Referring Image Segmentation via Text-to-Image Diffusion Models

Final Project Report

Mahdi Ranjbar
mahdi.ranjbar@ens-paris-saclay.fr

January 15, 2024

Abstract

Diffusion models, notably text-to-image types, have revolutionized generative modeling with their high-level synthesis capabilities, enabled by extensive vision-language pretraining. In this report, I present an analysis of semantic information of a text to image diffusion model pretrained on large-scale image-text pairs for referring image segmentation task.

1. Introduction

Large text-to-image diffusion models can implicitly learn both high-level and low-level visual concepts from massive image-text pairs. Using VPD (Visual Perception with Pretrained Diffusion Models Visual Perception Diffusion) [2], the idea is to exploit the pre-trained knowledge in the denoising UNet to provide semantic guidance for downstream visual perception tasks.

The main focus of this project is the task of referring image segmentation which aims to find the related object given a natural language expression from an image.

My project contribution is divided in three studies:

- I reproduce the results of the pretrained VPD on the RefCOCO dataset
- I investigate and modify the model's architecture
- I analyse the impact of noise on the performance of the model

2. VPD

The Visual Perception Diffusion (VPD) framework is designed to address the challenges in transfer learning for diffusion processes. This includes overcoming issues like the incompatibility between standard diffusion pipelines and visual perception tasks, as well as architectural disparities between UNet-like diffusion models and conventional visual backbone models.

2.1. Latent Diffusion Model

The Latent Diffusion Model (LDM) marks a significant advancement in the field of diffusion models, especially for text-to-image transformations. LDM's uniqueness stems from its method of converting images into a more manageable and compact format known as the latent space. This conversion is critical because the diffusion process is carried out within this latent space. The primary benefit of this approach is the significant reduction in complexity and computational requirements compared to traditional diffusion methods.

Central to the LDM is an autoencoder, which efficiently compresses image data into this latent form. Once compressed, the diffusion process introduces noise into the condensed image in a controlled manner, essential for the generation of new images. This method forms the core of the VPD framework, providing an effective means of integrating and interpreting visual and textual data.

2.2. VPD Architecture

Instead of using the step-by-step diffusion pipeline, VPD proposes to simply employ the autoencoder as a backbone model to directly consume the natural images without noise and perform a single extra denoising step with designed prompts to extract the semantic information. This is based on Stable Diffusion [1] models, which conduct the denoising process in a learned latent space with a UNet architecture. Figure 1 represents the architecture of denoising autoencoder backbone.

VPD extracts features from different hierarchies from the UNet decoder to construct visual representations of the input image and feed it plus cross attention maps to various visual decoder to perform visual perception tasks. It also uses a text adapter to refine the text features. The overall framework of VPD is depicted in Figure 2.

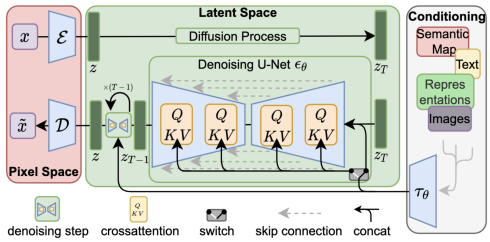

Figure 1. Latent Diffusion Model architecture.

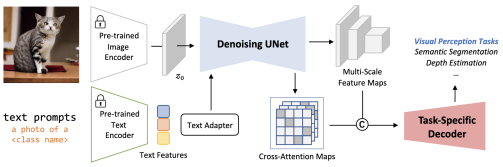

Figure 2. The overall framework of VPD.

2.3. Metric

The model performance is measured by Intersection over Union (IoU), MIoU and OIoU [4]. These are calculated as follows:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

$$\text{Mean IoU} = \frac{1}{N}\sum_{i=1}^{N}\text{IoU}_i$$

$$\text{Overall IoU} = \frac{\sum_{i=1}^{N}\text{Area of Intersection}_i}{\sum_{i=1}^{N}\text{Area of Union}_i}$$

3. Dataset

The dataset used in this project is RefCOCO [3] consist of $\sim$ 42000 training examples, 3800 validation examples and 3700 test examples. Key features of the RefCOCO include:

- **Images:** A collection of images with a wide range of objects in various contexts.
- **Referring Expressions:** Each image comes with multiple referring expressions that describe objects in the scene.
- **Object Annotations:** Precise object bounding boxes that enables model to learn the visual grounding of language.

