Visual Question Answering

刘家伟
2021.4.28

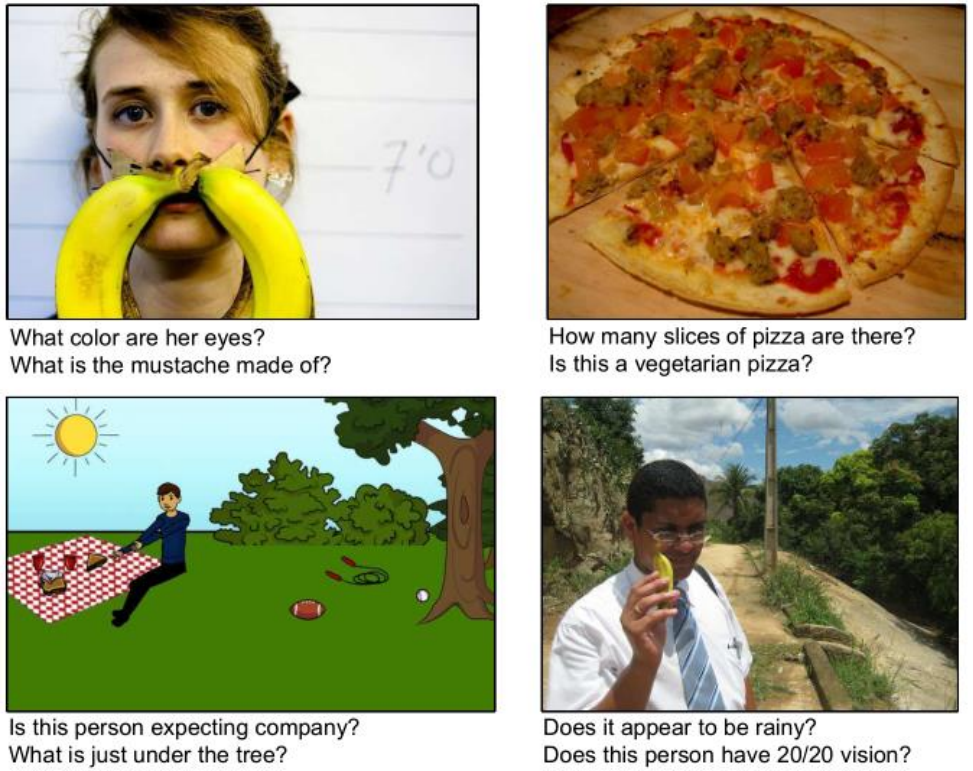
Outline

- Background
- Some Classic Architectures in VQA
- Recent Advances in Overcoming Language Priors Problem
- Recent Advances in Text VQA

Background

VQA (Visual Question Answering)

VQA是一项结合了CV和NLP的问答任务，
给定一张图片和一个问题，它的目标是从
图片的视觉信息中推理出问题的正确答案。



Background

Dataset

COCO-QA

数据集中的图像来自于MS-COCO数据集，主要包括123287张图像，其中72738张用于训练，38948用于测试，并且每张图像都有一个question/answer pair，每个answer都是一个单词。

Visual Genome

包含图像108077张和1445233个QA Pairs，图像来源是YFCC100M和COCO数据集，共有约540万张图像中的区域描述信息，这些信息能够达到精细的语义层次，问题类型是6W(what, where, how, when, who, why)。

VQA

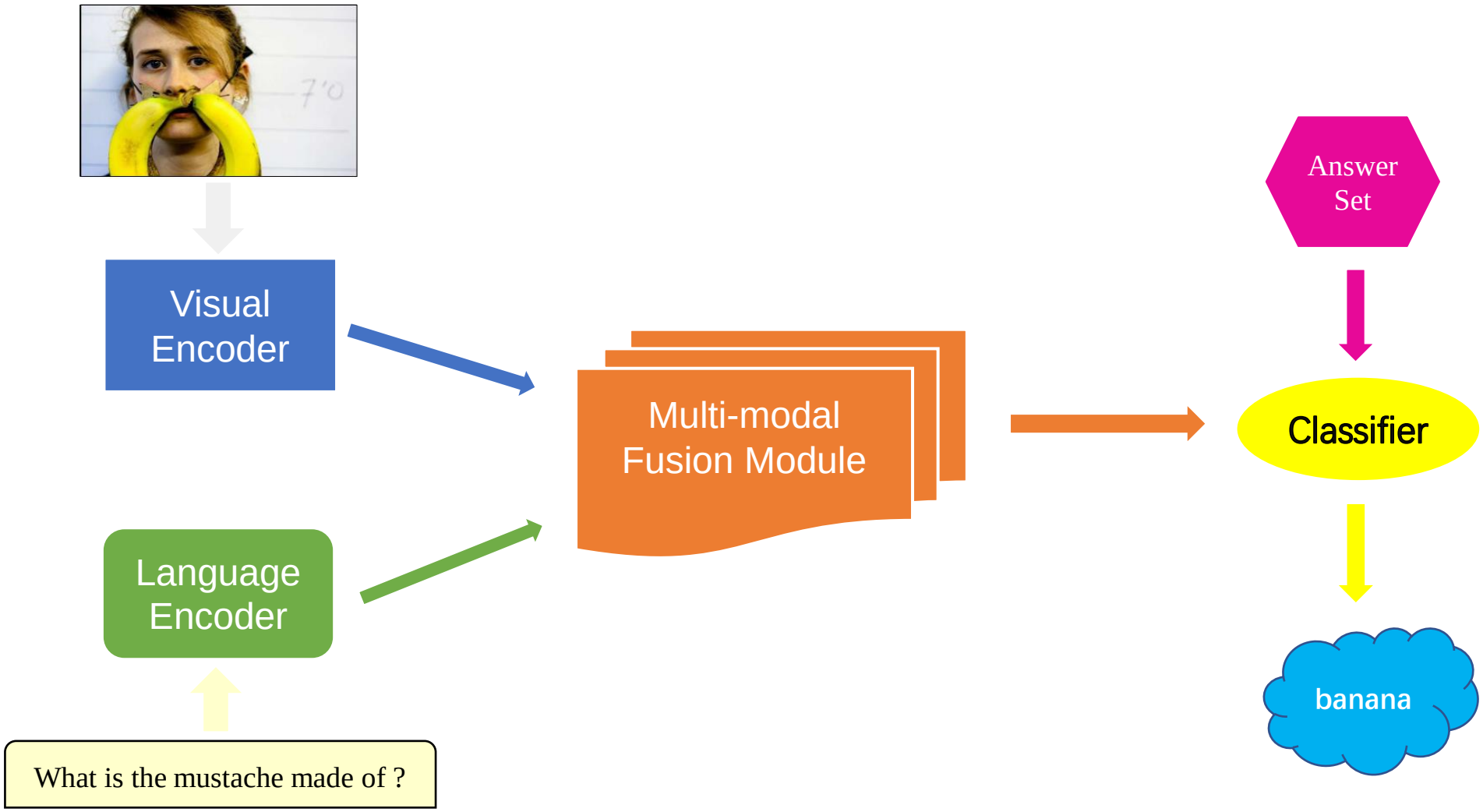
使用最广泛的数据集之一，2017年更新为VQA v2.0，包含使用真实图片的VQA-real和卡通图片的VQA-abstract。VQA-real包含123287 training和81424 test images from COCO，由真人提供开放型和是非型问题和多种候选答案，共614163个questions。VQA-abstract包括50000scenes，每个scene对应3个questions.

Evaluation Metric

WUPS：Wu-Palmer Similarity，在taxonomy tree中比较两者的common subsequence，当similarity超过某一阈值就认为是正确答案。

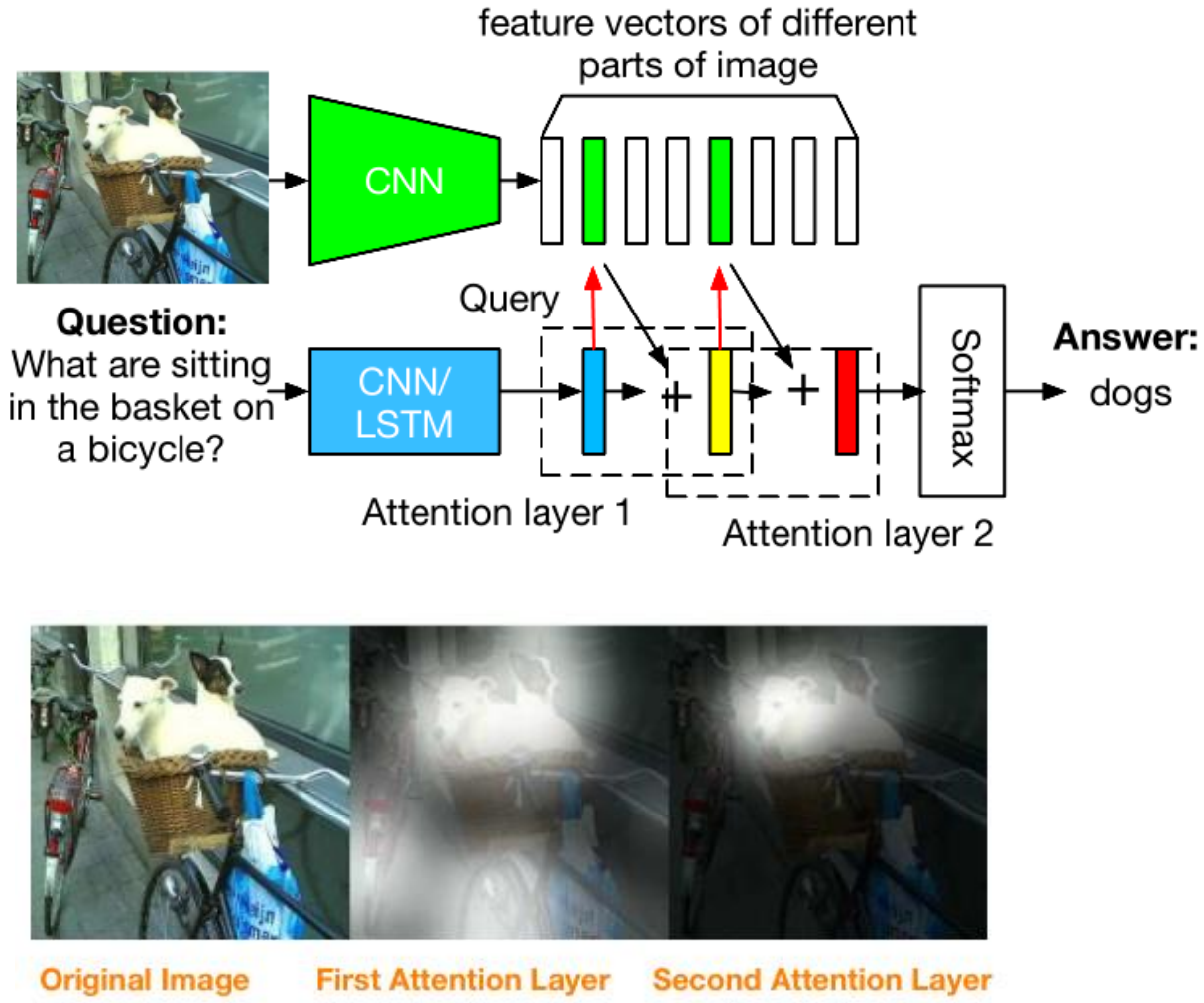
Accuracy：$\min\{\frac{\#\text{humans that said } ans}{3}, 1\}$

