SFU NatLangLab

CMPT 825: Natural Language Processing

Grounded Natural Language

Spring 2020
2020-04-09

Adapted from slides from Danqi Chen and Karthik Narasimhan

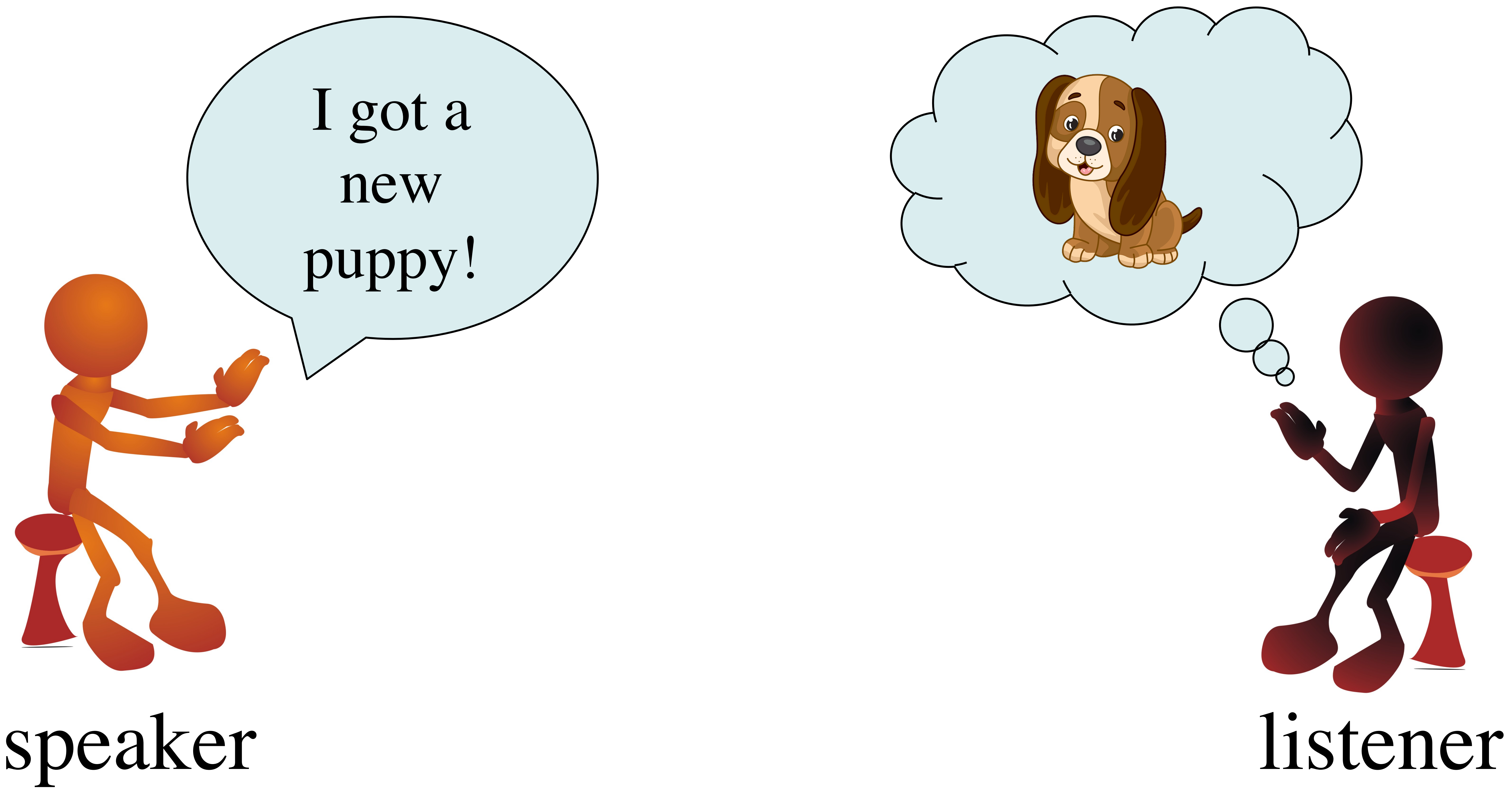
Topics in NLP research

- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Generation and Summarization
- Information Extraction and Question Answering
- Information Retrieval
- Language Resources and Evaluation
- Language and Vision
- Linguistic and Psycholinguistic Aspects of CL
- Machine Learning for NLP
- Machine Translation
- NLP for Web, Social Media and Social Sciences
- NLP-enabled Technology
- Phonology, Morphology and Word Segmentation
- Semantics
- Sentiment Analysis and Opinion Mining
- Spoken Language Processing
- Tagging, Chunking, Syntax and Parsing
- Text Categorization and Topic Models

Grounding

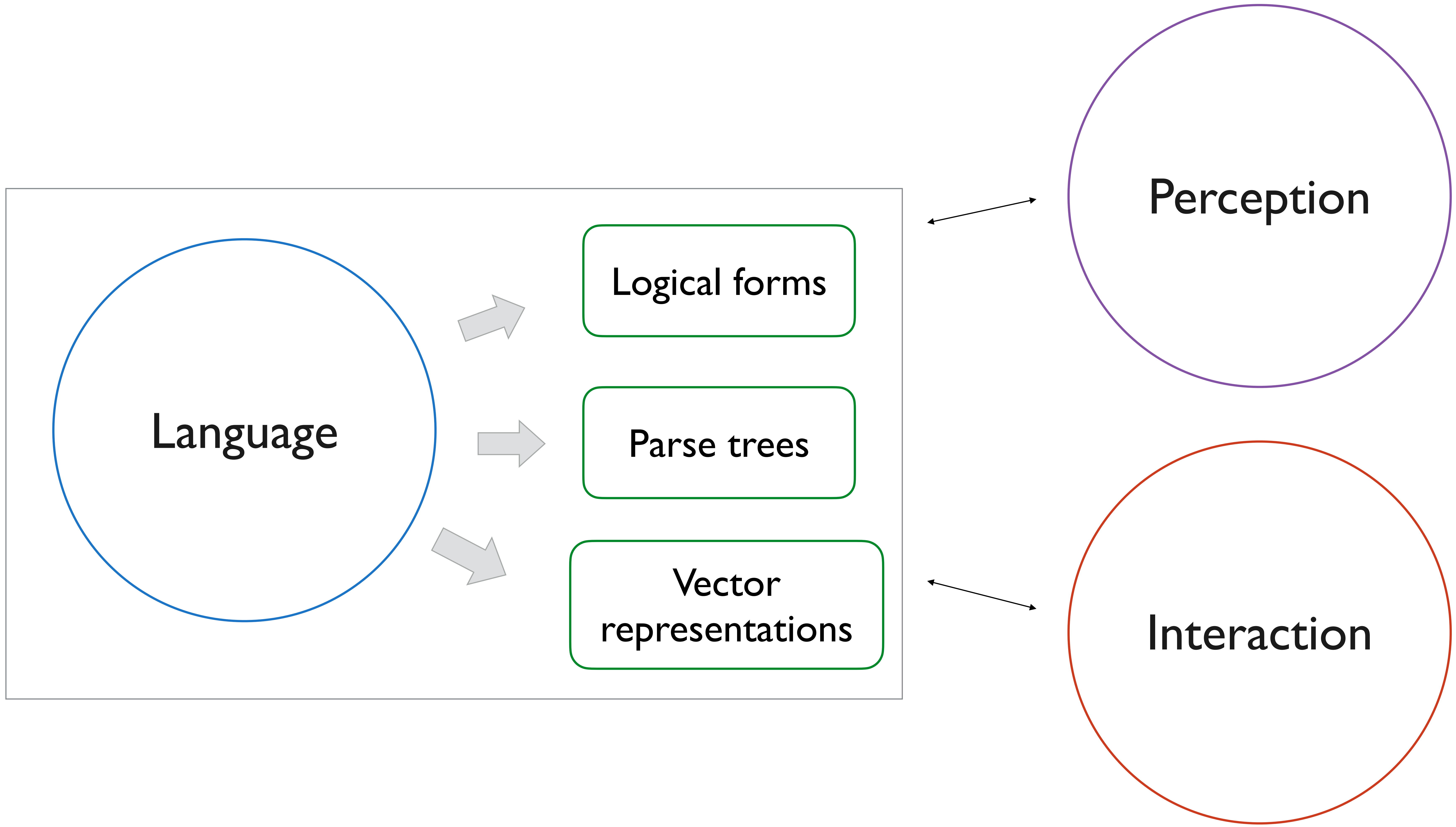
Language is used to communicate about the world

- Things, actions, abstract concepts



Connecting linguistic symbols to the physical world

Semantics does not exist in isolation



Children do not learn language from raw text or passively watching TV

Natural way to learn language in the context of its use in the physical and social world

