

cases	doc_1		doc_2		decision	id
	authors	<ul style="list-style-type: none">Aman ChadhaVinija Jain			DUPLICATES	226
	title	iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability	authors	<ul style="list-style-type: none">Andrew WangAman Chadha		
	publication_date	2021-06-25 02:56:34+00:00	title	iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability		
	source	SupportedSources.ARXIV	publication_date	2021-06-25 00:00:00		
	journal	None	source	SupportedSources.SEMANTIC_SCHOLAR		
	volume		journal	ArXiv		
	doi		volume	abs/2107.10300		
	urls	<ul style="list-style-type: none">http://arxiv.org/pdf/2107.10300v1http://arxiv.org/abs/2107.10300v1http://arxiv.org/pdf/2107.10300v1	doi			
	id	id6701397193203904184	urls	<ul style="list-style-type: none">https://www.semanticscholar.org/paper/4f4e0a52934cb91eded859e09b0ff145ac0828fa		
	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 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. 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.	id	id4562973105241463361		
			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 questionanswering, 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, iReasonâ€™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.	versions	
			versions			