Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.2: Structured Representations and Reasoning

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Administrative Stuff

Second Project Assignment (Due Sunday 10/8)

Main goals:

1. Help clarify and expand your research ideas
   - Build qualitative intuitions by directly studying the original data
   - Perform analyses on your dataset, relevant to your research ideas
2. Understand the structure in your data and modalities
   - Perform analyses and visualizations to understand each modality
   - Study representations from language and visual modalities

Two types of analyses:

- Idea-oriented analyses
- Modality-oriented analyses

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 8/29 & 8/31	**Course introduction** • Multimodal core challenges • Course syllabus	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/5 & 9/7 Read due: 9/9	**Unimodal representations** • Dimensions of heterogeneity • Visual representations	**Unimodal representations** • Language representations • Signals, graphs and other modalities
Week 3 9/12 & 9/14 Read due: 9/16 Proj. Due: 9/13	**Multimodal representations** • Cross-modal interactions • Multimodal fusion	**Multimodal representations** • Coordinated representations • Multimodal fission
Week 4 9/19 & 9/21 *Proj. due: 9/24*	**Multimodal alignment and grounding** • Explicit alignment • Multimodal grounding	**Alignment and representations** • Self-attention transformer models • Masking and self-supervised learning
Week 5 9/26 & 9/28 Read due: 9/30	**Multimodal transformers – Part 1** • Language pretraining • Multimodal transformers	**Multimodal Reasoning** • Hierarchical and graph representations • Modular and neuro-symbolic models
Week 6 10/3 & 10/5 *Proj. due: 10/8*	**Multimodal transformers – Part 2** • Image and video transformers • Vision-language transformers	***Multimodal language grounding*** • Guest lecturer: Jack Hessel • Vision, language and grounding

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.2: Structured Representations and Reasoning

Louis-Philippe Morency

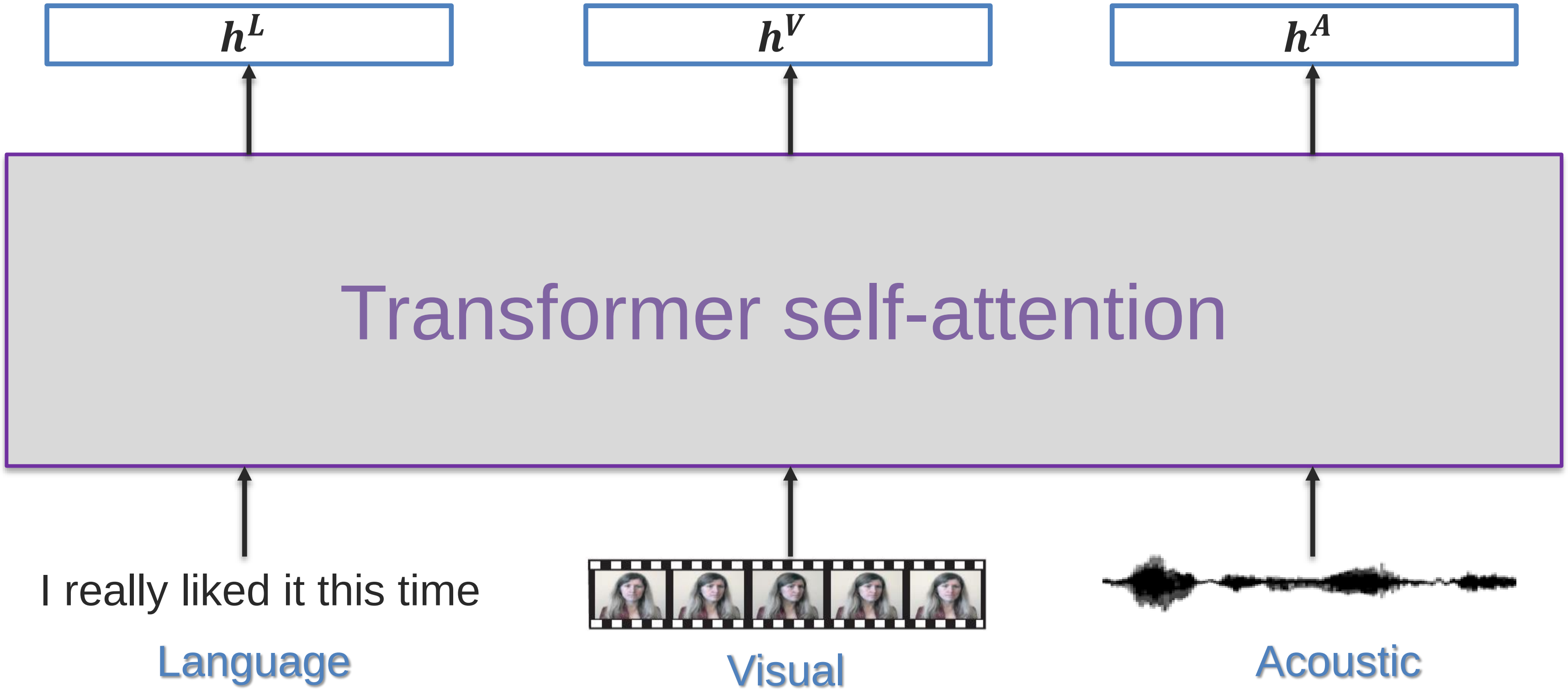
** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Objectives of today's class

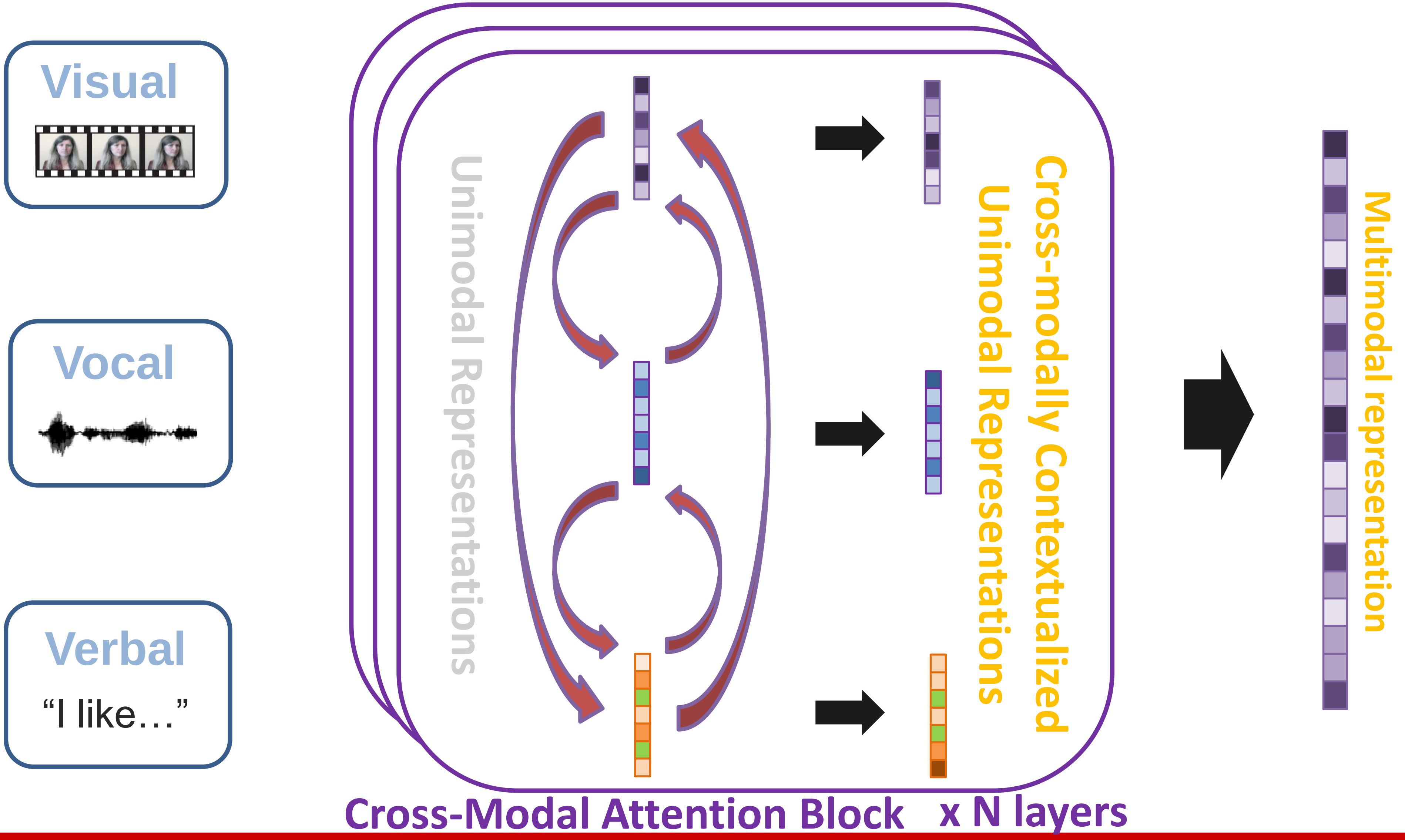
- Multimodal transformers
  - Modality-shift transformer (MAG-BERT)
- Sequence-to-sequence modeling with Transformers
- Going beyond sequences
  - Graph representations
    - Graph neural networks
  - Hierarchical representations
  - Modular representations
    - Neural module networks
    - Neuro-symbolic networks

Language-Vision Transformers

Simple Solution: Contextualized Multimodal Embeddings



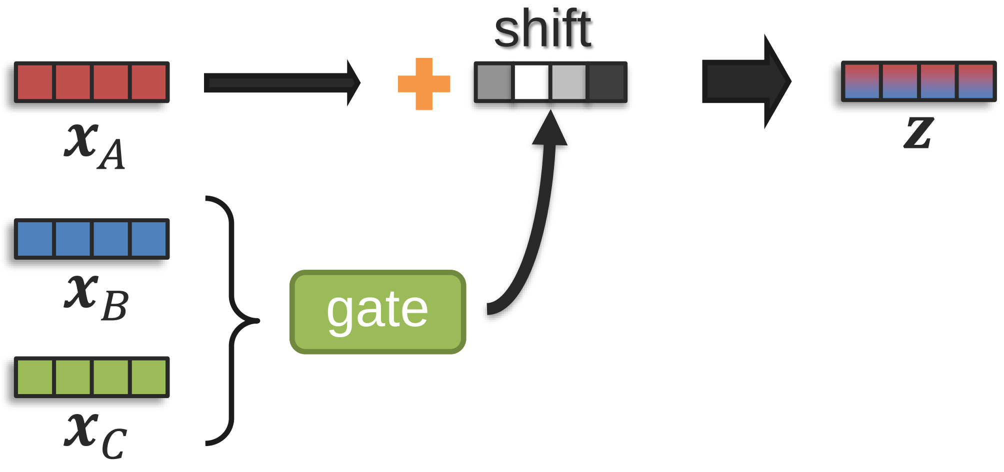
Multimodal Transformer – Pairwise Cross-Modal



Reminder: Modality-Shifting Fusion

Primary modality

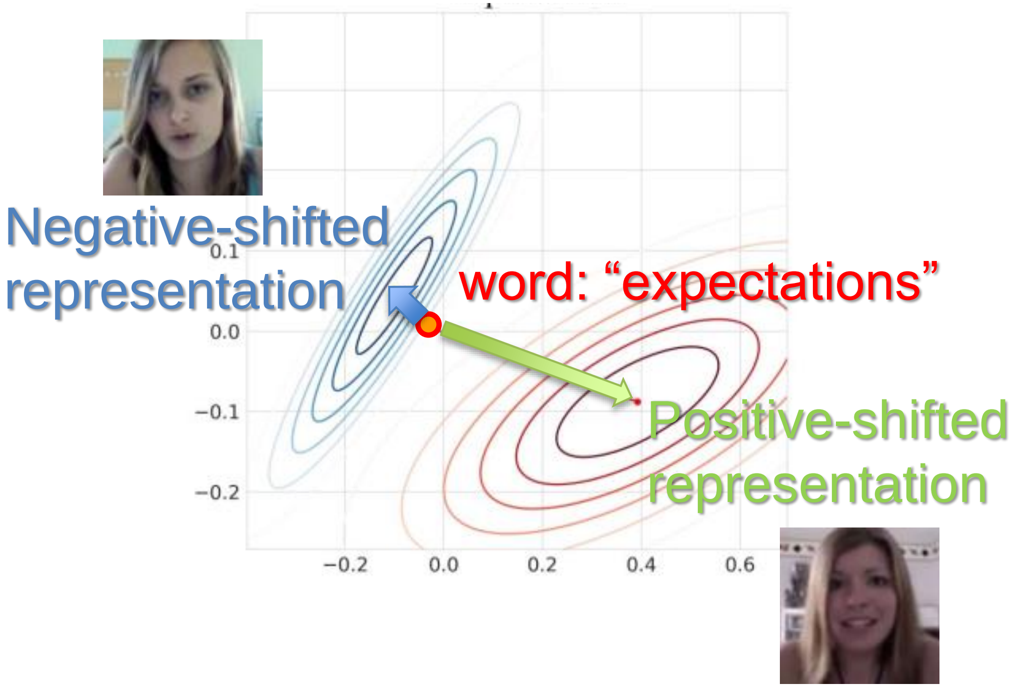
Secondary modalities



Example with language modality:

