2023

VILBERT :

Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.

FOM
Focus On Data Mining

FOM
Focus On Data Mining

INDEX

1. Introduction
2. Background
3. Method
4. Experiment
5. Conclusion

00 | 전주시 데이터 공모전

❖전주시 데이터 공모전 출전 확정

❖타이틀 : “이미지 캡셔닝을 활용한 SNS 게시물 생성기”

❖팀명 : **버트의 온고을 이야기**

❖06.16 까지 크롤링, 모델링 등 다양한 작업 필수

01 | Introduction

ViLBERT

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

Jiasen Lu[1], **Dhruv Batra**[1,2], **Devi Parikh**[1,2], **Stefan Lee**[1,3]
[1]Georgia Institute of Technology, [2]Facebook AI Research, [3]Oregon State University

Abstract

We present ViLBERT (short for Vision-and-Language BERT), a model for learning task-agnostic joint representations of image content and natural language. We extend the popular BERT architecture to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. We pretrain our model through two proxy tasks on the large, automatically collected Conceptual Captions dataset and then transfer it to multiple established vision-and-language tasks – visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval – by making only minor additions to the base architecture. We observe significant improvements across tasks compared to existing task-specific models – achieving state-of-the-art on all four tasks. Our work represents a shift away from learning groundings between vision and language only as part of task training and towards treating visual grounding as a pretrainable and transferable capability.

01 Introduction

ViLBERT

❖**ViLBERT**는 인공지능의 두 핵심 도메인인 **컴퓨터 비전**(Computer Vision)과 **자연 언어 처리**(NLP, Natural Language Processing)를 동시에 처리하는 BERT 기반 모델

➢ 해당 논문에서 주장하는 **가장 중요한 task**는 기존의 Vision & Language task transformer와 다른 Co-TRM을 사용

➢ image tokenized를 위해 OD task의 **Faster R-CNN**을 사용해 임베딩

❖이 모델의 목표는 이미지와 관련된 텍스트 정보를 동시에 이해하고, 다양한 시각-언어 작업에 대해 학습하는 것

❖이로써 이미지에 대한 설명을 생성하거나, 이미지 내의 특정 개체에 대한 질문에 대답하는 등의 작업을 수행 가능

02 | Background

How does it relate to the Bert model?

❖**ViLBERT는 Transformer** 아키텍처를 기반으로 하며, BERT의 성공적인 언어 모델링 전략을 활용

❖BERT는 언어 처리에서 큰 성과를 내왔는데, 이는 주변 문맥에 따라 *단어의 의미를 동적으로 해석*할 수 있기 때문

➢BERT는 MLM(Masked Language Model), NSP(Next Sentence Prediction)을 통해 ***양방향 학습***

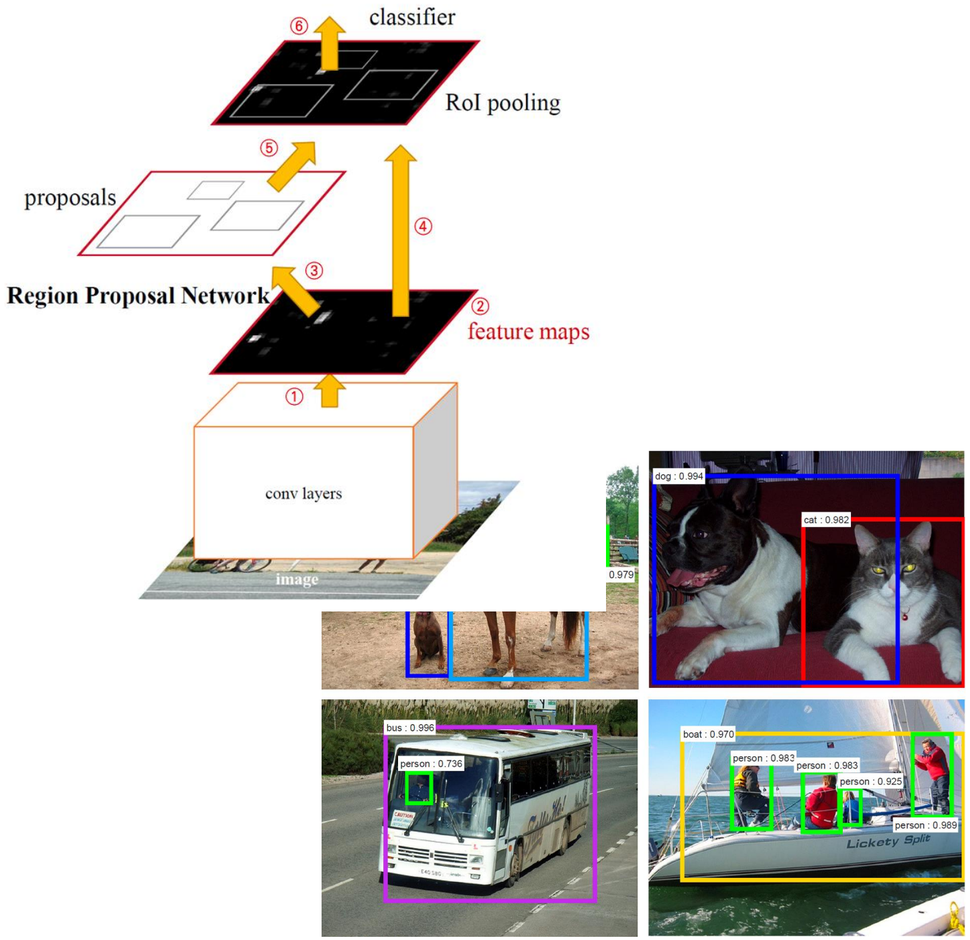
➢그러나 BERT는 텍스트 데이터만 처리할 수 있어, 이미지와 같은 비텍스트 데이터를 처리하려면 추가적인 기술이 필요

