Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Text-to-Image Synthesis using Conditional Stacked Generative Adversarial Network with Skip-Thought Vectors

Susmita Sharma [a], Subarna Shakya [b], Suwan Babu Bastola [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email:
[a] 074mscsk016.susmita@pcampus.edu.np, [b] drss@ioe.edu.np, [c] 074mscsk017.suwan@pcampus.edu.np

Abstract
Generative Adversarial Network is an emerging technology for synthesizing realistic images. Text to image synthesis have been used for many research studies however, it is very difficult to reflect the meaning in the images from given text descriptions. Samples generated from text-to-image synthesis may lack details and also produce low quality images. In this research, the text-to-image synthesis is further divided into three stages namely Text embedding, low resolution image generation and high resolution image generation. The text embedding is carried out by skip-thought vectors including Recurrent Neural Network encoder with Gated Recurrent Unit activation to create 4*4*1024 dimension text embedding. Two conditional GANs are stacked over each other where the first GAN generate low resolution 64*64 dimension images. The second GAN utilizes the low resolution image to generate more detailed 256*256 high resolution images. The Conditional Stacked Generative Adversarial Network with Skip-Thought Vectors uses Caltech-UCSD Birds 200 (CUB-200) data-set produced by California Institute of Technology. The same network with different skip-thought encoders produces varied visuals i.e. the combine-skip encoder outperforms the bi-skip and uni-skip encoders. With an inception score of 5.19$\pm$0.04 and FID of 46.92, the ST-CSGAN generates images from text descriptions, which is greater than other models for the CUB dataset.

Keywords
Generative Adversarial Networks, Skip-Thought Vectors, Conditional Stacked Generative Adversarial Network, Text-to-Image Synthesis, Caltech-UCSD Birds 200, Inception Score, Fréchet Inception Distance.

1. Introduction

The Text-To-Image Synthesis being the versatile topic, a lot of projects has been done in this field these days. Text to image synthesis is the process of creating an image from a text description. Because the source and target domains are dissimilar, the concept is both fascinating and challenging. Deep convolutional and recurrent network architectures have become popular for learning discriminative text feature representations by automatically generating realistic images from a given text description.

A GAN is a deep neural network architecture made up of two different networks, a generator network and a discriminator network.Through multiple cycles of generation and discrimination, generator is trained explicitly. The Gaussian latent space delivers a better outcome in Generator network because it uses the existing noisy data to generate new data from a randomly generated vector of integers called a latent space. Whereas, the discriminator network attempts to distinguish between genuine and generator-generated data. During training, a little amount of labelled data is merged with a large amount of unlabelled data in a semi-supervised learning technique.

NLP is a major field in this program since it has a changeable domain space, with the source being text and the goal being an image. The Skip-Thought is an unsupervised learning technique that learns phrase level vectors rather than word vectors. It is a pure language-based neural network model for learning fixed length sentence representations. It is an encoder-decoder model, in which the encoder receives the training phrases and outputs a vector, which is then transferred to the GAN network for additional processing to generate the image. In Skip-thought, text embedding is achieved by translating similar syntax and semantics to the same vector

representations.

The text-to-image synthesis is divided into two sub-problems in this method. The first is learning text representations that encode the visual qualities provided in the text, and the second is learning a model that can generate images using the text representations learnt.

2. Related Works

The approach of generating realistic-looking images from a text description is known as text-to-image (TTI) synthesis. It's tough to design a model that can generate visuals that reflect the true meaning of the text, therefore synthesizing images from text descriptions is a difficult task. In the previous few years, a lot of effort has been done in this sector because it is a very useful notion in today's environment.

Ian Goodfellow et al. [1] in the paper describes two models which were simultaneously trained: a generative model G that captures the data distribution and a discriminative model D that evaluates the chance that a sample came from the training data rather than the generator G. This framework correspond the two-player mini-max game where the training procedure for G is to minimize the probability of D making the mistake. These networks simultaneously compete against each other and trained with the back propagation so as to make the model strong.

Ryan Kiros et al. [2] describes an approach for unsupervised learning of a generic, distributed sentence encoder.They used the continuity of text from the books to train an encoder-decoder model that tries to recover the surrounding phrases of an encoded portion. Scott Reed et al. [3], created a new deep architecture and GAN formulation technique to bridge the gap between text and image modeling, efficiently transforming visual notions from characters to pixels. From precise text descriptions, the built model could produce credible visuals of birds and flowers. Mehdi Mirza et al.[4] discover condition label which is given to both the generator and discriminator networks in this conditioned variant of the generative adversarial model. The model generates digits conditioned on class labels using the MINIST digits dataset. They showed how the model might be used to create a multi-modal model and gave some rough examples of an application to picture tagging, demonstrating the ability to generate descriptive tags that weren't part of the training labels.

Tobias Hinz et al.[5] describes that Generative adversarial networks may generate realistic-looking images when conditioned on simple textual image descriptions. Furthermore, quantitative examination of the created image is challenging because practically all assessment measures focus solely on image quality rather than assessing the image's relationship to the given text descriptions. The Semantic Object Accuracy also showed that, despite their improved performance, most models are still unable to generate photo realistic images for a variety of domains. Jifeng Wang et al.[6] shows shadow detection and shadow removal are two tasks involved in analyzing shadows from a single image. They presented a multi-task perspective that aids in the collaborative detection and removal of both detection and removal tasks by mutually benefiting and strengthening the network model of both tasks. It's built on a Stacked Conditional Generative Adversarial Network (ST-CGAN), which is made up of two CGANs stacked on top of each other, each with two generators and two discriminators.

Tao Xu et al. [7], developed fine-grained text-to-image generation which is built on attention-driven, multi-stage refining. The AttnGAN can synthesize fine-grained details in different subregions of an image by focusing on relevant words in NLP, i.e. the network focuses on particular labels or text parameters so that the model can create the picture component appropriately. They suggested utilizing the deep intentional multimodal similarity model to compute a fine-grained image-text matching loss when training the generator. The neural layer in the AttnGAN automatically selects the label or the condition at word level which help in generating the different parts of the images. AfrozAhamad,[8] Generating Text through Adversarial Training using Skip-Thought Vectors utilizes GAN with word embedding for text generation. With the use of a GANs network, this model was able to generate text from the Skip-Thought sentence encoder-decoder model. BLEU-n, METEOR, and ROUGE are some of the automated evaluation metrics that the model outperforms. The project makes use of GANs with various word embeddings, such as conditional text generation and automatic language generation, to reproduce writing style in the generated text by modeling the method of expression at a sentence level

across all works.

Trevor Tsue et al.[9] in the paper IMAGAN: Learning Images from Captions build a GAN architecture where Oxford-102 Flowers Dataset was used to build images from the captions. The captions were encoded using skip-thought vectors and images were created using a conditional deep convolutional GAN (DCGAN) with conditional loss sensitivity (CLS). A fully connected text model was used using bi-directional LSTM trainable from the loss and also implemented various training techniques. They also implemented the boundary equilibrium GAN (BEGAN) which balances the generator and discriminator loss to vastly improve the visual quality.

Wang et al. [10] proposes recent GAN research in the field of image processing, such as image synthesis, image generation, image semantic editing, image-to-image translation, image super-resolution, image inpainting, and cartoon production, is discussed in this paper. They looked at and summarized the approaches that were employed in these applications to improve the outcomes. The goal of this review was to provide insight into GAN research and to present several GAN-based applications in diverse circumstances.

T. Salimans et al."Improved techniques for training GANs,"[11] presented the variety of new architectural features and training procedures for GAN. The research claims that analyzing the generated image in terms of both qualitative and quantitative analysis is difficult. The GANs network is primarily concerned with the semi-supervised learning technique and picture production, rather than with image relatedness. As a result, the study discussed a number of evaluation strategies for keeping the network stable during training.

3. Research methodology

3.1 Generative Adversarial Network (GAN)

Generative Adversarial Network is a class of neural network mostly used for unsupervised learning. The main idea is to learn the two neural network simultaneously. Initially,a random input vector z is fed to the generator G and maps to output image y then:

$$G : z- > y \tag{1}$$

Generator network G is trained to produce input-like results that cannot be classified by Discriminator network (D) whether it is an original or generated image. The discriminator network D is adversarial trained with G. In this way, G and D compete each other and adjust their parameters to become better at their task.

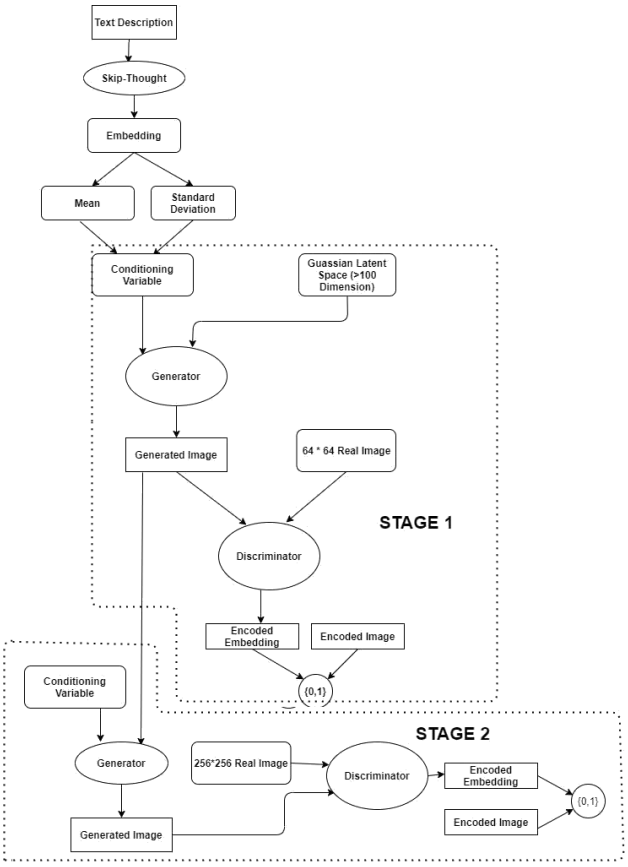

Figure 1: The System Flow Diagram for the Conditional Stacked Generative Adversarial Network with Skip-Thought Vectors (ST-CSGAN)

The objective function of standard GAN is given by:

$$F(G,D) = E_y[log(D(y))] + E_{x,y}[log(1-d(G(x,z)))] \tag{2}$$

Where, Generator G learns to minimize the cost function F(G, D) against an adversarial Discriminator D, which learns to maximize F(G, D).So, the objective of the network is summarized as:

$$G^* = min_G max_D F(G,D) \tag{3}$$

3.2 Conditional Generative Adversarial Network (CGAN)

Conditional GAN is an extension of GAN where both generator network and the discriminator network receive additional Conditioning variables c, yielding

G (z, c) and D (x, c). This formulation allows G to generate images Conditioned on variables c.

3.3 Skip-Thought Vectors

Skip-Thought, also known as Skip-Thought Vectors, is a neural network unsupervised learning technique for fixed length representations of any sentences. Skip-Thought is mostly used to generate previous and subsequent sentences from the present one. For this, the model uses one Encoder network and two Decoder network (previous Decoder network and next Decoder network). However, our model is text-to-image generation so we will only use the Encoder network for embedding the text vector. The RNN encoder with GRU activations where similar vector representations are assigned to phrases that share semantic and syntactic features. The image generation task continues after passing the pre trained 4800-dimensional 'skipthoughts' sentence embedding along with the random Gaussian noise vector of 100 dimensional to the generator network. The encoder uses 2400 GRU units with a word vector of dimensions 620 and combines the sentence embedding to produce 4800-dimensional skip vectors, with the uni-skip model's first 2400 dimensions and the bi-skip model's last 2400 dimensions. The encoder accepts the training phrase and displays a vector. Two decoders take the vector to be entered by both. The encoder and decoder are built from recurrent neural networks (RNN). Many encoder types, including uni-skip, bi-skip and combine-skip, are tried. In the forward direction, Uni-skip reads the sentence. Bi-skip reads the front and back sentence and brings the results together. Combined skips combine the uni- and bi-skip vectors. The input sentences are only minimally tokenized.

3.4 Conditional Stacked Generative Adversarial Network With Skip-Thought Vectors (ST-CSGAN)

A ST-CSGAN is a network that consists of two conditional GANs stacked together with the text embedding vector model to generate high-resolution images as shown in Figure 2. There are two stages to it: Stage I and Stage II. The Stage-I network produces low-resolution images with basic colors and rough sketches that are conditioned on a text embedding, whereas the Stage-II network transforms the image produced by the Stage-I network into a high-resolution image that is likewise conditioned on a word embedding. To produce a more realistic high-resolution image, the second network corrects defects and adds attractive elements.

The ST-CSGAN uses a sketch-refinement method to split the challenging problem into more manageable sub-problems because the images are generated in two steps. Based on the given text description, the Stage-I GAN generates low-resolution images by producing the object's low-detailed primitive shape and colors. Stage-II GAN uses Stage-I findings and text descriptions as inputs to create high-resolution images with photo-realistic details. It can repair flaws in Stage-I results and add compelling details using the refinement process.

4. The Model Architecture

The network consists of one text encoder, two generator network and two discriminator network as shown in Figure 2.

4.1 Text Embedding

First of all, the given text description is passed through the Skip-Thought Encoder, where the given text description is encoded into a 4800*1*1 Text Vector i.e. Text embedding (φ). This text embedding are connected to a fully connected layer to generate mean $\mu(\varphi)$ and standard deviation $\sigma(\varphi)$. Mean and standard deviation creates the diagonal covariance matrix $\Sigma(\varphi)$ and then the Gaussian distribution using:

$$N(\mu(\varphi), \Sigma(\varphi)) \tag{4}$$

Conditioning variable is calculated using:

$$c = \mu + \sigma N(0, I) \tag{5}$$

The conditioning variable is fed to the fully connected dense Leaky Rectified Linear Unit (L-ReLU) where the dimension is changed to 4*4*1024. The Noise variable of Gaussian Latent Space z(0,1) of 100 dimensions is concatenated with the text embedding and passed to the Generator network. The Gaussian Latent Space is a hyper sphere shape with the mean having zero and standard deviation of one which defines the shape and distribution of the input to the generator network for generating the images. The Gaussian Latent Space is recommended over the Uniform Latent Space which is of hyper cube shape

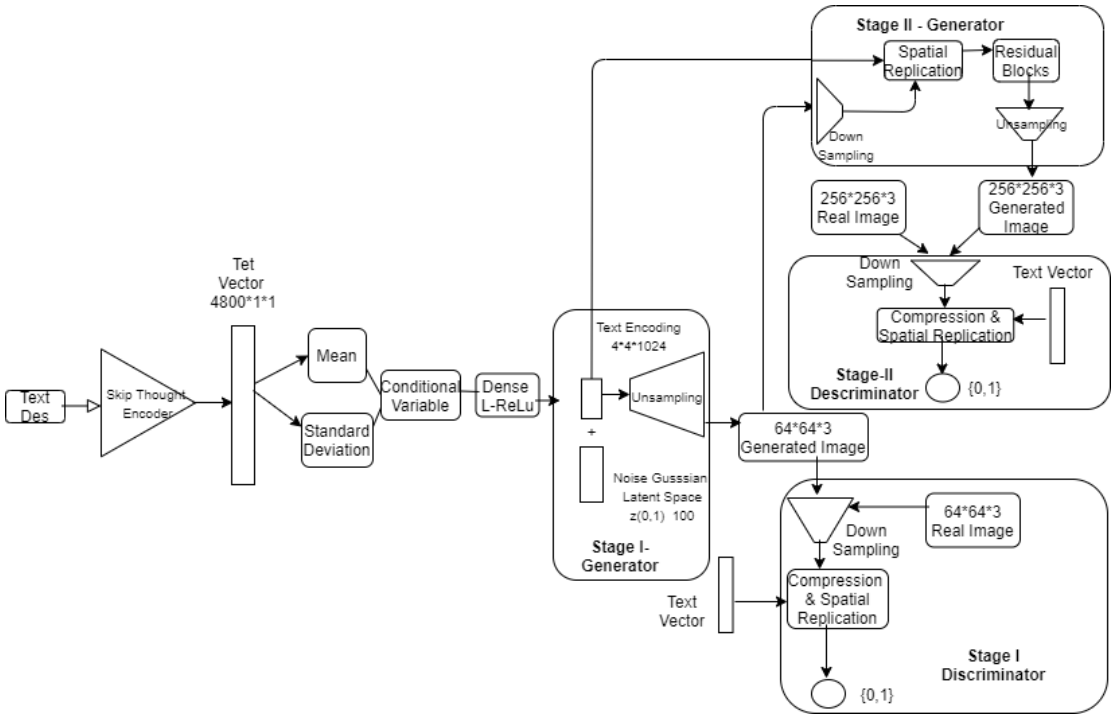

Figure 2: The Architecture for the Conditional Stacked Generative Adversarial Network with Skip-Thought Vectors (ST-CSGAN)

as the learning will be easier and effective for the network.

4.2 Low Resolution Image Generation

The Generator network is made up of several upsampling blocks for the sake of converting low dimensional text embedding to the higher dimensional image. The upsampling blocks generate the 64*64*3 dimensional image which is than fed to the discriminator network. The Discriminator network consists of several downsampling blocks where the generated image of 64*64 is passed along with the 64*64 real image for the comparison between the real and fake. The text vector and the output from the downsampling blocks are than passed to Compression and Spatial Replication block which outputs either 0 or 1 (fake=0 , real =1). The process continues up to the 600 epochs and each epoch produces the real looking images than the previous one. Finally the image is generated by the generator cannot be differentiated by the discriminator and the network reaches the Nash Equilibrium state.

4.3 High Resolution Image Generation

In this stage, the 64*64 generated image is passed to the Generator network. The extra downsampling layer is added to the network to convert high dimensional image into low dimension. The text encoder is passed to the spatial replication for the purpose of generating different conditional variable than that of Stage-I network. The concatenation of text embedding and the output from downsampling block is passed to the residual blocks where fully connected leaky ReLU and batch normalization changes the dimension. It is now fetched to the upsampling blocks which convert the low dimensional text to high dimensional image and hence the 256*256*3 dimensional image is generated from the network.

The generated image from the Stage-II generator along with the 256*256 real image is passed to the downsampling block in the Discriminator. The text vector and the output from the downsampling blocks are than passed to Compression and Spatial Replication block which outputs either 0 or 1 (fake=0 , real =1). The process continues for another 600 epochs where the generated image will be refined in the each next epoch and finally the image will be generated which the discriminator cannot differentiate either it is real or fake. In this way, the generator and discriminator network train each other to reach the Nash Equilibrium state.

5. Experiments

5.1 The Implementation Details

The network is made up of dense layer networks that are followed by batch normalization with gamma initializers of 1 and beta initializers of 0. Convolution 2D of 3*3 kernel size with strides of 1 for generator and 4*4 kernel size with strides of 2, zero padding 2D of padding 1*1. The network is expanded by adding the lambda layer, reshape layer, and add layer. The network also includes activation layers such as Rectified Linear Unit(ReLU), Leaky Rectified Linear Unit(LeakyReLU), Tanh, and Sigmoid functions, as well as flatten, compile, and Adam optimizer layers with learning rates of 0.0002[12], beta1 of 0.5, and beta2 of 0.999. Finally, it has concatenation layers, add layer and reshape layers. The guassian latent space noise vector was utilized to improve learning. Apart from the real and fake images, it also offers small batches. Instead of 1 and 0, label smoothing was performed by setting the real value to 0.9 and the false value to 0.1 for each incoming sample.

5.2 The Dataset

The other popular dataset (such as PASCAL VOC, Caltech-101, etc.) focus on basic level categories only, However the CUB-200 dataset which includes 11,788 images of birds, belonging to 200, mostly North American, bird species enable the study of subordinate categorization. In this dataset, each image is annotated with a bounding box, a rough bird segmentation or outline, and a set of attribute labels. The dataset provides an opportunity in computer vision by helping to classify objects that are unknown to them. The dataset contains of total number of 11,788 images where 8,855 images are used for training and remaining 2,933 images are used for testing.

The CUB-200 dataset consists of different 25 attributes like crown colour, nape colour, bill shape, head pattern, belly pattern, belly colour, wing shape, shape, primary colour, size, forehead colour, throat colour, eye colour, underparts colour, breast colour, upperparts colour, back pattern, back colour, leg colour, tail pattern, under tail colour, upper tail colour, wing pattern and wing colour.

The CUB-200 dataset consists 322 binary attributes labels along with discrete attributes labels needed for predictions. The dataset has list of part names with the location values, list of classes, list of images, list of image class labels and bounding box label for each images. These attributes provides the detail information to the dataset and thus helps in generation of images from text.

5.3 Evaluation metrices for the Model

The generator uses the discriminator as the loss function i.e. the generator is implicitly dependent upon the discriminator's loss function and is learned during training via the discriminator model while the discriminator network is updated like other neural network i.e. trained directly on the generated and real image. The aim of the generator network is to minimize the log of the inverse probability predicted by the discriminator for the fake images thus encouraging the generator to generate the images that will have a low probability of being fake. Generator: minimize log (1 - D(G(z))) The discriminator network seeks to maximize the average of the log probability of real images and the log of the inverse probability of the fake images. Discriminator: maximize log D(x) + log (1 - D(G(z)))

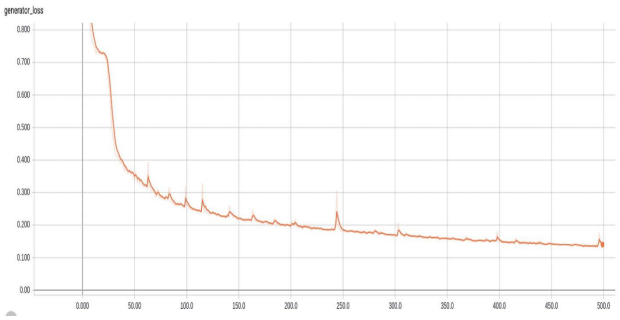

Figure 3: The Graph for the Generator Loss with respect to the Epochs

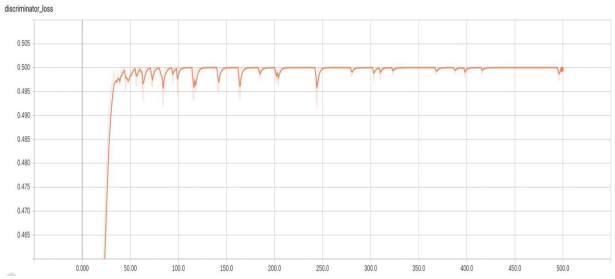

Figure 4: The Graph for the Discriminator Loss with respect to the Epochs

As shown in the Figure 3, the loss function for the generator is high at the beginning while it goes on decreasing as per the training phase continues. After

some epochs the generator network becomes good at generating the more real looking images and the generator loss decreases from 0.8 up to the 0.12 in the 500 epoch. Hence the main aim to minimize the loss is performed by our generator network model.

The Figure 4, describes the discriminator loss in respect to the epoch. After some epochs the discriminator network becomes good at classifying the real and fake images. The discriminator loss increases from 0.496 to 0.5 after being trained to 500 epochs. Thus the discriminator network in our model also performed its work to maximize the loss so that the network can be trained very well to generate the real looking images from the given text descriptions.

As the GAN network is the adversarial network i.e. the two networks (Generator network and Discriminator network) compete against each other to make themselves stronger and learn from competing against each other, they tries to become stronger by minimizing the opponent's performance.

6. Results

The images are generated in two stages from the given text description where the Stage-I generates 64*64 dimensional low resolution image that contains little information and Stage-II generates 256*256 dimensional high resolution image that contains detailed information. Also, the given text description can correctly fit the multiple number of images i.e. the space of plausible images given text description is multi modal. So, the text description is used to generate 8 different images in both the stages. The images generated using the uni-skip thought encoder vector, bi-skip thought encoder vector and combined-skip thought encoder vector are as shown below:

Figure 5: Images From Text Descriptions Using Combined-Skip Encoders.

Figure 6: Images From Text Descriptions Using Bi-Skip Encoders.

Figure 7: Images From Text Descriptions Using Uni-Skip Encoders.

6.1 Quantitative Results

In Generative Adversarial Networks, instead of being trained directly, the generative models are trained by the second model called the discriminator. Thus, there is no objective model for measuring the generator network. Because of these issues, it is difficult to figure out when the training should stop and hence the generator network are evaluated based on the quality of the generated images.

The quantitative analysis of the Generative Adversarial Network includes the calculation of specific numerical scores to measure the quality of the generated images. As the human judgement can vary from time to time, the fix judgement methodology is to measure the quality of the image.

Kullback-Leibler Divergence The Kullbac-Leibler (KL) divergence is a statistics metric for comparing and contrasting two probability distributions. We take the exponential of the KL divergence and then average it over all of our images to get the final score. The KL divergence between two probability distributions p(x) and q(x) is:

$$D_{KL}(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} dx \tag{6}$$

Inception Score The Inception Score is most commonly used to validate the GAN model's performance. It extracts the features of both generated and actual images using a pre-trained inception V3 network (trained on Imagenet).

The model is used to classify a huge number of generated photos. The chance of an image belonging

to each class is forecasted in detail. The inception score is created by combining these predictions. For N samples of images generated by the model, which is denoted as x^i, the marginal class distribution

$$p(y) = \int_x p(y|x)p_g(x) \tag{7}$$

The following steps are used to calculate the inception score:

The KL divergence and the predicted improvement calculated by:

$$IS(G) = exp(E_{x\sim p_g} D_{KL}(p(y|x)||p(y))) \tag{8}$$

Where, x represents a sample which is sampled from a distribution $p(y|x)$ is the conditional class distribution and $p(y)$ is the marginal class distribution.Finally, calculate the value of an exponential to get the inception score.

Table 1: Comparison of IS between different GAN Models

S.N.	GANs Models	IS
1	GAN-INT-CLS[13]	2.88±0.04
2	GAWWN[14]	3.62±0.07
3	AttnGAN[7]	4.29±0.05
4	DF-GAN[15]	5.10±0.06
5	ST-CSGAN (Uni-Skip)	4.12±0.21
6	ST-CSGAN (Bi-Skip)	4.58±0.10
7	**ST-CSGAN (Combined-Skip)**	**5.19±0.04**

Fréchet Inception Distance (FID) The Fréchet inception distance (FID) is a statistic for evaluating the quality of images produced by generative models such as generative adversarial networks (GAN). The FID compares the distribution of generated images with the distribution of real photos that were used to train the generator, unlike the earlier inception score (IS), which just analyzes the distribution of generated images. FID is more noise-resistant than IS. The distance will be great if the model only generates one image per class. As a result, FID is a more accurate measure of image diversity. The FID has a strong bias but a low variance. We should expect the FID between a training dataset and a testing dataset to be zero because both are real photos. Running the test with multiple batches of training samples, however, reveals that none of them have a zero FID. For a "multivariate" normal distribution, the Frechet Inception Distance is equal to:

$$FID = ||\mu_X - \mu_Y||^2 - T_r(\sum X + \sum Y - 2\sum X \sum Y) \tag{9}$$

where, X and Y are real and fake embeddings assumed to be two multivariate normal distributions. μ_X and μ_Y are the magnitudes of vectors X and Y. T_r is the trace of the matrix and $\sum$ X and $\sum$ Y are the covariance matrix for the vectors.

From the given analysis the Inception score for the images was obtained to be 5.19±0.04 and FID score is 46.92 for STCSGAN with combined skip-thought encoder network. When compared to other models, our model's Inception Score looks to be higher and FID is lower as indicated in Table 1 and Table 2.

Table 2: Comparison of FID between different GAN Models

S.N.	GANs Models	FID
1	StarGAN[16]	84.58
2	GAN-INT-CLS[13]	68.79
3	GAWWN[14]	67.22
4	ST-CSGAN (Uni-Skip)	70.13
5	ST-CSGAN (Bi-Skip)	57.82
6	**ST-CSGAN (Combined-Skip)**	**46.92**

Conclusions

To develop text-to-image synthesis, the Stacked Generative Adversarial Network with Skip-Thought Vectors uses several dense layers, as well as convolutional, upsampling, zero padding, batch normalization, activation function, and flatten layers. The Skip-Thought Vector generates text embedding vectors for text-to-image mapping using an encoder model. The model generates 2400 dimensional vectors using uni-skip and bi-skip encoders, then concatenates them to yield 4800 dimensional vectors. The low-resolution images generated in 64*64 dimensional images contain a basic sketch of the image, which is then used in the stage II network to generate a final image with a resolution of 256*256 that describes the detailed information. The conditional GAN simplifies the network by supplying conditional labels that favor the provision of meaningful information for image synthesis. On the CUB-200 Dataset, the Inception score and FID Score are used to assess the quantitative analysis of generated images. When compared to other models, the image's Inception score is 5.19±0.04 and FID score is 46.92 which indicates that the images formed have high similarity and minimal variances when compared to the original dataset.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014.

[2] Ryan Kiros, Yukun Zhu, R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015.

[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

[4] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[5] T. Hinz, S. Heinrich, and S. Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, sep 5555.

[6] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018.

[7] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[8] Afroz Ahamad. Generating text through adversarial training using skip-thought vectors. *ArXiv*, abs/1808.08703, 2019.

[9] Trevor Tsue Singhal, Karan. Imagan: Learning images from captions. cs230., 2016.

[10] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8:63514–63537, 2020.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

[12] Bhaskar Ghosh, Indira Dutta, Albert Carlson, Michael Totaro, and Magdy Bayoumi. An empirical analysis of generative adversarial network training times with varying batch sizes. 10 2020.

[13] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[14] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. *Advances in neural information processing systems*, 29:217–225, 2016.

[15] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[16] Matthew Amodio, Rim Assouel, Victor Schmidt, Tristan Sylvain, Smita Krishnaswamy, and Yoshua Bengio. Image-to-image mapping with many domains by sparse attribute transfer, 06 2020.