CUAI 프로젝트 NLP 2팀

2024.11.26

발표자: 박지후

팀원 소개

소프트웨어학과 **권하연**

미디어커뮤니케이션학부 **김지호**

응용통계학과 **박지후**

소프트웨어학과 **오재환**

목차

1. 프로젝트 개요

[주제]

AI 기술이 웹툰 산업에 미치는 영향과 상생 방안 모색 : 네이버 웹툰 댓글 분석을 중심으로

[방법]

- 데이터 수집 : 1회차부터 70회차까지 댓글 19,303개
  - 대댓글 미포함
- 구간 분리 : AI 관련 댓글 빈도에 따른 세 구간 설정
- 분석 기법 : 감성 분석, 연관어 분석, LDA 토픽 모델링

[결론]

- 후반부로 갈수록 AI 이용 자체에 대한 평가는 긍정적
- 그러나 **여전히 저작권에 대한 부정적 시각**은 후반부까지 이어짐

이미지 생성 AI로부터 저작권을 보호하려는 연구 탐색

2. 기술적 관점에서의 저작권 보호

Survey Paper

Abstract

1 Introduction

2 Copyright in Image Generation

3 Copyright in Text Generation

4 Other domains

5 Discussion

References

arXiv:2402.02333v2 [cs.CR] 24 Jul 2024

Copyright Protection in Generative AI: A Technical Perspective

Jie Ren
renjie3@msu.edu
Michigan State University USA

Han Xu
Michigan State University USA
xuhan1@msu.edu

Pengfei He
Michigan State University USA
hepengf1@msu.edu

Yingqian Cui
Michigan State University USA
cuiyingq@msu.edu

Shenglai Zeng
Michigan State University USA
zengshe1@msu.edu

Jiankun Zhang
School of Artificial Intelligence, Jilin University China
International Center of Future Science, Jilin University China
Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE China
zhangjk9920@mails.jlu.edu.cn

2. 기술적 관점에서의 저작권 보호

모델에서의 저작권 보호방법론

	추론단계	학습단계
GAN	- 올바른 정보 추출 방해 - 생성된이미지에 왜곡을 최대화하여 원본 보호	- FGSM를 활용해 적대적 예제를 생성하여 교란. 학습 과정에서 균형을 깨뜨림
Diffusion	- 적대적 예제를 찾아 분포에서 벗어난 샘플로 생성 - 원본과 다른 잠재표현을 가지고 편집과정 방해	- 예술작품 학습시 스타일을 제거하고 학습 - 범용적 교란 - 학습성능 저하로 고품질 이미지 생성방해

=> 그러나 일반화 X, 변화에 취약

2. 기술적 관점에서의 저작권 보호

워터마크를 이용한 저작권 보호방법론

*워터마크 : 저작권이 있는 데이터를 보호하기 위해 특정 "식별 가능한 정보"를 삽입하는 기술

텍스트-이미지 생성	- 캡션에 대응하는 트리거를 삽입 - 워터마크를 이미지의 스타일이나 객체보다 우선적으로 학습하여 구분하도록함
DiffusionShield	- 픽셀값을 최적화하여 모델학습 후에도 워터마크가 유지되도록함
DGM 워터마킹 개념	- **특정 정보(워터마크)**를 모델이나 생성된 콘텐츠에 삽입하는 기술 - 파라미터나 생성된 이미지 등에 워터마크를 삽입하여 생성된 이미지에 모델의 정보를 담도록 함 - 이미지만 보고도 사용모델과 원본을 추적할 수 있음

=> 역시 일반화 X, 변화에 취약

Unlearning과 De-duplication

Machine Unlearning	- 이미 학습한 데이터 중 특정 부분(예: 저작권이 있는 데이터)을 제거하여, 모델의 출력이 마치 해당 데이터를 학습하지 않은 것처럼 만드는 기술 - GAN : 판별기(Discriminator)가 저작권 데이터를 "거짓(Fake)" 샘플로 인식하도록 수정 - Diffusion : 노이즈 또는 무해한 조건으로 전환
Dataset De-duplication	- 데이터의 중복으로 인해 모델이 데이터를 암기하고 이를 그대로 생성하는 문제를 해결하기 위한 기술 - 중복이미지 탐색 및 제거 - 다양한 캡션 생성 - 에포크 수 감소

