Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding Supplementary Materials

Chaolei Tan[1] Zihang Lin[1] Jian-Fang Hu[1,2,3*] Wei-Shi Zheng[1,2,3] Jianhuang Lai[1,2,3]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{tanchlei,linzh59}@mail2.sysu.edu.cn, hujf5@mail.sysu.edu.cn,
wszheng@ieee.org, stsljh@mail.sysu.edu.cn

A. Training Loss in Detail

In the manuscript, we have briefly introduced the encoder loss and decoder loss as

$$\mathcal{L}_{enc} = \mathcal{L}_{enc}^{\mathcal{VW}} + \mathcal{L}_{enc}^{\mathcal{VS}} + \mathcal{L}_{enc}^{\mathcal{VP}} \tag{1}$$

$$\mathcal{L}_{dec} = \mathcal{L}_{dec}^{\mathcal{P}} + \mathcal{L}_{dec}^{\mathcal{S}} + \underbrace{\mathcal{L}_{union}^{\mathcal{W}} + \mathcal{L}_{subset}^{\mathcal{W}}}_{\text{weakly-supervised } \mathcal{L}_{dec}^{\mathcal{W}}} \tag{2}$$

Here, we provide more details about the employed encoder and decoder losses.

A.1. Encoder Loss

Our HSCNet encoder is trained by minimizing three semantic alignment losses $\mathcal{L}_{enc}^{\mathcal{VW}}$, $\mathcal{L}_{enc}^{\mathcal{VS}}$, and $\mathcal{L}_{enc}^{\mathcal{VP}}$ based on the video-word semantic similarity matrix $\mathbf{A}^{\mathcal{VW}}$, video-sentence semantic similarity matrix $\mathbf{A}^{\mathcal{VS}}$, and video-paragraph semantic similarity matrix $\mathbf{A}^{\mathcal{VP}}$, respectively.

The word-level semantic alignment loss $\mathcal{L}_{enc}^{\mathcal{VW}}$ is formulated as:

$$\mathcal{L}_{enc}^{\mathcal{VW}} = -\frac{1}{|\mathbb{V}|}\sum_{j\in\mathbb{V}} log\left(\sum_{i\in\mathbb{W}_j^+}\dot{\mathbf{A}}_{ij}^{\mathcal{VW}}\right) - \frac{1}{|\mathbb{W}|}\sum_{i\in\mathbb{W}} log\left(\sum_{j\in\mathbb{V}_i^+}\ddot{\mathbf{A}}_{ij}^{\mathcal{VW}}\right) \tag{3}$$

where $\mathbb{V}$ represents a set of indices for frames related to the words in the paragraph and $\mathbb{W}_j^+$ represents the indices of relevant words corresponding to the j-th frame. $\mathbb{W}$ represents the set of indices of words in the paragraph and $\mathbb{V}_i^+$ represents the indices of relevant frames corresponding to the i-th word. $\dot{\mathbf{A}}^{\mathcal{VW}}$ and $\ddot{\mathbf{A}}^{\mathcal{VW}}$ are obtained by performing column-wise and row-wise softmax on $\mathbf{A}^{\mathcal{VW}}$, respectively. Similarly, the sentence-level semantic alignment loss $\mathcal{L}_{enc}^{\mathcal{VS}}$ is formulated as:

$$\mathcal{L}_{enc}^{\mathcal{VS}} = -\frac{1}{|\mathbb{V}|}\sum_{j\in\mathbb{V}} log\left(\sum_{i\in\mathbb{S}_j^+}\dot{\mathbf{A}}_{ij}^{\mathcal{VS}}\right) - \frac{1}{|\mathbb{S}|}\sum_{i\in\mathbb{S}} log\left(\sum_{j\in\mathbb{V}_i^+}\ddot{\mathbf{A}}_{ij}^{\mathcal{VS}}\right) \tag{4}$$

where $\mathbb{V}$ represents the set of indices for frames related to the sentences and $\mathbb{S}_j^+$ represents the indices of relevant sentences corresponding to the j-th frame. $\mathbb{S}$ represents the set of indices of sentences in the paragraph and $\mathbb{V}_i^+$ represents the indices of relevant frames corresponding to the i-th sentence. $\dot{\mathbf{A}}^{\mathcal{VS}}$ and $\ddot{\mathbf{A}}^{\mathcal{VS}}$ are obtained by performing column-wise and row-wise softmax on $\mathbf{A}^{\mathcal{VS}}$, respectively. The paragraph-level semantic alignment loss $\mathcal{L}_{enc}^{\mathcal{VP}}$ is defined as:

