Semi-SWA: Semi-Supervised Stochastic Weight Averaging for Video Instance Segmentation

Fengliang Qi, Jing Xian, Zhuang Li, Bo Yan, YuChen Hu, Hongbin Wang
Ant Group
{qifengliang.qfl, xianjing.xj, jiangzi.lz, lengyu.yb, huyuchen.hyc, hongbin.whb}@antgroup.com

Abstract

Video instance segmentation is a comprehensive task related to the video understanding, dense prediction, and multi-object tracking. OVIS is the first large-scale dataset for occluded video instance segmentation, where most objects are more or less occluded in complicated scenes. In this work, we enhance the existing sota method on OVIS, i.e., IDOL, and achieve the **first** place on the 2nd Occluded Video Instance Segmentation Challenge, Multiple Object Tracking, and Segmentation in Complex Environments ECCV 2022 Workshop.

Introduction

Video instance segmentation aims at detecting, segmenting, and tracking object instances simultaneously in a given video. It attracted considerable attention after first defined (Yang, Fan, and Xu 2019) in 2019 due to the huge challenge and the wide applications in video understanding, video editing, autonomous driving, augmented reality, etc.

In Defense of OnLine models, termed IDOL, is the SOTA method which achieves 42.6 mAP on the validation set of OVIS. It argues that the per-clip segmentation doesn't outperform per-frame segmentation a lot in mask quality, and mask quality is also not the reason for the poor performance of online methods. We use the IDOL as our baseline framework and refine the open-source repository to make the inference process feasible. We also apply a stronger baseline, i.e., CBNet (Liang et al. 2021) to improve our performance.

Furthermore, we propose the Semi-SWA (S2WA) method. The S2WA method injects the pseudo-annotated dataset into the conventional SWA training procedure, which could guide the updating direction of model weight from the training set to the unseen target set.

Related Works

Video Instance Segmentation evolves from Image Instance Segmentation and aims to track objects from the whole video sequence. The initial method Mask-Track R-CNN (Yang, Fan, and Xu 2019) is built upon Mask R-CNN and introduces a tracking head to associate each instance in the video. SipMask (Cao et al. 2020) proposes a spatial preservation module to generate spatial coefficients for mask predictions based on the one-stage FCOS. CrossVIS (Yang et al. 2021) proposes a learning scheme that uses the instance feature in the current frame to pixel-wisely localize the same instance in other frames. SeqFormer (Wu et al. 2022a) dynamically allocates spatial attention on each frame and learns a video-level instance embedding, which greatly improves the performance. IDOL (Wu et al. 2022b) proposes a temporally weighted softmax score for instance matching and a memory bank-based association strategy to atain a strong instance association and improve the association quality of the online model, which can be applied to both ongoing and long videos and complex scenarios.

Our Method

Refined IDOL

In **D**efense of **On**Line models, termed IDOL, is the state-of-art framework for video instance segmentation. In our experiments, we use the IDOL as our baseline method and refine the open-source code for higher inference efficiency and lower storage overhead. To this end, we mainly optimize the post-processing procedure in IDOL where the predictions in each frame are filtered based on some pre-defined selection thresholds and the predictions across multi frames are associated to construct the final tracklets. Specifically, for each frame i to be inferenced, the forward network would output the information of N instances (e.g., $N = 300$ in IDOL). For each predicted instance j, the located bounding box $\mathcal{B}$, dense pixel mask $\mathcal{M}$, categorical logits $\mathcal{L}$ and instance-level embedding f are bundled. After all frames in video v are forwarded, we obtain the set of predicted instances:

$$\mathbb{S}^v = \{\mathbb{S}_i\}_{i=1}^{M^v}; \mathbb{S}_i = \{(\mathcal{B}_j, \mathcal{M}_j, \mathcal{L}_j, f_j)\}_{j=1}^{N},$$

where the M^v is the length of video v. The post-processing function (PPF) accepts $\mathbb{S}^v$ and output the final result for video v:

$$\mathbb{R}^v = \text{PPF}(\mathbb{S}^v, \boldsymbol{\tau}) = \{\mathcal{T}_k\}_{k=1}^{K},$$

where the $\boldsymbol{\tau}$ is the list of hyper-parameters, $\mathcal{T}_k$ is the tracklet for instance k and K is the number of final valid predicted instances after post-processing.

