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

Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee

Problem Statement

We have separate visual understandings and linguistic understanding, but there are no models that can actually relate to them.

It is of no use if we have a perfect visual representation of dog breeds if a downstream vision-and-language model fails to associate it with appropriate phrases like “beagle” or “shepherd”

Key points/takeaways

Introduces separate streams for vision and language processing that communicate through **co-attentional transformer** layers.

Can accommodate the differing processing needs of each modality and provides interaction between modalities at varying representation depths.

They demonstrate that this structure outperforms a single-stream unified model in our experiments.

Input representation

Text: BERT operates over sequences of discrete tokens comprised of vocabulary words and a small set of special tokens: SEP, CLS, and MASK. For a given token, the input representation is a sum of a token-specific learned embedding and encodings for position and segment.

Image: Extract bounding boxes and their visual features from a pre-trained object detection network. Encode spatial location instead, constructing a 5-d vector from region position (normalized top-left and bottom-right coordinates) and the fraction of image area covered.

Pretraining ViLBERT

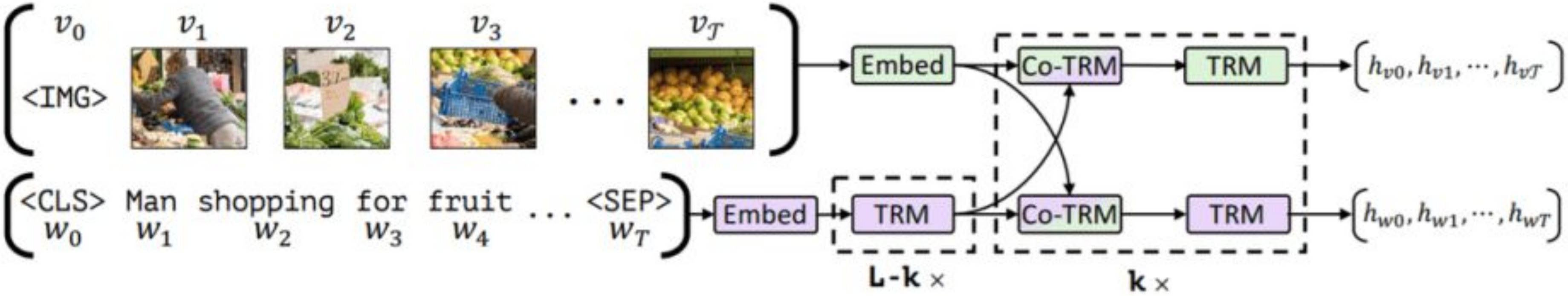
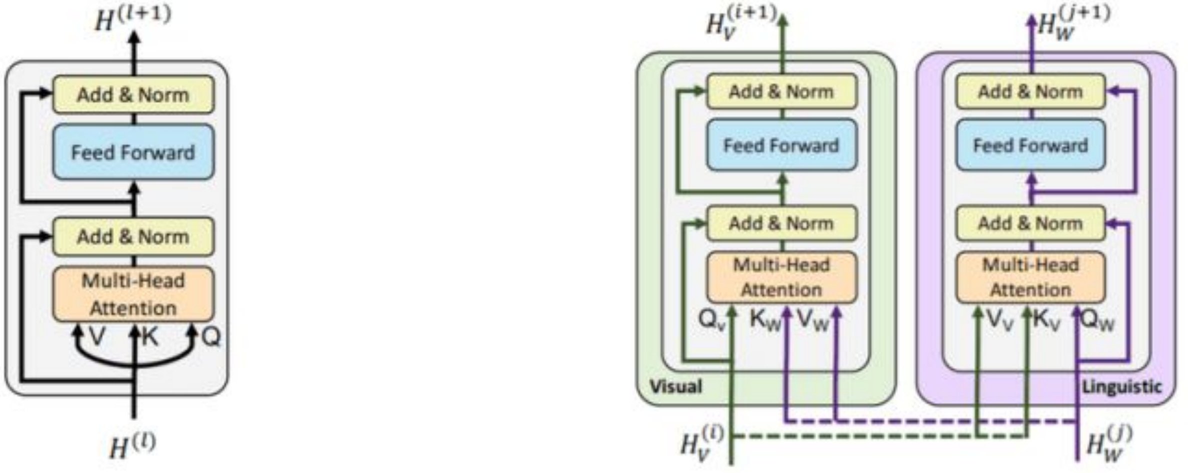

v_0
v_1
v_2
v_3
$v_\mathcal{T}$
<IMG>
...
<CLS> Man shopping for fruit ... <SEP>
w_0 w_1 w_2 w_3 w_4 w_T
Embed
Co-TRM
TRM
Embed
TRM
Co-TRM
TRM
$h_{v0}, h_{v1}, \cdots, h_{v\mathcal{T}}$
$h_{w0}, h_{w1}, \cdots, h_{wT}$
L-k ×
k ×

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.

Architecture



(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).

Co-attentional transformer layer

Query, key, and value matrices as in a standard transformer block

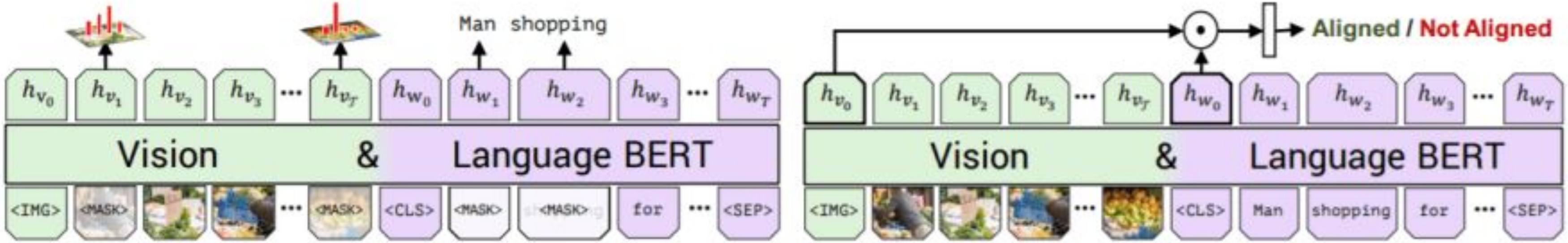
However, the keys and values from each modality are passed as input to the other modality’s multi-headed attention block.

The attention block produces attention-pooled features for each modality conditioned on the other – in effect performing image-conditioned language attention in the visual stream and language-conditioned image attention in the linguistic stream.

Analogy with BERT (pre-training)

1. Predicting the masked word
2. Check if the second sentence is the next sentence wrt the first one

1. Predicting the semantics of masked words and image regions given unmasked input
2. Check if image and sentence correspond to each other.


Man shopping
Aligned / Not Aligned
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ... h_{v_T} h_{w_0} h_{w_1} h_{w_2} h_{w_3} ... h_{w_T}
Vision & Language BERT
<IMG> <MASK> ... <MASK> <CLS> <MASK> <MASK> for ... <SEP>
Vision & Language BERT
<IMG> ... <CLS> Man shopping for ... <SEP>

(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

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.

Pretraining

The masked multi-modal modelling task- masking approximately 15% of both words and image region inputs and tasking the model with reconstructing them given the remaining inputs.

Rather than directly regressing the masked feature values, the model instead predicts a distribution over semantic classes for the corresponding image region. To supervise this, we take the output distribution for the region from the same pretrained detection model used in feature extraction. We train the model to minimize the KL divergence between these two distributions.

In the multi-modal alignment task {IMG, v1, . . . , v_T , CLS, w1, . . . , w_T , SEP} and must predict whether the image and text are aligned, i.e. whether the text describes the image.

An element-wise product between h_IMG and h_CLS and learn a linear layer to make the binary prediction whether the image and text are aligned.

Results

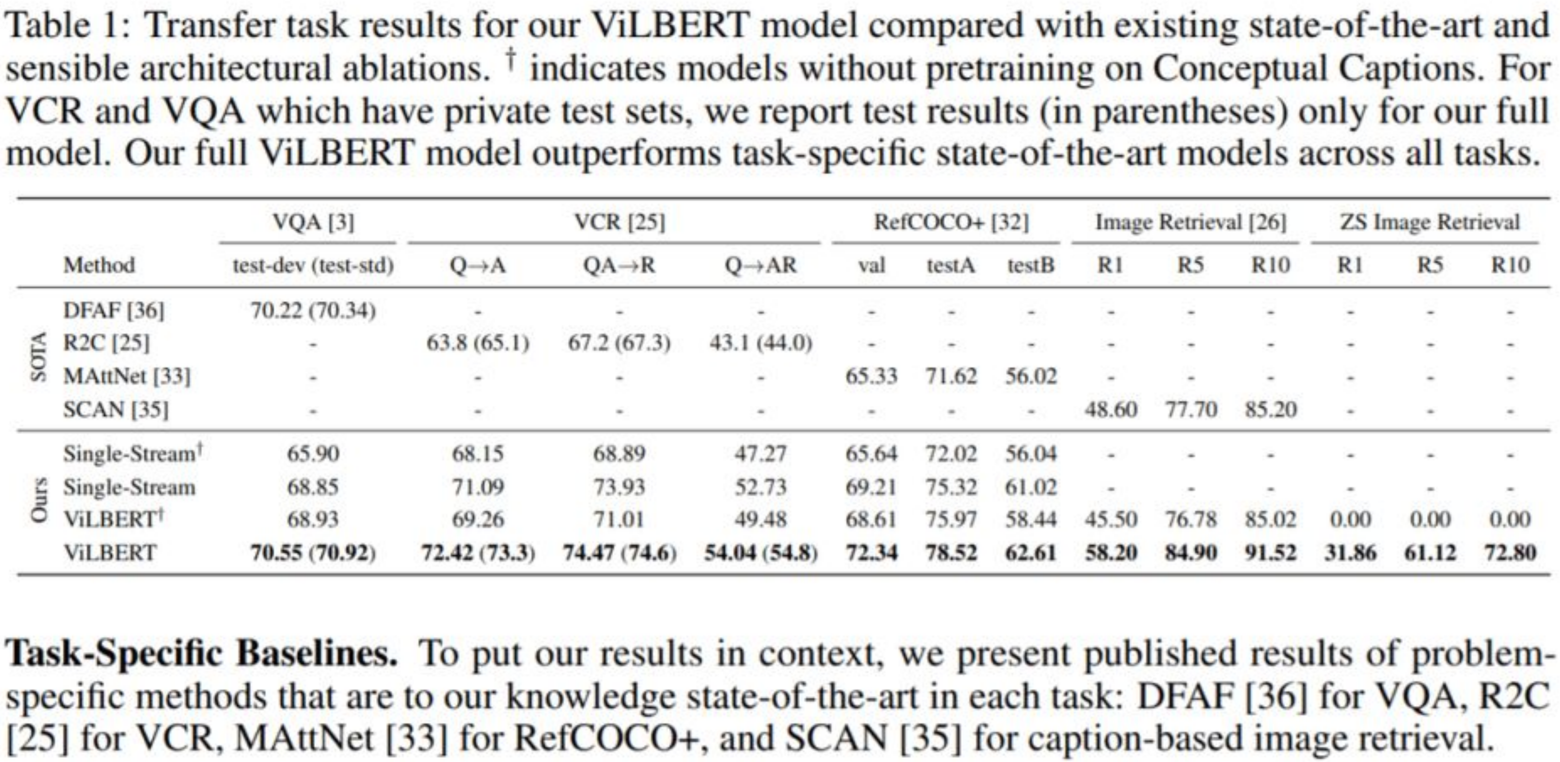


Table 1: Transfer task results for our ViLBERT model compared with existing state-of-the-art and sensible architectural ablations. † indicates models without pretraining on Conceptual Captions. For VCR and VQA which have private test sets, we report test results (in parentheses) only for our full model. Our full ViLBERT model outperforms task-specific state-of-the-art models across all 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**

Task-Specific Baselines. To put our results in context, we present published results of problem-specific methods that are to our knowledge state-of-the-art in each task: DFAF [36] for VQA, R2C [25] for VCR, MAttNet [33] for RefCOCO+, and SCAN [35] for caption-based image retrieval.