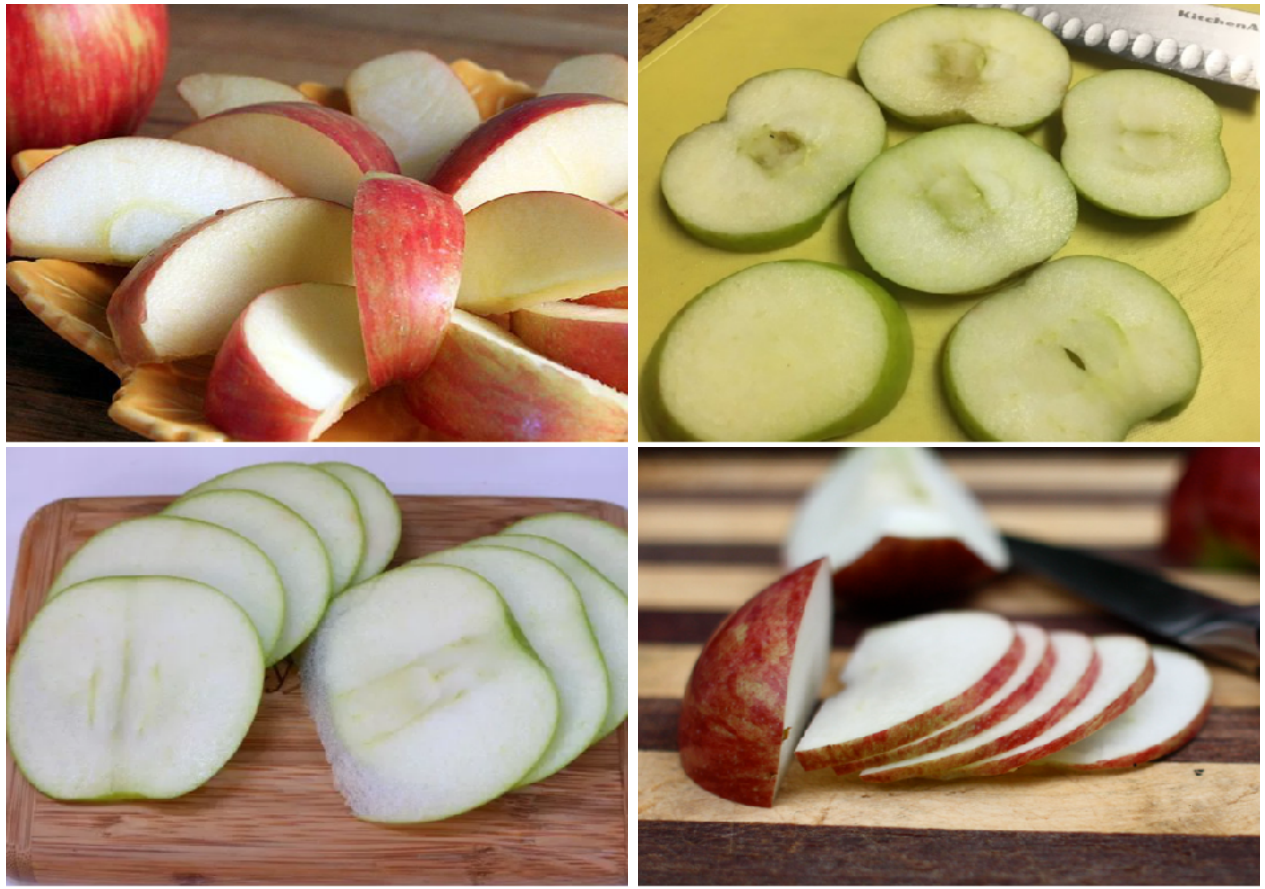
What Action Causes This? Towards Naive Physical Action-Effect Prediction

Qiaozi Gao[1], Shaohua Yang[1], Joyce Y. Chai[1], Lucy Vanderwende[2]

[1] Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI 48824
[2] Microsoft Research, Redmond, WA 98052

Motivation

- What action causes this?



Motivation

- What is the result state of “open box”?

Understanding Cause-Effect

The developing understanding that one event brings about another

8 months	18 months	36 months
At around eight months of age, children perform simple actions to make things happen, notice the relationships between events, and notice the effects of others on the immediate environment.	At around 18 months of age, children combine simple actions to cause things to happen or change the way they interact with objects and people in order to see how it changes the outcome.	At around 36 months of age, children demonstrate an understanding of cause and effect by making predictions about what could happen and reflect upon what caused something to happen. (California Department of Education [CDE] 2005)

From: cde.ca.gov. (California Department of Education)

Naïve Physical Action-Effect Prediction

Action to Effect

Action
(squeeze-bottle)

Naïve Physical Action-Effect Prediction

Action to Effect

Action
(squeeze-bottle)

Naïve Physical Action-Effect Prediction

Effect to Action

Action (peel-carrot) ?

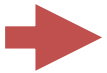
Action (juice-carrot) ?

Action (grate-carrot) ?

Action (chop-carrot) ?

Naïve Physical Action-Effect Prediction

Effect to Action



Action (peel-carrot) x	Action (juice-carrot) x
Action (grate-carrot) x	Action (chop-carrot) ✓

Related Work

- The **NLP** community
  - Most existing studies focus on the causal relations between high-level events. E.g., "the collapse of the housing bubble" causes the effect of "stock prices to fall". (Yang and Mao, 2014; Sharp et al., 2016)
  - This paper studies the basic cause-effect knowledge related to concrete actions and their effects to the world.
- Recent advances in **Computer Vision** and **Robotics**
  - Object physical state prediction (Zhou and Berg, 2016; Wu et al., 2017)
  - Action recognition through detection of state changes (Yang et al., 2013)
  - Robot following natural language commands (She et al, 2014; Misra et al., 2015)

This Work

- Introduce a new task on physical action-effect prediction and create a dataset for this task.
  - Data collection and analysis
- Propose an approach that harnesses the large amount of image data available on the web with minimum supervision.
  - Web images acquisition
  - Bootstrapping strategy
- Automatic prediction of effect knowledge for novel actions.

Action-Effect Data

- **Actions** (Verb-Noun Pairs)
  - 140 verb-noun pairs
  - 62 unique verbs (e.g., bend, boil, chop, crack, fold, grind, ignite, kick, peel, soak, trim)
  - 39 unique nouns (e.g., apple, baseball, book, car, chair, cup, flower, orange, shoe)
- **Effects**
  - Effects described in **language**
  - Effects depicted by **images**

Effects Described in Language

- Action effect is often presupposed in our communication and not explicitly stated.
- Crowd-sourcing data collection
  - Workers were shown a verb-noun pair, and were asked to describe what changes might occur to the object as a result of the action.
  - 1400 effect descriptions (10 for each verb-noun pair)
  - Examples:

Action	Effect Text
ignite paper	The paper is on fire.
soak shirt	The shirt is thoroughly wet.
fry potato	The potatoes become crisp and golden.
stain shirt	There is a visible mark on the shirt.

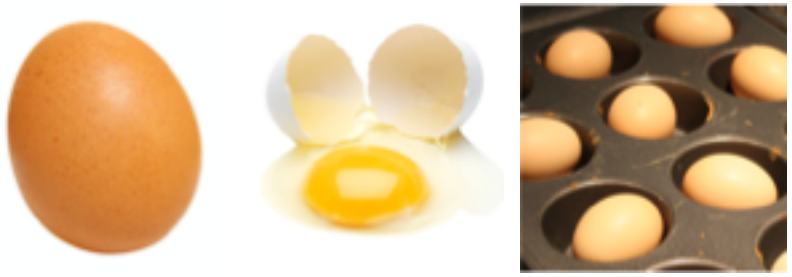
Effects Depicted by Images

- Human labeled image set: 4163 images (Data available on the project webpage.)
  - **Positive** images are those capturing the resulting world state of the action.
  - **Negative** images are those deemed to capture some state of the related nouns, but are not the resulting state of the corresponding action.

Action: Fry-Egg

Positive:

Negative:



Web Search Images

- Searching **keywords**: phrases extracted from language effect descriptions
  - Phrases were extracted using syntactic patterns:

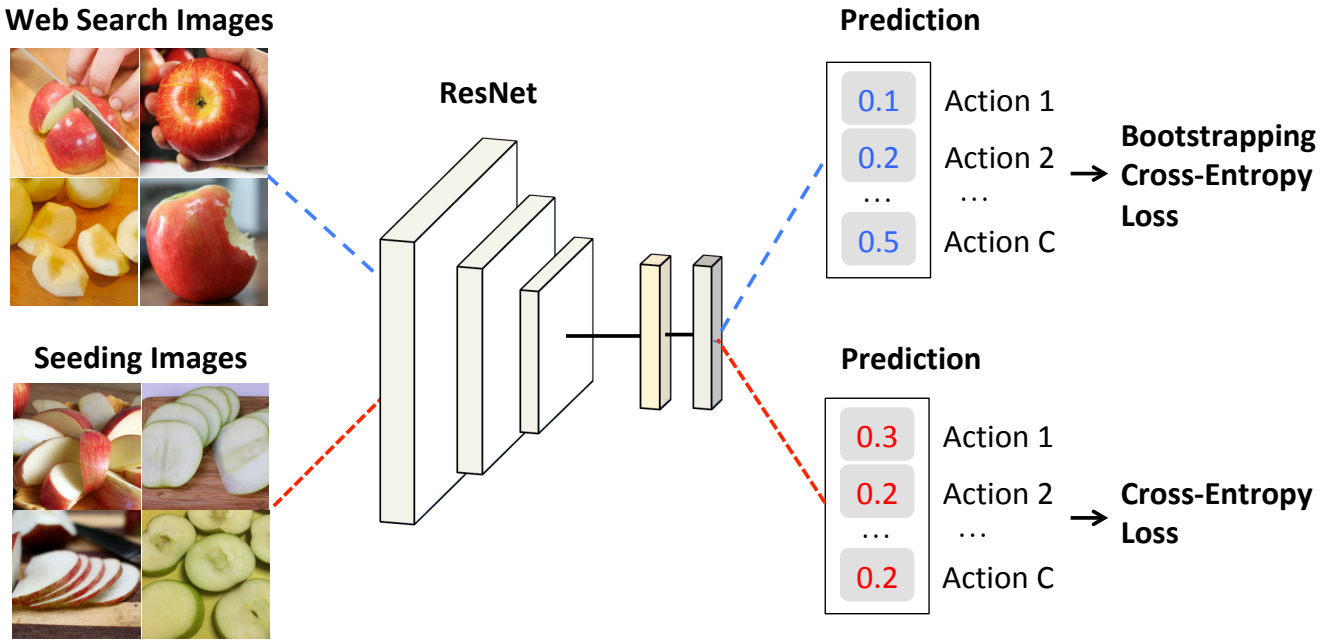
Example patterns	Example *Effect Phrases* (bold) extracted from effect descriptions
VP with a verb ∈ {be, become, turn, get}	The ship **is destroyed**.
VP + PRT	The wall is **knocked off**.
VP + ADVP	The door **swings forward**.
ADJP	The window would begin to get **clean**.

book

book is on fire

book is set aflame

Bootstrapping Approach



Cross-entropy loss: $$\mathcal{L}(\mathbf{t}, \mathbf{q}) = \sum_{i=1}^{C} t_i \log(q_i)$$

Bootstrapping cross-entropy loss: $$\mathcal{L}(\mathbf{t}', \mathbf{q}) = \sum_{i=1}^{C} [\beta t'_i + (1-\beta) z_i] \log(q_i)$$

(Reed et al., 2014)

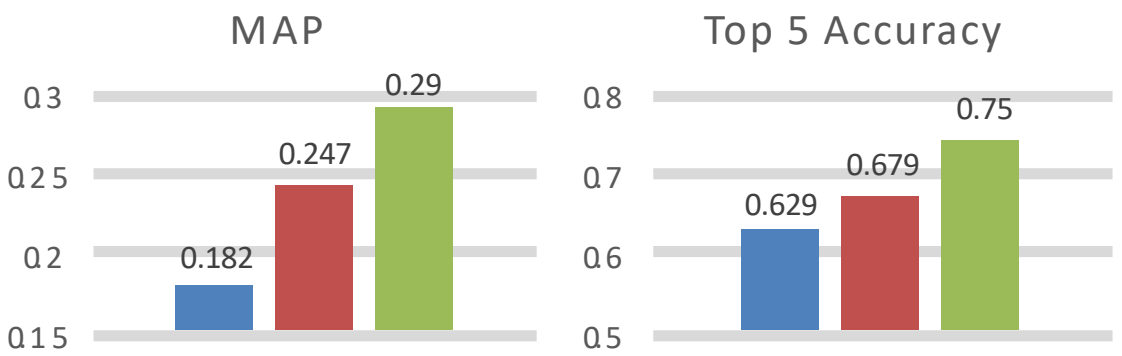
Evaluations

- **Human annotated image data**: use 10% as seeding images (training), 30% for development and 60% for test.
  - On average, each verb-noun pair only has 3 seeding images
- **Web search images**: over 60,000 images were downloaded using around 2,000 effect phrases as searching keywords.
- **Methods for comparison**
  - *Seed*
  - *Seed+Act+Eff*
  - *BS+Seed+Act+Eff*

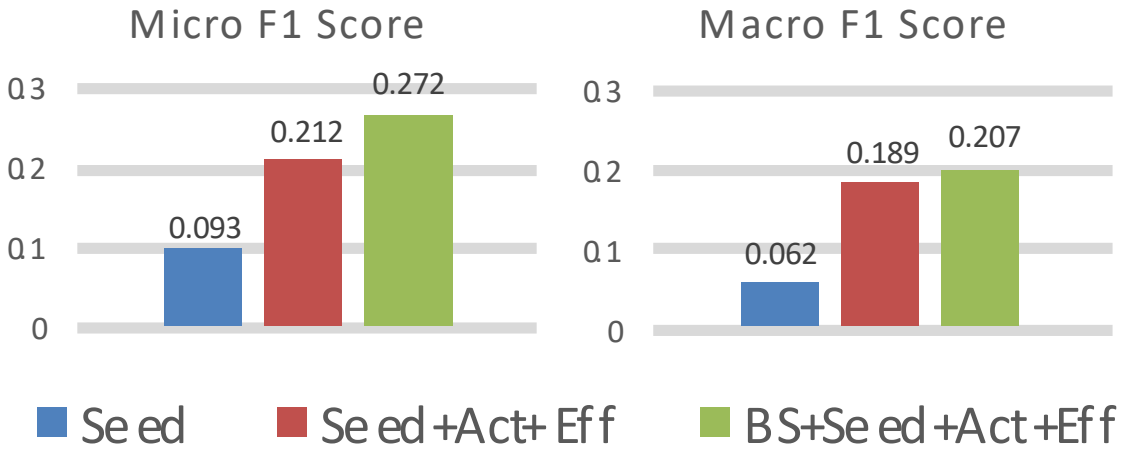
BS: bootstrapping approach; Seed: seed images;
Act: web images downloaded using verb-noun as keywords;
Eff: web images downloaded using effect phrases as keywords.

Evaluation Results

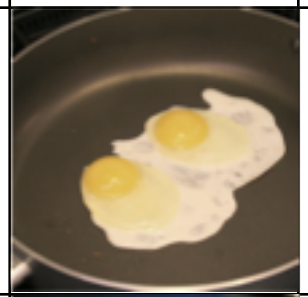
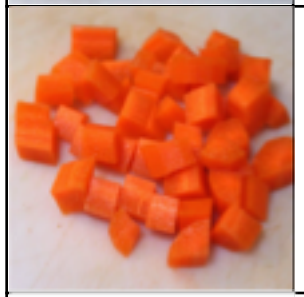
Action to Effect:



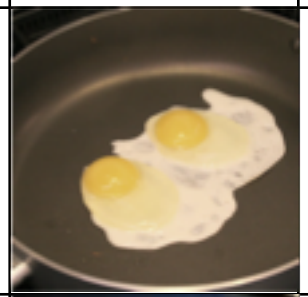
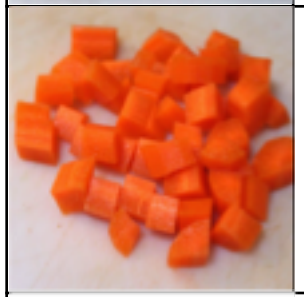
Effect to Action:



Examples

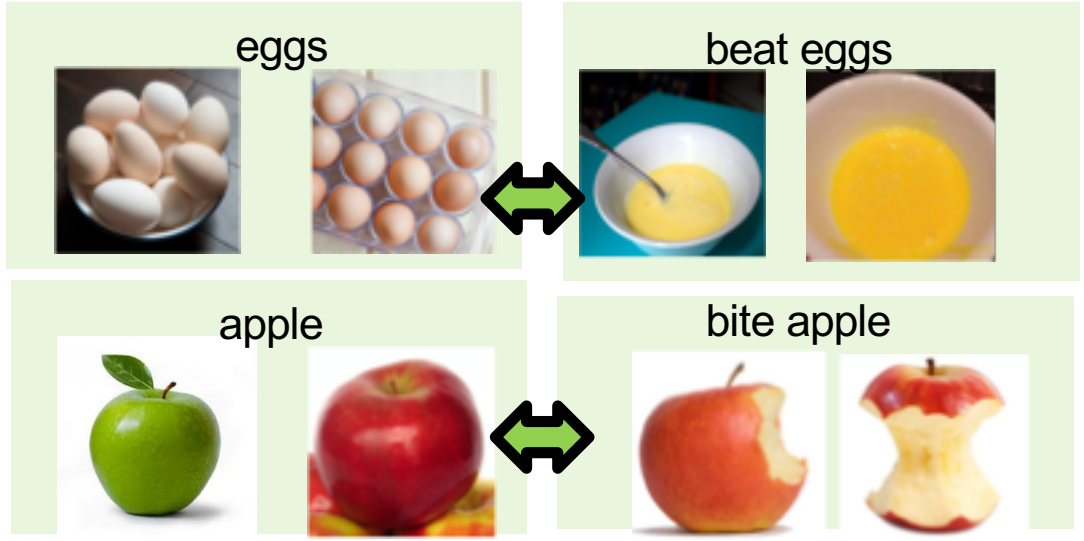
	Top Action Predictions			Top Action Predictions	
	**bite apple** background cut apple peel apple			**fry egg** background crack egg mix eggs	
	background **chop carrot** grate carrot peel carrot			background **insert key** close drawer fasten door	
	**background** cut potato fry potato mash potato			**pile books** background wrap book roll paper	

Examples

	Top Action Predictions	Top Effect Predictions		Top Action Predictions	Top Effect Predictions
	**bite apple** background cut apple peel apple	apple is eaten apple is being cut apple is chewed apple in tiny pieces		**fry egg** background crack egg mix eggs	egg into a harder substance cup into smaller pieces egg edible
	background **chop carrot** grate carrot peel carrot	carrot into tiny pieces carrot is being cut carrot into many smaller pieces		background **insert key** close drawer fasten door	key in the keyhole drawer without a key door is locked door is being bolted
	**background** cut potato fry potato mash potato	potato into a pot potato is being sliced potato for potato edible		**pile books** background wrap book roll paper	books in a stack book on books in a large stack books in a pile

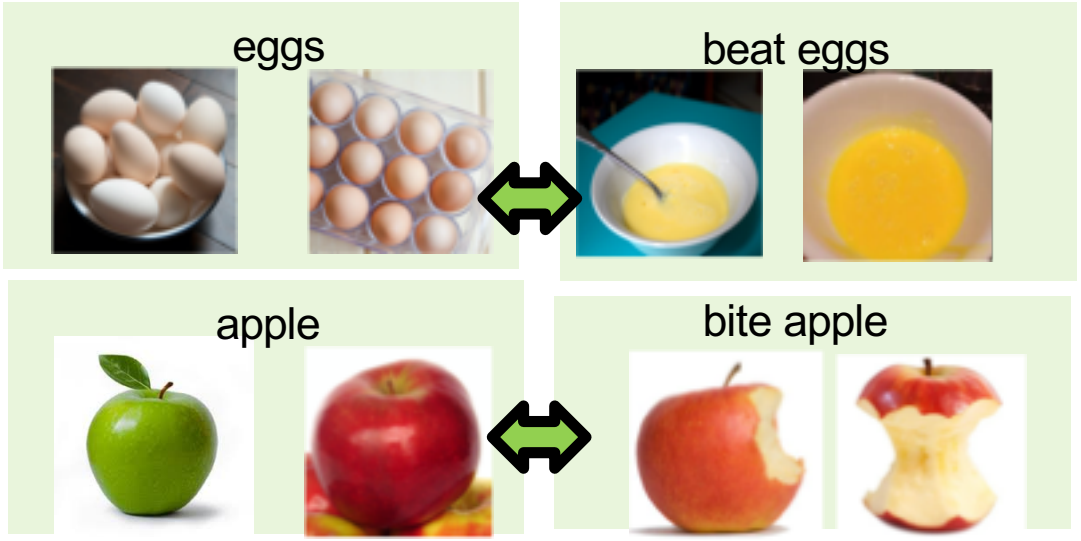
Examples

Action	AP
beat eggs	0.783
pile boxes	0.766
bite apple	0.484
slice onion	0.470

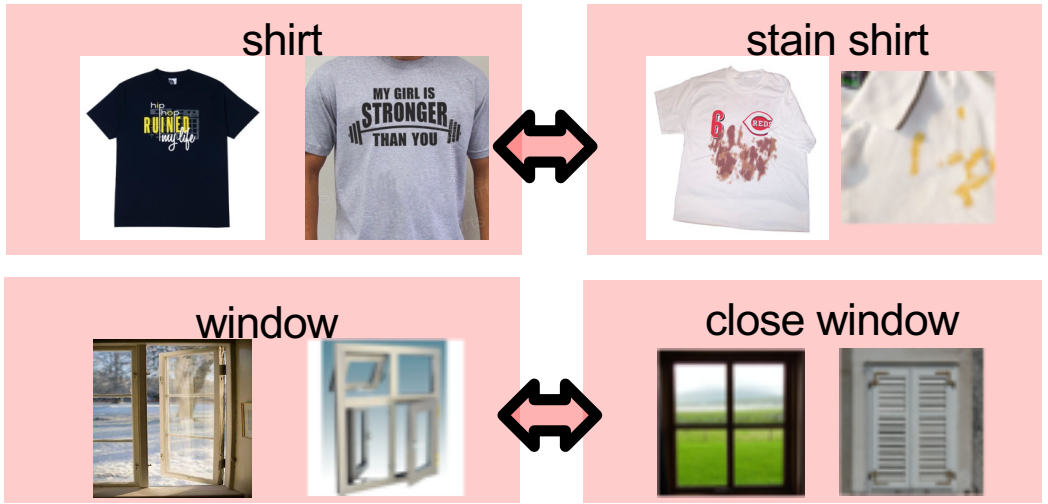


Examples

Action	AP
beat eggs	0.783
pile boxes	0.766
bite apple	0.484
slice onion	0.470

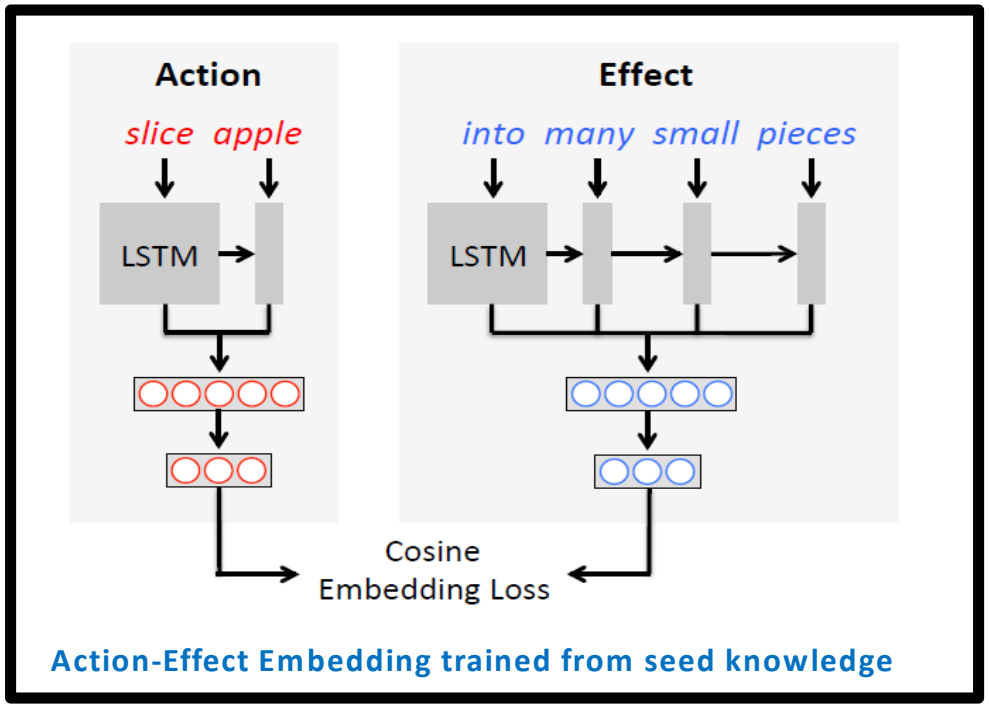


Action	AP
crack glass	0.047
lock drawer	0.037
stain shirt	0.023
close window	0.087



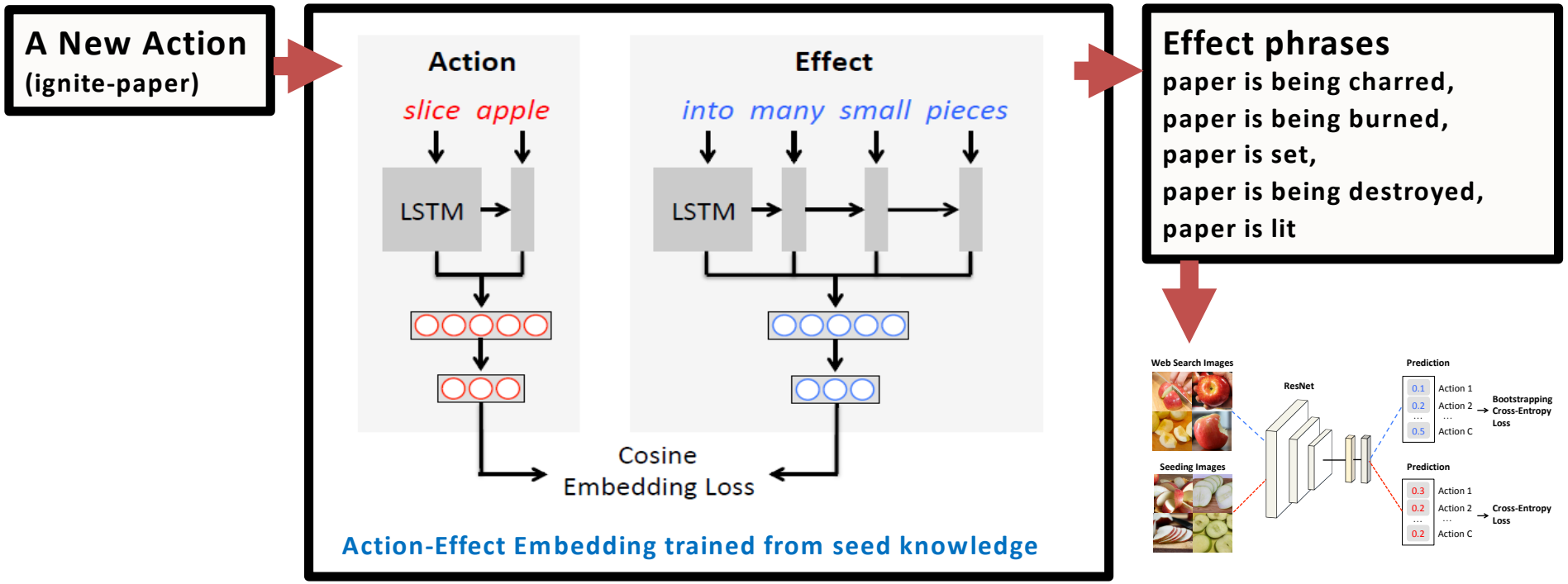
Handling Unseen Verb-Noun Pairs

- Generalize effect knowledge to new verb-noun pairs through an embedding model.



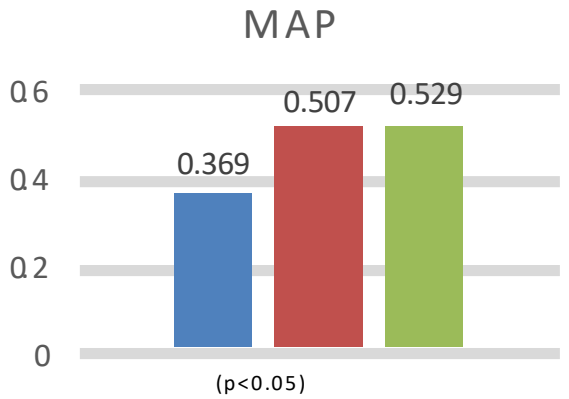
Handling Unseen Verb-Noun Pairs

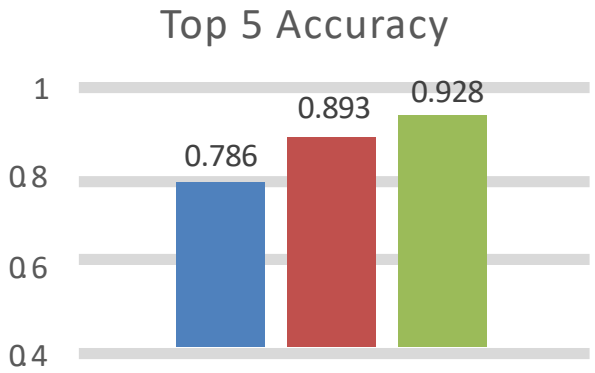
- Generalize effect knowledge to new verb-noun pairs through an embedding model.



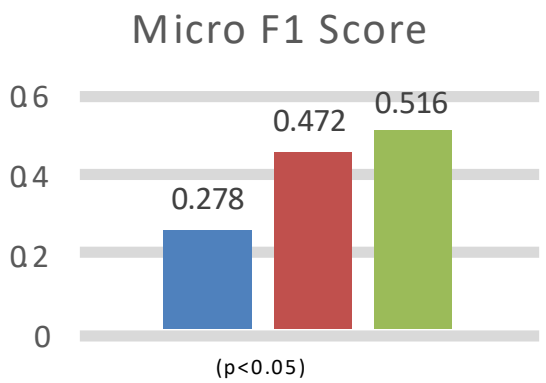
Evaluation Results

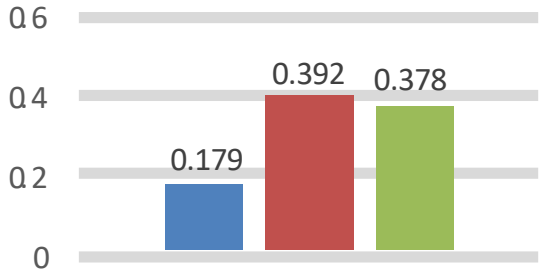
Action to Effect:





Effect to Action:

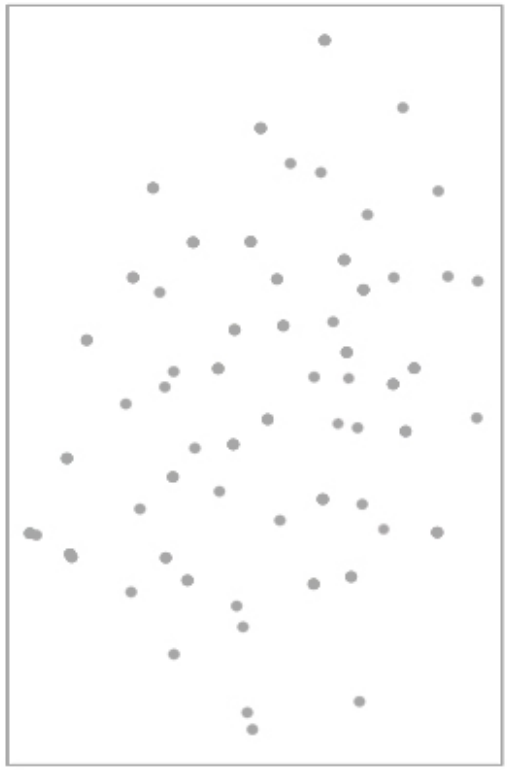




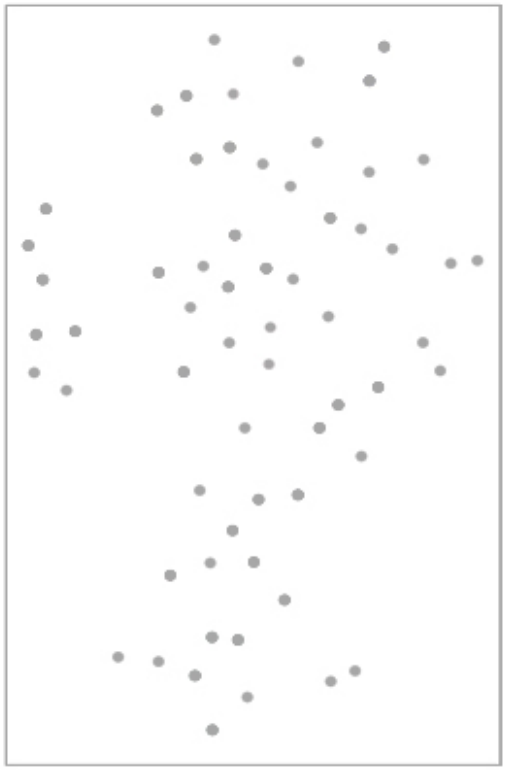
pEff: web images downloaded using the predicted effect phrases.

Action-Effect Embedding Space

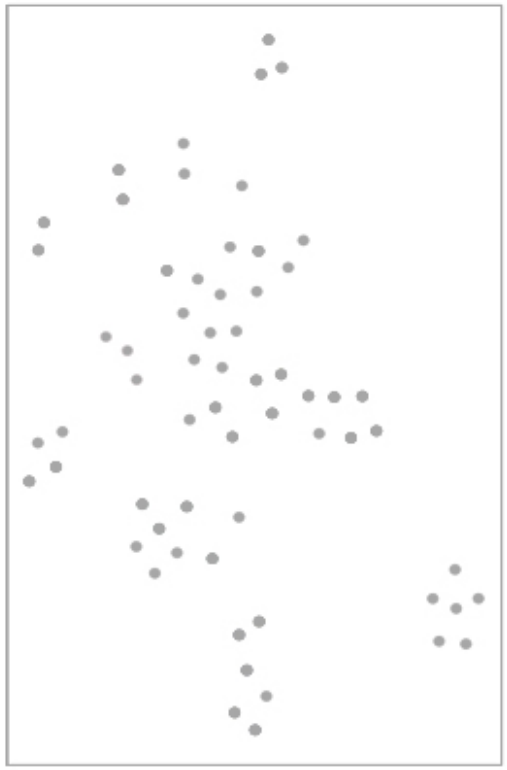
GloVe Verb



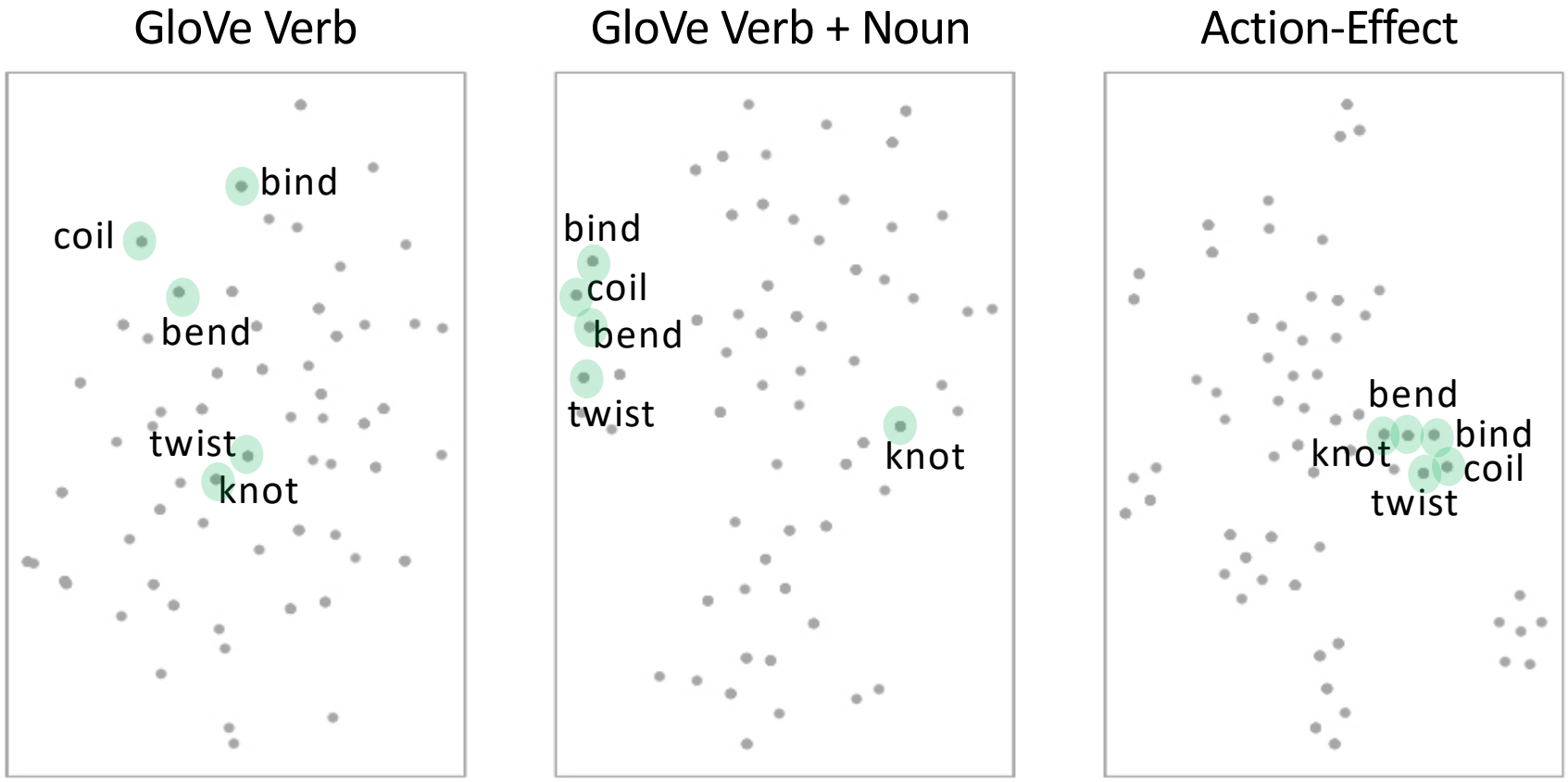
GloVe Verb + Noun



Action-Effect

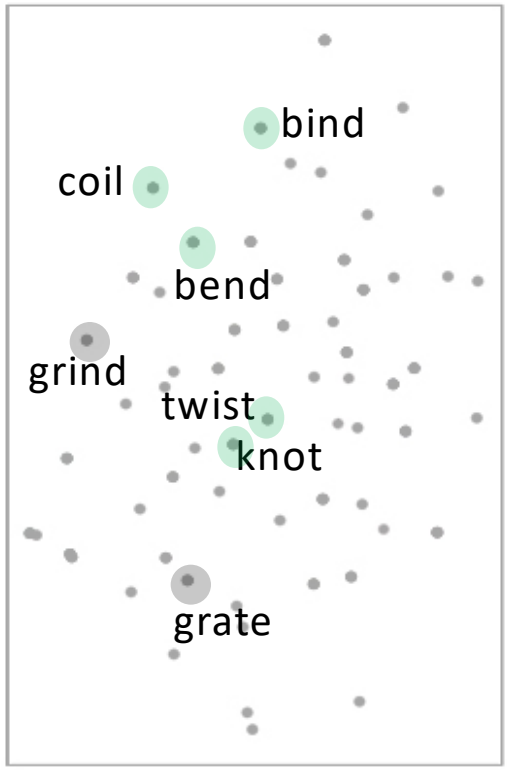


Action-Effect Embedding Space

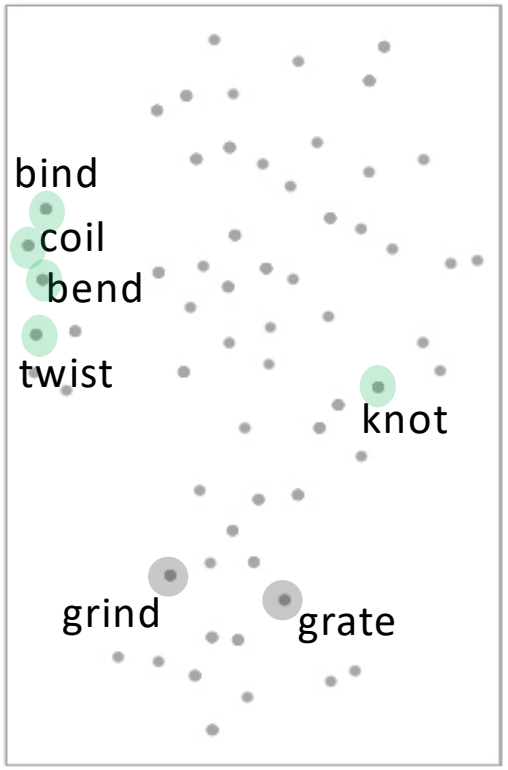


Action-Effect Embedding Space

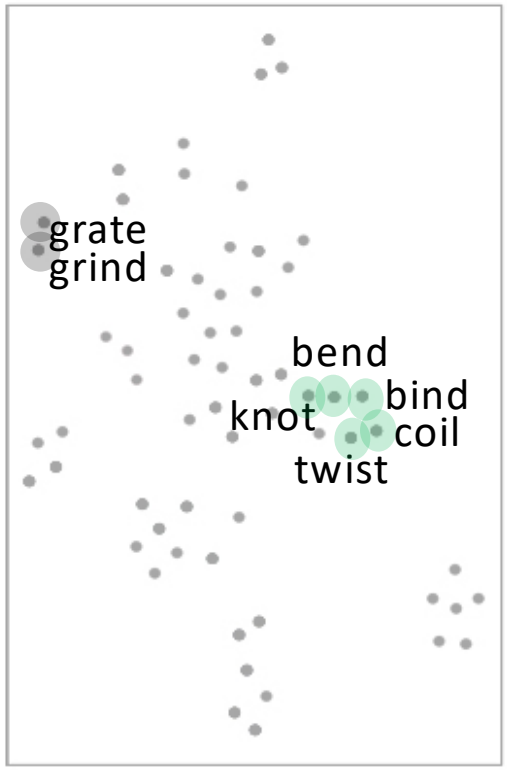
GloVe Verb



GloVe Verb + Noun

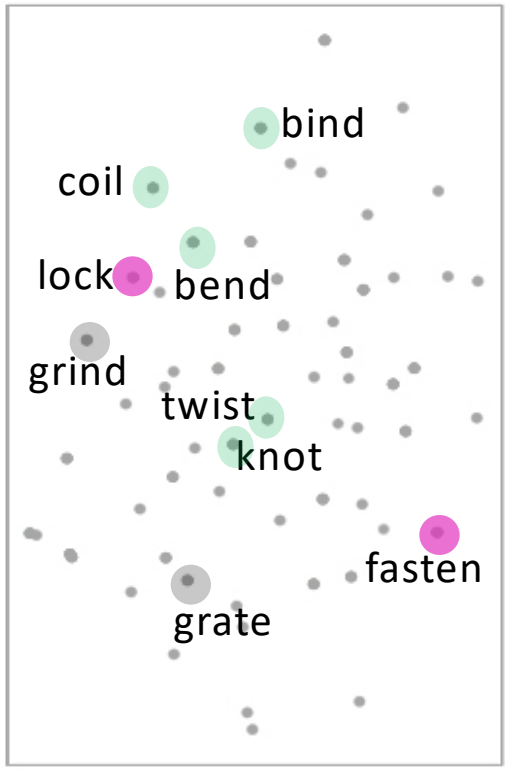


Action-Effect

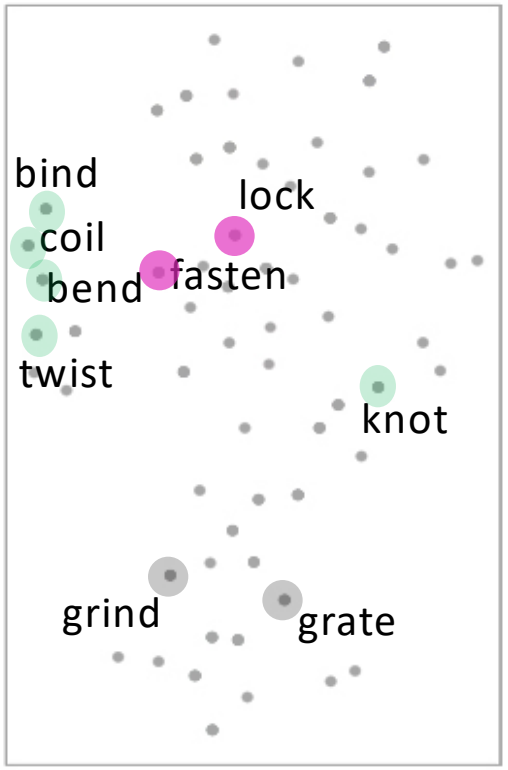


Action-Effect Embedding Space

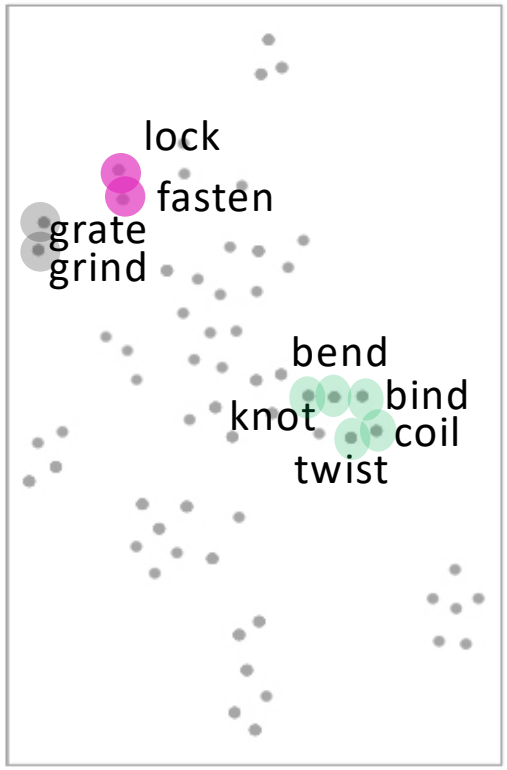
GloVe Verb



GloVe Verb + Noun

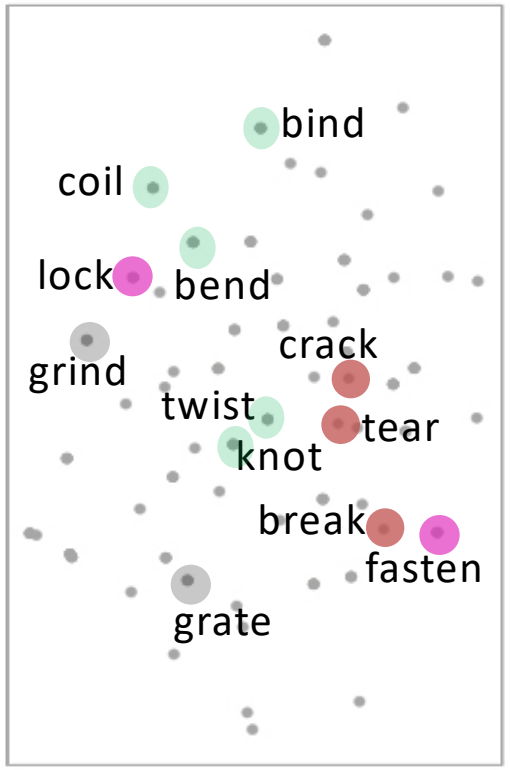


Action-Effect

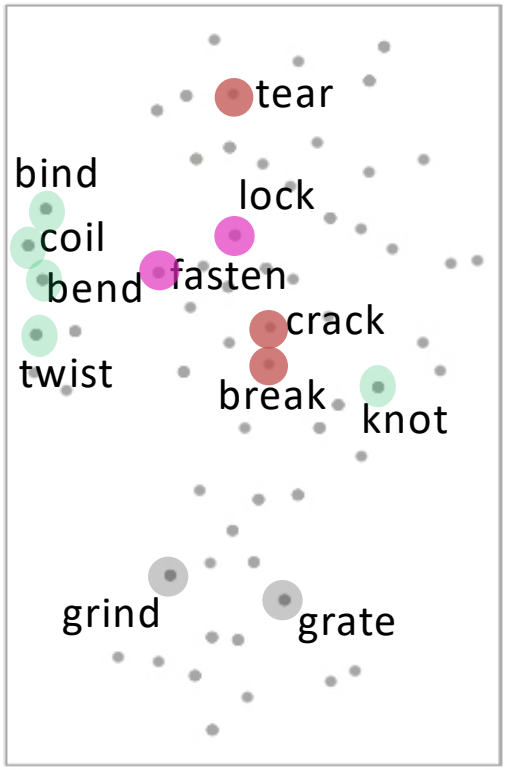


Action-Effect Embedding Space

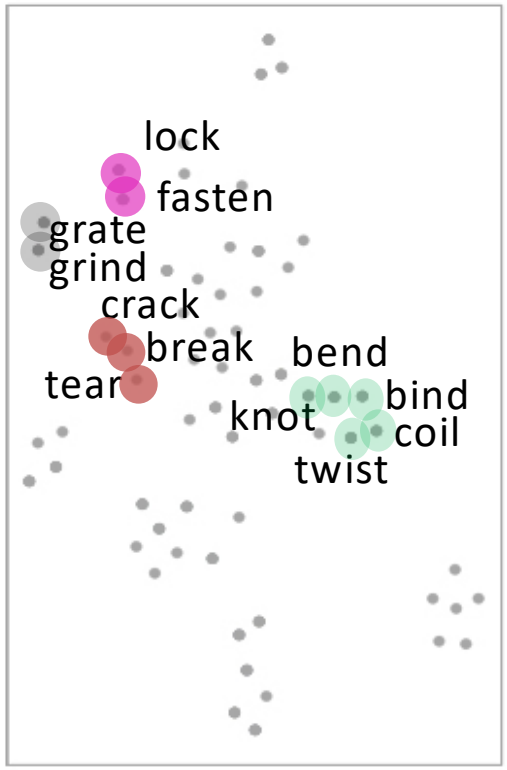
GloVe Verb



GloVe Verb + Noun

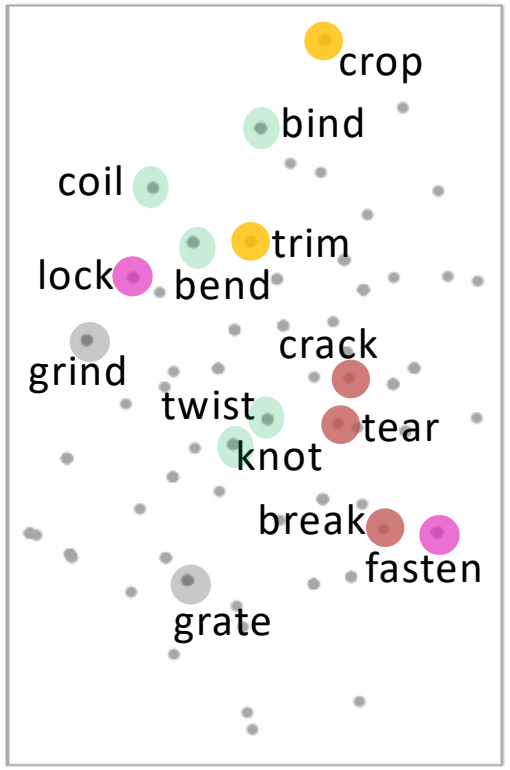


Action-Effect

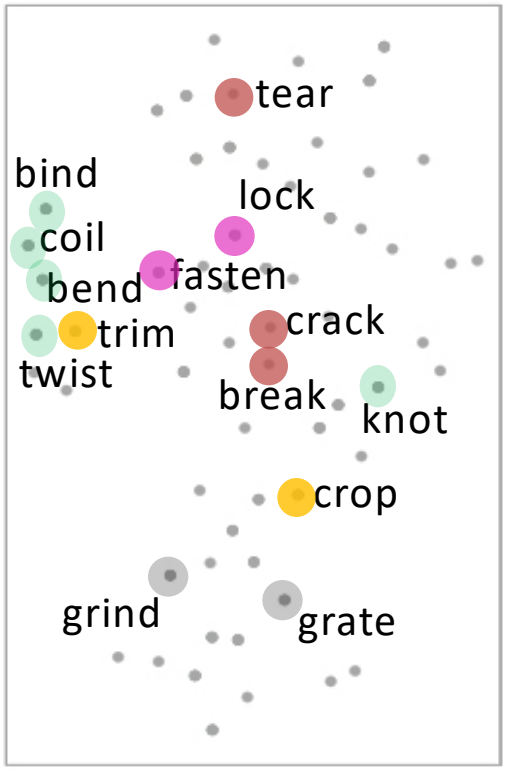


Action-Effect Embedding Space

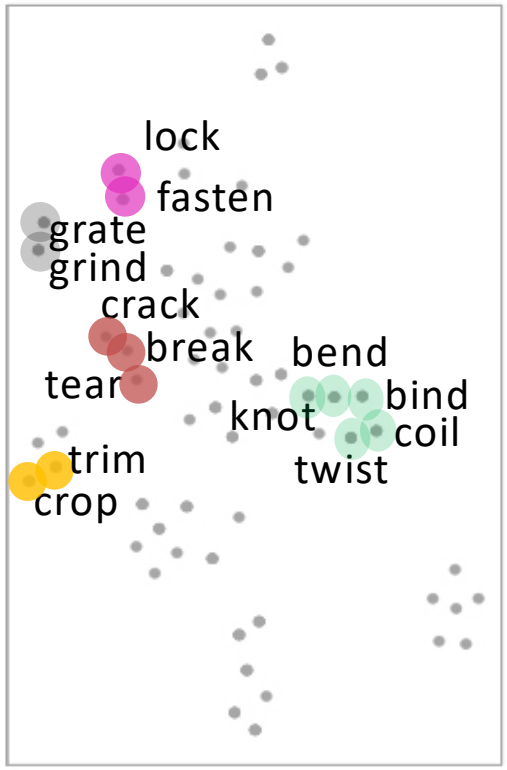
GloVe Verb



GloVe Verb + Noun

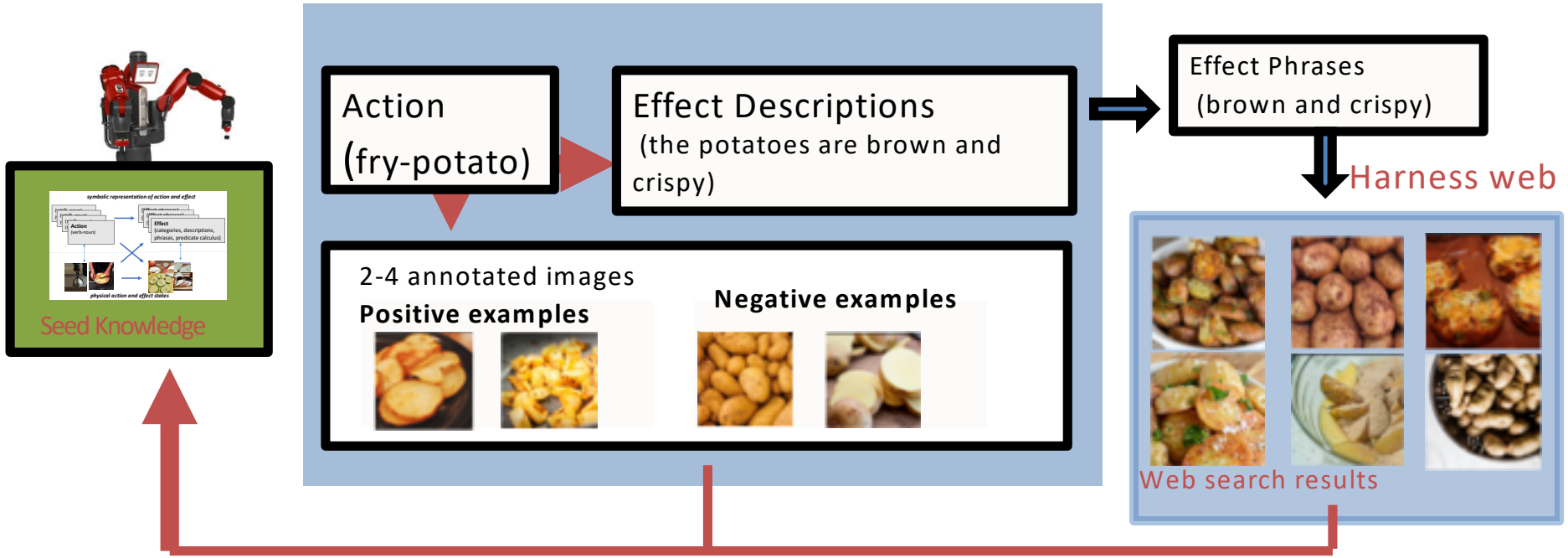


Action-Effect

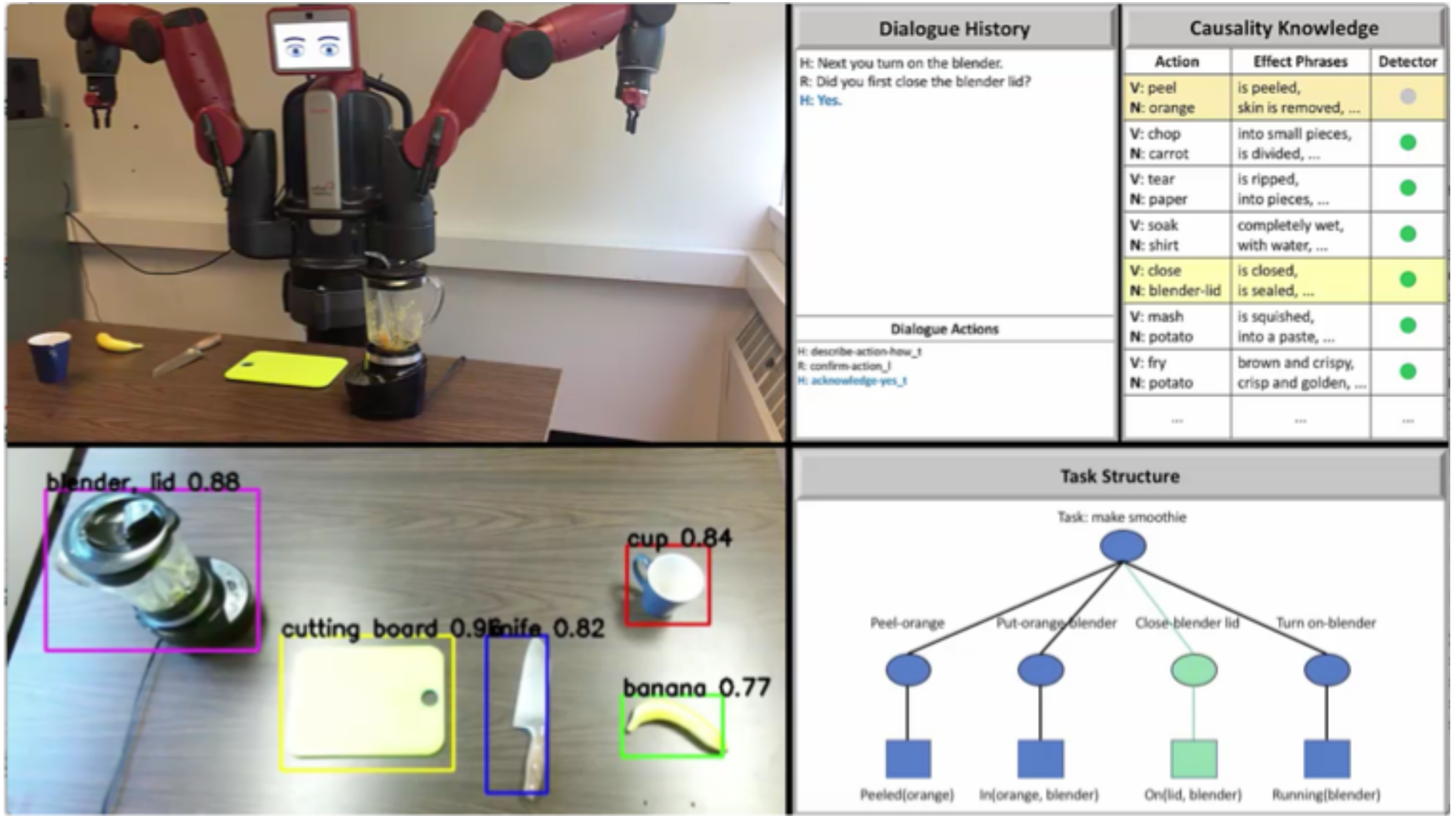


Learning from a few examples

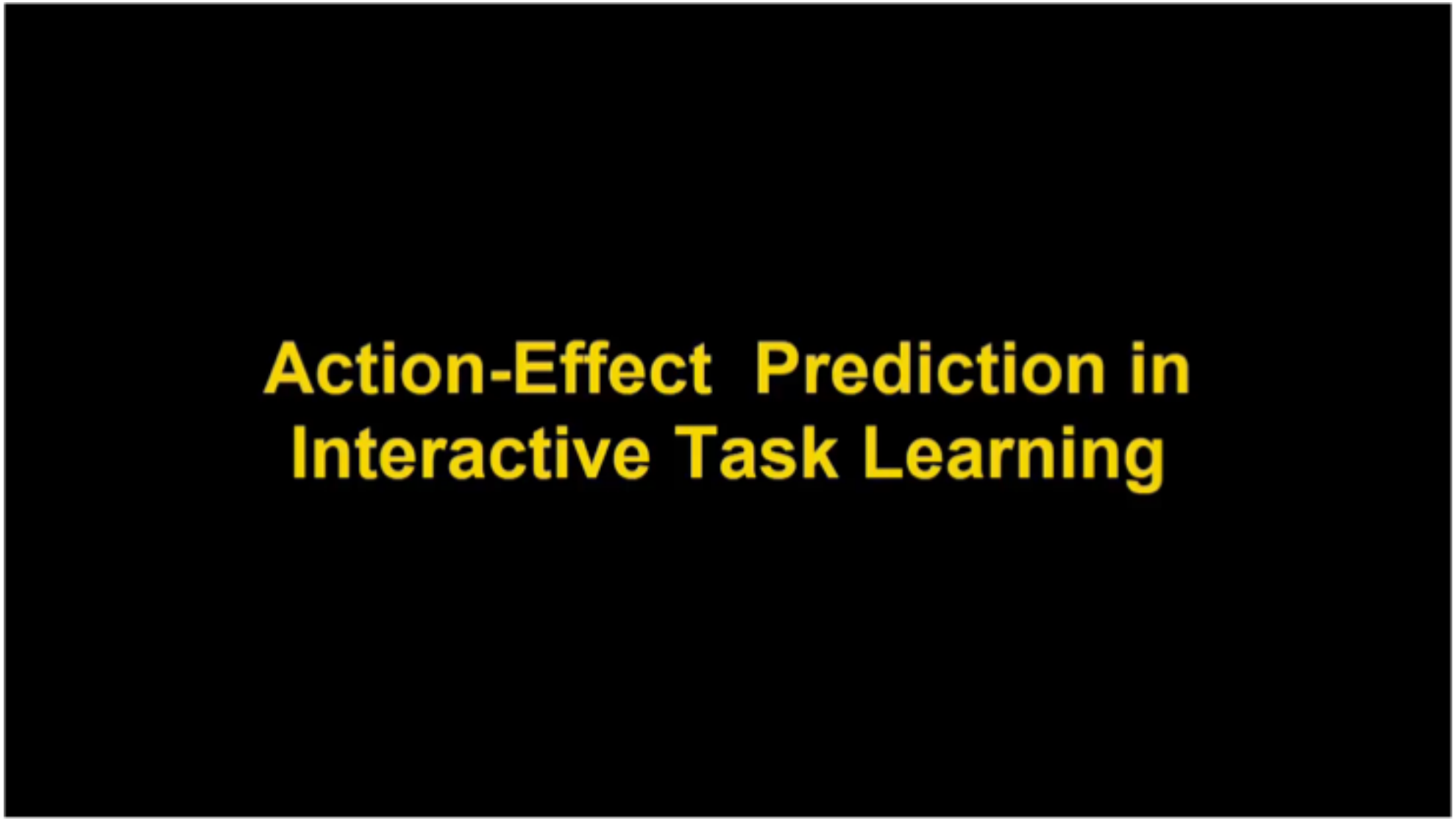
Goal: learn from a few examples to make it possible for humans to teach agents for tasks at hand.



Action-Effect Prediction in Interactive Task Learning



Action-Effect Prediction in Interactive Task Learning



Summary

- Presented an initial investigation on action-effect prediction.
- Explored method using web image data to facilitate the training of action-effect prediction models.
- Explored using semantic embedding space to extend effect knowledge to new verb-noun pairs.
- Future Directions
  - Develop better models to improve task performance
  - Extend action-effect prediction to video data

Thank you !