iReason: 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

Abstract

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 not immediately available in the input. Prior work has unraveled spurious observational biases that models fall prey to in the absence of causality. While language representation models such as BERT and GPT-2 can preserve rich knowledge within their learned embeddings, they lack exploiting causal relationships during the training process. Upon factoring in causal relationships, downstream tasks such as dense video captioning, video question-answering etc. tend to perform much better owing to the insight causal relationships bring about. Recently, there have been several proposed models that have undertaken the task of mining causal data from either the visual or textual modality. However, there does not exist widespread prevalent research that infers and mines causal relationships by juxtaposing the visual and textual modalities. While videos provides a rich, readily-available and naturally time-ordered resource for us to mine causality knowledge from, textual information offers details that could be implicitly implied in videos. We propose iReason, a model that is able to localize events in videos, draw on canonical frames that represent these events and learn causality from both videos and textual captions. Furthermore, our model architecture integrates a causal rationalization module to aid the process of explainability, error analysis/correction and bias detection. Finally, we demonstrate the effectiveness of our technique by comparing our results to our baseline and show that our approach furthers the state-of-the-art.

1 Key Information to include

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

2 Approach

2.1 Network Architecture

Figure 1 offers an architectural overview of iReason. We propose an end-to-end trainable model that offers the following novel advantages:

1. **Canonical frame detection module**: We adopted and successfully interfaced the event localization module in [1] with our baseline model in [2] to enable iReason to automatically identify a canonical image each from two events (vs. statically chosen as in [2]) as an

input to our model. This enabled our model to natively accept videos, learn deeper causal relationships and facilitated end-to-end training with video input.

2. **Causality rationalization module**: We plan to enhance the interpretability and robustness of the model proposed in our baseline [2] by integrating the causality rationalization architecture proposed in Rajani et al. [3]. This module is currently work-in-progress.

Similar to [2], we use BERT [4] to encode textual representations. Furthermore, we leverage Faster R-CNN [5], trained on MS-COCO [6], to perform object detection on the canonical frames.

2.2 Loss function

For each positive example in the Vis-Causal dataset [2], we randomly select one negative example and use cross-entropy as the loss function. Formally,

$$L = CrossEntropy\left(I'_i, I_j\right)$$

where, I' is the i^{th} positive sample and I_i is the j^{th} randomly selected negative sample.

3 Experiments

3.1 Data

For causality inference, we used the Vis-Causal dataset [2], which contains 4,000 image pairs and 12,000 annotated events. For causality rationalization, we plan to use the pre-trained model in [3] trained on the Common Sense Explanations (CoS-E) dataset.

3.2 Experimental details

Table 1 presents details of the training knobs and hyperparameters.

Table 1: Training config

Optimizer	Stochastic gradient descent
Parameter initialization	Random
Learning rate	10^{-4}
Training time	22 min/epoch $\times$ 10 epochs = 3.8 hours
Total trainable parameters	112.1 million (including 109.48 million from BERT-base)

3.3 Results

Table 2 compares iReason for the various context categories with our baseline [2]. We used Recall@N (R@N) as our evaluation metric, where N denotes whether the correct causal event is covered by the top one, five, or ten ranked events.

Table 2: Performance comparison of iReason. Bold numbers indicate best performance.

Model	Metric	Sports	Socializing	Household	Personal Care	Eating	Overall
VCC [2]	R@1	8.78	7.27	6.78	11.11	27.27	8.87
	R@5	37.16	36.36	28.81	33.33	45.45	34.75
	R@ 10	64.86	58.18	62.71	55.56	72.73	63.12
iReason	R@1	**9.27**	**8.09**	**7.91**	**12.72**	**28.89**	**9.21**
	R@5	**38.71**	36.36	**29.92**	**34.73**	**45.75**	**35.87**
	R@ 10	**65.12**	**58.52**	62.71	**55.86**	72.73	**63.51**

4 Future work

1. Interfacing the causality rationalization module [3] with the current model and implementing the causality rationalization loss. If time permits, comparing results based on accuracy.
2. We also plan to carry out the ablation experiments (no visual context, no automated canonical frame detection and no attention) to isolate their effect on the performance of our model.

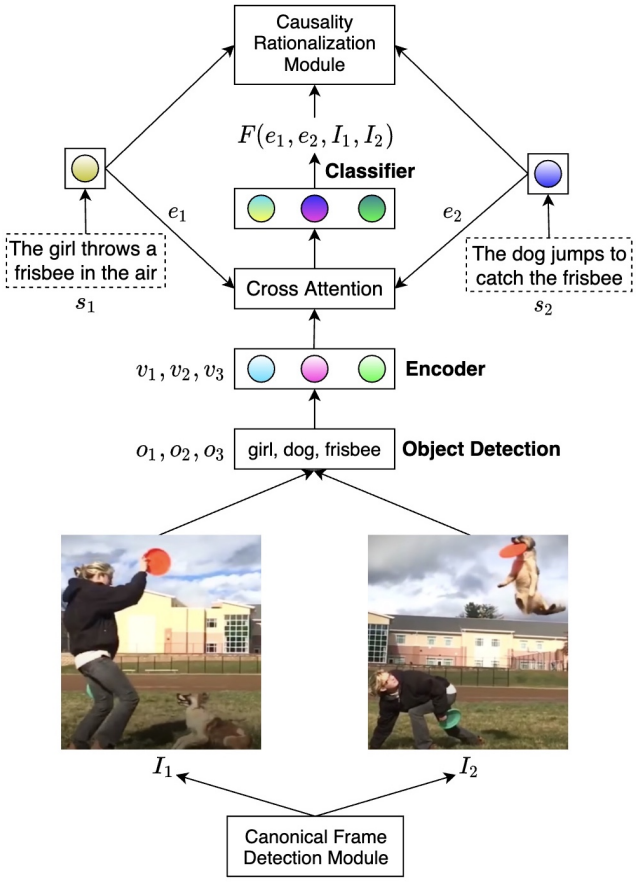

Figure 1: Network architecture

References

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

[2] Hongming Zhang, Yintong Huo, Xinran Zhao, Yangqiu Song, and Dan Roth. Learning contextual causality from time-consecutive images. *arXiv preprint arXiv:2012.07138*, 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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.