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

LU, Jiasen, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 2019, 32

경영과학연구실 전재현

2024. 02. 05

Why this Paper?

- Paper about processing multi-modal(image-text) input.
- Also about making the base-model like Scheduler-GPT.

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

J Lu, D Batra, D Parikh, S Lee

Advances in neural information processing systems, 2019 · proceedings.neurips.cc

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-

자세히 보기 ⌄

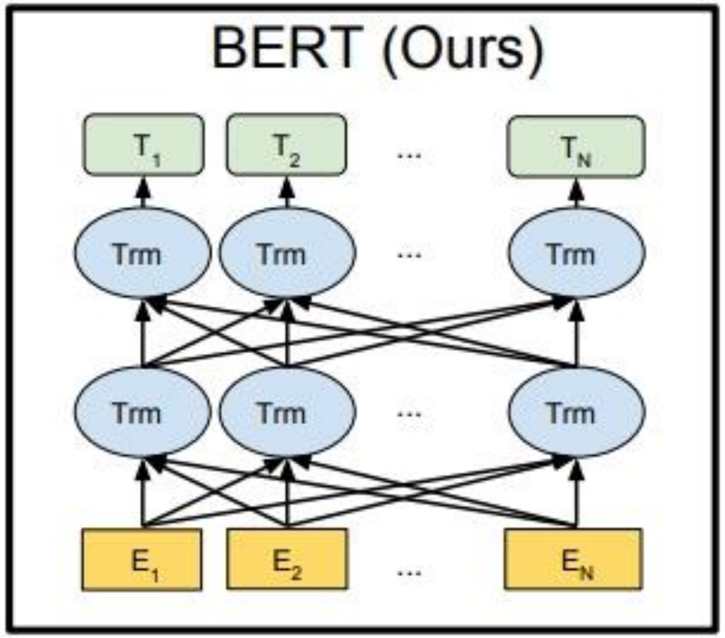
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

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

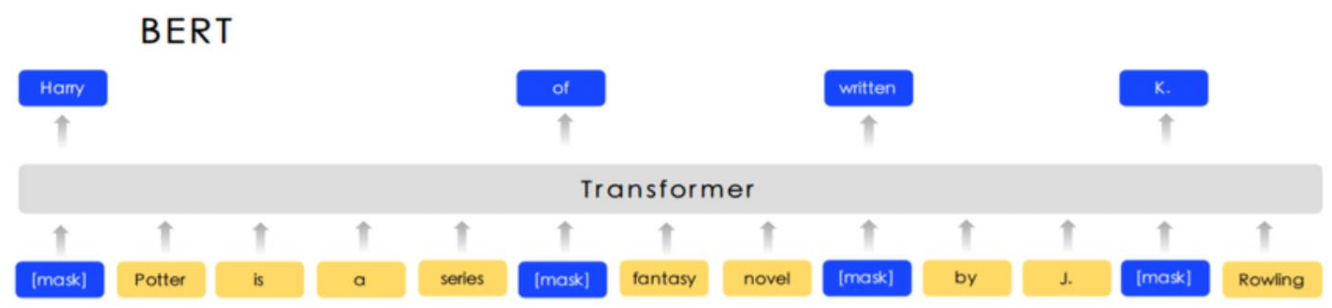
What is BERT?

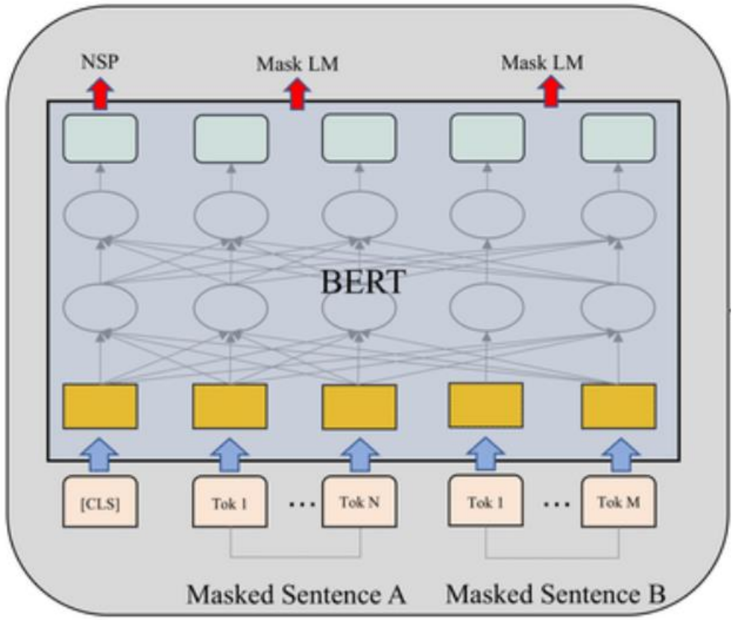
- BERT stands for Bidirectional Encoder Representations from Transformers
- Pretrained on extensive datasets such as Wikipedia and BooksCorpus, BERT utilizes unlabeled data to develop a versatile base model.
- Only from task-specific fine-tuning, BERT reached peak performance levels on variety tasks.



BERT's Pretraining Method

- BERT's pretraining encompasses two distinct phases with unlabeled data.
  - Masked Language Model : randomly hides 15% of input tokens, prompting the model to infer the masked words within its training context.
  - Next Sentence prediction : requires the model to ascertain if two sentences are sequentially connected.





Creating a robust base model capable of processing multi-modal inputs that combine vision and language data effectively

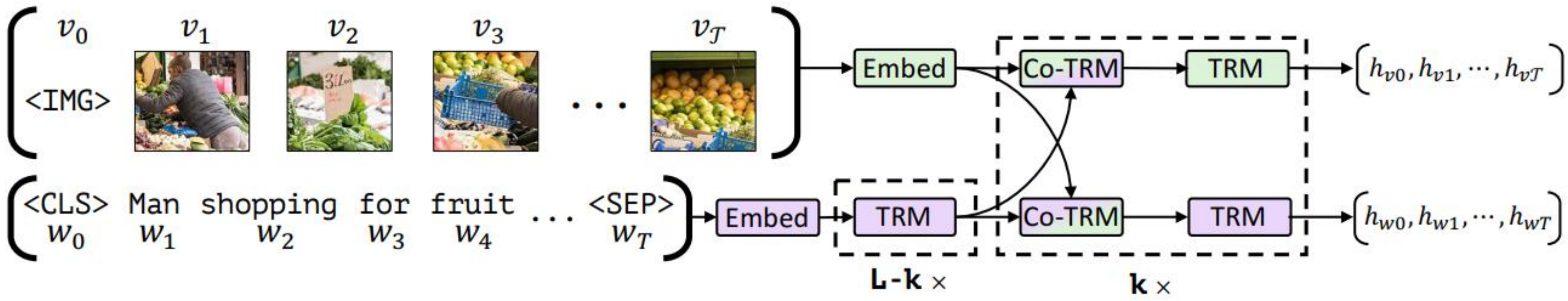
**Application of a Co-attention mechanism
within the transformer layers,
which processes the keys and values of
vision and text modalities interchangeably**

Related Works

Published year	Author	Paper
2021	Alec Radford et al. (Open AI)	Learning Transferable Visual Models From Natural Language Supervision
2020	Weijie su et al. (University of Science and Technology of China)	VL-BERT: Pretraining of generic visual linguistic representations
2019	Liunian Harold Li et al. (University of California)	VisualBERT: A Simple and performant baseline for Vision and language

The ViLBERT Model Architecture

- Expands upon BERT to concurrently represent visual and textual information.
- Comprises two parallel processing streams: **visual** and **linguistic**.
- Image : Utilizes cropped images defined by bounding boxes.
  Image features are extracted using Faster R-CNN built on ResNet-101.
  Each selected region i, vi is defined as the mean-pooled convolution feature.
- Text : Leverages the $\mathrm{BERT}_{\mathrm{BASE}}$ model for linguistic processing.

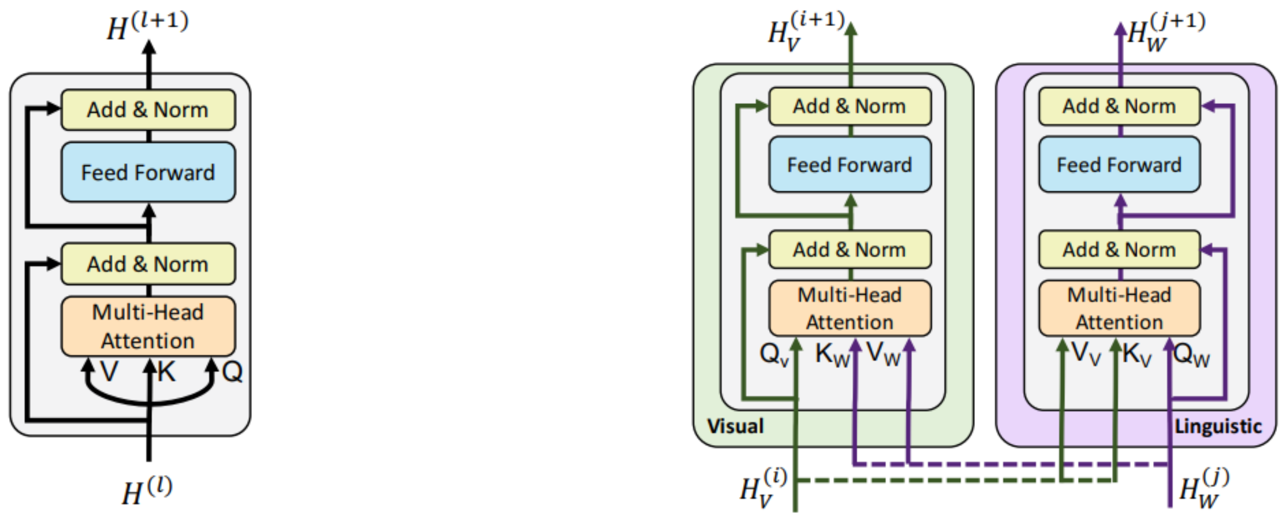


The Co-attention Transformer Layer

- Exchanges key-value pairs of two stream(Image ↔ language)
- Enables vision-attended language features to be incorporated into visual representations, and likewise for linguistic elements.

Image-conditioned language attention

Language -conditioned image attention

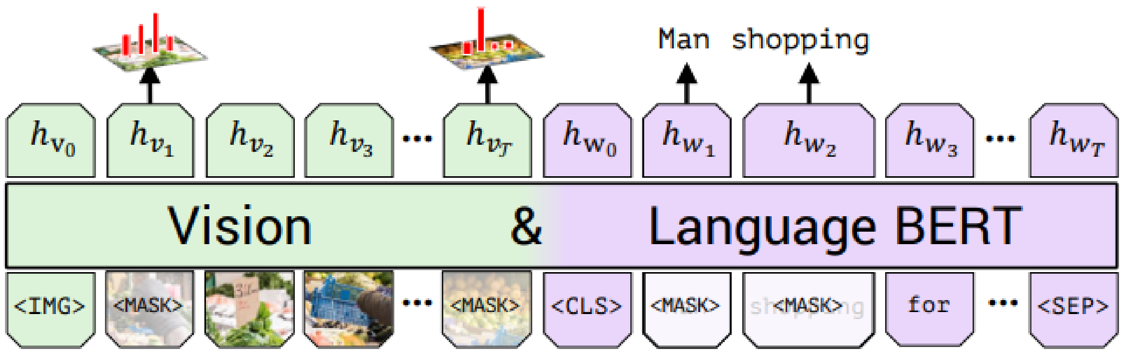


(a) Standard encoder transformer block

(b) Our co-attention transformer layer

Pretraining - Mask multi-modal learning

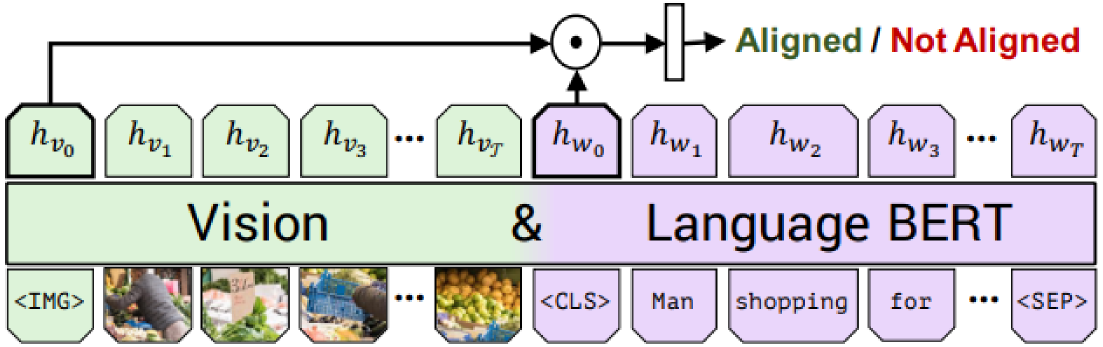
- During pretraining, 15% of the input from both images and language streams is masked, and the model learns to predict the masked portions.
- For text the pretraining method follows the conventional approach used by BERT.
- For image predicts the distribution over semantic classes for each corresponding image region. Aims to minimize the KL divergence between the predicted and true distributions.
- Trains the model to infer textual context through visual cues and vice versa.



(a) Masked multi-modal learning

Pretraining - Multi-modal alignment prediction

- Trains the model to predict whether text descriptions align accurately with the interpreted images.
- Engages in binary prediction training to determine if holistic representations, such as h_{v_0} and h_{w_0} correspond with each other.
- Through this method, learns to discern the relational dynamics between each image and its associated text.



(b) Multi-modal alignment prediction

Experiment Settings

- The pretrained ViLBERT model was fine-tuned across four distinct tasks.
- 4 tasks are as follows:
  - VQA : Answering questions based on a given image
  - VCR : Answering questions with a commonsense explanation based on visual cues(Q→A, QA→R, Q→AR)
  - Referring Expressions : Localizing an image region given a natural language reference.
  - Caption-Based Image Retrieval : Searching for the most relevant image from a given pool based on textual descriptions

Visual Question Answering **Visual Commonsense Reasoning** **(+ Zero-shot)**

Baselines

- Baselines
  - Single-Stream : One stream architecture without dividing image and text
  - Single-Stream^{+} : single-stream without pretraining
  - ViLBERT^{+} : ViLBERT without pretraining

- Task-Specific Baselines

Task	**Baselines**
VQA	DFAF
VCR	R2C
Referring Expressions (RefCOCO+)	MAttNet
Caption-based image retrieval	SCAN

Comparison against other algorithms

- Compares the recent state-of-the-art (SOTA) with ViLBERT, leveraging transfer learning across four distinct tasks.
- ViLBERT demonstrates superior performance across all tasks evaluated.
- The results underscore the effectiveness of a robust base model trained on vision-text operations, outperforming models specialized in individual tasks.

		VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
	Method	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	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**

The Impact of [Co-TRM → TRM] Blocks on Performance

- The optimal number of [Co-TRM → TRM] blocks varies across different tasks.
- Increasing the number of layers does not necessarily correlate with better performance.
- A higher count of [Co-TRM → TRM] implies more extensive context aggregation, suggesting that tasks involving a greater computational fusion of text and vision features tend to benefit from additional blocks.

	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
Method	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**

The Impact of Pretraining Dataset Size

- The dataset size was varied during the pretraining phase.
- An increase in the pretraining dataset size correlates with improved results post-finetuning.
- Implies that learning diverse relationships between images and text during pretraining positively impacts performance when transferring knowledge to different tasks.

	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
Method	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**

Examples of Image Descriptions from Pretrained ViLBERT

- Some examples of image descriptions from a VilBERT without task-specific finetuing.
- Without task-specific fine-tuning, the model can already utilize its pretrained knowledge to generate descriptions that are relevant to the images.
- ViLBERT's advantage in pretraining lies in its ability to leverage both text and image modalities to enhance the understanding of content.

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.

Conclusions

- Utilization of co-attention mechanisms allows it to excel by learning joint representations of visual and textual information, outperforming models that are narrowly focused on single-modality tasks.

- The incorporation of co-attention layers enables ViLBERT to effectively fuse and leverage multimodal features.

- Models trained on diverse datasets that encourage a broader contextual understanding show superior performance.

Q & A