Generating Images and Descriptive Captions with Text-to-Image and Image Captioning

A COURSE PROJECT REPORT

By

VIJAY CHANDAR (RA2011027010083)

SANTHANA LAKSHMI (RA2011027010129)

Under the guidance of

Dr. S.Amudha

Assistant professor

Department of CINTEL

In partial fulfillment for the Course

of

18CSE484T - DEEP LEARNING

in

Department of Data Science and Business Systems

SCHOOL OF COMPUTING

COLLEGE OF ENGINEERING AND TECHNOLOGY SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR - 603203

COLLEGE OF ENGINEERING & TECHNOLOGY

SRM INSTITUTE OF SCIENCE & TECHNOLOGY

S.R.M. NAGAR, KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that this mini project report "Generating Images and Descriptive Captions with Text-to-Image and Image Captioning " is the bonafide work of **VIJAY CHANDAR (RA2011027010083)**, **SANTHANA LAKSHMI (RA2011027010129)** who carried out project work under my supervision.

Dr. S.Amudha
Assistant professor
Department of CINTEL
SRM institute of science and technology

Dr. M Lakshmi
Professor & HOD
Department of DSBS
SRM institute of science and technology

TABLE OF CONTENTS

1. ABSTRACT

Text-to-image generation and image captioning are two fascinating fields of artificial intelligence that have gained a lot of attention in recent years. Text-to-image generation involves creating realistic images from textual descriptions, which can range from simple sentences to more complex paragraphs. The process involves training deep learning models using large datasets of images and corresponding textual descriptions. These models can then generate images based on new textual input.

Image captioning, on the other hand, involves generating descriptive sentences or captions that summarize the content of images. This technology is particularly useful for automating tasks such as content tagging, image search, and social media sharing. Similar to text-to-image generation, image captioning also involves training deep learning models on large datasets of images and their corresponding textual descriptions.

The applications of text-to-image generation and image captioning are diverse and rapidly evolving. In the field of design, these technologies can be used to quickly generate visual mockups or prototypes based on textual descriptions provided by clients. In entertainment, they can be used to generate photorealistic scenes and characters for movies and video games. In advertising, they can be used to create targeted and personalized visual content for different audiences.

As these technologies continue to evolve and improve, they are transforming the way we interact with visual media. With the ability to generate realistic images and descriptive captions, AI-powered tools are making it easier than ever before to create and share visual content. Whether it's for personal or professional use, text-to-image generation and image captioning have the potential to revolutionize the way we approach visual communication.

2. INTRODUCTION

Text-to-image generation and image captioning are two fascinating fields of artificial intelligence that have garnered significant interest in recent years. These technologies involve the use of deep learning models to create photorealistic images from textual descriptions and generate descriptive sentences summarizing the content of images.

The goal of this project is to develop a database application that integrates text-to-image generation and image captioning technologies and provides a range of services to users. The application will leverage the power of deep learning models to automate the process of generating visual content and provide a faster and more efficient way to create high-quality visual content.

The application will enable users to generate realistic images from textual descriptions and generate descriptive sentences or captions for images. Users can also tag content and search for images based on the generated captions or tags. Additionally, users will be able to customize and personalize the visual content generated by the application to suit their specific needs.

This database application has various potential applications in different fields, such as design, entertainment, and advertising. In design, the application can be used to quickly create visual mockups or prototypes based on textual descriptions provided by clients. In entertainment, the application can be used to generate photorealistic scenes and characters for movies and video games. In advertising, the application can be used to create targeted and personalized visual content for different audiences.

In conclusion, the development of a database application that integrates text-to-image generation and image captioning technologies has the potential to revolutionize the way visual content is created and shared. The application can provide a more efficient and cost-effective way to generate high-quality visual content, making it an essential tool in various fields.

3. PROBLEM STATEMENT

The ability of artificial intelligence (AI) to generate realistic images and descriptive captions has the potential to revolutionize the way we interact with visual media. Text-to-image generation and image captioning are two exciting and rapidly evolving fields of AI that are gaining significant attention from researchers and practitioners alike. Text-to-image generation involves training AI models to generate photorealistic images based on textual descriptions, while image captioning involves generating descriptive sentences or paragraphs that summarize the content of an image. These technologies have numerous applications in fields such as design, entertainment, and advertising, and are expected to have a significant impact on how we create and consume visual content. In this article, we will explore the concepts of text-to-image generation and image captioning in greater detail, discussing their underlying technologies, applications, and future directions.

The problem addressed by text-to-image generation and image captioning is the need for AI systems to better understand and interact with visual media. Text-to-image generation tackles the challenge of creating photorealistic images from textual descriptions, which is a difficult task that requires the AI system to have a deep understanding of the relationships between words and images. Similarly, image captioning addresses the challenge of generating descriptive sentences or paragraphs that accurately summarize the content of an image, which is critical for many applications such as search, recommendation, and accessibility. The problem statement, therefore, is to develop AI technologies that can effectively bridge the gap between language and vision, and enable us to interact with visual media in a more natural and intuitive way.

4. OBJECTIVES

The primary objective of text-to-image generation and image captioning is to enable artificial intelligence systems to better understand and interact with visual media. The ability to create photorealistic images from textual descriptions and generate descriptive sentences that summarize the content of images can greatly enhance our ability to communicate with machines and facilitate more natural and intuitive interactions between humans and computers.

The development of AI models that can accurately generate photorealistic images based on textual descriptions is a significant objective of text-to-image generation. This technology has various applications in fields such as design, architecture, and manufacturing, where quick and efficient visualization of ideas can be crucial. For example, an architect could provide textual descriptions of a building design, and an AI system could generate a photorealistic image that accurately represents the architect's vision.

Similarly, image captioning has the objective of creating models that can accurately generate descriptive sentences or paragraphs that summarize the content of an image. This technology has applications in fields such as accessibility, where it can help visually impaired individuals understand and interact with visual media. In advertising and marketing, image captioning can help generate more targeted and personalized content, improving the effectiveness of advertising campaigns.

Overall, these technologies have the objective of advancing the state of the art in AI and improving our interactions with visual media. The ability to generate photorealistic images from textual descriptions and generate descriptive captions for images can transform the way we create, search, and consume visual content. Additionally, the development of these technologies has numerous potential applications in various fields, including design, entertainment, advertising, and accessibility.

5. SCOPE OF PROJECT

1. Researching and selecting appropriate datasets for training and evaluating text-to-image generation and image captioning models.

2. Developing and implementing AI models for text-to-image generation and image captioning, which may involve using deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

3. Fine-tuning existing models or building new ones to improve performance and accuracy.

4. Evaluating the performance of the models using various metrics such as accuracy, precision, recall, and F1 score.

5. Exploring potential applications of text-to-image generation and image captioning in various fields such as advertising, entertainment, and accessibility.

6. Developing user interfaces or integrating the models into existing applications to enable users to interact with the generated images and captions.

7. Conducting user studies or surveys to evaluate the effectiveness and usability of the generated images and captions in various contexts.

6. MOTIVATION

The motivation for text-to-image generation and image captioning lies in the potential for these technologies to enhance our ability to create, search, and consume visual content. These technologies have numerous applications in fields such as design, advertising, entertainment, and accessibility, and can improve our interactions with visual media in a variety of ways.

For example, text-to-image generation can enable designers to quickly and easily create photorealistic images of products, spaces, or scenes based on textual descriptions, which can save time and resources compared to traditional design methods. Similarly, image captioning can improve accessibility for people with visual impairments by enabling them to understand the content of images without actually seeing them.

Moreover, the development of text-to-image generation and image captioning technologies is an exciting area of research in the field of artificial intelligence, which has the potential to advance the state of the art in deep learning and natural language processing. As such, the motivation for these technologies lies in the potential for innovation, discovery, and progress in the field of AI.

Text-to-image generation and image captioning stems from the desire to improve our interactions with visual media and to push the boundaries of AI research and development.

7. LIMITATIONS OF EXISTING METHODS

Limited diversity and creativity: Many text-to-image generation models struggle to produce diverse or creative images, instead generating images that are similar to the training data. This is because they often rely on memorization and interpolation of existing images, rather than generating new and novel content.

