TEXT TO IMAGE TO TEXT (T2I2T)

TEAM MENTOR: Aayush
TEAM MEMBERS: Umang | Nidhi | Varun | Chiranjeet | Abhranil

Abstract Introduction

The aim of this project is to convert text to image, and vice-versa. In other words, if a text prompt is provided as input, the aim is to generate high-quality, artistic random images coherent with the input. And if an image is provided as input, captions will be generated that are consistent with the description of the image.

Methodology

- For the text to image part, three pre-trained models are used. First, a pre-trained Dall-E Mini model is used to convert the input text to encoded images. Second, a pre-trained DCGAN model is used which decodes the encoded images. Finally a pre-trained CLIP model scores the predictions(the score is called CLIP score). The predicted images are then displayed in descending order of CLIP score.

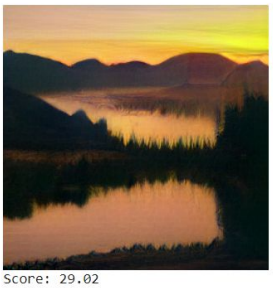
- For the image to text part, we have used an encoder-decoder architecture. The encoder part consists of a CNN architecture that is used to extract features from the input image, and the extracted features are then passed to the decoder part where a RNN architecture is used to generate a text output corresponding to the image input.

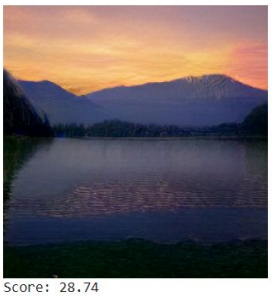
Reference

1. https://huggingface.co/spaces/dalle-mini/dalle-mini
2. https://arxiv.org/pdf/2102.12092.pdf
3. https://prvnk10.medium.com/encoder-decoder-model-for-image-captioning-e01c9392ea7f
4. https://arxiv.org/abs/2104.08718

Results

- Text to Image part (prompt used is "sunset over the lake in the mountains")


Score: 29.02


Score: 28.74


Score: 27.80


Score: 27.61

- Image to Text part (the following picture shows both the input image and the output text):


A surfer dives into the ocean


A person is walking along a beach with a big dog

Conclusion

Computational limitations have hindered our ability to build and train our own models for both the parts. For the text to image part, we have managed to overcome the limitation by using pre-trained models. Considering computational hindrances, Dall-E Mini is the best choice for this part, though other architectures with comparable performances can also be considered. If computational limitations are absent, Dall-E 2, can be considered as an excellent upgrade over Dall-E mini.

For the image to text part, the problem is partially solved by reducing the size of the dataset. This causes the output text to be slightly inaccurate, but it is not significant enough to be considered over the ease in training the architecture that it provides.