# 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, C3D($e$) and Vis($e$), to the event RNN, where C3D($e$) is a set of segment-level feature descriptors in an interval of an event $e$, and 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 the 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:

$$h_0^e = r, \tag{1}$$
$$x_t = W_{\text{wemb}} w_t, \tag{2}$$
$$z_t = \text{TDA}(\text{C3D}(e), \text{Vis}(e), h_{t-1}^e), \tag{3}$$
$$o_t = \text{CG}(z_t, \text{Vis}(e), x_t, h_{t-1}^e), \tag{4}$$
$$h_t^e = \text{LSTM}_e(o_t, x_t, h_{t-1}^e), \tag{5}$$
$$p_t = \text{Softmax}(W_p h_t^e), \tag{6}$$

where $W_{\text{wemb}}$ and $W_p$ are learnable parameters, $h^e$ means a hidden state of $\text{LSTM}_e$ 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

$$\alpha_s^t = W_\alpha \tanh(W_c \text{C3D}(e_s) + W_v \text{Vis}(e) + W_h h_{t-1}^e), \tag{7}$$
$$a_s^t = \frac{\exp(\alpha_s^t)}{\sum_{s=1}^{S} \exp(\alpha_s^t)}, \tag{8}$$
$$z_t = \sum_{s=1}^{S} a_s^t \text{C3D}(e_s), \tag{9}$$

where $W_\alpha$, $W_c$, $W_v$ 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:

$$\bar{z}_t = \tanh(W_z z_t), \tag{10}$$
$$\bar{v} = \tanh(W_{\bar{v}} \text{Vis}(e)), \tag{11}$$
$$k_t = \sigma(W_k[\bar{z}; \bar{v}; x_t; h_{t-1}^e]), \tag{12}$$
$$o_t = [(1 - k_t) \odot \bar{z}_t; k_t \odot \bar{v}], \tag{13}$$

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

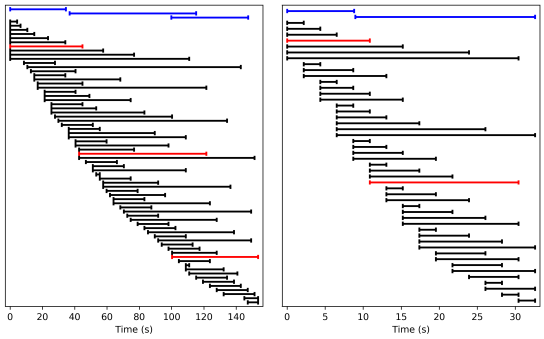


Figure A. Examples of the selected event proposals (**red**) out of the candidates (**black**) in the proposed event sequence generation network and the ground-truth events (**blue**).

## 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.