Towards Debiasing Frame Length Bias in Text-Video Retrieval via Causal Intervention

Burak Satar 1,2, Hongyuan Zhu 1, Hanwang Zhang 2, Joo Hwee Lim 1,2
1 Institute of Infocomm Research, A*STAR, 2 SCSE, Nanyang Technological University

Introduction
We present a systematic study on a temporal bias due to frame length discrepancy between training and test sets of trimmed video clips of EK-100, YC2 and MSRV-TT.

Problem.
Irrelevant clips are retrieved due to bias coming from the discrepancy. AFLC: Avg Frame Length of a Class.

Bias verification.
We calculate the discrepancy among <verb, noun> pairs (classes) by the average frame length difference between the training and test sets.

Debiasing Method
We propose a novel causal intervention method to remove this spurious correlation.

Structural causal model.
$L \rightarrow Y$: Natural effect on similarity matrix.
$L \rightarrow V$: Bias in videos, which should be cut off.

Model architecture.

Causal Intervention Formula.
\[ E[Y|do(V, Q)] = \sum_{M} P(L = |V, Q|)E[Y|V, Q, L = |] \]
\[ \Delta \sum_{k=1}^{M} (L_k)f_k \{V, Q\} \]

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>nDCG</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epic-Kitchens-100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>39.40</td>
<td>38.91</td>
</tr>
<tr>
<td>Baseline + RmvRand</td>
<td>39.69</td>
<td>38.42</td>
</tr>
<tr>
<td>Baseline + RmvAll</td>
<td>40.06</td>
<td>39.44</td>
</tr>
<tr>
<td>Baseline + Ensemble</td>
<td>40.38</td>
<td>39.76</td>
</tr>
<tr>
<td>Baseline + Ours</td>
<td>42.73</td>
<td>41.67</td>
</tr>
</tbody>
</table>

Qualitative.

Contributions.
1) The first to verify the frame length bias,
2) The first to propose a debiasing method with causal inference in text-video retrieval task,
3) The nDCG metric shows that the bias is mitigated.