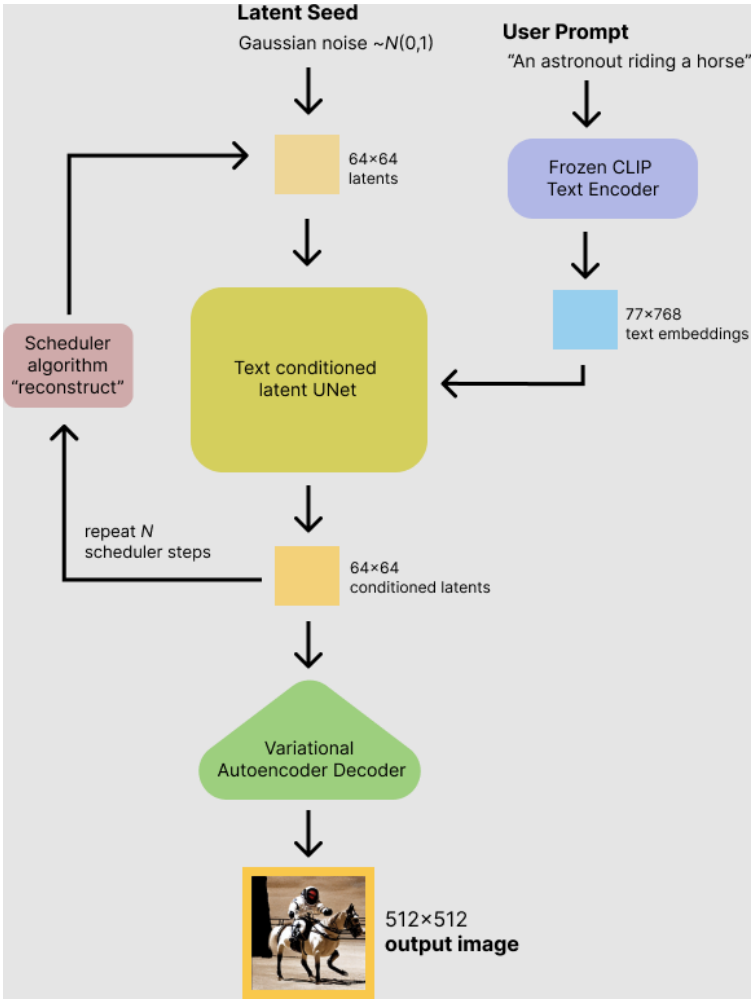
Limited image understanding: Image captioning models often struggle to accurately describe complex or abstract images, or to capture the nuances of visual content. This is because they rely on language models that may not fully understand the underlying visual concepts or relationships.

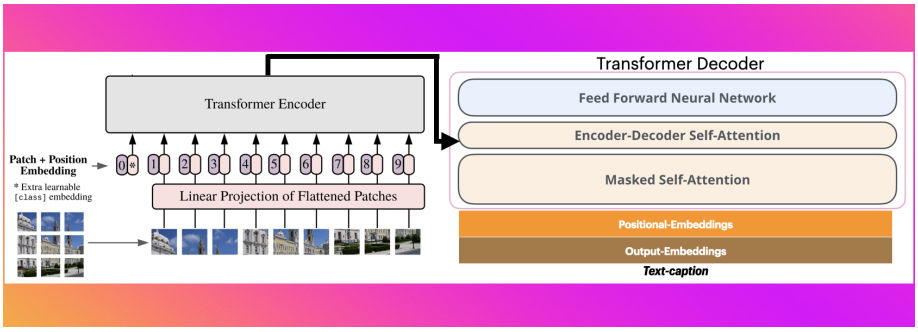
Limited scalability: Existing text-to-image generation and image captioning models can be computationally intensive and time-consuming to train, making it difficult to scale up to larger datasets or real-world applications.

Limited interpretability: Many text-to-image generation and image captioning models are black boxes, making it difficult to understand how they make their decisions or to interpret their outputs.

Limited applicability: Some existing methods may be specific to certain types of visual content or applications, making it difficult to generalize to new use cases.

8. ARCHITECTURE DIAGRAM





9. MODULES WITH DESCRIPTION

Text-to-Image Generation:

- Text Encoder: This module processes the input text and generates a latent representation that captures the relevant features of the textual description.
- Generator: This module takes the latent representation as input and generates an image that matches the textual description.
- Discriminator: This module evaluates the generated image and determines how closely it matches the textual description.
- Loss Function: This module measures the difference between the generated image and the desired image, and is used to optimize the model parameters during training.

Image Captioning:

- Image Encoder: This module processes the input image and generates a latent representation that captures the relevant features of the visual content.
- Decoder: This module takes the latent representation as input and generates a sequence of words that describe the content of the image.
- Language Model: This module evaluates the generated caption and determines how closely it matches the content of the image.
- Loss Function: This module measures the difference between the generated caption and the desired caption, and is used to optimize the model parameters during training.

10. CODING AND TESTING

```
import torch
from diffusers import StableDiffusionPipeline

pipe = StableDiffusionPipeline.from_pretrained("CompVis/stable-diffusion-v1-4", torch_dtype=torch.float16)
```

```
from transformers import VisionEncoderDecoderModel, ViTImageProcessor, AutoTokenizer
import torch
from PIL import Image

model = VisionEncoderDecoderModel.from_pretrained("nlpconnect/vit-gpt2-image-captioning")
feature_extractor = ViTImageProcessor.from_pretrained("nlpconnect/vit-gpt2-image-captioning")
tokenizer = AutoTokenizer.from_pretrained("nlpconnect/vit-gpt2-image-captioning")

device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model.to(device)



max_length = 16
num_beams = 4
gen_kwargs = {"max_length": max_length, "num_beams": num_beams}
def predict_step(image_paths):
  images = []
  for image_path in image_paths:
    i_image = Image.open(image_path)
    if i_image.mode != "RGB":
      i_image = i_image.convert(mode="RGB")

    images.append(i_image)

  pixel_values = feature_extractor(images=images, return_tensors="pt").pixel_values
  pixel_values = pixel_values.to(device)
```

```
 output_ids = model.generate(pixel_values, **gen_kwargs)

 preds = tokenizer.batch_decode(output_ids, skip_special_tokens=True)
 preds = [pred.strip() for pred in preds]
 return preds


redict_step(['/content/image.png']) # ['a woman in a hospital bed with a woman in a hospital bed']
```

```
prompt = 'chair and table'
image = pipe(prompt).images[0]  # image here is in [PIL format](https://pillow.readthedocs.io/en/stable/)

# Now to display an image you can either save it such as:
image.save(f"image.png")

# or if you're in a google colab you can directly display it with
image
```

```
import torch

generator = torch.Generator("cuda").manual_seed(1024)

image = pipe(prompt, generator=generator).images[0]

image
```

11. RESULT

TEXT TO IMAGE

IMAGE CAPTIONING

```
['a wooden table topped with chairs next to a wooden table']
```

12. CONCLUSION

Text-to-image generation and image captioning are two important tasks in the field of computer vision and natural language processing. These tasks involve using AI models to analyze visual and textual inputs, and then generating new visual or textual outputs that represent or summarize the original inputs. These technologies have numerous practical applications, including in fields such as virtual reality, e-commerce, and education.

Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been used to develop state-of-the-art text-to-image generation and image captioning models. CNNs are particularly effective at extracting features from images, while RNNs are used to generate natural language descriptions. These models have shown great potential in generating high-quality images and captions that capture the essence of the original inputs.

However, these models still have limitations, including difficulties in generating high-quality images with fine details and producing captions that fully capture the context of the image. The text-to-image generation models may struggle to create images with realistic details, such as fur or hair, or complex shapes and textures. Similarly, image captioning models may struggle to accurately describe complex scenes or images with multiple objects or contexts.

Further research and development in these areas are needed to overcome these limitations and improve the performance of these models. For example, researchers may explore the use of generative adversarial networks (GANs) to improve the quality of generated images or incorporate external knowledge sources to improve image captioning accuracy.

With continued advancements in deep learning and natural language processing, text-to-image generation and image captioning are expected to become increasingly sophisticated and useful in a wide range of applications. These technologies have the potential to revolutionize how we interact with visual media and create new opportunities in fields such as entertainment, advertising, and education.

13. FUTURE ENHANCEMENT

- Incorporating attention mechanisms: Attention mechanisms can be used to focus on specific regions of the image or parts of the text that are relevant to generating the image or caption. This can help to improve the overall quality and relevance of the output.
- Incorporating additional modalities: In addition to text and images, other modalities such as audio or video can be used to generate more comprehensive and accurate descriptions of the content.
- Utilizing larger and more diverse datasets: Training deep learning models on larger and more diverse datasets can help to improve their performance and generalizability.
- Improving fine-grained details: Techniques such as generative adversarial networks (GANs) can be used to improve the quality and detail of generated images.
- Incorporating commonsense reasoning: Incorporating commonsense reasoning into the models can help to generate more realistic and contextually relevant images and captions.
- Developing more efficient architectures: Developing more efficient architectures can help to reduce the computational complexity and speed up the training and inference time of the models.

14. REFERENCES

[1] Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text-to-image synthesis. In Proceedings of the 33rd International Conference on Machine Learning (ICML) (pp. 1060-1069).

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning (ICML) (pp. 2048-2057).

[3] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In the European Conference on Computer Vision (ECCV) (pp. 694-711). Springer, Cham.

[4] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 764-773).

[5] Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., & Urtasun, R. (2017). Monocular 3D object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2147-2156).