$$\mathcal{L}_{enc}^{\mathcal{VP}} = -log\left(\sum_{j\in\mathbb{V}_p}\ddot{\mathbf{A}}_{ij}^{\mathcal{VP}}\right) \tag{5}$$

where $\mathbb{V}_p$ represents the indices of frames corresponding to the holistic paragraph. $\ddot{\mathbf{A}}^{\mathcal{VP}}$ is obtained by performing row-wise softmax on $\mathbf{A}^{\mathcal{VP}}$. A frame will be associated with the paragraph if it falls within the time interval corresponding to any one of the sentences in the paragraph.

A.2. Decoder Loss

The paragraph-level decoder loss $\mathcal{L}_{dec}^{\mathcal{P}}$, sentence-level decoder loss $\mathcal{L}_{dec}^{\mathcal{S}}$ share the similar formulation consisting of a L_1 distance term and a GIoU term, which has been discussed in the manuscript. For the word-level decoding

loss, it is defined as a union loss $\mathcal{L}^{\mathcal{W}}_{union}$ and a subset loss $\mathcal{L}^{\mathcal{W}}_{subset}$ computed between the word-level predictions $\widetilde{\mathbf{T}}^{\mathcal{W}}$ and sentence-level ground-truth $\mathbf{T}$ in a weakly-supervised manner. Specifically, we first obtain the temporal union of word-wise timestamps within each sentence as $\widetilde{\mathbf{T}}^{\mathcal{U}}$, then $\mathcal{L}^{\mathcal{W}}_{union}$ can be defined as:

$$\mathcal{L}^{\mathcal{W}}_{union} = \frac{1}{N^{\mathcal{S}}} \sum_{i=1}^{N^{\mathcal{S}}} \left(\|\widetilde{\mathbf{T}}^{\mathcal{U}}_{s,i} - \mathbf{T}_{s,i}\|_1 + \|\widetilde{\mathbf{T}}^{\mathcal{U}}_{e,i} - \mathbf{T}_{e,i}\|_1 \right) \quad (6)$$

where $\|\cdot\|_1$ indicates L_1 distance, $N^{\mathcal{S}}$ is the number of sentences. $\mathbf{T}_{s,i}$ and $\mathbf{T}_{e,i}$ denote the starting and ending timestamps corresponding to the i-th sentence, respectively. The subset loss $\mathcal{L}^{\mathcal{W}}_{subset}$ is defined as:

$$\mathcal{L}^{\mathcal{W}}_{subset} = 1 - \frac{1}{N^{\mathcal{S}}} \sum_{i=1}^{N^{\mathcal{S}}} \frac{1}{N^{\mathcal{W}}_i} \sum_{j=1}^{N^{\mathcal{W}}_i} \frac{L^{\mathcal{I}}_{i,j}}{L^{\mathcal{W}}_{i,j}} \quad (7)$$

where $N^{\mathcal{W}}_i$ indicates the number of words in the i-th sentence. $L^{\mathcal{I}}_{i,j}$ indicates the temporal length of intersection between the prediction of the j-th word in the i-th sentence and the ground-truth of the i-th sentence. $L^{\mathcal{W}}_{i,j}$ indicates the length of the prediction of the j-th word in the i-th sentence.

B. Visualization Results of More Examples

In this section, we further visualize the grounding results of our HSCNet on some representative cases to demonstrate the effectiveness of our method.

Firstly, in the case presented in Figure S1, the input video has a long duration up to 11 minutes and the paragraph is composed of 10 sentences referring to different events. Overall, we can see that most of the sentences are successfully grounded by temporal boundaries around the ground-truth. Note that for the third sentence in this paragraph, its relevant event only occupies less than 5% of the total video duration, which makes it extremely difficult to be correctly grounded. However, our model gives a considerably precise prediction that are well overlapped with the target moment, demonstrating the effectiveness of our method.

In the second case presented in Figure S2, each sentence in the paragraph query describes a complex activity composed of multiple actions. These complex activities require the model to fairly well understand the fine-grained correspondence between the video and paragraph so that precise temporal grounding results can be obtained. Once more, our HSCNet gives a series of temporal boundaries that reasonably locate around the ground-truth temporal intervals of different events. It's worth noting that the fifth sentence in this paragraph is favorably grounded by our HSCNet, although it describes a more complex activity consisting of three consecutive actions. Additionally, we observe that our model performs not so well in some situations. For instance, the predicted starting time of the seventh event in Figure S2 is earlier than the ground-truth starting time to a certain extent. The reason might be that the model cannot reason well from the description "finished cutting the peels off the pineapple" to figure out when the man started to peel the pineapple for the last time in the video.

C. Hyperparameter Experiments

Number of encoder layers. In the hierarchical encoder, the number of word-level, sentence-level and paragraph-level layers are set as C_1, C_2 and C_3, respectively. To obtain the best configuration for encoder depth, we first defined C_1=C_2=C_3=c to reduce the huge search space for simplicity. Then we searched different c to observe its influence on model performance. As shown in Table S1, model performance saturates at a small number of encoder layers (i.e., c=1) on ActivityNet-Captions while the model obtains more gains with deeper encoder network and reaches its sweet point at c=3 on TACoS. This may be because the video duration and paragraph length on TACoS are both longer than those on ActivityNet-Captions, and the more complex structure of video-text relations requires more iterations of multi-model interactions to capture its characteristics.

Table S1. Ablation studies on encoder depth c in mIoU metric.

ActivityNet-Captions				TACoS			
c=1	c=2	c=3	c=4	c=1	c=2	c=3	c=4
59.71	59.63	59.38	59.21	37.90	39.39	**40.61**	39.87

Number of video clips. As mentioned in the manuscript, we sample a fixed number of short clips from the video at equal intervals. Here we also provide experimental results on the impact of the number of video clips $N_{\mathcal{V}}$. Specifically, we searched $N_{\mathcal{V}}$ within the range of $\{128, 256, 512, 768\}$ across all the datasets. As shown in Table S2, the best choice of $N^{\mathcal{V}}$ turns out to be 256 and 512 for ActivityNet-Captions and TACoS, respectively. For ActivityNet-Captions, a smaller number of video clips is more suitable because of its relatively short video length. For TACoS, we observed that a small number of video clips hurts the performance to some extent, which is largely due to the long video length in its data distribution.

Table S2. Ablation studies on number of clips $N^{\mathcal{V}}$ in mIoU metric.

ActivityNet-Captions				TACoS			
128	256	512	768	128	256	512	768
59.26	**59.71**	59.27	59.02	36.37	38.51	**40.61**	40.17

Selection of temperature parameter. We investigate the impact of different choices of temperature by searching τ^{ℓ} in $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ on TACoS dataset, as shown in Table S3. It could be seen that model performance

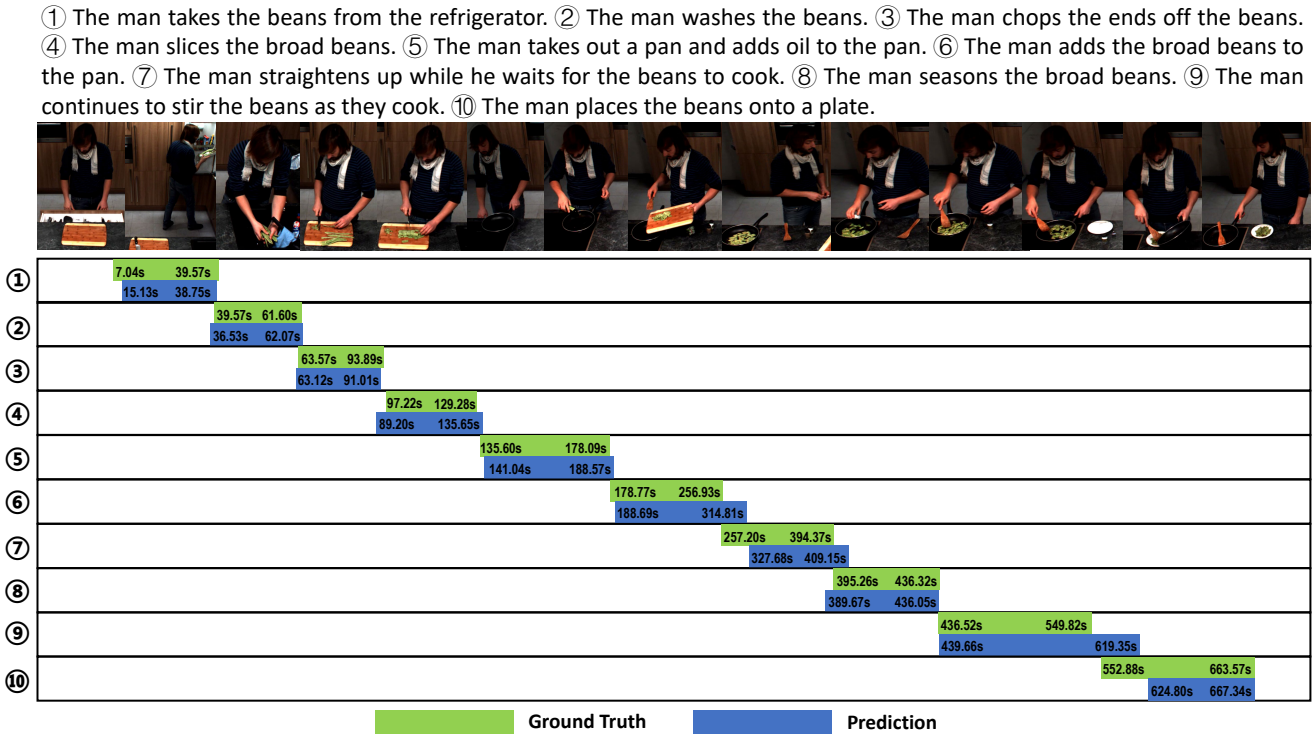

Figure S1. Grounding results of multiple events in a typically long video.

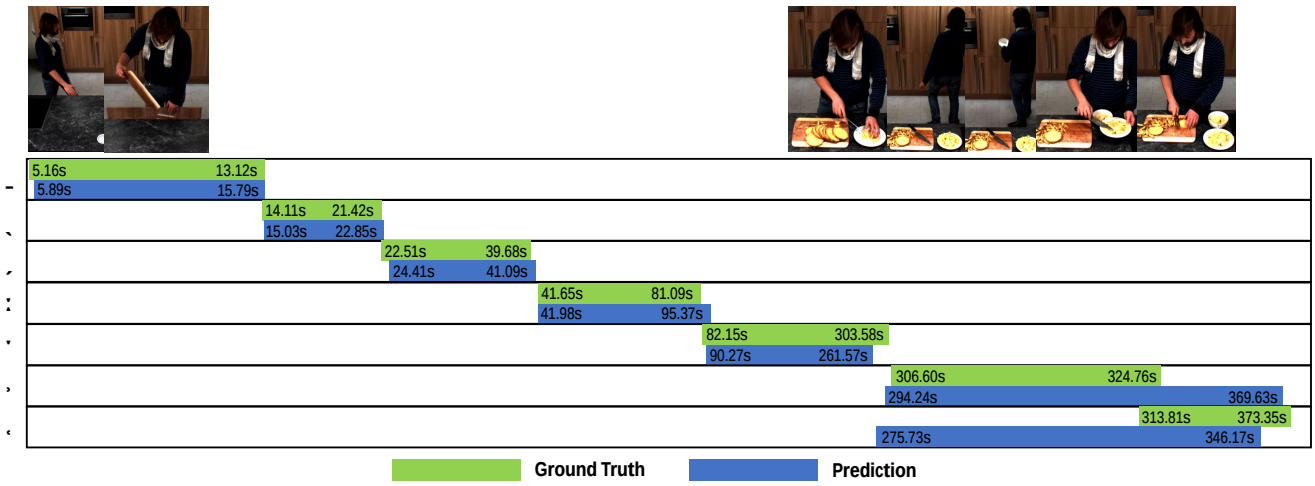

Figure S2. Grounding results of multiple complex events, each of which contains multiple actions.

is quite stable within a reasonable range of temperature, i.e., from 0.05 to 0.2. However, we also found the model performance start to degrade with the temperature being too high, which may attribute to the over-smoothing issue. We adopt $\tau^{\ell} = 0.2$ in all settings of our method because it performs relatively better compared with other choices.

Table S3. Ablation studies on temperature ` in mIoU metric.

τ^{ℓ}	0.05	0.1	0.15	0.2	0.25	0.3
mIoU	40.31	40.21	40.39	**40.61**	39.61	38.89