Connecting language and vision Summary pre-report

ByeongGeun Shin[1],
[1]Pusan National University.

1 Learning Transferable Visual Models From Natural Language Supervision(CLIP) Paper Summary

1.1 Abstract

최근의 딥러닝 모델들은 auto-regressive, MLM과 같이 task-agnostic objective를 사용하되 모델사이즈를 키워 성능을 매우 올려왔으며, 기존의 crowd-labeled 데이터셋 보다 웹상에서 모은 데이터셋을 활용하는게 더 효과적이었다. Computer vision의 경우 여전히 label set이 고정된 형태의 데이터가 많이 사용되고 있었으며 때문에 확장성이 있는 모델을 만드는데 한계가 있었다. 자연어 표현을 label로 활용한 연구들도 있었지만 zero-shot 성능이 매우 떨어지는 한계도 있었다.

본 연구에서는 natural languague supervision, scale(모델 크기), large web data를 적용한 pre-trained image classifier의 성능을 연구했으며 그 결과 GPT 계열과 마찬가지로 OCR, geol-ocalization 등과 같은 다양한 태스크에서 좋은 성능을 보이는 것을 확인하였다.

1.2 Introduction

최근 딥러닝 모델이 Vision Model 과 Language Model 두가지 카테고리로 발전이 되어가는 도중, Vision Model 은 어떻게 모델을 구성해야, 이미지를 입력받았을때 더 좋은 표현을 학습하는지가 주된 연구 포인트였다. 따라서, Inception, ResNet 과 같이 효율적으로 더 깊은 모델을 만드는 방법을 고민했다. SENet, BAM,CBAM 등은 어텐션 모듈을 사용하는 방법을 제시했다. 또, Visual Model 의 트렌드는 Transformer 구조를 적용한 Vision Transformer 등이 발표되었다. 하지만, 이미지만 학습한 모델은 일반화 능력이 부족하고, 노이즈들에 취약한 문제들이 있었다.

Language Model 은 2017년 Transformer 의 발표에 힘입고, 큰 발전을 이루었는데, 트랜스포머 구조로 긴문장도 효과적으로 처리할 수 있게 되었다. 그러나 Vision 분야에서는 아직 ImageNet과 같은 제한된 크기의 Labeled dataset이 관행적으로 사용되고 있었다. ImageNet의 데이터셋의 양이 적은 편은 아니나, 사람이 직접 사전 처리를 해주어야 하는 labeled dataset으로 학습하는 방식은 사용할 수 있는 데이터의 개수에 한계가 있었다. NLP에서와 같이, Vision 분야에서도 대규모의 raw data(Image-text data)를 이용하여 학습(Pre-train)하는 방식을 차용한다는 생각에서 기존의 한계를 극복하고자 했다.

1.3 Method

1.3.1 Natural Language Supervision

핵심 아이디어는 자연어에 담겨있는 인간의 인지, 지각(perception)을 학습하는 것이다(learning perception from supervision contained in natural language). 대부분의 연구들은 이를 unsupervised, self-supervised, 혹은 supervised와 같이 표현하는데 CLIP 논문에서는 자연어 label 존재 여부와 관계 없이 자연어 자체를 사용하는 방법을 사용하였다(supversion). 이러한 방법은 매우 많은 데이터로 scale을 키울 수 있고, 자연어의 개념이 담긴 representation을 학습하여 zero-shot 성능을 향상시킬 수 있다.

1.3.2 Creating a Sufficiently Large Dataset

기존의 computer vision 데이터셋들은 많은 비용을 들여 crowd sourcing으로 데이터에 label을 달았다. CLIP에서는 웹에서 (image, text) 데이터를 모으되 무의미한 text를 거르기 위해 등장 빈도가 비교적 높은 단어들로 필터링하여 데이터셋을 구축하였다.

1.3.3 Selecting an Efficient Pre-Training Method

CLIP에서는 image-text pre-training을 위해 효과적인 학습 방법을 찾는 것이 매우 중요하였다고 한다. 첫번째로 시도한 방법은 CNN & text transformer + autoregressive caption text 학습방법이었는데 비효율적이었다고 한다. 이미지마다 정확한 token을 학습하는 것은 학습 효율성 뿐만 아니라 이미지의 표현 다양성을 해치기도 한다.

