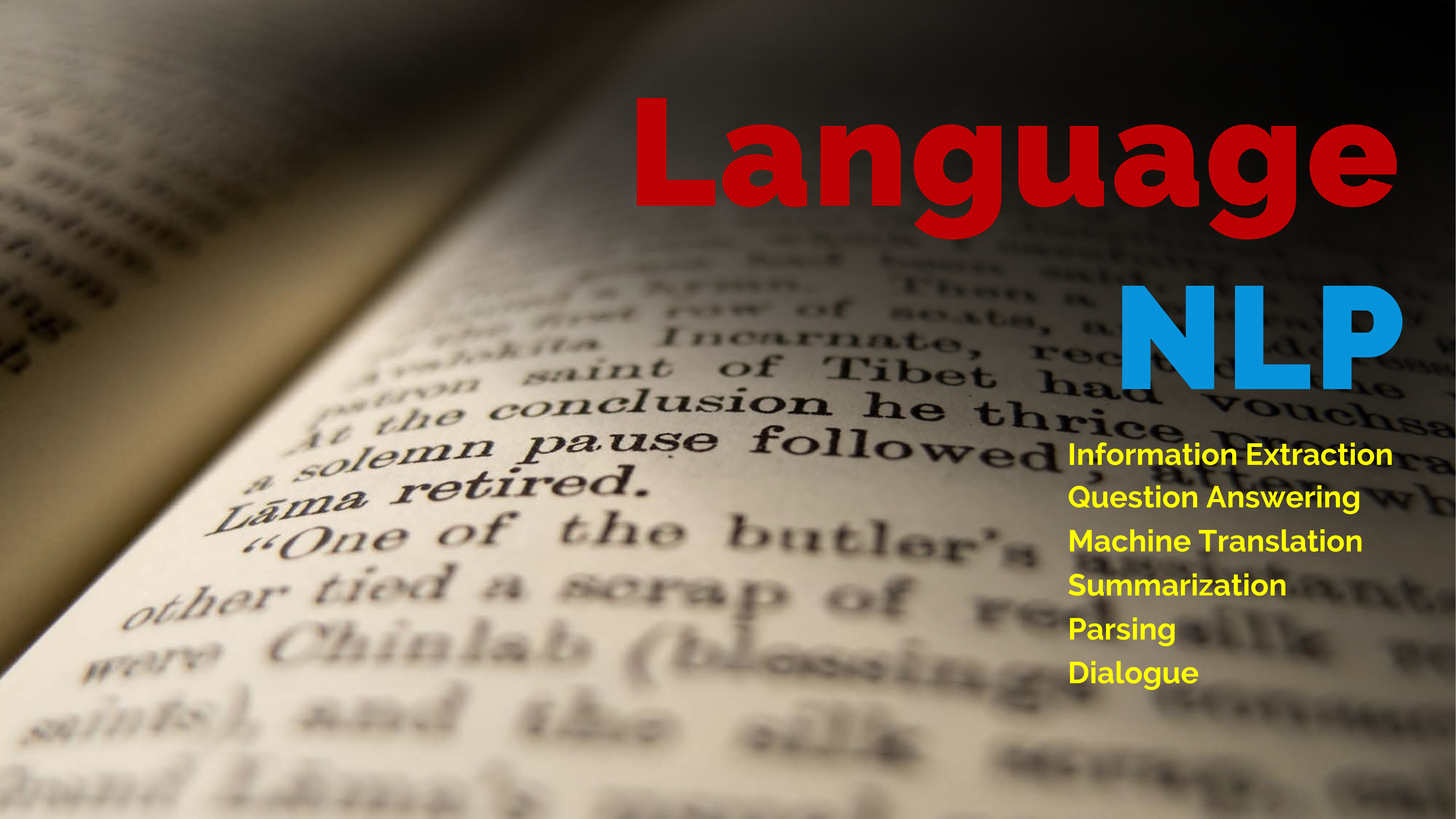
Learning By Abstraction:
The Neural State Machine

Drew Hudson & Christopher Manning

Stanford University

NeurIPS 2019

Language



Language

NLP

- Information Extraction
- Question Answering
- Machine Translation
- Summarization
- Parsing
- Dialogue

Language

Encodes **information**

Means of **communication**

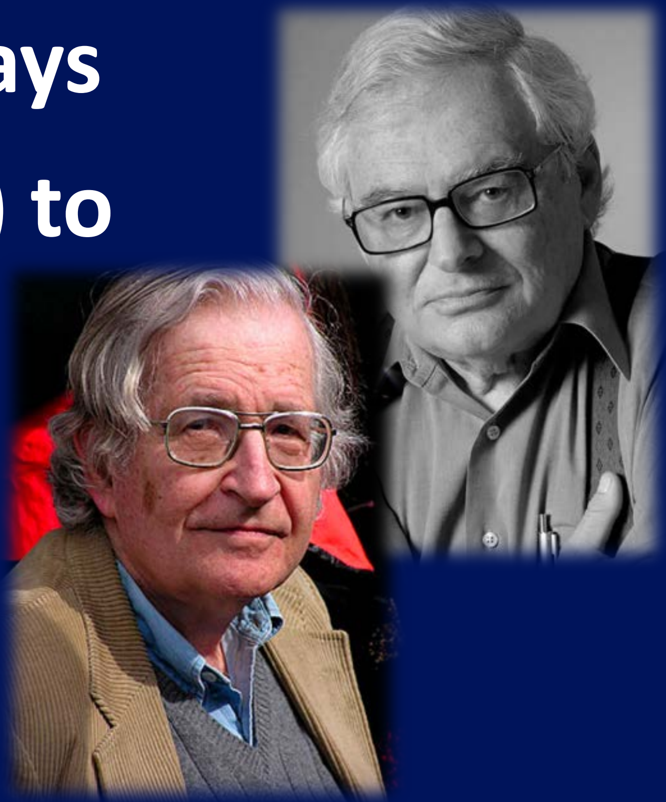
Processing **data, input**

Language Capacity

think through language, to represent information and experiences through a compositional discrete system

Language of Thought

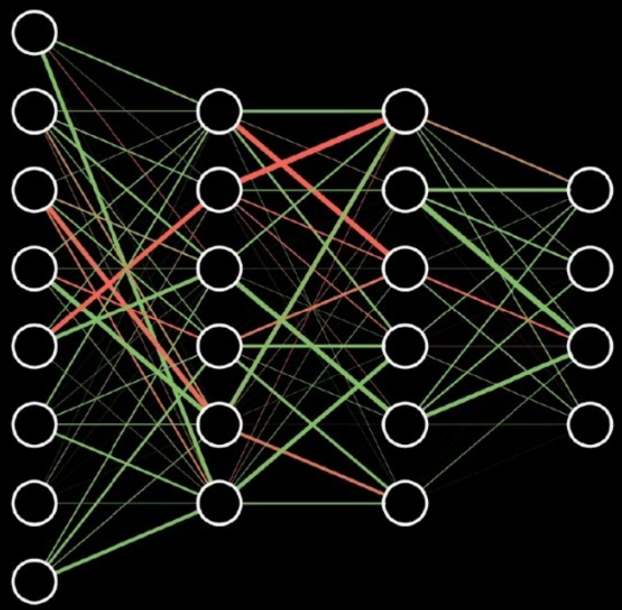
Thought itself posses a "language-like" compositional structure (*mentalese*), where simple concepts combine in systematic ways (akin to the rules of grammar in language) to build thoughts. In its most basic form, the theory states that thought, like language, has syntax.



Language of Thought

- **Generalization** (to a new concept, transfer)
- **Data Efficiency:** Learning from few examples
- **Interpretable:** Express our thought process
- **Abstraction:** support human-unique capability of **reasoning and abstract thinking**

Neural Networks

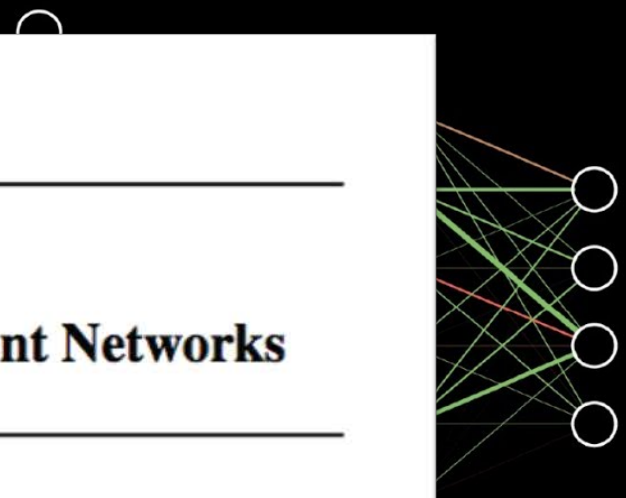


Continuous Computation

- They confuse **correlation** with causation
- They need a **lot of data** for training
- They **don't generalize** to unseen conditions
- They are **hard to interpret**

are a Black Box

Ne



Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks

Brenden Lake [1 2] Marco Baroni [2]

Abstract

Humans can understand and produce new utterances effortlessly, thanks to their compositional skills. Once a person learns the meaning of a new verb "dax ... derstand the m... dax." In this p... main, consistin... navigation com... ing action seq... generalization

then dax again." This type of compositionality is central to the human ability to make strong generalizations from very limited data (Lake et al., 2017). In a set of influential and controversial papers, Jerry Fodor and other researchers have

jump	⇒	JUMP
jump left	⇒	LTURN JUMP
jump around right	⇒	RTURN JUMP RTURN JUMP RTURN JUMP RTURN JUMP
turn left twice	⇒	LTURN LTURN
jump thrice	⇒	JUMP JUMP JUMP
jump opposite left and walk thrice	⇒	LTURN LTURN JUMP WALK WALK WALK
jump opposite left after walk around left	⇒	LTURN WALK LTURN WALK LTURN WALK LTURN WALK LTURN LTURN JUMP

APPROXIMATING CNNS WITH BAG-OF-LOCAL-FEATURES MODELS WORKS SURPRISINGLY WELL ON IMAGENET

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

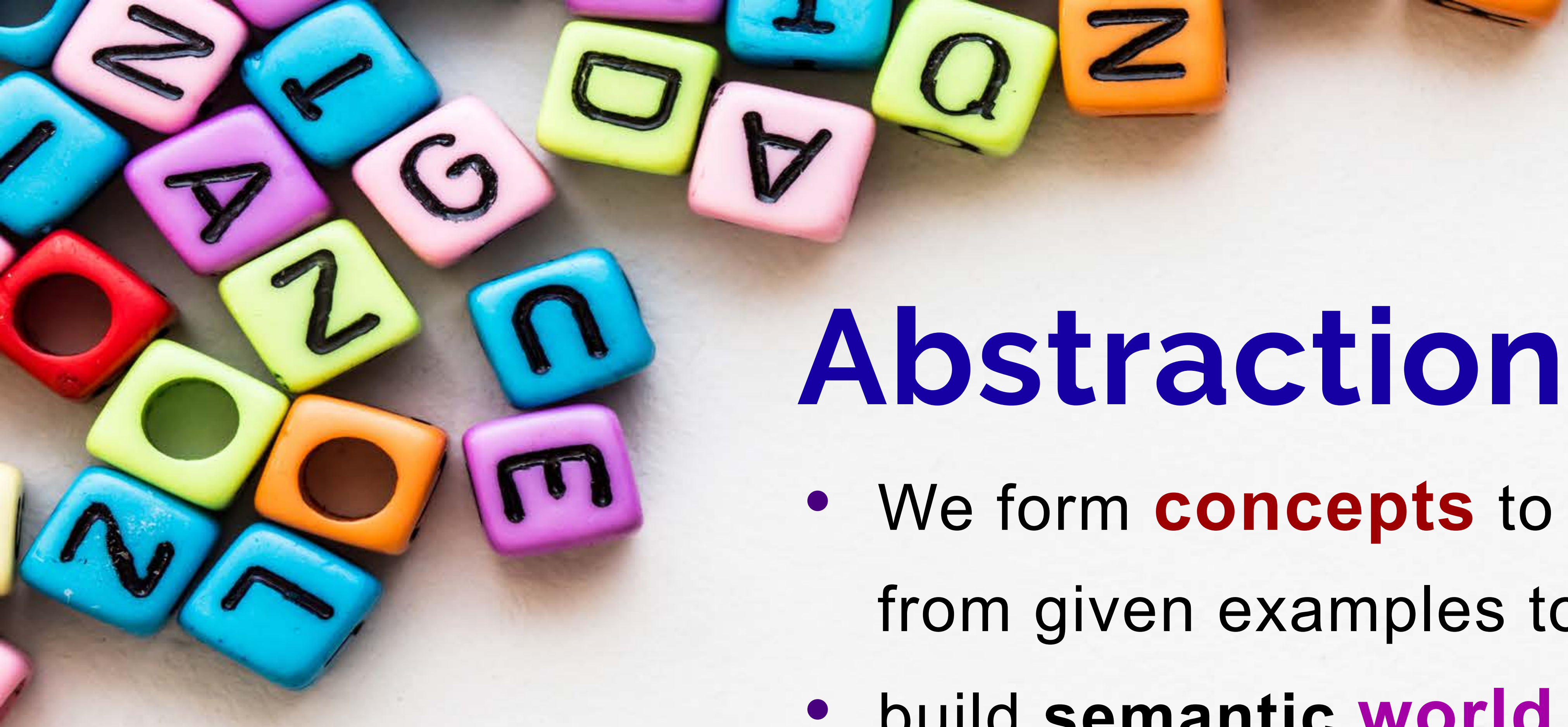
71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Neural models are more biased towards texture and local patterns rather than shape and global structure

Abstraction

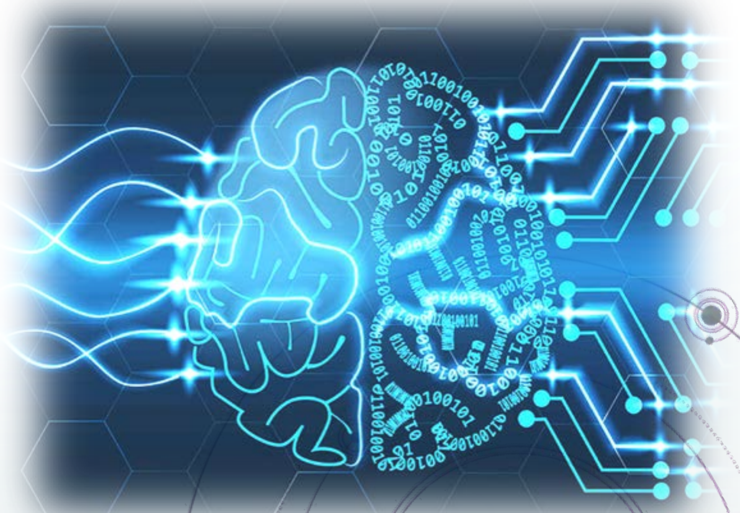


Abstractions

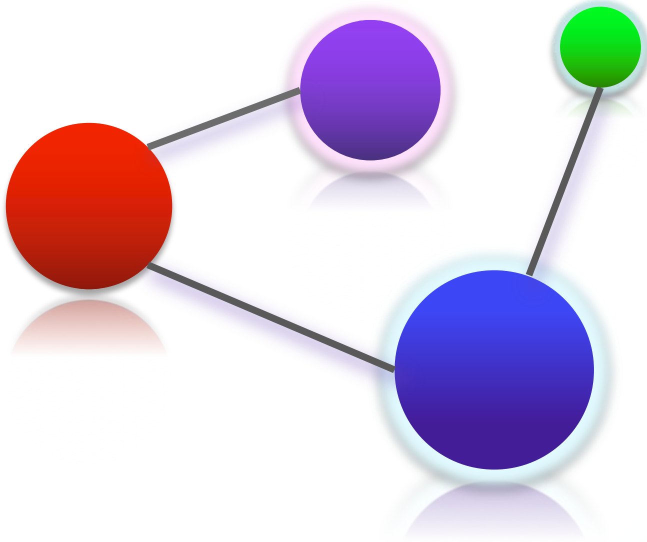
- We form **concepts** to **generalize** from given examples to new ones
- build **semantic** **world models** to represent our environment
- draw **inferences** to proceed from **facts to conclusions**

The Neural State Machine

- A **differentiable graph-based** model the simulates the operation of a **state machine**.
- Combines the strengths of **neural** and **symbolic** approaches.
- Explore the model in the context of **visual reasoning** and **question answering**.



