Compositional Attention Networks For Machine Reasoning

Drew A. Hudson, Christopher D. Manning

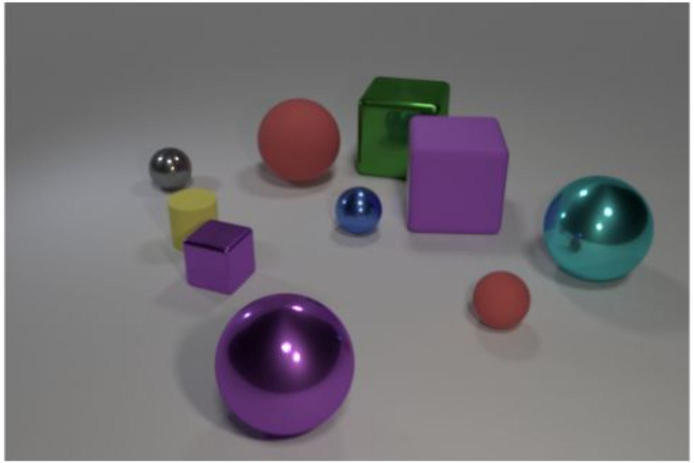
ICLR, 2018

Deniz Beser (dbeser@seas.upenn.edu)
03.27.2019

Penn Engineering

Problem & Motivation (1)

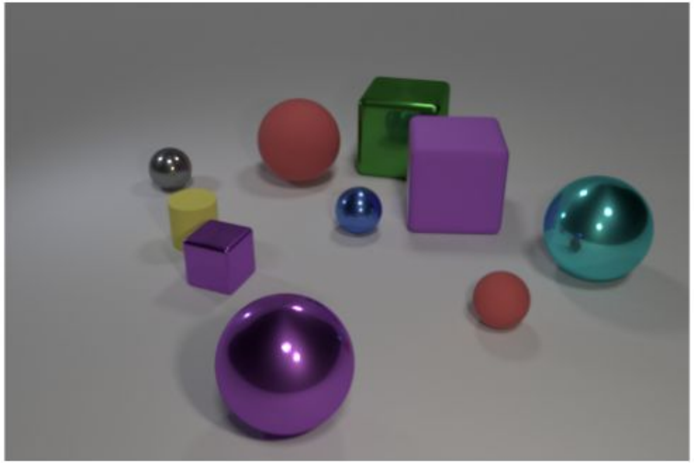
- Current neural networks are good at pattern recognition but not reasoning.
- E.g.: The networks can detect the spheres, but not operate on relations between spheres and cubes



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*

Problem & Motivation (2)

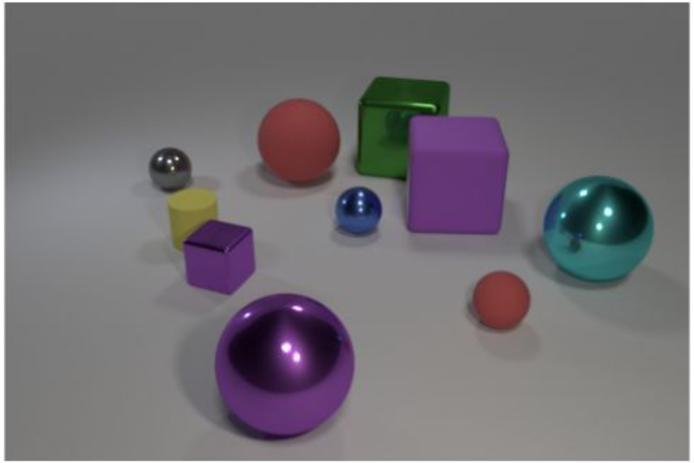
- Transparency and interpretability
- As opposed to elusive black-box architectures
- E.g.: What about the network's behavior explains the answer "No"?



