CVPDL HW 3: Data Augmentation

R13946003 李承彥

Abstract

This study delves into the application of data augmentation in image generation, focusing on the systematic analysis of caption-based templates and layout integration. By utilizing BLIP2 models to generate precise image descriptions and combining them with diffusion models for image generation, this study evaluated the impact of different BLIP models, template designs, and layout integrations on the quality of generated images. The experimental results demonstrate that the choice of models and methods significantly affects the Fréchet Inception Distance (FID), particularly when GLIGEN layout integration is employed, leading to notable performance improvements. This study provides valuable guidance for constructing more efficient data augmentation pipelines.

Introduction

Data augmentation is a critical technique for enhancing machine learning model performance by introducing data diversity to improve generalization. In the field of image generation, the combination of modern description generation models and diffusion techniques can significantly enhance the quality and realism of generated images. This study focuses on the following objectives:

- Investigating the performance differences of various BLIP2 model configurations in description generation.
- Analyzing the role of template design in determining the quality of generated images.
- Evaluating the impact of layout integration on the realism and fidelity of generated images.

This study utilized "Salesforce/blip2-opt-2.7b" and "Salesforce/blip2-opt-6.7b" as description generation models and employed "stable-diffusion-v1-5/stable-diffusion-v1-5" and "masterful/gligen-1-4-generation-text-box" for image generation. The

generated results were evaluated using FID as the primary metric.

Experimental Workflow

In the experimental workflow, this study followed the steps. **1) Model Selection and Comparison:** Generate descriptions using two BLIP models and evaluate their performance based on FID scores calculated from 2,160 generated images, selecting the best-performing model. **2) Template Design Experiments:** Using the selected BLIP model, design two different templates to generate 2,160 images for each and compare their FID scores. **3) Layout Integration:** Building upon the best-performing template, integrate the GLIGEN model with layout design (including bbox parameters) to generate 2,160 images and evaluate their quality.

Experimental Results

Experiment 1: BLIP Model Comparison

This experiment compared the quality of images generated using Stable Diffusion from descriptions produced by two BLIP models:

- Model 1: "Salesforce/blip2-opt-2.7b"
- Model 2: "Salesforce/blip2-opt-6.7b"

Template Settings

- **Caption Template (for BLIP model):** "Question: Describe this workplace image with exact details about: workers (clothing, positions), tools, and spatial relationships between elements. Answer:"
- **Enhanced Template (for genreration model):** “detailed photograph in real workplace, photorealistic: {generate from BLIP}”
- **Negative Template (for genreration model):** “blurry, low quality, distorted, unrealistic, cartoon, anime, sketchy, duplicate, double image, mutated, deformed”

BLIP model	FID Score
Salesforce/blip2-opt-2.7b	75.81
Salesforce/blip2-opt-6.7b	75.63

Table 1. Result of Experiment 1

Experiment 2: Template Design Comparison

This experiment explored the impact of different template designs on the quality of generated images using Salesforce/blip2-opt-6.7b for caption generation and Stable Diffusion for image generation.

Template #1: Same as Experiment 1.

Template #2:

- **Enhanced Template (for genreration model):** "detailed photograph in real workplace, photorealistic: {generate from BLIP}"
- **Negative Template (for genreration model):** "blurry, low quality, distorted, unrealistic, cartoon, anime, sketchy, duplicate, double image, mutated, deformed"

Template	FID Score
Template #1	75.63
Template #2	74.73

Table 2. Result of Experiment 2

Experiment 3: Layout Integration

This experiment assessed the impact of integrating GLIGEN's layout design, including bounding box (bbox) parameters, on image generation using Template #2 and the BLIP model "Salesforce/blip2-opt-6.7b." The findings revealed that the optimal combination for achieving the highest quality image generation was utilizing BLIP Model 2 ("Salesforce/blip2-opt-6.7b"), Template #2, and GLIGEN layout integration. This combination resulted in the best FID score of 55.02, demonstrating the significant improvements brought by incorporating structured layout design into the generation process.

Generation model	FID Score
masterful/gligen-1-4-generation-text-box	55.02

Table 3. Result of Experiment 3

Summary of Results

In the model overview, advanced techniques for text grounding and layout-to-image synthesis were employed to evaluate the performance of the BLIP and GLIGEN models. Specifically, text grounding for predefined templates was performed using the BLIP model, “Salesforce/blip2-opt-6.7b,” in conjunction with Stable Diffusion to ensure coherent and contextually rich image generation. For layout-to-image synthesis, the same BLIP model was integrated with GLIGEN to incorporate spatial layout guidance, enhancing the model's ability to generate images adhering to specified spatial constraints.

	Text Grounding		Layout-to-Image
Template	Template #1	Template #2	Template #2 + Layout
FID	75.63	74.73	55.02

Table 4. Summary of Results

Conclusion

Key findings from this study include:

- Impact of Model Configuration: “Salesforce/blip2-opt-6.7b” outperformed "Salesforce/blip2-opt-2.7b" in terms of description accuracy and image quality.
- Template Design: Template 2, focusing on realism, slightly outperformed Template 1 in terms of FID scores.
- Layout Integration: The introduction of GLIGEN significantly improved the quality of generated images, reducing the FID score from 74.73 to 55.02, underscoring the importance of layout design in structured image generation.
- Optimal Workflow: Combining “Salesforce/blip2-opt-6.7b”, Template #2, and GLIGEN layout integration constitutes the most effective generation pipeline.

This study highlights the potential of integrating advanced description generation models with image generation techniques in data augmentation. Future work could explore further optimizations in template design, additional description models, and

more sophisticated layout design techniques to enhance the utility and diversity of generated images.

Reference

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., ... & Lee, Y. J. (2023). Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22511-22521).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).