Progressive Multimodal Interaction Network for Referring Video Object Segmentation

Zihan Ding[1] Tianrui Hui[2] Shaofei Huang[2] Si Liu[1]
Xuan Luo[3] Junshi Huang[3] Xiaoming Wei[3]
[1] Institute of Artificial Intelligence, Beihang University
[2] Institute of Information Engineering, Chinese Academy of Sciences [3] Meituan

Abstract

Referring video object segmentation aims to segment the target object in the video referred by a natural language description. Existing methods perform the coarse late multimodal fusion to align visual and linguistic modalities, identifying the referent matched with the description. Then the memory attention among frames is conducted to refine the results in other frames. To achieve finer multimodal feature fusion, we propose a Progressive Multimodal Interaction Network (PMINet) which performs multimodal feature fusion in each stage of visual backbone, enabling the progressive learning of visual features under the guidance of linguistic features. Afterwards, we conduct other post-processing techniques to refine the mask prediction among all the frames, yielding the temporal consistency of segmentation result of the whole video. Our proposed method achieves the second place on the Track 3: Referring Video Object Segmentation of the 2021 YouTube VOS Challenge.

1. Introduction

Video object segmentation (VOS) is an important task in the computer vision community and obtains increasing attention in recent years. The VOS task enjoys a wide range of applications such as video editing and surveillance video analysis, etc. At present, there are two popular benchmarks for VOS task, i.e., DAVIS [4] and YouTube VOS [6], which contains complicated scenes with multiple objects of the same category in real world.

Traditional semi-supervised setting of VOS task requires the ground-truth mask of the first frame of each video, which restricts the potential application of VOS since the mask annotation consumes much human labor and time. Recently, a new task named referring video object segmentation (RVOS) is proposed in [5] to tackle the limitation of traditional VOS task. Instead of using ground-truth mask, RVOS adopts natural language description to refer the target object in the first frame so that models need to first comprehend the description to identify which object to segment, and then track this object along the rest frames of the video.

Since natural language has various forms and complicated meanings, the model is supposed to achieve comprehensive multimodal understanding to correctly identify the referent in the first frame. However, the method proposed in [5] conducts cross-model attention only in the last stage of visual backbone, which may capture the multimodal context insufficiently with this coarse multimodal feature fusion process. Therefore, in this paper, we propose a Progressive Multimodal Interaction Network (PMINet) to perform multimodal interaction in each stage of the visual backbone, and the multimodal feature is incorporated back to the visual backbone to guide the progressive learning of visual feature. After obtaining each stage of multimodal feature, we further integrate each stage with its former stage in our decoder to progressively upsample the resolution of feature map until reaching the original size of input frame. When the referent in each frame of the video is identified, we exploit MCN [3] to localize each referent again with bounding boxes, then utilize these bounding boxes to select one frame which may produce the best mask prediction of the referent. Finally, we feed the selected frame into an off-the-shelf VOS method CFBI [7] to propagate the high quality mask along the rest frames of the video to obtain the final prediction.

2. Method

The overall architecture of our method is illustrated in Figure 1. Given a video clip $V = \{V_1, V_2, ..., V_T\}$ and a query $L = \{L_1, L_2, ..., L_N\}$, we use a encoder-decoder framework with our proposed CMEM Module to form the PMINet. The coarse mask of each frame is denoted as $M = \{M_1, M_2, ..., M_T\}$. At the same time, we use the MCN to get the coarse detection mask(pixels within the detected box is filled with 1, others filled with 0) of each

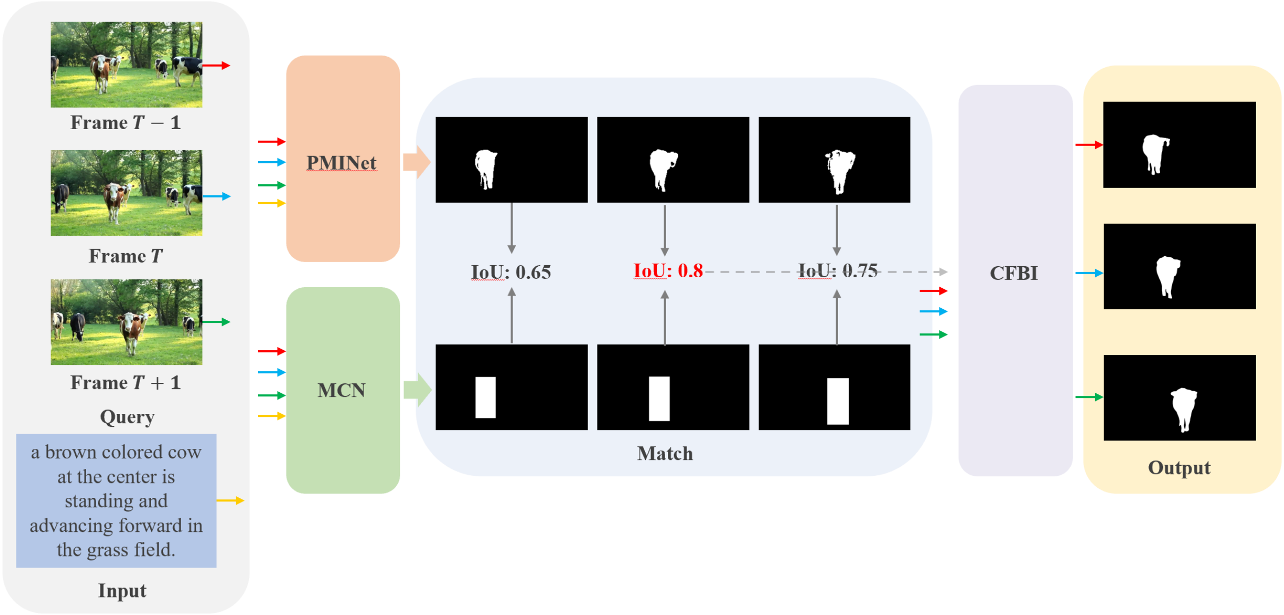

Figure 1. Illustration of our algorithm process.

frame, denoted as $B = \{B_1, B_2, ..., B_T\}$. Then, we calculate the iou between each coarse mask and their detection mask, the frame whose mask has the highest iou will be selected as the reference frame R. Finally, we use the reference frame R and the video clip $V = \{V_1, V_2, ..., V_T\}$ as the input of CFBI to get the refine mask of each frame $F = \{F_1, F_2, ..., F_T\}$.

2.1. Cross-Modal Excitation Modulation

We propose CMEM to conduct interaction between visual and linguistic features for highlighting visual features that are matched with the corresponding linguistic clues. As shown in Figure 2, the CMEM module is inserted into into each stage of the encoders. In order to introduce details of the CMEM module, we take the i-th stage of the encoder as an example and omit the superscript i for simplicity. As illustrated in Figure 3, given the visual feature $V \in \mathbb{R}^{H\times W\times C_V}$ of the target frame and the linguistic feature $L \in \mathbb{R}^{N\times C_L}$of the language query, we first get an attention map $A \in \mathbb{R}^{N\times HW}$ by conducting cross-modal attention, which measures the feature relevance between each word and the target frame. Concretely, V and L are first transformed to the same subspace by linear transformation, denoted as $V^{'} \in \mathbb{R}^{H\times W\times C_M}$ and $L^{'} \in \mathbb{R}^{N\times C_M}$. Then, $V^{'}$ is reshaped to $\mathbb{R}^{HW\times C_M}$ to match the matrix dimensions. We further perform matrix product between $V^{'}$ and $L^{'}$ to obtain attention map A as follows:

$$A = L^{'} \otimes V^{'T} \tag{1}$$

where $\otimes$ denotes matrix product.

Then we add all the values on the HW dimension and normalize it as follows:

$$w = \sum_{j=1}^{HW} A^j, \\ \tilde{w} = Softmax(\frac{w}{\|w\|_2}), \tag{2}$$

where $\|\cdot\|_2$ denotes the L_2 norm of a vector, $A^j \in \mathbb{R}^N$ is the feature relevance between the j-th spatial location and N words, and $\tilde{w} \in \mathbb{R}^N$ is the normalized global feature relevance between each word and the whole target frame. Therefore, the adaptive language feature can be attained by linearly re-combine features of N words $l = \sum_{k=1}^{N}(\tilde{w}^k L^k) \in \mathbb{R}^{C_L}$.

Afterwards, we adopt a linear layer and $sigmoid$ function are adopted to transform l to $\mathbb{R}^{C_V}$ dimensions and generate channel-wise modulation weights $\tilde{l} \in \mathbb{R}^{C_V}$.

Finally, we multiply $\tilde{l}$ with feature of the target frame V to high-light sentence-relevant visual feature channels and add the modulated feature with original V to ease optimization:

$$\tilde{V} = V + V \odot \tilde{l}, \tag{3}$$

where $\odot$ denotes element-wise product, $\tilde{V}$ is the output of the CMEM module and serves as the input feature of the next stage in the encoder.

