iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability

Stanford CS224N Custom Project | Mentor: Andrew Wang

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 causal reasoning, which is often derived using some form of commonsense knowledge not immediately available in the input but implicitly inferred by humans. Prior work has unraveled spurious observational biases that models fall prey to in the absence of causality. While language representation models preserve contextual knowledge within learned embeddings, they do not factor in causal relationships during training. By blending causal relationships with the input features to an existing model that performs visual cognition tasks (such as scene understanding, video captioning, video question-answering, etc.), better performance can be achieved owing to the insight causal relationships bring about. Recently, several models have been proposed that have tackled the task of mining causal data from either the visual or textual modality. However, there does not exist widespread prevalent research that mines causal relationships by juxtaposing the visual and language modalities. While images offer a rich and easy-to-process resource for us to mine causality knowledge from, videos are denser and consist of naturally time-ordered events. Also, textual information offers details that could be implicit in videos. As such, we propose iReason, a framework that infers visual-semantic commonsense knowledge using both videos and natural language captions. Furthermore, iReason's architecture integrates a causal rationalization module to aid the process of interpretability, error analysis and bias detection. We demonstrate the effectiveness of iReason using a two-pronged comparative analysis with language representation learning models (BERT, GPT-2) as well as current state-of-the-art multimodal causality models. Finally, we present case-studies attesting to the universal applicability of iReason by incorporating the "causal signal" in a range of downstream cognition tasks such as dense video captioning, video question-answering and scene understanding and show that iReason outperforms the state-of-the-art.

1 Introduction

"On the contrary, Watson, you can see everything. You fail, however, to reason from what you see."
- Sherlock Holmes, The Adventure of the Blue Carbuncle

As humans, a lot is said without explicit connotation [1, 2]. Humans often possess basic know-how about facts related to the environment we are in and the world at large. 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 people are walking on the road, they're using their legs to do so. Humans learn commonsense in an unsupervised fashion by exploring the physical world, and until machines imitate this learning path by imbibing the contextual property of causal knowledge in their understanding, there will be an inevitable "gap" between man and machine.

The aforementioned implicit knowledge fosters *commonsense of causality* [3] in everyday life. Causality helps identify the cause-and-effect relationship between events, which enables gaining deeper insights about not only the casual connections between the events themselves but also of the environment in which these events occur. This has the effect of improving the understanding of the happenings in a real-life events depicted through a video or a natural language snippet, not just for humans but also for deep-learning models [4, 5].

Prior work [6, 7] has unraveled spurious observational biases that models fall prey to in the absence of causality. Causal relationships remedy this by helping point out contextual attributes - such as if there's *barking* noise, a *dog* should be present - that are usually implied for humans. The problem of causality-in-AI [4] thus has broad applicability to a wide gamut of vision and text-based tasks. Specifically, causality-in-AI can help improve the robustness of downstream tasks that suffer from limited performance owing to the lack of understanding causal relationships such as dense video captioning [8, 9, 10], video question-answering [11, 12, 13, 14], and a plethora of other NLP tasks [15, 16, 17, 18, 19, 4], etc. This makes it valuable to impart the notion of causality to machines.

While most other work in the domain of causality-in-AI requires expensive hand annotation [20, 19, 5] to acquire commonsense knowledge owing to their exclusive use of the text modality, relatively little work exists in literature that utilizes visual modalities. On the flipside, while there is work in the field [4] that generates commonsense using self-supervised methods (thus obliterating the need for expensive hand annotation), but it is limited to the visual modality (and thus doesn't imbibe learnings from natural language snippets – say captions – using NLP). Zhang et al. [18] propose a unique direction in causality-in-AI by proposing a vision-contextual causal (VCC) model that utilizes both the visual and language modality to infer commonsense knowledge. However, under the visual modality, [18] limits commonsense knowledge generation from images and cannot natively accept videos as input, which are a much more prevalent source of commonsense knowledge compared to images and text. Instead, the model utilizes a pair of randomly-selected images to be able to infer commonsense knowledge which limits its effectiveness.

Furthermore, the ability to *rationalize* causal relationships is instrumental to not only ease the process of error analysis but also instill confidence in the model's predictions. VCC [18] doesn't offer rationalization for its causality inference, thereby hindering the model's interpretability. 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., act as a 'debug signal' to develop an understanding of the model's (mis)learnings.

To push the envelope for causality-in-AI, we propose iReason, a framework that generates commonsense knowledge by inferring the causal relationships using two of the most knowledge-rich modalities – videos and text. This enables the model to seek intrinsic causal relationships between objects within events in a video sequence and supplement the knowledge thus gained using natural language snippets, i.e., captions of the aforementioned events. To demonstrate that iReason furthers the state-of-the-art, we offer hands-on evaluation by comparing our results to textual representation learning models (BERT, GPT-2) in addition to the current state-of-the-art causality models. Furthermore, we present case-studies by incorporating the "causal signal" in downstream cognition tasks such as dense video captioning and video question-answering and show that imbibing causality knowledge using iReason into the aforementioned tasks helps them outperform the current state-of-the-art.

In summary, our key contributions are centered around the following.

1. **Commonsense reasoning using videos and natural language:** iReason infers causal knowledge grounded in videos and natural language. We envision this as a step towards human-level causal learning. As such, iReason as a dual-grounded causality learning approach offers the following advantages:
   (a) **Causality using the visual 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 [3]. For e.g., consider a video of a girl throwing a frisbee in the air (event X) and a dog jumping to catch it (event Y). In this case, there exists a causal relationship between the two events (event $X \rightarrow$ event Y). While a textual caption of the entire sequence would be helpful in understanding the events, it would typically fail to explicitly specify this relationship. However, the fact that the girl threw a frisbee (event X) *led* to the dog jumping (event Y) would be apparent from the video. As such, both modalities hold their unique importance in the

task of learning causal relationships. iReason thus seeks to blend both the visual and text modality to mine commonsense. Figure 1 traces this example through iReason.

   (b) **Exploit the time-ordered nature of videos:** There exists a strong correlation between temporal and causal relations (say A is the cause and B is the effect, then A has to precede B in time). Since events in most video sequences are naturally time-ordered, they are an apt resource for us to mine cause-and-effect relationships from.
   (c) **Use objects in videos to develop an idea of contextual causality:** Objects in videos can be used as environmental context to understand causal relations in the scene.

2. **Offer interpretability and error detection:** iReason can rationalize causal relationships and thus help us understand its learnings using natural text. This would help perform error analysis and more importantly, also spot biases in the dataset.
3. **Universal applicability to cognition tasks:** iReason's commonsense features can be incorporated in downstream tasks that require cognition by supplementing them with the input features (cf. Section). It is thus noteworthy that they have a certain universality and are not limited to the realizations of DVC and VideoQA discussed in this work. As such, they can be easily adapted for other video-based vision tasks such as scene understanding [21], panoptic segmentation [22], etc.

iReason thus infers visual-semantic causal knowledge by blending videos and natural language to perform multimodal commonsense reasoning. This is accomplished by localizing events in videos, drawing on canonical frames that represent these events, and learning contextual causal relationships using both videos and text captions. Furthermore, iReason's architecture integrates a causal rationalization module to aid the process of interpretability, error analysis and bias detection.

2 Related Work

Inferring causal relationships to bolster machine intelligence has been an area that has been under the spotlight in recent times owing to it being a significant step towards artificial general intelligence (AGI) [3, 23, 24]. Commonsense knowledge can be derived using either the language and/or visual (images or video) modality and literature in the field can thus be reviewed in a similar fashion.

2.1 Causality in Natural Language

Several approaches [25, 26, 27, 20, 19, 5] have been proposed that extract causal knowledge using natural language. These approaches either mine textual snippets such as captions, text blurbs or large-scale knowledge bases such as Wikipedia. However, causality grounded in natural language has the obvious disadvantage of being limited by the reporting bias [1, 2] (for e.g., washing a car *leads to* the car being clean is something that is not explicitly mentioned in text, but can easily be visually inferred) and thus suffer from sub-standard performance.

2.2 Causality in Vision

There has been a recent surge of interest in coupling the complementary strengths of computer vision and causal reasoning [28, 29]. The union of these fields has been explored in several contexts, including image classification [30, 31], reinforcement learning [32], adversarial learning [33], visual dialog [34], image captioning [35] and scene/knowledge graph generation [36, 37]. While these methods offer limited task-specific causal inference, current research that tackles the task of building a generic commonsense knowledge base from visual input mainly falls into two categories: (i) learning from images [38, 1, 39, 40] and (ii) learning actions from videos [41]. While the former limits learning to human-annotated knowledge which restricts its effectiveness and outreach, the latter is essentially learning from correlation.

2.3 Multimodal Causality

With the recent success of pre-trained language models [42, 43, 44] in NLP, several approaches [45, 46, 47, 48] have emerged that utilize weakly-supervised learning models using large, unlabelled, multimodal (images and natural language) data to encode visual-semantic knowledge. However, the inordinate memory cost for task-specific finetuning is a significant barrier-to-entry for such systems.

3 Task Definition

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

1. The input to the model is a video, 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 frames I_1 and I_2, sampled from the video V, in temporal order (i.e., I_1 appears before I_2, and thus I_1 and I_2 are the cause and effect frames respectively).
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: (i) identifying events in frames and, (ii) identifying causal relations between the said events.
3. The canonical frame identification module (cf. Section 4) enables the first sub-task. For the second sub-task, we assume that the set of events contained in I_1 is denoted as $\mathcal{E}_1$ and the set of events contained in all frames 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 ($e_1 \rightarrow e_2$).
4. The output of the model is a causality score prediction $C \in [0, 1]$ interpreted as a probability measure. By setting a compliance threshold for c (say, 0.5), positive ($e_1 \rightarrow e_2$) and negative ($e_1 \not\rightarrow e_2$) causal relationships can be inferred.
5. Finally, we rationalize our causality output using the causality rationalization module which accepts the output c from the prior step along with e_1, e_2 and the outputs a string explaining our rationale behind the prediction.

4 Network Architecture

We propose an end-to-end trainable model that utilizes the following modules: canonical frame detection, which identifies representative frames corresponding to events; textual event and object encoder, which encodes the two input events and detected objects from the two context frames into vectors; cross attention [49], which seeks to find context and event representation; and causality rationalization, which fosters model interpretability. Figure 1 illustrates a top level view of iReason.

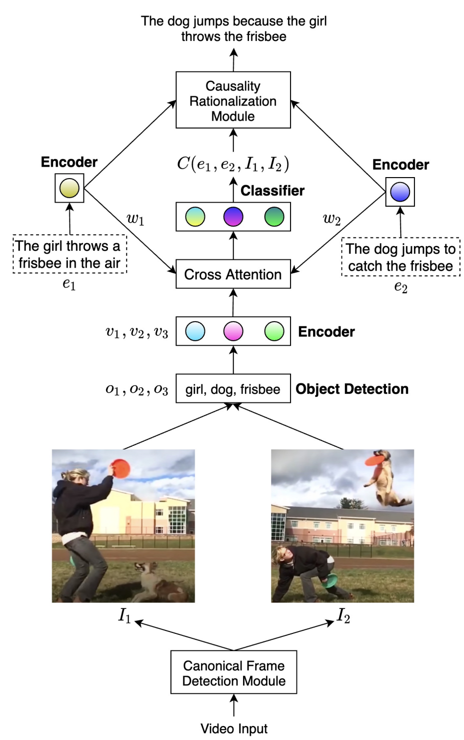


Figure 1: Architectural overview of iReason.

4.1 Canonical Frame Detection Module

While VCC [18] utilizes a pair of event frames as input to the model, they are sampled at uniform intervals and the events corresponding to the images are thus essentially randomly picked (since there is little correlation between the timing of event occurrences in different videos, say a dog jumping *in response to* a frisbee thrown in the air vs. the opening of a door *as a result of* the doorbell ringing). Statically selected event frames are thus error-prone, do not accommodate overlapping events and are limited in magnitude (2-3 events per video on an average). To remedy this, we utilize the event localization module in [4] to derive a set of events from the input video, which are much more exhaustive (7-10 events per video on an average) and contain both overlapping and non-overlapping events. Furthermore, we propose a canonical frame detection algorithm which enables the model to pick a representative frame given an event (vs. statically chosen as in [18]).

The canonical frame detection algorithm performs activity detection on frames within a particular localized event to infer visual context. We do so using a pre-trained model from [50]. Next, we perform a unigram-based BLEU score [51] match to identify the first frame with the best match of constituent objects compared to those in the event caption. Formally,

$$BLEU = BP \times \exp\left(\lambda_n \log p_n\right) \tag{1}$$

where, BP is the brevity penalty which is set to 1.

This enables iReason to natively accept videos, learn deeper causal relationships and facilitates end-to-end training with video input. Our canonical frame detection module thus has the distinct novel advantage of being able to establish causal relationships between a broad range of events – overlapping/non-overlapping, short/long, few/many, etc.

4.2 Textual Encoder Module

Similar to [18], we use BERT [42] to encode textual representations of events e_1 and e_2, denoted by w_1 and w_2. Following scene-graph approaches [52, 53], we leverage a pre-trained Faster R-CNN [54] model trained on MS-COCO [55], to perform object detection on the canonical frames and thus establish visual context. Next, we choose the top m object predictions sorted by their confidence score, where m is a hyperparameter (cf. Section 5.3). Finally, we encode the vector representations of the selected objects o_1, o_2, . . ., o_n using BERT and denote them as v_1, v_2, . . ., v_n.

4.3 Cross-Attention Module

The cross-attention module seeks to select (i) objects associated with events thus establishing context, and (ii) events associated with the aforementioned context.

Context Representation. For each event e, whose tokens' vector representations are $w_1, w_2, \ldots, w_n$, we first take the average of all tokens and denote the resultant average vector as $\widetilde{w}$. Next, with the set of all objects denoted as $\mathcal{O}$, we compute the context representation as:

$$o = \sum_{o' \in \mathcal{O}} a_{\widetilde{w},o'} \cdot o' \tag{2}$$

$$a_{\widetilde{w},o'} = NN_a\left([\tilde{w}, o']\right)$$

where $a_{\widetilde{w},o'}$ is the attention weight of $\widetilde{w}$ on object o' computed by a two-layer feed forward neural network NN_a and $[,]$ indicates concatenation.

Event Representation. Assuming that the set of events e is denoted by $\mathcal{W}$, we can get the event representation using a similar attention structure:

$$e = \sum_{w' \in \mathcal{W}} b_{o,w'} \cdot w' \tag{3}$$

$$b_{o,w'} = NN_b\left([o, w']\right)$$

where $b_{o,w'}$ is the attention weight computed by another feed forward neural network NN_b.

4.4 Causality Score Prediction

Assuming that the context representations for e_1 and e_2 are denoted as o_{e_1} and o_{e_2} respectively, we can predict the final causality score using a binary classifier NN_c as:

$$C\left(e_1, e_2, I_1, I_2\right) = NN_c\left([w_1, w_2, o_{e_1}, o_{e_2}]\right) \tag{4}$$

4.5 Causality Rationalization Module

We enhance the interpretability and robustness of iReason by adapting the commonsense auto-generated explanations (CAGE) framework proposed by Rajani et al. [19] which generates explanations for commonsense reasoning using natural language (cf. Section 5.1) for causal events e_1 and e_2. Given a question q, four answer choices c_0, c_1, c_2, and a labeled answer a, CAGE generates explanations e_i according to a conditional language modeling objective as follows:

$$-\sum_{i} \log P\left(e_i \mid e_{i-k}, \ldots, e_{i-1}, C_{txt}; \theta\right) \tag{5}$$

where, C_{txt} is the input context defined as “q, c_0, c_1 or c_2? commonsense says <explanation>” and k is its size; θ is the set of parameters conditioned on C_{txt} and previous explanation tokens;

We adapt CAGE for iReason by forming a question using e_2 by prepending it with “Why does …”. Next, we perform activity detection using a pre-trained model from [50] on I_1 and fetch the top three ranked activities. Table 1 offers insights into CAGE’s inputs/outputs using the example in figure 1.

Table 1: Adapting CAGE for iReason.

Input	Question	Why does the dog jump to catch the frisbee?
	Choices	girl throws frisbee, dog sits on grass, girl stands on grass
Output	CAGE	The dog jumps because the girl throws the frisbee.

4.6 Loss Function

For each positive example in the Vis-Causal dataset [18], we randomly select one negative example and use cross-entropy as the loss function. Formally,

$$J = CrossEntropy\left(I_i', I_j\right) \tag{6}$$

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

5 Experiments

5.1 Data

For learning causal relationships, we propose iReasonData, a dataset that amalgamates videos from the ActivityNet dataset [56] (which contains short videos from YouTube) and causal annotations from the Vis-Causal dataset [18] (which contains causal relationships between events in frames sampled from the aforementioned videos). This enables us to run the canonical frame detection module directly on the ActivityNet videos which serve as an input to iReason, thereby enabling iReason to select representative frames from the selected events. The output is a binary label indicating the causal relationship between the events pictured in the sampled canonical frames. We chose the same 1,000 videos obtained from the ActivityNet dataset [56] that Vis-Causal used and the corresponding causal annotations from Vis-Causal for the videos. Our training/validation/test split was 80%/10%/10%. Table 2 presents details of the iReasonData.

