Scene Text Visual Question Answering

Ali Furkan Biten*[1], Rubèn Tito*[1], Andrés Mafla*[1], Lluis Gomez[1], Marçal Rusiñol[1], Ernest Valveny[1], C.V. Jawahar[2], Dimosthenis Karatzas[1]

[1]Computer Vision Center, UAB, Spain [2]CVIT, KCIS, IIIT Hyderabad, India

Q: What is the price displayed in large letters on the sign?
A: 14.99

Q: What is written on the sign?
A: Stop

Q: Where is this train going?
A: To New York
A: New York

Q: What is the exit number on the street sign?
A: 2
A: Exit 2

Q: What word in black comes below 1/2 price?
A: sale

Q: What company's logo is on the coffee cup?
A: STARBUCKS COFFEE

Q: What is written in the black rectangle?
A: Do not block driveway

Q: Which street sign is higher than the other?
A: HIGH

Q: What brand of alcohol is served at this establishment?
A: Guinness

Q: What is the name of the library one of the signs is pointing to?
A: Lee Wee Nam Library

Q: What is the name of the toy gun?
A: Bang

Q: What is something found in the Bible?
A: The Ten Commandments

Q: What is the automobile sponsor of the event?
A: kia

Overview

- Motivation: Current VQA datasets do not include **rich semantic information conveyed as text in an image** when reasoning about an image/question pair.
- Images are taken from several commonly used computer vision datasets.
- Question/Answer pairs were collected with Amazon Mechanical Turk.
- **New metric** is proposed that smoothly captures OCR precision as well as reasoning capabilities.
- **A competition and three novel tasks** are proposed with increasing levels of complexity.
- Several baselines are proposed and evaluated as a starting point, demonstrating the need of a generative model.

ST-VQA
Dataset and Leaderboard

ST-VQA Gathering and Statistics

Original Dataset	Images	Questions
Coco-text	7,520	10,854
Visual Genome	8,490	11,195
VizWiz	835	1,303
ICDAR	1,088	1,423
ImageNet	3,680	5,165
IIIT-STR	1,425	1,890
Total	23,038	31,791

- **Criteria**: Images with at least 2 text occurrences.

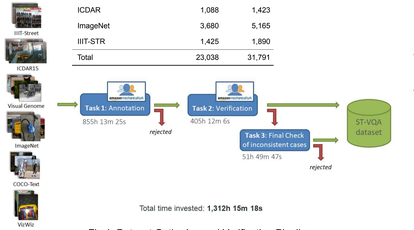

Fig 1. Dataset Gathering and Verification Pipeline.

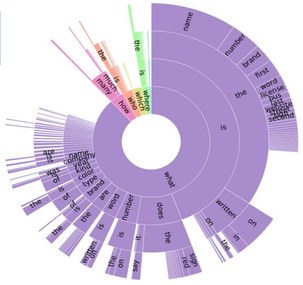

Fig 2. Distribution of questions in the ST-VQA train set by their starting 4-grams (ordered from center to outwards).

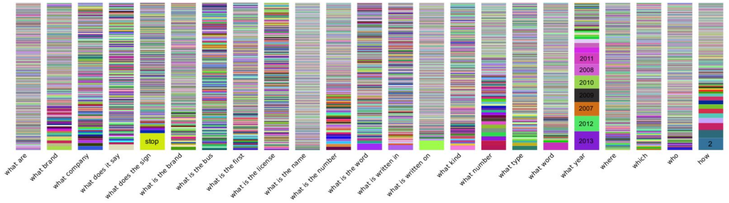

Figure 3. Distribution of answers for different types of questions in the ST-VQA train set. Each color represents a different unique answer.

Comparison with TextVQA

- The two datasets are highly complementary as the image sources used do not intersect with each other.
- We explicitly required a minimum amount of two text instances to be present, while in TextVQA images were sampled on a category basis.
- The answers in ST-VQA are always the text found in the image, whereas in Text-VQA, the answer can be anything as long as text is relevant to answer.

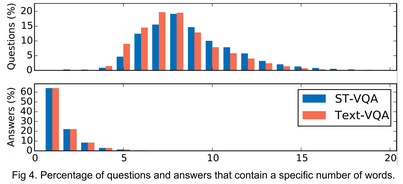

Fig 4. Percentage of questions and answers that contain a specific number of words.

Metric: Average Normalized Levenshtein Similarity (ANLS)

- Tracks accuracy.
- Based on string edit distance.
- Deals with border cases (Correct reasoning but partially wrong recognition).

$$ANLS = \frac{1}{N}\sum_{i=0}^{N}\left(\max_j s(a_{ij}, o_{qi})\right) \quad \text{, where} \quad s(a_{ij}, o_{qi}) = \begin{cases} \left(1 - NL(a_{ij}, o_{qi})\right) & if\ NL(a_{ij}, o_{qi}) < \tau \\ 0 & if\ NL(a_{ij}, o_{qi}) \geq \tau \end{cases}$$

Question	What is the number of Boat?			Who is the sponsor of this tennis match?			What does the black label say?			Who is the author of this book?		
GT	25686			KIA			GemOro			Maxwell Anderson		
	Output	Accuracy	ANLS	Output	Accuracy	ANLS	Output	Accuracy	ANLS	Output	Accuracy	ANLS
METHOD A	5686	0	0.8	kin	0	0.666	gemors	0	0.83	lorraine	0	0
METHOD B	3	0	0	kia	100	1	gemoro	100	1	anderson	0	0.5

ST-VQA Tasks

What is written on the piece of material above to walking men?
K LUNC

What is the wattage of the vacuum cleaner?
1400W

How do the Omnibuses stop?
BY REQUEST

What did the sign say originally?
ONE WAY

TASK 1

Strongly Contextualized Dictionary (100 words per image)

SERIES, NATIONAL, BACK, SECOND, ANOTHER, OUR, JULY, MEMBER, **K**, POWER, CUP, WEST, SINCE, DUE, **LUNC**, 1, HIGH, CHAMPIONSHIP, MARCH, AROUND, ...

Strongly Contextualized Dictionary (100 words per image)

SERIES, IMAGE, SITE, DUE, HIGH, USING, SHIP, OCTOBER, **1400W**, END, BUILT, SERVICE, AREA, WEST, BEGAN, SINCE, SYSTEM, LONG, MEMBER, 0, PROGRAM, ...

Strongly Contextualized Dictionary (100 words per image)

FOUR, ANOTHER, STATION, STILL, SEVERAL, USE, UNITED, BUILT, BEGAN, **REQUEST**, DUE, LONG, MEMBER, AIR, MAIN, TYPE, OMNIBUSES, GOVERNMENT, WATER, YORK, ...

Strongly Contextualized Dictionary (100 words per image)

ONE, RIGHT, FORCE, OFTEN, FAMILY, POINT, NATIONAL, BACK, **WAY**, BUILT, AREA, BEGAN, DUE, LONG, FRANCE, MUCH, PARTY, LIFE, CHAMPIONSHIP, MARCH, ...

TASK 2

Weakly Contextualized Dictionary (30,000 words, same for all test set images)

CHAMPIONSHIP, LAST, MARCH, SIDE, NATIONAL, OUR, FAMILY, FRANCE, CAR, JOHN, MEMBER, WORK, **REQUEST**, NOW, BUILDING, ISLAND, AMERICAN, MEMBER, SIDE, NATIONAL, **WAY**, MUCH, BRITISH, SYSTEM, SERIES, LINE, RIGHT, SINCE, HOLE, CHAMPIONSHIP, WATER, FORCE, ONEGA, PARTY, LUNC, SINCE, LIFE, BACK, **LUNC**, DAY, BEGAN, SHIP, DAY, DUE, USING, WILL, **K**, SEVERAL, NOW, LEFT WORK, BUILT, MEMBER, MARCH, UNITED, OCTOBER, **1400W**, CUP, **ONE**, SECOND, WILL, SOUTH, FOUR, TRAIN, HIGH, AMERICAN, BRITISH, LONG, OMNIRUSES, FOUR, ANOTHER, LONG, KING, TYPE, WATER, DUE, DUE, GOVERNMENT, ELBA, RIVER, **BY**, PROGRAM, PLACE, BUILDING, AIR, SOUTH, BEGAN, LARGE, NOW, WORK, END, USE, SEVERAL, . . .

