TOPIC: Stable Diffusion - Image to Prompts

Deduce the prompts that generated our "highly detailed, sharp focus, illustration, 3d renders of majestic, epic" images

GOAL: The project aims to predict or anticipate the matching prompts to produce the pictures in the dataset

DATASET: The prompts range in complexity from simple to sophisticated, requiring many items and modifiers, and were produced using unknown methods. The pictures were produced in 50 steps at a resolution of 768x768 pixels using Stable Diffusion 2.0 (768-v-ema.ckpt). The generated photos were reduced in size for the competition dataset to 512x512 pixels. The default settings have been kept untouched in the script that generated the photos.

INTRODUCTION:

A text-to-image model using deep learning called Stable Diffusion was launched in 2022. Although it may be used for various tasks, including inpainting, outpainting, and creating image-to-image translations directed by text prompts, its primary application is to produce detailed visuals conditioned on text descriptions.

Stability AI created stable diffusion in cooperation with academic scholars and nonprofit groups. It can function on most consumer hardware outfitted with a modest GPU and at least 8 GB VRAM, and its code and model weights have also been made open source.

Stable Diffusion created the generative AI text-to-image online software known as DreamStudio. Like DALL-E2, it creates graphics from prompts using natural language processing and gives users input controls to personalize the image further. It is seen as DALL-E2's rival.

We aim to develop a model that can accurately reverse the diffusion process that produces a particular picture, i.e., a matching prompt that can produce the images produced using stable diffusion should be identified. The embeddings of our prompts will be used to calculate prompt similarity robustly.

LITERATURE REVIEW:

PAPER 1: Diversity is Definitely Needed: Improving Model-Agnostic Zero-shot Classification via Stable Diffusion

AUTHORS: (Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes)

In this paper, the authors offer insights that MA-ZSC(ModelAgnostic Zero-Shot Classification) performance may be enhanced by boosting the variety of photos in the produced dataset. Their study gives a new viewpoint on the MA-ZSC issue. Using a diffusion model that has already been trained, they suggest several changes to the text-to-image creation procedure that they refer to as our "bag of tricks" to improve variety. The method proposed significantly improved several classification designs, and the results are on par with those of cutting-edge models like CLIP. To validate their methodology, they tested CIFAR10 and CIFAR100.

PAPER 2: What's in a Text-to-Image Prompt? The Potential of Stable Diffusion in Visual Arts Education

AUTHORS: Nassim Dehouche, Kullathida Dehouche

The authors have codified this novel way of producing art and assessed its potential for instructing art history, aesthetics, and methods by analyzing a collection of 72,980 Stable Diffusion prompts. According to this study, text-to-image AI can revolutionize how art is taught by opening up new, accessible channels for experimentation and artistic expression. It illustrates how AI may be used to improve art instruction. But it also raises important questions about who owns artistic works. As more and more work is created with the help of these instruments to protect artists' rights, it will be crucial to develop new legal and business structures.

PAPER 3: CoCa- Contrastive Captioners are Image-Text Foundation Models

AUTHORS: Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini Yonghui Wu

Yonghui Wu

The authors present CoCa, a simple design to combine contrastive loss with captioning loss while pretraining an image-text encoder-decoder foundation model, using model capabilities from generative techniques like SimVLM and contrastive approaches like CLIP. An image encoder extracts characteristics from pictures, and a text encoder, which encodes the captions, makes up the conventional Contrastive Captioner paradigm. The CoCa model is trained to increase the similarity between picture features and their related accurate captions while decreasing the similarity between wrong captions and image features. On a variety of downstream tasks, including visual recognition (ImageNet, Kinetics400/600/700, Moments-in-Time), crossmodal retrieval (MSCOCO, Flickr30K, MSR-VTT), multimodal understanding (VQA, SNLI-VE, NLVR2), and image captioning (MSCOCO, NoCaps), CoCa empirically achieves state-of-the-art performance with zero-shot transfer or minimal task-specific adaptation. With a frozen encoder and trained classification head, CoCa achieves zero-shot top-1 accuracy of 86.3%, 90.6%, and 91.0%, respectively.

PAPER 4: BLIP- Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

AUTHORS: Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi -Salesforce Research

The authors have introduced BLIP, a novel VLP framework adaptable to vision-language interpretation and generation operations. Bootstrapping the captions—where a captioner creates artificial text captions, and the filter pulls out the noisy ones—allows BLIP to use the noisy online data efficiently. They obtained cutting-edge outcomes on various vision-language tasks, including image-text retrieval, image captioning, and the VQA. BLIP also exhibits significant generalization abilities when applied in a zero-shot fashion to video language problems.

RELATED WORK:

The COCA, CLIP, and BLIP models are used in our effort to create text prompts from images, and we describe studies that are closely linked to them in this part.

Paper 1: by Radford et al. (2021), "Learning Transferable Visual Models From Natural Language Supervision"

The CLIP (Contrastive Language-Image Pretraining) model, which trains visual representations under natural language supervision, is presented in this study. The model is developed using a sizable dataset of photos and the textual descriptions that go with them. The CLIP model is made to excel at a variety of visual understanding tasks, including object identification, captioning images, and answering visual questions. The CLIP model is used in our project to create text prompts from photos. The paper also covers several CLIP model shortcomings. For example, the model tends to produce reasonable but wrong replies for several visual question-answering tasks and is sensitive to small changes in the wording of textual prompts.

Paper 2: by Xu et al. (2015), "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention"

This paper presents a neural network-based model for creating picture captions. The model uses an attention mechanism to concentrate on particular input image areas during captioning. The model creates more accurate and in-depth captions by focusing on different areas of the image with selective attention. Because it illustrates the potential advantages of including attention mechanisms in image-to-text generation models, this work is pertinent to our endeavor.

Paper 3: "Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy et al. (2015)

The algorithm described in this paper learns to create textual descriptions of images by matching visual regions with words in a phrase. The model uses a deep neural network that concurrently learns to produce textual descriptions and recognizes items and their relationships in an image. This research is pertinent to our study since it shows how deep neural networks can create text prompts from images by coordinating visual and semantic information.

DATA EXPLORATION:

Dataset:

The prompts.csv contains the images in the training dataset and their corresponding prompts that were used initially to derive the images.

```
import matplotlib.pyplot as plt
import seaborn as sns

# Load the prompts.csv file
prompts_df = pd.read_csv("../input/stable-diffusion-image-to-prompts/prompts.csv")

# Print the first few rows of the dataframe
print(prompts_df.head())
```

```
      imgId                                             prompt
0  20057f34d  hyper realistic photo of very friendly and dys...
1  227ef0887  ramen carved out of fractal rose ebony, in the...
2  92e911621  ultrasaurus holding a black bean taco in the w...
3  a4e1c55a9  a thundering retro robot crane inks on parchme...
4  c98f79f71  portrait painting of a shimmering greek hero, ...
```

Shape of the dataframe:

The dataframe contains 7 rows which are the 7 images derived as a result of stable diffusion and the 2 columns are the imgId and the corresponding prompts

```
print(prompts_df.shape)
```

```
(7, 2)
```

DATA PREPROCESSING:

The preprocessing steps include loading images from a directory, resizing them to a target size, normalizing pixel values, and storing the preprocessed images in a list. These steps are crucial for ensuring that the input data is consistent and compatible with the model's requirements. The preprocessing steps are as follows:

Image Resizing: To ensure that all images have the same dimensions, we resize each image to a predefined target size (target_size = (256, 256)). This step is essential because most deep learning models require consistent input dimensions to process the images effectively.

Pixel Value Normalization: After resizing the images, we normalize the pixel values to fall within the range of 0 to 1 by dividing each pixel value by 255. This normalization step is crucial for deep learning models, as it helps in improving the model's convergence during training and stabilizing the gradients.

PROBLEM FORMULATION:

Targeted Problem: The main challenge is to produce pertinent and meaningful prompts that match to images produced by stable diffusion methods. These prompts should describe or relate to the content, features, or themes present in the images.

Goal: The main goal is to create a model that can precisely and reliably produce appropriate prompts for a given collection of stable diffusion pictures. The prompts should adequately explain the photos and be varied and logical. We intend to develop a model, that accurately and reliably provides relevant and varied prompts that explain or connect to the content, characteristics, or themes present in a series of pictures produced utilizing stable diffusion techniques while taking into account the project's limits.

Constraints: A project's constraints may include the availability of labeled data (pictures with associated prompts) there are only 7 images given and the corresponding prompts used to derive those images are given. There are no much models and research papers to support this problem. Managing the computing resources, time restrictions for the development of prompts, and the required level of prompt specificity or originality or other constraints in the Kaggle competition.

MODEL SELECTION:

CLIP Interrogator:

The CLIP Interrogator is a tool for creating text prompts that combines Salesforce's BLIP and OpenAI's CLIP to make text prompts that are optimized to match a picture.Using DreamStudio we can create an amazing images using the resultant prompts.

CLIP Interrogator pipeline looks as follows:

- In order to acquire the primary description, a picture is supplied to BLIP's input.

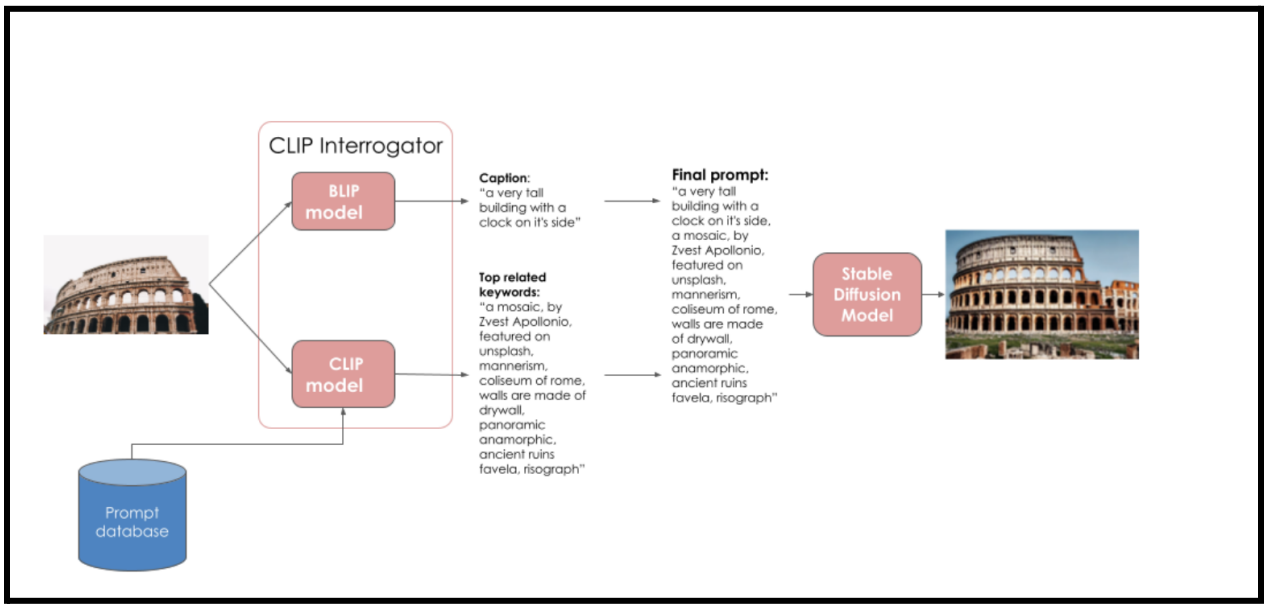
- In order to get an image's embedding, CLIP receives an image as input.
- The top 4 embeddings with the highest similarity are chosen after embeddings from the picture and labels from the lists are compared.
- The outgoing prompt for the CLIP portion is composed of four primary lists: artists.txt (list of artists), flavors.txt (main list for picture description), mediums.txt (image kind), movements.txt (image style), and sites (popular art websites). The output score may be greatly raised by eliminating the artists.txt and the sites lists, as I said previously.

- The resultant sentences are combined to create an image description (or the promt from which a picture was created), which is then returned.
- **Input**:

  The input picture to test is called image (image file). The CLIP Interrogator will evaluate this image and provide an optimum text question.

  The CLIP model to employ for the analysis is clip_model_name (string). ViT-L-14/openai and ViT-H-14/laion2b_s32b_b79k are the two alternatives available for Stable Diffusion 1 and Stable Diffusion 2, respectively.

  The mode for prompt creation is mode (string). The options are "best" and "fast." While the "fast" option is speedier, requiring only 1-2 seconds, the "best" mode produces higher quality results but takes 10–20 seconds to finish.

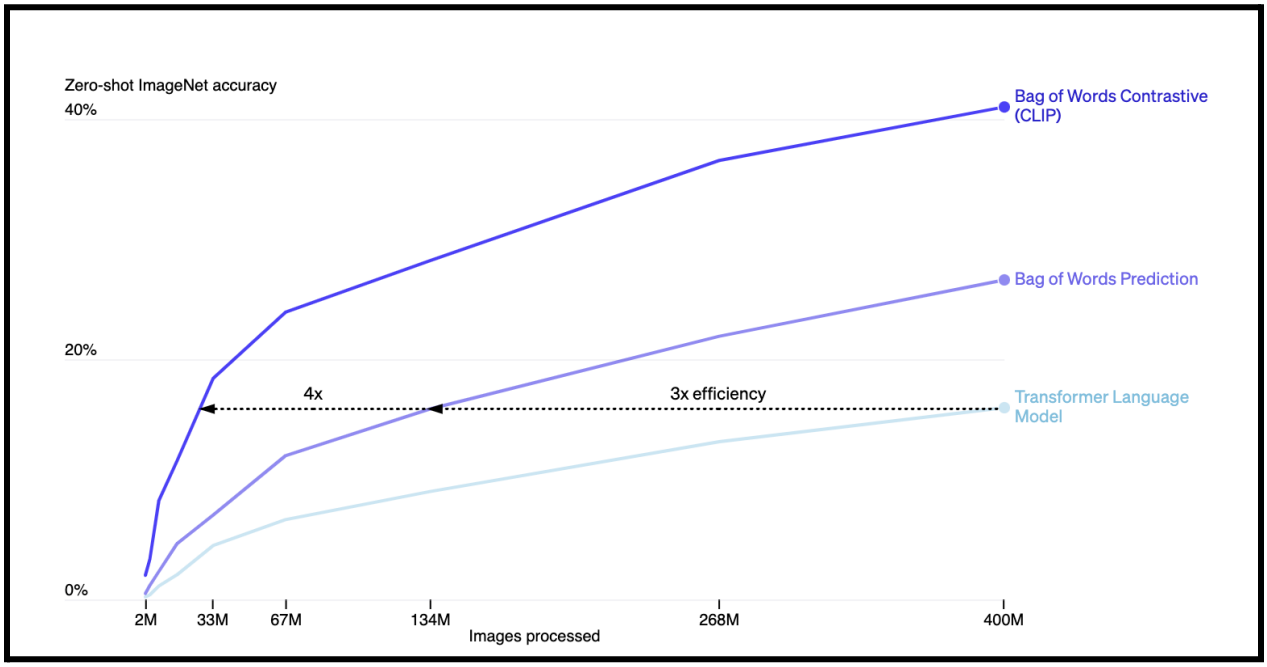


Sourced from https://medium.com/@silkworm/diversify-photo-database-with-clip-interrogator

CLIP(Contrastive Language–Image Pre-training):

A new neural network called CLIP is proposed by Open AI and it effectively picks up visual notions from natural language supervision. Similar to the "zero-shot" features of GPT-2 and GPT-3, CLIP may be used to any visual classification benchmark by only giving the names of the visual categories to be recognized. CLIP is highly efficient because when employed with CLIP, the contrastive objective is four to ten times more effective at zero-shot ImageNet categorization. The computation efficiency is three timed more efficient than normal ResNet. Final results has shown that CLIP model trains on 256 GPUs for two weeks, which is

comparable to current large-scale picture models. Additionally, CLIP does not generalize well for photos not included in its pre-training dataset. Zero-shot CLIP only achieves 88% accuracy when tested on handwritten digits from the MNIST dataset, far less than the dataset's 99.75% human accuracy.



Sourced from https://openai.com/research/clip

BLIP(Bootstrapping Language-Image Pre-training):

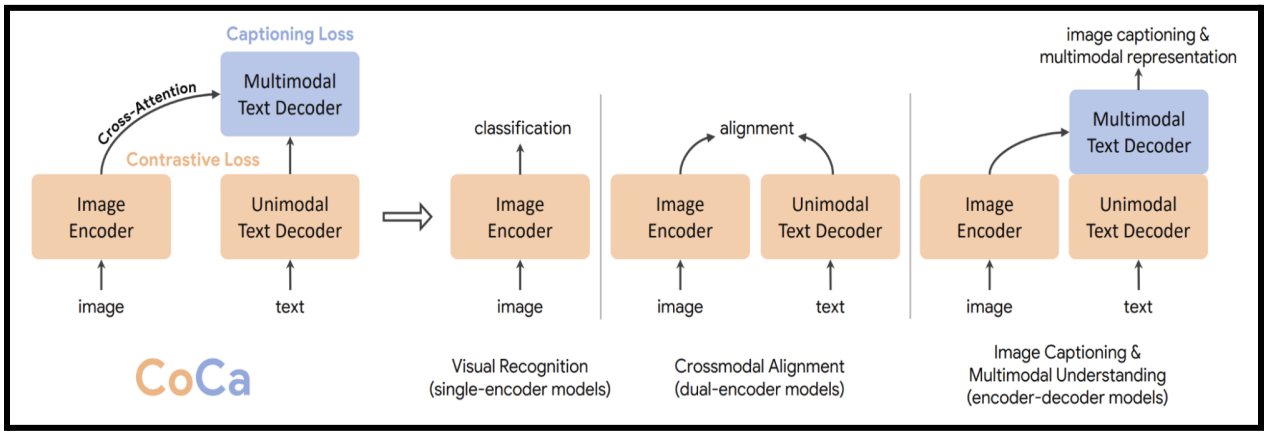
The multi-modal activities that the BLIP model is capable of doing includes Answering Questions Visually, image-text matching (image-text retrieval), captioning of images. This model effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones. In BLIP, an encoder-decoder and a multi-task model are combined into a multimodal system that can perform three modes.

- **Unimodal encoders**:Encoders that encode text and graphics individually. A vision transformer is the image encoder. It uses the same text encoder as BERT.
- **Image-grounded text decoder**: This replaces the bi-directional self-attention layers in the text encoder with causal self-attention layers.
- **Image-grounded text encoder**: This injects visual information by inserting a cross-attention layer between the self-attention layer and the feed-forward network for each transformer block of the text encoder.

CoCa(Contrastive Captioners):

Creating text suggestions connected to images is crucial in many applications, including caption generation, image interpretation, and visual storytelling. Contrastive Captioners are image

captioning models that use contrastive learning strategies to provide precise, varied, and resilient descriptions for pictures. They are made to be able to generalize to new, unheard-of pictures by learning from both successful and unsuccessful instances. A text encoder is in charge of encoding the captions while an image encoder is utilized to extract features from the pictures in a Contrastive Captioner model. The model has been trained to increase similarity between picture features and their related accurate captions while decreasing similarity between wrong captions and image features. With this contrastive learning strategy, images and captions that are comparable in terms of semantic content are placed near together, while those that are different are placed apart.



CLIP model (CoCa-ViT-L-14):

A neural network model called CLIP (Contrastive Language-Image Pretraining) was created by OpenAI to learn to align images and text in a zero-shot scenario. The "coca_ViT-L-14" pre-trained CLIP model, which was optimized on the MSCOCO dataset, is utilized in our code. A Vision Transformer with 14 layers makes up this model. In order for Vision Transformers to function, a picture must first be divided into patches, which are then linearly embedded and processed through a number of Transformer layers. Tasks requiring picture identification work particularly well with this design.

Images and text prompts are both encoded using the CLIP paradigm, producing embeddings that may be compared in a common embedding space. As a result, the model may establish a connection between the pictures and their associated written instructions.

Sentence Transformer model (MiniLM-L6-v2):

Using Transformer models like BERT, RoBERTa, and MiniLM, the Sentence Transformers Python package enables us to produce dense sentence embeddings. Our code makes use of the

pre-trained model "all-MiniLM-L6-v2". The original BERT architecture has six layers; MiniLM has fewer layers (L6 represents six layers). Compared to the original BERT model, this model is optimized for producing sentence embeddings at a lower computational cost.

The sentence embeddings for the text prompts are computed in our code using the Sentence Transformer model.

Coca + Clip:

The CLIP model functions by latently encoding both the images and the sentences. The model is employed in this use case to identify the most pertinent prompt for a given image. The dataset's preprocessed photos are encoded into image embeddings. The most similar text prompt for each image is then found by comparing these image embeddings with the text embeddings produced by the SentenceTransformer model.

To sum up, the OpenAI CLIP model is used to encode images and texts into the same latent space, and the SentenceTransformer model is utilized to generate text embeddings for the dataset's prompts. The most pertinent question for each image in the collection is then determined by comparing these embeddings.

Blip :

The Blip The model receives an input image, converts it to pixel values, and then creates a caption (prompt) for that image. After that, other tasks like computing embeddings are performed using the generated prompt.
The SentenceTransformer model computes an embedding for a sentence given as input (in this case, the generated prompt). After that, the embeddings are put to use in downstream tasks like clustering and similarity judging.

In conclusion, the SentenceTransformer model computes embeddings for these prompts and the BlipForConditionalGeneration model generates prompts (captions) for the photos. Combining these models enables the code to process both picture and text input, and the resulting embeddings may be applied to a number of different tasks, including ranking, grouping, and comparison.

EVALUATION:

The assessment metric used to compare the generated prompts to the ground truth prompts was cosine similarity. The higher the cosine similarity, the generated prompts are more semantically similar the ground truth prompts. Used cosine similarity in our project to compare the embeddings of the generated prompts to the ground truth prompts in order to assess the quality of the generated prompts. The ground truth prompt embeddings and the created prompt embeddings have a calculated cosine similarity of 0.5283.

Our finding shows that the generated prompts and the ground truth prompts are quite comparable, which is good news for our model. There is still room for development, though. We might experiment with other model topologies, fine-tuning tactics, or training data augmentation techniques to improve the performance of our model. Additionally, we might examine the generated prompts with lower similarity scores to find any particular problems or trends that the model's future iterations would want to take into account.

To determine the cosine similarity between a collection of reference questions and the created prompts. An angle between two non-zero vectors in a multi-dimensional space, in this case the embeddings of the generated prompts and the reference prompts, is measured using the cosine similarity metric. There is more resemblance between the two prompts when the cosine similarity score is larger.

In order to create a list, the reference prompts must first be taken from the prompts_df DataFrame. Then, a similarity_scores matrix with the shape (number of generated prompts, number of reference prompts) is initialized with zeros.

Cosine similarity for Blip + Clip Model

```
Generated Prompt 1 has highest similarity with Reference Prompt 1: 0.21
Generated Prompt 2 has highest similarity with Reference Prompt 2: 0.46
Generated Prompt 3 has highest similarity with Reference Prompt 3: 0.38
Generated Prompt 4 has highest similarity with Reference Prompt 4: 0.48
Generated Prompt 5 has highest similarity with Reference Prompt 5: 0.68
Generated Prompt 6 has highest similarity with Reference Prompt 6: 0.55
Generated Prompt 7 has highest similarity with Reference Prompt 7: 0.32
```

Cosine similarity for Coca + Clip Model

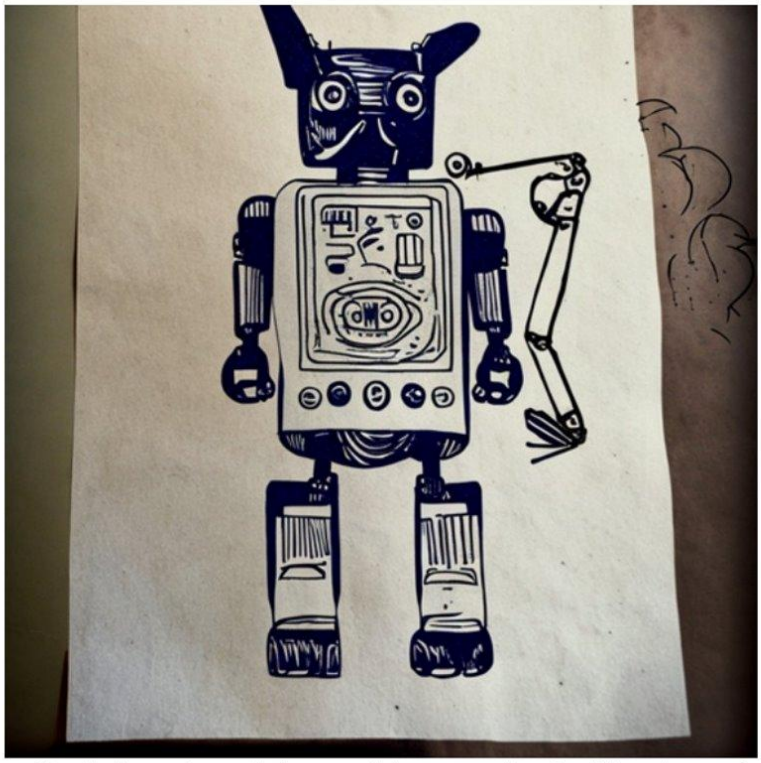
```
Generated Prompt 1 has highest similarity with Reference Prompt 1: 0.44
Generated Prompt 2 has highest similarity with Reference Prompt 6: 0.25
Generated Prompt 3 has highest similarity with Reference Prompt 3: 0.42
Generated Prompt 4 has highest similarity with Reference Prompt 5: 0.26
Generated Prompt 5 has highest similarity with Reference Prompt 4: 0.46
Generated Prompt 6 has highest similarity with Reference Prompt 5: 0.39
Generated Prompt 7 has highest similarity with Reference Prompt 6: 0.23
```

RESULT ANALYSIS:

The below image shows the generated prompts for both the models

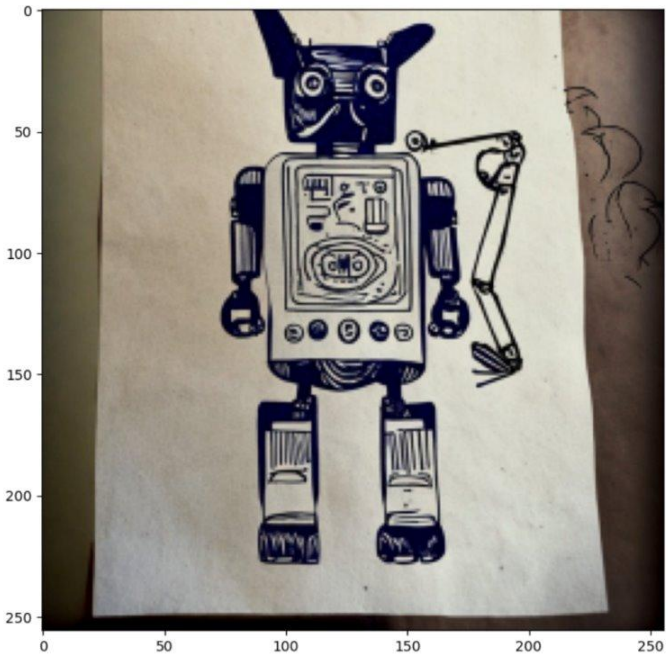
Results for Coca + Clip Model:

Below image shows the original image with prompt for the highest cosine similarity



a thundering retro robot crane inks on parchment with a droopy french bulldog

Below image shows the Generated image with prompt for the highest cosine similarity



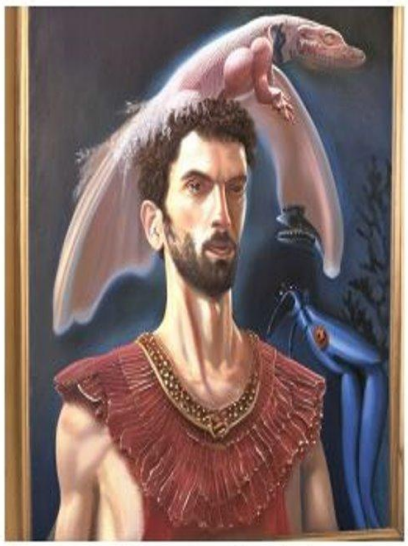
a drawing of a toy robot with scissors on it

Results for Blip Model:

Below image shows the original image with prompt for the highest cosine similarity

portrait painting of a shimmering greek hero, next to a loud frill-necked lizard

Below image shows the Generated image with prompt for the highest cosine simila



Grenerated Prompt: ['painting of a man with a snake skin on his head and a snake skin on his head']

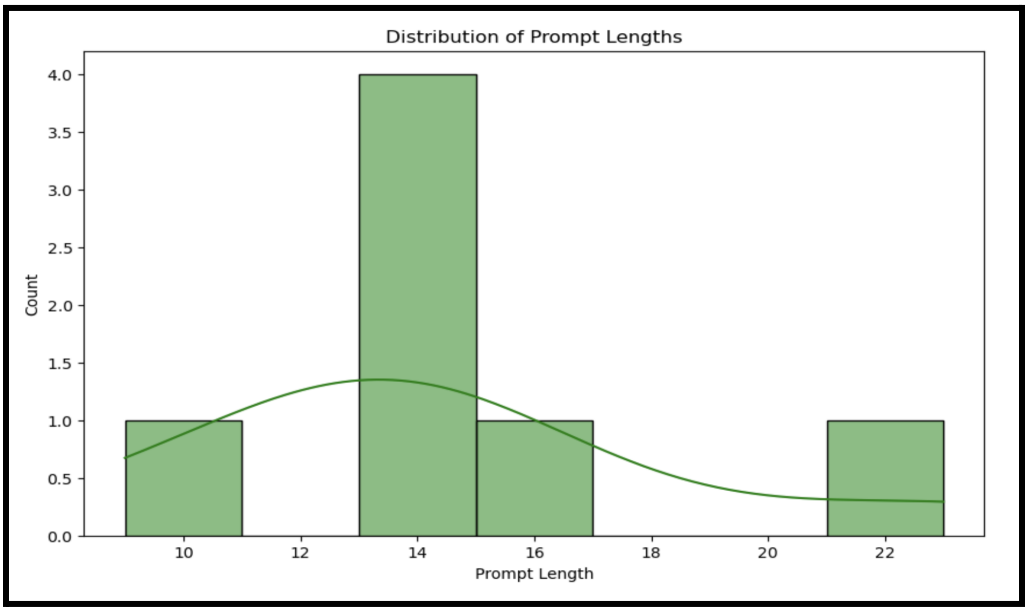
VISUALIZATION:

Data Examples and Corresponding prompts:

ultrasaurus holding a black bean taco in the woods, near an identical cheneosaurus

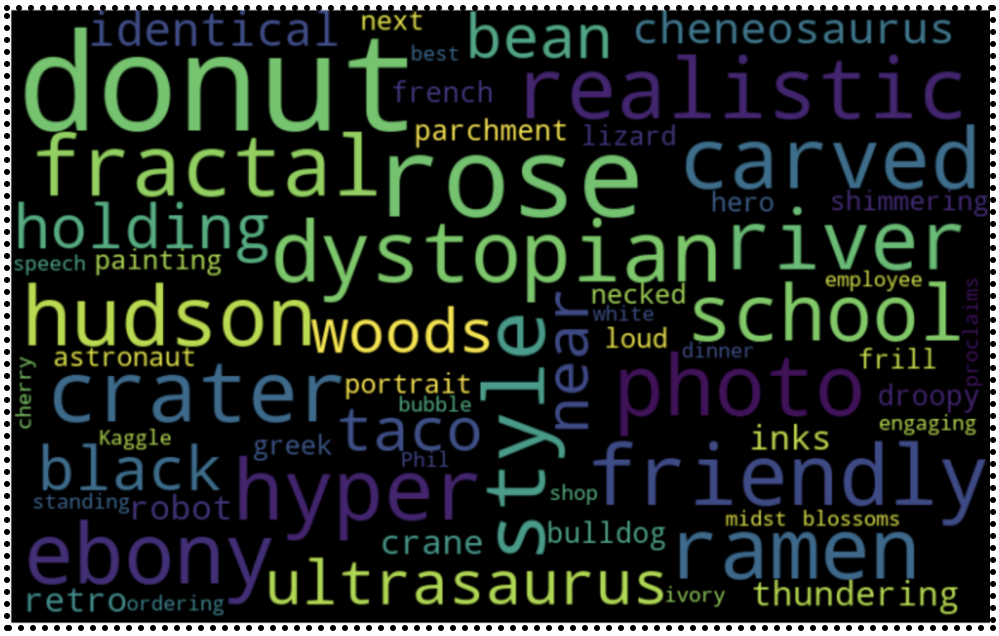
Prompt Length Distribution:

The maximum length of the prompt is 22 and the minimum length is 10 and the below plot shows the distribution of the length of the prompt.



Word cloud for the words in the prompt:

The below figure shows terms from the prompts.csv word cloud format, with the size of each word denoting its frequency in the text. Less often used words will be smaller and more infrequent ones will be bigger. The most recurring themes or subjects in the text data can be spotted immediately by using this visualization.



Model Justification:

Performance: In producing insightful text prompts for images across multiple domains, both Blip + Clip and Coca + Clip have shown strong performance. They are a good fit for this project because it aims to create interesting and accurate textual representations for a variety of photos because they can deliver high-quality prompts.
Flexibility: The Blip + Clip and Coca + Clip models are flexible to many application fields because they can be tweaked on a variety of datasets. This adaptability enables the investigation of many use cases and the potential to address particular difficulties in producing text prompts for images.
Ease of Integration: Because pre-trained models are available and well-known deep learning frameworks like PyTorch offer support, it is simple to integrate the models into the project pipeline. The development process can be accelerated because to the integration's simplicity, and the models' implementation and evaluation take less time and effort.

Real world Applications:

Image captioning: This project can be used to provide captions for pictures automatically. Such a feature would be beneficial in a number of areas, such as social media platforms, content management systems, and accessibility tools for users who are blind or visually impaired.

Visual Storytelling: Creating textual narratives or stories based on a series of photos can be done using the project as a foundation. This can be used to summarize movies, write children's novels, or come up with unique descriptions for photo albums.

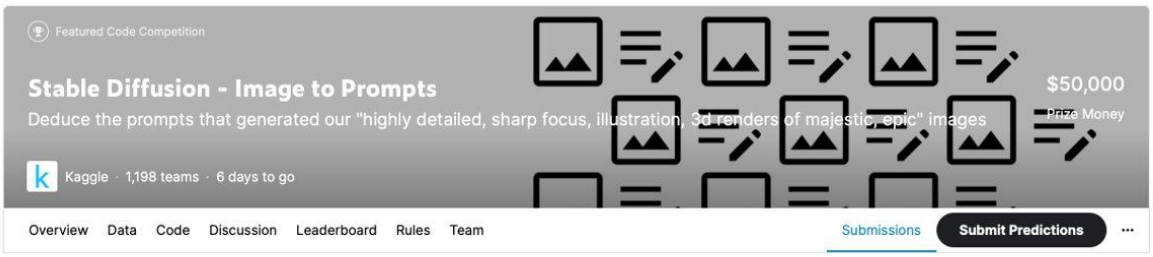
Advertising and Marketing: This technology can be used to produce compelling marketing content, like as ad copy or product descriptions, to increase audience engagement and boost sales by providing contextually relevant and interesting prompts for photos.

Kaggle Score and Position:

1095	Kiazgu	0.40057	1	2mo
1096	Pankaj Pansari	0.40057	8	2mo
1097	Dmitrii Shubin	0.40057	6	2mo
1098	Damir	0.40057	2	1mo
1099	prompters	0.40057	3	7d
1100	**Divya Nalam**	0.40057	3	1h

Your Best Entry!
Your submission scored 0.40057, which is not an improvement of your previous score. Keep trying!

1101	kevin:)	0.39866	5	2mo
1102	Darshan Makwana	0.39844	3	2mo
1103	mt	0.39818	1	2mo
1104	adypd	0.39805	10	1mo
1105	e-toppo	0.39526	4	2mo



Submissions

Select up to 2 submissions that will count towards your final leaderboard score. If less than 2 are selected, Kaggle will automatically select from your best scoring submissions. Learn More

0/2

Auto-selection candidates

All Successful Selected Errors

Recent

Submission and Description	Public Score	Select
submission_blip_pretrained - Version 4 Succeeded · sajit08 · 1h ago · Notebook submission_blip_pretrained | Version 4	0.40057	