Stable Diffusion - Image to Prompts

Pin-Yen Liu
Department of Communications Engineering
pinyen.ee11@nycu.edu.tw

Po-Chun Huang
Graduate Degree Program of Cybersecurity
qmao.cs11@nycu.edu.tw

Po-Chuan Chen
Department of Electronics and Electrical Engineering
present90308.ee11@nycu.edu.tw

Abstract

This project is a Kaggle competition called **Image to Prompts** Ashley Chow [2023]. In this competition, we need to **reverse the typical direction of a generative text-to-image model**: instead of generating an image from a text prompt. We want to create a model which can predict the text prompt given a generated image. And making predictions on a dataset containing a wide variety of (prompt, image) pairs generated by Stable Diffusion 2.0, to understand how reversible the latent relationship is. With our method, we can reach the SBERTReimers and Gurevych [2019] score 0.57733 in this competition.

1 Introduction

The goal of this competition is to reverse the typical direction of a generative text-to-image model: instead of generating an image from a text prompt, can you create a model which can predict the text prompt given a generated image? You will make predictions on a dataset containing a wide variety of `(prompt, image)` pairs generated by Stable Diffusion 2.0, in order to understand how reversible the latent relationship is.

The popularity of text-to-image models has spurned an entire new field of prompt engineering. Part art and part unsettled science, ML practitioners and researchers are rapidly grappling with understanding the relationships between prompts and the images they generate. Is adding "4k" to a prompt the best way to make it more photographic? Do small perturbations in prompts lead to highly divergent images? How does the order of prompt keywords impact the resulting generated scene? In here, we will use three different models to handlw this issue.

2 Method

Our method is to ensemble the CLIP Interrogator, OFA model,and ViT model.

Here's the ratio for three different model

1. Vision Transformer (ViT) model: 74.88%
2. CLIP Interrogator: 21.12%
3. OFA model fine-tuned for image captioning: 4%

Vision Transformer (ViT). The Vision Transformer (ViT) Dosovitskiy et al. [2021] is a transformer encoder model (BERT-like), and pretrained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224. Vision Transformer (ViT) model pre-trained on ImageNet-21k (14 million images, 21,843

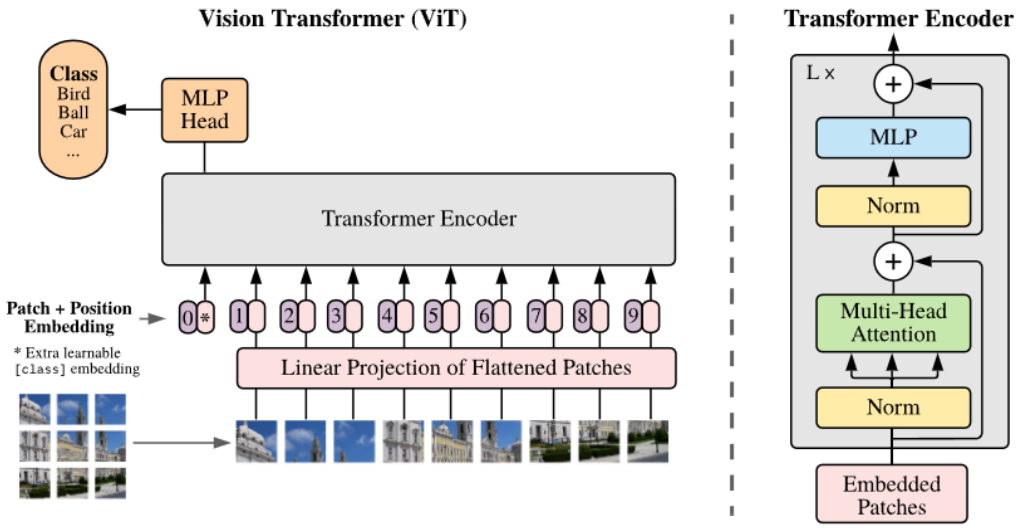

Figure 1: Vision Transformer Architecture

classes) at resolution 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224.

An overview of the model is depicted in Figure 1. The standard Transformer receives as input a 1D sequence of token embeddings. To handle 2D images, we reshape the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times \left(P^2 \cdot C\right)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. The Transformer uses constant latent vector size D through all of its layers, so we flatten the patches and map to D dimensions with a trainable linear projection. We refer to the output of this projection as the patch embeddings.

CLIP Interrogator. The CLIP Interrogator is a prompt engineering tool that combines OpenAI's CLIP Radford et al. [2021] and Salesforce's BLIP Li et al. [2022] to optimize text prompts to match a given image. In Figure 2 is the pipeline for CLIP Interrogator. CLIP Interrogator pipeline looks as follows:

1. An image is passed to the input to BLIP to obtain the main description.
2. An image is passed to the input to CLIP to receive its embedding.
3. Embeddings received from the image are compared with embeddings received from labels from the lists and the top 4 with the greatest similarity are selected.
4. There are 4 main lists on which the outgoing prompt for the CLIP part is formed: `artists.txt` (list with artists), `flavors.txt` (main list for image description), `mediums.txt` (image type), `movements.txt` (image style) and `sites.txt` (popular artwork sites). In here, if we remove `artists.txt` and `sites.txt`, the performance will improve.

The resulting texts are concatenated and returned as an image description (or promt on which an image was generated).

OFA model fine-tuned for image captioning. OFA model Wang et al. [2022] is a unified multimodal pretrained model that unifies modalities (i.e., cross-modality, vision, language) and tasks (e.g., image generation, visual grounding, image captioning, image classification, text generation, etc.) to a simple **sequence-to-sequence learning framework**. In here, we use **OFA-large-caption** for image captioning for pre-trained model, it uses COCO dataset for training the model.

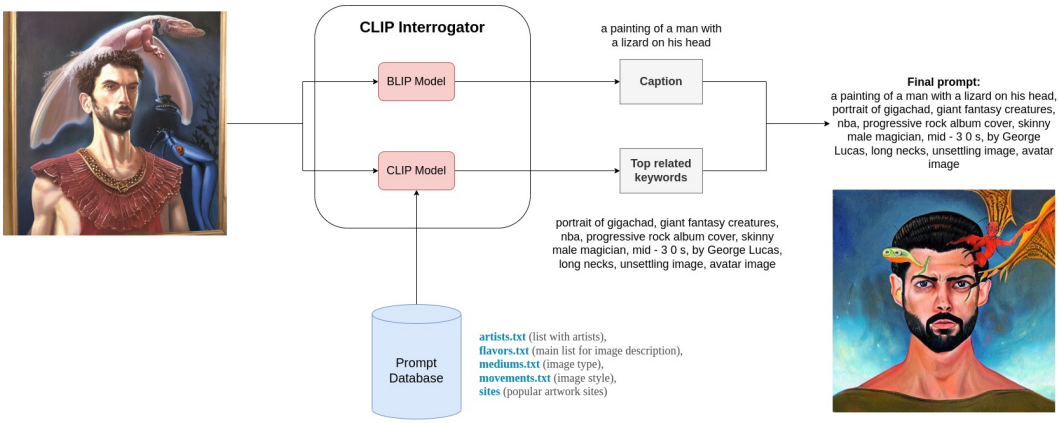

Figure 2: CLIP interrogator pipeline

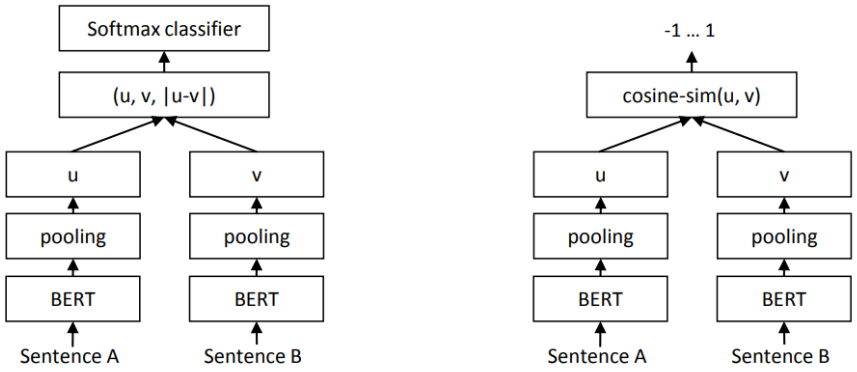

Figure 3: SBERT architecture. (Left) With classification objective function. (Right) In inference mode.

3 Experiments

Because our method is to use pre-trained model. So there has no dataset for our method.

But, there have some datasets that can be used in this competition. The dataset is build in (image, prompt) pairs. For example:

- DiffusuinDB Reimers and Gurevych [2019] : 14 million image-text pairs
- COCO (Microsoft Common Objects in Context) Lin et al. [2015] : 2.5 million labeled instances in 328k images
- Laion2B-en Webster et al. [2023] : 2.3 billion image-text pairs

The evaluation metric is from a paper Sentence-BERT Reimers and Gurevych [2019]. In the Figure 3. We can know that the input will be the prompt sentence, output will be the embedding. We can get the embedding with using sentence transformer.

Here has an example in Figure 4, if we have an image like this below We may need to generate a prompt **ultrasaurus holding a black bean taco and a piece of cheese in the woods** for this image.

Here are a sample figure 5 that using generated prompts from CLIP Interrogator and OFA model. The left side image is generated from OFA's prompt. And the right side image is generated from CLIP Interrogator's prompt. Finally, we can see the SBERT score for three different prompts in Table 1.

Figure 4: Sample image

Figure 5: Reconstruct diffusion image from prompts

Model	Caption	SBERT Score
ViT	comic book, triceratops, coffee mug, jigsaw puzzle, book jacket, dust cover, dust jacket	0.553
CLIP Interrogator	a cartoon dinosaur with a piece of cheese in its mouth, boardgamegeek, gaia, woodland background, pancake flat head, inspired by Charles Fremont Conner, mobile learning app prototype, is essentially arbitrary, blue scales, museum item, nachos, prehistory, plates	0.739
OFA	A blue dinosaur eating a piece of cheese in a forest	0.772

Table 1: SBERT score for a sample image

4 Discussions

This competition aims to explore the reversibility of the latent relationship between text prompts and generated images, challenging participants to predict the text prompt from a given image.

It highlights the importance of prompt engineering in text-to-image models and the potential for understanding the intricate relationships between prompts and images.

The combination of CLIP Interrogator, OFA, and ViT models provides a robust approach for analyzing and predicting prompt embeddings.

5 Future work

But the method for this competition can still be improved.

Using the generation of a large dataset of images and prompts, along with modifications and training of models, demonstrates the potential for further improvements in the field of text-to-image generation and prompt engineering (Score: 0.66316, Rank 1) bestfitting [2023].

Or, the utilization of a ViT-based method and the creation of a customized dataset, indicating the importance of supervised learning in predicting sentence embeddings and its relevance to advancing text-to-image models (Score: 0.65179, Rank 2) KaizaburoChubachi [2023].

References

Will Cukierski Ashley Chow, inversion. Stable diffusion - image to prompts, 2023. URL https://kaggle.com/competitions/stable-diffusion-image-to-prompts.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL http://arxiv.org/abs/1908.10084.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b, 2023.

bestfitting. 1st place solution, 2023. URL https://www.kaggle.com/competitions/stable-diffusion-image-to-prompts/discussion/411237.

KaizaburoChubachi. 2nd place solution, 2023. URL https://www.kaggle.com/competitions/stable-diffusion-image-to-prompts/discussion/410606.