Streamlined Dense Video Captioning

Supplementary Material

Jonghwan Mun[1,5] Linjie Yang[2] Zhou Ren[3] Ning Xu[4] Bohyung Han[5]
[1]POSTECH [2]ByteDance AI Lab [3]Wormpex AI Research [4]Amazon Go [5]Seoul National University

A. Details of Event RNN

As described in Section 3.4 of the main paper, the event RNN is in charge of generating a description given an event and a context information, and returning features for the generated captions. This section discusses the caption generation process of the event RNN in detail.

Following [A1], we provide two kinds of event information, $\text{C3D}(e)$ and $\text{Vis}(e)$, to the event RNN, where $\text{C3D}(e)$ is a set of segment-level feature descriptors in an interval of an event e, and $\text{Vis}(e)$ is a visual representation obtained from the event proposal network, SST. We first set an initial hidden state of the event RNN to the context vector of episode r given by the episode RNN. Then, at each time step of the event RNN, we perform Temporal Dynamic Attention (TDA) to obtain an attentive segment-level feature from $\text{C3D}(e)$, followed by Context Gating (CG) to adaptively model relative contributions of the attentive segment-level feature and the visual feature and return a gated event feature. Based on the gated event feature, the event RNN generates a word, and returns the hidden state as the caption feature g when generating the END token .

The whole caption generation process in the event RNN is summarized by the following sequence of operations:

$$
\begin{aligned}
h^e_0 &= r, && (1)\\
x_t &= W_{\text{wemb}} w_t, && (2)\\
z_t &= \text{TDA}(\text{C3D}(e), \text{Vis}(e), h^e_{t-1}), && (3)\\
o_t &= \text{CG}(z_t, \text{Vis}(e), x_t, h^e_{t-1}), && (4)\\
h^e_t &= \text{LSTM}_e(o_t, x_t, h^e_{t-1}), && (5)\\
p_t &= \text{Softmax}(W_p h^e_t), && (6)
\end{aligned}
$$

where W_{wemb} and W_p are learnable parameters, h^e means a hidden state of LSTM_e in the event RNN, and w_t, x_t, z_t, o_t and p_t denote an input word, a word embedding vector, an attentive segment-level feature vector, a gated event feature vector and a probability distribution over vocabulary at time t, respectively. At time step t, given an event with S segments, TDA computes the attentive vector z_t by

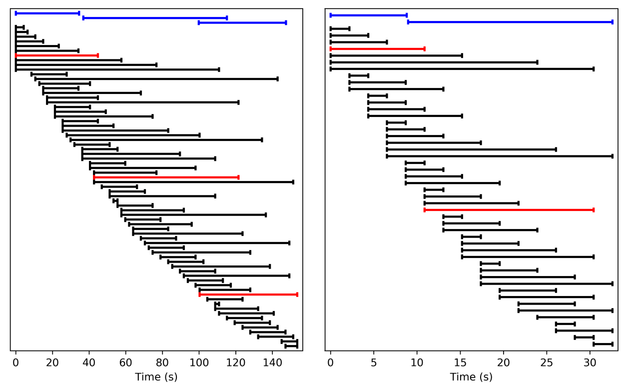

Figure A. Examples of the selected event proposals (**red**) out of the candidates (**black**) in the proposed event sequence generation network and the ground-truth events (**blue**).

$$
\begin{aligned}
\alpha^t_s &= W_\alpha \tanh(W_\text{c}\text{C3D}(e_s) + W_\text{v}\text{Vis}(e) + W_h h^e_{t-1}), && (7)\\
a^t_s &= \frac{\exp(\alpha^t_s)}{\sum_{s=1}^{S} \exp(\alpha^t_s)}, && (8)\\
z_t &= \sum_{s=1}^{S} a^t_s \text{C3D}(e_s), && (9)
\end{aligned}
$$

where W_α, W_c, W_v and w_h are learnable parameters, and e_s indicates the s^{th} segment in event e. Once obtaining the attentive segment-level feature, CG computes the gating vector k_t and the gated event vector o_t as follows:

$$
\begin{aligned}
\bar{z}_t &= \tanh(W_\text{z} z_t), && (10)\\
\bar{v} &= \tanh(W_{\bar{\text{v}}}\text{Vis}(e)), && (11)\\
k_t &= \sigma(W_k[\bar{z}; \bar{v}; x_t; h^e_{t-1}]), && (12)\\
o_t &= [(1 - k_t) \odot \bar{z}_t; k_t \odot \bar{v}], && (13)
\end{aligned}
$$

where W_z, $W_{\bar{v}}$ and W_k are learnable parameters, σ is a sigmoid function, $[\cdot;\cdot]$ denotes vector concatenation and $\odot$ means element-wise multiplication.

B. Visualization of Event Selection

Fig. A illustrates the event selection results. Our proposed event sequence generation network successfully identifies the event proposals out of the candidates, which highly overlap with the ground-truths.

References

[A1] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018.