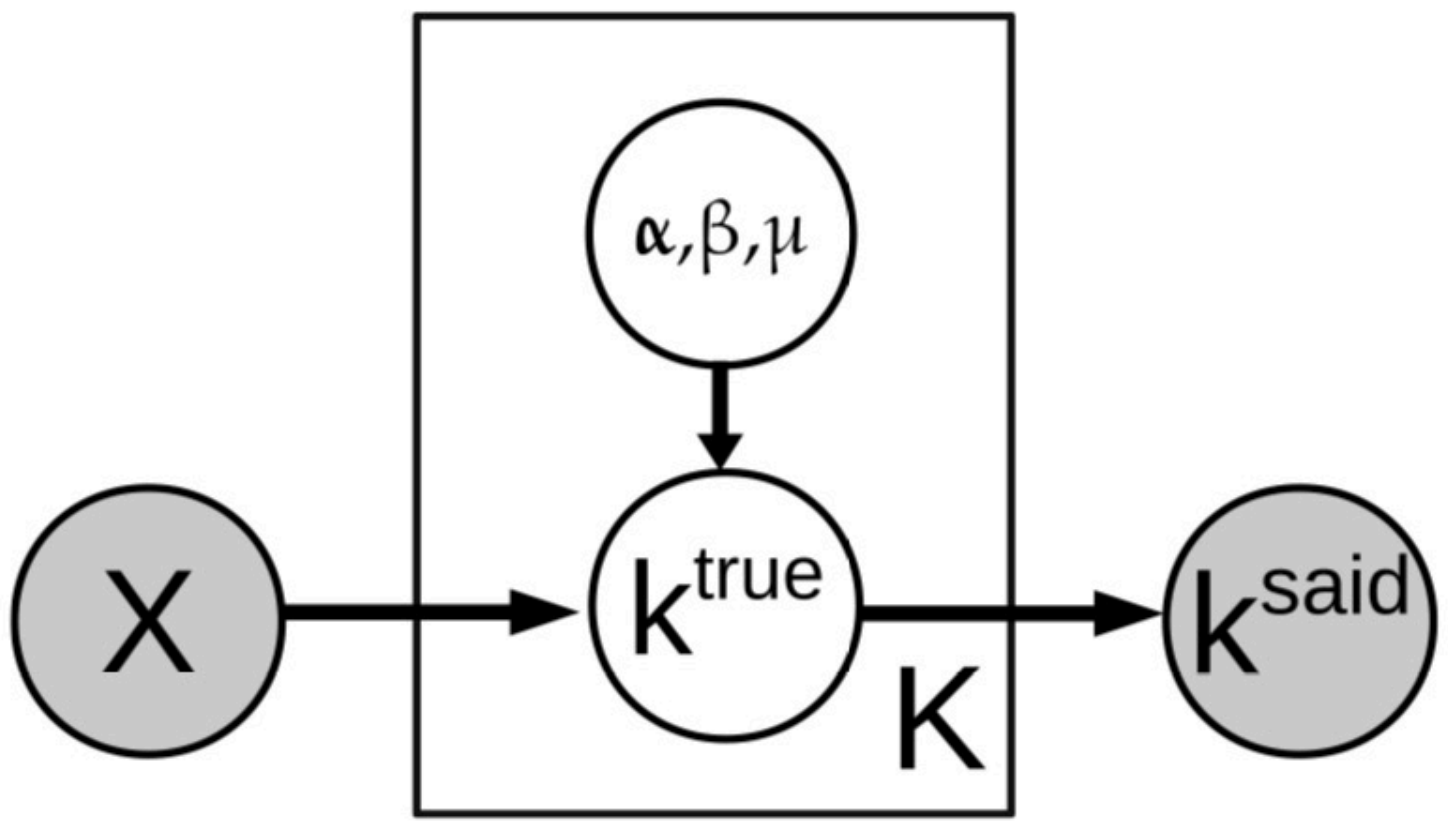
This requires inferring the meaning of utterances from their perceptual context

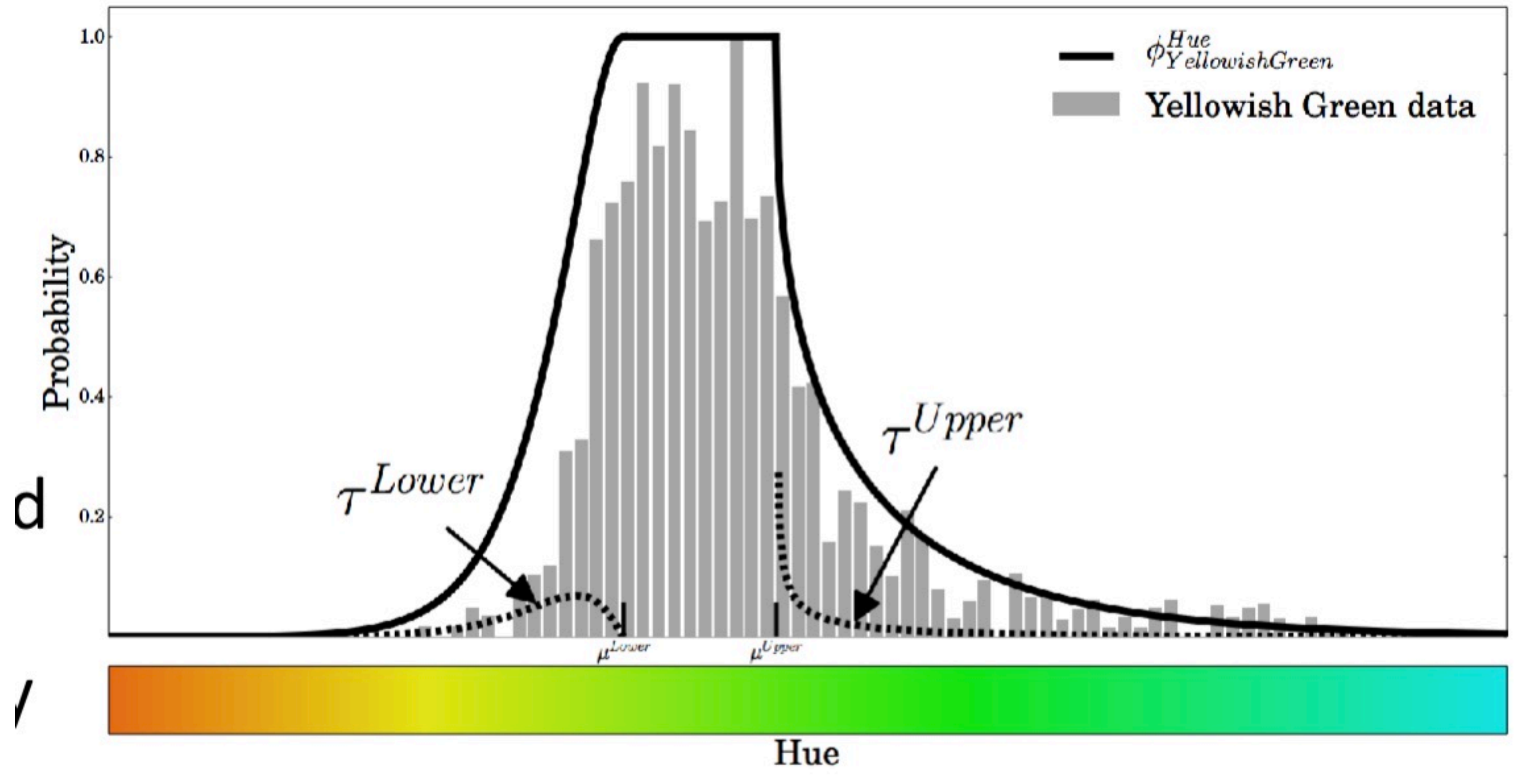
Color test

- What color is this?

Grounding color

- Bayesian model for grounded color semantics
- 829 color descriptions

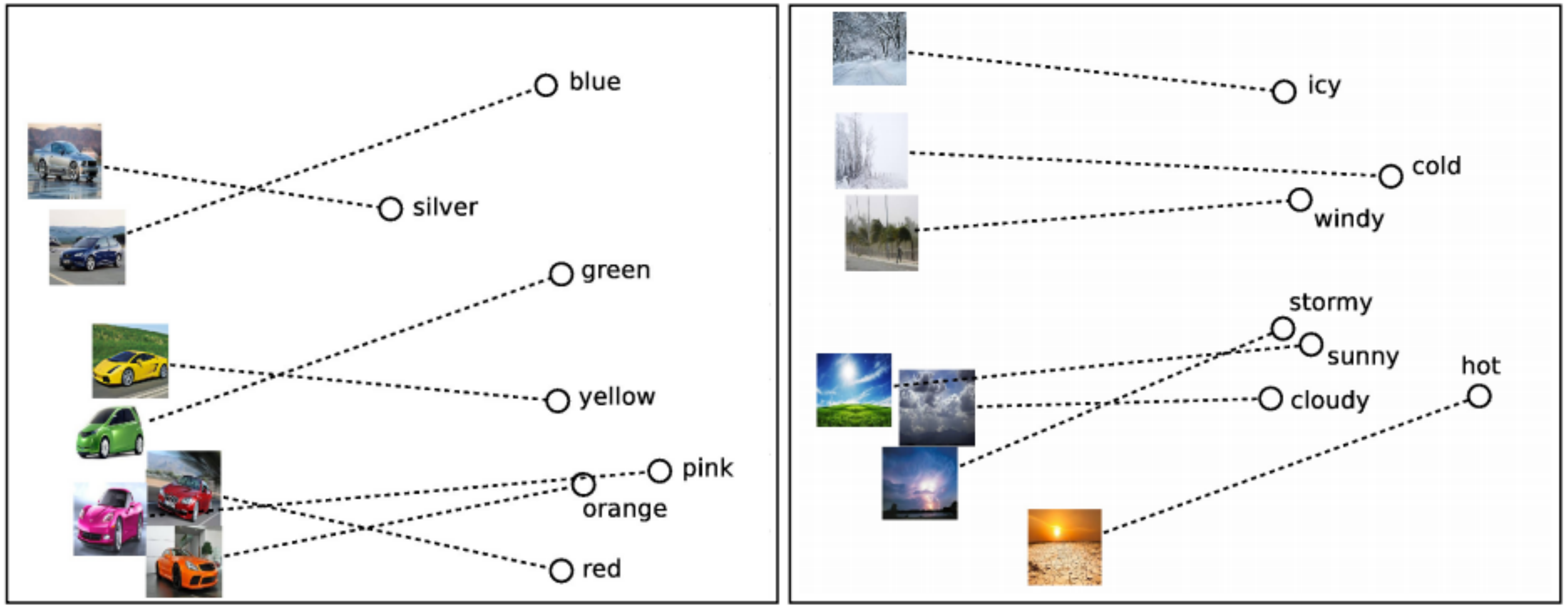




(McMahan and Stone, 2014)

Cross-modal Embeddings

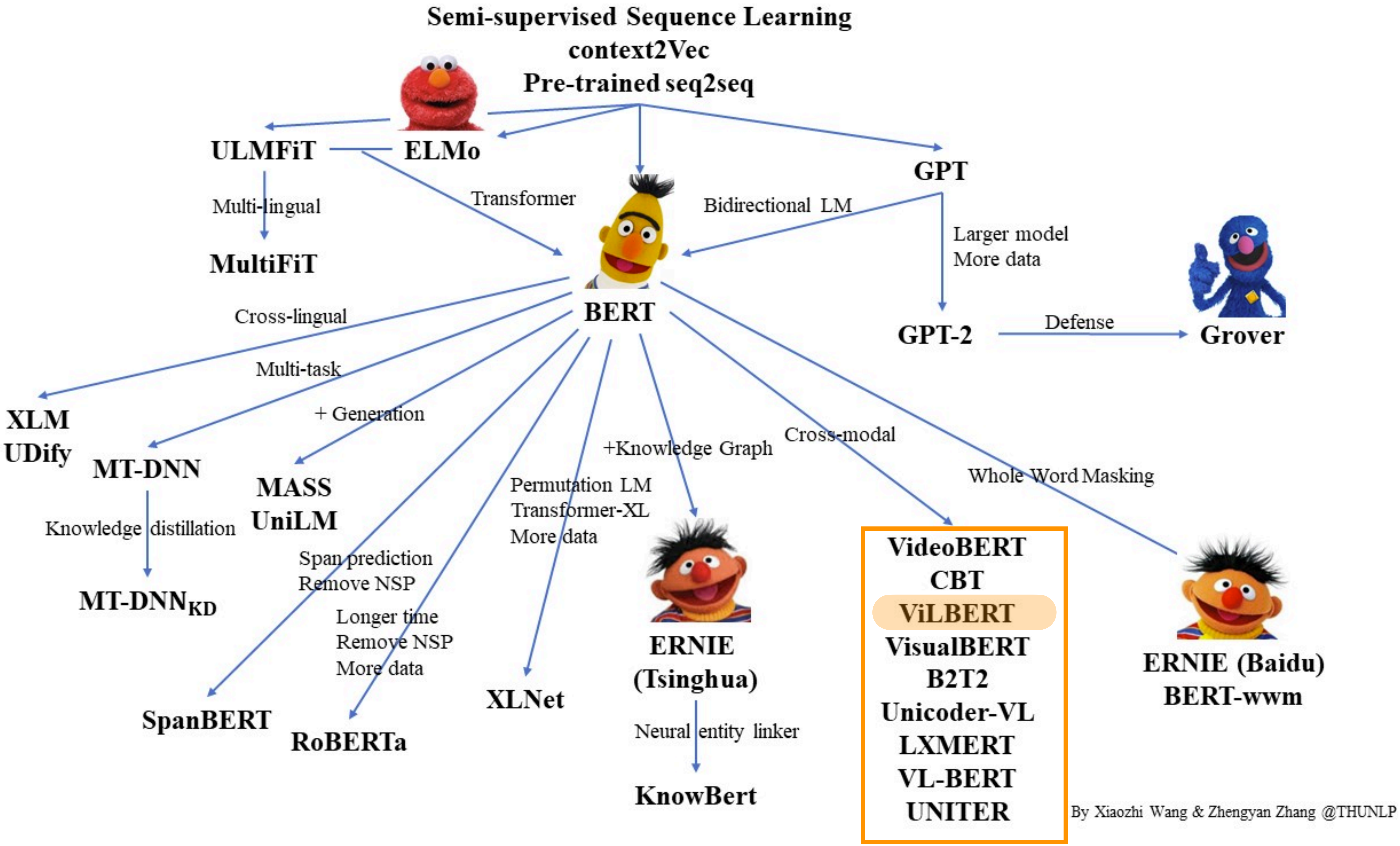
Common representation for language and vision: vectors!



(a) Colors (b) Weather

Figure 5: PCA projection of the 300-dimensional word and image representations for (a) cars and colors and (b) weather and temperature.

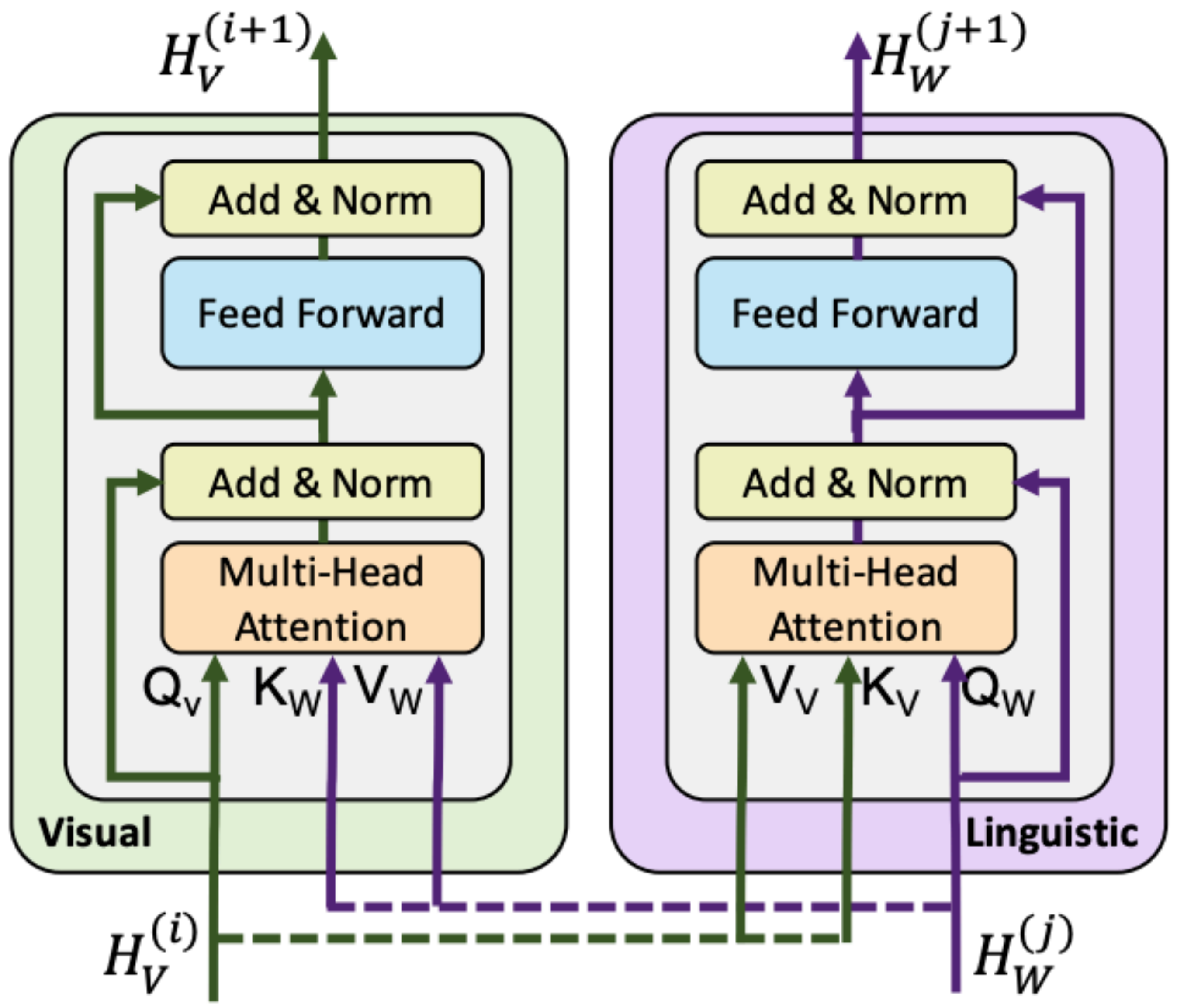
Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
[Kiros, Salakhutdinov, Zemel TACL 2015, https://arxiv.org/pdf/1411.2539.pdf]

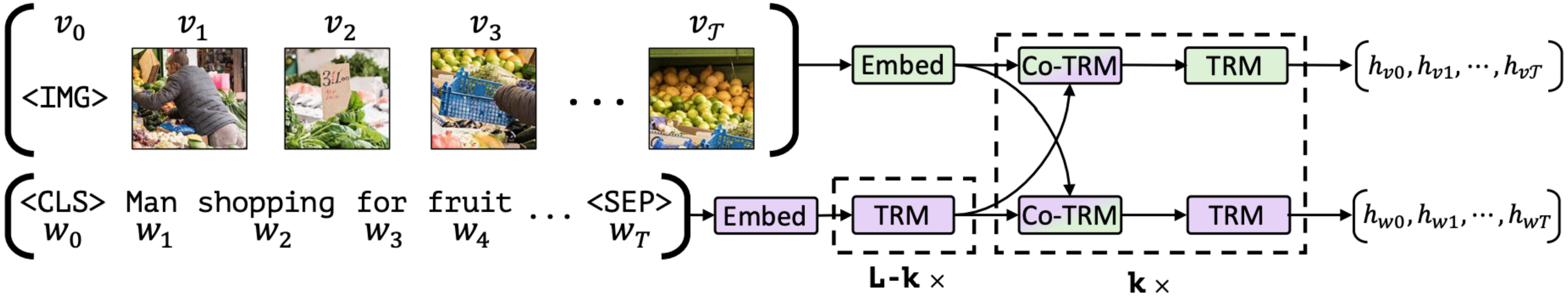

Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
Multi-lingual
MultiFiT
Transformer
BERT
GPT
Bidirectional LM
Larger model
More data
GPT-2
Defense
Grover
Cross-lingual
Multi-task
XLM
UDify
MT-DNN
+ Generation
MASS
UniLM
Knowledge distillation
MT-DNNKD
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

Pretrained representations for vision and language

Image represented as

- series of **image region features** (extracted from pre-trained object detection network)
- **Region position** encoded as $5d$ vector





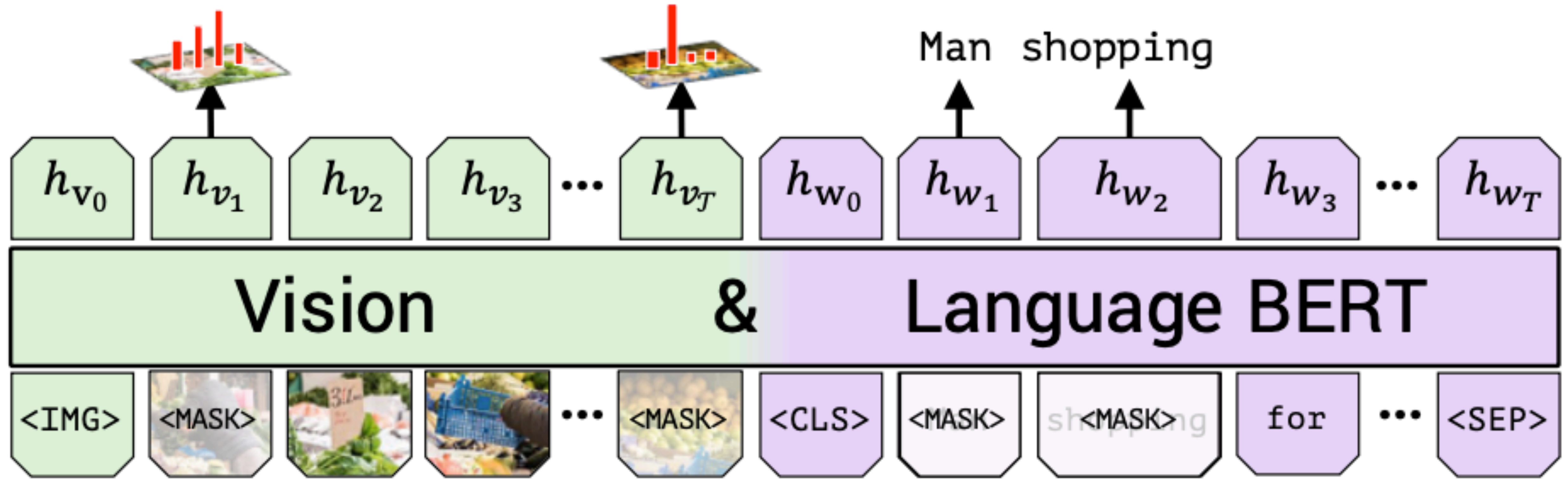
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks [Lu et al 2019, https://arxiv.org/pdf/1908.02265.pdf]

Pretrained representations for vision and language

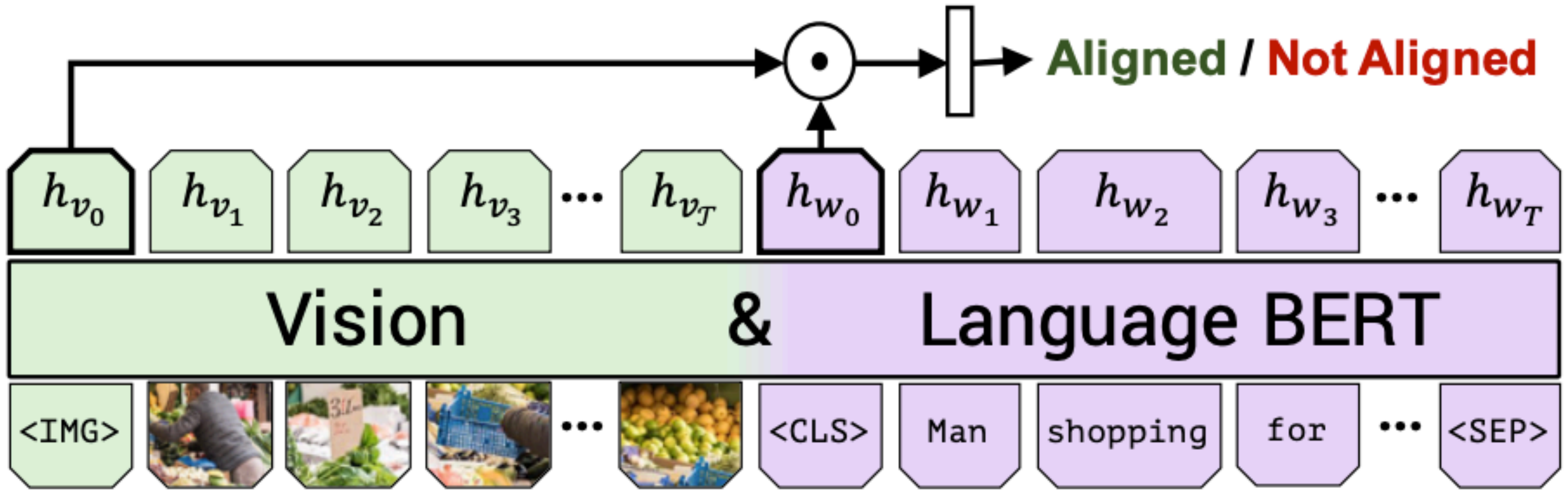
Predict semantic class distribution

Trained on

- Conceptual captions (~3.3M images with captions cleaned from alt-text labels)
- Two tasks to predict:
  - masked out words and semantic class distribution for masked out image regions
  - Is the image/description aligned?



