UNITER: UNiversal Image-TExt Representation Learning

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy
Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu

ICLR 2020’

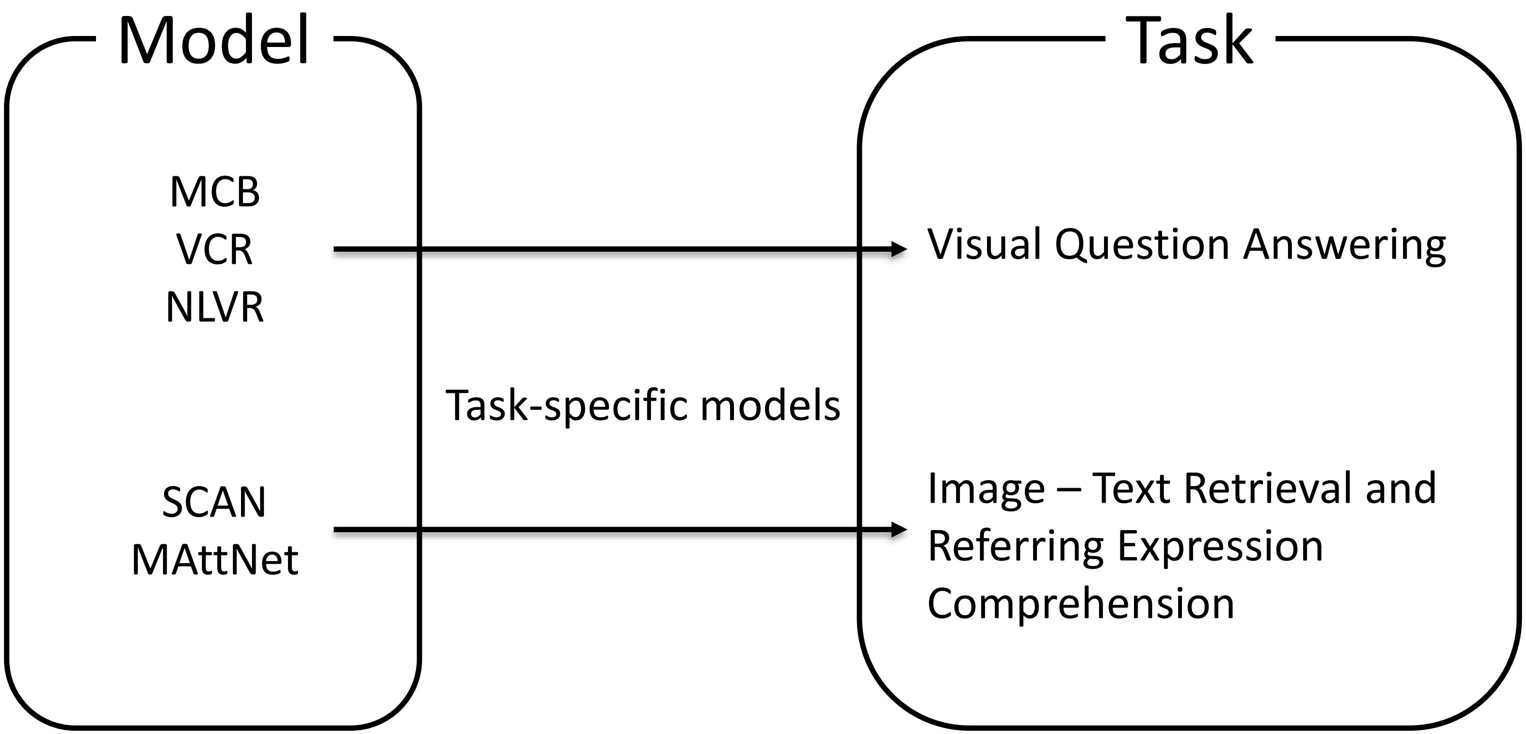
Microsoft Dynamics 365 AI Research

Presented by Dong Hui Im
ehdgnl101@korea.ac.kr
Data Intelligence Laboratory, Korea University
2nd January, 2021

Problem Statement

How to learn a UNiversal Image-TExt Representation for all V + L tasks?

Motivation



Raising a million-dollar question :
Can we learn a universal image-text representation for all V+L tasks?

Motivation

+

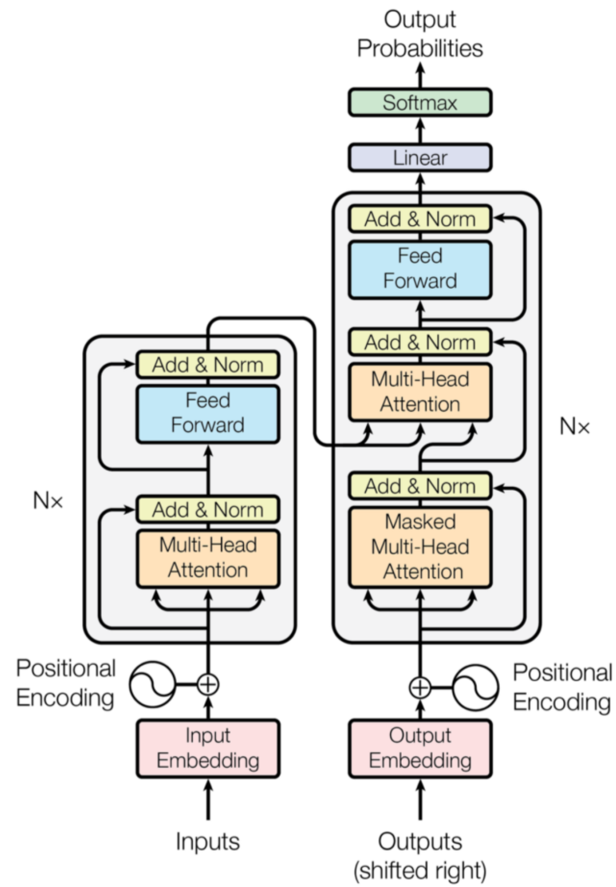


Figure 1: The Transformer - model architecture.

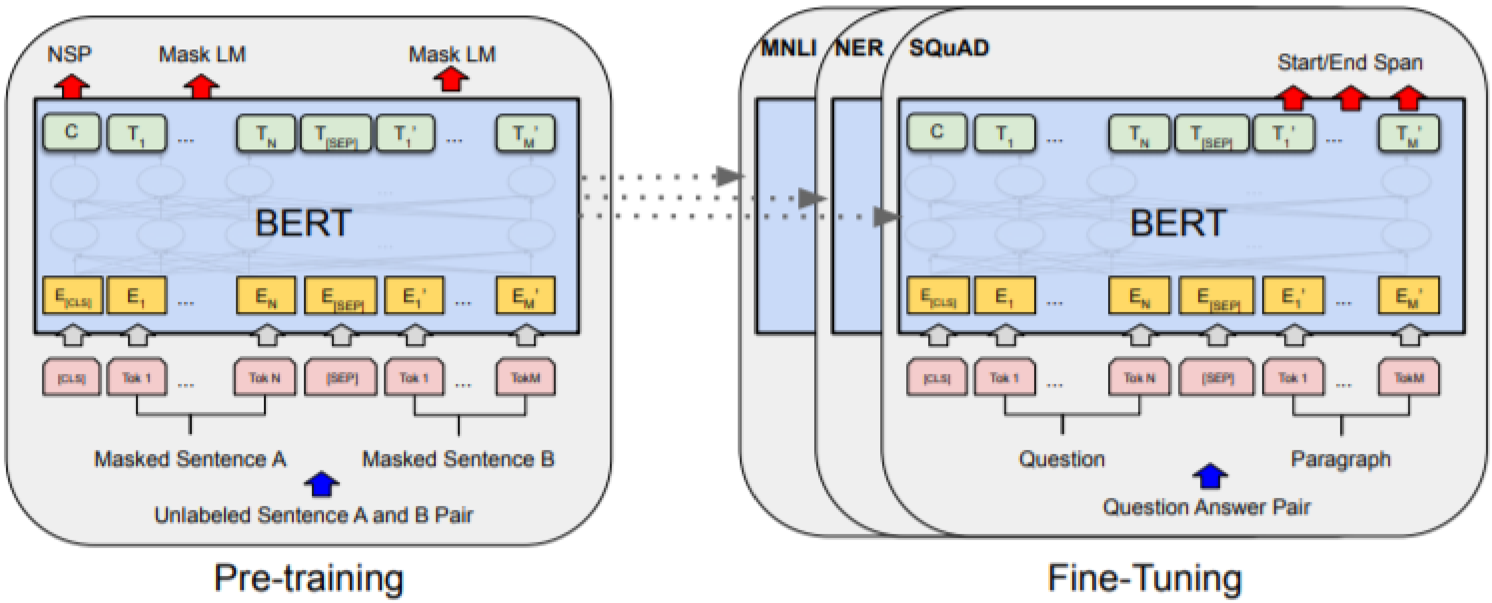
Two keys to great advance in NLP task

1. Effective pre-training task over large language corpus.
2. Use of Transformer for learning contextualized text representations.

