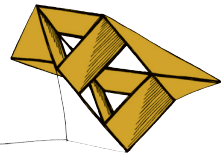
FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

STUDENT
SCIENTIFIC CONFERENCE 2023



Encoder semantic space adaptation between two semantic segmentation models

Jakub Straka[1]

1 Introduction

Semantic segmentation is a computer vision task that aims to classify each pixel in an image in a predefined class, in other words, for each class, find a segmentation mask that indicates all objects of that class in the image.

The most common and widely used convolutional neural network (CNN) for this task is U-Net introduced by Ronneberger et al. (2015). U-Net consists of two parts, an encoder, and a decoder. The encoder aims to create a semantic representation of the image in a high-dimensional space. This is achieved by passing the image through convolutional layers, which gradually reduce the spatial dimension and increase the number of channels. The decoder does the opposite process. It gradually reduces the number of channels and increases the spatial dimension of the image. Its goal is to decode semantic information into masks of individual classes.

Recently was introduced segmentation model by Kirillov et al. (2023) called the Segment Anything Model (SAM). The model is based on the architecture of the transformer, therefore inherently adopting encoder-decoder architecture. Similarly to U-Net, the SAM encoder aims to create a semantic representation of an input image. In this case, the encoder is Vision Transformer (ViT) with 12 layers. Unlike U-Net, SAM also contains a prompt encoder block. This block allows the user to interactively select the objects in the image that should be segmented. As prompt can be used point, box, mask, or text. SAM decoder combines semantic information from the encoder and prompt information from the prompt encoder and generates for each prompt three masks. SAM decoder is based on a standard transformer decoder, but only 2 layers are used. The main idea of this model is to pre-compute semantic information of an image by slow ViT and then interactively use prompts from the user for generating masks by the fast decoder. The model was trained on a large dataset and is therefore capable of generating masks for a large number of objects.

Even though the model was trained on a huge amount of data, it is not able to segment all objects, an example can be specific segmentation tasks such as segmentation of medical data. One of the options for adapting the model for a specific dataset is to fine-tune the model. This can be computationally challenging due to the large encoder on the input. But the question arises, would it be possible to transform the semantic space of an already trained model on a specific dataset into the semantic space of the SAM encoder and then use fast SAM decoder with the prompt encoder to generate masks?

[1] student of the doctoral degree program Applied Sciences, field of study Cybernetics, e-mail: strakajk@kky.zcu.cz

2 Experimental setup

To address this issue first was necessary to choose an experimental dataset. We chose Plant Phenotyping datasets introduced by Minervini et al. (2016) which contains approximately one thousand photos of plants taken from above. The second step was to train U-Net on this dataset to distinguish between plants and backgrounds. The last step was to replace the SAM encoder with the U-Net encoder. As transformation between semantic spaces was chosen 1x1 convolution layer which was placed behind the U-Net encoder. Then except for the adaptation layer all, parameters of the model were frozen and the model was trained on the dataset. Since the model expects a prompt, the center of the plant ground truth mask was used. The model is shown in Figure 1. Because SAM decoder generates three masks, it is necessary to select only one. During training the mask with the smallest loss was selected and during prediction, the mask with the largest intersection over union (IoU) with ground truth mask was selected.

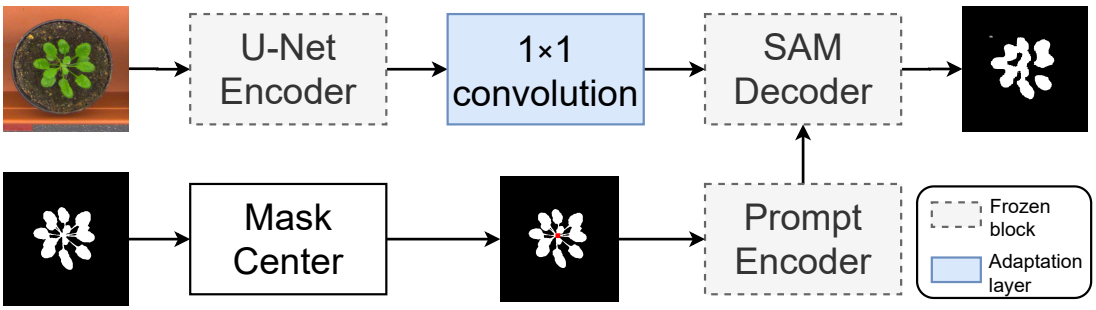


Figure 1: Diagram of the proposed experiment.

3 Results and conclusion

The pre-trained SAM model was initially evaluated on the selected datasets, establishing a baseline performance with an IoU of 0.573. Subsequently, the U-Net model was trained on the same dataset, achieving an IoU of 0.710. Finally, a combined model from the previous two models with an adaptation layer was created. After the training of the adaptation layer model achieved 0.762 IoU.

This result suggests that the transformation between the semantic spaces of the two encoders is possible. And it can be inferred that a sufficiently general decoder for mask generation can be combined with a task-specific encoder using and training only one convolutional layer.

Acknowledgement

The work was supported by the University of West Bohemia, project No. SGS-2022-017.

References

Ronneberger, O., Fischer, P. and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

Minervini, M., Fischbach, A., Scharr, H. and Tsaftaris, S.A., 2016. Finely-grained annotated datasets for image-based plant phenotyping. Pattern recognition letters, 81, pp.80-89.