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

Qiaozì Gào¹, Shaohua Yang¹, Joyce Y. Chai¹, Lucy Vanderwende²

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

<table>
<thead>
<tr>
<th>8 months</th>
<th>18 months</th>
<th>36 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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.</td>
<td>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)</td>
</tr>
</tbody>
</table>

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)

- Correct: Bottle intact, no effect.
- Incorrect: Bottle broken, effect observed.
- Incorrect: Bottle intact, no effect.
- Correct: Bottle intact, no effect.
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:

<table>
<thead>
<tr>
<th>Action</th>
<th>Effect Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>ignite paper</td>
<td>The paper is on fire.</td>
</tr>
<tr>
<td>soak shirt</td>
<td>The shirt is thoroughly wet.</td>
</tr>
<tr>
<td>fry potato</td>
<td>The potatoes become crisp and golden.</td>
</tr>
<tr>
<td>stain shirt</td>
<td>There is a visible mark on the shirt.</td>
</tr>
</tbody>
</table>
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**

<table>
<thead>
<tr>
<th>Positive:</th>
<th>Negative:</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Fried Egg" /></td>
<td><img src="image2" alt="Raw Egg" /></td>
</tr>
<tr>
<td><img src="image3" alt="Fried Egg" /></td>
<td><img src="image4" alt="Raw Egg" /></td>
</tr>
<tr>
<td><img src="image5" alt="Fried Egg" /></td>
<td><img src="image6" alt="Raw Egg" /></td>
</tr>
</tbody>
</table>
Web Search Images

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

<table>
<thead>
<tr>
<th>Example patterns</th>
<th>Example Effect Phrases (bold) extracted from effect descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP with a verb ∈ {be, become, turn, get}</td>
<td>The ship is destroyed.</td>
</tr>
<tr>
<td>VP + PRT</td>
<td>The wall is <strong>knocked off</strong>.</td>
</tr>
<tr>
<td>VP + ADVP</td>
<td>The door <strong>swings forward</strong>.</td>
</tr>
<tr>
<td>ADJP</td>
<td>The window would begin to get <strong>clean</strong>.</td>
</tr>
</tbody>
</table>

- book
- book is on fire
- book is set aflame
Bootstrapping Approach

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

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

(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

<table>
<thead>
<tr>
<th>Top Action Predictions</th>
<th>Top Action Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bite apple</strong>&lt;br&gt;background&lt;br&gt;cut apple&lt;br&gt;peel apple</td>
<td><strong>fry egg</strong>&lt;br&gt;background&lt;br&gt;crack egg&lt;br&gt;mix eggs</td>
</tr>
<tr>
<td>background&lt;br&gt;<strong>chop carrot</strong>&lt;br&gt;grate carrot&lt;br&gt;peel carrot</td>
<td>background&lt;br&gt;<strong>insert key</strong>&lt;br&gt;close drawer&lt;br&gt;fasten door</td>
</tr>
<tr>
<td>background&lt;br&gt;<strong>cut potato</strong>&lt;br&gt;fry potato&lt;br&gt;<strong>mash potato</strong></td>
<td><strong>pile books</strong>&lt;br&gt;background&lt;br&gt;wrap book&lt;br&gt;roll paper</td>
</tr>
</tbody>
</table>
# Examples

<table>
<thead>
<tr>
<th>Top Action Predictions</th>
<th>Top Effect Predictions</th>
<th>Top Action Predictions</th>
<th>Top Effect Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bite apple</strong>&lt;br&gt;background&lt;br&gt;cut apple&lt;br&gt;peel apple</td>
<td>apple is eaten&lt;br&gt;apple is being cut&lt;br&gt;apple is chewed&lt;br&gt;apple in tiny pieces</td>
<td><strong>fry egg</strong>&lt;br&gt;background&lt;br&gt;crack egg&lt;br&gt;mix eggs</td>
<td>egg into a harder substance&lt;br&gt;cup into smaller pieces&lt;br&gt;egg edible</td>
</tr>
<tr>
<td>background&lt;br&gt;<strong>chop carrot</strong>&lt;br&gt;grate carrot&lt;br&gt;peel carrot</td>
<td>carrot into tiny pieces&lt;br&gt;carrot is being cut&lt;br&gt;carrot into many smaller pieces</td>
<td>background&lt;br&gt;<strong>insert key</strong>&lt;br&gt;close drawer&lt;br&gt;fasten door</td>
<td>key in the keyhole&lt;br&gt;drawer without a key&lt;br&gt;door is locked&lt;br&gt;door is being bolted</td>
</tr>
<tr>
<td>background&lt;br&gt;<strong>cut potato</strong>&lt;br&gt;fry potato&lt;br&gt;mash potato</td>
<td>potato into a pot&lt;br&gt;potato is being sliced&lt;br&gt;potato for potato edible</td>
<td>background&lt;br&gt;<strong>pile books</strong>&lt;br&gt;wrap book&lt;br&gt;roll paper</td>
<td>books in a stack&lt;br&gt;book on&lt;br&gt;books in a large stack&lt;br&gt;books in a pile</td>
</tr>
</tbody>
</table>
### Examples

<table>
<thead>
<tr>
<th>Action</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>beat eggs</td>
<td>0.783</td>
</tr>
<tr>
<td>pile boxes</td>
<td>0.766</td>
</tr>
<tr>
<td>bite apple</td>
<td>0.484</td>
</tr>
<tr>
<td>slice onion</td>
<td>0.470</td>
</tr>
</tbody>
</table>

- **eggs**
- **beat eggs**
- **apple**
- **bite apple**
### Examples

<table>
<thead>
<tr>
<th>Action</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>beat eggs</td>
<td>0.783</td>
</tr>
<tr>
<td>pile boxes</td>
<td>0.766</td>
</tr>
<tr>
<td>bite apple</td>
<td>0.484</td>
</tr>
<tr>
<td>slice onion</td>
<td>0.470</td>
</tr>
</tbody>
</table>

- **Eggs**: beat eggs → eggs
- **Apple**: bite apple → apple
- **Shirt**: stain shirt → stain
- **Window**: close window → close

<table>
<thead>
<tr>
<th>Action</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>crack glass</td>
<td>0.047</td>
</tr>
<tr>
<td>lock drawer</td>
<td>0.037</td>
</tr>
<tr>
<td>stain shirt</td>
<td>0.023</td>
</tr>
<tr>
<td>close window</td>
<td>0.087</td>
</tr>
</tbody>
</table>

- **Eggs**: beat eggs → eggs
- **Apple**: bite apple → apple
- **Shirt**: stain shirt → stain
- **Window**: close window → close
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.

**A New Action** (ignite-paper)

**Effect phrases**
- paper is being charred,
- paper is being burned,
- paper is set,
- paper is being destroyed,
- paper is lit

**Action-Effect Embedding trained from seed knowledge**
Evaluation Results

Action to Effect:

<table>
<thead>
<tr>
<th>MAP</th>
<th>Top 5 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.369</td>
<td>0.786</td>
</tr>
<tr>
<td>0.507</td>
<td>0.893</td>
</tr>
<tr>
<td>0.529</td>
<td>0.928</td>
</tr>
</tbody>
</table>

Effect to Action:

<table>
<thead>
<tr>
<th>Micro F1 Score</th>
<th>Macro F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.278</td>
<td>0.179</td>
</tr>
<tr>
<td>0.472</td>
<td>0.392</td>
</tr>
<tr>
<td>0.516</td>
<td>0.378</td>
</tr>
</tbody>
</table>

pEff: web images downloaded using the predicted effect phrases.
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

- bind
- coil
- bend
- twist
- knot

GloVe Verb + Noun

- bind
- coil
- bend
- twist
- knot
- grate
- grind

Action-Effect

- grate
- grind
- bend
- knot
- coil
- bind
- twist
- grate
- grind
Action-Effect Embedding Space

GloVe Verb

GloVe Verb + Noun

Action-Effect

bind
coil
bend
twist
knot
bind
coil
twist
knot
bind
twist
knot
bind
twist
knot
bind
twist
knot

grind
grate
grind
grate
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
lock
fasten
Action-Effect Embedding Space

GloVe Verb

GloVe Verb + Noun

Action-Effect
Action-Effect Embedding Space

GloVe Verb

GloVe Verb + Noun

Action-Effect

- grind
- grate
- grind
- grate
- lock
- fasten
- fasten
- lock
- crop
- crop
- trim
- trim
- crack
- crack
- break
- break
- tear
- tear
- crop
- crop
- trim
- trim
- twist
- twist
- coil
- coil
- bind
- bind
- bend
- bend
- knot
- knot
- twist
- twist
- break
- break
- tear
- tear
- trim
- trim
- twist
- twist
- coil
- coil
- bind
- bind
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** (fry-potato)
- **Effect Descriptions** (the potatoes are brown and crispy)
- **2-4 annotated images**
  - Positive examples
  - Negative examples

- **Harness web**
  - Effect Phrases (brown and crispy)
  - Web search results
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 !