TAP: Text-Aware Pre-training for Text-VQA and Text-Caption

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin,
Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, Jiebo Luo

CVPR VIRTUAL JUNE 19-25

Scene Text Vision Language Tasks

Vision-language models that can read

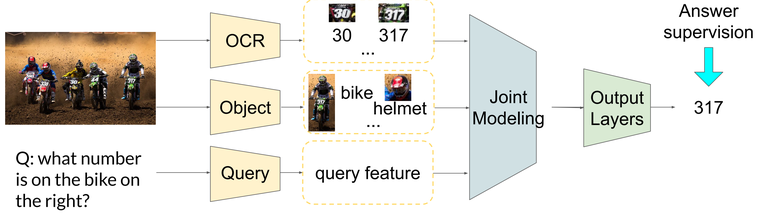
Question: what **number** is on the bike on the right? ---- A: the number is **317**

Text-VQA

A group of motorcyclists with **number 317, 44, 30, 338, 598** racing outdoor.

Text-Captioning

Common Pipeline and Limitations

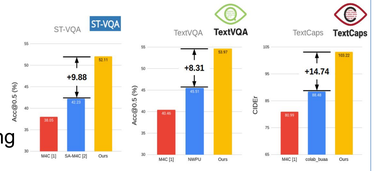


- Answer/caption loss alone is insufficient for Text-VQA/Captioning
- Previous vision-language pre-training does not consider scene text

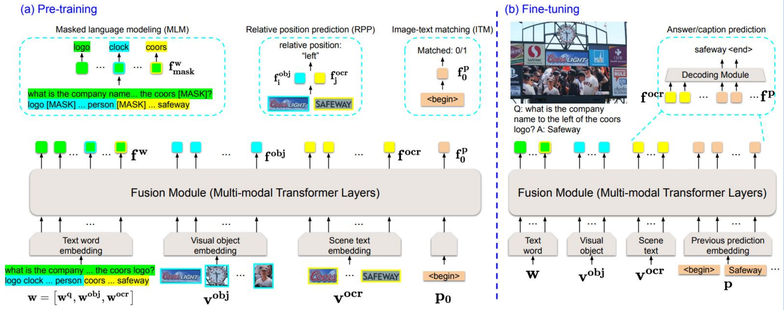
Contributions

Text-Aware Pre-training for Text-VQA and Text-Captioning

- Scene-text aware pre-training tasks design
- OCR-CC: text-related dataset for pre-training



Text-Aware Pre-training (TAP)



Text-aware pre-training for joint representation learning

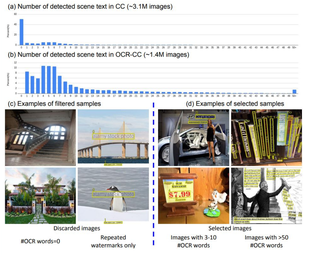
- Text-Visual (Object and Scene text): masked language modeling (MLM), image-text matching (ITM)
- Object-Scene text: relative position prediction (RPP)

- MLM, ITM:
  - Limited scene text word in original question/caption
  - Adding OCR/object words as additional text input

- RPP
  - Importance of visual regions
  - Given two sampled regions, predicting their relative spatial relationship

Pre-training with Extra Data (OCR-CC)

- Images with scene text from the Conceptual Captioning (CC) dataset



Dataset	TextVQA	ST-VQA	TextCaps	CC-OCR
Training images	**22K**	**19K**	**22K**	**1.37M**
Text	35K OCR-QA pairs	26K OCR-QA pairs	110K OCR-Caption	One caption per image
Image source	Open Image	ICDAR 2013/15, ImageNet, VizWiz, IIIT Scene Text Retrieval, Visual Genome, COCO-Text	TextVQA	Conceptual captions
# OCR	mean: 23.1, med:12	mean: 19.2, med:10	mean: 23.1, med:12	mean: 11.4, med: 6

Quantitative Results

Results on TextVQA

Method	OCR System	Extra Data	Val Acc.	Test Acc.
LoRRA [47]	Rosetta-ml	✗	26.56	27.63
MM-GNN [17]	Rosetta-ml	✗	31.44	31.10
M4C [20]	Rosetta-en	✗	39.40	39.01
SMA [16]	Rosetta-en	✗	40.05	40.66
CRN [36]	Rosetta-en	✗	40.39	40.96
LaAP-Net [19]	Rosetta-en	✗	40.68	40.54
M4C† [20]	Rosetta-en	✗	39.55	-
TAP (Ours)	Rosetta-en	✗	44.06	-
M4C [20]	Rosetta-en	ST-VQA	40.55	40.46
LaAP-Net [19]	Rosetta-en	ST-VQA	41.02	40.54
SA-M4C [25]	Google-OCR	ST-VQA	45.4	44.6
SMA [16]	SBD-Trans OCR	ST-VQA	-	45.51
M4C† [20]	Microsoft-OCR	✗	44.50	44.75
M4C† [20]	Microsoft-OCR	ST-VQA	45.22	-
TAP (Ours)	Microsoft-OCR	✗	49.91	49.71
TAP (Ours)	Microsoft-OCR	ST-VQA	50.57	50.71
TAP†† (Ours)	Microsoft-OCR	ST-VQA, TextCaps, OCR-CC	**54.71**	**53.97**

Results on ST-VQA

Method	Val Acc.	Val ANLS	Test ANLS
SAN+STR [8]	-	-	0.135
M4C [20]	38.05	0.472	0.462
SA-M4C [25]	42.23	0.512	0.504
SMA [16]	-	-	0.466
CRN [36]	-	-	0.483
LaAP-Net [19]	39.74	0.497	0.485
M4C† [20]	42.28	0.517	0.517
TAP (Ours)	45.29	0.551	0.543
TAP†† (Ours)	**50.83**	**0.598**	**0.597**

Results on TextCaps

Method	Val CIDEr	Test CIDEr
BUTD [4]	41.9	33.8
AoANet [22]	42.7	34.6
M4C [46]	89.6	81.0
MMA-SR [54]	98.0	88.0
CNMT [49]	-	93.03
M4C† [46]	99.89	93.36
TAP (Ours)	105.05	99.49
TAP†† (Ours)	**109.16**	**103.22**

Qualitative Results