Multimodal Machine Learning

Lecture 7.1: Alignment and Translation

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.*
Spring 2021 edition taught by Yonatan Bisk

Administrative Stuff

Midterm Project Report Instructions

- **Goal:** Evaluate state-of-the-art models on your dataset and identify key issues through a detailed error analysis
  - It will inform the design of your new research ideas
- **Report format:** 2 column (ICML template)
  - The report should follow a similar structure to a research paper
  - Teams of 3: 8 pages, Teams of 4: 9 pages, Teams of 5: 10 pages.
- **Number of SOTA models**
  - Teams of 3 should have at least two baseline models
  - Teams of 4 or 5 should have at least three baseline models
- **Error analysis**
  - This is one of the most important part of this report. You need to understand where previous models can be improved.

Examples of Possible Error Analysis Approaches

- Visualization (e.g., TSNE) of the correct and incorrect predictions
- Manually inspect the samples that are incorrectly predicted
  - What are the commonalities?
  - What are differences with the correct ones?
- Ablation studies to understand what model components are important

Midterm Project Report Instructions

Main report sections:

- Abstract
- Introduction
- Related work
- Problem statement
- Multimodal baseline models
- Experimental methodology
- Results and discussion
- New research ideas

The structure is similar to a research paper submission ☺

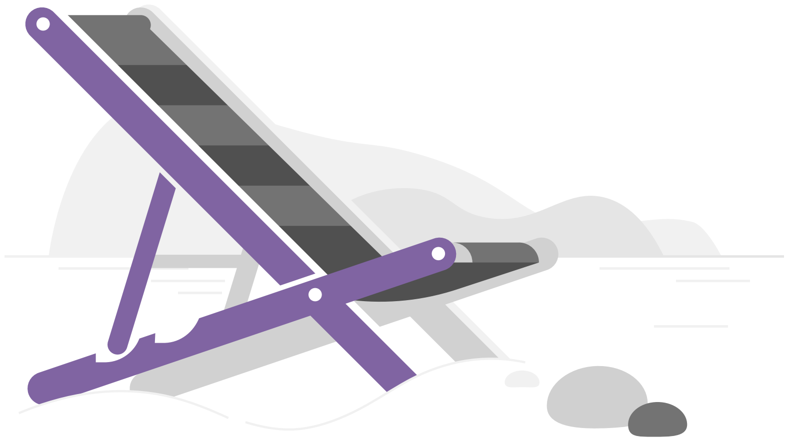
Upcoming Deadlines

- Reading assignments this week and next week
- Thursday October 21st : Project session (no lecture)
- Sunday October 31st: Midterm report deadline
- Tuesday and Thursday (11/2 and 11/4): midterm presentations
  - All students are expected to attend both presentation sessions in person
  - Each team will present either Tuesday or Thursday
  - The focus of these presentations is about your research ideas
  - Feedback will be given by all students, instructors and TAs

Mid-Semester Break

No lecture on Thursday (Oct 13)

CMU official holiday!



Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 7.1: Alignment and Translation

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.*
Spring 2021 edition taught by Yonatan Bisk

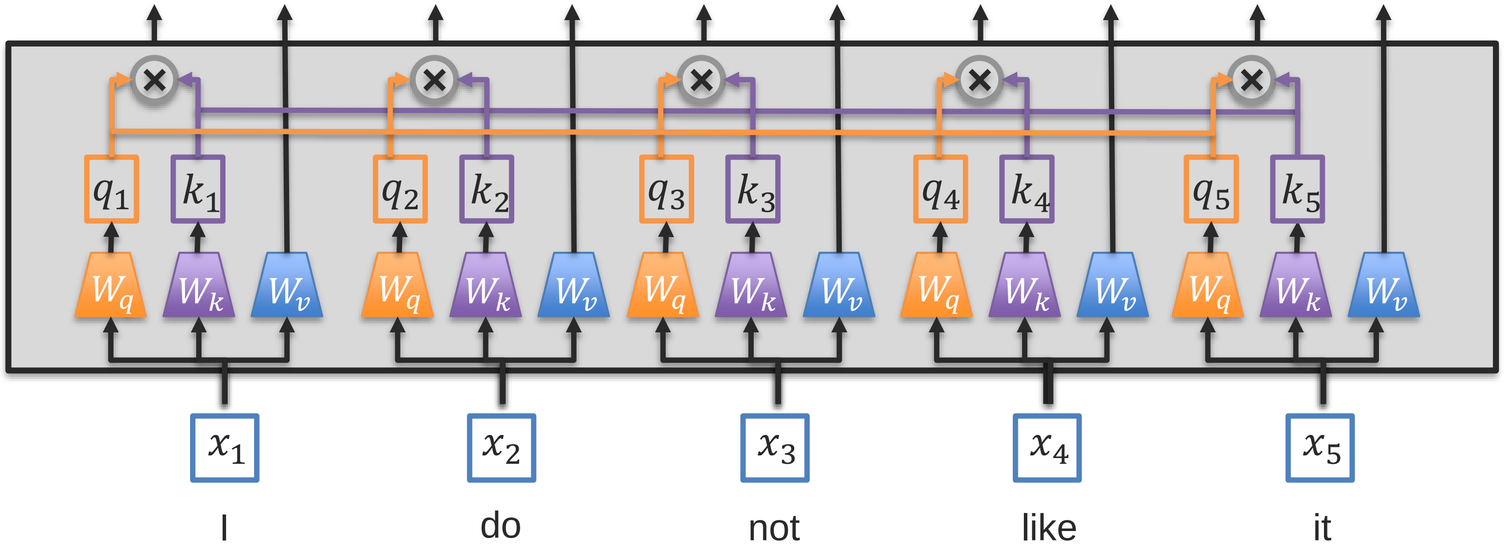
Learning Objectives of Today's Lecture

- Graph Representations
  - Graph Neural Networks
  - Graph Convolution Networks
- Multimodal Translation
  - Visual Question Answering
    - Co-attention, Stacked attention
  - Neural module networks
    - Neural-symbolic learning
    - Neural State Machine
  - Biases in VQA models
- Visual Dialogue
  - Causal Graph
  - Multi-Step Reasoning

Going Beyond Sequences: Graph Representations

*slides adapted from Leskovec, Representation Learning on Networks. WWW 2018

Transformers – Fully-Connected Sequences

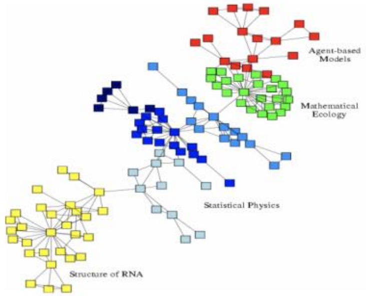


Should everything be connected to everything?

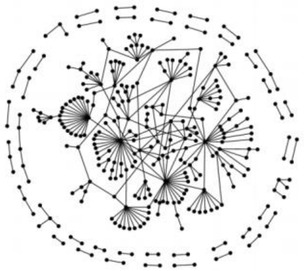
What if we have domain knowledge about connections?

Graphs (aka “Networks”)

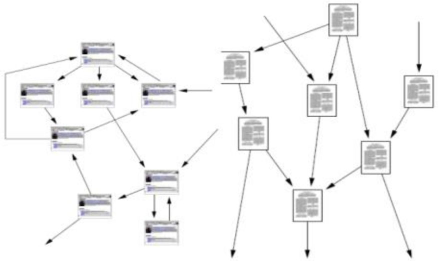
Social networks



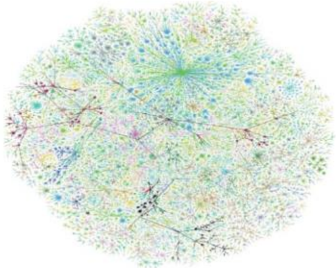
Economic networks



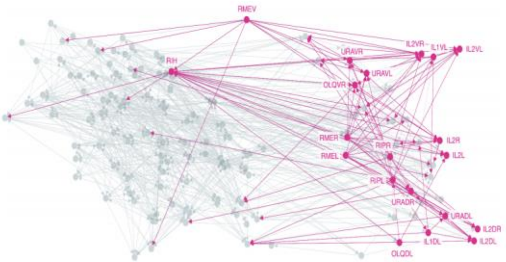
Biomedical networks



Information networks:
Web & citations



Internet



Networks of neurons

Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019

Graph Representation

Assume we have a graph **G:**

V is the set of vertices

A is the binary adjacency matrix

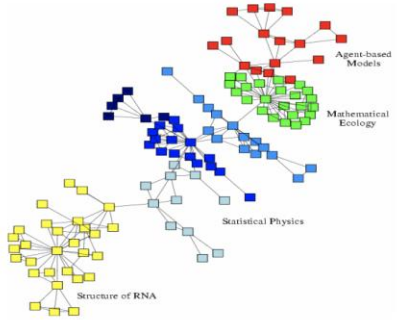
X is a matrix of node features:

- Categorical attributes, text, image data

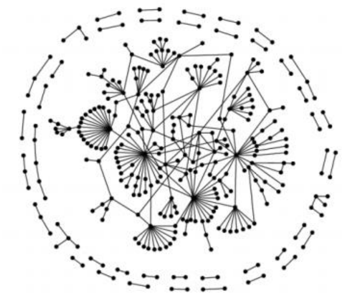
  e.g. profile information in a social network
- …

Y is a vector of node labels (optional)

Social networks



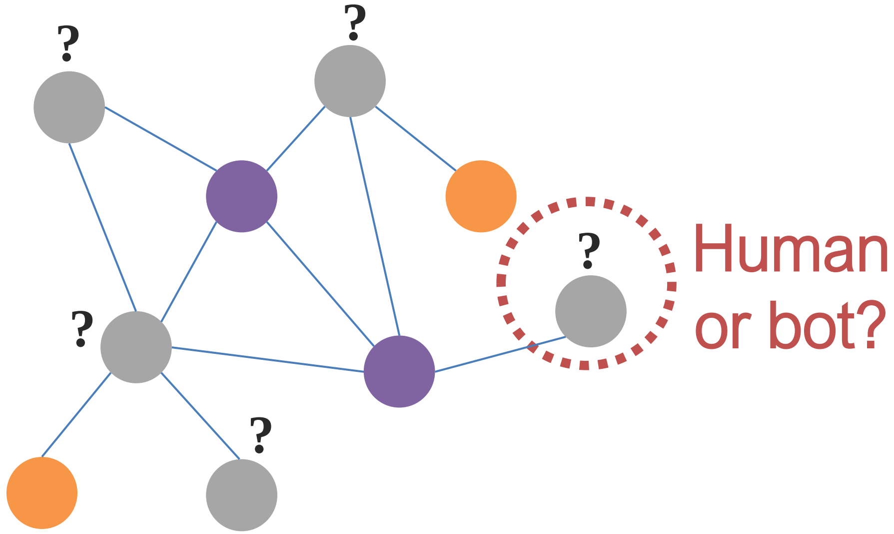
Economic networks



Biomedical networks

Graphs – Supervised Task

Goal: Learn from labels associated with a subset of nodes (or with all nodes)



