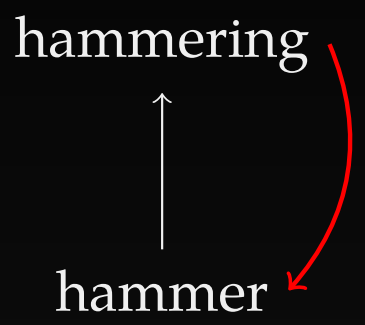
From language to vision and back again

Center for Brains,
Minds & Machines

Andrei Barbu

Torralba *et al* 2003


16 d 3-1/4" 8.255 cm
COATED SINKER
30 LBS. 13.60 KG


hammering
hammer

Perception is unreliable.

Top-down knowledge affects our perception.

One integrated representation for many tasks.

Recognition

Retrieval

Generation

Question answering

Disambiguation

Acquiring language

Images, not videos

Translation

Planning

Theory of mind

...

Humans perform language–vision tasks all the time

Give me the cup.

Which chair should I sit in?

This is an apple.

To win this game you have to make a straight line out of your pieces.

Recognition $S(\text{sentence}, \text{video})$

Retrieval

Generation

Question answering

Disambiguation

Acquiring language

Images, not videos

Translation

Planning

Theory of mind

...

The person rode the skateboard leftward.

object detector, tracker, event recognizer


Tracks
Events
Sentences

Object detection

Figure removed due to copyright restrictions. Please see the video.
Source: Barbu, Andrei, Aaron Michaux, Siddharth Narayanaswamy, and Jeffrey Mark Siskind. "Simultaneous object detection, tracking, and event recognition." Advances in Cognitive Systems: 203-220 (2012).

Felzenszwalb *et al* 2008

Object detection

Figure removed due to copyright restrictions. Please see the video.
Source: Barbu, Andrei, Aaron Michaux, Siddharth Narayanaswamy, and Jeffrey Mark Siskind. "Simultaneous object detection, tracking, and event recognition." Advances in Cognitive Systems: 203-220 (2012).

Felzenszwalb *et al* 2008

Object detection

Figure removed due to copyright restrictions. Please see the video.
Source: Barbu, Andrei, Aaron Michaux, Siddharth Narayanaswamy, and Jeffrey Mark Siskind. "Simultaneous object detection, tracking, and event recognition." Advances in Cognitive Systems: 203-220 (2012).

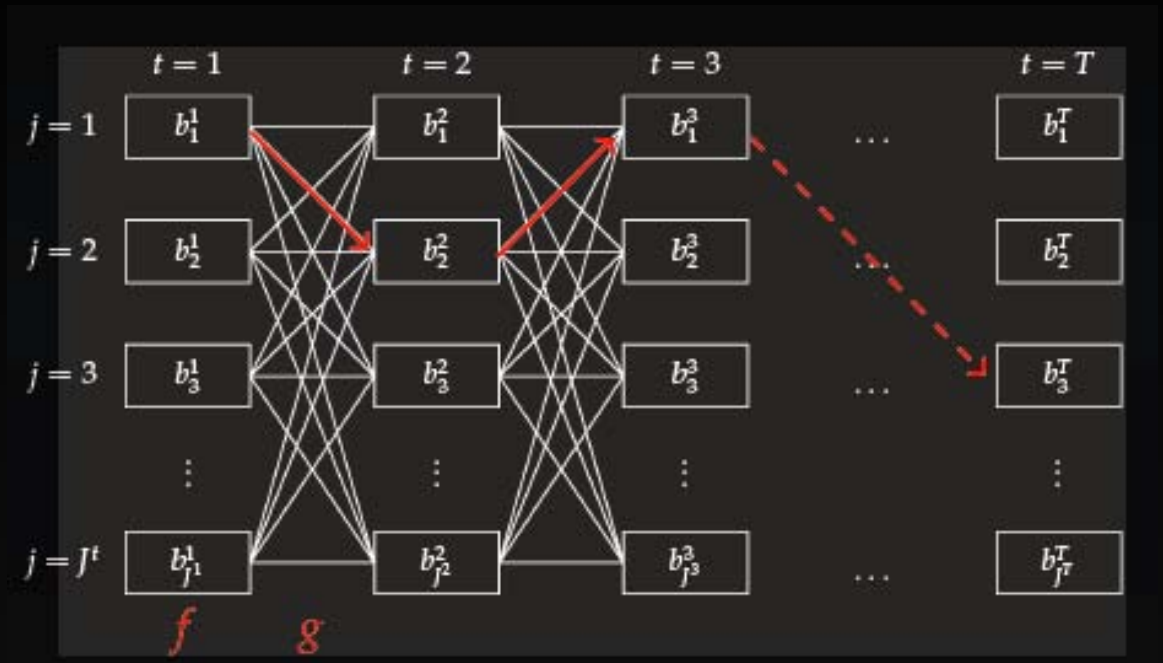
Felzenszwalb *et al* 2008

Object detectors work poorly

Figure removed due to copyright restrictions. Please see the video.
Source: Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115, no. 3 (2015): 211-252.

Russakovsky *et al* 2015

Fixing bad detectors with higher-level knowledge



detection / object / frame

temporally coherent track

- object detector confidence (f)
- motion coherence (g)

optimal path through the lattice of detections

dynamic programming

Bellman (1957), Viterbi (1967)

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t)$$

Barbu *et al* 2012

Fixing bad detectors with higher-level knowledge

Courtesy of Andrei Barbu, Alexander Bridge, Dan Coroian, Sven Dickinson, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Want, Jinlian Wei, Yifan Yin & Zhiqi Zhang. Used with permission.
Source: Barbu, Andrei, Alexander Bridge, Dan Coroian, Sven Dickinson, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi et al. "Large-scale automatic labeling of video events with verbs based on event-participant interaction." arXiv preprint arXiv:1204.3616 (2012).

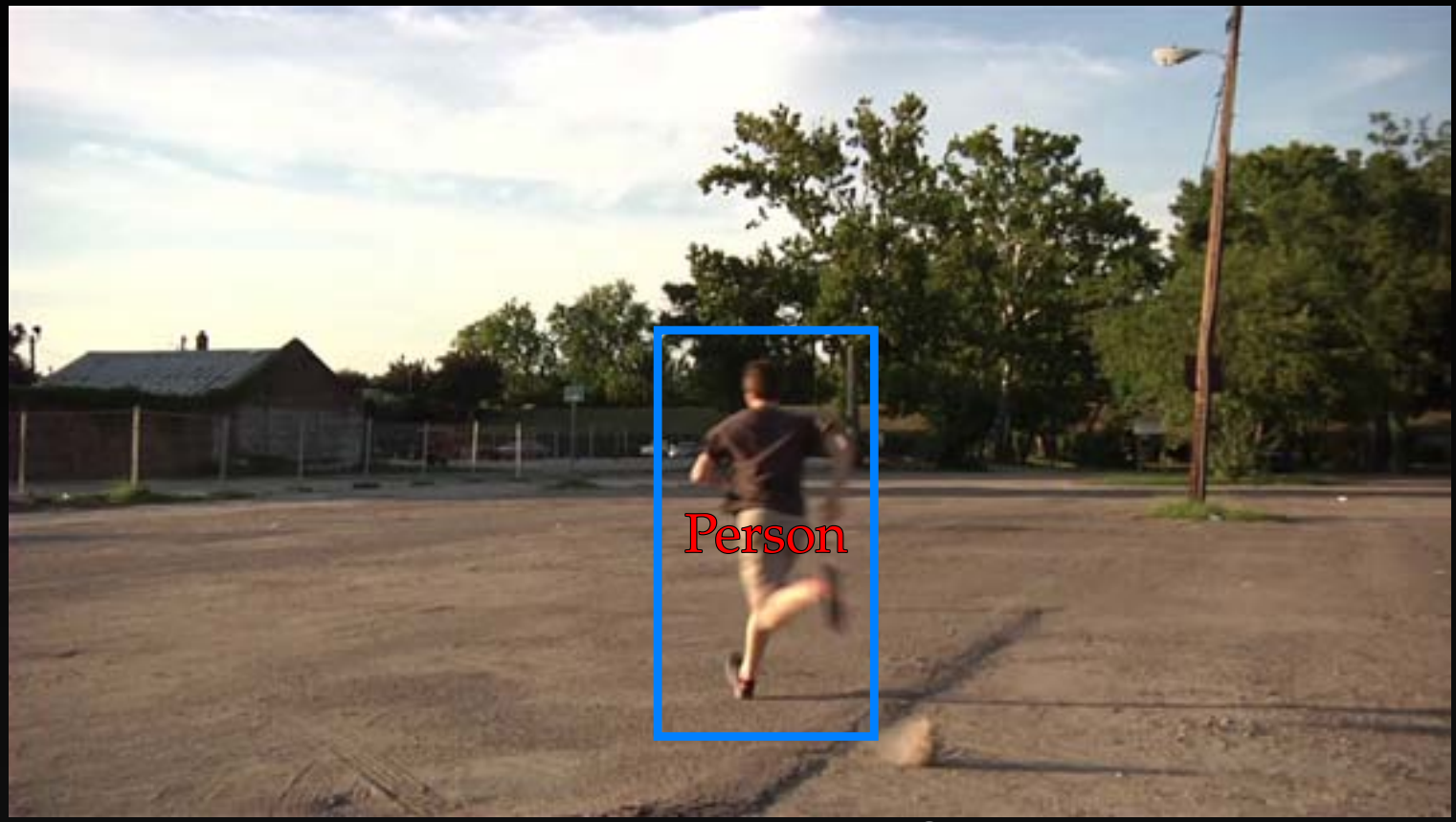
- detection / object / frame
- temporally coherent track
  - object detector confidence (f)
  - motion coherence (g)
- optimal path through the lattice of detections
- dynamic programming

Bellman (1957), Viterbi (1967)

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^t_{j^t})$$

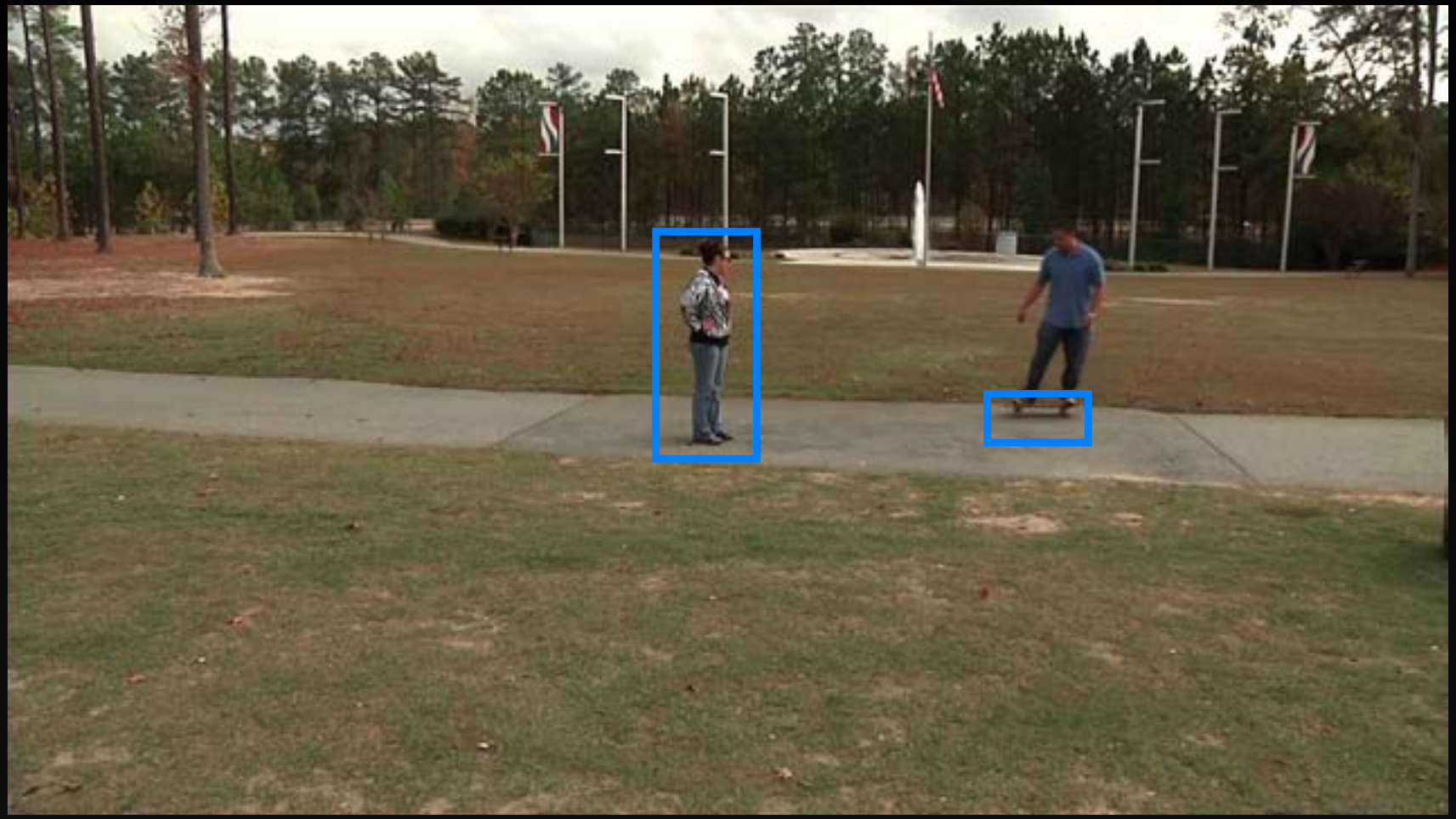
Barbu *et al* 2012

Feature vector — single participant



position	$\frac{\mathrm{d}}{\mathrm{d}t}$position	$\frac{\mathrm{d}^2}{\mathrm{d}t^2}$position	
aspect ratio	$\frac{\mathrm{d}}{\mathrm{d}t}$aspect ratio	area	$\frac{\mathrm{d}}{\mathrm{d}t}$area
object class	root-filter index		

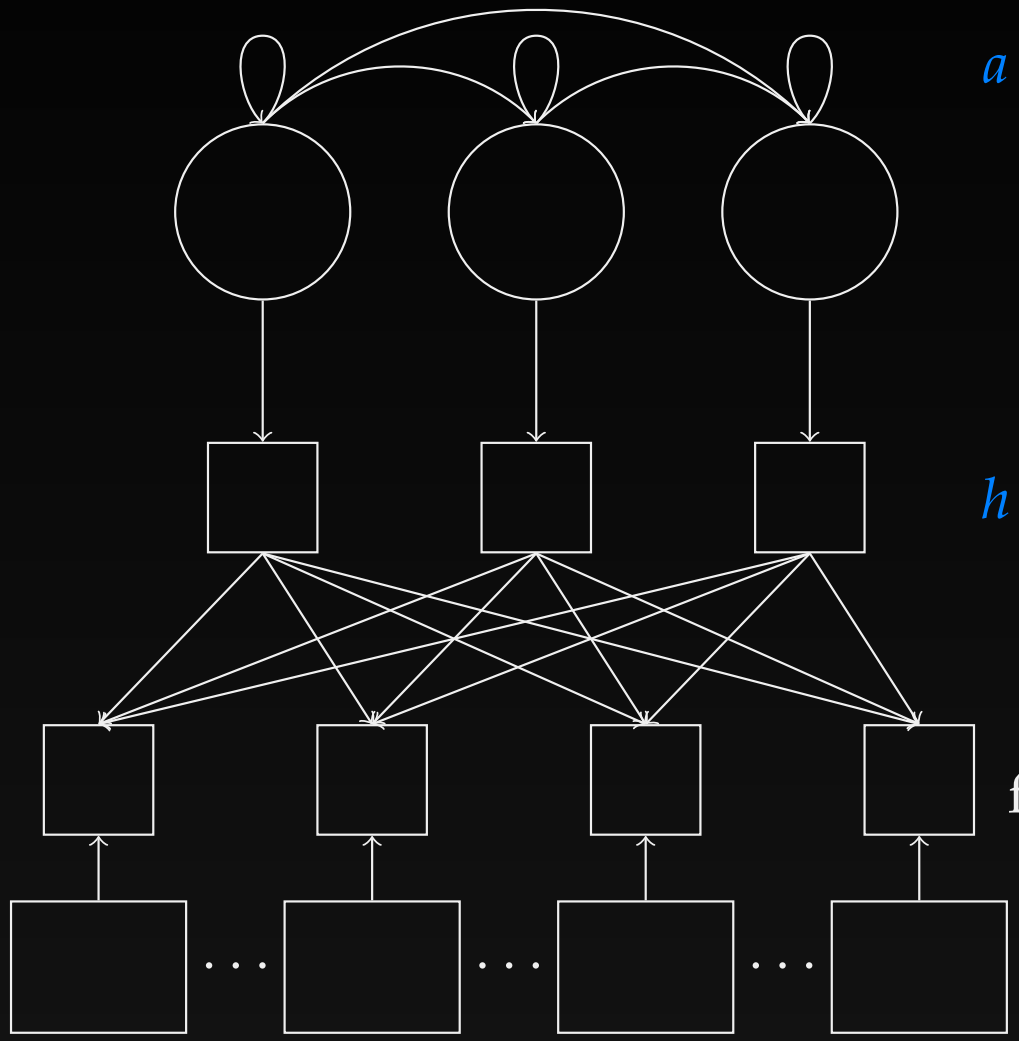
Feature vector — dual participant



$$\text{distance} \quad \frac{\mathrm{d}}{\mathrm{d}t}\text{distance} \quad \text{orientation} \quad \frac{\mathrm{d}}{\mathrm{d}t}\text{orientation}$$

person riding skateboard | person approaching person | skateboard approaching person

Event recognition



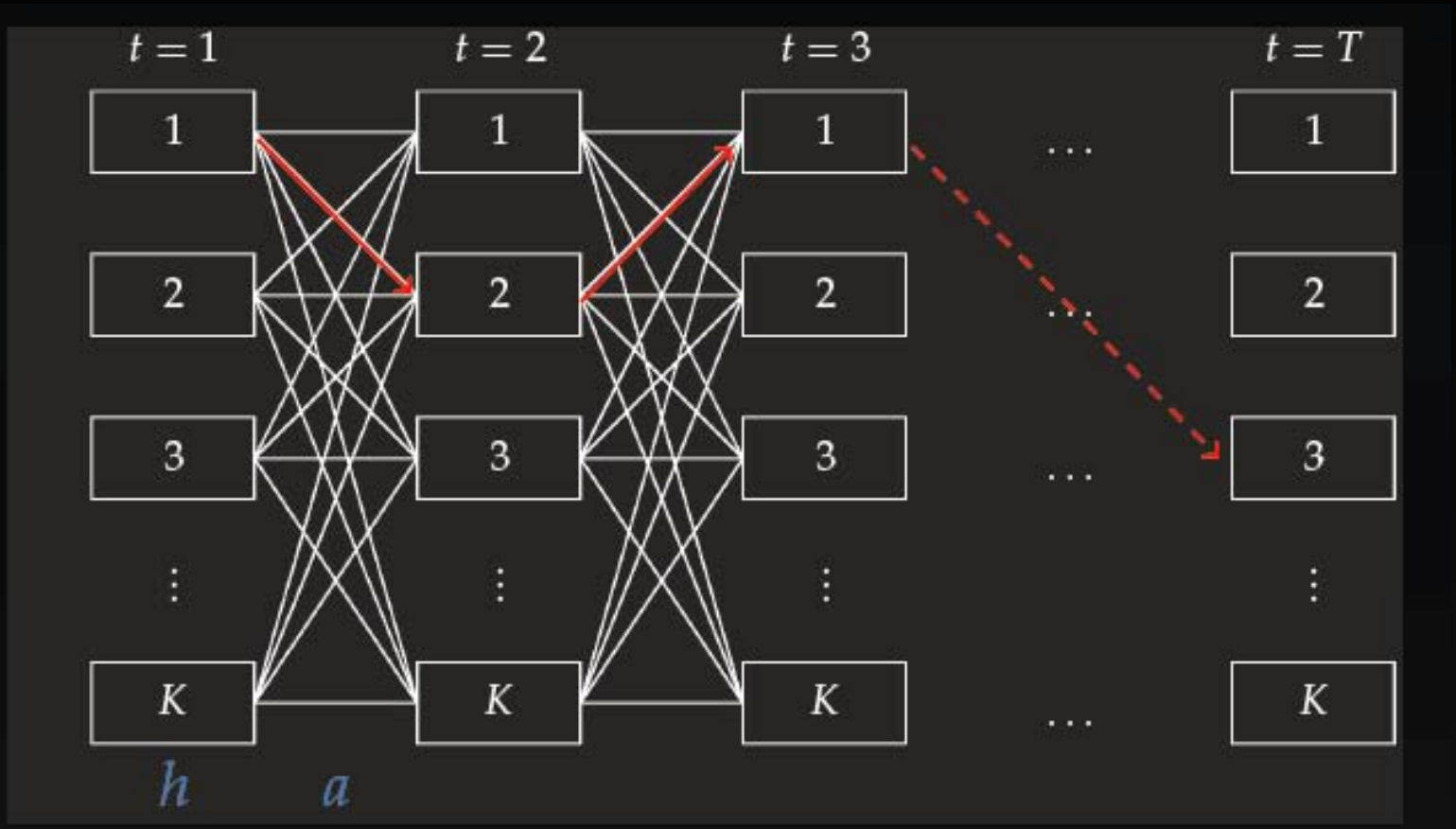
HMMs

Baum and Petrie (1966)

One HMM per event class

Try each HMM

Event recognition



$$\max_{k^1,\ldots,k^T} \sum_{t=1}^{T} h(k^t, b^t_{\hat{j}^t}) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Barbu *et al* 2012


Tracks
Events
Sentences

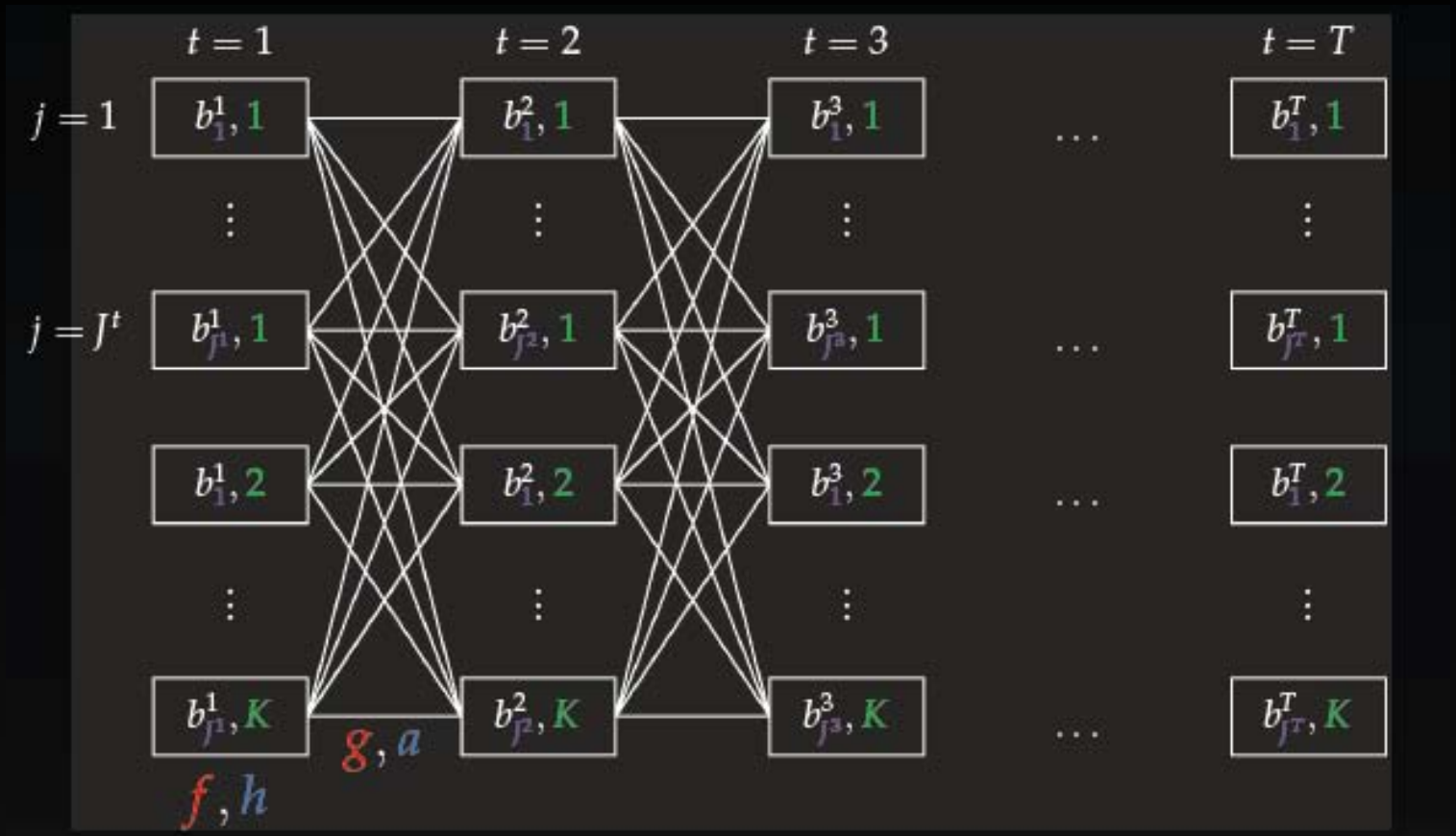
Tracking in the context of event recognition

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t) + \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} h(k^t, b_{\hat{j}^t}^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Tracking in the context of event recognition

$$\max_{j^1,\ldots,j^T} \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^t_{j^t}) + \sum_{t=1}^{T} h(k^t, b^t_{j^t}) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Tracking in the context of event recognition



$$\max_{j^1,\ldots,j^T} \max_{k^1,\ldots,k^T} \sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^t_{j^t}) + \sum_{t=1}^{T} h(k^t, b^t_{j^t}) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Tracking in the context of event recognition in action

tracking

tracking and event recognition

Courtesy of Andrei Barbu, Alexander Bridge, Dan Coroian, Sven Dickinson, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Want, Jinlian Wei, Yifan Yin & Zhiqi Zhang. Used with permission.
Source: Barbu, Andrei, Alexander Bridge, Dan Coroian, Sven Dickinson, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi et al. "Large-scale automatic labeling of video events with verbs based on event-participant interaction." arXiv preprint arXiv:1204.3616 (2012).

Tracking in the context of event recognition in action

tracking

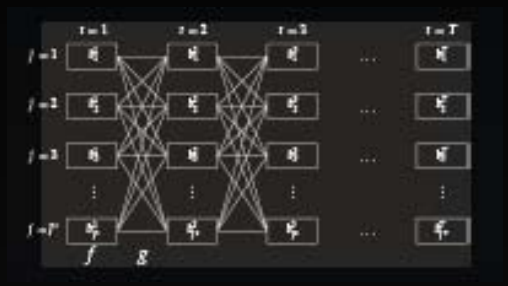
tracking and event recognition


Tracks
Events
Sentences

Building sentences out of trackers and words

Viterbi tracker

track 1



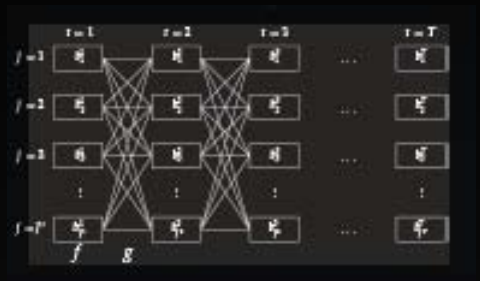
$$\max_{j_1^1,\ldots,j_1^T} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t)$$

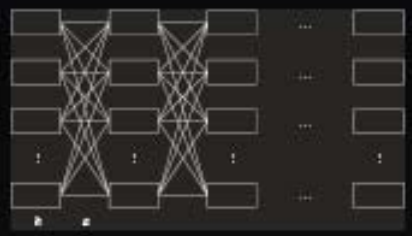
Siddharth *et al* 2014

Building sentences out of trackers and words

Event tracker for intransitive verbs

track 1





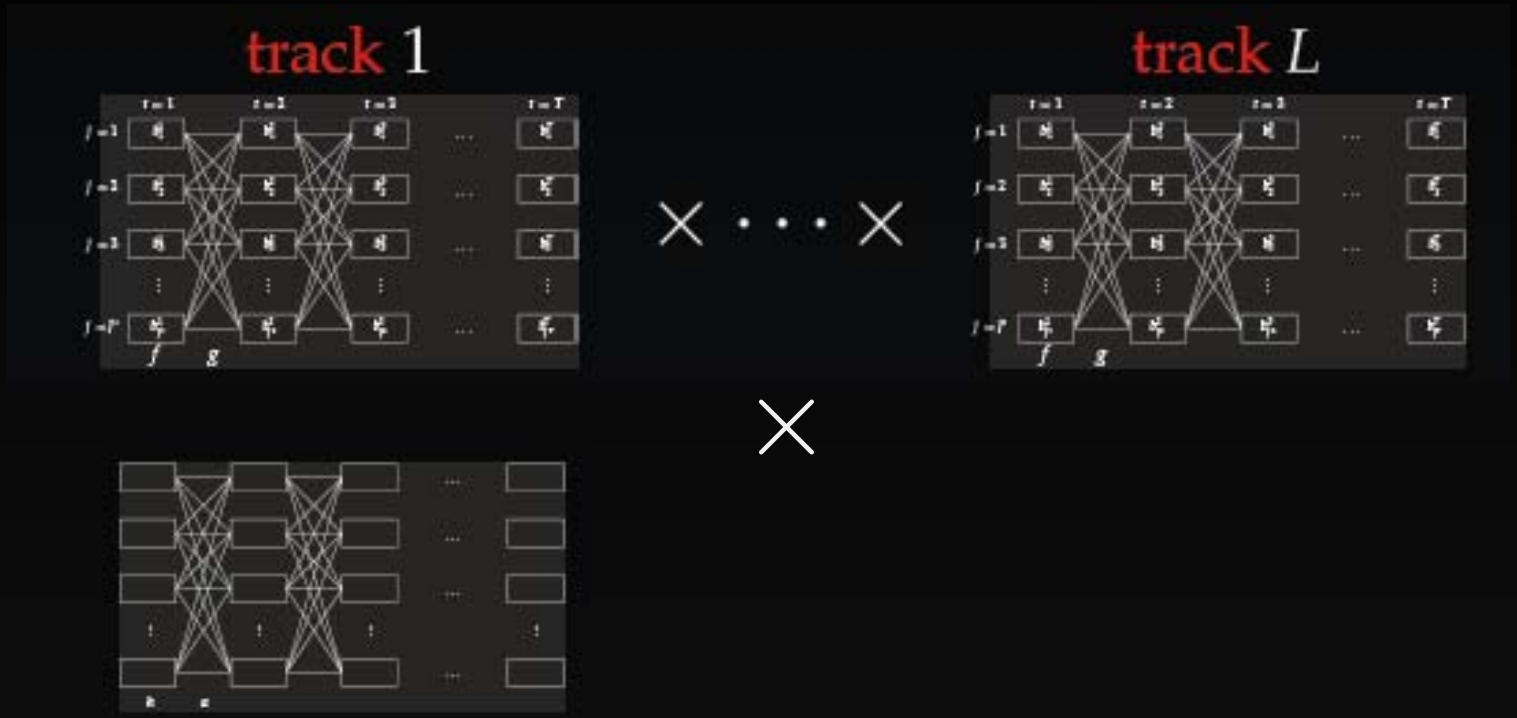
word 1

$$\max_{j_1^1,\ldots,j_1^T} \max_{k_1^1,\ldots,k_1^T} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \sum_{t=1}^{T} h(k^t, b_{j^t}^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Siddharth *et al* 2014

Building sentences out of trackers and words

Event tracker

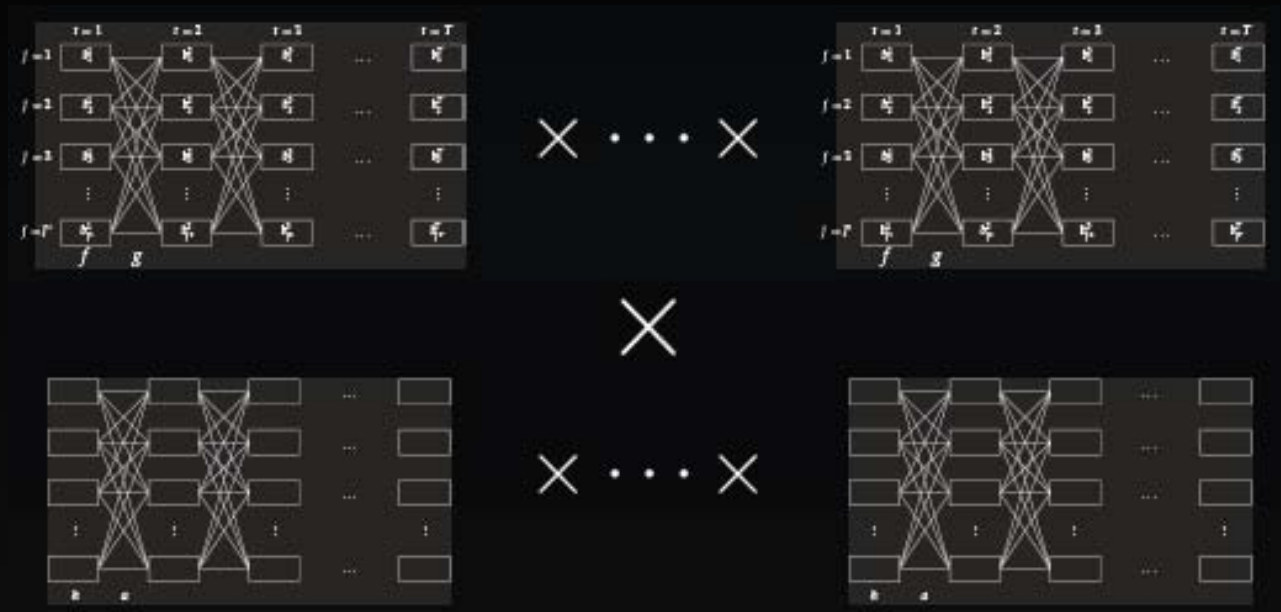


$$\max_{\substack{j_1^1,\ldots,j_1^T \\ \vdots \\ j_L^1,\ldots,j_L^T}} \max_{k_1^1,\ldots,k_1^T} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \sum_{t=1}^{T} h(k^t, b_{j_{\theta^1}^t}^t, b_{j_{\theta^2}^t}^t) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Siddharth *et al* 2014

Building sentences out of trackers and words

Sentence tracker

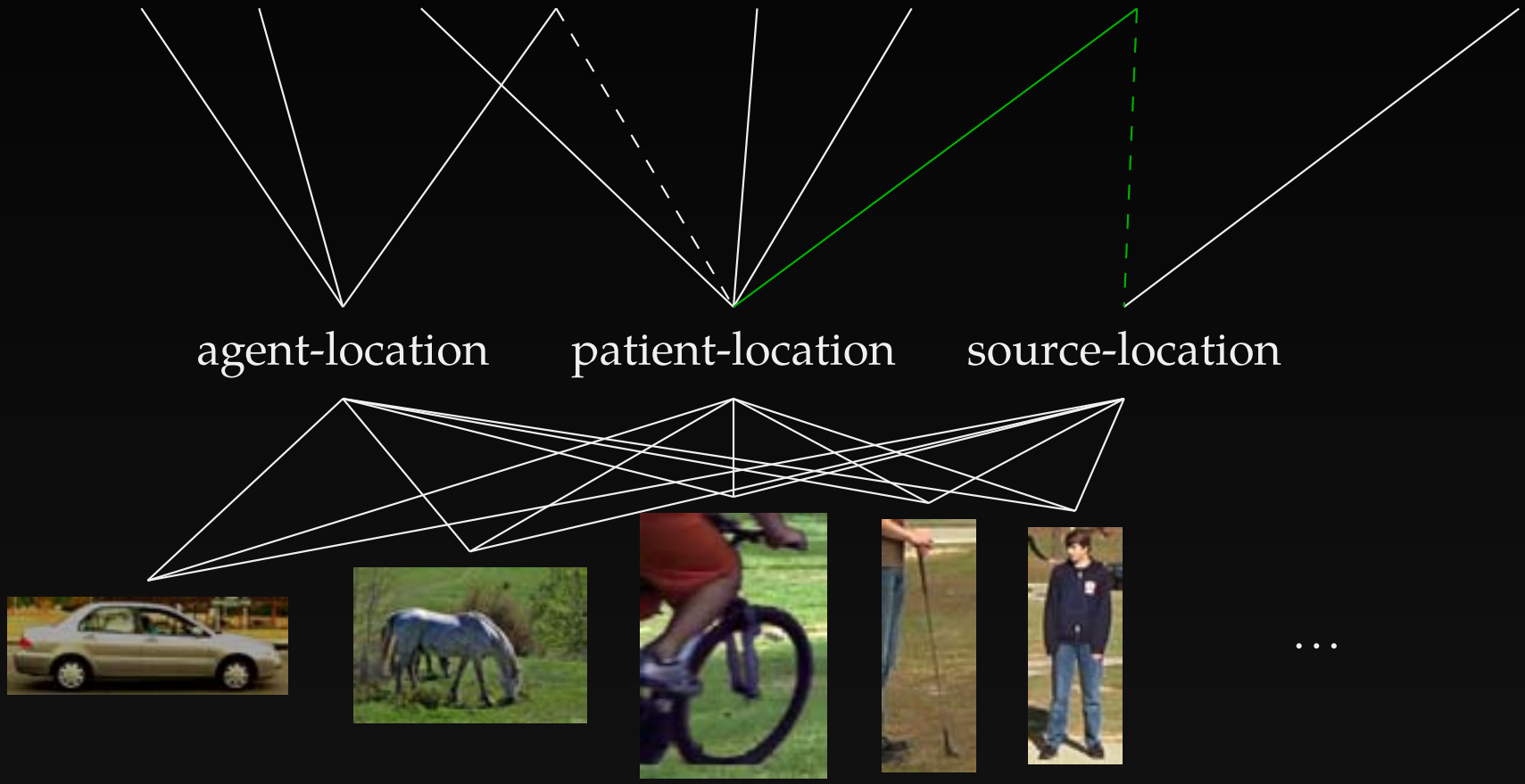


$$\max_{\substack{j_1^1,\ldots,j_1^T \\ \vdots \\ j_L^1,\ldots,j_L^T}} \max_{\substack{k_1^1,\ldots,k_1^T \\ \vdots \\ k_W^1,\ldots,k_W^T}} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b^t_{j_l^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j_l^{t-1}}, b^t_{j_l^t}) + \sum_{w=1}^{W}\sum_{t=1}^{T} h_w(k_w^t, b^t_{j_{\theta_w^1}^t}, b^t_{j_{\theta_w^2}^t}) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t)$$