4. Experiments

4.1. Pretrained Checkpoint

I evaluated the fine-tuned model of VPD on the validation set in RefCOCO dataset at distinct precision at k (P@k) thresholds and noticed that the obtained results match approximately the published results in the paper. The quantitative results are represented in Table 1.

Model	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	OloU	MIoU
VPD	85.52	83.02	78.45	68.53	36.31	73.46	75.67
Inf	84.74	81.71	76.58	66.25	33.34	72.11	74.55

Table 1. Comparative results between pretrained checkpoint in inference and published results in the paper.

4.2. Training Strategy

After evaluating the pre-trained model, I planned to train the model from the scratch, but I ran into memory issues that made it impossible to do so. To solve this, I made a few important changes. First, I decided to stop any further training to the whole SD backbone by freezing it. I also took out the text adapter as it seemed not a crucial part and even I doubted not correctly implemented.

Another important step I took was to change the CLIP model's pre-trained checkpoint from the large version to the base model.

4.2.1 Training Parameters

In order to avoid memory issues the whole set up was scaled down.

Parameter	Value
Loss	Cross Entropy
Batch size	1
Optimizer	Adam
Number of workers	0
Learning rate	0.00005
Train dataset length	5000
Validation dataset length	500
Test dataset length	5000

Table 2. Training Parameters

4.2.2 Adding Noise

As in the original paper the experiment are done without adding noise, I decided to extend my exploration by adding 2 different noise scales T=5 and T=500. The noise is added once the image is in the latent space, right before entering the denoising Unet.

5. Results

5.1. Without Noise

Table 3 represents the result of the trained model without taking into account any noise. It's worth nothing that

the model takes approximately 3 hours for each epoch and the results are not satisfactory because the model is highly scaled down.

Epoch	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	OloU	MIoU
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.37	0.09	0.00	0.00	0.00	1.68	1.96
10	3.14	1.95	0.58	0.21	0.11	3.68	4.10

Table 3. Training results of the model after 1, 5 and 10 epochs.

5.2. With Noise

I implemented two noise scaled T=5 and T=500 and trained the model from scratch for 5 epochs. In appendix an example of image with noise schedule is shown.

Noise	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	OloU	MIoU
0	0.37	0.09	0.00	0.00	0.00	1.68	1.96
5	0.22	0.05	0.00	0.00	0.00	1.21	0.83
500	8.59	6.57	5.19	3.61	1.93	4.45	5.28

Table 4. Training results of the model with different noise scales.

Based on the table 4, it can be generalized that adding noise can help improve the performance of the model. (At least in this scaled down set up with low number of epochs.)

6. Conclusion

This report has thoroughly investigated the capabilities and limitations of a text-to-image diffusion model, specifically focusing on the referring image segmentation task. The experiments conducted using the VPD framework on the RefCOCO dataset provide valuable insights into the adaptability and efficiency of pre-trained diffusion models in understanding and processing visual and textual data.

6.1. Key Findings

The key findings of this study can be summarized as follows:

- **Reproduction of Pretrained VPD Results:** The successful replication of results using the pretrained VPD on the RefCOCO dataset demonstrates the model's robustness and reliability. This replication can be used a solid foundation for further exploration and modification of the model.
- **Architectural Modifications:** The adjustments made to the model's architecture, such as freezing the SD backbone and removing the text adapter, were essential in managing resource constraints. These changes, while necessary, also highlighted the importance of computational resources in working with such advanced models.
- **Impact of Noise:** The experiments with varying levels of noise introduced interesting results. Particularly, the noise levels at T=500 showed a marked improvement in the model's performance. This suggests that adding noise is can be used beneficially in diffusion models.

