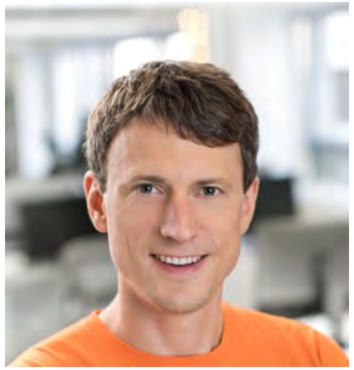
Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

Ronghang Hu[1,2], Amanpreet Singh[1], Trevor Darrell[2], Marcus Rohrbach[1]

[1]Facebook AI Research (FAIR) [2]University of California, Berkeley



Background – the TextVQA task

(traditional) Visual Question Answering

Question: What color is the hydrant?
Answer: yellow

our task:

TextVQA – understanding texts in images

Question: what is the speed limit?
Answer: 75 mph

The challenges of the TextVQA task

Challenge 1 – how to jointly reason over *three modalities*:

- **the question**
- **visual objects** in the image
- **text (OCR) tokens** in the image

Question: **what is the speed limit?**
Answer: 75 mph

The challenges of the TextVQA task

Challenge 2 – how to represent text (OCR) tokens:

- ***semantic*** cue – what does it mean
- ***spatial*** cue – where is it
- ***visual*** cue – what does it look like

Question: what is the speed limit?
Answer: 75 mph

The challenges of the TextVQA task

Challenge 3 – how to predict multi-word answers and mix words from two sources:

- predicted from fixed vocabulary
- copied from OCR results

Question: what is the speed limit?
Answer: 75 mph

Our model – M4C

The contributions of M4C:

1. **handling 3 modalities** (question, visual object, OCR) with joint embedding and multimodal transformers

2. **rich OCR features** – capturing semantic, spatial, and visual cues

3. **iterative answer decoding** – predict the answer word by word with a decoder along with a pointer network for OCR copying

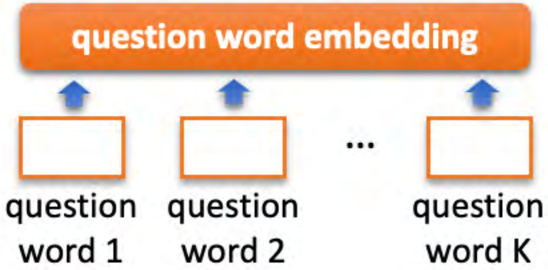
25%+ improvement on 3 datasets (TextVQA, ST-VQA, OCR-VQA)

The M4C model – multimodal joint embedding

- Step 1: embed all input modalities to a common semantic space

question: what is the speed limit of this road ?

answer:



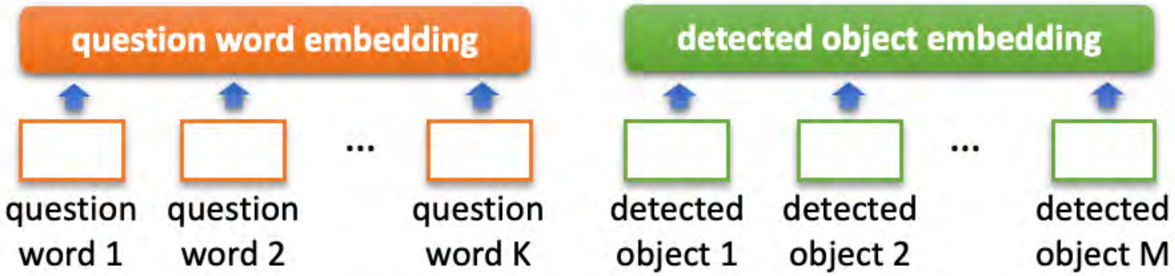
The M4C model – multimodal joint embedding

- Step 1: embed all input modalities to a common semantic space

question: what is the speed limit of this road ?

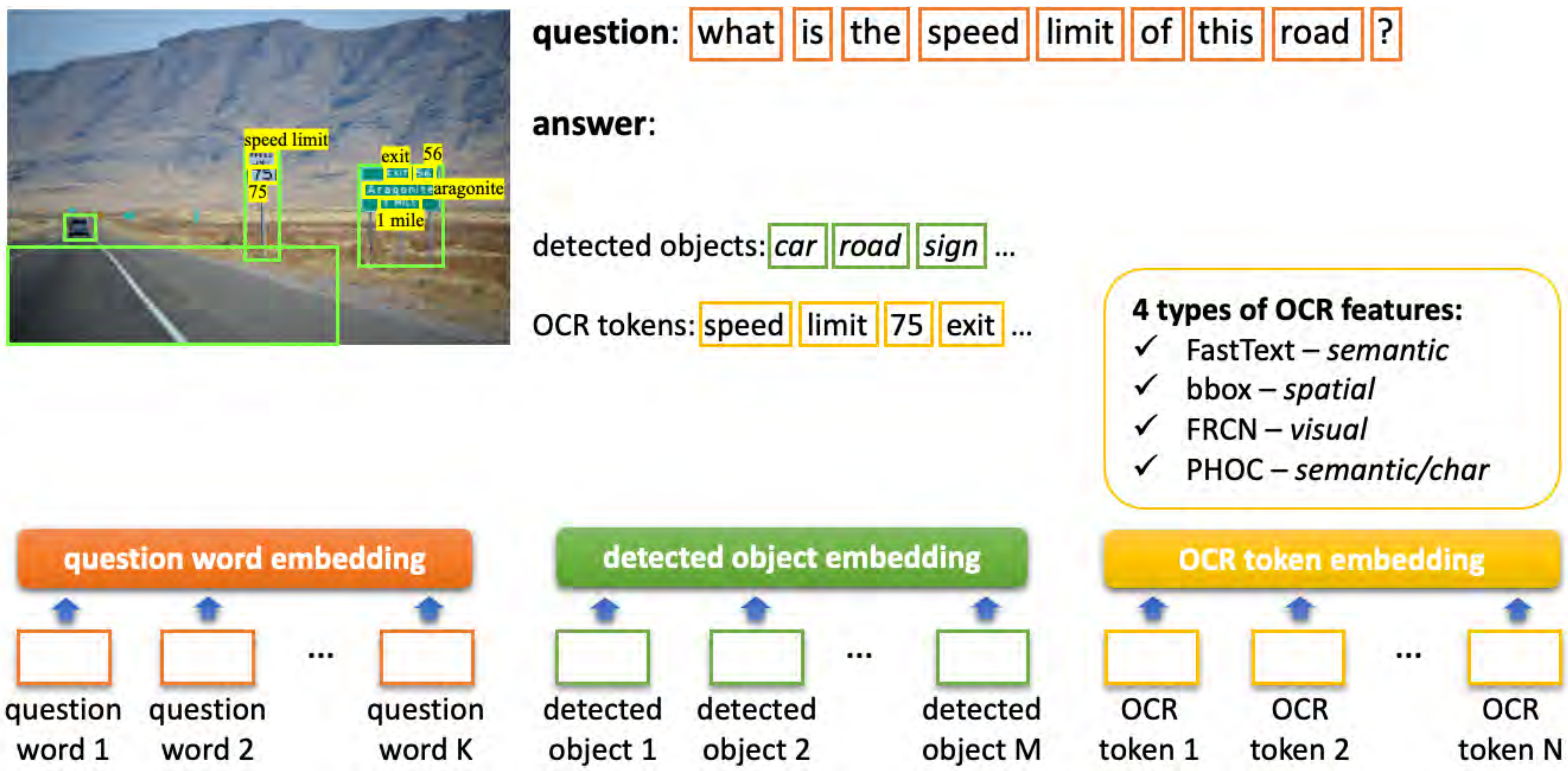
answer:

detected objects: *car* *road* *sign* ...



