International Journal of Computer Applications Technology and Research
Volume 9–Issue 11, 287-294, 2020, ISSN:-2319–8656

Image Modification using Text with GANs

Fenil Doshi
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Parth Doshi
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Jimit Gandhi
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Khushmann Dwivedi
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Dr. Ramchandra Mangrulkar
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Abstract: This paper works towards finding an effective solution for the task of image feature manipulation using natural language commands. The authors aim to modify the relevant features of an image using a natural language description of the target image, such that the irrelevant features are not modified. Majority of the research in this domain focuses on generating completely new images using natural language description, and the few methodologies which attempt manipulation of existing images super in a number of aspects such as modification of irrelevant features or the aesthetic quality of the generated image. The authors propose an architecture that combines the best components of existing techniques to create an effective system to solve the stated task. The proposed architecture generates images at a high resolution to maintain the aesthetic quality of the image and ensures that the irrelevant content of the original image is not affected. The authors present a qualitative and quantitative analysis of the system as compared to the existing baselines and demonstrate the system for a relevant application such as virtual trial of clothes.

Keywords: Language-Based Image Editing (LBIE), Generative Adversarial Networks, Text Adaptive Discriminator, Feature Wise Linear Transformation, Bilinear Residual Layer

1. INTRODUCTION

Images have become an inseparable part of everyone's lives. With the widespread usage of images, the need for instant manipulation of images based on user requirements has grown at the same time. There are several existing solutions, but these tools are highly advanced and not easy to use for an amateur. For example, if a person wishes to see how they would look in a particular set of clothes, they should not have to use highly complicated and advanced tools. In such a scenario, being able to use textual descriptions or natural language commands would be the easiest way out.

The authors of this paper propose a model to solve this problem by combining features from several different architectures and applying it on a very common application such as the virtual trial of clothes. The focus of this paper is to manipulate various characteristics of an existing image using textual descriptions through the use of generative models.

The rest of the paper is organized as follows: Section 2 gives a brief summary of all the work undertaken relevant to the problem statement, Section 3 proposes a novel approach to tackle the problem and Section 4 explains the implementation of the proposed model. This is followed by Section 5 that discusses the results obtained. Finally, Section 6 gives a possible Future Scope and Section 7 provides a conclusion for the paper.

2. LITERATURE REVIEW

Majority of the existing research closely related to the task focuses on the creation of new images based on a textual description. The recent push in research in this domain can be attributed to the success of generative architectures such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE).

Goodfellow et al. [4] were the first ones to introduce the idea of Generative Adversarial Networks (GANs). Previous work in the field of creating images by estimating the probabilistic distribution of data included using Variational Autoencoders (VAE). They trained two processes simultaneously with opposing loss functions (adversarial loss)- (i) A Generator G that captures the distribution of data and (ii) A Discriminator D that tries to distinguish between generated and data samples. Another variant of GANs - Conditional GANs was proposed by Mirza et al. [11]. They introduced a new model - a conditional version of GANs - where the Generator is conditioned on some external feature to generate the sample.

GANs and Conditional GANs have been used in several architectures for the generation of images from random noise vectors (latent space). These architectures have focused on generating entirely new images from random noise vectors conditioned on some variable. For example, Riviere et al. [14] focus on the creation of entirely original images which are inspired by an image which is fed as the conditioning variable, while Wang et al. [16] propose an architecture for the generation of photo-realistic high resolution images from semantic label maps using Conditional GANs. While these architectures provide an interesting insight into how conditional GANs can be effectively used for generation of new images based on some condition, these do not focus on manipulation of existing images based on user interaction.

Zhu et al. [22] present an interesting solution, in which users can manipulate images using sketching tools normally available in painting applications. However, it uses manipulation of latent space vectors, and the results are not always predictable or generated as per the user's intention. The approach suggested by Lample et al. [8] reconstructed images by extracting the information of the values of attributes of the image directly in the latent space. Instead of using natural language, they used sliding knobs to modify specific attributes of the image.

He et al. [6] present another architecture involving manipulation of the original images using latent space vectors. It proposes an architecture where an attribute classification constraint is applied directly on the generated image to check whether the desired features have been modified or not, while style controllers are used to control the amount of attribute editing. It is an improvement over the Fader Networks proposed in [8], but does not take into account any textual description of the image, which leads us to look at architectures related to text-to-image synthesis.

