<table>
<tr><th rowspan="2">cases</th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr>
<td>
<table>
<tr><th>authors</th><td>Aman Chadha
Vinija Jain</td></tr>
<tr><th>title</th><td>iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability</td></tr>
<tr><th>publication_date</th><td>2021-06-25 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.INTERNET_ARCHIVE</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>https://web.archive.org/web/20210725210643/https://arxiv.org/pdf/2107.10300v1.pdf</td></tr>
<tr><th>id</th><td>id-246673312439342318</td></tr>
<tr><th>abstract</th><td>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.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>
<table>
<tr><th>authors</th><td>Andrew Wang
Aman Chadha</td></tr>
<tr><th>title</th><td>iReason: Multimodal Commonsense Reasoning using Videos and Natural Language with Interpretability</td></tr>
<tr><th>publication_date</th><td>2021-06-25 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.SEMANTIC_SCHOLAR</td></tr>
<tr><th>journal</th><td>ArXiv</td></tr>
<tr><th>volume</th><td>abs/2107.10300</td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>https://www.semanticscholar.org/paper/4f4e0a52934cb91eded859e09b0ff145ac0828fa</td></tr>
<tr><th>id</th><td>id4562973105241463361</td></tr>
<tr><th>abstract</th><td>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. Finally, we present case-studies attesting to the universal applicability of iReason by incorporating the â€œcausal signalâ€ 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.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>DUPLICATES</td>
<td>227</td>
</tr>
</table>