

Building neural network models that can reason

Stanford



Christopher Manning and Drew Hudson

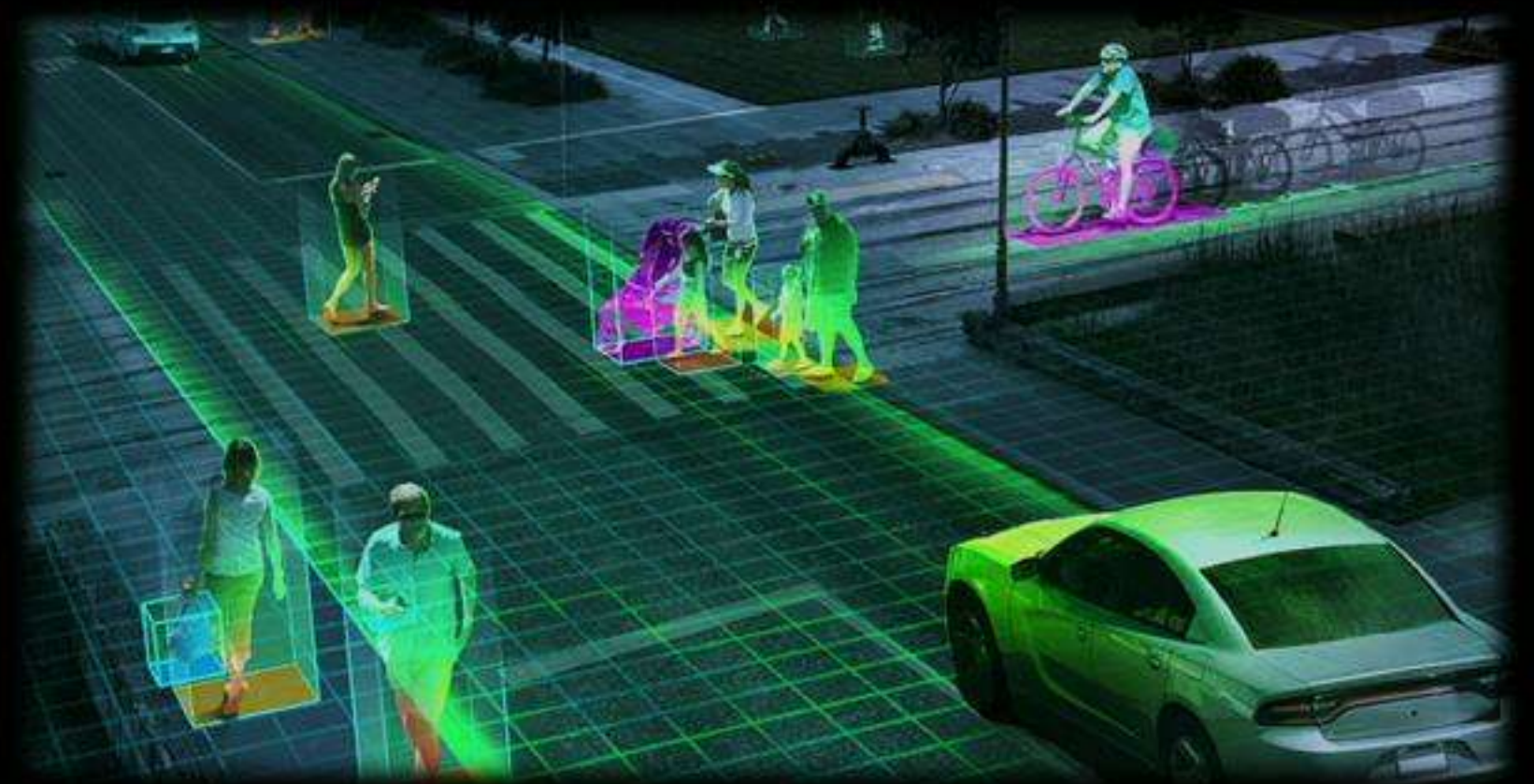
Stanford University

@chrmanning * @stanfordnlp





Ich fliege nach Kanada
Tengo sed
I will fly to Canada
I am thirsty



But what about reasoning?

But what about reasoning?



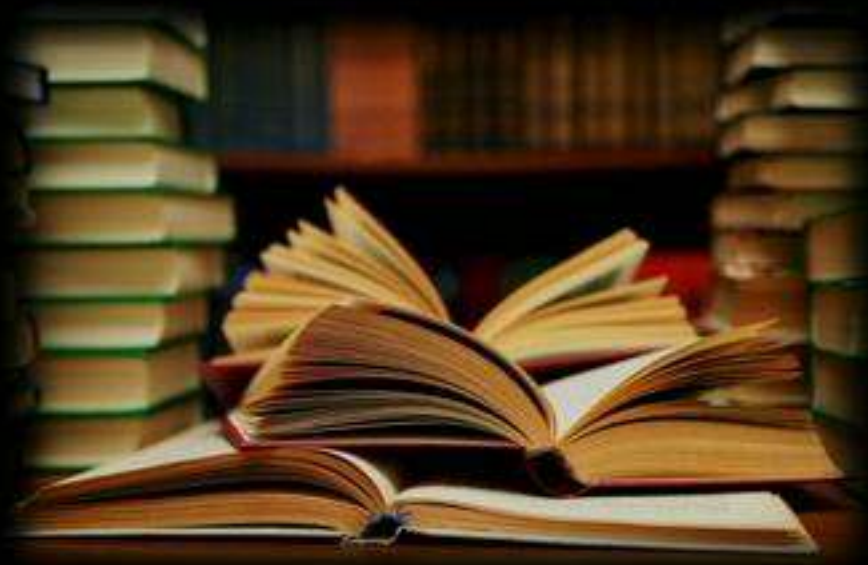
But what about reasoning?



But what about reasoning?



But what about reasoning?



What is Reasoning? [Bottou 2011]

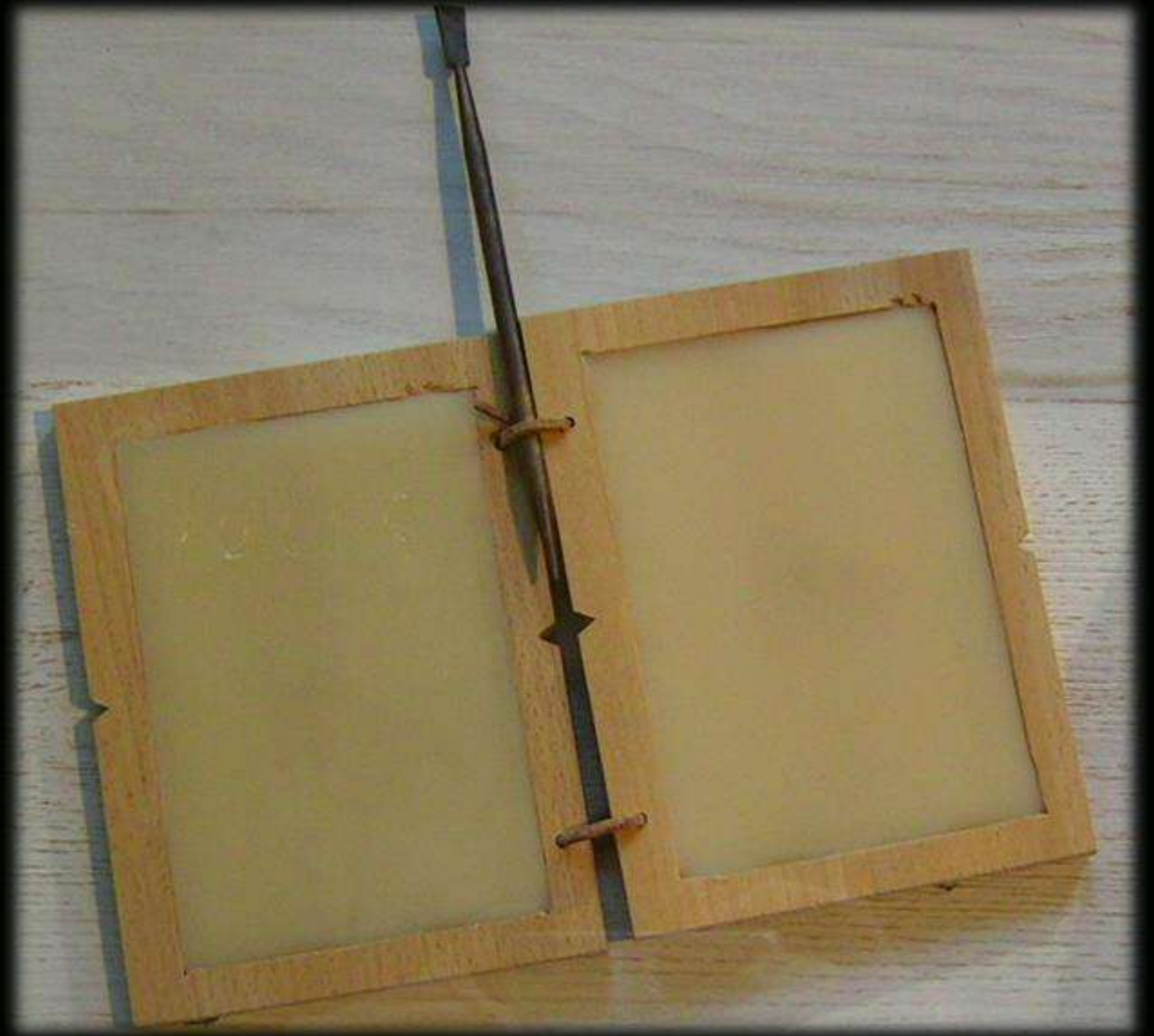


What is Reasoning? [Bottou 2011]



- **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



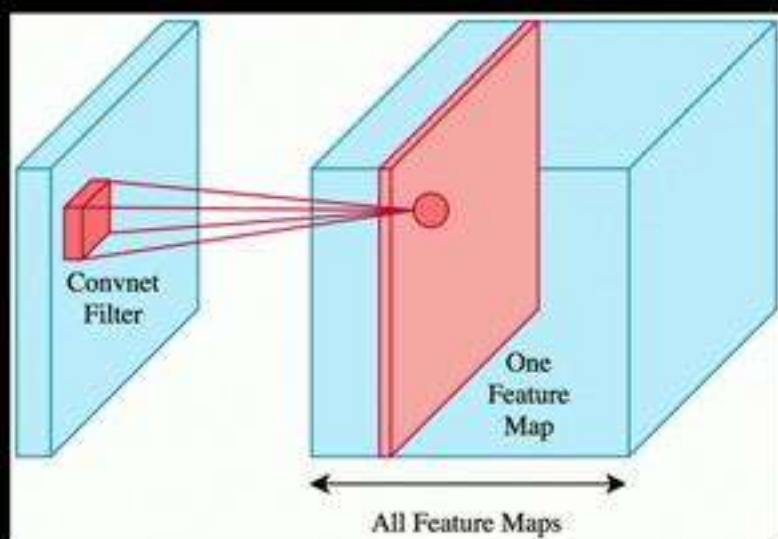
Worshipping the
tabula rasa

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

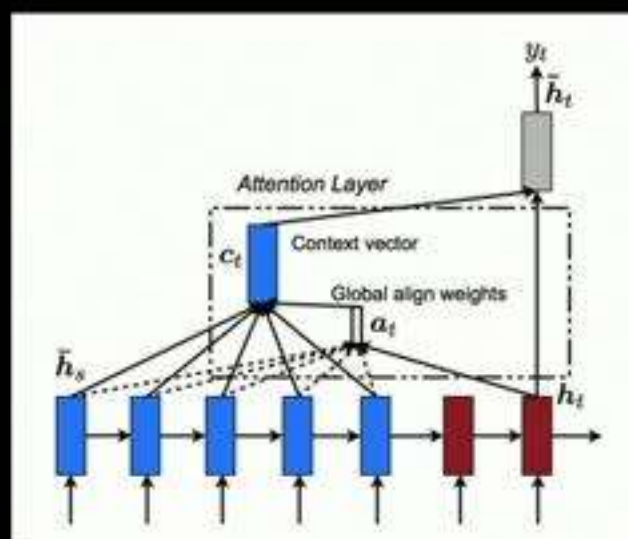


Appropriate structural priors

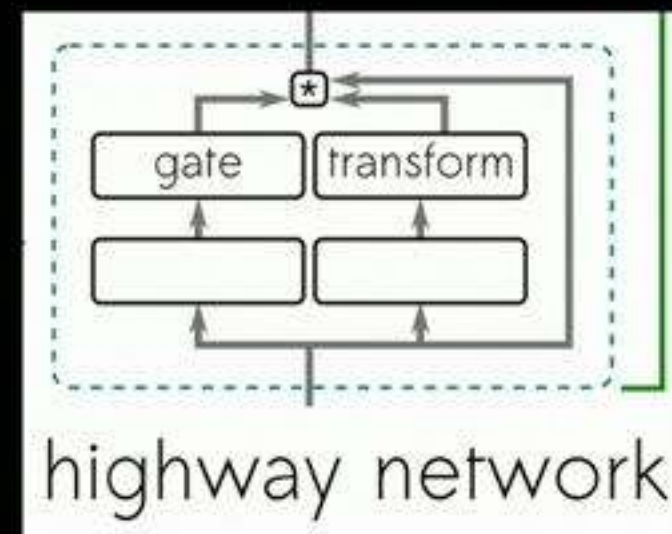
Appropriate structural priors



Convolution

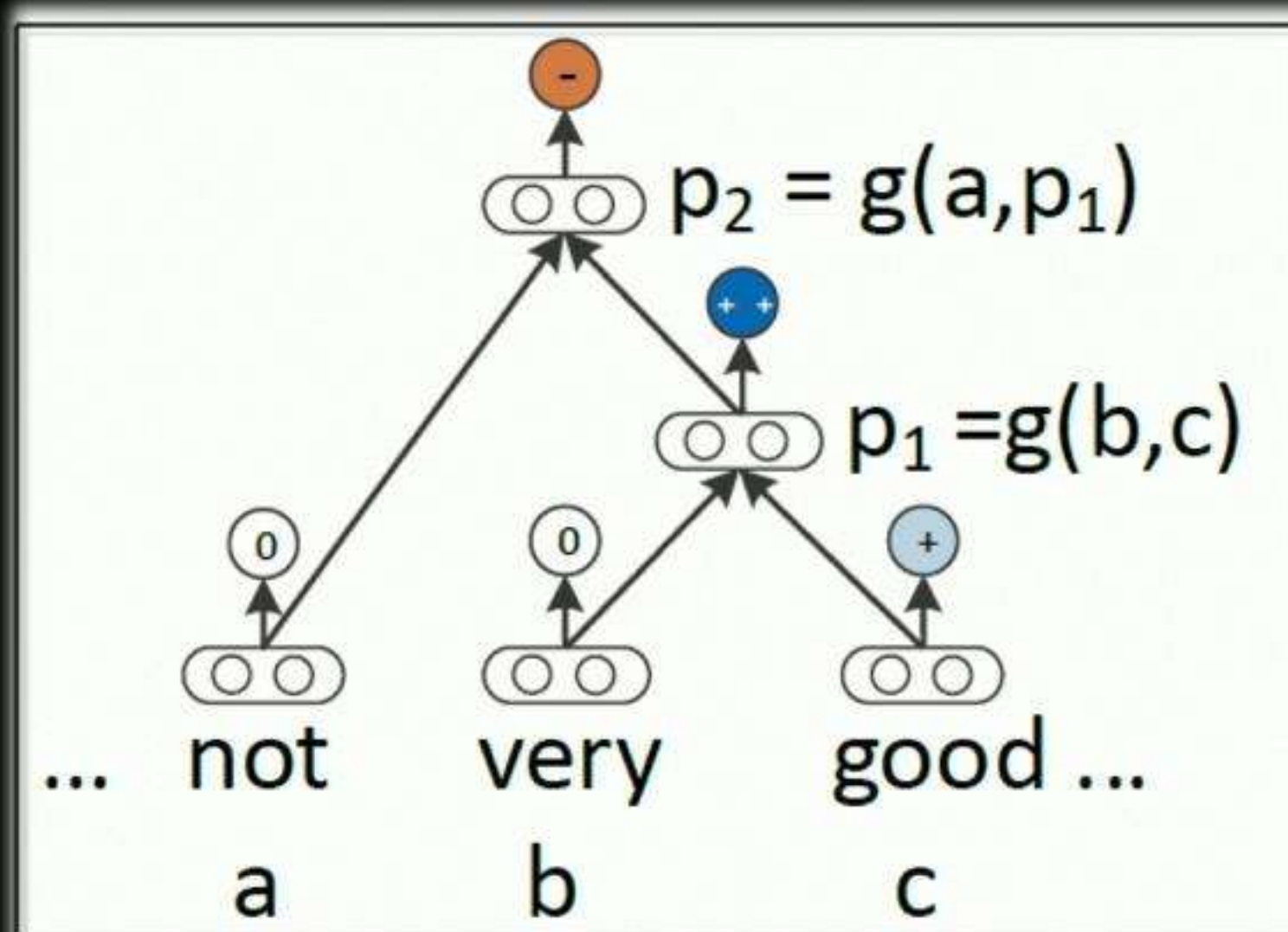


Attention



Gating (LSTM/highway)

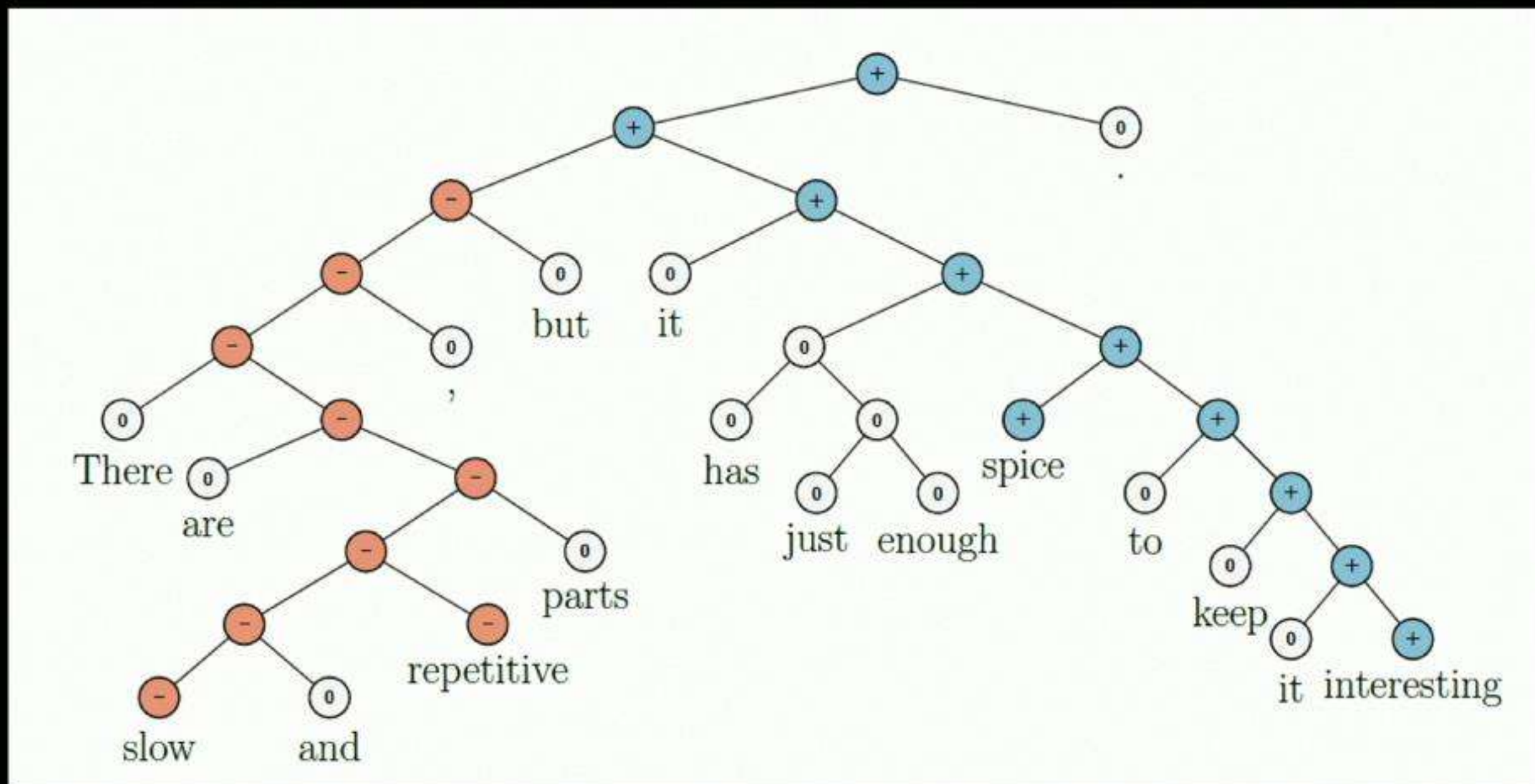
Tree-structured models



[Socher et al. 2010ff]

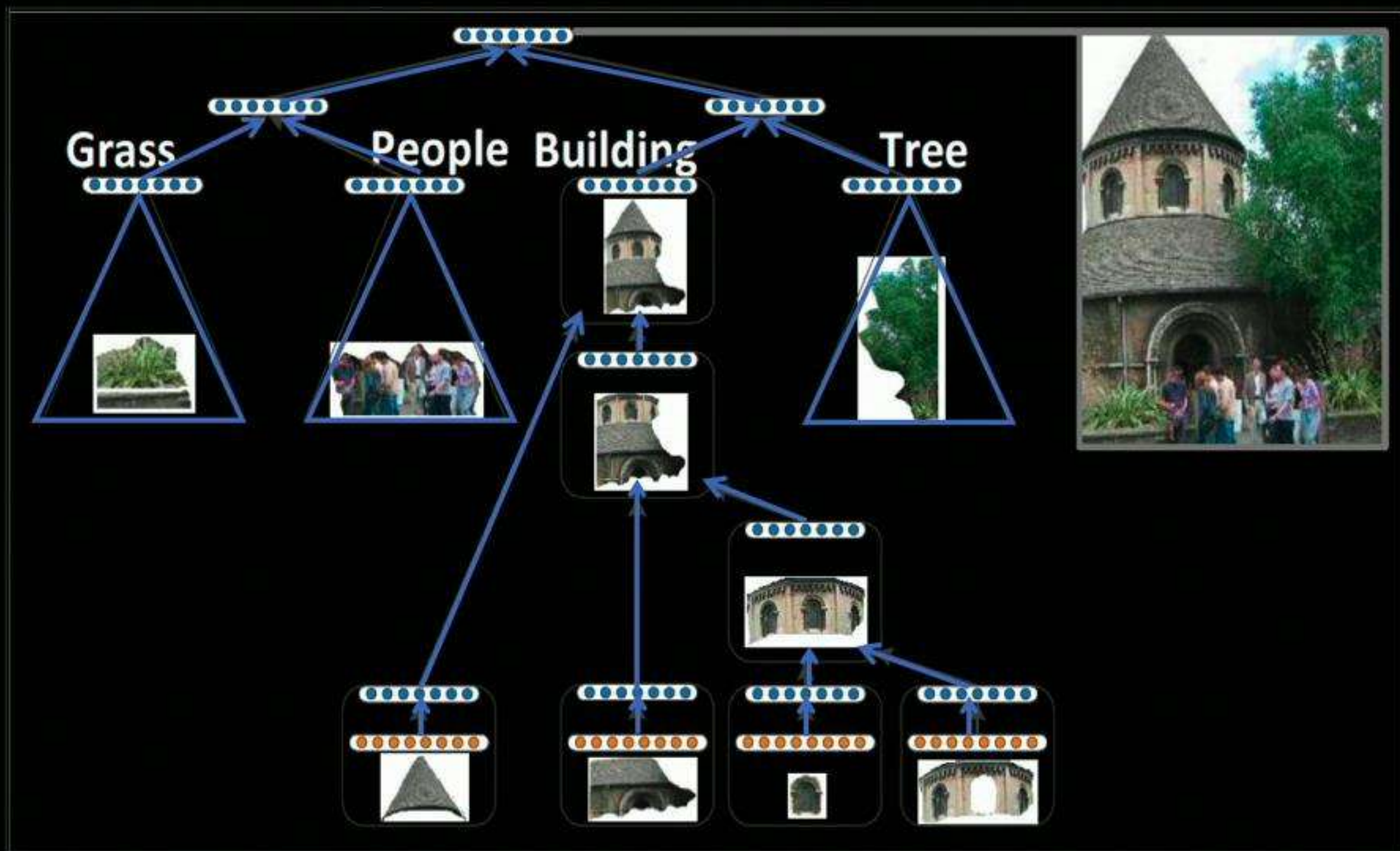
[Tai et al. 2015]

Tree-structured models



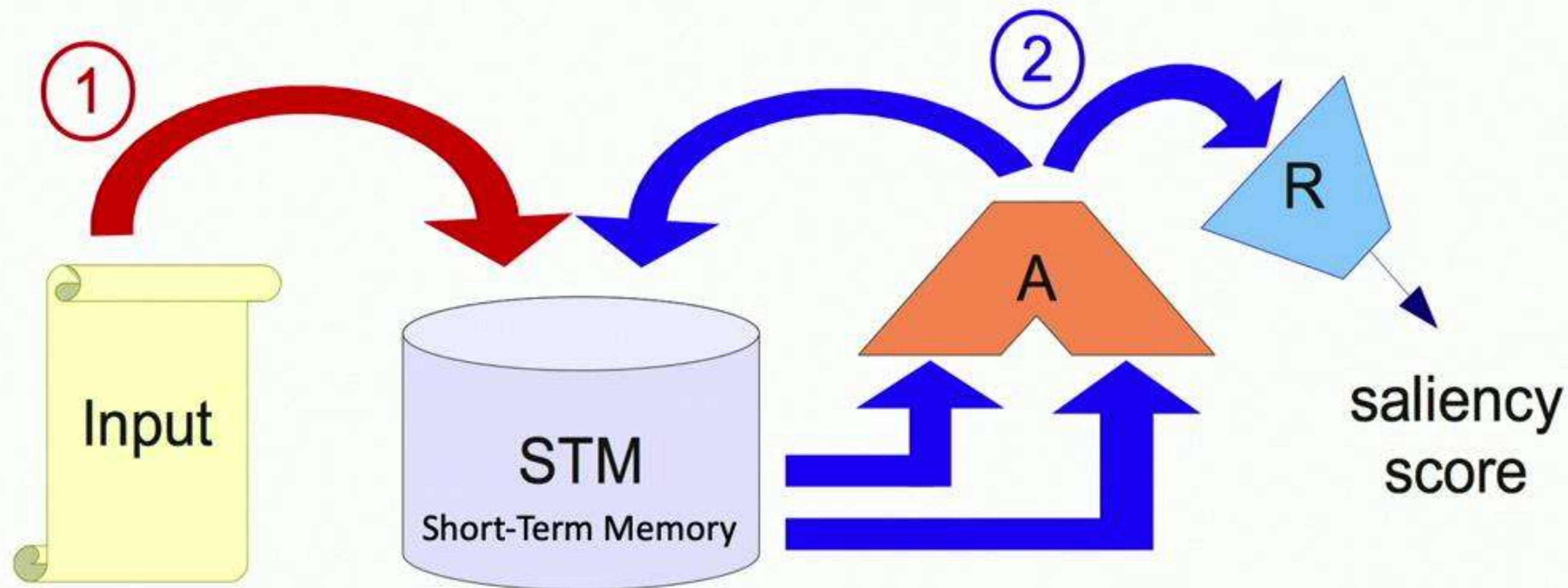
[Socher et al. 2010ff, Tai et al. 2015]

