Vision and Language Pretrained Models

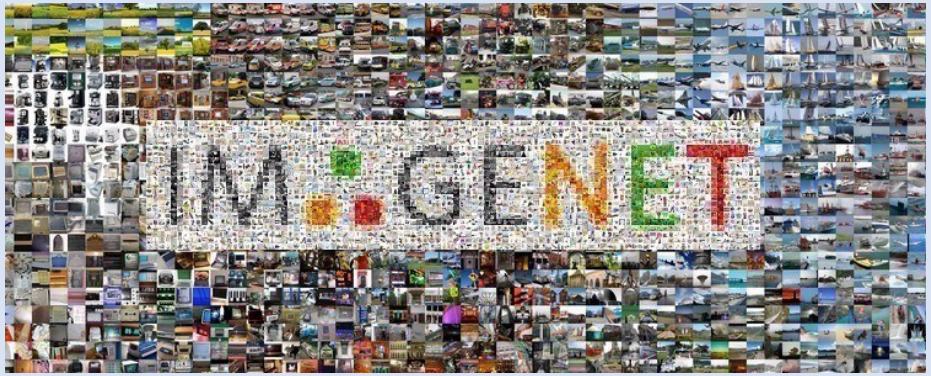
Luke Ye

52194506006

Outline

- Background
  - Pretrained Model
  - Vision and Language Pretrain
- Recent Advances in Vision and Language Pretrained Models
- Summary and Thinking

Pretrained Model



Object Detection

Instance Segmentation

Semantic Segmentation

...

BERT (Devlin et al, 2018)

Wikipedia

BookCorpus

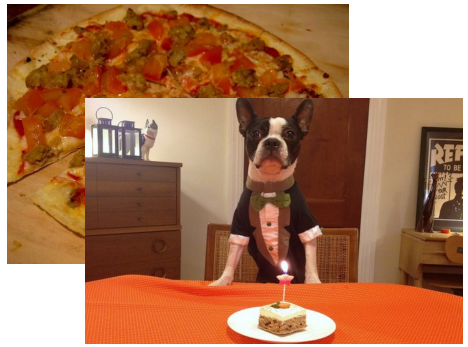
Question Answering

Natural Language Inference

Sentence Classification

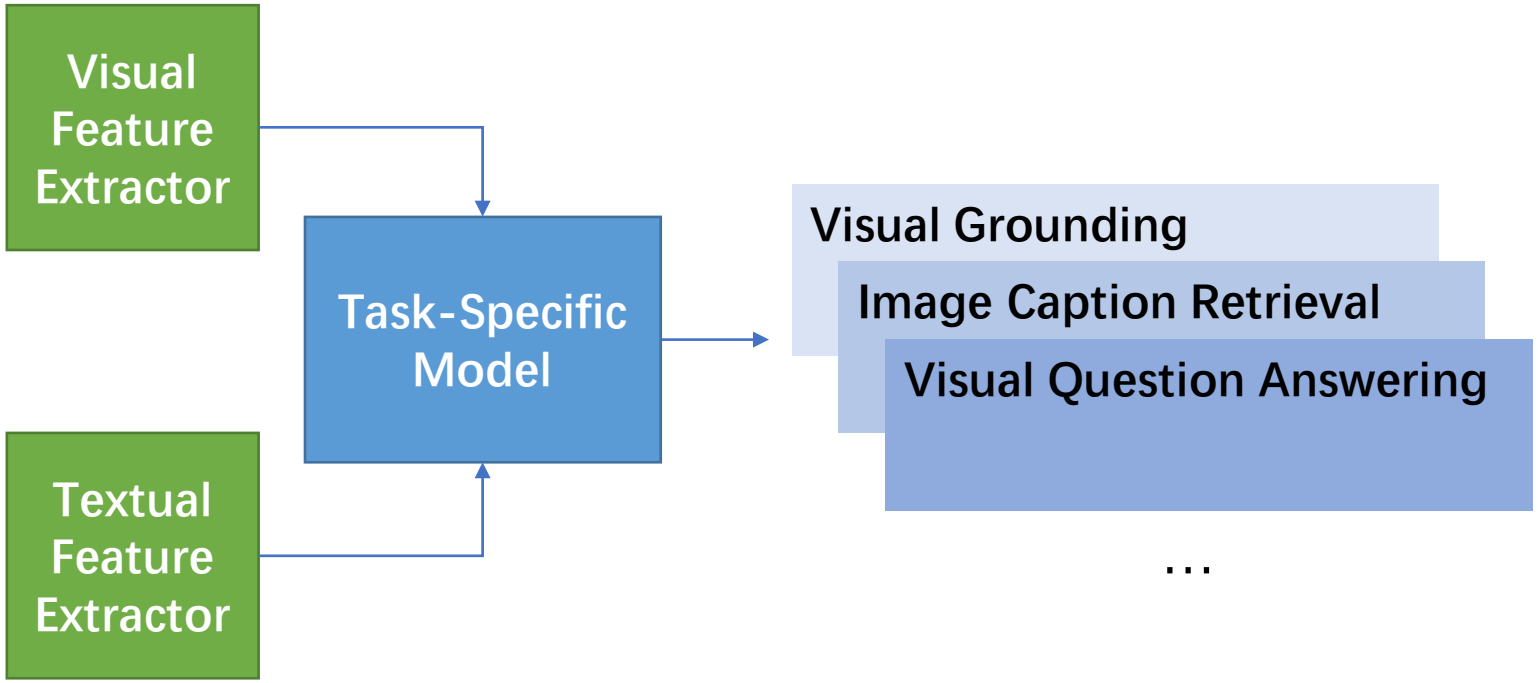
...

Vision and Language Pretrain



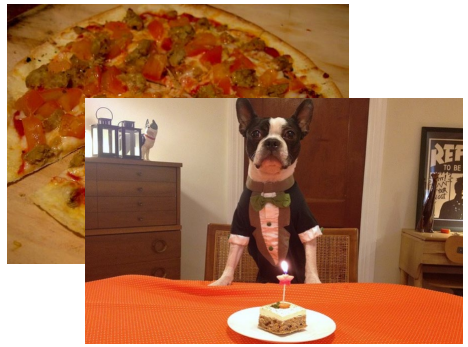
How many slices of pizza are there?

A puppy with a tie is sitting at table with a cake.



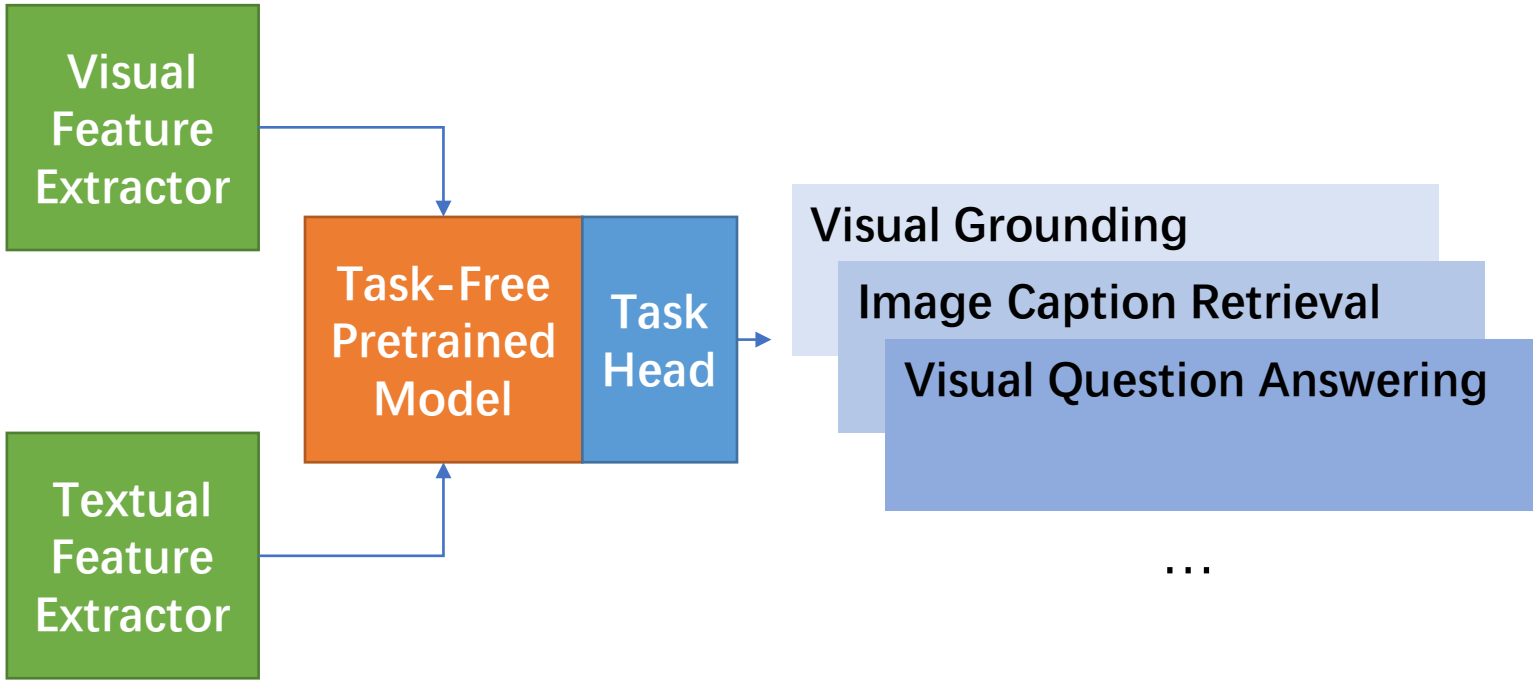
Background

Vision and Language Pretrain



How many slices of pizza are there?

