석사학위 청구논문

지도교수 조 용 범

생성형 AI를 활용한 효과적인 딥러닝 학습용 데이터셋 생성에 대한 연구

2024년 2월

건국대학교 대학원

전자·정보통신공학과

유태영

생성형 AI를 활용한 효과적인 딥러닝 학습용 데이터셋 생성에 대한 연구

Research on Effective Dataset Generation Using Generative AI for Deep Learning Training

이 논문을 전자공학 석사학위 청구논문으로 제출합니다.

2023년 11월

건국대학교 대학원

전자·정보통신공학과

유태영

유태영의 공학 석사학위 청구논문을 인준함.

심사위원장 이 기승 (인)

심사위원 조 용범 (인)

심사위원 김 원준 (인)

2023년 12월

건국대학교 대학원

목차

표 목차

그림 목차

ABSTRACT

Research on Effective Dataset Generation Using Generative AI for Deep Learning Training

Yu, Tae Yeong
Department of Electronics, Information and communication Engineering
Graduate School of Konkuk University

This paper explores the possibility of using synthetic images generated by Text-to-Image (t2i) models as a substitute for deep learning training datasets. Previous research has explored the potential for dataset expansion through image synthesis and 3D simulation, but these methods have shown practical limitations. This study utilizes the continuously advancing Diffusion Model to generate high-quality synthetic images and validates their practical application by applying them to various deep learning architectures and assessing image classification performance on three general-purpose datasets.

Additionally, to address the lack of output diversity highlighted in previous research, this paper introduces a new Image-to-Text (i2t) model, HypoNet, capable of generating multiple subcategories from a single sample image. The study also compares the results of images generated using the latest techniques such as FreeU and LoRA. The findings reveal that training with images created using prompts extracted by HypoNet shows equivalent or

superior performance compared to training with real images. This research suggests that dataset creation using the Diffusion Model without additional training on existing models can be a practical alternative, presenting a new avenue for dataset creation that significantly reduces reliance on large-scale real-world data collection.

Keywords: Dataset Generation, Deep learning, Diffusion Model, Image classification

제1장 서론

기계학습 분야에서 딥러닝 모델의 효율적인 학습을 위해 다양한 데이터셋을 확보하는 것은 딥러닝 모델의 선택만큼 중요한 일이다[1,2,3,4]. 양질의 이미지 수집은 복잡하고 어려운 과정이며, 이는 1990년대에 연구자들이 직접 사진을 찍어 데이터셋을 구축하던 시절부터[5,6,7], 2000년대 인터넷 크롤링을 통한 데이터 수집에 이르기까지 지속적으로 발전해왔다[8]. 그러나 이러한 방식으로 구축된 데이터셋은 노이즈와 사회적 편향을 반영할 위험이 있으며, 이를 정제하기 위한 편집 작업은 상당한 비용을 수반한다.

이에 따라 이미지 생성 모델을 활용한 데이터셋 구축[9,10,11] 또는 3D 시뮬레이터를 활용한 데이터셋 구축에 관한 연구[12,13,14,15,16,17]가 진행되었다. 그러나 이러한 방법들은 데이터의 다양성, 품질, 그리고 실제 환경 적용성 부족이라는 한계로 인해 실제 학습 데이터로 사용되지 않고 있다.

본 연구에서는 디퓨전 모델[18, 19]과 같은 혁신적인 이미지 생성 기술을 활용하여, 이전 연구에서 지적된 다양성 및 품질 문제를 해결하고자 한다. 다양한 기술을 적용하여 생성된 이미지 데이터셋을 비교 분석함으로써, 가장 실용적인 데이터셋 생성 방법을 탐구한다.

이 연구의 주요 기여는 새로운 Image-to-Text (i2t) 모델인 HypoNet의 개발 및 구현이다. HypoNet은 단일 샘플 이미지에서 여러 하위 카테고리를 생성하며, 이를 주요 오픈 소스 Text-to-Image (t2i) 모델인 Stable Diffusion[20]에 적용 후 생성된 데이터셋으로 학습한 모델과 전통적인 데이터셋으로 학습한 모델 간의 이미지 분류 정확도를 비교 분석하여, HypoNet의 실용성을 입증한다.

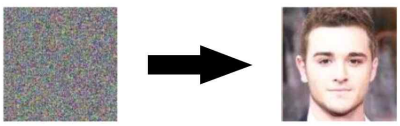
결과적으로, HypoNet을 이용하여 생성된 데이터셋으로 학습된 모델이 전통적인 데이터셋으로 학습된 모델과 비교하여 동등하거나 그 이상의 이미지 분류 정확도를 달성했다는 점에서, 본 연구는 딥러닝 학습용 데이터셋 생성 방법론에 새로운 지평을 제시한다.

제2장 배경 지식

제1절 디퓨전 모델(Diffusion Model)

디퓨전 모델[18, 19]은 최근 딥러닝과 확률론을 결합한 연구에서 주목받고 있는 방법론으로, 특히 이미지 생성 분야에서 전통적인 생성적 적대 신경망(GANs)[21]보다 뛰어난 결과를 보이는 것으로 알려져 있다. 이 모델은 데이터를 점진적으로 노이즈로 변환하는 과정과 그 역과정을 학습하여, 고품질의 샘플을 생성할 수 있다. 이러한 방식은 이미지, 오디오 및 기타 복잡한 데이터 타입에 효과적으로 적용될 수 있다[22].

<그림 2-1>은 디퓨전 모델의 동작을 나타내는 예시이다.



<그림 2-1> Diffusion Model의 동작 예시

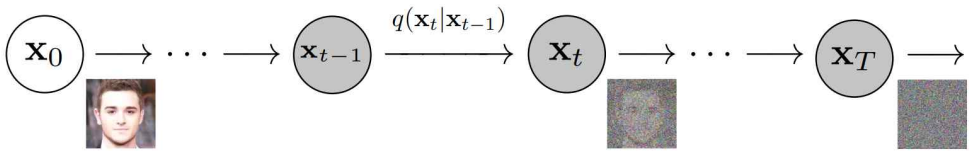
1. 디퓨전 과정 (Forward Process)

(식 2-1)은 디퓨전 과정을 수식으로 나타낸 것으로, 원본 데이터x_0을 점차적으로 노이즈 ϵ로 변환하는 과정이다. 이 변환은 일련의 단계 T를 거치며, 각 단계에서 데이터 x_t에 노이즈를 추가한다[19].

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim N(0, I) \qquad \text{(식 2-1)}$$

(식 2-1)에서 α_t는 각 단계에서의 노이즈 비율을 결정하는 계수이며, $N(0, I)$는 평균이 0이고 단위 분산을 가지는 정규 분포를 나타낸다.

<그림 2-2>는 디퓨전 과정을 시각화한 예시를 보여준다.

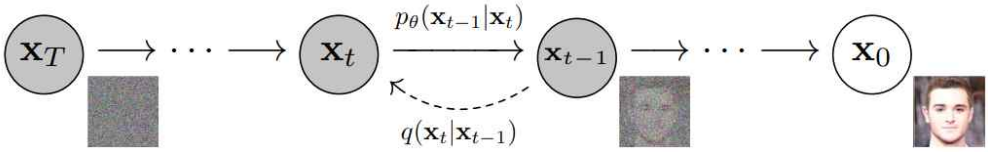


<그림 2-2> Forward Process0

2. 역 디퓨전 과정 (Reverse Process)

역 디퓨전 과정은 노이즈화된 데이터 x_T로부터 원본 데이터 x_0를 복구하는 과정이다. 이 과정은 딥러닝 모델을 사용하여 노이즈를 제거하는 방식으로 이루어진다. 모델은 (식 2-2)와 같이 학습된다:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(x_t,t)), \quad \overline{\alpha_t} = \prod_{i=1}^{t}\alpha_i \qquad \text{(식 2-2)}$$



<그림 2-3> Reverse Process

(식 2-2)에서 $\epsilon_\theta(x_t,t)$는 학습된 모델을 통해 추정된 노이즈이며, $\overline{\alpha_t}$는 시간 t까지의 누적 노이즈 비율을 나타낸다.

<그림 2-3>은 역 디퓨전 과정을 시각화한 예시를 보여준다.

3. 학습 및 최적화

모델은 원본 데이터 x_0와 노이즈 ϵ사이의 차이를 최소화 하는 방향으로 학습된다. 이는 다음과 같은 손실 함수를 최소화함으로써 달성된다:

$$L(\theta) = \boldsymbol{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(x_t,t)||^2] \qquad \text{(식 2-3)}$$

(식 2-3)에서 $\epsilon_\theta(x_t, t)$는 모델이 추정한 노이즈이며, ϵ은 실제 노이즈이다. 이 손실 함수는 모델이 데이터의 실제 분포를 더 잘 이해하고 학습할 수 있도록 도와준다.

이 방법은 GAN과 달리 학습 과정에서의 불안정성이나 모드 붕괴(mode collapse)와 같은 문제들을 효과적으로 완화시킨다[23]. 또한, 디퓨전 모델은 특히 고해상도 이미지 생성에서 GAN보다 우수한 성능을 보이는 것으로 평가되고 있으며, 이는 다양한 응용 분야에서 그 잠재력을 발휘할 수 있음을 의미한다[24,25].

제2절 잠재 디퓨전 모델(Latent Diffusion Model)

스테이블 디퓨전(Stable Diffusion)과 같은 잠재 디퓨전 모델은 이미지 생성 분야에서의 새로운 패러다임을 제시한다[26]. 이 모델들은 고차원의 이미지 공간 대신 잠재 공간(Latent Space)에서 작동함으로써 기존 디퓨전 모델의 한계를 극복하고, 연산 효율성을 크게 개선한다[27].

1. 잠재 디퓨전 모델의 접근법

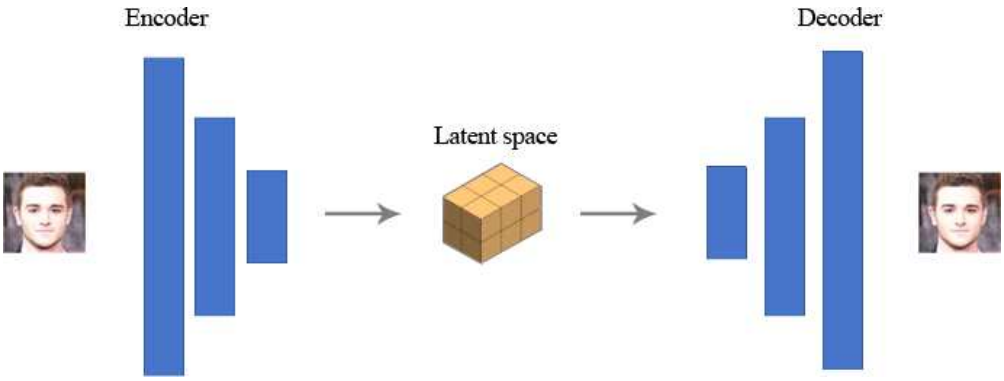
고차원 이미지 공간에서의 작업은 컴퓨팅 자원과 시간 면에서 큰 부담을 주며 상당한 어려움을 야기한다[28]. 예를 들어, 512×512 픽셀의 이미지는 262,144개의 픽셀로 구성되며, 각 픽셀은 RGB 채널을 포함하여 총 786,432개의 차원을 갖는다[29]. 이러한 높은 차원의 데이터는 처리하기 어렵고, 시간이 많이 소요되는 문제가 있다.

잠재 확산 모델(Latent Diffusion Model)은 이미지를 잠재 공간으로 압축하여 이 문제에 접근한다. 이 과정에서 원본 이미지는 상대적으로 낮은 차원의 데이터로 변환되어 처리될 수 있다[30]. 잠재 공간의 차원은 원본 이미지 공간의 차원에 비해 훨씬 작다는 장점이 있다[31].

잠재 공간으로의 압축은 가변 자동 인코더(Variational Auto Encoder, VAE)를 사용한다. VAE[27]는 인코더와 디코더 두 부분으로 구성되며, 인코더는

이미지를 잠재 공간의 낮은 차원으로 변환하고, 디코더는 이 잠재 표현을 다시 원본 이미지로 복원하는 역할을 한다[33].