One of the seminal papers in this domain, Reed et al. [13] propose an architecture for generation of low-resolution images from a descriptive image caption. The need for high resolution images is satisfied by stacking multiple GANs, as proposed in [19][20]. Spatial attention is introduced in this task by Xu et al. in the approach proposed in [18], which uses an attention module to automatically select word level conditions for generating different parts of the image. It also uses multiple generators to generate images at different resolutions and scales. However, these architectures cannot be used directly for manipulation of existing images, as these are focused on text-to-image synthesis.

The problem statement mainly focuses on this architecture where the conditioning parameter is a natural language description on which the image needs to be modified. Over the years, there have been many proposed models that particularly focus on this type of problem. This conditioning parameter can either control the latent space (Latent Space GANs) or the generator network in GANs (Conditional GANs). This paper looks into the latter architecture in more detail.

Shinagawa et al. [15] manipulated the latent space vector of the original image using the embedded vector of the natural language command. They constructed a neural network that handled image vectors in latent space to transform the source vector to the target vector by using the vector of instruction.

Dong et al. [2] experimented with a standard conditional GAN architecture where the image was encoded with encoder network, then fused together with representations of text from a text encoder network by concatenating both representations, followed by a decoder network that generated the required image. A simple sentence-level discriminator would provide feedback to the generator about the correctness of the generated image.

Nam et al. [12] used a novel approach of TAGAN (Text adaptive generative adversarial networks) and modified the discriminator architecture in order to make it more adaptive to the editing text. Previously, the works focused on single sentence-level discriminators. In TAGAN, these sentence-level discriminator is split into multiple local word-level discriminator which are then aggregated with text attention. This ensures fine-grained training feedback to the generator which then only modifies text-relevant content of the image. The experimental results of this architecture gave state of the art performance and outperformed existing methods.

Concatenation of features, as done in [2], was not the most efficient way to fuse together the image and text representations. Hence, Gunel et al. [5] used a FiLMedGAN architecture where the generator used Feature-wise linear transformations in order to combine the image and text representations. This significantly reduced the parameter space without any loss in the accuracy. Moreover, Mao et al. [10] introduced a new fusing module - BRL layers (Bilinear Residual Layers) to provide richer representations than linear models by learning second order interaction. The experimental results show that these models outperform the aforementioned models when the editing required is much more complex.

Language-based Image editing (LBIE) has been used in a various number of applications where the architectures are tweaked catering to the specific application in context.

Zhou et al. [21] used Conditional GAN architecture to modify a person's pose and other visual attributes using a natural language description. The architecture consists of two systems, namely a pose inference system to infer the pose that text refers and an image generation network that transfers the pose and attributes from text to the input image to output the required image.

G□nel et al. [5] used the FiLMedGAN architecture in order to edit the out of the person in the input image based on textual descriptions. This has various applications in the fashion industry.

El-Nouby et al. [3] extends the work done on Conditional GANs by presenting a recurrent image generation model which takes the generated image up to the current step and the natural language based instruction into account for the generation of a new image. This presents an architecture for iterative editing, which is based on conversational dialogue between the user and the system.

3. PROPOSED MODEL

The authors of this paper propose a new architecture of GANs where the fusing layer in Generator is BRL layer as discussed in [10] and the Discriminator is Text Adaptive Discriminator as proposed in [2]. Both of these techniques are tried independently but have never been examined simultaneously.

For the task, the authors propose the following model - Let input be <x, t, t'>where <x>represents the input image to be modified, <t>represents the positive text that correctly describes the image and <t'>represents negative text according to which <x>has to be manipulated to produce image <x'>where <t'>is a positive text for image <x'>. The generator objective is to produce image <x'>whereas the discriminator objective is to discriminate between <x> and <x'> (Output 1 for <x>i.e. Real image from dataset and 0 for <x'> i.e. generate fake image not from dataset.

