Human-centric Spatio-Temporal Video Grounding via the Combination of Mutual Matching Network and TubeDETR

Fan Yu\textsuperscript{1}, Zhixiang Zhao\textsuperscript{1}, Yuchen Wang\textsuperscript{1}, Yi Xu\textsuperscript{2}, Tongwei Ren\textsuperscript{1,*}, Gangshan Wu\textsuperscript{1}
\textsuperscript{1}State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
\textsuperscript{2}School of Computer Science and Technology, Soochow University, Suzhou, China
\{yf.zhaozx,181250147\}@smail.nju.edu.cn, yxu9910@gmail.com, \{rentw,gswu\}@nju.edu.cn

ABSTRACT

In this technical report, we represent our solution for the Human-centric Spatio-Temporal Video Grounding (HC-STVG) task of the 4th Person in Context (PIC) workshop and challenge. Our solution is built on the basis of TubeDETR and Mutual Matching Network (MMN). Specifically, TubeDETR exploits a video-text encoder and a space-time decoder to predict the starting time, the ending time and the tube of the target person. MMN detects persons in images, links them as tubes, extracts features of person tubes and the text description, and predicts the similarities between them to choose the most likely person tube as the grounding result. Our solution finally finetunes the results by combining the spatio localization of MMN and the temporal localization of TubeDETR. In the HC-STVG track of the 4th PIC challenge, our solution achieves the third place.

1 INTRODUCTION

Human-centric Spatio-Temporal Video Grounding (HC-STVG) task \cite{10} is one of the three tracks in the 4th Person in Context (PIC) workshop and challenge. HC-STVG is a further exploration of visual grounding, which aims to locate the object of a given query with its bounding box \cite{3, 14}. Video grounding requires to localize the starting and ending time of the given video according to a query \cite{2, 15}. Given a sentence depicting an object, spatio-temporal video grounding (STVG) \cite{11, 16} extracts the spatio-temporal tube of the object. The query of an input video in HC-STVG is a sentence describing a person in terms of the appearance, the action and the interaction with the environment. Similar to STVG, HC-STVG needs to localize the target person, \textit{i.e.}, the starting and ending time with the bounding boxes of the target person during the video clip.

The first proposed method for HC-STVG is STGVT \cite{10}, which detects region proposals in frames, links the bounding boxes in consecutive frames to form spatio-temporal tube proposals and then uses a visual Transformer combining features extracted from videos and textual descriptions to match and trim the tubes with the given textual description. Su \textit{et al.} \cite{7} propose a unified STVG framework named STVGBert, which also exploits the Transformer model to encode visual and textual features but does not require to generate tube proposals in the beginning. In the 2021 PIC challenge, three more solutions were proposed for HC-STVG. Tan \textit{et al.} \cite{8} propose to first localize the temporal segment with the Augmented 2D-TAN model and then predict the spatial location of the target person in each frame. Yu \textit{et al.} \cite{1} propose to extract human information from the query text, \textit{i.e.}, gender, clothing color and clothing type, generate human tubes from the corresponding video, and finally exploit a Transformer to encode visual and textual features to perform tube-description matching and tube trimming. Wang \textit{et al.} \cite{12} introduces metric learning \cite{17} on the basis of visual features extraction from linked human tubes and textual features extraction from the given query. Moreover, TubeDETR \cite{13} is proposed as a unified framework for HC-STVG, which uses video-text encoders to combine visual and textual features and predicts starting time, ending time and the spatio-temporal tube with a space-time decoder.

Our solution is built on the basis of TubeDETR \cite{13} and MMN \cite{12}. We observe that TubeDETR achieves desired results of spatio localization and MMN has better performance of temporal localization. Thus, we keep the temporal results of MMN and replace its spatio results with TubeDETR’s.

2 DATASET

The first dataset for the HC-STVG task is \textit{HC-STVG}, where each video is of 20 seconds and is labeled with a sentence describing a person and the corresponding spatio-temporal localization. The spatio-temporal localization in \textit{HC-STVG} is represented by the staring frame, the ending frame and the bounding boxes during the segment. \textit{HC-STVG} dataset has been updated to the third version. Compared with \textit{HC-STVG} 1.0, data in \textit{HC-STVG} 2.0 are expanded and the labels are cleaned. In \textit{HC-STVG} 2.1, noisy data are further manually re-annotated and some videos are moved from the test set to the validation set. The difference among the three versions of data composition is shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Number of video clips in different versions of \textit{HC-STVG}.</th>
</tr>
</thead>
<tbody>
<tr>
<td>version</td>
</tr>
<tr>
<td>training set</td>
</tr>
<tr>
<td>validation set</td>
</tr>
<tr>
<td>test set</td>
</tr>
</tbody>
</table>

\*Corresponding author.
As illustrated in Figure 1, our solution combines the temporal valid moment candidates, N visual feature and textual feature are both used with metric learning prospective. The framework of MMN contains two stages: the first stage aims to extract features and the second stage matches textual description with tube candidates and trims the target tube. MMN detects humans in frames with Faster R-CNN and links the human bounding boxes following ACT to generate tube candidates. Each candidate tube is split into 16 clips and each tube clip is considered as a unit. The visual feature of each unit is generated by CSN and a 2D moment map is constructed for each tube candidate to predict the IoU score of a candidate sub-tube for the groundtruth tube with the maximum value of the multiplied IoU score. During training, the total loss is the summation of iou loss, video loss and sentence loss:

\[ L_{TD} = L_{TD}^{iou} + L_{TD}^{box} + L_{TD}^{sen} + \frac{1}{|B|} \sum_{b \in B} L_{TD}(b, \hat{b}), \]

where \( L_{TD}^{iou} \) is the IoU loss, \( L_{TD}^{box} \) is the bounding box loss, \( L_{TD}^{sen} \) is the sentence loss, and \( L_{TD} \) is the total loss for a clip. During training, the total loss is the summation of bounding box loss, iou loss, Kullback-Leibler divergence loss and guided attention loss:

\[ L_{TD}^{iou} = \alpha L_{TD}^{iou} + \beta L_{TD}^{iou} + \gamma L_{TD}^{iou} + \theta L_{TD}^{iou}, \]

\[ L_{TD}^{box} = \sum_{b \in B} (1 - \frac{I}{U} + \frac{A^c - U}{A^c}) \]

\[ L_{TD}^{sen} = \sum_{i = 1}^{n} (1 - \delta_{r < s_i \leq r^e}) \log(1 - a_i) \]

The input video is segmented into 20 clips, and the duration of each clip is 1 second. Visual features extracted from video clips are combined with the textual feature extracted from the corresponding query in video-text encoders. A space-time decoder then takes the time-sequentially combined features as input and predicts the probability of starting and ending along with the tube for each clip. During training, the total loss is the summation of bounding box loss, iou loss, Kullback-Leibler divergence loss and guided attention loss:

3 SOLUTION

As illustrated in Figure 1, our solution combines the temporal localization result of MMN and the spatio localization result of TubeDETR.

MMN. MMN performs cross-modal mutual matching in the metric-learning prospective. The framework of MMN contains two stages: the first stage aims to extract features and the second stage matches textual description with tube candidates and trims the target tube. MMN detects humans in frames with Faster R-CNN and links the human bounding boxes following ACT to generate tube candidates. Each candidate tube is split into 16 clips and each tube clip is considered as a unit. The visual feature of each unit is generated by CSN and a 2D moment map is constructed for each tube candidate to predict the IoU score of a candidate sub-tube for the groundtruth tube with the maximum value of the multiplied IoU score. During training, the total loss is the summation of iou loss, video loss and sentence loss:

\[ L_{TD} = L_{TD}^{iou} + \lambda L_{TD}^{td} + L_{TD}^{sen}, \]