최근 연구들에서는 contrastive representation learning이 효과적인 방법임이 증명되고 있고 이미지의 generative model을 학습시키는 방법보다 계산량이 적다. CLIP에서는 이에 착안하여 text 자체를 학습하는게 아닌 이미지가 어떤 whole text와 쌍인지를 학습하는 방법을 사용하였다. 그림처럼 N개의 (image, text) 쌍이 있을 때 N개의 정답에 대한 cosine similarity는 극대화하고 N$\hat{2}$ - N개의 쌍의 cosine similarity는 낮추는 방법으로 학습하였다.

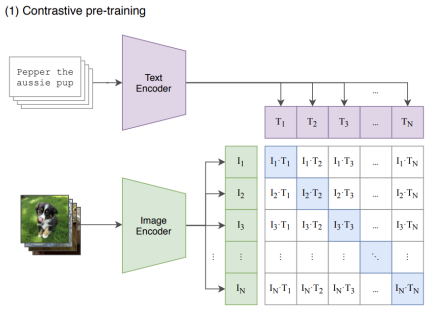


Figure 1: Contrastive Pre-Training.

1.3.4 Choosing and Scaling a Model

CLIP에서는 image encoder로 몇 가지 변형이 추가된 ResNet-50, ViT(Vision Transformer) 두개를 사용하였고 text encoder로는 Transformer를 활용하였다.

1.4 Experiments

1.4.1 Zero-Shot Transfer & Linear Probing

Image-Natural Language text 쌍에 대해 Pre-train된 CLIP을 이용하여, Zero-shot prediction (Classification)을 수행하도록 할 수 있다. dataset에서, dataset에 있는 모든 class(label)의 name을 text snippet으로 변경하고, input image와 함께 encoding하여 cosine similarity를 계산하여 가장 높은 similarity를 가진 text snippet을 선택하면 된다.

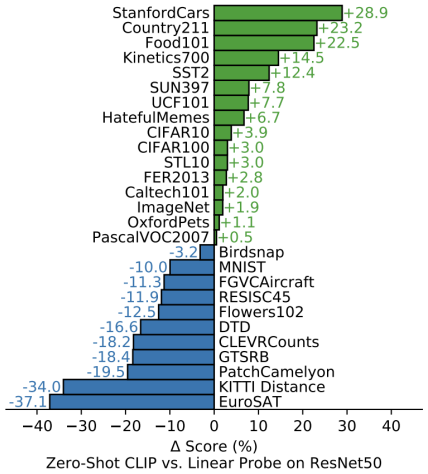


Figure 2: Zero-Shot CLIP vs Linear probe on Resnet50.

이러한 CLIP에서의 Zero-shot 성능이, ResNet에서의 Linear Probing 보다 여러 dataset에 대해 (위의 결과에서는 절반 이상의) 성능이 좋다는 결과가 도출되었다. * Linear Probe란 학습이 완료된 Encoder를 가져와 Supervised Learning으로 Classifier만 재학습해주는 방법이다. 만약 Encoder가 좋은 표현을 많이 학습했다면 단순히 Classifier만 재조정 해주어도 높은 성능이 나올 것이라는 전제가 깔려있는 방법이다.

1.5 Result

다른 모델들과의 비교에서도 CLIP 모델은 성능, 계산 효율성 측면에서 모두 매우 좋은 성능을 보여준다. 저자들은 CLIP이 pre-train 과정에서 end-to-end fine-tuning 보다 더 많은 task들의 정보를 학습했기 때문이라고 분석하였다. 변형된 Natural Distribution Shift dataset에서도 다른 표준 ImageNet 모델들보다 더욱 월등한 성능을 보인다. 따라서 robust함도 증명하였다. 따라서 CLIP모델은 Label 데이터를 전혀 사용하지 않고도 Label 정보를 사용하여 학습한 동일한 모델보다 더 좋은 성능을 보여주었기 때문에 기존의 SimCLR, BiT 등 좋은 표현을 학습한다고 알려진 다른 방법들보다 좋은 표현을 학습한다는 점이 검증되었다.

2 Linearly Mapping from Image to Text Space Paper Summary

2.1 Abstract

이 논문은 텍스트 전용 언어 모델(LMs)과 비전 전용 모델의 개념적 표현이 구조적으로 유사하다는 가설을 탐구한다. 이를 위해 LiMBeR (Linearly Mapping Between Representation spaces)라는 방법을 제안하며, 단일 선형 변환을 통해 이미지 표현을 연속적인 프롬프트로 변환하여 고정된 LM에 입력으로 제공한다. 이 방법은 이미지 캡셔닝 및 시각적 질문 응답 과제에서 경쟁력 있는 성능을 보여준다. 결과적으로 이미지 인코더 사전 학습 중 언어적 감독의 중요성을 입증하며, 비전 모델과 언어 모델 간 구조적 유사성을 뒷받침한다.

