LAION-400M: OPEN DATASET OF CLIP-FILTERED 400 MILLION IMAGE-TEXT PAIRS

ABSTRACT

수억 개의 이미지-텍스트 쌍(예: CLIP, DALL-E)으로 학습된 다중모달 언어-비전 모델은 최근 급격히 주목받았으며, 대상 이미지 데이터에 대한 샘플별 레이블이 없어도 제로샷 또는 몇샷 학습과 전이에서 뛰어난 성능을 보인다. 이러한 경향에도 불구하고, 현재까지 이러한 모델을 처음부터 학습시키기 위한 충분한 규모의 공개 데이터셋은 존재하지 않았다. 이 문제를 해결하기 위해, 커뮤니티의 노력을 통해 우리는 CLIP로 필터링한 4억 이미지-텍스트 쌍, 해당 CLIP 임베딩 및 효율적인 유사도 검색을 가능하게 하는 kNN 인덱스를 포함한 LAION-400M을 구축하여 공개한다.

1 INTRODUCTION

멀티모달 언어-비전 모델은 최근 샘플별 레이블 없이도 새로운 데이터셋으로의 강한 전이 능력을 보였다[1, 2, 3]. 이러한 능력은 사전학습 동안 충분히 큰 모델 및 데이터 규모를 필요로 한다. 데이터 규모만 증가시켜도 종종 모델 성능이 향상될 수 있다[4]. 여기에 모델과 연산 예산 규모를 함께 증가시키면, 데이터 규모에 의해 병목되지 않는 한 스케일링 법칙은 일반화 및 전이 성능의 추가 향상을 시사한다[5, 6, 7, 8]. 다양한 모델을 최적으로 대규모화하기 위해 거대한 데이터셋을 구축한 최근 연구들이 다수 존재한다[9, 1, 2, 3]. 그러나 이러한 거대 데이터셋들은 여러 이유로 거의 공개되지 않았다.

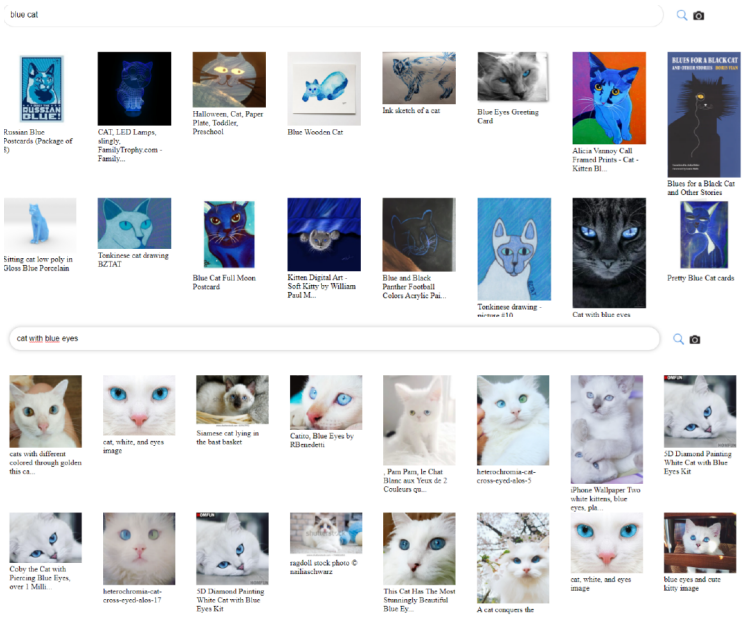

Figure 1: 웹 데모에서 ”blue cat” 또는 ”cat with blue eyes” 쿼리로 검색된 샘플 이미지들이 문제를 해결하기 위해 우리는 CLIP으로 필터링한 4억 개의 이미지-텍스트 페어, 해당 CLIP 임베딩 및 kNN 인덱스를 포함한 LAION-400M을 구축하고 공개한다. 데이터셋 생성 절차를 기술하고

DALL-E 아키텍처의 성공적인 학습을 시연한다. 충분히 큰 규모를 갖춘 이 데이터셋은 멀티모달 언어-비전 모델에 대한 연구를 넓은 커뮤니티에 개방하는 계기를 제공한다.

2 Dataset and Methods

LAION-400M 개요. 우리는 LAION-400M 프로젝트 아래 다음 패키지를 공식 공개한다:

- · 4억 쌍의 이미지 URL과 해당 메타데이터
- · 4억 쌍의 CLIP 이미지 임베딩과 해당 텍스트
- · 데이터셋 내 빠른 검색을 가능하게 하는 여러 세트의 kNN 인덱스
- · 최소한의 자원으로 수백만 장의 이미지와 메타데이터를 URL 목록으로부터 효율적으로 크롤링하고 처리할 수 있는 img2dataset 라이브러리
- · LAION-400M에 대한 이미지-텍스트 검색 웹 데모 (Fig. 1, 2)

이미지 URL과 메타데이터 쌍에 대해서는 각 쌍에 대해 다음 속성들로 구성된 parquet 파일을 제공한다: 샘플 ID, URL, Creative Commons 라이선스 유형(해당되는 경우), NSFW 태그(CLIP으로 감지된), 텍스트와 이미지 임베딩 간 코사인 유사도 점수 및 이미지의 높이와 너비. 사용자들은 NSFW 태그로 필터링할 수 있으며, 감지된 이미지가 1

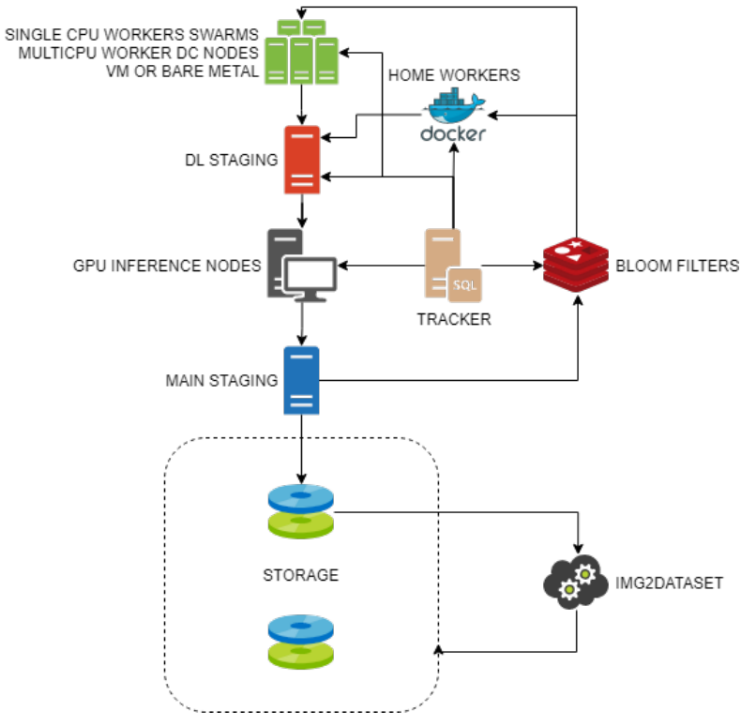

Figure 2: 획득 workflow

획득(Acquisition). 획득은 Fig. 2의 흐름도에 따라 진행되며 두 주요 구성요소로 나눌 수 있다:

- · 일치하는 URL과 캡션 모음을 생성하는 페타바이트 규모 Common Crawl 데이터셋의 분산 처리.
- · 데이터를 단일 노드에서 후처리하는 단계로, 훨씬 가볍고 수일 내에 실행되어 최종 데이터셋을 생성한다.

2.1 Distributed processing of Common Crawl

이미지-텍스트 쌍을 생성하기 위해 Common Crawl의 WAT 파일을 파싱하여 alt-text 속성을 포함하는 모든 HTML IMG 태그를 추출한다. 파싱한 URL로부터 원시 이미지는 Trio 및 Asks 라이브러리를 사용한 비동기 요청으로 다운로드한다.

2.1.1 Filtering out unsuitable image-text pairs

Common Crawl에서 WAT 파일을 다운로드한 후 다음 필터링 조건을 적용했다:
· Alt-text 길이가 5자 미만이거나 이미지 크기가 5 KB 미만인 모든 샘플을 삭제한다.
· URL과 alt-text를 기준으로 bloom filter를 사용해 중복 제거를 수행한다.
· CLIP을 사용해 이미지와 alt-text의 임베딩을 계산한다. 그런 다음 두 임베딩의 코사인 유사도를 계산하고 코사인 유사도가 0.3 미만인 모든 샘플을 삭제한다. 이 임계값은 사람의 검사를 기반으로 선택되었다.
· 이미지와 텍스트의 CLIP 임베딩을 사용해 불법 콘텐츠를 필터링한다.

Table 1: Image size distribution of LAION-400M

Number of unique samples	413M
Number with height or width ≥ 1024	26M
Number with height and width ≥ 1024	9.6M
Number with height and width ≥ 512	67M
Number with height or width ≥ 512	112M
Number with height and width ≥ 256	211M
Number with height or width ≥ 256	268M

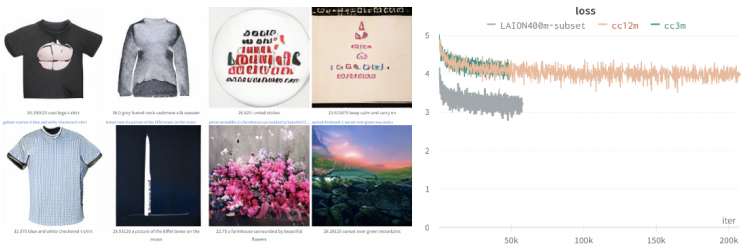

Figure 3: : DALL-E 실험 결과. (왼쪽) DALL-E 모델을 7.2M개의 무작위 LAION-400M 샘플로 RTX 2070 Super (8GB VRAM) 에서 1 epoch 학습했을 때 생성된 예시. (오른쪽) DALL-E 모델을 Conceptual Captions 3M (녹색), Conceptual Captions 12M (주황색), LAION-400M의 3M 하위 집합 (회색) 으로 실행한 결과.

2.1.2 img2dataset

우리는 주어진 URL 집합으로부터 편리하게 이미지를 다운로드하고, 리사이즈하며 캡션을 webdataset 형식으로 저장하기 위해 img2dataset 라이브러리를 개발했다.3 이는 단일 노드(1Gbps 연결 속도, 32GB RAM, 16코어 i7 CPU)로 URL 목록에서 1억 장의 이미지를 20시간 만에 다운로드할 수 있게 하며, 누구나 전체 데이터셋이나 더 작은 하위 집합을 얻을 수 있도록 한다.

3 Analysis & Results

웹 데모와 유사도 검색. 사용자가 입력 이미지 또는 텍스트의 CLIP 임베딩과 사전 계산된 kNN 인덱스를 이용해 이미지와 텍스트를 검색할 수 있도록 웹 데모를 만들었다. 이는 LAION-400M에서 찾을 수 있는 이미지와 캡션의 다양성뿐만 아니라 높은 의미적 관련성을 보여준다 (Fig. 1). Tab. 1은 LAION-400M의 이미지 크기 분포를 보여준다. 고해상도 이미지가 풍부하므로, 다양한 맞춤형 모델을 학습하기 위한 이미지 하위집합을 생성할 수 있으며, 특정 학습 목적에 적합한

이미지 해상도를 선택할 수 있다.

DALL-E 모델 학습. 우리는 텍스트-투-이미지 모델을 학습할 수 있는 데이터셋의 능력을 평가하기 위해 DALLE-pytorch [11]를 실행했다. 이미지 토큰을 인코딩하기 위해 ImageNet에서 사전학습된 VQGAN [12]을 사용했다. 생성 시에는 캡션당 총 128 샘플 중 상위 8개를 순위 매기기 위해 CLIP ViT-B/16 [1]을 사용했다. 단일 에포크에 약 720만 이미지의 하위집합만 본에도 불구하고 다양한 범주에서 빠른 수렴을 관찰했다. 모델에서 생성된 샘플은 충분한 품질을 보이며 성공적인 학습 진행의 증거를 제공한다 (Fig. 3).

4 CONCLUSION

4억 개의 이미지-텍스트 쌍을 포함하는 공개 데이터셋을 공개함으로써 DALL-E와 CLIP과 같은 최첨단 언어-비전 모델을 훈련하는 데 필요했던 독점적 대규모 데이터셋과의 격차를 해소했다. 개념 증명으로서, 데이터셋의 하위 집합을 사용해 DALL-E 모델을 훈련시켜 충분한 품질의 샘플을 생성할 수 있음을 보였다. 이 데이터셋은 이전에 독점적 대규모 데이터셋에 접근할 수 있는 사람들로 제한되었던 언어-비전 모델의 대규모 훈련 및 연구의 길을 폭넓은 커뮤니티에 열어준다.