Primary modality: language

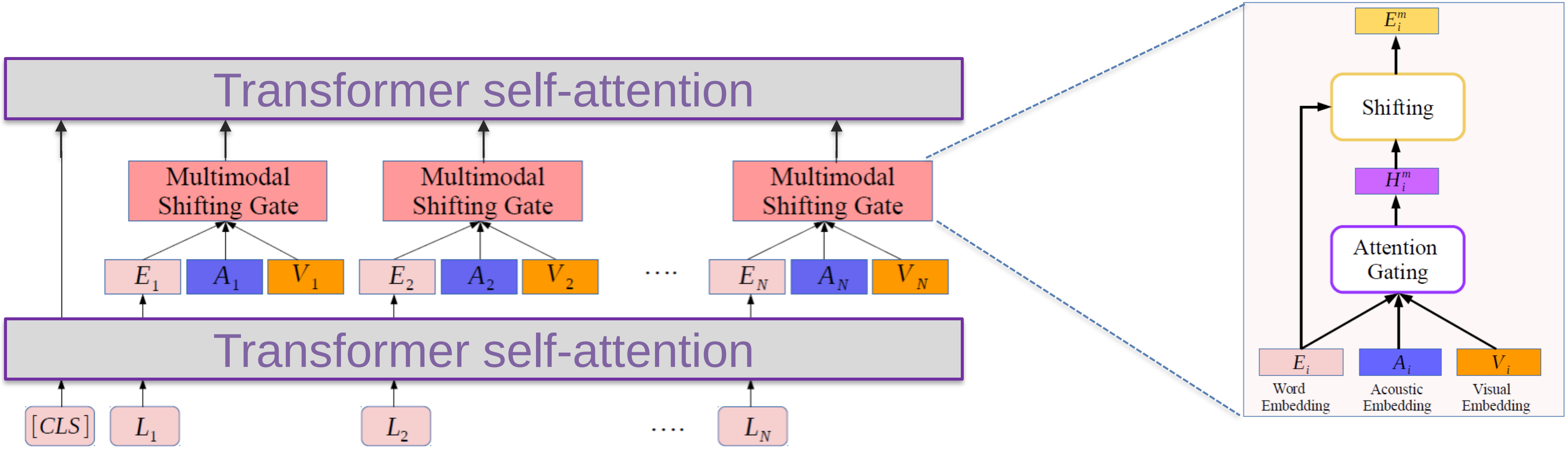
Secondary modalities: acoustic and visual



Wang et al., Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors, AAAI 2019

Modality-Shifting with Transformers

Multimodal Adaptation Gate (MAG) + BERT

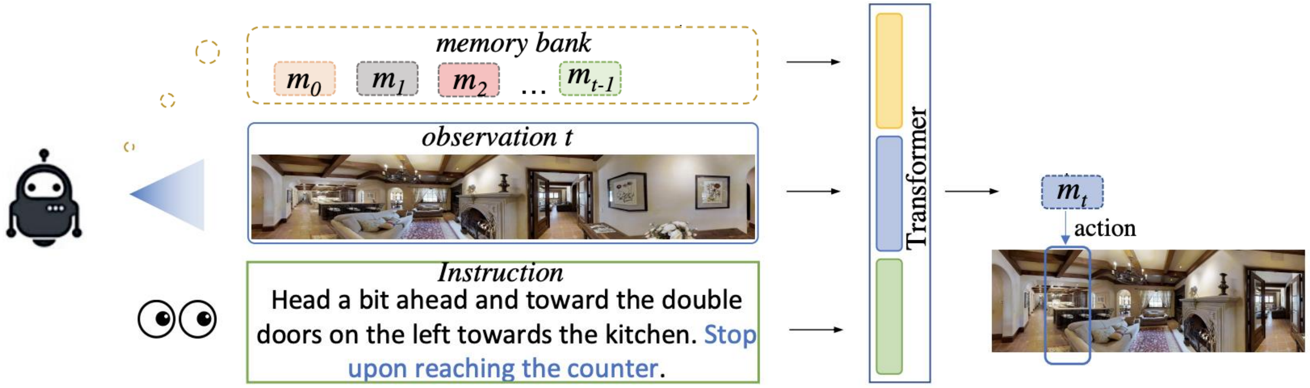


Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers, ACL 2020

Memory for Multimodal Sequences

Memory + aligned contextualized representations

Where have I visited previously?

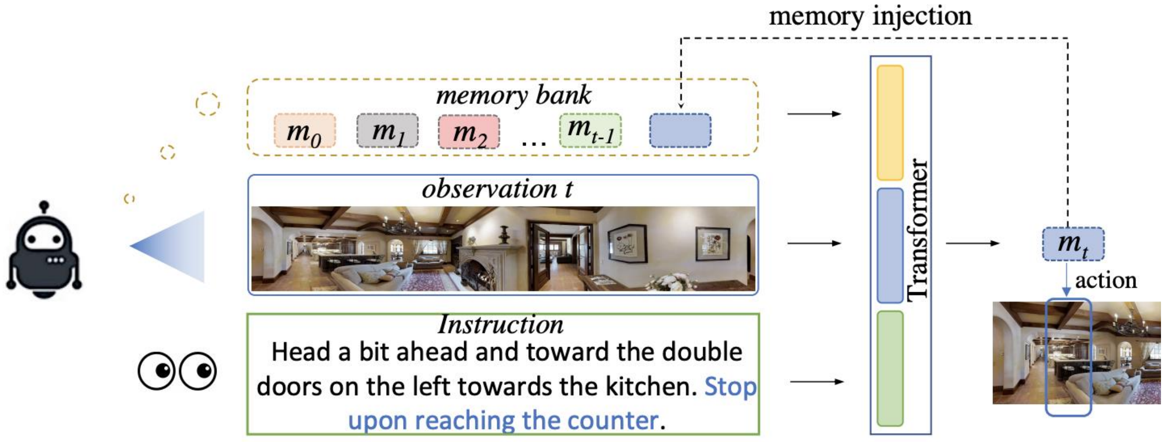


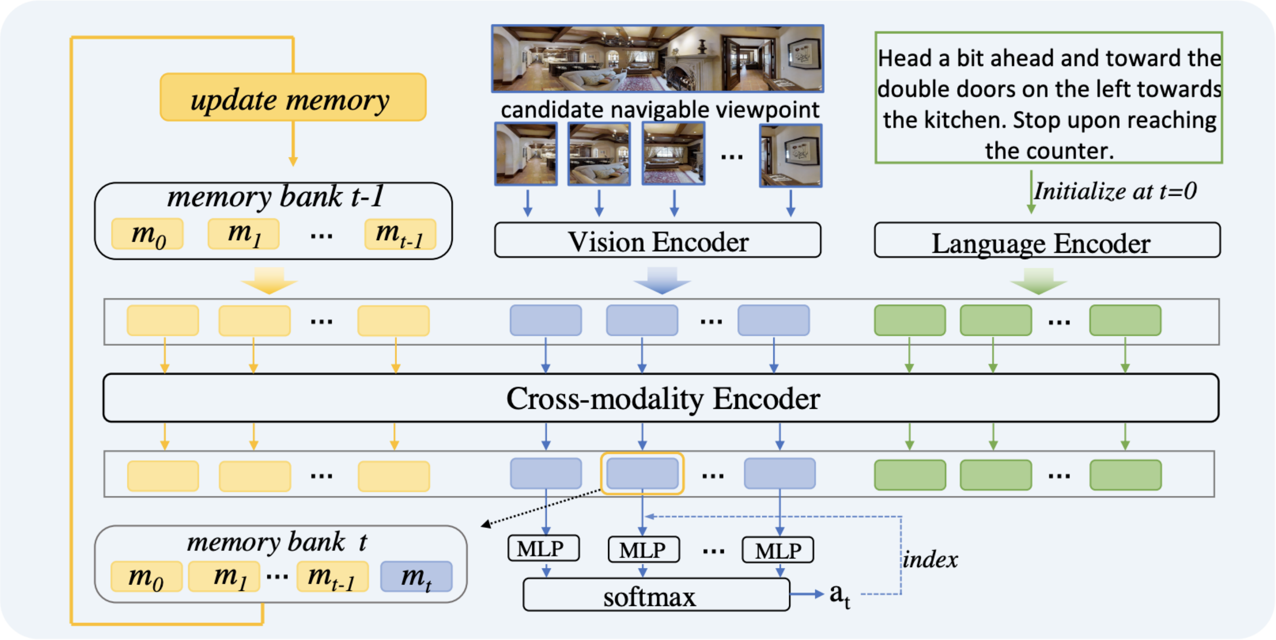
[Chen et al., History Aware Multimodal Transformer for Vision-and-Language Navigation. NeurIPS 2021]
[Lin et al., Multimodal Transformer with Variable-length Memory for Vision-and-Language Navigation. ECCV 2022]

Memory for Multimodal Sequences

Memory + aligned contextualized representations

Where have I visited previously?





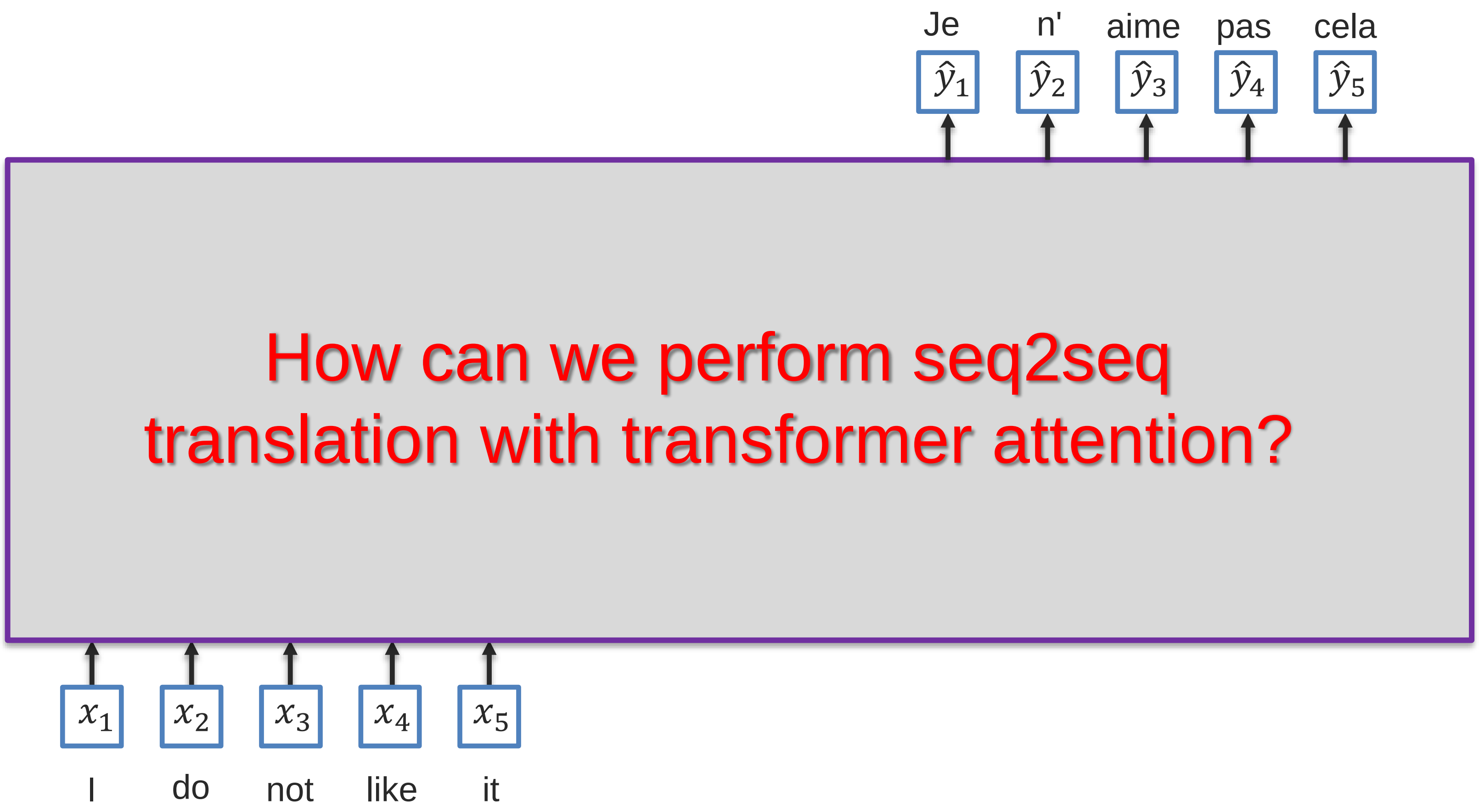
+ Contextualized representations

+ Memory mechanisms

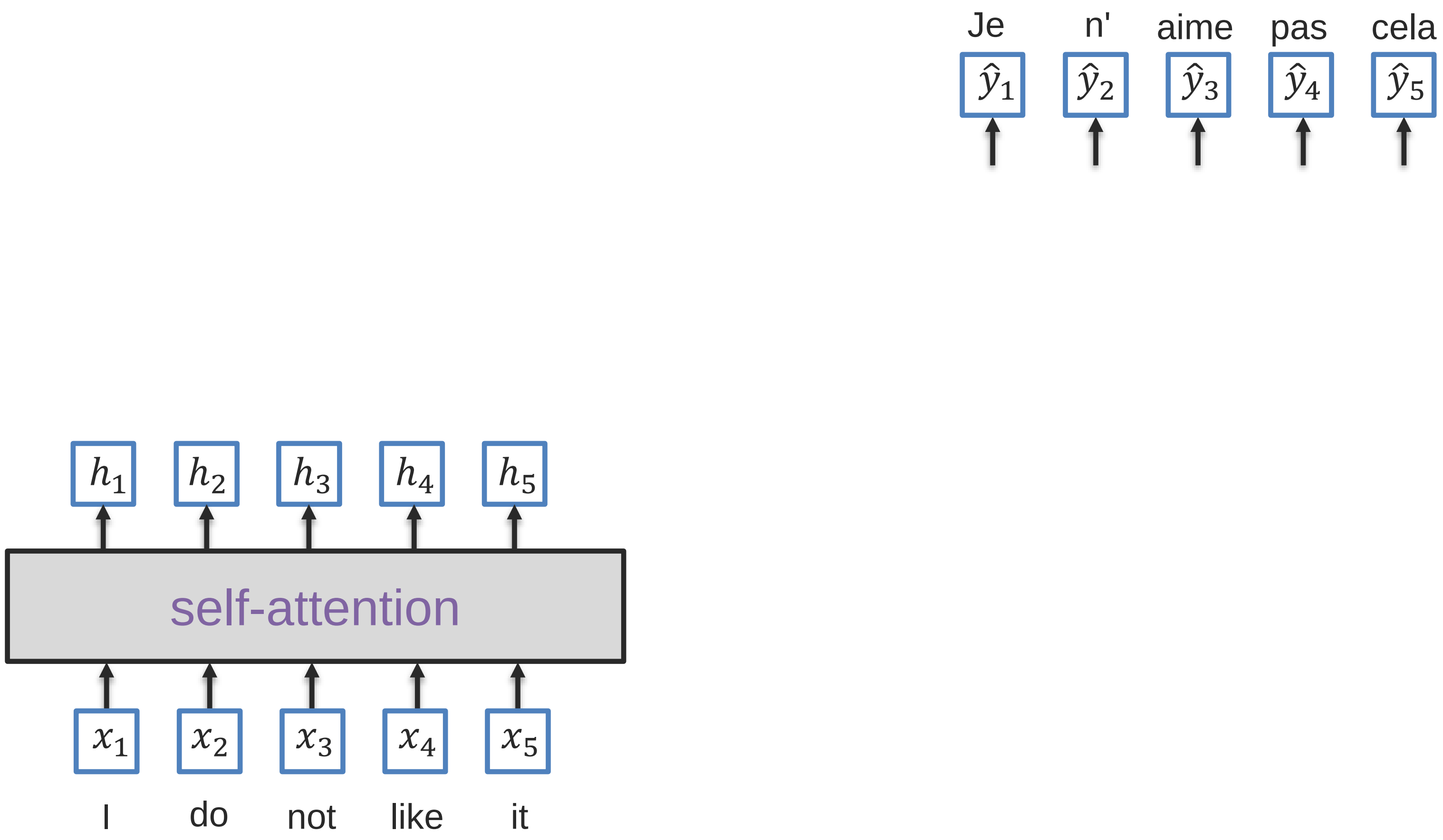
[Chen et al., History Aware Multimodal Transformer for Vision-and-Language Navigation. NeurIPS 2021]
[Lin et al., Multimodal Transformer with Variable-length Memory for Vision-and-Language Navigation. ECCV 2022]

Sequence-to-Sequence Using Transformer

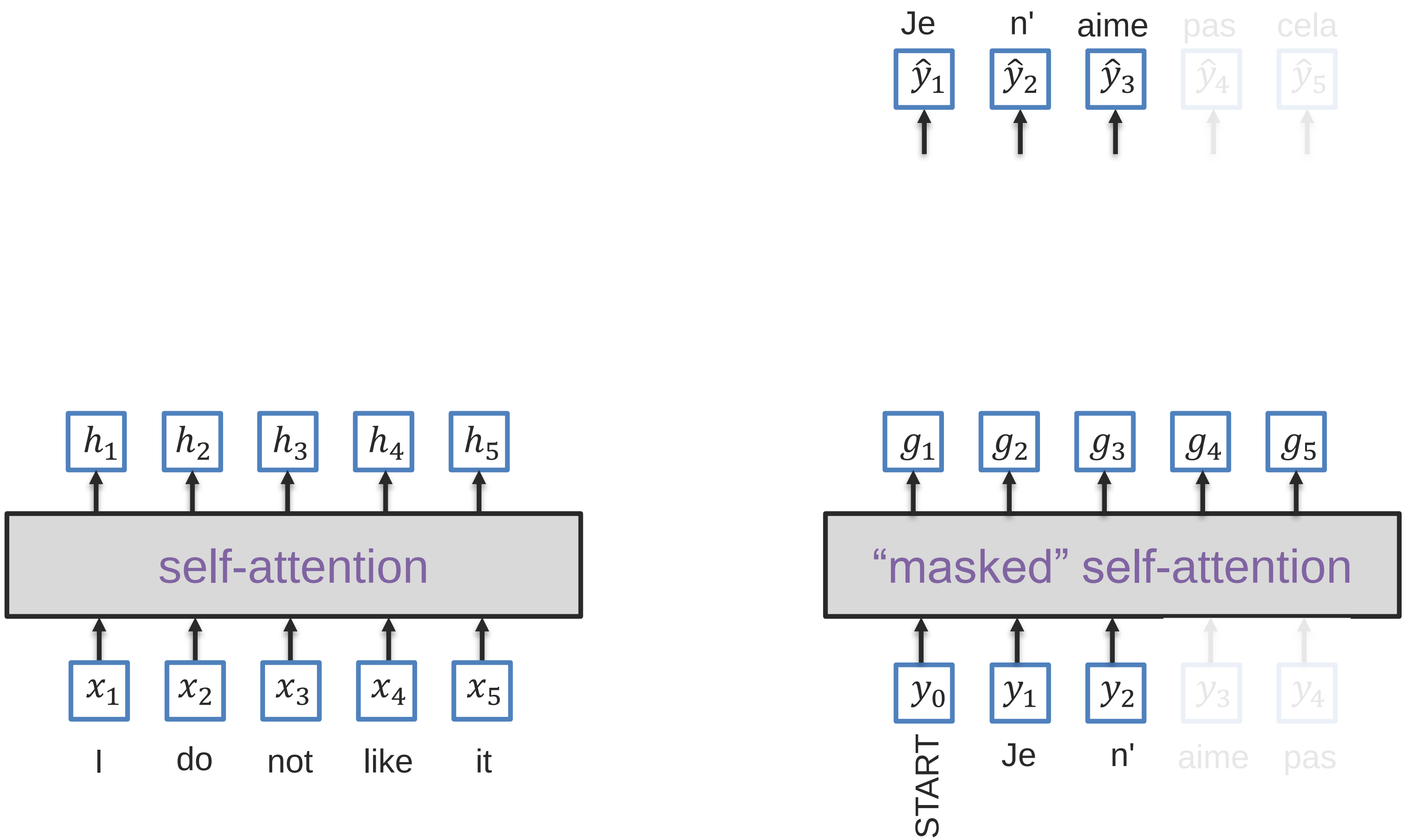
Sequence-to-Sequence Modeling



Seq2Seq with Transformer Attentions

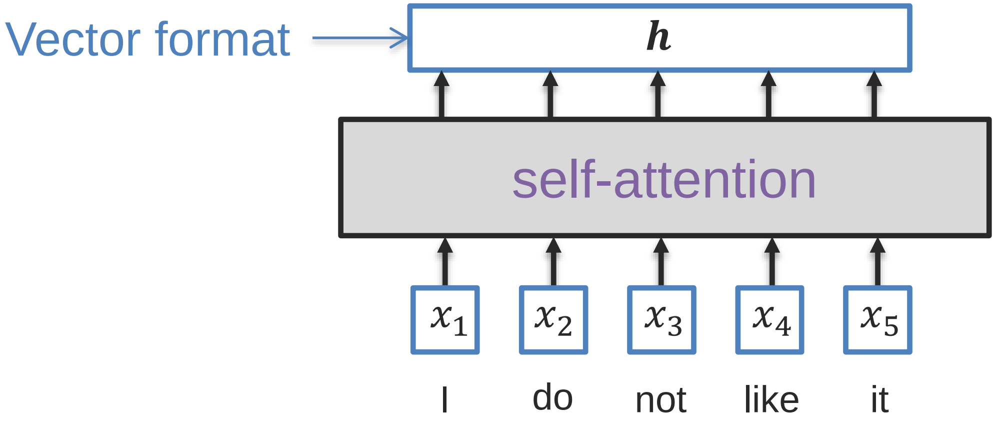


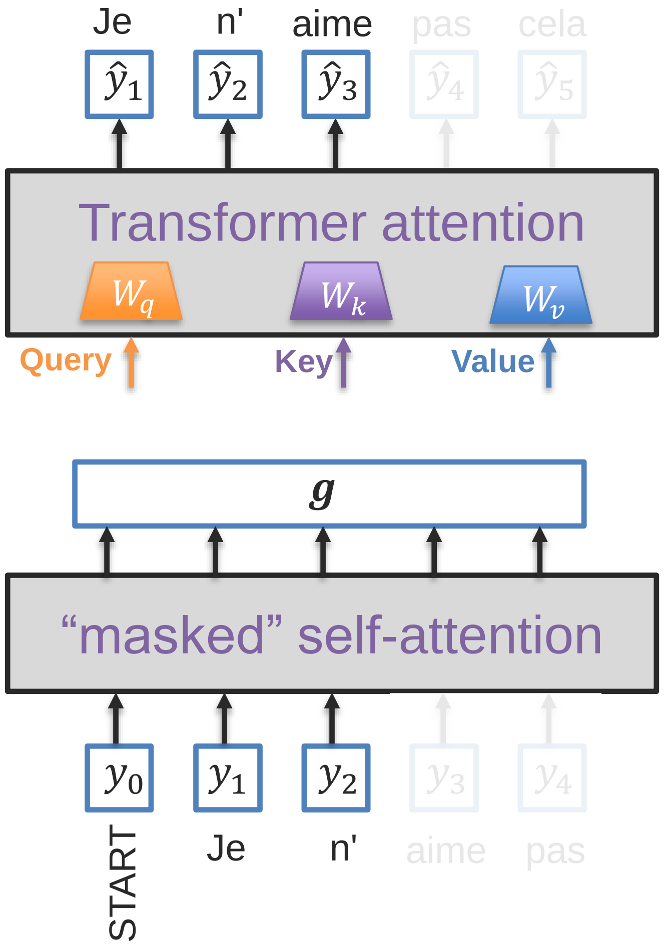
Seq2Seq with Transformer Attentions



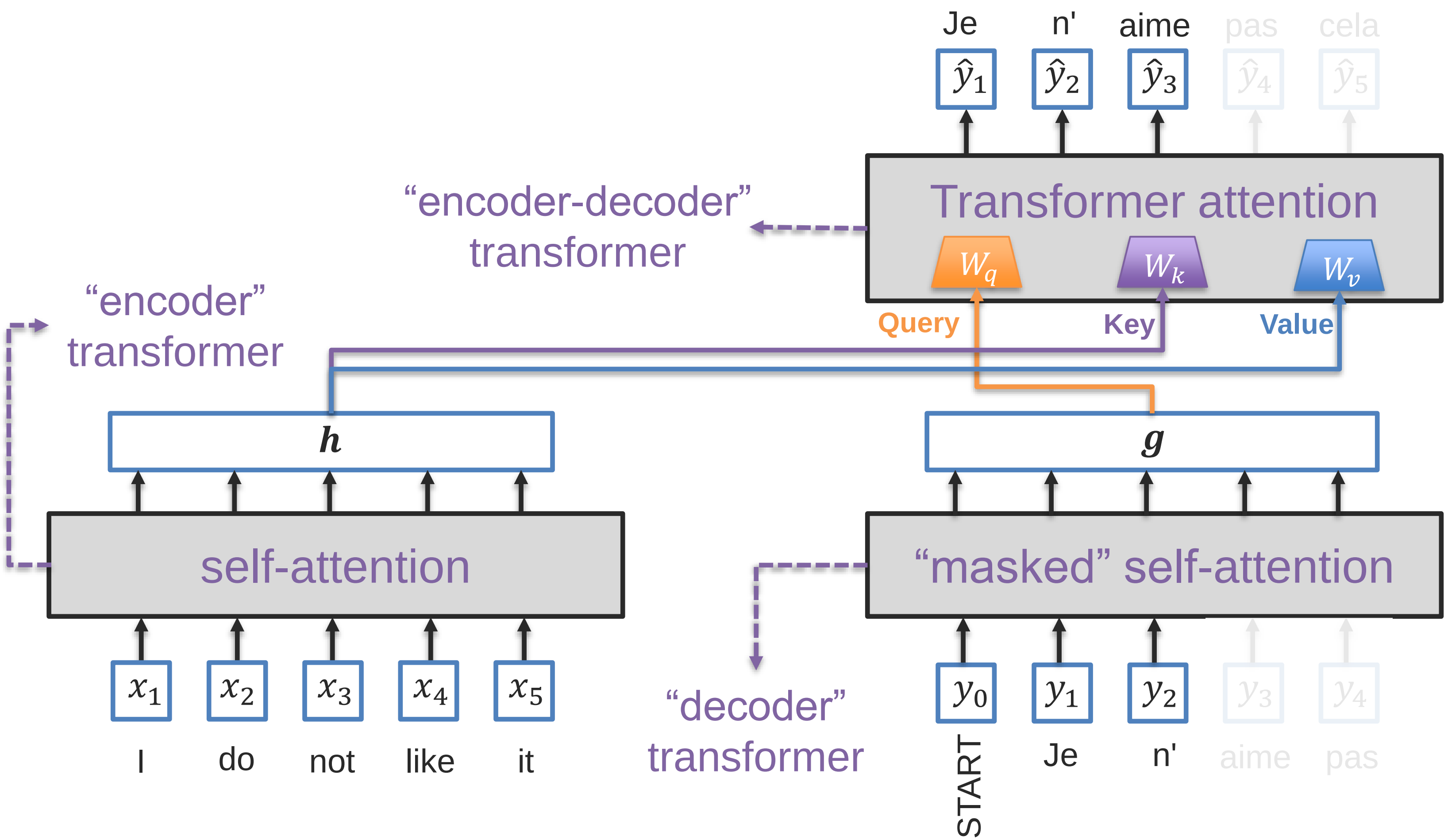
Seq2Seq with Transformer Attentions

How should we connect the encoder and decoder self-attention to the transformer attention?





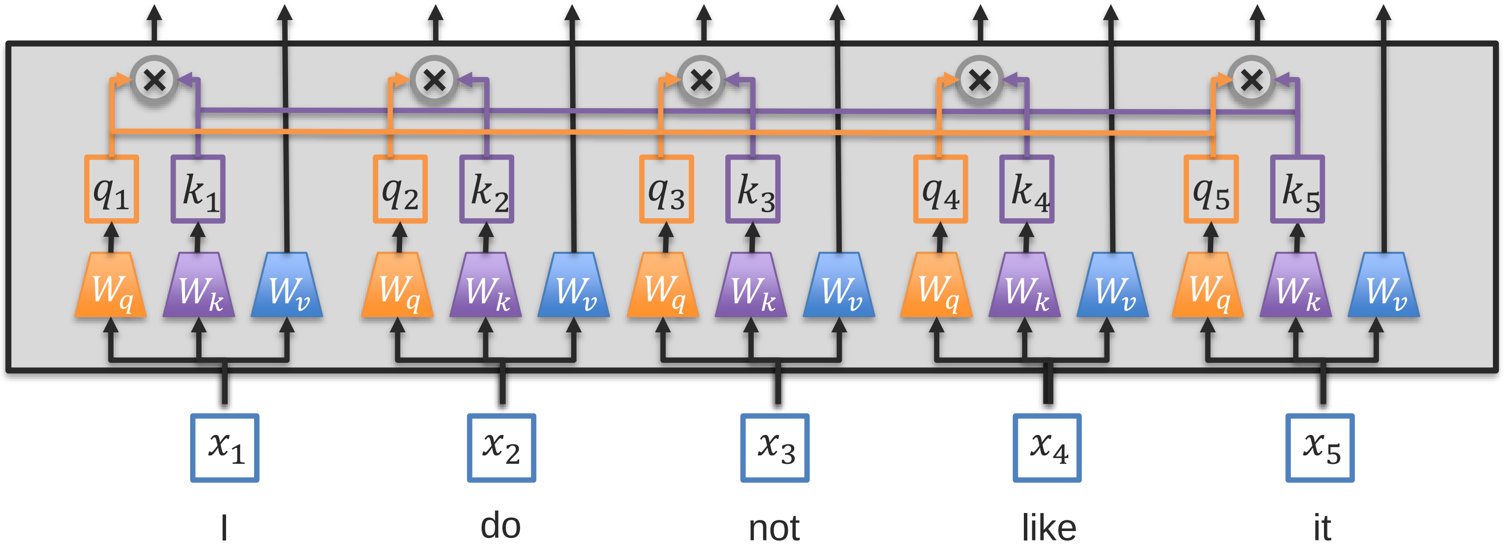
Seq2Seq with Transformer Attentions



Going Beyond Sequences: Graph Representations

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

Transformers – Fully-Connected Sequences

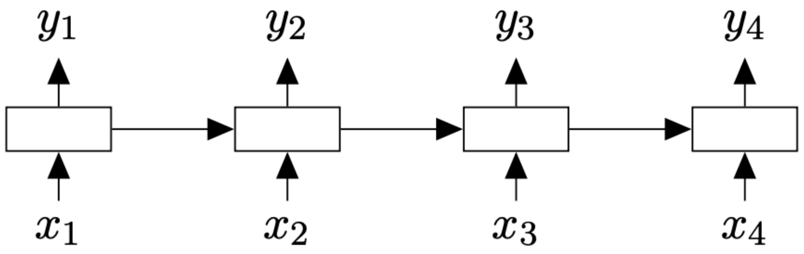


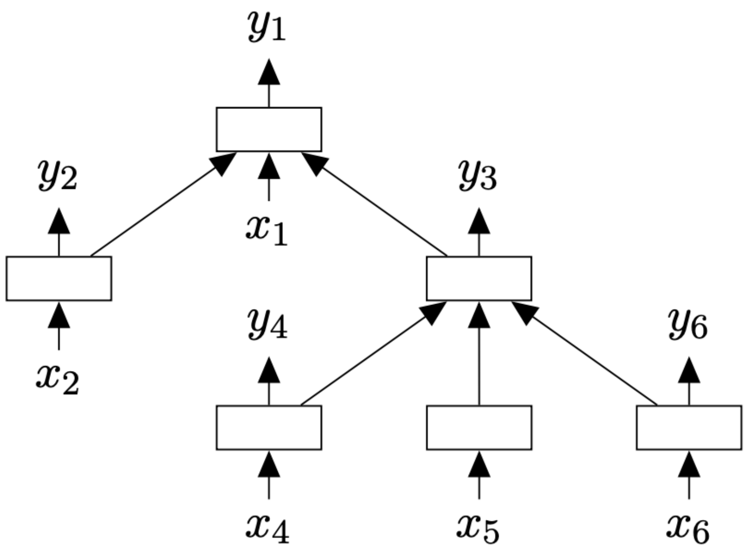
Should everything be connected to everything?

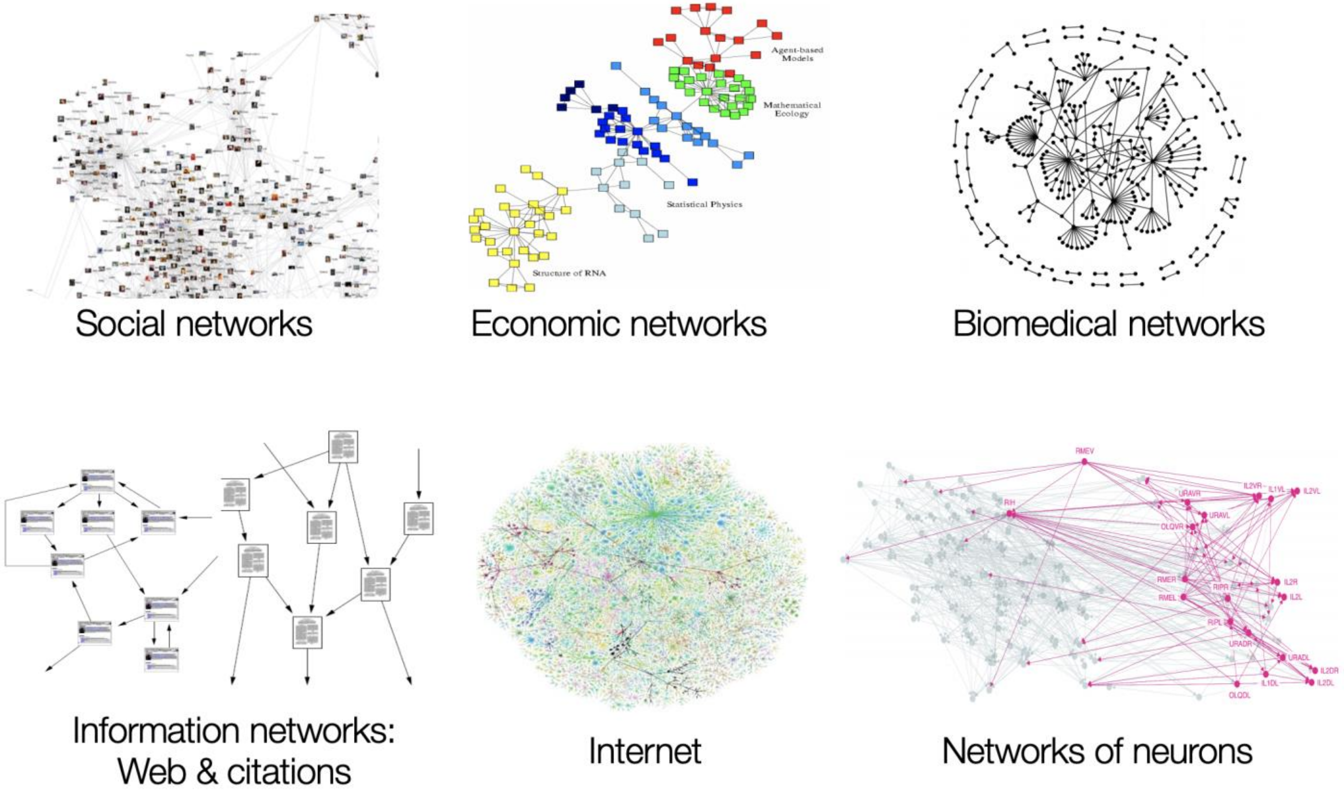
What if we have domain knowledge about connections?

Tree and Graph Networks

From linear chain models to tree and graph-structured models



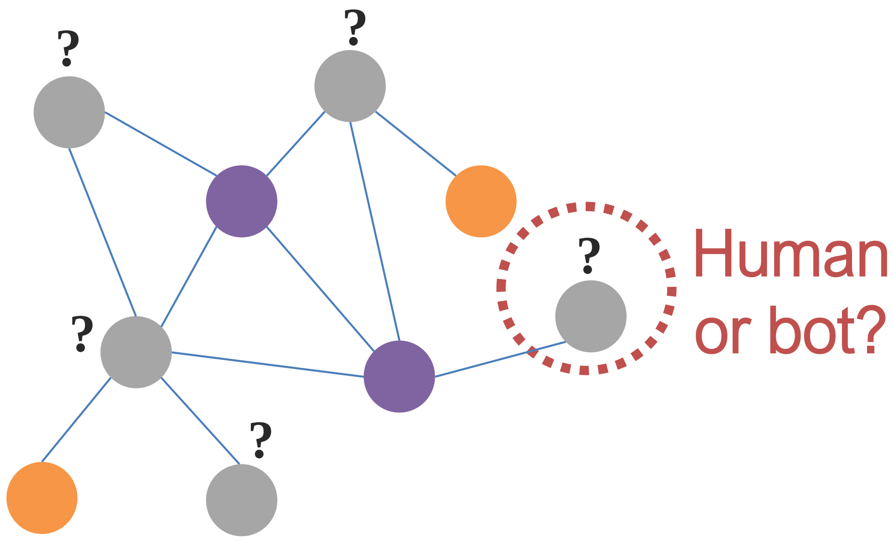




[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graphs – Supervised Task

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



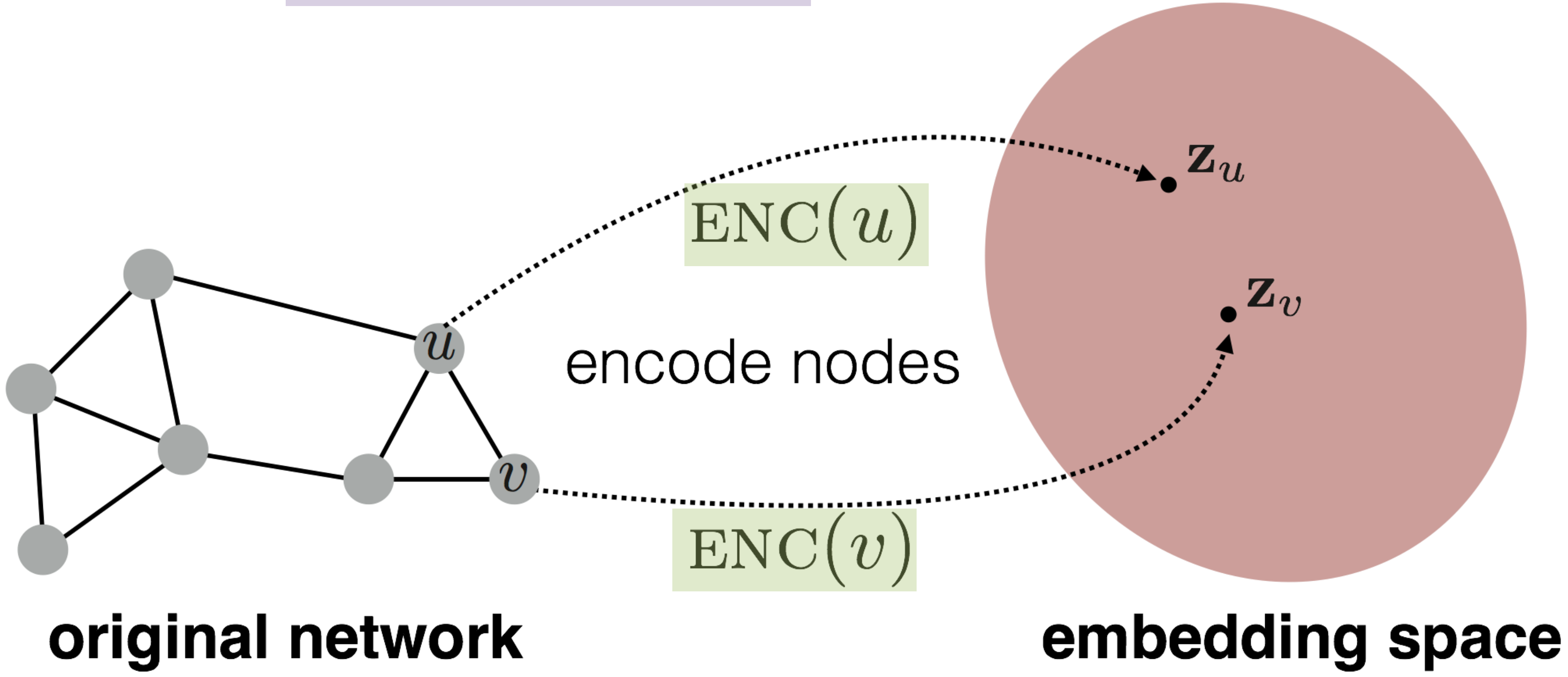
e.g., an online social network

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

Graphs – Unsupervised Task

Goal: Learn an embedding space where

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



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

Graph Neural Nets

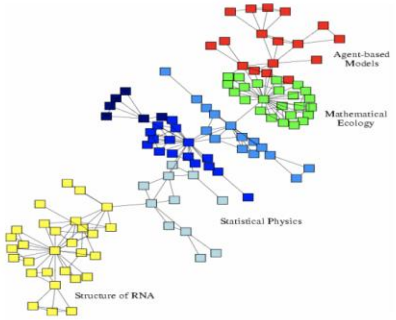
Assume we have a graph **G:**

V is the set of vertices
A is the binary adjacency matrix
X is a matrix of node features:

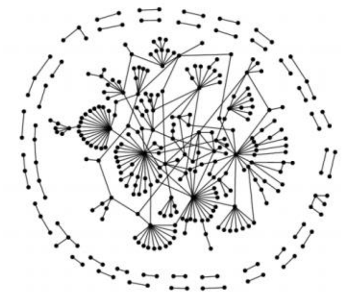
- Categorical attributes, text, image data
  e.g. profile information in a social network
- …

Y is a vector of node labels (optional)

Social networks



Economic networks

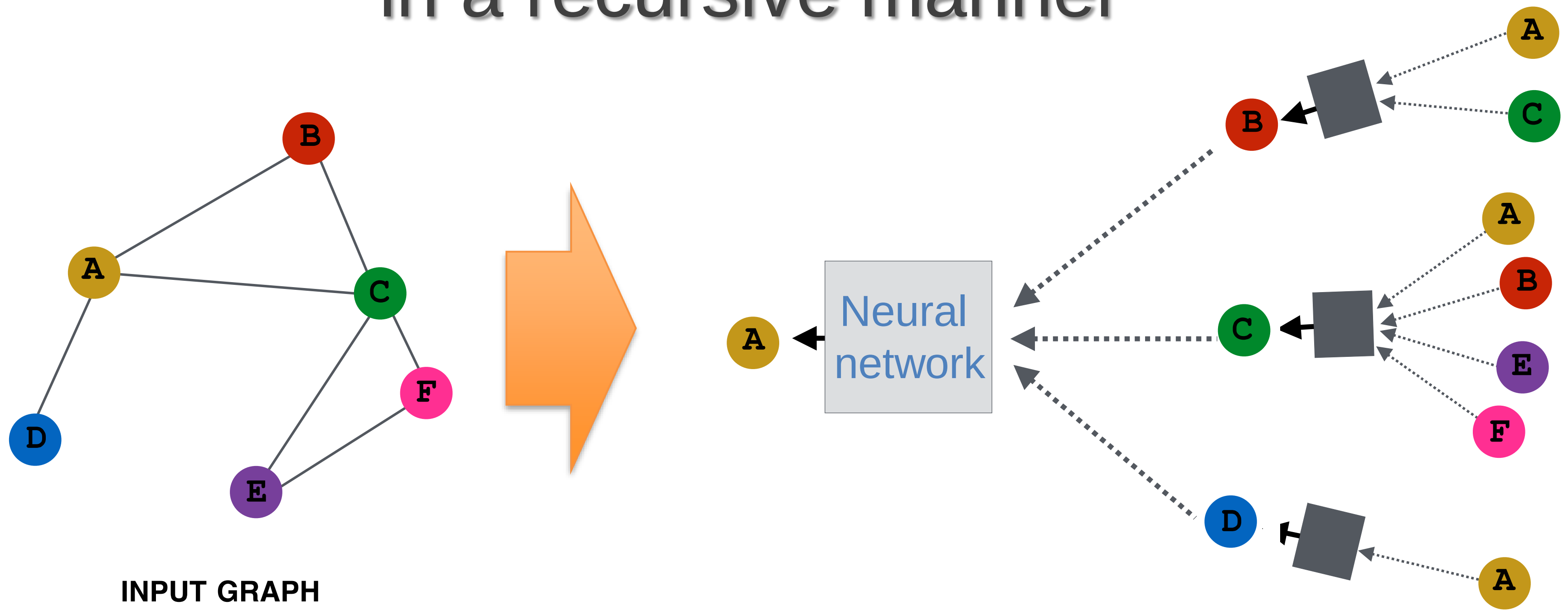


Biomedical networks

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

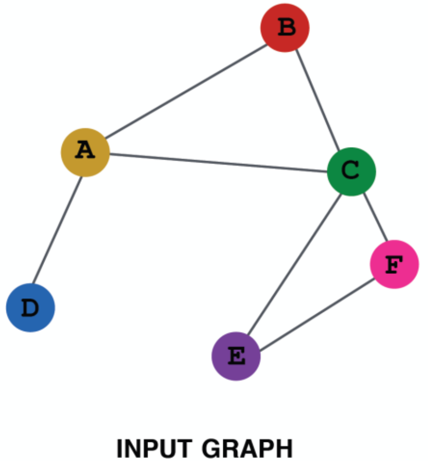
Graph Neural Nets

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

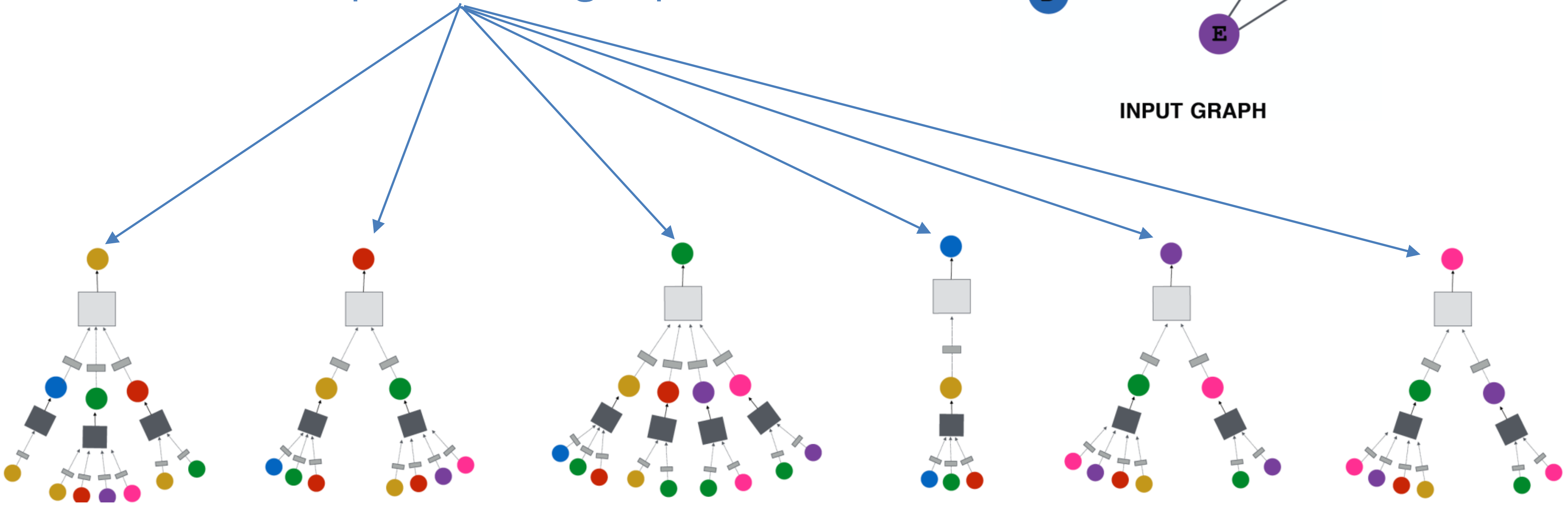


[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets



Every node defines a unique computation graph!

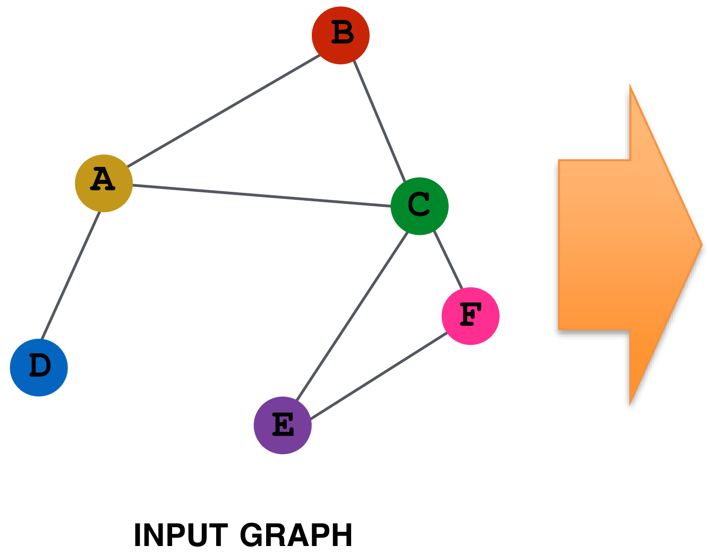


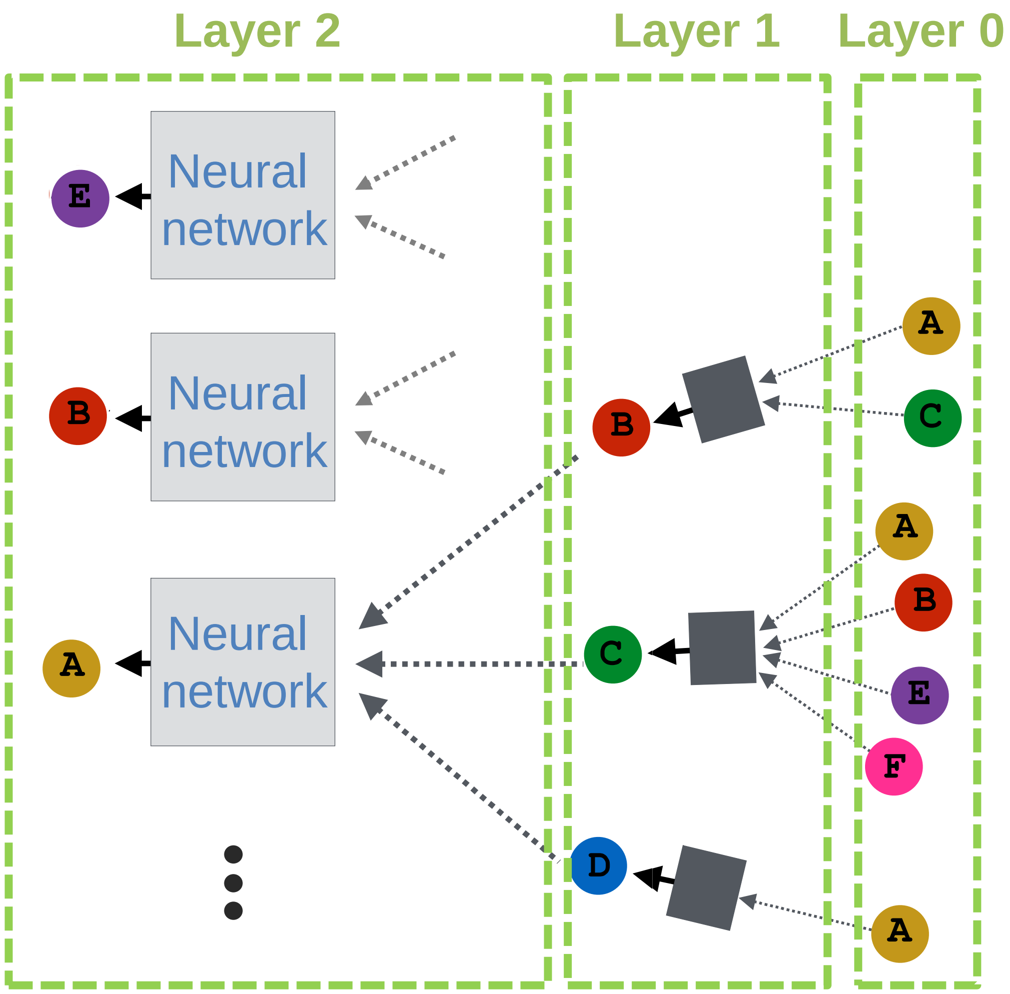
[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

And multiple layers!

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

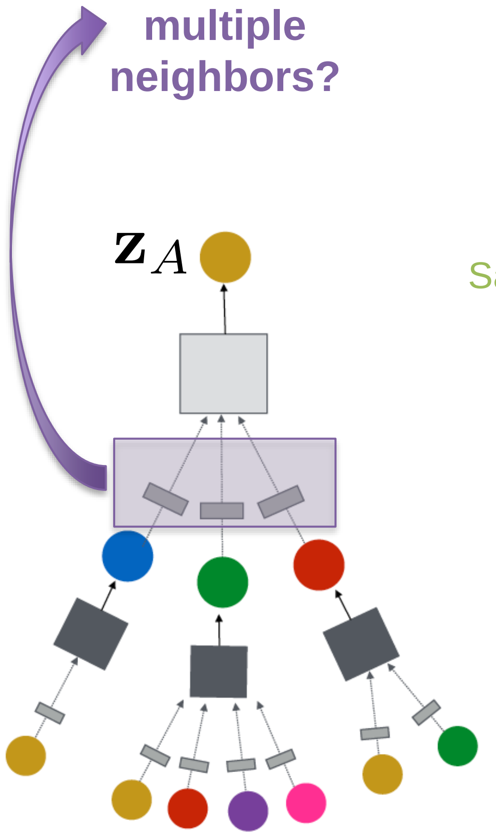




[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets – Neighborhood Aggregation

How to aggregate multiple neighbors?



Average pooling (Scarselli et al., 2005)

Different weights for neighbors and self

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

K is num layers

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)

Attention weights

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

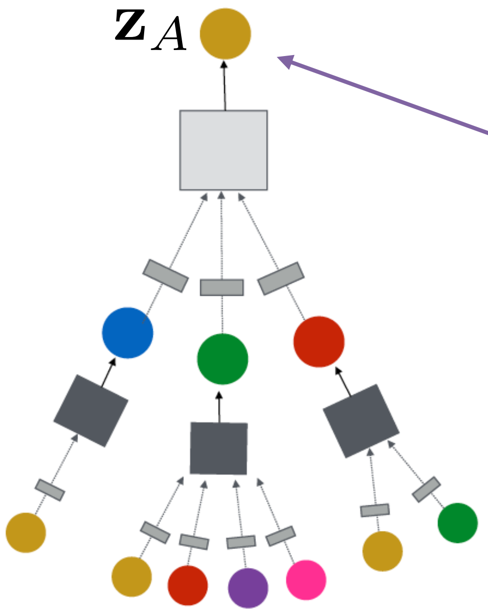
Very similar to a self-attention transformer

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

Graph Neural Nets – Supervised Training

classification weights

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



output node embedding

node class label

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

Going Beyond Sequences: Hierarchical Structure

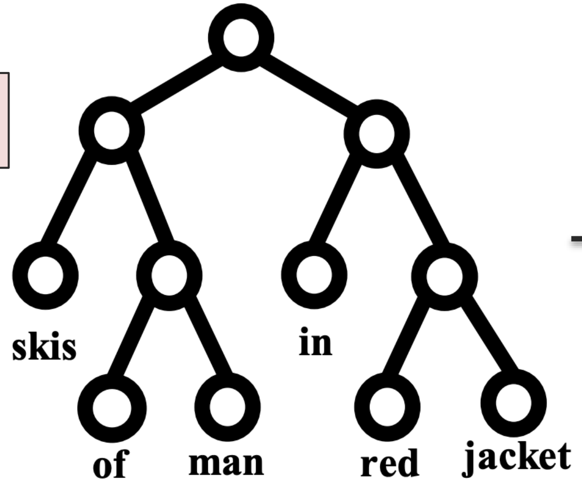
*slides adapted from Leskovec, Representation Learning on Networks. WWW 2018

Hierarchical Structure

Leverage syntactic structure of language

Parse

Skis of man in red jacket.



Object detection

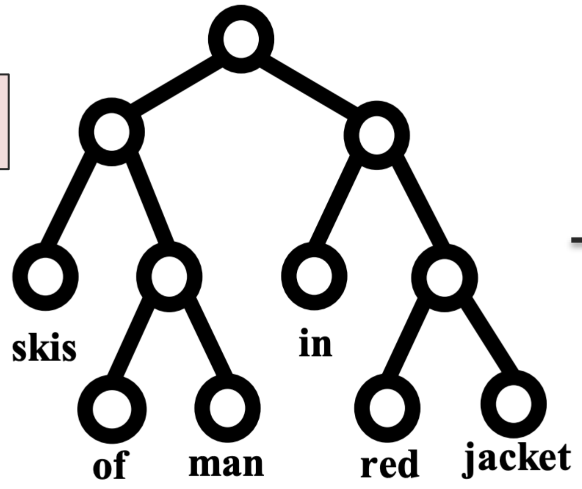
[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

Hierarchical Structure

Leverage syntactic structure of language

Parse

Skis of man in red jacket.



Object detection

Coordination

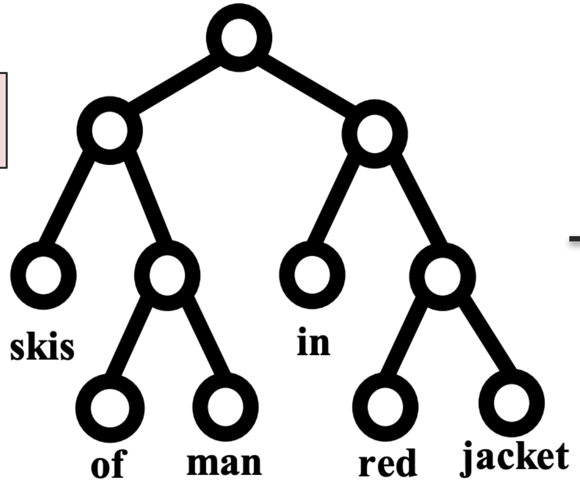
[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

Hierarchical Structure

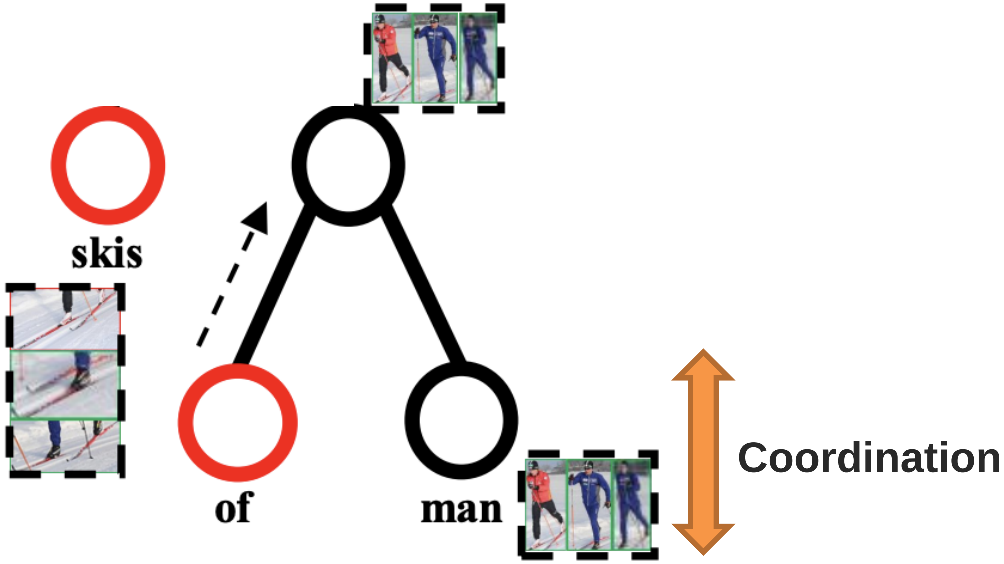
Leverage syntactic structure of language

Parse

Skis of man in red jacket.



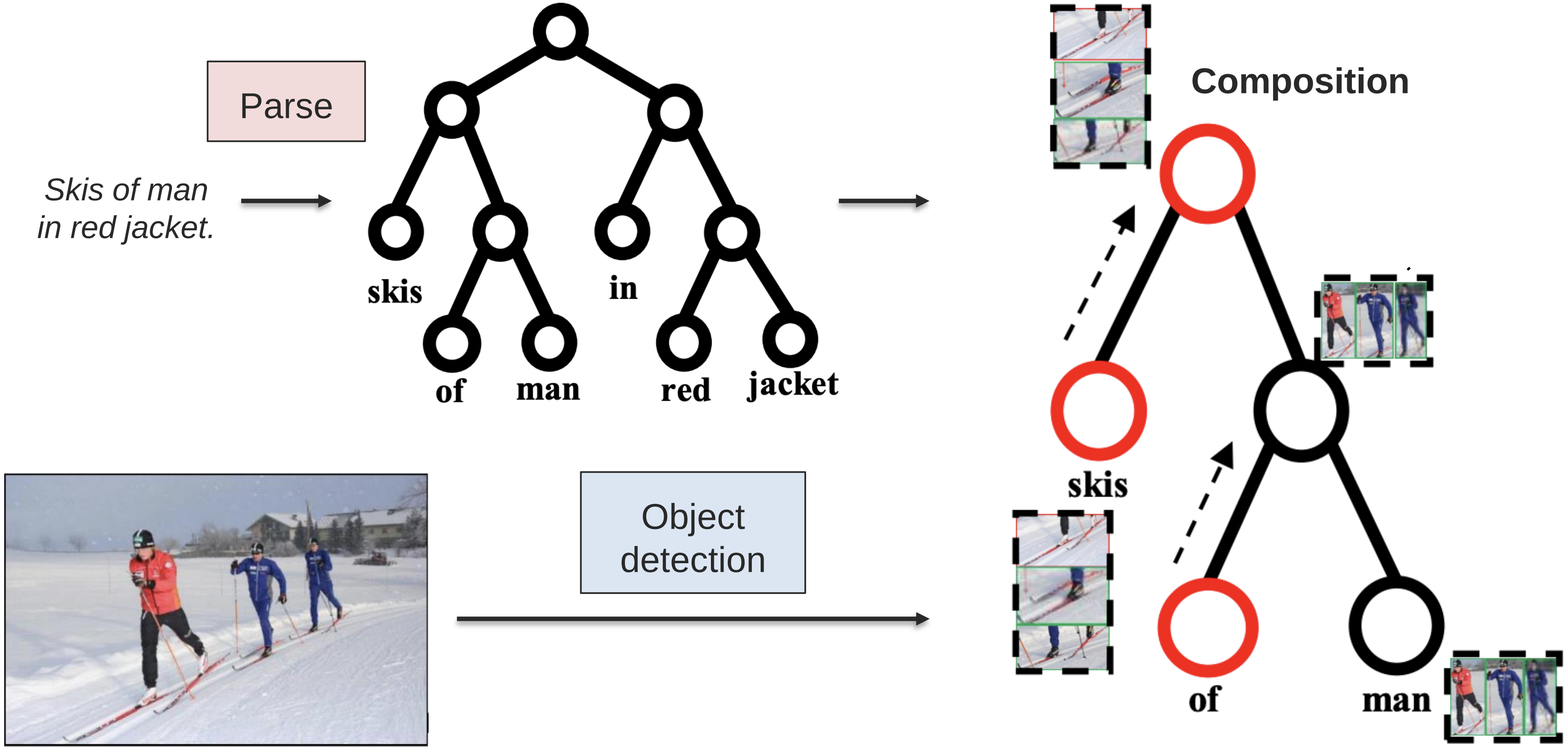
Object detection



[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

Hierarchical Structure

Leverage syntactic structure of language



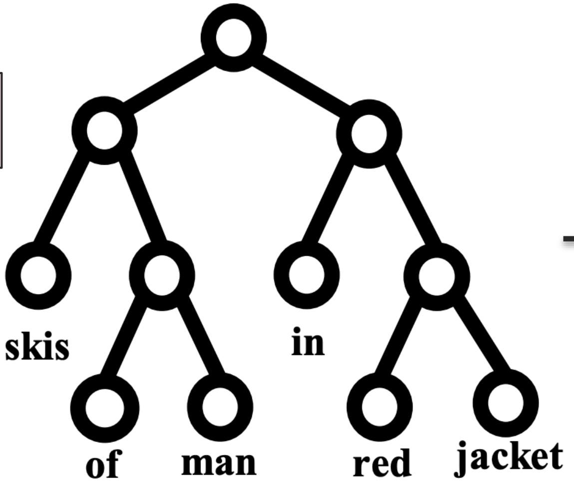
[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

Hierarchical Structure

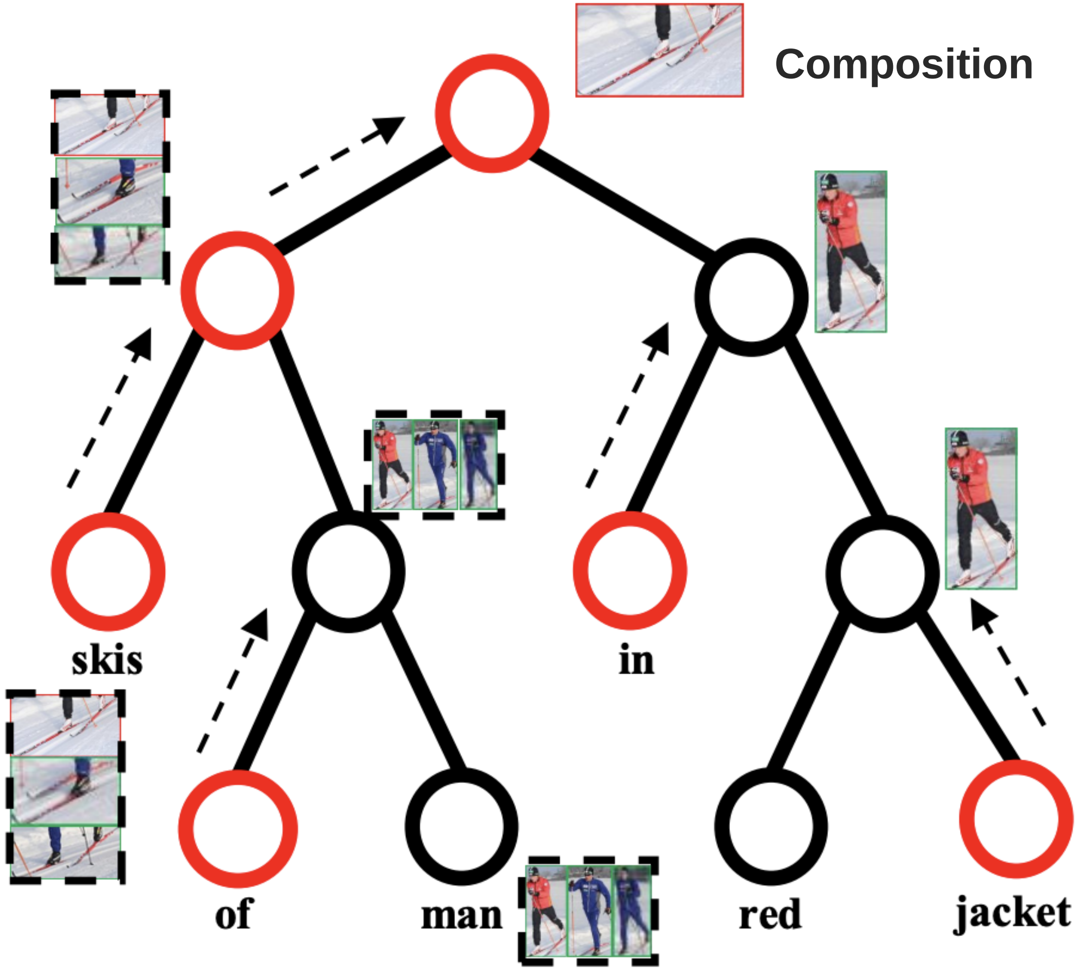
Leverage syntactic structure of language

Parse

Skis of man in red jacket.



Object detection

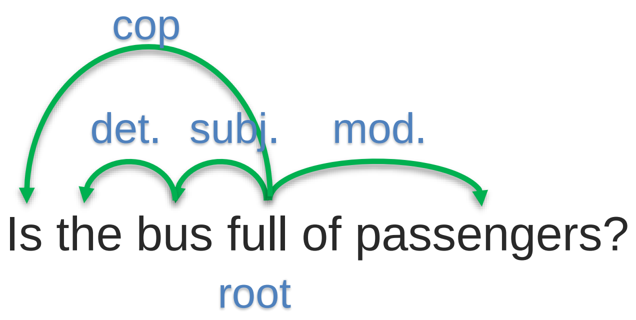


[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

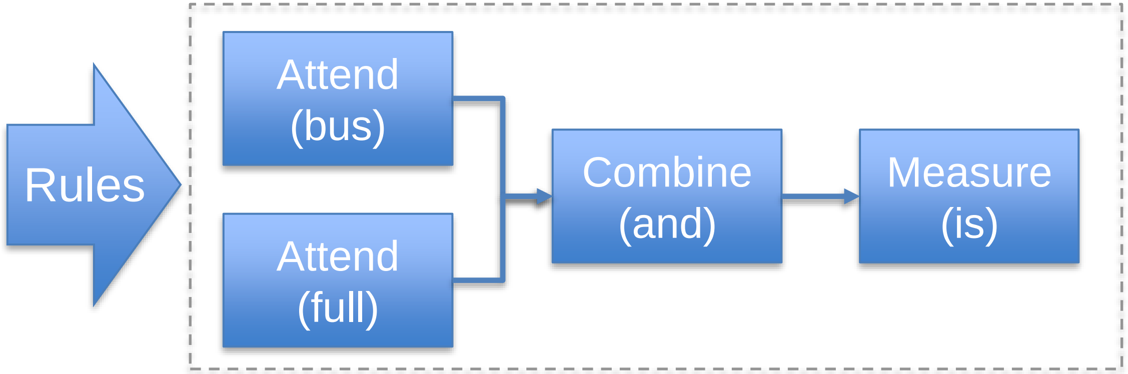
Going Beyond Sequences: Modular Structure

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

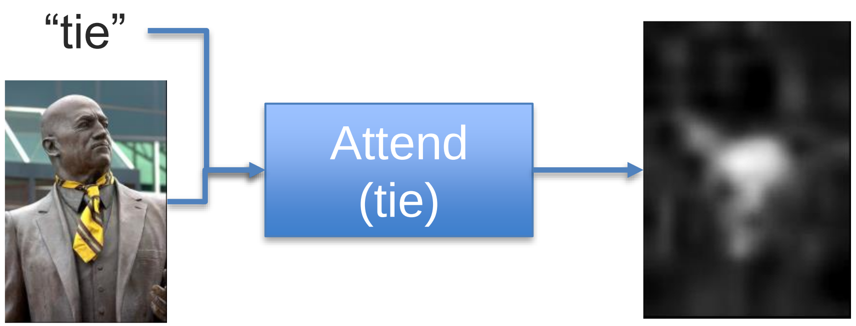
Neural Module Network



Computation layout



Each module work on the attention map(s):

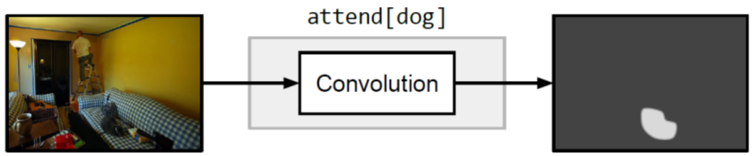


Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

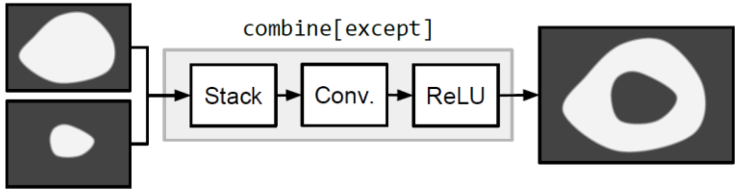
Predefined Set of Modules

1) Analyze the image:

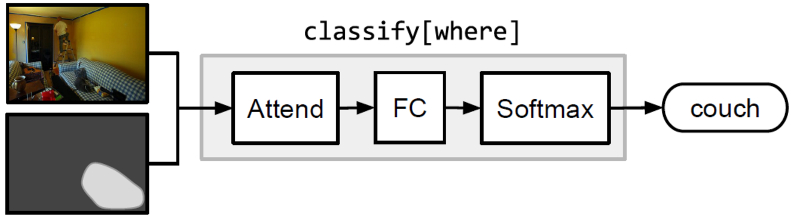
$\texttt{attend} : \mathit{Image} \rightarrow \mathit{Attention}$



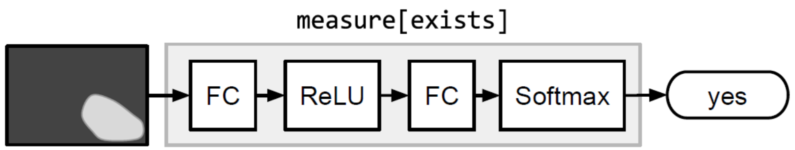
$\texttt{combine} : \mathit{Attention} \times \mathit{Attention} \rightarrow \mathit{Attention}$



2) Make a prediction



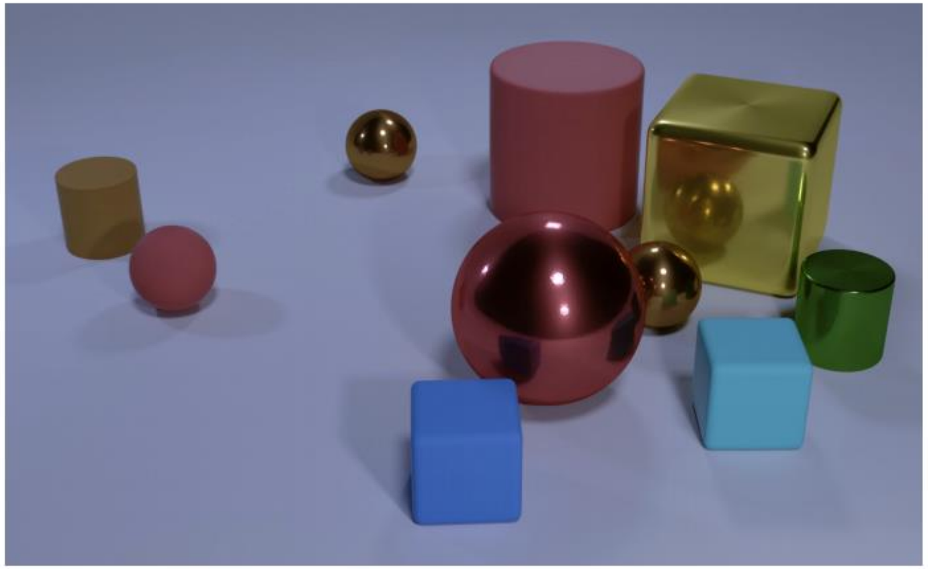
$\texttt{measure} : \mathit{Attention} \rightarrow \mathit{Label}$



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

CLEVR: Dataset for Visual Reasoning

Perfect for a neural module network!



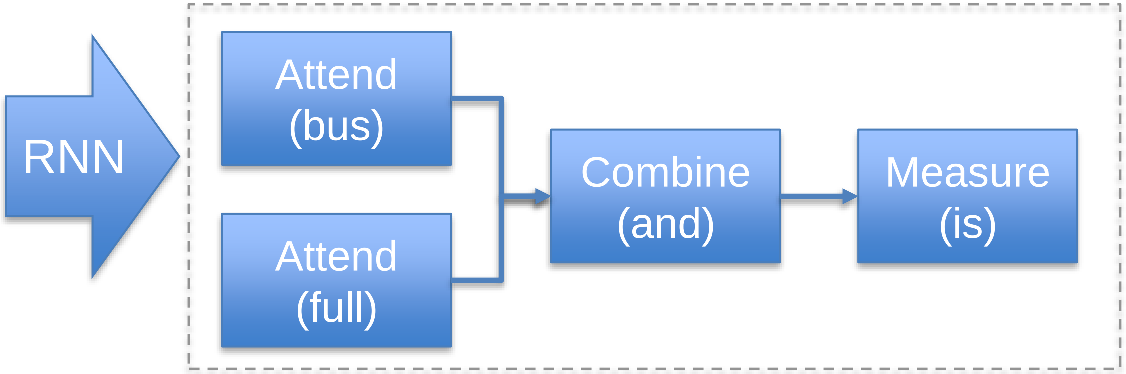
Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many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

Computation layout

Is the bus full of passengers?



No need to parse the question!

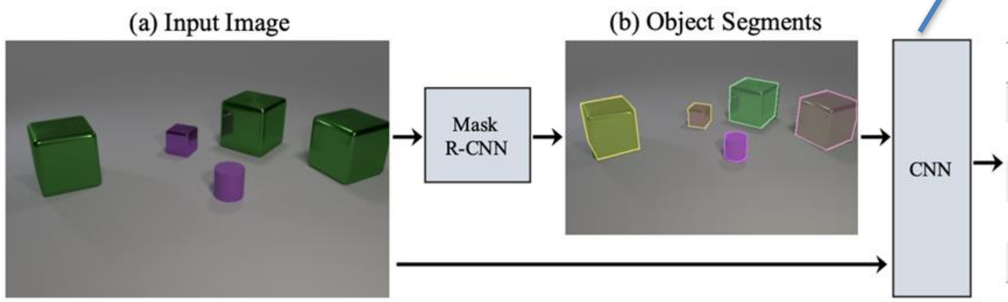
No rule-based creation of the layout!

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

Module Network 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

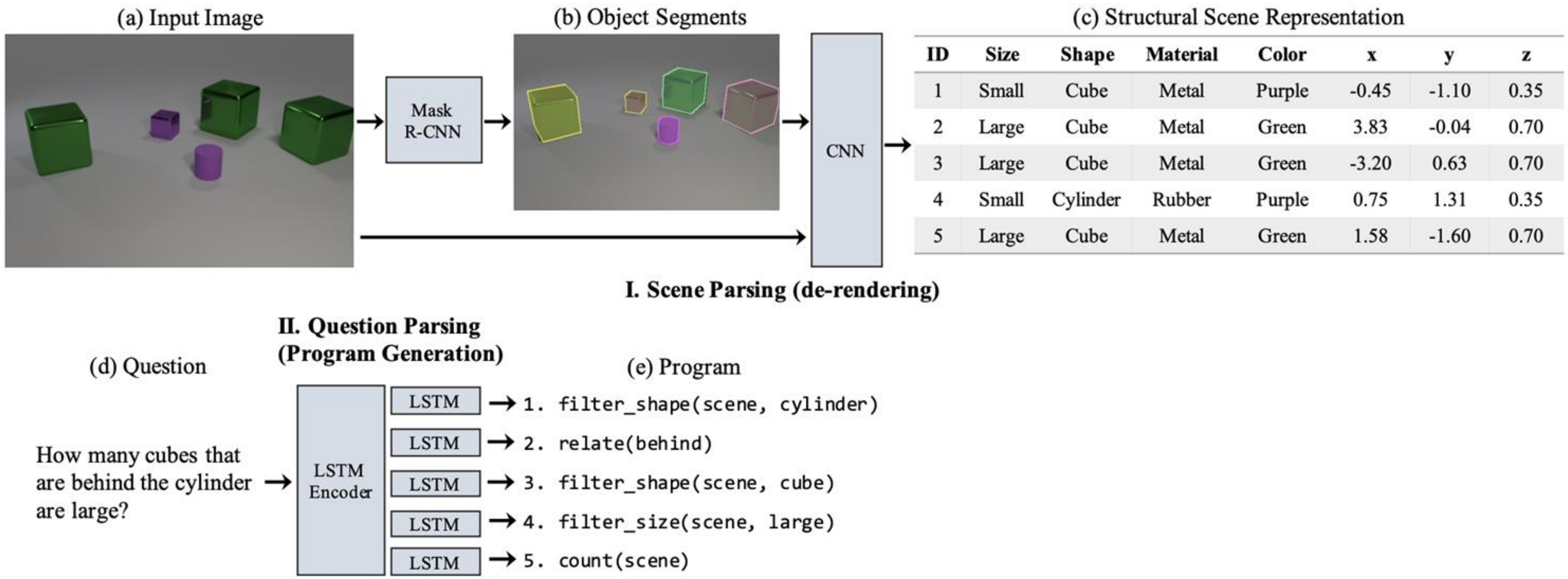
I. Scene Parsing (de-rendering)

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



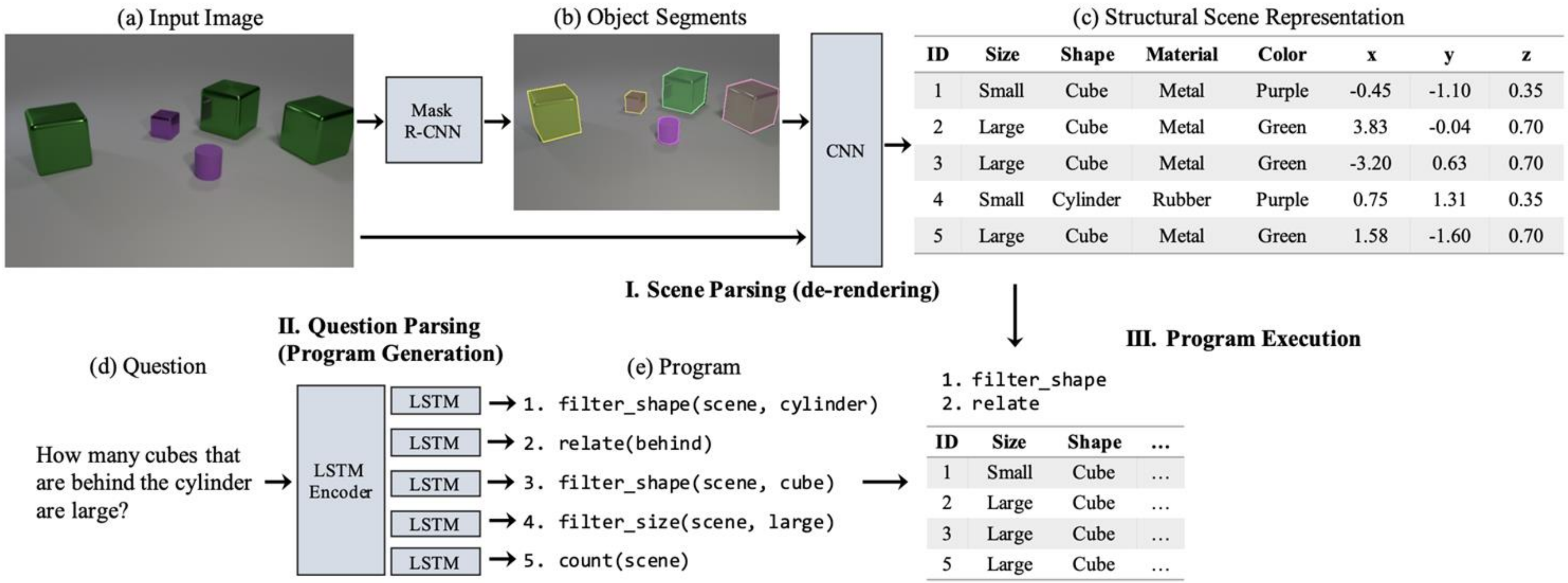
ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

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 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.

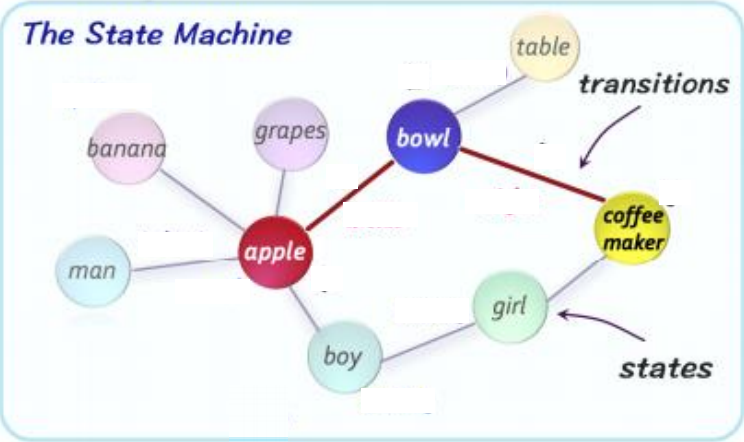
2. Treat the graph as a **state machine** and simulate iterative computation over it to *answer questions* or *draw inferences*.

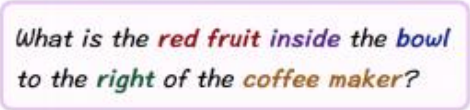
3. Natural language questions are translated into *soft instructions* and used to perform sequential reasoning over the scene graph/state machine.

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

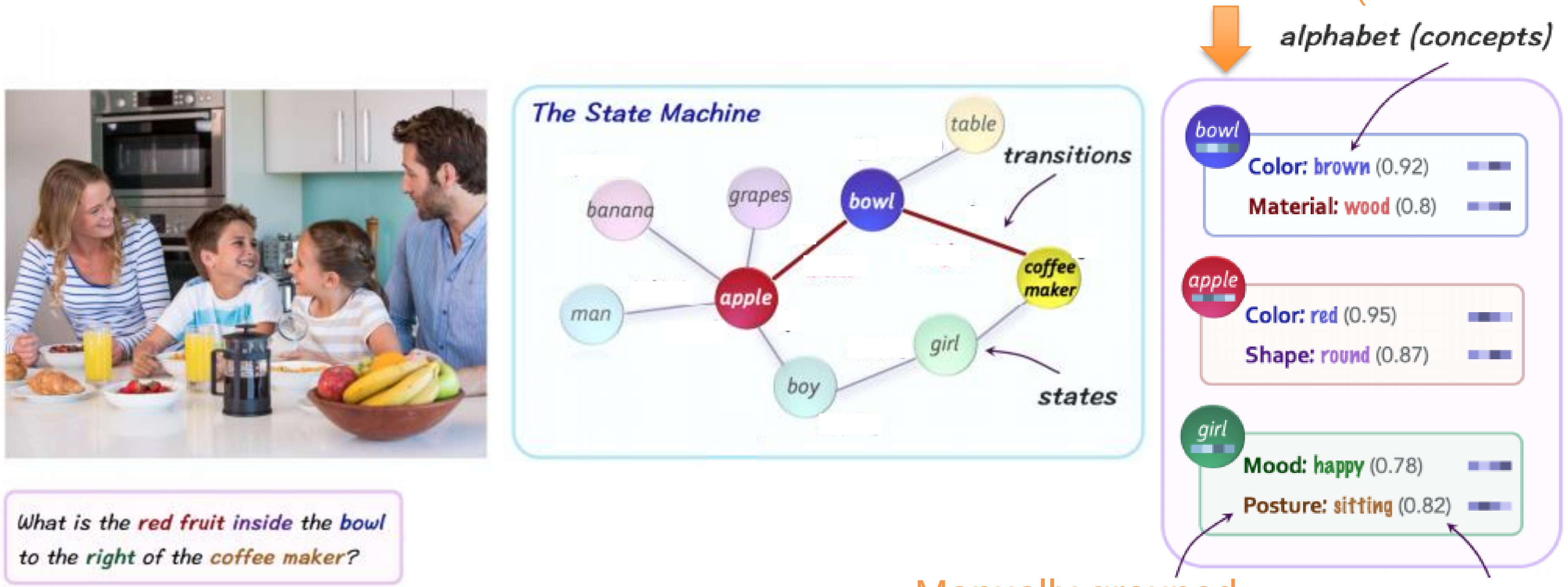




Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine.“ NeurIPS 2019

Module Networks V4: The Neural State Machine

Pre-trained an alphabet of concepts
(Visual Genome)



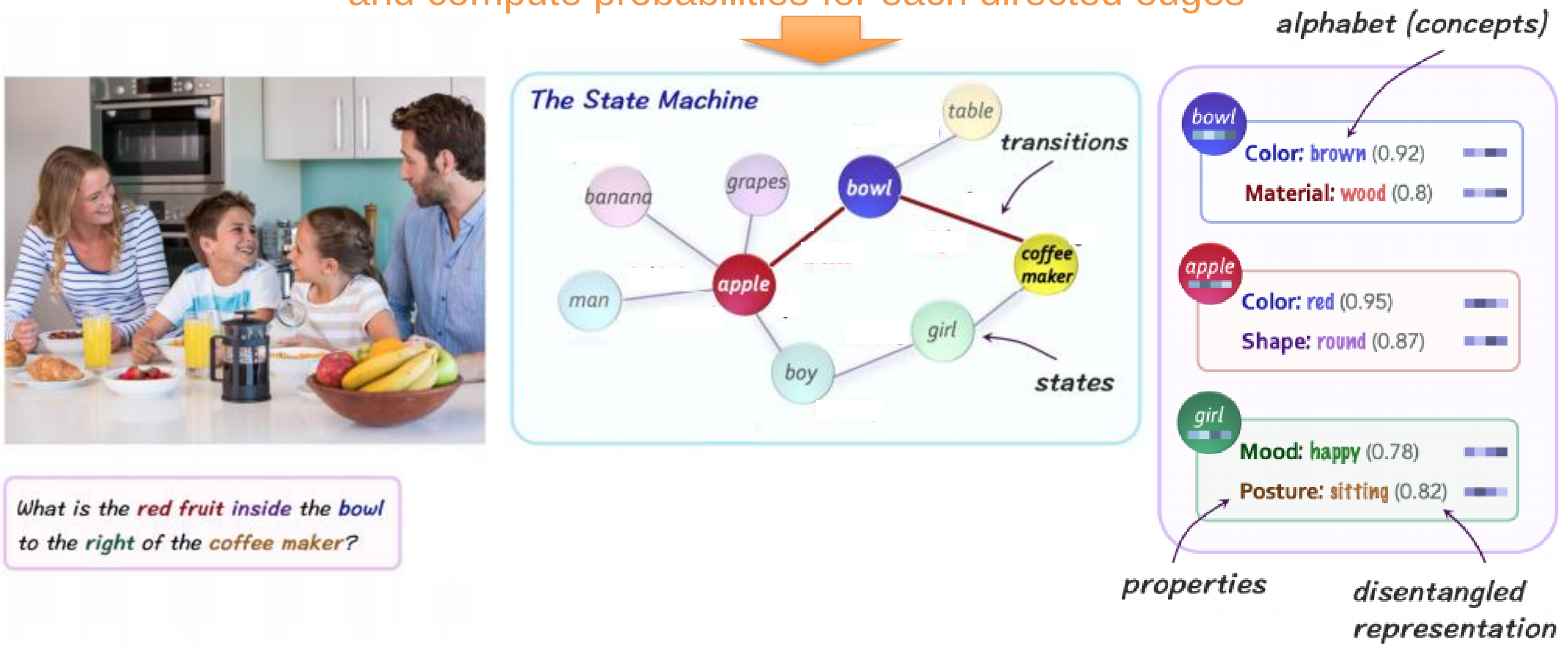
Manually grouped by “properties”

Probabilities computed at runtime for each object instance

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

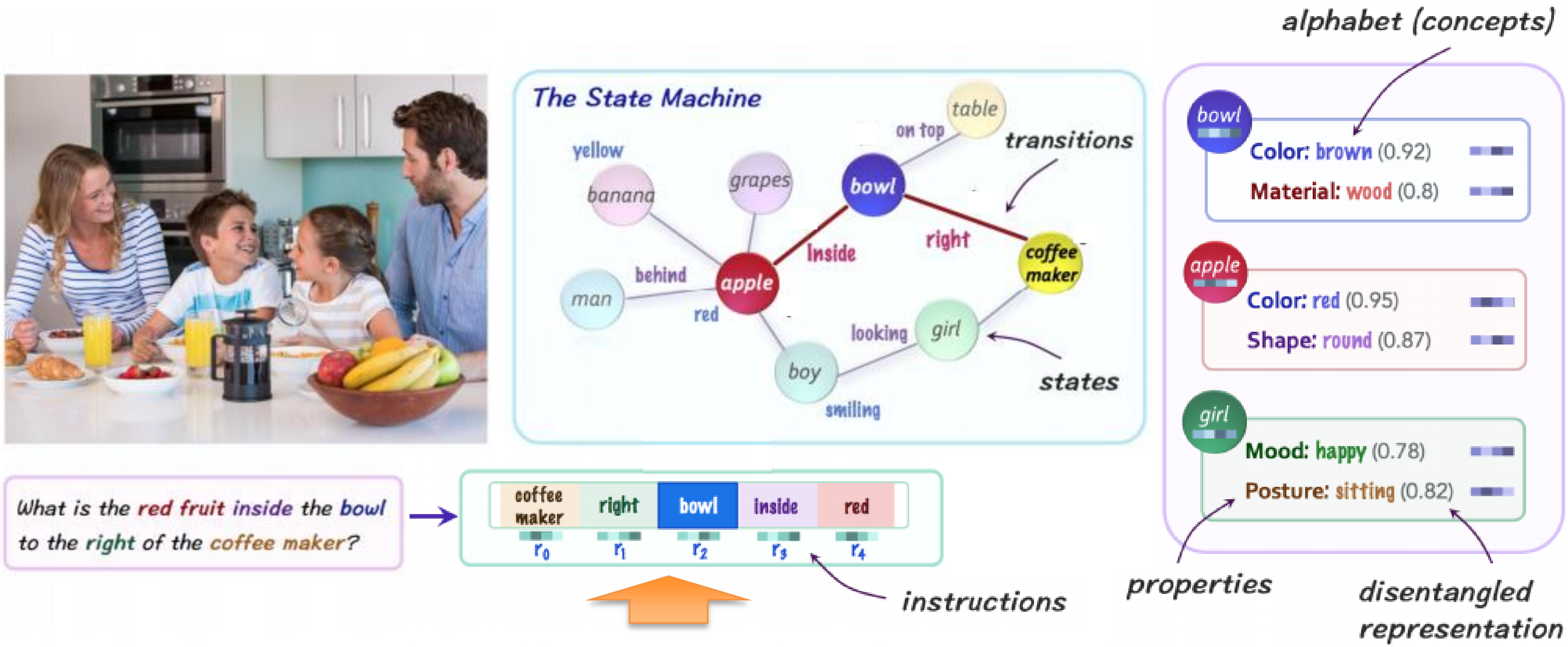
Module Networks V4: The Neural State Machine

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

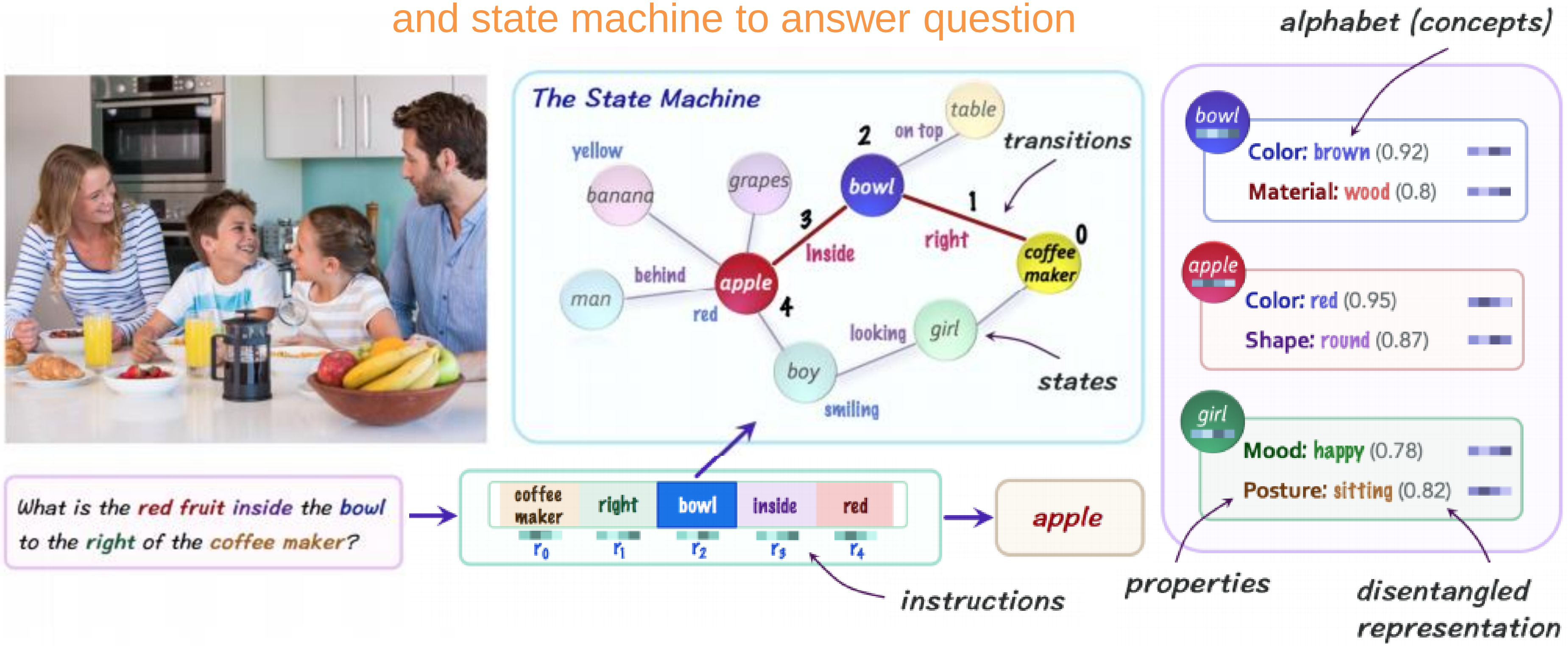


Translate each word in a concept-based representation
and group in a fixed number of instruction steps

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