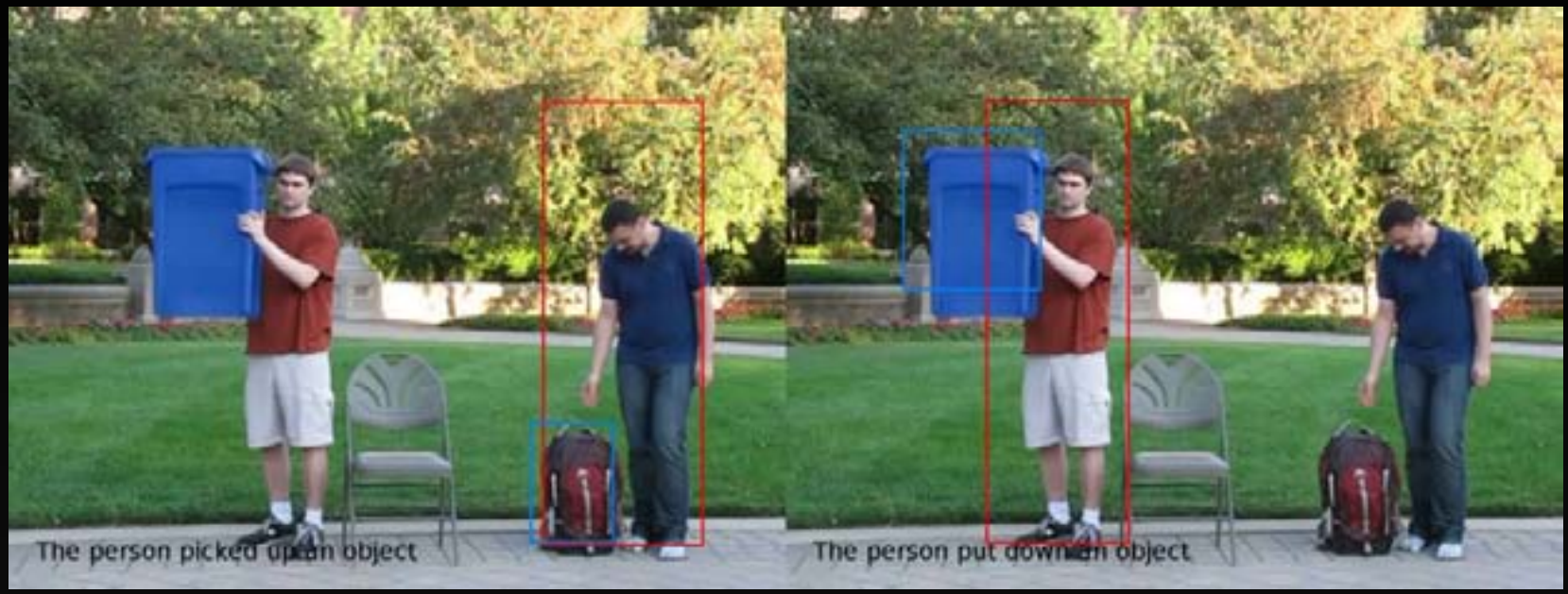
Siddharth *et al* 2014

Sentences

The tall person quickly rode the horse leftward away from the other horse.



Understanding sentences as a whole



Source: Yu, Haonan, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. "A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video." J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.

We can differentiate events based on the verb:

The person picked up an object.
The person put down an object.

Siddharth *et al* 2014

Understanding sentences as a whole

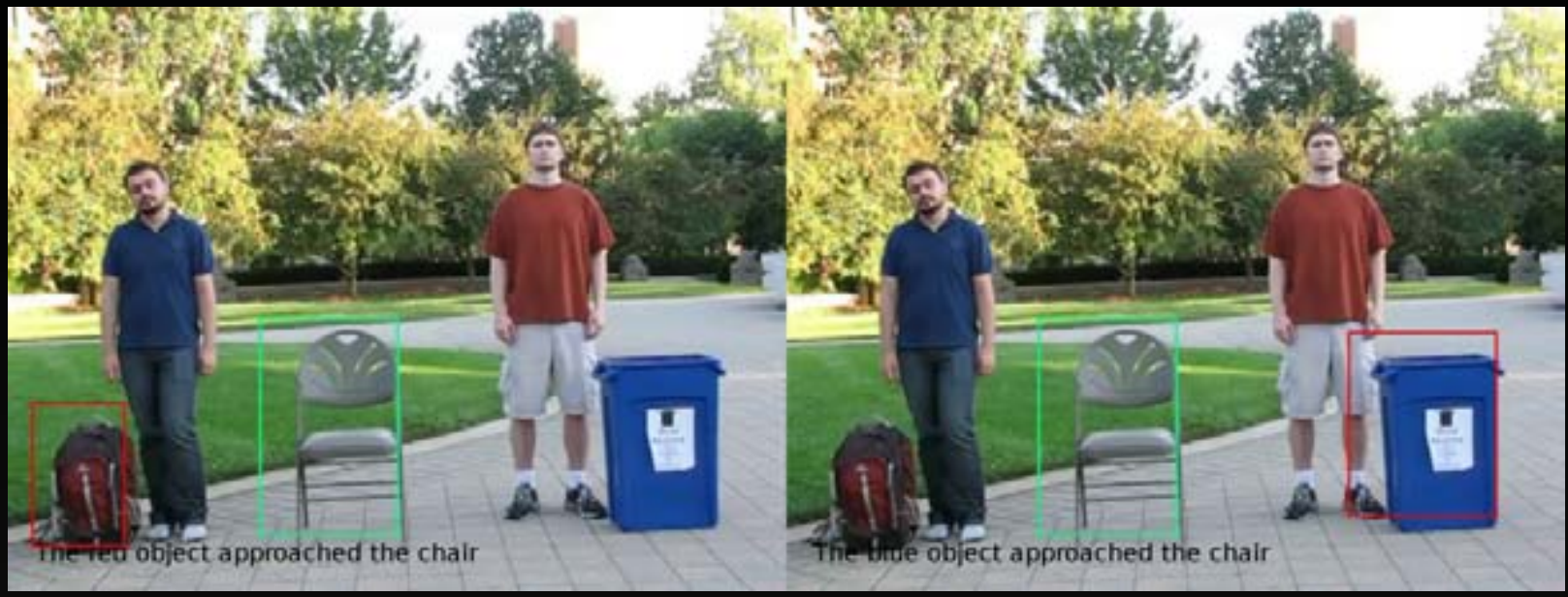
Source: Yu, Haonan, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. "A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video." J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.

We can differentiate events based on the subject noun:

The backpack approached the bin.
The chair approached the bin.

Siddharth *et al* 2014

Understanding sentences as a whole



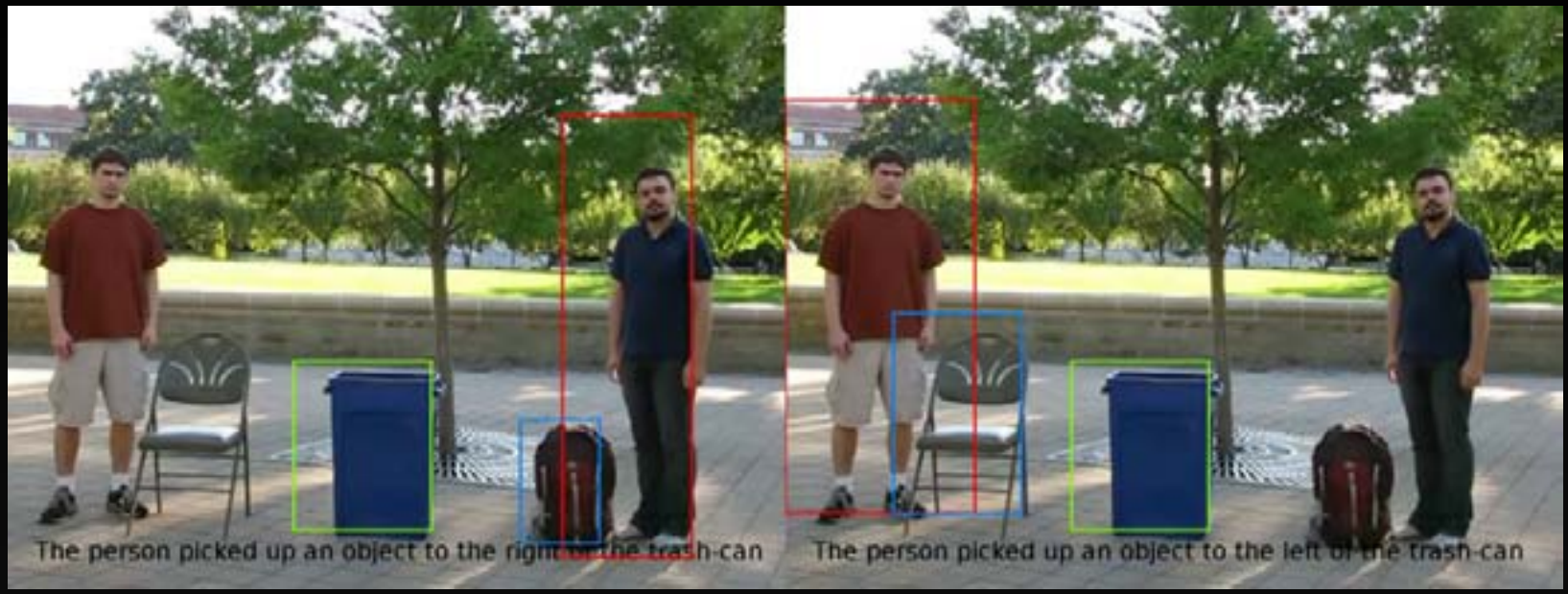
Source: Yu, Haonan, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. "A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video."
J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.

We can differentiate events based on an adjective in the subject NP:

The red object approached the chair.
The blue object approached the chair.

Siddharth *et al* 2014

Understanding sentences as a whole



Source: Yu, Haonan, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. "A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video."
J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.

We can differentiate events based on a preposition in the object NP:

The person picked up an object to the right of the bin.
The person picked up an object to the left of the bin.

Siddharth *et al* 2014

Recognition $S(\text{sentence}, \text{video})$

Retrieval $\underset{v \in V}{\operatorname{argmax}}\, S(s, v)$

Generation

Question answering

Disambiguation

Acquiring language

Images, not videos

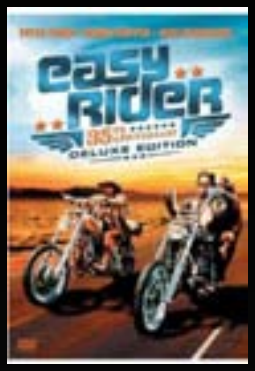
Translation

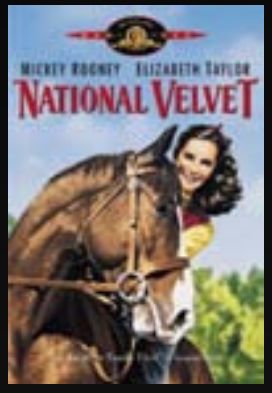
Planning

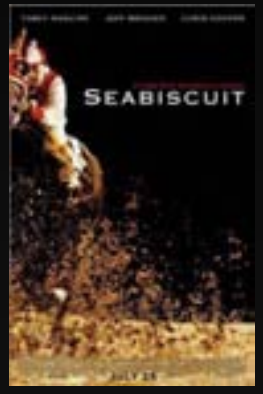
Theory of mind

...

Sentential retrieval







Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	
Disambiguation	
Acquiring language	
Images, not videos	
Translation	
Planning	
Theory of mind	
...	

Generating sentences

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D [A] N [PP]} \\
\text{D} &\rightarrow \textit{an} \mid \textit{the} \\
\text{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{VP} &\rightarrow \text{V NP [Adv] [PP}_\text{M}\text{]} \\
\text{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\text{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PP}_\text{M} &\rightarrow \text{P}_\text{M}\ \text{NP} \\
\text{P}_\text{M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

"the person carried the backpack"

Generated sentences

The person to the right of the bin picked up the backpack.

Source: Yu, Haonan, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. "A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video." J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S'(s, s_q, v)$
Disambiguation	
Acquiring language	
Images, not videos	
Translation	
Planning	
Theory of mind	
...	

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	
Acquiring language	
Images, not videos	
Translation	
Planning	
Theory of mind	
...	

Question answering

What did the person put on top of the red car?
The person put NP on top of the red car.
The person put the pear on top of the red car.

Question answering

Who put an object on top of the red car?

Generation for question answering

Who put an object on top of the red car?

NP put an object on top of the red car.

. . .

The person put an object on top of the red car.

Generation for question answering

Who put an object on top of the red car?

NP put an object on top of the red car.

. . .

~~The person put an object on top of the red car.~~

Discriminative generation for question answering

Who put an object on top of the red car?

NP put an object on top of the red car.

. . .

~~The person put an object on top of the red car.~~

. . .

The person on the left of the car put an object on top of the red car.

Question answering

Who put an object on top of the red car?

NP put an object on top of the red car.

The person on the left of the car put an object on top of the red car.

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	
Images, not videos	
Translation	
Planning	
Theory of mind	
...	

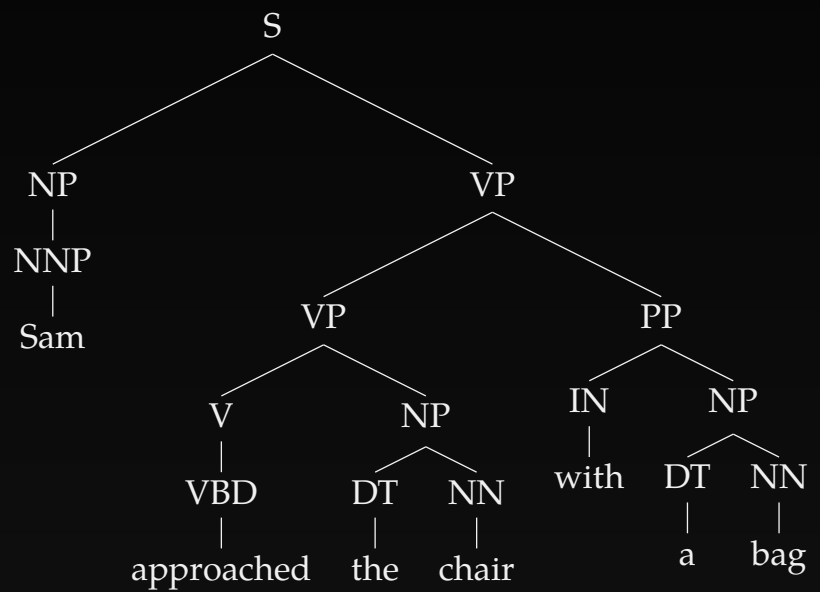
Disambiguation

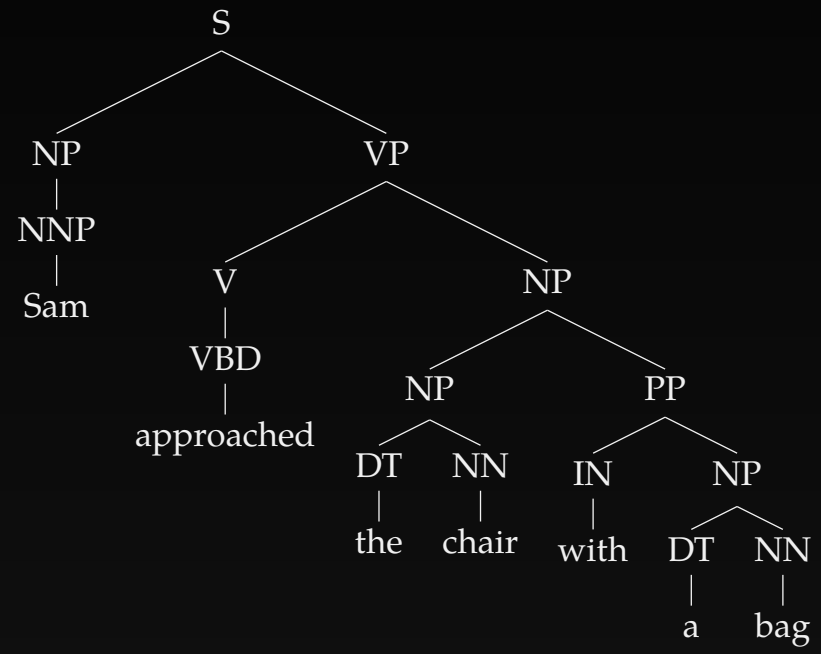
Danny approached the chair with a bag.

Berzak *et al* 2015

Disambiguation

Danny approached the chair with a bag.





Berzak *et al* 2015

Disambiguation

PP Attachment

Danny looked at Andrei with a telescope.

Berzak *et al* 2015

Disambiguation

PP Attachment

VP Attachment

Andrei approached the person holding a green chair.

Berzak *et al* 2015

Disambiguation

PP Attachment

VP Attachment

Conjunction

Danny and Andrei picked up the yellow bag and chair.

Source: Berzak, Yevgeni, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. "Do you see what i mean? visual resolution of linguistic ambiguities." arXiv preprint arXiv:1603.08079 (2016).

Berzak *et al* 2015

Disambiguation

PP Attachment

VP Attachment

Conjunction

Logical Form

Someone put down the bags.

Berzak *et al* 2015

Disambiguation

PP Attachment

VP Attachment

Conjunction

Logical Form

Anaphora

Danny picked up the bag and the chair. It is yellow.

Berzak *et al* 2015

Disambiguation

- PP Attachment
- VP Attachment
- Conjunction
- Logical Form
- Anaphora
- Ellipsis

Danny left Andrei. Also Yevgeni.

Source: Berzak, Yevgeni, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. "Do you see what i mean? visual resolution of linguistic ambiguities." arXiv preprint arXiv:1603.08079 (2016).

Berzak *et al* 2015

Not just parse trees

Danny and Andrei moved a chair.

Danny and Andrei move the same chair.

$$\textbf{chair}(x)$$
$$\textbf{move}(\text{Danny}, x), \textbf{move}(\text{Andrei}, x)$$

Danny and Andrei move different chairs.

$$\textbf{chair}(x), \textbf{chair}(y)$$
$$\textbf{move}(\text{Danny}, x), \textbf{move}(\text{Andrei}, y), x \neq y$$

Not just parse trees

Danny and Andrei moved a chair.

Danny and Andrei move the same chair.

$$\textbf{chair}(x), \textbf{person}(y), \textbf{person}(z), y \neq z$$
$$\textbf{move}(y, x), \textbf{move}(z, x)$$

Danny and Andrei move different chairs.

$$\textbf{chair}(x), \textbf{chair}(y), \textbf{person}(z), \textbf{person}(w), z \neq w$$
$$\textbf{move}(\mathrm{z}, x), \textbf{move}(\mathrm{w}, y), x \neq y$$

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	
Translation	
Planning	
Theory of mind	
...	

Language learning

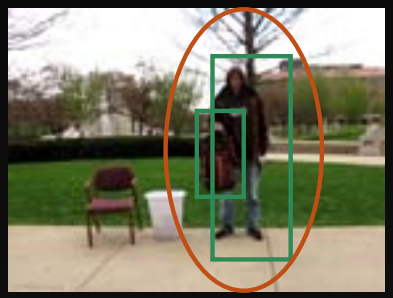
Split into two tasks:

- Learning word meaning
- Learning syntax

Language learning: word meaning

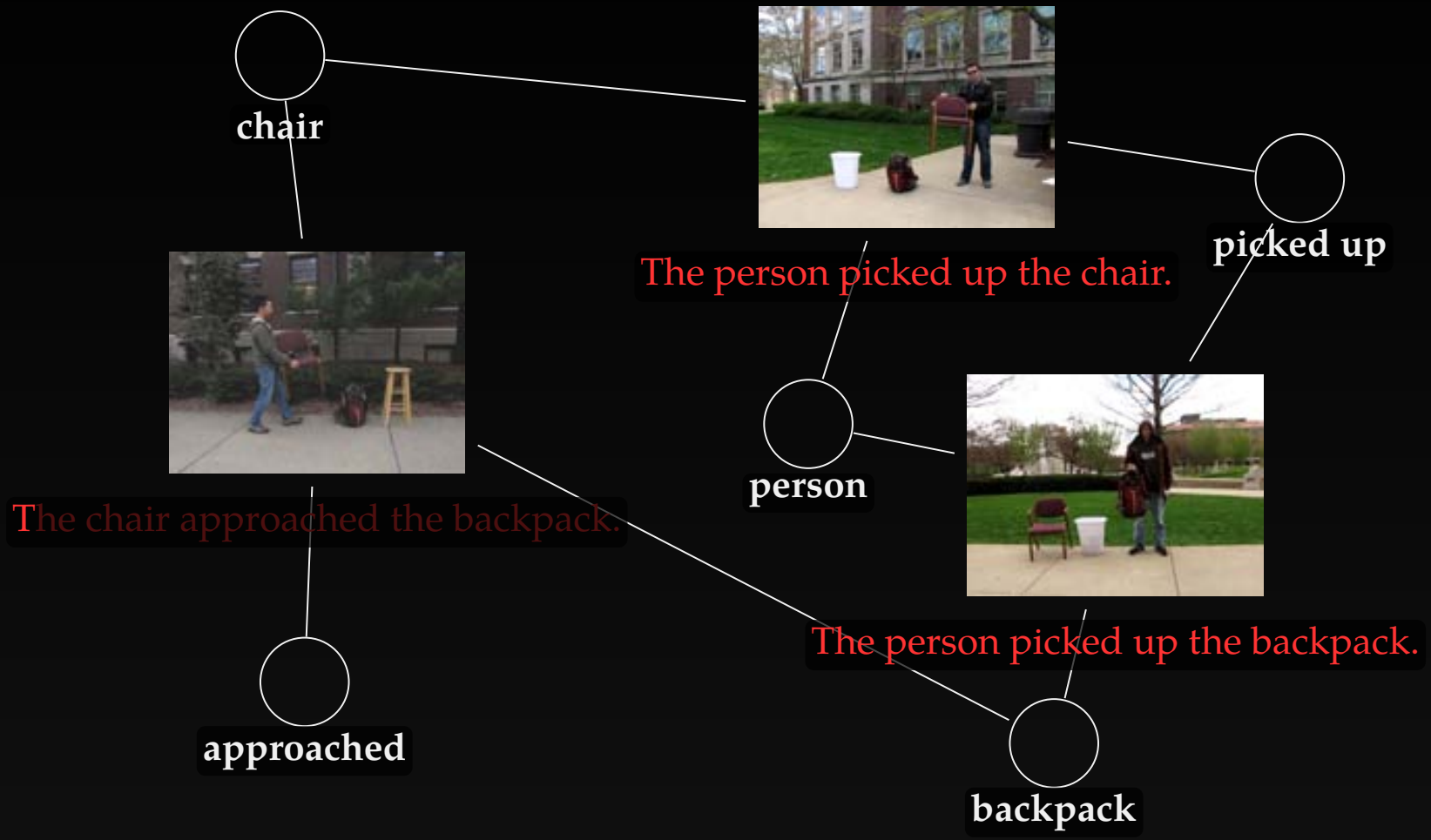
The person picked up the chair.

The chair approached the backpack.



The person picked up the backpack.

Language learning: word meaning



Source: Yu, Haonan, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. "A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video." J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.

Yu *et al* 2014

Language learning

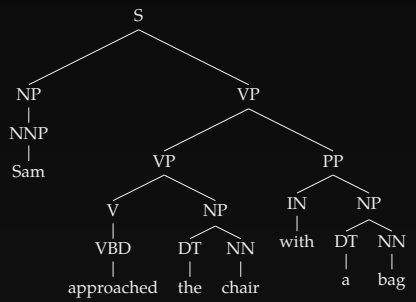
Split into two tasks:

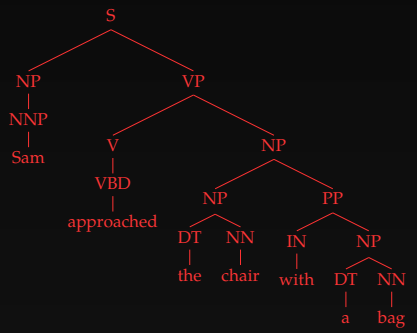
- Learning word meaning
- Learning syntax

Language learning: syntax; in progress

Danny approached the chair with a bag.

parser

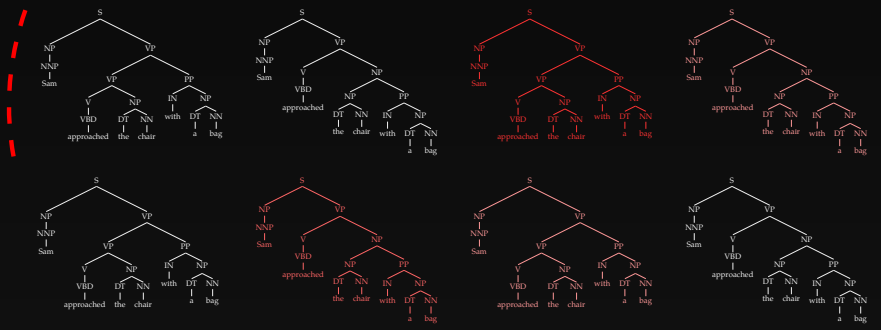




Language learning: syntax; in progress

Danny approached the chair with a bag.

≈parser



• • •

Pilley and Reid 2011

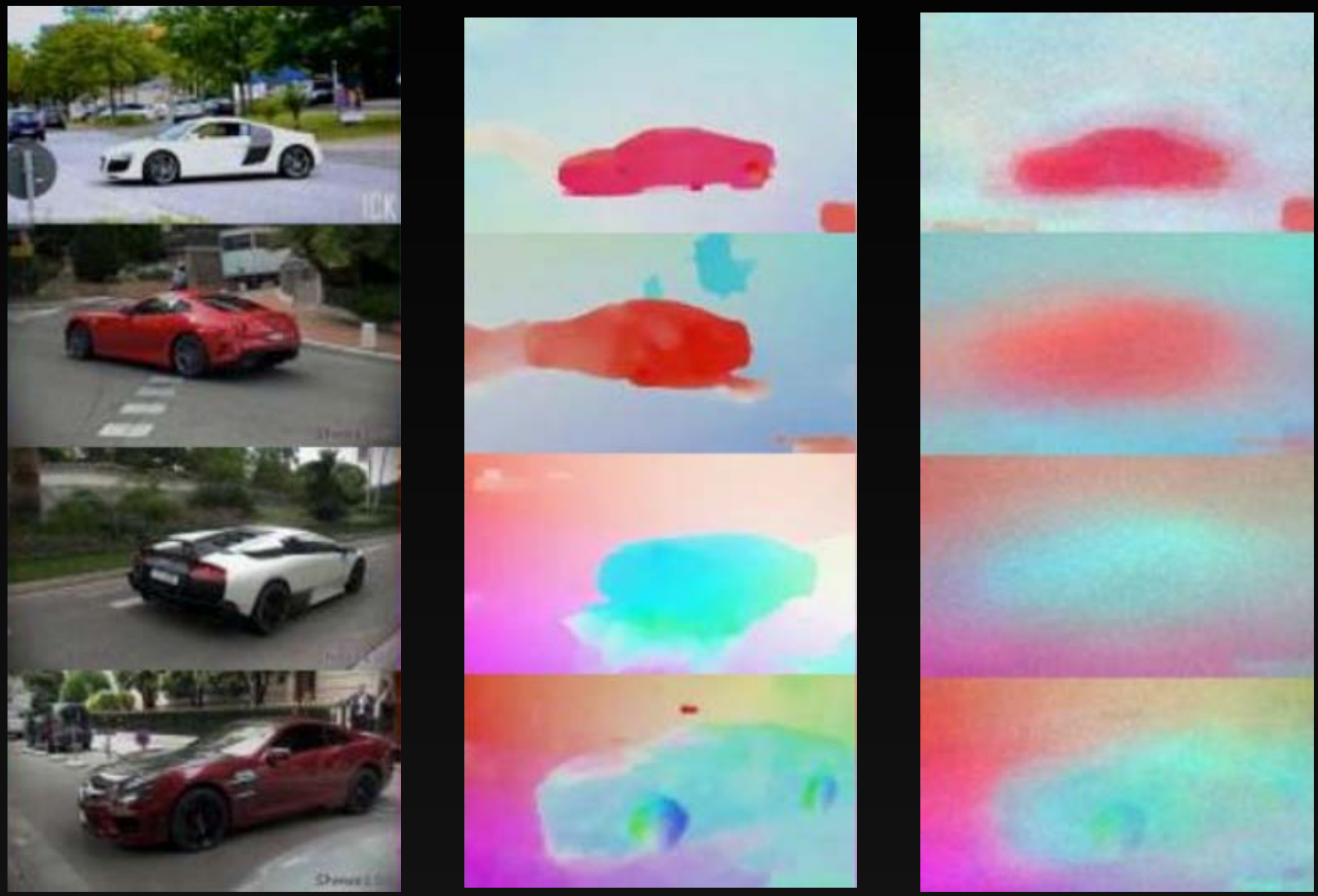
Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{video})$
Translation	
Planning	
Theory of mind	
...	

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	
Planning	
Theory of mind	
...	

Single-frame optical flow

Single-frame optical flow

Input | Flow | Predicted



Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}}\, f(s, s')$
Planning	
Theory of mind	
...	

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}} \int_v \lvert S(s', v) - S(s, v) \rvert$
Planning	
Theory of mind	
...	

Statistical machine translation

Sam was happy

Sam a fost fericita

Сэм была счастлива

In Thai you specify siblings by age not gender.

In English you specify relative time but you don't need to in Chinese.

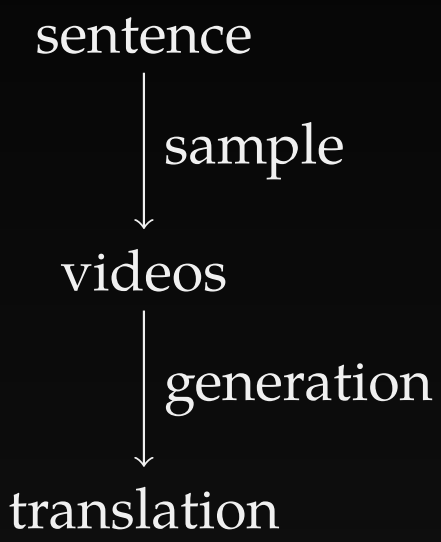
Guugu Yimithirr language only uses absolute directions.

Many languages don't distinguish blue/green.

Swahili specifies color as "the color of X".

In Turkish you have to report if something is hearsay.

Translation by imagination



Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}} \int_v \lvert S(s', v) - S(s, v) \rvert$
Planning	$\underset{s \in L}{\text{argmax}} \int_v S(s, v_0 : v : v_n)$
Theory of mind	
...	

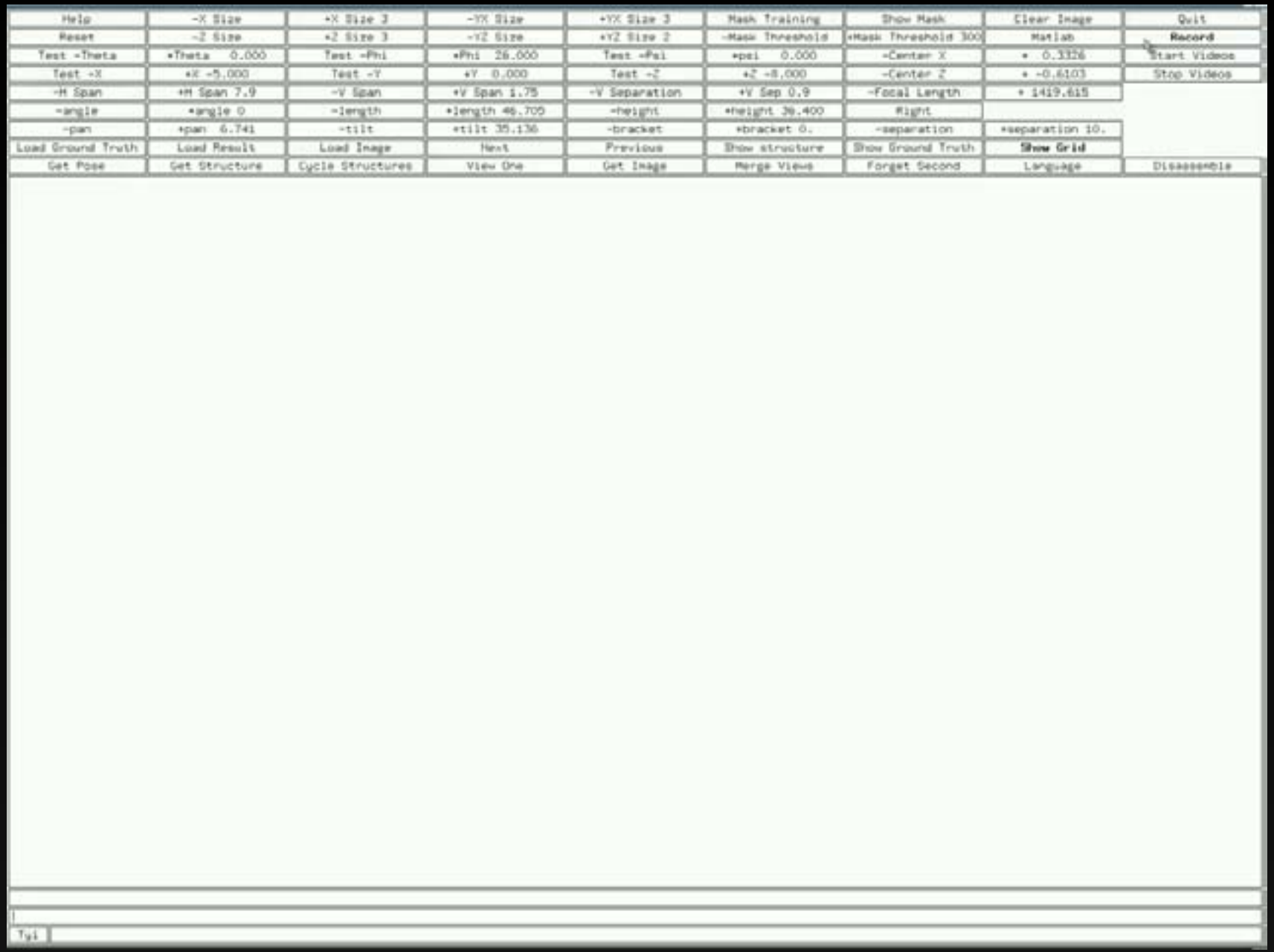
Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}} \int_v \lvert S(s', v) - S(s, v) \rvert$
Planning	$\underset{s \in L}{\text{argmax}} \int_v S(s, v_0 : v : v_n)$
Theory of mind	$S(s, v)$
...	

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}} \int_v \lvert S(s', v) - S(s, v) \rvert$
Planning	$\underset{s \in L}{\text{argmax}} \int_v S(s, v_0 : v : v_n)$
Theory of mind	$S(s, \text{tracks}) S(\text{tracks}, v)$

...

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}} \int_v \lvert S(s', v) - S(s, v) \rvert$
Planning	$\underset{s \in L}{\text{argmax}} \int_v S(s, v_0 : v : v_n)$
Theory of mind	$S(\text{planner}, \text{tracks}) S(s, \text{tracks}) S(\text{tracks}, v)$

...



Siddharth *et al* 2012

The long road ahead . . .

Coherent stories

3D

Forces & contact relations

The long road ahead . . .

Coherent stories

3D

Forces & contact relations

Segmentation

Parts and low-level features

The long road ahead . . .

- Coherent stories
- 3D
- Forces & contact relations
- Segmentation
- Parts and low-level features
- Theory of mind
- Physics
- Modification

The long road ahead . . .

Coherent stories

3D

Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Physics

Modification

The vast majority of verbs: absolve, admire, anger, approve, bark, bend, chase, cheat, complete, concede, discover, fire, follow, fumble, hurry, race, recruit, reject, scratch, steal, taste, want, etc.

Metaphoric extension

etc.

Thanks to many great collaborators

Yevgeni Berzak, Danny Harari, Maximilian Nickel,
Candance Ross, Victor Carbarera, Santiago Perez,
Boris Katz, Shimon Ullman, Tomaso Poggio

Siddharth Narayanaswamy, Jeffrey Siskind,
Sven Dickinson, Song Wang,
Haonan Yu, Caiming Xiong

Recognition	$S(\text{sentence}, \text{video})$
Retrieval	$\underset{v \in V}{\text{argmax}}\, S(s, v)$
Generation	$\underset{s \in L}{\text{argmax}}\, S(s, v)$
Question answering	$\underset{s \in L}{\text{argmax}}\, S(\text{Q}(s, s_q), v)$
Disambiguation	$\underset{i \in \text{parser}(s)}{\text{argmax}}\, S(i, v)$
Acquiring language	$\underset{p}{\text{argmax}} \prod_{s,v} S(s(p), v)$
Images, not videos	$S(\text{sentence}, \text{Flow}(\text{image}))$
Translation	$\underset{s' \in L_b}{\text{argmin}} \int_v \lvert S(s', v) - S(s, v) \rvert$
Planning	$\underset{s \in L}{\text{argmax}} \int_v S(s, v_0 : v : v_n)$
Theory of mind	$S(\text{planner}, \text{tracks}) S(s, \text{tracks}) S(\text{tracks}, v)$

...

Resource: Brains, Minds and Machines Summer Course

Tomaso Poggio and Gabriel Kreiman