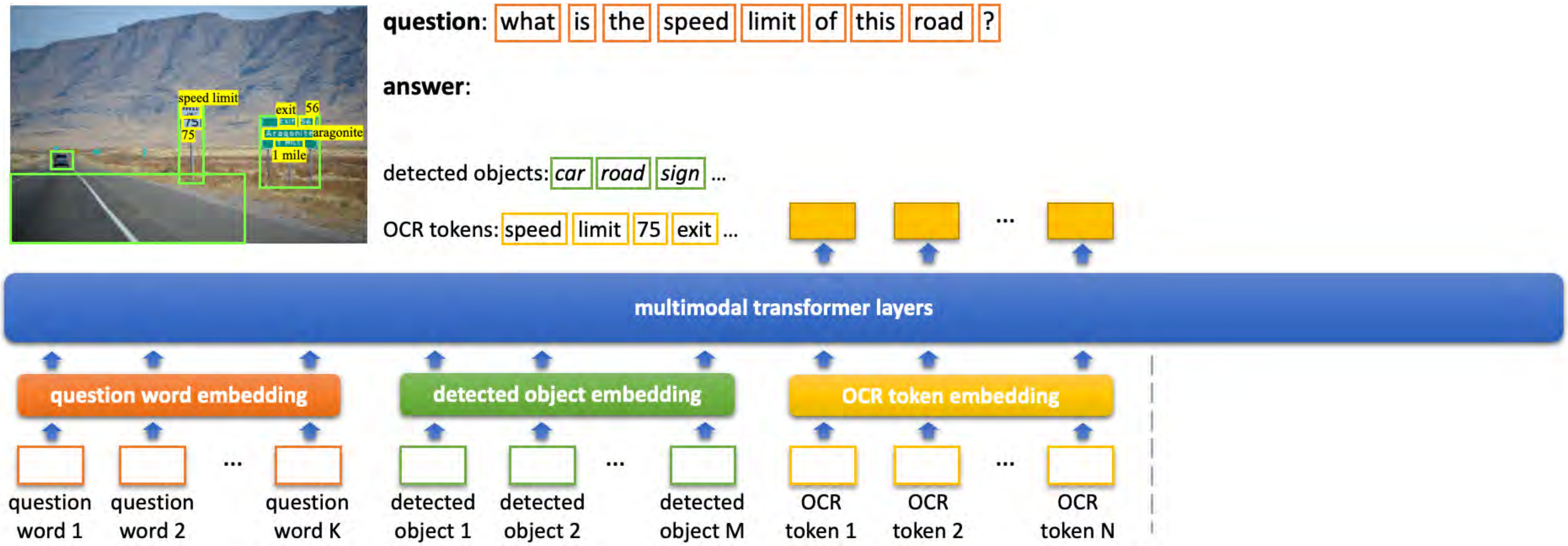
The M4C model – multimodal joint embedding

- Step 1: embed all input modalities to a common semantic space



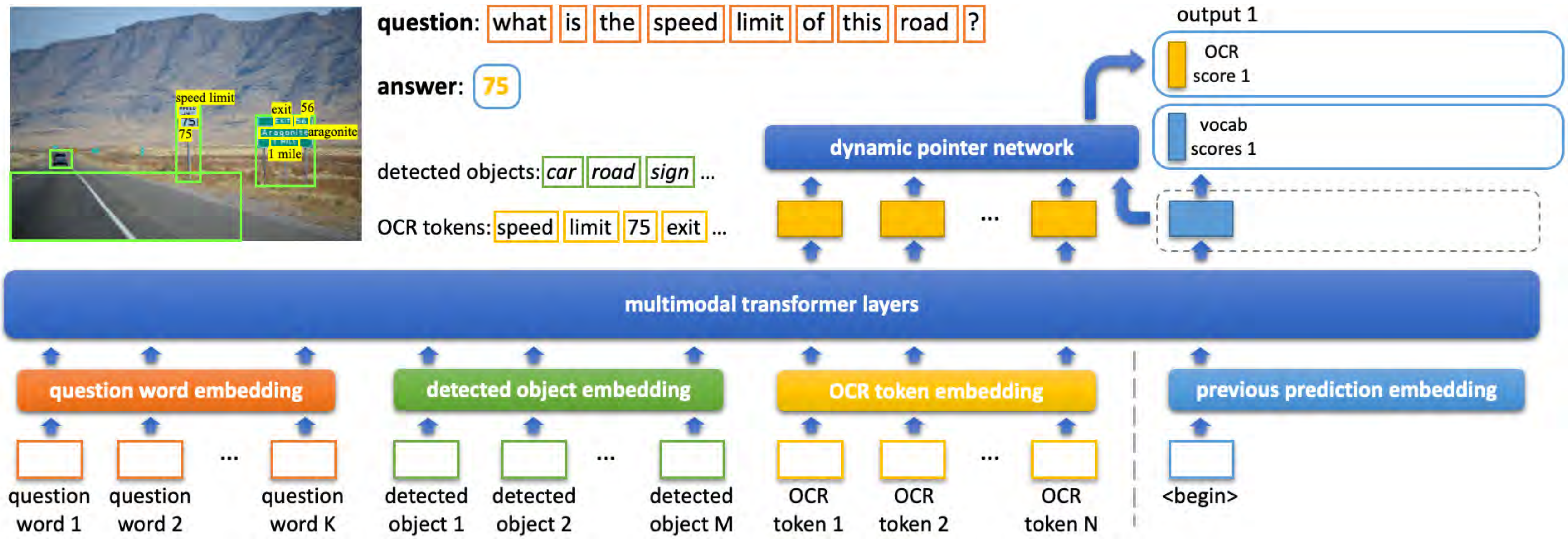
The M4C model – multimodal transformers

- Step 2: rich fusion – self-attention over embeddings from all modalities



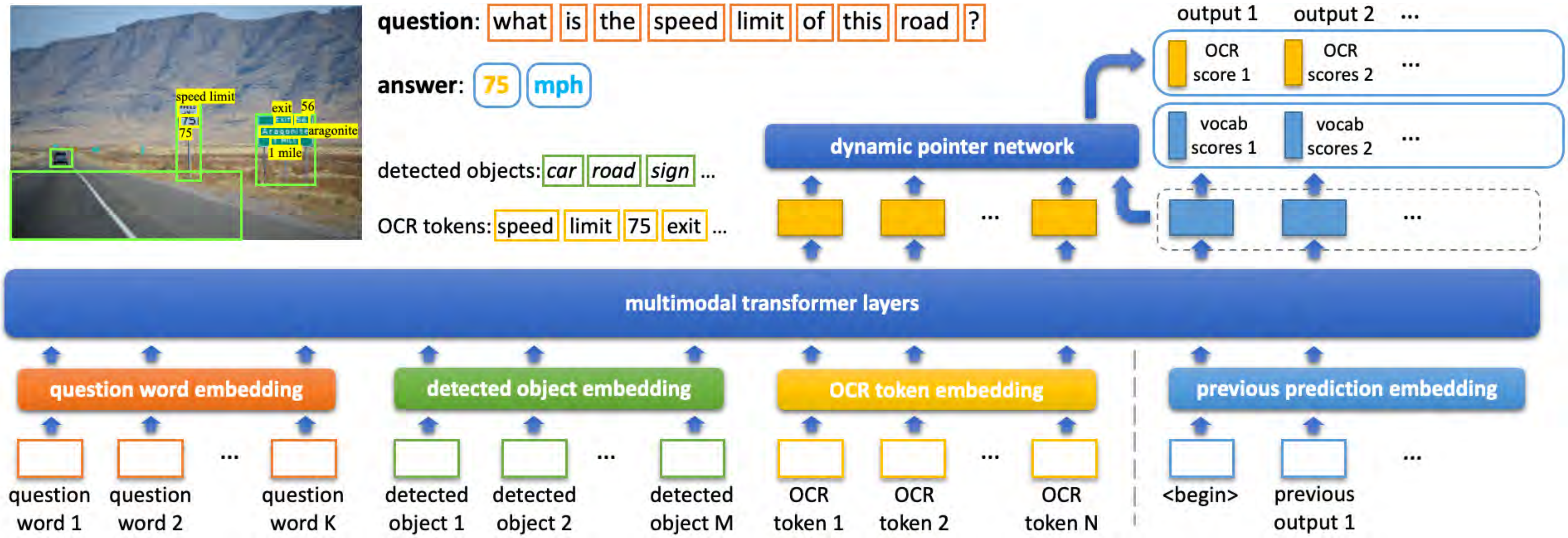
The M4C model – iterative answer prediction

- Step 3: decode the answer word-by-word with dynamic pointers



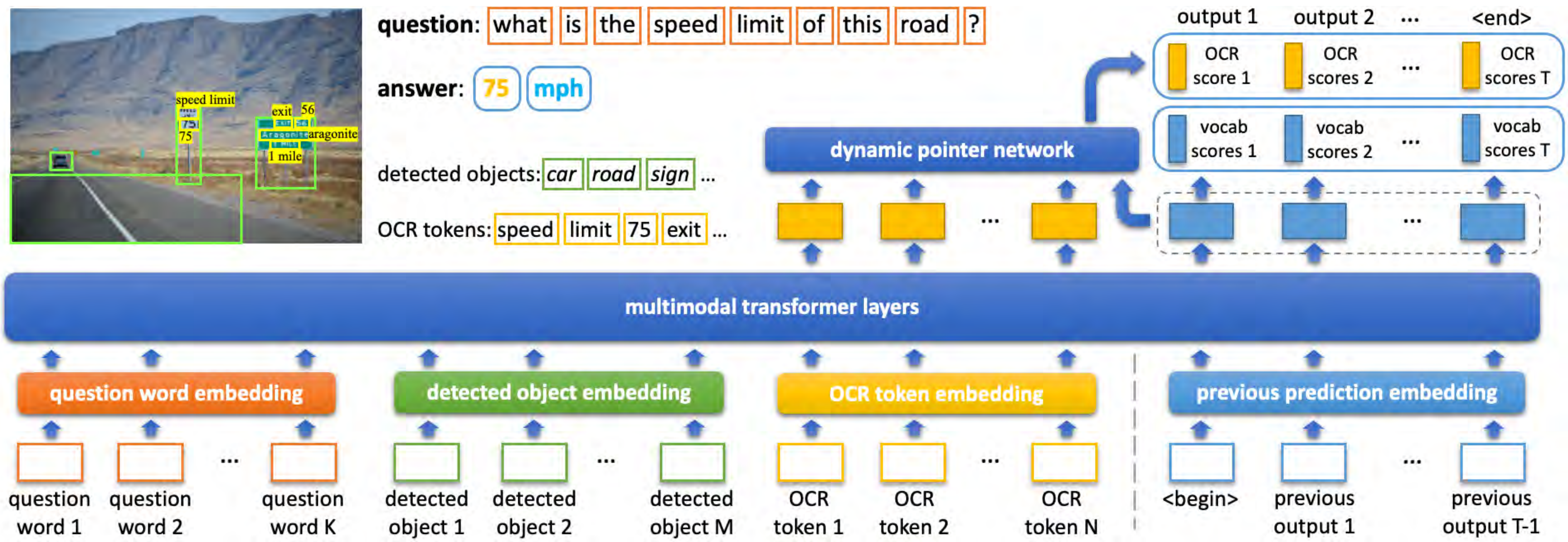
The M4C model – iterative answer prediction

- Step 3: decode the answer word-by-word with dynamic pointers

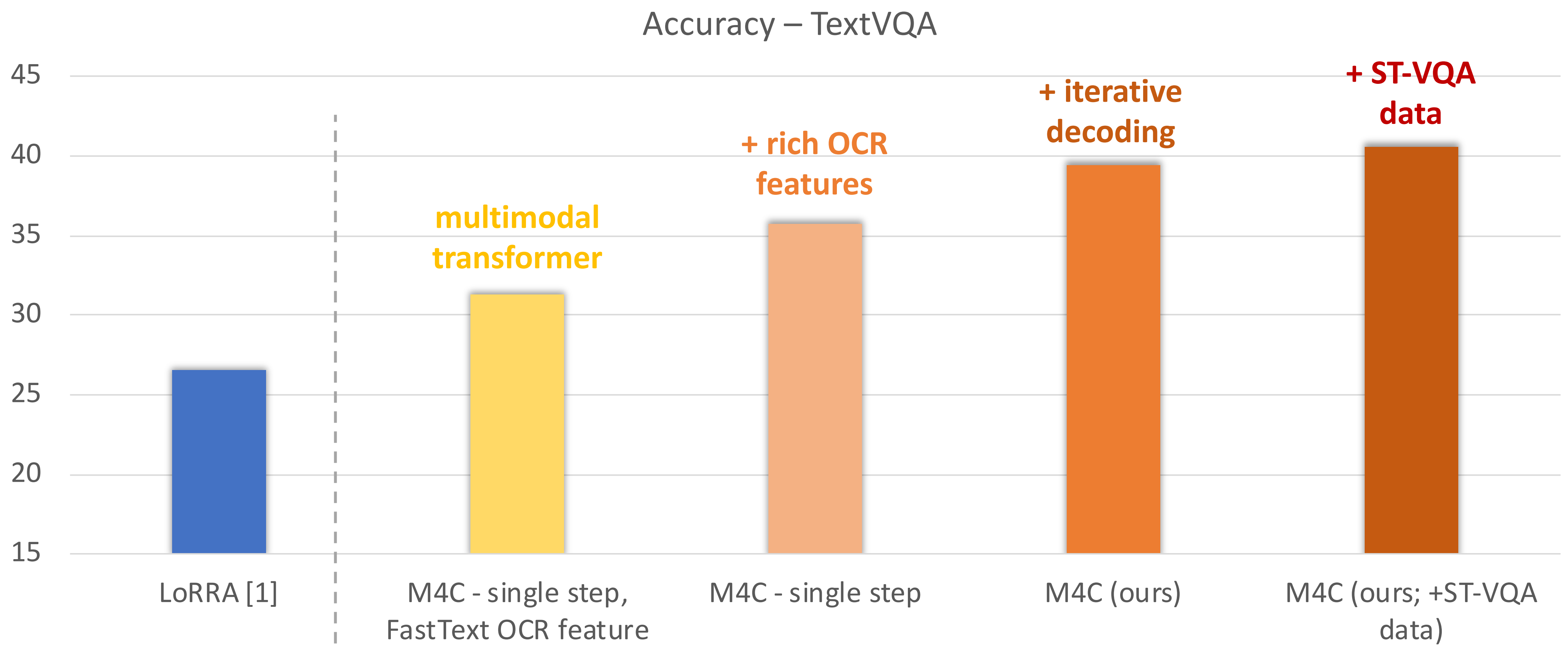


The M4C model – iterative answer prediction

- Step 3: decode the answer word-by-word with dynamic pointers

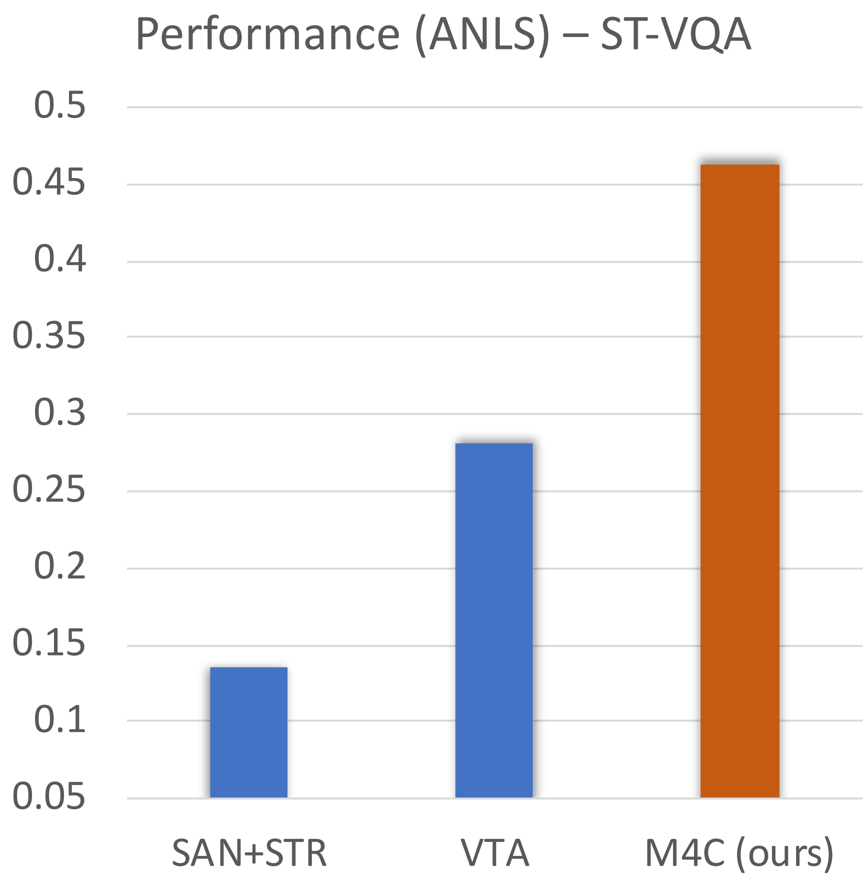


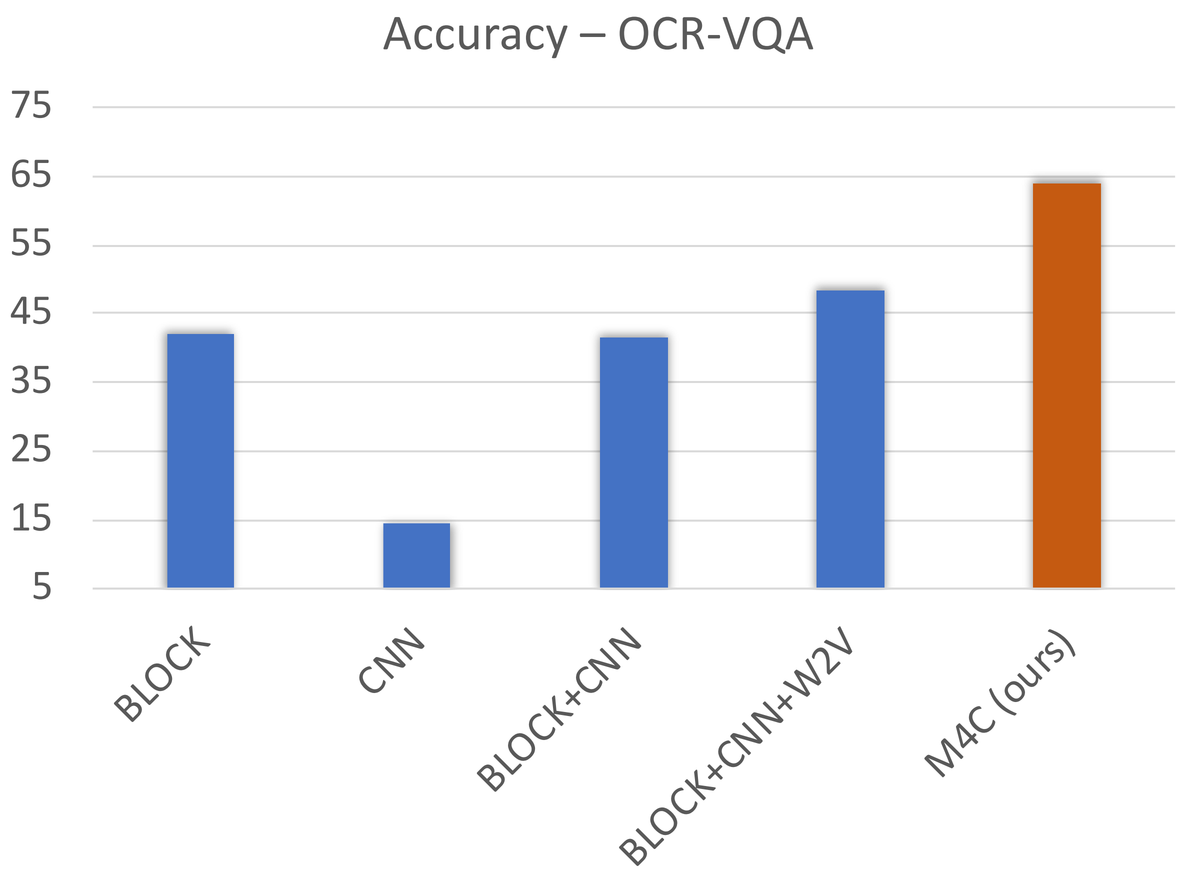
Quantitative results – the TextVQA dataset



[1] Singh, Amanpreet, et al. "Towards VQA Models That Can Read." in CVPR 2019.

Quantitative results – ST-VQA & OCR-VQA

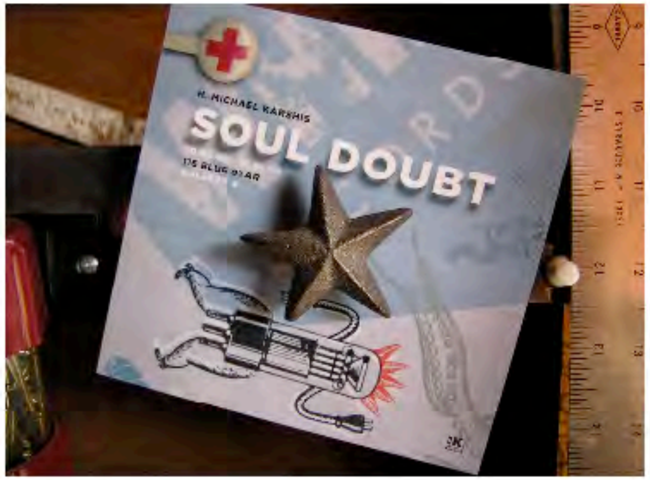




Qualitative results – the TextVQA dataset

What does the light sign read on the farthest right window?
prediction: bud light
GT: **bud light**

Who is usa today's bestselling author?
prediction: cathy williams
GT: **cathy williams**



What is the name of the band?
prediction: soul doubt
GT: **soul doubt**

blue: words from fixed vocabulary; orange: words from OCR results

Qualitative results – the ST-VQA dataset

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**

blue: words from fixed vocabulary; orange: words from OCR results

Follow-up work – Applying M4C to the image captioning task

- image captioning based on reading comprehension (TextCaps)
- our adapted M4C-Captioner model outperforms strong baselines

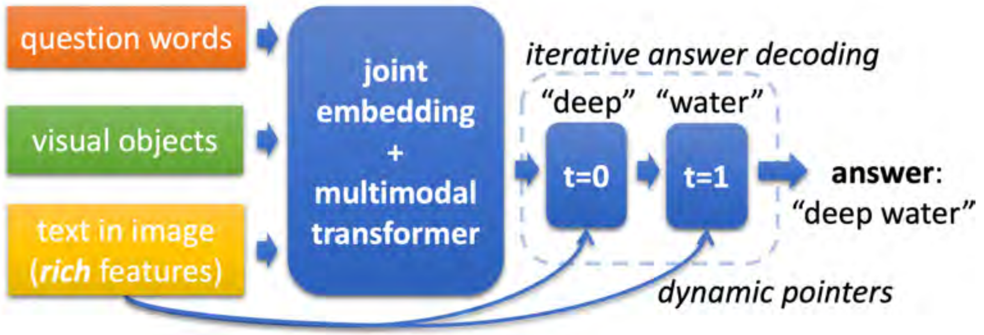
check the paper [1] for details!

M4C-Captioner (ours): The front and back of an LG phone that is on October 19

[1] Sidorov, Oleksii, et al. "TextCaps: a Dataset for Image Captioning with Reading Comprehension." arXiv:2003.12462 https://arxiv.org/abs/2003.12462

blue: words from fixed vocabulary; orange: words from OCR results

Summary of M4C



- fuse 3 modalities with **joint embedding** and **multimodal transformer**
- represent text (OCR) tokens by **a rich representation** with 4 features
- **decode the answer iteratively** beyond one-step classification

Code and models: https://github.com/facebookresearch/mmf/tree/master/projects/m4c

SQuINTing at VQA Models: Introspecting VQA Models with Sub-Questions


Ramprasaath R. Selvaraju


Purva Tendulkar


Devi Parikh

Georgia Institute of Technology


Eric Horvitz


Marco Ribeiro

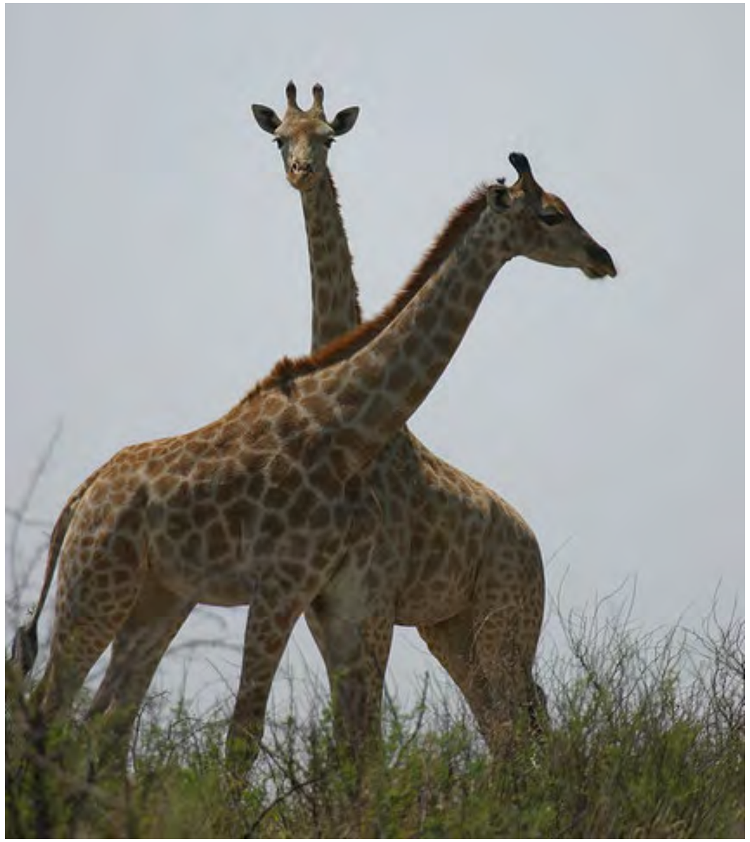

Besmira Nushi


Ece Kamar

Microsoft Research

Motivation

- VQA models face consistency issues



Main Reasoning Question

Are these giraffes in captivity?

No

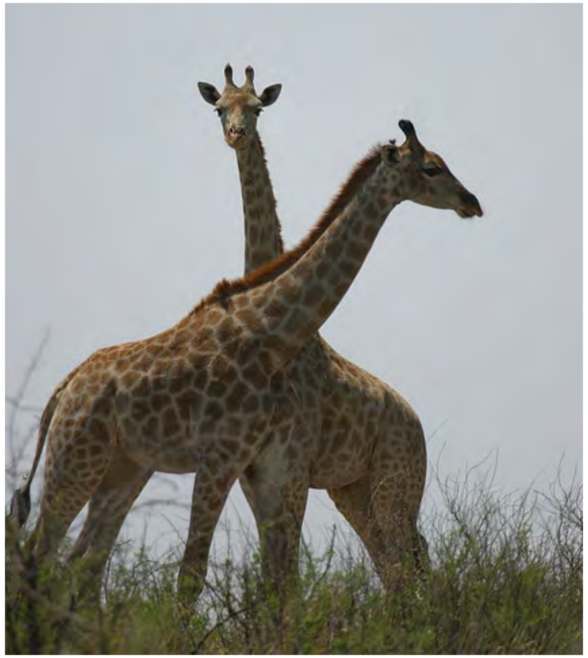
Perception Sub-Question

Is there a fence around the giraffes?

Yes

Motivation

- Questions with varying levels of complexities are considered equivalent during evaluation and learning



Are the giraffes in captivity?

Reasoning

What color is the man’s shirt?

Perception

Most questions only require perceptual skills

- **Physical properties of objects/entities:**
  - What color is the couch?
- **Existence:**
  - Is there a fork?
- **Simple activities:**
  - Is the man sitting?
- **Spatial relationship:**
  - What is to the right of the bed?
- **Text/Symbol recognition:**
  - What does the sign say?

82% of the VQA2.0 dataset

* Identified through simple regex rules

Few require reasoning over perceptual concepts

How is this train powered?

Electricity

Is this a good idea for a rainy day?

No

What perceptual concepts are needed for answering the reasoning question?

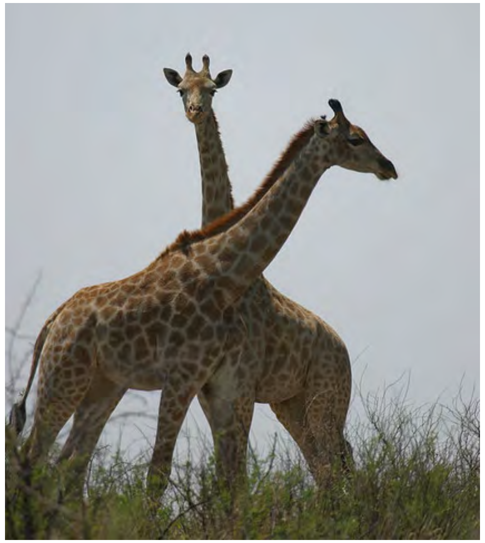
VQA-Introspect Dataset

Main Reasoning Question:

- Is this a keepsake photo? “Yes”

Perception Sub-questions:

- Is this a black and white photo? “Yes”
- Is the woman wearing a white veil and holding flowers? “Yes”
- Is the woman wearing a veil? “Yes”
- What is the woman next to the man wearing? “Gown”



Main Reasoning Question:

- Are these giraffes in captivity? “No”

Perception Sub-questions:

- Is there a fence? “No”
- is there anything man-made near the giraffes? “No”
- is there a fence around the giraffes? “No”

200K sub-questions for 86K reasoning questions

Are VQA models consistent in their reasoning process?

Evaluating Pythia

Overall : 60.26%

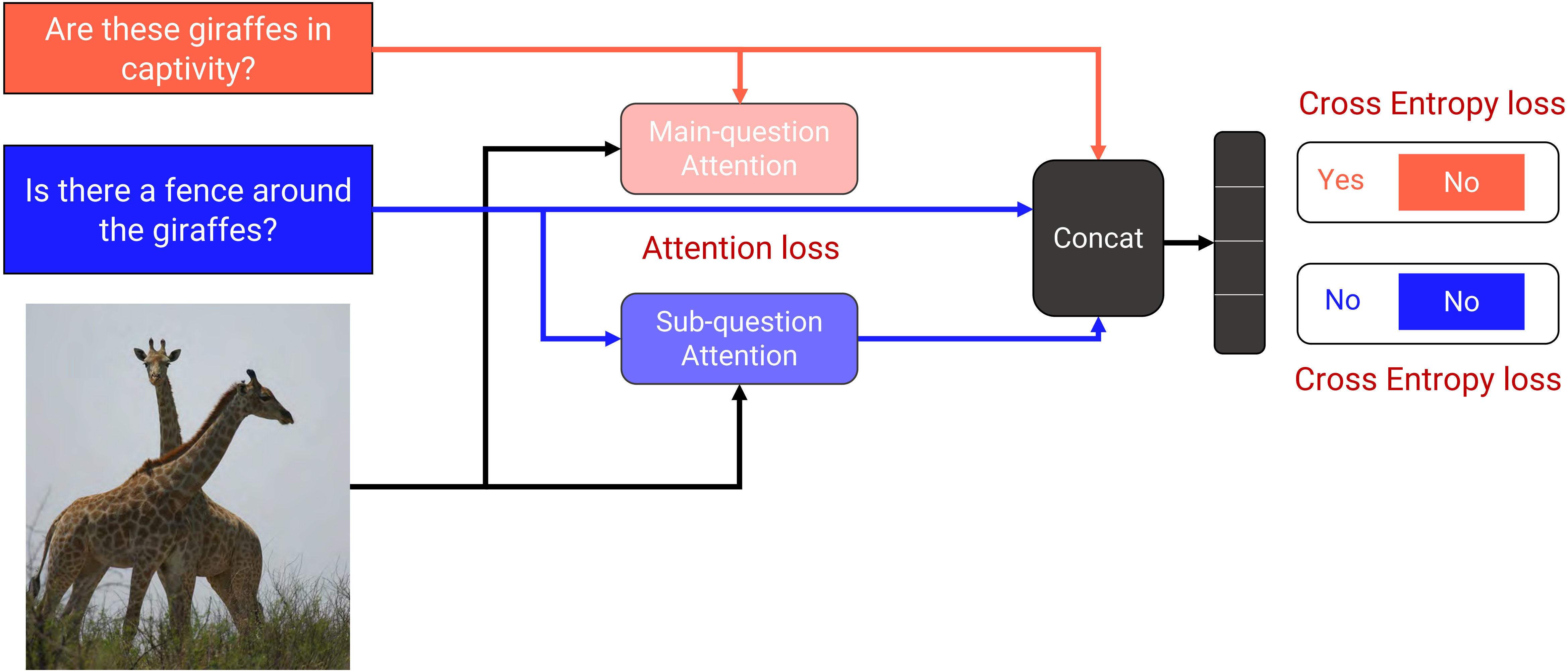
Both main & sub-question correct M✓ S✓ 47.72%	Main correct & sub-question incorrect M✓ S✗ 18.48%
Main incorrect & sub-question correct M✗ S✓ 20.54%	Both main & sub-question incorrect M✗ S✗ 13.26%

$$\text{Consistency} = \frac{\text{M✓ S✓}}{\text{M✓ S✓} + \text{M✓ S✗}}$$

Model is consistent only ~72% of the times

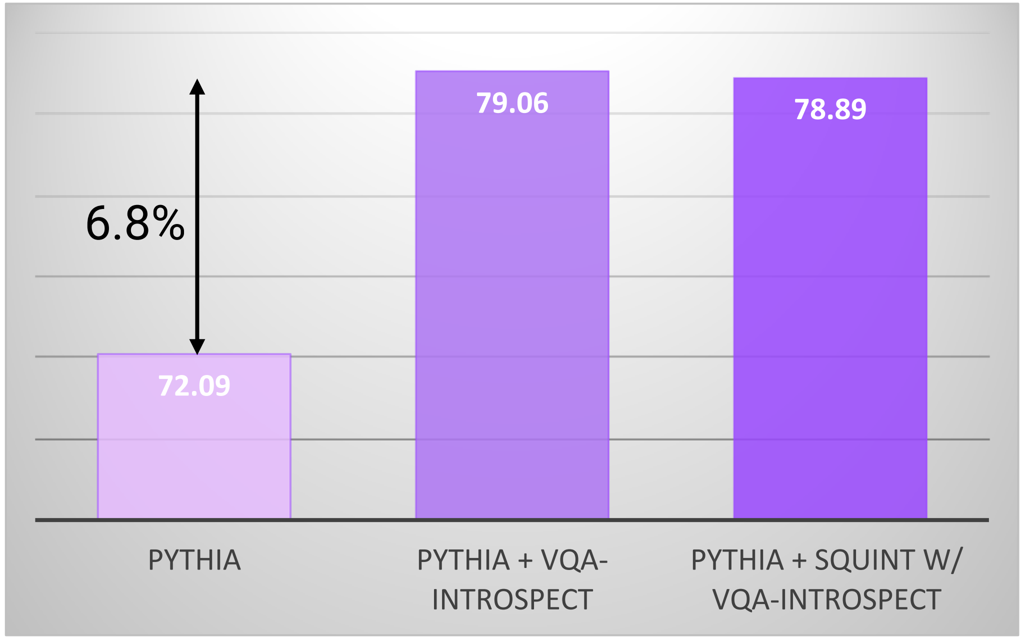
* Results are for Pythia trained on VQAv1 evaluated on VQA-Introspect v0.7