Tree-structured models



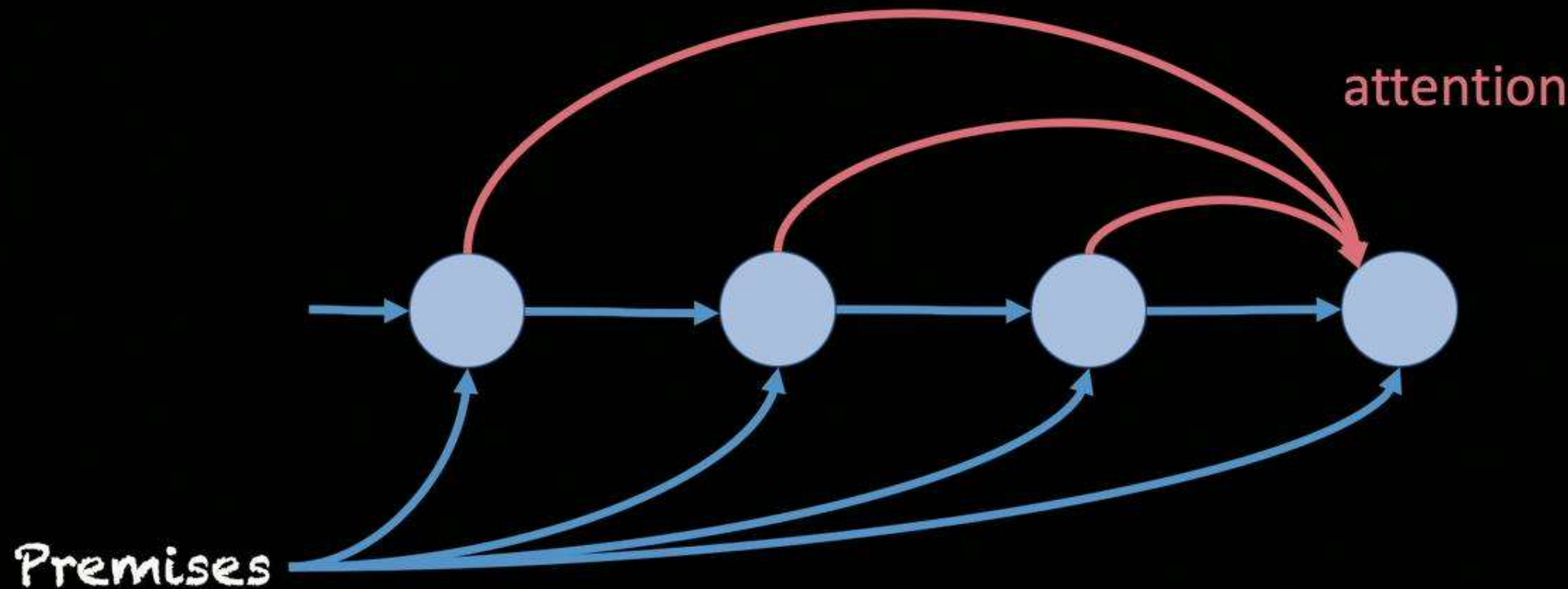
[Socher et al. 2011]

Compositional reasoning tree



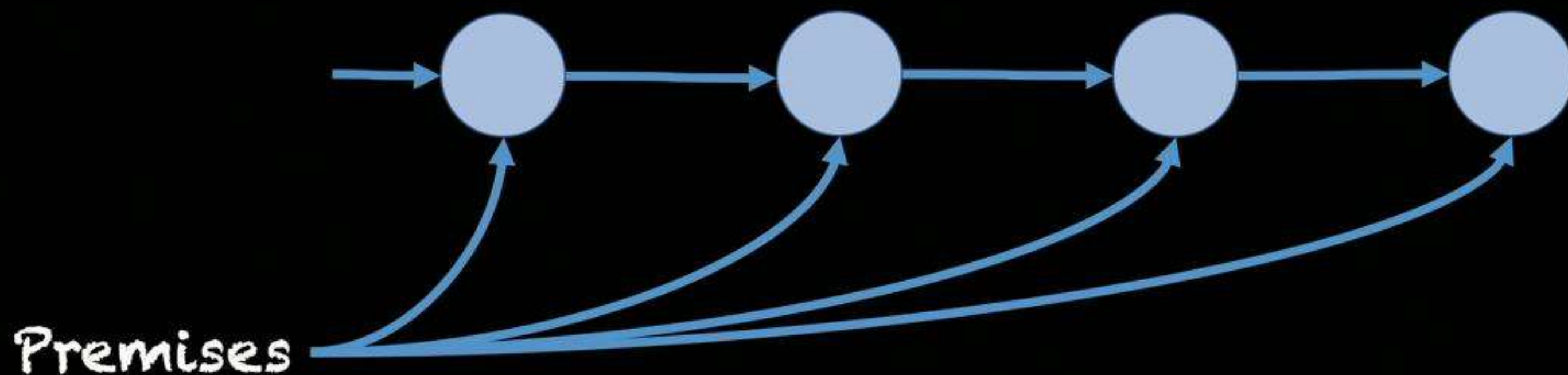
Compositional reasoning without trees

Compositional reasoning without trees



Compositional reasoning without trees

If $f: (X \times Y \times Z) \rightarrow N$, then $\text{curry}(f): X \rightarrow (Y \rightarrow (Z \rightarrow N))$



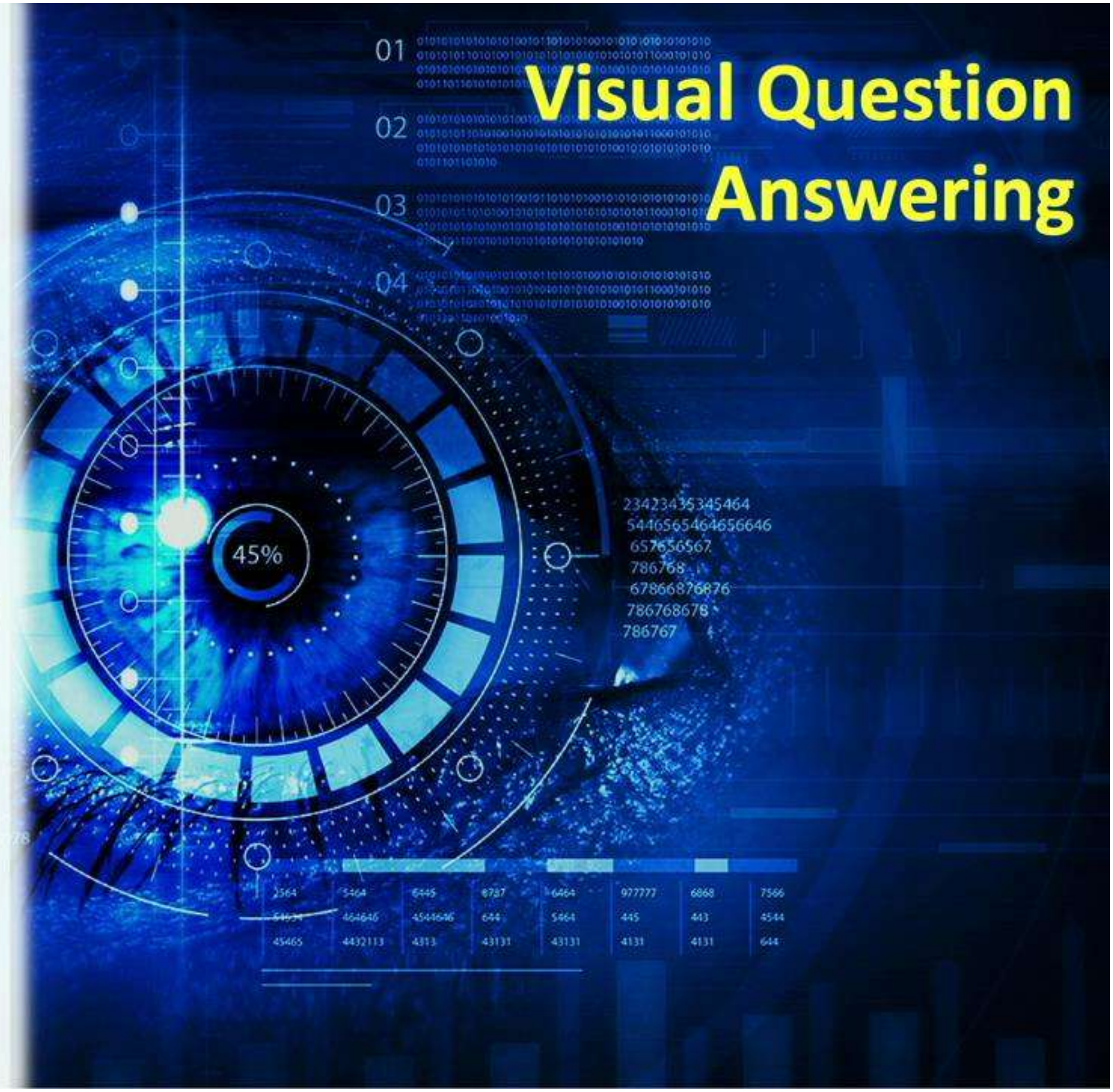
Our Goal

Rather than using standard machine learning **correlation engines**, the goal is improved neural network designs

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

“When a person understands a story, [they] can demonstrate [their] understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”

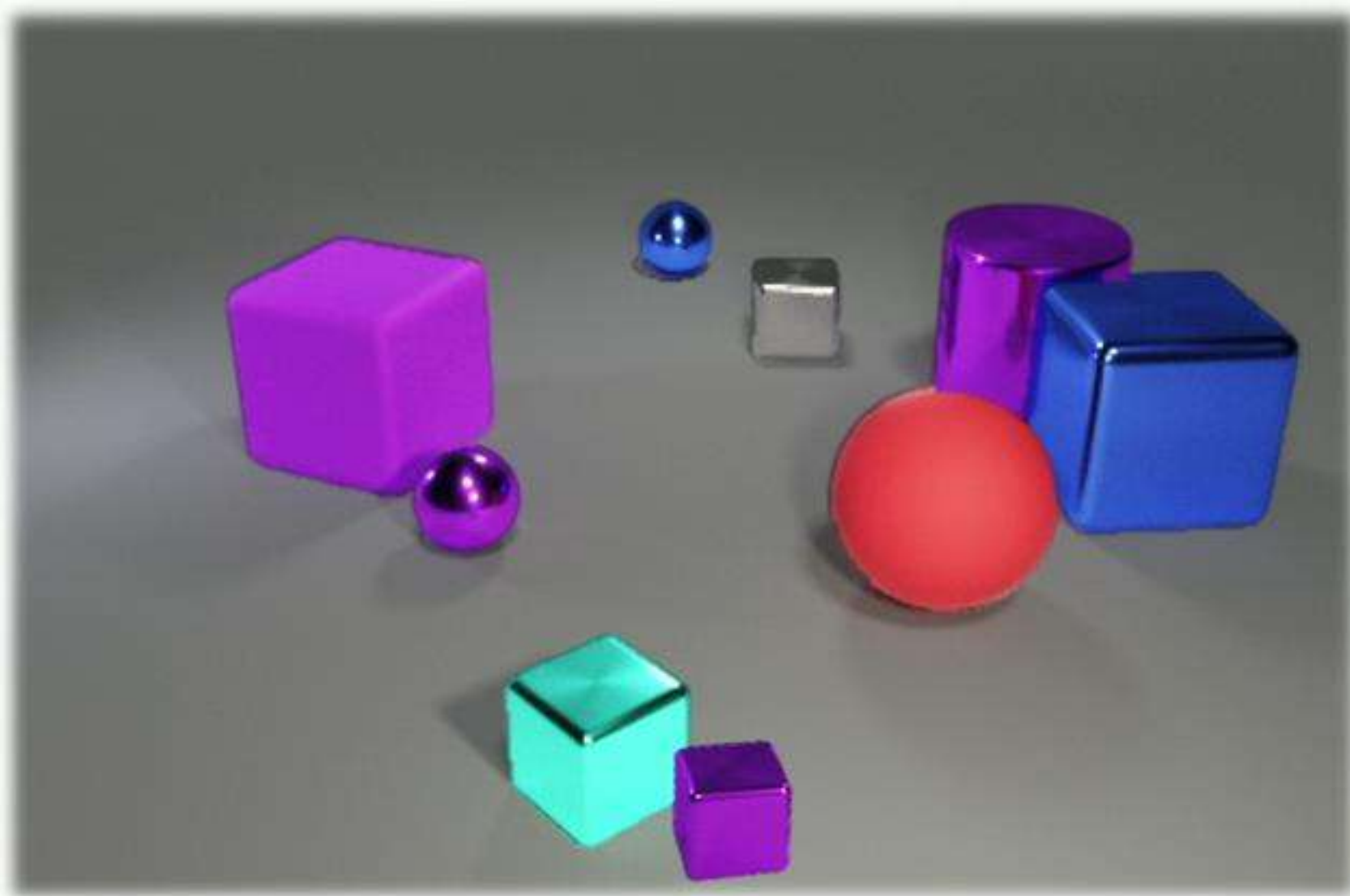
— Wendy Lehnert (PhD, 1977)



Talk Outline

- ✓ From Machine Learning to Machine Reasoning
- **MAC networks on the CLEVR task**
 - The GQA dataset for VQA
 - Neural State Machines for VQA

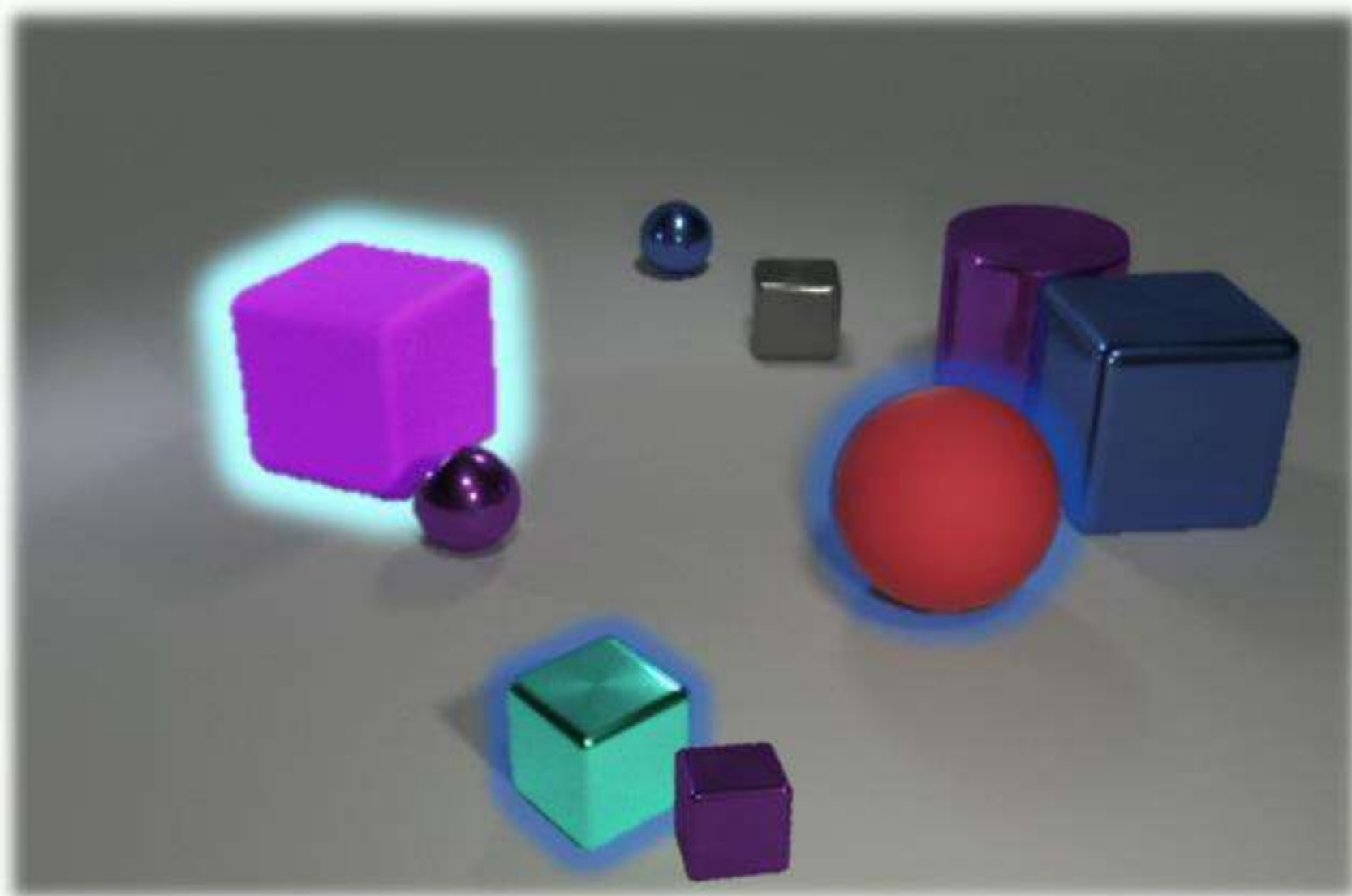
CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning



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

[Johnson et al, CVPR 2017]

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

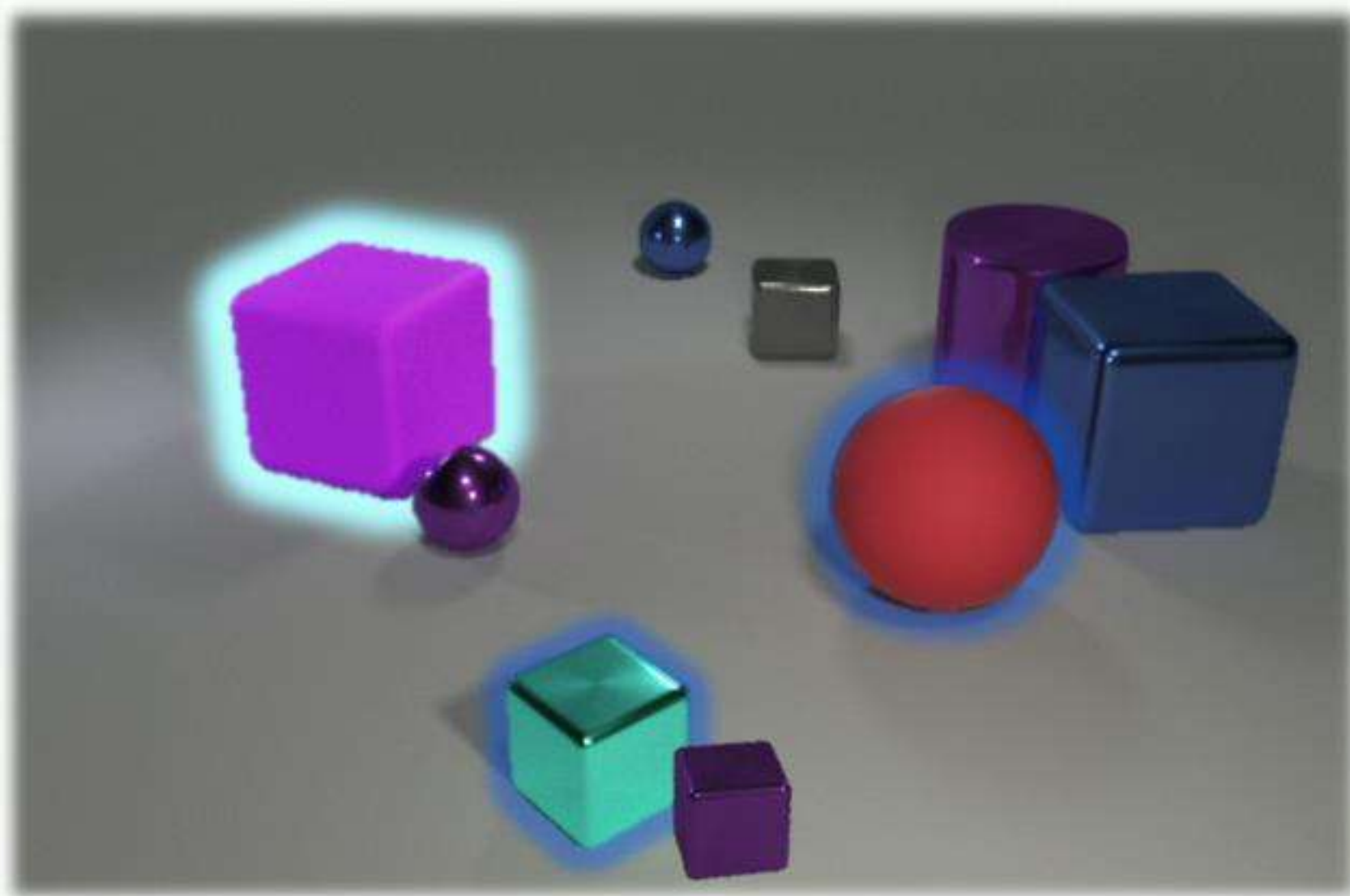


query: material
filter: purple
filter: cube
relate: behind
filter: metal
relate: left
filter: large
filter: ball

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

[Johnson et al,
CVPR 2017]

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



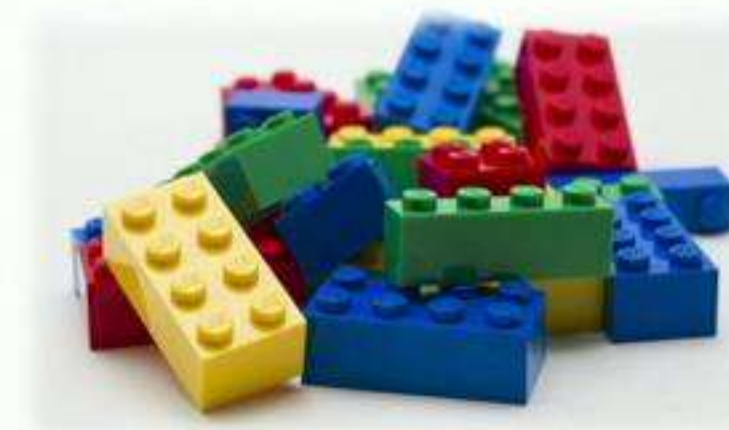
query: material
filter: purple
filter: cube
relate: behind
filter: metal
relate: left
filter: large
filter: ball

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

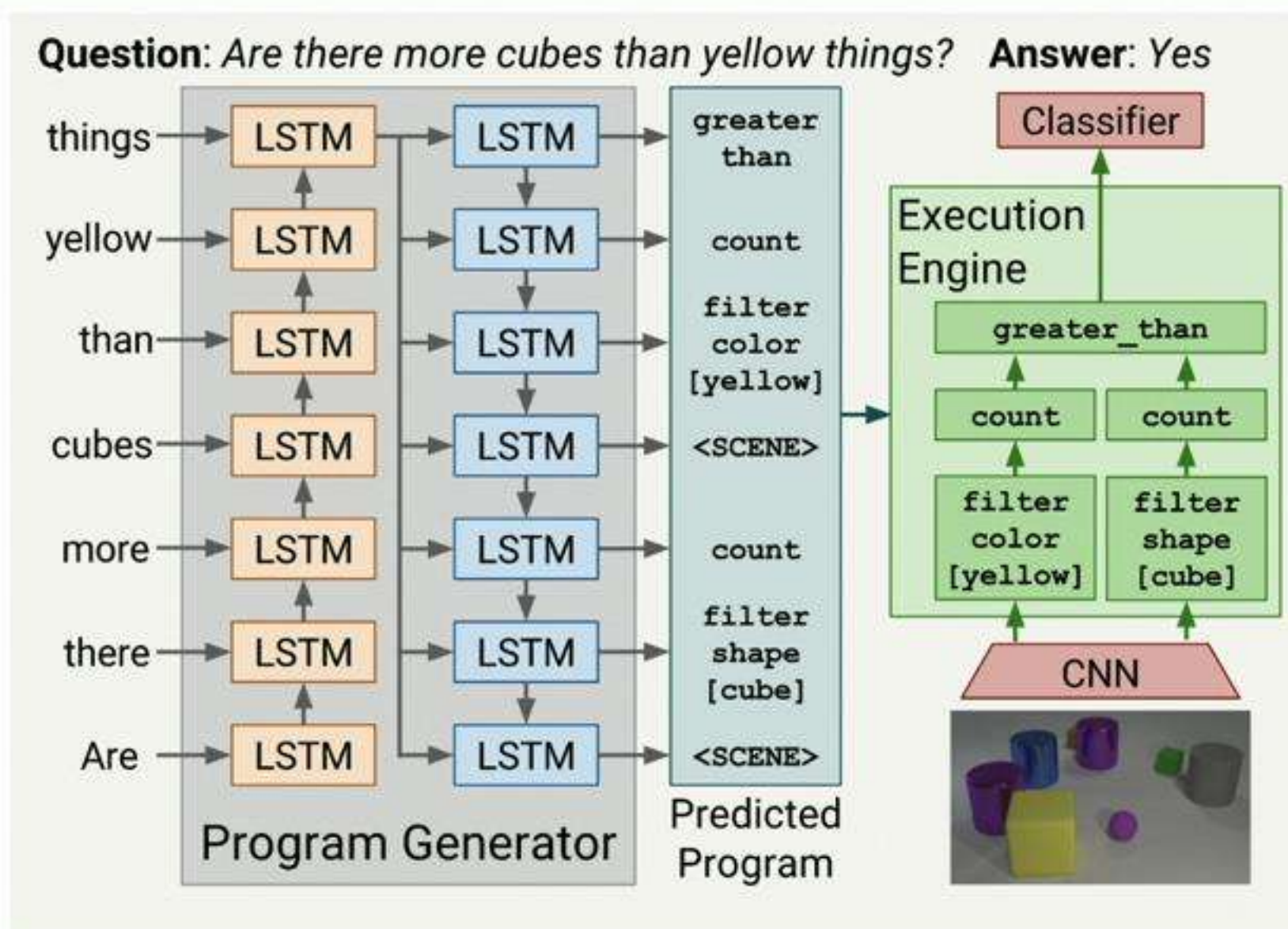
[Johnson et al,
CVPR 2017]

One Existing Approach...

Neural Module Networks



- **Partially differentiable** models that rely on **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 et al, CVPR 2016; Johnson et al, ICCV 2017]

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



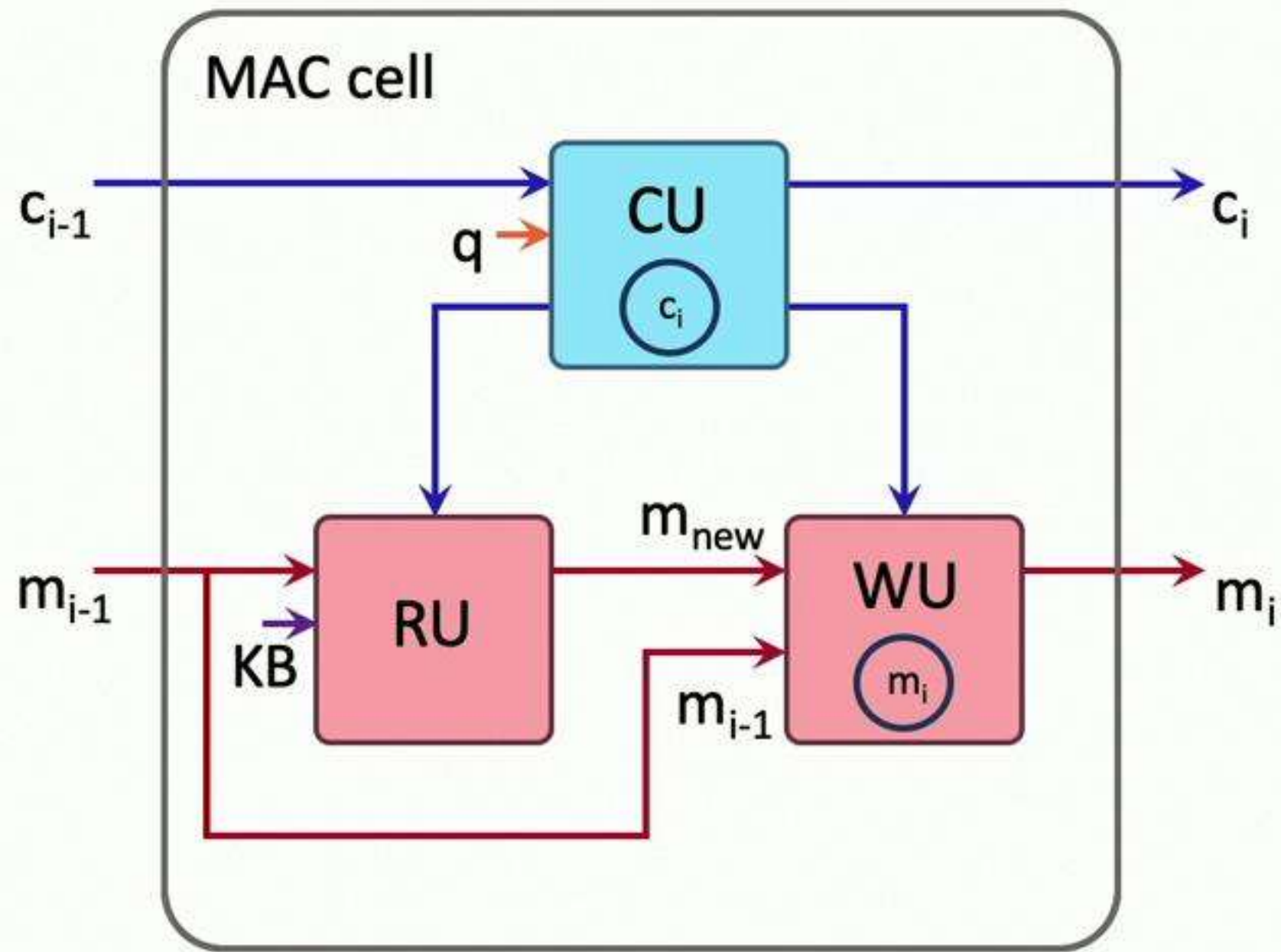
Each **MAC cell** is responsible for performing **one reasoning step at a time**. It maintains **dual recurrent states**:

- **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)



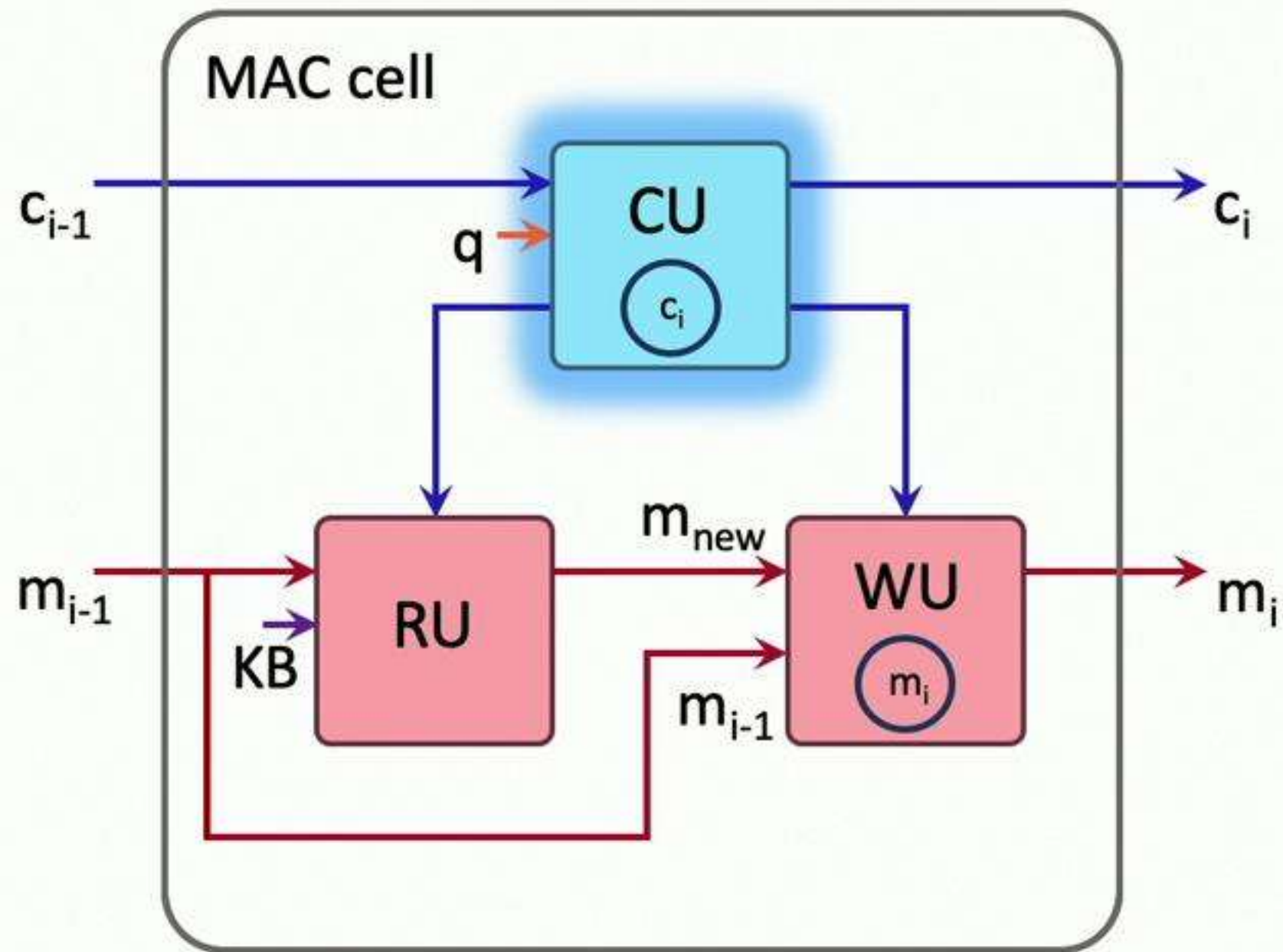
Memory, Attention, Composition.

The MAC cell



Memory, Attention, Composition.

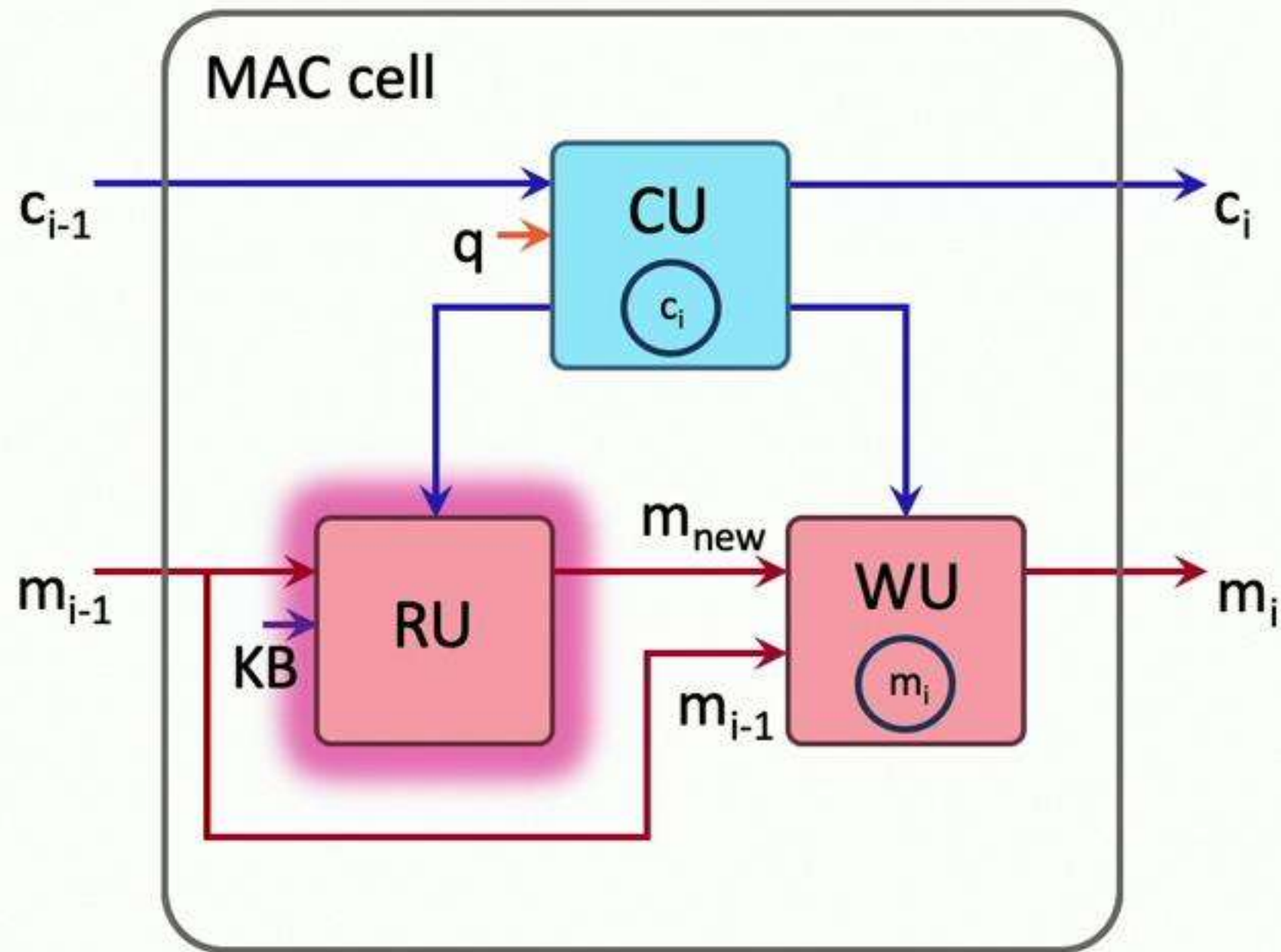
The MAC cell



- **Control Unit (CU) computes a control state, extracting an instruction that focuses on some aspect of the query**

Memory, Attention, Composition.

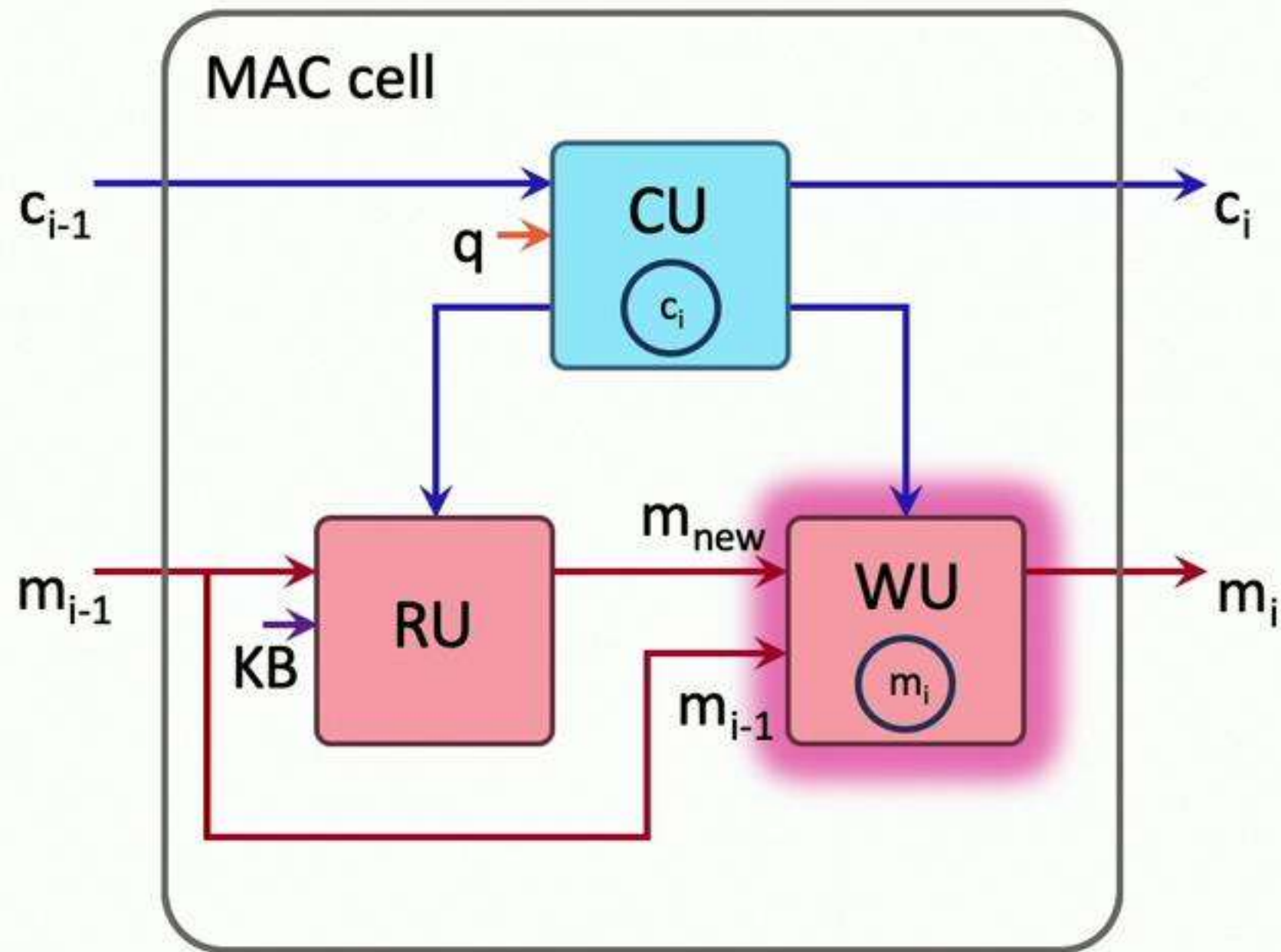
The MAC cell



- **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**

Memory, Attention, Composition.

The MAC cell



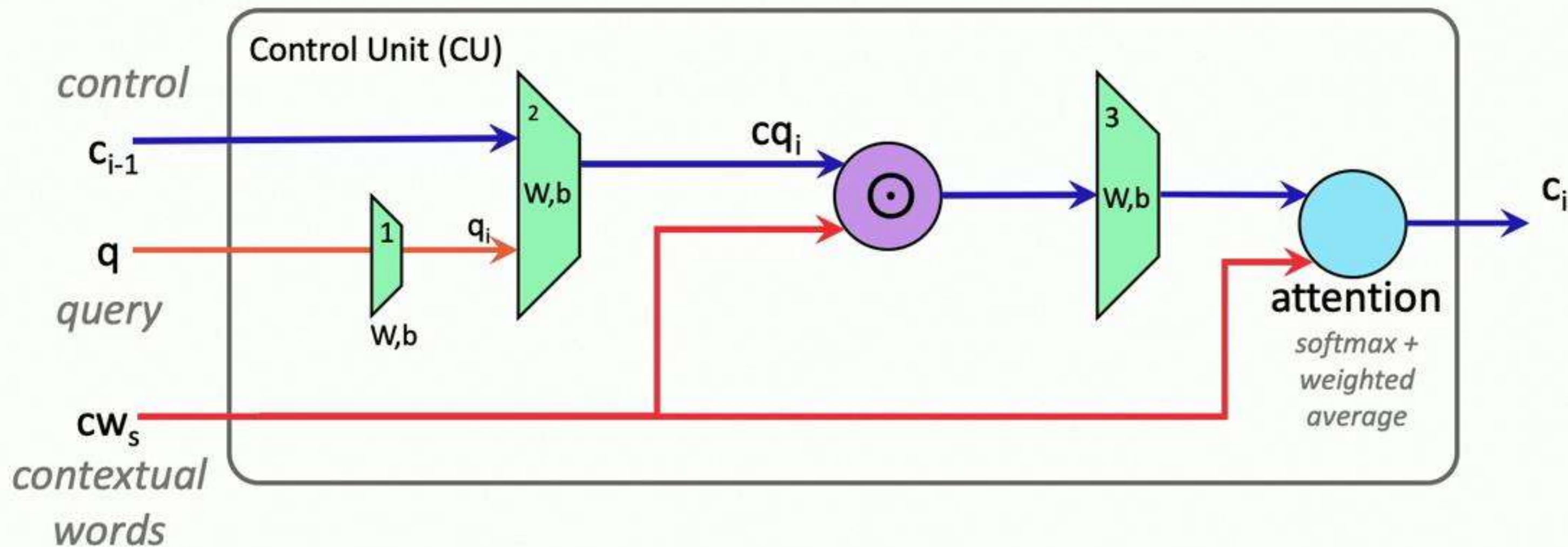
- **Control Unit (CU)** computes a **control** state, extracting an **instruction** that **focuses** on some **aspect of the query**
- **Read Unit (RU)**: **retrieves information** from the **knowledge base** given the **current control** state and **previous memory**
- **Write Unit (WU)**: **updates**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

The MAC cell

The Control Unit (CU)



Extract an instruction (control) from the question

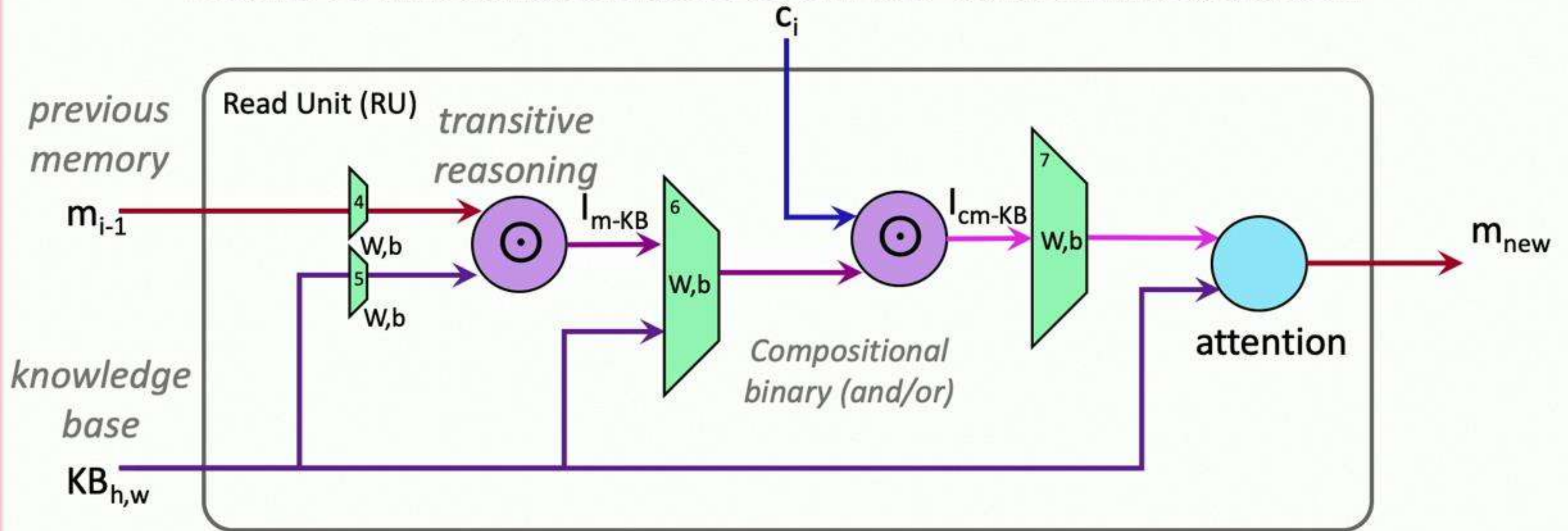


The MAC cell

The Read Unit (RU)



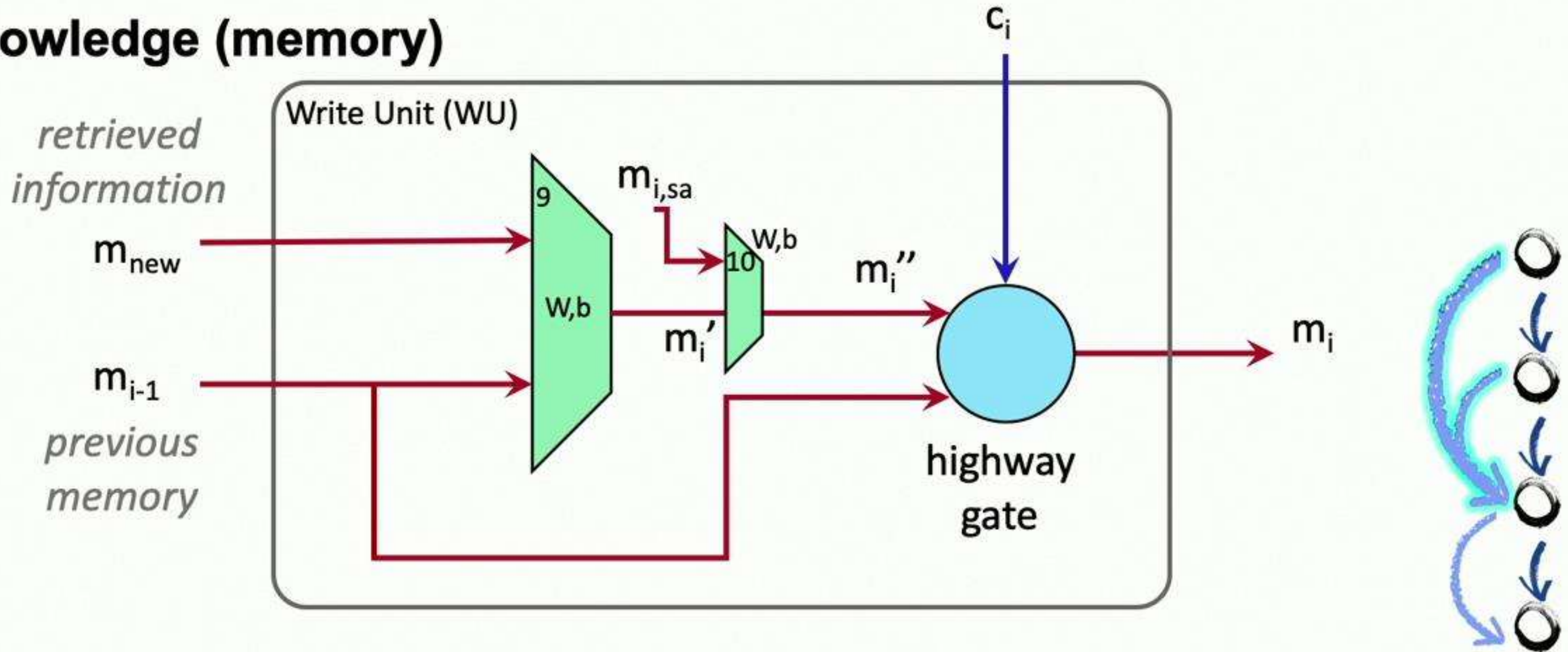
Retrieve information based on the current instruction



The MAC cell

The Write Unit (WU)

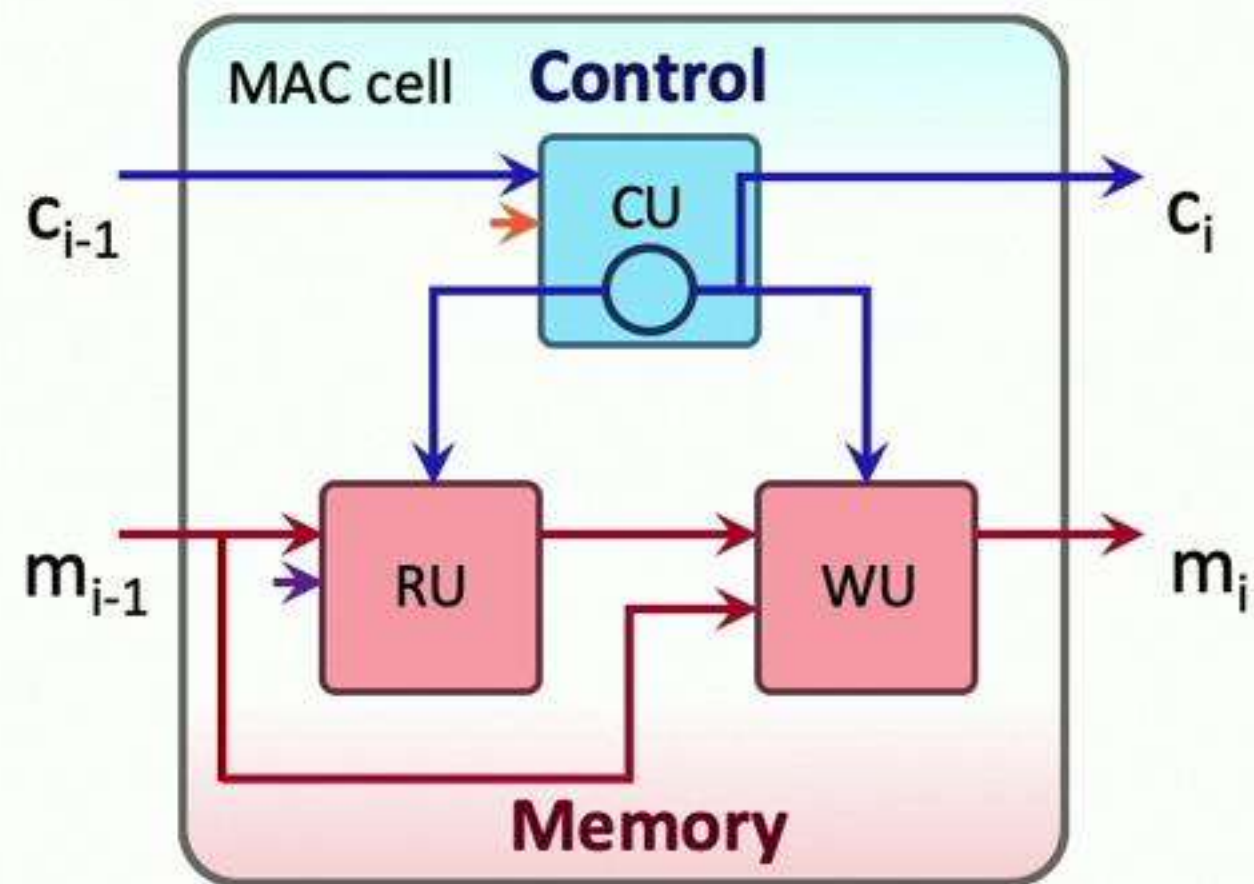
Combine retrieved information with accumulated knowledge (memory)



The MAC net

From Cell to Network

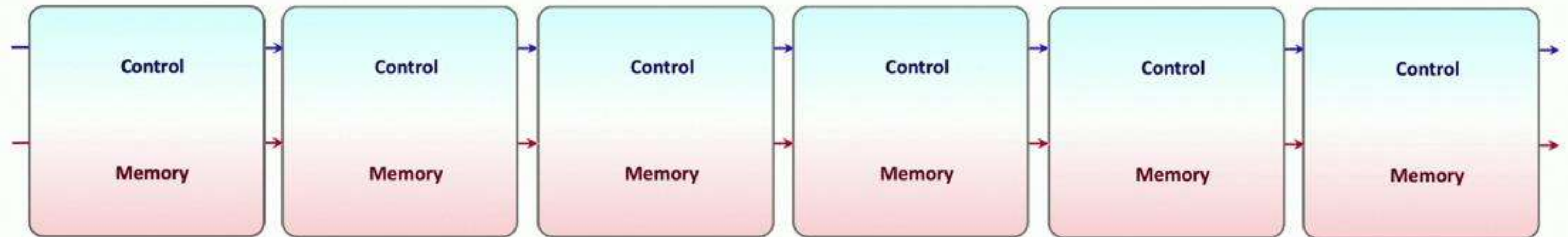
A MacNet is a **soft-attention sequence** of p MAC cells



The MAC net

From Cell to Network

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

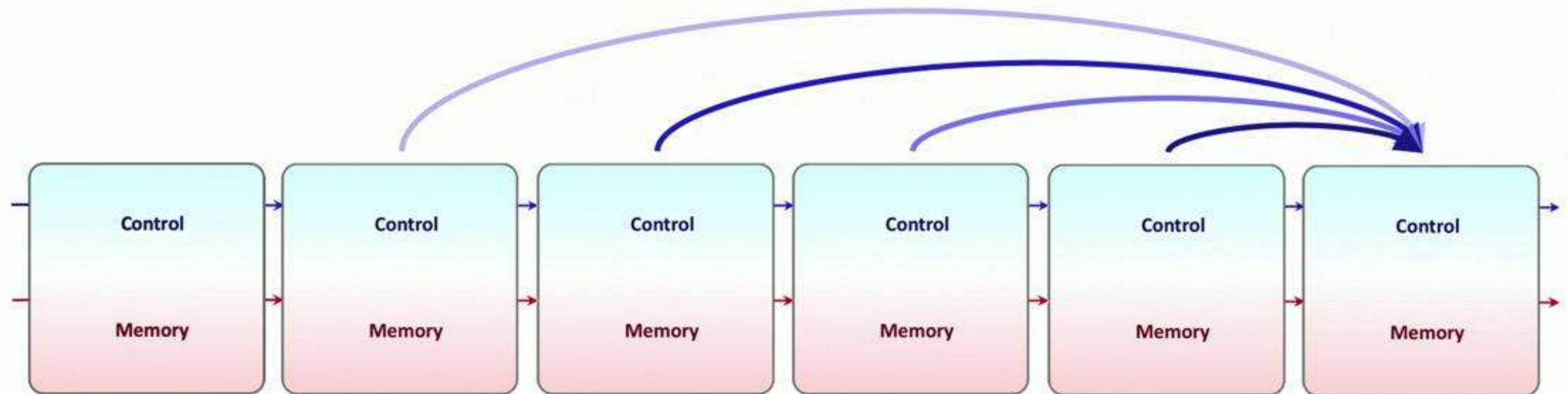


Uniform **sequential structure** for all queries;
efficient, easy to deploy, and fully differentiable

The MAC net

From Cell to Network

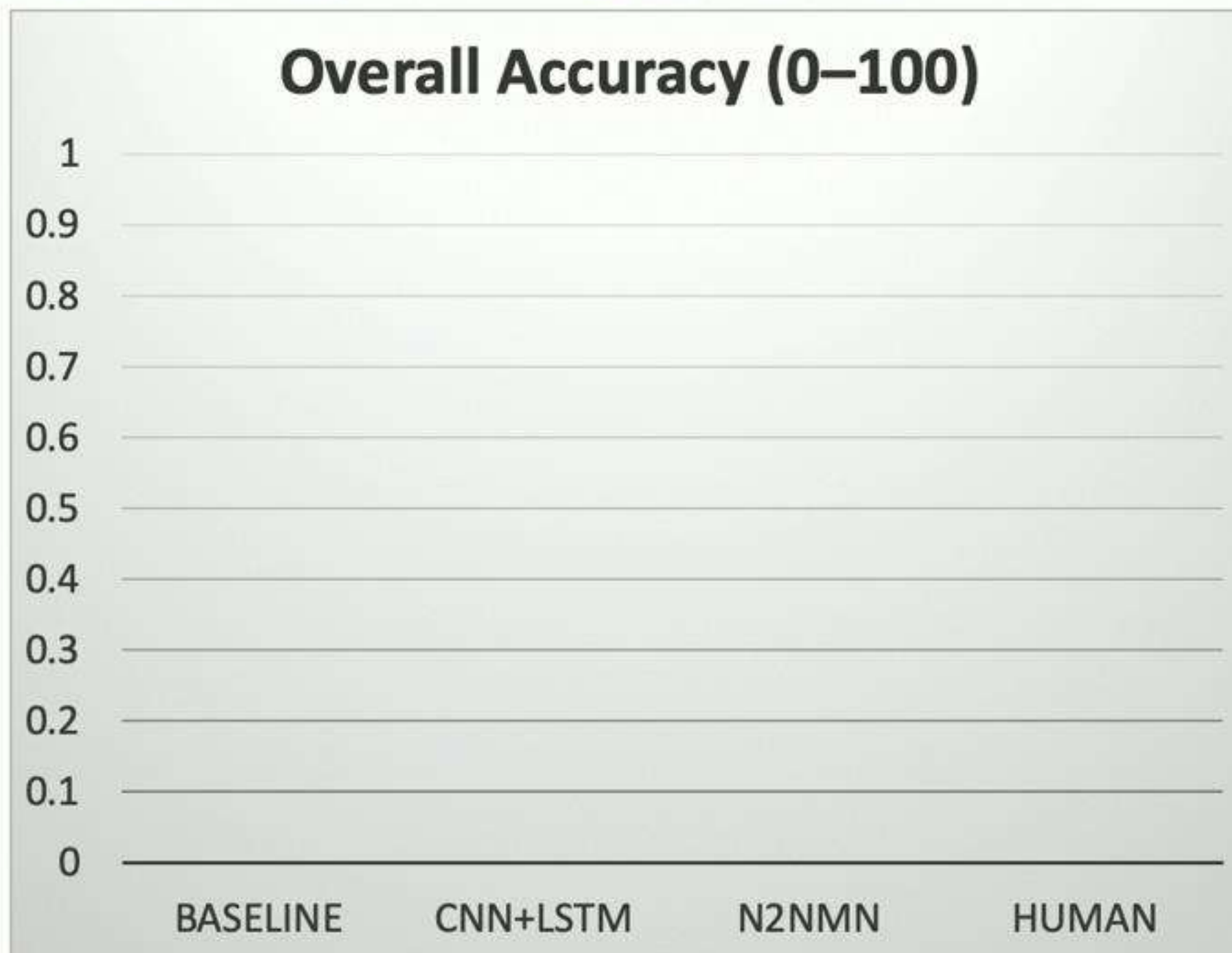
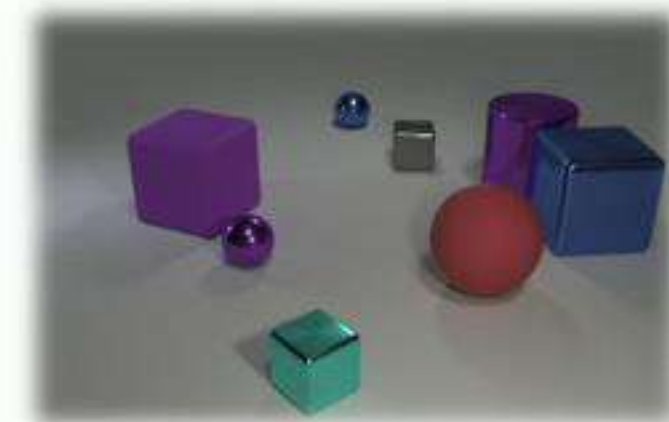
A MacNet is a **soft-attention sequence** of p MAC cells



A **capacity** to represent arbitrarily complex reasoning **Directed Acyclic Graphs** (DAGs)

Experiments

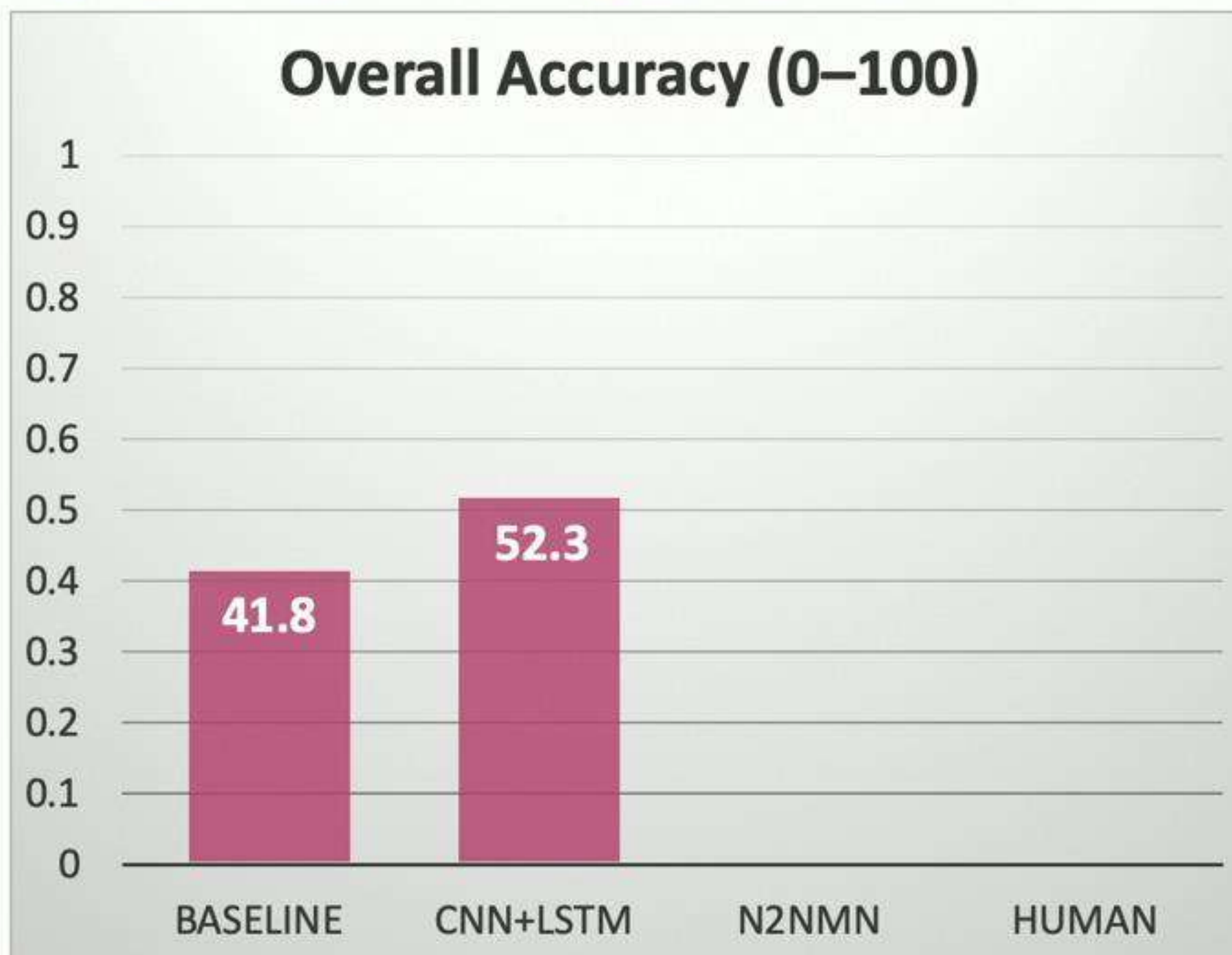
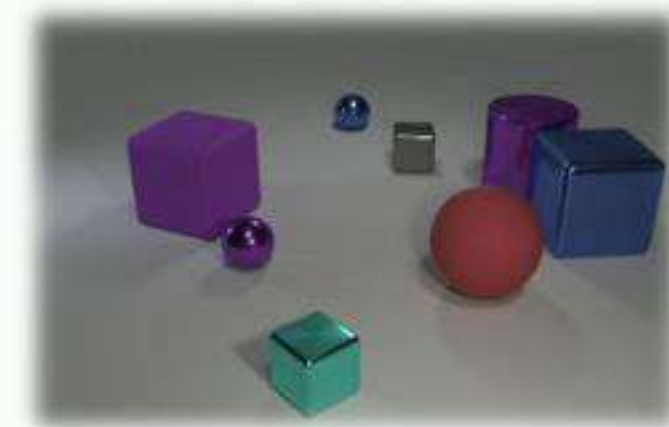
CLEVR Overall Results



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

Experiments

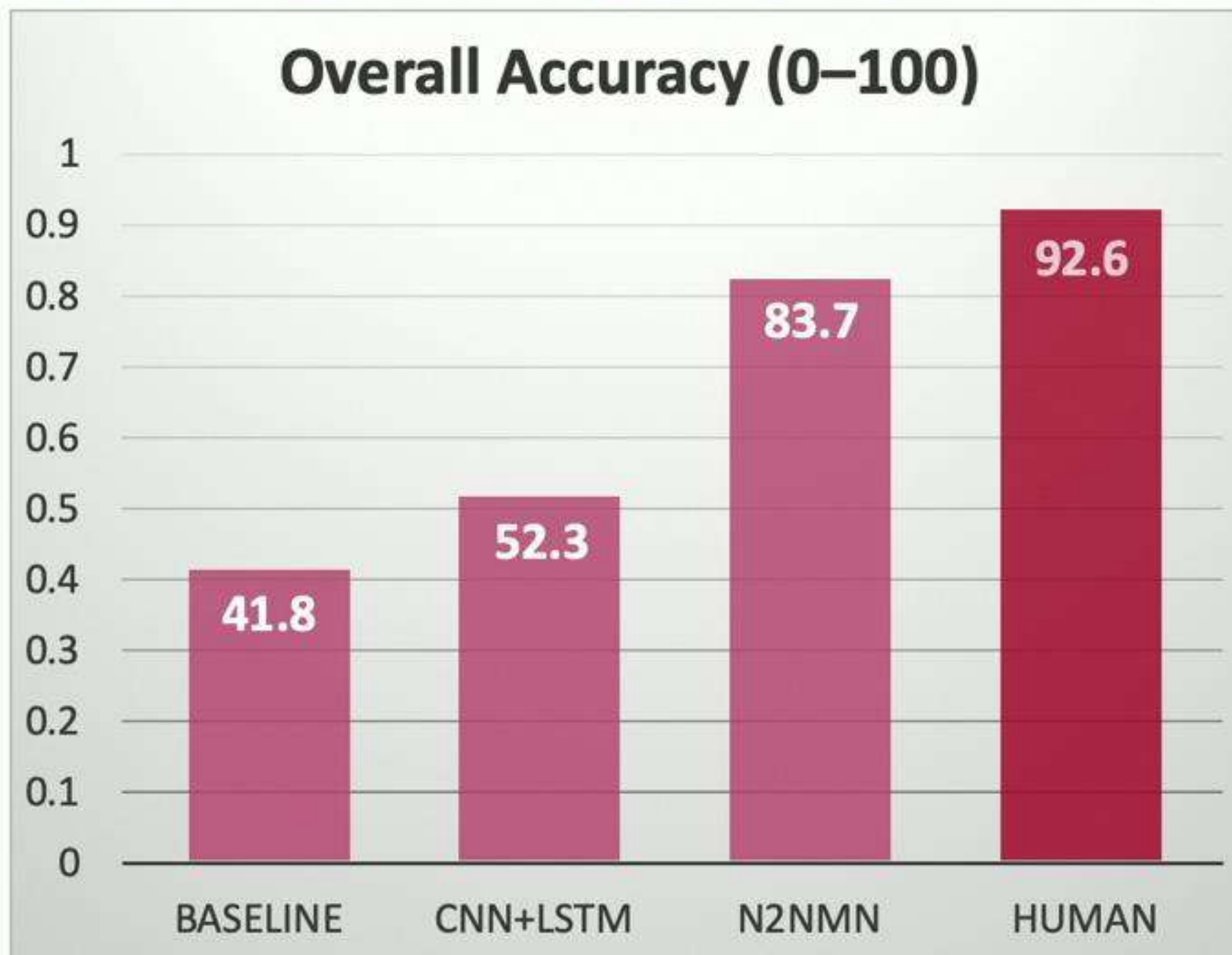
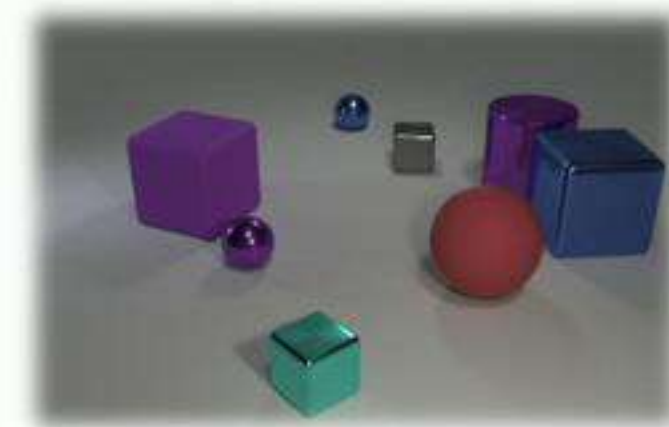
CLEVR Overall Results



- 700k Training set
- 150k Test set
- 28 candidate answers
- **Baseline:** the most frequent answer for each question type

Experiments

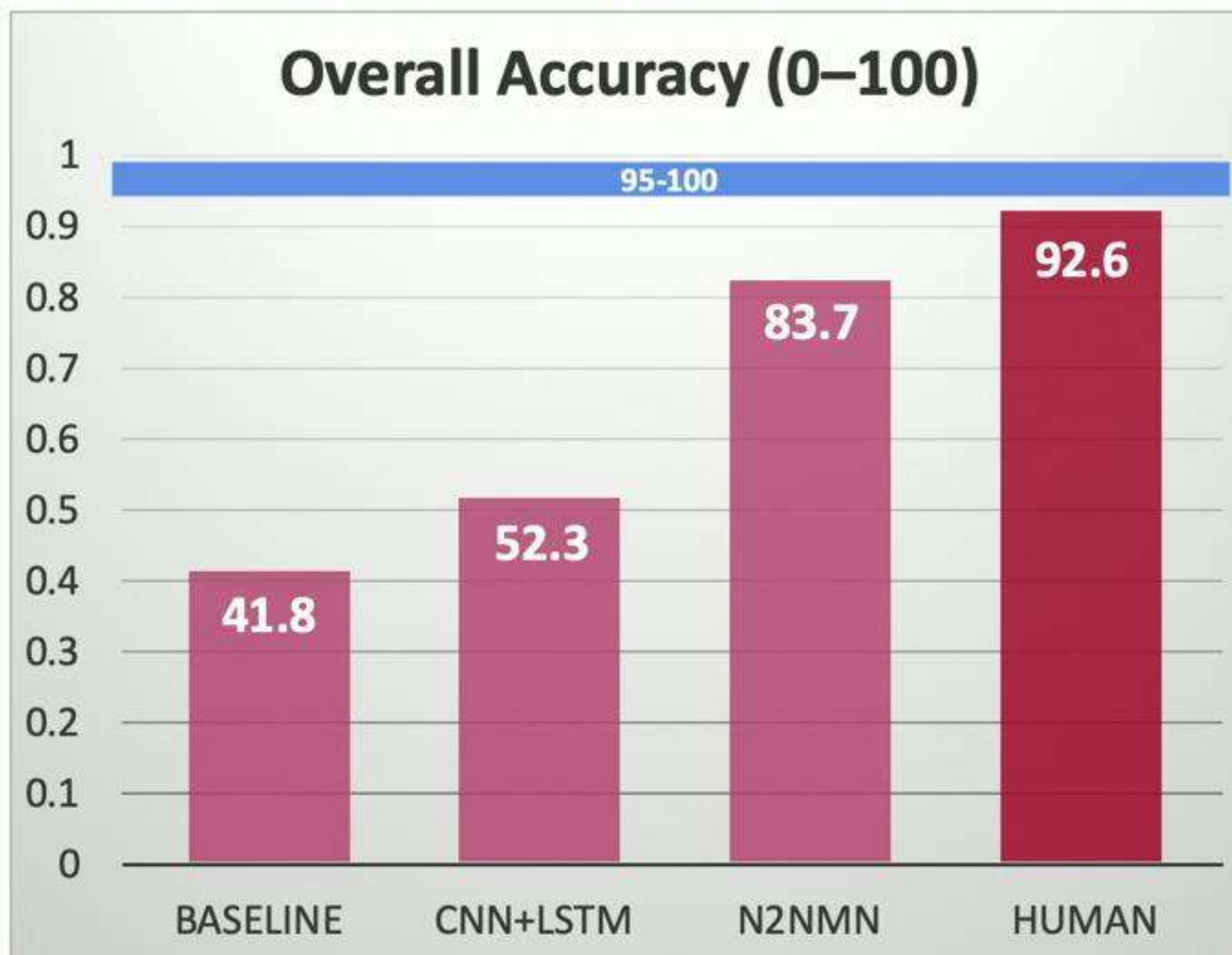
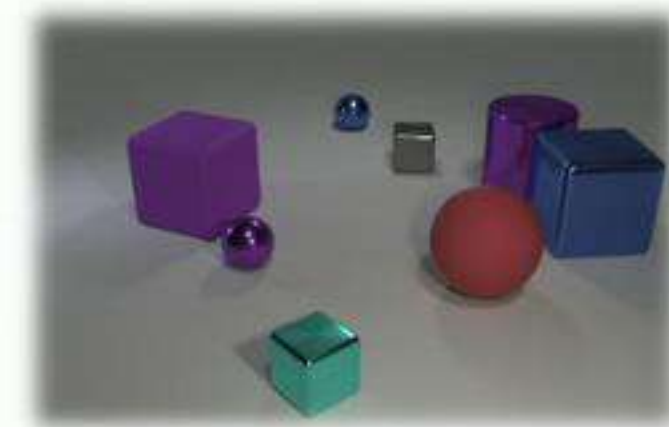
CLEVR Overall Results



- 700k Training set
- 150k Test set
- 28 candidate answers
- **Baseline:** the most frequent answer for each question type

Experiments

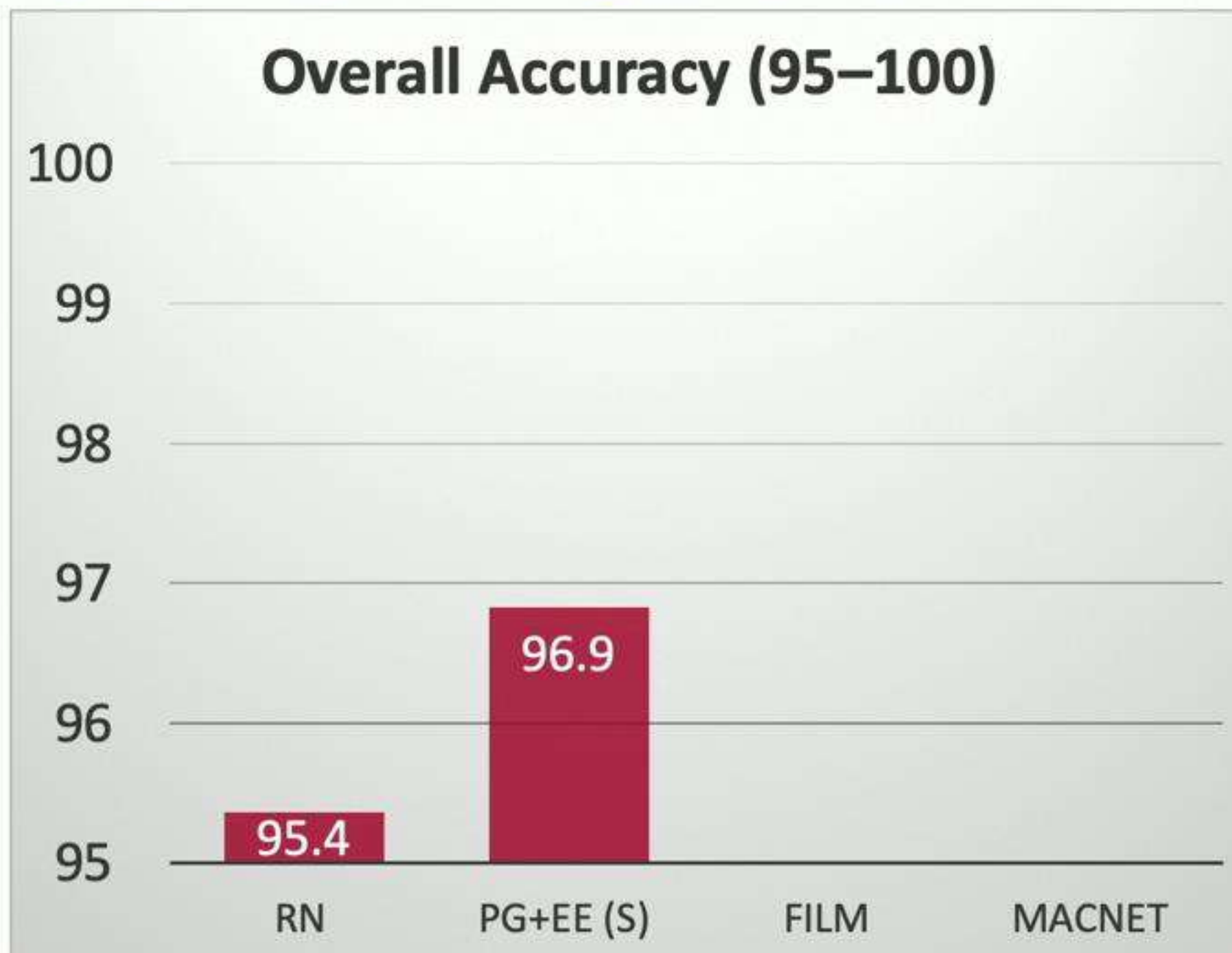
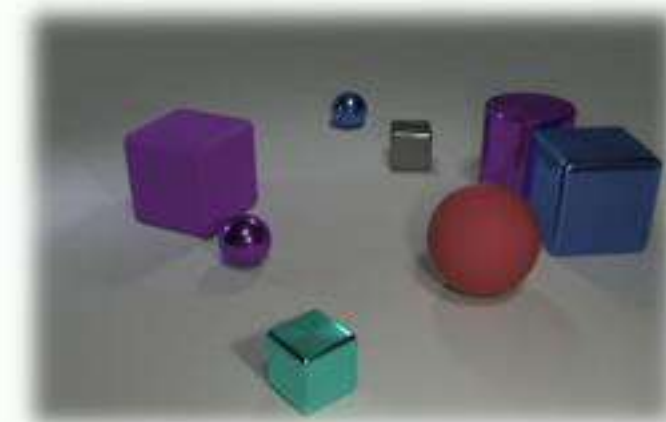
CLEVR Overall Results



- 700k Training set
- 150k Test set
- 28 candidate answers
- **Baseline:** the most frequent answer for each question type

Experiments

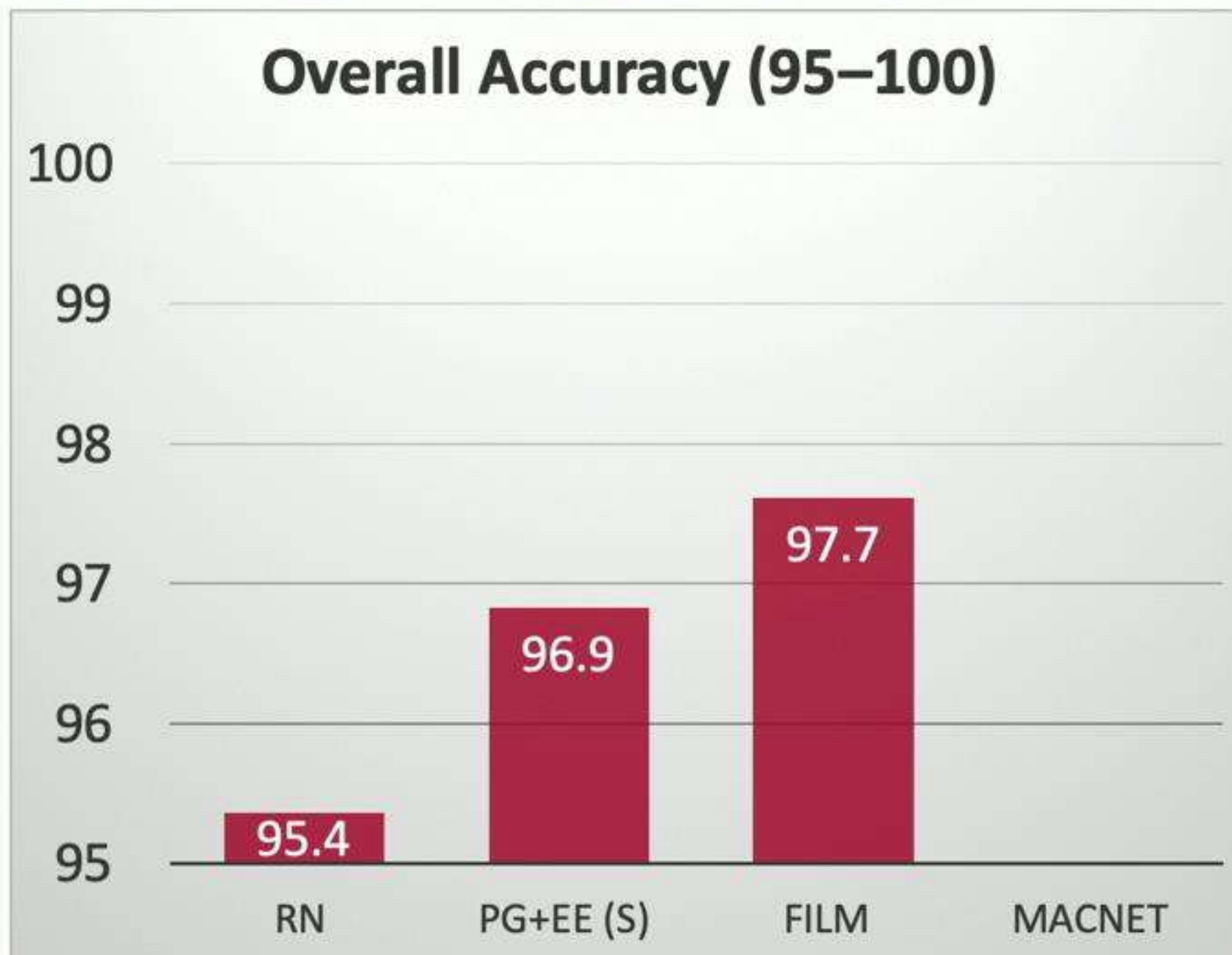
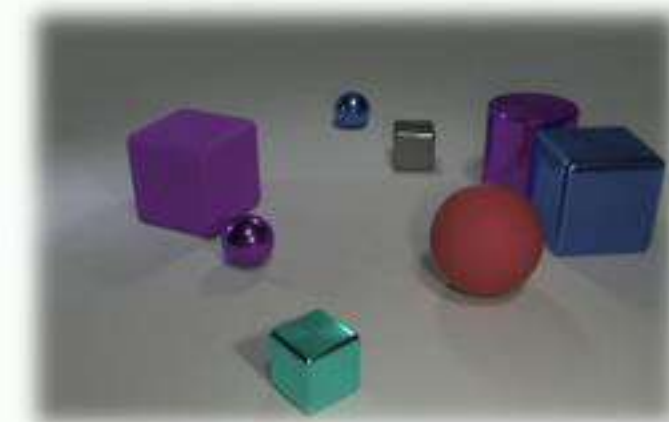
CLEVR Overall Results



■ **(S)**: strongly supervised

Experiments

CLEVR Overall Results



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

Existing Approaches

Relation Nets and FiLM

Large CNN stacks interleaved with
specialized layers

RN [Santoro et al, 2017]

FiLM [Perez et al, 2017]

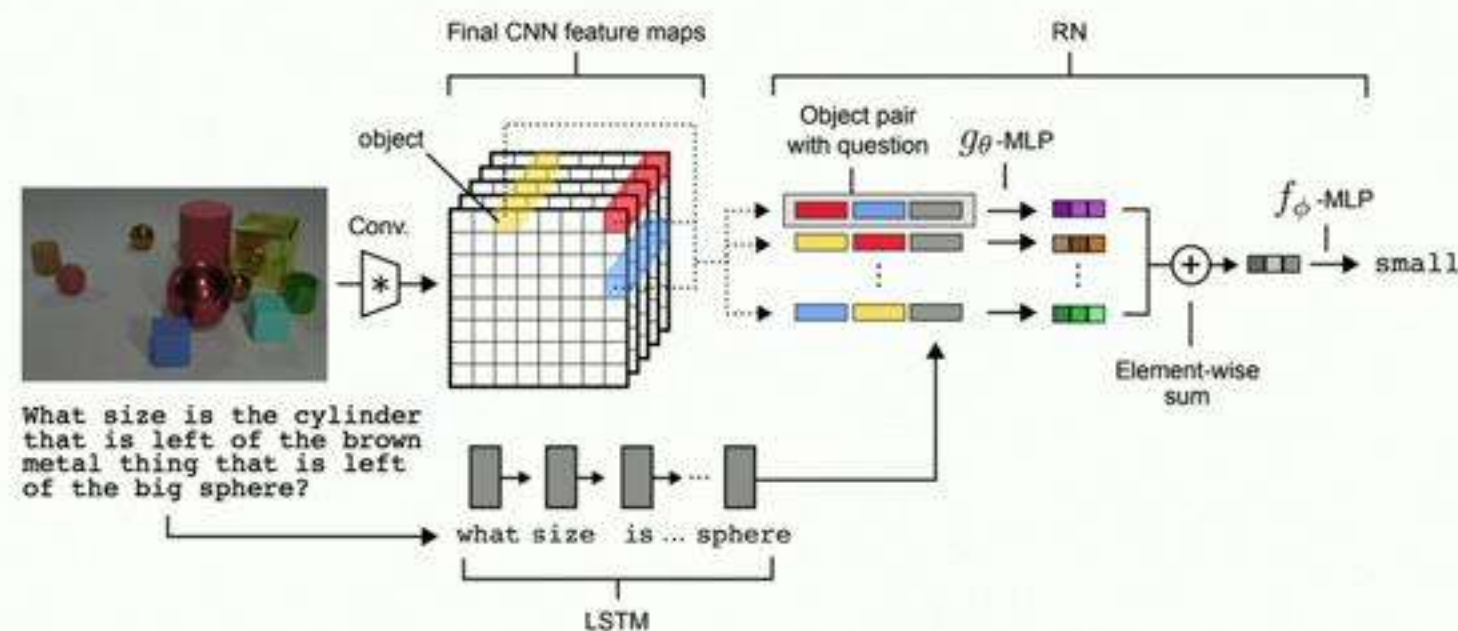
Existing Approaches

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



RN [Santoro et al, 2017]

FiLM [Perez et al, 2017]

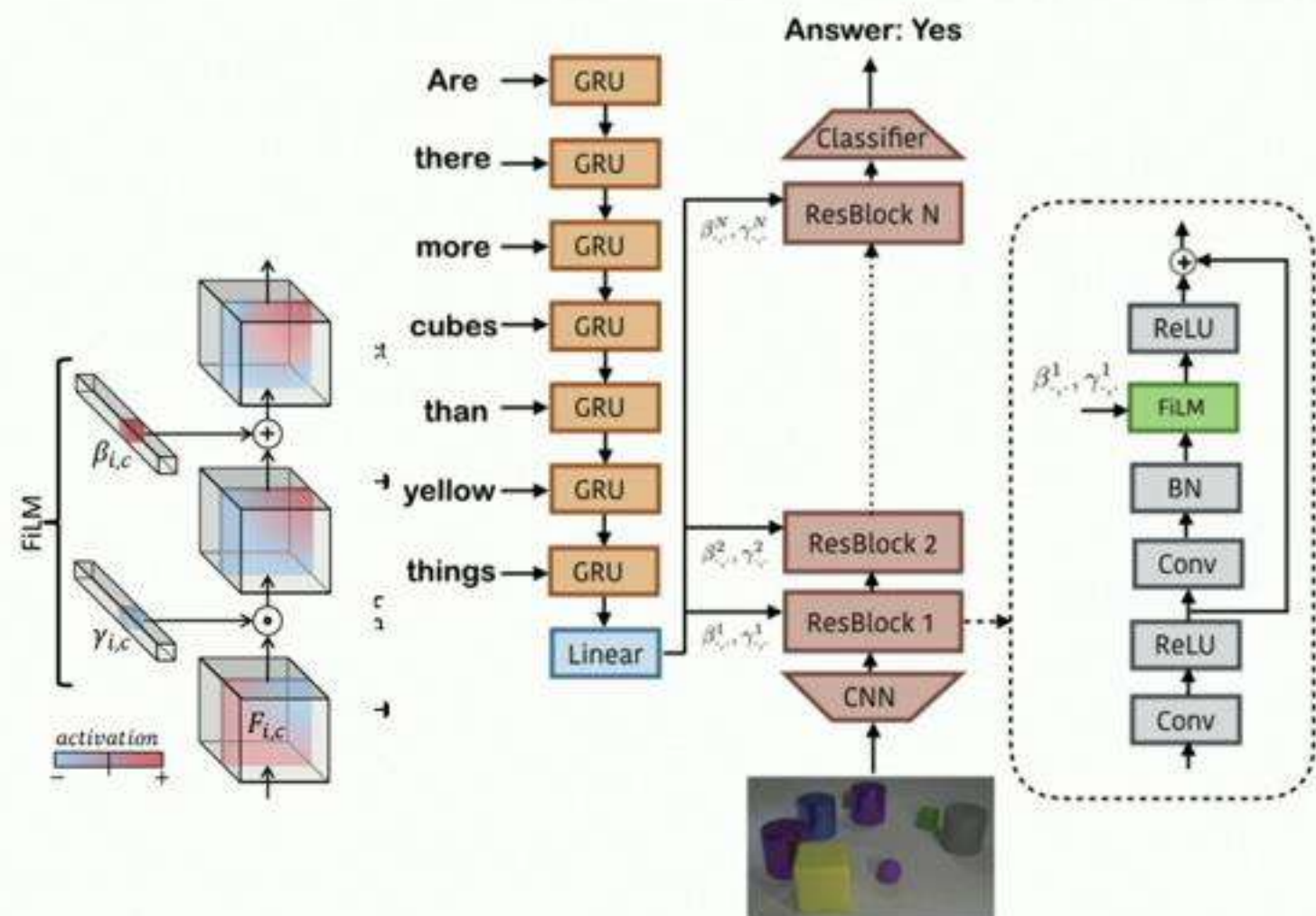
Existing Approaches

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

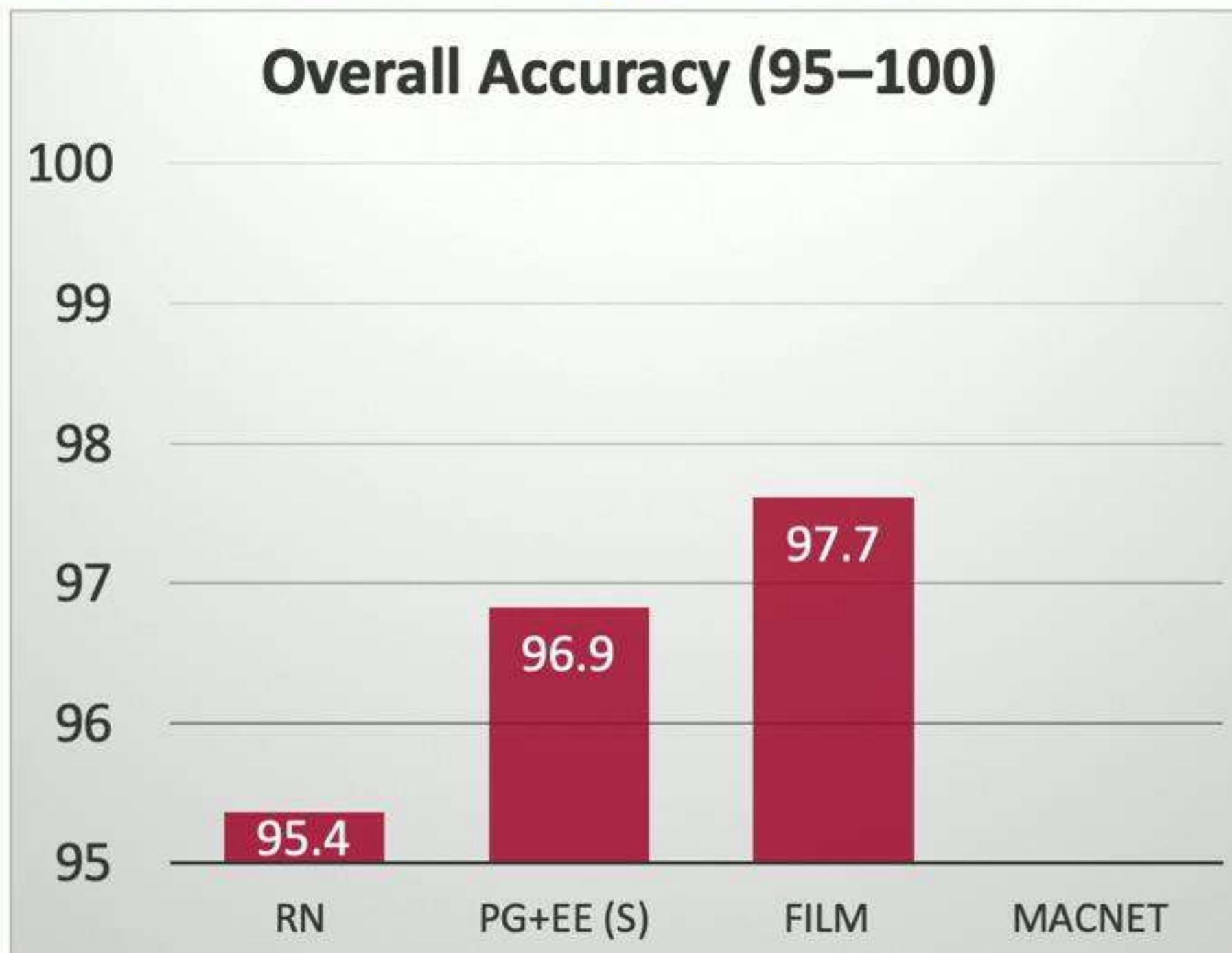
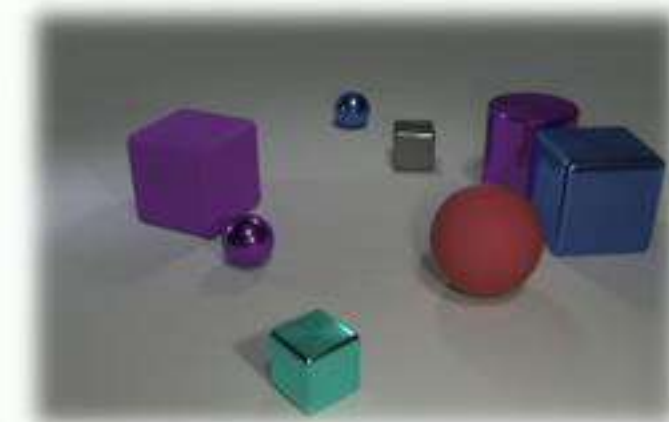


RN [Santoro et al, 2017]

FiLM [Perez et al, 2017]

Experiments

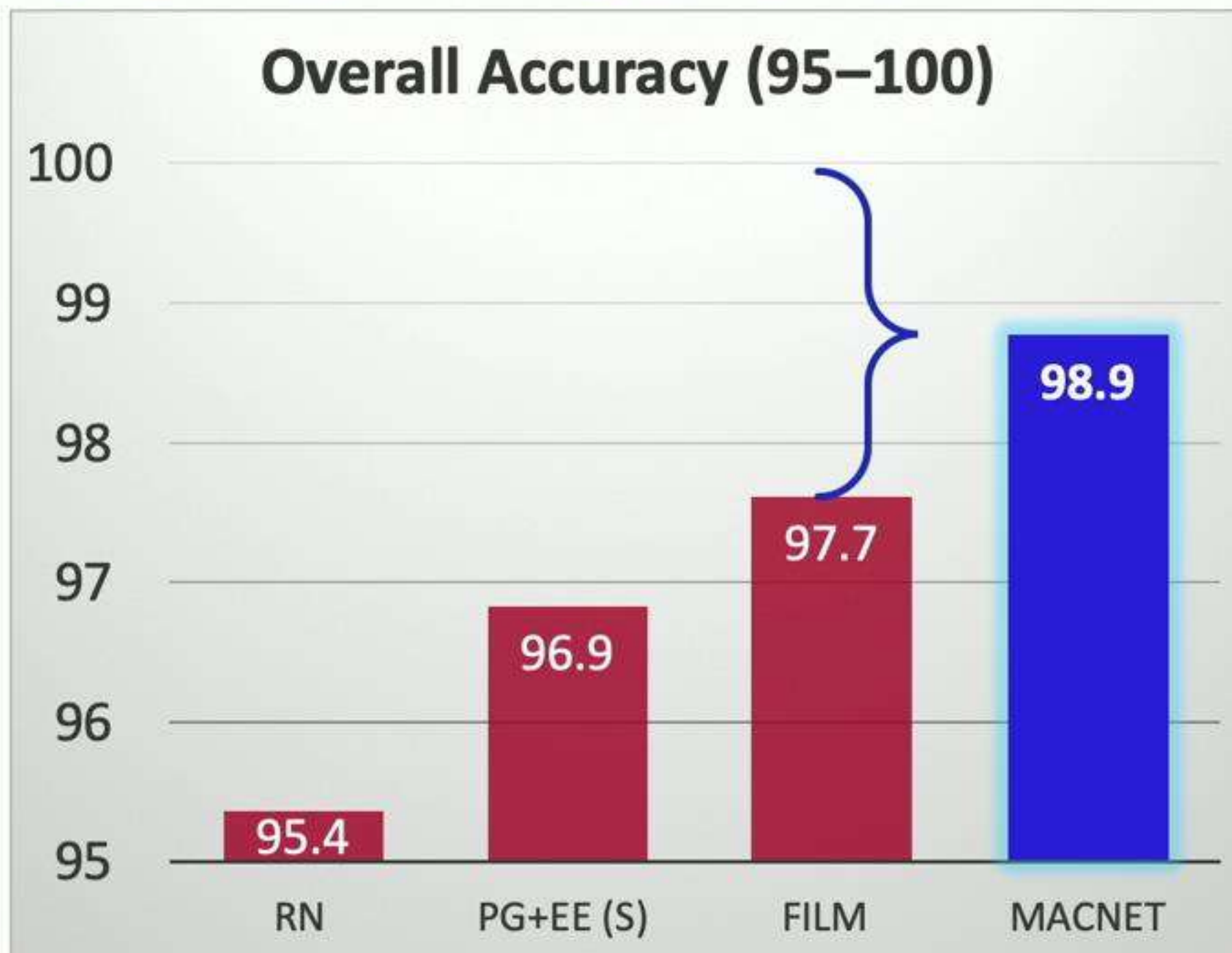
CLEVR Overall Results



■ (S): strongly supervised

Experiments

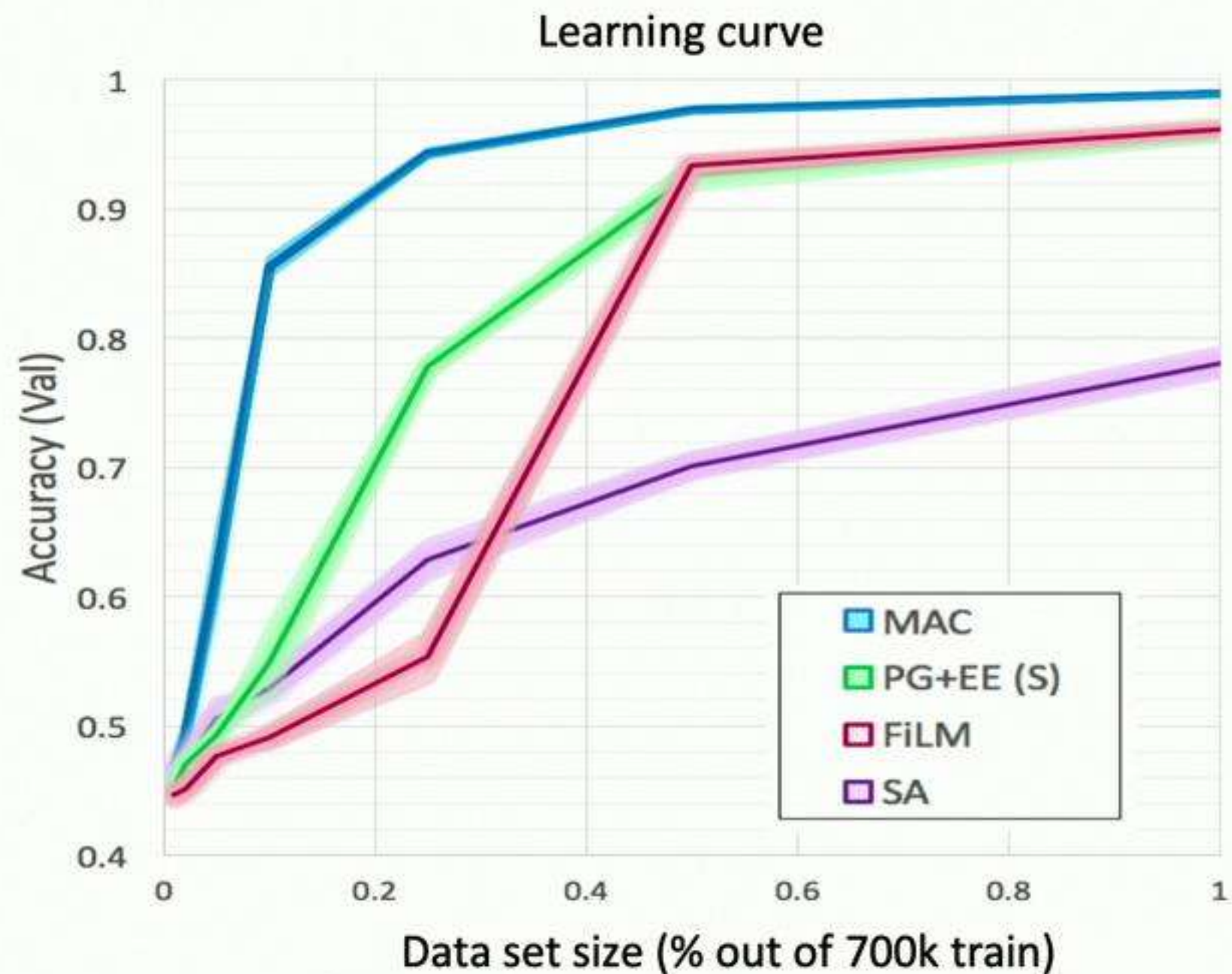
CLEVR Overall Results



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

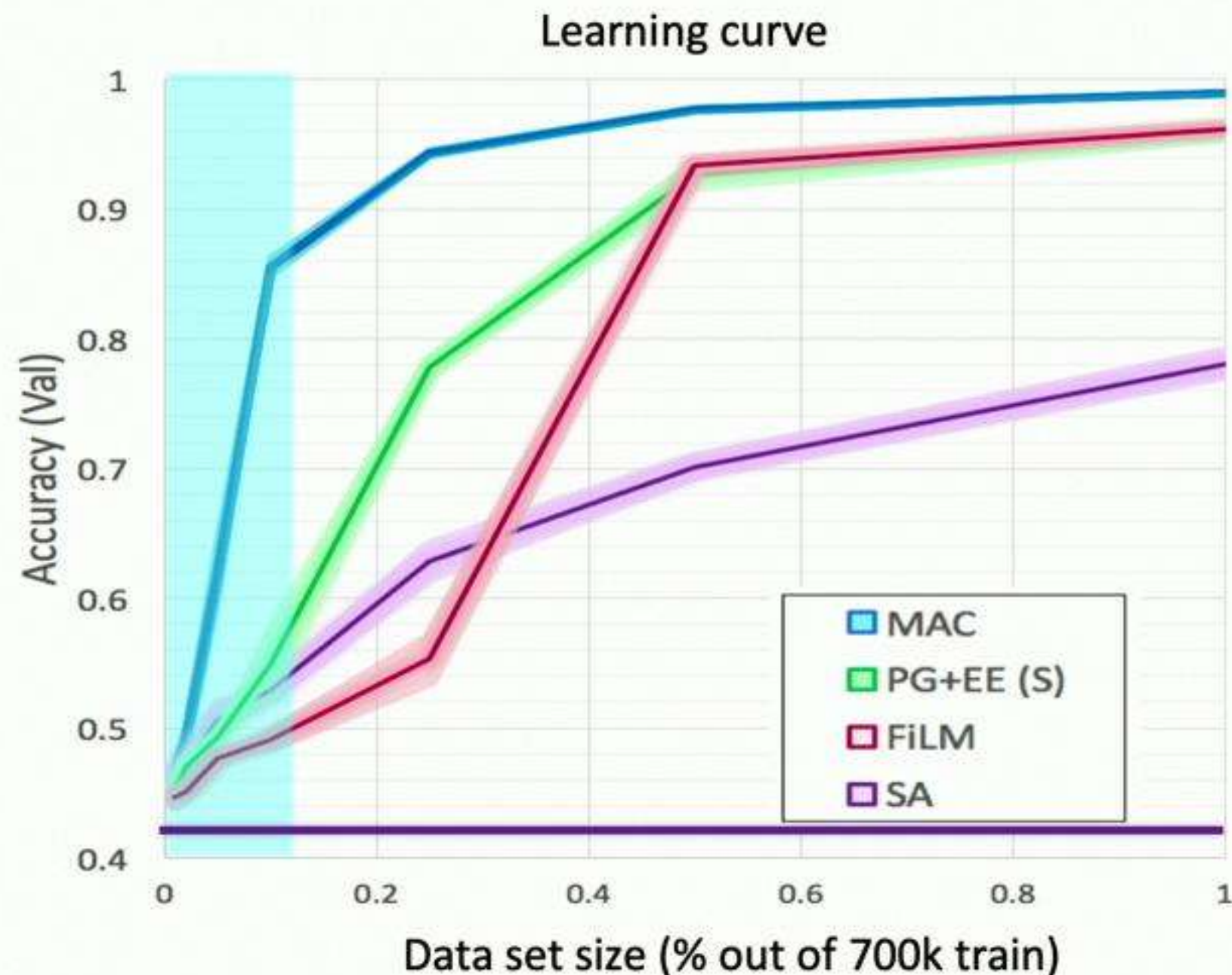
Experiments

Data Efficiency



Experiments

Data Efficiency



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

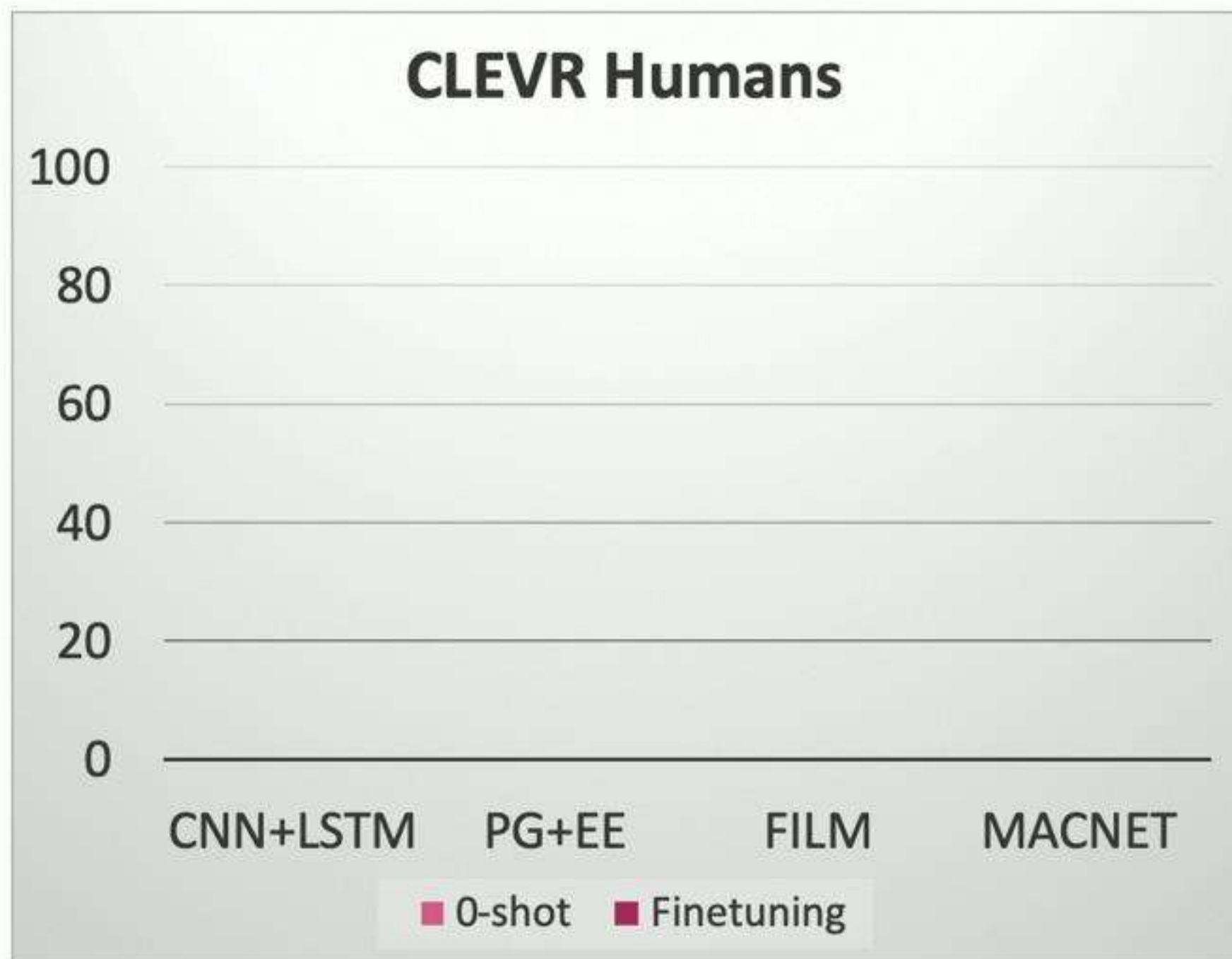
- **MacNet** achieves **86%**
- **Other approaches** obtain **51.6%** at best
- **Baseline** achieves **41.8%**

Baseline

Most Frequent Answer for Question Type

Experiments

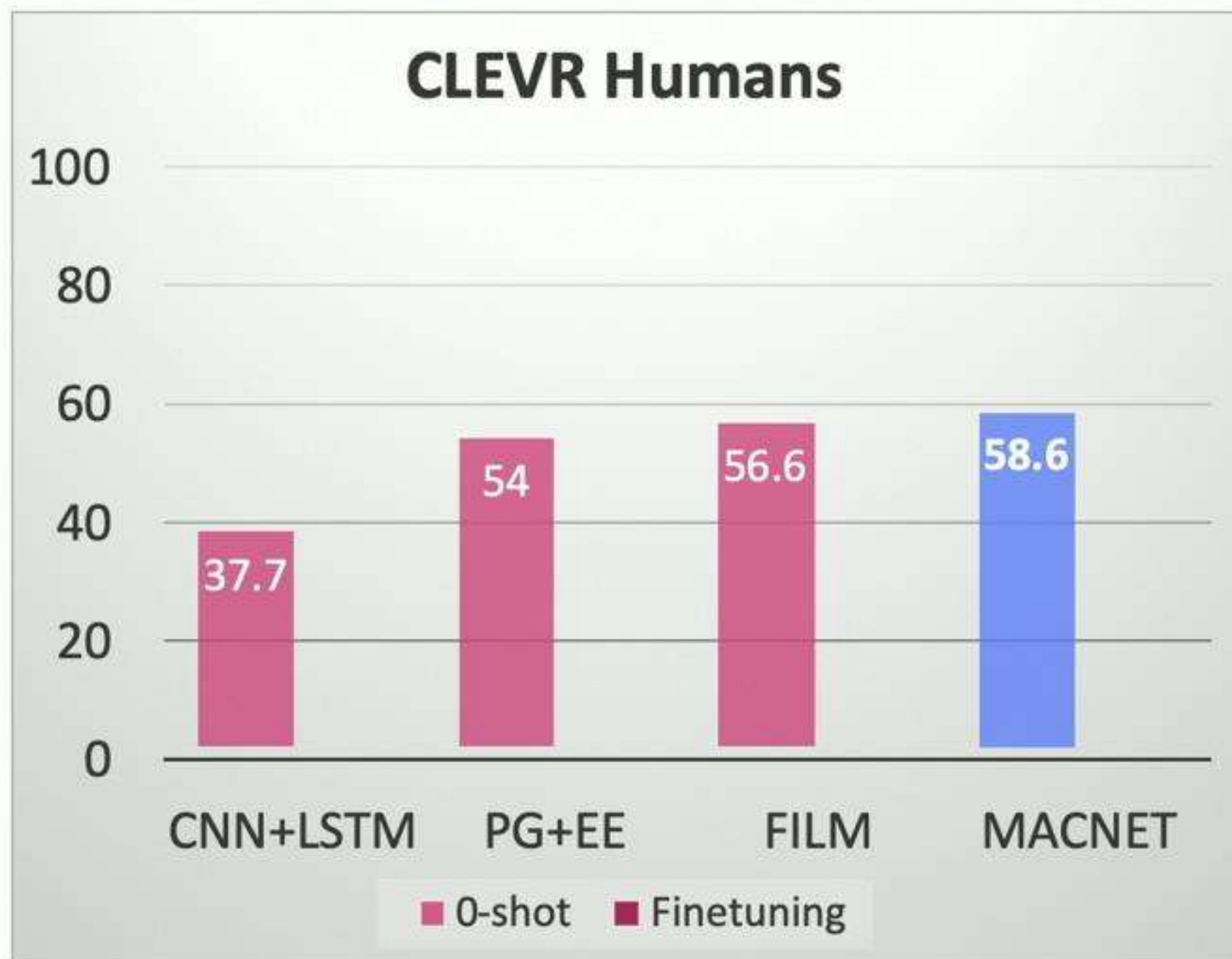
CLEVR-Humans



- 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**
- Has a small training set for fine-tuning

Experiments

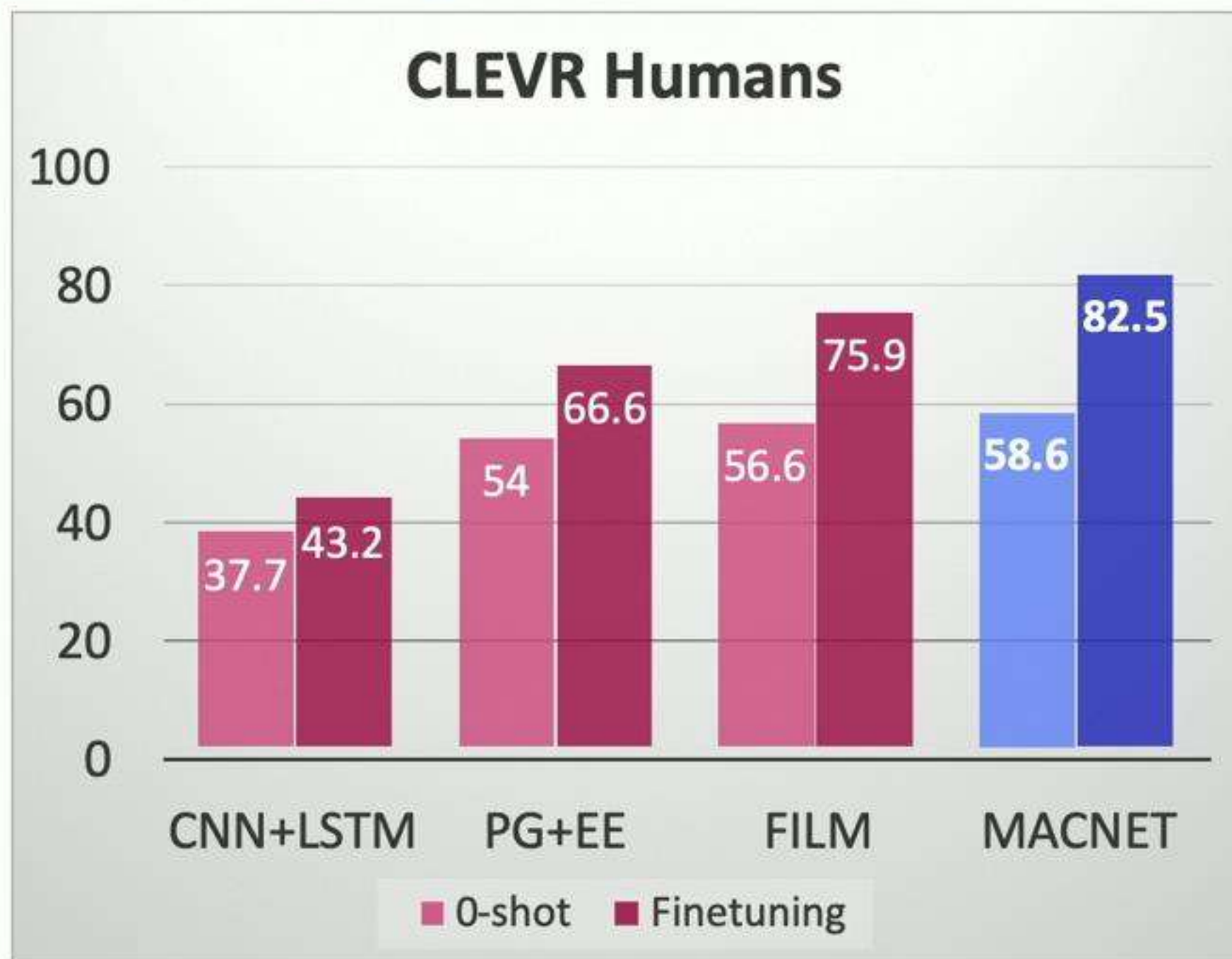
CLEVR-Humans



- 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**
- Has a small training set for fine-tuning

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**
- Has a small training set for fine-tuning

A neural compositional reasoning engine

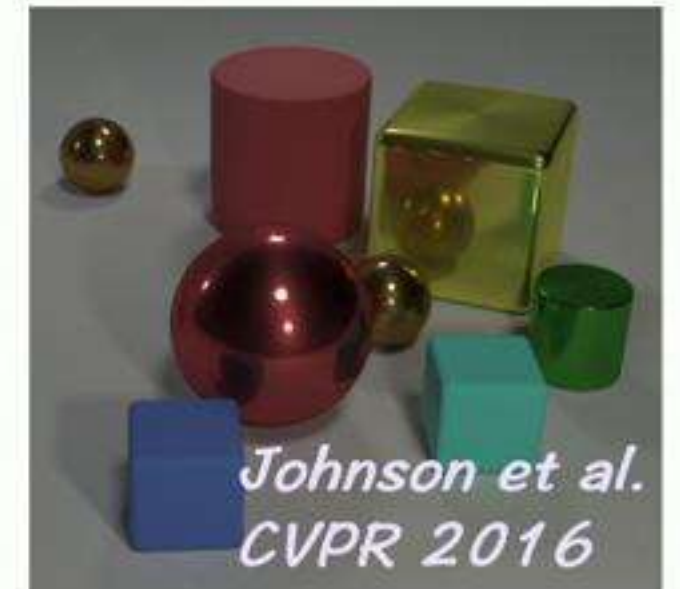
- **An initial design for a compositional reasoning engine**
A constrained sequence model, separating control and memory and exploiting attention is a good prior for reasoning
- **Strong compositional reasoning skills**
Halves the previous lowest error rate
Generalizes much better from more modest training data
Generalizes better to new tasks in CLEVR-Humans
- **Generic, fully differentiable, end-to-end model**

Earlier reasoning datasets are **limited**

Artificial images and/or language

A very **small space** of possible **objects and attributes**

High capacity models may **memorize** all combinations, **reducing effective compositionality**



Current VQA Benchmarks are **problematic**

Strong *language* real-world biases

models *guess* based on *language priors*

Visual biases

models overly focus on salient objects

Unclear error sources

noisy language; lack of object *grounding*

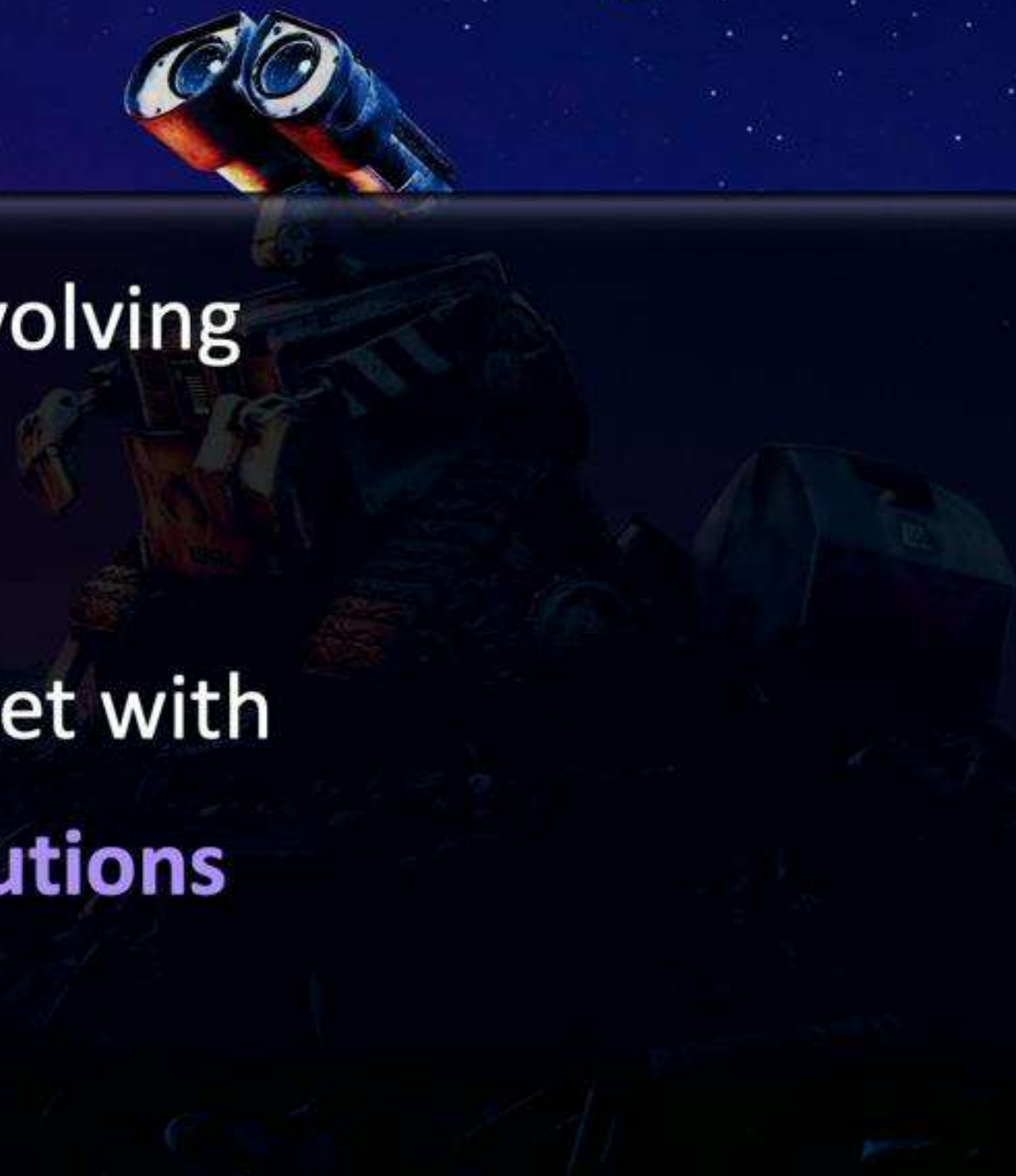
Little reasoning/compositionality required



GQA

a new dataset for compositional question answering over real-world images

- **10M compositional questions** involving a diverse set of **reasoning skills**
- A **balanced 1.5M-questions** dataset with **closely controlled answer distributions**



GQA

a new dataset for compositional question answering over real-world images

- Each **image** comes with a **scene graph** to represent its semantics
- Each **question** comes with a **functional program** to represent its semantics, grounded in the scene graph

GQA

a new dataset for compositional question answering over real-world images

- Questions are **generated** using a (traditional, rule-based) multi-step **question engine** focusing on **linguistic diversity** and a **large vocabulary**
- A **suite of new metrics** exploit the known grounding to shed light on model behaviors in various aspects



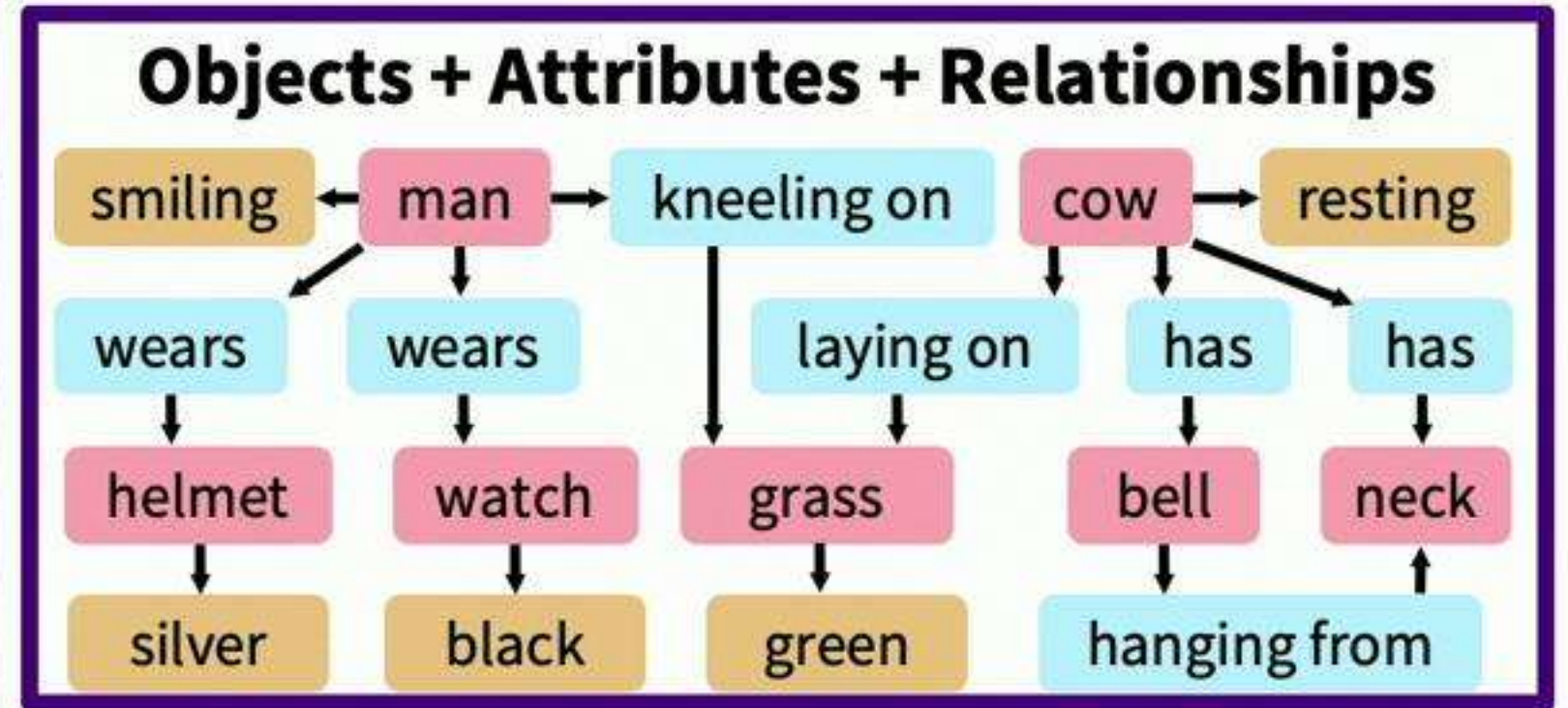
Visual Genome



[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]



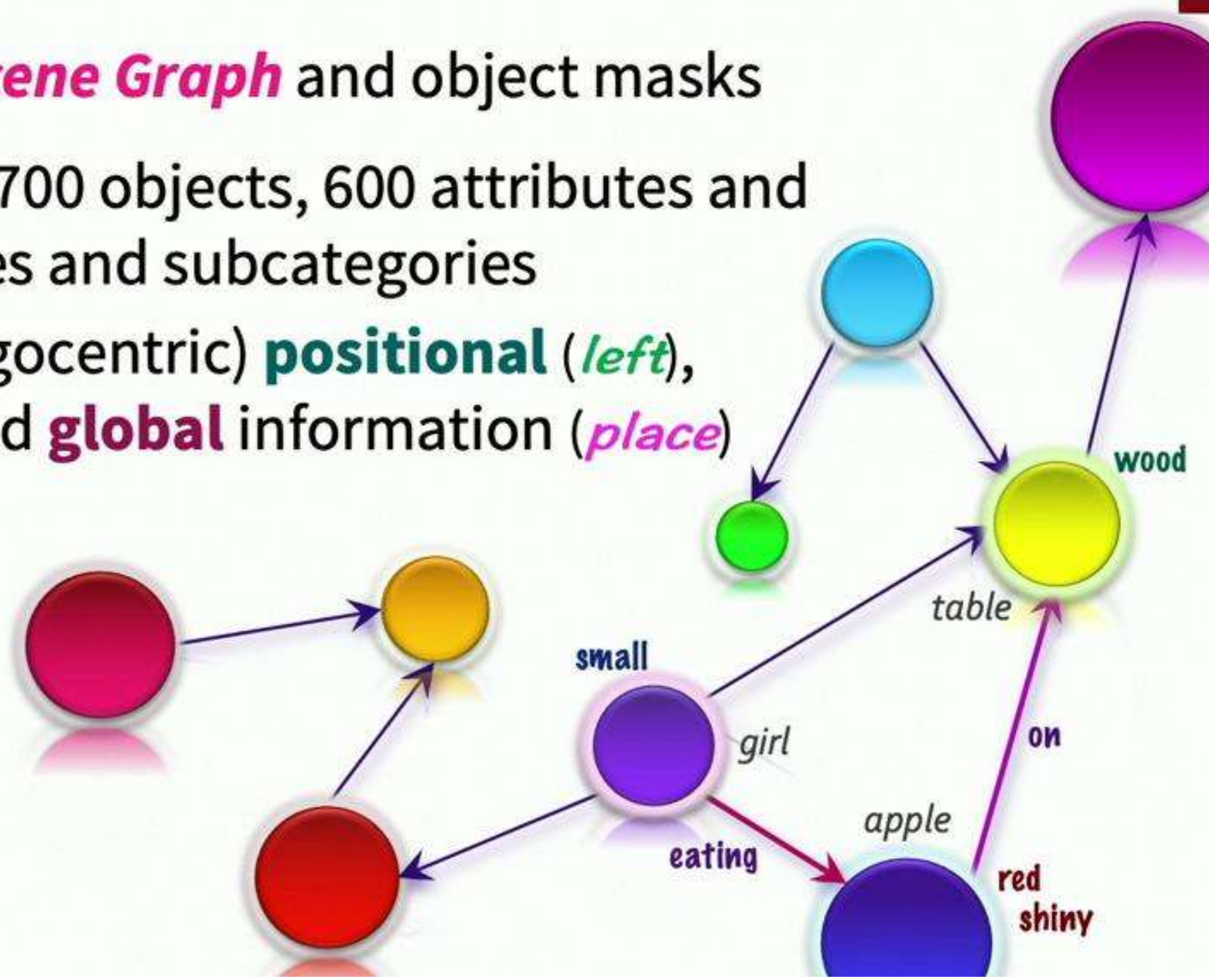
Visual Genome Scene Graph



[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

Improved Visual Genome

- **108k images**, each with a **Scene Graph** and object masks
- Use ontology of concepts: 1700 objects, 600 attributes and 330 relations, in 60 categories and subcategories
- Augment the graphs with (egocentric) **positional** (*left*), **comparative** (*same color*) and **global** information (*place*)

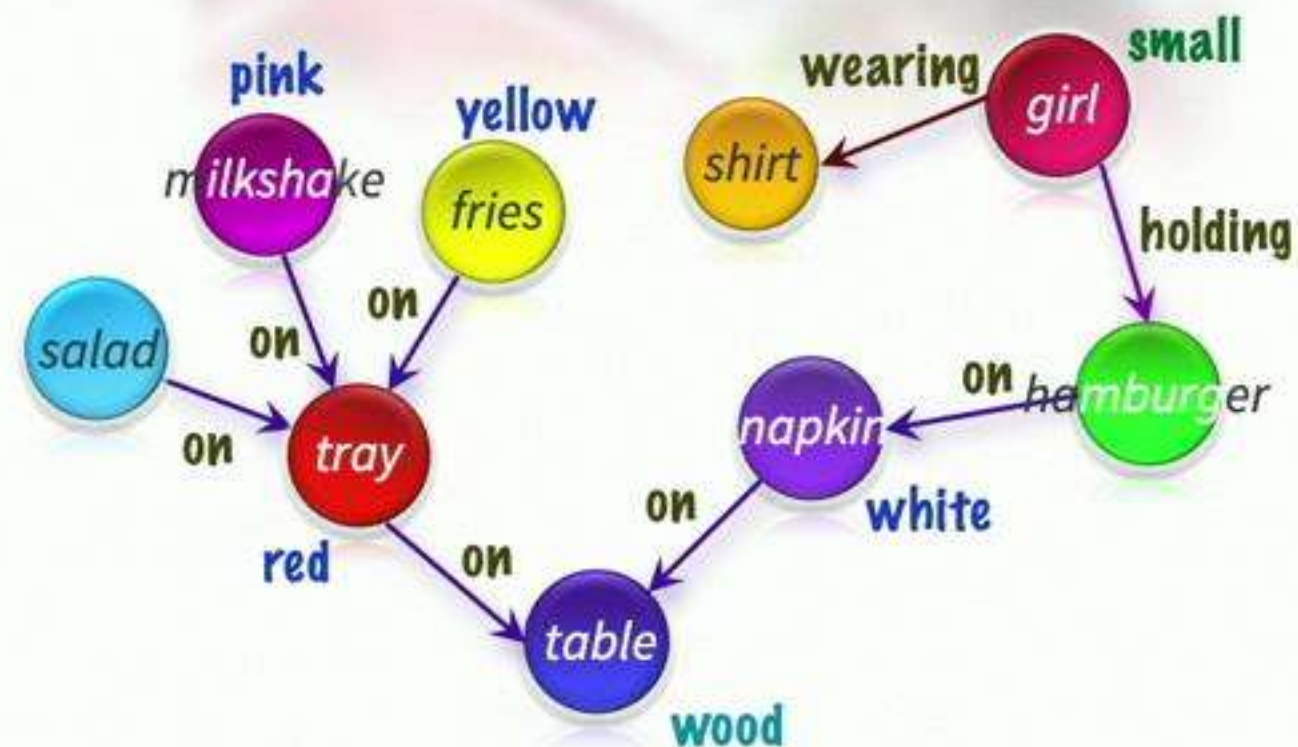


Question generation from graphs

Patterns: 500 probabilistic patterns, give a *high-level question outline*

What|Which *<type>* [do you think] *<is>* *<dobject>*,
<attr> or *<decoy>*?

Select: *<dobject>* → **Choose**
<type>: *<attr>*|*<decoy>*



Example Questions



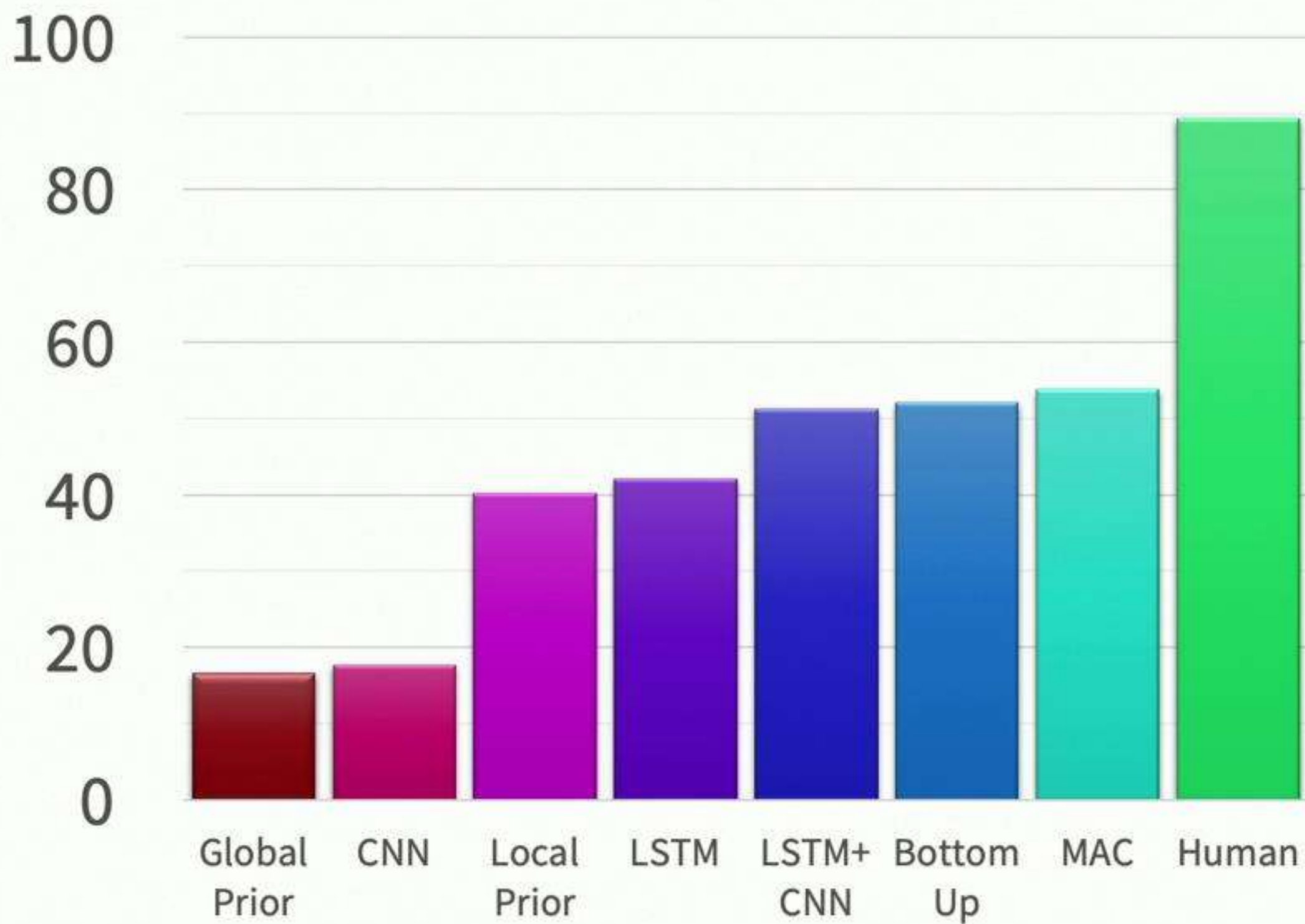
VQA

1. Does this **man** need a **haircut**?
2. What color is the **guy's tie**?
3. What is different about the **man's suit** that shows this is for a special occasion?

GQA

1. Is the **person's hair** long and brown?
2. What **appliance** is to the left of the **man**?
3. **Who** is in front of the **refrigerator** on the left?
4. Is there a **necktie** in the picture that is not red?
5. Is the color of the **vest** different than **shirt**?

Baseline Accuracies



VQA

Language

VQA

A diagram illustrating the components of a Video Question Answering (VQA) system. The letters 'VQA' are displayed in a large, bold, white font. A white bracket is positioned above the 'QA' portion of the text, with the word 'Language' written in white above the bracket. This indicates that 'QA' represents the language component of the system, while 'V' represents the video component.

Language

VQA

Language of Thought

V Abstraction: Towards a Language of Thought



We see and reason with concepts,
not visual details, 99% of the time
“Scene gists”

- A man
 - A cyclist
 - Wearing glasses, gloves, watch
- A cow
- Grassland
- Sky ... clouds

V Abstraction: Towards a Language of Thought

- We use **concepts** to organize our sensory experience
- We build semantic **world models** relating concepts to represent our environment
- Used to **generalize** from given examples to new ones
- Used to draw **inferences** from facts to conclusions

The **hope of deep neural models is to
learn higher-level abstractions**

Abstractions **disentangle factors of
variation, improving generalization**

Content-based attention over concepts

- Attention allows focus on a few elements out of a large set
- But we need attention over **concept space**, not over **pixel space**



- Cf. Yoshua Bengio's so-called "Consciousness Prior"
 - Learn a deep representation that disentangles abstract explanatory factors
 - The conscious state is then a very low-dimensional vector, an attention mechanism applied on the deep representation

A Neural State Machine

- A **differentiable graph-based** model that simulates the operation of a **state machine**
- Aims to combine the strengths of **neural** and **symbolic** approaches

A Neural State Machine

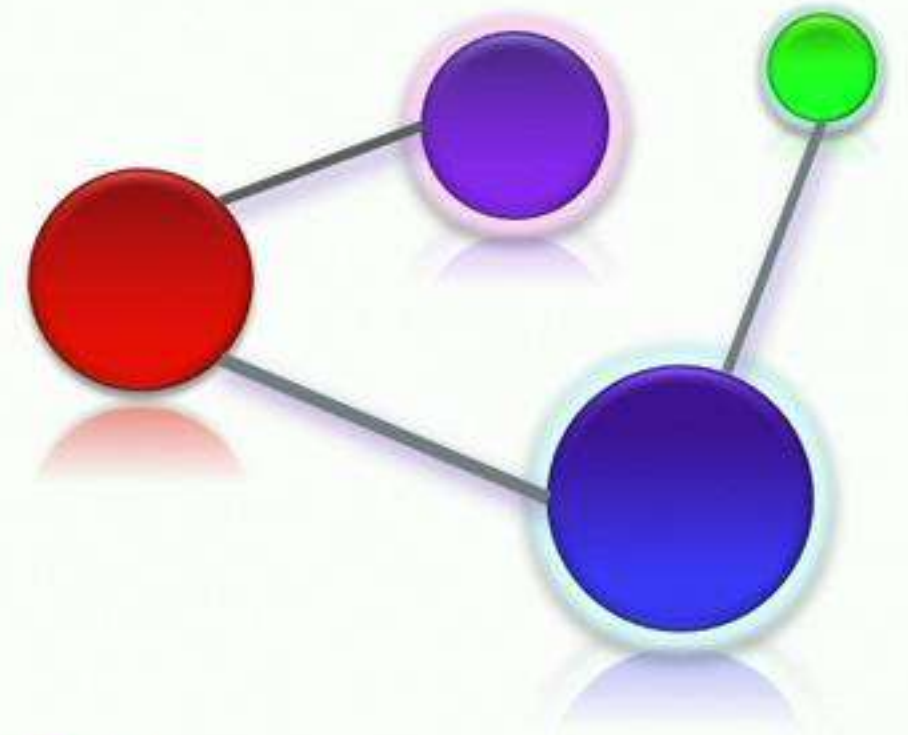
Two stages of **construction** and **inference**:

- 1) **Construction**: transforms the raw inputs into **abstract** semantic representations, building *the state machine*

Image → *Scene graph*, *Question* → *Instructions*

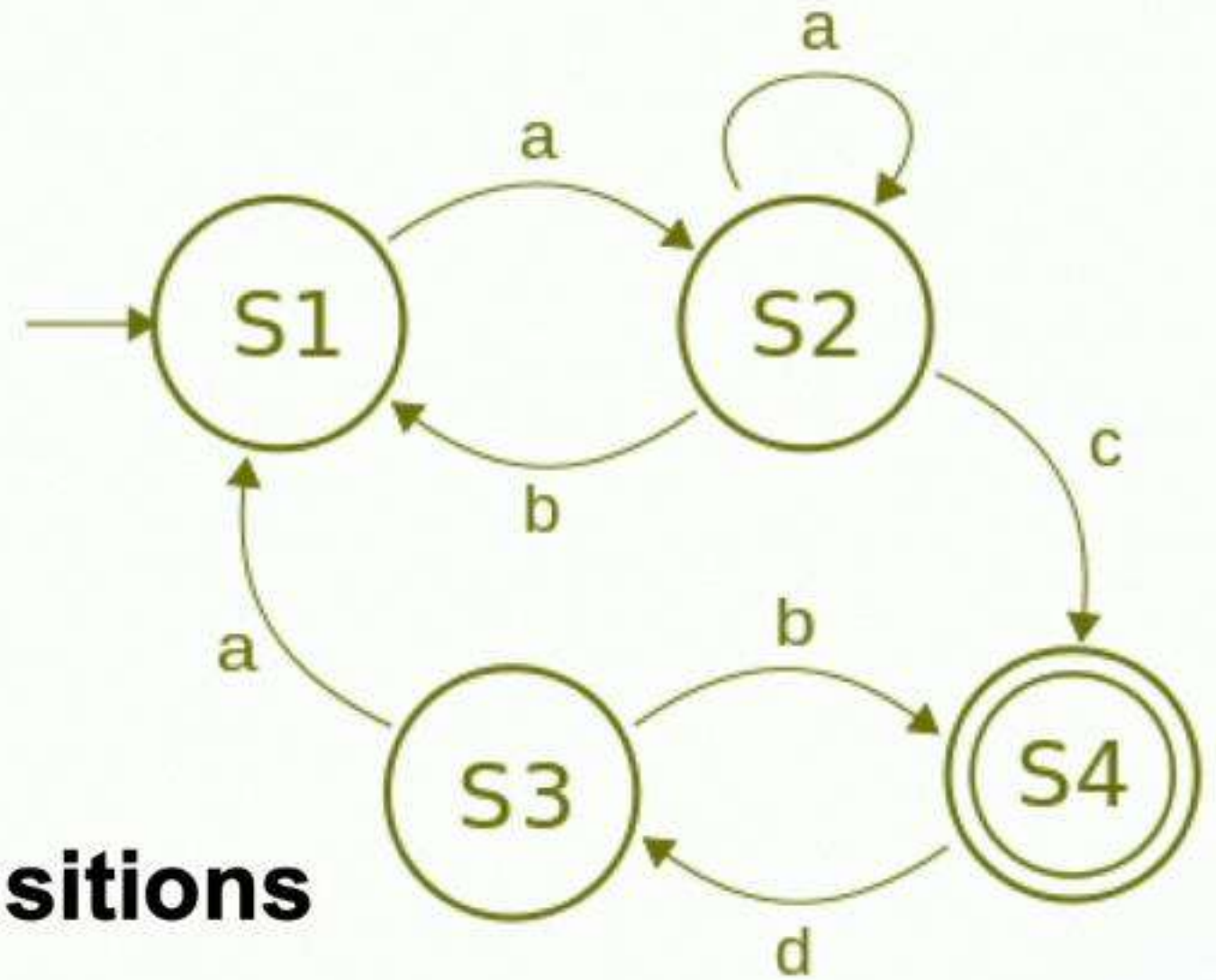
- 2) **Inference**: *simulates an iterative computation* over the machine, sequentially traversing the states until completion.

Reasoning over the scene graph to compute an answer

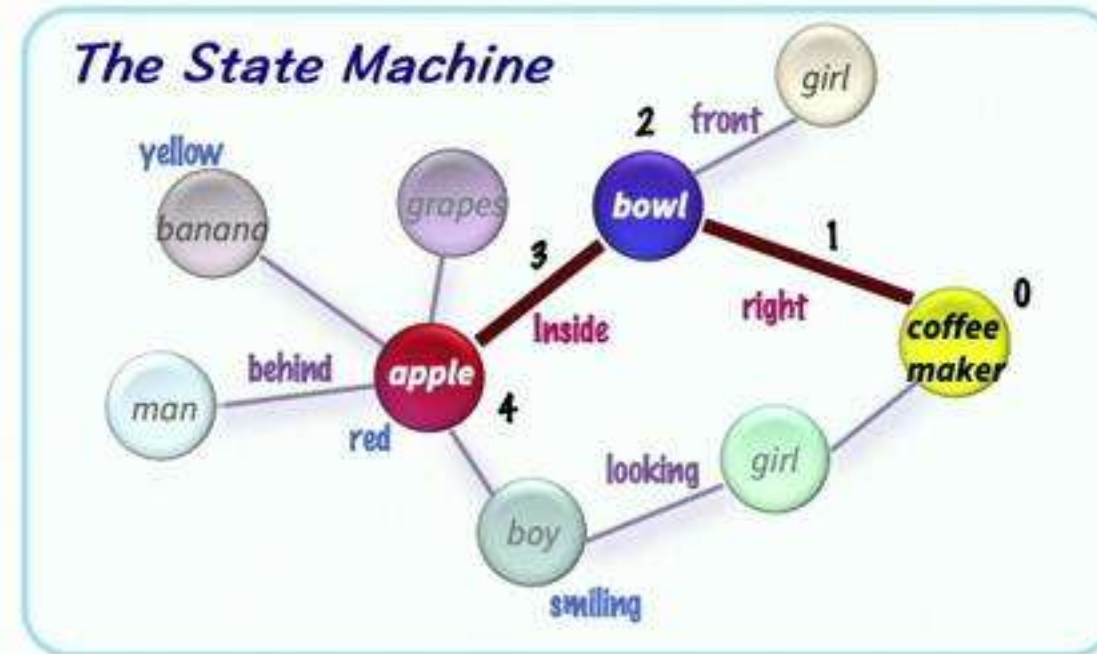


Formal Definition

- \mathcal{C} the model's **alphabet**
(**embedded concepts**)
- \mathcal{S} a set of **states**
- \mathcal{E} a set of **edges** for valid **transitions**
- $r_i, i \leq n$, **instruction** sequence
- p_0 distribution over the **initial state**
- $\delta: p_i \times r_i \rightarrow p_{i+1}$ a **neural state transition function**



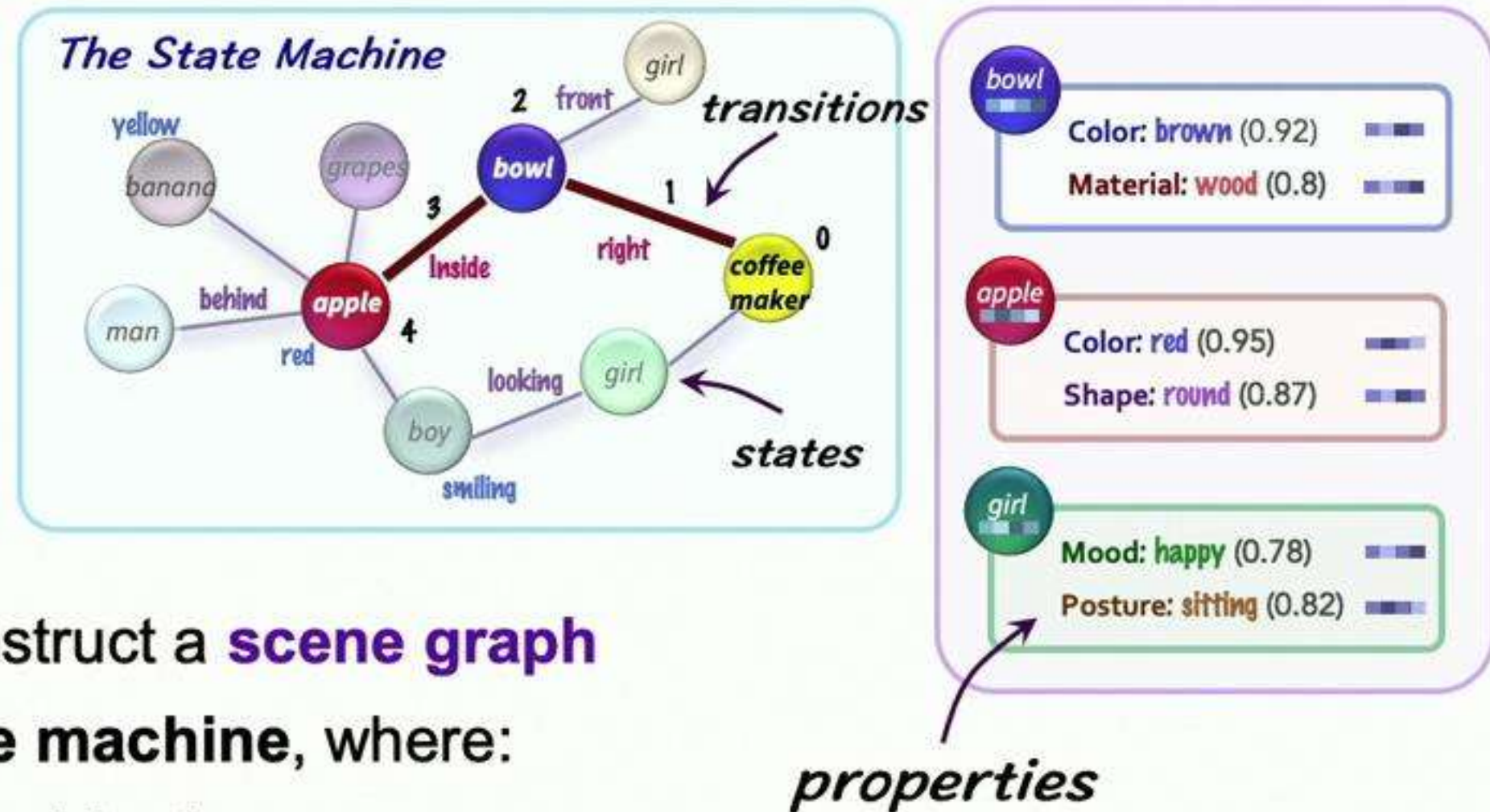
Reasoning with Abstractions



Given an **image**, we construct a **scene graph**

Treat it as a **neural state machine**, where:

Reasoning with Abstractions

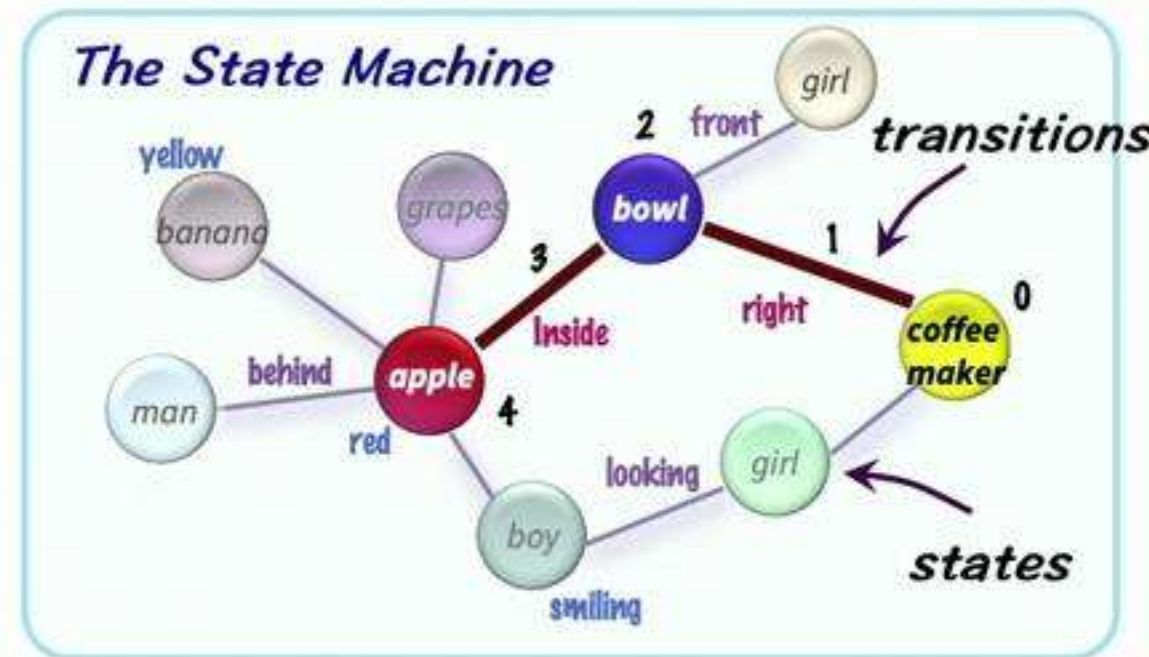


Given an **image**, we construct a **scene graph**

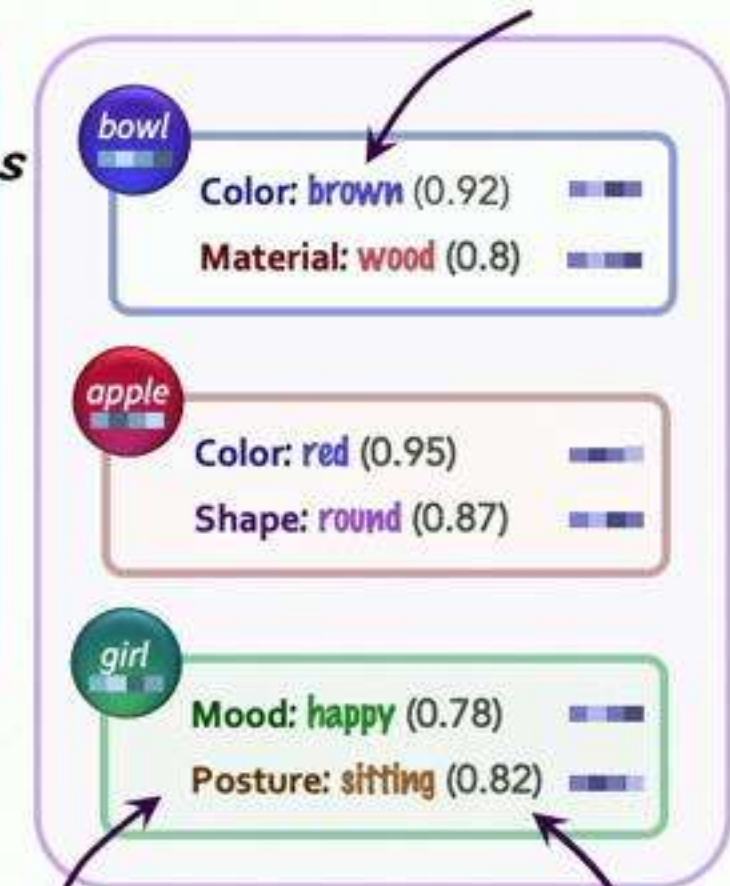
Treat it as a **neural state machine**, where:

- **States** correspond to **objects**
- **Transitions** correspond to **relations**
- States have different **(soft) properties (attributes)** via **attention**

Reasoning with Abstractions



alphabet (concepts)

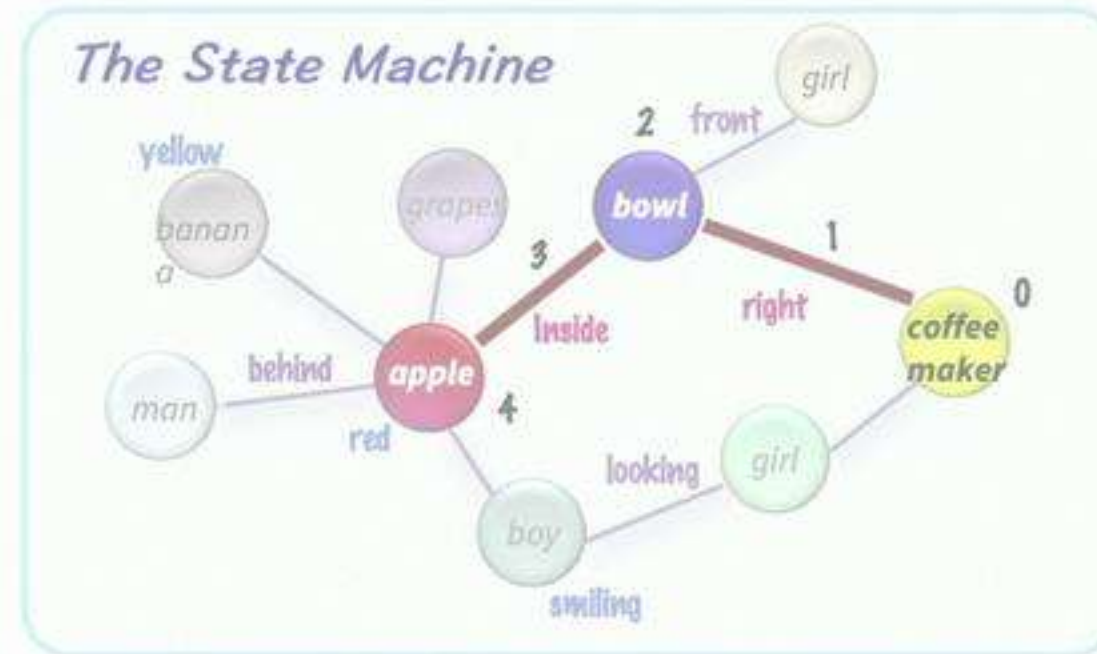


properties

disentangled representation

Objects are represented through a **factorized distribution** over **semantic properties** (*color, shape, material*), defined over the **concept vocabulary**

Reasoning with Abstractions



alphabet (concepts)



*What is the **red fruit** inside of the **bowl** to the right of the **coffee maker**?*



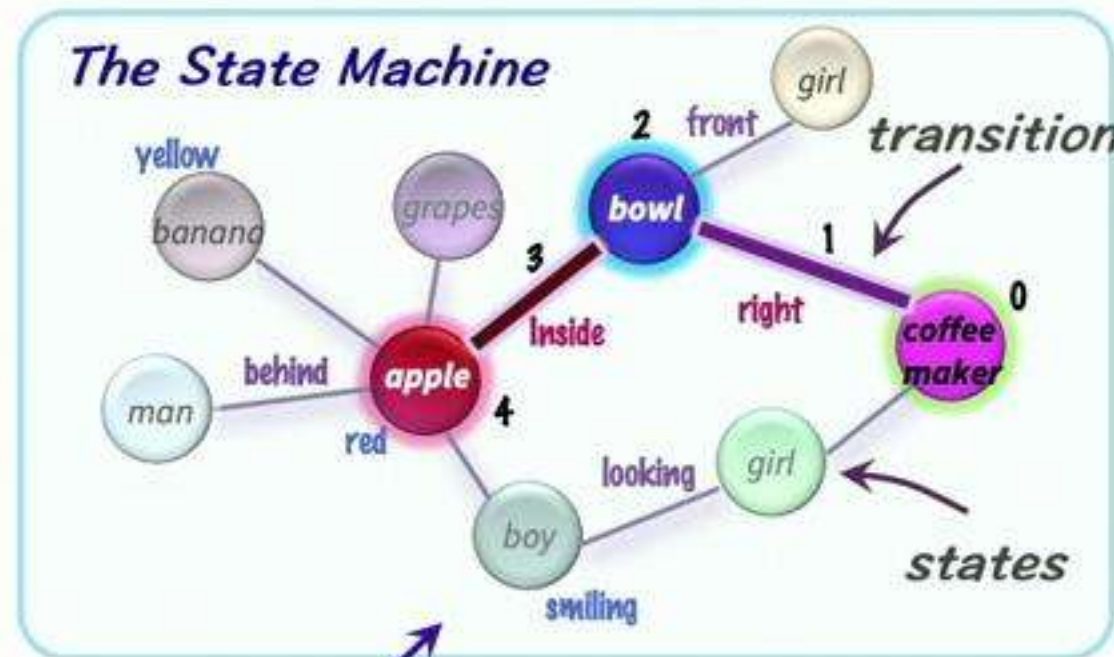
instructions

properties

disentangled representation

The question is translated into a **series of instructions** (with an attention-based encoder-decoder), defined over the **concepts**

Reasoning with Abstractions



alphabet (concepts)



What is the red fruit inside of the bowl to the right of the coffee maker?

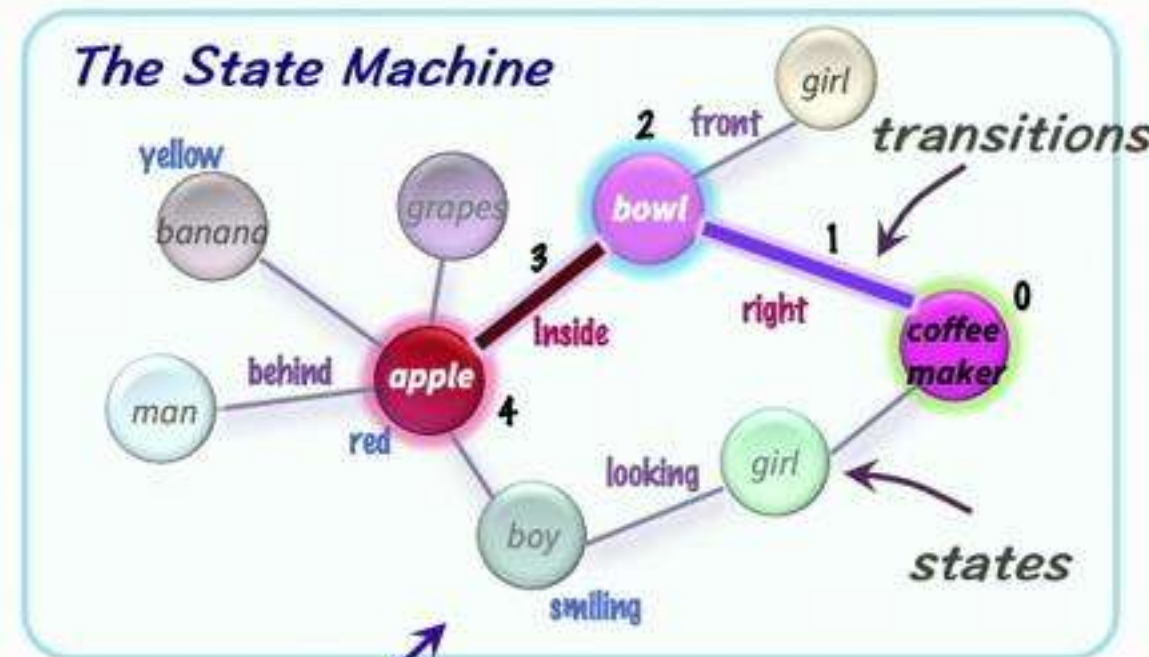


instructions

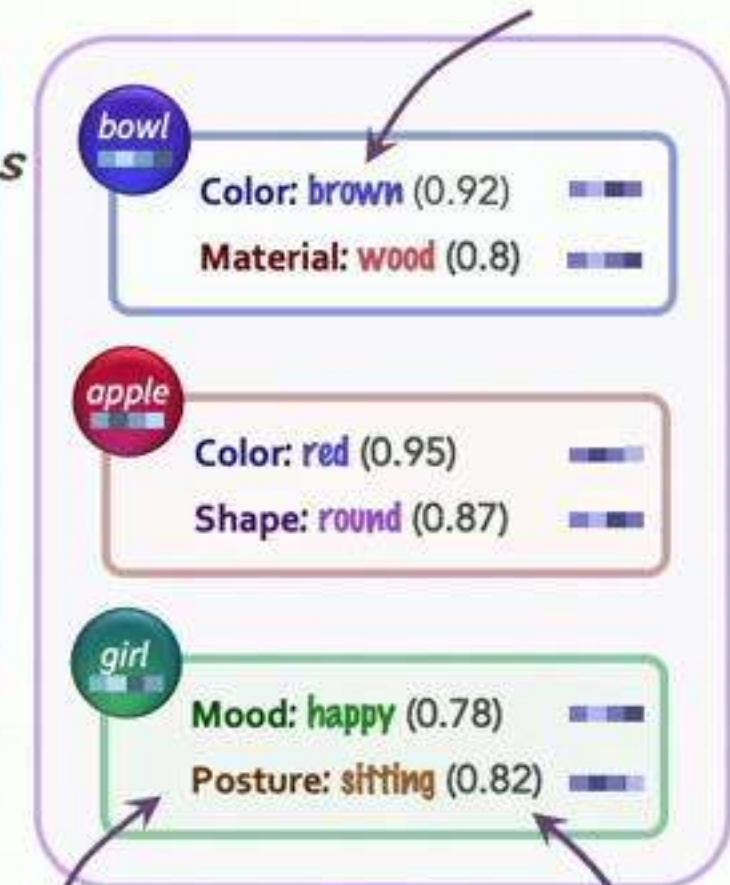
properties

disentangled representation

Reasoning with Abstractions



alphabet (concepts)



What is the red fruit inside of the bowl to the right of the coffee maker?