TASK 3

Open Dictionary (no words provided)

Baseline Qualitative Results

Q: What brand are the machines?
A: bongard
STR (bbox): 1
Scene Image OCR: zbongard
SAAA+STR: aeropostale
SAN(CNN): ray
VTA: boar
USTB-TQA: boar
QAQ: bongard
Clova AI OCR: bongard

Q: Where is the high court located?
A: delhi
STR (bbox): delhi
Scene Image OCR: high
SAAA+STR: delhi
SAN(CNN): delhi
VTA: delhi court
USTB-TQA: delhi
QAQ: delhi
Clova AI OCR: facal facal

Q: What is the route of the bus?
A: purple route
STR (bbox): purple
Scene Image OCR: 1208
SAAA+STR: crosstown
SAN(CNN): 66
VTA: purple route
USTB-TQA: purple
QAQ: north
Clova AI OCR: arts- arts-

Q: What's the street name?
A: place d'armes
STR (bbox): dames
Scene Image OCR: d'armes
SAAA+STR: W 159 st
SAN(CNN): 10th st
VTA: brasington
USTB-TQA: d'
QAQ: d'armes
Clova AI OCR: place d'armes

Baseline Results

Method	OCR	Q	V	Task 1 ANLS	Task 1 Acc.	Task 2 ANLS	Task 2 Acc.	Task 3 ANLS	Task 3 Acc.	Upper bound ANLS	Upper bound Acc.
Random	✗	✗	✗	0.015	0.96	0.001	0.00	0.00	0.00	-	-
STR (retrieval)	✓	✗	✗	**0.171**	**13.78**	0.073	5.55	-	-	0.782	68.84
STR (bbox)	✓	✗	✗	0.130	7.32	0.118	6.89	0.128	7.21	-	-
Scene Text Image OCR	✓	✗	✗	0.145	8.89	0.132	8.69	**0.140**	8.60	-	-
SAAA (1k cls)	✗	✓	✓	0.085	6.36	0.085	6.36	0.085	6.36	0.571	31.96
SAAA+STR (1k cls)	✓	✓	✓	0.091	6.66	0.091	6.66	0.091	6.66	0.571	31.96
SAAA (5k cls)	✗	✓	✓	0.087	6.66	0.087	6.66	0.087	6.66	0.740	41.03
SAAA+STR (5k cls)	✓	✓	✓	0.096	7.41	0.096	7.41	0.096	7.41	0.740	41.03
SAAA (19k cls)	✗	✓	✓	0.084	6.13	0.084	6.13	0.084	6.13	0.862	52.31
SAAA+STR (19k cls)	✓	✓	✓	0.087	6.36	0.087	6.36	0.087	6.36	0.862	52.31
QA+STR (19k cls)	✓	✓	✗	0.069	4.65	0.069	4.65	0.069	4.65	0.862	52.31
SAN(LSTM)	✗	✓	✓	0.102	7.78	0.102	7.78	0.102	7.78	0.740	41.03
SAN(LSTM)+STR (5k cls)	✓	✓	✓	0.136	10.34	**0.136**	**10.34**	0.136	**10.34**	0.740	41.03
SAN(CNN)+STR (5k cls)	✓	✓	✓	0.135	10.46	**0.135**	**10.46**	0.135	**10.46**	0.740	41.03

ICDAR ST-VQA Competition Results

Method	Task 1 Shared	Task 1 Specific	Task 1 Total	Task 1 Acc.	Task 1 Dict.	Task 2 Shared	Task 2 Specific	Task 2 Total	Task 2 Acc.	Task 2 Dict.	Task 3 Shared	Task 3 Specific	Task 3 Total	Task 3 Acc.
VTA	0.507	0.501	**0.506**	**43.52**	100	0.280	0.268	**0.279**	**17.77**	48.91	0.280	0.285	**0.282**	18.13
USTB-TQA	0.457	0.445	0.445	39.98	97.05	0.168	0.196	0.173	13.34	84.11	0.168	0.183	0.170	13.14
USTB-TVQA	0.129	0.100	0.124	10.09	20.55	0.093	0.094	0.093	6.59	83.76	0.093	0.108	0.095	6.86
Focus	0.300	0.275	0.295	24.45	68.84	0.080	0.081	0.080	4.16	58.84	0.088	0.089	0.088	4.42
VQA-DML	0.142	0.138	0.141	11.63	99.97	-	-	-	-	-	-	-	-	-
TMT	0.076	0.045	0.055	4.53	13.80	-	-	-	-	-	-	-	-	-
QAQ	-	-	-	-	-	-	-	-	-	-	0.255	0.265	0.256	**19.19**
Clova AI OCR	-	-	-	-	-	-	-	-	-	-	0.213	0.224	0.215	12.53

Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Mathew, M., Jawahar, CV, Valveny, E. & Karatzas, D. (2019). ICDAR 2019 Competition on Scene Text Visual Question Answering. *15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*