

ViLbert

pretraining task-agnostic-vision-linguistic representations
for vision-and-language-tasks

시각-언어 표현 공동처리 작업을 위한
task에 관계없이 적용 가능한 시각-언어표현 사전학습방법

Demo: <https://vilbert.cloudcv.org/>

Paper: <https://arxiv.org/abs/1908.02265>

Code: <https://github.com/facebookresearch/vilbert-multi-task>

2021. 06. 23

발표자: 이세영

목차

I. Multi-modal

II. Introduction

III. ViLBERT

i. Training Tasks

ii. Co-attention Transformer Layer

IV. Experiment

i. Training

ii. Evaluation Tasks

V. Results

VI. Conclusion

I. Multi-modal

Multi-modal

: 컴퓨터와 사용자가 Interaction(상호작용)하는 과정에서 음성, 펜, 키보드 등을 통해 정보를 주고받는 것을 말함.

Modal

: 사용자가 컴퓨터와 의사소통하는 채널. ex. 음성, 이미지, 언어 등

Multi-modal AI

: 시각, 청각 등 다양한 Modality를 동시에 받아들여 처리하는 AI

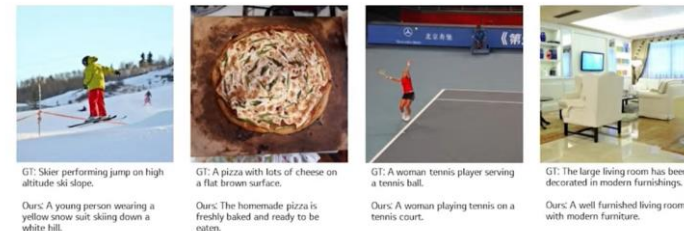


DALL-E

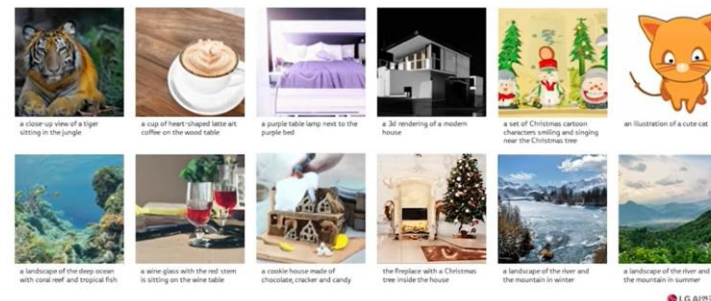
<https://openai.com/blog/dall-e/>

Image-to-Text Generation

Examples of image-to-text generation on MS-COCO with corresponding ground-truth



Text-to-Image Generation



LG EXAONE(Expert AI for everyone)

<http://www.aitimes.com/news/articleView.html?idxno=141958>

II. Introduction

Vision and Language Pre-Trained Models

Computer Vision • 24 methods



Involves models that adapt pre-training to the field of Vision-and-Language (V-L) learning and improve the performance on downstream tasks like visual question answering and visual captioning.

According to [Du et al. \(2022\)](#), information coming from the different modalities can be encoded in three ways: fusion encoder, dual encoder, and a combination of both.

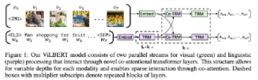
References:

- [A Survey of Vision-Language Pre-Trained Models](#)
- [Vision Language models: towards multi-modal deep learning](#)

Methods

Method	Year	Papers
 CLIP Learning Transferable Visual Models From Natural Language Supervision	2021	211
 ALIGN Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision	2021	176
 LXMERT LXMERT: Learning Cross-Modality Encoder Representations from Transformers	2019	25
 ViLBERT ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks	2019	21
 OSCAR Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks	2020	16
 VisualBERT VisualBERT: A Simple and Performant Baseline for Vision and Language	2019	14
 VL-BERT ViLBERT: Pre-training of Generic Visual-Linguistic Representations	2019	4

<https://paperswithcode.com/methods/category/vision-and-language-pre-trained-models>



ViLBert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks

[J Lu](#), [D Batra](#), [D Parikh](#), [S Lee](#) - Advances in neural ..., 2019 - [proceedings.neurips.cc](#)

... 3.1 Training **ViLBERT** To train our full **ViLBERT** model, we apply the training tasks ... We initialize the linguistic stream of our **ViLBERT** model with a BERT language model pretrained on ...

☆ 저장 77 인용 1331회 인용 관련 학술자료 전체 7개의 버전 77

Vi-bert: Pre-training of generic visual-linguistic representations

[W Su](#), [X Zhu](#), [Y Cao](#), [B Li](#), [L Lu](#), [F Wei](#), [J Dai](#) - arXiv preprint arXiv ..., 2019 - [arxiv.org](#)

... **VL-BERT** is designed to be a generic feature representation for various visual-linguistic ... simple to finetune **VL-BERT** for various downstream tasks. We simply need to feed **VLBERT** with

☆ 저장 77 인용 725회 인용 관련 학술자료 전체 6개의 버전 77

Visualbert: A simple and performant baseline for vision and language

[LH Li](#), [M Yatskar](#), [D Yin](#), [CJ Hsieh](#)... - arXiv preprint arXiv ..., 2019 - [arxiv.org](#)

... **VisualBERT**, a simple and flexible framework for modeling a broad range of vision-and-language tasks. **VisualBERT** ... objectives for pre-training **VisualBERT** on image caption data. ...

☆ 저장 77 인용 597회 인용 관련 학술자료 전체 4개의 버전 77

Lxmert: Learning cross-modality encoder representations from transformers

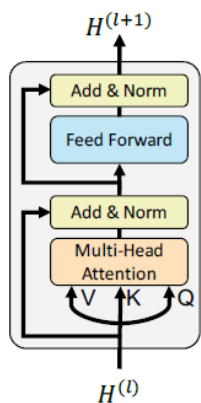
[H Tan](#), [M Bansal](#) - arXiv preprint arXiv:1908.07490, 2019 - [arxiv.org](#)

... -masked words in the language modality, **LXMERT**, with its cross-modality model architecture, ... We also show that loading BERT parameters into **LXMERT** will do harm to the pre-training ...

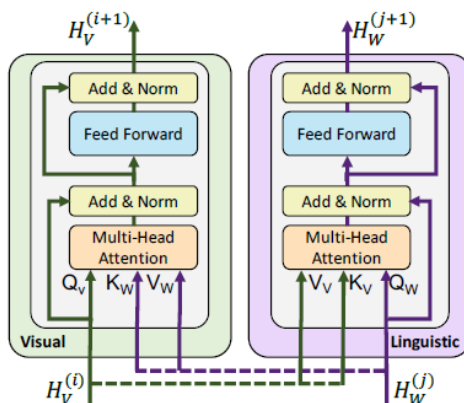
☆ 저장 77 인용 951회 인용 관련 학술자료 전체 8개의 버전 77

II. Introduction

- ◆ Vision-Language task는 대부분 pretrained large vision model + pretrained large language model
 - Vision-Language 정보를 따로 학습함.
 - 학습한 Vision-Language 정보가 관련이 없는 경우가 있음.
- ◆ Vision-Language 정보의 connection을 학습하고, vision-language와 관련된 다양한 task에 광범위하게 활용할 수 있는 vision-language 공통 처리 모델 개발을 하려고 함.
 - vision, language 라는 서로 다른 Modality 처리 가능
 - Modality 간의 상호작용 제공
 - ViLBERT모델이 Single-stream을 통합한 모델보다 성능이 좋았음.



(a) Standard encoder transformer block



(b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

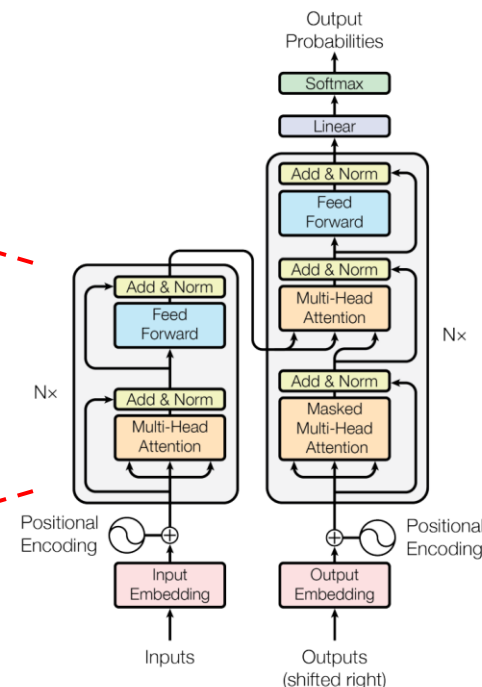


Figure 1: The Transformer - model architecture.

Attention is All You Need

III. ViLBERT

◆ ViLBERT: Vision & Language BERT

◆ 동작 원리

- Data representation, training Tasks and objectives
- Co-Attentional Transformer Layers

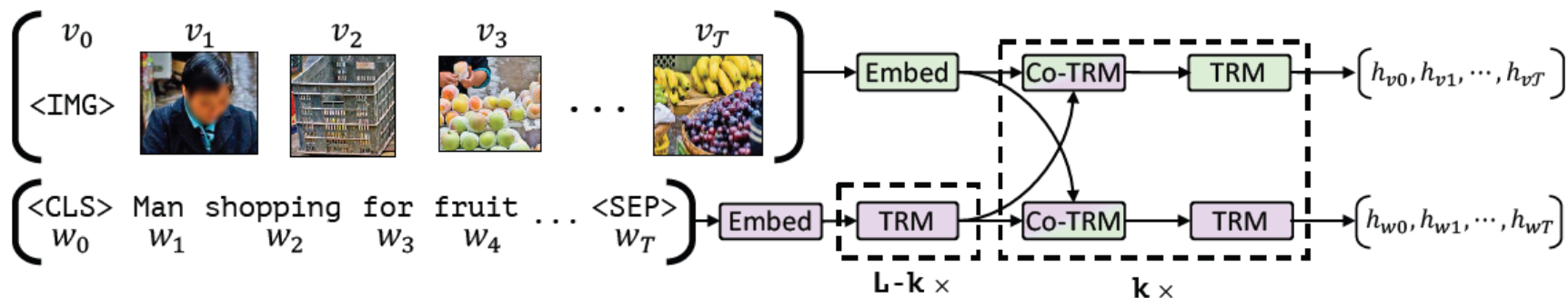


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

Image vector \rightarrow embedding \rightarrow Co-TRM \rightarrow TRM \rightarrow hidden img vector

Word vector \rightarrow embedding \rightarrow Transformer \rightarrow Co-TRM \rightarrow TRM \rightarrow hidden word vector

III. ViLBERT

◆ Data representation

❖ Text

- BERT와 동일.

someone is checking out produce at a supermarket.

-> [CLS], someone, is, checking, out, produce, at, a supermarket, [SEP]

-> [101, 1800, 1110, 9444, 1149, 3133, 1120, 170, 20247, 119, 102]

input_ids,

token_type_ids,

position_ids=position_ids

❖ Image



mean-pooled visual features with a spatial encoding corresponding to the entire image

전체이미지에 대한 위치 인코딩을 가진 mean-pooled visual features.

각 이미지 feature의 공간위치(상단왼쪽, 하단오른쪽)영역으로 부터 5차원 벡터 구성.

두 feature vectors를 concat
{IMG, v1, ..., vT, CLS, w1, ..., wT, SEP}

III. ViLBERT

◆ Training tasks and objectives

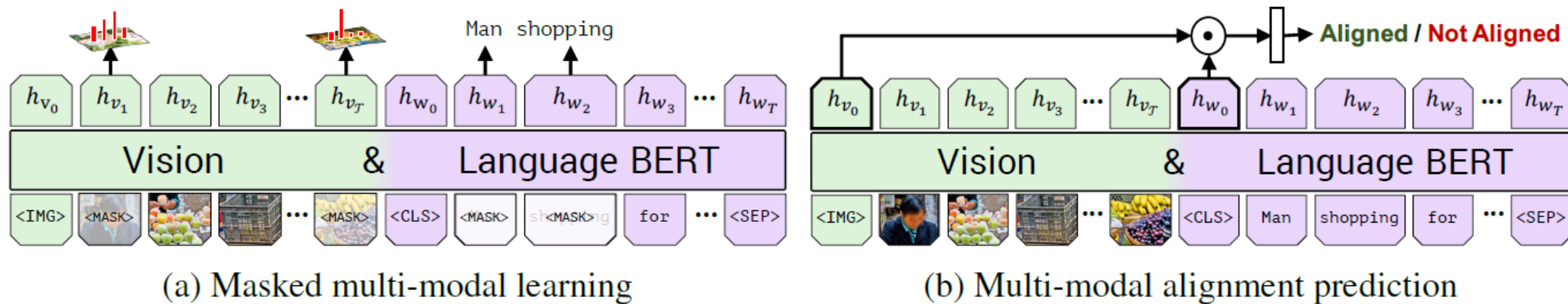


Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

- Masked Multi-modal modeling

: feature의 15%를 masking, 90%확률로 0으로 만듦. Masking된 부분을 복원하며 학습
이미지의 경우, masking된 이미지 영역에 대한 시맨틱 클래스에 대한 분포예측
=> 원래 영역에 대한 클래스 출력분포와 모델이 예측한 분포를 유사하게 하기위해 KLD loss로 학습

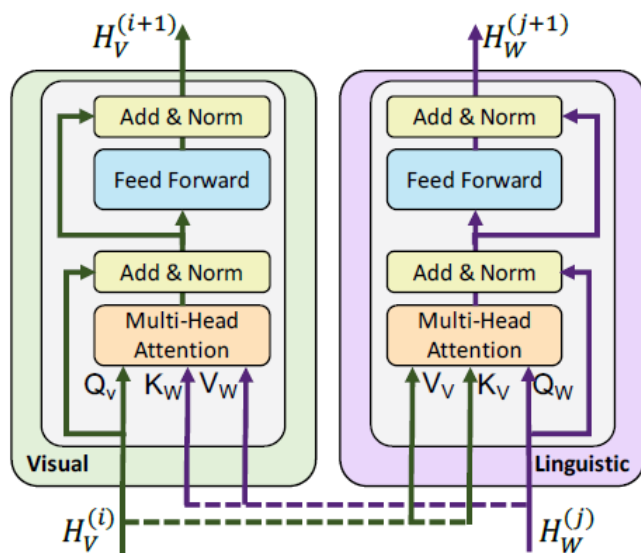
$$KLD\ loss = y_{true} * \log \frac{y_{true}}{y_{pred}}$$

- Multi-modal alignment prediction

: 이미지와 텍스트가 연관성이 있는 데이터인지(Aligned) 아닌지(Not aligned) 학습

III. ViLBERT

◆ Co-Attentional Transformer Layers



(b) Our co-attention transformer layer

- $H_v^{(i)}, H_w^{(j)}$ 이 주어지면 모듈은 일반적인 transformer 처럼 Q, K, V matrix 를 계산한다.
- K, V 는 자신과 다른 모달리티(V- \rightarrow L, L- \rightarrow V)의 Multi-Head Attention 블록에 입력으로 전달된다.
- 결과적으로 각 attention 블록은 다른 방식으로 conditioned 된 각각의 양식에 대해 attention-pooling된 feature($H_v^{(i+1)}, H_w^{(j+1)}$) 를 생성한다.
- Transformer 블록의 나머지 부분은 BERT 모델과 같이 진행된다.

IV. Experiment

◆ Training

- Dataset: Conceptual Captions dataset 330만 개의 데이터 중 어색한 문장, 설명이 너무 짧은 문장은 제외하고 310만개의 데이터 사용



by Joi Ito

the trail climbs steadily uphill most of the way.



by Danail Nachev

the stars in the night sky.



by Justin Higuchi

musical artist performs on stage during festival.



by Viaggio Routard

popular food market showing the traditional foods from the country.

- pretrained BERT-BASE로 ViLBERT의 **Linguistic Stream** 을 initialization 함.
 - hidden_states: 762, Layers: 12, Multi-head Attention: 12
 - 훈련 시간 때문에 BERT-BASE를 사용함.
 - BERT-LARGE를 사용하면 성능 향상을 기대할 수 있다.
- Visual Stream** 은 Visual Genome 데이터셋에서 pretrained Faster R-CNN 사용하여 이미지 영역 feature 추출.
 - 추출된 영역 feature 중 클래스 확률이 threshold 를 초과하는 영역을 선택.

IV. Experiment

◆ Training

- Visual Stream의 Transformer 및 Co-attentional Transformer
 - Hidden_state: 1024, attention-head: 8
- Training details
 - 10 epochs
 - batch_size: 512
 - 8개의 Titan X Gpu
 - 초기 learning rate가 $1e-4$ 인 Adam optimizer 사용
 - linear decay learning rate schdule with warm up 사용

IV. Experiment

◆ Evaluation Tasks

❖ VQA(Visual Question Answering)

- COCO 이미지에 대한 110만개 QnA 데이터셋에 대해 훈련, 평가

❖ VCR(Visual Commonsense Reasoning)

- 모델에는 이미지, 개체, 질문 및 네 가지 답변 선택이 제공됨. 모델은 어떤 답이 옳은지 결정.
- 그런 다음 네 가지 이론적 선택이 주어지며 그 중 어떤 것이 답이 옳은지를 설명하는 가장 좋은 이유를 결정해야 합니다.

❖ Grounding Referring Expressions.

- 제공된 설명 및 이미지에 해당하는 인스턴스 주위에 bounding box 배치

❖ Caption-Based Image Retrieval.

- 제공된 캡션에 해당하는 이미지를 검색하는 작업



IV. Experiment

◆ Evaluation Tasks

❖ VQA(Visual Question Answering)

- COCO 이미지에 대한 110만개 QnA 데이터셋에 대해 훈련, 평가

```
class SimpleClassifier(nn.Module):  
    def __init__(self, in_dim, hid_dim, out_dim, dropout):  
        super().__init__()  
        self.logit_fc = nn.Sequential(  
            nn.Linear(in_dim, hid_dim),  
            GeLU(),  
            BertLayerNorm(hid_dim, eps=1e-12),  
            nn.Linear(hid_dim, out_dim),  
        )  
  
    def forward(self, hidden_states):  
        return self.logit_fc(hidden_states)
```

<https://github.com/facebookresearch/vilbert-multi-task/blob/main/vilbert/vilbert.py>



이 제공됨. 모델은 어떤 답이 옳은지 결정.
어떤 것이 답이 옳은지를 설명하는 가장 좋은 이유

bounding box배치



V. Results

◆ results

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)	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-
	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-
	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	-	-	-	-	-
	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-
	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
	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

Baselines

VQA:DFAF, VCR: R2C, RefCOCO+: MAttNet, ImageRetrieval: SCAN

Single-Stream+, ViLBERT+: Non-pretrained

Results

- 단일 스트림 모델에 비해 성능을 향상
- 우리의 사전 훈련 작업은 시각 언어 표현을 향상

V. Results

◆ Effect of Visual Stream Depth.

- Image Retrieval task 에서 Co-TRM 의 깊이가 깊을 수록 성능이 향상되는 것을 확인함.
- 반면 VCR, RefCOCO+의 경우 얇은 모델의 성능이 더 좋았음.

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

V. Results

◆ Benefits of Large Training Sets.

- 데이터를 25%, 50%, 100%로 무작위로 나누어 성능확인
- 사전 학습된 데이터가 많을 수록 좋은 성능을 보임.

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 (0 %)	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
ViLBERT (25 %)	69.82	71.61	73.00	52.66	69.90	76.83	60.99	53.08	80.80	88.52	20.40	48.54	62.06
ViLBERT (50 %)	70.30	71.88	73.60	53.03	71.16	77.35	61.57	54.84	83.62	90.10	26.76	56.26	68.80
ViLBERT (100 %)	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

V. Results

◆ What does ViLBERT learn during pretraining?

- 데이터를 0%, 25%, 50%, 100%로 무작위로 나누어 성능확인
- 사전 학습된 데이터가 많을 수록 좋은 성능을 보임.

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)	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-
	MattNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-
	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	-	-	-	-	-
	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-
	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
	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

- Zero-shot 성능은(표 1, 오른쪽) 미세 조정된 모델(Z:31.86 vs F: 58.20, R1)보다 현저히 낮음. (이전 SOTA의 경우 Z:31.86 vs F:48.60, R1)
- 하지만, Flickr30k 이미지 또는 캡션을 보지 않고도 합리적으로 수행하므로 ViLBERT가 사전 교육 동안 시각과 언어 사이의 의미 있는 connection 을 학습했음을 나타낸다.

VI. Conclusion

◆ Contributions

- ❖ Visual features 와 Linguistic features 를 동시에 처리할 수 있는 모델 제안
- ❖ 새로운 사전학습 방법 제안
- ❖ BERT에 간단한 변형으로 성능이 향상되었다는 것을 실험을 통해 증명
- ❖ 다양한 Task에 적용되는 모델
- ❖ 기본 모델에 Classifier 만 더하는 간단한 변형으로 다양한 Task 에 적용될 수 있음.

Thank You