Here,

$$< x'>= G (< x >, < t'>) = dec (brl (enc (< x >), w (< t' >))) \quad (1)$$

where G = Generator function, dec = Decoder Network, brl = Bilinear Residual Layer for fusing image and text representation, enc = Encoder network for Image <x>, w = representation of text <t'>

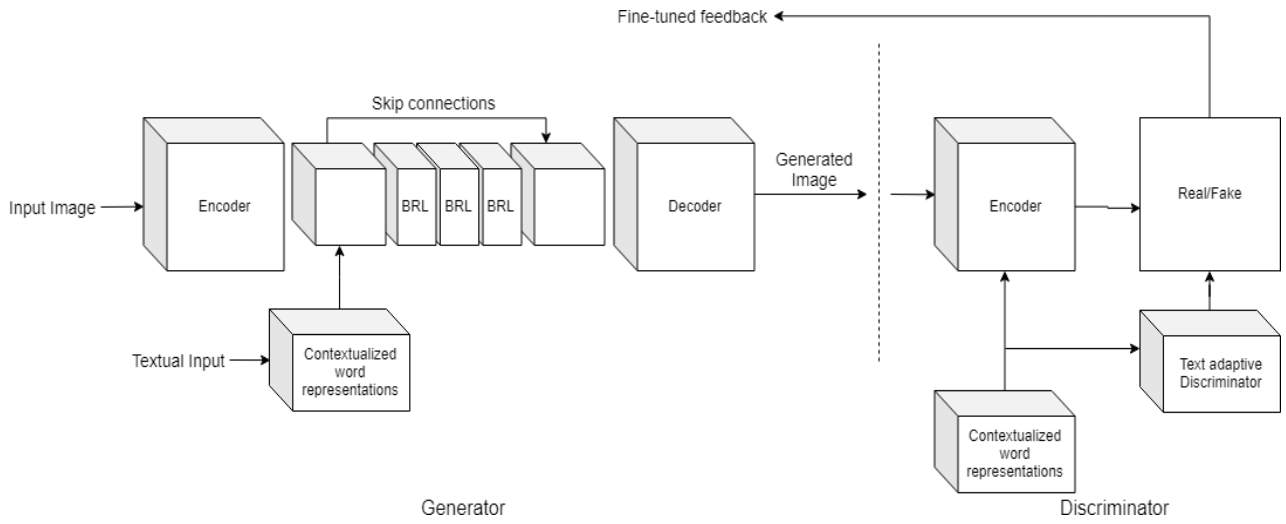

Fig. 1. Proposed Architecture

The generator loss function is given by –

$$LG = [\log(D(x)) + \lambda_1 * \log(D(G(x, t'), t'))] + \lambda_2 * L_{rec} \quad (2)$$

where Lrec = Recreational loss to preserve the text-irrelevant contents of the original image, λ_1 = weight to control the importance assigned to the discriminator's ability to correctly identify a true image and description pair, λ_2 = weight of recreational loss and D = Discriminator network as defined below.

At the discriminator end, the discriminator takes in the text and the image and provides feedback to the generator. This is done by considering output from many word-level local discriminators for each visual attribute. The text-adaptive discriminator takes in input as <x, t, t'>and tries to produce 1(real) for D (<x, t>) and D (<x>). Similarly, it tries to output 0 (fake) for D (<x, t'>) and D (G (<x, t'>)). That is, the discriminator tries to classify the generated image as well as original image on negative text as fake samples whereas original image on positive text as positive sample.

Thus, the discriminator objective is to minimize the following loss function –

$$LD = \log(D(x)) + \lambda_1 * (\log(D(x, t) + \log(1 - D(x, t')))) + \log(1 - D(G(x, t'))) \quad (3)$$

Hence, the proposed GAN structure adversarially tries to minimize both these losses to produce the output image <x'>which is conditioned on the given text <t'>.

4. IMPLEMENTATION

Algorithms and Methods Used

Fasttext Embeddings[1][7]: The user provided text is represented in a mathematical format using Word Embeddings. The authors used pretrained Fasttext embeddings for this part. Each word is represented as a vector in a 300- dimensional space. These embeddings use character level information while training and hence, can handle rare words more efficiently.
One of the major drawbacks of these embeddings is the high memory requirement as it breaks down the text into n-grams to incorporate words that are not seen while training.

Bilinear Residual Layer (BRL)[10]: The conditioning of two features in a generative network is generally accomplished by concatenation of both the feature vectors or using FiLM (Feature-wise Linear Modulation). This could have been easily done with concatenating the image encodings and word embeddings.

However, the interaction that follows between the two vectors in not deep and complex interactions cannot emerge with

simple concatenation. This is due to the fact that the concatenation or FiLM is a simple linear transformation of the features over the conditioning features.

The authors apply a more complex bilinear transformation using BRL which learns second order interaction and provides richer representation.

Using BRL, the output feature will be given as –

$$I_o = I_f * W * I_c \quad (4)$$

where I_o = Output feature, I_f = the feature that is being conditioned (here, image representation), I_c = the conditioning feature (here, text representation) and W = learnable weight matrix.

The authors chose the weight matrix to be of low rank (Hence, low rank bilinear residual) and decompose it into two matrices of that lower rank in order to relax the computations.

Generative Adversarial Networks (GANs) [4]: Generative Adversarial Networks comprises two networks - Generator and Discriminator. Both are trained on opposite tasks with opposing

loss functions (Hence, Adversarial). Generators try to approximate (mimic) the probability distribution of input space (from the dataset) and Discriminator tries to distinguish between the two distributions (approximated and real distribution).

The authors use a variant of GANs known as conditional GANs where the generation of data samples is conditioned on some external feature. The data samples are images and the conditioning feature is a natural description of the required image.

The generator and discriminator architectures need to be modified accordingly as discussed above while formulating the loss functions.

Generators - The Generator network consists of Encoder network, Fusing Network and Decoder Network. The Encoder network is a series of Convolutions, Pooling and Batch Normalization layers. The Fusing Network uses Residual layers and BRL layers (as discussed above). The Decoder Network takes in the fused vector and upsamples it using Transposed Convolutions and unpooling layers along with Batch Normalization.

Discriminators - The Discriminator is made up of several local discriminators which are created using Recurrent neural networks with Gated Recurrent Unit (GRU) cells and then attention is applied over them. More details will be followed in the next section.

Both the Generators and Discriminator network are trained using ReLU activation functions with dropout regularization to avoid overfitting. The loss functions of the networks are mentioned above.

Text Adaptive Discriminators [6]: In order to consider the effect of each word on the text, the idea is to have word-level discriminators. The generator will then receive a combined feedback from each `N' discriminator where `N' is the number of words in text.

The final classification decision to distinguish between whether the image is fake (from generator) or real (from dataset) comes via taking attention on the output vector of each of the local discriminators. This would ensure to give relatively less importance to insignificant words such as `the', `under', etc. and more importance to the words that de ne the change such as `red blouse', `green shirt', `blue jacket', etc. After considering the weighted sum using attention, the final output is then calculated.

4.2 Experimentation and Training

Datasets and Creating Training Samples: The training was performed on two datasets –

Caltech-UCSD Birds 200 [17]: It consists of 11,788 images of birds in their natural habitat along with their captions. Each of the images is associated with 10 different captions that de ne the image. The authors select any of these captions randomly.
The total images are divided into 200 classes. The authors aim to modify the body and color of the birds based on the textual description.

DeepFashion: Fashion Image Synthesis Dataset [9] [23]: There are a total 78,979 images of fashion outfits along with The training set pairs were created by –

1. The image and its caption formed the pair <x, t>which was fed to the Discriminator in order to output 1 (Real).

2. The image and a random caption from some other image was chosen to form the pair <x, t'>. The true label for Discriminator for such a pair would be
0 (Fake). The Generator was fed the pair <x, t'>to create a new fake image <x'> (Generated Fake Sample conditioned on text <t'>). Additionally, <x'>was compared with <x>for computing Reconstruction loss in order to preserve the background.

3. The new pair of <x', t'>was also fed to Discriminator and Discriminator was trained to output 0 (Fake) for such an input pair. Simultaneously, Generator was trained to get an output 1 (Real) from the Discriminator.

Training Setup: After generating the training samples, Generator and Discriminator were trained consequently one after the other. While updating the weights of the Generator, the corresponding gradients of Discriminator were not updated. Similarly, the Generator was not used while training the Discriminator.

The samples were fed in the batch size of 32 to the model. The model was trained on CUB dataset for 600 epochs while the DeepFashion dataset was trained for 220 epochs because of the difference in size of datasets. It was observed that there was no considerable decrease in loss and not much improvement of the model in its performance.

The model was trained on a single core GPU (RTX 2080 Ti) with 18.3 TFLOPS with 64 GB storage RAM. The machine was rented on Vast.ai and was trained for 2.5 days for training the Birds model and for 4-5 days for training the Fashion Synthesis model.

5. RESULTS AND DISCUSSION

5.1 Analysis on CUB Dataset

To analyse the results of the model, the authors first trained the model on the CUB-200 dataset [17]. To test, images and input texts were randomly selected and first they were run on the TAGAN model proposed by Nam et al. [12]. Then, the same pairs of images and text were tested on the model and the results were collated together. Some results are visible in Fig. 2. The authors' model is referred to as Model 1 and the TAGAN Model [12] is referred to as Model 2.

Because of the absence of some uniform universal metric to analyze the efficacy of the results, the authors chose to perform comparative analysis.

In the comparative analysis, the authors showed the original image, the input text and the output image of both the TAGAN model as well as the authors' model. After observing these three

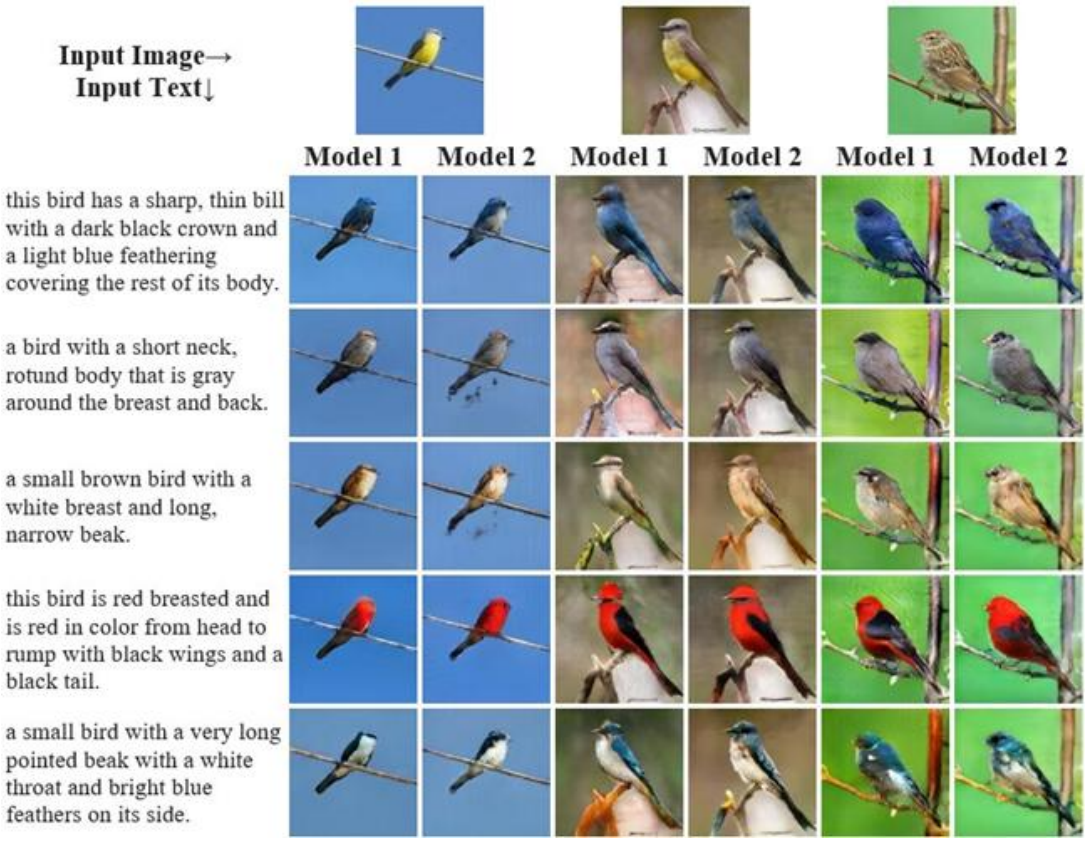

Fig. 2. Results obtained on the CUB-200 Dataset

images, the volunteers were asked to choose which model performed better in the following characteristics:

1. Which model was successfully able to edit the image according to the textual description?
2. Which model preserved the background of the image?
3. In which model is the bird distinguishable naturally?

The choices of the volunteers for each of these questions were recorded and then analyzed. Fig. 3 shows the pie chart for responses of the questions asked.

As can be seen from the pie chart, 56.5% volunteers found Model 1 to successfully edit the image according to the textual description, as opposed to 32.4% for Model 2. Further, according to 56.5% of the volunteers, the bird was naturally distinguishable in the output of Model 1, remarkably more than the 32.4% of Model 2. However, the background of the image was preserved in a more effective way by Model 2, as 46.3% chose Model 2 and only 44.4% chose Model 1. The number of responses for each question can be seen from Fig 4.

Hence, from this extensive comparative analysis, the authors conclude that their model (Model 1) outperforms Nam et. al's model [12] (Model 2) in two of the three metrics analysed by the authors. That is, their model successfully performs the task of editing the image according to the text better and the output image of the bird is more distinguishable in their model. However, Nam et. al's model still preserves the background of the image in a better way than the model proposed by the authors.

This suggests that the BRL layer does increase the interaction between the text and image which consequently leads to better and focussed editing of the image. However, it might be the case that the reconstruction error is overwhelmed by this and the model focuses much more on editing the image and less on preserving the background.

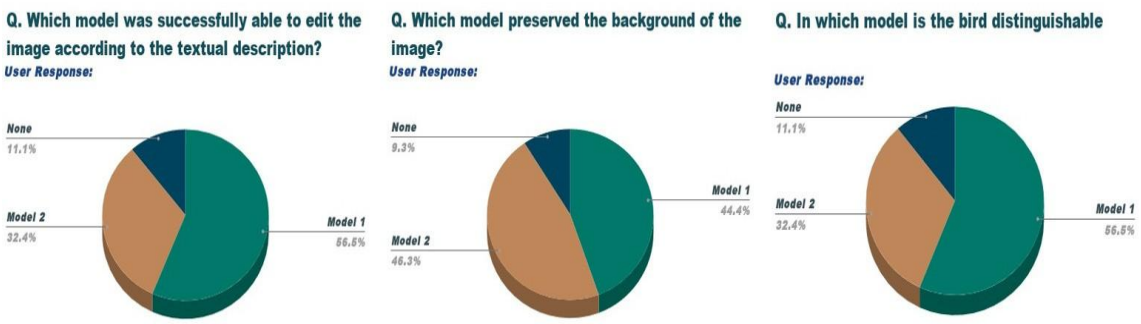

Fig. 3. Pie chart for responses to the questions

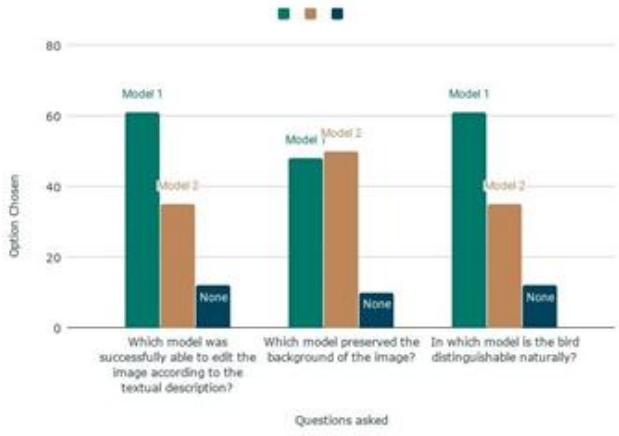

Fig. 4. Bar Graph displaying choices chosen by volunteers

5.2 Analysis on DeepFashion Dataset

After performing analysis on the CUB-200 dataset, the authors trained their model on the DeepFashion dataset [9] [23]. The model was then tested by randomly selecting images and input statements and collating the results together. Fig. 5 shows some of the results obtained.

Due to the absence of a quantitative metric to judge the accuracy of the model on the DeepFashion dataset, the author's decided to perform qualitative analysis. In this, the chosen volunteers were given the input image, input text and the output and were asked to rate the image manipulation for the following metrics:

1. Was the model successfully able to edit the input image based on the given textual description?
2. Did the model preserve the background of the image (features such as face, hair, etc.)?
3. How good is the naturalness of the image and does the image look similar to a person?

The volunteers were asked to rate the image manipulation on a scale of 1-5 where 1 signifies the most erroneous conversion whereas 5 signifies the most accurate conversion. The results were collected and can be understood by Fig 6.

Hence, the volunteers gave an average score of 3.79/5 to the model's accuracy in successfully editing the input image based on the given textual description. They gave an average score of 3.61/5 to the model for preserving the background of the image and an average score of 3.03 to the model for maintaining the naturalness of the image.

Fig. 5. Results obtained on the DeepFashion Dataset

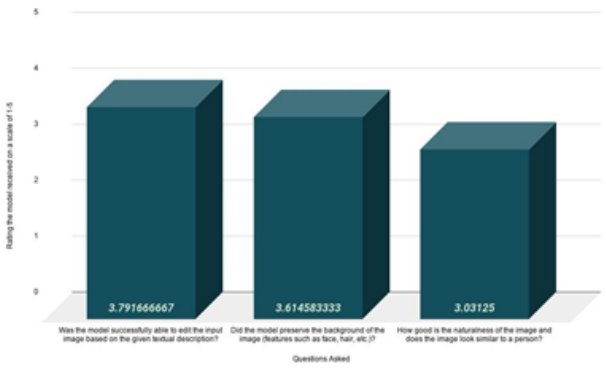

Fig. 6. Average results for the questions asked on the DeepFashion Dataset

Hence, while the model performs considerably well in performing the task of image manipulation based on the input text, it sometimes does not preserve the features of the women in the images. The hair, the accessories like watches, hats, sunglasses, etc. sometimes get altered or disappear completely. Another area where the model lacks slightly is preserving the naturalness of the image. Sometimes the face in the output image fails to be naturally identified as a human face. However, given the complexity of the human body and human faces, the model performs considerably well in retaining the overall posture of the body and effectively performs manipulation on the dresses.

6. FUTURE SCOPE

There are many directions that can be taken for furthering this work. One direction is the trial implementation of a prototype in a real shop, thus testing the system against real-world noisy data. Another direction is to improve the size and quality of images generated through the use of deeper neural networks, requiring massive computational resources.

7. CONCLUSION

Thus the architecture combines the best features of existing architectures and can be used in the fashion domain for virtual trial of clothes. The architecture overcomes the drawbacks of the existing approaches of image manipulation and makes it easier for an amateur to use textual descriptions to convert an existing image into the desired version.

The results show that the model generates better results than the baseline models that the authors have based it on. The model

combines the best features of different models and allows high-quality results to be generated with accurate modifications as per the user requirements.

8. References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135-146 (2017).

   https://doi.org/10.1162/tacl a 00051, https://www.aclweb.org/anthology/Q17-1010

2. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5706-5714 (2017)

3. El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L.E., Kahou, S.E., Bengio, Y., Taylor, G.W.: Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10304-10312 (2019)

4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672-2680 (2014)

5. G□unel, M., Erdem, E., Erdem, A.: Language guided fashion image manipulation with feature-wise transformations. arXiv preprint arXiv:1808.04000 (2018)

6. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing 28(11), 5464-5478 (2019)

7. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for e cient text classication. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427-431. Association for Computational Linguistics, Valencia, Spain (Apr 2017),

   https://www.aclweb.org/anthology/E17-2068

8. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: Manipulating images by sliding attributes. In: Advances in neural information processing systems. pp. 5967-5976 (2017)

9. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

10. Mao, X., Chen, Y., Li, Y., Xiong, T., He, Y., Xue, H.: Bilinear representation for language-based image editing using conditional generative adversarial networks.

    In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2047-2051. IEEE (2019)

11. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

12. Nam, S., Kim, Y., Kim, S.J.: Text-adaptive generative adversarial networks: Manipulating images with natural language. In: Advances in neural information processing systems. pp. 42-51 (2018)

13. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)

14. Riviere, M., Teytaud, O., Rapin, J., LeCun, Y., Couprie, C.: Inspirational adversarial image generation. arXiv preprint arXiv:1906.11661 (2019)

15. Shinagawa, S., Yoshino, K., Sakti, S., Suzuki, Y., Nakamura, S.: Interactive image manipulation with natural language instruction commands. arXiv preprint arXiv:1802.08645 (2018)

16. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: Highresolution image synthesis and semantic manipulation with conditional gans. In:Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798-8807 (2018)

17. Welinder, P., Branson, S., Mita, T., Wah, C., Schro, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)

18. 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. pp. 1316-1324 (2018)

19. 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. pp. 5907-5915 (2017)

20. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence 41(8), 1947-1962 (2018)

21. Zhou, X., Huang, S., Li, B., Li, Y., Li, J., Zhang, Z.: Text guided person image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3663-3672 (2019)

22. Zhu, J.Y., Krahenbuhl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European conference on computer vision.

    pp. 597-613. Springer (2016)

23. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: Fashion synthesis with structural coherence. In: International Conference on Compute

    Vision (ICCV) (October 2017)