❖이에, ViLBERT는 이미지 처리에 능한 모델과 BERT를 결합하여 이미지와 텍스트를 함께 이해할 수 있게 만듦

"Bert 모델은 Transformer의 Encoder-Only 모델"

02 | Background

What is Faster R-CNN?



❖**Faster R-CNN**은 이미지 내의 개별 객체를 감지하고 그들의 영역(region)을 추출하는 데 특화된 *컴퓨터 비전 모델*

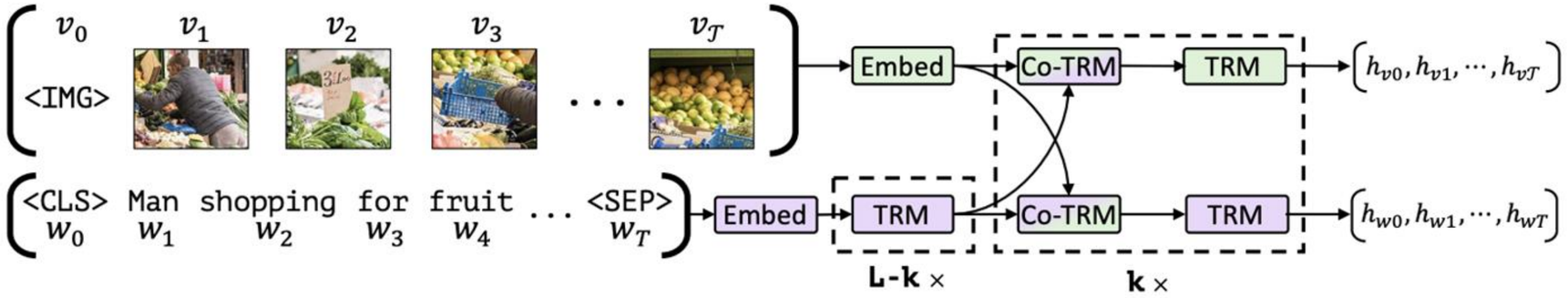
❖이미지를 여러 개의 영역(region)으로 분할하고 각 영역이 무엇을 나타내는지를 파악하게 됨

➢추출된 영역별 정보는 이미지의 **특성 벡터**로 변환

➢이 벡터들은 그 다음에 VilBERT의 시각 인코더에 입력으로 제공되며, 이 과정을 통해 이미지와 텍스트 정보가 함께 처리됨

03 | Method

ViLBERT Model



❖ViLBERT는 두 개의 별도의 Transformer 인코더를 사용하는 'Two-Stream' 모델을 채택 :

➢ 하나는 시각적 입력을 처리하고, 다른 하나는 언어적 입력을 처리 각 스트림은 자체 입력을 독립적으로 처리

➢ **특정 단계**에서 두 스트림 사이에 정보를 교환하게 됨

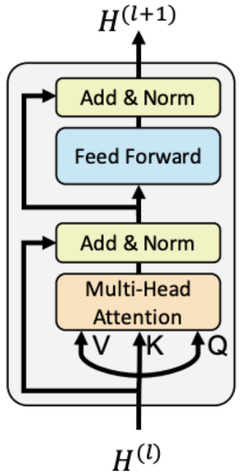
❖이 'co-attentional' 메커니즘은 이미지의 특정 부분과 연관된 텍스트, 또는 그 반대의 정보를 모델이 파악하는 데 도움

➢ 또한, 이 메커니즘이 각각의 스트림이 각자의 도메인에 대해 특화된 표현을 학습

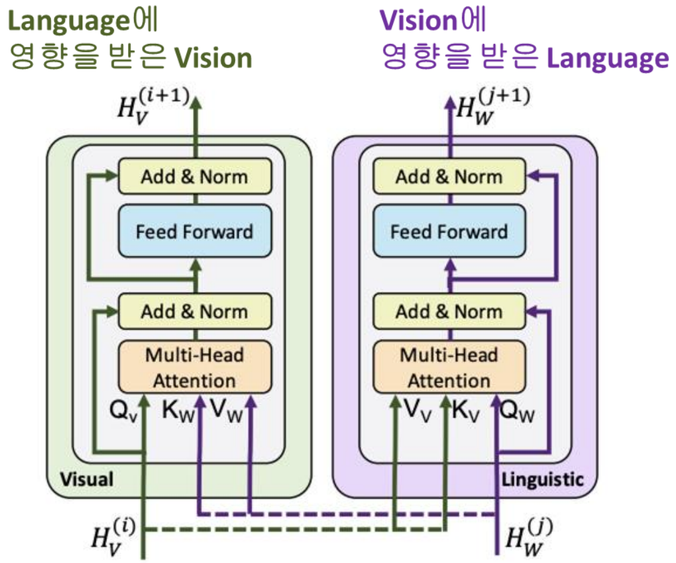
➢ 다른 도메인과의 연결을 유지하게 만듬

03 | Method

Encoder-Only Transformer in ViLBERT



(a) Standard encoder transformer block



(b) Our co-attention transformer layer

(a) **BERT** 모델에서의 Encoder-Only 모델 :

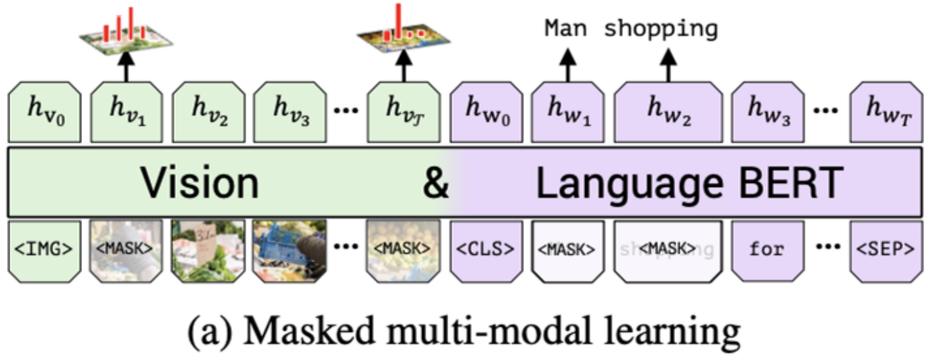
embeding된 text data MHA에서 각각의 Q,K,V가 계산 됨

(b) **ViLBERT** 모델에서의 Encoder-Only 모델 :

embeding된 text data와 image data가 co-attention 형식으로 Q,K,V가 계산됨

03 | Method

Masked multi-modal modelling



(a) Masked multi-modal learning

❖BERT의 **MLM**처럼 입력값을 마스킹하며, 언어에 대해서는 동일하고 image의 경우 15%의 영역을 선택, 그 중 90%의 영역을 지움

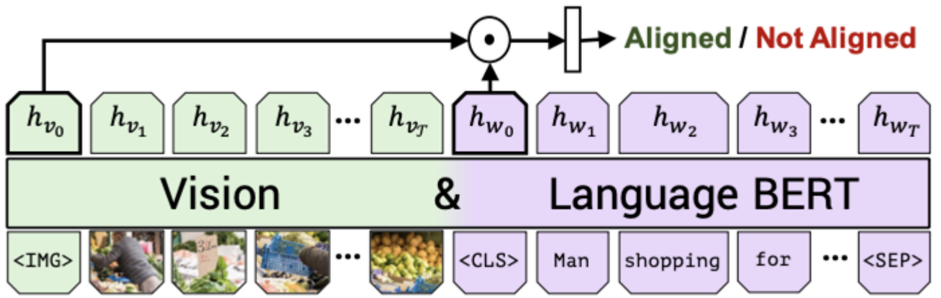
❖ 이미지에서는 masking된 feature의 값을 예측하기 보다는, 해당 영역의 **semantic class에 대한 분포**를 예측

