UNIVERSITY OF SOUTH FLORIDA

A PREEMINENT RESEARCH UNIVERSITY

AI+X INSTITUTE FOR ARTIFICIAL INTELLIGENCE

Open World Reasoning with Canonical Representations from Grenander's Pattern Theory

Sudeep Sarkar, U of South Florida, USF
Anuj Srivastava, Florida State University, FSU.
Fillipe deSouza, USF, now at Intel
Sathyanarayan Aakur, USF, now at Oklahoma St U

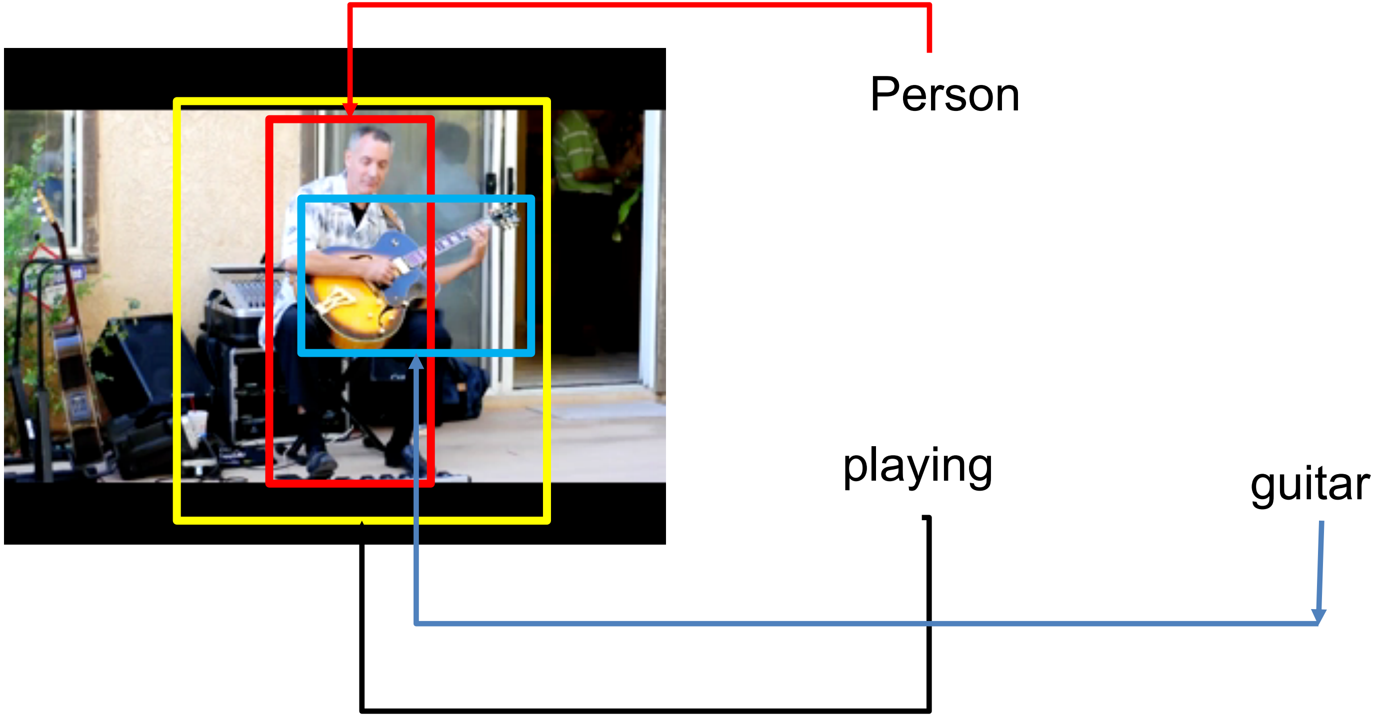
Events are central to human experience

An event is a segment of time at a given location that is perceived by an observer to have a beginning and an end.

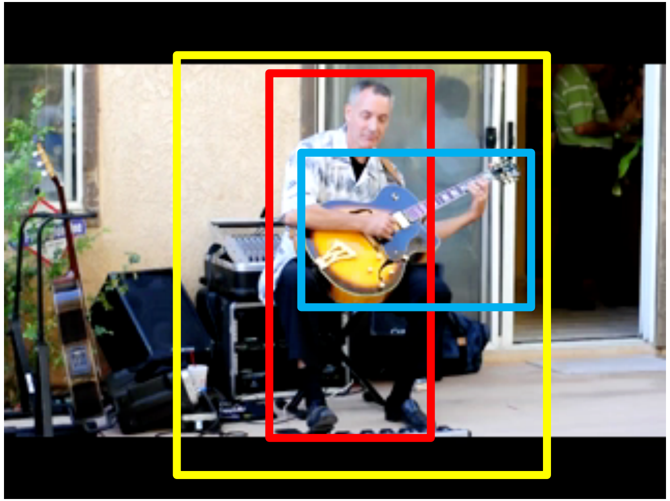
--Zacks, Tversky, and Iyer, 2001

It is described by
- Who – nouns
- What – actions, activity
- Where – location
- When – temporal
- Why – intention

Deep Learning 1.0



Deep Learning 2.0

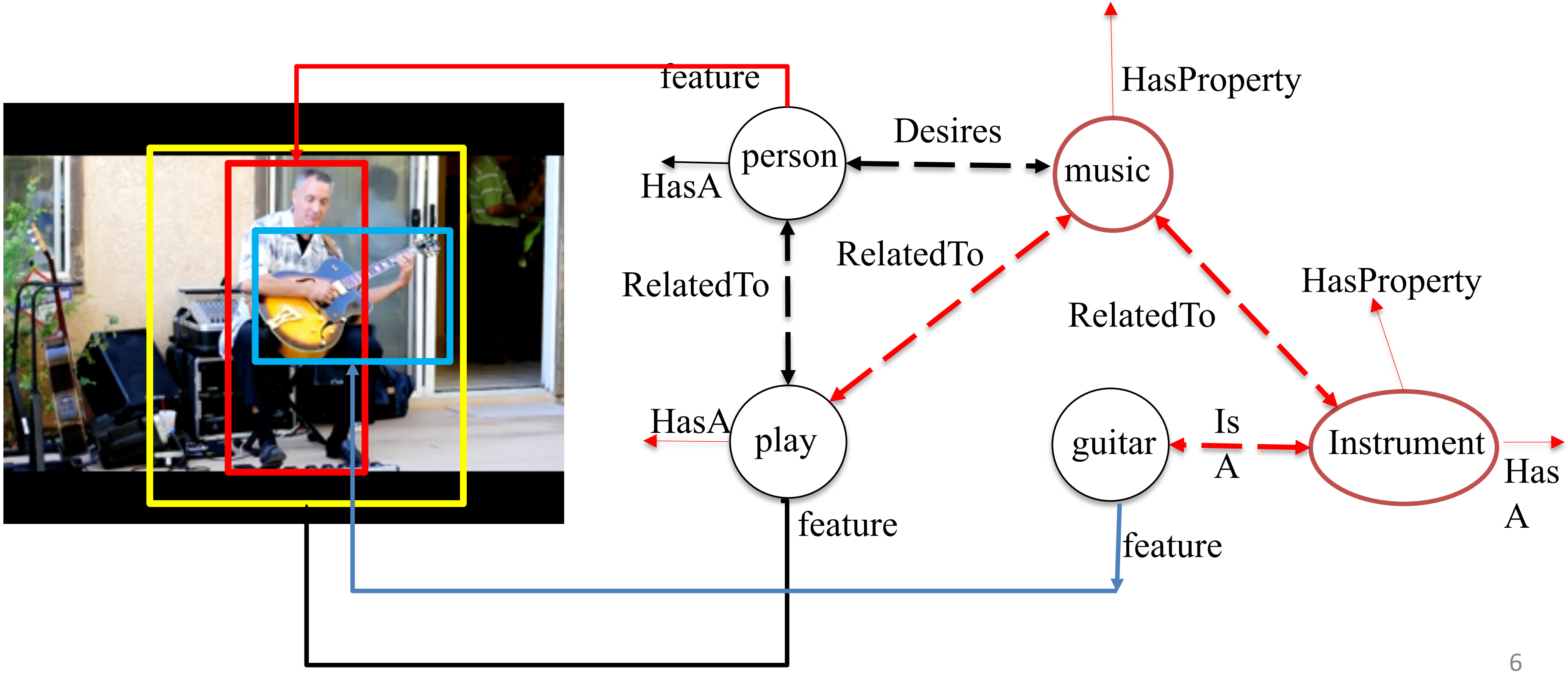


Person playing guitar

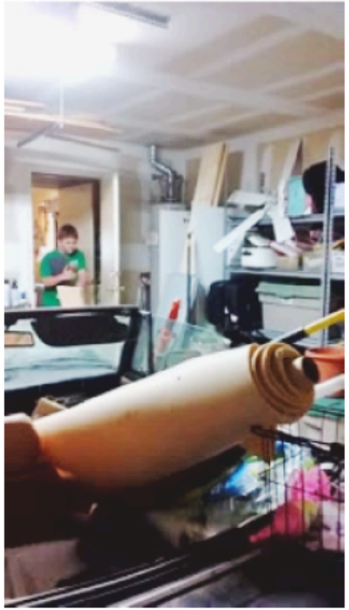
Open World Event understanding

making connection to past knowledge and creating an event model that goes beyond what is sensed.

