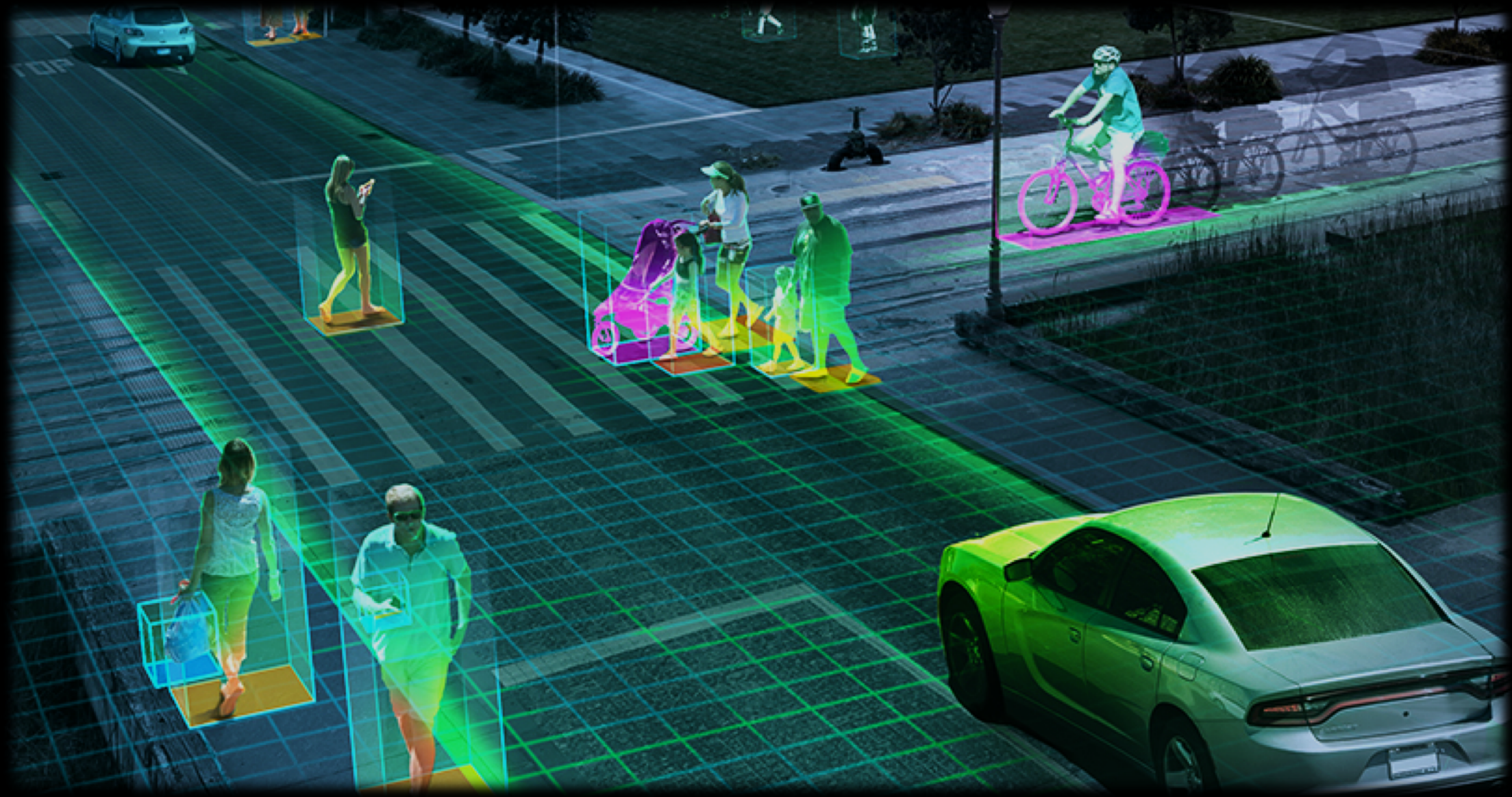
Compositional Attention Networks for Machine Reasoning

Christopher Manning & Drew Hudson

Stanford University

@chrmanning • @stanfordnlp

ICLR 2018

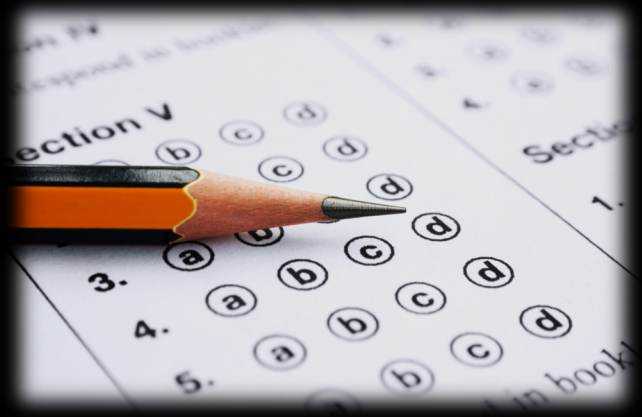



Ich fliege nach
ill fly to Canada
Tengo se
am thirsty

Machine Learning

Current NN/ML systems excel on intuitive learning tasks

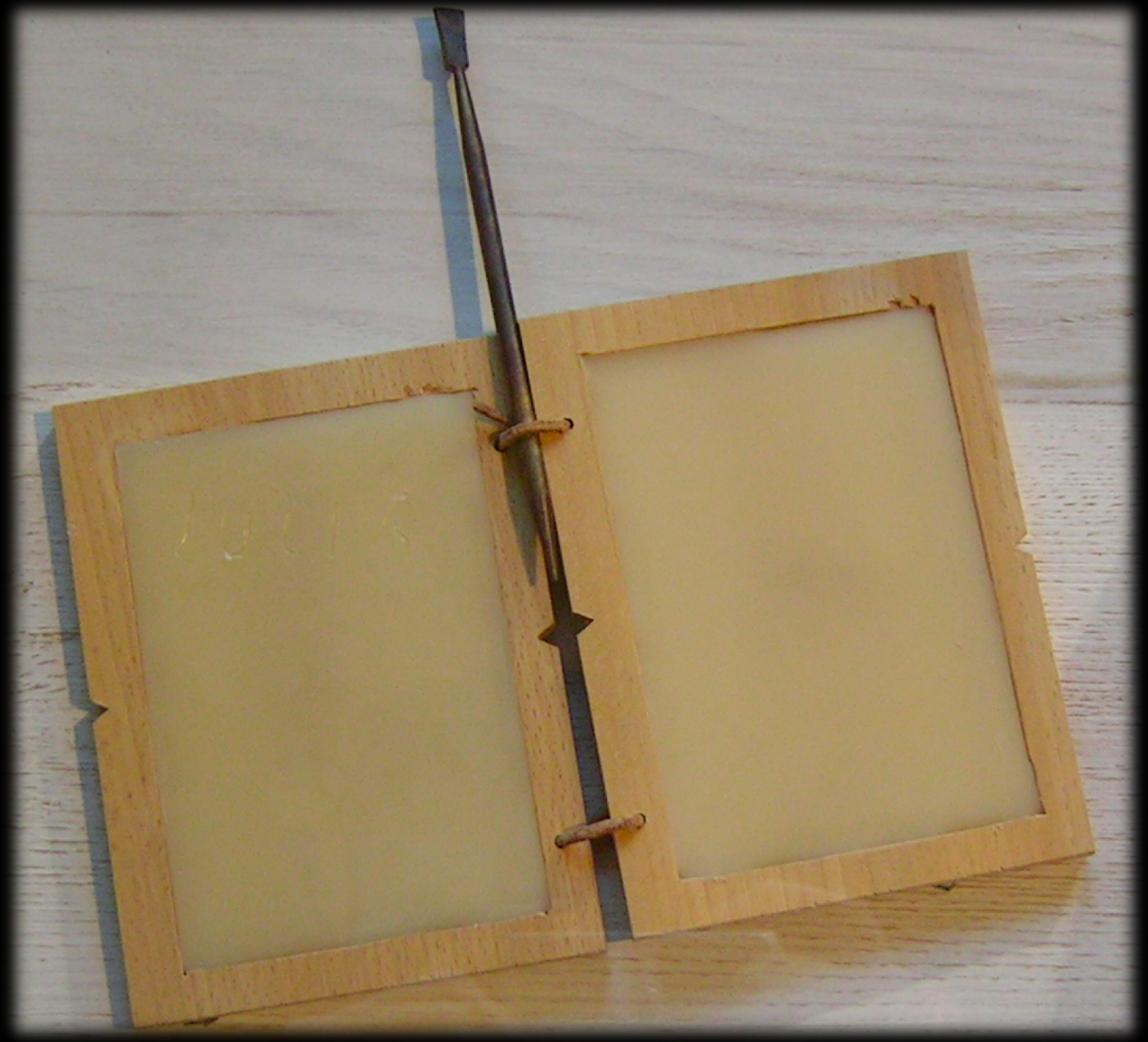
But what about reasoning?



What is Reasoning? [Bottou 2011]

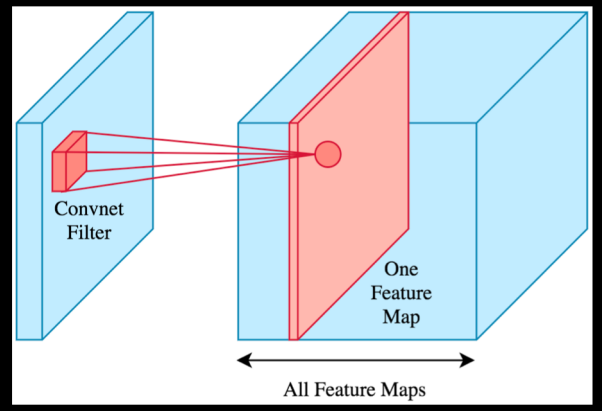
- **Algebraically manipulating** previously acquired **knowledge** in order to **answer a new question**
- **Is not necessarily achieved** by making **logical inferences**
- **Continuity** between **algebraically rich inference** and **connecting together trainable learning** systems
- Central to **reasoning** **is** **composition rules** to guide the combinations of modules to address new tasks

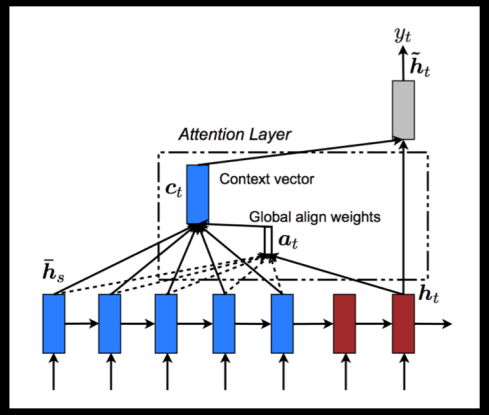
Worshipping the tabula rasa

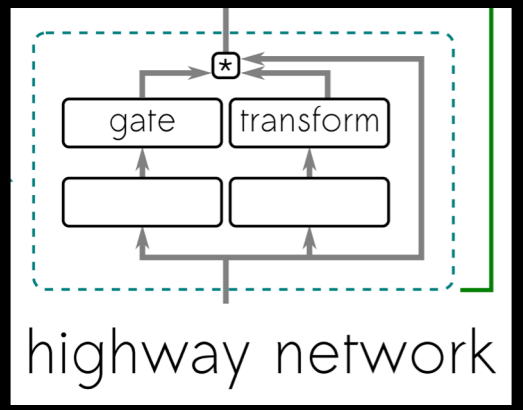


A good inductive bias improves your ability to learn (quickly and well)

