# BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

https://arxiv.org/pdf/2201.12086

## 0. Introduction

- 비전-언어 모델 BLIP는 이미지와 텍스트를 함께 이해하고 생성하기 위해 설계됨
- 기존 모델들이 이미지와 텍스트를 따로 다루거나 대규모 데이터 필요로 하는 문제를 해결하고자 함
- BLIP는 사전 학습과 부트스트래핑 방식으로 효과적인 멀티모달 표현 학습을 목표로 함
- 이를 통해 이미지 캡션 생성, 이미지-텍스트 검색, 비주얼 질문응답 등 다양한 작업에서 뛰어난 성능을 보임
- 모델 크기와 학습 데이터 효율성을 개선해 실제 응용에 적합하게 설계됨

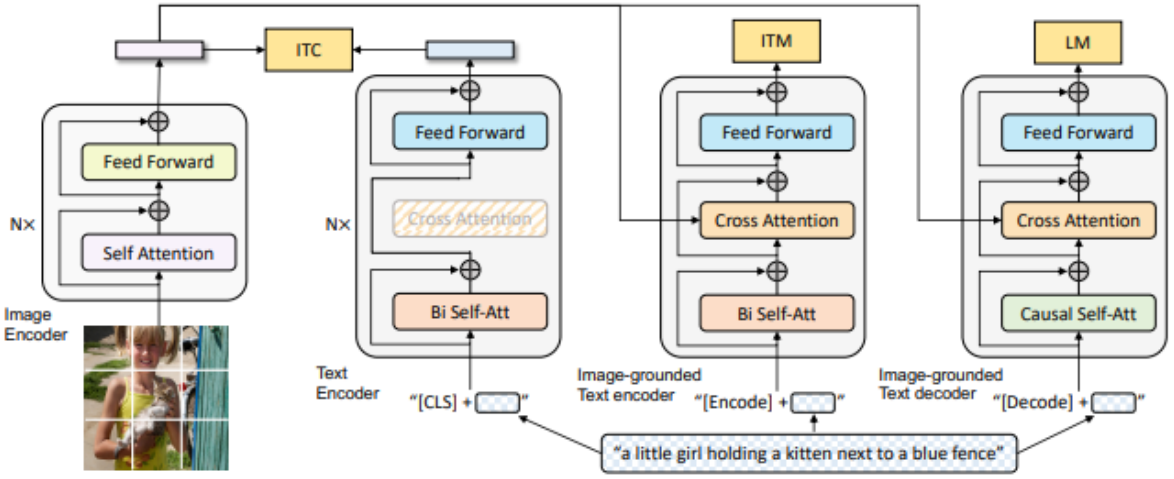
## 1. Overview

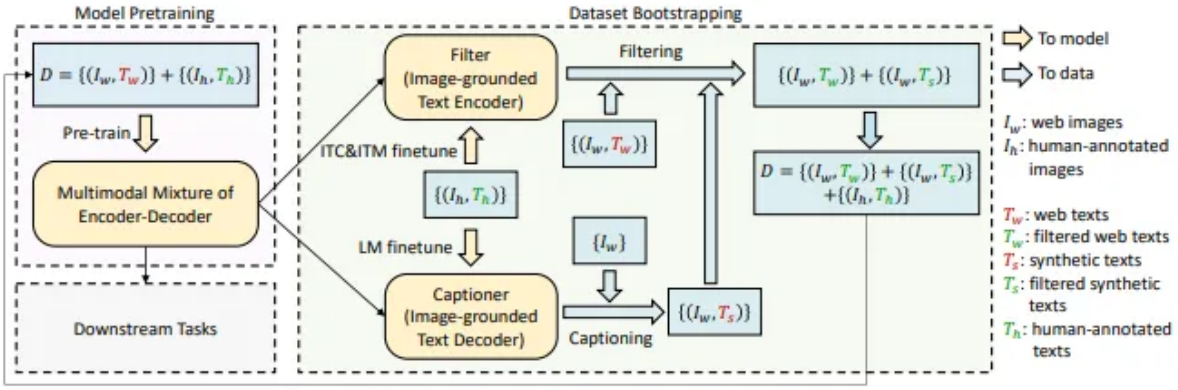
- BLIP는 이미지와 텍스트를 통합해 다룰 수 있는 멀티모달 모델임
- 사전학습 단계에서 이미지-텍스트 쌍을 활용해 시각-언어 표현을 동시에 학습함
- 세 가지 주요 구성요소로 설계됨: 이미지 인코더, 텍스트 인코더, 그리고 두 모달리티를 연결하는 모듈
- 부트스트래핑 학습 방식을 사용해, 자기 지도 학습과 지도 학습을 효과적으로 결합함
- 다양한 비전-언어 작업(이미지 캡션, 이미지-텍스트 검색, 비주얼 질문응답 등)에 대해 높은 범용성과 성능을 달성함

