Natural Language Processing and Robotics

Jesse Thomason

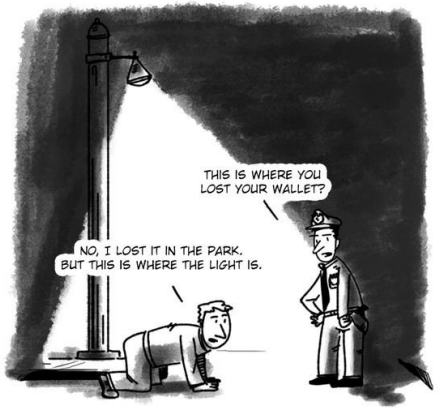
[http://glamor.rocks/]

USC Viterbi School of Engineering

✨GLAMOR

Streetlight Effect

A policeman sees a drunk man searching for something under a streetlight and asks what the drunk has lost. He says he lost his keys and they both look under the streetlight together. After a few minutes the policeman asks if he is sure he lost them here, and the drunk replies, no, and that he lost them in the park. The policeman asks why he is searching here, and the drunk replies, "this is where the light is".

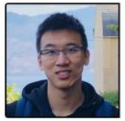


https://en.wikipedia.org/wiki/Streetlight_effect

GLAMOR Lab

- **G**rounding **L**anguage in **A**ctions, **M**ultimodal **O**bservations, and **R**obots

PhD Students



Bill Zhu
2nd year PhD student (co-Robin Jia)
Compositional Semantics
Vision-Language Navigation

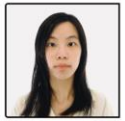
Ishika Singh
2nd year PhD student
VLN and VL Interaction

Tejas Srinivasan
2nd year PhD student
Vision-Language Alignment
VL Continual Learning

Leticia Pinto-Alva
2nd year PhD student
Vision-and-Language

Anthony Liang
2nd year PhD student
Language-guided RL
RL for Dialogue

Lee Kezar
3rd year PhD student
VL for ASL



Ting-Yun Chang
2nd year PhD student (co-Robin Jia)
Pre-trained Model Understanding

Abrar Anwar
2nd year PhD student
Speech and Gesture
Assistive Robotics

Undergraduate and Masters Students

Elle Szabo
Senior Undergraduate @ USC
Computer Science

Chu Fang
Senior Undergraduate @ USC
Computer Science (Games)
URAP

Flora Jia
Junior Undergraduate @ USC
Computer Science and Applied Mathematics
CURVE Fellowship

Junu Song
Junior Undergraduate @ USC
Computer Science
CURVE Fellowship

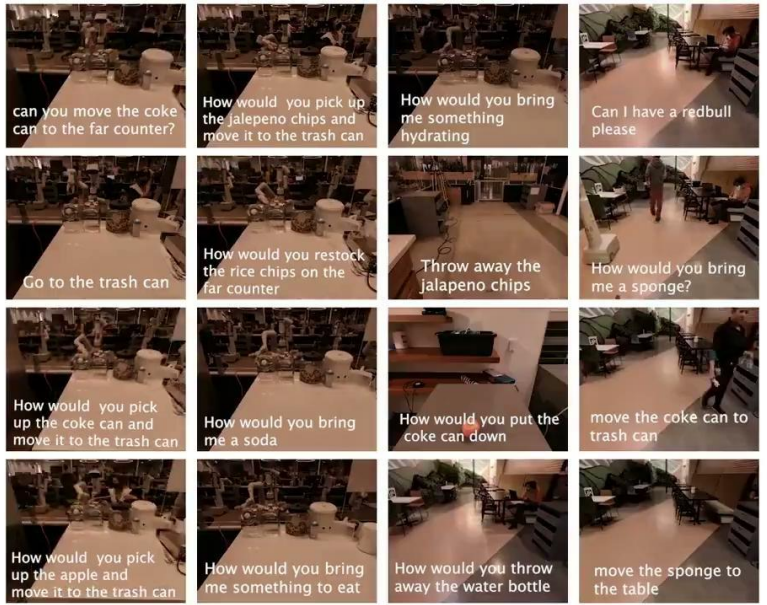
Yuan Huang
2nd year Masters Student @ USC
Electrical and Computer Engineering
Specializing in Machine Learning and Data Science

Leslie Moreno
Sophomore Undergraduate @ USC
Computer Engineering and Computer Science
CURVE Fellowship (co-Maja Mataric)

Julie Kim
Sophomore Undergraduate @ USC
Computer Engineering and Computer Science
CURVE Fellowship (co-Maja Mataric)

http://glamor.rocks/

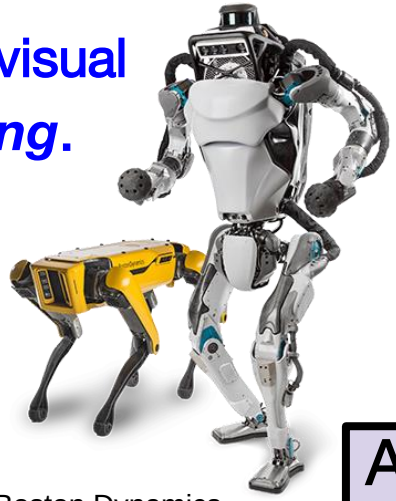
Why Would a Robot Need Language?



https://emprise.cs.cornell.edu/hrcom/ ; https://say-can.github.io/

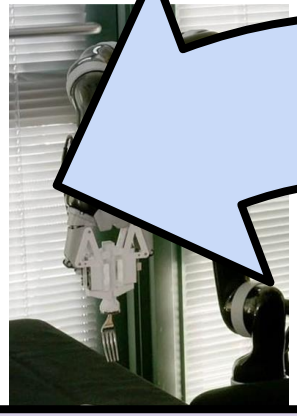
Why Aren't Our Robots Language-Guided Already?

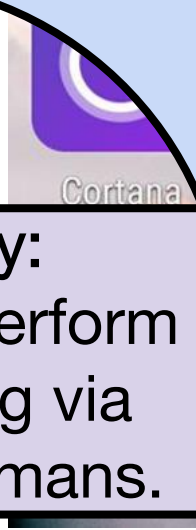
Require visual *grounding*.



Boston Dynamics

iRobot





???

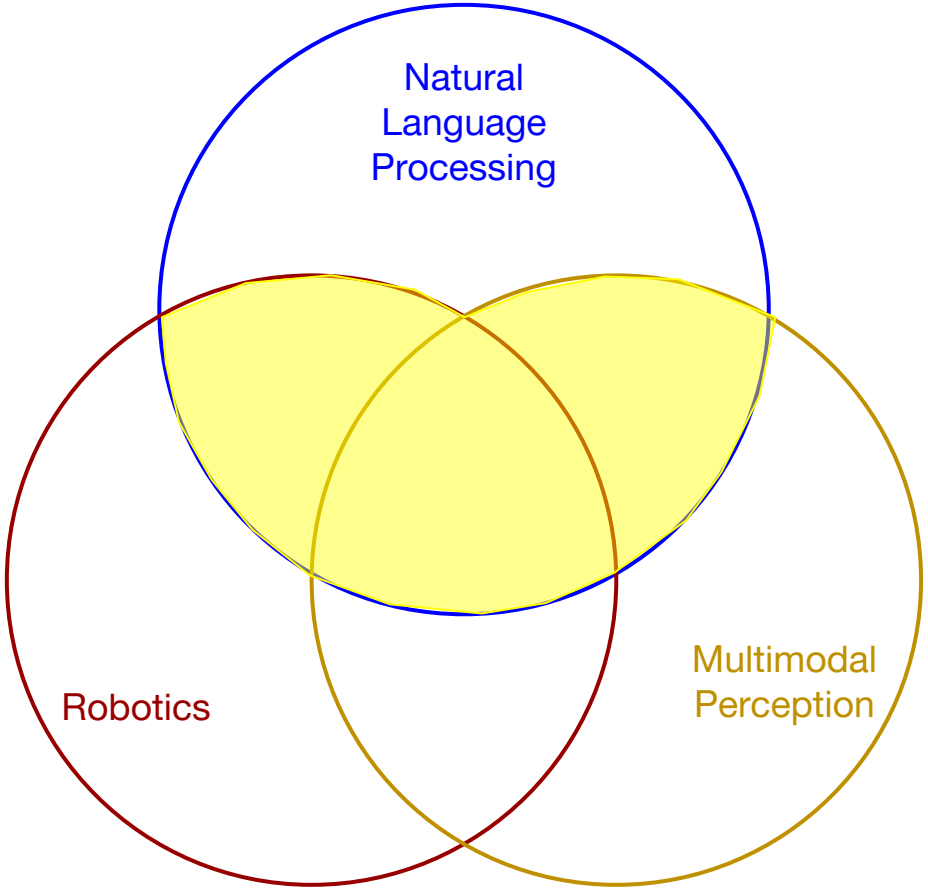
No visual connections.

Amazon

Google

Approach Summary: Enable robots to perform language grounding via interaction with humans.

