Term Project Final Report: Personalized Image Generation using Textual Inversion

B09611007 Po-Lin Chen

ABSTRACT

This report discusses the use of deep learning techniques to generate personalized visuals using textual inversion. The goal is to create high-quality images of a specific object, represented by a unique token, with minimal training data. In this study, four images of a Corgi were used to train a special token, enabling the model to generate images of the Corgi in various contexts based on simple and scene-based textual prompts. The model's performance was evaluated using CLIP Image Score and CLIP Text Score, which assessed how well the generated images aligned with the given prompts. The results demonstrate the effectiveness of textual inversion in generating realistic object-specific pictures and highlight the model's strengths and weaknesses in both isolated and contextual environments. Further optimizations are discussed to enhance image generation consistency.

Keywords: Personalization, Diffusion Model, Deep Learning

1. INTRODUCTION

1.1. Textual Inversion for Personalized Images

Textual Inversion is a technique that facilitates personalized image generation by learning a unique embedding for a specific object using only a few images. This embedding is then associated with a special token, allowing the model to generate accurate representations of the object when used in textual prompts.

1.2. Deep Learning Models for Image Generation

Deep learning models for image generation have advanced significantly, particularly with the rise of diffusion models. A key model in this area is Stable Diffusion, built on Denoising Diffusion Probabilistic Models (DDPM) and the more efficient Denoising Diffusion Implicit Models (DDIM). In this study, we use Stable Diffusion to generate personalized images of a Corgi based on minimal training data.

1.3 Objectives

The objectives of this study are:

1. Train a special token using four images of a Corgi to capture its unique features.
2. Use textual prompts to generate images of the Corgi in various contexts.
3. Assess the quality of generated images using the CLIP Image Score and CLIP Text Score.
4. Examine and discuss the performance of multi-concept image generation with another dataset.

2. MATERIALS AND METHODS

2.1 Data Preparation

Four images of a Corgi were selected for training the special token, as shown in Figure 1. The backgrounds were deliberately blurred to minimize distractions and ensure the model focuses on learning the Corgi's distinctive features, such as its shape and fur texture. This approach ensures that the model captures the object's characteristics without interference from environmental factors.

Figure 1. The input images for training the Corgi token.

2.2 Text-to-Image Generation using Textual Inversion

Textual Inversion is a technique that enables a model to generate images of specific objects using a minimal number of training images (typically 4-6). It works by learning a unique embedding (*) for the object, which is then associated with a special token. This embedding captures the distinct visual features of the object. Once trained, the token can be used in textual prompts to generate images of the object in various contexts.

The process allows the model to understand and generate the object's features, even with limited data, by conditioning the image generation process on the learned embedding. This makes Textual Inversion particularly useful for tasks where only a small set of images is available, but high-quality, object-specific image generation is required.

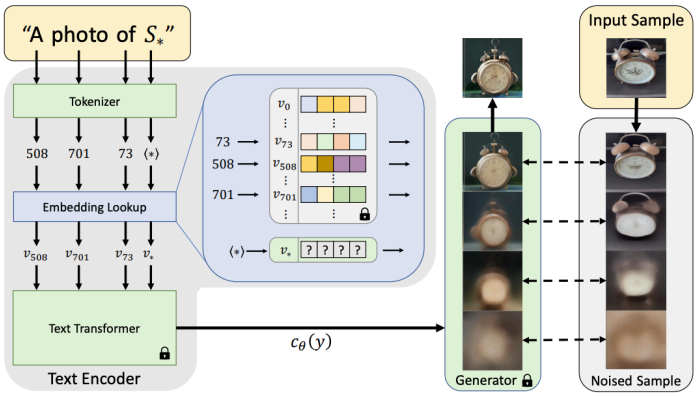

Figure 2. Outline of the text-embedding and inversion process.

2.3 Performance Analysis

The evaluation of generated images is inherently abstract and can be assessed using various standard metrics. In this study, we use the following key indicators to evaluate the quality and relevance of the generated images:

1. **CLIP Image Score**: This score is computed using the CLIP model to assess how well the generated image aligns with the semantic meaning of the corresponding textual description.
2. **CLIP Text Score**: This score evaluates how well the generated image aligns with the given textual prompt. By comparing the generated image's corresponding text with the provided description, the CLIP model measures the coherence between the text and the image content.

For our analysis, we compare the performance under two types of prompts:

1. **Simple prompts**: "An illustration of my *", "An illustration of a clean *", "A close-up photo of the *", and "A photo of a nice *".
2. **Scene-based prompts**: "A * perched on a park bench with the Colosseum looming behind" and "A * posing proudly on a hilltop with Mount Fuji in the background".

These two categories of prompts allow us to evaluate the model's ability to generate images both in isolation and within specific contextual environments.

2.4 Comparison and Optimization

Conduct comparative analyses of the generated results under various conditions, including training hyperparameters, image generation settings, and prompting strategies.

3. RESULTS AND DISCUSSIONS

3.1 Image Generation with Simple Prompts

Figure 3 shows the generated images using simple prompts: "An illustration of my *", "An illustration of a clean *", "A close-up photo of the *", and "A photo of a nice *". The generated images successfully depict the Corgi in various styles. However, there are instances where two Corgis appear in the same image. This may result from the model's inability to fully constrain the number of objects or a result of the random nature of the generation process.

Figure 3. The generated images using simple prompts: "An illustration of my * ", "An illustration of a clean * ", "A close-up photo of the * ", "A photo of a nice * ", respectively.

3.2 Performance Analysis of Scene Integration

For the prompt "A * perched on a park bench with the Colosseum looming behind.", the generated image achieved a CLIP Image Score of 73.72 and a CLIP Text Score of 32.94. The model successfully captured the key scene element of the Colosseum looming in the background, aligning with the textual description.

However, as shown in Figure 4, despite the presence of a "park" setting, the "bench" was not depicted in the generated image. This suggests that while the model is effective at integrating larger, more prominent scene elements, it may struggle with smaller, more specific details in the prompt, such as particular objects within the scene.

Figure 4. The generated images using the prompt: “A * perched on a park bench with the Colosseum looming behind.”

For the prompt “A * posing proudly on a hilltop with Mount Fuji in the background.”, as shown in Figure 5, the generated image achieved a CLIP Image Score of 78.98 and a CLIP Text Score of 33.66. The model effectively captured the concept of “Mount Fuji” in the background, and the “hilltop” concept was also clearly represented. However, the color of the Corgi in the image appears to blend with the background, possibly due to the lighter tones of the object and the surrounding environment.

Figure 5. The generated images using the prompt: “A * posing proudly on a hilltop with Mount Fuji in the background.”

3.3 Discussion on Multi-Concept Personalized Image Generation

This section examines the performance of the text-to-image framework in multi-concept scenarios, where at least two special tokens are employed during generation. To analyze its effects, we trained another special token using a dataset of four cat images, as illustrated in Figure 6.

Figure 6. Another dataset to train the special token: #

The generated images for the prompt "A * next to a #" are presented in Figure 7. These outputs reveal significant challenges, including attribute leakage and conceptual confusion, such as the emergence of dog-like shapes and the grey textures of the cat. Additionally, the framework often prioritizes one concept, leading to incomplete or distorted representations. Future work could explore advanced strategies for multi-concept integration, as proposed by Kumari [3] and Yao [4].

Figure 7. The generated images using the prompt: "A * next to a #".

4. CONCLUSION

This study demonstrates the effectiveness of textual inversion for generating personalized images with minimal training data. By training a unique token for a Corgi using just four images, the model was able to generate high-quality, object-specific visuals based on both simple and scene-based textual prompts. The evaluation using CLIP Image and Text Scores revealed that the model effectively captured the key features of the Corgi and integrated contextual elements. However, challenges remain in ensuring consistency and accurately depicting all scene details. Furthermore, this study also discusses the challenges in multi-concept scenarios.

REFERENCES

[1] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618

[2] 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).

[3] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., & Zhu, J. Y. (2023). Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1931-1941).

[4] Yao, Z., Feng, F., Li, R., & Wang, X. (2024). Concept Conductor: Orchestrating Multiple Personalized Concepts in Text-to-Image Synthesis. *arXiv preprint arXiv:2408.03632*.

PROGRAM LINK

https://drive.google.com/drive/folders/1F14904S3G5cJF7OVUHEbl5DKzLH6yOC-?usp=sharing