Interactive Attribution-based Explanations for Image Segmentation

C. Humer, M. Elharty, A. Hinterreiter, and M. Streit

Johannes Kepler University Linz, Austria

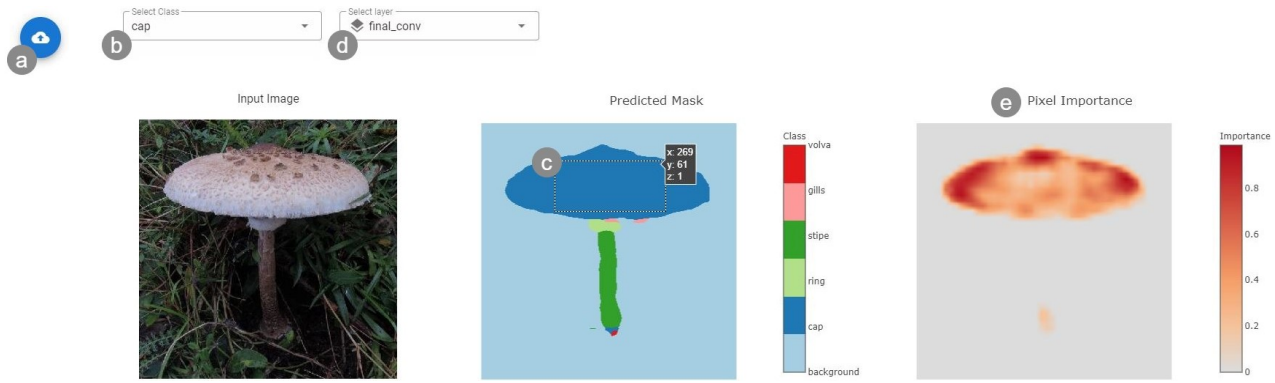

Figure 1: *The web-based application allows users to interactively explore Grad-CAM explanations of segmentation networks. Users can (a) upload images, (b) choose a class for which the explanation is calculated, (c) choose single pixels or areas in the predicted segmentation mask to calculate the explanation, and (d) choose the convolutional layer for which the Grad-CAM is calculated. The (e) pixel importance shows the resulting Grad-CAM as a heatmap, which users can leverage for their analysis. The example of a mushroom segmentation shows that the network focuses mainly on the border pixels of the mushroom cap to segment the cap. Input image adapted from [S*22, Vis18].*

Abstract

Explanations of deep neural networks (DNNs) give users a better understanding of the inner workings and generalizability of a network. While the majority of research focuses on explanations for classification networks, in this work we focus on explainability for image segmentation networks. As a first contribution, we introduce a lightweight framework that allows generalizing certain attribution-based explanations, originally developed for classification networks, to also work for segmentation networks. The second contribution is a web-based tool that utilizes this framework and allows users to interactively explore segmentation networks. We demonstrate the approach using a self-trained mushroom segmentation network.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** • ***Computing methodologies*** → ***Image segmentation;***

1. Introduction

Image classification networks usually deliver one classification result for their input. In contrast, segmentation networks result in one classification for each pixel of the image (i.e., the output is of the form *imgHeight* × *imgWidth* × *nChannels*, where *nChannels* corresponds to the number of classes). Due to this difference in architecture, it is not straightforward to take existing explanation methods—originally created for classification tasks—and apply them to segmentation networks. Previous work by Vinogradova et al. [VDM20] shows how the common explainability method Grad-CAM [SCD*17] can be adapted to also work for segmentation networks. Having attribution-based explanations—usually provided in form of heatmaps—is important for better understanding deep neural networks (DNNs). However, static explanations alone are insufficient for gaining a deeper understanding of the inner workings of such networks. Therefore, our contribution is twofold:

- We propose a lightweight framework that can be used to adapt segmentation networks in a way that makes them resemble classification networks. Many of the existing explanation techniques can then be applied to the adapted segmentation network.
- We present an interactive tool that facilitates the exploration of attribution-based explanations for image segmentation networks.

To demonstrate our approach, we present example outputs of the tool with a self-trained mushroom segmentation network.

2. Related work

Segmentation networks are usually based on fully convolutional networks (FCN) [LSD15], which have been improved over the years to generate finer and more accurate results [NHH15, RFB15, ZSQ*17, GWX*19]. A large body of research addresses attribution-based explanation methods in the domain of image classification [SVZ14, ZF14, BBM*15, SCD*17], which are used to explain the importance of regions in the input image to the prediction. Previous work [HMK*19, VDM20] adapted such attribution-based methods and applied them to image segmentation networks to generate static explanations. Although there is work on interactive exploration of image segmentation networks [FMH16, JSO19], these do not allow the exploration of attribution-based explanations.

3. Methods

This section gives details about our approach to adapt segmentation networks and how we use this approach in an interactive tool for exploring the generated explanations. A simple use case that analyzes the architecture of a segmentation network trained on mushroom images can be found in the supplementary materials.

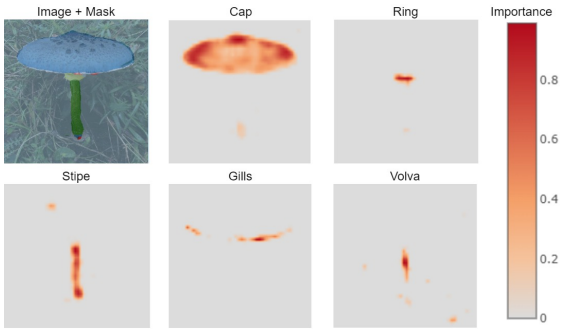

Figure 2: *Sample image overlaid with its predicted mask (top left) and corresponding Grad-CAM explanations for each class of a mushroom segmentation. All pixels of the corresponding class were averaged and the resulting value was used for calculating the Grad-CAM. Input image adapted from [S*22, Vis18].*

3.1. Segmentation network adaptation

Vinogradova et al. [VDM20] presented a method that enhances Grad-CAM explanations to also work for segmentation networks. We propose a more general approach that does not adapt a specific explanation technique, but instead transforms a trained segmentation network to resemble the architecture of a classification network. We believe that with this approach many image classification explanation techniques can be also used for image segmentation networks without the need of adapting them. To bring a segmentation network in a form that resembles a classification network, the output shape of the network has to be reduced to give only one classification result instead of one result per pixel.
To achieve this, we could just use a global average pooling layer that calculates the average value for each channel, which carries the segmentation result for one particular class. This new network already resembles the architecture of a classification network, and could already be used with existing explanation techniques. However, with this approach users are not able to specify which part of the segmentation output they want to investigate. We address this limitation by adding a custom layer that can be applied on top of the final segmentation layer of a network. This custom layer takes a binary mask as parameter, which indicates the regions of interest, and multiplies it with the output of the segmentation network. We implemented an example of this approach with TensorFlow [AAB*15] available at `https://github.com/ginihumer/segmentation-explanation-adapter`. Using TensorFlow's *multiply* and *global average pooling* functions, we manage to maintain a derivable model that can also be used for gradient-based explanation methods. A comparison of results using our approach versus the approach proposed by Vinogradova et al. [VDM20] can be found in the supplementary materials.

3.2. Interactive tool

We present an interactive web application that builds on top of the framework introduced in Section 3.1. The tool allows users to upload images (see Figure 1a) showing their predicted segmentation mask and the corresponding Grad-CAM explanation (see Figure 1e). Users are then able to select either a whole class (see Figure 1b), which automatically creates a binary mask that activates all pixels predicted as this class as shown in Figure 2, or users can directly select pixels (see Figure 1c) in the visualization that shows the segmentation mask. For segmenting the mushroom cap, the network seems to mainly look at the edge-pixels of the cap, as can be seen in Figure 2. Interestingly, although there are no gills visible in the image, the network seems to expect them to occur directly under the cap. For the volva, the explanation suggests that the network mainly locates it by identifying the stem. As a final interaction, users can select the convolutional layer (see Figure 1d) of the segmentation network, for which the Grad-CAM is calculated. This enables users to also explore the architecture of the network (see supplementary materials for an example).

4. Limitations and future work

We believe that the tool is a valuable addition to current research on image segmentation explanations. However, so far we tried our proposed approach with Grad-CAM only. Other explanation methods—originally created for classification—have to be investigated and it has to be verified that they work as expected with the proposed approach. To increase the usefulness of our tool, we plan to integrate the option to upload custom models. Furthermore, the tool is currently limited to TensorFlow models. Potential areas of future work in this context include enabling the application of models from other machine learning frameworks. Additionally, user studies could be conducted to validate our approach over various use cases.

5. Acknowledgments

This work was supported in part by the State of Upper Austria and the Austrian Federal Ministry of Education, Science and Research via the LIT – Linz Institute of Technology (LIT-2019-7-SEE-117), and the Austrian Science Fund (FWF DFH 23–N). We thank Patrick Adelberger, Klaus Eckelt, and Christian Alexander Steinparz for providing feedback on the abstract text.

