A UNROLLING CLASSIFIER

In the following, we use $\texttt{FC}(v)$ to denote a linear (fully-connected) transformation on vector v. The parameters of different $\texttt{FC}$'s are distinguished by their subscripts.

We took global average pooling followed by a flatten operator on spatial dimension to obtain a 1D vector of y_t (see eq. 17) as $\overline{y_t}$. The unrolling LSTM is then performed on the $\overline{y_t}$ to unfold the predictions. The hidden and memory states of the unrolling LSTM are initialized by the $\overline{y_t}$ with additional transformations:

$$\xi = \texttt{ReLU}(\texttt{FC}_{\xi}(\overline{y_t})) \tag{A.1}$$
$$feat = \texttt{ReLU}(\texttt{FC}_{feat}(\xi)) \tag{A.2}$$
$$c_0^{url} = \texttt{FC}_c(feat) \tag{A.3}$$
$$h_0^{url} = \texttt{FC}_h(feat) \tag{A.4}$$

the states are unrolled T steps, where $T = \tau_a$ (anticipate time) $\times$ fps, with the consistent input $\overline{y_t}$,

$$(h_T^{url}, c_T^{url}) = \texttt{LSTMCell}(\overline{y_t}, (h_{T-1}^{url}, c_{T-1}^{url})) \tag{A.5}$$

We leverage the final unrolling output h_T^{url} for further logits calculation:

$$\text{logit}_a = \texttt{FC}_a(h_T^{url}) \tag{A.6}$$
$$\text{logit}_v = \texttt{FC}_v(h_T^{url}) + \texttt{FC}_{a2v}(\text{logit}_a) \tag{A.7}$$
$$\text{logit}_n = \texttt{FC}_n((h_T^{url}) + \texttt{FC}_{a2n}(\text{logit}_a)) \tag{A.8}$$

where the logit_a also serves as biases to logit_v and logit_n for explicitly building the relationship of action to the verb and noun.

The cross-entropy loss is used to compare between logits and ground-truth labels. The overall loss function is the summation over the individual loss of each action, verb, and noun, for every anticipation time.