

Possibility of Disaster Image Through Fine-Tuning

2023.11.09

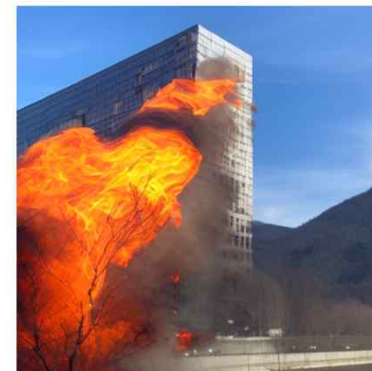
UST-ETRI

Media Intelligence Lab

Minji Choi

cmj@etri.re.kr/ji_min9245@naver.com

Previous Work



Previous Work

[Summary]

- The goal of this research is to generate disaster images using image generation models.
- Among the current image generation models, diffusion-based models exhibit superior performance.
- The most widely used diffusion-based model is the open-source Stable Diffusion.
- Diffusion-based models sometimes struggle to accurately represent the terrain or may omit keywords, making them currently unsuitable for disaster image generation.

[Future Works]

- Collection of domestic terrain image datasets.
- Attempting fine-tuning of image generation models.
- Exploring editing and synthesis techniques rather than image generation.

Contents

1

Introduction

2

Background

3

Fine Tuning

4

Experiment Results

5

Evaluation Results

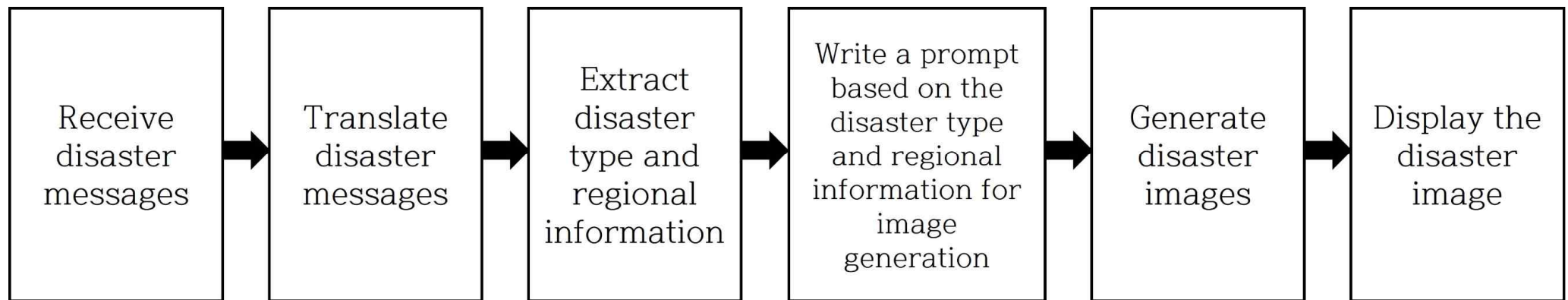
6

Conclusion

1

Introduction

How to create disaster images based on disaster messages



Designing an image creation method based on disaster text messages

It is determined that a Korean-style disaster image creation model is necessary (that can harmoniously represent Korea's topography and types of disasters)

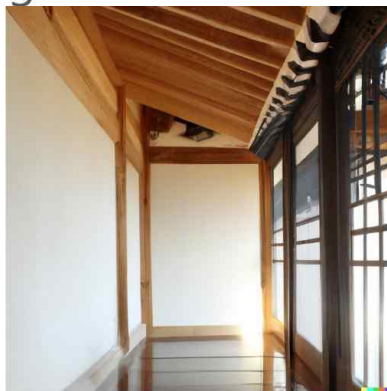
2

Background

Reason for Selecting Training Image Data



Stable Diffusion



DALL·E 2



Playground


Input prompt : Hanok

Since image generation models have been trained on large-scale image datasets, they are expected to represent mountains and rivers relatively well. However, when prompted to generate images of Hanok (traditional Korean houses), the models often produce buildings that resemble traditional Chinese or Japanese houses. This suggests a lack of Hanok data in the existing models. Based on this, the goal is to further train the existing models with Hanok images, analyze the results, and explore the potential for a Korea-specific disaster image generation model.

2

Background

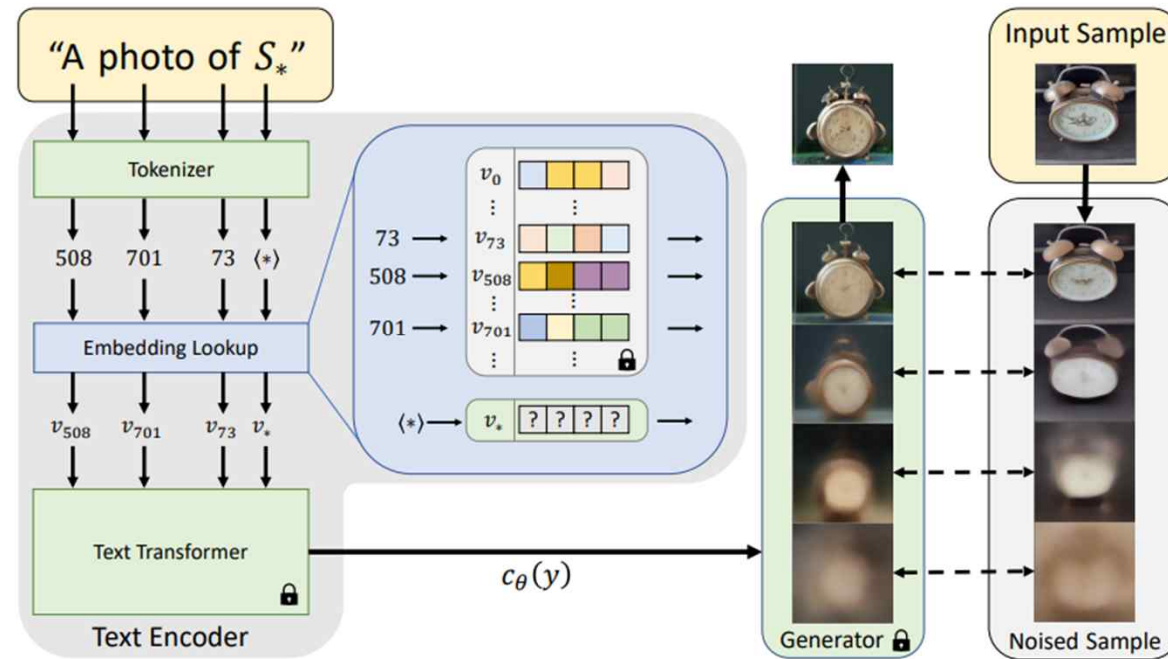
Reason for Selecting Training Image Data

Korea	China	Japan
		
<ul style="list-style-type: none">- Monochrome or wooden-colored main gates and pavilions are numerous- Curvature, simplicity- The eaves have a gentle curve- Mostly single-story buildings	<ul style="list-style-type: none">- Features shades of red/green/gold- Flamboyant, intense, linear- Large in size and majestic	<ul style="list-style-type: none">- Low-saturation main gates- Understated, linear- Many two-story structures- White and gray buildings- Many structures have gardens

3

Fine Tuning

Textual inversion

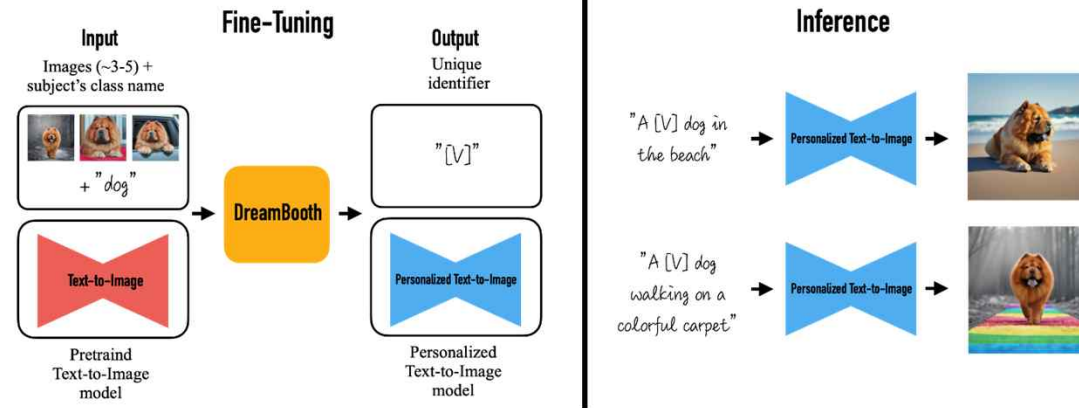


The goal is to find new embedding vectors that represent new concepts. When an image is provided, it finds a single-word embedding that leads to the reconstitution of the image from a small dataset.

3

Fine Tuning

Dreambooth



A proposed method aims to solve two problems that occur with traditional fine-tuning:

1. Topic-based image generation. Synthesizing new context images while maintaining high fidelity to the visual characteristics of the topic with just a few simple photos.
2. Preserving existing semantic knowledge while fine-tuning a Text-to-Image (T2I) diffusion model with just a handful of images.

3

Fine Tuning

LoRA

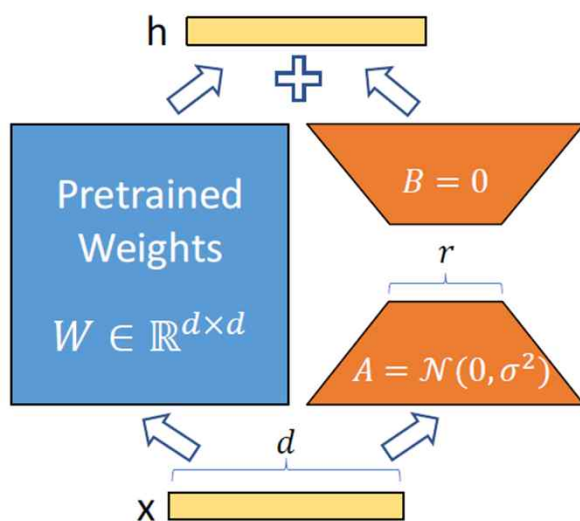


Figure 1: Our reparametrization. We only train A and B .

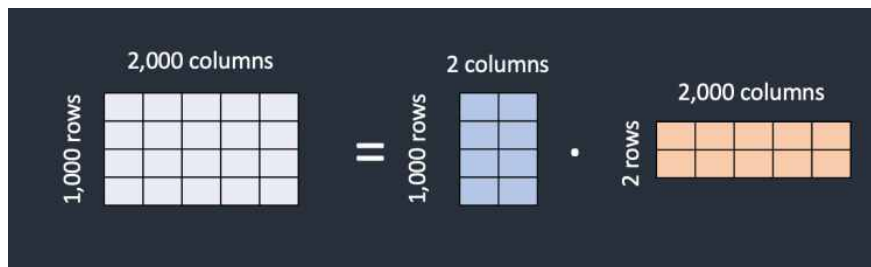
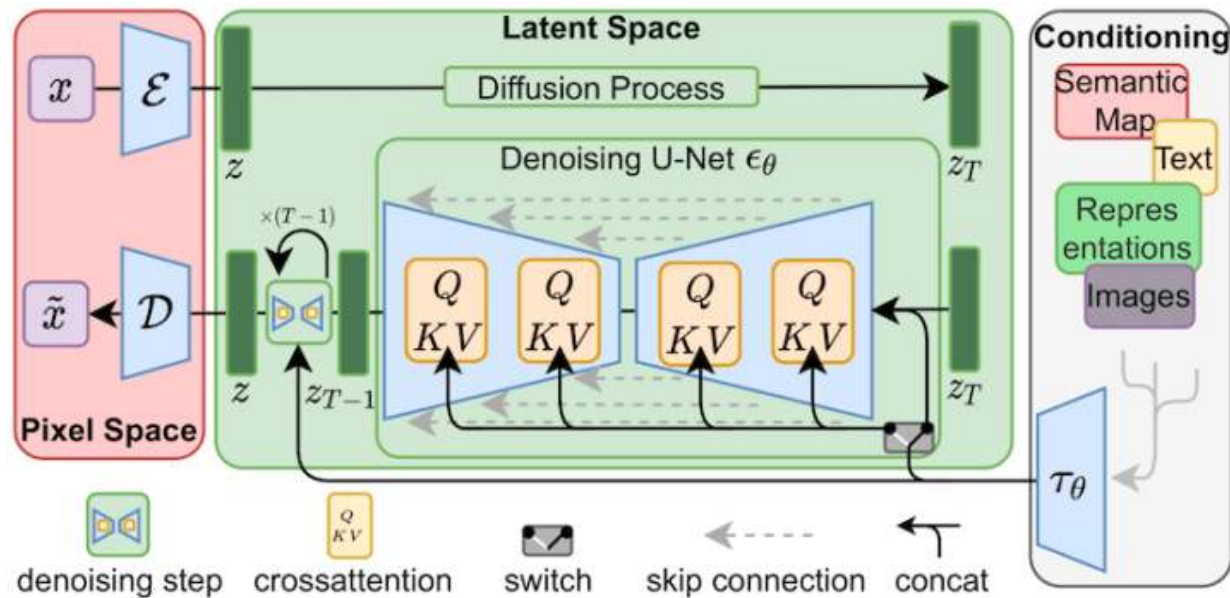
Instead of fine-tuning all weights of the pretrained model, this approach introduces adding trainable rank decomposition matrices to each layer of the Transformer architecture, with all previous weights frozen.

