Generating Image from Text

Miss. Asmita Krishna Kadam[1] Miss. Titiksha Dhanpal Ingale[2] Miss. Prajakta Rupesh Kalyani[3] Miss. Snehal Mohan Chougule[4] Dr. Prof. D. A. Nikam[5]
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Dr. J. J. Magdum College of Engineering, Jaysingpur, India

Abstract— Generating photo-realistic images from text is an important problem and has tremendous applications, including photo-editing, computer-aided design, etc[5]. Recently, Generative Adversarial Networks (GAN) has shown promising results in synthesizing real-world images.[3] Conditioned on given text descriptions, conditional GANs are able to generate images that are highly related to the text meanings. Synthesizing images from text descriptions is a challenging problem in computer vision and has many practical applications. Although Generative Adversarial Networks (GANs) have shown remarkable success in various tasks, they have modelled image-generation of general objects in everyday-life[3].

Keywords: Sentiment Analysis, Classification, Machine Learning

I. INTRODUCTION

A. *Introduction*

In this project, we aim to generate an image depending on the subject mentioned in the text input using GANs. In order to provide more realistic, accurate and detailed images, the project focuses on the specific domain of face images.[3] The system takes text input describing various features of face like shape of face, hair color etc and generates corresponding image.

This provides a wider range of applications including detailed visual-aids in education, in research for visualization of un-captured conditions, in various Machine Learning modules, computer-aided design, etc. Photographic text-to-face synthesis is a mainstream problem with potential applications in image editing, video games, or for accessibility.[2] We are motivated by the potential of automated face generation to impact and assist critical tasks such as criminal face reconstruction.

The task can be addressed as learning a mapping from a semantic text space describing the facial features[2] e.g., "Pointy Nose" and "Wavy hair" to the RGB pixel space.. The input to the system will be description of facial features like:

- Structure of face - round face, ovalface, double chin, etc
- Hairstyle Bangs, wavy hair, bald, brown hair, straight hair, etc
- Facial Hairstyle - Goatee, Moustache, Sideburns, Beard, etc.
- Attributes that enhance the appearance - Attractive, Smiling, etc.

The text will be the input for GAN to generate the image as output according to the description provided about features of face. Generating an image from a given text description has two goals: visual realism and semantic consistency. Although significant progress has been made in generating high-quality and visually realistic images using generative adversarial networks, guaranteeing semantic consistency between the text description and visual content remains very challenging[2].

B. *Challenges*

Face challenges in generating detailed images of domain-specific entities. In this project, we aim to generate an image depending on the on the subject mentioned in the text input using GANs. In order to provide more realistic, accurate and detailed images, the project focuses on the specific domain of face images. The system takes text input describing various features of face like shape of face, hair color etc. and generates corresponding image. This provides a wider range of applications including detailed visual-aids in education, in research for visualization of un-captured conditions, in various Machine Learning modules, computer-aided design, etc.

II. NECESSITY OF WORK

The objective of the project is:

- To synthesize images from text description[5].
- To learn the generality and discriminative power of text description.
- To combine the advances in both the fields by making a model through the formulation of GAN.
- To learn mapping from words and characters to image pixels such that generated images can be mistaken as real by human which looks like photographs or photographic realistic image.