Stable Diffusion 모델의 잠재 공간 차원은 4×64×64로[26], 16,384개의 차원을 가지며, 이는 원본 이미지 공간 차원인 786,432의 1/48에 불과하다. 이 저차원의 특성은 연산을 더 빠르고 효율적으로 만들며, 고해상도 이미지 생성에 필요한 컴퓨팅 자원을 줄일 수 있다[34].



<그림 2-4> 가변 자동 인코더(Variational Auto Encoder, VAE)

<그림 2-4>는 가변 자동 인코더(Variational Auto Encoder, VAE)의 동작을 시각적으로 나타내며, 입력 이미지가 Encoder를 통해 Latent Space로 변환되고 Decoder를 통해 원본 이미지로 복원되는 것을 보여준다.

잠재 디퓨전 모델에서의 디퓨전 과정은 잠재 공간상에서 발생한다. 이 과정에서 생성되는 잡음은 잠재 공간상의 이미지 표현을 변형시키는 데 사용된다. 이 변형 과정은 원본 이미지 공간보다 훨씬 적은 연산으로 수행될 수 있다.

2. 조건 부여 및 텍스트 프롬프트 처리

잠재 디퓨전 모델에서 특정 이미지를 생성하기 위해서는 조건부여 과정이 필요하다. 조건부여 없이는 잠재 공간에서 학습된 데이터로부터 무작위 이미지만 생성된다. 사용자가 원하는 특정한 주제나 스타일의 이미지를 생성

하기 위해서는 텍스트 기반의 조건부여가 필수적이다. 프롬프트에 포함된 단어들은 먼저 토큰 생성기에 의해 토큰으로 변환되고, 이 토큰들은 임베딩 과정을 거쳐 벡터 형태로 변환된다. 처리된 임베딩은 텍스트 변환기를 거쳐 잡음 예측기로 전달된다. U-Net[35] 구조의 잡음 예측기는 텍스트 변환기의 출력을 여러 번 사용한다. 이러한 조건부여 과정은 잠재 디퓨전 모델이 사용자의 입력에 따라 구체적이고 의도된 이미지를 생성할 수 있도록 한다.

각 토큰은 고유한 임베딩 벡터를 갖는다. 이러한 임베딩은 단어 간의 연관성을 포착한다. 예를 들어, “man“, “gentleman“, “guy“는 서로 연관된 임베딩을 가진다. 이러한 임베딩은 텍스트와 이미지 간의 더 세밀한 매칭을 가능하게 한다. 처리된 임베딩은 텍스트 변환기(text transformer)를 거쳐 잡음 예측기(noise predictor)로 전달된다. 이 변환기는 다양한 유형의 입력을 처리할 수 있으며, 잡음 예측기로 전달되는 데이터를 조정한다.

U-Net 구조의 잡음 예측기는 텍스트 변환기의 출력을 여러 번 사용한다. 이때 교차 인지 메커니즘을 통해 텍스트 프롬프트와 이미지 사이의 상호 작용이 이루어진다. “갈색 머리의 남자”라는 프롬프트를 예시로 들었을 때, “갈색”과 “머리”의 두 단어를 교차 인지를 통해 연관지어 갈색 머리의 남자 이미지를 생성한다.

이러한 조건부여 과정은 잠재 디퓨전 모델이 사용자의 입력에 따라 구체적이고 의도된 이미지를 생성할 수 있도록 한다. 이는 모델의 다양성과 유연성을 크게 향상시키며, 사용자가 원하는 특정한 이미지나 스타일을 정확하게 생성할 수 있게 한다.

3. 분류 자유도 척도 (Classifier-Free Guidance, CFG)

분류기 안내 척도(Classifier Guidance Scale)[33]는 디퓨전 과정이 얼마나 레이블을 밀접하게 따를지를 결정하는 매개변수다. 높은 값을 설정하면, 모델은 더 명확하고 모호성이 적은 이미지를 생성한다. 즉, “cat“을 요청할 경우, 다른 요소 없이 오직 고양이만 있는 이미지를 생성할 가능성이 높아진다.

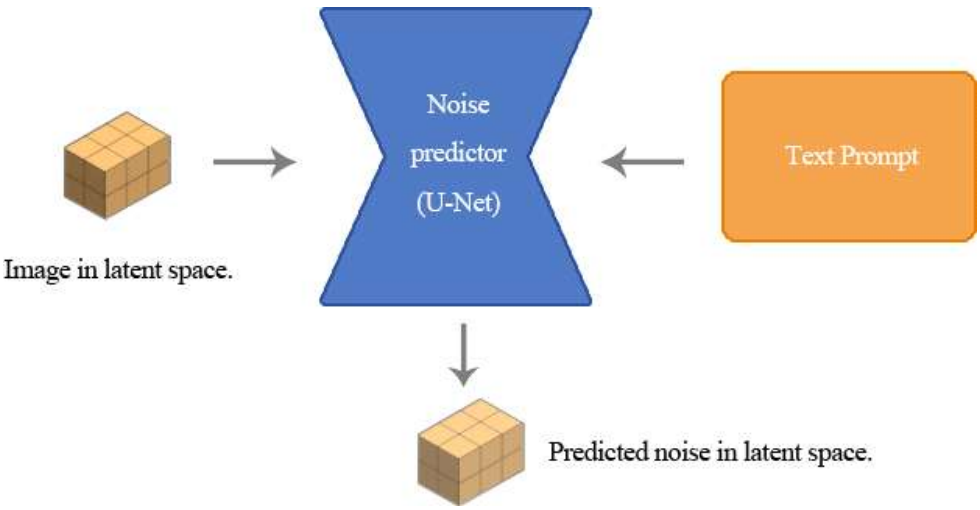
분류기 안내는 강력하지만, 추가적인 모델이 필요하며 학습 과정이 복잡해질 수 있다. 무분류기 안내(Classifier-Free Guidance)는 분류기가 없는 상태에서도 유사한 결과를 달성하기 위한 접근 방식이다. 이는 클래스 레이블이

나 별도의 모델을 사용하지 않고, 이미지 캡션과 같은 텍스트 기반의 조건부여를 사용하여 디퓨전 모델을 훈련한다.

무분류기 안내는 U-Net[35]과 같은 잡음 예측기의 조건부화 기능을 활용하여 별도의 이미지 분류기 없이도 특정 방향성을 부여한다. 텍스트 프롬프트는 이 과정에서 핵심 역할을 하며, text-to-image 변환 과정에서 지향점을 설정하는 데 기여한다.

분류 자유도 척도(Classifier-Free Guidance, CFG)[33] 값은 텍스트 프롬프트 기반의 조건부여를 얼마나 따를지를 결정한다. 이 값이 0이면, 프롬프트가 무시되어 조건부여 없이 이미지가 생성된다. 반대로 높은 값을 설정하면, 디퓨전 과정은 프롬프트에 더 밀접하게 응답하여, 사용자의 요구에 맞는 이미지를 생성하게 된다.

<그림 2-5>는 텍스트 프롬프트에 따른 잠재 공간 에서의 노이즈 예측을 시각화한 예시이다.



<그림 2-5> 잠재 공간에서의 노이즈 예측

제3장 디퓨전 모델 기반 데이터셋 생성 방법

본 장에서는 기존에 연구된 디퓨전 모델 기반 이미지 생성 방법들에 대해 설명한다. 제1절에서는 다양성을 해결하기 위한 생성 전략인 동적 프롬프트의 원리와 와일드카드(wildcard)에 대해 설명하며, 제2절에서는 프롬프트에 충실한 이미지를 출력하기 위해 고안된 전략인 긍정 프롬프트와 부정 프롬프트에 대해 설명한다. 제3절에서는 제2절과 마찬가지로 정확한 이미지를 출력하기 위해 고안된 최신 기법인 FreeU에 대해 설명한다. 제4절에서는 기존 LLM(Large Language Model)에서 사용된 파인튜닝 기법인 LoRA(Low Rank Adaptation)를 디퓨전 모델에 적용한 사례를 설명한다.

제1절 동적 프롬프트

동적 프롬프트(Dynamic Prompt)는 텍스트 기반 이미지 생성에서 입력 텍스트의 동적 조정을 통하여 생성되는 이미지의 다양성을 향상시키는 기법이다[34]. 이 방법은 기존의 고정된 텍스트 지시어 대신, 입력 텍스트를 변형하여 다양한 시나리오를 모사할 수 있는 유연성을 제공한다[33].

기존 이미지 생성의 문제점 중 하나인 다양성의 부족은 <그림 3-1>과 <그림 3-2>에서 확인할 수 있다. 각기 다른 랜덤 노이즈에서 생성된 이미지 이지만 서로 비슷한 타겟 오브젝트를 생성함을 확인할 수 있다. 그러나 동적 프롬프트를 적용한 예시인 <그림 3-2>에서는, 다양한 종류의 타겟 오브젝트와 배경, 구도를 포함하여 이미지의 다양성이 향상 된 것을 확인할 수 있다.

1. 동작 원리

<알고리즘 3-1>은 전체적인 동적 프롬프트의 동작 방식을 설명한다. 동적 프롬프트는 기본 프롬프트(basicPrompt)와 변형 목록(variations)을 입력으로 받는다. 각 변형에 대해 가능한 모든 조합을 생성하여 새로운 프롬프트를 형성한다. 알고리즘에서의 예시로, [“red”, “blue”], [“airplane”,

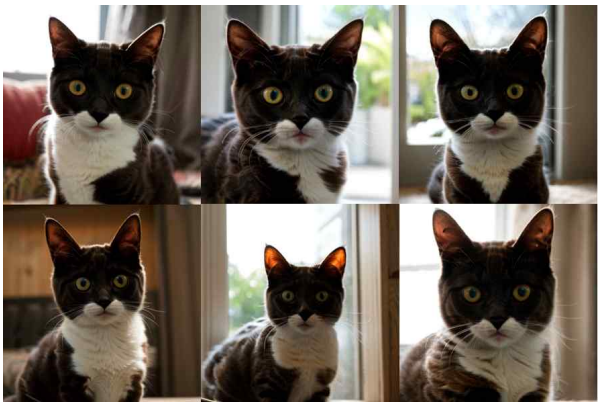
“truck”]의 동적 프롬프트를 사용한다면 “red airplane”, “blue airplane”, “red truck”, “blue truck”의 네 가지 조합의 프롬프트를 생성할 수 있게되며, 이 중 하나를 무작위 하게 디퓨전 모델의 입력 프롬프트로 사용하게 된다.

이와 같이 동적 프롬프트는 주어진 기본 텍스트 프롬프트에 여러 변형을 적용한다. 이러한 변형은 특정 주제, 속성, 상황 등을 반영하여 이미지 생성 과정에 다양한 맥락을 제공한다.

2. 와일드카드 (Wildcard)

본 논문에서는 다양한 변형 목록(variations)를 효율적으로 사용하기 위해 와일드카드(Wildcard)의 개념을 사용한다. 와일드카드는 하나의 상위 분류로서, 여러 하위분류를 포함한 텍스트 파일로 구성된다.

동적 프롬프트의 데이터셋 생성 효율성을 검증하기 위해 view, color, location의 총 세 가지의 와일드카드를 고정적으로 사용했으며, 각 하위분류의 값은 <표3-1>과 같다.



<그림 3-1> 동적 프롬프트 미적용 시

<그림 3-2> 동적 프롬프트 적용 시

```
Function generateDynamicPrompt(basicPrompt, variations)
    newPrompt = basicPrompt
    foreach variation in variations
        randomIndex = random(0, length(variation) - 1)
        newPrompt += " " + variation[randomIndex]
    return newPrompt
// 예시
basicPrompt = "A photo of"
variations = [["red", "blue"], ["airplane", "truck"]]
dynamicPrompt = generateDynamicPrompt(basicPrompt, variations)
```

<알고리즘 3-1> 동적 프롬프트

view 와일드카드의 하위분류에는 물체를 보는 시점에 대한 5개의 하위 분류가 포함되어 있으며, color 와일드카드의 하위분류에는 물체의 색깔에 대한 45개의 하위분류가 포함되어 있다. 마지막으로 location 와일드카드의 하위분류에는 45개의 무작위한 위치 정보가 포함되어 있다. 이 세 가지 와일드카드를 고정적으로 사용함으로써 목표하는 물체에 대한 다양한 상황의 이미지를 생성할 수 있게된다.

