Exploring Novel Image Generation via Script-Directed Scene Formation

Kaveesh Khattar, Tejas Goyal, KP Dhruv, Aditya Hombal, and Dr. Mamatha H R[1]

Department of Computer Science and Engineering,
PES University, Bengaluru, Karnataka, India
[1]mamathahr@pes.edu

Abstract. Deep Fusion Generative Adversarial Networks (DF GANs) have shown to be effective tools for producing realistic pictures. In this research, we demonstrate our implementation that creates unique representations of objects and human beings using a pretrained DF GAN. With the help of labelled input photos and identification tokens, our model was able to create new representations of objects and show people doing interesting things. By using script-based prompts, our system was able to generate scenarios with many pictures that made sense. We also made accessibility easier with a Gradle-implemented user-friendly interface. We illustrate the efficiency and adaptability of our method for producing a wide range of visual material through experimental validation and qualitative evaluation.

Keywords: Image Models · Text-to-Image · Generative AI.

1 Introduction

The creation of high-fidelity pictures from massive datasets has been revolutionised by Generative Adversarial Networks (GANs), especially Deep Fusion GANs. In this work, we explore new picture creation using a pretrained DF GAN. Our primary objectives are to: increase the model's capacity to generate new object representations and human actions; and secondly allow the creation of scenes that are prompted by text. We challenge the model to produce human movements and unseen object descriptions by providing labelled input pictures and tokens, therefore expanding the limits of its generalisation abilities. We also include scripted building cues for cohesive scene development, with the goal of achieving dynamic visual narrative. To guarantee usability, we employ Gradle to create an interface that is easy to use and accessible to a wide range of user demographics. By advancing pretrained DF GANs, we contribute to the evolution of image synthesis techniques, fostering innovation in artificial intelligence and computer vision realms.

2 Dataset

MS-COCO (Microsoft Common Objects in Context): The MS-COCO benchmark dataset is highly advantageous for tasks in object identification, segmentation, and captioning within the field of computer vision. Since it has over 200,000 photos, it provides data that can help in building robust models. Thorough annotations of images in the dataset such as segmentation masks give the dataset its advantage. With a combination of large number of images and high quality annotation it gives the model what it needs to provide more accurate results.

Furthermore, MS-COCO is used by state of the art models in the field of computer vision, which further solidifies our stance in choosing it as our primary dataset. Also the dataset's accessibility makes research contributions to it more seamless and helps it stay relevant for further use cases.

Models trained on MS-COCO perform extraordinarily well. Improved performance can be achieved by efficiently fine-tuning on smaller, task-specific datasets by pre-training on this dataset.

Finally, the photos from MS-COCO's real-world relevance, which show commonplace scenarios, improve the applicability of models trained on this dataset. These include anything from robots and self-driving cars to general visual comprehension. For this reason, the MS-COCO dataset remains a fundamental tool that propels progress and creativity in computer vision research.

3 Architecture

3.1 Sourcing Images and performing transformations

Fig. 1. Providing Captions for Images

We obtained a variety of datasets from publicly accessible sources that included photographs of both human individuals and objects. To ensure a broad range of visual representations, we carefully selected a collection of high-resolution photographs from different categories for items. To provide more context about

humans to the model, we chose pictures that had people in multiple settings performing different activities.

We pre-processed the photos by converting them to a specified resolution the model could take and cropping them to size for compatibility. Then we give these images to the pre-trained Deep Fusion Generative Adversarial Network (DF-GAN). The process involved normalization techniques to produce pixel values in-line with all the images in the dataset and re-scaling the photos to 512x512 pixels.

To give the model context and direction throughout the creation process, the photos were annotated with relevant labels and tokens. This step aims to enable the model to produce output that is more coherent and contextually relevant.

Finding and preparing the photographs was only one of the many tasks involved in getting the dataset ready for training and evaluation. These procedures prepared the way for the pre-trained DF-GAN to be applied with success in trials involving picture generation.

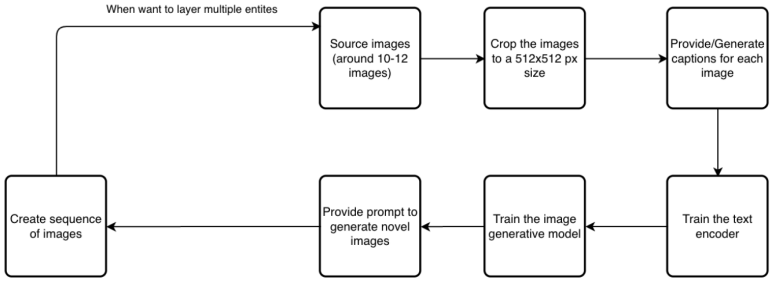

Fig. 2. Simplified flow of the model

3.2 Train text encoder

Interpreting and using the provided labels, and tokens is the most important phase in the training process for the text encoder. Textual inputs, such as labels indicating item categories, tokens for picture identification, and prompts guiding scene development, must be transformed into a format that can be combined with the image data by the text encoder. Through iterative training, the text encoder learns to precisely extract relevant characteristics and context from the textual input, which makes it easier for the model to understand and comprehend the intended instructions. By improving the text encoder in combination with the