\[ L_{TD}^{iou} = -\sum_{i = 1}^{N} \log(p_{v_i}^{iou} + (1 - y_{v_i}) \log(1 - p_{v_i}^{iou})) \]

\[ L_{TD}^{td} = -\sum_{i = 1}^{N} \log(p_{i}^{td} \mid s_i), \]

\[ L_{TD}^{sen} = -\sum_{i = 1}^{N} \log(p_{i}^{sen} \mid v_i), \]

where \( \lambda \) is the weight parameter, \( C \) is the total number of valid moment candidates, \( N \) is the total number of moment-sentence pairs for training, \( i_u \) and \( i_e \) denote the instance-level classes of the \( i_u \)th moment and the \( i_e \)th sentence respectively, \( p_{v_i}^{iou} \) and \( y_{v_i} \) denote the predicted and groundtruth iou of the \( i_u \)th moment respectively, and \( v_i \) and \( s_i \) refer to the \( i_u \)th moment and the \( i_e \)th sentence respectively.

TubeDETR. Different from MMN, TubeDETR is a unified framework with the encoder-decoder architecture. The input video is segmented into 20 clips, and the duration of each clip is 1 second. Visual features extracted from video clips are combined with the textual feature extracted from the corresponding query in video-text encoders. A space-time decoder then takes the time-sequentially combined features as input and predicts the probability of starting and ending along with the tube for each clip. During training, the total loss is the summation of bounding box loss, iou loss, Kullback-Leibler divergence loss and guided attention loss:

\[ L_{TD}^{iou} = \alpha L_{TD}^{iou} + \beta L_{TD}^{iou} + \gamma L_{TD}^{iou} + \theta L_{TD}^{iou}, \]

\[ L_{TD}^{box} = \sum_{b \in B} (1 - \frac{I}{U} + \frac{A_c - U}{A_c}) \]

\[ L_{TD}^{sen} = \sum_{i = 1}^{n} (1 - \delta_{r < s_i \leq r^e}) \log(1 - a_i) \]

where \( \alpha, \beta, \gamma, \theta \) are weight parameters, \( B \) is the set of the groundtruth bounding boxes, \( b \) is the predicted bounding box associated with a groundtruth bounding box element \( b \), \( L1 \) represents L1 loss, \( I \) and \( U \) is the intersection and union area of the predicted bounding box and the groundtruth bounding box respectively, \( A^c \) represents the area of the smallest enclosing box, \( D_{KL} \) is the Kullback-Leibler divergence, \( \tau^s \) and \( \tau^e \) refer to the probabilities of the start and end of the output video tube respectively, \( \tau^s \) and \( \tau^e \) refer to the target start and target end distribution respectively, \( \delta \) is the Kronecker delta and \( a_i \) is the \( i_u \)th column in the attention matrix \( A \). In our solution, we use the MDETR [5] as the pretrained model, which assists the TubeDETR to achieve the best performance.
**Finetuning.** The bounding box results of TubeDETR is directly predicted by the space-time decoder together with the starting time and ending time and the network for jointly spatio-temporal prediction is trained on the HC-STVG dataset. However, the person tubes and the corresponding features in MMN are generated with pre-trained models. Thus, the spatio localization of TubeDETR is more accurate than that of MMN. The temporal location results of MMN are predicted with a starting-ending moment 2D matrix while the starting time and ending time are predicted in TubeDETR independently. Thus, MMN can achieve better performance in temporal localization. For these reasons, we keep the temporal results of MMN and replace the spatio results with TubeDETR’s.

4 EXPERIMENTS

4.1 Metrics

To evaluate the performance of solutions for HC-STVG, three types of metrics are used.

- **tIoU.** tIoU is used to evaluate the performance of temporal localization:
  \[
  tIoU = \frac{|S_i|}{|S_u|},
  \]
  where \(S_i\) is the set of frames in the intersection of predicted and ground truth tube, \(S_u\) is the set of frames in the union of predicted and ground truth tube.

- **vIoU.** vIoU evaluates both temporal localization and spatio trajectory:
  \[
  vIoU = \frac{1}{|S_u|} \sum_{t \in S_i} IoU(Box^t, Box'^t),
  \]
  where \(Box^t\) and \(Box'^t\) are the predicted bounding box and ground truth bounding box of frame \(t\).

- **vIoU@R.** vIoU@$R$ represents the percentage of samples whose vIoU is larger than \(R\), and vIoU@$0.3$ and vIoU@$0.5$ are used in this report.

### Table 2: Comparison results on the HC-STVG 2.1 validate set. It is worth noting that the final result of ours in leaderboard of HC-STVG 2022 is the result on the test set of MMN (corresponding to the first line).

<table>
<thead>
<tr>
<th>Methods</th>
<th>vIoU@0.3</th>
<th>tIoU@0.3</th>
<th>vIoU@0.5</th>
<th>tIoU@0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMN</td>
<td>0.280</td>
<td>0.503</td>
<td>0.449</td>
<td>0.227</td>
</tr>
<tr>
<td>TubeDETR</td>
<td>0.285</td>
<td>0.445</td>
<td>0.426</td>
<td>0.192</td>
</tr>
<tr>
<td>TubeDETR+MMN</td>
<td>0.255</td>
<td>0.445</td>
<td>0.375</td>
<td>0.154</td>
</tr>
<tr>
<td>MMN+TubeDETR</td>
<td>0.313</td>
<td>0.503</td>
<td>0.501</td>
<td>0.252</td>
</tr>
</tbody>
</table>

4.2 Quantitative Analysis

We compare the results of MMN and Tube along with the finetuned results in Table 2. Compared with MMN, TubeDETR achieves better performance in vIoU but has worse performance in tIoU. “TubeDETR+MMN” represents the method that uses the temporal localization of TubeDETR and the spatio localization of MMN, all metrics of which are worse than those of both MMN and TubeDETR. However, “MMN+TubeDETR”, which represents the method that uses the temporal result of MMN and replaces its spatio result with TubeDETR’s, has the best performance in all metrics. These experimental data validate the effectiveness of our solution, which combines the temporal localization of MMN and the spatio localization of TubeDETR.

4.3 Qualitative Analysis

Two visualization examples (Figure 2 and Figure 3) show the performance difference between the solutions in Table 2. As shown in Figure 2, MMN has accurate temporal localization but detects the wrong person, TubeDETR has accurate spatio localization but its prediction of temporal localization is undesired. “TubeDETR+MMN” still detects the wrong person since it keeps the spatio result of MMN, while “MMN+TubeDETR” can detect the right person on the

![Figure 2: An example result of MMN, TubeDETR and MMN+TubeDETR. Spatio and temporal annotations in groundtruth are in yellow, and those in prediction are in red.](image)
“TubeDETR+MMN” uses the temporal result of TubeDETR as the target time, but TubeDETR is more accurate in bounding box detection than MMN. Since “TubeDETR+MMN” uses the temporal result of TubeDETR and the spatio result of MMN, spatio localization is missing in almost half of its target time. “MMN+TubeDETR” keeps the accurate temporal localization of MMN and also uses the better spatio localization of TubeDETR, thereby achieving can achieve good performance in both spatio and temporal evaluation. These examples shows that combining the temporal prediction of MMN and the spatio prediction of TubeDETR is more effective.

5 CONCLUSIONS

In this report, we represented our solution for the HC-STVG track in PIC 2022 challenge. Our solution is built on the basis of the MMN and TubeDETR method, keeping the temporal localization result of MMN and the spatio localization result of TubeDETR. Experiments are conducted on the HC-STVG 2.1 dataset and validated the effectiveness of our solution.

ACKNOWLEDGEMENT

This work is supported by National Science Foundation of China (62072232), Natural Science Foundation of Jiangsu Province (BK20191248) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES


