Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer (Supplementary Material)

Hao Jiang
Peking University
jianghao@stu.pku.edu.cn

Yadong Mu *
Peking University
myd@pku.edu.cn

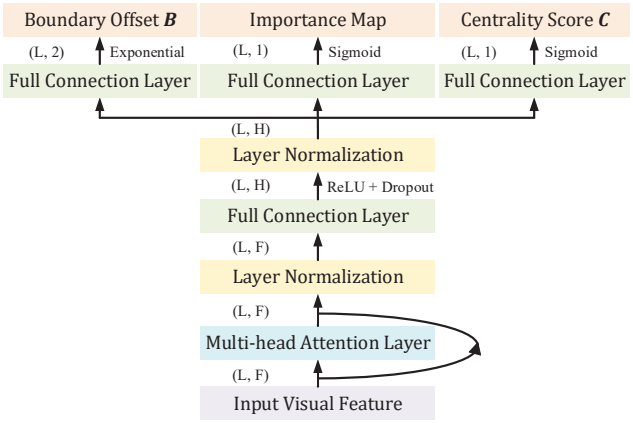


Figure 1. Illustration of the neural network used in the video summarization module, where L represents the sequence length of input visual features, F is the input feature size, and H denotes the hidden size.

A. Neural Designs of SM

Figure 1 shows the neural architecture we used in the video summarization module. It is mainly composed of the multi-head attention layer and the full connection layer, and the layer normalization is employed to help the training of the neural network. The importance map output by the network is finally used for video summarization.

B. Neural Designs of LM

Figure 2 illustrates the neural network architecture of the video moment localization module. It is mainly composed of the linear projection layer, the convolutional layer, and the multi-head attention layer. The layer normalization is also used to help the training of the neural network.

The Context-Query Attention layer [3, 4] is used to capture the cross-modal interactions between visual and textual features. Given input visual features $\boldsymbol{V} \in \mathbb{R}^{n \times d_v}$ and textural features $\boldsymbol{Q} \in \mathbb{R}^{m \times d_q}$ of this layer, it first uses the matrix

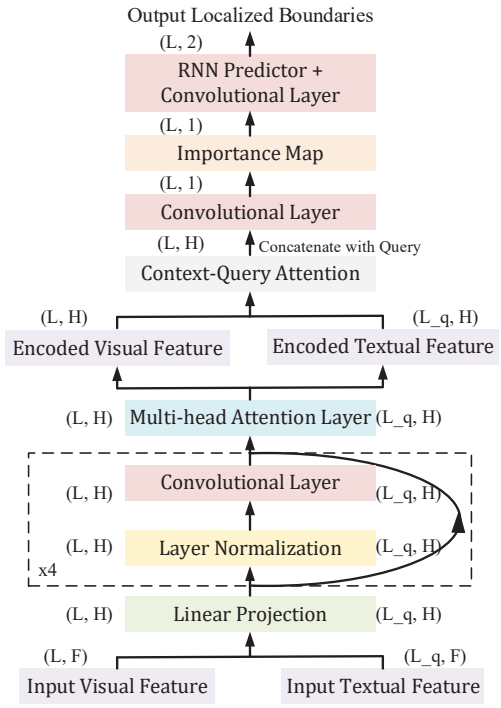


Figure 2. Illustration of the neural network used in the video moment localization module, where L represents the sequence length of the input visual feature, L_q represents the sequence length of the input textual feature, F is the input feature size, and H denotes the hidden size.

product to calculate the similarity matrix $\boldsymbol{C} \in \mathbb{R}^{n \times m}$ between $\boldsymbol{V}$ and $\boldsymbol{Q}$, and then the textual-to-visual attention $\boldsymbol{T}$ is calculated as follows:

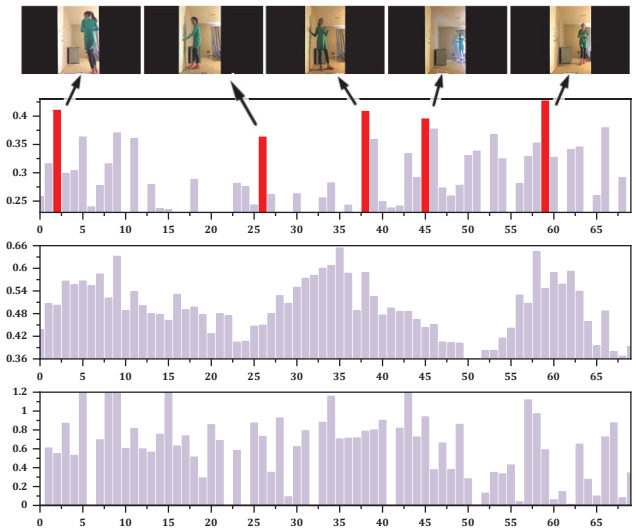
$$\boldsymbol{T} = \mathrm{RowNorm}(\boldsymbol{C}) \cdot \mathrm{ColNorm}(\boldsymbol{C})^T \cdot \boldsymbol{V} \quad (1)$$