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

ViLBERT

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
[Lu et al 2019, https://arxiv.org/pdf/1908.02265.pdf]

Pretrained representations for vision and language

		VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
	Method	test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
	SCAN [35]	-	-	-	-	-	-	-	48.60	77.70	85.20	-	-	-
Ours	Single-Stream†	65.90	68.15	68.89	47.27	65.64	72.02	56.04	-	-	-	-	-	-
	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
	ViLBERT†	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
	ViLBERT	**70.55 (70.92)**	**72.42 (73.3)**	**74.47 (74.6)**	**54.04 (54.8)**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Pretraining improves performance on variety of vision+language tasks!

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
[Lu et al 2019, https://arxiv.org/pdf/1908.02265.pdf]

Types of grounding

- **Perception**
  - Visual: *green* = [0,1,0] in RGB
  - Auditory: *loud* =   >120 dB
  - Taste: sweet = >some threshold level of sensation on taste buds
  - High-level concepts:

cat

dog

Types of grounding

- **Temporal concepts**
  - *late evening* = after 6pm
  - *fast, slow* = describing rates of change
- **Actions**

running

eating

Types of grounding

- **Relations**
  - **Spatial:**
    - *left, on top of, in front of*
  - **Functional:**
    - *Jacket:* keeps people warm
    - *Mug*: holds water
  - **Size:**
    - Whales are *larger* than lions

A chair

A chair

green

light

armless

fragile

medium size

used to sit on

plush

Context is very important!

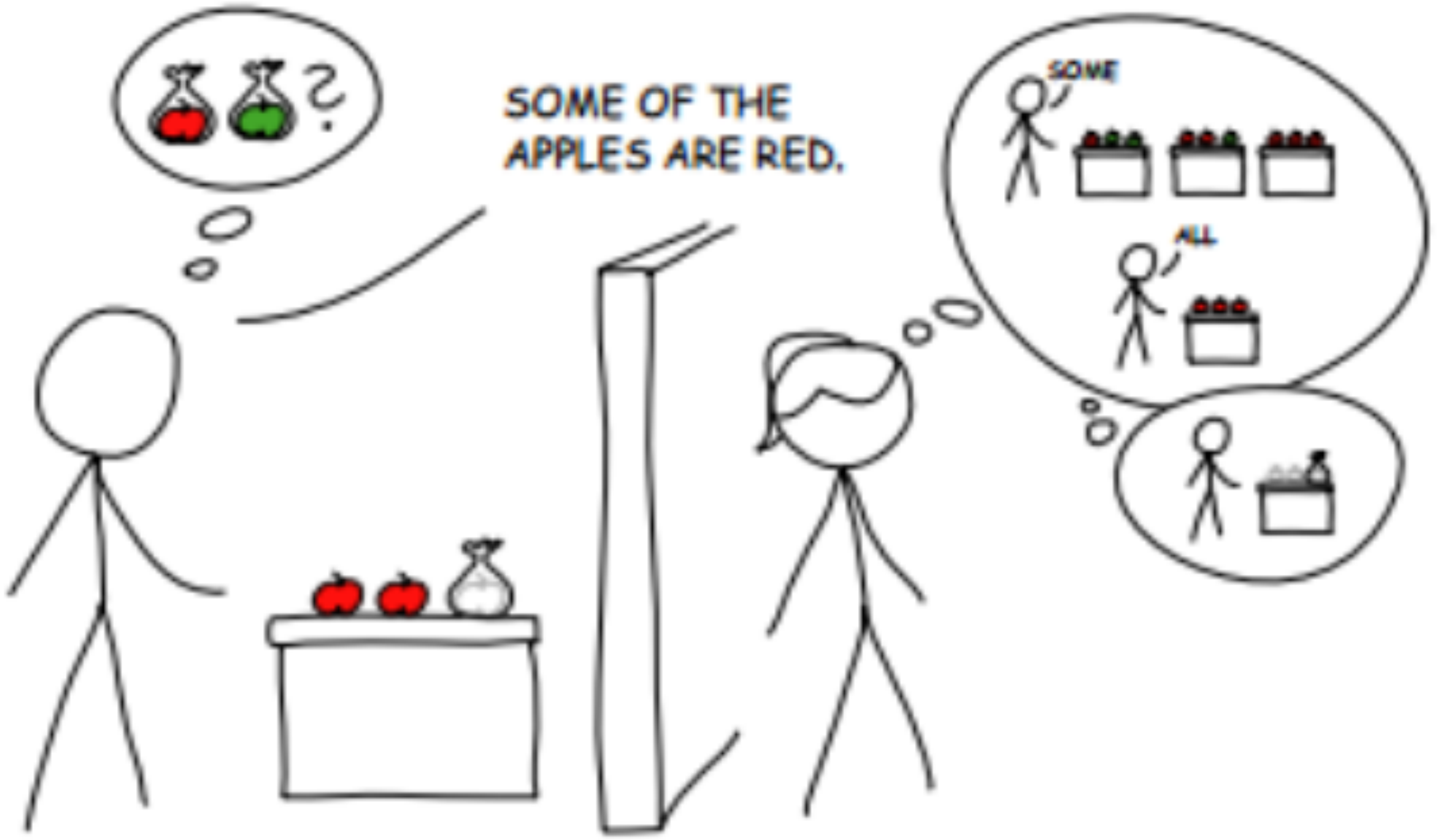
Language game

Wittgenstein. 1953. Philosophical Investigations:

Language derives its meaning from use.

'block' 'pillar' 'slab' 'beam'.

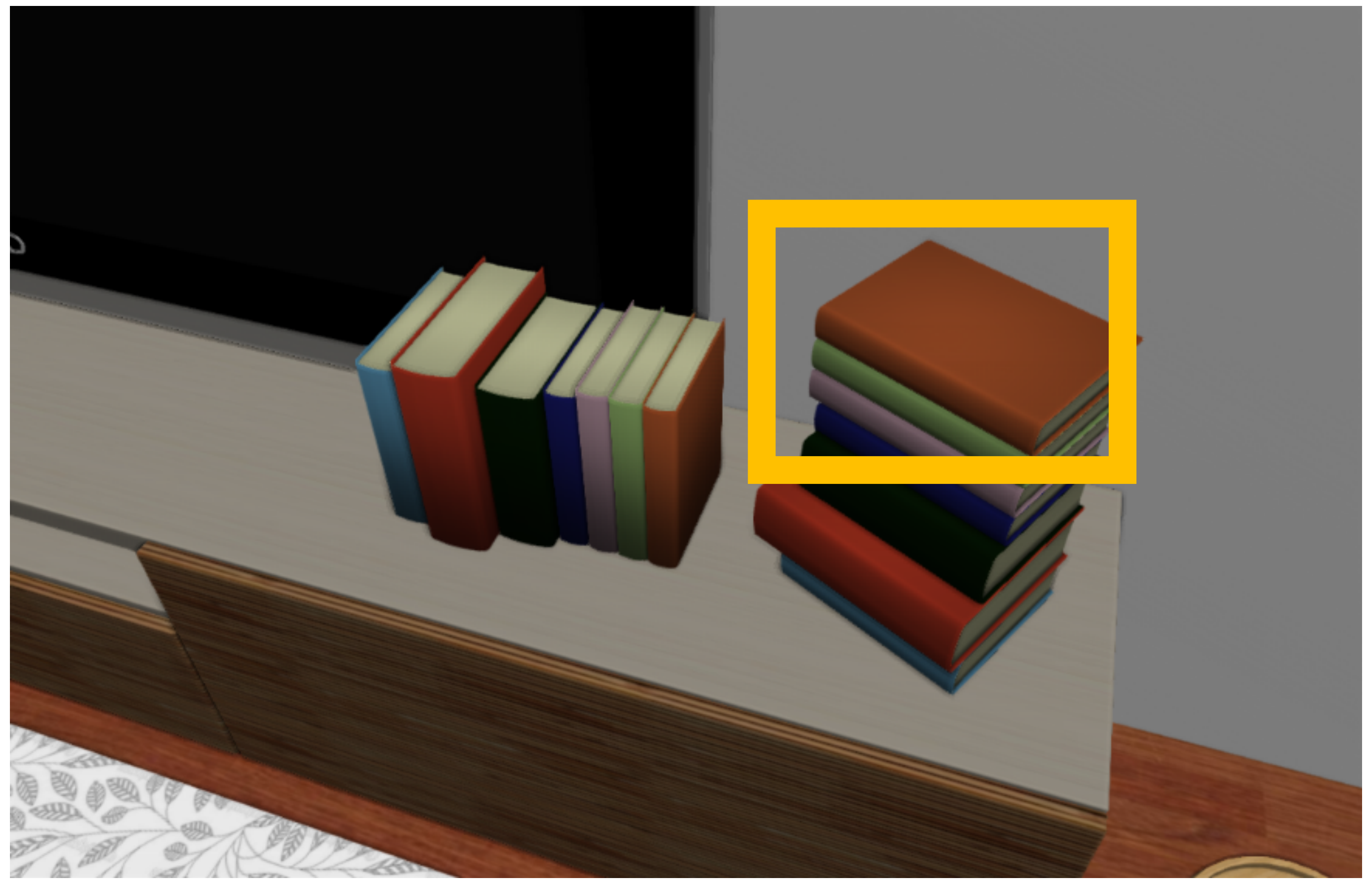
Pragmatics




Can you give me the orange book on top?

Need mental model of the other person

Speaker

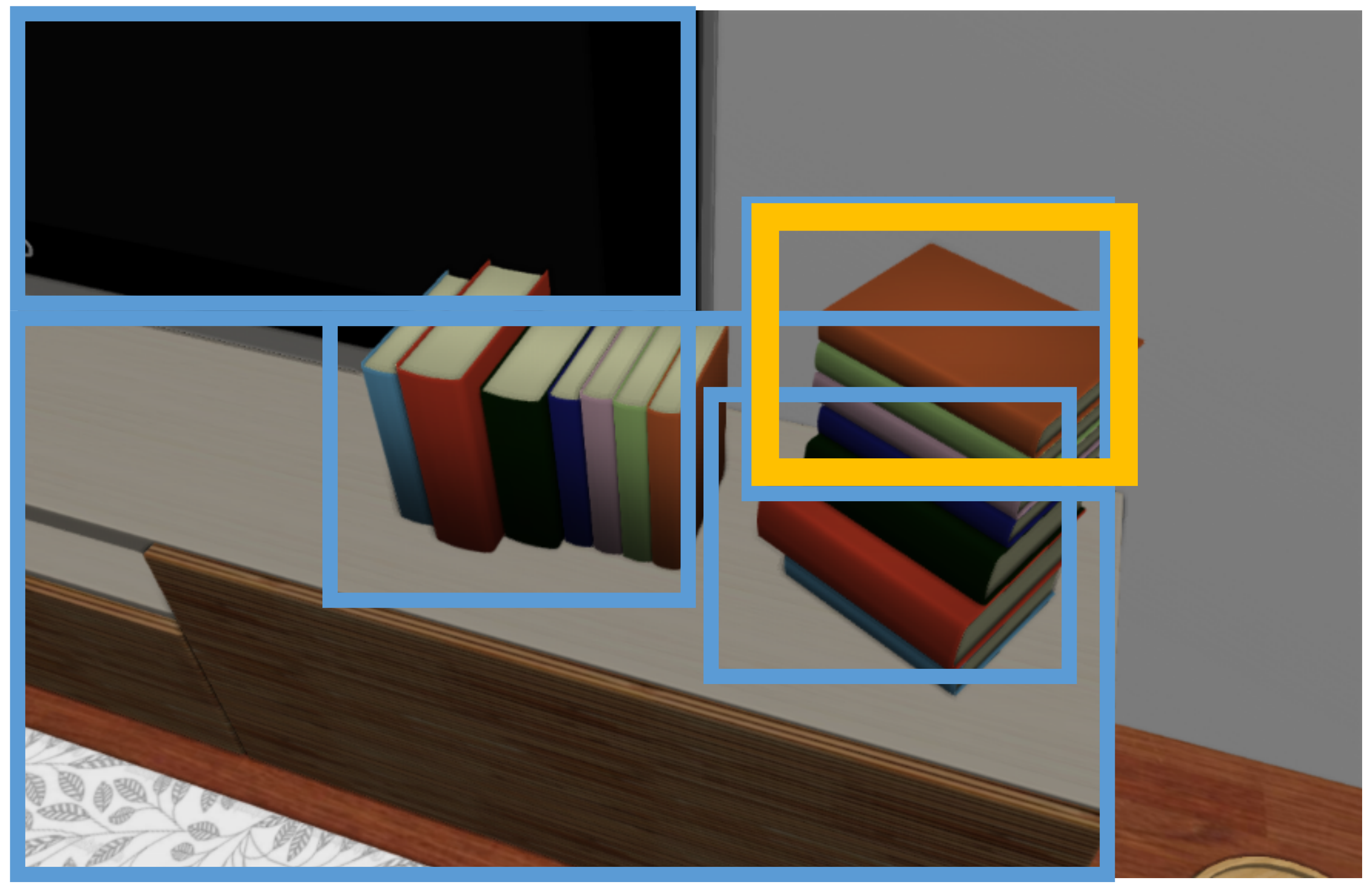


- Input: Image I with region R
- Output: Description $D*$

$$D* = \arg\max_{D} p(D \mid R, I)$$

Listener

- Input: Image I, with description D
  Generate candidate regions C
- Output: Region $R*$



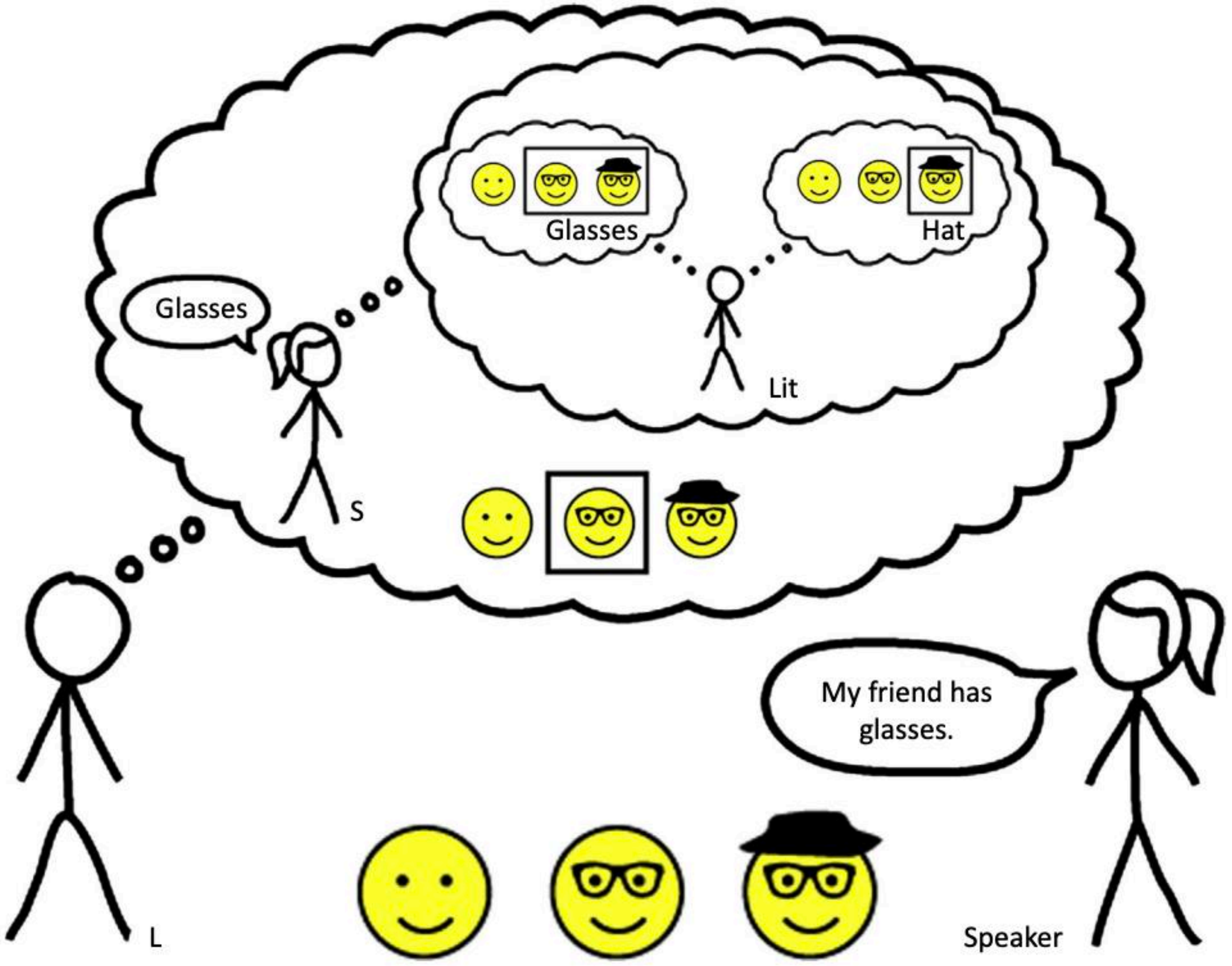
$$R* = \arg\max_{R \in C} p(R \mid D, I)$$

Speaker model

$$p(R \mid D, I) = \frac{p(D \mid R, I)\, p(R \mid I)}{\sum_{R' \in C} p(D \mid R', I)\, p(R' \mid I)}$$

Gricean maxims

- Cooperative, effective communication
- **Maxim of quantity:** Give as much information as need, and no more
- **Maxim of quality:** Provide truthful information, supported by evidence
- **Maxim of relation:** Be relevant, say things pertinent to discussion
- **Maxim of manner:** Be clear, brief and orderly, avoid obscurity and ambiguity



[Pragmatic Language Interpretation as Probabilistic Inference, Goodman and Frank 2016, http://langcog.stanford.edu/papers_new/goodman-2016-tics.pdf]

Is understanding language fundamental to solving AI?

Does solving AI mean solving language?

Reddit Ask Me Anything (Sept 2014) with Michael I. Jordan (UC Berkeley)

- AMA: “If you got a billion dollars to spend on a huge research project that you get to lead, what would you like to do?”
- michaelijordan: I'd use the billion dollars to build a **NASA-size program focusing on natural language processing (NLP)**, in all of its glory (semantics, pragmatics, etc). Intellectually I think that NLP is fascinating, allowing us to focus on highly-structured inference problems, on issues that go to the core of "what is thought" but remain eminently practical, and on a technology that surely would make the world a better place. Although current deep learning research tends to claim to encompass NLP, I'm (1) much less convinced about the strength of the results, compared to the results in, say, vision; (2) much less convinced in the case of NLP than, say, vision, the way to go is to couple huge amounts of data with black-box learning architectures.

[https://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan/]

Does solving AI mean solving language?

Reddit Ask Me Anything (Nov 2014) with Geoff Hinton (U Toronto and Google)

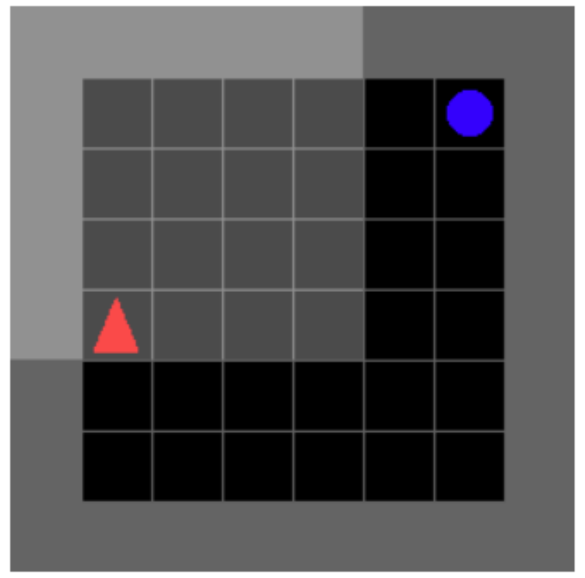
- I think that the most exciting areas over the next five years will be really understanding text and videos. I will be disappointed if in five years' time we do not have something that can watch a YouTube video and tell a story about what happened. In a few years time we will put [Deep Learning] on a chip that fits into someone's ear and have an English-decoding chip that's just like a real Babel fish.

[https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton/]

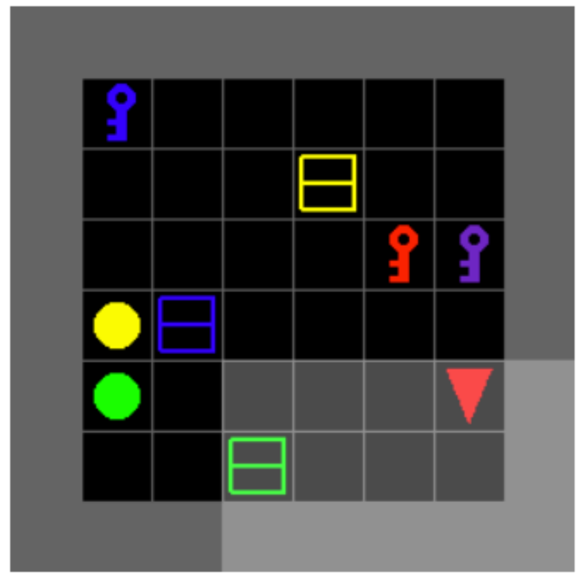
Frameworks for understanding grounded language (with perception and actions)

BabyAI

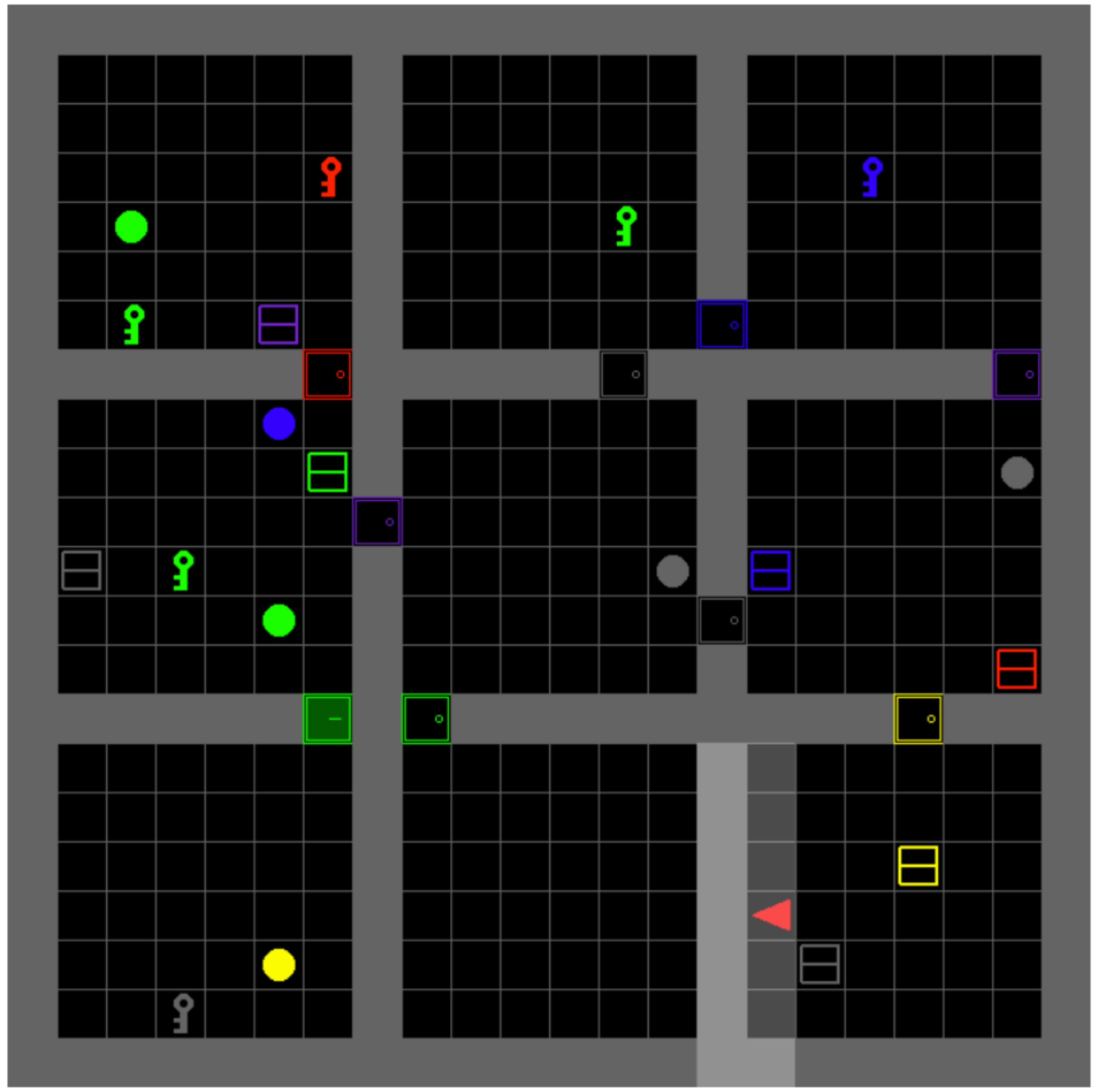
- Grid Environment
- Generated (synthetic language) using grammar
- Easy to hard levels
- Studies grounding and compositionality



(a) GoToObj: "go to the blue ball"



(b) PutNextLocal: "put the blue key next to the green ball"

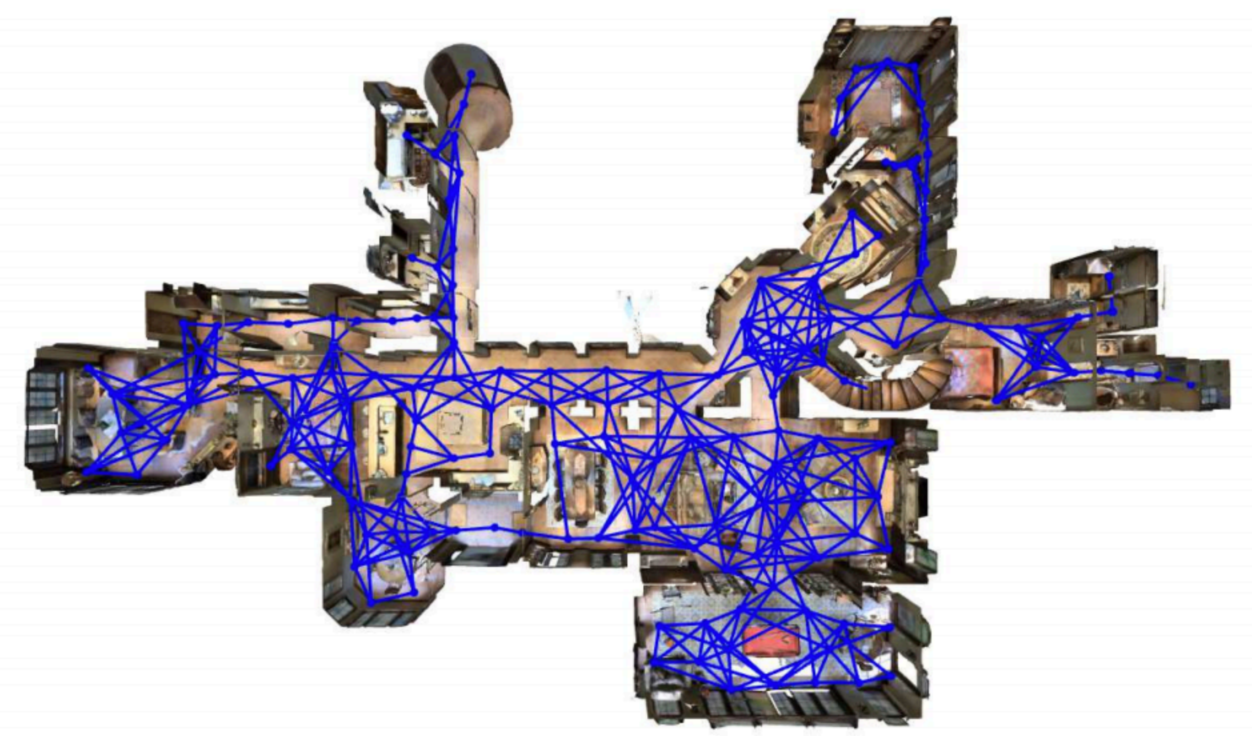


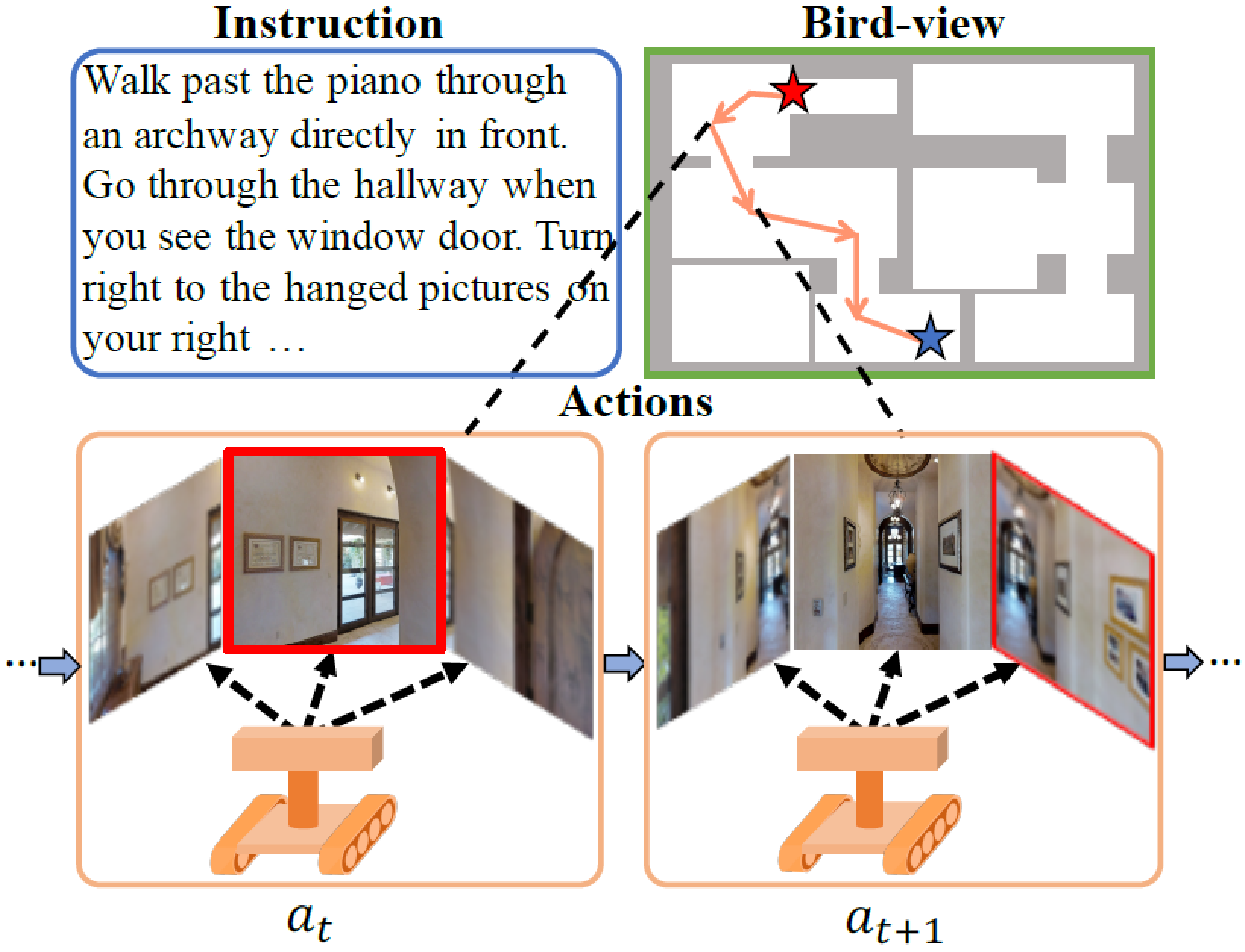
(c) BossLevel: "pick up the grey box behind you, then go to the grey key and open a door". Note that the green door near the bottom left needs to be unlocked with a green key, but this is not explicitly stated in the instruction.

BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning [Chevalier-Boisvert et al 2018, https://arxiv.org/pdf/1810.08272.pdf]

Vision-and-language Navigation

- More realistic houses
- Human instructions navigation
- Discrete action space
- Navigation graph

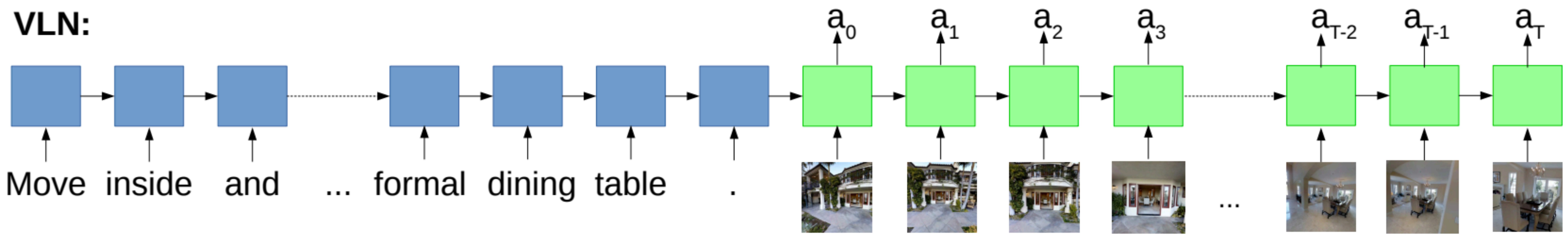




Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments
[Anderson et al 2018, https://bringmeaspoon.org/]

Vision-and-language Navigation

- Sequence of words to sequence of actions!



Input Images at each time step

Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments [Anderson et al 2018, https://bringmeaspoon.org/]

Vision-and-language Navigation

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

ALFRED

- More realistic houses
- Sequence of human instructions for common household tasks
- Study embodied language understanding

ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks
[Shridhar et al 2019, https://askforalfred.com/]