동적 프롬프트의 적용을 통해 생성된 데이터셋은 다양한 시나리오에 걸친 목표 객체의 다양성을 증대시킴으로써 이미지 생성의 범위를 넓힌다. 이는 데이터 과적합(overfitting)과 표본 편향(sample bias)을 방지하고, 모델의 일반화 능력을 강화할 것으로 기대된다. 이러한 접근은 특히 학습 데이터셋의 다양성이 중요한 딥러닝 애플리케이션에서 핵심적인 가치를 지니며, 이를 통해 모델의 성능과 신뢰도를 향상시킬 수 있을 것으로 예상된다.

<표 3-1> 고정 와일드카드와 그 하위분류

view	color	location
	Red	City
	Blue	Village
	Green	Island
	Yellow	Mountain Range
	Black	Desert
	White	Forest
	Orange	Beach
	Purple	Lake
	Pink	River
	Brown	Valley
	Gray	Ocean
	Violet	Countryside
	Gold	Suburb
	Silver	Jungle
	Crimson	Cave
	Teal	Canyon
	Lavender	Waterfall
	Beige	Volcano
	Turquoise	Swamp
	Indigo	Fjord
Side View	Maroon	Plains
Frontal View	Olive	Hill
Rear View	Coral	Glacier
Top-Down View	Magenta	Meadow
Close-Up View	Lime	Peninsula
	Cyan	Wetlands
	Dark Blue	Savanna
	Light Blue	Hot Springs
	Dark Green	Cliff
	Light Green	Marshland
	Mustard	Lagoon
	Peach	Estuary
	Rose	Reef
	Emerald	Delta
	Sapphire	Sand Dunes
	Scarlet	Plateau
	Rust	Prairie
	Burgundy	Badlands
	Mint Green	Tundra
	Sky Blue	Archipelago
	Fuchsia	Ravine
	Amber	Geyser
	Ivory	Rainforest
	Navy Blue	Sahara Desert
	Charcoal	Everglades

제2절 긍정 프롬프트와 부정 프롬프트

긍정 프롬프트(Positive Prompt)와 부정 프롬프트(Negative Prompt)의 사용은 이미지 생성의 정확성을 극대화하는데 중요한 역할을 한다. 긍정 프롬프트는 원하는 요소를 강조하며 부정 프롬프트는 원치 않는 요소를 제거하는데 사용된다.

1. 긍정 프롬프트 (Positive Prompt)

Text-to-Image모델에서 사용되는 입력 프롬프트는 긍정 프롬프트를 의미한다. 긍정 프롬프트는 생성하고자 하는 이미지에 대한 텍스트를 포함하며, 디퓨전 모델의 이미지 생성 방향을 구체적으로 가이드하는 역할을 수행한다. 따라서 긍정 프롬프트의 정확하고 상세한 사용은 이미지의 질과 주제에 대한 관련성을 향상시키는 중요한 기여를 한다.

2. 부정 프롬프트 (Negative Prompt)

부정 프롬프트는 긍정 프롬프트의 반대 개념으로서, 생성을 원치 않는 이미지에 대한 텍스트를 포함하며, 이 또한 생성 방향을 가이드하는 역할을 수행한다. 부정 프롬프트의 원리를 이해하기 위해서는 먼저 긍정 프롬프트만 사용했을 때의 한계를 이해하는 것이 중요하다. <그림 3-3>와 같이 “Portrait photo of a man“의 긍정 프롬프트를 사용하면, 예상대로 남성의 초상화를 생성하지만, <그림 3-4>와 같이 “Portrait photo of a man without mustache“의 긍정 프롬프트를 사용하면, 오히려 더 두드러진 콧수염을 가진 남성 이미지가 생성될 수 있다. 이는 디퓨전 모델이 “without“과 “mustache“를 연관지어 해석하지 못하고, 단순히 “man“과 “mustache“로 이해하는 문제에서 기인한다.

따라서, 원하지 않는 요소를 제거하고자 할 때 부정 프롬프트를 사용하는 것이 효과적이다. <그림 3-5>에서는 긍정 프롬프트 “Portrait photo of a man“와 부정 프롬프트 “mustache“를 함께 사용하여 콧수염이 없는 남성 이미지를 성공적으로 생성하였다. 이는 부정 프롬프트가 디퓨전 모델을 통해

원하지 않는 요소를 제거하는 데 중요한 역할을 한다는 원리를 시사한다 [50].

3. 프롬프트의 상호작용

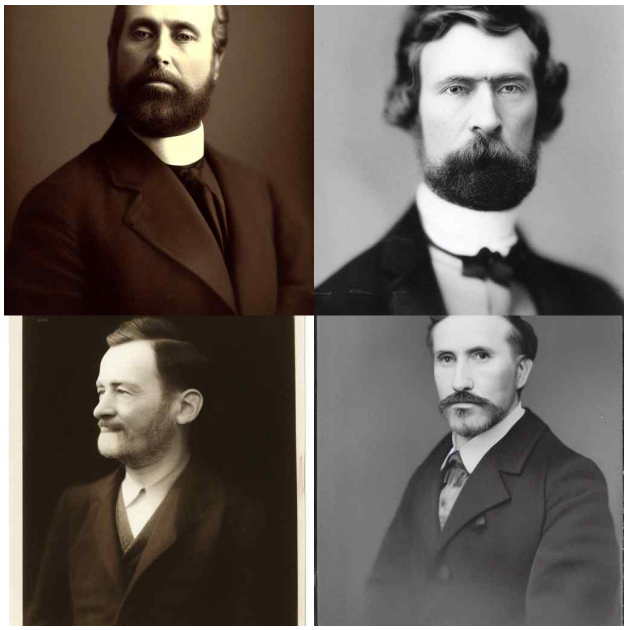
긍정 프롬프트만으로 샘플링을 수행할 때, 알고리즘은 먼저 텍스트 프롬프트에 의해 가이드되는 조건부 샘플링을 통해 이미지를 디노이즈한다. 이후, 텍스트 프롬프트를 사용하지 않는 것처럼 무조건적인 샘플링으로 동일한 이미지를 다시 디노이즈한다.

긍정 프롬프트와 부정 프롬프트 사용할 때, 샘플링 단계는 긍정 프롬프트에 따라 이미지를 생성하는 방향으로, 부정 프롬프트에서 설명하는 것으로부터 멀어지는 방향으로 진행된다.

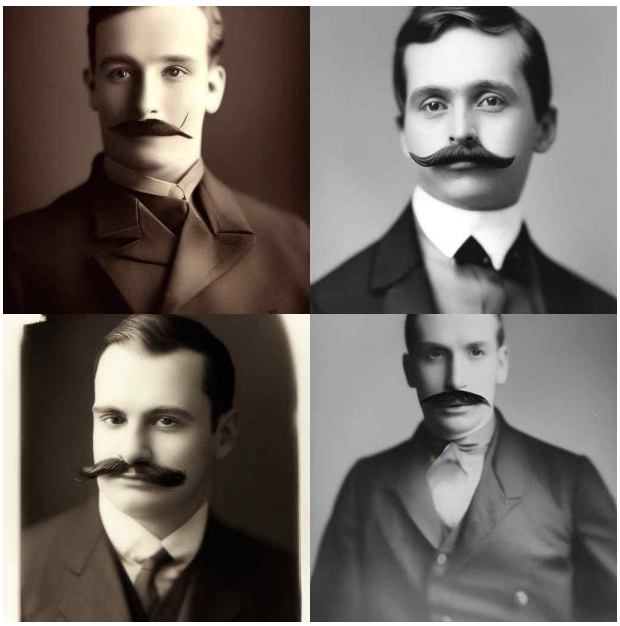
결과적으로, 긍정 프롬프트는 디퓨전을 관련 이미지 쪽으로 유도하는 반면, 부정 프롬프트는 디퓨전을 그것으로부터 멀어지게 한다. 이러한 디퓨전과정은 잠재 공간에서 발생하며, 이미지 공간에서의 이러한 표현은 단지 설명을 위한 것이다.

긍정과 부정 프롬프트를 혼합하여 사용할 때에는 주의가 필요하다. 긍정과 부정 프롬프트가 서로 충돌하거나 유사한 경우, 이미지에 불필요한 아티팩트가 생성될 수 있다. 따라서 명확하고 조직적인 프롬프트 작성이 중요하다.

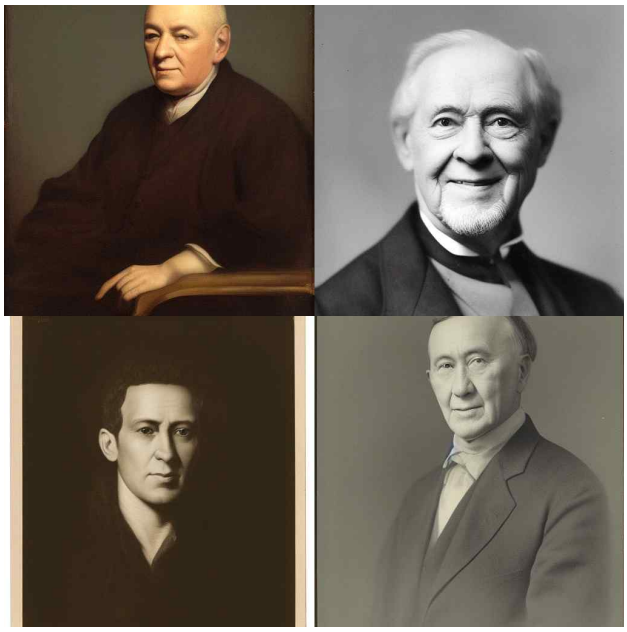
긍정 프롬프트와 부정 프롬프트를 적절히 적용하여 생성된 데이터셋은 기존 합성 이미지의 문제인 부정확한 생성을 방지하여 정확한 목표 이미지 생성을 가능케 하고, 높은 품질의 이미지를 생성할 수 있다. 결과적으로, 이 접근법의 데이터셋 생성은 딥러닝 학습에서의 데이터 레이블 잡음(Label Noise), 학습 정확도 저하(Decreased Training Accuracy), 예측 신뢰도 감소(Decreased Prediction Reliability)를 방지할 것으로 기대할 수 있다.



<그림 3-3>
긍정 프롬프트: Portrait photo of a man.



<그림 3-4>
긍정 프롬프트: Portrait photo of a man without mustache.

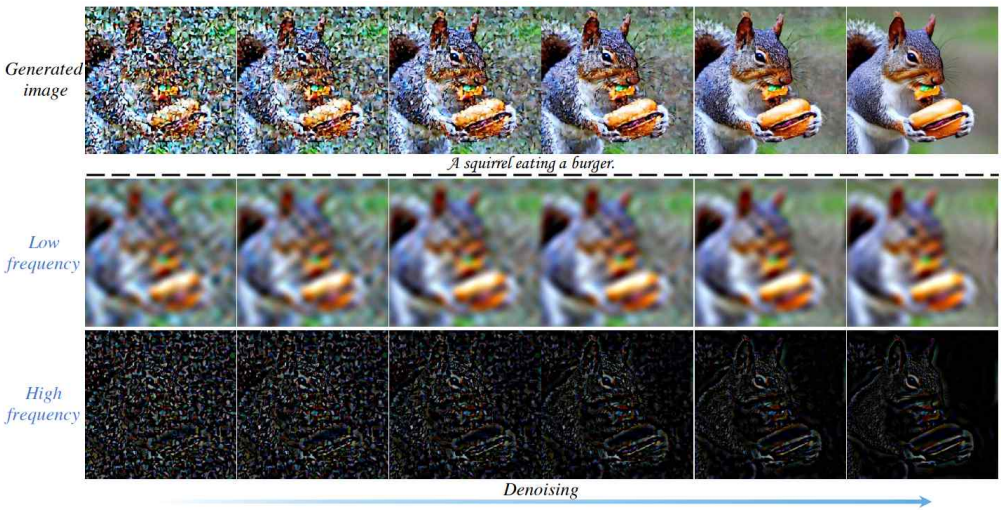


<그림 3-5> 긍정 프롬프트: Portrait photo of a man.
부정 프롬프트: mustache.

