Text-to-Image Generation Project Proposal

Group Name: WonderVision

Group Members:

- Xinyi Ji, 1003798713
- Chenhao Gong, 1004144598
- Chengmin Jiang, 1004199510
- Ruijia Wang, 1003803164

Short Abstract

(Write one or two sentences describing your main idea, approach and deliverable.)

Automatically generating text into the image is always an interesting topic in our daily life. DALL·E implemented by OpenAi is one of the approaches to realize this idea. We are going to reproduce this model in our project and, time permitting, attempt to improve this model in different aspects like the position of more than 3 objects, or the colour of different objects.

The approach we are going to use in our project is just like stated in the DALL·E paper, a combination of a discrete variational auto-encoder and an autoregressive transformer written in python in google colab. To be more specific, the dVAE is first used to compress the original training image in order to save the memory. We will then concatenate the text tokens with the image tokens. And the autoregressive transformer is used to model the joint distribution over the text and image tokens.

Finally, we will implement a model which will generate a consistent image if you pass a text description into the model.

Extended Abstract

(Write one or two pages describing your proposed project in more detail.)

As stated in the approach part of the short abstract, the overall procedure can be viewed as maximizing the evidence lower bound on the joint likelihood of the model distribution over images x, captions y and the tokens z for the encoded image. With the

help of the DALL·E paper, we could model this distribution using the factorization $p_{\theta,\psi}(x, y, z) = p_{\theta}(x|y, z)p_{\psi}(y, z)$, which yields the lower bound

$$ln p_{\theta,\psi}(x, y) \geq E_{z\sim q(z|x)}(ln p_{\theta}(x|y, z) - \beta D_{KL}(q_{\phi}(y, z|x), p_{\psi}(y, z)))$$

In the expression above:

- q_{ϕ}denotes the distribution over the image tokens generated by the dVAE encoder given the original RGB imagex^2;
- p_{θ}denote the distribution over the RGB images generated by the dVAE decoder given the image token;
- p_{ψ} denotes the joint distribution over the text and image tokens modelled by the transformer.

The first part of our project is to learn the visual codebook. In practice, we train a dVAE on the images alone in order to maximize the evidence lower bound with respect to φ and θ. The initial prior p_{ψ}would be set to the uniform categorical distribution over the codebook vectors and q_{ϕ} would be set to be categorical distributions parameterized by the logits at the same spatial position in the output by the encoder.

The second part of our project is to learn the prior distribution over the text and image tokens, i.e. to train the autoregressive transformer. By training the transformer, we could maximize the evidence lower bound with respect to ψ after fixing φ and θ. The transformer is a decoder-only model in which each image token can attend to all text tokens in all of its self-attention layers. There are three different kinds of self-attention masks used in the model. The part of the attention masks corresponding to the text-to-text attention is the standard causal mask, and the part for the image-to-image attention uses either a row, column, or convolutional attention mask.

The third part of our project is the sample generation part where we will rerank the samples drawn from the transformer using a pre-trained contrastive model. Ideally, when given a caption and a candidate image, the contrastive model assigns a score based on how well the image matches the caption.

Summary of Technical Aspect

(Describe in a two or three paragraphs, with some equations, the main technical ideas for your project. If you propose a new model, or an application of existing models,

summarize your approach. If it is unclear, please emphasize the relevance to our course Probabilistic Learning and Reasoning.)

We noticed that there is much acknowledgement in the DALL·E paper that will be covered in the scope of CSC412 lectures in the near future:

1. The evidence lower bound which we will use to design the loss function as stated in the extended abstract.
2. Kullback–Leibler divergence which was used in the loss function after constructing the evidence lower bound.
3. The discrete variational auto-encoder which was used to compress the original image. We heard that we would learn variational auto-encoder during the following class. Besides this, we would also learn the discrete version by ourselves. (Here, we suppose that the codebook described in stage one is the latent space representation of the auto-encoder. Hence, in the scenario that the auto-encoder fails to obtain the expected results, we can still attempt on alternatives including PCA or t-SNE algorithm. However, we shall expect worse results using these alternatives because of their linear property in comparison with VAE.)

There is also knowledgement beyond the scope of the course. We need to study individually:

1. Gumbel-softmax relaxation and log-laplace distribution which is used in the dVAE.
2. Contrastive model which is used to perform sample generation.

Note that we might need to reduce the size of our training dataset in order to be run on Google Colab, considering that the scale of the original dataset used in the DALL·E paper may be too excessive for a successful execution within the limits of performance provided by Google.

Project Deliverable Goals

(Describe concrete results and artifacts that your group will consider a successful project to deliver within the constraints of the semester. Describe concretely what experiments you plan to run, what results you plan to collect, and how you plan to communicate those results.)

We deem that our model achieves the minimum deliverable goal if it is capable of:

1) Modifying several of an object's attributes, as well as the number of times that it appears and see how the model behaves;
2) Controlling multiple objects, their attributes, and their spatial relationships.

Moreover, it would be more favorable if the model could also achieve the following aspects like the DALL·E model:

1) Visualizing perspective and three-dimensionality.
2) Visualizing internal and external structure.
3) Inferring contextual details.
4) Applications of preceding capabilities.
5) Combining unrelated concepts.

If our model fails to achieve the target, we will analyze the outliers (i.e. the one that fails) and make adjustments to our model accordingly.

Nice-to-haves

(If you have additional ideas, or ambitious goals that would be unlikely to achieve within the project timeline you can describe them here.

If your initial results are extremely promising and you have additional energy you can include these.)

After reviewing the result of the DALL·E we found that it will be less concise when we try to plot more than 3 objects. The image would also fail when there are more than 4 positions or colours. The model also has poor performance during plotting long strings. Hence, if we have more time (possibly during the summer vacation) or if our initial results are extremely promising (achieving a similar performance with the DALL.E model) we would proceed with our project and improve the model's performance in these aspects.

(If your initial results are less promising you can use these alternative directions, possible experiments, animation ideas and plots to continue or understand.)

After viewing the paper of DALL·E, we realized the difficulty for us to reproduce the whole process in such a limited available time with our existing members in the group. Hence it is inevitable that our initial results might be less promising. If this occurs, we hope our model could generate less complicated sentences like "There is a red hat on the desk" and later on improve our model based on our existing achievements.

(Note on Assessment: Your group will not be graded on whether your results are "good". It is likely you will have a successful project even if you have not successfully implemented your described nice-to-haves, or even project deliverable goals.)

Review of related work

(Describe 3 related resources to your proposed project.

Ideally at least 1 resource will be a textbook chapter or review article giving introduction and context to the primary technical aspects of your project. If you propose any applications involving non-standard data, please include a resource describing similar projects. This is primarily that your considering any application that has data appropriately available to allow your attention to be primarily on aspects relevant to our course.

If you are using this project as an opportunity to investigate these methods on your primary research interest, include a resource that describes any connections or attempts at this application.
)

1. The paper which will be the base of our project: https://arxiv.org/pdf/2102.12092.pdf (Zero-Shot Text-to-Image Generation)

   The code of the DALL·E model:

   https://github.com/openai/DALL-E/issues?q=is%3Aissue+is%3Aclosed (we definitely won't copy this code)

2. The paper which introduces the dVAE:

   https://arxiv.org/abs/1609.02200

3. The paper which centers on the transformer:

   https://arxiv.org/abs/1706.03762

Thank you so much for reading! :)