Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

Jihyeon Lee, *July 21th, 2020*

Motivation

VQA(Visual Question Answering) models can not read! -> TextVQA

- VQA tasks tend to ignore a crucial modality -- text in the images – that carries essential information
- TextVQA tasks explicitly require understanding and reasoning about text in the image
- Input question, visual contents, and the text in the image
- Gap between human performance and machine performance is significantly larger on TextVQA than on VQA 2.0

TextVQA

Question: *What is the danger?*

Previous work: *water*

Our model: *deep water*

Related Work

LoRRA(Look, Read, Reason & Answer)

- Towards VQA Models That Can Read (CVPR 2019)
- Extends previous VQA models with an OCR attention branch and adds OCR tokens as a vocab to the answer classifier
- Input question, visual contents, and the text in the image

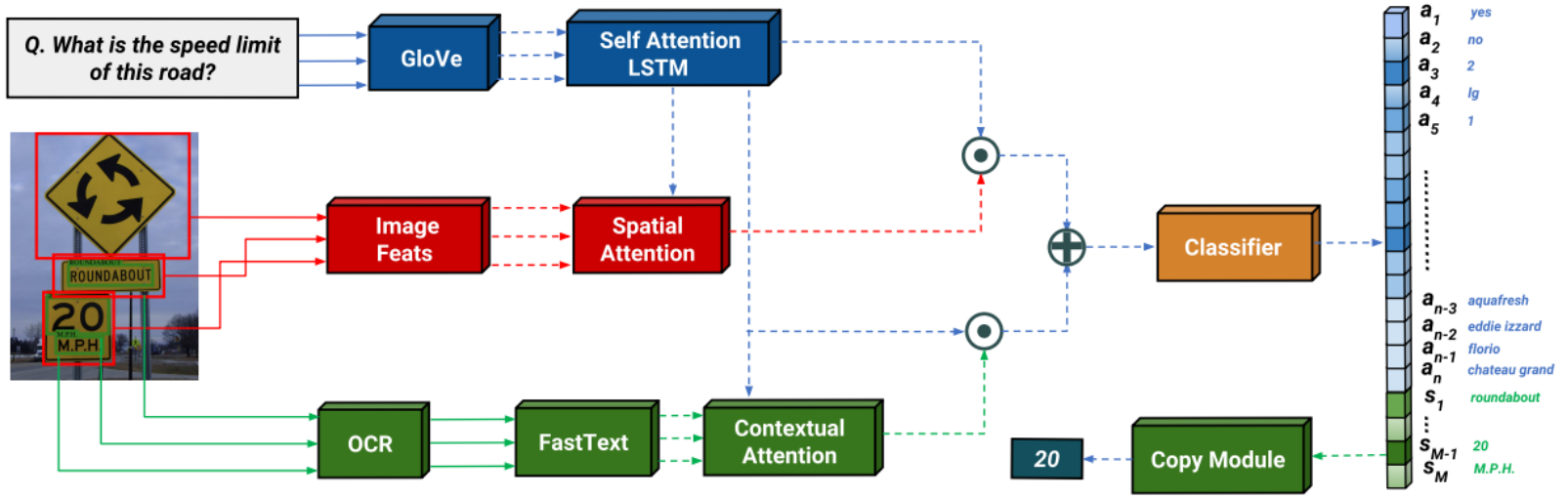


Figure 2: **Overview of our approach *Look, Read, Reason & Answer* (LoRRA)**. Our approach looks at the image, reads its text, reasons about the image and text content and then answers, either with an answer a from the fixed answer vocabulary or by selecting one of the OCR strings s. Dashed lines indicate components that are not jointly-trained. The answer cubes on the right with darker color have more attention weight. The OCR token "20" has the highest attention weight in the example.