References

[AAB*15] ABADI M., AGARWAL A., BARHAM P., BREVDO E., CHEN Z., CITRO C., CORRADO G. S., DAVIS A., DEAN J., DEVIN M., GHEMAWAT S., GOODFELLOW I., HARP A., IRVING G., ISARD M., JIA Y., JOZEFOWICZ R., KAISER L., KUDLUR M., LEVENBERG J., MANÉ D., MONGA R., MOORE S., MURRAY D., OLAH C., SCHUSTER M., SHLENS J., STEINER B., SUTSKEVER I., TALWAR K., TUCKER P., VANHOUCKE V., VASUDEVAN V., VIÉGAS F., VINYALS O., WARDEN P., WATTENBERG M., WICKE M., YU Y., ZHENG X.: TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. URL: https://www.tensorflow.org/. 2

[BBM*15] BACH S., BINDER A., MONTAVON G., KLAUSCHEN F., MÜLLER K.-R., SAMEK W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One 10*, 7 (2015), e0130140. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4498753/, doi:10.1371/journal.pone.0130140. 2

[FMH16] FRÖHLER B., MÖLLER T., HEINZL C.: GEMSe: Visualization-Guided Exploration of Multi-channel Segmentation Algorithms. *Computer Graphics Forum 35*, 3 (2016), 191–200. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12895, doi:10.1111/cgf.12895. 2

[GWX*19] GENG L., WANG J., XIAO Z., TONG J., ZHANG F., WU J.: Encoder-decoder with dense dilated spatial pyramid pooling for prostate MR images segmentation. *Computer Assisted Surgery 24*, sup2 (2019), 13–19. URL: https://doi.org/10.1080/24699322.2019.1649069, doi:10.1080/24699322.2019.1649069. 2

[HMK*19] HOYER L., MUNOZ M., KATIYAR P., KHOREVA A., FISCHER V.: Grid Saliency for Context Explanations of Semantic Segmentation. In *Advances in Neural Information Processing Systems* (2019), vol. 32, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/hash/6950aa02ae8613af620668146dd11840-Abstract.html. 2

[JSO19] JANIK A., SANKARAN K., ORTIZ A.: *Interpreting Black-Box Semantic Segmentation Models in Remote Sensing Applications*. The Eurographics Association, 2019. URL: https://diglib.eg.org:443/xmlui/handle/10.2312/mlvis20191158, doi:10.2312/mlvis.20191158. 2

[LSD15] LONG J., SHELHAMER E., DARRELL T.: Fully Convolutional Networks for Semantic Segmentation. *arXiv:1411.4038 [cs]* (2015). URL: http://arxiv.org/abs/1411.4038. 2

[NHH15] NOH H., HONG S., HAN B.: Learning Deconvolution Network for Semantic Segmentation. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1520–1528. doi:10.1109/ICCV.2015.178. 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (2015). URL: http://arxiv.org/abs/1505.04597. 2

[S*22] STATENS NATURHISTORISKE MUSEUM, ET AL.: Danmarks svampeatlas, 2022. Accessed March 29, 2022. URL: https://svampe.databasen.org/. 1, 2

[SCD*17] SELVARAJU R. R., COGSWELL M., DAS A., VEDANTAM R., PARIKH D., BATRA D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 618–626. doi:10.1109/ICCV.2017.74. 1, 2

[SVZ14] SIMONYAN K., VEDALDI A., ZISSERMAN A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]* (2014). URL: http://arxiv.org/abs/1312.6034. 2

[VDM20] VINOGRADOVA K., DIBROV A., MYERS G.: Towards Interpretable Semantic Segmentation via Gradient-Weighted Class Activation Mapping (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 10 (2020), 13943–13944. URL: https://ojs.aaai.org/index.php/AAAI/article/view/7244, doi:10.1609/aaai.v34i10.7244. 1, 2

[Vis18] VISIPEDIA: 2018 FGCVx fungi classification challenge, 2018. Accessed March 29, 2022. URL: https://www.kaggle.com/competitions/fungi-challenge-fgvc-2018/. 1, 2

[ZF14] ZEILER M. D., FERGUS R.: Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Lecture Notes in Computer Science, Springer International Publishing, pp. 818–833. doi:10.1007/978-3-319-10590-1_53. 2

[ZSQ*17] ZHAO H., SHI J., QI X., WANG X., JIA J.: Pyramid Scene Parsing Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 6230–6239. doi:10.1109/CVPR.2017.660. 2