Text to Image GANs with RoBERTa and Fine-grained Attention Networks

Siddharth M, R Aarthi
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India

***Abstract*—Synthesizing new images from textual descriptions requires understanding the context of the text. It is a very challenging problem in Natural Language Processing and Computer vision. Existing systems use Generative Adversarial Network (GAN) to generate images using a simple text encoder from their captions. This paper consist synthesizing images from textual descriptions using Caltech-UCSD birds datasets by baselining the generative model using Attentional Generative Adversarial Networks (AttnGAN) and using RoBERTa pre-trained neural language model for word embeddings. The results obtained are compared with the baseline AttnGAN model and conduct various analyses on incorporating RoBERTa text encoder concerning simple encoder in the existing system. Various performance improvements were noted compared to baseline Attention Generative networks. The FID score has decreased from 23.98 in AttnGAN to 20.77 with incorporation of RoBERTa model with AttnGAN.**

***Keywords*—Natural language processing; computer vision; GANs; AttnGAN; RoBERTa**

I. Introduction

Text to Image generation is an application of Generative Networks. The underlying problem comprises recognising the context of the text description and generating a realistic image that matches the caption. It is a multi-modal problem that challenges natural language Processing for context understanding and Computer Vision for Images. There are numerous applications in arts and design and have advanced considerably in recent years. Text to image generation can assist game developers to generate more distinct characters or skins with ease. Artists could use them to create starter comics from descriptions of the scene.

GANs are primarily used as generative networks for Image Synthesis from text and use Deep Convolutional GANs [1]. Recently, enormous progress has been made in synthesising images from texts for a single class of datasets like the Caltech-UCSD Birds-200-2011 dataset [2] or Oxford 102 Flower dataset. This paper uses the baseline AttnGAN model [3] that make use of Long Short term Memory for processing the text description. The latest language models developed lately has proved to be very efficient for text related problems. Thus it is required to utilize the latest transformer models so that better attention can be obtained within the natural language and hence increase the overall performance of the problem.

This problem can be divided into different parts and approached individually as a module. Text description is taken with details as input. The description includes features of its appearance, like the colour of particular body parts and its length. Text like, "this small bird has a short, pointy orange beak and white belly", are provided as input. The RoBERTa language neural model [4] to capture attention and understand the context of the description.

Fig. 1. Result of a Generated Bird from the Generative Model using RoBERTa Text Encoder and the Attention Captured during the Generation of Image at Epoch 600.

RoBERTa language model is used for embeddings words into a feature representation. Transform models use an attention mechanism to capture critical details associated with the word, and they can link them using the attention heads [5]. AttnGAN are used to train the generative networks. With each stage, higher resolution images are synthesized. Another component used in the AttnGAN is a Deep Attentional Multimodal Similarity Model (DAMSM). The attention mechanism and the DAMSM are used to find the similarity between the image generated by the GANs and the sentence using both the global sentence level and the fine-grained word-level information. The DAMSM component provides a fine-grained image and text comparison loss that can be used to train the generator [3].

The goal is to explore the current state-of-the-art model that can generate images based on the description from text using the Attention mechanism. RoBERTa language model is incorporated and experiments are performed on how the

existing system is affected. The same is analyzed and compare each result using the Fréchet inception distance (FID) score obtained from the base model. The focus is on the CUB Bird dataset for this paper, and the dataset also provides us with boundary box segmentation of bird images. Segmentation [6,7,8] can help train the model for generating a specific object in the boundary of images used from training . Analysis also performed on how RoBERTa embeddings have an effect in an interval of epochs on generating the images. The models are implemented using deep convolutional neural networks because they enhance the processed image [9,10]. The final output of the generative network is a high-resolution image matching the text description. Various attentions mapped are recorded with the experiments conducted and we obtain results as in Fig. 1.

The major benefits on exploring research on this problem lay in understanding the use of transformers with basic attention based generative system. This will help us explore how latest language models that uses attention heads can help understand the text association with the image better. A larger intuition could be developed in natural language association with text generation and scene prediction. This will help artists, game developers, animation industries to develop characters based on the textual description provided by the artists.

II. Related Work

Generative Adversarial networks are generative models that can synthesize images and are used for generative learning. In this network, a generator generates images based on the input from the embeddings and the noise. The discriminator help discriminate the picture as real or fake. Both generator and discriminator improve over time. The generator aims to generate images to fool the discriminator into thinking and classifying the images as real. These models were the first approach in generative networks [11] by Ian Goodfellow. There have been a variety of works on generative networks, and GANs have been able to generate photorealistic images with very high resolutions lately. Lately, generating images from text descriptions has been an area of research, and have few novel approaches to this problem.

The first approach to this problem was synthesizing low-resolution images from captions using Deep Convolutional GANs [1]. This system however could not completely produce image that looked realistic enough. Many images that were synthesised didn't exactly match the description either. This lead the author [12] to introduce Generative Adversarial What-Where Network (GAWWN). It exposed the control with the object's bounding box in the image and focused on particular parts. It modelled the distribution on various components like the tail and beak to obtain efficient results by focusing on that area. This proved to help identify key objects for generation but couldn't exactly focus on key details of objects like its poses or structure. A conditional Pixel Convolutional Neural Network (CNN) was used to synthesize images from text and used a multi-scale model structure. An image closer to the text description could be generated and a starting point to research the text to image using GANs. The image quality generated was an issue motivating Stack GANs, which used a stacked approach to improve the image resolution in different stages and generated 256 x 256 sized photorealistic images. The initial model was able to generate 64 x 64 resolution images. This approach generated an initial 64 x 64 images and was trained with GANs in two stages to get 128 x 128 images in Stage-I and 256 x 256 images in Stage-II. Each stage had an aim to improve the image quality to gain a photorealistic effect. StackGANs could generate photorealistic birds and flower images [13].

While this generated photorealistic images, proper context extraction was lacking. It paved the way to AttnGAN, which took a new approach by using an attention mechanism from the text description. Attention capturing helped the network to a closer understanding of context and better generation of images from text. AttnGAN used word embeddings and could capture important words from descriptions of birds [2] in the Caltech-UCSD Birds-200-2011 dataset. The attentional generative network could synthesize fine-grained details at various image subregions by providing detailed attention to the relevant words provided in the text description. This paper also introduced a deep attentional multimodal similarity model (DAMSM), used as the loss function and matched with text description and the generated image features. Word level condition selection was introduced to synthesize image details [3]. AttnGAN made use of bidirectional LSTM for the natural language processing. Similar to this work is the Controllable Text-to-Image Generation, which is used to synthesize high-quality images effectively by controlling parts of the image generation concerning natural language descriptions. In addition to the attention mechanism followed in AttnGAN, this paper used a channel-wise attention module and a word-level discriminator. It adopted a perceptual loss [14] in the text-to-image synthesis. Experimental researches have been performed by updating the architecture within the GANs by connecting generated image with the input description. The method of redescription was performed in MirrorGANs and Cycle GANs [15,16] using the BERT language model, where the authors obtained a great performance enhancement on complex datasets. Lately transformers have enhanced the neural language processing and is widely used in most of the latest intelligent systems. While the works performed till now has shown great result, it is very important to understand how these models could perform with the latest transformer models. This lead us to using latest neural language model RoBERTa as a pretrained model and incorporate it with the AttnGAN network instead of using basic LSTM system at language processing end. The aim of this paper is to analyze how well the AttnGAN model improves its performance using this system.

III. Dataset

Table I shows the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [2] image dataset that contains 200 categories of birds were used for experiments. There are a total of 11,788 images with annotations. The images and annotations together size up to 1.1 gigabytes. This dataset is used as a benchmark dataset for all the text to image synthesis research works. The images along with them have boundary boxes provided and are of various sizes each.

This dataset [2] contains images of North American birds from 200 different species of ranges. The dataset (CUB-200) was created in 2010 and contains approximately 6000 photos of each of the 200 bird types. Additional label data, such as

TABLE I. STATISTICS OF DATASETS

Dataset	Train	Test
No: of Samples	8,855	2,933
No: of Captions	10	10

bounding boxes, crude segmentations, and additional features, accompanied this. The dataset was updated in 2011 (CUB-200–2011) to include new photos, bringing the total number of images in the dataset to around 12,000 (CUB-200–2011). 15 component locations, 312 binary attributes, and a bounding box per image were added to the accessible attributes . The photos and class labels will be used to create and train networks for predicting bird class for the majority of this series.

IV. METHODS

A. Attention GANs

Fig. 2 shows the architecture of the AttnGAN networks with RoBERTa neural language model. AttnGAN make use of an attention mechanism that embeds the generated caption from the Birds dataset and run through the RoBERTa model to generate word and sentence vectors. The text encoder takes the caption, which is a T words sentence. The sentence features contribute to the global vector, which is passed on to the noise vector. The sentence feature is the final hidden state with a dimension D.

$$\bar{e} \in R^D \tag{1}$$

Similarly, word features are extracted separately. It produces a hidden state from all timesteps for the T-word sentence.

$$e \in R^{DXT} \tag{2}$$

The conditioning augmentation has the randomly sample latent variables from the Gaussian distribution. So $\bar{e}$, which is received as an input to caption feature, is split into μ and σ with a fully connected linear layer. This is the mean and variance from the sentence embedding. The mean and variance generated are used to parameterize the normal distribution from which a sentence embedding sample gets generated to be passed on to the generative network.