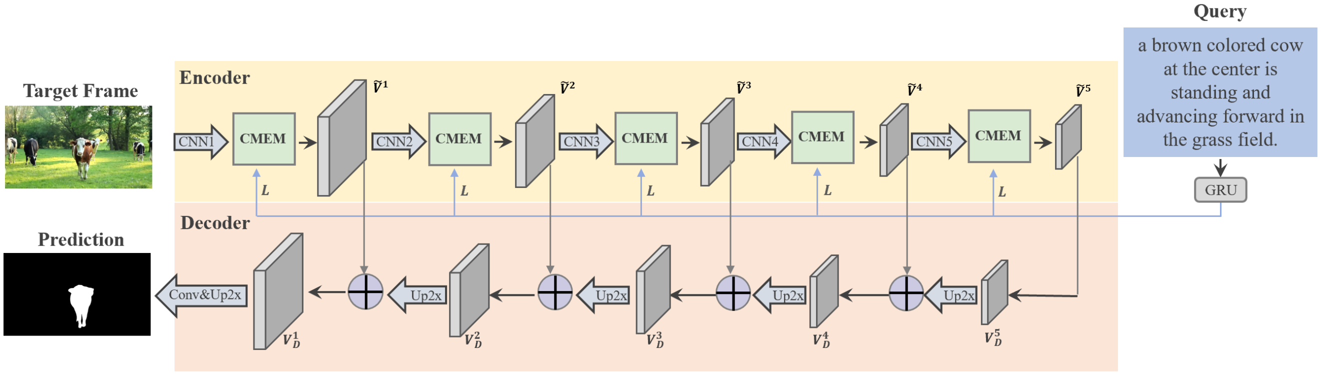

Figure 2. Illustration of our PMINet.

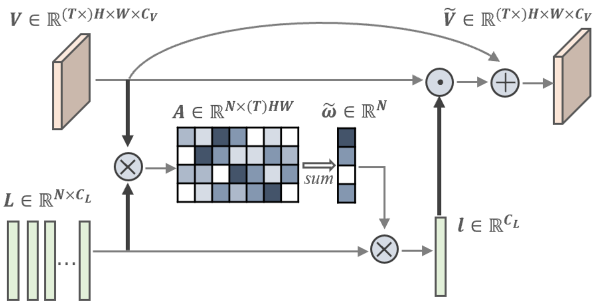
$V \in \mathbb{R}^{(T\times)H\times W\times C_V}$
$\tilde{V} \in \mathbb{R}^{(T\times)H\times W\times C_V}$
$A \in \mathbb{R}^{N\times(T)HW}$
$\tilde{\omega} \in \mathbb{R}^N$
sum
$L \in \mathbb{R}^{N\times C_L}$
$l \in \mathbb{R}^{C_L}$

Figure 3. Illustration of our CMEM module.

mcn	cfbi	Overall	J	F
		0.482	0.467	0.496
	✓	0.530	0.515	0.545
✓	✓	0.542	0.530	0.555

Table 1. Ablation study in YouTube-VOS 2021 validation set.

Team	Overall	J	F
leonnnop	0.607	0.594	0.620
nowherespyfly	0.494	0.484	0.503
feng915912132	0.482	0.474	0.490
Merci1	0.412	0.406	0.418
wangluting	0.407	0.395	0.418
dongming.wu	0.363	0.355	0.371

Table 2. Results in YouTube-VOS 2021 test set. Our method achieves an overall second place.

3. Experiments

3.1. Training Details

We utilize ResNeSt [8] pretrained on ImageNet dataset as the visual backbone of our PMINet. And the linguistic feature is encoded by GRU [1]. The maximum length of the input sentence is set as 20. The input frames are resized and padded to 320 $\times$ 320. We adopt Adam [2] as the optimizer and the initial learning rate is set as $5e^{-4}$. We train the network for 9 epochs and reduce the learning rate by $10x$ at the 8-th epoch. MCN is trained using the same setting as our PMINet and CFBI is only used for testing.

3.2. Components analysis

We evaluate our model on YouTube-VOS 2021 validation set. Region similarity J and the contour accuracy F are used as metrics, following the official test scripts.

As shown in Table 1, introducing CFBI as post-processing tool to refine the coarse masks from the Encoder-Decoer Network can bring a 0.048 improvement(second row). When there is no matching between the coarse masks and the detection boxes from MCN, the frame with highest foreground average score is selected as the reference frame. After adding the matching part(third row), the performance is further boosted by 0.012, which demonstrate that the detection boxes from MCN can be a good reference for the target described by language description.

3.3. Model ensemble

We also utilize the model ensemble which consists of 4 models with different hyperparameter settings to promote model performance. Simply, we average the predicted score map after $sigmoid$ of different models. After the model ensemble, we achieve 0.548 global mean on YouTube-VOS 2021 validation set. The results of the same model are submitted onto the test server an obtain a global mean of 0.494. As shown in Table 2, our method ranks first on YouTube-VOS 2021 semi-supervised video object segmentation challenge on both seen and unseen objects

4. Conclusion

In this paper, we propose an Progressive Multimodal Interaction Network (PMINet) for referring video object segmentation. Our approach achieves an overall score of 0.494, ranking second place on YouTube-VOS 2021 referring video object segmentation.

References

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[3] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020. 1

[4] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 1

[5] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[6] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1

[7] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *Proceedings of the European Conference on Computer Vision*, 2020. 1

[8] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 3