2.2 Introduction

최근 NLP 분야에서는 텍스트 전용 데이터로 학습된 언어 모델이 비언어적 개념을 얼마나 잘 학습할 수 있는지에 대한 논의가 활발하다. 본 논문에서는 이러한 모델들이 비전 모델과의 구조적 유사성을 통해 개념적 표현을 공유할 수 있다는 가설을 제안한다. 이를 검증하기 위해 LiMBeR라는 간단한 프레임워크를 도입하여 이미지 표현을 선형 변환을 통해 텍스트 입력 공간으로 매핑한다. 이 접근법은 모델의 매개변수를 수정하지 않으면서, 텍스트와 비전 모델 간의 구조적 유사성을 탐구하는 데 중점을 둔다.

2.3 Method

LiMBeR는 이미지 표현을 텍스트 모델의 입력 공간으로 선형 변환하는 접근법이다. 이전의 연구에서는 이미지를 언어 모델의 소프트 프롬프트로 변환하여 멀티모달 학습을 시도했으나, 이러한 변환의 메커니즘을 제한적으로 이해하려는 시도는 없었다.

본 연구에서는 이미지 인코더에서 생성된 고차원 표현을 텍스트 언어 모델의 입력 공간으로 매핑하기 위해, 단일 선형 변환 레이어인 P를 학습한다. 이 변환된 입력은 텍스트의 개별적인 토큰과 일치하지 않으며, "soft prompts"로 해석된다.

Figure 3: input space of a language model to produce captions describing images.

2.3.1 Image Encoders

연구진는 언어 모델이 학습하는 개념적 표현이 이미지 인코더에서 학습되는 표현과 선형 변환을 통해 일치한다고 가정한다. 이미지 인코더는 세 가지로 나누어 실험을 진행했다.

첫번째는 BEIT로 언어적 감독 없이 자체 지도 학습 방식으로 이미지의 시각적 토큰을 예측하도록 학습된 모델이다. 이는 가장 단순한 비전 전용 모델로, 언어와의 연결이 거의 없다.

두번째 모델은 NF-ResNet라는 ImageNet에서 분류 작업을 통해 학습된 모델로, 하이퍼니즘 구조를 활용해 간접적인 언어적 감독을 포함한다. 이 모델은 텍스트-비전 간 연결성을 일부 학습한다.

마지막은 이전 리뷰 논문에서 다룬 CLIP모델로 이미지와 텍스트 캡션의 정렬을 학습하는 방식으로 사전 학습된 모델이다. 자연어와 이미지를 연결하기 때문에 가장 높은 수준의 언어적 감독을 가진다.

이 이미지 인코더들로부터 추출한 이미지 표현을 선형 변환을 통해 텍스트 입력 공간으로 매핑하며, 이를 통해 이미지 캡셔닝 작업을 수행한다.

2.3.2 Language Model

제안하는 LiMBeR는 60억 개 매개변수를 가진 GPT-J를 언어 모델로 사용한다. 언어 모델의 입력 공간은 4096차원의 벡터로 구성되며, 이는 선형 변환된 이미지 표현을 입력으로 받아 문장을 생성한다. GPT-J는 고정된 상태에서 선형 변환된 이미지 입력으로 캡션을 생성하거나 질문에 답하는 능력을 평가받는다.

2.3.3 Linear Projection

LiMBeR 모델의 핵심은 이미지 인코더의 출력 공간을 언어 모델 입력 공간으로 매핑하는 단일 선형 변환 레이어이다. 이 레이어는 고정된 이미지 인코더의 출력과 언어 모델의 입력 차원을 맞추며, "soft prompts"로 불리는 연속적인 벡터로 변환한다. 이러한 프롬프트는 텍스트 토큰이 아닌 연속적인 표현으로, 모델 학습 중 캡셔닝 작업에 최적화된다.

