TAG: Boosting Text-VQA via Text-aware Visual Question-answer Generation

Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph F. JaJa, Larry S. Davis

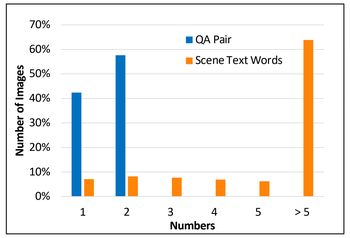
BMVC 2022

Introduction

Q: What is the speed limit of this road?
A: 75 mph

- Analyze issues of current Text-VQA models due to **biases of data annotation** and propose possible solutions to mitigate the issues
- **First architecture** to explore new **question-answer (QA) pairs generation** for Text-VQA task without additional annotated data
- We present consistently accurate results across **two vanilla Text-VQA methods on two datasets**

Motivation - Biases of Data Annotation



Statistics in TextVQA Training Set

Training Example

More Generated QA Pairs

Q: what is the alcoholic content of the middle bottle?
A: 9.2%

Q: what is the brand of the wine?
A: Trappistes

Q: what is the name on the first bottle from the left?
A: 100% satisfaction blend

Q: what is written on the left side of the red label on the left?
A: Sa4stacrie

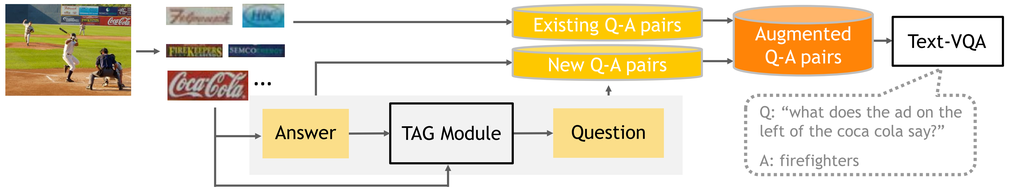
Q: what 3 numbers come after "d"?
A: 803

Q: what is the title of this book?
A: Oktett

Q: what is the slogan printed on the "thanks" advertisement behind these rugby players?
A: it's how we connect

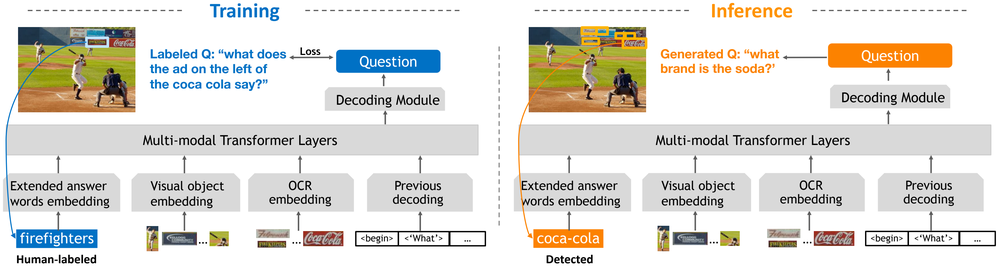
Q: what does the sign say on the left?
A: NK

Overall Architecture



The proposed **Text-VQA framework**. It consists of **two parts**: a text-aware visual question-answer generation module (TAG), followed by a downstream Text-VQA model. TAG takes **a single image and text words (the answer)** as input, and outputs a **newly-generated question** corresponding to the input answer. The generated QA pairs from TAG together with the originally labeled data are subsequently used **to train Text-VQA models**, leading to better Text-VQA performance.

TAG Overview



Generated QA Pairs

Q: what is the first ingredient listed?
A: water

Q: what is the name of this drink?
A: SAIGON

Q: what kind of tickets can you buy?
A: ferry

Q: what letter is in the green?
A: F

Q: what type of store is this?
A: duty free

Q: what is the name of the store?
A: URERTY

Q: what is the van's license plate number?
A: ls09 vgk

Q: what does the van say?
A: CTYPOLICE

Experiments on TextVQA

Method	OCR system	Extra Data	Val Acc.	Test Acc.
CRN [27]	Rosetta-en	×	40.39	40.96
LaAP-Net [16]	Rosetta-en	×	40.68	40.54
SMA [12]	SBD-Trans OCR	×	43.74	44.29
SSBaseline [44]	SBD-Trans OCR	×	43.95	44.72
LOGOS [28]	Microsoft-OCR	×	50.79	50.65
M4C† [17]	Microsoft-OCR	×	44.50	44.75
M4C† + TAG	Microsoft-OCR	×	**45.68**	**45.96**
TAP [41]	Microsoft-OCR	×	49.91	49.71
TAP + TAG	Microsoft-OCR	×	**52.54**	**52.57**
LaAP-Net [16]	Rosetta-en	ST-VQA	41.02	41.41
SA-M4C [20]	Google-OCR	ST-VQA	45.40	44.60
SMA [12]	SBD-Trans OCR	ST-VQA	44.58	45.51
SSBaseline [44]	SBD-Trans OCR	ST-VQA	45.53	45.66
LOGOS [28]	Microsoft-OCR	ST-VQA	51.53	51.08
M4C† [17]	Microsoft-OCR	ST-VQA	45.22	-
M4C† + TAG	Microsoft-OCR	ST-VQA	**46.33**	**46.38**
TAP [41]	Microsoft-OCR	ST-VQA	50.57	50.71
TAP + TAG	Microsoft-OCR	ST-VQA	**53.63**	**53.69**

Ablation Study on TextVQA

Each input modality matters

Ans.	Obj.	OCR.	Val Acc.
✓			48.76
✓		✓	48.95
✓	✓		49.13
✓	✓	✓	**52.54**

Impact of answer selection

Answer Selection	Val Acc.
random	49.26
largest	52.54
top three	52.73
top five	52.19

Experiments on ST-VQA

Method	Extra Data	Val Acc.	Val ANLS	Test ANLS
CRN [27]	×	-	-	0.483
LaAP-Net [16]	×	39.74	0.497	0.485
SMA [12]	×	-	-	0.486
SA-M4C [20]	×	42.23	0.512	0.504
SSBaseline [44]	×	-	-	0.509
LOGOS [28]	×	44.10	0.535	0.522
M4C† [17]	×	42.28	0.517	0.517
M4C† + TAG	×	**44.52**	**0.540**	**0.529**
TAP [41]	×	45.29	0.551	0.543
TAP + TAG	×	**50.18**	**0.595**	**0.586**
SSBaseline [44]	TextVQA	-	-	0.550
LOGOS [28]	TextVQA	48.63	0.581	0.579
M4C† [17]	TextVQA	46.60	0.560	0.552
M4C† + TAG	TextVQA	**48.69**	**0.579**	**0.571**
TAP†† [41]	TextVQA, TextCaps, OCR-CC	50.83	0.598	0.597
TAP + TAG	TextVQA	**53.53**	**0.620**	**0.602**