- Memory usage is tripled, and parameters are reduced by a factor of 10,000.
- The pretrained model remains unchanged and is shared, while only the newly trained small modules are swapped out for learning new tasks.
- Only the small matrices added to the layers are trained, allowing for efficient memory usage.
- It can be used without any additional latency during the inference process.
- Offers the flexibility to be used alongside many existing methods.

3

Fine Tuning

LoRA



$$1000 \times 2000 = 2,000,000$$

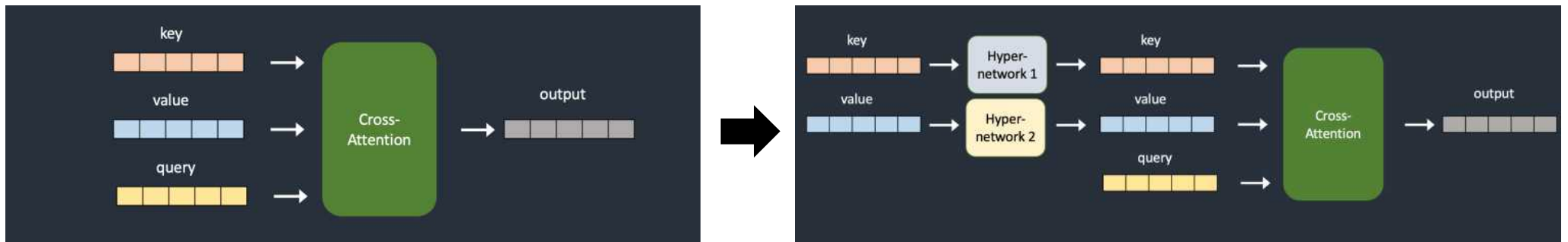
$$1000 \times 2 + 2 \times 2000 = 6000$$

Creates matrices that are 333 times smaller.

3

Fine Tuning

Hypernetworks



This represents an entirely new development, distinct from existing hypernetworks, specifically engineered by Novel AI for stable diffusion.

Incorporated within the U-Net cross-attention module, it involves the insertion of two networks that transform keys and queries. This application is exclusive to cross-attention, leaving the rest of the U-Net architecture untouched.

It cannot function independently and must be used in conjunction with a checkpoint model.

3

Fine Tuning

Experiment Process

Varations in Fine tuning process

In the fine-tuning of Stable Diffusion, unique 'initialization text' is critical for targeting specific learning outcomes, using a placeholder '*' to avoid confusion with pre-existing concepts. BLIP captioning is utilized to generate captions, transforming typical descriptions to match the initialization text, thus facilitating the learning of Hanok images through this new vector.

Ex) "A building with a roof in the fire"
-> "*" in the fire"

Background for selecting fire simulation

Fire is chosen for its easy detection in image generation, marked by the intense red color and smoke, and it's the most common building disaster, constituting 32% of cases.

Importance of the number of datasets

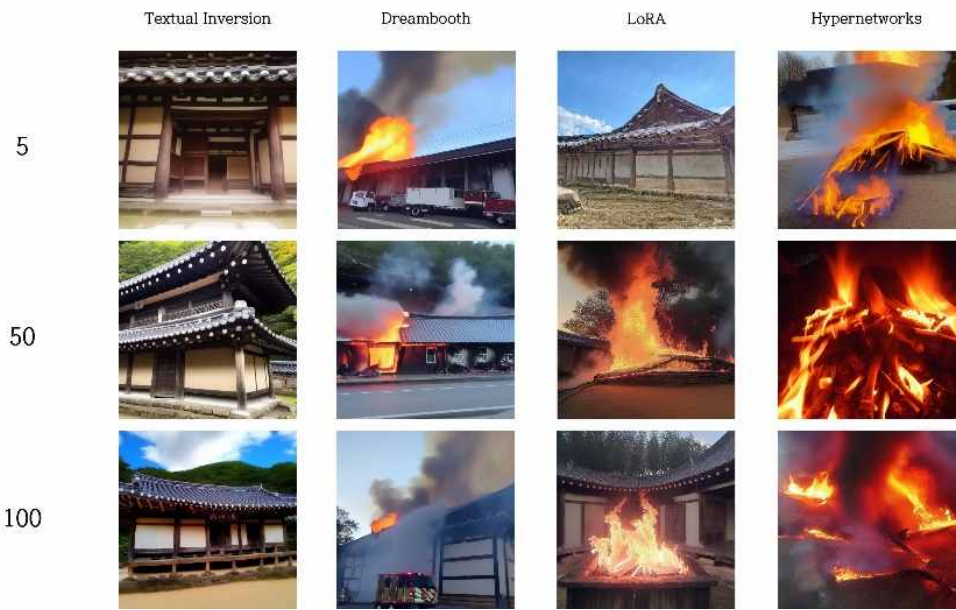
The optimal dataset size varies for each model.

Adjusting these sizes will be instrumental in deciding the required dataset magnitude for developing a Korea-specific disaster image generation model in subsequent efforts.

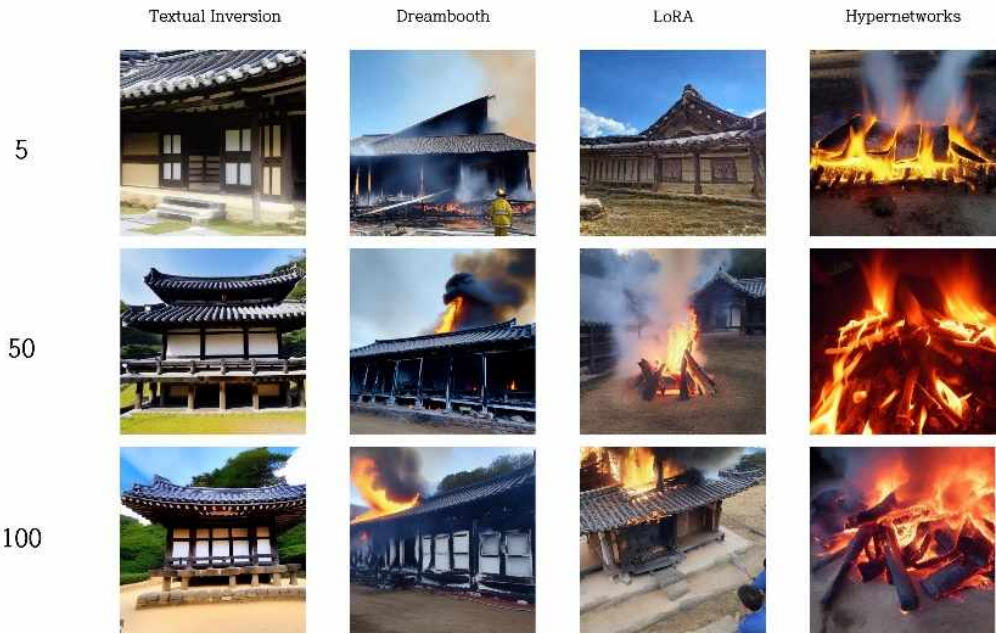
4

Experiment Results

Experiment Results



A prompt("*, fire")

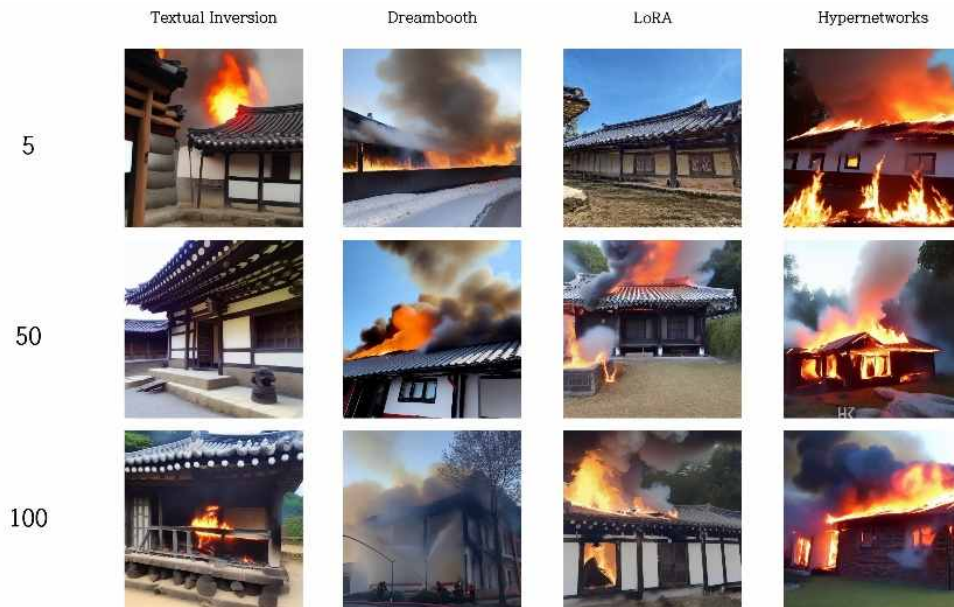


B prompt("* in the fire")

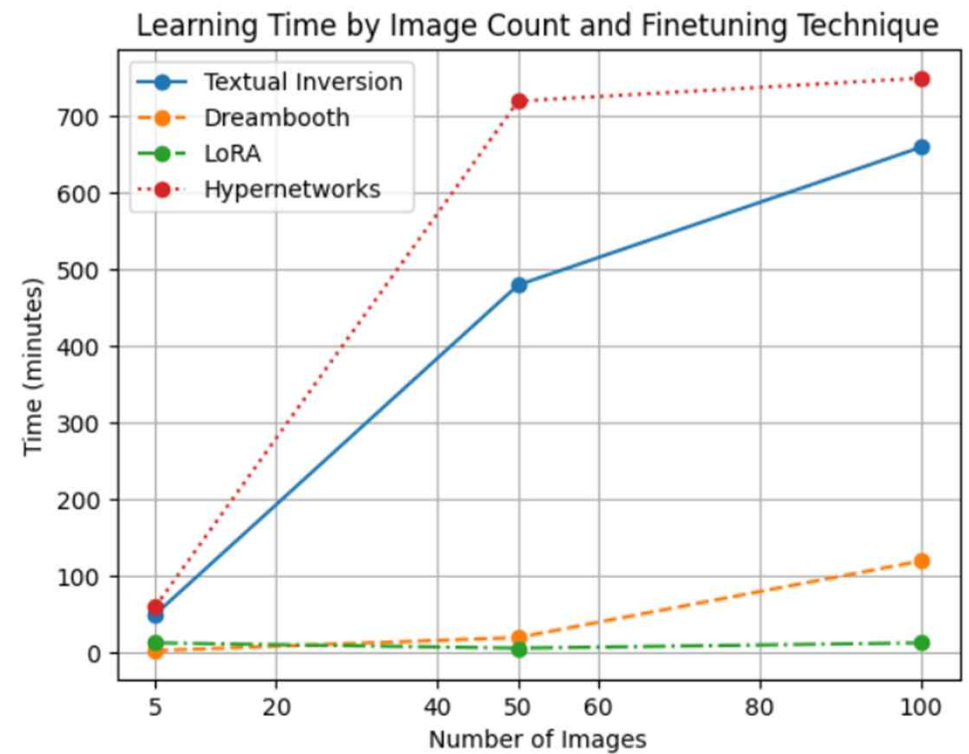
4

Experiment Results

Experiment Results



C prompt("A fire broke out in *")



5

Evaluation Results

Prompt A	Textual Inversion	Dreambooth	LoRA	Hypernetworks
5	0.24	0.30	0.24	0.26
50	0.25	0.30	0.28	0.27
100	0.25	0.30	0.28	0.27
Average	0.247	0.30	0.267	0.267

Prompt B	Textual Inversion	Dreambooth	LoRA	Hypernetworks
5	0.26	0.28	0.26	0.24
50	0.25	0.30	0.26	0.26
100	0.23	0.27	0.29	0.25
Average	0.247	0.283	0.27	0.25

Prompt C	Textual Inversion	Dreambooth	LoRA	Hypernetworks
5	0.29	0.28	0.21	0.29
50	0.21	0.30	0.27	0.28
100	0.25	0.30	0.30	0.28
Average	0.25	0.293	0.26	0.283

In all outcomes, Dreambooth received the highest scores.

5

Evaluation Results

Analysis of Causes by Model

Textual inversion

- Textual inversion Fine-tuning technology to identify the optimal embedding that can represent the concept to be tuned.
- Therefore, the resulting images are limited to being generated only within the existing model output domain. This means it is impossible to create new concepts that the original model has not learned.
- It is speculated that inference capability decreases when the prompt is lengthened or becomes more complex.
- If fine-tuning with Textual Inversion is required, additional research into prompt engineering techniques must be conducted to produce optimal result images from the original model's dataset.

5

Evaluation Results

Analysis of Causes by Model

Dreambooth

- It demonstrates outstanding generative results for fires regardless of the number of training images, also achieving the highest clip score among the four fine-tuning techniques.
- However, Dreambooth has disadvantages in terms of size and speed because it fine-tunes all weights within the U-Net and text encoder of the model. In fact, research indicates that using Dreambooth with Stable Diffusion adjusts over 1GB of parameters, and takes about 5 minutes for 1,000 training iterations.
- Consequently, a minimum of 12GB of VRAM and a high-performance GPU are required for its use, necessitating research to reduce capacity and speed.

5

Evaluation Results

Analysis of Causes by Model

LoRA

- While most results were successful in generating buildings in the form of Hanok, the representation related to fire was inconsistent, with issues of campfire-like flames appearing.
- Furthermore, in the results of training with 5, 50, and 100 images, the problem of Hanok buildings appearing in similar locations, appearances, and color schemes occurred.
- The actual LoRA technology aims to reduce the trainable parameters by exclusively learning the rank decomposition matrix without altering the existing model's parameters; hence, this method does not extend to larger matrices and cannot efficiently learn feature variations in the feature space.
- It is anticipated that research into prompts that consistently and accurately represent terrain and disaster scenarios with caution against overfitting will be necessary.

5

Evaluation Results

Analysis of Causes by Model

Hypernetworks

- Most images failed to capture the characteristics of Hanok.
- In Figures 3 and 4, the representation of fire tended to be independent, resembling campfires rather than harmoniously integrating with the buildings.
- These results suggest that the hypernetwork is unable to specifically reflect the learned terrain forms.
- Hypernetworks influence only the cross-attention part without fine-tuning the entire model, making it difficult to learn styles or themes compared to other techniques.
- Therefore, further research is needed on optimizing the training dataset and fine-tuning the prompts.

6

Conclusion

Summary & Future Works

[Summary]

- The goal of this research is to generate disaster images using image generation models.
- Currently, image generation models are composed of large datasets, which makes it challenging to generate terrain images specialized for a specific country.
- To train existing image generation models data they don't know and to assess the results, a dataset with distinctive Korean traditional houses, Hanok, was compiled.
- The results of adding Hanok images to the training using four different fine-tuning techniques were observed and analyzed.
- Ultimately, it was found that using the Dreambooth model yielded the best results in terms of time and performance, although this may vary depending on computer specifications.

[Future Works]

- Exploring editing and synthesis techniques rather than image generation.
- Attempt to construct a disaster image generation algorithm.

Thank you

Q&A