ECCV'20 ONLINE
23-28 AUGUST 2020

Spatially Aware Multimodal Transformers for TextVQA

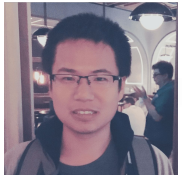

Yash Kant


Dhruv Batra


Peter Anderson


Alex Schwing


Devi Parikh


Jiasen Lu

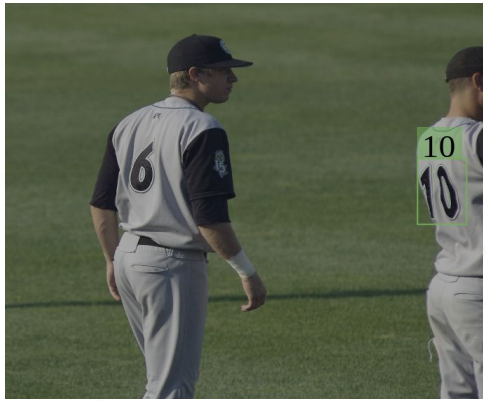

Harsh Agrawal

ECCV'20 ONLINE
23-28 AUGUST 2020

TextVQA

Question: **Who is the author of the book at the top of the stack?**

Answer: **Nate Bolt**



Question: **What is the number of the player on the right?**

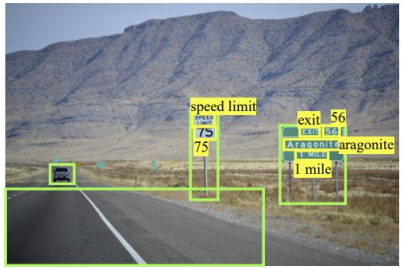
Answer: **10**

Question: **What brand of watch is this?**

Answer: **SEIKO**

[*Singh et al.* CVPR 2018]

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

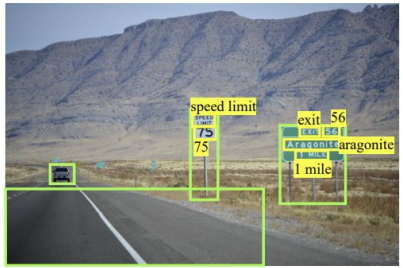
answer: 75 | mph

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

OCR tokens: speed | limit | 75 | exit | ...

Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

answer: 75 | mph

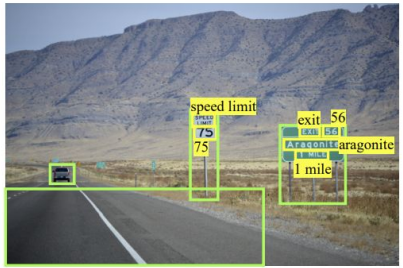
detected objects: *car* | *road* | *sign* | ...

OCR tokens: speed | limit | 75 | exit | ...

multimodal transformer layers

Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

answer: 75 | mph

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

OCR tokens: speed | limit | 75 | exit | ...

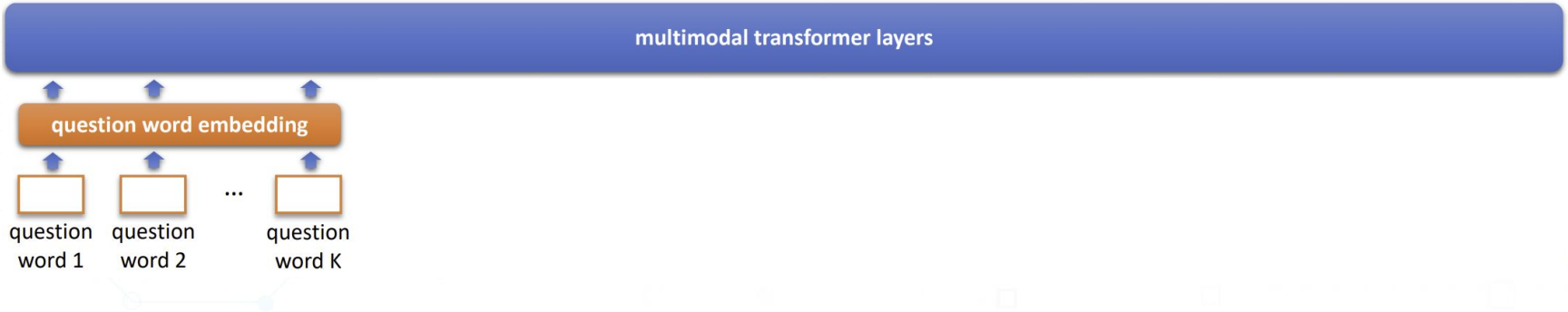
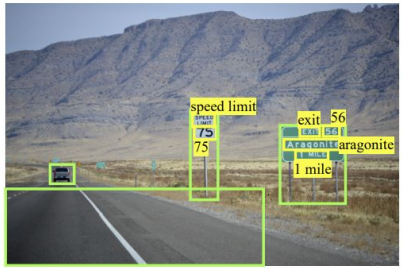


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

answer: 75 | mph

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

OCR tokens: speed | limit | 75 | exit | ...

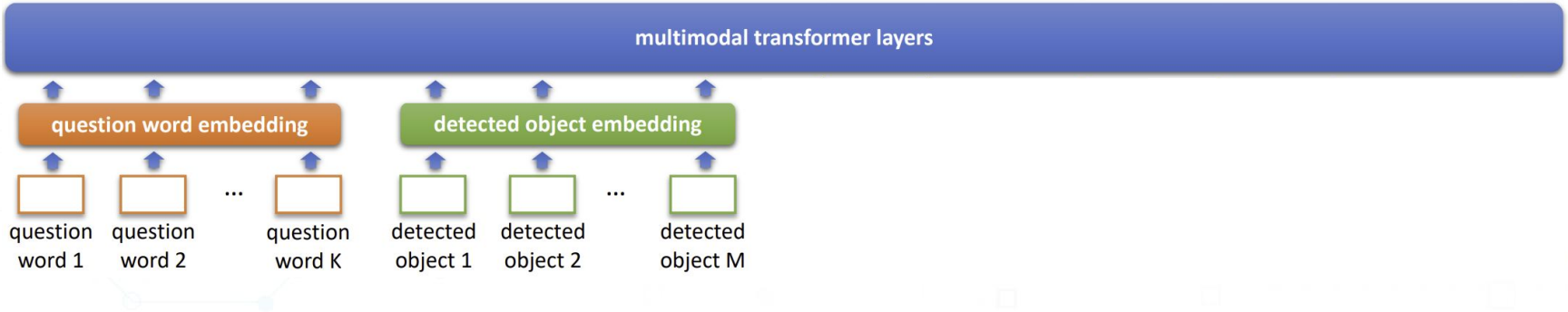
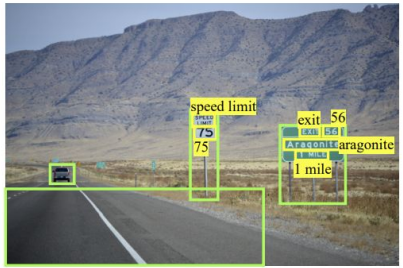


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

answer: 75 | mph

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

OCR tokens: speed | limit | 75 | exit | ...

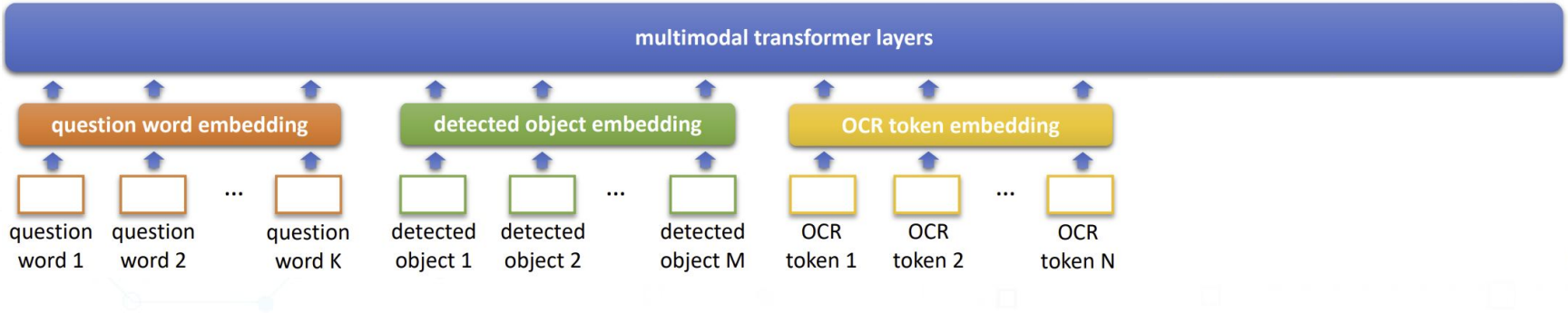
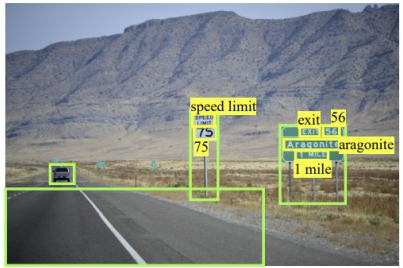


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

answer: 75 | mph

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

OCR tokens: speed | limit | 75 | exit | ...

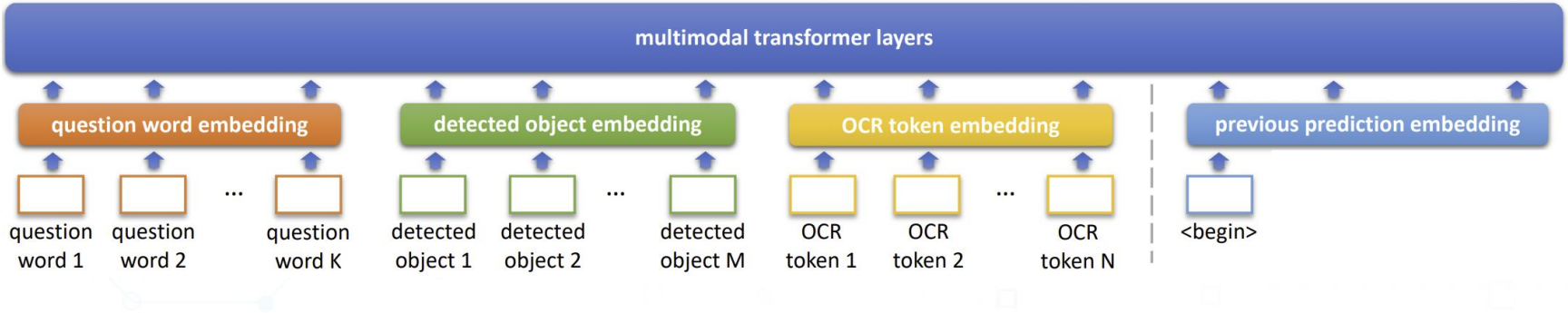
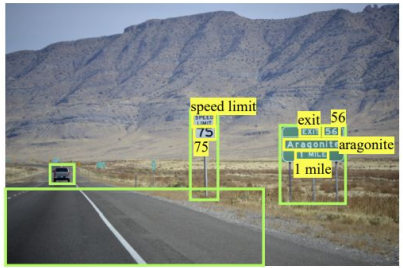


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]



question: what | is | the | speed | limit | of | this | road | ?

answer: 75 | mph

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

OCR tokens: speed | limit | 75 | exit | ...

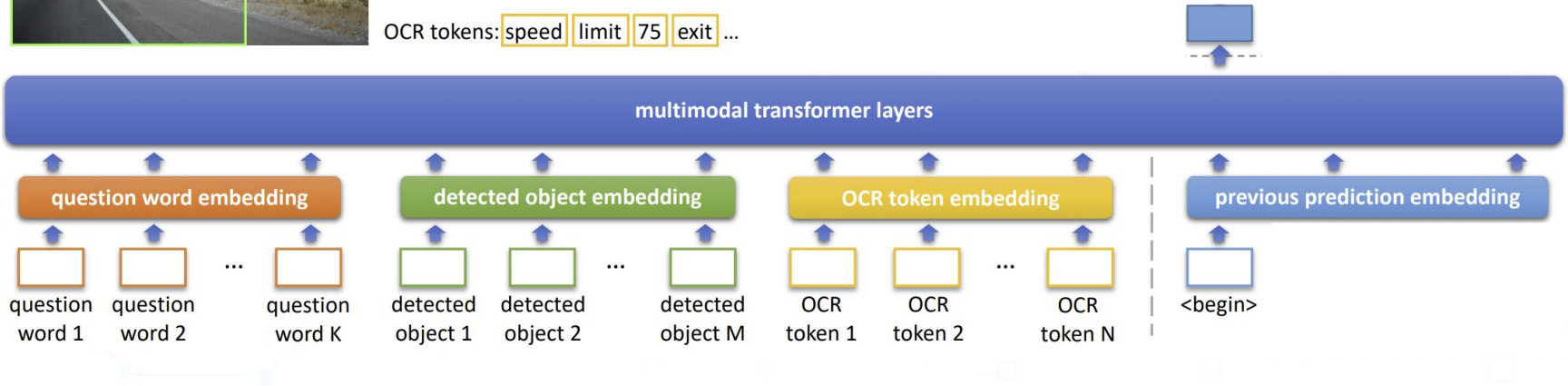


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]

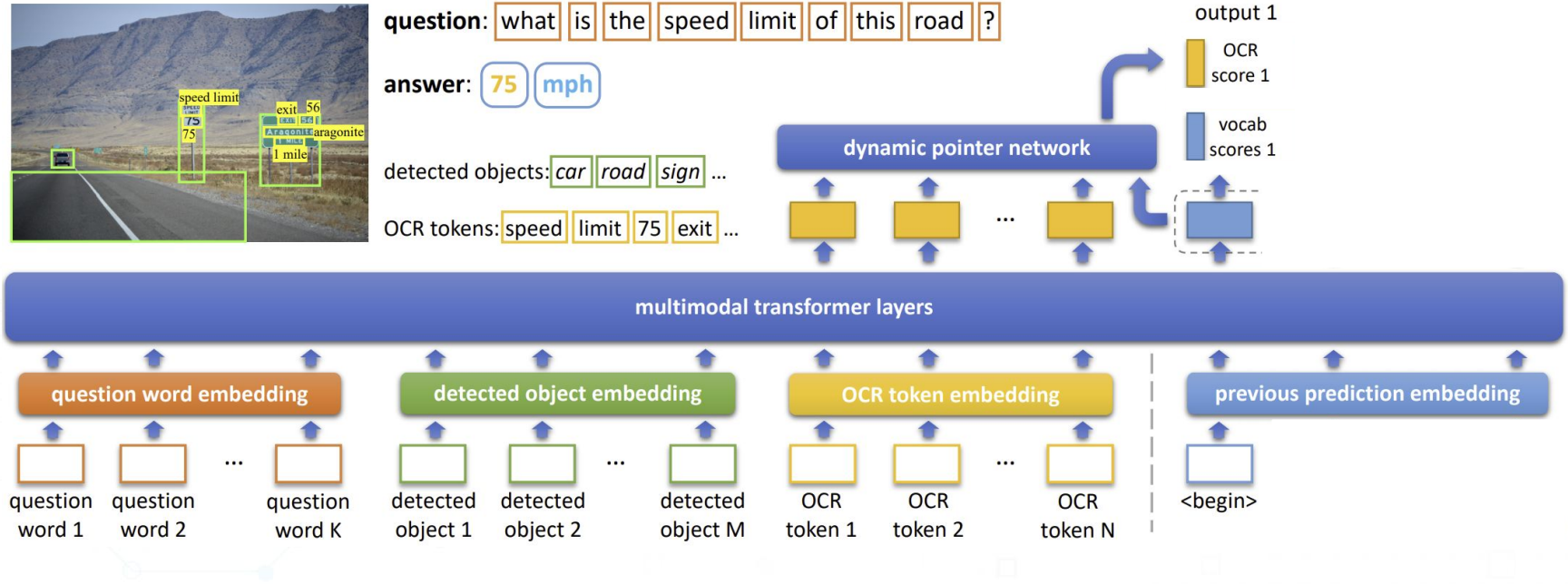


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]

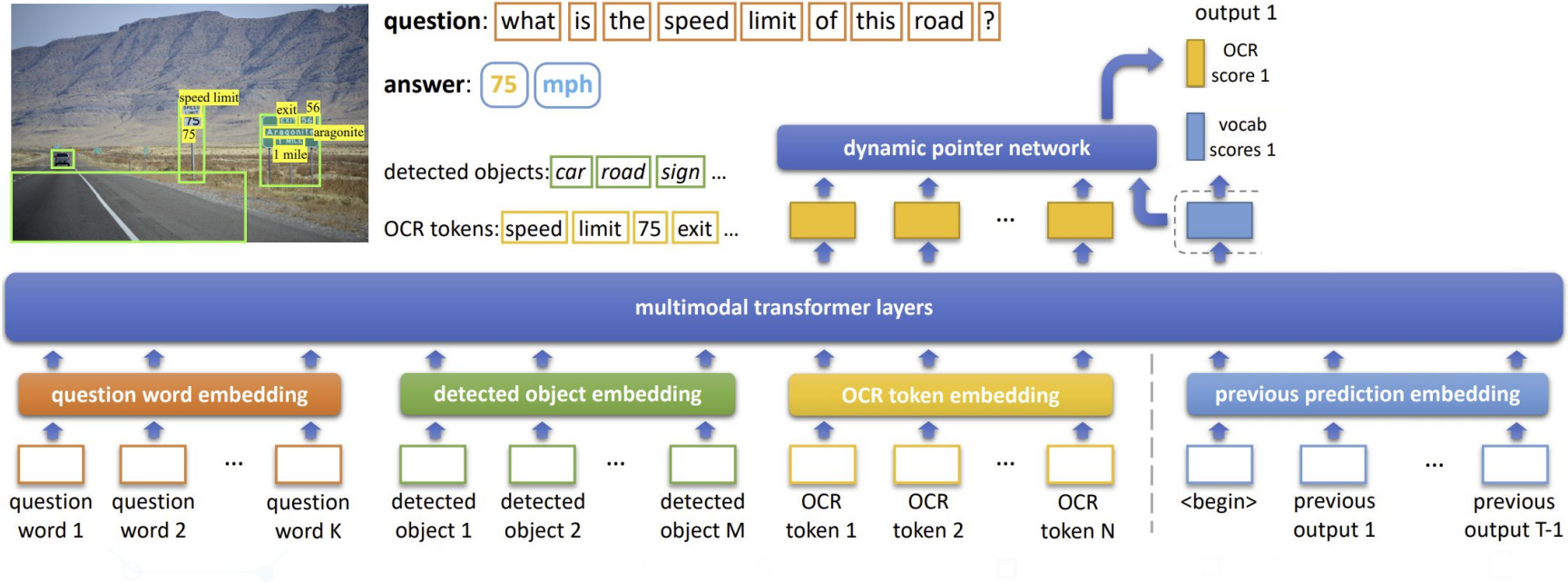


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

M4C Architecture [Hu et al. CVPR 2020]

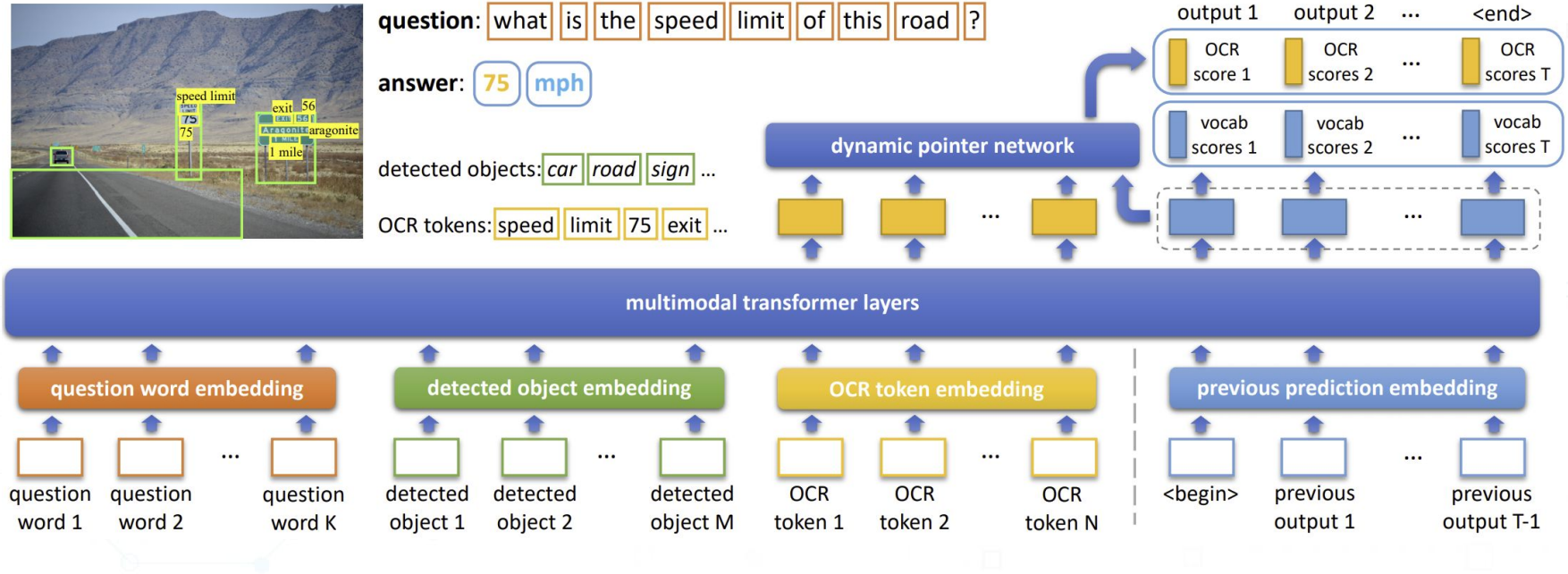
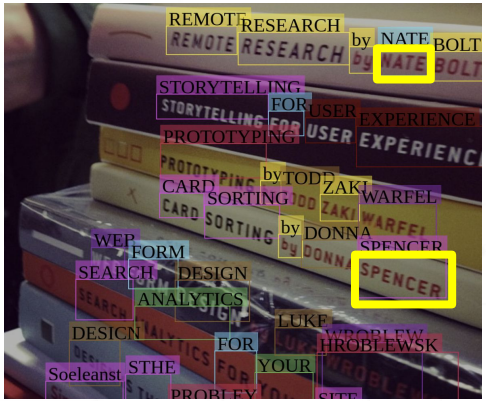


Image Credit: Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

Failure Modes of M4C

Question: **What number is the right one?**

M4C: **8953**
GT Answer: **8954**



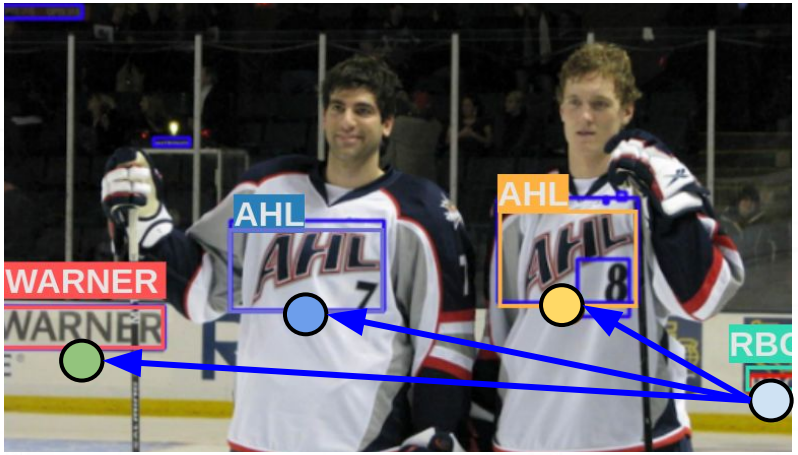
Question: **Who is the author of the book at the top of the stack?**

M4C: **Nate Spencer**
GT Answer: **Nate Bolt**

Question: **What is the top word on the sign on the left?**

M4C: **Burenwurst**
Answer: **Krainerwurst**

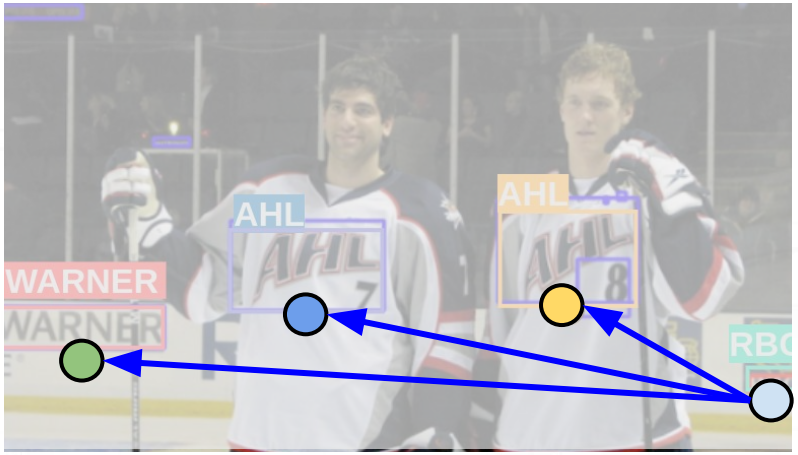
Spatially Aware Transformer Layer [Walkthrough Example]



Question: **What sponsor is to the right of the players?**

M4C: **AHL**
Answer: **RBC**

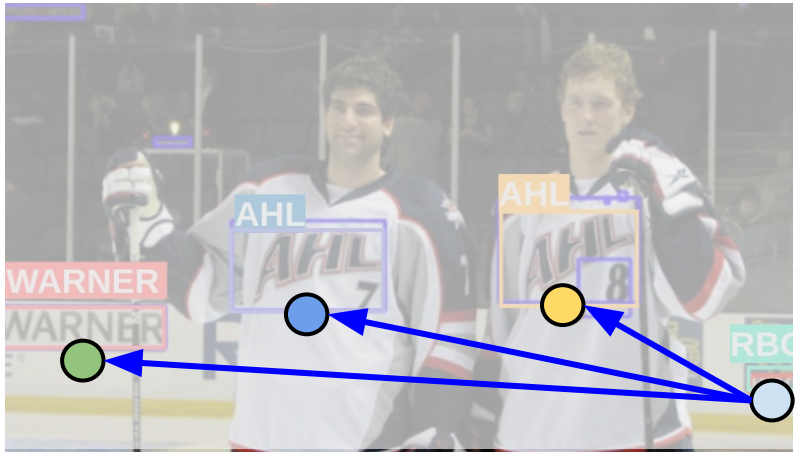
Spatially Aware Transformer Layer [Walkthrough Example]



Question: **What sponsor is to the right of the players?**

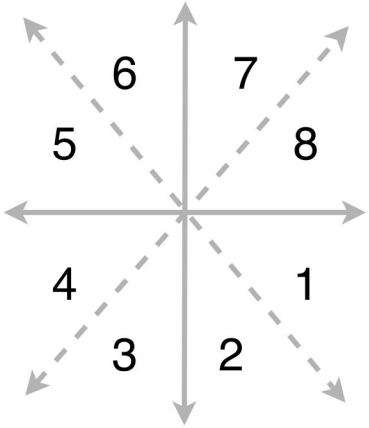
M4C: **AHL**
Answer: **RBC**

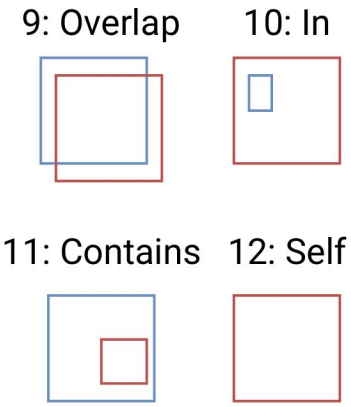
Spatially Aware Transformer Layer [Walkthrough Example]



Question: **What sponsor is to the right of the players?**

M4C: **AHL**
Answer: **RBC**

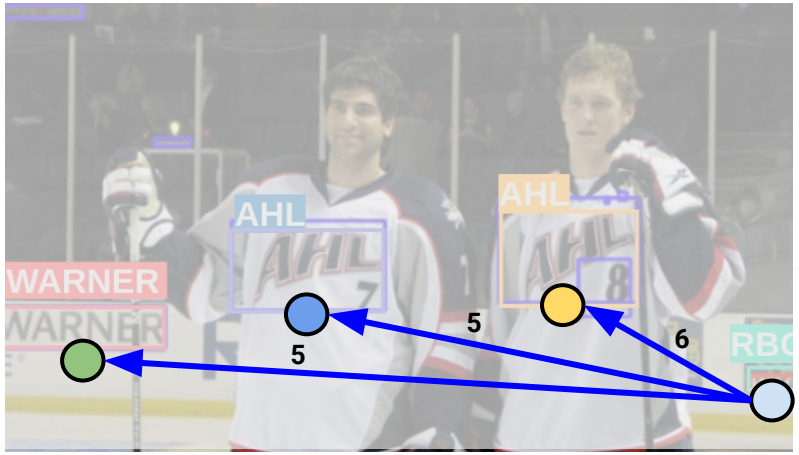




Spatial Relation Categories

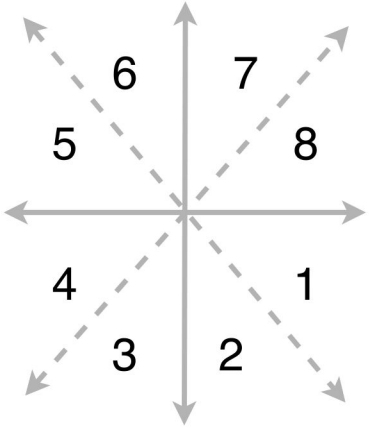
Reference: Exploring Visual Relationship for Image Captioning

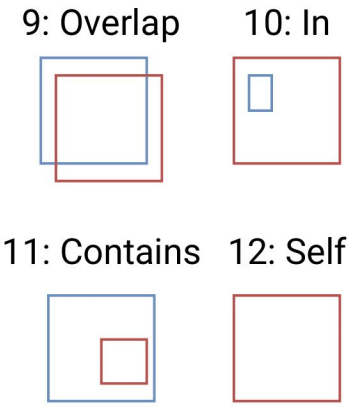
Spatially Aware Transformer Layer [Walkthrough Example]



Question: **What sponsor is to the right of the players?**

M4C: **AHL**
Answer: **RBC**

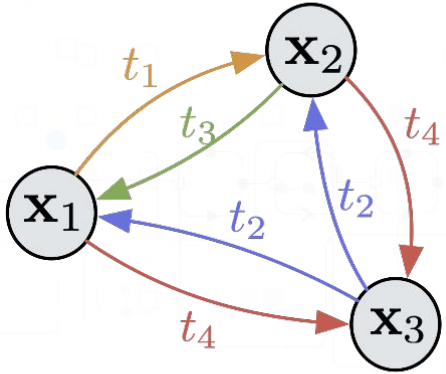




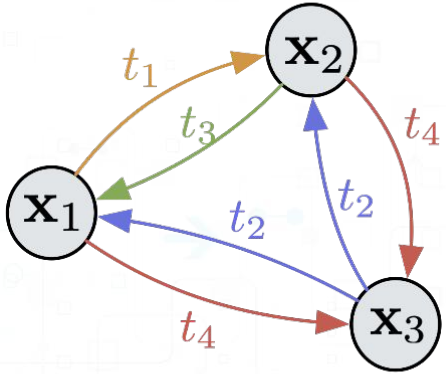
Spatial Relation Categories

Reference: Exploring Visual Relationship for Image Captioning

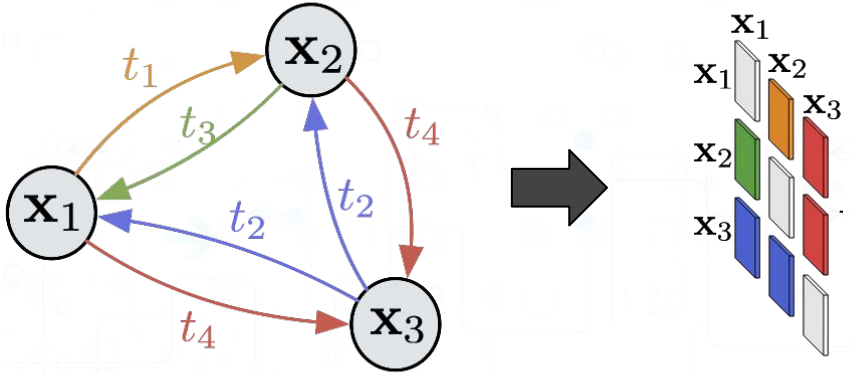
Spatially Aware Transformer Layer [Walkthrough Example]



