Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 9, No. 4, 2021, 475-485.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Multi-Sentence Hierarchical Generative Adversarial Network GAN (MSH-GAN) for Automatic Text-to-Image Generation

Elham Pejhan[1,2] and Mohammad Ghasemzadeh[1]*

1. Computer Engineering Department, Yazd University, Yazd, Iran.
2. Computer Science Department, University of Copenhagen, Copenhagen, Denmark.

Article Info

Article History:
Received 15 May 2021
Revised 21 June 2021
Accepted 08 July 2021

DOI:10.22044/JADM.2021.10837.2224

Keywords:
Generative Adversarial Networks (GANs), Deep Learning, Natural Language, Processing (NLP).

**Corresponding author: m.ghasemzadeh@yazd.ac.ir (M. Ghasemzadeh).*

Abstract

This research work is related to the development of technology in the field of automatic-text-to-image generation. In this regard, two main goals are pursued. First, the generated image should look as real as possible, and secondly, the generated image should be a meaningful description of the input text. Our proposed method is a multi-sentence hierarchical generative adversarial network (MSH-GAN) for the text-to-image generation. In this research project, we consider two main strategies: 1) produce a higher quality image in the first step, and 2) use two additional descriptions in order to improve the original image in the next steps. Our goal is to focus on using more information to generate images with a higher resolution using more than one sentence input text. We propose different models based on GANs and memory networks. We also use a more challenging dataset called ids-ade. This is the first time; this dataset has been used in this area. We evaluate our models based on the IS, FID, and R-precision evaluation metrics. The experimental results obtained demonstrate that our best model performs favorably against the basic state-of-the-art approaches like StackGAN and AttGAN.

1. Introduction

In the field of image processing, it is well-known that “a picture is worth a thousand words”. Since images can represent the events better and they can create deeper concepts, they have been used to describe the concepts and display information. With the advancement of the new technologies, the text-to-image generation problem has become an important area due to its applications in various fields such as automated content generation. This is a common field in various fields of science and technology including computer vision and natural language processing (NLP) [1].

One of the most common and challenging issues in the field of NLP and computer vision is the text-to-image generation. In this case, the goal is to generate an image from the given description automatically. From a high-level perspective, this problem can be considered as an example of the linguistic translation problem. In this way, different concepts and information can be expressed in two different languages, text and image, and each can be translated into another. However, these two issues are quite different from a language translation. In fact, text-to-image generation and captioning are considered the multi-dimensional issues. For example, suppose that we want to translate the simple phrase ``this is a beautiful flower" into French. In this case, a limited number of valid sentences can be presented as an acceptable translation, whereas if we want to produce an image that fits this sentence, a group of images may match it.

Although this multi-dimensional behavior also exists in the problem of image captioning, due to the coherence in the language, this problem is simpler than the problem of text-to-image generation. In the image-captioning problem, previous words can also be used in order to produce the next words, while this is not the case for the text-to-image generation.

In the matter of producing an image from a text, there are two main purposes: 1) the generated image should look as real as possible; 2) the generated image should describe the description of the input text. In the recent years, the generative adversarial networks [2] have been introduced that can produce a wide range of content such as images, text, and audio [3-5]. In this network, one deep neural network generates fake data, and another deep network is responsible for identifying whether the input data is real or fake. The scope of operation can be used to create an acceptable painting, poem, or piece of music [6]. Most of the proposed models for the text-to-image generation problem are designed based on the generative adversarial networks (GANs) [7-9]. One of the basic models is StackGAN, which uses a hierarchical method in order to generate an image [8]. Then other models based on the hierarchical methods have been proposed [10-12], which have added the attention mechanism to be able to produce images with a higher resolution and quality.

Although these methods can generate acceptable results, they still face challenges, which are listed in the following two main cases. First, the final generated images depend on the generated image in the first step. The hierarchical method will not be very successful if the first image is not appropriate. The second challenge is that each word in the input description contains a different level of information for generating the final image. The visual information should be aware of the importance of each word in each step to improve the generated image [7].

The methods presented so far have used only one sentence in order to generate the initial images. They use this sentence to improve the generated images in the next steps. They have used one sentence, while in the datasets used in this field, there exist at least five descriptions for each image.

In this paper, we introduce four different hierarchical methods based on GANs. As mentioned earlier, there exists more than one description for each image so we use three sentences to generate images and improve them. Our purpose is to focus on using more information from the training data to have higher resolution images. The differences between the proposed methods are 1) the way of selecting three sentences from the descriptions that exist for an image, 2) the approach of combining three different selected sentences in order to generate an image in the first step and improve it in the second and third combination steps. The basic structure of all networks is in such a way that first we use one sentence or a combination of the selected sentences to generate a low-resolution image. Then in the next two steps, the generated image from the previous step will be improved by two other sentences or a combination of them.

The structure of the existing memory provides the conditions at each step to retrieve the information that is more important for improving the generated image based on the attention mechanism.

In order to evaluate the performance of our proposed models, several sets of experiments are performed on the CUB-200 [13] and ids-ade [14] datasets, and the quality of the generated images is evaluated by the IS [15], FID [16], and R-precision [7] metrics.

The main contributions of this paper are as follow:

- We present the hierarchical GANs combined with the memory and attention mechanism using three sentences to generate the higher resolution images.
- We introduce the loss functions that fit our proposed models. These functions can more accurately assess the relevance of the generated images and the selected sentences.
- We focus on the new dataset, ids-ade, which is more complicated than CUB-200 and contains more than one object.

