



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 9.1: Reasoning – Model Design

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Major revision co-developed with Paul Liang. Co-lecturers
included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

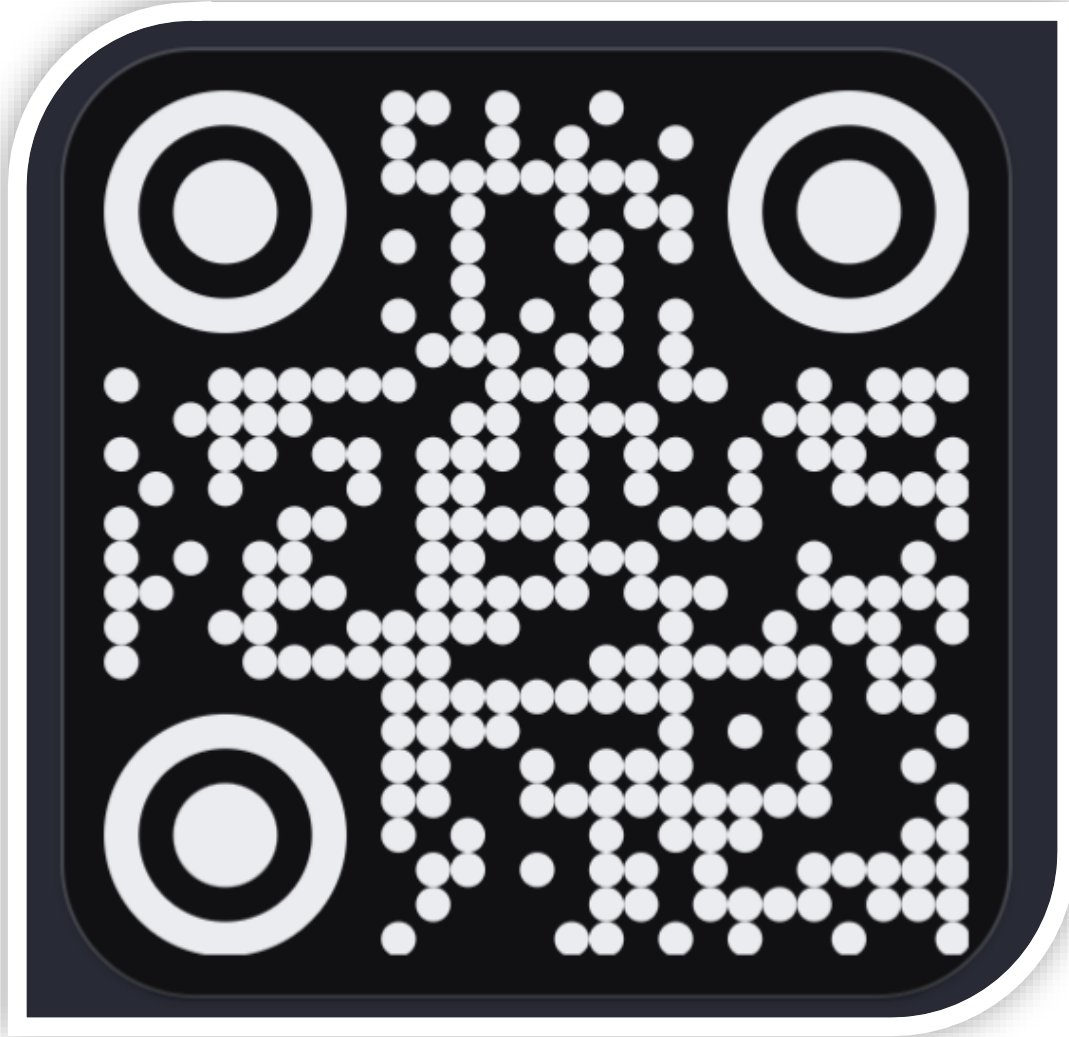
Midterm project report (due this Sunday 10/25 at 8pm ET)

- For each N-1 research idea, implement a first (baseline) model
 - The emphasis is on the result discussion (to update research ideas)
- Read carefully the instructions!!

Midterm project presentations (next week)

- In-class presentations: Tuesday 10/27 and Thursday 10/29
 - One presentation per team, maximum of N+1 minutes
 - Focus on the updated research ideas (from report)
- Attendance required for both sessions
 - Give feedback to at least 6 teams per session
- Upload your slides on shared Google Drive
 - Same laptop used for all presentations

Attendance using Poll Everywhere



- 1 Scan QR Code
or
<https://pe.app/mmm1>
- 2 Enter your AndrewID
(as your “name tag”)
- ✗

Do not try to sign-in
- 3 Click “Share location”



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 9.1: Reasoning – Model Design

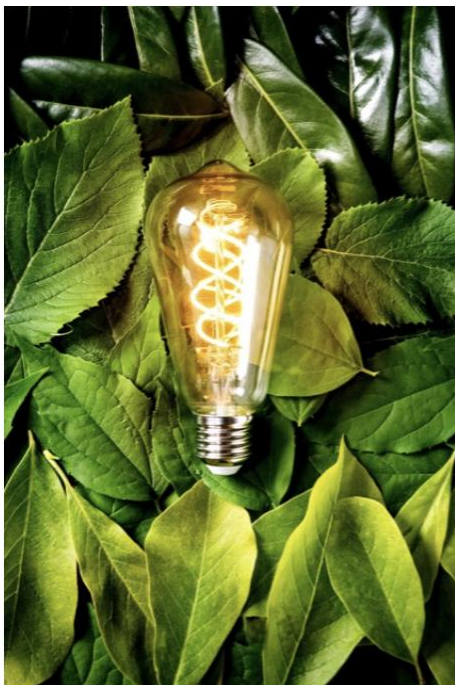
Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Major revision co-developed with Paul Liang. Co-lecturers
included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

- Reasoning: definition and four sub-challenges
 - Compositionality as a challenge for multimodal models
- Structure modeling
 - Graphs and graph neural networks
 - Module networks: end-to-end learning, neural-symbolic VQA
 - The Neural State Machine, and visual programming
- Intermediate concepts
 - Hard attention, concept alphabets, language
- External knowledge
 - Knowledge graphs, and retrieval-augmented models

The Challenge of Compositionality



(a) some plants
surrounding a
lightbulb

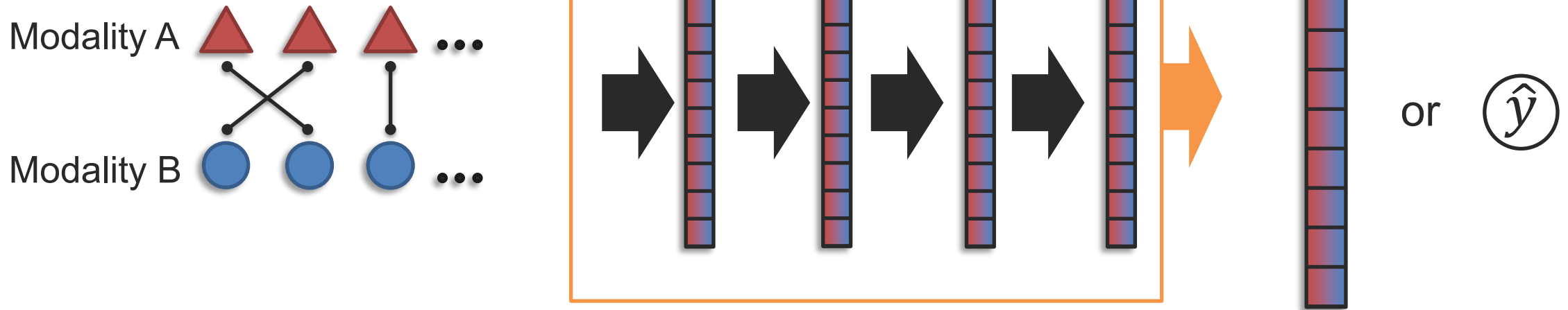


(b) a lightbulb surrounding some plants

[Thrush et al., Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. CVPR 2022]

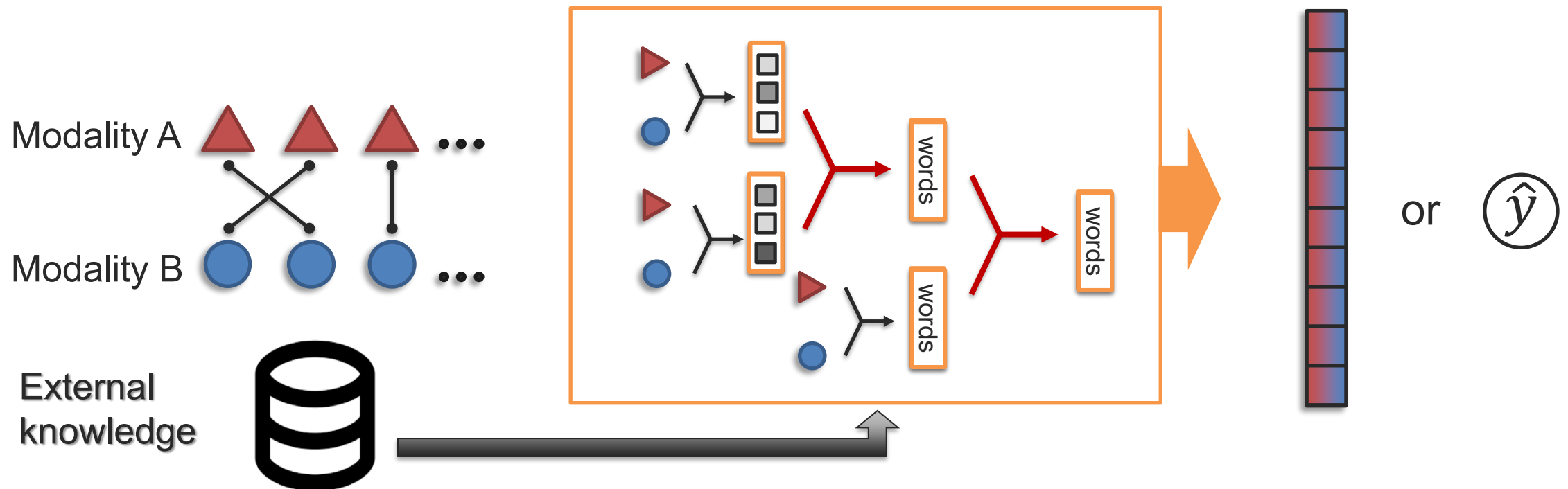
Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

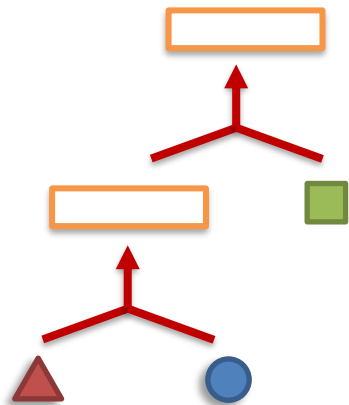


Reasoning

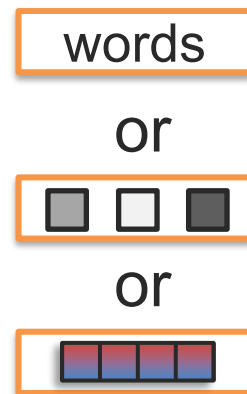
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

