Diffusion Inversion을 활용한 이미지 창의성의 정량적 측정 모델 개발

신영민
학번: 20201092

2025년 12월

Abstract

본 연구는 확산 모델(Diffusion Model)의 통계적 특성을 활용하여 이미지의 창의성을 정량적으로 측정하는 방법론을 제안한다. 창의성을 "학습된 분포로부터의 일탈"로 정의하고, DDIM Inversion을 통한 재구성 오차를 측정 지표로 삼았다. WikiArt 인간 작품 3,000점과 Stable Diffusion 1.5 생성 이미지 3,000점을 분석한 결과, 초기 가설과 달리 AI 이미지가 더 높은 복원 오차(평균 0.684 대 0.628, $p < 10^{-47}$)를 나타냈다. 이는 WikiArt 회화가 Stable Diffusion 학습 데이터에 포함되어, 오히려 인간 작품이 모델에 익숙하기 때문으로 해석된다. 복합 점수(1-코사인유사도, LPIPS, 로그정규화 MSE의 평균)와 시대별·화가별 분석을 통해 이 현상의 일관성을 검증하였다.

주요어: 창의성 측정, 확산 모델, DDIM Inversion, 재구성 오차, AI 생성 이미지, WikiArt

Contents

1 서론

1.1 연구 배경

생성 모델의 발전으로 Stable Diffusion, DALL-E, Midjourney 등은 전문 예술가의 작품과 구별하기 어려운 이미지를 생성한다 [1]. 이에 따라 "창의성"이라는 개념을 AI에 적용할 수 있는지에 대한 논의가 활발해졌다.

창의성 연구에서는 새로움과 유용성을 핵심 요소로 보지만, 이를 계산 가능한 형태로 정량화하는 방법은 아직 확립되지 않았다.

1.2 연구 목적

본 연구의 핵심 질문은 다음과 같다:

> **"AI가 생성한 이미지와 인간이 창작한 이미지의 창의성을 정량적으로 구분할 수 있는가?"**

이를 위해 본 연구는 다음과 같은 구체적 목표를 설정한다:

(i) 확산 모델의 역변환(Inversion) 특성을 활용한 창의성 측정 프레임워크 개발

(ii) 다중 지표(복합 점수)를 통한 강건한 창의성 평가 체계 구축

(iii) AI 생성 이미지와 인간 작품 간의 통계적 차이 분석

(iv) 화가별, 시대별 세부 분석을 통한 패턴 발견

1.3 이론적 배경: 철학적 관점

1.3.1 중국어 방 논증과 시스템 논변

존 설(John Searle)의 중국어 방 논증은 AI가 진정한 이해 없이도 지능적인 행동을 모방할 수 있다고 주장한다. 그러나 시스템 논변은 개별 구성 요소가 이해하지 못하더라도 전체 시스템이 창발적 이해를 보일 수 있다고 반박한다.

본 연구는 실용주의적 관점에서, AI가 내부 원리를 "이해"하는지 여부와 관계없이, **결과적으로 인간과 구별할 수 없는 수준의 창의적 성과물을 생성한다면 창의적이라고 간주**할 수 있다는 입장을 채택한다.

1.3.2 창의성의 조작적 정의

본 연구에서 창의성은 다음과 같이 조작적으로 정의된다:

> **정의 (창의성):** 창의성은 주어진 데이터 분포(학습 분포)로부터의 **통계적 일탈**의 정도로 측정된다. 학습된 분포와 일치할수록 창의성이 낮고, 분포 밖에 가까울수록 창의적이다.

이 정의에 따르면, 확산 모델이 "예측하기 어려운" 이미지일수록 더 창의적인 것으로 간주된다.

2 관련 연구

2.1 확산 모델

잡음 제거 확산 확률 모델(Denoising Diffusion Probabilistic Models, DDPMs)은 데이터에 점진적으로 노이즈를 추가한 후, 이를 역으로 제거하는 과정을 학습하는 생성 모델이다 [2]. DDIM(Denoising Diffusion Implicit Models)은 결정론적 샘플링을 통해 더 빠른 생성과 정확한 역변환을 가능하게 한다 [3].

2.2 Stable Diffusion

잠재 확산 모델(Latent Diffusion Models, LDMs)은 픽셀 공간이 아닌 잠재 공간에서 확산 과정을 수행하여 계산 효율성을 크게 향상시킨다 [1]. Stable Diffusion v1.5는 LAION-5B 데이터셋으로 학습되었으며, 고품질 이미지 생성 능력으로 널리 사용되고 있다.

