

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

Qiaoz Gao[†] Shaohua Yang[†] Joyce Y. Chai[†] Lucy Vanderwende[‡]

[†]Department of Computer Science and Engineering, Michigan State University

[‡]Microsoft Research, Redmond, Washington

{gaoqiaoz, yangshao, jchai}@msu.edu

lucy_vanderwende@live.com

Abstract

Despite recent advances in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the ability to understand basic action-effect relations regarding the physical world, for example, the action of cutting a cucumber most likely leads to the state where the cucumber is broken apart into smaller pieces. If artificial agents (e.g., robots) ever become our partners in joint tasks, it is critical to empower them with such action-effect understanding so that they can reason about the state of the world and plan for actions. Towards this goal, this paper introduces a new task on naive physical action-effect prediction, which addresses the relations between concrete actions (expressed in the form of verb-noun pairs) and their effects on the state of the physical world as depicted by images. We collected a dataset for this task and developed an approach that harnesses web image data through distant supervision to facilitate learning for action-effect prediction. Our empirical results have shown that web data can be used to complement a small number of seed examples (e.g., three examples for each action) for model learning. This opens up possibilities for agents to learn physical action-effect relations for tasks at hand through communication with humans with a few examples.

1 Introduction

Causation in the physical world has long been a central discussion to philosophers who study causal reasoning and explanation (Ducasse, 1926; Gopnik et al., 2007), to mathematicians or com-

puter scientists who apply computational approaches to model cause-effect prediction (Pearl et al., 2009), and to domain experts (e.g., medical doctors) who attempt to understand the underlying cause-effect relations (e.g., disease and symptoms) for their particular inquiries. Apart from this wide range of topics, this paper investigates a specific kind of causation, the very basic causal relations between a concrete action (expressed in the form of a verb-noun pair such as “cut-cucumber”) and the change of the physical state caused by this action. We call such relations *naive* physical action-effect relations.

For example, given an image as shown in Figure 1, we would have no problem predicting what actions can cause the state of the world depicted in the image, e.g., slicing an apple will likely lead to the state. On the other hand, given a statement “slice an apple”, it would not be hard for us to imagine what state change may happen to the apple. We can make such action-effect prediction because we have developed an understanding of this kind of basic action-effect relations at a very young age (Baillargeon, 2004). What happens to machines? Will artificial agents be able to make the same kind of predictions? The answer is not yet.

Despite tremendous progress in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the understanding of naive causal relations regarding the physical world. This is one of the bottlenecks in machine intelligence. If artificial agents ever become capable of working with humans as partners, they will need to have this kind of physical action-effect understanding to help them reason, learn, and perform actions.

To address this problem, this paper introduces a new task on naive physical action-effect prediction. This task supports both *cause* predic-



Figure 1: Images showing the effects of “slice an apple”.

tion: given an image which describes a state of the world, identify the most likely action (in the form of a verb-noun pair, from a set of candidates) that can result in that state; and *effect prediction*: given an action in the form of a verb-noun pair, identify images (from a set of candidates) that depicts the most likely effects on the state of the world caused by that action. Note that there could be different ways to formulate this problem, for example, both causes and effects are in the form of language or in the form of images/videos. Here we intentionally frame the action as a language expression (i.e., a verb-noun pair) and the effect as depicted in an image in order to make a connection between language and perception. This connection is important for physical agents that not only can perceive and act, but also can communicate with humans in language.

As a first step, we collected a dataset of 140 verb-noun pairs. Each verb-noun pair is annotated with possible effects described in language and depicted in images (where language descriptions and image descriptions are collected separately). We have developed an approach that applies distant supervision to harness web data for bootstrapping action-effect prediction models. Our empirical results have shown that, using a simple bootstrapping strategy, our approach can combine the noisy web data with a small number of seed examples to improve action-effect prediction. In addition, for a new verb-noun pair, our approach can infer its effect descriptions and predict action-effect relations only based on 3 image examples.

The contributions of this paper are three folds. First, it introduces a new task on physical action-effect prediction, a first step towards an under-

standing of causal relations between physical actions and the state of the physical world. Such ability is central to robots which not only perceive from the environment, but also act to the environment through planning. To our knowledge, there is no prior work that attempts to connect actions (in language) and effects (in images) in this nature. Second, our approach harnesses the large amount of image data available on the web with minimum supervision. It has shown that physical action-effect models can be learned through a combination of a few annotated examples and a large amount of un-annotated web data. This opens up the possibility for humans to teach robots new tasks through language communication with a small number of examples. Third, we have created a dataset for this task, which is available to the community¹. Our bootstrapping approach can serve as a baseline for future work on this topic.

In the following sections, we first describe our data collection effort, then introduce the bootstrapping approach for action-effect prediction, and finally present results from our experiments.

2 Related Work

In the NLP community, there has been extensive work that models cause-effect relations from text (Cole et al., 2005; Do et al., 2011; Yang and Mao, 2014). Most of these previous studies address high-level causal relations between events, for example, “the collapse of the housing bubble” causes the effect of “stock prices to fall” (Sharp et al., 2016). They do not concern the kind of naive physical action-effect relations in this paper. There is also an increasing amount of effort on capturing commonsense knowledge, for example, through knowledge base population. Except for few (Yatskar et al., 2016) that acquires knowledge from images, most of the previous effort apply information extraction techniques to extract facts from a large amount of web data (Dredze et al., 2010; Rajani and Mooney, 2016). DBpedia (Lehmann et al., 2015), Freebase (Bollacker et al., 2008), and YAGO (Suchanek et al., 2007) knowledge bases contain millions of facts about the world such as people and places. However, they do not contain basic cause-effect knowledge related to concrete actions and their effects to the world. Recent work started looking into phys-

¹This dataset is available at <http://lair.cse.msu.edu/lair/projects/actioneffect.html>

ical causality of action verbs (Gao et al., 2016) and other physical properties of verbs (Forbes and Choi, 2017; Zellers and Choi, 2017; Chao et al., 2015). But they do not address action-effect prediction.

The idea of modeling object physical state change has also been studied in the computer vision community (Fire and Zhu, 2016). Computational models have been developed to infer object states from observations and to further predict future state changes (Zhou and Berg, 2016; Wu et al., 2016, 2017). The action recognition task can be treated as detecting the transformation on object states (Fathi and Rehg, 2013; Yang et al., 2013; Wang et al., 2016). However these previous works only focus on the visual presentation of motion effects. Recent years have seen an increasing amount of work integrating language and vision, for example, visual question answering (Antol et al., 2015; Fukui et al., 2016; Lu et al., 2016), image description generation (Xu et al., 2015; Vinyals et al., 2015), and grounding language to perception (Yang et al., 2016; Roy, 2005; Tellez et al., 2011; Misra et al., 2017). While many approaches require a large amount of training data, recent works have developed zero/few shot learning for language and vision (Mukherjee and Hospedales, 2016; Xu et al., 2016, 2017a,b; Tsai and Salakhutdinov, 2017). Different from these previous works, this paper introduces a new task that connects language with vision for physical action-effect prediction.

In the robotics community, an important task is to enable robots to follow human natural language instructions. Previous works (She et al., 2014; Misra et al., 2015; She and Chai, 2016, 2017) explicitly model verb semantics as desired goal states and thus linking natural language commands with underlying planning systems for action planning and execution. However, these studies were carried out either in a simulated world or in a carefully curated simple environment within the limitation of the robot’s manipulation system. And they only focus on a very limited set of domain specific actions which often only involve the change of locations. In this work, we study a set of open-domain physical actions and a variety of effects perceived from the environment (i.e., from images).

3 Action-Effect Data Collection

We collected a dataset to support the investigation on physical action-effect prediction. This dataset consists of actions expressed in the form of verb-noun pairs, effects of actions described in language, and effects of actions depicted in images. Note that, as we would like to have a wide range of possible effects, language data and image data are collected separately.

Actions (verb-noun pairs). We selected 40 nouns that represent everyday life objects, most of them are from the COCO dataset (Lin et al., 2014), with a combination of food, kitchen ware, furniture, indoor objects, and outdoor objects. We also identified top 3000 most frequently used verbs from Google Syntactic N-gram dataset (Goldberg and Orwant, 2013) (Verbargs set). And we extracted top frequent verb-noun pairs containing a verb from the top 3000 verbs and a noun in the 40 nouns which hold a *dobj* (i.e., direct object) dependency relation. This resulted in 6573 candidate verb-noun pairs. As changes to an object can occur at various dimensions (e.g., size, color, location, attachment, etc.), we manually selected a subset of verb-noun pairs based on the following criteria: (1) changes to the objects are visible (as opposed to other types such as temperature change, etc.); and (2) changes reflect one particular dimension as opposed to multiple dimensions (as entailed by high-level actions such as “cook a meal”, which correspond to multiple dimensions of change and can be further decomposed into basic actions). As a result, we created a subset of 140 verb-noun pairs (containing 62 unique verbs and 39 unique nouns) for our investigation.

