

2주차 스터디

Zero-Shot Text-to-Image Generation

2.6 샘플 생성



Razavi et al. (2019)과 유사하게, 사전 훈련된 대조 모델을 사용하여 변압기에서 추출한 샘플의 순위를 다시 매깁니다(Radford et al., 2021). 캡션과 캔디 데이트 이미지가 주어지면 대조 모델은 다음을 기반으로 점수를 할당합니다.

(MS-COCO의 캡션에 대한 이전 작업(DF-GAN)과 모델에 대한 인체 평가(온도 감소 없이 평가된 제로샷) 비교. 5전 2선승제 투표에서 모델의 샘플은 90.0%의 확률로 가장 사실적인 것으로 선택되었으며, 93.3%의 경우 공유 캡션과 가장 일치하는 이미지로 선정되었습니다. 이미지가 캡션과 얼마나 잘 일치하는지. 그림 6은 상위 k 이미지를 선택하는 샘플 N의 수를 늘리는 효과를 보여줍니다.)가 나오는 이유와 원리를 설명해줘

이 프로세스는 일종의 언어 안내 검색으로 볼 수 있으며(Andreas et al., 2017), Xu et al. (2018)이 제안한 보조 텍스트-이미지 일치 손실과도 유사합니다. 달리 명시되지 않는 한, 정 성적 및 정량적 결과에 사용된 모든 샘플은 온도 감소(즉, $t = 1$ 사용) 없이 얻어지며(그림 2 제외) $N = 512$ 로 재순위를 사용합니다.

💡 핵심 문장 요약

"5전 2선승제 투표에서 DALL·E는 DF-GAN보다 90.0% 더 사실적으로, 93.3% 더 텍스트와 잘 맞는 이미지로 평가되었다."

"그림 6은 더 많은 샘플(N)을 생성한 후 상위 k 개만 선택했을 때 일치도 (성능)가 올라감을 보여준다."

🧠 이게 말하는 원리 핵심

1. ✅ Human Evaluation (인간 평가) 방식

- 5전 2선승제 투표: 평가자가 두 이미지(예: DALL·E vs DF-GAN)를 보고,
 - 둘 중 더 사실적인 이미지 또는
 - 둘 중 텍스트와 더 잘 맞는 이미지를 선택
- 이걸 한 캡션당 5명의 평가자가 보고 최소 2명 이상이 고른 쪽이 '승자'로 결정됨.
- 결과:
 - 90.0% → DALL·E가 더 사실적으로 선택된 비율
 - 93.3% → DALL·E가 캡션과 더 일치한다고 선택된 비율

📌 의미: 단순 수치(FID, IS) 말고도, 사람들이 보기에 진짜 좋다고 느낀 비율도 높다는 뜻!

2. 📊 그림 6의 원리: N 개의 샘플 중 상위 k 개 선택

- N : 하나의 캡션에 대해 생성한 총 이미지 수 (예: 1개, 4개, 16개, 32개 등)
- k : 그 중에서 가장 텍스트와 잘 맞는 '상위 k 개' 이미지만 선택

실험 시나리오:

1. "고양이가 자전거를 타고 있다" 같은 캡션에 대해


→ DALL·E가 32장의 이미지를 생성

2. 그 중 가장 텍스트와 잘 맞는 상위 1~k개를 선택

→ 사람이 보거나, 대조 모델을 이용해 선택

3. 평가 결과:

- N이 커질수록 더 좋은 이미지가 나올 가능성이 높아지고
- 그 중 상위 k개를 선택하면, 텍스트 일치도가 확연히 증가

 **의미:** 모델이 한 번에 완벽한 이미지를 생성하진 않아도,

충분히 많이 생성하면, 그 중에서 "가장 잘 맞는 것"을 골라낼 수 있다!

