Members: Mansi Sudhirbhai Joshi, Nikhil Bharadwaj Narayanam and Pranav Barathwaj Padmanaban
INFO-H 518: Deep Learning Neural Networkss
Ph.D. Sunandan Chakraborty

TEXT2PICBOT: Generating Visuals From Words with GANs

Introduction

In the ever-evolving landscape of artificial intelligence and computer vision, our project ventures into the synthesis of lifelike images from textual descriptions. Fueled by the potential of Generative Adversarial Networks (GANs) and innovative training strategies, our objective is to craft realistic bird images that seamlessly emerge from detailed textual descriptions. This introduction sets the stage for our exploration into pushing the boundaries of image synthesis.

Problem Description

The core challenge our project addresses is the efficient and reliable synthesis of lifelike bird images based on comprehensive textual descriptions. To tackle this, we leverage the robust capabilities of Generative Adversarial Networks (GANs) along with an innovative and optimized training strategy. Our primary goal is to achieve a seamless translation of descriptive text into visually convincing bird images, encompassing diverse species and intricate details with an exceptional level of realism. This section lays out the specific problem we aim to solve and the key technologies we employ to address it.

Description of Data

Our project relies on the extensive Caltech-UCSD Birds-200-2011 dataset, an augmented version of the original CUB-200 dataset, for the synthesis of realistic bird images. This dataset, obtained from https://www.vision.caltech.edu/datasets/cub_200_2011/, is a rich resource featuring approximately double the images per category and detailed part location annotations, making it a comprehensive repository for avian imagery (Perona Lab - CUB-200-2011, 2011). Meticulous annotations for each image include 15 specific part locations, facilitating a nuanced understanding of avian anatomy, 312 binary attributes providing detailed information about various bird features, and a single bounding box annotation offering a precise delineation of each bird's spatial extent in the images (Perona Lab - CUB-200-2011, 2011). With 200 distinct bird categories, the dataset encompasses a broad spectrum of avian diversity, enabling our synthesis model to learn and replicate the characteristics of different bird species. These annotations and categories are valuable for training our model to generate lifelike bird images.

Methodology

Data Collection

Our data collection involved curating the Caltech-UCSD Birds-200-2011 (Perona Lab - CUB-200-2011, 2011) dataset. Complementing visual data, we integrated Image Text Descriptions https://drive.google.com/file/d/0B3y_msrWZaXLT1BZdVdycDY5TEE/view?resourcekey=0-sZrhftoEfdvHq6MweAeCjA, offering ten captions per image, class names, and filenames. This enriched dataset combines visual and textual context, pivotal for training our model. The meticulous curation provides a strong foundation, enabling our model to learn intricate relationships between textual descriptions and visual representations, advancing the synthesis of realistic bird images.

Data Pre-processing

Our data preprocessing pipeline plays a pivotal role in preparing the raw dataset for synthesis model training. We initiated the process by loading filenames, class IDs, and bounding box information. Filenames enabled the retrieval of corresponding visual data, while class IDs provided categorical details about depicted bird species. Precise bounding box information was crucial for subsequent cropping and resizing operations. Application of the loaded bounding box details facilitated focused cropping and resizing of images, isolating relevant bird regions. Processed data, including images (X), class IDs (y), and pre-trained sentence embeddings, were then appended to lists. These lists were further converted to NumPy arrays, optimizing numerical operations and ensuring compatibility with machine learning frameworks. This meticulous pipeline lays the groundwork for our synthesis model's effective training on curated and processed data.

```
['this bird has a bright yellow body, with brown on its crown and wings.\n',
 'this bird has a red breast and belly as well as a small bill.\n',
 'small, roundish bird with off white breast and belly, light brown crown, brown and black
colored wings.\n',
 'A white bird with a black crown and yellow beak\n',
 'the bird has gray crown, belly and white abdomen, with black tarsus and feet\n',
 'a colorful bird with a bright yellow body, a black crown and throat, orange bill, and
black primaries and secondaries.\n']
```

Figure 1: Text description after word-embedding

Model Architecture

StackGAN Architecture

The objective of the StackGAN architecture is to excel in text-to-image synthesis, focusing on the generation of high-resolution and photo-realistic images from textual descriptions. The architecture comprises two stages: Stage-I GAN, which takes a low-resolution image and corresponding text description as input, producing refined images capturing basic details (Zhang et al., 2017). Stage-II GAN, which takes the output of Stage-I (low-resolution image) along with the text description, generates high-resolution, photo-realistic images. This two-stage workflow progressively enhances image resolution and realism, contributing to the seamless translation of textual descriptions into visually convincing outputs (Zhang et al., 2017).

GAN-INT-CLS

GAN-INT-CLS, an early exploration into text-to-image synthesis utilizing Generative Adversarial Networks (GANs), features an RNN encoder coupled with a GAN decoder. The model operates by taking text embeddings as input, employing an RNN encoder, and utilizing a generative model as a decoder (Reed et al., 2016). However, a notable drawback is evident—the generated images lack sharpness and fall short of effectively replacing real images (Reed et al., 2016). This underlines the model's limitations and points towards the continuous evolution in the field of text-to-image synthesis

The StackGAN architecture was chosen for our project due to its advanced two-stage design, which has proven effective in generating realistic images from textual descriptions. The model's ability to capture both basic and intricate details aligns well with our objective of seamlessly translating the comprehensive text into visually convincing bird images.

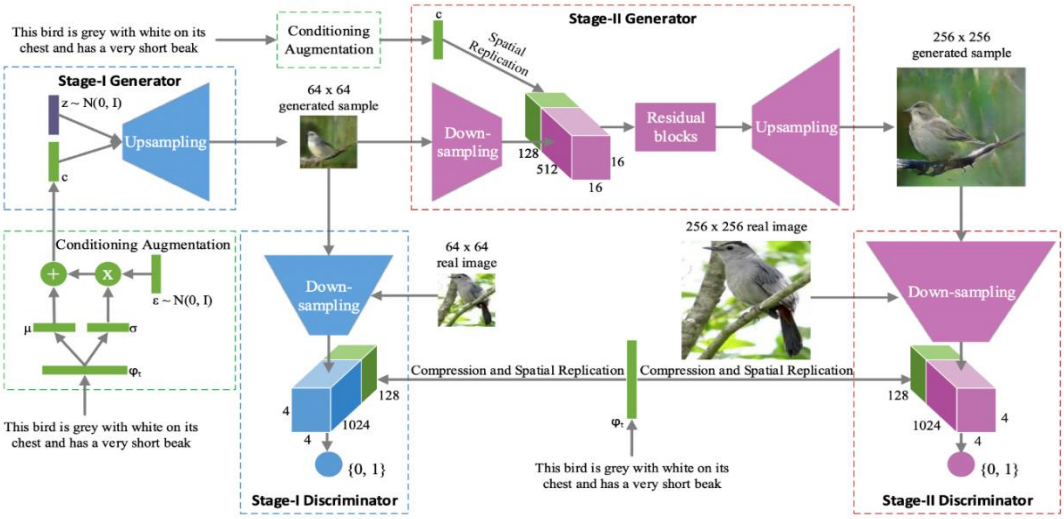

Figure 2: StackGAN Architecture used in the Project (Zhang et al., 2017)

Model Training

The model training process encompasses several key components tailored for synthesizing realistic bird images. The Conditional Augmentation (CA) Block plays a pivotal role in transforming input noise into a structured space, specifically designed for handling class labels and embeddings (Zhang et al., 2017). This customization contributes to the generation of diverse and contextually relevant bird images. The two-stage GAN architecture includes Stage 1 and Stage 2 Generators, progressively creating images with Stage 1 focusing on lower-resolution outputs and Stage 2 refining them for higher quality and realism (Reed et al., 2016). These generators collaboratively contribute to the multi-stage refinement process, ensuring the generation of visually convincing bird images. Aligned with each generator stage, the Stage 1 and Stage 2 Discriminators are tasked with distinguishing between real and fake images, providing continuous feedback to enhance the adversarial training process (Zhang et al., 2017). The Embedding Compressor Models process additional information, such as text embeddings related to bird descriptions, enhancing the integration of textual details and context for realistic image generation. The Adversial Models, constituting the Stacked GAN architecture represent the overall framework, combining generators and discriminators in a stacked configuration (Brownlee, 2019). This collaborative architecture facilitates the seamless translation of textual descriptions into high-quality and realistic bird images.

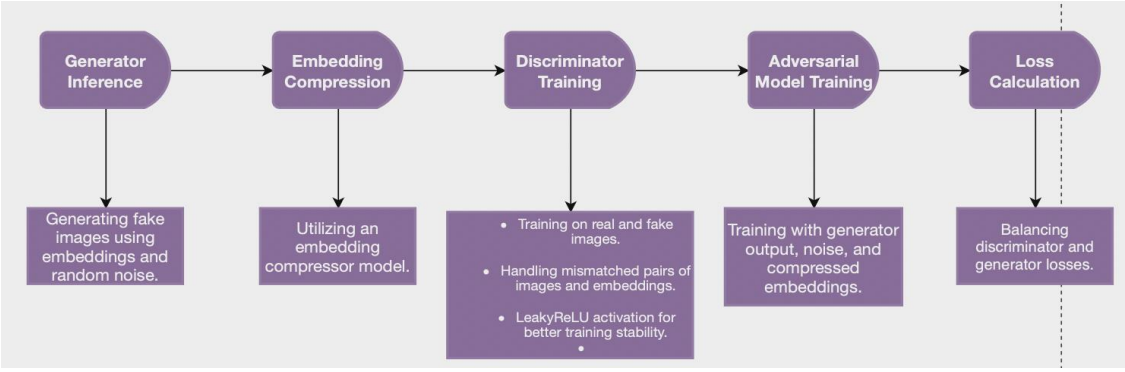

Figure 3: Training Process

Model Evaluation

Our evaluation strategy, which leans towards qualitative assessment for image synthesis, fits well with the creative nature of tasks that prioritize visual appeal and contextual relevance. Qualitative evaluation allows us to explore artistic details and context, aspects that quantitative metrics, might find challenging to capture effectively (Brownlee, 2019). This method offers a holistic understanding of how well the model translates textual descriptions into visually engaging and contextually fitting images. To gather participant feedback, we're considering a survey-based approach, prompting evaluations on how well the generated images align with textual descriptions. Despite the existence of quantitative metrics like Inception Score (IS) and Frechet Inception Distance (FID), their inclusion poses challenges. IS, while informative about diversity, doesn't directly assess visual quality or realism, potentially leading to less visually appealing results (Brownlee, 2019). Additionally, FID's computational intensity, especially with large datasets, makes it resource-demanding (Brownlee, 2019). Given the strengths and limitations of both approaches, the decision to stick to qualitative evaluation is grounded.

Results

Below are some generated images:

Figure 4: Image Generated 1

Figure 5: Image Generated 2

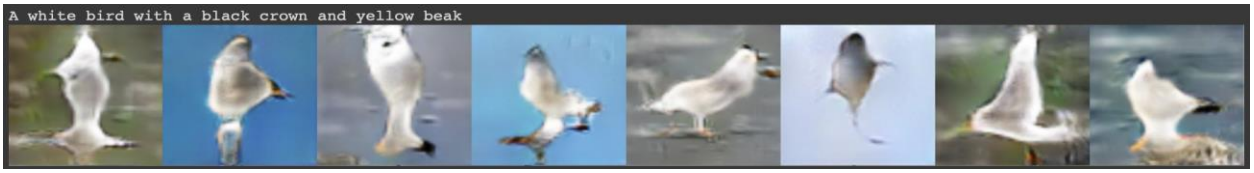

Figure 6: Image Generated 3

The above images were included in our survey and below are the observations from our survey.

