CLIP method for image generation, which can be found in this notebook. The following are some examples of images genenerated with this architecture:

Input text: “Nature”

Input text: “Miss climate change takes over”

With this experiment, I wanted to see what a personification of a natural phenomenon would look like. The result was far from expected, as one can see, it is clear that the word “miss” led the architecture to generate the image of a beauty contest participant. Still, you can see the model generated what seems to be some fallen trees and a top-view images of a storm/hurricane.

Input text: “Climate Change”

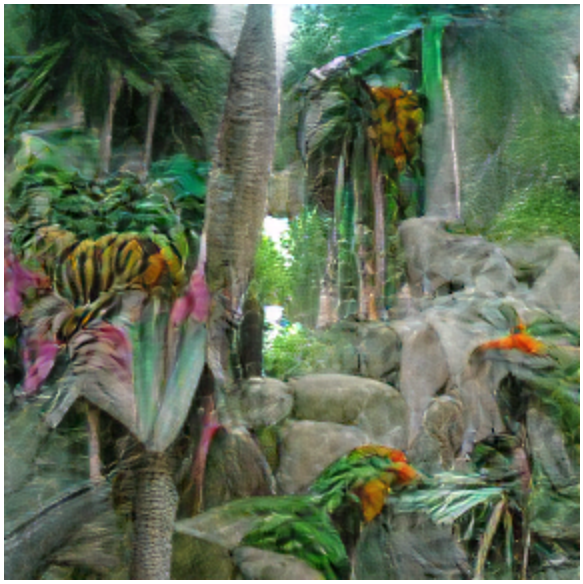
Input text: “Global Warming”

Input text: “Forest of old trees”

Input text: “Old tree on grassy field”

Input text: “Old-growth tree”

Input text: “The most beautiful garden”



Input text (both): “Tropical jungle”

Each image represents the output of fitting the image generation to the input text. Each example is the output after 1000 iterations except for the first one which is the output after 2500 iterations. Their resolution was set to either 300x300 or 480x480 pixels (constrained to a somewhat low resolution to decrease training time during experimentation).

As expected from the nature of CLIP’s training, the best images are those generated from descriptive text, while more abstract concepts lead to more white space or noise in the final output (compare “Climate Change” and “Global Warming” to the images that follow).

Quantitative analysis of performance

Assessing the performance of the program quantitatively is insightful. We can interpret such assessments in order to decide how many iterations of the training to run while generating images. We wish to optimize the image generation process; output high-quality images with the least number of iterations. In order to make such quantitative measurements, we must first understand the model metrics implemented by CLIP. We focus on CLIP since this is the architecture that drives the weight updating process. At the end of each iteration, CLIP matches the input text with the generated image, computes a loss function score, and informs VQGAN of the result for weight updating.

