Creating a Captioner for Synthetic Data Annotation

Director de these:
Damien Rohmer, LIX, Ecole Polytechnique, IP Paris (damien.rohmer@polytechnique.edu)
Cosupervisors:
Vicky Kalogeiton, LIX, Ecole Polytechnique, IP Paris (vicky.kalogeiton@polytechnique.edu)
David Picard, IMAGINE, ENPC (david.picard@enpc.fr)

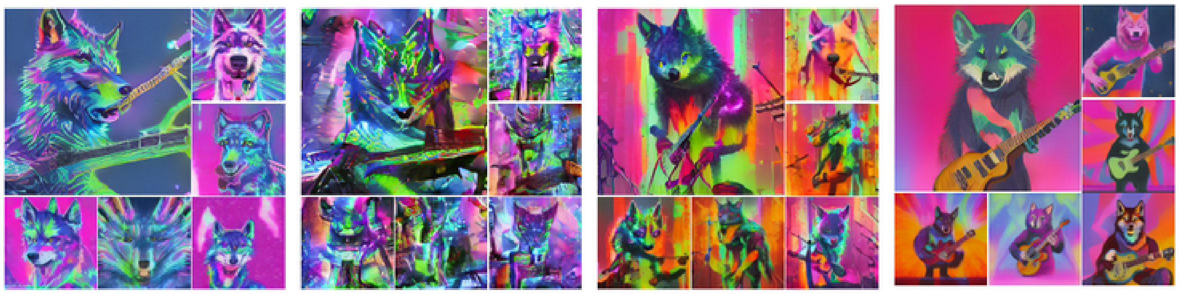
Date: October 2024 - September 2027

Employer : these COFRA, Ecole polytechnique, LIX et Le ministère des Armées

October 1, 2024

Introduction

Text-guided models have revolutionized computer vision in recent years, with CLIP [2] demonstrating the effectiveness of leveraging noisy Web-scraped data to train robust image representations. CLIP is trained in a weakly supervised manner on alt-text text data associated with images found on the web. This large-scale approach allows the network to filter out noise and become highly versatile. Text conditioning has also been successfully applied to generative models such as text-to-image [3, 4, 5, 6] and text-to-motion [1] models. This conditioning strategy has proven to be not only effective for training but also practical for downstream users. Artists have developed prompt engineering techniques to create beautiful artworks using text conditioning.



Prompt: *A vibrant, multicolored furry wolf with neon highlights playing an electric guitar on stage; trending on artstation*

Figure 1: Example of artistic images generated by text-to-image diffusion models

Challenges

Despite its advantages, the noisy nature of alt-text data poses several challenges. The most significant challenge is the requirement of massive datasets, often exceeding a billion images, which can be computationally expensive to process. This has limited the development of text-to-image models to a few industrial labs with access to extensive computing resources. Recently, [9] showed that synthetic captions can be used to improve the performance of CLIP models, indicating that synthetic data can be competitive with alt-text data. Synthetically captioned images have also been shown to improve caption understanding and data efficiency in text-to-image generation [7, 10, 11].

Research Objectives

This thesis aims to address these challenges and advance the development of efficient text-conditioned models. (1) The primary objective is to leverage visual language models (VLMs) to create a more effective captioner capable of generating rich and diverse captions. Modern available captioners, such as BLIP 2 [8], are limited in their ability to capture artistic information and provide multiple, diverse textual descriptions (captions) of the same image, each capturing different aspects or perspectives of the visual content. However, VLMs offer the potential to overcome these limitations. Our first goal is to utilize VLMs to extract relevant information from images and enrich coarse captions. This will require careful safeguards to avoid losing information from the alt-text while minimizing the introduction of synthetic noise. (2) The second objective is to investigate the impact of synthetic captions on the training of text-to-image diffusion models. We will explore how synthetic captions affect data efficiency, whether they can be used for data augmentation, and how caption evolution should be adapted to image augmentation. Additionally, we will study the relationship between caption precision and diffusion model performance. To address these questions, we will train text-to-image models aiming to achieve state-of-the-art performance while minimizing data requirements.

Proposed plan

Modern text-to-image models lack reproducibility because of the decay of web-datasets. With a good captioner, one could leverage standard computer vision datasets, such as Imagenet, Imagenet 21K, SAM dataset, YFCC100M, Places365, and OpenImages. Furthermore, one could also enhance the specificity of these datasets. This can be achieved, for instance, by incorporating ImageNet 21K label names into the captioning process and expanding the vocabulary of the dataset. Domain-specific data sets such as OSV-5M for geographical data, iNaturalist for biological data, and the PlantNet dataset could be integrated. This approach aims to enable the utilization of any data bank that lacks captioning.

Currently, augmenting image datasets may lead to mismatches between the image and its caption (cropping for instance). To mitigate this, the idea is to implement recaptioning on-the-go during training. This approach would enable various types of data augmentation while ensuring consistency between images and captions. Generating multiple captions per image and alternating between them during training could also be a way to enhance the robustness and diversity of the model.

The study of the importance of good captions aims to derive inverse scaling laws to demonstrate that less data are required to achieve similar results with good captions. It asks whether high-resolution captions are essential during diffusion training and explores the concept of curriculum learning in this context. In addition, the study examines the comparative challenges posed by incorrect captions versus concise captions.

Expected Outcomes

This thesis will result in the development of a novel captioner that utilizes VLMs to generate more informative web dataset. This captioner will be evaluated for its effectiveness in improving the training of text-to-image diffusion models, particularly in terms of data efficiency and semantic accuracy. The findings of this research will contribute to the advancement of text-conditioned models and pave the way for more efficient and versatile image generation techniques.

References

[1] Mathis Petrovich, Michael J. Black, and Gül Varol. *Action-Conditioned 3D Human Motion Synthesis with Transformer VAE*. 2021. arXiv: `2104.05670 [cs.CV]`.

[2] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: `2103.00020 [cs.CV]`.

[3] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: `2204.06125 [cs.CV]`.

[4] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models.* 2022. arXiv: 2112.10752 [cs.CV].

[5] Chitwan Saharia et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.* 2022. arXiv: 2205.11487 [cs.CV].

[6] Yogesh Balaji et al. *eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers.* 2023. arXiv: 2211.01324 [cs.CV].

[7] Junsong Chen et al. *PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis.* 2023. arXiv: 2310.00426 [cs.CV].

[8] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.* 2023. arXiv: 2301.12597 [cs.CV].

[9] Thao Nguyen et al. *Improving Multimodal Datasets with Image Captioning.* 2023. arXiv: 2307.10350 [cs.LG].

[10] Eyal Segalis et al. *A Picture is Worth a Thousand Words: Principled Recaptioning Improves Image Generation.* 2023. arXiv: 2310.16656 [cs.CV].

[11] James Betker et al. “Improving Image Generation with Better Captions”. In: URL: https://api.semanticscholar.org/CorpusID:264403242.