The Third Place Solution for CVPR2022 Referring Youtube-VOS challenge

Peng Liu Quanlong Zheng Zhengxia Zou Yandong Guo
OPPO Research
{arlen.liu, zhengquanlong}@oppo.com, zhengxiazou@gmail.com, yandong.guo@live.com

Abstract

Referring video object segmentation aims to segment language-referred objects in the video. This task requires understanding both semantic and textual information and combining them together. ReferFormer[9] get excellent results with end-to-end framework on this task. We build our model based on ReferFormer and make some changes. In this report we describe the technical details of the implementation we submitted. Our method achieves the third place on CVPR2022 Referring Youtube-VOS challenge.

1. Introduction

Referring video object segmentation(RVOS) is an emerging and challenging multi-modal task. In the task of referring image segmentation, the text description mainly focuses on the appearance features or spatial relationships of objects in a single image. The RVOS task describes the continuous and abstract actions of the target in multiple frames, which requires the model to have stronger spatiotemporal modeling capabilities and to ensure the consistency of the segmentation target in terms of temporal association.

2. Approach

2.1. ReferFormer

Our method is based on ReferFormer. ReferFormer proposes a simple and unified, Transformer-based end-to-end RVOS framework. It uses language descriptions as constraints on queries so that tasks can be accomplished with a small number of queries. As of the start of the competition, it has achieved state-of-the-art performance on on Ref-Youtube-VOS[8], Ref-DAVIS17[2], A2D-Sentences and JHMDB-Sentences[1].

As shown in Figure 1, ReferFormer is mainly composed of four parts: Backbone, Transformer, cross-modal FPN and instance segmentation. In our experiments, we use multiple visual backbones for the model ensemble.

Method	*J&F*
Baseline	64.9
Increase the training frame-number to 6	65.0
Our frame selection strategy	65.3
Compare continuous expressions	66.3

Table 1. Ablation study on the test-dev set. This experiment is based on video swin base.

2.2. Backbone

Many classical segmentation network architectures are designed based on CNN encoder to extract visual features. In recent years, with the emergence of transformer, many transformer backbones for visual tasks, such as Swin-Transformer[4], show excellent performance on vision task. Based on Swin-Transformer, Video-Swin-Transformer[5] works great in the field of video vision. In our experiment, we use swin large, video swin base, video swin small and video swin tiny as our visual backbones.

As for given language description, we use $RoBERTa_Base$[3] to extract the test feature.

2.3. Frame selection strategy

Following the ablation study on ReferFormer, the performance get better when the frame number gets larger from 1 to 5. It shows using more frames to form a clip helps the model better aggregate the temporal action-related information. Based on this idea, we increased the number of frames to 6, and achieved a slight improvement.

The frame selection strategy on ReferFormer divides the selected frame into local frame and global frame. We find the images are taken every 5 frames from videos in Ref-Youtube-VOS. It means there is a lot of variation between consecutive frames, especially when the movement is intense. So we try to increase the compactness of local selected frames. However, taking all the nearer local frames will lose some video timing information so we still select a portion of the global frame. Finally we try several way and find adding two adjacent frames directly get the best results.

The overall strategy is as follows

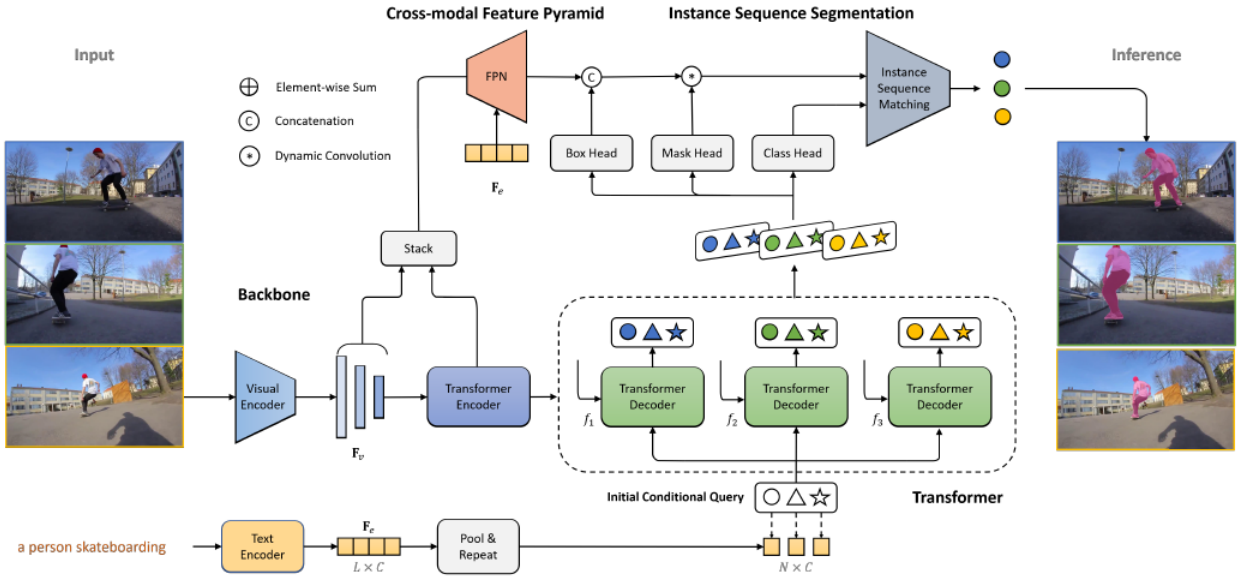

Figure 1. ReferFormer, figure taken from [9]

Algorithm 1 Frame selection strategy

1: Take one frame every 6 frames from the video as the reference frame and record its number as i. Add reference frame into selected frames.
2: Add (i-1)th and (i+1)th frames into selected frames if they exist.
3: Random $j \in (1, 4]$, if (i-j)th frame exists and not in selected frames, add it into selected
4: Random $k \in (1, 4]$, if (i+k)th frame exists and not in selected frames, add it into selected
5: **repeat**
6: Random select from $((-\infty, i-4) \cup (i+4, \infty)) \cap Allframes$ into selected frames
7: **until** Selected frames are enough.

3. Experiments

3.1. Training details

Our training setting is based on ReferFormer. We use joint-training models on the image referring segmentation datasets Ref-COCO[10], Ref-COCOg[10] and Ref-COCO+[7] from ReferFormer as our pretrained models. Then we fine-tune four models with different visual encoders, i.e., swin large, video swin base, video swin small and video swin tiny on Ref-Youtube-VOS.

We employ photometric distortion, random horizontal flip, random scale and random crop on the input images to augment the dataset in training process. Our model is optimized using AdamW[6] optimizer with the weight decay of 5×10^{-4}. The initial learning rate is set to 5×10^{-6} for visual backbone and 10^{-5} for the rest. The fine-tuning procedure runs for 3 epochs on Ref-Youtube-VOS with the learning rate decays divided by 10 at epoch 1.

For video visual backbones, we increase the training frame-number to 6 and use our new frame selection strategy. For image visual backbone, we use basic experiment setting. The text encoder is froze all the time.

3.2. Test details

We adopt our the multi-scale testing {288, 320, 360, 384, 416} for four models. Limited by the memory size of GPU, we get only a part of test results when inference size or model is too large. Then we average results from different scales. We try two size combinations and find even though size-416 is missing part of the images, the result of {320, 360, 384, 416} is better than {288, 320, 360, 384}.

We find that one object almost corresponds to two expressions in training, validation and test dataset. We consider two expression correspond to the same object if their segmentation mask IOU is larger than a predefined threshold. Then we replace the low-confidence results with the higher one.

After post-processing, we ensemble all the 16 results (4 scales x 4 encoders).

Size	*J&F*
288	64.2
320	65.0
360	65.3
384	65.5
416	-
{288, 320, 360, 384}	65.9
{320, 360, 384, 416}	66.6

Table 2. Results. Different single size test and multi-scale test on the test-dev set. This experiment is based on video swin base

Team	*J&F*	*J*	*F*
Bo____	64.1	62.2	66.1
jiliushi	61.7	59.8	63.6
PENG	60.8	58.9	62.7
ds-hohhot	59.6	57.9	61.2
JQK	59.4	57.7	61.1
nero	58.0	56.1	59.9

Table 3. Results in Ref-YouTube-VOS 2022 test-challenge set. Our method achieves a third place.

4. Conclusion

In this report, we propose a frame selection strategy based on ReferFormer and a multi-model ensemble method with multi-scale test. Our approach achieves the third place with a *J&F* score of 60.8% on test-challenge set on referring YouTube-VOS 2022 challenge.

References

[1] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018.

[2] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2018.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[5] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[7] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[8] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pages 208–223. Springer, 2020.

[9] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. *arXiv preprint arXiv:2201.00487*, 2022.

[10] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.