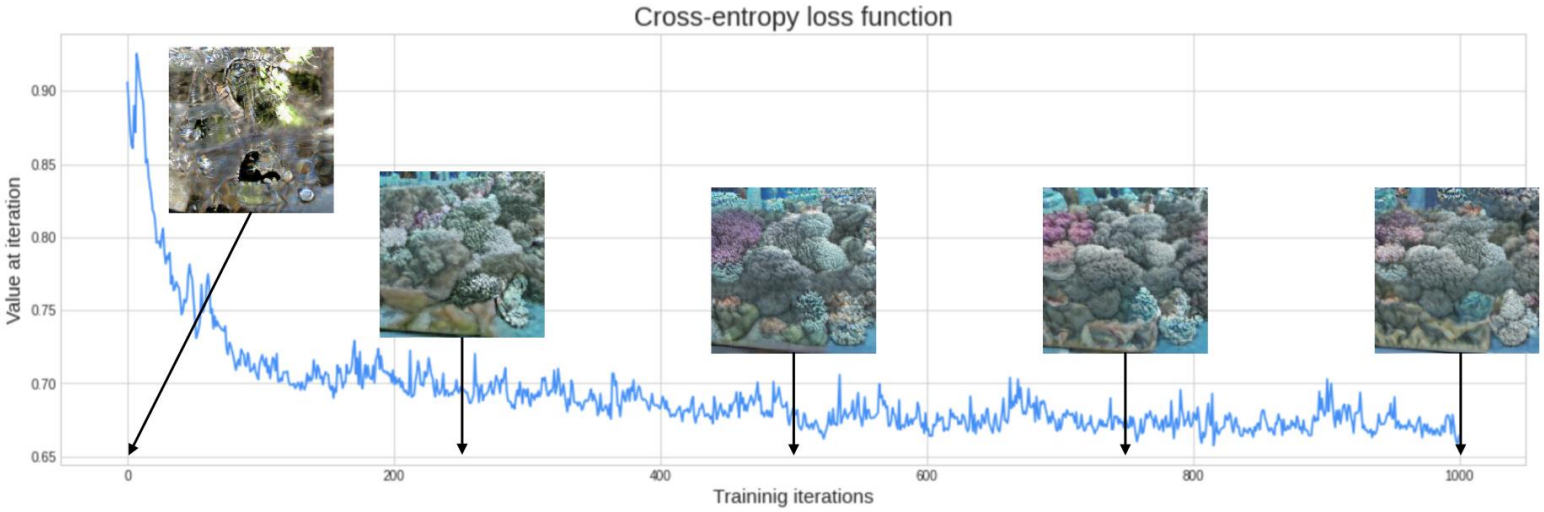
Given a batch of N (text,image) pairs, CLIP is trained to predict which of the N × N possible (image, text) pairings across a batch actually occurred (the closer match). CLIP will learn a multi-modal embedding space by jointly training an image and a text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch (matches) while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. The resulting similarity scores are used while optimizing a symmetric cross entropy loss (Radford, et al., 2021).

Cross-entropy is a measure of how different two probability distributions are from each other. It represents the average number of bits needed to encode data coming from a source with distribution P when we use model Q. Let us define P as a target probability distribution (e.g. images from CLIP training that relate to the input text), and Q as an approximation to the target distribution (e.g. the generated image). Then, the cross-entropy of Q from P is the number of additional bits to represent an event using Q instead of P. The result will be a positive number, the closer to zero, the more similar the distributions are. Cross-entropy is commonly used as a

loss function when optimizing classification models, and is more efficient than other methods like sum-of-squares (faster training and improved generalisation) (Brownlee, 2019).

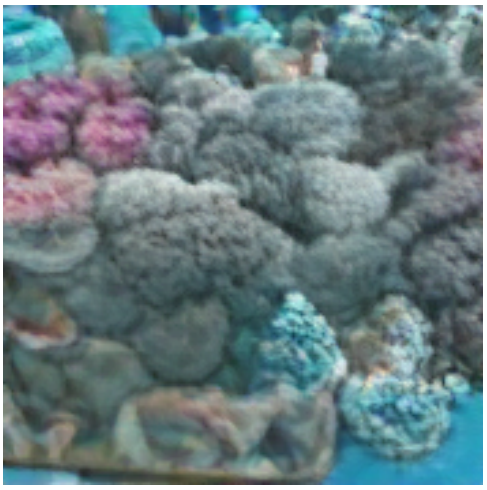
The original measure is not symmetric, meaning H(P, Q) != H(Q, P); meaning that the cross-entropy score of modeling distribution P from distribution Q is different to the cross-entropy score of modeling distribution Q from distribution P. Nevertheless, the model metric used by CLIP is a symmetric version of the original cross-entropy. According to Wang, et al. (2019), the asymmetric, original version of the metric is not efficient when working with noisy (incorrect) labels. Empirical data shows that cross-entropy tends to exhibit overfitting when implemented on “easy” labels (quick to converge), and underlearning when implemented on “hard” labels (those that take longer to converge because of close similarity to other classes). Not being able to manage noisy labels is a great problem while training a model like CLIP given that its training set is extremely large and definitely includes incorrectly-labeled images. Symmetric cross-entropy loss is able to solve the underlearning problem when training includes similar/overlapping labels, and thus has become a common measure while dealing with large, noisy traning data sets.

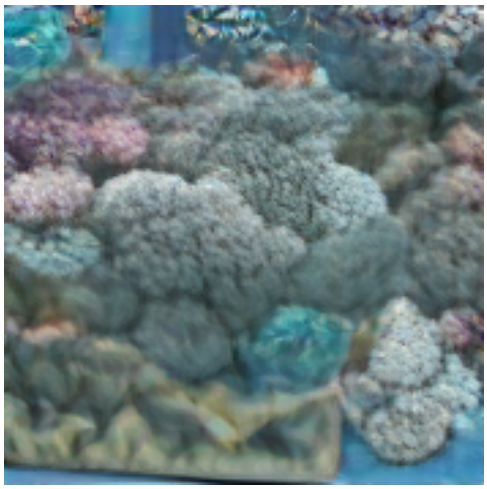
We analyzed the behavior of the symmetric cross entropy score through iterations of the training by running the algorithm and appending to a list the score computed after every step. First, we experimented with a smaller image (180x180 pixels), from the input text “coral reef” and 1000 training iterations. The following figure presents the evolution of the score through time:



The symmetric cross-entropy score started at around 0.90, which is expected since the image at step 0 was a noise vector. Then, the loss fluctuates and generally drops to around 0.70 after the first 180 training iterations. The generated image at step 250 was:

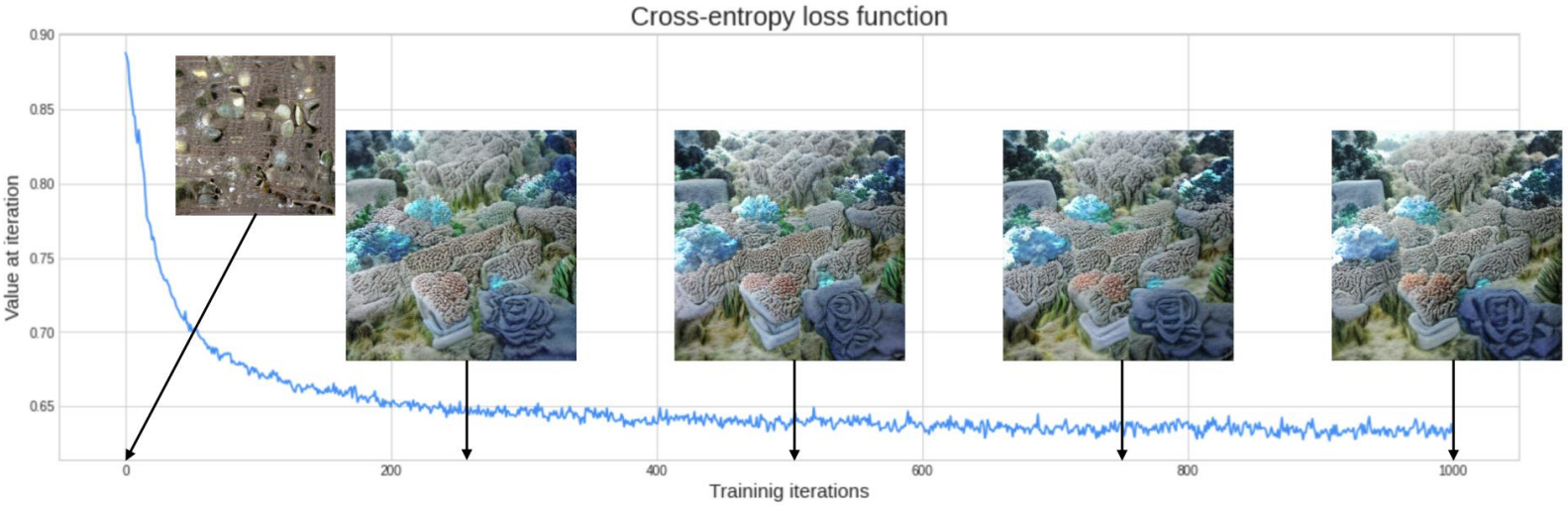
The output already has the general structure, and colors of final generated image. However, it misses the details and textures we expect from a high-quality output. We keep iterating the training and the loss decreases at a much slower rate. It takes more than 200 more iterations to get to 0.68. The improvement rate keeps slowing down, and the loss score only decreases to 0.65 after the 1000 training iterations. The outputs at 750 and 1000 steps were the following (left and right, respectively):





As the reader can appreciate, the images are very similar. The only noticeable qualitative improvement from step 750 to step 1000 is the addition to finer textures to the coral shapes.

A second experiment with the same input text, but larger size (360x360pixels) follows the same general trend of quantitative and qualitative improvement after 1000 training iterations (see figure below).[7]



[7] #dataviz see Appendix A

The loss score starts at around 0.90 and quickly drops to 0.68 after a little over 100 iterations. Then, the quantitative improvement rate slows down and gets to 0.63 after 1000 iterations.

The generated images at step 250 and 1000 were:

The images are strikingly similar in spite of being 750 iterations away. Both outputs have the same general structure, colors, and textures.

Comparing the evolution of loss scores between the low and high resolution experiments yields interesting observations. As seen by comparing both graphs, the second, higher-resolution experiment has less fluctuation of the loss score around the general decreasing trend. Furthermore, the loss score decreases faster and to a lower value in the second experiment compared to the first. In other words, the VQGAN+CLIP method is more efficient while generating large high-quality images than smaller ones. This can be explained by the nature of the transformer architecture, which is highly expressive (able to find patterns globally rather than locally as CNNs). Constraining the model to generate a low-resolution image will lead to overlap between local and global patterns. Working with larger resolutions (although counter-intuitive)

will allow the architecture to take full advantage of the locally-biased learning style of the convolutional network and the global-oriented attention of the transformer. Ultimately leading to quicker and higher-quality outputs.

From the present analysis we learned that while implementing VQGAN+CLIP for image generation a goof rule of thumb will be to constrain image generation to larger resolutions (around 300x300 pixels) and fewer training iterations (250 to 500) given that the quality of the image will not improve drastically after that point of training.

Example of presentation of the artworks

Caption:

This image was created by Artifical Intelligence.

Have you ever wondered how computers see the world? This is what a coral reef looks like for a computer.

Coral reefs support more species than any other marine environment, they clean the ocean from carbon dioxide, protect coastlines from storms and erosion, and over half a billion people depend on them for food, income, and protection. Unfortunately, because of climate change, more than half of all coral reefs have been lost, and the rest are experiencing rapid extinction.

This image is part of DSTRTD, a series of computer-generated snapshots of nature meant to question how our generations' perception of the natural world is shaped by our dependence on digital media.

Is your perception DSTRTD (distorted)?[8]

[8] #sciencecommunication see Appendix A

References

Asch, S. E. (1956). Studies of independence and conformity. I. A minority of one against a unanimous majority. Psychol Monogr 70(9):1–70.

Barnosky, A. D., Hadly, E., Bascompte, J., Berlow, E. L., Brown, J. H., et al. (2012) Approaching a state shift in Earth's biosphere. Nature 486: 52–58

Barrat, R. (2018). art-DCGAN. Retrieved from: https://github.com/robbiebarrat/art-DCGAN

Beisner, B. E. (2012). Alternative stable states. Nature Education Knowledge, 3(10), 33. Retrieved from https://www.nature.com/scitable/knowledge/library/alternative-stable-states-78274277

Borji, A. (2018). Pros and Cons of GAN Evaluation Measures. Retrieved from: https://arxiv.org/pdf/1802.03446.pdf

Brownlee, J. (2019). A Gentle Introduction to Cross-Entropy for Machine Learning. Retrieved from: https://machinelearningmastery.com/cross-entropy-for-machine-learning/

Brownlee, J. (2019). How to Evaluate Generative Adversarial Networks. Retrieved from: https://machinelearningmastery.com/how-to-evaluate-generative-adversarial-networks/

Brownlee, J. (2021). How to Identify and Diagnose GAN Failure Modes. Retrieved from: https://machinelearningmastery.com/practical-guide-to-gan-failure-modes/

Candido, R. (2020). Generative Adversarial Networks: Build Your First Models. Real Python. Retrieved from: https://realpython.com/generative-adversarial-networks/

Denchak, M., & Turrentine, J. (2021). Climate Change: What You Need to Know The lowdown on the earth’s central environmental threat. Retrieved from: https://www.nrdc.org/stories/global-climate-change-what-you-need-know#facts

Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. (2017). CAN: Creative Adversarial Networks Generating “Art” by Learning from Styles and Deviating from Style Norms. Rutgers University and Facebook AI Research. Retrieved from: https://arxiv.org/pdf/1706.07068.pdf

Esser, P., Rombach, R., and Ommer, B. (2021). Taming Transformers for High-Resolution Image Synthesis. Retrieved from: https://compvis.github.io/taming-transformers/paper/paper.pdf

Flores, C. (2019). The disconnection between humans and nature. Center for humans and nature. Retrieved from: https://www.humansandnature.org/what-happens-when-we-see-ourselves-as-separate-from-or-as-a-part-of-nature-the-disconnection-between-humans-and-nature

Garduno, C. R. (2020). Coding Art: Generative Adversarial Networks. Minerva University. Unpublished manuscript.

Herring, D., & Lindsey, R. (2021). Can we slow or even reverse global warming? National Oceanic and Atmospheric Administration. Retrieved from: https://www.climate.gov/news-features/climate-qa/can-we-slow-or-even-reverse-global-warming

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kesebir, S., and Kesebir P. (2017). How Modern Life Became Disconnected from Nature. Greater Good Magazine. Retrieved from: https://greatergood.berkeley.edu/article/item/how_modern_life_became_disconnected_from_nature

Nugent, C. (2018). The Painter Behind These Artworks Is an AI Program. Do They Still Count as Art? Retrieved from: https://time.com/5357221/obvious-artificial-intelligence-art/

Lesen, A. E., Rogan, A., & Blum, M. J. (2016). Science Communication Through Art: Objectives, Challenges, and Outcomes. Trends in Ecology & Evolution, 31(9), 657–660. doi.org/10.1016/j.tree.2016.06.004

Onkaos. (2018). Memories of Passersby I (Companion Version), 2018. Retrieved from: https://www.artsy.net/artwork/mario-klingemann-memories-of-passersby-i-companion-version-1

Ornes, S. (2020). Computers are changing how art is made. Software can inspire and even guide the creativity of artists. Retrieved from: https://www.sciencenewsforstudents.org/article/computers-are-changing-how-art-is-made

Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Retrieved from: https://arxiv.org/pdf/2103.00020.pdf

Scheffer, M., Bascompte, T. K. Bjordam, S. R. Carpenter, L. B. Clarke, C. Folke, P. Marquet, N. Mazzeo, M. Meerhoff, O. Sala, and F. R. Westley. (2015). Dual thinking for scientists. Ecology and Society 20(2): 3. http://dx.doi.org/10.5751/ES-07434-200203

Scheffer, M., Bascompte, J., Brock, W. et al. (2009). Early-warning signals for critical transitions. Nature 461, 53–59. Retrieved from: https://doi.org/10.1038/nature08227

Scheufele, D. A. (2014). Science communication as political communication. Proceedings of the National Academy of Sciences 13585-13592. DOI: 10.1073/pnas.1317516111

Shibsankar, D. (2021). 6 GAN Architectures You Really Should Know. Retrieved from: https://neptune.ai/blog/6-gan-architectures

Steffen, W. et al. (2011) The Anthropocene: from global change to planetary stewardship. AMBIO 40, 739–761.

Steinbrück, A. (2021). VQGAN+CLIP — How does it work? Retrieved from: https://alexasteinbruck.medium.com/vqgan-clip-how-does-it-work-210a5dca5e52#:~:text=VQGAN%2BCLIP%20is%20a%20neural,(and%20some%20other%20parameters).

Thakur, A. (2021). Taming Transformers for High-Resolution Image Synthesis. The efficiency of convolutional approaches with the expressivity of transformers. Retrieved from: https://wandb.ai/ayush-thakur/taming-transformer/reports/Taming-Transformers-for-High-Resolution-Image-Synthesis---Vmlldzo0NjEyMTY

US Environmental Protection Agency. (2022). The Effects: Dead Zones and Harmful Algal Blooms. Retrieved from: https://www.epa.gov/nutrientpollution/effects-dead-zones-and-harmful-algal-blooms#:~:text=Dead%20zones%20are%20areas%20of,excess%20nutrients%20from%20upstream%20sources.

Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. (1997). Human domination of Earth's ecosystems. Science 277, 494–499

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J. (2019). Symmetric Cross Entropy for Robust Learning With Noisy Labels. 322-330. 10.1109/ICCV.2019.00041.

World Economic Forum. (2021). The global risks report 2021. World Economic Forum. Retrieved from: https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2021.pdf

Appendix A. LO and HC applications

LO applications

Course: NS166: Keeping Earth Habitable

#disturbances: While characterizing the problem in the *background and introduction* section, I conceptualized the issue through an ecological state theory framework. This allowed me to describe the urgency of solving the problem. When we think of humans as a major disturbance to the biosphere's processes (the system) it is easy to see how immediate action is needed. Furthermore, it is useful to apply the concept of hysteresis while forecasting the possible future of the climate crisis if significant rapid change is not achieved.

#socialframes: Through both presented sections I discussed why a political frame is necessary to consider while developing an effective solution to the climate crisis. Drawing from historical examples of the overlap between natural and political science, as well as arguing science communication can be a powerful tool for political advocacy and advice. Understanding the problem through an ecological lens is not sufficient for developing a long-term solution. Acknowledging why it is fundamental to spread the word about climate change in order to trigger the requited cultural shift.

Course: NS164: Solutions From and For Life

#sciencecommunication: The purpose of the capstone project is rooted in science communication. As explained in the *Introduction and Background* section, the project is meant to bring scientific knowledge about the natural world closer to an audience who is growingly dependent on digital mediums to interact and perceive the world around them. The project

applies the principles of #scienceoflearning in order to make scientific content more accessible and engaging. An example of the final public-facing deliverable can be found in the last section. The image and opening line are meant to catch the attention of the scrolling audience, delivering quick facts about the importance of coral reefs and a minimal description of the project. The posts share crucial information about ecosystems in a practical, digestible way. After engaging with the post, the audience will have learned about coral reefs and (hopefully) question how technology shapes their perception of nature. Note that the caption is meant to be minimal with the purpose of retaining more of the audience (people are more likely to read a short than a long caption).

Course: CS156: Machine Learning for Science and Profit

#neuralnetworks: The Learning Journal explains and analyzes several architectures conformed of different types of neural networks. Furthermore, the narrative records my experimentation and learning process, starting from the simple GAN, then the VQGAN+Transformer network method, and finally the VQGAN+CLIP implementation. Throught the completion of this project, I have been able to learn more about different kinds of neural networks, their advantages and drawbacks.

#modelmetrics: The *Research Stage* section describes quantitative and qualitative model metrics for assessing the performance of a generative architecture. The *Quantitative analysis of performance* section describes the loss function that drives weight updating of the final architecture (symmetric cross-entropy loss). The section explains how this is different from the regular (asymmetric) cross-entropy loss and why the symmetric version is preferred for this kind

of problem. Finally, the same section analyzes the performance of the architecture quantitatively, comparing the performance across different resolutions and interpreting results. Ther use of model metrics has directly informed the parameters that will be used for image generation.

HC applications

#rightproblem: The *introduction and background* section clearly defines the nature of the current environmental crisis (problem), describing its alarming likely future consequences in terms of ecological state theory and, as well as describing the need of tackling the issue at different scales (local, regional, global).

#purpose: The *about the project* section explicitly states the purpose of the project. "To appeal to the audience's emotions and nudge them into questioning their connection to nature and how it may be distorted by their postmodern lifestyle and dependency on digital technologies." Such purpose statement is drawn from the discussed research on the increasing disconnection between human minds and nature. Furthermore, the purpose statement directly addresses the problem identified in the previous section, justifying its approach (appealing to system I instead of system II).

#audience: The *about the project* section describes the intended audience of the artwork, considering the audience-wise trends of similar projects in the US, and explaining how a general audience will maximize the reach of the pieces. Furthermore, the section also includes a discussion of why exactly a general audience is effective, explaining how from a psychological perspective, it is necessary to get as many people on board as possible to facilitate a cultural shift that, in turns, pressures political action. I plan on enriching this description while applying #designthinking. More specifically, after analyzing the impact of the final product on a general audience.

#systemdynamics: The *Avoiding hysteresis at the global scale* section conceptualizes the Earth's biosphere as a steady-state system. A system which can shift from one state to another when critical thresholds are reached. The section describes why it is that the state of the biosphere is about to reach a tipping point after which the disturbance caused on the system will require more energy to come back to than the energy required to shift in the first place.

#algorithms: A deep understanding of the algorithms behind neural networks was fundamental while interpreting and developing the Machine Learning pipelines. Understanding the way in which these architectures work and how many times they attempt to imitate cognitive processes (e.g. attention, creativity, discrimination, etc.) required me to conceptualize these problems through algorithmic thinking (input, internal logic, output).

#evidencebased and #sourcequality: I have effective and efficiently used high-quality sources to support the claims I describe to justify and describe each section of the project. Given the interdisciplinary nature of the project, I was able to interact with high-quality sources across fields (complex systems, science communication, psychology, political science, environmental science, computer science, mathematics, among others).

#designthinking: As documented in the Learning journal section, while selecting the ML architecture that was the best fit for my project, I apply the principles of design thinking. Namely, I had three iterations (different implementations) where I would choose, investigate and experiment with a ML architecture, then evaluate the quality and fit of the outputs, and inform the following iteration with such information.

#dataviz: While presenting the loss function analysis in the *Quantitative analysis of performance* section I designed clear and comprehensive data visualizations that whow the evolution of the loss score through training iterations, as well as presenting some of the snapshots of the generated image. These data visualizations are able to present and help the audience understand the connections between the quantitative and qualitative assessments of the image generation process.

#variables: In order to interpret, annotate, and write code efficiently, it was always useful to think of each function in terms of its independent (e.g. inputs, parameters) and dependent variables (outputs). When you think about the data available and the goals for initializing the next step of the pipeline, it becomes clear what the purpose and mechanisms of each part of the code are.

#professionalism and #organization: The present document attains to the standard professional and technical conventions of a project in the field. Taking advantage of the learning journal structure, I am able to present a narrative about the challenges and steps taken throughout the completion of project.

#selfawareness: The completion of the project has required me to continuously assess and understand my strengths and weaknesses. A good example of this was during last semester when I decided to pivot from my initial to the current project after realizing the project I was pursuing was not feasible given my knowledge and experience. This resulted into effectively

crafting a project I could complete on time and within the constraints and requirements of a capstone project.

#emotionaliq: A thorough understanding of my emotions has been critical for the project completion, particularly while self-regulating, dealing with stress and anxiety, and staying motivated. Knowing my triggers and developing healthy coping mechanisms has helped me ensure I am in the right mental space for working on capstone.

#scienceoflearning: The use of computer-generated artworks sparks from the recognition of the affective domain of learning as central for decision making. As analyzed in the *Educating through art* section, while traditional learning focuses on rationality, psychological research shows that appealing to the learner's emotions tends to be more engaging. Bridging art and science is not only a creative but a also a more effective way of teaching. Especially outside of traditional teaching settings (school vs social media).