Visual Question Answering and Visual Reasoning

Thao Minh Le

Oct 3, 2020

Joint work with **Vuong Le**, **Svetha Venkatesh** and **Truyen Tran**.

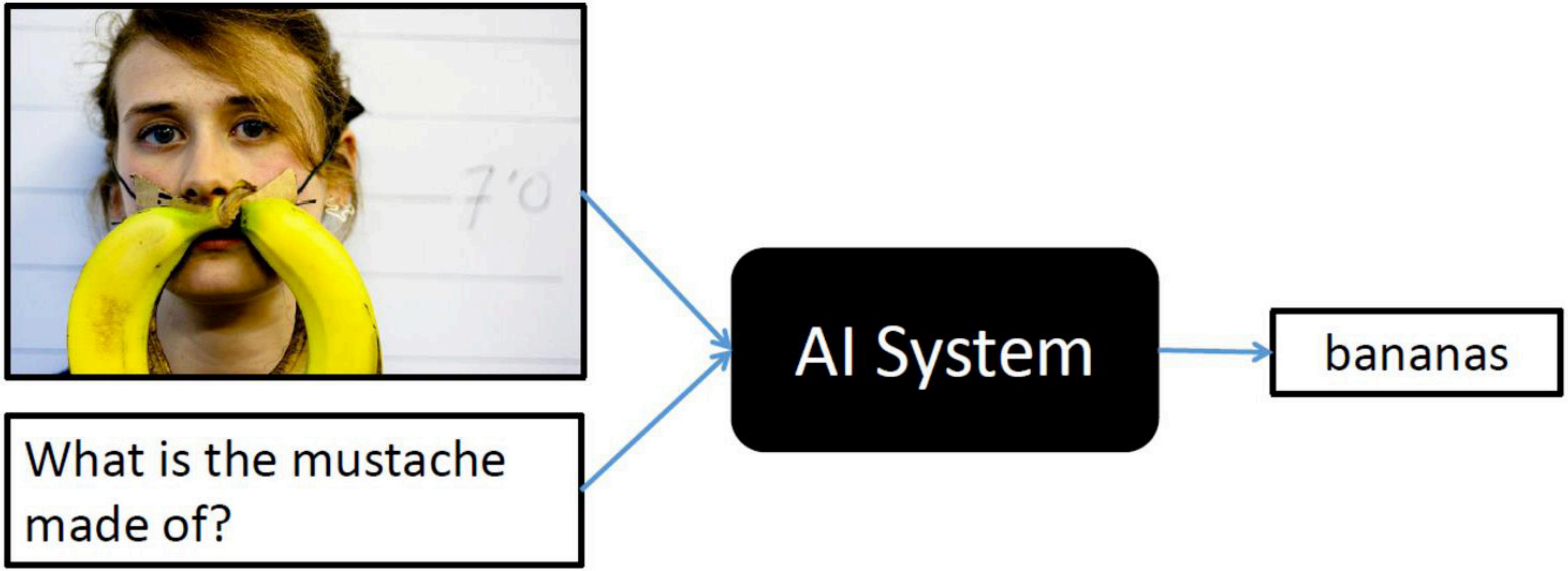
APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

Agenda

- Task overview
- Common approaches
- Dynamic language binding in relational visual reasoning
- Hierarchical conditional relation networks for video question answering
- Summary

Task overview



Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE international conference on computer vision*. 2015.

Try VQA demo by Georgia Tech

https://vqa.cloudcv.org/

Motivation: Why vision + language?

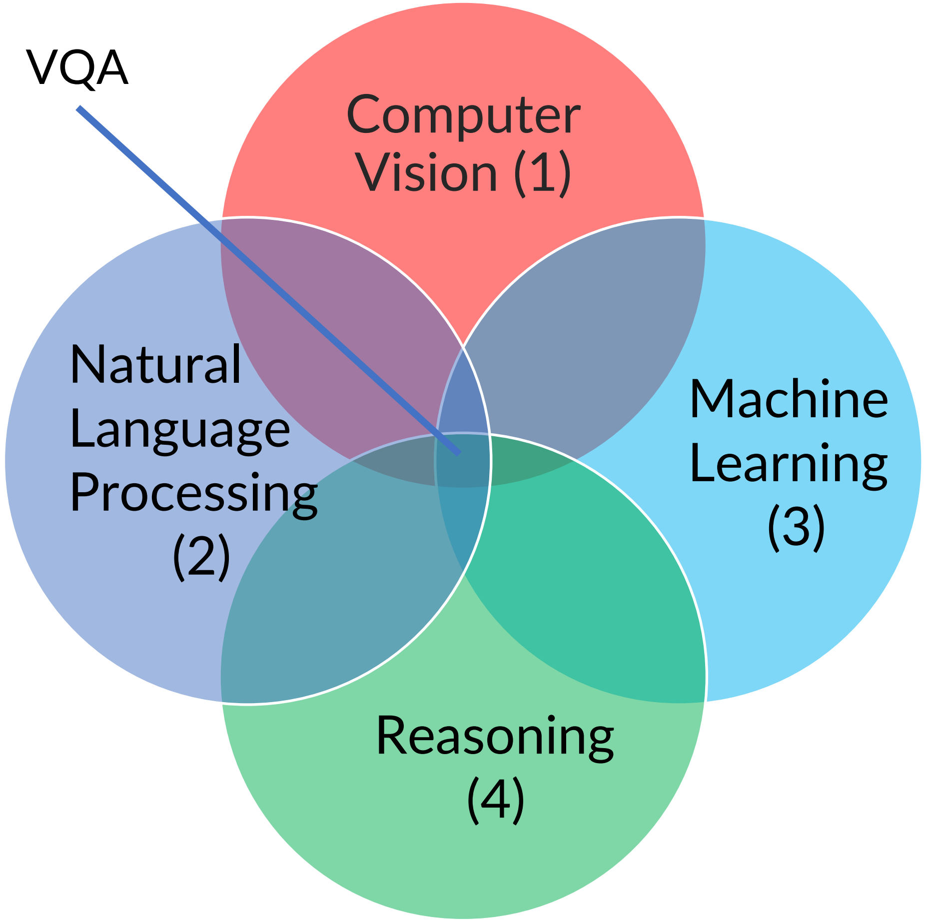
- Pictures/videos are everywhere.
- Words are how humans communicate.