The Neural State Machine



Two stages of **learning** and **inference:**

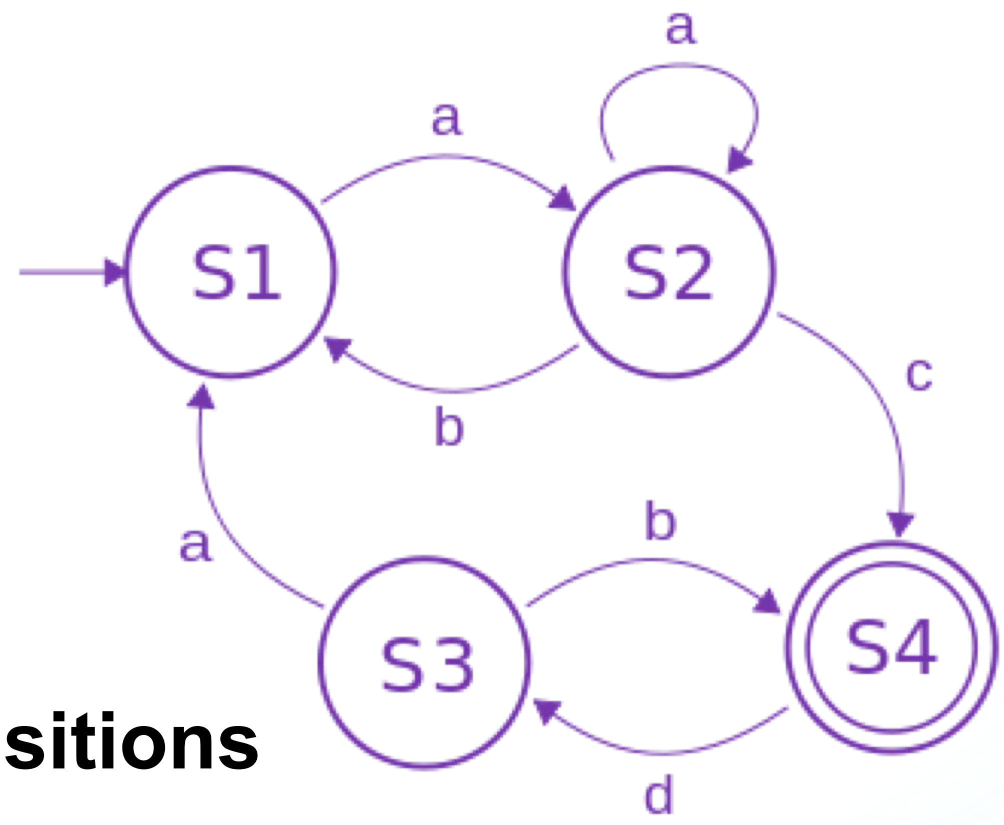
1) **Modeling:** transforms the raw inputs into **abstract** semantic representations, and ***construct the state machine***.

   Image → Scene graph, Question → Instructions

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

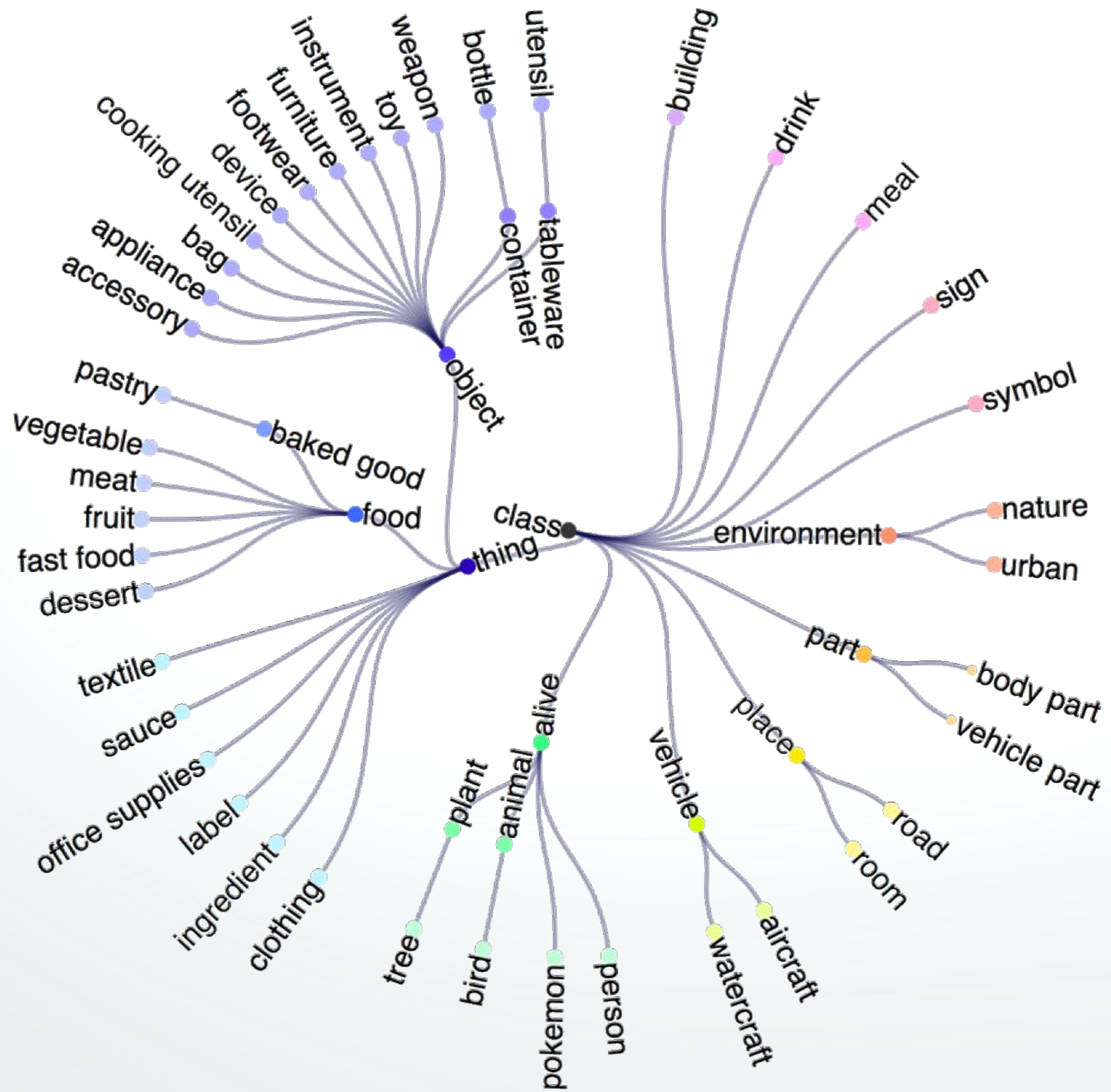
   Reasoning over the scene graph to compute an answer