제3절 FreeU

FreeU[36]는 기존의 VAE[27], GAN[21], 벡터 양자화 접근 방식과 구별되는 새로운 디퓨전 모델 기반 생성 패러다임을 제시한다. 이 접근법은 고정된 마르코프 체인을 통해 잠재 공간을 매핑하며, 데이터셋 내의 구조적 복잡성을 포착하는 복잡한 매핑을 용이하게 한다.

FreeU의 핵심은 U-Net 아키텍처의 효율성을 강화하는 것이다. 이는 저주파 및 고주파 성분 사이의 관찰된 차이에 기반하여 이루어진다. 저주파 성분은 이미지의 글로벌 구조와 특성을 구현하고, 고주파 성분은 이미지의 가장자리와 텍스처 등 급격한 변화를 담는다. FreeU는 이러한 두 성분 사이의 균형을 조절하여 디퓨전 프레임워크 내에서 U-Net 아키텍처의 기능을 강화한다.



<그림 3-6> 디퓨전 모델의 디노이징 과정

<그림 3-6>의 상단 행에는 디노이징 과정을 거치는 동안의 연속적인 반복을 통해 생성된 이미지가 나타난다[36]. 하단의 두 행은 역 푸리에 변환을 적용한 후, 각각의 단계에 따른 저주파 및 고주파 성분의 공간 도메인 정보를 시각화하고 있다. 이 그림에서 명확히 관찰할 수 있는 점은, 저주파 성분이 점진적으로 변조되며 부드러운 변화를 보여주는 반면, 고주파 성분은 디노이징 과정 전반에 걸쳐 더욱 뚜렷하고 역동적인 변화를 나타낸다는 사실이다.

저주파 성분은 본질적으로 이미지의 글로벌 레이아웃과 부드러운 색상을 포함하는 글로벌 구조와 특성을 구현한다. 이 성분은 이미지의 본질과 표현을 구성하는 주요 글로벌 요소를 담당한다. 디노이징 과정에서 이러한 저주파 성분의 급격한 변화는 일반적으로 불합리하며, 이미지의 기본적인 본질을 변형시킬 수 있어, 이는 디노이징 프로세스의 목표와 부합하지 않는다.

반면, 고주파 성분은 이미지의 가장자리, 텍스처 등 급격한 변화를 담당한다. 이러한 세밀한 디테일은 잡음에 매우 민감하며, 이미지에 잡음이 도입되면 랜덤한 고주파 정보로 표현될 수 있다. 따라서, 디노이징 과정에서는 이러한 복잡한 디테일을 유지하면서 잡음을 효과적으로 제거하는 것이 필요하

다.

디퓨전 모델에서 U-Net 디코더의 각 단계에서는 스킵 연결(skip connection)의 스킵 특징(skip feature)과 백본 특징(backbone feature)이 결합된다. 연구에 따르면 U-Net의 주요 백본(backbone)은 주로 디노이징 과정에 기여하는 것으로 나타났다. 반면, 스킵 연결은 디코더 모듈에 고주파 특징(feature)을 도입하여, 세밀한 의미적(semantic) 정보를 전달하고 입력 데이터의 복구를 용이하게 한다. 그러나 이러한 과정은 추론 단계에서 U-Net의 본래의 디노이징 능력을 약화시킬 수 있다.

이에 대한 해결책으로, FreeU는 U-Net 아키텍처의 주요 백본과 스킵 연결의 기능 균형을 맞추기 위한 두 가지 특수 변조 인자(factor)를 사용한다. 첫 번째인 백본 특징 인자(backbone feature factor)는 백본의 특징 맵(feature map)을 증폭하여 디노이징 과정을 강화하는 데 초점을 맞춘다. 그러나 이러한 증폭은 때때로 텍스처의 과도한 평활화를 야기할 수 있다. 이를 완화하기 위해, 두 번째 인자인 스킵 특징 스케일링 인자(skip feature scaling factor)를 도입하여 텍스처의 과도한 평활화 문제를 줄이는 것을 목표로 한다.

<그림 3-7>은 Stable Diffusion을 기반으로 FreeU의 적용 전후 생성 이미지를 비교한 예시를 나타낸다. FreeU를 적용한 이미지는 프롬프트의 충실도가 높으며, 이미지 품질이 비교적 높은 것을 확인할 수 있다.

이미지 데이터셋 생성에 있어서 FreeU에 기대할 수 있는 이점은, 데이터셋 품질의 향상과 레이블 오류 최소화이다. 또한 추가적인 학습이나 fine-tuning 없이, 단 두 개만의 인자를 추가함으로써 샘플 품질을 크게 향상시킬 수 있는 이점을 지니고 있다.

<그림 3-7> FreeU 미적용 및 적용 샘플 이미지

제4절 LoRA(Low Rank Adaptation)

LoRA(Low Rank Adaptation)는 대규모 언어 모델(Large Language Models, LLM)에 적용되어 성공적인 결과를 보인 기법으로, 이후 딥러닝 모델의 효율적인 미세 조정(Fine-Tuning) 방법으로 발전하였다. LoRA의 주된 목적은 모델의 재학습 없이도 빠르고 효과적으로 새로운 작업이나 데이터셋에 모델을 적응시키는 것이다. 이를 위해, LoRA는 기존의 가중치 행렬 W에 낮은 차원의 수정을 적용하는 방법을 사용한다[37].

$$W' = W + BA \qquad \text{(식 3-1)}$$

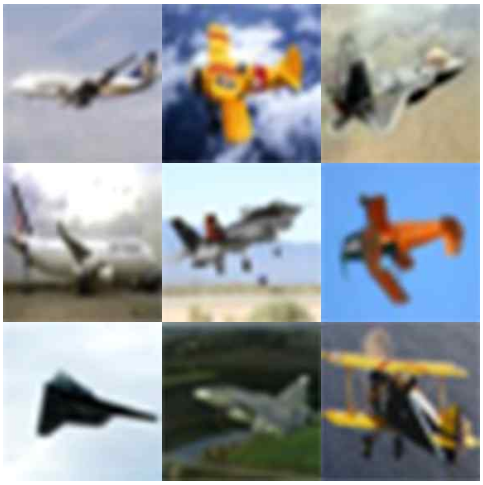
(식 3-1)은 LoRA의 수학적 표현으로, W는 기존 모델의 웨이트 행렬이며 B와 A는 새롭게 도입되는 낮은 차원의 행렬이다. 기존 행렬 W가 $d \times k$일 때,

B는 $d \times r$, A는 $r \times k$이며, $r \ll \min(d,k)$이다. 이러한 변조는 기존 모델의 구조를 그대로 유지하면서도 필요한 부분에만 미세한 조정을 가하여 모델의 성능을 향상시킨다. 이를 통해 모델은 새로운 데이터셋이나 작업에 대해 효율적으로 적응하고, 추가적인 학습이나 복잡한 최적화 과정 없이도 빠른 결과를 도출할 수 있다.

<그림 3-8>은 Stable Diffusion 1.5[20]를 사용하여 생성된 512×512 크기의 이미지를 나타낸다. <그림 3-10>은 <그림 3-9>의 CIFAR-10[38] 학습 데이터셋으로 미세 조정된 LoRA 모델을 Stable Diffusion 1.5에 적용하여 생성한 512×512 크기의 이미지를 보여준다. 이 예시에서 LoRA의 적용이 CIFAR-10 데이터셋의 스타일을 모방하여 이미지를 생성하는 능력을 Stable Diffusion 모델에 부여함을 확인할 수 있다. 노이즈를 추가하는 필터의 개념과 유사해 보일 수 있지만, 해당 데이터셋의 물체 형태 또한 모방한다는 점에서 차이가 있다. 이는 LoRA가 디퓨전 모델의 적응성과 유연성을 증가시키는 강력한 기법임을 시사한다.

디퓨전 모델은 학습 과정에서 상당한 자원을 요구하는데, LoRA를 적용함으로써 이러한 자원 요구를 현저히 감소시킬 수 있다. 이 기법을 통해, 적은 자원을 사용하여 디퓨전 모델을 효과적으로 미세 조정할 수 있으며, 이는 목표 이미지 생성 작업에서 모델의 성능을 향상시킬 것으로 기대된다. 이러한 접근은 사용자가 특정한 데이터셋을 생성하는 데 있어서 범용적이지 않은 요구 사항에도 능동적으로 대응할 수 있게 해주며, 이는 모델의 적용 범위를 넓히고 효율성을 높이는 데 기여할 것으로 기대할 수 있다.

<그림 3-8> Stable Diffusion 1.5



<그림 3-9> CIFAR-10 Train

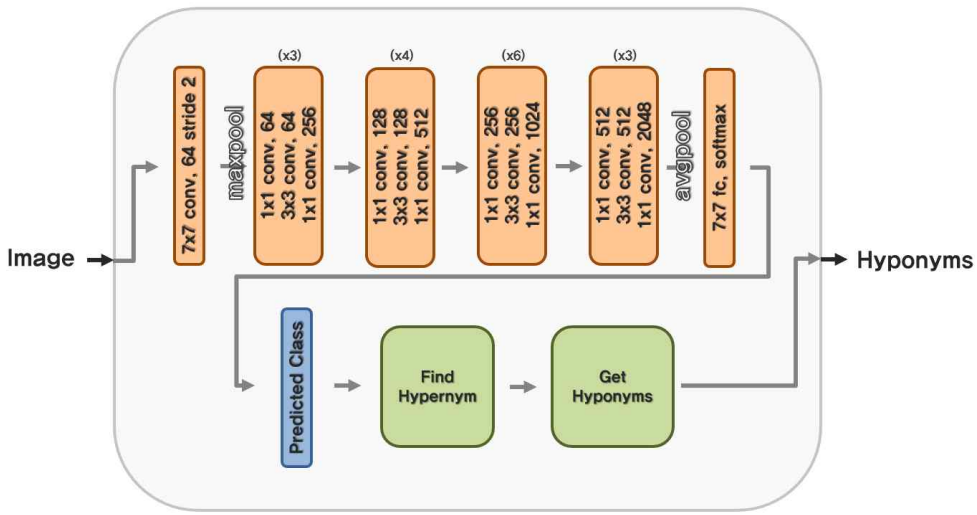
<그림 3-10> Stable Diffusion 1.5 + LoRA

제4장 HypoNet

본 장에서는 디퓨전 모델 기반 효율적인 딥러닝 학습용 데이터셋 생성을 위해 본 논문에서 제안하는 Image-to-Text(i2t)모델인 HypoNet을 설명한다.

HypoNet은 기존 Text-to-Image 모델을 통한 데이터셋 생성의 주요 문제인 다양성의 부족을 해결하기 위해 고안된 Image-to-Text 모델이다. HypoNet은 단일 샘플 이미지를 입력으로 받아 해당 이미지 클래스의 다양한 하위분류(Hyponyms)들을 생성한다. HypoNet이 출력하는 이러한 하위분류들을 Stable Diffusion과 같은 Text-to-Image 모델의 동적 프롬프트로 적용함으로써, 딥러닝 학습에 효과적인 이미지 생성을 가능하게 한다. 이 접근법은 데이터셋의 다양성을 증진시키고, 딥러닝 모델의 성능 향상에 기여할 것으로 기대된다.

제1절 구조 및 원리



<그림 4-1> HypoNet 구조

HypoNet의 기본 구조는 <그림 4-1>과 같으며 원리는 <알고리즘 4-1>과 같다. 이 모델은 단일 이미지를 입력받아 ImageNet-1k[39]로 사전 학습된 ResNet-50[40]을 기반으로 하는 합성곱 신경망을 사용하여 이미지 내의 객체를 식별하고 분류한다. Softmax[41]의 출력에서 가장 높은 점수를 보이는 Top-1 클래스 객체의 상위 개념(Hypernym)을 WordNet[42]을 기반으로 추론하고, 이어서 해당 상위 개념의 하위분류들(Hyponyms)을 최종 출력한다.