3. 구체적인 방법 소개

개념지우기

- 모델

Erasing concepts from Diffusion Model

Rohit Gandikota*1, Joanna Materzyńska*2, Jaden Fiotto-Kaufman1, David Bau1
1Northeastern University, 2MIT CSAIL; *Equal contribution

Stable Diffusion 모델을 통한 방대한 데이터 세트 모방 훈련 ⇨ 딥페이크 및 저작권 침해 문제 발생 ⇨ erasing model을 통한 Concept 제거

3. 구체적인 방법 소개

개념지우기

- 모델 **text-to-image diffusion model**

$$\epsilon_\theta(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) - \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)]$$

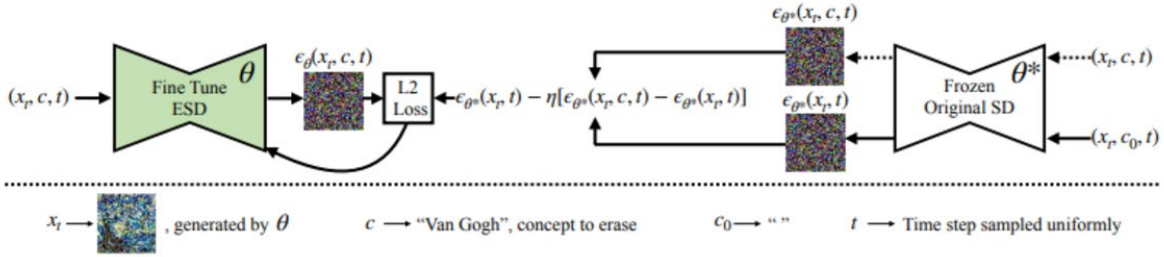


Figure 2: The optimization process for erasing undesired visual concepts from pre-trained diffusion model weights involves using a short text description of the concept as guidance. The ESD model is fine-tuned with the conditioned and unconditioned scores obtained from frozen SD model to guide the output away from the concept being erased. The model learns from its own knowledge to steer the diffusion process away from the undesired concept.

3. 구체적인 방법 소개

개념지우기

- 사용 데이터

‘반고흐’ 프롬프트 data

	case_number	prompt	evaluation_seed	artist
0	0	A Wheatfield, with Cypresses by Vincent van Gogh	2219	Vincent van Gogh
1	1	Almond Blossoms by Vincent van Gogh	4965	Vincent van Gogh
2	2	Bedroom in Arles by Vincent van Gogh	2795	Vincent van Gogh
3	3	Bridge at Trinquetaille by Vincent van Gogh	3370	Vincent van Gogh
4	4	Café Terrace at Night by Vincent van Gogh	2776	Vincent van Gogh
5	5	Cypresses by Vincent van Gogh	2410	Vincent van Gogh
6	6	Enclosed Field with Rising Sun by Vincent van Gogh	2768	Vincent van Gogh
7	7	Entrance to a Quarry by Vincent van Gogh	4274	Vincent van Gogh
8	8	Fishing Boats on the Beach at Saintes-Maries by Vincent van Gogh	3485	Vincent van Gogh

esd_diffusers 모델 학습

```
!python esd_diffusers.py --erase_concept "Van Gogh" --train_method "xattn"
```

3. 구체적인 방법 소개

개념지우기

- 결과

생성형 이미지 생성- **Original model**

Original model

```
[ ]  seed = 1234
     images = diffuser("image of a cat in the style of Van Gogh",
              img_size=512,
              n_steps=50,
              n_imgs=1,
              generator=torch.Generator().manual_seed(seed),
              guidance_scale=7.5
              )[0][0]
     images
```

image of a cat in the style of Van Gogh

3. 구체적인 방법 소개

개념지우기

- 결과

생성형 이미지 생성- **Erasing model**

erasing model

```
[ ]  with finetuner:
         images = diffuser("image of a cat in the style of Van Gogh",
                  img_size=512,
                  n_steps=50,
                  n_imgs=1,
                  generator=torch.Generator().manual_seed(seed),
                  guidance_scale=7.5
                  )[0][0]
     images
```

image of a cat in the style of Van Gogh

3. 구체적인 방법 소개

개념지우기

- 결과

	image of sunflower in the style of Van Gogh	image of scream boy in the style of Van Gogh	image of room in the style of Van Gogh
Original model			
Erasing model			

3. 구체적인 방법 소개

개념지우기

'Marvel Comics' 프롬프트 생성 후, 모델 학습 추가 진행

- 결과

	image of a cat in the style of Marvel	image of a dog in the style of Marvel	image of New York in the style of Marvel
Original model			
Erasing model			

3. 구체적인 방법 소개

reward이용

- reward이용

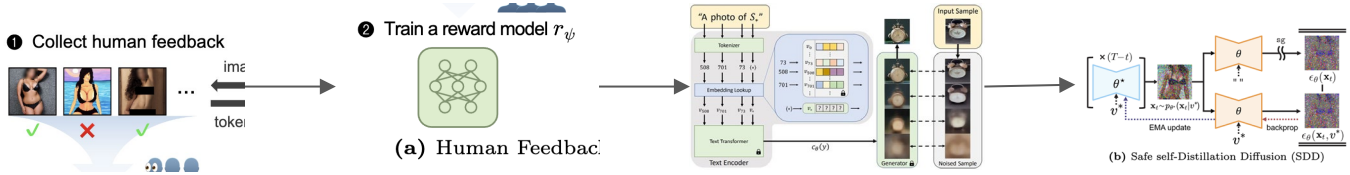
v1 [cs.CV] 17 Jul 2024

Safeguard Text-to-Image Diffusion Models with Human Feedback Inversion

Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee

Korea Advanced Institute of Science and Technology (KAIST)
{nannullna,heon2203,balhaekim,ms.choi,jinwoos,juholee}@kaist.ac.kr

Abstract. This paper addresses the societal concerns arising from large-scale text-to-image diffusion models for generating potentially harmful or copyrighted content. Existing models rely heavily on internet-crawled data, wherein problematic concepts persist due to incomplete filtration processes. While previous approaches somewhat alleviate the issue, they often rely on text-specified concepts, introducing challenges in accurately capturing nuanced concepts and aligning model knowledge with human understandings. In response, we propose a framework named Human Feedback Inversion (HFI), where human feedback on model-generated images is condensed into textual tokens guiding the mitigation or removal of problematic images. The proposed framework can be built upon existing techniques for the same purpose, enhancing their alignment with human judgment. By doing so, we simplify the training objective with a self-distillation-based technique, providing a strong baseline for concept



4. 국내 웹툰 산업에서의 AI 활용 동향

네이버의 웹툰 AI 연구 진행방향

- 지난 AI 웹툰 활용 논란 이후
  저작권 문제를 해소하기 위해
  작가 별 맞춤형 AI 생성 툴 개발 착수

- 작가 별 작가 자신의 그림 데이터
  +저작권 문제가 없는 그림 데이터를
  활용한 생성 툴 연구 진행중

네이버 웹툰 창작도구 AI 연구 진행상황

■ 웹툰 AI 페인터(Webtoon AI Painter)

딥러닝 기술을 활용해 자연스러운 채색 지원 프로그램 (2021년 10월 베타 출시)

■ 생성형 AI 창작 지원 툴

작품별 맞춤형 학습을 통해 작가별 생성 AI 제작 지원 도구 (현재 연구 중)

■ 웹툰 AI 에디터(Webtoon AI Editor)

웹툰 전용 편집 툴. AI 활용 누끼따기 작업 등 반복작업 대체 (현재 개발 중)

감사합니다