XAI-driven Model Improvements in Interpretable Image Segmentation

Rokas Gipiškis[1,*]

[1]Vilnius University, Institute of Data Science and Digital Technologies, 4 Akademijos St, Vilnius, Lithuania

Abstract

Semantic image segmentation is the most fine-grained task in computer vision. Its applications range from autonomous vehicles to medical imaging. Despite its deployments in critical areas, interpretable image segmentation remains an underexplored field, especially when compared to explainable AI (XAI) solutions in classification and object detection. Even less attention has been paid to the use of XAI in non-explainability-related scenarios, where XAI methods are applied not for interpretability *per se*, but rather for other instrumental reasons, such as improving a model's performance. Such use cases can potentially extend to AI safety, specifically in the case of adversarial attacks, self-supervised learning, neural architecture search (NAS), and continual learning (CL). Most of these areas have never been investigated in the context of interpretable segmentation. This work outlines key developments in the field of interpretable image segmentation, with a particular focus on XAI-driven model improvements. We also consider potential uses of interpretable image segmentation for model compression in the case of NAS, and instance-based memory compression in the case of CL.

Keywords

Explainable AI, Interpretable AI, Image segmentation, XAI

1. Introduction

Image segmentation is a predominant task in computer vision, with applications ranging from medicine to industry. However, evaluation metrics for deep learning (DL) models do not provide a complete view of their performance. Although a model's performance might be good, it could still focus on undesirable spurious correlations. Furthermore, even if the metric is accurate, it does not offer insights into the internal mechanisms of the model. The need for explainable and trustworthy systems is ever-increasing. Yet, there appears to be a clear gap in the explainable AI (XAI) literature between classification and segmentation.

One could argue that segmentation could be considered a subset or rather an extension of explainable classification. However, this does not resolve the fact that explainable segmentation has its own unique challenges. Firstly, there is the question of how to efficiently generate explanations for segmentation when multiple pixels are involved. Secondly, there is the challenge of how to interpret those explanations when multiple pixels are involved. Another important question, not directly related to enhancements in interpretability, is whether current XAI

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

rokas.gipiskis@mif.vu.lt (R. Gipiškis)

0000-0001-5166-0920 (R. Gipiškis)

techniques in image segmentation can be used to improve model performance, specifically in compression-based approaches in neural architecture search (NAS) and continual learning (CL).

Based on this reasoning, the current doctoral research is framed by the following questions:

1. How do we improve explainable segmentation techniques for better explainability? This question is directly related to the enhancement of explainable segmentation techniques for a better understanding of segmentation models.
2. How do we use explainable segmentation to enhance performance in tasks not directly related to explainability? This question is related to model improvements by indirectly using XAI-based techniques.

By combining the results of these two application areas, we can not only identify the most suitable XAI candidates in image segmentation, but also evaluate their potential use in creating more efficient models, both in terms of their weights and memory utilization. Such efficient solutions could be particularly useful in edge computing and in Internet of Things devices.

2. Related Work

The research into explainable image segmentation began in the late 2010s. Since then, it has seen incremental growth in new methods and applications. For the purposes of this paper, we can divide related works into two groups, based on the end-purpose of XAI application. The first group uses XAI methods to better understand a model's performance or enhance its interpretability. The second group employs XAI techniques not primarily for explainability, but as tools to enhance a model's performance, whether by compressing its weights, improving its memory utilization, or, in the case of self-supervised segmentation, limiting the need for manual annotations.

2.1. Explanations for Image Segmentation

Two influential early applications of XAI in segmentation can be found in [1] and [2]. A perturbation-based explainability approach is introduced in [1] as a method to detect contextual bias. In [2], on the other hand, a gradient-based Seg-Grad-CAM solution is proposed and evaluated using U-Net [3]. Perturbation-based XAI techniques are further investigated in [4] and [5]. In [4], the focus is on occlusions in the input space, while [5] explores gradient-free perturbations in the activation space. Simple gradient and Smoothgrad [6] extensions for segmentation are presented in [7], where they are investigated for applications in cyber-physical systems. These methods are further discussed in Section 5. Currently, gradient-based post-hoc approaches are more prevalent in explainable segmentation due to their lower computational costs.