상위 개념과 하위분류들은 WordNet 데이터 파일 내에 포함되어 있는 동의어(Synset) 정보와, 관계 유형 정보, 품사 정보를 바탕으로 탐색한다. 탐색된 상위 개념 중 첫 번째로 탐색 된 상위 개념을 기반으로 HypoNet의 최대 탐색 파라미터 값만큼의 하위분류를 출력한다. 이러한 방식으로 HypoNet은 시각적 정보를 언어적 차원으로 확장하는 고도의 추론 작업을 수행한다.

이 과정에서 모델은 주어진 이미지에 대한 분류 예측뿐만 아니라, 해당 분류의 상위 및 하위 개념을 탐색하여, 이미지에 대한 추상적 및 구체적 이해를 동시에 제공한다. 이러한 양방향 접근법은 이미지의 단일 객체 인식을 넘어서, 그 객체가 속한 범주와 그 범주 내 다른 객체들 사이의 관계를 이해하는 데 중요한 역할을 한다.

<그림 4-2>는 Stable Diffusion 모델에 HypoNet을 적용한 예시이다. 이 예시는 HypoNet이 이미지 생성의 다양성을 증진시키는 데 효과적임을 보여준다. 본 논문에서 소개된 HypoNet의 적용은 Text-to-Image 모델과의 결합을 통해 합성 이미지 기반 데이터셋 생성과 같은 다양한 응용 분야에서 중요한 기여를 할 것으로 기대된다. 이는 딥러닝 모델의 학습 데이터셋의 품질과 다양성을 향상시키는 데 큰 잠재력을 지닌다.

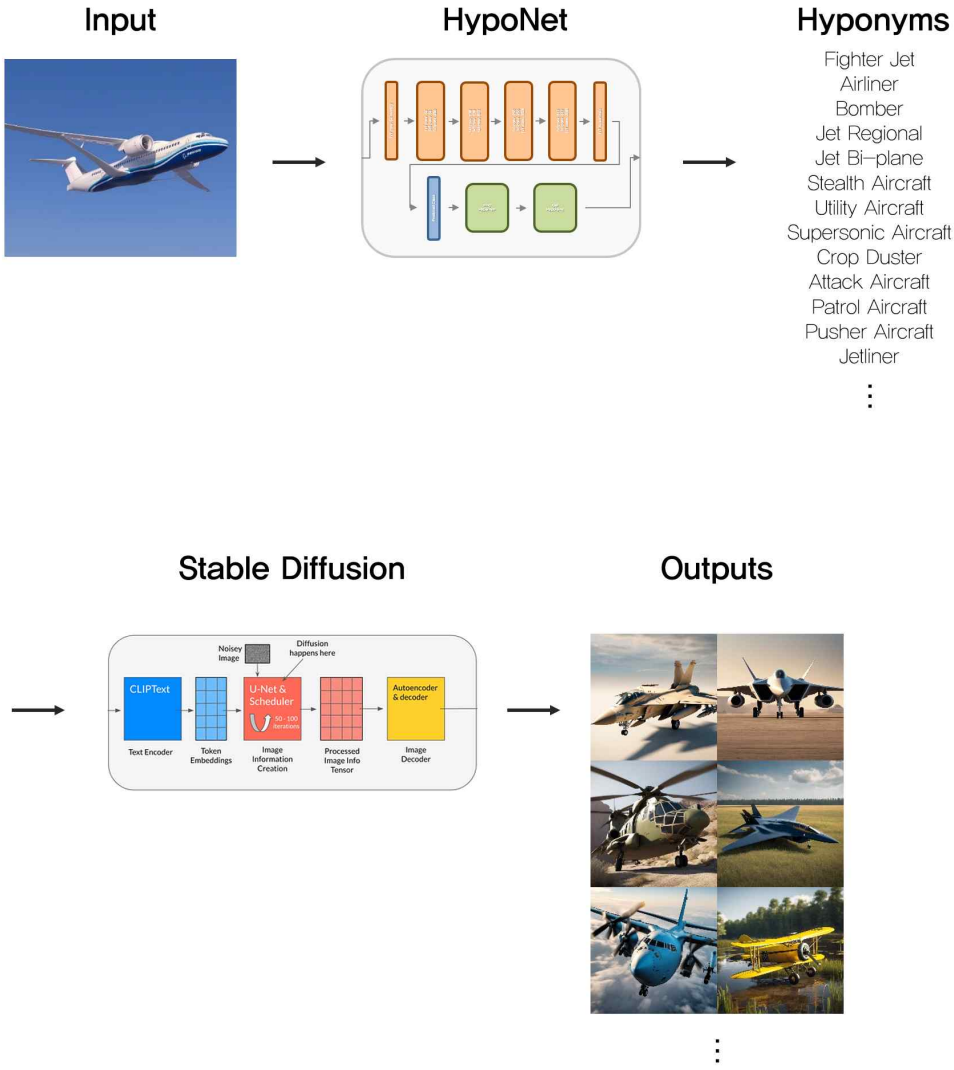
```
Function predict_image_class(image_data)
    model = load ResNet-50 model with ImageNet-1K weights
    set model to evaluation mode
    output = model forward pass with image_data
    _, predicted_class = find maximum value index in output
    return predicted_class

Function find_hypernyms(class_index, class_index_json)
    class_idx_to_label = load JSON file class_index_json
    specific_class_label = find label for class_index in class_idx_to_label
    synsets = find all WordNet synsets for specific_class_label
    hypernym_sets = empty set
    for each synset in synsets
        for each hypernym of synset
            add hypernym to hypernym_sets
    select a hypernym from hypernym_sets
    return selected hypernym

Function get_hyponyms(synset, max_count)
    hyponyms = empty set
    for each hyponym of synset
        add lemma names of hyponym to hyponyms
    if max_count < size of hyponyms
        randomly select max_count hyponyms from hyponyms
    return list of selected hyponyms

Function classify_and_find_hyponyms(image, max_count)
    class_index = predict_image_class(image)
    hypernym_synset = find_hypernyms(class_index, "imagenet_class_index.json")
    hyponyms = get_hyponyms(hypernym_synset, max_count)
    return hyponyms as a list
```

<알고리즘 4-1> HypoNet 알고리즘



<그림 4-2> HypoNet + Stable Diffusion

제2절 적용 결과

본 절에서는 HypoNet의 실제 적용 결과에 대해 논의한다. HypoNet은 Stable Diffusion 모델의 버전 1.4, 1.5, 및 XL에 적용되어 이미지들을 생성하는 데 사용되었다. 이러한 적용을 통해 얻어진 결과는 HypoNet의 효과성과 다양성 증진 능력을 입증한다.

<그림 4-3>에서 <그림 4-17>까지는 HypoNet이 Stable Diffusion 모델에 적용되어 생성된 이미지들의 예시를 보여준다. 각 그림은 5가지의 클래스인 “automobile”, “cat”, “horse”, “ship”, “truck”에 대한 100장의 이미지 예시를 포함한다. 이들 예시는 모델의 버전에 따른 결과의 차이를 보여주는 동시에, HypoNet이 각 클래스에 대해 어떻게 다양한 하위분류를 생성하는지를 시각적으로 입증한다. 예시에서 사용된 HypoNet의 입력 이미지는 ImageNet-1K 학습 데이터셋에서 생성하고자하는 클래스에 해당되는 무작위한 하나의 이미지를 사용하였다.

1. Stable Diffusion 1.4

<그림 4-3>부터 <그림 4-6>까지는 Stable Diffusion 1.4 버전에 HypoNet을 적용한 결과를 보여준다. 이 결과들에서 HypoNet은 각 범주의 이미지를 다양한 스타일과 컨텍스트에서 생성함으로써, 딥러닝 모델의 학습에 필요한 다양한 데이터를 제공한다.

2. Stable Diffusion 1.5

<그림 4-7>부터 <그림 4-11>까지는 Stable Diffusion 1.5 버전에 적용된 HypoNet의 결과를 보여준다. 이 버전에서는 더 개선된 이미지의 품질과 다양성을 관찰할 수 있다.

3. Stable Diffusion XL

<그림 4-13>부터 <그림 4-17>까지는 Stable Diffusion XL 버전에 적용된 HypoNet의 결과를 보여준다. 이 버전에서는 더 높은 해상도와 세부적인 이

미지 생성이 가능하다.

이러한 결과들을 통해 HypoNet이 이미지 생성의 다양성을 향상시키는 데 기여한다는 것을 명확히 볼 수 있다. 또한, HypoNet의 적용은 Stable Diffusion 모델의 다양한 버전에 걸쳐 일관된 효과를 보여주며, 딥러닝 모델 학습을 위한 데이터셋 생성의 새로운 패러다임을 제시한다. 이러한 적용 결과들은 HypoNet이 딥러닝 모델의 성능 향상에 기여할 수 있는 잠재력이 있음을 보여준다.

<그림 4-3> Stable Diffusion 1.4 + HypoNet: automobile

<그림 4-4> Stable Diffusion 1.4 + HypoNet: cat

<그림 4-5> Stable Diffusion 1.4 + HypoNet: horse

<그림 4-6> Stable Diffusion 1.4 + HypoNet: ship

<그림 4-7> Stable Diffusion 1.4 + HypoNet: truck

<그림 4-8> Stable Diffusion 1.5 + HypoNet: automobile

<그림 4-9> Stable Diffusion 1.5 + HypoNet: cat

<그림 4-10> Stable Diffusion 1.5 + HypoNet: horse

<그림 4-11> Stable Diffusion 1.5 + HypoNet: ship

<그림 4-12> Stable Diffusion 1.5 + HypoNet: truck

<그림 4-13> Stable Diffusion XL + HypoNet: automobile

<그림 4-14> Stable Diffusion XL + HypoNet: cat

<그림 4-15> Stable Diffusion XL + HypoNet: horse

<그림 4-16> Stable Diffusion XL + HypoNet: ship

<그림 4-17> Stable Diffusion XL + HypoNet: truck

제5장 실험 및 성능 평가

본 장에서는 이미지 생성 모델을 통해 생성된 합성 이미지의 실용성 및 품질을 비교 분석한다. 제1절에서는 실험 환경을 설명하며 제2절에서는 다양한 기법을 적용하여 데이터셋을 구축하고, 이를 통해 학습된 모델을 세 가지 대중적인 실제 데이터셋에 대해 분류 정확도를 실험하여 합성 이미지의 실용성을 확인한다. 이러한 분석으로써 딥러닝 학습용 데이터셋 구축의 가장 실용적이고 효과적인 방법을 탐색한다. 제3절과 제4절에서는 합성 이미지의 품질을 분석하기 위해 프레셰 인셉션 거리(Frechet Inception Distance, FID)[43] 및 인셉션 점수(Inception Score, IS)[44]과 같은 합성 이미지의 품질 측정 기준을 활용하여, 다양하고 실제 이미지와 유사한 고품질의 합성 이미지가 생성되는 상황을 분석한다.

제1절 실험 환경

본 장의 실험은 동일한 하드웨어 및 소프트웨어 구성에서 실행되었다. 디퓨전 모델을 활용하여 이미지 데이터셋 구축을 위해서는 많은 GPU 메모리가 요구된다. 실험에서, 다양한 데이터셋을 구축하여 비교하기 위해 많은 GPU 메모리를 보유한 NVIDIA RTX 4090을 두 개가 사용되었다. 또한, 학습에 사용된 ResNet-18[40] 모델은 모든 데이터셋에 대해 동등한 비교를 위해 동일한 하이퍼파라미터를 사용하여 훈련을 실시하였다. <표 5-1>은 실험에서의 하드웨어 및 소프트웨어 사양에 대한 상세한 정보를 보여준다.

<표 5-1> 하드웨어 및 소프트웨어 사양

가) 하드웨어 사양

CPU	16 Cores, 32 Threads @ 2.00GHz
RAM	250 GB
GPU	NVIDAIA RTX 4090 * 2
GPU Memory	24 GB
CUDA version	12.0

나) 소프트웨어 사양

PyTorch	2.0.1+cu118
Torchvision	0.15.2+cu118
Python	3.8.10

분류 정확도를 분석하기 위해 ResNet-18에서 사용한 학습 파라미터는 <표 5-2>와 같다.

<표 5-2> ResNet-18 학습 파라미터

Batch size	32
Learning rate	10^{-3}
Opimiser	AdamW
Epoch	50
Downsample	Bicubic
Augmentation	-

동적 프롬프트를 적용하여 이미지 생성시 사용된 와일드카드와 각 하위분류는 <표 5-3>와 같다. 괄호의 값은 해당 와일드카드의 하위분류 갯수를 의미한다. 해당 와일드카드의 하위분류는 HypoNet을 통해 추출되었다.

〈표 5-3〉 객체별 와일드카드 및 하위분류

와일드카드	하위분류
automobile (37)	Sedan
	SUV
	Coupe
	Hatchback
	Convertible
	Minivan
	Pickup Truck
	Sports Car
	Electric Car
	Hybrid Car
	Station Wagon
	Off-Road Vehicle
	Luxury Car
	Compact Car
	Jeep
	Subcompact Car
	Midsize Car
	Van
	Multi-Purpose Vehicle
	Racing Car
	Diesel Car
	Muscle Car
	Microcar
	Antique Car
	Classic Car
	Concept Car
	Rally Car
	Roadster
	Police Car
	Targa Top
	Panel Van
	Delivery Van
	Passenger Van
	Mini-SUV
	Hot Rod
	Lowrider
	Taxi
cat (40)	Domestic Shorthair
	Siamese Cat
	Maine Coon
	Persian Cat
	Bengal
	British Shorthair
	Abyssinian
	Scottish Fold
	Norwegian Forest Cat
	Oriental Shorthair
	American Shorthair

와일드카드	하위분류
horse (38)	Arabian Horse
	Thoroughbred
	Clydesdale
	Andalusian
	Morgan Horse
	American Paint Horse
	Appaloosa
	Quarter Horse
	Shire Horse
	Lipizzaner
	Belgian Draft
	Tennessee Walking Horse
	Hanoverian
	Friesian
	Gypsy Vanner
	Dutch Warmblood
	Standardbred
	Lusitano
	Haflinger
	Saddlebred
	Peruvian Paso
	Akhal-Teke
	Trakehner
	Irish Draught
	Welsh Pony
	Palomino
	Connemara Pony
	American Cream Draft
	Missouri Fox Trotter
	Holsteiner
	Rocky Mountain Horse
	Australian Stock Horse
	Swedish Warmblood
	Cremello
	Barb Horse
	Suffolk Punch
	American Curly Horse
	Racking Horse
ship (25)	Fishing Boat
	Yacht
	Sailing Ship
	Tugboat
	Research Vessel
	Patrol Boat
	Galleon
	Clipper Ship
	Luxury Yacht
	PT Boat

	Cornish Rex
	Birman
	Devon Rex
	Russian Blue
	Manx
	Balinese
	Turkish Van
	Exotic Shorthair
	Egyptian Mau
	Tonkinese
	Turkish Angora
	Chartreux
	Nebelung
	British Longhair
	Ocicat
	Korat
	American Curl
	American Bobtail
	Japanese Bobtail
	Selkirk Rex
	California Spangled
	Cymric
	Australian Mist
	Burmese
	Chausie
	Colorpoint Shorthair
	Domestic Longhair
	German Rex
	Khao Manee

	Amphibious Assault Ship
	Rowboat
	Swift Boat
	Jet Ski
	Rigid Inflatable Boat
	Dredger
	Bulk Carrier
	Schooner
	Clipper Ship
	Galleon
	Patrol Boat
	Trimaran
	Riverboat
	Ferry
	Container Ship
truck (23)	Semi-Trailer Truck
	Tipper Truck
	Box Truck
	Dump Truck
	Flatbed Truck
	Tow Truck
	Fire Truck
	Garbage Truck
	Refrigerated Truck
	Utility Truck
	Boom Truck
	Delivery Truck
	Roll-Off Truck
	Wrecker
	Straight Truck
	Cabover Truck
	Recovery Truck
	Moving Truck
	Emergency Truck
	Chipper Truck
	Ladder Truck
	Reefer Truck
	Service Truck

FreeU[36]를 적용하여 이미지 생성시 사용한 파라미터는 <표 5-4>과 같다.

<표 5-4> FreeU 파라미터

파라미터	Stable Diffusion (1.4, 1.5)	Stable Diffusion XL
Start At Step	0	0
Stop At Step	1	1
Transition Smoothness	0	0
Stage 1		
Backbone 1 Scale	1.2	1.1
Backbone 1 Offset	0	0
Backbone 1 Width	0.5	0.5
Skip 1 Scale	0.9	0.6
Skip 1 High End scale	1	1
Skip 1 Cutoff	0	0
Stage 2		
Backbone 2 Scale	1.4	1.2
Backbone 2 Offset	0	0
Backbone 2 Width	0.5	0.5
Skip 2 Scale	0.2	0.4
Skip 2 High End scale	1	1
Skip 2 Cutoff	0	0
Stage 3		
Backbone 3 Scale	1	1
Backbone 3 Offset	0	0
Backbone 3 Width	0.5	0.5
Skip 3 Scale	1	1
Skip 3 High End scale	1	1
Skip 3 Cutoff	0	0

LoRA[37]를 적용하여 생성시 LoRA모델 학습 파라미터는 <표 5-5>와 같다.

<표 5-5> LoRA 파라미터

파라미터	Stable Diffusion (1.4, 1.5)	Stable Diffusion XL
Batch size	5	5
Epoch	7	7
Mixed precision	bf16	bf16
LR Scheduler	constant	constant
Optimizer	Adafactor	Adafactor
Learning rate	0.0012	0.0012
LR warmup	0	0
Max resolution	512,512	1024,1024
Min bucket resolution	256	256
Max bucket resolution	2048	2048

LoRA 모델을 학습시, CIFAR-10 train의 10개 클래스에서 40장의 무작위 이미지를 추출 후 BLIP[45]을 통해 해당 이미지를 Captioning 하였다. 또한 각 이미지는 R-ESRGAN 4x+[46]를 통해 512x512사이즈로 upsampling 하였다.

Stable Diffusion[20]을 통해 이미지 생성시 사용된 파라미터값은 <표 5-6>와 같다.

<표 5-6> Stable Diffusion 파라미터

파라미터	Stable Diffusion (1.4, 1.5)	Stable Diffusion XL
Stpes	30	30
CFG scale	7.5	7.5
Sampler	DPM++ 2M Karras	DPM++ 2M Karras
Seed	-1	-1
Size	512,512	1024,1024
Refiner switch	-	0.8

제2절 분류 정확도

해당 절에서는 본 논문의 제3장에서 소개된 기법들의 효과를 평가하기 위해 Stable Diffusion[20]의 세 가지 버전(1.4, 1.5, XL)을 기반으로 적용하여, “automobile“, “cat“, “horse“, “ship“, “truck“의 5개 클래스에 대한 이미지를 생성한다. 각 기법이 적용된 모델에서 생성한 데이터셋을 ResNet-18[40] 아키텍처를 사용하여 학습시킨 후 CIFAR-10 test, STL-10 test, ImageNet-1k validation set의 세 가지 데이터셋으로 검증하여 분류 정확도를 분석하는 것이다[38,47,39]. 이를 통해 각 기법이 딥러닝 모델의 학습 데이터셋 생성에 어떠한 영향을 미치는지, 그리고 어떤 기법이 가장 효과적인 학습 데이터셋을 생성하는지를 파악할 수 있다.

<표 5-7>는 이미지 생성시 사용된 고정 긍정 프롬프트와 고정 부정 프롬프트이며, LoRA[37]를 적용하여 생성시 고정 긍정 프롬프트를 사용하였으며, HypoNet을 적용하여 생성시 고정 긍정 프롬프트와 고정 부정 프롬프트를 사용하였다. 특정 기법들에서는 긍정 및 부정 프롬프트를 적용했을 때, 출력 이미지에 부정적인 결과가 관찰되었다. 이러한 고정 프롬프트는 경험적 연구(Empirical research)를 통해 선정되었으며, 특정 시나리오에서 예상치 못한 결과를 초래한 것으로 나타났다. 또한 1.4와 같은 가중치는 프롬프트의 중요도를 상승시키는데, 이 또한 경험적 연구를 통해 결정된 값이다.

<표 5-7> 고정 긍정 프롬프트와 고정 부정 프롬프트

Fixed Positive prompt (Fp)	best quality, 8k, realistic, photo-realistic:1.4
Fixed Negative prompt (Fn)	worst quality:1.4, low quality:1.4, normal quality:1.4, watermark:1.4, nsfw:1.4, drawings, abstract art, cartoons, surrealist painting, conceptual drawing, graphics, bad proportions, human, person, people, man, woman, girl, boy

각 기법을 적용하였을 때, 사용된 프롬프트는 <표 5-8>과 같으며, 괄호안의 프롬프트는 와일드카드를 의미한다. 또한, “cat”과 “horse”클래스에 대해서는 “color”와일드카드를 사용하지 않았다.

<표 5-8> 긍정 프롬프트와 부정 프롬프트

	긍정 프롬프트	부정 프롬프트
none	automobile cat horse ship truck	-
FreeU	automobile cat horse ship truck	-
LoRA	(view) of a (color) (automobile):1.4 in (location), Fp (view) of a (color) (cat):1.4 in (location), Fp (view) of a (color) (horse):1.4 in (location), Fp (view) of a (color) (ship):1.4 in (location), Fp (view) of a (color) (truck):1.4 in (location), Fp	-
HypoNet	(view) of a (color) (automobile):1.4 in (location), Fp (view) of a (color) (cat):1.4 in (location), Fp (view) of a (color) (horse):1.4 in (location), Fp (view) of a (color) (ship):1.4 in (location), Fp (view) of a (color) (truck):1.4 in (location), Fp	Fn

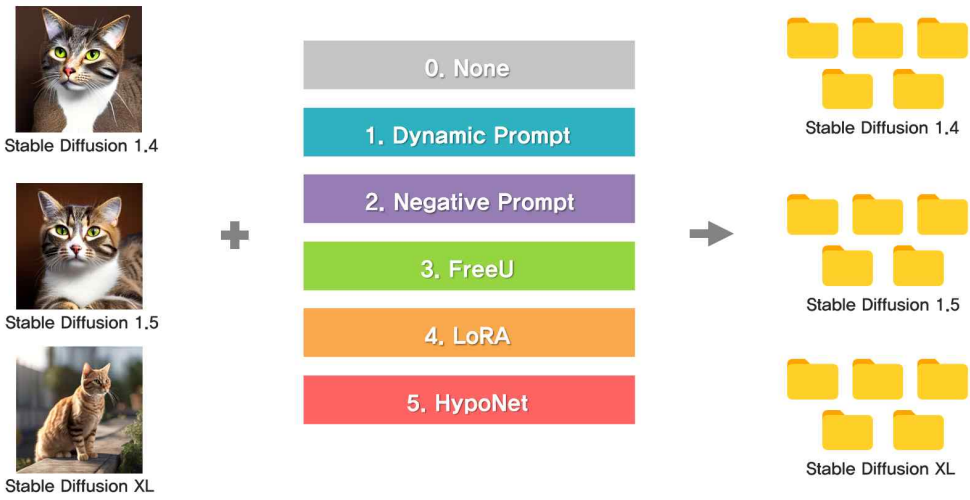
1. 순수 합성 데이터로의 학습

본 실험에서는 실제 데이터를 사용하지 않고 오로지 합성 이미지만을 이용한 학습의 분류 정확도를 측정한다. 또한 다양한 기법으로 생성된 이미지가 딥러닝 학습에 얼마나 효과적이고 실용적인지를 평가할 수 있다. 이러한 접근은 합성 이미지를 활용한 학습 방법론의 실질적인 가치와 적용 가능성을 탐색하는 데 중요한 기여를 한다.

<그림 5-1>은 본 실험의 데이터셋 생성 과정을 보여주며, <표 5-9>는 본

논문에서 가장 핵심 내용을 포함하고있는 학습 데이터별 분류 정확도를 나타낸다. CIFAR-10 Train[38] 데이터셋으로 학습시 3가지 데이터셋에 대한 평균 정확도는 70.50%였으며, 이와 동등하거나 그 이상의 값을 가질 경우 합성 데이터로의 학습 데이터셋 대체 가능성이 존재함을 나타낸다. 실험 결과, HypoNet을 활용하여 동적 프롬프트를 사용하고 적절한 부정 프롬프트를 사용시에 가장 높은 평균 정확도를 보임을 확인할 수 있었다.

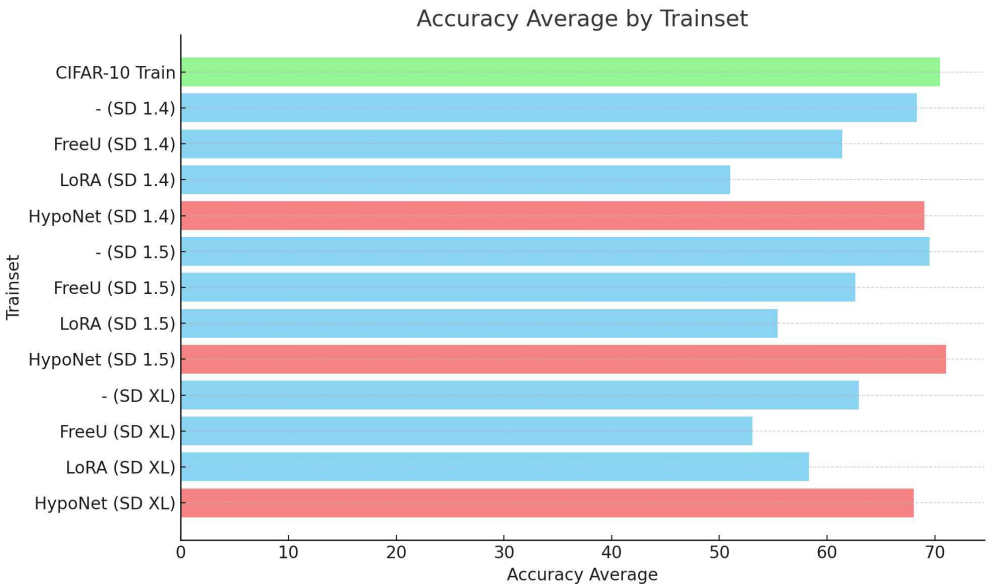
특히, Stable Diffusion 1.5[20]의 경우 CIFAR-10 Train 데이터셋을 사용하여 학습한 결과보다 미량 높은 정확도를 보였다. 이는 HypoNet을 통한 동적 프롬프트 사용과 적절한 부정 프롬프트 사용시에 가장 효과적인 이미지 데이터셋을 생성할 수 있으며, 합성 데이터가 실제 데이터를 대체하여 딥러닝 학습용 데이터셋을 생성할 수 있음을 시사한다.



<그림 5-1> 실험을 위한 데이터셋 구축 과정

<표 5-9> 학습 데이터셋별 분류 정확도와 평균 정확도(%)

Trainset \ Testset	CIFAR-10 ↑	ImageNet-1k ↑	STL-10 ↑	Average ↑
CIFAR-10 Train	84.78	66.11	60.61	70.50
Stable Diffusion 1.4				
-	69.86	71.33	63.75	68.31
FreeU	63.16	58.67	62.38	61.40
LoRA	53.32	54.44	45.20	50.99
HypoNet(ours)	75.32	**73.11**	58.70	69.04
Stable Diffusion 1.5				
-	68.96	70.00	69.50	69.49
FreeU	66.56	61.11	60.13	62.60
LoRA	61.42	41.11	63.65	55.39
HypoNet(ours)	74.42	**73.11**	**65.62**	**71.05**
Stable Diffusion XL				
-	61.18	64.44	63.23	62.95
FreeU	61.32	55.78	42.15	53.08
LoRA	59.50	56.67	58.75	58.31
HypoNet(ours)	72.50	**76.22**	55.38	68.03



<그림 5-2> 학습 데이터셋별 평균 분류 정확도 그래프

2. 실제 데이터와 합성 데이터로의 증강 후 학습

CIFAR-10 Train[38] 데이터셋은 10개의 클래스로 이루어져 있으며 클래스당 5,000장의 이미지가 포함되어있다. 해당 데이터셋에 합성 이미지를 추가하여 학습함으로써 실제 이미지에 대한 합성 이미지의 증강이 정확도에 어떠한 영향을 끼치는지를 분석한다.

<표 5-10>는 CIFAR-10 Train 데이터셋에 HypoNet을 적용하여 Stable Diffusion 1.5[20] 모델로 생성한 합성 이미지를 증강 데이터로 추가했을 때의 정확도 변화를 보여준다. 표에 나타난 값들은 각 클래스별 이미지의 개수를 나타내며, 실험 결과에서는 실제 데이터셋에 합성 데이터를 추가함으로써 정확도가 향상되는 경향을 관찰할 수 있다. 이러한 결과는 적절한 합성 데이터가 데이터 증강에 긍정적인 영향을 미침을 시사한다. 또한, 증강 이미지의 개수 증가와 정확도 상승 사이에 직접적인 비례 관계가 항상 성립하지 않는다는 발견도 이루어졌다. 이는 데이터 증강의 효과가 항상 선형적이지 않음을 보여주며, 최적의 증강 전략을 수립하는 데 중요한 기준이 될 수 있다.

<표 5-10> 증강 합성 이미지 개수별 정확도(%)

CIFAR-10 Train / per class	Stable Diffusion 1.5 (HypoNet) / per class	정확도↑
5,000	0	84.78
5,000	1,000	**88.22**
5,000	2,000	**87.64**
5,000	3,000	**88.88**
5,000	4,000	**89.24**
5,000	5,000	**88.94**

제3절 FID

FID(Fréchet Inception Distance)[43]는 최근 고해상도 이미지 생성 모델의 성능을 평가하는 데 널리 사용되는 지표이다. 해당 절에서는 FID를 활용하여 실험 결과를 분석한다. FID는 합성 이미지와 실제 이미지 사이의 통계적 거리를 측정하여 이미지의 품질을 평가하는 데 사용된다. 이 지표를 통해 합성 이미지의 질적 수준과 CIFAR-10 Train[38] 데이터셋과의 유사성을 정량적으로 평가하고, 이를 통해 본 논문에서 설명된 기법들이 생성한 이미지의 질적 특성을 면밀히 분석한다. FID 점수가 낮을수록 합성 이미지의 품질이 더 높고, CIFAR-10 Train 데이터셋에 더 근접함을 나타낸다.

$$FID = |\mu - \mu_w|^2 + tr(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2}) \quad \text{(식 5-1)}$$

(식 5-1)은 FID의 수학적 정의를 나타낸다. 이 식에서, $|\mu - \mu_w|^2$는 두 분포의 평균 간의 거리를 나타내며, $tr(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2})$는 두 공분산 행렬 간의 차이를 측정한다.

일변수(Univariate) 분포의 경우, 공분산 행렬은 대각 성분 외에 모든 요소가 0인 형태로 간소화된다. 이때 FID는 (식 5-2)과 같이 단순하게 표현된다.

$$FID = |\mu - \mu_w|^2 + |\sigma - \sigma_w|^2 \quad \text{(식 5-2)}$$

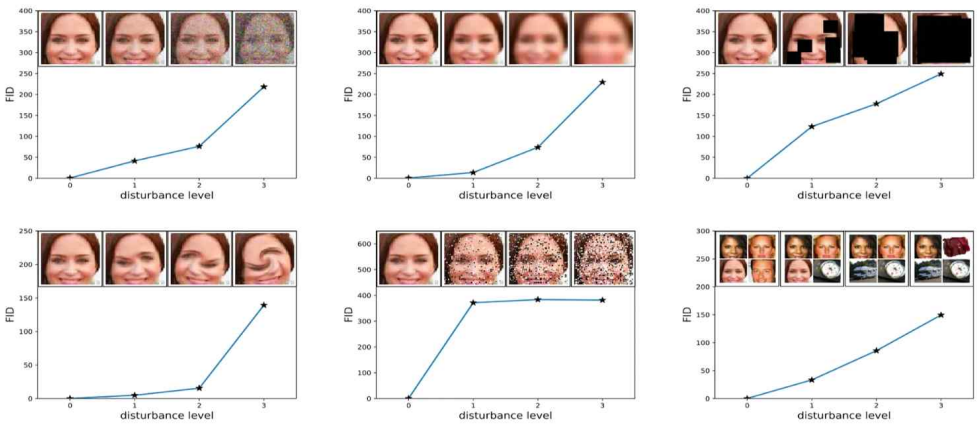
(식 5-2)에서, $|\sigma - \sigma_w|^2$는 두 분포의 표준편차 간의 차이를 나타낸다. 이러한 접근 방식은 단순 픽셀 값의 차이를 측정하는 pixel distance보다 훨씬 의미 있는 비교를 제공한다.

FID를 계산하기 위해 사용되는 값은 Inception v3[48] 모델의 coding layer, 즉 이미지 클래스를 분류하기 직전의 가장 마지막 pooling layer에서 추출된 값이다. 이는 다른 사전 훈련된 모델에도 적용될 수 있으며, 분류 직전 layer의 값을 사용함으로써 FID를 산출할 수 있다.

이러한 방식으로 FID는 실제와 생성된 이미지 간의 시각적 유사성을 효과적으로 측정하며, 고해상도 이미지 생성 모델의 성능을 정량적으로 평가하는 데 중요한 역할을 한다.

또한 <그림 5-3>은 이미지 왜곡과 FID의 상관관계를 보여준다. 여기서 왜곡이 증가할수록 FID 점수가 상승함을 확인할 수 있다.

<표 5-11>은 각 데이터셋에 대한 FID 측정 결과를 보여준다. CIFAR-10 Train 데이터셋을 기준으로 한 FID 측정에서, CIFAR-10 Train 데이터셋의 FID 점수는 0으로 나타난다. 이는 두 데이터셋 간 완벽한 일치를 의미한다. 반면, ImageNet-1k Train 데이터셋은 CIFAR-10 Train 데이터셋과 비교하여 약 38.13의 FID 값을 가진다. 특히 주목할 점은, 모든 Stable Diffusion 모델에서 HypoNet을 적용한 데이터셋이 가장 낮은 FID 점수를 기록했다는 것이다. CIFAR-10 스타일을 모방한 LoRA 데이터셋은 시각적으로는 CIFAR-10과 유사해 보일 수 있으나, LoRA 학습시의 업샘플링 및 FID 측정시의 다운샘플링으로인한 데이터 노이즈 등의 요인으로 인해 상대적으로 높은 FID 점수를 가진 것으로 예상된다.



<그림 5-3> 이미지 왜곡과 FID 점수의 상관관계

<표 5-11> 데이터셋별 Fréchet inception distance(FID) 비교

Dataset	FID↓
CIFAR-10 Train	0
CIFAR-10 Test	3.24
ImageNet-1k Train	38.13
Stable Diffusion 1.4	
-	34.05
FreeU	44.12
LoRA	40.53
HypoNet(ours)	**25.99**
Stable Diffusion 1.5	
-	33.19
FreeU	43.09
LoRA	26.08
HypoNet(ours)	**24.08**
Stable Diffusion XL	
-	38.14
FreeU	51.75
LoRA	27.77
HypoNet(ours)	**24.91**

제4절 IS

IS(Inception Score)[44]는 이미지 생성 모델의 성능을 평가하는 데 주로 사용되는 지표로, 특히 생성된 이미지의 다양성과 명확성을 동시에 측정한다. 이 절에서는 IS를 사용하여 실험 결과를 분석한다. IS는 이미지가 얼마나 다양한 클래스에 속하는지, 그리고 각 클래스에 속하는 이미지가 얼마나 명확하게 분류되는지를 평가한다. 이를 통해 생성된 이미지의 질적 수준과 해당 클래스와의 일치도를 정량적으로 평가하며, 이로써 본 논문에서 소개된 기

법들이 생성한 이미지의 질적 특성을 면밀히 분석한다. IS 점수가 높을수록 이미지의 다양성과 명확성이 높음을 나타낸다.

