International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Text to image generation using BERT and GAN

Kavitha Soppari, Bhanu Vangapally *, Syed Sameer Sohail and Harish Dubba

Department of CSE (Artificial Intelligence and Machine Learning), ACE Engineering College, India.

International Journal of Science and Research Archive, 2025, 14(01), 720-725

Publication history: Received on 04 December 2024; revised on 13 January 2025; accepted on 15 January 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.1.0137

Abstract

Generating text to images is a difficult task that combines natural language processing and computer vision. Currently available generative adversarial network (GAN)-based models usually employ text encoders that have already been trained on image-text pairs. Nevertheless, these encoders frequently fall short in capturing the semantic complexity of unread text during pre-training, which makes it challenging to produce images that accurately correspond with the written descriptions supplied.Using BERT, a very successful pre-trained language model in natural language processing, we present a novel text-to-image generating model in order to address this problem. BERT's aptitude for picture generating tasks is improved by allowing it to encode rich textual information through fine-tuning on a large text corpus. Results from experiments on a CUB_200_2011 dataset show that our approach performs better than baseline models in both qualitative and quantitative measures.

Keywords: Text to image generation; Multimodal data; BERT; GAN; High quality

1. Introduction

Since the actual world is intrinsically multimodal, research in multimodal deep learning is essential to the advancement of artificial intelligence, even if many deep learning techniques were created for single-modality problems. One important example of multimodal learning is text-to-picture generation, where combining text and image modalities presents special difficulties. Frequently, images come with tags or descriptions that explain their significance and offer context. However, because the input (text) and output (picture) have essentially distinct properties, learning to create images from text is intrinsically difficult. For text-to-image

generation to be effective, three key issues must be resolved. Learning text representations that highlight visually meaningful characteristics is crucial first. Second, these language features need to be used to create realistic-looking, high-quality graphics.Third, for previously encountered text during training, significant feature extraction is required. In order to address these issues, text-to-image generation models usually consist of a GAN for picture production based on the embedded characteristics and a text encoder for embedding. Learning mappings that allow for the creation of various images that are in line with the semantic information in the text is one of the main objectives. In this research, we propose a text-to-image generation model that combines high-quality image generation using StackGAN with BERT-based embeddings. Because they rely on pre-trained text encoders made for zero-shot visual identification tasks, existing models frequently include gaps in the text manifold. By optimizing the pre-trained BERT for the text-to-image creation task, our method overcomes this constraint.

2. Literature analysis

Sutskever and colleagues (2014) The sequence -to-sequence (Seq2Seq) model was first presented in this seminal paper, which efficiently maps input sequences to output sequences using Long Short-Term Memory (LSTM) networks. By

[*] Corresponding author: Bhanu Vangapally.

tackling machine translation issues, it produced cutting-edge outcomes for English-to-French translations in the WMT'14 dataset. The model outperformed phrase-based systems with a BLEU score of 34.8. Reversing input sequences to enhance learning and using beam search for decoding were two significant developments. The foundation for managing sequential data in a variety of fields was established by this concept.

In 2016, Reed et al. In this groundbreaking study, Reed et al. provide a brand-new technique for employing Generative Adversarial Networks (GANs) to create images from textual descriptions. They take textual descriptions and convert them into visual representations using deep convolutional networks. The deployment of a deep convolutional GAN (DC-GAN), in which the discriminator and generator are both conditioned on text encodings obtained from recurrent neural networks (RNNs), is demonstrated in the paper.The paper's main contribution is the incorporation of text embeddings, which capture the description's semantic information, into a GAN framework. This results in the creation of realistic visuals that correspond to the text that is supplied. For instance, they demonstrate the potential of text-to-image synthesis by demonstrating how their approach can produce realistic images of flowers and birds from in-depth verbal descriptions.

Mansimov and colleagues, 2016: Mansimov et al. present a novel method in this research that uses attention mechanisms to generate visuals from text descriptions.The method's main concept is the employment of an iterative drawing process, in which the model simultaneously attends to pertinent words in the input caption and creates images by painting patches on a canvas. This attention mechanism aids the model in concentrating on particular caption elements that are essential for producing various image elements. A deep recurrent attention-based generative model is used by the alignDRAW model.In order to iteratively improve the image, a recurrent neural network (RNN) is used to construct patches and focus on certain caption sections at each stage. The Microsoft COCO dataset, which has many photos with informative captions, is used to train the system. A high degree of generalization is demonstrated by the model's ability to produce visuals that complement textual descriptions and outcomes that can adjust to previously unknown captions.

