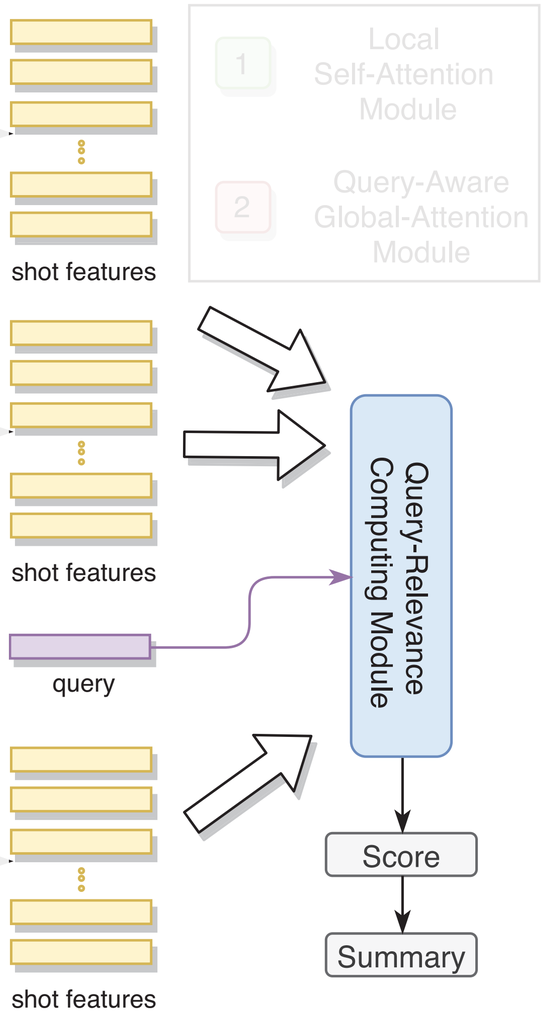
Proposed Method

Query-Relevance Computing Module

- Compute the similarity score between video shot and user's query.
- The module takes the shot-level features $(\hat{v}_1^f, \hat{v}_2^f, \ldots, \hat{v}_n^f)$ generate by feature encoding network and concepts as inputs.
- Given a specific concept c, using pretrained language model to obtain its embedding feature f_c.
- Calculate f_c and shot-level features $\hat{v}_i^f$ of i-th shot distance-based similarity: $d_i = W_f \hat{v}_i^f \odot W_c f_c$
  - W_f, W_c: parameter matrices project the visual & textual features into same vector space



Proposed Method

Query-Relevance Computing Module

- Then let the output pass a MLP and get the concept-relevant score between i-th segment and concept c.
- The average of two concept-relevant score is taken as the query-relevant score $s = \{s_1, s_2, \ldots, s_n\}$
- Given the ground truth annotations $\hat{s} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n\} \in [0,1]$
- Loss: $L_{summ} = \frac{1}{T}\sum_{t=1}^{T} \hat{s}_t \log s_t + (1-\hat{s}_t)\log(1-s_t)$
- By minimizing the loss, module can focus on the most concept-related video shots.