picture production model, we ensure that the generated outputs are consistent with the supplied textual direction. This results in representations that are both contextually relevant and visually appealing. The synergy between textual and visual modalities is increased and the model's ability to generate a wide range of outstanding pictures that match preset requirements is improved through this iterative training process.

3.3 Train image encoder

Training the image encoder is one of the many steps involved in developing dependable and effective picture production models.

Once the text encoder is trained on the captions and tokens assigned to an image, the image encoder is then trained on the corresponding image. Before this step, as mentioned before, the image is pre-processed and transformed to meet the model's compatibility needs.

The model's primary task is to extract salient visual features from the raw input photographs, identifying details and patterns that are required to remember the subject of the image. Then later on, the model's primary task is to produce coherent and visually content that is realistic and contextually appropriate.

Throughout the training phase, the image encoder is optimised to extract relevant visual information from the input photos through a series of runs. The parameters of the image encoder were tuned in tandem with those of other components of the DF-GAN framework, such as the text encoder so that it only learns the tokens associated with that image and binds that learning to an image. Continuous enhancement of the image encoder enables the model to better learn how to integrate visual cues with written instructions to generate precise representations of the input data.

The image encoder and the text encoder work hand in hand to achieve high consistency results between the textual and visual modalities. Through collaborative training, the model gains an understanding of the input data. This enables it to create realistic images that match the provided written descriptions. The iterative approach used in this paper strengthens the model's capacity to provide high-quality visual information and enhances contextual coherence in the generated pictures.

3.4 Prompt and epoch configurations for optimum outputs

To optimize for image generation, epoch parameters, and text prompt settings must be given throughout the training phase. To make impactful prompts we write brief textual descriptions that instruct the model to generate desired visual outcomes. These prompts, which act as contextual signals, help assist the model in focusing its attention on certain distinctive aspects of images or scenes. Correct prompt formulation and selection help ensure that the photos generated reflect the intended creative concept.

The extent and rigor of model training are also heavily controlled by the epoch configurations used. The training dataset is separated into epochs, or iterations, with each epoch allowing the model to improve performance and make parameter adjustments as needed. Training efficiency and model convergence are balanced by selecting appropriate period settings. By modifying these parameters, we would improve the model's performance and ensure that the pictures created contain the desired properties, such as realism, relevance, and contextual importance.

Prompt and epoch configurations must be adjusted by ongoing testing and validation. We iteratively tweak these parameters to assist the model's ability to generate high-quality photographs that match preset requirements. We can fully utilize the model's capabilities and adjust its behavior via an iterative method to satisfy specific picture-producing criteria and creative ambitions.

3.5 Scene generation for prompts with multiple sentences

During setting up scenes for multiple sentence prompts while training, it is vital to set up scene-creation configuration with multi-sentence prompts to allow for the model's capability to generate cohesive contextually rich, and correct narratives. When we give the model multi-sentence prompts, we need to provide a sequence of textual or verbal descriptions that explain complex and multi-faceted scenarios or stories. This instructs the model to generate multiple pictures that collectively depict the intended scene and sequence. These prompts serve as a sort of scaffolding for the story, guiding the model while it creates the scene and ensuring coherence and consistency in the images that are generated.

To learn how to employ multi-sentence prompts for scene generation most effectively, let's examine the following examples: In one scene, we see two men drinking orange juice in a cosy house, farmers gathering oranges in a grove before dawn, and orange boxes on an old truck parked on a road. These lines serve as the model's guide, creating a sequence of images that depict the process from harvest to consumption.

Another use case features a close-up shot of a burger smeared in ketchup. The setting of a dinner table with the ketchup bottle. As the narrative progresses, a man is seen eating this burger in a rather crowded restaurant. These few words effectively set the tone for a series of pictures that portray the joy and enthusiasm of a hearty meal which can be used in many marketing placeholders.

Every line that is produced using multi-sentence prompts gives a different facet to the larger plot, such as a place, a character, an event, or a dialogue. These sentences are deliberately placed differently to allow the scene to unfold gradually. Subsequently, the model would build upon these signals to generate coherent images that keep the storyline consistent. Utilizing this method, the model can create a range of dynamic situations by generating rich and well-structured visual tales from multi-sentence prompts.

4 Results

We show the experimental findings demonstrating the effectiveness of our work in making a diverse range of visual media using pre-trained Deep Fusion Generative Adversarial Networks (DF-GANs). We measured the model's capability to generate unique object representations by giving it labeled photographs as input and distinguishing tokens. The results successfully demonstrate the pre-trained DF GAN's capacity to extrapolate and generalize using existing data by generating realistic representations of items/objects that had not yet been seen in the dataset. We also noticed how effectively the model would use script-based prompts to build logically consistent scenarios. Short descriptions of diverse scenarios were given to the pre-trained DF-GAN, requesting it to generate many pictures that correctly matched the given context.

5 Performance Metrics

5.1 Inception Score(IS)

The inception score (IS) is a mathematical algorithm used to measure or determine the quality of images created by a generative adversarial network (GAN). The inception score offers objective and consistent measures of generated images; and by extension, the quality and capability of the underlying generative model.

5.2 Fréchet Inception Distance(FID)

This metric measures the distribution similarity of feature representations from a pre-trained Inceptionv3 model for real and produced images by combining Inception Distance and Frechet Distance. Higher quality image production is indicated by a smaller Frenchet Inception Distance.

5.3 Mean Opinion Score(MOS)

Utilize the Mean Opinion Score (MOS), which is derived from human participants rating images according to perceived fidelity or quality, to assess the quality of generated photos. The MOS is obtained by averaging these assessments; higher numbers denote more visually appealing or realistic images, while lower values denote inferior quality or fidelity.

6 Future Research Directions

Text to image generation models have come a long way, but there are still a number of areas that need more research and development. This section highlights potential future research directions based on the current state of models and identified areas for improvement

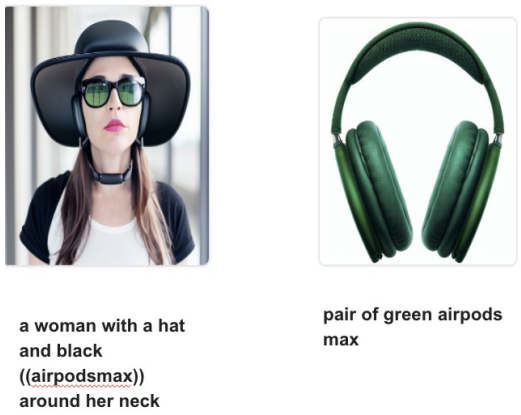

Fig. 3. Novel representation

Fig. 4. Before providing images for training v/s after providing images for training

Fig. 5. Scene visualisation

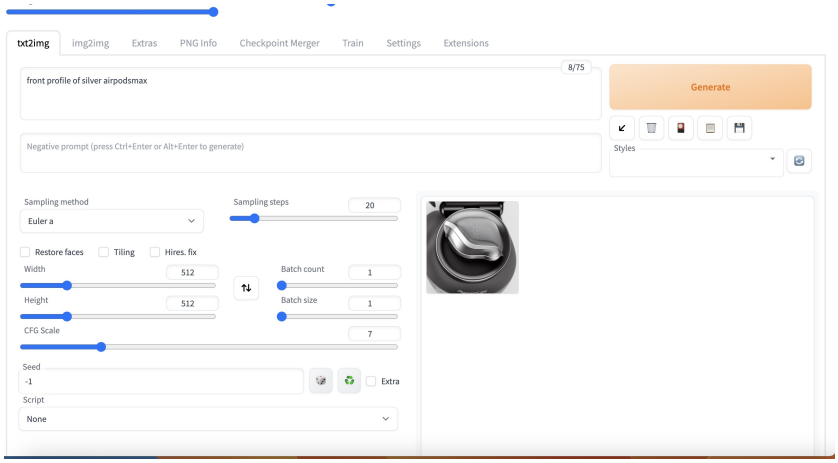

Fig. 6. Dashboard for Fine-Tuning

Improved Semantic Understanding: Subsequent studies on text-to-image generation ought to improve semantic comprehension by utilizing sophisticated natural language processing methods such as knowledge graphs or pre-trained language models. By using knowledge graphs created from text embeddings, models may be able to produce visuals that are more closely matched with the intended meaning of the input text, potentially improving positional and contextual understanding..

Increased Resolution and Realism: Though current models have come a long way in producing high-quality photographs, resolution and photo-realism may still use some work. Future research could focus on developing techniques to generate images at higher resolutions, allowing for more detailed and visually appealing results. Additionally, exploring advanced loss functions or perceptual similarity metrics could further enhance the realism of generated images, making them indistinguishable from real photographs.

Fine-grained Control and Manipulation: Current text-to-image models often lack fine-grained control over generated images. Future research could investigate methods to enable precise control and manipulation of image attributes, such as object positions, colors, and styles, based on textual input. This could involve exploring novel conditioning techniques or incorporating additional information during the generation process to produce images that align with specific user requirements.

Handling Ambiguity and Multi-modal Outputs: In order to achieve a variety of outputs that might capture many meanings, future research in text-to-image generation should investigate strategies for handling ambiguity in textual

descriptions. One could use strategies like adversarial learning, variational techniques, and uncertainty estimation. To potentially reduce ambiguity, one way is to train the image generator using the prompt and the parsed scene graph output.

Incorporating User Feedback and Interactive Generation: Interactive text-to-image generation systems that incorporate user feedback and preferences hold great potential for enhancing user satisfaction and enabling personalized image generation. Subsequent investigations may concentrate on creating models that can adjust and enhance their creation procedure in response to user interactions, enabling users to offer input and direct the picture synthesis procedure in real-time.

7 Conclusion

Through this effort, we successfully show the ability of pre-trained Deep Fusion Generative Adversarial Networks (DF-GANs) that help expand the possibilities of image generation. We show significant work in the generation of unseen novel objects and human actions, as well as in the development of coherent text-prompted situations, by using these advanced models. With the help of comprehensive testing and qualitative research, our method has shown its value and versatility in creating novel and contextually relevant visual media.

Our findings portray the effectiveness of pre-trained DF GANs as efficient tools for creative picture synthesis. These models contain large potential for a diverse range of applications, including but not limited to content creation, building virtual worlds and narratives, and the creation of human behaviors and novel object representations. They also facilitate the creation of scenario-driven narratives.

Additionally, we consider accessibility and usability important, as seen by the Gradle user interface. demonstrating our commitment to helping a variety of academics, practitioners, and artists have access to cutting-edge AI technology.

In the future, deep learning and artificial intelligence research might lead to the creation of videos in addition to further advancements in picture production techniques. By building on the foundation created in this study, we want to further enhance pre-trained DF GANs and encourage additional computer vision research.

References

1. Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K.-H., Poland, D., Borth, D., Li, L.-J.: YFCC100M: The New Data in Multimedia Research. Available at https://dl.acm.org/doi/10.1145/2812802
2. Lin, Microsoft COCO: Common Objects in Context. Available at https://arxiv.org/abs/1405.0312

3. Wah, The Caltech-UCSD Birds-200-2011 Dataset. Available at http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
4. Nilsback, Oxford 102 Dataset; Automated flower classification over a large number of classes. Available at http://www.robots.ox.ac.uk/ vgg/data/flowers/102/
5. Schuldt, KTH Action Recognition; Recognizing human actions: a local SVM approach. Available at http://www.nada.kth.se/cvap/actions/
6. Rodriguez, UCF Sports; Action mach a: a new representation for human action recognition. Available at https://www.crcv.ucf.edu/data/ucf_sports_actions/
7. Ding, M.: CogView: Mastering Text-to-Image Generation via Transformers. arXiv preprint arXiv:2106.13700, 2021
8. Biswas, Biswajit, Ghosh, Swarup Kr, Ghosh, Anupam", "DVAE: Deep Variational Auto-Encoders for Denoising Retinal Fundus Image",2020
9. Xu, Tao and Zhang, Pengchuan and Huang, Qiuyuan and Zhang, Han and Gan, Zhe and Huang, Xiaolei and He, Xiaodong, AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks, 2018
10. Tsue, T., Sen, S., Li, J.: Cycle Text-To-Image GAN with BERT. arXiv preprint arXiv:2003.12137 (2020)
11. Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: Learning Text-to-image Generation by Redescription. arXiv preprint arXiv:1903.05854 (2019)
12. Tao, M., Tang, H., Wu, F., Jing, X., Bao, B., Xu, C.: DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. arXiv preprint arXiv:2008.05865 (2022)
13. Srivastava, Nitish and Mansimov, Elman and Salakhudinov, Ruslan: Unsupervised Learning of Video Representations using LSTMs, Proceedings of the 32nd International Conference on Machine Learning, 2015
14. Zhang, Han and Xu, Tao and Li, Hongsheng and Zhang, Shaoting and Wang, Xiaogang and Huang, Xiaolei and Metaxas, Dimitris N., StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks, 2017
15. Ramesh, A., Chu, C., Dhariwal, P., Nichol, A., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125, 2022
16. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, Oran Gafni and Adam Polyak and Oron Ashual and Shelly Sheynin and Devi Parikh and Yaniv Taigman, 2022,