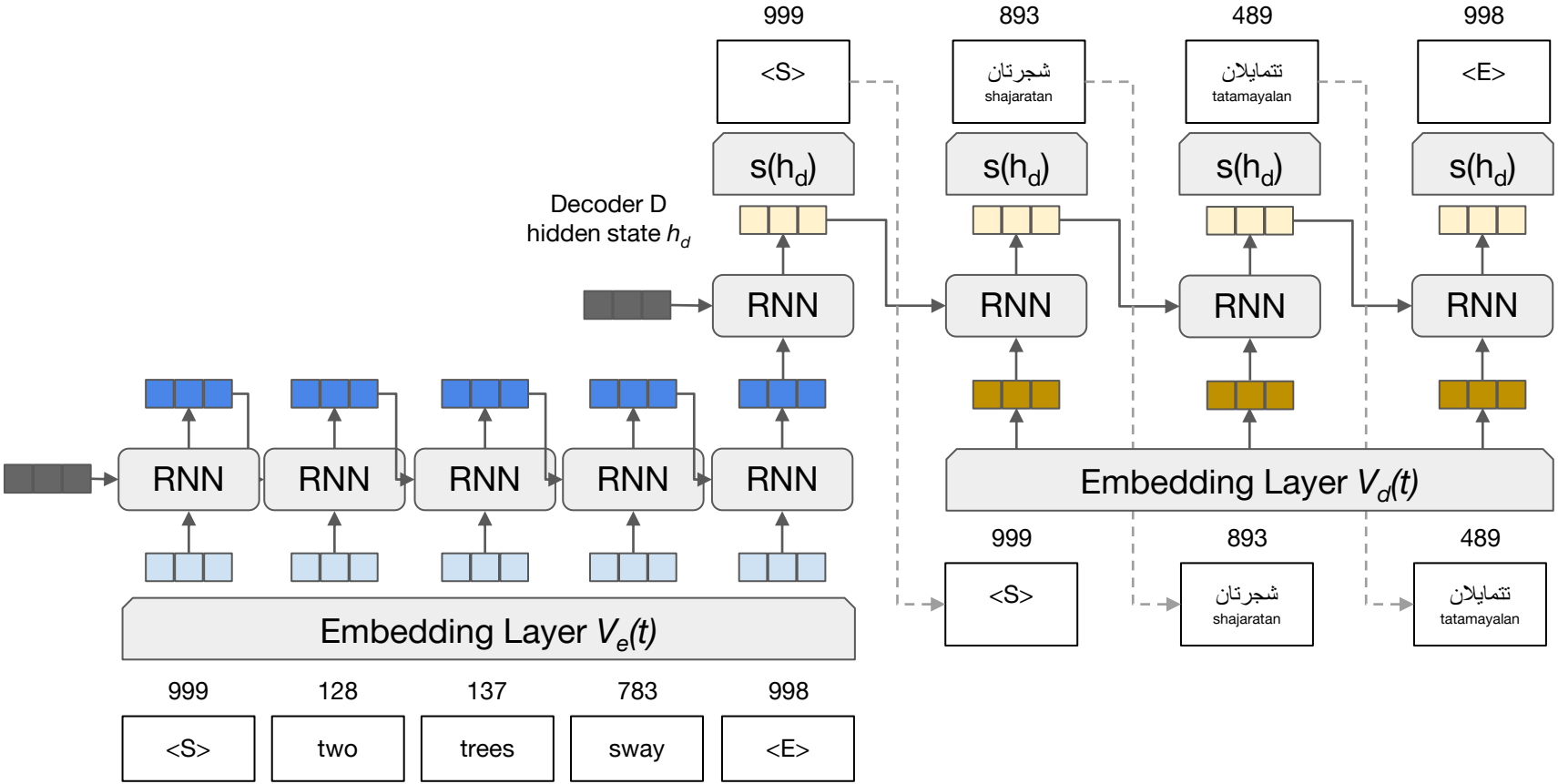
GLAMOR Lab Work



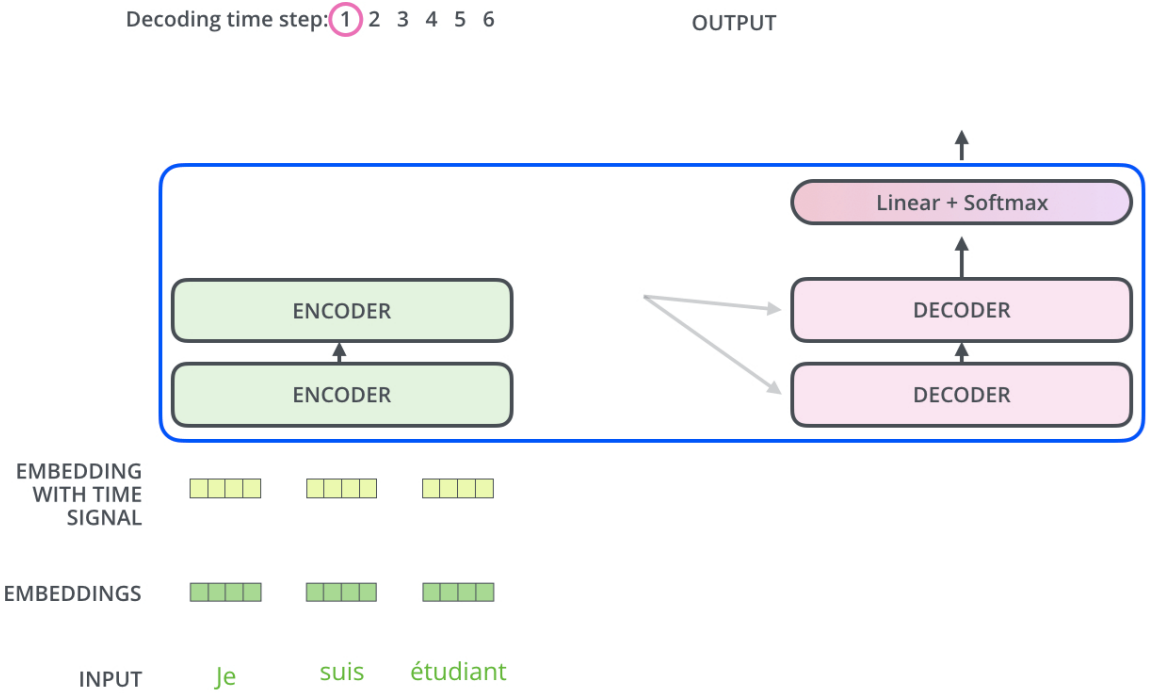
Natural Language Processing

- Concerned with turning “in the wild” human language into something machine readable for downstream tasks like
  - Text classification
    - Sentiment analysis, fake news detection, hate speech detection, spam filters, …
  - Textual language modeling
    - Text generation (e.g., GPT-3), autocomplete on your phone, spelling correction suggestions, machine translation, …

Machine Translation: Basic Idea

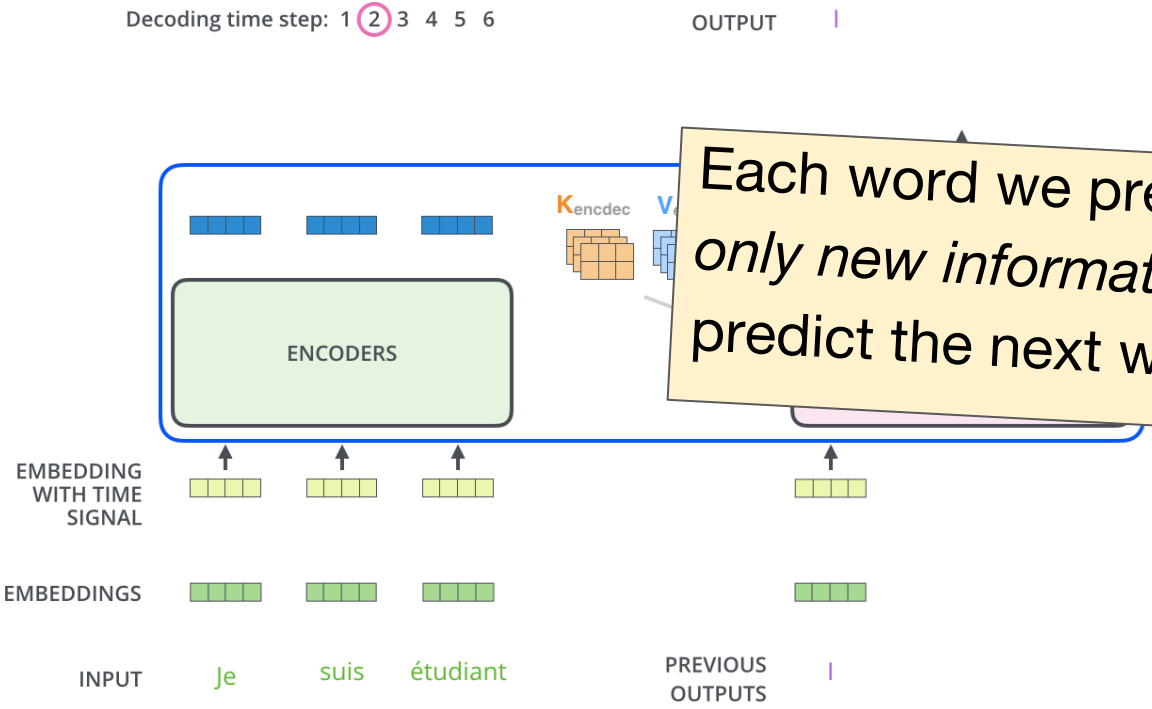


Machine Translation: Transformer Models



https://jalammar.github.io/illustrated-transformer/

Machine Translation: Transformer Models

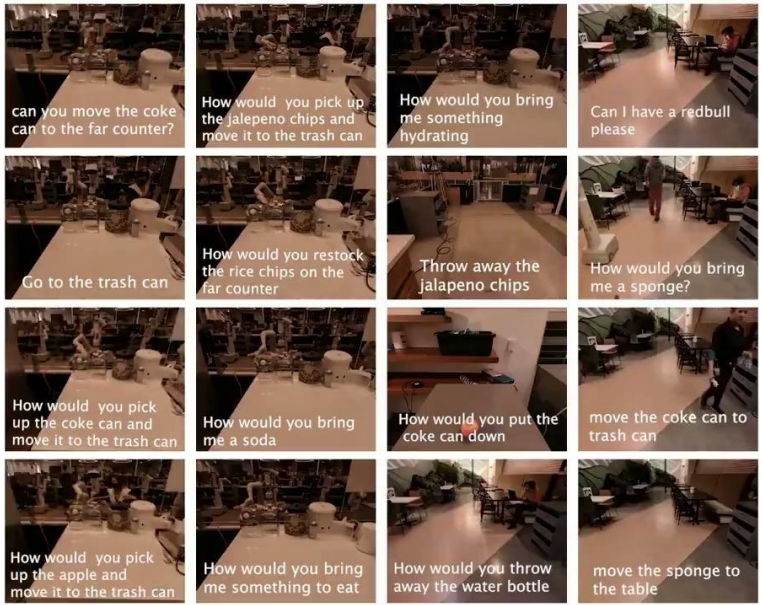


Each word we predict is the *only new information we get* to predict the next word.

https://jalammar.github.io/illustrated-transformer/

Machine Translation for Robots

- Source language input:
  - “Can you move the coke can to the far counter?”
- Target language output:
  - `look down, rotate right, move forward, move forward, rotate right, grasp coke can, rotate right, move forward, move forward, rotate right, place coke can`



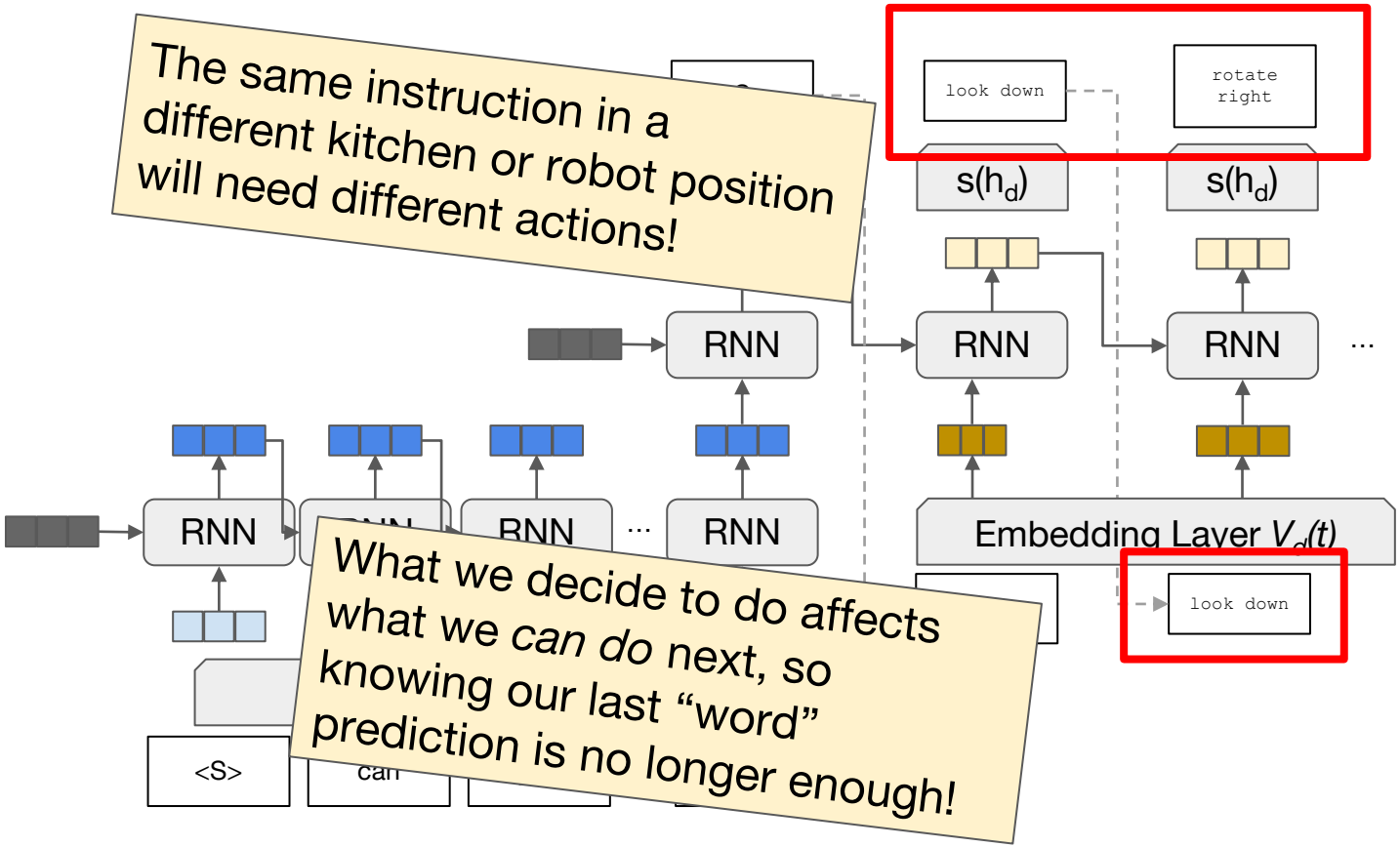
https://say-can.github.io/

Machine Translation for Robots

- Source language input:
  - “Can you move the coke can to the far counter?”
- Target language output:
  - `look down, rotate right, move forward, move forward, rotate right, grasp coke can, rotate right, move forward, move forward, rotate right, place coke can`

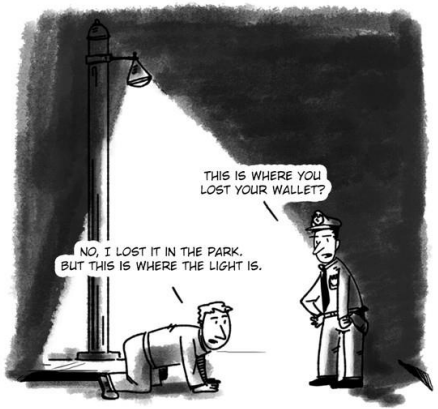
- How many input/output examples would you need to learn a *policy* that reliably translates English to sequences of robot actions?

Machine Translation for Robots



Streetlight Effect

A policeman sees a drunk man searching for something under a streetlight and asks what the drunk has lost. He says he lost his keys and they both look under the streetlight together. After a few minutes the policeman asks if he is sure he lost them here, and the drunk replies, no, and that he lost them in the park. The policeman asks why he is searching here, and the drunk replies, "this is where the light is".



https://en.wikipedia.org/wiki/Streetlight_effect

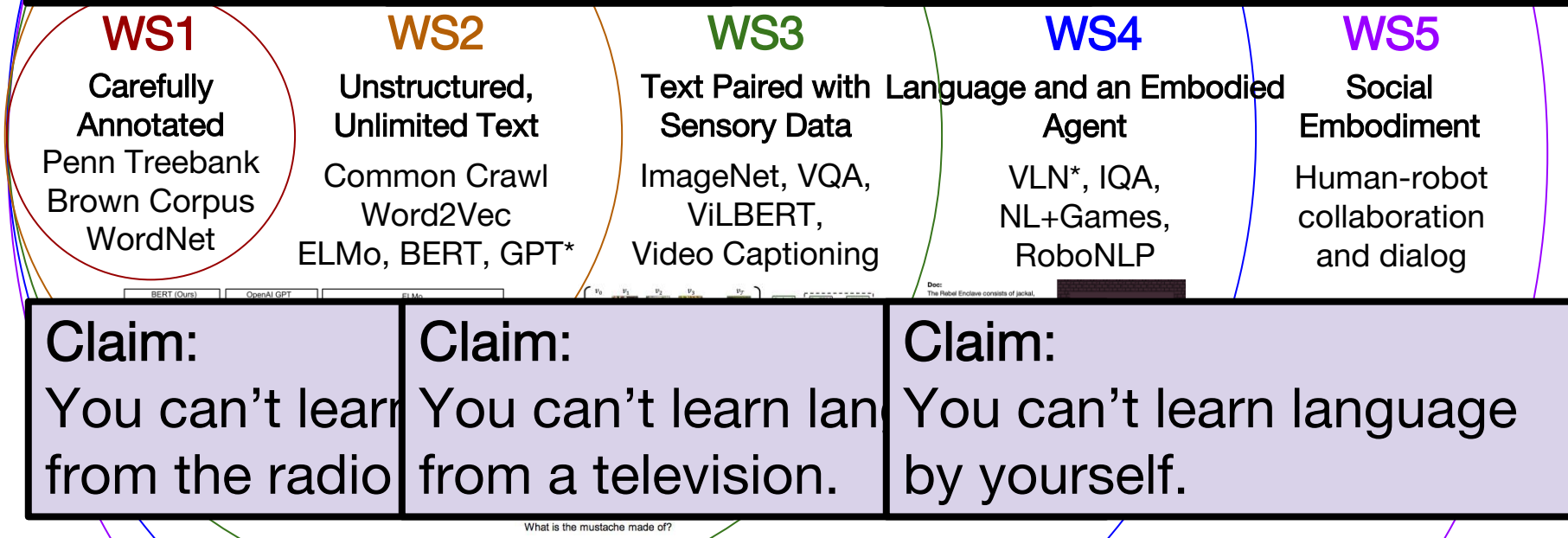
[Bisk, Holtzman, **Thomason**, et al., EMNLP'20]

Experience Grounds Language - World Scopes

Upshot:
Most current
operates at th

Upshot:
Most L folks playing
and vice versa are

Upshot:
Maybe this is AI Complete;
but so is "language"



Claim:
You can't learn
from the radio

Claim:
You can't learn lan
from a television.

Claim:
You can't learn language
by yourself.

[Bisk, Holtzman, **Thomason**, et al., EMNLP'20]

Experience Grounds Language - World Scopes

Common Crawl

WS1

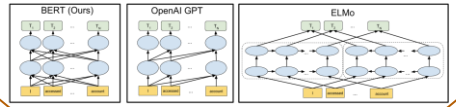
Carefully Annotated

Penn Treebank
Brown Corpus
WordNet

WS2

Unstructured Unlimited Text

Common Crawl
Word2Vec
ELMo, BERT, GPT*



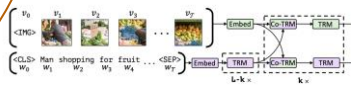
the trail climbs steadily uphill most of the way.

the stars in the night sky.

WS3

Text Paired with Sensory Data

ImageNet, VQA,
ViLBERT,
Video Captioning



What color are her eyes?
What is the mustache made of?

WS4

Language and an Embodied Agent

VLN*, IQA,
NL+Games,
RoboNLP

WS5

Social Embodiment

Human-robot collaboration and dialog

Images and Text for Retrieval

Text Query

"A tropical bird perches in the jungle."

Candidate Images






Images and Text for Retrieval

Image Query

Candidate Captions

“A rabbit sits in the palm of a hand.”

“Men are talking on a basketball court.”

“A tropical bird perches in the jungle.”

“Children play soccer in a field.”

“A white fox is looking at the camera.”

“Sports equipment is staged for a photo.”

Formulating the Retrieval Problem

Images	$\mathcal{I}$
Captions	$\mathcal{L}$
Scoring Function	$F : \mathcal{I} \times \mathcal{L} \to \mathbb{R}$

Formulating the Retrieval Problem

$$F : \mathcal{I} \times \mathcal{L} \rightarrow \mathbb{R}$$

F(# red pixels, # word “red”) = 1 if (a > 0, b > 0) else 0

F(RGB bins, color word counts) =
sum of #(color word, bin)/(#color word + #bin) in training data

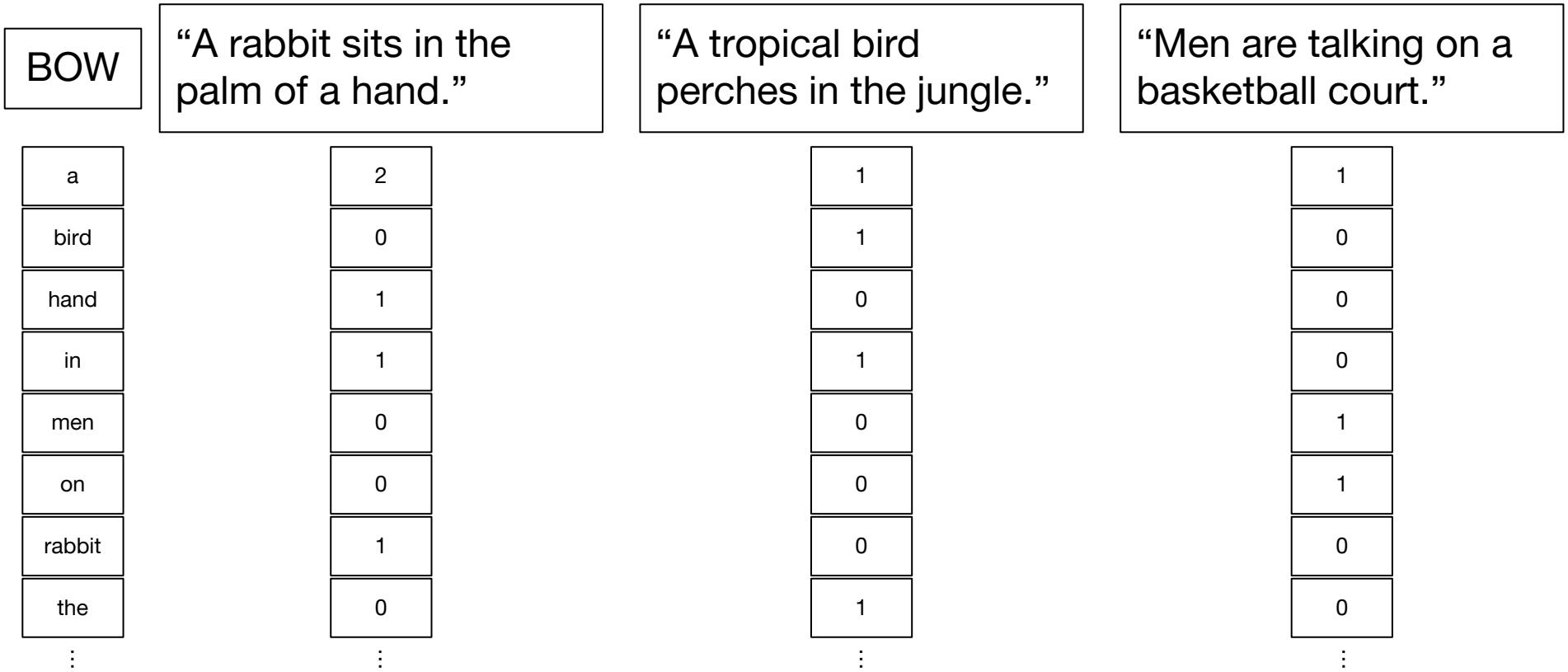
F(detected objects, word counts) =
sum of #(object, word)/(#object + #word) in training data

F(image pixels, token sequence) =
NN trained with contrastive matching loss

Formulating the Retrieval Problem

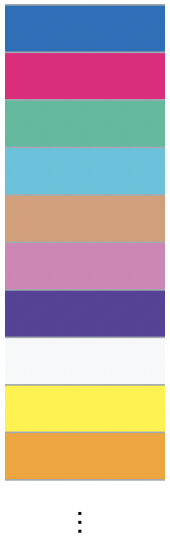
Image Features	$\psi : \mathcal{I} \rightarrow \mathbb{R}^n$
Caption Features	$\omega : \mathcal{L} \rightarrow \mathbb{R}^m$
Scoring Function	$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$

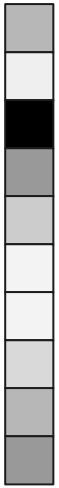
Feature Extraction for Language

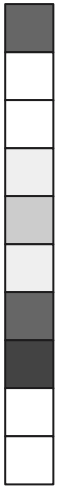


Feature Extraction for Vision

RGB
kNN bins





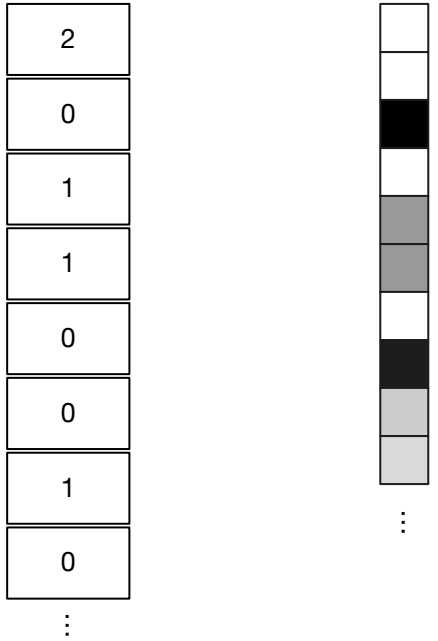


A Simple Retrieval Solution

$$\psi : \mathcal{I} \rightarrow \mathbb{R}^n$$

$$\omega : \mathcal{L} \rightarrow \mathbb{R}^m$$

$$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$$



Contextual Information

You shall know a word by the company it keeps (Firth, J. R. 1957:11)

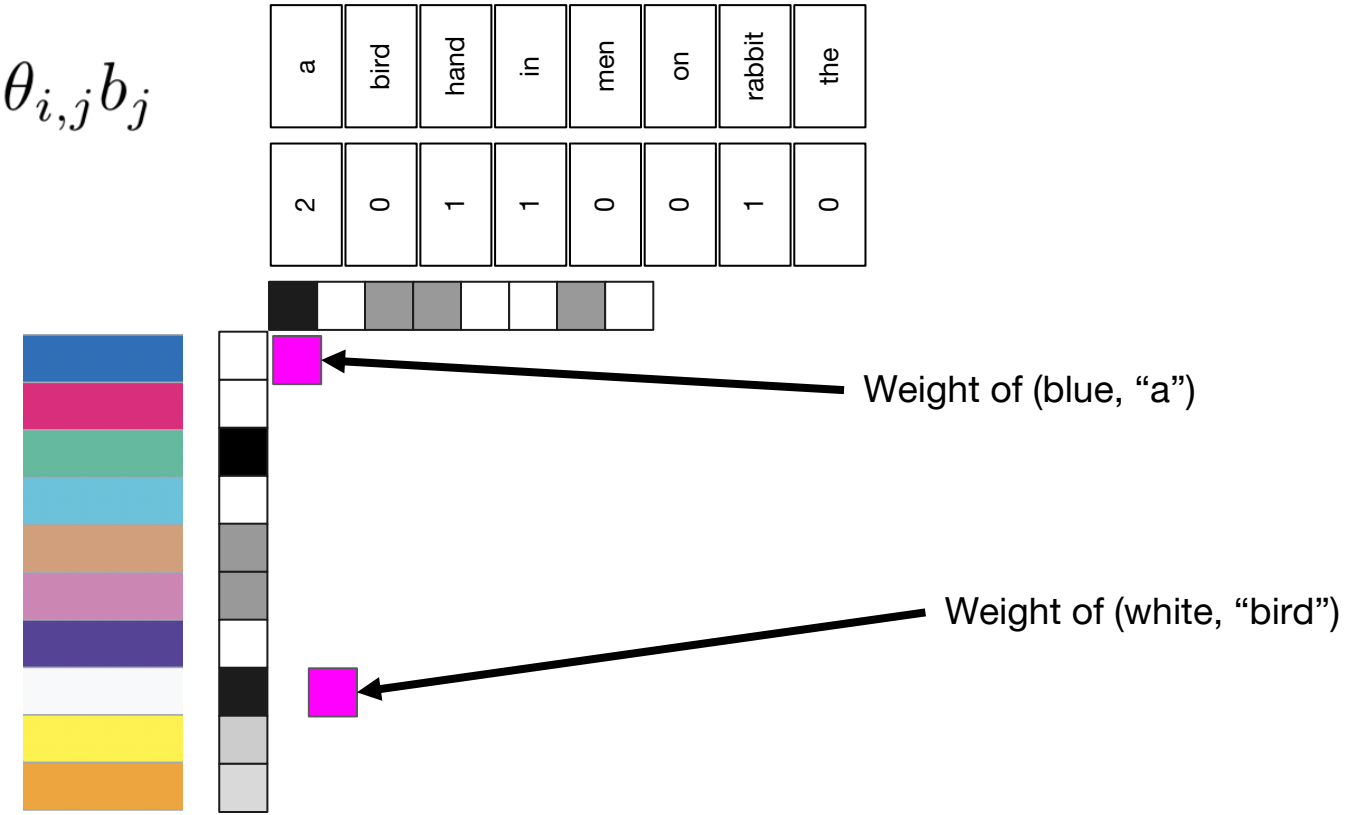
You shall know the features of modality A by the company they keep in the features of modality B. And vice versa.

Formulating the Retrieval Problem as a Linear Model

$$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$$

Formulating the Retrieval Problem as a Linear Model

$$\sum_{i=1,j=1}^{n,m} a_i \theta_{i,j} b_j$$



Pointwise Mutual Information

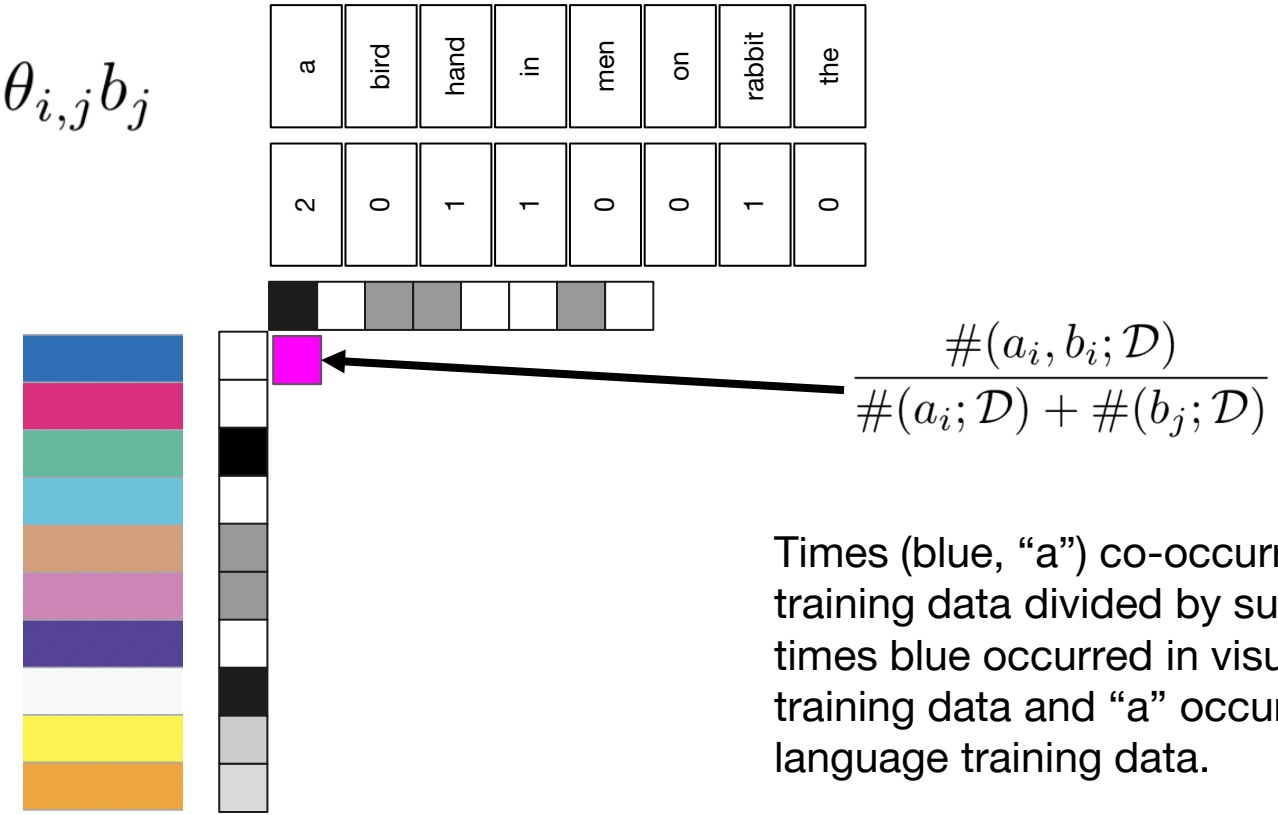
$$\mathrm{pmi}(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$$

$$F(A,B) = \mathrm{pmi}(A;B); A \in \mathcal{I}, B \in \mathcal{L}$$

$$
\begin{aligned}
F(\phi(A), \omega(B)) &= \text{pmi}(\vec{a}; \vec{b}) \\
&= \log \frac{p(\vec{a}, \vec{b})}{p(\vec{a})p(\vec{b})}
\end{aligned}
$$

Formulating the Retrieval Problem as a Linear Model

$$\sum_{i=1,j=1}^{n,m} a_i \theta_{i,j} b_j$$

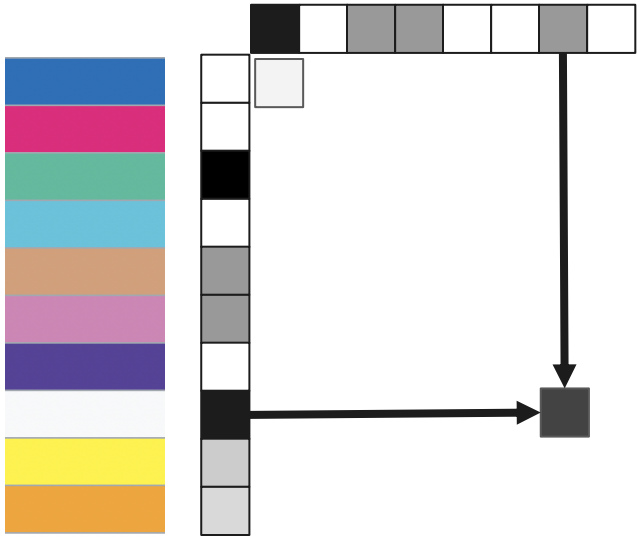


Times (blue, “a”) co-occurred in training data divided by sum of times blue occurred in visual training data and “a” occurred in language training data.

Formulating the Retrieval Problem as a Linear Model

$$\sum_{i=1,j=1}^{n,m} a_i \theta_{i,j} b_j$$

a	bird	hand	in	men	on	rabbit	the
2	0	1	1	0	0	1	0



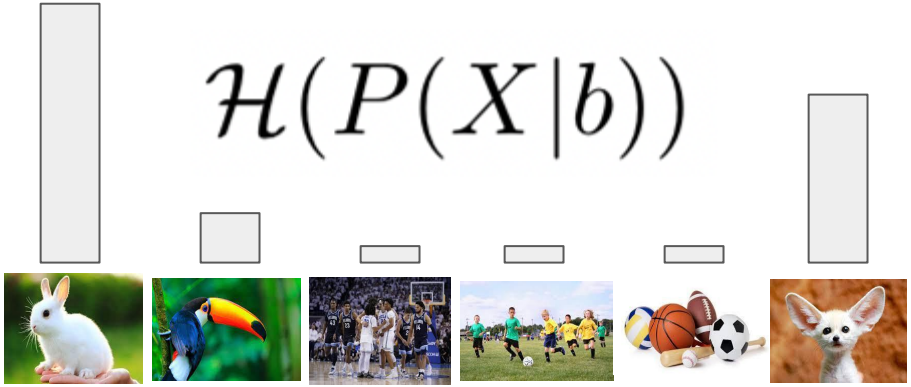
Feature Extraction and Representation Learning

$$X \subset \mathcal{I}, Y \subset \mathcal{L}$$

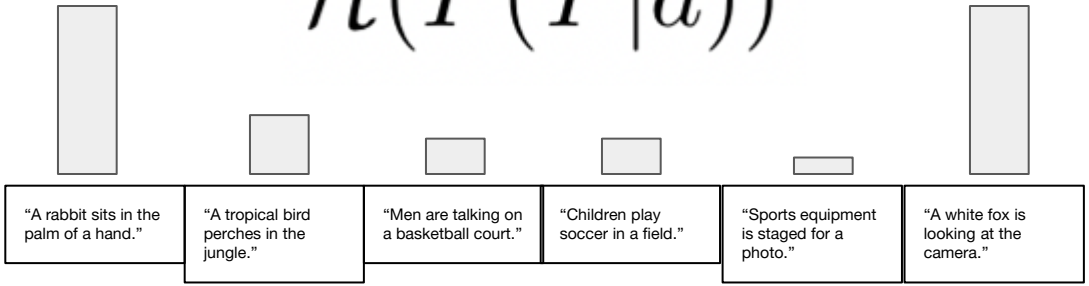
$$a \in \psi(\mathcal{I}), b \in \omega(\mathcal{L})$$

$$\mathcal{H}(P(Y|a))$$

$$\mathcal{H}(P(X|b))$$



b indicates “rabbit”



a indicates detected furry ear

Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

$$\mathcal{H}(P(X|b))$$

Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

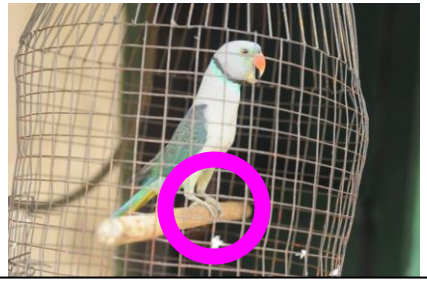
“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

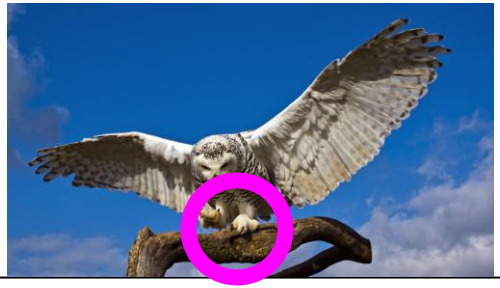
Stemming	a	tropic	bird	perch	in	the	jungl	.
	perch	parakeet	in	a	wire	frame	cage	.
	a	snow	owl	land	on	wood	perch	.

Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”



“Perching parakeet in a wire frame cage.”



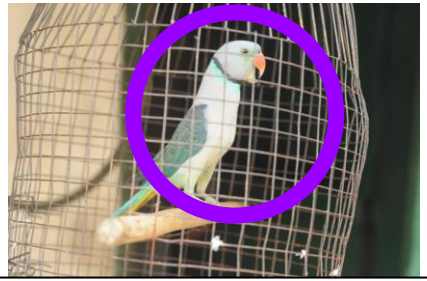
“A snow owl lands on a wooden perch.”

Stemming

a	tropic	bird	perch	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	perch	.

Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”



“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

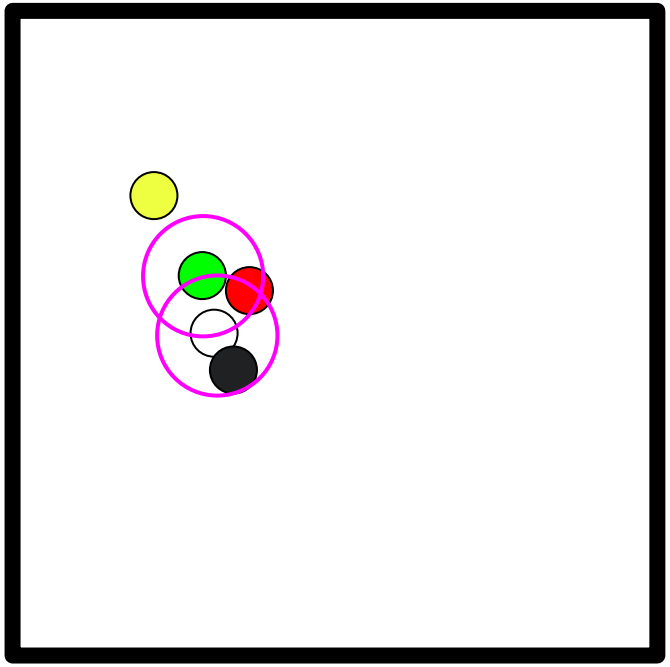
a	tropic	bird	perch	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	perch	.

Pretrained Language Token Embeddings

- Cosine similarity of “bird”, “owl”, “parakeet” helps share information across training data
- Taking a guess: what are the nearest neighbors of “green” in word2vec embedding space?
  - Blue, white, red, yellow, black, grey, purple, pink, light, gray
- What can we learn for highly polysemous words like “play”?
  - “play guitar”, “play piano”, “play basketball”, “play tag”
- Text-based embeddings help us share information ***only to the extent that words used in a similar context share a visual representation***
  - which is true for, say, birds or trees, but not colors

Pretrained Language Token Embeddings

$\omega(\mathcal{L})$



Training

Inference

“A white bunny rabbit held in a green field.”

“A black rabbit watches a red sunset.”

Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

$$\mathcal{H}(P(Y|a))$$

Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

Object Classification

toucan

cockatoo

great grey owl

https://deeplearning.cms.waikato.ac.nz/user-guide/class-maps/IMAGENET/

Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

Object Detection

bird	potted plant
bird	boat
bird	aeroplane

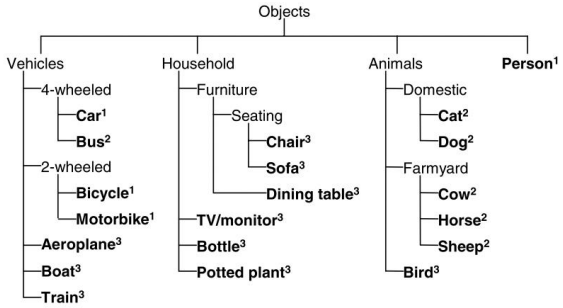
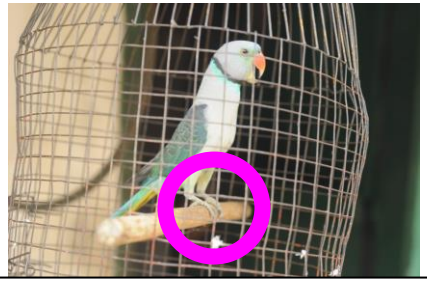
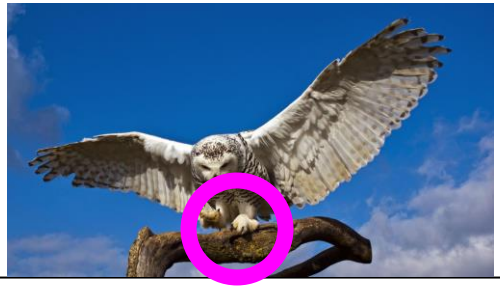


Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”



“Perching parakeet in a wire frame cage.”



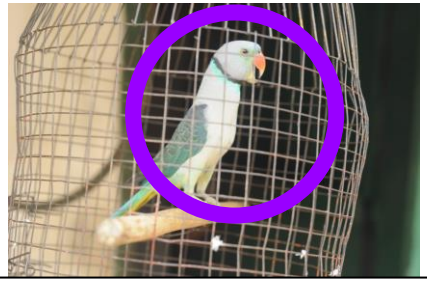
“A snow owl lands on a wooden perch.”

bird	potted plant
bird	boat
bird	aeroplane

a	tropic	bird	perch	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	perch	.

Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”



“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

bird	potted plant
bird	boat
bird	aeroplane

a	tropic	bird	perch	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	perch	.

"A rabbit sits in the palm of a hand."

"A white fox is looking at the camera."

"Children play soccer in a field."

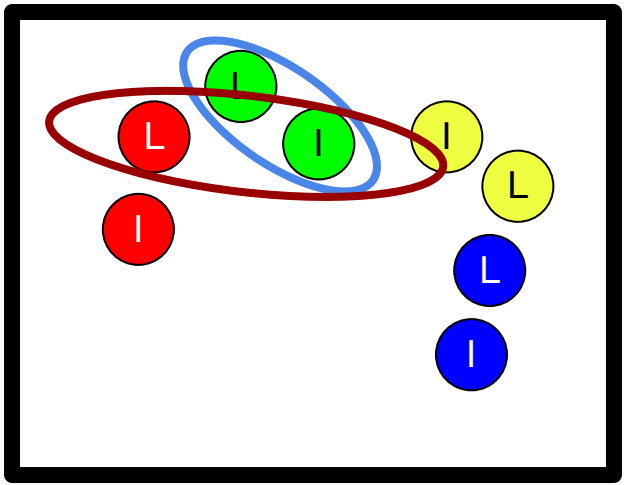
"Men are talking on a basketball court."

Image Embedder $\psi(\mathcal{I})$

Caption Embedder $\omega(\mathcal{L})$

Pull matching imag embeddings together.

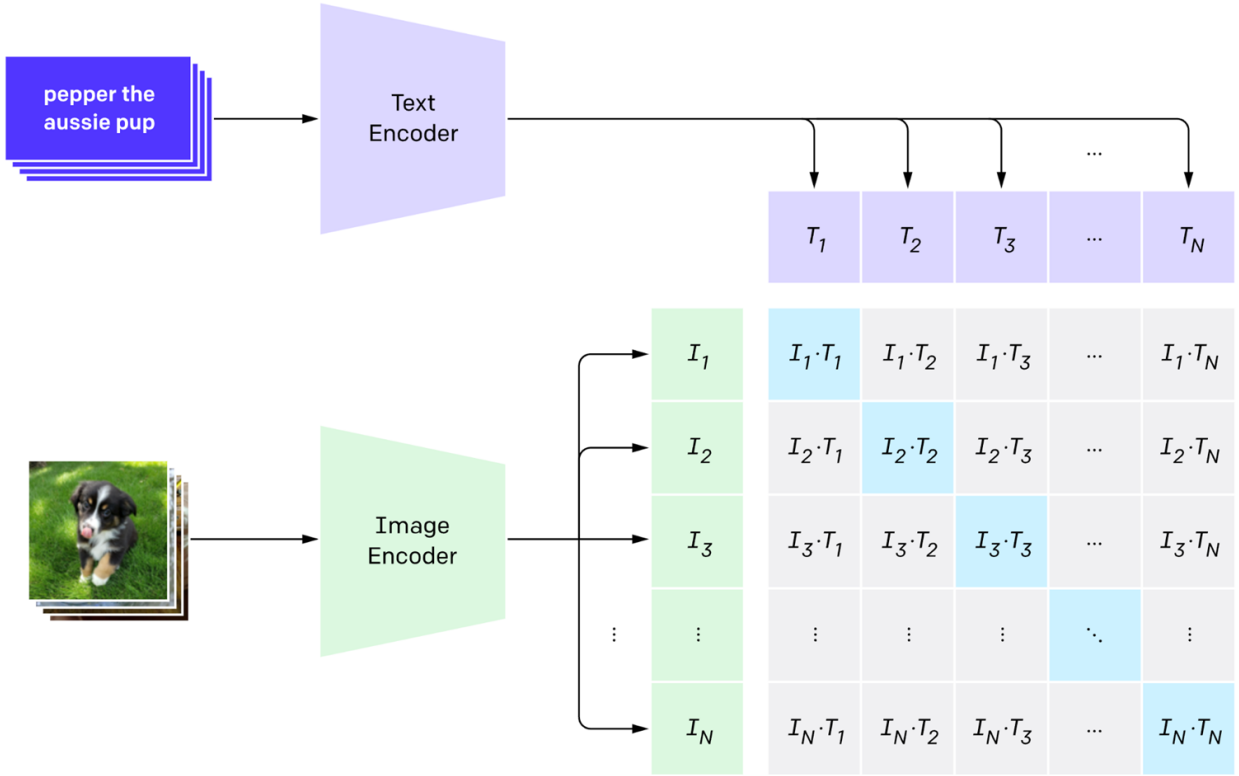
Learned Projection



Push distractor on embeddings apart.

Learned Projection

Contrastive Language–Image Pre-training (CLIP)



https://openai.com/blog/clip/

[Bisk, Holtzman, **Thomason**, et al., EMNLP'20]

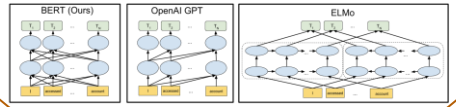
Experience Grounds Language - World Scopes

Common Crawl

WS1

Carefully Annotated

Penn Treebank
Brown Corpus
WordNet



WS2

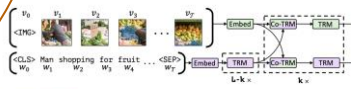
Unstructured Unlimited Text

Common Crawl
Word2Vec
ELMo, BERT, GPT*

WS3

Text Paired with Sensory Data

ImageNet, VQA,
ViLBERT,
Video Captioning



WS4

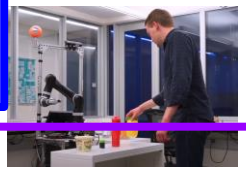
Language and an Embodied Agent

VLN*, IQA,
NL+Games,
RoboNLP

WS5

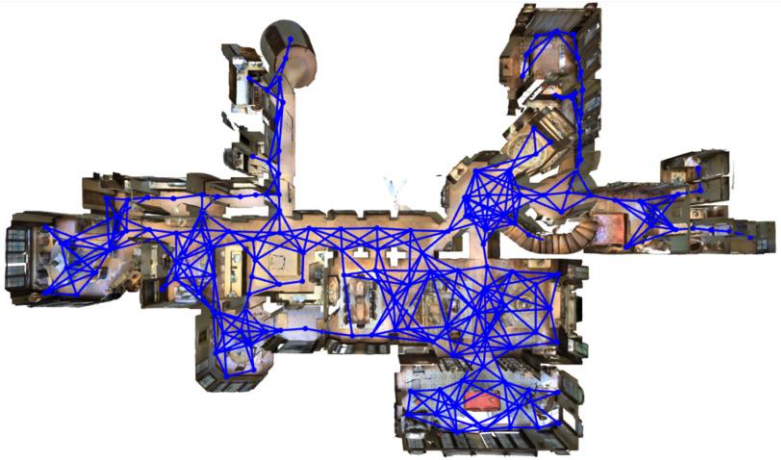
Social Embodiment

Human-robot collaboration and dialog



Vision-and-Language Navigation (VLN)

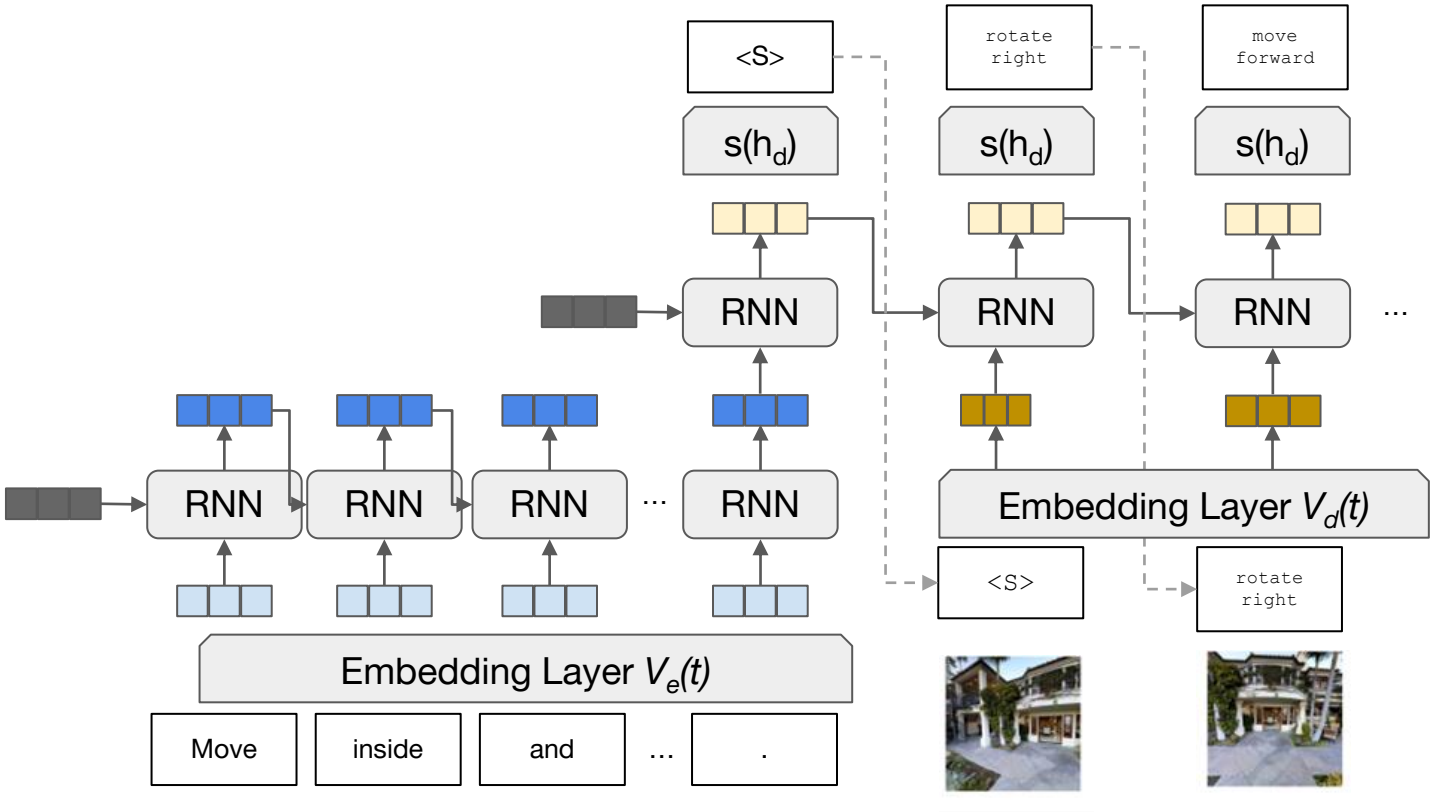
Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.



- Source language input: instructions in English
- Target language output: sequence of actions like “forward”, “rotate left”, “rotate right”, and “stop”

https://bringmeaspoon.org/

Machine Translation for Robots



Try to find the goal room given the text (left click and drag to move rooms, right click (like a phone) to scroll). Ask your partner for help using the chat. Your partner can only see movements on their screen.

Alex: The goal room connects a wall.

Alex	Partner
Can the hall be too difficult?	
	Left into the hall, it takes a lot a long room
Should I go outside?	

Your partner has asked for help in the chat window! View the best next few steps by clicking "Show Next Route" below, and refer on screen back to them.

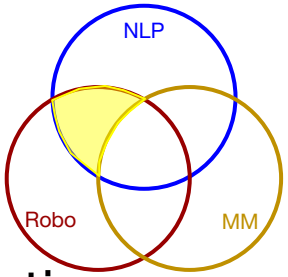
Alex: The goal room connects a wall.

Alex	Partner
	Can the hall be too difficult?
Left into the hall, it takes a lot a long room	
	Should I go outside?

Activity: Cooperative Vision-and-Dialogue Navigation

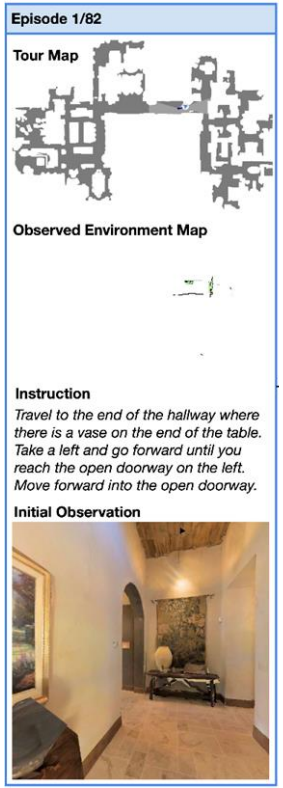
- If we ask a computer to do it, we have to be able to do it too!
- https://cvdn.dev/
  - “Two-player Demo”
  - Scroll to the bottom, click “Start Task”

Language-guided Robots with Vision



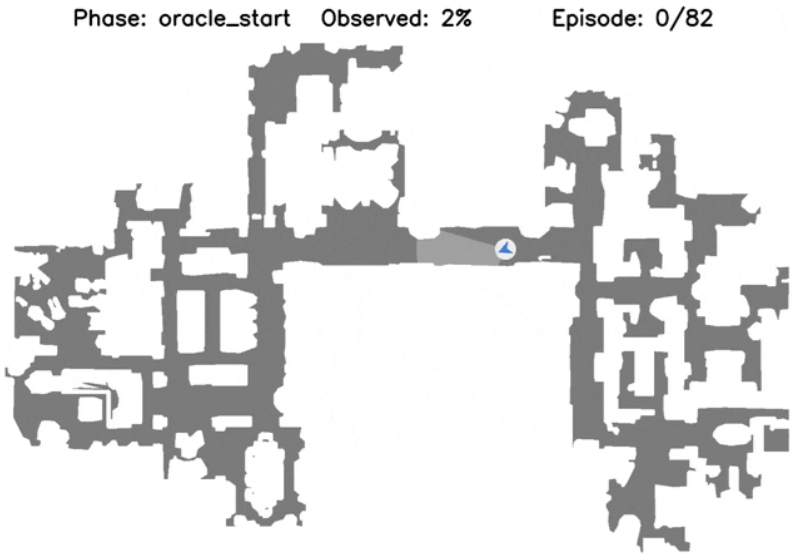
- We are exploring three directions in the space of translating language instructions to agent movement plans:
  - *Memory*: following instructions and building maps together.
  - *Physical Control*: overcoming the gap between simulation and real robot control.
  - *Planning*: translating underspecified language into coherent plans to get a job done.

Memory: Iterative Vision-and-Language Navigation [*in submission*]



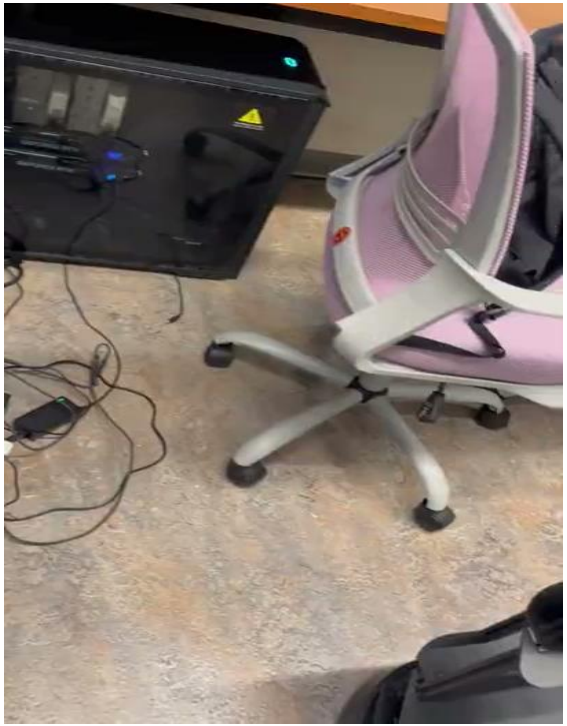
https://jacobkrantz.github.io/ivln

Memory: Iterative Vision-and-Language Navigation



https://jacobkrantz.github.io/ivln

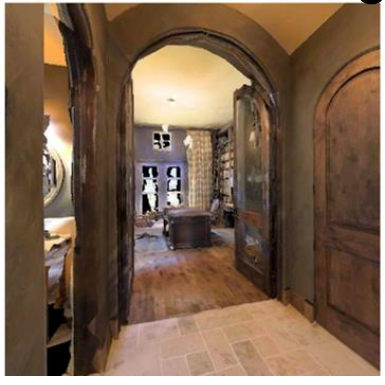
Physical Control: VLN With A Physical Robot



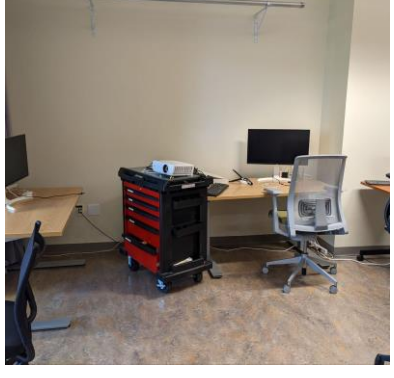
“Go to the couch and stop.”

Physical Control: VLN With A Physical Robot

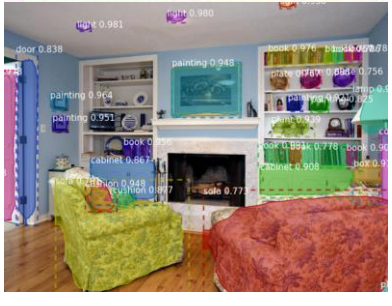
Simulation Training



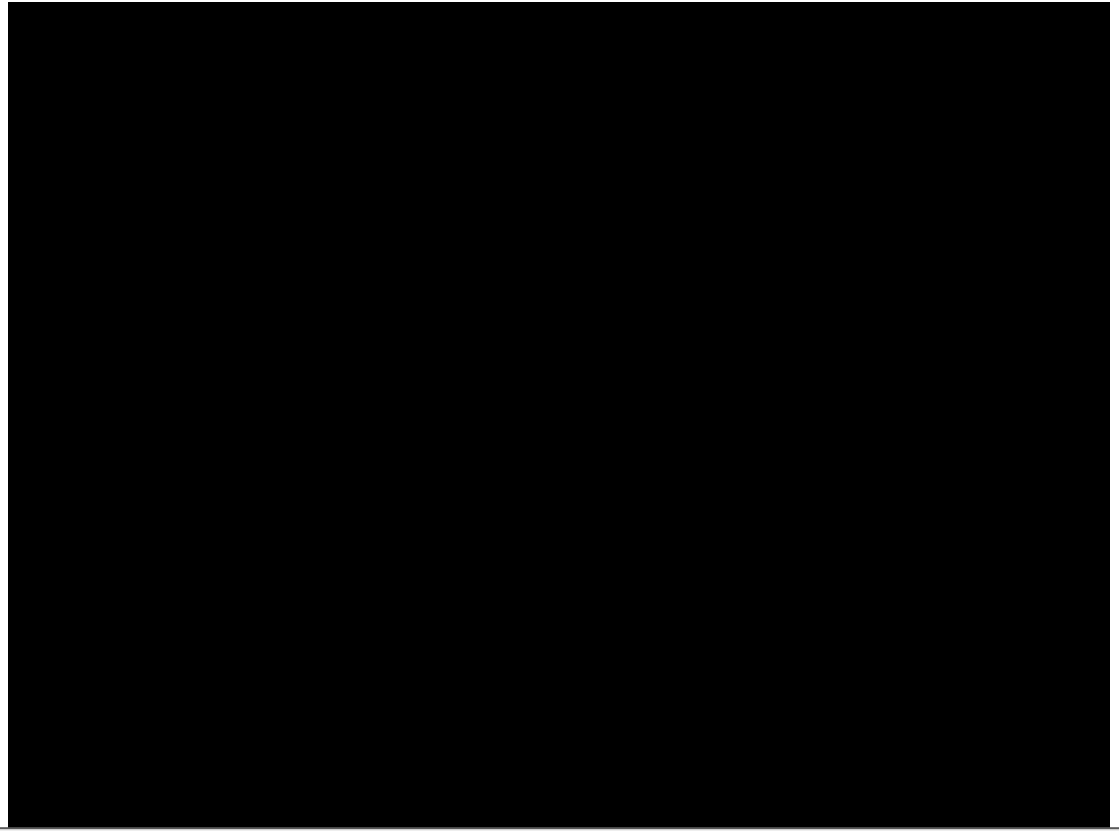
My Lab Space



- Use *mid-level* vision to align virtual and real world spaces



Planning: Prompting for Robot Control [*in submission*]



- LLMs generate actions robots can't do with objects that aren't around
- Pythonic *prompts* can specify robot actions and world objects

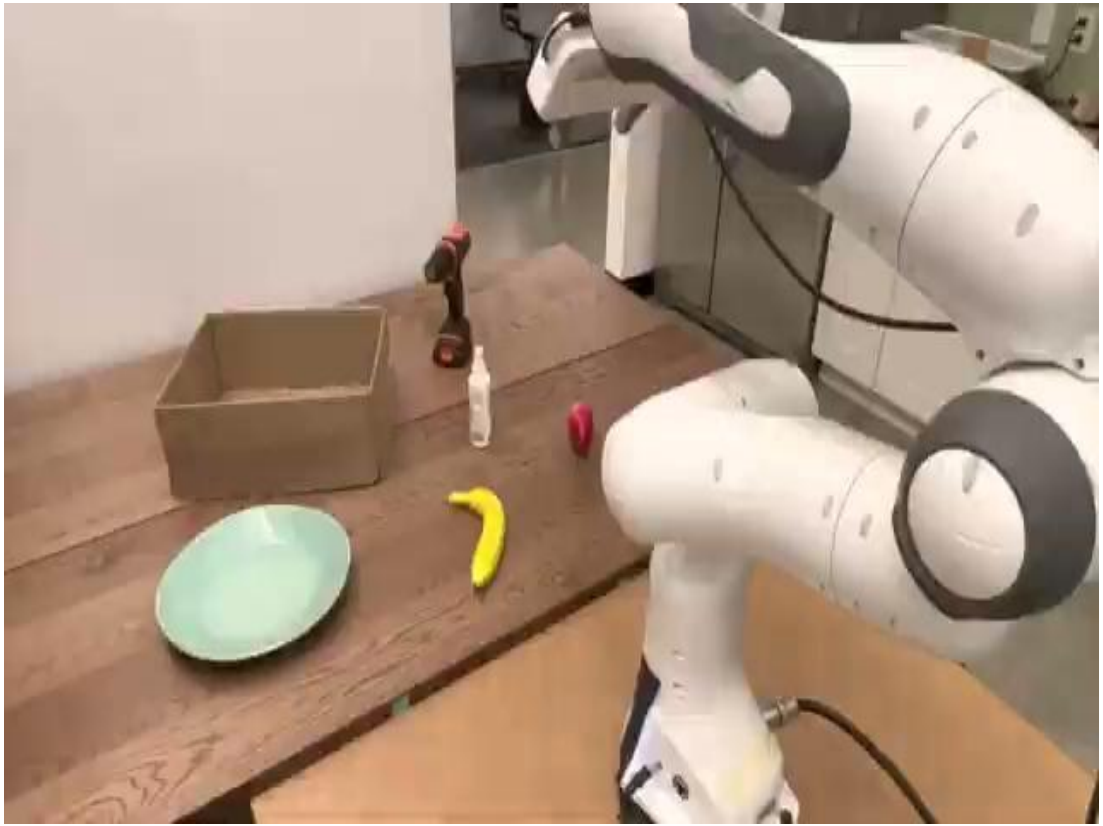
https://progprompt.github.io/

Planning: Prompting for Robot Control

- Used to control an agent in a virtual environment
- Perform new tasks in a zero-shot setting

https://progprompt.github.io/

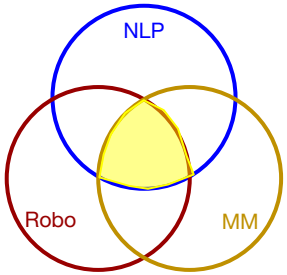
Planning: Prompting for Robot Control



- Also enables generating pick-and-place robot plans!

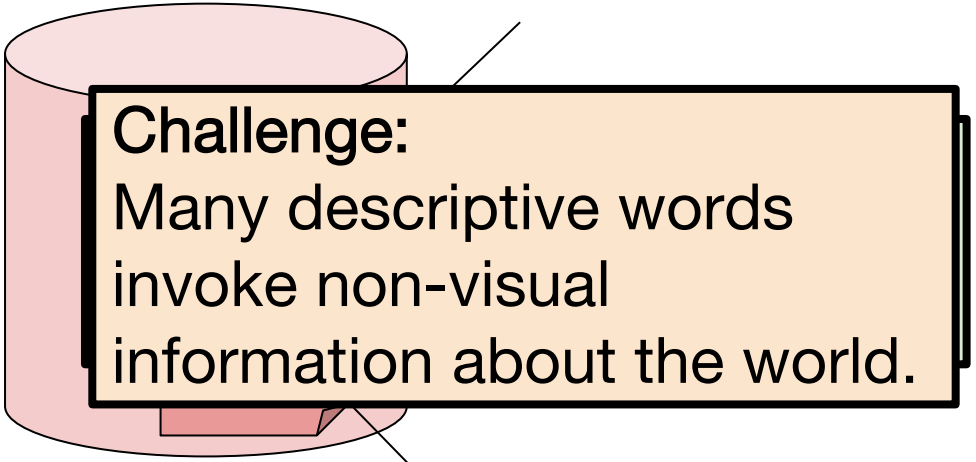
https://progprompt.github.io/

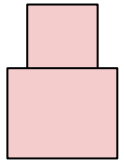
Language-guided Embodied, Multimodal Agents



- Sometimes robot *vision* is not enough.
- A robot is more than a mounted camera; it exists in the physical world and enables *physical sensory experience*
- We can combine language instructions and dialogue for learning with machine vision and perception for robots that learn from people and all their senses

Language-guided Embodied, Multimodal Agents





Physical Robot Agent

Sensory Perception Beyond Vision

Look

color, ***shape***, and deep ***visual*** features extracted from images of the object.

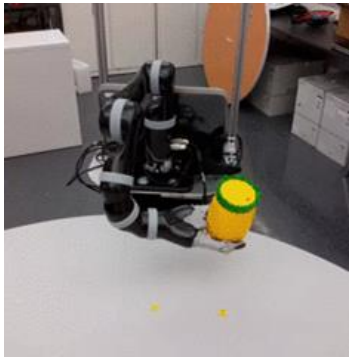
Push

Press

Grasp

Lift / Lower

Drop

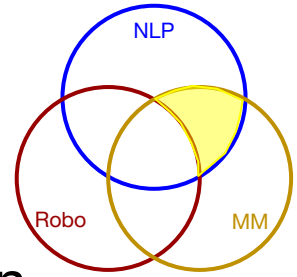


haptic and ***audio*** features captured from arm and microphone.

Grounding Multimodal Sensory Perception with Dialogue

Jointly Improving Parsing and Perception for Natural Language Commands through Human-Robot Dialog

Multimodal NLP: Other Times Text is not Enough



- We are exploring the integration of language-and-vision representations with other sensory and linguistic signals for:
  - *Continual Learning*: creating systems that learn from multiple tasks that involve combinations of language and vision input [*to appear* @ Neurips'22!]
  - *Sign Recognition*: using linguistic priors to better identify ASL signs from videos [*in submission*]
  - *Dementia Detection*: using both language transcriptions and speech signals to detect disfluencies indicating early stage dementia.

Natural Language Processing and Robotics

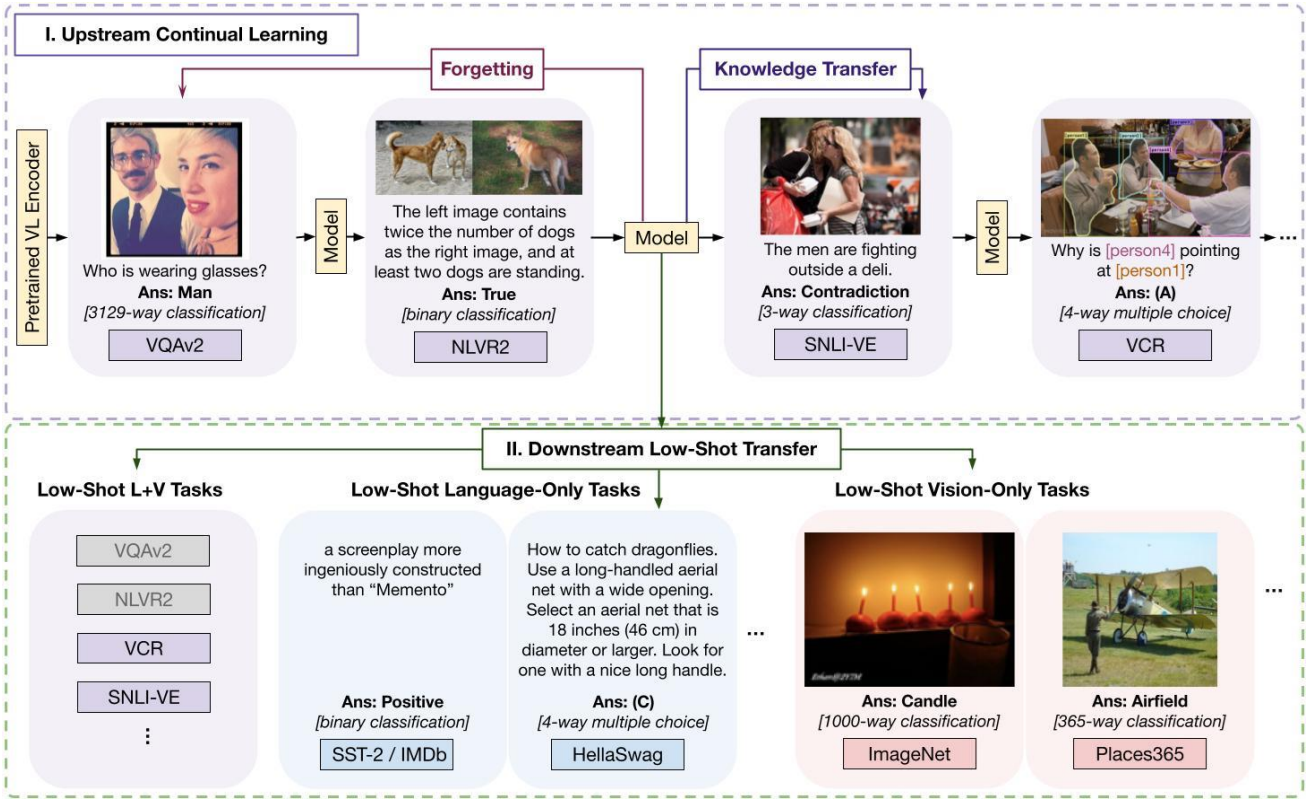
Jesse Thomason

[http://glamor.rocks/]

USC Viterbi School of Engineering

✨GLAMOR

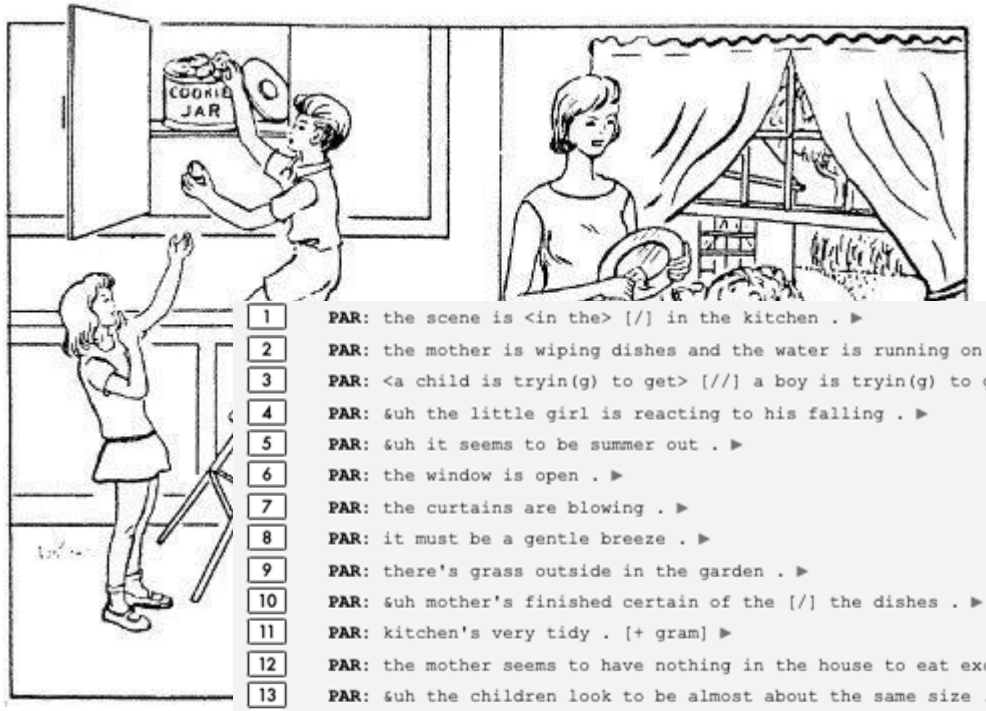
Multimodal Continual Learning



- Explore *adapter fusion* and *split adapters* to learn future tasks, even when a modality disappears

https://github.com/GLAMOR-USC/CLiMB

Early Detection of Dementia



1	**PAR:** the scene is <in the> [/] in the kitchen . ▶
2	**PAR:** the mother is wiping dishes and the water is running on the floor . ▶
3	**PAR:** <a child is tryin(g) to get> [//] a boy is tryin(g) to get cookies outta [: out of] a jar and he's about to tip over on a stool . ▶
4	**PAR:** &uh the little girl is reacting to his falling . ▶
5	**PAR:** &uh it seems to be summer out . ▶
6	**PAR:** the window is open . ▶
7	**PAR:** the curtains are blowing . ▶
8	**PAR:** it must be a gentle breeze . ▶
9	**PAR:** there's grass outside in the garden . ▶
10	**PAR:** &uh mother's finished certain of the [/] the dishes . ▶
11	**PAR:** kitchen's very tidy . [+ gram] ▶
12	**PAR:** the mother seems to have nothing in the house to eat except cookies in the cookie jar . ▶
13	**PAR:** &uh the children look to be almost about the same size . ▶
14	**PAR:** perhaps they're twins . ▶
15	**PAR:** they're dressed for summer warm weather . ▶

- Predict based on both text *and* audio signals

https://dementia.talkbank.org/

Isolated Sign Language Recognition

- Predict signed word from video frames
- Adding auxiliary predictions can help!
- Handshape
- Location

ASL-LEX