Learning Contextual Causality between Daily Events from Time-consecutive Images

Anonymous CIV 2021 submission

Paper ID 7

Abstract

Conventional textual-based causal knowledge acquisition methods typically require laborious and expensive human annotations. As a result, their scale is often limited. Moreover, as no context is provided during the annotation, the resulting causal knowledge records (e.g., ConceptNet) typically do not consider the context. In this paper, we jump out of the textual domain to explore a more scalable way of acquiring causal knowledge and investigate the possibility of learning contextual causality from the visual signal. In detail, we first propose a high-quality dataset Vis-Causal and then conduct experiments to demonstrate that with good language and visual representations, it is possible to discover meaningful causal knowledge from the videos. Further analysis also shows that the contextual property of causal relations indeed exists, taking which into consideration might be crucial if we want to use the causal knowledge in real applications.

1. Introduction

Humans possess a basic knowledge about facts and understandings for commonsense of causality in our everyday life. For example, if we leave five minutes late, we will be late for the bus; if the sun is out, it's not likely to rain; and if we are hungry, we need to eat. Causal relations in the commonsense domain typically appear between daily eventualities (i.e., events and states) and are generally contributory and contextual [1]. By contributory,[1] we mean that the cause is neither necessary nor sufficient for the effect, but it strongly contributes to the effect. By contextual, we mean that some causal relations only make sense in a certain context. The contextual property of causal relations is important for both the acquisition and application of causal knowledge. For example, if some people tell the AI assistant (e.g., Siri) "they are hungry" in a meeting, a basic assistant may suggest that they order food because it knows that "being hungry" causes "eat food." A better assistant may recommend ordering food **after** the meeting because it knows that the causal relation between 'being hungry' and 'eat food' may not be plausible in the meeting context. Without understanding the contextual property of causal knowledge, achieving such a level of intelligence would be challenging.

To help machines better understand the causality commonsense, many efforts have been devoted to developing the causal knowledge bases. For example, ConceptNet [5] leverages human-annotation to acquire the causal knowledge. Even though the annotated causal knowledge bases are of high quality, their effect is limited by the small scales. Moreover, none of these KGs take the aforementioned contextual property of causal knowledge into consideration, which may restrict their usage in downstream tasks.

In this paper, we propose to ground causal knowledge into the real world and explore the possibility of acquiring causal knowledge from visual signals (i.e., images in time sequence, which are cropped from videos). By doing so, we have three significant advantages: (1) Videos can be easily acquired and can cover rich commonsense knowledge that may not be mentioned in the textual corpus; (2) Events contained in videos are naturally ordered by time. As discussed by [6], there exists a strong correlation between temporal and causal relations, and thus such time-consecutive images can become a dense causal knowledge resource; (3) Objects from the visual signals can act as the context for detected causal knowledge, which can remedy the aforementioned "lack of contextual property" issue of existing approaches.

2. The Task Definition

The goal is to acquire contextual causal knowledge from videos. However, as current models cannot afford to process videos directly, we simplify the task into mining causal knowledge from time-consecutive frames (i.e., images), which are cropped from the video. Thus, we formally define the task as follows. Each image pair $P \in \mathcal{P}$, where $\mathcal{P}$ is

[1]The other two levels are absolute causality (the cause is necessary and sufficient for the effect) and conditional causality (the cause is necessary but not sufficient for the effect), which commonly appear in the scientific domain rather than our daily life.

