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

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

Scene Text Vision Language Tasks

- Vision-language models that can read
- Text-VQA, Text-Captioning

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

Text-VQA [1]

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

Text-Captioning [2]

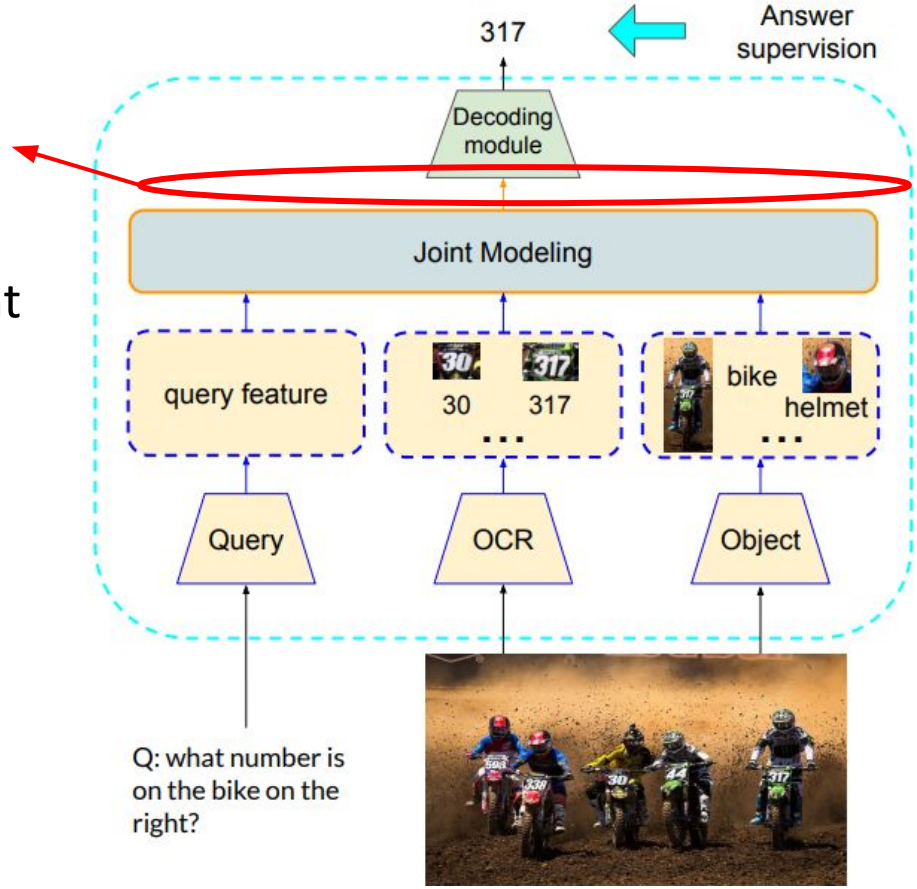
[1] Singh, Amanpreet, et al. "Towards vqa models that can read." In CVPR 2019.
[2] Sidorov, Oleksii, et al. "TextCaps: a Dataset for Image Captioning with Reading Comprehension." In ECCV 2020.

Related Works

- **Limitations**

Answer supervision alone is insufficient

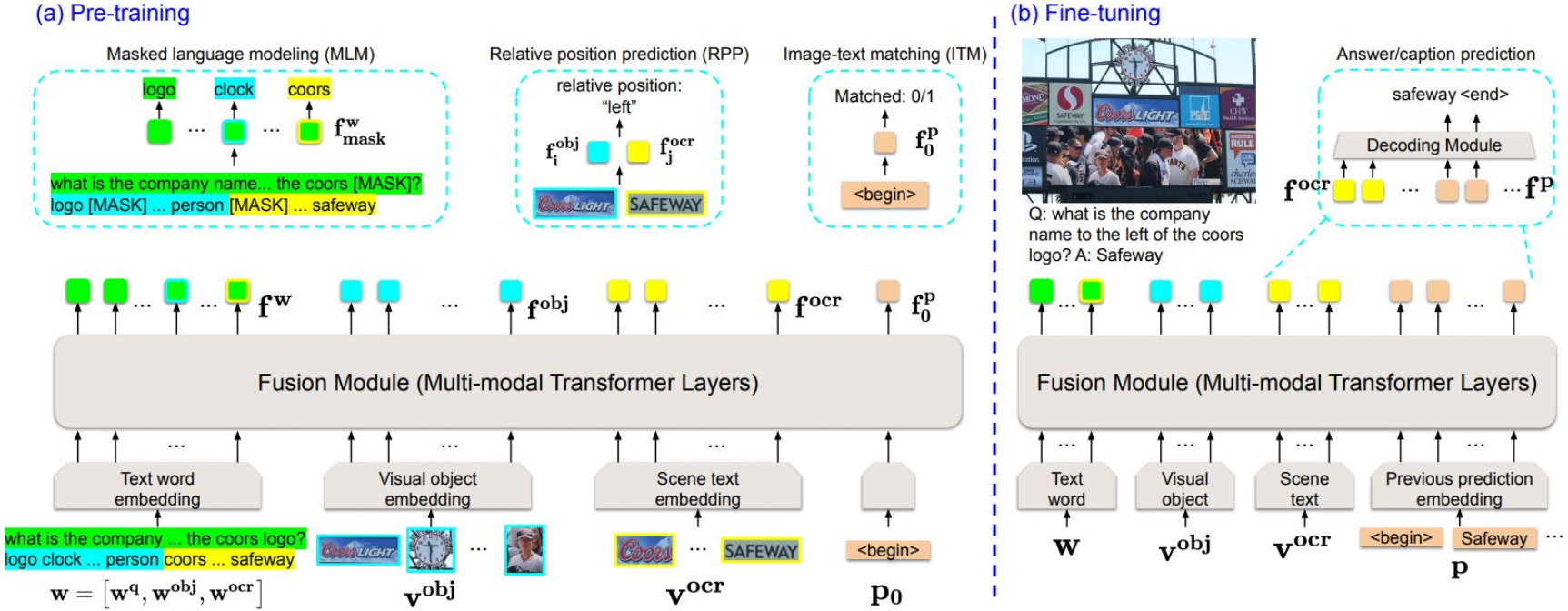
- number ⇔  
- bike ⇔  
- bike on the right ⇔ 
-  ⇔ 



Method

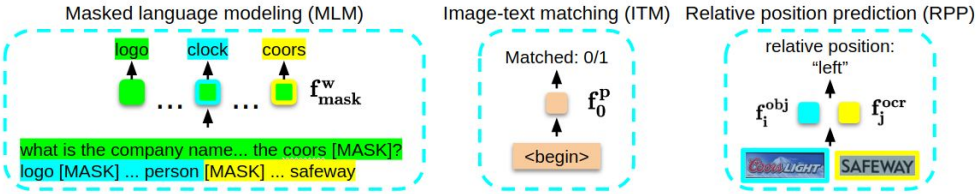
- **Framework Overview**

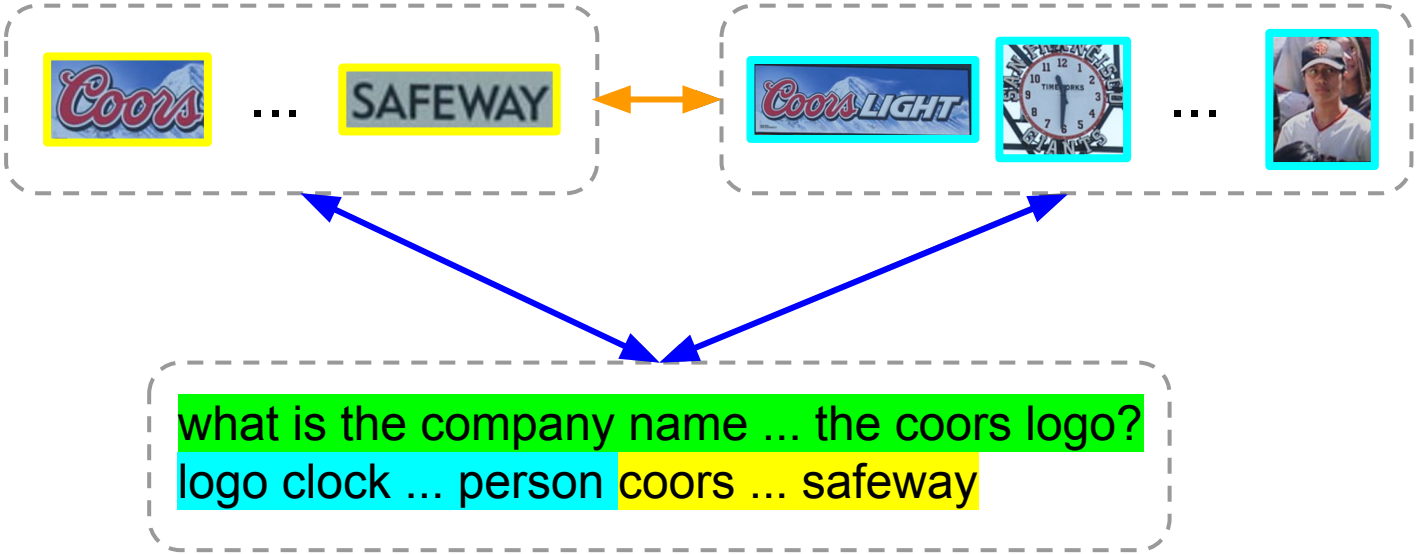
- Text-aware pre-training (TAP) for aligned representation learning



Method

- **Pre-training Tasks Design**



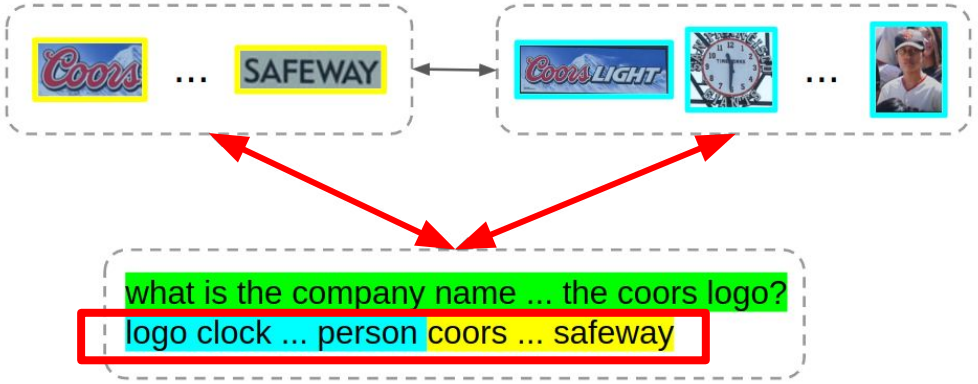


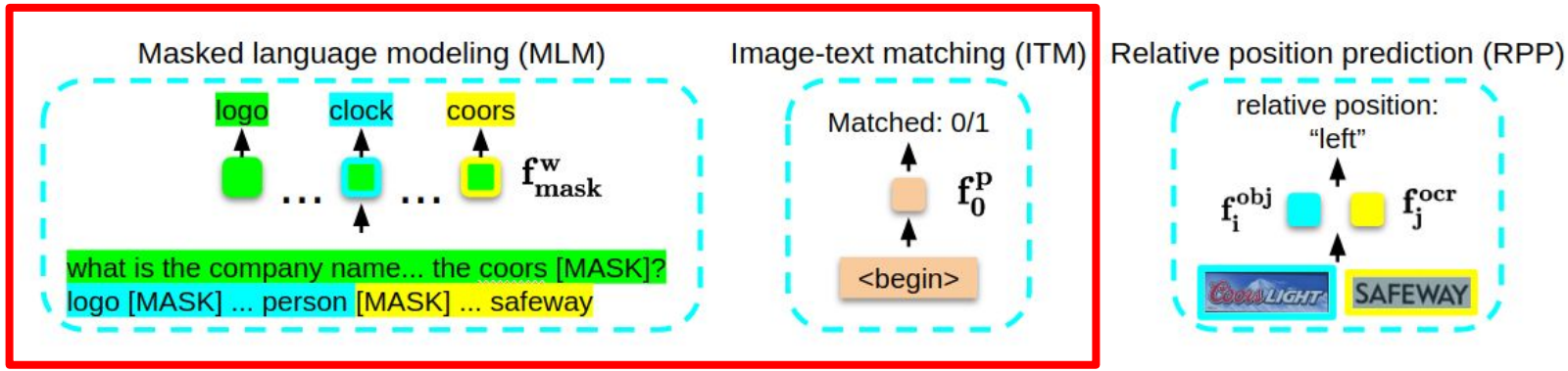
- Text word, visual object, <u>and scene text</u>