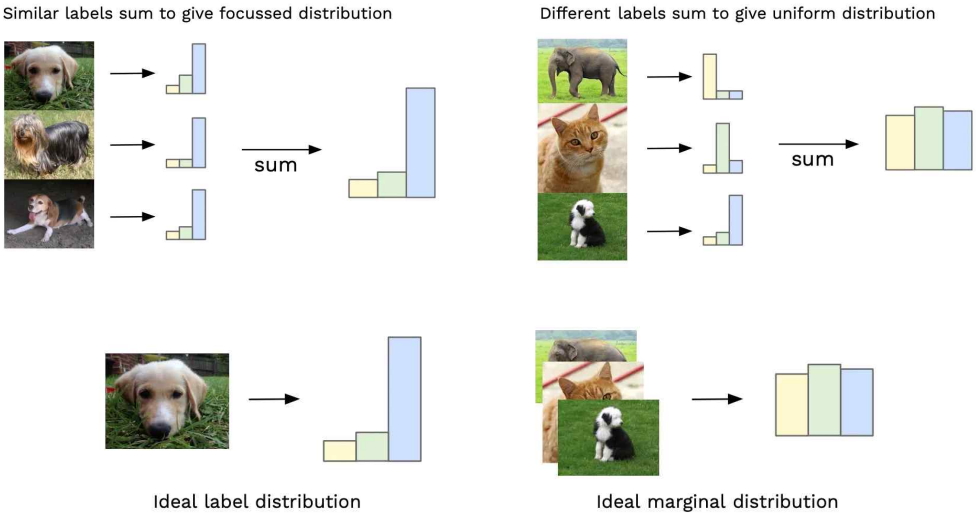
IS의 계산은 Inception v3[48] 모델을 사용하여 각 이미지가 얼마나 특정 클래스에 속할 확률이 높은지를 측정함으로써 이루어진다. 이 확률 분포의 엔트로피를 계산하여 이미지의 명확성을 평가하고, 이를 여러 이미지에 대해 평균내어 이미지 세트의 다양성을 평가한다. 수학적 정의는 (식 5-3)과 같다.

$$IS(G) = \exp(Ex \sim p_g D_{KL}(p(y|x) \| p(y))) \quad \text{(식 5-3)}$$

D_{KL}은 KL발산(Kullback-Leibler divergence)[49]을 나타내며, $Ex \sim p_g$는 모든 결과의 합계와 평균을 의미한다. 생성된 이미지를 Inception v3 모델을 통해 처리하여 조건부 확률 분포 $p(y|x)$를 도출하고 주변 확률 분포 $p(y)$를 계산한다. 이어서 $p(y)$와 $p(y|x)$사이의 KL발산을 측정한다. 각 클래스에 대한 합계를 산출하고 모든 이미지에 대해 평균 점수를 계산한다. 최종적으로 이 모든 결과의 평균값$Ex \sim p_g$를 계산하고 지수를 취하여 IS를 산출한다.

이러한 방식으로 IS는 생성된 이미지의 질적 수준을 효과적으로 평가하며, 이미지 생성 모델의 성능을 정량적으로 평가하는 데 중요한 역할을 한다.

<그림 5-4>는 IS를 시각적으로 설명하며, <표 5-12>는 데이터셋별 IS 측정 결과를 나타낸다. HypoNet을 적용하여 생성된 데이터셋은 일관되게 높은 IS 점수를 기록하였으며, 특히 FreeU[36]가 적용된 경우에는 IS에서 탁월한 점수를 보였다. 이러한 결과는 HypoNet과 FreeU가 이미지 생성 과정에서 다양한 고품질의 이미지를 생성하는 데에 긍정적인 영향을 미침을 시사한다.



<그림 5-4> Inception Score

<표 5-12> 데이터셋별 Inception Score(IS) 비교

Dataset	IS ↑
CIFAR-10 Train	10.23
Stable Diffusion 1.4	
-	11.48
FreeU	13.11
LoRA	9.10
HypoNet(ours)	12.95
Stable Diffusion 1.5	
-	11.46
FreeU	12.82
LoRA	11.11
HypoNet(ours)	12.34
Stable Diffusion XL	
-	10.35
FreeU	9.52
LoRA	9.65
HypoNet(ours)	10.63

제5장 결론

본 연구에서는 디퓨전 모델을 활용한 효과적인 딥러닝 학습용 데이터셋 생성에 대한 새로운 가능성을 탐구하였다. 최근 급격한 발전을 보이는 디퓨전 모델을 이용하여 고품질의 합성 이미지를 생성하고, 이를 딥러닝 아키텍처에 적용하여 다양한 범용적인 데이터셋에서 이미지 분류 성능을 평가함으로써 실질적인 활용 가능성을 입증하였다.

기존 이미지 생성 모델의 문제인 출력 다양성 부족 문제를 극복하기 위해 단일 샘플 이미지에서 다수의 하위 카테고리를 생성할 수 있는 새로운 Image-to-Text (i2t) 모델인 HypoNet을 제안하였다. 이 모델과 함께 적절한 부정 프롬프트를 적용하여 이미지를 생성할 경우 다른 최신 기법들과의 비교에서 우수한 결과를 보여주었으며, 추가적인 학습 없이 온전히 합성 이미지만으로 구축한 데이터셋임에도 실제 이미지 데이터셋을 통한 학습과 비교하여 동등하거나 그 이상의 성능을 나타냈다.

Stable Diffusion XL과 같은 거대 모델은 높은 품질의 이미지를 생성했지만 실험 결과 낮은 정확도와 품질 점수를 받았다. 이는 해당 모델이 큰 사이즈의 이미지를 생성하기 때문에 ResNet 및 InceptionNet과 같은 아키텍처에 입력되기 전 이미지 다운샘플링 과정의 노이즈와, 해상도 감소로 인한 미세한 디테일이 손실이 전체적인 이미지 품질에 영향을 미친 것으로 파악된다.

본 연구 결과는 디퓨전 모델을 통한 데이터셋 생성이 실용적 대안으로서의 가능성을 제시하며, 대규모 실세계 데이터 수집에 대한 의존도를 크게 감소시킬 수 있는 새로운 방법론을 제시하였다. 즉, 기존의 복잡한 데이터셋 구축의 방식에서 벗어나 합성 이미지로의 데이터셋 구축이라는 새로운 경로를 밝혔다. 이러한 결과는 미래의 딥러닝 연구 및 응용 분야에 중요한 영향을 미칠 것으로 기대된다.

참고문헌

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115:211-252, 2015.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. arXiv preprint arXiv:2304.02643, 2023.

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the Point: Semantic Segmentation with Point Supervision. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII 14, pages 549-565. Springer, 2016.

[4] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-Modal Contrastive Learning for Text-to-Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 833-842, 2021.

[5] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). Tech Report, 1996.

[6] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In IEEE workshop on applications of computer vision, 1994.

[7] Athinodoros S Georghiades, David J Kriegman, and PN Belhurneur. Illumination cones for recognition under variable lighting: Faces. In CVPR, 1998.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[9]. Stöckl, A. Evaluating a Synthetic Image Dataset Generated with Stable

Diffusion. 2022.
[10] Tian, Y., Fan, L., Isola, P., Chang, H., & Krishnan, D. StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. 2023.
[11] Voetman, R., Aghaei, M., & Dijkstra, K. The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Detection Models. 2023.
[12] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. 2016.
[13] Rajpura, P. S., Bojinov, H., & Hegde, R. S. Object Detection Using Deep CNNs Trained on Synthetic Images. 2017.
[14] Tremblay, J., To, T., & Birchfield, S. SIDOD: A synthetic dataset for 3D object detection and pose estimation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2018-June, 2119-2122. 2018.
[15] Hurl, B., Czarnecki, K., & Waslander, S. Precise Synthetic Image and LiDAR (PreSIL) Dataset for Autonomous Vehicle Perception. 2019.
[16] Anderson, J. W., Ziolkowski, M., Kennedy, K., & Apon, A. W. Synthetic Image Data for Deep Learning. 2022.
[17] Li, X., Wang, Y., Yan, L., Wang, K., Deng, F., & Wang, F. Y. ParallelEye-CS: A New Dataset of Synthetic Images for Testing the Visual Intelligence of Intelligent Vehicles. IEEE Transactions on Vehicular Technology, 68(10), 9619-9631. 2019.
[18] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In ICML, 2015.
[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020.
[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser,

and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. Communications of the ACM, 63(11):139-144, 2020.

[22] Dhariwal, P., & Nichol, A. Diffusion Models Beat GANs on Image Synthesis. 2021.3115109-2951

[23] Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862.

[24] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. arXiv preprint arXiv:1503.03585.

[25] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. arXiv preprint arXiv:2105.05233.

[26] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. arXiv preprint arXiv:2112.10752.

[27] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[28] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[29] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[30] Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Advances in Neural Information Processing Systems (pp. 10236-10245).

[31] Van den Oord, A., Li, Y., Vinyals, O., Kavukcuoglu, K., Dyer, C., &

Bengio, S. (2017). Neural discrete representation learning. In Advances in neural information processing systems (pp. 6306-6315).

[32] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2016). Beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations.

[33] Ho, J., Jain, A., & Abbeel, P. (2019). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.

[34] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092.

[35] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[36] Si, C., Huang, Z., Jiang, Y., & Liu, Z. (2023). FreeU: Free Lunch in Diffusion U-Net. arXiv preprint arXiv:2309.11497.

[37] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In International Conference on Learning Representations. Available at

[38] Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Toronto.

[39] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248-255). IEEE.

[40] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[41] Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Neurocomputing: Algorithms, Architectures and Applications (pp. 227-236). Springer, Berlin, Heidelberg.

[42] Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39-41.

[43] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Advances in Neural Information Processing Systems (pp. 6626-6637).

[44] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. In Advances in Neural Information Processing Systems (pp. 2234-2242).

[45] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Dong, J. (2022). BLIP: Bootstrapped Language Image Pretraining for Vision-Language Foundation Models. arXiv preprint arXiv:2201.12086.

[46] Wang, X., Yu, K., Dong, C., & Loy, C. C. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. arXiv preprint arXiv:2107.10833.

[47] Coates, A., Ng, A., & Lee, H. (2011). An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 215-223).

[48] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[49] Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

[50] Andrew, How does negative prompt work?,

https://stable-diffusion-art.com/how-negative-prompt-work/, (2023, 9, 27)

국문초록

생성형 AI를 활용한 효과적인 딥러닝 학습용 데이터셋 생성에 대한 연구

본 논문은 Text-to-Image (t2i) 모델을 이용하여 생성된 합성 이미지를 딥러닝 학습용 데이터셋으로의 대체 가능성을 탐구한다. 기존 연구에서는 이미지 합성과 3D 시뮬레이션을 통한 데이터셋 확장의 가능성을 모색했지만, 실용성에 있어서 한계가 있었다. 본 연구에서는 최근 발전을 거듭하고 있는 Diffusion Model을 활용하여 고품질의 합성 이미지를 생성하고, 이를 딥러닝 아키텍처에 적용 후 세 가지 범용적인 데이터셋에서 이미지 분류 성능을 평가함으로써 실질적인 사용 가능성을 검증한다.

또한, 기존 연구에서 지적된 출력의 다양성 부족을 개선하기 위해, 단일 샘플 이미지로부터 다수의 하위 카테고리를 생성할 수 있는 새로운 Image-to-Text (i2t) 모델인 HypoNet을 제시한다. 또한 FreeU, LoRA와 같은 최신 기법을 적용하여 생성한 이미지들과의 평가 결과를 비교한다. 결과적으로, 모델의 추가적인 학습 없이 HypoNet을 적용하여 생성한 경우, 온전히 합성 이미지만으로 구축한 데이터셋임에도 실제 이미지 데이터셋을 통한 학습과 비교하여 동등하거나 그 이상의 성능을 보였다. 본 연구는 기존 모델에서 추가적인 학습없이도 Diffusion Model을 통한 데이터셋 생성은 실용적 대안이 될 수 있음을 시사한다. 이는 대규모 실세계 데이터 수집에 대한 의존도를 크게 줄일 수 있는 데이터셋 생성의 새로운 길을 제시한다.

주제어 : 데이터셋 생성, 딥러닝, 디퓨전 모델, 이미지 분류