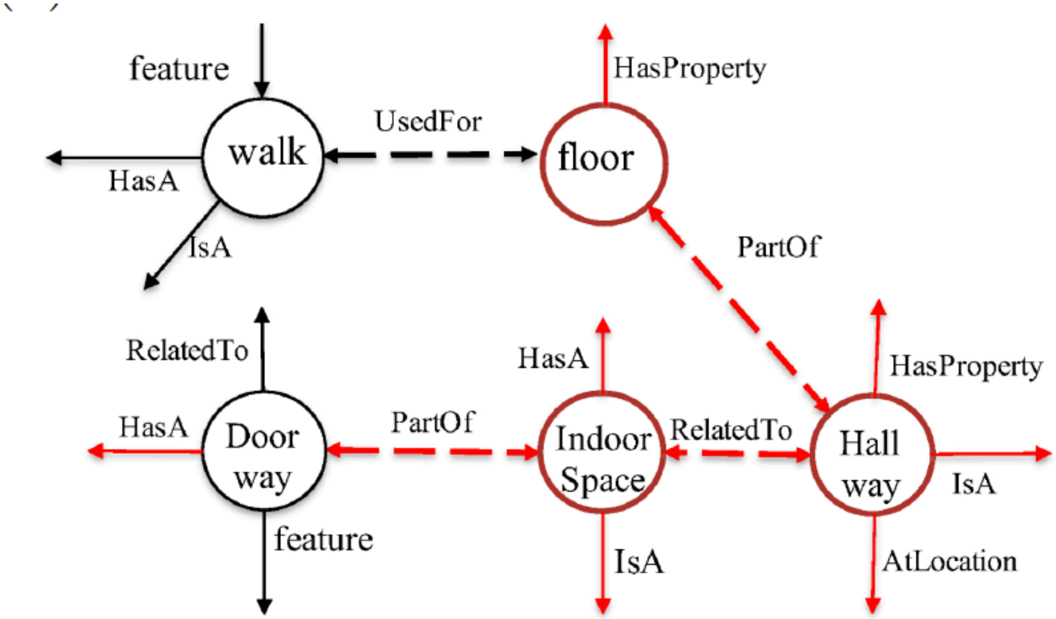
Beyond labels, beyond what is sensed…



Rich, open world interpretation



Walk through doorway

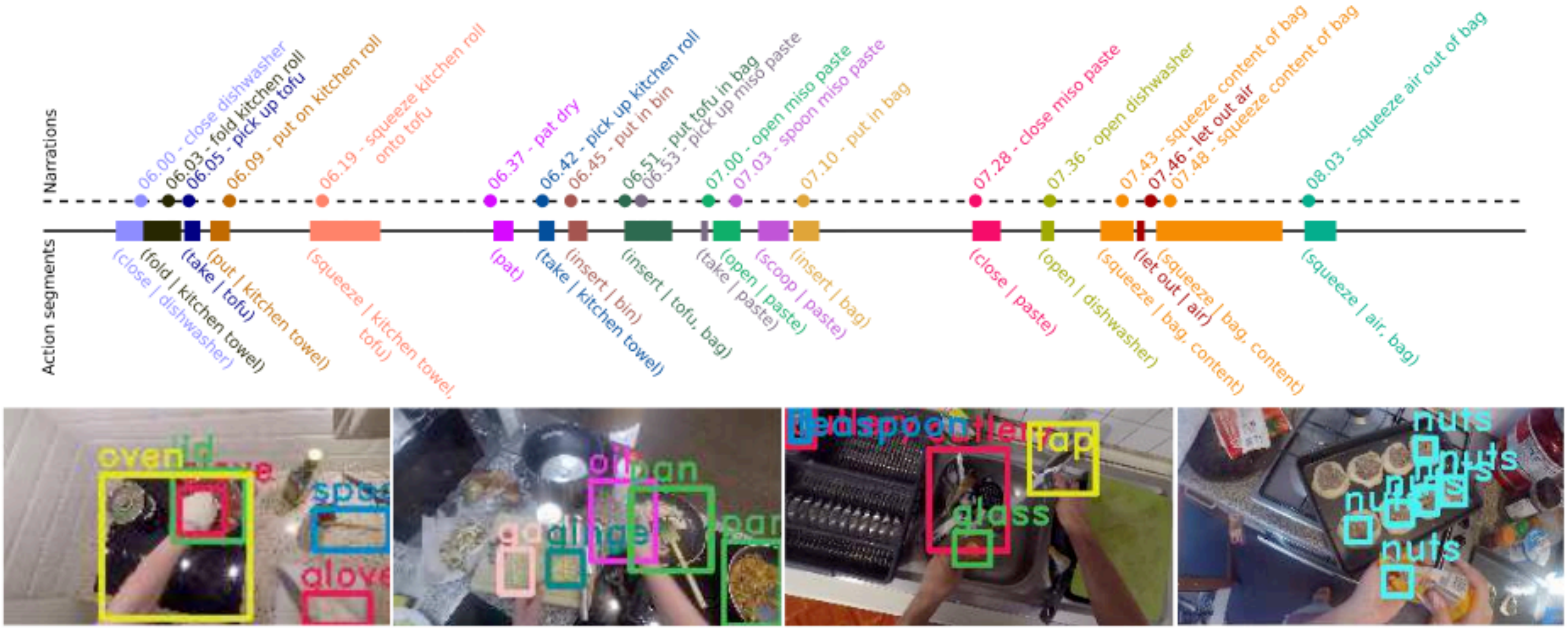


Walk through doorway

The ability to support open world inference is limited by three main aspects:

- the source of semantics,
- the underlying representation, and
- the ability to continuously learn or adapt.

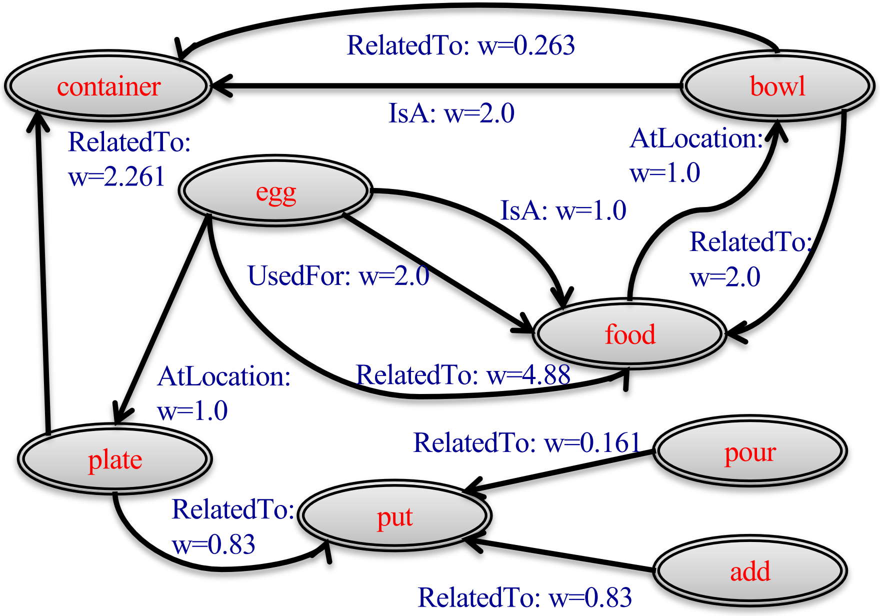
Over reliance on annotated data for semantics



© Scaling Egocentric Vision: The EPIC-KITCHENS Dataset, ***Dima Damen et al.*** ECCV, 2018.

Use Symbolic Knowledge-bases for Semantics

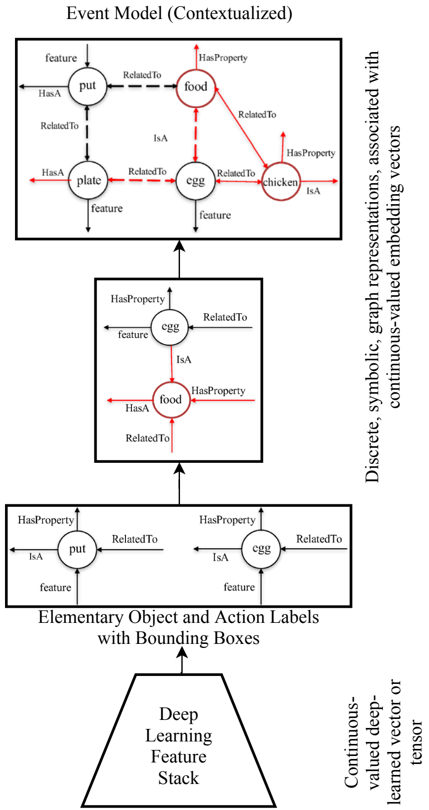
- Crowd-sourced knowledge-base mined from:
  - Wiktionary and Wikipedia
  - DBPedia (Auer et al., 2007)
  - Freebase (Bollacker et al., 2008)
  - WordNet (Fellbaum, 1998)
- It contains 12.5 million edges, representing about 8.7 million assertions connecting 3.9 million concepts (different languages).



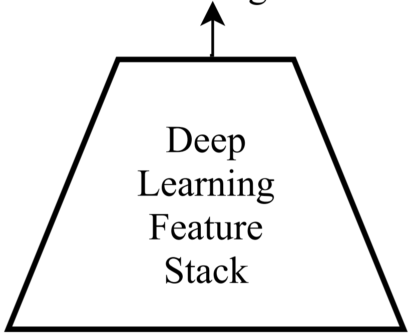
Speer, Robert, and Catherine Havasi. "Representing General Relational Knowledge in ConceptNet 5." In LREC, pp. 3679-3686. 2012

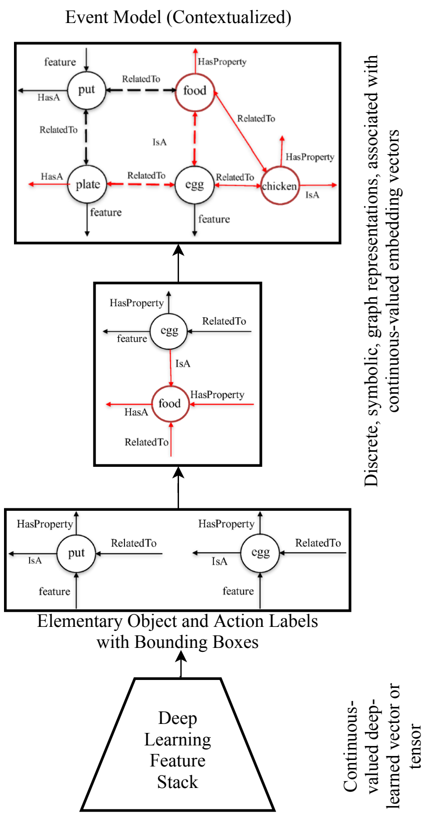
Rise of Neuro-Symbolic Approaches

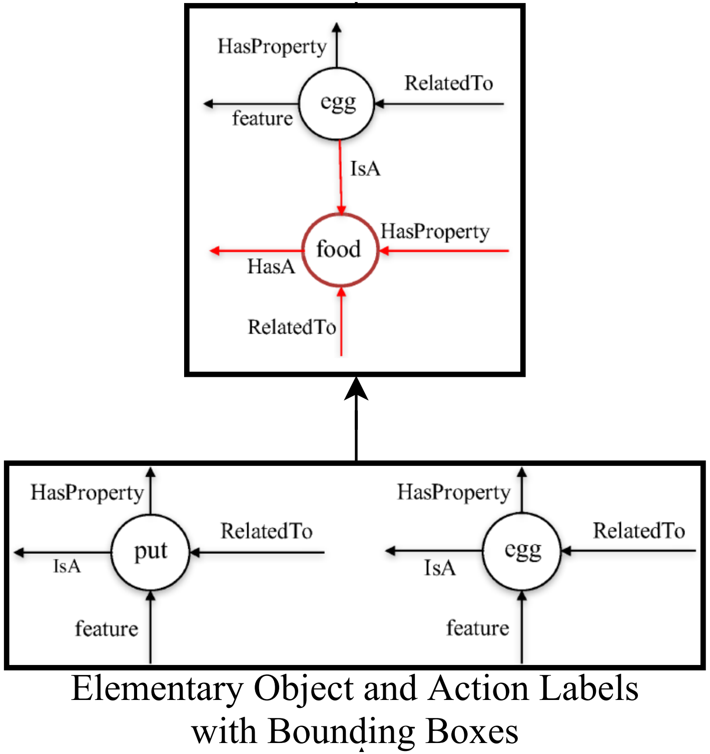
- Stack of continuous-valued vectors
  - The top-level vectors mapped to desired concepts
  - Concepts: Labels, phrases, sentences
- Graph-based, explicit, symbolic
  - Explainable, can be targeted for different applications



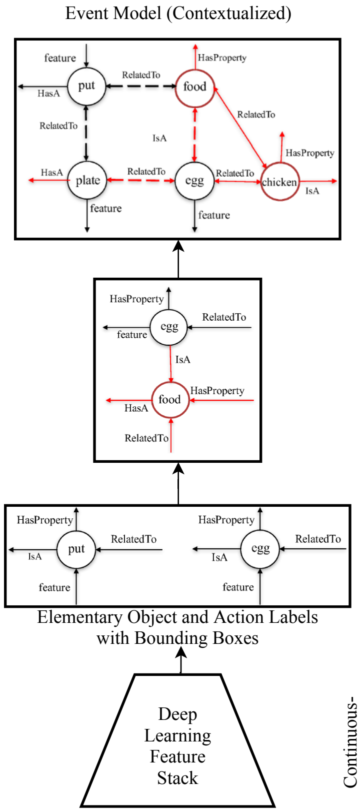
Elementary Object and Action Labels
with Bounding Boxes




Event Model (Contextualized)
feature
put
HasA
RelatedTo
food
HasProperty
RelatedTo
IsA
RelatedTo
HasProperty
HasA
plate
RelatedTo
egg
RelatedTo
chicken
IsA
feature
feature
HasProperty
egg
RelatedTo
feature
IsA
food
HasProperty
HasA
RelatedTo
HasProperty
put
RelatedTo
IsA
feature
HasProperty
egg
RelatedTo
IsA
feature
Elementary Object and Action Labels with Bounding Boxes
Deep Learning Feature Stack
Discrete, symbolic, graph representations, associated with continuous-valued embedding vectors
Continuous-valued deep-learned vector or tensor

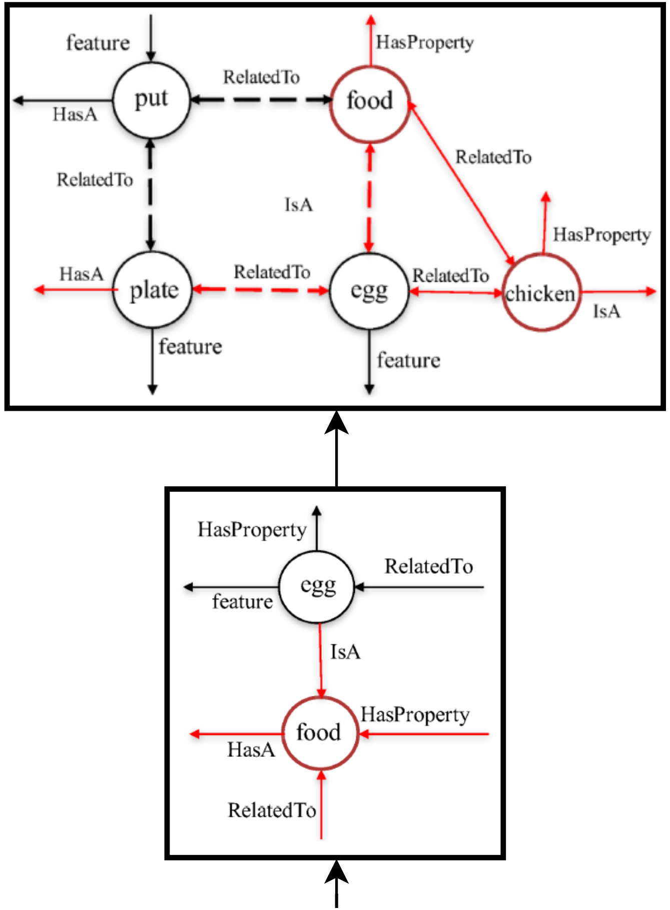

HasProperty
egg
RelatedTo
feature
IsA
food
HasProperty
HasA
RelatedTo
HasProperty
put
RelatedTo
IsA
feature
HasProperty
egg
RelatedTo
IsA
feature
Elementary Object and Action Labels with Bounding Boxes

Event Model (Contextualized)

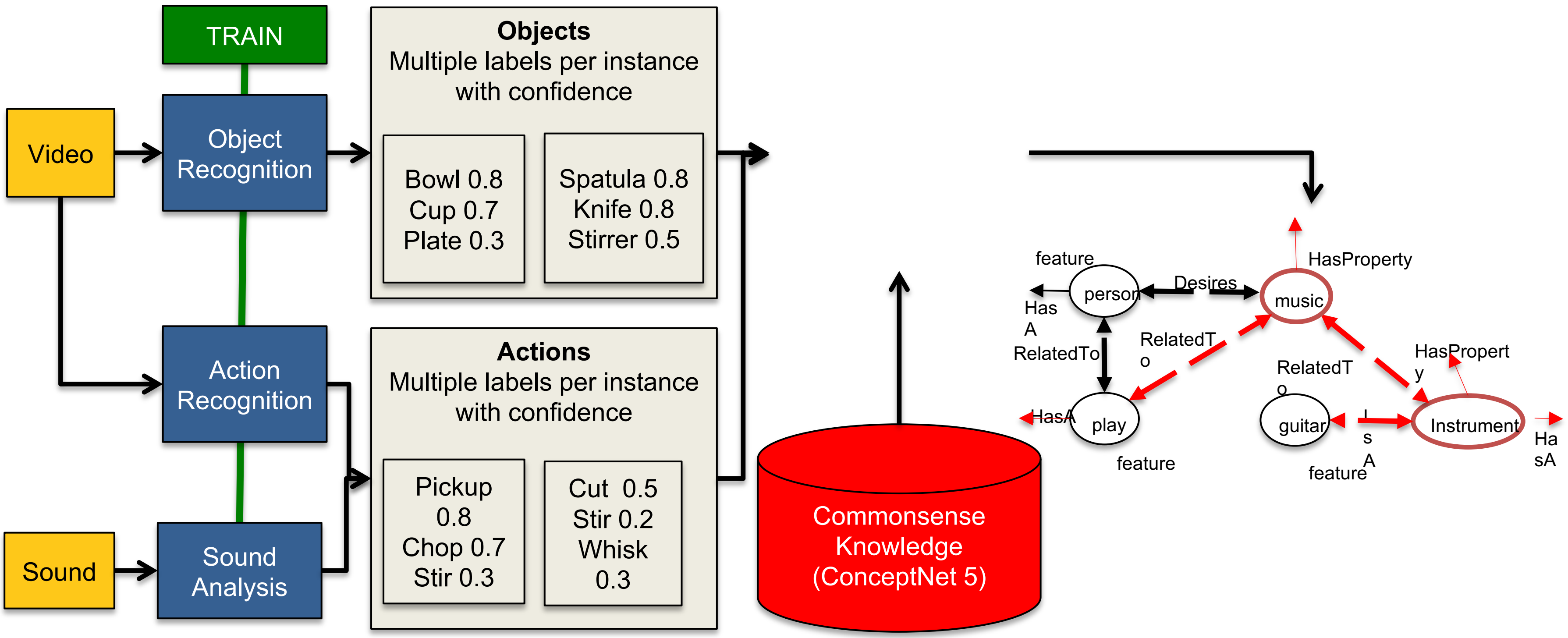


Discrete, symbolic, graph representations, associated with continuous-valued embedding vectors

Continuous-valued deep-learned vector or tensor

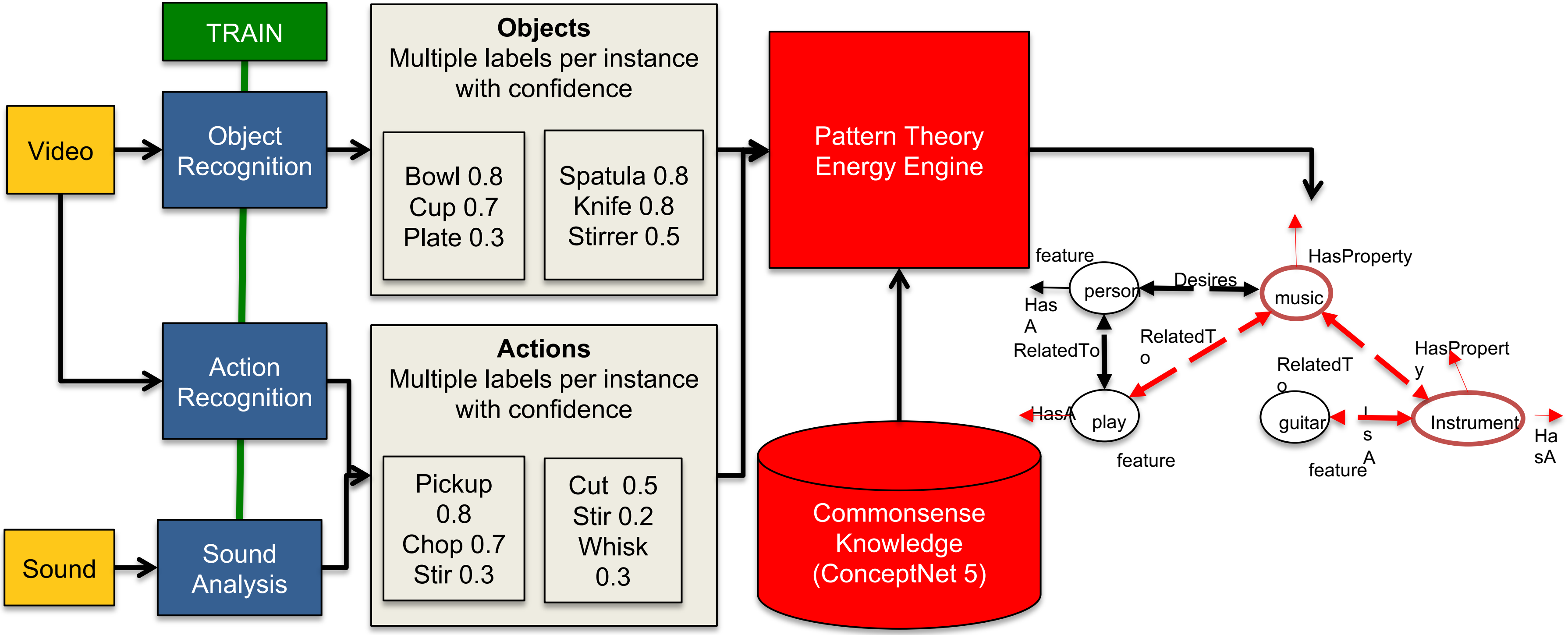


Discrete, symbolic, graph representations, associated with continuous-valued embedding vectors



Semantic Mapping

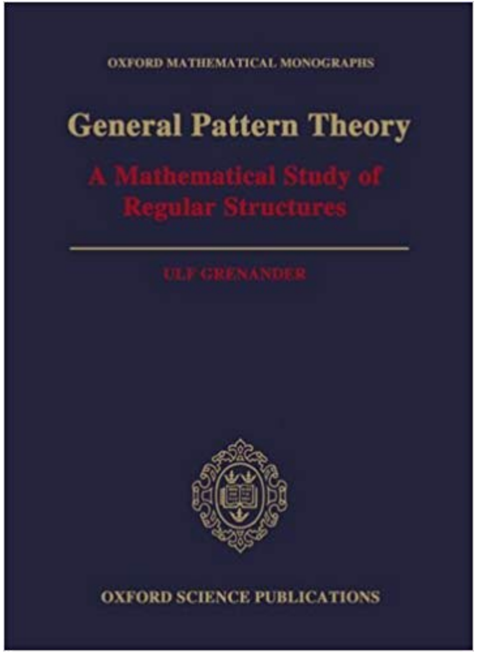
Semantic Representation

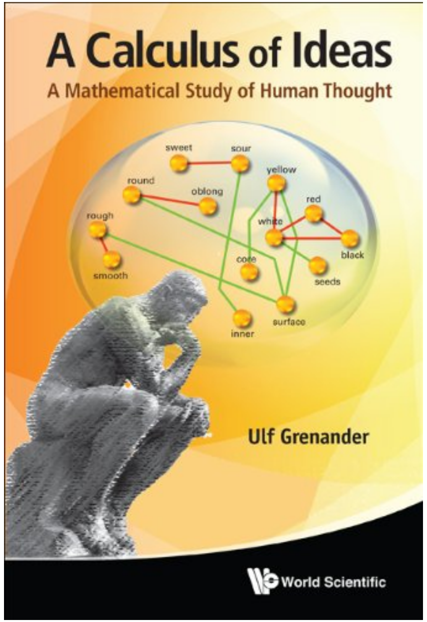


Semantic Mapping

Semantic Representation

Symbolic Reasoning using Grenander's Canonical Representations





Sketch of the approach

- Pattern theory is combinatorial in nature
  - Complex structures are built from simpler ones.
  - Much like elements combine to make molecules, proteins, etc.
- Symbols can interact with each other form larger combinations.
- The interactions are constrained by how symbols interact locally and by the characteristic of the overall graph structure,
- Probabilistic structures on the representations allow for expressing the variation of natural patterns.
- A unified manner for viewing DAGs, MRFs, Gaussian random fields and probabilistic formal languages.

Canonical Representation Involves

Basic symbols
(Generators)

Equivalences among symbols
(Similarity group)

$$R(G, S, \rho, \Sigma)$$

How symbols interact with each other.
(Local regularity)

Constraint on the overall structure
(Global Regularity)

Lastly, define a probability space over these structures

Generators are the basic units of representation)

- Generator Space G

$$G = \{g_1, g_2, \dots, g_n\}$$

g_i

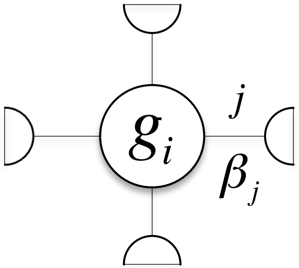
- Elementary symbols are our generators

$$R(G, S, \rho, \Sigma)$$

Generators
(Basic Units of Representation)

- Generator Space G

$$G = \{g_1, g_2, \ldots, g_n\}$$



- Bonds

$$\beta_j(g_i) \in B, j = 1, \ldots, w(g_i)$$

$$R(G, S, \rho, \Sigma)$$

Generators
(Basic Units of Representation)

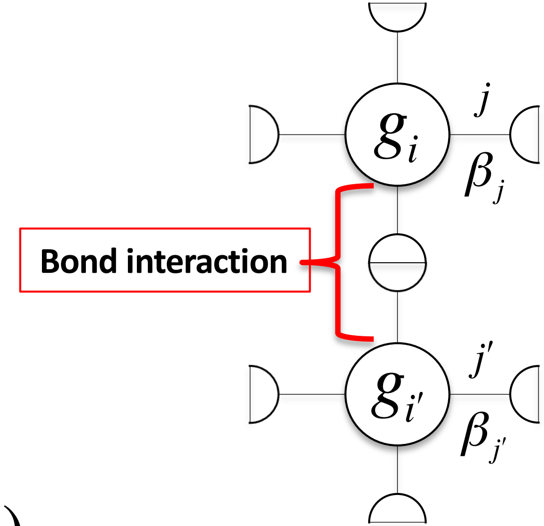
- Generator Space G

$$G = \{g_1, g_2, \ldots, g_n\}$$

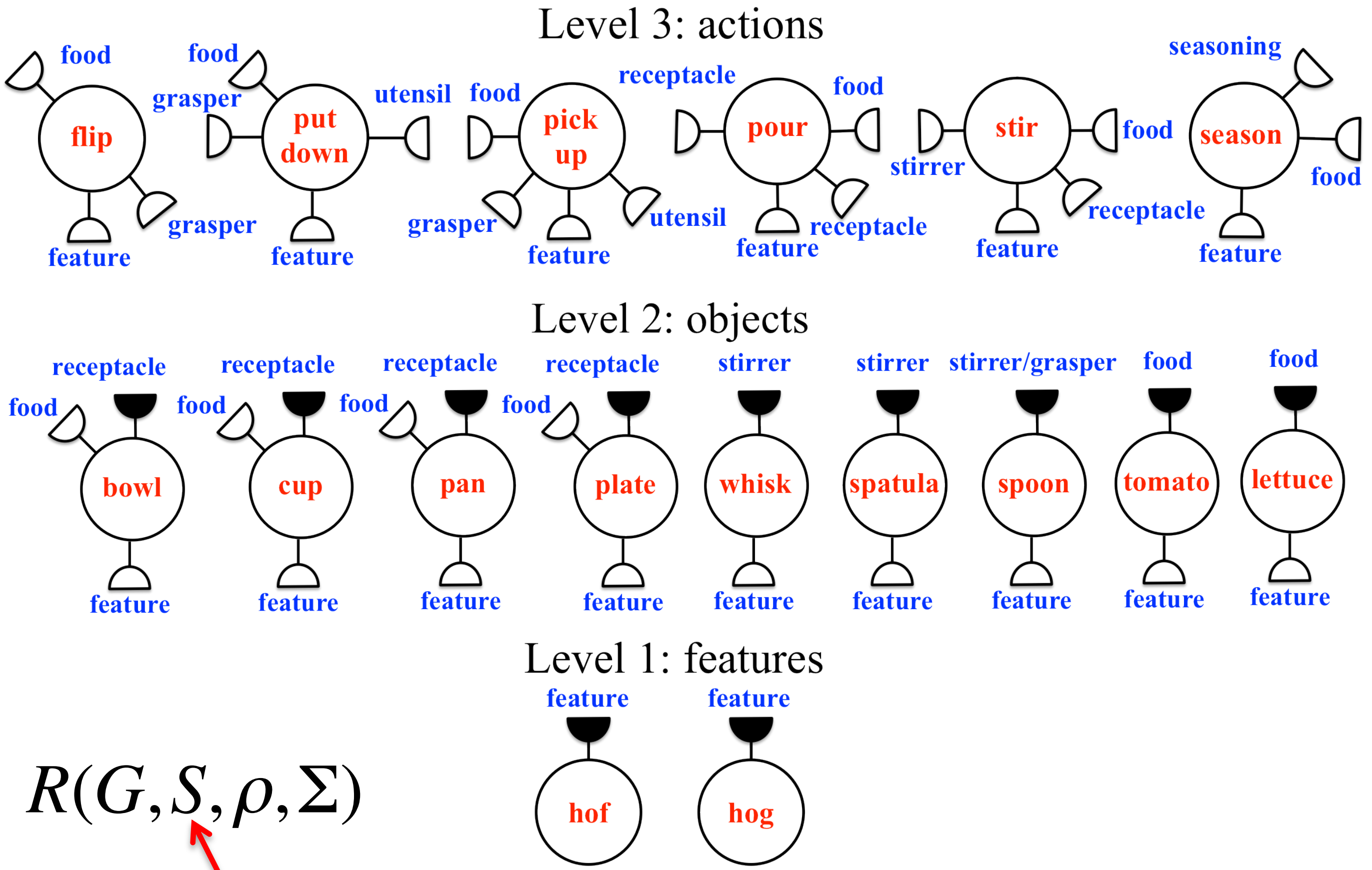
- Bonds

$$\beta_j(g_i) \in B, j = 1, \ldots, w(g_i)$$

$$R(G, S, \rho, \Sigma)$$



Example Generators



$$R(G, S, \rho, \Sigma)$$

Similarity Group

- Define a similarity group S as

$$s : G \rightarrow G \mid s \in S$$

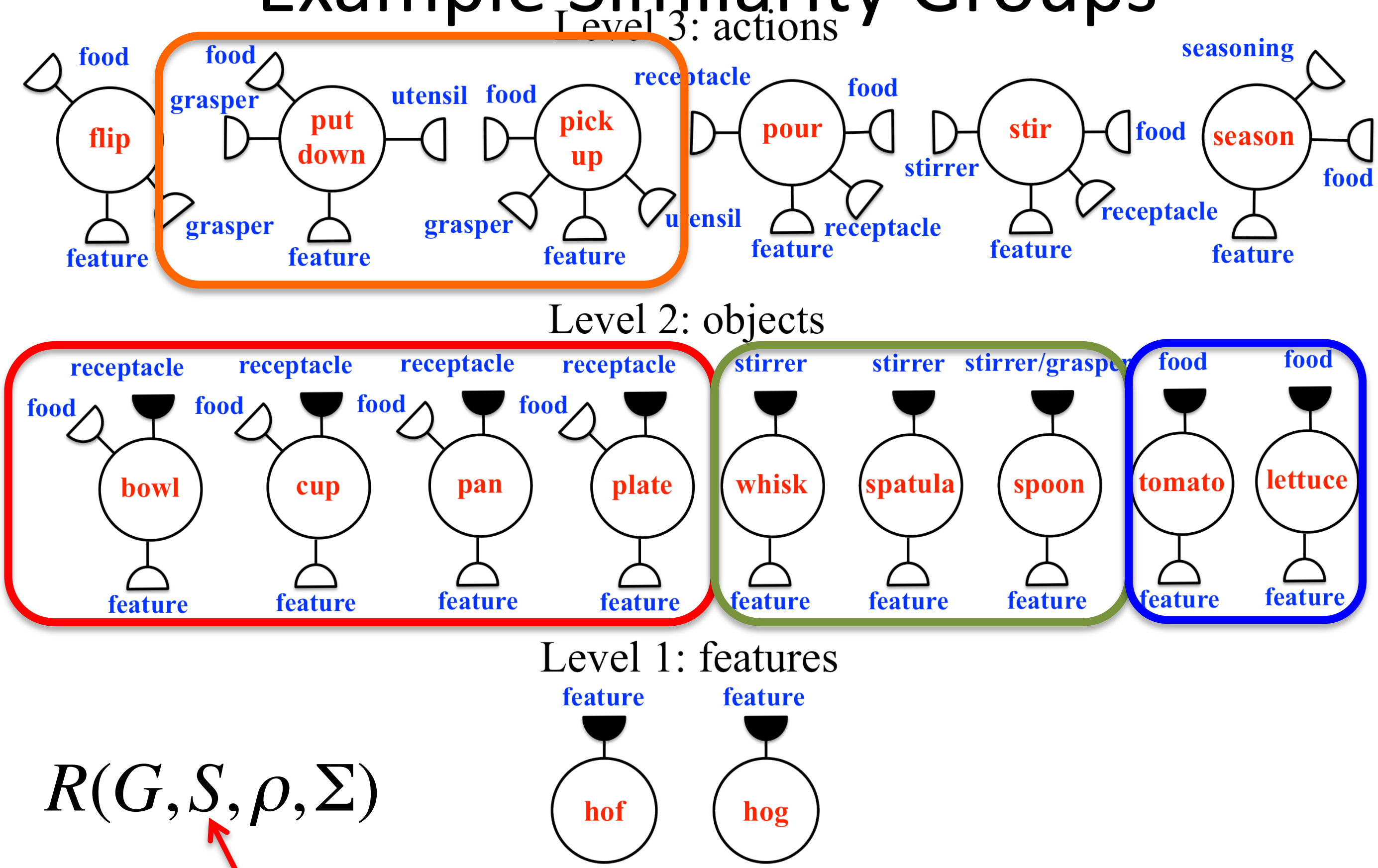
- Induces a partition of the generator space into equivalence classes (disjoint groups Gα)

$$G = \bigcup_{\alpha \in A} G^{\alpha}$$

- Such that $sg = g' \mid g, g' \in G^{\alpha}$

$$R(G, S, \rho, \Sigma)$$

Example Similarity Groups



$$R(G, S, \rho, \Sigma)$$

Bond Relation (Local Regularity)

- Bond relation ρ specifies the rules of combination among generators, formally defined as

$$\rho : B \times B \rightarrow \{TRUE, FALSE\}$$

- Determines local regularity of a connected structure of generators

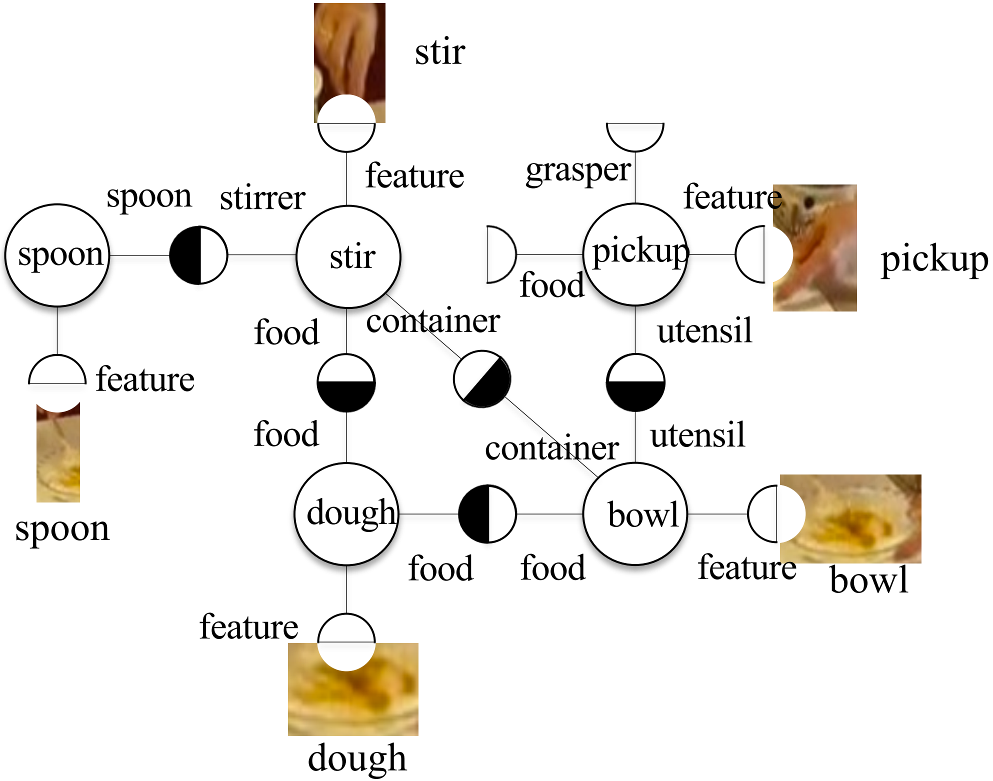
$$R(G, S, \rho, \Sigma)$$

Configurations

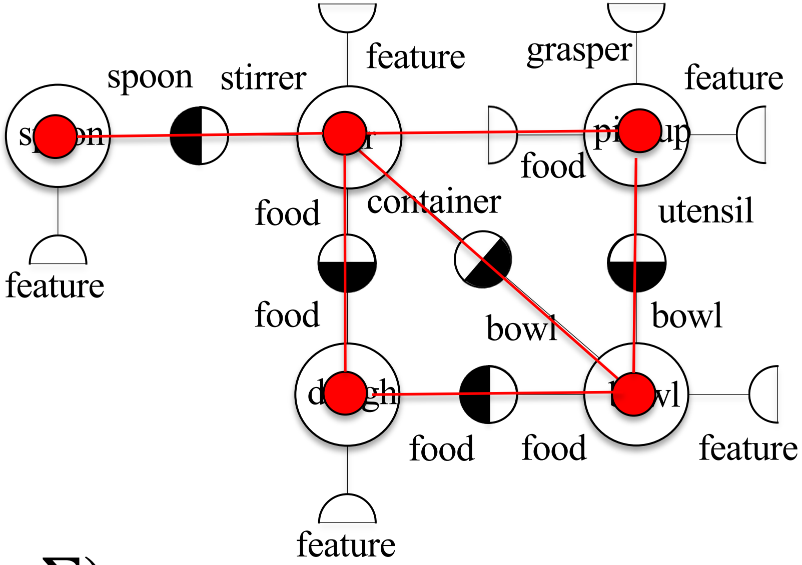
Picking up bowl with dough. Stirring dough in a bowl using a spoon.

$$\boxed{c = \sigma(g_1, g_2, \ldots, g_n)}$$

$$R(G, S, \rho, \Sigma)$$



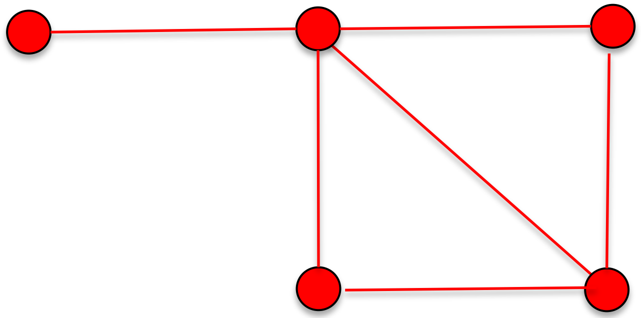
Connection Graph (Global Regularity)



$$R(G, S, \rho, \Sigma)$$

Connection Graph (Global Regularity)

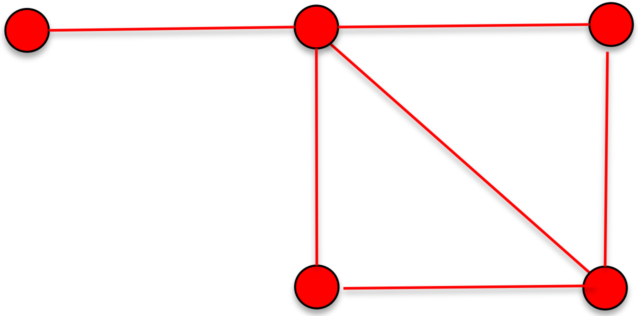
$$\sigma \in \Sigma$$



$$R(G, S, \rho, \Sigma)$$

Connection Graph (Global Regularity)

$$\sigma \in \Sigma$$



$$R(G,S,\rho,\Sigma)$$

- If S is fixed then we have MRF or Bayesian network, if directed
- Could be a tree structure like AND-OR graphs

Connection Type (Global Regularity)

- Σ represents the connection type, in our example, Σ=POSET (partially ordered set)
- Partial ordering is based on the hierarchy of the representation
- More general than MRF, Bayesian networks, AND-OR, etc.

$$R(G, S, \rho, \Sigma)$$

Relationship to Other Formalisms

- Undirected bonds, pre-specified, fixed structure, lattice connection graph → MRF
- Directed bonds, pre-specified, fixed structure, DAG connection graph → Bayesian Networks
- Undirected bonds, pre-specified, fixed structure, AND-OR tree connection graph → AND-OR graph (Zhu et al. at UCLA)
- Grammar rules as generators, tree-structure as connection graph → Context Free Grammar

PROBABILITY MEASURE ON THE CONFIGURATION SPACE

90

Knowledgebase (ConceptNet)

$$P(c|C_N, F) = \frac{1}{Z} e^{-E(F|c) - E(c|C_N)}$$

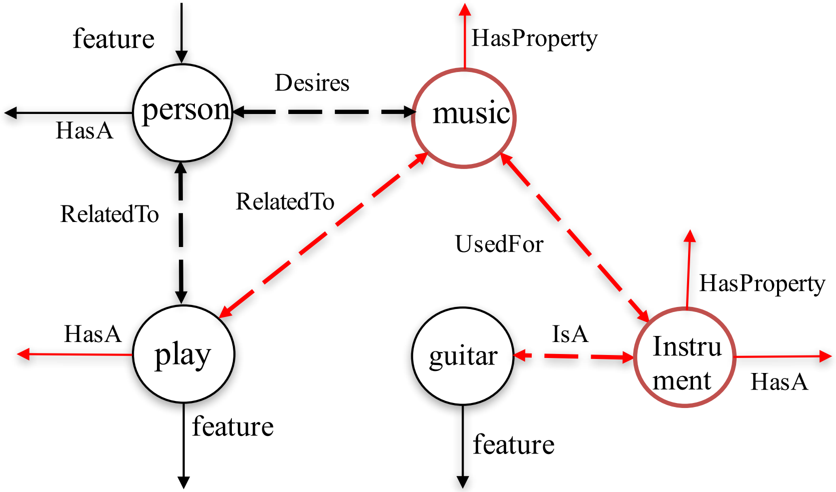
configuration

Features detected

Direct Video Support

Prior Support

Components of the energy function



$$E(c) = -\sum_{(\beta',\beta'')\in c} a_{sup}(\beta'(g_i), \beta''(g_j)) - \sum_{(\beta',\beta'')\in c} a_{sem}(\beta'(g_i), \beta''(g_j)) + k \sum_{\bar{g}_i \in G'} \sum_{\beta^j_{out} \in \bar{g}_i} [D(\beta^j_{out}(\bar{g}_i))]$$

"feature" links | dashed links | Unconnected solid links

$$P(c|C_N, F) = \frac{1}{Z} e^{-E(F|c) - E(c|C_N)}$$

The partition function involves a double sum. $\sum_\sigma \sum_c$

Inference through Stochastic Search

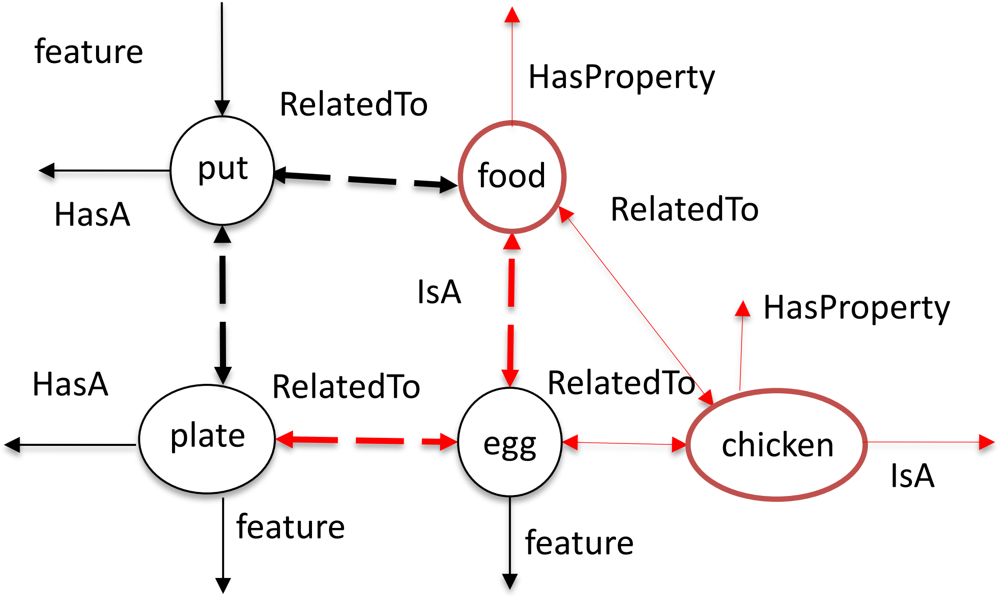
- Searches over connection structures AND generators
  - Gibbs ensemble
  - Note difference with other graphical approaches that freeze the structure after training.
- Local and global proposals
- Combinatorics controlled by similarity structure and structure of the prior knowledge.

Algorithm 1: MCMC based simulated annealing for inference

```
MCMC Simulated Annealing (F, G, U, α, p, k_max, T_0);
c ← resetConfiguration(F, G, U)
best ← c
for k ← 1 ... k_max: do
    t ← UniformSample(0, 1)
    if t < p then
        c' ← resetConfiguration(F, G)
    end
    else
        c' ← groundedSwitch(c, G, U)
    T ← T_0 × α^k
    if E(c') < E(c) then
        c ← c'
    end
    else
        z ← UniformSample(0, 1)
        if z < exp(−(E(c') − E(c))/T) then
            c ← c'
        end
    if E(c) < E(b) then
        best ← c
    end
end
return best
```

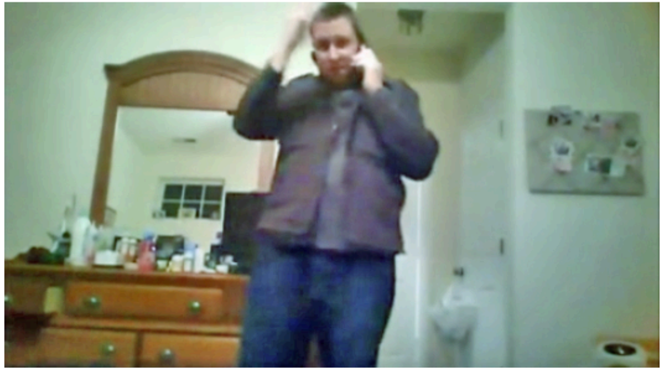
Global Move

Local Move

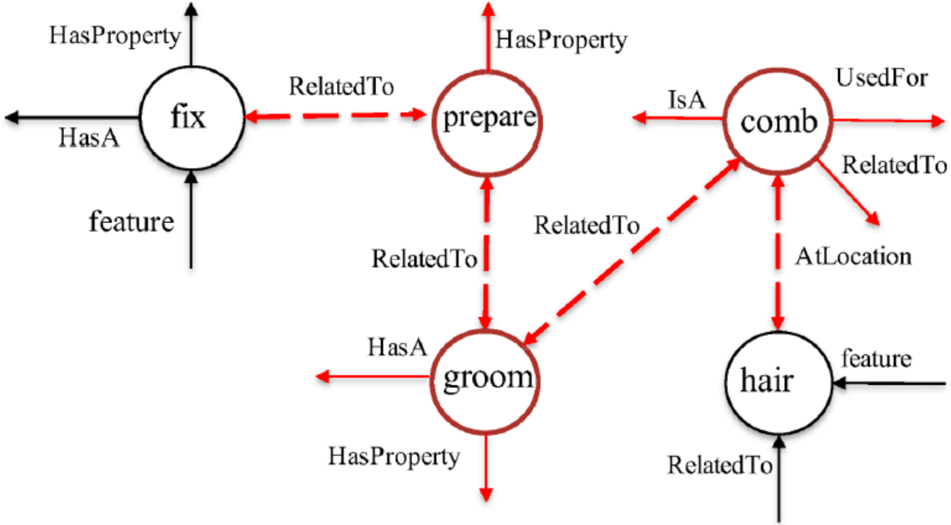


Aakur, S., de Souza, F., & Sarkar, S. (2019). Generating open world descriptions of video using common sense knowledge in a pattern theory framework. Quarterly of Applied Mathematics, 77(2), 323-356.

Rich, open world interpretation



Fix hair

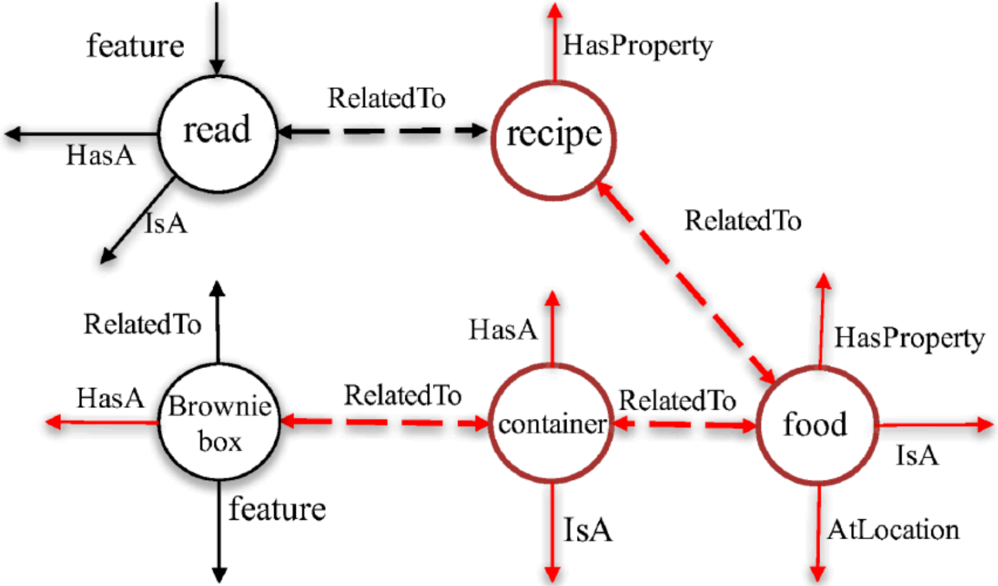


Fix hair

It is also to be noted that for many of the interpretations, the label with the highest confidence score was not the one used in its final (best) interpretation.

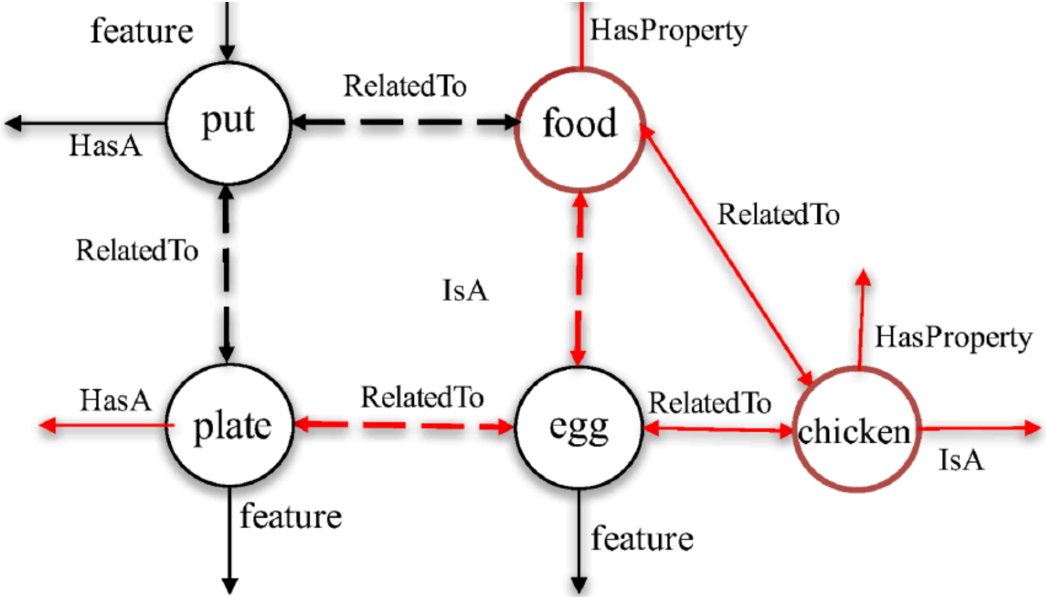
Groundtruth

Output Interpretation



It is also to be noted that for many of the interpretations, the label with the highest confidence score was not the one used in its final (best) interpretation.

Put egg on plate

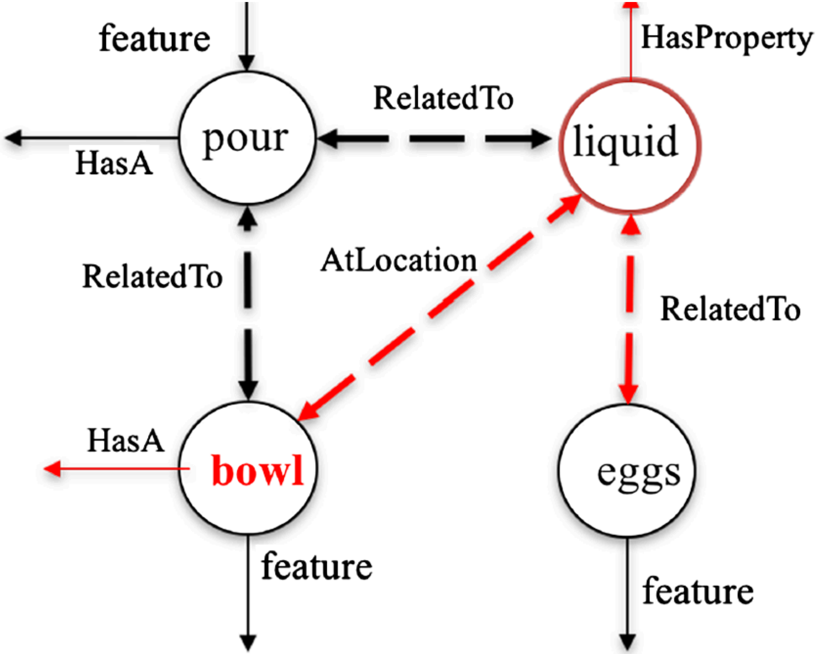


Put egg on plate

(b)

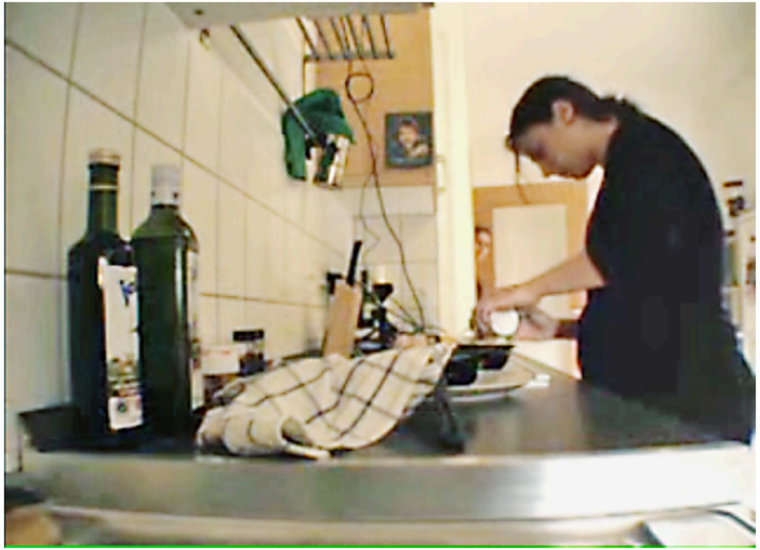
Even "erroneous" interpretations are not semantically "bad"

Pour eggs on pan.

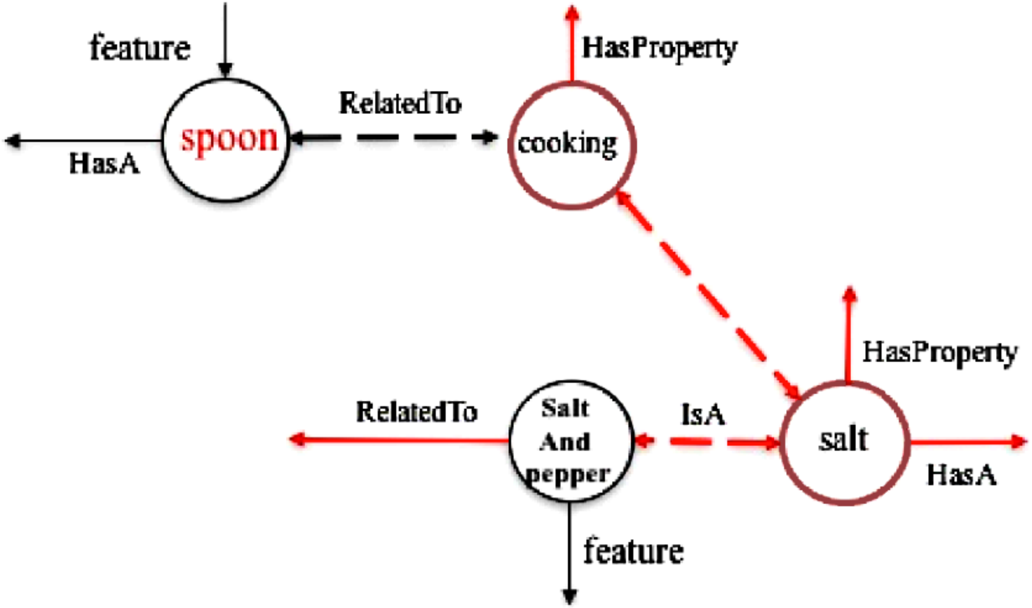


Pour eggs into bowl.

Error but semantics okay

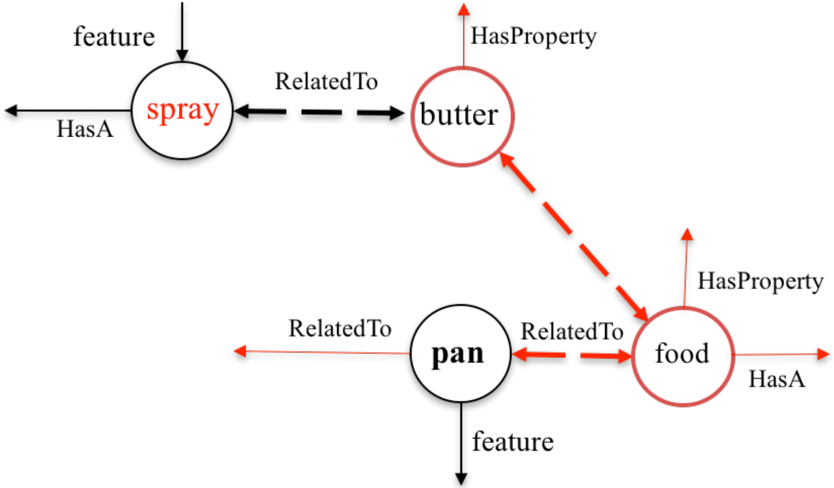


Add salt and pepper.



Spoon salt and pepper

Egocentric videos…



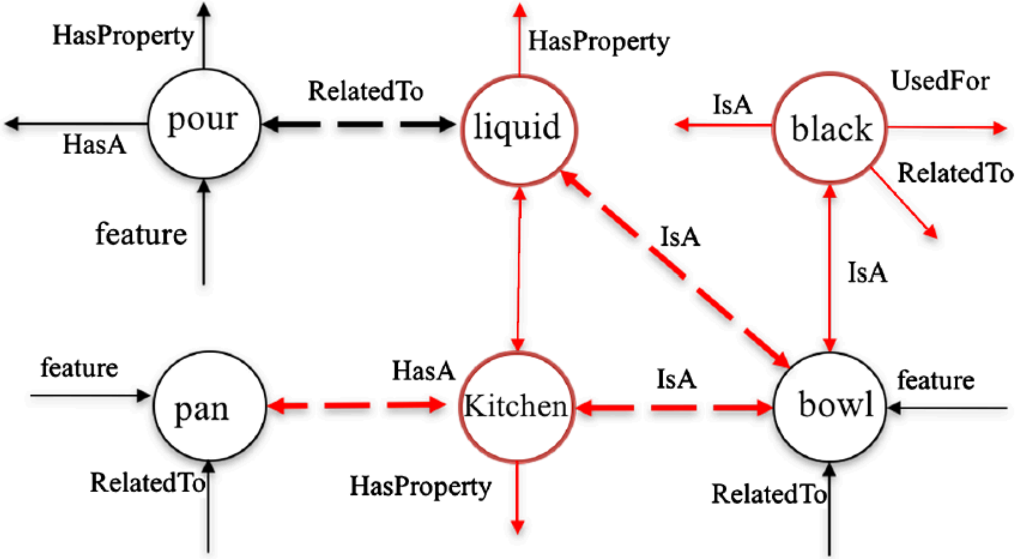
Aakur, S., de Souza, F., & Sarkar, S. (2019). Generating open world descriptions of video using common sense knowledge in a pattern theory framework. Quarterly of Applied Mathematics, 77(2), 323-356.

Works on ego-centric videos

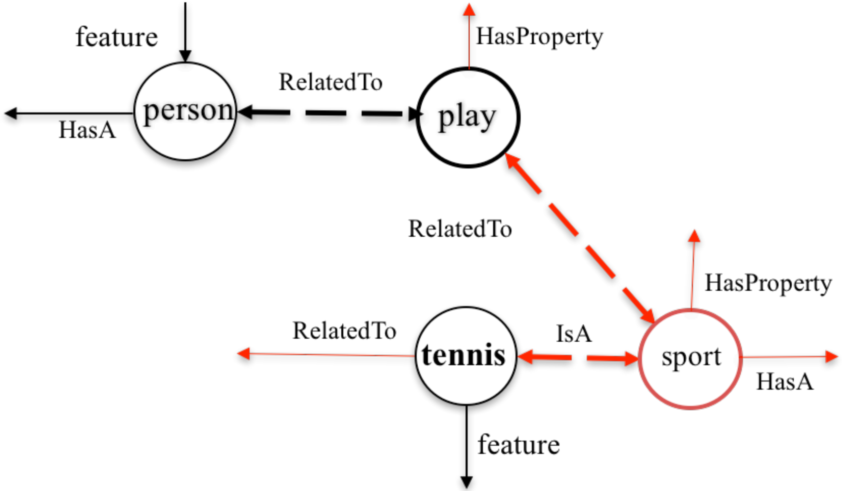
Groundtruth

Pour from bowl to pan

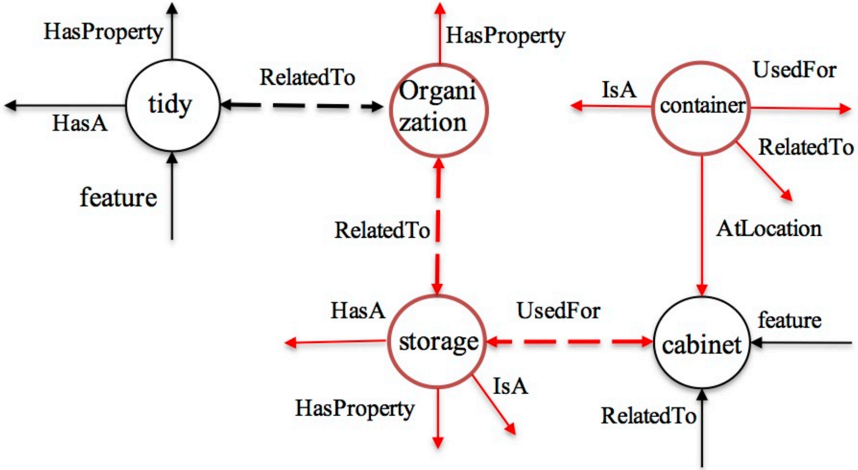
Output Interpretation



Pour from bowl to pan



Aakur, S., de Souza, F., & Sarkar, S. (2019). Generating open world descriptions of video using common sense knowledge in a pattern theory framework. Quarterly of Applied Mathematics, 77(2), 323-356.



Aakur, S., de Souza, F., & Sarkar, S. (2019). Generating open world descriptions of video using common sense knowledge in a pattern theory framework. Quarterly of Applied Mathematics, 77(2), 323-356.

Other graphical approaches

TABLE 3. Results on Breakfast Action dataset. Top 10 means that we consider the best of 10 interpretations generated by the approach.

Approach	Precision
HMM [37]	14.90%
CFG + HMM [37]	31.8%
RNN + ECTC [33]	35.6%
RNN + ECTC (Cosine) [33]	36.7%
PT+weights (Top 10) [20]	33.40%
PT+training (Top 10) [20]	38.60%
Our Approach (Top 10)	**41.87%**

No training on Object-action pairs

1000 recipe videos, consisting of different scenarios with a combination of 10 recipes

MSVD: 1,970 videos taken from YouTube

Approach	BLEU Score
Probabilistic Factor Graph, [Thomason 2014 COLING]	13.68%
LSTM + Transfer Learning,, trained on Youtube [Venugopalan 2014 arXiv]	31.19%
LSTM + Transfer Learning, trained on FLICKR [Venugopalan 2014 arXiv]	32.03%
LSTM + Transfer Learning, trained on COCO [Venugopalan 2014 arXiv]	33.29%
LSTM + Transfer Learning, trained on COCO+FLICKR [Venugopalan 2014 arXiv]	33.29%
GoogleNet+3D CNN, [Yao 2015 ICCV][31]	41.92%
Hierarchical RNN, [Pan 2016 CVPR] [18]	43.60%
Energy Minimizing Formulation using Pattern Theory, (Ours)	**42.98%**

Table 2: BLEU scores on "videos in the wild" Microsoft Video Description Corpus (MSVD) dataset. CCN: Convolutional Neural Networks, RNN: Recurrent Neural Networks, LSTM: Long-Short Term Memory RNN.

No training needed other than for the basic categories of objects and actions

Aakur, S., de Souza, F., & Sarkar, S. (2019). Generating open world descriptions of video using common sense knowledge in a pattern theory framework. Quarterly of Applied Mathematics, 77(2), 323-356.

9,848 videos across
157 action classes

Table 1. Results on Charades dataset. ATF refers to Asynchronous Temporal Fields method. PT + ConceptNet semantics refers to the proposed approach. Trained semantics will indicate the use of training annotations to capture semantics between concepts. Note: All models use 2-stream features extracted from the videos as input, unless otherwise indicated.

Approach	mAP
LSTM	17.80%
RGB + ATF + trained semantics, no intent, no temporal	17.30%
RGB + ATF + trained semantics, no temporal, intent	17.40%
RGB + ATF + trained semantics, temporal, no intent	17.40%
ATF + trained semantics, intent, temporal	22.40%
PT + ConceptNet semantics, no intent, no temporal	**29.69%**
LSTM + PT + ConceptNet semantics, no intent, no temporal	**32.56%**

Easily Extendible Formalism

- Object Clutter – introduce spatial proximity bond between action and object (IJCV 2016)
- Simultaneous events – introduce spatial proximity bond between action and object (IJCV 2016)
- Activity Sequence – introduce temporal bonds (CVPR 2015)

Acknowledgement

- This material is based upon work supported by the National Science Foundation under Grant Nos: NSF IIS 1217676, NSF CNS 1513126, NSF CMMI 1826258
- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.