Related Work

LoRRA(Look, Read, Reason & Answer)

- Rely on multimodal mechanisms only between two modalities -> limit the types of possible interactions
- Answer prediction as a single-step classification problem -> ex) McDonald's burger
- Miss important cues such as appearance and location

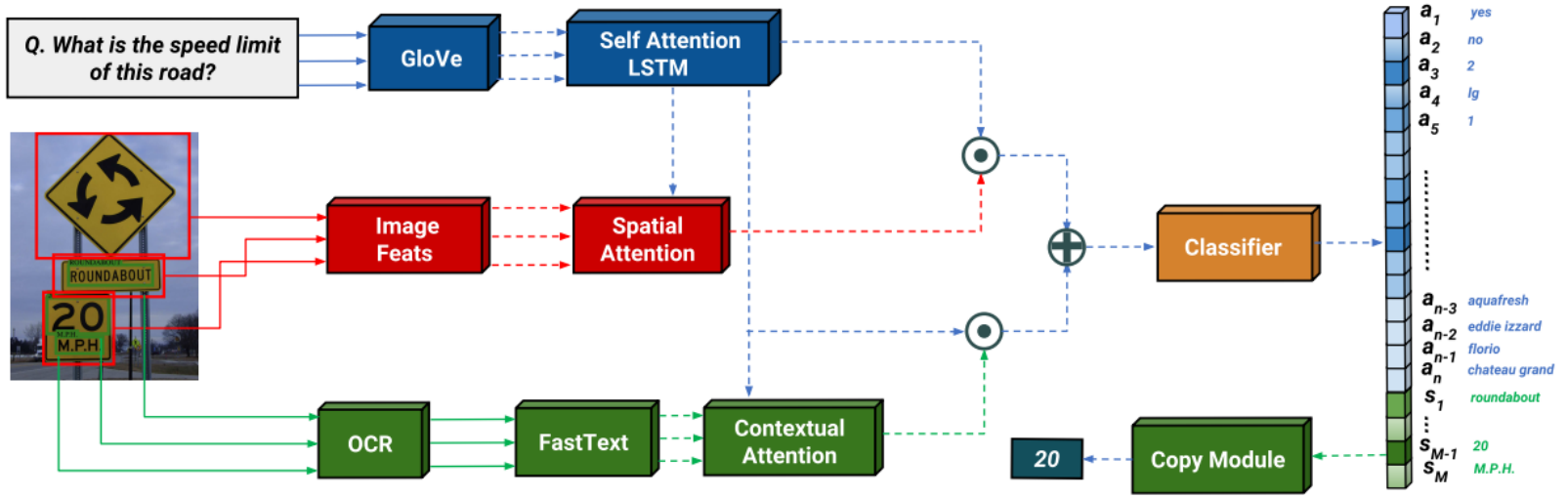
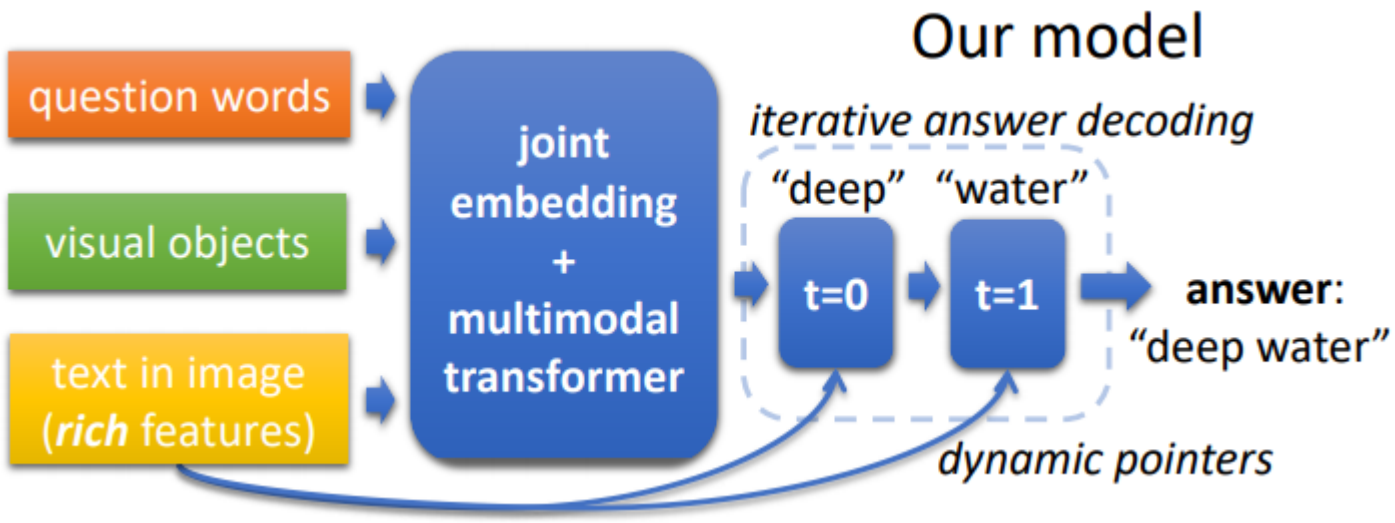