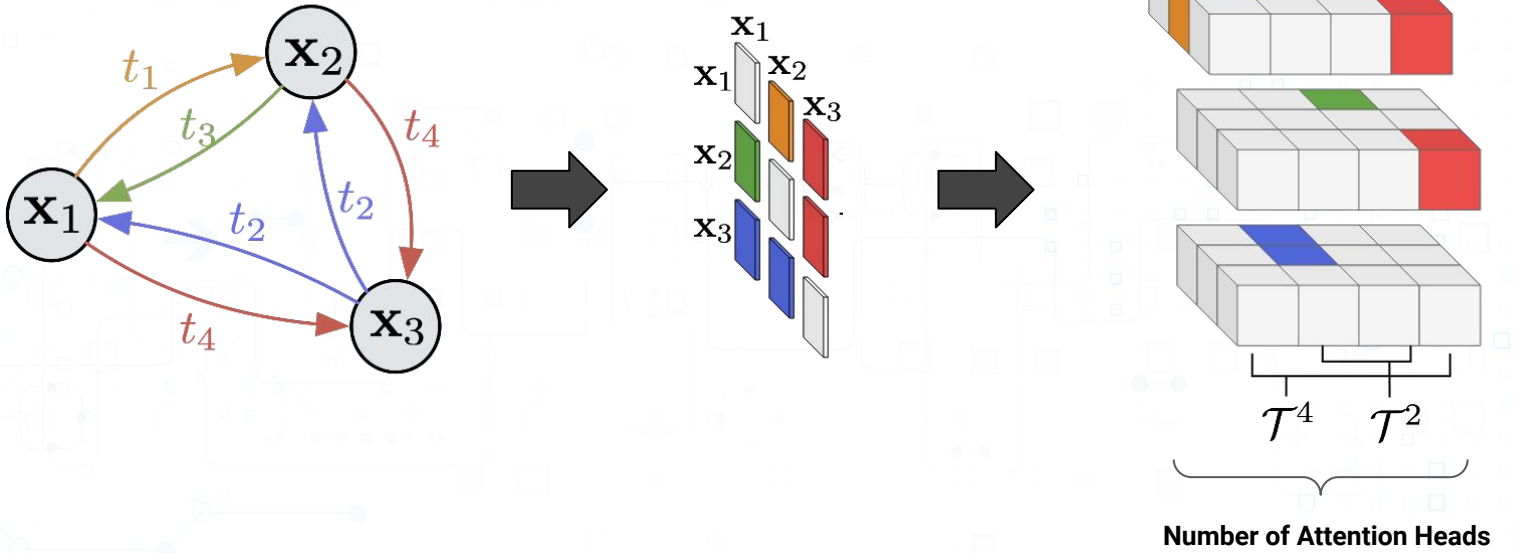
Spatially Aware Transformer Layer [Walkthrough Example]



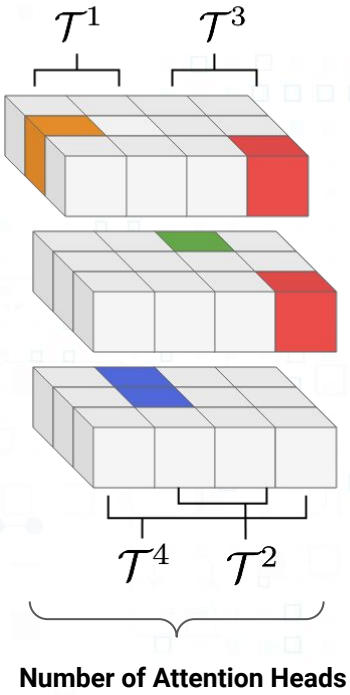
Spatially Aware Transformer Layer [Walkthrough Example]



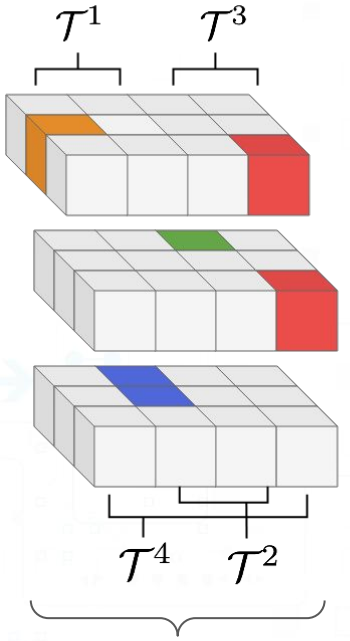
Spatially Aware Transformer Layer [Walkthrough Example]



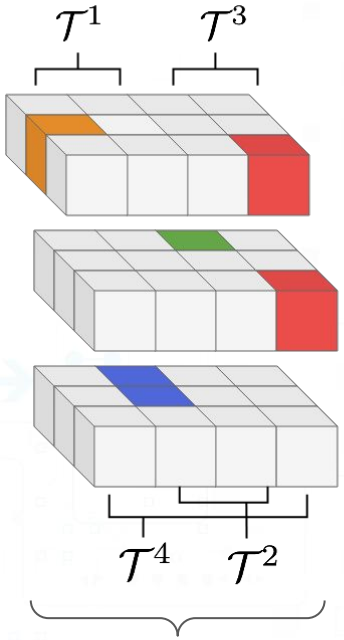
Spatially Aware Transformer Layer [Walkthrough Example]



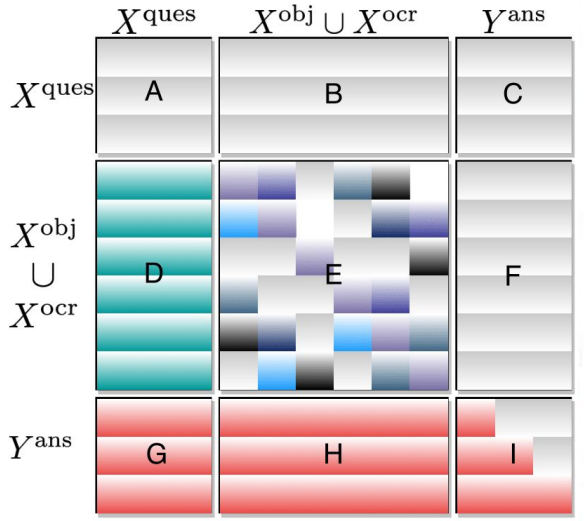
Spatially Aware Transformer Layer [Walkthrough Example]



Spatially Aware Transformer Layer [Walkthrough Example]

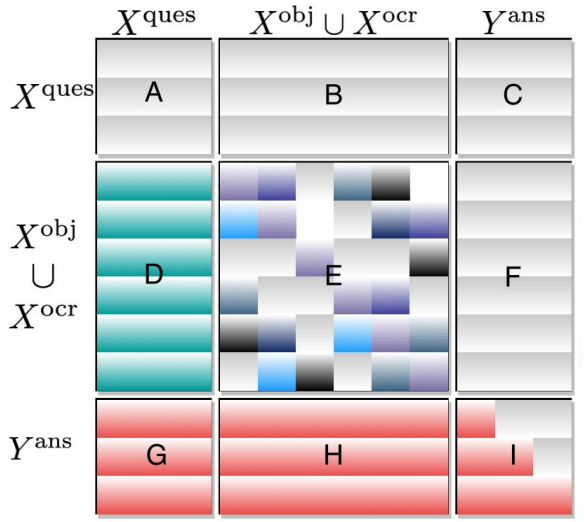


Number of Attention Heads



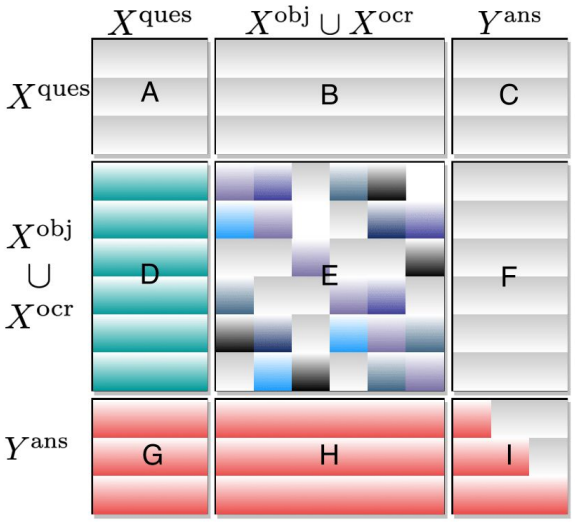
Sparse Self-Attention Matrix

Spatially Aware Transformer Layer [Walkthrough Example]



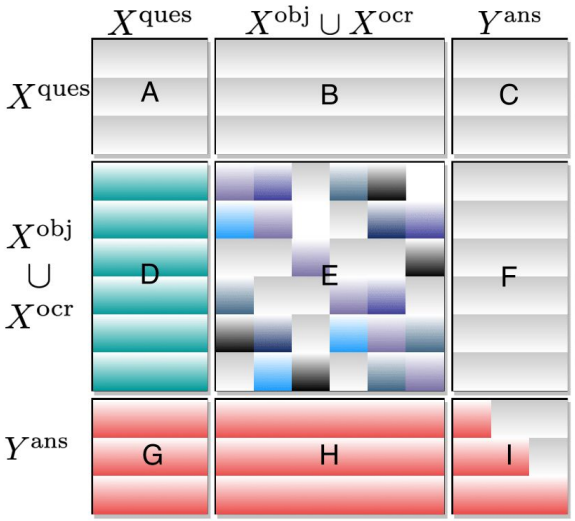
Sparse Self-Attention Matrix

Spatially Aware Transformer Layer [Walkthrough Example]



Sparse Self-Attention Matrix

Spatially Aware Transformer Layer [Walkthrough Example]



Sparse Self-Attention Matrix

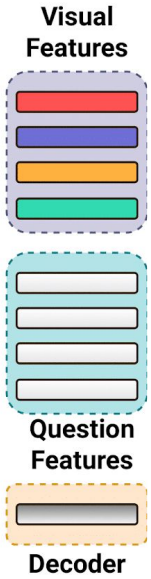
$$b_{i,j}^{h} = \begin{cases} 0 & e_{i \to j} \text{ of type } t_h, \quad \mathbf{x}_i, \mathbf{x}_j \in X \\ -\infty & \text{otherwise} \end{cases}$$

Bias Function

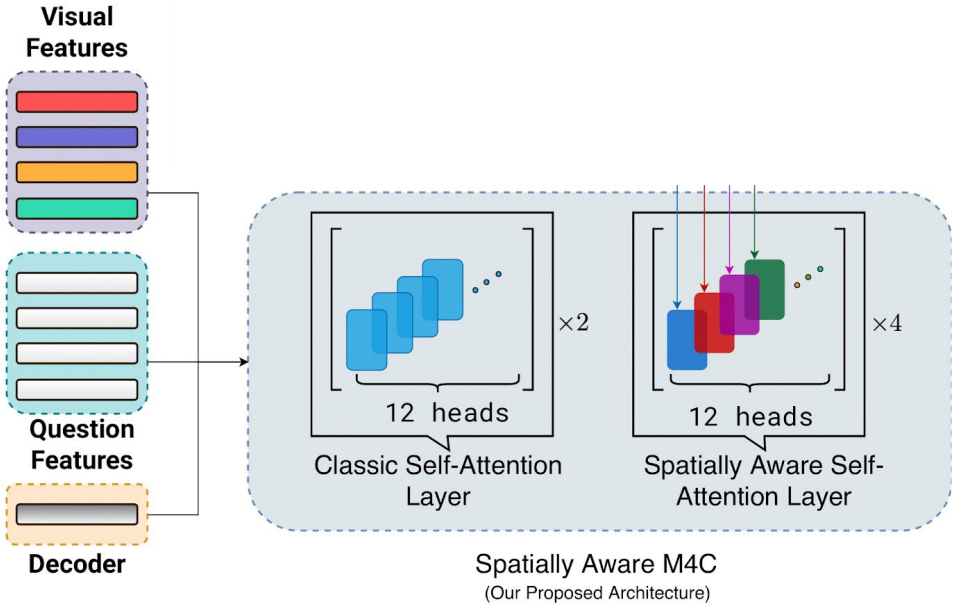
$$\alpha_{ij}^{h} = \text{Softmax}\left(\frac{\mathbf{q}_i^h (\mathbf{k}_j^h)^T + b_{i,j}^h}{\sqrt{d_h}}\right)$$

Modified Attention

Spatially Aware M4C (SA-M4C)

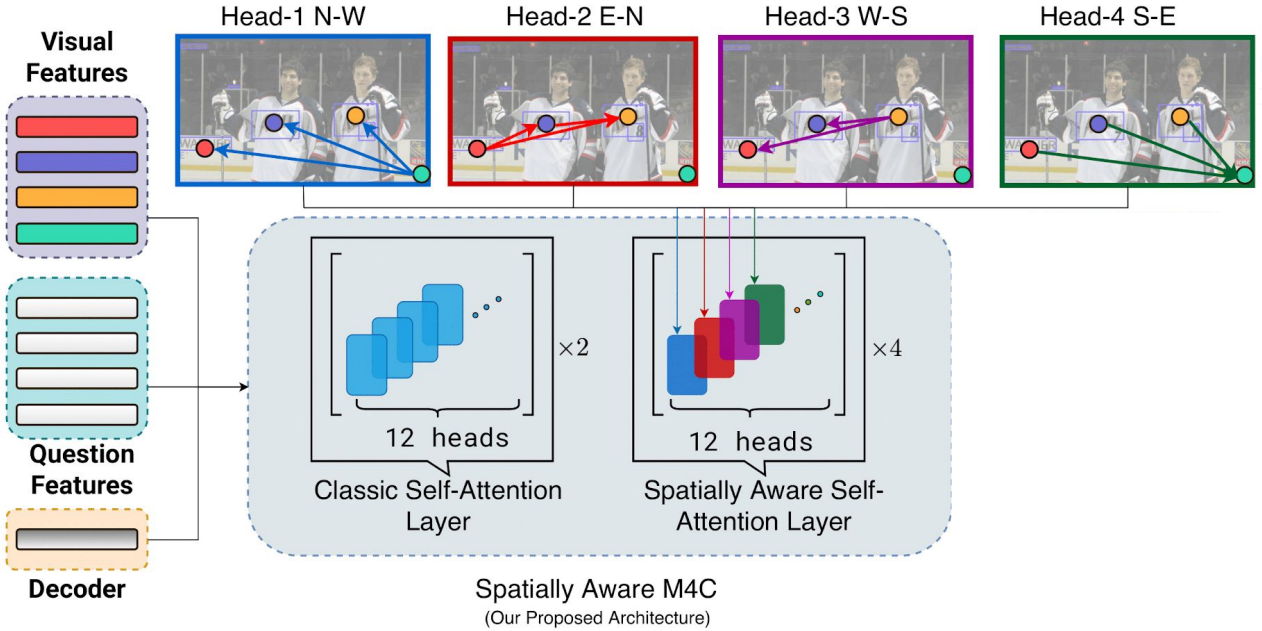


Spatially Aware M4C (SA-M4C)

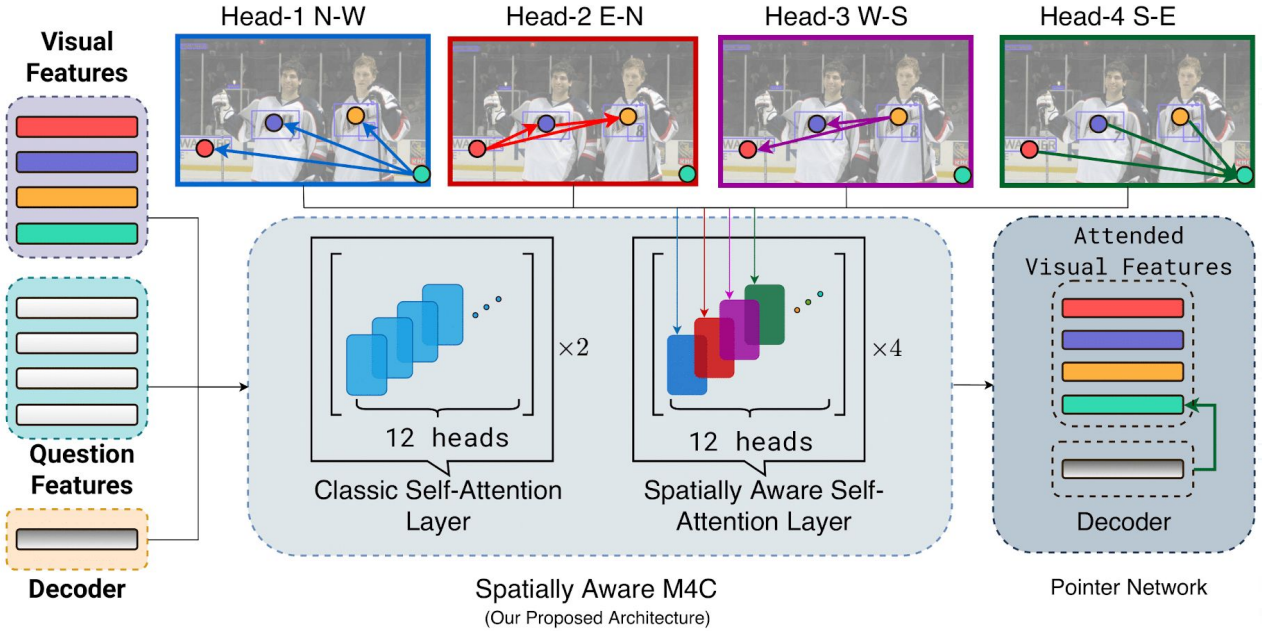


Spatially Aware M4C
(Our Proposed Architecture)

Spatially Aware M4C (SA-M4C)



Spatially Aware M4C (SA-M4C)



TextVQA

		Method	Structure	OCR system	DET backbone	Beam size	Accu. on val	Accu. on test
Previous Works	1	LoRRA [38]	-	R-ml	ResNet	-	26.5	27.6
	2	M4C [14]	4N	R-en	ResNet	1	40.5	40.4
Improved Baseline	3	M4C [14]†	6N	G	ResNeXt	1	43.3	-
	4	M4C [14]††	6N	G	ResNeXt	5	43.8	42.4
SA-M4C	5	SA-M4C (ours)	2N→4S	G	ResNeXt	1	45.1	-
	6	SA-M4C (ours)	2N→4S	G	ResNeXt	5	**45.4**	**44.6**

† Indicates our ablations for improved baseline.
†† Indicates the best model from improved baseline.

TextVQA

		Method	Structure	OCR system	DET backbone	Beam size	Accu. on val	Accu. on test
Previous Works	1	LoRRA [38]	-	R-ml	ResNet	-	26.5	27.6
	2	M4C [14]	4N	R-en	ResNet	1	40.5	40.4
Improved Baseline	3	M4C [14]†	6N	G	ResNeXt	1	43.3	-
	4	M4C [14]††	6N	G	ResNeXt	5	43.8	42.4
SA-M4C	5	SA-M4C (ours)	2N→4S	G	ResNeXt	1	45.1	-
	6	SA-M4C (ours)	2N→4S	G	ResNeXt	5	**45.4**	**44.6**

2%

† Indicates our ablations for improved baseline.
†† Indicates the best model from improved baseline.

TextVQA

		Method	Structure	OCR system	DET backbone	Beam size	Accu. on val	Accu. on test
Previous Works	1	LoRRA [38]	-	R-ml	ResNet	-	26.5	27.6
	2	M4C [14]	4N	R-en	ResNet	1	40.5	40.4
Improved Baseline	3	M4C [14]†	6N	G	ResNeXt	1	43.3	-
	4	M4C [14]††	6N	G	ResNeXt	5	43.8	42.4
SA-M4C	5	SA-M4C (ours)	2N→4S	G	ResNeXt	1	45.1	-
	6	SA-M4C (ours)	2N→4S	G	ResNeXt	5	**45.4**	**44.6**

2%

4.2%

† Indicates our ablations for improved baseline.
†† Indicates the best model from improved baseline.

ST-VQA

		Method	Struc.	Beam size	VQA Accu.	ANLS on val	ANLS on test
Previous Works	1	SAN+STR [7]	-	-	-	-	0.135
	2	VTA [6]	-	-	-	-	0.282
	3	M4C [14]	4N	1	38.05	0.472	0.462
Improved Baseline	4	M4C [14]†	6N	1	40.71	0.499	-
SA-M4C	5	SA-M4C (ours)	2N→4S	1	42.12	0.510	-
	6	SA-M4C (ours)	2N→4S	5	**42.23**	**0.512**	**0.504**

4.2%

ECCV'20 ONLINE
23-28 AUGUST 2020

ST-VQA

	Method	Struc.	Beam size	VQA Accu.	ANLS on val	ANLS on test
Previous Works	1 SAN+STR [7]	-	-	-	-	0.135
	2 VTA [6]	-	-	-	-	0.282
	3 M4C [14]	4N	1	38.05	0.472	0.462
Improved Baseline	4 M4C [14]†	6N	1	40.71	0.499	-
SA-M4C	5 SA-M4C (ours)	2N→4S	1	42.12	0.510	-
	6 SA-M4C (ours)	2N→4S	5	**42.23**	**0.512**	**0.504**

More Baselines

ECCV'20 ONLINE
23-28 AUGUST 2020

ST-VQA

	Method	Struc.	Beam size	VQA Accu.	ANLS on val	ANLS on test
Previous Works	1 SAN+STR [7]	-	-	-	-	0.135
	2 VTA [6]	-	-	-	-	0.282
	3 M4C [14]	4N	1	38.05	0.472	0.462
Improved Baseline	4 M4C [14]†	6N	1	40.71	0.499	-
SA-M4C	5 SA-M4C (ours)	2N→4S	1	42.12	0.510	-
	6 SA-M4C (ours)	2N→4S	5	**42.23**	**0.512**	**0.504**

More Baselines

	Method	Struc.	Context	Accu.(val)
Ablations on Number of Layers	1 M4C [14]†	6N	-	42.70
	2 SA-M4C (ours)	4N→2S	1	43.19
	3 SA-M4C (ours)	2N→4S	1	43.80
Baselines from VQA and NLP works	4 M4C-Random	2N→4S	1	42.60
	5 M4C-Top-60 [51]	2N→4T	-	43.26
	6 M4C-ReGAT [23]	2N→4Re	-	43.20
	7 SA-M4C (ours)	2N→4S	2	**43.90**

ECCV'20 ONLINE
23-28 AUGUST 2020

ST-VQA

	Method	Struc.	Beam size	VQA Accu.	ANLS on val	ANLS on test
Previous Works	1 SAN+STR [7]	-	-	-	-	0.135
	2 VTA [6]	-	-	-	-	0.282
	3 M4C [14]	4N	1	38.05	0.472	0.462
Improved Baseline	4 M4C [14]†	6N	1	40.71	0.499	-
SA-M4C	5 SA-M4C (ours)	2N→4S	1	42.12	0.510	-
	6 SA-M4C (ours)	2N→4S	5	**42.23**	**0.512**	**0.504**

More Baselines

	Method	Struc.	Context	Accu.(val)
Ablations on Number of Layers	1 M4C [14]†	6N	-	42.70
	2 SA-M4C (ours)	4N→2S	1	43.19
	3 SA-M4C (ours)	2N→4S	1	43.80
Baselines from VQA and NLP works	4 M4C-Random	2N→4S	1	42.60
	5 M4C-Top-60 [51]	2N→4T	-	43.26
	6 M4C-ReGAT [23]	2N→4Re	-	43.20
	7 SA-M4C (ours)	2N→4S	2	**43.90**

-1.2%

ECCV'20 ONLINE
23-28 AUGUST 2020

ST-VQA

	Method	Struc.	Beam size	VQA Accu.	ANLS on val	ANLS on test
Previous Works	1 SAN+STR [7]	-	-	-	-	0.135
	2 VTA [6]	-	-	-	-	0.282
	3 M4C [14]	4N	1	38.05	0.472	0.462
Improved Baseline	4 M4C [14]†	6N	1	40.71	0.499	-
SA-M4C	5 SA-M4C (ours)	2N→4S	1	42.12	0.510	-
	6 SA-M4C (ours)	2N→4S	5	**42.23**	**0.512**	**0.504**

More Baselines

	Method	Struc.	Context	Accu.(val)
Ablations on Number of Layers	1 M4C [14]†	6N	-	42.70
	2 SA-M4C (ours)	4N→2S	1	43.19
	3 SA-M4C (ours)	2N→4S	1	43.80
Baselines from VQA and NLP works	4 M4C-Random	2N→4S	1	42.60
	5 M4C-Top-60 [51]	2N→4T	-	43.26
	6 M4C-ReGAT [23]	2N→4Re	-	43.20
	7 SA-M4C (ours)	2N→4S	2	**43.90**

-0.54%

ECCV'20 ONLINE
23-28 AUGUST 2020

ST-VQA

	Method	Struc.	Beam size	VQA Accu.	ANLS on val	ANLS on test
Previous Works	1 SAN+STR [7]	-	-	-	-	0.135
	2 VTA [6]	-	-	-	-	0.282
	3 M4C [14]	4N	1	38.05	0.472	0.462
Improved Baseline	4 M4C [14]†	6N	1	40.71	0.499	-
SA-M4C	5 SA-M4C (ours)	2N→4S	1	42.12	0.510	-
	6 SA-M4C (ours)	2N→4S	5	**42.23**	**0.512**	**0.504**

More Baselines

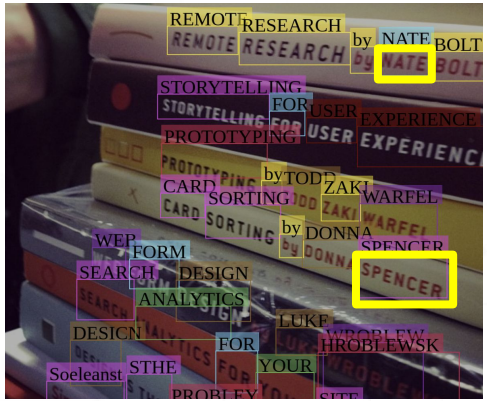
	Method	Struc.	Context	Accu.(val)
Ablations on Number of Layers	1 M4C [14]†	6N	-	42.70
	2 SA-M4C (ours)	4N→2S	1	43.19
	3 SA-M4C (ours)	2N→4S	1	43.80
Baselines from VQA and NLP works	4 M4C-Random	2N→4S	1	42.60
	5 M4C-Top-60 [51]	2N→4T	-	43.26
	6 M4C-ReGAT [23]	2N→4Re	-	43.20
	7 SA-M4C (ours)	2N→4S	2	**43.90**

-0.6%

Analysis: Spatial Reasoning

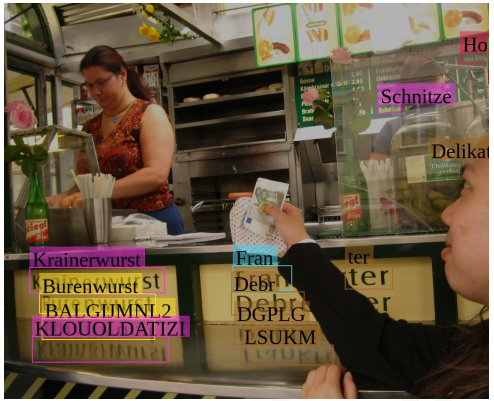
Question: **What number is the right one?**

M4C: **8953**
GT Answer: **8954**
SA-M4C: **8954**



Question: **Who is the author of the book at the top of the stack?**

M4C: **Nate Spencer**
GT Answer: **Nate Bolt**
SA-M4C: **Nate Bolt**



Question: **What is the top word on the sign on the left?**

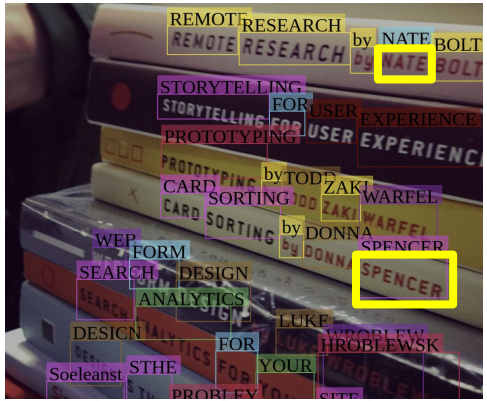
M4C: **Burenwurst**
Answer: **Krainerwurst**
SA-M4C: **Krainerwurst**

Analysis: Spatial Reasoning

+ 4.62%

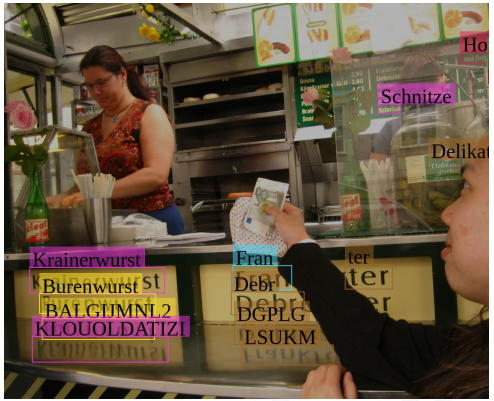
Question: **What number is the right one?**

M4C: **8953**
GT Answer: **8954**
SA-M4C: **8954**



Question: **Who is the author of the book at the top of the stack?**

M4C: **Nate Spencer**
GT Answer: **Nate Bolt**
SA-M4C: **Nate Bolt**



Question: **What is the top word on the sign on the left?**

M4C: **Burenwurst**
Answer: **Krainerwurst**
SA-M4C: **Krainerwurst**

Qualitative Samples

Original Question: What is the word written in capitals on the top right?

M4C: oh
Ours: oui
GT: oui

Flipped Question: What is the word written in capitals on the bottom left?

M4C: socialistes
Ours: www.ps-ge.ch
GT: www.ps-ge.ch

Original Question: What does it say on the top left of the sign?

M4C: andre
Ours: operation campus
GT: operation campus

Flipped Question: What does it say on the bottom right of the sign?

M4C: saint-gely-du-fesc
Ours: herault
GT: www.herault.fr

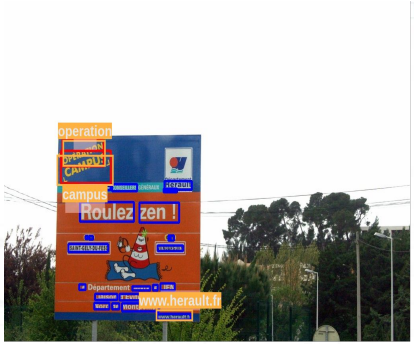
Qualitative Samples

Original Question: What is the word written in capitals on the top right?

M4C: oh
Ours: oui
GT: oui

Flipped Question: What is the word written in capitals on the bottom left?

M4C: socialistes
Ours: www.ps-ge.ch
GT: www.ps-ge.ch



Original Question: What does it say on the top left of the sign?

M4C: andre
Ours: operation campus
GT: operation campus

Flipped Question: What does it say on the bottom right of the sign?

M4C: saint-gely-du-fesc
Ours: herault
GT: www.herault.fr

Summary

We introduce transformer layer which:

1. Consumes a spatial graph to guide and sparsify the attention between object and ocr-tokens.
2. Prevents the attention from diluting across object/ocr-tokens.
3. Does not let the attention heads learn redundant features.

Thank You

Arxiv: https://arxiv.org/abs/2007.12146