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*

Problem & Motivation (3)

- NNs are generally tabula rasa (i.e. clean slate, minimal priors, very versatile), which is often compensated with big-data.
- E.g.: How can we teach what a sphere is by using fewer images?
- Goal: faster learning



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*

Penn Engineering

Motivation

- A compositional model
  - that can reason about relations with its inductive biases
  - that is interpretable/transparent
  - that learns quickly (requires less data)
  - that achieves better results

Contents:

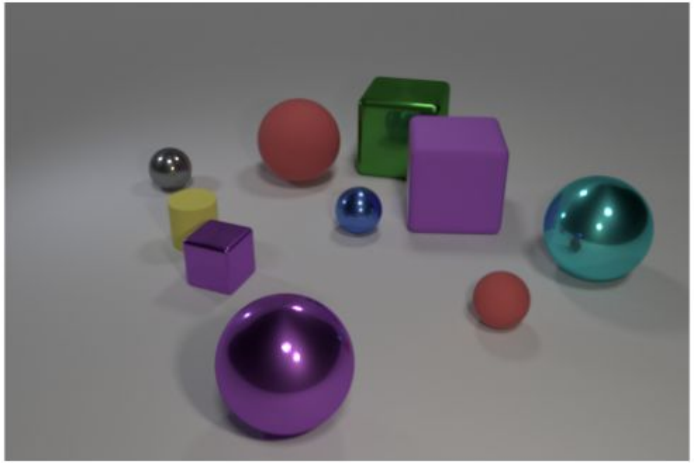
- Task & Evaluation
- Previous approaches
- Contributions
- Model Details
- Results
- Analysis
  - Learning efficiency
  - Interpretability
- Conclusions & Future Work

Task & Evaluation

- CLEVR Visual Question Answering Dataset
- Natural language questions about images.

  E.g. “ Are any objects gold? A: yes”

  E.g. “What color ball is close to the small purple cylinder? A: gray”
- 700k examples w/ 28 possible answers



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*

Penn Engineering

Previous approaches

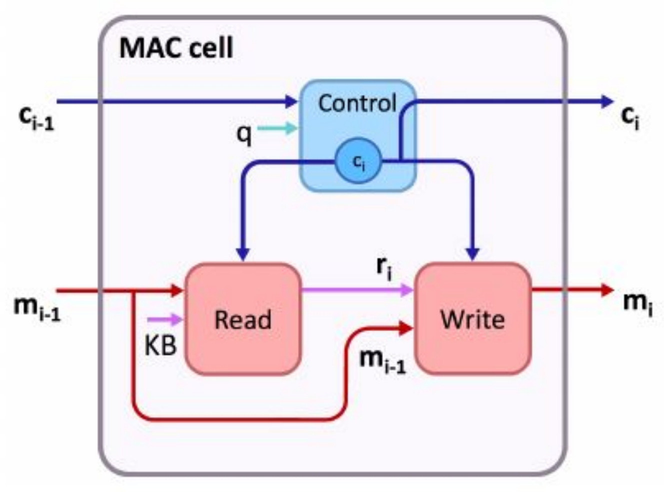
- Generally a mix of CNN+LSTM
- Modular approaches such as Module Networks (Andreas et al., 2016)

    - **not fully differentiable!**

- **Can't count:** Counting and aggregation skills tend challenging for previous models (Santoro et al., 2017; Hu et al., 2017; Johnson et al., 2017b)

Contributions

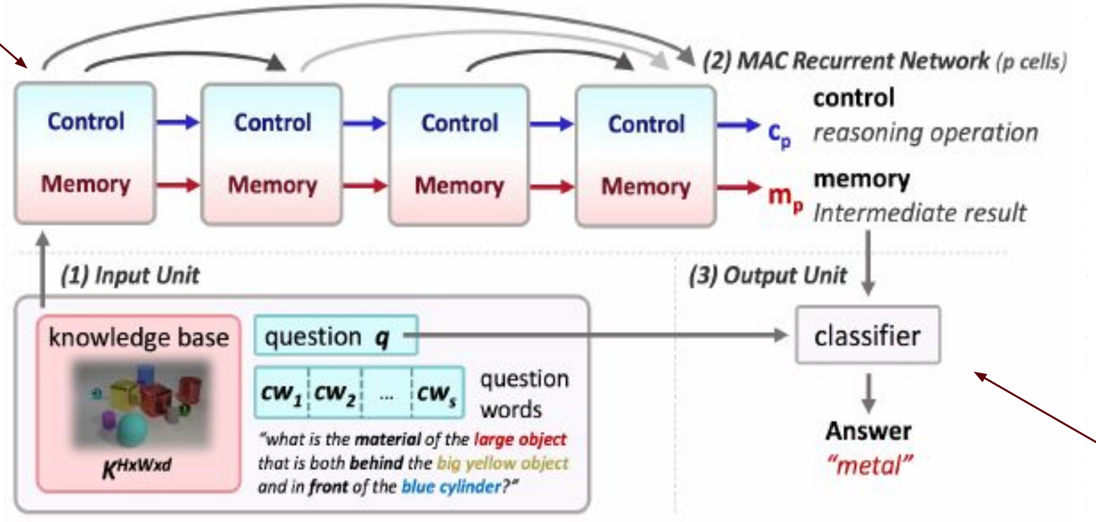
- Memory, Attention, and Composition (MAC) cell
- A cell deliberately designed to capture an elementary, yet general-purpose reasoning step, inspired by computers
- State-of-the-art accuracy and efficiency on CLEVR
- Discussion of significance of inductive biases



