Reasoning over unstructured sets

https://neuralreasoning.github.io/

Presented by Vuong Le

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 or even translation?

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

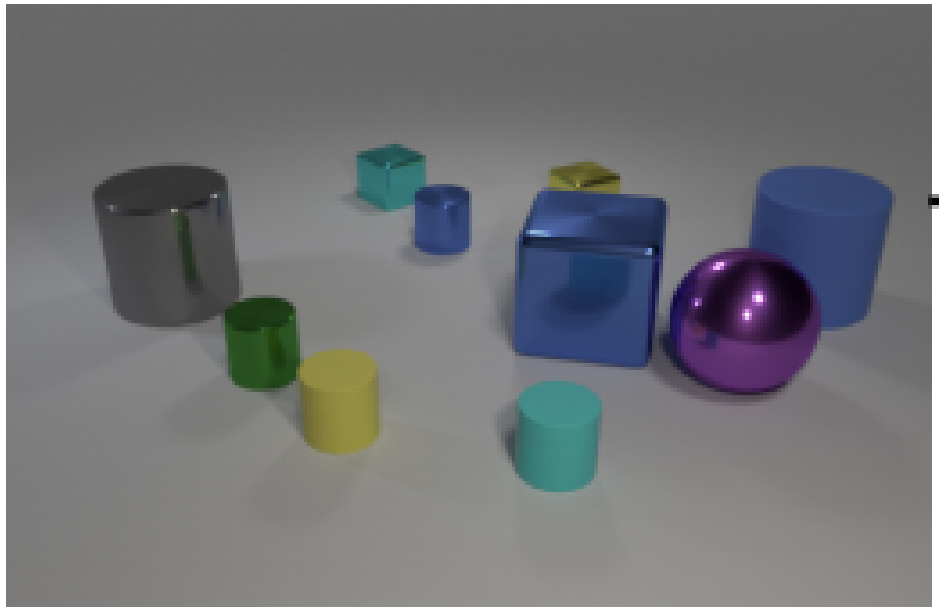
→ We can solve (some parts of) it with good structures and inference strategies

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 specific details of Image QA: Lecture 7



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

The two main approaches in Image QA

- 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*
- **Compositional reasoning (This lecture + Lecture 5)**

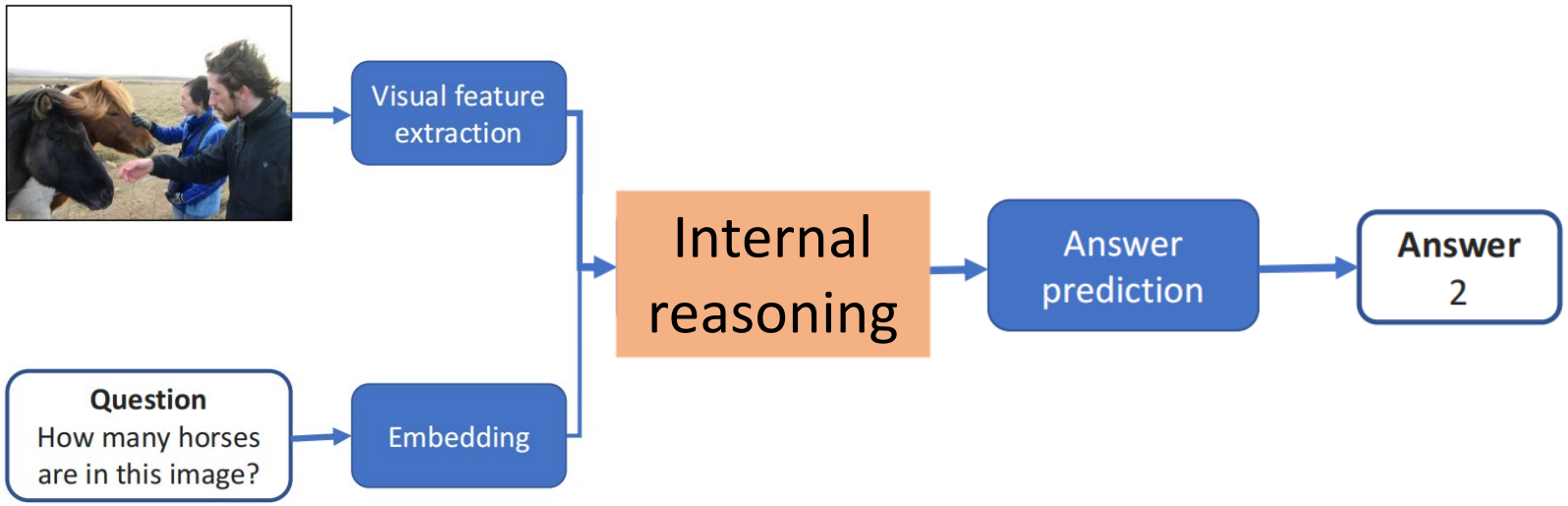
what color is the vase?