2.2. XAI-driven Model Improvements

To our knowledge, there are no XAI-driven model improvements specifically for NAS or CL in segmentation. In classification, [8] proposes a NAS model based on class activation mapping (CAM). The teacher and student models are incorporated into the evolutionary search. The less

complex student model has to generate an explanation map that closely approximates the one generated by the teacher model, as measured by the inverse of the Euclidean distance. In [9], an input saliency-based NAS is introduced as a way of reweighing different data points. However, the proposed solution only focuses on the features in the input space, leaving investigation of the activation space features for further research. This approach is suitable for differentiable NAS methods, but further investigation is needed for non-differentiable methods, such as evolutionary-algorithm-based NAS. Additional modifications or selecting a non-gradient based optimization algorithm would be required.

Explainable segmentation techniques have also been investigated for safety and robustness evaluations in segmentation models [7]. XAI methods can be potentially used in detecting adversarial attacks targeting segmentation. However, the study does not investigate architecture-based changes, so that the safety-critical measures could be incorporated into the model itself. Another non-explainability-related area that uses XAI in model training is self-supervised learning as well as weakly-supervised segmentation. Although, after the initial use of classification saliencies for weakly-supervised object localization [10], several other applications have been proposed, the solutions focus on XAI techniques in classification and are not directly related to interpretable segmentation. Typically, explanatory heatmaps for a selected class are used as imprecise segmentation masks that could be employed to increase dataset size, since manual annotations are expensive and time-consuming.

3. Research

3.1. Research Questions

Current research questions encompass XAI use for both explainability-related (the first question) and non-explainability-related tasks (the latter two questions):

1. How can explainability techniques for image segmentation be improved, either in terms of evaluative XAI metrics or computational costs?
2. Can more efficient segmentation models be found by incorporating explainable segmentation techniques in NAS, specifically in the case of teacher-student architectures?
3. Can explainable segmentation techniques be implemented in continual learning for compression in experience replay?

3.2. Hypothesis

The underlying assumption is that efficient explainable segmentation techniques can identify those regions in the input space or those feature maps in the activation space that are most important for the decision-making of a model and, by extension, its accuracy. Since XAI techniques primarily focus on these areas, their results could be used for model compression in NAS, or memory compression in CL.

3.3. Objectives

To better investigate XAI-driven segmentation model enhancements, we define the following objectives:

1. Identify the most suitable XAI techniques in segmentation based on computational requirements and quantitative XAI metrics.
2. Investigate whether the CAM-NAS application in classification can be successfully extended to segmentation.
3. Evaluate the performance of various explainable segmentation techniques, focusing on their potential uses in NAS.
4. Explore the use of explainable segmentation techniques for memory compression in experience replay by storing only the image areas centered around the most important input features, as identified by selected XAI techniques.

4. Approach

Firstly, suitable XAI techniques have to be selected for the experiments. Based on previous research [4], [5], [7], gradient-based methods are preferable for the proposed use-cases due to their lower computational costs. Speed is an important factor when extracting saliency maps, especially when multiple iterations are required. This is further supported by the CAM-NAS [8] experiments in classification, where gradient-based methods achieve the best results. A simple gradient-based saliency map technique can be used as a baseline. Different implementations of Seg-Grad-CAM [2] can also be investigated. Since gradient-based techniques can generate a lot of noise, noise-reduction techniques, such as thresholding a certain percentage of pixels based on their importance, might also be considered. Especially when manual human-in-the-loop supervision is involved.

NAS focuses on automating the design of neural network architectures. Following [8], the initial teacher-student model will be extended to semantic segmentation models. The explanations will be generated based on the summed-up pre-Softmax prediction scores for the selected class of interest (Fig. 1). A well-trained segmentation model (the teacher) will be paired up with a less complex model (the student). Then, explainable segmentation maps will be generated for the same input images and compared in terms of a similarity score. If the teacher model has truly learned the most important representations in an unbiased way, and if the selected XAI technique can capture the most important features for the model's decision-making, then we would like to have a student model that is also sensitive to the same features. This could be viewed as a knowledge transfer from the teacher model to the student. The original CAM-NAS implementation [8] uses evolutionary algorithms for the generation of search submodels, and it could serve as an initial starting point for the experiments.

It is less clear whether XAI-driven model enhancements can be implemented in the case of CL, specifically for memory compression in experience replay. CL focuses on how an already trained model can learn new tasks without forgetting the previous ones. Experience replay is an efficient CL strategy that allows storing the most important examples from old tasks inside the memory so that the model can still be exposed to them in the future. In classification, it is possible to

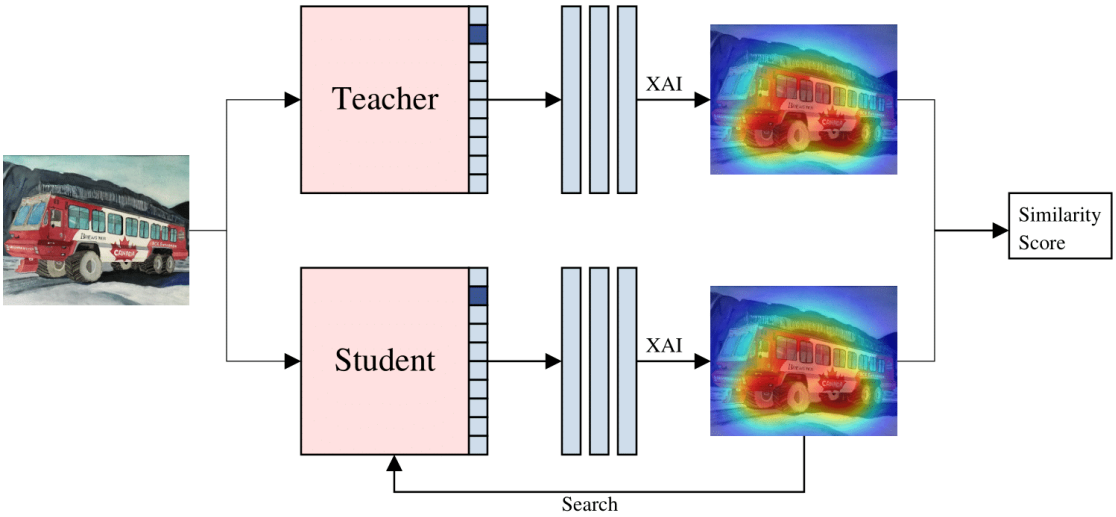

Figure 1: The pipeline for CAM-NAS in segmentation, based on the original implementation [8] for classification tasks. This is an idealized scenario where the saliency maps generated by both models are identical.

reduce memory utilization by storing just the most salient regions of the data samples [11]. By cropping the image so that it is centered around the most important regions, memory can be utilized more efficiently. However, it is unclear if the cropping strategy could work in the case of segmentation, as it is a dense prediction task that, unlike image classification, could not be completed if part of the image was missing. In this particular context, compared to compression in classification, segmentation appears to be more sensitive to partial data. Enough critical contextual information would have to be stored for the segmentation to be successful. Perhaps less salient regions could be downsampled, as described in [12] in the case of classification. Then, enough contextual information could still be preserved to complete segmentation, especially if the right contextual information was identified by the explainable segmentation technique. Following [12], once the most important image regions are identified, they can be occluded by a bounding box. The resulting image with unoccluded non-discriminative pixels is then downsampled. Then, the previously occluded salient region is summed up to the downsampled image. The final image occupies significantly less space in memory. To our knowledge, similar experiments have not yet been conducted for CL in image segmentation.

5. Results

So far, the investigation has focused on explainable segmentation for a better understanding of the model, with an additional focus on XAI safety and robustness [7]. The first comprehensive survey on XAI in image segmentation has been prepared [13], including the initial taxonomy, graphical representations of XAI pipelines for different XAI method categories, and a detailed literature analysis based on evaluation metrics, application domains, and used datasets. Surveyed XAI methods in segmentation have been grouped into gradient-based, perturbation-based,

prototype-based, architecture-based, and counterfactual techniques.

Input perturbation experiments, also known as occlusion sensitivity, have been performed on different segmentation models in [4]. Both the size and color of the occlusion filter have been investigated using deletion curves, a quantitative XAI evaluation metric. The results indicate that segmentation models are sensitive to varying occlusion colors and sizes, and that this is related to the original colors in the unperturbed input image as well as the ratio between the foreground object of interest and its background. It has also been observed that different occlusion colors can greatly affect the segmentation outputs, even when the same filter size is used and when applied to the same input region. More neutral Gaussian-based filters appear to cause less unnatural distortions to the model's output. In contrast, depending on the background and foreground colors, the black occlusion filter can be mistaken as part of the segmented object. Another interesting finding is that compared to perturbation-based explanations in classification, there is significantly less variance generated in the evaluation scores. Therefore, normalization techniques, such as min-max scaling, can be used to generate clearer explanations with higher color intensities. It is also worth noting that input perturbation methods require significant computational costs, as each occlusion requires an additional inference.