→ 왜냐하면, 이미지에 대한 설명은 각 이미지 영역 자체의 feature를 설명하는 게 아니고, 높은 레벨의 맥락을 다루기 때문

03 | Method

Multi-modal alignment prediction

- 이미지-텍스트 쌍이 입력으로 주어졌을 때, 텍스트가 이미지를 설명하는 것이 **맞는지(aligned) 아닌지(not aligned)**를 판별
- ⟨IMG⟩, ⟨CLS⟩ 토큰을 BERT의 ⟨CLS⟩ 토큰 처럼 사용하며,
  이 두 개의 토큰들의 element wise product를 두 모달리티를 거친 입력값의 전체적인 표현으로 사용해 판별 과제를 수행



(b) Multi-modal alignment prediction

03 | Method

ViLBERT

VQA

VCR Q→A VCR QA→R

Referring Expressions

Caption-Based Image Retrieval

04 Experiment

ViLBERT

❖ViLBERT는 여러 가지 시각-언어 작업에서 테스트되었으며, 이에는 이미지 설명 생성, 시각적 질문 응답, 시각적 추론 등이 포함

❖이러한 실험에서 ViLBERT는 각각의 작업을 동시에 학습하고 전이 학습을 활용하여 다양한 작업에 적용할 수 있음을 보임

❖이러한 실험 결과는, 모델이 이미지와 텍스트에 걸쳐 있는 복잡한 패턴을 동시에 학습하고 이해할 수 있음을 보여줌

04 | Experiment

ViLBERT

	Method	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
		test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
SOTA	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
SOTA	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
SOTA	SCAN [35]	-	-	-	-	-	-	-	48.60	77.70	85.20	-	-	-
Ours	Single-Stream†	65.90	68.15	68.89	47.27	65.64	72.02	56.04	-	-	-	-	-	-
Ours	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
Ours	ViLBERT†	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
Ours	ViLBERT	**70.55 (70.92)**	**72.42 (73.3)**	**74.47 (74.6)**	**54.04 (54.8)**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Method	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
	test-dev	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
ViLBERT (2-layer)	69.92	72.44	**74.80**	**54.40**	71.74	**78.61**	62.28	55.68	84.26	90.56	26.14	56.04	68.80
ViLBERT (4-layer)	70.22	**72.45**	74.00	53.82	72.07	78.53	**63.14**	55.38	84.10	90.62	26.28	54.34	66.08
ViLBERT (6-layer)	**70.55**	72.42	74.47	54.04	**72.34**	78.52	62.61	58.20	84.90	**91.52**	31.86	61.12	72.80
ViLBERT (8-layer)	70.47	72.33	74.15	53.79	71.66	78.29	62.43	**58.78**	**85.60**	91.42	**32.80**	**63.38**	**74.62**

➢ 이 때 당시엔 **SoTA**였으나, 현재는 VLBert -〉 UNITER 순으로 성능이 상승

04 Experiment

ViLBERT

The concept comes to life with a massive display of fireworks that will fill the grounds.

Happy young successful business woman in all black suit smiling at camera in the modern office.

A grey textured map with a flag of country inside isolated on white background .

New apartment buildings on the waterfront, in a residential development built for cleaner housing.

Figure 5: Qualitative examples of sampled image descriptions from a ViLBERT model after our pretraining tasks, but before task-specific fine-tuning.

05 | Conclusion

ViLBERT

❖ViLBERT는 이미지와 텍스트를 동시에 이해하는 인공지능 모델로서, 다양한 **시각-언어 작업**에서 뛰어난 성능을 보임

➢ 시각과 언어 사이의 상호 작용을 이해하고 이를 활용할 수 있는 인공지능의 중요성을 보여주며, 이러한 융합 모델의 중요성을 강조함

➢ *본 논문에서 image data를 embedding하는 부분의 내용이 부족하다는 단점*

❖이런 종류의 모델은 인공지능이 사람처럼 복잡한 시각적 장면을 이해하고 설명할 수 있게 만드는데 큰 역할을 할 것으로 기대함

❖FOM 프로젝트를 진행하며 이미지 캡셔닝 관련 공부를 할 수 있어 뜻 깊었음

Q & A

감사합니다