Sub-Question Importance-aware Network Tuning (SQuINT)

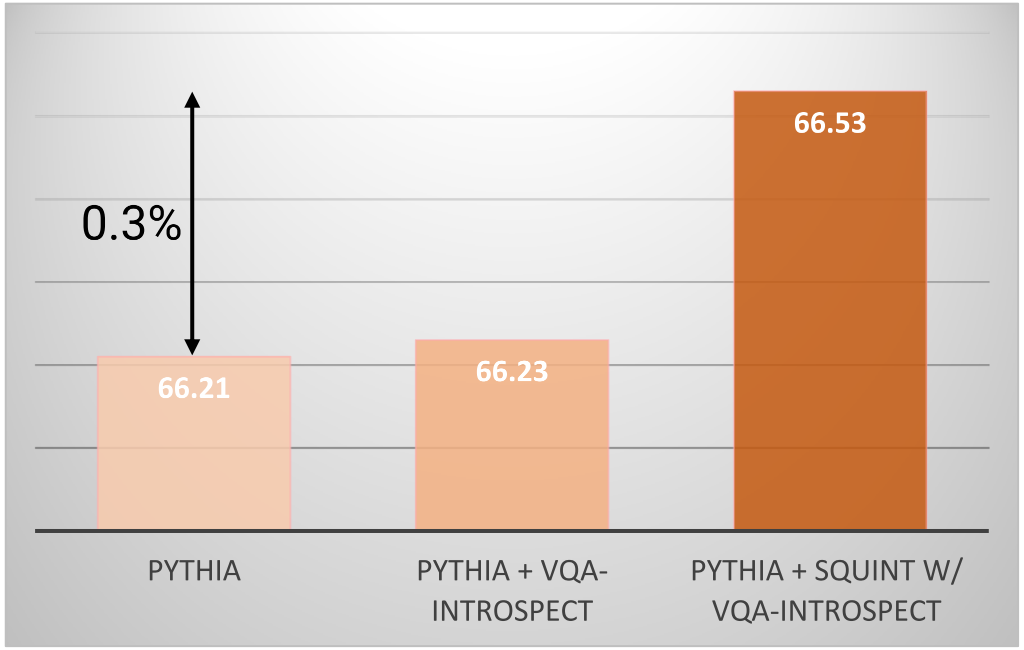


Results

Consistency

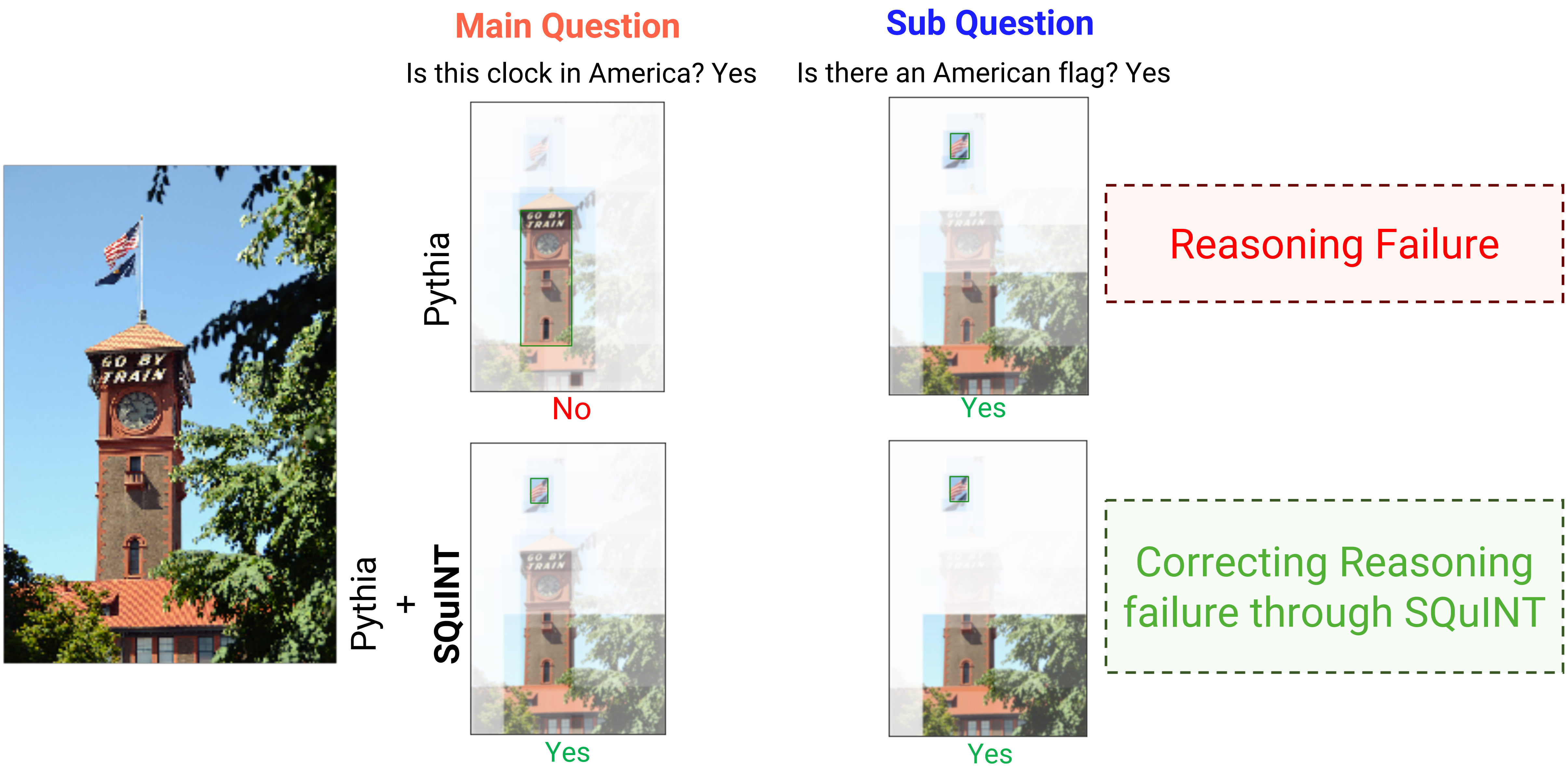


Reasoning Accuracy



* Results are for Pythia trained on VQAv1 evaluated on VQA-Introspect v0.7

Qualitative results



Take-aways

Dataset

Paper

- New split of the VQA dataset – Perception vs Reasoning
- Releasing a new dataset, **VQA-Introspect** containing ~200K sub-questions for 86K reasoning questions
- We find that even top-performing models are inconsistent ~28% times
- Introduce SQuINT
    - Encouraging models to look at sub-question regions when answering main question improves consistency