Effects Described in Language. The basic knowledge about physical action-effect is so fundamental and shared among humans. It is often presupposed in our communication and not explicitly stated. Thus, it is difficult to extract naive action-effect relations from the existing textual data (e.g., web). This kind of knowledge is also not readily available in commonsense knowledge bases such as ConceptNet (Speer and Havasi, 2012). To overcome this problem, we applied crowd-sourcing (Amazon Mechanical Turk) and collected a dataset of language descriptions describing effects for each of the 140 verb-noun pairs. The workers were shown a verb-noun pair, and were asked to use their own words and imag-

| 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. |

Table 1: Example action and effect text from our collected data.

inations to describe what changes might occur to the corresponding object as a result of the action. Each verb-noun pair was annotated by 10 different annotators, which has led to a total of 1400 effect descriptions. Table 1 shows some examples of collected effect descriptions. These effect language descriptions allow us to derive *seed effect knowledge* in a symbolic form.

Effects Depicted in Images. For each action, three students searched the web and collected a set of images depicting potential effects. Specifically, given a verb-noun pair, each of the three students was asked to collect at least 5 positive images and 5 negative images. Positive images are those deemed to capture the resulting world state of the action. And negative images are those deemed to capture some state of the related object (i.e., the nouns in the verb-noun pairs), but are not the resulting state of the corresponding action. Then, each student was also asked to provide positive or negative labels for the images collected by the other two students. As a result each image has three positive/negative labels. We only keep the images whose labels are agreed by all three students. In total, the dataset contains 4163 images. On average, each action has 15 positive images, and 15 negative images. Figure 2 shows several examples of positive images and negative images of the action *peel-orange*. The positive images show an orange in a *peeled* state, while the negative images show oranges in different states (orange as a whole, orange slices, orange juice, etc.).

4 Action-Effect Prediction

Action-effect prediction is to connect actions (as causes) to the effects of actions. Specifically, given an image which depicts a state of the world, our task is to predict what concrete actions could cause the state of the world. This task is different from traditional action recognition as the underlying actions (e.g., human body posture/movement) are not captured by the images. In this regard, it is also different from image description generation.

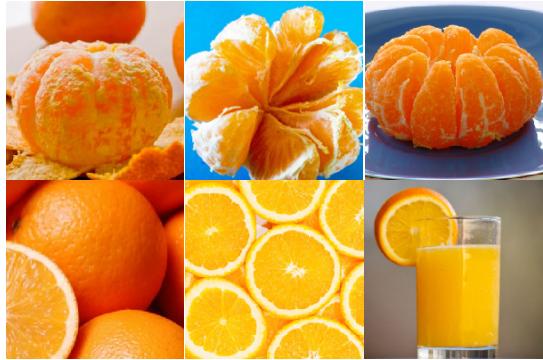


Figure 2: Positive images (top row) and negative images (bottom row) of the action *peel-orange*.

We frame the problem as a few-shot learning task, by only providing a few human-labelled images for each action at the training stage. Given the very limited training data, we attempt to make use of web-search images. Web search has been adopted by previous computer vision studies to acquire training data (Fergus et al., 2005; Kennedy et al., 2006; Berg et al., 2010; Otani et al., 2016). Compared with human annotations, web-search comes at a much lower cost, but with a trade-off of poor data quality. To address this issue, we apply a bootstrapping approach that aims to handle data with noisy labels.

The first question is what search terms should be used for image search. There are two options. The first option is to directly use the action terms (i.e., verb-noun pairs) to search images and the downloaded web images are referred to as *action web images*. As desired images should be depicting effects of an action, terms describing effects become a natural choice. The second option is to use the key phrases extracted from language effect descriptions to search the web. The downloaded web images are referred to as *effect web images*.

4.1 Extracting Effect Phrases from Language Data

We first apply chunking (shallow parsing) using the SENNA software (Collobert et al., 2011) to break an effect description into phrases such as noun phrases (NP), verb phrases (VP), prepositional phrases (PP), adjectives (ADJP), adverbs (ADVP), etc. After some examination, we found that most of the effect descriptions follow simple syntactic patterns. For a *verb-noun* pair, around 80% of its effect descriptions start with the same noun as the subject. In an effect description, the

| Example patterns | Example <i>Effect Phrases</i> (bold) extracted from effect descriptions |
|-------------------------------------------------------|----------------------------------------------------------------------------------|
| VP with a verb $\in \{\text{be, become, turn, get}\}$ | 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 . |
| PP + NP | The eggs are divided into whites and yolks . |

Table 2: Example patterns that are used to extract effect phrases (**bold**) from sample sentences.

change of state associated with the noun is mainly captured by some key phrases. For example, an adjective phrase usually describes a physical state; verbs like *be*, *become*, *turn*, *get* often indicate a description of change of the state. Based on these observations, we defined a set of patterns to identify phrases that describe physical states of an object. In total 1997 *effect phrases* were extracted from the language data. Table 2 shows some example patterns and example effect phrases that are extracted.

4.2 Downloading Web Images

The purpose of querying search engine is to retrieve images of objects in certain effect states. To form image searching keywords, the effect phrases are concatenated with the corresponding noun phrases, for example, “apple + into thin pieces”. The image search results are downloaded and used as supplementary training data for the action-effect prediction models. However, web images can be noisy. First of all, not all of the automatically extracted effect phrases describe visible state of objects. Even if a phrase represents visible object states, the retrieved results may not be relevant. Figure 3 shows some example image search results using queries describing the object name “book”, and describing the object state such as “book is on fire”, “book is set aflame”. These state phrases were used by human annotators to describe the effect of the action “burn a book”. We can see that the images returned from the query “book is set aflame” are not depicting the physical effect state of “burn a book”. Therefore, it’s important to identify images with relevant effect states to train the model. To do that, we applied a bootstrapping method to handle the noisy web images as described in Section 4.3. For an action (i.e., a verb-noun pair), it has multiple corresponding effect phrases, and all of their image search results are treated as training images for this action.

Since both the human annotated image data (Section 3) and the web-search image data were obtained from Internet search engines, they may



Figure 3: Examples of image search results.

have duplicates. As part of the annotated images are used as test data to evaluate the models, it is important to remove duplicates. We designed a simple method to remove any images from the web-search image set that has a duplicate in the human annotated set. We first embed all images into feature vectors using pre-trained CNNs. For each web-search image, we calculate its cosine similarity score with each of the annotated images. And we simply remove the web images that have a score larger than 0.95.

4.3 Models

We formulate the action-effect prediction task as a multi-class classification problem. Given an image, the model will output a probability distribution \mathbf{q} over the candidate actions (i.e., verb-noun pairs) that can potentially cause the effect depicted in the image.

Specifically for model training, we are given a set of human annotated seeding image data $\{\mathbf{x}, \mathbf{t}\}$ and a set of web-search image data $\{\mathbf{x}', \mathbf{t}'\}$. Here \mathbf{x} and \mathbf{x}' are the images (depicting effect states), and \mathbf{t} and \mathbf{t}' are their classification targets (i.e., actions). Each target vector is the observed image label, $\mathbf{t} \in \{0, 1\}^C$, $\sum_i t_i = 1$, and C is the number of classes (i.e., actions). The human annotated targets \mathbf{t} can be trusted. But the targets of web-search images \mathbf{t}' are usually very noisy. Bootstrapping method has been shown to be an effective method to handle noisy labelled data (Rosenberg et al., 2005; Whitney and Sarkar, 2012; Reed et al., 2014). The objective of the

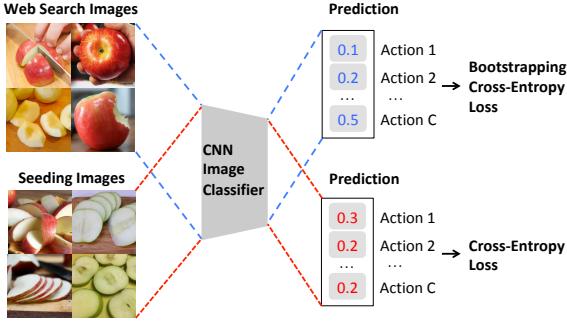


Figure 4: Architecture for the action-effect prediction model with bootstrapping.

cross-entropy loss is defined as follows:

$$\mathcal{L}(\mathbf{t}, \mathbf{q}) = \sum_{i=1}^C t_i \log (q_i), \quad (1)$$

where \mathbf{q} are the predicted class probabilities, and C is the number of classes. To handle the noisy labels in the web-search data $\{\mathbf{x}', \mathbf{t}'\}$, we adopt a bootstrapping objective following Reed’s work (Reed et al., 2014):

$$\mathcal{L}(\mathbf{t}', \mathbf{q}) = \sum_{i=1}^C [\beta t'_i + (1 - \beta) z_i] \log (q_i), \quad (2)$$

where $\beta \in [0, 1]$ is a model parameter to be assigned, \mathbf{z} is the one-hot vector of the prediction \mathbf{q} , $z_i = 1$, if $i = \text{argmax } q_k$, $k = 1 \dots C$.

The model architecture is shown in Figure 4. After each training batch, the current model will be used to make predictions \mathbf{q} on images in the next batch. And the target probabilities is calculated as a linear combination of the current predictions \mathbf{q} and the observed noisy labels \mathbf{t}' . The idea behind this bootstrapping strategy is to ensure the consistency of the model’s predictions. By first initializing the model on the seeding image data, the bootstrapping approach allows the model to trust more on the web images that are consistent with the seeding data.

4.4 Evaluation

We evaluate the models on the action-effect prediction task. Given an image that illustrates a state of the world, the goal is to predict what action could cause that state. Given an action in the form of a verb-noun pair, the goal is to identify images that depict the most likely effects on the state of the world caused by that action.

For each of the 140 verb-noun pairs, we use 10% of the human annotated images as the seeding image data for training, and use 30% for development and the rest 60% for test. The seeding image data set contains 408 images. On average, each verb-noun pair has less than 3 seeding images (including positive images and negative images). The development set contains 1252 images. The test set contains 2503 images. The model parameters were selected based on the performance on the development set.

As a given image may not be relevant to any effect, we add a background class to refer to images where effects are not caused by any action in the space of actions. So the total of classes for our evaluation model is 141. For each verb-noun pair and each of the effect phrases, around 40 images were downloaded from the Bing image search engine and used as candidate training examples. In total we have 6653 action web images and 59575 effect web images.

Methods for Comparison

All the methods compared are based on one neural network structure. We use ResNet (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) to extract image features. The extracted image features are fed to a fully connected layer with rectified linear units and then to a softmax layer to make predictions. More specifically, we compare the following configurations:

(1) *BS+Seed+Act+Eff*. The bootstrapping approach trained on the seeding images, the action web images, and the effect web images. During the training stage, the model was first trained on the seeding image data using vanilla cross-entropy objective (Equation 1). Then it was further trained on a combination of the seeding image data and web-search data using the bootstrapping objective (Equation 2). In the experiments we set $\beta = 0.3$.

(2) *BS+Seed+Act*. The bootstrapping approach trained in the same fashion as (1). The only difference is that this method does not use the effect web images.

(3) *Seed+Act+Eff*. A baseline method trained on a combination of the seeding images, the web action images, and the web effect images, using the vanilla cross-entropy objective.

(4) *Seed+Act*. A baseline method trained on a combination of the seeding images and the action web images, using the vanilla cross-entropy objective.

| | Top Action Predictions | Top Effect Descriptions | | Top Action Predictions | Top Effect Descriptions |
|--|----------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|--|----------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| | bake potato peel potato boil potato fry potato | potato crispy potato is crushed eggs get beaten potato browned | | wrap book tear book fold paper shave hair | book is ripped paper become creased book into smaller pieces meat is being prepped |
| | peel carrot cut wood chop carrot grate carrot | carrot into little sections tree into pieces carrot into tiny pieces wood is being chopped | | stain paper close drawer squeeze bottle crack bottle | bottle is pressed together meat is exposed paper around itself drawer is pushed back |
| | chop onion cut onion slice onion background | onion is being cut onion in banana is made banana is removed | | chop onion cook onion grate potato background | onion is heated onion into small pieces onion into multiple pieces onion is chopped |

Figure 5: Several example test images and their predicted actions and predicted effect descriptions. The actions in bold are ground-truth labels.

| | MAP | Top 1 | Top 5 | Top 20 |
|-----------------|--------------|--------------|--------------|---------------|
| BS+Seed+Act+Eff | 0.290 | 0.414 | 0.750 | 0.921 |
| BS+Seed+Act | 0.252 | 0.414 | 0.721 | 0.893 |
| Seed+Act+Eff | 0.247 | 0.314 | 0.679 | 0.886 |
| Seed+Act | 0.241 | 0.371 | 0.650 | 0.814 |
| Seed | 0.182 | 0.329 | 0.629 | 0.807 |

Table 3: Results for the action-effect prediction task (given an action, rank all the candidate images).

| | MAP | Top 1 | Top 5 | Top 20 |
|-----------------|--------------|--------------|--------------|---------------|
| BS+Seed+Act+Eff | 0.660 | 0.523 | 0.843 | 0.954 |
| BS+Seed+Act | 0.642 | 0.508 | 0.802 | 0.924 |
| Seed+Act+Eff | 0.289 | 0.176 | 0.398 | 0.625 |
| Seed+Act | 0.481 | 0.301 | 0.724 | 0.926 |
| Seed | 0.634 | 0.520 | 0.765 | 0.892 |

Table 4: Results for the action-effect prediction task (given an image, rank all the actions).

(5) *Seed*. A baseline method that was only trained on the seeding image data, using the vanilla cross-entropy objective.

Evaluation Results

We apply the trained classification model to all of the test images. Based on the matrix of prediction scores, we can evaluate action-effect prediction from two angles: (1) given an action class, rank all the candidate images; (2) given an image, rank all the candidate action classes. Table 3 and 4 show the results for these two angles respectively. We report both mean average precision (MAP) and top prediction accuracy.

Overall, *BS+Seed+Act+Eff* gives the best performance. By comparing the bootstrap approach with baseline approaches (i.e., *BS+Seed+Act+Eff*

vs. *Seed+Act+Eff*, and *BS+Seed+Act* vs. *Seed+Act*), the bootstrapping approaches clearly outperforms their counterparts, demonstrating its ability in handling noisy web data. Comparing *BS+Seed+Act+Eff* with *BS+Seed+Act*, we can see that *BS+Seed+Act+Eff* performs better. This indicates the use of effect descriptions can bring more relevant images to train better models for action-effect prediction.

In Table 4, the poor performance of *Seed+Act+Eff* and *Seed+Act* shows that it is risky to fully rely on the noisy web search results. These two methods had trouble in distinguishing the background class from the rest.

We further trained another multi-class classifier with web effect images, using their corresponding effect phrases as class labels. Given a test image, we apply this new classifier to predict the effect descriptions of this image. Figure 5 shows some example images, their predicted actions based on our bootstrapping approach and their predicted effect phrases based on the new classifier. These examples also demonstrate another advantage of incorporating seed effect knowledge from language data: it provides state descriptions that can be used to better explain the perceived state. Such explanation can be crucial in human-agent communication for action planning and reasoning.

5 Generalizing Effect Knowledge to New Verb-Noun Pairs

In real applications, it is very likely that we do not have the effect knowledge (i.e., language effect descriptions) for every verb-noun pair. And annotat-

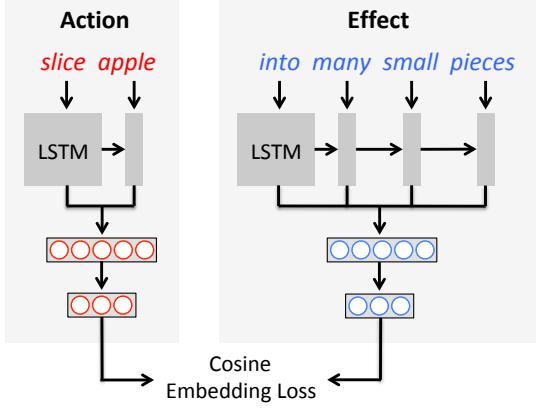


Figure 6: Architecture of the action-effect embedding model.

ing effect knowledge using language (as shown in Section 3) can be very expensive. In this section, we describe how to potentially generalize seed effect knowledge to new verb-noun pairs through an embedding model.

5.1 Action-Effect Embedding Model

The structure of our model is shown in Figure 6. It is composed of two sub-networks: one for verb-noun pairs (i.e., *action*) and the other one for effect phrases (i.e., *effect*). The *action* or *effect* is fed into an LSTM encoder and then to two fully-connected layers. The output is an action embedding \mathbf{v}_c and effect embedding \mathbf{v}_e . The networks are trained by minimizing the following cosine embedding loss function:

$$\mathcal{L}(\mathbf{v}_c, \mathbf{v}_e) = \begin{cases} 1 - s(\mathbf{v}_c, \mathbf{v}_e), & \text{if } (c, e) \in T \\ \max(0, s(\mathbf{v}_c, \mathbf{v}_e)), & \text{if } (c, e) \notin T \end{cases}$$

$s(\cdot, \cdot)$ is the cosine similarity between vectors. T is a collection of action-effect pairs. Suppose c is an input for *action* and e is an input for *effect*, this loss function will learn an action and effect semantic space that maximizes the similarities between c and e if they have an action-effect relation (i.e., $(c, e) \in T$). During training, the negative action-effect pairs (i.e., $(c, e) \notin T$) are randomly sampled from data. In the experiments, the negative sampling ratio is set to 25. That is, for each positive action-effect pair, 25 negative pairs are created through random sampling.

At the inference step, given an unseen verb-noun pair, we embed it into the action and effect semantic space. Its embedding vector will be used to calculate similarities with all the embedding vectors of the candidate effect phrases.

| | MAP | Top 1 | Top 5 |
|------------------|--------------|--------------|--------------|
| BS+Seed+Act+Eff | 0.529 | 0.643 | 0.928 |
| BS+Seed+Act+pEff | 0.507 | 0.642 | 0.893 |
| BS+Seed+Act | 0.435 | 0.643 | 0.964 |
| Seed | 0.369 | 0.678 | 0.786 |

Table 5: Results for the action-effect prediction task (given an action, rank all the candidate images).

| | MAP | Top 1 | Top 5 |
|------------------|--------------|--------------|--------------|
| BS+Seed+Act+Eff | 0.733 | 0.574 | 0.947 |
| BS+Seed+Act+pEff | 0.729 | 0.551 | 0.961 |
| BS+Seed+Act | 0.724 | 0.557 | 0.933 |
| Seed | 0.705 | 0.557 | 0.898 |

Table 6: Results for the action-effect prediction task (given an image, rank all the actions).

5.2 Evaluation

We divided the 140 verb-noun pairs into 70% training set (98 verb-noun pairs), 10% development set (14) and 20% test set (28). For the action-effect embedding model, we use pre-trained GloVe word embeddings (Pennington et al., 2014) as input to the LSTM. The embedding model was trained using the language effect data corresponding to the training verb-noun pairs, and then it was applied to predict effect phrases for the unseen verb-noun pairs in the test set. For each unseen verb-noun pair, we collected its top five predicted effect phrases. Each predicted effect phrase was then used as query keywords to download web effect images. This set of web images are referred to as *pEff* and will be used in training the action-effect prediction model.

For each of the 28 test (i.e., new) verb-noun pairs, we use the same ratio 10% (about 3 examples) of the human annotated images as the seeding images, which were combined with downloaded web images to train the prediction model. The remaining 30% and 60% are used as the development set, and the test set. We compare the following different configurations:

- (1) *BS+Seed+Act+pEff*. The bootstrapping approach trained on the seeding images, the action web images, and the web images downloaded using the predicted effect phrases.
- (2) *BS+Seed+Act+Eff*. The bootstrapping approach trained on the seeding images, the action web images, and the effect web images (downloaded using ground-truth effect phrases).
- (3) *BS+Seed+Act*. The bootstrapping approach trained on the seeding images and the action web

| Action Text | Predicted Effect Text |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------|
| chop carrot | carrot into sandwiches, carrot is sliced, carrot is cut thinly, carrot into different pieces, carrot is divided |
| ignite paper | paper is being charred , paper is being burned, paper is set, paper is being destroyed, paper is lit |
| mash potato | potato into chunks, potato into sandwiches, potato into slices, potato is chewed, potato into smaller pieces |

Table 7: Example predicted effect phrases for new verb-noun pairs. Unseen verbs and nouns are shown in bold.

images.

(4) *Seed*. A baseline only trained on the seeding images.

Table 5 and 6 show the results for the action-effect prediction task for unseen verb-noun pairs. From the results we can see that $BS+Seed+Act+pEff$ achieves close performance compared with $BS+Seed+Act+Eff$, which uses human annotated effect phrases. Although in most cases, $BS+Seed+Act+pEff$ outperforms the baseline, which seems to point to the possibility that semantic embedding space can be employed to extend effect knowledge to new verb-noun pairs. However, the current results are not conclusive partly due to the small testing set. More in-depth evaluation is needed in the future.

Table 7 shows top predicted effect phrases for several new verb-noun pairs. After analyzing the action-effect prediction results we notice that generalizing the effect knowledge to a verb-noun pair that contains an unseen verb tends to be more difficult than generalizing to a verb-noun pair that contains an unseen noun. Among the 28 test verb-noun pairs, 12 of them contain unseen verbs and known nouns, 7 of them contain unseen nouns and known verbs. For the task of ranking images given an action, the mean average precision is 0.447 for the unseen verb cases and 0.584 for the unseen noun cases. Although not conclusive, this might indicate that, verbs tend to capture more information about the effect states of the world than nouns.

6 Discussion and Conclusion

When robots operate in the physical world, they not only need to perceive the world, but also need to act to the world. They need to understand the current state, to map their goals to the world state, and to plan for actions that can lead to the goals. All of these point to the importance of the ability to understand causal relations between actions and the state of the world. To address this issue, this paper introduces a new task on action-effect prediction.

Particularly, we focus on modeling the connection between an action (a verb-noun pair) and its effect as illustrated in an image and treat natural language effect descriptions as side knowledge to help acquiring web image data and bootstrap training. Our current model is very simple and performance is yet to be improved. We plan to apply more advanced approaches in the future, for example, attention models that jointly capture actions, image states, and effect descriptions. We also plan to incorporate action-effect prediction to human-robot collaboration, for example, to bridge the gap of commonsense knowledge about the physical world between humans and robots.

This paper presents an initial investigation on action-effect prediction. There are many challenges and unknowns, from problem formulation to knowledge representation; from learning and inference algorithms to methods and metrics for evaluations. Nevertheless, we hope this work can motivate more research in this area, enabling physical action-effect reasoning, towards agents which can perceive, act, and communicate with humans in the physical world.

Acknowledgments

This work was supported by the National Science Foundation (IIS-1617682) and the DARPA XAI program under a subcontract from UCLA (N66001-17-2-4029). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

- Renée Baillargeon. 2004. Infants' physical world. *Current directions in psychological science*, 13(3):89–94.
- Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.
- Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015. Mining semantic affordances of visual object categories. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4259–4267. IEEE.
- Stephen V Cole, Matthew D Royal, Marco G Valtorta, Michael N Huhns, and John B Bowles. 2005. A lightweight tool for automatically extracting causal relationships from text. In *SoutheastCon, 2006. Proceedings of the IEEE*, pages 125–129. IEEE.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.
- Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.
- Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics.
- Curt J Ducasse. 1926. On the nature and the observability of the causal relation. *The Journal of Philosophy*, 23(3):57–68.
- Alireza Fathi and James M Rehg. 2013. Modeling actions through state changes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2579–2586. IEEE.
- Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from google's image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1816–1823. IEEE.
- Amy Fire and Song-Chun Zhu. 2016. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23.
- Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 266–276.
- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
- Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Y Chai. 2016. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1814–1824.
- Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 241–247.
- Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. 2007. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Lyndon S Kennedy, Shih-Fu Chang, and Igor V Kozintsev. 2006. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258. ACM.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.
- Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1015–1026. Association for Computational Linguistics.
- Dipendra Kumar Misra, Kejia Tao, Percy Liang, and Ashutosh Saxena. 2015. Environment-driven lexicon induction for high-level instructions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 992–1002.
- Tanmoy Mukherjee and Timothy Hospedales. 2016. Gaussian visual-linguistic embedding for zero-shot recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 912–918.
- Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer.
- Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Nazneen Fatema Rajani and Raymond J Mooney. 2016. Combining supervised and unsupervised ensembles for knowledge base population. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*.
- Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.
- Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 29–36. IEEE.
- Deb Roy. 2005. Grounding words in perception and action: computational insights. *Trends in cognitive sciences*, 9(8):389–396.
- Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 138–148.
- Lanbo She and Joyce Chai. 2016. Incremental acquisition of verb hypothesis space towards physical world interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 108–117.
- Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1634–1644.
- Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 89–97.
- Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.
- Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 1, page 2.
- Yao-Hung Hubert Tsai and Ruslan Salakhutdinov. 2017. Improving one-shot learning through fusing side information. *arXiv preprint arXiv:1710.08347*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.
- Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. 2016. Actions~ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667.
- Max Whitney and Anoop Sarkar. 2012. Bootstrapping via graph propagation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 620–628. Association for Computational Linguistics.

Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. 2016. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, volume 2, page 7.

Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. 2017. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems*, pages 152–163.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Xun Xu, Timothy Hospedales, and Shaogang Gong. 2017a. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333.

Xun Xu, Timothy M Hospedales, and Shaogang Gong. 2016. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer.

Zhongwen Xu, Linchao Zhu, and Yi Yang. 2017b. Few-shot object recognition from machine-labeled web images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159.

Xuefeng Yang and Kezhi Mao. 2014. Multi-level causal relation identification using extended features. *Expert Systems with Applications*, 41(16):7171–7181.

Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. 2013. Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2563–2570. IEEE.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of NAACL-HLT*, pages 193–198.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yipin Zhou and Tamara L Berg. 2016. Learning temporal transformations from time-lapse videos. In *European Conference on Computer Vision*, pages 262–277. Springer.