EECS 595 Project Report

Yanfu Guo
University of Michigan
yanfuguo@umich.edu

Yijie Shi
University of Michigan
yijieshi@umich.edu

Abstract

The causal inference has drawn a lot of attention in recent years due to the research and application significance. However, artificial intelligence still lacks the ability to infer action-effect relations in real-world problems. For example, the action of dropping a glass will result in the glass breaking or lying on the floor. If artificial agents don't have this ability, they won't be able to reason about the state of the world or choose the appropriate action. To solve this task, in this paper, we follow a task on naive physical action-effect prediction, which is introduced in previous work. The goal is to address the relations between actions in the form of verb-noun pairs and corresponding effects in the form of physical world images. We proposed a new method to collect a web image dataset for this task and improved the model by using a pre-trained transformer model. The experiment results have shown that transformer performs well on this task and our dataset can provide more realistic multi-label data so that the trained model can better cope with complex real problems.

1 Introduction

Thanks to the success of deep neural networks and multi-modal learning in the past few years, great progress has been made in action recognition. However, little effort has been devoted to the interplay between action and effect (Yoo et al., 2021). Despite recent advances in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the ability to understand the relationship between action and effect in real-world problems. For artificial agents, being able to understand the relationship between actions and effects is key to simulating actions and reasoning in the real world. For example, it is easy for humans to infer that the action of dropping a glass will result in the glass breaking or lying on the floor, but this task remains challenging for robots.

In recent years, the performance of artificial intelligence in tasks such as representation and reasoning has greatly improved. For example, given the words 'sliced apple', an agent could easily judge whether an image shows apples sliced into pieces. However, it is still hard for AIs to build cause-effect relations between phrases and images. This kind of ability could help agents understand the world in the way of humans and further enable agents to learn through communications with humans to make reasonable and human-like decisions when facing new conditions.

In this paper, we focus on the task that aims to determine the action-effect causal relation between actions (expressed in the form of verb-noun pairs) and the corresponding effects (expressed in the form of several images), introduced in previous work (Gao et al., 2018).

To be more specific, given an image of the effect and several candidate actions, the goal of the AI is to determine the probabilities of causing that effect for each candidate's action. Also, given an action (verb-noun pair) and several candidate effect images, the AI should be able to rank the likelihood of all effect images.

The main contributions are summarized as follows:

1. We created our own dataset using Huggingface images search API, rather than the Bing image search engine.

2. We used verb-noun pairs as action keywords to search effect images, rather than human-annotated effect descriptions.

3. We replaced the simple ResNet model used in previous work with a pre-trained transformer provided by Microsoft in Huggingface.

The rest of this paper is organized as follows. We will review related works for this project in Section 2. Next, in Section 3, we will introduce

Gao's dataset (Gao et al., 2018) used in our work. Then, in Section 4, we will give a brief introduction to our task. In Section 5, we will discuss the preprocessing methods and compare them with Gao's approaches. In Section 6, we will describe and propose our new approaches and compare them with Gao's work. In Section 7, we will discuss how we evaluate our models and approaches. In Section 8, we will discuss our experimental results and conduct an analysis of them. Finally, we conclude our works and summarize our contribution in Section 9.

2 Related Works

The Cause-Effect relations has been mentioned and studied in some paper (Cole et al., 2006; Do et al., 2011; Yang and Mao, 2014). However, most papers that dig into cause-effect relations focus on the high-level relations (Sharp et al., 2016) while paying less attention to the physical direct cause-effect relations. In this paper, we will connect the images with words. However, in recent papers, there are only a few paper (Yatskar et al., 2016) that focus on this work, and most of the work are mainly extracting facts and information from a huge amount of web data, or knowledge base (Dredze et al., 2010).

Modeling the state of items has been studied in computer vision (Zhou and Berg, 2016; Wu et al., 2016). The state of the item can be inferred through observation, and this observation could provide evidence for future state changes (Fathi and Rehg, 2013).

With the rise of interest in exploring the connection between images and words. Problems such as visual question answering (Fukui et al., 2016), image description generation (Xu et al., 2015), and grounding language to perception (Yang et al., 2016) have been proposed and studied. Isola et al. (2015) collected a dataset of scenes, objects, and materials, each being in different transformed states. Their proposed tasks are threefold; Firstly, given a noun, find the related changes it can undergo (e.g., a potato can undergo slicing, cooking, etc.). Secondly, given an image, assign states that can be sensed (e.g., slice, original, etc.). Thirdly, arrange a set of images with a common object between a pair of states according to their properties (e.g. sausage images from raw to ripe, potatoes from unripe to ripe, etc.).