## 2. Challenges

- 이미지와 텍스트의 복합적 관계를 효과적으로 모델링하기 어려움
- 대규모 멀티모달 데이터 없이도 강력한 사전학습을 수행하는 게 어려움
- 기존 모델들은 이미지와 텍스트 정보를 별도로 처리해 통합 표현 학습에 한계가 있음
- 다양한 비전-언어 작업에 모두 잘 대응할 수 있는 범용 모델 설계가 어려움
- 데이터 잡음과 불완전한 레이블로 인한 학습 효율 저하 문제 존재

# 3. Method




- BLIP는 이미지 인코더와 텍스트 인코더로 각각 시각 및 언어 정보를 추출함
- 두 모달리티를 연결하는 cross-modal 모듈을 통해 이미지와 텍스트 간 상호작용을 학습함
- 부트스트래핑 학습 전략 사용: 초기에는 약한 라벨 또는 생성된 캡션으로 자기지도 학습을 수행

- 이후 점진적으로 신뢰할 수 있는 데이터로 지도학습을 강화해 성능 개선
- 학습 과정에서 이미지-텍스트 매칭, 이미지 캡션 생성, 비주얼 질문응답 등 다양한 목표 함수를 활용
- 사전학습 후, 특정 태스크에 맞춰 미세조정(fine-tuning) 수행해 범용성과 특화 성능 모두 확보

# 4. Experiments

- 사용한 데이터셋: COCO, Flickr30k, Visual Genome 등 이미지-텍스트 페어 데이터
- 평가 태스크: 이미지 캡션 생성, 이미지-텍스트 검색, 비주얼 질문응답(VQA)
- 비교 대상 모델: CLIP, ViLBERT, UNITER 등 멀티모달 사전학습 모델
- 평가 지표: BLEU, CIDEr, METEOR(캡션), Recall@K(검색), 정확도(VQA)
- 사전학습 후 태스크별 미세조정(fine-tuning)으로 실험 진행
- 다양한 태스크와 데이터셋에서 범용성과 효율성 평가됨

# 5. Results

| Method | Pre-train #Images | NoCaps validation | | | | | | | | COCO Caption Karpathy test | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | in-domain | | near-domain | | out-domain | | overall | | | |
| | | C | S | C | S | C | S | C | S | B@4 | C |
| Enc-Dec (Changpinyo et al., 2021) | 15M | 92.6 | 12.5 | 88.3 | 12.1 | 94.5 | 11.9 | 90.2 | 12.1 | - | 110.9 |
| VinVL† (Zhang et al., 2021) | 5.7M | 103.1 | 14.2 | 96.1 | 13.8 | 88.3 | 12.1 | 95.5 | 13.5 | 38.2 | 129.3 |
| $\text{LEMON}_{\text{base}}$† (Hu et al., 2021) | 12M | 104.5 | 14.6 | 100.7 | 14.0 | 96.7 | 12.4 | 100.4 | 13.8 | - | - |
| $\text{LEMON}_{\text{base}}$† (Hu et al., 2021) | 200M | 107.7 | 14.7 | 106.2 | 14.3 | 107.9 | 13.1 | 106.8 | 14.1 | **40.3** | **133.3** |
| BLIP | 14M | 111.3 | 15.1 | 104.5 | 14.4 | 102.4 | 13.7 | 105.1 | 14.4 | 38.6 | 129.7 |
| BLIP | 129M | 109.1 | 14.8 | 105.8 | 14.4 | 105.7 | 13.7 | 106.3 | 14.3 | 39.4 | 131.4 |
| $\text{BLIP}_{\text{CapFilt-L}}$ | 129M | **111.8** | **14.9** | **108.6** | **14.8** | **111.5** | **14.2** | **109.6** | **14.7** | 39.7 | **133.3** |
| $\text{LEMON}_{\text{large}}$† (Hu et al., 2021) | 200M | 116.9 | 15.8 | 113.3 | 15.1 | 111.3 | 14.0 | 113.4 | 15.0 | 40.6 | 135.7 |
| $\text{SimVLM}_{\text{huge}}$ (Wang et al., 2021) | 1.8B | 113.7 | - | 110.9 | - | 115.2 | - | 112.2 | - | 40.6 | 143.3 |
| $\text{BLIP}_{\text{ViT-L}}$ | 129M | 114.9 | 15.2 | 112.1 | 14.9 | 115.3 | 14.4 | 113.2 | 14.8 | 40.4 | 136.7 |

| Method | MRR↑ | R@1↑ | R@5↑ | R@10↑ | MR↓ |
|---|---|---|---|---|---|
| VD-BERT | 67.44 | 54.02 | 83.96 | 92.33 | 3.53 |
| VD-ViLBERT† | 69.10 | 55.88 | 85.50 | 93.29 | 3.25 |
| BLIP | **69.41** | **56.44** | **85.90** | **93.30** | **3.20** |

| Method | MSRVTT-QA | MSVD-QA |
|---|---|---|
| *zero-shot* | | |
| VQA-T (Yang et al., 2021) | 2.9 | 7.5 |
| BLIP | 19.2 | 35.2 |
| *finetuning* | | |
| HME (Fan et al., 2019) | 33.0 | 33.7 |
| HCRN (Le et al., 2020) | 35.6 | 36.1 |
| VQA-T (Yang et al., 2021) | 41.5 | 46.3 |

| Method | R1↑ | R5↑ | R10↑ | MdR↓ |
|---|---|---|---|---|
| *zero-shot* | | | | |
| ActBERT (Zhu & Yang, 2020) | 8.6 | 23.4 | 33.1 | 36 |
| SupportSet (Patrick et al., 2021) | 8.7 | 23.0 | 31.1 | 31 |
| MIL-NCE (Miech et al., 2020) | 9.9 | 24.0 | 32.4 | 29.5 |
| VideoCLIP (Xu et al., 2021) | 10.4 | 22.2 | 30.0 | - |
| FiT (Bain et al., 2021) | 18.7 | 39.5 | 51.6 | 10 |
| BLIP | **43.3** | **65.6** | **74.7** | **2** |
| *finetuning* | | | | |
| ClipBERT (Lei et al., 2021) | 22.0 | 46.8 | 59.9 | 6 |
| VideoCLIP (Xu et al., 2021) | 30.9 | 55.4 | 66.8 | - |

- BLIP가 대부분 태스크에서 기존 SOTA 대비 우수한 성능 기록
- 이미지 캡션 생성에서 BLEU, CIDEr 등 지표 크게 향상
- 이미지-텍스트 검색에서 Recall@K 지표에서 높은 정확도 달성
- VQA 태스크에서도 정확도 면에서 경쟁력 있는 결과 보여줌
- 다양한 데이터셋과 태스크에서 일관되고 강력한 성능 입증
- 사전학습과 미세조정의 조합이 모델 성능 향상에 크게 기여함

# 6. Insight

- 부트스트래핑 학습으로 지도학습과 자기지도학습을 효과적으로 결합해 멀티모달 표현 강화 가능
- 이미지와 텍스트를 동시에 이해하고 생성하는 통합 모델이 다양한 비전-언어 작업에 범용적이고 강력함
- 데이터 효율성을 높여 비교적 적은 데이터로도 우수한 성능 달성 가능
- 멀티모달 사전학습이 실제 응용에서 이미지 캡션, 검색, 질문응답 등 다양한 분야에 크게 기여할 수 있음
- 다만, 복잡한 학습 구조와 계산 비용, 그리고 일부 작업에서의 미세한 성능 차이는 개선 여지가 있음