Previous work

[Pretrain-then-transfer learning]

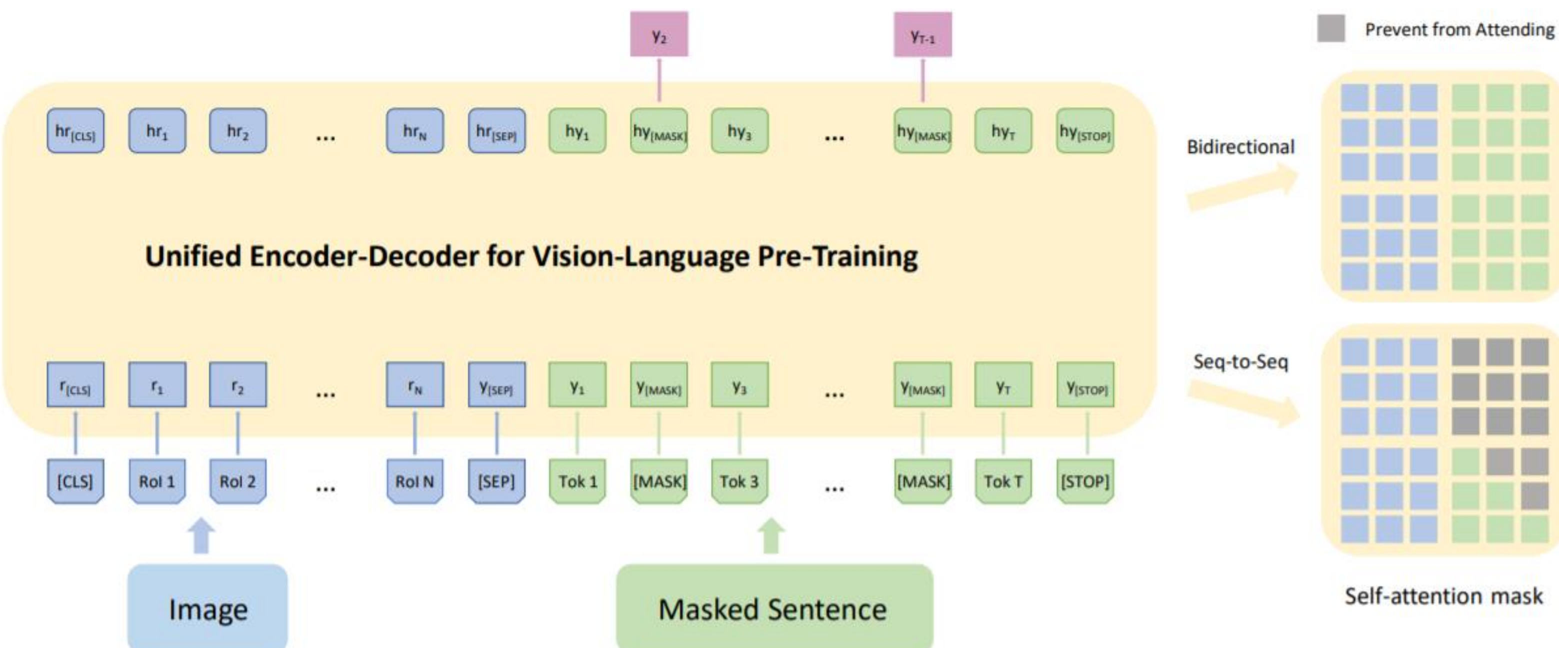
<BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding>



Previous work

[Pretrained model for V + L tasks]

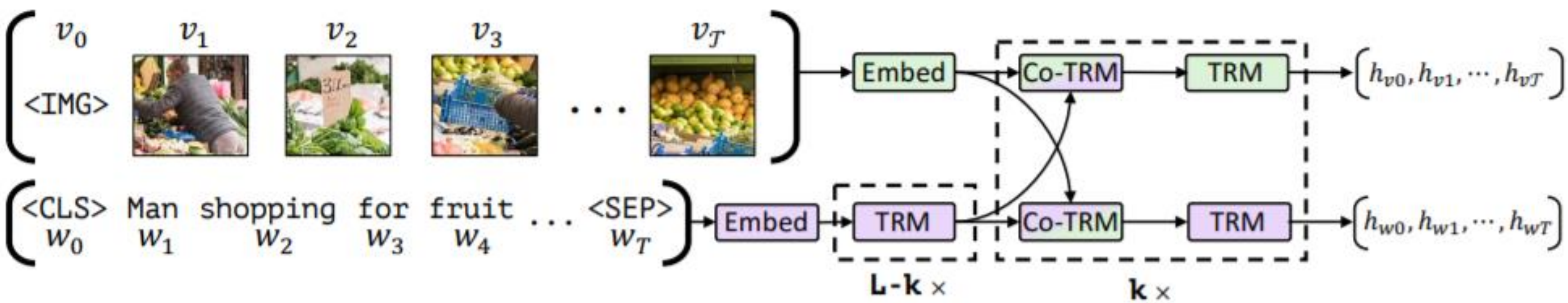
<Unified Vision-Language Pre-Training for Image Captioning and VQA>



VLP applied pre-trained models to both image captioning and VQA.

