Computer Vision 2 Project

An improvement on “Guiding Text-to-Image Diffusion Model Towards Grounded Generation” by Z. Li et al.

Ankit, Walter Simoncini, Ioanna Gogou, Didier Merk, Tsatsral Mendsuren, André de Brandt

Motivation

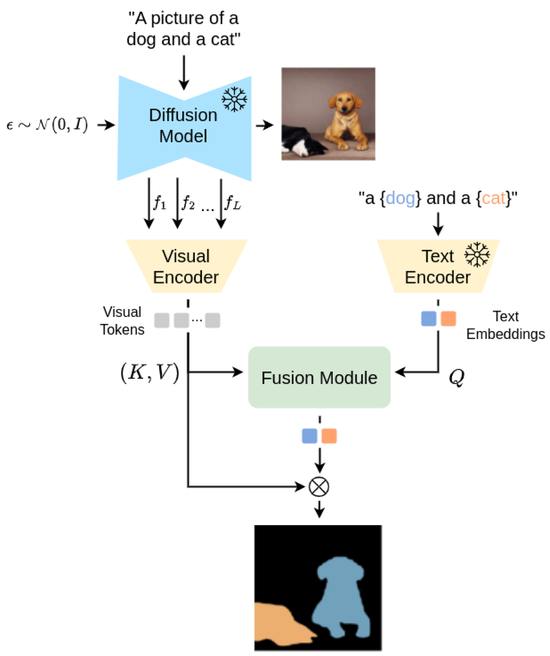
Grounded Generation - Z. Li et al. use Stable Diffusion to generate images of objects from text prompts, and propose a model to segment objects from visual and text embeddings.

Problem - Segmenting multiple objects in the same image produces imprecise masks.

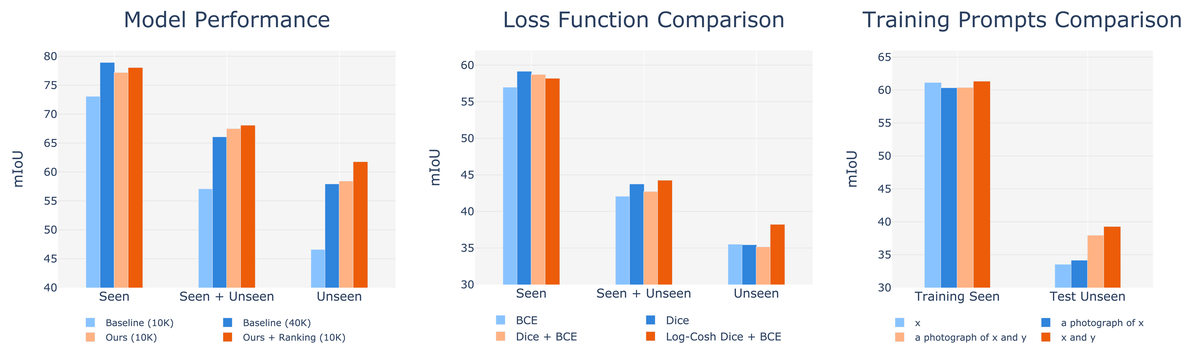
Contributions

1. Compared generations and grounding quality of Stable Diffusion v2, v1.5 and its fine-tuned checkpoint.
2. Enhanced the fusion module training via prompt engineering.
3. Boosted mIoU dual-object image segmentation by training using log-cosh loss, dropout, and multiple epochs.

Architecture



Quantitative Results



Qualitative Results

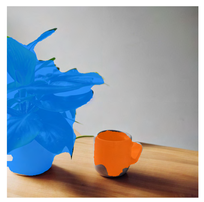
The Good

The Bad

The Ugly

a photograph of a **basketball** with a **cat** standing on top of it on a field long shot

a photograph of a **dog** sitting on top of a **boat** sailing on a river, lit by a warm natural light long shot



a photograph of a **mug** and a **potted plant** side by side sitting on top of a table in a kitchen with strong natural light

a photograph of a red **car** and an **eagle** flying in the sky

BoW Behavior

a photograph of a **mug** and a **pottedplant**

a photograph of a **mug** and a **potted plant**

Limitations

1. Training assumes target objects are in the image. Otherwise the mask will cover the whole image
2. Generating images with more objects is possible but quality decreases.
3. CLIP embeddings behave as a bag of words, so for a prompt "x and y" the masks might overlap.

Methodology

Dataset - We generated a dataset of 10k images with corresponding masks from Mask R-CNN, each containing 2 objects selected at random between 15 classes of Pascal VOC.

Prompt Engineering - We evaluated different prompts for training: “x”, “x and y”, “a photograph of x and y”, “a photograph of x”.

Stable Diffusion - We updated the Stable Diffusion (SD) model from v1 to v1.5. We also assessed the performance of SD v2, with worse results. The fine-tuned checkpoint of v1.5 showed promising results.

Training - Log-cosh dice loss + BCE loss, induced dropout regularization and trained models for 10 epochs, unlike Z. Li et al.

$$L_{\text{BCE}}(m_i^{\text{gt}}, m_i) = -\frac{m_i^{\text{gt}} \log(m_i) + (1 - m_i^{\text{gt}}) \log(1 - m_i)}{N}$$

$$L_{\text{DICE}}(m_i^{\text{gt}}, m_i) = 1 - \frac{2 m_i^{\text{gt}} m_i + 1}{m_i^{\text{gt}} + m_i + 1}$$

$$L(m_i^{\text{gt}}, m_i) = \log\Big(\cosh\big(L_{\text{DICE}}\big)\Big) + L_{\text{BCE}}$$

Testing - three different setups:

1. Seen: only objects from classes seen in training
2. Seen+Unseen: one object class seen in training, one unseen
3. Unseen: two objects, neither class seen in training

Experimental Setup

The model was trained using 10k images with two objects for 10 epochs, using a 0.1 dropout regularization. The best model was selected using a validation set of 1k images.

Testing was done on 1k images for the seen, seen+unseen and seen+unseen setups. We also tested on the best 800 images by quality as scored by CLIP-IQA+ to emulate how Z. Li et al. manually remove bad generations from the test set