End-to-End Vision and Language Models for Commonsense Reasoning with Explainability

Stanford CS224N Custom Project

Aman Chadha
Department of Computer Science
Stanford University
achadha@stanford.edu

1 Key Information to include

- External collaborators (if you have any): None.
- Mentor (custom project only): Mentorship requested.
- Sharing project: No.

2 Research paper summary (max 2 pages)

Title	Learning Contextual Causality from Time-Consecutive Images
Venue	arXiv pre-print, under review at ACL 2021
Year	2020
URL	https://arxiv.org/abs/2012.07138

Table 1: Bibliographical information for the chosen baseline paper [1].

Background. Causality knowledge is vital to building robust AI systems. Deep learning models often perform poorly on tasks that require causality-based reasoning, which is often derived using some form of commonsense knowledge from information not immediately present in the input. However, conventional causality systems rely solely on textual information which is laborious and expensive to annotate. As a result, their effectiveness is often limited. The paper proposes a scalable way of learning contextual causality using both the visual and text modality. To obtain data for the visual modality, the paper simplifies the task into mining causality knowledge from time-consecutive images, which are uniformly sampled from the video. For the text modality, the paper utilizes descriptions of events in the video.

Summary of contributions. The paper offers two major contributions:

1. Compared to pure text-based causality inference approaches, the paper offers the following advantages owing to its proposed idea of learning commonsense knowledge using both the visual and text modality:
   (a) **Causality using the visual modality**: Commonsense knowledge is rarely expressed in text but is rich in the visual modality;
   (b) **Exploit the time-ordered nature of videos**: Events in most video sequences are naturally time-ordered, which provides a rich and readily-available resource for us to mine causality knowledge from;
   (c) **Use objects to develop contextual causality**: Objects in the video can be used as context to develop a contextual understanding of the scene and thus infer causal relations.

2. Furthermore, the paper also proposes a dataset called Vis-Causal which offers image pairs and localized events with binary labels indicating causal relationships between events.

Limitations and discussion. The paper suffers from the following drawbacks:

1. **Videos? Sorry!**: A major drawback of Zhang et al. [1] is that it cannot natively accept videos as input, which are a much more prevalent source of commonsense knowledge compared to images and text. Instead, the model requires a pair of randomly-selected images to be able to infer commonsense knowledge. This dampens our goal of commonsense generation natively using videos and also prevents end-to-end training.
2. **Lacks the ability to rationalize causal relationships**: The proposed model doesn't offer rationalization for its causality inference, thereby making the model less interpretable. This also makes it difficult to understand the source of error/bias when analyzing results since the rationalization can offer a peak into the model's modus operandi, i.e., a 'debug signal' to develop an understanding of the model's (mis)learnings.

Why this paper? While most other work in the domain of causality-in-AI requires expensive hand annotation [2, 3, 4] to acquire commonsense knowledge owing to their exclusive use of the text modality, this paper proposes a unique direction in the field by utilizing both the visual and text modality to infer commonsense knowledge. On the flipside, there is work in the field [5] that generates commonsense using self-supervised methods (thus obliterating the need for expensive hand annotation), but is limited to the visual modality. The paper thus serves as a great baseline for pushing the envelope for causality-in-AI.

Wider research context. Causality helps identify the cause-and-effect relationship between events, which can improve a model's understanding of the happenings in a given video sequence. Furthermore, the problem of causality-in-AI has broad applicability to several vision and text applications. The proposed model in the paper can help improve the robustness of downstream tasks that suffer from limited performance owing to the lack of causal relationships such as video captioning, video question answering, etc. [5].

3 Project description (1-2 pages)

Goal. The top-level goals of this project are:

1. A model that can infer causal relationships using both the visual (specifically, videos) and text modality. Videos prevalently contain commonsense knowledge that cannot be easily inferred using just text because such information is not usually explicitly specified in textual form. For e.g., consider a video wherein a girl throws a frisbee in the air (event X) and a dog jumps to catch it (event Y). In this case, there exists a causal relationship between the two events (event $X \rightarrow$ event Y). In other words, the fact that the girl threw a frisbee (event X) led to the dog jumping (event Y) is apparent from a video of the activity but a textual caption of the entire sequence would usually fail to explicitly specify this relationship. Figure 1 traces this example through our proposed model.
2. A model that can rationalize and thus help us understand it's learnings using natural text. This would help perform error analysis but more importantly, also biases in the dataset.

Task. The ultimate goal of this work is to mine contextual causality knowledge from videos. We formally define the task as follows:

1. The input to the model is a video sequence, which is fed to the canonical frame identification module, which outputs an image pair $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of image pairs from each event $e \in E$, where E is the set of all events in the video. P thus consists of two images I_1 and I_2, sampled from the video V, in temporal order (i.e., I_1 appears before I_2).
2. For each P, our goal is to identify all possible causal relations between I_1 and I_2. Normally, this task contains two sub-tasks: identifying events in images and identifying causality relation between contained events. The canonical frame identification module enables the first sub-task.

3. For the second sub-task, we assume that the event sets contained in I_1 is denoted as $\mathcal{E}_1$ and the event sets contained in all images sampled from V_1 is denoted as $\mathcal{E}_v$. For each event $e_1 \in \mathcal{E}_1$, our goal is finding all events $e_2 \in \mathcal{E}_v$ such that e_1 causes e_2.
4. Finally, we rationalize our causality output using the causality rationalization module which accepts our causality prediction from the prior step along with e_1, e_2 and the outputs a string explaining our rationale behind the prediction.

Data. We plan to use the Vis-Causal dataset [1] for training our model, which contains 4,000 image pairs sampled from 1,000 videos obtained from ActivityNet [6]. Each video thus has four uniformly sampled screenshots. Further, each of the images have three annotations leading to 12,000 events for the aforementioned 4,000 image pairs. Overall, the Vis-Causal dataset contains (i) uniformly sampled frames from raw video data and, (ii) causal annotations for events localized from the frames. We do not plan to perform any preprocessing on the dataset.

Link to dataset: `https://github.com/HKUST-KnowComp/Vis_Causal`

Methods. Figure 1 illustrates the overall block diagram of our proposed model.

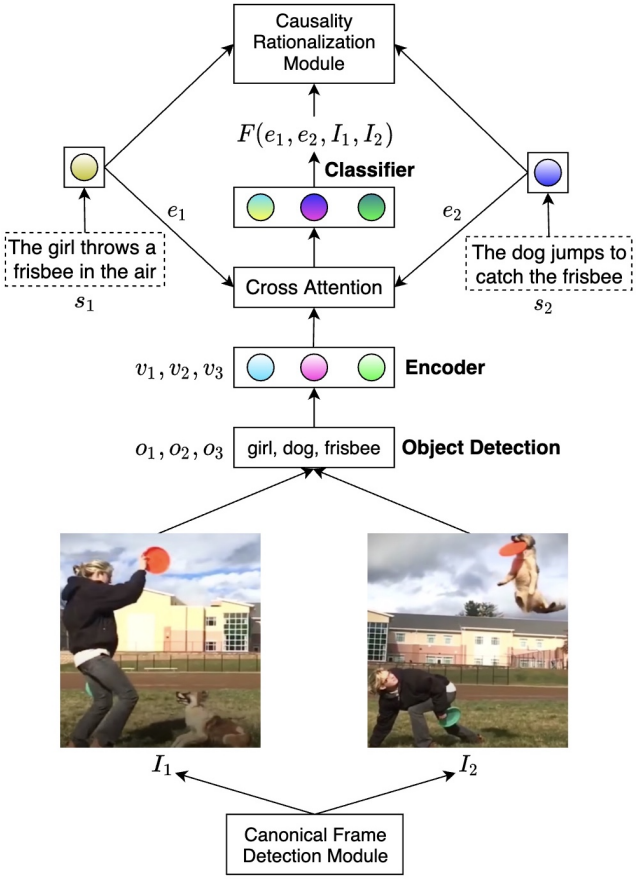

Figure 1: Network architecture

We propose a novel, enhanced and end-to-end trainable model that offers the following advantages:

1. **Canonical frame detection module**: To remedy this, we plan to adopt the event localization algorithm proposed in Chadha et al. [5] and enable the model to pick a representative image each from two adjacent events to feed the model as an input. Note that while [1] also utilize a pair of images as input to the model, the events are hand-picked. They are thus error-prone, lack overlapping events and are simply extremely limited (two to three events per video on an average) in comparison to what the event localization module in [5] can identify, which are much more exhaustive (seven to ten events per video on an average) and contain a lot of

overlapping sequences. This would enable our model to natively accept videos, learn much more causal relationships and facilitate end-to-end training. We plan to work on interfacing the algorithm in [5] with the baseline model in [1].

2. **Causality rationalization module**: We enhance the interpretability and robustness of the model proposed in the primary paper [1] by integrating the causality rationalization architecture proposed in [3]. This would not only ease the process of error analysis and correction, but also instill confidence in the model's predictions. Again, this module would be end-to-end trained as part of the model.

Baselines. We plan to use [1] as our primary baseline. If time permits, we also plan to compare [3] using variations of Commonsense AutoGenerated Explanations (CAGE) proposed in the paper. Comparisons would be with the published scores in the respective papers.

We also plan to carry out the following ablation experiments to isolate the effect of the following aspects on the performance of our model:

1. **No visual context:** Predict causal relationships without using the visual modality, i.e., limit the model to only use text.
2. **No automated canonical frame detection:** Disable canonical frame detection module and use static canonical frames as proposed in [1].
3. **No attention:** Remove the cross-attention module and use the average word embeddings of all selected objects to represent the context.

Evaluation. To compare with Zhang et al. [1], we plan to use Recall@N (R@N) as our evaluation metric, where N denotes whether the correct caused event is covered by the top one, five, or ten ranked events. [1] achieved 8.87, 34.75 and 63.12 on R@1, R@5 and R@10 respectively. If time permits, we plan to compare with Rajani et al. [3] using accuracy as our evaluation metric. [3] report 58.2% accuracy. on We expect our model to perform better owing to the automated canonical frame detection module (which enables video as native inputs, rather than hand-selected frames) and rationalizing module (which offers better error analysis and correction).

References

[1] Hongming Zhang, Yintong Huo, Xinran Zhao, Yangqiu Song, and Dan Roth. Learning contextual causality from time-consecutive images. *arXiv preprint arXiv:2012.07138*, 2020.

[2] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning, 2020.

[3] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

[4] Yuhao Wang, Vlado Menkovski, Hao Wang, Xin Du, and Mykola Pechenizkiy. Causal discovery from incomplete data: a deep learning approach. *arXiv preprint arXiv:2001.05343*, 2020.

[5] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering, 2020.

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.