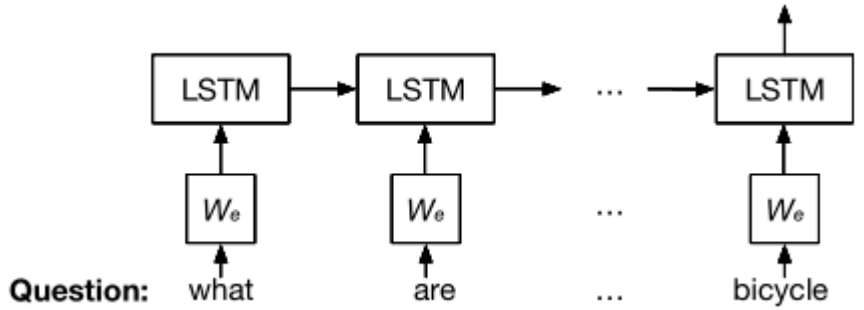
Background

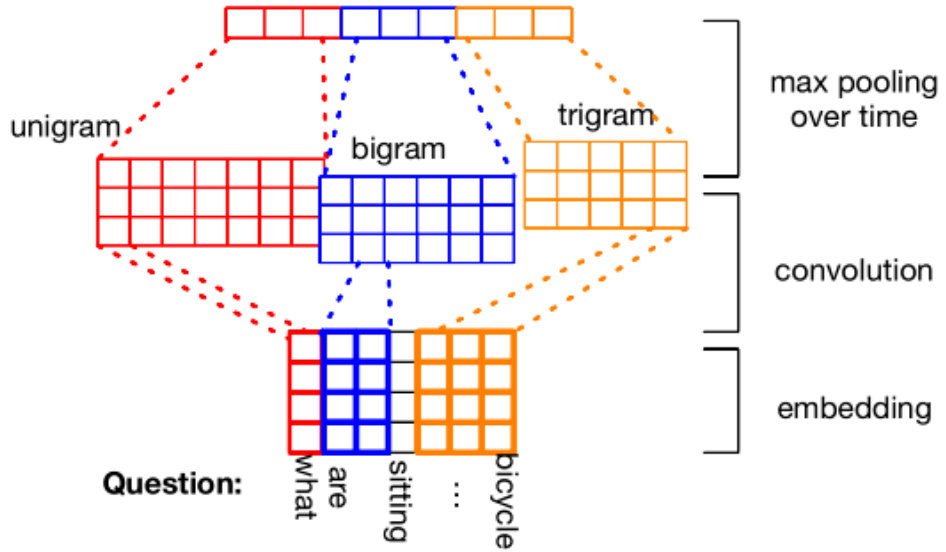

Visual Encoder
Language Encoder
What is the mustache made of ?
Multi-modal Fusion Module
Answer Set
Classifier
banana

Some Classic Architectures in VQA

SAN







CVPR 2016 Stacked Attention Networks for Image Question Answering

Some Classic Architectures in VQA

Methods	Accuracy	WUPS0.9	WUPS0.0
VSE: [21]			
GUESS	6.7	17.4	73.4
BOW	37.5	48.5	82.8
LSTM	36.8	47.6	82.3
IMG	43.0	58.6	85.9
IMG+BOW	55.9	66.8	89.0
VIS+LSTM	53.3	63.9	88.3
2-VIS+BLSTM	55.1	65.3	88.6
CNN: [17]			
IMG-CNN	55.0	65.4	88.6
CNN	32.7	44.3	80.9
Ours:			
SAN(1, LSTM)	59.6	69.6	90.1
SAN(1, CNN)	60.7	70.6	90.5
SAN(2, LSTM)	61.0	71.0	90.7
SAN(2, CNN)	**61.6**	**71.6**	**90.9**

Table 3: COCO-QA results, in percentage

Methods	Objects	Number	Color	Location
VSE: [21]				
GUESS	2.1	35.8	13.9	8.9
BOW	37.3	43.6	34.8	40.8
LSTM	35.9	45.3	36.3	38.4
IMG	40.4	29.3	42.7	44.2
IMG+BOW	58.7	44.1	52.0	49.4
VIS+LSTM	56.5	46.1	45.9	45.5
2-VIS+BLSTM	58.2	44.8	49.5	47.3
Ours:				
SAN(1, LSTM)	62.5	49.0	54.8	51.6
SAN(1, CNN)	63.6	48.7	56.7	52.7
SAN(2, LSTM)	63.6	**49.8**	57.9	52.8
SAN(2, CNN)	**64.5**	48.6	**57.9**	**54.0**

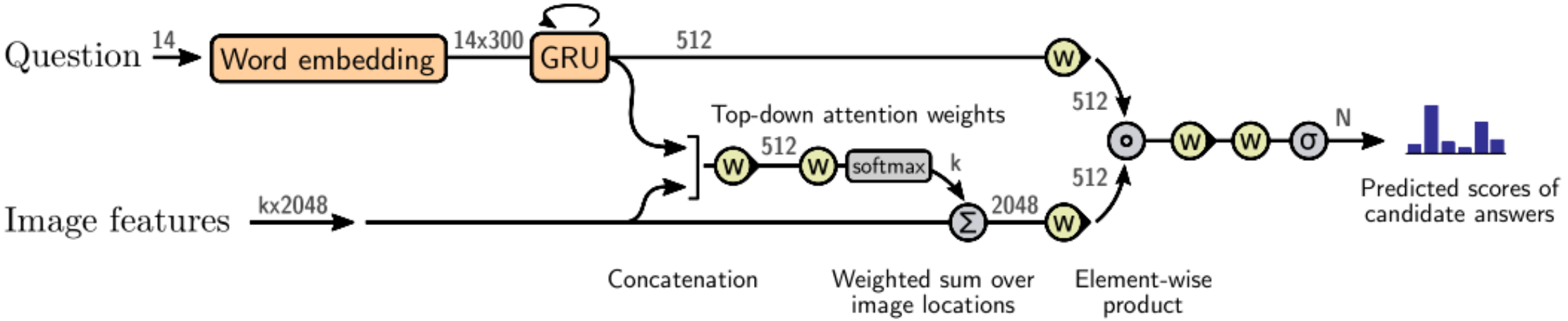
Table 4: COCO-QA accuracy per class, in percentage

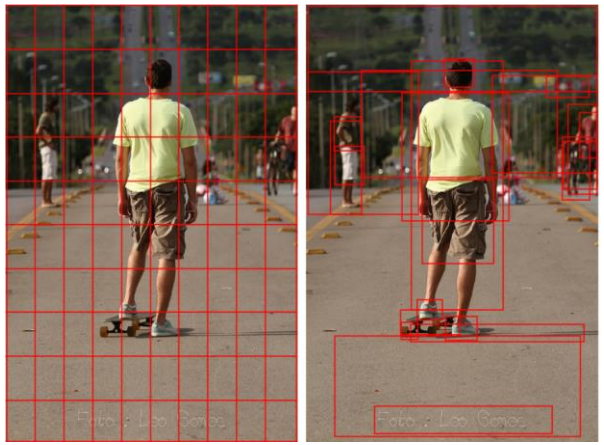
	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

CVPR 2016 Stacked Attention Networks for Image Question Answering

UpDn





CVPR 2018 Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Some Classic Architectures in VQA

	Yes/No	Number	Other	Overall
Prior [12]	61.20	0.36	1.17	25.98
Language-only [12]	67.01	31.55	27.37	44.26
d-LSTM+n-I [26, 12]	73.46	35.18	41.83	54.22
MCB [11, 12]	78.82	38.28	53.36	62.27
UPMC-LIP6	82.07	41.06	57.12	65.71
Athena	82.50	44.19	59.97	67.59
HDU-USYD-UNCC	84.50	45.39	59.01	68.09
Ours: Up-Down	**86.60**	**48.64**	**61.15**	**70.34**

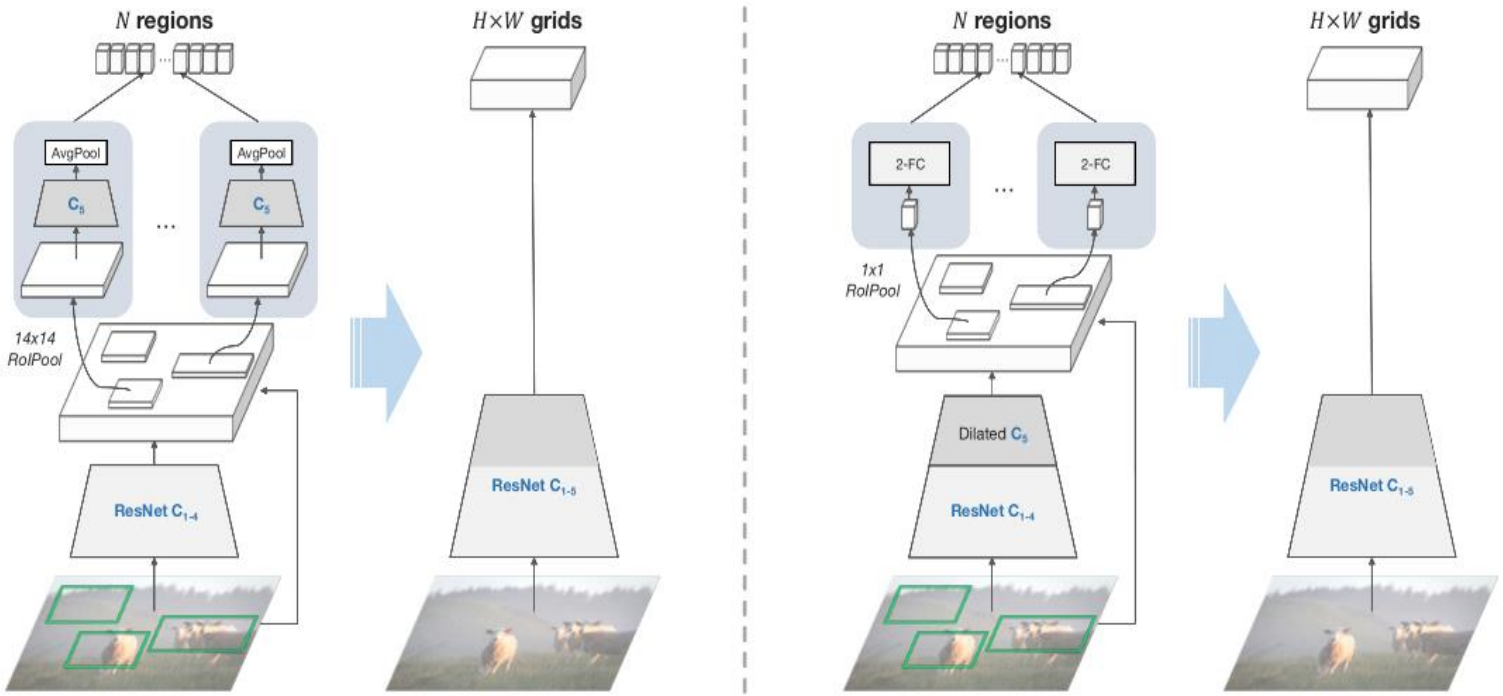
Results on VQA V2.0 test set

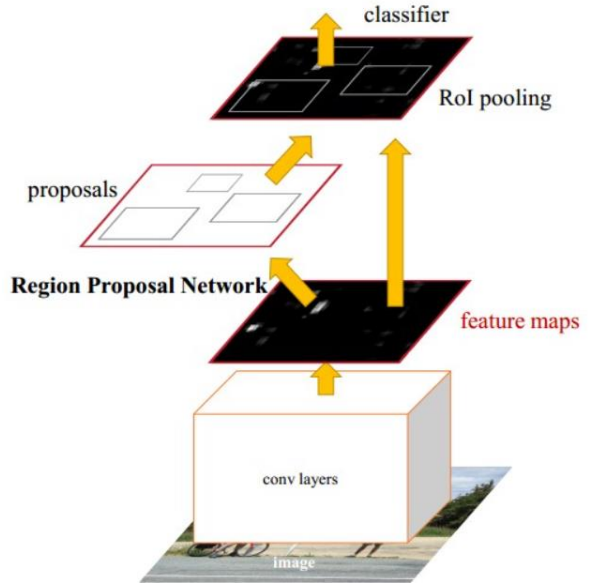
	Yes/No	Number	Other	Overall
Ours: ResNet (1×1)	76.0	36.5	46.8	56.3
Ours: ResNet (14×14)	76.6	36.2	49.5	57.9
Ours: ResNet (7×7)	77.6	37.7	51.5	59.4
Ours: Up-Down	**80.3**	**42.8**	**55.8**	**63.2**
Relative Improvement	3%	14%	8%	6%

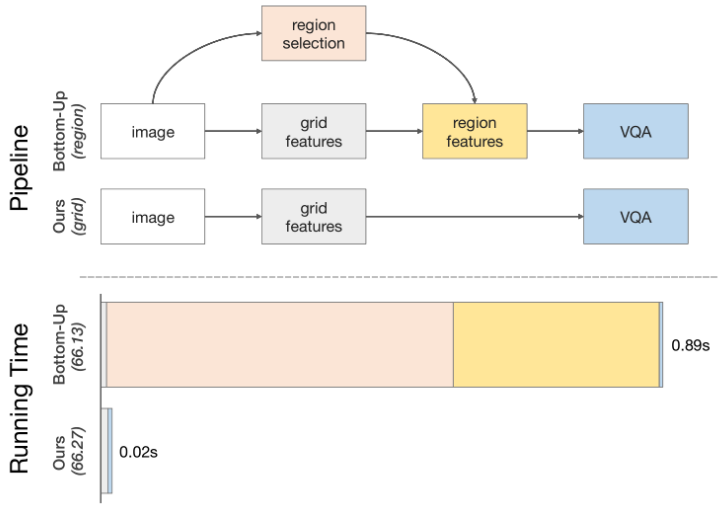
Results on VQA V2.0 validation set

CVPR 2018 Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Some Classic Architectures in VQA





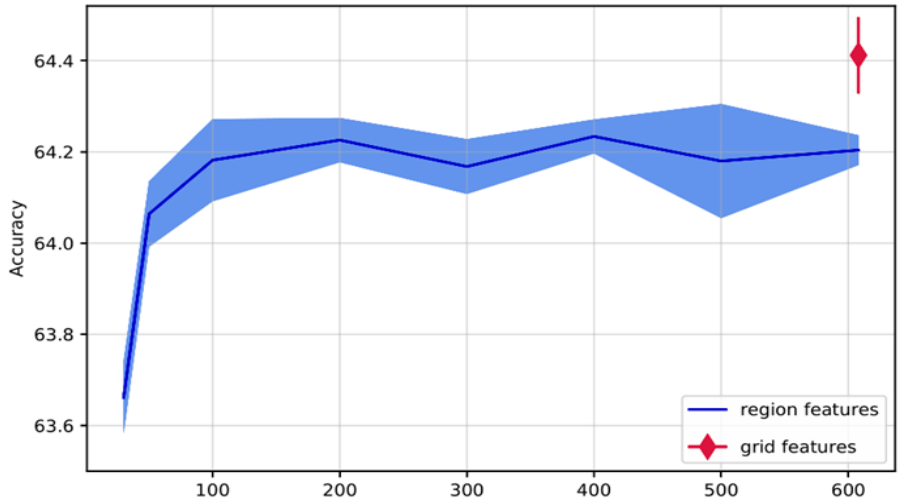


Some Classic Architectures in VQA

#	feature	VG detection pre-train: RoIPool	VG detection pre-train: region layers	VG detection pre-train: AP	VQA: accuracy	VQA: Δ
1	R [2]	14×14	C_5 [15]	4.07	64.29	-
2	R [2]	1×1	2-FC	2.90	63.94	*-0.35*
3	G	14×14	C_5	4.07	63.64	*-0.65*
4	G	1×1	2-FC	2.90	**64.37**	*0.08*
5		ImageNet pre-train			60.76	*-3.53*

	dataset	input size: shorter side	input size: longer side	# features N	accuracy
G	ImageNet	448	448	196	60.76
G	ImageNet	448	746	336	61.21
G	ImageNet	600	1000	608	61.52
G	ImageNet	800	1333	1050	61.52
G	VG	448	448	196	63.24
G	VG	448	746	336	63.81
G	VG	600	1000	608	64.37
G	VG	800	1333	1050	64.61

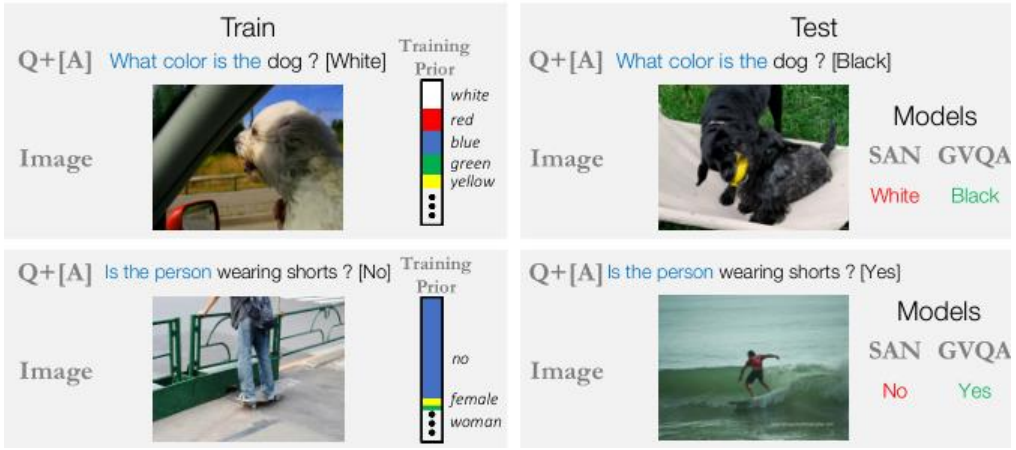
	# features (N)	test-dev accuracy	inference time breakdown (ms): shared conv.	region feat. comp.	region selection	VQA	total
R	100	66.13	9	326	548	6	889
R	608	66.22	9	322	544	7	882
G	608	66.27	11	-	-	7	18

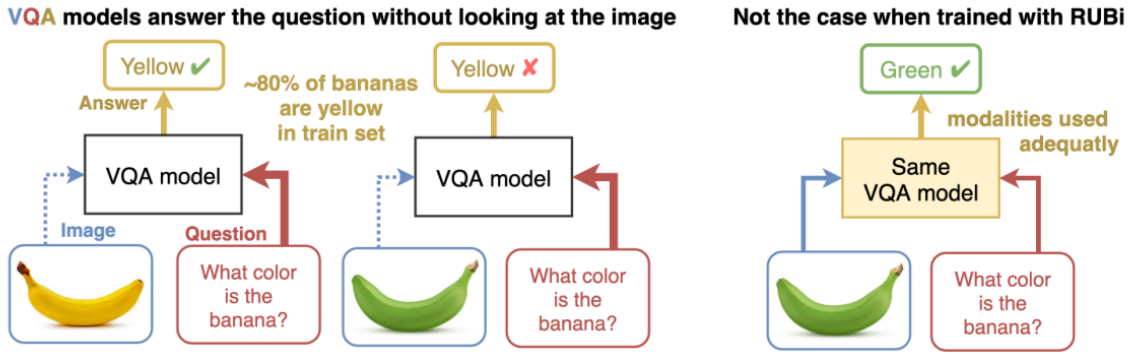


Language Priors Problem

语言先验性就是对于训练的Question与Image数据，模型并没有学会依照Image来回答问题，而只是简单的依赖answer的先验。

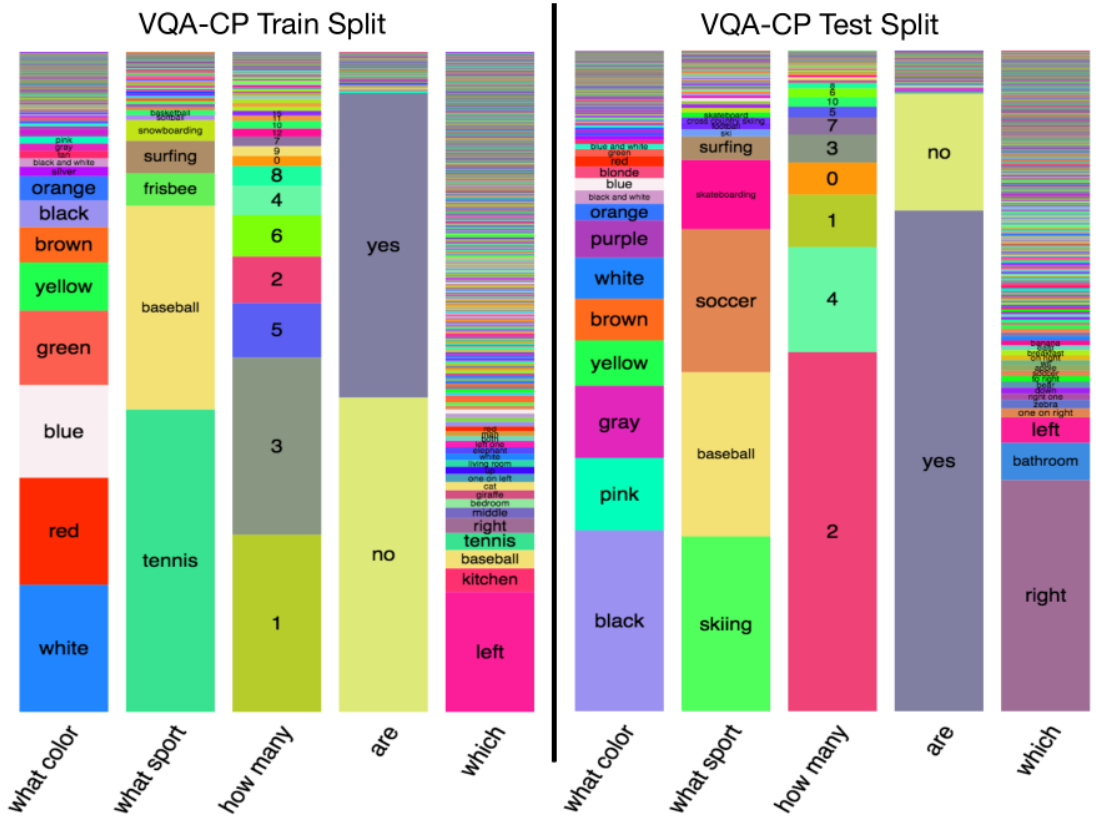
比如对于what color这类question，答案为white占比为80%，那么当输入这类问题，模型就直接回答为white，而完全不需要依照Image，且这样的正确率很高。





Recent Advances in Overcoming Language Priors Problem

VQA CP dataset

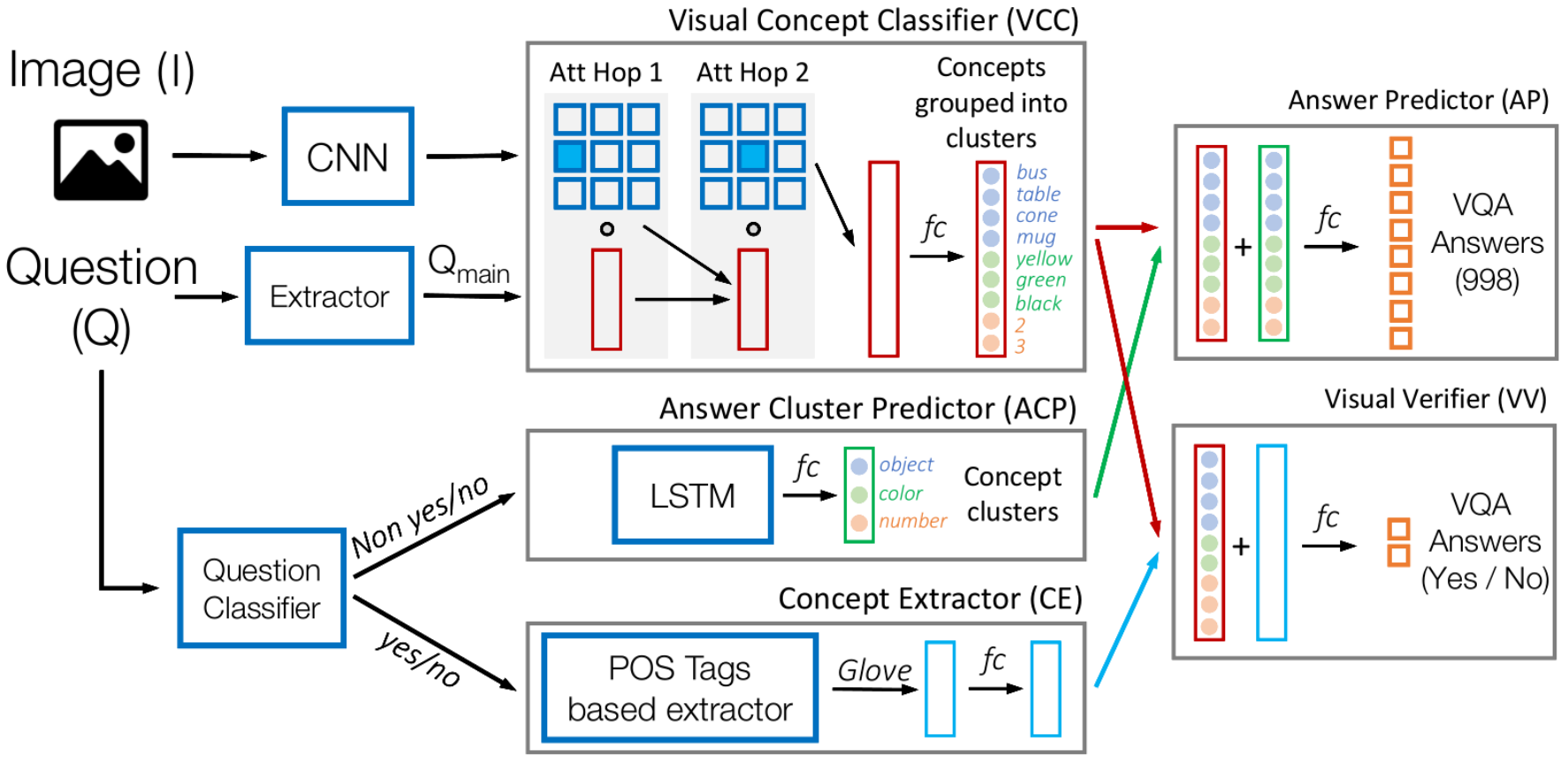


Model	Dataset	Overall	Yes/No	Number	Other	Dataset	Overall	Yes/No	Number	Other
per Q-type prior [5]	VQA v1	35.13	71.31	31.93	08.86	VQA v2	32.06	64.42	26.95	08.76
	VQA-CP v1	08.39	14.70	08.34	02.14	VQA-CP v2	08.76	19.36	11.70	02.39
d-LSTM Q [5]	VQA v1	48.23	79.05	33.70	28.81	VQA v2	43.01	67.95	30.97	27.20
	VQA-CP v1	20.16	35.72	11.07	08.34	VQA-CP v2	15.95	35.09	11.63	07.11
d-LSTM Q + norm I [24]	VQA v1	54.40	79.82	33.87	40.54	VQA v2	51.61	73.06	34.41	39.85
	VQA-CP v1	23.51	34.53	11.40	17.42	VQA-CP v2	19.73	34.25	11.39	14.41
NMN [3]	VQA v1	54.83	80.39	33.45	41.07	VQA v2	51.62	73.38	33.23	39.93
	VQA-CP v1	29.64	38.85	11.23	27.88	VQA-CP v2	27.47	38.94	11.92	25.72
SAN [39]	VQA v1	55.86	78.54	33.46	44.51	VQA v2	52.02	68.89	34.55	43.80
	VQA-CP v1	26.88	35.34	11.34	24.70	VQA-CP v2	24.96	38.35	11.14	21.74
MCB [11]	VQA v1	60.97	81.62	34.56	52.16	VQA v2	59.71	77.91	37.47	51.76
	VQA-CP v1	34.39	37.96	11.80	39.90	VQA-CP v2	36.33	41.01	11.96	40.57

CVPR 2018 Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

Recent Advances in Overcoming Language Priors Problem

GVQA



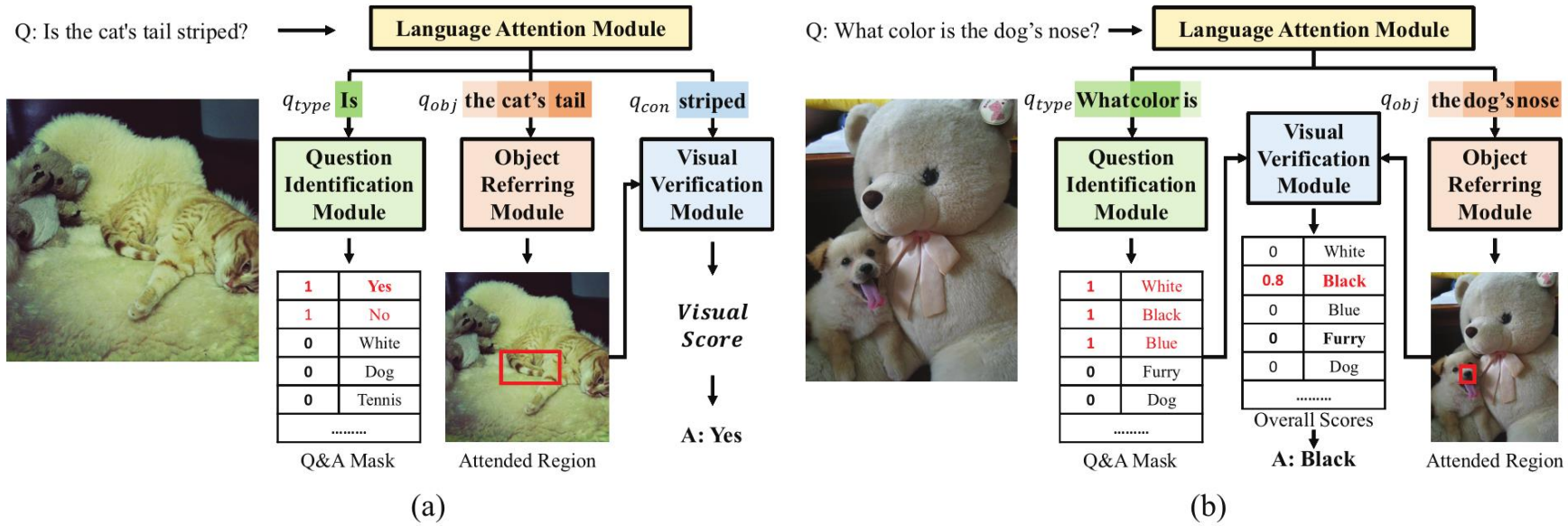
CVPR 2018 Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

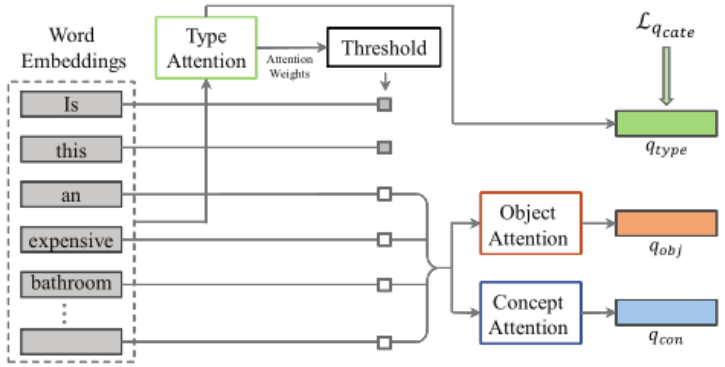
Recent Advances in Overcoming Language Priors Problem

Dataset	Model	Overall	Yes/No	Number	Other
VQA-CP v1	SAN [39]	26.88	35.34	11.34	24.70
	GVQA (Ours)	**39.23**	**64.72**	**11.87**	**24.86**
VQA-CP v2	SAN [39]	24.96	38.35	11.14	21.74
	GVQA (Ours)	**31.30**	**57.99**	**13.68**	**22.14**

CVPR 2018 Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

Recent Advances in Overcoming Language Priors Problem

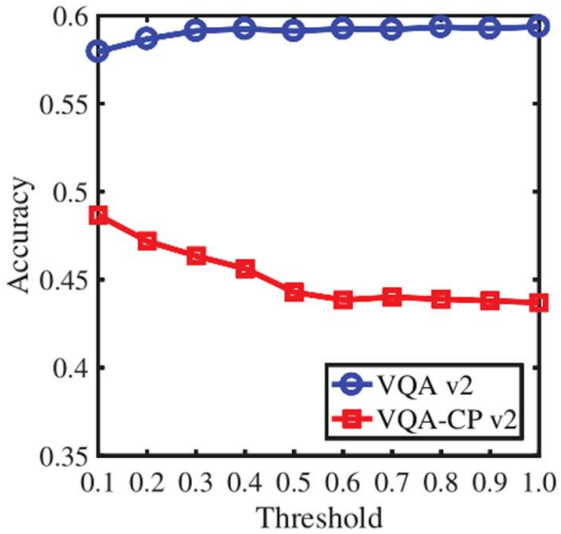




CVPR 2020 Overcoming Language Priors in VQA via Decomposed Linguistic Representations

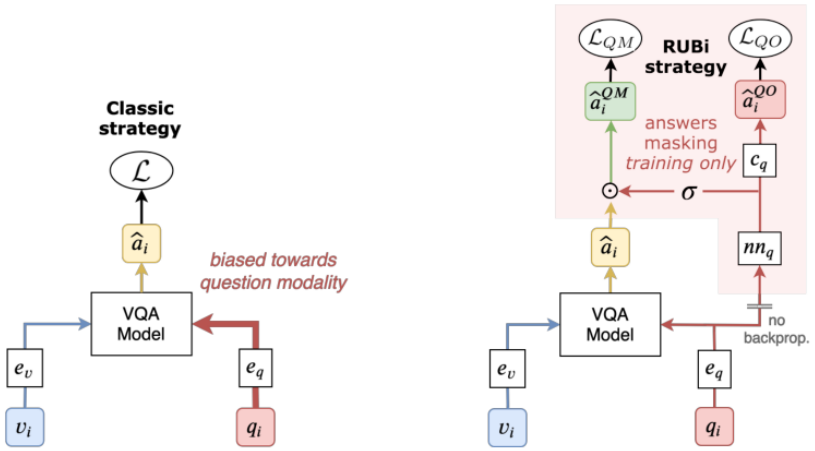
Recent Advances in Overcoming Language Priors Problem

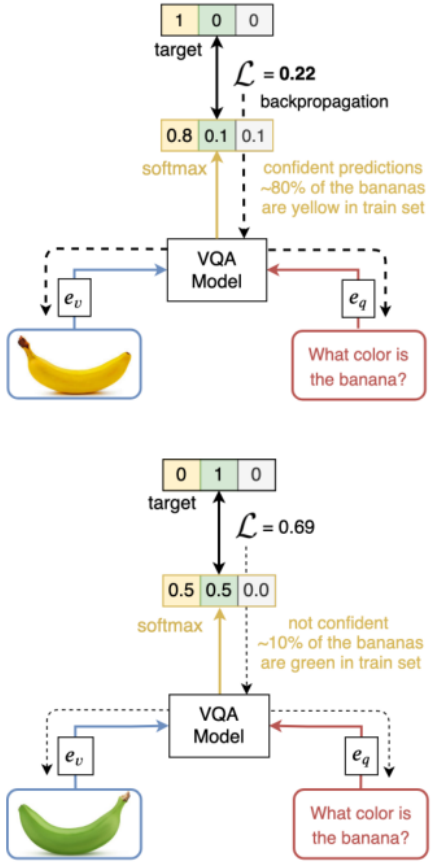
Model	*VQA-CP v2 test*				*VQA v2 val*			
	Overall	Yes/No	Numbers	Other	Overall	Yes/No	Numbers	Other
SAN (Yang et al. 2016)	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
UpDn (Anderson et al. 2018)	39.49	45.21	11.96	42.98	62.85	80.89	42.78	54.44
GVQA (SAN) (Agrawal et al. 2018)	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
AdvReg (UpDn) (Ramakrishnan et al. 2018)	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
HINT (UpDn) (Selvaraju et al. 2019)	47.7	70.04	10.68	**46.31**	62.35	80.49	41.75	54.01
Ours (SAN)	34.83	57.28	15.11	28.48	49.27	66.71	32.47	40.43
Ours (UpDn)	**48.87**	**70.99**	**18.72**	45.57	57.96	76.82	39.33	48.54



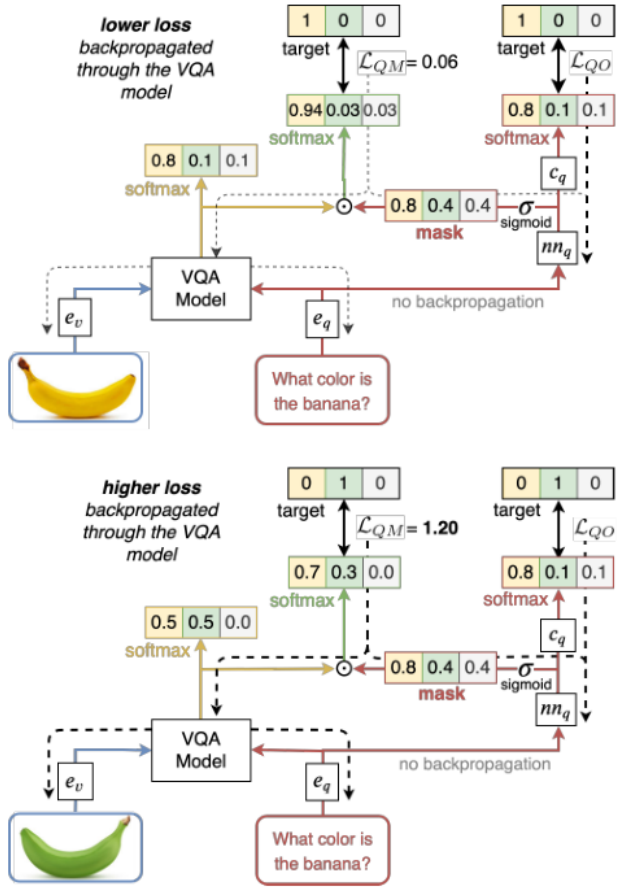
CVPR 2020 Overcoming Language Priors in VQA via Decomposed Linguistic Representations

Recent Advances in Overcoming Language Priors Problem





(a) Classical learning strategy



(b) RUBi learning strategy

NeurIPS 2019 RUBi: Reducing Unimodal Biases in Visual Question Answering

Recent Advances in Overcoming Language Priors Problem

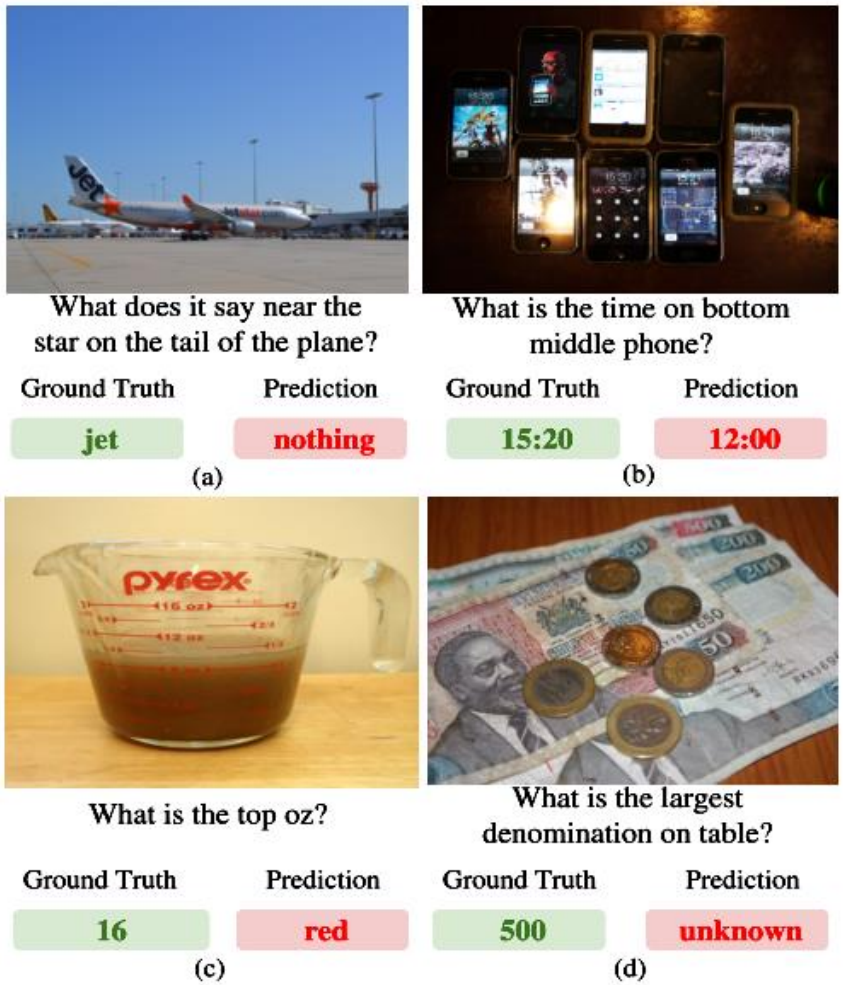
Model	Overall	Answer type		
		Yes/No	Number	Other
Question-Only [10]	15.95	35.09	11.63	7.11
UpDn [15] **	38.01	.	.	.
RAMEN [45]	39.21	.	.	.
BAN [19] **	39.31	.	.	.
MuRel [16]	39.54	42.85	13.17	45.04
UpDn [15] *	39.74	42.27	11.93	**46.05**
UpDn + Q-Adv + DoE [25]	41.17	65.49	15.48	35.48
Balanced Sampling	40.38	57.99	10.07	39.23
Q-type Balanced Sampling	42.11	61.55	11.26	40.39
Baseline architecture (ours)	38.46 ± 0.07	42.85 ± 0.18	12.81 ± 0.20	43.20 ± 0.15
RUBi (ours)	**47.11** ± 0.51	**68.65** ± 1.16	**20.28** ± 0.90	43.18 ± 0.43

Model	val	test-dev
Baseline (ours)	**63.10**	**64.75**
RUBi (ours)	61.16	63.18

NeurIPS 2019 RUBi: Reducing Unimodal Biases in Visual Question Answering

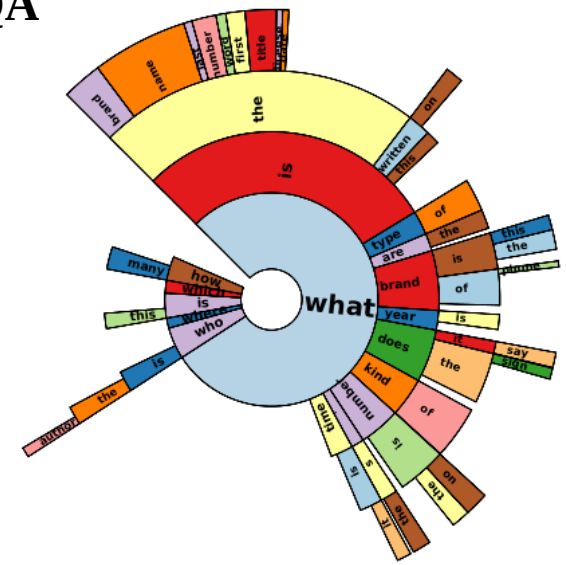
Text VQA

TextVQA要求模型阅读和推理图像中的文本，以回答有关它们的问题。具体来说，模型需要合并图像中出现的文本，并通过推理来回答TextVQA问题。



Recent Advances in Text VQA

TextVQA



数据集：28,408 张图像，45,336 个问题 (37,912)，453,360 个答案 (26,263)
训练集：21,953 张图像，34,602 个问题
验证集： 3,166 张图像， 5,000 个问题
测试集： 3,289 张图像， 5,734 个问题

数据来源：Open Images v3 dataset

每张图像 1-2 个问题，每个问题 10 个答案，问题的平均长度为 7.18 个单词，答案的平均长度为 1.58 个单词

ST-VQA

数据集：23,038张图像，31,791个问题
训练集：19,027张图像，26,308个问题
测试集： 2,993 张图像， 4,163个问题

数据来源： Coco-Text, Visal Genome, VizWiz, ICDAR(13+15), ImageNet, IIIT-STR

$$\text{ANLS} = \frac{1}{N}\sum_{i=0}^{N}\left(\max_{j} s(a_{ij}, o_{q_i})\right) \qquad (1)$$

$$s(a_{ij}, o_{q_i}) = \begin{cases} (1 - NL(a_{ij}, o_{q_i})) & \text{if } NL(a_{ij}, o_{q_i}) < \tau \\ 0 & \text{if } NL(a_{ij}, o_{q_i}) \geqslant \tau \end{cases}$$

CVPR 2019 Towards VQA Models That Can Read
CVPR 2019 Scene Text Visual Question Answering

Recent Advances in Text VQA

EST-VQA (Chinese + English)

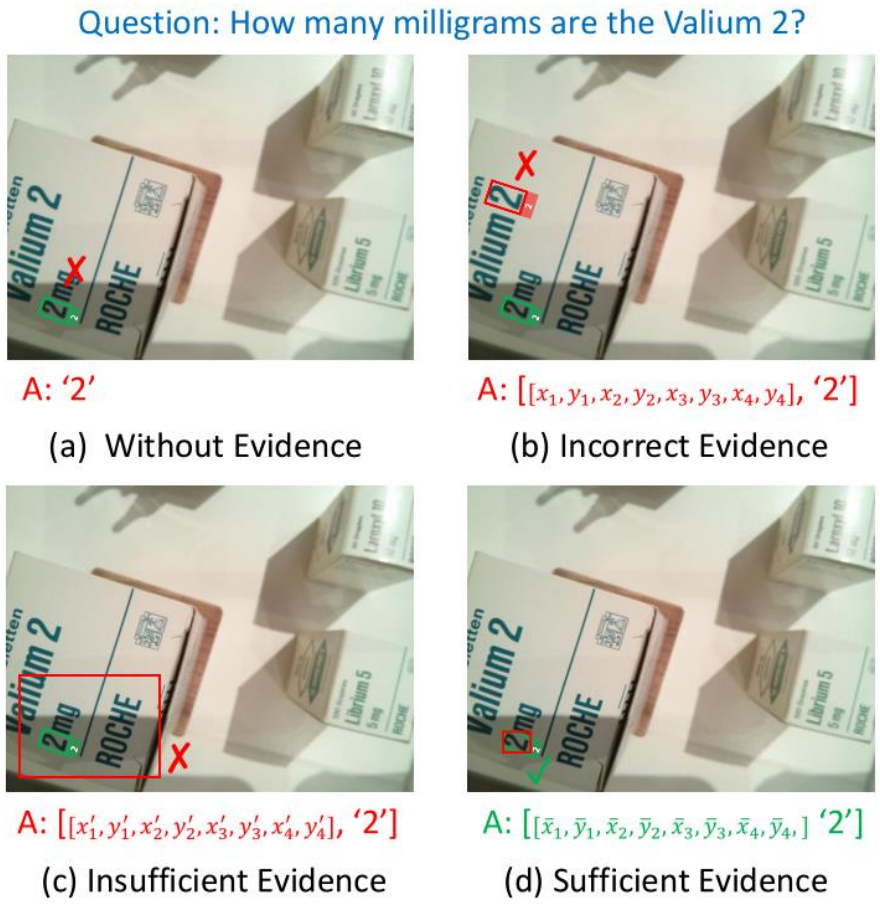
数据集：25239张图像，28062个问题
训练集：20757张图像，23062个问题
测试集： 4482张图像， 5000个问题

数据来源：
英文数据：Total-Text, ICDAR2013, ICDAR2015, CTW1500, MLT, COCO-Text
中文数据：LSVT

15056个英文问题和13006个中文问题；
只可以通过图像中文本回答，并且还标注了该答案对应的矩形边界框（证据）

数据集的中英分布：

Set	English		Chinese		All	
	# I	# Q	# I	# Q	# I	# Q
Train	11,383	12,556	9,374	10,506	20,757	23,062
Test	2,267	2,500	2,215	2,500	4,482	5,000
Total	13,650	15,056	11,589	13,006	25,239	28,062

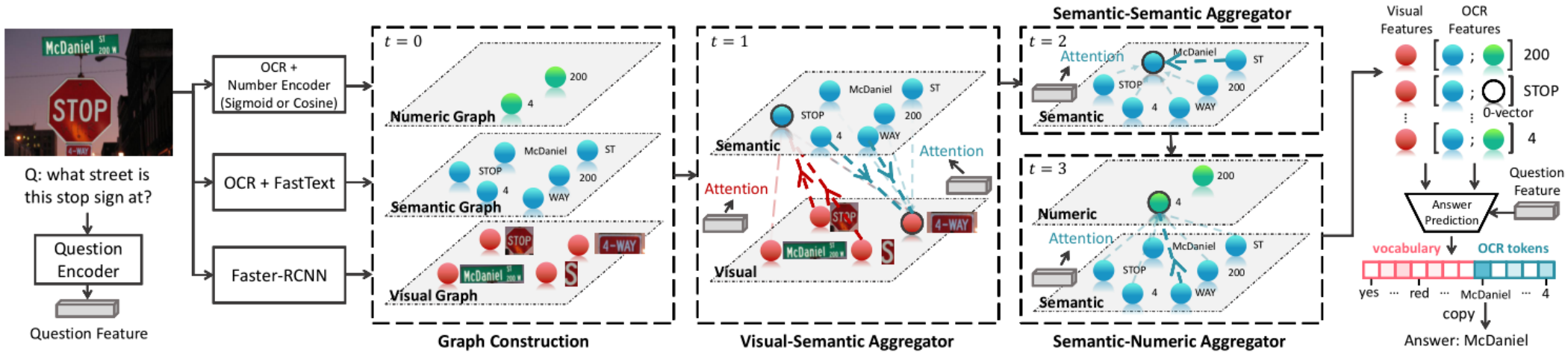


$$E^i_\tau = f(\frac{B_{gt} \cap B_{det}}{B_{gt} \cup B_{det}}) = \begin{cases} Incorrect, & E = 0 \\ Insufficient, & 0 < E < \theta \\ Sufficient, & E \geq \theta \end{cases}$$

$$s_e(ans, gt, E) = \begin{cases} s_l, & \text{if } E \text{ sufficient} \\ 0, & \text{else} \end{cases}$$

CVPR 2020 On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering

MM-GNN



CVPR 2020 Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text

Recent Advances in Text VQA

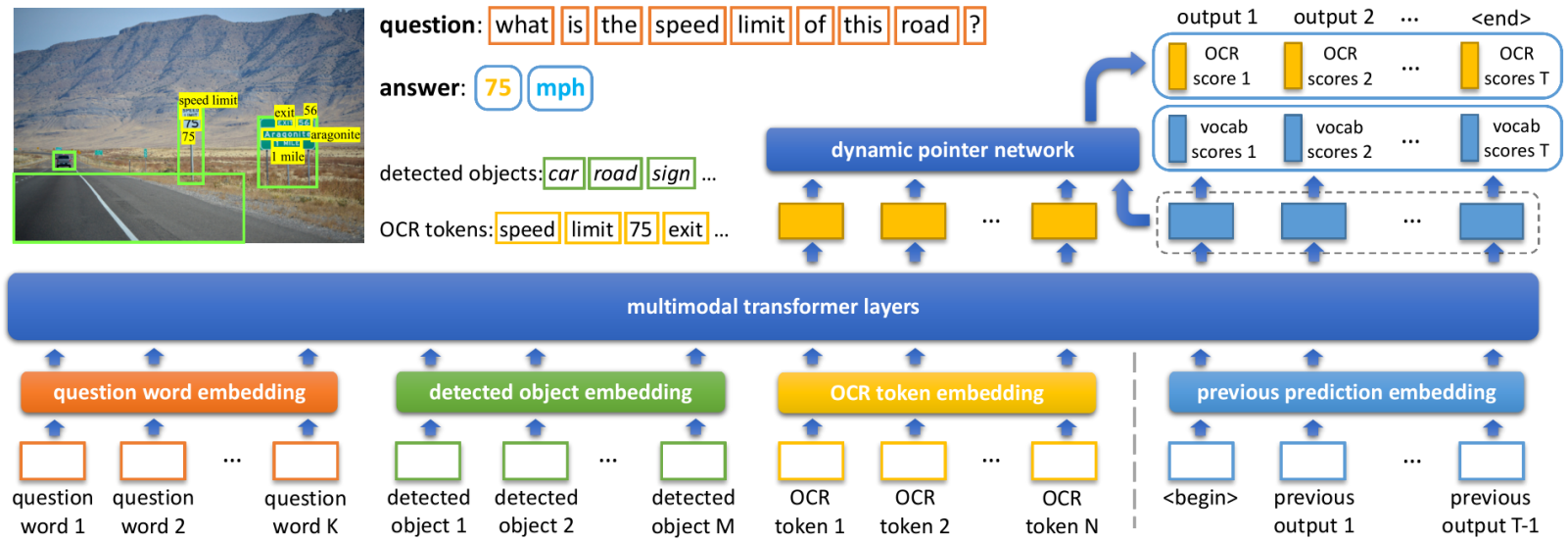
Method	Val	Test
Pythia	13.04	14.01
LoRRA (BAN)	18.41	-
LoRRA (Pythia)	26.56	27.63
BERT + MFH	28.96	-
MM-GNN (ours)	**31.44**	**31.10**
BERT + MFH (ensemble)	31.50	31.44
MM-GNN (ensemble) (ours)	**32.92**	**32.46**
LA+OCR UB	67.56	68.24

Method	Weakly Contextualized		Open Dictionary	
	ANLS	Acc.	ANLS	Acc.
SAAA	0.085	6.36	0.085	6.36
SAAA+STR	0.096	7.41	0.096	7.41
SAN(LSTM)+STR	0.136	10.34	0.136	10.34
SAN(CNN)+STR	0.135	10.46	0.135	10.46
VTA [7]	0.279	17.77	0.282	18.13
MM-GNN (ours)	0.203	15.69	0.207	16.00

CVPR 2020 Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text

Recent Advances in Text VQA

M4C

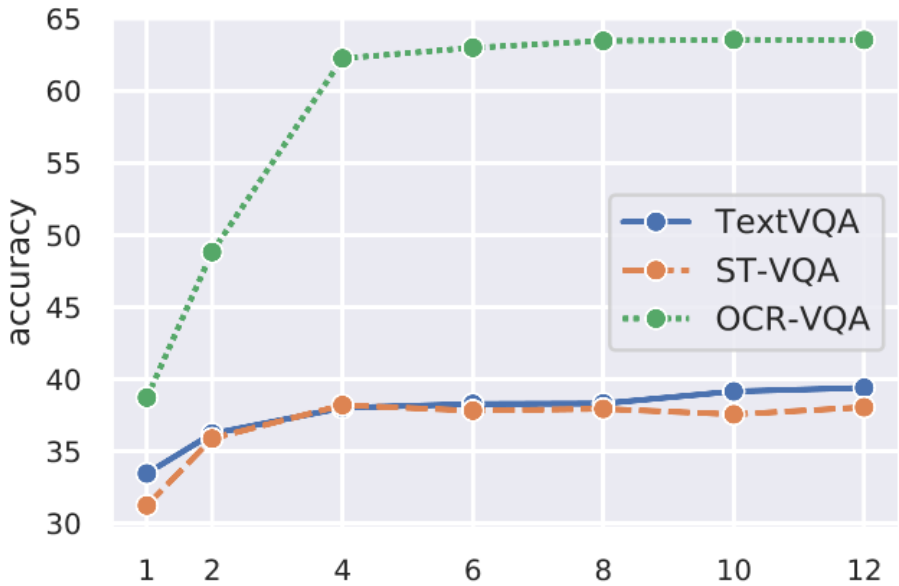


CVPR 2020 Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

Recent Advances in Text VQA

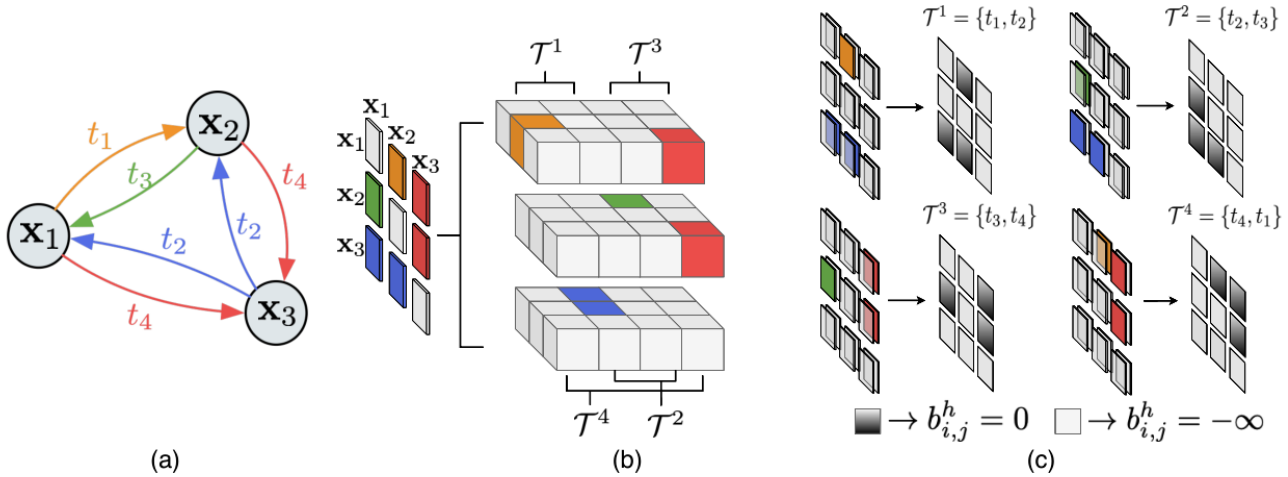
#	Method	Question enc. pretraining	OCR system	OCR token representation	Output module	Accu. on val	Accu. on test
1	LoRRA [44]	GloVe	Rosetta-ml	FastText	classifier	26.56	27.63
2	M4C w/o dec.	GloVe	Rosetta-ml	FastText	classifier	29.36	–
3	M4C w/o dec.	(none)	Rosetta-ml	FastText	classifier	29.55	–
4	M4C 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**

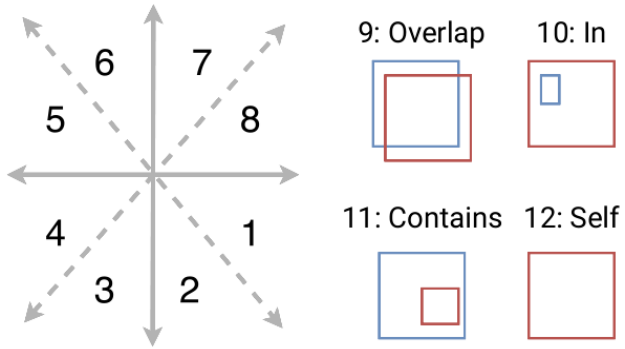
#	Method	Output module	Accu. on val	ANLS on val	ANLS on test
1	SAN+STR [8]	–	–	–	0.135
2	VTA [7]	–	–	–	0.282
3	M4C w/o dec.	classifier	33.52	0.397	–
4	M4C (ours)	decoder	**38.05**	**0.472**	**0.462**



CVPR 2020 Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA

SA-M4C





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

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

$$b_{i,j}^h = \begin{cases} \beta_{t_l}^h & t_l \in \mathcal{T}^h, \quad \mathbf{x}_i, \mathbf{x}_j \in X \\ -\infty & \text{otherwise} \end{cases},$$

ECCV 2020 Spatially Aware Multimodal Transformers for TextVQA

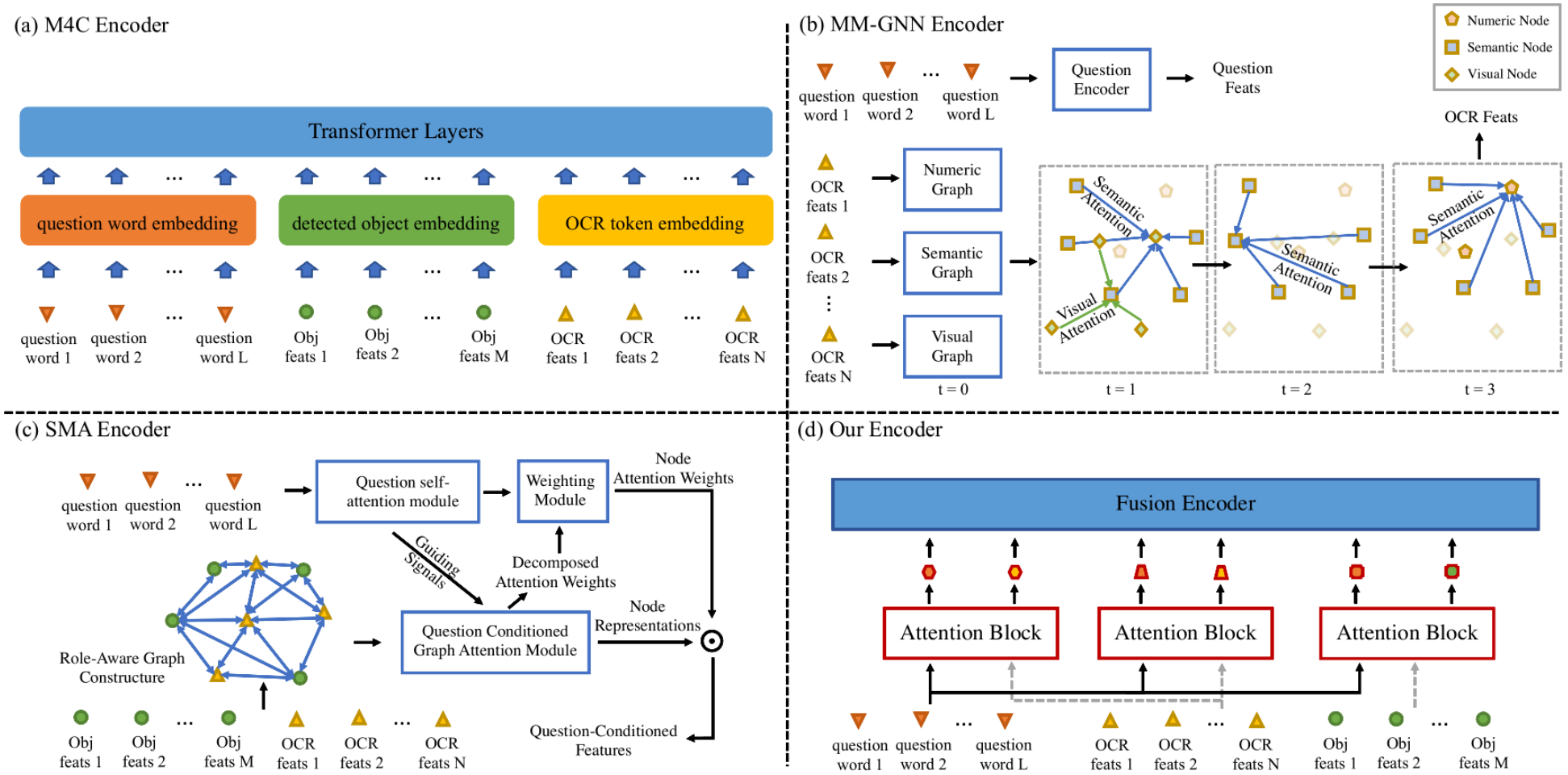
Recent Advances in Text VQA

	Method	Structure	OCR system	DET backbone	w/ ST-VQA	Beam size	Accu. on val	Accu. on test
1	LoRRA [38]	-	R-ml	ResNet	✗	-	26.5	27.6
2	DCD [25]	-	-	-	-	-	31.4	31.4
3	MSFT [40]	-	-	-	-	-	32.9	32.4
4	M4C [14]	4N	R-en	ResNet	✗	1	39.4	39.0
5	M4C [14]	4N	R-en	ResNet	✓	1	40.5	40.4
6	M4C [14]†	4N	G	ResNet	✗	1	41.8	-
7	M4C [14]†	4N	G	ResNeXt	✗	1	42.0	-
8	M4C [14]†	6N	G	ResNeXt	✗	1	42.7	-
9	M4C [14]†	6N	G	ResNeXt	✓	1	43.3	-
10	M4C [14]††	6N	G	ResNeXt	✓	5	43.8	42.4
11	SA-M4C (ours)	2N→4S	G	ResNeXt	✗	1	43.9	-
12	SA-M4C (ours)	2N→4S	G	ResNeXt	✓	1	45.1	-
13	SA-M4C (ours)	2N→4S	G	ResNeXt	✓	5	**45.4**	**44.6**

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

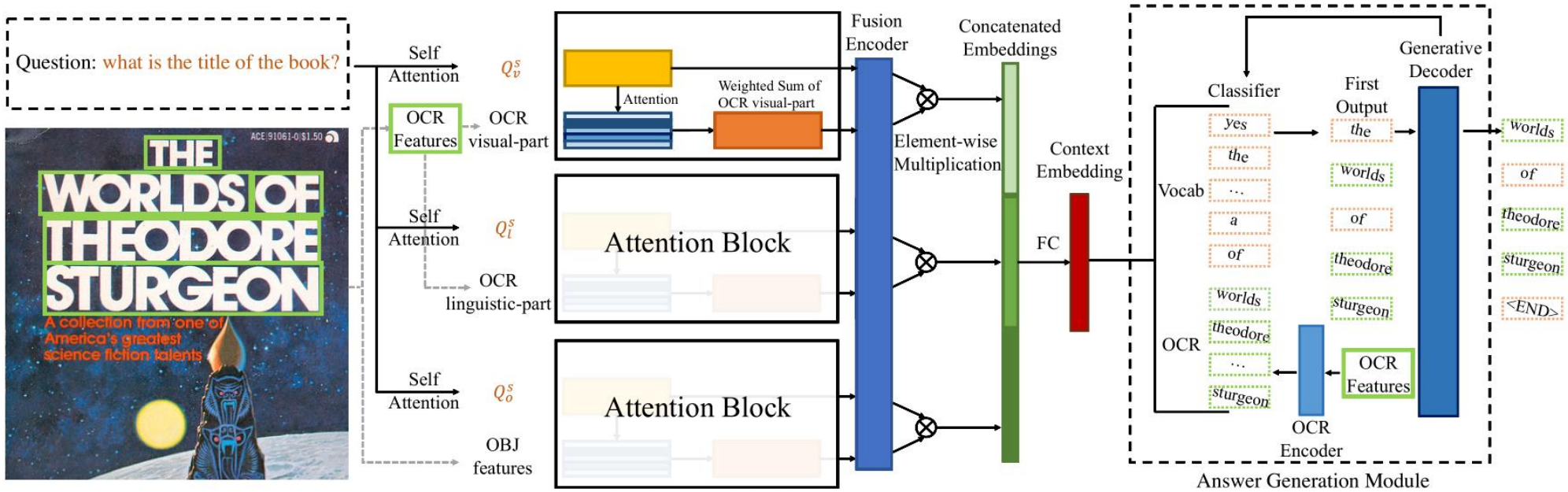
ECCV 2020 Spatially Aware Multimodal Transformers for TextVQA

Recent Advances in Text VQA



2020 Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps

Recent Advances in Text VQA



2020 Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps

Recent Advances in Text VQA

#	Method	OCR system	Accu. on val	Accu. on test
1	LoRRA (Singh et al. 2019)	Rosetta-ml	26.56	27.63
2	DCD ZJU (Lin et al. 2019)	-	31.48	31.44
3	MSFT VTI (Anonymous 2019)	-	32.92	32.46
4	M4C (Hu et al. 2019)	Rosetta-en	39.40	39.01
5	SA-M4C (Kant et al. 2020)	Google OCR	45.40	44.60
6	SMA (Gao et al. 2020a)	SBD-Trans	44.58	45.51
7	ours (three-block)	Rosetta-en	40.38	40.92
8	ours (three-block w/ST-VQA)	SBD-Trans	**45.53**	**45.66**

#	Method	ANLS on test1	ANLS on test2	ANLS on test3
1	M4C (Hu et al. 2019)	-	-	0.4621
2	SA-M4C (Kant et al. 2020)	-	0.4972	0.5042
3	SMA (Gao et al. 2020a)	0.5081	0.3104	0.4659
4	ours	0.5060	0.5047	0.5089
5	ours(w/TextVQA)	**0.5490**	**0.5513**	**0.5500**

2020 Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps