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

Key Contributions

- Introduce UNITER, a powerful UNiversal Image-TExt Representations for Vision-and-Language tasks.
- Achieved new state of the art (SOTA) on multiple V+L benchmarks

Features

Image embeddings

- Use Faster R-CNN to extract the visual features (pooled ROI features) for each region. Also the location features for each region via a 7-dimensional vector.
- Both visual and location features then projected into the same embedding space.

Text embeddings

- Sum up word embedding and position embedding, followed by another LN layer

Model

- UNITER takes the visual regions of the image and textual tokens of the sentence as the input.
- Use an Image Embedder and a Text Embedder to extract their respective embeddings.
- These embeddings are fed into a multi-layer self-attention Transformer to learn a cross-modality contextualized embedding

Note: No use of multiple transformers for different modality.

Model architecture

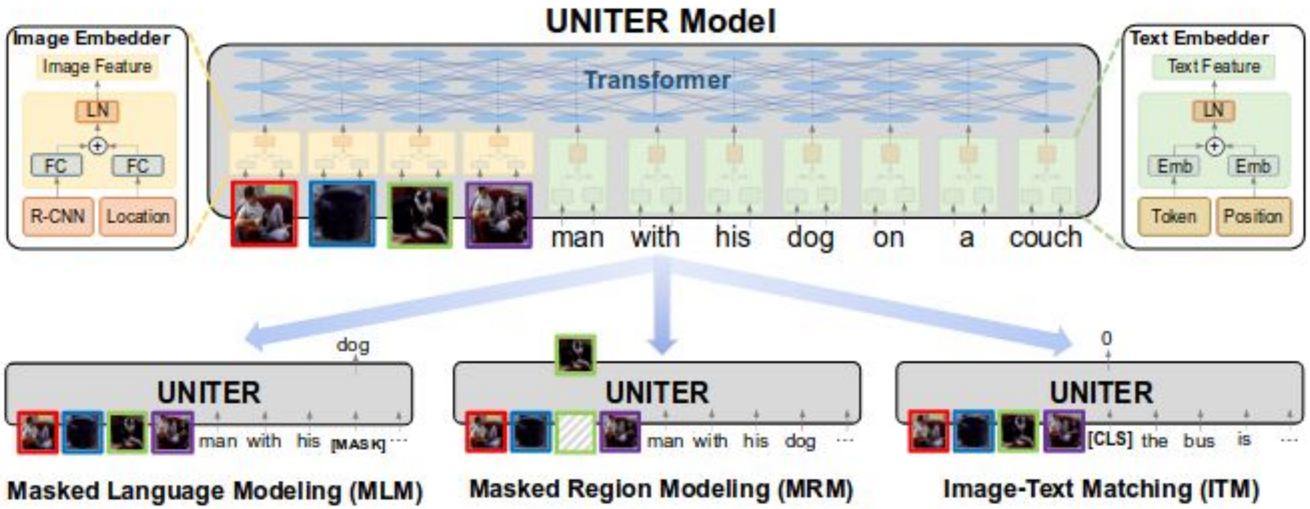
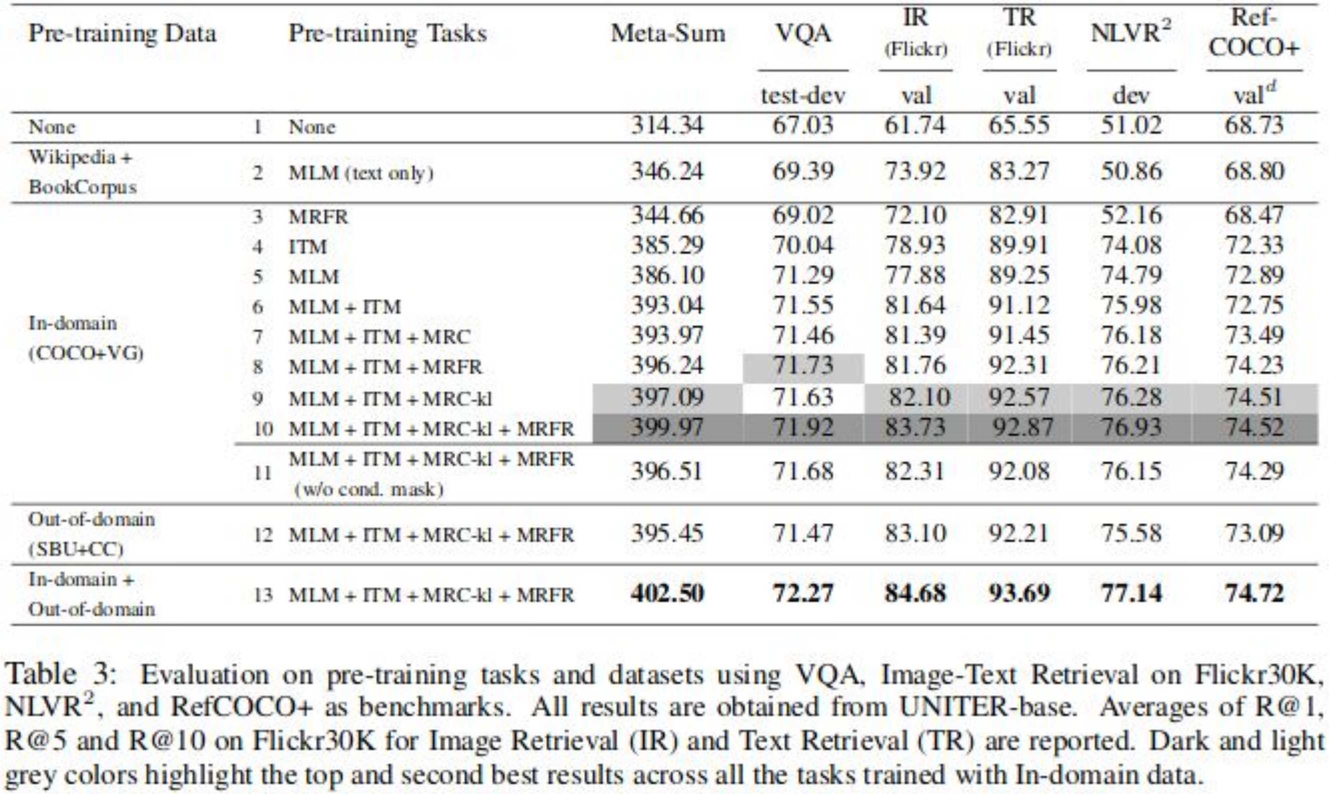


Figure 1: Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer self-attention Transformer, learned through three pre-training tasks.

Pretraining tasks

1. Masked Language Modeling
2. Masked Region Modeling
   a. Masked Region Feature Regression (MRFR)
   b. Masked Region Classification (MRC)
   c. Masked Region Classification with KL-Divergence (MRC-kl)
3. Image-Text Matching (ITM)

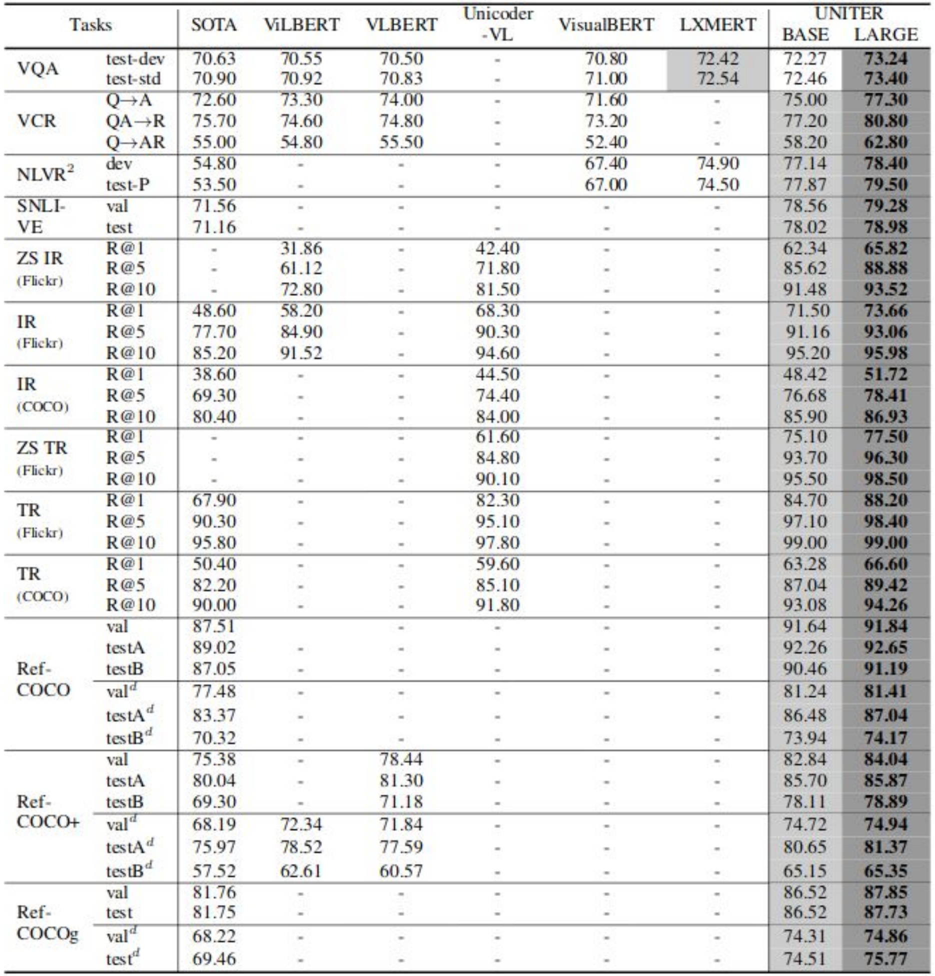
Pre training results



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 (w/o cond. mask)	396.51	71.68	82.31	92.08	76.15	74.29
Out-of-domain (SBU+CC)	12	MLM + ITM + MRC-kl + MRFR	395.45	71.47	83.10	92.21	75.58	73.09
In-domain + Out-of-domain	13	MLM + ITM + MRC-kl + MRFR	**402.50**	**72.27**	**84.68**	**93.69**	**77.14**	**74.72**

Table 3: Evaluation on pre-training tasks and datasets using VQA, Image-Text Retrieval on Flickr30K, NLVR^2, and RefCOCO+ as benchmarks. All results are obtained from UNITER-base. Averages of R@1, R@5 and R@10 on Flickr30K for Image Retrieval (IR) and Text Retrieval (TR) are reported. Dark and light grey colors highlight the top and second best results across all the tasks trained with In-domain data.

Downstream tasks results



Tasks		SOTA	ViLBERT	VLBERT	Unicoder	VisualBERT	LXMERT	UNITER	
					-VL			BASE	LARGE
VQA	test-dev	70.63	70.55	70.50	-	70.80	72.42	72.27	**73.24**
	test-std	70.90	70.92	70.83	-	71.00	72.54	72.46	**73.40**
VCR	Q→A	72.60	73.30	74.00	-	71.60	-	75.00	**77.30**
	QA→R	75.70	74.60	74.80	-	73.20	-	77.20	**80.80**
	Q→AR	55.00	54.80	55.50	-	52.40	-	58.20	**62.80**
NLVR^2	dev	54.80	-	-	-	67.40	74.90	77.14	**78.40**
	test-P	53.50	-	-	-	67.00	74.50	77.87	**79.50**
SNLI-VE	val	71.56	-	-	-	-	-	78.56	**79.28**
	test	71.16	-	-	-	-	-	78.02	**78.98**
ZS IR (Flickr)	R@1	-	31.86	-	42.40	-	-	62.34	**65.82**
	R@5	-	61.12	-	71.80	-	-	85.62	**88.88**
	R@10	-	72.80	-	81.50	-	-	91.48	**93.52**
IR (Flickr)	R@1	48.60	58.20	-	68.30	-	-	71.50	**73.66**
	R@5	77.70	84.90	-	90.30	-	-	91.16	**93.06**
	R@10	85.20	91.52	-	94.60	-	-	95.20	**95.98**
IR (COCO)	R@1	38.60	-	-	44.50	-	-	48.42	**51.72**
	R@5	69.30	-	-	74.40	-	-	76.68	**78.41**
	R@10	80.40	-	-	84.00	-	-	85.90	**86.93**
ZS TR (Flickr)	R@1	-	-	-	61.60	-	-	75.10	**77.50**
	R@5	-	-	-	84.80	-	-	93.70	**96.30**
	R@10	-	-	-	90.10	-	-	95.50	**98.50**
TR (Flickr)	R@1	67.90	-	-	82.30	-	-	84.70	**88.20**
	R@5	90.30	-	-	95.10	-	-	97.10	**98.40**
	R@10	95.80	-	-	97.80	-	-	99.00	**99.00**
TR (COCO)	R@1	50.40	-	-	59.60	-	-	63.28	**66.60**
	R@5	82.20	-	-	85.10	-	-	87.04	**89.42**
	R@10	90.00	-	-	91.80	-	-	93.08	**94.26**
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	-	78.44	-	-	-	82.84	**84.04**
	testA	80.04	-	81.30	-	-	-	85.70	**85.87**
	testB	69.30	-	71.18	-	-	-	78.11	**78.89**
	val^d	68.19	72.34	71.84	-	-	-	74.72	**74.94**
	testA^d	75.97	78.52	77.59	-	-	-	80.65	**81.37**
	testB^d	57.52	62.61	60.57	-	-	-	65.15	**65.35**
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**

Table 4: Results on downstream V+L tasks from UNITER model, compared with task-specific state-of-the-art (SOTA) and concurrent pre-trained models. ZS: Zero-Shot, IR: Image Retrieval and TR: Text Retrieval.