Recent work started to address action-effect prediction (Gao et al., 2018). This research uses images to address the relationship between actions and effects on the state of the physical world. They use web image data for action effect prediction with distant supervision. Their research shows that using effect descriptions leads to better behavioral performance and effect prediction.

After that, a cycle-reasoning model that can effectively reason about the precondition and effect and can enhance action recognition performance has been proposed (Yoo et al., 2021). There is a developmental approach that allows a robot to interpret and describe the actions of human agents by reusing previous experience (Saponaro et al., 2020). Also, Yang et al. (2021) proposed the Visual Goal-Step Inference (VGSI) task and created a dataset using wikiHow. Given a high-level textual goal and a set of candidate images, the model learns to identify the images which constitute a reasonable step toward the given goal.

This paper follows the task (Gao et al., 2018) that connects language with a vision for physical action-effect prediction and improves the previous approaches.

3 Dataset

The dataset that we use for this task is the Action-Effect dataset (Gao et al., 2018) on Huggingface. This dataset contains action-effect information for 140 verb-noun pairs. It has two parts: effects described by natural language and effects depicted in images.

The language data contains verb-noun pairs and their effects described in natural language. For each verb-noun pair, its possible effects are described by 10 different annotators. Effect phrases were automatically extracted from their corresponding effect sentences.

The image data contains images depicting action effects. For each verb-noun pair, positive images and negative images were collected. 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.

In this project, we mainly focus on the image part. Each sample in the image set has five columns, verb noun, effect_sentence_list, effect_phrases_list, positive_image_list, negative_image_list. The detailed explanation of each column is shown in

Table 1.

4 Problem Description

Action-to-Effect prediction builds up the connection between actions and the effect of these actions. This task is different from traditional image description generation or action recognition because it contains causal inference in this task. We need artificial intelligence to understand basic action-effect relations regarding the physical world. For example, the action of slicing apples most likely leads to the state where the apples are broken apart into smaller pieces. Therefore, our aim is to help AI build up such kind of connection.

5 Preprocessing Work

Gao's dataset for each verb-noun pair contains less than 3 seeding images and around 27 web-search images. Those web-search images are downloaded using human-annotated effect descriptions as keywords. To better train our model, we need a big number of images. Therefore, we want to crawl images from image search APIs.

In Gao's work, the dataset is generated using the Bing image search engine, and now we want to switch to the hugging face image search API because it is less noisy. We tried several image APIs and manually checked the quality. After comparing the image quality from google image search API, bing image search API, and hugging face image search API, we decided to use the hugging face API.

However, we found that if we use these effect descriptions as keywords, some images we get are noisy. Table 2 shows actions and corresponding description texts.

Taking action stain shirt as an example, figure 1 shows the search result of 'stain shirt' and Figure 2 shows the result of 'There is a visible mark on the shirt.'. We can see that the previous result is more accurate.

Figure 1: Search result of 'stain shirt'

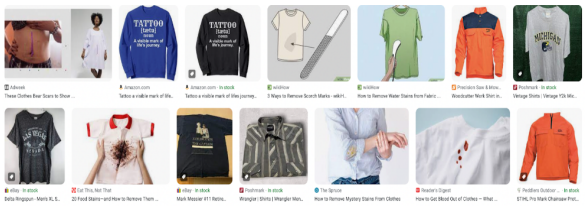

Figure 2: Search result of 'There is a visible mark on the shirt'

One of the possible reasons is that the search engine may not get the point given a relatively long sentence. Therefore, we choose to use verb-noun pairs to get images, and we find that the outcome is better.

For 140 verb-noun pairs in Gao's dataset, we downloaded 50 images for each label. Since each image downloaded from the image search API may vary in size, we further crop these images into the same size (224×224) so that it would be convenient for us to process them as input and extract features from these images.

Then we normalize the image mean and standard deviation for the model architecture we are going to use. We instantiate what is called a feature extractor with the $AutoFeatureExtractor.from\_pretrained$ method.

6 Approaches

Gao's work (Gao et al., 2018) raises the action-effect prediction problem for artificial intelligence, and, in addition, Gao offers a simple way to solve it. The architecture of the action-effect prediction model is basically a CNN image classifier with bootstrapping objective following Reed's work (Reed et al., 2014). Here, for the two tasks, we provide new models for each task and expect the new models would achieve better outcomes.

6.1 Effect to action prediction task

In the effect-to-action task, the model is given an image, and we will let the model output a probability distribution of all candidate actions. The structure of the model is shown in Figure 3. Here we use a transformer instead of CNN image classifier. The transformer here is based on swin-transformer. It will take features of figures as input and output a probability distribution of all candidate actions. The sum of all dimensions is equal to 0. The value on i^{th} column shows the likelihood that the effect of i^th action is shown in the input image.

Column Name	Type	Meaning
verb noun	string	action described as a verb and a noun
effect_sentence_list	sequence	sentences that describe the effect of this action
effect_phrases_list	sequence	words extracted from each sentence in effect_sentence_list that represents this action
positive_image_list	sequence	images that correctly show the effect of the action
negative_image_list	sequence	images with the same topic, but the item is not in the correct state

Table 1: information about each column in the dataset

Action	Description text
ignite paper	The paper is on fire
soak shirt	The shirt is thoroughly wet
fry potato	The potatoes become crispy
stain shirt	There is a visible mark on the shirt

Table 2: Example action and description text

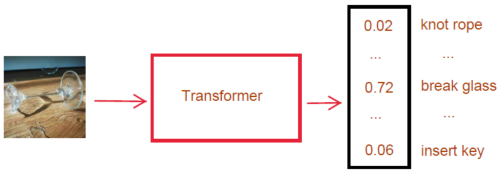

Figure 3: Architecture for effect-to-action prediction model

6.2 Action to effect prediction task

In the action-to-effect task, we give an action(verb-noun pair) and a set of images to the model and let it choose the most likely image that depicts the effect of the given action.

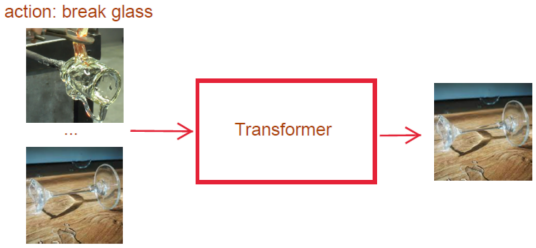

Figure 4: Architecture for the action to effect prediction model

Similar to the effect-to-action model, we choose transformer instead of CNN image classifier as our model. This transformer will take all features of images and make predictions for each image. Then based on all these distributions, it will rank the probability corresponding to the true label and find the most likely image that depicts the effect of the action.

6.3 Build up Dataset

Besides, we decided to enlarge Gao's human-annotated seeding image data. To be more specific, we would add 14 verb-noun pairs together with corresponding effect descriptions and images to the dataset. Both the verb and noun would be unique. This means that we enlarge the verb set by 22.5% and the noun set by 35.9%. All of these new data points would be added to our new training set. We believe doing so can provide more robustness to our model.

After finishing building up our dataset, we divided it into the training set and testing set and uploaded them to the hugging face dataset. For each verb-non pair, our dataset contains 50 images with the verb-noun pair label and the index of this label.

7 Evaluation

We evaluate our model on two tasks, action-to-effect prediction and effect-to-action prediction. For 140 verb-noun pairs given, we divide the data of each label into the training set and testing set. The training set takes about 90%, and the testing set is the remaining 10%.

For both the action-to-effect prediction task and the effect-to-action prediction task, we use a pre-trained transformer model. In this project, we fine-tuned the model from the 'swin-tiny-patch4-window7-224' pre-trained transformer checkpoint provided by Microsoft. This pre-trained model is based on the Swin Transformer model, which is trained on ImageNet-1k at resolution 224x224. It was introduced in the paper Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (Liu et al., 2021). The Swin Transformer is a type of Vision Transformer that serves as a general-purpose backbone for both image classification and dense recognition tasks.

7.1 Effect-to-action Prediction

For the effect-to-action task, our model is given an image and will output a probability distribution over the candidate actions that can potentially cause the effect depicted. Based on the probability distribution, we could know what the most likely action that may cause the effect in the image is. We compare the result among random guesses, ResNet, ResNet + bootstrapping, and our transformer model result. The comparing result is shown in Table 3. The result contains three columns. Since the model outputs a probability distribution among all candidate actions, and we could evaluate the performance by checking what ranking the true label is. The Top k accuracy means that the model chooses the true label in rank first k choices.

7.2 Action-to-effect Prediction

For the action-to-effect task, the model is given a set of images and an action. Then it will choose an image that most likely shows the effect of the given action. We compare the result among ResNet, ResNet + bootstrapping, and our transformer model result. The comparing result is shown in Table 4. The result also contains three columns. Given a set of images and the action, the model will make a prediction on each image and output the probability of each possible label. Assume the true label is on index i. We will rank the i^{th} column of each image's probability distribution on candidate actions. The Top k accuracy here means that the true answer is ranked in first k positions in the probability of label i.

8 Discussion

For ease of use, we pushed our dataset generated by Huggingface APIs and verb-noun action pairs to Huggingface Hub as a public dataset. We also provided python scripts on a Github public repository that can easily load our dataset or Gao's dataset, train a transformer on either dataset and evaluate a trained model on either task 1 (effect to action) or task 2 (action to effect)

We trained the transformer model on Gao's dataset and test on it for two tasks. Also, We trained the same pre-trained transformer model on our dataset with the same hyper-parameters and test on it for two tasks, respectively. Section 7 describes our evaluation process. The accuracy results are shown in Table 3,4.

By comparing the Gao's dataset results (i.e., *ResNet+Gao's dataset* v.s. *ResNet+Bootstrap+Gao's Dataset* v.s. *Transformer+Gao's Dataset*), the top 1 accuracy increased to 0.6195 on task 1 and 0.78 on task 2. Therefore, we can conclude that using a transformer gives much better performance than using a CNN model (ResNet). Transformer without bootstrapping approaches clearly outperforms ResNet with and without bootstrapping, demonstrating its ability to understand naive causal relations.

In the last two rows of Table 3,4, by comparing different dataset results (i.e., *Transformer+Gao's Dataset* v.s. *Transformer+Our Dataset*), it is obvious that the top 1 accuracy on our dataset in both tasks is lower than the top 1 accuracy on Gao's dataset. It may indicate that: (1) the causal relations in our dataset is harder for transformer to learn; (2) our dataset contains too many noisy web search images. However, we can see that the top 5 accuracy on our dataset is much better, and the gap between two datasets has been narrowed.

After careful analysis by manually checking our dataset, we do observe some noise data in our dataset. However, we found the main reason for this phenomenon is: in our dataset, some downloaded images are actually multi-labeled. For example, Figure 5 shows an example image in our dataset which should be labeled as both 'arrange flowers' and 'arrange chairs.' In other words, it is reasonable to consider this image as a physical world effect of the action 'arrange flowers' or the action 'arrange chairs.' Therefore, the top 1 accuracy gap is much larger than the top 5 accuracy gap. Due to such data, we believe Top 5 accuracy is better than the Top 1 accuracy.

Besides, we believe the model trained on our dataset can better cope with real complex problems since our dataset provides realistic multi-label data. In reality, robots rarely encounter a single object in their field of view. It's rare, for example, to have an empty floor with neatly arranged chairs and nothing else. Most of the time, we'll see a table next to the arranged chairs, a vase on the table, cutlery, food, someone next to the table, and a little trash on the floor. The artificial agents should infer several possibilities: someone might have littered, someone might have laid out the flowers, someone might have just cooked the meal, and someone might have laid out the chairs. Artificial intelligence needs to

Method	Top 1 Accuracy	Top 5 Accuracy	Top 20 Accuracy
Random Guess	0.0071	0.0357	0.143
ResNet+Gao's Dataset	0.176	0.398	0.625
ResNet+Bootstrap+Gao's Dataset	0.523	0.843	0.954
Transformer+Gao's Dataset	0.6195	0.8976	0.9902
Transformer+Our Dataset	0.4423	0.8122	0.9621

Table 3: Results for the effect-action prediction (given an effect image, rank all actions).

Method	Top 1 Accuracy	Top 5 Accuracy	Top 20 Accuracy
ResNet+Gao's Dataset	0.314	0.679	0.886
ResNet+Bootstrap+Gao's Dataset	0.414	0.750	0.921
Transformer+Gao's Dataset	0.78	0.93	1.0
Transformer+Our Dataset	0.37	0.88	0.99

Table 4: Results for the action-effect prediction (given an action, rank all candidate images).

Figure 5: An example in our dataset that should be labeled as both 'arrange flowers' and 'arrange chairs'

be able to deal with such complex problems before it can be a partner for humans.

9 Conclusion

For artificial agents, reasoning actions is important because it helps them predict whether a series of actions will lead them to achieve their desired goals; explain an observation in terms of possible actions, and identify actions that may result in an undesirable situation.

In this paper, we follow the naive physical action-effect prediction task introduced in previous work. The goal is to determine the causal relations between an action in the form of a verb-noun pair and the corresponding effect in the form of a physical world image. We propose a novel approach to collect a web image dataset and improve the model by using pre-trained Transformer models. Experimental results show that the Transformer performs well on this task, and our dataset can provide more realistic multi-label data, enabling the trained model to better deal with complex real-world problems.

As future work, it would be useful to investigate further how to logically connect actions and corresponding effects. So far, our model only works on pre-defined 140 classes (i.e., 140 actions), and the model can be considered as a (multi-label) classifier. Ideally, we expect the artificial agents can correctly guess the effect of an action based on prior knowledge even though the agents have never seen the effect of this action. In other words, we want the model to be able to handle new actions. For example, the artificial agent knows that glass has the property 'fragile', and it knows the action 'dropping glass' may result in the effect of 'the glass broke into pieces'. Now when it encounters a new object A that has the property 'fragile', it should be able to infer that the action 'dropping A' may result in the effect 'A broke into pieces'. We want to feed the properties of objects to models, or let the model judge the properties by itself. For instance, agent can infer that an object has property 'soft' by touching it with a robotic arm with sensors, or look at the object with a vision sensor, and then search the knowledge base for known objects that look similar. This may be a step away from the realm of natural language processing, but it's still an meaningful topic.

Also, in the real world, some actions are expected to occur more frequently than other actions,

which creates the class imbalance problem. Besides, we believe it makes sense to include videos in the datasets, which can help the model understand causal relations.

References

Stephen V Cole, Matthew D Royal, Marco G Valtorta, Michael N Huhns, and John B Bowles. 2006. A lightweight tool for automatically extracting causal relationships from text. In *Proceedings of the IEEE SoutheastCon 2006*, pages 125–129. IEEE.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, Tim Finin, et al. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Alireza Fathi and James M Rehg. 2013. Modeling actions through state changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2586.

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, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945, Melbourne, Australia. Association for Computational Linguistics.

P. Isola, J. J. Lim, and E. H. Adelson. 2015. Discovering states and transformations in image collections. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391, Los Alamitos, CA, USA. IEEE Computer Society.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping.

Giovanni Saponaro, Lorenzo Jamone, Alexandre Bernardino, and Giampiero Salvi. 2020. Beyond the self: Using grounded affordances to interpret and describe others' actions. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):209–221.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. *arXiv preprint arXiv:1609.08097*.

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.

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. PMLR.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce 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.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikihow.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198.

Hongsang Yoo, Haopeng Li, Qiuhong Ke, Liangchen Liu, and Rui Zhang. 2021. Precondition and effect reasoning for action recognition. *CoRR*, abs/2112.10057.

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

A Appendices

A.1 Work Division

This is a two-member team project. Both team members, Yanfu Guo and Yijie Shi, participate in designing models, writing codes that train and test the models, and writing the final project report. In addition to that, we divide our work as follows:

- Yanfu Guo individual works: (1) wrote the code that loaded and re-organized Gao's dataset; (2) wrote the Abstract, Introduction, Related Works, Discussion, Conclusion and Work Division of this final project report; (3) wrote the code that preprocesses data, trains model and evaluates the trained model on both task 1 and task 2.
- Yijie Shi individual works: (1) wrote the code that harnesses web image data to generate our dataset; (2) read related papers and see whether there exist some techniques that can be applied to this problem and further improve our models; (3) wrote the Dataset, Problem Description, Preprocessing Work, Approaches and Evaluation of this final project report.