Perturbations in the activation space have been investigated in [5]. A new gradient-free Seg-Ablation-CAM technique has been proposed as an extension of [14]. The method is based on partial or full deactivations of activation maps from the selected neural network layer. The results of foreground and background occlusions indicate that foreground occlusions are more important for the model's output. Based on the qualitative results, the proposed approach provides less noisy and more concentrated saliency visualizations compared to gradient-based XAI methods in segmentation. However, just like other perturbation-based methods, it is computationally expensive, as multiple inferences are required.

Gradient-based explainable segmentation methods have been described in [7], where vanilla gradient and its SmoothGrad [6] extension have been investigated for industrial applications. This is the first study to focus on adversarial attacks targeting explainable segmentation. It has been demonstrated that segmentation models can be attacked so that the model's output does not change, but its corresponding explanation is arbitrarily manipulated. This is achieved by introducing a three-term loss function which ensures that the perceptible noise is not introduced in the input, that the output is not significantly changed, and that the generated explanation is close to the targeted one, as specified before the attack.

6. Further Research

The next research steps will focus on the implementation of CAM-NAS for semantic segmentation. It needs to be evaluated whether the proposed strategy is feasible for higher-resolution images. The final contribution could lead to more efficient image segmentation models. Further research directions could explore how to retain both interpretability and NAS efficiency while investigating the potential trade-offs between the two. This might be related to the fact that some computationally expensive XAI techniques, such as Seg-Ablation-CAM [5], are less noisy compared to less computationally demanding techniques, like simple gradients. Other studies in interpretable image segmentation could also investigate whether safety-critical components

can be automatically identified by XAI techniques and then incorporated into the proposed architecture.

Acknowledgments

I extend my gratitude to Prof. Olga Kurasova, my doctoral advisor, and Prof. Chun-Wei Tsai for their support and counsel.

References

[1] L. Hoyer, M. Munoz, P. Katiyar, A. Khoreva, V. Fischer, Grid saliency for context explanations of semantic segmentation, Proceedings of the Advances in Neural Information Processing Systems 32 (2019).

[2] K. Vinogradova, A. Dibrov, G. Myers, Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13943–13944.

[3] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proceedings of the Medical image computing and computer-assisted intervention–MICCAI 2015, 2015, part III 18, 2015, pp. 234–241.

[4] R. Gipiškis, O. Kurasova, Occlusion-based approach for interpretable semantic segmentation, in: Proceedings of the Iberian Conference on Information Systems and Technologies, 2023, pp. 1–6.

[5] R. Gipiškis, D. Chiaro, D. Annunziata, F. Piccialli, Ablation studies in activation maps for explainable semantic segmentation in Industry 4.0, in: Proceedings of the IEEE EUROCON 2023-20th International Conference on Smart Technologies, 2023, pp. 36–41.

[6] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: Removing noise by adding noise, arXiv preprint arXiv:1706.03825 (2017).

[7] R. Gipiškis, D. Chiaro, M. Preziosi, E. Prezioso, F. Piccialli, The impact of adversarial attacks on interpretable semantic segmentation in cyber–physical systems, IEEE Systems Journal (2023).

[8] Z. Zhang, Z. Wang, I. Joe, CAM-NAS: An efficient and interpretable neural architecture search model based on class activation mapping, Applied Sciences 13 (2023) 9686.

[9] R. Hosseini, P. Xie, Saliency-aware neural architecture search, Proceedings of the Advances in Neural Information Processing Systems 35 (2022) 14743–14757.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[11] G. Saha, K. Roy, Online continual learning with saliency-guided experience replay using tiny episodic memory, Machine Vision and Applications 34 (2023) 65.

[12] Z. Luo, Y. Liu, B. Schiele, Q. Sun, Class-incremental exemplar compression for class-incremental learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11371–11380.

[13] R. Gipiškis, C.-W. Tsai, O. Kurasova, Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey, arXiv preprint arXiv:2405.01636 (2024).
[14] S. Desai, H. G. Ramaswamy, Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 983–991.