III. DIAGRAM

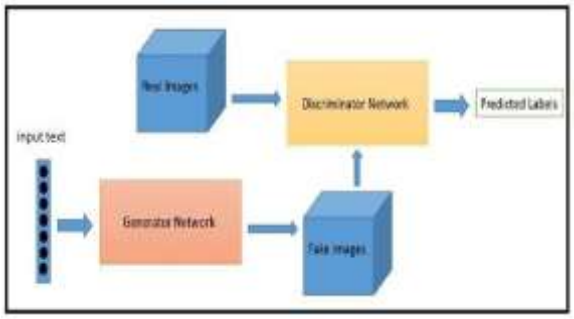

Fig. 1: Architecture of system

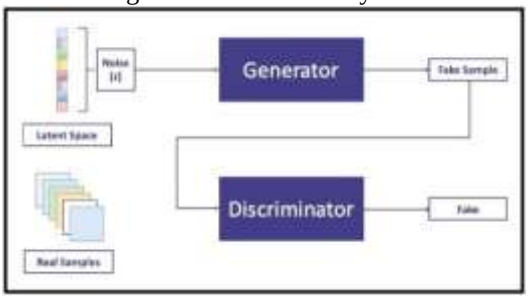

Fig. 2: Formulation between Generator and Discriminator

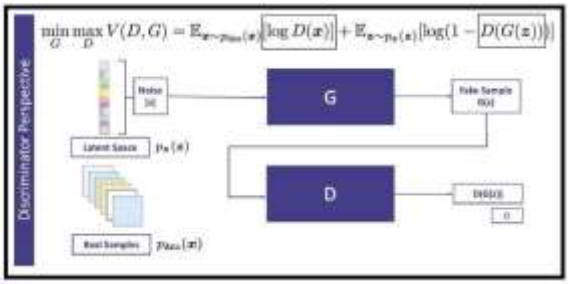

Fig. 3: Discriminator Perspective

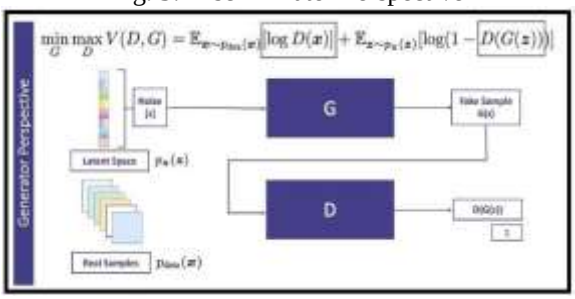

Fig. 4: Generator perspective

Main modules of our system are:-

- Generator Network
- Discriminator Network

A. Fig 2. Formulation between Generator and Discriminator

The beauty of this formulation is the adversarial nature between Generator and Discriminator. Discriminator wants to do its job in best possible way, when a fake sample [which are generated by Generator] is given to a Discriminator, it wants to call it out as fake but the Generator wants to generate samples in a way so that the Discriminator makes a mistake in calling it out as a real one. In some sense, the Generator is trying to fool the Discriminator. Let us have a quick look at the objective function and how the optimization does is done. It's a min-max optimization formulation where Generator wants to minimize the objective function whereas Discriminator wants to maximize the same objective function.

B. Fig 3. Discriminator Perspective

Discriminator wants to drive the likelihood of D(G(z)) to 0. Hence it wants to maximize (1-D(G(z))) whereas the Generator wants to force the likelihood of D(G(z)) to 1 so that Discriminator makes a mistake in calling out generated sample as real. Hence Generator wants to minimize (1-D(G(z)).

For Discriminator, instead of having only image as input, a pair of image and text embeddings are sent as input. Output signals are either 0 or 1. Earlier Discriminator's responsibility was just to predict whether given image is real or fake.

Now, Discriminator has one more additional responsibility. Along with identifying the given image is read or fake, it also predicts the likelihood of whether the given image and text aligned with each other. This formulation force Generator to not only generate images which looks real but also to generate images which are aligned with input textual description.

C. Fig 4. Generator perspective

This min-max formulation of the objective function has a global optimum when data distribution and model distribution is same which means if the optimization function converges to global minimum then the model had learnt the underlying data distribution present in input training dataset.

IV. ADVANTAGES

GANs are an unsupervised learning method: Acquiring labeled data is a manual process that takes a lot of time. GANs don't require labeled data; they can be trained using unlabeled data as they learn the internal representations of the data. GANs generate data: One of the best things about GANs is that they generate data that is similar to real data. Because of this, they have many different uses in the real world. They can generate images, text, audio, and video that is indistinguishable from real data. Images generated by GANs have applications in marketing, e-commerce, games, advertisements, and many other industries.

V. METHODOLOGY

Methods the whole work of this dissertation is categorized into three major parts or stages as discussed below:

- Pre-processing Stage
- Processing Stage
- Post Processing Stage

Pre-processing stage eliminates challenges created by noise, blurring effect and uneven lighting which makes performing text detection, extraction and recognition that are embedded in image documents simple and better. In this stage, the image input (scanned or captured by different devices or browsed from drive) is processed to remove any noise that may affect the image during the time of acquisition or during the time of transmission. A colors (RGB) image will be converted to a gray scale image and then thresholding edge thinning, and noise removal process will be done. The image is then converted to a binary image with suitable threshold in order to simplify extraction process. Processing stage have different steps in which the image is checked whether it contains text or not, identify to locate the text area on image, and differentiating foreground and background of that image text is done. The following steps are most common:

- Text Detection: takes enhanced image as input and decides it contains text or not and identifies the text regions in an image.
- Text Localization: Merges the text regions to formulate the text objects and define the tight bounds around the text objects. Text detection, localization and tracking modules are closely related to each.
- Text Tracking: is also used to speed up the text extraction process by not applying the Binarization and recognition step to every detected object.
- Text Binarization: used to segment the text object from the background in the bounded text objects. It converts gray scale image to binary image, where text pixels and background pixels appear in two different binary levels like white text on dark background or vice versa. Binarization can also be done before the other steps.

VI. CONCLUSION

We have proposed a conditional generative adversarial network (Conditional GAN), which can generate And manipulate the generation of images based on natural language descriptions. Our Conditional GAN Can successfully disentangle different visual attributes and allow parts of the synthetic image to be msanipulated accurately, while preserving the generation of other content.

In this the presented captions for the Celeb A dataset was used to facilitate face synthesis from text. We Then used Generative Adversarial Network (GAN) to learn the conditional multimodality in synthesis of face from captions.

A. For future work:

1) Input filter for wrong inputs using NLP techniques.
2) Feature extraction for unknown features using spacey pipelines
3) Image synthesis

REFERENCES

[1] Usha Tiwari, Shivani Gupta, Nisha Basudevan and Piyush D. Shahani (2014) "Text Extraction from Images", Noida, India.
[2] N. K. Gundu, S. M. Jadhav, T. S. Kulkarni and A. S. Kumbhar (2014), "Text Extraction from Image and Displaying its Related Information", Pune, India.
[3] Tingting Qiao, Jing Zhang, Duanqing Xu and Dacheng Tao (2008), "Learn, Image and Create: Textto-Image Generation from Prior Knowledge", China.
[4] Vivek Dhanpal Sapate (2010),"A Survey: Text Extraction from Images and Video", India.
[5] C. Mishra, P. K. Swain and J. K. Mantri (2012), "Text Extraction and Recognition from Imageusing Neural Network", India.
[6] Important Links:- https://www.kaggle.com/jessicali9530/celeba- dataset https://paperswithcode.com/paper/text2facegan-face-generation-from-fine/review.