where $\mathrm{RowNorm}(\boldsymbol{C})$ and $\mathrm{ColNorm}(\boldsymbol{C})$ represent the row normalization and column normalization of matrix $\boldsymbol{C}$, respectively. The visual-to-textual attention $\overline{\boldsymbol{T}}$ is calculated by:

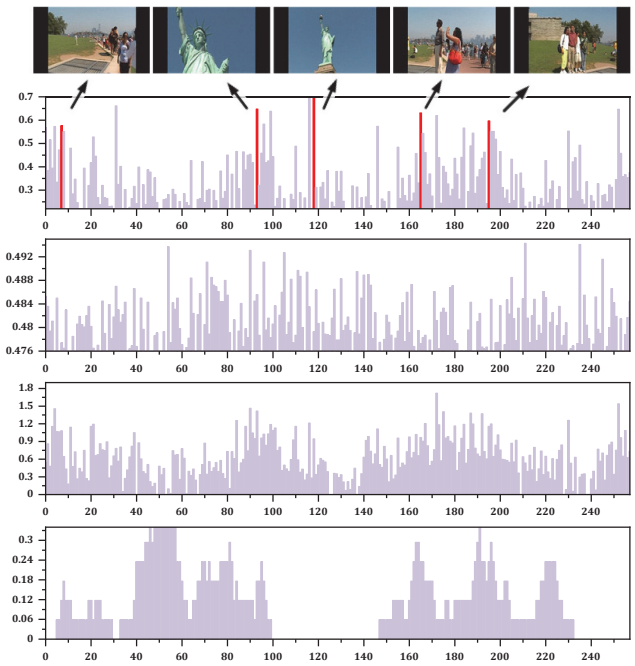
$$\overline{\boldsymbol{T}} = \mathrm{RowNorm}(\boldsymbol{C}) \cdot \boldsymbol{Q} \quad (2)$$

Context-Query Attention layer outputs the feature vector

*Corresponding author.



(a) Qualitative results on a randomly-sampled video from the moment localization dataset Charades (thus no ground truth for video summarization).



(b) Qualitative results on the video drawn from the SumMe dataset.

Figure 3. Illustration of the qualitative results on both the video summarization and video moment localization datasets. In (a), the first row represents the selected video frames based on the generated video summaries, the second row denotes the importance map generated by the video summarization module, the third row is the importance map generated by the moment localization module, and the fourth row is the importance map after propagation. In (b), the first four rows have the same meaning as in (a), and the last row represents the ground-truth summary annotated by users.

$\boldsymbol{S}$ that adopts the attention weights to encode both visual and textual features, which is given by the following formula:

$$\boldsymbol{S} = \text{FeedForwardNet}(\boldsymbol{V} \oplus \overline{\boldsymbol{T}} \oplus \boldsymbol{V} \odot \boldsymbol{T} \oplus \boldsymbol{V} \odot \overline{\boldsymbol{T}}) \quad (3)$$

where $\oplus$ represents the concatenation operation, and $\odot$ denotes the Hadamard production.

The importance map generated by the LM neural network is used for video moment localization and subsequent collaborative teaching.

C. Qualitative Results

In addition to the quantitative performance comparison experiments presented in the main text, we have also conducted a few qualitative experiments for the proposed method. As shown in Figure 3, we visualize the importance maps generated by SM and LM on both the video summarization and video moment localization datasets, together with the results after importance propagation. Figure 3 chooses the video 406LH in the Charades [1] dataset and the 20th video in the SumMe [2] dataset in (a) and (b) respectively. The results of qualitative experiments provide more intuitive understanding of the effectiveness of the proposed method.

References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 2

[2] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014. 2

[3] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *ICLR*, 2018. 1

[4] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *TPAMI*, 2021. 1