There are three main steps in the post-processing procedure: BBox-aware instance filter, Mask-aware instance filter, and Embedding-aware instance association. The first one follows the conventional IoU-based Non-maximum Suppression (NMS) (Girshick et al. 2013) strategy where a

classification threshold $\tau_0 = 0.1$ firstly truncates the low-confidence predictions and then if the IoU between two bounding boxes exceeds the threshold (e.g., 0.9 in IDOL), the prediction with lower classification confidence will also be eliminated. However, the sequential and recursive operations in NMS result in non-negligible latency. For mask NMS in the second step, this drawback is further magnified. Compared to the bounding box, it takes more time to compute the IoU of each mask pair, thus leading to a large overhead. We address this problem by introducing Matrix NMS (Wang et al. 2020), which performs NMS with parallel matrix operations in one shot.

Matrix NMS views the mask deduplication process by considering how a predicted mask m_j is being suppressed. For m_j , its decay factor is affected by: (a) The penalty of each prediction m_i on m_j ($s_i > s_j$), where s_i and s_j are the confidence scores; and (b) the probability of m_i being suppressed. For (a), the penalty of each prediction m_i on m_j could be easily computed by $f(\text{iou}_{i,j})$. For (b), the probability of m_i being suppressed is not so elegant to be computed. However, the probability usually has positive correlation with the IoUs. So here we directly approximate the probability by the most overlapped prediction on m_i as

$$f(\text{iou}_{\cdot,i}) = \min_{\forall s_k > s_i} f(\text{iou}_{k,i}),$$

To this end, the final decay factor becomes

$$\text{decay}_j = \min_{\forall s_i > s_j} \frac{f(\text{iou}_{i,j})}{f(\text{iou}_{\cdot,i})},$$

We use the linear decremented functions $f(\text{iou}_{i,j}) = 1 - \text{iou}_{i,j}$ and only the decay_j larger than 0.5 will survive.

CBNet Backbone

Composite Backbone Network (CBNet) (Liang et al. 2021) is a simple and novel composition approach to use existing pre-trained backbones under the pretraining fine-tuning paradigm. Unlike most previous methods that focus on modular crafting and require pre-training on ImageNet to strengthen the representation, CBNet improves the existing backbone representation ability without additional pre-training by grouping multiple identical backbones together. Specifically, parallel backbones (named assisting backbones and lead backbone) are connected via composite connections. The output of each stage in an assisting backbone flows to the parallel and lower-level stages of its succeeding sibling. Finally, the features of the lead backbone are fed to the neck and detection head for bounding box regression and classification. In our experiment, we connect two identical Swin-L backbones and use the pretraining weight recommended in IDOL as our initial optimization point.

Stochastic Weight Averaging

Stochastic Weight Averaging (SWA) (Izmailov et al. 2018) is proposed to achieve the wider solutions than the optima found by SGD. The loss on the train set would be shifted with respect to the test error. The SGD generally converges to a point near the boundary of the wide flat region of optimal points. SWA on the other hand is able to find a point centered in this region, often with slightly worse train loss but with substantially better test error. In our experiment, we firstly train the model for 12000 iterations and reduce learning rate by a factor of 10 at the 8000 iterations. Then we further train the model for 12000 iterations on the training set with cyclical learning rate, whose value linearly drop from 1×10^{-4} to 0 for every 2000 iterations. We save the checkpoint once the learning rate reaches 0.0 and 6 weights are equally averaged to get the final SWA model.