Table 2: Specifics of the iReasonData dataset.

Input	# of videos (from the ActivityNet dataset)	1,000
	# of frames sampled from each video	4
	Total # of frame pairs	1,000 (videos) $\times$ 4 (frames/video) = 4,000
	# of event annotations for each video	3
	Total # of event annotations	4,000 (videos) $\times$ 3 (annotations/video) = 12,000
Output	# of causal relationships	12,000 (annotations) $\times$ 1 (causal relationship) = 12,000

For learning causality rationalization, we used the pre-trained model in [19] trained on the CoS-E dataset, which consists of 9741 question-choices-answer samples (with a training/validation/test split of 80%/10%/10%). Table 3 explores the structure of a sample within CoS-E.

Table 3: Structure of a sample within the CoS-E dataset.

Input	Question; three answer choices; ground-truth answer
Output	Highlighted relevant words in the question that justify the ground-truth answer Brief explanation based on the highlighted justification that serves as the commonsense reasoning

5.2 Evaluation method

Since each event in the I_1 could cause multiple events in the I_2, we evaluate different causality extraction models (cf. Section 5.4) with a ranking-based evaluation metric [52]. Given an event e in I_1, models are required to rank all candidate events based on how likely they think these events are caused by e. We then evaluate different models using Recall@N (R@N), where N denotes whether the correct causal event is covered by the top one (R@1), five (R@5), or ten (R@10) ranked events.

5.3 Experimental Details

Table 4 presents details of the training knobs and hyperparameters [18].

Table 4: iReason's training config.

Optimizer	Stochastic gradient descent
Parameter initialization	Random
Learning rate	10^{-4}
Training platform	Azure NC6s_V3 (6x Intel Xeon E5-2690 and 1x NVIDIA Tesla V100)
Training time	22 min/epoch $\times$ 10 epochs = 3.8 hours
Early stopping	Yes
Number of detected objects considered	10
Total trainable parameters	112.1 million (including 109.48 million from BERT-base)

5.4 Results

Table 5 compares iReason for the various context categories with VCC [18], while table 6 compares iReason with language representation learning models (BERT [42], GPT-2 [57]). Table 7 offers a visual walkthrough of iReason with positive and negative examples.

5.5 Ablation Experiments

To prove that the multimodal context is crucial for learning causality, we carried out the following ablation experiments. Table 5 offers an in-depth view into these experiments.

1. **No lingual context:** Predict causal relationships without using the language modality, i.e., limit the model to only use videos.
2. **No visual context:** Predict causal relationships without using videos, i.e., limit the model to only use natural language captions.

Table 5: Performance analysis of iReason and ablated versions compared to VCC [18]. Bold numbers indicate best performance.

Model	Metric	Sports	Socializing	Household	Personal Care	Eating	Overall
Random guess	R@1	0.67	3.64	1.69	0.00	9.09	2.13
	R@5	14.19	16.36	15.25	11.11	27.27	15.25
	R@10	28.38	38.18	27.12	33.33	27.27	30.14
No lingual context (use only videos)	R@1	7.47	7.17	6.28	10.88	24.55	7.17
	R@5	35.15	34.12	27.12	31.22	39.11	30.64
	R@10	60.27	54.18	58.15	52.11	66.22	58.23
No visual context (use only natural language captions)	R@1	7.88	7.66	6.45	11.01	24.57	8.13
	R@5	36.11	34.36	28.77	32.17	42.77	31.22
	R@10	61.11	55.15	58.22	52.25	67.73	60.14
VCC [18]	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**