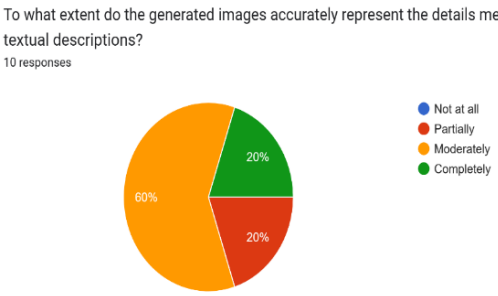

Figure 7: Question 1 of the Survey

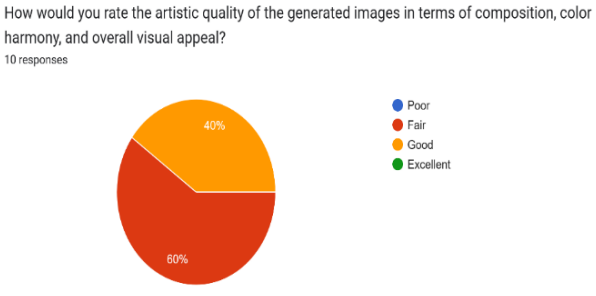

Figure 8: Question 2 of the Survey

As observed, most believe that the generated images are moderately accurate with respect to the textual descriptions. It is also noticeable that the majority consider the artistic quality of the generated images to be fair. We can interpret both of these responses as signs that the images are fairly good. However, there is still a need for better overall visual appeal so that anyone who sees the image can identify it without having to look for details. (Survey Link: https://docs.google.com/forms/d/e/1FAIpQLScd24JFUUOOks7BYtyknc6Y836_mmavWDu8EeXaNhr_rOZbyQ/viewform)

Limitations

Computational Resources

Despite upgrading to Colab Pro Plus, the vast CUB dataset still poses challenges for model training due to resource constraints. The increased computational resources haven't significantly alleviated limitations in running the model for extended epochs, hindering learning depth and potentially impacting optimal results. Additionally, restricted resources limit the model's exposure to the full dataset diversity.

Hyperparameter Tuning

GANs are notorious for their sensitivity to hyperparameters, and finding the optimal set can be a challenging and time-consuming process. Difficulty in tuning hyperparameters may lead to suboptimal model performance, affecting the quality and realism of the generated images. Fine-tuning is crucial for achieving the desired balance in the adversarial training dynamics.

Generalization to Unseen Data

StackGANs may struggle to generalize to unseen data, particularly if the training dataset lacks diversity, impacting the model's performance when faced with new, unseen data. This limitation could hinder the model's practical applicability in scenarios requiring accurate representation of diverse and unseen bird species, affecting the quality of its results.

Future Works

Experimenting on Larger Computational Resources for More Epochs

Leveraging larger computational resources can enhance the model training process, enabling more extensive sessions to capture intricate patterns and improve StackGAN performance. Longer training durations empower the model to learn complex representations, enhancing its capability to generate high-quality and realistic bird images, particularly beneficial for large and detailed datasets like CUB-200-2011.

Quantitative Evaluation Framework

Establishing a comprehensive quantitative evaluation framework, incorporating metrics such as Inception Score (IS) and Frechet Inception Distance (FID), will be crucial (Brownlee, 2019). Quantitative metrics provide numerical insights into the diversity, realism, and quality of generated images. Balancing both qualitative and quantitative evaluations ensures a more nuanced understanding of the model's strengths and areas for improvement.

Exploring Attention GAN for Better Results

Attention mechanisms have shown promise in improving the quality and focus of generated images. Integrating an attention mechanism, such as in an Attention GAN, can guide the model to pay more attention to relevant parts of the input text, resulting in more contextually accurate and visually appealing image synthesis.

References

Caltech-UCSD birds-200-2011 (CUB-200-2011). Perona Lab - CUB-200-2011. (2011). https://www.vision.caltech.edu/datasets/cub_200_2011/

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). *Generative Adversarial Text to Image Synthesis*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). *StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D (2018, February 25). *Stackgan-Pytorch*. GitHub. https://github.com/hanzhanggit/StackGAN-Pytorch

Brownlee, J. (2019, July 12). *How to evaluate generative Adversarial Networks*. MachineLearningMastery.com. https://machinelearningmastery.com/how-to-evaluate-generative-adversarial-networks/

Goodfellow, I., Bengio, Y., & Courville, A. (2023). *Deep learning*. Alanna Maldonado.

Appendix 1: Contributions from each member

Mansi Sudhirbhai Joshi

Introduction & Problem Description, Data Collection, Model Architecture, Model Training, Limitations, Future Work, Project Presentation & Report

Nikhil Bharadwaj Narayanam

Introduction & Problem Description, Data Pre-Processing, Model Architecture, Model Training, Limitations, Future Work, Project Presentation & Report

Pranav Barathwaj Padmanaban

Introduction & Problem Description, Model Architecture, Model Evaluation, Limitations, Future Work, Project Presentation & Report