Odena and others, 2017) In order to produce high-resolution, lifelike photographs, this research presents Auxiliary Classifier GANs (AC-GANs), a variant of Generative Adversarial Networks (GANs) that include label conditioning. By adding an auxiliary classifier, AC-GANs enhance image synthesis by assisting the generator in creating diverse and class-conditional images. Compared to models that only use scaled low-quality photos, this model produces images with global coherence and 128x128 resolution.The addition of two new measures of image quality—discriminability and diversity—is a significant contribution. Compared to merely expanding lower-quality photos, the study demonstrates that higher resolution images are more discriminable and offer greater class information. With outcomes similar to actual ImageNet data, AC-GANs also demonstrate a significant degree of variation across 1000 ImageNet classes.

Ramzan, Sadia.This study investigates the creation of visuals from textual descriptions using deep learning methods, specifically GANs. An overview of the various architectures used in text-to-image synthesis, including StackGAN, AttnGAN, and other deep learning techniques, is given by the authors.Their research focuses on using RNNs and deep convolutional networks to encode text and provide visual outputs that correspond to it. The study contrasts different model performances and emphasizes the difficulties in preserving image quality and semantic coherence.

Table 1 Comparative Study of Existing Methodologies

Sequence to sequence learning with Neural Networks	Encoder-Decoder architecture	RNN	Moderate	-	-	ReLU, Tanh	Limited	Medium	Limited
Generative Adversarial Text-to-Image Synthesis	GAN	GAN	Moderate	-	-	ReLU LeakyReLU	Limited	High	Limited

Generating Images from Captions with Attention	CNN-RNN With Attention	CNN-RNN	Moderate	-	-	ReLU, Tanh	Limited	Low	Limited
Conditional Image Synthesis with Auxiliary Classifier GANs	GAN	GAN	Moderate	-	-	ReLU LeakyReLU	Limited	Limited	Limited
Text-to-Image Generation Using Deep Learning	CNN-RN N	CNN-RN N	Moderat e	-	-	ReLU, Tanh	Limited	Moderate	Limited
Text to Image Generation with Semantic-Spatial Aware GAN	GAN	GAN	High (85-90%)	Moderate	Moderate	ReLU, LeakyReLU	Limited	Moderate	Complex training requirements
StackGAN	Progressive GAN, two-stage model	GAN	Moderate (70-75%	Moderate	Moderate	ReLU, LeakyReLU	No	High	Struggle s with long text inputs
Ours: BERT+StackGAN	Bert,Stack Gan	GAN	Improved (85-90%)	Moderate	Moderate	ReLU, LeakyReLU	Yes	High	computationally significant hardware resources required.

3. Architecture

Mainly here the Architecture is divided into two Parts.The part is consisting the BERT it is an Natural language Processing (NLP) which will take the input from the user and converts the given text description into Numerical vectors (embeddings).These sentence embeddings are then used as input for the subsequent image generation stages, providing crucial information about the text to guide the image generation process.Second Part consisting the Generative Adversarial Network(GAN) there are so many types of Gan's are availabe.the Gan is consisting a Generator and a discriminator.the role of the generator is to generate the image from the given text and then the role of the discriminator is to validate the generated image is perfectly aligned with the given text or not .Here in this System we are using mainly the Stacked Generative Adversarial Network it is a deep- learning algorithm Architecture which is used to convert the given numerical vectors(generated by the BERT)into image.It addresses the challenge of generating high-resolution images from textual descriptions by employing a two-stage approach.Stage-1 generates the image from the given feed by the BERT Transformer and it generates the Low-Quality Images. Stage-2 generates High-Quality images which will take the input as the image generated by the Stage-1 generator.the role of the Stage-1 and Stage-2 Discriminator is to validate the images generated by the generators if they are aligning with the given text discription or not