6 Analysis of Results

1. From table 5, we observe that iReason outperforms VCC [18] on R@1 across-the-board, and on R@5 and R@10 for most categories (while still maintaining the performance levels of VCC on R@5 and R@10 for the other categories). The fact that iReason outperforms VCC on R@1 implies that iReason is much more effective than VCC to identify the correct causal event using only the first ranked event. Put simply, in some cases, VCC offers the same performance levels as iReason for R@5 and R@10 (which implies requiring atleast five and ten of the top ranked events respectively to obtain the correct causal event) since the additional events increase the probability of VCC to identify the right answer.
2. From table 5, iReason outperforms the "no visual/lingual context" ablated models for all settings, which proves the importance of visual-semantic context. Also, since using only lingual context outperforms using only visual context, lingual context is more useful than visual context to learn commonsense knowledge. However, they offer complementary knowledge in most cases leading to iReason performing better than either ablated model.
3. From table 6, we see that all models significantly outperform the "random guess" baseline in almost all settings, which shows that the models have learned to extract meaningful semantic and/or causal knowledge from their respective input modalities.
   (a) Compared with the "random guess" baseline, BERT and GPT-2 achieve slightly better performance. This implies that even though these pre-trained language representation models encode semantic information about events, they can only identify the probabilistic relevance between the events, rather than their causal relationship.
   (b) iReason outperforms VCC [18] which indicates that canonical frame detection helps mine additional causal knowledge compared to VCC.

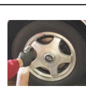
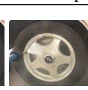
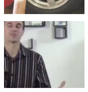
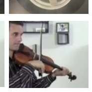
7 Conclusion

We proposed iReason, a multimodal causal knowledge generation framework. iReason outperforms not only language representation models but also state-of-the-art causal models owing to its canonical frame detection module (which enables video as native inputs, rather than randomly-selected frames) and causal rationalization module (which offers interpretability). With ablation experiments, we assessed the role played by multimodal context. To improve iReason, a couple of ideas can be explored. iReason's performance is a function of the object detection module and the language representation model. Using a better object detector and language representation model could improve context availability and event coding respectively for iReason and thus offer enhanced performance. Furthermore, to even remotely match the magnitude of a human-level commonsense knowledge base requires a dataset that covers a multitude of scenarios. While iReasonData is effective, more data is needed to cover additional causal interactions that humans easily infer (sub)consciously.

Table 6: Performance comparison with BERT and GPT-2. Bold numbers indicate best performance.

	R@1	R@5	R@10
Random Guess	2.13	15.25	30.14
BERT	2.13	22.34	39.00
GPT-2	3.55	17.73	34.40
VCC (BERT)	8.87	34.75	63.12
VCC (GPT-2)	7.80	31.56	56.03
iReason (BERT)	**9.21**	**35.87**	**63.51**
iReason (GPT-2)	8.90	35.21	63.23

Table 7: Visually inspecting positive and negative examples from iReason.

Canonical frame detection output (I_1, I_2)	Event captions	Causal prediction	Causality rationalization output
 	e_1: Tire shine is being applied. e_2: The rims are shiny.	Yes	The rims are shiny because tire shine was applied.
 	e_1: A man is talking. e_2: A man plays the violin.	No	No causal rationalization since events are not causal.

References

[1] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE international conference on computer vision*, pages 2542–2550, 2015.

[2] Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2984–2993, 2015.

[3] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

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

[5] 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.

[6] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

[7] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019.

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[9] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018.

[10] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198, 2018.

[11] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[12] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017.

[13] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[14] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[15] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra-and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743, 2013.

[16] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, 2014.

[17] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. Joint reasoning for temporal and causal relations. *arXiv preprint arXiv:1906.04941*, 2019.

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

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

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

[21] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. In *European Conference on Computer Vision*, pages 767–785. Springer, 2020.

[22] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020.

[23] Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, 2014.

[24] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[25] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60, 2019.

[26] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*, 2015.

[27] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, 2020.

[28] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[29] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

[30] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017.

[31] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014.

[32] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019.

[33] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

[34] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. *arXiv preprint arXiv:1911.10496*, 2019.

[35] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. *arXiv preprint arXiv:2004.00390*, 2020.

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. *arXiv preprint arXiv:2002.11949*, 2020.

[37] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. *arXiv preprint arXiv:2003.13942*, 2020.

[38] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 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, 2016.

[39] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014.

[40] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017.

[42] 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.

[43] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[44] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[46] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[47] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[48] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[50] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. T-c3d: Temporal convolutional 3d network for real-time action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[52] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[53] Das Abhishek, Kottur Satwik, Gupta Khushi, Singh Avi, and Yadav Deshraj. Devi parikh and dhruv batra. visual dialog. in. In *CVPR*, pages 326–335, 2017.

[54] 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.

[55] 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.

[56] 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.

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.