Motivation: AI research

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

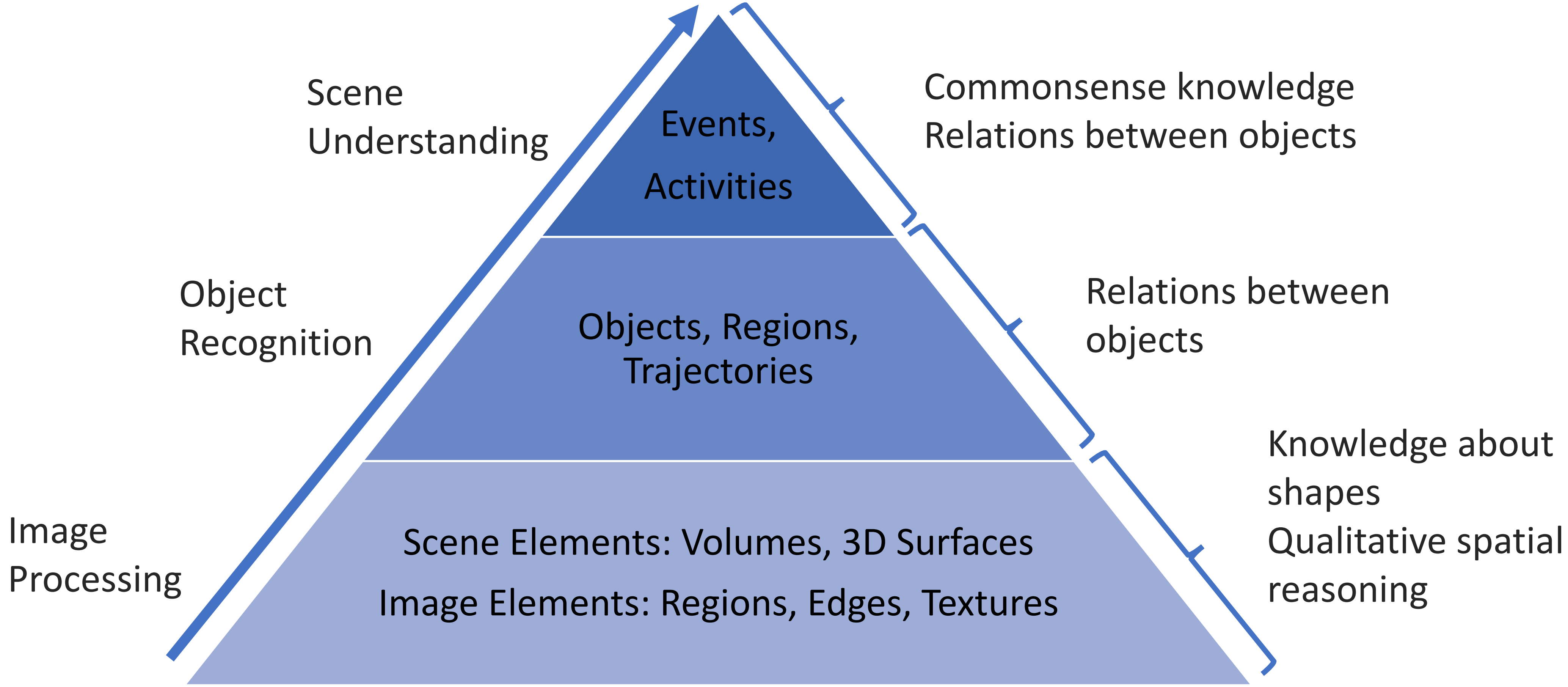
Answer: Firefighting

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

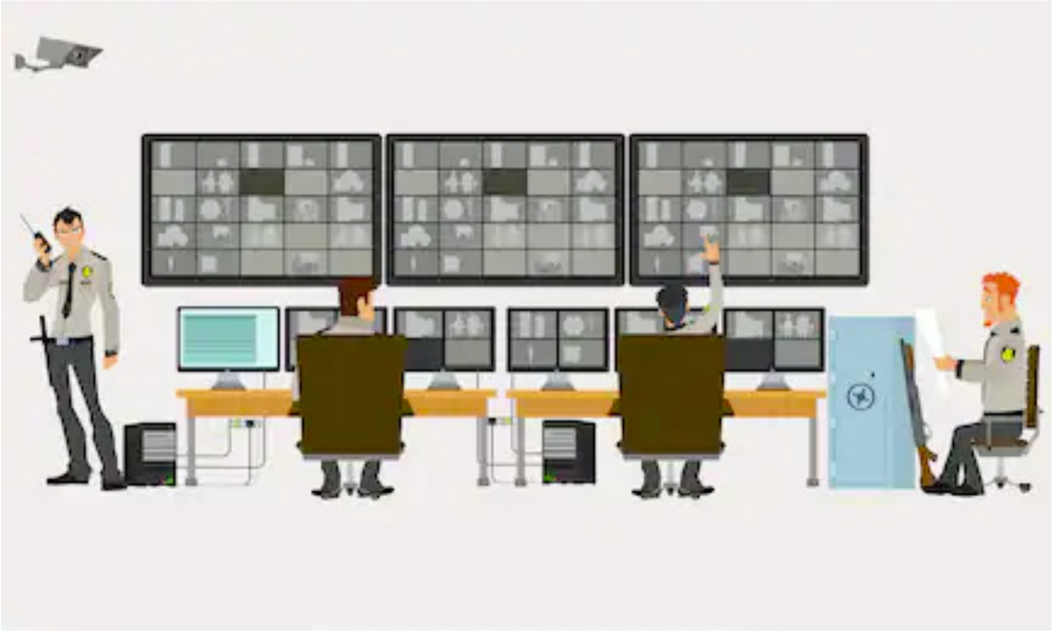
Image credit: ARIA

Applications of VQA

- Surveillance and visual data summarization

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

Image credit: journalistsresource.org



shutterstock.com • 289173068

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?

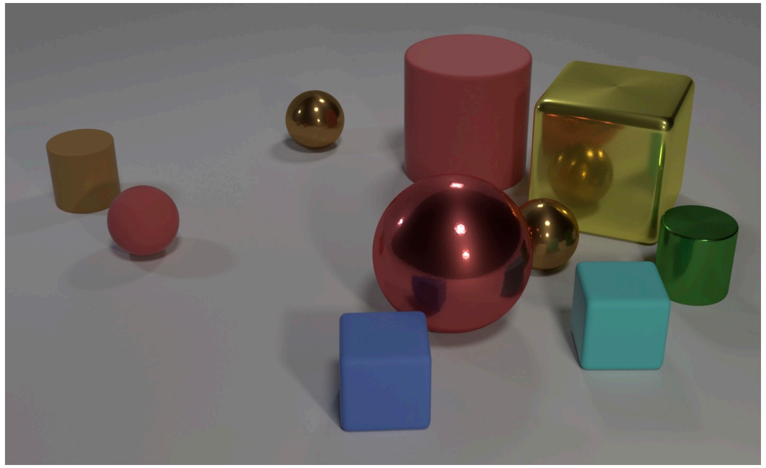
Perception

(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)

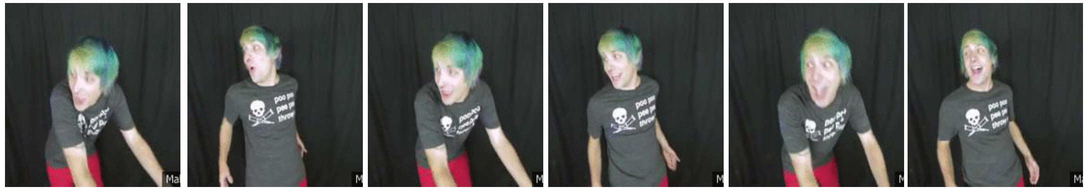


(Q) How many objects are either small cylinders or metal things?
(Q) Are there an equal number of large things and metal spheres?

Multi-step reasoning

VQA: Video QA datasets

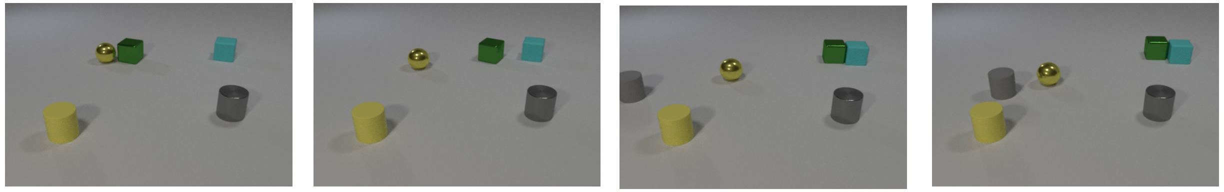
(TGIF-QA, Jang et al., 2018)



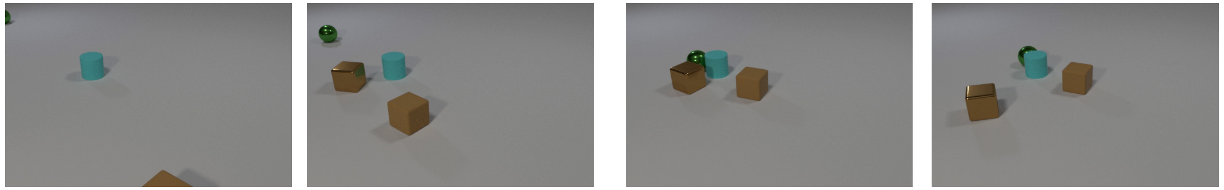
Q: What does the man do 5 times?
A: (0) step
(2) sway head
(3) bounce
(4) knod head
(5): move body to the front

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

VQA: Common approach

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

Visual feature extraction

Question
How many horses are in this image?

Embedding

Merge

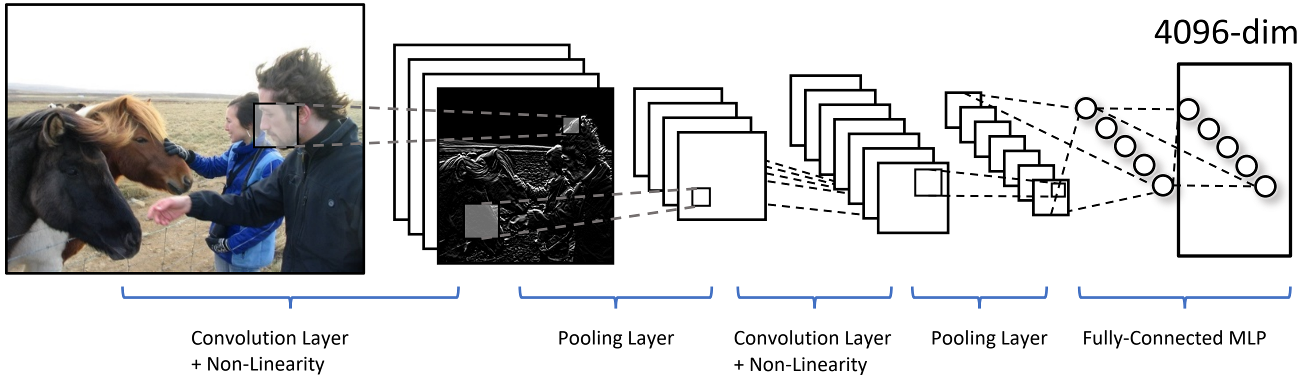
Answer prediction

Answer
2

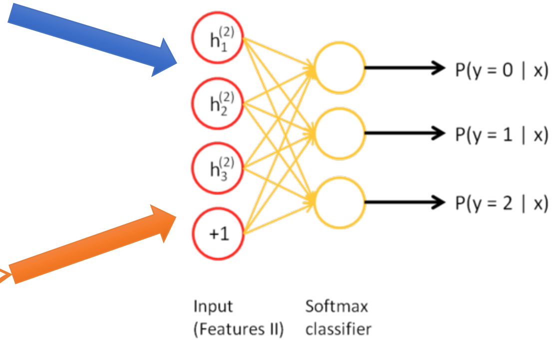
A very large body of methods have been proposed to solve VQA!

[VQA, Agrawal et al., 2015]

Image Embedding (VGGNet)

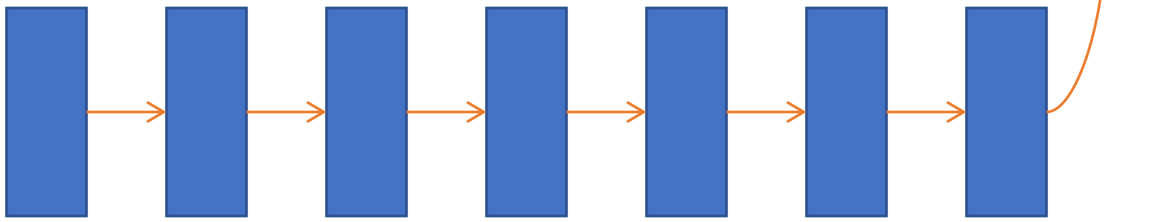


Neural Network
Softmax
over top K answers



Question Embedding (LSTM)

"How many horses are in this image?"

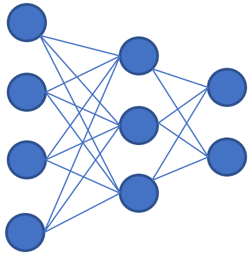


Relational Reasoning in Image QA

Thao Minh Le, Vuong Le, Svetha Venkatesh and Truyen Tran, "Dynamic Language Binding in Relational Visual Reasoning", *To appear at IJCAI'20*.

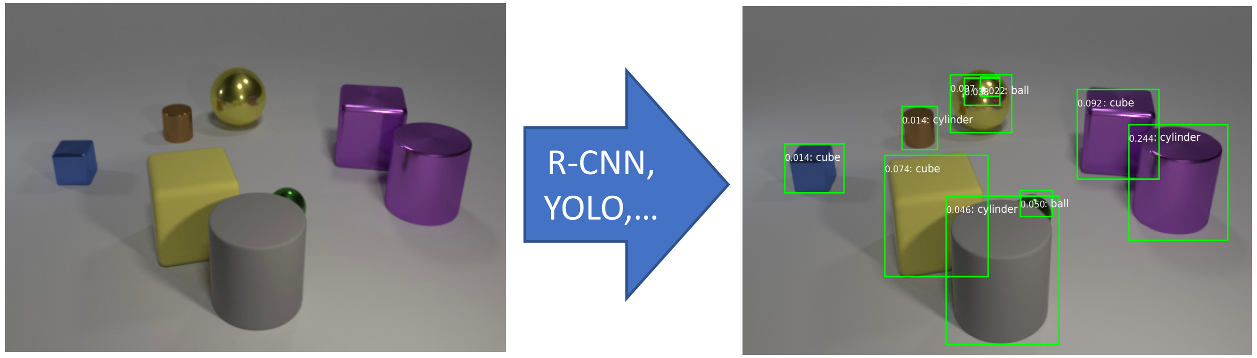
Our focus: Visual reasoning

From recognition to visual reasoning



a cat

What is this?



Where do the objects locate?

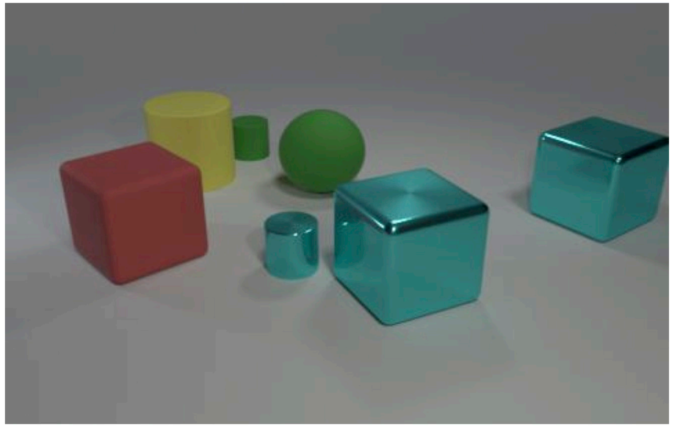
Object recognition

Object detection

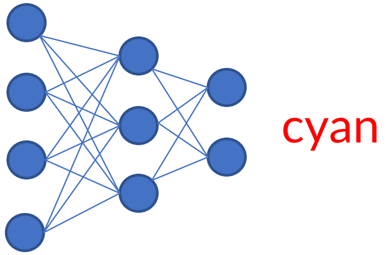
Image courtesy: https://dcist.com/

Our focus: Visual reasoning

Why things do not go well?



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



- 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).

Reasoning is to deduce knowledge from previously acquired knowledge in response to a query (or a cue) [Roni et al., 1997]

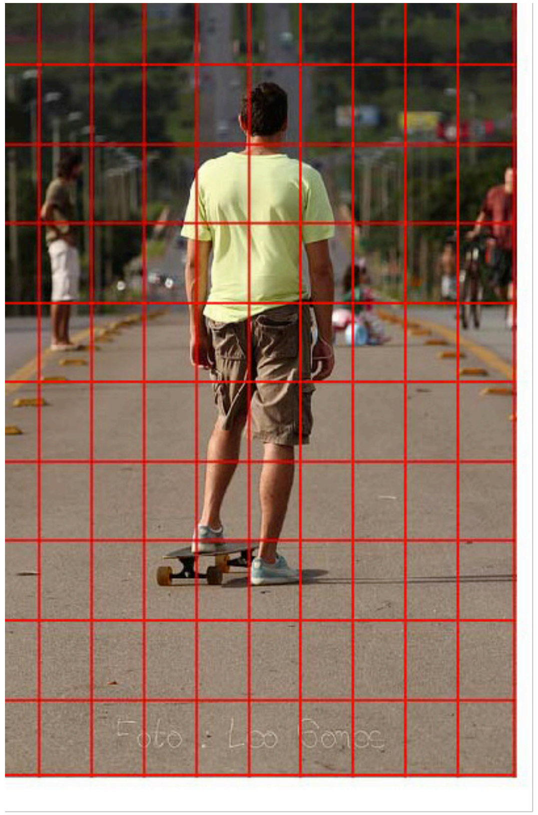
Reasoning with structured representation of spatial relations

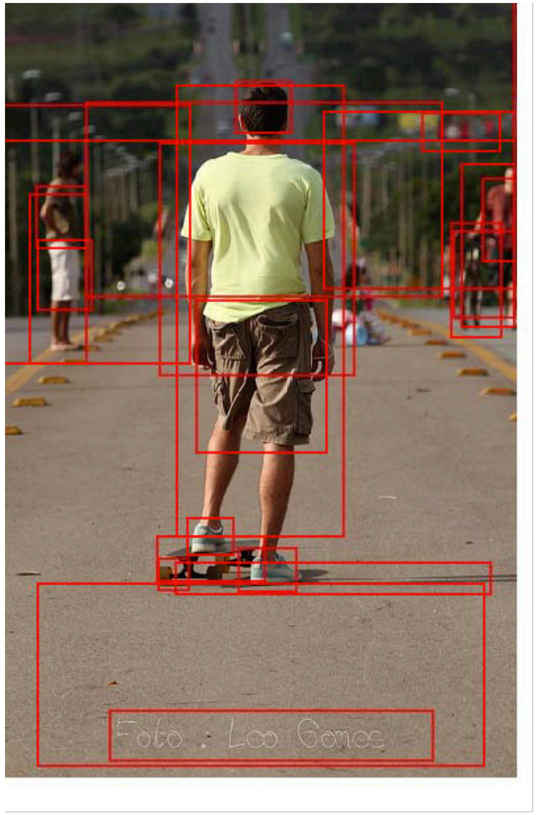
Key insight: *Reasoning is chaining of relational predicates to arrive at a final conclusion*

- Needs to uncover spatial relations, conditioned on query (query-conditioned scene graph).
- Chaining is query-driven
- Objects/language need(s) binding
- Everything is end-to-end differentiable

From grid features to region features

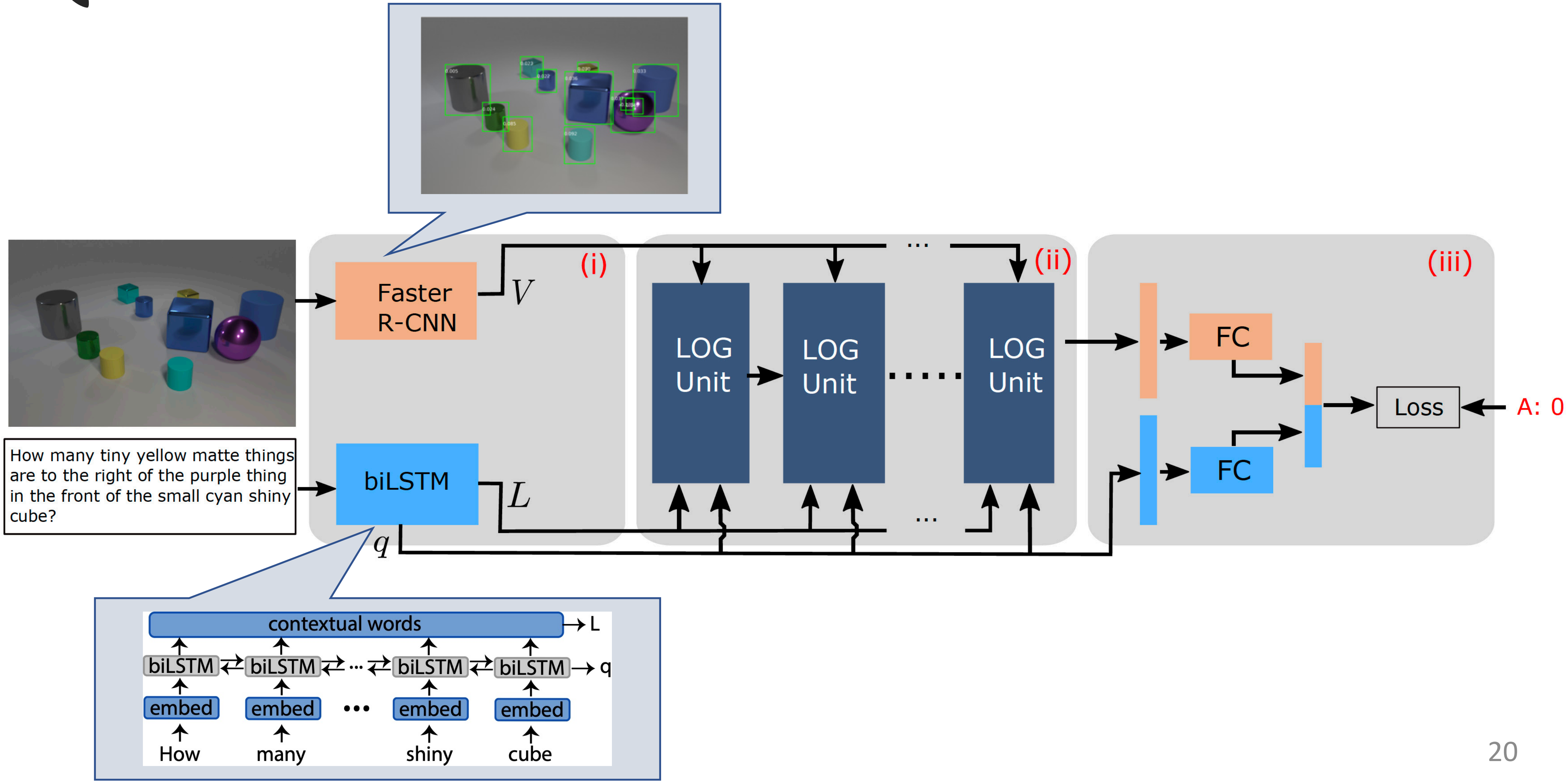
- Grid representation is irrespective of the fine-grained semantics of images.
- Region proposals are of the same semantic abstract with words.
- Interpretability.



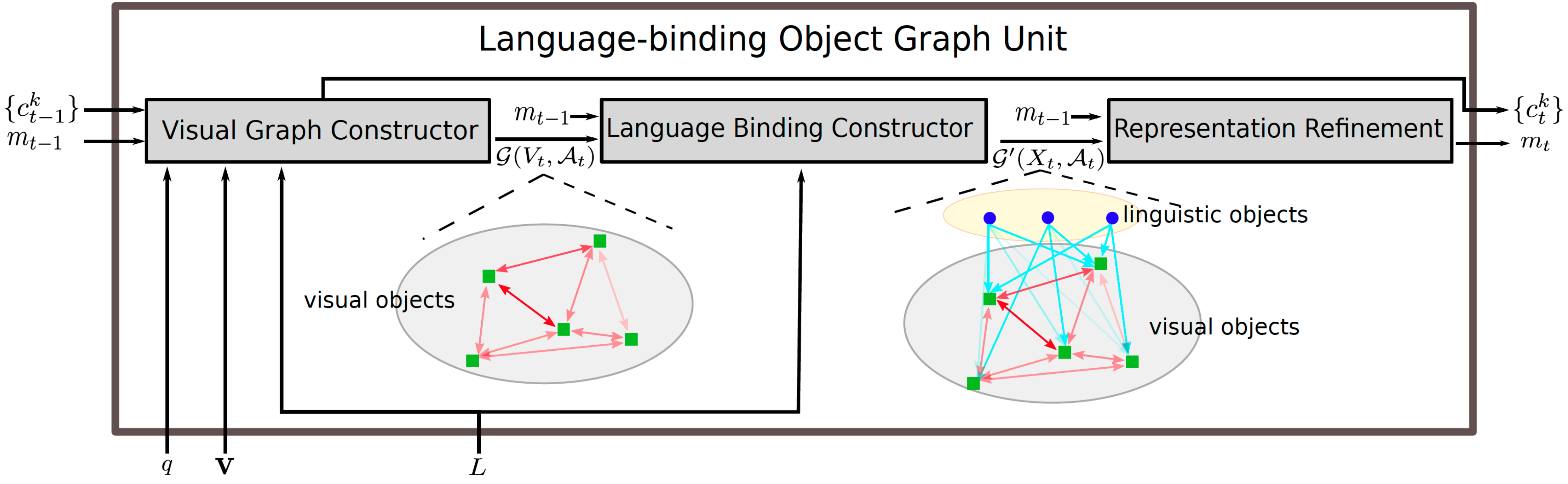


Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *CVPR'*18.

Language-binding Object Graph Model for VQA

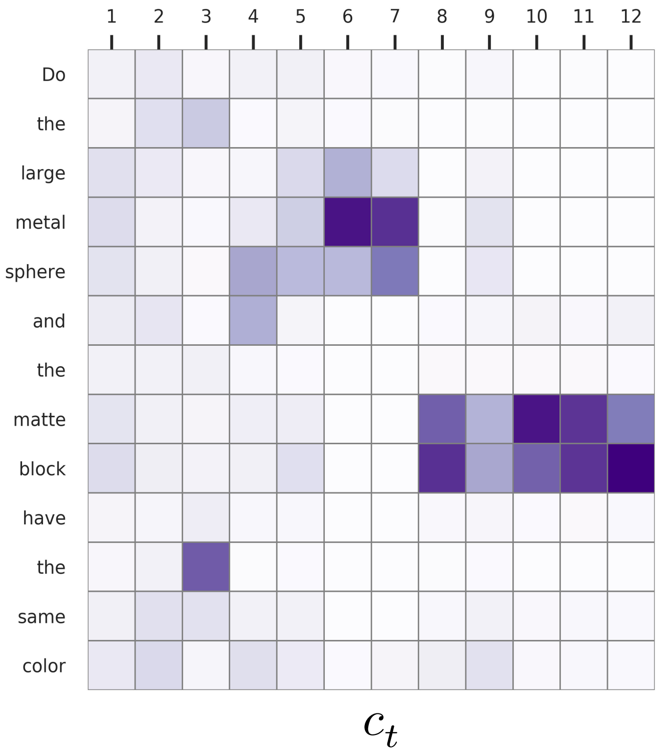


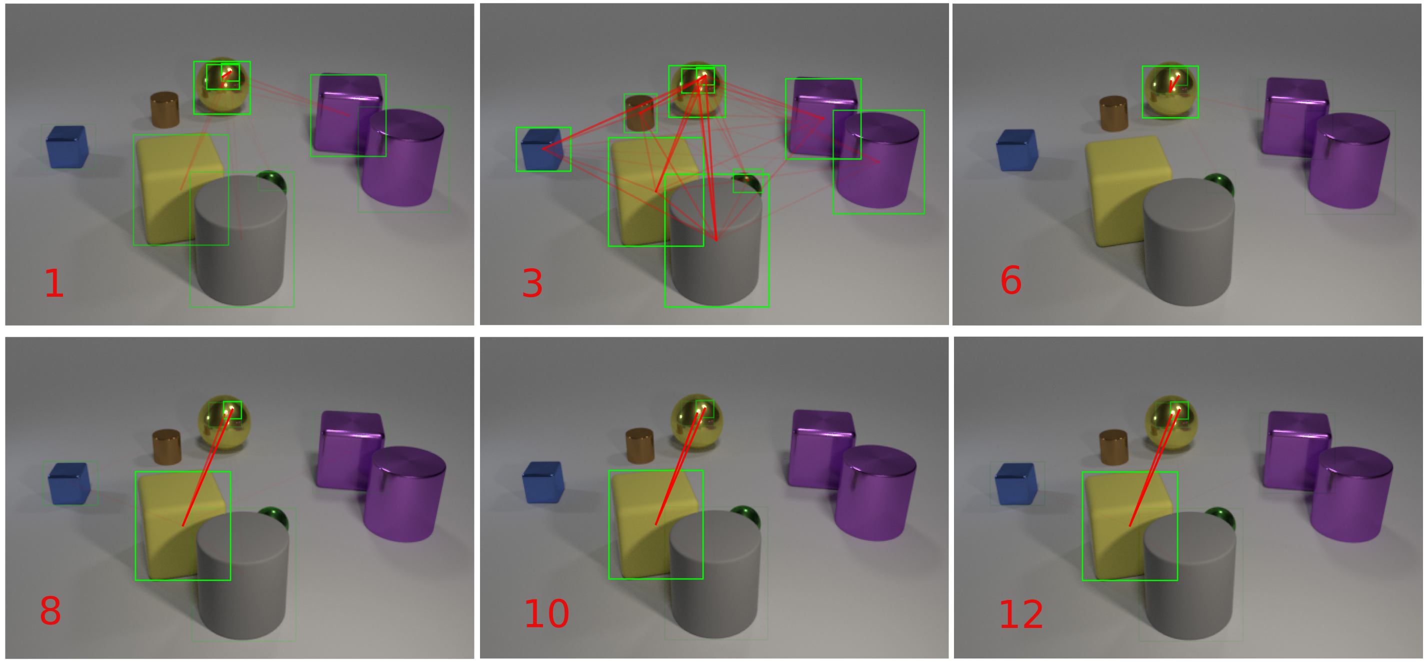
Language-binding Object Graph Unit (LOG)



Visual Graph Constructor

- c_t: question subset for t-th LOG unit.
- V: set of visual objects.
- $\bar{V}_t = V \odot c_t$: language-driven visual features.
- $A_t = \bar{V}_t^{\intercal} \bar{V}_t$: adjacency matrix





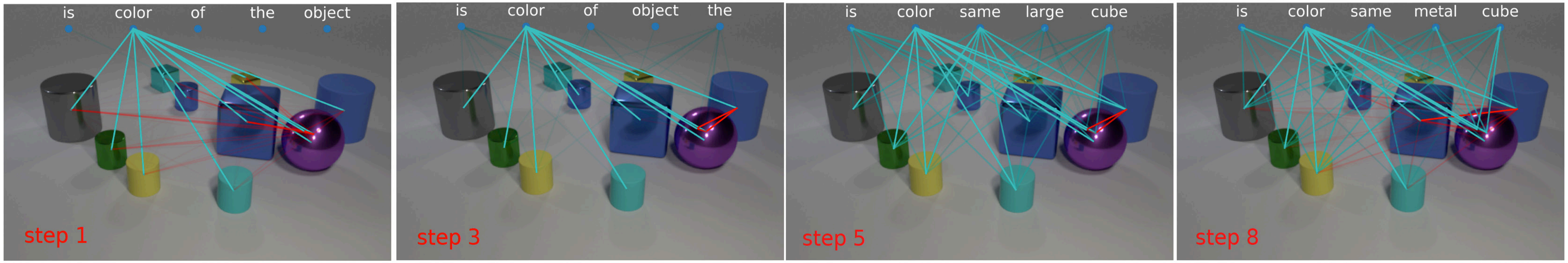
Language binding constructor

- Visual representation depends on context (query).
  - E.g: “What is the man holding a glass of wine wearing?” in a scene of multiple men visible.
- Update node representation in consideration of word bindings.

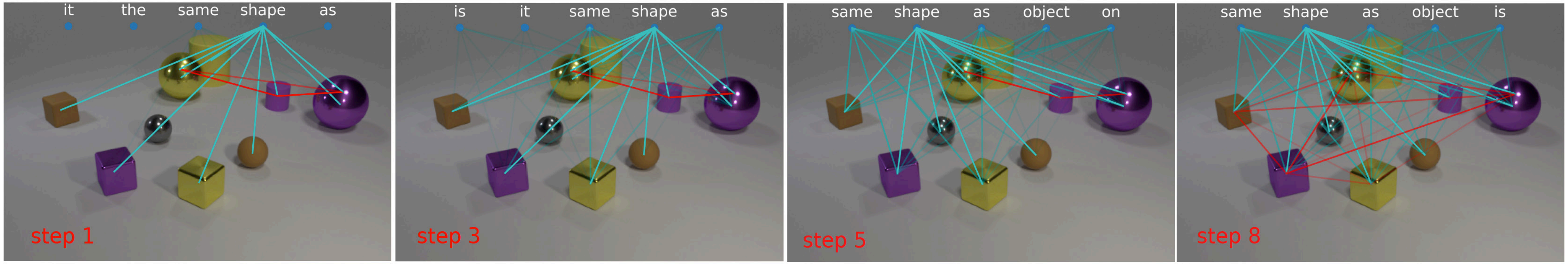
Node representation refinement

- Update the representation of a node based on information of its neighbors.
- Utilize skip-connect graph convolutional network.

LOGNet's Output

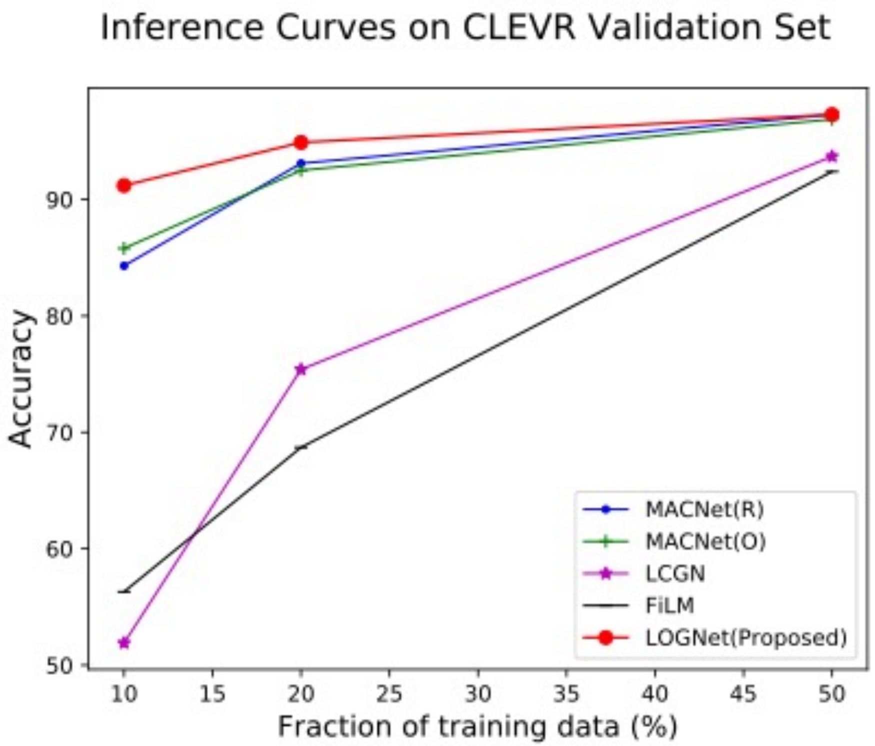


Question: Is the color of the big matte object the same as the large metal cube?
Prediction: yes **Answer**: yes



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

Results



Comparison with SOTAs on CLEVR dataset of different data fractions.

Method	Val. Acc. (%)
FiLM	56.6
MACNet(R)	57.4
LCGN [Hu *et al.*, 2019]	46.3
BAN [Shrestha *et al.*, 2019]	60.2
RAMEN [Shrestha *et al.*, 2019]	57.9
LOGNet	**62.3**

Performance comparison on CLEVR-Human.

Results (Cont.)

Method	Accuracy (%)	
	val	test
Full training data		
CNN+LSTM	49.2	46.6
Bottom-Up [Anderson *et al.*, 2018]	52.2	49.7
MACNet(O)	57.5	54.1
LCGN [Hu *et al.*, 2019]	63.9	56.1
LOGNet	63.3	55.2
Subset 50% training data		
LCGN	60.6	-
LOGNet	60.7	-
Subset 20% training data		
LCGN	53.2	-
LOGNet	55.6	-

Performance on GQA

Method	Val. Acc. (%)
XNM [Shi *et al.*, 2019]	43.4
MACNet(R)	40.7
MACNet(O)	45.5
LOGNet	**46.8**

Performance on

VQA v2 subset of long questions

Relational Reasoning in Video QA

Thao Minh Le, Vuong Le, Svetha Venkatesh and Truyen Tran, "Hierarchical conditional relation networks for video question answering", Appeared at *CVPR'20 (Oral)*.

From Image QA to Video QA

- Understanding **temporal reasoning** in addition to **visual reasoning**.
- Videos are **richer** than images, can be incorporated with **additional channels** such as subtitles, speech etc.

Q1: What does the boy with a brown hoodie do before running away ? ***A**: flip to the front side*

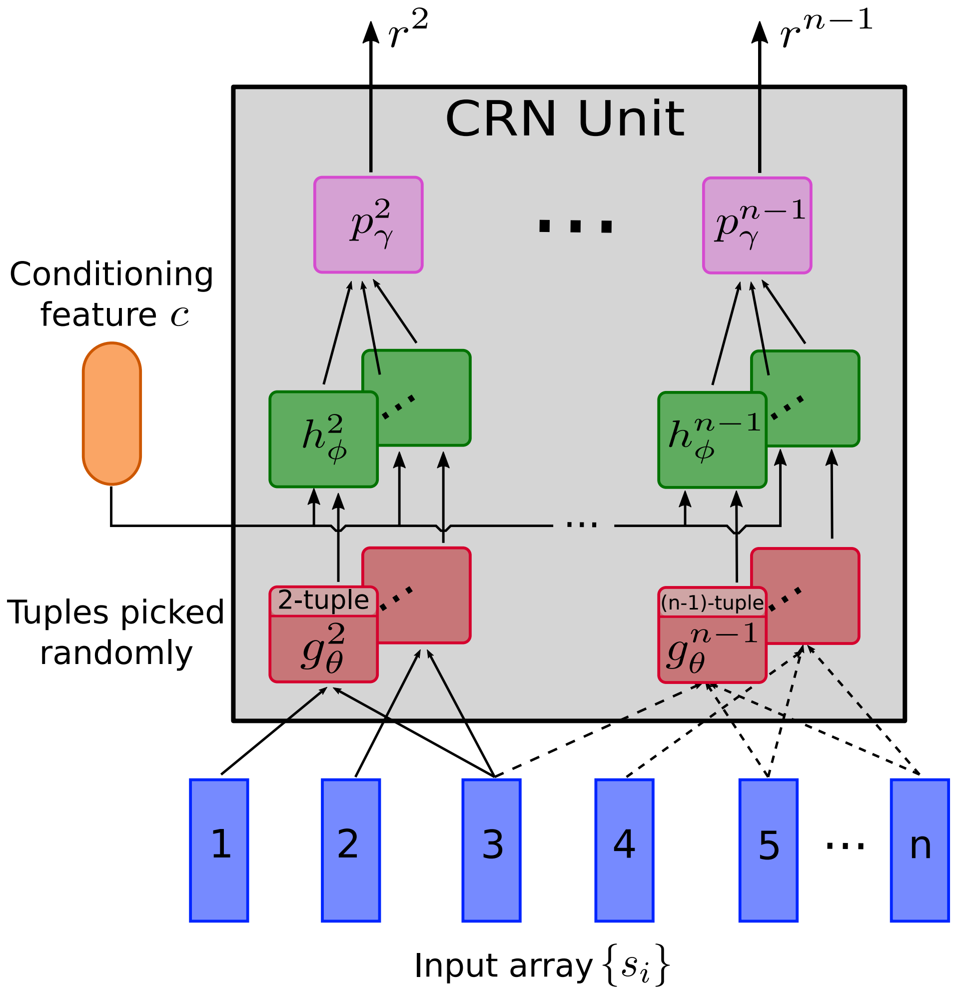
Q2: What does the boy with a brown hoodie do after flipping to the front side? ***A**: run away*

Q3: Where is the boy with brown jacket running? ***A**: street*

We aim for a reasoning engine that

- Effectively reflects the **long-short temporal relation**, **hierarchy**, **compositionality** of videos.
- Be readily extended to handle **additional information channels**.
- Eases the model building process by simple **rearrangements** and **block stacking** with a **generic unit** similar to most of the breakthough neural architectures.

Conditional Relation Network Unit (CRN)



Algorithm 1: CRN Unit

Input : Array $\mathcal{S} = \{s_i\}_{i=1}^n$, conditioning feature c

Output : Array R

Metaparams : $\{k_{\max}, t \mid k_{\max} < n\}$

Build all sets of subsets $\{Q^k \mid k = 2, 3, ..., k_{\max}\}$ where Q^k is set of all size-k subsets of $\mathcal{S}$

Initialize $R \leftarrow \{\}$

for $k \leftarrow 2$ **to** $k_{\max}$ **do**

  Q^k_{selected} = randomly select t subsets from Q^k

  for each *subset* $q_i \in Q^k_{\text{selected}}$ **do**

    $g_i = g^k(q_i)$

    $h_i = h^k(g_i, c)$

  end

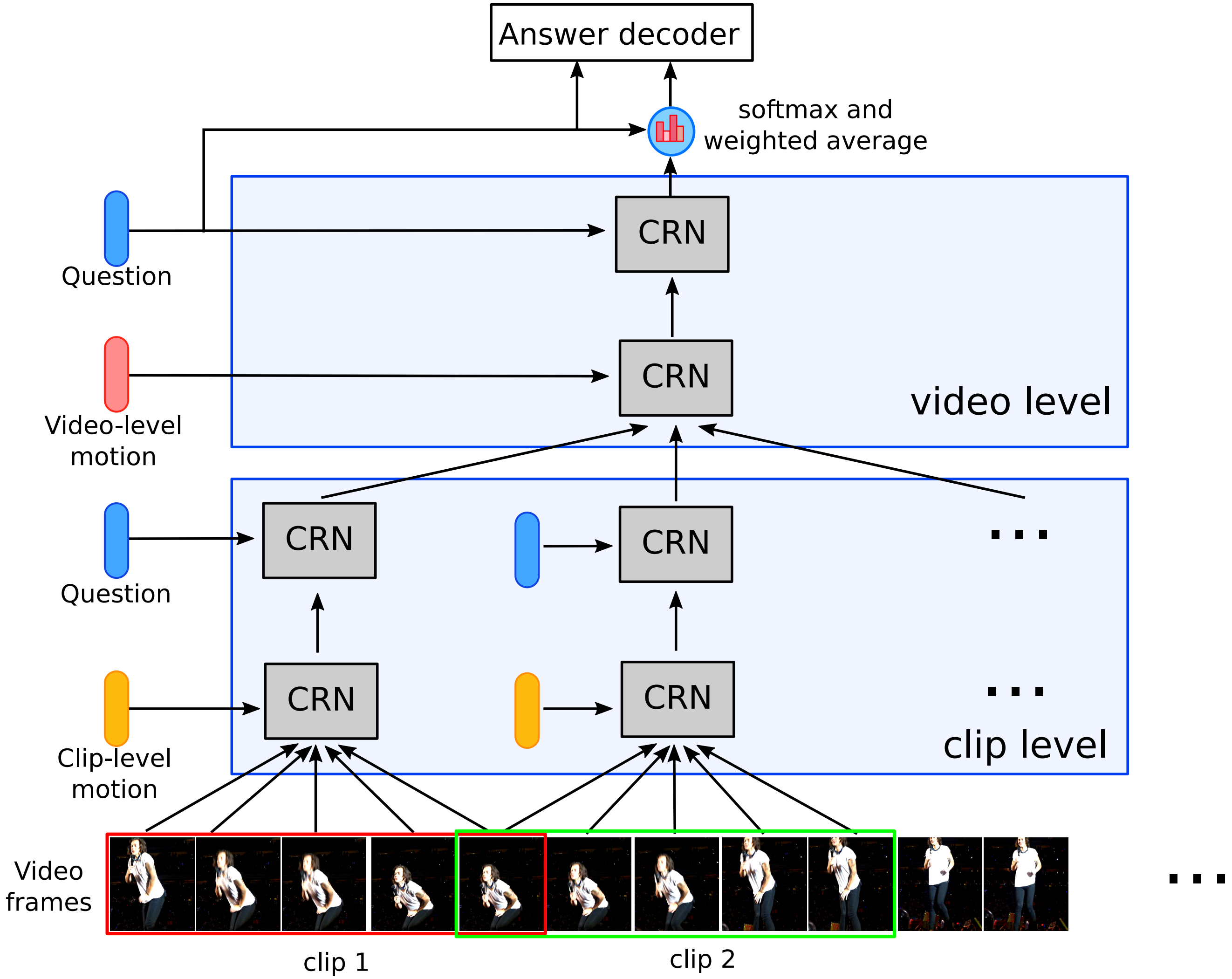
  $r^k = p^k(\{h_i\})$

  add r^k to R

end

Hierarchical Conditional Relation Networks for Video QA

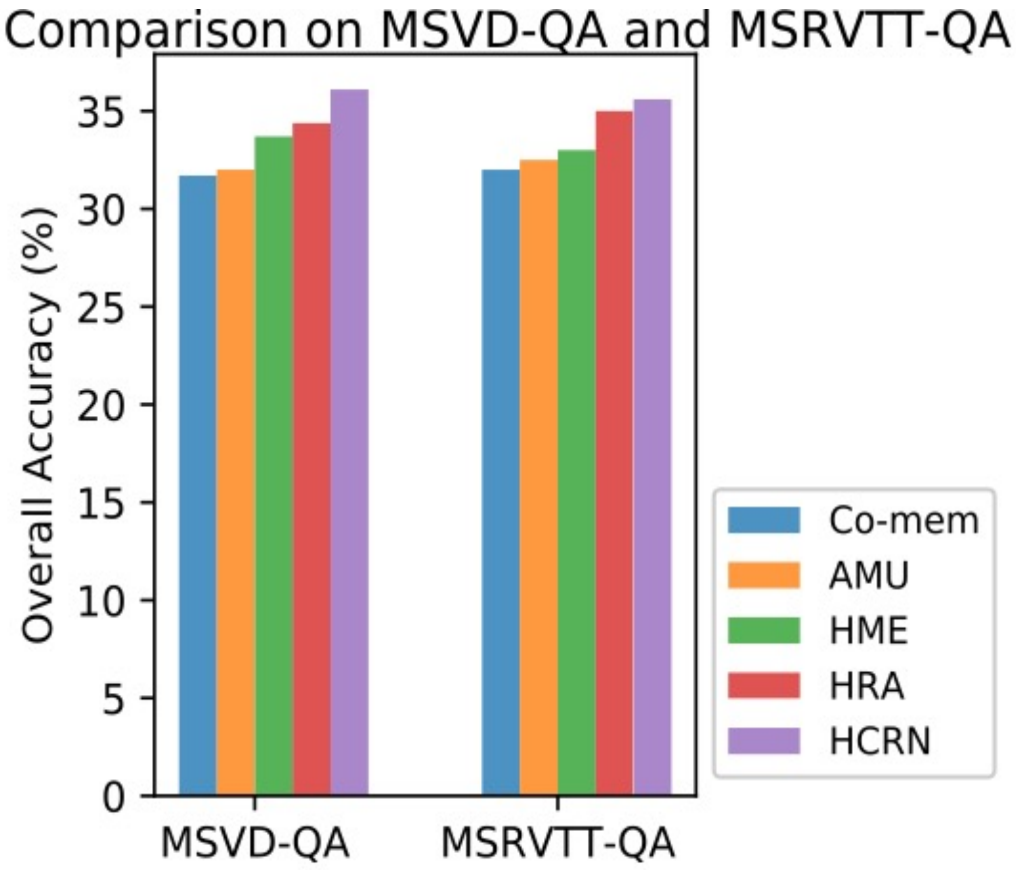
- Frame-wise appearance features: extracted by ResNet101 pretrained on ImageNet.
- Motion features: extracted by an 3D ResNet pretrained on Kinetics.
- Linguistic representation: a BiLSTM on GloVe word embeddings.



Results

Model	Action	Trans.	F.QA	Count
ST-TP	62.9	69.4	49.5	4.32
Co-Mem	68.2	74.3	51.5	4.10
PSAC	70.4	76.9	55.7	4.27
HME	73.9	77.8	53.8	4.02
HCRN	**75.0**	**81.4**	**55.9**	**3.82**

TGIF-QA dataset



Results (Cont.)

Model	Action	Trans.	F.QA	Count
Relations (k_{max}, t)				
$k_{max} = 1, t = 1$	65.2	75.5	54.9	3.97
Hierarchy				
1-level, video CRN only	66.2	78.4	56.6	3.94
Motion conditioning				
w/o motion	70.8	79.8	56.4	4.38
Linguistic conditioning				
w/o linguistic condition	66.5	75.7	56.2	3.97
Gating				
w/o gate	74.1	82.0	55.8	3.93
Full 2-level HCRN	75.1	81.2	55.7	3.88

Highlighted Ablation Studies on TGIF-QA dataset

Qualitative Results

Q: What does the girl do 9 times?

HCRN: blocks a person's punch

Baseline: walk

Q: What does the man do before turning body to left?

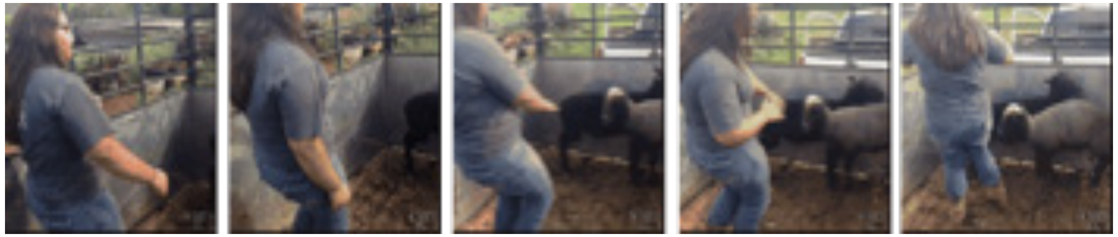
HCRN: breath

Baseline: pick up the man's head

Q: What does the person do after kissing finger?

HCRN: wave them at the camera

Baseline: sit



Q: How many times does the woman reach forward with her hands?

HCRN: 3

Baseline: 2

Extension: HCRN for Movie QA

Subtitle:

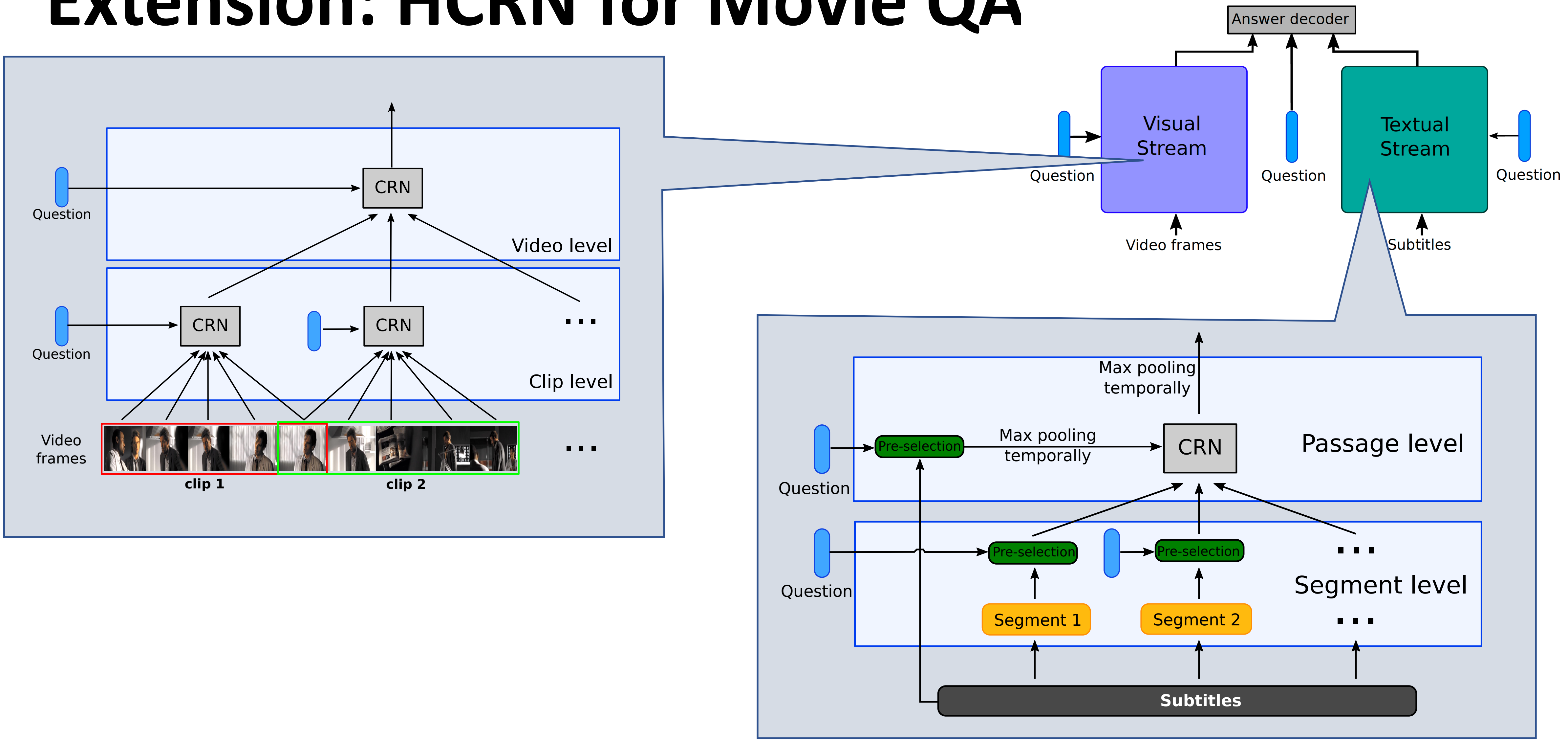
00:00:0,395 --> 00:00:1,896	00:00:1,897 --> 00:00:4,210	00:00:8,851 --> 00:00:10,394
(Keith:) I'm not gonna stand here and let you accuse me	(Keith:) of killing one of my best friends, all right?	(Castle:) You hear that sound?

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

Extension: HCRN for Movie QA



Future directions

- Generalization and consistency in VQA.
- Transformer-based methods to solve vision-and-language tasks, especially for video-and-language tasks.
- Object-centric for Video QA and video understanding. Object and event interaction are good in terms of algorithmic transparency and interpretability.
    - Deal with object tracking, relation across space-time, causality of events.

Summary

TL;DR:

- Introduction to VQA and its applications.
- Multi-step **relational reasoning** in Image QA.
- Difficulties when switching from Image QA to Video QA.
- A hierarchical network architecture reflecting **temporal relations**, **multimodal interactions** for Video QA.

The team @A2I2, Deakin University

- Lead: Prof. Svetha Venkatesh, A/Prof. Truyen Tran.
- One of the top Australian AI research institutes.
- Research on both AI fundamentals and applications.
- Fully scholarships available.
- Learn more at: https://a2i2.deakin.edu.au
  https://truyentran.github.io

THANK YOU!

Thao Minh Le

Email: lethao@deakin.edu.au

Personal Site: https://thaolmk54.github.io

Applied Artificial Intelligence Institute,

Deakin University, Australia