In order to illustrate our proposed models, in Section 2, we introduce the Generative Adversarial Networks (GANs) and memory-based structures. In Section 3, we briefly describe some of the research works in this area. The details of the proposed methods are in Section 4. Section 5 shows the experiment details. Finally, Section 6 presents the results of the experiments.

2. Preliminaries

This section provides the basic knowledge required to better understand the proposed models. At first, we describe the generative adversarial networks, which is the base of our models. Then we briefly show an introduction to the dynamic memory and attention mechanism.

2.1. Generative Adversarial Networks (GANs)

GANs are a very popular group of generative networks introduced in 2014 by Jan Goodflow [2]. These networks are based on the game theory approach. One deep neural network called Generator (G) competes with another network in the adversary process. Another deep network called Discriminator (D) tries to distinguish the

samples generated from the G network and the original data. The competition between these two networks will ultimately lead to a better learning and an improved performance for both networks. Equation (1) shows the competition between D and G, which is a type of Min-Max game.

$$V(D,G) = E_{x:P_{data(x)}}[LogD(x)] + E_{x:P_{z(z)}}[Log(1-G(z))] \quad (1)$$

In this equation, the first term is the entropy of the real data passing through D, and the network D tries to maximize it. In contrast, the second term is the entropy of the random data passing through G, and network D tries to bring it closer to zero. The function of network G is quite the opposite of the behavior of network D, and it tries to minimize the expression.

GANs can also extend to the conditional networks if both the generative network and the discriminative network contain additional information. The condition can be any kind of auxiliary information such as class labels or any other data. This condition can apply as an additional input layer to both the G network and the D network. Both networks use these additional inputs to configure and learn their parameters. In the text-to-image generation problem, the condition of the generative network is the input text.

2.2. Deep attentional Multi-modal Similarity Model (DAMSM)

The DAMSM model [11], with a text encoder and an image encoder, maps the words in the sentence and different regions of the image into a common space, and then it can calculate the text-image similarity. A text encoder is a bi-directional Long Short-Term Memory (LSTM) [17]. Bi-directional LSTM is a special type of recurrent neural network that can extract semantic vectors from the input description and learn long-term dependencies. In this network, each word corresponds to two hidden states that are in two directions, and the concatenation of these two vectors represents a semantic representation of the word. Each word in the description is the vector $e \in R^{D \times T}$, where T is the number of words in the description and D is the length of the word feature vector. The concatenation of the last hidden state of the bi-LSTM models is the sentence feature vector.

The image encoder is a Convolutional Neural Network (CNN) layer called Inception-v3 [18]. The resulting matrix of Inception-v3 contains the features related to the image regions. Each region of the image is the vector $f \in R^{N \times R}$, where R is the number of image regions and N is the length of the feature vector of each region. The last average pooling layer of Inception-v3 is the global feature vector of the whole image. Finally, in order to calculate the similarity of image regions and sentence words, the text and image feature vectors are mapped to a common space with equal dimensions using a perceptron layer.

2.3. Dynamic Memory Networks

A memory-based network [19] first stores the information in the external memory, and then uses the information in the next steps. In the recent years, some of these networks have been used in the structure of GANs using the key-value memory [20]. In this model, each value of the memory module has a weight called the memory key, which is used when calculating the output. A dynamic memory-based network [19] is a network that has been used in the recent research works on the text-to-image generation problem.

In this model, in each step, the words that are most relevant to the generated image are written in the memory. Thus, the generated images are more relevant to the input description. When the model is reading the memory, the image feature vector is used in the form of a query to retrieve information from the memory. In the next steps, the model uses the retrieved information from the memory to improve the quality of the initial image.

3. Related Works

Generating images from the text is a hard and complex problem in the fields of machine learning and computer vision that has been considered over the recent years. We have divided the research works in this field in two categories, as follow:

3.1 Traditional Text-to-image Retrieval

The early methods [20, 21] used a combination of search and supervised learning methods to retrieve the images related to the input text. The strategy is to calculate the correlation of the words in the text and the image regions, and select the more relevant words. The model then uses the selected words in order to retrieve the most relevant images. The problem with this solution is that it cannot generate the images with a new content.

3.2 GAN-based Text-to-image Generation

For the limitation of image retrieval, in the recent years, many types of research works have been introduced to solve the text-to-image generation problem with GANs [8, 9, 11, 12]. The research in

generative models has advanced significantly, and delivers solutions to learn from the training images and produce a new visual content. These methods are known as the multi-dimensional methods that combine the features of different models, algorithms, and ideas to improve problem-solving [10, 23]. In 2016, Reed *et al*. [7], for the first time, introduced a new architecture for image modeling that could transfer the characters to pixels using GANs. Their proposed model can produce plausible images of flowers and birds with a resolution of 64×64. Their proposed method cannot show the small details of the objects in the images like bird eyes. Their model also cannot generate higher resolution images such as 128×128. They tried to generate images by one GAN network in just one step, and it was the limitation of their method because they were not able to add more details to the generated images.

In order to solve that limitation, Zheng *et al*. [8] proposed a model called StackGAN that, for generating higher-quality images, divides the problem into two smaller sub-problems and solves them using GANs. In the first step, the model specifies the initial form and color of the objects in the text, and the output of this step will be a low-resolution image. In the second step, the model gets the generated image of the previous step and the text as the inputs to improve the image and to generate a high-quality image. They repeat this strategy for the third step to improve the resolution of the image. They use multiple GANs embedded in a tree structure to improve the resolution of the generated images in the GAN network. On the other way, this method generates images at different scales from different tree branches.

