Predicting Stable Diffusion Prompts Based on Generated Images

Alison Reed abr9982@nyu.edu

David Glaser djg448@nyu.edu

Naman Soni ns5239@nyu.edu

See Github repository here

Introduction

Image generation has become a popular focus in deep learning research. Using large quantities of image samples and high-performing computing, deep learning models have displayed an unprecedented ability to learn image features and generate original images of high quality.

Generative Adversarial Networks (GANs) were a major breakthrough in image generation. In 2014, Goodfellow et al. demonstrated that by pitting two deep neural networks against one another (a generator and a discriminator), GANs can generate highly realistic images [9].

Diffusion models were the next big breakthrough in image generation. Diffusion models are probabilistic models which train by learning the addition (and the removal) of noise to an image until it is indistinguishable from random noise. Unlike GANs, diffusion models significantly mitigate mode collapse, a phenomenon where the model effectively memorizes a training example, deceiving itself into concluding that it is in fact generating a realistic image. Diffusion models are more lightweight than GANs which require min-max model optimization.

In combination with CLIP models, which are able to liken meaning in text with the subject of an image, diffusion models have grown in popularity as a means to generate images from text. One such popular implementation of diffusion models is Stable Diffusion. Released in 2022 by Stability AI, Stable Diffusion's code is publicly available and able to be run on consumer hardware with basic GPU, making it attractive as a means for research on diffusion models [10].

Due to the impressive images generated by Stable Diffusion models, ongoing research seeks to further our understanding of the nature of diffusion models. One collective research effort was spurred by a Kaggle competition in 2023, entitled "Stable Diffusion – Image to Prompts"[7]. It encouraged participants to utilize deep learning models to work backward from Stable Diffusion: given an inputted Stable Diffusion-generated image, output the likely prompt that created the image. This paper outlines our attempts at reverse engineering Stable Diffusion. We used CLIP and generative language models, among others, to do so.

Literature Review

CLIP Model

OpenAI's Contrastive Language-Image Pretraining (CLIP) [4] model represents a significant advancement in the field of multimodal learning, which involves training models to understand and generate both text and images. CLIP is designed to learn visual concepts from natural language descriptions. This is achieved by pretraining the model on a vast dataset of image-caption pairs collected from the internet. The model's architecture consists of two separate encoders: one for images and one for text. These encoders are trained to project images and their corresponding text descriptions into a shared embedding space where similar images and text are located close to each other.

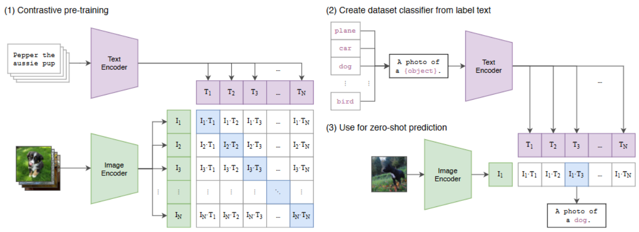
(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Fig. 1 CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

The strength of CLIP lies in its ability to perform zero-shot classification. This means it can recognize objects in images without having been explicitly trained on those specific categories. Instead, it relies on the natural language descriptions provided during training. CLIP's versatility and robustness make it a valuable tool for a variety of applications, including image classification, object detection, and even generating prompts based on images, which is particularly relevant for the Stable Diffusion – Image to Prompt competition.

GPT-2

The Generative Pre-trained Transformer 2 (GPT-2)[5] is an advanced language model developed by OpenAI, designed to generate coherent and contextually relevant text based on a given input. The architecture of GPT-2 is based on the transformer, a type of deep learning model that uses attention mechanisms to weigh the influence of different words on each other's context. Unlike earlier models that process words sequentially, transformers process all words in parallel, which allows for more complex understanding and generation of text.

GPT-2 is characterized by its large number of parameters and its ability to perform a wide variety of language tasks without task-specific tuning. It was trained using an unsupervised learning method on a diverse dataset compiled from the Internet. This extensive training enables the model to have a broad understanding of human language, making it capable of tasks such as translating text, answering questions, summarizing passages, and continuing from a given text prompt.

One of the most remarkable aspects of GPT-2 is its zero-shot learning capabilities, where the model performs tasks without having seen them during training. For example, when given the first half of a sentence, GPT-2 can generate multiple plausible continuations based on the patterns it has learned during its training. This ability makes GPT-2 highly versatile and useful for a range of applications from creating conversational AI agents to assisting in content creation.

CLIP Interrogator

The CLIP Interrogator leverages the combined capabilities of the CLIP and BLIP [3] models to generate detailed and contextually appropriate textual descriptions from input images. First, the CLIP model encodes the image into a high-dimensional vector within a shared embedding space where both images and text coexist. This encoding process captures the visual features of the image in a way that aligns with textual descriptions. The BLIP model, which is designed to bootstrap the language and image pre-training process, then utilizes these encodings to enhance the textual output, ensuring the generated descriptions are not only accurate but also contextually rich and coherent.

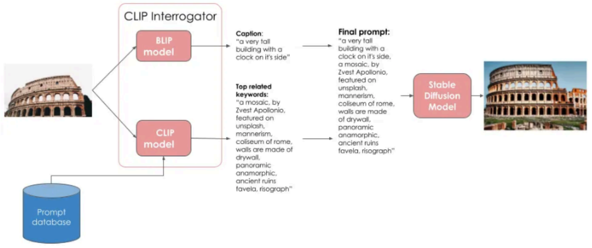
CLIP Interrogator
BLIP model
CLIP model
Prompt database
Caption:
"a very tall building with a clock on it's side"
Top related keywords:
"a mosaic, by Zvest Apollonio, featured on unsplash, mannerism, coliseum of rome, walls are made of drywall, panoramic anamorphic, ancient ruins favela, risograph"
Final prompt:
"a very tall building with a clock on it's side, a mosaic, by Zvest Apollonio, featured on unsplash, mannerism, coliseum of rome, walls are made of drywall, panoramic anamorphic, ancient ruins favela, risograph"
Stable Diffusion Model

Fig. 2 The above figure demonstrates the CLIP Interrogator Pipeline [6]

In practical use, the CLIP interrogator feeds the image through the CLIP model to obtain an initial set of potential textual descriptions based on their similarity scores with the image embedding. These candidate descriptions are then refined using the BLIP model, which fine-tunes the language generation process to produce more detailed and context-aware prompts. By combining the strengths of CLIP's robust visual understanding with BLIP's advanced language pre-training capabilities, the CLIP Interrogator excels in generating high-quality, descriptive text that closely matches the content and context of the input images. This integration is particularly valuable in applications like the Stable Diffusion – Image to Prompt competition, where precise and meaningful image-to-text translations are essential.

OFA

OFA [1] (One For All) represents a significant advancement in the field of AI by offering a unified architecture capable of addressing a variety of vision and language tasks. OFA leverages a single, cohesive model architecture to handle diverse tasks such as image captioning, visual question answering, text-to-image generation, and more. This approach reduces the need for task-specific models, enabling more efficient transfer learning and simplifying the deployment and maintenance of AI systems. OFA's design is particularly advantageous in resource-constrained environments, as it facilitates the use of a single model for multiple applications.

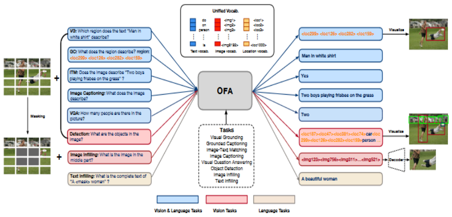
Unified Vocab.
do
on
person
is
Text vocab.
<img1>
<img2>
<img3>
<img8192>
Image vocab.
<loc1>
<loc2>
<loc3>
<loc1000>
Location vocab.
VG: Which region does the text "Man in white shirt" describe?
GC: What does the region describe? region: <loc299> <loc126> <loc282> <loc159>
ITM: Does the image describe "Two boys playing frisbee on the grass" ?
Image Captioning: What does the image describe?
VQA: How many people are there in the picture?
Detection: What are the objects in the image?
Image Infilling: What is the image in the middle part?
Text Infilling: What is the complete text of "A <mask> woman" ?
+
Masking
+
OFA
Tasks
Visual Grounding
Grounded Captioning
Image-Text Matching
Image Captioning
Visual Question Answering
Object Detection
Image Infilling
Text Infilling
<loc299> <loc126> <loc282> <loc159>
Visualize
Man in white shirt
Yes
Two boys playing frisbee on the grass
Two
<loc187><loc47><loc381><loc74>car<loc299><loc126><loc282><loc159>person ...
Visualize
<img123><img756><img311>...<img521>
Decoder
A beautiful woman
Vision & Language Tasks
Vision Tasks
Language Tasks

Fig. 3 A demonstration of the pretraining tasks, including visual grounding, grounded captioning, image-text matching, image captioning, VQA, object detection, image infilling as well as text infilling

The model achieves this versatility through extensive pre-training on diverse datasets, equipping it with a rich understanding of both visual and linguistic representations. For instance, in tasks like image captioning and visual question answering, OFA can generate contextually relevant responses and descriptions, underscoring its comprehension of visual content and natural language. By utilizing a Transformer-based sequence-to-sequence framework, OFA unifies task and modality representations, allowing it to perform well across different domains. This unified approach not only enhances performance on individual tasks but also enables seamless integration and interaction between different modalities, paving the way for more sophisticated AI systems.

Methodology

Datasets

SDDBM2 is a dataset of 2 million images produced by Stable Diffusion and their associated text prompts [8]. Due to hard drive constraints, we trained on 1000 of these image-prompt pairs and produced bizarre results. This may be due to the small dataset used and the fragmented nature of the prompts. Many prompts are lists of unlikely pairings of adjectives and nouns. Because of the unusual content and structure of the prompts, we believe training could have improved with more examples and epochs.

COCO is a dataset of about 80,000 image and caption pairs, of which we used 2000 [2]. We hoped the more typical sentence structure of the captions would make fine tuning our pre-trained GPT-2 model faster.

Models

We based the architecture of our first model on BLIP-2. To encode input images, we used OpenAI's pretrained "clip-vit-base-patch32" model. We passed the image embedding output from the CLIP model into a transformer which generated tokens, representing the image's most prominent features, that could be interpreted by an LLM. We then input the tokens into a pre-trained GPT-2 model to generate a coherent caption (i.e. prompt) for the image.

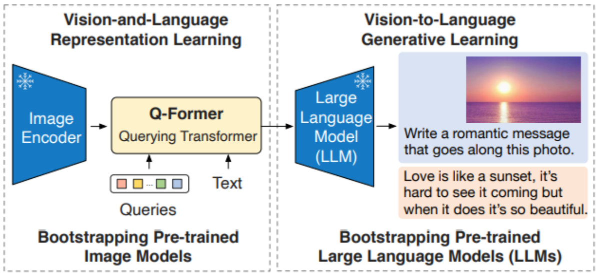
Vision-and-Language
Representation Learning
Image Encoder
Q-Former
Querying Transformer
Text
Queries
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Large Language Models (LLMs)

Fig. 4: BLIP-2 architectural diagram [3]

In training, we froze the CLIP model and all but the last two layers of the GPT-2 model. Thus, our training focused on teaching the bridging transformer and fine tuning GPT-2 to output sentences similar to those in our diffusion prompt and caption datasets. We calculated the loss from GPT-2's output sentence and the datasets' caption/prompt using negative log-likelihood.

Our hyperparameter experimentation was limited because we ran our code in Kaggle and quickly ran out of memory. Batches of size 256 and 128 exceeded our CUDA limit. With the given memory constraints, we were able to successfully train with a batch size of 64. This smaller batch size increased the amount of time required for our model to converge. Thus, we experimented with larger numbers of training epochs (30 and 50), expecting the results to improve with more epochs.

The second model we explored in our research is the CLIP Interrogator. We utilized a pret-rained CLIP model, ViT-H-14, alongside a pretrained BLIP model, model_large_caption, to facilitate our experiments. Additionally, we employed Sentence Transformers as the embedding model to generate text embeddings.

We subsequently developed a custom interrogator function, which utilized the pre-trained models to generate descriptive prompts from the images. This approach allowed us to efficiently and accurately produce contextually relevant captions, demonstrating the efficacy of our modified similarity measure in practical applications.

In addition, we tested a pre-trained OFA model. Here, the inputs consisted of images and the textual query, "What does the image describe?" The text was tokenized using the OFA Tokenizer. Sentence Transformers were again employed for embedding generation and for processing the submission data. This approach enabled us to leverage the robust capabilities of the OFA model in generating comprehensive descriptions based on the given images and textual inputs, further enriching our comparative analysis of various model performances.

Results

Below are the plots representing training and test losses for our CLIP + transformer + GPT-2 models. The model's test loss reached a minimum at epoch 12 so we used the checkpoint from this epoch to test the output of the model. Our final cosine similarity score is 0.0059 for the model trained on the COCO dataset and 0.0695 for the model trained on the SDDBM2 dataset. This appears to be inconsistent with the sample output prompt shown in fig. 7, where training on the COCO dataset promoted higher quality prompt generation.