Method

- **Text-Object; Text-Scene text**

- Modifications over MLM, ITM
  - Adding OCR/object words

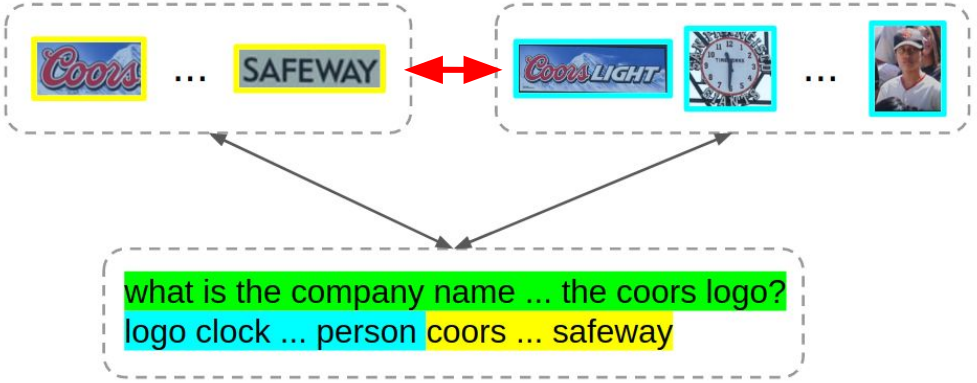


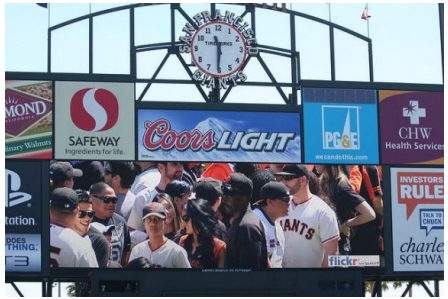


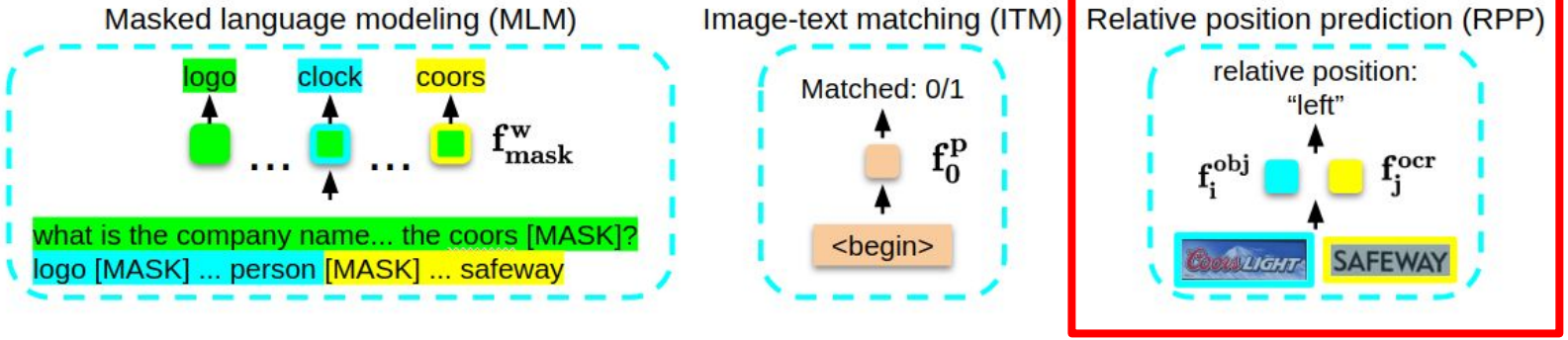
Method

- **Object-Scene text**

- Relative position prediction

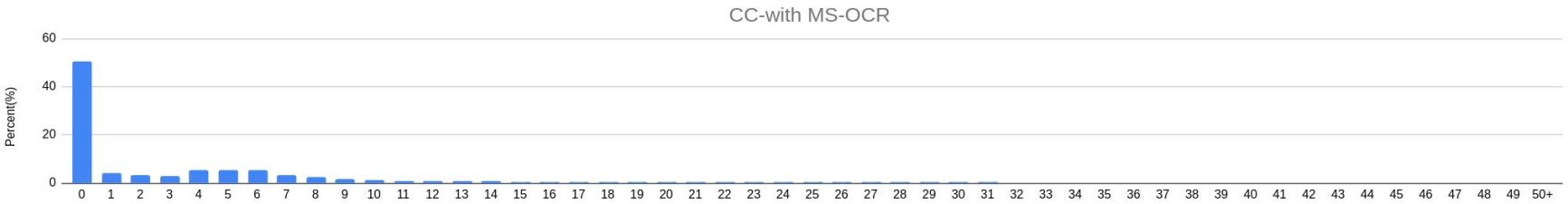


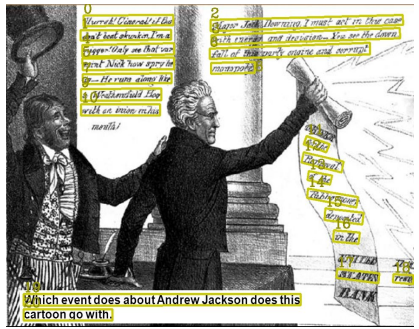




OCR-CC