Sub-challenges:

Structure Modeling



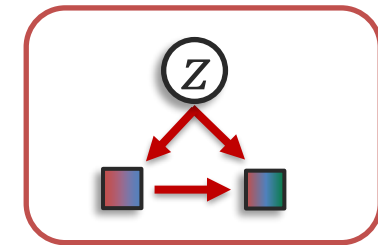
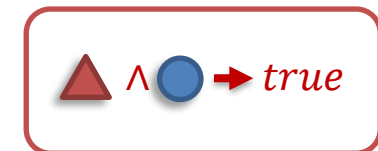
Intermediate concepts



External Knowledge



Inference Paradigm



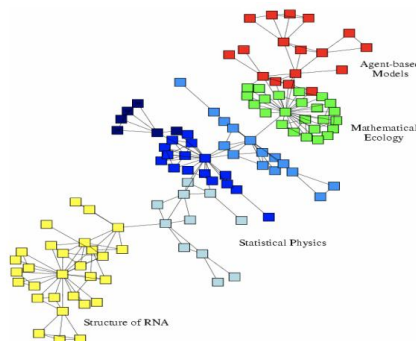
Structure Modeling

(... beyond sequences)

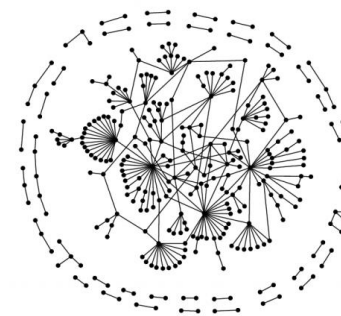
Structure Beyond Sequences: Trees and Graphs



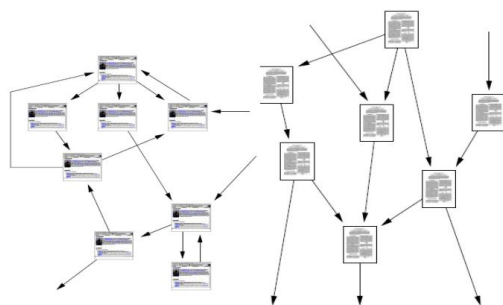
Social networks



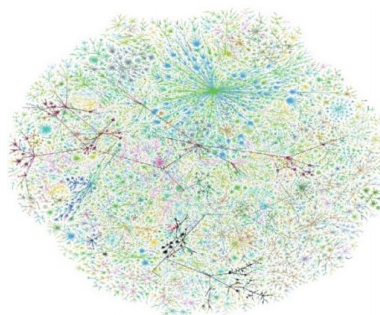
Economic networks



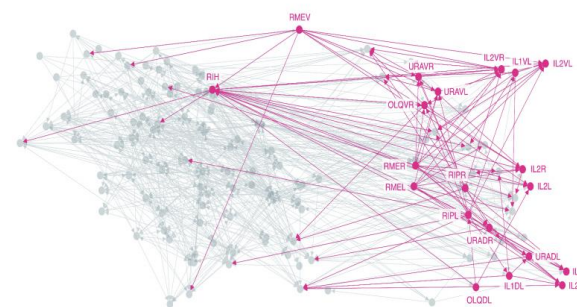
Biomedical networks



Information networks:
Web & citations



Internet



Networks of neurons

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

Assume we have a graph $\mathbf{G}$:

$\mathbf{V}$ is the set of vertices

$\mathbf{A}$ is the binary adjacency matrix

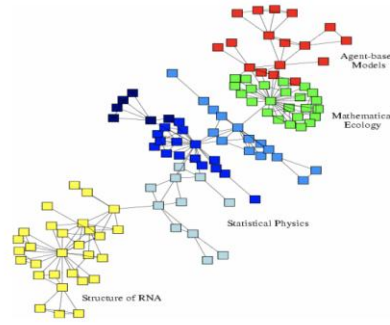
$\mathbf{X}$ is a matrix of node features:

- Categorical attributes, text, image data
e.g. profile information in a social network
- ...

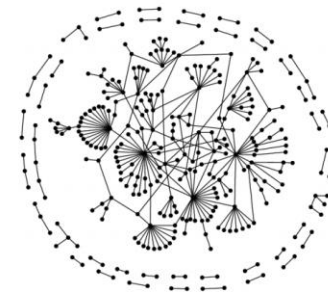
$\mathbf{Y}$ is a vector of node labels (optional)



Social networks



Economic networks

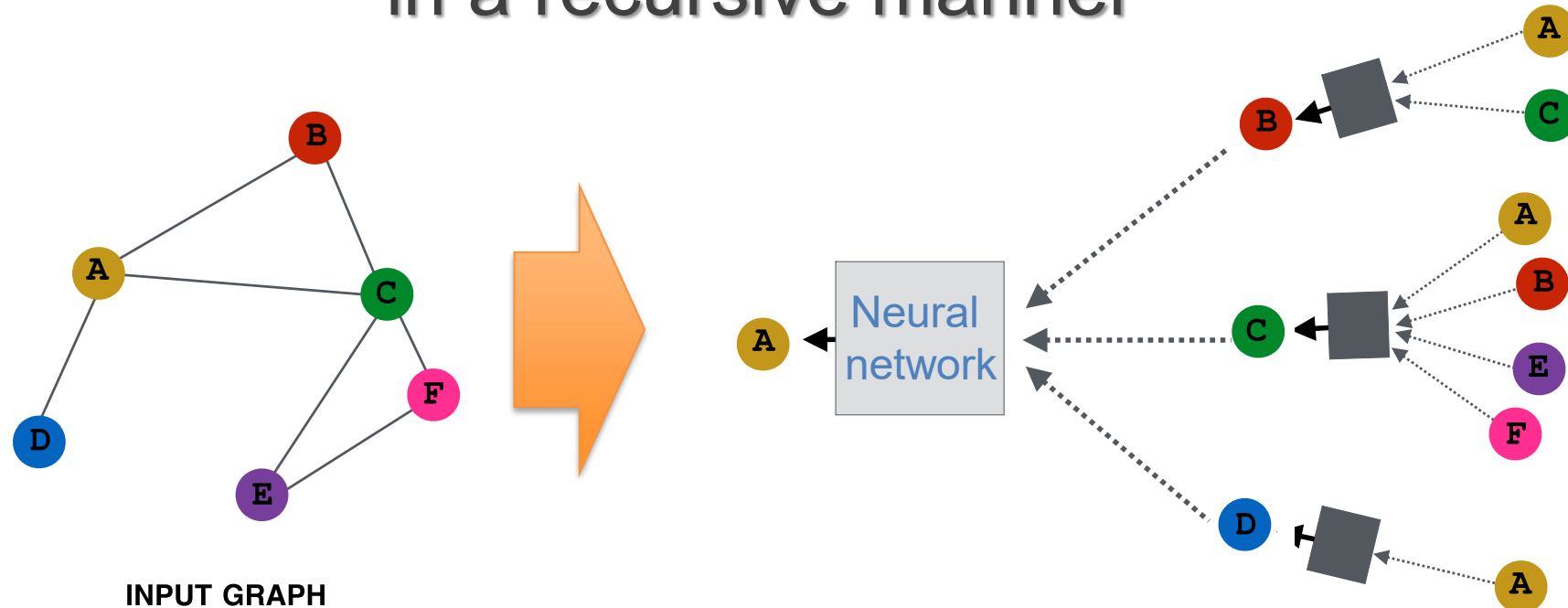


Biomedical networks

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

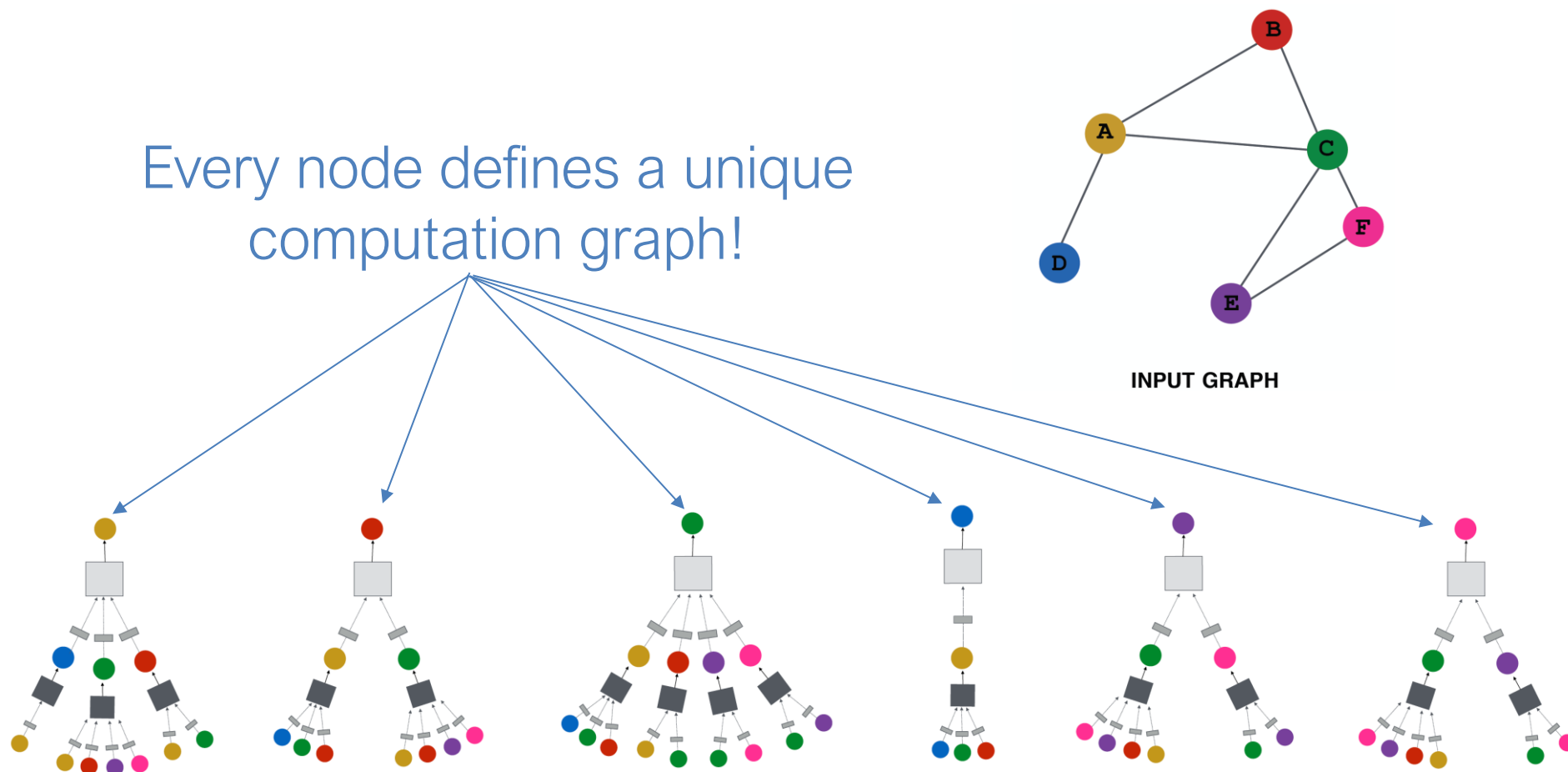
Key idea: Generate node embeddings based on local neighborhoods in a recursive manner



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

Every node defines a unique computation graph!