Model - Network Overview

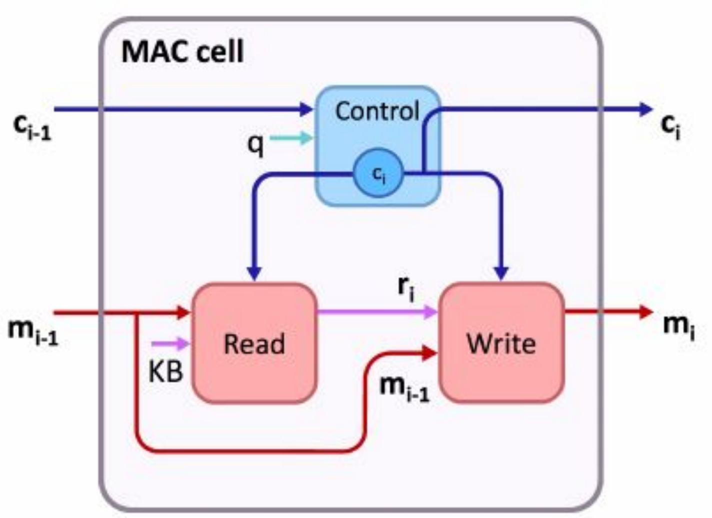
Similar to RNNs

1) Input unit
2) Recurrent cells
3) Output unit



Model - The MAC Cell Architecture

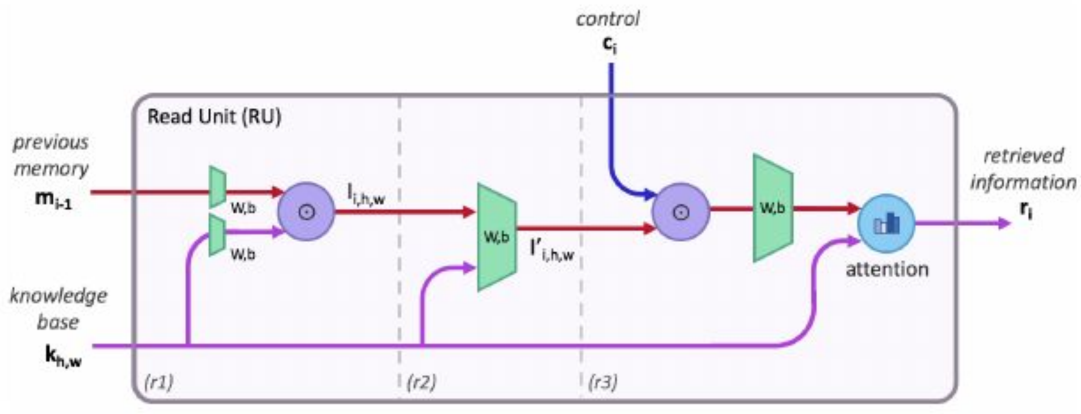
- Control, Read, and Write units
  - Control and memory hidden states.
- The control unit successively attends to different parts of the task (i.e. question)
- The read unit extracts information from knowledge base (i.e. image), guided by control.
- The write unit integrates the retrieved information into the memory state, yielding the new intermediate result (reasoning)
- Inspired by computer architecture!



Penn Engineering

Model - The Read Unit

- Retrieves information using KB and memory
- Uses this information and the control state to generate attention over KB



$$I_{i,h,w} = [W_m^{d\times d} \boldsymbol{m_{i-1}} + b_m^d] \odot [W_k^{d\times d} \boldsymbol{k}_{h,w} + b_k^d] \quad \text{(r1)}$$

$$I'_{i,h,w} = W^{d\times 2d}\left[I_{i,h,w}, \boldsymbol{k}_{h,w}\right] + b^d \quad \text{(r2)}$$

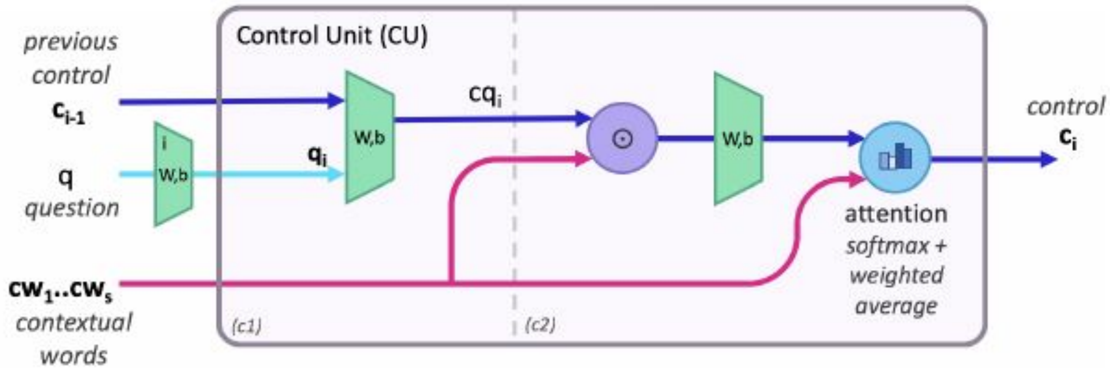
$$ra_{i,h,w} = W^{d\times d}(\boldsymbol{c_i} \odot I'_{i,h,w}) + b^d \quad \text{(r3.1)}$$

$$rv_{i,h,w} = \text{softmax}(ra_{i,h,w}) \quad \text{(r3.2)}$$

$$\boldsymbol{r_i} = \sum_{h,w=1,1}^{H,W} rv_{i,h,w} \cdot \boldsymbol{k}_{h,w} \quad \text{(r3.3)}$$

Model - The Control Unit

- Uses previous control state and the task question to apply attention over question words to generate new control state



$$cq_i = W^{d\times 2d}[\boldsymbol{c_{i-1}}, \boldsymbol{q_i}] + b^d \qquad \text{(c1)}$$

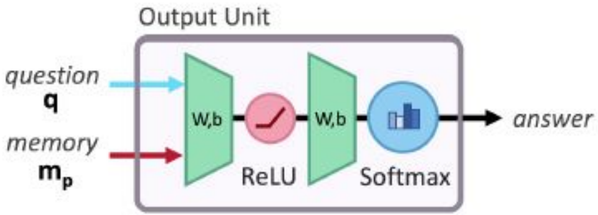
$$ca_{i,s} = W^{1\times d}(cq_i \odot \boldsymbol{cw_s}) + b^1 \qquad \text{(c2.1)}$$

$$cv_{i,s} = \text{softmax}(ca_{i,s}) \qquad \text{(c2.2)}$$

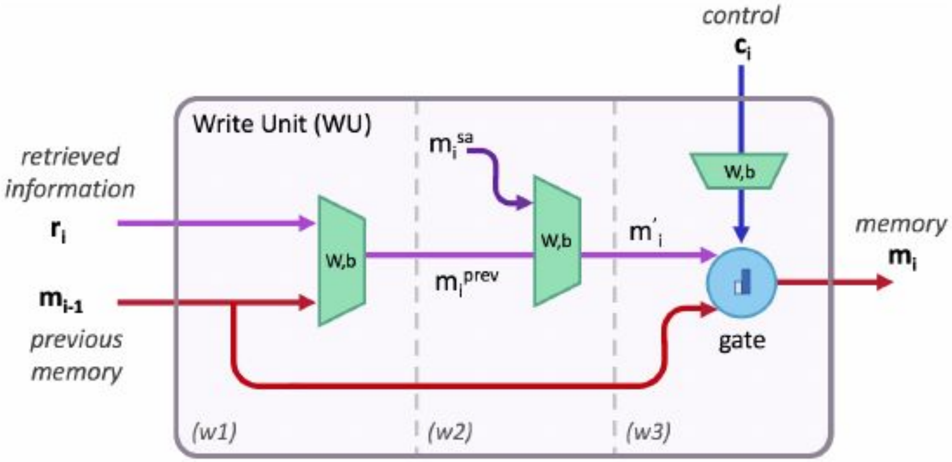
$$\boldsymbol{c_i} = \sum_{s=1}^{S} cv_{i,s} \cdot \boldsymbol{cw}_s \qquad \text{(c2.3)}$$

Penn Engineering

Model - The Write Unit



- Integrates the retrieved information into the memory state
- Control decides the gating (i.e. maintaining previous memory)



$$m_i^{info} = W^{d\times 2d}[\boldsymbol{r_i}, \boldsymbol{m_{i-1}}] + b^d \qquad \text{(w1)}$$

$$sa_{ij} = \text{softmax}\left(W^{1\times d}(\boldsymbol{c_i} \odot \boldsymbol{c_j}) + b^1\right) \qquad \text{(w2.1)}$$

$$m_i^{sa} = \sum_{j=1}^{i-1} sa_{ij} \cdot \boldsymbol{m_j} \qquad \text{(w2.2)}$$

$$m'_i = W_s^{d\times d} m_i^{sa} + W_p^{d\times d} m_i^{info} + b^d \qquad \text{(w2.3)}$$

$$c'_i = W^{1\times d}\boldsymbol{c_i} + b^1 \qquad \text{(w3.1)}$$

$$\boldsymbol{m_i} = \sigma\left(c'_i\right)\boldsymbol{m_{i-1}} + \left(1 - \sigma\left(c'_i\right)\right) m'_i \qquad \text{(w3.2)}$$

Penn Engineering

Results

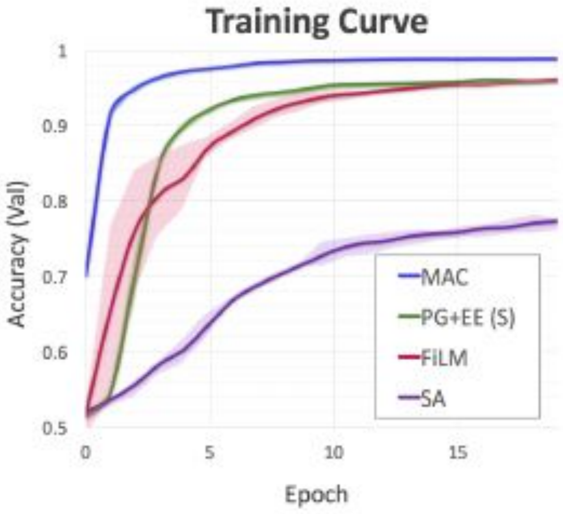
- SotA in all categories in CLEVR
  - Overall, count, exist, compare attribute, compare numbers...
- Nonetheless, many models perform well on CLEVR
  - High 90s in many categories

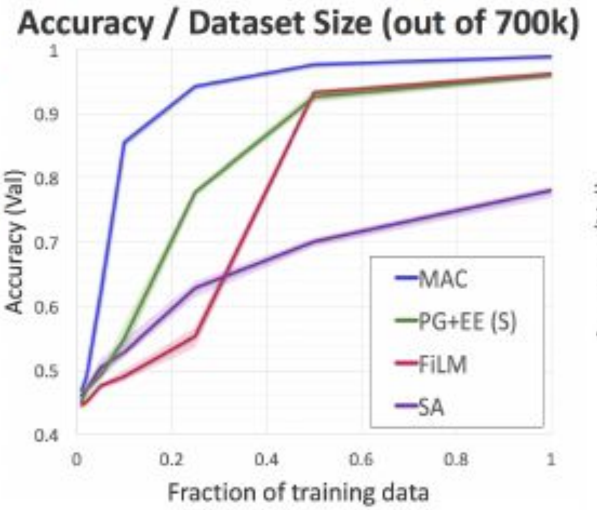
Model	Overall	Count	Exist	Compare Attribute
Previous SotA	97.7	94.3	99.3	99.3
MAC	**98.9**	**97.1**	**99.5**	**99.5**

Penn Engineering

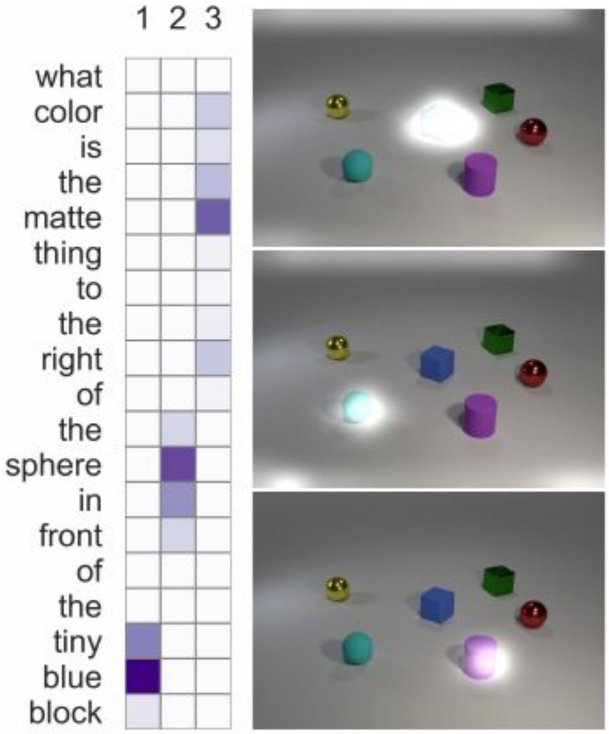
Analysis - Learning Efficiency

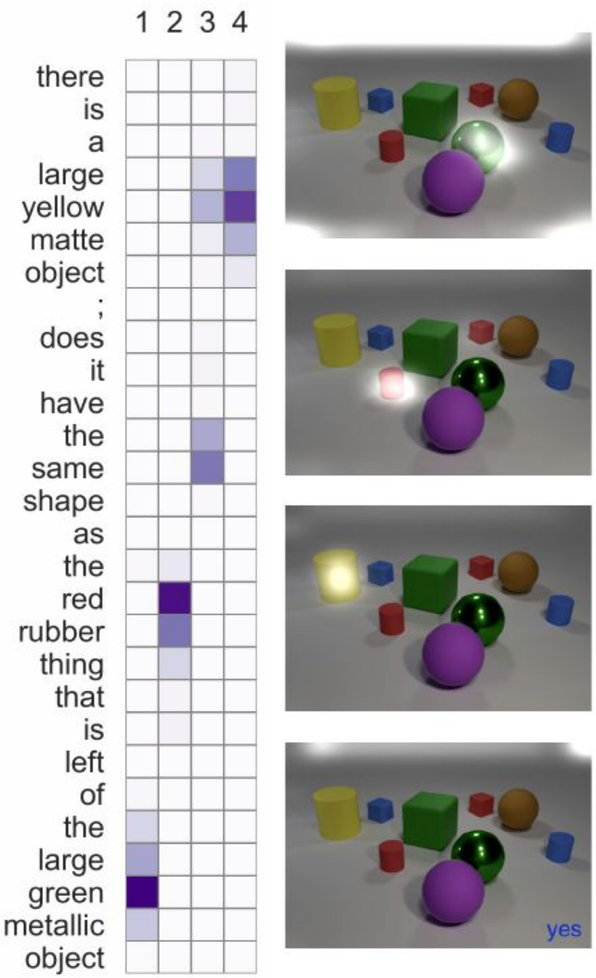
Accuracy as a function of training data:





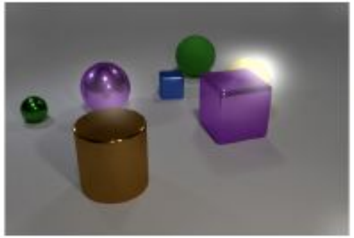
Analysis - Interpretability



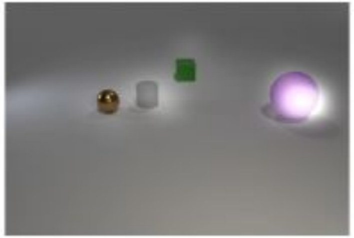


Analysis - Interpretability

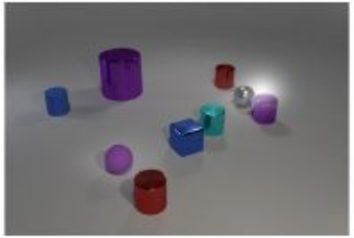
- Model performs well on questions collected through crowdsourcing (that are not in original CLEVR dataset)



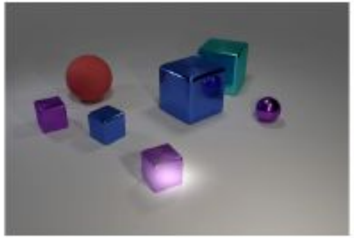
Q: *What is the shape of the large item,* ***mostly occluded*** *by the metallic cube?* **A:** *sphere* ✓



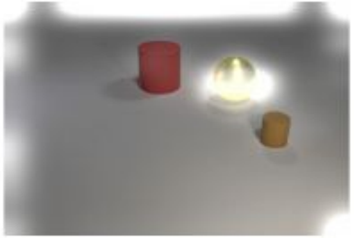
Q: *What color is the object that is a* ***different*** *size?* **A:** *purple* ✓



Q: *What color ball is* ***close to*** *the small purple cylinder?* **A:** *gray* ✓



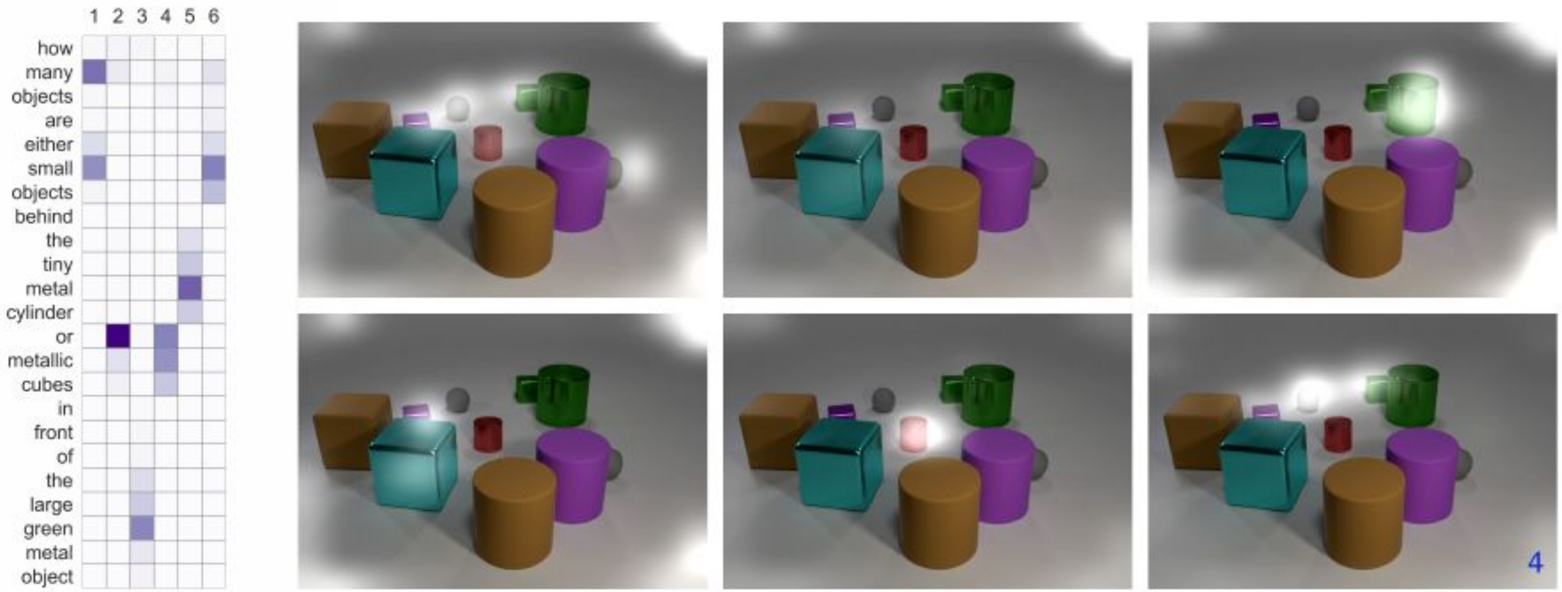
Q: *What color block is* ***farthest front****?* **A:** *purple* ✓



Q: *Are any objects* ***gold****?* **A**: *yes* ✓

Analysis - Interpretability

- **Counting:** On which words does the model focus first?



Conclusions and Future Work

- State-of-the-art accuracy and efficiency for VQA
- **Inductive biases** can
  - improve accuracy (halved error on CLEVR)
  - increase efficiency - more significant given other models' accuracy

- Apply this model to different tasks
  - Other evaluations needed - many models perform well
- Further analysis of the distinguishing qualities of this model
  - How useful is the computer-architecture?

- Code: https://github.com/stanfordnlp/mac-network