It is combined with a noise vector so that the generated images show higher variation for a single caption. The c vector is concatenated with the Z noise vector, and this is used in further stages for the generation of various features of birds in the network. Similarly, word features are extracted separately. It produces a hidden state from all timesteps for the T-word sentence.

$$\begin{aligned} &\bar{e} \longrightarrow \mu, \sigma \\ &c = \mu + \sigma * \varepsilon, \varepsilon \sim N(0, I) \end{aligned} \tag{3}$$

The first generative network is mainly responsible for upsampling. The nearest neighbour interpolation is used to upsample with a scaling factor of 2. The output is generated with a 64 X 64 image. This stage does not use any word-level features that are extracted using the RoBERTa model. It utilizes the sentence level features, which is taken as input from the noise vector space.

$$\begin{aligned} &h \in R^{\hat{D} \times N} \\ &h_0 = F_0\left(z, F^{ca}(\bar{e})\right) \end{aligned} \tag{4}$$

The first attention model combines word features e, with the previous stage context h_{i-1}. The word features before combination are brought into a common space. This is represented using e' and obtained by adding a new perceptron layer. $e' = Ue$, where $U \in R^{\hat{D} \times D}$. Each column of h is a feature vector of a sub-region of the image.

$$s'_{j,i} = h_j^T e'_i \tag{5}$$

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \text{ where } \beta_{j,i} = \frac{\exp\left(s'_{j,i}\right)}{\sum_{k=0}^{T-1} \exp\left(s'_{j,k}\right)} \tag{6}$$

Combining them with the context, it generates a score for a particular sub-region j, and a word i. So a combination is brought out with a particular word with a sub-region and it's used for the word-context vector for that region. This process is repeated for each region. This provides us with the output of the attention network.

$$F^{attn}(e, h) = (c_0, c_1, \ldots, c_{N-1}) \in R^{\hat{D} \times N} \tag{7}$$

The second generator also is used for upsampling of the image and it obtains an image of 128 X 128. Here, along with the previous output as input from the first generator which carries the context vector, the word embeddings through the attention networks are also added which carries the word context vectors. The residual blocks here, make the network deeper and train them without degradation. Similarly, one more generator was used to upscale the image up to 256 X 256 and it takes input similar to that of the second generator.

In the end, 256 X 256 image is passed to an image encoder. In the image encoder, local image features can be extracted and this is converted to a common space to match the text encoder feature. These two are combined to make the Deep Attentional Multimodal Similarity Model (DAMSM) and this is trained with attention loss. The DAMSM model is pre-trained for stability in the system.

There are three discriminators each attached with its respective generators. The sentence level features is taken without noise vector as input to each discriminator. Two forms used in the network is the unconditional form that tells if the image is real or fake and the conditional form that tells if the image and caption are of the same pair. In unconditional pair, a result close to 1 is obtained if both the pair are matching.

1) Text Encoder: RoBERTa makes use of transformers that has attention mechanism which learns the contextual relationship between the words within the sentences.