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

green (green)

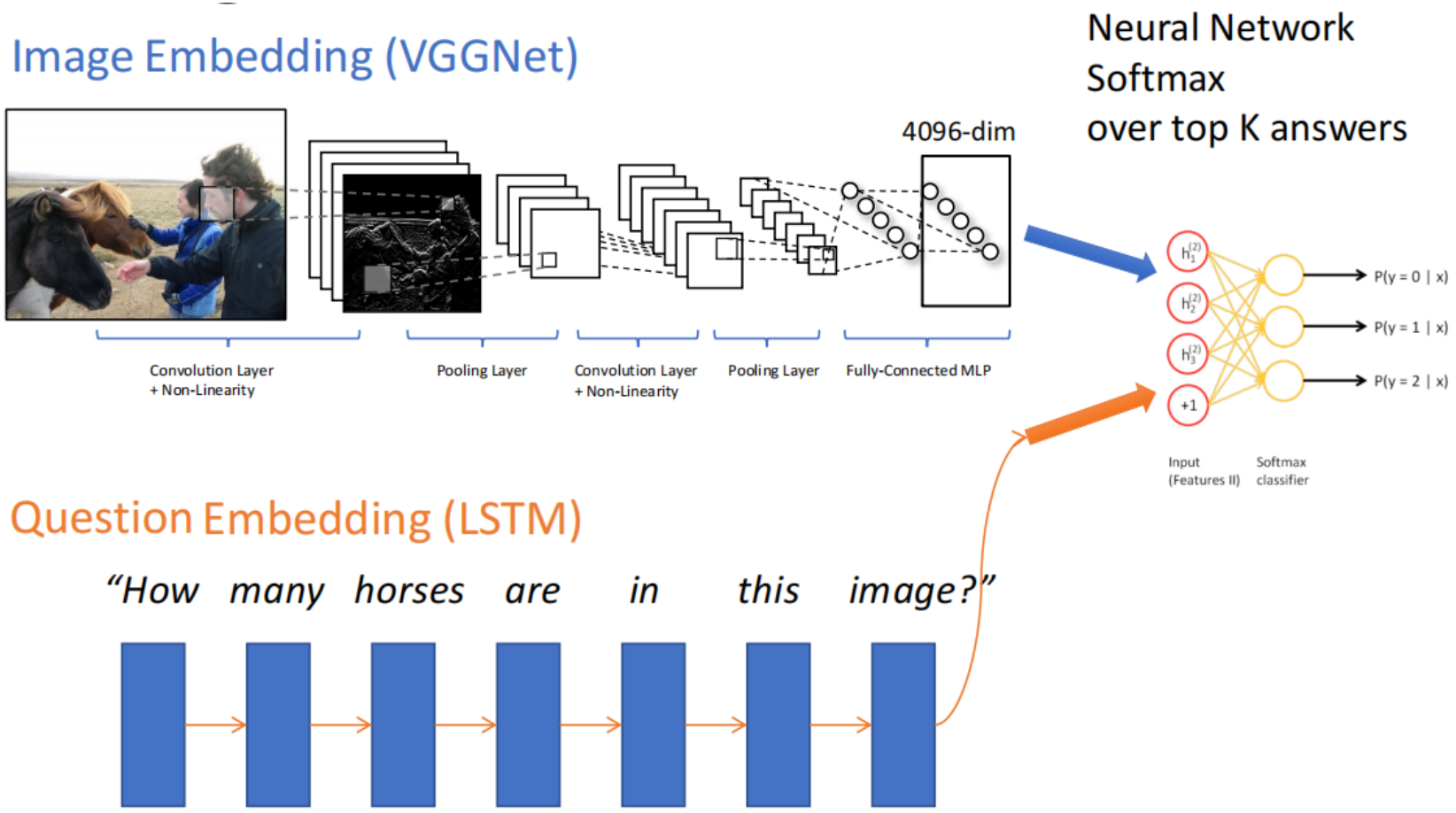
Compositional reasoning

- 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: Attention?**

Agenda

- Cross-modality reasoning, the case of vision-language integration.
- **Reasoning as set-set interaction.**
- Relational reasoning
- Temporal reasoning
  - Video question answering.

Reasoning as set-set interaction

- C: a set of context objects

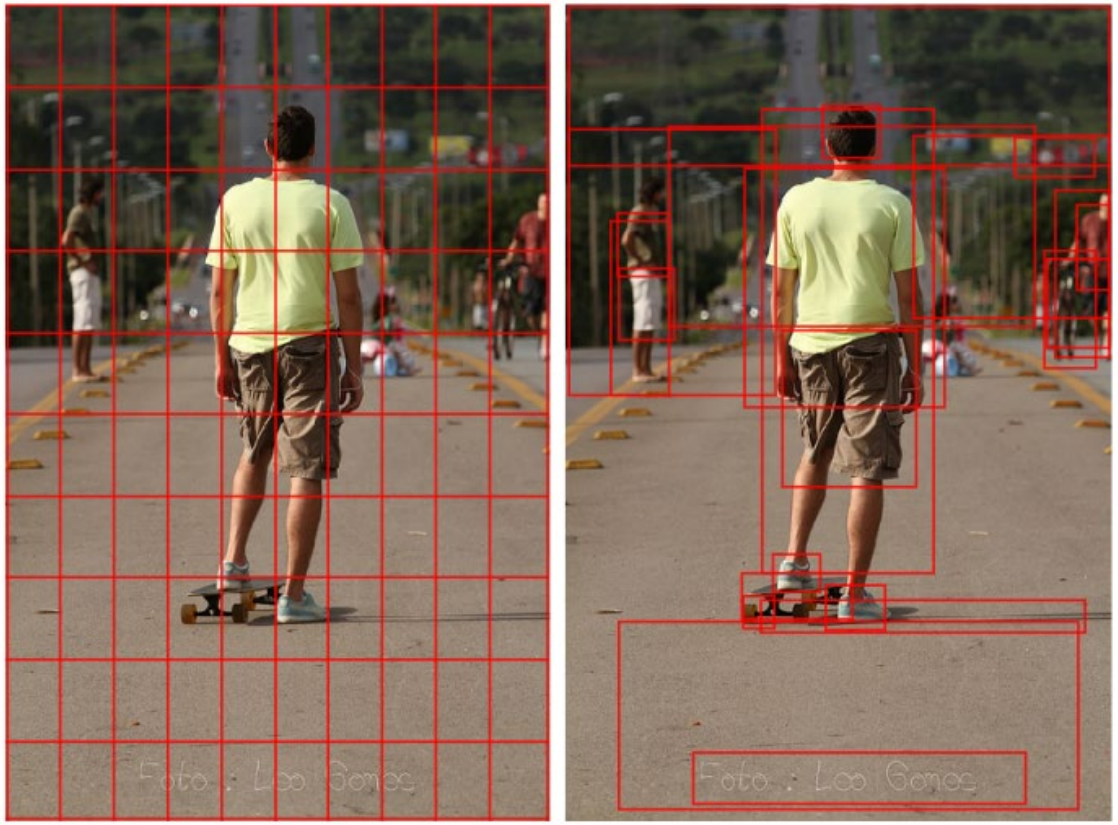
$$C = \{o_1, o_2, ..., o_n\}$$

  - 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)

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

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

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

→ 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:

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

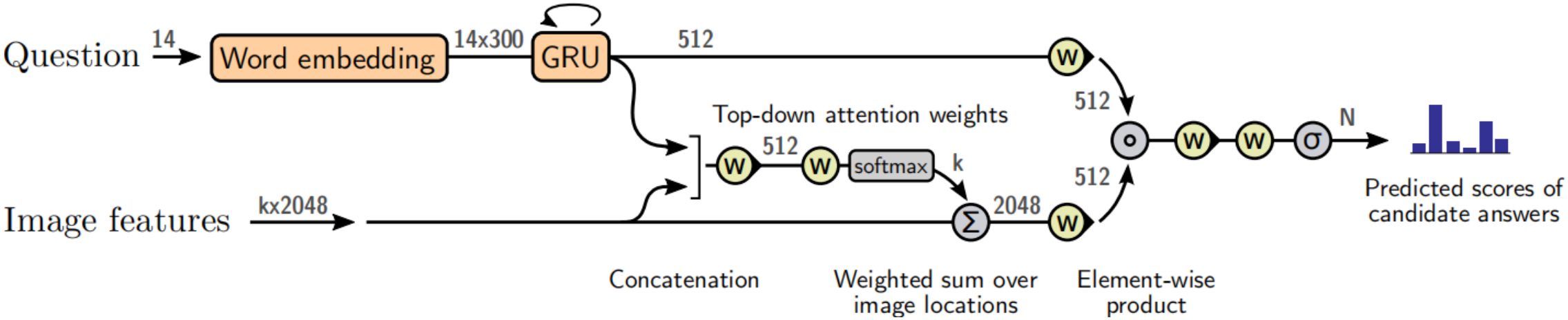
$$\alpha_i = \frac{\exp(\mathbf{W} s_i)}{\sum_j \exp(\mathbf{W} s_j)}$$

  - 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

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?

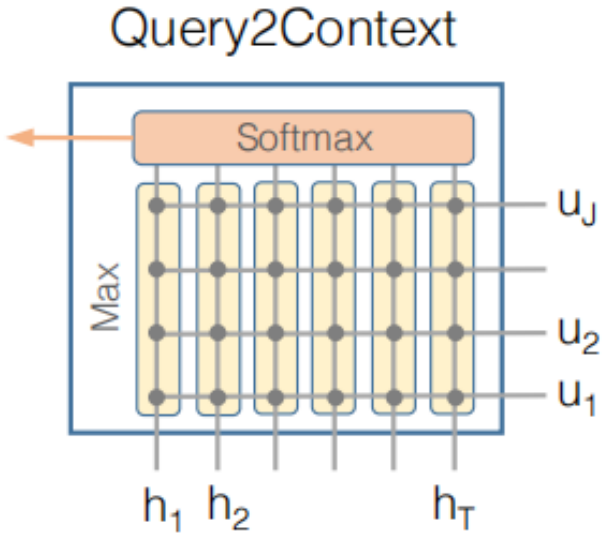
Bi-directional attention

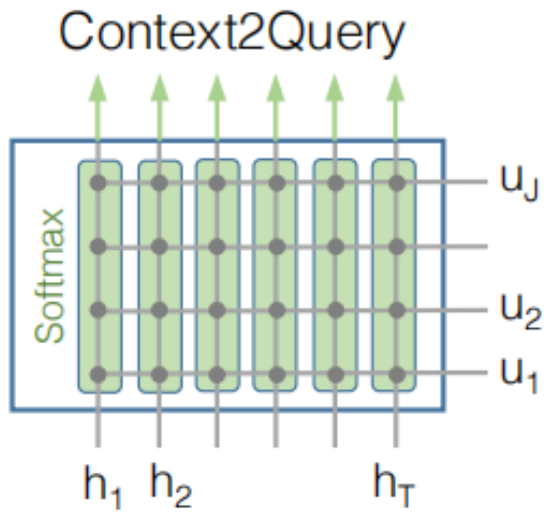
- Question-context similarity measure

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

- Question-guided context attention
  - Softmax across columns

- Context-guided question attention
  - Softmax across rows





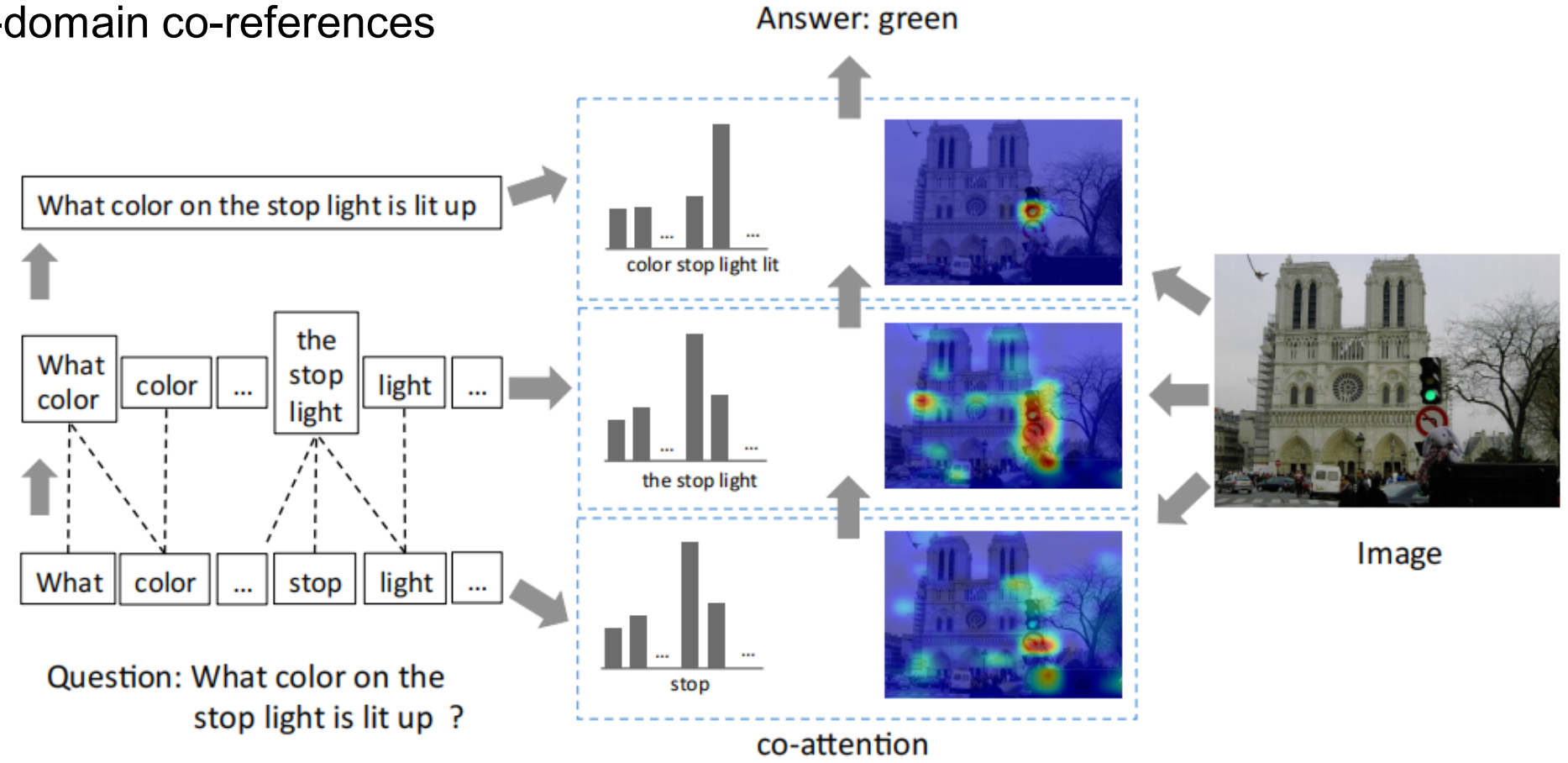
→ Q: Probably not working for image qa where single words does not have the co-reference with a region?

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



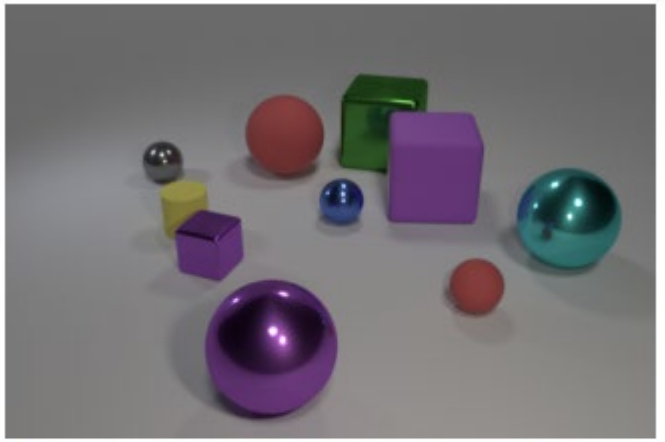
→ Q: Can this be done on text qa as well?

→ 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?

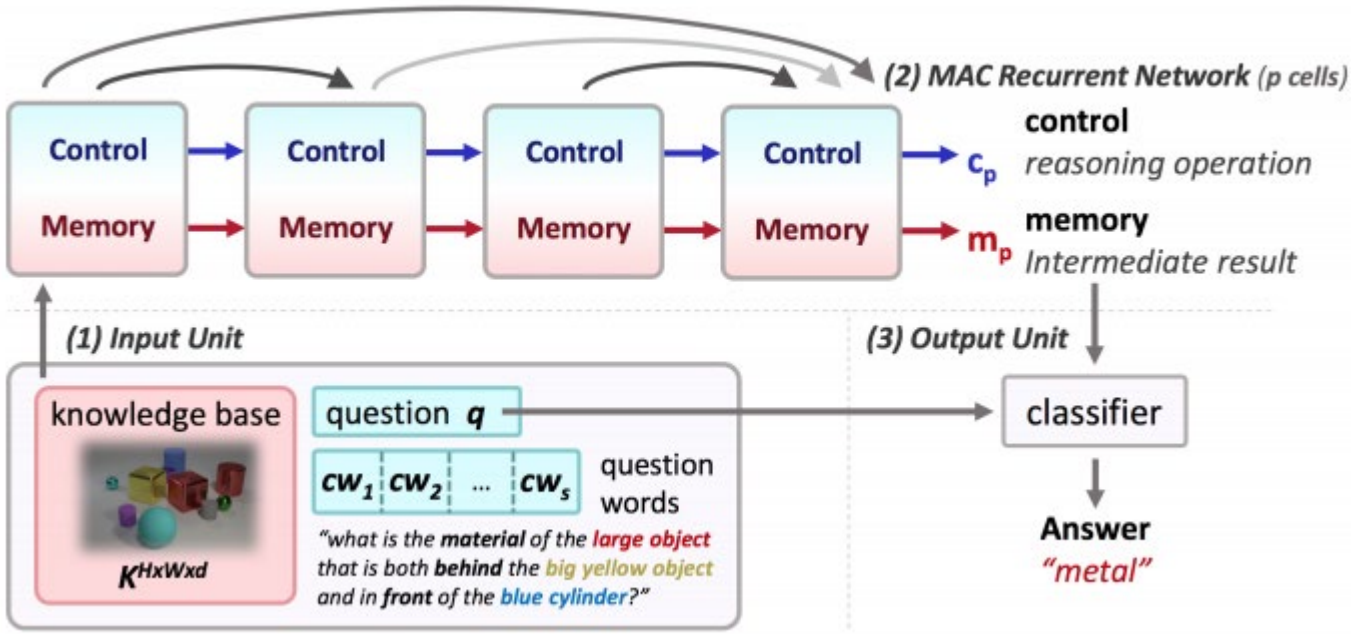


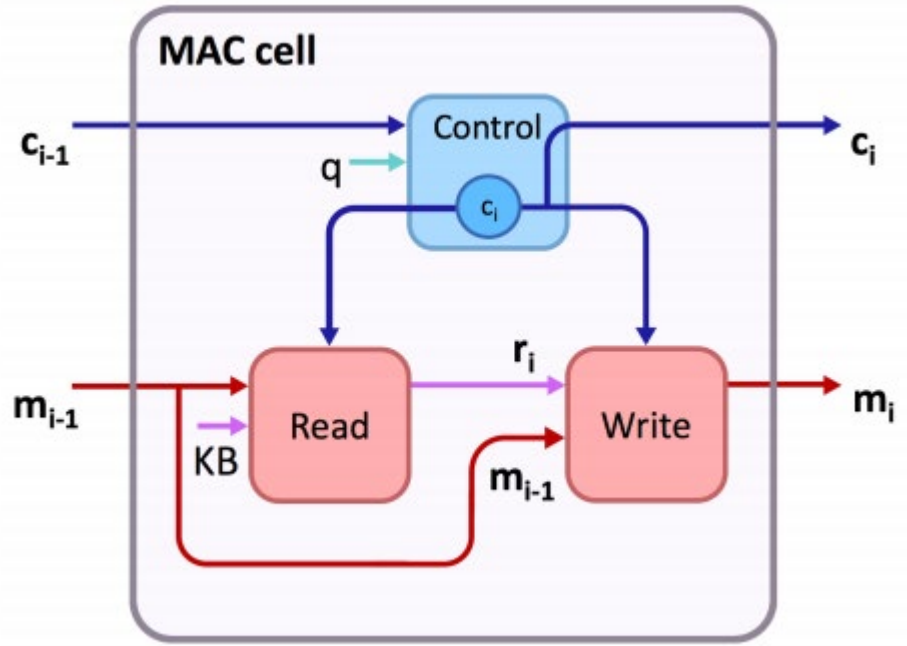
Q: *Do the block in* ***front*** *of the tiny yellow cylinder and the tiny thing that is to the* ***right*** *of the large green shiny object have the* ***same color****?* **A:** *No*

Figure: Hudson and Manning – ICLR 2018

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

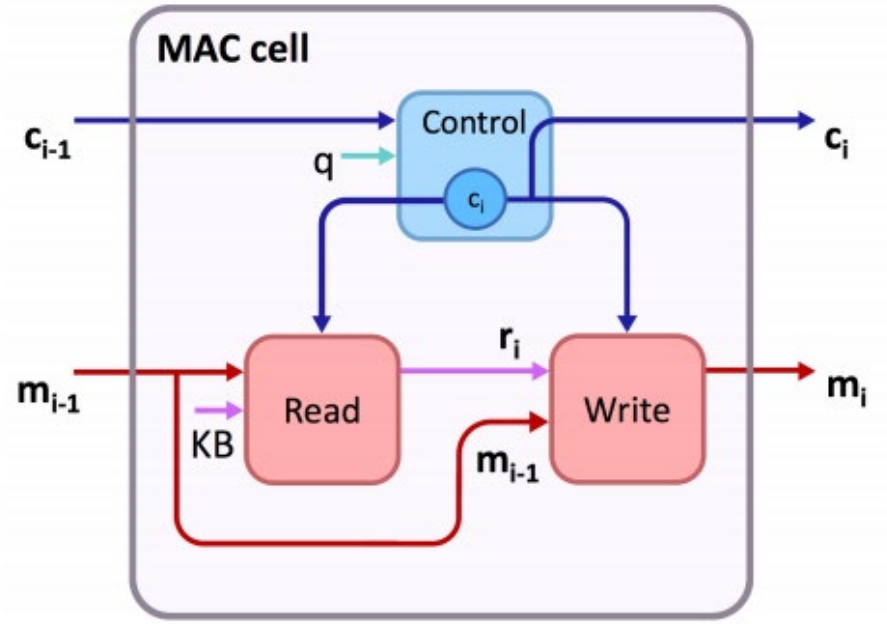
- Attention reasoning is done through multiple sequential steps.
- Each step is done with a recurrent neural cell
- *What is the key differences to the normal RNN (LSTM/GRU) cell?*
  - *Not a sequential input, it is sequential processing on static input set.*
  - *Guided by the question through a controller.*





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



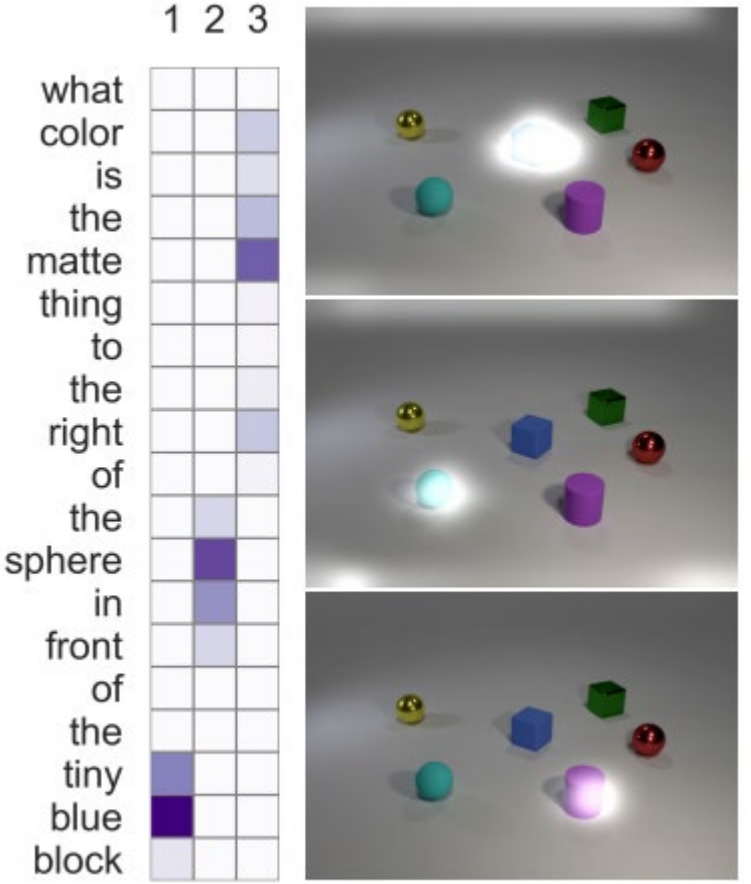
MAC network, Hudson and Manning – ICLR 2018

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)

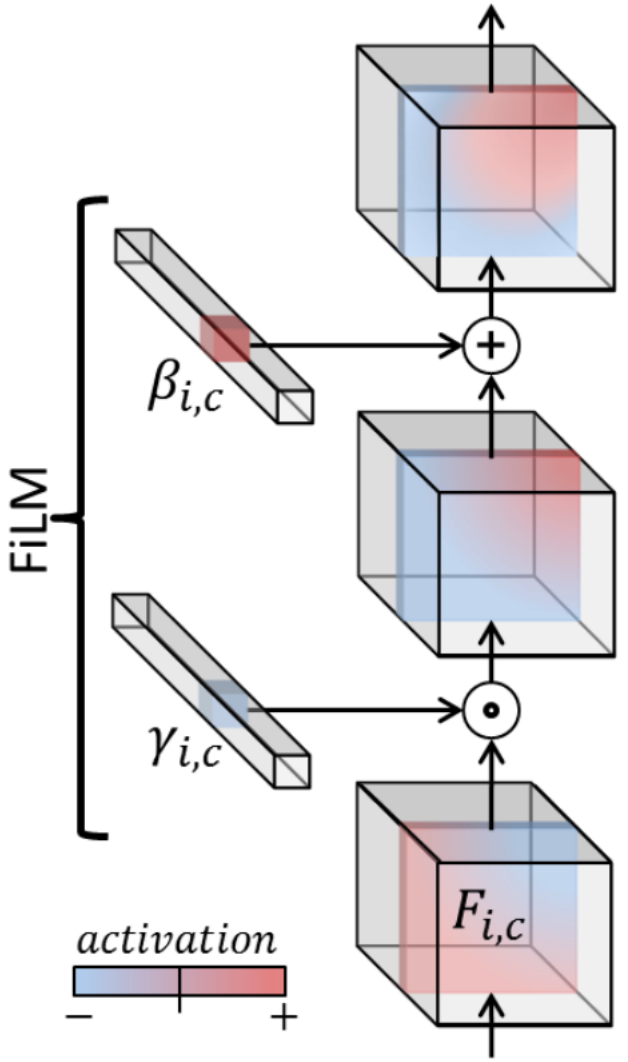
- Influence of input x to network features

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

- The modulation is done with an affine transform

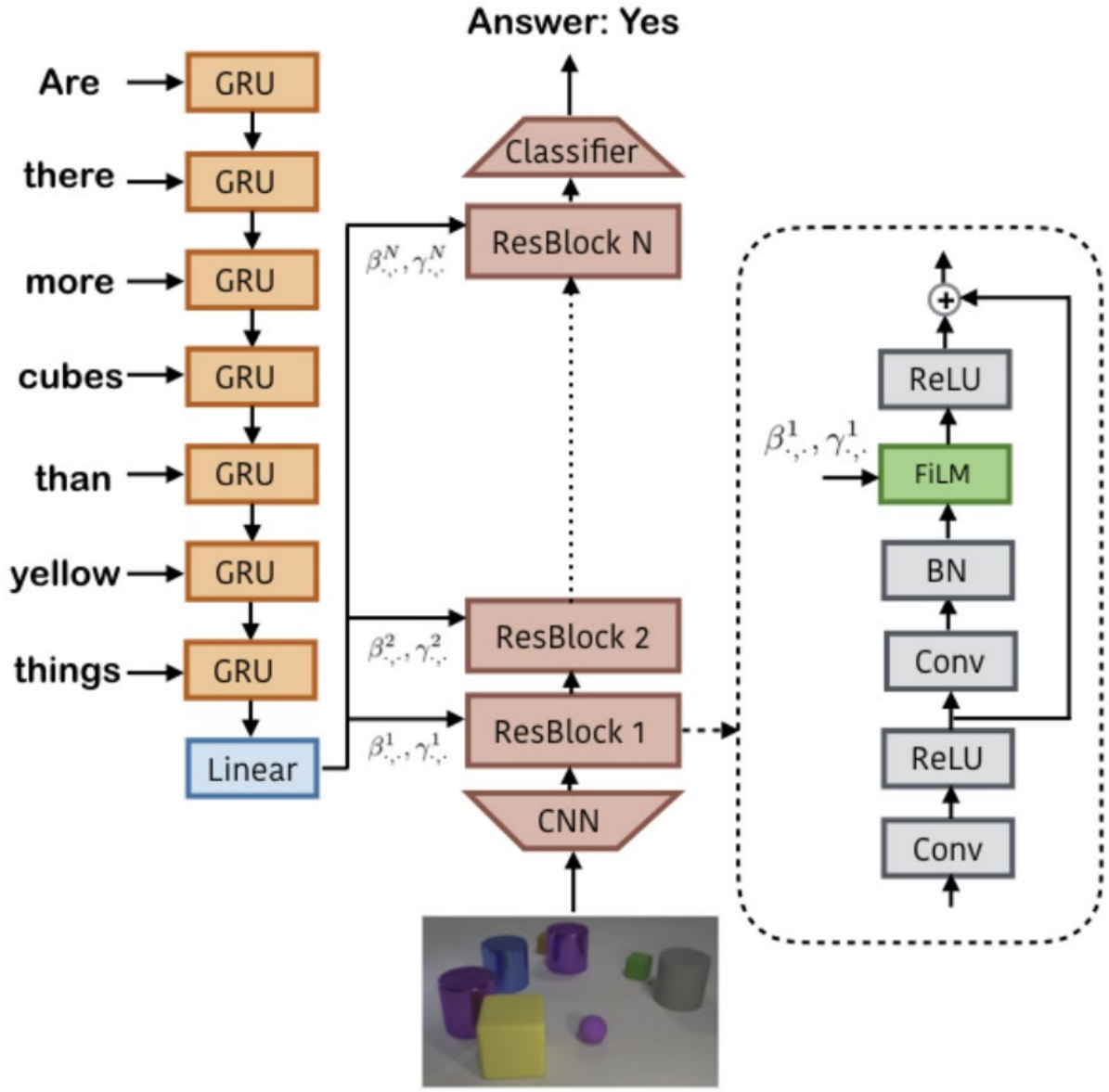
$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}$$

