APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

Machines that learn to talk about what they see

A/Prof **Truyen Tran**
& Thao Minh Le
Deakin University

truyen.tran@deakin.edu.au
truyentran.github.io
@truyenoz
letdataspeak.blogspot.com
goo.gl/3jJ1O0

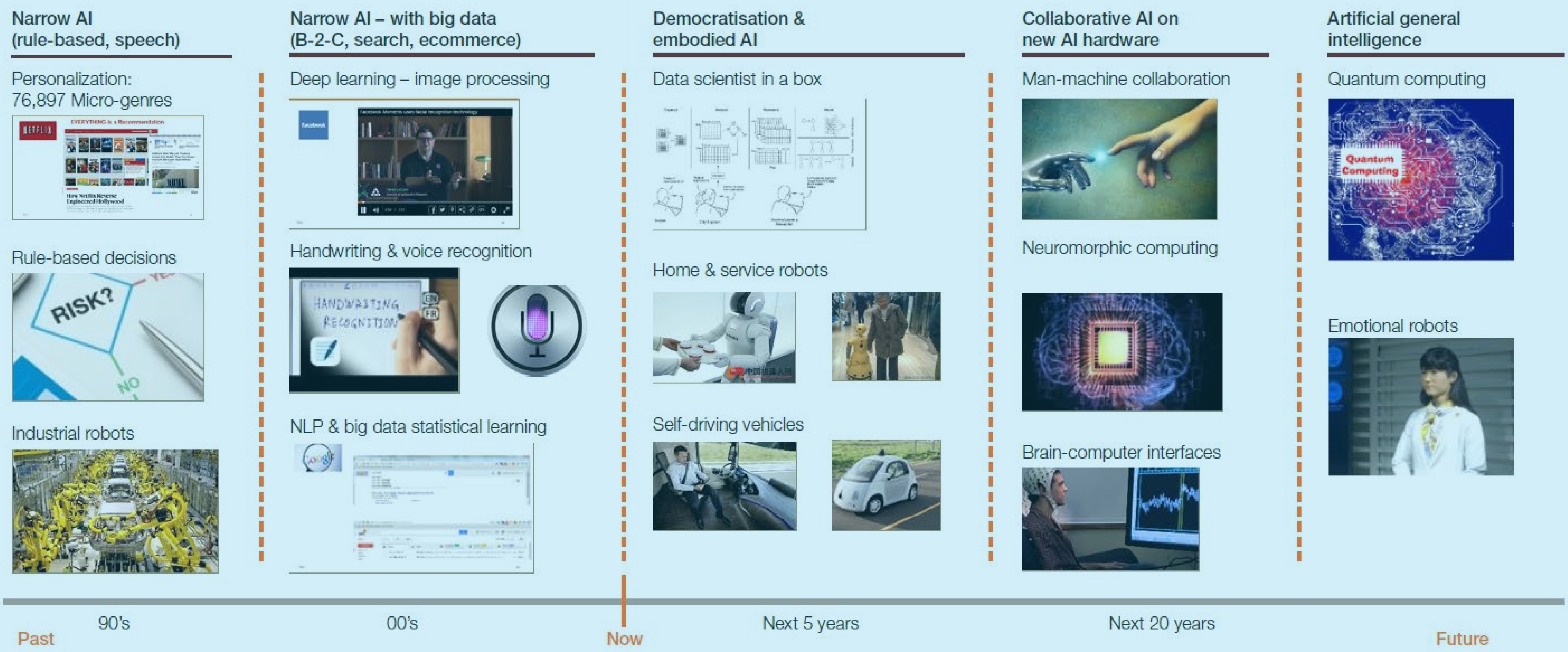
Hanoi, Dec 2019

AI is there but yet to be solved

Among the most challenging scientific questions of our time are the corresponding analytic and synthetic problems:

- How does the brain function ?
- Can we design a machine which will simulate a brain ?

-- *Automata Studies*, 1956.

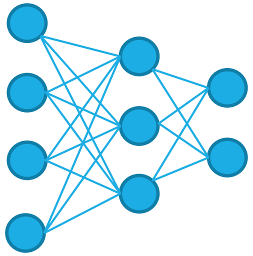

Narrow AI
(rule-based, speech)
Personalization:
76,897 Micro-genres
Rule-based decisions
RISK?
NO
Industrial robots
Narrow AI – with big data
(B-2-C, search, ecommerce)
Deep learning – image processing
Handwriting & voice recognition
HANDWRITING
RECOGNITION
NLP & big data statistical learning
Democratisation &
embodied AI
Data scientist in a box
Home & service robots
Self-driving vehicles
Collaborative AI on
new AI hardware
Man-machine collaboration
Neuromorphic computing
Brain-computer interfaces
Artificial general
intelligence
Quantum computing
Quantum
Computing
Emotional robots
Past
90's
00's
Now
Next 5 years
Next 20 years
Future

Source: PwC

Agenda

- Visual reasoning
- Neural memories
- Dual-process theories

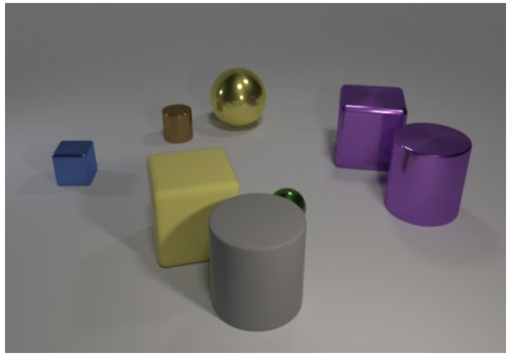
Visual recognition, high performance

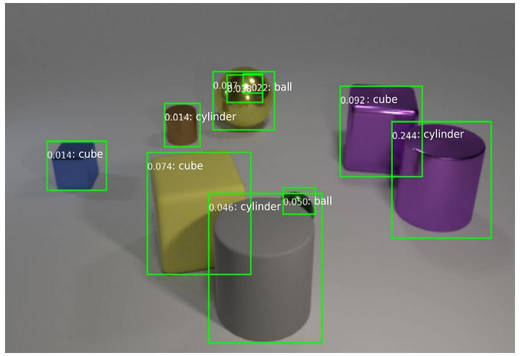


a cat

What is this?

Object recognition



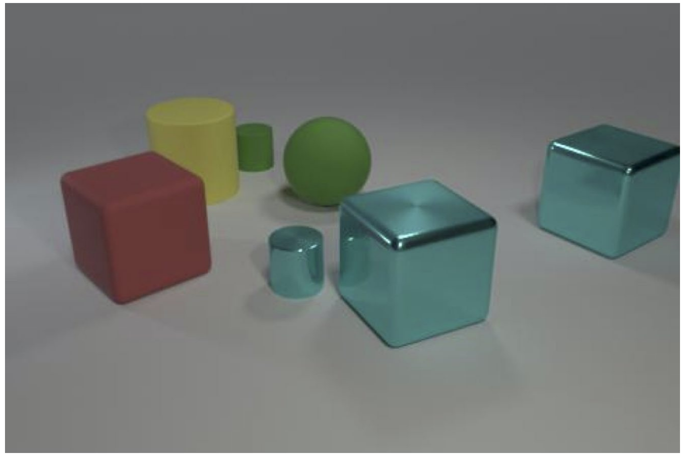


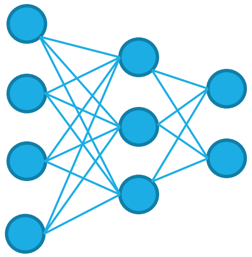
Where are objects, and what are they?

Object detection

Image courtesy: https://dcist.com/

Visual QA, not that realistic, yet

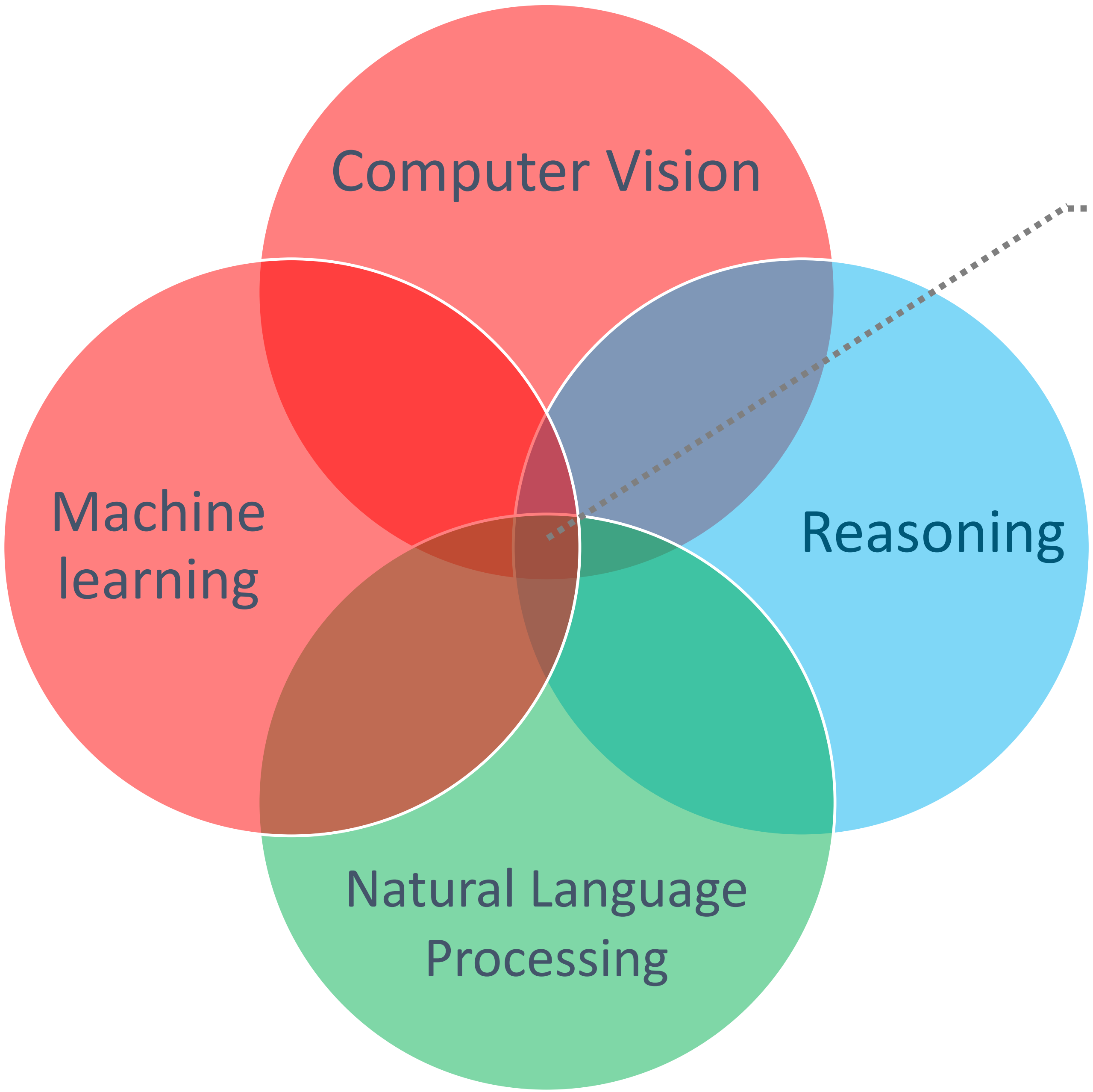




blue

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

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

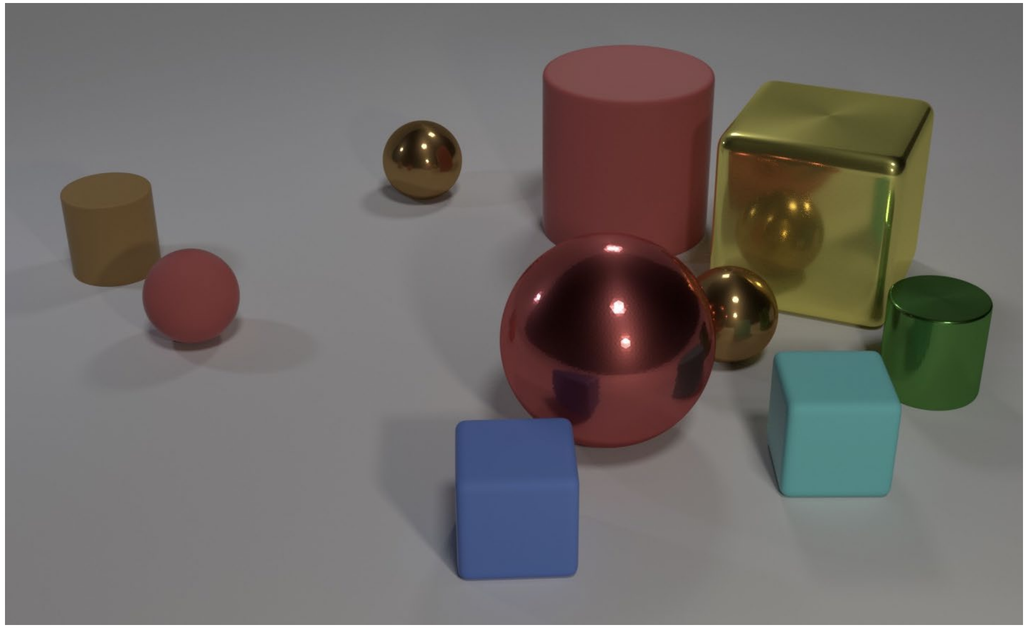


Visual Reasoning

Visual QA is a great testbed

Examples of Visual QA tasks

(CLERV, 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?

(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?

Examples of Video QA tasks

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

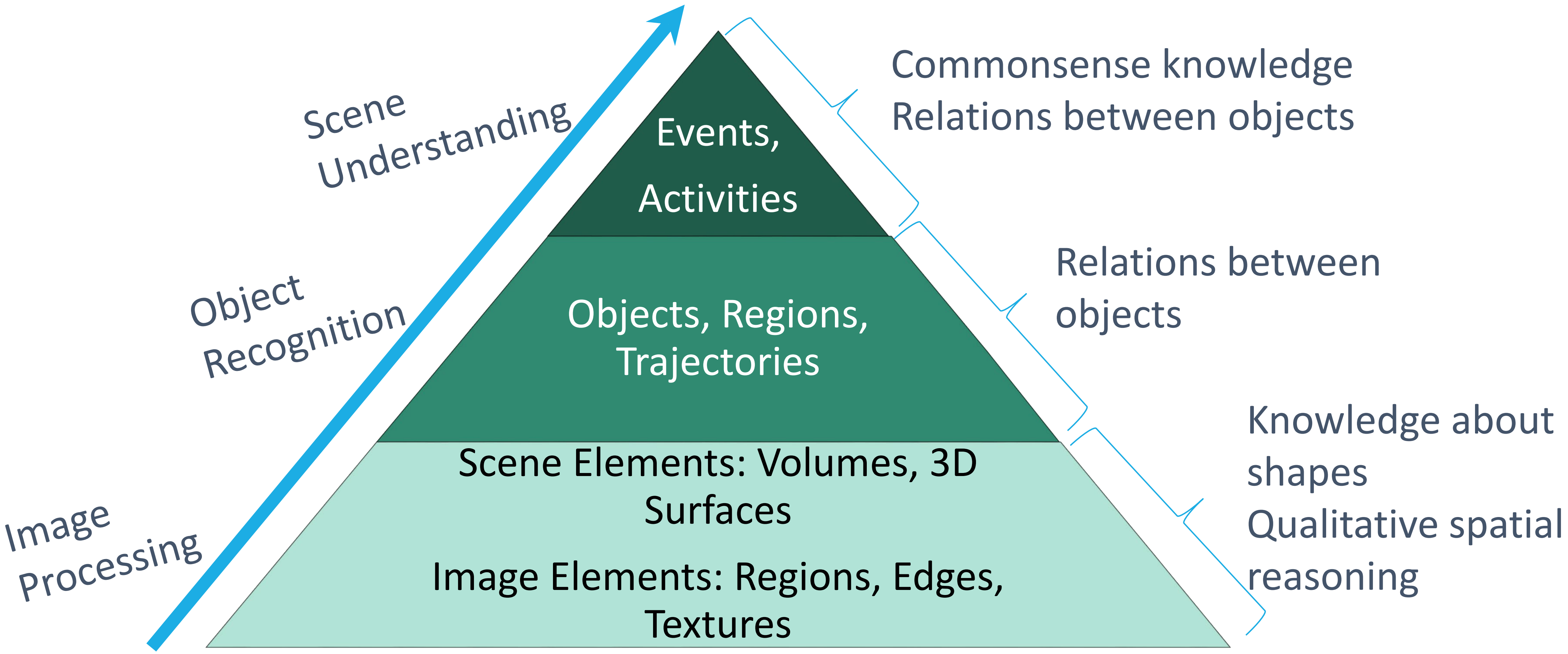
(2) pick up the man's hand (4) raise hand

(5): breath

(Data: TGIF-QA)

Reasoning over visual modality

- Context:
  - Videos/Images: **compositional** data
  - Hierarchy of visual information: objects attributes, **relations,** commonsense knowledge.
  - Relation of abstract objects: spatial, temporal, semantic aspects.
- Research questions:
  - How to represent **object relations** in images and videos?
  - How to reason with **object relations**?



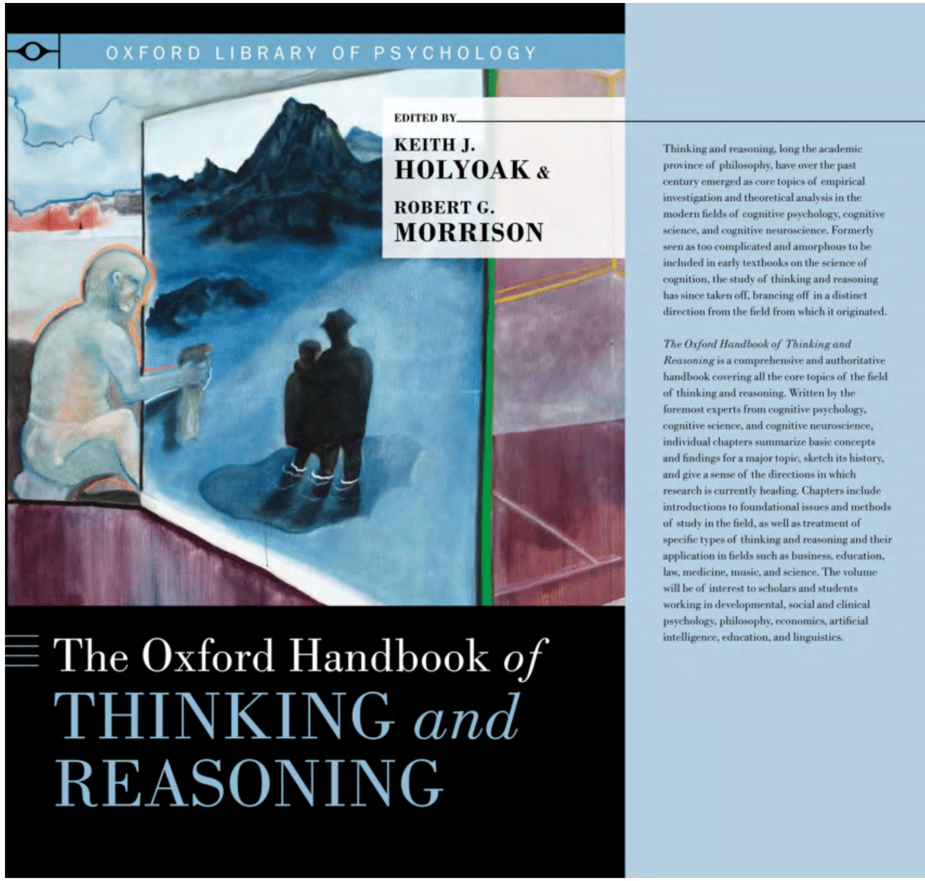
Adapted from (Aditya et al., 2019)

Many facets of reasoning

Analogical
Relational
Inductive
Deductive
Abductive
Judgemental
Causal

Legal
Scientific
Moral
Social
Visual
Lingual
Medical
Musical

Problem solving
Theorem proving
One-shot learning
Zero-shot learning
Counterfactual



Learning | Reasoning

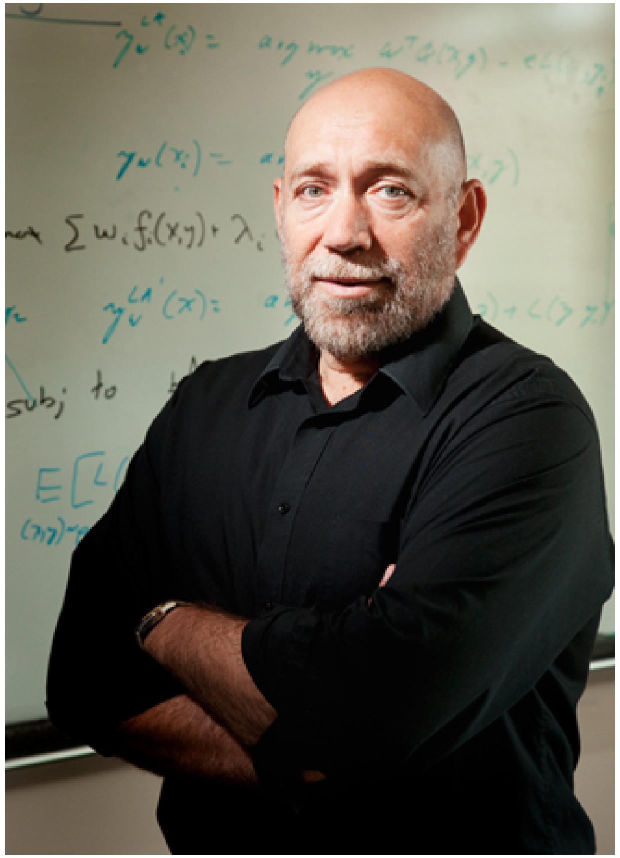
Learning is to improve itself by experiencing ~ acquiring knowledge & skills

Reasoning is to deduce knowledge from previously acquired knowledge in response to a query (or a cues)

Early theories of intelligence (a) focuses solely on reasoning, (b) learning can be added separately and later! (Khardon & Roth, 1997).

Learning precedes reasoning or the two interacting?

Can reasoning be learnt? Are learning and reasoning indeed different facets of the same mental/computational process?



(Dan Roth; *ACM Fellow; IJCAI John McCarthy Award*)

Khardon, Roni, and Dan Roth. "Learning to reason." *Journal of the ACM (JACM) 44.5 (1997): 697-725.*

Bottou's vision

Is not necessarily achieved by making logical inferences

Continuity between algebraically rich inference and connecting together trainable learning systems

Central to reasoning is composition rules to guide the combinations of modules to address new tasks

"When we observe a visual scene, when we hear a complex sentence, we are able to explain in formal terms the relation of the objects in the scene, or the precise meaning of the sentence components. However, there is no evidence that such a formal analysis necessarily takes place: we see a scene, we hear a sentence, and we just know what they mean. **This suggests the existence of a middle layer, already a form of reasoning, but not yet formal or logical**."

Bottou, Léon. "From machine learning to machine reasoning." *Machine learning* 94.2 (2014): 133-149.

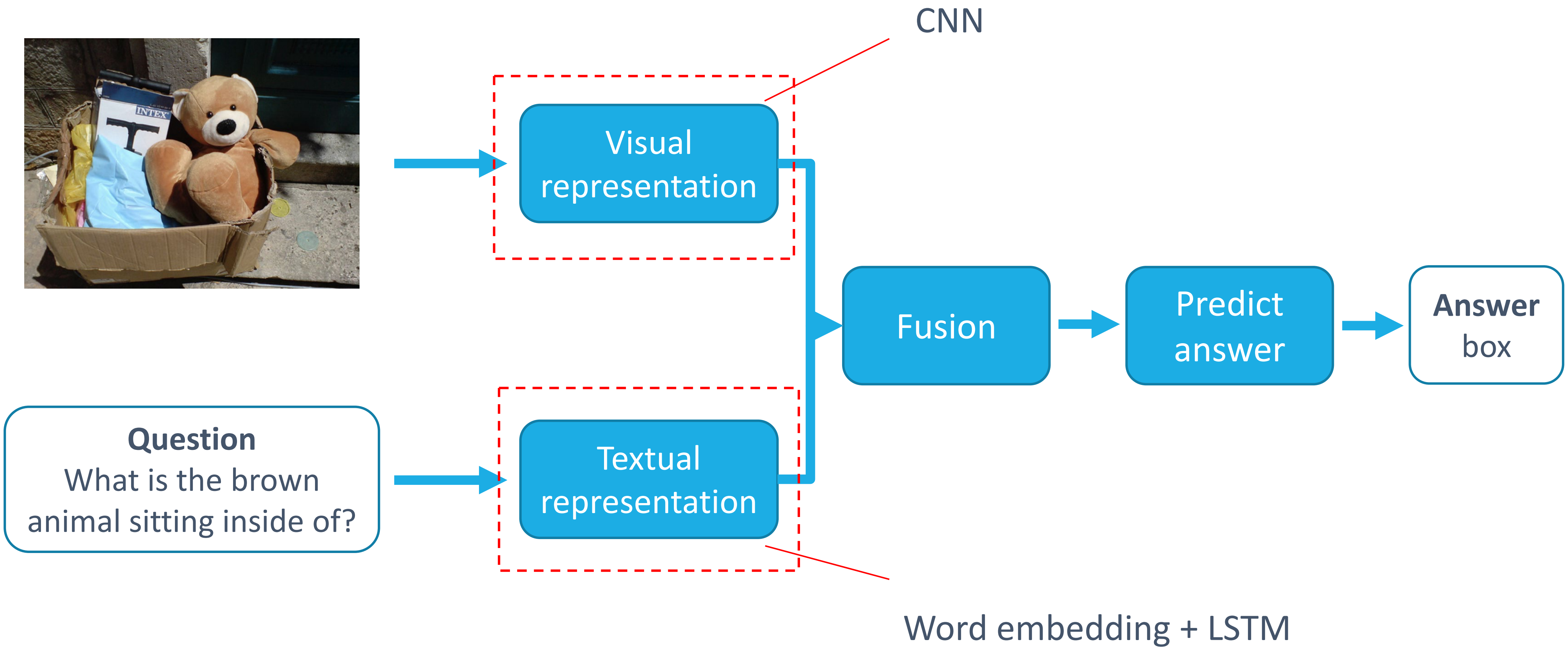
Learning to reason, formal def

Definition 2.1.1. An algorithm A is an *exact reasoning* algorithm for the reasoning problem $(\mathcal{F}, \mathcal{Q})$, if for all $f \in \mathcal{F}$ and for all $\alpha \in \mathcal{Q}$, when A is presented with input (f, α), A runs in time polynomial in n and the size of f and α, and answers "yes" if and only if $f \models \alpha$.

E.g., given a video f, determines if the person with the hat turns before singing.

Khardon, Roni, and Dan Roth. "Learning to reason." *Journal of the ACM (JACM)* 44.5 (1997): 697-725.

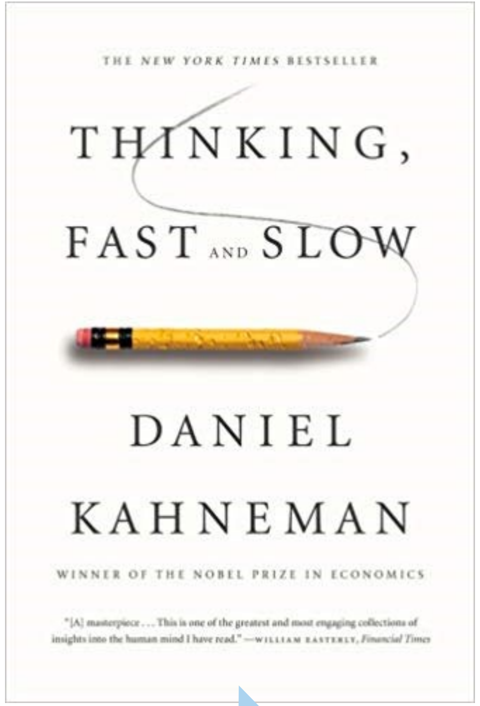
A simple VQA framework that works surprisingly well



Agenda

- Visual reasoning
- Dual-process theories
- Neural memories

Dual-process theories



Multiple

System 1: Intuitive

- Fast
- Implicit/automatic
- Pattern recognition

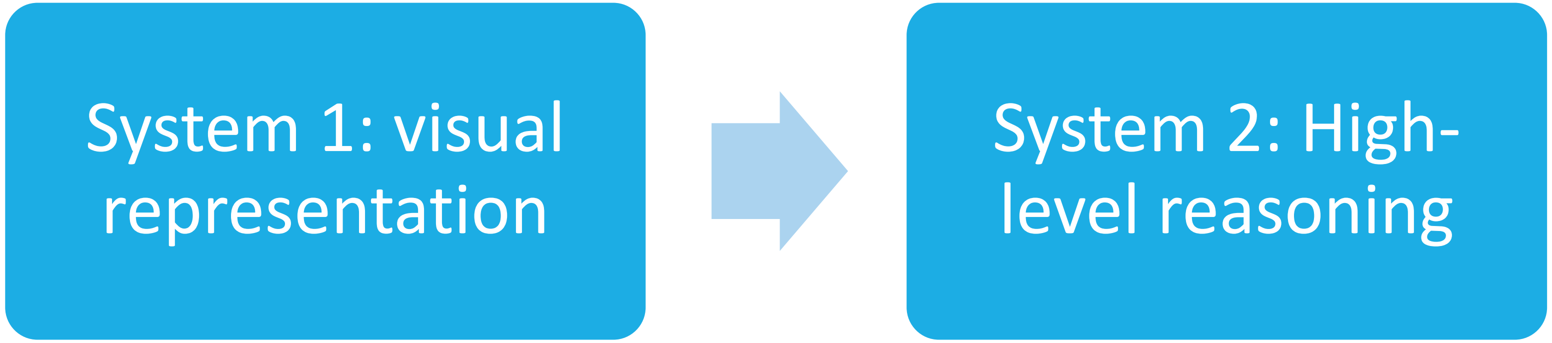
Single

System 2: Analytical

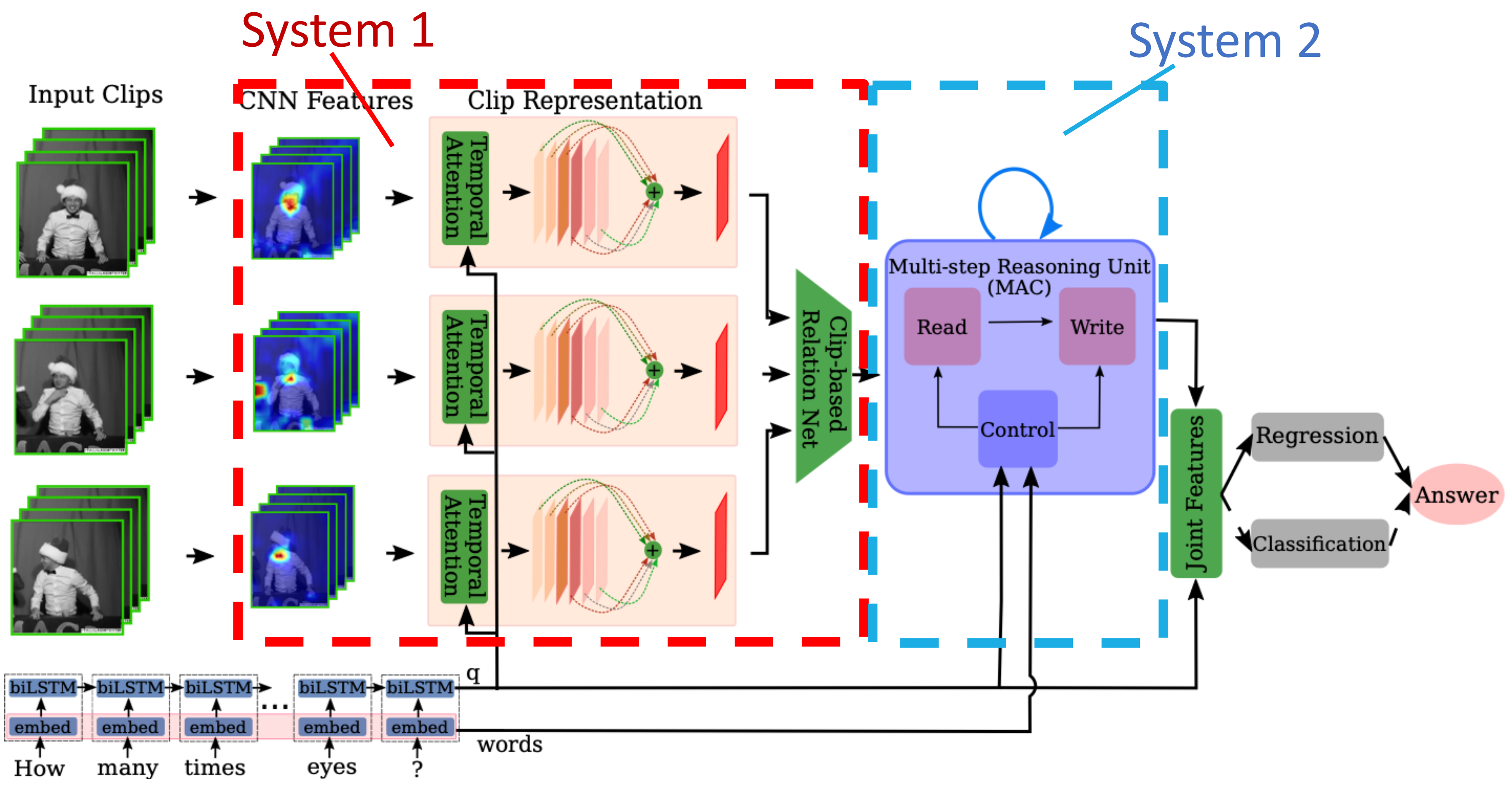
- Slow
- Deliberate/rational
- Careful analysis
- Sequential

Prelim Idea: Separate reasoning process from perception

- Video QA: **inherent dynamic nature** of visual content over time.
- Recent success in visual reasoning with **multi-step inference** and handling of **compositionality**.



Prelim dual-process architecture



Le, Thao Minh, Vuong Le, Svetha Venkatesh, and Truyen Tran. "Learning to Reason with Relational Video Representation for Question Answering." *arXiv preprint arXiv:1907.04553* (2019).

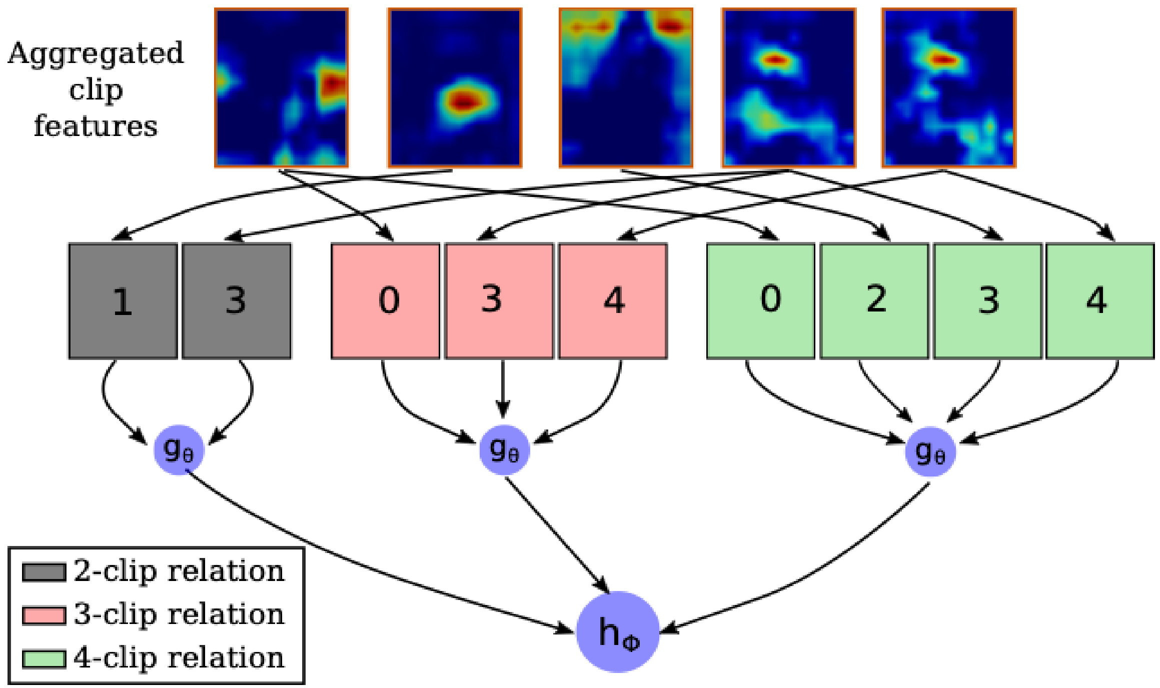
System 1: Clip-based Relation Network

Why temporal relations?

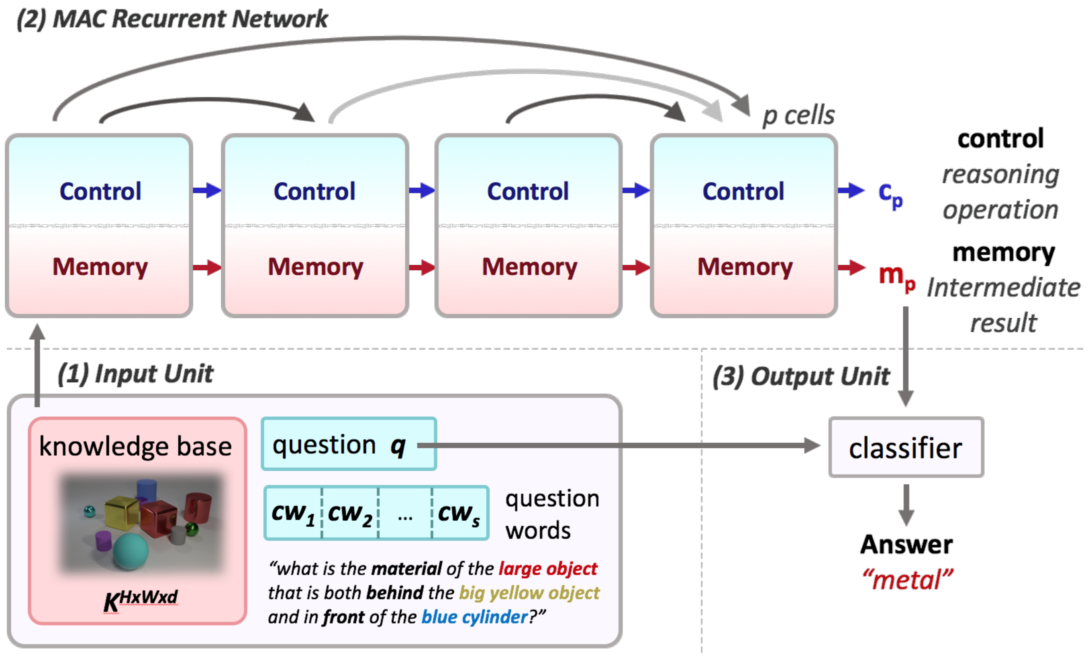
- Situate an event/action in relation to events/actions in the past and formulate hypotheses on future events.
- Long-range sequential modeling.

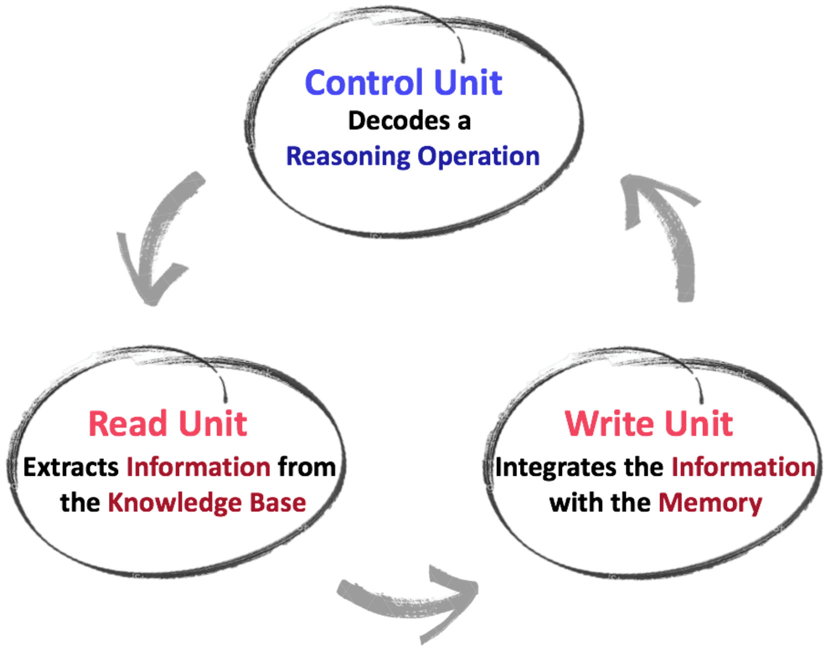
$$R^{(k)}(C) = h_{\Phi}\left(\sum_{i_1<i_2\ldots<i_k} g_{\theta}\left(\bar{C}_{i_1}, \bar{C}_{i_2}, \ldots, \bar{C}_{i_k}\right)\right)$$

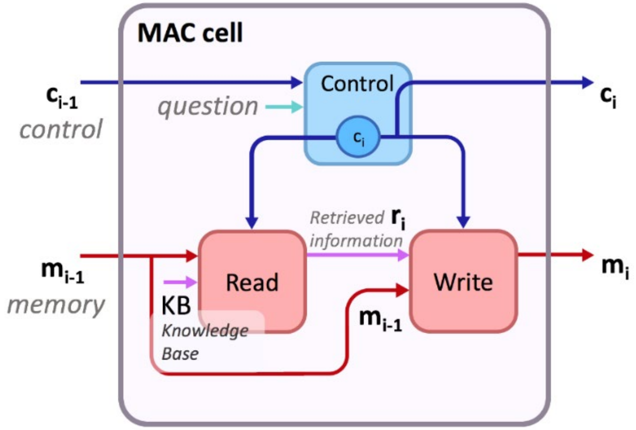
For $k = 2,3,\ldots,K$ where h_ϕ and g_θ are linear transformations with parameters ϕ and θ, respectively, for feature fusion.



System 2: MAC Net







Hudson, Drew A., and Christopher D. Manning. "Compositic networks for machine reasoning." *arXiv preprint arXiv:1803.03067* (2018).

Results on SVQA dataset.

Model	Exist	Count	Integer Comparison			Attribute Comparison					Query					All
			More	Equal	Less	Color	Size	Type	Dir	Shape	Color	Size	Type	Dir	Shape	
SA(S)	51.7	36.3	72.7	54.8	58.6	52.2	53.6	52.7	53.0	52.3	29.0	54.0	55.7	38.1	46.3	43.1
TA-GRU(T)	54.6	36.6	73.0	57.3	57.7	53.8	53.4	54.8	55.1	52.4	22.0	54.8	55.5	41.7	42.9	44.2
SA+TA	52.0	38.2	74.3	57.7	61.6	56.0	55.9	53.4	57.5	53.0	23.4	63.3	62.9	43.2	41.7	44.9
CRN+MAC	72.8	56.7	84.5	71.7	75.9	70.5	76.2	90.7	75.9	57.2	76.1	92.8	91.0	87.4	85.4	75.8

Results on TGIF-QA dataset.

Model	Action	Trans.	FrameQA	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
CRN+MAC	**71.3**	**78.7**	**59.2**	4.23

Better System 1: Conditional Relation Network Unit

Problems:

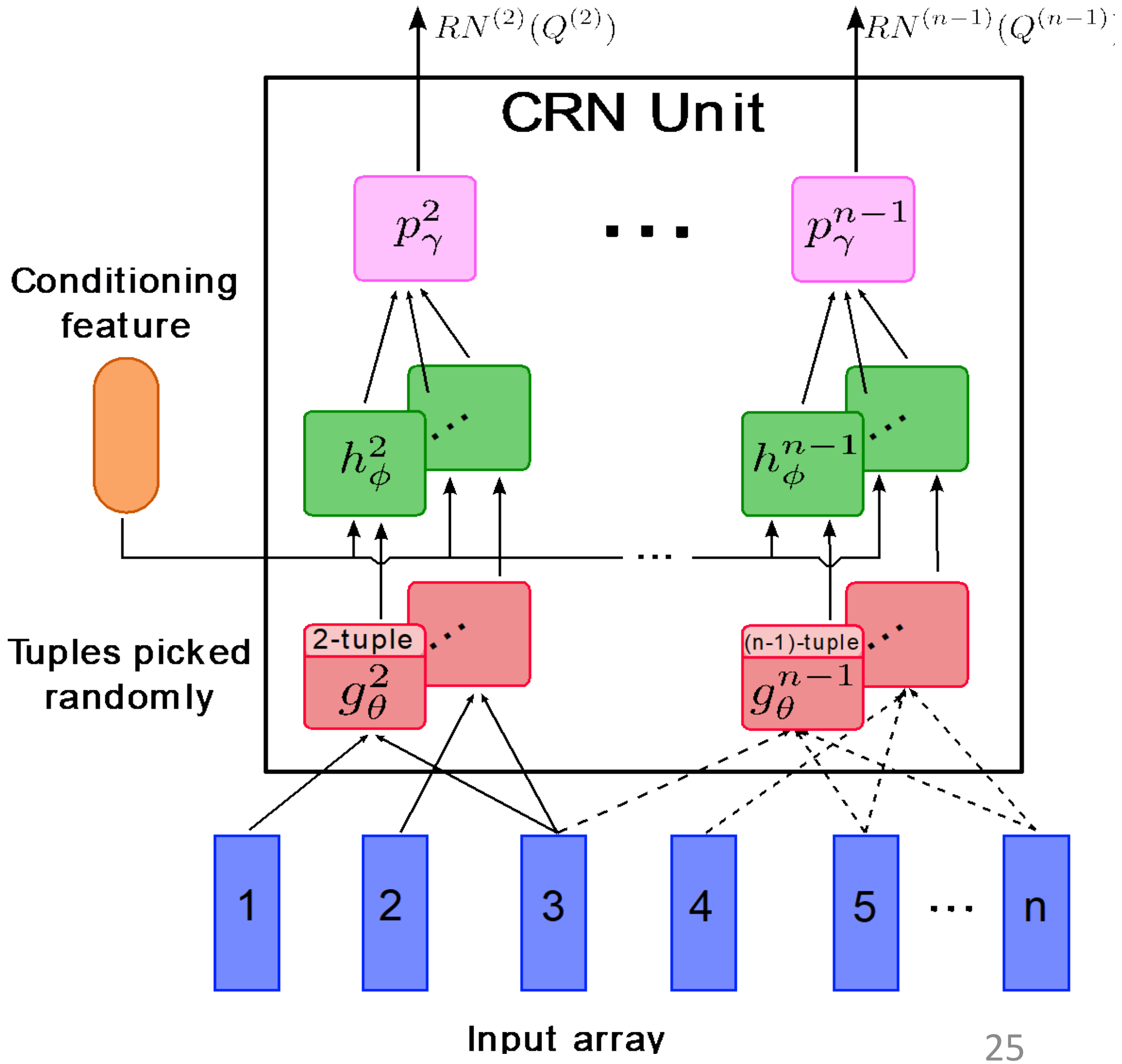
- Lack of a generic mechanism for modeling the interaction of multimodal inputs.
- Flaws in temporal relations of breaking local motion in videos.

Inputs:

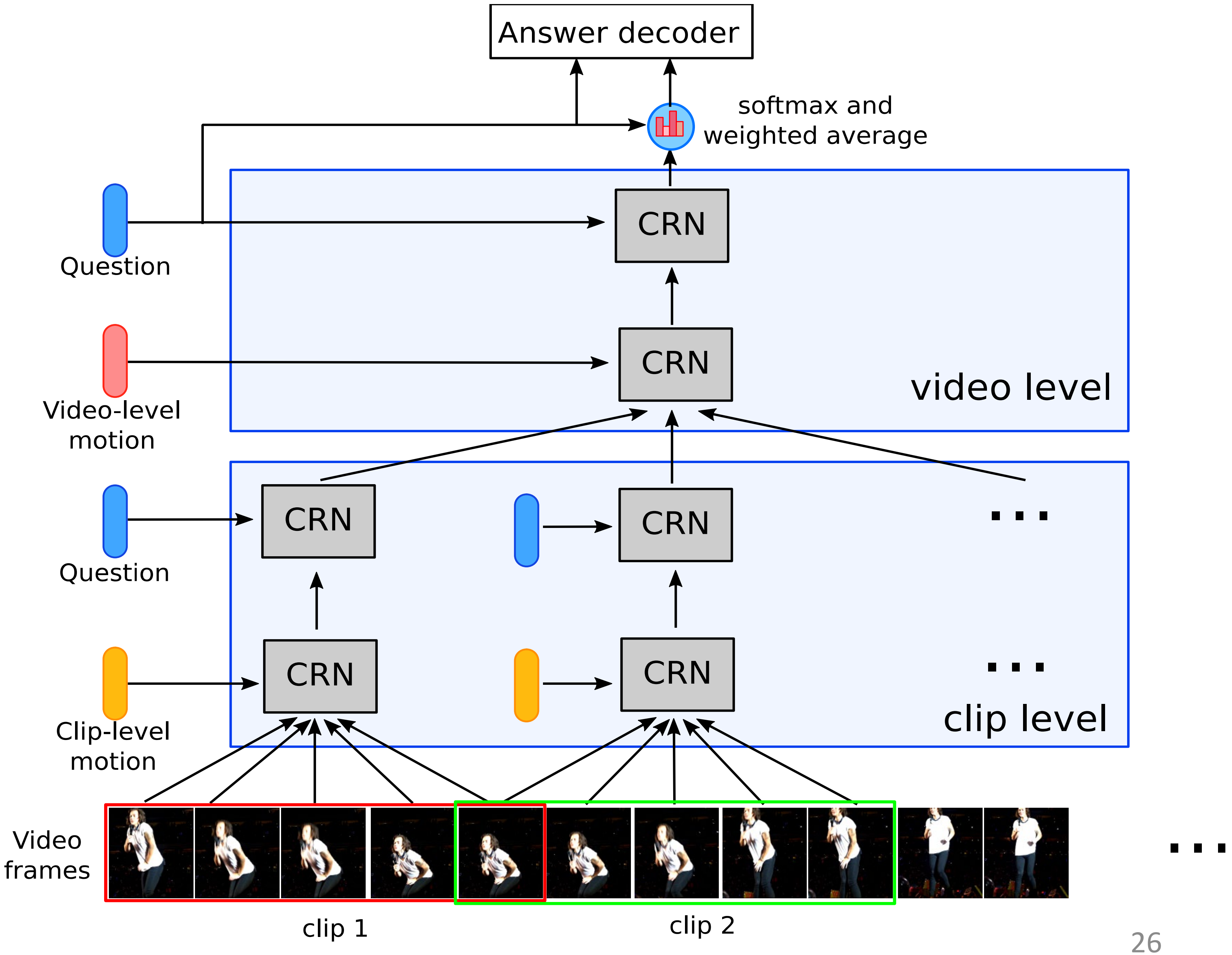
- An array of n objects
- Conditioning feature

Output:

- An array of m objects



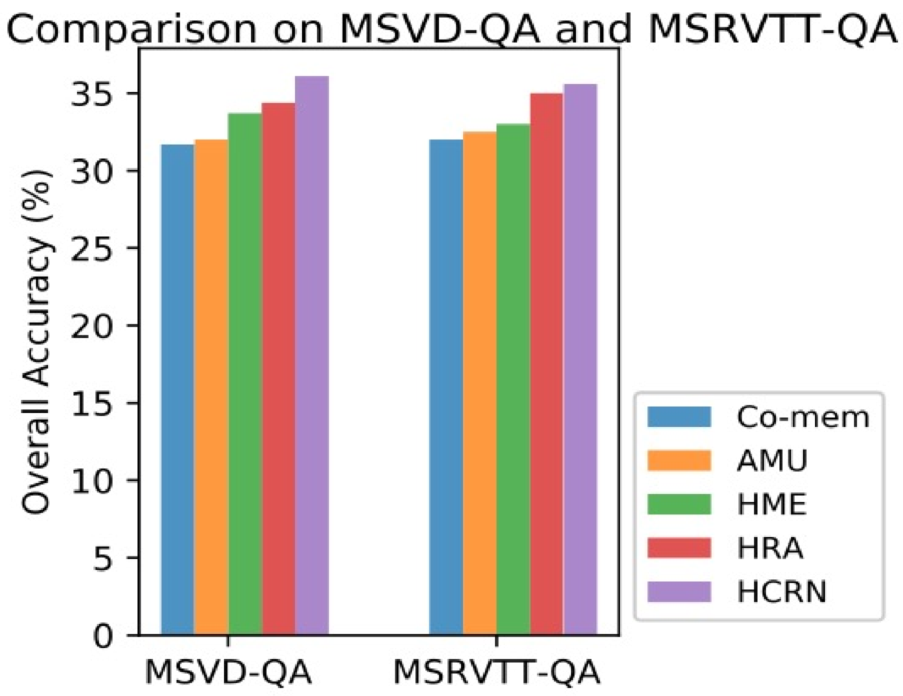
A system for VideoQA: Hierarchical Conditional Relation Networks



Results

Model	Action	Trans.	FrameQA	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**

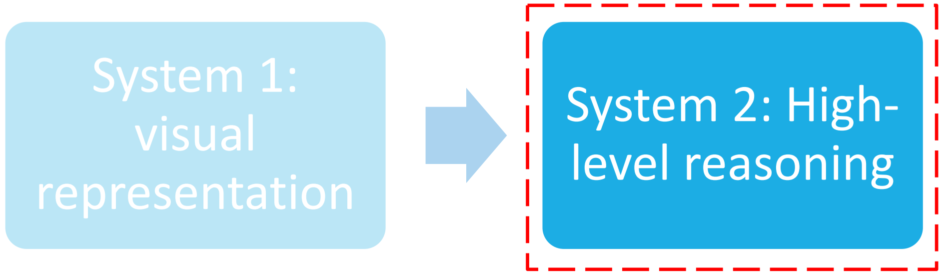
TGIF-QA dataset



MSVD-QA and MSRVTT-QA datasets.

Better System 2: Reasoning with structured representation of spatial relation

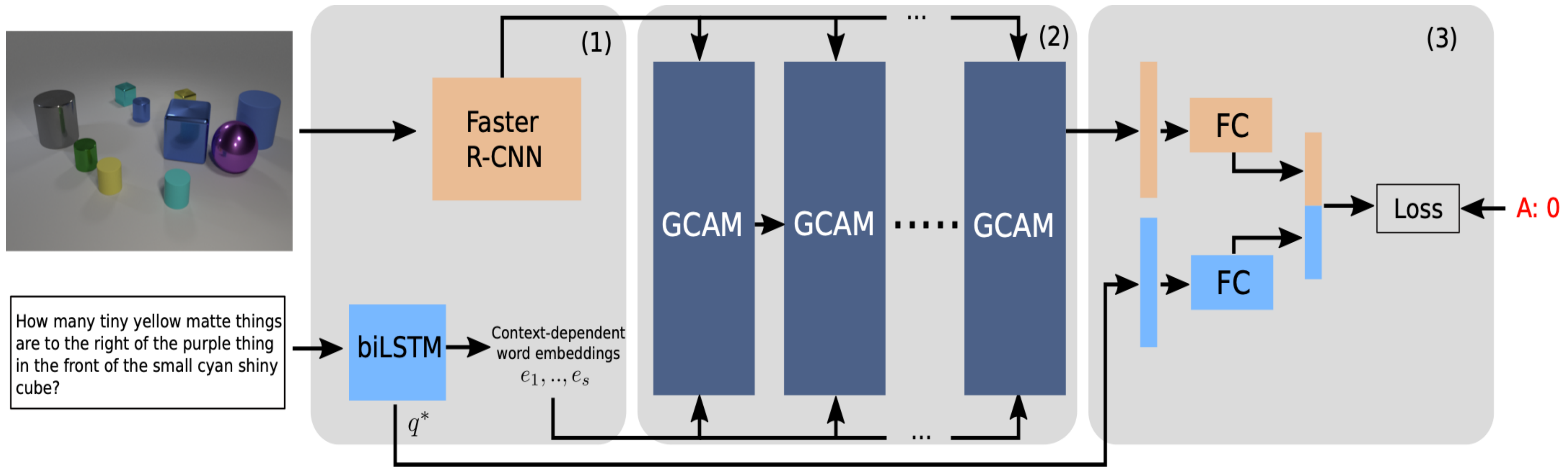
- Image representation is mature and well-studied.
- **Spatial relation:** key for ImageQA.
- Great potentials of **structured representation** of knowledge for reasoning.
- Advance the reasoning capability of **System 2** in the dual-view system.



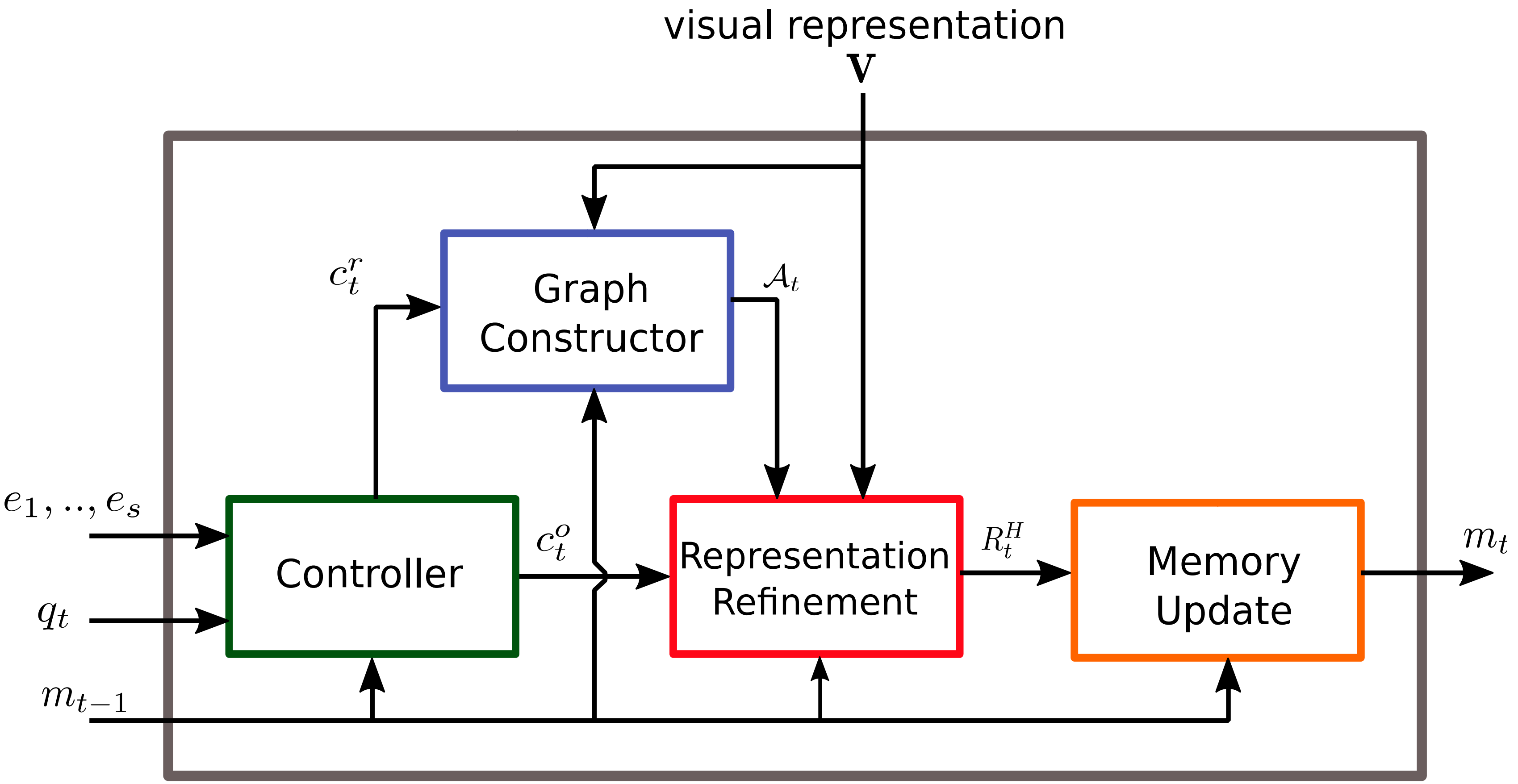
Approach:

- Incorporate spatial relations of concrete objects in the decision-making process.

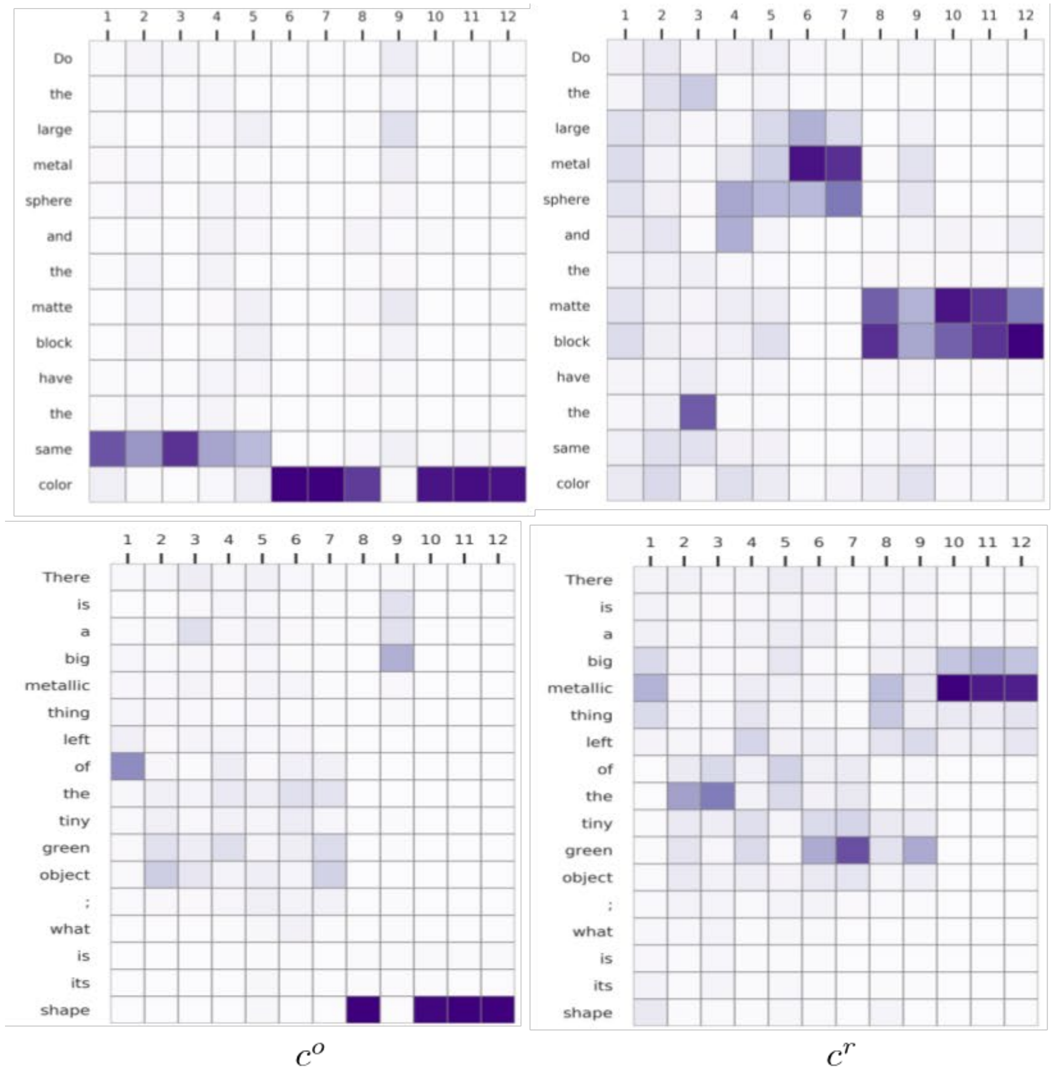
Relation-aware Co-attention Networks for VQA

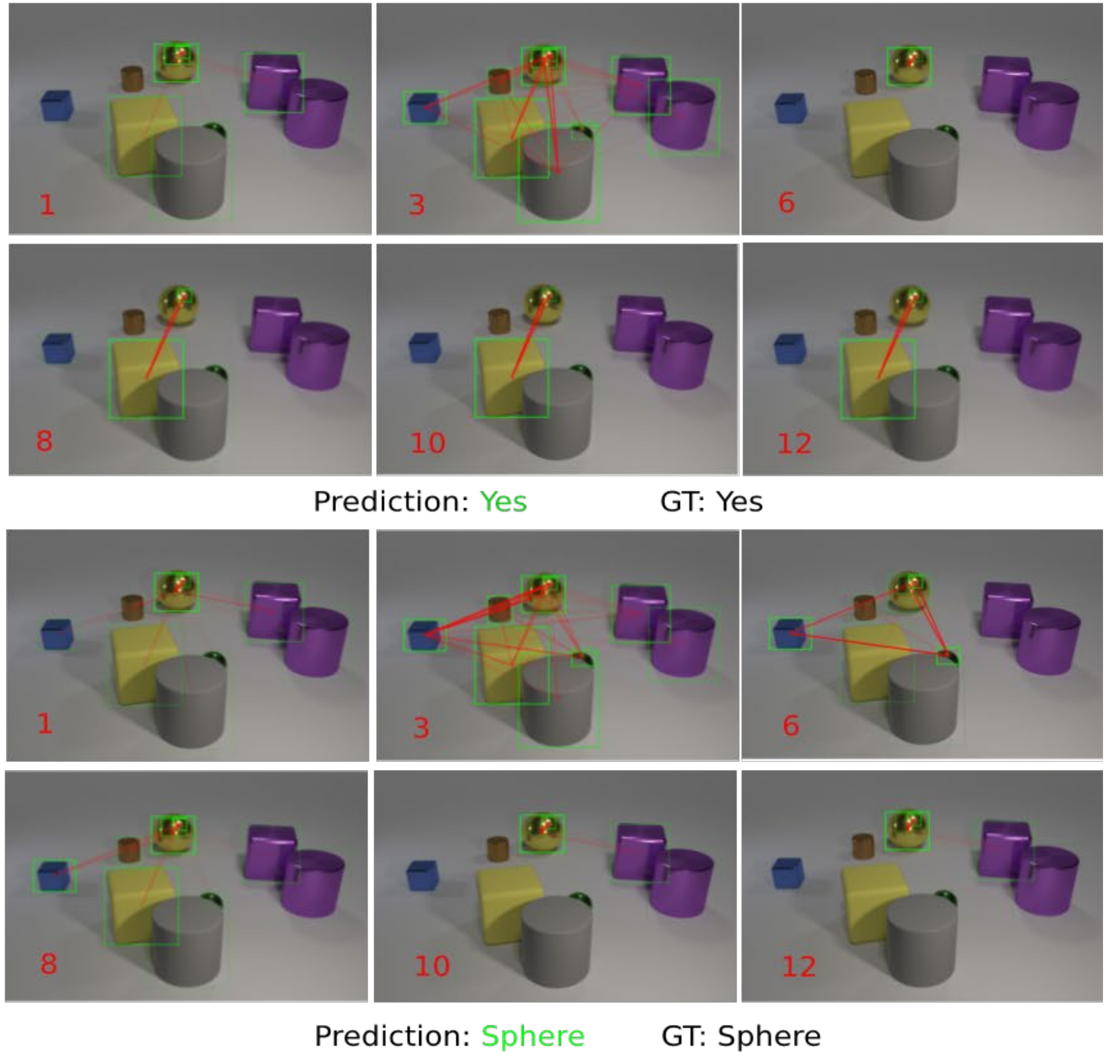


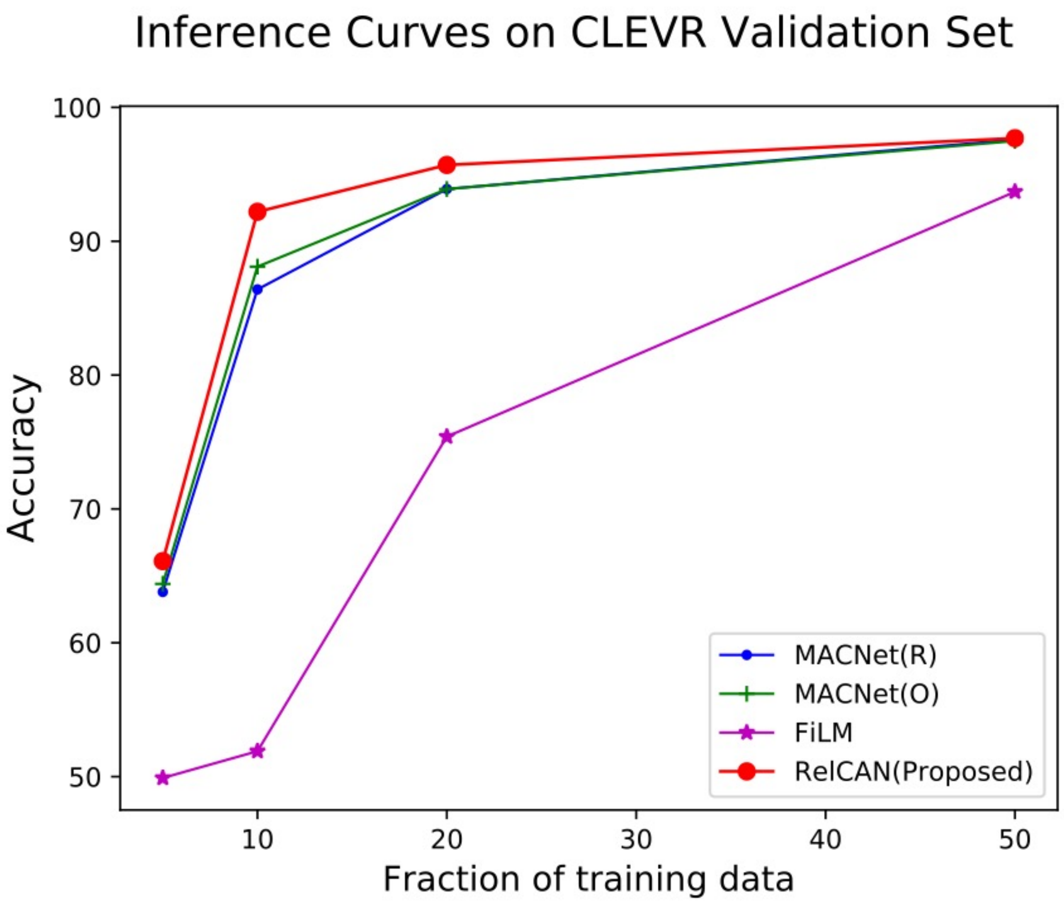
Graph-Structured Co-Attention Module (GCAM)



Results in CLEVR dataset







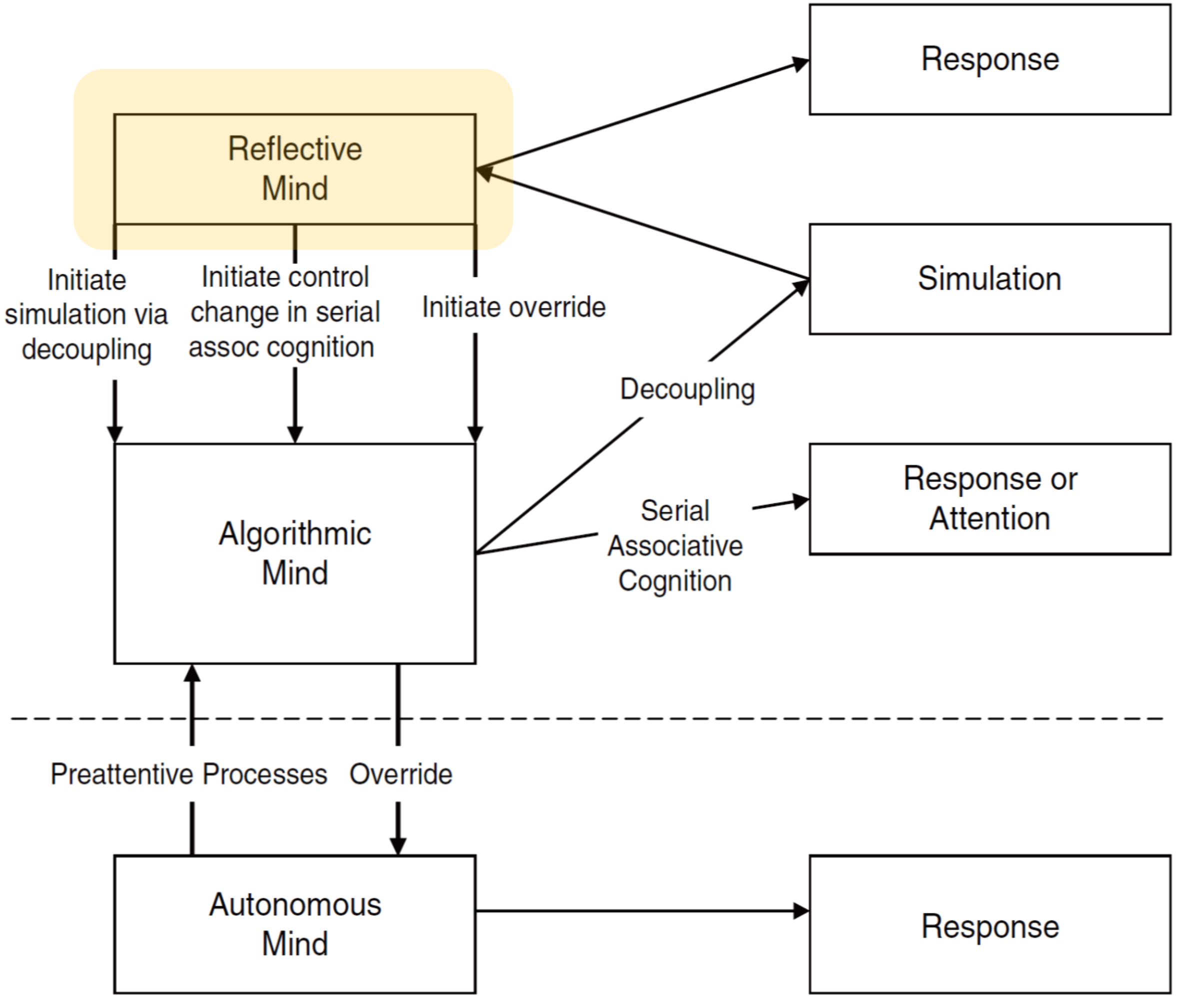
Comparison with the on CLEVR dataset of different data fractions.

Model	Action
XNM(Objects)	43.4
MAC(ResNet)	40.7
MAC(Objects)	45.5
HCRN(Objects)	**46.6**

Performance comparison on VQA v2 subset of long questions.

Towards a ~~dual~~ tri-process theory

Photo credit: mumsgrapevine



Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory. *In two minds: Dual processes and beyond*, 55-88.

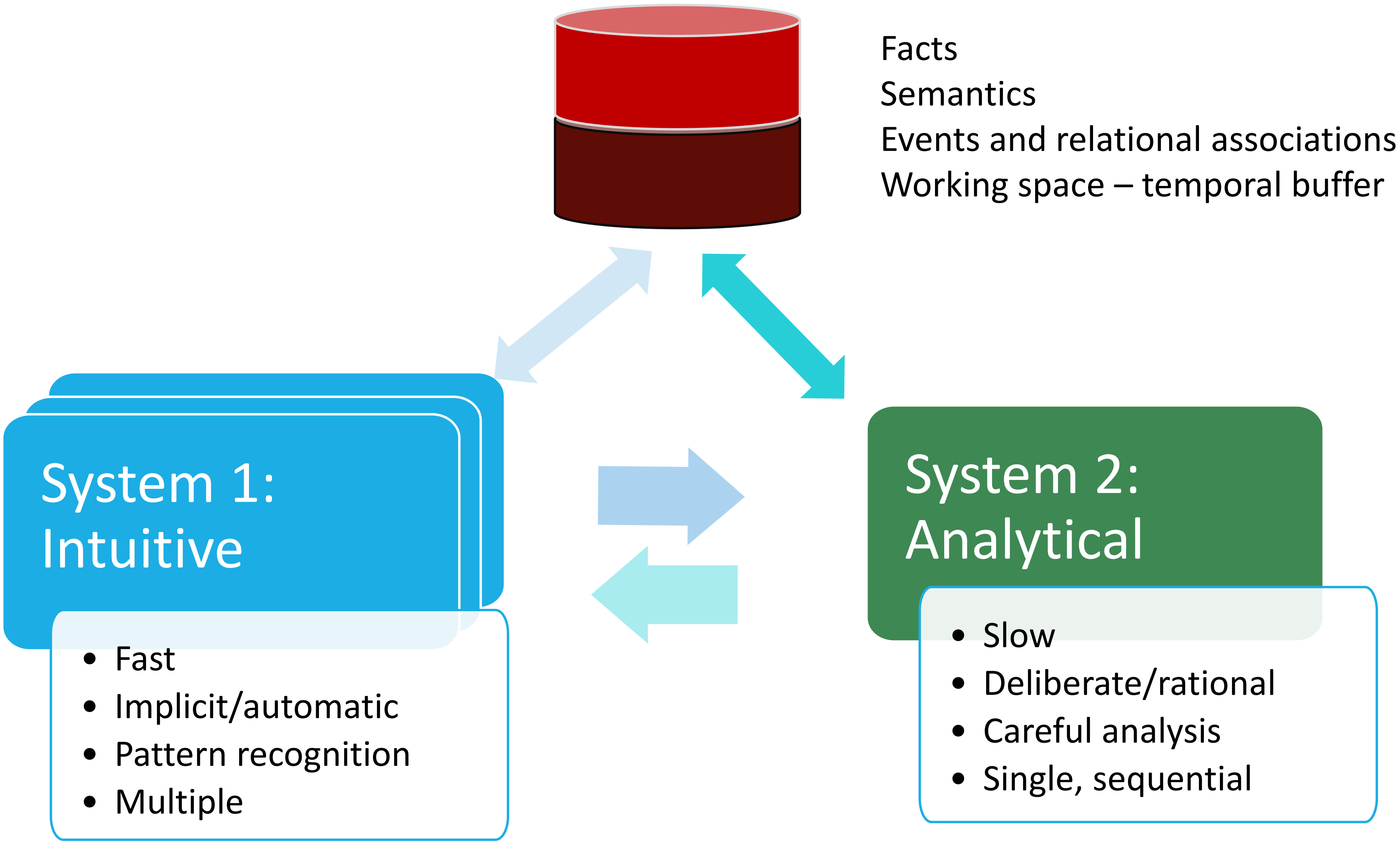
Agenda

Visual reasoning

Dual-process theories

Neural memories

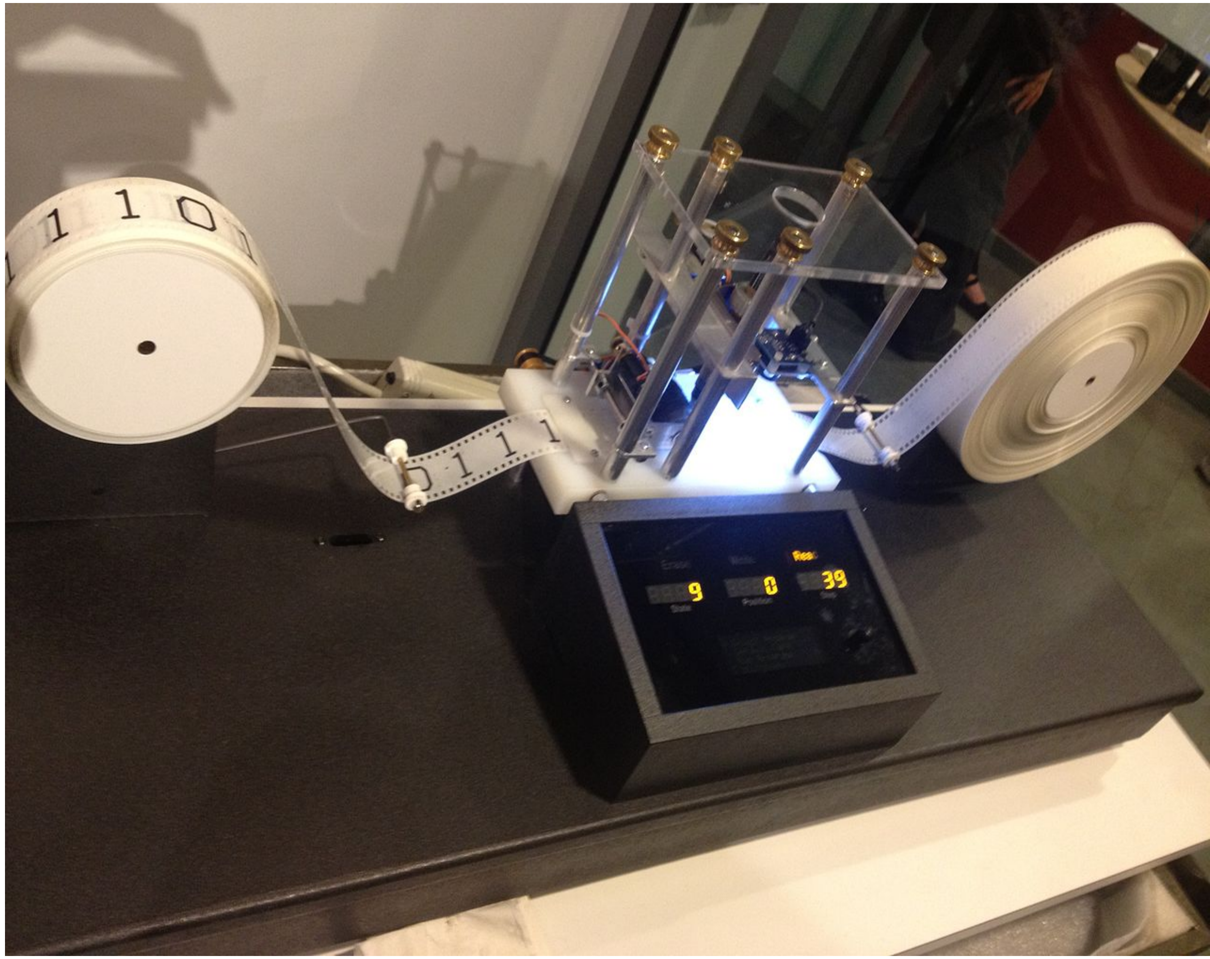
Memories for dual-process system



Learning a Turing machine

→ Can we learn a (neural) program that learns to program from data?

Visual reasoning is a specific program of two inputs (visual, linguistic)



Neural Turing machine (NTM)

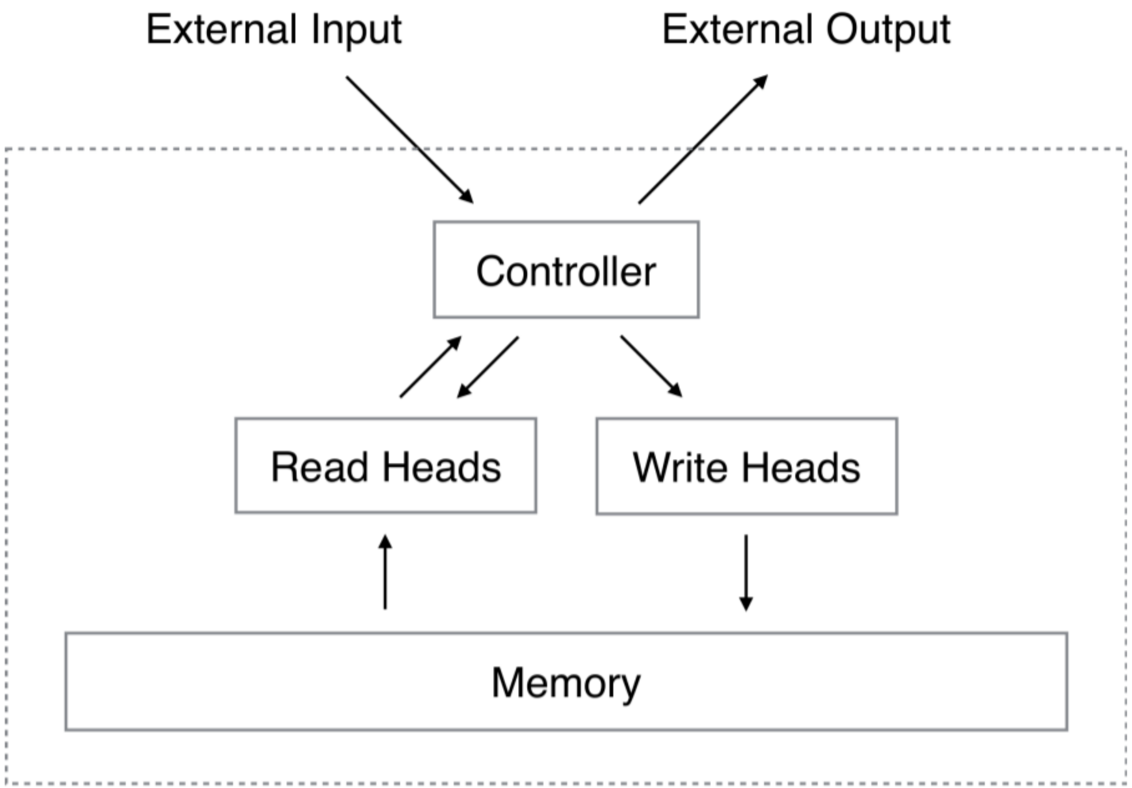
A memory-augmented neural network (MANN)

A controller that takes input/output and talks to an external memory module.

Memory has read/write operations.

The main issue is where to write, and how to update the memory state.

All operations are differentiable.



https://rylanschaeffer.github.io/content/research/neural_turing_machine/main.html

Computing devices vs neural counterparts

FSM (1943) ↔ RNNs (1982)

PDA (1954) ↔ Stack RNN (1993)

TM (1936) ↔ NTM (2014)

UTM/VNA (1936/1945) ↔ NUTM--ours (2019)

The missing piece: A memory to store programs

→ Neural stored-program memory

NUTM = NTM + NSM

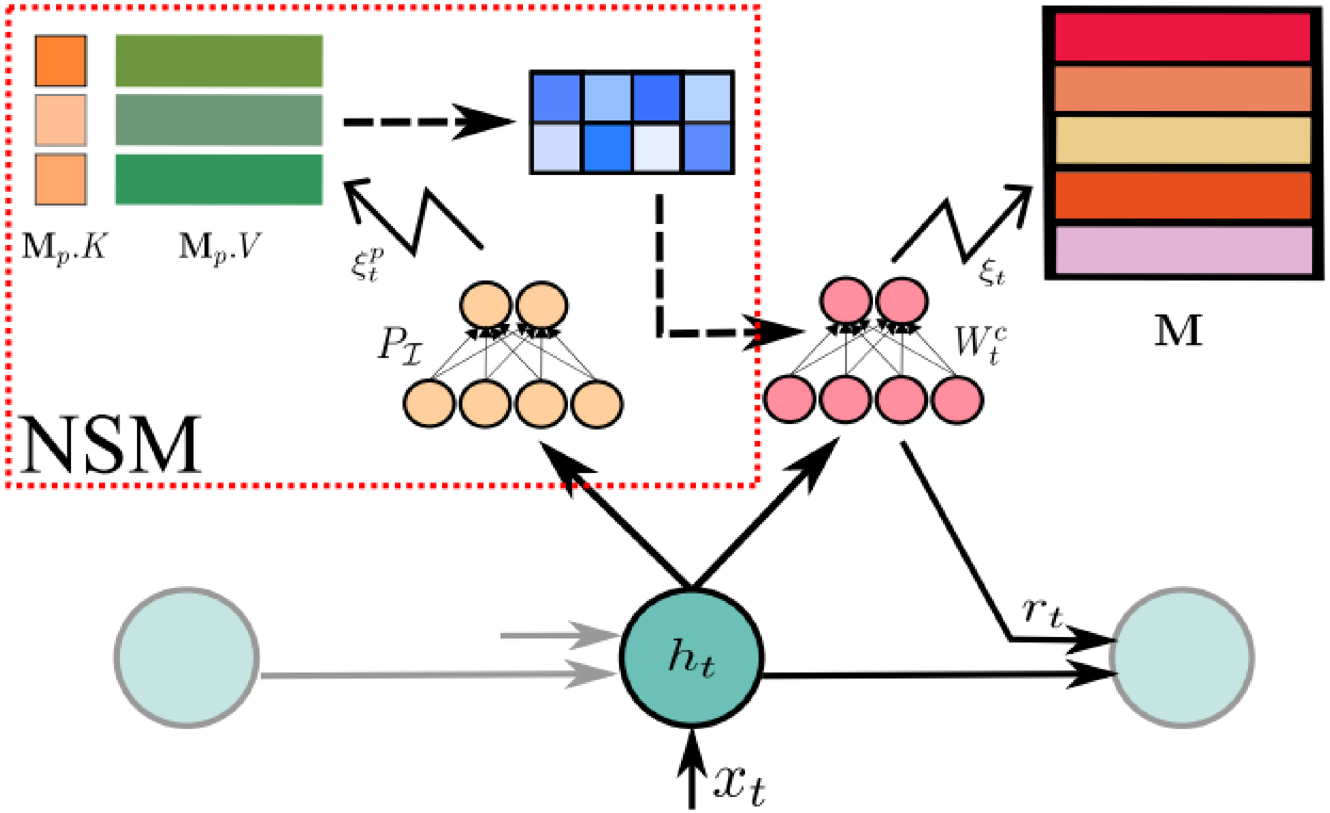


Figure 1: Introducing NSM into MANN. At each timestep, the program interface ($P_{\mathcal{I}}$) receives input from the state network and queries the program memory $\mathbf{M}_p$, acquiring the working weight for the interface network (W_t^c). The interface network then operates on the data memory $\mathbf{M}$ as normal.

Question answering (bAbI dataset)

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

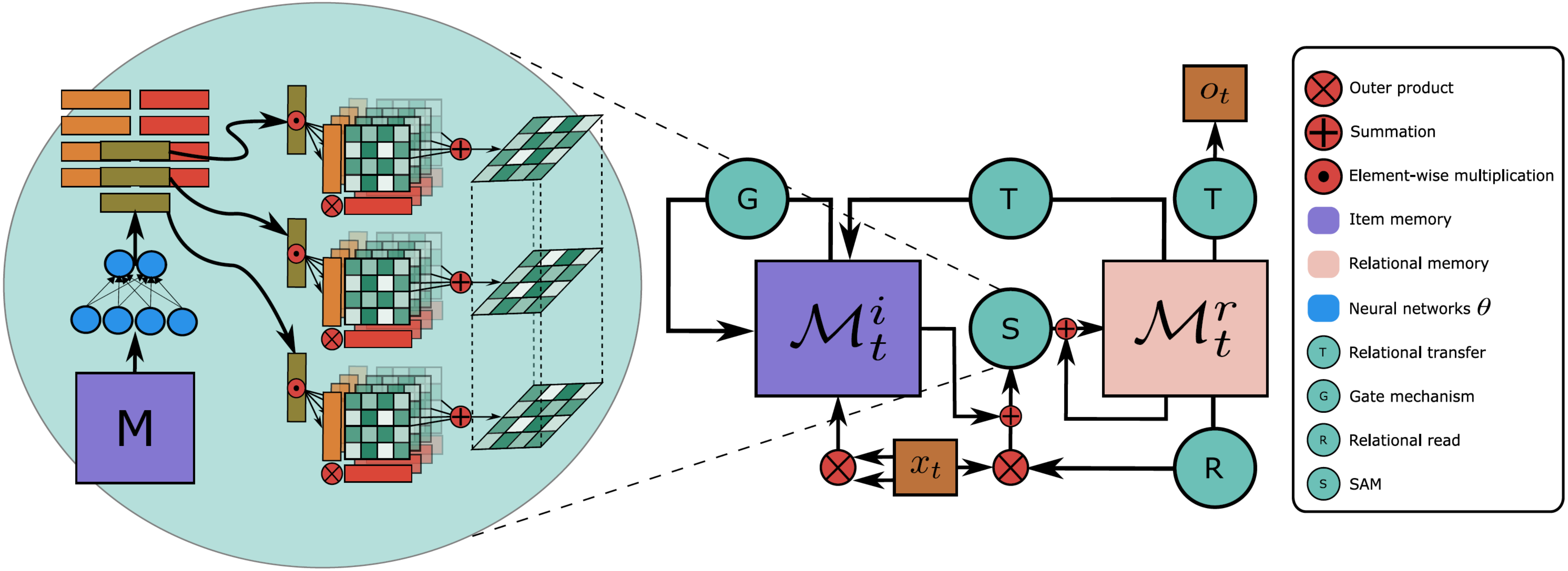
Credit: hexahedria

DNC[12]	SDNC[20]	ADNC[9]	DNC-MD[8]	NUTM (DNC core)	
				$p = 2$	$p = 4$
16.7 ± 7.6	6.4 ± 2.5	6.3 ± 2.7	9.5 ± 1.6	7.5 ± 1.6	$\mathbf{5.6 \pm 1.9}$

Table 3: Mean and s.d. for bAbI error (%).

Self-attentive associative memories (SAM)

Learning relations automatically over time



Multi-step reasoning over graphs

Model	#Parameters	Convex hull $N=5$	Convex hull $N=10$	TSP $N=5$	TSP $N=10$	Shortest Path	Minimum Spanning Tree
LSTM	4.5 M	89.15	82.24	73.15 (*2.06*)	62.13 (*3.19*)	72.38	80.11
ALSTM	3.7 M	89.92	85.22	71.79 (*2.05*)	55.51 (*3.21*)	76.70	73.40
DNC	1.9 M	89.42	79.47	73.24 (*2.05*)	61.53 (*3.17*)	83.59	82.24
RMC	2.8 M	93.72	81.23	72.83 (*2.05*)	37.93 (*3.79*)	66.71	74.98
SAM	1.9 M	**96.85**	**91.88**	**73.96 (*2.05*)**	**69.43 (*3.03*)**	**93.43**	**94.77**

Yet to be solved …

Common-sense reasoning

Reasoning as program synthesis with callable, reusable modules

Systematicity, aka systematic generalization

Knowledge-driven VQA, knowledge as semantic memory

- Differentiable neural-symbolic systems for reasoning

Visual dialog

- Active question asking

Higher-order thought (e.g., self-awareness and consciousness)

A better prior for reasoning

The reasoning team @

A/Prof Truyen Tran

Dr Vuong Le

Mr Hung Le

Mr Tin Pham

Mr Thao Minh Le

Mr Dang Hoang Long

Thank you

Truyen Tran

truyen.tran@deakin.edu.au

truyentran.github.io

@truyenoz

letdataspeak.blogspot.com

goo.gl/3jJ1O0

APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY