A Comprehensive Study On Custom Image Generation Using Textual Inversion

{ Benjamín Farías V. } Pontificia Universidad Católica de Chile

DCC Departamento de Ciencia de la computación

Introduction

Recently, large scale text-to-image models have proved to be really capable when it comes to reasoning over natural language descriptions and producing novel images. However, this approach is limited by the user's ability to express the desired target through text. In this work, we study a method named *Textual Inversion* **[3]**, which overcomes this limitation by **finding** new words in the textual embedding space of these models.

Related Work

LDM [1]: *Latent Diffusion Models* work by decomposing the image formation process into a sequential application of denoising *autoencoders*, which work in a latent space of low dimensional vectors.

Textual Inversion [3]: Similar to *DreamBooth* **[2]**, it learns a new representation for a custom concept. However, it only uses *fine-tuning* on the embedding layer of the text encoder, considerably improving efficiency at the cost of some performance.

Method

The *Textual Inversion* method adds a new special token to the embedding layer of the pre-trained text encoder (*CLIP*). Then, it directly optimizes these embeddings, by minimizing the *MSE* loss over images sampled from the training set. To condition the generation, we use prompts that generically describe the concept represented by the special token.

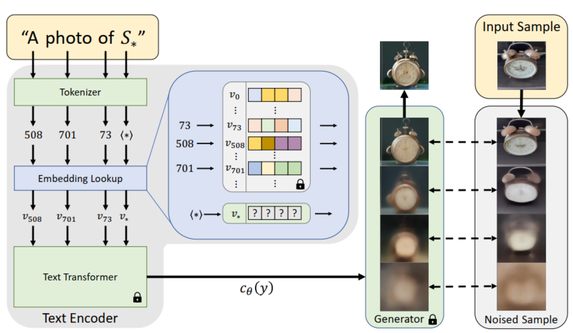

Figure 1: Outline of the *Textual Inversion* process.

Experiments: Qualitative

Figure 2: Ablation: Effect of number of training steps on generated images (0 to 5000 steps, from left to right, top to bottom). The object is a small version of the fictional character *Groot*. **Result:** Ideal number of steps from 1000 to 3500.

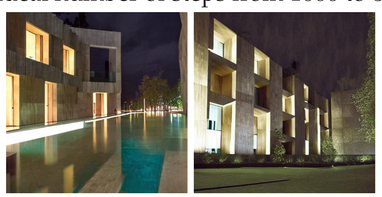

Figure 3: Ablation: Effect of the initial embedding used for training. The object is the *PUC Innovation Center*. On the left, the initial reference is the word *building*, and on the right, the word *tower*. **Result:** Changing the initial reference can add variability.

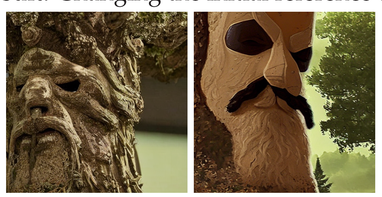

Figure 4: Ablation: Effect of two different randomization seeds on training. The object is a fictional tree called the *Deku Tree*. **Result:** Different seeds can provide substantial changes in appearance.

Figure 5: Comparison between vanilla Stable Diffusion (left) and Textual Inversion (right). The prompt is: *A <deku-tree> tree in the forest.* **Initializer:** *Tree*.

Figure 6: Comparison between vanilla Stable Diffusion (left) and Textual Inversion (right). The prompt is: *A <puc-innovation-center> in the campus.* **Initializer:** *Building*.

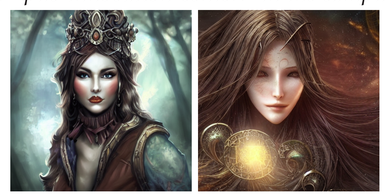

Figure 7: Comparison between vanilla Stable Diffusion (left) and Textual Inversion (right). The prompt is: *A woman in the style of <elden-ring>.* **Initializer:** *Fantasy*.

Figure 8: Comparison between vanilla Stable Diffusion (left) and Textual Inversion (right). The prompt is: *A lake in the style of <elden-ring>.* **Initializer:** *Fantasy*.

Experiments: Quantitative

The following table shows the scores for **Reconstruction** and **Editability**, for both the vanilla *Stable Diffusion* and the *Textual Inversion* model:

Model	Reconstruction	Editability
SD	0.7622	0.2639
SD + TI	0.8077	0.1611

Table 1: Quantitative performance comparison.

Conclusions

From the results, we can see that this method is very effective when it comes to learning new concepts, but **better quality requires sacrificing editability**, due to the model overfitting to the few-shot dataset. Future research could be focused on finding dataset characteristics that improve performance during training, in a way that the model is able to learn the concept without losing editability. Additionally, more experiments could be done in terms of the type of concept to learn.

References

[1] R. Rombach, A. Blattmann, and D. Lorenz et al. High-resolution image synthesis with latent diffusion models. 2021.

[2] N. Ruiz, Y. Li, and V. Jampani et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[3] R. Gal, Y. Alaluf, and Y. Atzmon et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2022.