Local-Global Video-Text Interactions for Temporal Grounding

Supplementary Material

Jonghwan Mun[1,2] Minsu Cho[1] Bohyung Han[2]
[1]Computer Vision Lab., POSTECH, Korea
[2]Computer Vision Lab., ASRI, Seoul National University, Korea
[1]{jonghwan.mun,mscho}@postech.ac.kr [2]bhhan@snu.ac.kr

This supplementary document first presents the architecture of our model without semantic phrase extraction (*i.e.*, LGI–SQAN) used for in-depth analysis on the local-global video-text interactions. We also present additional qualitative examples of our algorithm.

1. Architectural Details of LGI–SQAN

Compared to our full model (LGI), LGI–SQAN does not explicitly extract semantic phrases from a query as presented in Fig. A; it performs local-global video-text interactions based on the sentence-level feature representing whole semantics of the query.

In our model, the sentence-level feature ($\mathbf{q}$) is copied to match its dimension with the temporal dimension (T) of segment-level features ($\mathbf{S}$). Then, as done in our full model, we perform local-global video-text interactions—1) segment-level modality fusion, 2) local context modeling, and 3) global context modeling—followed by the temporal attention based regression to predict the time interval $[t^s, t^e]$. Note that we adopt a masked non-local block or a residual block for local context modeling, and a non-local block for global context modeling, respectively.

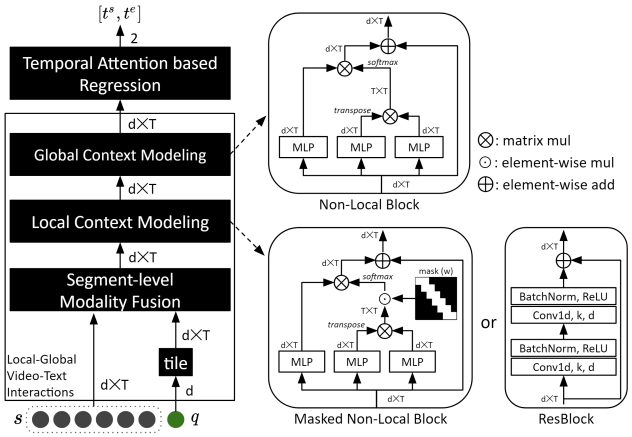

Figure A. Illustration of architecture of LGI–SQAN. In LGI–SQAN, we use sentence-level feature $\mathbf{q}$ to interact with video.

2. Visualization of More Examples

Fig. B and Fig. C illustrate additional qualitative results in the Charades-STA and ActivityNet Captions datasets, respectively; we present two types of attention weights—temporal attention weights $\mathbf{o}$ (T-ATT) and query attention weights $\mathbf{a}$ (Q-ATT)—and predictions (Pred.). T-ATT shows that our algorithm successfully attends to relevant segments to the input query while Q-ATT depicts that our sequential query attention network favorably identifies semantic phrases from the query describing actors, objects, actions, etc. Note that our model often predicts accurate time intervals even from the noisy temporal attention.

Fig. D demonstrates the failure cases of our algorithm. As presented in the first example of Fig. D, our method fails to localize the query on the confusing video, where a man looks like smiling at multiple time intervals. However, note that the temporal attention of our method captures the segments relevant to the query at diverse temporal locations in a video. In addition, as presented in the second example of Fig. D, our model sometimes fails to extract proper semantic phrases; 'wooden' and 'floorboards' are captured at different steps although 'wooden floorboards' is more natural, which results in the inaccurate localization.

Figure B. Qualitative results of our algorithm on the Charades-STA dataset. T-ATT and Q-ATT stand for temporal attention weights and query attention weights, respectively.

Figure C. Qualitative results of our algorithm on the ActivityNet Captions dataset. T-ATT and Q-ATT stand for temporal attention weights and query attention weights, respectively.

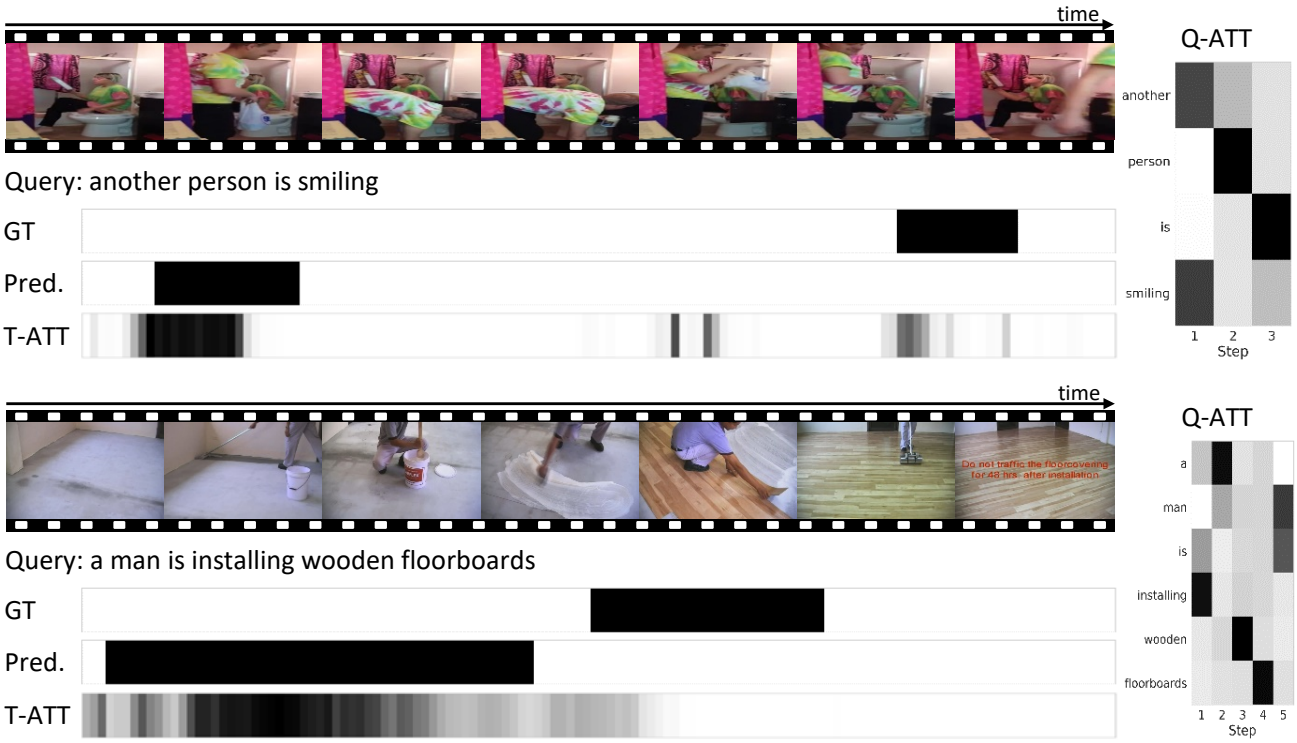

Figure D. Failure case of our algorithm. Examples in the first and second row are obtained from the Charades-STA and Activity Captions datasets, respectively.