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

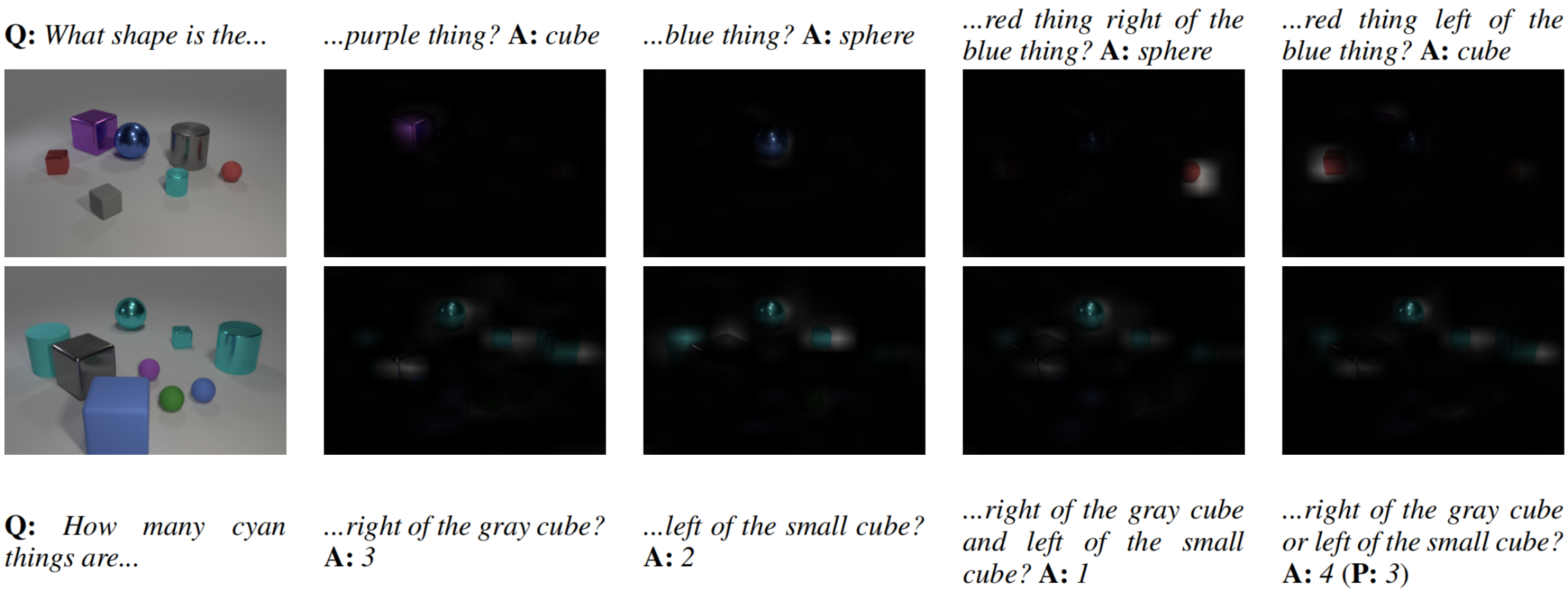


FiLM for question answering

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



FiLM – visualization of result



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?

→ Q: Set-set interaction is inadequate for questions about *relations between objects*