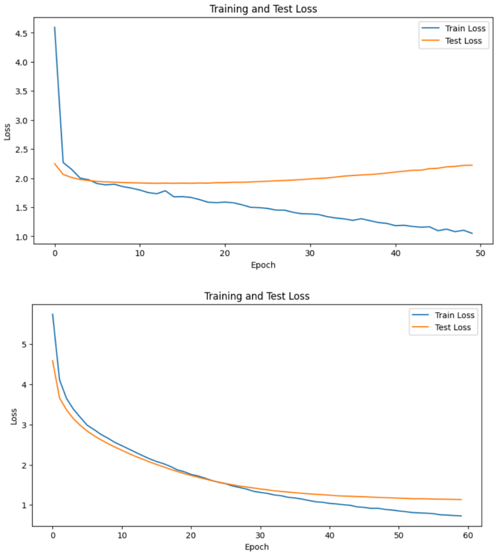

Fig. 5 : Training and test loss curves using the CLIP + transformer + GPT-2 model. The top plot was trained on the COCO dataset, while the bottom plot wastrained on the SDDBM2 dataset.

We tested our models on an image of a seal (fig. 6) generated by a stable diffusion model. The chart in fig. 7 shows the prompt each of our models generated for the seal image, the cosine similarity score for the generated prompt and the actual prompt, and the model's score in the Kaggle competition.

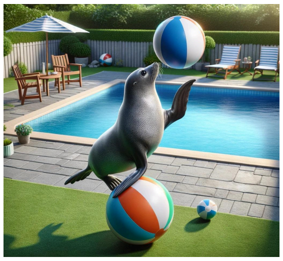

Fig. 6: Image of a seal generated by Stable Diffusion. Input prompt was "photo of a seal balancing on a beach ball while balancing a beach ball on its nose in someone's backyard by an above ground swimming pool"

Model	Generated caption	Cosine similarity score	Kaggle score
CLIP + transformer + GPT-2 (COCO dataset)	"baby and a dog sitting on top of a couch."	0.0059	0.157
CLIP + transformer + GPT-2 (SDDBM2 dataset)	"mits of dishonor, dishonor, dishonor, dishonor, dishonor, dishonor, dishonor, dishonor, dishonor, dishonor, and dishonorable character, in the light of the truth, in the light of the truth"	0.0695	0.063
CLIP Interrogator	"a seal playing with a beach ball near a pool, a photorealistic painting, photorealism, global illumination. vfx, photorealistic cgi, ultra realistic 3d illustration"	0.5476	0.458
OFA	"a seal balancing on a ball in a backyard"	0.787	0.426

Fig. 7: Comparison of models

As shown in the chart above, our CLIP + transformer + GPT-2 models performed very poorly, especially when compared to the CLIP Interrogator and OFA models.

Conclusion

In this project, our objective was to generate a matching prompt given an image produced by Stable Diffusion. While our models successfully generated text from images, the generated text did not correlate well with the images. We believe that this shortfall is primarily due to the small size of our training dataset. Models like OFA and the CLIP Interrogator are trained on millions of image-text pairs, whereas Kaggle's memory limitations constrained us to a maximum of two thousand image-text pairs.

In the "Stable Diffusion – Image to Prompts" competition [7], the highest achieved score was 0.66. Our model attained a maximum score of 0.458, indicating significant room for improvement. Future research will focus on expanding our dataset and exploring more sophisticated architectures to enhance the model's performance.

References

[1] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022). *OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework.* Retrieved from https://arxiv.org/pdf/2202.03052

[2] Awsaf49. (2023). *COCO 2017 Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset

[3] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. Retrieved from https://arxiv.org/pdf/2301.12597

[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. Retrieved from https://arxiv.org/abs/2103.00020

[5] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[6] Tran, T. (2023, April 26). Diversify Photo Database with CLIP Interrogator. Medium. Retrieved from https://trungtranthanh.medium.com/diversify-photo-database-with-clip-interrogator-5dd1833be9f5

[7] Kaggle Competition: "Stable Diffusion: Image to Prompts". Available at: https://www.kaggle.com/competitions/stable-diffusion-image-to-prompts/overview.

[8] PoloClub. (2023). DiffusionDB [Data set]. Hugging Face. Retrieved from https://huggingface.co/datasets/poloclub/diffusiondb

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. arXiv. https://arxiv.org/abs/1406.2661

[10] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. arXiv. https://arxiv.org/abs/2112.10752