Related work

Two stream architecture – ViLBERT

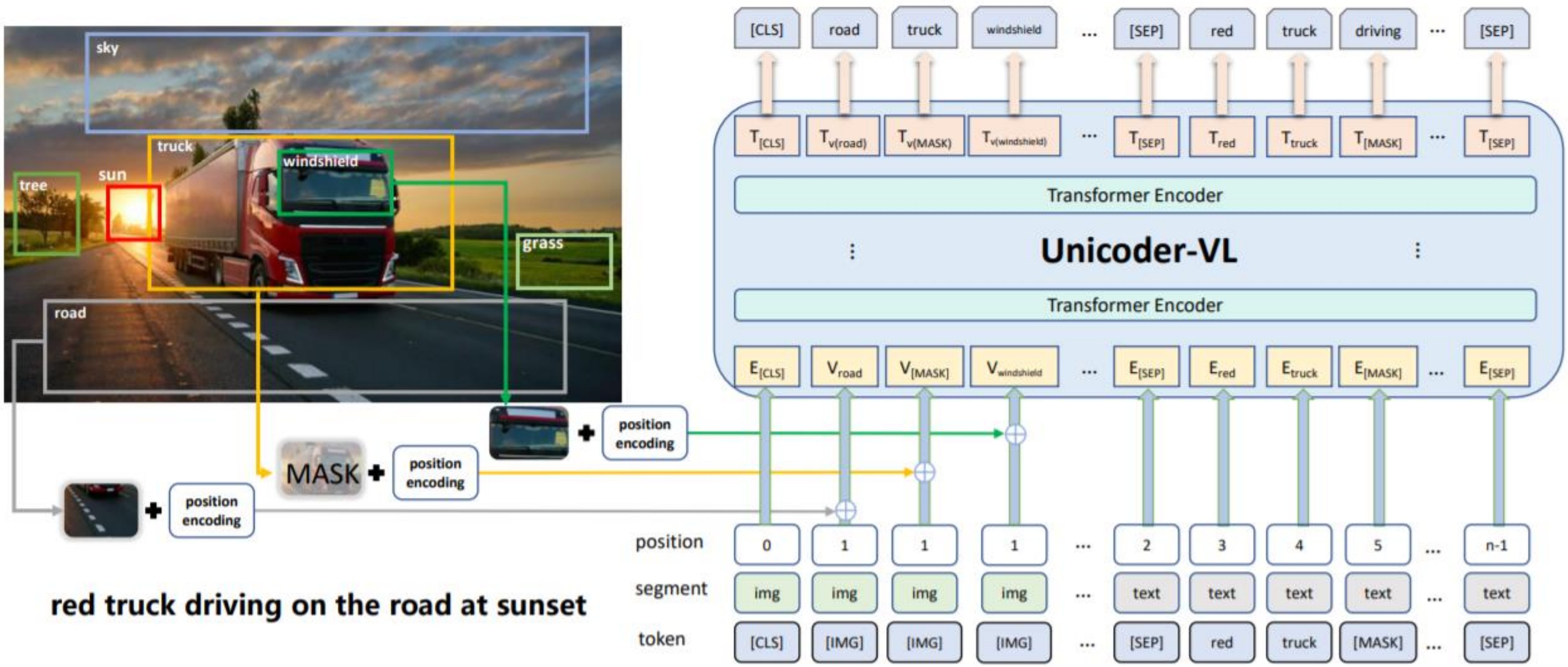


Two separate stream, interact through co-attentional transformer layer.

Pre-trained with Masked Multi-modal modeling and Multi-modal alignment prediction

Related work

Single stream architecture – Unicoder-VL



Pre-trained with Masked Language Modeling,
Masked Object Classification and Visual-linguistic Matching

Contribution

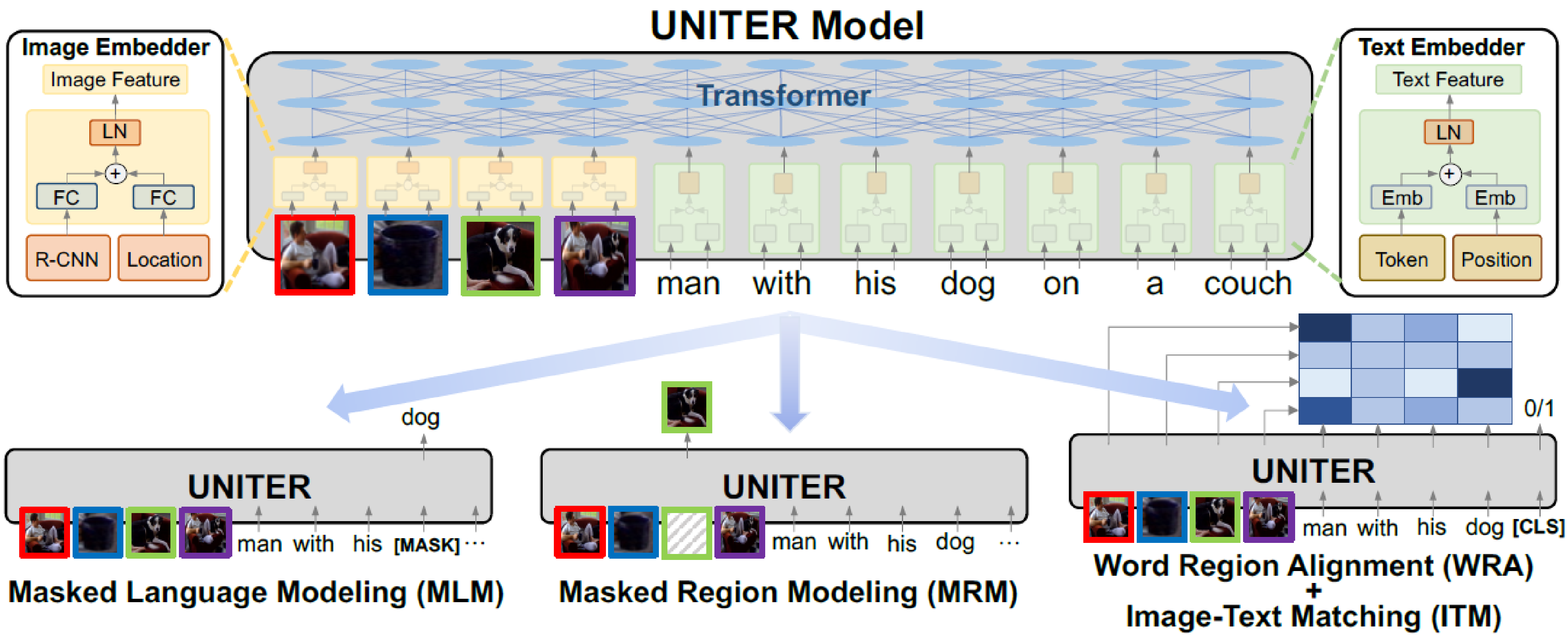
1. They Introduce powerful Universal Image-Text Representation for V + L task.

2. They present conditional Masking for masked language/region modeling and propose a novel optimal-Transport-based Word-region Alignment task for pre-training.

3. They achieve new SOTA on a wide range of V + L benchmarks, outperforming existing multimodal pre-training methods by a large margin.

Model

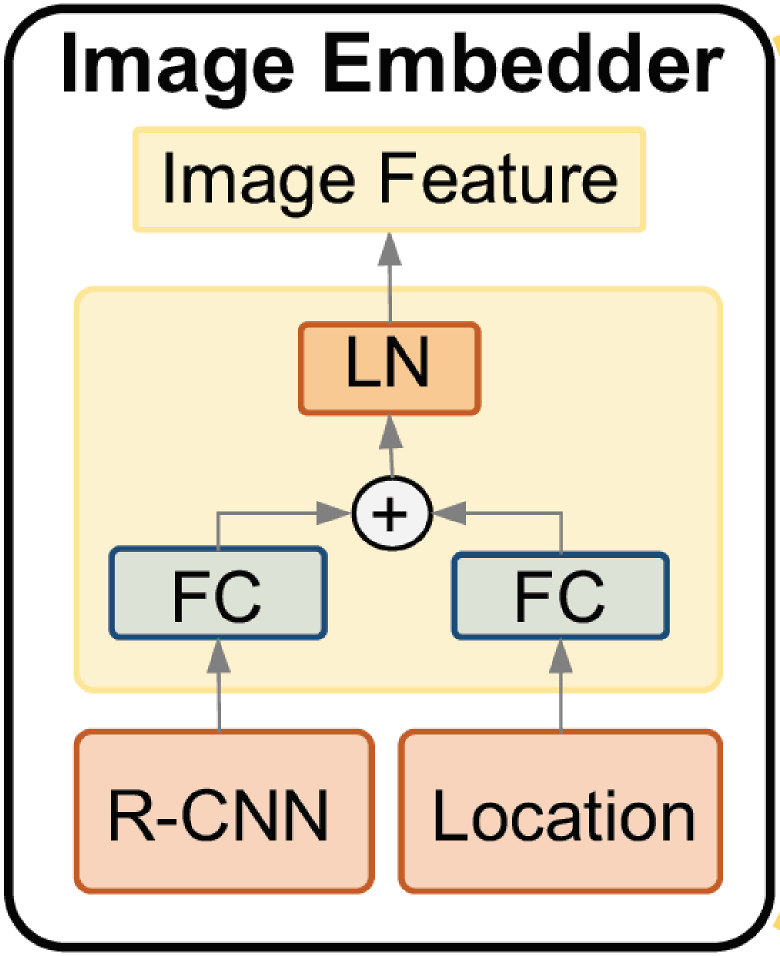
[The entire architecture]



Uniter model consists of Image Embedder, Text Embedder and Transformer module.
Input : image – sentence pair data, with region and words token

Model

[Image Embedder]



Represent location feature
via a 7-dimensional vector

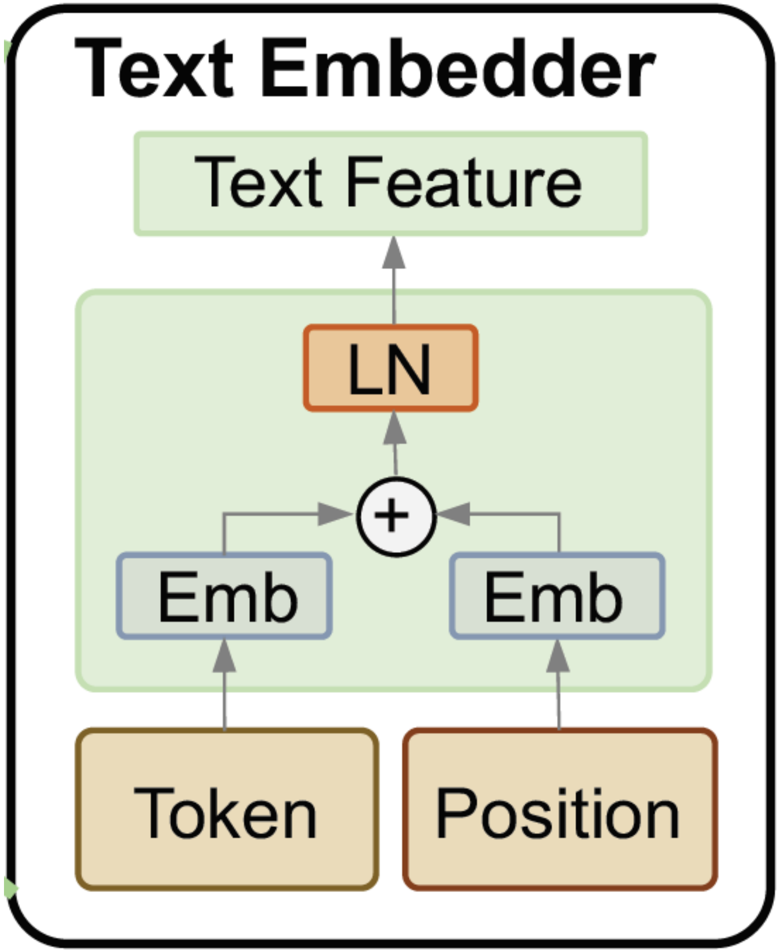
By Fully connected layer,
projected Same embedding space.

Passing through a layer
normalization layer.

Faster R-CNN was pre-trained on Visual
Genome object + attribute data

Model

[Text Embedder]

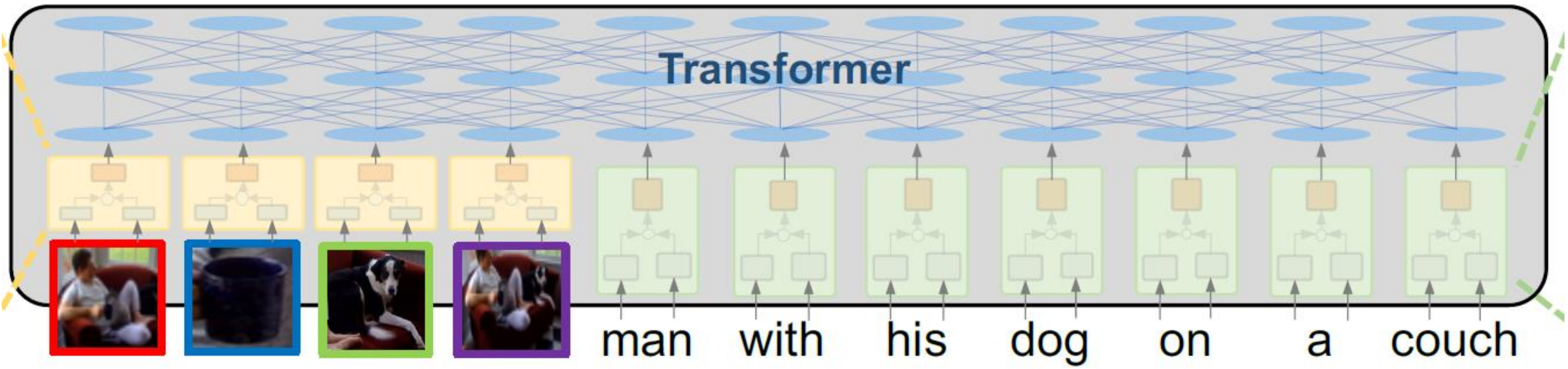


Tokenize the input sentence into Word-Pieces.

Use position feature because transformer is order-less.

Model

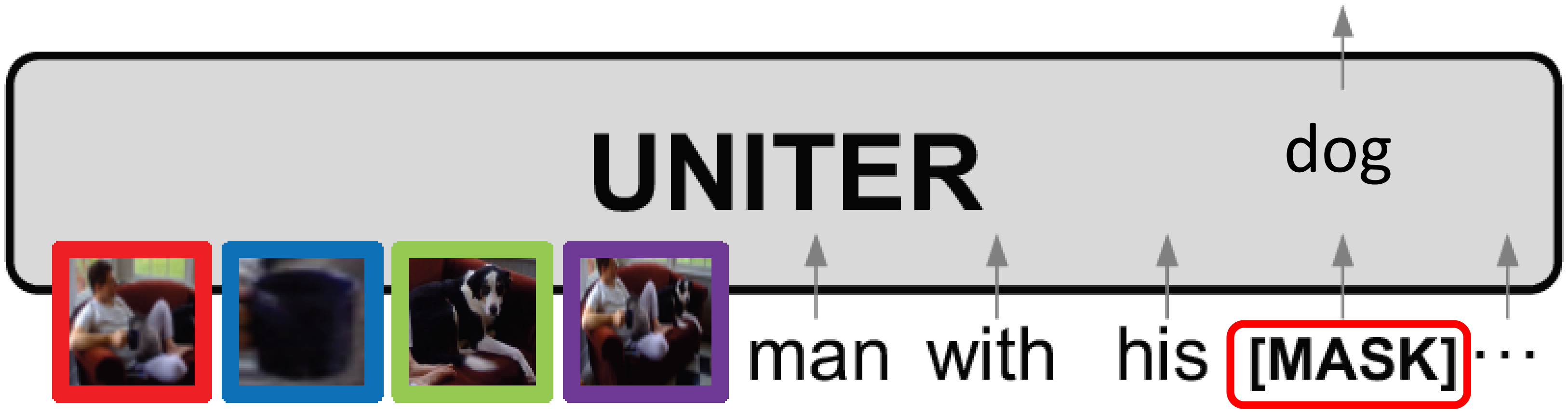
[Transformer]



Randomly sample one task for each mini-batch,
Train on only one objective per SGD update.

Model

[Pre-training Task : MLM]

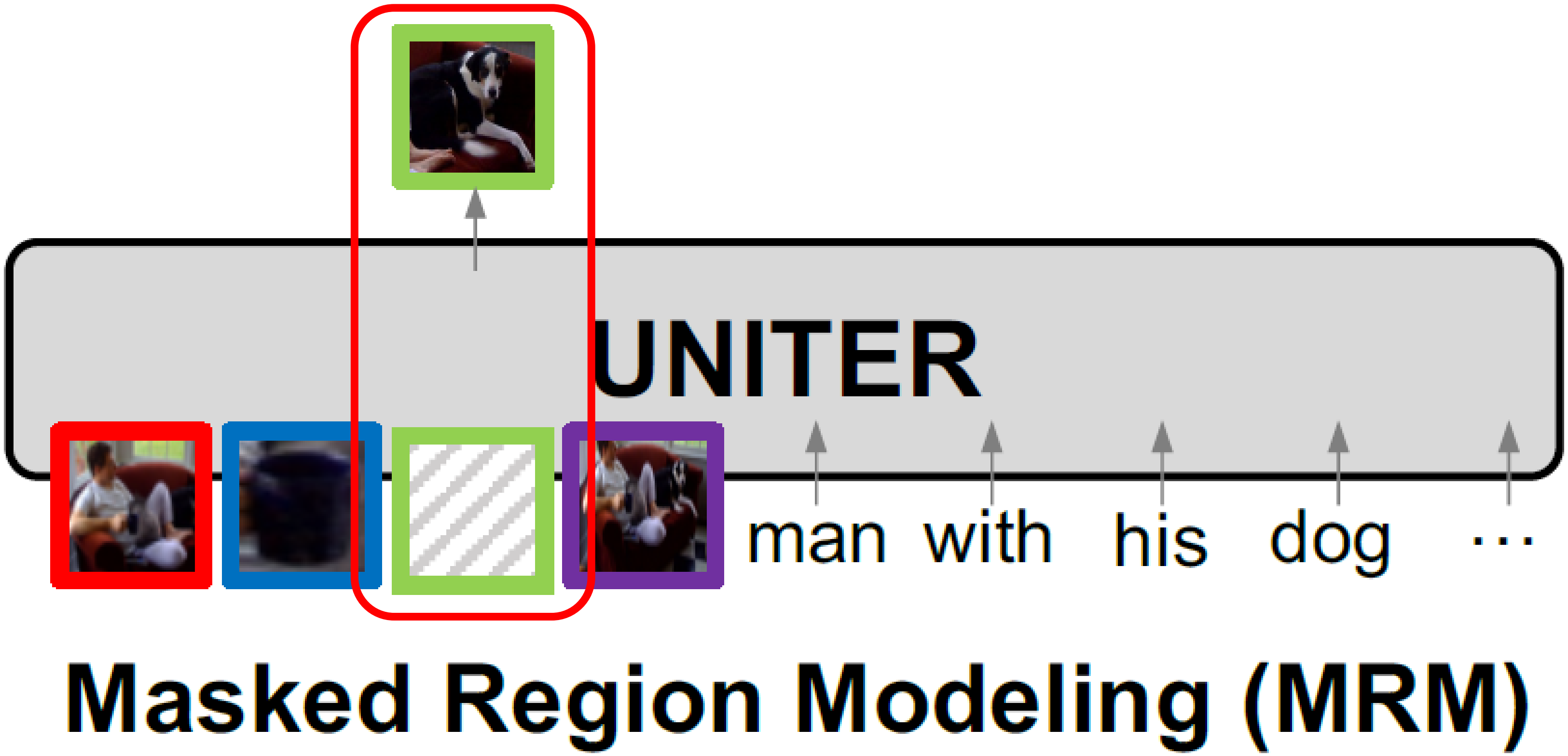


Masked Language Modeling (MLM)

Randomly mask with 15% probability

Model

[Pre-training Task : MRM]



Visual features are high-dimensional and continuous
-> Propose three variants for MRM

Model

[Pre-training Task : MRM]

1. Masked Region Feature Regression(MRFR)

- Predict region feature by L2 regression

2. Masked Region Classification(MRC)

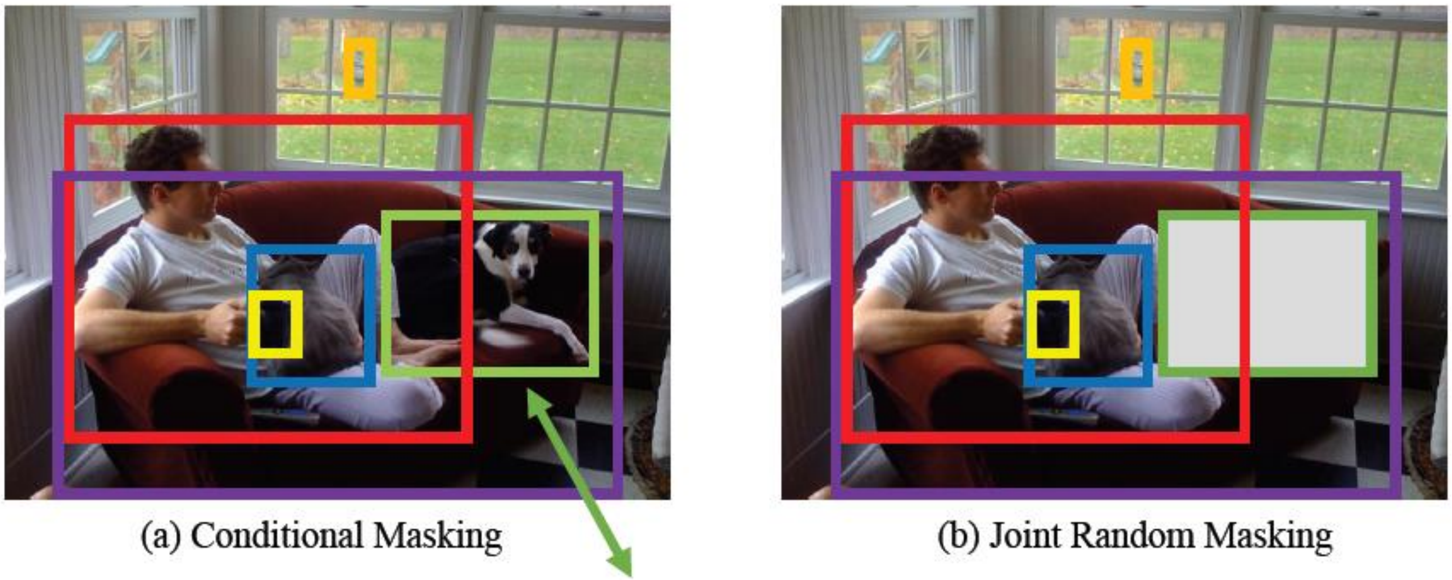
- Predicted as one of K object classes.

3. Masked Region Classification with KL-Divergence(MRC – kl)

- Similar with MRC, but not hard label – (1, 0)
- Minimize the real distribution and Predict distribution

Model

[Pre-training Task : MLM & MRM]



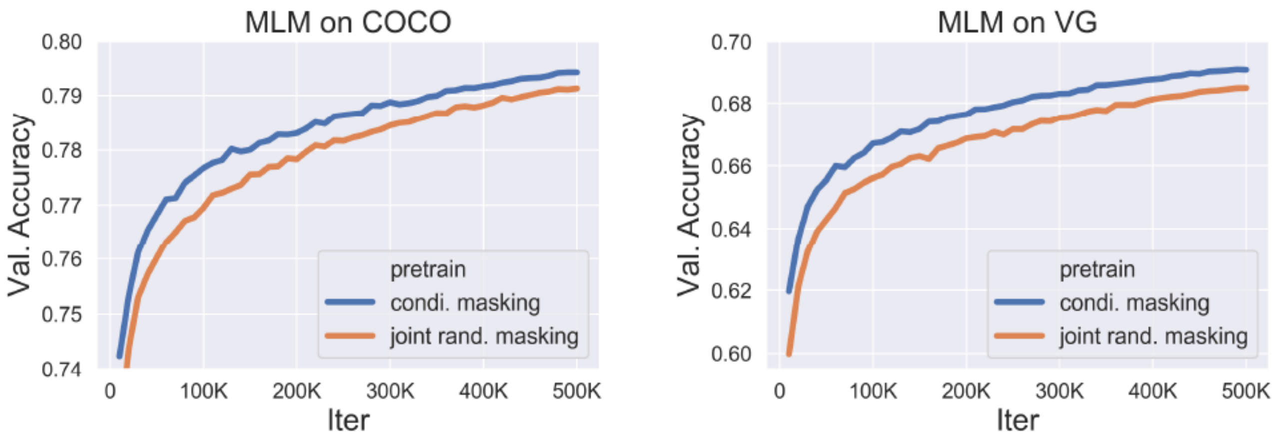
(a) Conditional Masking (b) Joint Random Masking

a man with his <MASK> and cat sitting on the sofa

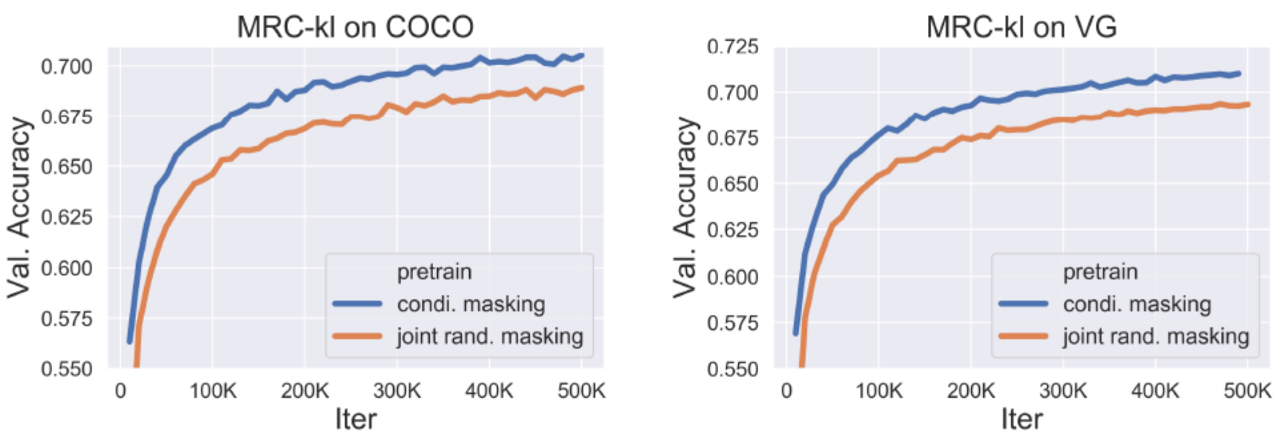
On joint mask random masking,
Problems arise when both the region and image are masked

Model

[Pre-training Task : MLM & MRM]



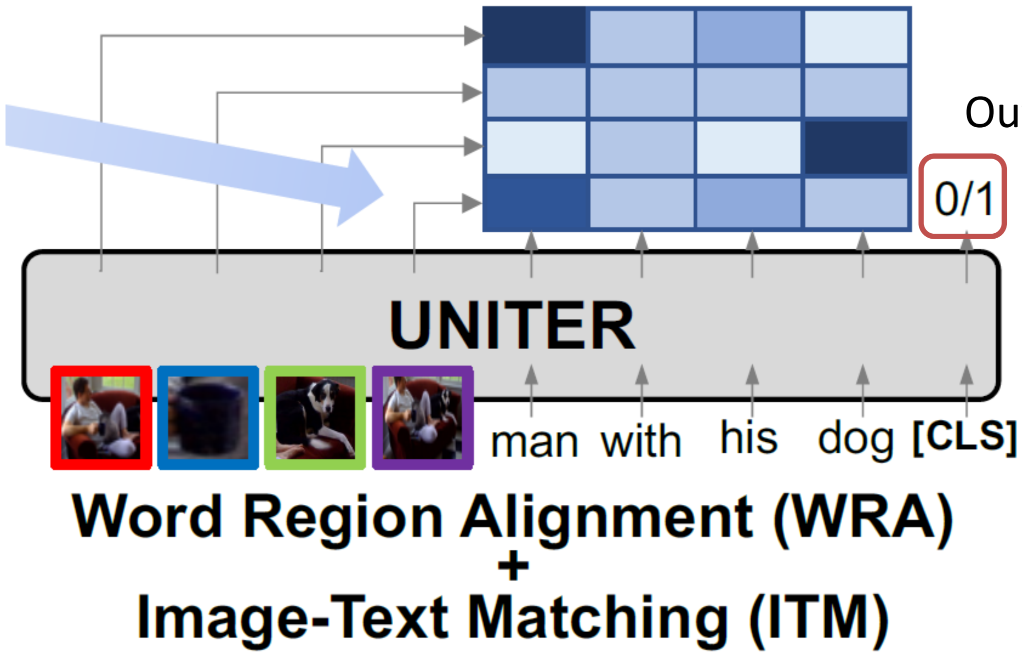
(a) Validation accuracy of MLM on COCO and VG datasets



Val accuracy is higher when using conditional masking.

Model

[Pre-training Task : ITM]



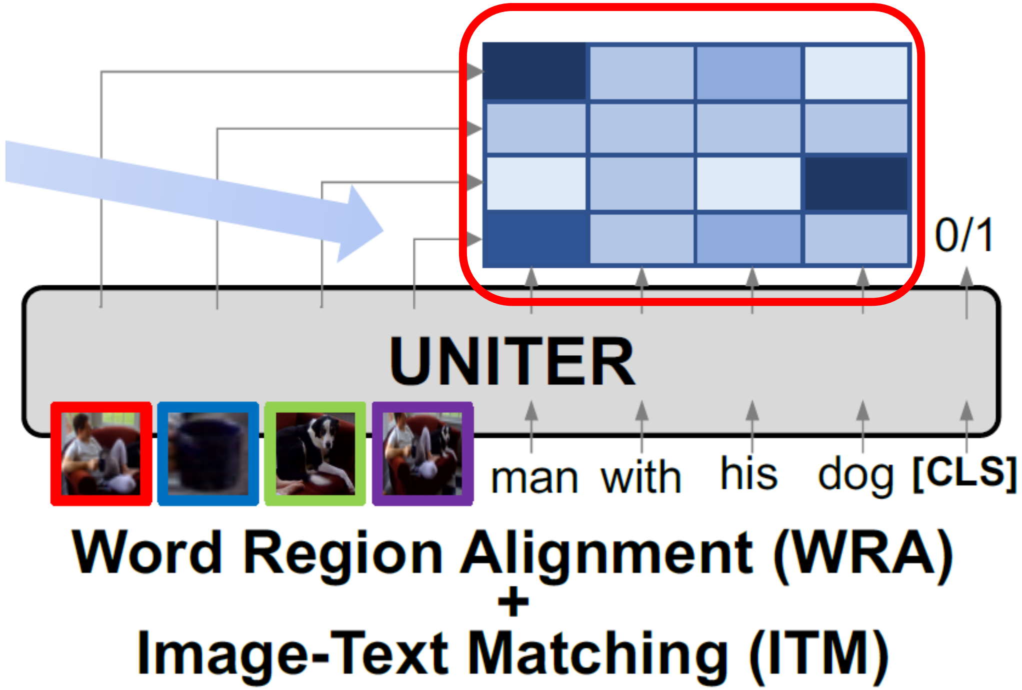
Make sure the given image and text fit well.

Classification as 0/1 across the sigmoid function.

An additional special token [CLS] fed into model.

Model

[Pre-training Task : WRA]



Optimal Transport for WRA

1) Self-normalization : sum of elements of matrix is 1

2) Robust : matrix contains (2r – 1) non – zero elements at most.

3) Efficiency : matrix vector product are efficiency for pre-train.

Q & A

Datasets / Code

[Dataset]

COCO	https://cocodataset.org/#home
Visual Genome	https://visualgenome.org/
Conceptual Captions	https://ai.google.com/research/ConceptualCaptions
SBU Captions	http://www.cs.virginia.edu/~vicente/sbucaptions/
Flickr30k	http://nlp.cs.illinois.edu/
VQA2.0	https://visualqa.org/
VCR	https://visualcommonsense.com/
NLVR	http://lil.nlp.cornell.edu/nlvr/
SNLI-VE	https://github.com/necla-ml/SNLI-VE

[Code]

Publicly available	https://github.com/ChenRocks/UNITER

Setting

1. UNITER-base

L = 12, H = 768, A = 12, Total Parameters = 86M

2. UNITER-large

L = 24, H = 1024, A = 16, Total Parameters = 303M

L : number of stacked Transformer blocks;

H: hidden activation dimension

A: number of attention heads

Baselines

1. ViLBERT

2. VLBERT(Large)
 - Faster R-CNN and Geometry Embedding.

3. Unicoder-VL

4. VisualBERT
 - Used CNN and BERT

5. LXMERT
- 2 stream architecture with cross-modality encoder

6. Task Specific SOTA Models

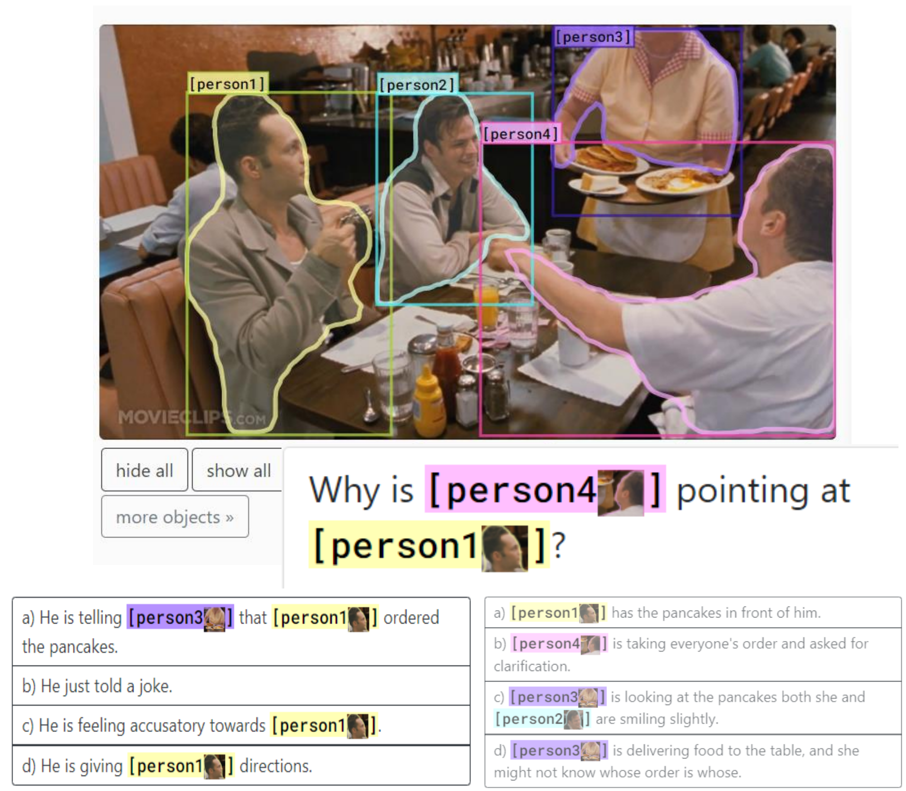
Tasks

[Task 1: Visual Question and Answering(VQA)]

Input : Image and Question

Output : Single word

[Task 2 : Visual Commonsense Reasoning (VCR)]

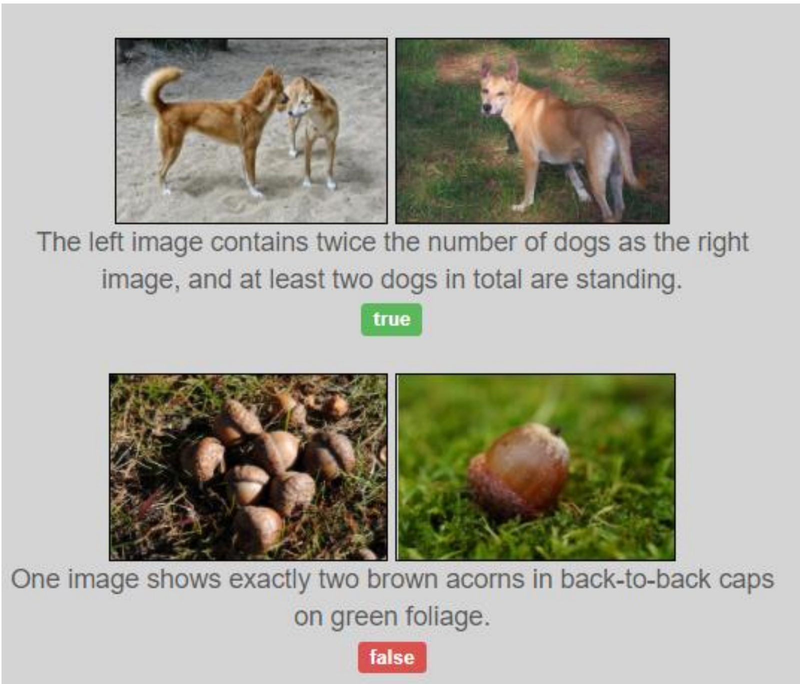


Input : Image + (Question or Q&A)

Output : Answer/Reason/A & R

Tasks

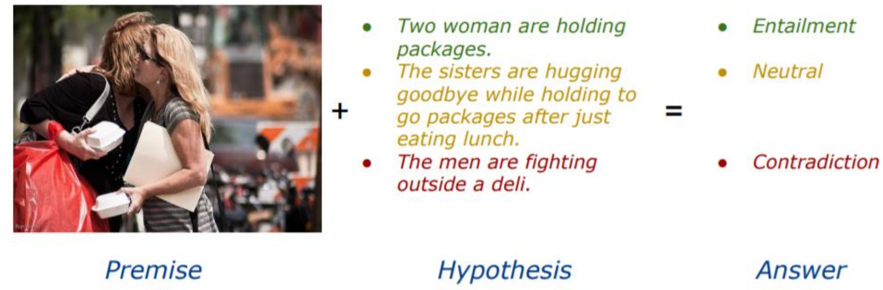
[Task 3 : Natural Language for Visual Reasoning(NLVR)]

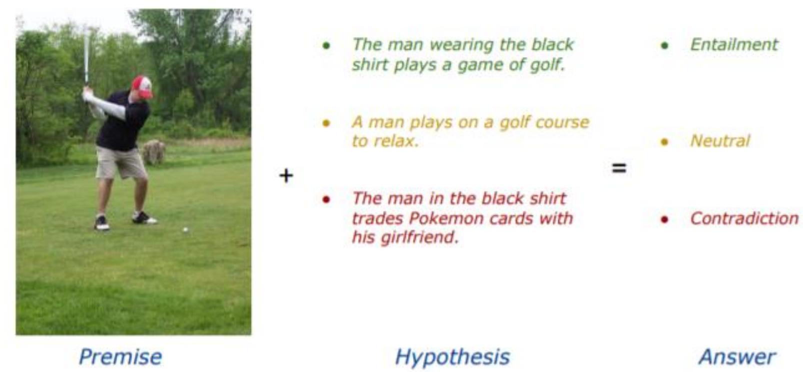


Input : Image and sentence

Output : Judge of correctness of description.

[Task 4 : Stanford Natural Language Inference Visual Entailment(SNLI - VE)]





Input : Image and hypothesis

Output : Answer

Tasks

[Task 5 : Image - Text retrieval]

(a)

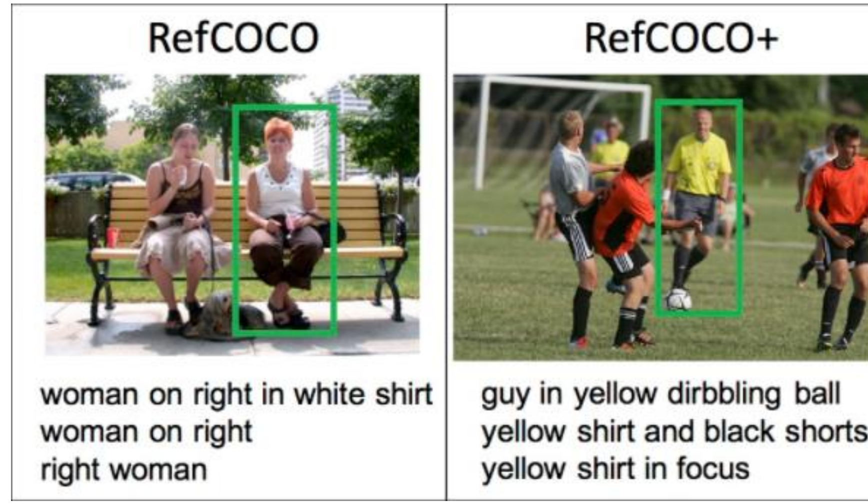
1:Older women and younger girl are opening presents up . ✓
2:Two ladies and a little girl in her pajamas opening gifts ✓
3:A family opening up their Christmas presents . ✓
4:A mother and two children opening gifts on a Christmas morning . ✓
5:A little girl opening a Christmas present . ✓

Query: *A man riding a motorcycle is performing a trick at a track .*

Input : Image and some sentences
Or Sentence and some Images

Output : Answer

[Task 6 : Referring Expression Comprehension]



Input : Reference and image regions

Output : The most relevant image region

Experiments

[Pre-training tasks]

Pre-training Data		Pre-training Tasks	Meta-Sum	VQA	IR (Flickr)	TR (Flickr)	NLVR^2	Ref-COCO+
				test-dev	val	val	dev	val^d
None	1	None	314.34	67.03	61.74	65.55	51.02	68.73
Wikipedia + BookCorpus	2	MLM (text only)	346.24	69.39	73.92	83.27	50.86	68.80
In-domain (COCO+VG)	3	MRFR	344.66	69.02	72.10	82.91	52.16	68.47
	4	ITM	385.29	70.04	78.93	89.91	74.08	72.33
	5	MLM	386.10	71.29	77.88	89.25	74.79	72.89
	6	MLM + ITM	393.04	71.55	81.64	91.12	75.98	72.75
	7	MLM + ITM + MRC	393.97	71.46	81.39	91.45	76.18	73.49
	8	MLM + ITM + MRFR	396.24	71.73	81.76	92.31	76.21	74.23
	9	MLM + ITM + MRC-kl	397.09	71.63	82.10	92.57	76.28	74.51
	10	MLM + ITM + MRC-kl + MRFR	399.97	71.92	83.73	92.87	76.93	74.52
	11	MLM + ITM + MRC-kl + MRFR + WRA	400.93	72.47	83.72	93.03	76.91	74.80
	12	MLM + ITM + MRC-kl + MRFR (w/o cond. mask)	396.51	71.68	82.31	92.08	76.15	74.29
Out-of-domain (SBU+CC)	13	MLM + ITM + MRC-kl + MRFR + WRA	396.91	71.56	84.34	92.57	75.66	72.78
In-domain + Out-of-domain	14	MLM + ITM + MRC-kl + MRFR + WRA	**405.24**	**72.70**	**85.77**	**94.28**	**77.18**	**75.31**

With Out of domain dataset, Meta-Sum score is larger than
When use only In-domain dataset.

Experiments

[Down-stream tasks]

Tasks		SOTA	ViLBERT	VLBERT (Large)	Unicoder -VL	VisualBERT	LXMERT	UNITER Base	UNITER Large
VQA	test-dev	70.63	70.55	71.79	-	70.80	72.42	72.70	**73.82**
	test-std	70.90	70.92	72.22	-	71.00	72.54	72.91	**74.02**
VCR	Q→A	72.60	73.30	75.80	-	71.60	-	75.00	**77.30**
	QA→R	75.70	74.60	78.40	-	73.20	-	77.20	**80.80**
	Q→AR	55.00	54.80	59.70	-	52.40	-	58.20	**62.80**
NLVR^2	dev	54.80	-	-	-	67.40	74.90	77.18	**79.12**
	test-P	53.50	-	-	-	67.00	74.50	77.85	**79.98**
SNLI-VE	val	71.56	-	-	-	-	-	78.59	**79.39**
	test	71.16	-	-	-	-	-	78.28	**79.38**
ZS IR (Flickr)	R@1	-	31.86	-	48.40	-	-	66.16	**68.74**
	R@5	-	61.12	-	76.00	-	-	88.40	**89.20**
	R@10	-	72.80	-	85.20	-	-	92.94	**93.86**
IR (Flickr)	R@1	48.60	58.20	-	71.50	-	-	72.52	**75.56**
	R@5	77.70	84.90	-	91.20	-	-	92.36	**94.08**
	R@10	85.20	91.52	-	95.20	-	-	96.08	**96.76**
IR (COCO)	R@1	38.60	-	-	48.40	-	-	50.33	**52.93**
	R@5	69.30	-	-	76.70	-	-	78.52	**79.93**
	R@10	80.40	-	-	85.90	-	-	87.16	**87.95**

UNITER achieve SOTA for all V+L Tasks.
UNITER-base model outperforms SOTA by approximately +2.8% for VCR on Q -> AR, +2.5% for NLVR$_2$, +7% for SNLI-VE, +4% on R@1 for Image-Text Retrieval (+15% for zero-shot setting), and +2% for RE Comprehension.

Experiments

[Down-stream tasks]

Tasks		SOTA	ViLBERT	VLBERT (Large)	Unicoder -VL	VisualBERT	LXMERT	UNITER Base	UNITER Large
ZS TR (Flickr)	R@1	-	-	-	64.30	-	-	80.70	**83.60**
	R@5	-	-	-	85.80	-	-	**95.70**	**95.70**
	R@10	-	-	-	92.30	-	-	**98.00**	97.70
TR (Flickr)	R@1	67.90	-	-	86.20	-	-	85.90	**87.30**
	R@5	90.30	-	-	96.30	-	-	97.10	**98.00**
	R@10	95.80	-	-	99.00	-	-	98.80	**99.20**
TR (COCO)	R@1	50.40	-	-	62.30	-	-	64.40	**65.68**
	R@5	82.20	-	-	87.10	-	-	87.40	**88.56**
	R@10	90.00	-	-	92.80	-	-	93.08	**93.76**
Ref-COCO	val	87.51		-	-	-	-	91.64	**91.84**
	testA	89.02	-	-	-	-	-	92.26	**92.65**
	testB	87.05	-	-	-	-	-	90.46	**91.19**
	val^d	77.48	-	-	-	-	-	81.24	**81.41**
	$testA^d$	83.37	-	-	-	-	-	86.48	**87.04**
	$testB^d$	70.32	-	-	-	-	-	73.94	**74.17**
Ref-COCO+	val	75.38	-	80.31	-	-	-	83.66	**84.25**
	testA	80.04	-	83.62	-	-	-	86.19	**86.34**
	testB	69.30	-	75.45	-	-	-	78.89	**79.75**
	val^d	68.19	72.34	72.59	-	-	-	75.31	**75.90**
	$testA^d$	75.97	78.52	78.57	-	-	-	81.30	**81.45**
	$testB^d$	57.52	62.61	62.30	-	-	-	65.58	**66.70**
Ref-COCOg	val	81.76	-	-	-	-	-	86.52	**87.85**
	test	81.75	-	-	-	-	-	86.52	**87.73**
	val^d	68.22	-	-	-	-	-	74.31	**74.86**
	$test^d$	69.46	-	-	-	-	-	74.51	**75.77**

Despite that UNITER is single stream model with a fewer parameter than two-stream model, it achieved SOTA for many Tasks.

Conclusions

1. They present large-scale pre-trained model providing UNiversal Image-TExt Representations for Vision-and-Language tasks

2. Four main pre-training tasks are proposed and evaluated through extensive ablation studies.

3. Trained with both in-domain and out-of-domain datasets, UNITER outperforms state-of-the-art models over multiple V+L tasks by a significant margin.

Q & A