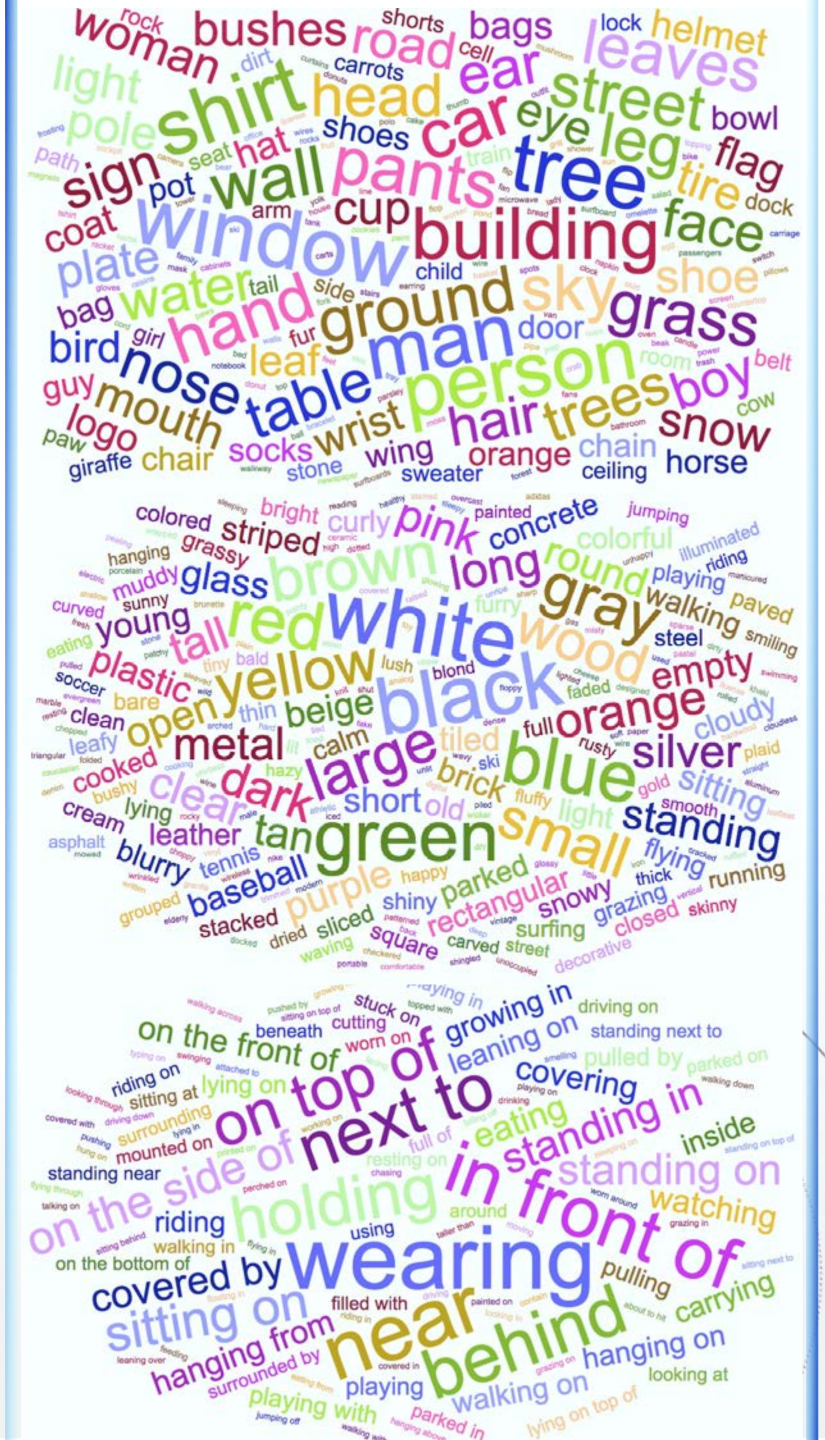
Formal Definition



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

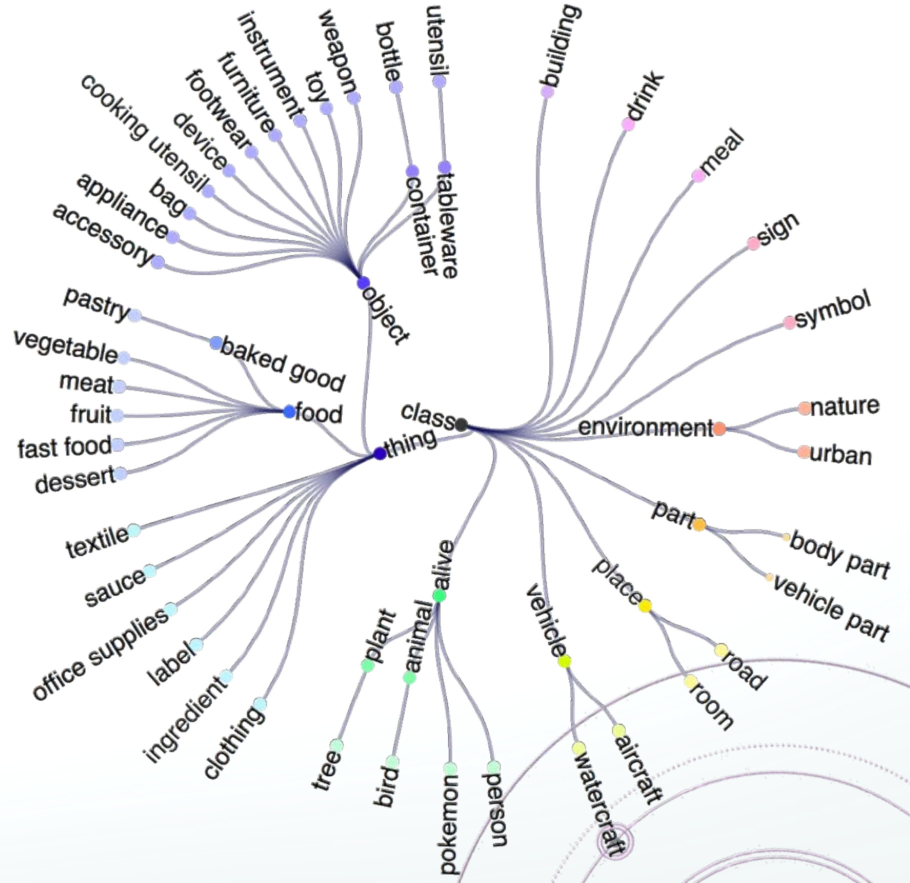
Concepts Vocabulary



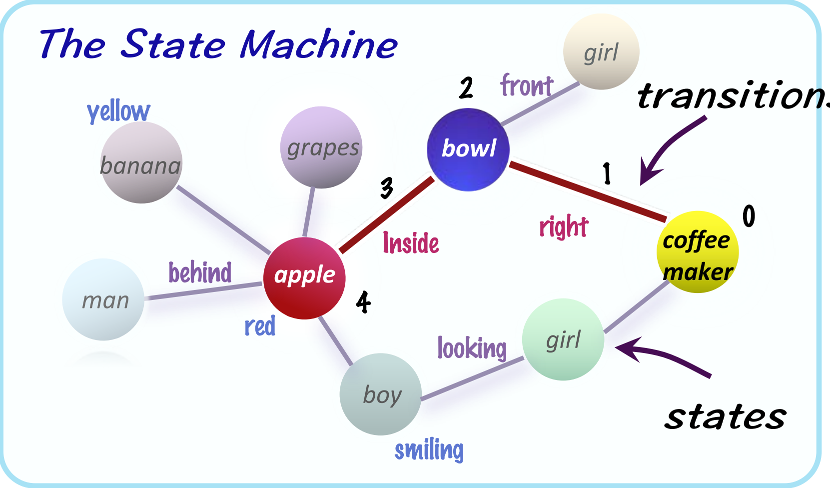


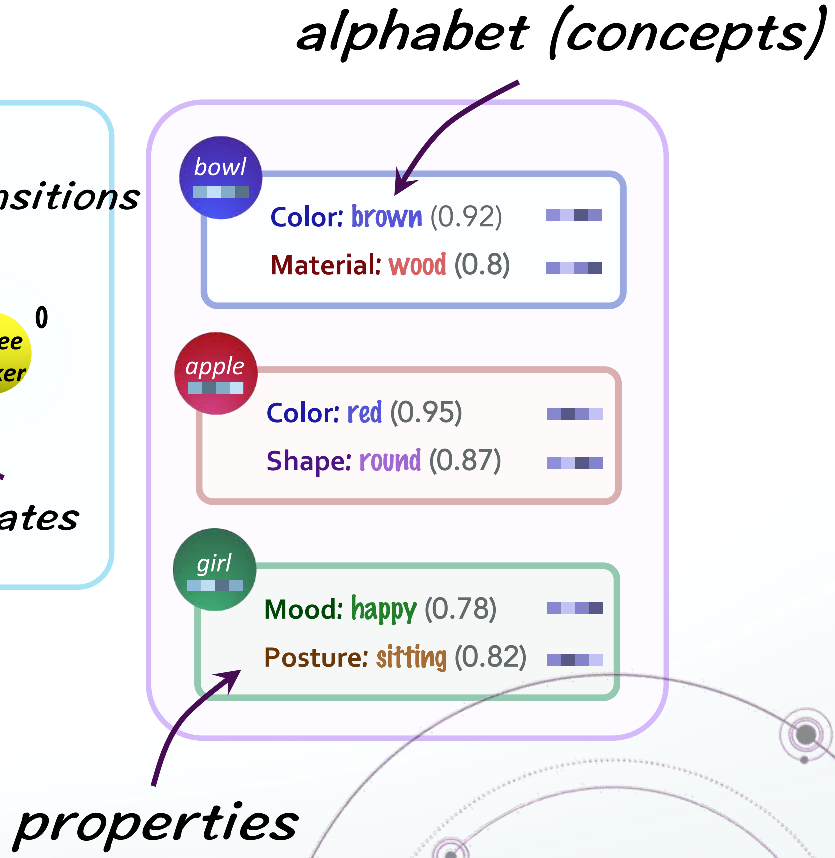
Concepts Vocabulary

- The model operates over a vocabulary of **embedded concepts**, **atomic semantic units** that **represent** aspects of **the world**.
- **Translate** both **modalities** (image and question) to "speak the same language".
- **Abstraction over the raw dense features**
- Inspired by **concept learning** in humans (cognitive science)



The Neural State Machine *for VQA*





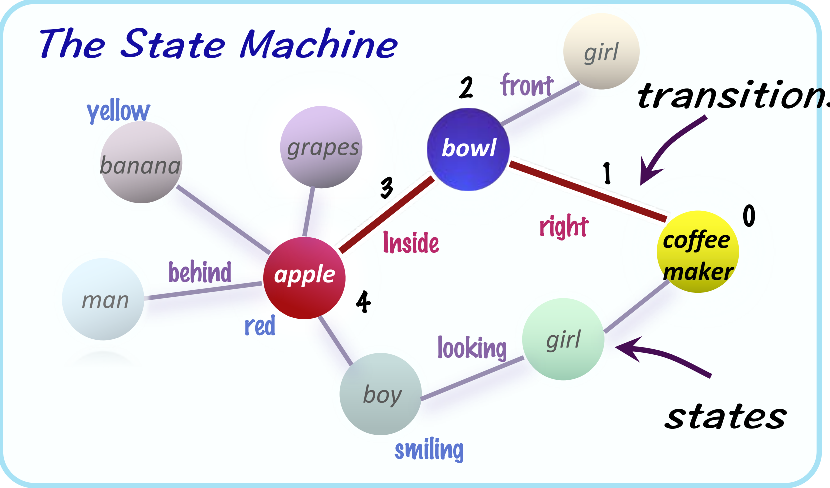
Given an **image**, we first construct a **scene graph**

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

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

The Neural State Machine *for VQA*

alphabet (concepts)

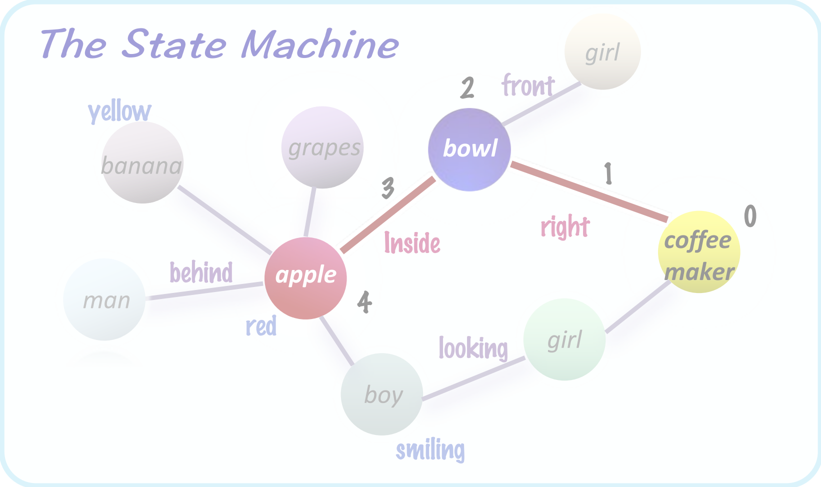


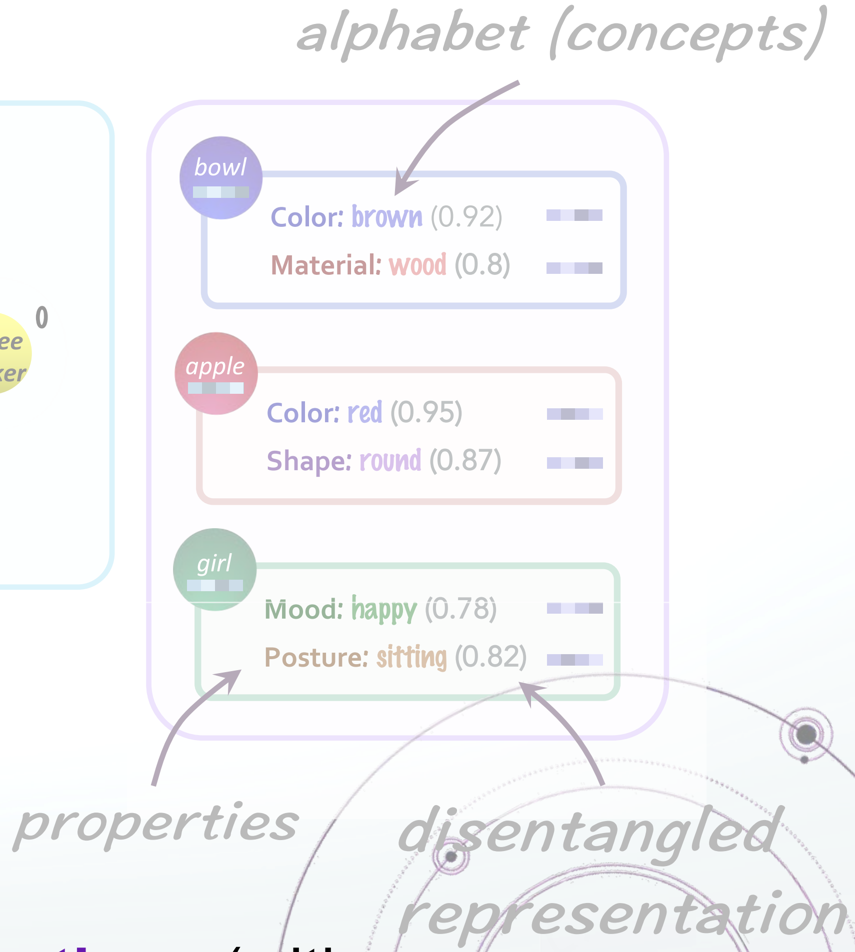
Objects are represented through a **factorized distribution** over **semantic properties** *(color, shape, material)*, defined over the **concept vocabulary**.

properties

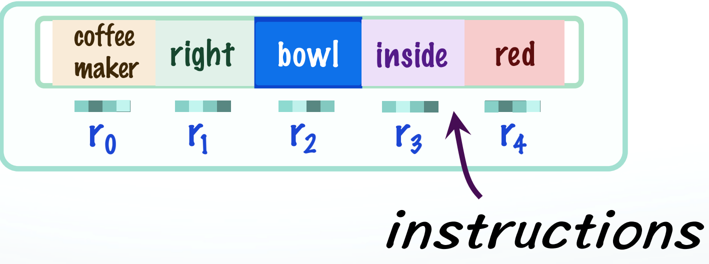
disentangled representation

The Neural State Machine *for VQA*





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



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

Abstract Decoder

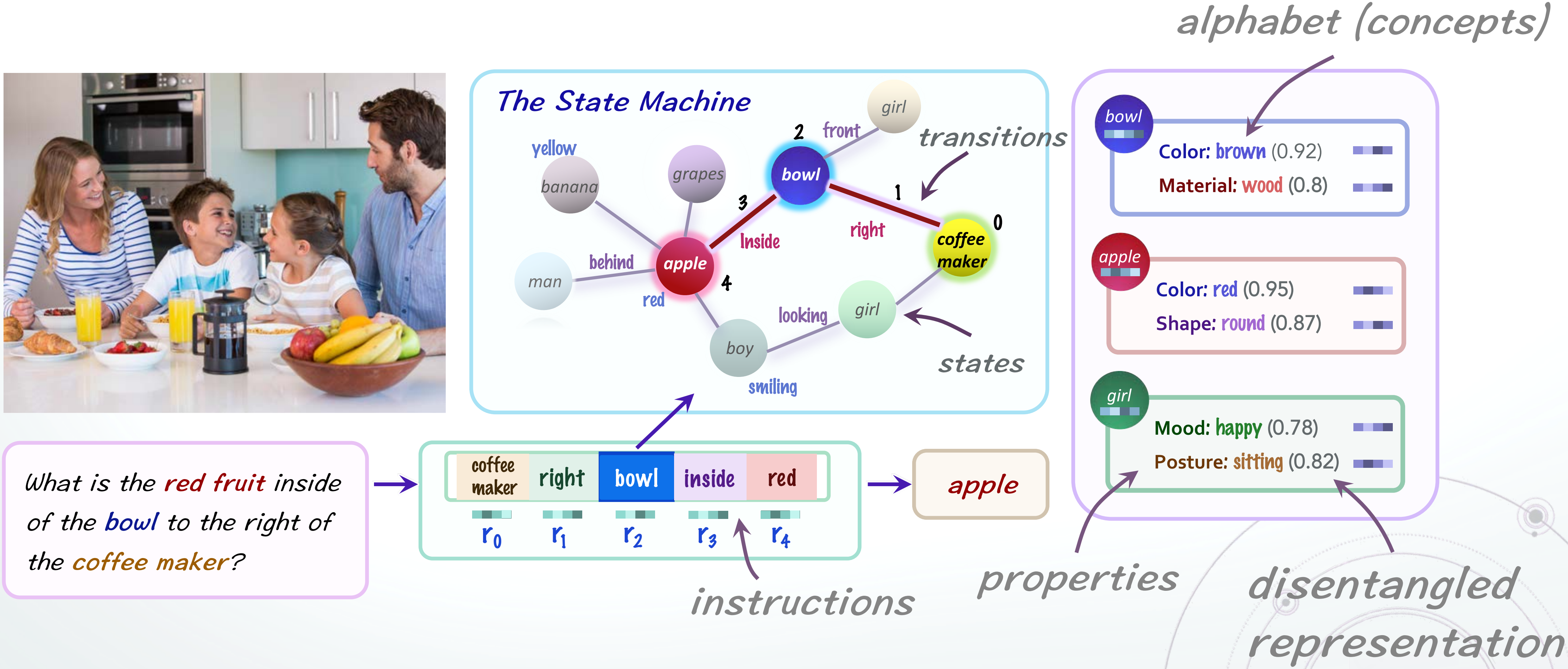
What color is the cat behind the red chair?

What ATTR is the OBJ REL the ATTR OBJ?

OBJ	ATTR	REL	OBJ	ATTR
chair	red	behind	cat	color

Disentangles structure and content

The Neural State Machine *for VQA*



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

Machine Simulation *(Traversal)*

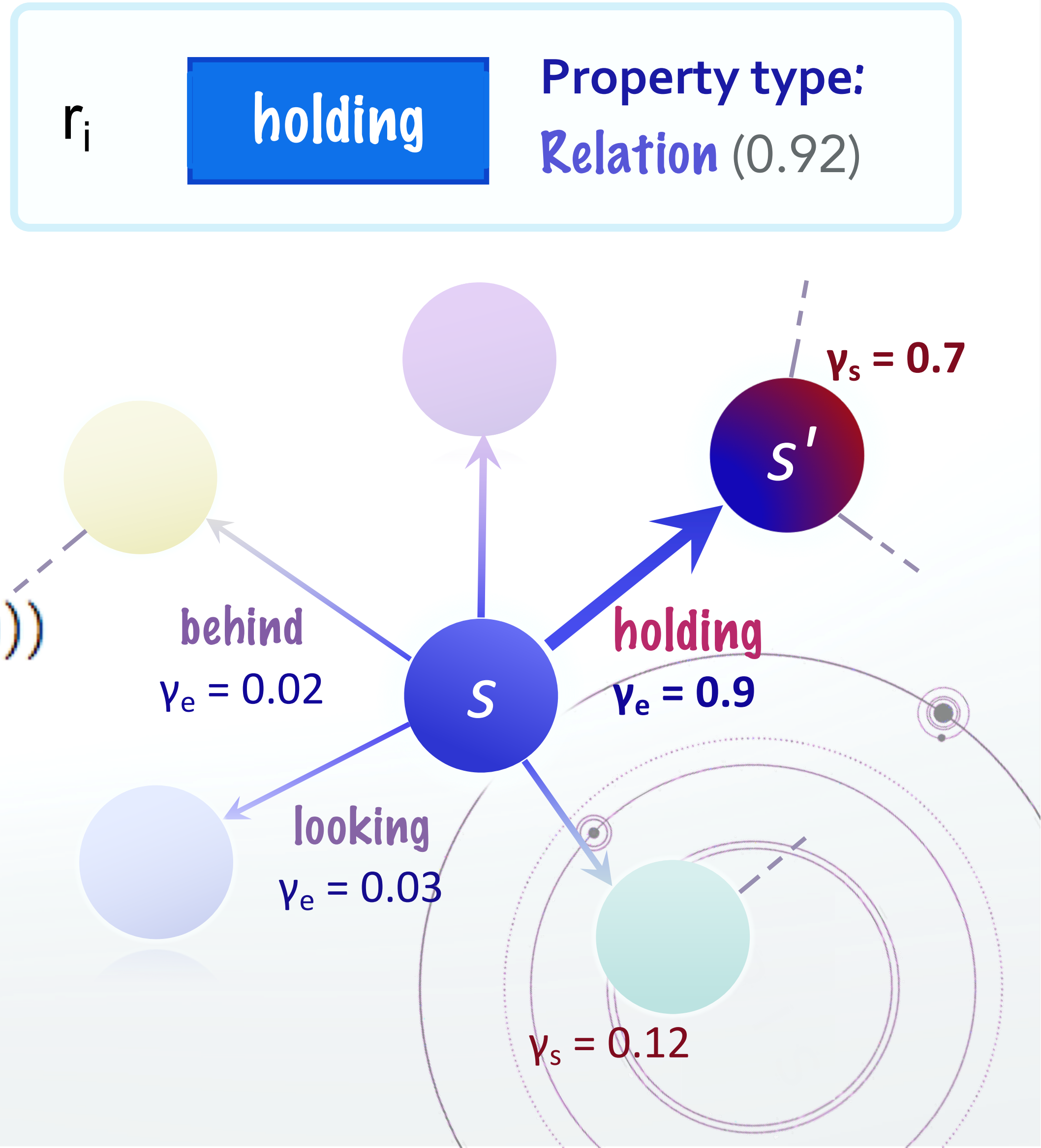
$$\gamma_i(s) = \sigma\Big(\sum_{j=0}^{L} R_i(j)(r_i \circ \mathbf{W}_j s^j)\Big)$$

$$\gamma_i(e) = \sigma\big(r_i \circ \mathbf{W}_{L+1} e'\big)$$

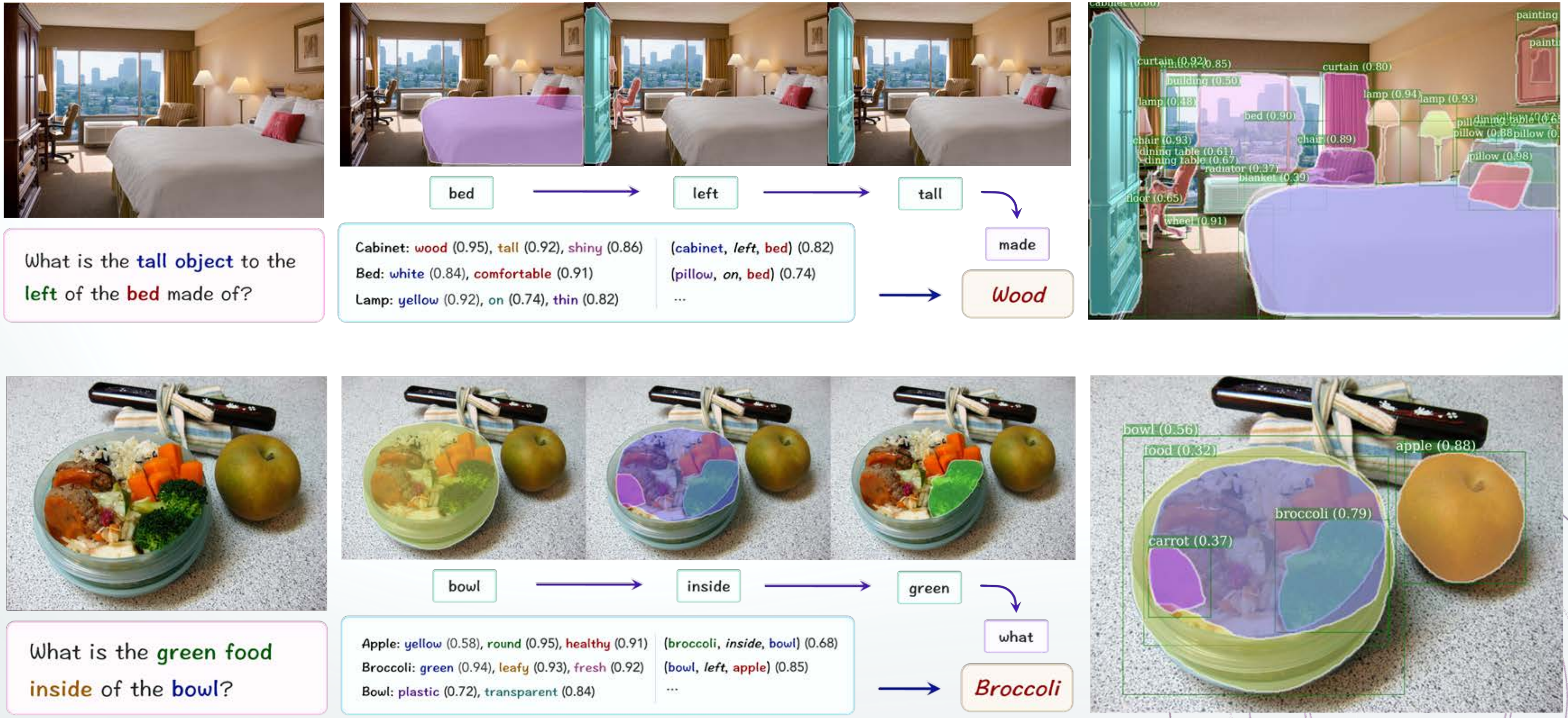
$$p^s_{i+1} = \mathrm{softmax}_{s \in S}(\mathbf{W}_s \cdot \gamma_i(s))$$

$$p^r_{i+1} = \mathrm{softmax}_{s \in S}\Big(\mathbf{W}_r \cdot \sum_{(s',s)\in E} p_i(s') \cdot \gamma_i((s',s))\Big)$$

$$p_{i+1} = r'_i \cdot p^r_{i+1} + (1 - r'_i) \cdot p^s_{i+1}$$



Qualitative Results



Question Examples

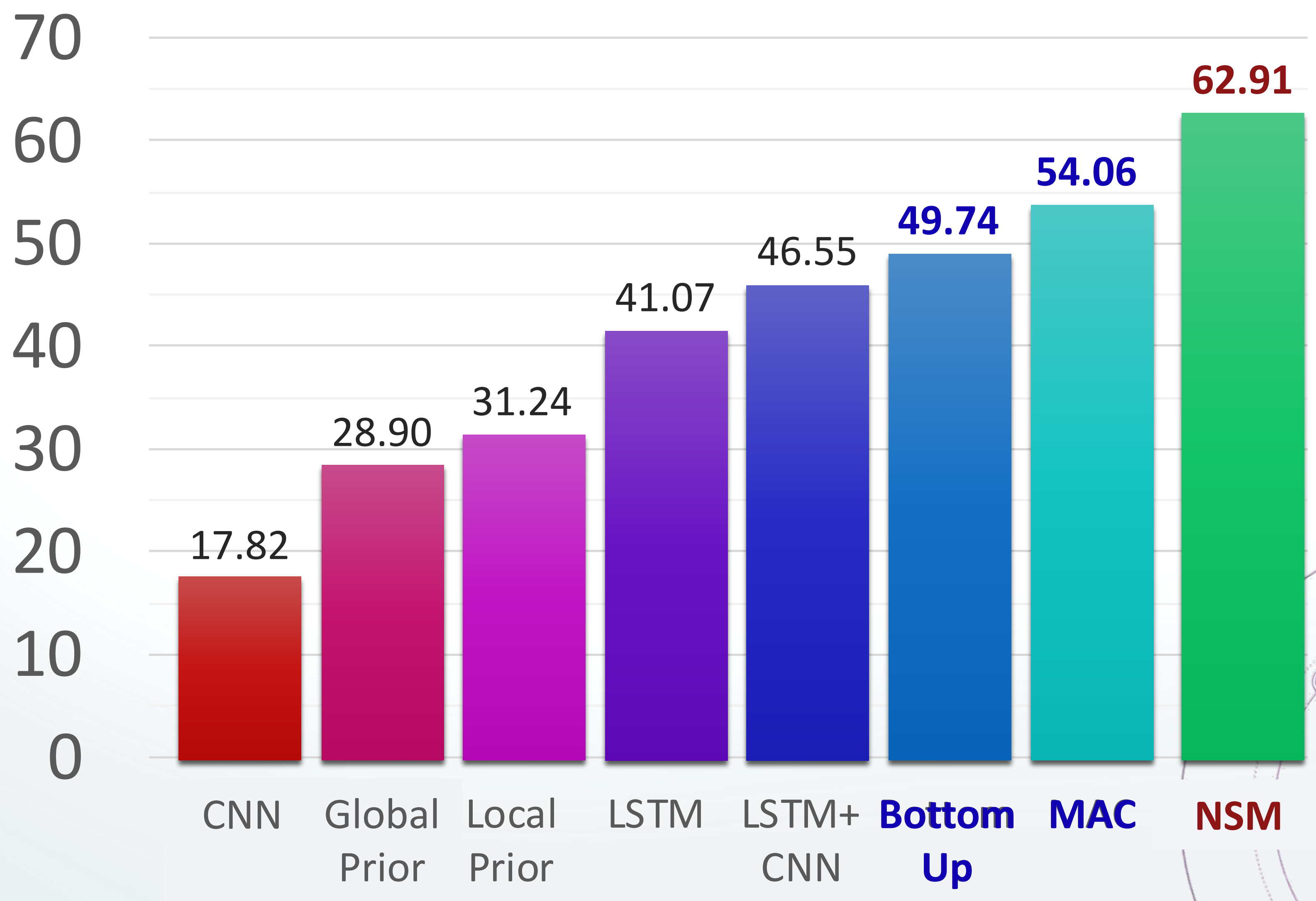
1) What is the giraffe looking at? ***person*** ✓
2) Is the fence in front of the giraffe made of metal? ***no*** ✓
3) Is the woman's shirt blue or yellow? ***blue*** ✓
4) On which side of the image is the person? ***right*** ✓
5) Is there a child behind the giraffe? ***no*** ✗

1) What is the fruit to the right of the salad? ***strawberries*** ✓
2) Is the fork to the right of the salad? ***no*** ✓
3) Is the plate white and square? ***no*** ✓
4) Is the cup behind the round plate? ***yes*** ✓
5) What is the plate made of? ***paper*** ✗

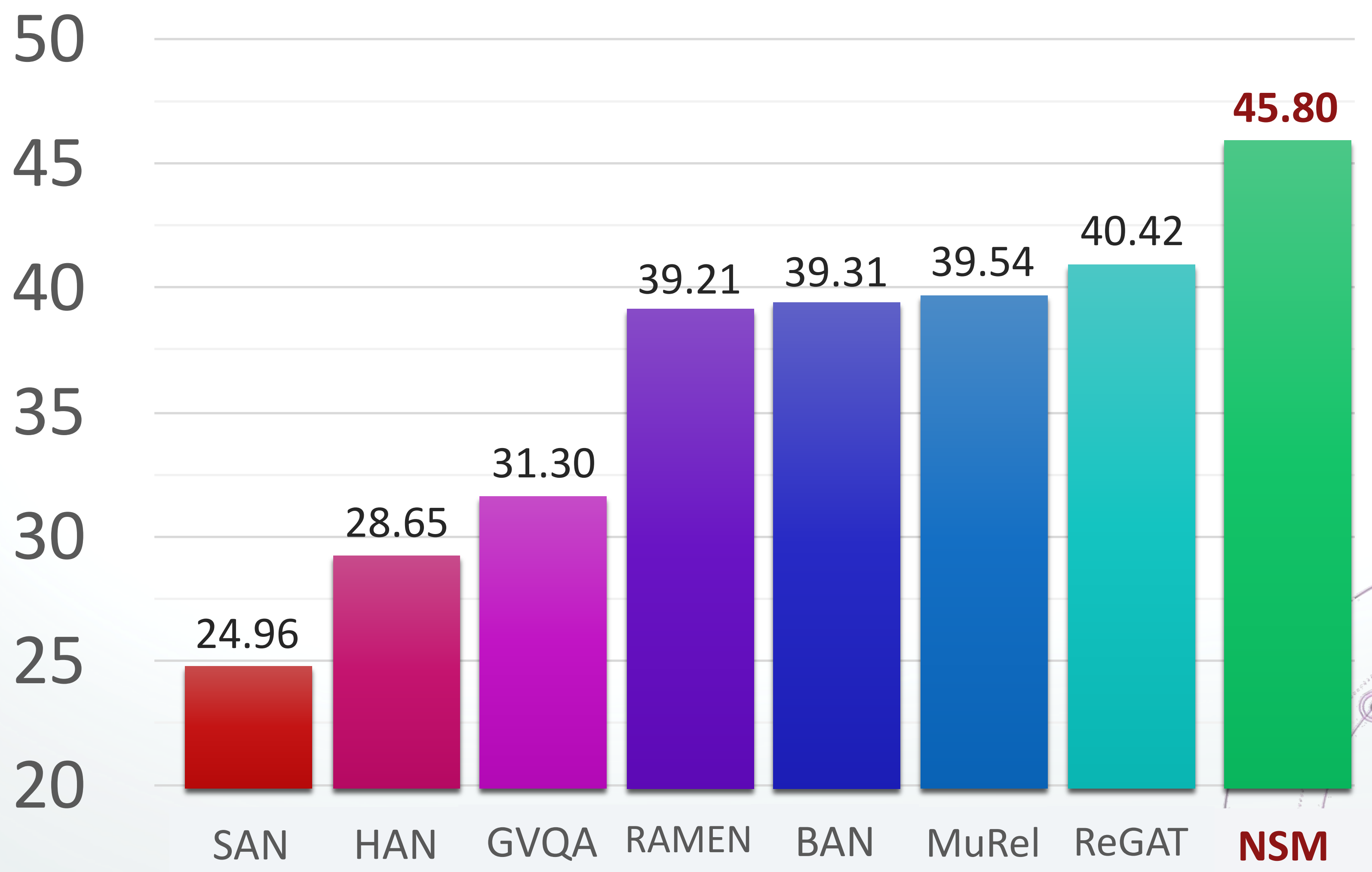
1) Are there either scarves or hats that are not pink? ***no*** ✓
2) Do the bear's dress and the person's shirt have the same color? ***yes*** ✓
3) Is the bear sitting or standing? ***sitting*** ✓
4) What is the green object that the bear is sitting on? ***book*** ✓
5) Is the bear wearing white shoes? ***yes*** ✗

1) Are there either a chair or a clock in the image? ***no*** ✓
2) Are there any flowers behind the bed on the left of the room? ***yes*** ✓
3) What color is the appliance on the right? ***black*** ✓
4) Is the carpet brown or blue? ***brown*** ✓
5) Is the TV turned on? ***yes*** ✗

Quantitative Results (GQA)



Generalization (VQA-CP)



New Generalization Splits

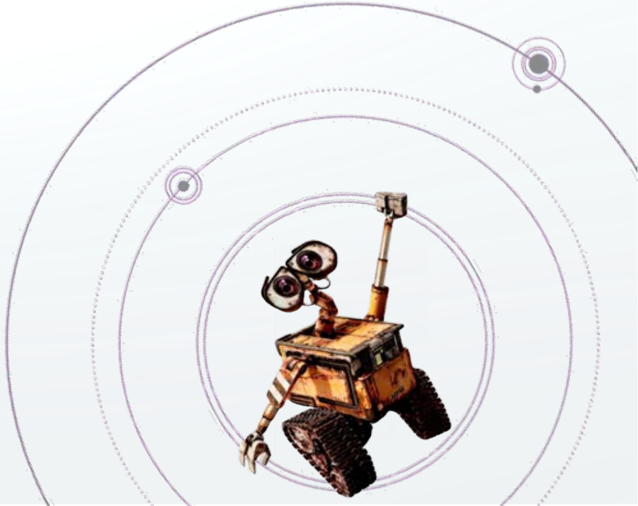
	training	testing
structure	What is the <obj> **covered by**? **Is there a** <obj> in the **image**? What is the <obj> **made of**? **What's the name** of the <obj> **that is** <attr>?	What is **covering the** <obj>? **Do you see any** <obj>s in the **photo**? What **material makes up** the <obj>? **What 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)

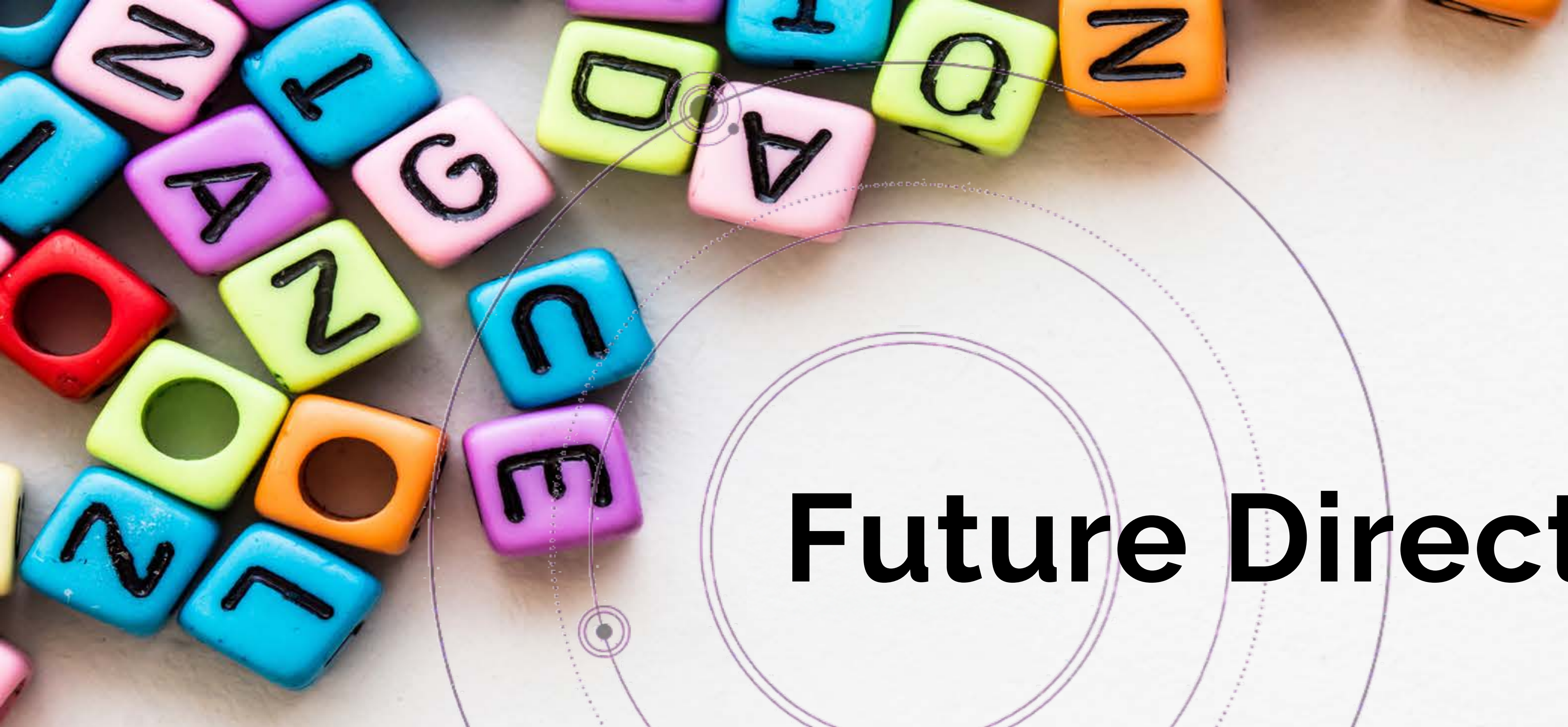
Generalization Results

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**

Summary

- **Construct** and **simulate** a neural state machine.
- A **neural traversal** over the **scene graph** guided by the **instructions** derived from the **questions** – **Sequential graph-based reasoning**.
- **Both** visual and linguistic **modalities** are transformed into the shared abstract language of concepts.
- Combines the strengths of **abstraction** and **compositionality**.
- *A neural implementation of a classical model of computation!*





Future Directions

Thank you! ☺