<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><ul><li>Mayu Otani</li><li>Yuta Nakashima</li><li>Esa Rahtu</li><li>J. HeikkilÃ¤</li><li>N. Yokoya</li></ul></td></tr>
<tr><th>title</th><td>Learning Joint Representations of Videos and Sentences with Web Image Search</td></tr>
<tr><th>publication_date</th><td>2016-08-08 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.SEMANTIC_SCHOLAR</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td>10.1007/978-3-319-46604-0_46</td></tr>
<tr><th>urls</th><td><ul><li>https://www.semanticscholar.org/paper/aa76f655c2ad655080593a191c4b479ab9f18117</li></ul></td></tr>
<tr><th>id</th><td>id-7750524676858768777</td></tr>
<tr><th>abstract</th><td>None</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td><td>
<table>
<tr><th>authors</th><td><ul><li>Mayu Otani</li><li>Yuta Nakashima</li><li>Esa Rahtu</li><li>Janne HeikkilÃ¤</li><li>Naokazu Yokoya</li></ul></td></tr>
<tr><th>title</th><td>Learning Joint Representations of Videos and Sentences with Web Image Search</td></tr>
<tr><th>publication_date</th><td>2016-08-08 09:54:15+00:00</td></tr>
<tr><th>source</th><td>SupportedSources.ARXIV</td></tr>
<tr><th>journal</th><td>None</td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td><ul><li>http://arxiv.org/pdf/1608.02367v1</li><li>http://arxiv.org/abs/1608.02367v1</li><li>http://arxiv.org/pdf/1608.02367v1</li></ul></td></tr>
<tr><th>id</th><td>id2121869158855415783</td></tr>
<tr><th>abstract</th><td>Our objective is video retrieval based on natural language queries. In addition, we consider the analogous problem of retrieving sentences or generating descriptions given an input video. Recent work has addressed the problem by embedding visual and textual inputs into a common space where semantic similarities correlate to distances. We also adopt the embedding approach, and make the following contributions: First, we utilize web image search in sentence embedding process to disambiguate fine-grained visual concepts. Second, we propose embedding models for sentence, image, and video inputs whose parameters are learned simultaneously. Finally, we show how the proposed model can be applied to description generation. Overall, we observe a clear improvement over the state-of-the-art methods in the video and sentence retrieval tasks. In description generation, the performance level is comparable to the current state-of-the-art, although our embeddings were trained for the retrieval tasks.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td><td>DUPLICATES</td><td>373</td></tr>
</table>