Dense Captioning Events in Videos

VCMR

December 10, 2024

Problems

- Current methods of captioning can only focus on one event, ignoring other events → limitation in detail
- If we want to caption multiple events → multiple frames → there can be events overlapping each other in terms of times → PROBLEM.

Problems

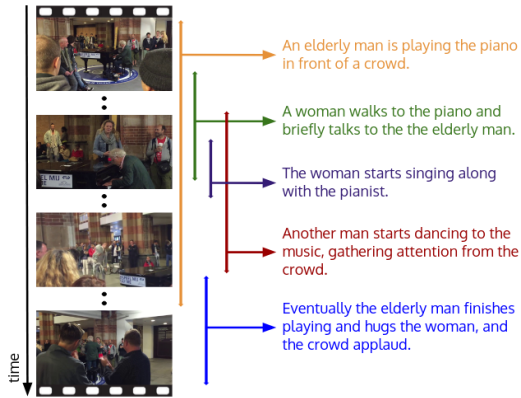

Figure: detecting multiple events that occur in a video and describing each event using natural language

Assumption

- Video contains multiple events overlapping + relating to each other
- The research does not consider the situation that there's gonna be transition of scene forward and backward (Ex: there is scene A, then there is an interview scene for a seconds, then changing back to scene A).
- They only assume that the video would consistently present one topic.

Challenges

- Step 1: Detect all events in the video
- Step 2: Utilize context from the past, current, and future to find the relation (e.g cause-effect) between them.

$\rightarrow$ Rely on components of other works in action propasal and social human tracking to solve.

Overview their approach

- Input is a series of frames $\text{frame}_t, t \in 0, \ldots, T-1$
- First define the vocab set V, using for generating sentence.
- In on video, there are different N events, each event $i \in N$ is describe with one sentence s_i. And this event-sentence pair has the start time t^{start} and end time t^{end}.

Overview their approach

- There are N events, with event $i \in N$, we have s_i for text, and also we have to encode the series of frames of the i-th event. They use Propasal module to extract vector $h_i \in \mathbb{R}^d$.

Input frames $\rightarrow$ Module $\rightarrow$ output proposals P

$$P = \{t_i^{\text{start}}, t_i^{\text{end}}, \text{score}_i, h_i\} \tag{1}$$

- The score_i is used to filter which context that language model should use to generate sentence description s_i.