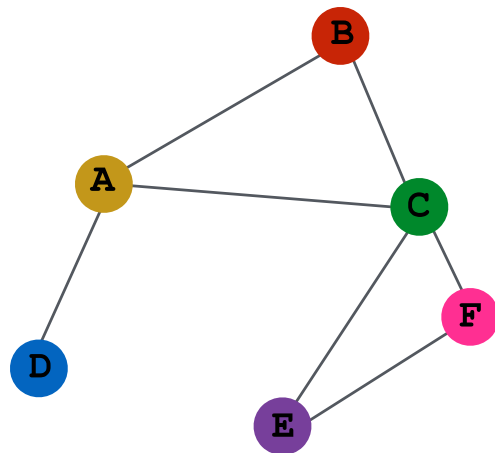


[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

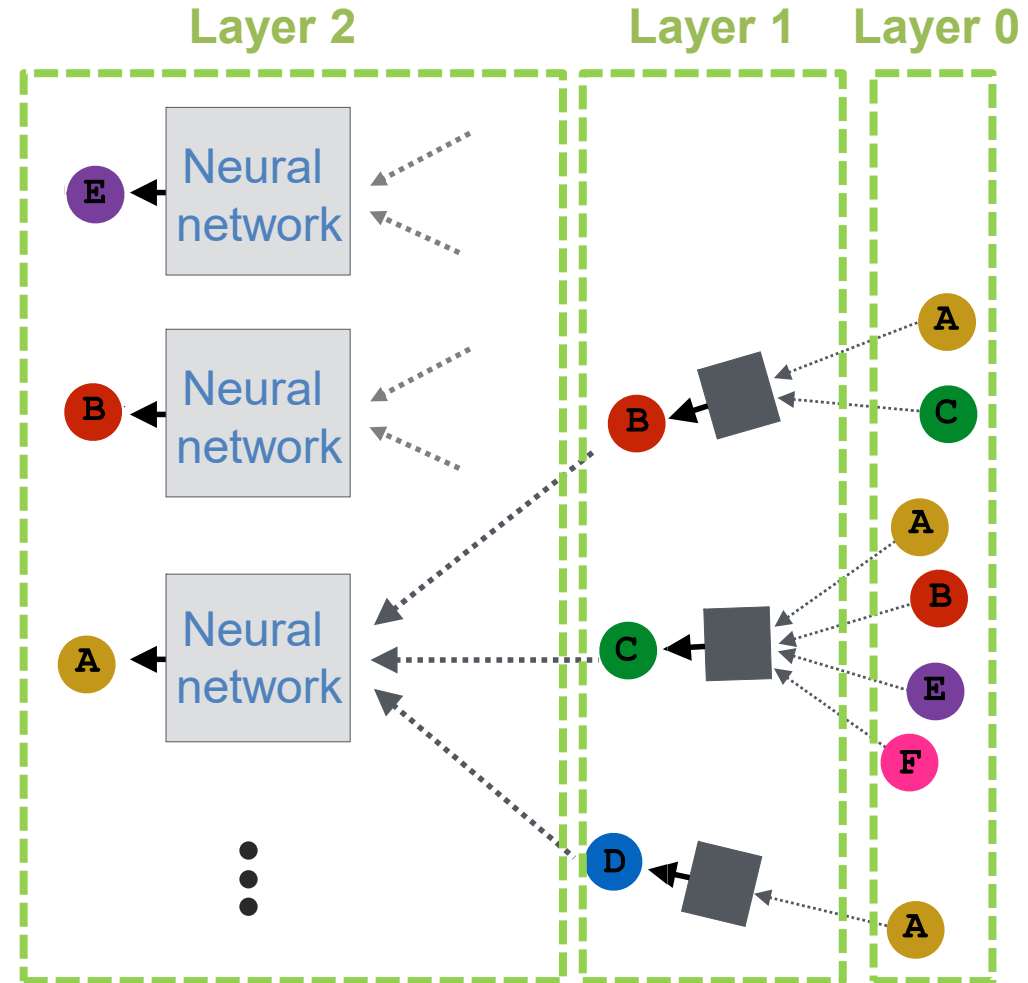
Graph Neural Nets

And multiple layers!

- ➡ Shared parameters within a specific layer
- ➡ “layer-0” is the input feature x_u

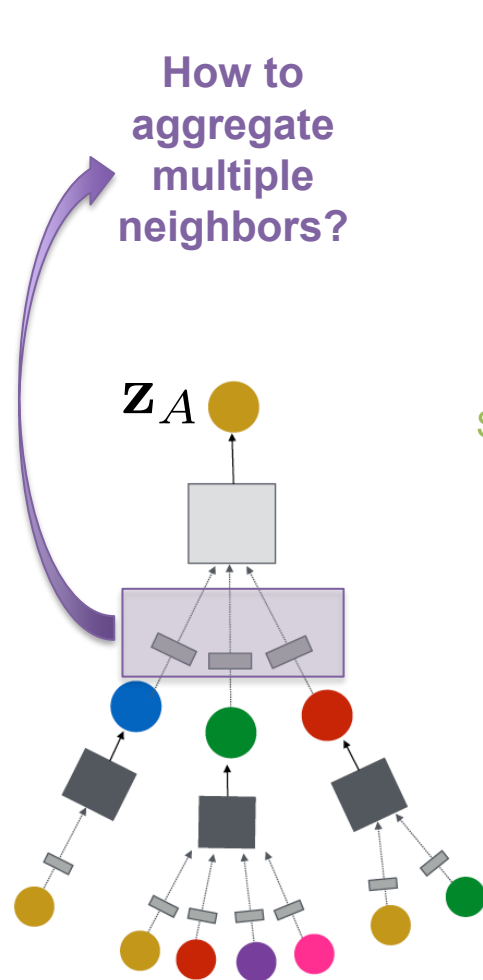


INPUT GRAPH



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets – Neighborhood Aggregation



Average pooling (Scarselli et al., 2005)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1} \right)$$

K is num layers

Different weights for neighbors and self

Graph Convolution Network (Kipf et al., 2017)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)| |N(v)|}} \right)$$

Same weights

Different normalization

It can be efficiently implemented

Graph Attention Network (Velickovic et al., 2018)

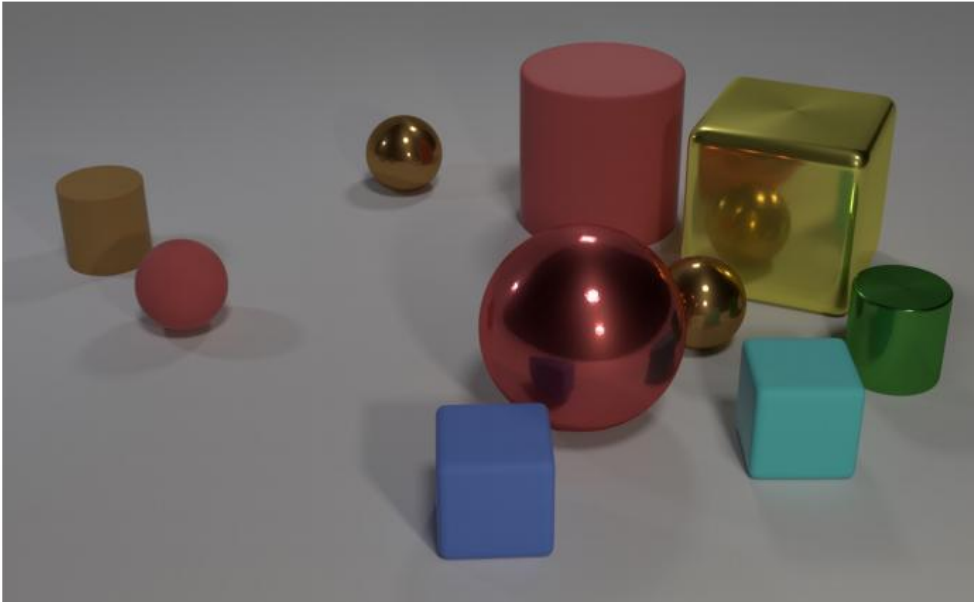
$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\alpha_{uv} \mathbf{h}_u^{k-1}}{\sqrt{|N(u)| |N(v)|}} \right)$$

Attention weights

Very similar to a self-attention transformer

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

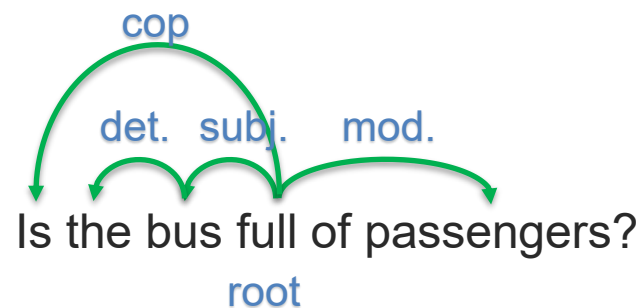
CLEVR: Dataset for Visual Reasoning



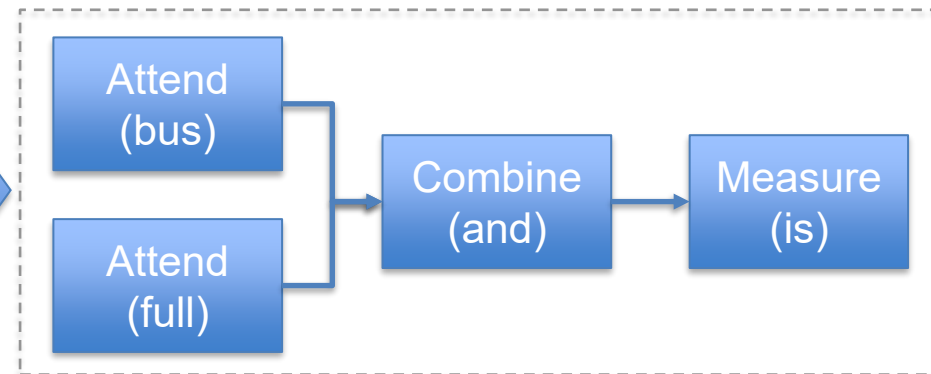
Q: Are there an **equal number** of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

Can we transform text into a graph,
aligned with the image?

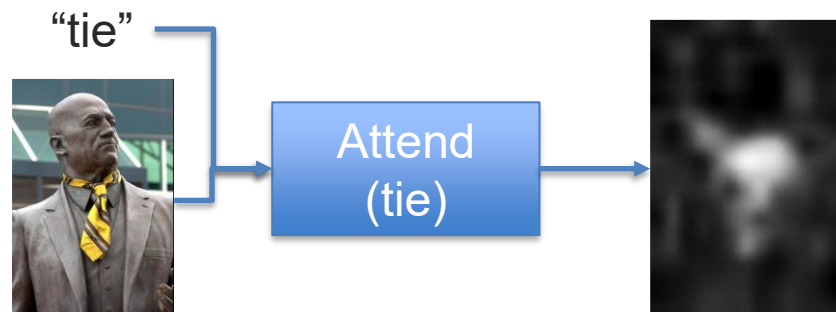
Neural Module Network



Computation layout



Each module work on the attention map(s):

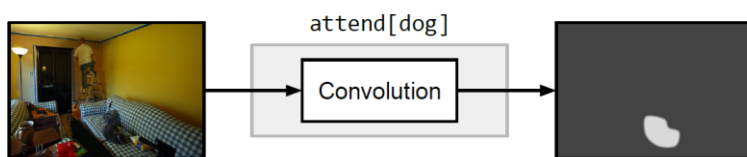


Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

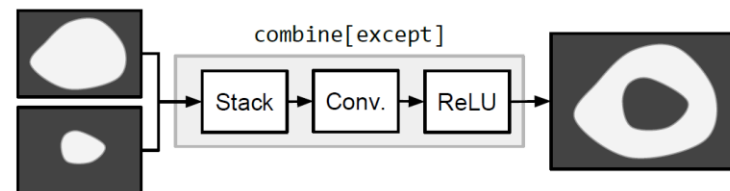
Predefined Set of Modules

1) Analyze the image:

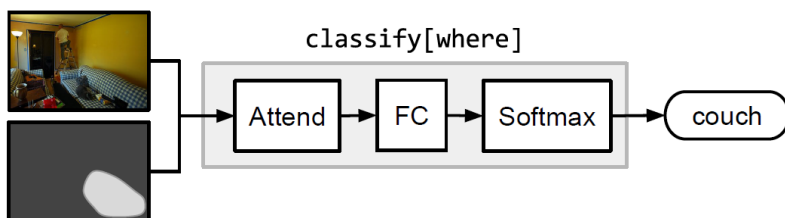
$\text{attend} : \text{Image} \rightarrow \text{Attention}$



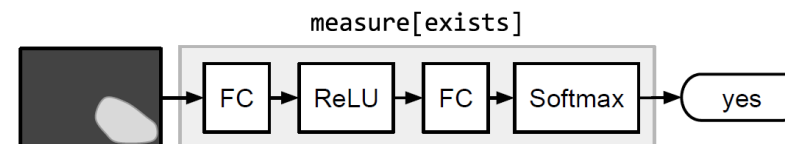
$\text{combine} : \text{Attention} \times \text{Attention} \rightarrow \text{Attention}$



2) Make a prediction

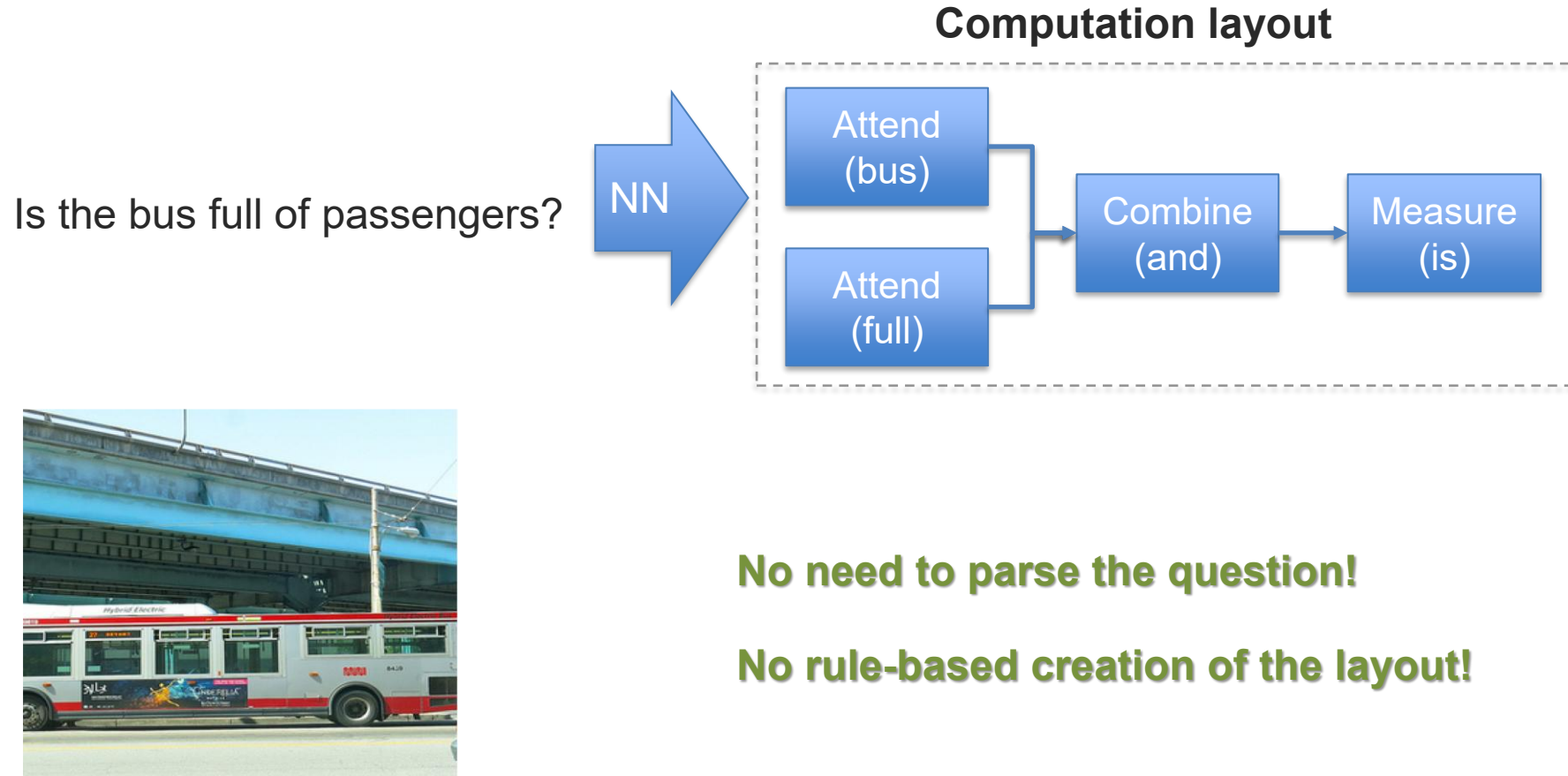


$\text{measure} : \text{Attention} \rightarrow \text{Label}$



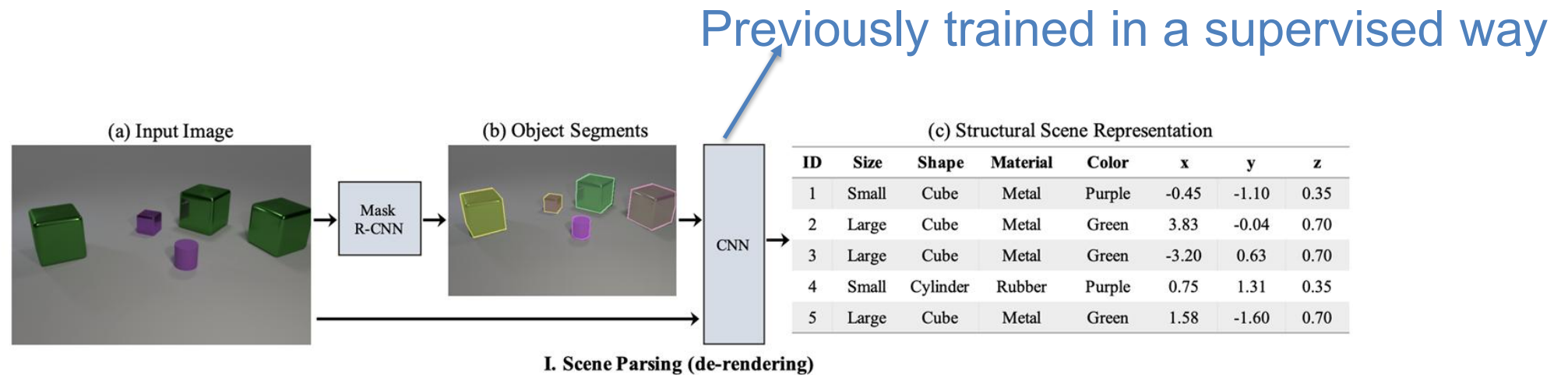
Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

Module Network V2: End-to-End Learning



Module Network V3: Neural-symbolic VQA

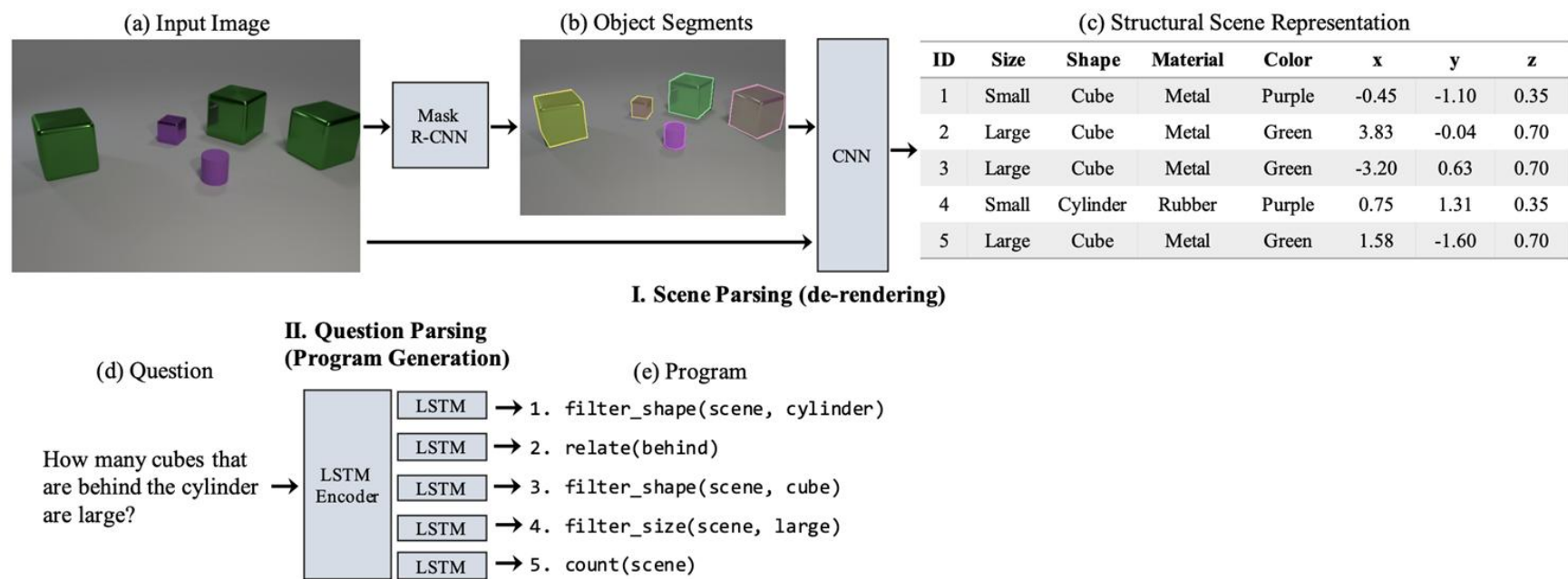
1) Image Attributes



Module Network V3: Neural-symbolic VQA

2) Parsing questions into programs

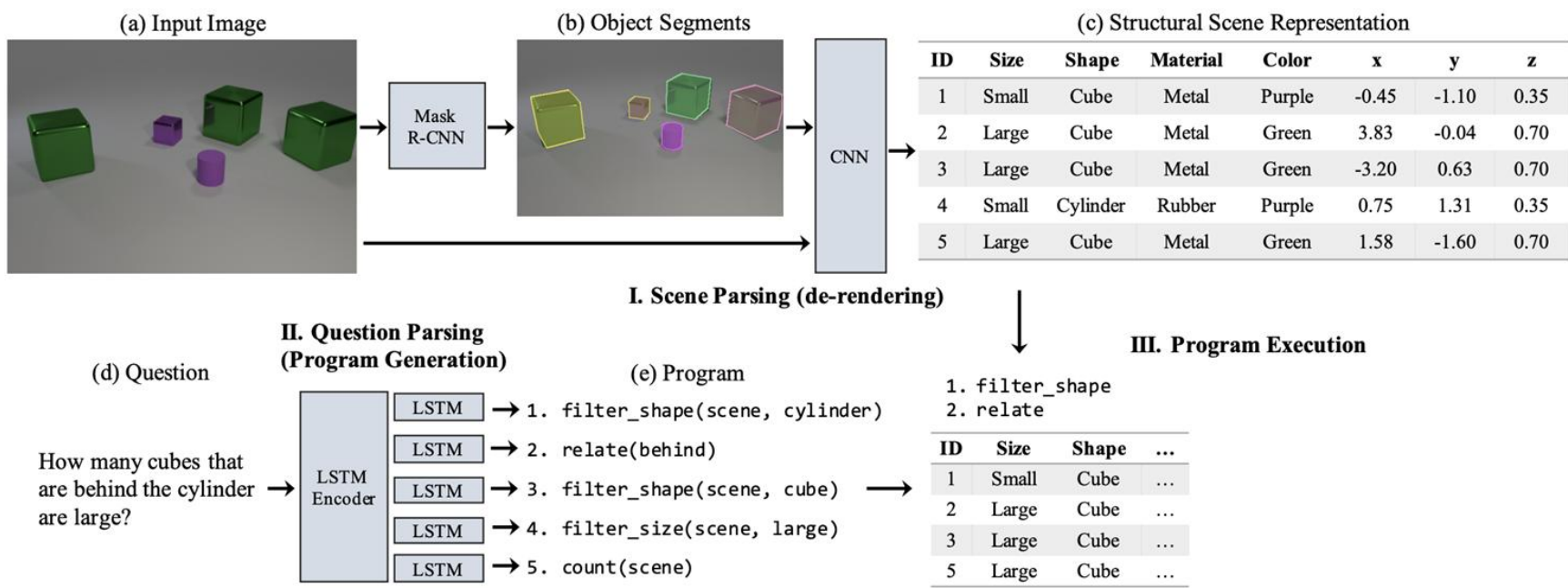
Similar to neural module networksk



Module Network V3: Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the “symbolic” representation of the image



Module Networks V3b: The Neural State Machine

How to solve this question
using visual reasoning?

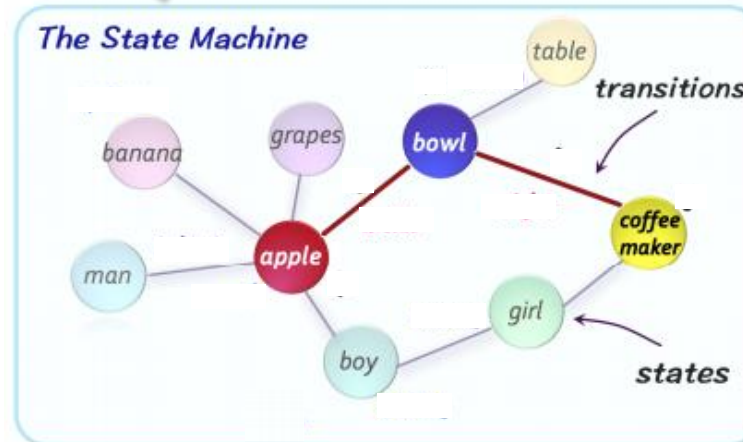


What is the *red fruit* inside the *bowl*
to the *right* of the *coffee maker*?

1. Given an **image**, generate a probabilistic **scene graph** that captures the semantic concepts.
2. Treat the graph as a **state machine** and simulate iterative computation over it to *answer questions or draw inferences*.
3. Natural language questions are translated into *soft instructions* and used to perform sequential reasoning over the scene graph/state machine.

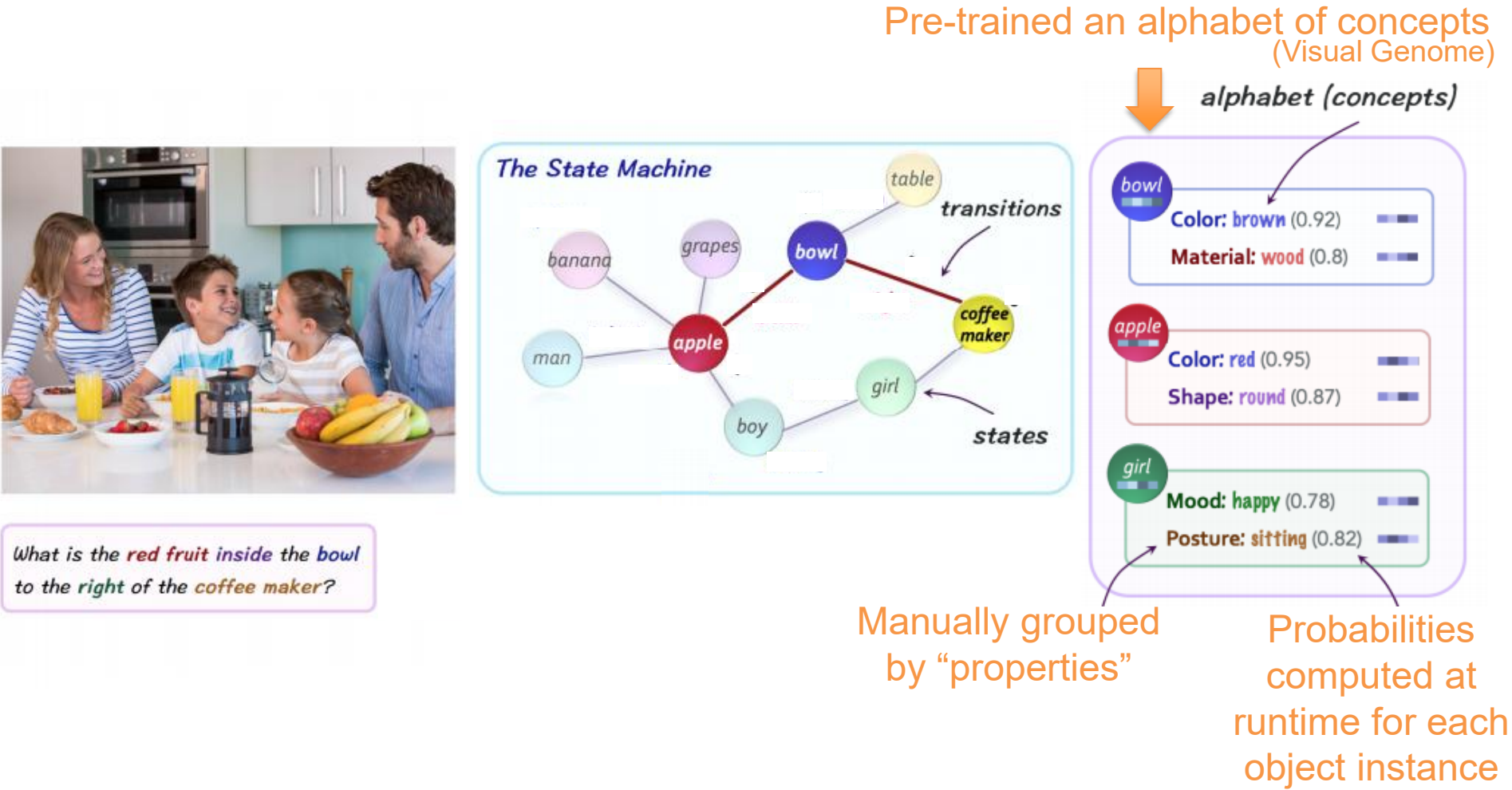
Module Networks V3b: The Neural State Machine

Detect objects and create proximity graph



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

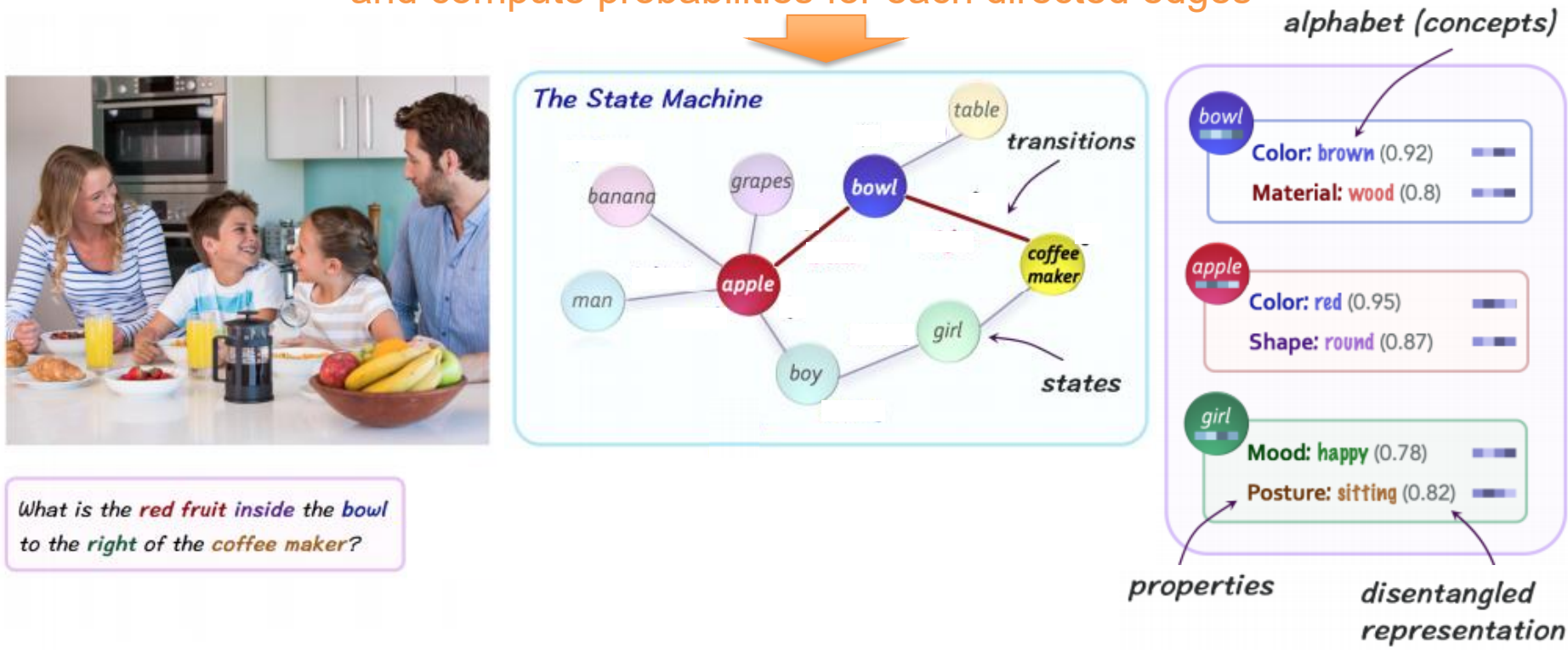
Module Networks V3b: The Neural State Machine



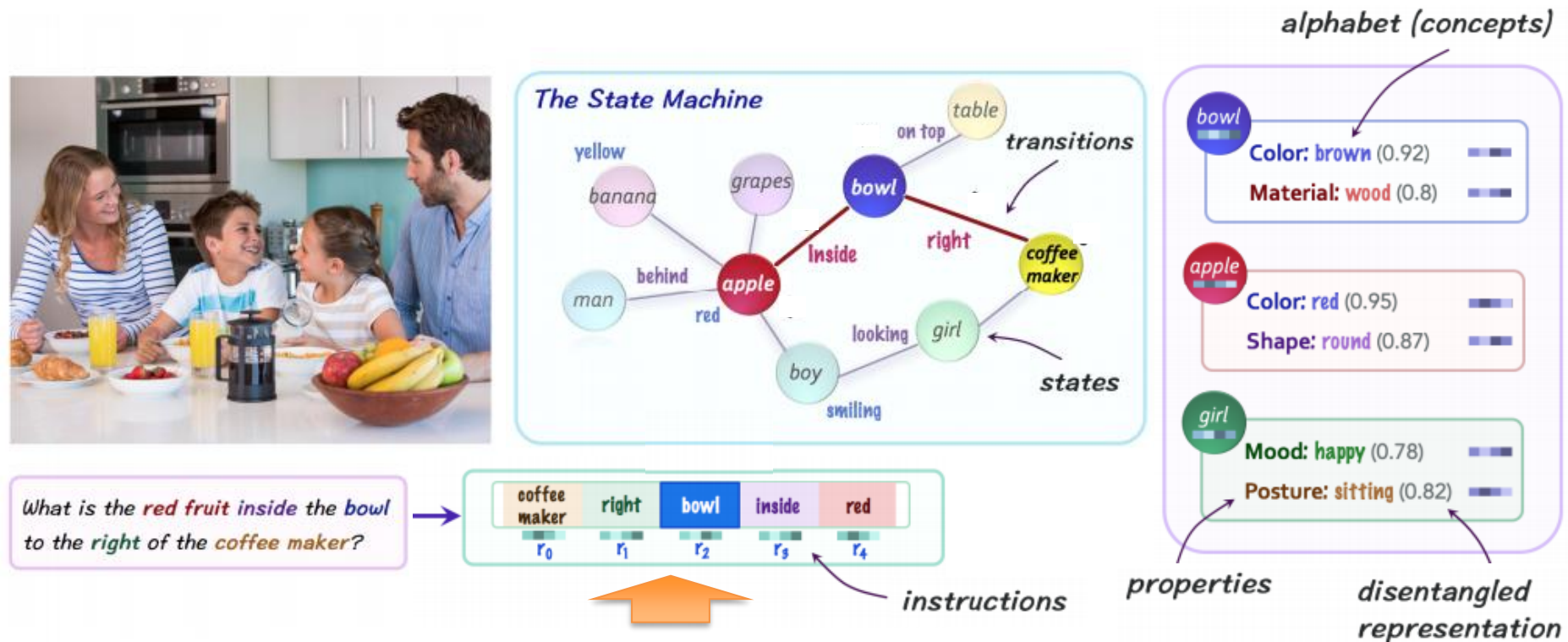
Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Module Networks V3b: The Neural State Machine

Predefined an alphabet of relations
and compute probabilities for each directed edges



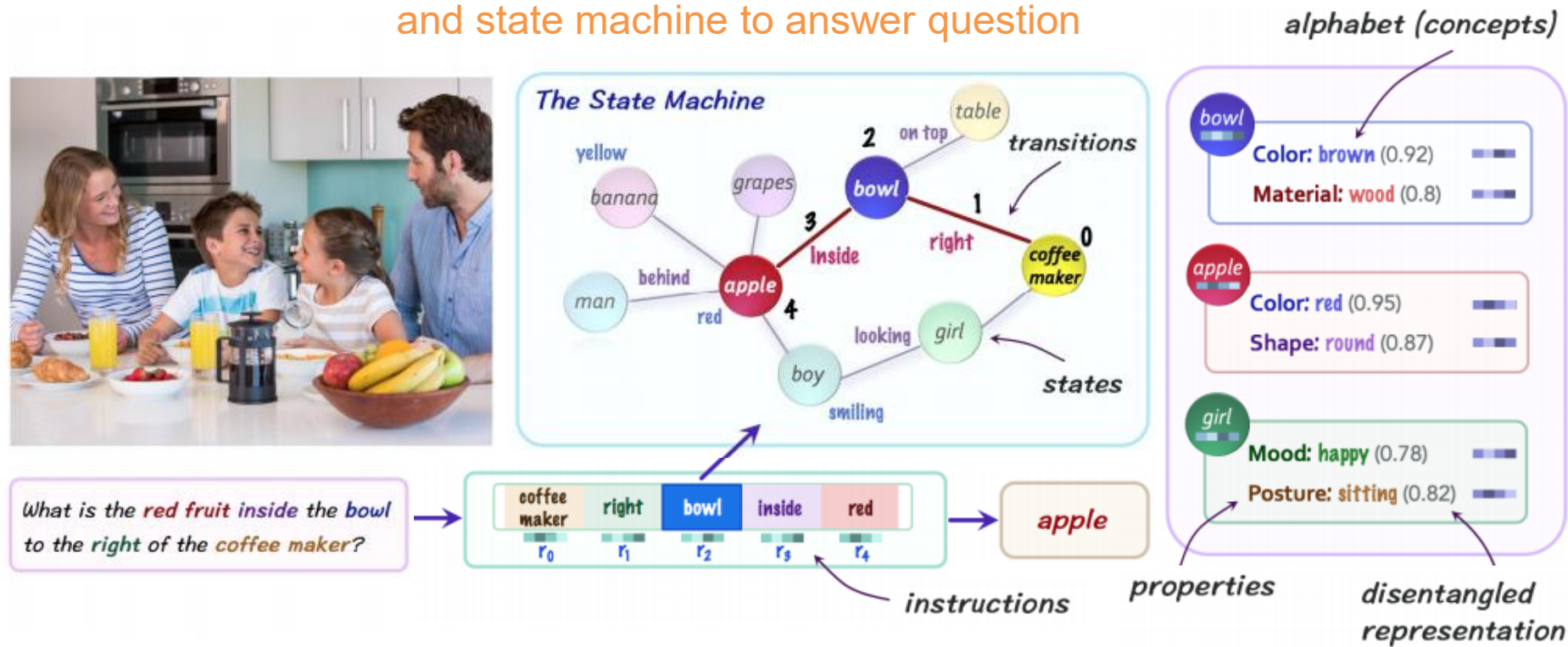
Module Networks V3b: The Neural State Machine



Translate each word in a concept-based representation
and group in a fixed number of instruction steps

Module Networks V3b: The Neural State Machine

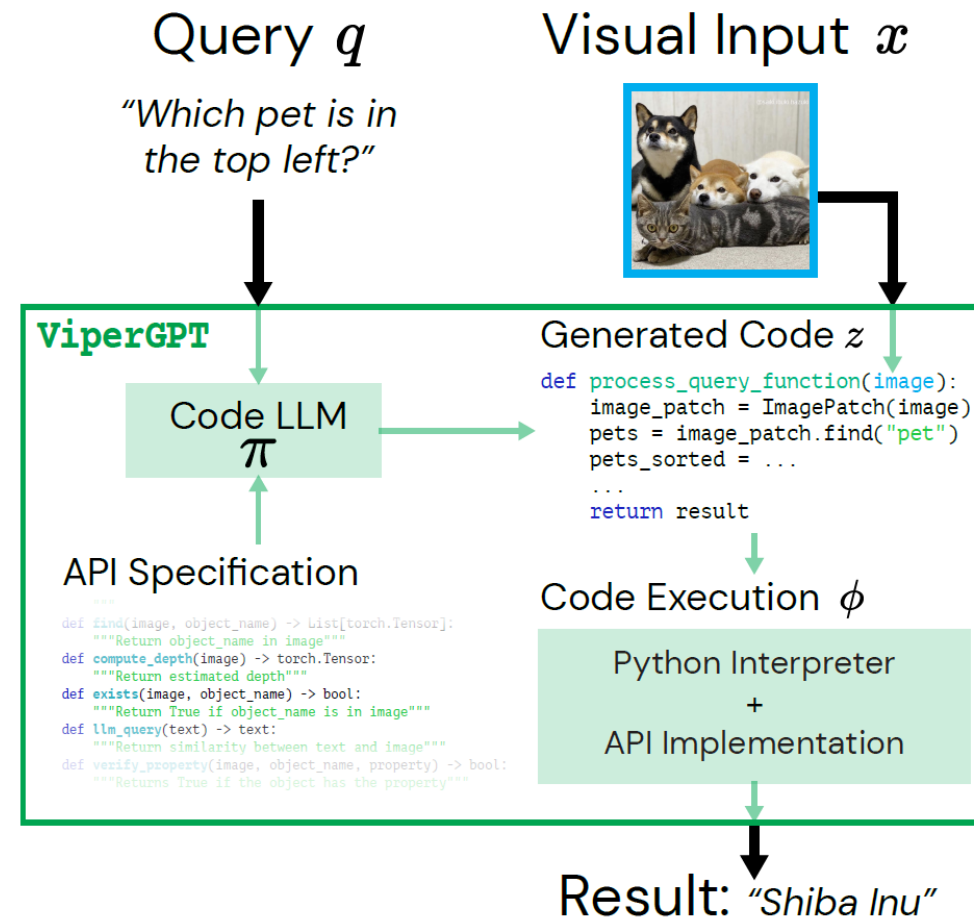
Finally, perform reasoning using instructions and state machine to answer question



Module Networks V4: Visual Programming

Predefined modules and subroutines

Image Understanding	Loc OWL-ViT	FaceDet DSFD (pypi)	Seg MaskFormer	Select CLIP-ViT	Classify CLIP-ViT	Vqa ViLT
Image Manipulation	Replace Stable Diffusion	ColorPop PIL.convert() cv2.grabCut()	BgBlur PIL.GaussianBlur() cv2.grabCut()	Tag PIL.rectangle() PIL.text()	Emoji AugLy (pypi)	
	Crop PIL.crop()	CropLeft PIL.crop()	CropRight PIL.crop()	CropAbove PIL.crop()	CropBelow PIL.crop()	
Knowledge Retrieval	List GPT3	Arithmetic & Logical		Eval eval()	Count len()	Result dict()



Gupta and Kembhavi, Visual Programming: Compositional visual reasoning without training, 2023

Suris et al., ViperGPT: Visual Inference via Python Execution for Reasoning, 2023

Intermediate Concepts

Intermediate Concepts via Hard Attention

Hard attention 'gates' (0/1) rather than soft attention (softmax between 0-1)

- Can be seen as discrete layers in between differentiable neural net layers

Sentiment analysis,
emotion recognition



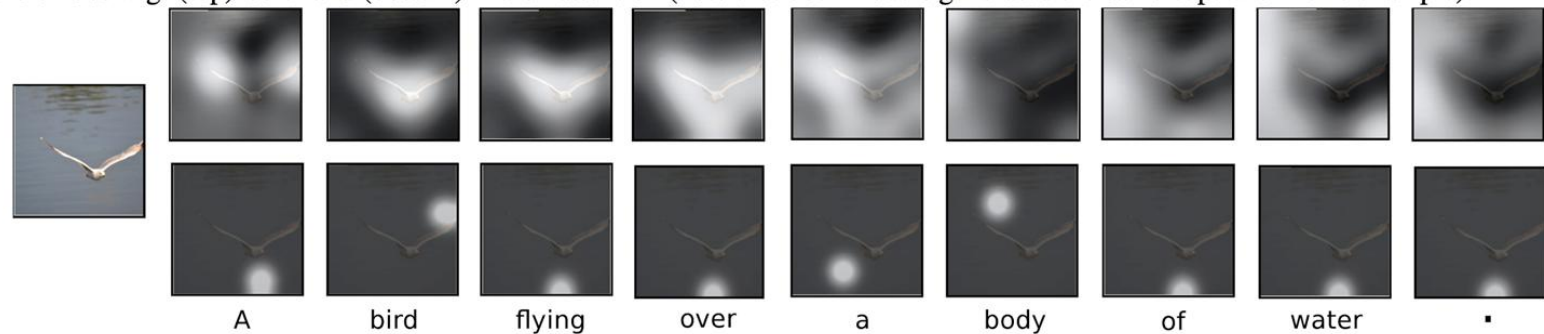
Reject

Pass

Reject

Figure 3. Visualization of the attention for each generated word. The rough visualizations obtained by upsampling the attention weights and smoothing. (top) "soft" and (bottom) "hard" attention (note that both models generated the same captions in this example).

Image captioning



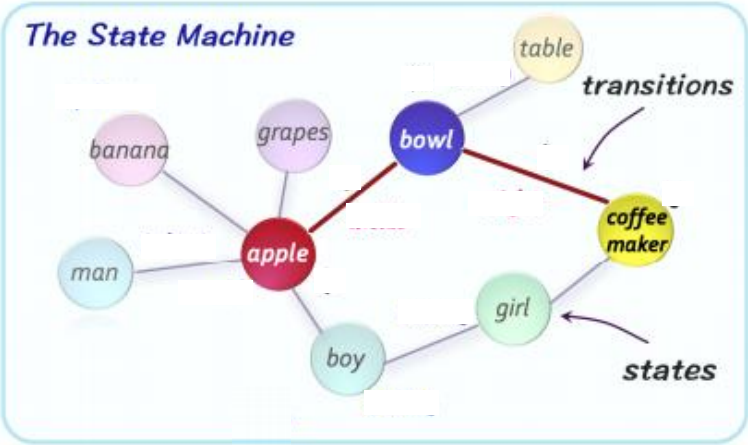
[Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015]

[Chen et al., Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning. ICMI 2017]

Fixed Concept Alphabet



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

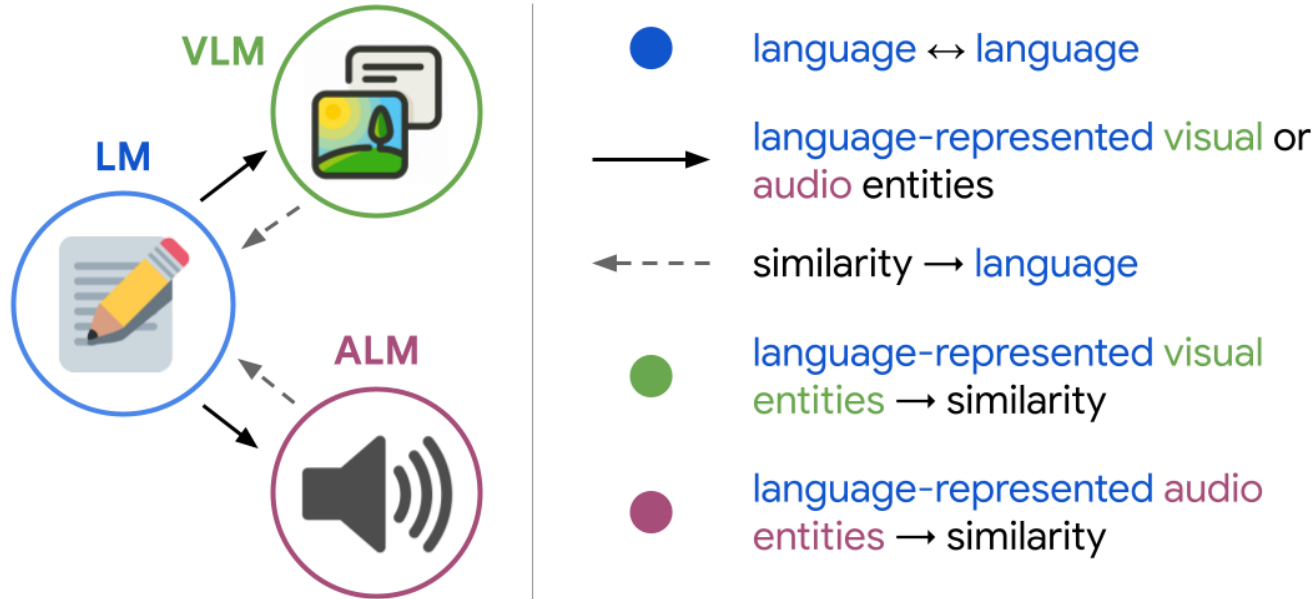


Pre-trained an alphabet of concepts
(Visual Genome)

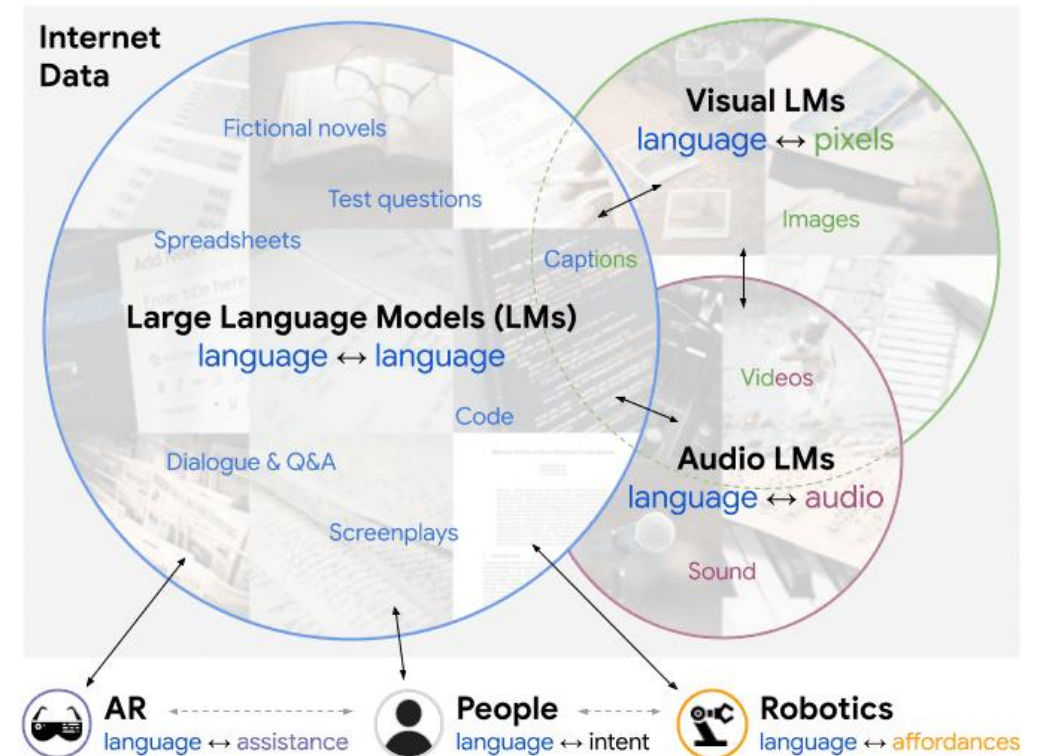


Language as Intermediate Concepts

- Large language/video/audio models interacting with each other
- Each language model has its own distinct *domain knowledge*
- Interaction is scripted and zero-shot



Guided multimodal discussion

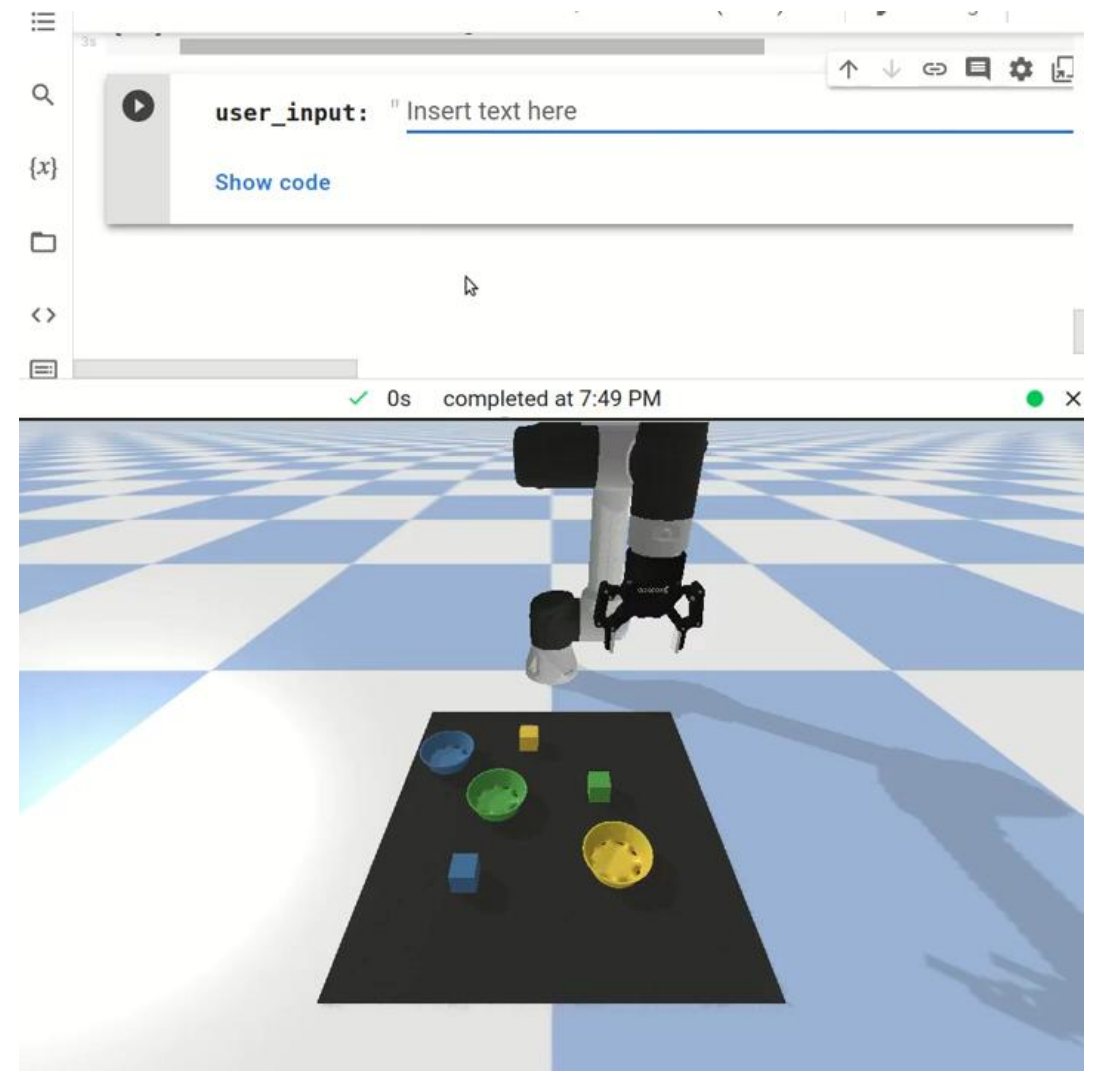


Combining domain knowledge

[Zeng et al., Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. arXiv 2022]

Language as Intermediate Concepts

Zero-Shot
Socratic
Internet
Image
Captioning



[Zeng et al., Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. arXiv 2022]

External Knowledge

External Knowledge: Why Models Need It

Knowledge can also be gained from external sources



What kind of board is this?

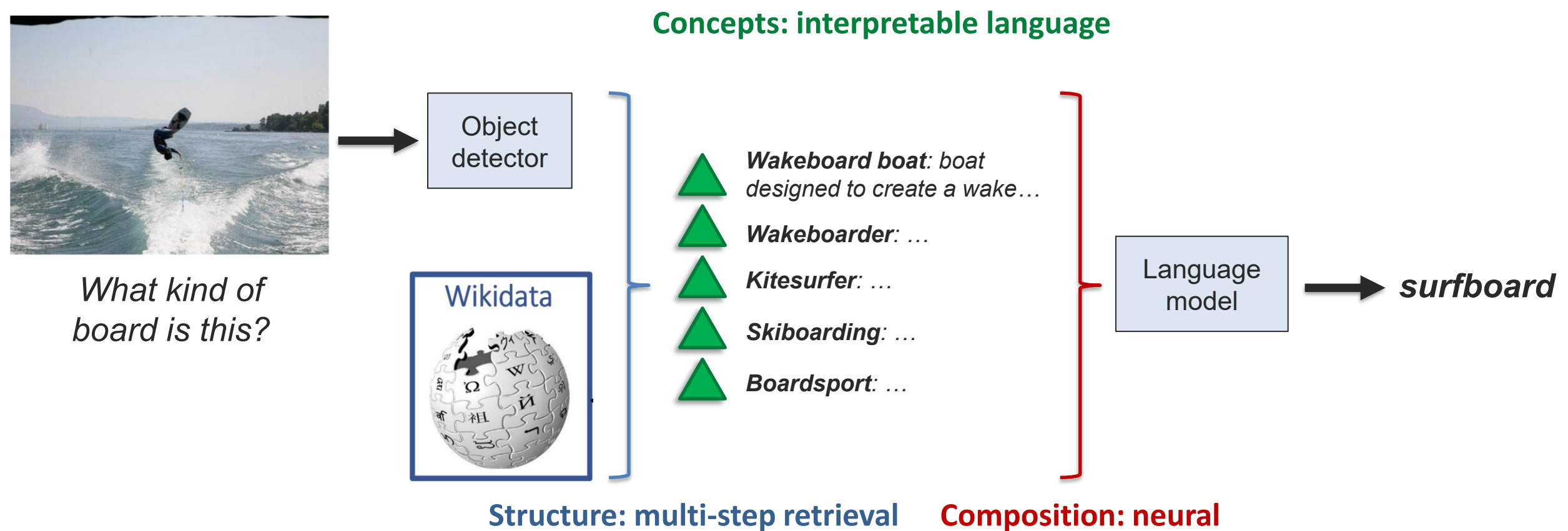
Requires knowledge of water sports, sports equipment, etc.

Existing models struggle when external knowledge is needed.
How can we leverage external knowledge?

[Marino et al., OK-VQA: A visual question answering benchmark requiring external knowledge. CVPR 2019]

External Knowledge: Knowledge Graphs (KAT)

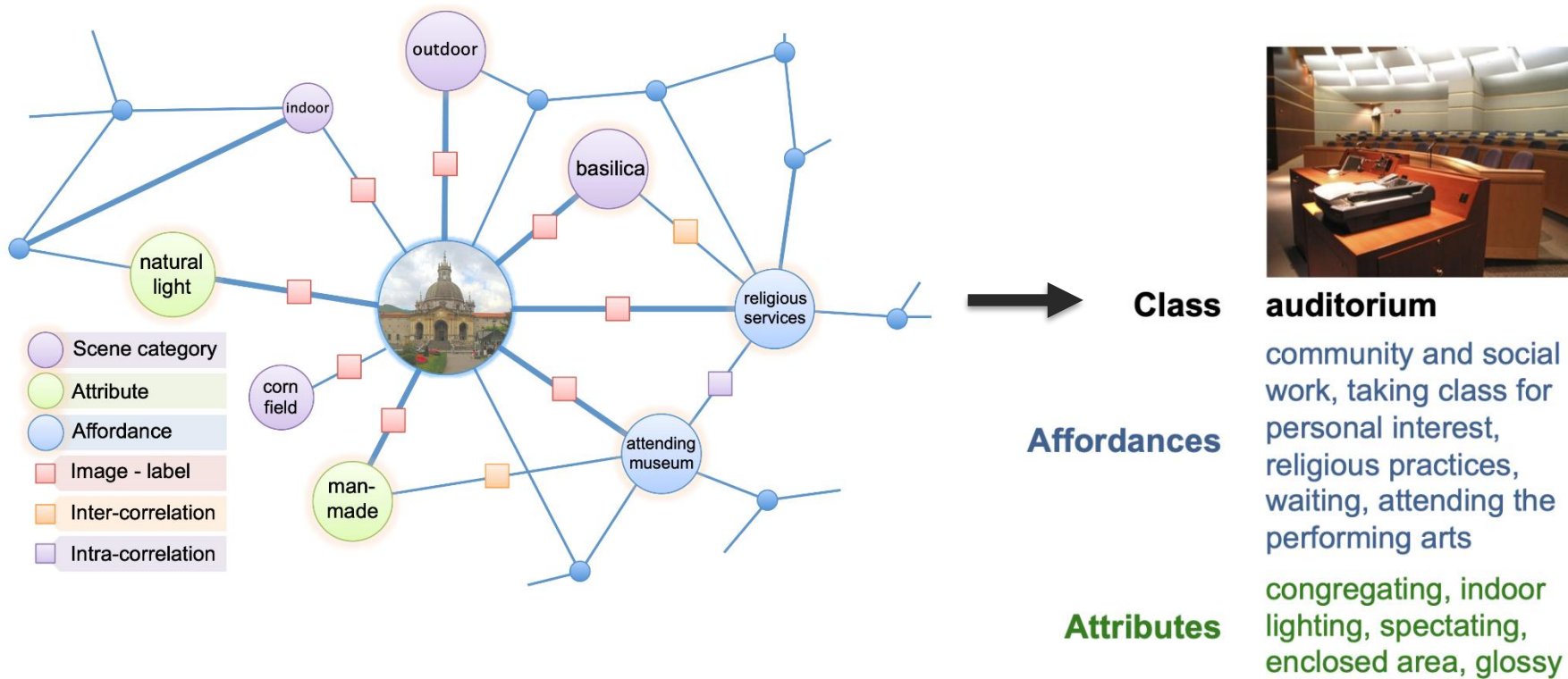
Knowledge can also be gained from external sources



[Gui et al., KAT: A Knowledge Augmented Transformer for Vision-and-Language. NAACL 2022]

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources



Concepts: interpretable
Structure: multi-step inference
Composition: graph-based

[Zhu et al., Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries. arXiv 2015]

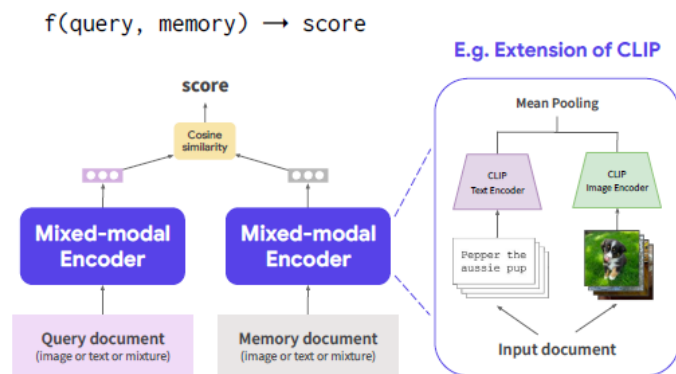
External Knowledge: Retrieval



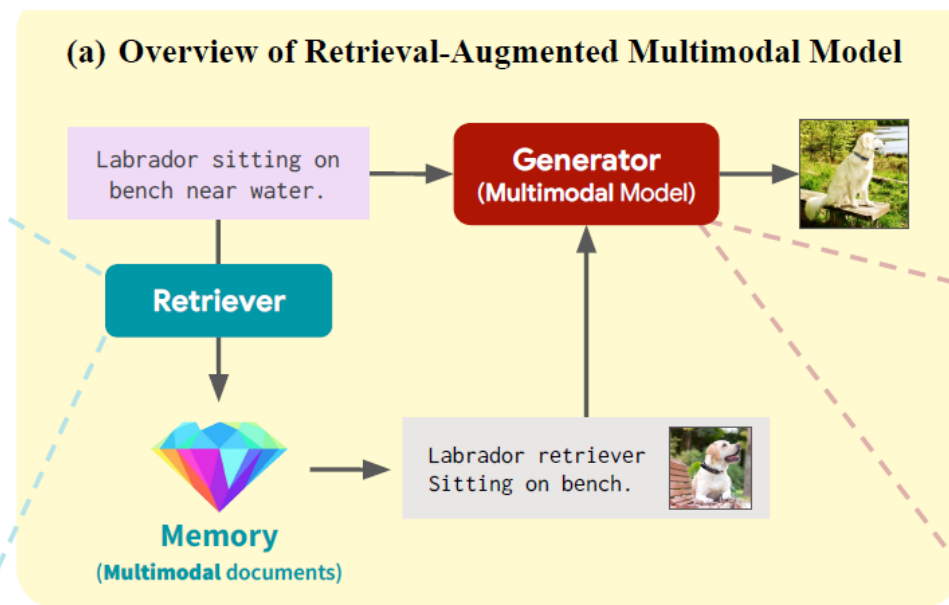
[Hu et al., REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory, 2023]

External Knowledge: Retrieval + Generation

(b) Dense Multimodal Retriever



(a) Overview of Retrieval-Augmented Multimodal Model



(c) Retrieval-Augmented Generator

