Exploring Generative Adversarial Networks for Text-to-Image Generation with Evolution Strategies

Victor Costa, Nuno Lourenço, João Correia, Penousal Machado
University of Coimbra, CISUC, DEI
Coimbra, Portugal
{vfc,naml,jncor,machado}@dei.uc.pt

ABSTRACT
Text-to-image generation has achieved impressive results, featuring a variety of models trained on extensive datasets comprising text-image pairs. However, some methods depend on pre-trained models, using gradient-based approaches to update latent vectors in the latent space. In this work, we propose the use of Covariance Matrix Adaptation Evolution Strategy to explore the latent space of a Generative Adversarial Network. Our experimental study compares our approach with gradient-based and hybrid strategies, using diverse text inputs for image generation. We adapt an evaluation method that projects the generated samples into a two-dimensional grid to assess the diversity of the distribution. Results evidence that the evolutionary method produces more diverse samples across different grid regions, while the hybrid method combines gradient-based and evolutionary approaches, enhancing result quality.

CCS CONCEPTS
• Computing methodologies → Genetic algorithms; Neural networks.

KEYWORDS
evolution strategies, generative adversarial networks, generative models

ACM Reference Format:
Victor Costa, Nuno Lourenço, João Correia, Penousal Machado. 2023. Exploring Generative Adversarial Networks for Text-to-Image Generation with Evolution Strategies. In Genetic and Evolutionary Computation Conference Companion (GECCO ’23 Companion), July 15–19, 2023, Lisbon, Portugal. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583133.3590549

1 INTRODUCTION
Recently, generative models gained a lot of attention from the community. In this context, several models were proposed for image generation based on text inputs. Some rely on pre-trained generative models, such as Generative Adversarial Networks (GANs) [4], combined with a language processing system capable of describing an input image using natural language. For this, a mapping mechanism from the output of the language model to the latent space of a GAN is paramount for achieving a meaningful representation of the resulting images.

In [1], Contrastive Language-Image Pre-training (CLIP) [8] and BigGAN [2] are used to convert texts to images. In this work, we propose the use of Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [5] instead of Adam to explore the latent space of GANs for text-to-image generation, aiming to achieve a diverse representation for the text input.

We compare the evolutionary approach, the backpropagation training, and a hybrid solution. To represent our results, we adapt the evaluation method of [3] for projecting the generated samples into a two-dimensional grid. In this way, we provide insight into the distribution achieved by the methods evaluated in our study not only by visual inspection but also by a quantitative metric.

2 PROPOSED MODEL
In [1], two pre-trained models are used, CLIP and BigGAN, and backpropagation with the Adam optimizer [6] is used to optimize the latent space. BigGAN is trained on the ImageNet dataset and is used as the generative model. CLIP is trained on 400 million pairs of images and texts.

Figure 1 shows an overview of the architecture of the framework. The objective of the training process is to find the best latent inputs for BigGAN that can generate an image corresponding to the input text. First, CLIP encodes the input text into a feature vector of 512 elements that will be used to compare how the generated images approximate the input text. Then, at each iteration, the latent input...
We adapt the evaluation method proposed in [3] to analyze the performance of the other methods.

To take into account both art and elements of contemporary culture, in this work, we use the generator embedding from BigGAN to project the class conditioning vector to the same dimension of the latent input (typically set to 128). Thus, the input used for BigGAN consists of \((1 + [h] \times 2 \times [z])\) elements, where \([h]\) is the number of hidden layers and \([z]\) is the latent dimension, resulting in an input of 3840 elements.

Three approaches are used to optimize this input for the desired text-to-image task: backpropagation with Adam, CMA-ES, and a hybrid solution. The backpropagation version uses the Adam optimizer with the cosine similarity between the feature spaces of the text and images as the loss function. CMA-ES uses the input of 3840 elements as the population of individuals, where each one represents a latent vector for BigGAN. The CMA-ES algorithm produces new generations of individuals according to its mutation strategy, being guided by the fitness function that also uses the cosine similarity. In the hybrid approach, we apply \(k\) steps of backpropagation before applying the mutation step from CMA-ES.

4 RESULTS

Figure 2 presents the distribution of images in the two-dimensional grid produced by the evaluation method using the first input text: "A painting of Superman by Van Gogh". The first grid (Figure 2(a)) contains the images generated by BigGAN when using Adam to optimize the latent space. We can see that images are concentrated in a specific region, indicating that the diversity of the created samples is limited when using this approach. When inspecting the CMA-ES in Figure 2(b), we can see that the images are more distributed through the grid. This is an indication that the CMA-ES is promoting a better exploration of the search space, allowing for more diverse results. Another interesting observation emerges when we compare CMA-ES with Adam. We can see that they are exploring non-overlapping regions, indicating that they follow different paths when performing the optimization of the latent space. When combining Adam with CMA-ES, we leverage the results from both approaches. Thus, we can see in Figure 2(c) a mixture of the results of both approaches.

To confirm our observations, we apply the Jaccard Index using the projected grids as proposed in [3]. The Jaccard Index measures the intersection over union of the samples from the distributions. For this, we use the best performing method as the baseline, i.e., the Hybrid approach. Given this baseline, the results for the Jaccard Index for the Adam and CMA-ES approaches are 0.2978 ± 0.0111 and 0.3710 ± 0.0124, respectively. As bigger values indicate that the distribution is closer to the baseline, our results evidence that the CMA-ES approach better approximates the Hybrid method. The Adam version achieved lower performance, evidencing the poor capacity of diversity in the exploration of the latent space. However, the Hybrid method is able to combine the two distributions, leading to a broader exploration of the search space.

In Figure 3, we present samples created using the three different methods. Although the samples are not perfect, we can see elements from the input text in all images, such as texture and colors. It is important to note that we use BigGAN trained on the ImageNet dataset in our experiments. Thus, the generative model is not trained specifically on the tasks given through the input texts.
Exploring Generative Adversarial Networks for Text-to-Image Generation with Evolution Strategies

GECCO ’23 Companion, July 15–19, 2023, Lisbon, Portugal

Figure 2: Two-dimensional grid revealing the distribution of images generated using Adam (a), CMA-ES (b), and the Hybrid approach (c) for the text input "A painting of Superman by Van Gogh".

Figure 3: Samples from Adam (top row), CMA-ES (middle row), and the Hybrid (bottom row) approach for the text input "A painting of Superman by Van Gogh".

Figure 4: The cosine similarity for best images generated using Adam, CMA-ES, and the Hybrid approach for the text input "A painting of Superman by Van Gogh". As we have a different number of iterations and generations for each approach, we represent on the horizontal axis the iteration percentage. The values reported in this chart are the average cosine similarities for each approach during 500 executions with a confidence interval of 95%. The results show that Adam achieved the best results regarding this metric. However, this behavior does not correspond to the visual inspection of the images. The CMA-ES solution achieved the lowest results, but the samples appear to be better than the Adam version. The Hybrid approach is closer to the Adam results, combining the benefits of both approaches in terms of quality and diversity of the generated samples. This suggests that relying on the similarity metric alone may limit the diversity of the results.

Figure 5 shows the distributions for the second text input: "A painting of Darth Vader in the style of Matisse", where CMA-ES and the Hybrid approaches also present more diversity than Adam. Considering the hybrid approach as the baseline, the Jaccard Index for Adam and CMA-ES are $0.3015 \pm 0.0119$ and $0.3671 \pm 0.0132$, respectively. We can also see that the Adam approach produces a relevant number of samples that contain only a single color without any textures, failing to capture the characteristics of the text input. This indicates that the evolutionary algorithm also contributes to discarding poor latent vectors, focusing on areas that actually give some significant attributes with respect to the input text.

We highlight in Figure 6 samples achieved from each strategy applied in our experiments. The top, middle, and bottom rows contain samples generated by applying Adam, CMA-ES, and the Hybrid approach. We can see in these samples the intense use of colors, as in the paintings of Henri Matisse, and the fictional character cited in the sentence, evidencing that the model successfully captures the characteristics of the input text.

Figure 7 shows the cosine similarity for the best images generated during the latent space exploration using Adam, CMA-ES, and the Hybrid approach for this text input. In this case, the Hybrid approach is able to overcome Adam. One possible explanation is that the monochromatic samples generated by Adam are degrading the results.
5 CONCLUSION

In this work, we explore the latent space of Generative Adversarial Networks by using Evolution Strategies to improve the diversity of the generated samples. For this, we adapt a model that uses CLIP and BigGAN for text-to-image generation. We design an experimental study comparing the exploration of latent spaces by backpropagation with Adam, an evolutionary algorithm using CMA-ES, and a hybrid approach using Adam and CMA-ES. To evaluate the results, we adapt a method to visualize and quantify the distribution of samples by projecting them into two-dimensional grids.

Our results show that the CMA-ES algorithm contributes to a better exploration of the latent space, achieving better diversity when compared to Adam. The hybrid leverages the quality and diversity of the samples, outperforming the standalone CMA-ES and Adam approaches.

ACKNOWLEDGMENTS

This research was supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI. This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (FIDADE), within the scope of CBUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020.

REFERENCES

[1] [n. d.]. The Big Sleep: BigGANxCLIP. https://colab.research.google.com/drive/1NCceXzmlhKOSLAd 0/j7U7na9USkK5SNWR. Accessed: 2022-06-21.
[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In International Conference on Learning Representations.
[3] Victor Costa, Nuno Lourenço, João Correia, and Penousal Machado. 2021. Demonstrating the Evolution of GANs Through t-SNE. In Applications of Evolutionary Computation, Pedro A. Castillo and Juan Luis Jiménez Laredo (Eds.). Springer, Cham, 618–633.
[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In NIPS Curran Associates, Inc., 2672–2680.
[5] Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. Evolutionary computation 9, 2 (2001), 159–195.
[6] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR).
[7] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning research 9, Nov (2008), 2579–2605.
[8] Alec Radford, Jong Woon Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning. PMLR, 8748–8763.