7. Challenges

This project encountered several significant challenges, impacting the results and offering insights for future research:

- **Extended Epoch Duration:** The complexity of the VPD model, combined with limited computational resources, led to extremely long training epochs. This issue slowed down the process of model tuning and experimentation.
- **Model Heaviness and Code Adaptation:** Due to its size and complexity, the VPD model was computationally heavy and difficult to adapt. Code modifications for specific tasks were challenging, given the intricate model architecture and dependency to other projects.
- **Data Limitation Impact:** Using a small subset of the RefCOCO dataset limited the model's learning capacity. Optimal performance of large-scale models like VPD relies on extensive data, and a reduced dataset hampers this, leading to poorer results.
- **Single-threaded Processing:** The choice to use 0 workers for data loading meant not using multi-core processors for parallel processing, resulting in increased training times.

These challenges underscore the importance of resource optimization, data sufficiency, and efficient code adaptation in developing and applying advanced generative models.

8. Future Perspectives

Moving forward, several avenues for further research and development are evident:

- **Exploring Architectural Innovations:** Investigating more efficient and resource-light architectures that maintain high performance would be beneficial, especially for applications with limited computational capabilities.
- **Extended Noise Exploration:** Further studies on the role and optimization of noise in diffusion models could lead to significant improvements in model performance and generalizability.
- **Broader Dataset Application:** Applying the modified models to a wider range of datasets could provide deeper insights into their versatility and adaptability across different contexts and applications.

References

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. 1

[2] Zhao, Wenliang, et al. "Unleashing text-to-image diffusion models for visual perception." arXiv preprint arXiv:2303.02153 (2023). 1

[3] Yu, Licheng, et al. "Modeling context in referring expressions." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016. 2

[4] Rezatofighi, Hamid, et al. "Generalized intersection over union: A metric and a loss for bounding box regression." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. 2

Appendix

A. Results for Pre-trained Model

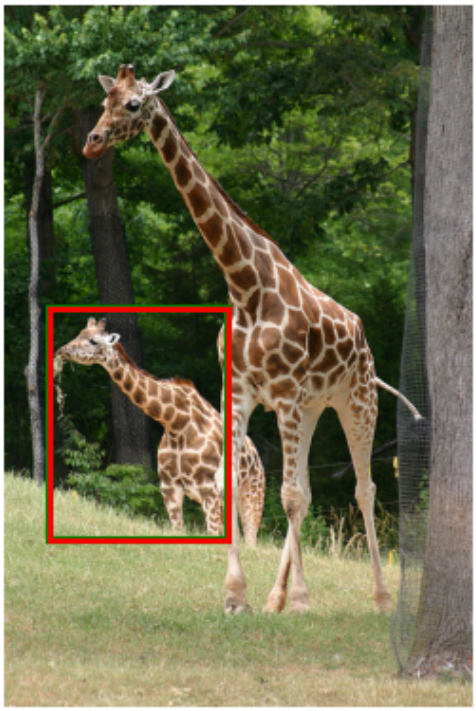

Figure 3. The referring sentence is *"the giraffe on the left"*. It can be seen that IoU is 1 and the target and predicted box overlaps.

B. Results for Trained Model

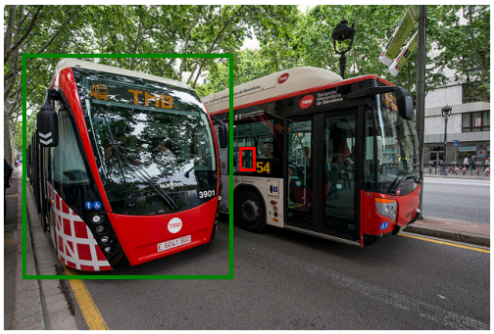

Figure 4. Example of the performance of the model after 10 epochs of training. It can be noted that the IoU is 0 and the performance is not good at all. Green is the target and red is the prediction of the model.

C. Noise Schedule

Figure 5. T=0

Figure 6. T=5

Figure 7. T=500