Appropriate structural priors





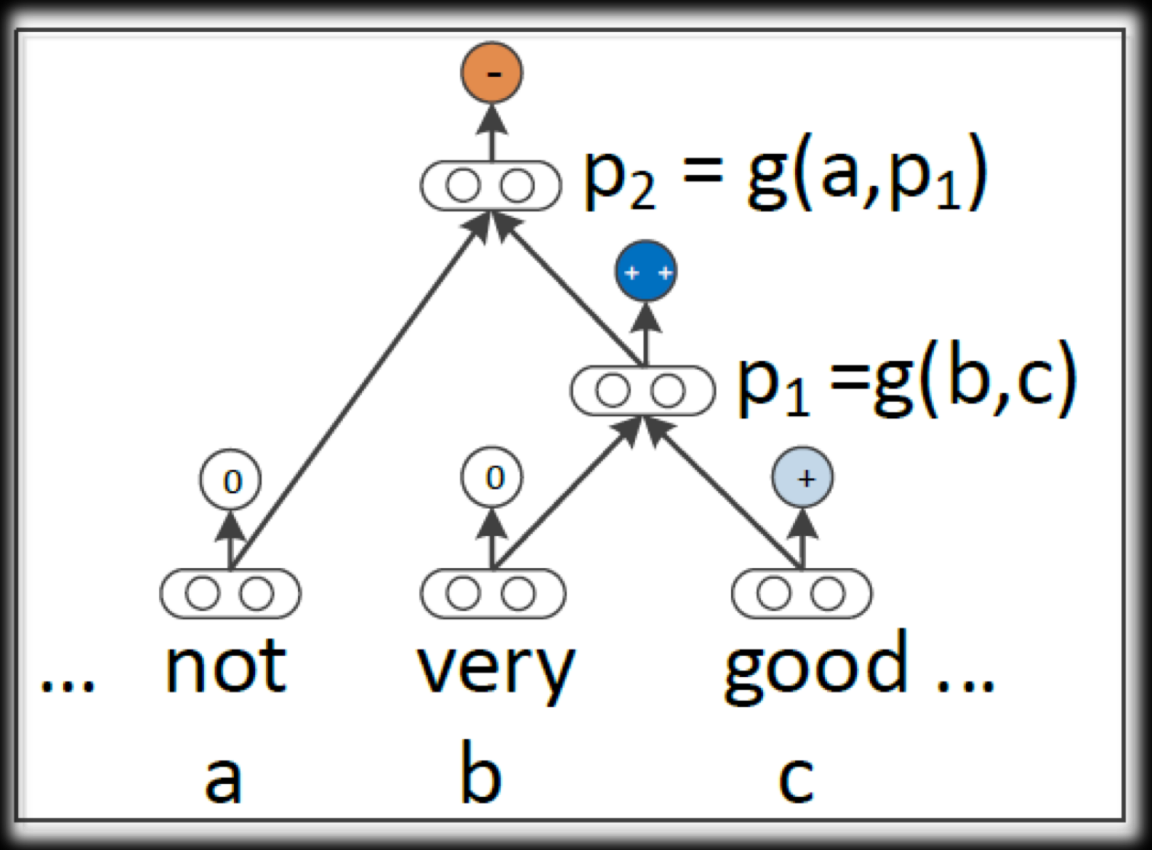


Convolution

Attention

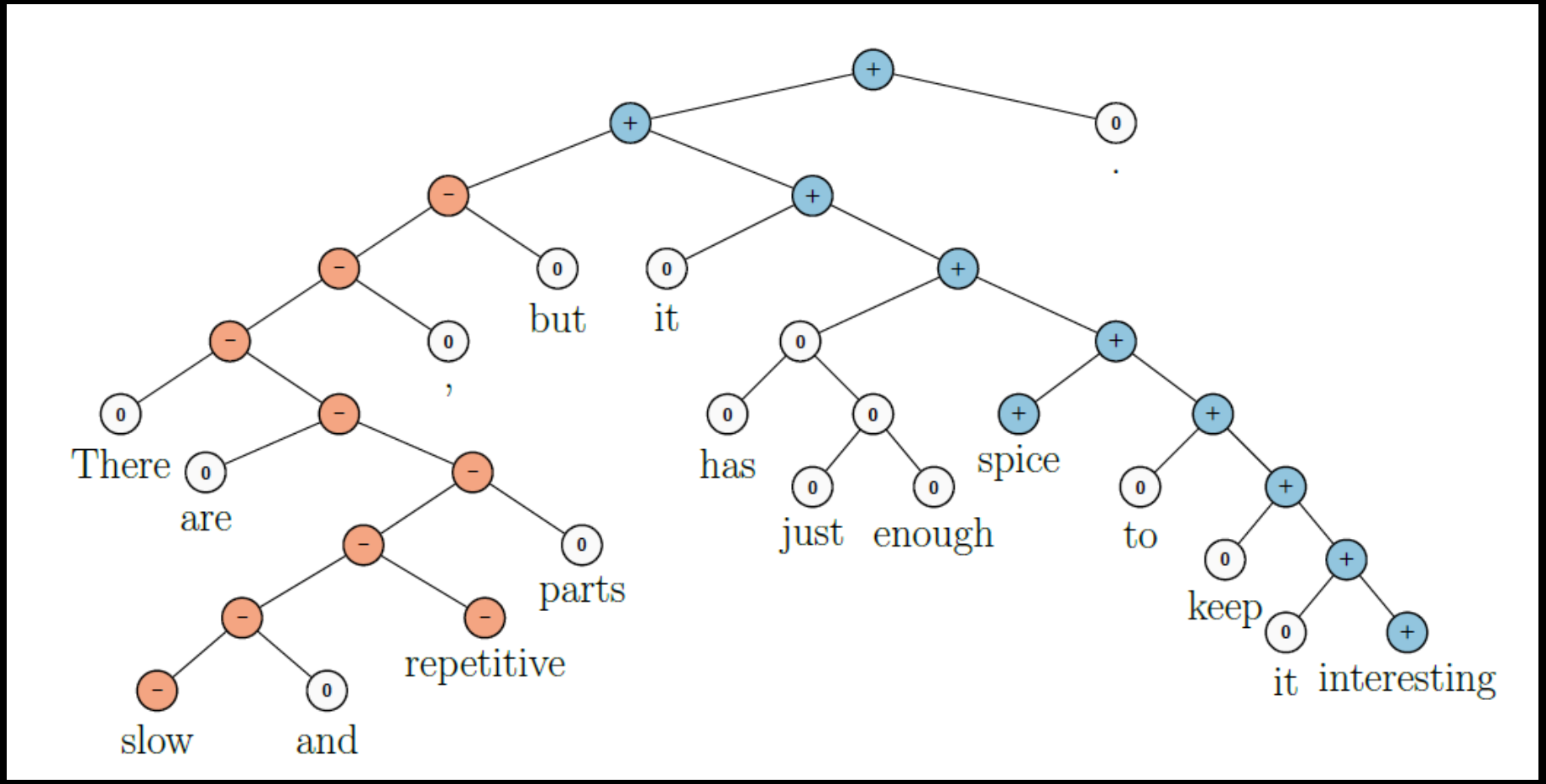
Gating (forget/highway)

Tree-structured models



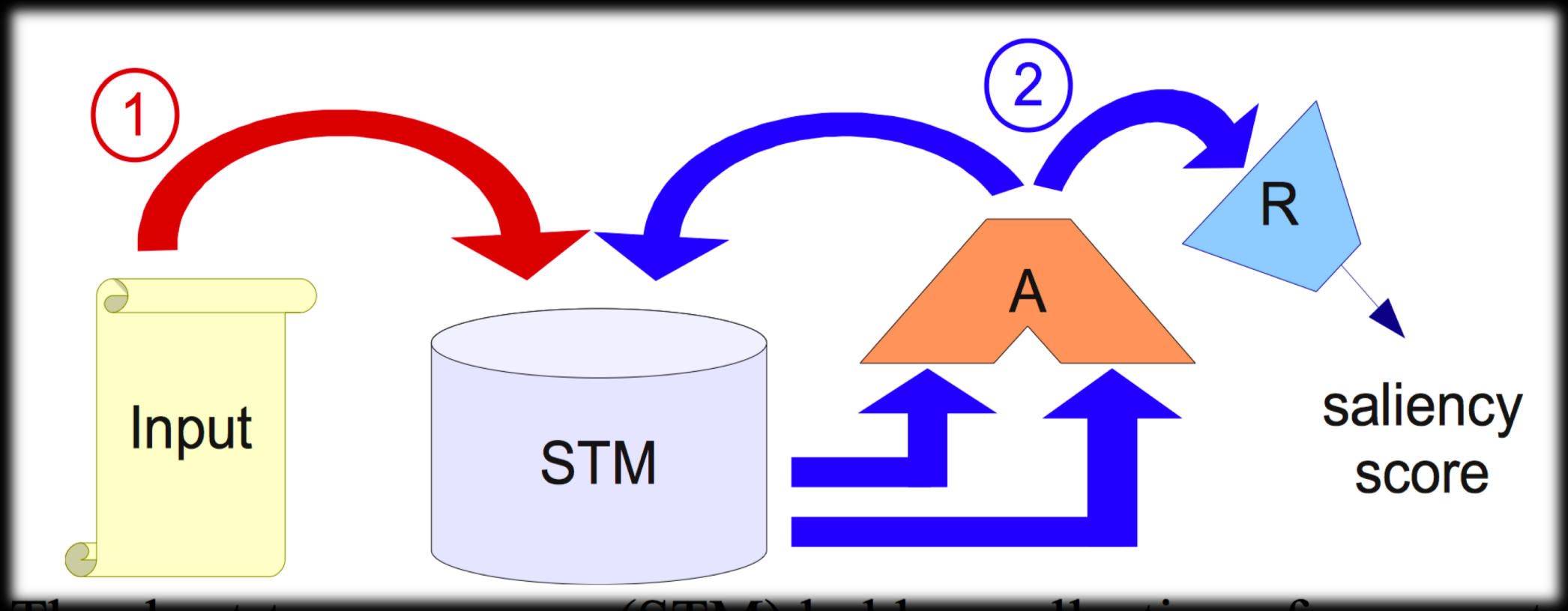
[Socher et al. 2010ff]
[Tai et al. 2015]

Tree-structured models

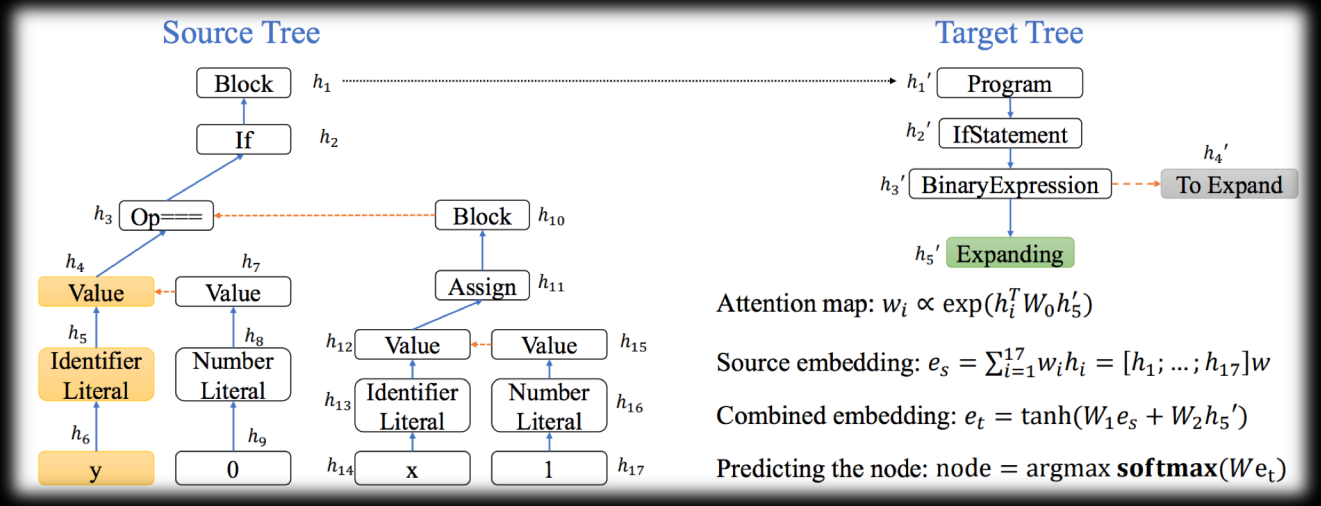


[Socher et al. 2010ff, Tai et al. 2015]

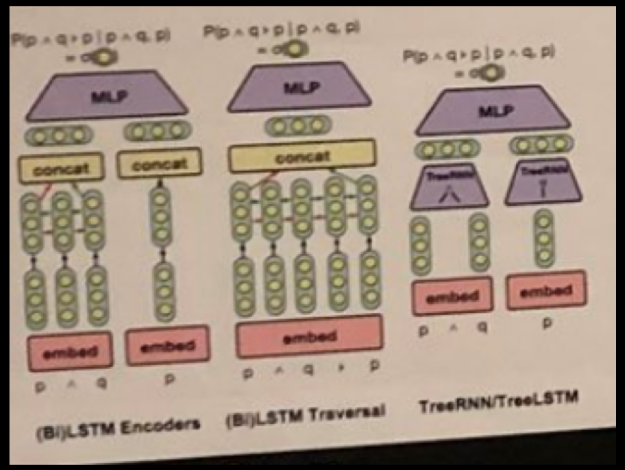
Compositional reasoning tree



Tree-structured models

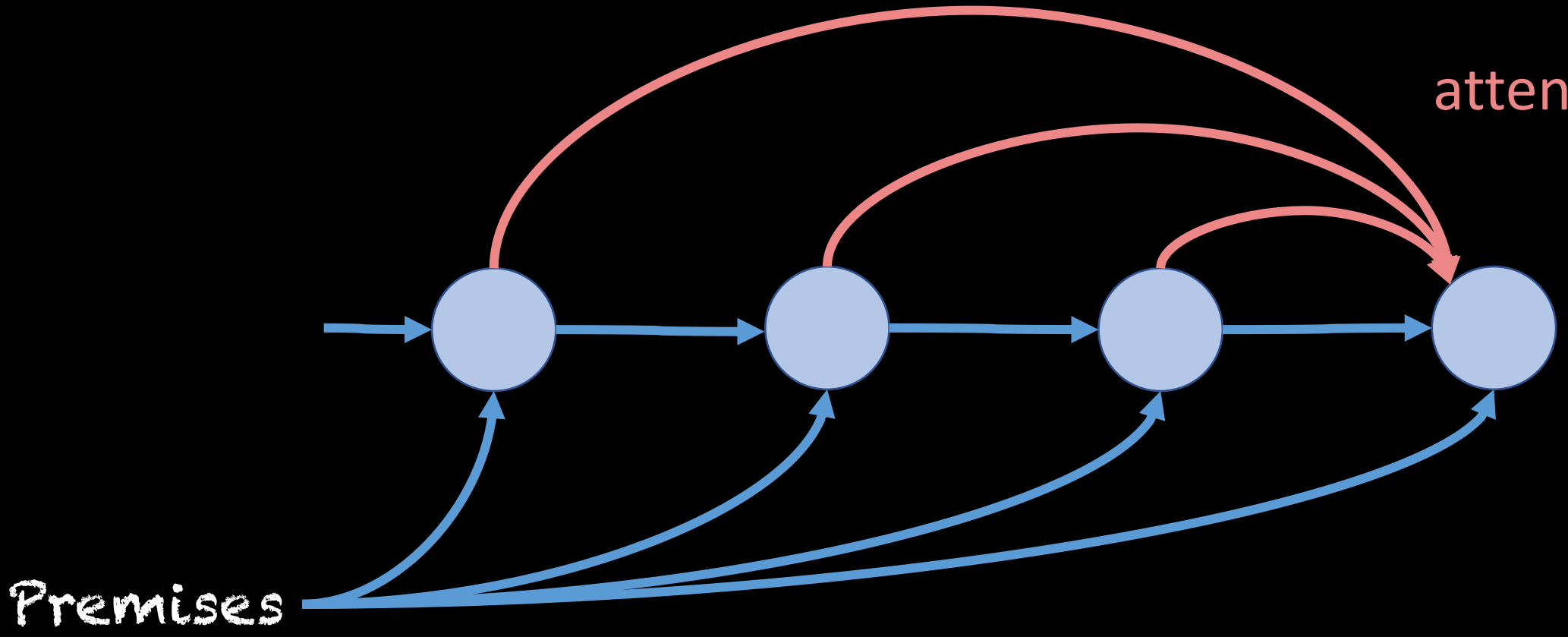


Program Translation
Xinyun Chen, Chang Liu & Dawn Song
ICLR 2018



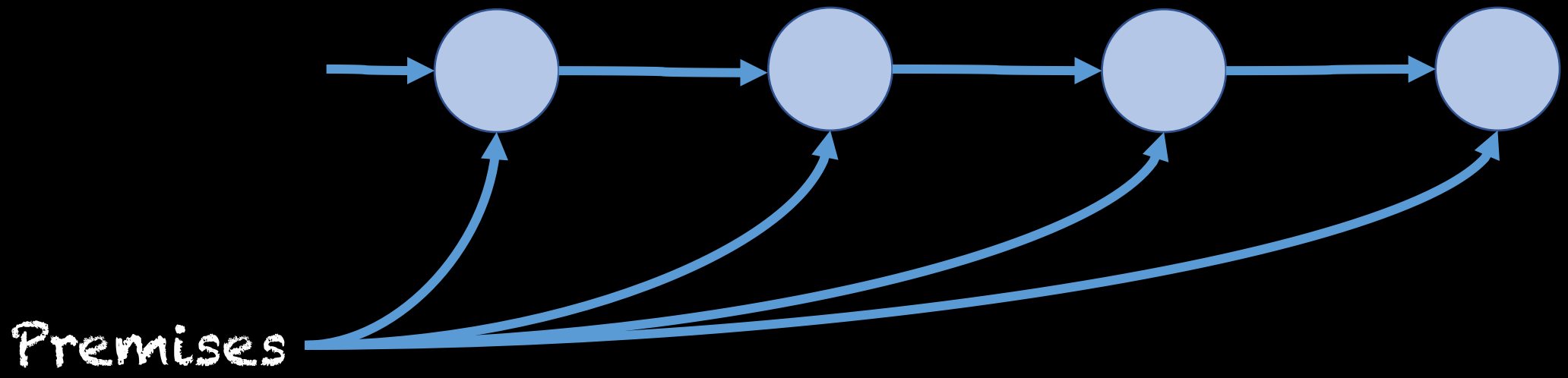
Can Neural Networks Understand Logical Entailment?
Evans, Saxton, Amos, Kohli & Grefenstette
ICLR 2018

Compositional reasoning without trees



Compositional reasoning without trees

If *f*: (X × Y × Z) → N, then curry(*f*): X → (Y → (Z → N))



Our Goal

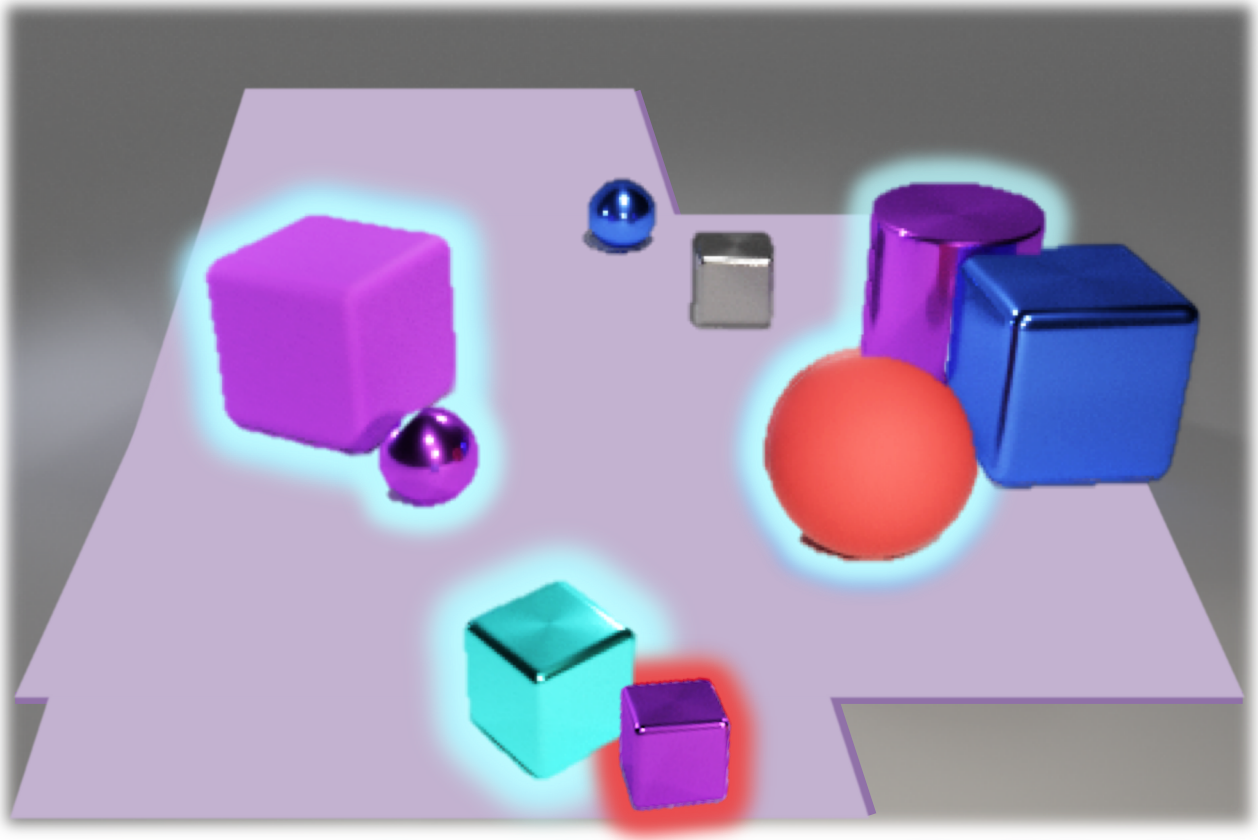
Rather than using standard machine learning correlation engines, the goal is a new neural network design

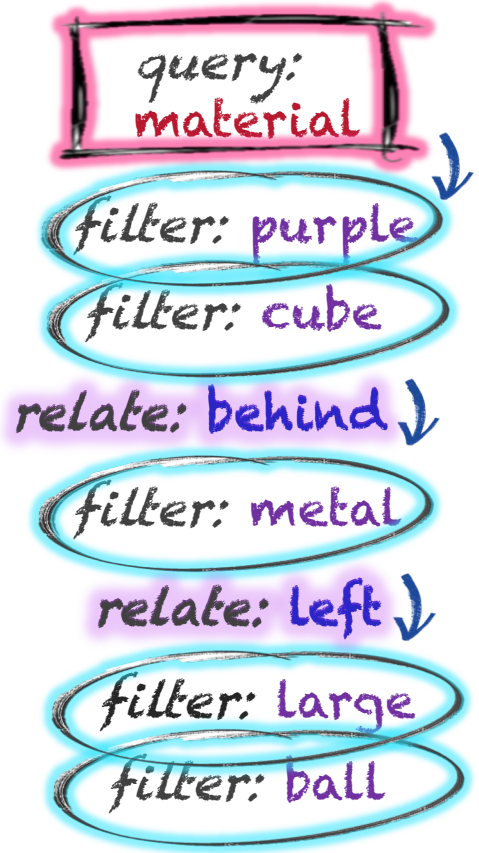
- With a structural prior encouraging **compositional and transparent multi-step reasoning**
- While retaining **end-to-end differentiability** and **scalability to real-world problems**

Talk Outline

- ✓ From Machine Learning to Machine Reasoning
- ➢ The CLEVR task
- ○ Memory-Attention-Composition Networks (MAC nets)
- ○ Experiments and Discussion

CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

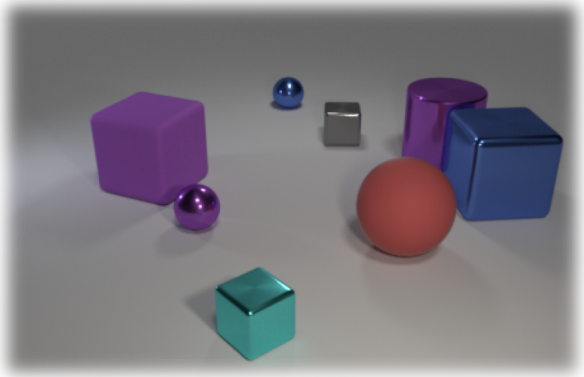




*There is a **purple cube** that is **behind** a **metal** object **left** to a **large ball**; what **material** is the cube?* ***Rubber***

[Johnson et al, CVPR 2017]

CLEVR: The Dataset

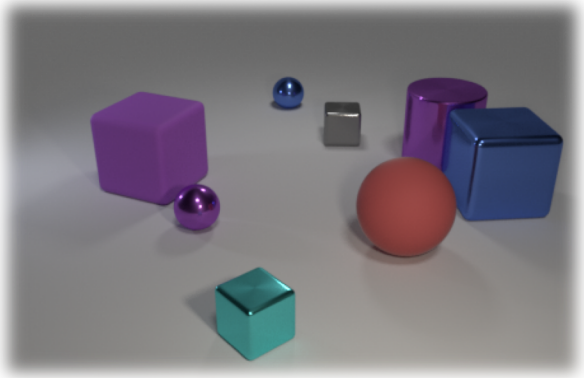


1. **Photorealistic** images of 3D objects in **a few shapes, colors, materials and sizes**
2. Highly **compositional multi-step questions**
3. **Tree-structured functional program** representations provided

- Requires variety of **reasoning skills**, such as **transitive inference**, **counting**, and **comparison** between attributes and amounts.
- **CLEVR-humans** features **natural language questions**

[Johnson et al, CVPR 2017]

CLEVR: Strengths

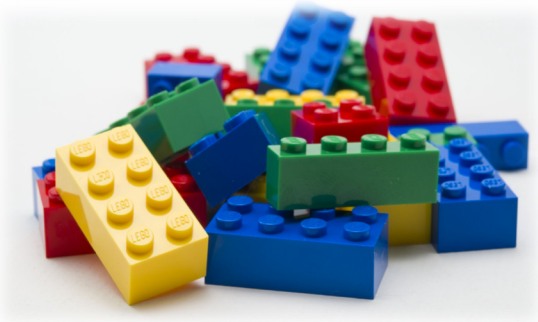


- Allows **thorough analysis** of performance based on the **question**'s functional **structure and type**.
- **Reduces** question-conditional **biases**, compared to standard VQA, thus **reducing the risk** for **spurious** or **superficial reasoning**.

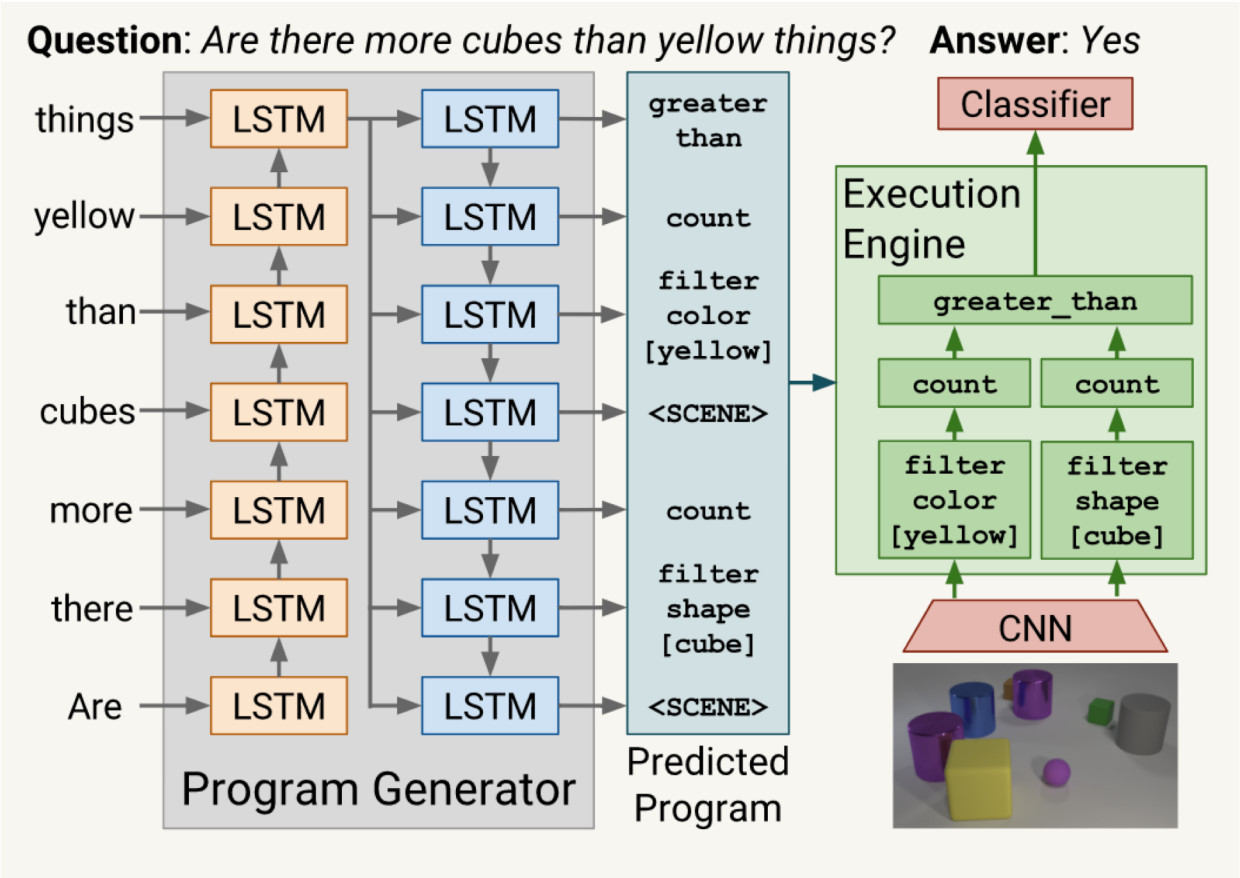
What covers the ground?

[Johnson et al, CVPR 2017]

Existing Approaches
Neural Module Networks



- **Partially differentiable** models that rely on the **strong supervision** to translate queries into a **tree**-structured functional **program**
- The programs are used to compose a corresponding neural network out of a **discrete collection** of **specialized neural modules**

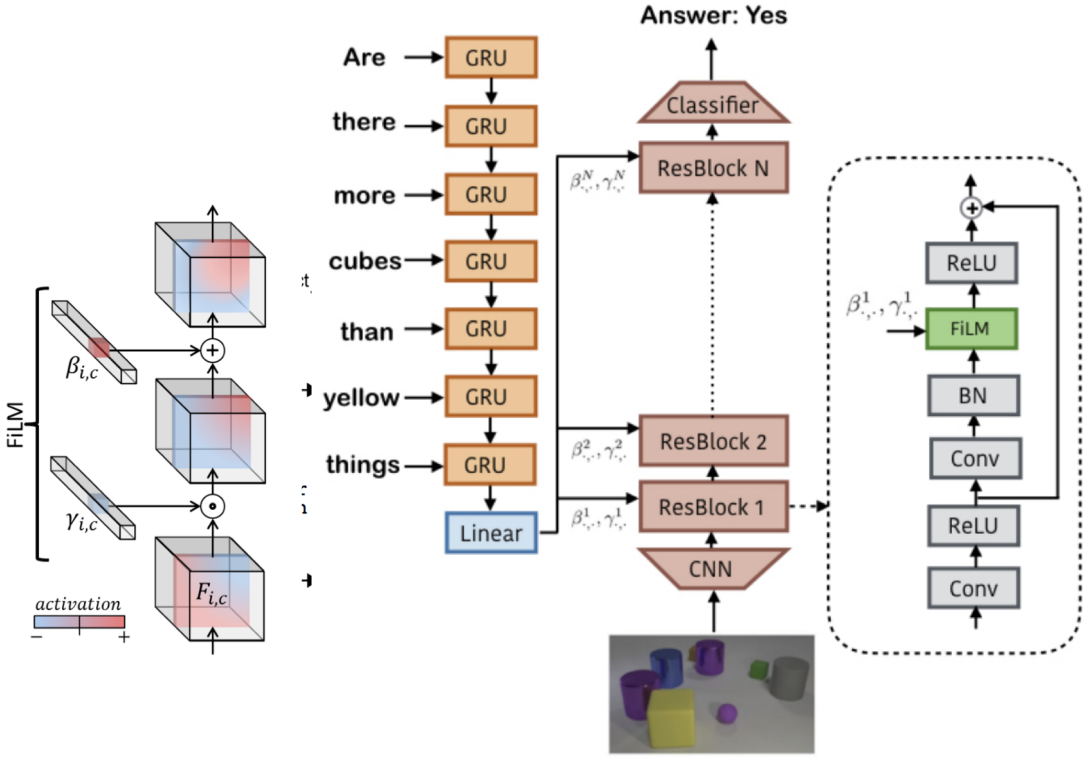


[Andreas at al, CVPR 2016; Johnson et al, ICCV 2017]

Existing Approaches
Relation Nets and FiLM

Large CNN stacks interleaved with specialized layers

- **Relation Net:** Inspects every **pair of pixels** in order to make predictions based on **binary relations**
- **FiLM:** Inserts **conditional linear normalization** layers that **tilt the activations** based on the **question**



RN [Santoro et al, 2017]

FiLM [Perez et al, 2017]

Talk Outline

- ✓ From Machine Learning to Machine Reasoning
- ✓ The CLEVR task
- ➢ Memory-Attention-Composition Networks (MAC nets)
- ○ Experiments and Discussion
- ○ Conclusion

Memory. Attention. Composition. The MAC Network

A **neural model** for **problem solving** and **reasoning** tasks

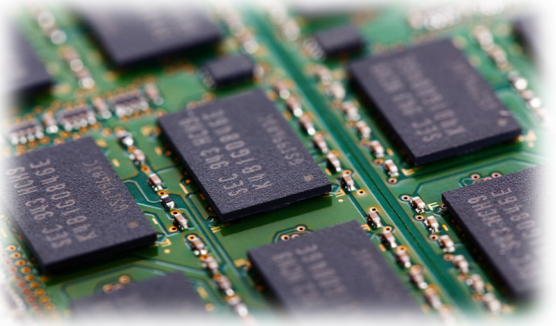
- **Decomposes** a problem into a **sequence of explicit reasoning steps**, each performed by a **Memory-Attention-Composition** (MAC) cell
- One **universal recurrent MAC cell** is used throughout all the steps, where its behavior is **versatile**, adapting to the context in which it is applied
- The network can represent **arbitrarily complex reasoning graphs** in a **soft** manner (***self-attention***), maintaining an **end-to-end differentiability**

Memory. Attention. Composition. The MAC Network

Each **MAC cell** is responsible for performing **one reasoning step at a time**. It maintains **recurrent dual states**:

- ***Control*** c_i: this step's **reasoning operation**

  Attention-based average *of a given* ***query*** *(question)*



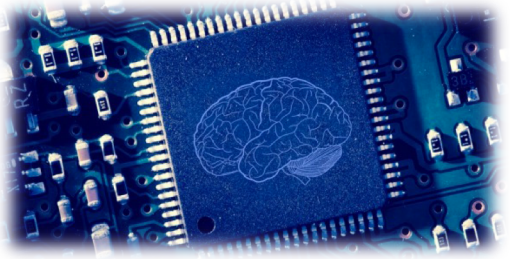
- ***Memory*** m_i: **retrieved information** relevant to a query, accumulated over steps

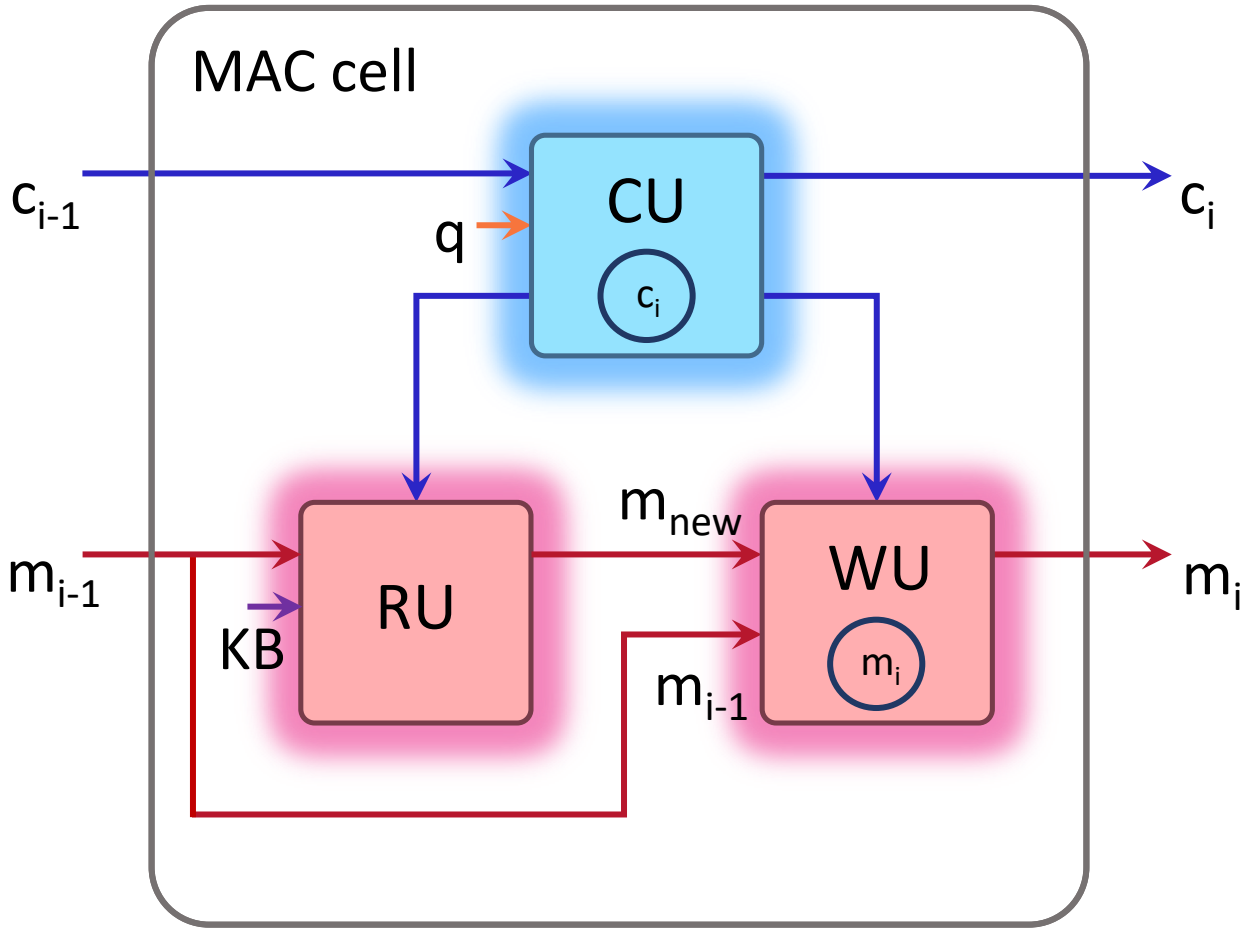
  Attention-based average *of a given* ***KB*** *(image)*

Network Inputs

- ***Query***: processed by a **biLSTM**, yielding:
  - A series of **output states** that we call ***contextual words***, $cw_1, \ldots, cw_n$
  - An **overall query represenation**, $\boldsymbol{q} = \left[\overleftarrow{h_1^B}, \overrightarrow{h_n^F}\right]$ the **concatenation of the final forward and backward hidden states**
- ***Knowledge Base***: For VQA, an **image** represented by ResNet-101 **extracted features** preprocessed by 2-layer CNN, yielding $\boldsymbol{KB_V}$ of **dimensions** $[\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{d}]$

Memory. Attention. Composition.
The MAC cell



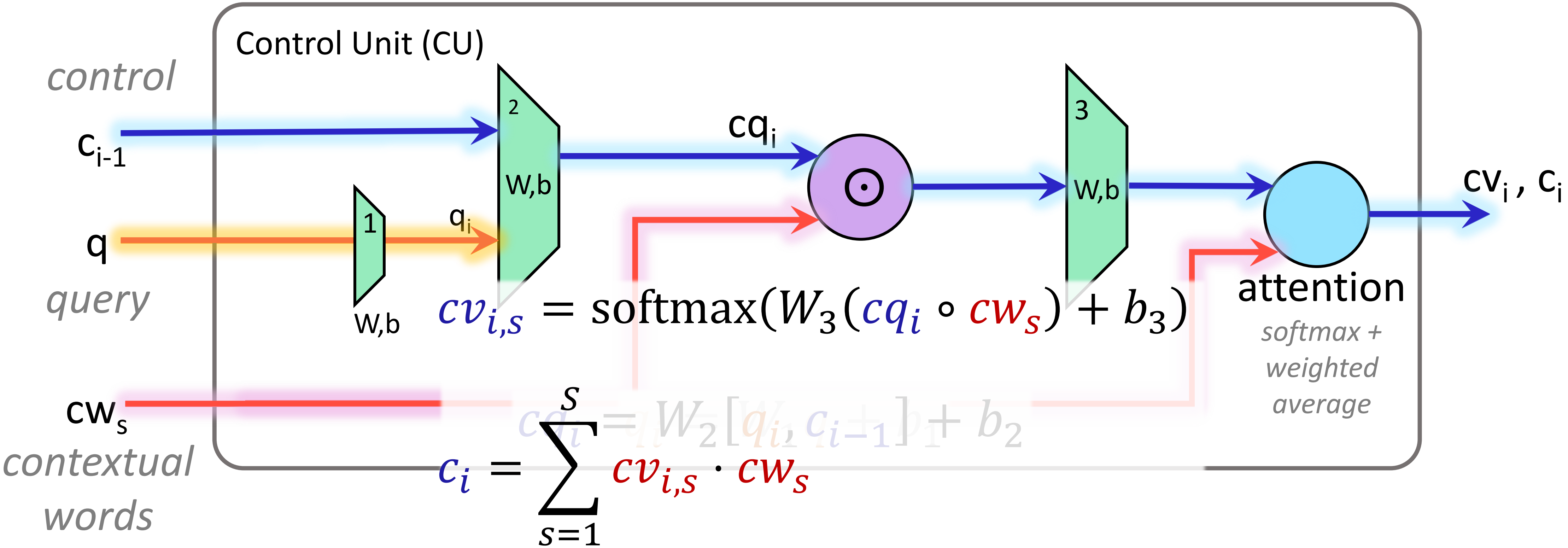


- ***Control Unit (CU)*** **computes** a ***control*** state, extracting an ***instruction*** that **focuses** on some **aspect of the query**
- ***Read Unit (RU)****:* **retrieves information** from the ***knowledge base*** given the **current *control*** state and **previous *memory***
- ***Write Unit (WU)****:* **updates** the new ***memory*** state, **merging old** and **new** information

The MAC cell

The Control Unit (CU)

Extract an instruction (control) from the question



$$cv_{i,s} = \text{softmax}(W_3(cq_i \circ cw_s) + b_3)$$

$$cq_i = W_2[q_i, c_{i-1}] + b_2$$

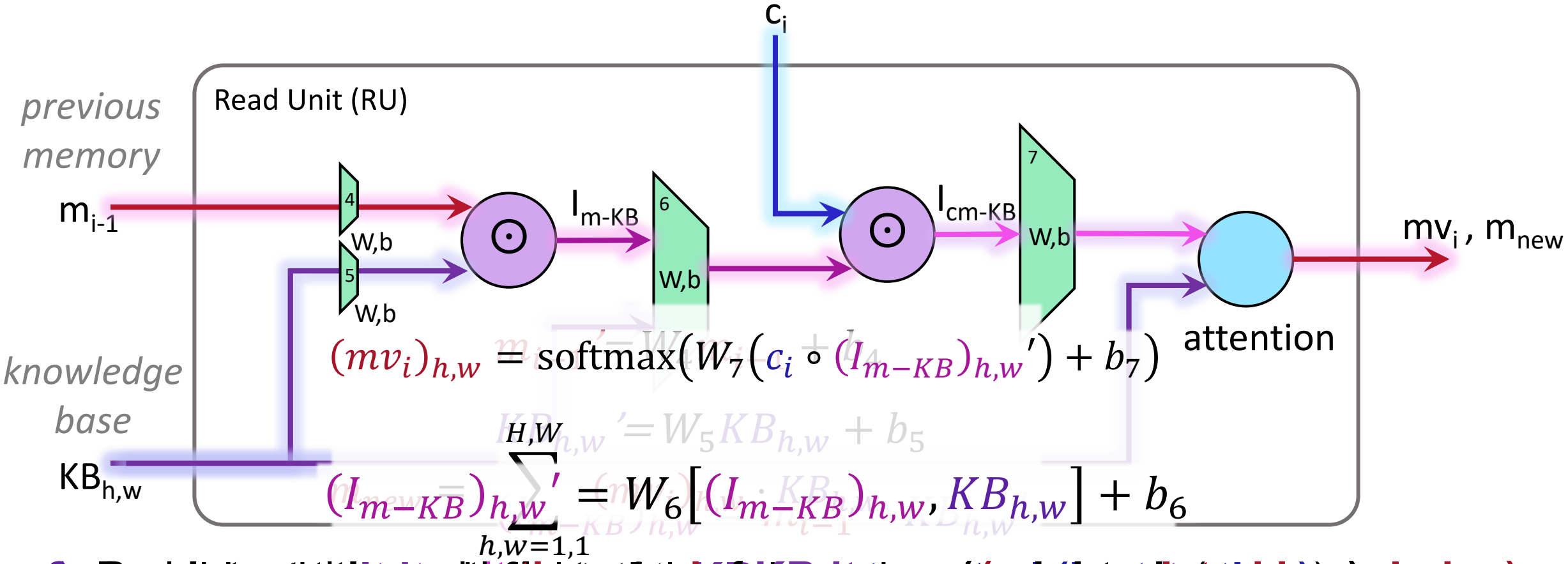
$$c_i = \sum_{s=1}^{S} cv_{i,s} \cdot cw_s$$

… (weighted-average)

The MAC cell

The Read Unit (RU)

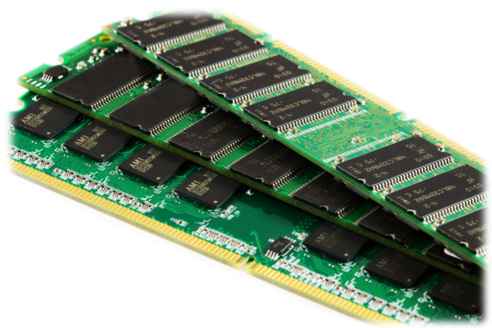
Retrieve information based on the current instruction



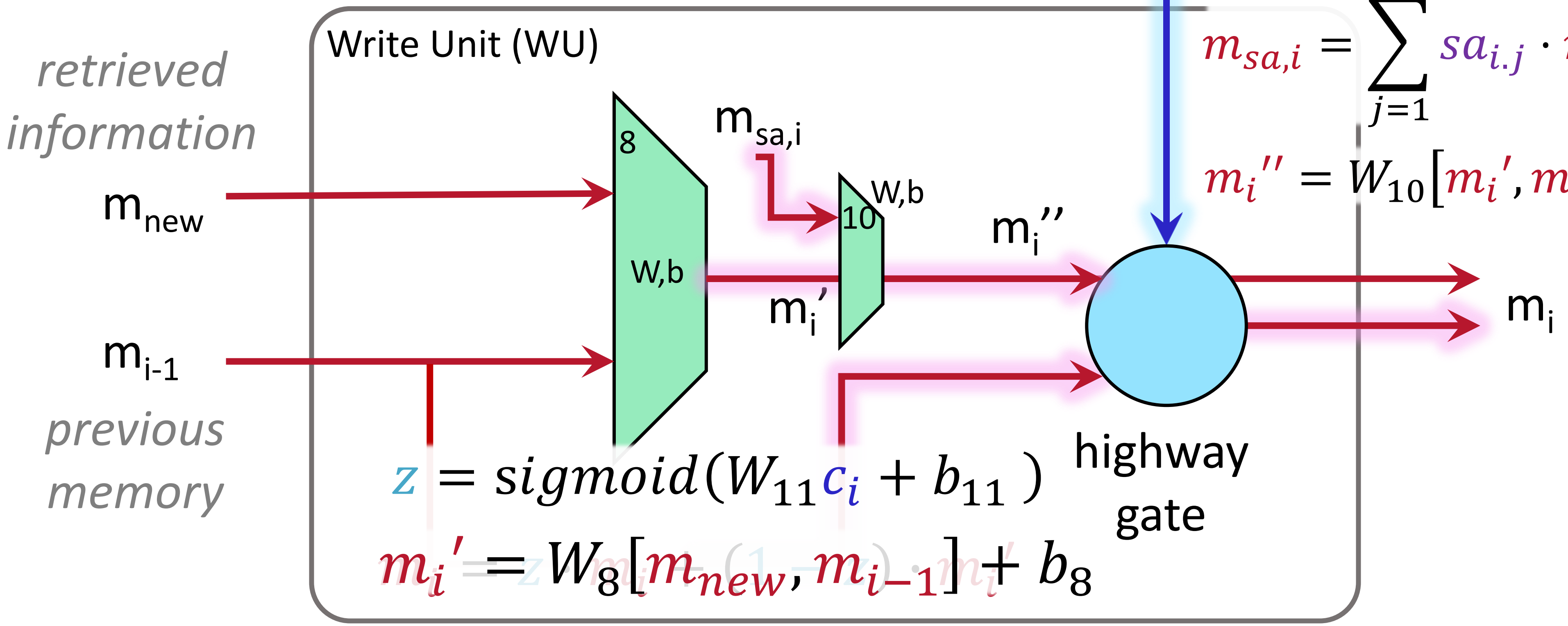
2. Use the previous KB memory (accumulated knowledge)

The MAC cell

The Write Unit (WU)



Combine **retrieved information** **with** **accumulated knowledge (memory)**



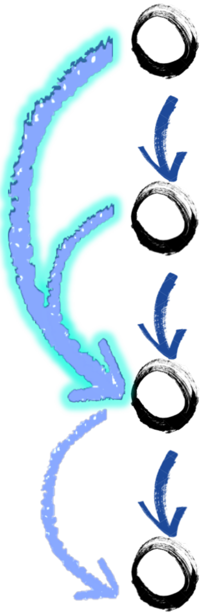
$$sa_{i,j} = \text{softmax}\left(W_9\left(c_i \odot c_j\right) + b_9\right)$$

$$m_{sa,i} = \sum_{j=1}^{i-1} sa_{i,j} \cdot m_j$$

$$m_i'' = W_{10}\left[m_i', m_{sa,i}\right] + b_{10}$$

$$z = sigmoid(W_{11} c_i + b_{11})$$

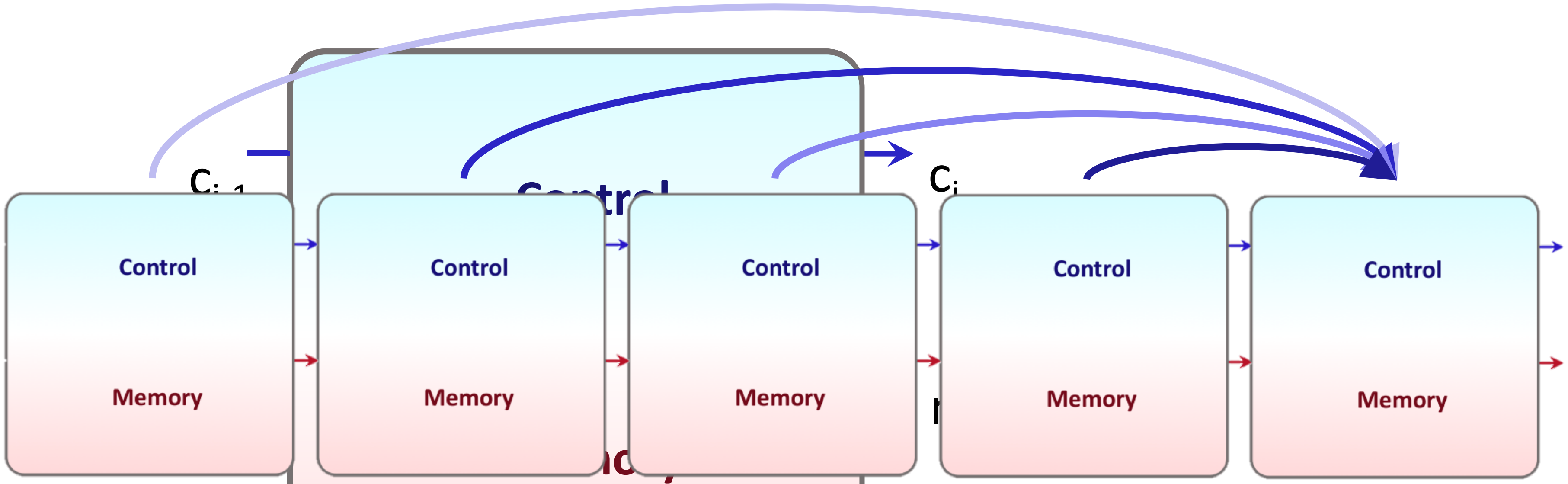
$$m_i' = W_8\left[m_{new}, m_{i-1}\right] + b_8$$



- Self-Attention: non-sequential computations (trees, DAGs)
- Gating: adapt the length of computation to question complexity

The MAC net
From Cell to Network

A **MacNet is a soft-attention sequence** of *p* MAC cells



Uniform: **capacity** to represent **sequential st** ... **s**;

reasoning, **efficient**, **easy to deploy**, **Directed Acyclic Graphs** (DAGs), and **fully differentiable**

Network Outputs

The result is predicted using the ***query*** and **final *memory***

*For CLEVR, a **softmax classifier** predicts one-word answer among the 28 candidates*

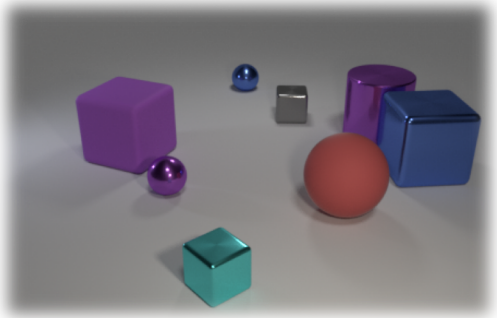
Considering both the final *memory* and the *query* is critical.

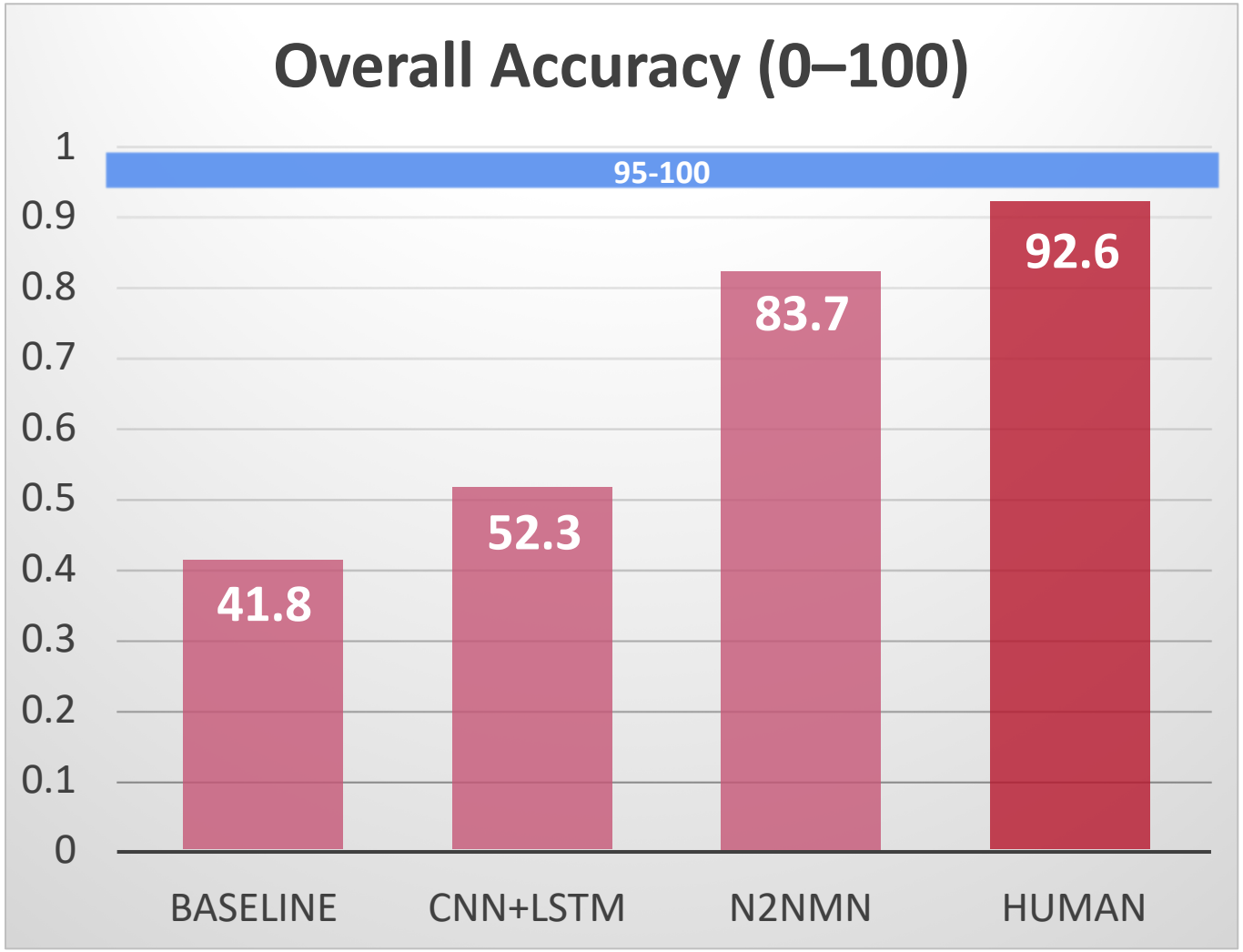
- The ***memory*** represents the **information retrieved from the *KB*** deemed relevant to answer the query
- It **does not directly contain** information about **the *query***
- Thus, the model has to recall the specifics of the query, embedded in its **continuous representation** q, to address it correctly

Talk Outline

- ✓ From Machine Learning to Machine Reasoning
- ✓ The CLEVR task
- ✓ Memory-Attention-Composition Networks (MAC nets)
- ➢ Experiments and Discussion
- ○ Conclusion

