Vision and Language Representation Learning

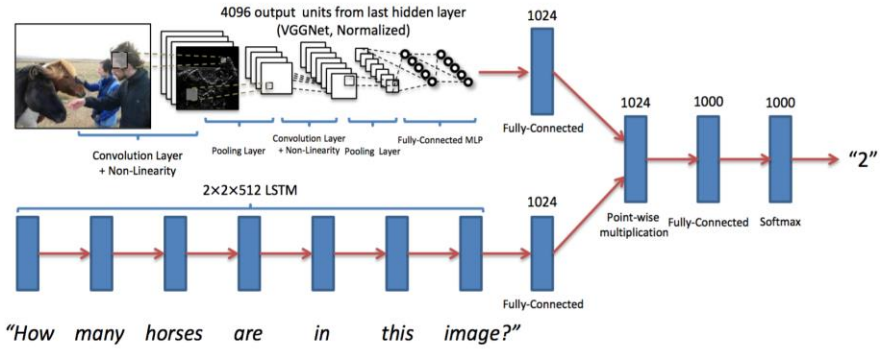
– Self Supervised Pretraining and Multi-Task Learning

Jiasen Lu

April 21, 2020

Allen Institute for AI

Vision and Language



Visual Question Answering

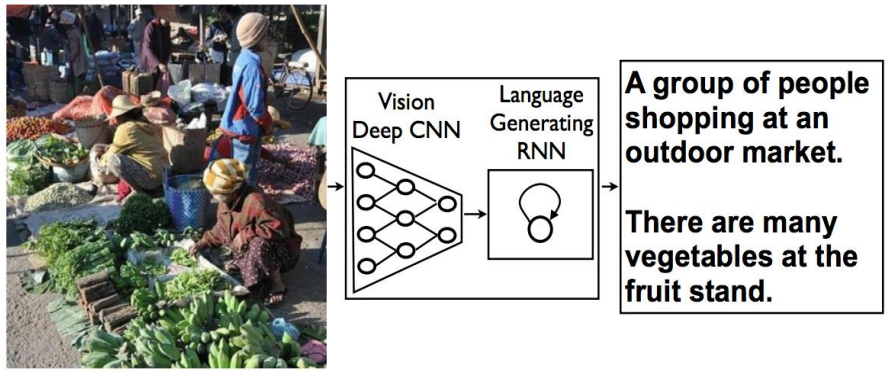
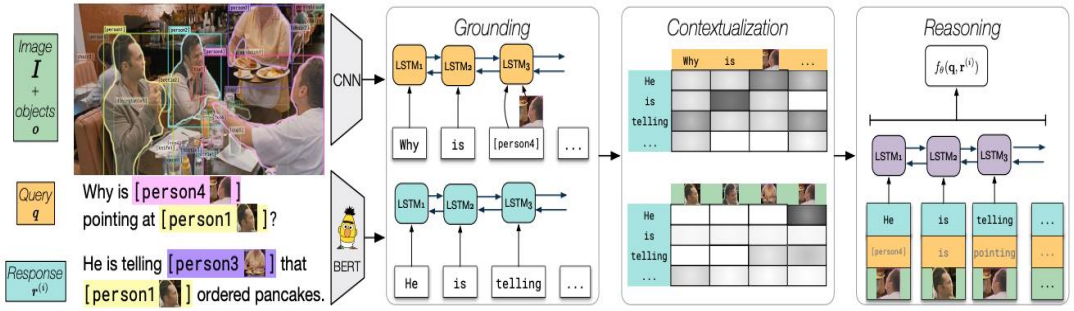
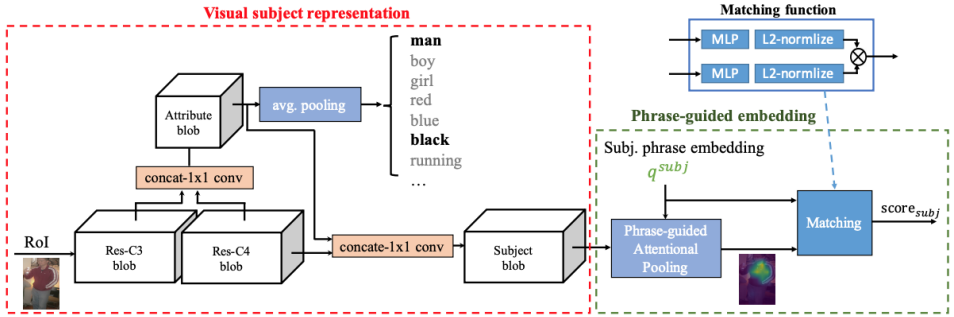


Image Captioning



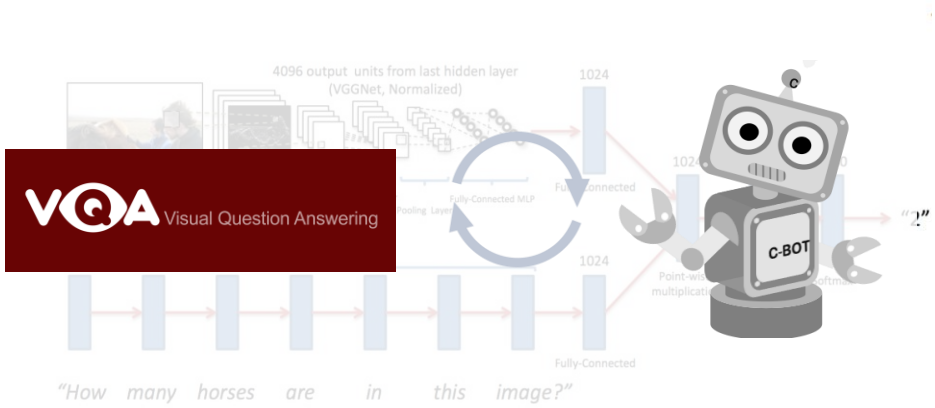
Visual Commonsense Reasoning



Refer Expression

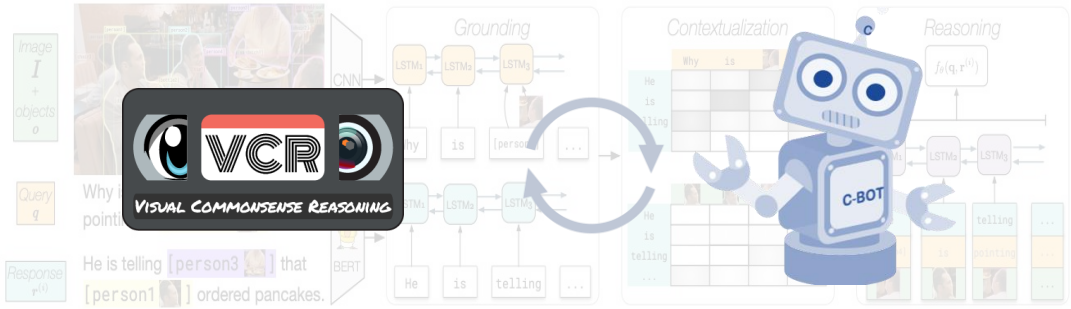
[Antol et.al. 2015, Vinyals et.al. 2015, Zellers et.al. 2018, Yu et.al 2018]

Vision and Language

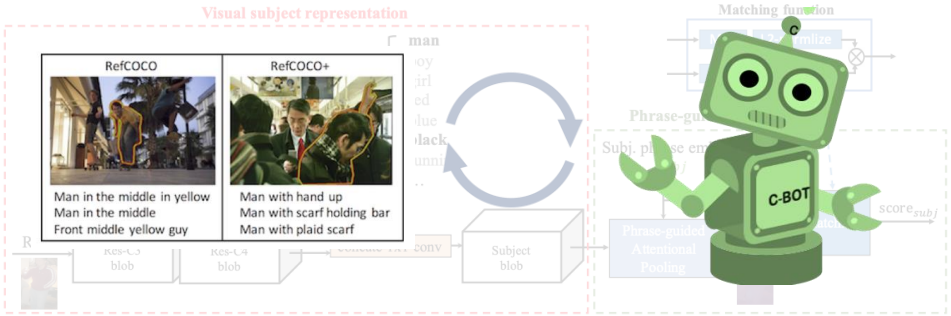


Visual Question Answering

Image Captioning



Visual Commonsense Reasoning

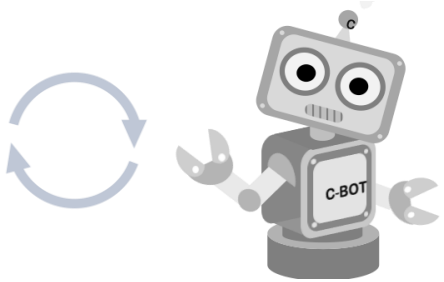


Refer Expression

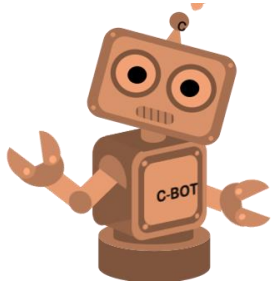
[Antol et.al. 2015, Vinyals et.al. 2015, Zellers et.al. 2018, Yu et.al 2018]

Visual Grounding

VQA Visual Question Answering



COCO Common Objects in Context



Q: What type of plant is this?

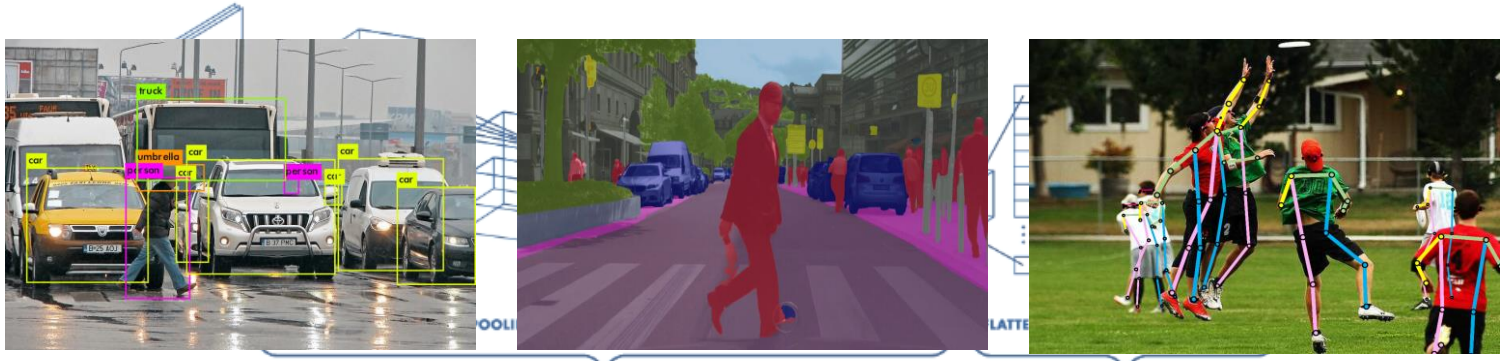
A: Banana

C: A bunch of red and yellow flowers on a branch.

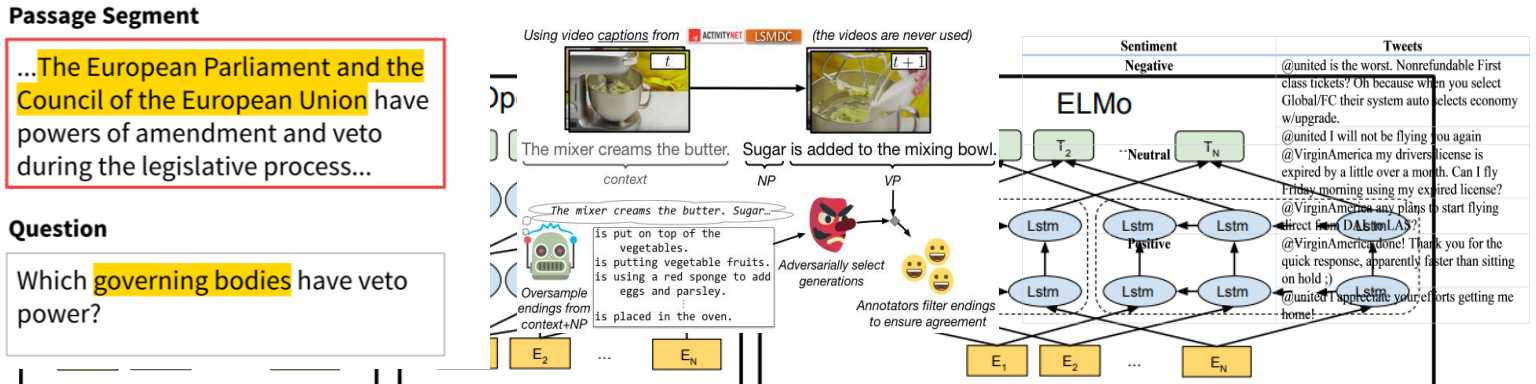
Common model for visual grounding and leverage them on a wide array of vision-and-language tasks

[Shen et.al 2018]

Pretrain-Transfer



Object Detection Semantic Segmentation Pose Estimation



Question Answering Commonsense Inference Sentiment Analysis

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

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.

Conceptual Caption Dataset

- Aligned image-caption pairs.
- 3.3 million image compared to 0.12 million in COCO caption.
- Automatically collected.

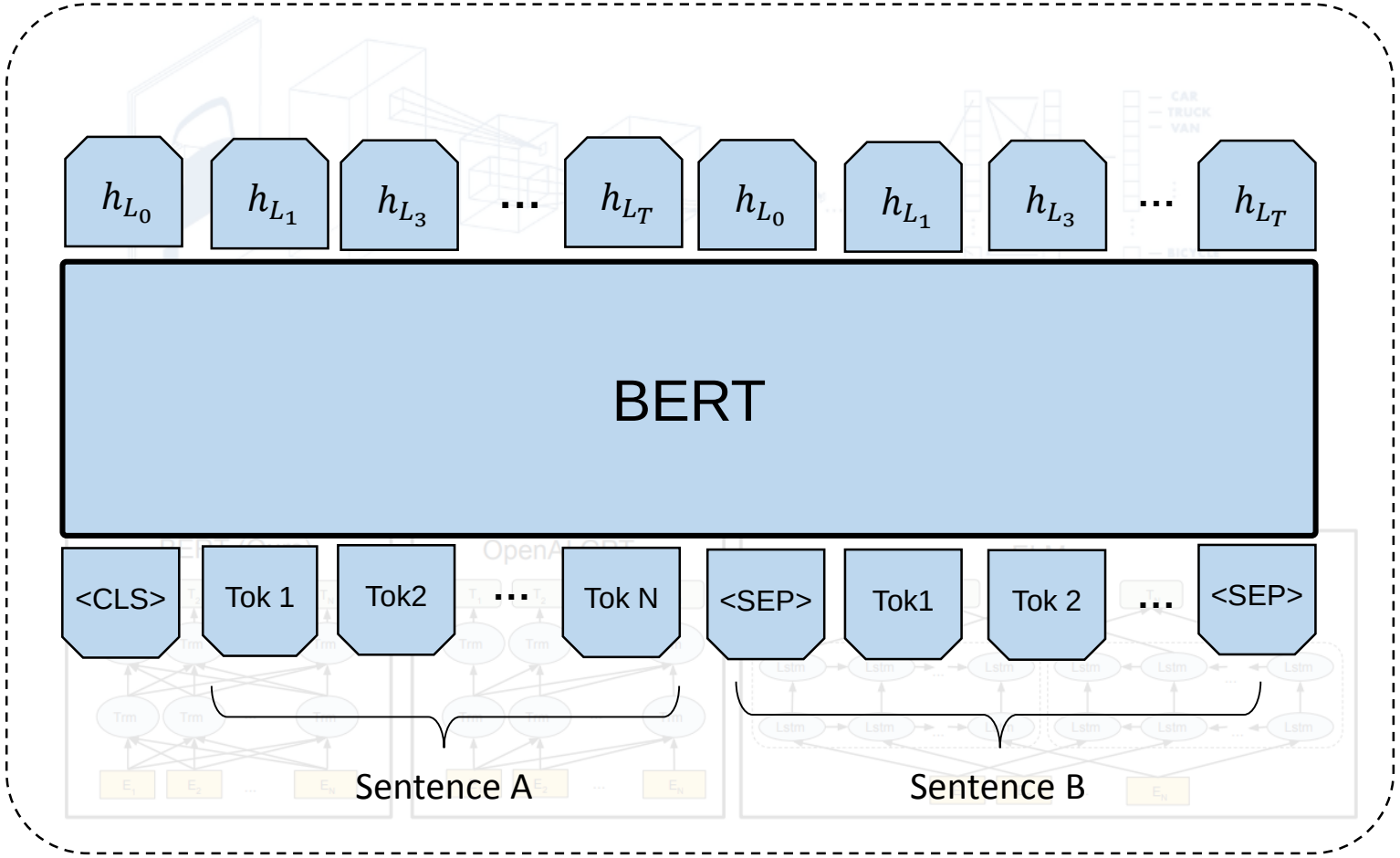
[Sharma et.al 2018]

BERT

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.

