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

Jiasen Lu Dhruv Batra Devi Parikh Stefan Lee
Georgia Institute of Technology
Oregon State University
Facebook AI Research

NeuralPS '19

Noah Lee (noahlee357@korea.ac.kr)
Donghui Im (ehdgnl101@korea.ac.kr)
Data Intelligence Lab, Korea University
2021 02. 19.

01 Problem Statement

How can we learn task-agnostic joint representations of vision and language?

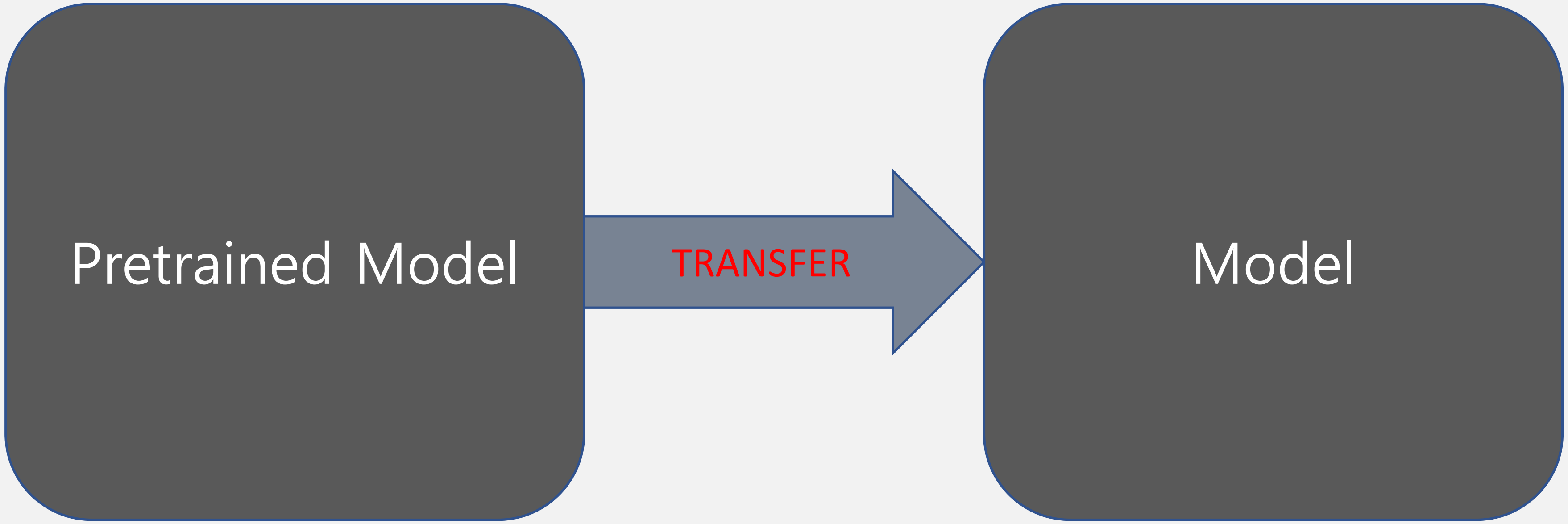
02 Motivation

- Common need to align natural language and visual stimuli
- The need for **visual grounding**

Extract visual contents to distinguish !!

02 Motivation

- Widespread usage of pretrain-then-transfer – performing *proxy tasks*
- Recent success with self-supervised learning ELMo, BERT, GPT



02 Motivation

"Shepherd"
"Beagle"

Achieve joint representation of image content and natural language

BUT is visual grounding pretrainable & transferable?

03 Contribution

1) Proposes a joint model for learning task-agnostic **visual grounding** to jointly reason about text and images

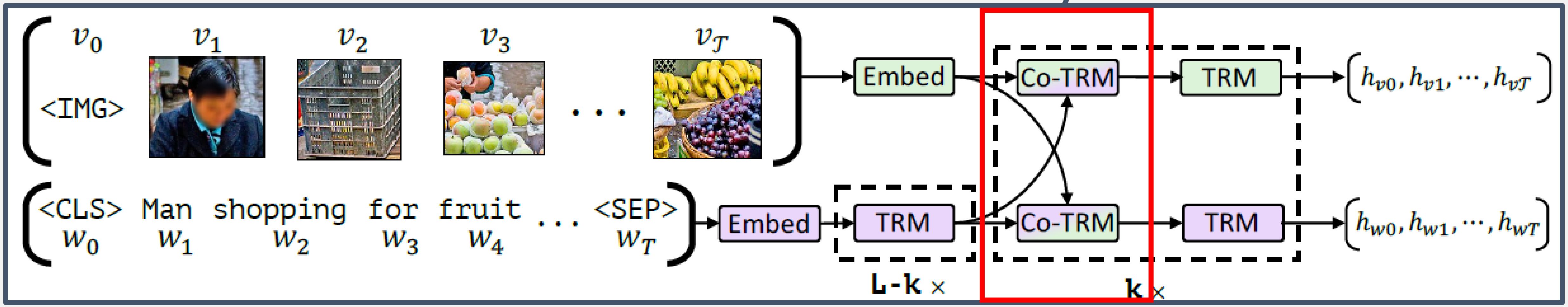
2) Utilizes **co-attention transformer layers** by introducing separate streams for vision and language processing

3) Outperforms four established vision-and-language tasks

06 Model

Model overview

Visual (green)

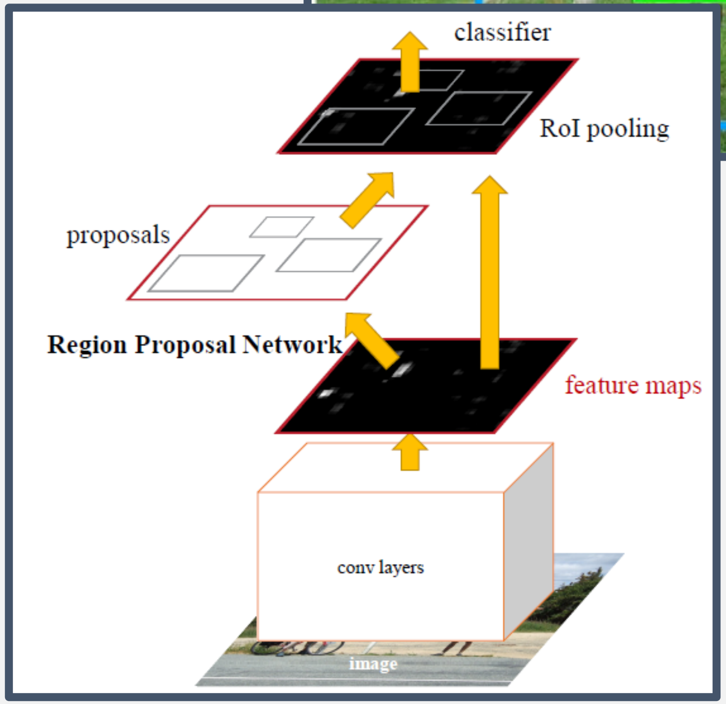


Linguistic (purple)

03 Model: Visual Representation

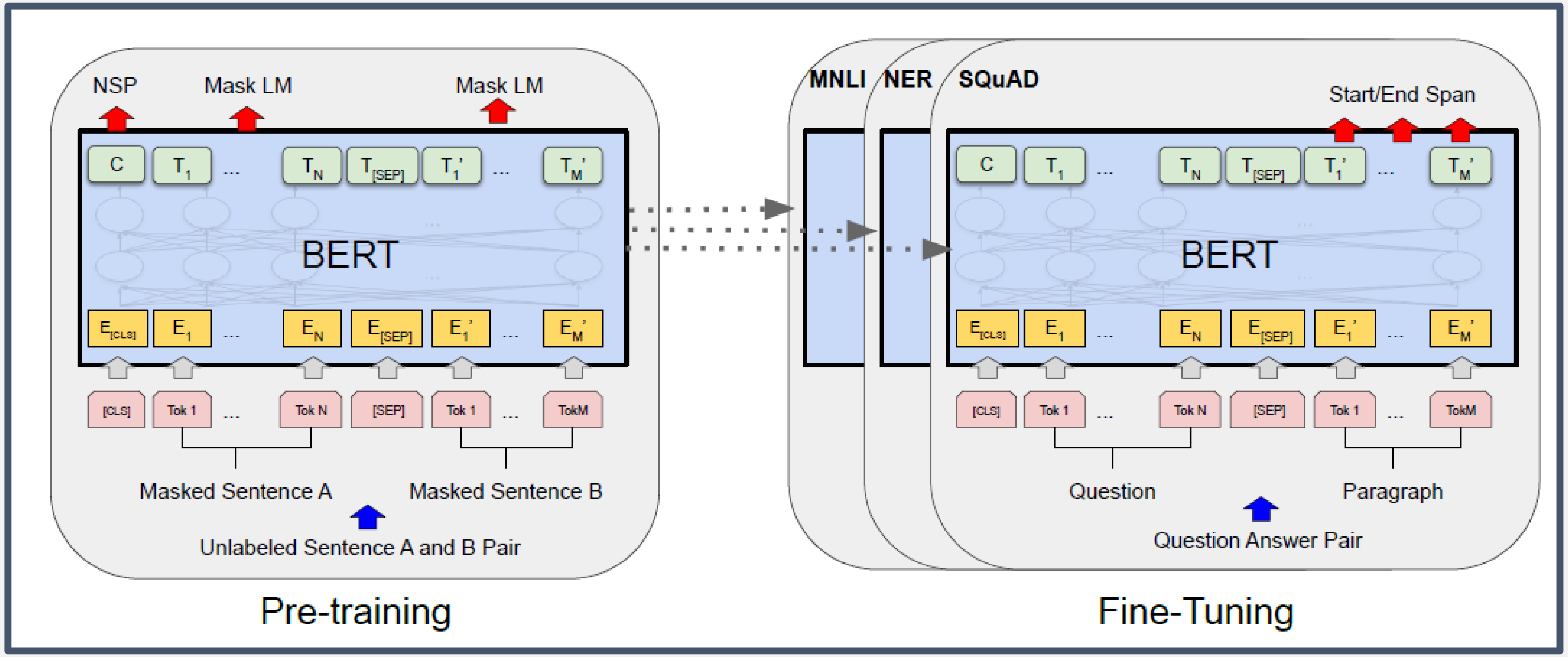
Faster R-CNN

- w/ ResNet-101 backbone pretrained on Visual Genome
- Encode spatial location (5-d vector)
- Extract bounding boxes & visual features (10~35 high scoring)



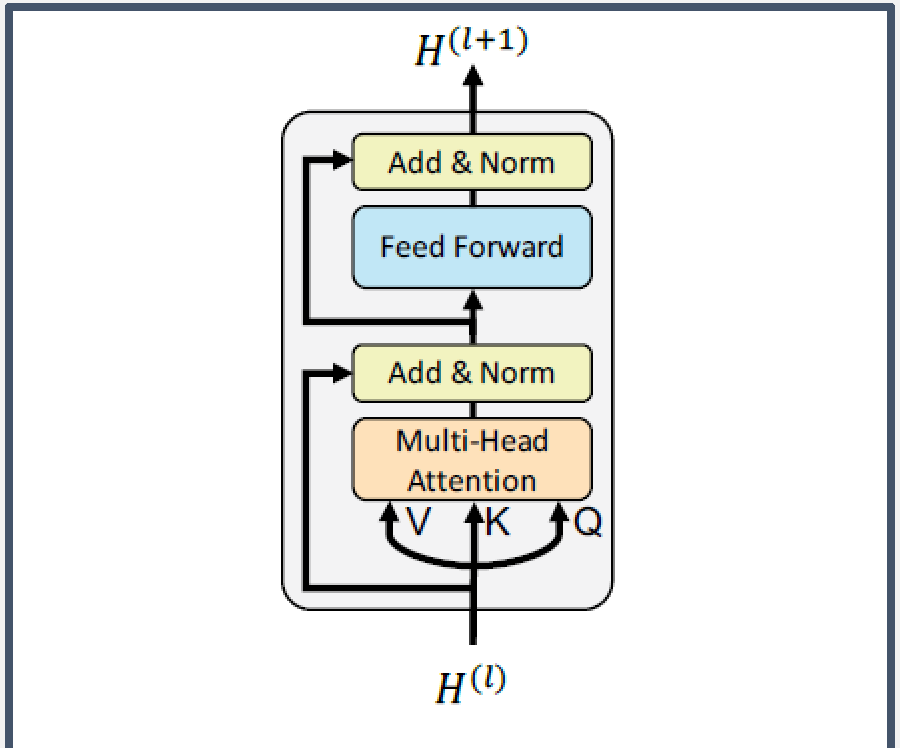
03 Model: Linguistic Representation

BERT

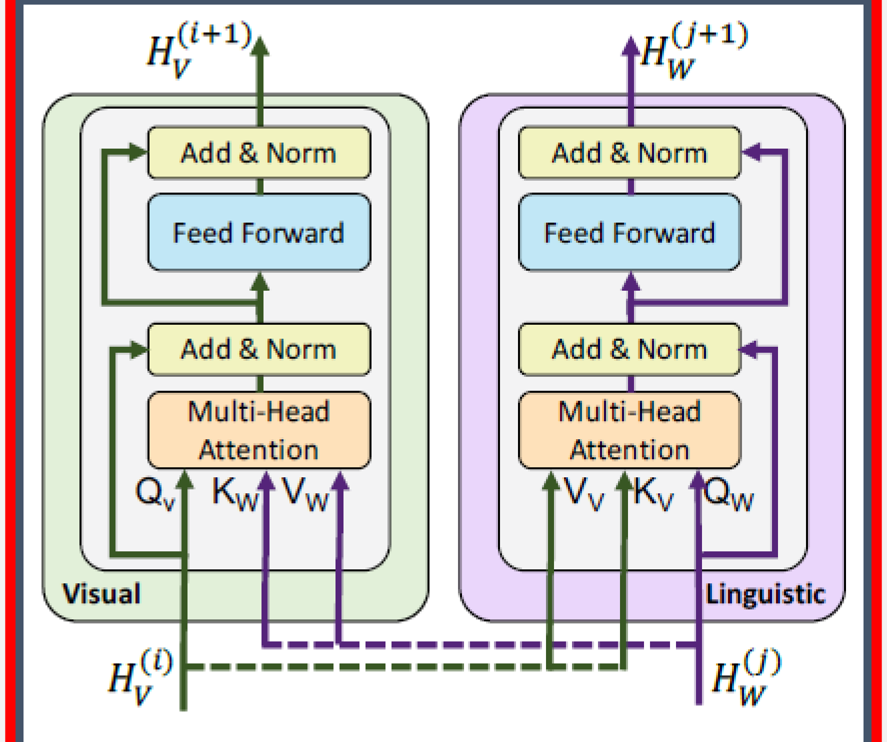


Pretrained on BookCorpus & Wikipedia

Co-attention



(a) Standard encoder transformer block



(b) Our co-attention transformer layer

Exchanging key-value pairs in multi-head attentions

03 Model: Pretraining Dataset

Conceptual Captions

- 3.3 million images
- Weakly-associated descriptive captions
- Trained on two proxy tasks

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

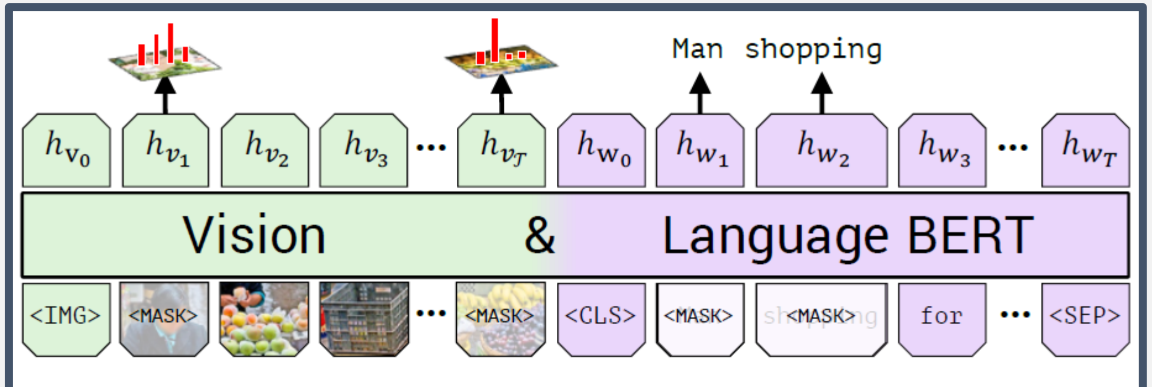
Conceptual Captions: pop artist performs at the festival in a city.

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

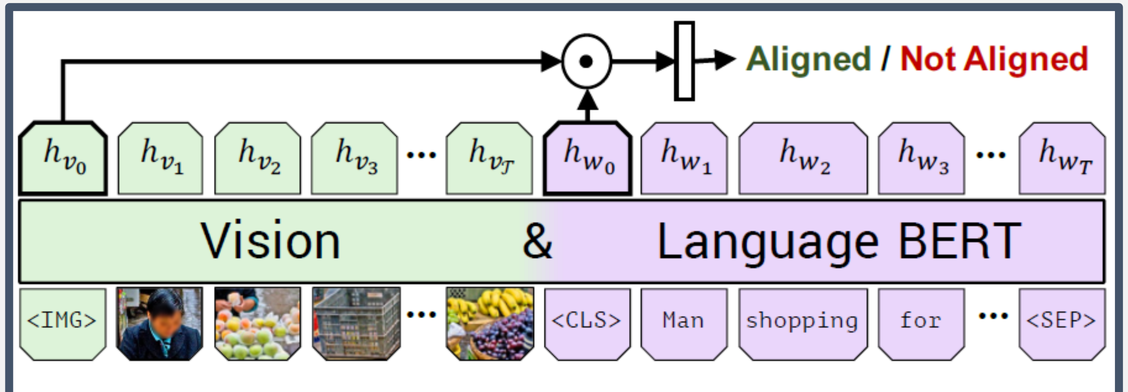
Conceptual Captions: a worker helps to clear the debris.

06 Model: Pre-training

Two pretraining tasks



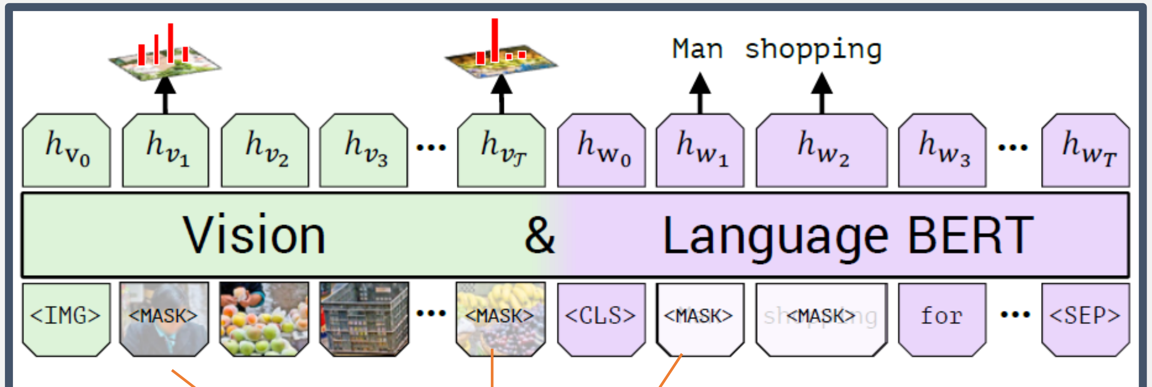
(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

06 Model: Pre-training

Masked multi-model learning



(a) Masked multi-modal learning

Mask 15% from both stream

[Language]

- Setting identical to BERT
- Masked text input
  - 80% : <mask>
  - 10% : random
  - 10% : unaltered

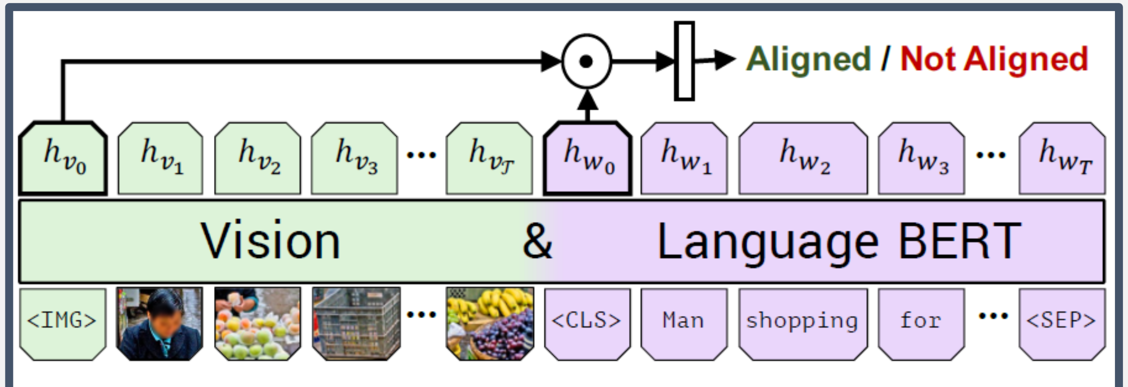
[Vision]

- Masked input
  - 90%: zeroed out
  - 10%: unaltered

06 Model: Pre-training

Multi-modal alignment prediction

- Predict whether image and text are aligned (binary prediction)
- Negatives generated by randomly replacing either the image or caption



(b) Multi-modal alignment prediction

05 Dataset/Code

[Dataset] – all publicly available

Visual Question Answering: VQA 2.0
https://visualqa.org/

Visual Commonsense Reasoning: VCR
https://visualcommonsense.com/

Grounding Referring Expressions: RefCOCO+
http://tamaraberg.com/referitgame/

Caption-Based Image Retrieval, Zero-Shot Caption-Based Image Retrieval: Flickr30k
http://nlp.cs.illinois.edu/

[Code] – publicly available

https://github.com/facebookresearch/vilbert-multi-task

07 Settings

[Embedding]
- Visual: Faster R-CNN w/ ResNet-101 backbone
- Linguistic: $\text{BERT}_{\text{BASE}}$

[Hyperparameter]
- Batch size: 512
- Epochs: 10
- Optimizer: Adam
- Learning rate: 1e-4
- Learning rate decay: scheduled linear decay

08 Baselines

[Ablative]

- Single-Stream: single BERT
- Single-Stream†: single stream w/o pretraining
- ViLBERT†: VilBERT w/o pretraining

[Task-Specific]

- Visual Question Answering - DFAF
- Visual Commonsense Reasoning – R2C
- Grounding Referring Expressions - MAttNet
- Caption-Based Image Retrieval: SCAN
- Zero-shot Caption-Based Image Retrieval

09 Experiment

Comparison to the SOTAs

	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	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**

- Outperforms SOTAs by a margin of 2~13%
- Improved visiolinguistic representations with pretraining and finetuning

09 Experiment

Effect of pretraining

	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	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**

- Zero-shot performance outperforming drastically (vs ViLBERT†)
  -> ViLBERT learned semantically meaningful visiolinguistic alignment during pretraining

09 Experiment

Size of pretraining dataset

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**

- Same ViLBERT setup but control the percentage of the conceptual caption dataset being used
- Monotonic increase of accuracy as the amount of data increases

10 Conclusion

1) Proposes a joint model, **ViLBERT** for learning task-agnostic **visual grounding** to jointly reason about text and images

2) Utilizes **co-attentional transformer layers** by introducing separate streams for vision and language processing

3) Compares with four established vision-and-language tasks and achieves significant improvements with 7~10% margin

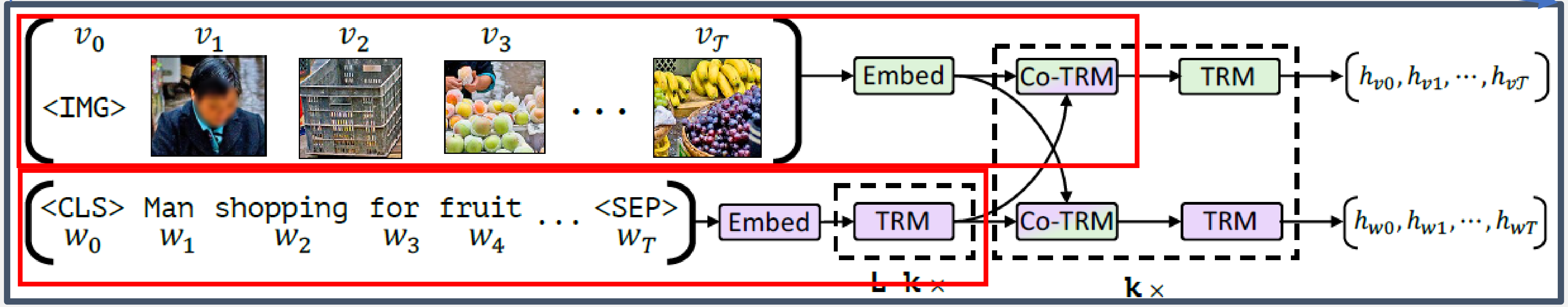
Q&A

11 Progress report

Main model

Evaluation

Training



- Image layer – Noah Lee
- Text layer – Donghui Im