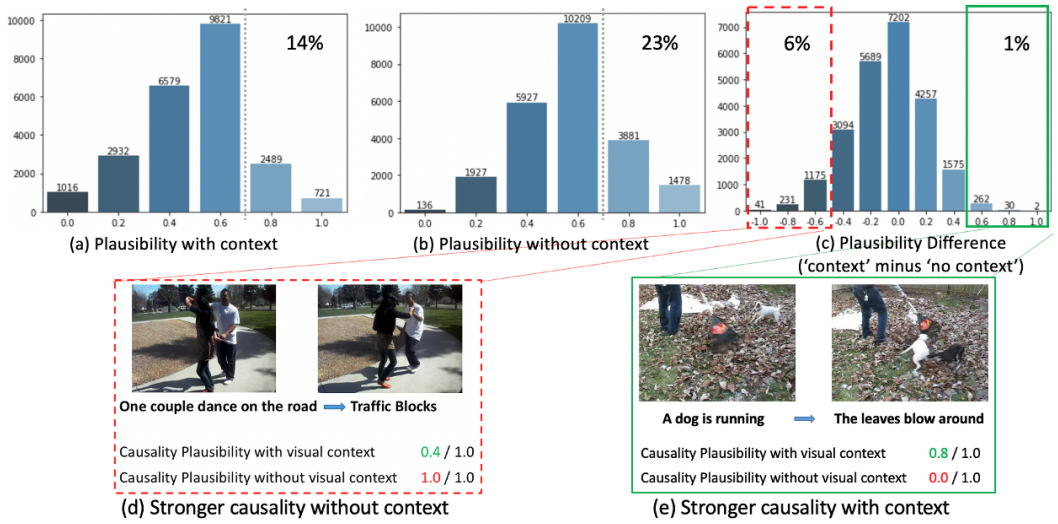


Figure 1: Distribution of plausibility scores under different settings and their difference.

the overall image pair set, consists of two images I_1 and I_2, sampled from the same video, in temporal order (i.e., I_1 appears before I_2). For each P, our goal is to identify all possible causal relations between the contained images. Normally, this task contains two sub-tasks: identifying events in images and identifying causality relation between contained events. As there exists a huge overlap between the event identification task and the scene graph generation task [8], which has been extensively studied [9], in this work, we focus on the second sub-task. We assume that the event set contained in I_1 is denoted as $\mathcal{E}_1$, and the set of all events 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.

3. The Vis-Causal Dataset

In this section, we introduce the details about the creation of Vis-Causal as the following steps[2].

1. **Data Pre-processing**: We use ActivityNet [4], which contains short videos from YouTube, as the video resource. We randomly select 1,000 videos. We take five uniformly sampled screen-shots for each video and take adjoined screen-shots as pairs of time-consecutive images to capture the chained events better. As a result, we collected 4,000 image pairs.

2. **Event Identification**: We then invite annotators to write down any events they can identify in the first image. We invite three annotators for each image pair, resulting in 12,000 events in total for 4,000 image pairs.

3. **Causality Annotation**: For each pair of time-consecutive images, we select all three identified events in the first image, and for each one of them, we ask annotators to describe one event that happens in the second image and is caused by the selected event, which occurs in the first image. For each question, we invite three different annotators to provide annotations. After filtering out answers that contain 'None' or have less than two words, we obtain 23,558 event pair candidates. To investigate the contextual property of causal knowledge, we invite annotators to annotate whether the visual context can influence the causal relation in their mind. Specifically, for each identified event pair, we invite five annotators to annotate if there is a causal relation with or without the context. We employ Inter Annotator Agreement (IAA), which computes the average agreement of an annotator with the average of all other annotators, to evaluate the overall annotation quality. As a result, we achieve 78% and 76% IAA scores for "*with context*" and "*without context*" settings, respectively.

The distribution of annotation results for both settings are shown in Figure 1(a) and Figure 1(b) respectively. For each pair of events, we compute the plausibility based on voting. In general, we can see that the majority of the candidate events pairs have weak causal relations for both settings, and only a small portion of the candidates contain strong causal relations, especially for the "*with context*" setting. To investigate the contextual property of causal relations, we show the distribution of plausibility difference ("*with context*" minus "*without context*") in Figure 1(c). From the result, we can observe that about 6% of event pairs, which is indicated with the dashed box, have stronger causal relations without any context, while about 1% of event pairs, which is indicated with the solid box, have a stronger causal relation

[2]We select Amazon Mechanical Turk as the annotation platform and clear instructions were given for all annotations.

