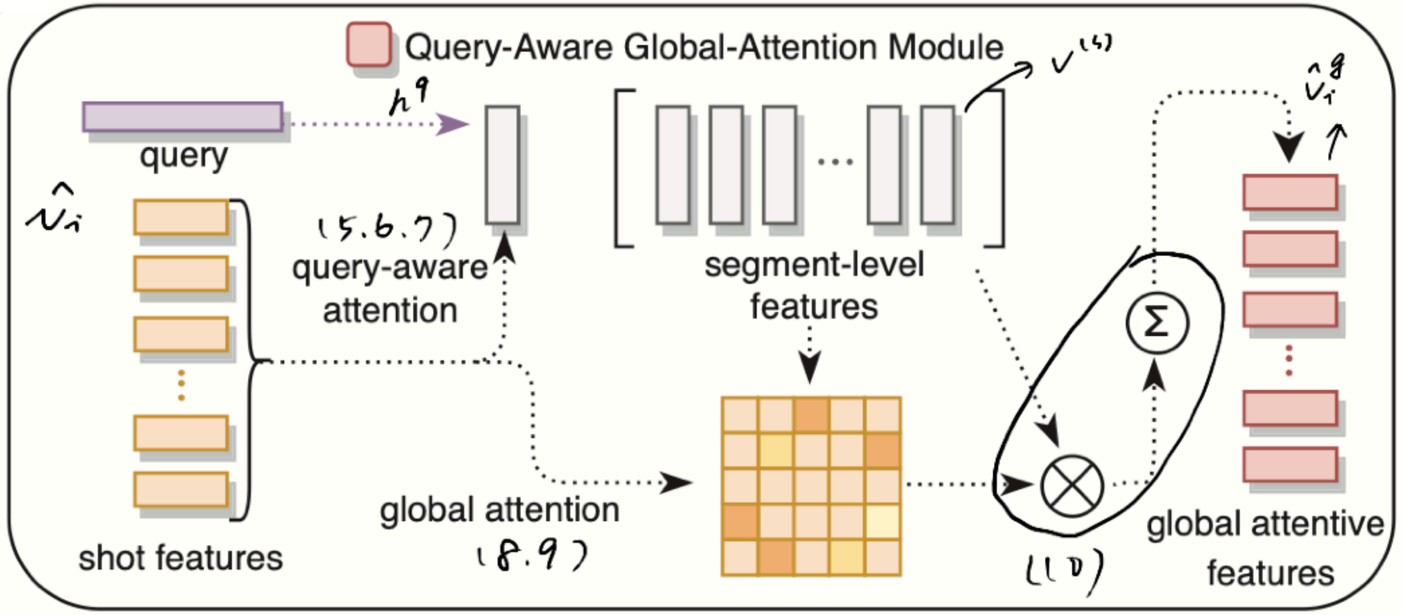
Proposed Method

Query global-attention module

- Compute the query-aware global -attentive representation for each shot.
- Given visual feature $\hat{v}_i$ & all segment-level visual representation $(v_1^{(s)}, v_2^{(s)}, \ldots, v_m^{(s)})$
  - m: number of video segments



(8) $e_j^g = v^T \tanh(W_1^g \hat{v}_i + W_2^g v_j^{(s)} + b)$

(9) $r_j^g = \dfrac{\exp(e_j^g)}{\sum_{k=0}^{m} \exp(e_k^g)}$

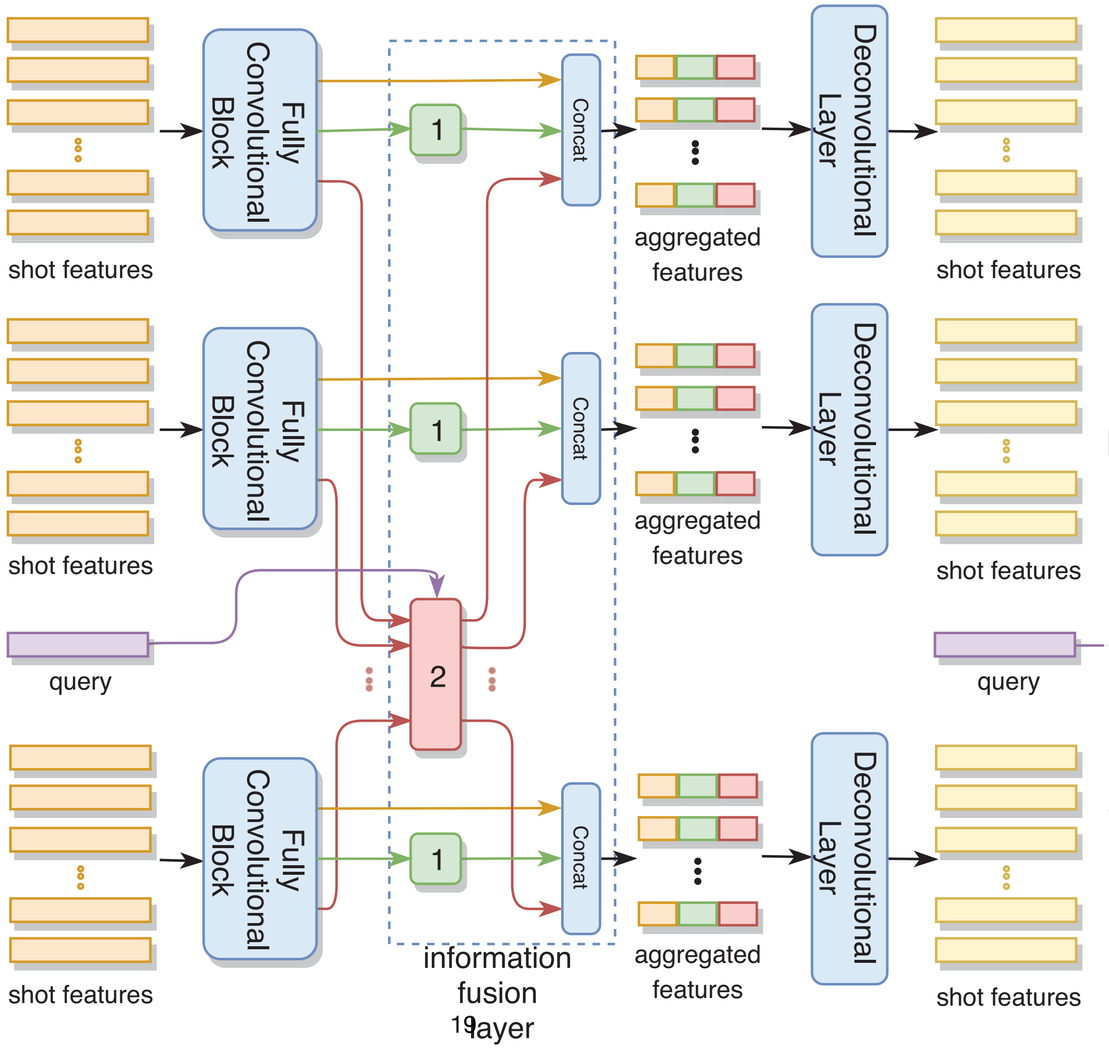
(10) Global-attentive representation for i-shot:

$$\hat{v}_j^g = \sum_{j=0}^{m} r_j^g v_j^s$$

Proposed Method

Deconvolutional Layer

- Use fully convolutional block to decrease the number on temporal axis of video features.



- Propose several 1D deconvolutional layers to recover the original number of video shots.