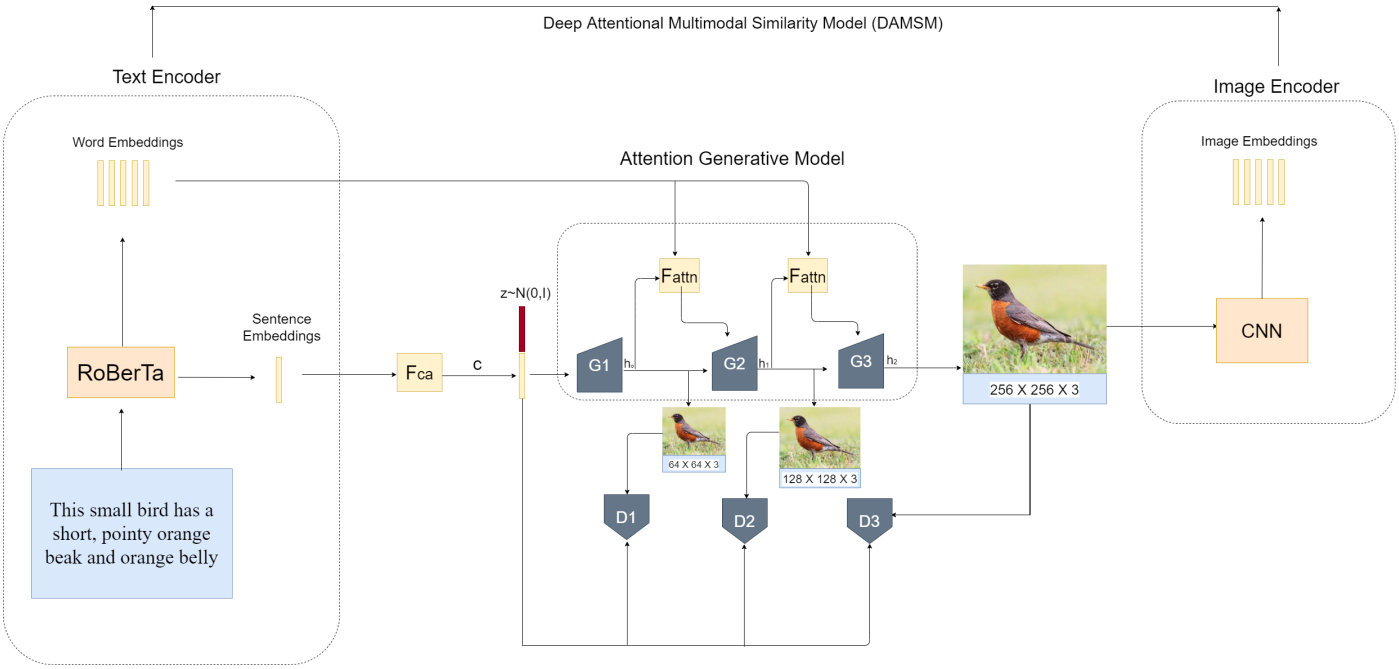


Fig. 2. Architecture of the Proposed System. RoBERTa Text Encoder is for Word Embeddings passed to Generative Network with Fine Grained Attention Networks; Text-Image Matching Loss Generated with DAMSM for the Generative Networks.

2) Image Encoder: The Image encoder is used with DAMSM as a convolutional neural network to extract the features out so that it can map to a common space. With CNN, the intermediate layers can learn various features associated with the different sub-regions of the image and the latter learns about the global features associated with them. A pre-trained Inception-v3 model on ImageNet was used as the image encoder. 768 is the local features dimension and it resizes the image to 299 X 299 pixels, to get 289 sub-regions in the image. In the end, these features are converted to similar space to that of the text encoder by adding perceptron layers.

3) Loss: For every generation, G_i a discriminator D_i and the loss is a combination of both conditional and unconditional at each stage. The embeddings of sentences is being conditioned on. The unconditional loss brings the generated images sampled from the generator of the particular distribution and is passed to the discriminator. The loss is minimized here so that the discriminator is fooled to think the image coming is real. For the conditional loss, passed $\bar{e}$ along with the generated image to the discriminator.

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(D_i\left(\hat{x}_i\right)\right]\right.}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(D_i\left(\hat{x}_i, \bar{e}\right)\right]\right.}_{\text{conditional loss}} \tag{8}$$

The discriminator uses cross-entropy loss and has data from the original distribution and the generated distribution. The discriminator will try to bring the original distribution close to 1 and the generated images output close to 0 to minimize the discriminator loss.

$$\mathcal{L}_{D_i} = \underbrace{-\frac{1}{2}E_{x_i \sim p_{\text{data}_i}}\left[\log D_i\left(x_i\right)\right] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(\hat{x}_i\right)\right)\right] +}_{\text{unconditional loss}}$$
$$\underbrace{-\frac{1}{2}E_{x_i \sim p_{\text{data}_i}}\left[\log D_i\left(x_i, \bar{e}\right)\right] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(\hat{x}_i, \bar{e}\right)\right)\right]}_{\text{conditional loss}} \tag{9}$$

B. RoBERTa

Attention GANs use basic RNN, which is a bidirectional LSTM. LSTM is used on the text description to extract the semantic vectors. With bi-directional LSTM, each word corresponds to two hidden states representing one for each direction [3]. RoBERTa: A Robustly Optimized BERT Pre-training Approach is the latest language model introduced by Facebook that optimizes the existing BERT architecture. It introduces the dynamic masking, hence the masked token changes during the training epochs. RoBERTa uses 160 GB of text for pre-training, including large Books Corpus and English Wikipedia are used in BERT. The additional data included CommonCrawl News dataset, Web text corpus and Stories from Common Crawl. The RoBERTa makes use of similar architecture as BERT Model but uses the byte-level BPE as the tokenizer [4]. The 'roberta-base' is the model used for prediction. The model was trained with an embedding dimension of 768. The pre-trained RoBERTa model is used to obtain the word and sentence embeddings and pass them with a fully connected layer before remaining in the Attention GANs architecture. The pre-trained model is 12 layered with

12 heads for the attention mechanism of the transformer and has around 125M parameters.

C. Deep Attentional Multimodal Similarity Model

Deep Attentional Multimodal Similarity Model (DAMSM) verifies if the generated image follows the description. It accompanies various steps to check this and update the network. The image features are brought f and $\bar{f}$ into a common space by adding a perceptron layer. The dimmension D is similar to the text encoders dimmension.

$$v = Wf, \quad \bar{v} = \bar{W}\bar{f} \tag{10}$$

$$v \in, R^{DX289}\bar{v} \in R^{D} \tag{11}$$

The matching score is driven by attention for the text and image features and calculate them as a pair The similarity matrix is calculated first for every pair in the sub-region of the image using

$$s = e^{T}v \tag{12}$$

In this equation, $s \in, R^{TX289}$ and i^{th} word in that sentence with the j^{th} sub-region available in that image. The normalized matrix for better result and stability,

$$\bar{s}_{i,j} = \frac{\exp\left(s_{i,j}\right)}{\sum_{k=0}^{T-1}\exp\left(s_{k,j}\right)} \tag{13}$$

Region-context vector c_i, is calculated. Earlier the interest was in generating the image, so it went through all the words and found the sub-region at each time. But, here it can be found if that particular word has any significance in the generation of that particular image. So for all the sub-region, it is needed to be checked one word at a time. This is summed here for all 289 sub-regions. $gamma_1$ is the attention scaling factor used in the equation and a score is generated for that and multiplied with the image features.

$$c_i = \sum_{j=0}^{288}\alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp\left(\gamma_1 \bar{s}i, j\right)}{\sum k = 0^{288}\exp\left(\gamma_1 \bar{s}_{i,k}\right)} \tag{14}$$

Word level relevance of i^{th} word is calculated with cosine similarity. It uses the current words, region-context vector and words vector. $R\left(c_i, e_i\right)$ tells us the score of each of those words on how important they are in generating the actual image.

$$R\left(c_i, e_i\right) = \left(c_i^{T}e_i\right) / \left(\|c_i\| \, \|e_i\|\right) \tag{15}$$

The word-level features are used to calculate the final global level scores. This image description score is calculated using word-level features with the hyperparameter γ_2, which signifies word to region context pair importance.

$$R(Q, D) = \log\left(\sum_{i=1}^{T-1}\exp\left(\gamma_2 R\left(c_i, e_i\right)\right)\right)^{\frac{1}{\gamma_2}} \tag{16}$$

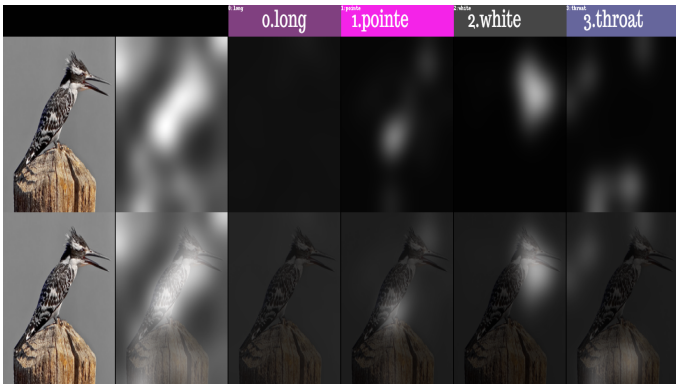


Fig. 3. Attention Map Generated by DAMSM on using RoBERTa Text Encoder. The Figure shows one Part of the Entire Caption Captured.

Similarly, using sentence-level features, using cosine similarity between global sentence and image features.

$$R(Q, D) = \left(\bar{v}^{T}\bar{e}\right) / \left(\|\bar{v}\|\|\bar{e}\|\right) \tag{17}$$

In the training process, calculation is done for the DAMSM loss for all the pairs. Thus with multiple descriptions and multiple images. The posterior probability is calculated of D_i matching with Q_i. So this gives a probability of how likely is that a description will be selected out of all the descriptions available. γ_3 is a hyperparameter for smoothing and stability in training the DAMSM. Similarly, it is also found the posterior probability when there is description and the images needs to be found.

$$P\left(D_i \mid Q_i\right) = \frac{\exp\left(\gamma_3 R\left(Q_i, D_i\right)\right)}{\sum_{j=1}^{M}\exp\left(\gamma_3 R\left(Q_i, D_j\right)\right)} \tag{18}$$

$$P\left(Q_i \mid D_i\right) = \frac{\exp\left(\gamma_i R\left(Q_i, D_i\right)\right)}{\sum_{j=1}^{M}\exp\left(\gamma_3 R\left(Q_j, D_i\right)\right)} \tag{19}$$

D. Total Loss

Total DAMSM loss in the network is calculated by

$$\mathcal{L}_{DAMSM} = \mathcal{L}_1^{w} + \mathcal{L}_2^{w} + \mathcal{L}_1^{s} + \mathcal{L}_2^{s} \tag{20}$$

L_1^{w} provides the word-level loss with respect to the description given the image and is the negative summation of log value of $P\left(D_i \mid Q_i\right)$. L_2^{w} provides the word-level loss with respect to the image given the description and is the negative summation of log value of $P\left(Q_i \mid D_i\right)$. L_1^{s} provides the sentence-level loss with respect to the description given the image. L_2^{s} provides the sentence-level loss with respect to the image given the description. Both the sentence level loss is same as word level loss except it use $\bar{e}$ instead of e.

Total loss in the entire network

$$\mathcal{L} = \mathcal{L}_G + \lambda\mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=1}^{3}\mathcal{L}_{G_i} \tag{21}$$

Here, L_G is the generator loss summed with L_{DAMSM} multiplied by a hyperparameter λ for smooth training.

Fig. 4. Image of an Indigo Bunting Generated at Every 50 Epochs by the Model.

Fig. 5. Comparison of Image Generated by the Model and the Ground Truth Image.

Fig. 6. Bird Generated by the Model for the Caption "This Bird has Wings that are Black and has a Red Belly".

E. Frechet Inception Distance (FID)

To analyze the model generated images, the Frechet Inception Distance (FID) score [17] is used as the metric. FID score is the best metric that can be used for the evaluation of the system as it measures the distance between the feature vectors of both real and generated images. The baseline model's paper has used the inception score as a metric for evaluating the GANs. The problem associated with the inception score being taken as a metric is that it does not find how the generated images compare with the actual images. With FID, it evaluate the generated images based on the generation distribution with the actual image in that particular target domain. For the FID score, the lower, the better. A lower score indicates that the generated images are closer to authentic images and are higher quality images and features with real one's match.

V. Experiments and Result

Multiple experiments were performed with the proposed Generative network using a pre-trained RoBERTa language model. The comprehensive network was initially pre-trained for DAMSM up to 200 epochs. Tesla V100 GPU with 16GB VRAM and 24GB CPU RAM for training with worker set as 4 were used for experimentation. The batch size for training was kept at 48 with a learning rate at encoder at 0.00005, and gradient clipping of 0.25 was kept to make sure training was stable. Various smoothing parameters were set during training, which helps train the various losses in multiple steps followed at DAMSM. $\gamma 1 = 4$, $\gamma 2 = 5$, $\gamma 3 = 10$, respectively. The parameters choosen for experimentation are taken from AttnGAN model and used for direct comparison. (GF_DIM) is the number of conv filters in the first layer of the generator and (DF_DIM) is the number of conv filters in the first layer

of the discriminator.

Ten captions were taken per image for training with the number of dimensions for the latent representation of the text embedding as 768. In previous experiments that were conducted in AttnGAN used bidirectional LSTM, which only required 256 embeddings. the base size was set as 299 which captured the attention map generated while pretraining the DAMSM with the pre-trained RoBERTa text encoder. The text encoder model was adopted from the hugging face library, providing the tokenizer for RoBERTa. It was observed that word vectors like colours get clustered together in vector space. The training of the transformer started from the pre-trained 'roberta-base' model with a RoBerTa tokenizer. The training of CNN started from the ImageNet pre-trained Inception-v3 model. Fig. 3 shows an attention map captured from image of a bird. At each frame a part of bird is being captured based on the text assosiated with it. The language model finds the relationship linking the body part and the colour designated for it. Attention mechanism explicates how each word corresponds to synthesizing a selective part of the bird image. Once the pre-training is completed, the DAMSM model generates a text encoder and an image encoder. This is used in the training of the AttnGAN architecture. AttnGAN network was trained using RoBERTa text encoder for 600 epochs. Due to limited resource allocation, kept the number of convolutional filters in the first layer of the generator (GF_DIM) and the number of convolutional filters in the first layer of the discriminator (DF_DIM) as 32.

The batch size was restricted to 8. The discriminator and generator learning rate was set to 0.0002. The generator and discriminator models were saved at every 50 epochs for analysis. For performance comparison with the baseline AttnGAN model, which trained the model with λ set as 5. The dimension of the RoBERTa text encoder is 768, amidst ten captions per image. The number of dimensions of the Noise vector was kept as 100 throughout the training process. Fig. 4 shows an Image generated at every 50 epoch by the AttnGAN with RoBERTa language model network. It is observable that around 200 epochs, the generator learns to generate an image close to a real-life bird. Images generated after 400 epochs looks realistic. By 600 epochs, it concluded the training and the models were saved. Fig. 5 explicates how the image generated by the model resembles the ground truth image. The model has learnt well to capture the essential details of birds like the body parts like wings, beaks, eyes, and feathers and understand its colour. It has also captured the pose of a bird to a reasonable extent. With more GPU power, it can use more generative networks to convert the image to higher quality. Fig. 6 was generated by the model around 600 epochs. The text as "this bird has black wings and a red belly", were provided to the generator synthesized an image matching the text description. 'roberta-base' was used as the model for capturing the context from the text description. The natural language model's main idea is to find attention heads and associate words in a bidirectional way. Fig. 7 envision the attention head for essential words in the sentence and how the RoBERTa model builds the attention mechanism. With the hugging face xbert tool, visualizing how the roberta-base model works in associating each word within the sentence with each other is simpler. The 'roberta-base' model uses 12 attention heads for generating a semantic relationship between each word in any direction. The word "this" is associated with itself and many other words within the sentence. The RoBERTa model learns that bird is the best associated and predictable word as the model is trained. There can be seen a strong connection between "this" and "bird". The word "bird" with the other words in the sentence is associated with particular words like "wings", "black", and "belly".These are the semantic relationship found by RoBERTa, and these get correlated with each other.

This assists in image generation and particularly in developing the text encoder in DAMSM loss. The word "black" here is a colour, and "wings" and "belly" are the two strong words that "belly" correlates. Colours like "red" get strongly related to "belly" in the sentence. Fig. 6 shows us the generated image by the model. The belly is red, and have black wings. This shows us how the RoBERTa model associates each word and what level of attention is provided for each word which nurses in the generation of the image with that particular features like body parts or colour.

TABLE II. FID SCORE GENERATED FOR EVERY 50 EPOCHS

Epoch	FID Score
0	275.833490
50	45.298608
100	32.540591
150	29.566392
200	25.616694
250	27.216560
300	24.648624
350	26.616447
400	22.377503
450	23.709113
500	22.760225
550	**20.773709**
600	21.468151

The Frechet Inception Distance(FID) score was calculated as part of the validation to compare the results with previous models as shown in Table II. Multiple epochs were ran and the FID score for λ = 5 were recorded. The number of generated images used for FID calculation was 2928, and the number of real images to be used in FID calculation was 11788. For epoch 0, had initially got a score of 275.833490. As each epoch is being trained, it can be seen the FID score keep reducing (Fig. 9).

A lower FID score indicates the model can generate more realistic images and various distributions of images. Around 550 epoch, a score of 20.773709 was obtained, the lowest and best for this model. This shows a good improvement from the baseline AttnGAN module [3] that used bidirectional LSTM as it got an FID score of 23.98 around the same number of epochs. Experiments were conducted on different values for λ at 100 epochs. Table III shows how the model performed without DAMSM and, by altering values of DAMSM, how the FID score or generation of different results is affected. A score of 35.44 was obtained. While tuning in too much attention value, the stability is lost in training, and it ends up with a score of 54.77 for λ set to 100. λ value of 1 seems to be stable for training with the AttnGAN and generating the bird images. A λ value set to 1 seems to work well for the model with RoBERTa embeddings and AttnGAN network. The λ value may fit differently for different datasets and should be

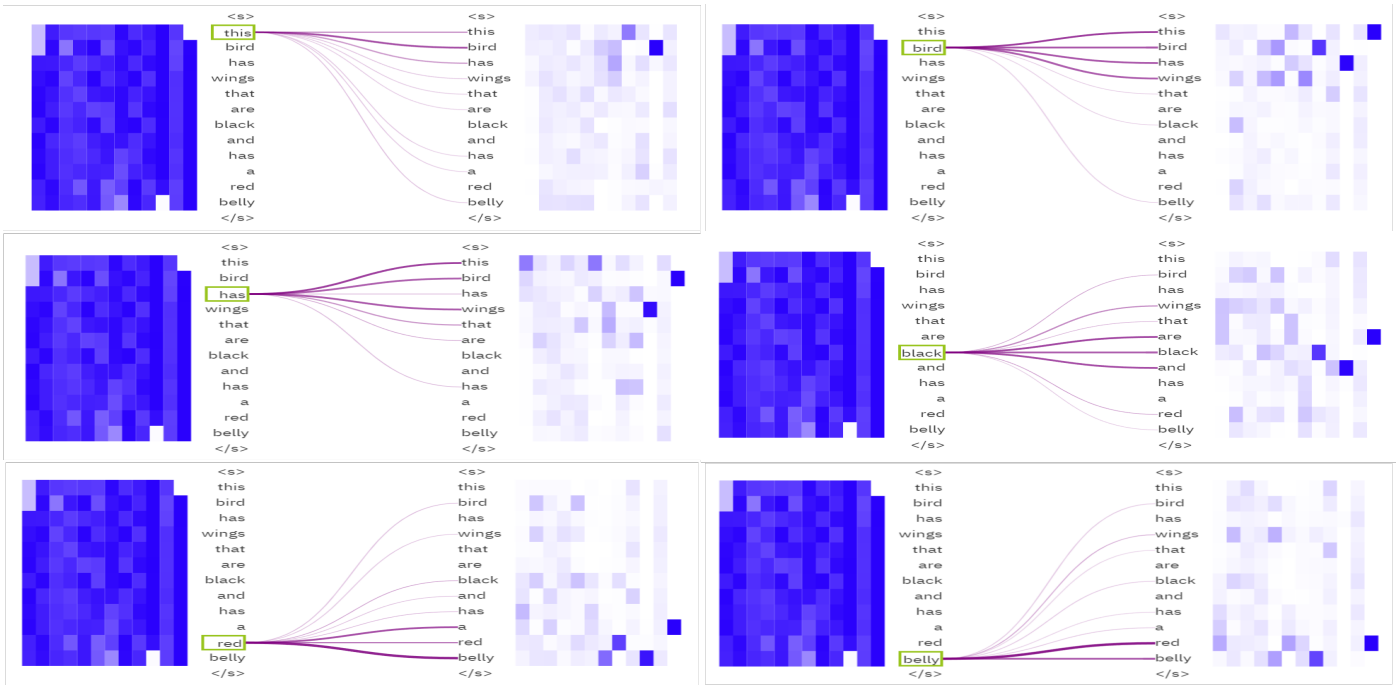


Fig. 7. Attention Head Generated by "Roberta-base" Pre-trained Model on the Text "This Bird has Black Wings and a Red Belly".

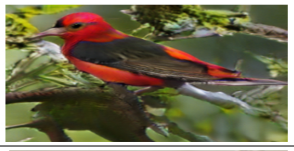
this bird has a red body and black wings with a small beak	
a small bird with a yellow belly and a small bill that curves down	
this bird is brown with white belly and long pointy beak	
bird with an all white body and a long black beak	
this is a yellow bird with a white head and a pointy beak	

Fig. 8. Images of Birds Generated by the Model for Various Provided Captions.

Fig. 9. Frechet Inception Distance(FID) Score for λ = 5.

explicitly experimented with that dataset. Table IV shows we have got a FID score of 20.77 with λ value set as 5 comparing to various other models. Fig. 8 shows some examples of text description and the images generated by the model.

TABLE III. FID SCORE GENERATED FOR 100 EPOCHS FOR DIFFERENT VALUES OF λ

Epoch	FID Score
0	35.440683
0.1	30.596095
1	28.663923
5	32.540591
10	34.538827
50	46.275016
100	54.775857

TABLE IV. COMPARISON OF FID SCORE WITH VARIOUS MODELS WHEN $\lambda = 5$

Model	GAWWN [12]	StackGANs [13]	AttnGAN [3]	RoBERTa GAN
FID Score	67.22	51.89	23.98	**20.77**

VI. CONCLUSION

This paper used the baseline AttnGAN model with the latest pre-trained language model RoBERTa. It used transformers with the generative network to analyze the fine-grained text to image generation. The generative network takes in captions in word and sentence embeddings level and uses the latent space of noise vector to synthesize birds images matching the text description. With the help of a deep attentional multimodal similarity model, and found the fine-grained image-text matching loss. This loss was further used to train the generator. Pre-trained Inception V3 model was used for the Image encoder along with pre-trained RoBERTa for Text Encoder. The baseline AttnGAN model had achieved a Frechet Inception Distance (FID) score of 23.98. The model with RoBERTa text encoder improved this performance and obtained a score of 20.77 on the CUB dataset. Various experiments were performed and recorded the results for the proposed architecture of generative networks.

REFERENCES

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

[2] Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S., Perona, P. "Caltech-UCSD Birds 200". California Institute of Technology. CNS-TR-2010-001. 2010.

[3] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[6] Sikha, O. K., Kumar, S. S., & Soman, K. P. (2018). Salient region detection and object segmentation in color images using dynamic mode decomposition. Journal of Computational Science, 25, 351-366.

[7] Subbiah, U., Kumar, D. K., Thangavel, S. K., & Parameswaran, L. (2020, September). An Extensive Study and Comparison of the Various Approaches to Object Detection using Deep Learning. In 2020 International Conference on Smart Electronics and Communication (ICOSEC) (pp. 183-194). IEEE.

[8] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks", in 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2017

[9] Aarthi, R., & Harini, S. (2018). "A Survey of Deep Convolutional Neural Network Applications in Image Processing". International Journal of Pure and Applied Mathematics, Vol. 118 No. 7, pp. 185-190.

[10] Brunda, R. & Divyashree, B. & Rani, N Shobha. (2018). Image segmentation technique- A comparative study. International Journal of Engineering and Technology(UAE). 7. 3131-3134. 10.14419/ijet.v7i4.18445.

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[12] Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016). Learning what and where to draw. Advances in neural information processing systems, 29, 217-225.

[13] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

[14] Li, B., Qi, X., Lukasiewicz, T., & Torr, P. H. (2019). Controllable text-to-image generation. arXiv preprint arXiv:1909.07083.

[15] Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1505-1514).

[16] Tsue, T., Sen, S., & Li, J. (2020). Cycle Text-To-Image GAN with BERT. arXiv preprint arXiv:2003.12137.

[17] Chong, M. J., & Forsyth, D. (2020). Effectively unbiased fid and inception score and where to find them. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6070-6079).