Figure 2: **Overview of our approach *Look, Read, Reason & Answer* (LoRRA)**. Our approach looks at the image, reads its text, reasons about the image and text content and then answers, either with an answer a from the fixed answer vocabulary or by selecting one of the OCR strings s. Dashed lines indicate components that are not jointly-trained. The answer cubes on the right with darker color have more attention weight. The OCR token "20" has the highest attention weight in the example.

Contribution

1. Show that multiple input modalities can be naturally fused through **multimodal transformer architecture**
2. Unlike previous work, model predicts the answer through **pointer-augmented multi-step decoder**
3. Adopt a **rich feature representation** for text tokens in images
4. Model **significantly outperforms** previous work on three challenging datasets for TextVQA



Datasets

TextVQA

What is the largest denomination on table?

Ground Truth	Prediction
500	unknown

What is the top oz?

Ground Truth	Prediction
16	red

ST-VQA

Q: What is the price of the bananas per kg?
A: $11.98

Q: What does the red sign say?
A: Stop

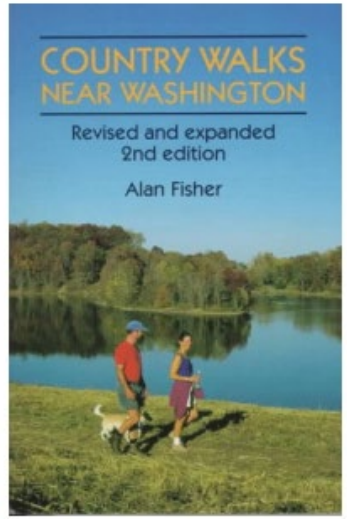
Q: Where is this train going?
A: To New York
A: New York

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

OCR-VQA

(c)

Q. Which year's calendar is this?
A. 2016
Q. Is this an exam preparation book?
A. No



(d)

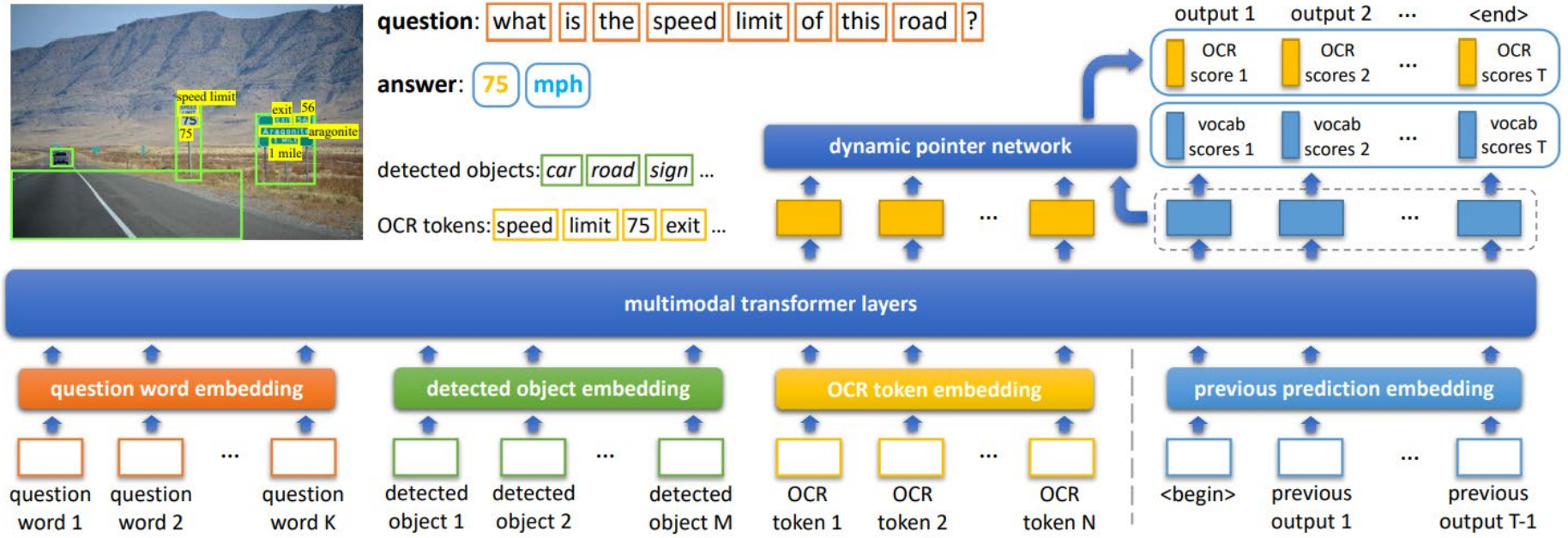
Q. Who wrote this book?
A. Alan Fisher
Q. What is the edition of this book?
A. 2

Model

M4C(Multimodal Multi-Copy Mesh)

1) Common d-dimensional semantic space for all modalities through domain specific embedding

- x_k^{ques} : pretrained BERT
- $x_m^{obj} = LN(W_1 x_m^{fr} + W_2 x_m^b)$: Faster R-CNN
- $x_n^{ocr} = LN\left(W_3 x_n^{ft} + W_4 x_n^{fr} + W_5 x_n^p\right) + LN(W_6 x_n^b)$: FastText + Faster R-CNN + PHOC

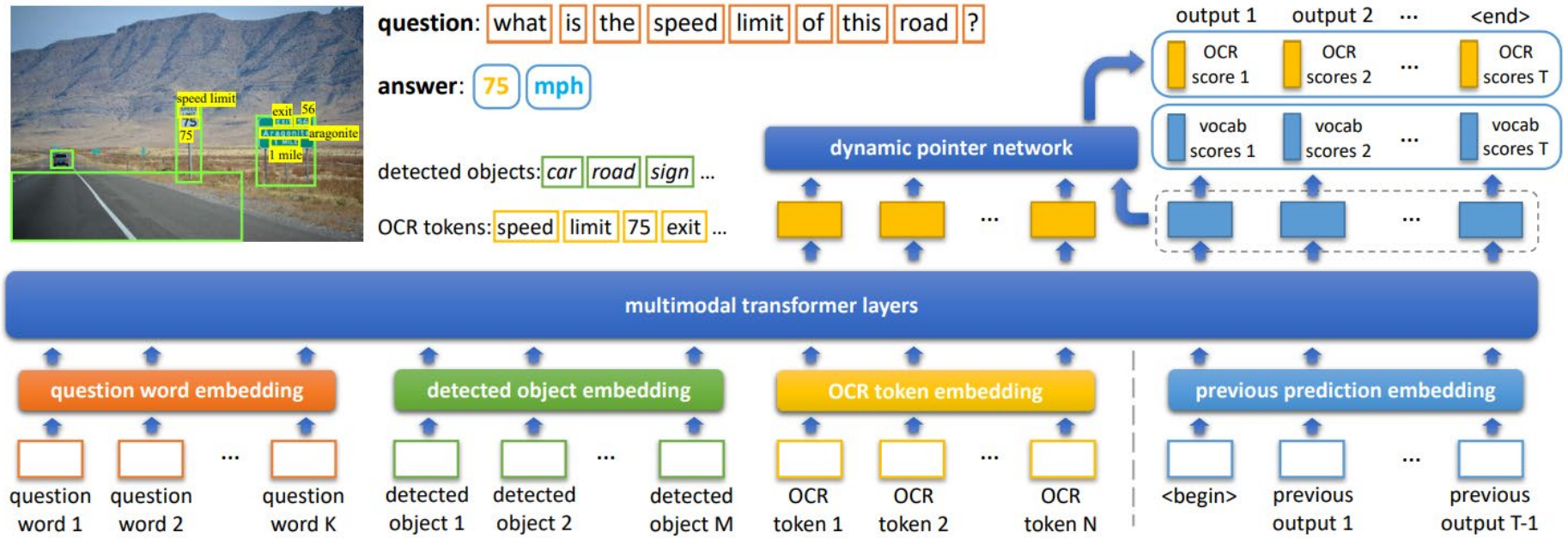


Model

M4C(Multimodal Multi-Copy Mesh)

2) Multimodal fusion and iterative answer prediction with pointer-augmented transformers

- L transformer layers over the $K + M + N$ entities from $\{x_k^{ques}\}, \{x_m^{obj}\}$, and $\{x_n^{ocr}\}$
- Enables modeling both inter- and intra- modality relations in a homogeneous way through same set of parameters
- Decode the answer in an auto-regressive manner for a T steps, each word may be either an OCR token or fixed vocab

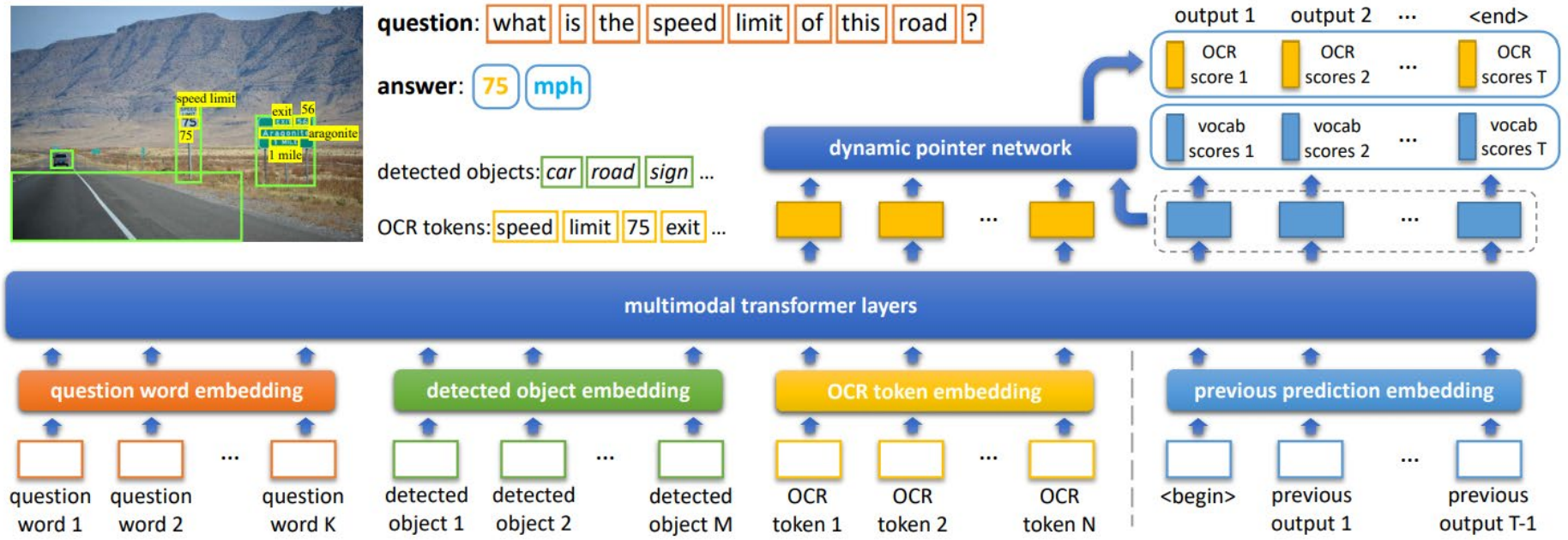


Model

M4C(Multimodal Multi-Copy Mesh)

2) Multimodal fusion and iterative answer prediction with pointer-augmented transformers

- $y_{t,i}^{voc} = (w_i^{voc})^T z_t^{dec} + b_i^{voc}$; $y_{t,n}^{ocr} = (W^{ocr} z_n^{ocr} + b^{ocr})^T (W^{dec} z_t^{dec} + b^{dec})$; $y_t^{all} = [y_t^{voc}; y_t^{ocr}]$
- Feed in x_n^{ocr} or w_i^{voc} as the next step's input x_{t+1}^{dec}
- Multi-label sigmoid loss (instead of softmax loss)



Experiments

Evaluation on TextVQA datset

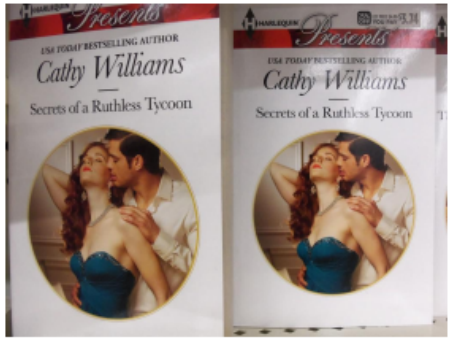
#	Method	Question enc. pretraining	OCR system	OCR token representation	Output module	Accu. on val	Accu. on test
1	LoRRA [44]	GloVe	Rosetta-ml	FastText	classifier	26.56	27.63
2	M4C w/o dec.	GloVe	Rosetta-ml	FastText	classifier	29.36	–
3	M4C w/o dec.	(none)	Rosetta-ml	FastText	classifier	29.55	–
4	M4C w/o dec.	BERT	Rosetta-ml	FastText	classifier	30.15	–
5	M4C w/o dec.	BERT	Rosetta-en	FastText	classifier	31.28	–
6	M4C w/o dec.	BERT	Rosetta-en	FastText + bbox	classifier	33.32	–
7	M4C w/o dec.	BERT	Rosetta-en	FastText + bbox + FRCN	classifier	34.38	–
8	M4C w/o dec.	BERT	Rosetta-en	FastText + bbox + FRCN + PHOC	classifier	**35.70**	–
9	M4C (ours - ablation)	(none)	Rosetta-ml	FastText + bbox + FRCN + PHOC	decoder	36.06	–
10	M4C (ours - ablation)	BERT	Rosetta-ml	FastText + bbox + FRCN + PHOC	decoder	37.06	–
11	M4C (ours)	BERT	Rosetta-en	FastText + bbox + FRCN + PHOC	decoder	**39.40**	39.01
12	DCD_ZJU (ensemble) [32]	–	–	–	–	31.48	31.44
13	MSFT_VTI [46]	–	–	–	–	32.92	32.46
14	M4C (ours; w/ ST-VQA)	BERT	Rosetta-en	FastText + bbox + FRCN + PHOC	decoder	**40.55**	**40.46**

Table 1. On the TextVQA dataset, we ablate our M4C model and show a detailed comparison with prior work LoRRA [44]. Our multimodal transformer (line 3 vs 1), our rich OCR representation (line 8 vs 5) and our iterative answer prediction (line 11 vs 8) all improve the accuracy significantly. Notably, our model still outperforms LoRRA by 9.5% (absolute) even when using fewer pretrained parameters (line 9 vs 1). Our final model achieves 39.01% (line 11) and 40.46% (line 14) test accuracy without and with the ST-VQA dataset as additional training data respectively, outperforming the challenge-winning DCD_ZJU method by 9% (absolute). See Sec. 4.1 for details.

Experiments

Evaluation on TextVQA datset

What does the light sign read on the farthest right window?
LoRRA: **exit**
M4C (ours): **bud light**
human: **bud light**; **all 2 liters**



Who is usa today's bestselling author?
LoRRA: **roger zelazny**
M4C (ours): **cathy williams**
human: **cathy williams**

What is the name of the band?
LoRRA: **7**
M4C (ours): **soul doubt**
human: **soul doubt**; **h. michael karshis**; **unanswerable**

what is the time?
LoRRA: **1:45**
M4C (ours): **3:44**
human: **5:40**; **5:41**; **5:42**; **8:00**

Figure 4. Qualitative examples from our M4C model on the TextVQA validation set (orange words are from OCR tokens and blue words are from fixed answer vocabulary). Compared to the previous work LoRRA [44] which selects one answer from training set or copies only a single OCR token, our model can copy multiple OCR tokens and combine them with its fixed vocabulary through iterative decoding.

Experiments

Evaluation on ST-VQA datset

#	Method	Output module	Accu. on val	ANLS on val	ANLS on test
1	SAN+STR [8]	–	–	–	0.135
2	VTA [7]	–	–	–	0.282
3	M4C w/o dec.	classifier	33.52	0.397	–
4	M4C (ours)	decoder	**38.05**	**0.472**	**0.462**

What is the name of the street on which the Stop sign appears?
prediction: 45th parallel dr
GT: **45th parallel dr**

What does the white sign say?
prediction: tokyo station
GT: **tokyo station**

How many cents per pound are the bananas?
prediction: 99
GT: **99**

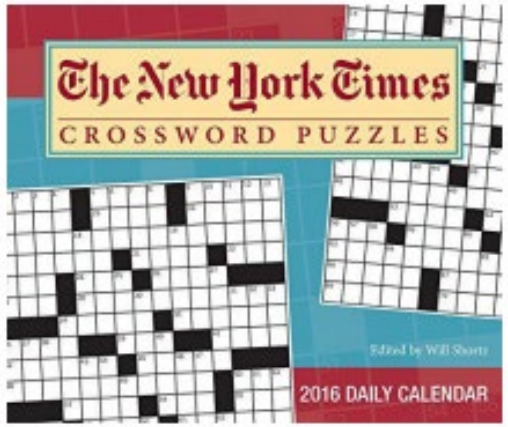
What kind of stop sign is in the image?
prediction: stop all way
GT: **all way**

Figure 5. Qualitative examples from our M4C model on the ST-VQA validation set (orange words from OCR tokens and blue words from fixed answer vocabulary). Our model can select multiple OCR tokens and combine them with its fixed vocabulary to predict an answer.

Experiments

Evaluation on OCR-VQA datset

#	Method	Output module	Accu. on val	Accu. on test
1	BLOCK [37]	–	–	42.0
2	CNN [37]	–	–	14.3
3	BLOCK+CNN [37]	–	–	41.5
4	BLOCK+CNN+W2V [37]	–	–	48.3
5	M4C w/o dec.	classifier	46.3	–
6	M4C (ours)	decoder	**63.5**	**63.9**



Who is the author of this book?
prediction: the new york times
GT: **the new york times**

Is this a pharmaceutical book?
prediction: no
GT: **no**

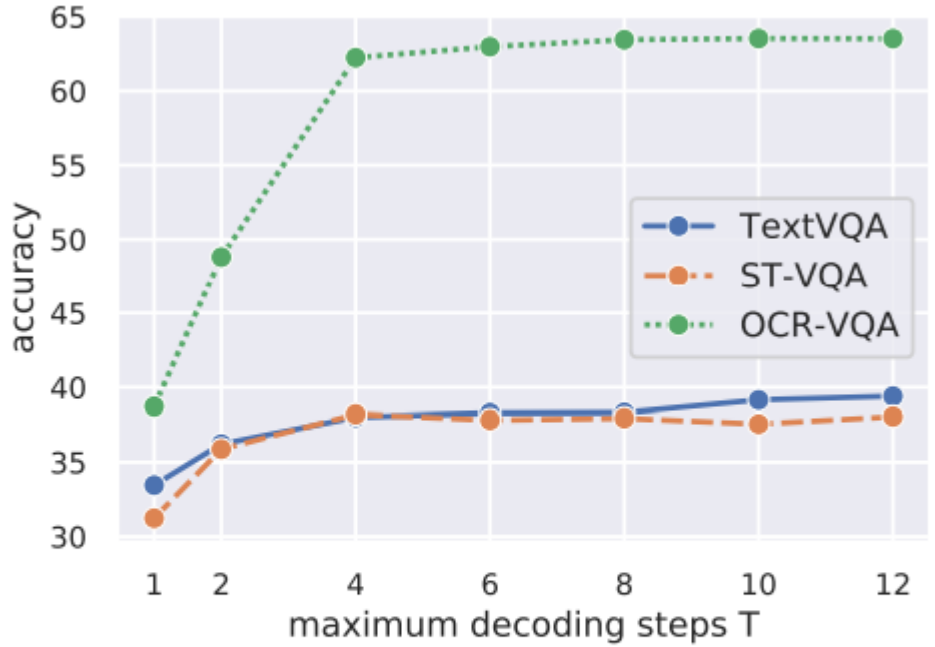


Figure 3. Accuracy under different maximum decoding steps T on the validation set of TextVQA, ST-VQA, and OCR-VQA. There is a major gap between single-step ($T = 1$) and multi-step ($T > 1$) answer prediction. We use 12 steps by default in our experiments.

Conclusion

- Show that multiple input modalities can be naturally fused through **multimodal transformer architecture**
- Unlike previous work, model predicts the answer through **pointer-augmented multi-step decoder**
- Adopt a **rich feature representation** for text tokens in images
- Model **significantly outperforms** previous work on three challenging datasets for TextVQA

❖ ***It is efficient to handle multiple modalities through domain-specific embedding followed by homogeneous self-attention and to generate complex answers as multi-step decoding instead of one-step classification!***