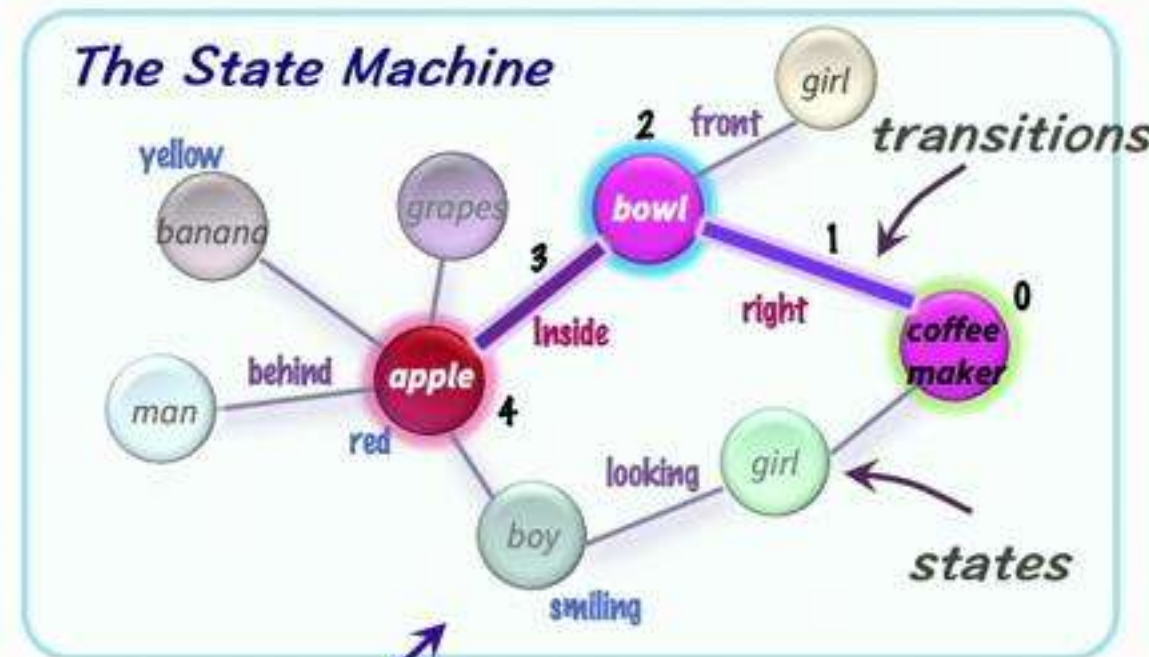
instructions

properties

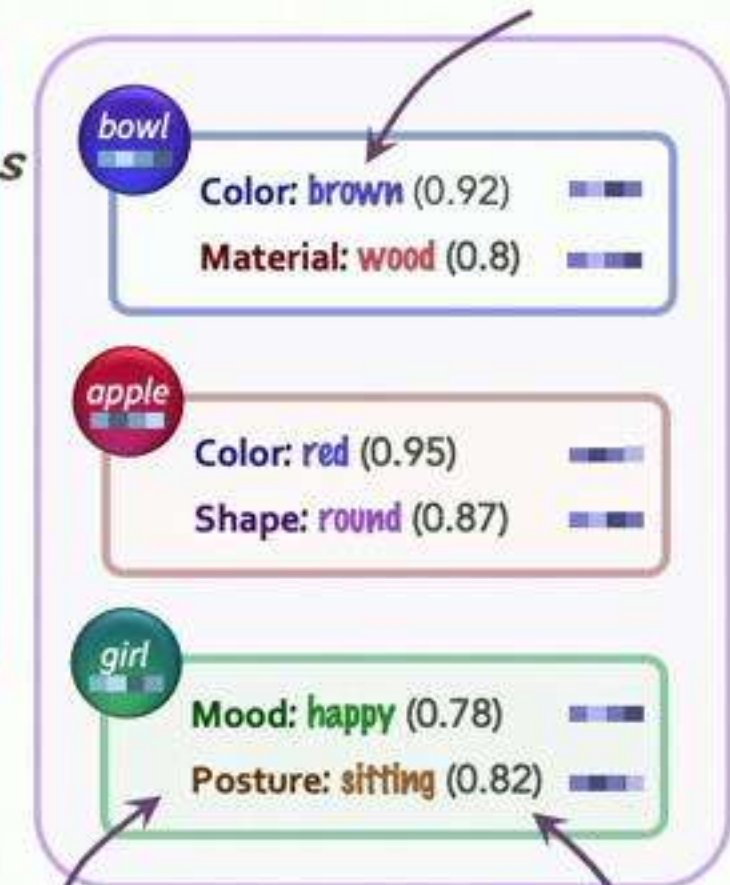
disentangled representation

We **simulate** a computation as a neural state machine, feeding one **instruction** at a time and **traversing the states** until completion.

Reasoning with Abstractions



alphabet (concepts)



What is the red fruit inside of the bowl to the right of the coffee maker?



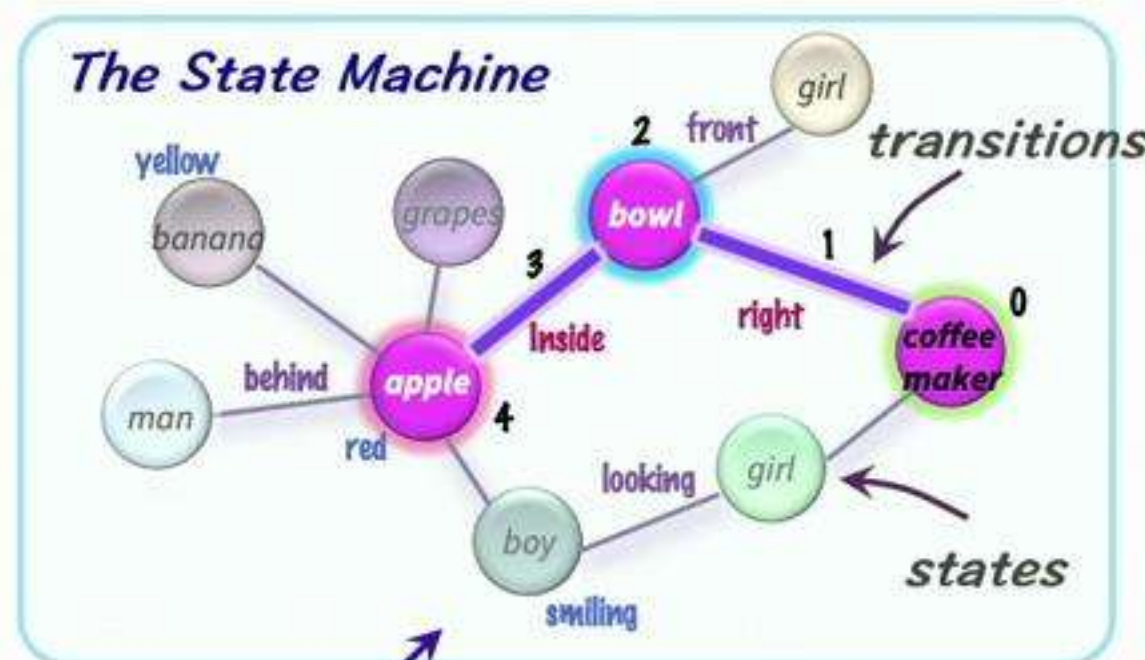
instructions

properties

disentangled representation

We **simulate** a computation as a neural state machine, feeding one **instruction** at a time and **traversing the states** until completion.

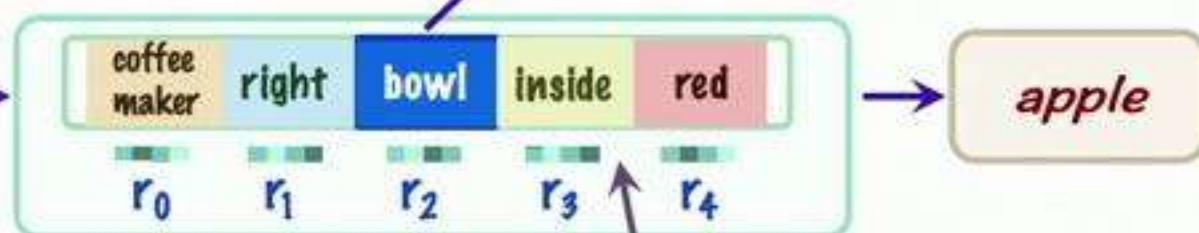
Reasoning with Abstractions



alphabet (concepts)



What is the red fruit inside of the bowl to the right of the coffee maker?



instructions

properties

disentangled representation

We **simulate** a computation as a neural state machine, feeding one **instruction** at a time and **traversing the states** until completion.

One more example



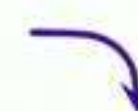
bed



left



tall



made

What is the tall object to the left of the bed made of?

Cabinet: wood (0.95), tall (0.92), shiny (0.86)

Bed: white (0.84), comfortable (0.91)

Lamp: yellow (0.92), on (0.74), thin (0.82)

(cabinet, left, bed) (0.82)

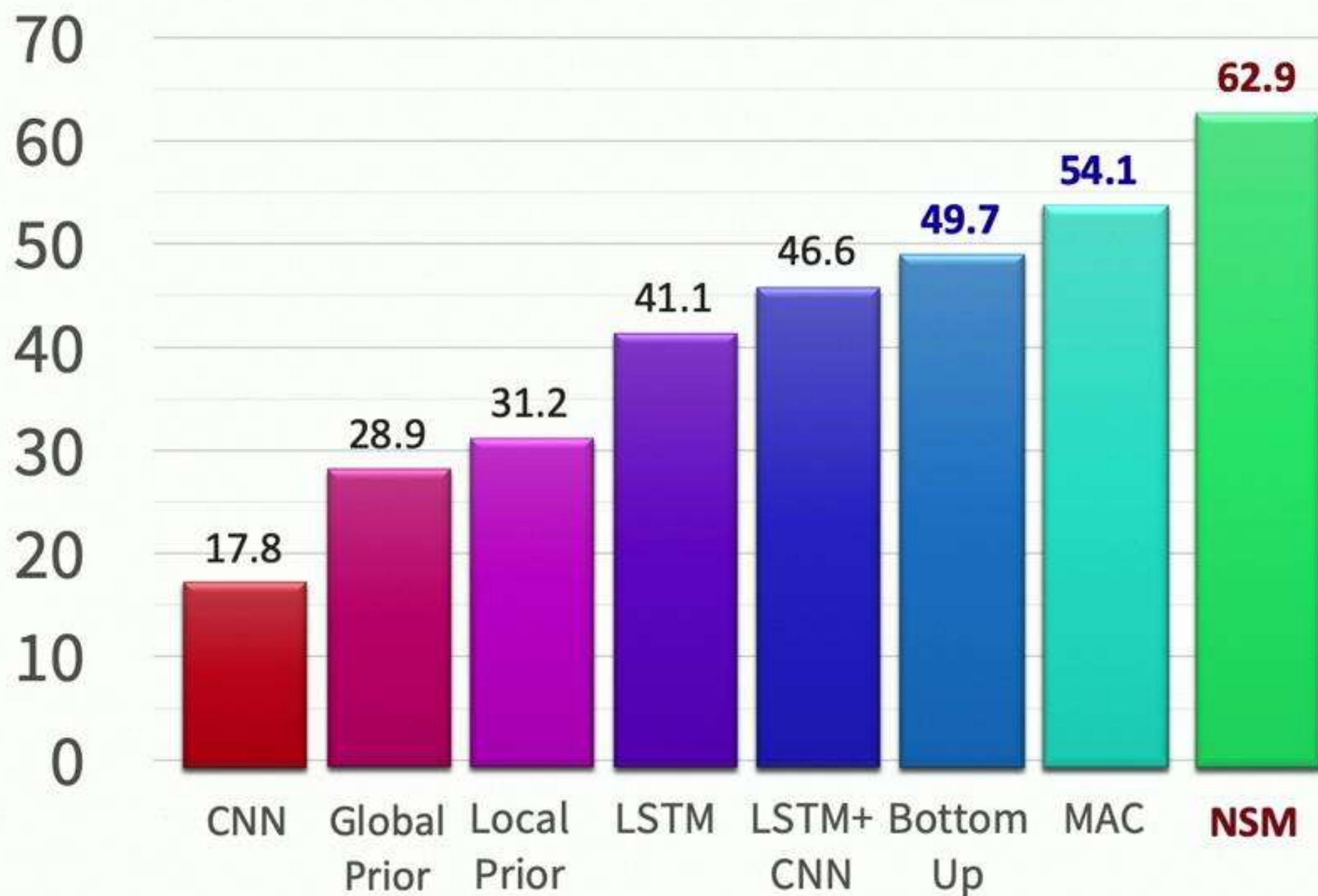
(pillow, on, bed) (0.74)

...



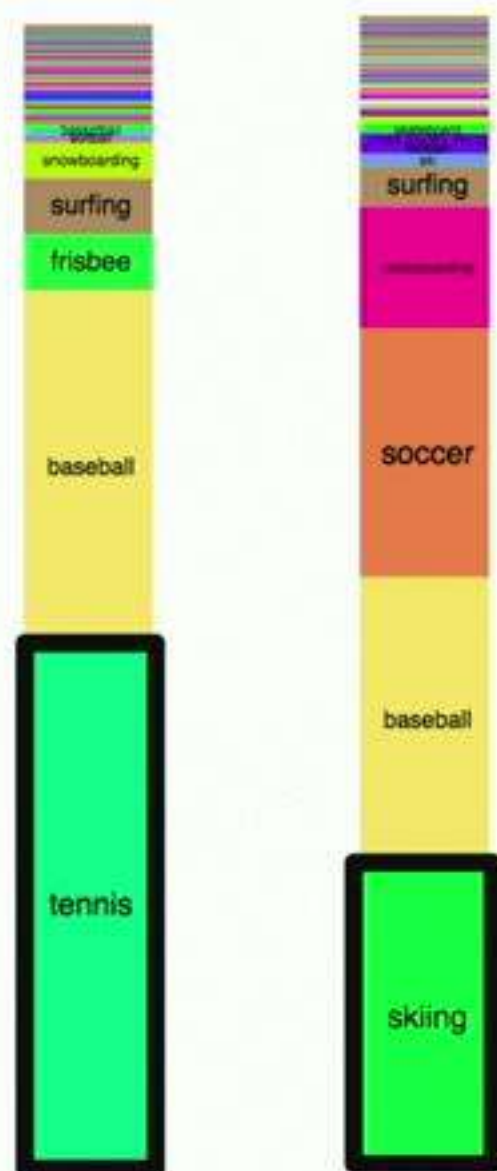
Wood

NSM accuracy on GQA



Testing Disentanglement (\approx Understanding) — VQA-CP: VQA under Changing Priors [Agrawal et al. 2017]

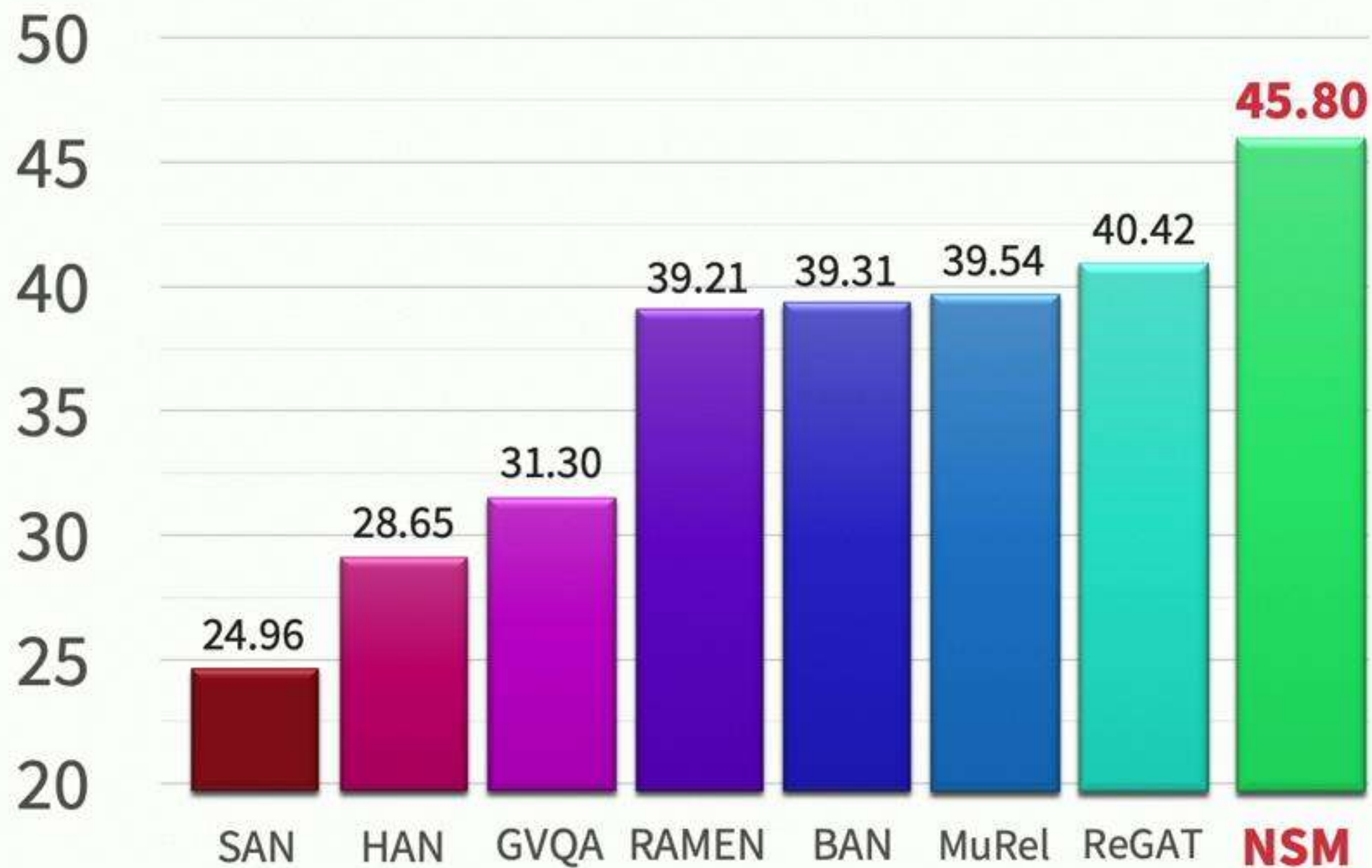
Train Split Test Split



What sport ...?

Model	Dataset	Overall score
d-LSTM Q + norm I (Antol et al. ICCV 2015)	VQA v1	54.40
	VQA-CP v1	23.51 -31%
NMN (Andreas et al. CVPR 2016)	VQA v1	54.83
	VQA-CP v1	29.64 -25%
SAN (Yang et al. CVPR 2016)	VQA v1	55.86
	VQA-CP v1	26.88 -29%
MCB (Fukui et al. EMNLP 2016)	VQA v1	60.97
	VQA-CP v1	34.39 -27%

Generalization on VQA-CP v2



GQA Generalization Splits

training

testing

structure

What is the <obj> **covered by**?

Is there a <obj> in the **image**?

What is the <obj> **made of**?

What's the name of the <obj> **that is** <attr>?

What is **covering the** <obj>?

Do you see any <obj>s in the **photo**?

What **material makes up** the <obj>?

What is the <attr> <obj> **called**?

content

Only questions that **do not** refer to any type of **food** or **animal** (do not have any word from these categories)

Only questions that refer to **foods** or **animals** (have a word from that one of these categories)

GQA Generalization Results

Model	Content	Structure
Global Prior	8.51	14.64
Lobal Prior	12.14	18.21
Vision	17.51	18.68
Language	21.14	32.88
Lang+Vision	24.95	36.51
BottomUp	29.72	41.83
MAC	31.12	47.27
NSM	40.24	55.72

Language

V Q A

A diagram illustrating the components of a VQA (Video Question Answering) task. The letters 'V', 'Q', and 'A' are displayed in a large, bold, white font. A white bracket is positioned above the 'Q' and 'A' characters, with the word 'Language' centered above the bracket. This indicates that the question and answer parts of the task are related to the concept of language.

Language

VQA

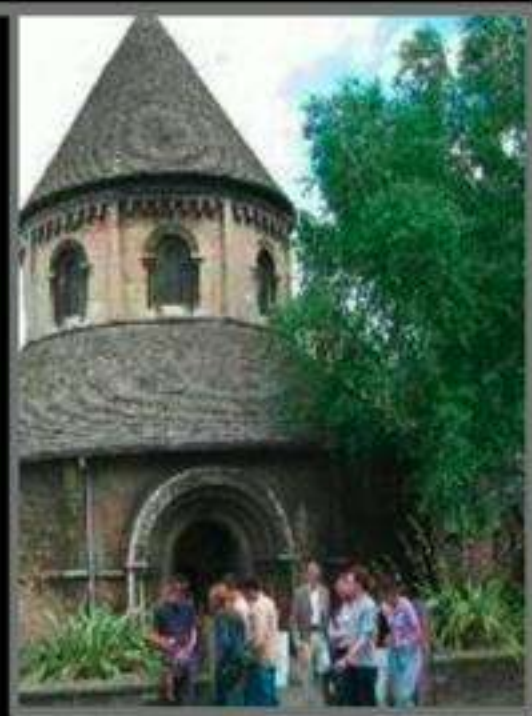
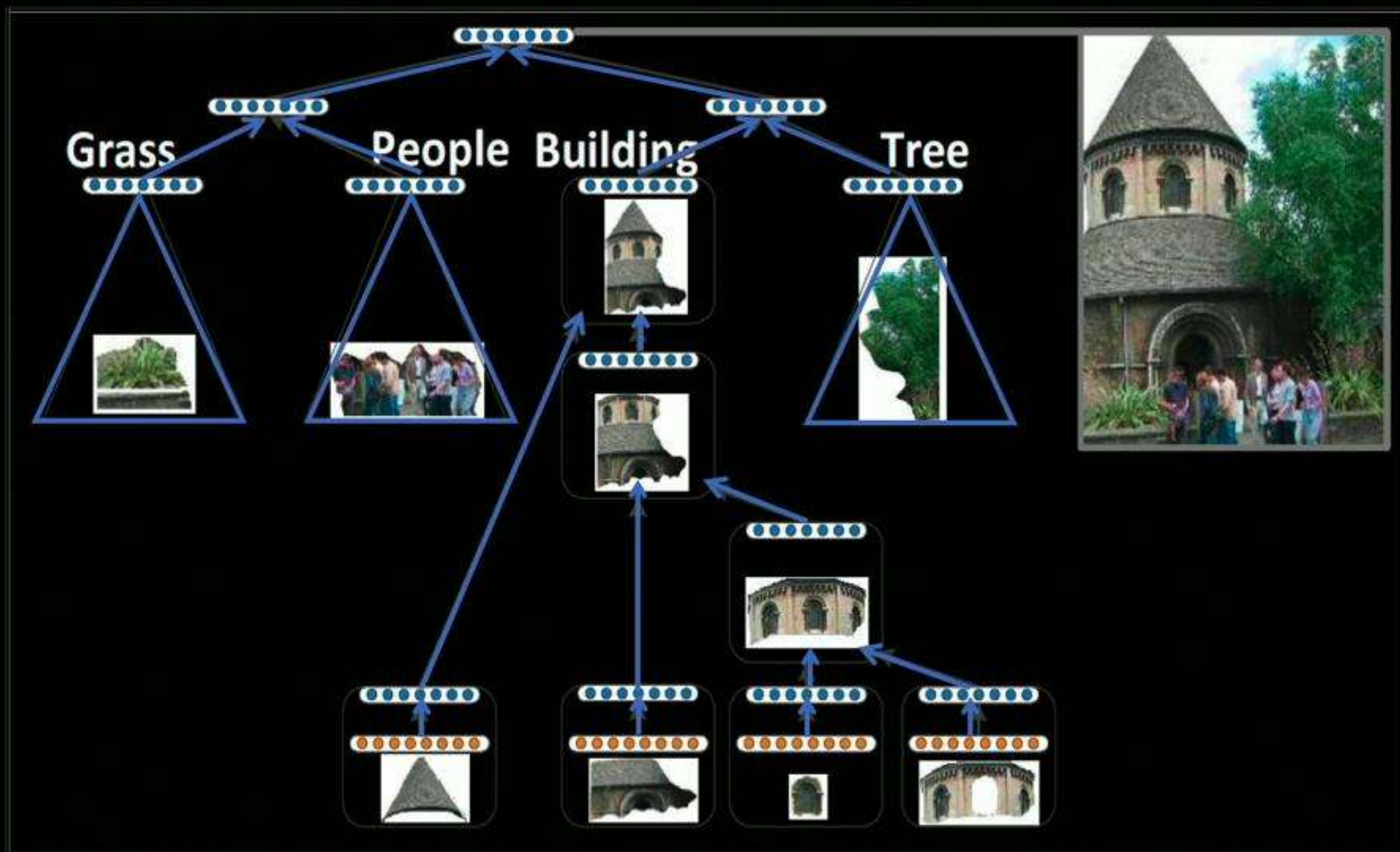
Language of Thought

**We should seek tasks involving
understanding and
multi-step compositional reasoning**

Let's build neural networks that think!

**By iterative attention over
abstracted, disentangled concepts**

Tree-structured models



[Socher et al. 2011]

GQA

a new dataset for compositional question answering over real-world images

- Questions are **generated** using a (traditional, rule-based) multi-step **question engine** focusing on **linguistic diversity** and a **large vocabulary**
- A **suite of new metrics** exploit the known grounding to shed light on model behaviors in various aspects

Visual Genome



[Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017]

A Neural State Machine

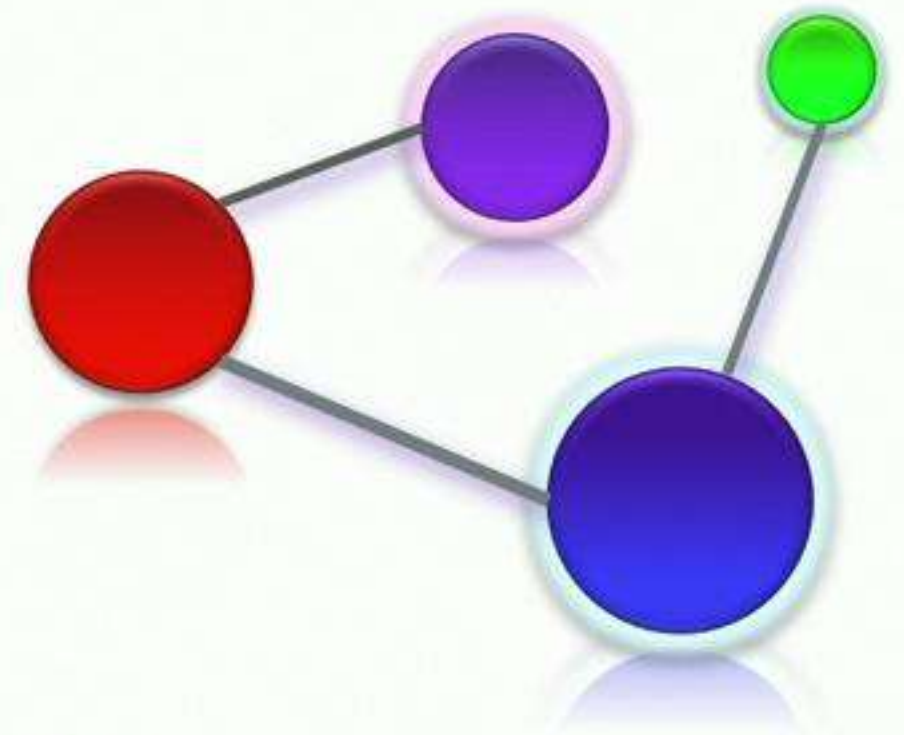
Two stages of **construction** and **inference**:

- 1) **Construction**: transforms the raw inputs into **abstract** semantic representations, building *the state machine*

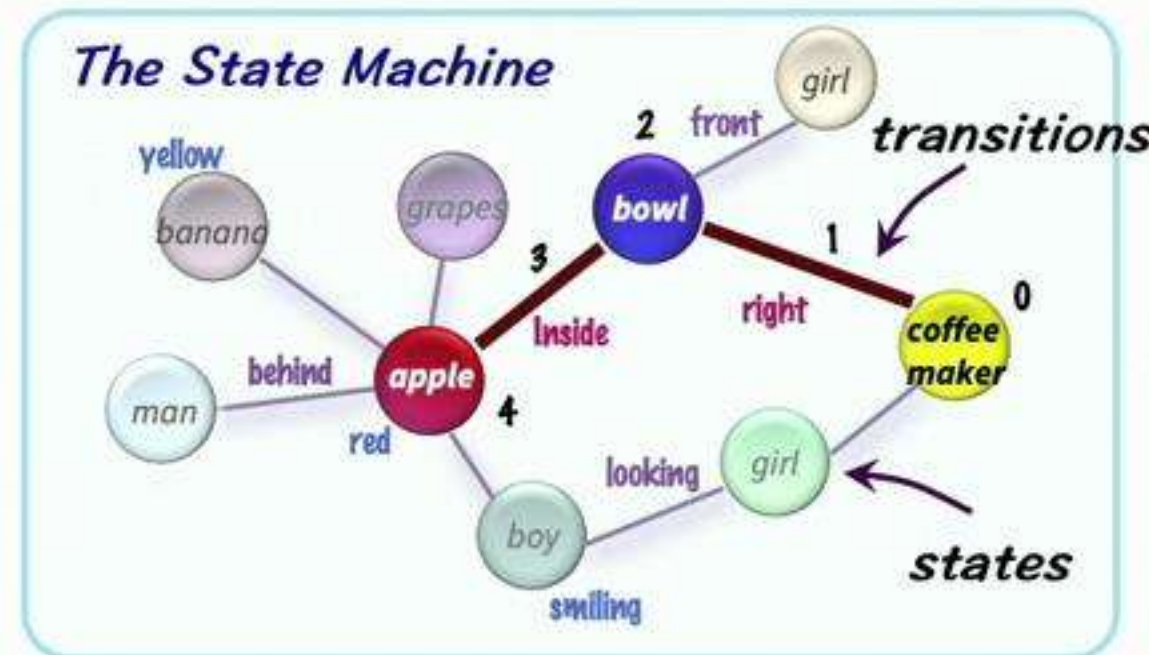
Image → *Scene graph*, *Question* → *Instructions*

- 2) **Inference**: *simulates an iterative computation* over the machine, sequentially traversing the states until completion.

Reasoning over the scene graph to compute an answer



Reasoning with Abstractions



Given an **image**, we construct a **scene graph**

Treat it as a **neural state machine**, where:

- **States** correspond to **objects**
- **Transitions** correspond to **relations**
- States have different **(soft) properties (attributes)** via **attention**