Experiments
CLEVR Overall Results

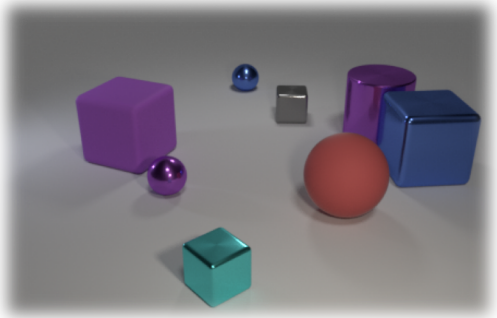


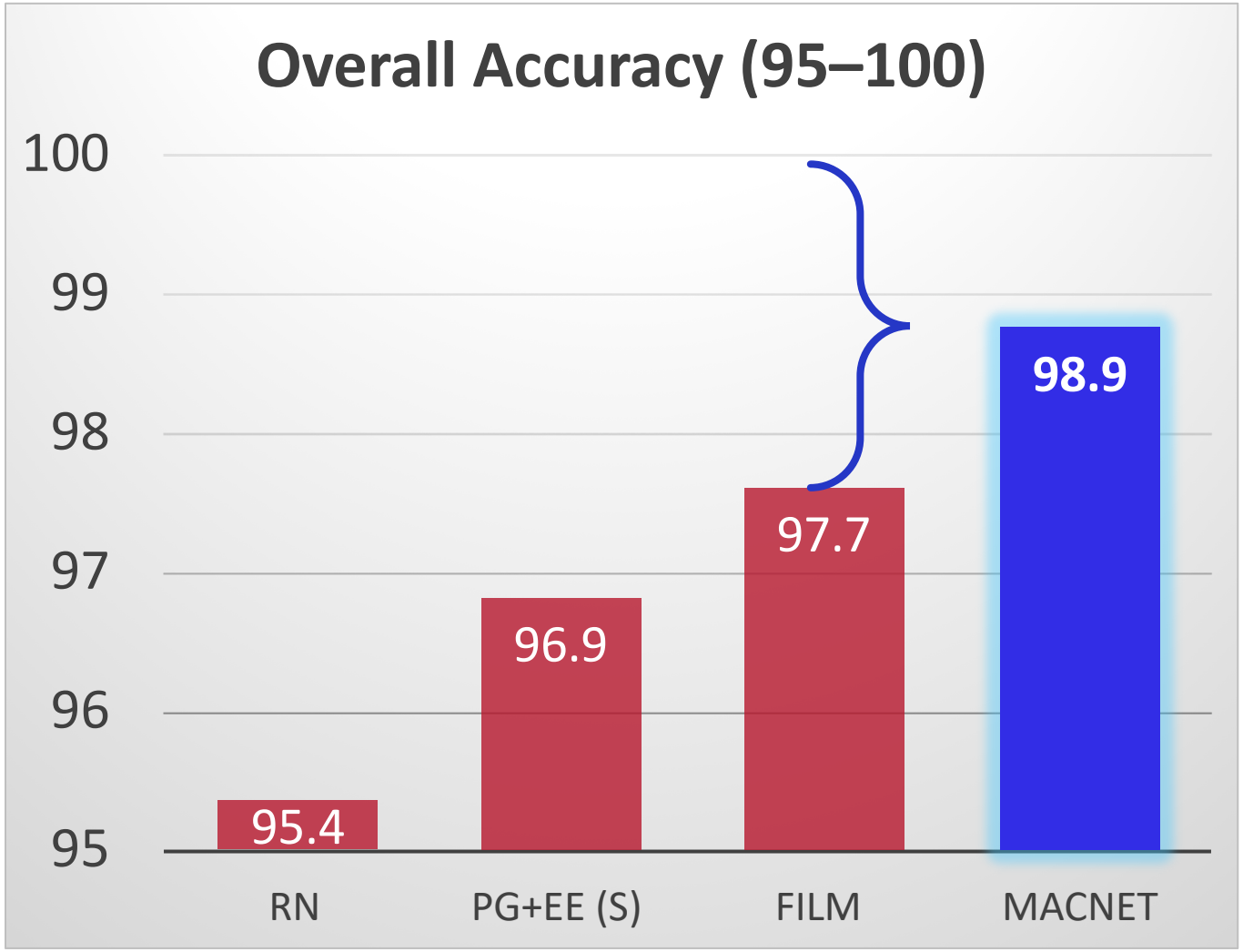


- **700k** Training set
- **150k** Test set
- **28** candidate answers

- ***Baseline***: **the most frequent answer** for each **question type**

Experiments
CLEVR Overall Results

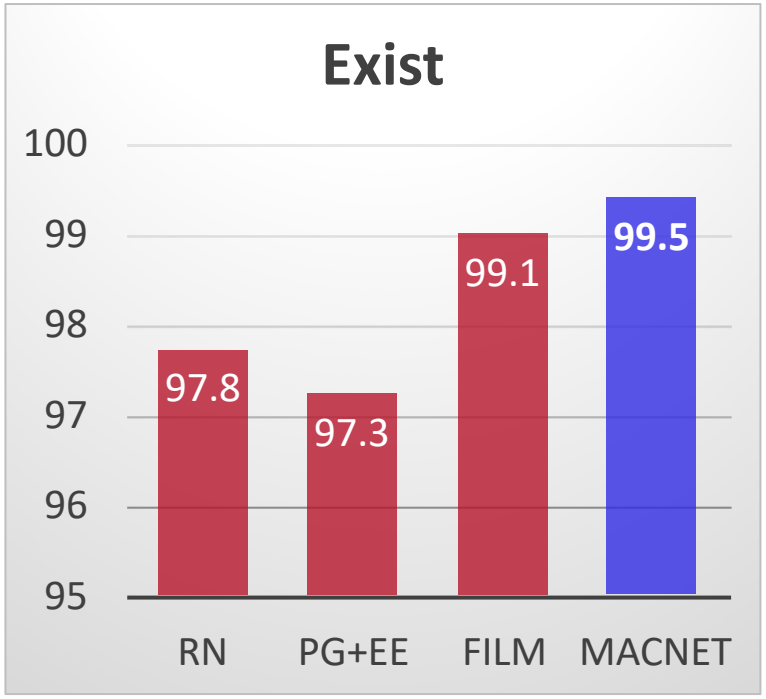


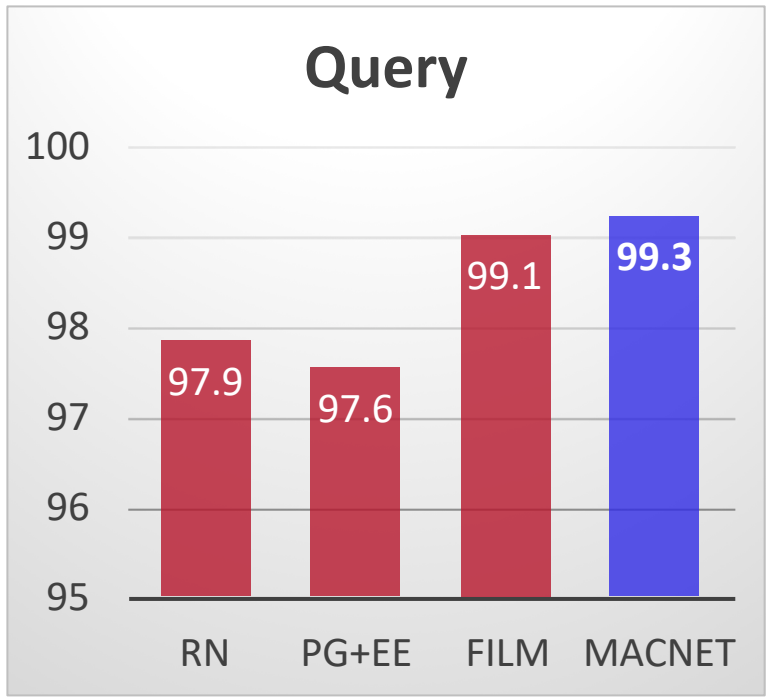


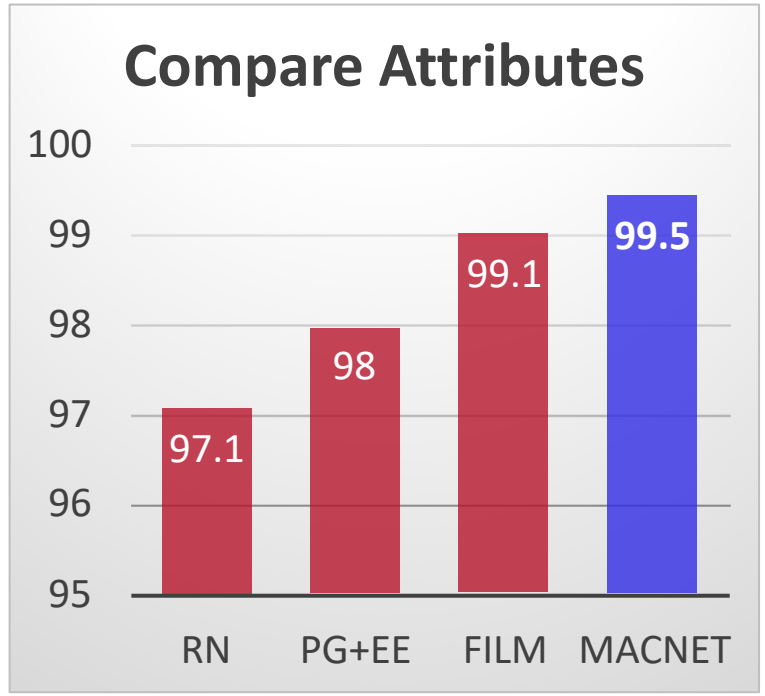
- **(S)**: strongly supervised
- MAC net **halves** the previous best **error rate**

Experiments

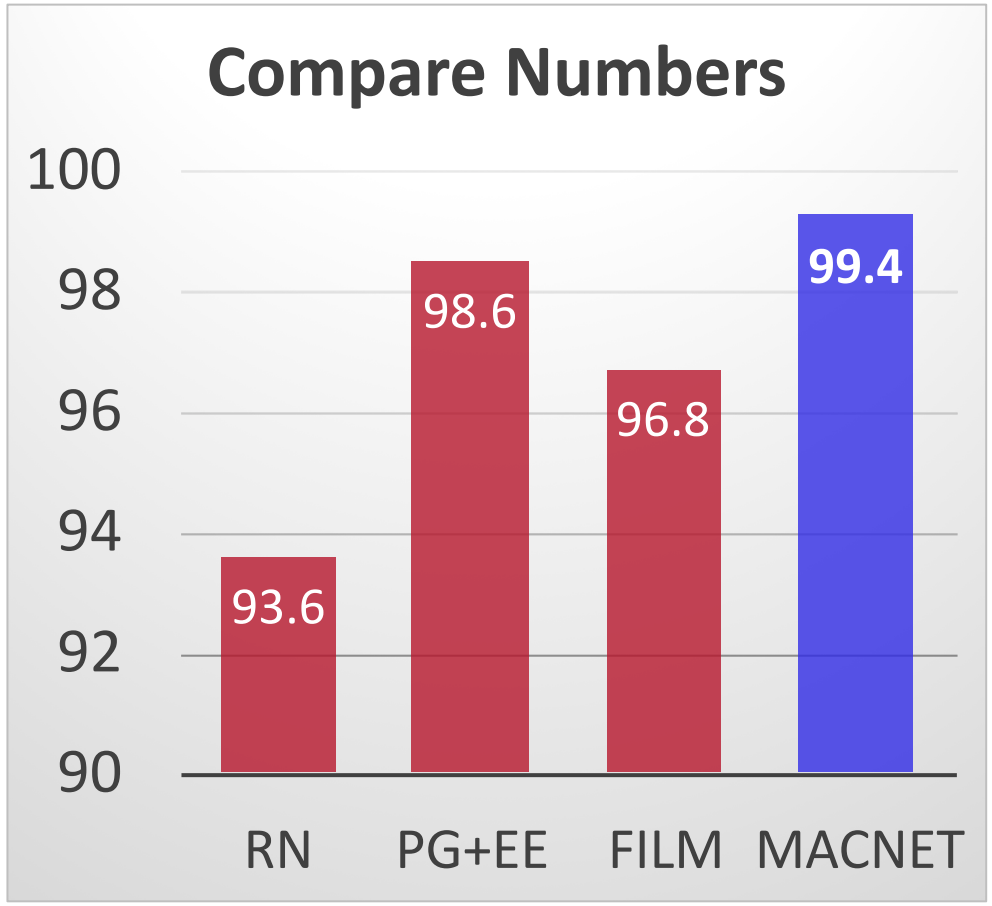
Accuracy Per Type

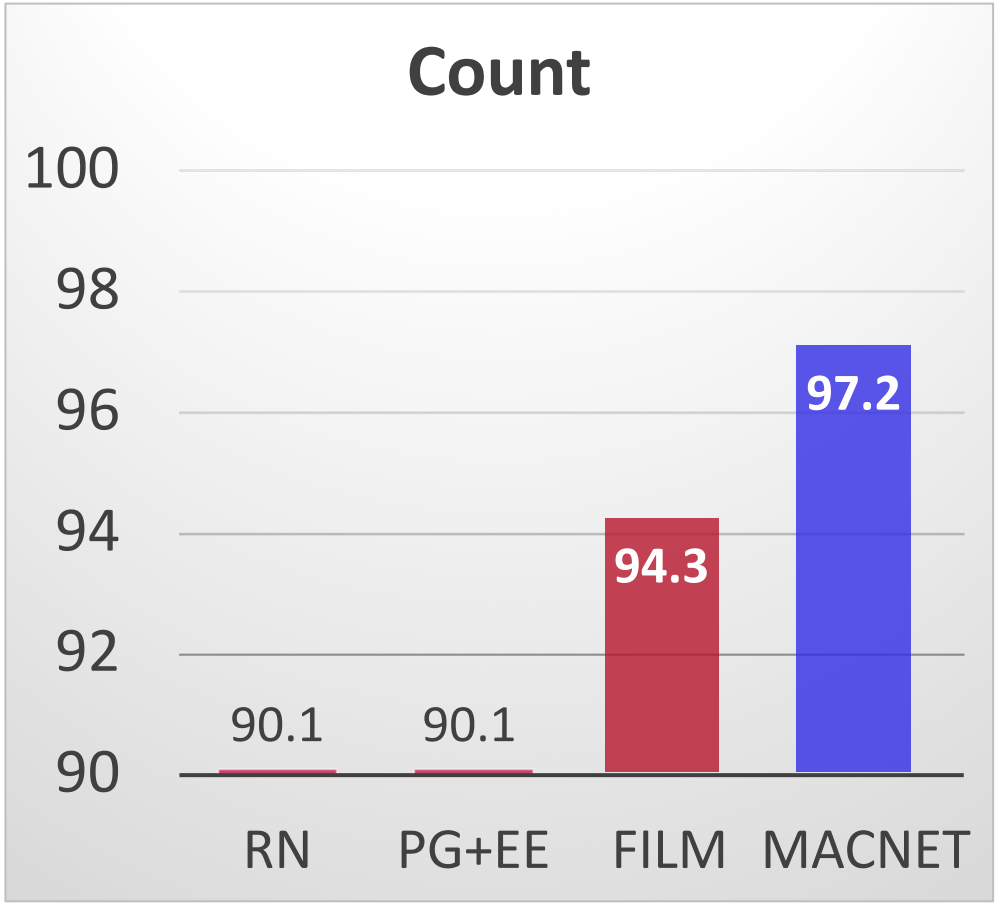




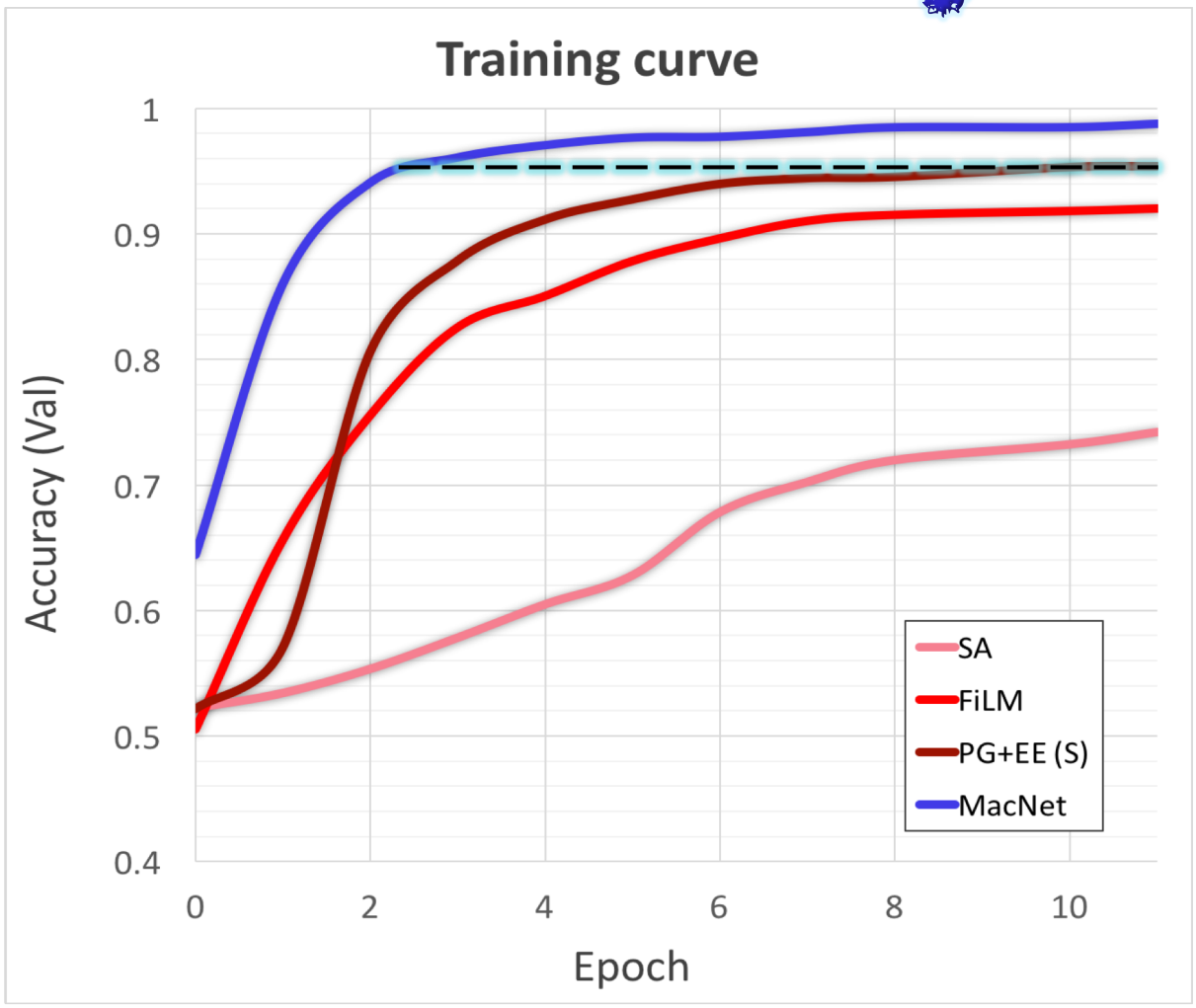


Experiments
Counting and Numbers



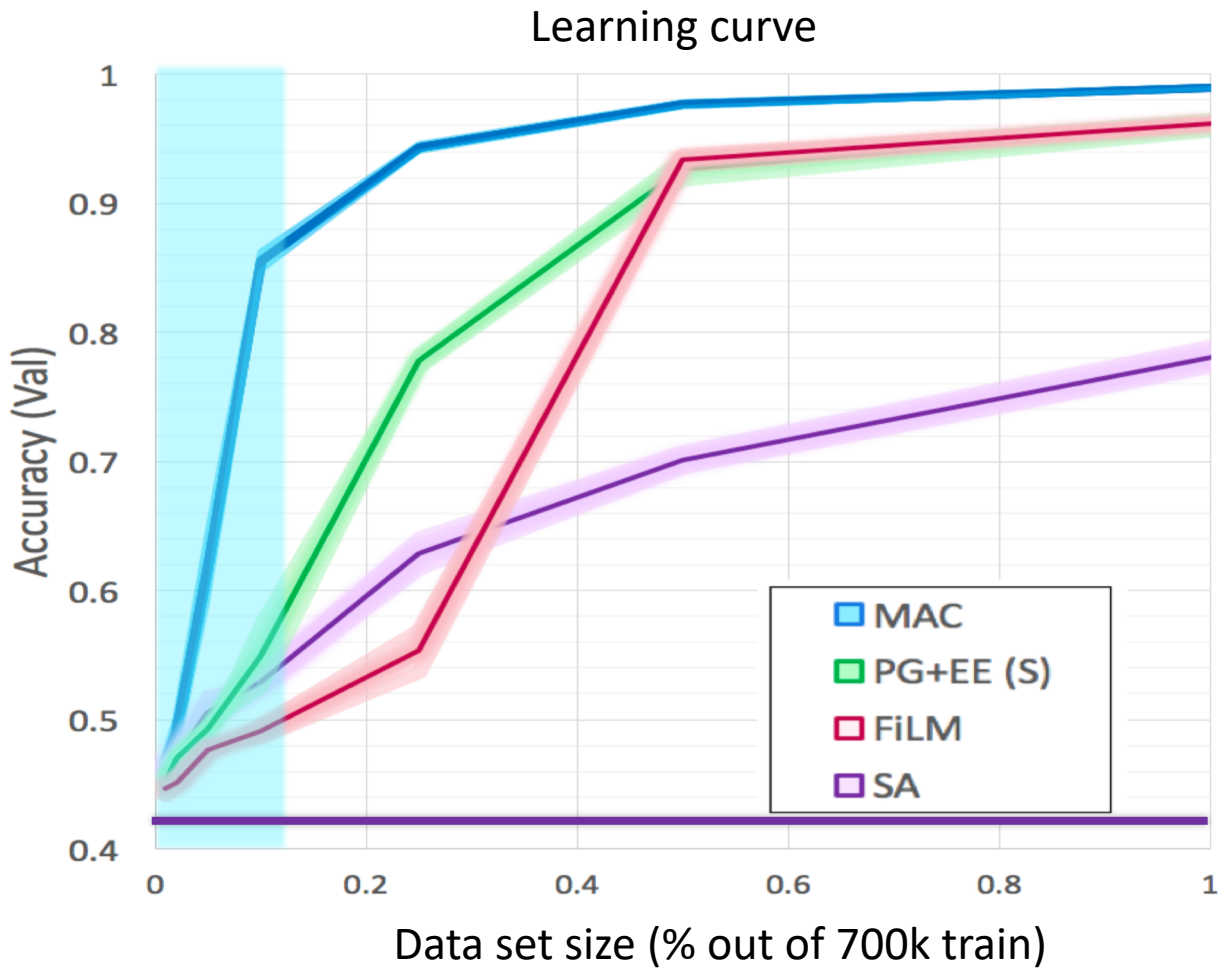


Experiments
Generalizability



- **40x** training **speedup** (timewise) compared to **Relation Networks (RN)**.
- **10x speedup** (timewise) compared to **FiLM**, to match its best score.

Experiments
Data Efficiency



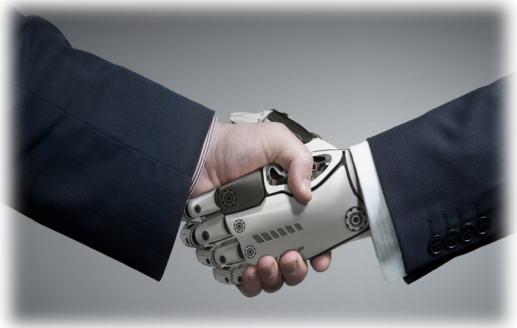
For **10% of the CLEVR** dataset, **70k examples**:

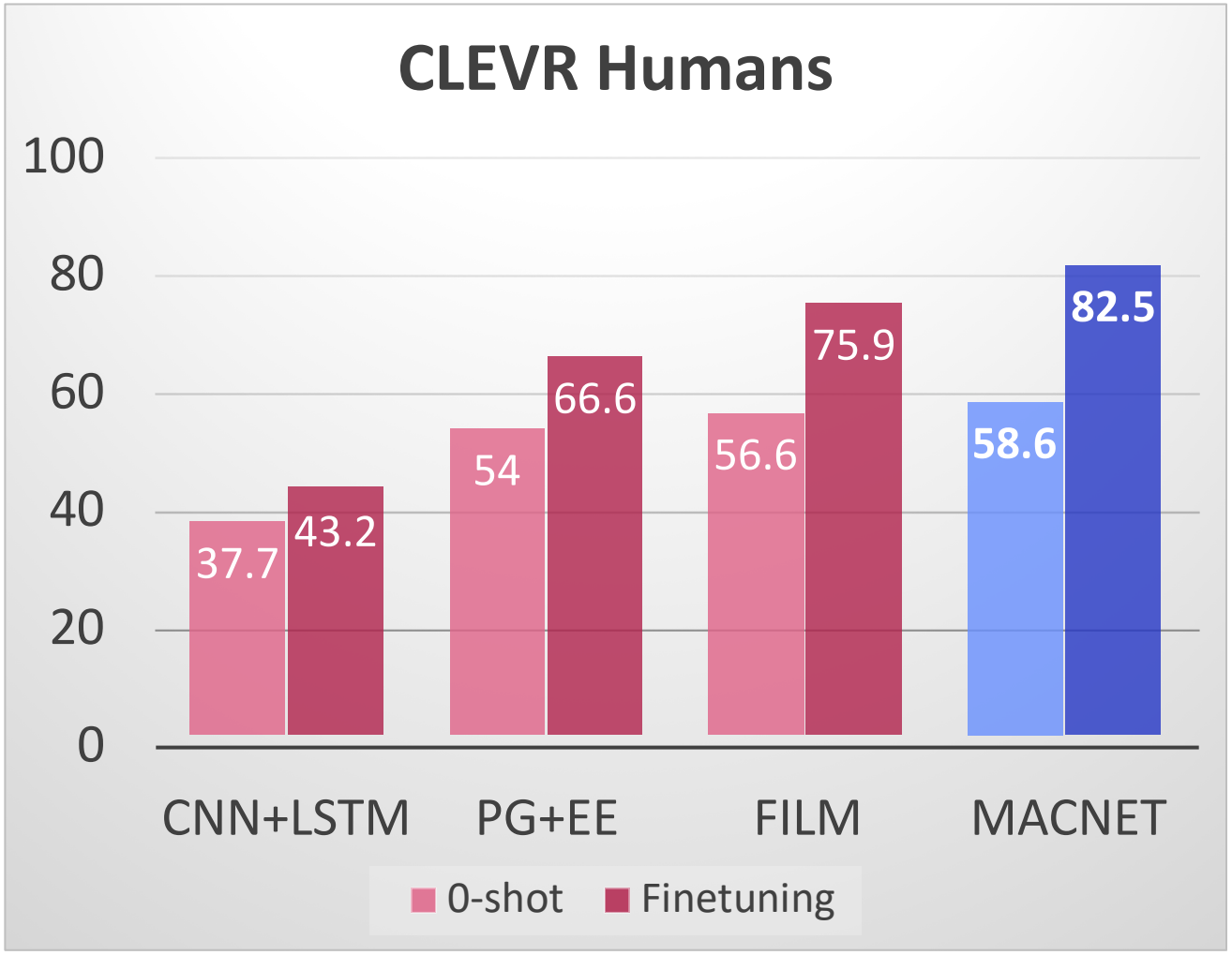
- **MacNet** achieves **86%**
- **Other approaches** obtain **51.6%** at best
- **Baseline** achieves **41.8%**

Baseline

Most Frequent Answer for Question Type

Experiments
CLEVR-Humans

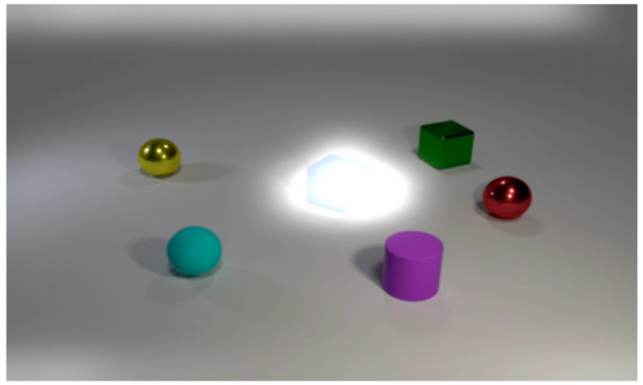


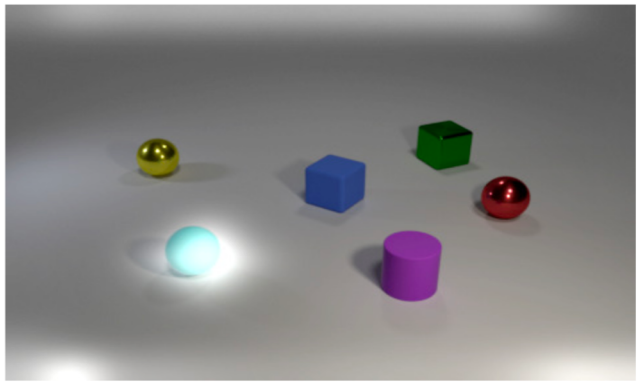


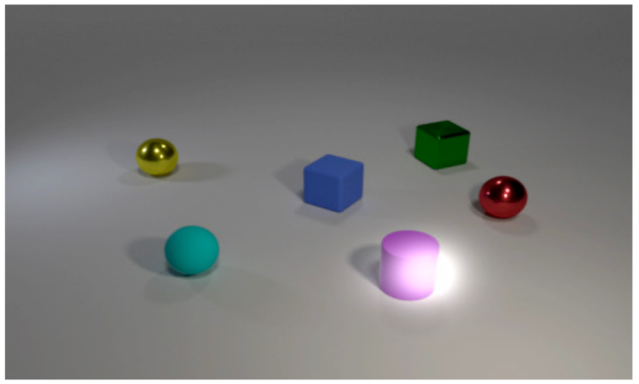
- CLEVR-Humans is **18k natural language questions** collected through **crowdsourcing**
- They wrote *"**questions hard for a smart robot to answer**"*
- Dataset has **diverse vocabulary** and **linguistic variation**; demands more **varied reasoning skills**

Attention visualizations

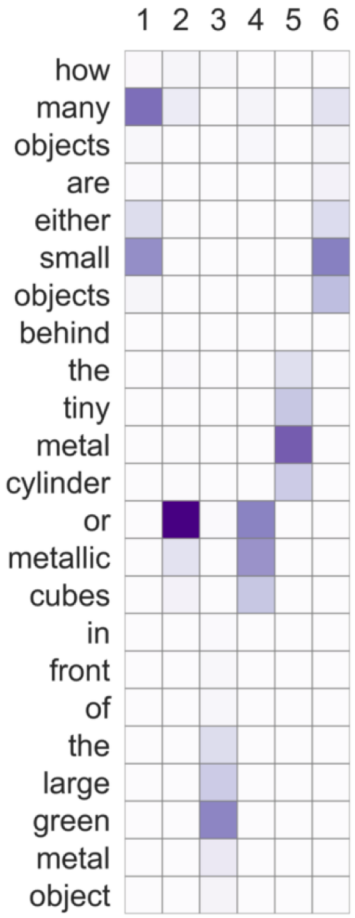
What color is the matte thing to the right of the sphere in front of the tiny blue block? Purple

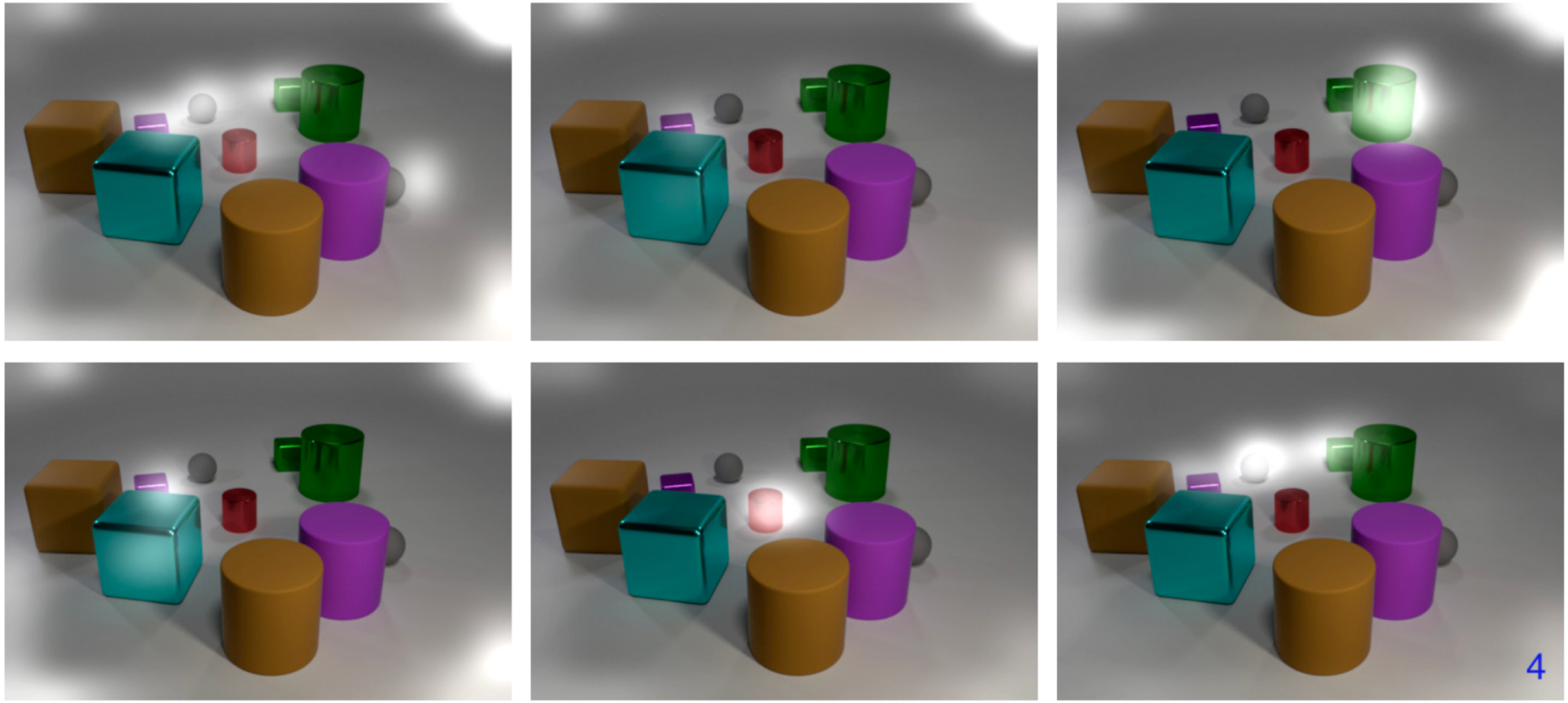




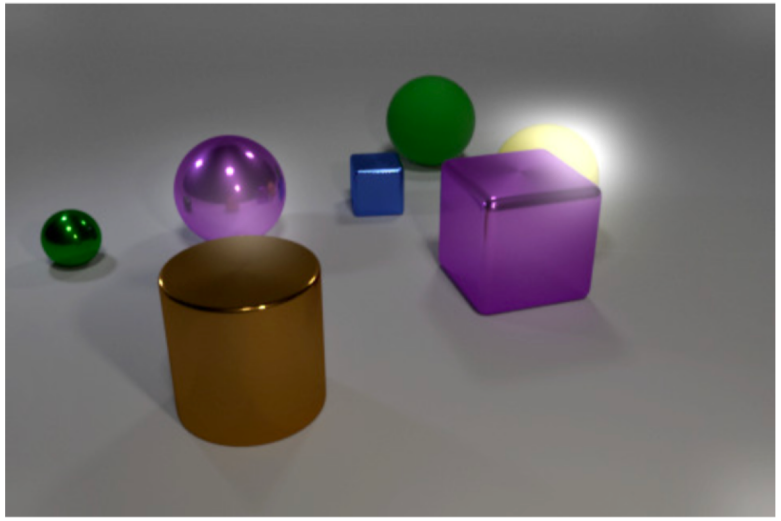


Attention visualizations

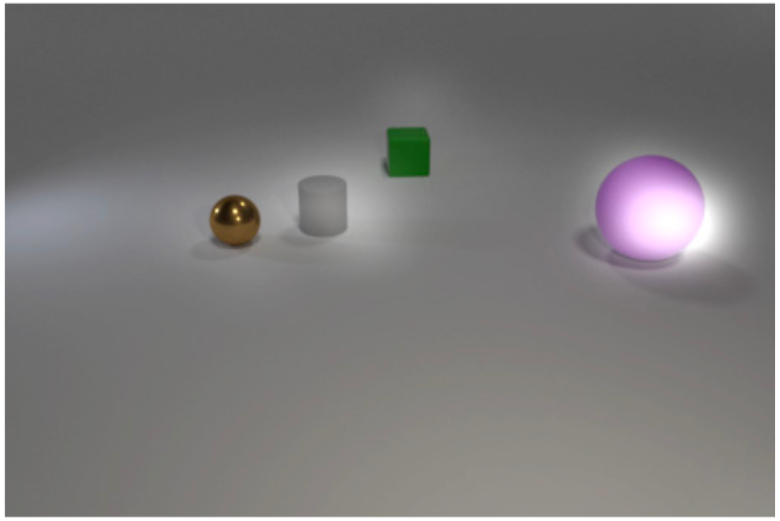




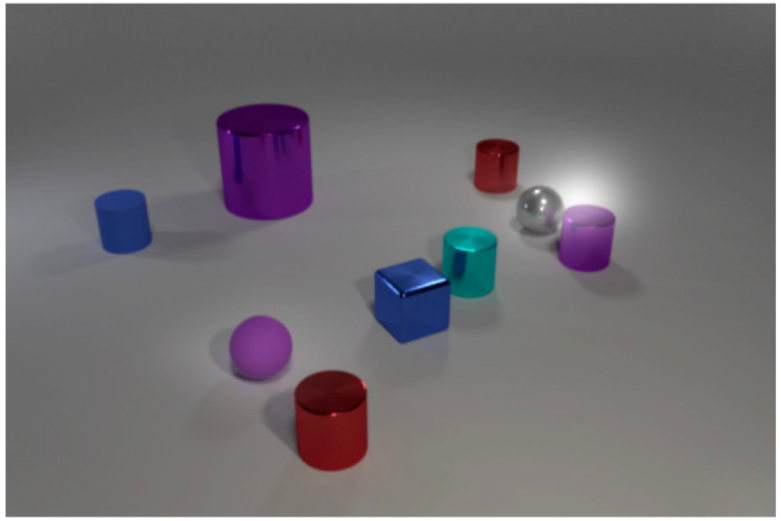
CLEVR-Humans examples



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* ✓

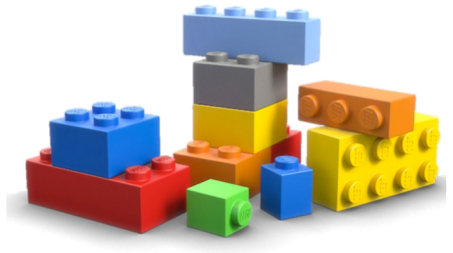


Q: *What color ball is* ***close to*** *the small purple cylinder?* **A:** *gray* ✓

Universality
MAC nets vs Module Nets

- The **MAC cell** is **universal** and **versatile** – the **same cell is used across all states**

  - **Sharing** both **architecture** and **parameters**
  - **Continuously adapting** its behavior to the context in which it is applied
- **Module Networks** advocate for a **discrete** and **fixed inventory** of (task-specific) **specialized modules**, with **distinct parameters** and even architectures



Distant modulation

MAC nets and FiLM

The ***query*** and ***KB*** representations are ***not*** **transformed into the same vector space**

- **MAC net**: The interaction between them is **mediated** by **discrete probability distributions**
- **FiLM**: The ***question*** affects the computation over the ***image*** by applying **conditional normalization layer** to a **CNN pipeline**
  - The influence is **identical across all image regions**
  - Doesn't allow **selective behavior** based on **feature values** or **positions**

A compositional reasoning engine

- **A new design for a compositional reasoning engine**

  A constrained sequence model, separating control and memory and exploiting attention is a good prior for reasoning

- **Strong compositional reasoning skills**

  Halves the previous lowest error rate

  Generalizes much better from more modest training data

  Generalizes better to new tasks in CLEVR-Humans

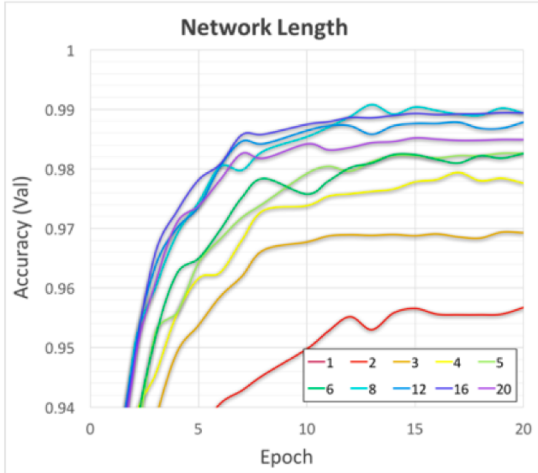
- **Generic, fully differentiable end-to-end model**

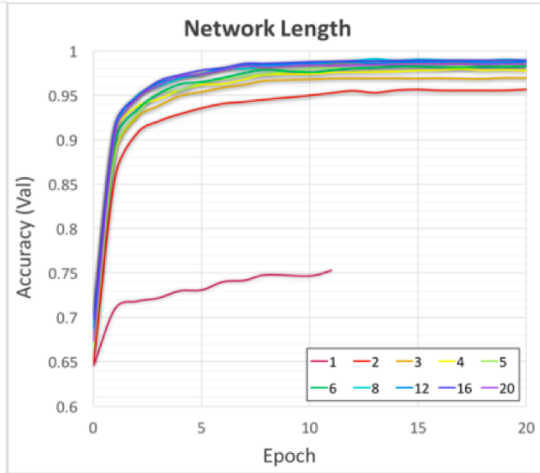
Thank you!

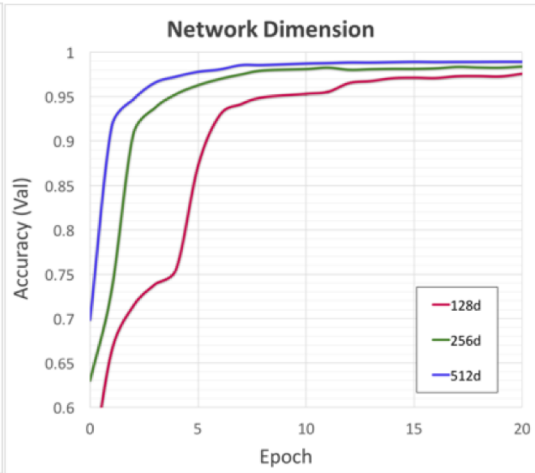
Based on the paper:

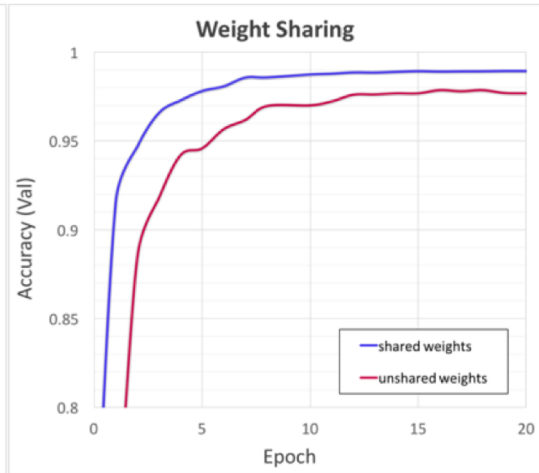
“Compositional Attention Networks for Machine Reasoning”

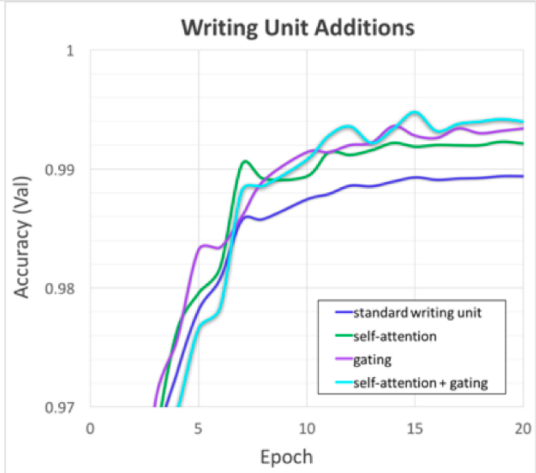
Drew A. Hudson and Christopher D. Manning, ICLR 2018 (arXiv)

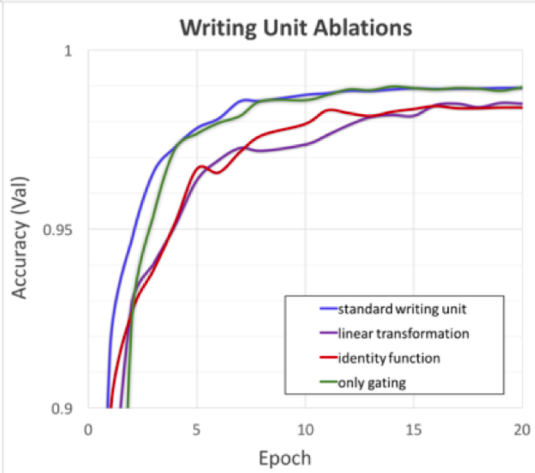

Network Length
Accuracy (Val)
Epoch
1
2
3
4
5
6
8
12
16
20

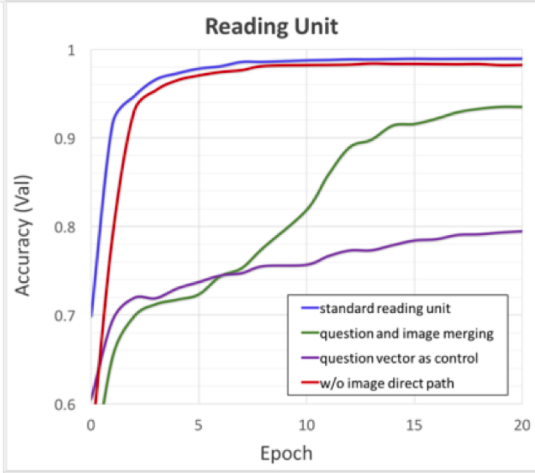

Network Length
Accuracy (Val)
Epoch
1
2
3
4
5
6
8
12
16
20

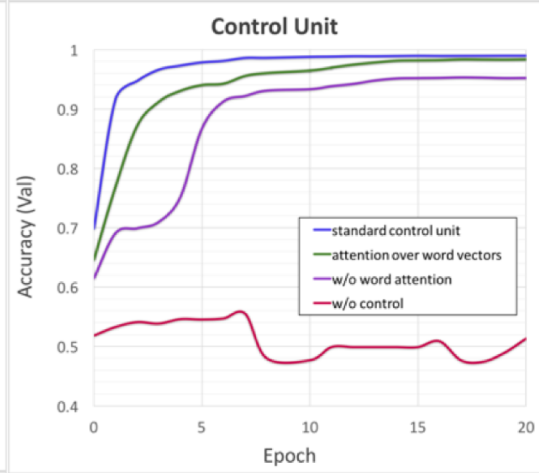

Network Dimension
Accuracy (Val)
Epoch
128d
256d
512d

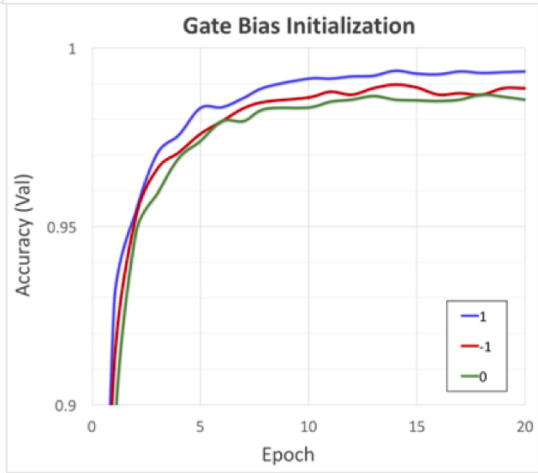

Weight Sharing
Accuracy (Val)
Epoch
shared weights
unshared weights

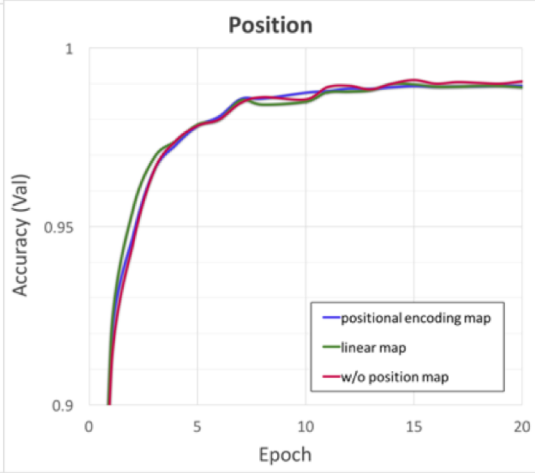

Writing Unit Additions
Accuracy (Val)
Epoch
standard writing unit
self-attention
gating
self-attention + gating

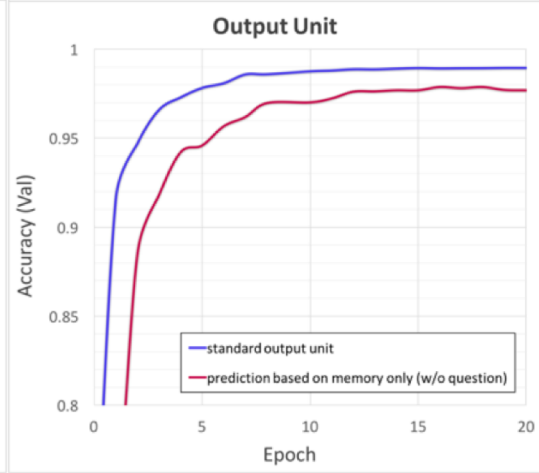

Writing Unit Ablations
Accuracy (Val)
Epoch
standard writing unit
linear transformation
identity function
only gating


Reading Unit
Accuracy (Val)
Epoch
standard reading unit
question and image merging
question vector as control
w/o image direct path


Control Unit
Accuracy (Val)
Epoch
standard control unit
attention over word vectors
w/o word attention
w/o control


Gate Bias Initialization
Accuracy (Val)
Epoch
1
-1
0


Position
Accuracy (Val)
Epoch
positional encoding map
linear map
w/o position map


Output Unit
Accuracy (Val)
Epoch
standard output unit
prediction based on memory only (w/o question)