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





Discarded images

#OCR words=0

Repeated watermarks only

Selected images

Images with 3-10 #OCR words

Images with >50 #OCR words

Datasets

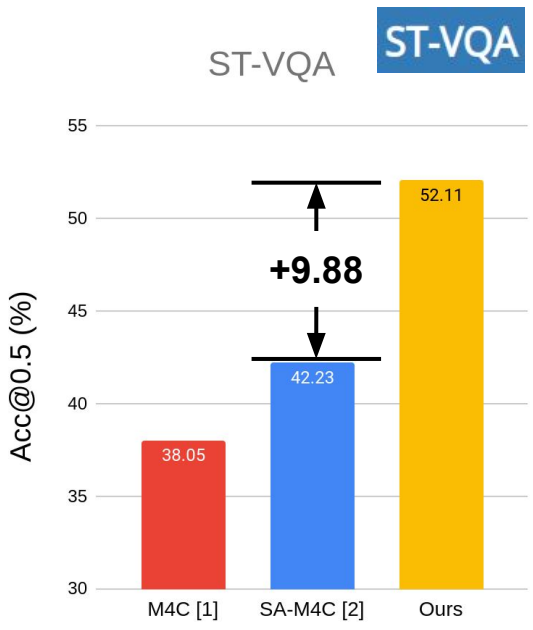
- Text-VQA: TextVQA, ST-VQA
- Text-captioning: TextCaps

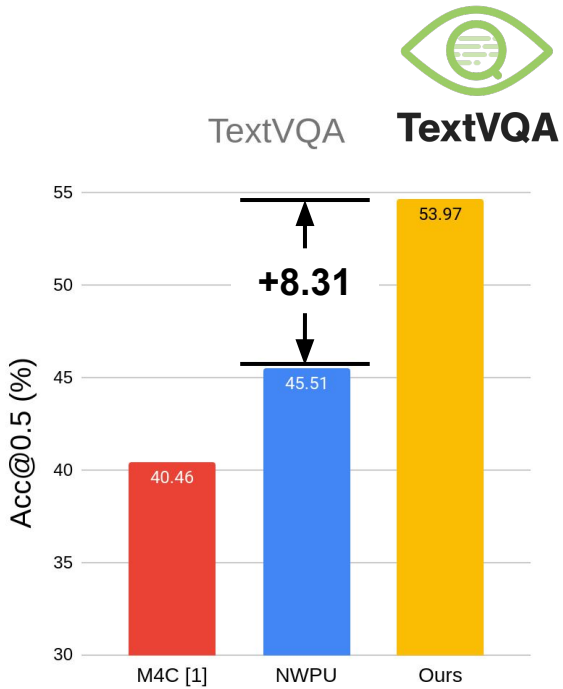
Question: what **number** is on the bike on the right? ---- A: the number is **317**

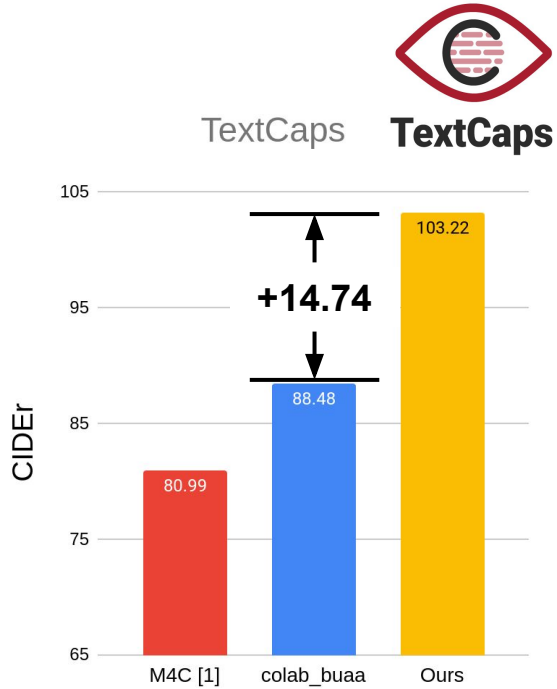
A group of motorcyclists with **number** **317, 44, 30, 338, 598** racing outdoor.

[1] Singh, Amanpreet, et al. "Towards vqa models that can read." In CVPR 2019.
[2] Biten, Ali Furkan, et al. "Scene text visual question answering." In CVPR 2019.
[3] Sidorov, Oleksii, et al. "TextCaps: a Dataset for Image Captioning with Reading Comprehension." In ECCV 2020.

Comparison to Other Methods







Experiment Results

- **Co-reference**

Coref Type	W/O TAP	With TAP
Text Word → Scene Text	0.0477	**0.3514**
Scene Text → Text Word	0.0473	**0.5206**
Visual Object → Scene Text	0.0045	**0.0130**
Scene Text → Visual Object	0.0337	**0.0680**

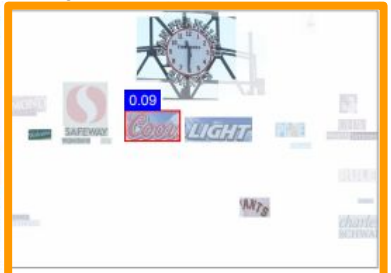
(a) who must survive?
M4C†: survive
GT: yaam

must

survive

(b) what is the company name to the left of the coors logo?
M4C†: coors light
GT: safeway

W/o TAP



coors

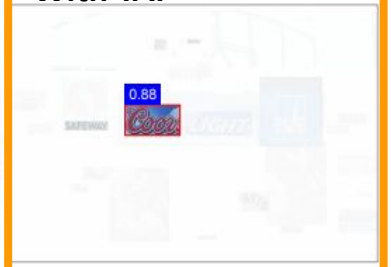
With TAP

Ours: yaam
GT: yaam

must

survive

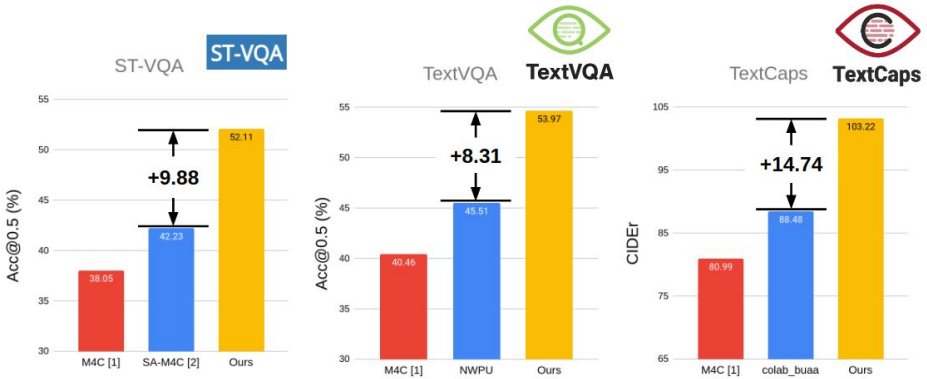
Ours: safeway
GT: safeway



coors

Text-aware pre-training for TextVQA and TextCaps

Key idea: joint representation learning for scene text vision language tasks



Code & data:
https://github.com/zyang-ur/TAP

Contact:
Zhengyuan Yang
zhengyuan.yang13@gmail.com