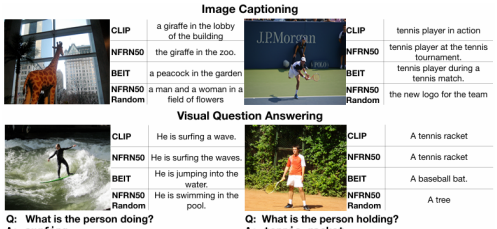
2.4 Training Procedure

LiMBeR의 훈련은 Conceptual Captions 3M 데이터셋을 사용하여 15,000 스텝 동안 학습되었다. AdamW 옵티마이저를 사용하며, 이미지와 캡션 쌍이 모델 입력으로 제공된다. 모든 모델은 이미지 표현을 언어 모델의 입력으로 변환하기 위해 선형 변환 레이어만 학습하고, 이미지 인코더와 언어 모델은 고정된 상태로 유지된다. 이를 통해 우리는 이미지 인코더의 표현 공간과 언어 모델의 입력 공간 간의 구조적 유사성을 분석할 수 있다. 캡션 생성의 정확도를 높이기 위해 프롬프트로 "A picture of"를 추가하여 언어 모델이 보다 정확한 설명을 생성할 수 있도록 유도한다.

2.5 Limitations

본 연구에서의 주요 한계 중 하나는 각 이미지 인코더에 대한 프롬프트 길이(k)의 제어가 부족하다는 점이다. Tsimpoukelli et al. (2021)은 프롬프트 길이(k)가 모델 간 비교에 큰 영향을 미치지 않는다고 보고한 바 있다. 그러나 CLIP과 BEIT에서 k 값이 상대적으로 더 크기 때문에, 본 연구에서는 이를 철저하게 통제하지 않았다. 이는 모델 성능에 미치는 영향을 최소화하려 했지만, 여전히 일부 모델 간 차이에 영향을 줄 수 있는 요소로 작용할 수 있다.

또한, 언어 모델의 "runoff" 현상도 한계로 고려된다. 예를 들어, 언어 모델이 "the beach"와 같은 개념을 인식하고 관련 단어를 생성한 경우, 그 뒤에 "building a sandcastle"과 같은 연관된 단어가 생성될 수 있다. 이로 인해 모델이 이미지의 모든 요소를 인식한 것처럼 보일 수 있지만, 실제로는 이미지에 나타난 요소를 정확히 반영하지 못할 수 있다. 이 문제는 일부 데이터셋에서는 발생할 수 있으나, 다수의 대형 데이터셋에서 우리의 결과는 여전히 이미지 정보를 복원할 수 있음을 보여준다. 우리는 이를 시각적 질문 응답 분석에서 '블라인드' 모델을 사용하여 통제하였다.



2.6 Experiments

LiMBeR는 이미지 캡셔닝과 시각적 질문 응답 작업에서 평가되었다. MSCOCO, NoCaps, VQA2 데이터셋을 사용하여 수행된 실험 결과, LiMBeR는 MAGMA와 같은 복잡한 멀티모달 모델과 유사한 성능을 보였다. 특히, CLIP 인코더를 사용할 때 가장 높은 성능을 기록했으며, 이는 언어적 감독이 중요한 역할을 한다는 것을 보여준다. 반면, BEIT는 세부적인 단어 카테고리 전이에 어려움을 겪었지만, 시각적 속성을 잘 전달했다.

LiMBeR는 VQA에서의 성능도 평가하였으며, 여러 "샷"(shots) 설정에 따른 정확도를 기록했다. CLIP 모델은 다른 모델들보다 우수한 성능을 보였으며, NFRN50이나 BEIT보다 VQA 작업에서 뛰어난 결과를 보였다. 그러나 BEIT 모델은 복잡한 시각적-언어적 추론을 처리하는 데 어려움을 겪었다.

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

VQA n-shots	0	1	2	4
Blind	20.60	35.11	36.17	36.99
NFRN50 Tuned	27.15	37.47	38.48	39.18
MAGMA (ours)	24.62	39.27	40.58	41.51
MAGMA (reported)	32.7	40.2	**42.5**	43.8
NFRN50 Random	25.34	36.15	36.79	37.43
BEIT	24.92	34.35	34.70	31.72
NFRN50	27.63	37.51	38.58	39.17
CLIP	33.33	39.93	**40.82**	40.34

2.7 Conclusion

본 연구는 이미지 표현을 텍스트 모델의 입력 공간으로 변환하는 간단한 선형 변환이 효과적이라는 점을 입증하였다. 이를 통해 비전 모델과 언어 모델 간의 개념적 표현이 구조적으로 유사함을 확인할 수 있었으며, 이미지 인코더의 사전 학습에서의 언어적 감독 수준이 전이 성능에 중요한 역할을 한다는 것을 보여주었다. LiMBeR는 복잡한 멀티모달 모델에 대한 유효한 기준선을 제공하며, 다양한 데이터 유형 간의 표현적 유사성을 이해하는 데 중요한 출발점을 제시한다.