https://aka.ms/vqa-introspect

ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS

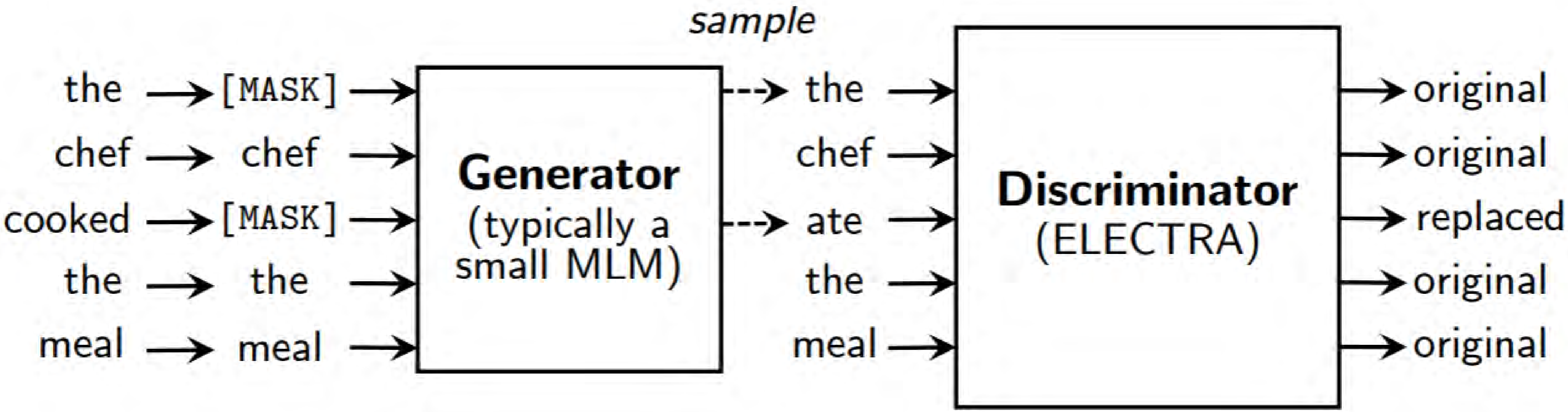
Kevin Clark
Stanford University
kevclark@cs.stanford.edu

Minh-Thang Luong
Google Brain
thangluong@google.com

Quoc V. Le
Google Brain
qvl@google.com

Christopher D. Manning
Stanford University & CIFAR Fellow
manning@cs.stanford.edu

ICLR 2020



$$\min_{\theta_G,\theta_D} \sum_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}_{\text{MLM}}(\boldsymbol{x},\theta_G) + \lambda\mathcal{L}_{\text{Disc}}(\boldsymbol{x},\theta_D)$$

1. Train only the generator with $\mathcal{L}_{\text{MLM}}$ for n steps.
2. Initialize the weights of the discriminator with the weights of the generator. Then train the discriminator with $\mathcal{L}_{\text{Disc}}$ for n steps, keeping the generator's weights frozen.

Model	Train / Infer FLOPs	Speedup	Params	Train Time + Hardware	GLUE
ELMo	3.3e18 / 2.6e10	19x / 1.2x	96M	14d on 3 GTX 1080 GPUs	71.2
GPT	4.0e19 / 3.0e10	1.6x / 0.97x	117M	25d on 8 P6000 GPUs	78.8
BERT-Small	1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	75.1
BERT-Base	6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	82.2
ELECTRA-Small	1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	79.9
50% trained	7.1e17 / 3.7e9	90x / 8x	14M	2d on 1 V100 GPU	79.0
25% trained	3.6e17 / 3.7e9	181x / 8x	14M	1d on 1 V100 GPU	77.7
12.5% trained	1.8e17 / 3.7e9	361x / 8x	14M	12h on 1 V100 GPU	76.0
6.25% trained	8.9e16 / 3.7e9	722x / 8x	14M	6h on 1 V100 GPU	74.1
ELECTRA-Base	6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	85.1

Table 1: Comparison of small models on the GLUE dev set. BERT-Small/Base are our implementation and use the same hyperparameters as ELECTRA-Small/Base. Infer FLOPs assumes single length-128 input. Training times should be taken with a grain of salt as they are for different hardware and with sometimes un-optimized code. ELECTRA performs well even when trained on a single GPU, scoring 5 GLUE points higher than a comparable BERT model and even outscoring the much larger GPT model.

Model	Train FLOPs	Params	CoLA	SST	MRPC	STS	QQP	MNLI	QNLI	RTE	Avg.
BERT	1.9e20 (0.27x)	335M	60.6	93.2	88.0	90.0	91.3	86.6	92.3	70.4	84.0
RoBERTa-100K	6.4e20 (0.90x)	356M	66.1	95.6	**91.4**	92.2	92.0	89.3	94.0	82.7	87.9
RoBERTa-500K	3.2e21 (4.5x)	356M	68.0	96.4	90.9	92.1	92.2	90.2	94.7	86.6	88.9
XLNet	3.9e21 (5.4x)	360M	69.0	**97.0**	90.8	92.2	92.3	90.8	94.9	85.9	89.1
BERT (ours)	7.1e20 (1x)	335M	67.0	95.9	89.1	91.2	91.5	89.6	93.5	79.5	87.2
ELECTRA-400K	7.1e20 (1x)	335M	**69.3**	96.0	90.6	92.1	**92.4**	90.5	94.5	86.8	89.0
ELECTRA-1.75M	3.1e21 (4.4x)	335M	69.1	96.9	90.8	**92.6**	**92.4**	**90.9**	**95.0**	**88.0**	**89.5**

Model	Train FLOPs	Params	SQuAD 1.1 dev		SQuAD 2.0 dev		SQuAD 2.0 test	
			EM	F1	EM	F1	EM	F1
BERT-Base	6.4e19 (0.09x)	110M	80.8	88.5	–	–	–	–
BERT	1.9e20 (0.27x)	335M	84.1	90.9	79.0	81.8	80.0	83.0
SpanBERT	7.1e20 (1x)	335M	88.8	94.6	85.7	88.7	85.7	88.7
XLNet-Base	6.6e19 (0.09x)	117M	81.3	–	78.5	–	–	–
XLNet	3.9e21 (5.4x)	360M	**89.7**	**95.1**	87.9	**90.6**	87.9	90.7
RoBERTa-100K	6.4e20 (0.90x)	356M	–	94.0	–	87.7	–	–
RoBERTa-500K	3.2e21 (4.5x)	356M	88.9	94.6	86.5	89.4	86.8	89.8
ALBERT	3.1e22 (44x)	235M	89.3	94.8	87.4	90.2	88.1	90.9
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	–	–
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	–	–
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	–	–
ELECTRA-1.75M	3.1e21 (4.4x)	335M	**89.7**	94.9	**88.0**	**90.6**	**88.7**	**91.4**

Table 4: Results on the SQuAD for non-ensemble models.