Article

A Study on Generating Webtoons using Multilingual Text-to-Image Models

Kyungho Yu [1], Hyungho Ju [1], Jeongin Kim[1], Chanjun Chun[1], Pankoo Kim[1,*]

Department of Computer Engineering, Chosun University, 309 Pilmun-Daero, Dong-Gu, Gwangju 61452, Korea;infinitegh@chosun.ac.kr(K.Y.); snowlisakim@gmail.com(H.J.); jungingim@gmail.com(J.K.); cjchun@chosun.ac.kr(C.C.)

* Correspondence: pkkim@chosun.ac.kr(P.K.)

Abstract: Recent advances in deep learning technology have led to increased interest in text-to-image technology, which enables computers to create images from text by simulating the human process of forming mental images. The GAN-based text-to-image technology involves the extraction of features from input text, which are combined with noise and then used as input to a GAN that generates images that are similar to the original images through competition between the generator and discriminator. Although generating images from English text is a mature area of research, text-to-image technology based on multilingualism, such as Korean, is still in its early stages of development. Webtoon is a digital comic format that allows comics to be viewed online. The creation process for webtoons is divided into story planning, content/sketching, coloring, and background drawing. Since each stage of webtoon production requires human intervention, it is both time-consuming and expensive. As a result, deep learning technologies such as automatic coloring and automatic line drawing are being used to reduce human involvement. However, there is a shortage of technology that can assist authors with story creation in webtoon production. Therefore, this study proposes a multilingual text-to-image model capable of generating webtoon images when presented with multilingual input text. The proposed model employs Multilingual BERT to extract feature vectors for multiple languages, and trains a DCGAN in conjunction with the images. The experimental results demonstrate that the model can generate images that are similar to the original images when presented with multilingual input text after training.

Keywords: Multilingual BERT; Text-to-image; DCGAN; Webtoon; GAN

1. Introduction

With the advancement of deep learning techniques, research on training computers to generate images has been actively conducted[1]. Image generation technology has rapidly progressed, starting with Generative Adversarial Networks (GANs), which generate fake images similar to the real ones by an adversarial learning process between generator and discriminator networks. Currently, it can generate photorealistic images that are difficult to be distinguished from those drawn by humans.

Text-to-image technology refers to the process of enabling computers to generate images based on the input text[1,2]. Deep learning-based text-to-image technology creates images that reflect the contextual features of the text used to condition the image generator[3]. Traditional text-to-image approaches extract keywords from the sentence and synthesize images that correspond to those keywords, and require human intervention. However, deep learning-based text-to-image models extract features from the input text and generate images based on those features. Initially, text-to-image models trained on generative adversarial networks could create images similar to the input text by conditioning the image generator with the sentence's feature vector. However, this method failed to reflect the contextual meaning of individual words in the sentence and generated low-quality images. To overcome this limitation, attention mechanisms were introduced in AttnGAN [4], which improved the quality of the generated images. Since then, text-to-image models such as

 doi:10.20944/preprints202304.0935.v1

StackGAN [5], MirrorGAN [6]have been developed. Despite the development of GAN-based text-to-image models, they suffer from unstable image generation due to the training imbalance between the generator and discriminator. Recently, text-to-image technology has focused on multimodal learning and diffusion models to overcome these limitations. Multimodal learning involves representing information in various forms, such as images, sounds, and text, and clustering similar data through contrastive learning. Dalle-2 [7], developed by OpenAI, and DreamBooth [17], developed by Google, utilize a multimodal space as an encoder and a diffusion model as a decoder to generate images based on the predicted image embedding from the input text. Currently, most deep learning-based text-to-image methods couple English language text with image data, which limits their accessibility in countries where English is not spoken as a primary language. Furthermore, translating the native language into English as a pre-processing step will incur additional complexity and may not achieve the same efficiency. Therefore, the development of multilingual text-to-image technology that can generate images of similar quality regardless of the input language is necessary.

A deep learning-based text-to-image model that generates images similar to those drawn by humans can be used in the entertainment industry for creating virtual characters, animations, and more. Webtoons, a portmanteau of "web" and "cartoon," are comics that are published on the internet. The creation process of webtoons is divided into story planning, storyboarding/sketching, coloring, and background drawing stages. As each stage of webtoon production requires human involvement, it is time-consuming and costly. To minimize human intervention, artificial intelligence-based methods such as automatic coloring, automatic line drawing, and style transfer techniques are being utilized in different stages of webtoons creations.

In order for webtoons to be attractive to readers, the overall story of the webtoon is important. To keep readers interested and immersed as the webtoon progresses, it is important to provide tension and resolution. To ensure that readers' interest and immersion are not lost, the author goes through the stage of planning the story before creating the webtoon. In the planning stage, the author sets the genre, characters, and world view. Then, before drawing the webtoon, the author describes a scene in detail in writing, called a treatment. Treatment is a term used not only for webtoons but also for content production such as movies and dramas. When producing movies, dramas, or webtoons, treatments are used as a reference to describe the key elements of a scene such as the time, place, and characters involved. Thus, treatments contain rich information about the scenes of the final webtoon product, making them suitable input data for deep learning-based text-to-image models.

To overcome these limitations, we propose a multilingual text-to-image model that can generate webtoon images not only from English but also from other languages as input. To develop the multilingual text-to-image model, we construct and train a webtoon dataset consisting of English, and Korean texts, and image data. The training process involves extracting features from multilingual text input using a multilingual BERT and training a GAN-based text-to-image model with the image data. Once the training is completed, the multilingual text-to-image model can generate webtoon images when given multilingual text input. The main contributions of this work are: 1) Unlike existing text-to-image technology that generates images from text written in English, the proposed multilingual text-to-image model uses multilingual BERT as a text encoder to generate images from multilingual input. 2) Additionally, since the webtoon dataset is trained on the text-to-image model, it is expected to contribute to the creation of webtoon content by allowing the model to generate webtoon images from multilingual text input.

The structure of this paper is as follows: Section 2 describes the self-supervised pre-training models and deep learning-based text-to-image technology used in this study, while Section 3 proposes the construction method for the multilingual treatment-webtoon dataset and the GAN-based text-to-image model. Section 4 trains the constructed multilingual text-to-image dataset on the text-to-image model and verifies the results. Finally, Section 5 concludes the study and discusses future research.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

2. Related work

2.1 Self-supervised Pre-trained Models

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on teaching machines to understand human language. For this purpose, it is necessary to transform human language into a suitable format, which is called word embedding [11]. Word embedding refers to mapping the words that make up a text to a real-valued vector. When words are input into this type of word embedding, they are vectorized in a multi-dimensional space, which allows for the measurement of the degree of similarity between words. Embedding models such as word2vec and GloVe predict intermediate or surrounding words within a predefined window size in a sentence, resulting in word embedding. However, this method cannot fully capture the context of words in sentences, resulting in the same embedding being assigned to words with different contextual meanings. To address this issue, Embeddings from Language Model (ELMo) was developed, which is a pre-trained language model that uses bidirectional LSTMs to reflect contextual meaning in sentence embedding models [12]. Later, BERT was developed as a large-scale pre-trained language model based on the transformer architecture [13]. BERT is a pre-trained language model that uses attention mechanisms to prevent information loss that can occur with RNN or LSTM models as the length of the input sentence increases. The BERT model consists of a transformer encoder composed of multi-head attention and fully connected layers. The input sentence is tokenized, segmented, and position-embedded before being partially masked and input into the transformer. Training is performed by predicting the masked tokens, enabling the model to understand the context. The trained model can then extract feature vectors from the input sentences, which can be then used to perform several NLP tasks such as sentence classification and question answering [15]. Many transformer-based derivative models have been developed after BERT, and they have demonstrated excellent performance in SOTA.

2.2 Deep learning-based text-to-image generation

Deep learning-based text-to-image generates images through the semantic information of the input sentence. GAN-based text-to-image generation can be divided in two stages: first is the extraction of semantic information from the text and the second stage is the image generation. In the first, the sentence is embedded with word embedding and the features are extracted by inputting it into an RNN or transformer structure. In the image generation stage, the GAN inputs the semantic information of the text and noise of the same size as that of the output generated image to the Generator to create an image. In contrast, the discriminator compares the generator's generated images with the real images. Through adversarial learning of the two neural networks that constantly create fake images and judge whether the generated image is real or not, they generate fake images that are similar to the real ones. The initial deep learning-based text-to-image model used deep convolutional GAN (DCGAN) [3]. The DCGAN-based text-to-image model generates an image by using the conditional vector of the sentence as a condition. Although the generated image is similar to the meaning of the input sentence, it has the disadvantage of not reflecting the contextual meaning of each word in the sentence and generating low-quality images. To overcome these disadvantages, AttnGAN, which introduces the attention mechanism to image generation, was developed. AttnGAN uses the feature vector of the sentence to create an image first, and when generating the next stage image, it combines the attention map of the word with the image vector to create an improved image step by step. AttnGAN could express each word in the sentence in a detailed manner compared to DCGAN, and subsequently, stackGAN, MirrorGAN, and DM-GAN were developed, which can express the meaning of the input sentence more precisely and generate high-resolution images [4-6].

Recently, diffusion model-based text-to-image methods have shown better performance than GAN-based methods. GANs suffer from the problem of imbalanced training between the generator and discriminator, which leads to the issue of collapsing. When the generator only creates images that are easy to fool the discriminator, it stops learning. On the other hand, diffusion models generate images by iteratively adding noise to the training data in the forward process and recovering data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

from noise in the reverse process [8]. After training, diffusion models generate images similar to the training data by using the reverse process. The DALL-E 2 model, released by OpenAI, uses the Contrastive Language-Image Pre-training (CLIP) as an encoder and the diffusion model as a decoder to generate images from text [7]. CLIP extracts features from both text and image data and determines the similarity between them using contrastive learning. After training, the text and image features with similar characteristics are densely packed in one multimodal space [16]. CLIP extracts the image vector that is similar to the input text from the pre-trained model and generates an image by inputting it to the diffusion model. Diffusion model-based text-to-image methods have shown superior performance compared to GAN-based methods on benchmark datasets in generative modeling. Examples of diffusion model-based text-to-image generation methods include Dalle-2, DreamBooth, and Imagegen [17].

GAN and diffusion models have different strengths and weaknesses [18, 19]. GAN has the advantage of shorter training time compared to diffusion models, but it has a risk of model collapse due to the learning imbalance between the generator and discriminator. In contrast, diffusion models have the advantage of being able to generate a wider variety of high-quality images than GANs, despite requiring longer training time. In this study, we train a DCGAN-based text-to-image model on a multilingual text-image webtoon dataset specialized in the webtoon domain, to enable the application of deep learning-based text-to-image technology in the webtoon industry with a shorter training time.

3. Text-to-image generation using Multilingual BERT

In this section, we propose a multilingual text-to-image model that generates webtoon images when given multilingual text input. The proposed text-to-image model for webtoon generation aims to generate webtoons similar to the given text input in English and Korean, and proceeds in two steps. In the first step, we construct a multilingual text-to-webtoon dataset by using the MSCOCO dataset, a benchmark dataset, and the cartoon GAN. In the second step of the training process, we use Multilingual BERT to extract sentence vectors from the multilingual text-to-webtoon dataset and use them as conditions for DCGAN training.

3.1 Webtoon dataset

To generate webtoon images by using treatment as an input of text-to-image generation models, it is necessary to train the model on both treatment and webtoon datasets. As it is difficult to construct a large-scale dataset consisting of actual treatment and webtoons, the benchmark dataset commonly used in generative models, the MSCOCO dataset, is transformed into cartoon images using a pre-trained CartoonGAN [20]. The MSCOCO dataset is a description-photo dataset released by Microsoft, consisting of 123,287 training and validation images, each accompanied by five descriptions [21]. The officially provided dataset only includes English, so the Korean MSCOCO dataset, which was translated from the MSCOCO dataset by AI hub in Korea, was additionally used to construct a multilingual treatment dataset [22]. The Korean MSCOCO dataset also consists of five Korean descriptions per image, resulting in 10 sentences per image, five in English and five in Korean. The pre-trained CartoonGAN is trained on four styles, as shown in Figure 1. Thus, it is possible to construct 1,232,870 treatment-webtoon data pairs for each style, resulting in a total of 4,931,480 data pairs for all four styles.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

Figure 1. Examples of datasets transformed using CartoonGAN

3.2. Text-to-image generation with Multilingual BERT

A deep learning-based text-to-image model generates images from input text through two stages. The first stage is feature extraction of the input text, and the second is training a GAN model by combining the extracted text features with noise. Extracting text features that reflect the contextual characteristics embedded in the text is crucial, as the quality of the generated images is heavily influenced by the quality of the learned features. In other words, if the model fails to extract features that capture the essence of the key words between the text and the image, the generated image may deviate from the intended meaning of the text. Therefore, in this study, as shown in Figure 2, we utilize a pre-trained Multilingual BERT model, which has demonstrated high performance in natural language processing (NLP), to extract feature vectors of the sentences. The sentence feature vectors are obtained by using the "cls" token in the BERT model as the sentence vector.

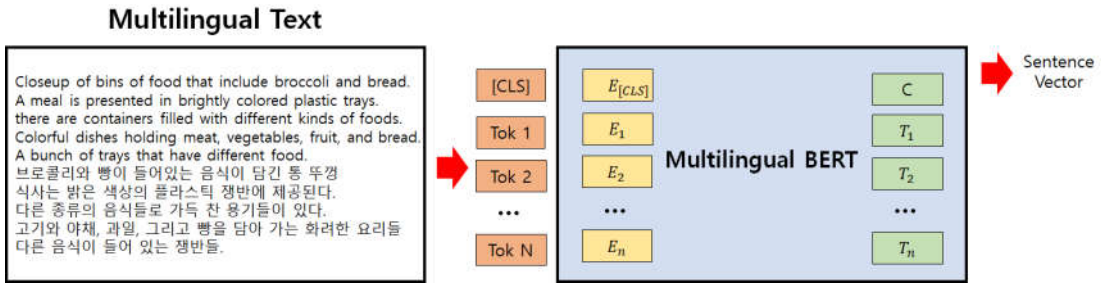


Figure 2. Sentence vector extraction using Multilingual BERT

To generate webtoons based on the extracted features of multilingual treatments, a GAN-based DCGAN model is used. Figure 3 shows the structure of the Text-to-Image model using DCGAN. GAN generates images similar to the original data through competition between two neural networks called generator and discriminator. The generator learns by reflecting the distribution of input data and generates fake data from random vectors, while the discriminator compares the data generated by the generator with the original data to determine whether it is real or fake. To generate webtoons from multilingual treatments, the sentence vectors extracted through Multilingual BERT are combined with noise and input to DCGAN for training. The proposed generator model consists of six blocks composed of convolutional layer, dropout layer, batch normalization layer, and relu layer, and upsamples as it passes through each block. After passing through the final layer, it generates an image with a size of $3 \times 64 \times 64$. The discriminator consists of six blocks composed of convolutional layer, batch normalization layer, and leaky relu layer, and downsamples as it passes through each block. Finally, it combines with the text features and uses sigmoid to determine whether it is real or fake.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

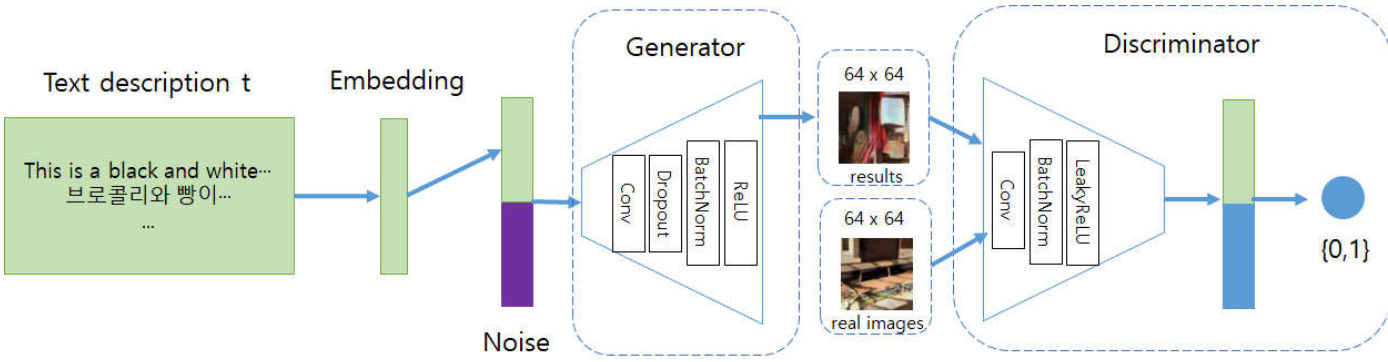


Figure 3. Architecture of Text-to-Image model using DCGAN

The loss function of Text-to-Image using DCGAN is given in Equation 1, where G represents the generator function and D represents the discriminator function, x is the input vector, z is the random vector, and φ(t) represents the feature vector of the input text. The discriminator is trained to have a larger value for Equation 1 to distinguish between real and fake data, while the generator is trained to have a smaller value for Equation 1 to generate fake data that is similar to real data. The discriminator of the DCGAN used in this study receives three types of data as inputs and is trained to classify real image-right text as real and real image-wrong text and fake image-real text as fake. To prevent overfitting, label smoothing and feature matching were used in the experiments, and the loss function is given as

$$\min_{D}\max_{G} \mathrm{E}(\mathrm{D},\mathrm{G}) = \mathrm{E}_{(\mathrm{x},\mathrm{t})\sim \mathrm{p}_{\mathrm{data}}(\mathrm{x},\mathrm{t})}\left[\log \mathrm{D}\left(\mathrm{x},\varphi(\mathrm{t})\right)\right] + \mathrm{E}_{\mathrm{z}\sim \mathrm{p}_{\mathrm{z}}(\mathrm{z}),t\sim \mathrm{p}_{t}(t)}\left[\log\ 1-\mathrm{D}\left(\mathrm{G}\left(\mathrm{z},\varphi(\mathrm{t})\right)\right)\right] + \mathrm{E}_{\mathrm{x}\sim \mathrm{p}_{x}(x),\hat{t}\sim \mathrm{p}_{t}(t)}\left[\log\ 1-\mathrm{D}\left(\mathrm{x},\left(\mathrm{z},\varphi(\hat{t})\right)\right)\right] \quad (1)$$

4. Experiment

4.1 Datasets

In this section, a webtoon dataset for training a multilingual text-to-image model is constructed using the benchmark dataset. To create a multilingual webtoon dataset, the MSCOCO dataset released by Microsoft and the Korean MSCOCO dataset translated and released by AI Hub were used, and the CartoonGAN was used to transform them into cartoon images as mentioned in Section 3.1. The original MSCOCO dataset consists of 82,783 training images and 40,504 validation images, each with five English descriptions. The Korean MSCOCO dataset was added to this to create a total of 1,232,870 pairs of multilingual webtoon data, with each image having ten descriptions (five in English and five in Korean). The CartoonGAN can transform images into four different styles, so a total of 1,232,870 pairs of webtoon data were constructed by pairing one image with one treatment. Figure 4 shows examples of multilingual treatment-webtoon datasets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

Train image	Caption
	A tennis player in action on the court. a male in a white shirt is playing tennis a man that is playing tennis on a court A male tennis player hits the ball on a grass court. A man in motion hitting a tennis ball with a tennis racket on a tennis court. 테니스 선수가 코트에서 뛰고 있다. 흰 셔츠를 입은 남자가 테니스를 치고 있다. 코트에서 테니스 치는 남자 한 남자 테니스 선수가 잔디 코트에서 공을 친다. 테니스 코트에서 테니스 라켓을 가지고 테니스 공을 치며 움직이는 남자
	Close-up of bins of food that include broccoli and bread. A meal is presented in brightly colored plastic trays. there are containers filled with different kinds of foods Colorful dishes holding meat, vegetables, fruit, and bread. A bunch of trays that have different food. 브로콜리와 빵이 들어 있는 음식이 담긴 통 뚜껑 식사는 밝은 색상의 플라스틱 쟁반에 제공된다. 다른 종류의 음식들로 가득 찬 용기들이 있다. 고기와 야채, 과일, 그리고 빵을 담아 가는 화려한 요리들 다른 음식이 들어 있는 쟁반들.

Figure 4. Examples of multilingual treatment-webtoon datasets

4.2 Webtoon generation using DCGAN

To generate webtoons by inputting multilingual text into a deep learning-based text-to-image model, the multilingual webtoon dataset was trained using a DCGAN model. As training a DCGAN with real-time pre-trained multilingual BERT to extract feature vectors for multilingual text takes a longer time, the feature vectors were extracted and stored in advance, which were then loaded and used during training. The experiment was conducted using 820,752 multilingual text-webtoon data pairs for one style, excluding incomplete data. The validation data consisted of 405,040 data points, which were divided into validation and test data sets in an 8:2 ratio. The size of the images used in the experiment was $3 \times 64 \times 64$, the noise dimension was 100, the batch size was 32, and the learning rate was set to 0.0002 for both the generator and the discriminator, and the optimizer used was Adam.

Figure 5 shows the images generated by the generator when the validation data text is input to it as the number of training epochs increases. As shown in Figure 5 (a), when the number of training epochs is 1, the generated images are vague. As the training progresses, the generator begins to draw the shape of objects, as shown in Figure 5 (e) up to 100 epochs. However, it can be observed that the generator is unable to generate any further images after a certain point. Therefore, it can be concluded from our experiments that the model that has been trained for 75 epochs is suitable for the webtoon dataset constructed in this study, based on the loss of the generator and discriminator, as well as the images output using the validation data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

Figure 5. Image generation using validation data according to epoch

Figures 6 to 8 show images generated by inputting test data into the generator model trained for 75 epochs. Figure 5 and 6 show webtoons generated when Korean and English treatments were input, respectively, and Figure 7 shows the webtoons generated when treatments of the same meaning in Korean and English were input. Although the similarity with the original images is not high, it can be observed that the generator is still able to produce some shapes that are expressed in the text. When analyzing objects, it can be seen from the experimental results that the generator does not express well for objects such as humans and animals, but relatively well for objects that are things such as chairs. By looking at Figure 7, it can be observed that although the generated webtoons may be different from each other, the semantic shapes of the words expressed in the treatments are similarly represented.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

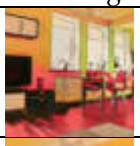
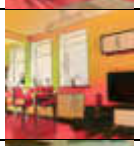
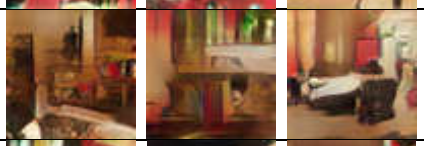
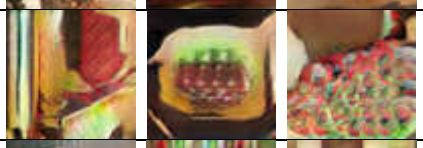
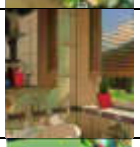
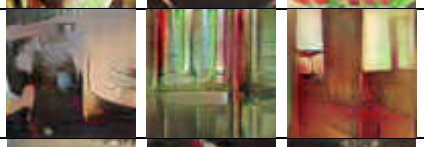
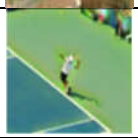
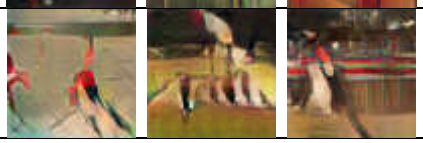
Caption	Real image	Fake image
의자와 테이블과 여자가 있는 방		
텔레비전과 테이블이 있는 생활 공간		
치즈 브로콜리와 닭고기가 들어 있는 흰색 접시		
거울 아래에 있는 욕실 싱크대		
한 남자가 경기에서 테니스 공을 서브하기 위해 준비한다		

Figure 6. Webtoon generated from Korean treatment input

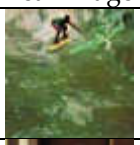
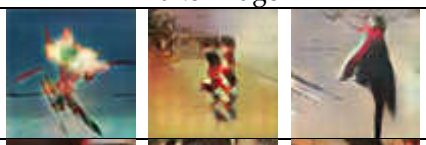
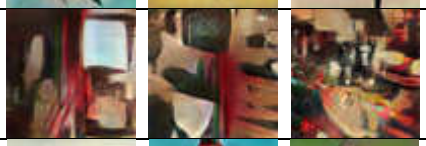
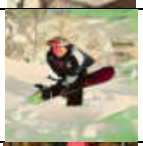
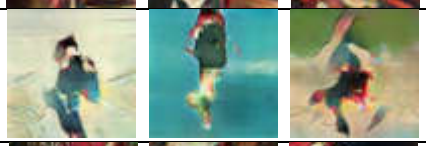
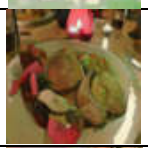
Caption	Real image	Fake image
youth_surfing_on_a_body_of_water_outside		
This_is_a_black_and_white_picture_of_a_chester_bench		
A smiling person holding a snowboard standing on a snow covered hill		
A turkey dinner shows corn, peas, mashed potatoes and biscuits all on one plate		
A business called Ray's Tavern with motorcycles sitting outside it		

Figure 7. Webtoon generated from English treatment input

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2023 doi:10.20944/preprints202304.0935.v1

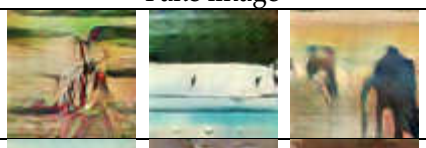
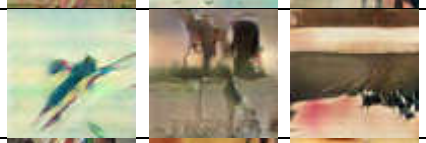
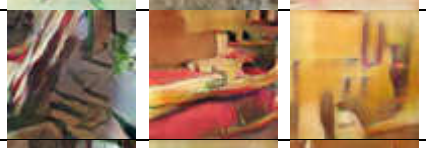
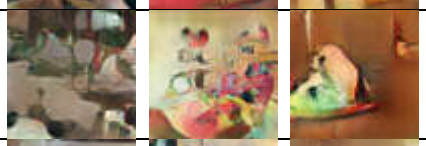
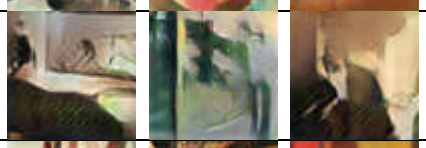
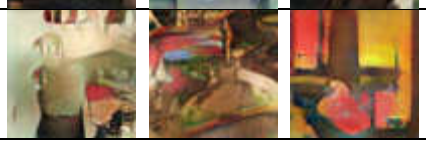
Caption	Real image	Fake image
한 남자가 거품이 이는 바다에서 서핑을 하고 놀았다		
A man surfs and plays in the foamy ocean		
두툼한 빵 위에 야채와 치즈를 얹은 샌드위치		
A plate of food with bread, grape tomatoes, cheese, cucumbers and sauce on it		
밝은 색과 초록색 그리고 흰색 커튼이 있는 방에 있는 두개의 침대		
Two beds in a room with a light and green and white curtain		

Figure 8. Webtoon generated from multilingual treatment input with the same meaning

Table 1 summarizes the proposed multilingual text-to-image model on the test dataset in terms of the Inception score and Frechet Inception Distance (FID) score. The Inception Score evaluates the quality and diversity of the generated images, and the FID score compares the mean and covariance values of the feature values of the real and generated images. The Inception Score and FID score of the multilingual text-to-image model are 4.992 and 22.212, respectively. The Inception Score of the DCGAN model trained on the MSCOCO dataset is 7.88 [4]. The lower Inception Score of the DCGAN model trained on the multilingual webtoon dataset is expected due to the distortion of the represented shapes in the images during the cartoonization process.

Table 1. Performance evaluation of the proposed Multilingual Text-to-Image Model.

Dataset	Inception score	FID score
Multilingual Webtoon	4.992	22.212

5. Conclusions

In this study, we propose a multilingual text-to-image model that can generate webtoons when given multilingual inputs. To construct a multilingual webtoon dataset, we used the Korean MSCOCO dataset from AI Hub and transformed it into webtoon images using CartoonGAN. The resulting webtoon dataset consists of Korean and English treatments in four different art styles, and we constructed a total of 1,232,879 pairs of multilingual text-webtoon data. We used a GAN-based DCGAN model for the text-to-image model and trained it on a dataset of 820,752 pairs from one style of the multilingual webtoon dataset. The DCGAN used for training consists of a generator and discriminator, where the generator takes the feature vector of the treatment and noise as input to generate a fake image, which is then evaluated by the discriminator to update the network weights and ultimately generate an image similar to the input webtoon image.

We validated the model by generating images using validation data and found that up to 100 epochs of training, the model can accurately represent the shapes present in the treatments. However, we observed the model's stopped learning beyond this point. Therefore, we consider the generator

trained for 75 rounds, which produced high-quality generated images, to be suitable for our multilingual webtoon dataset. When we evaluated the generated images using the test data, we obtained an Inception score of 4.992 and an FID score of 22.212. Although the generated images generally do not accurately represent the original data's shapes, we confirmed that the shapes present in the input treatments can be expressed in the generated webtoons. We observed that the DCGAN trained on the webtoon dataset performed worse than the DCGAN trained on the MSCOCO dataset. We speculate that this result is due to the distortion of the shapes in the images when the MSCOCO data was cartoonized, affecting the training. However, when we input the same multilingual sentences, we confirmed that the generated images represent the semantic shapes of the words expressed in the text similarly.

As future research, we plan to train the multilingual webtoon dataset on the latest text-to-image models to generate high-quality webtoons. This will help the webtoons artist to preview the scenes they envision for storytelling in advance, which can aid the overall storytelling process. Additionally, generating webtoons from treatments allows authors to modify and use them, leading to time and cost savings in webtoon production.

Author Contributions: K.Y. conducted the deep learning model experiments and manuscript writing for this paper. H.J. and J.K. handled the data set collection and transformation tasks. C.C. contributed to the experimental design and writing of the manuscript. P.K. supervised the development of the idea as well as overseeing the funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported as a 'Technology Commercialization Collaboration Platform Construction' project of the INNOPOLIS FOUNDATION (Project Number: 1711177250)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

Conflicts of Interest: Not applicable

References

1. Agnese, J., Herrera, J., Tao, H., & Zhu, X. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020,*10*(4), e1345
2. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., ... & Zollhöfer, M. State of the art on neural rendering. In *Computer Graphics Forum*, 2020, May., Vol. 39, No. 2, pp. 701-727
3. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. Generative adversarial text to image synthesis. In *International conference on machine learning*, 2016, June, pp. 1060-1069
4. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316-1324
5. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907-5915
6. Qiao, T., Zhang, J., Xu, D., & Tao, D. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505-1514
7. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M.Hierarchical text-conditional image generation with clip latents, 2022, *arXiv preprint arXiv:2204.06125*
8. Ho, J., Jain, A., & Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020,*33*, 6840-6851
9. Kim, G., Kwon, T., & Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426-2435
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684-10695
11. Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient estimation of word representations in vector space. 2013. *arXiv preprint arXiv:1301.3781*.

 doi:10.20944/preprints202304.0935.v1

12. Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 2021, *304*, 114135.
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, *arXiv preprint arXiv:1810.04805*.
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 2017, *30*
15. Tenney, I., Das, D., & Pavlick, E. BERT rediscovers the classical NLP pipeline. 2019, *arXiv preprint arXiv:1905.05950*.
16. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021, July, pp. 8748-8763
17. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022, *arXiv preprint arXiv:2208.12242*
18. Dhariwal, P., & Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021, *34*, 8780-8794.
19. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., ... & Norouzi, M, Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, July, pp. 1-10
20. Chen, Y., Lai, Y. K., & Liu, Y. J. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9465-9474
21. MSCOCO, Available online: https://cocodataset.org/ (accessed on 1 January 2023)
22. AI hub, Available online: https://aihub.or.kr/ (accessed on 1 January 2023)
23. Pires, T., Schlinger, E., & Garrette, D. How multilingual is multilingual BERT?,2019, *arXiv preprint arXiv:1906.01502*.