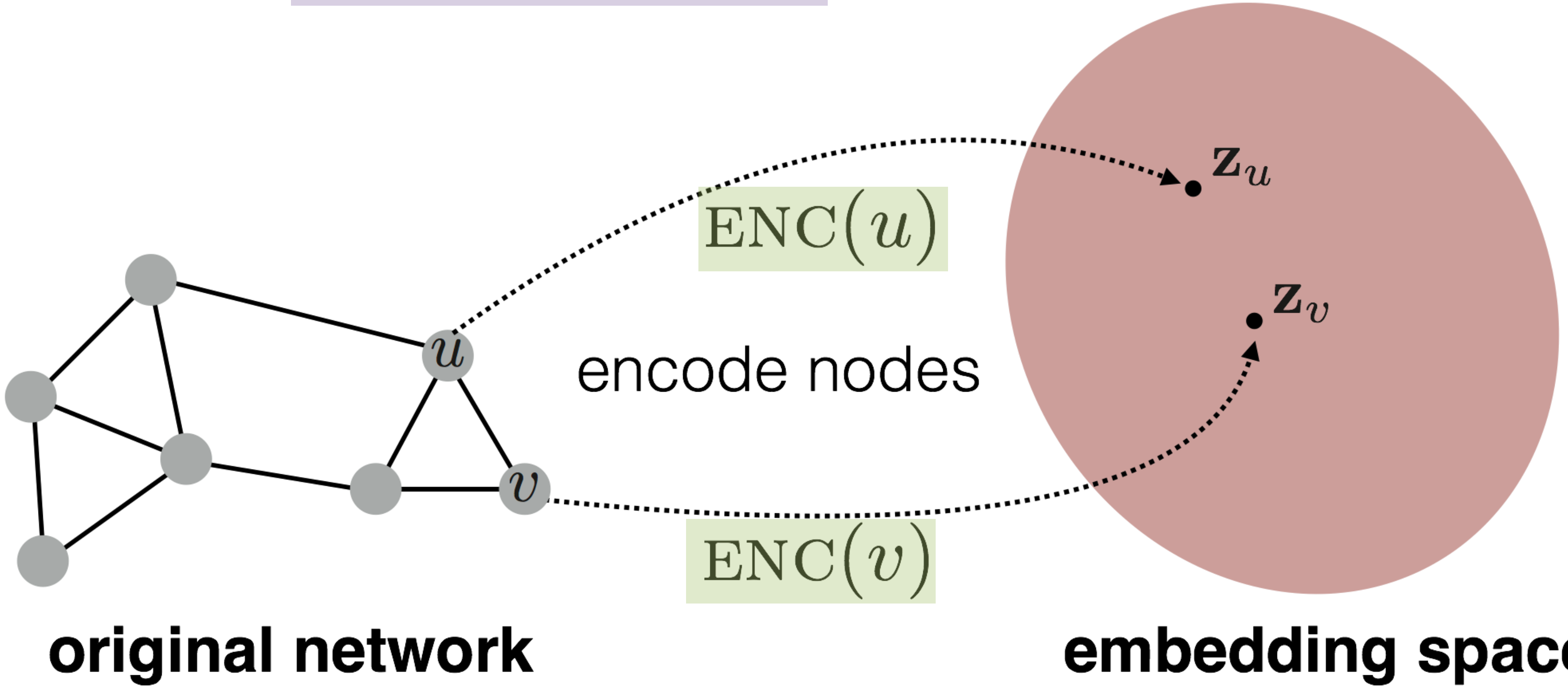
e.g., an online social network

14

Graphs – Unsupervised Task

Goal: Learn an embedding space where

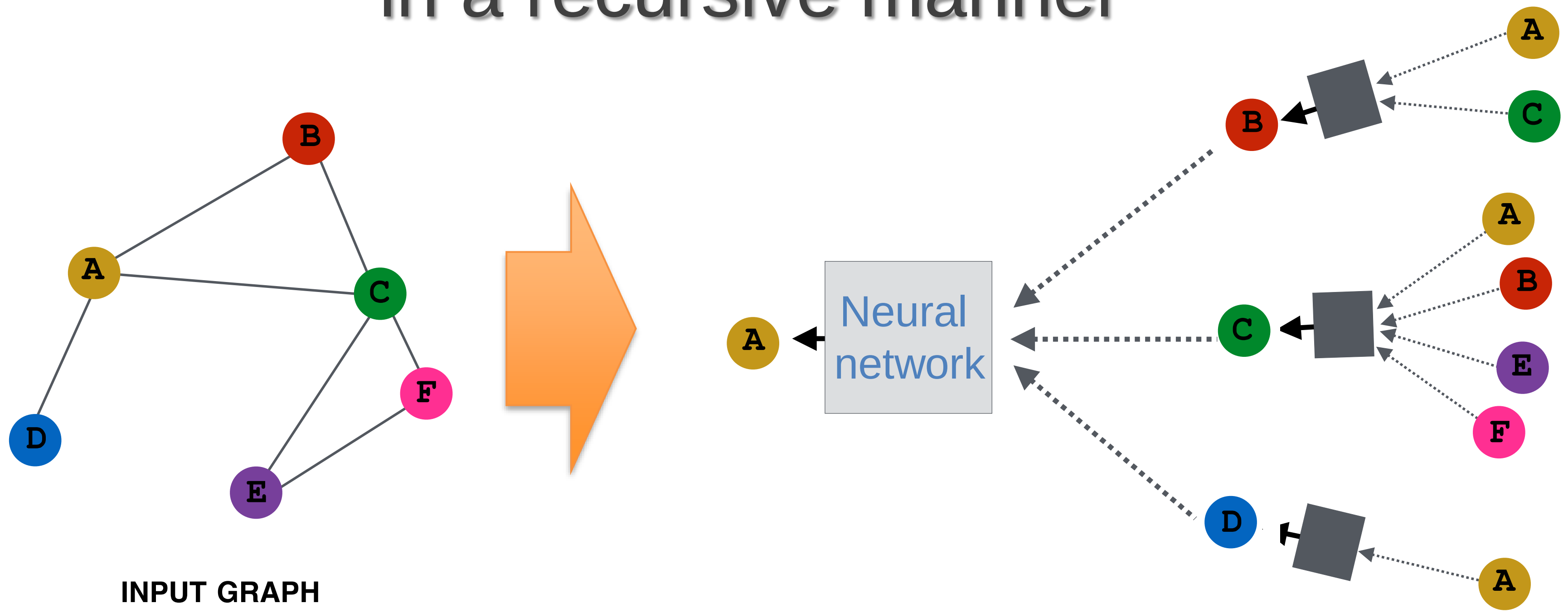
$$\text{similarity}(u, v) \approx \mathbf{z}_v^\top \mathbf{z}_u$$



15

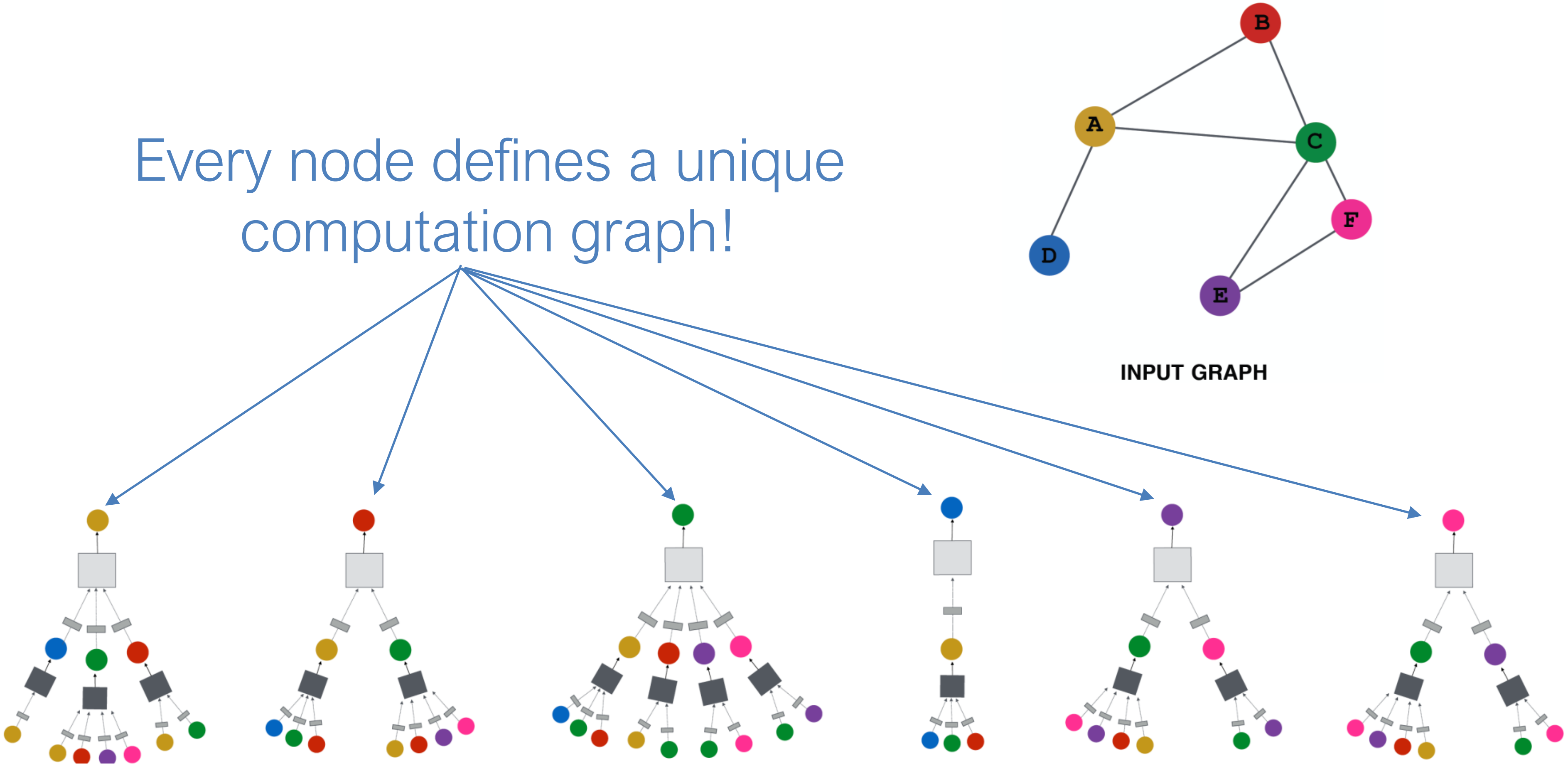
Graph Neural Nets

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



Graph Neural Nets

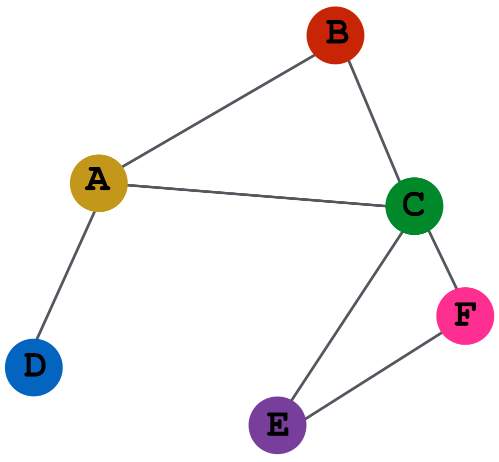
Every node defines a unique computation graph!



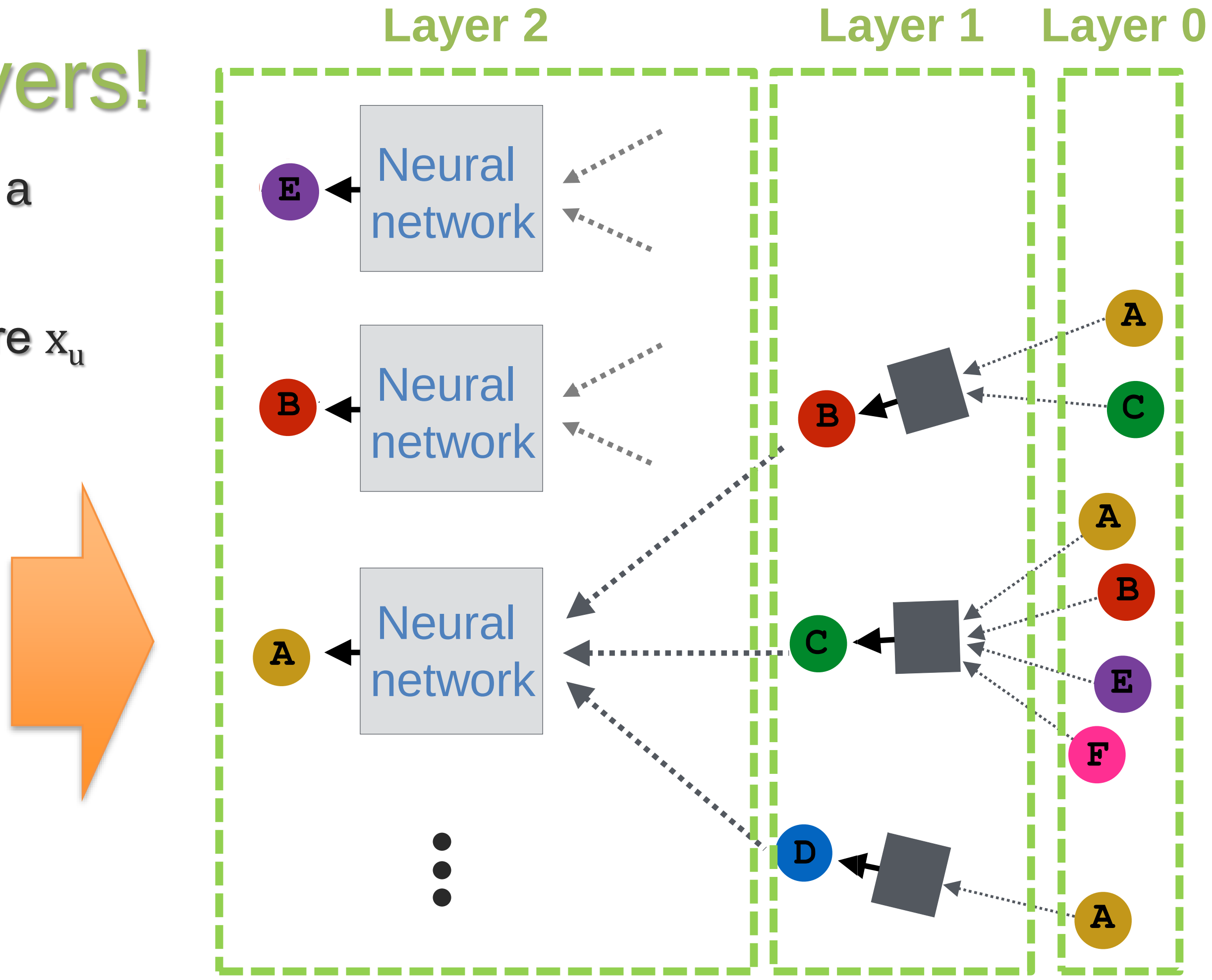
Graph Neural Nets

And multiple layers!

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



INPUT GRAPH



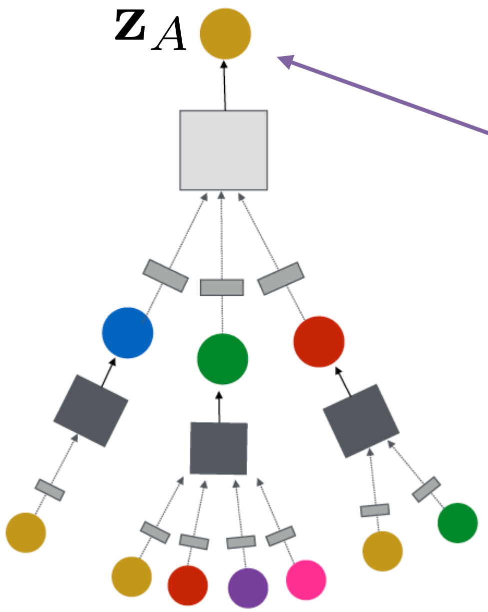
How do we train it?

Graph Neural Nets – Supervised Training

classification weights

Human or bot?

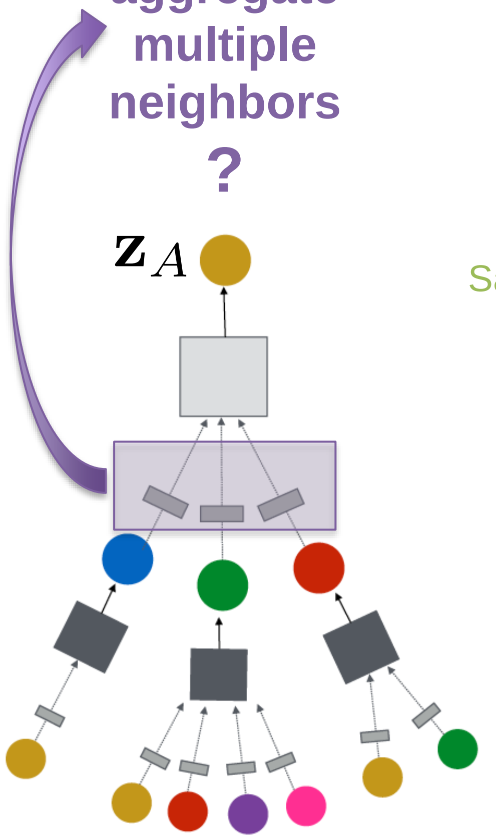
$$\mathcal{L} = \sum_{v \in V} \boxed{y_v} \log(\sigma(\boxed{\mathbf{z}_v}^\top \boxed{\boldsymbol{\theta}})) + (1 - \boxed{y_v}) \log(1 - \sigma(\boxed{\mathbf{z}_v}^\top \boxed{\boldsymbol{\theta}}))$$



output node embedding

node class label

Key Technical Challenge: Neighborhood Aggregation



Average pooling (Scarselli et al., 2005)

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

Different weights for neighbors and self

Graph Convolution Network (Kipf et al., 2017)

Same weights

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

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

Multimodal Translation

Visual Question Answering (VQA)

Visual Question Answering

Question

Is the skateboard airborne?

Image

Answer

yes

How can we use attention?

VQA and Attention

Question

Is the skateboard airborne?

Language can
be used to
attend the image

Image

Answer

yes

VQA and Attention

Question

Is the skateboard airborne?

Image could also be used to attend the text

Image

Answer

yes

Co-attention

Question

Is the skateboard airborne?

Image

Or do both!

Answer

yes

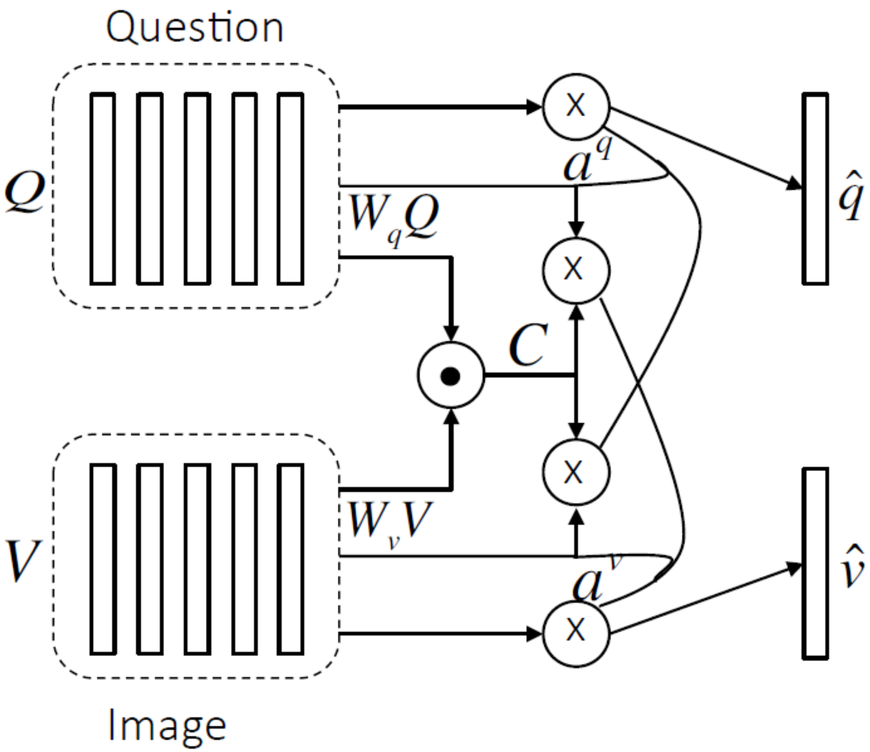
Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

Co-attention

Question

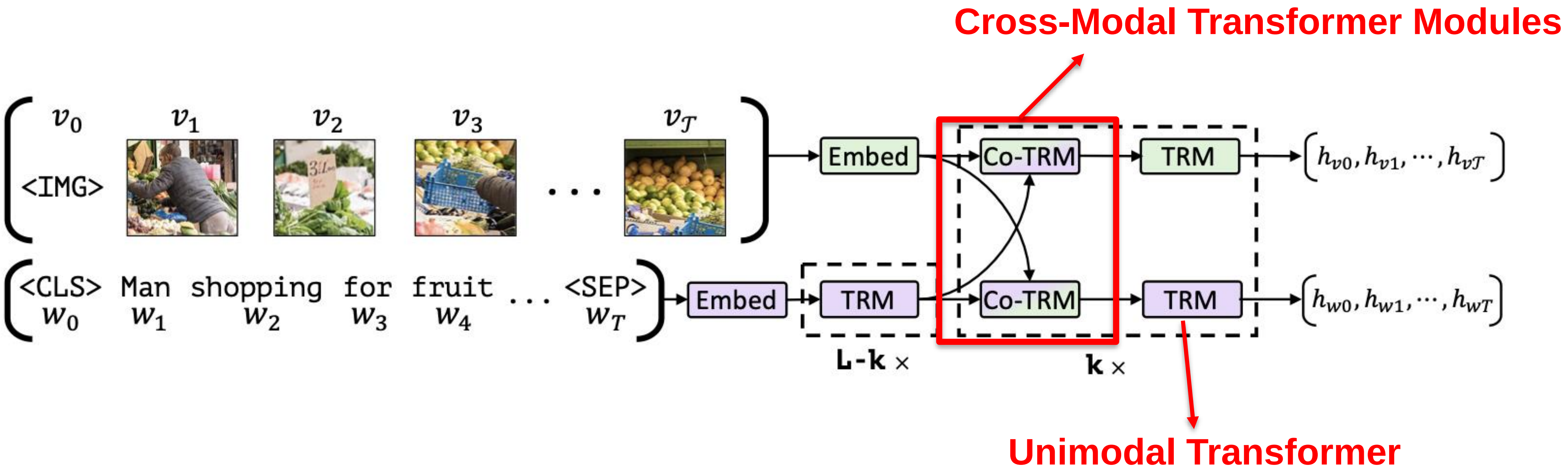
Is the skateboard airborne?

Image



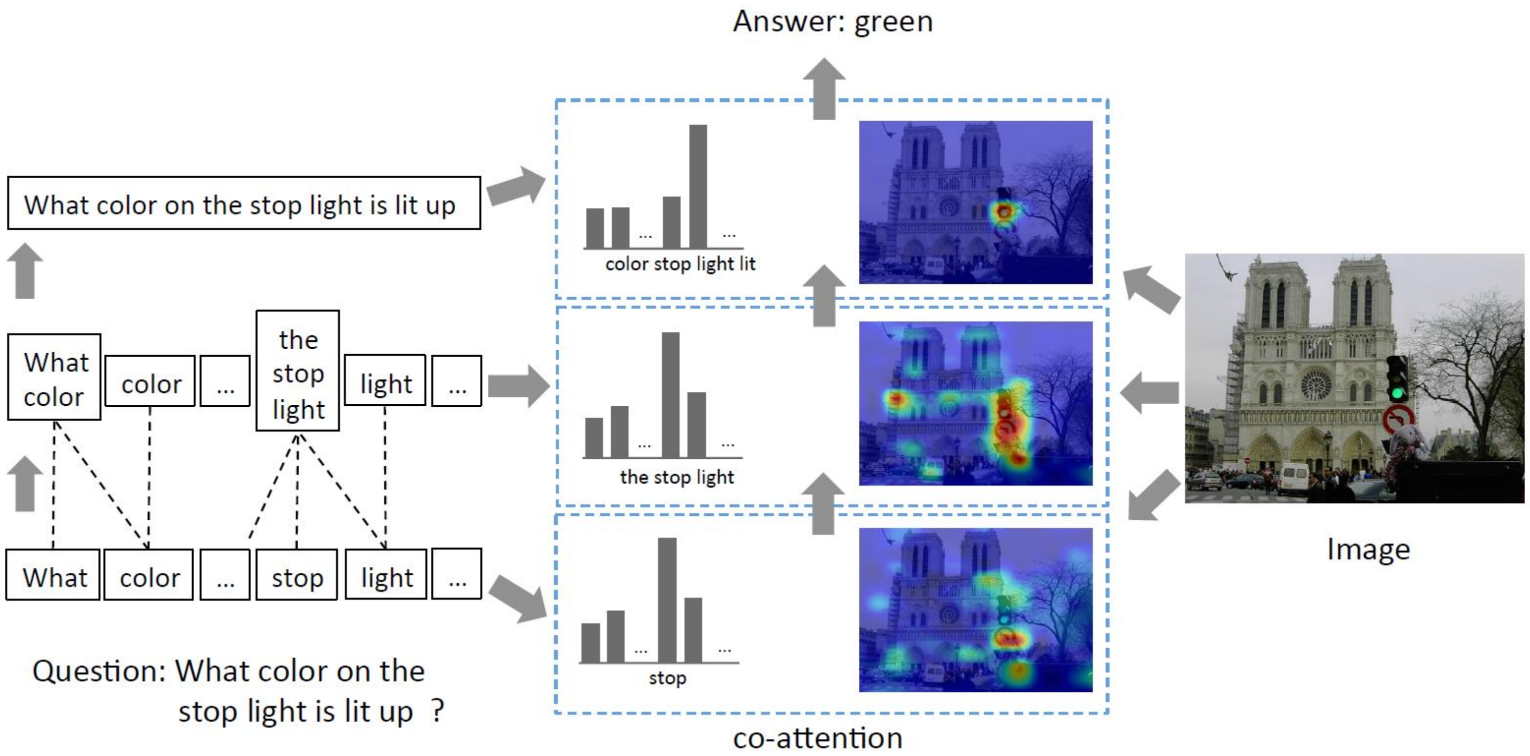
Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

Transformed-based “Co-Attention”: ViLBERT



Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *arXiv* (August 6, 2019).

Hierarchical Co-attention



Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

Stacked Attentions

Question

What are sitting in the basket on a bicycle?

Image

Attention 1 Attention 2

Answer

dogs

Attention 1 Attention 2

Yang et al., Stacked Attention Networks for Image Question Answering, CVPR 2016

Other Attention-based Models for VQA

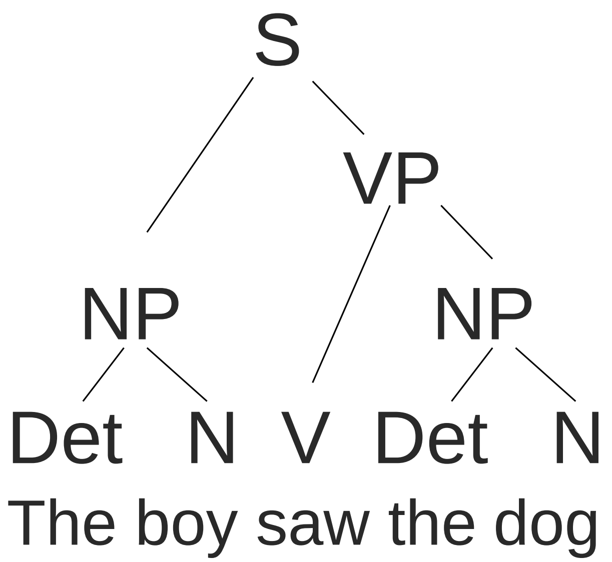
- Bottom-up and top-down attention for image captioning and visual question answering, CVPR 2018
  - Adds the idea of object-based representations
- Bilinear Attention Pooling, NIPS 2018
  - Extend low-rank bilinear pooling to multimodal
- Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE TNNLS, 2018

How can we make this attention process more interpretable?

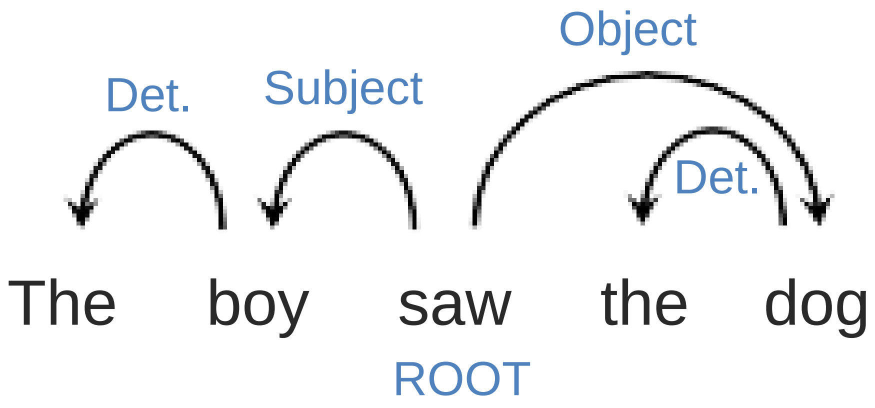
Can we take advantage of prior knowledge (e.g., language structure)?

Neural Module Networks

Syntax and Language Structure

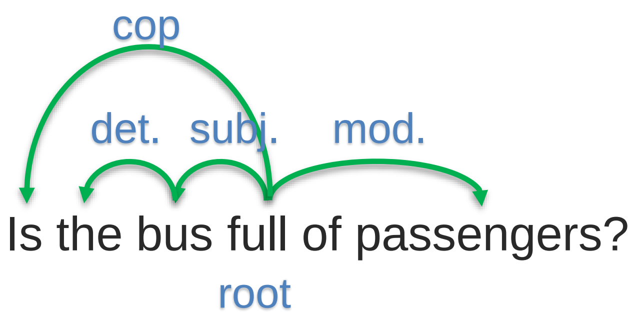


Constituency Parsing

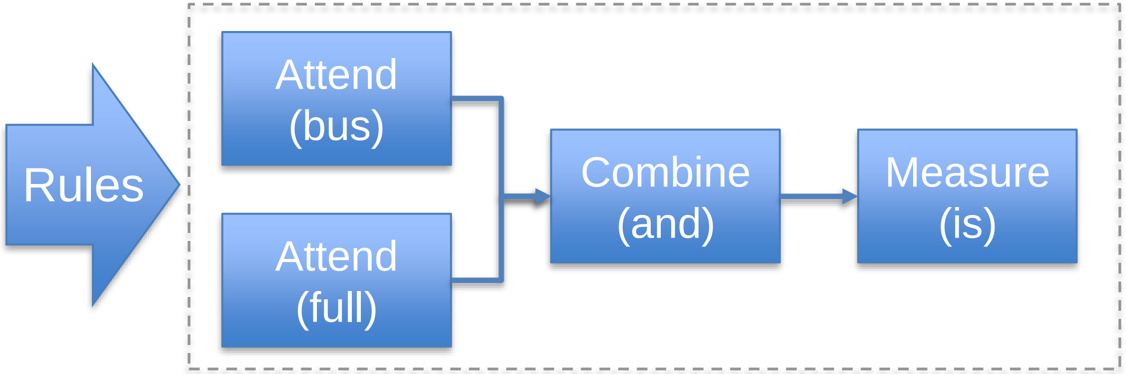


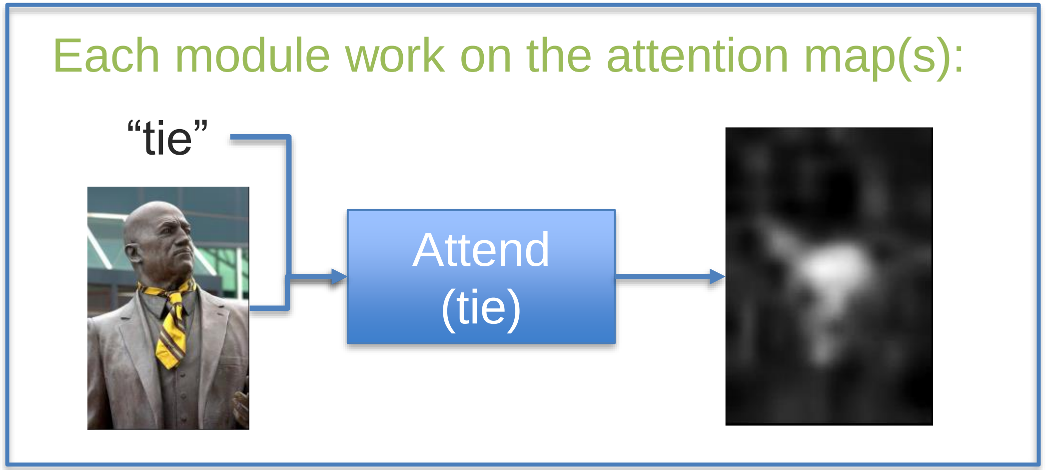
Dependency Parsing

Neural Module Network



Computation layout



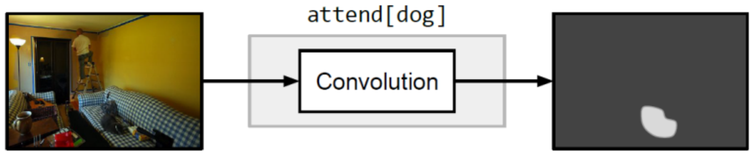


Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

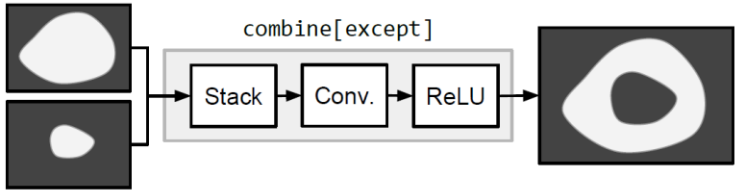
Predefined Set of Modules

1) Analyze the image:

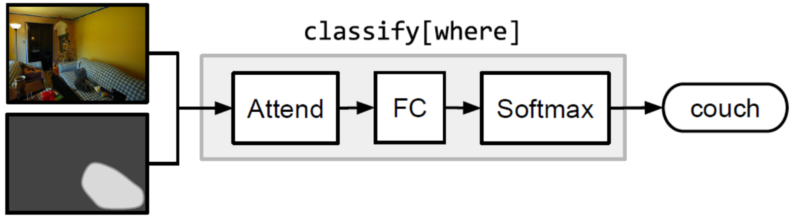
$\texttt{attend} : \mathit{Image} \to \mathit{Attention}$



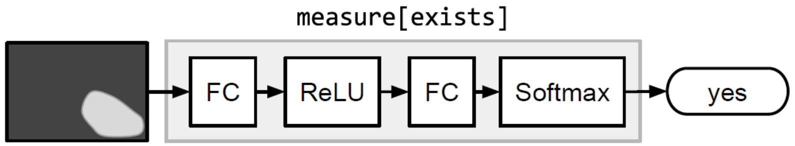
$\texttt{combine} : \mathit{Attention} \times \mathit{Attention} \to \mathit{Attention}$



2) Make a prediction



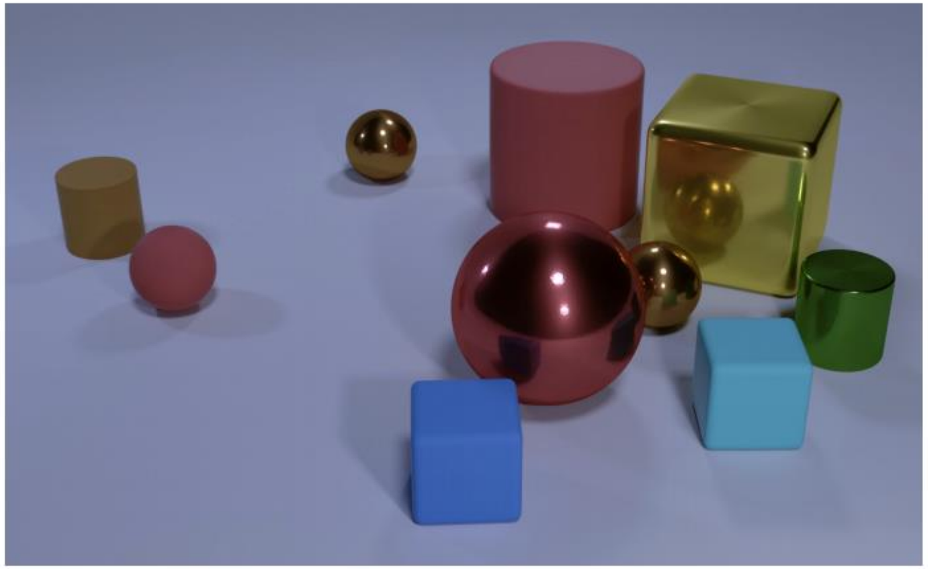
$\texttt{measure} : \mathit{Attention} \to \mathit{Label}$



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

CLEVR: Dataset for Visual Reasoning

Perfect for a neural module network!



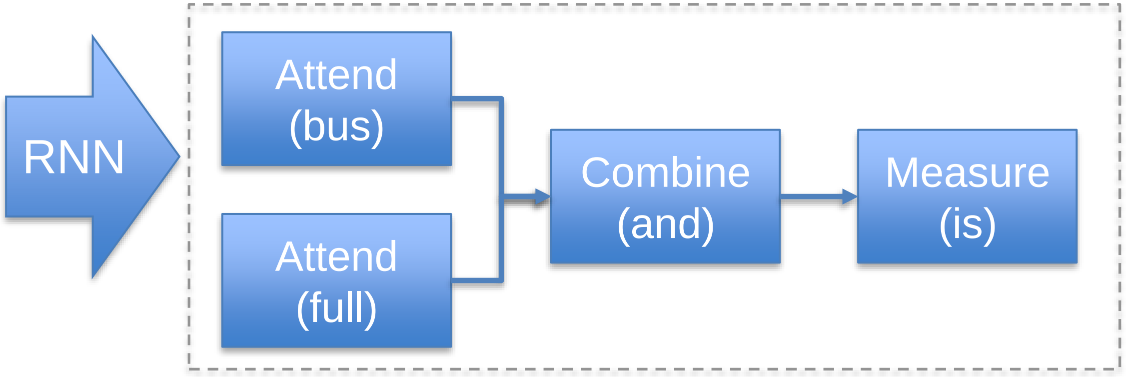
Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many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

Johnson et al., CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017

Module Network V2: End-to-End Learning

Is the bus full of passengers?

Computation layout

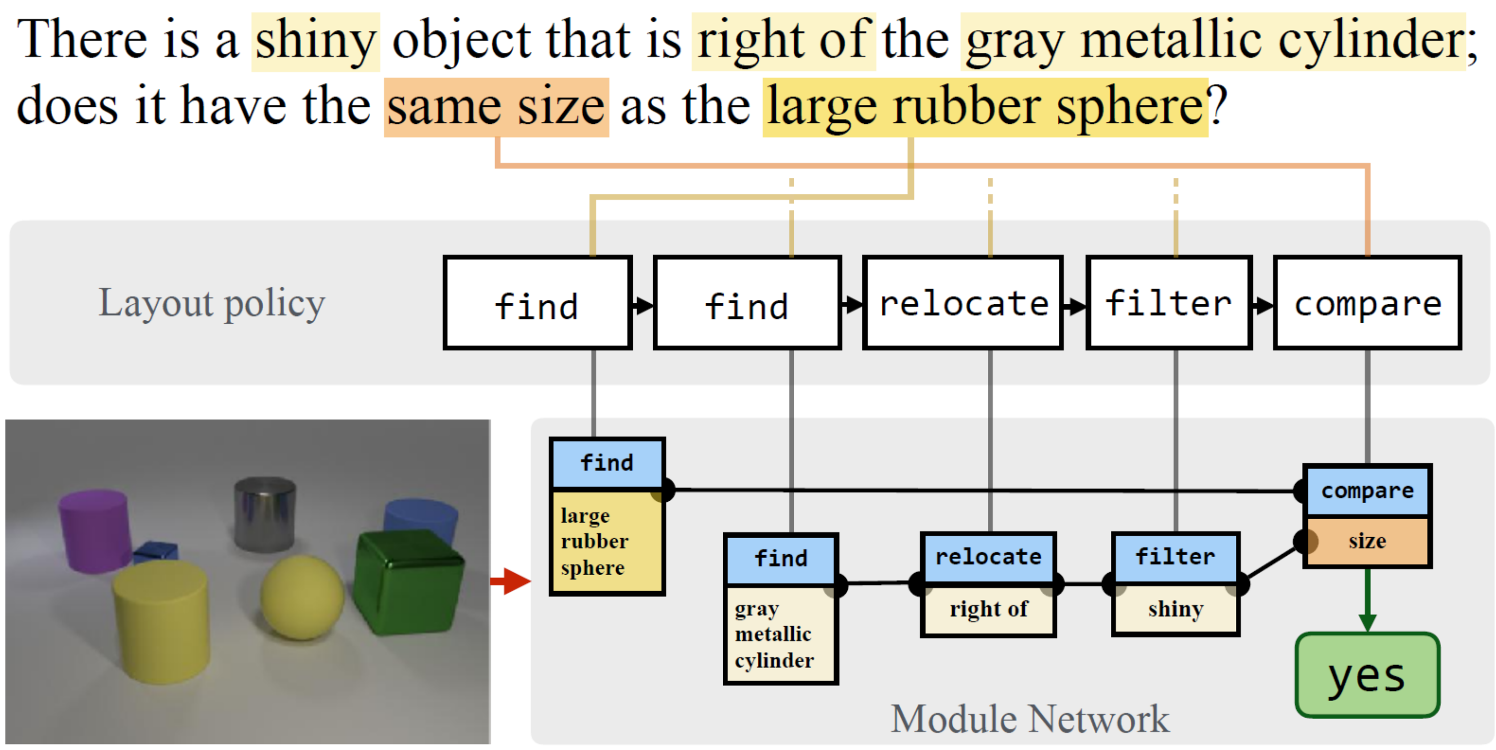


No need to parse the question!

No rule-based creation of the layout!

Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering, 2017

Module Network V2: End-to-End Learning

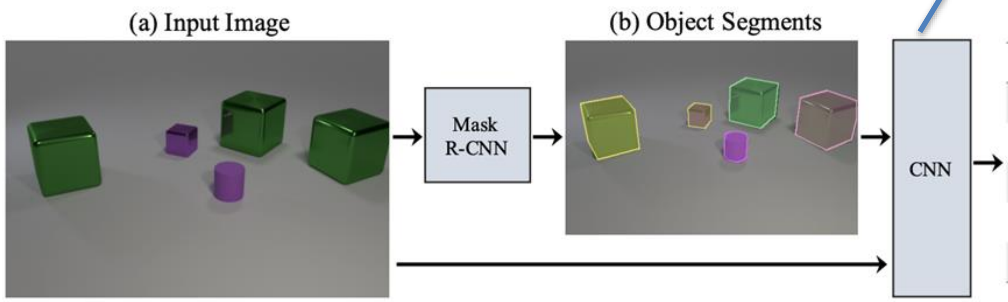


Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering, 2017

Module Network V3: Neural-symbolic VQA

1) Image Attributes

Previously trained in a supervised way



(c) Structural Scene Representation

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

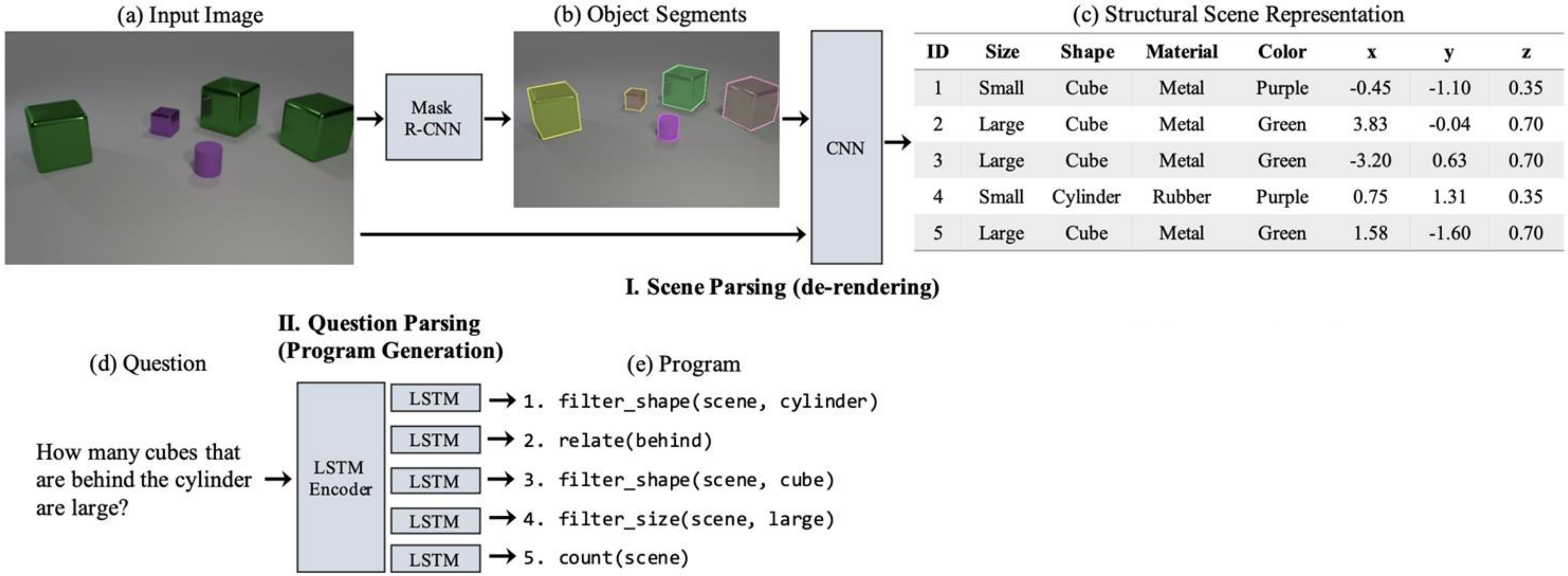
I. Scene Parsing (de-rendering)

Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Module Network V3: Neural-symbolic VQA

2) Parsing questions into programs

Similar to neural module networsk

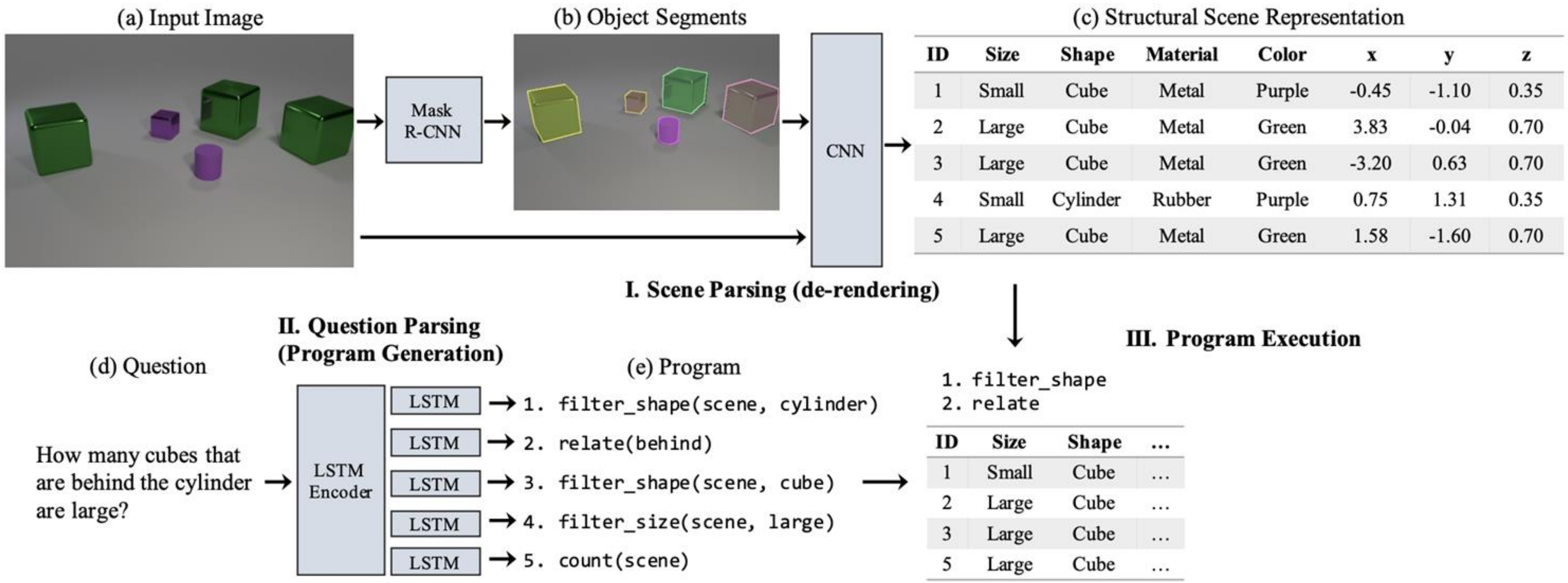


Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

Module Network V3: Neural-symbolic VQA

3) Program execution

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

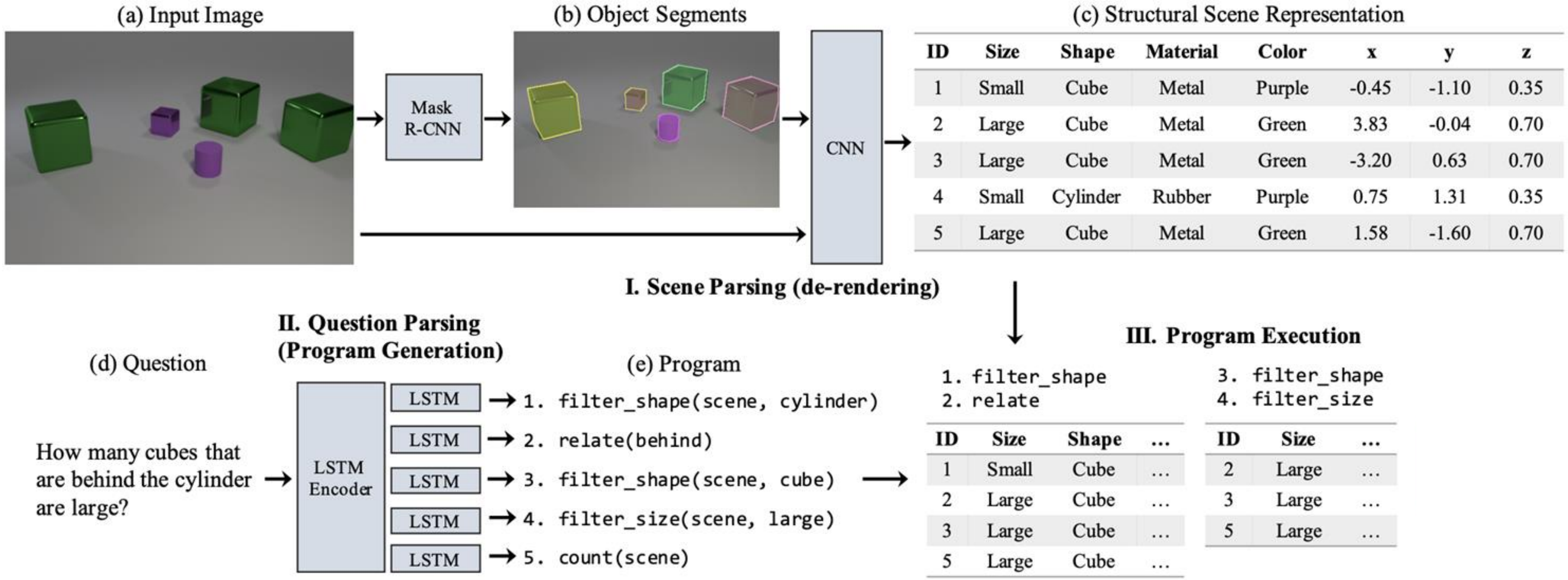


Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Module Network V3: Neural-symbolic VQA

3) Program execution

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

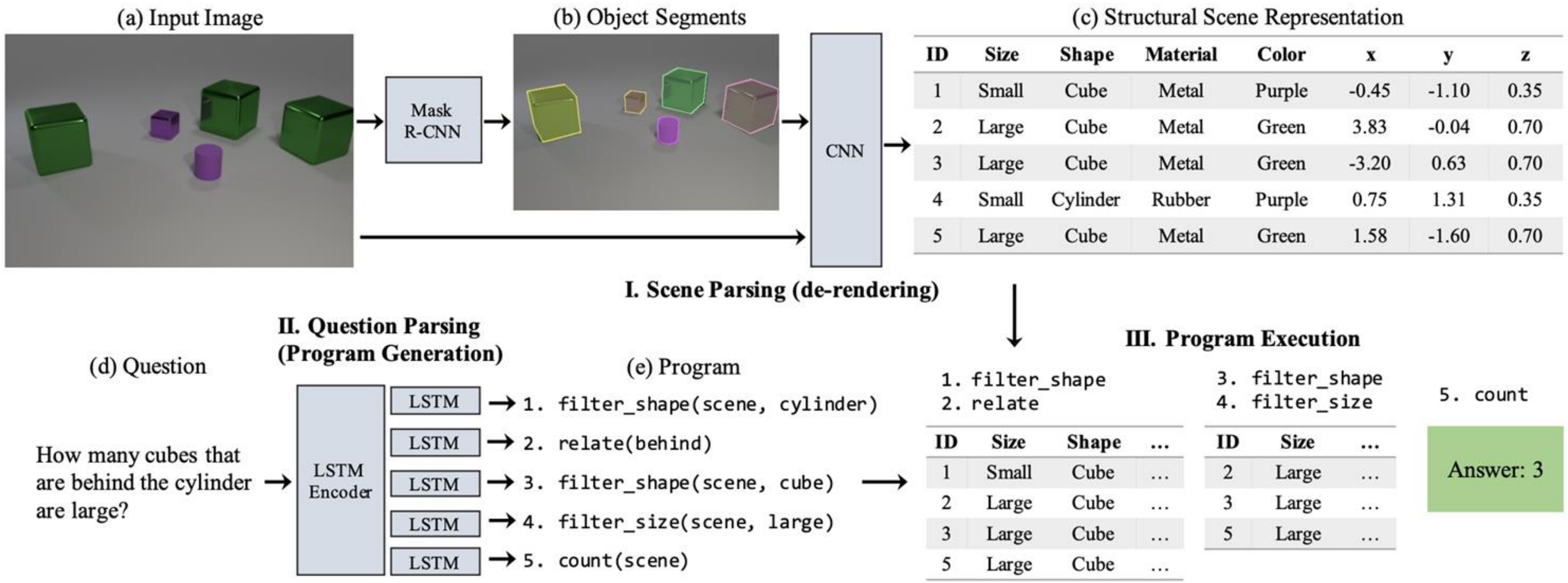


Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Module Network V3: Neural-symbolic VQA

3) Program execution

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

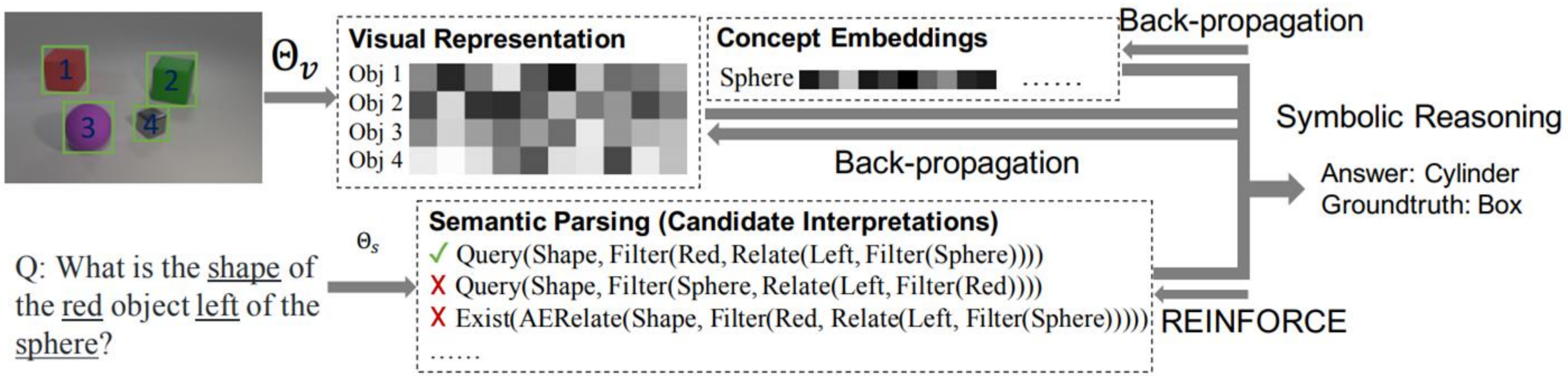


Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

The Neuro-symbolic Concept Learner

Extension from Neural-symbolic VQA:

Learns **visual concepts**, words, and semantic parsing of sentences without explicit supervision on any of them, but just by looking at images and reading paired questions and answers



Jiayuan Mao , et al. “The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision.” ICLR 2019

Module Networks V4: 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.

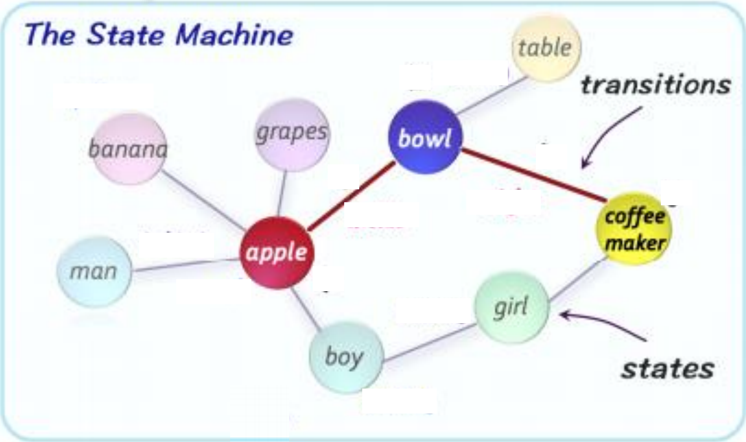
2. Treat the graph as a **state machine** and simulate iterative computation over it to *answer questions* or *draw inferences*.

3. Natural language questions are translated into *soft instructions* and used to perform sequential reasoning over the scene graph/state machine.

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

Module Networks V4: The Neural State Machine

Detect objects and create proximity graph

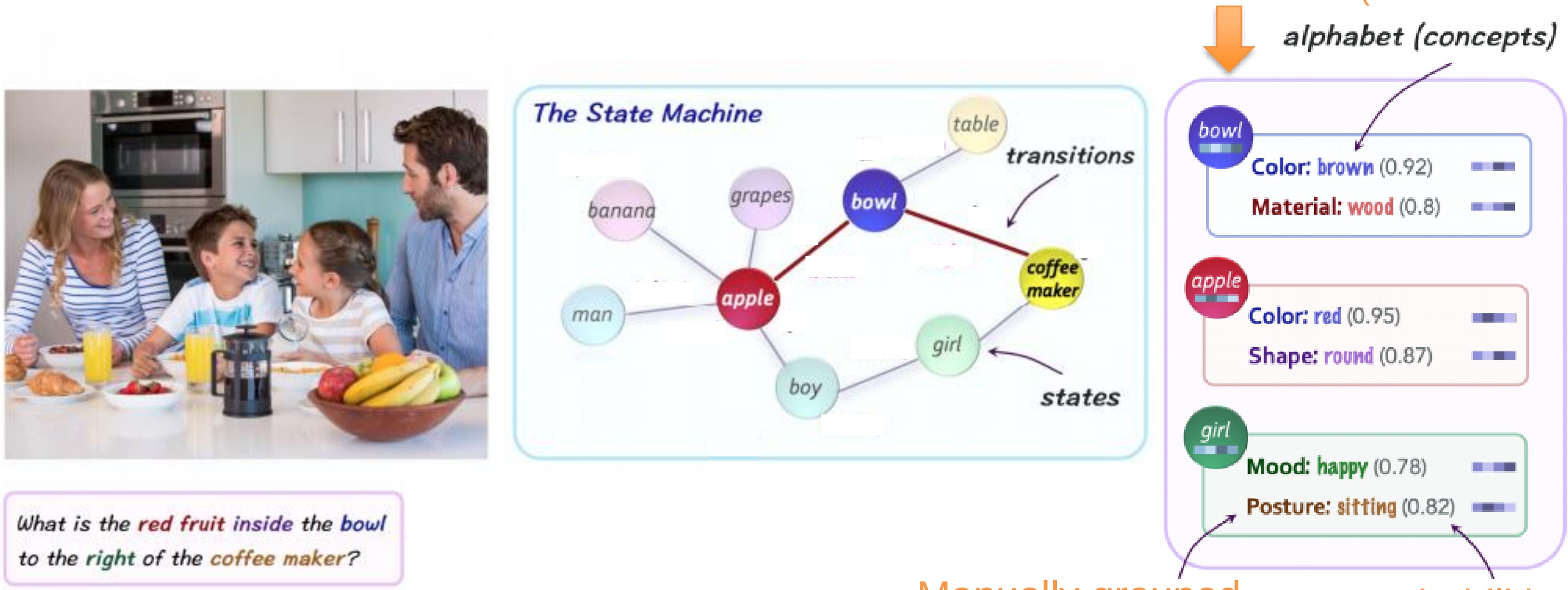


What is the red fruit inside the bowl to the right of the coffee maker?

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

Module Networks V4: The Neural State Machine

Pre-trained an alphabet of concepts
(Visual Genome)



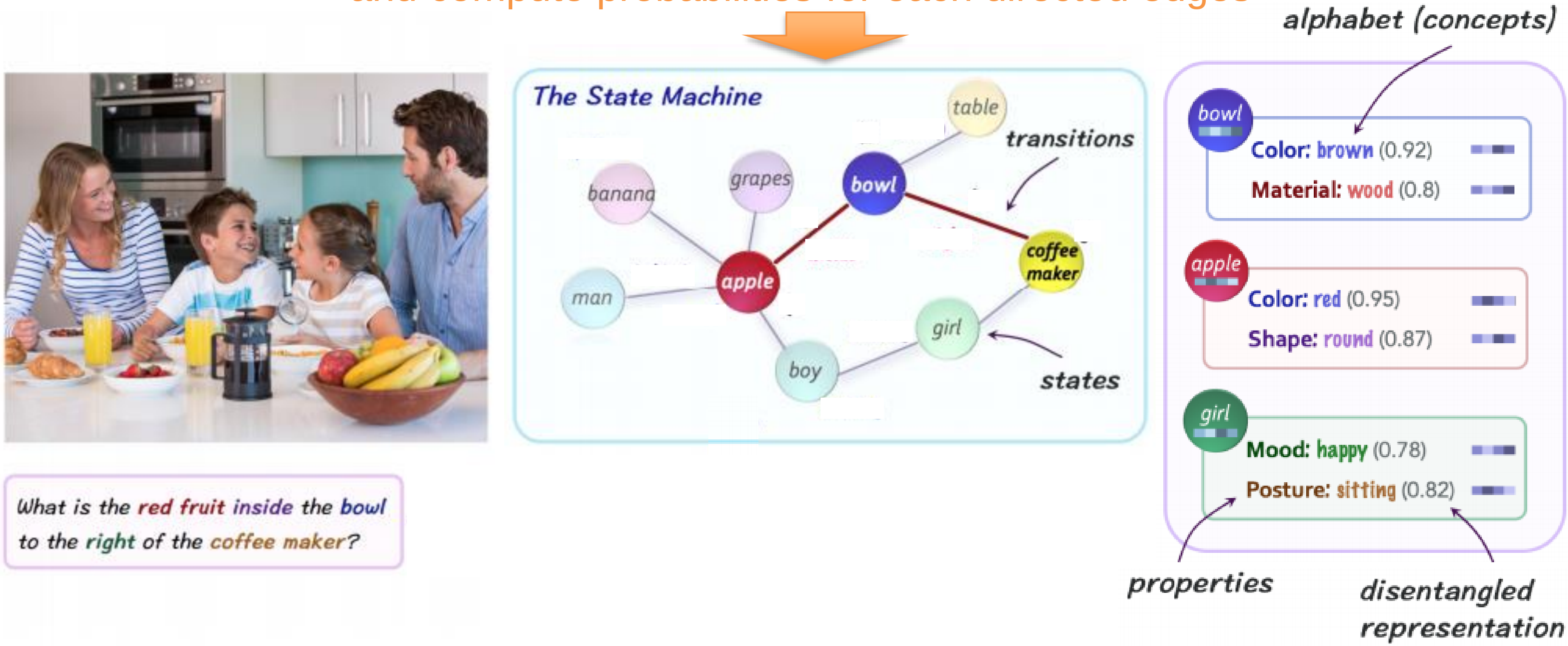
Manually grouped by "properties"

Probabilities computed at runtime for each object instance

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

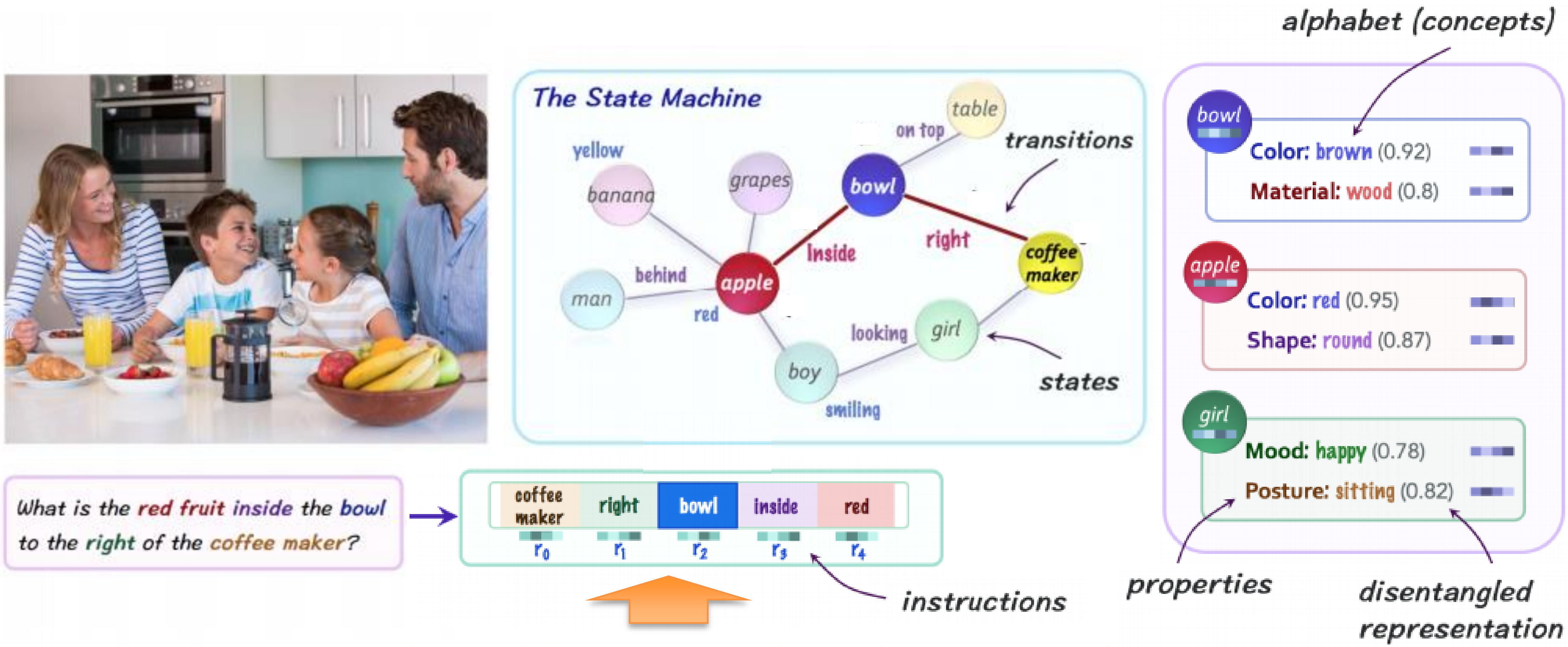
Module Networks V4: The Neural State Machine

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



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

Module Networks V4: The Neural State Machine

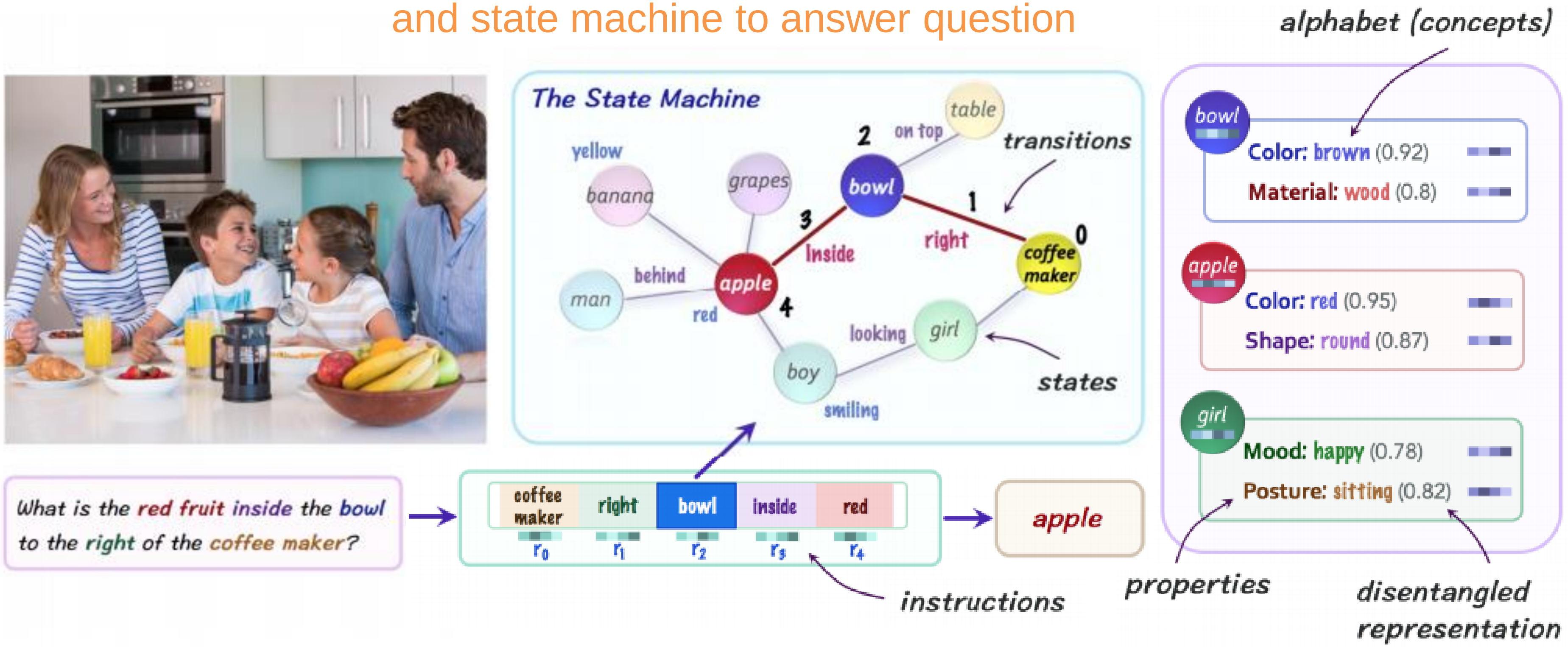


Translate each word in a concept-based representation
and group in a fixed number of instruction steps

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

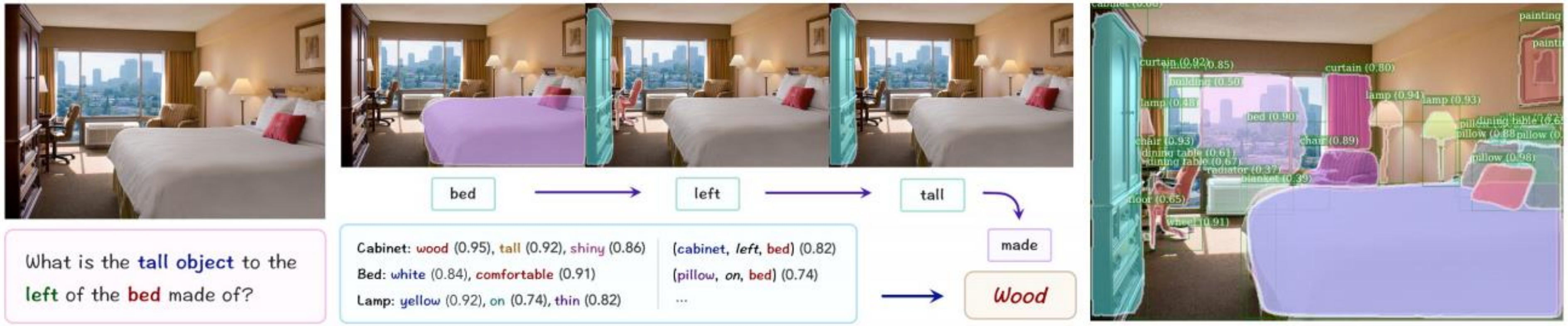
Module Networks V4: The Neural State Machine

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



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

Module Networks V4: The Neural State Machine



1. Compute the scene graph (blue boxes & image on the right)
2. Convert the question into a sequence of instructions (bed, left, tall, made)
3. Reason over the scene graph by attending to the relevant nodes using the instructions.

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

Studying Biases in VQA models

Studying Biases in VQA Models

	Prediction
What is in the basket?	banana
What is contained in the basket?	pizza
What can be seen inside the basket?	remote
What does the basket mainly contain?	paper

Why one question was correctly answered and not the others?

VQA models may be finding spurious correlations (e.g., confounding variables)

Research idea: Try to remove visual objects to see if they are confounding variables. + Propose a new evaluation metric to measure it.

Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

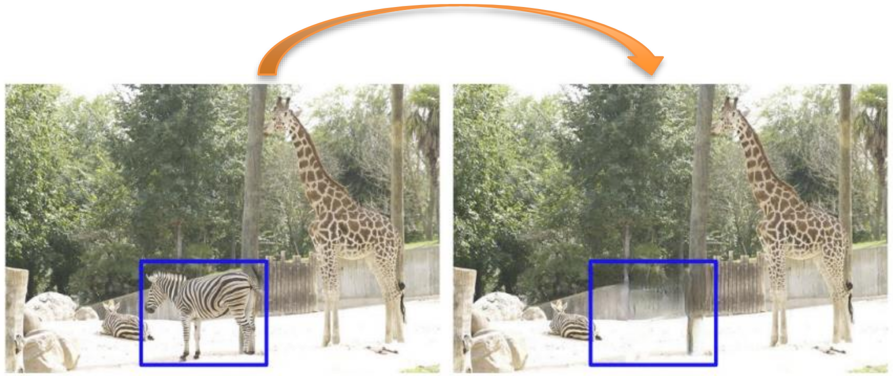
Studying Biases in VQA Models

***Consistency* metric:** Study the change in performance when individual objects are removed from the image

➡ using GAN to manipulate the images

Q: Is this a kitchen?

A: no *toilet removed*; A: no



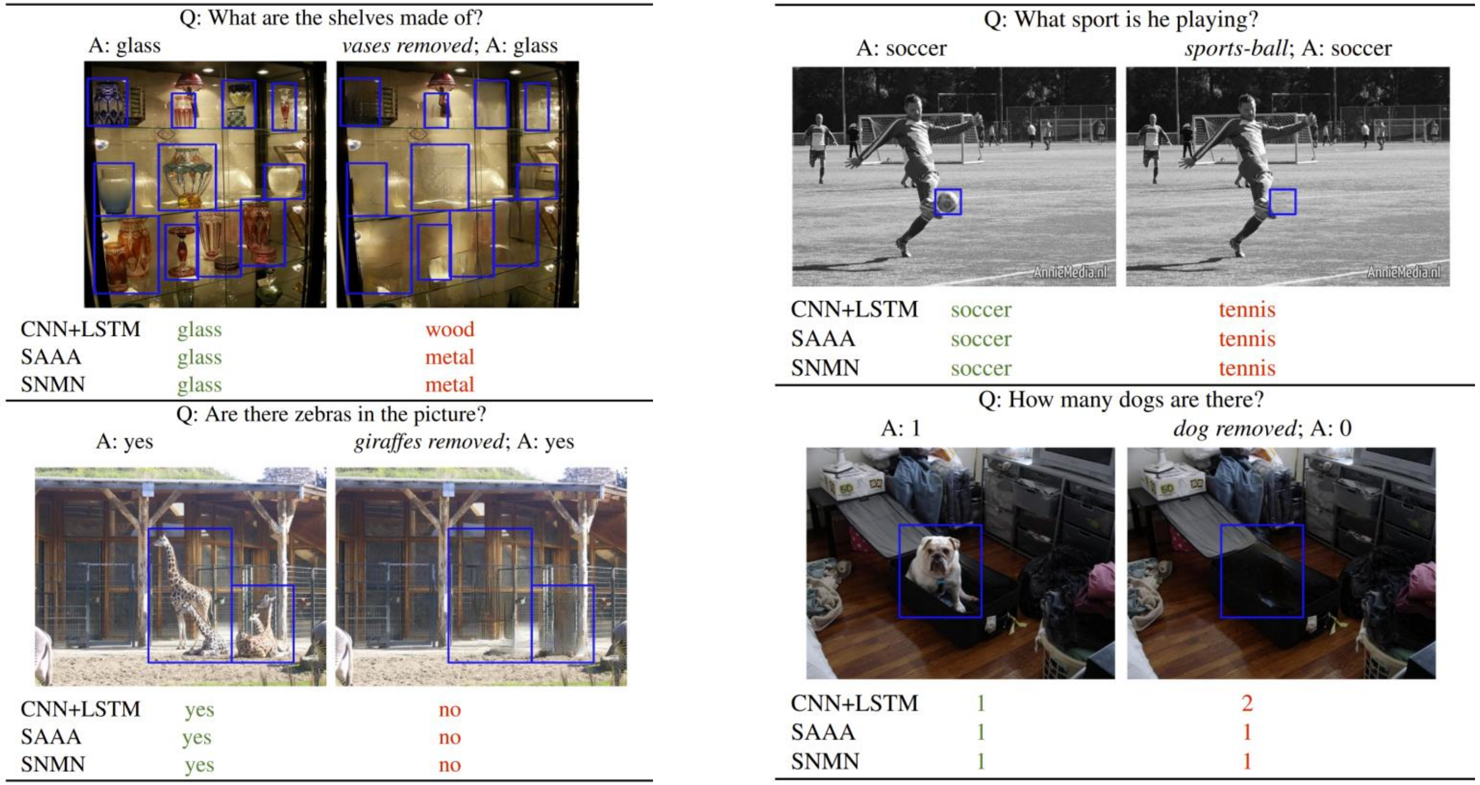
Q: How many zebras are there in the picture?

A: 2 *zebra removed* A: 1

Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

Studying Biases in VQA Models

State-of-the-art models often exploit spurious correlations…



Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

Studying Biases in VQA Models

Proposed solution: training the model on original VQA datasets plus synthetic datasets, consisting of images with removed objects.

Q:Is there a bowl on the table?

A: no — *cup removed*; A: no

	real	real+edit	real	real+edit
CL	no	no	yes	no
SAAA	no	no	yes	no
SNMN	no	no	yes	no

Q: How many people are in the water?

A: 1 — *person removed*; A: 0

	real	real+edit	real	real+edit
CL	1	1	1	0
SAAA	1	1	1	0
SNMN	1	1	1	0

Agarwal, Vedika, Rakshith Shetty, and Mario Fritz. "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing."

Visual Dialog

Visual Dialogue

Q

"is he wearing shorts ?"

I

H

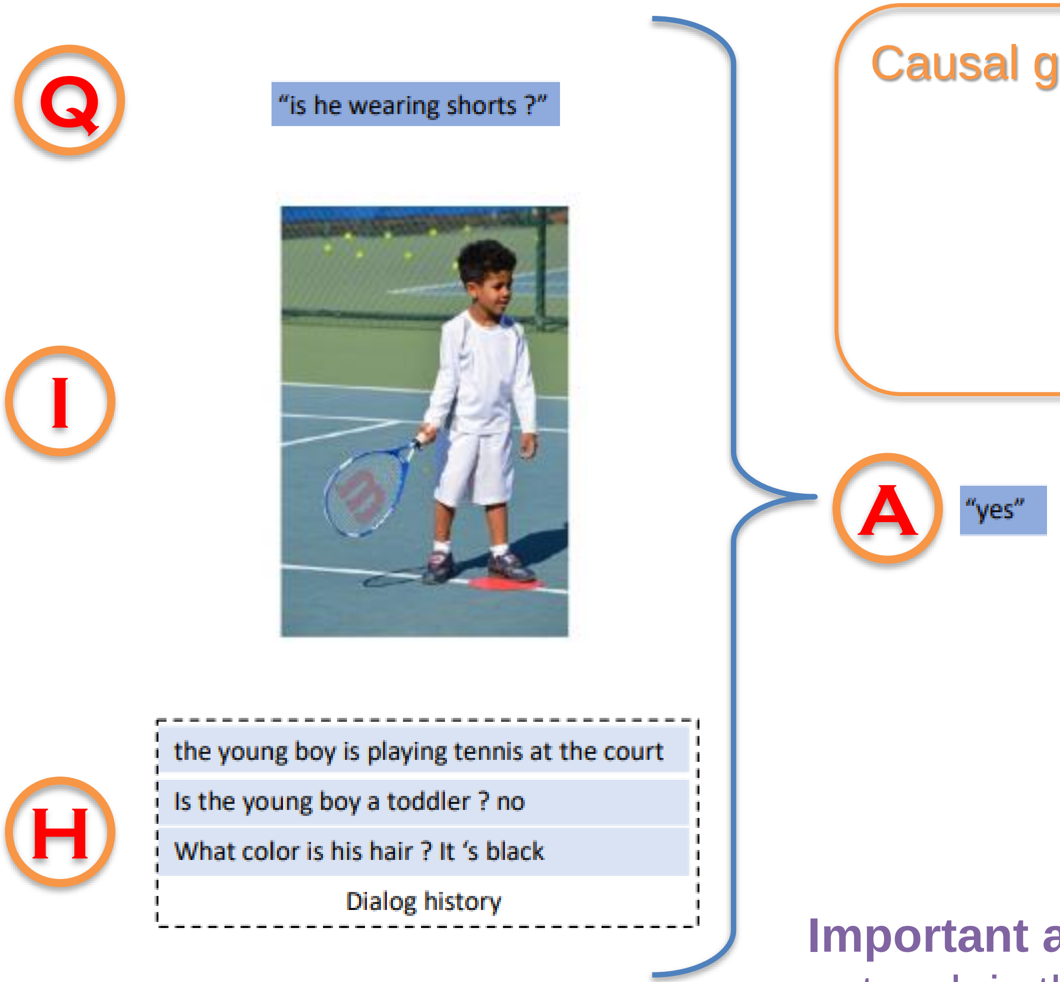
the young boy is playing tennis at the court

Is the young boy a toddler ? no

What color is his hair ? It 's black

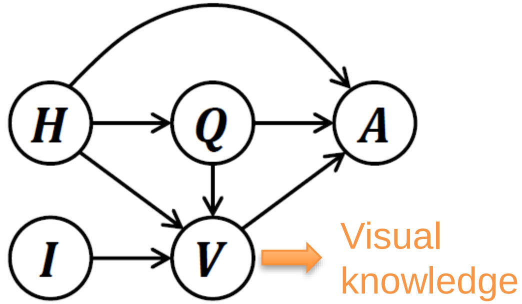
Dialog history

Visual Dialogue Expressed with Causal Graph



Causal graph: acyclic graph where nodes denote variables and edges denote causal relationships

How to represent this visual dialogue problem?



Important assumption: the output of a neural network is the *effect* of the input (the *cause*)

Two Causal Principles for Improving Visual Dialog

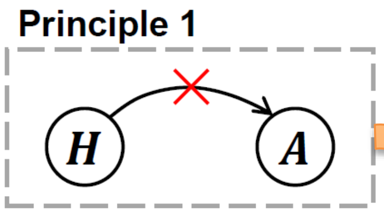
Two causal principles that are holding back Visual Dialolg models:

1. **Harmful shortcut bias** between dialog history (H) and the answer (A)
2. **Unobserved confounder** between H, Q and A leading to spurious correlations.

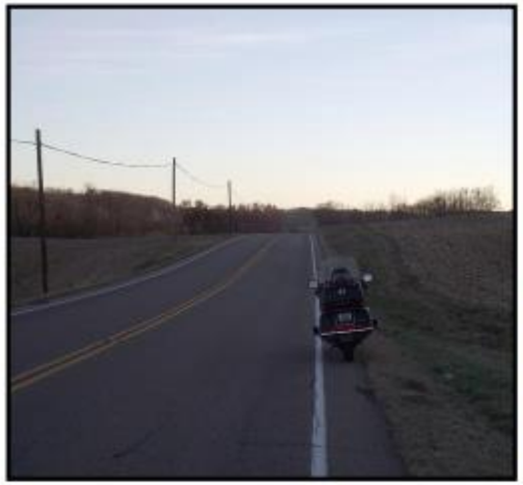
Qi, Jiaxin, et al. "Two causal principles for improving visual dialog.“ CVPR 2020

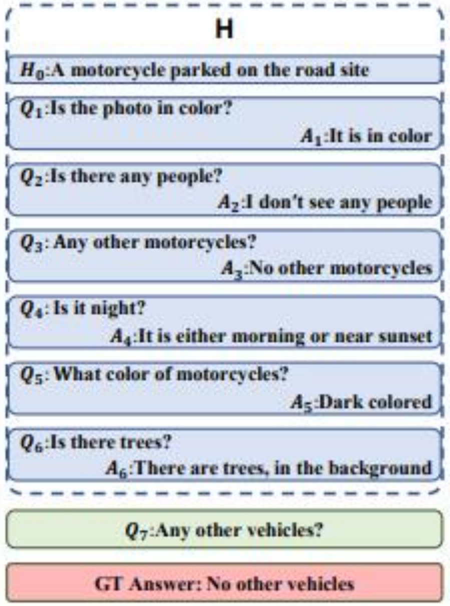
Two Causal Principles for Improving Visual Dialog

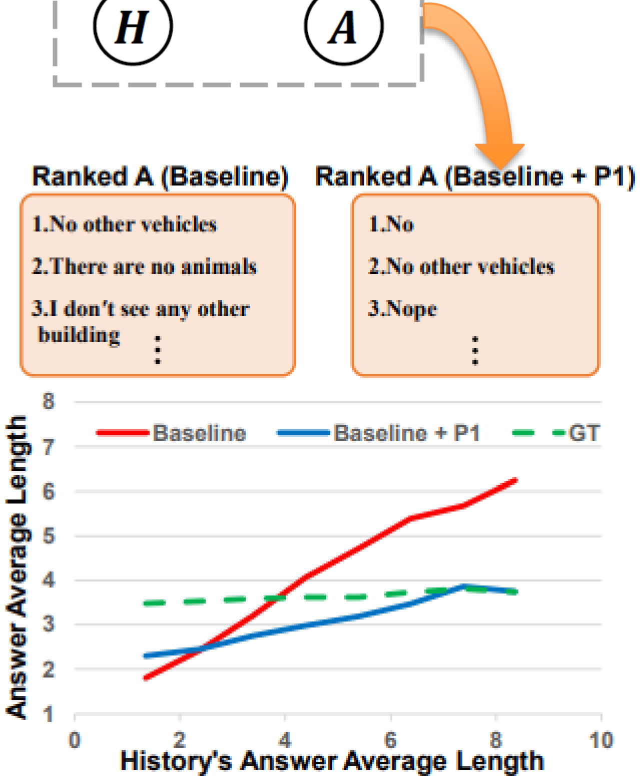
Principle 1: Harmful shortcut bias between dialog history (H) and the answer (A)



Dataset bias example:



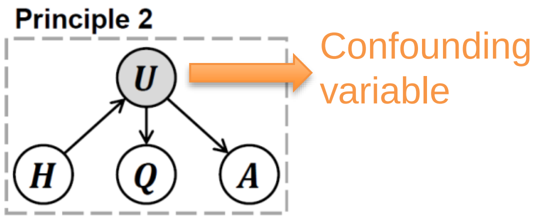




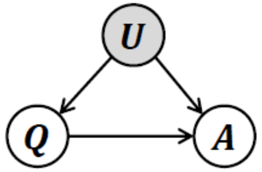
Qi, Jiaxin, et al. "Two causal principles for improving visual dialog.“ CVPR 2020

Two Causal Principles for Improving Visual Dialog

Principle 2: Unobserved confounder between H and A (as well as between H and Q) leading to spurious correlations.



Explaining confounding variable:



We may think that Q is primarily causing A, but U is a common cause for both Q and A

U has a *spurious* relation with Q and A

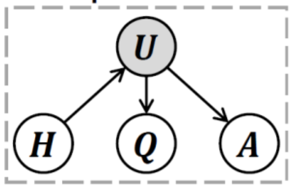
In our case, U is *unobserved*, and most likely because answerers (aka "users") could see the history.

Qi, Jiaxin, et al. "Two causal principles for improving visual dialog." CVPR 2020

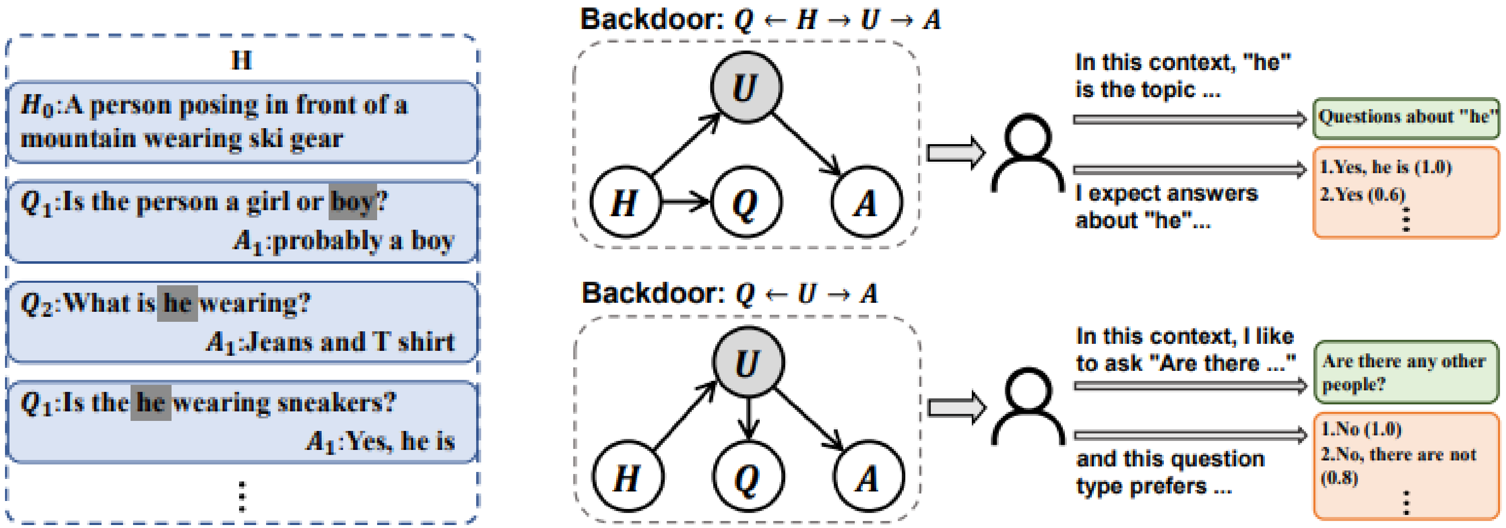
Two Causal Principles for Improving Visual Dialog

Principle 2: Unobserved confounder between H, Q and A leading to spurious correlations.

Principle 2



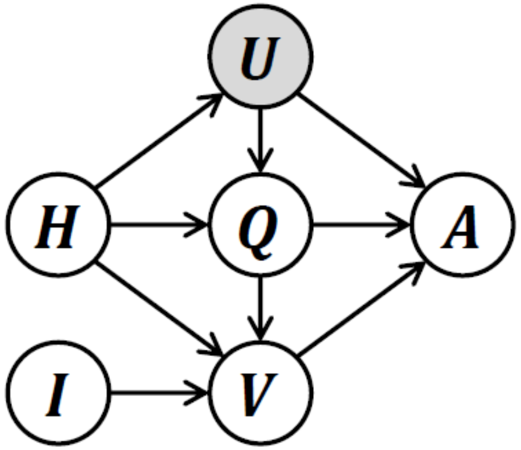
Dataset bias example:



Qi, Jiaxin, et al. "Two causal principles for improving visual dialog." CVPR 2020

Two Causal Principles for Improving Visual Dialog

Proposed method



1. Removes the **Harmful shortcut bias** between dialog history (H) and the answer (A)

2. Explicitly model the **unobserved confounder** between H, Q and A

Qi, Jiaxin, et al. "Two causal principles for improving visual dialog.“ CVPR 2020

Visual Dialog – Another Challenge

Hypothesis: The failure of visual dialog is caused by the inherent weakness of single-step reasoning.

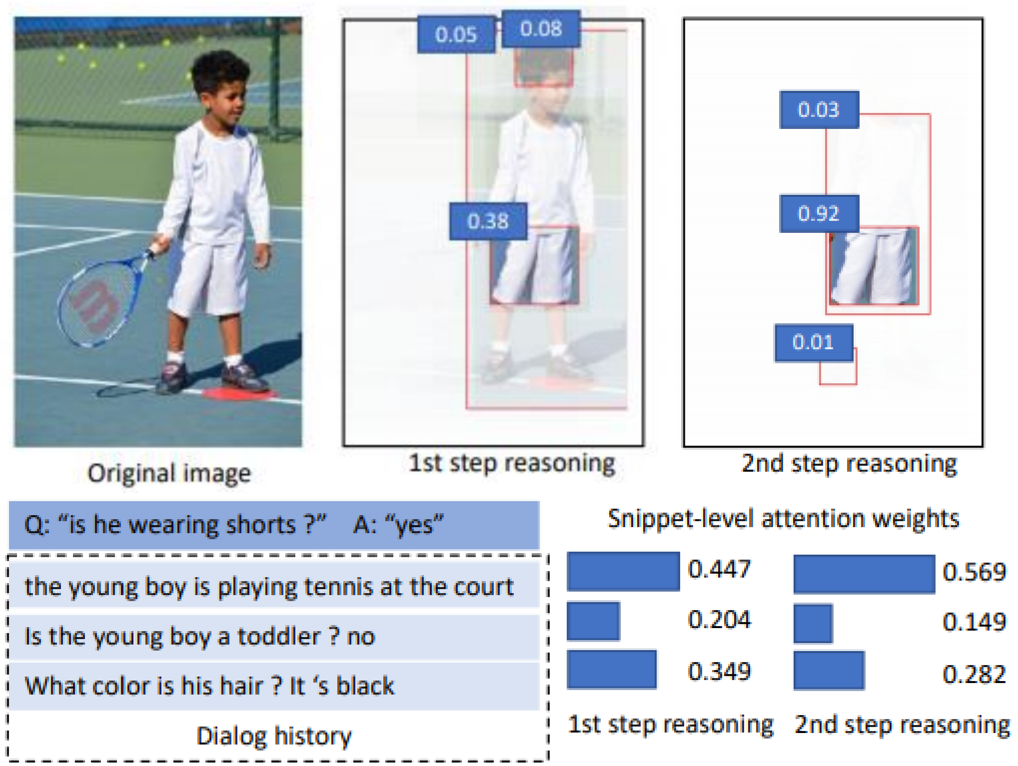
Intuition: Humans take a first glimpse of an image and a dialog history, before *revisiting* specific parts of the image/text to understand the multimodal context.

Proposal: Apply *Multi-step reasoning* to visual dialog by using a recurrent (aka multi-step) version of attention (aka reasoning). This is done on both text and questions (aka, dual).

➡ Recurrent Dual Attention Network

Gan, Zhe, et al. "Multi-step reasoning via recurrent dual attention for visual dialog.“ ACL 2019

Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog



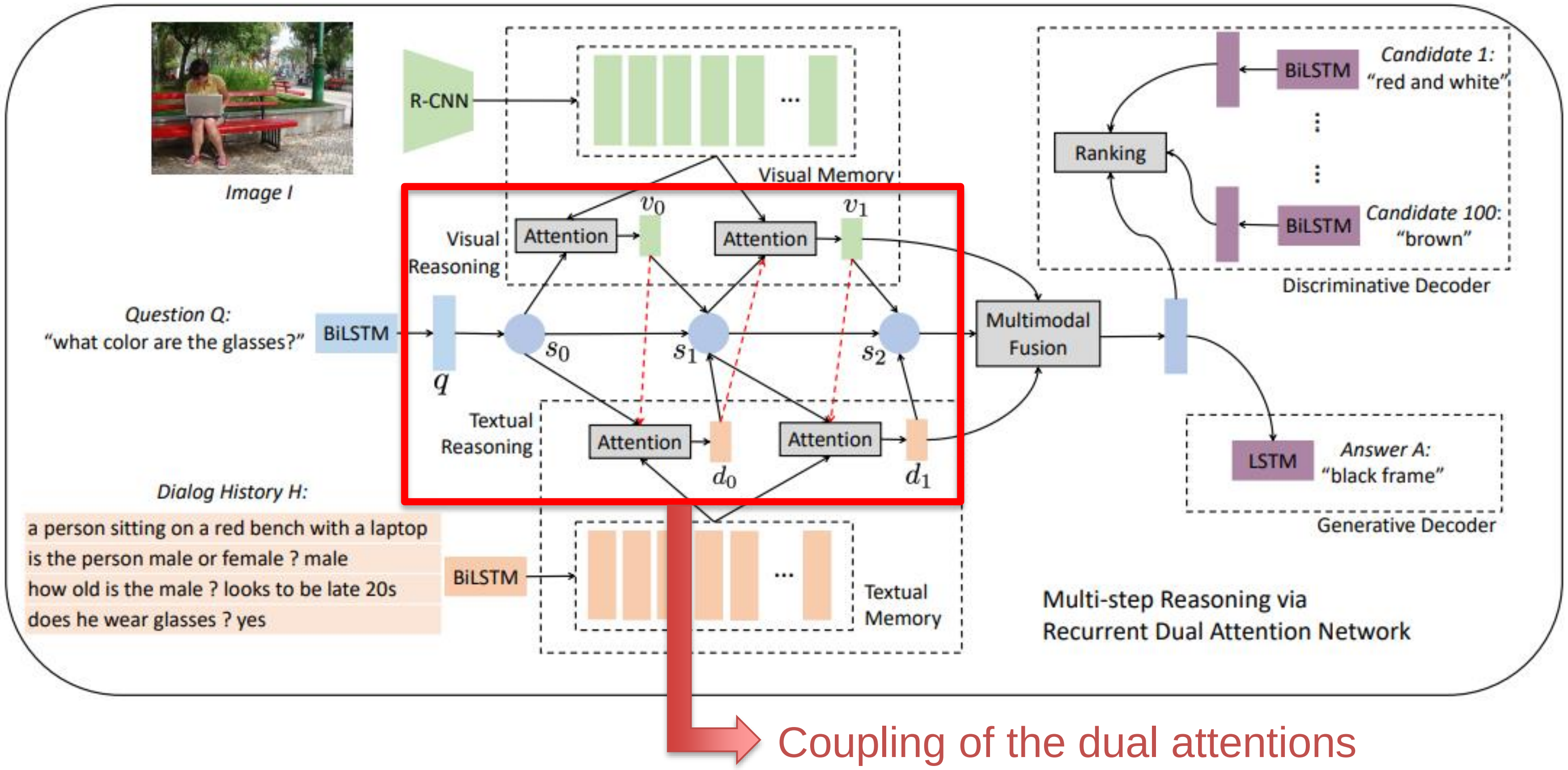
1st Step Reasoning: Attend to *all relevant* objects and dialog turns.

2nd Step Reasoning: Narrow down to context relevant regions (shorts, young boy).

In the 2nd step, the attention becomes sharper.

Gan, Zhe, et al. "Multi-step reasoning via recurrent dual attention for visual dialog." ACL 2019

Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog



Gan, Zhe, et al. "Multi-step reasoning via recurrent dual attention for visual dialog." ACL 2019