Image Synthesis

Venkata Anantha Reddy Arikatla
Pranav Pinjarla
College of Engineering
Northeastern University
Boston, MA 02115 arikatla.v@northeastern.edu
pinjarla.p@northeastern.edu

Abstract

This project presents an application of advanced machine learning techniques to generate high-quality images from textual descriptions by integrating pre-trained models into an efficient synthesis pipeline. Utilizing the Stable Diffusion v2 model, a cutting-edge diffusion-based text-to-image generation framework available through Hugging Face, in conjunction with the GPT-2 text generation model, this study showcases the practical implementation of these technologies. The project automates the entire process, from generating prompts to creating images, resulting in diverse and contextually rich visual outputs. By leveraging Hugging Face's comprehensive model repository and diffusers library, the project demonstrates how state-of-the-art models can be easily deployed for creative and research purposes. The work not only highlights the capabilities of these models but also explores their limitations and the ethical considerations inherent in their use. This integration underscores the potential of combining natural language processing with image synthesis, offering valuable insights for applications in content creation, digital art, and beyond.

1 Introduction

The convergence of Natural Language Processing (NLP) and Computer Vision has opened up new avenues for creative and research applications, particularly in the field of text-to-image synthesis. As models and algorithms continue to advance, it has become possible to generate high-quality, detailed images based on simple textual descriptions, a task that was previously thought to be far beyond the reach of machines.

This project explores the implementation of a text-to-image synthesis pipeline, leveraging cutting-edge pre-trained models to automate the process of generating images from user-defined text prompts. The core of this project is the integration of the Stable Diffusion v2 model and the GPT-2 text generation model, both of which are accessed and deployed via Hugging Face's diffusers library.

Stable Diffusion v2 is a state-of-the-art latent diffusion model designed for generating high-resolution images from textual inputs. Developed by Robin Rombach and colleagues, the model utilizes a powerful text encoder to translate textual descriptions into corresponding visual representations. The diffusion process, which underlies the model's functionality, iteratively

refines a noisy image into a clear and coherent final output, guided by the textual input provided. This model was trained on large-scale datasets, ensuring that it can handle a wide range of subjects and styles, although it does have limitations, particularly in handling non-English prompts and complex compositional tasks.

GPT-2, developed by OpenAI, is a transformer-based language model that generates coherent and contextually relevant text from a given seed prompt. In this project, GPT-2 is employed to create diverse and imaginative textual prompts that serve as the input for the Stable Diffusion model. The synergy between these two models allows for the automatic generation of a variety of images, each reflecting different interpretations of the same initial concept.

The integration of these models, facilitated by Hugging Face's diffusers library, provides a seamless and efficient workflow for text-to-image synthesis. The diffusers library allows for easy deployment and customization of these powerful models, making it accessible even for those with limited machine learning experience. Additionally, by using pre-trained models, the project circumvents the need for extensive computational resources, making the process more feasible and scalable.

This project aims to demonstrate the practical application of these technologies by automating the pipeline from prompt generation to image creation. By doing so, it highlights the potential of text-to-image synthesis in various domains, including digital art, content creation, and educational tools. The generated images are evaluated for their quality, diversity, and contextual relevance, providing insights into the strengths and limitations of the models used. Furthermore, the project considers the ethical implications of using generative models, particularly in relation to biases and potential misuse, underscoring the importance of responsible deployment.

In conclusion, this project not only showcases the technical capabilities of modern text-to-image synthesis models but also provides a foundation for further exploration and application of these technologies in creative and research contexts.

2 Method

This section details the methodology employed in the project, encompassing the dataset preparation, model architecture, training process, and implementation of the image generation pipeline. The integration of these components facilitated the automatic generation of images based on textual prompts.

The architecture of this project is based on two main pre-trained models: **Stable Diffusion v2** for image generation and **GPT-2** for text generation. Both models are accessed through Hugging Face's diffusers and transformers libraries, which simplify the integration and deployment of these powerful models.

Stable Diffusion v2 Model:

- **Model Type**: Latent Diffusion Model (LDM).
- **Function**: Generates images from textual descriptions by encoding the text into latent representations and iteratively refining noisy images into coherent visual outputs.
- **Training**: The model was fine-tuned for 150,000 steps on a 512x512 resolution dataset and further refined for 140,000 steps on a 768x768 resolution dataset.

GPT-2 Model:

- **Model Type**: Transformer-based language model.
- **Function**: Generates textual prompts that are used as input for the Stable Diffusion model.
- **Training**: Trained on a large corpus of internet text, allowing it to generate contextually appropriate prompts from a given seed text.

Pipeline Integration:

- The project utilizes the diffusers library to create an integrated pipeline where GPT-2 generates multiple text prompts based on an initial seed. These prompts are then passed to the Stable Diffusion model, which generates corresponding images. The pipeline ensures that the process is seamless and can be easily adapted for various applications.

3.3 Training

Since pre-trained models were used in this project, the emphasis was on **inference** rather than training. The pre-trained models were directly fetched from Hugging Face's repository and utilized within the pipeline to generate images. Below are the key steps involved in the process:

1. **Model Initialization**:
    - The models were initialized using the StableDiffusionPipeline and pipeline functions from the diffusers and transformers libraries, respectively. The Stable Diffusion model was configured with specific parameters, such as guidance_scale and num_inference_steps, to control the quality and style of the generated images.
2. **Text Prompt Generation**:
    - The GPT-2 model generated a series of text prompts based on an initial input seed. This was achieved through the generate_prompts function, which used the prompt_generator object to create multiple variations of the seed text.
3. **Image Generation**:
    - The generated prompts were fed into the Stable Diffusion model via the create_image_from_text function. This function utilized the model to generate images based on the textual descriptions provided by GPT-2.
    - The generate_and_save_images function automated the process of generating and saving multiple images based on different prompts, ensuring that each output was documented and stored.
4. **Logging and Output Management**:
    - A logging mechanism was set up to track the generation process, recording each prompt and its corresponding image in a log file (generation.log). This step ensured that the entire process was transparent and traceable.
5. **Image Display**:
    - Finally, the generated images were displayed in a grid format using the display_images function. This function read the saved images from the output directory and organized them into a visual grid, making it easy to evaluate the results.

4 Results

The results of this project demonstrate the effectiveness of the text-to-image synthesis pipeline in generating visually compelling images based on a user-defined prompt. The images generated from the prompt "Northeastern University" highlight the model's ability to interpret and visualize complex concepts, providing diverse outputs that align with the essence of the given prompt.

Generated Images Overview

The generated images exhibit a range of perspectives and interpretations of the "Northeastern University" prompt. The outputs include aerial views of urban campuses, detailed architectural representations of buildings, and more abstract interpretations that reflect the model's flexibility and creative potential. The images vary in style and composition, showcasing the model's capability to produce diverse and contextually rich visual outputs from a single textual input.

Description of the Generated Images:

1. **Top-Left Image**: An aerial view of what appears to be a modern university campus, with a mix of green spaces and large academic buildings. The image suggests a vibrant, well-organized campus environment, typical of a large university setting.
2. **Top-Center Image**: A close-up of a contemporary building facade, with glass windows and clean architectural lines. The image likely represents a modern university building, possibly an administrative or academic block, highlighting the urban and professional aspect of the university.
3. **Top-Right Image**: Another aerial view, this time focusing on a different section of the university. The image features a highway adjacent to the campus, suggesting a well-connected urban environment. The buildings appear large and institutional, reinforcing the impression of a significant academic establishment.
4. **Bottom-Left Image**: This image offers a detailed perspective of a section of the campus, showcasing both the diversity in building design and the surrounding green spaces. The circular structure in the center adds a unique architectural element, making this output distinct.
5. **Bottom-Center Image**: A similar aerial view, emphasizing the extensive layout of the university campus. The buildings are interconnected, and the presence of greenery suggests an environment that balances academic infrastructure with natural elements.
6. **Bottom-Right Image**: Unlike the other images, this one portrays a person standing against a wall, possibly an abstract interpretation or an alternative perspective related to the university's community or culture. The person appears to be in a contemplative pose, adding a human element to the series of architectural images.

Evaluation

The images generated by the Stable Diffusion v2 model reflect the robustness of the text-to-image synthesis process. The variety in the images demonstrates the model's ability to produce

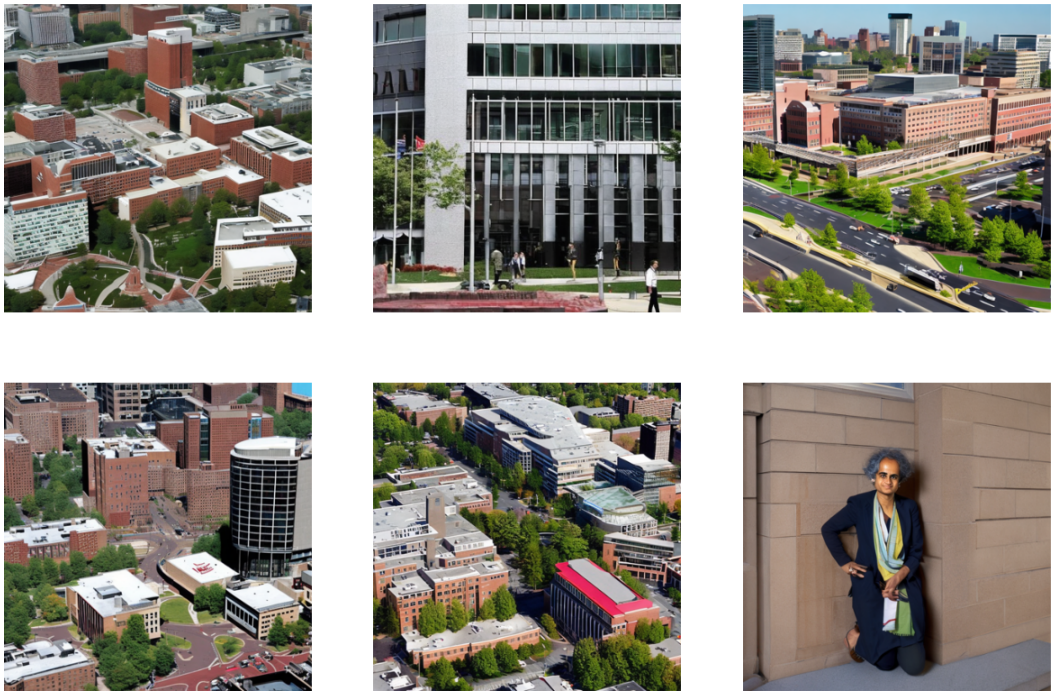
multiple interpretations of the same prompt, providing a rich set of visual outputs that can be used in various contexts, from academic presentations to creative projects.

Strengths:

- **Diversity of Outputs**: The generated images offer different perspectives and styles, showing the model's capacity to explore various visual interpretations of a single concept.
- **Quality of Images**: The images are detailed and well-structured, with a good balance between architectural precision and artistic creativity.

Limitations:

- **Abstract Interpretations**: While the majority of the images align with the expected output, some images, like the bottom-right one featuring a person, may appear abstract or unrelated to the main theme. This reflects the inherent unpredictability of generative models, which can sometimes produce unexpected results.
- **Contextual Relevance**: Although the images generally represent a university setting, the specific connection to "Northeastern University" is more abstract and generalized, as the model does not have direct knowledge of specific institutions.



5. Discussion

The results of this project demonstrate the capabilities of modern text-to-image synthesis models in generating diverse and contextually rich images from textual prompts. The integration of the Stable Diffusion v2 and GPT-2 models into a cohesive pipeline has proven effective in automating the image generation process, with outputs that vary in style and perspective. However, the project also highlights some challenges and limitations that can be addressed in future work.

Future Experiments/Work

Given more time and resources, several avenues for further experimentation and enhancement could be explored:

1. **Model Fine-tuning**:
    - **Domain-Specific Fine-tuning**: Fine-tuning the Stable Diffusion v2 model on a dataset specific to a particular domain (e.g., academic campuses, urban architecture) could improve the contextual relevance of the generated images. This would allow the model to produce outputs that are more closely aligned with the specific features and characteristics of the target environment.
2. **Enhanced Text-to-Image Integration**:
    - **Interactive Prompt Refinement**: Implementing an interactive system where users can refine and adjust prompts based on preliminary outputs could lead to more precise and desired results. This would involve real-time feedback loops where the user can iteratively modify the text prompts and observe changes in the generated images.
3. **Incorporation of Multimodal Inputs**:
    - **Image and Text Input Combination**: Future experiments could involve combining text prompts with existing images to create even more contextually accurate outputs. This multimodal approach would allow users to guide the model more precisely by providing visual cues alongside text.
4. **Exploration of Ethical and Bias Mitigation Strategies**:
    - **Bias Detection and Mitigation**: Further work could focus on developing strategies to detect and mitigate biases inherent in the models, especially when generating images involving diverse communities and cultures. This could include incorporating fairness metrics and debiasing techniques into the pipeline.
5. **Scaling and Deployment**:
    - **Scalable Cloud-Based Deployment**: Deploying the pipeline in a scalable cloud environment could facilitate the generation of large volumes of images, making the system suitable for industrial applications such as content generation and virtual environment creation.
6. **Integration with Creative Tools**:
    - **Creative and Educational Applications**: Integrating the model with creative tools like Adobe Photoshop or educational platforms could provide users with a seamless experience for generating and refining visual content. This could be particularly valuable in fields like digital art, design, and education.

Supplementary Materials

To support further research and replication of this work, several supplementary materials are recommended:

1. **Jupyter Notebooks**:
    - The Jupyter Notebook used in this project can be provided as a supplementary material to allow other researchers and practitioners to replicate the process and experiment with different configurations.
2. **Sample Datasets**:

- While the project used pre-trained models, providing sample datasets, such as subsets of LAION-5B or other image-text pairs, could be valuable for researchers interested in fine-tuning the models.

3. **Model Configuration Files**:
   - Configuration files detailing the specific parameters and settings used for both the Stable Diffusion and GPT-2 models can be provided. This would help others to replicate the results or adapt the pipeline for different use cases.
4. **Generated Image Gallery**:
   - A gallery of the generated images, categorized by prompt and model settings, could be included as a supplementary material. This would provide a visual reference for the quality and diversity of outputs produced by the model.

6. Conclusion

This project successfully demonstrated the application of pre-trained machine learning models for text-to-image synthesis. By integrating the Stable Diffusion v2 model with GPT-2, a seamless pipeline was created that could generate diverse and high-quality images based on textual prompts. The generated images highlighted the model's capability to interpret and visualize complex concepts, offering a valuable tool for creative and research purposes.

While the project achieved its primary objectives, it also uncovered areas for further exploration, including fine-tuning the models, enhancing the text-to-image integration process, and addressing ethical considerations related to bias. The work serves as a foundation for future research and development in the field of text-to-image synthesis, with potential applications in digital art, content creation, education, and more.

7. Works Cited

1. **Rombach, Robin, et al. "High-Resolution Image Synthesis With Latent Diffusion Models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 10684-10695.**
   - This paper details the development and application of Latent Diffusion Models, including the Stable Diffusion v2 model used in this project.
2. **Radford, Alec, et al. "Language Models are Unsupervised Multitask Learners." OpenAI, 2019.**
   - This paper discusses the development of GPT-2, the text generation model used for generating prompts in this project.
3. **Hugging Face. "Stable Diffusion v2 Model Card." Hugging Face, 2022.**
   - This model card provides an overview of the Stable Diffusion v2 model, including details about its training data, architecture, and intended use cases.
4. **Hugging Face. "GPT-2 Model Card." Hugging Face, 2019.**
   - This model card provides information about the GPT-2 model, including its architecture, training data, and potential applications.
5. **LAION-5B: "A Large-Scale Dataset for Image-Text Training." NeurIPS, 2022.**
   - This paper provides an overview of the LAION-5B dataset, which was used in the training of the Stable Diffusion model.

8. Acknowledgement

This project was made possible by the pre-trained models and resources provided by Hugging Face, specifically the Stable Diffusion v2 and GPT-2 models. Special thanks to the developers and researchers who contributed to these models, as well as the Hugging Face community for making these tools accessible to the public. Their work has significantly advanced the field of text-to-image synthesis, enabling projects like this one to be realized with minimal computational overhead.

Appendix A: Code Snippets

This section includes key parts of the code used in the project, allowing others to understand or replicate the workflow.

A.1 Configuration Settings

In [4]:
```python
# Configuration settings
class Config:
    device_type = "cuda"
    random_seed = 42
    torch_gen = torch.Generator(device_type).manual_seed(random_seed)
    num_steps = 35
    model_id = "stabilityai/stable-diffusion-2"
    output_image_size = (400, 400)
    guidance_factor = 9
    text_model_id = "gpt2"
    dataset_size = 6
    max_text_length = 12
    output_dir = "generated_images"
    log_file = "generation.log"
```

A.2 Model Initialization

In [6]:
```python
# Load the image generation model
diffusion_pipeline = StableDiffusionPipeline.from_pretrained(
    Config.model_id, torch_dtype=torch.float16,
    revision="fp16", use_auth_token='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',
    guidance_scale=Config.guidance_factor
)
diffusion_pipeline = diffusion_pipeline.to(Config.device_type)
```

In [7]:
```python
# Load the text generation model for prompt generation
prompt_generator = pipeline("text-generation", model=Config.text_model_id)
set_seed(Config.random_seed)
```

A.3 Image Generation and Saving

In [14]:
```python
# Function to generate an image based on a text prompt
def create_image_from_text(text_prompt, pipeline_model):
    result_image = pipeline_model(
        text_prompt, num_inference_steps=Config.num_steps,
        generator=Config.torch_gen,
        guidance_scale=Config.guidance_factor
    ).images[0]

    resized_image = result_image.resize(Config.output_image_size)
    return resized_image
```

In [15]:
```python
# Function to generate prompts using GPT-2
def generate_prompts(prompt_seed, num_prompts):
    prompts = prompt_generator(
        prompt_seed,
        max_length=Config.max_text_length,
        num_return_sequences=num_prompts
    )
    return [p['generated_text'] for p in prompts]
```

In [17]:
```python
# Function to generate and save multiple images
def generate_and_save_images(prompts, model, save_dir):
    if not os.path.exists(save_dir):
        os.makedirs(save_dir)

    for idx, prompt in enumerate(prompts):
        image = create_image_from_text(prompt, model)
        image_path = os.path.join(save_dir, f"image_{idx+1}.png")
        image.save(image_path)
        logging.info(f"Generated image for prompt: '{prompt}' and saved to '{image_path}'")
```

Appendix B: System and Environment Setup

B.1 Required Installations

- The following Python packages are required to run the project:

```
In [2]: # Required installations
        !pip install --upgrade diffusers transformers -q
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.7/43.7 kB 2.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.6/2.6 MB 40.4 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 9.5/9.5 MB 69.2 MB/s eta 0:00:00
```

B.2 Environment Details

- The code was run in a Google Colab environment, which provides GPU support (CUDA) for efficient model inference. The specific version of CUDA used in the environment was 12.1.

B.3 Model and Resource Links

- **Stable Diffusion v2**: Hugging Face - Stable Diffusion v2
- **GPT-2**: Hugging Face - GPT-2
- **Diffusers Library**: Hugging Face - Diffusers

Appendix C: Generated Image Samples

C.1 Northeastern University Prompt Outputs

- The images generated from the "Northeastern University" prompt are included in this report. Each image represents a different interpretation of the prompt, demonstrating the model's versatility in generating diverse outputs.

C.2 Output Directory Structure

- All generated images are stored in the `generated_images` directory with filenames indicating their corresponding prompts: