Lecture 4: Reasoning over unstructured sets

https://neuralreasoning.github.io/

Presented by Vuong Le

Learning to Reason - Pratical formulation

- `query,database → answer`
- This is very general:
  - Classification: Query = *what is this?* Database = *data.*
  - Regression: Query = *how much?* Database = *data.*
  - QA: Query = *NLP question.* Database = *context/image/text.*
  - Multi-task learning: Query = *task ID.* Database = *data.*
  - Zero-shot learning: Query = *task description.* Database = *data.*
  - Drug-protein binding: Query = *drug*. Database = *protein.*
  - Recommender system: Query = *User (or item)*. Database = *inventories (or user base)*;

→ Reasoning problem: query changes, and only available at runtime

Learning to Reason formulation

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

$$\tilde{a} = \arg\max_{a \in \mathbb{A}} \mathcal{P}_{\theta}\left(a \mid C, q\right)$$

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

"What affects her mobility?"

Is it simply an optimization problem like recognition, detection, translation?

→No, because the query q is unknown until the run time

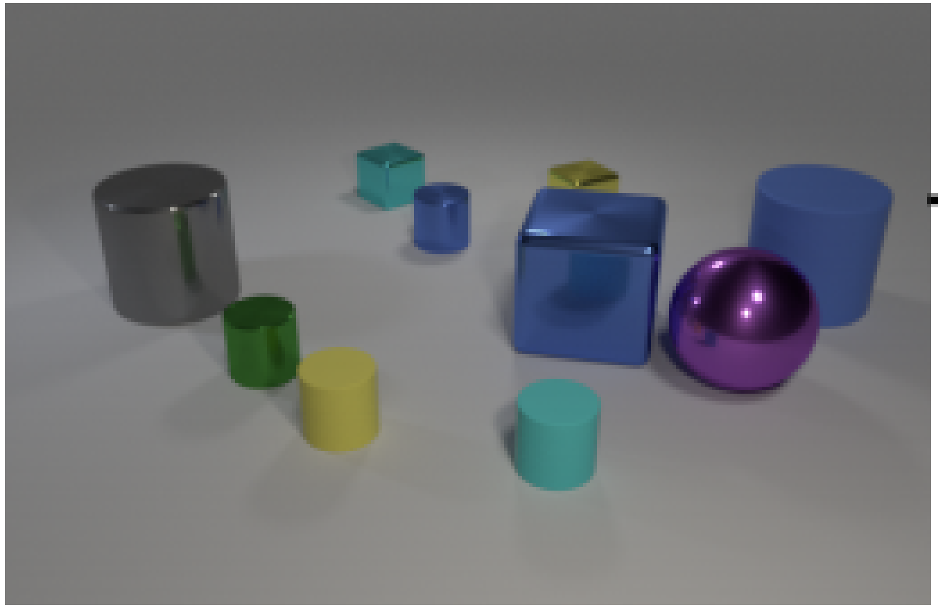
→We need to count for it adaptively under the model's structures and inference strategies

A case study: Image Question Answering

$$\tilde{a} = \arg\max_{a \in \mathbb{A}} \mathcal{P}_\theta \left(a \mid C, q \right)$$

- Specs:
  - C: visual content of an image
  - q: a linguistic question
  - a: a linguistic phrase answering q regarding C
- Challenges
  - Reasoning through facts and logics
  - Cross-modality integration

- Further details of Image QA: Lecture 8



How many tiny yellow matte things are to the right of the purple thing in the front of the small cyan shiny cube?

The main approaches in Image QA

- Symbolic logical reasoning
  - Parse the question into a “program” of logical inference steps
  - The logical inference follow the program

  \+ Explicit and interpretable

  \+ Close to human’s logical inference

  \- Brittle, cannot recover from mistakes

  \- Struggling with nuances of language and visual context

  \- *Leon Bottou: Reasoning needs not to be logical inferences*
- **Compositional reasoning (This lecture + Lecture 5)**
- **Neural symbolic reasoning (Lecture 6)**

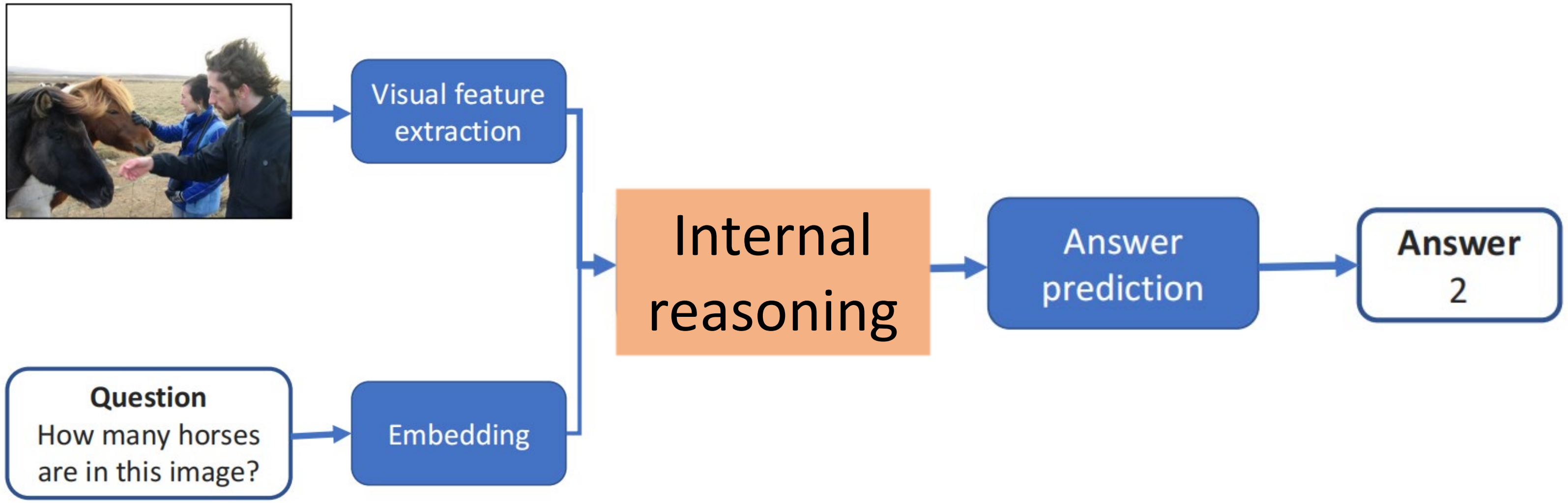
what color is the vase?

```
classify[color](
  attend[vase])
```

green (green)

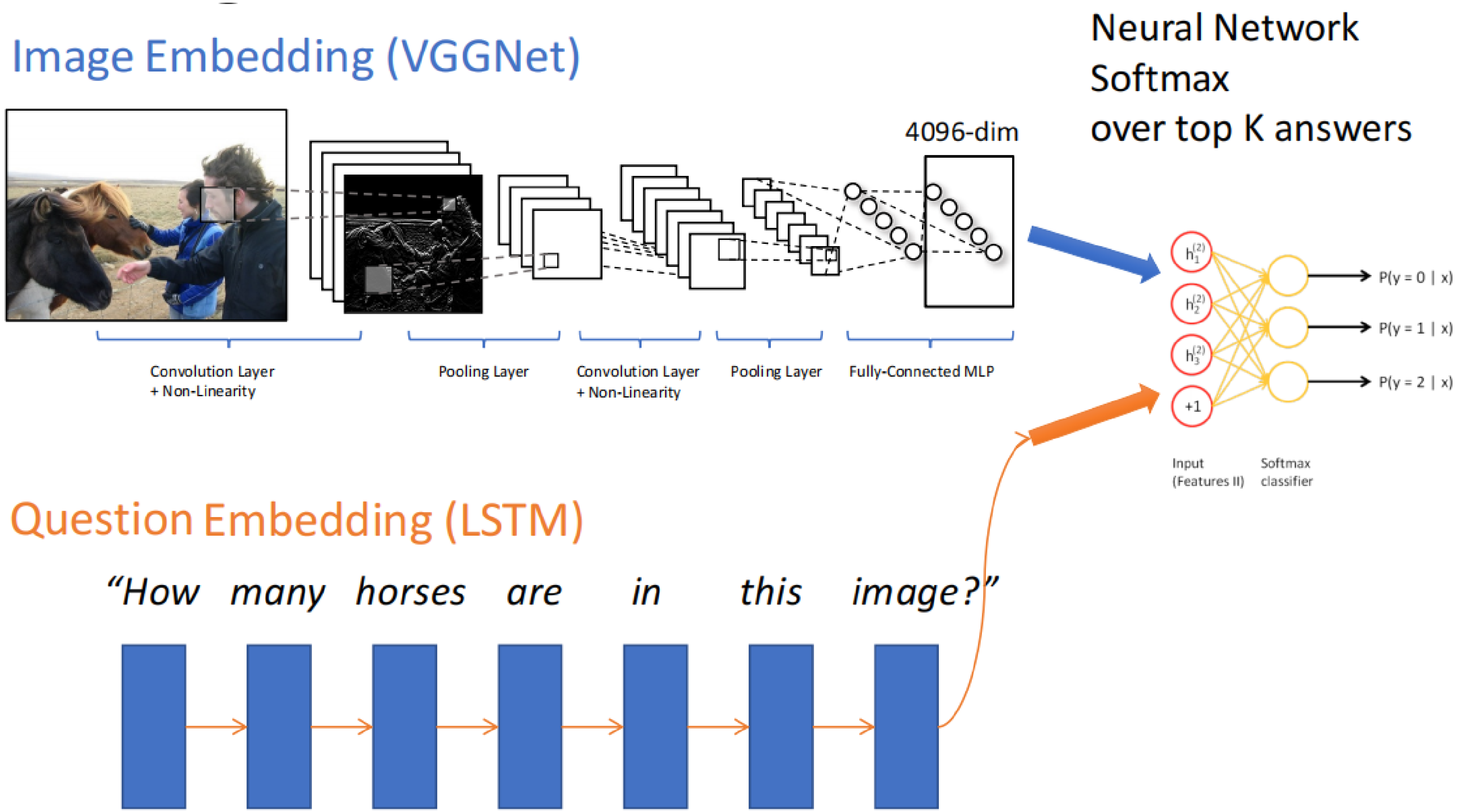
Compositional reasoning

- Extract visual and linguistic individual- and joint- representation
- Reasoning happens on the structure of the representation
  - Sets/graphs/sequences
- The representation got refined through multi-step compositional reasoning



Also resembling one way that human thinks and decides.
(My personal take: this is the more prominent way that we think with)
Q: Can compositional reasoning be combined with neural symbolic? Maybe. It is a promising path to go!

A simple approach



→Issue: This is very susceptible to the variations and nuances of images and questions

→We must be able to concentrate on relevant parts of image: Set of concepts? Attention?

Reasoning as set-set interaction

- C: a set of context objects

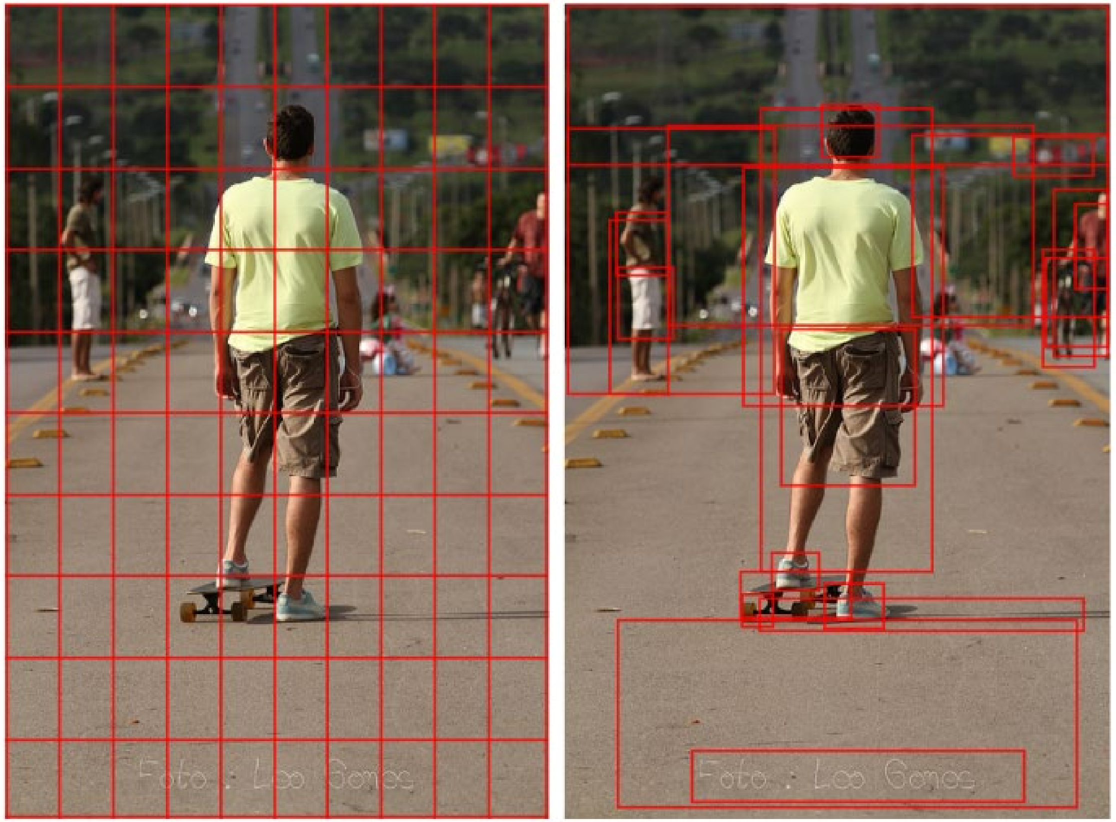
$$C = \{o_1, o_2, ..., o_n\}$$

  - Faster-RCNN regions
  - CNN slices
- q: a set of linguistic objects

$$Q = \{w_1, w_2, ..., w_n\}$$

- biLSTM embedding of q

$$\mathbf{w}_i^q = [\overrightarrow{\text{LSTM}}(\mathbf{e}_i^q); \overleftarrow{\text{LSTM}}(\mathbf{e}_i^q)]$$



→ Reasoning is formulated as the interaction between the two sets O and L for the answer a

Figure credit: BUTD, Anderson et.al.

