Tutorial at IJCAI, August 19th /20th 2021

Neural Machine Reasoning

Lecture 7+8+9: Applications

Truyen Tran, Vuong Le, Hung Le and Thao Le
{truyen.tran,vuong.le,thai.le,thao.le}@deakin.edu.au

https://neuralreasoning.github.io

Introduction

QA as Standardized Tests for Machine Reasoning

- **Question Answering** = computer systems that **automatically** answer **natural language questions** about knowledge by humans.
  - Not simple search-and-retrieve.
    - E.g. "what affects her mobility?"
- **Why question answering?**
  - Humans learn by answering questions.
  - QA can be used to formulate other tasks
    - E.g. "what is present in the image?" (recognition), "what action has the person in the video performed?"

Q: "What affects her mobility?"

Learning to Reason in QA form

- Input:
  - A knowledge context C
  - A query q
- Output: an answer satisfying

$$\tilde{a} = \underset{a \in \mathbb{A}}{\mathrm{argmax}} \mathcal{P}_\theta(a \mid q, C)$$

- C can be
  - structured: knowledge graphs
  - unstructured: text, image, sound, video

Q: "What affects her mobility?"

Q: Is it simply an optimization problem like recognition, detection or even translation?

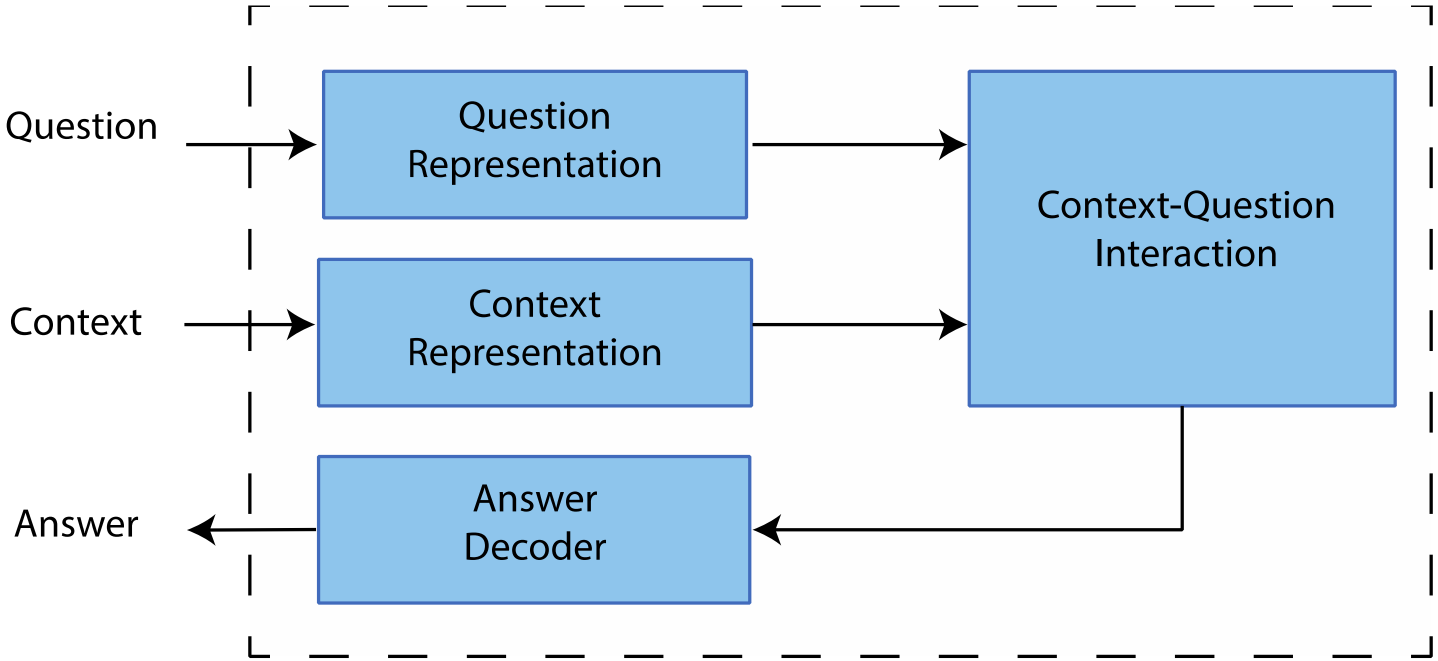
→ No, because the logics from C, q into a is more complex than other solved optimization problems

→ We can solve (some parts of) it with good structures and inference strategies

[Slide credit: Vuong Le]

Overall Architecture of General QA

- Question representation
- Context representation (domain specific)
- Context-question interaction (reasoning)
- Answer decoder



Lecture 7: Textual QA

Tasks: Standford Question Answering Dataset (SQuAD)

Text passage

Private schools, also known as independent schools, non-governmental, or nonstate schools, [...]; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, [...].

QA

Q: Rather than taxation, what are private schools largely funded by?

A: charging their students tuition

Size: 151,054 samples
Task: given context information as a paragraph, predict the text span contains the correct answer.

Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." *ACL'18.*

Other Tasks

Cloze

P: You will need 3/4 cup of black berries ... Pour the mixture into cups and insert a popsicle stick in it or pour it in a posicle maker. Place the cup ... in the freezer. ...

Q: Choose the best title for the missing blank to correctly complete the recipe. *Ingredients,__ , Freeze, Enjoying*

Candidates: (A) Cereal Milk Ice Cream (B) Ingredients (C) Pouring (D) Oven
Answer: C

Multiple choice

P: It was Jessie Bear's birthday. She ...

Q: Who was having a birthday?

Candidates: (A) Jessie Bear (B) no one (C) Lion (D) Tiger
Answer: A

Open-ended

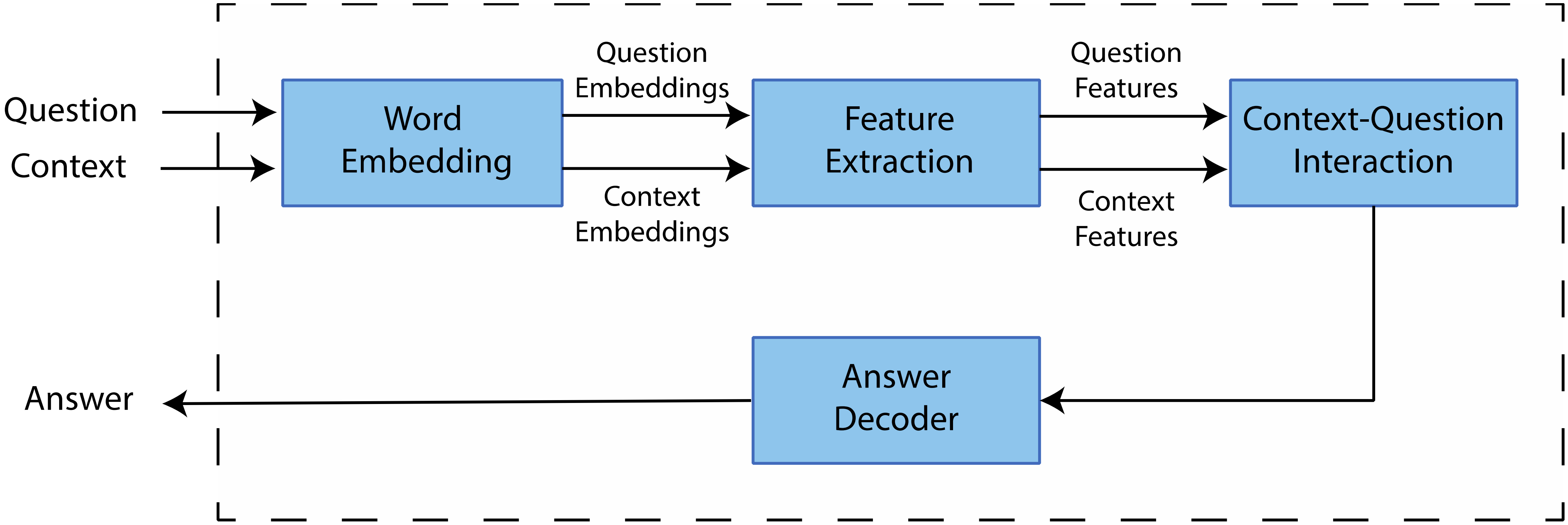
P: ...Mark decides to broadcast his final message as himself. They finally drive up to the crowd of protesting students, ... The police step in and arrest Mark and Nora....

Q: What are the students doing when Mark and Nora drive up?

Answer: Protesting

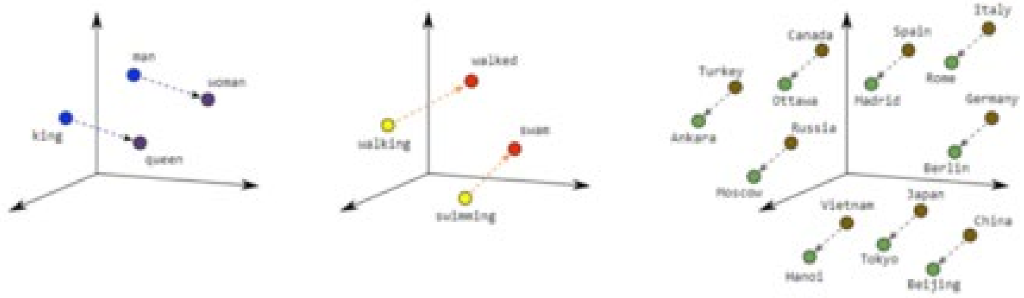
Dzendzik, Daria, Carl Vogel, and Jennifer Foster. "English Machine Reading Comprehension Datasets: A Survey." *2021.*

Machine Comprehension Test



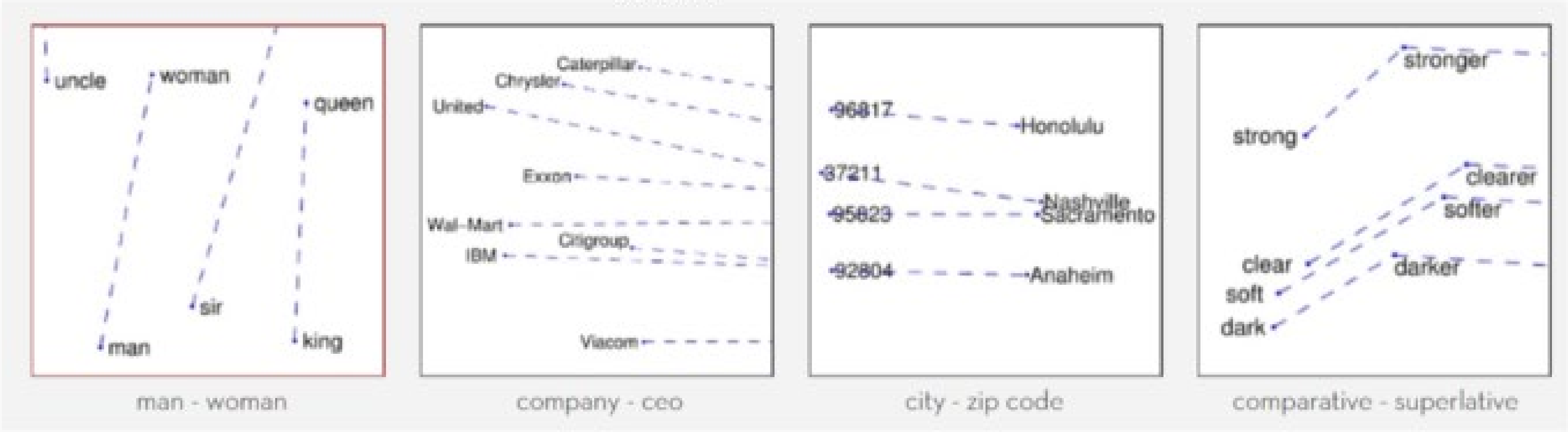
Word Representations

Word2Vec



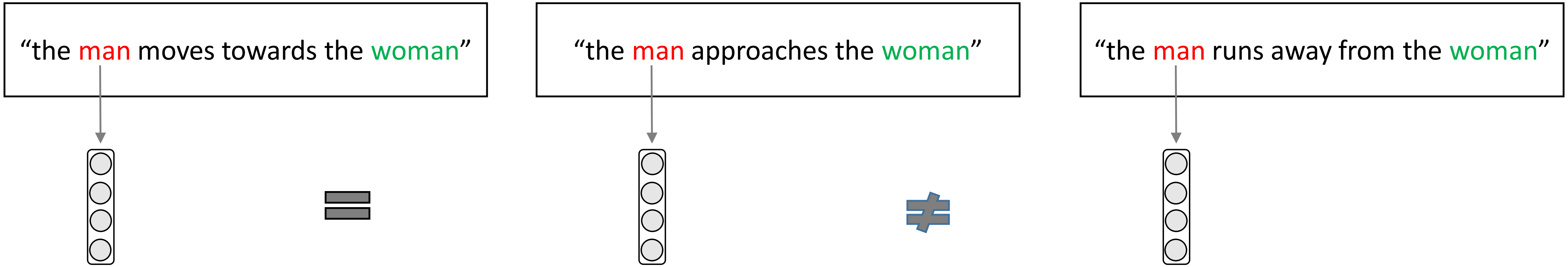
Male-Female Verb Tense Country-Capital

GloVe



[Word Embeddings for NLP. Understanding word embeddings and their… | by Renu Khandelwal | Towards Data Science]

Word Representations in Context

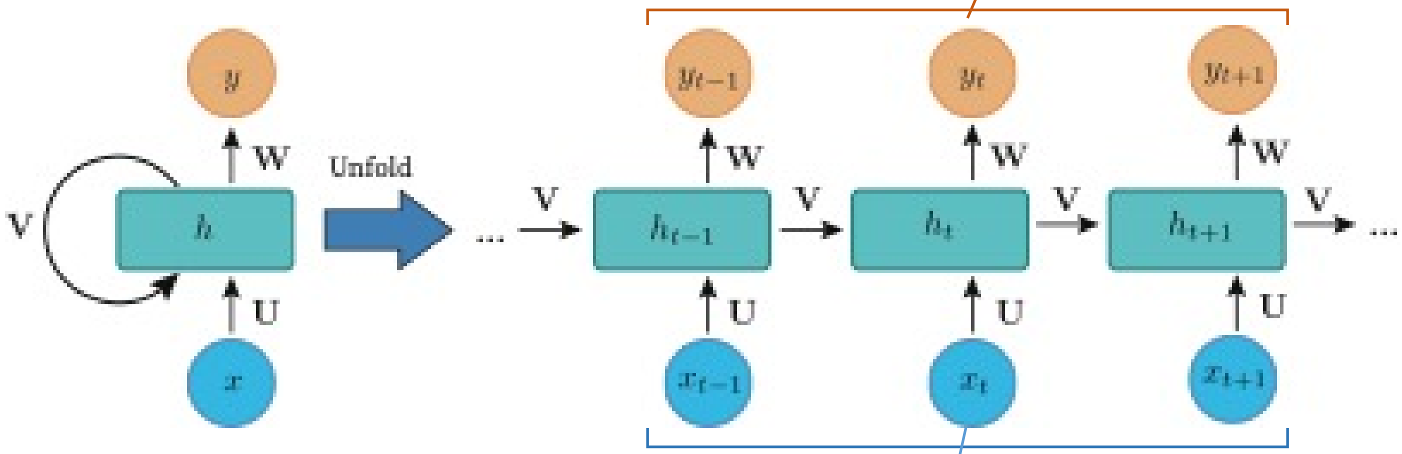


- Word representations should vary depending on context

Recurrent Neural Networks

- Variants: LSTM, GRU
- Advantages:
  - Good for sentences/short text
  - Robust in practice
- Disadvantages:
  - Slow, computational costly
  - Cannot parallelize
  - Not good for very long sequences.

Context-dependent embeddings

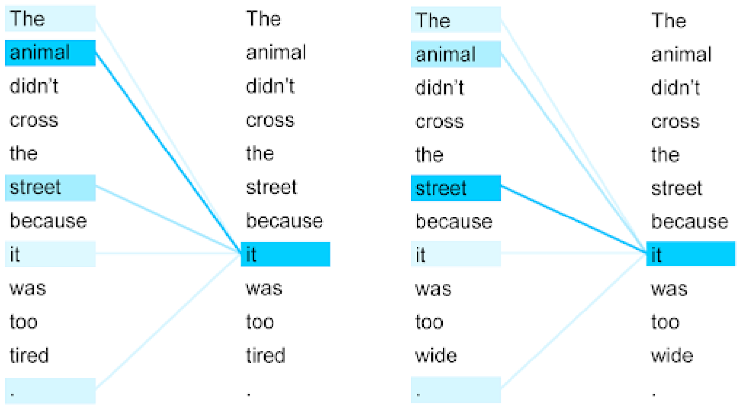


$$y_t = f(x_t, y_{t-1})$$

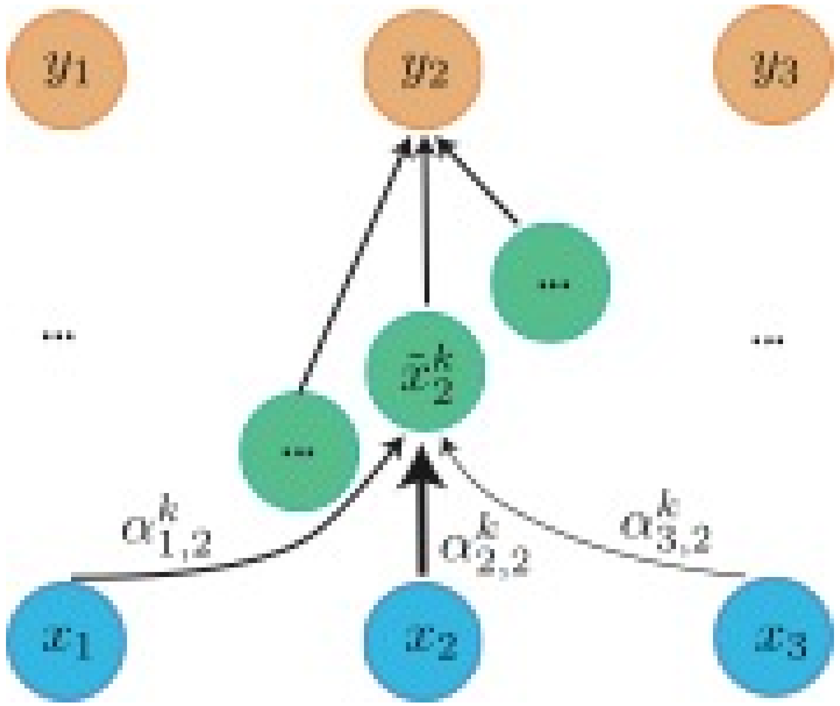
Context-independent embeddings

Self-Attention

- Advantages:
  - Good at capturing long range dependencies
  - Can capture co-reference chains
  - Parallelizable and fast
- Disadvantages:
  - Memory intensive
  - Hyper-parameters tuning



Transformer Self-Attention Coreference Visualization
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html



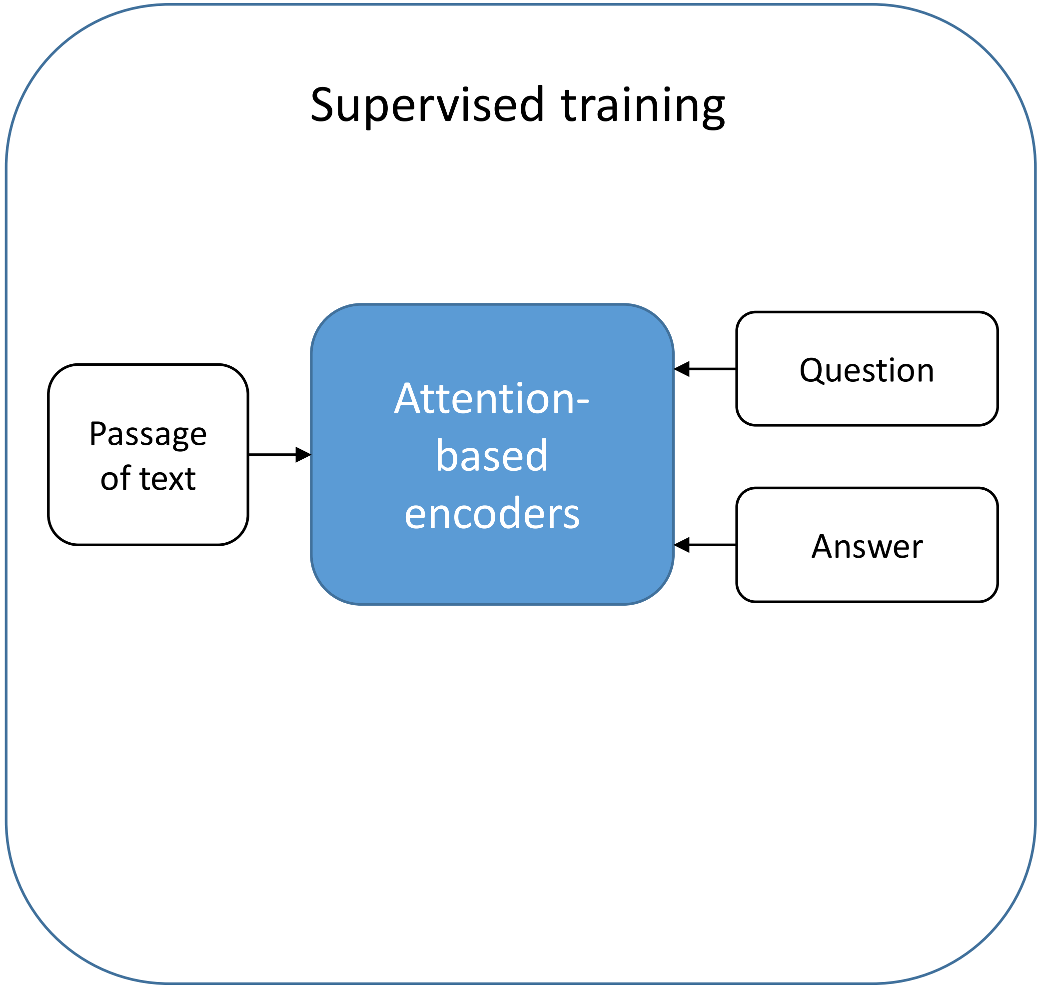
- K: attention heads
- T: sequence length
- $\alpha_{j,t}$: self-attention weights

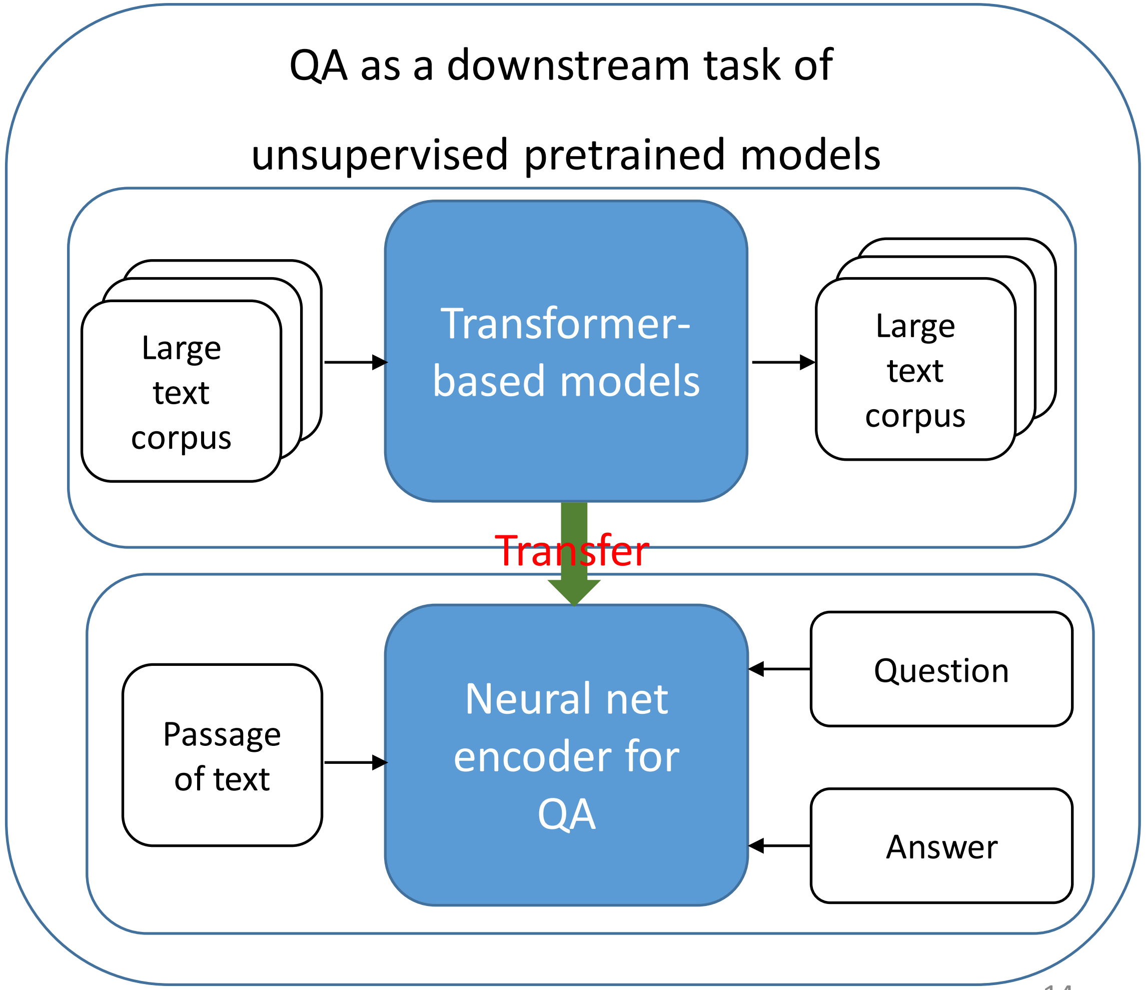
$$\tilde{x}_t^k = \sum_{j=1}^{T} \alpha_{j,t}^k x_j$$

$$y_t = f(\tilde{x}_t^1, \ldots, \tilde{x}_t^K)$$

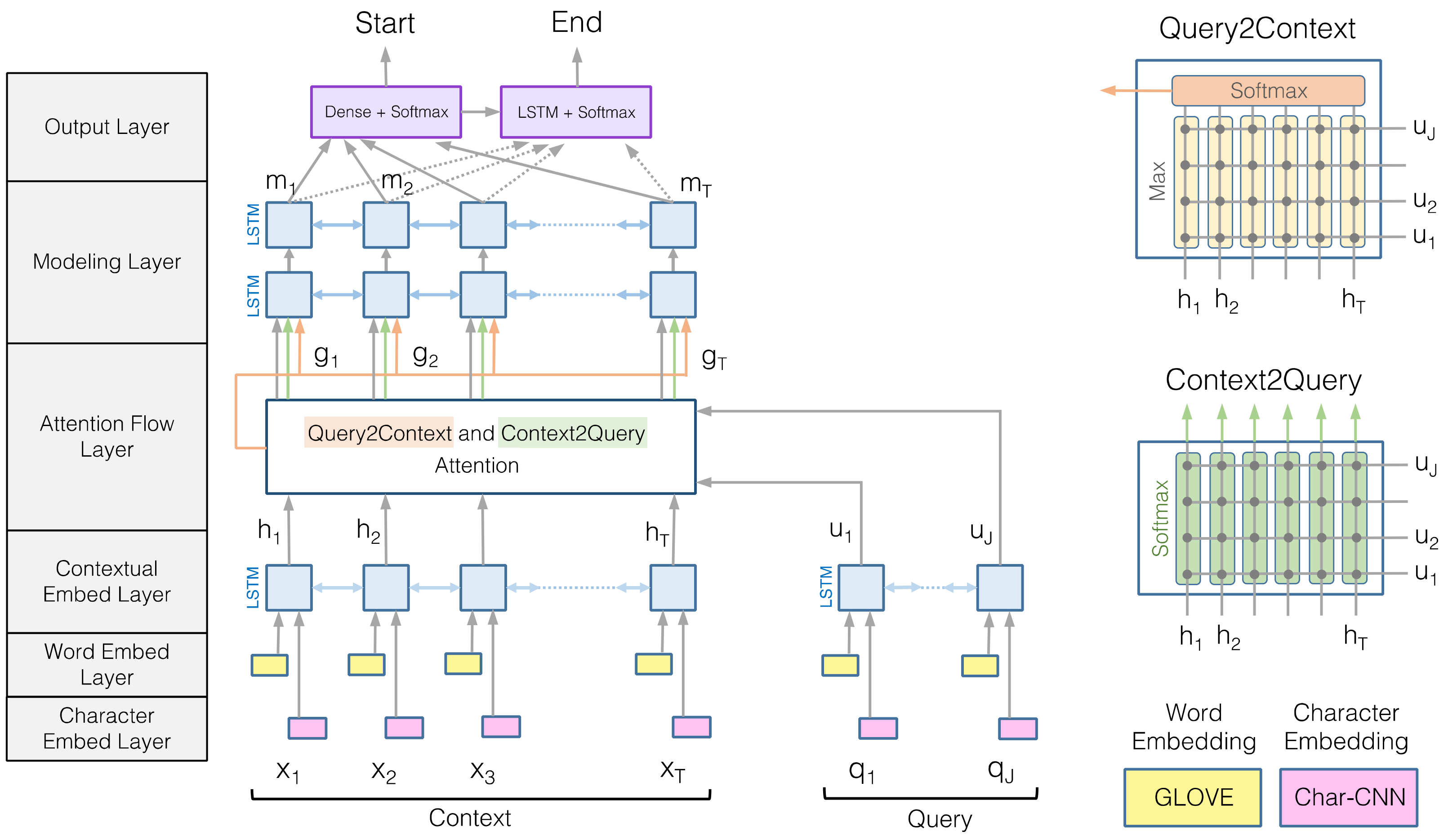
[Slide credit: Pierre-Emmanuel Mazaré]

Reasoning Approaches in MRC





Bi-Directional Attention Flow Model

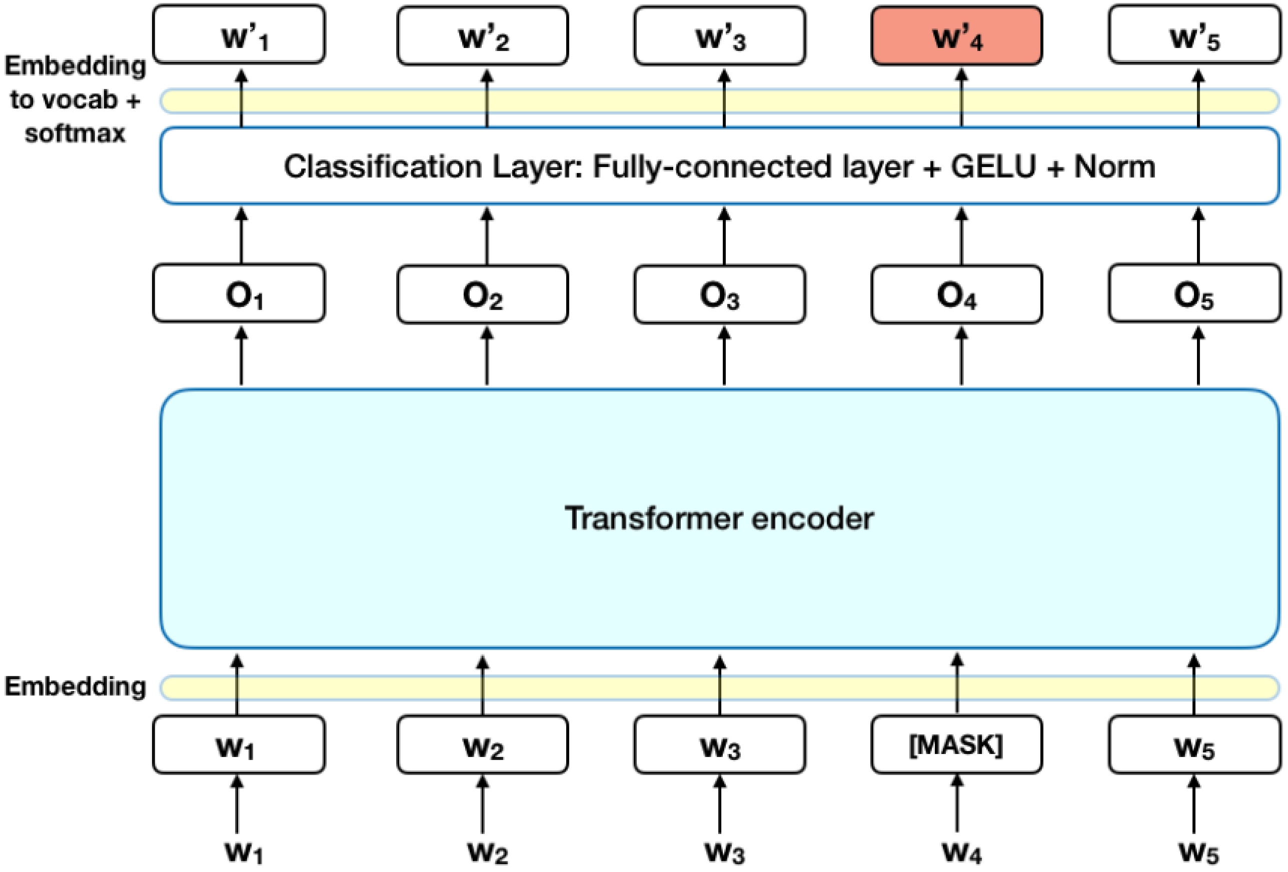


Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." *ICLR'17.*

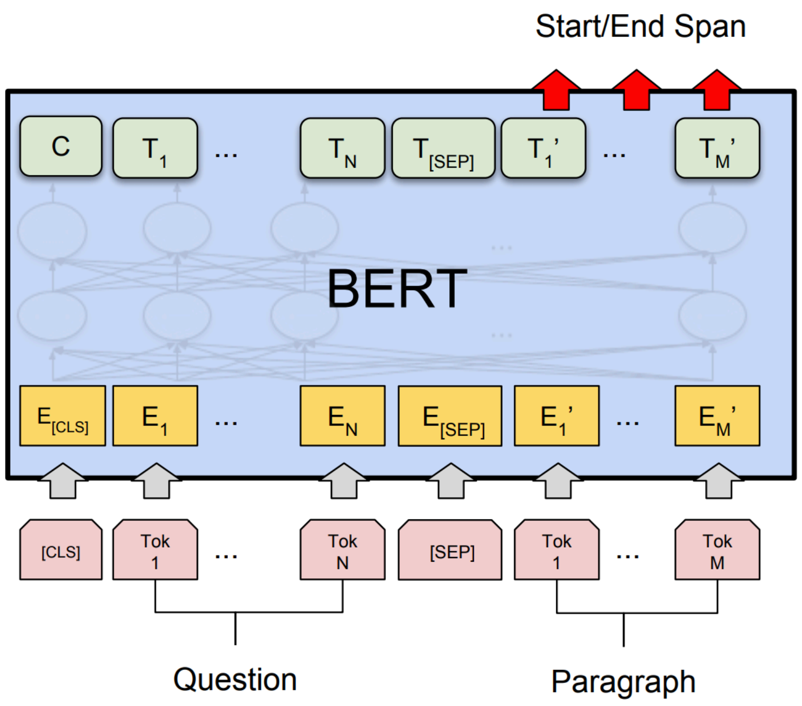
BERT: Transformer That Predicts Its Own Masked Parts

BERT is like parallel approximate pseudo-likelihood

- ~ Maximizing the conditional likelihood of some variables given the rest.
- When the number of variables is large, this converses to MLE (maximum likelihood estimate).



BERT – Fine-tuning for MRC



(c) Question Answering Tasks:
SQuAD v1.1

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

[Slide credit: Pierre-Emmanuel Mazaré]

Figures from Devlin et al. 18'

Lecture 8: Image Question Answering

Recall the Learning to Reason formulation

- Input:
  - Context C given by an **image**
  - A query q
- Output: an answer satisfying

$$\tilde{a} = \underset{a \in \mathbb{A}}{\operatorname{argmax}} \mathcal{P}_\theta(a \mid q, C)$$

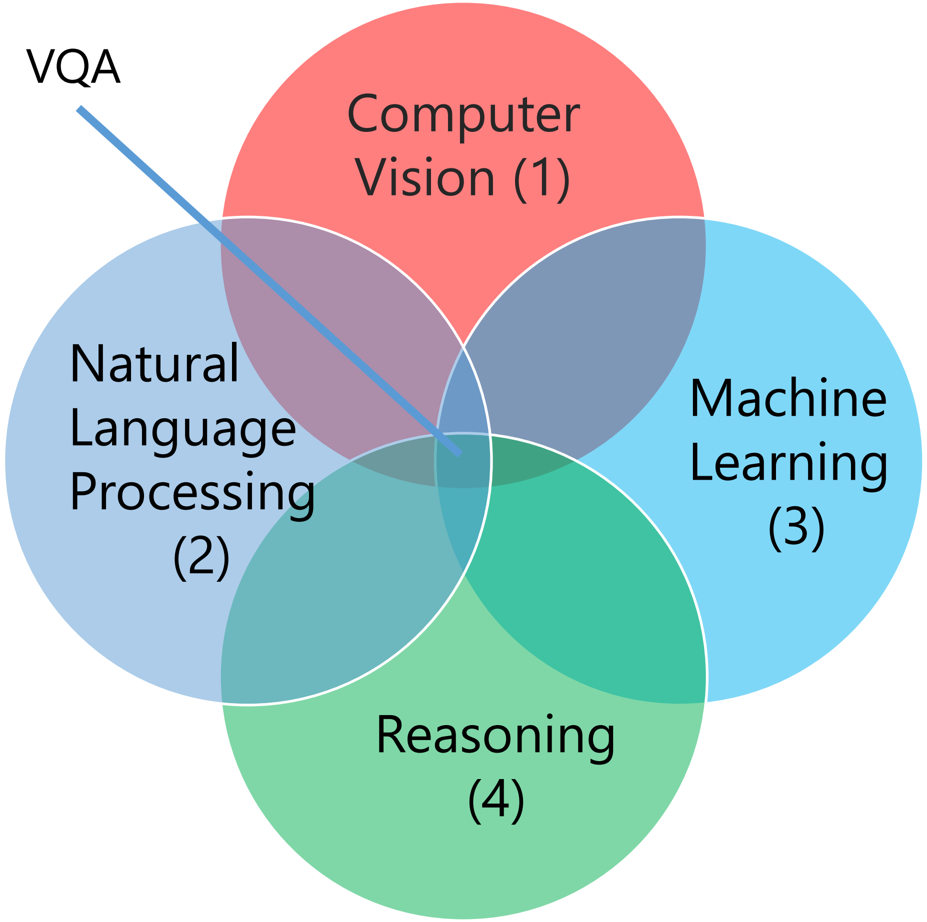
Q: "What affects her mobility?"

Why VQA is an AI testbed?

Question: What can the red object on the ground be used for ? (2)

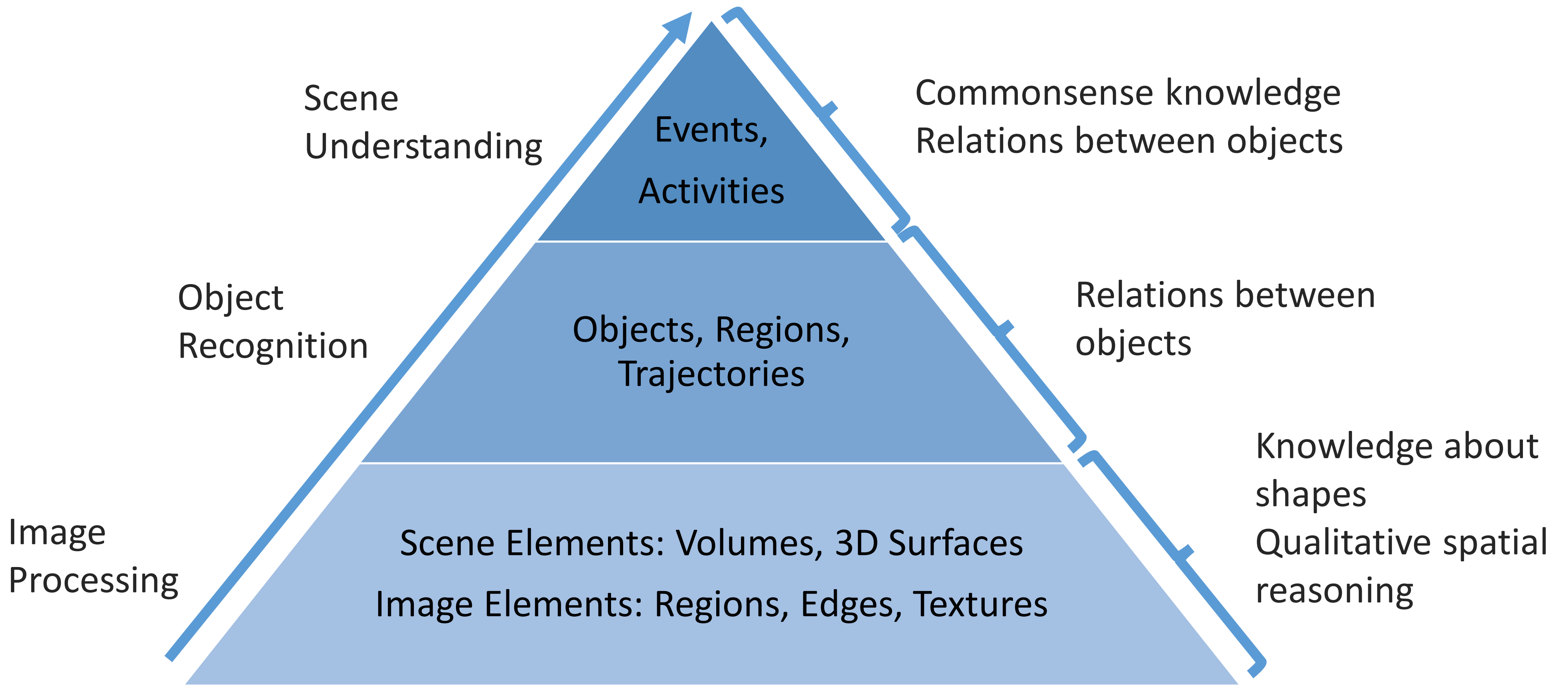
Answer: Firefighting

Support Fact: Fire hydrant can be used for fighting fires. (2, 4)



Wang, Peng, et al. "Fvqa: Fact-based visual question answering." TPAMI 2018

Why VQA is an AI testbed?



Adapted from [Somak et al., 2019]

Applications of VQA

- Aid visually-impaired users

Image credit: ARIA

Applications of VQA

- Surveillance and visual data summarization

What did the man in red shirt do before entering the building?

Image credit: journalistsresource.org

VQA: Question types

Open-ended

- Is this a vegetarian pizza?
- What is the red thing in the photo?

Multi-choice

(Q) What is the red thing in the photo?

(A) (1) capsicum (2) beef (3) mushroom (4) cheese

Counting

- How many slices of pizza are there?

(VQA, Agrawal et al., 2015)

VQA: Image QA datasets

(VQA, Agrawal et al., 2015)

(Q) What is in the picture?
(Q) Is this a vegetarian pizza?

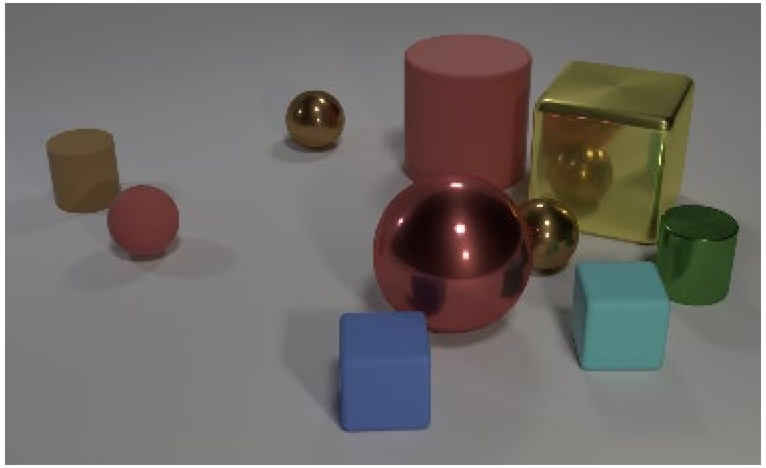
Perception

(GQA, Hudson et al., 2019)

(Q) What is the brown animal sitting inside of?
(Q) Is there a bag to the right of the green door?

Relational reasoning

(CLEVR, Johnson et al., 2017)



(Q) How many objects are either small cylinders or metal things?
(Q) Are there an equal number of large things and metal spheres?

Multi-step reasoning

Dual-view System for VQA

(Deep feature extraction, e.g. CNN, GloVe/BERT etc.)

Multiple

System 1: Intuitive

- Fast
- Implicit/automatic
- Pattern recognition
- Multiple

(Cross-modality interaction)

Single

System 2: Analytical

- Slow
- Deliberate/rational
- Careful analysis
- Single, sequential

Attention-based VQA Methods

- Unidirectional attention
  - Find relation score between parts in the context C to the question q:

$$s_i = f(\mathbf{q}, \mathbf{w}_j^c)$$

Options for f:

- $s_i \;=\; \tanh(\mathbf{W}^c \mathbf{w}_i^c + \mathbf{W}^q \mathbf{q})$ Hermann et al. (2015)
- $s_i \;=\; \mathbf{q}^\top \mathbf{W}^s \mathbf{w}_i^c$ Chen et al. (2016)

  - Normalized by softmax into attention weights

$$\alpha_i \;=\; \frac{\exp(\mathbf{W} s_i)}{\sum_j \exp(\mathbf{W} s_j)}$$

  - Attended context vector:

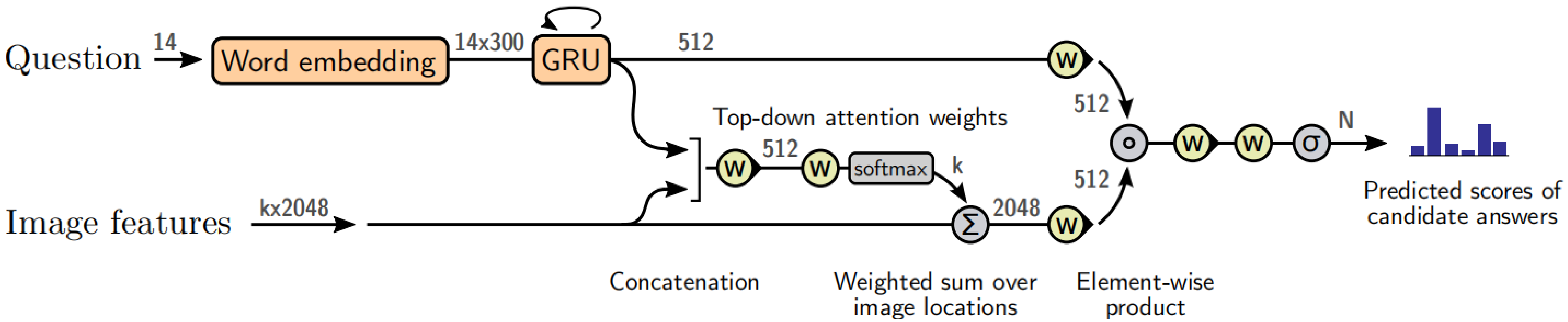
$$\mathbf{i} \;=\; \sum_i \alpha_i \mathbf{w}_i^c$$

→ We can now extract information from the context that is “relevant” to the query

[Slide credit: Vuong Le]

Bottom-up-top-down attention (Anderson et al 2017)

- Bottom-up set construction: Choosing Faster-RCNN regions with high class scores
- Top-down attention: Attending on visual features by question



→ Q: How about attention from vision objects to linguistic objects?

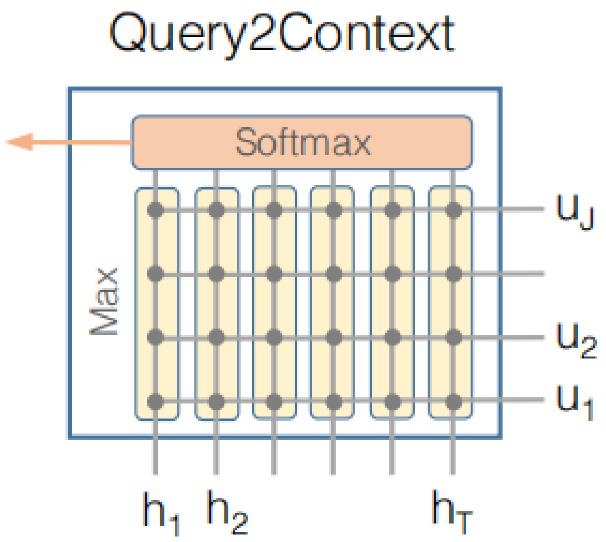
[Slide credit: Vuong Le]

Bi-directional attention

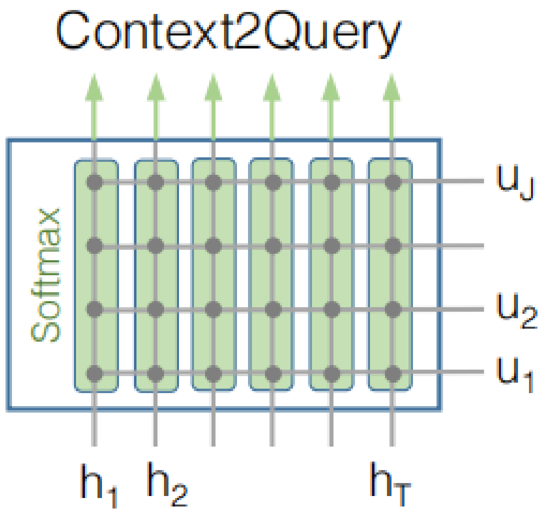
- Question-context similarity measure

$$s_i = f(\mathbf{q}, \mathbf{w}_j^c)$$

- Question-guided context attention
  - Softmax across columns



- Context-guided question attention
  - Softmax across rows



→ Q: Probably not working for image QA where single words does not have the co-reference with a region?

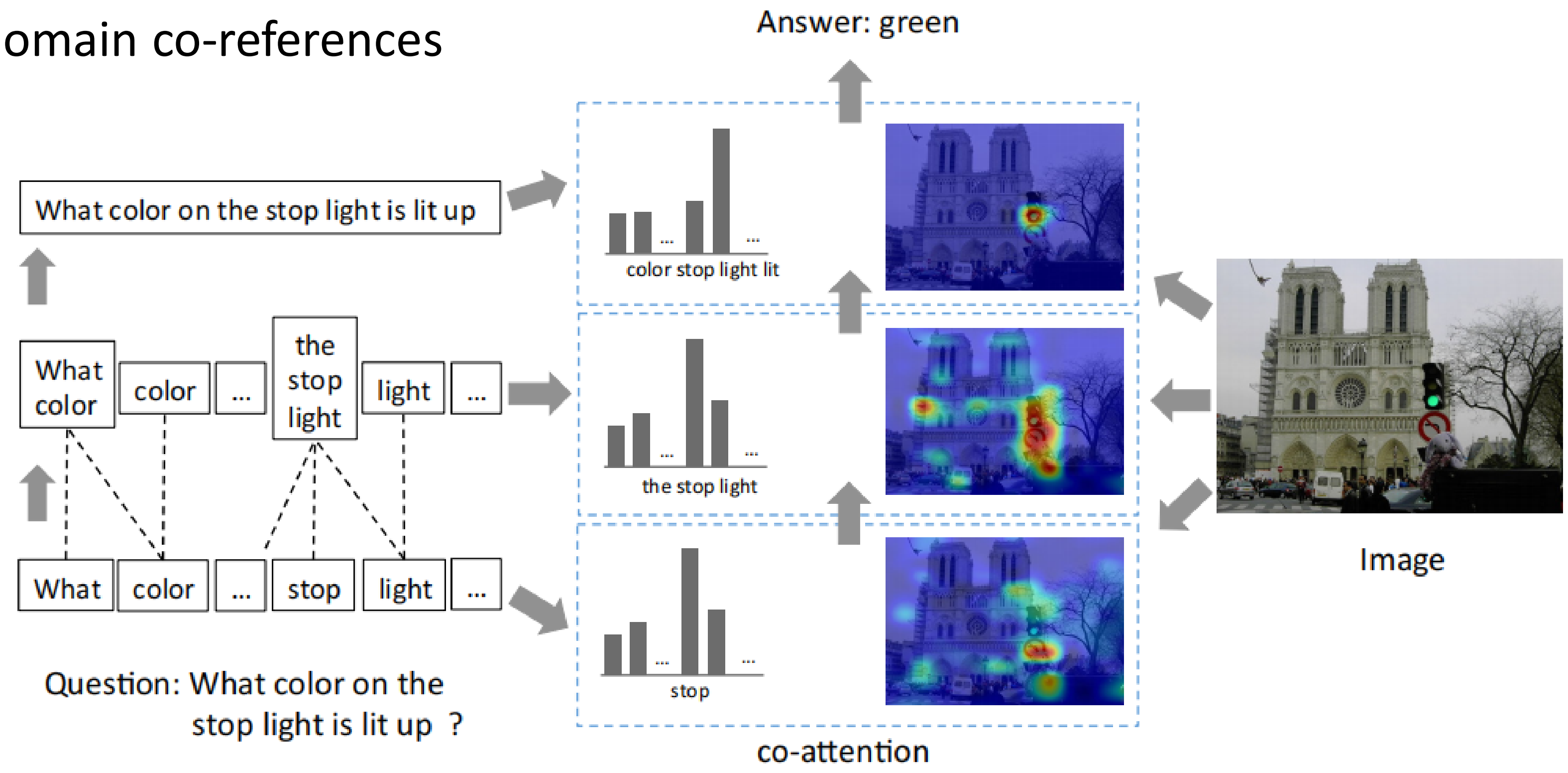
Dynamic coattention networks for question answering (Seo et al., ICLR 2017)

[Slide credit: Vuong Le]

Hierarchical co-attention for ImageQA

- The co-attention is found on a word-phrase-sentence hierarchy

→ better cross-domain co-references



→ Q: Can this be done on text QA as well?

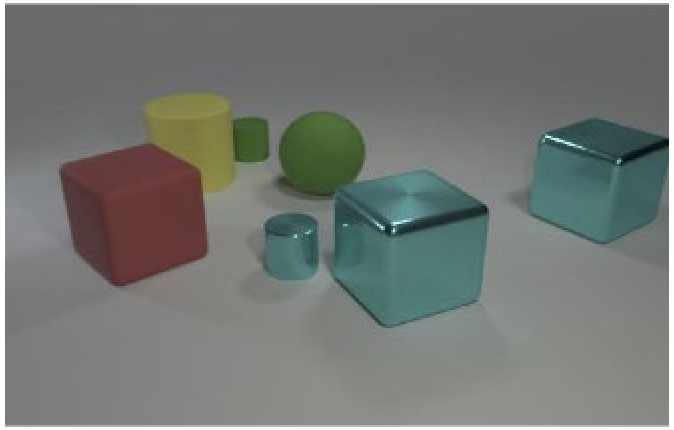
→ Q: How about questions with many reasoning hops?

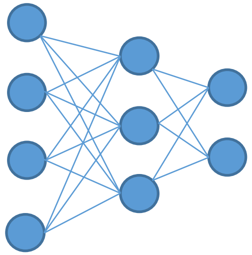
[Slide credit: Vuong Le]

Compositional Reasoning: Why do we care?

- Visual data and text data are compositional by nature.
- **Principle of compositionality**: "the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them"

[Principle of compositionality - Wikipedia]





cyan

What color is the thing with the same size as the cyan cylinder?

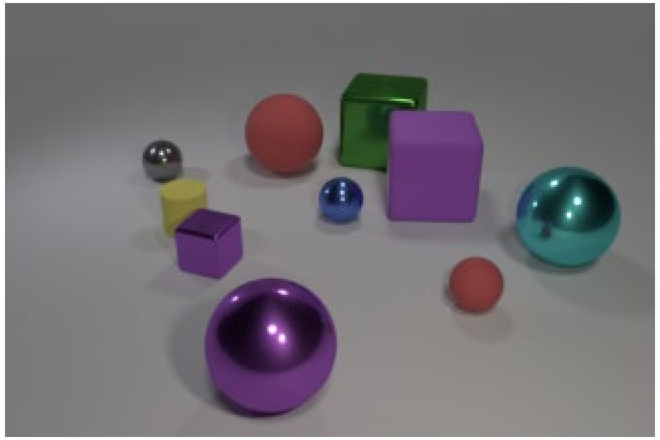
Figure from CLEVR dataset

- The network guessed the most common color in the image.
- Linguistic bias.
- Requires ***multi-step reasoning***: find cyan cylinder ➔ locate another object of the same size ➔ determine its color (green).

Multi-step Compositional Reasoning

- Complex question need multiple hops of reasoning
- Relations inside the context are multi-step themselves
- Single shot of attention won't be enough
- Single shot of information gathering is definitely not enough

→ Q: How to do multi-hop attentional reasoning?



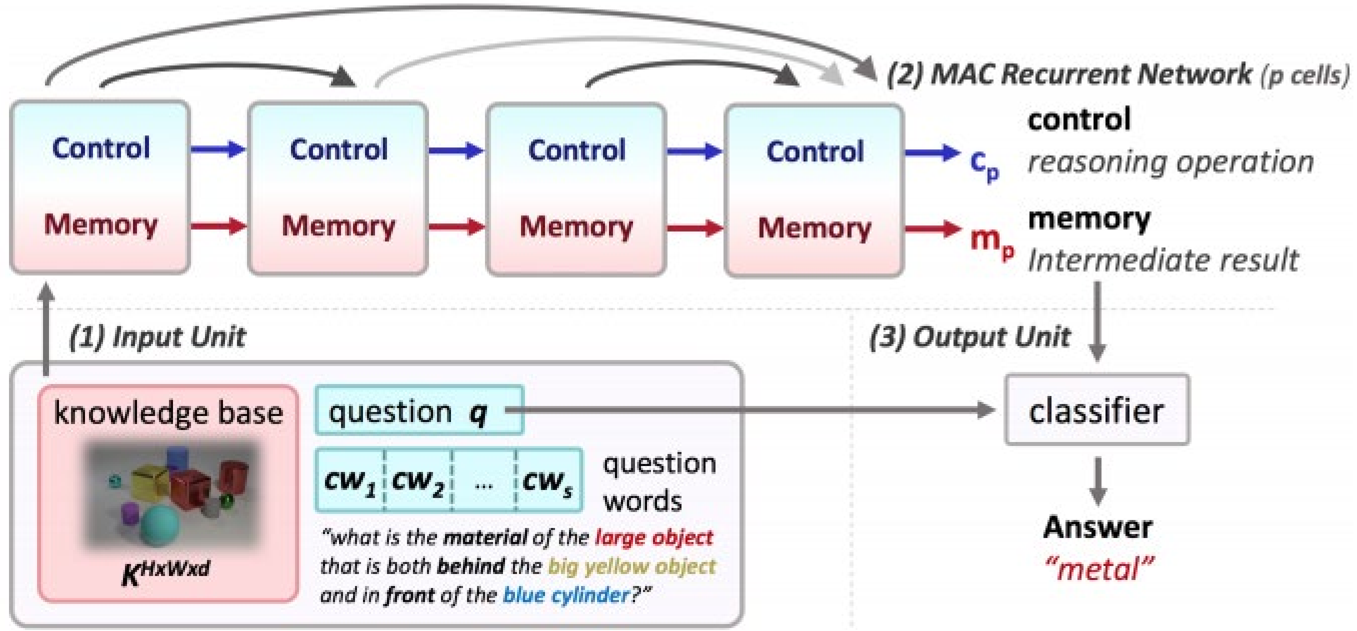
Q: *Do **the block** in **front** of **the tiny yellow cylinder** and **the tiny thing** that is to the **right** of **the large green shiny object** have the **same color**?* **A:** *No*

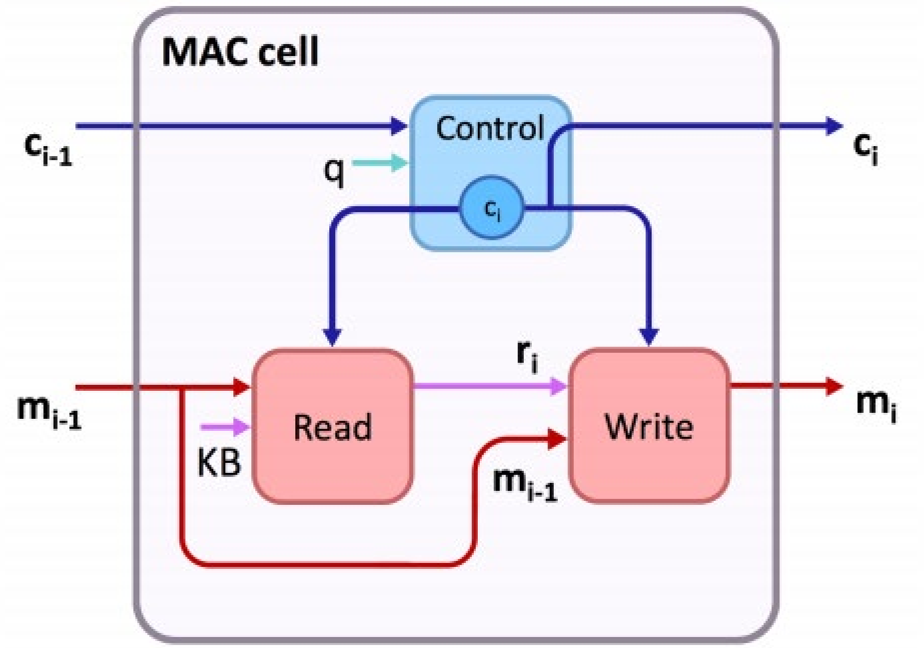
Figure: Hudson and Manning – ICLR 2018

[Slide credit: Vuong Le]

Multi-step reasoning - Memory, Attention, and Composition

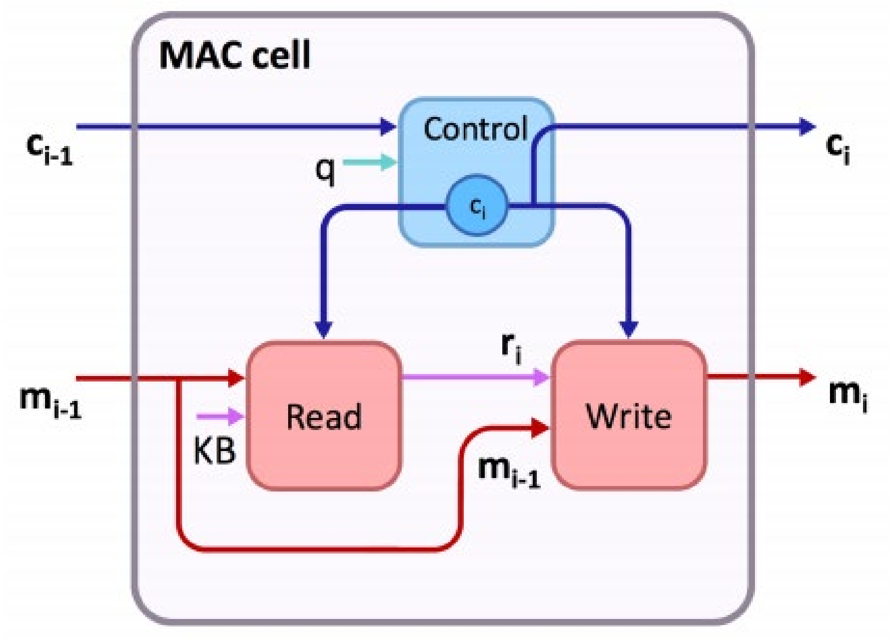
- Attention reasoning is done through multiple sequential steps.
- Each step is done with a recurrent neural cell
- *What is the key differences to the normal RNN (LSTM/GRU) cell?*
  - *Not a sequential input, it is sequential processing on static input set.*
  - *Guided by the question through a controller.*





Multi-step Attentional Reasoning

- At each step, the controller decide what to look next
- After each step, a piece of information is gathered, represented through the attention map on question words and visual objects
- A common memory kept all the information extracted toward an answer

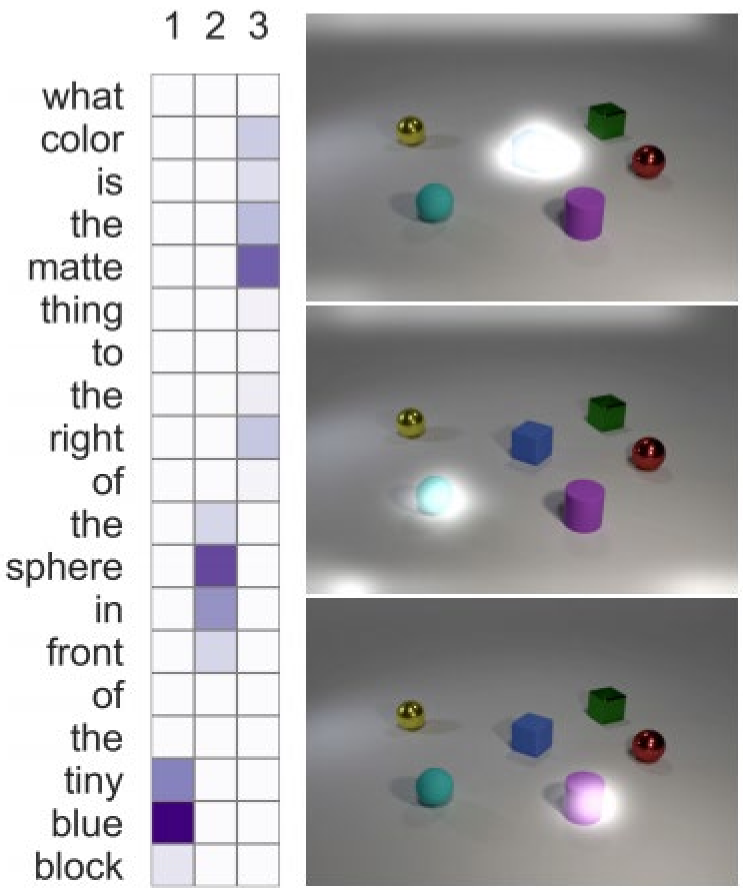


MAC network, Hudson and Manning – ICLR 2018

[Slide credit: Vuong Le]

Multi-step Attentional Reasoning

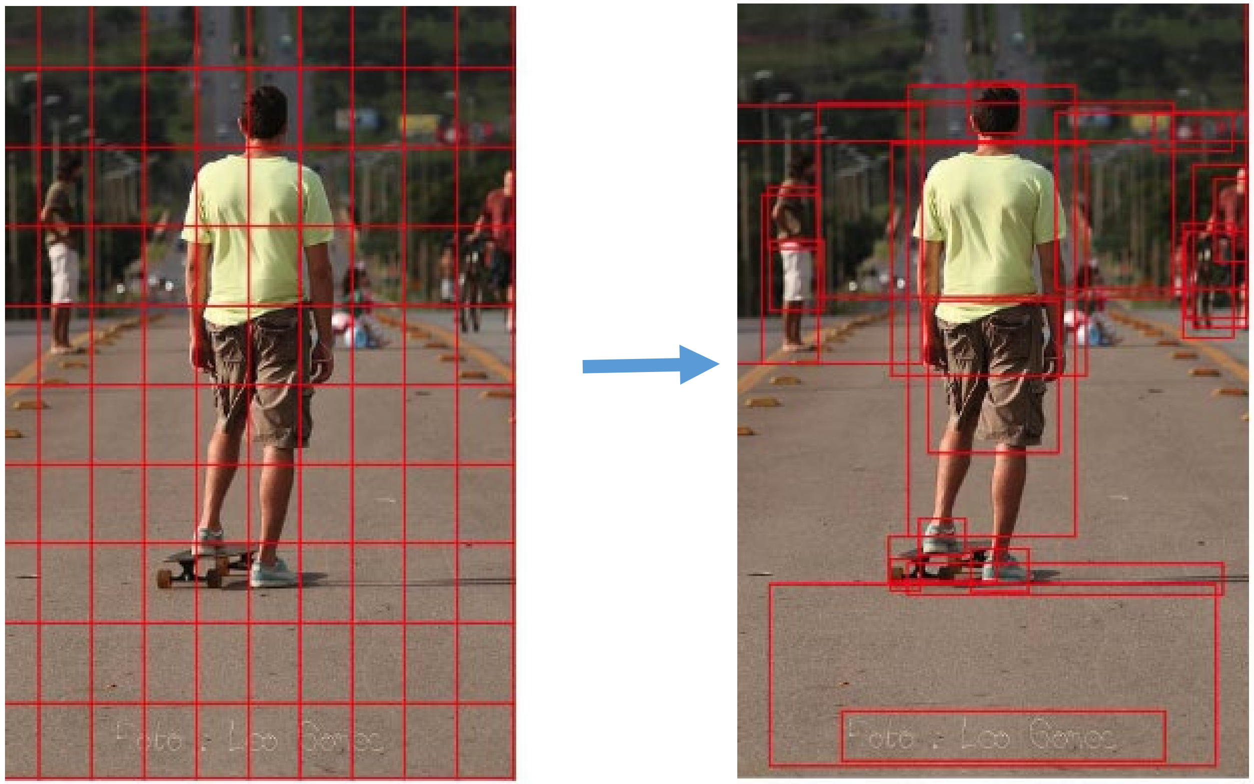
- Step 1: attends to the ***"tiny blue block"***, updating ***m*1**
- Step 2: look for ***"the sphere in front"*** ***m*2**.
- Step3: traverse from the cyan ball to the final objective – ***the purple cylinder***,



[Slide credit: Vuong Le]

From Spatial Reasoning to Object-centric Reasoning

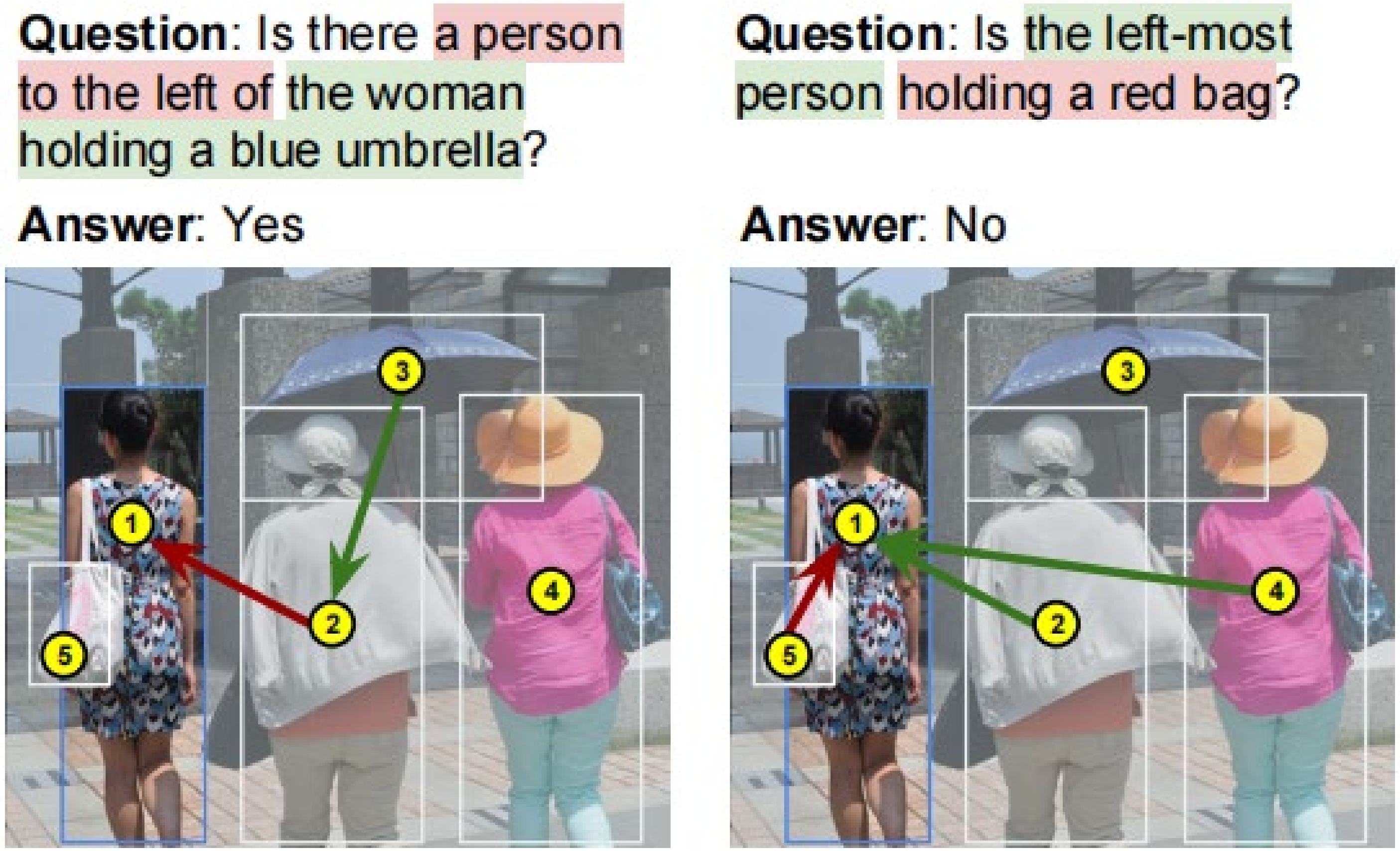
- Grid representation is irrespective of the fine-grained semantics of images.
- Region proposals are of the same semantic abstract with words -> help visual grounding.
- Interpretability.



Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *CVPR*'18.

Dynamic Reasoning Graphs

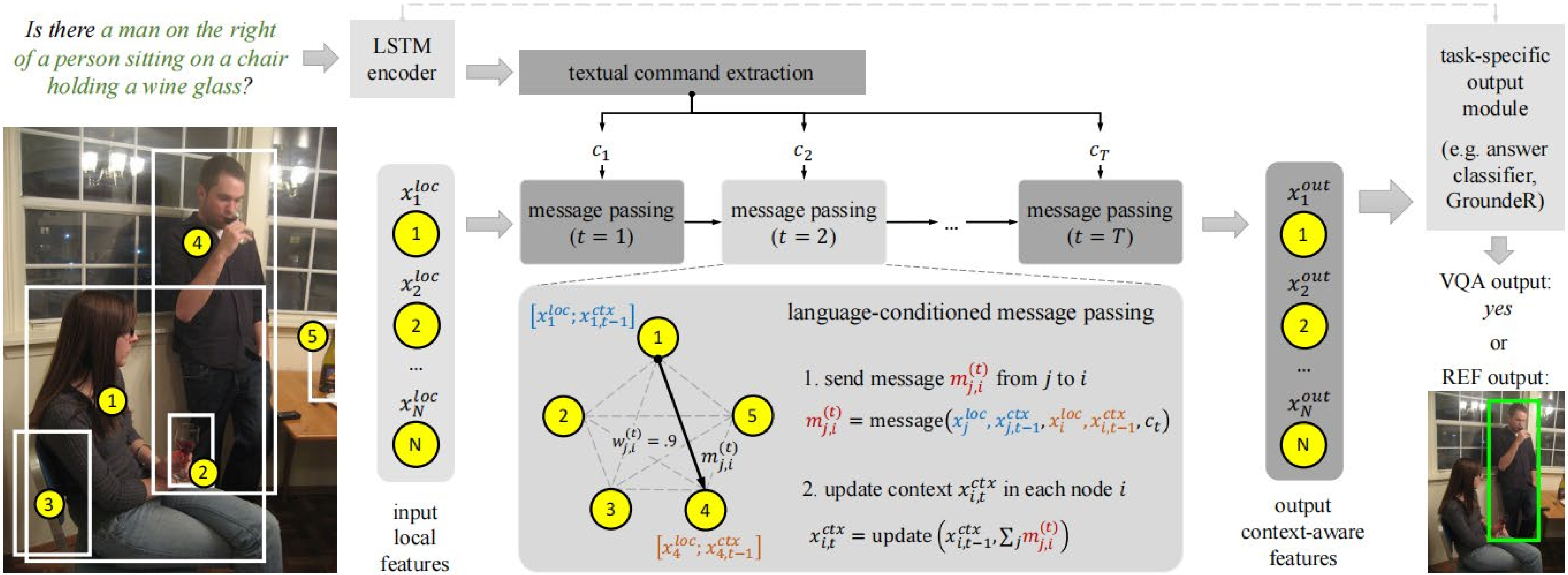
- On complex questions, multiple sets of relations are needed
- We need not only multi-step but also multi-form structures
- Let's do multiple dynamically–built graphs!



LCGN, Hu et.al. ICCV19

[Slide credit: Vuong Le]

Dynamic Reasoning Graphs



→ The questions so far act as an unstructured command in the process

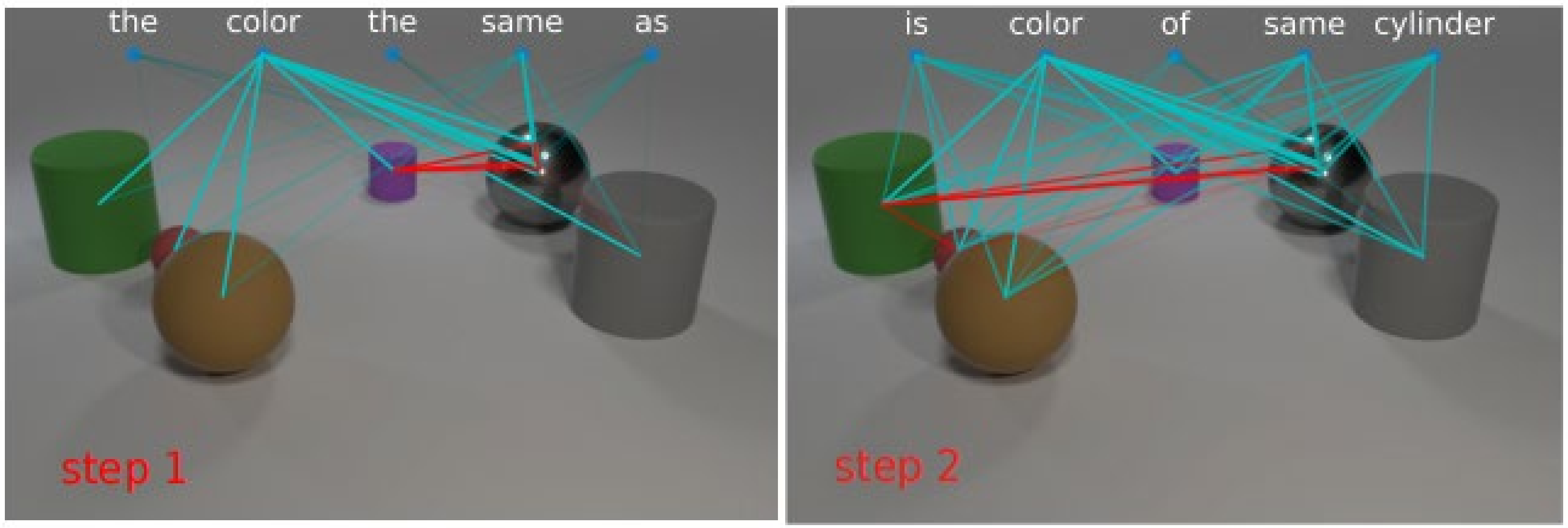
→ *Aren't their structures and relations important too?*

LCGN, Hu et.al. ICCV19

[Slide credit: Vuong Le]

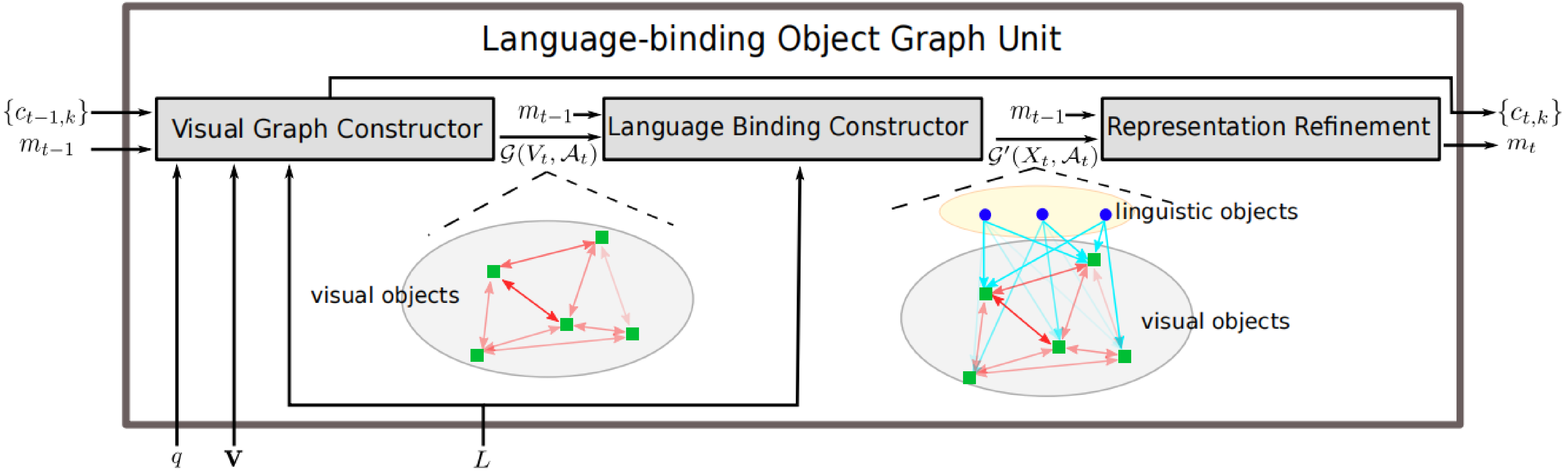
Cross-modality Graph Interactions for VQA

- Two types of nodes: Linguistic entities and visual objects
- Two types of edges:
  - Visual
  - Linguistic-visual binding *(as a fuzzy grounding)*
- Adaptively updated during reasoning



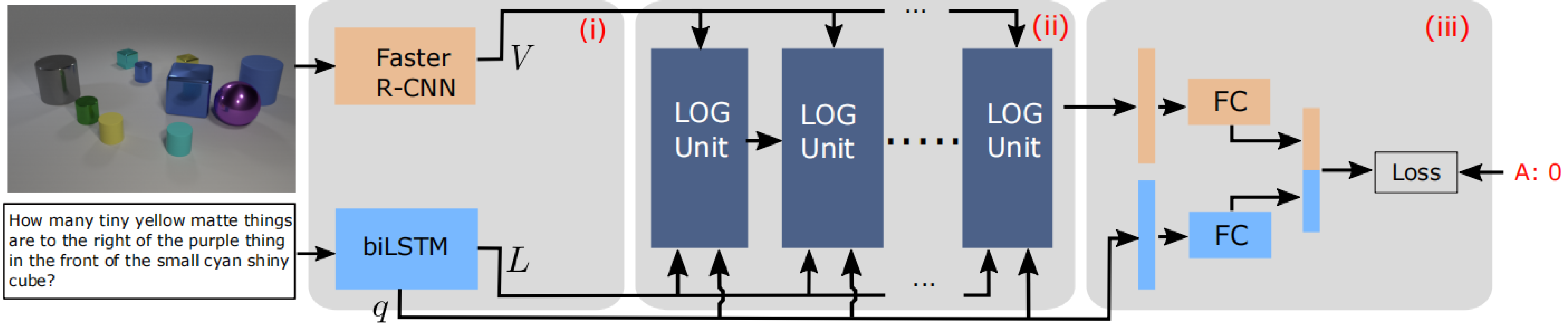
Language-binding Object Graph (LOG) Unit

- Graph constructor: build the dynamic vision graph
- Language binding constructor: find the dynamic L-V relations

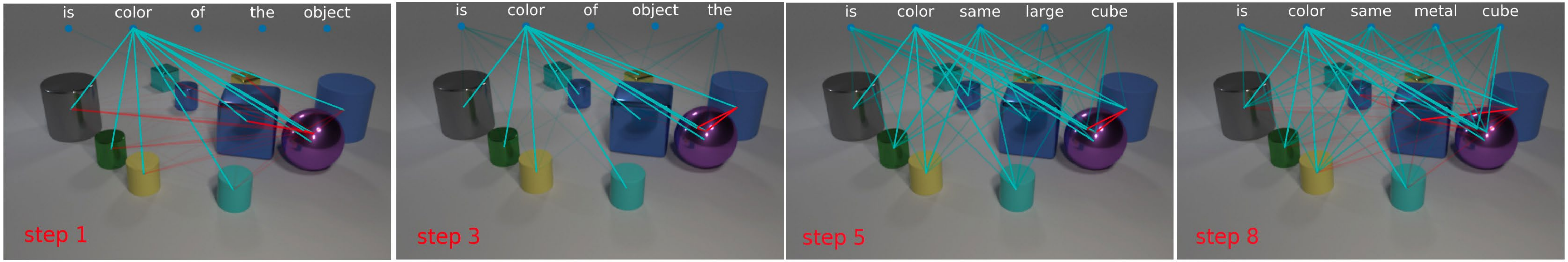


LOGNet: multi-step visual-linguistic binding

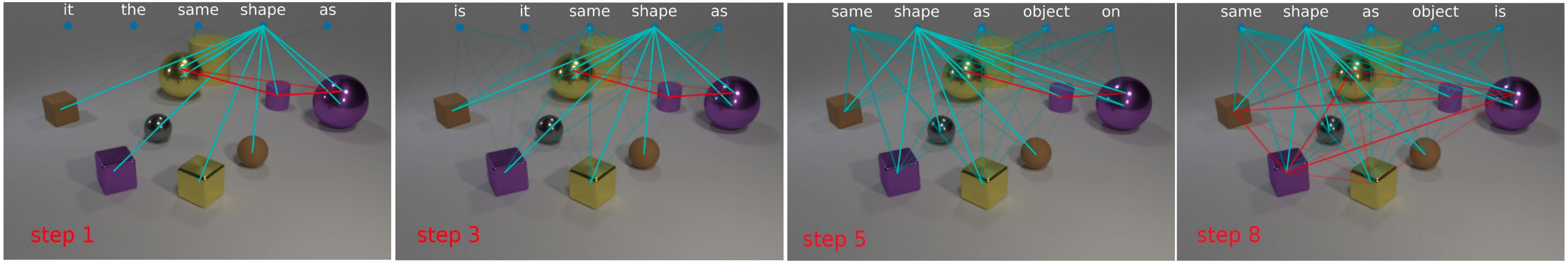
- Object-centric representation ✓
- Multi-step/multi-structure compositional reasoning ✓
- Linguistic-vision detail interaction ✓



Dynamic Language-Vision Graphs in Actions



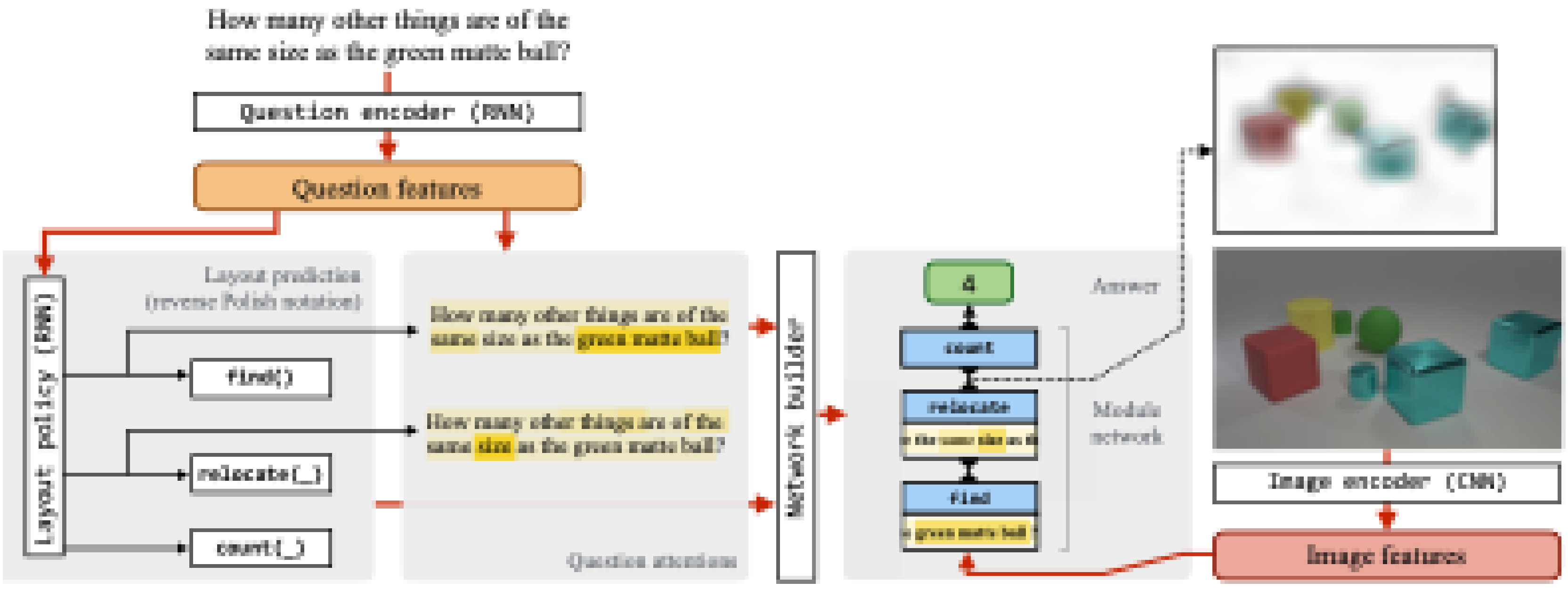
Question: Is the color of the big matte object the same as the large metal cube?
Prediction: yes **Answer**: yes



Question: There is a tiny purple rubber thing; does it have the same shape as the brown object that is on the left side of the rubber sphere?
Prediction: no **Answer**: no

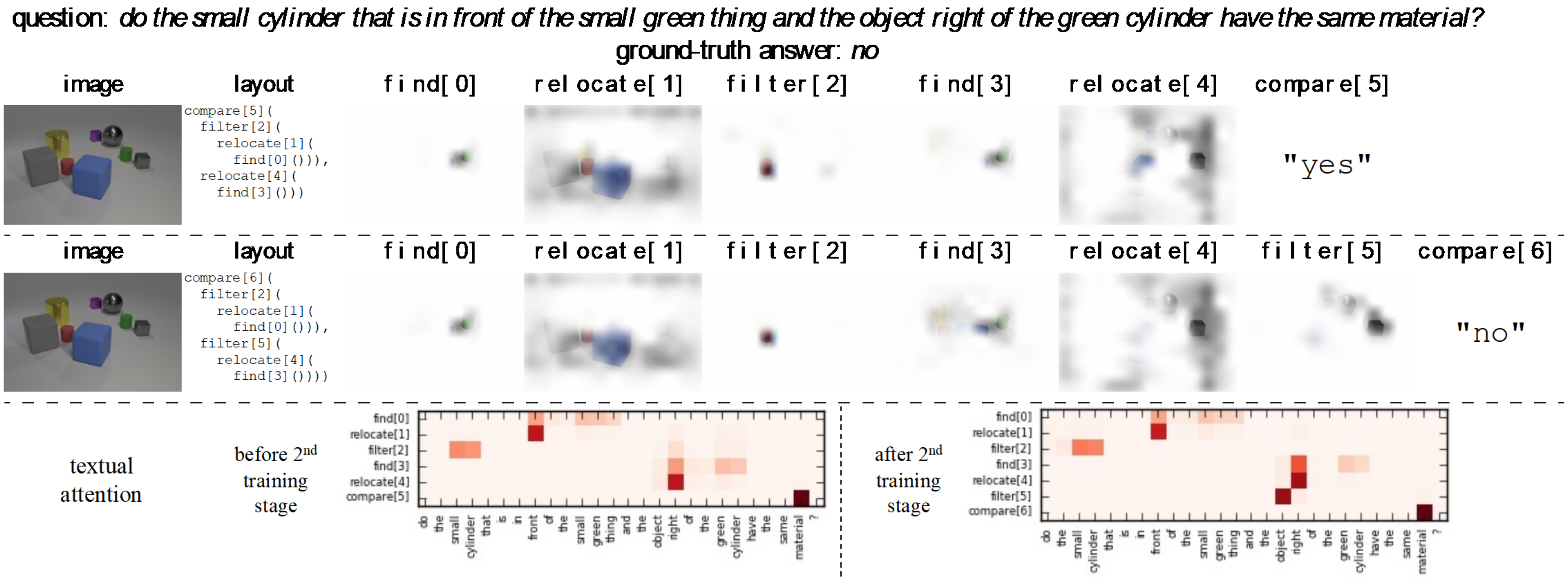
Reasoning as Query-driven Program

- Reasoning as laying out modules to reach an answer
- Composable neural architecture → question parsed as program (layout of modules)
- A module is a function (x → y), could be a sub-reasoning process ((x, q) → y).



Hu, Ronghang, et al. "Learning to reason: End-to-end module networks for visual question answering." *ICCV'17*.

What Do the Modules Learn?

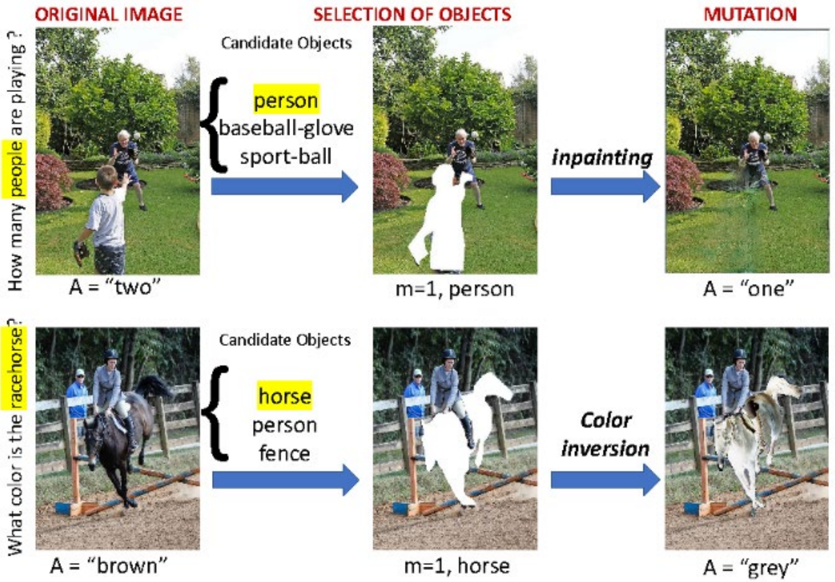


Hu, Ronghang, et al. "Learning to reason: End-to-end module networks for visual question answering." *ICCV'17*.

Current Trend in VQA: Learning to reason with less labels

- Data augmentation with analogical and counterfactual examples
- Question generation
- Self-supervised learning for question answering
- Learning with external knowledge graphs

Data Augmentation with Analogical and Counterfactual Examples



Visual counterfactual example

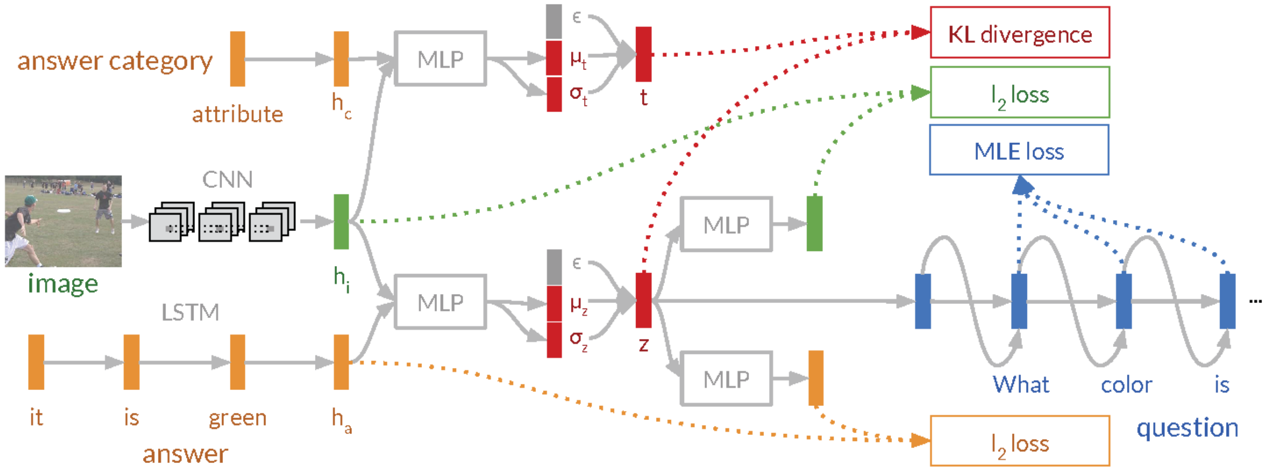
- **Poor generalization** when training under independent and identically distributed assumption.
- **Intuition**: augmenting counterfactual samples to allow machines to understand the critical changes in the input that lead to changes in the answer space.
  - Perceptually similar, yet
  - Semantically dissimilar realistic samples

Mutation Type	Question	Answer
Original	Is the lady holding the baby?	Yes
Substitution (Negation)	Is the lady not holding the baby?	No
Substitution (Adversarial)	Is the cat holding the baby?	No
Original	How many people are there?	Three
Deletion (Masking)	How many [MASK] are there?	"Number"
Original	What is the color of the man's shirt?	Blue
Substitution (Negation)	What is not the color of the man's shirt?	Magenta
Deletion (Masking)	Is the [MASK] holding the baby?	Can't say
Original	What color is the umbrella ?	Pink
Deletion (Masking)	What color is the [MASK]?	"color"

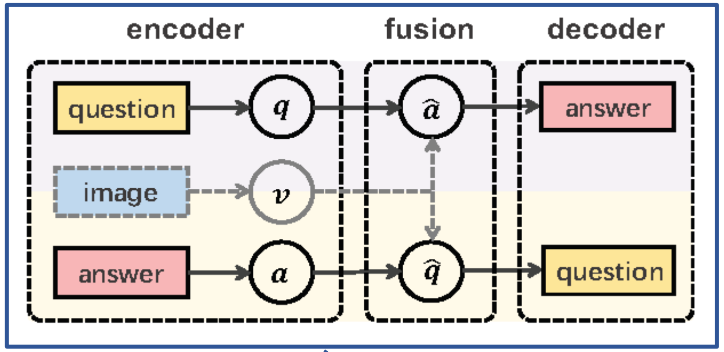
Language counterfactual examples

Gokhale, Tejas, et al. "Mutant: A training paradigm for out-of-distribution generalization in visual question answering." *EMNLP'20*.

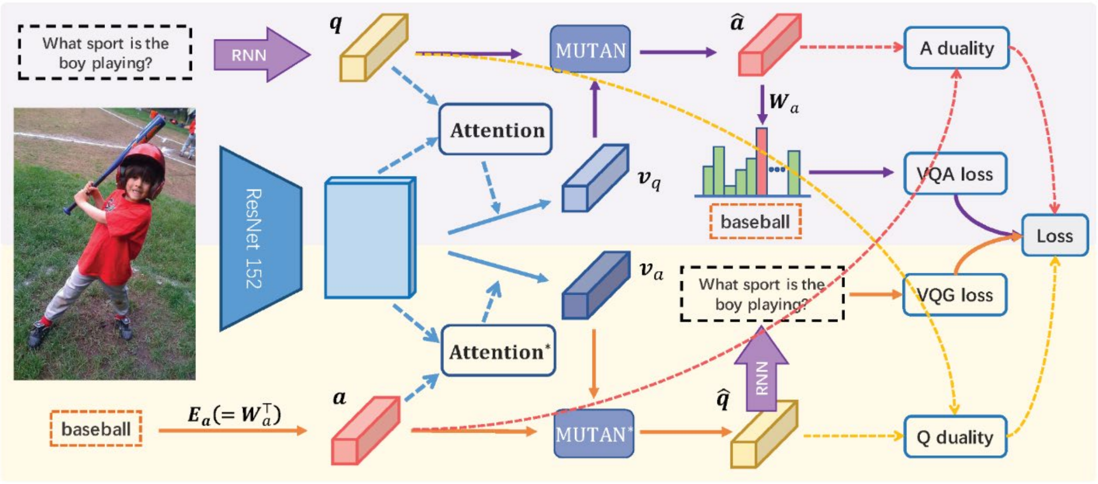
Question Generations



Krishna, Ranjay, Michael Bernstein, and Li Fei-Fei. "Information maximizing visual question generation." *CVPR*'19.



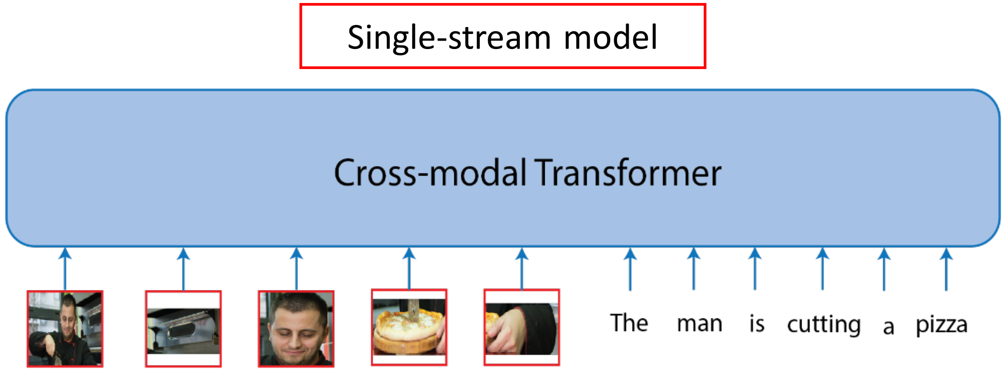
- Question answering is a few-shot learning problem. Question generation helps cover a wider range of concepts.
- Question generation can be done with either supervised and unsupervised learning.



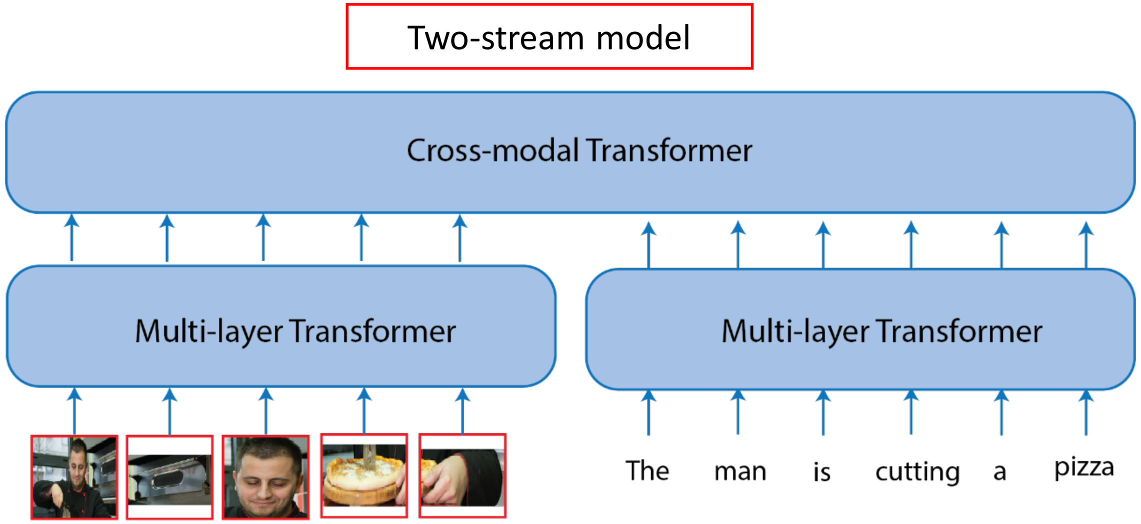
Li, Yikang, et al. "Visual question generation as dual task of visual question answering." *CVPR*'18.

Visual QA as a Down-stream Task of Visual-Language BERT Pre-trained Models

Numerous pre-trained visual language models during 2019-2021.



VisualBERT (Li, Liunian Harold, et al., 2019)
VL-BERT (Su, Weijie, et al., 2019)
UNITER (Chen, Yen-Chun, et al., 2019)
12-in-1 (Lu, Jiasen, et al., 2020)
Pixel-BERT (Huang, Zhicheng, et al., 2019)
OSCAR (Li, Xiujun, et al., 2020)



ViLBERT (Lu, Jiasen, et al. , 2019)
LXMERT (Tan, Hao, and Mohit Bansal, 2019)

[Slide credit: Licheng Yu et al.]

Learning with External Knowledge

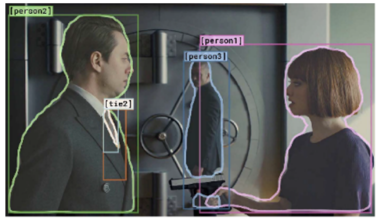
Why external knowledge for reasoning?

- Questions can be beyond visual recognition (e.g. firetrucks usually use a fire hydrant).
- Human's prior knowledge for cognition-level reasoning (e.g. human's goals, intents etc.)

Q: What sort of vehicle uses this item?
A: firetruck

Q: What is the sports position of the man in the orange shirt?
A: goalie/goalkeeper

Marino, Kenneth, et al. "Ok-vqa: A visual question answering benchmark requiring external knowledge." *CVPR'19*.



Why is [person1] pointing a gun at [person2]?

a) [person1] wants to kill [person2].(1%)

b) [person1] and [person3] are robbing the bank and [person2] is the bank manager. (71%)

c) [person2] has done something to upset [person1]. (18%)

d) Because [person2] is [person1]'s daughter. [person1] wants to protect [person2]. (8%)

b) is right because...

a) [person1] is chasing [person1] and [person3] because they just robbed a bank. (33%)

b) Robbers will sometimes hold their gun in the air to get everyone's attention. (5%)

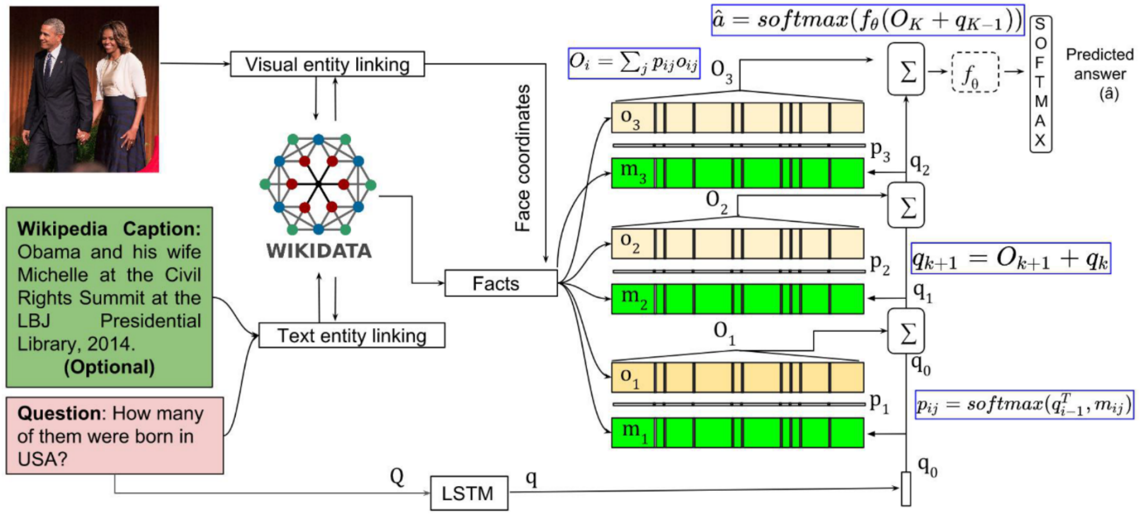
c) The vault in the background is similar to a bank vault. [person3] is waiting by the vault for someone to open it. (49%)

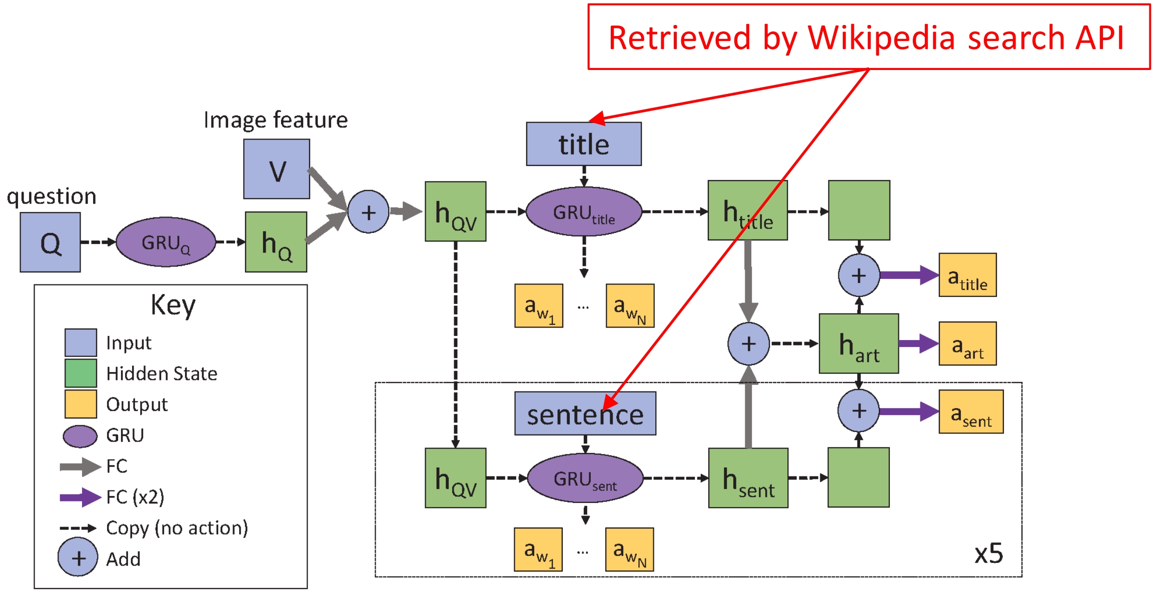
d) A room with barred windows and a counter usually resembles a bank. (11%)

Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." *CVPR'19*.

Learning with External Knowledge

Shah, Sanket, et al. "Kvqa: Knowledge-aware visual question answering." *AAAI'19*.





Marino, Kenneth, et al. "Ok-vqa: A visual question answering benchmark requiring external knowledge." *CVPR'19*.

Lecture 9: Video/Movie Question Answering

Recall the Learning to Reason formulation

- Input:
  - Context C is a **dynamic scene**
  - A query q
- Output: an answer satisfying

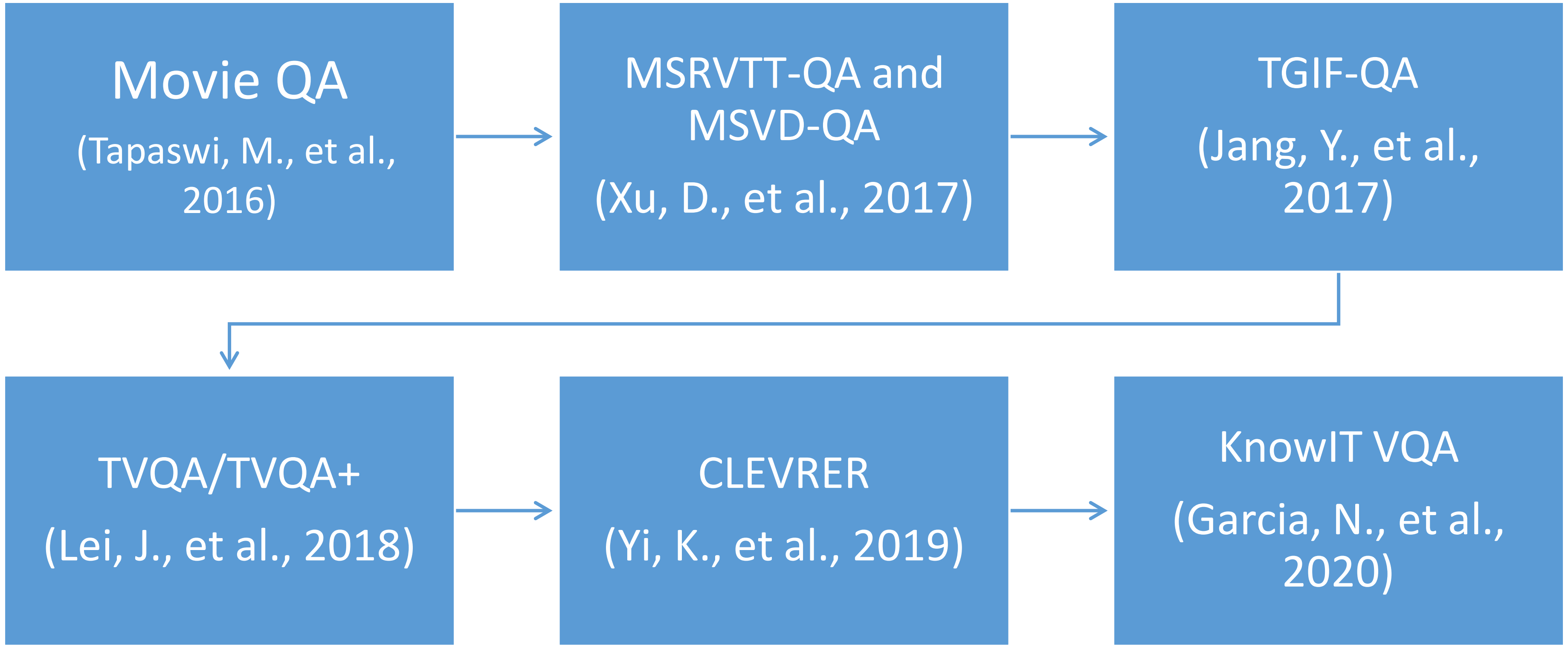
$$\tilde{a} = \underset{a \in \mathbb{A}}{\mathrm{argmax}} \mathcal{P}_\theta(a \mid q, C)$$

Q: What does the boy with a brown hoodie do before running away ? ***A****: flip to the front side*

Challenges

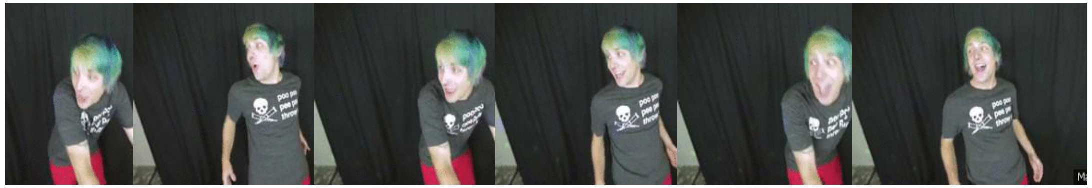
- Difficulties in data annotation.
- Content for performing reasoning spreads over space-time and multiple modalities (videos, subtitles, speech etc.)

Video QA Datasets



Video QA datasets

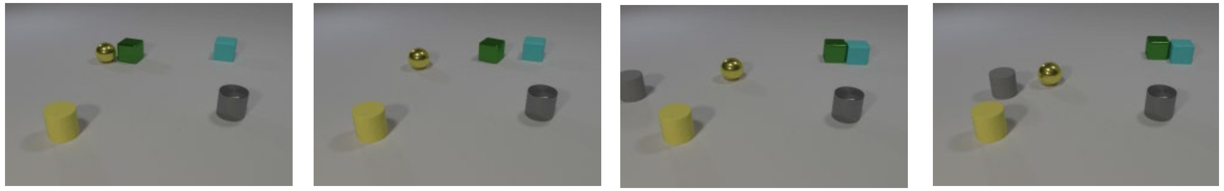
(TGIF-QA, Jang et al., 2018)



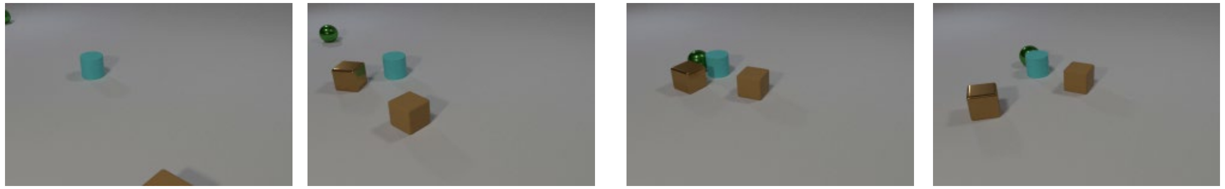
Q: What does the man do 5 times?
A: (0) step (3) bounce
(2) sway head (4) knod head
(5): move body to the front

Q: What does the man do before turing body to left?
A: (0) run a cross a ring (3) flip cover face with hand
(2) pick up the man's hand (4) raise hand
(5): breath

(CLEVRER, Yi, Kexin, et al., 2020)

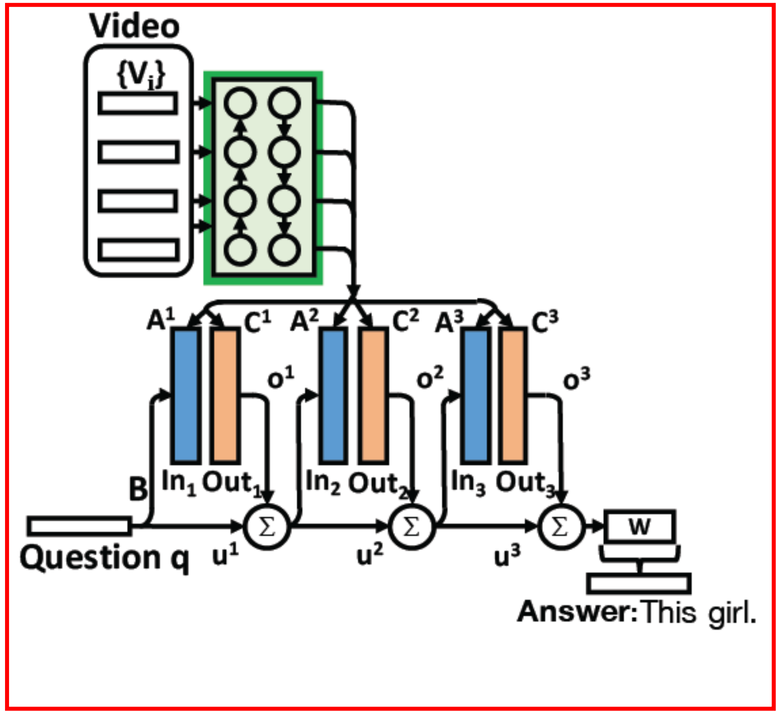


Q: What color is the last object to collide with the green cube?
A: cyan

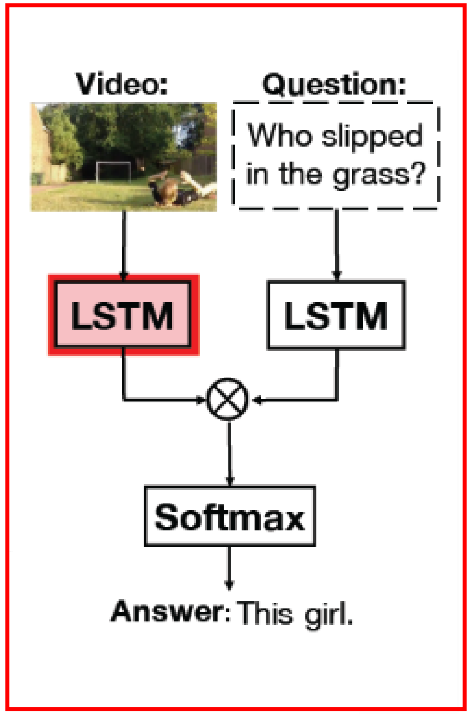


Q: Which of the following is responsible for the collision between the metal cube and the cylinder?
A: (a) The presence of the brown rubber cube
(b) The sphere's colliding with the cylinder
(c) The rubber cube's entrance
(d) The collision between the metal cube and the sphere

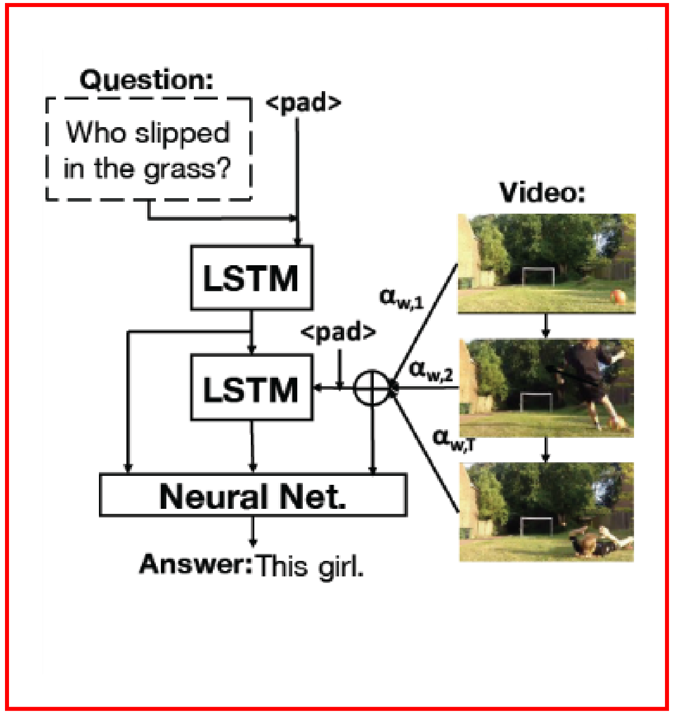
Video QA as a spatio-temporal extension of Image QA



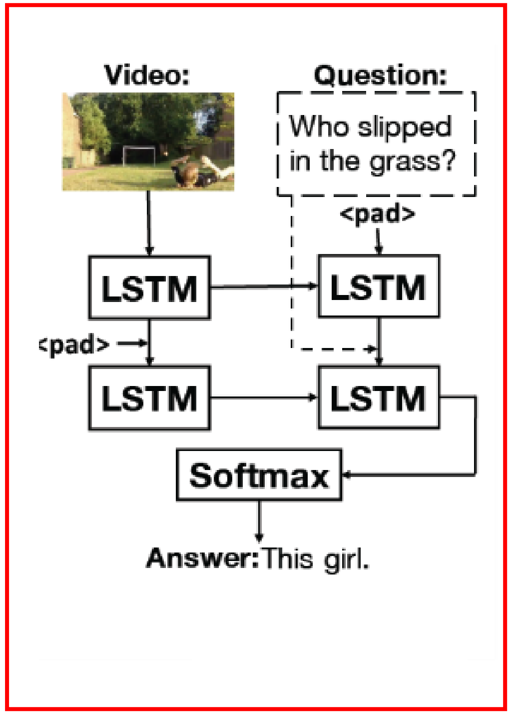
(a) Extended end-to-end memory network



(b) Extended simple VQA model



(c) Extended temporal attention model



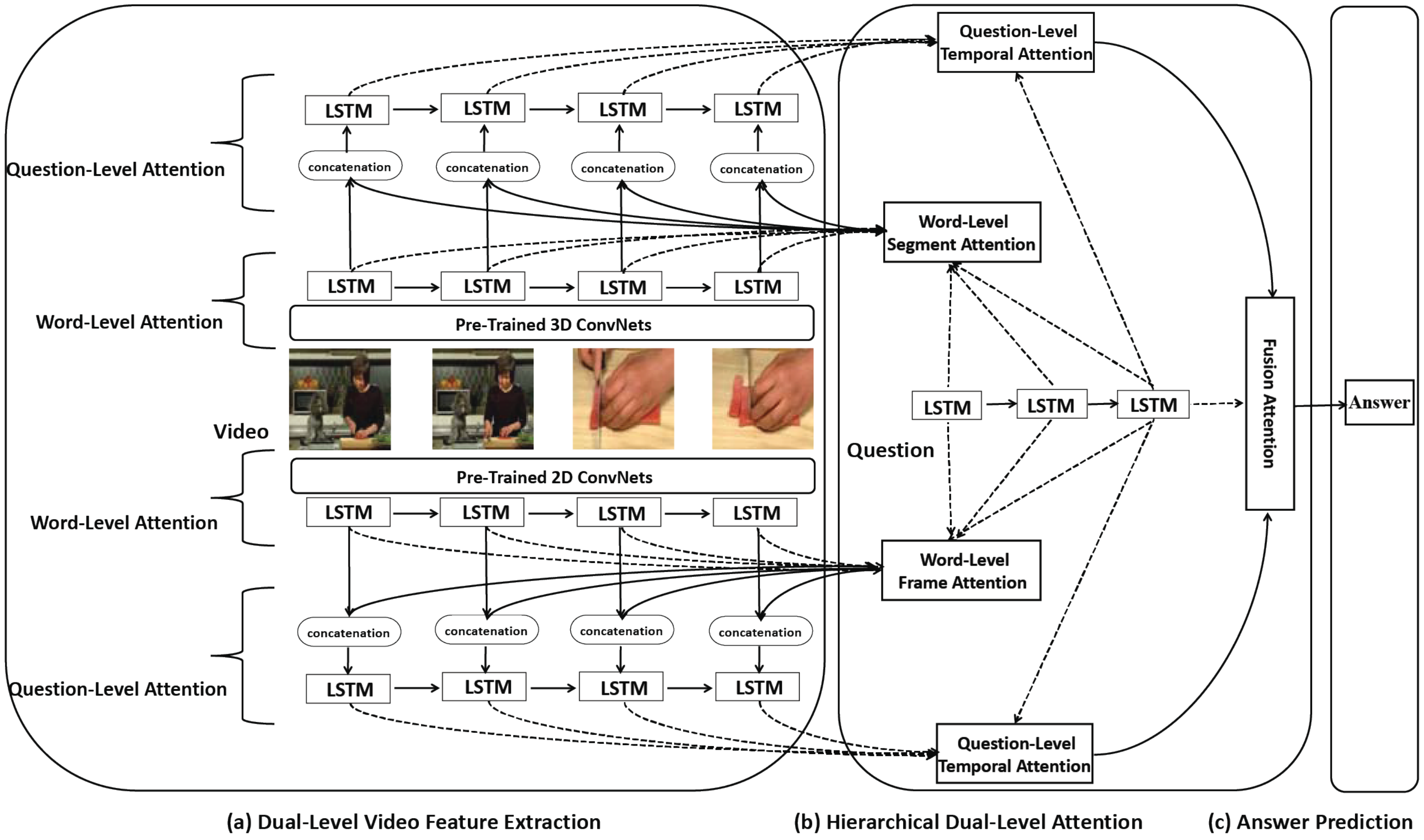
(d) Extended sequence-to-sequence model

Zeng, Kuo-Hao, et al. "Leveraging video descriptions to learn video question answering." *AAAI'17*.

Spatio-temporal cross-modality alignment

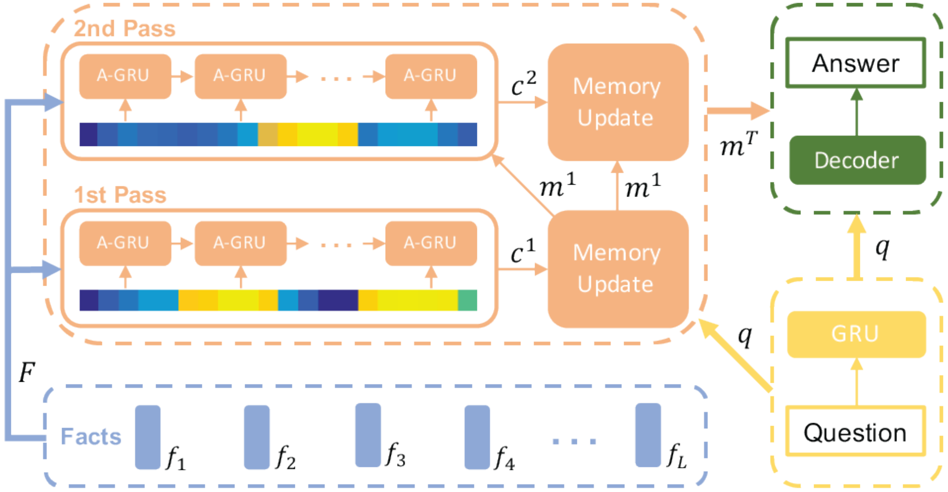
Key ideas:

- Explore the correlation between vision and language via attention mechanisms.
- Joint representations are query-driven spatio-temporal features of a given videos.

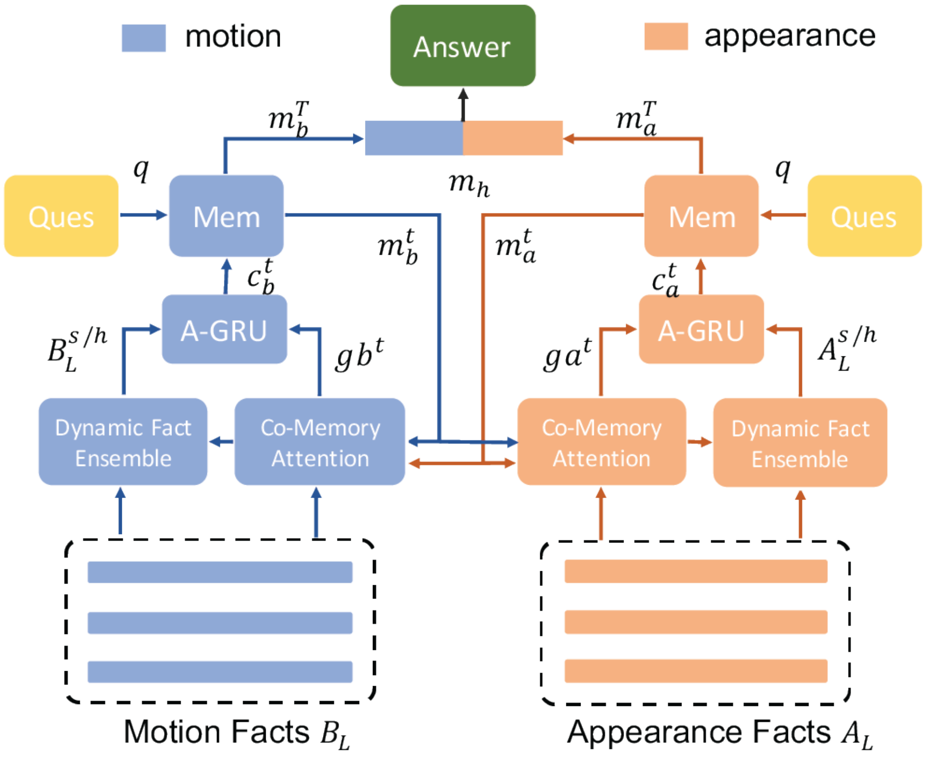


Zhao, Zhou, et al. "Video question answering via hierarchical dual-level attention network learning." *ACL'17*.

Memory-based Video QA



General Dynamic Memory Network (DMN)



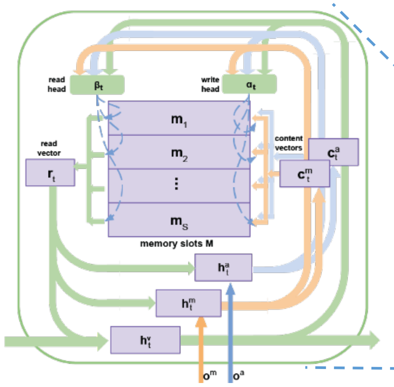
Co-memory attention networks for Video QA

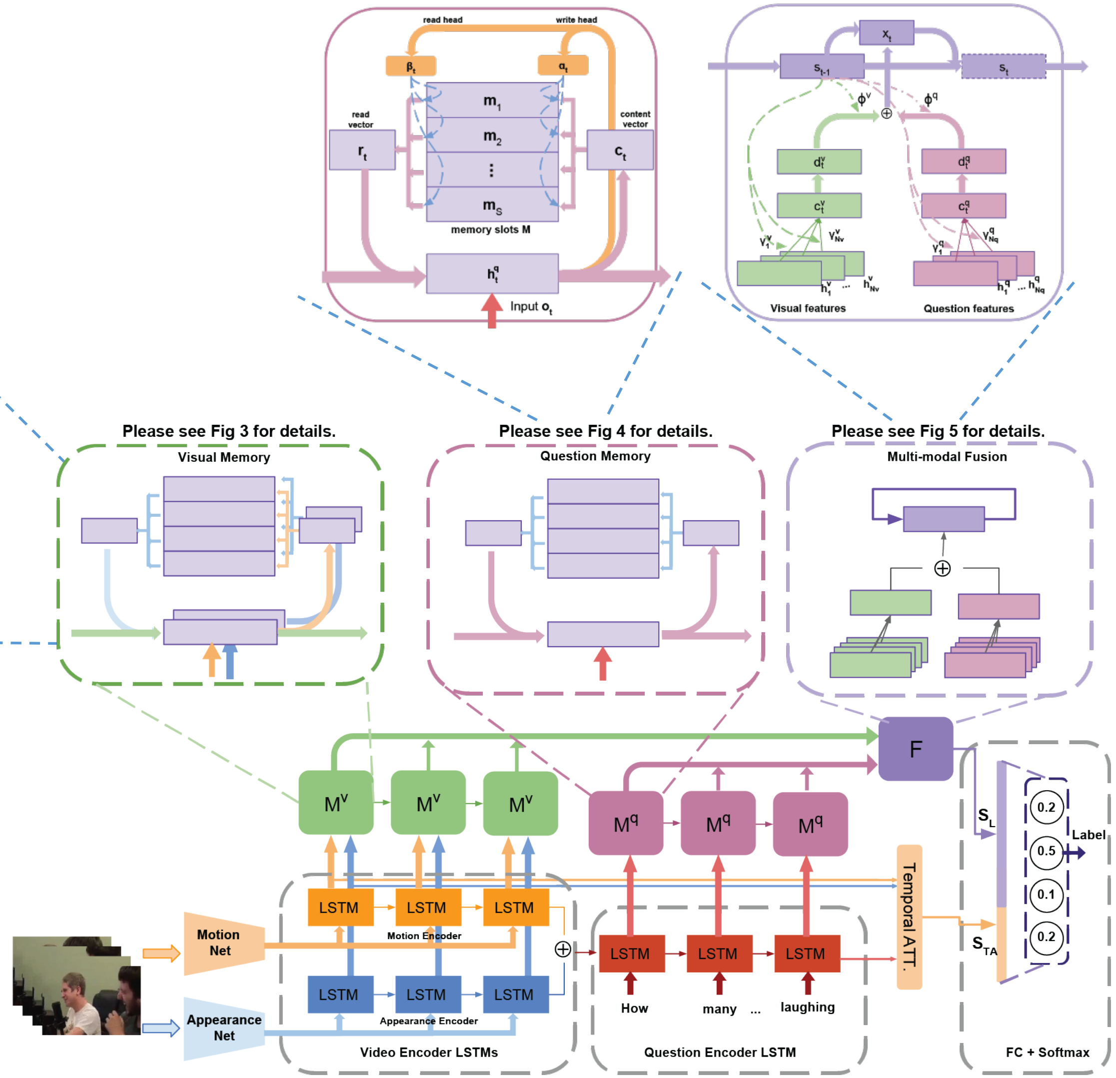
Key ideas:

- DMN refines attention over a set of facts to extract reasoning clues.
- Motion and appearance features are complementary clues for question answering.

Gao, Jiyang, et al. "Motion-appearance co-memory networks for video question answering." *CVPR'18*.

Memory-based Video QA





Heterogeneous video memory for Video QA

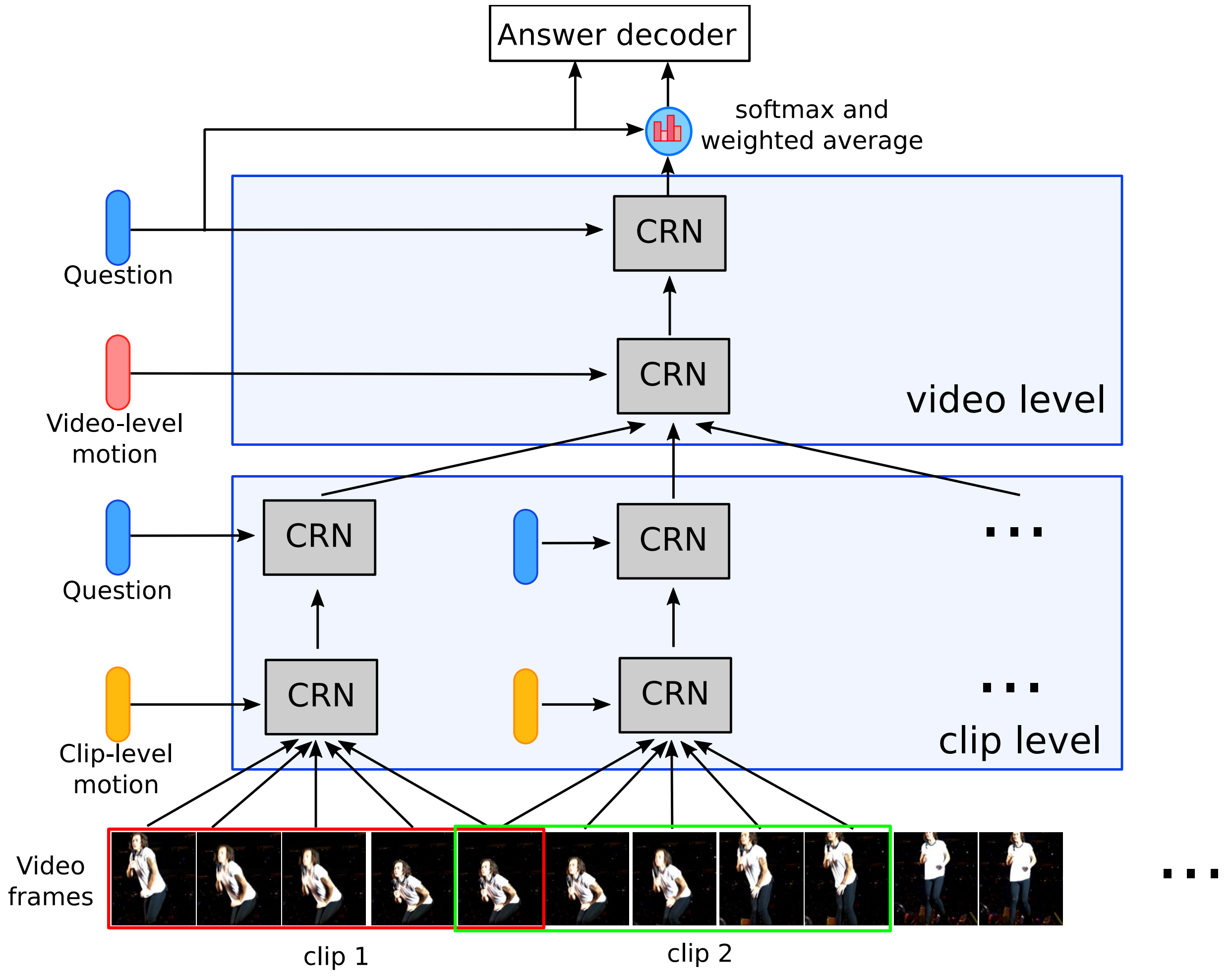
Key differences:

- Learning a joint representation of multimodal inputs at each memory read/write step.
- Utilizing external question memory to model context-dependent question words.

Fan, Chenyou, et al. "Heterogeneous memory enhanced multimodal attention model for video question answering." *CVPR'19*.

Multimodal reasoning units for Video QA

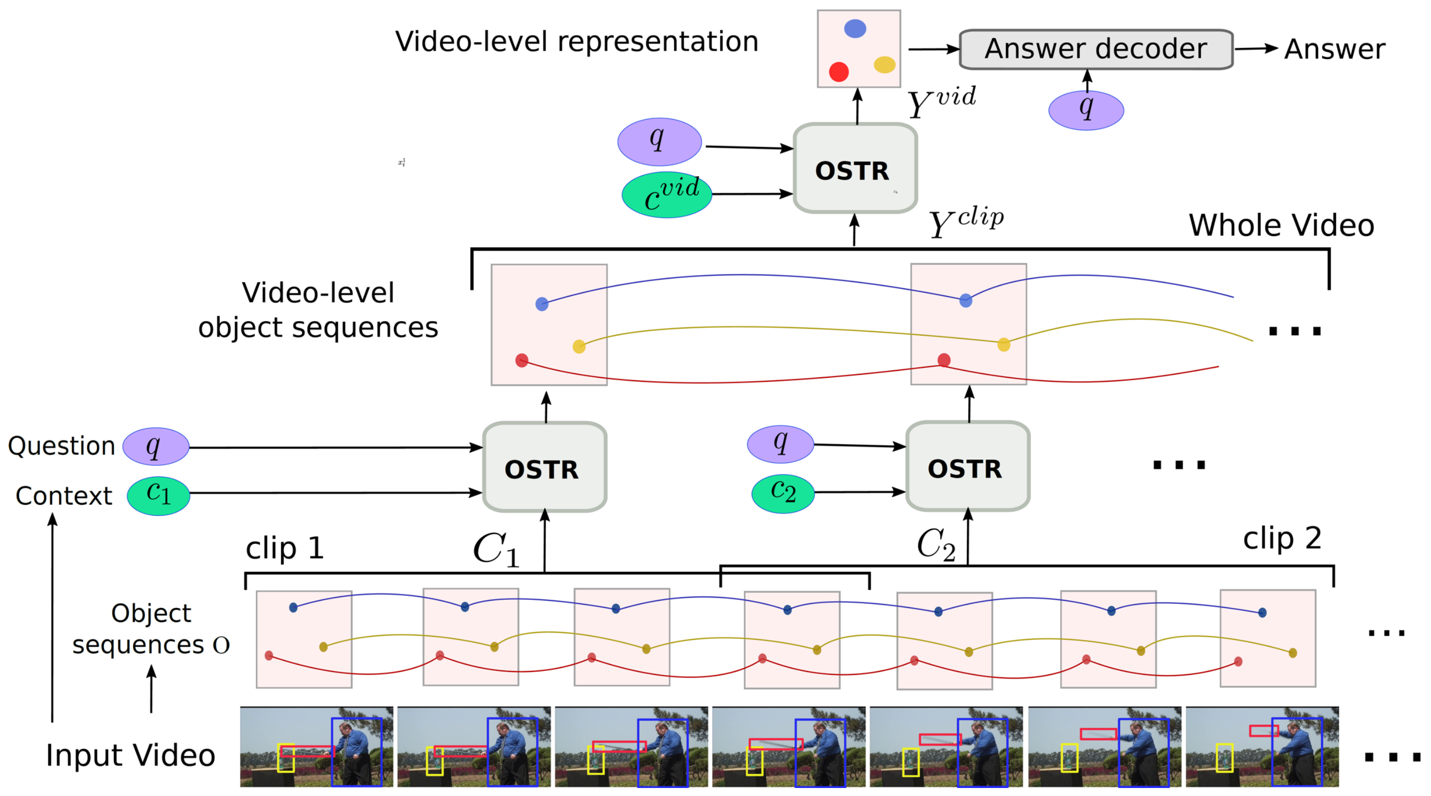
- CRN: Conditional Relation Networks.
- Inputs:
    - Frame-based appearance features
    - Motion features
    - Query features
- Outputs:
    - Joint representations encoding temporal relations, motion, query.



Le, Thao Minh, et al. "Hierarchical conditional relation networks for video question answering." CVPR'20

Object-oriented spatio-temporal reasoning for Video QA

- OSTR: Object-oriented Spatio-Temporal Reasoning.
- Inputs:
  - Object lives tracked through time.
  - Context (motion).
  - Query features.
- Outputs:
  - Joint representations encoding temporal relations, motion, query.



Video QA as a down-stream task of video language pre-training

VideoBERT
Apr., 2019

HowTo100M
Jun., 2019

MIL-NCE
Dec., 2019

UniViLM
Feb., 2020

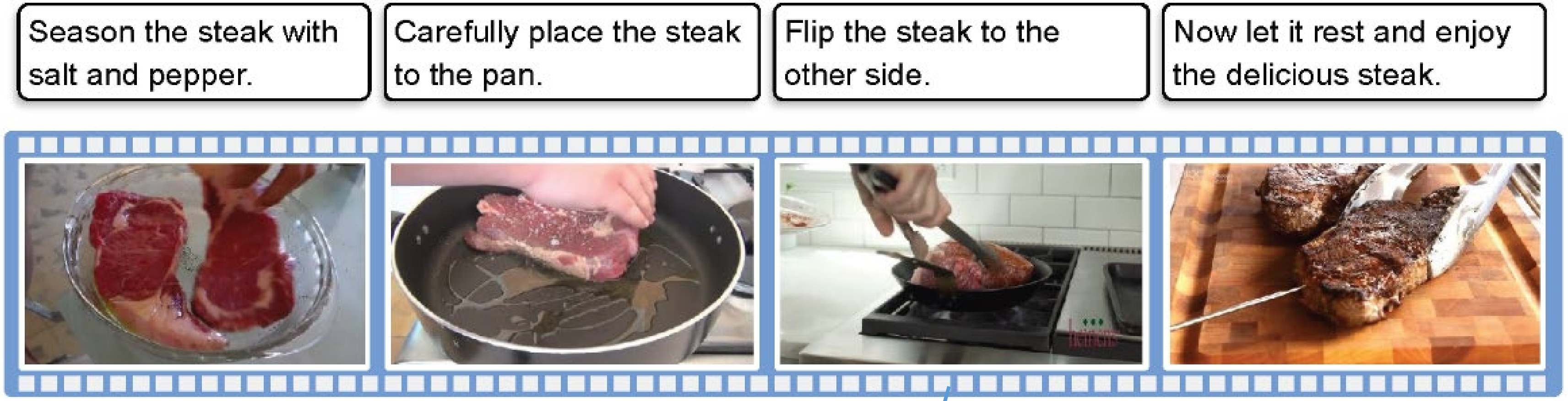
HERO
May, 2020

ClipBERT
Feb., 2021

VideoBERT: a joint model for video and language representation learning

- Data for training: Sample videos and texts from YouCook II.

Instructions in text given by ASR toolkit

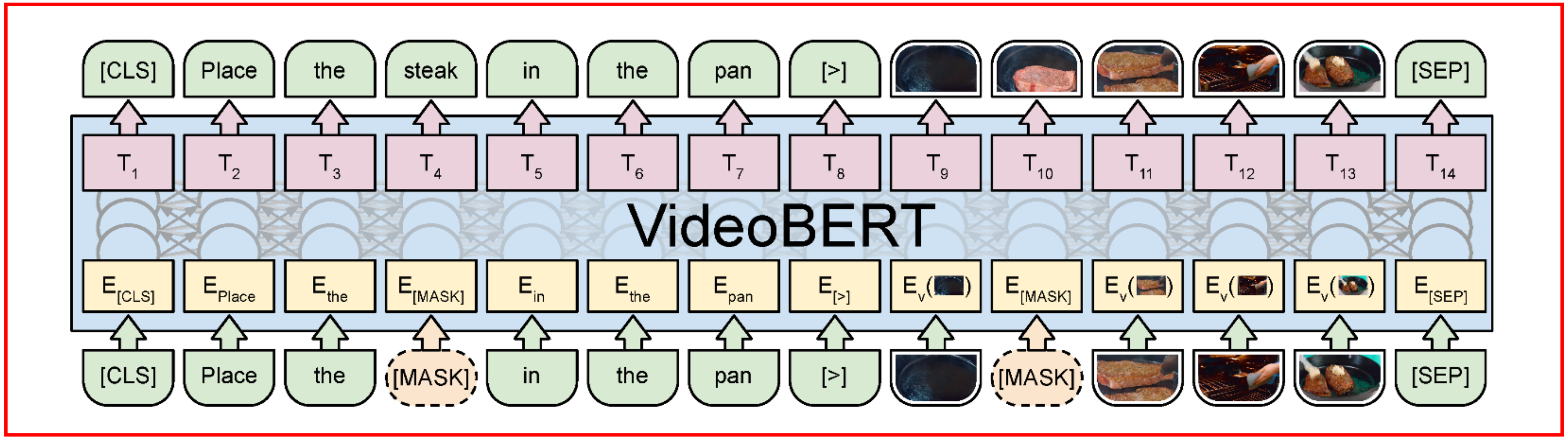


Subsampled video segments

Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." ICCV'19.

VideoBERT: a joint model for video and language representation learning

Pre-training

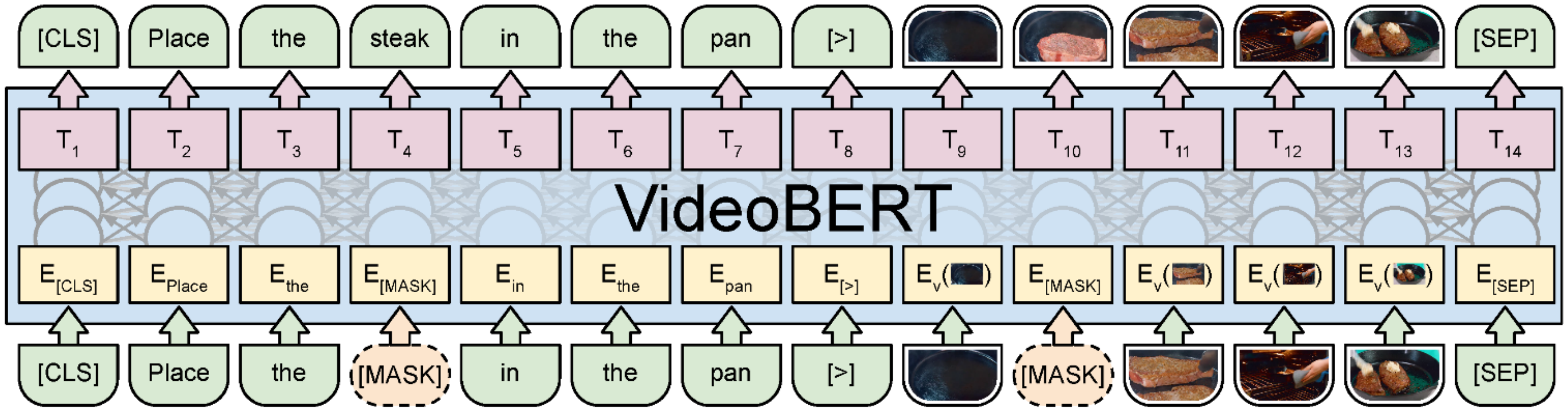


- Linguistic representations:
  - Tokenized texts into WordPieces, similar as BERT.
- Visual representations:
  - S3D features for each segmented video clips.
  - Tokenized into clusters using hierarchical k-means.

Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." ICCV'19.

VideoBERT: a joint model for video and language representation learning

Pre-training



Down-stream tasks

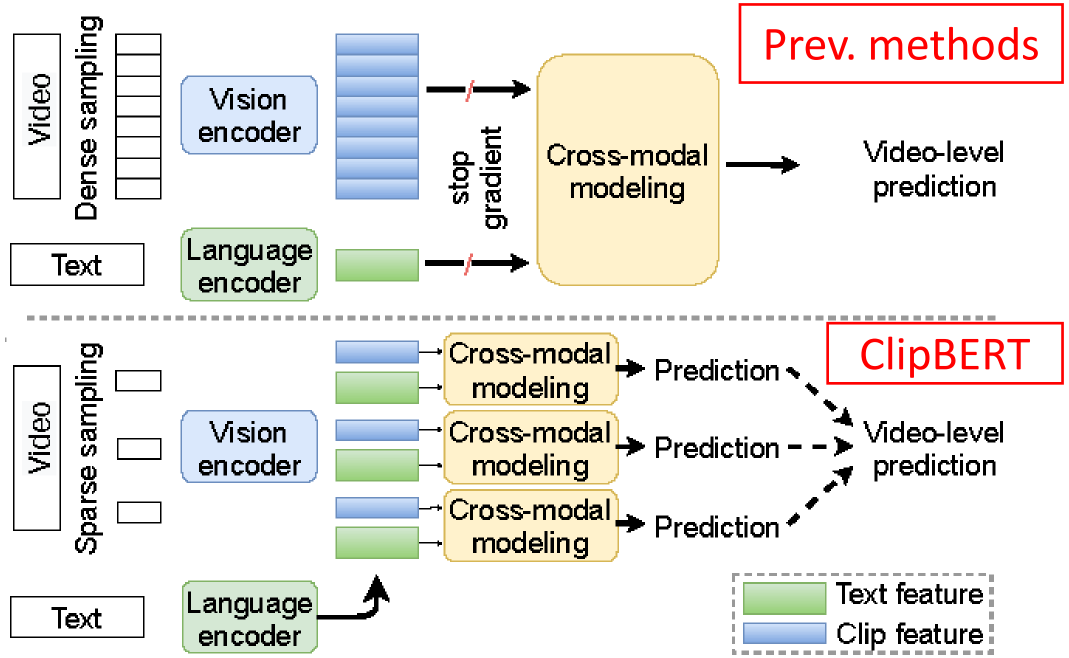
Video captioning

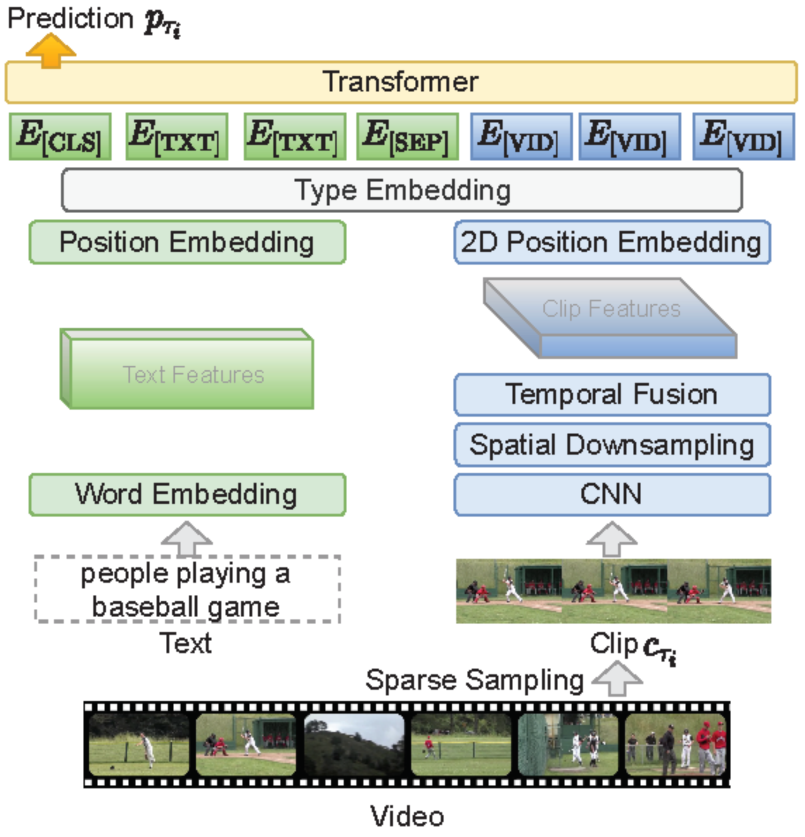
Video question answering

Zero-shot action classification

Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." ICCV'19.

CLIPBERT: video language pre-training with sparse sampling





Procedure:

- Pretraining on large-scale image-text datasets.
- Finetuning on video-text tasks.

ClipBERT overview

Lei, Jie, et al. "Less is more: Clipbert for video-and-language learning via sparse sampling." *CVPR'21.*

From short-form Video QA to Movie QA

Long-term temporal relationships

Multimodal inputs

Subtitle:

00:00:0.395 --> 00:00:1.896
(Keith:) I'm not gonna stand here and let you accuse me

00:00:1.897 --> 00:00:4.210
(Keith:) of killing one of my best friends, all right?

00:00:8.851 --> 00:00:10.394
(Castle:) You hear that sound?

Question: What did Keith do when he was on the stage?

Choice 1: Keith drank beer
Choice 2: Keith played drum
Choice 3: Keith sing to the microphone
Choice 4: Keith played guitar
Choice 5: Keith got off the stage and walked out

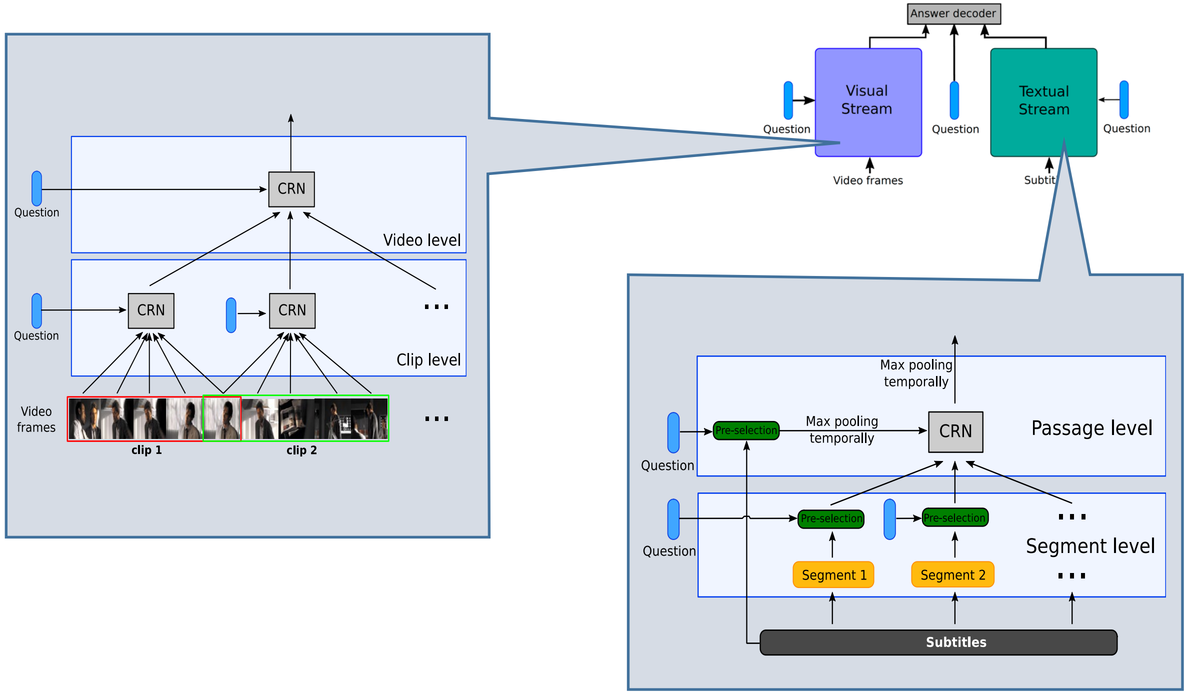
Baseline: Keith played guitar
HCRN: Keith got off the stage and walked out
Ground truth: Keith got off the stage and walked out

Lei, Jie, et al. "Tvqa: Localized, compositional video question answering." *EMNLP'*18.

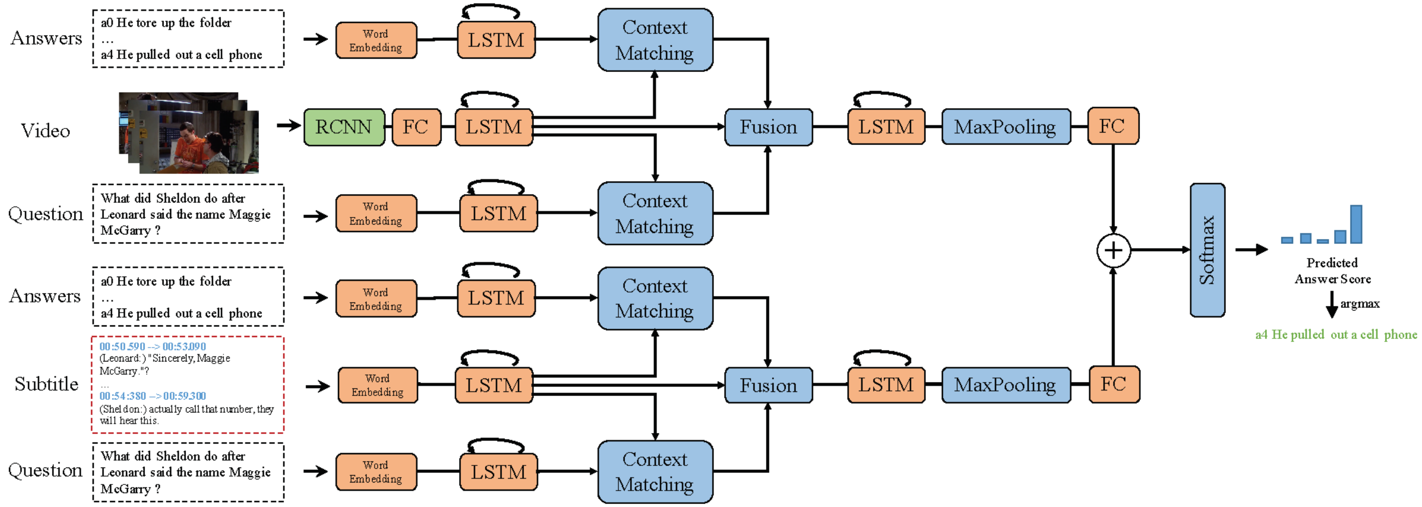
Conventional methods for Movie QA

Question-driven multi-stream models:

- Short-term temporal relationships are less important.
- Long-term temporal relationships and multimodal interactions are key.
- Language is dominant over visual counterpart.



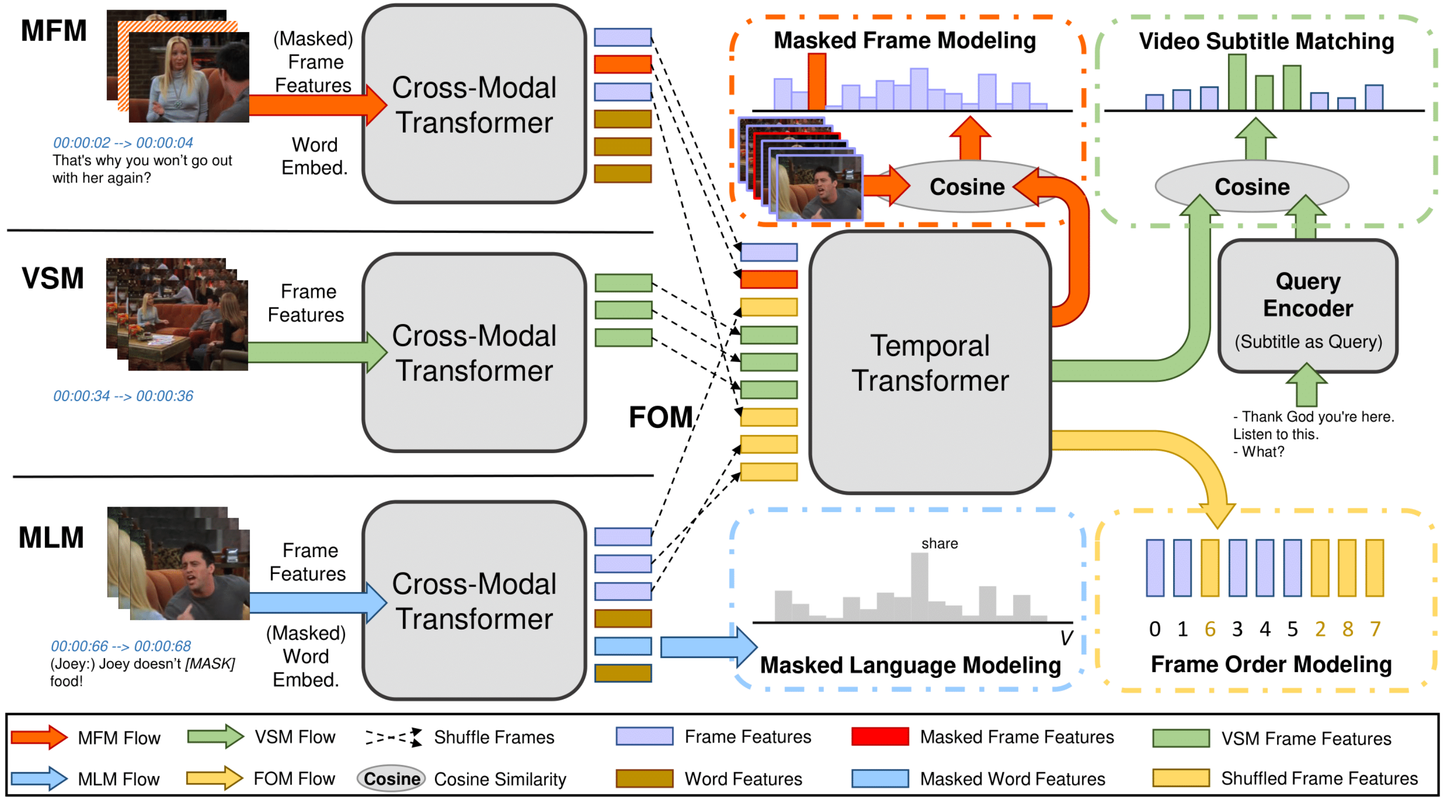
Le, Thao Minh, et al. "Hierarchical conditional relation networks for video question answering." IJCV'21.



Lei, Jie, et al. "Tvqa: Localized, compositional video question answering." *EMNLP'*18.

HERO: large-scale pre-training for Movie QA

- Pre-trained on 7.6M videos and associated subtitles.
- Achieved state-of-the-art results on all datasets.



Method \Task	TVR			How2R			TVQA	How2QA	VIOLIN	TVC			
	R@1	R@10	R@100	R@1	R@10	R@100	Acc.	Acc.	Acc.	Bleu	Rouge-L	Meteor	Cider
SOTA Baseline	3.25	13.41	30.52	2.06	8.96	13.27	70.23	-	67.84	10.87	32.81	16.91	45.38
HERO	**6.21**	**19.34**	**36.66**	**3.85**	**12.73**	**21.06**	**73.61**	**73.81**	**68.59**	**12.35**	**34.16**	**17.64**	**49.98**

Li, Linjie, et al. "Hero: Hierarchical encoder for video+ language omni-representation pre-training." *EMNLP'20*.

End of Lecture 7+8+9

https://neuralreasoning.github.io