After that, many studies used the generative adversarial network hierarchically in order to generate the images with a higher quality [11, 12]. As mentioned earlier, in this process, first, the model generates a low-quality image, and in the next steps, improves the quality of the generated image. StackGAN uses the GAN network in the next two steps, and after generating the initial image, in the second step, again by using the GAN network, it produces the image with a resolution of 128×128. The second version, called StackGAN ++ [9], introduced a tree structure instead of the GAN grid in order to improve the image and use it in several steps to generate an image with a higher quality.

After that, Tao Xu *et al*. introduced the AttGAN network [11]. It is similar to the two networks mentioned in that it also uses the attention mechanism. The method of this network is that in the first step, based on the embedding of the text related to the whole sentence, the model generates a low-resolution image. In the next steps, the model improves the initial image with the attention mechanism. The attention mechanism retrieves the more important words in the input text. With this process, we first have an overview, and then the details will be added to the image over time. Using the attentional GAN was a very important contribution, allowing the model to focus on a specific region in the generated image and improve that region. Another worthy contribution of their model is to use DAMSM we described in Section 2.2. They used DAMSM after the result of the final stage to calculate the similarity between the generated image and the text embedding at both the sentence level and the more fine-grained word level.

After AttGAN, Zhang *et al*. proposed hierarchically-nested adversarial network (HDGAN) [1] in order to tackle the difficult problem of dealing with the photographic images from the semantic text descriptions. The main contributions of HDGANs include the introduction of a visual-semantic similarity measure.

In the networks mentioned so far, the final image quality depends on the quality of the initial generated image. If the initial generated image is not of acceptable quality, in the next steps, the model cannot improve the quality of the image well. In order to solve this problem, in 2019, Zhou and colleagues introduced the DM-GAN network [12]. The network improves the quality of the initial generated image using the dynamic key-value memory [20]. The memory module uses the initial image features as the search key. The model selects the words associated with the generated image dynamically at each step and writes on the memory.

Since generating high-resolution images is a difficult problem, the models we mentioned are trying to enhance the resolution of GANs. Our proposed model is also in this way, and is trying to generate the images using different GAN models in different steps.

Besides, there are some other models that use other methods. For example, in 2020, Jing Yu Koh *et al*. proposed a sequential model called TRECS [24]. TRECS uses the descriptions to retrieve the segmentation masks and predict the object labels aligned with mouse traces. The model selects the position of the masks using these alignments, and generates a fully covered segmentation canvas. In the last step, the model

uses the generated canvas as an input of a segmentation-to-image generator in order to generate the final image. As these methods are not the basic of our model, we just mentioned one sample.

Given the recent advances in the use of the hierarchical structure of GANs and dynamic memory in solving the problem of text-to-image production, we propose a method based on this. Unlike the other methods, our proposed method uses three sentences instead of one in order to produce and improve the image.

4. Our proposed Methods

Our proposed method is a multi-sentence hierarchical generative adversarial network (MSH-GAN) for the text-to-image generation problem. In this paper, we look at two key options: 1) produce a higher quality image in the first step, and 2) use two additional sentences to improve the original image in the next steps. In order to achieve the first, we use two additional sentences to generate the initial image.

In the datasets in this field, there are at least five descriptions for each image. In order to achieve the second goal, unlike the previous works that have used only one sentence to generate and improve the image, in this work, we use three sentences. Due to the high volume of calculations and the requirement for a more powerful hardware, it was not possible to perform the experiments with five sentences. In the following, we first describe the different architectures we have proposed. After that, we introduce the cost function provided for the proposed models.

4.1 Model Architecture

The proposed model is a hierarchical model that generates the image from low to high resolution. First, the model generates an image containing the generalities related to the description, and then the details will add to it. The hierarchical architecture has made a significant progress in the recent years [25-27]. We have presented and tested five different models and architectures, as follow:

1- MSH-BASE
2- MSH-CAT
3- MSH-SUM
4- MSH-Hybrid-V1
5- MSH-Hybrid-V2

As mentioned earlier, we use three sentences to generate the image in all models. One challenge for our problem is to choose three sentences out of five. We tested three different strategies for each one of the above models, which are listed below:

- CS123: the first three sentences in the training data in the same order.
- CS1RR: the first sentence is the same as the first sentence in the training data, and the two other sentences are selected randomly from the remaining four sentences.
- CSRRR: the three sentences are selected randomly from the five training sentences.

For all the proposed models, the text encoder, a Bi-LSTM model, generates the text embedding of the input description. The number of texts embedded in the training dataset is small. In order to increase the generality of the output model, the text embedding of the first step is sent to the conditional augmentation module [8] to generate the additional conditional variables. The conditional variable of this module is taken from the Gaussian distribution space $N(\mu(\varphi_t), \sum(\varphi_t))$; φ_t is the text embedding related to the input sentence, and $\mu(\varphi_t)$ and $\sum(\varphi_t))$ are the mean and covariance of the embedding vector, respectively. In this way, the problem of overfitting is reduced, and we will have a more powerful model.

In the next steps, instead of text embedding at the sentence level, we use its feature vector at the word level. Then the feature vector of the image produced in the previous step and the text embedding of the input are given to the dynamic memory to retrieve the words that are more important for improving the initial image. At this step, the generator generates a 128×128 image using the attention mechanism. This process is repeated in the next step using the third input text to generate an image with a resolution of 256×256. In the next sub-sections, we will describe the proposed models in details, and show the results obtained in the next section.

4.1.1. MSH-BASE Model

In this model, the first three sentences are selected from five sentences. In the first step, the feature vector of the first sentence at the sentence level passes through the conditional module, joined to the noise vector, and then sent to the generator network in order to generate the image in the size of 64×64. In the second step, the second sentence at the word level and the initial generated image are given to the second generator network in order to generate an image with the quality of 128×128. The third step receives the image produced in the second step and the feature vector at the word level for the third sentence. The third

generator network generates an image with a quality of 256×256.

4.1.2. MSH-CAT Model

In this model, after selecting three sentences, the feature vectors related to the three sentences are joined at the sentence level. The vector resulting from the conditional module is passed, and then connected to the noise vector and given to the first generator network in order to generate the initial image in the first step. Then for the second step, the feature vector of the second sentence at the vocabulary level is given to the second generator, and this process is performed for the third sentence. The logic used in this and the subsequent methods is that the sentences in the ids-ade dataset are independent from each other in that each sentence describes a different object. As mentioned earlier, in the first step, the main structure of the initial image is generated. Thus, we concatenate the feature vectors of three sentences to retrieve more knowledge for generating the initial image.

4.1.3. MSH-SUM Model

This method is similar to the MSH-CAT model, except that instead of concatenating the feature vectors of three sentences, the feature vectors are added together at the sentence level. First, the feature vectors are added together, passed through the conditional module, and then joined by a noise vector, and the rest of the path is similar to the previous method.

4.1.4. MSH-Hybrid-V1 Model

In this method, we propose to use all the knowledge of three sentences for all three steps. For this purpose, we first pass the feature vector of each one of the three sentences at the sentence level separately from the conditional module. We combine the resulting three vectors with the noise vector, and the first generating network produces the initial image based on this vector. In the next step, the network uses the feature vector of the second sentence in order to improve the quality of the initial image. The third sentence is also used to improve the image produced in the second step. The difference between this method and the MSH-CAT category is that instead of first joining three sentences together and passing through the conditional module and then connecting with the noise vector, each one of the sentences passes through the conditional module separately, and they are then joined together and connected to the noise vector.

4.1.5. MSH-Hybrid-V2 Model

This method is similar to the previous one, except that instead of joining three sentences and using them directly, the most effective words are selected from the whole three-sentence words using an attention mechanism. The method is that first, the feature vector related to the sentences are joined together, and then the network based on the attention mechanism selects 15 words; 15 is the mean length of the sentences in the dataset.

4.2 Cost Functions

The cost function for the generator and the discriminator are defined separately, which are described below.

The cost function of the generator consists of three parts, which are given in Equation (2).

$$L = L_G + \lambda_1 L_{CA} + \lambda_2 L_{DAMSM} \quad (2)$$

In this relation, L_G is the sum of the cost functions of the three generators in the proposed networks, each one of which is obtained from Equation (3).

$$L_{G_i} = -\frac{1}{2} E_{\hat{I}_i \sim P_{G_i}} D_i^{uc}(\hat{I}_i) - \frac{1}{2} E_{\hat{I}_i \sim P_{G_i}} D_i^{c}(\hat{I}_i, \varphi_t) \quad (3)$$

The first expression in this relation represents the conditional cost function, and the second expression represents the unconditional cost function. The unconditional cost function tries to make the generated image look as real as possible. The conditional function tries to increase the compatibility of the generated image with the input text.

The second statement in Equation (2) is the cost function for the conditional augmentation module [8] mentioned in Section 4.1, and is given in Equation (4). This relationship is the criterion of Kolbeck-Labler divergence, and shows the similarity of the conditional module distribution behavior with the normal distribution.

$$L_{CA} = D_{KL}(N(\mu(\varphi_t), \sum(\mu(\varphi_t)) \, \mathrm{P} N(0,1) \quad (4)$$

The third expression is the DAMSM cost function [11], which shows a measure of the relationship between the generated image and the input description. This phrase evaluates the relationship between the input description and the image produced at the sentence level and at the word level.

The cost function for each discriminator is obtained independently according to Equation (5).

$$
\begin{aligned}
L_{D_i} = \frac{1}{2}[&E_{\hat{I}_i : P_{data_i}} \max(0.1 - D_i^{uc}(I_i)) + \\
&E_{I_i : P_{data_i}} \max(0.1 - D_i^{c}(I_i, \varphi_t))] + \\
\frac{1}{3}[&E_{\hat{I}_i : P_{G_i}} \max(0.1 - D_i^{uc}(\hat{I}_i)) + \\
&E_{\hat{I}_i : P_{G_i}} \max(0.1 - D_i^{c}(\hat{I}_i, \varphi_t))] \\
&E_{\hat{I}_i : P_{data_i}} \max(0.1 - D_i^{c}(\hat{I}_i, \bar{\varphi}_t))
\end{aligned}
\qquad (5)
$$

In the above relation, φ_t is the input description, I_i is the original image related to the description in the training data, and $\hat{I}_i$ is the generated image in step i. It should be noted that the last phrase indicates the relationship of a wrong sentence (sentences from another description of the input of the training data) with the input image, which improves the learning power.

5. Experiments

The experiments were conducted in PyTorch on a cluster plat-format in the University of Copenhagen. For all models, we used the pre-trained Bi-LSTM with a size 256. In the CUB-200 dataset, for image encoding, we also used the pre-trained Inception-v3 model. Since the images of the ids-ade dataset were different from the ImageNet data, we fine-tuned Inception-v3 as an image encoder for the ids-ade dataset. In the learning process, ADAM was used as an optimizer with a batch size of 10. The number of iterations for the CUB-200 data was 600, and for the ids-ade data was 2000.

5.1 Datasets

In order to evaluate our proposed models, we carried out the experiments on the two datasets CUB-200 [13] and ids-ade [14]. In order to follow one of our goals, we used ids-ade with more complexity, objects, and details. Ids-ade had 3528 training data and 441 test data. This dataset had five dependent descriptions for each image. The first sentence was a general description of the image, which usually refers to the category to which the image belongs. An example of an image with its descriptions of this dataset is shown in Figure 1. As mentioned earlier, the present work is the first study to use the ids-ade dataset in this area.

5.2 Evaluation Metrics

We quantified our proposed models in terms of Inception Score (IS), Frechet Inception Distance (FID), and R-precision. Each model generated 30,000 images conditioning on the text descriptions from the test set for evaluation.

In GANs, the IS score [15] is used to evaluate the quality and variety of the generated images. This metric was calculated by Equation (6).

$$IS(G) = \exp(E_{X : P_g} D_{KL}(p(y \mid x) \,\|\, p(y))) \qquad (6)$$

In this equation, the term $p(y \mid x)$ means the conditional distribution of y concerning x, in which *y* is the label predicted by the Inception-v3 model. According to Equation (6), the closer the distribution of the generated images to the distribution of the training data has a greater IS. A large IS means that the generated outputs have a high diversity of images for all classes, and each image clearly belongs to a specific class. The IS score has no control over the degree of similarity of the generated images to the training images, and it just calculates the quality of the generated images. Therefore, we used another metric, FID, in order to overcome this limitation.

FID [16] computes the Frechet distance between the generated images and the real-world images based on the extracted features from an Inception-v3 network. FID is calculated by Equation (7).

$$
\begin{aligned}
F(r, g) = \| \mu_r - \mu_g \|^2 + trace(\Sigma_r + \Sigma_g \\
- 2(\Sigma_r + \Sigma_g)^{1/2})
\end{aligned}
\qquad (7)
$$

1: *This is a large bedroom with two large windows, a bed, and a two person chaise lounge.*
2: *The windows have striped curtains in front of them and a curtain rod that goes over both windows.*
3: *There is a ceiling light and fan in the center of the room.*
4: *There are two large pictures above the bed and dark colored nightstands on both sides.*
5: *There are table lights on the nightstands and several plants throughout the room.*

Figure 1. An example of the ids-ade dataset.

In this equation, r and g refer to the real images and generated images, respectively; μ_r and μ_g refer to the feature-wise mean of the real and generated images; and Σ_r and Σ_g are the

covariance matrix for the real and generated feature vectors, often referred to as sigma. A lower FID implies a closer distance between the generated image distribution and the real-world image distribution.

Since these metrics cannot find the dependency between the generated image and the input text, the R-precision [7] measure is introduced. The R-precision is measured by retrieving the relevant text given an image query. We computed the cosine distance between a global image vector and 100 candidate sentence vectors. The candidate text descriptions included R ground truth and 100-R randomly selected mismatching descriptions. For each query, if r results in the top R ranked retrieval descriptions are relevant, then the R-precision is r/R. In practice, we compute the R-precision with $R=1$.

5.3. Experimental Results and Analysis

As described in Section 4, we have five different models. In this section, we show our evaluations in terms of the three measures we mentioned. We have to mention that for any one of our models, the mean time for training was nine days, and the mean test time for generating the images and evaluating the metrics was one day with the hardware we had access from the University of Copenhagen.

For the first step, we run all our proposed models with different variants involving CS123, CS1RR, and CSRRR to see which one of our models was the best model. All the variants were done on the ids-ade dataset. Table 1 shows the results of the experiments for all the proposed models. We have to mention that we run an extra version for the MSH-Hybrid-V1 and MSH-Hybrid-V2 models called 'Single'. We proposed this version to use a three-sentence knowledge for generating the initial image, and also for improving the images in the second and third steps. This version is an enhanced version of MSH-Hybrid-V1-CSRRR, except that it does not use the second and third sentence feature vectors at the word level to improve the images in the second and third steps. Instead, it uses the same feature vector as the first step.

Table 1 shows that "MSH-Hybrid-V1-Single" is the best model. This model has the highest value for R-precision and IS as well as the minimum value for the FID criteria, and we choose it as the best model. Figure 2 shows the evaluation values for the best iteration. As mentioned earlier, this is the first time that the ids-ade dataset has been used for the text-to-image generation problem.

Table 1. Results of our models on the ids-ade dataset.

Model	Version	FID↓	IS↑	R-precision↓
MSH-BASE	CS123	20.72	3.99	0.6583
	CS1RR	24.05	4.04	0.6692
	CSRRR	17.32	4.11	0.7036
MSH-CAT	CS123	23.68	4.54	0.6968
	CS1RR	20.84	4.73	0.7057
	CSRRR	24.09	4.56	0.7182
MSH-SUM	CS123	17.61	4.27	0.677
	CS1RR	15.68	3.85	0.7087
	CSRRR	22.67	4.05	0.7093
MSH-Hybrid-V1	CS123	17.76	4.53	0.7426
	CS1RR	31.83	4.23	0.7462
	CSRRR	18.19	4.53	0.7574
	Single	**17.76**	**5.17**	**0.7973**
MSH-Hybrid-V2	Single	31.44	4.91	0.9764
	CSRRR	42.52	4.93	0.6469

Therefore, in order to compare our proposed model with the previous methods, we propose two solutions, which are explained in the following. For comparison with the other models, we selected our best model, which was "MSH-Hybrid-V1-Single".

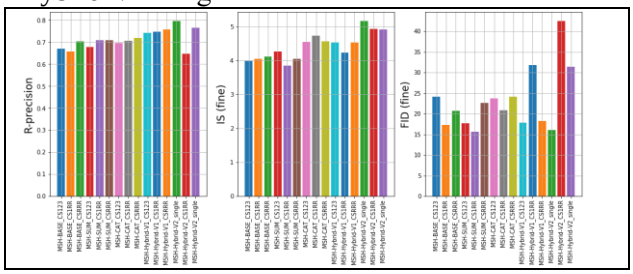

Figure 2. Best iteration evaluation results for the proposed models.

First approach: In this approach, we run one of the best models in this area, called the DM-GAN model, on the ids-ade dataset. We used the corresponding code, which was available online on GitHub (https://github.com/MinfengZhu/DM-GAN). Since the DM-GAN method uses only one sentence to generate the image, we used two methods for selecting the sentence: 1) random selection (DM-GAN-CSR) and 2) first sentence selection (DM-GAN-CS1). Table 2 shows the values of IS and R-precision for DM-GAN along with our best model.

As shown in Table 2, our proposed model performs better, and for the ids-ade dataset, it produces higher quality images, and the generated images have a better relation with the input text.

On the other hand, for the DM-GAN model, the DM-GAN-CSR version, which randomly selects a sentence, performs better than the DM-GAN-CS1 version, which uses only the first sentence. This result was predictable because, as mentioned earlier, the first sentence in this set of data contains general information that usually refers to the image category, while the rest of the sentences describe the objects with more details in the image.

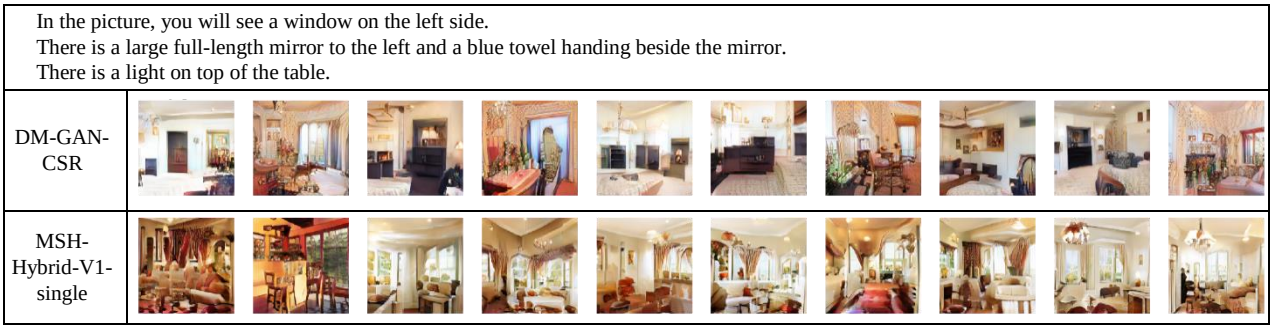

Figure 3. An example of the generated images of DM-GAN-CSR and MSH-Hybrid-V1-Single for the ids-ade dataset.

Figure 3 shows an example of the generated images for the best version of the DM-GAN model and our best model for the ids-ade dataset.

Table 2. Comparison of MSH-Hybrid-V1-Single model with DM-GAN model on the ids-ade dataset.

Model	R-precision↑	IS↑
DM-GAN-CSR	69.73	4.61
DM-GAN-CS1	65.3	4.35
MSH-Hybrid-V1-single	**79.73**	**5.17**

Second approach: In this approach, we implemented our best proposed model, MSH-Hybrid-V1-single, on the CUB-200 dataset, and compared the results with the state-of-the-art methods. Table 3 shows the results of three evaluation metrics. For FID and IS, we used the pre-trained Inception-v3. The values for the other methods were derived from the results published in the related articles. Table 3 shows that our proposed method on the CUB-200 dataset has better values for R-precision and IS than the other methods. DM-GAN has the best value for FID but our proposed method has a small distance ratio with that.

Table 3. Comparison of the proposed model with the state-of-the-art methods on the CUB-200 dataset.

Model	R-precision↑	FID↓	IS↑
GAN-INT-CLS [7]	-	-	2.88
GAWWN [1]	-	-	3.62
StackGAN [8]	-	-	3.70
AttnGAN [11]	67.82	23.98	4.36
DM-GAN [12]	72.31	**16.09**	4.75
MSH-Hybrid-V1-single	**79.27**	18.04	**4.80**

Based on the results obtained, we think that DM-GAN works better because the descriptions on the CUB-200 are not very different, and using three sentences does not improve the results than the models that only use one sentence. Figure 4 shows two examples of the generated images by our proposed model and the DM-GAN model.

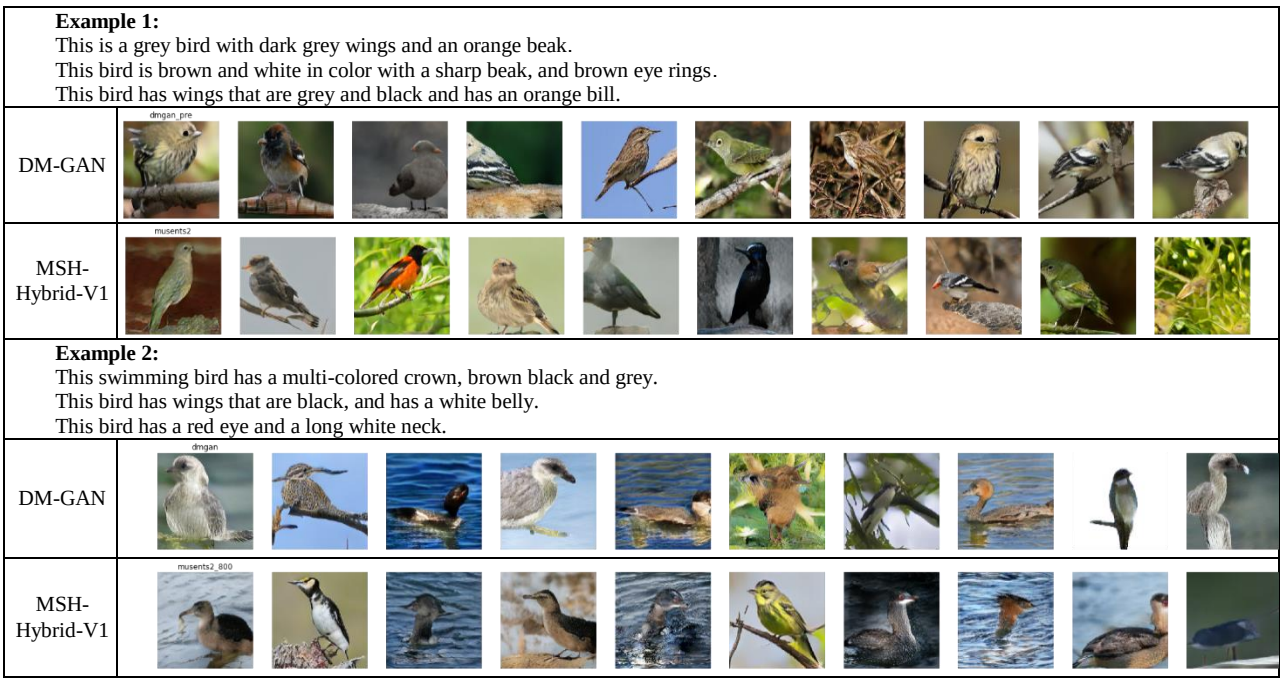

Figure 2. Examples of the generated images of DM-GAN-CSR and MSH-Hybrid-V1-Single for the CUB-200 dataset.

For each example, the first row contains three descriptions that have been selected randomly for our proposed method, and the first of these three sentences has been used for the DM-GAN method. Each example was executed ten times, and in each row, ten resulting images were displayed for that input text.

7. Conclusions

In this paper, we introduced different hierarchical memory-based methods using Generative Adversarial Networks (GANs) in order to solve the text-to-image generation problem. Compared to the previous models, we used three sentences instead of one sentence to generate the images. We trained our models on the ids-ade dataset. The ids-ade dataset is a complex one, and it includes the images with more than one object and details that are more complicated. We tested our models on the CUB-200 and ids-ade datasets in terms of different measures in order to evaluate the quality of the generated images and the relation between the generated images and the input texts. The experimental results obtained show that our best-proposed method performs better than the previous methods. Despite the improvements, the proposed model still faces challenges. The results obtained are particularly sensitive to the type of sentence selection for the ids-ade dataset, where the descriptions are dependent. The ids-ade dataset we used in this field for the first time has fewer training examples than the other datasets previously used in this field. This challenge makes the training more difficult for the GANs network. Therefore, in some cases, our proposed model cannot produce an image of high quality. At the beginning of this research work, we were thinking of using five sentences instead of three but there were more than 25 million parameters for learning, and we had a GPU limitation, and we just used three models.

In this area, the hardware limitation is one of the main problems. Not having enough data for training and time-cost were the other challenges we had in this research work. In the future, we will try to design a more powerful model to generate the initial images with a better quality. We used a simple attention mechanism for one of our proposed methods. We will also try to use an attention mechanism that can help the training phase of the model more. We will try the parameter sharing method in our models in order to make our model simpler and see if we can use more than three sentences to improve the quality of the generated images.

Acknowledgment

This research work was supported by the 'Iran Ministry of Science, Research and Technology'. Part of the work was carried out when the first author was visiting the University of Copenhagen. She gratefully acknowledges the hospitality from the host university.

References

[1] Z. Zhang, Y. Xie, and L. Yang, "Photo-graphic text-to-image synthesis with a hierarchically-nested adversarial network", *in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 6199-6208, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks", *in Advances in neural information processing systems*, pp. 2672-2680, 2014.

[3] Y. Li, Y. Chen, and Y. Shi, "Brain tumor segmentation using 3D generative adversarial networks", *International Journal of Pattern Recognition and Arti_cial Intelligence*, p. 2157002, 2020.

[4] Y. Li, Z. He, Y. Zhang, and Z. Yang, "High-quality many-to-many voice conversion using transitive star generative adversarial networks with adaptive instance normalization" *Journal of Circuits, Systems and Computers* (2020).

[5] A. Fakhari. and K. Kiani. "An image restoration architecture using abstract features and generative models." *Journal of AI and Data Mining*. Vol. 9, No. 1, pp. 129-139, 2021.

[6] M.M. Haji-Esmaeili and G. Montazer, "Automatic coloring of grayscale images using generative adversarial networks", *Journal of Signal and Data Processing (JSDP)*, Vol. 16 (1), pp. 57-74, 2019.

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis", arXiv preprint arXiv:1605.05396, 2016.

[8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X.Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks", *in Proc. of the IEEE int. conference on computer vision*, pp. 5907-5915, 2017.

[9] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks", *IEEE transactions on pattern analysis and machine intelligence*, 41, 1947-1962, 2018.

[10] K.J. Joseph, A. Pal, S. Rajanala, and V.N. Balasubramanian, "C4synth: Cross-caption cycle-consistent text-to-image synthesis", *in IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 358-366, 2019.

[11] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks", *in Proc. of the IEEE conf. on computer vision and pattern recognition*, pp. 1316-1324, 2018.

[12] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis", *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802-5810, 2019.

[13] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, The caltech-ucsd birds-200-2011 dataset, 2011.

[14] N. Ilinykh, S. ZarrieB, and D. Schlangen, "Tell me more: A dataset of visual scene description sequences", *in Proc. of the 12th International Conference on Natural Language Generation*, pp. 152-157, 2019.

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans", *in Advances in neural information processing systems (NIPSs)*, pp. 2234-2242, 2016.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium", arXiv preprint arXiv:1706.08500. 2017.

[17] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification", *in Proceedings of the 54th annual meeting of the association for computational linguistics*, pp. 207-212, 2016.

[18] C. Szegedy, V. Vanhoucke, S. Io_e, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", *in Proc. of the IEEE conf. on computer vision and pattern recognition*, pp. 2818-2826, 2016.

[19] C. Gulcehre, S. Chandar, K. Cho, and Y. Bengio, "Dynamic neural Turing machine with continuous and discrete addressing schemes", *Neural computation*, 30, 857-884, 2018.

[20] A. Miller, A. Fisch, J. Dodge, A. H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents", *in Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[21] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes", *in European Conf. on Computer Vision*, pp. 776-791, 2016.

[22] X. Zhu, A.B. Goldberg, M. Eldawy, C.R. Dyer, and B. Strock, "A text-to-picture synthesis system for augmenting communication", *in (AAAI)* , pp. 1590-1595, 2007.

[23] A. Dash, J.C.B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "Tac-gan-text conditioned auxiliary classifier generative adversarial network", arXiv preprint arXiv:1703.06412, 2017.

[24] J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Text-to-image generation grounded by fine-grained user attention". *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 237-246, 2021.

[25] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multi-modal machine learning: A survey and taxonomy". *IEEE transactions on pattern analysis and machine intelligence* 41, 423-443, 2018.

[26] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training", *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12174-12182, 2019.

[27] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation", *in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2327-2336., 2019.

شبکه خصمانه سلسله مراتبی چندجمله‌ای برای تولید خودکار متن از تصویر

الهام پژهان[۱و۲] و محمد قاسم‌زاده[۱*]

[۱] دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران.

[۲] گروه علوم کامپیوتر، دانشگاه کپنهاگ، کپنهاگ، دانمارک.

ارسال ۲۰۲۱/۰۵/۱۵؛ بازنگری ۲۰۲۱/۰۶/۲۱؛ پذیرش ۲۰۲۱/۰۷/۰۸

چکیده:

این کار تحقیقاتی در ارتباط با توسعه فناوری در زمینه تولید متن از روی تصویر به صورت خودکار است. در این راستا دو هدف اصلی دنبال می‌شود. اولاً، تصویر تولید شده باید تا حد ممکن واقعی به نظر برسد، و دوم اینکه، تصویر تولید شده باید توصیف معنی‌داری از متن ورودی باشد. روش پیشنهادی این پژوهش، به‌کارگیری یک شبکه رقابتی مولّد سلسله مراتبی چندجمله‌ای (MSH-GAN) برای تولید متن از تصویر است. در این پروژه تحقیقاتی، دو استراتژی اصلی را در نظر گرفته‌ایم: ۱) تولید تصویر باکیفیت بالا در مرحله اول و ۲) استفاده از دو توضیح اضافی برای بهبود تصویر اصلی در مراحل بعدی. هدف تمرکز بر استفاده از اطلاعات بیشتر برای تولید تصاویر با وضوح بالاتر با استفاده از متن ورودی با بیش از یک جمله است. در این پژوهش مدل‌های مختلفی بر اساس GAN ها و شبکه‌های حافظه پیشنهاد شده‌اند. ضمناً یک مجموعه داده چالشی به نام ids-ade برای ارزیابی نتایج لحاظ شدند. در این رابطه مدل‌های پیشنهادی بر اساس معیارهای IS، FID و R-precision ارزیابی شدند. نتایج آزمایشی حاکی از آن است که مدل پیشنهادی در برابر رویکردهای پیشرفته مانند StackGAN و AttGAN عملکرد مطلوبی دارد.

کلمات کلیدی: شبکه‌های رقابتی مولّد، یادگیری عمیق، پردازش زبان طبیعی.