CVPDL HW3 r12631001 許喬淇

Relevant Papers

- BLIP-2

I used BLIP-2 (Bootstrapping Language-Image Pre-training), a vision-language model released in 2023. It uses a Q-Former architecture that connects vision and language models, achieving good performance while requiring fewer parameters than previous models. The model performs well in several tasks including image-to-text generation, visual question answering, and image-text matching.

Pre-trained on large-scale image-text pairs, BLIP-2 can work together with language models like OPT and T5. These capabilities make it useful for applications that need to process both visual and text information.

- Grounded Language-Image Generation

In my project, I used GLIGEN (Grounded Language-Image Generation), which is a text-to-image generation model built on top of Stable Diffusion. GLIGEN adds the capability to control object placement through bounding boxes, allowing users to specify both the content and location of objects in the generated image.

The model's architecture incorporates grounded language features to understand spatial relationships and object positions. This functionality makes GLIGEN suitable for creating complex scenes where layout control is important. While other text-to-image models focus solely on content generation, GLIGEN maintains accurate object placement while producing natural-looking images.

Image Captioning and Prompt Design

A. **Compare the performance of 2 selected different pre-trained models in**

generating captions, and use the one you find the most effective for later problems.

blip2-opt-6.7b-coco相較其他OPT模型有較多的參數量，並在coco資料集上進行微調，blip2-flan-t5-xl則是基於Flan T4-xl訓練。由於此兩權重檔是基於不同的語言模型進行訓練，故選擇Salesforce/blip2-opt-6.7b-coco和Salesforce/blip2-flan-t5-xl比較。

為了比較兩者，實驗方法如下: 1）首先建立9種不同的prompt 2) 選擇分別生成最好的描述 3）使用描述由Gligen生成圖片。

1) 九種不同的prompt:
    a) A very detailed description:
    b) A very detailed description of the people and the actions:
    c) A detailed description of objects and actions:
    d) List the protected body parts and their corresponding safety equipment visible in this image:
    e) List safety equipment
    f) What is the scene of the photo? What are the people doing in the photo?
    g) What is the scene of the photo? What are the people doing or the item appeared in the photo?
    h) What is the scene of the photo? What are the people doing or the item and tool appeared in the photo?
    i) What is the scene or background of the photo? What are the people doing or the item and tool appeared in the photo?
2) 最好的描述是根據生成描述— a.不會重複設計的promt b.不會在描述一張圖像時不斷出現冗言贅字(像是口吃) c.不會生成空白描述—的能力去選擇，而過程中是由人工進行篩選。

    以下範例為重複設計的promt: What is the scene of the photo. What are the people doing in the photo? what is the scene of the photo. (o

f image pexels-photo-4904559.jpeg)

以下範例為重複詞語:a pink wall, a blue wall, a green wall, a white wall, a pink wall, a blue wall, a green wall, a white wall, a pink wall, a blue wall, a green wall, a white wall, a pink wall, a blue wall, a green wall. (of image pexels-photo-7659778.jpeg)

在觀察描述時，發現blip2-opt-6.7b-coco對於what和where的提問形式難以理解，多數產生空白描述;同時，若請求列出人與物，也會產生不自然的描述。而Salesforce/blip2-flan-t5-xl對於提問則有較優秀的表現，對於prompt有較加的語意理解性。

因此，對於blip2-opt-6.7b-coco使用第1.a個prompt生成描述; 對於Salesforce/blip2-flan-t5-xl使用第1.i個prompt生成描述。

3) 使用兩模型生成之描述由Gligen生成圖片，其中將Gligen的參數—gligen_scheduled_sampling_beta—

設為0，以不使用到其他類型的grounding input，並使用FID評估，結果如下。

Blip2 Model	blip2-opt-6.7b-coco	Salesforce/blip2-flan-t5-xl
Prompt	A very detailed description.	What is the scene or background of the photo? What are the people doing or the item and tool appeared in the photo?
FID Score	51.68	59.42

根據結果，在兩模型皆選擇分別較有效的prompt的情況下，blip2-opt-6.7b-coco有較好的表現，其FID Score達到51.68。因此在接下來的題目中，選擇blip2-opt-6.7b-coco進行討論。

B. Design templates of prompts for later generating comparison.

a. A very detailed description.

b. A very detailed description, high quality, highly detail and photorea listic.

c. A very detailed description. + --negetive prompt = " bad anatomy, bad hands, missing fingers."

Text-to-Image Generation

	Text grounding			Layout-to-Image
Prompt	Template #a	Template #b	Template #c	Template #a
FID	**51.68**	52.31	54.64	77.01