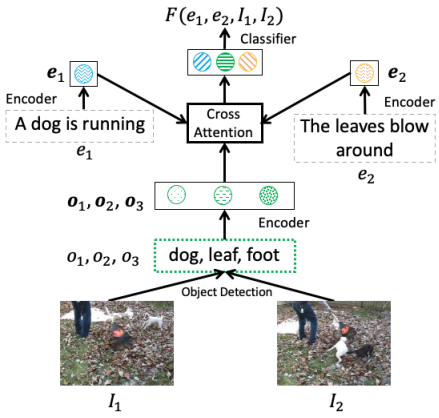


Figure 2: Demonstration of the proposed model.

when the visual context is provided. Two examples of both cases are shown in Figure 1(d) and 1(e) respectively.

We split the dataset into the train, dev, and test sets based on the original split of ActivityNet [4] and collect 800, 100, and 100 videos, respectively. We select positive causal relations based on the annotation under the "*with context*" setting. If at least four of five annotators think there exists a causal relation between a pair of events given the context, we will treat it as a positive example. As a result, we got 2,599, 329, and 282 positive causal pairs for the train, dev, and test set, respectively. On average, each event pair contains 11.41 words, and the total vocabulary size is 10,566.

4. The VCC Model

This section introduces the proposed Vision-Contextual Causal (VCC) Model, which leverages both the visual context and contextual representation of events to predict the causal relations. We show the overall framework in Figure 2. In total, we have three major components: event encoding, visual context encoding, and cross attention. The details about these components are introduced as follows.

4.1. Textual Event Encoding

As both e_1 and e_2 are represented with natural language, we begin with converting them into vector representations. In this work, we leverage a pre-trained language representation model BERT [2] to encode all events. Assuming that after the tokenization, event e contains n tokens $w_1, w_2, ..., w_n$, we denote their contextualized representations after BERT as $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n$.

4.2. Visual Context Encoding

Following the standard approach in multi-modal approaches [8], we first leverage an object detection module to detect objects from images and use all extracted objects to represent the visual context. Assuming that for I_1 and I_2, we extract m_1 and m_2 objects, respectively. After combining all objects from two images together and sorting them based on the confidence score provided by the object detection module, we keep the top m objects and denote them as $o_1, o_2, ..., o_m$. As all objects are in the form of words, to align with events, we use the same pre-trained language representation model to extract the vector representation[3] of selected objects and denote them as $\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_m \in \mathcal{O}$.

4.3. Cross-Attention Module

The cross-attention module aims to minimize the influence of noise by selecting important context objects with events and informative tokens in events with the context. Thus, the cross-attention module contains two sub-steps: (1) context representation; (2) event representation.

Context Representation: For each event e, whose tokens' vector representations are $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n$, we first take the average of all tokens and denote the resulted average vector as $\widetilde{\mathbf{w}}$. As the vector representation set of all selected objects is denoted as $\mathcal{O}$, we compute the overall context representation as:

$$\mathbf{o} = \sum_{\mathbf{o}' \in \mathcal{O}} a_{\widetilde{\mathbf{w}}, \mathbf{o}'} \cdot \mathbf{o}', \tag{1}$$

where $a_{\widetilde{\mathbf{w}}, \mathbf{o}'}$ is the attention weight of $\widetilde{\mathbf{w}}$ on object o'. Here we compute the attention weight as:

$$a_{\widetilde{\mathbf{w}}, \mathbf{o}'} = NN_a([\widetilde{\mathbf{w}}, \mathbf{o}']), \tag{2}$$

where NN_a is a standard two-layer feed forward neural network and $[,]$ indicates the concatenation.

Event Representation: After getting the context representation, the next step is computing the event representation. Assuming that the vector set of e is $\mathcal{W}$, we can get the event representation with a similar attention structure:

$$\mathbf{e} = \sum_{\mathbf{w}' \in \mathcal{W}} b_{\mathbf{o}, \mathbf{w}'} \cdot \mathbf{w}',$$
$$b_{\mathbf{o}, \mathbf{w}'} = NN_b([\mathbf{o}, \mathbf{w}']),$$

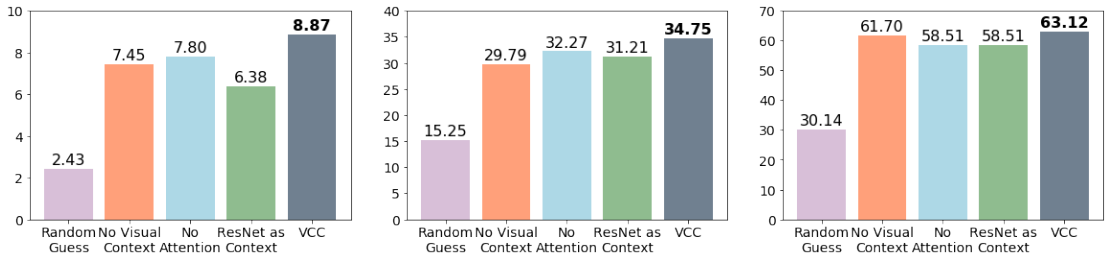
where b is the attention weight we computed with another feed forward neural network NN_b.

4.4. Causality Prediction

Assuming that the context representations with e_1 and e_2 as attention signal are denoted as $\mathbf{o}_{e_1}$ and $\mathbf{o}_{e_2}$ respectively and the overall representations of e_1 and e_2 are $\mathbf{e}_1$ and $\mathbf{e}_2$, we can then predict the final causality score as follows:

$$F(e_1, e_2, I_1, I_2) = NN_c([\mathbf{e}_1, \mathbf{e}_2, \mathbf{o}_{e_1}, \mathbf{o}_{e_2}]). \tag{3}$$

[3] If an object word is tokenized to multiple tokens, we take their average representation as the token representation.



(a) Recall@1 (b) Recall@5 (c) Recall@10

Figure 3: Experimental results on Recall@1, Recall@5, and Recall@10, respectively.

5. The Experiment

As each event in the first image could cause multiple events in the second image, we evaluate different causality extraction models with Recall@1, Recall@5, and Recall@10, and compare with the following models:

1. **No Visual Context**: Directly predicts the causal relation between events without considering the visual context. We take the average of word representations as the event representation and concatenate the representations of two events together for the final prediction for each event.

2. **No Attention**: Removes the cross-attention module and uses the average word embeddings of all selected objects to represent the context.

3. **ResNet as Context**: Removes the object detection module and uses the average image representation extracted by ResNet-152 [3] as the context representation. We acquire event representation as same as the 'no context' setting and concatenate the representation of context and events together for the final prediction.

We use the pre-trained BERT [2] as the textual representation model and follow the previous scene graph generation work [8] to leverage a Faster R-CNN network [7] to detect objects. We set the hidden state size in the feed-forward neural network to be 200 and the number of selected objects m to be 10. During the training phase, for each positive example, we randomly select one negative example and use cross-entropy as the loss function. We employ stochastic gradient descent (SGD) as the optimizer. All parameters are initialized randomly, and the learning rate is set to be 10^{-4}.

From the results in Figure 3 we can make the following observations: (1) All models significantly outperform the "random guess" baseline, which shows that models can learn to extract meaningful causal knowledge from these time-consecutive images; (2) With the help of the context information, VCC outperforms the baseline 'No Context' model in most experiment settings, proving the importance of visual context and is consistent with our previous observation that some causality only makes sense in certain contexts; (3) All proposed components contribute to the final success.

6. Conclusion

In this paper, we explore the possibility of learning causal knowledge from time-consecutive images. To do so, we first formally define the task and then create a high-quality dataset Vis-Causal. On top of the collected dataset, we propose a Vision-Contextual Causal (VCC) model to demonstrate that with the help of strong pre-trained textual and visual representations and careful training, it is possible to directly acquire contextual causality from visual signals. Both the dataset and code will be released to encourage research on causality acquisition.

References

[1] Mario Bunge. *Causality and modern science*. Routledge, 2017. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019. 3, 4

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR 2016*, pages 770–778, 2016. 4

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of CVPR 2015*, pages 961–970, 2015. 2, 3

[5] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 1

[6] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. Joint reasoning for temporal and causal relations. In *Proceedings of ACL 2018*, pages 2278–2288, 2018. 1

[7] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 4

[8] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of CVPR 2017*, pages 3097–3106, 2017. 2, 3, 4

[9] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *Proceedings of ECCV 2018*, pages 690–706, 2018. 2