A puppy with a tie is sitting at table with a cake.



ViLBERT

Data

- Conceptual Captions

3.3 *million image-caption pairs automatically scraped from alt-text enabled web images*

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.

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.

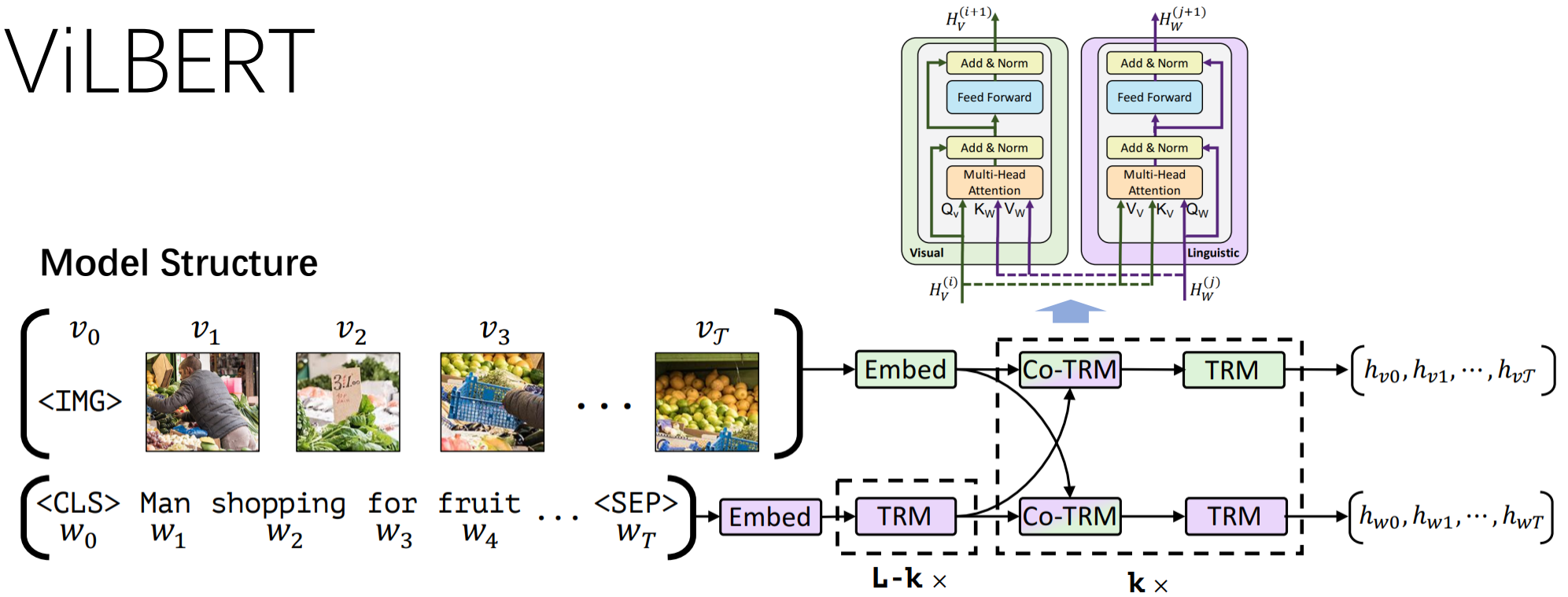
Task

- Masked multi-modal modelling
- Multi-modal alignment prediction

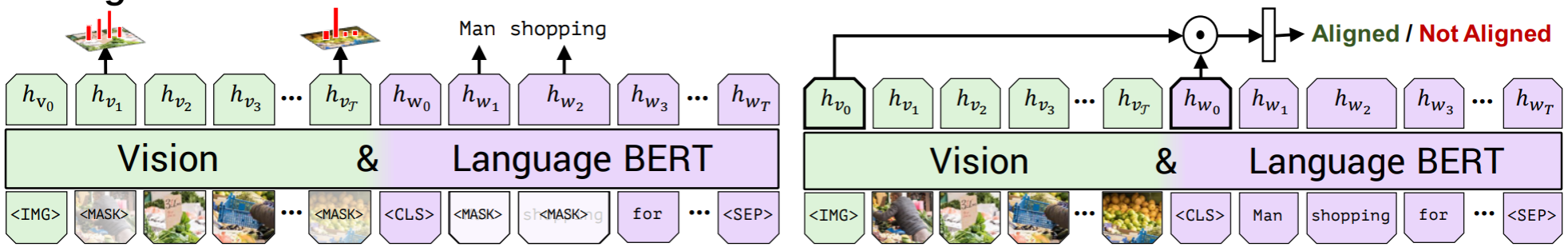
NIPS 2019, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

ViLBERT

Model Structure



Training Task



(a) Masked multi-modal learning (b) Multi-modal alignment prediction

NIPS 2019, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

ViLBERT

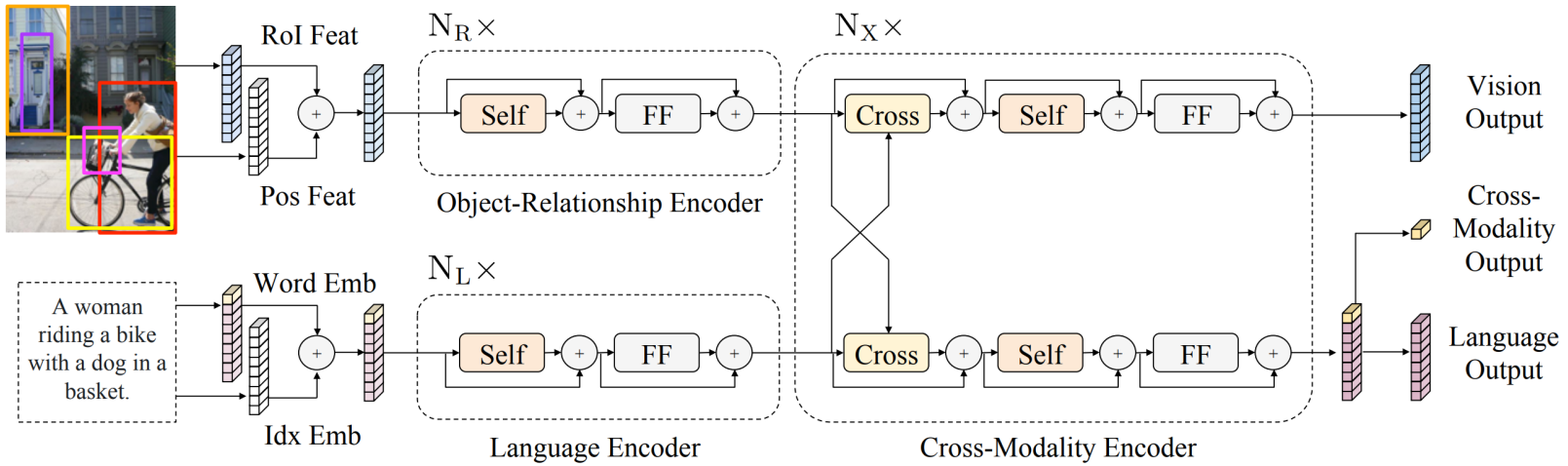
		VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
	Method	test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
SOTA	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
SOTA	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
SOTA	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	-	-	-	-	-	-
Ours	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
Ours	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
Ours	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**

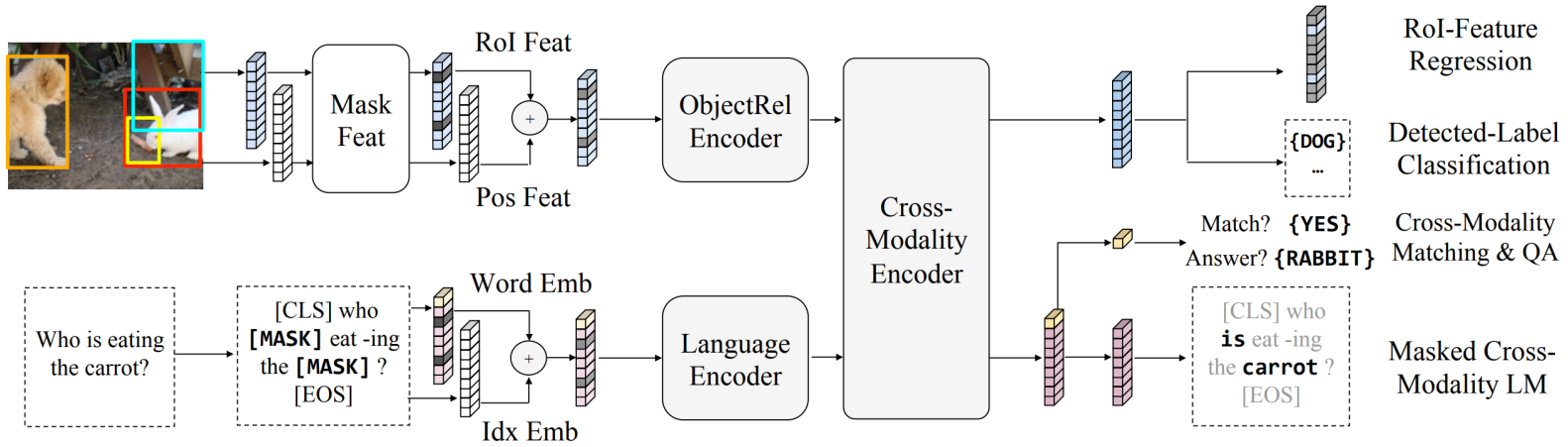
†： 不使用Conceptual Captions预训练

Single-Stream：使用单个BERT模型处理视觉和文本输入

NIPS 2019, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

LXMERT





Pre-training Datasets

- MS COCO
- Visual Genome
- VQA v2.0
- GQA
- VG-QA

For QA Task

ACL 2019, LXMERT: Learning Cross-Modality Encoder Representations from Transformers

LXMERT

Comparison

(Test-set result)

Method	VQA				GQA			$NLVR^2$	
	Binary	Number	Other	**Accu**	Binary	Open	**Accu**	Cons	**Accu**
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT	**88.2**	**54.2**	**63.1**	**72.5**	**77.8**	**45.0**	**60.3**	**42.1**	**76.2**

70.92 (ViLBERT)

Ablation studies

(Dev-set result)

Method	VQA	GQA	$NLVR^2$
1. P20 + DA	68.0	58.1	-
2. P20 + FT	68.9	58.2	72.4
3. P10+QA10 + DA	69.1	59.2	-
4. P10+QA10 + FT	**69.9**	**60.0**	**74.9**

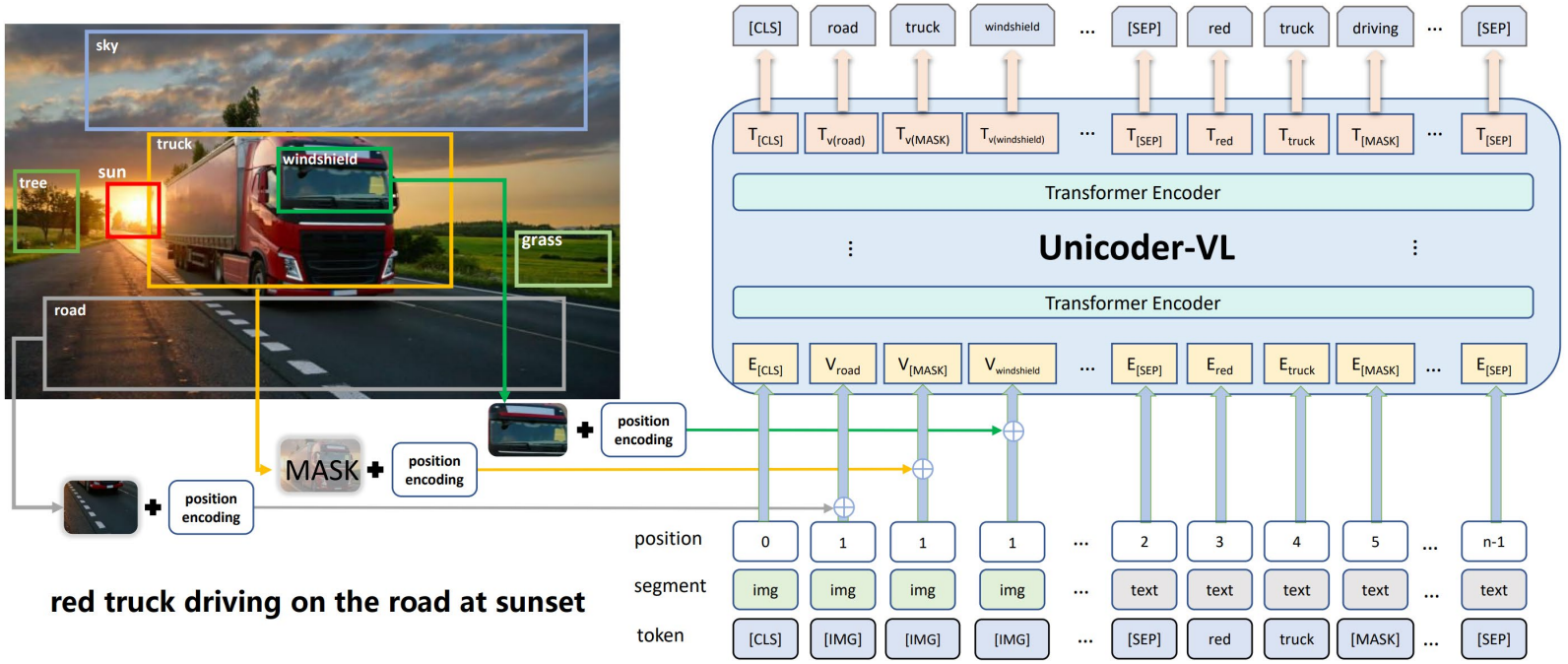
Method	VQA	GQA	$NLVR^2$
1. No Vision Tasks	66.3	57.1	50.9
2. Feat	69.2	59.5	72.9
3. Label	69.5	59.3	73.5
4. Feat + Label	**69.9**	**60.0**	**74.9**

ACL 2019, LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Unicoder-VL

Improvement

- use a single-stream architecture
- use object label classification instead minimize KL divergence

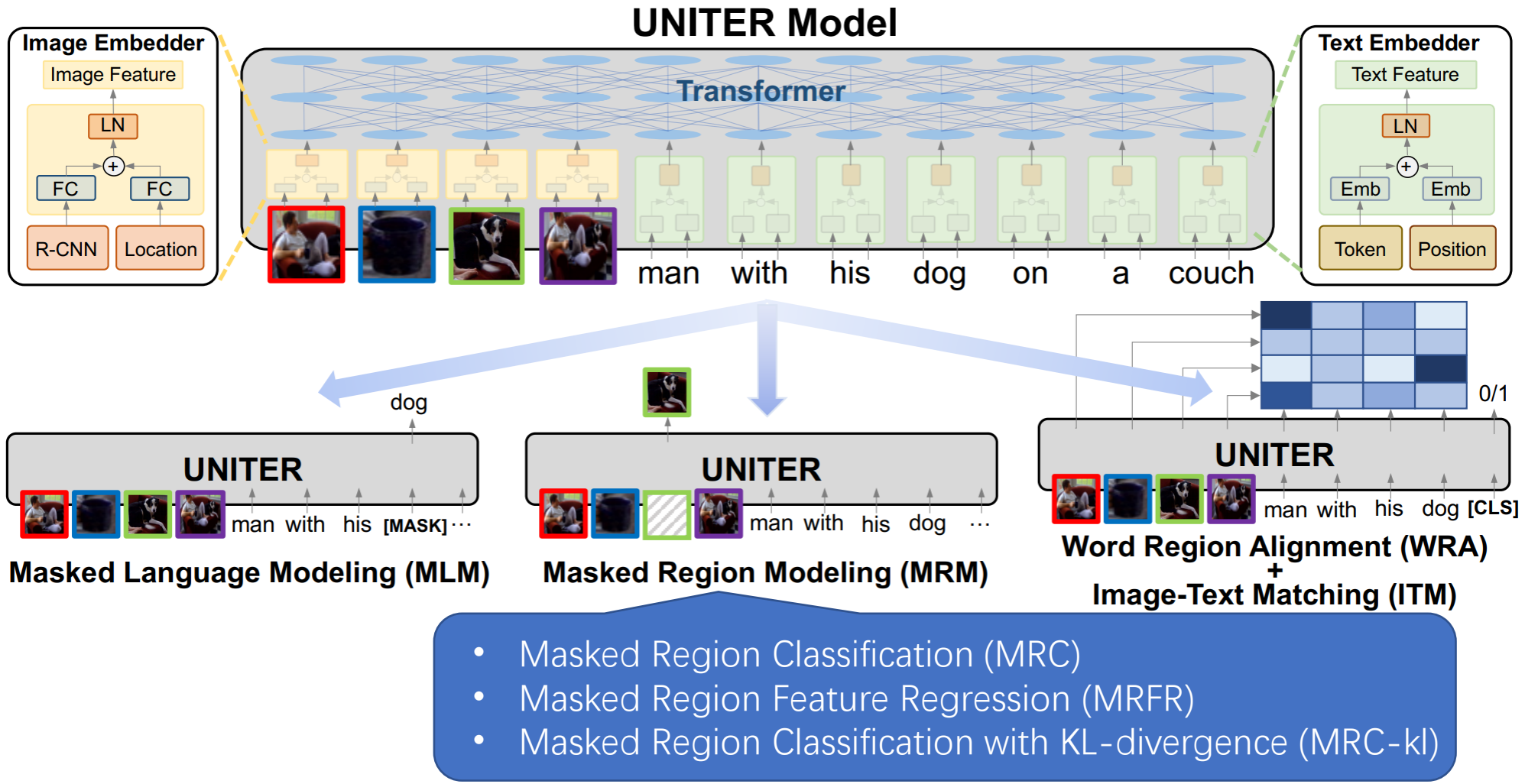


Unicoder-VL

Methods	MSCOCO						Flickr30k					
	Sentence Retrieval			Image Retrieval			Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1K Test set												
DVSA (Karpathy and Fei-Fei 2015)	38.4	69.9	80.5	27.4	60.2	74.8	22.2	48.2	61.4	15.2	37.7	50.5
m-CNN (Ma et al. 2015)	42.8	73.1	84.1	32.6	68.6	82.8	33.6	64.1	74.9	26.2	56.3	69.6
DSPE (Wang, Li, and Lazebnik 2016)	50.1	79.7	89.2	39.6	75.2	86.9	40.3	68.9	79.9	29.7	60.1	72.1
VSE++ (Faghri et al. 2017)	64.7	-	95.9	52.0	-	92.0	52.9	79.1	87.2	39.6	69.6	79.5
SCAN (Lee et al. 2018)	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
SCG (Shi et al. 2019)	76.6	96.3	99.2	61.4	88.9	95.1	71.8	90.8	94.8	49.3	76.4	85.6
PFAN (Wang et al. 2019)	76.5	96.3	99.0	61.6	89.6	95.2	70.0	91.8	95.0	50.4	78.7	86.1
ViLBERT (Lu et al. 2019)†	-	-	-	-	-	-	-	-	-	58.2	84.9	91.5
UNITER (Chen et al. 2019)†	-	-	-	-	-	-	84.7	**97.1**	99.0	71.5	**91.2**	**95.2**
Unicoder-VL (zero-shot)	54.4	82.8	90.6	43.4	76.0	87.0	64.3	85.8	92.3	48.4	76.0	85.2
Unicoder-VL (w/o pre-training)	75.1	94.3	97.8	63.9	91.6	96.5	73.0	89.0	94.1	57.8	82.2	88.9
Unicoder-VL	**84.3**	**97.3**	**99.3**	**69.7**	**93.5**	**97.2**	**86.2**	96.3	**99.0**	**71.5**	90.9	94.9
5K Test set												
SCAN (Lee et al. 2018)	50.4	82.2	90.0	38.6	69.3	80.4	-	-	-	-	-	-
SCG (Shi et al. 2019)	56.6	84.5	92.0	39.2	68.0	81.3	-	-	-	-	-	-
UNITER (Chen et al. 2019)†	**63.3**	87.0	**93.1**	**48.4**	**76.7**	**85.9**	-	-	-	-	-	-
Unicoder-VL	62.3	**87.1**	92.8	46.7	76.0	85.3	-	-	-	-	-	-

AAAI 2020, Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training

UNITER



ECCV 2020, UNITER: Universal Image-Text Representation Learning

Recent Advances in Vision and Language Pretrained Models

UNITER

Tasks		SOTA	ViLBERT	VLBERT (Large)	Unicoder -VL	VisualBERT	LXMERT	UNITER Base (12 layer)	UNITER Large (24 layer)
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**
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**

3.3M
1.8M

Pre-training Dataset
- Conceptual Captions
- SBU Captions
- COCO
- Visual Genome

Out-of-domain

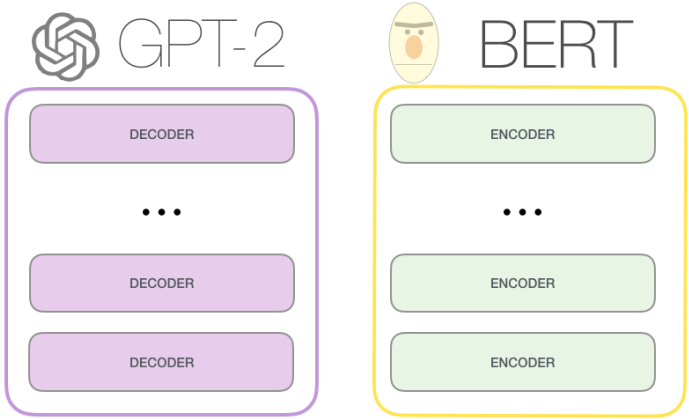
In-domain

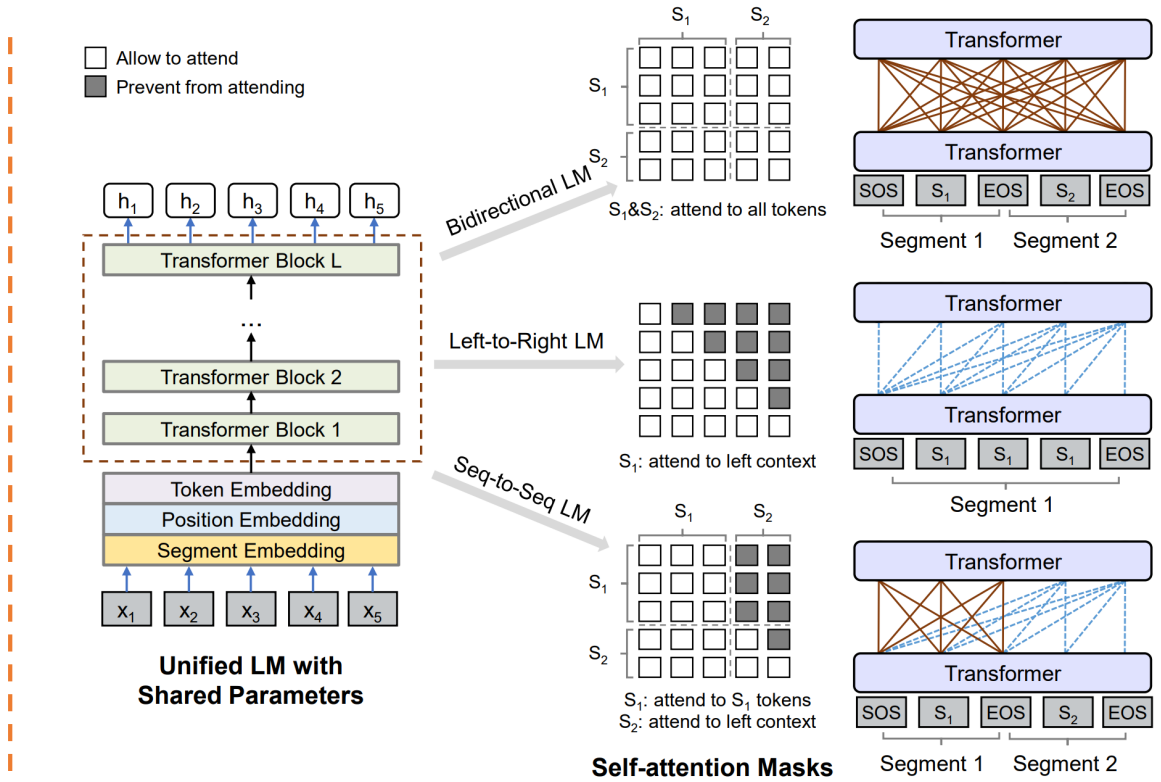
AVG of R@1, R@5 R@10

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**

ECCV 2020, UNITER: Universal Image-Text Representation Learning

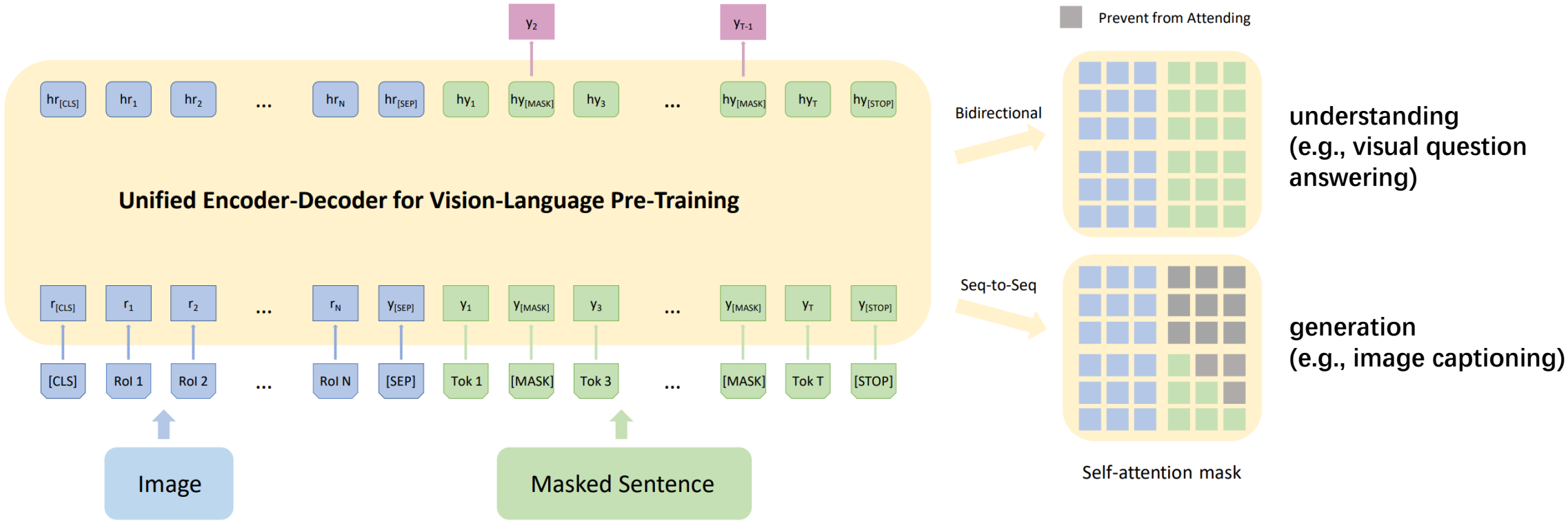
Unified VLP - Background





NIPS 2019, Unified Language Model Pre-training for Natural Language Understanding and Generation

Unified VLP

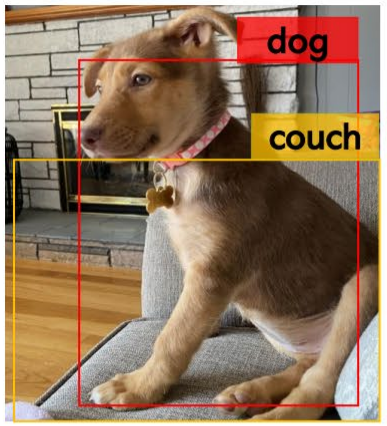

Unified Encoder-Decoder for Vision-Language Pre-Training
Image
Masked Sentence
Prevent from Attending
Bidirectional
Seq-to-Seq
Self-attention mask
understanding
(e.g., visual question answering)
generation
(e.g., image captioning)

Unified VLP

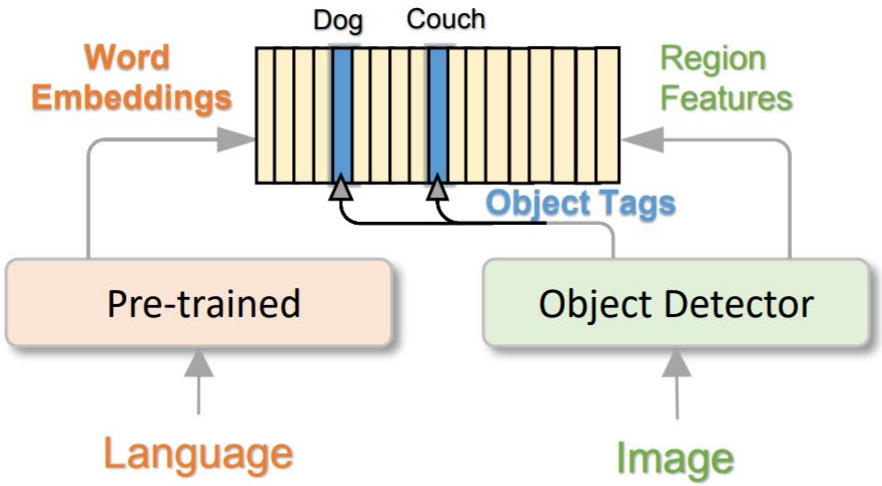
Method	COCO				VQA 2.0 (Test-Standard)				Flickr30k			
	B@4	M	C	S	Overall	Yes/No	Number	Other	B@4	M	C	S
BUTD (Anderson et al. 2018)	36.2	27.0	113.5	20.3	65.7	-	-	-	27.3	21.7	56.6	16.0
NBT (with BBox) (Lu et al. 2018)	34.7	27.1	107.2	20.1	-	-	-	-	27.1	21.7	57.5	15.6
GCN-LSTM (spa) (Yao et al. 2018)	**36.5**	27.8	115.6	20.8	-	-	-	-	-	-	-	-
GCN-LSTM (sem)	**36.8**	27.9	116.3	20.9	-	-	-	-	-	-	-	-
GVD (Zhou et al. 2019)	-	-	-	-	-	-	-	-	26.9	22.1	60.1	16.1
GVD (with BBox)	-	-	-	-	-	-	-	-	27.3	22.5	62.3	16.5
BAN (Kim, Jun, and Zhang 2018)	-	-	-	-	70.4	85.8	**53.7**	**60.7**	-	-	-	-
DFAF (Gao et al. 2019)	-	-	-	-	70.3	-	-	-	-	-	-	-
AoANet* (Huang et al. 2019)	37.2	28.4	119.8	21.3	-	-	-	-	-	-	-	-
ViLBERT* (Lu et al. 2019)	-	-	-	-	70.9	-	-	-	-	-	-	-
LXMERT* (Tan and Bansal 2019)	-	-	-	-	72.5	88.2	54.2	63.1	-	-	-	-
Ours												
w/o VLP pre-training (baseline)	35.5	28.2	114.3	21.0	70.0	86.3	52.2	59.9	27.6	20.9	56.8	15.3
seq2seq pre-training only	**36.5**	**28.4**	**117.7**	**21.3**	70.2	86.7	52.7	59.9	**31.1**	**23.0**	**68.5**	**17.2**
bidirectional pre-training only	36.1	28.3	116.5	21.2	**71.3**	**87.6**	**53.5**	**61.2**	**30.5**	22.6	63.3	16.9
Unified VLP	**36.5**	**28.4**	**116.9**	**21.2**	**70.7**	**87.4**	52.1	60.5	30.1	**23.0**	**67.4**	**17.0**

*** indicates unpublished works**

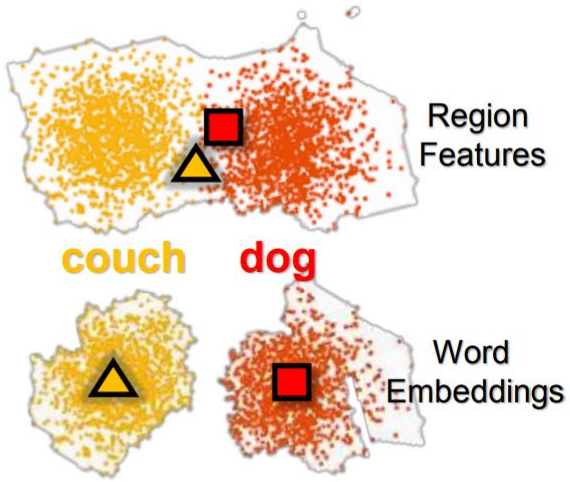
Oscar



a) Image-text pair



(b) Objects as anchor points



(c) Semantics spaces

ECCV 2020, Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks

Oscar

$$\mathcal{L}_{\mathrm{C}} = -\mathbb{E}_{(\boldsymbol{h}',\boldsymbol{w})\sim\mathcal{D}} \log p(y|f(\boldsymbol{h}',\boldsymbol{w}))$$

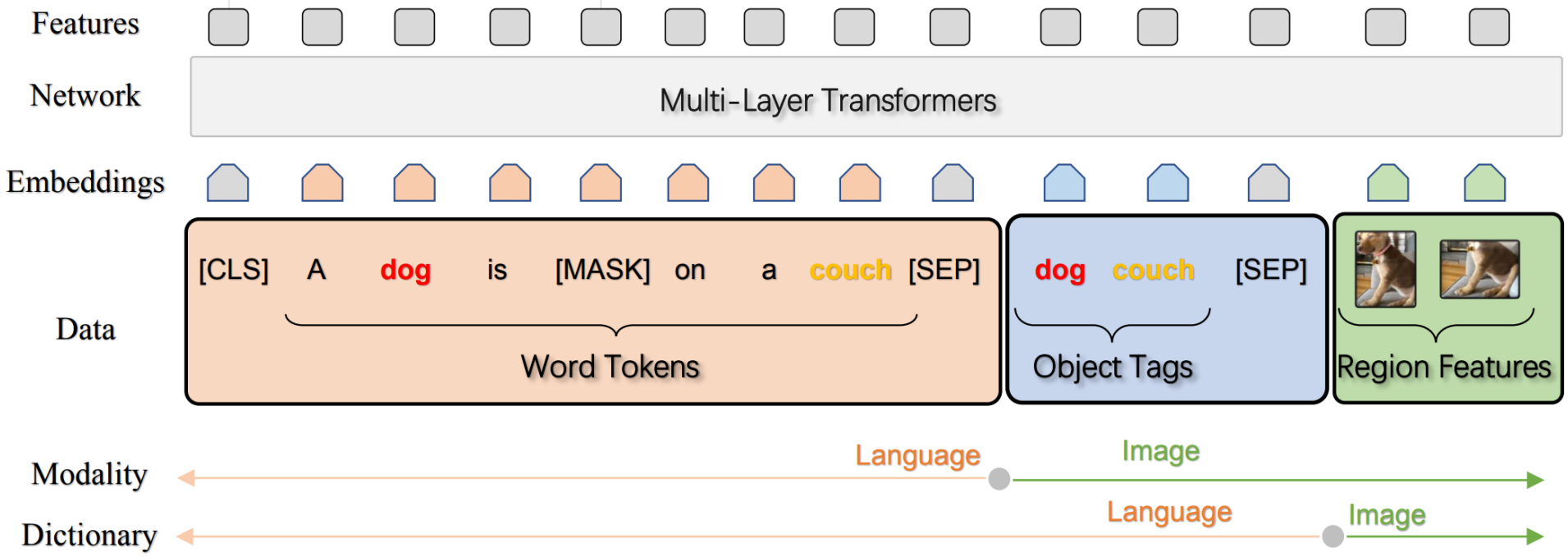
$$\boldsymbol{h}' \triangleq [\boldsymbol{q},\boldsymbol{v}]$$

Contrastive Loss

$$\boldsymbol{h} \triangleq [\boldsymbol{w},\boldsymbol{q}]$$

$$\mathcal{L}_{\mathrm{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i|\boldsymbol{h}_{\setminus i},\boldsymbol{v})$$

Masked Token Loss



The image-text pair was represent as a triple [**word tokens** , **object tags** , **region features**]

ECCV 2020, Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks

Oscar

Method	Size	Text Retrieval R@1	Text Retrieval R@5	Text Retrieval R@10	Image Retrieval R@1	Image Retrieval R@5	Image Retrieval R@10	Text Retrieval R@1	Text Retrieval R@5	Text Retrieval R@10	Image Retrieval R@1	Image Retrieval R@5	Image Retrieval R@10
		1K Test Set						5K Test Set					
DVSA [14]	-	38.4	69.9	80.5	27.4	60.2	74.8	-	-	-	-	-	-
VSE++ [7]	-	64.7	-	95.9	52.0	-	92.0	41.3	-	81.2	30.3	-	72.4
DPC [46]	-	65.6	89.8	95.5	47.1	79.9	90.0	41.2	70.5	81.1	25.3	53.4	66.4
CAMP [42]	-	72.3	94.8	98.3	58.5	87.9	95.0	50.1	82.1	89.7	39.0	68.9	80.2
SCAN [18]	-	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
SCG [33]	-	76.6	96.3	99.2	61.4	88.9	95.1	56.6	84.5	92.0	39.2	68.0	81.3
PFAN [41]	-	76.5	96.3	99.0	61.6	89.6	95.2	-	-	-	-	-	-
Unicoder-VL [19]	B	84.3	97.3	99.3	69.7	93.5	97.2	62.3	87.1	92.8	46.7	76.0	85.3
12-in-1 [24]	B	-	-	-	65.2	91.0	96.2	-	-	-	-	-	-
UNITER [5]	B	-	-	-	-	-	-	63.3	87.0	93.1	48.4	76.7	85.9
UNITER [5]	L	-	-	-	-	-	-	66.6	89.4	94.3	51.7	78.4	86.9
OSCAR	B	88.4	**99.1**	**99.8**	75.7	95.2	98.3	70.0	91.1	95.5	54.0	80.8	88.5
	L	**89.8**	98.8	99.7	**78.2**	**95.8**	**98.3**	**73.5**	**92.2**	**96.0**	**57.5**	**82.8**	**89.8**

(a) Image-text retrieval

Method	ViLBERT	VL-BERT	VisualBERT	LXMERT	12-in-1	UNITER_B	UNITER_L	OSCAR_B	OSCAR_L
Test-dev	70.63	70.50	70.80	72.42	73.15	72.27	**73.24**	73.16	**73.61**
Test-std	70.92	70.83	71.00	72.54	–	72.46	73.40	**73.44**	**73.82**

(b) VQA

Method	MAC	VisualBERT	LXMERT	12-in-1	UNITER_B	UNITER_L	OSCAR_B	OSCAR_L
Dev	50.8	67.40	74.90	–	77.14	**78.40**	78.07	**79.12**
Test-P	51.4	67.00	74.50	78.87	77.87	**79.50**	78.36	**80.37**

(c) NLVR2

Method	Test-dev	Test-std
LXMERT [39]	60.00	60.33
MMN [4]	–	60.83
12-in-1 [24]	–	60.65
NSM [12]	–	63.17
OSCAR_B	61.19	61.23
OSCAR_B*	**61.58**	**61.62**

(d) GQA

Method	cross-entropy optimization B@4	M	C	S	CIDEr optimization B@4	M	C	S
BUTD [2]	36.2	27.0	113.5	20.3	36.3	27.7	120.1	21.4
VLP [47]	36.5	28.4	117.7	21.3	39.5	29.3	129.3	23.2
AoANet [11]	37.2	28.4	119.8	21.3	38.9	29.2	129.8	22.4
OSCAR_B	36.5	**30.3**	**123.7**	**23.1**	**40.5**	**29.7**	**137.6**	**22.8**
OSCAR_L	**37.4**	**30.7**	**127.8**	**23.5**	**41.7**	**30.6**	**140.0**	**24.5**

(e) Image captioning on COCO

Method	in-domain CIDEr	in-domain SPICE	near-domain CIDEr	near-domain SPICE	out-of-domain CIDEr	out-of-domain SPICE	overall CIDEr	overall SPICE
UpDown [1]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1
UpDown + CBS [1]	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
UpDown + ELMo + CBS [1]	79.3	**12.4**	73.8	11.4	71.7	9.9	74.3	11.2
OSCAR_B	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
OSCAR_B + CBS	80.0	12.1	80.4	**12.2**	75.3	**10.6**	79.3	**11.9**
OSCAR_B + SCST + CBS	**83.4**	12.0	**81.6**	12.0	**77.6**	**10.6**	**81.1**	11.7
OSCAR_L	79.9	**12.4**	68.2	11.8	45.1	9.4	65.2	11.4
OSCAR_L + CBS	78.8	12.2	78.9	**12.1**	77.4	10.5	78.6	**11.8**
OSCAR_L + SCST + CBS	**85.4**	11.9	**84.0**	11.7	**80.3**	10.0	**83.4**	11.4

(f) Evaluation on NoCaps Val. Models are trained on COCO only without pre-training.

Training and inference tricks in image caption task

- Constrained beam search (CBS)
- self-critical sequence training (SCST)

ECCV 2020, Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks

Oscar

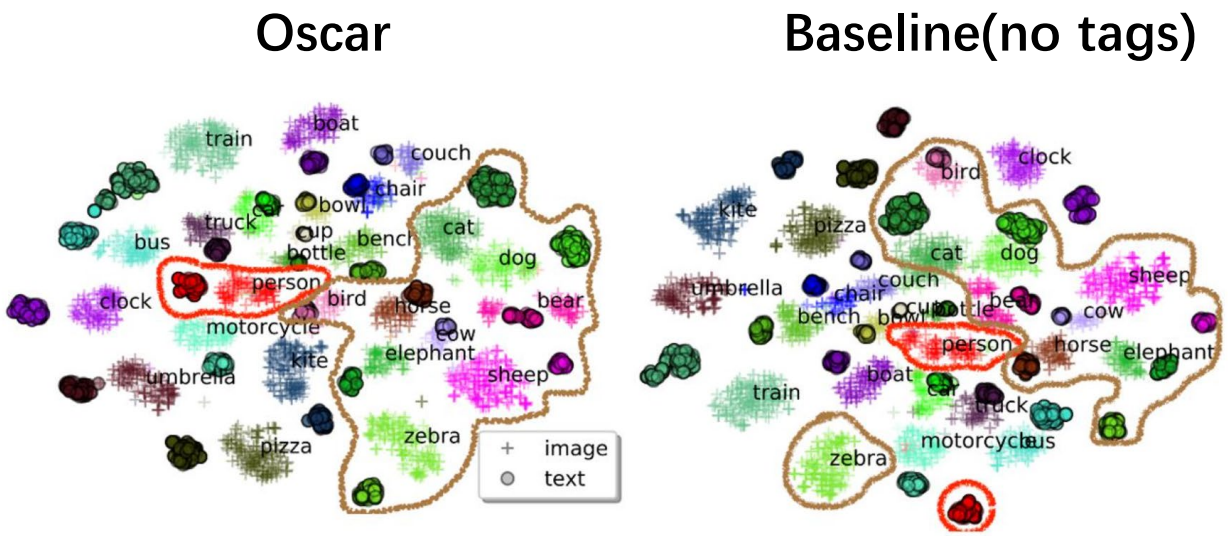


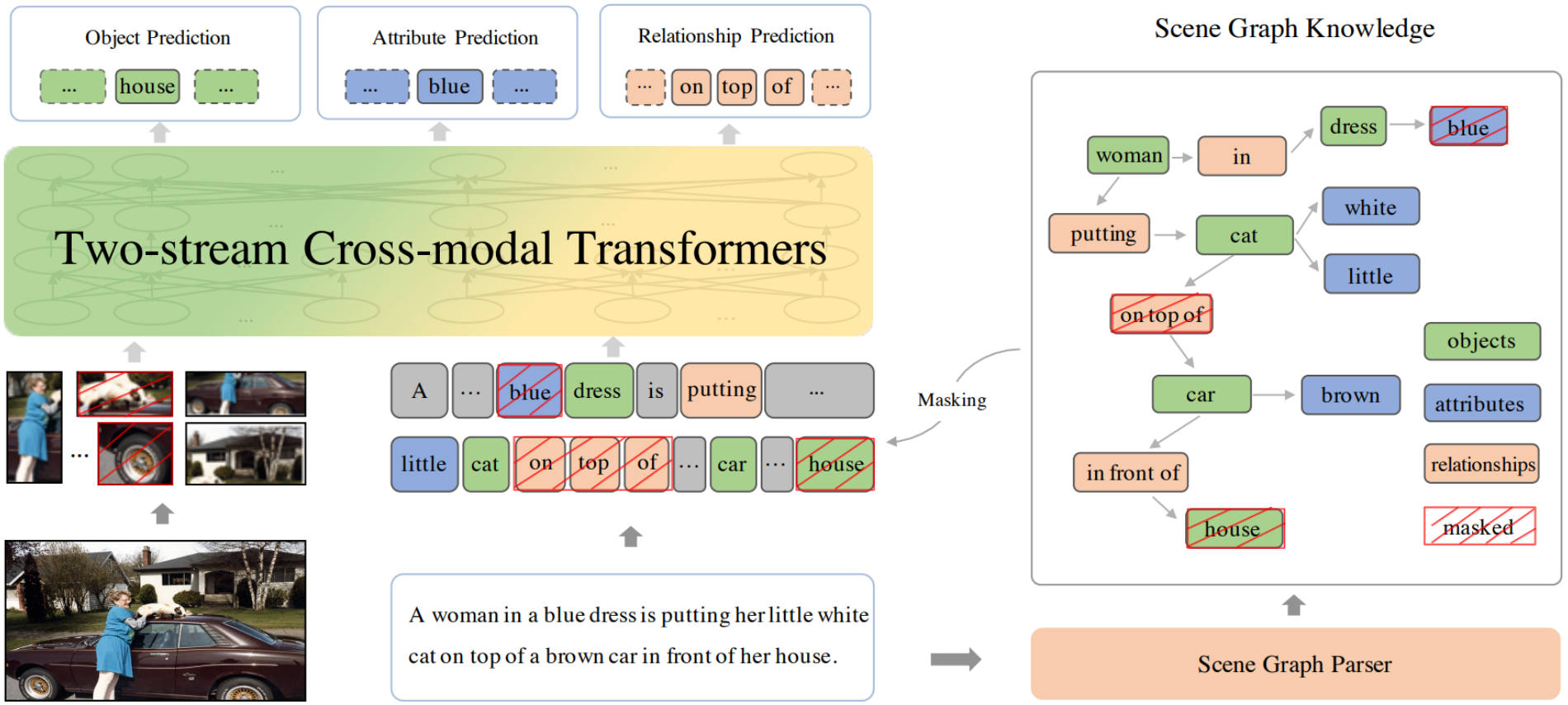
Figure 4: 2D visualization using t-SNE. The points from the same object class share the same color. Oscar (left) improves the cross-domain alignment over the baseline without object tags (right). Red and grey curves cover the objects of the same and related semantics, respectively.

Table 4: Results with various pre-training schemes.

Pre-train	VQA	Text Retrieval			Image Retrieval			Image Captioning			
	dev	R@1	R@5	R@10	R@1	R@5	R@10	B@4	M	C	S
BASELINE (NO TAGS)	70.93	84.4	98.1	99.5	73.1	94.5	97.9	34.5	29.1	115.6	21.9
OSCAR^{VG}	71.70	88.4	99.1	99.8	75.7	95.2	98.3	36.4	30.3	123.4	23.0
OSCAR^{OI}	71.15	85.9	97.9	99.5	72.9	94.3	97.6	35.3	29.6	119.5	22.6

visual genome
open images

ERNIE-VIL



arXiv 2020, ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph

ERNIE-VIL

Models		VCR			RefCOCO+		
		Q→A	QA→R	Q→AR	val	testA	testB
Out-of-domain	ViLBERT-base	72.42 (73.3)	74.47 (74.6)	54.04 (54.8)	72.34	78.52	62.61
	Unicoder-VL-base	72.6 (73.4)	74.5 (74.4)	54.4 (54.9)	-	-	-
	VLBERT-base	73.8 (-)	74.4 (-)	55.2 (-)	71.60	77.72	60.99
	UNITER-base	-	-	-	72.78	-	-
	VLBERT-large	75.5 (75.8)	77.9 (78.4)	58.9 (59.7)	72.59	78.57	62.30
	ERNIE-ViL-base	74.37 (77.0)	79.65 (80.3)	61.24 (62.1)	74.02	80.33	**64.74**
	ERNIE-ViL-large	**78.52(79.2)**	**83.37(83.5)**	**65.81(66.3)**	**74.24**	**80.97**	64.70
Out-of-domain + in-domain	UNITER-base	74.56 (75.0)	77.03 (77.2)	57.76 (58.2)	75.31	81.30	65.58
	VILLA-base	75.54 (76.4)	78.78 (79.1)	59.75 (60.6)	76.05	81.65	65.70
	UNITER-large	77.22 (77.3)	80.49 (80.8)	62.59 (62.8)	75.90	81.45	66.70
	VILLA-large	78.45 (78.9)	82.57 (82.8)	65.18 (65.7)	**76.17**	81.54	66.84
	ERNIE-ViL-large	**78.62** (-)	**83.42** (-)	**65.95**(-)	75.95	**82.07**	**66.88**

Models		VQA		IR-Flickr30K			TR-Flickr30K		
		test-dev	test-std	R@1	R@5	R@10	R@1	R@5	R@10
Out-of-domain	ViLBERT-base	70.55	70.92	58.20	84.90	91.52	-	-	-
	Unicoder-VL-base	-	-	71.50	90.90	94.90	86.20	96.30	99.00
	VLBERT-base	71.16	-	-	-	-	-	-	-
	UNITER-base	71.56	-	-	-	-	-	-	-
	VLBERT-large	71.79	72.22	-	-	-	-	-	-
	ERNIE-ViL-base	72.62	72.85	74.44	92.72	95.94	86.70	97.80	99.00
	ERNIE-ViL-large	**73.78**	**73.96**	**75.10**	**93.42**	**96.26**	**88.70**	**97.30**	**99.10**
Out-of-domain + in-domain	UNITER-base	72.70	72.91	72.52	92.36	96.08	85.90	97.10	98.80
	OSCAR-base	73.16	73.61	-	-	-	-	-	-
	VILLA-base	73.59	73.67	74.74	92.86	95.82	86.60	97.90	**99.20**
	12-in-1-base	73.15	-	67.90	-	-	-	-	-
	UNITER-large	73.82	74.02	75.56	94.08	96.76	87.30	**98.00**	**99.20**
	OSCAR-large	73.44	73.82	-	-	-	-	-	-
	VILLA-large	74.69	74.87	76.26	**94.24**	**96.84**	87.90	97.50	98.80
	ERNIE-ViL-large	**74.75**	**74.93**	**76.70**	93.58	96.44	**88.10**	**98.00**	**99.20**

arXiv 2020, ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph

Visual Dialog

Task	Num of sentences
Dialogue	up to 10 rounds sequences
Caption	1 sentence
Image-Text Retrieval	
VQA	2 sentences (question-answer pair)

Captioning
Two people are in a wheelchair and one is holding a racket.

VQA
Q: How many people on wheelchairs ?
A: Two

Q: How many wheelchairs ?
A: One

Visual Dialog
Q: How many people are on wheelchairs ?
A: Two
Q: What are their genders ?
A: One male and one female
Q: Which one is holding a racket ?
A: The woman

Visual Dialog
Q: What is the gender of the one in the white shirt ?
A: She is a woman
Q: What is she doing ?
A: Playing a Wii game
Q: Is that a man to her right
A: No, it's a woman

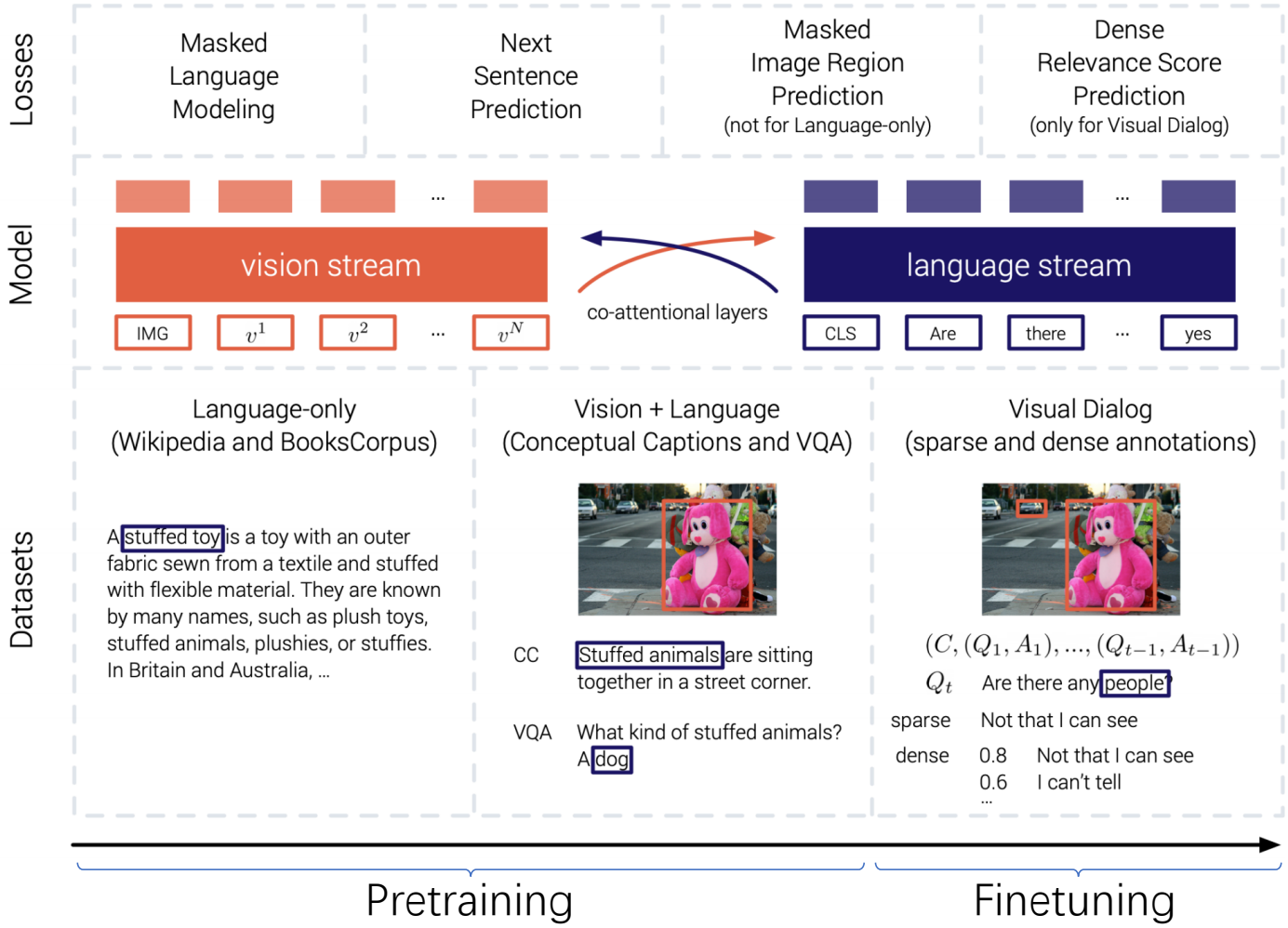
Candidate answers:

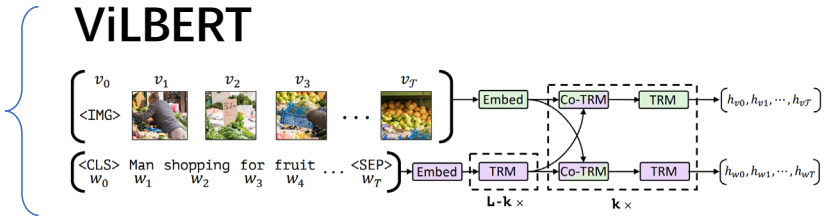
$A_t = \left\{A_t^{(1)}, \ldots, A_t^{(100)}\right\}$

CVPR 2017, Visual Dialog

Recent Advances in Vision and Language Pretrained Models

VisDial-BERT





ECCV 2020, Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline

VisDial-BERT

BERT

ViLBERT

	Model	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MR ↓
L-only	NSP	56.17	63.37	49.17	80.62	89.42	4.21
L-only	NSP + MLM	57.26	64.40	50.30	81.60	90.43	4.01
+vision	w/ L-only	62.64	67.86	54.54	84.34	92.36	3.44
+vision	w/ CC [3]	60.80	67.13	53.59	84.39	92.49	3.44
+vision	w/ CC [3]+VQA [4]	64.94	**69.10**	**55.88**	**85.50**	**93.29**	**3.25**
+dense	CE	**75.24**	52.22	39.92	65.05	80.63	6.17
+dense	CE + NSP	69.24	65.88	53.41	80.92	90.18	4.24

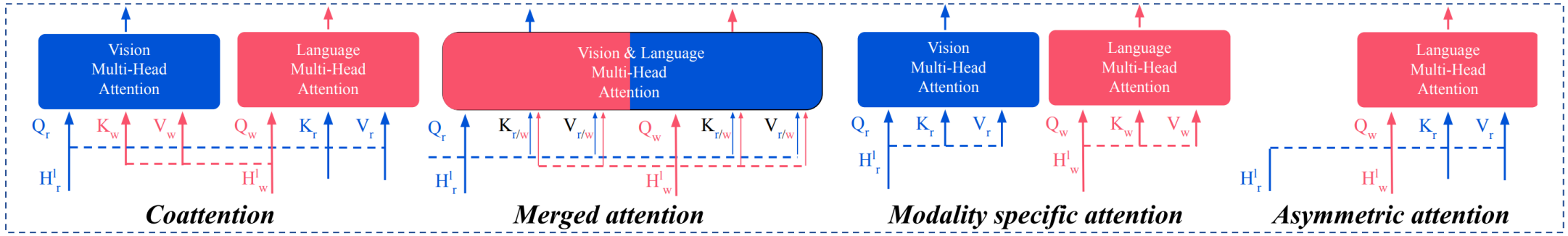
Table 3: Results on VisDial v1.0 val. ↑ indicates higher is better

	Model	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MR ↓
Published Results	GNN [15]	52.82	61.37	47.33	77.98	87.83	4.57
Published Results	CorefNMN [12]	54.70	61.50	47.55	78.10	88.80	4.40
Published Results	RvA [14]	55.59	63.03	49.03	80.40	89.83	4.18
Published Results	HACAN [22]	57.17	64.22	50.88	80.63	89.45	4.20
Published Results	NMN [12]	58.10	58.80	44.15	76.88	86.88	4.81
Published Results	DAN [17]	57.59	63.20	49.63	79.75	89.35	4.30
Published Results	DAN† [17]	59.36	64.92	51.28	81.60	90.88	3.92
Published Results	ReDAN [18]	61.86	53.13	41.38	66.07	74.50	8.91
Published Results	ReDAN+† [18]	64.47	53.74	42.45	64.68	75.68	6.64
Published Results	DualVD [25]	56.32	63.23	49.25	80.23	89.70	4.11
Published Results	FGA [16]	56.93	66.22	52.75	82.92	91.08	3.81
Published Results	DL-61 [23]	57.32	62.20	47.90	80.43	89.95	4.17
Published Results	DL-61† [23]	57.88	63.42	49.30	80.77	90.68	3.97
Published Results	MReal - BDAI⋆ [24]	74.02	52.62	40.03	68.85	79.15	6.76
Leaderboard Entries	LF	45.31	55.42	40.95	72.45	82.83	5.95
Leaderboard Entries	HRE	45.46	54.16	39.93	70.45	81.50	6.41
Leaderboard Entries	MN	47.50	55.49	40.98	72.30	83.30	5.92
Leaderboard Entries	MN-Att	49.58	56.90	42.43	74.00	84.35	5.59
Leaderboard Entries	LF-Att	51.63	60.41	46.18	77.80	87.30	4.75
Leaderboard Entries	MS ConvAI	55.35	63.27	49.53	80.40	89.60	4.15
Leaderboard Entries	USTC-YTH	56.47	61.44	47.65	78.13	87.88	4.65
Leaderboard Entries	UET-VNU	57.40	59.50	45.50	76.33	85.82	5.34
Leaderboard Entries	square	60.16	61.26	47.15	78.73	88.48	4.46
Leaderboard Entries	MS D365 AI	64.47	53.73	42.45	64.68	75.68	6.63
Ours	w/ CC [3]+VQA [4]	63.87	**67.50**	**53.85**	**84.68**	**93.25**	**3.32**
Ours	CE	**74.47**	50.74	37.95	64.13	80.00	6.28
Ours	CE + NSP	68.08	63.92	50.78	79.53	89.60	4.28

Table 4: Summary of results on VisDial v1.0 test-std. ↑ indicates higher is better. ↓ indicates lower is better. † denotes ensembles ⋆ denotes the winning team of the 2019 Visual Dialog Challenge.

Summary

- Dataset and data processing
  - Automatically collecting datasets
  - Processing images into a sequence involves defining “visual tokens”
- Multimodal Attention



- Loss Functions
  - masked region modeling loss
    - minimizes the KL-divergence
    - minimizes the cross-entropy loss
  - masked sub-word modeling loss
  - image-text matching loss
    - $-y\log(\sigma(s_\theta(\boldsymbol{r}^m,\boldsymbol{w}^m)))$ / $-(1-y)\log(1-\sigma(s_\theta(\boldsymbol{r}^m,\boldsymbol{w}^m)))$ binary classification loss
    - $-\log\left(\frac{e^{s_\theta(\boldsymbol{r}^m,\boldsymbol{w}^m))}}{e^{s_\theta(\boldsymbol{r}^m,\boldsymbol{w}^m)}+\sum_{(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{w}})\sim\mathcal{N}}e^{s_\theta(\tilde{\boldsymbol{r}}^m,\tilde{\boldsymbol{w}}^m)}}\right)$ contrastive loss

Thinking and further work

- Exploring cognitive understanding level task

Title: **Why am I so tired?**

Generated caption: **A woman sitting on top of a bed in a room.**

- Knowledge enhanced multimodal pretrain

1 is quickly walking down the sidewalk with a sandwich in her hand

intent	without event	get to her destination quickly
	with event	get to work on time
	ground truth	get to her destination quickly save time by eating lunch while walking
before	without event	walk out of the building
	with event	pick up the sandwich
	ground truth	purchase the sandwich at a deli take the subway to her current location be late for work be eating a sandwich while walking to work
after	without event	walk into a building
	with event	eat the sandwich
	ground truth	cross the street at a crosswalk throw away the trash from the sandwich look down at the sandwich thinking it's terrible finish eating the sandwich anyway