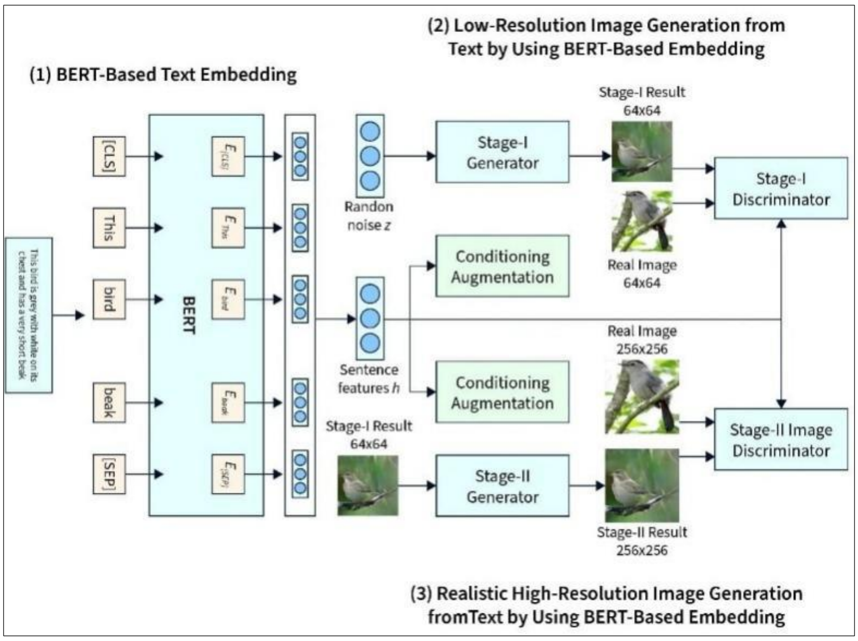
(2) Low-Resolution Image Generation from Text by Using BERT-Based Embedding
(1) BERT-Based Text Embedding
This bird is grey with white on its chest and has a very short beak
[CLS]
This
bird
beak
[SEP]
BERT
$E_{[CLS]}$
E_{This}
E_{bird}
E_{beak}
$E_{[SEP]}$
Randon noise z
Sentence features h
Stage-I Generator
Conditioning Augmentation
Conditioning Augmentation
Stage-II Generator
Stage-I Result 64x64
Real Image 64x64
Real Image 256x256
Stage-I Result 64x64
Stage-II Result 256x256
Stage-I Discriminator
Stage-II Image Discriminator
(3) Realistic High-Resolution Image Generation fromText by Using BERT-Based Embedding

Figure 1 Generating realistic images from text using BERT-based embeddings

4. Proposed Methodology

4.1. BERT+StackGAN

We have proposed the combination of the two models i.e the BERT is an NLP model and the StackGAN is an Deep-learning model offers several significant advantages for image generation system. BERT's exceptional ability to understand and represent the semantic meaning of text provides a strong foundation for image generation. By incorporating BERT-generated sentence embeddings, the StackGAN architecture can leverage deep semantic understanding of the text description.In StackGan basically the Architecture of this model is a Hierarchical manner so we can able to generate the image for the Larger text Description.

So, we combined this BERT and StackGan to generate the images aligned with the give Text description Perfectly and to generate the High-quality realistic images.

5. Dataset

CUB_200_2011 is the dataset that we are utilizing.There are 200 classes in the CUB dataset. Each of the 11,788 bird photos has ten textual descriptions. Approximately 80% of the CUB dataset has an object proportion that is smaller than half the size of the image. Using a bounding box for the item, preprocessing was done in the experiment described in this study so that the object's ratio was greater than 0.75.

6. Algorithm

6.1. Bert-Embedding Algorithm

- Step 1: Set up the pre-trained parameters BERT θBERT
- Step 2: Define the number of fine-tuning steps (Sf) and dataset descriptions T is equal to {t1, t2,..., tn}
- Step 3: For f = 1,..., Sf: perform
- Step 4: Execute the Adam update on θBERT using T.
- Step 5: Close for

6.2. StackGan-Algorithm

6.2.1. Stage 1

Let x_{row} be the low-resolution image, z be the random vector, D1 be the discriminator of stage 1, G1 be the generator, h be the text embedding of the provided text description via fine-tuned BERT, h_{ca} be the text embedding taken from the Gaussian conditioning variable, with λ serving as the regularization parameter. The stage 1 generator is taught to reduce the loss function of Equation (2), whereas the stage 1 discriminator is trained to maximize Equation (1). Equation (2)'s regularization term, the KL divergence, D_{KL}(N (μ_0(h), $\sum_0$(h))k (0, X_{row})), recovers latent text representation from an independent Gaussian distribution. Randomness is introduced by sampling the Gaussian conditioning variable h_{ca} from N(μ_0(h), $\sum_0$(h)).

$$LD_1 = \mathrm{E}(xrow, h) \sim Pdata\,[\log D1(xrow, h)] + \mathrm{E}z \sim pz, h \sim pdata\,\left[\log\left(1 - D1(G1(z, hca), h)\right)\right] \quad \ldots (1)$$

$$LG_1 = \mathrm{E}z \sim pz, h \sim pdata[\log\,(1 - D1(G1(z, hca), h) + \lambda DKL(N\,(\mu 0(h), \textstyle\sum 0\,(h))||N\,(0, Xrow) \ldots\ldots. \quad (2)$$

7. Quantitative Results

7.1. Stage 2

Assume G2 is the stage 2 generator and D2 is the stage 2 discriminator. Equation (3) is maximized by training the stage 2 discriminator, and Equation (4) is minimized by training the stage 2 generator. Only in stage 1 is random noise utilized, not at this point. Rather, stage 1 xrow uses a low-resolution image produced by the generator. By entering the low-resolution image produced by the stage1 generator and the text embedding vector taken from the text encoder, a high-resolution image, xhigh, is produced. By entering an image and text embedding vector, the discriminator ascertains whether the image and text match.

Table 2 IS and FID of existing and our Proposed model on the CUB dataset**.**

Dataset	**Model**	**Inception Score**	**FréchetInception Distance**
CUB	Sequence to Sequence Learning with Neural Networks	2.88	68.79
CUB	Generative Adversarial Text-to- Image Synthesis	3.62	67.22
CUB	Generating Images from Captions with Attention	3.7	51.89
CUB	Conditional Image Synthesis with Auxiliary Classifier GANs	4.36	NaN
CUB	stackGAN	4.16	38.73
CUB	StackGAN-BERT	4.44	37.79

$$LD_2 = \mathrm{E}(X_{row}, h) \sim P_{data}\,[\log D_2(x, h)] + \mathrm{E}x_{row} \sim P_{G1}\,, h \sim P_{data}\,[\log(1 - D_2\,(G2(x_{row}, h_{ca}), h))] \ldots.. (3)$$

$$LG_2 = \mathrm{E}x_{row} \sim p_{G1} h \sim p_{data}[\log(1 - D_2\,(G_2(x_{row}, h_{ca}), h))] + \lambda D_{kl}(N\,\left(\mu 0(h), \textstyle\sum(h)\right)N\,(0, X) \quad \ldots\ldots..(4)$$

8. Sample Generated Images

Text The bird is short and stubby with yellow on its body

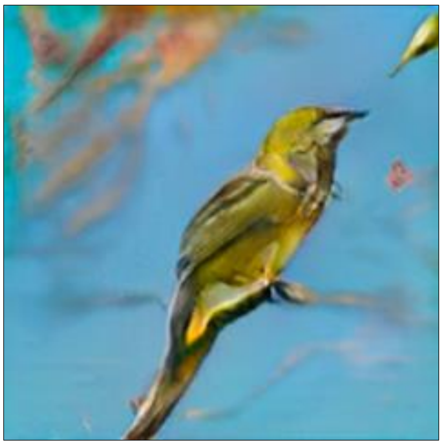

Figure 2 Image generated by the proposed model

9. Conclusion

StackGAN is a deep learning model that generates photo-realistic images from text descriptions using a two-stage process. First, it creates a low-resolution sketch based on the text. Then, it refines the sketch with intricate details and colors to produce a high-resolution image. By incorporating "BERT Text Embeddings," StackGAN ensures the generated images accurately reflect the input text, leading to significant improvements in image quality, realism, and diversity.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Sequence to Sequence Learning with Neural Networks"(Sutskever et al., 2014).

[2] Generative Adversarial Text-to- Image Synthesis"(Reed et al., 2016)

[3] Generating Images from Captions with Attention"(Mansimov et al., 2016)

[4] Conditional Image Synthesis with Auxiliary Classifier GANs"

[5] Text-to-Image Generation Using Deep Learning"(Sadia Ramzan 1,* , Muhammad Munwar Iqbal 1 and Tehmina Kalsum 2)