



بنام خداوند جان و خرد

پرسش و پاسخ تصویری

ارائه دهنده: مریم سادات هاشمی

استاد راهنما: دکتر سید صالح اعتمادی

شرح مسئله



شرح مسئله



پیراهن کودک چه رنگی است؟

شرح مسئله



پیراهن کودک چه رنگی است؟

سیستم پرسش و
پاسخ تصویری

شرح مسئله



پیراهن کودک چه رنگی است؟

سیستم پرسش و
پاسخ تصویری

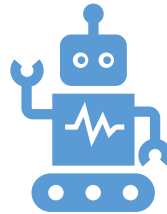
قرمز

کاربردها



دستیار هوشمند برای
افراد کم بینا و نابینا

کاربردها



تعامل با ربات‌ها

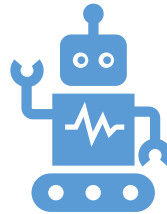


دستیار هوشمند برای
افراد کم بینا و نابینا

کاربردها



کمکی به پزشکان برای
تفسیر تصاویر پیچیده پزشکی



تعامل با ربات‌ها



دستیار هوشمند برای
افراد کم بینا و نابینا

چالش‌ها

- مسئله پرسش و پاسخ تصویری **پیچیدگی بیشتری** نسبت به مسئله پرسش و پاسخ متنی دارد:

چالش‌ها

- مسئله پرسش و پاسخ تصویری **پیچیدگی بیشتری** نسبت به مسئله پرسش و پاسخ متنی دارد:

۱. زیرا تصاویر **بعد بالاتر و نویز بیشتری** نسبت به متن دارند.

چالش‌ها

- مسئله پرسش و پاسخ تصویری **پیچیدگی بیشتری** نسبت به مسئله پرسش و پاسخ متنی دارد:

۱. زیرا تصاویر **بعد بالاتر و نویز بیشتری** نسبت به متن دارند.

۲. تصاویر **فاقد ساختار و قواعد دستوری زبان** هستند.

چالش‌ها

• مسئله پرسش و پاسخ تصویری **پیچیدگی بیشتری** نسبت به مسئله پرسش و پاسخ متنی دارد:

۱. زیرا تصاویر **بعد بالاتر و نویز بیشتری** نسبت به متن دارند.

۲. تصاویر **فاقد ساختار و قواعد دستوری زبان** هستند.

۳. تصاویر **غنای بیشتری** از دنیای واقعی را ضبط می‌کنند.

چالش‌ها

- مسئله پرسش و پاسخ تصویری **پیچیدگی بیشتری** نسبت به مسئله پرسش و پاسخ متنی دارد:

۱. زیرا تصاویر **بعد بالاتر و نویز بیشتری** نسبت به متن دارند.

۲. تصاویر **فاقد ساختار و قواعد دستوری زبان** هستند.

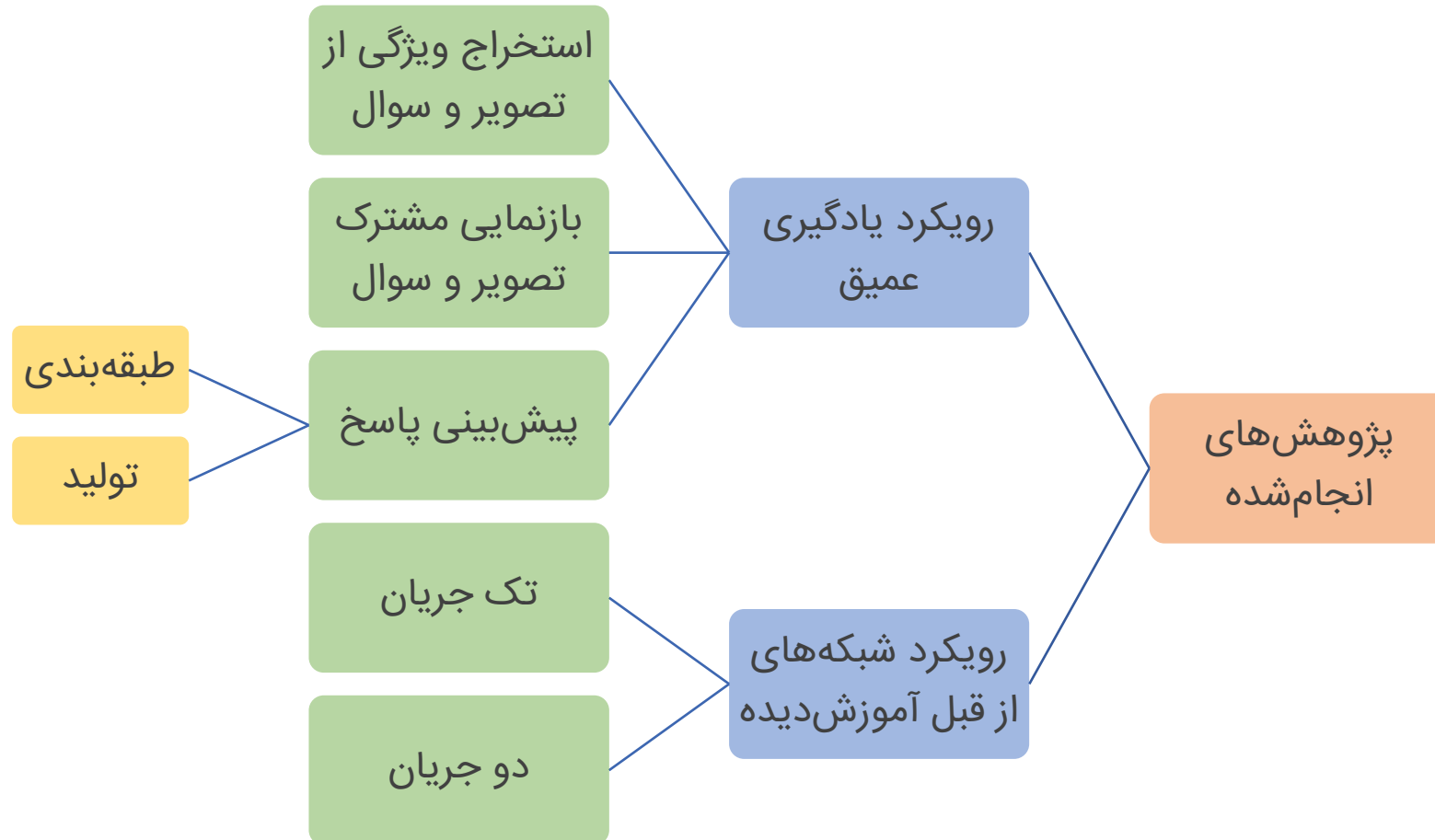
۳. تصاویر **غنای بیشتری** از دنیای واقعی را ضبط می‌کنند.

- مسئله پرسش و پاسخ تصویری نیاز به **حل مسائل پایه‌ای و فرعی** دارد مانند تشخیص اشیا، تشخیص فعالیت، طبقه‌بندی صفات، شمارش و روابط مکانی بین اشیا

مجموعه دادگان

- دادگان DAQUAR
- دادگان VQA
- دادگان Visual Madlibs
- دادگان Visual7W
- دادگان CLEVR
- دادگان Tally-QA
- دادگان KVQA

کارهای انجام شده



رویکرد یادگیری عمیق: استخراج ویژگی از تصویر

مدل پرسش و پاسخ تصویری	AlexNet	VGGNet	GoogleNet	ResNet
Image_QA		✓		
Talk_to_Machine			✓	
VQA		✓		
Vis_Madlibs	✓	✓		
VIS + LSTM		✓		
Ahab		✓		
ABC-CNN		✓		
Comp_QA		✓		
DPPNet		✓		
Answer_CNN		✓		
VQA-Caption		✓		
Re_Baseline				✓
MCB				✓
SMem-VQA			✓	
Region_VQA		✓		
Vis7W		✓		
Ask_Neuron	✓	✓	✓	✓
SCMC				✓
HAN				✓
StrSem		✓		
AVQAN				✓
CMF				✓
EnsAtt				✓
MetaVQA				✓
DA-NTN				✓
QGHC				✓
QTA				✓
WRAN				✓
QAR				✓

رویکرد یادگیری عمیق: استخراج ویژگی از تصویر

مدل پرسش و پاسخ تصویری	AlexNet	VGGNet	GoogleNet	ResNet
Image_QA		✓		
Talk_to_Machine			✓	
VQA		✓		
Vis_Madlibs	✓	✓		
VIS + LSTM		✓		
Ahab		✓		
ABC-CNN		✓		
Comp_QA		✓		
DPPNet		✓		
Answer_CNN		✓		
VQA-Caption		✓		
Re_Baseline				✓
MCB				✓
SMem-VQA			✓	
Region_VQA		✓		
Vis7W		✓		
Ask_Neuron	✓	✓	✓	✓
SCMC				✓
HAN				✓
StrSem		✓		
AVQAN				✓
CMF				✓
EnsAtt				✓
MetaVQA				✓
DA-NTN				✓
QGHC				✓
QTA				✓
WRAN				✓
QAR				✓

VGGNet و **ResNet** به طور گسترده‌ای در سیستم‌های پرسش و پاسخ تصویری مورد استفاده قرار گرفته‌اند.

رویکرد یادگیری عمیق: استخراج ویژگی از تصویر

مدل پرسش و پاسخ تصویری	AlexNet	VGGNet	GoogleNet	ResNet
Image_QA		✓		
Talk_to_Machine			✓	
VQA		✓		
Vis_Madlibs	✓	✓		
VIS + LSTM		✓		
Ahab		✓		
ABC-CNN		✓		
Comp_QA		✓		
DPPNet		✓		
Answer_CNN		✓		
VQA-Caption		✓		
Re_Baseline				✓
MCB				✓
SMem-VQA			✓	
Region_VQA		✓		
Vis7W		✓		
Ask_Neuron	✓	✓	✓	✓
SCMC				✓
HAN				✓
StrSem		✓		
AVQAN				✓
CMF				✓
EnsAtt				✓
MetaVQA				✓
DA-NTN				✓
QGHC				✓
QTA				✓
WRAN				✓
QAR				✓

VGGNet و **ResNet** به طور گسترده‌ای در سیستم‌های پرسش و پاسخ تصویری مورد استفاده قرار گرفته‌اند.

یکی از دلایلی که محققان VGGNet را ترجیح می‌دهند این است که ویژگی‌هایی را استخراج می‌کند که **عمومیت بیشتری** دارد.

رویکرد یادگیری عمیق: استخراج ویژگی از تصویر

مدل پرسش و پاسخ تصویری	AlexNet	VGGNet	GoogleNet	ResNet
Image_QA		✓		
Talk_to_Machine			✓	
VQA		✓		
Vis_Madlibs	✓	✓		
VIS + LSTM		✓		
Ahab		✓		
ABC-CNN		✓		
Comp_QA		✓		
DPPNet		✓		
Answer_CNN		✓		
VQA-Caption		✓		
Re_Baseline				✓
MCB				✓
SMem-VQA			✓	
Region_VQA		✓		
Vis7W		✓		
Ask_Neuron	✓	✓	✓	✓
SCMC				✓
HAN				✓
StrSem		✓		
AVQAN				✓
CMF				✓
EnsAtt				✓
MetaVQA				✓
DA-NTN				✓
QGHC				✓
QTA				✓
WRAN				✓
QAR				✓

VGGNet و **ResNet** به طور گسترده‌ای در سیستم‌های پرسش و پاسخ تصویری مورد استفاده قرار گرفته‌اند.

یکی از دلایلی که محققان VGGNet را ترجیح می‌دهند این است که ویژگی‌هایی را استخراج می‌کند که **عمومیت بیشتری** دارد.

نکته‌ی قابل توجه دیگر در جدول، روند مهاجرت از VGGNet به ResNet در مقالات اخیر است. زیرا در سال‌های اخیر، **منابع محاسباتی** کافی با **هزینه مناسب** در دسترس محققان می‌باشد.

رویکرد یادگیری عمیق: استخراج ویژگی از سوال

مدل پرسش و پاسخ تصویری	one-hot	CBOW	Skip-gram/Word2vec	GloVe	CNN	LSTM	GRU
Image_QA			✓				
Talk_to_Machine						✓	
VQA		✓					
Vis_Madlibs			✓				
VIS + LSTM						✓	
ABC-CNN						✓	
Comp_QA						✓	
DPPNet							✓
Answer_CNN					✓		
VQA-Caption						✓	
Re_Baseline			✓				
MCB						✓	
SMem-VQA		✓					
Region_VQA			✓				
Vis7W	✓						
Ask_Neuron		✓			✓	✓	✓
SCMC					✓		
HAN						✓	
StrSem						✓	
AVQAN	✓						
CMF				✓		✓	
EnsAtt				✓			
MetaVQA				✓			✓
DA-NTN							✓
QGHC							✓
WRAN							✓
QAR				✓			

رویکرد یادگیری عمیق: استخراج ویژگی از سوال

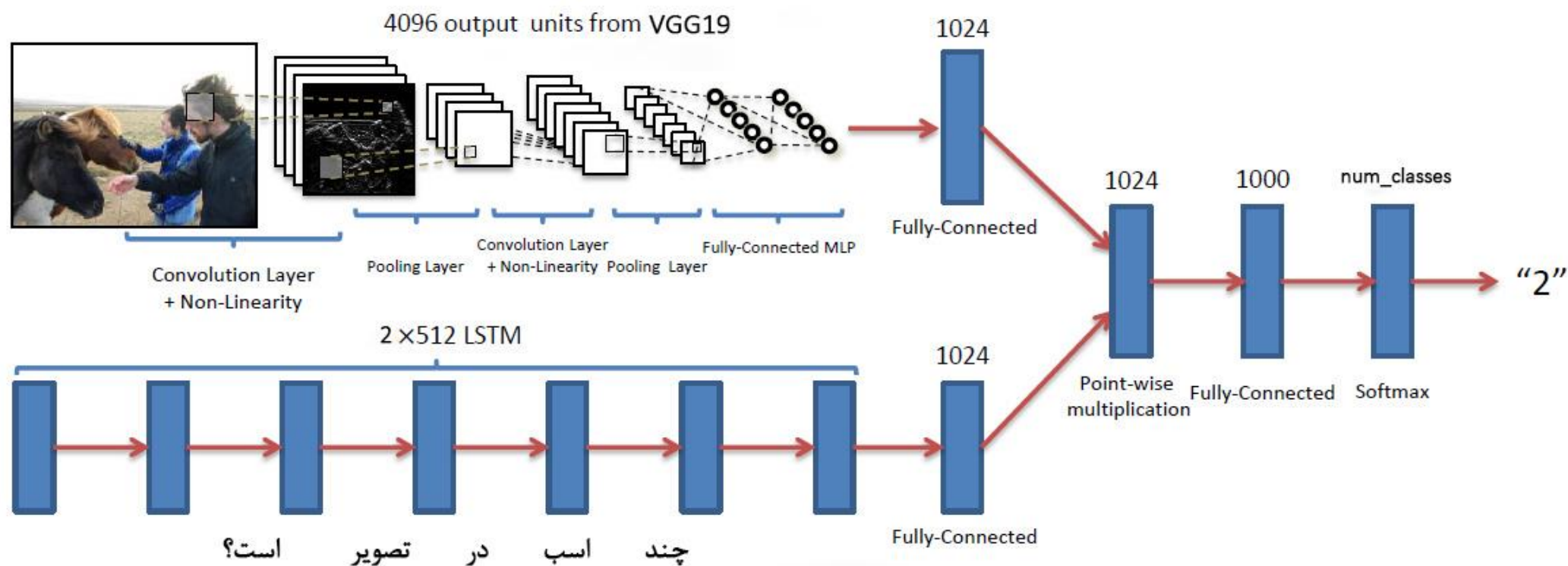
مدل پرسش و پاسخ تصویری	one-hot	CBOW	Skip-gram/Word2vec	GloVe	CNN	LSTM	GRU
Image_QA			✓				
Talk_to_Machine						✓	
VQA		✓					
Vis_Madlibs			✓				
VIS + LSTM						✓	
ABC-CNN						✓	
Comp_QA						✓	
DPPNet							✓
Answer_CNN					✓		
VQA-Caption						✓	
Re_Baseline			✓				
MCB						✓	
SMem-VQA		✓					
Region_VQA			✓				
Vis7W	✓						
Ask_Neuron		✓			✓	✓	✓
SCMC					✓		
HAN						✓	
StrSem						✓	
AVQAN	✓						
CMF				✓		✓	
EnsAtt				✓			
MetaVQA				✓			✓
DA-NTN							✓
QGHC							✓
WRAN							✓
QAR				✓			

محققان حوزه‌ی پرسش و پاسخ تصویری ترجیح می‌دهند؛ برای استخراج ویژگی از متن و بازنمایی آن از **LSTM** استفاده کنند. آن‌ها معتقد هستند که **RNN**ها عملکرد بهتری نسبت به روش‌های مستقل از دنباله‌ی کلمات دارند.

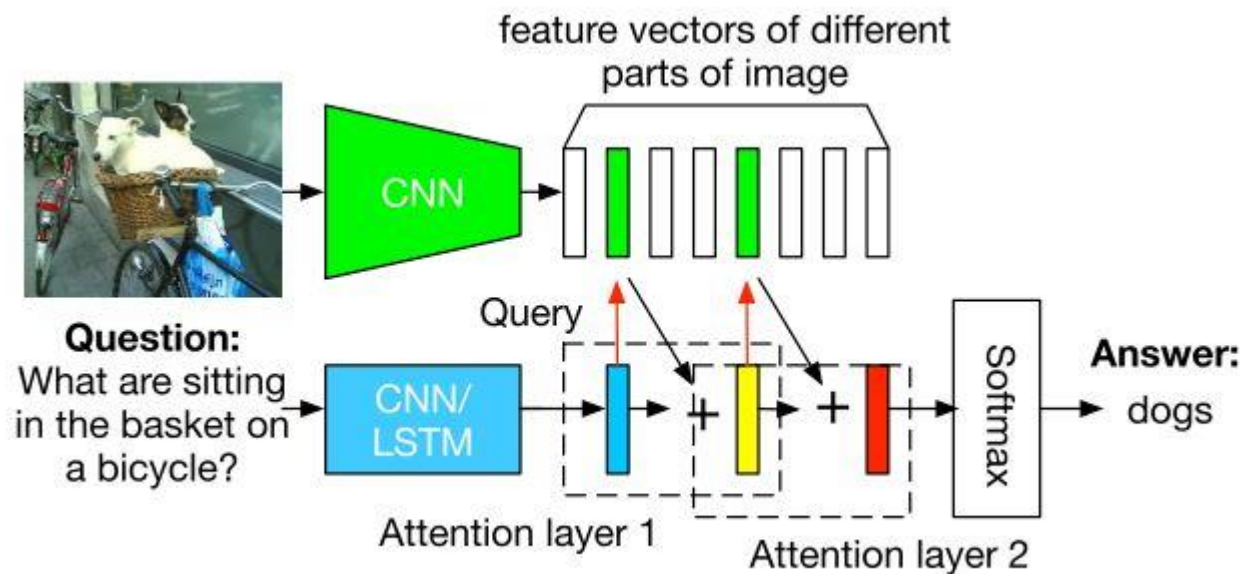
بازنمایی مشترک تصویر و سوال

- برای بدست آمدن بازنمایی مشترک بین تصویر و سوال از **روش‌های ساده** مانند ضرب ویژگی‌ها تا **روش‌های پیچیده‌تر** مانند مکانیزم توجه استفاده می‌شود.

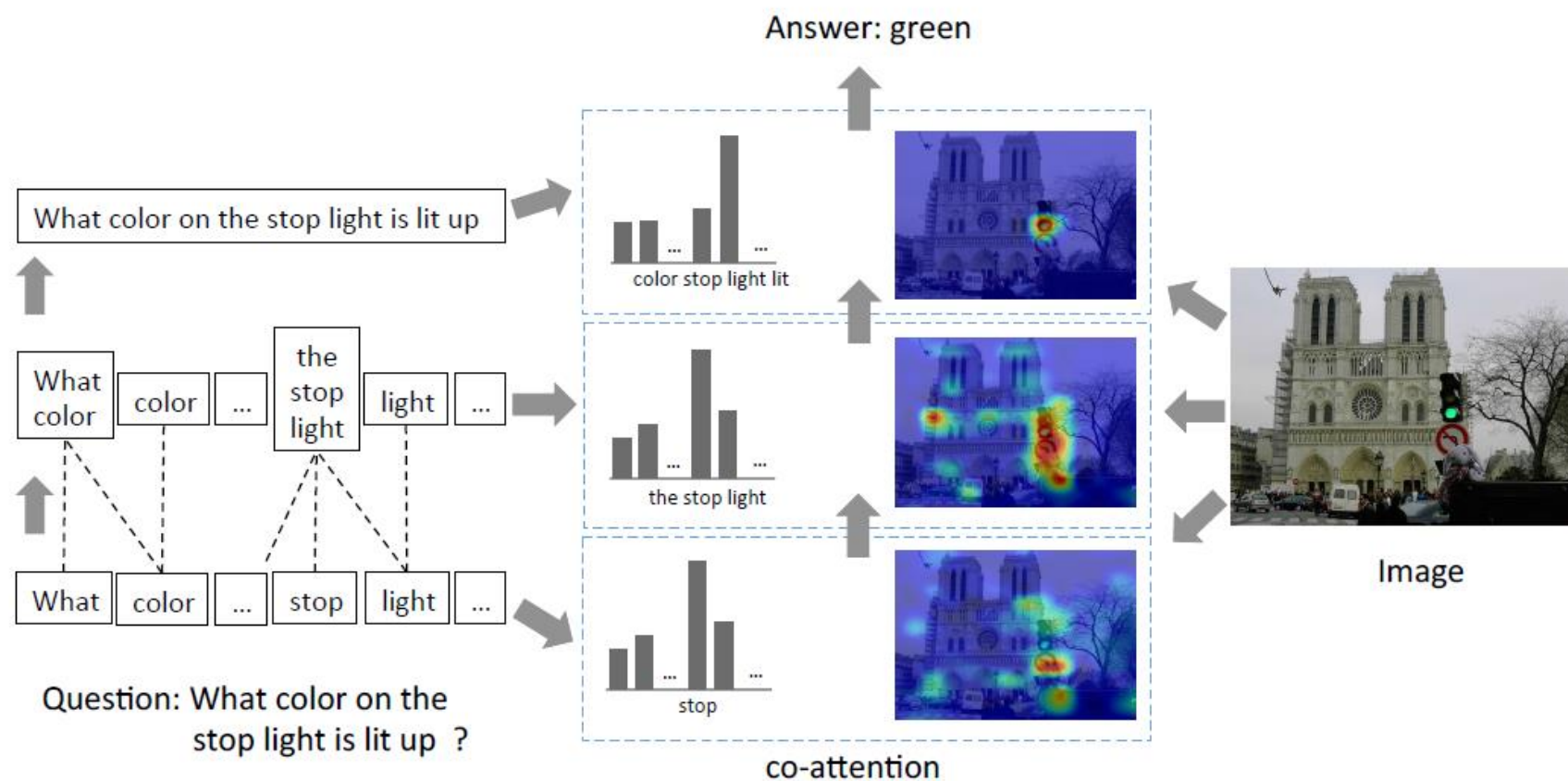
بازنمایی مشترک تصویر و سوال: LSTM Q + norm l



بازنمایی مشترک تصویر و سوال: SAN



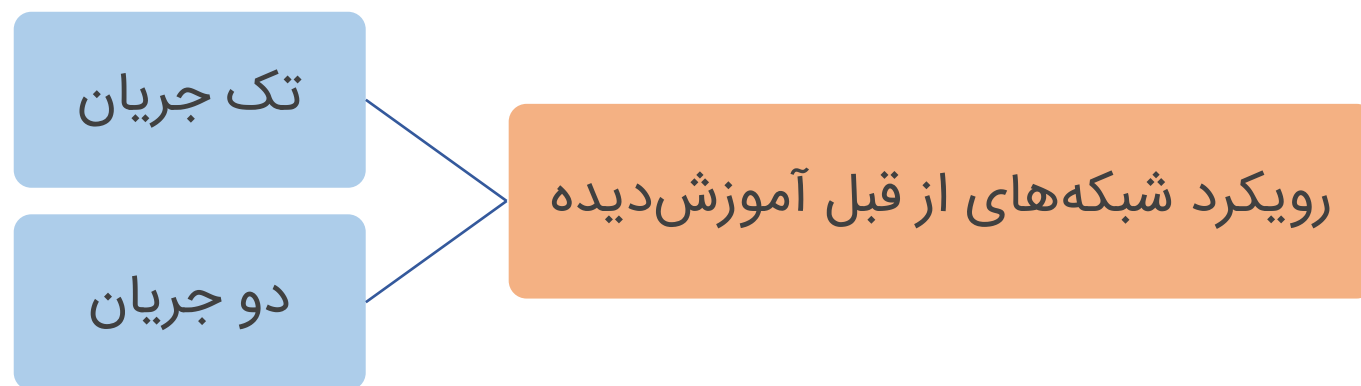
بازنمایی مشترک تصویر و سوال: HieCoAttention



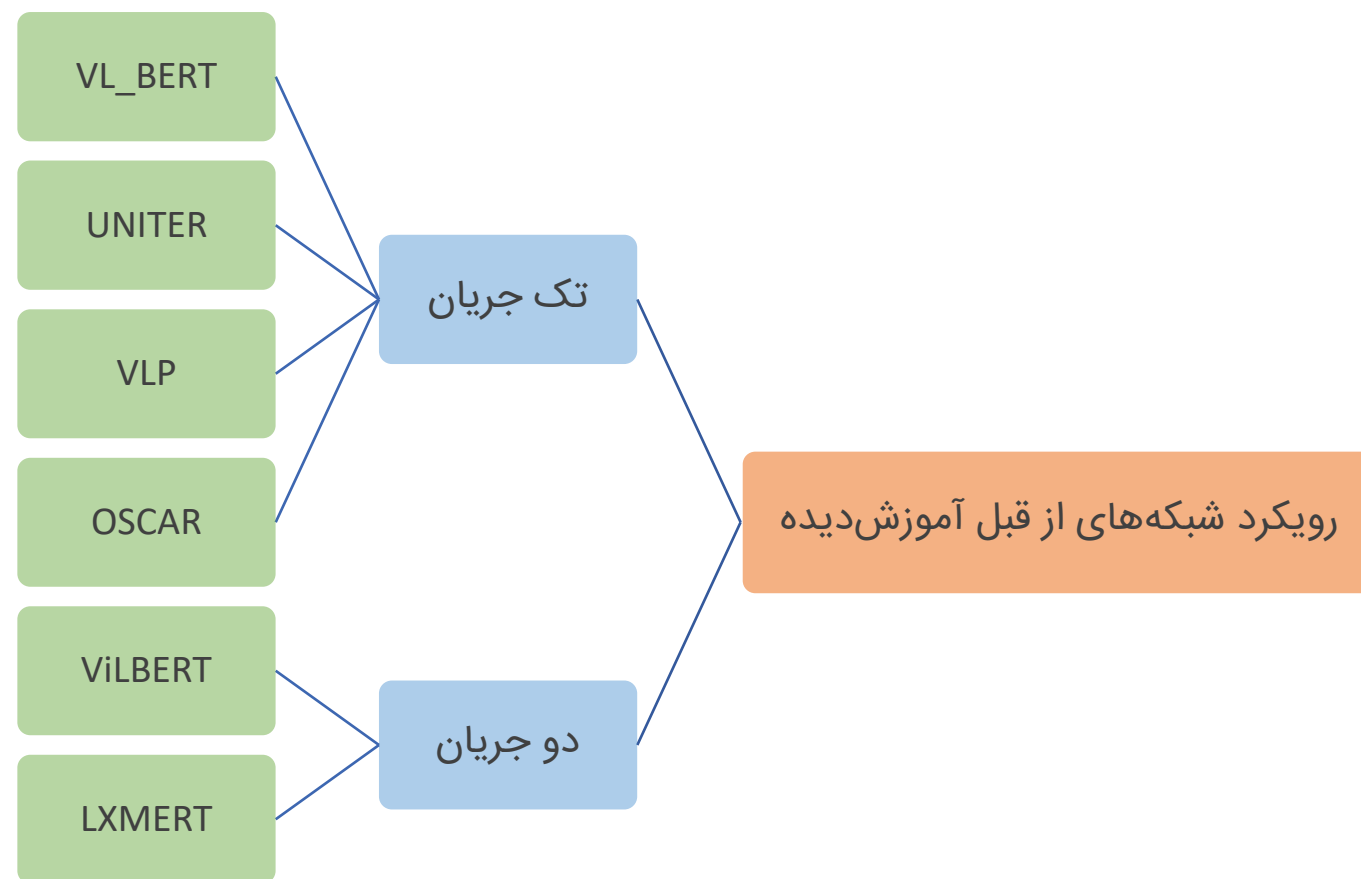
پیش‌بینی پاسخ

تولید	طبقه‌بندی	مدل پرسش و پاسخ تصویری
✓		Talk_to_Machine
✓	✓	VQA
	✓	HieCoAttention
	✓	MCB
✓	✓	Ask_Neuron
	✓	Mutan
	✓	MCAN
	✓	AnswerAll

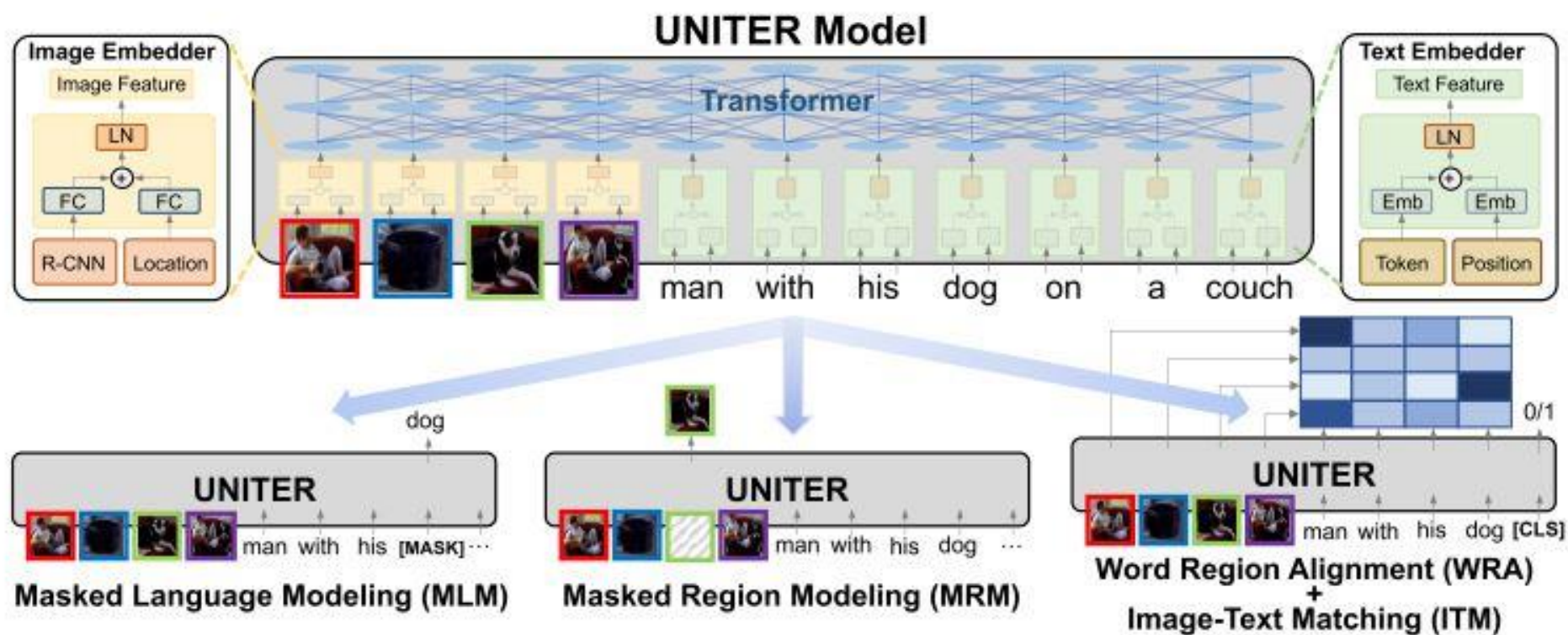
شبکه‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر



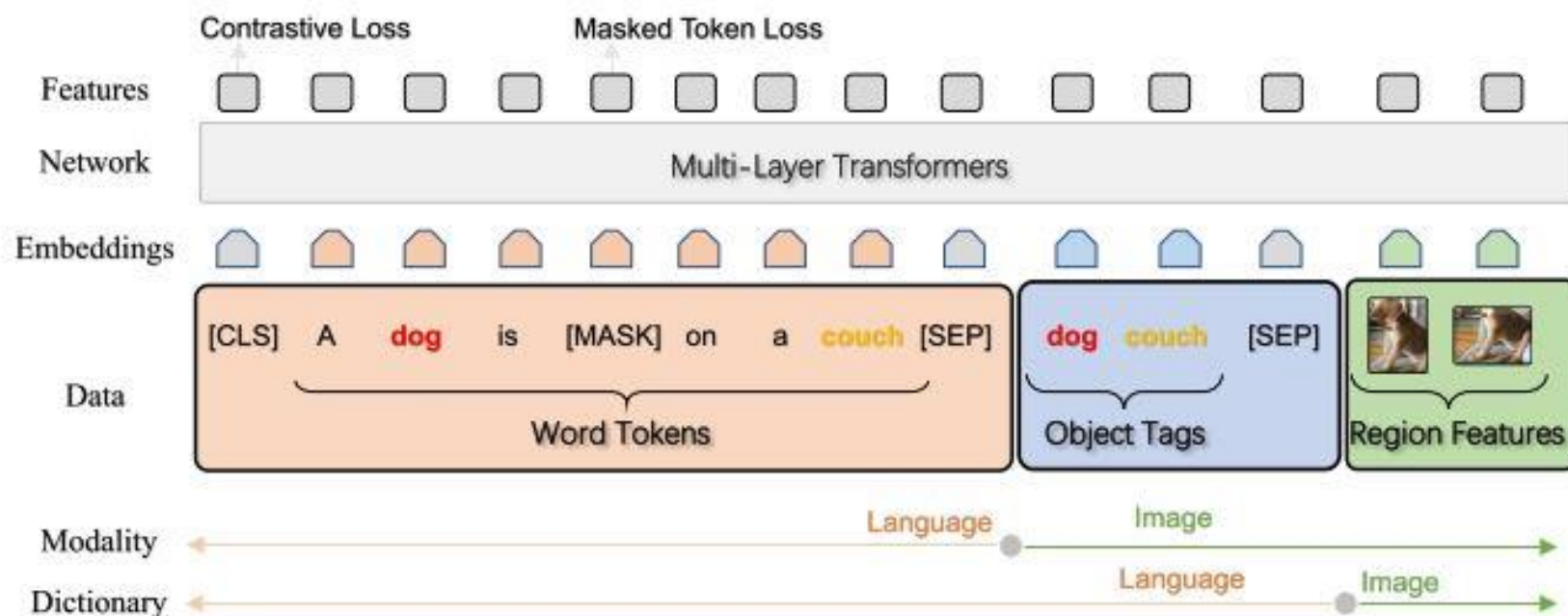
شبکه‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر



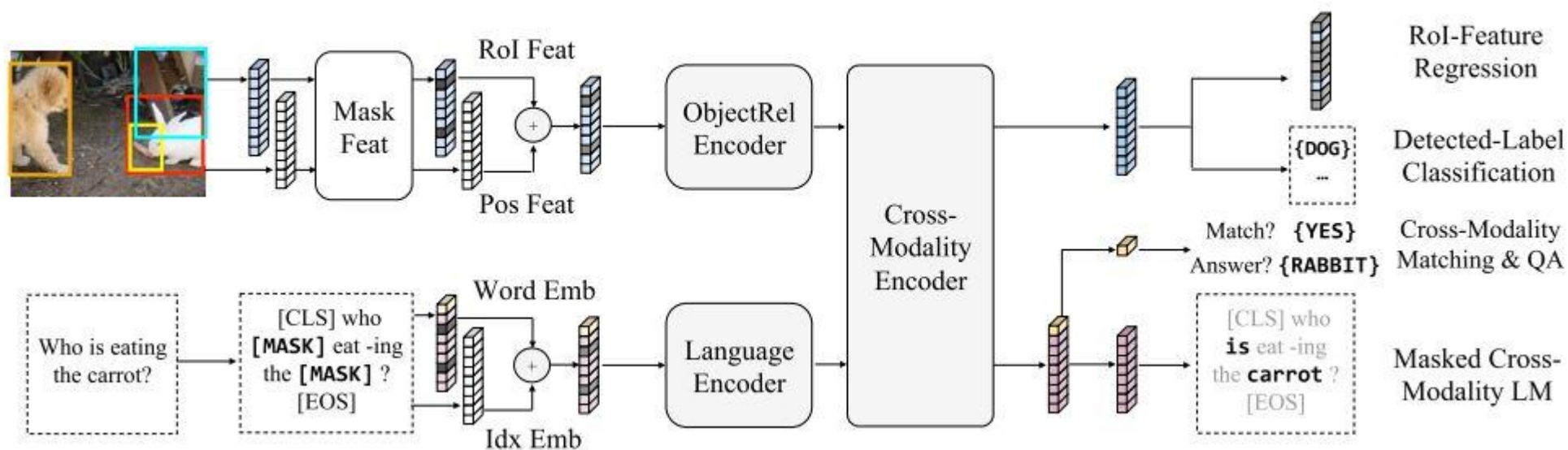
شبکه‌های از قبل آموزش دیده UNITER



شبکه‌های از قبل آموزش دیده OSCAR



شبکه‌های از قبل آموزش دیده LXMERT



مقایسه بین شبکه‌های از قبل آموزش‌دیده بر روی زبان طبیعی و تصویر

روش	معماری	ورودی	مجموعه‌دادگان استفاده شده برای آموزش	توابع هدف
VL-BERT	تک جریان	کلمات جمله + ROI های تصویر	Conceptual Captions + BooksCorpus + English Wikipedia	text-based MLM + visual-based MLM
UNITER	تک جریان	کلمات جمله + ROI های تصویر	COCO + Visual Genome + Conceptual Captions + SBU Captions	text-based MLM + visual-based MLM + Image-Text Matching + Word-Region Alignment
VLP	تک جریان	کلمات جمله + ROI های تصویر	Conceptual Captions	bidirectional + seq2seq
OSCAR	تک جریان	کلمات جمله + ROI های تصویر + برچسب اشیا	COCO + Conceptual Captions + SBU captions + flicker30 + GQA	Masked Token Loss + Contrastive Loss
ViL-BERT	دو جریان	کلمات جمله + ROI های تصویر	Conceptual Captions	text-based MLM + visual-based MLM + Image-Text Matching
LXMERT	دو جریان	کلمات جمله + ROI های تصویر	MS COCO + Visual Genome + VQA v2.0 + GQA balanced version + VG-QA	text-based MLM + visual-based MLM + Image-Text Matching + Image Question Answering

دقت شبکه‌های از قبل آموزش‌دیده بر روی دادگان VQA v2.0 (test-std)

روش	سوالات بله/خیر	سوالات شمارشی	سایر سوالات	دقت کل
VLP	۸۷/۴	۵۲/۱	۶۰/۵	۷۰/۷
ViL-BERT	–	–	–	۷۰/۹۲
VL-BERT	–	–	–	۷۲/۲۲
LXMERT	۸۸/۲	۵۴/۲	۶۳/۱	۷۲/۵
OSCAR	–	–	–	۷۳/۸۲
UNITER	–	–	–	۷۴/۰.۲

دقت شبکه‌های از قبل آموزش‌دیده بر روی دادگان VQA v2.0 (test-std)

دقت کل	سایر سوالات	سوالات شمارشی	سوالات بله/خیر	روش
۷۰/۷	۶۰/۵	۵۲/۱	۸۷/۴	VLP
۷۰/۹۲	–	–	–	ViL-BERT
۷۲/۲۲	–	–	–	VL-BERT
۷۲/۵	۶۳/۱	۵۴/۲	۸۸/۲	LXMERT
۷۳/۸۲	–	–	–	OSCAR
۷۴/۰.۲	–	–	–	UNITER

مدل‌های **تک جریان** نتایج بهتری نسبت به مدل‌های **دو جریان** بدست آوردند در حالی که **تعداد پارامترهای** مدل‌های تک جریان نسبت به مدل‌های دو جریان **کمتر** است.

معیارهای ارزیابی مسئله پرسش و پاسخ تصویری

- معیار دقت
- معیار شباهت Wu-Palmer
- معیار اجماع
- معیار MPT
- معیار BLEU
- معیار METEOR

کارهای آینده

- فهمیدن **روند درک مدل‌هایی فعلی از زبان و تصویر** به منظور پیشنهاد مدلی برای پاسخ به سوالاتی که نیاز به **استدلال طولانی** دارند.

کارهای آینده

- فهمیدن **روند درک مدل‌هایی فعلی از زبان و تصویر** به منظور پیشنهاد مدلی برای پاسخ به سوالاتی که نیاز به **استدلال طولانی** دارند.
- استفاده از **ترنسفرمرها با چندین لایه رمزگذار و رمزگشا** با هدف **تولید پاسخ** نه طبقه‌بندی پاسخ.

کارهای آینده

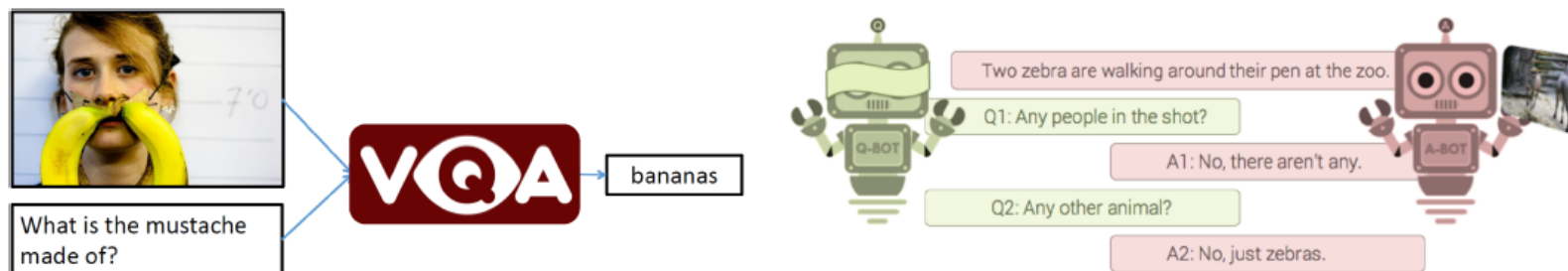
- فهمیدن **روند درک مدل‌هایی فعلی از زبان و تصویر** به منظور پیشنهاد مدلی برای پاسخ به سوالاتی که نیاز به **استدلال طولانی** دارند.
- استفاده از **ترنسفرمرها با چندین لایه رمزگذار و رمزگشا** با هدف **تولید پاسخ** نه طبقه‌بندی پاسخ.
- تهیه و جمع‌آوری مجموعه‌دادگان متناسب با واقعیت و **کاربردهای عملی** و بدون **بایاس**.

کارهای آینده

- فهمیدن **روند درک مدل‌هایی فعلی از زبان و تصویر** به منظور پیشنهاد مدلی برای پاسخ به سوالاتی که نیاز به **استدلال طولانی** دارند.
- استفاده از **ترنسفرمرها با چندین لایه رمزگذار و رمزگشا** با هدف **تولید پاسخ** نه طبقه‌بندی پاسخ.
- تهیه و جمع‌آوری مجموعه‌داده‌گان متناسب با واقعیت و **کاربردهای عملی** و بدون **بایاس**.
- استفاده از **ترنسفرمرهای بهبودیافته** در معماری شبکه‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر.

کارهای آینده

- فهمیدن **روند درک مدل‌هایی فعلی از زبان و تصویر** به منظور پیشنهاد مدلی برای پاسخ به سوالاتی که نیاز به **استدلال طولانی** دارند.
- استفاده از **ترنسفرمرها با چندین لایه رمزگذار و رمزگشا** با هدف **تولید پاسخ** نه طبقه‌بندی پاسخ.
- تهیه و جمع‌آوری مجموعه‌دادگان متناسب با واقعیت و **کاربردهای عملی** و بدون **بایاس**.
- استفاده از **ترنسفرمرهای بهبودیافته** در معماری شبکه‌های از قبل آموزش دیده بر روی زبان طبیعی و تصویر.
- **تهیه و جمع‌آوری دادگان فارسی** برای مسئله پرسش و پاسخ تصویری و آموزش یک مدل کارآمد براساس آن.



با تشکر از توجه شما