Set operations

- Reducing operation (eg: sum/average/max)

$$\mathbf{c} \;=\; h_{\boldsymbol{\theta}}\left(\{\mathbf{o}_1, \mathbf{o}_2, .., \mathbf{o}_N\}\right)$$

- Attention-based combination (Bahdanau et al. 2015)

$$\mathbf{c} \;=\; \sum_{i=1}^{N} \alpha_i \mathbf{o}_i \qquad\qquad \alpha_i \;=\; \frac{\exp(\mathbf{W}^o \mathbf{o}_i)}{\sum_{j=1}^{N} \exp(\mathbf{W}^o \mathbf{o}_j)}$$

- Attention weights as query-key dot product (Vaswani et al., 2017)

$$\mathbf{c} \;=\; \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

→ Attention-based set ops seem very suitable for visual reasoning

Attention-based reasoning

- 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:

- $s_i = \tanh(\mathbf{W}^c \mathbf{w}_i^c + \mathbf{W}^{\mathbf{q}} \mathbf{q})$ 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

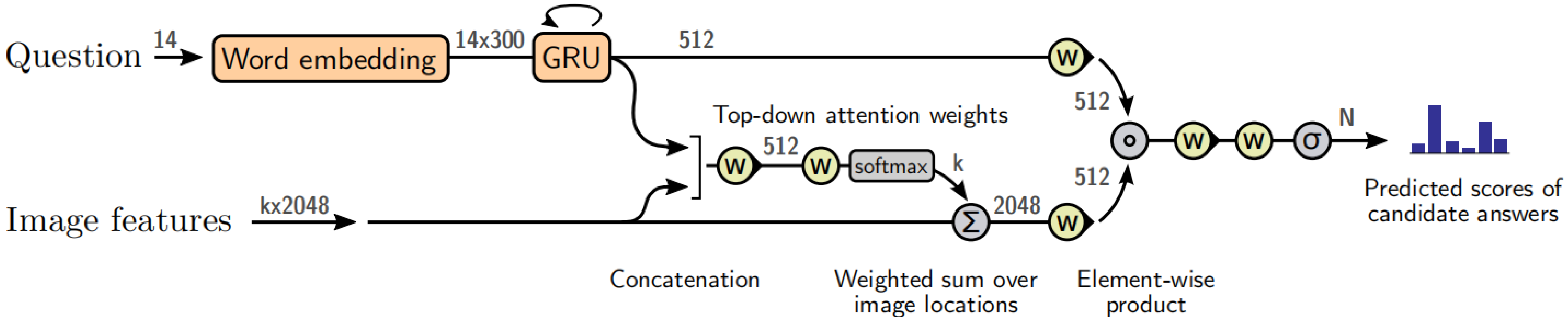
$$\alpha_i = \frac{\exp(\mathbf{W} s_i)}{\sum_j \exp(\mathbf{W} s_j)}$$

  - Attended context vector: $\mathbf{i} = \sum_i \alpha_i \mathbf{w}_i^c$

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

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

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



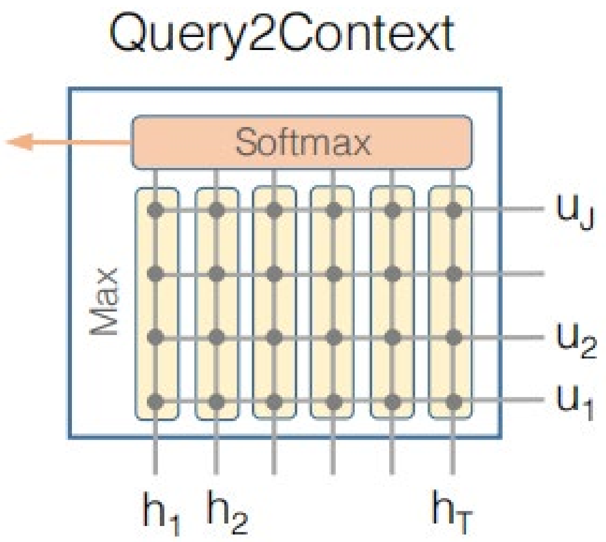
→ Q: How about attention from vision objects to linguistic objects?

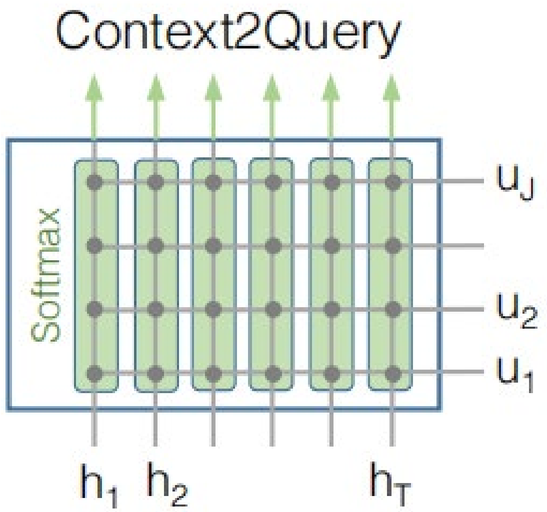
Bi-directional attention

- 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





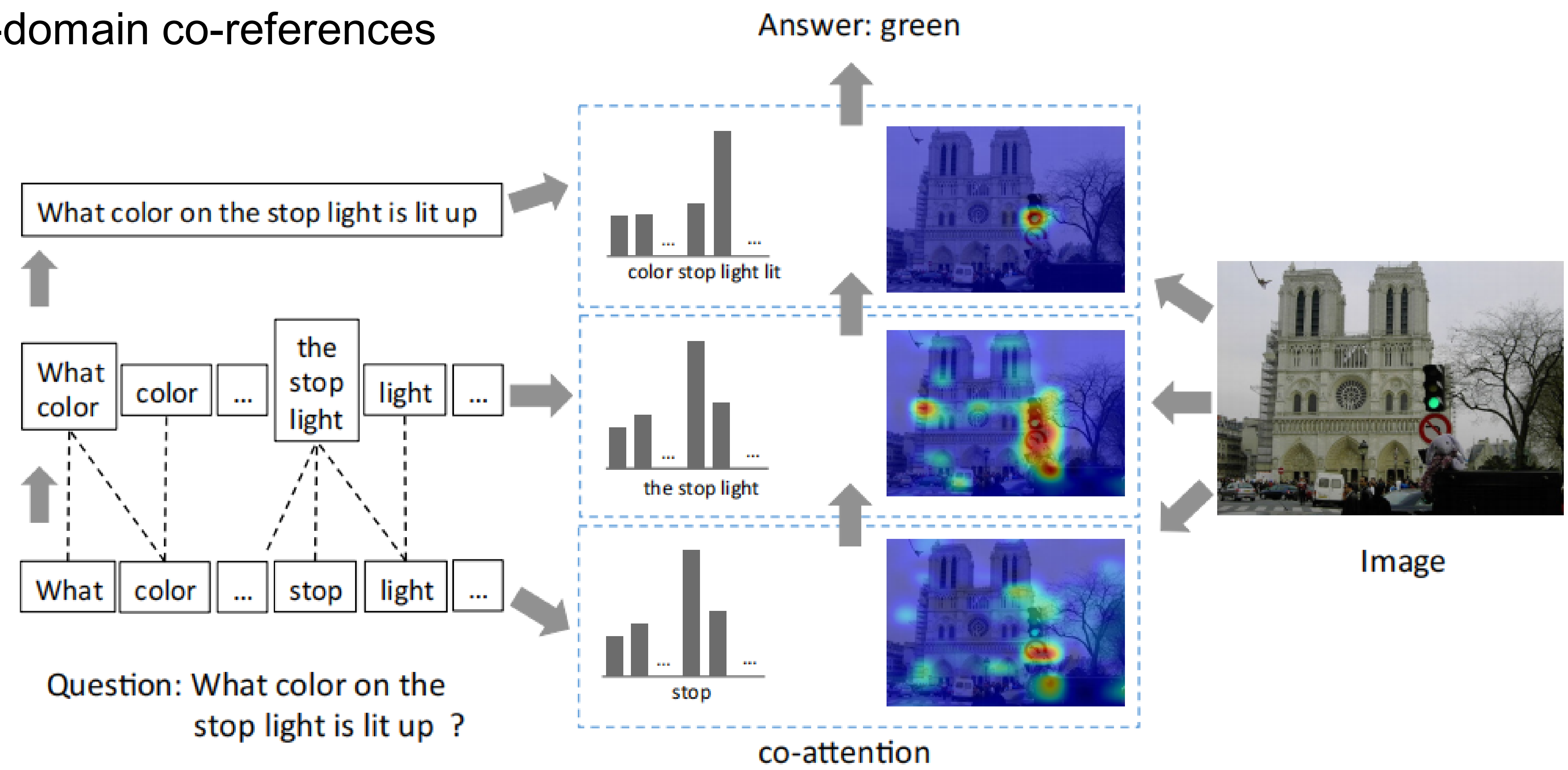
→ Q: Probably not working for image qa where single words does not have the co-reference with a region?

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

Hierarchical co-attention for Image QA

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

→ better cross-domain co-references



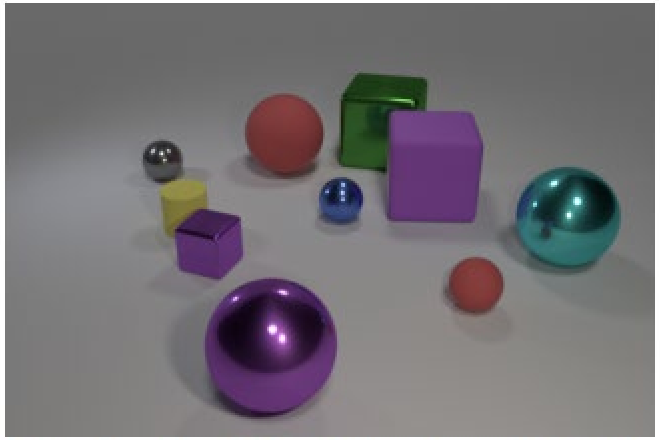
→ Q: Can this be done on text qa as well?

→ Q: How about questions with many reasoning hops?

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?

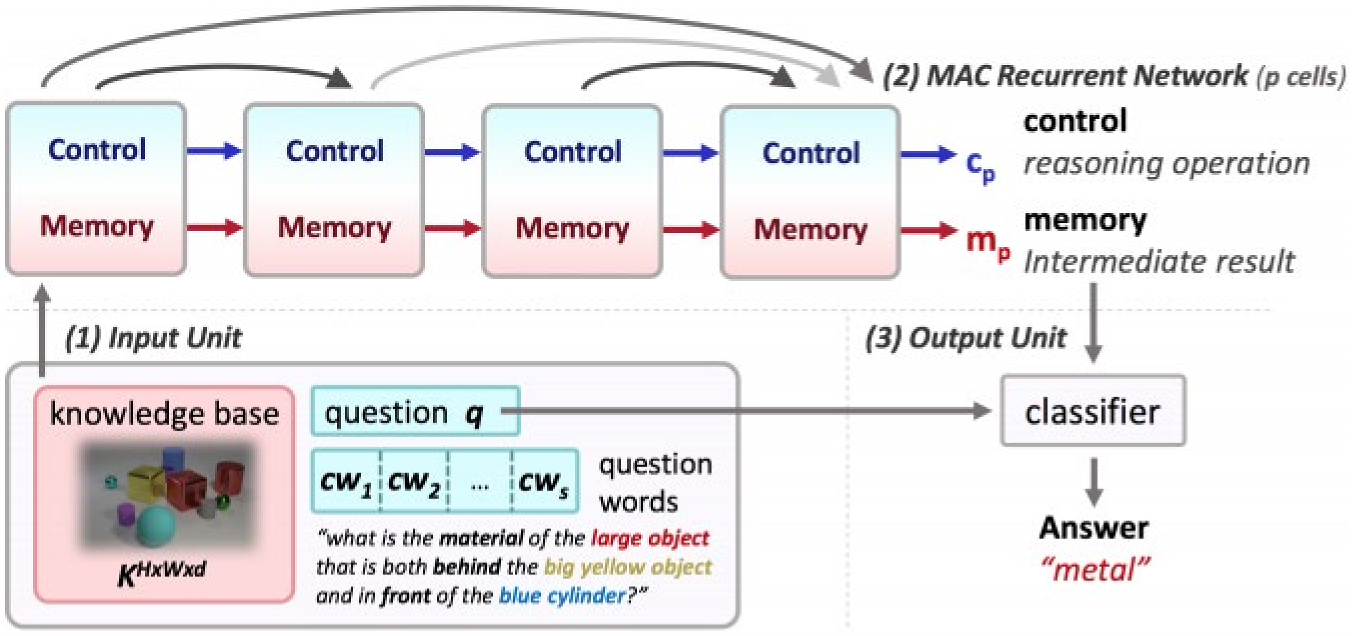


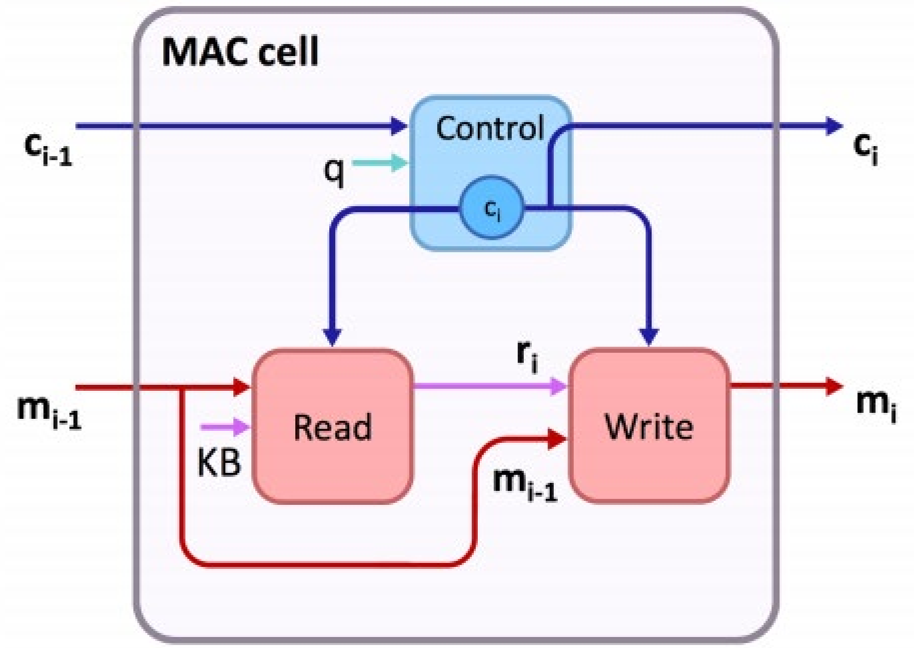
Q: *Do **the block** in **front** of **the tiny yellow cylinder** and **the tiny thing** that is to the **right** of **the large green shiny object** have the **same color**?* **A:** *No*

Figure: Hudson and Manning – ICLR 2018

Multi-step reasoning - Memory, Attention, and Composition (MAC Nets)

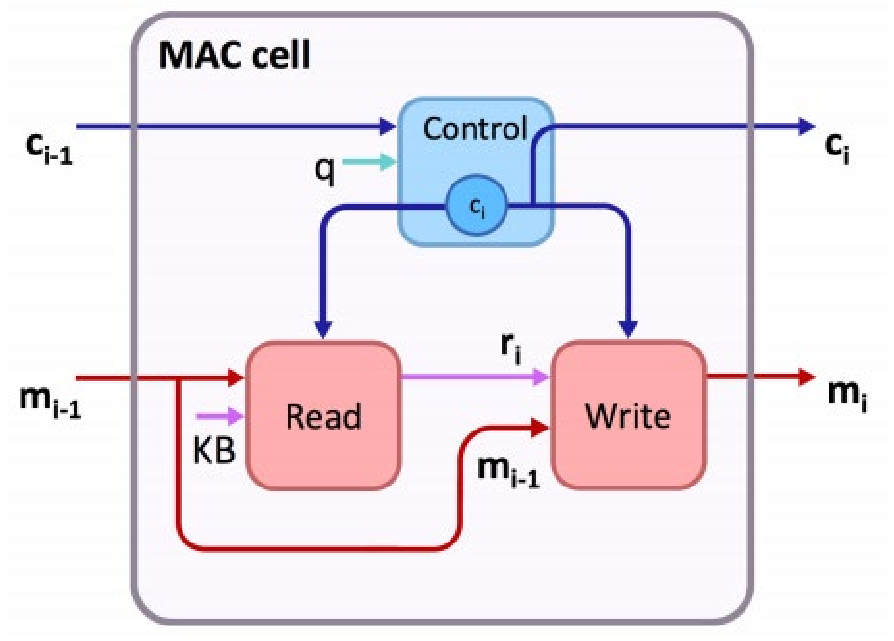
- Attention reasoning is done through multiple sequential steps.
- Each step is done with a recurrent neural cell
- *What is the key differences to the normal RNN (LSTM/GRU) cell?*
    - *Not a sequential input, it is sequential processing on static input set.*
    - *Guided by the question through a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





Multi-step attentional reasoning

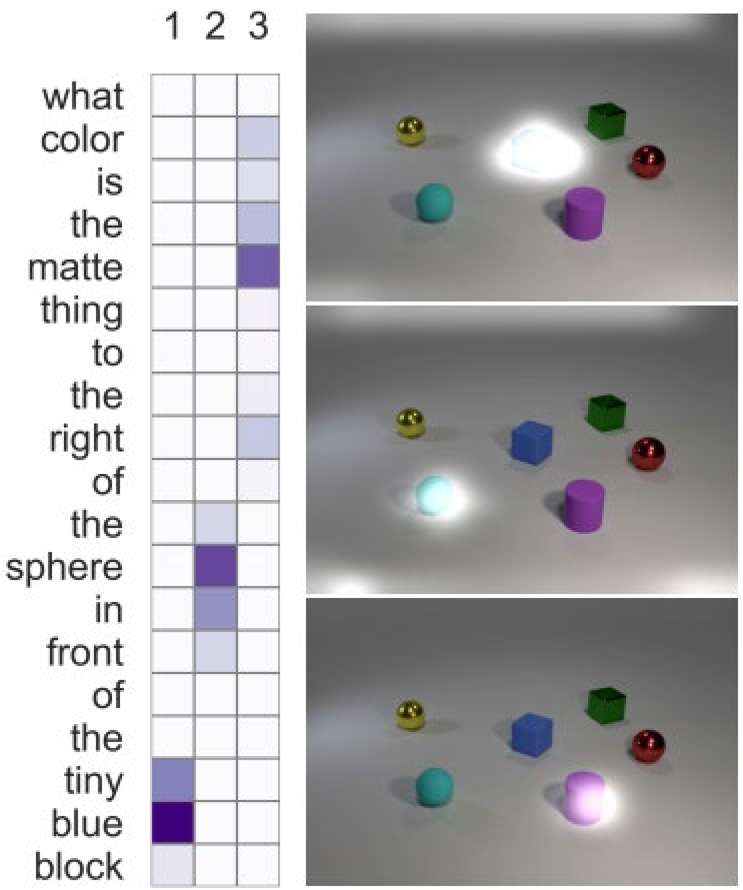
- At each step, the controller decide what to look next
- After each step, a piece of information is gathered, represented through the attention map on question words and visual objects
- A common memory kept all the information extracted toward an answer



MAC network, Hudson and Manning – ICLR 2018

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***,



→ Multi-step refinement seems to be a good reasoning strategy

→ Can we do it out of attention scheme?

Feature-wise Linear Modulation (FiLM)

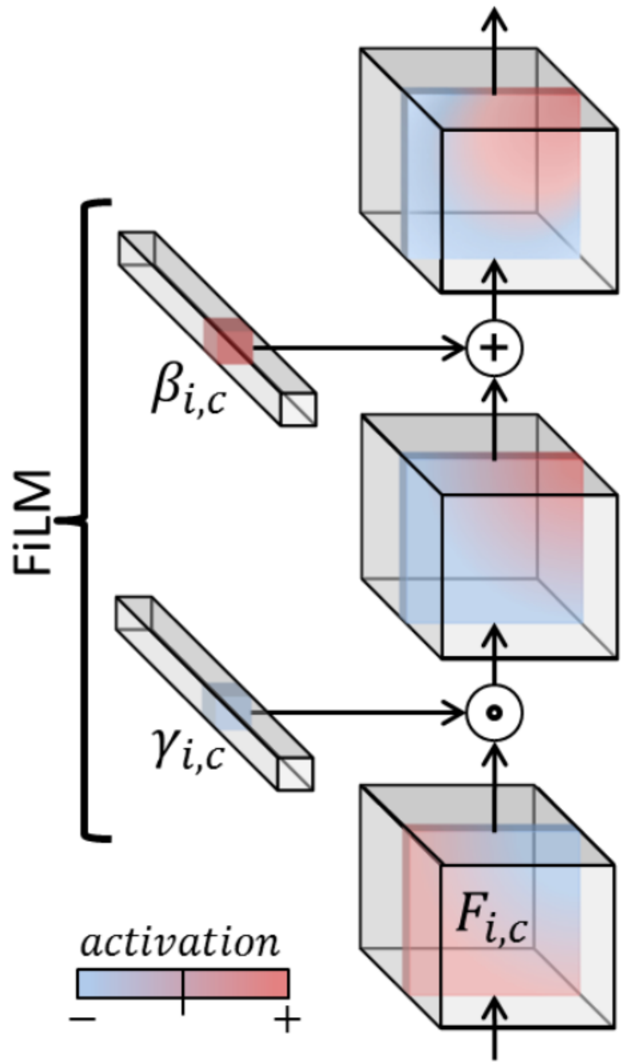
- Influence of input x to network features

$$\gamma_{i,c} = f_c(\boldsymbol{x}_i) \qquad \beta_{i,c} = h_c(\boldsymbol{x}_i)$$

- The modulation is done with an affine transform

$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}$$

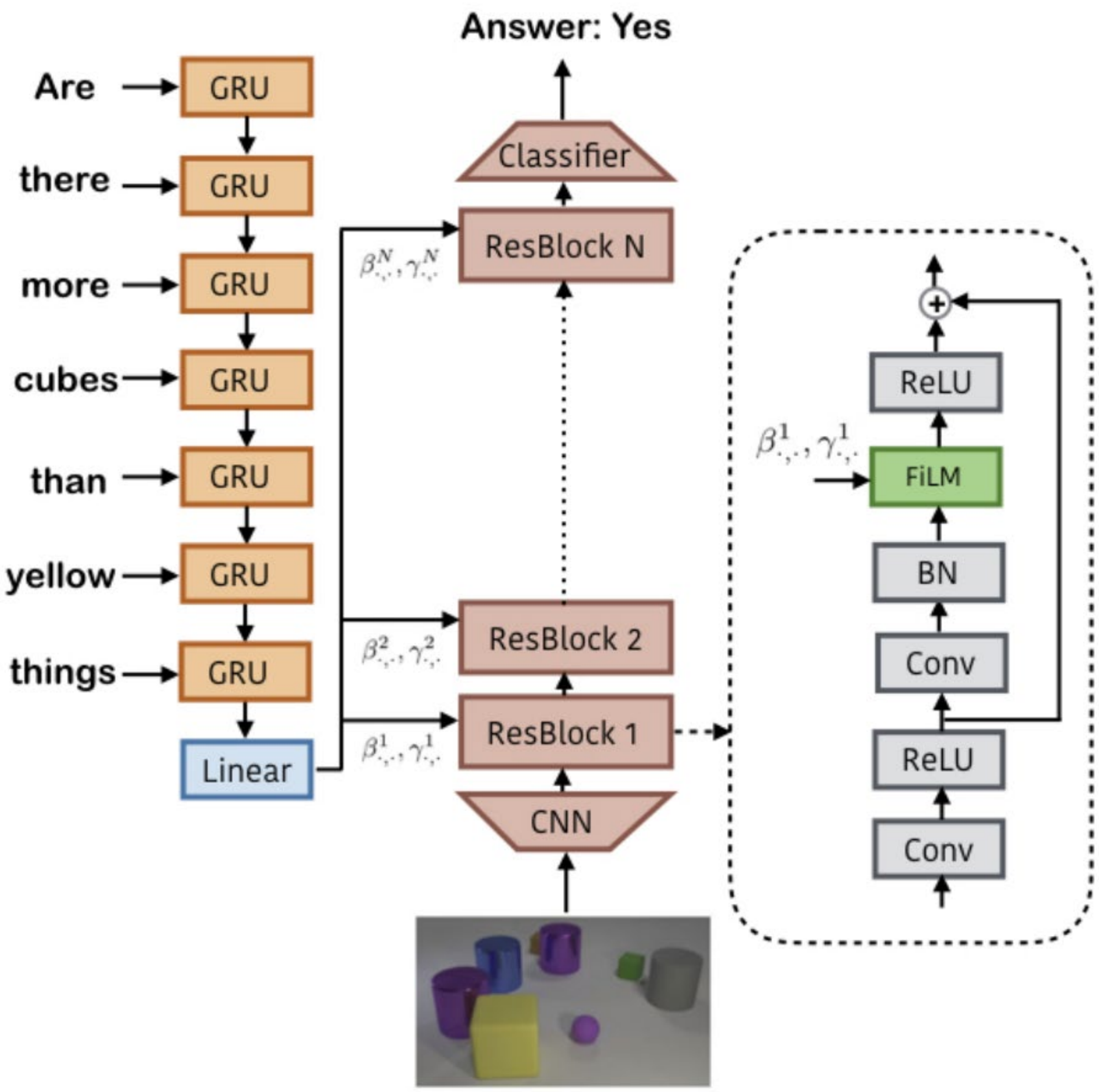
- For CNNs, f and h modulate the per-feature-map distribution of activations based on x_i, agnostic to spatial location



FiLM, Perez et.al. AAAI2018

FiLM for question answering

- Input x of modulation cues is from the question
- It is used to modulate the output of each layer of the CNN



FiLM, Perez et.al. AAAI2018

Reasoning as set-set interaction – a look back

- C : a set of context objects

$$C = \{o_1, o_2, ..., o_n\}$$

- q: a set of linguistic objects

$$Q = \{w_1, w_2, ..., w_n\}$$

- Reasoning = interaction of C and Q for the answer a
- Information refinement is the key outcome of multi-step compositional reasoning

Q:What is the brown animal sitting inside of?

→ Does it work for questions about *relations between objects*

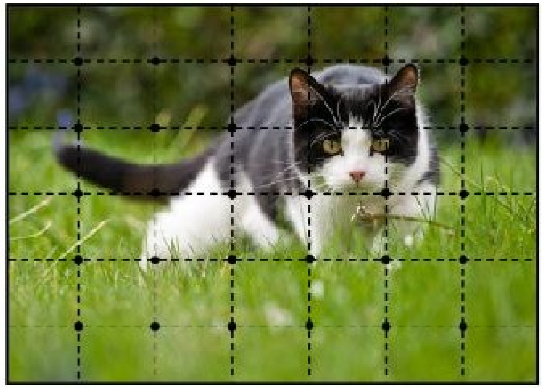
Lecture 5:
Reasoning over graphs

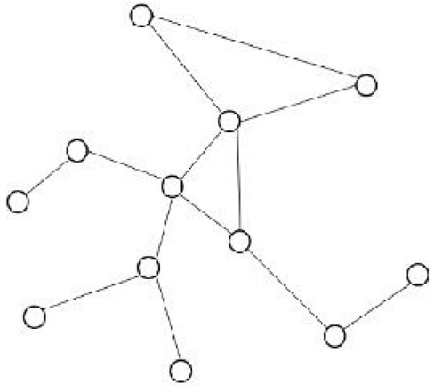
https://neuralreasoning.github.io/

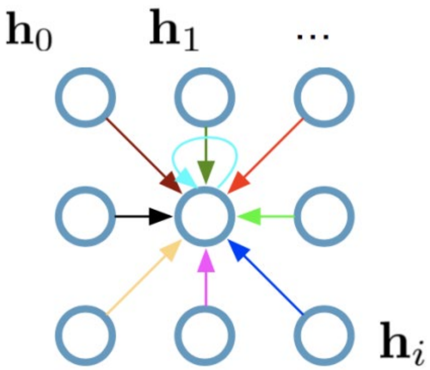
Presented by Vuong Le

Graph representation of visual data

- CNNs is a model run on an implicit grid-based graph
  - Local connections
  - Efficient weights
  - Easy to have multiple layers
- Too uniform
  - less concentration
  - less object-centric
  - restricted to locality of relations







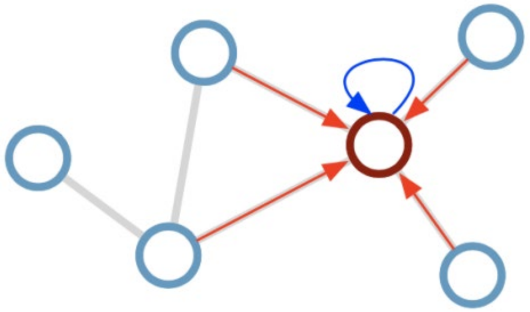


Figure credit: Thomas Kipf

Reasoning on Graphs

- Relational questions: requiring explicit reasoning about the relations between multiple objects

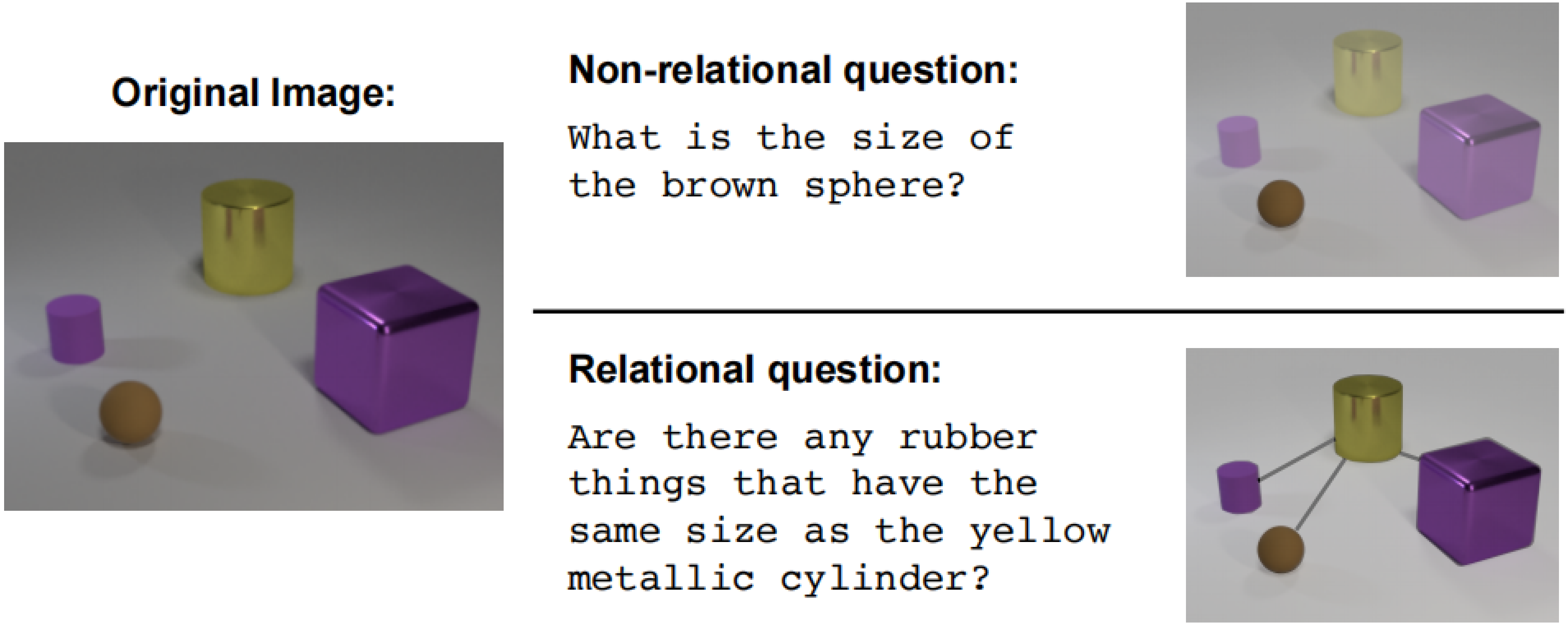
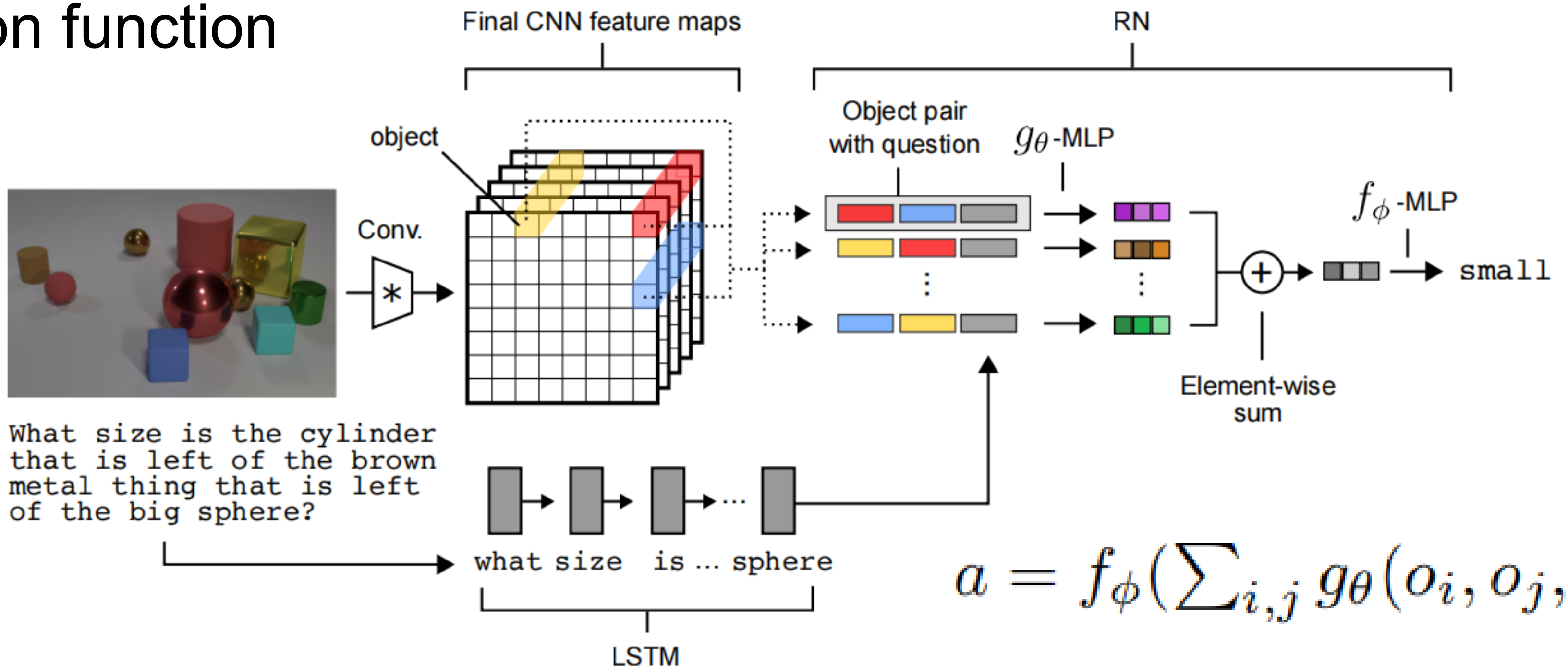


Figure credit: Santoro et al 2017

Relation networks (Santoro et al 2017)

- Relation networks $\mathrm{RN}(O) = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j)\right)$
- f_ϕ and g_θ are neural functions
- g_θ generate "relation" between the two objects
- f_ϕ is the aggregation function



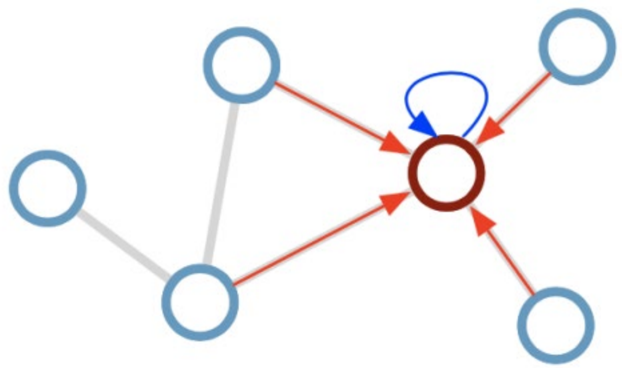
$$a = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j, q)\right)$$

→ The relations here are implicit, over-complete, pair-wise

→ inefficient, and lack expressiveness

Graph Convolutional Networks

- Update each node representation based on neighboring nodes and connected edges

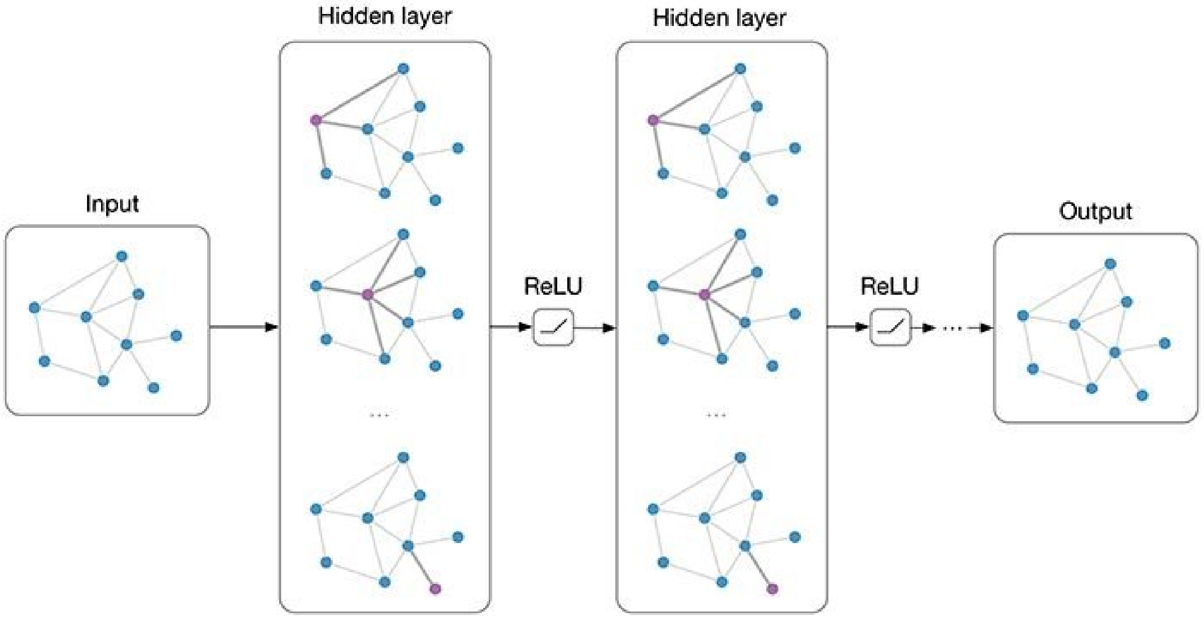


- Share the efficiency of CNN by shared weights

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

Multi-layer GCN

- Capture the dependence via message passing between nodes
- Refine node (and edge) representations
- Used for
  - node/graph classification
  - translation
  - relation discovery
  - generative models



What does it do fundamentally?

→It resembles an information refinement scheme!

→*But we need to be able to pass an arbitrary query in?*

Figure credit: Thomas Kipf

Reasoning with Graph convolution networks

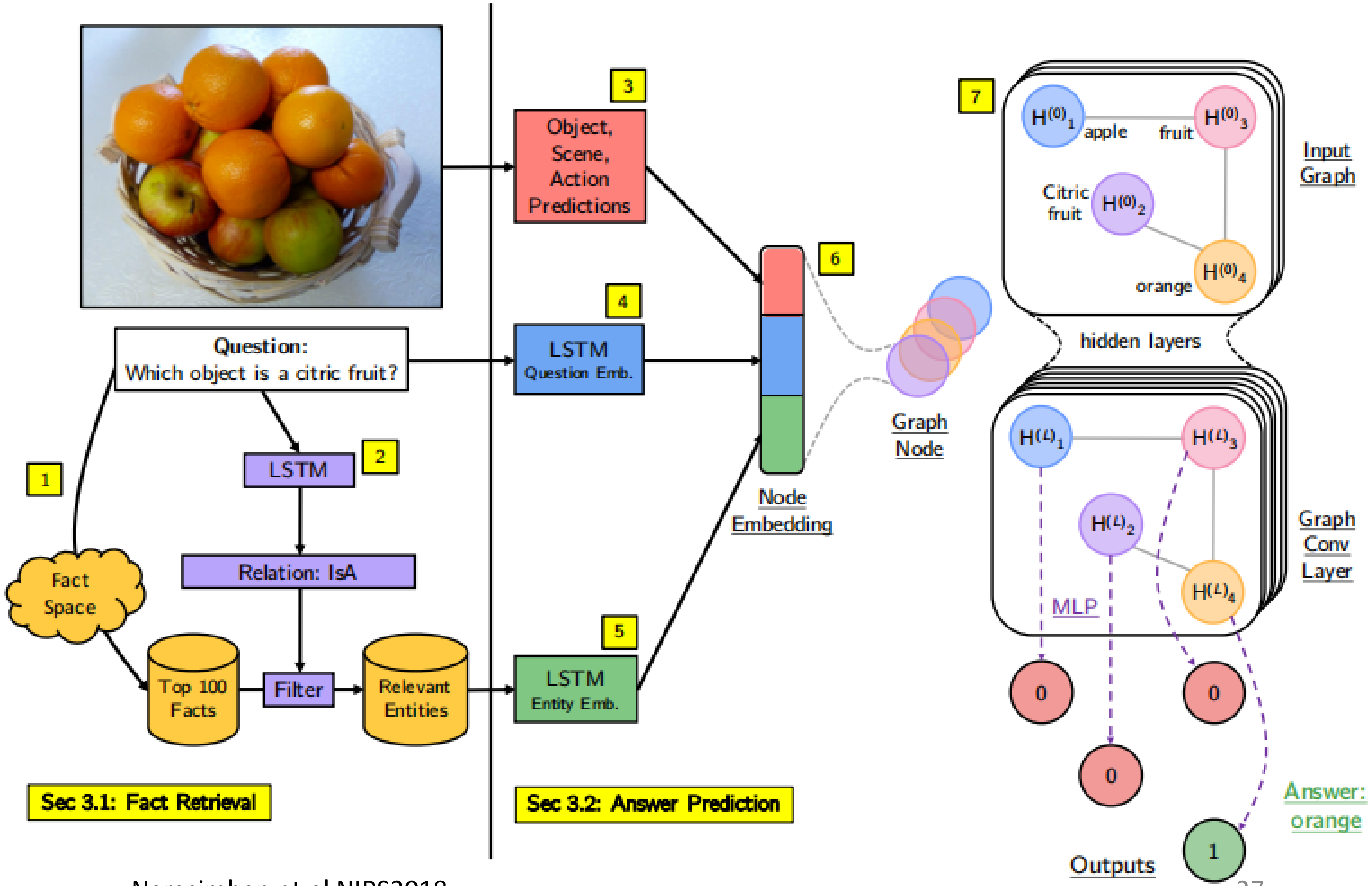
- Input graph is built from image entities and question
- GCN is used to gather facts and produce answer

→ The relations are now explicit and pruned

→ But the graph building is very stiff:

- Unrecoverable from mistakes
- Information during reasoning are not used to build graphs

→ The graphs should be dynamically constructed during reasoning



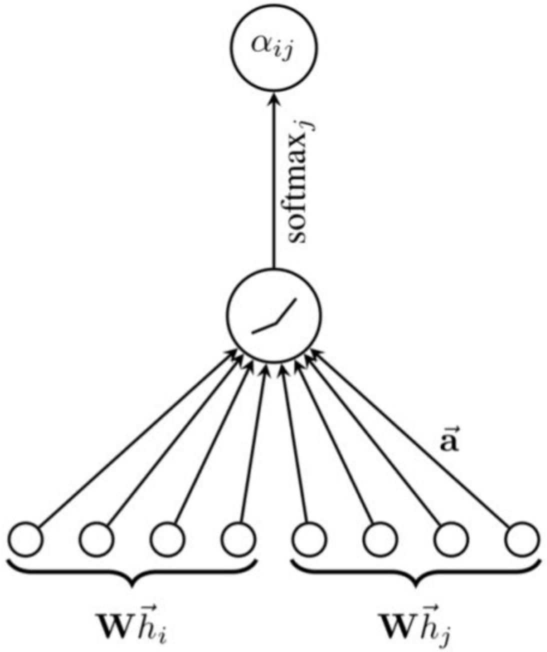
Narasimhan et.al NIPS2018

Graph Neural Networks with Attention

- Assigning different importances to nodes of a same neighborhood
- Implicitly model the edge reps
- Efficient in params
- Costly computation

(but still better than GNN with edge embeddings)

$$\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right)$$



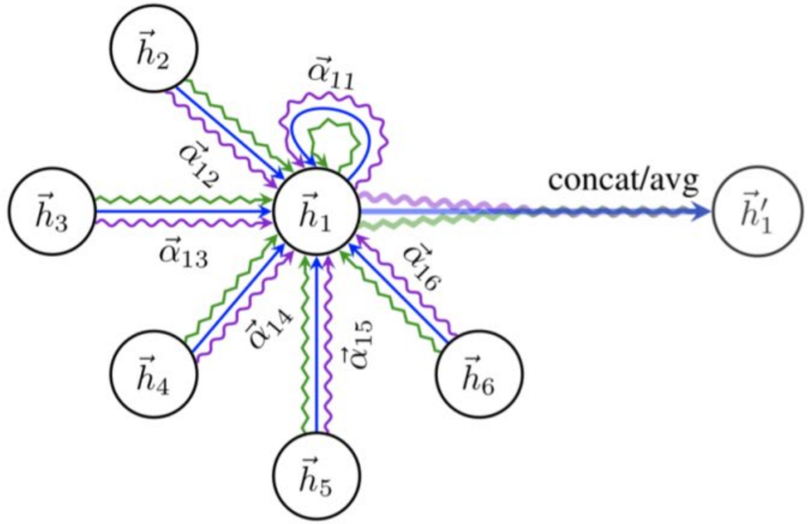


Figure credit: Graph Attention Networks Veličković et al. (ICLR 2018)

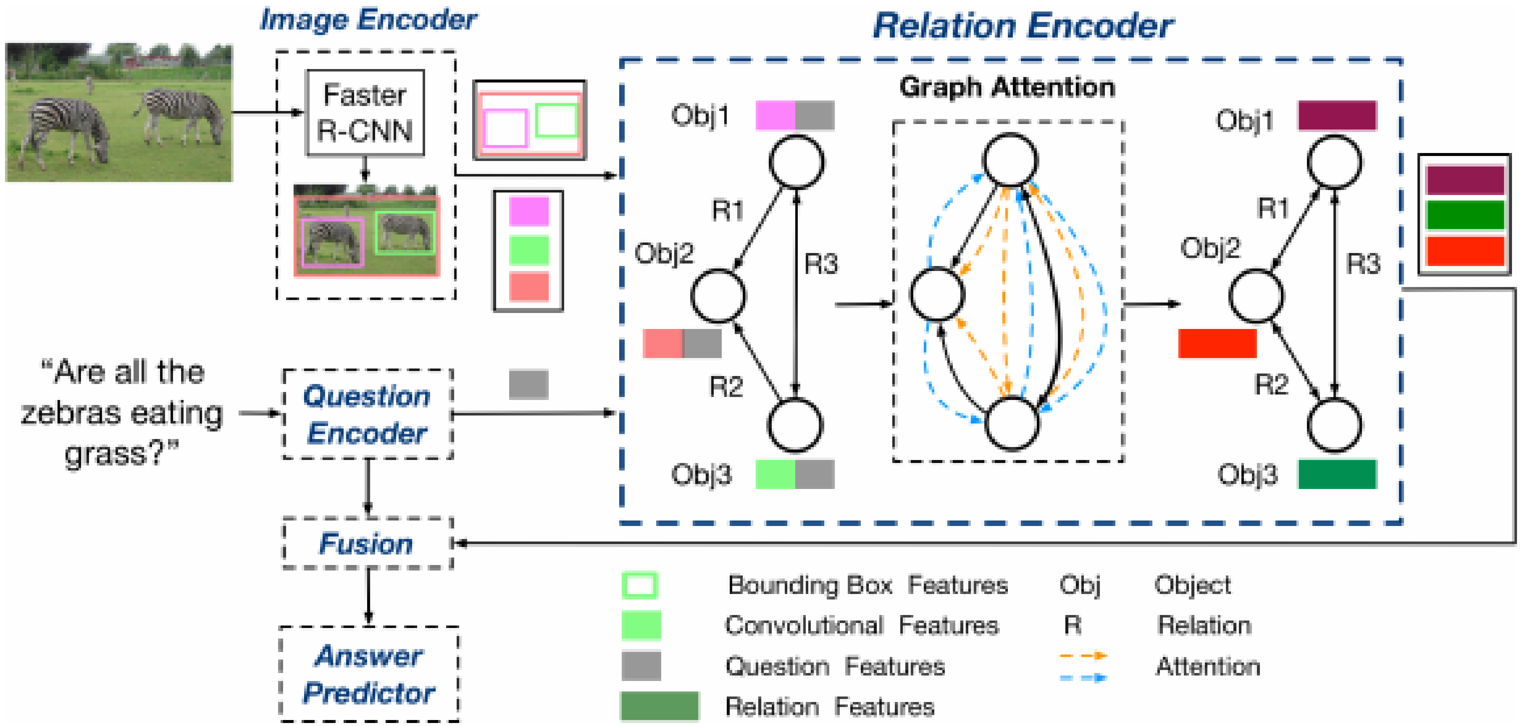
Reasoning with Graph attention networks

- The graph is determined during reasoning process with attention mechanism

→The relations are now adaptive and integrated with reasoning

→ Are the relations singular and static?

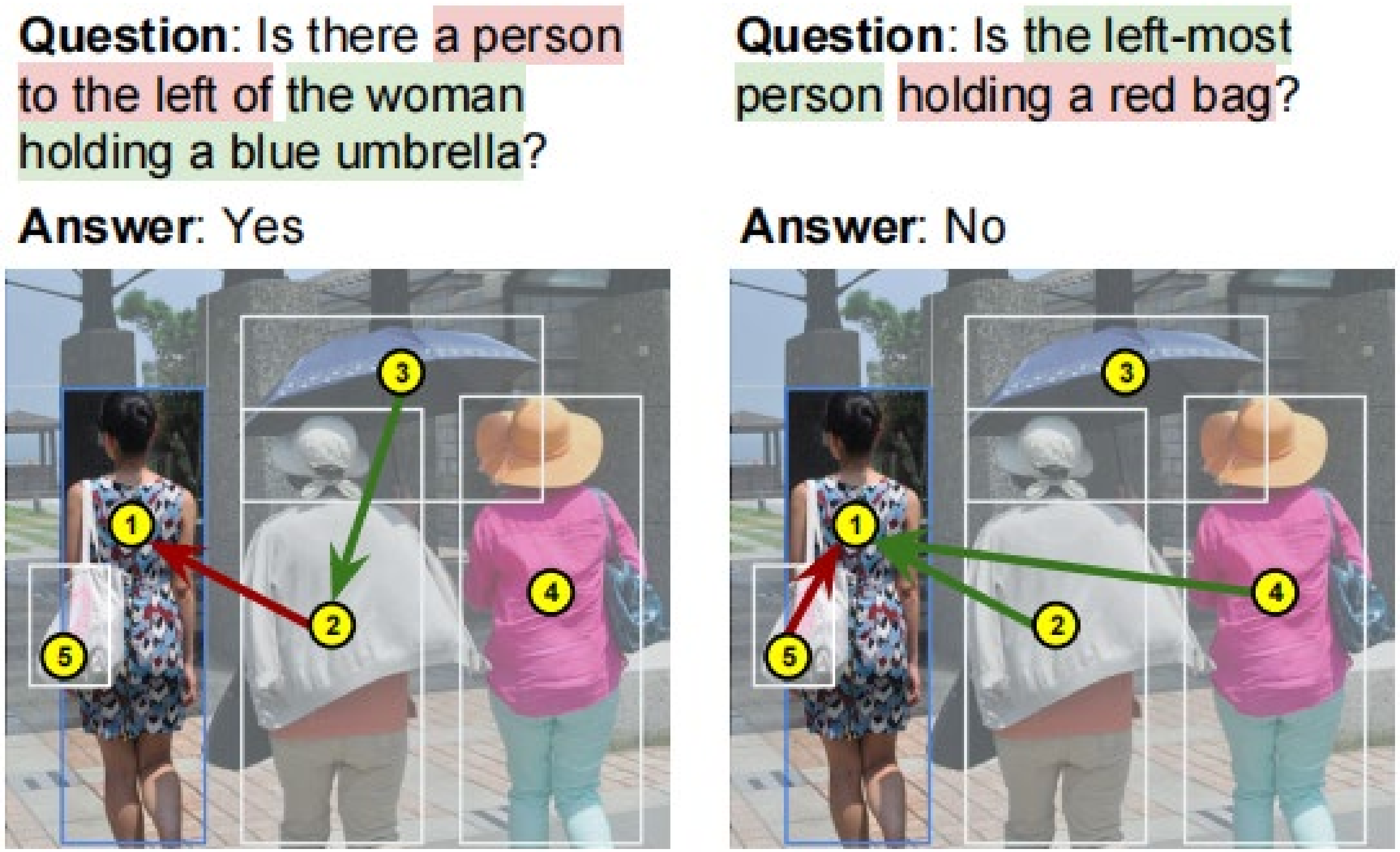
→ Reminder: reasoning is iterative!



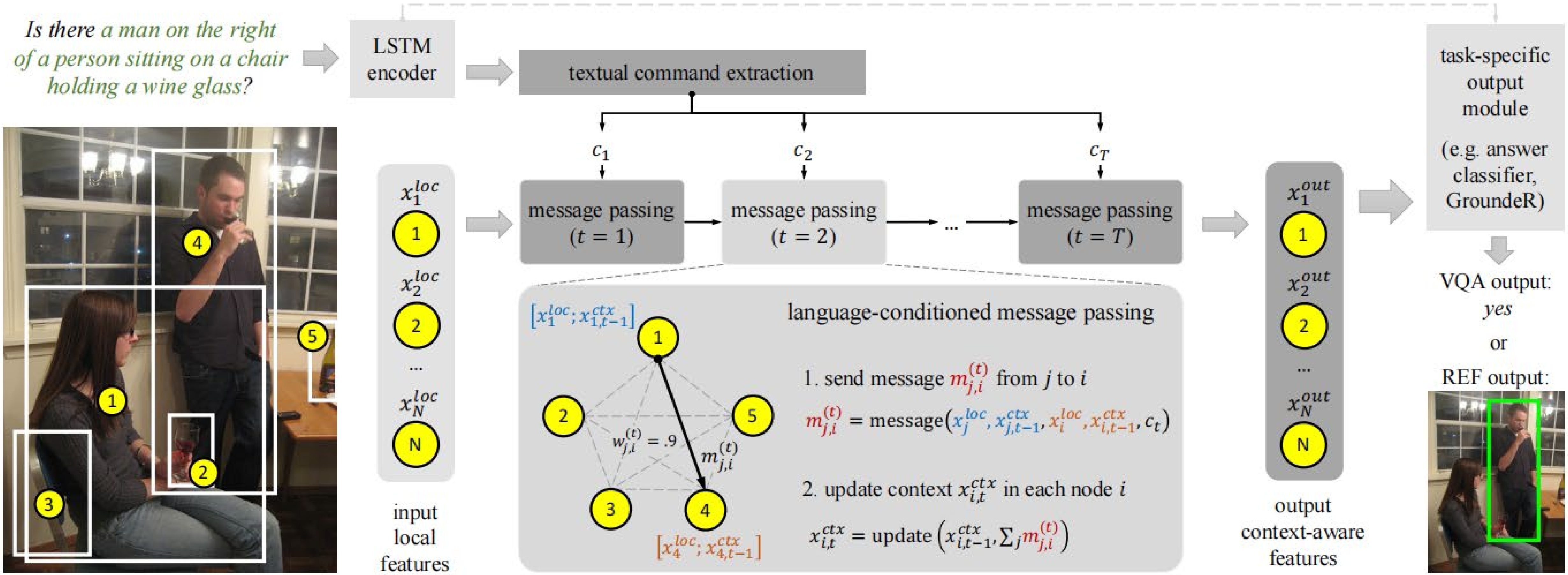
ReGAT model, Li et.al. ICCV19

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!



Dynamic reasoning graphs

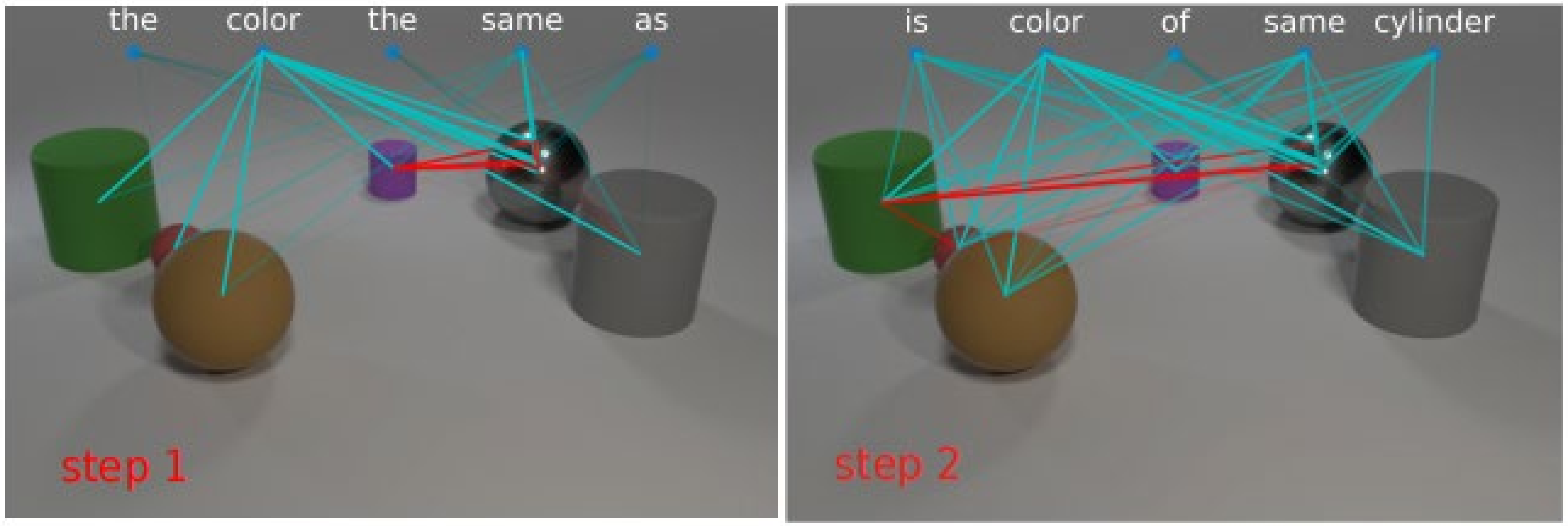


→ The questions so far act as an unstructured command in the process

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

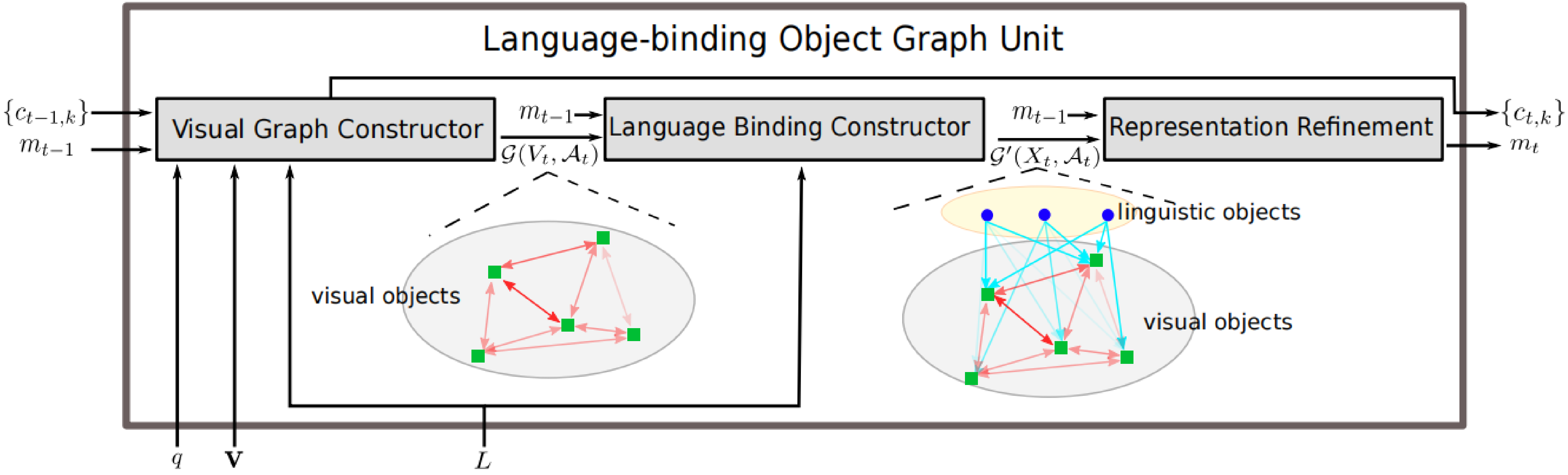
Reasoning on cross-modality graphs

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



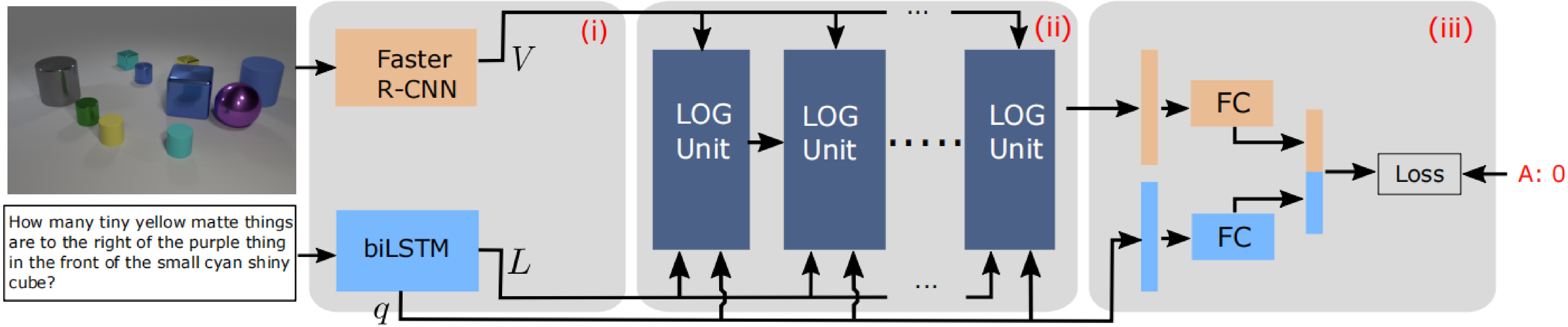
Language-binding Object Graph (LOG) Unit

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

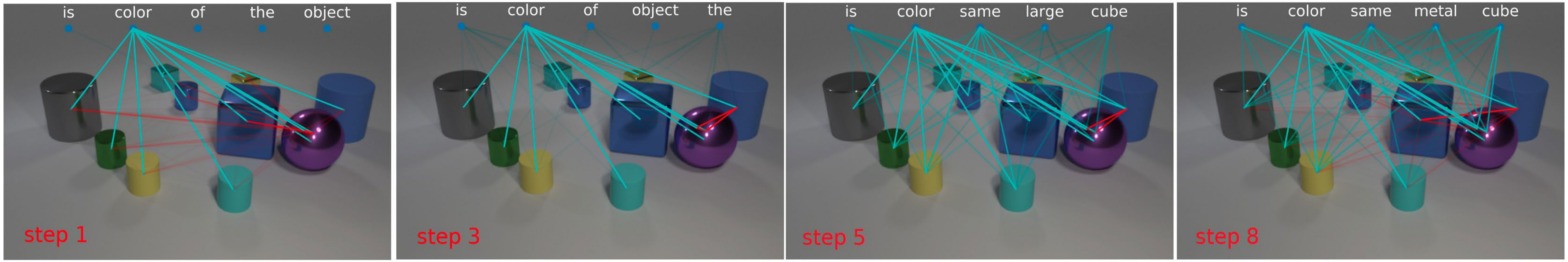


LOGNet: multi-step visual-linguistic binding

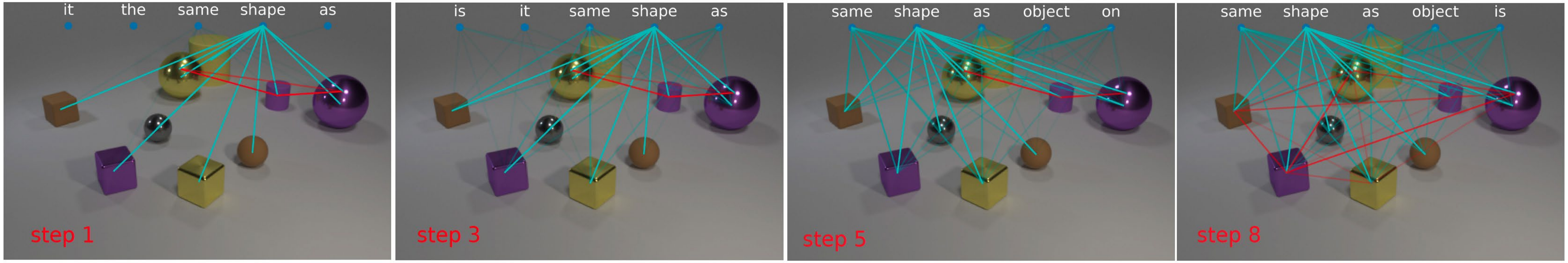
- Object-centric representation ✓
- Multi-step/multi-structure compositional reasoning ✓
- Linguistic-vision detail interaction ✓



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

We got sets and graphs, how about sequences?

- Videos pose another challenge for visual reasoning: the dynamics through time.
- Sets and graphs now becomes sequences of such.
- Temporal relations are the key factors
- The size of context is a core issue

→Lecture 8 will address these

(a) Question: What does the girl do 9 times?

Ground truth: blocks a person's punch

(b) Question: What does the man do before turning body to left?

Ground truth: breath

The two main approaches in Image QA

- **Compositional reasoning (Lecture 4 + 5)**
- Neuro-symbolic reasoning (Lecture 6)
  - Parse the question into a “program” of small logical inference steps
  - Learn the inference steps as *neural modules*
  - Use and reuse the modules for different programs

+ Explicit and interpretable

+ Close to human’s logical inference

\- Brittle, cannot recover from mistakes

\- Struggling with nuances of language and visual context

\- *Leon Bottou: Reasoning needs not to be logical inferences*

what color is the vase?

```
classify[color](
  attend[vase])
```

green (green)

Lecture 6:
Hybrid neuro-symbolic reasoning

https://neuralreasoning.github.io/

Presented by Vuong Le

The two main approaches in Image QA

- **Neuro-symbolic reasoning**
  - Parse the question into a “program” of small logical inference steps
  - Learn the inference steps as *neural modules*
  - Use and reuse the modules for different programs

  \+ Explicit and interpretable

  \+ Close to human’s logical inference

  \+ **Strongly support generalization**

  \- Brittle, cannot recover from mistakes

  \- Struggling with nuances of language and visual context
- Compositional reasoning

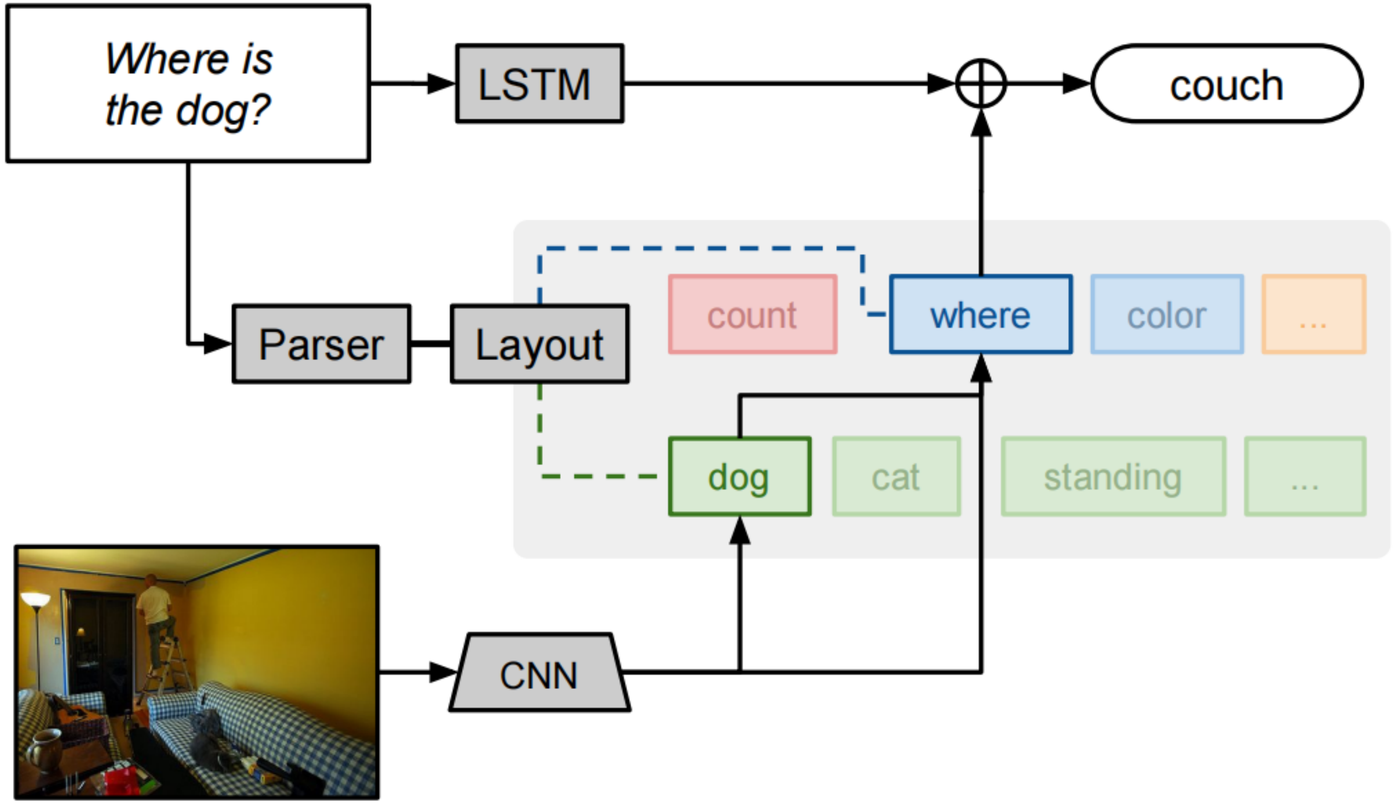
what color is the vase?

```
classify[color](
  attend[vase])
```

green (green)

Neural Module Networks

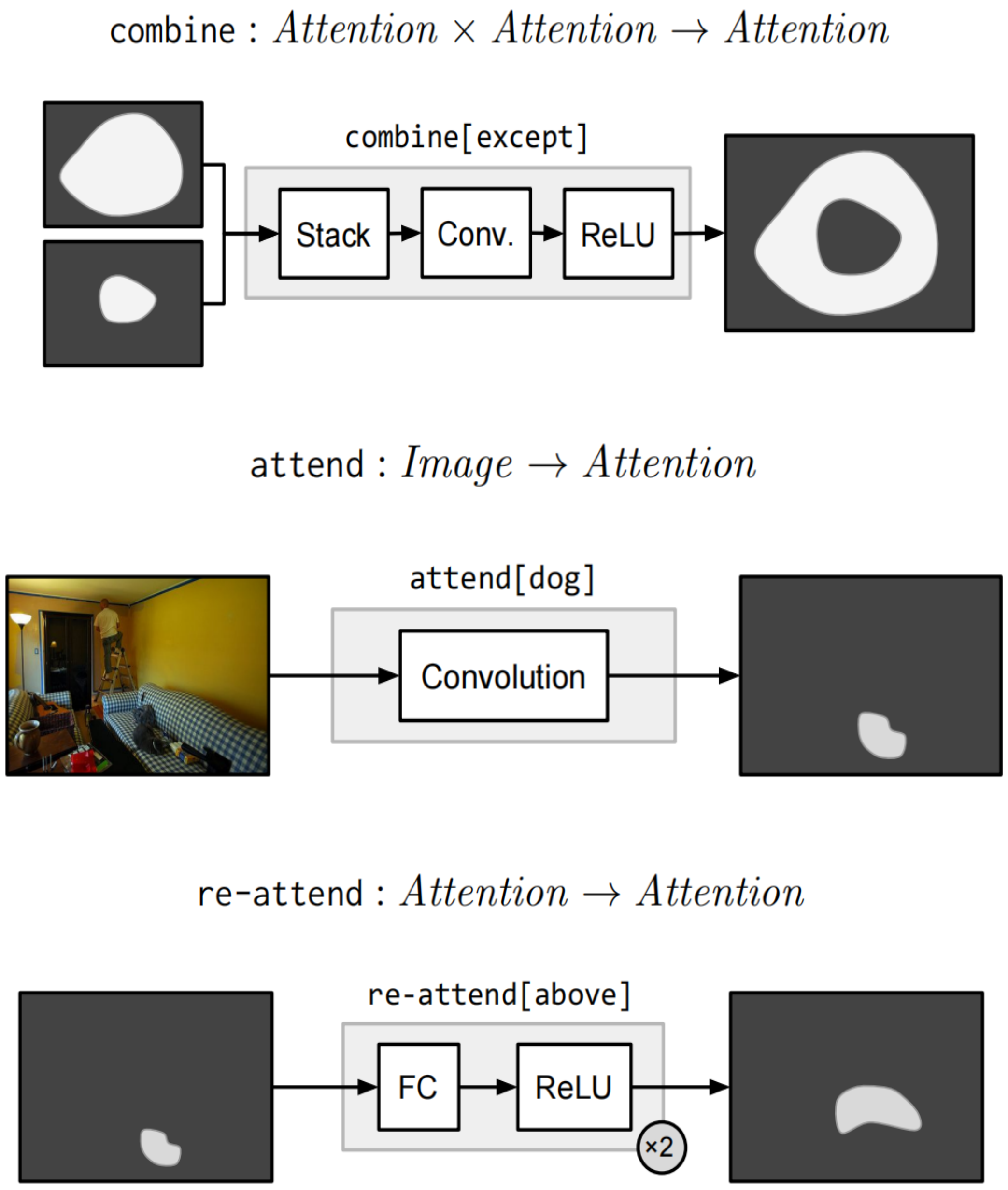
- NLP parser to build program
- The layout consists of modules which are learnable sub-networks
- Use attention as key compositional operator



Neural Module Networks, Andreas et.al CVPR 16

Modules

- attend[c] has weights distinct for each c to produce a heatmap
- re-attend[c] is MLP mapping from one attention to another.
- combine[c] merges two attentions
- into a single attention.

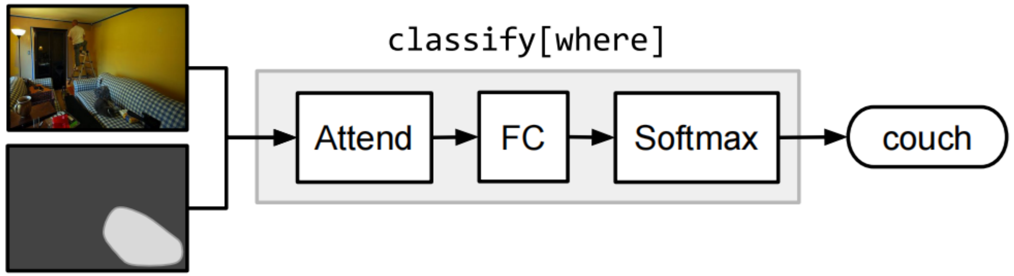


Neural Module Networks, Andreas et.al CVPR 16

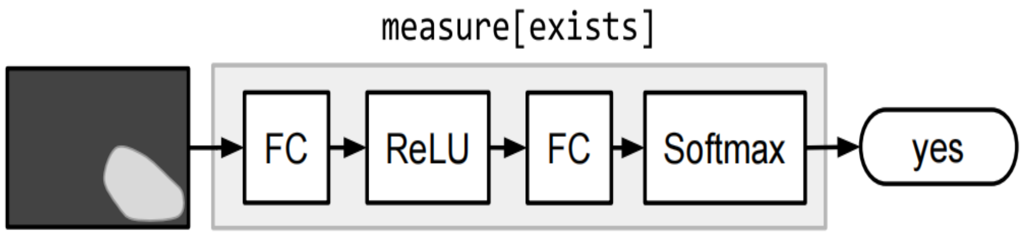
Modules

- classify[c] takes an attention and the input image and maps them to a distribution over labels.
- measure[c] takes an attention alone and maps it to a distribution over count labels

$\texttt{classify} : \mathit{Image} \times \mathit{Attention} \to \mathit{Label}$



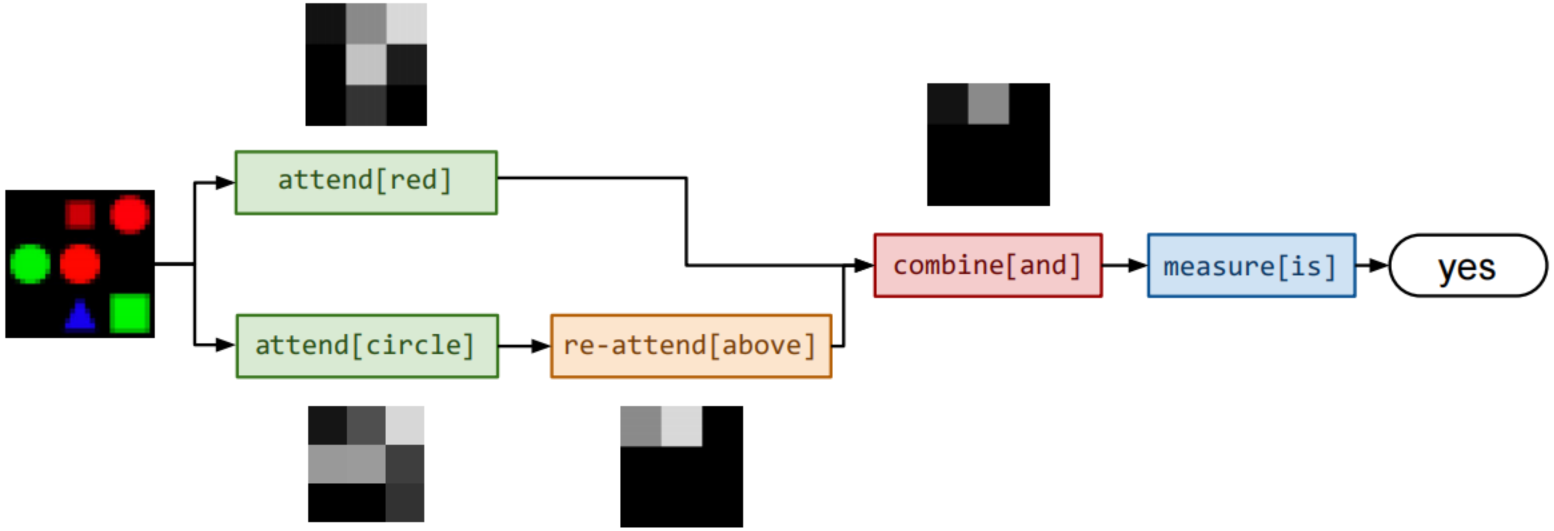
$\texttt{measure} : \mathit{Attention} \to \mathit{Label}$



Parsing

- Stanford parser: create grammatical dependency tree
- Forming the layout
  - Leaves become attend modules
  - Internal nodes become re-atten or combine
  - Root nodes become classify or measure depend on the question type

Neural Module Networks – example



Is there a red shape above a circle?

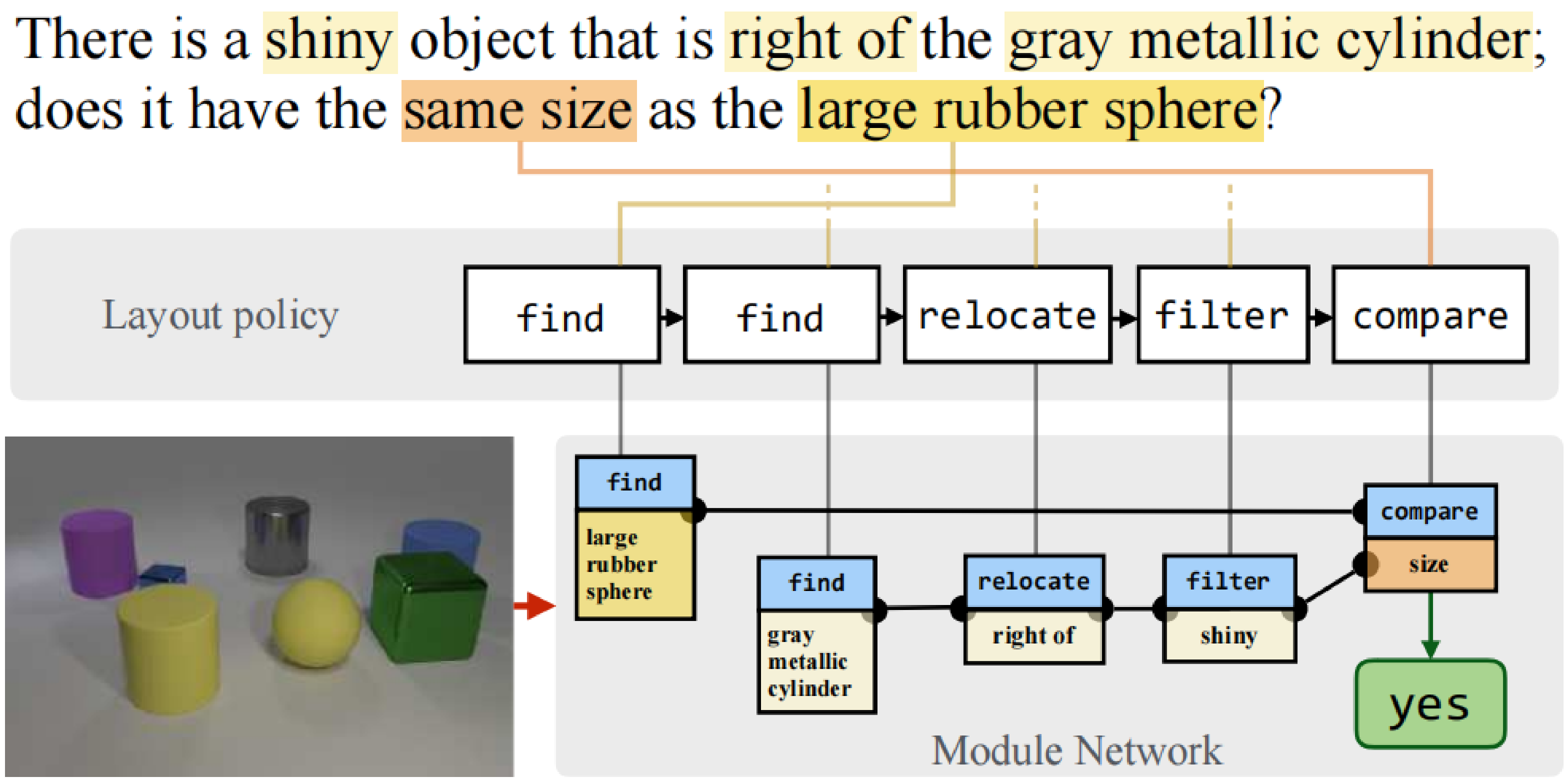
→ Relying on an off-the-shelf parser. What if it makes a mistake?

Can the two steps be connected?

Neural Module Networks, Andreas et.al CVPR 16

End-to-End Module Networks

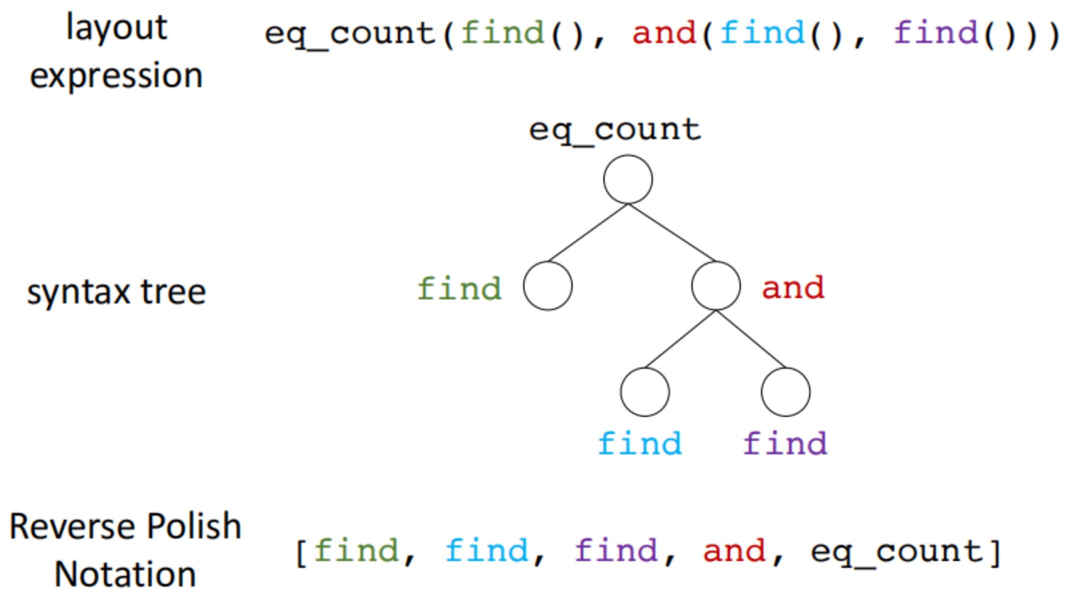
- Construct the program internally
- The two parts are jointly learnable



End-to-End Module Networks , Hu et.al., ICCV 17

Layout policy

- A layout can be linearized into a sequence
- Then a layout prediction turns into seq-2-seq problem
- And can be done by an RNN encoder-decoder arch.



End-to-End Module Nets

- Layout policy $p(l|q;\theta)$
- QA loss according to such policy $\tilde{L}(\theta, l; q, I)$
- End-to-end loss $L(\theta) = E_{l \sim p(l|q;\theta)}[\tilde{L}(\theta, l; q, I)]$

  - This loss is not fully differentiable as l is discrete

  →Policy gradient for non-diff parts, estimated through MC sampling

  - Still a very hard problem as the two parts are more or less independent.

  → Direct supervision of $p(l|q;\theta)$ using some expert policy

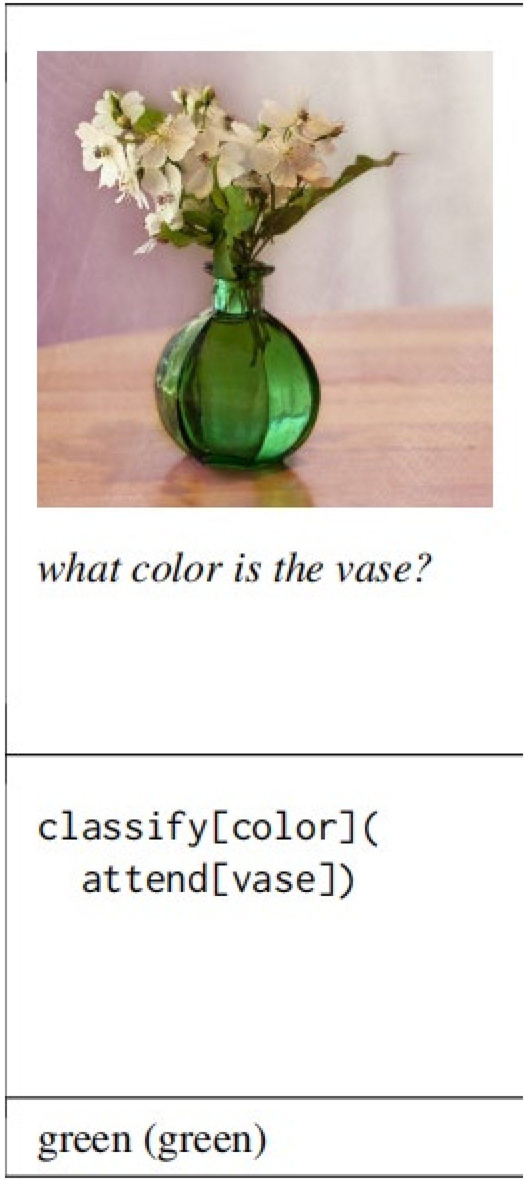
Combine the two main reasoning approaches

- Neuro-symbolic reasoning vs Compositional reasoning
  - \+ Explicit and interpretable
  - \+ Close to human's logical inference
  - \+ Strongly support generalization
  - \- Brittle, cannot recover from mistakes
  - \- Struggling with nuances of language and visual context

→Can we combine the two?

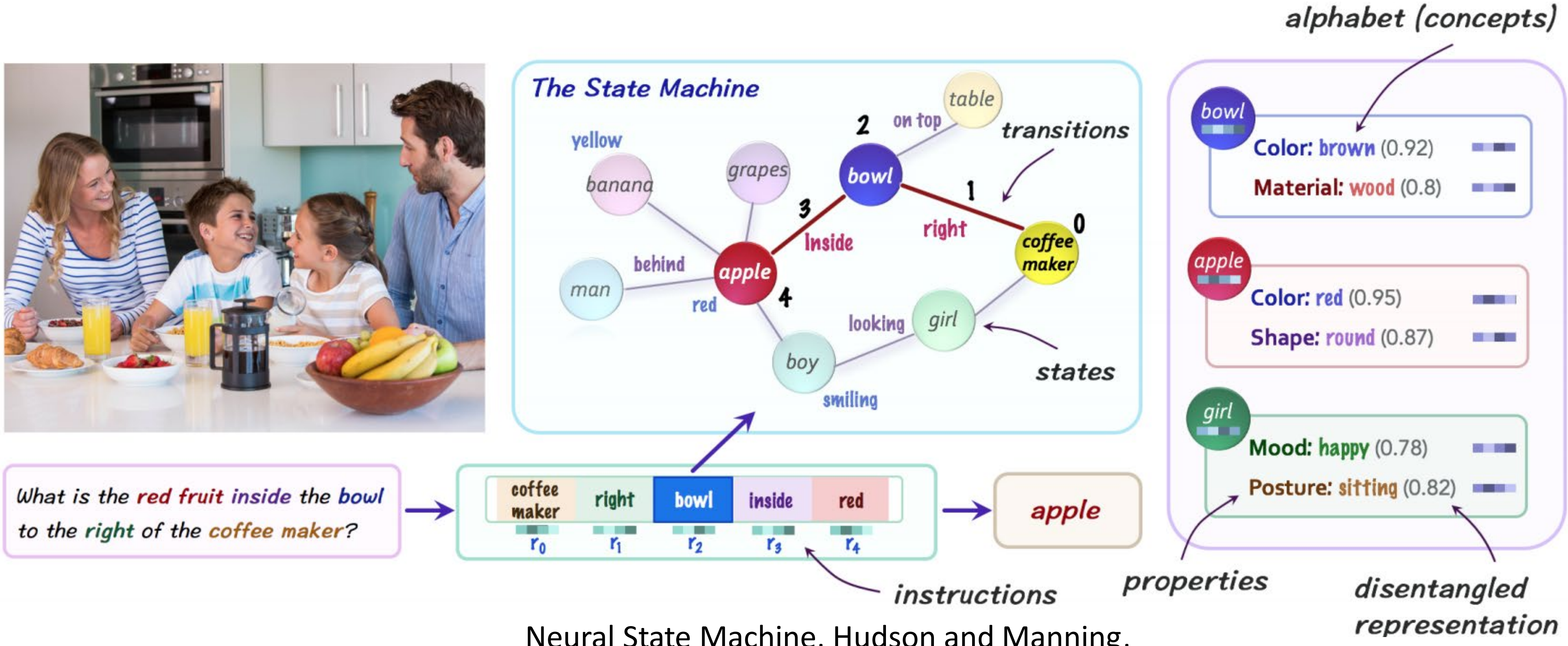
→Process questions into a series of symbolic instructions

→Use the instructions for guide the compositional reasoning process



Neural State Machine

- Generate a scene graph from image
- Translate question into a series of instructions
- Traverse the graph using the instruction toward the answer



Neural State Machine, Hudson and Manning, NeurIPS 2019

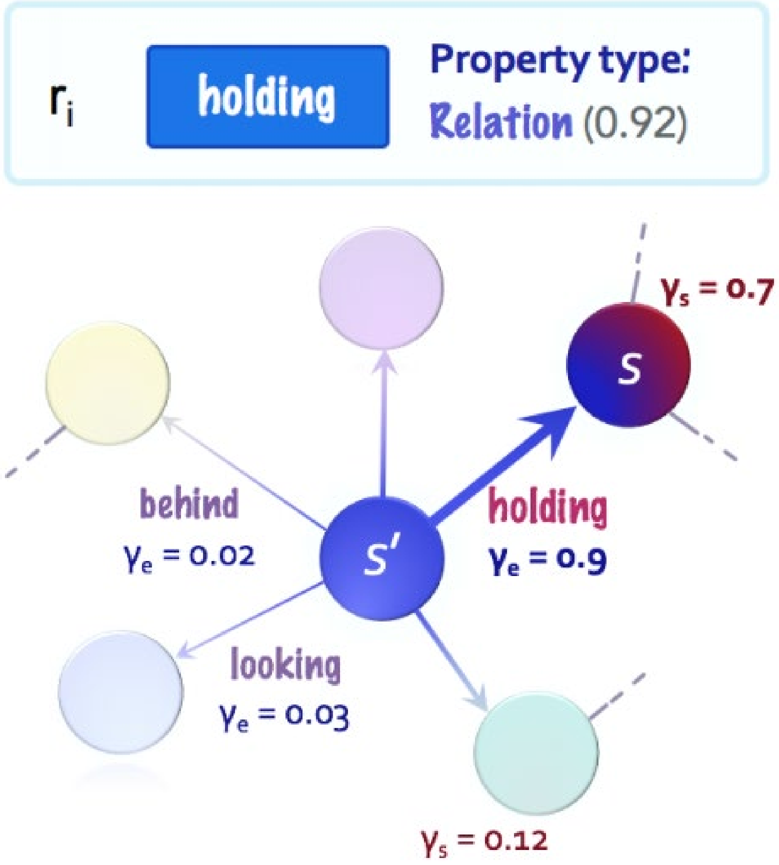
Neural State Machine $(C, S, E, \{r_i\}_{i=0}^{N}, p_0, \delta)$

- C: Concepts: *obj identity, attributes, relation*
- S: States: *objs detected in image*
- E: Transition edges between the states: *relations of objs*
- r_i a sequence of instructions: *encoded from the question*
- p_0 $: S \rightarrow [0, 1]$ distribution of the initial state.
- $\delta_{S,E}$ $: pi \times ri \rightarrow pi{+}1$ a state transition function
  - a neural module that at each step i
  - considers the distribution pi over the states as well as an input instruction ri
  - redistribute the probability along the edges, yielding an updated state distribution pi+1.

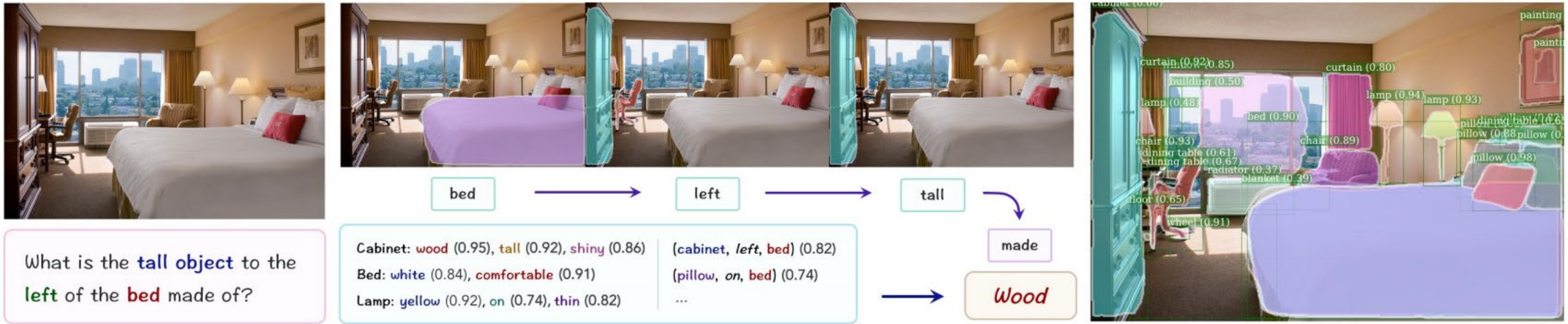
State transition

Attention is being shifted from one node to its neighbor along the most relevant edge.

- Explicit reasoning ✓
- Multi-step information refinement ✓
- Dynamic structure reasoning ✗



NSM in action



→Is the sequential order of reasoning necessarily the (inverse) order of the words in question?

→Is the reasoning state transitions only attention shifting?

→The gap between symbolic and compositional reasoning is still there