Conceptual Caption Dataset



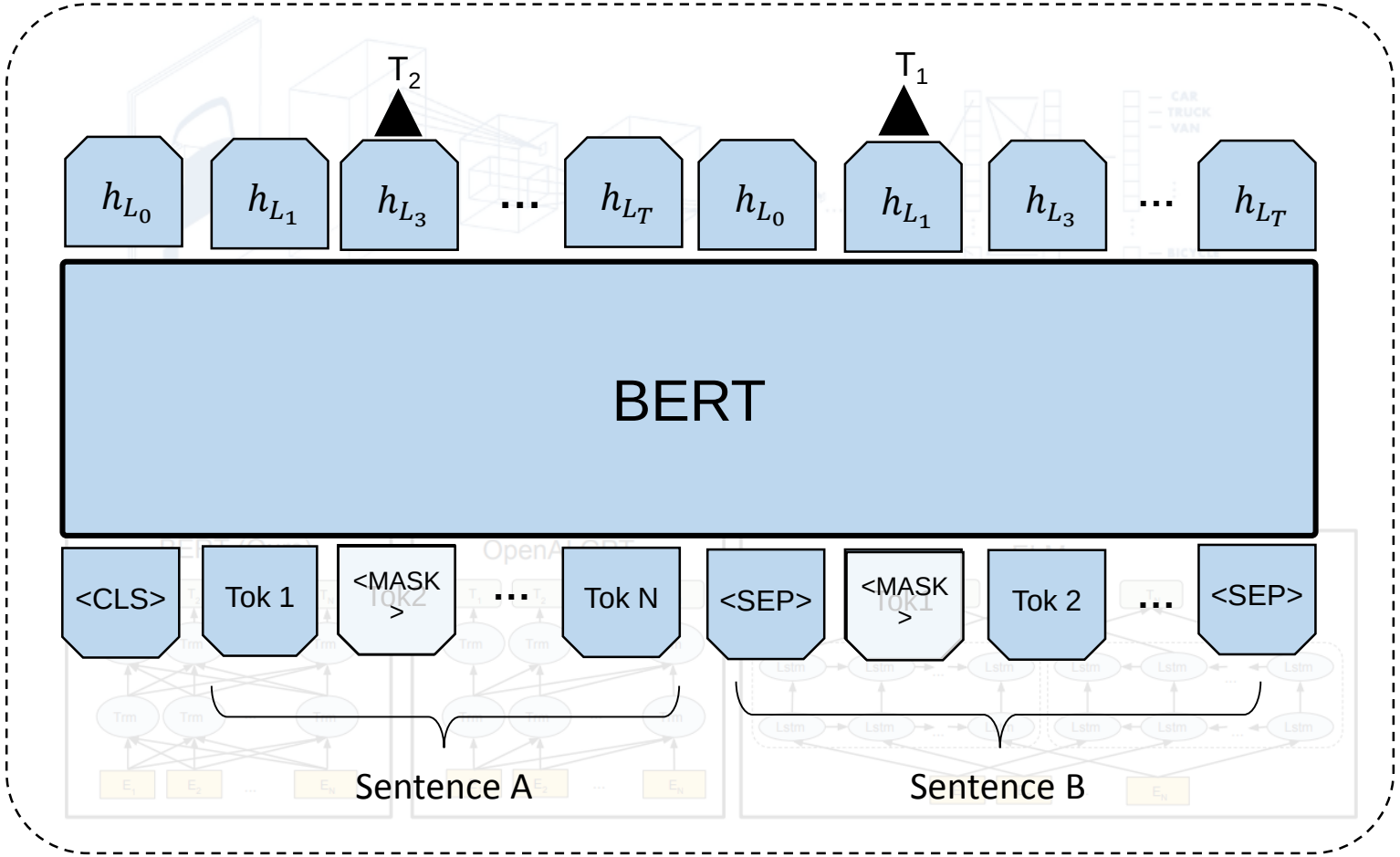
[Sharma et.al 2018, Devlin et.al 2018]]

ViLBERT

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.

Conceptual Caption Dataset



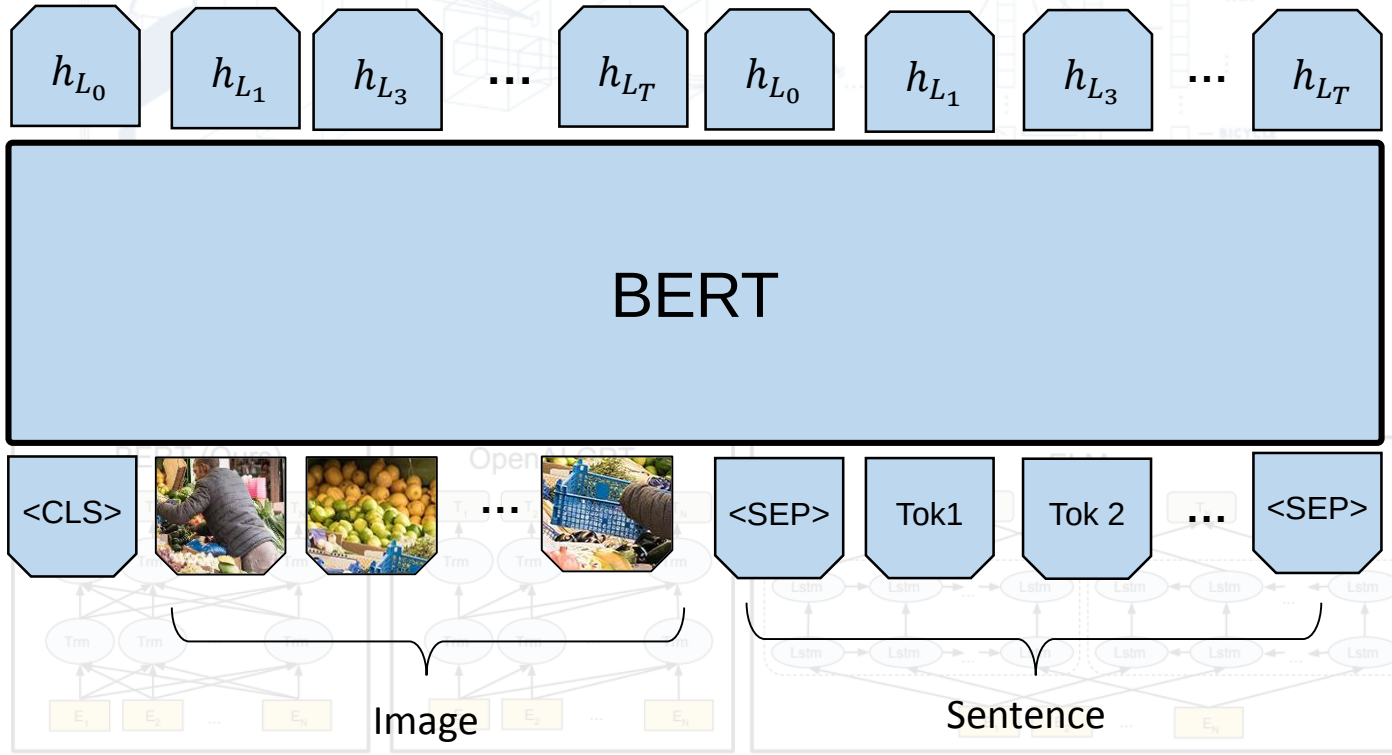
[Sharma et.al 2018, Devlin et.al 2018]]

Single-Stream model

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.

Conceptual Caption Dataset



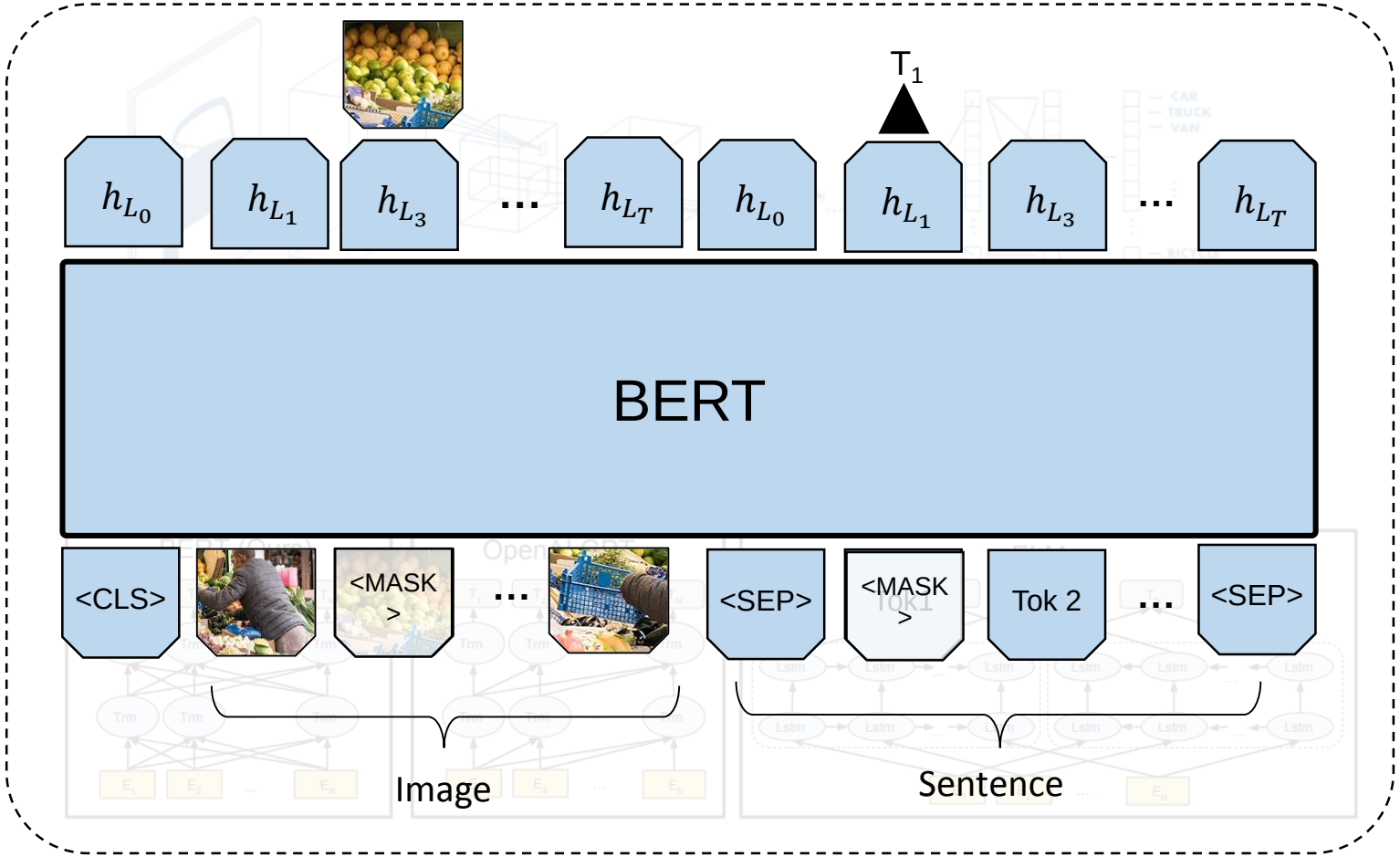
[Sharma et.al 2018, Devlin et.al 2018]]

Single-Stream model

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.

Conceptual Caption Dataset

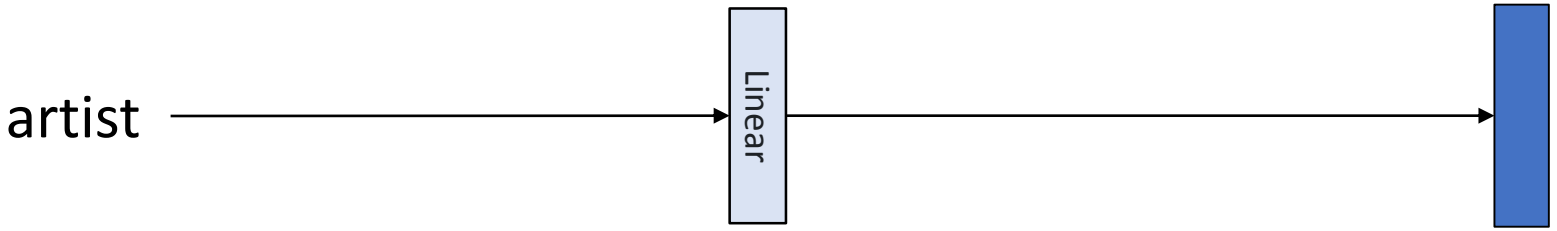


[Sharma et.al 2018, Devlin et.al 2018]]

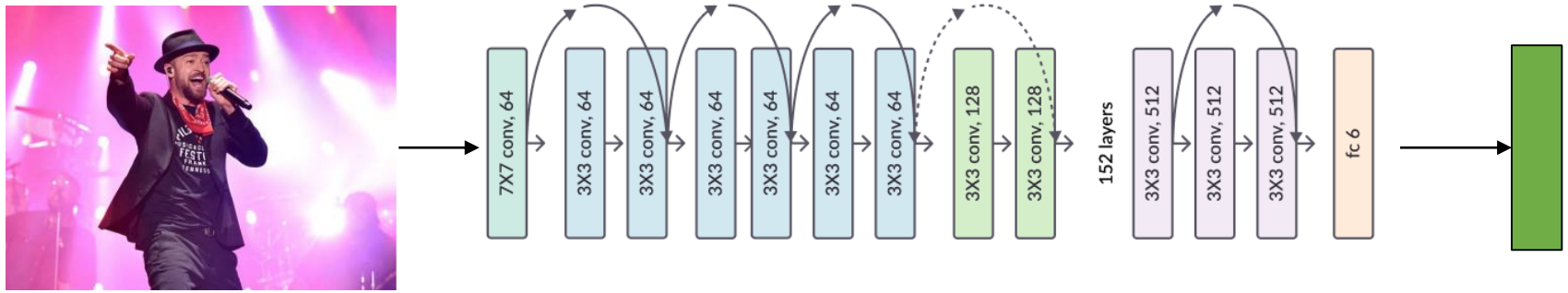
ViLBERT

Problem: Different modalities may require different level of abstractions.

- Linguistic stream:



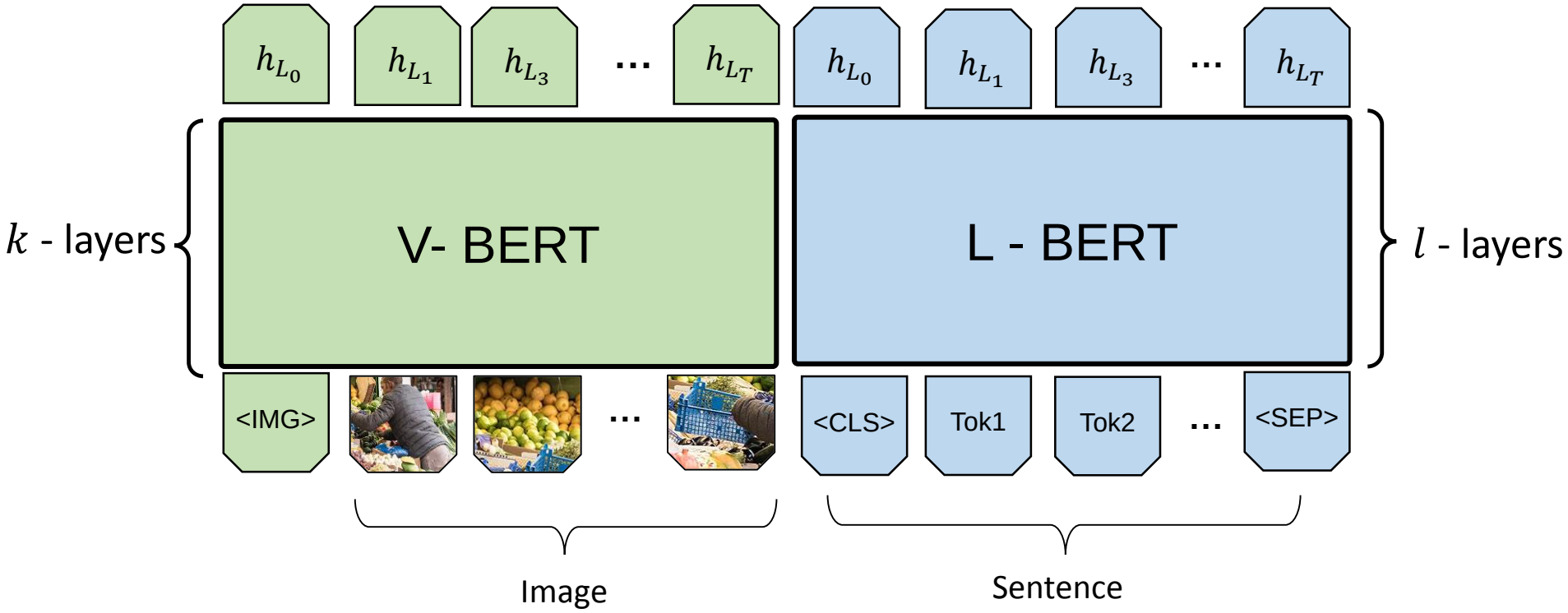
- Visual stream:



[He et.al. 2015]

ViLBERT

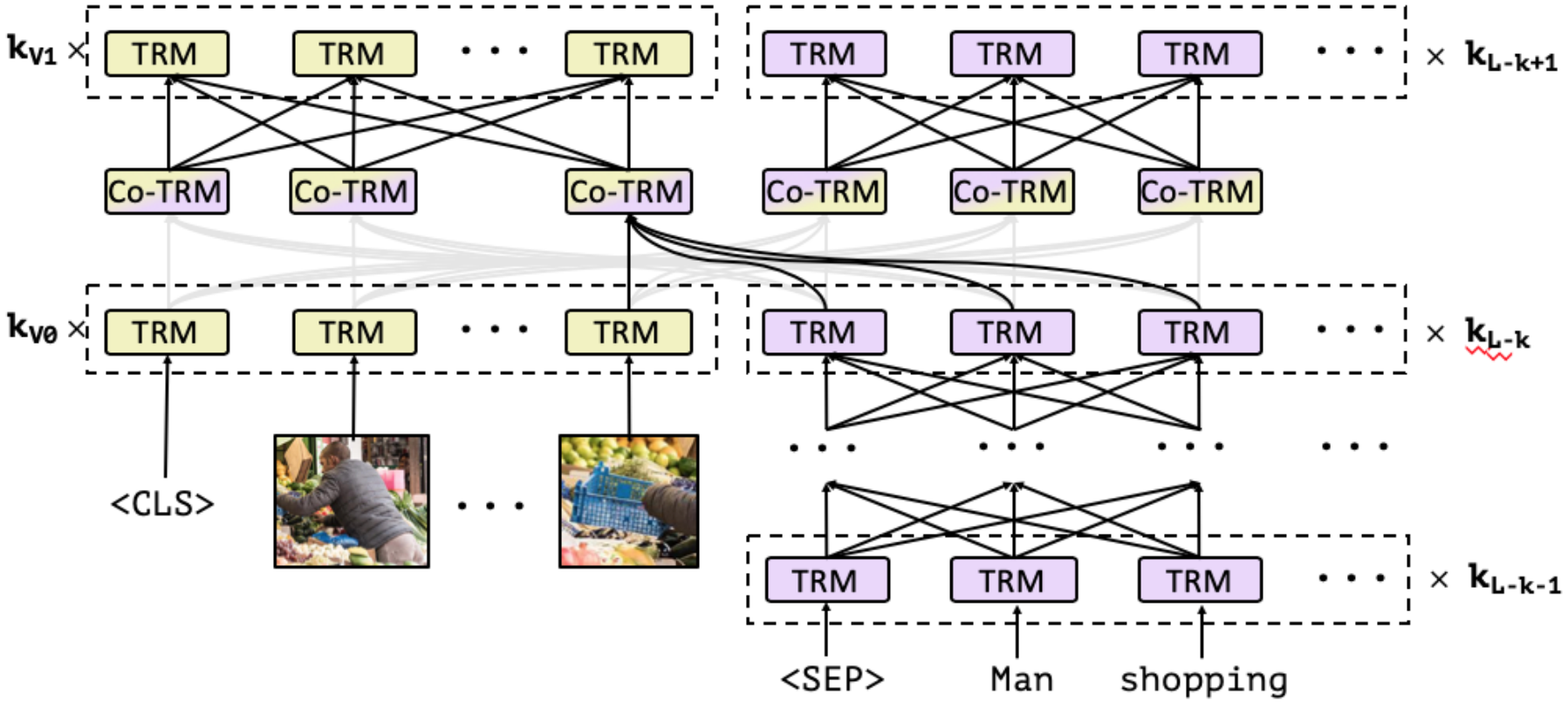
Solution: two-stream model which process visual and linguistic separately.



ViLBERT

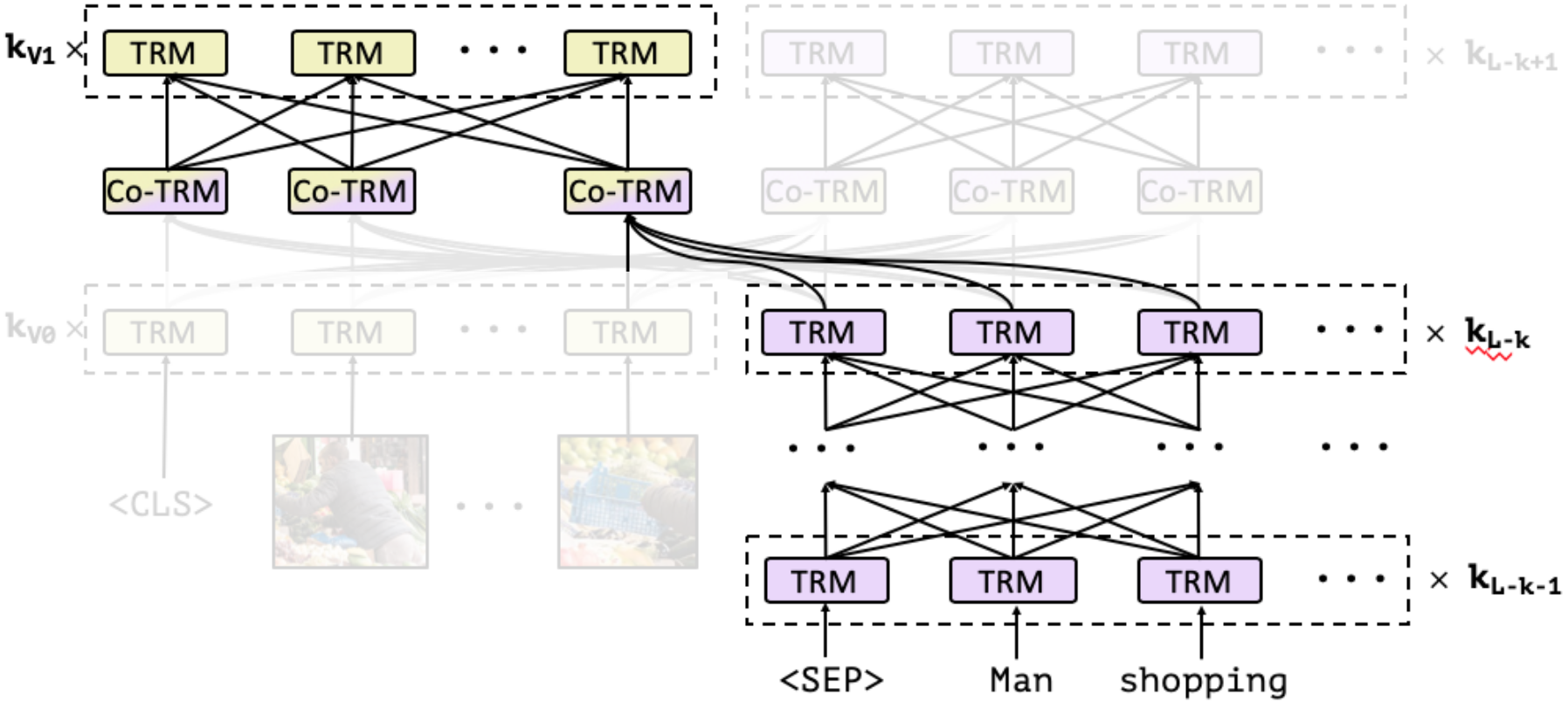
Problem: how to fuse two different modality?

Solution: use **co-attention** [*Lu et.al. 2016*] to fuse information between different source.



ViLBERT

Co-attention [*Lu et.al. 2016*] to fuse information between different source.



Pre-training Objective

Masked multi-modal modelling

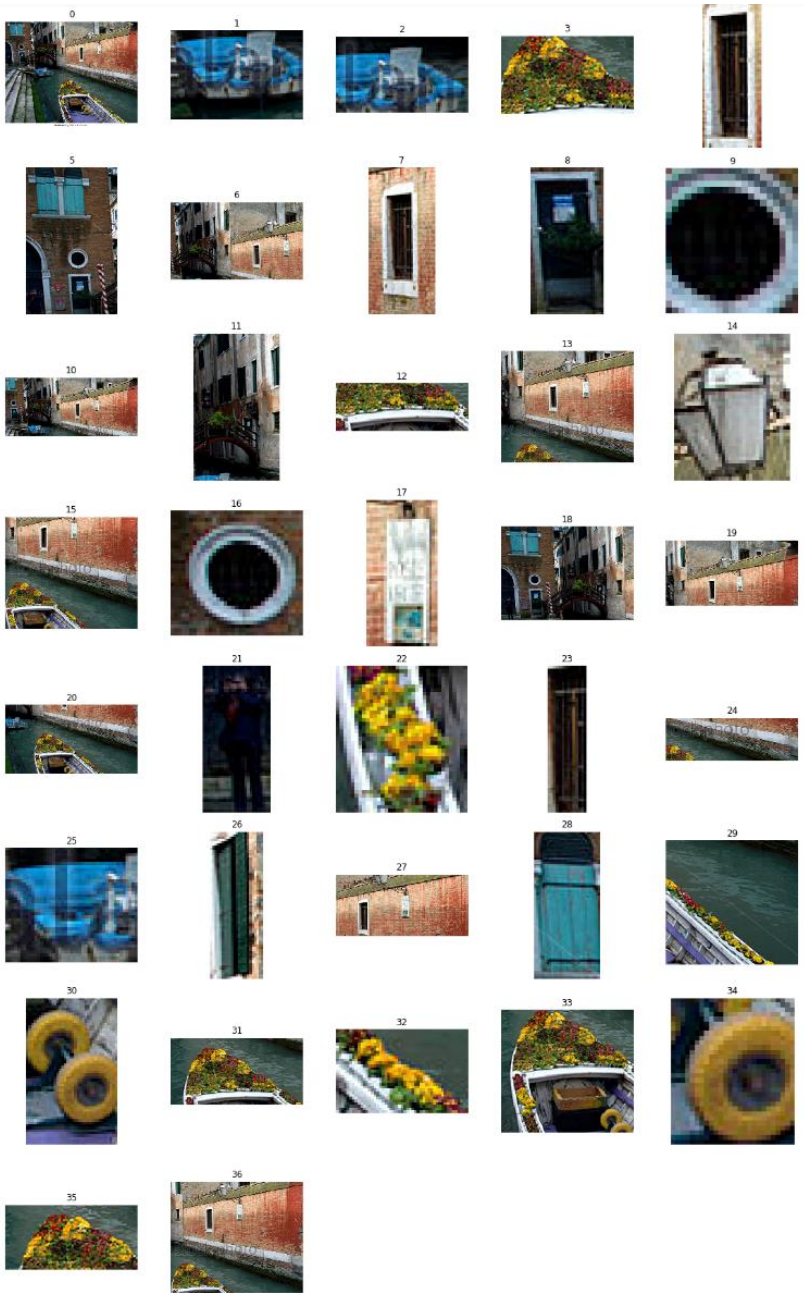
- Follows masked LM in BERT.
- 15% of the words or image regions to predict.
- Linguistic stream:
  - 80% of the time, replace with *[MASK]*.
  - 10% of the time, replace random word.
  - 10% of the time, keep same.
- Visual stream:
  - 80% of the time, replace with zero vector.

Multi-modal alignment prediction

- Predict whether image and caption is aligned or not

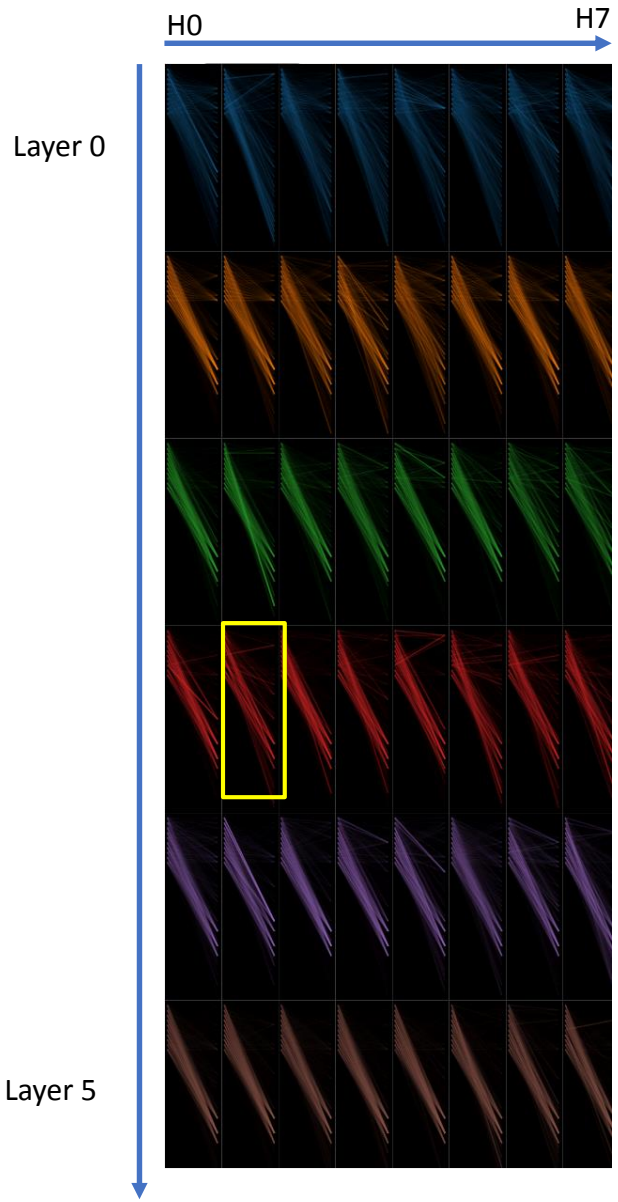
Visualizations

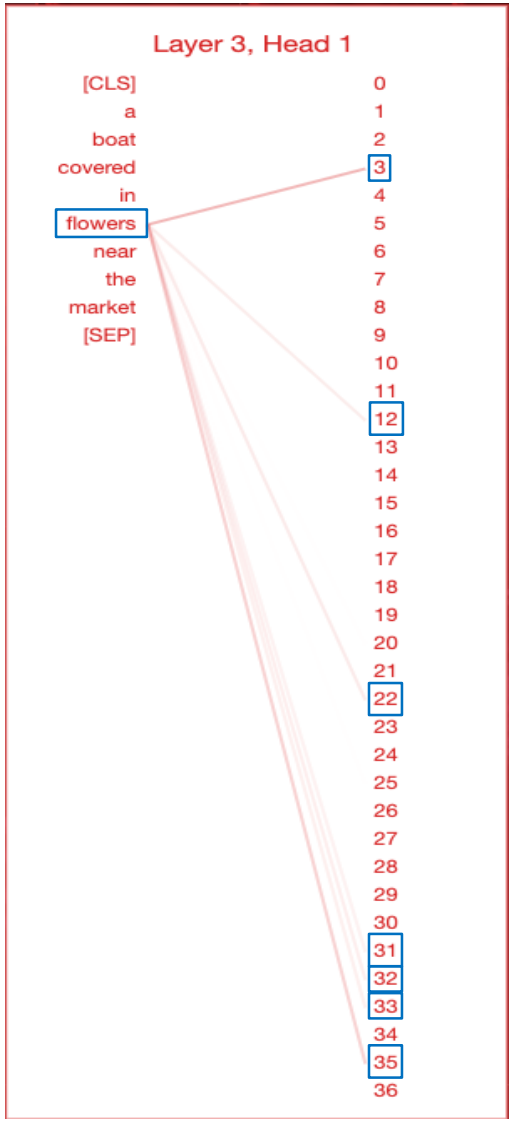
A boat covered in flowers near the market.



[Sharma et.al 2018]

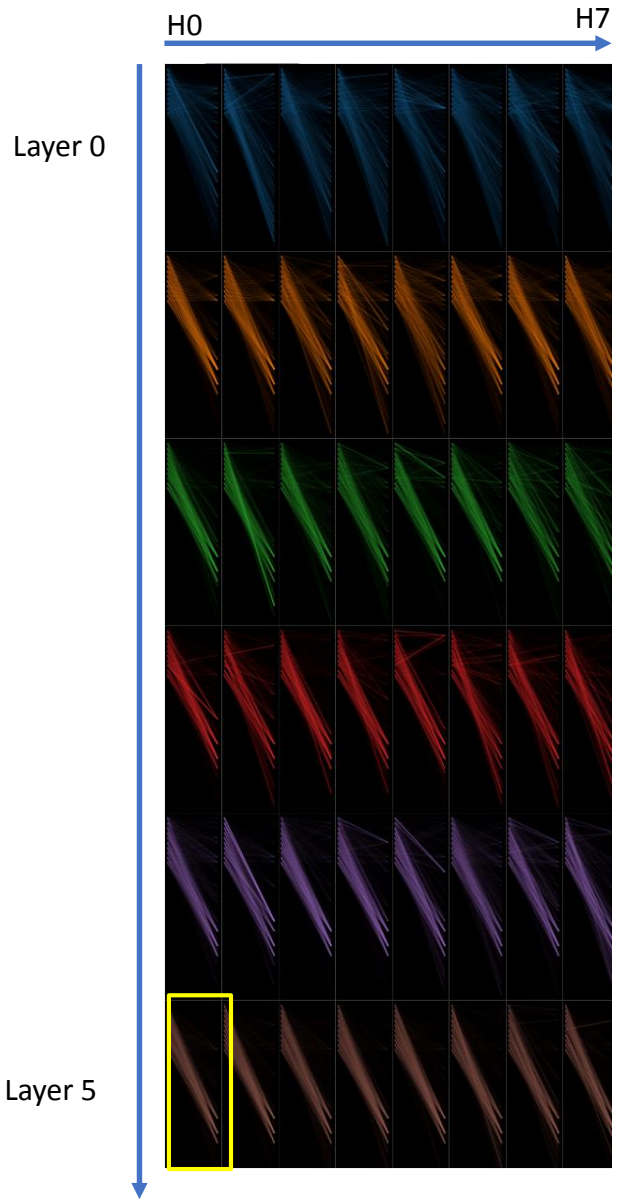
Sentence → Image

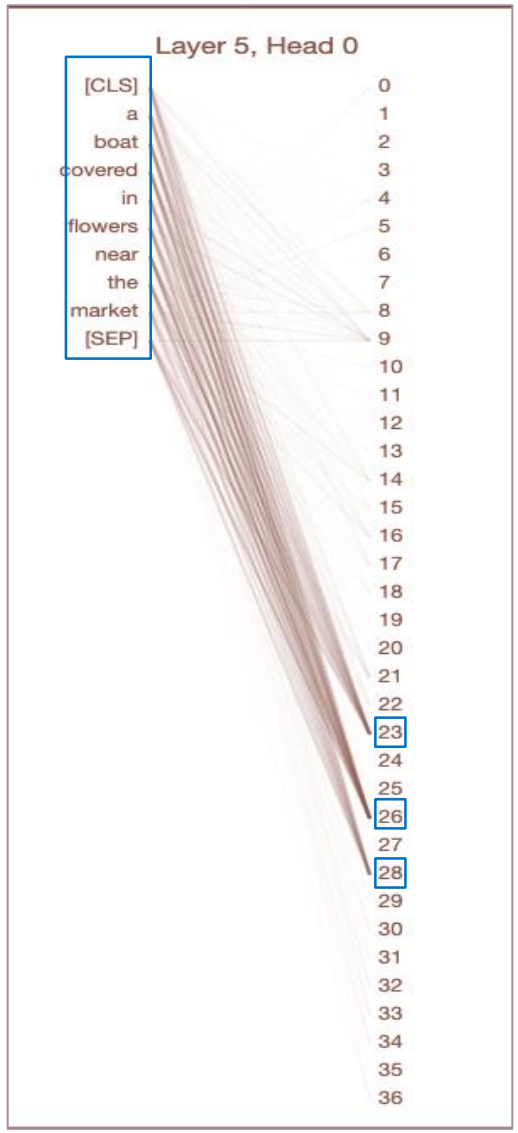




BertVis https://github.com/jessevig/bertviz

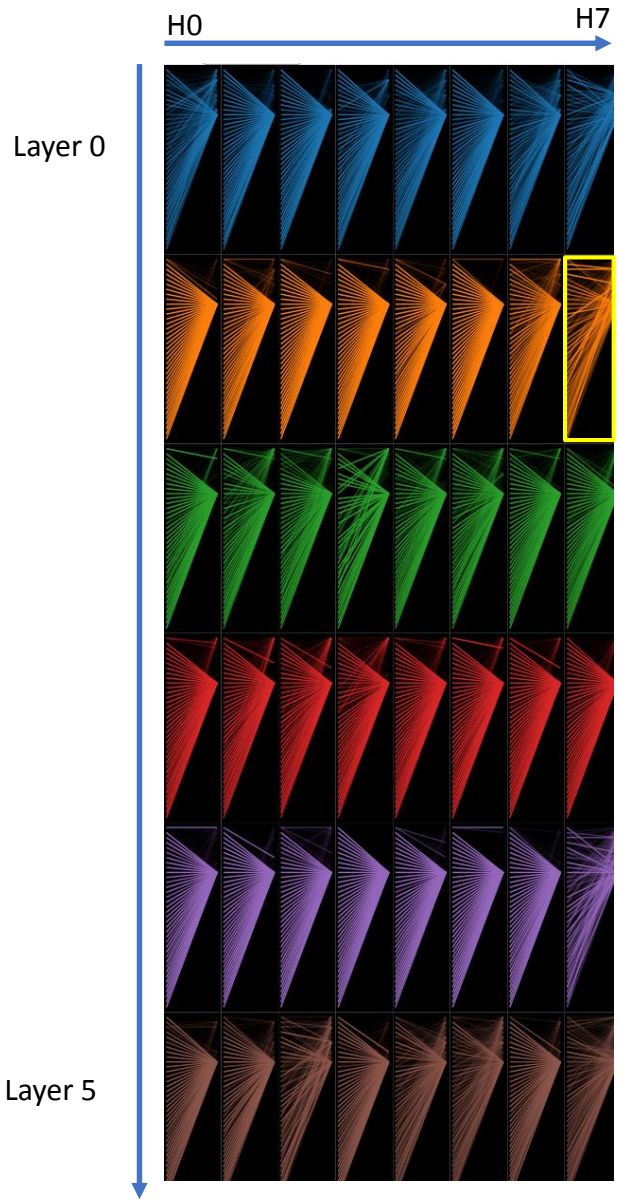
Sentence → Image

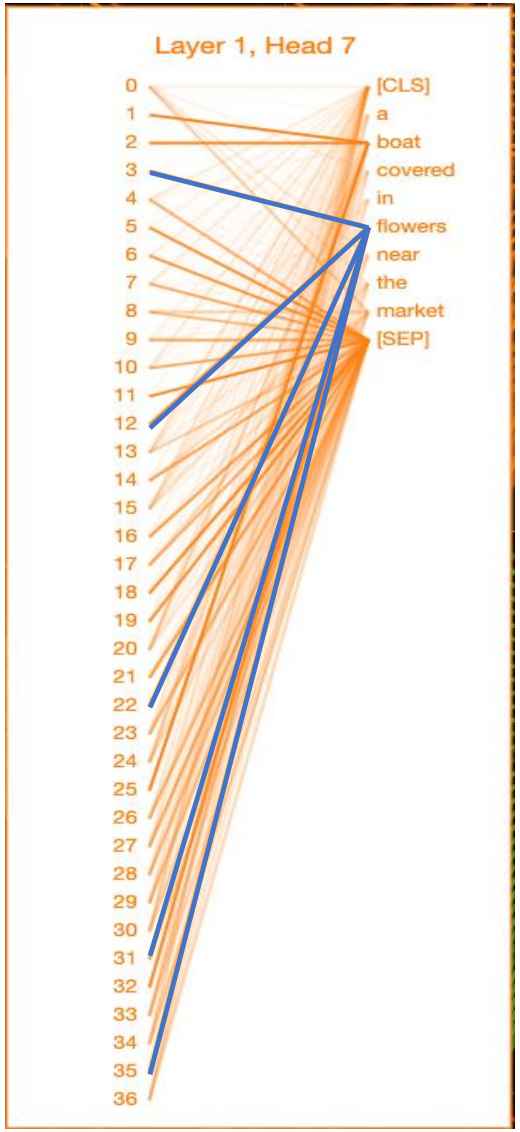


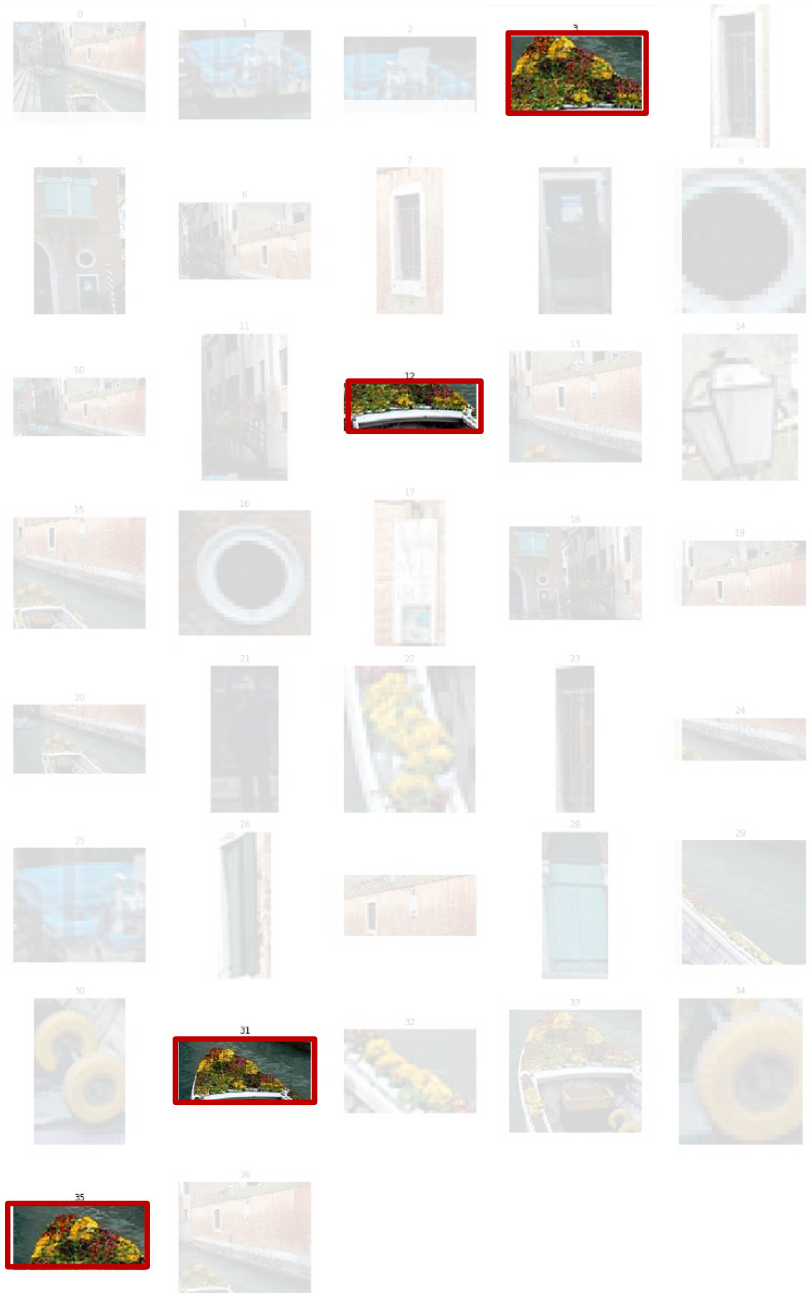


BertVis https://github.com/jessevig/bertviz

Image → Sentence

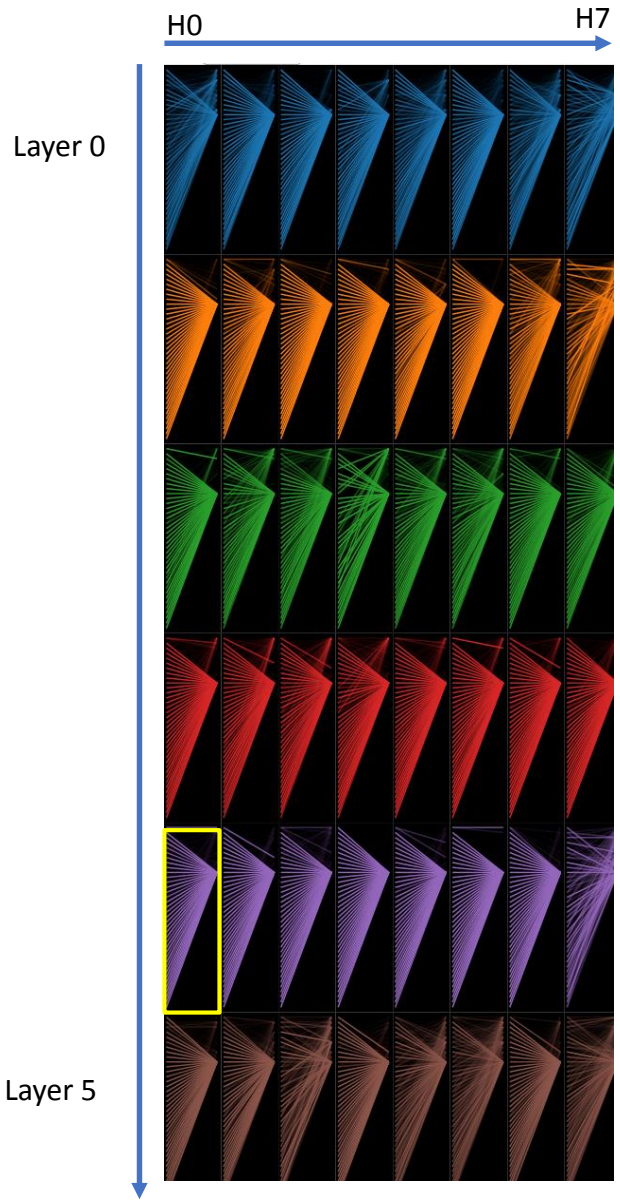


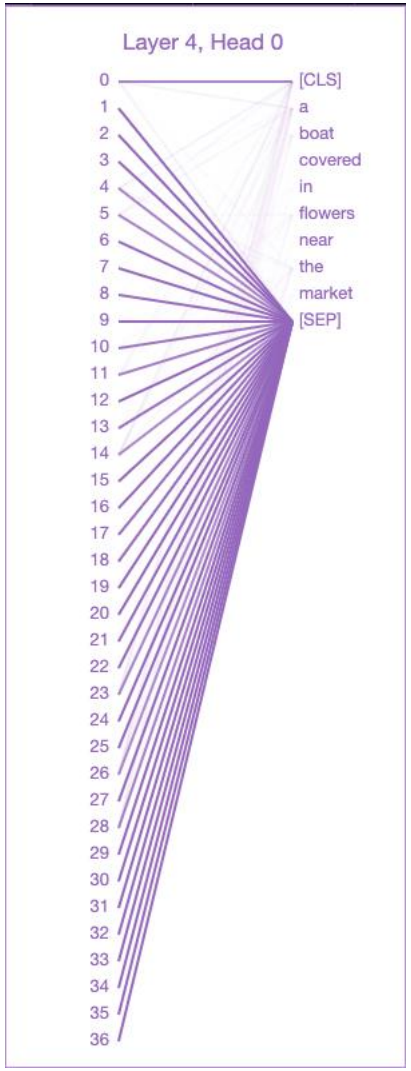


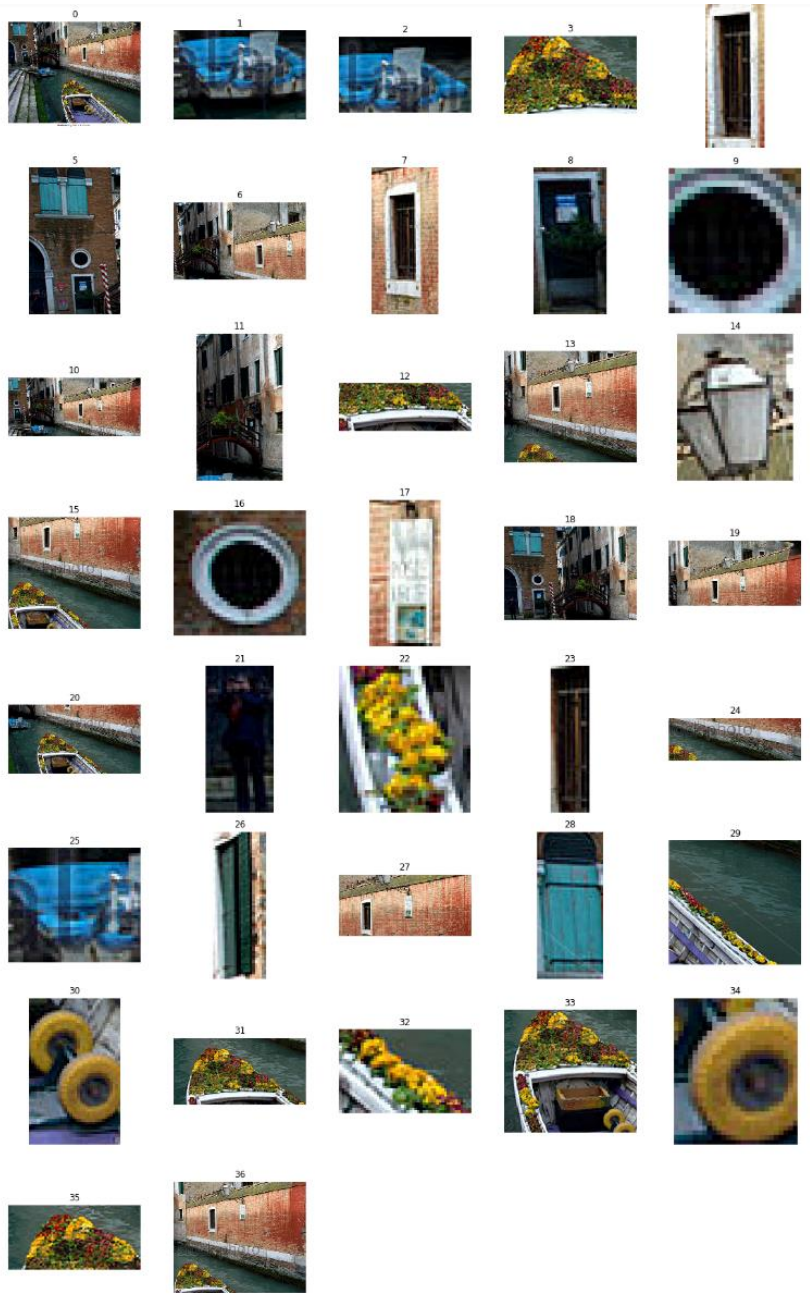


BertVis https://github.com/jessevig/bertviz

Image → Sentence

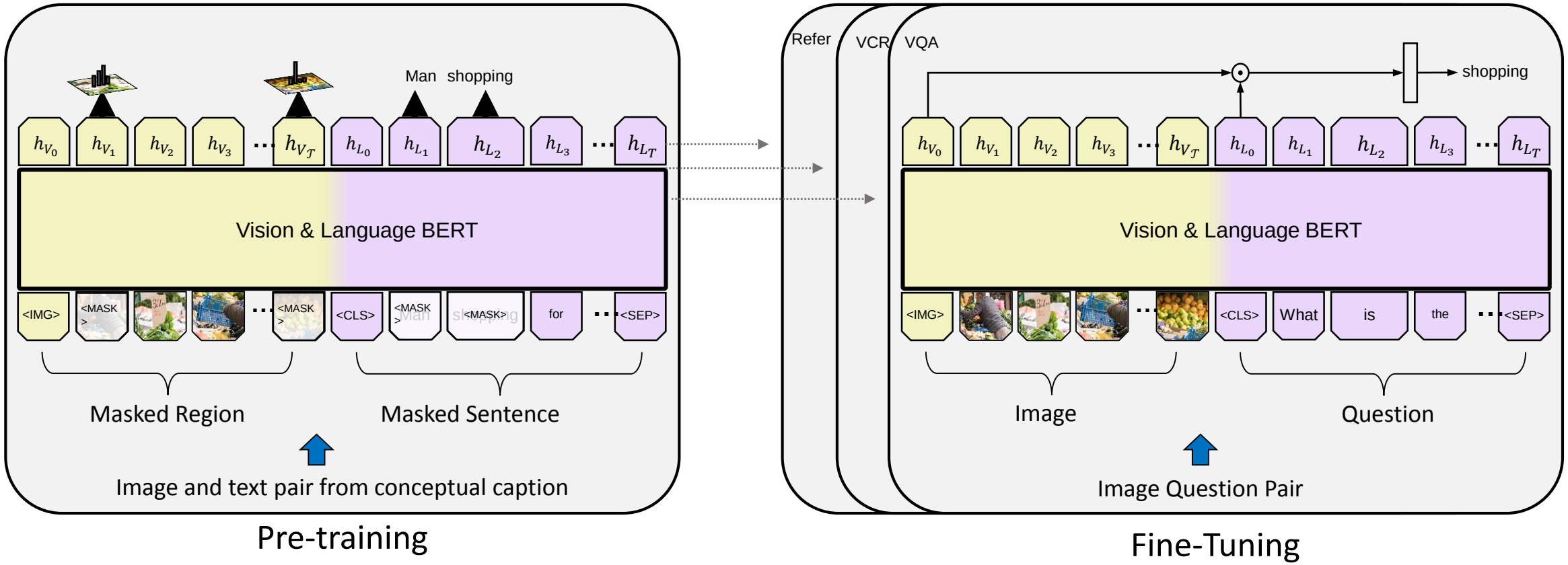






BertVis https://github.com/jessevig/bertviz

Fine-tuning Procedure



Tasks

VQA

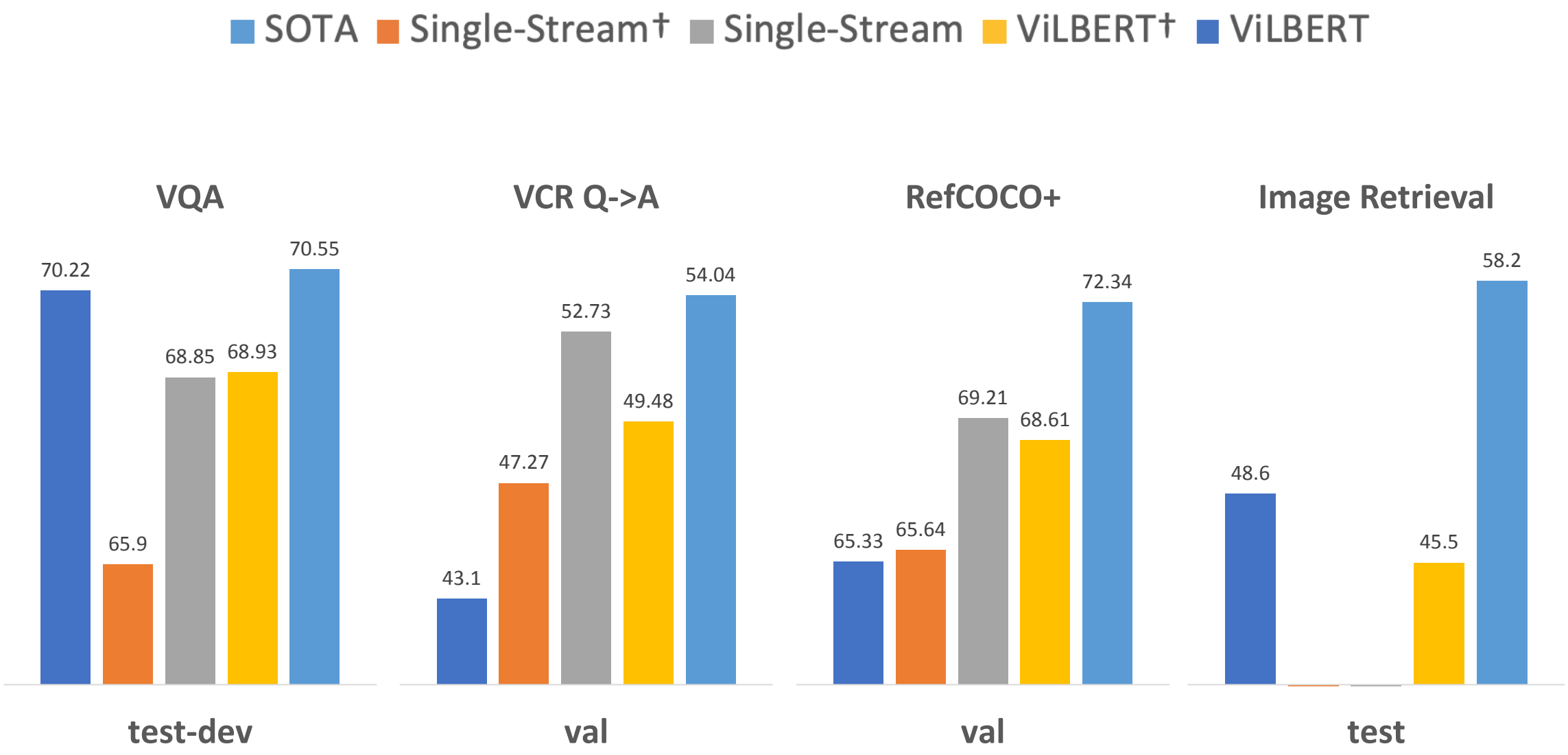
VCR Q→A VCR QA→R

Referring Expressions

Caption-Based Image Retrieval

[Antol 2015, zeller 2018, Yu 2016, Plummer 2015]

Results


SOTA
Single-Stream†
Single-Stream
ViLBERT†
ViLBERT
VQA
70.22
65.9
68.85
68.93
70.55
test-dev
VCR Q->A
43.1
47.27
52.73
49.48
54.04
val
RefCOCO+
65.33
65.64
69.21
68.61
72.34
val
Image Retrieval
48.6
45.5
58.2
test

Concurrent Work

VisualBERT: A Simple and Performant Baseline for Vision and Language

Liunian Harold Li[†], Mark Yatskar[*], Da Yin[°], Cho-Jui Hsieh[†] & Kai-Wei Chang[†]
[†]University of California, Los Angeles
[*]Allen Institute for Artificial Intelligence
[°]Peking University
liunian.harold.li@cs.ucla.edu, marky@allenai.org,
wade_yin9712@pku.edu.cn, {chohsieh, kwchang}@cs.ucla.edu

LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Hao Tan **Mohit Bansal**
UNC Chapel Hill
{haotan, mbansal}@cs.unc.edu

Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training

Gen Li [†*], Nan Duan [§], Yuejian Fang [†], Ming Gong [‡], Daxin Jiang [‡], Ming Zhou [§]
[†] School of Software & Microelectronics, Peking University, Beijing, China
[‡] STCA NLP Group, Microsoft, Beijing, China
[§] Microsoft Research Asia, Beijing, China
ligen.li@pku.edu.cn, fangyj@ss.pku.edu.cn
{nanduan, migon, djiang, mingzhou}@microsoft.com

VL-BERT: Pre-training of Generic Visual-Linguistic Representations

Weijie Su[1,2*], Xizhou Zhu[1,2*], Yue Cao[2], Bin Li[1], Lewei Lu[2], Furu Wei[2], Jifeng Dai[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia
{jackroos,ezra0408}@mail.ustc.edu.cn,binli@ustc.edu.cn
{yuecao,lewlu,fuwei,jifdai}@microsoft.com

Unified Vision-Language Pre-Training for Image Captioning and VQA

Luowei Zhou[1*], Hamid Palangi[2], Lei Zhang[3], Houdong Hu[4], Jason J. Corso[1], Jianfeng Gao[2]
[1] University of Michigan [2] Microsoft Research [3] Microsoft Cloud & AI [4] Microsoft AI & Research
{luozhou, jjcorso}@umich.edu {hpalangi, leizhang, houhu, jfgao}@microsoft.com

UNITER: Learning UNiversal Image-TExt Representations

Yen-Chun Chen,[*] Linjie Li,[*] Licheng Yu,[*]
Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, Jingjing Liu
Microsoft Dynamics 365 AI Research
{yen-chun.chen,lindsey.li,licheng.yu}@microsoft.com
{ahmed.elkholy,fiahmed,zhe.gan,yu.cheng,jingjl}@microsoft.com

[Li 2019, Tan 2019, Li 2019, Su 2019, Zhou 2019, Chen 2019]

Summary

Summary

Task-agnostic visiolinguistic representations pretraining for visual grounding

- Introduce ***pretrain-transfer*** to vision and language tasks.
- Achieve SOTA on multiple vision and language tasks.

Limitations

The model can still learn inconsistent grounding by task specific finetuning.

- Training multiple vision and language task together – multi-task V&L

Multi-Task V&L Learning

ViLBERT Problems:

- Inconsistent grounding by task specific finetuning.
- Four V&L tasks.
- Model is huge, overfitting.

What we want:

- Test on more tasks.
- Consistent Grounding across tasks.
- Explore the limit of the model.

One Model for V&L:

VQA
- VQA
- Genome QA
- GQA

Image Description
- Caption based Retrieval (COCO)
- Caption based Retrieval (COCO)

Referring Expression
- Ref COCO
- Ref COCO+
- Ref COCOg
- Visual 7w
- GuessWhat

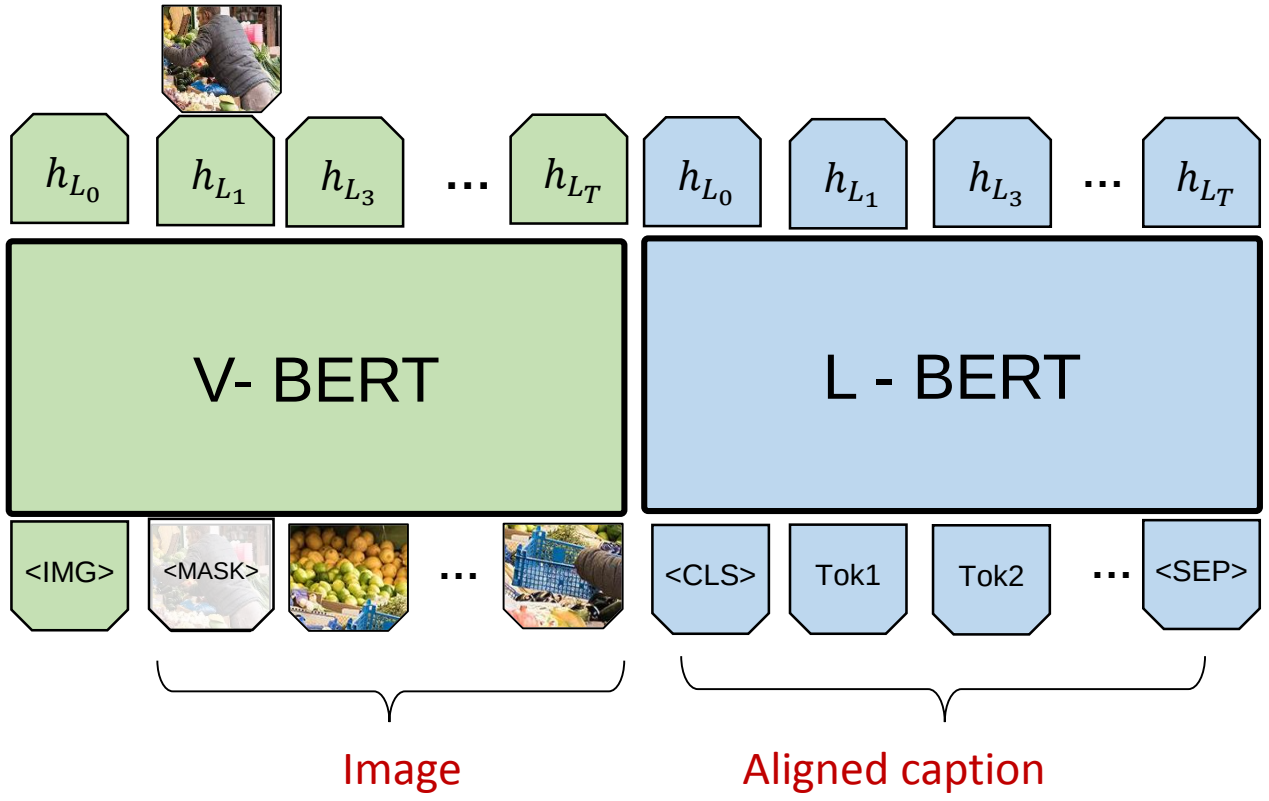
V&L Verification
- NLVR2
- Visual Entailment

Model improvements over ViLBERT

Multi-Task V&L Learning

Model improvements over ViLBERT

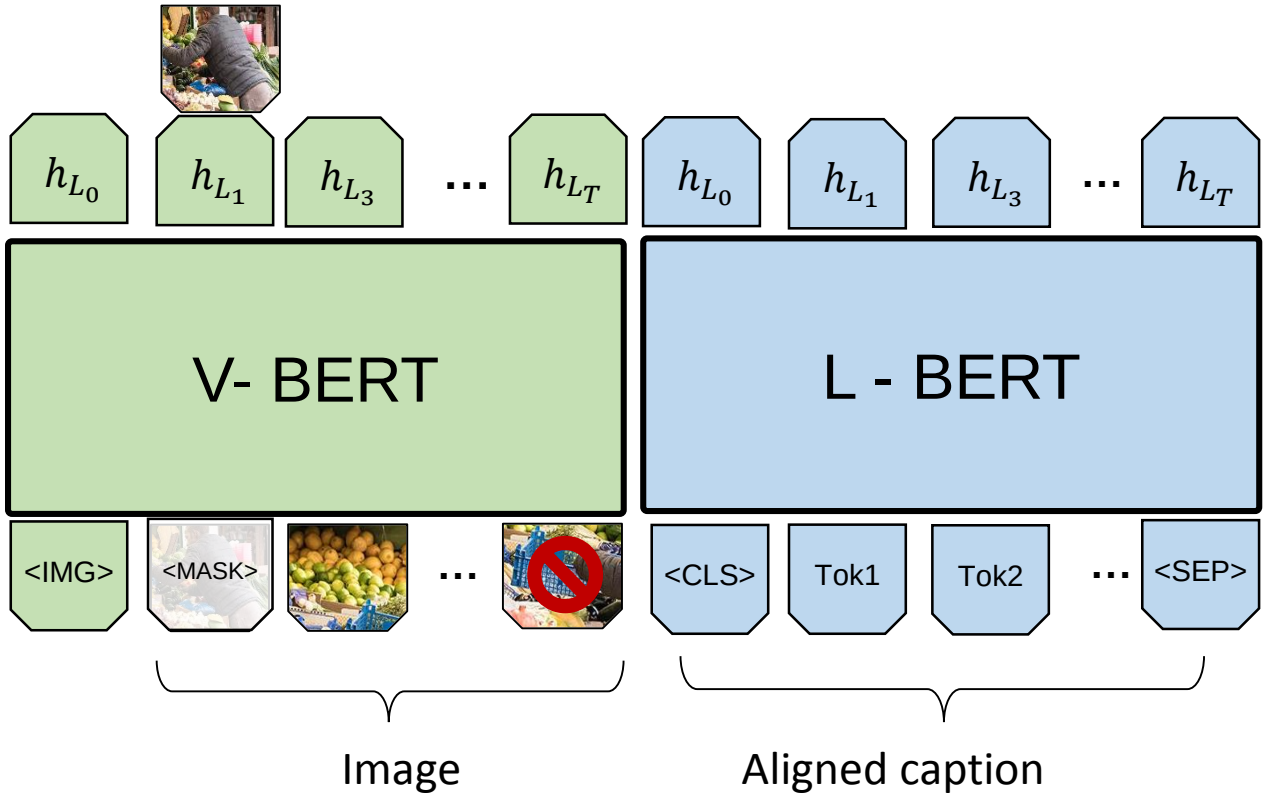
- Masked multi-modal modelling only for aligned image caption pairs.



Multi-Task V&L Learning

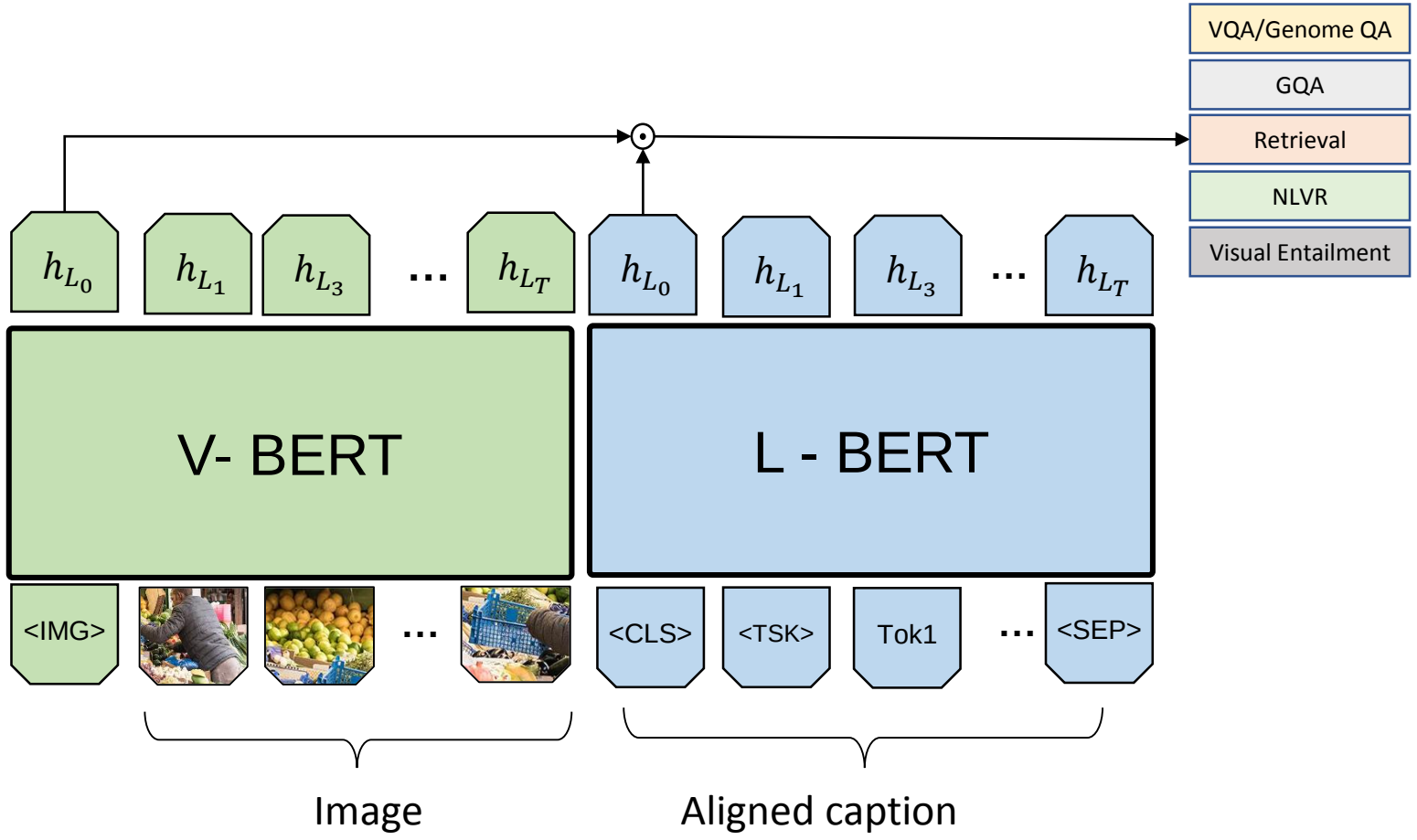
Model improvements over ViLBERT

- Masked multi-modal modelling only for aligned image caption pairs.
- Masking overlapped image regions (IOU > 0.4).



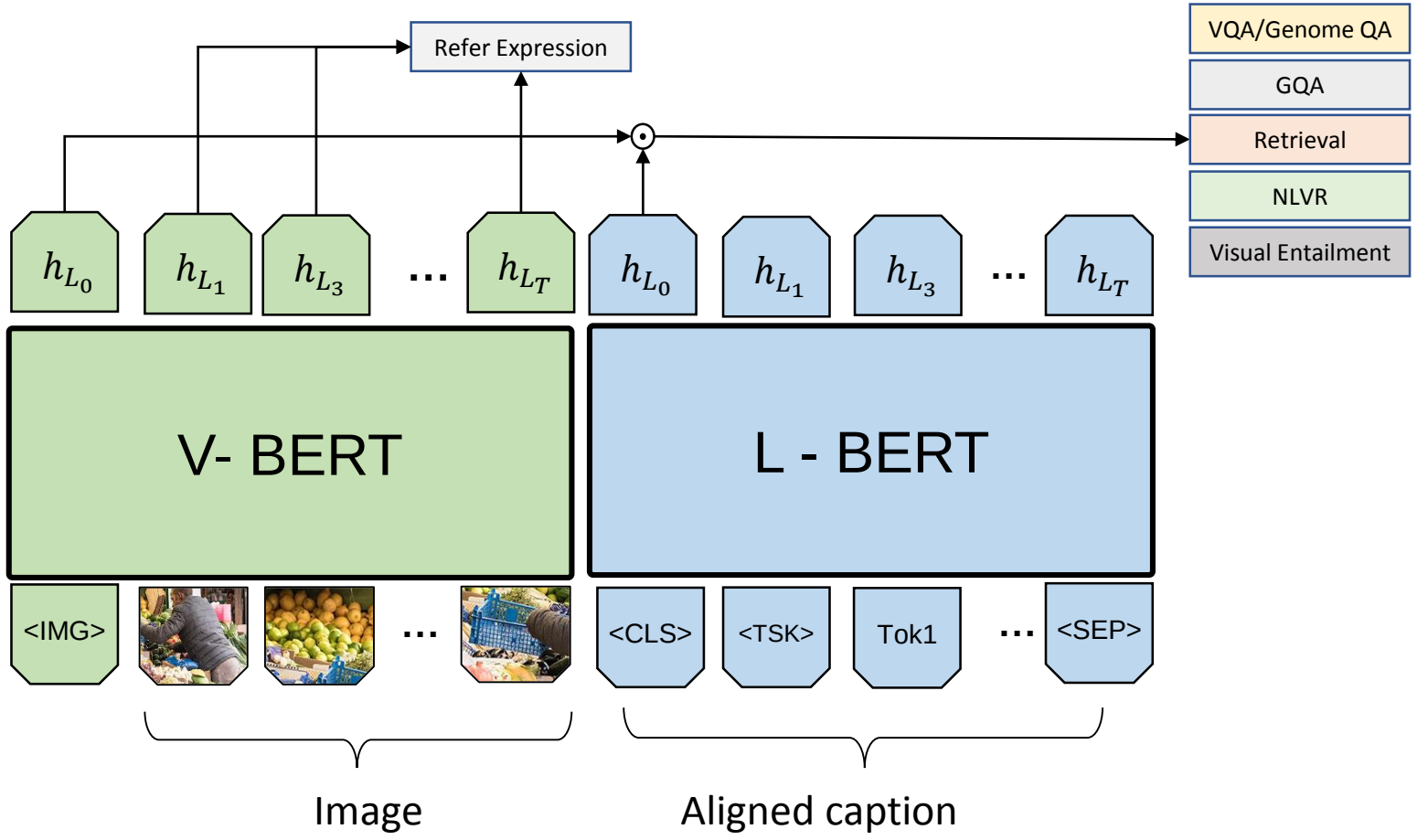
Multi-Task Vision and Language Learning

- Use different head but similar tasks share the same head.



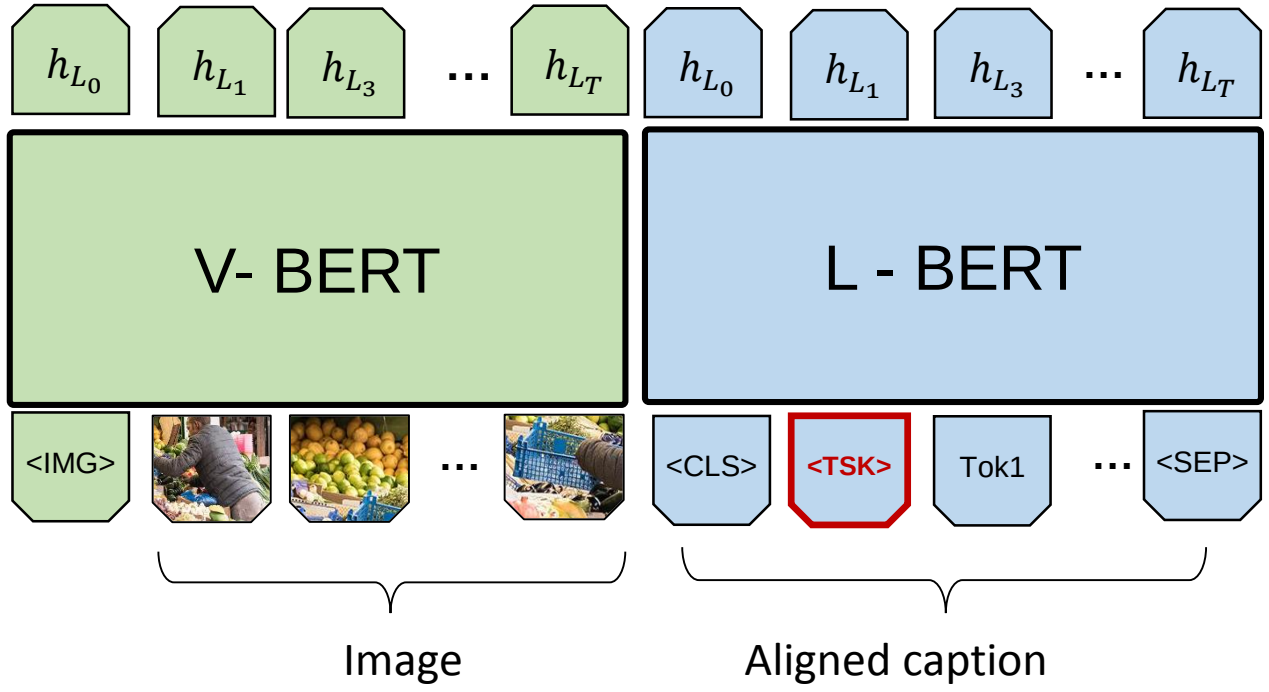
Multi-Task Vision and Language Learning

- Use different head but similar tasks share the same head.



Multi-Task Vision and Language Learning

- Use different head but similar tasks share the same head.
- Add <TSK> token for multi-task training.



Multi-Task Vision and Language Learning

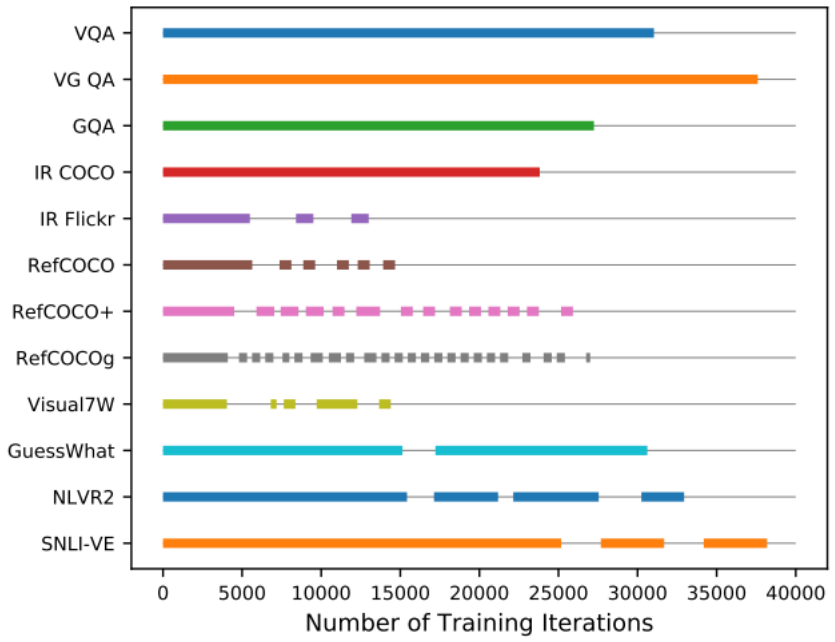
- Use different head but similar tasks share the same head.
- Add <TSK> token for multi-task training.
- Dynamic Stop and Go

Multi-Task Vision and Language Learning

- Use different head but similar tasks share the same head.
- Add <TSK> token for multi-task training.
- Dynamic Stop and Go

Algorithm 1: DSG for Multi-Task Learning

$n_t \leftarrow$ number of iterations per epoch for task t
$\Delta \leftarrow$ size of gap between iterations in `stop` mode
$\mathrm{DSG}_t \leftarrow$ `go`
for $i \leftarrow 1$ **to** *MaxIter* **:**
  for $t \in$ *Tasks* **:**
    if $\mathrm{DSG}_t =$ `go` **or** ($\mathrm{DSG}_t =$ `stop` **and** $i \bmod \Delta = 0$) **:**
      Compute task loss $L_t(\theta)$ and gradient $\nabla_t(\theta)$
      Update $\theta \leftarrow \theta - \epsilon \nabla_t(\theta)$, where $\theta = \theta_s \cup \theta_t$
    if $i \bmod n_t = 0$ **:**
      Compute validation score s_t on task t
      if $DSG_t =$ *go* **and** *Converged* (s_t) **:**
        $\mathrm{DSG}_t \leftarrow$ `stop`
      else if $DSG_t =$ *stop* **and** *Diverged* (s_t) **:**
        $\mathrm{DSG}_t \leftarrow$ `go`
  end
end



Multi-Task V&L Learning

Multi-Task Vision and Language Learning

- Use different head but similar tasks share the same head.
- Add <TSK> token for multi-task training.
- Dynamic Stop and Go

Training Procedure

- Conceptual caption pre-training.
- Multi-task training.
- Finetune on single task.

Compare with SOTA

ViLBERT VisualBERT VLBERT ViLBERT* LXMERT UNITER MT-VL MT-VL-FT

Model

Pretrained

VQA

RefCOCO+

test-dev

test

[Li 2019, Tan 2019, Li 2019, Su 2019, Chen 2019]

Compare with SOTA

	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream							
Pretrained	CC							

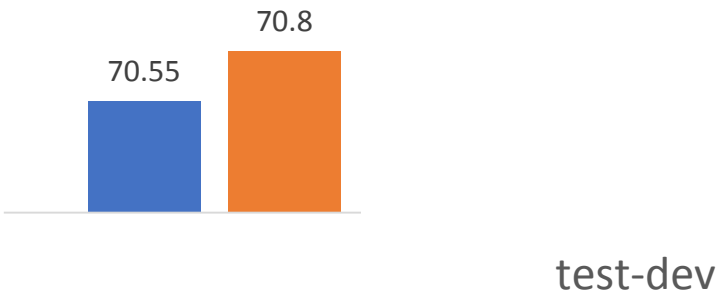
[Li 2019, Tan 2019, Li 2019, Su 2019, Chen 2019]

Compare with SOTA

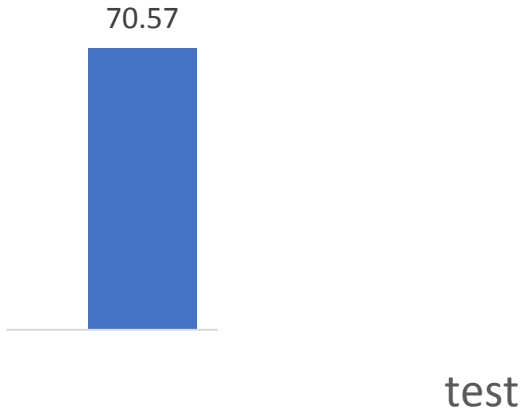
ViLBERT VisualBERT VLBERT ViLBERT* LXMERT UNITER MT-VL MT-VL-FT

Model	2-stream	1-stream
Pretrained	CC	CC+COCO

VQA



RefCOCO+

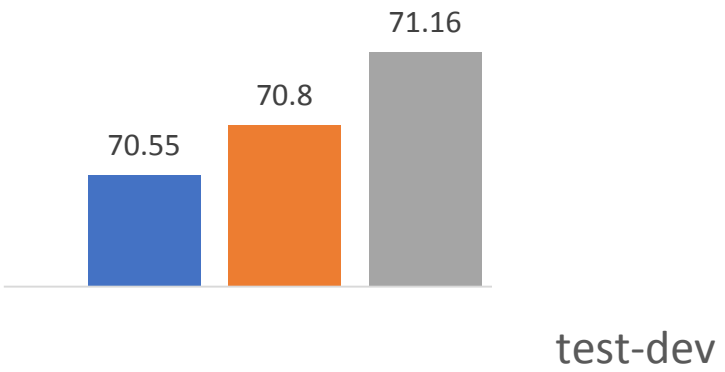


[Li 2019, Tan 2019, Li 2019, Su 2019, Chen 2019]

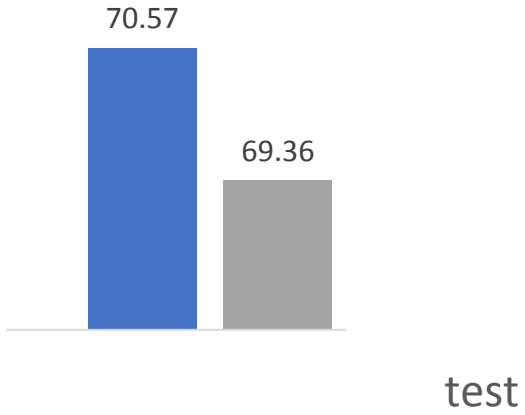
Compare with SOTA

	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream	1-stream	1-stream					
Pretrained	CC	CC+COCO	CC+wiki					

VQA



RefCOCO+

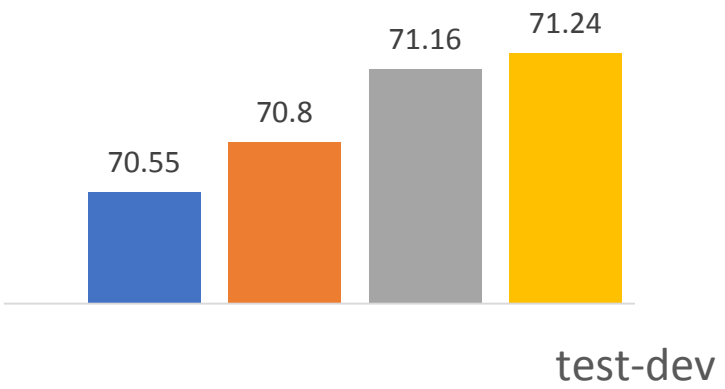


[Li 2019, Tan 2019, Li 2019, Su 2019, Chen 2019]

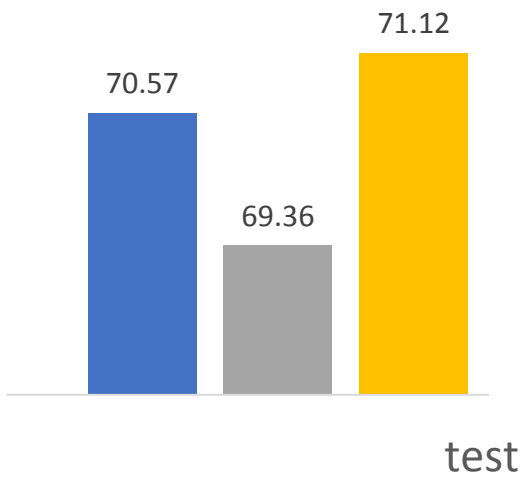
Compare with SOTA

	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream	1-stream	1-stream	2-stream				
Pretrained	CC	CC+COCO	CC+wiki	CC				

VQA



RefCOCO+

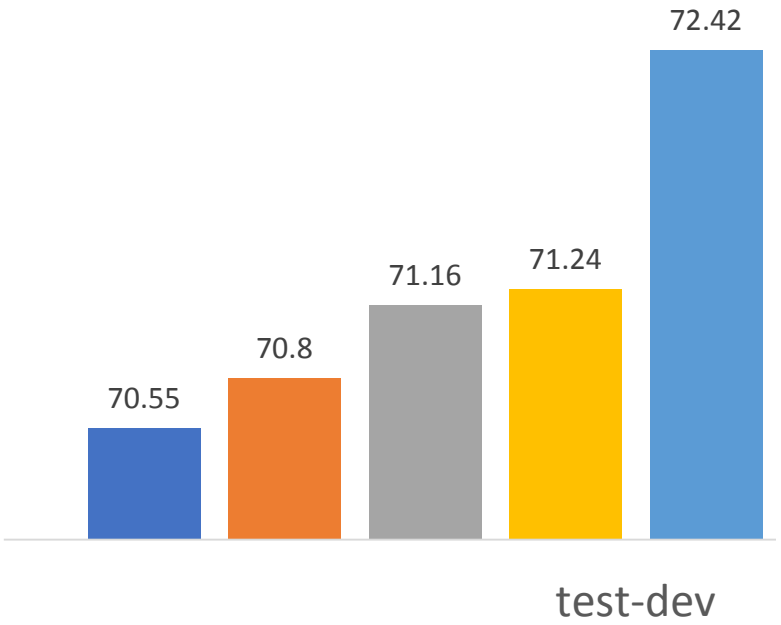


[Li 2019, Tan 2019, Li 2019, Su 2019, Chen 2019]

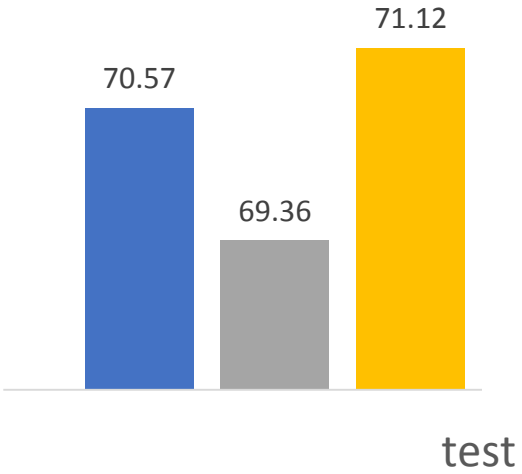
Compare with SOTA

	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream	1-stream	1-stream	2-stream	2-stream			
Pretrained	CC	CC+COCO	CC+wiki	CC	COCO+VG			

VQA



RefCOCO+

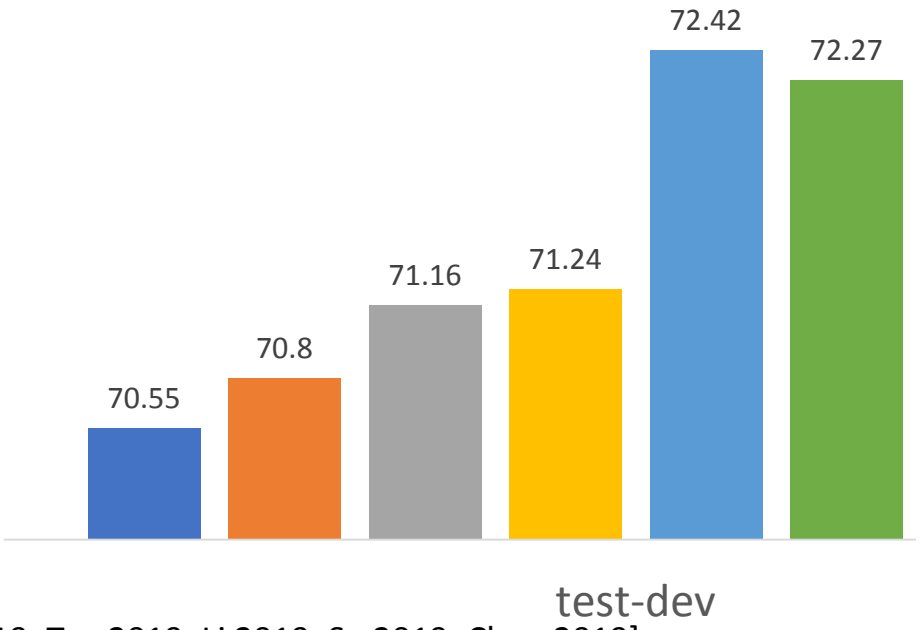


[Li 2019, Tan 2019, Li 2019, Su 2019, Chen 2019]

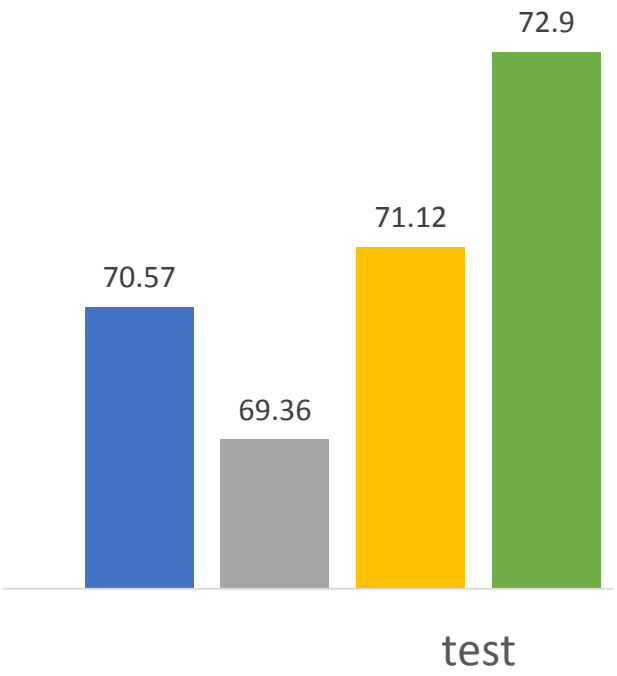
Compare with SOTA

	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream	1-stream	1-stream	2-stream	2-stream	1-stream		
Pretrained	CC	CC+COCO	CC+wiki	CC	COCO+VG	COCO+VG CC+SBU		

VQA



RefCOCO+

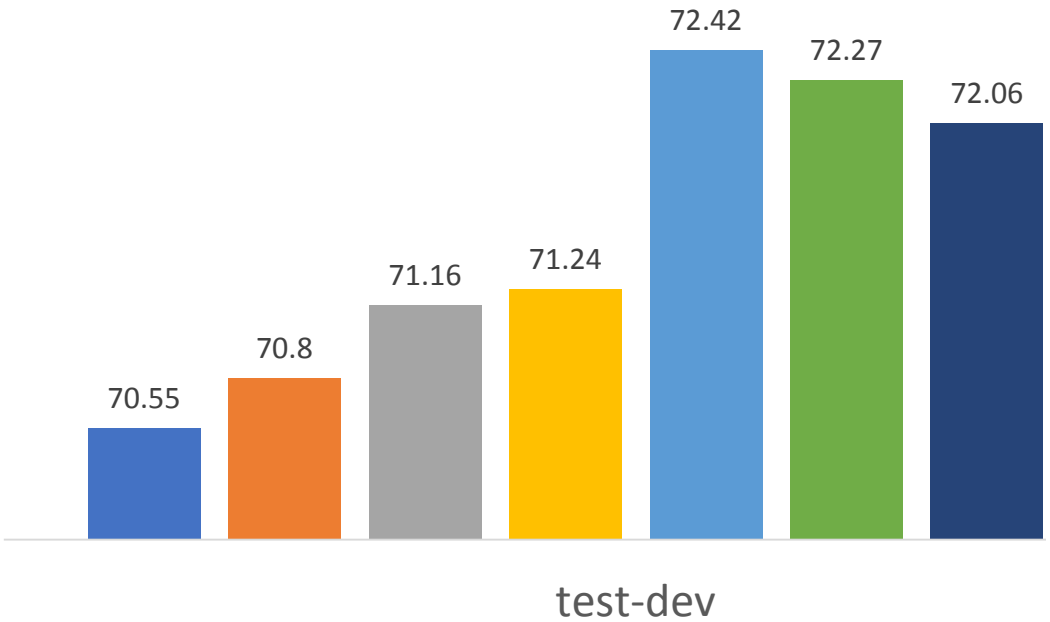


[Li 2019. Tan 2019. Li 2019. Su 2019. Chen 2019]

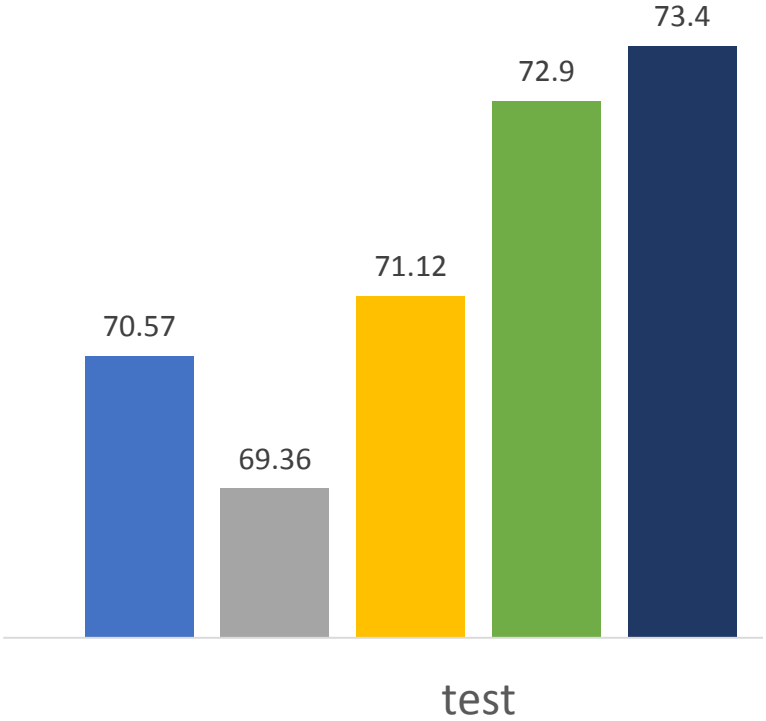
Compare with SOTA

	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream	1-stream	1-stream	2-stream	2-stream	1-stream	2-stream	
Pretrained	CC	CC+COCO	CC+wiki	CC	COCO+VG	COCO+VG CC+SBU	CC	

VQA



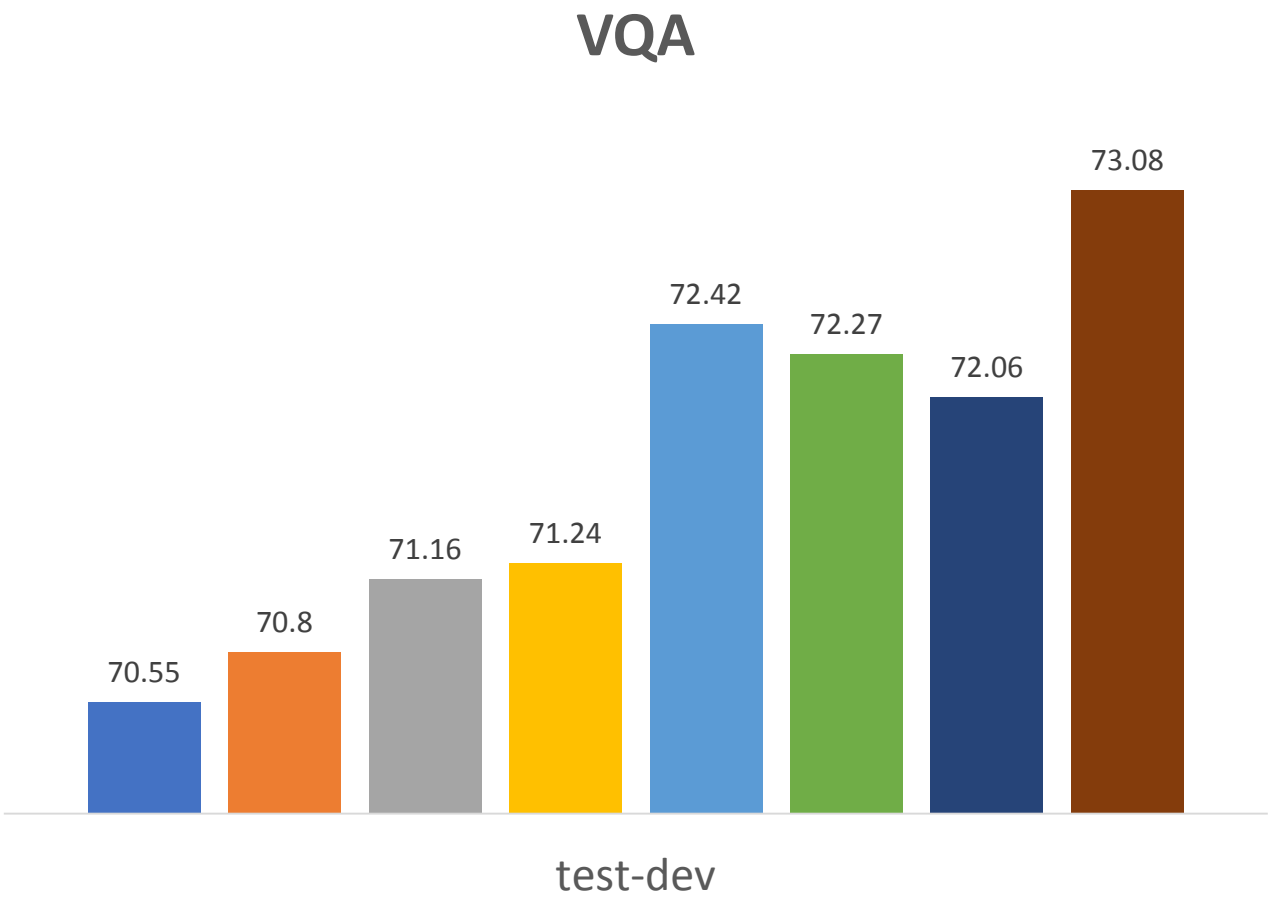
RefCOCO+

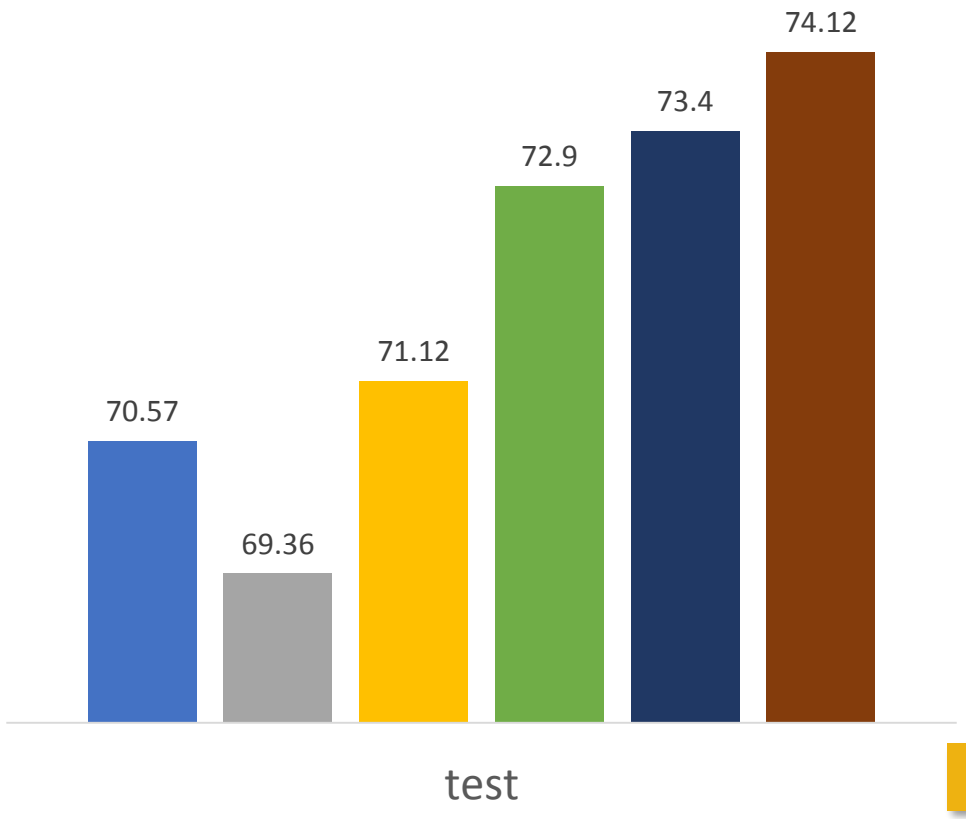


[Li 2019. Tan 2019. Li 2019. Su 2019. Chen 2019]

Compare with SOTA

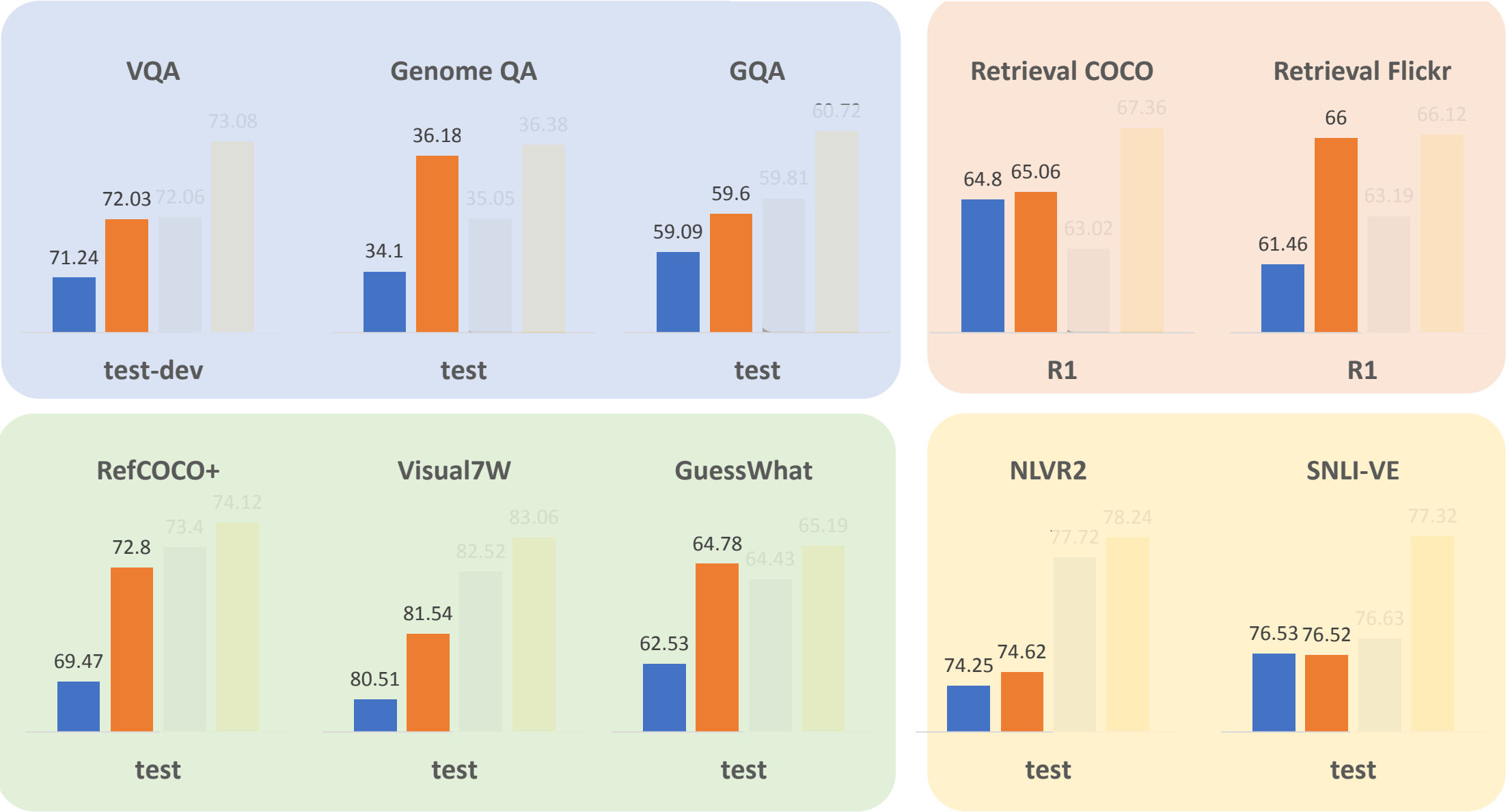
	ViLBERT	VisualBERT	VLBERT	ViLBERT*	LXMERT	UNITER	MT-VL	MT-VL-FT
Model	2-stream	1-stream	1-stream	2-stream	2-stream	1-stream	2-stream	2-stream
Pretrained	CC	CC+COCO	CC+wiki	CC	COCO+VG	COCO+VG CC+SBU	CC	CC + MT



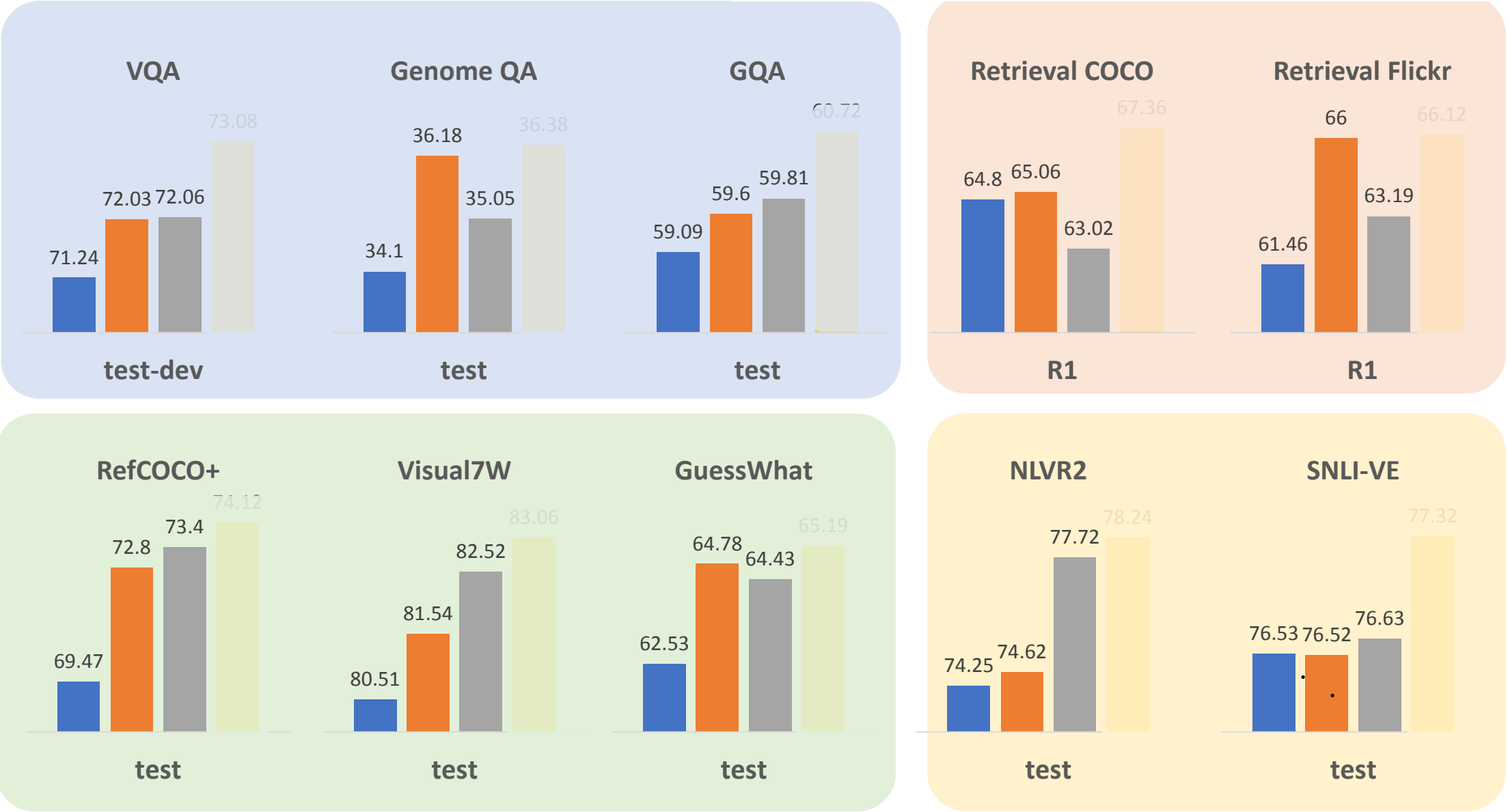


[Li 2019. Tan 2019. Li 2019. Su 2019. Chen 2019]

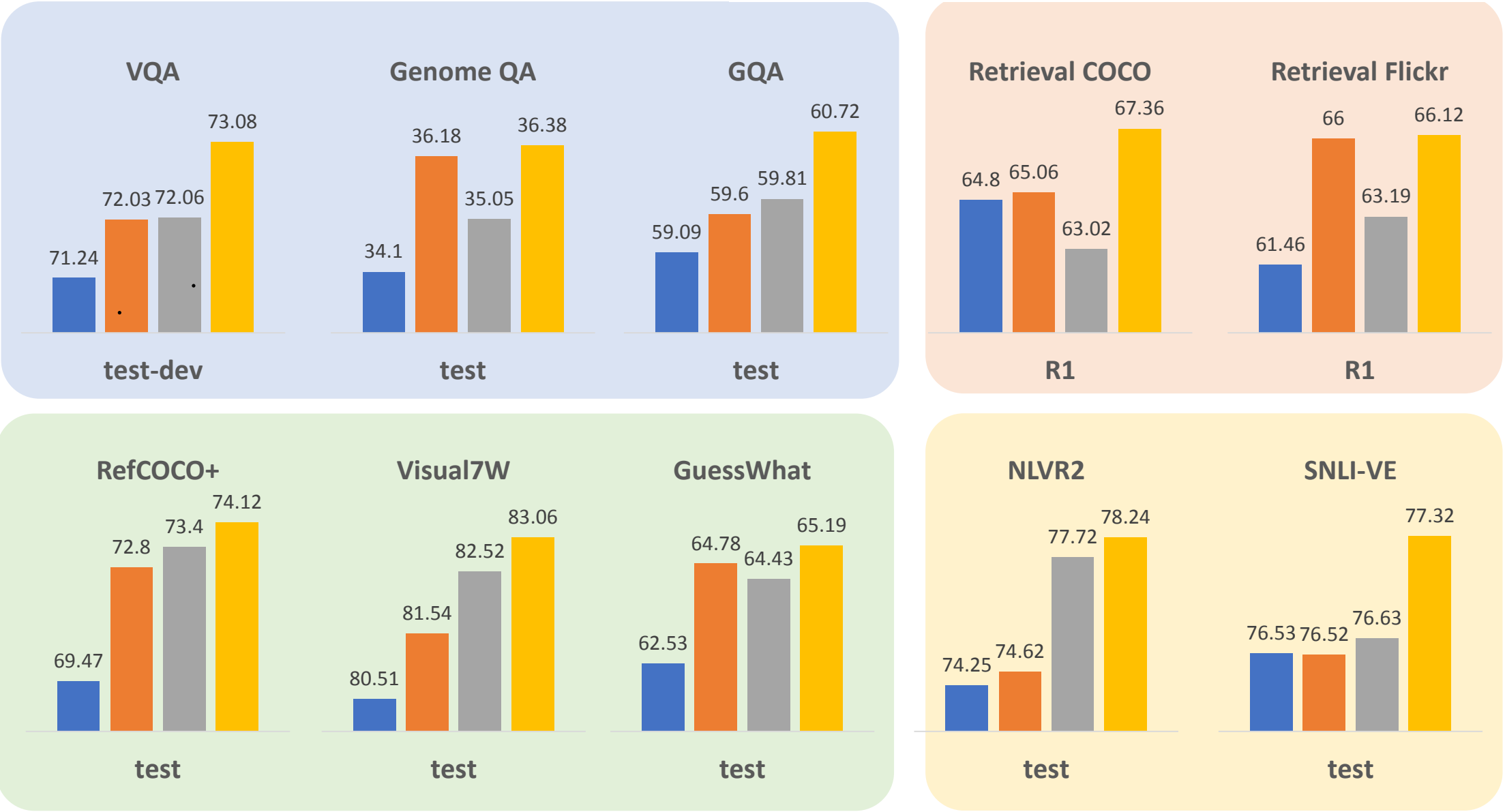
Ablation Study



Ablation Study



Ablation Study



Ablation Study

Task Performance with different Group

Relative Performance of	Trained with				Average
	VQA	**Image Retrieval**	**Refer Expression**	**V&L Verification**	
VQA	-	0.38	0.38	-0.2	0.19
Image Retrieval	0.46	-	0.23	-4.13	-1.15
Refer Expression	0.39	0.78	-	0.24	0.47
V&L Verification	2.29	1.47	0.67	-	1.48
Average	1.04	0.88	0.43	-1.36	-

Ablation Study

Task Performance with different Group

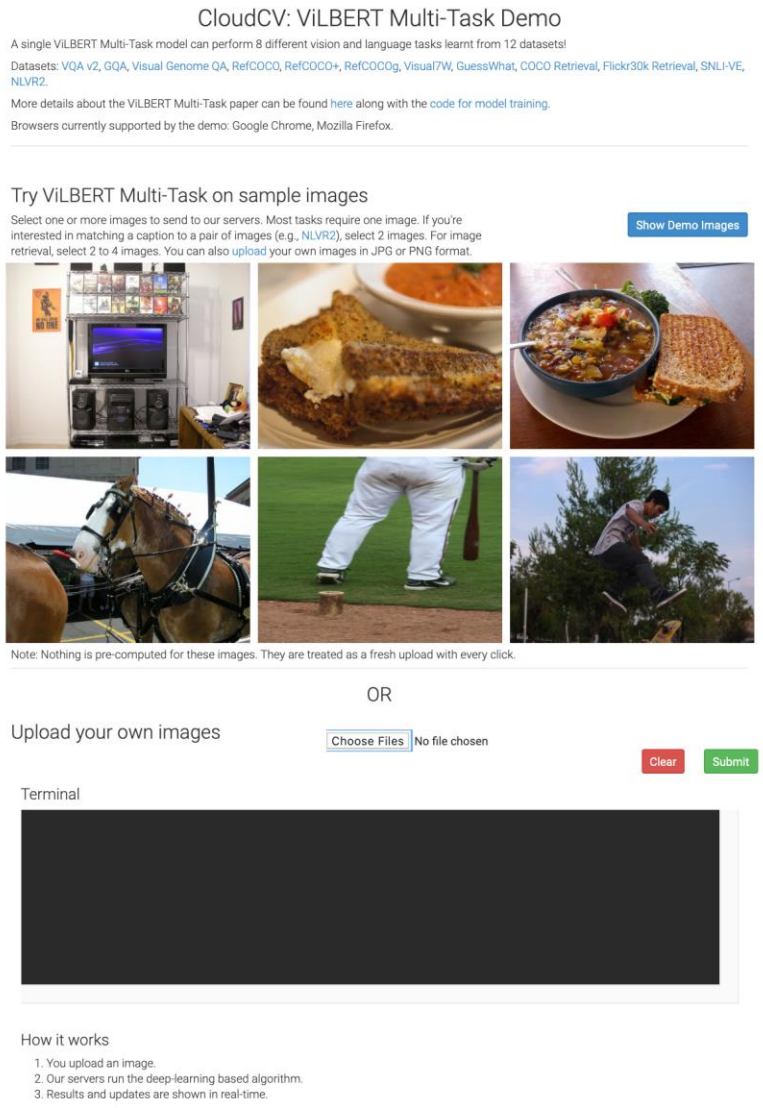
Relative Performance of	Trained with				Average
	VQA	**Image Retrieval**	**Refer Expression**	**V&L Verification**	
VQA	-	0.38	0.38	-0.2	0.19
Image Retrieval	0.46	-	0.23	-4.13	-1.15
Refer Expression	0.39	0.78	-	0.24	0.47
V&L Verification	2.29	1.47	0.67	-	1.48
Average	1.04	0.88	0.43	-1.36	-

Ablation Study

Task Performance with different Group

Relative Performance of	Trained with				Average
	VQA	**Image Retrieval**	**Refer Expression**	**V&L Verification**	
VQA	-	0.38	0.38	-0.2	0.19
Image Retrieval	0.46	-	0.23	-4.13	-1.15
Refer Expression	0.39	0.78	-	0.24	0.47
V&L Verification	2.29	1.47	0.67	-	1.48
Average	1.04	0.88	0.43	-1.36	-

DEMO



https://vilbert.cloudcv.org/

Summary

Summary

Explore multi-tasks vision and language learning.

- Introduce several tricks to improve ViLBERT.
- Add <TSK> tokens to improved the multi-task learning.
- Finetune on the multi-task representation lead to new SOTA.
- Study how different groups connect with each other.

Potential directions

- How to use the information across tasks for XAI?
- How to incorporate more modalities?
- How to make the model smaller and more efficient?
- How to combine symbolic reasoning with representation learning.

The END

Question?