SegmATRon: Embodied Adaptive Semantic Segmentation for Indoor Environment

Tatiana Zemskova
MIPT
Moscow, Russia
zema-tania@yandex.ru

Margarita Kichik
MIPT
Moscow, Russia
kichik.mg@phystech.edu

Dmitry Yudin
AIRI, MIPT
Moscow, Russia
yudin.da@mipt.ru

Aleksandr Panov
AIRI, MIPT
Moscow, Russia
panov.ai@mipt.ru

Abstract

This paper presents an adaptive transformer model named SegmATRon for embodied image semantic segmentation. Its distinctive feature is the adaptation of model weights during inference on several images using a hybrid multicomponent loss function. We studied this model on datasets collected in the photorealistic Habitat Simulator. We showed that obtaining additional images using the agent's actions in an indoor environment can improve the quality of semantic segmentation.

1. Introduction

Recently, embodied methods [2, 4, 5, 7] have appeared, which demonstrate that the information fusion from an image sequence during indoor navigation positively affects the quality of detection. However, the existing embodied approaches do not consider semantic segmentation, another important perception task for intelligent agents. Inspired by work [4], we propose and investigate an adaptive learning method with different action policies for improvement of semantic segmentation in the Habitat photorealistic environment.

2. Method

Adaptive Learning. The key idea of adaptive semantic segmentation is the adaptation of model weights during inference on several images using a hybrid multicomponent loss function $\mathcal{L}^{\phi}_{adapt}(\theta, \mathbf{F})$. The loss function is parameterized by ϕ and depends on parameters θ of a segmentation model and a sequence of frames $\mathbf{F}$. The goal during the training process is to minimize the ground-truth loss $\mathcal{L}_{segm}(\theta, \mathbf{F})$, where the parameters θ are updated by backpropagation through adaptive gradients, see Eq. (1):

$$\min_{\theta,\phi} \sum_{\mathbf{F}\in\mathbf{R}_{all}} \mathcal{L}_{segm}(\theta - \alpha\nabla_{\theta_{head}}\mathcal{L}^{\phi}_{adapt}(\theta, \mathbf{F}), \mathbf{F}). \quad (1)$$

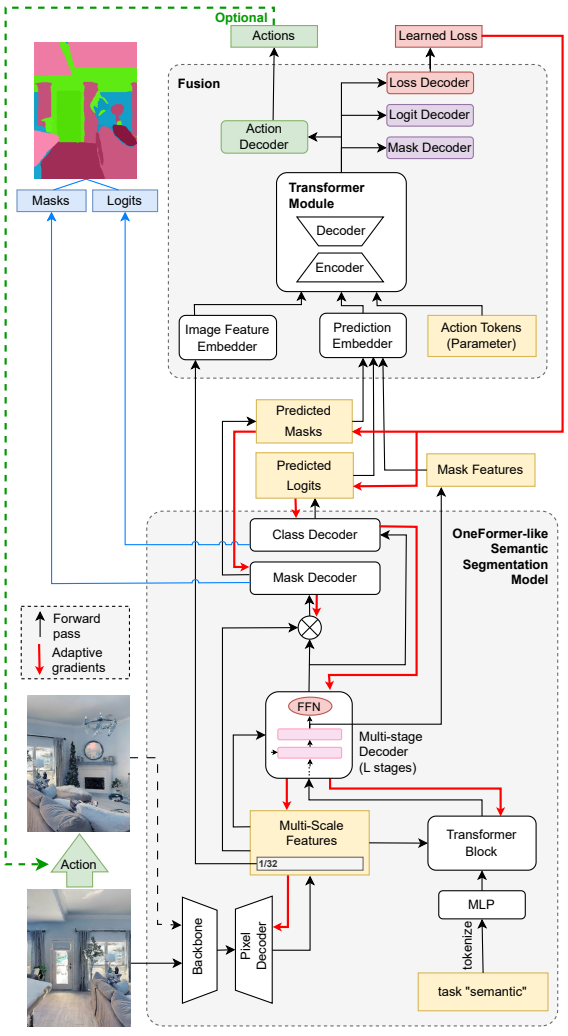

Figure 1. Inference scheme of proposed SegmATRon approach.

Due to the computational cost of backpropagation, we compute the adaptive gradients only for parameters θ_{head} of a semantic segmentation model head.

Table 1. Comparison of SegmATRon method with different action policies and Single Frame baseline on the Val. and Test datasets.

Method	Action Policy (Training)	Action Policy (Inference)	Dataset	$mIoU, \%$	$fwIoU, \%$	$mACC, \%$	$pACC, \%$
OneFormer	Single Frame	Single Frame	Val.	38.0 ± 0.5	67 ± 0.4	47.9 ± 1.1	78.5 ± 0.2
SegmATRon	Random Rotation	Random Rotation	Val.	$\mathbf{39.2 \pm 0.9}$ $(+3\%)$	$\mathbf{67.9 \pm 0.3}$ $(+1\%)$	$\mathbf{49.5 \pm 1.2}$ $(+3\%)$	$\mathbf{78.9 \pm 0.3}$ $(+0.5\%)$
SegmATRon	Look down	Look down	Val.	38.2 ± 1.7 $(+0.5\%)$	67.4 ± 0.3 $(+0.6\%)$	48.2 ± 2.3 $(+0.6\%)$	78.3 ± 0.2 (-0.3%)
OneFormer	Single Frame	Single Frame	Test	40.5 ± 1.1	67.0 ± 0.4	51.8 ± 1.3	78.5 ± 0.3
SegmATRon	Random Rotation	Random Rotation	Test	40.5 ± 1.1 $(+0\%)$	67.5 ± 0.4 $(+0.7\%)$	51.8 ± 1.4 $(+0\%)$	$\mathbf{78.6} \pm 0.3 (+0.1\%)$
SegmATRon	Random Rotation	Move backward	Test	39.9 ± 1.6 (-1%)	67.4 ± 0.2 $(+0.6\%)$	51.4 ± 2.2 (-0.7%)	78.3 ± 0.2 (-0.3%)
SegmATRon	Look down	Look down	Test	40.6 ± 1.2 $(+0.2\%)$	67.4 ± 0.2 $(+0.6\%)$	$52.1 \pm 1.9 (+0.6\%)$	78.4 ± 0.2 (-0.1%)
SegmATRon	Look down	Move backward	Test	$\mathbf{40.7 \pm 0.5}$ $(+0.5\%)$	$\mathbf{67.7 \pm 0.2}$ $(+1\%)$	$\mathbf{52.1 \pm 0.5} (+0.6\%)$	$\mathbf{78.6 \pm 0.2} (+0.1\%)$

Transformer model. As a segmentation model (see Fig. 1) we consider the modification of OneFormer [3], which is one of state-of-the-art methods for semantic segmentation. The off-the-shelf OneFormer that uses a single frame to make prediction of masks and labels represents a baseline approach for comparison with our SegmATRon model. Following the idea of Interactron [4] we choose a Transformer model to combine predictions and image features from two frames in order to predict the loss for the adaptive backward pass. We keep the architecture of Fusion module the same as it is provided by the authors of Interactron [4], therefore our Fusion module contains MLP decoders for the learned loss, masks, logits and actions. However, in our experiments we use only the learned loss output.

During training, the parameters ϕ of the Fusion module are updated by the ground-truth loss that is computed from the segmentation annotation and predictions made by OneFormer after backpropagation of adaptive gradients. Then, the parameters of OneFormer model are optimized in order to reduce the ground-truth loss with adapted weights. During inference, there is no ground truth and only the parameters of the head of OneFormer model are updated by the learned loss predicted by the Fusion module.

Datasets for Adaptive Learning in Habitat Environment. All datasets were collected in Habitat environment. OneFormer model was pretrained using 250K images, collected in random navigable points of train scenes from Habitat-matterport 3d semantics (HM3DSem) dataset [6] with 40 Matterport3D categories [1]. As initial points of view we considered random navigable points containing more than 4 instances of objects that are not wall, ceiling or floor. To train our models we collected a dataset of 944 points in train scenes of HM3DSem with possible additional points of view. A validation dataset (Val.) of 140 points was collected from validation scenes of HM3DSem. For train and validation dataset we considered 4 possible view points obtained with following agent actions: turn left, turn right, look up, look down. All rotations are made by 30°. Finally, we explored the application of our model to another set of actions by collecting the Test set of 129 random points with additional action corresponding to observing a scene from more distant point of view by moving backward.

3. Experimental Results

We train neural network models on a server with 1 Nvidia Tesla V100 GPU. We pretrain OneFormer model with Swin-L backbone, crop size 640×640, and batch size equal to 4. The weights are initialized by OneFormer model trained on ADE20k [8]. To train SegmATRon as well as Single frame baseline we follow a training procedure described by authors of Interactron [4], but we reduce the epoch number to 50 due to fast convergence of segmentation model. On inference we compute mean value of each metric (see. [8]) for 20 runs.

On the validation dataset the SegmATRon with Random rotation action policy significantly outperforms the baseline OneFormer approach (see Tab. 1). The heuristic policy of Looking down also demonstrates an improvement of $fwIoU$ compared to baseline approach. Since the SegmATRon approach requires the backpropagation of adaptive gradients during inference, more computing resources are needed for this method.

On the test set, the best results are demonstrated by the SegmATRon method trained with heuristic policy of Looking down and Moving backward from the initial point during inference (see Tab. 1). The heuristic policy of moving backward does not improve the quality of the SegmATRon approach trained with Random rotation policy. These results suggest that using certain policies to select the next view point can improve the quality of the SegmATRon model even if only one additional frame is used. Therefore, further study concerning the search for the optimal policy for choosing the next action is of interest.

4. Conclusion

Our results show that the semantic segmentation quality benefits from mechanism of multicomponent loss learning which allows us to use an additional point of view. We have also demonstrated that the action strategy has a significant impact on the result, while further research on the number of actions and their automatic learning are reasonable. A future perspective for SegmATRon approach would be action policy optimization via reinforcement learning based on segmentation loss, which we are currently working on.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[2] Wenhao Ding, Nathalie Majcherczyk, Mohit Deshpande, Xuewei Qi, Ding Zhao, Rajasimman Madhivanan, and Arnie Sen. Learning to view: Decision transformers for active object detection. *arXiv preprint arXiv:2301.09544*, 2023. 1

[3] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[4] Klemen Kotar and Roozbeh Mottaghi. Interactron: Embodied adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14860–14869, 2022. 1, 2

[5] Zhenyu Wu, Ziwei Wang, Zibu Wei, Yi Wei, and Haibin Yan. Smart explorer: Recognizing objects in dense clutter via interactive exploration. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6600–6607. IEEE, 2022. 1

[6] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023. 2

[7] Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2040–2050, 2019. 1

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2