Text Conditioned Image Generation Using Diffusion Models

Dhanya Srinivasan Maheshwari M R Dr. P. Mirunalini

Department of CSE, Sri Sivasubramaniya Nadar College of Engineering
G3_8, Final Year Project, May 2024

Highlights of Proposed Model

To develop a diffusion model

- To facilitate the generation of visually realistic and semantically aligned images.
- That utilizes the use of shared modalities representations to discover the relationship between image-text pairs.
- To generate a highly semantically-aligned image even after training on a small dataset

Challenges in generating realistic images:

- To train the VQ-VAE model using both text and picture embeddings instead of just utilizing images as is currently the case
- The objects might not appear consistent in the generated photos due to the use of artistic images.

Proposed Model architecture

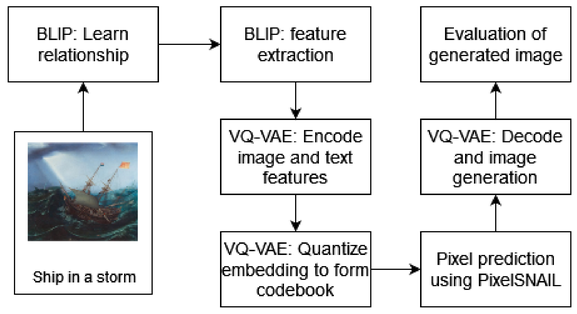

Figure 1. Overview of proposed model

Performance metrics of Image generation

The system assesses generated images using NLP scores for relevance and LPIPS for patch similarity.

Metric	Value
BLEU Score [5]	0.47138
ROUGE-1 score [4]	0.4312

Table 1. Performance Metrics for image generation

Model	LPIPS score
CycleGAN ([6])	0.134
DSVIB ([2]	0.129
Proposed system	0.124

Table 2. LPIPS score comparison

Functional Modules and Dataset Description

- Data Pre-processing
- Fine tuning BLIP for feature extraction
- VQ-VAE
- Code extraction
- Pixel predication using PixelSNAIL
- Image generation

- The MediaEval MUSTI 2023 data set [7] is used for this proposed system. The dataset consists of artistic paintings depicting various scenes and still lifes.
- The training dataset consists of 795 images, each with an id, an elaborate associated text and title.
- The testing dataset consists of 200 similar data. The titles associated with the images are used as the captions for the system.

Proposed Model for image generation

1. Fine tuning BLIP for feature extraction

- The BLIP model has been used to understand the relationship between multiple modalities
- The BLIP has 2 encoder models, the text and the vision model.
- Both these models are pre-trained individually and then integrated into a joint text-vision model which helps in cross-modal understanding
- Text and image features are extracted from the embeddings generated by the model.

2. VQ-VAE

- VQ-VAE is a probabilistic model that finds latent, low-dimensional representations of data
- The encoders in the VQ-VAE encode images and text into a lower dimensional latent space.
- The quantizer then takes this latent space as input and quantizes them into joint discrete embeddings.
- During the time of generation the decoder module reconstructs the quantized features back into the original features.

3. Code extraction

- In this step, the trained VQ-VAE is used to extract latent codes from the images and store them in a Lightning Memory-Mapped Database (LMDB)
- LMDB is used as it is highly efficient for data fetching from a large dataset and does not create any garbage data

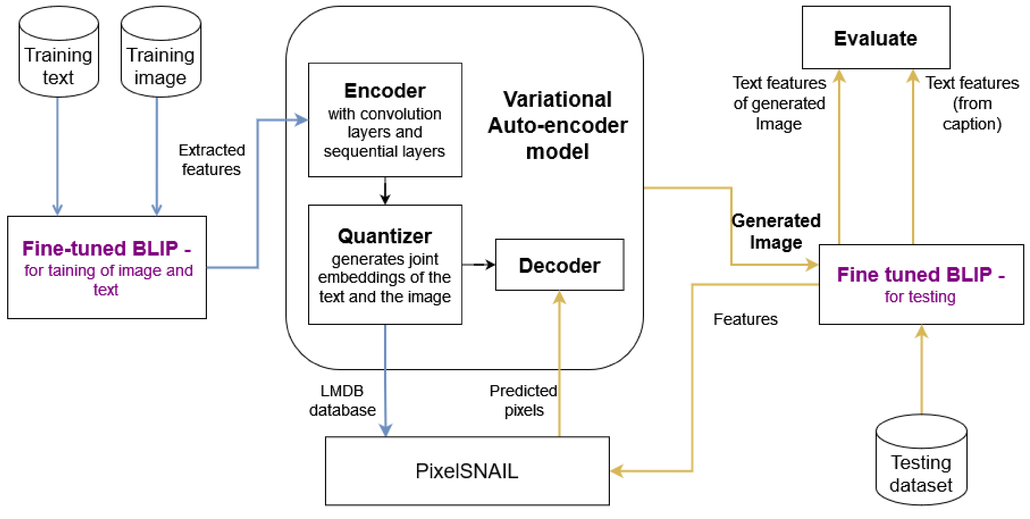

Figure 2. Model pipeline

4. Pixel prediction using PixelSNAIL

- PixelSNAIL is an autoregressive generative model that is used in density estimation tasks with high dimensional data.
- The latent space that is stored in LMDB is used to train the model which helps in predicting the pixels while capturing long-range dependencies and incorporating conditional information
- It works by leveraging causal convolutions, gated residual blocks, attention mechanisms, and conditional processing.

5. Image generation

- It is ensured that the predicted pixel values are within the valid value (-1 to 1).
- After these samples are generated the VQ-VAE decoder is used to up-sample and capture details in the pixels generated by the PixelSNAIL to generate image.

Text Prompt	Generated Image	Caption generated for evaluation
Ship in a storm		Ship in ocean

Figure 3. Sample of text prompt, the generated image and caption generated from it

Inference

- Accuracy, precision, recall and F1 score for the BLIP model indicates that the model is highly efficient in discerning the existence of a relationship between the image and text.
- The NLP techiques like BLEU and ROUGE-1 scores obtained show that the system achieves a high level of semantic alignment and visual realism.
- LPIPS score also show that the system is a significant improvement over models like CycleGAN [6] and DSVIB[2]

References

- Junnan Li et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900
- Shuyang Gu et al. "Vector quantized diffusion model for text-to-image synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10696–10706
- Kishore Papineni et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318