Spa Text: Spatio-Textual Representation for controllable Image generation(Abstract)

- The text- image generation models in recent times are very functional and also getting results with high quality

- As the entire aim of text to image generation is to get the mental image of the user to digital, So right now the user holds the responsibility to input a proper prompt to get the image that resides in his mind

- For every user it is not possible to follow the prompt and even though the user followed the prompt it is very uncertain that only a text with a proper prompt might not get him the exact regions and object shapes he thought about.

- This is where SpaText comes into picture.

- The main goal of the Spa-text model is to make user inputs more expressively so that the chance or the possibility of getting a mental image of the user will increase

- In text to image generation models the input will be a prompt and image will be generated but in this proposed model the image generation model will also be expecting a segmentation map.

- The Spa-text proposes a new method for text to image generation using **Open vocabulary scene control**

- Open vocabulary scene control refers to a concept or capabilty with in computer graphics and virtual environment that allows for flexible and unconstrained control over scene generation or manipulation without pre-defined limitations on the vocabulary or scene elements that can be used.

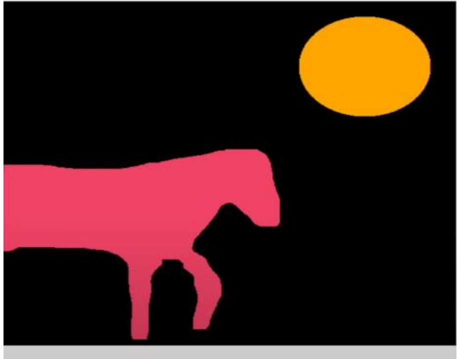
Introduction

- As the text to image is powerful tool in recent times, a single prompt is able to generate n number of possible outcomes

- How ever it is impossible to catch the regions, objects in the mental image of the user. With current state of art work

- “Make a Scene proposed to tackle this problem adding a segmentation map as additional input along with the text

- For ex: imagine a using a digital paint brush and drawing a horse and a moon in the sky



- Along with the user text prompt this will be turned into a segmentation map for the additional input.

- Which will act input more information and output more possible

- There are also some limitations that are overcame

- Rather than using using a fixed set of labels to represent it we are using a spatal free-form text which typically involves text that is entered or presented in a way that retains its spatial organization or layout.

- Rather than providing a dense segmentation map accounting for each pixel, Sparse map is proposed.

- So after providing the input like discussed, the prompt will be describing the image and spatio-textual scene will be specifying the interested regions, shapes, objects for the user.

- To train this model the data set is required, as acquiring the large data sets that contains free-form text is expensive, and that data sets might also not be available to acquire according to the paper

- So a novel CLIP based spatio-textual representation that enables a user to specify for each of their mental image

- And there are also several automatic evaluation metrics and use them to compare against baselines they adapted from existing methods.
- Several automatic evaluation metrics are performed to use them to compare against baselines we adapted from existing methods.

Related work

- Recently, we have witnessed great advances in the field of text-to-image generation. The seminal works based on RNNs and GANs
- Later, zero-shot open-domain models were achieved using transformer-based approaches
- All these methods do not tackle the problem of image generation with free-form textual scene control.
- **Local text-driven image editing**: train a designated inpainting model, whereas Blended Diffusion leverages a pretrained text-to-image model
- Combining these localized methods with a text-to-image model may enable scene-based image generation.
- This method can be compared to the method we are discussing now.

Method

- The model aim to provide the user with more fine-grained control over the generated image. In addition to a single global text prompt, the user will also provide a segmentation map, where the content of each segment of interest is described using a local free-form text prompt.

- Formally, the input consists of a global text prompt tglobal that describes the scene in general, and a H×W raw spatiotextual matrix RST.

- Our goal is to synthesize an H×W image I that complies with both the global text description global and the raw spatio-textual scene matrix RST

- As over the recent year text to image generation came into picture with high in use. Large-Scale data sets are generated by this community

- But these data sets cannot be used to train the proposed model, as they do not contain the local text descriptions for each segment in the images.

- To extract the objects in the images along with the textual description we can opt to use a pre-trained panoptic segmentation model along with a **CLIP** model.

- The clip model is used here because it is trained to emed images and text prompts into a rich shared latent space by contransitive learning on huge data

- The utilization of this shared latent space can be done by using image encoder im **CLIP**$_{img}$ to extract the local embeddings using pixels of the object in the image that should be generated in the model

- During inference time **CLIP**$_{txt}$ text encoder can be used to extract the local embeddings using text description provided by the user to form a raw spatio-textual matrix RST

- To reduce the domain gap between train and inference time these embeddings can be converted to **CLIP**$_{img}$ using prior model P. This is model is already trained to convert **CLIP** text embeddings to **CLIP** image embeddings using an image-text pair data set .

- Spatio-textual representation will be constructed

Incorporating Spatio-Textual Representation into SoTA Diffusion Models

- The diffusion model that should be discussed is **DALL-E2** (Distributed and autoregressive language learning with encoding) This model developed by open AI which consists of three diffusion models.

- **Prior-model(P)** is trained to translates the tuples(CLIP_{txt},(y), byte-pair encoding(y)) into CLIP_{img}(x) where pair(x, y) is a image text-pair

- **Decoder-model(D)** which holds the ability to convert the image into low resolution image .

- **Super-resolution-model(SR)** will be increasing the resolution of image to a higher quality possible.

- This model will be working in the flow of SR * D * P

- Concatenation To maintain spatial alignment between the noisy image (during diffusion steps) and the spatio-textual representation (ST), the noisy image (xt) and the textual representation (ST) are concatenated along the RGB channels. This means they combined into single input format that model can process

- To utilize the vast knowledge gathered during the training process, fine-tuning the decoder component (D) should be performed to the model to better adapt to specific textual descriptions and control image generation.

- At each diffusion step, the decoder refines the noisy image (xt-1) to produce a less noisy version (xt), incorporating information from the CLIPimg representation and the textual features.

Multi-Conditional Classifier-Free Guidance

- Multi-conditional classifier-free guidance is a method used in diffusion models to generate images based on different conditions (like text prompts or labels) without using explicit classifiers.

- Classifier-Free Approach: This method doesn't rely on traditional classifiers for conditional generation. Instead, it uses learned differences in generating with and without specific conditions.

- During training, each condition (e.g., text prompts or labels) is replaced with a null condition (∅). The model learns to generate samples both with and without specific conditions.
- When generating new images (inference), the model calculates the impact of each condition separately on the generated output.

- Individual guidance scales (si) are used for each condition during inference. The model combines the impacts of multiple conditions to control the direction of generation.

- Enables generating outputs based on multiple conditions simultaneously. Provides flexibility in conditional generation tasks without the need for explicit classifiers.

Experiments

- The pre trained models latent based and pixel based variants are fine tuned with 35M image text pairs, following make a scene while filtering according to the requirements

Qualitative and quantitative Comparisions

- The comparison intialy performed with two models, (1) **NTLB** (No token left behind) and also with (2) Make a scene

- NTLB proposes method that conditions text to image model on spatial locations using an optimization approach

- The raw spatio-textual matrix (RST), which contains spatial and textual information, is converted into separate masks to condition the image generation process.

- **MAS**(make a scene) is a method that conditions text to image model on global text form and dense segmentation map

- Instead of using a fixed segmentation map the local text form the spatio textual matrix are concatenated to global text prompt for conditioning

- Unlike MAS using dense segmentation map, sparse map is utilized
- Several numerical evaluations like (1) **FID** score, (2) **Global distance,** (3) **Local distance,** (4) **Local IOU** are performed
- The observation says that latent based variant out-performed the pixel based variant in all the metrics. The reason might also because of insufficient re-implementation of the DALL.E2 model .
- Still the pixel based model is also able to take into both the global text and as well as spatio-textual representation

User study

- A user study is conducted on amazon mechanical Turkish people as there are lot of aspects to be assessed.
- They have preferred this method over the baseline .

Mask sensitivity

- When conducting the experiments inaccurate alignment of the provided spatio textual representation is found.
- This inaccuracy is also found in Local instersection over union(lou) scores, which measure how well the generated image match the specific regions or shapes or masks.
- These characteristics will be advantageous when dealing with unrealistic and non accurate input mask
- One possible reason for this behavior is the downsampling of the input mask during training, where some detailed information is lost.
- Due to this loss of detail, the model learns to fill in the missing parts based on the text prompts provided during training.

Ablation study

- The ablation study systematically investigates the impact of different model components or configurations on overall performance metrics such as image quality, text matching, and feature representation. The results help identify key factors contributing to the model's effectiveness and guide improvements in the proposed method.

Limitations and conclusions

- In some cases, characteristics propagate to adjacent segments, e.g. (left), instead of a blue bowl the model generated a vase with a wooden color. In addition, the model tends to ignore tiny masks (right).

- In addition the model might not be able to notice tiny segments, or tiny descriptions in user terminology

- Fine tuning procedure might be the reason for the above limitation, when we fine tune a the model, we choose a very random number of segments that are above the size of threshold because CLIP embeddings may not be reasonable for every low resolution image.

- In conclusion this paper has explained the image to text generation with sparse scene control.

I would like to conclude saying that this paper has a potential to learn about lot of terminologies and methodologies which would fill and maintain the enthusiasm about the domain in one individual. My sincere apologies if there is any misconception of my understanding.

Thank you