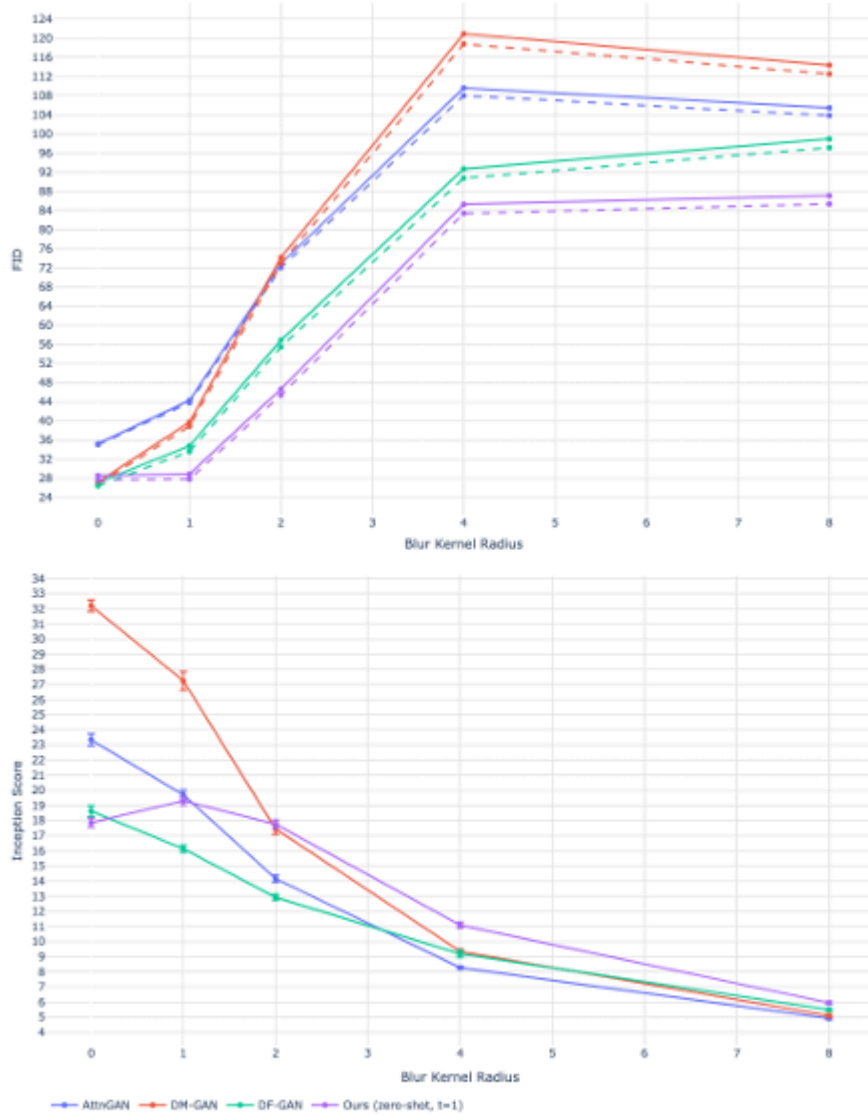
요약하면

요소	의미
5전 2선승제 평가	사람이 보기에 DALL·E가 DF-GAN보다 더 진짜 같고 텍스트에 잘 맞음
샘플 수 N 증가 → 상위 k 선택	이미지를 많이 만들고 고르면 성능이 올라간다 (특히 텍스트-이미지 일치도)
그림 6의 시사점	샘플 다양성 + 재정렬(선택) 조합이 성능을 결정짓는 중요한 전략임

3.1 정량적 평가

AttnGAN(Xu et al., 2018), DM GAN(Zhu et al., 2019), DF-GAN(Tao et al., 2020)의 세 가지 이전 접근 방식과 비교하여 모델을 평가하며, 이 중 마지막 접근 방식은 (MS-COCO에서 최고의 시작 점수(Salimans et al., 2016) 및 Fréchet 시작 거리(Heusel et al., 2017)를 보고합니다. 그림 3은 우리 모델의 samples를 이전 작업의 samples와 정성적으로 비교합니다.)

또한 Koh et al. (2021)에서 사용한 것과 유사한 인간 평가를 수행하여 DF-GAN에 대한 접근 방식을 비교했으며, 그 결과는 그림 7에 나와 있습니다. 캡션이 주어지면 모델의 샘플은 93%의 확률로 캡션과 더 잘 일치하는 과반수 투표를 받습니다. 또한 90%의 시간 동안 더 현실적이라는 이유로 과반수 투표를 받습니다.



(a) FID and IS on MS-COCO as a function of blur radius.

그림 9(a)는 우리 모델이 캡션에 대해 훈련된 적이 없음에도 불구하고 최상의 이전 접근 방식에서 2점 이내의 MS-COCO에서 FID 점수를 얻는다는 것을 보여줍니다. 훈련 데이터에는 그림 8과 같이 필터링된 YFCC100M 하위 집합이 포함되어 있습니다. CUB 데이터셋에 있는 모델의 제로샷 샘플. 그리고 다음 섹션에 설명된 중복 제거 절차에서 MS-COCO 검증 세트에 있는 이미지의 약 21%가 포함되어 있음을 발견했습니다. (이 효과를 분리하기 위해 이러한 이미지(실선)가 있는 이미지와 없는 이미지(파선) 모두에 대한 검증 세트에 대한 FID 통계량을 계산하며, 결과에는 큰 변화가 없습니다.)

dVAE en coder의 토큰에 대해 트랜스포머를 훈련하면 이미지를 시각적으로 인식할 수 있도록 하는 저주파 정보에 모델링 용량을 할당할 수 있습니다. 그러나 무거운 압축으로 인해 고주파 세부 사항을 생성할 수 없기 때문에 모델에 단점도 있습니다. (정량적 평가에 대한 이

의 효과를 테스트하기 위해 검증 이미지와 모델의 샘플 모두에 다양한 반경을 가진 가우스 필터를 적용한 후 그림 9(a)에서 FID 및 IS를 계산합니다.) 우리의 접근 방식은 반경 1의 약간의 흐림으로 약 6포인트의 여백으로 최상의 FID를 달성합니다. 우리의 접근 방식과 다른 접근 방식 사이의 격차는 흐림 반경이 증가함에 따라 넓어지는 경향이 있습니다. 또한 흐림 반경이 2보다 크거나 같을 때 가장 높은 IS를 얻습니다.