2.3 DDIM Inversion

DDIM Inversion은 주어진 이미지를 확산 모델의 잠재 공간으로 매핑하는 기법이다. 이 과정은 생성 과정의 역변환으로, 원본 이미지를 재구성할 수 있는 노이즈 벡터를 추출한다. 이 기법은 이미지 편집, 스타일 전이 등 다양한 응용에 활용된다.

2.4 창의성의 계산적 측정

기존 연구에서 창의성의 계산적 측정은 주로 다음 접근법을 사용한다:

- **새로움 기반**: 기존 샘플들과의 거리 측정 (예: FID, IS)
- **놀라움 기반**: 예측 모델의 손실 함수 활용
- **인간 평가**: 전문가 평가 또는 크라우드소싱

본 연구는 확산 모델 자체의 재구성 능력을 새로움의 대리 지표로 활용하는 새로운 접근법을 제안한다.

3 연구 방법

3.1 핵심 아이디어

본 연구의 핵심 가설은 다음과 같다:

> **가설:** 확산 모델이 학습한 분포 내에 있는 이미지는 역변환-재구성 과정에서 정확하게 복원되며, 분포 밖에 있는 이미지는 복원 오차가 크다.

이 가설에 기반하여, **재구성 오차**를 창의성의 지표로 사용한다.

3.2 DDIM Inversion 및 재구성

3.2.1 순방향 과정 (역변환)

주어진 원본 이미지 x_0를 DDIM Inversion을 통해 잠재 노이즈 x_T로 변환한다:

$$x_{t+1} = \sqrt{\alpha_{t+1}} \cdot f_\theta(x_t, t) + \sqrt{1 - \alpha_{t+1}} \cdot \epsilon_\theta(x_t, t) \tag{1}$$

여기서 ϵ_θ는 학습된 노이즈 예측 네트워크이다.

3.2.2 역방향 과정 (재구성)

추출된 잠재 노이즈 x_T를 동일한 모델로 역변환하여 재구성 이미지 x'_0를 얻는다:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \cdot \hat{x}_0^{(t)} + \sqrt{1 - \alpha_{t-1}} \cdot \epsilon_\theta(x_t, t) \tag{2}$$

3.2.3 오차 측정

원본 x_0와 재구성된 x'_0 사이의 차이를 다중 지표로 측정한다.

3.3 복합 창의성 점수

단일 지표의 한계를 극복하기 위해, 세 가지 상호보완적인 지표를 결합한 복합 점수를 제안한다:

$$\text{복합 점수} = \frac{1}{3}\left(S_{\cos} + S_{\text{LPIPS}} + S_{\text{MSE}}\right) \tag{3}$$

3.3.1 DINOv2 코사인 유사도

DINOv2 [4]는 자기지도 방식으로 학습된 비전 트랜스포머로, 의미론적 특징을 강력하게 포착한다.

$$S_{\cos} = 1 - \frac{\mathbf{f}(x_0) \cdot \mathbf{f}(x_0')}{\|\mathbf{f}(x_0)\| \cdot \|\mathbf{f}(x_0')\|} \tag{4}$$

여기서 $\mathbf{f}(\cdot)$는 DINOv2 임베딩 함수이다. 이 지표는 **고수준 의미적 변화**를 측정한다.

3.3.2 LPIPS (학습 기반 지각적 유사도)

LPIPS [5]는 인간의 지각적 유사도 판단과 높은 상관관계를 보이는 학습 기반 지표이다.

$$S_{\text{LPIPS}} = \text{LPIPS}(x_0, x_0') \tag{5}$$

AlexNet 백본을 사용하며, **구조적·시각적 변화**를 측정한다.

3.3.3 로그 정규화 MSE

픽셀 수준의 차이를 측정하되, 값의 범위를 조정하기 위해 로그 정규화를 적용한다:

$$S_{\text{MSE}} = \frac{\log_{10}(1 + \text{MSE}(x_0, x_0'))}{5.0} \tag{6}$$

여기서 분모 5.0은 최대 MSE($255^2 \approx 65{,}000$)의 로그값에 근접한 정규화 상수이다. 이 지표는 저수준 픽셀 변화를 측정한다.

3.4 실험 파이프라인

Algorithm 1 창의성 점수 계산 파이프라인

Require: 이미지 x_0, 확산 모델 $\mathcal{M}$, 단계 수 T
Ensure: 복합 창의성 점수 S

$z_0 \leftarrow$ 인코딩(x_0) {VAE 인코딩으로 잠재 공간 변환}
$z_T \leftarrow$ DDIM역변환(z_0, T) {가이던스=0으로 역변환}
$z_0' \leftarrow$ DDIM샘플링(z_T, T) {재구성}
$x_0' \leftarrow$ 디코딩(z_0') {VAE 디코딩}
$S_{\cos} \leftarrow 1 -$ 코사인유사도(DINOv2(x_0), DINOv2(x_0'))
$S_{\text{LPIPS}} \leftarrow \text{LPIPS}(x_0, x_0')$
$S_{\text{MSE}} \leftarrow$ 로그정규화MSE(x_0, x_0')
$S \leftarrow (S_{\cos} + S_{\text{LPIPS}} + S_{\text{MSE}})/3$
return S

4 실험 설정

4.1 데이터셋

실험에는 Hugging Face에 공개된 예술 이미지 데이터셋을 활용하였다.

4.1.1 WikiArt 데이터셋 (인간)

인간 작가의 작품인 대조군 데이터셋으로는 `Dant33/WikiArt-81K-BLIP_2-768x768`을 사용하였다. 이는 르네상스부터 현대 미술까지 다양한 시대와 스타일의 작품 약 81,000점을 포함하고 있으며, 본 연구에서는 이 중 3,000점을 무작위로 추출하여 사용하였다.

4.1.2 AI 생성 이미지 데이터셋

AI 생성 이미지인 실험군 데이터셋으로는 `Dant33/Wikiart_with_StableDiffusion`을 활용하였다. 해당 데이터셋은 Stable Diffusion v1.5 모델을 사용하여 WikiArt의 각 작품에 대응하도록 생성된 이미지들로 구성되어 있다. 프롬프트는 원본 작품의 스타일, 화가, 주제를 반영하여 설계되었으며, 인간 작품과 1:1로 매핑되는 쌍(Pair) 구성을 가진다.

4.1.3 데이터 구성

Table 1: 데이터셋 구성

레이블	샘플 수	출처 (Hugging Face Repository)
인간	3,000	`Dant33/WikiArt-81K-BLIP_2-768x768`
AI	3,000	`Dant33/Wikiart_with_StableDiffusion`
합계	**6,000**	

4.2 하드웨어 및 소프트웨어

Table 2: 실험 환경

구성요소	사양
GPU	AMD Radeon RX 6800 (16GB VRAM)
프레임워크	PyTorch 2.x
확산 모델	Stable Diffusion v1.5 (runwayml)
특징 추출기	DINOv2-Base (facebook)
지각적 지표	LPIPS (AlexNet 백본)
역변환 단계	20단계 (DDIM)

4.3 전처리

모든 이미지는 다음과 같은 전처리를 거친다:

1. **중앙 크롭**: 비율 왜곡 방지를 위해 중앙 기준 정방형 크롭
2. **크기 조정**: 512×512 해상도로 조정
3. **정규화**: $[-1, 1]$ 범위로 정규화

5 실험 결과

5.1 주요 결과: 가설 검증

5.1.1 초기 가설

AI 생성 이미지는 모델의 학습 분포 내에 존재하므로 복원율이 높을 것(낮은 오차)이며, 인간의 창의적 이미지는 분포 밖에 가까워 복원율이 낮을 것(높은 오차)이다.

5.1.2 실험 결과: 가설 기각

실험 결과는 초기 가설과 **정반대**로 나타났다.

Table 3: 주요 결과: 복합 창의성 점수 비교

레이블	샘플 수	평균 점수	표준편차	p-값
AI	3,000	**0.684**	0.159	$< 10^{-47}$
인간	3,000	0.628	0.131	

이 결과는 초기 가설과 반대로, AI 이미지의 복원 오차가 인간 작품보다 크다는 것을 보여준다.

5.2 세부 지표별 분석

Table 4: 개별 지표 비교

지표	AI (평균)	인간 (평균)	해석
1 - 유사도 (DINOv2)	0.817	0.780	AI가 의미적으로 더 많이 변질됨
LPIPS	0.586	0.499	AI가 구조적으로 더 많이 변질됨
정규화 MSE	0.648	0.606	AI가 픽셀 단위로 더 많이 깨짐
복합 점수	**0.684**	**0.628**	**AI가 전반적으로 창의성 점수가 높음**

세 지표 모두 일관되게 AI > 인간 패턴을 보여 결과의 일관성을 확인할 수 있다.

5.3 시대별 분석

Table 5: 예술 시대별 분석

시대	기간	AI 평균	인간 평균	차이	유의미
바로크/로코코	1600-1800	0.704	0.616	+0.088	✓
르네상스	1400-1600	0.712	0.671	+0.041	✓
낭만주의	1800-1870	0.679	0.628	+0.051	✓
인상주의	1870-1910	0.691	0.630	+0.061	✓
모던	1910-1950	0.704	0.639	+0.065	✓
현대	1950+	0.671	0.622	+0.049	✓

발견: 모든 시대에서 AI > 인간 패턴이 관찰되었으며, 특히 바로크/로코코 시대에서 가장 큰 차이 (+0.088)가 나타났다.

5.4 화가별 분석

Table 6: 주요 화가별 분석 (일부)

순위	화가	AI 평균	인간 평균	차이	p-값
1	이반 아이바조프스키	0.672	0.582	+0.090	0.006 ✓
2	파블로 피카소	0.716	0.626	+0.090	0.037 ✓
3	피에르 오귀스트 르누아르	0.692	0.606	+0.086	0.005 ✓
4	존 싱어 사전트	0.737	0.653	+0.084	0.012 ✓
7	빈센트 반 고흐	0.684	0.615	+0.069	0.005 ✓
8	클로드 모네	0.659	0.614	+0.045	0.042 ✓
예외 케이스 (인간 > *AI*):					
14	카미유 피사로	0.605	0.641	-0.036	0.294
15	귀스타브 도레	0.648	0.689	-0.041	0.094

예외 발견: 카미유 피사로와 귀스타브 도레의 경우 인간 > AI 패턴이 관찰되었다. 이는 해당 화가들의 작품이 Stable Diffusion의 학습 데이터에 상대적으로 덜 포함되어 있을 가능성을 시사한다.

5.5 연대별 추이

Table 7: 연대별 추이 분석

연대	AI 평균	인간 평균	차이	추세
1870년대	0.696	0.616	+0.080	AI ↑
1890년대	0.704	0.642	+0.062	AI ↑
1920년대	**0.728**	**0.634**	**+0.094**	AI ↑↑
1940년대	0.676	0.588	+0.088	AI ↑
1960년대	0.619	0.560	+0.059	AI ↑
1980년대	0.673	0.591	+0.082	AI ↑
2000년대	0.549	0.602	**-0.053**	**인간↑**
2010년대	0.547	0.484	+0.063	AI ↑

주목할 점: 2000년대는 유일하게 인간 > AI 패턴을 보인 시기이다. 이는 Stable Diffusion의 학습 컷오프(2022년) 직전 시기의 현대 미술이 상대적으로 학습 데이터에 덜 포함되었을 가능성을 시사한다.

6 논의

6.1 결과 해석

초기 가설과 반대 결과가 나온 원인을 세 가지로 분석할 수 있다.

6.1.1 WikiArt의 학습 분포 포함

Stable Diffusion 1.5는 LAION-Aesthetics 데이터셋으로 학습되었으며, 이 데이터셋은 인터넷에서 수집된 약 20억 개의 이미지-텍스트 쌍을 포함한다 [6]. WikiArt의 유명 회화들은 웹에 광범위하게 존재하므로, 학습 데이터에 상당수 포함되어 있을 가능성이 높다.

따라서 WikiArt(인간) 이미지는 역설적으로 확산 모델의 **분포 내**에 해당하며, 이로 인해 복원이 더 잘 된다.

6.1.2 AI 생성 이미지의 고주파 아티팩트

AI가 생성한 이미지에는 사람 눈에 잘 보이지 않는 고주파 패턴이나 미세한 아티팩트가 포함될 수 있다. DDIM 역변환 과정에서 이러한 특성이 손실되어, 결과적으로 재구성 오차가 증가한다.

6.1.3 모델 자기 일관성의 한계

동일한 모델(SD1.5)로 생성한 이미지라 하더라도, 역변환-재구성 과정에서 완벽한 복원이 보장되지 않는다. 특히 생성 시 사용된 가이던스 스케일과 역변환 시 설정(가이던스=0)의 차이로 인해 불일치가 발생할 수 있다.

6.2 창의성 정의의 재해석

본 연구 결과는 "재구성 오차 = 창의성"이라는 조작적 정의의 한계를 드러낸다. 높은 재구성 오차가 반드시 "창의적"임을 의미하지 않으며, 오히려 다음을 반영할 수 있다:

- 모델 학습 분포와의 불일치
- 이미지 내 고주파 노이즈나 아티팩트
- 세부 텍스처의 복잡성

6.3 지표 일관성

세 지표(DINOv2, LPIPS, MSE)가 동일한 방향의 결과를 보였다. 이는 결과가 특정 지표의 편향이 아닌 데이터 자체의 특성을 반영함을 뜻한다.

6.4 예외 케이스의 함의

일부 화가(카미유 피사로, 귀스타브 도레)와 특정 시기(2000년대)에서 인간 > AI 패턴이 관찰된 것은, 학습 데이터 분포의 비균일성을 시사한다. 이는 향후 연구에서 **학습 데이터 포함 여부**를 통제 변인으로 고려해야 함을 제안한다.

7 한계점

7.1 방법론적 한계

1. **단일 모델 의존성**: 본 연구는 Stable Diffusion 1.5만을 사용하였다. 다른 버전(SD 2.1, SDXL) 또는 다른 계열의 모델(DALL-E, Midjourney)에서 동일한 결과가 재현될지 검증이 필요하다.
2. **데이터셋 편향**: WikiArt 데이터셋이 학습 분포에 포함되어 있다는 점이 결과에 큰 영향을 미쳤다. 학습 데이터에 포함되지 않은 이미지로 추가 검증이 필요하다.
3. **역변환 충실도**: DDIM 역변환은 결정론적이지만 완벽하지 않다. Null-text Inversion 등 더 정밀한 기법의 적용이 필요할 수 있다.

7.2 개념적 한계

1. **창의성 정의**: "재구성 오차"를 창의성의 대리 지표로 사용하는 것은 직관적이지만, 인간의 창의성 판단과의 상관관계 검증이 부재하다.
2. **주관적 검증 부재**: 인간 평가자를 통한 창의성 평가와의 비교가 이루어지지 않았다.

8 향후 연구

8.1 다중 모델 검증

- Stable Diffusion 2.1, SDXL 등 다른 버전의 모델로 동일 실험 수행
- WikiArt로 미세조정된 모델 사용 시 결과 변화 분석
- DALL-E, Midjourney 등 타 계열 모델과의 비교

8.2 다양한 데이터셋 활용

- GenImage, DiffusionDB 등 다양한 AI/인간 이미지 쌍 데이터셋
- 학습 데이터에 확실히 포함되지 않은 최신 작품
- 비예술 이미지(사진, 스케치 등)로 확장

8.3 창의성 정의 정교화

- 인간 평가와의 상관관계 연구
- 다차원 창의성 지표 개발 (새로움, 놀라움, 가치 등)
- 전문가 평가와 자동화 지표 간 관계 분석

8.4 역변환 기법 개선

- Null-text Inversion, EDICT 등 더 정밀한 역변환 기법 적용
- 다단계 가이던스 전략 탐색

9 결론

본 연구는 DDIM Inversion 기반 재구성 오차를 창의성 측정 지표로 활용하는 방법을 제안하고, 6,000개 이미지로 검증하였다. 주요 발견은 다음과 같다:

1. AI 이미지가 인간 작품보다 높은 복원 오차를 보였다. WikiArt 회화가 Stable Diffusion 학습 데이터에 포함되어 있어, 오히려 인간 작품이 모델에 익숙하기 때문이다.
2. 세 지표(DINOv2, LPIPS, MSE)가 동일한 방향의 결과를 보여 일관성을 확인하였다.
3. 대부분의 시대·화가에서 AI > 인간 패턴이 관찰되었으나, 일부 예외가 존재하여 학습 분포의 불균일성을 시사한다.

본 연구는 "창의성의 정량적 측정"이라는 도전적 과제에 대한 실증적 접근을 시도하였으며, AI 창의성 연구의 새로운 방향을 제시한다. 향후 다양한 모델과 데이터셋을 통한 검증, 그리고 인간 평가와의 상관관계 연구를 통해 더욱 강건한 창의성 측정 프레임워크를 발전시킬 수 있을 것이다.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10674–10685.

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.

[3] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

[4] Oquab, M., Darcet, T., Moutakanni, T., et al. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[5] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

[6] Schuhmann, C., Beaumont, R., Vencu, R., et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35, 25278–25294.