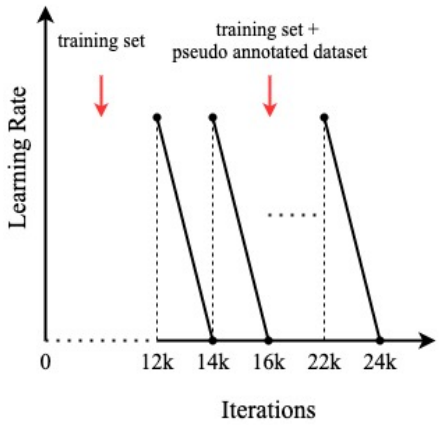

Figure 1: We inject the pseudo-annotated dataset to the SWA training procedure.

Semi-supervised Learning

To further improve the model's performance, we use semi-supervised learning on the validation set. To this end, we generate the pseudo labels on the target datasets. To avoid the adverse effect of the noisy pseudo annotations, we increase the classification threshold τ_0 from 0.1 to 0.5. After this, we additionally inject the pseudo-annotated datasets into the SWA training procedure as shown in Fig.1. We name this training strategy Semi-SWA (S2WA), which combines semi-supervised learning and stochastic weight averaging together.

Experiments

We conduct ablation experiments on the validation set of OVIS. The OVIS Consists of 296k high-quality instance masks, 25 commonly seen semantic categories, 901 videos with severe object occlusions, and 5,223 unique instances. We use average precision (AP) at different intersection-over-union (IoU) thresholds and average recall (AR) as our evaluation metrics. The IoU in video instance segmentation is the sum of the intersection area over the sum of the union area across the video.

Empirical Results

As illustrated in Tab1, the superior CBNet backbone improve the performance by 0.36. And the vanilla SWA training strategy further enhances the mAP to 42.81. Our proposed S2WA strategy reaches the best performance, which denotes the superiority of this method.

Method	Swin-L	CBNet	SWA	S2WA
mAP	41.39	41.75	42.81	43.81
ΔmAP	-	+0.36	+1.06	+1.00

Table 1: The performance of different methods on the validation set.

Conclusions

In this paper, we refine the inference procedure of IDOL using the matrix NMS method, which could parallelly remove the highly-overlapped mask. To further improve the performance, we propose the Semi-SWA method, where the pseudo-annotated datasets participate into the conventional SWA training process. We achieve the **first** place on the 2nd Occluded Video Instance Segmentation Challenge, Multiple Object Tracking and Segmentation in Complex Environments ECCV 2022 Workshop.

References

Cao, J.; Anwer, R.; Cholakkal, H.; Khan, F.; Pang, Y.; and Shao, L. 2020. Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 1, 4, 10, 11.

Girshick, R. B.; Donahue, J.; Darrell, T.; and Malik, J. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR* abs/1311.2524. URL http://arxiv.org/abs/1311.2524.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. *CoRR* abs/1803.05407. URL http://arxiv.org/abs/1803.05407.

Liang, T.; Chu, X.; Liu, Y.; Wang, Y.; Tang, Z.; Chu, W.; Chen, J.; and Ling, H. 2021. CBNetV2: A Composite Backbone Network Architecture for Object Detection. *CoRR* abs/2107.00420. URL https://arxiv.org/abs/2107.00420.

Wang, X.; Zhang, R.; Kong, T.; Li, L.; and Shen, C. 2020. SOLOv2: Dynamic, Faster and Stronger. *CoRR* abs/2003.10152. URL https://arxiv.org/abs/2003.10152.

Wu, J.; Jiang, Y.; Bai, S.; Zhang, W.; and Bai, X. 2022a. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2, 5, 9, 12, 13.

Wu, J.; Liu, Q.; Jiang, Y.; Bai, S.; Yuille, A.; and Bai, X. 2022b. In Defense of Online Models for Video Instance Segmentation. In *ECCV*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5188–5197.

Yang, S.; Fang, Y.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; and Liu, W. 2021. Crossover learning for fast online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2–14.