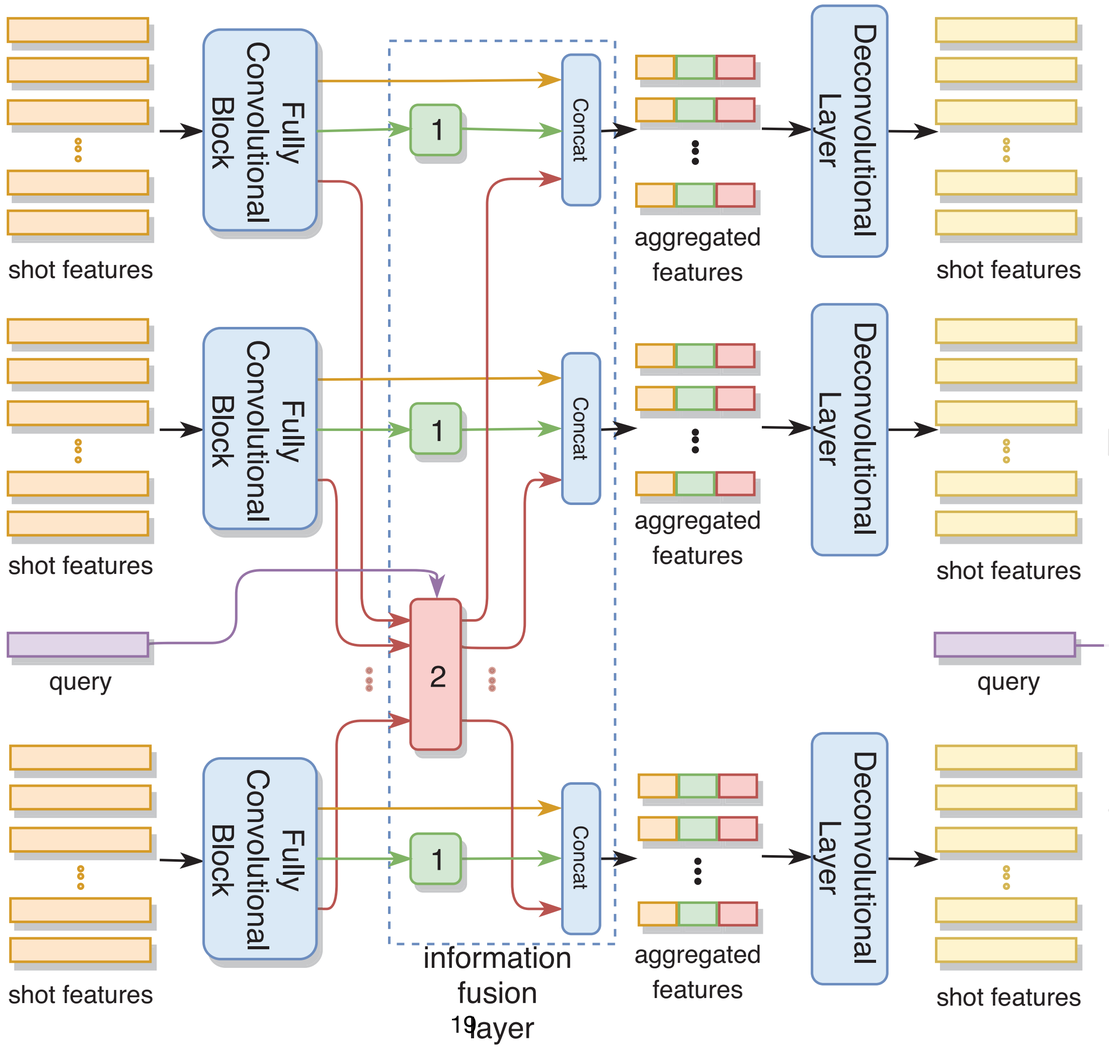
Proposed Method

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.

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