우리의 모델은 CUB 데이터 세트에서 훨씬 더 나빠졌으며, 이 경우 우리 모델과 선행 접근 방식 사이에 FID에서 거의 40포인트의 차이가 있습니다(그림 9(b)). 이 데이터 세트에 대해 12%의 중복률을 발견했으며, 이러한 이미지를 제거한 후에도 결과에 큰 차이가 없었습니다. 우리는 우리의 제로샷 접근 방식이 CUB와 같은 특수 분포에서 유리하게 비교될 가능성이 적다고 추측합니다. 우리는 미세 조정이 개선을 위한 유망한 방향이라고 믿으며 이 조사는 향후 작업에 맡깁니다. 이 데이터 세트의 캡션에 대한 모델의 샘플은 그림 8에 나와 있습니다.

마지막으로, (그림 9(c)는 대조 모델로 순위를 다시 매기는 데 사용되는 표본 크기가 증가함에 따라 MS-COCO에 대한 FID 및 IS의 명확한 개선을 보여줍니다. 이러한 경향은 표본 크기가 32까지 계속되며, 그 이후에는 흐림 반경의 함수로 MS-COCO에서 감소하는 것을 관찰할 수 있습니다.) 그림 9. MS-COCO 및 CUB에 대한 정량적 결과. 실선은 원래 검증 세트에 대해 계산된 FID를 나타내고, 파선은 겹치는 이미지가 제거된 검증 세트에 대해 계산된 FID를 나타냅니다(섹션 3.2 참조). MS-COCO의 경우 검증 세트에서 샘플링된 30000개의 캡션 하위 집합에서 모든 모델을 평가합니다. CUB의 경우 테스트 세트의 모든 고유 캡션에서 모든 모델을 평가합니다. <https://github.com/MinfengZhu/DM-GAN> 에서 사용할 수 있는 DM-GANcode를 사용하여 FID와 IS를 계산합니다.

✓ 먼저 용어 정리부터!

- **FID (Fréchet Inception Distance):** 생성된 이미지 분포와 실제 이미지 분포의 "거리"를 측정해, 이미지 품질을 수치로 나타냄. 낮을수록 좋음.
- **IS (Inception Score):** 앞서 설명했듯, 품질과 다양성을 동시에 측정. 높을수록 좋음.
- **MS-COCO:** 널리 쓰이는 이미지 캡셔닝 데이터셋. 논문에서는 이걸 기준으로 테스트합니다.
- **가우시안 블러(Gaussian Blur):** 이미지를 흐리게 만드는 필터. 이미지 품질을 인위적으로 떨어뜨릴 때 사용함.
- **샘플 수:** 생성된 이미지 중 "선택해서 평가에 사용하는 수". 예를 들어 32개 샘플만 골라 평가할 수도 있고, 128개로 늘릴 수도 있음.

💡 논문에서 말하는 실험의 핵심 원리

1. MS-COCO 검증 세트와의 중복 문제

- 생성된 이미지 중 **21%가 MS-COCO 검증 세트와 중복되어** 있다는 걸 발견했어요.
 - 그래서 평가에 편향이 생길 수 있어, **중복된 이미지와 중복되지 않은 이미지로 나누어 FID를 계산**해봤어요.
→ 결과적으로
큰 차이가 없었고, 평가 신뢰성엔 큰 영향 없음.
-

2. 가우스 필터 실험 – 흐림이 평가에 미치는 영향 (그림 9a)

- 생성 이미지와 진짜 이미지에 '흐림'을 인위적으로 적용해봤어요.
 - 이유는? FID나 IS가 진짜로 "품질 저하"를 잘 감지하는지 보기 위해서예요.
 - 결과: 흐릴수록 FID는 증가하고, IS는 감소함.
→ 즉,
이 두 지표가 이미지 품질 저하를 잘 반영한다는 걸 확인함.
-

3. 샘플 수 증가 실험 – 샘플 선택이 성능에 미치는 영향 (그림 9c)

- 생성된 이미지 중에서 **몇 개를 선택해서 FID/IS 평가에 쓸지**가 결과에 영향을 미치는지 실험.
 - 샘플 수가 **8 → 16 → 32개로 늘어날수록**, 성능(FID ↓, IS ↑)이 확실히 개선됨.
 - 이유: 더 많은 이미지 중 "**좋은 것만 선택해서**" 평가하기 때문이에요. (일종의 재정렬 / reranking)
-

🔄 결론 정리

- FID/IS는 이미지 품질 변화를 잘 잡아내는 유효한 평가 지표임을 확인.
- 중복 이미지는 일부 있지만, 평가에 큰 영향 없음.
- 샘플을 많이 뽑고 정렬하면 FID/IS가 개선되는 경향이 있다 → 즉, **선택적으로 좋은 샘플을 뽑는 것도 성능 향상 방법**이 될 수 있다는 걸 보여줌.