Interpretability of Image Segmentation Models

Degree programme: BSc in Computer Science | Specialisation: Computer Perception and Virtual Reality
Thesis advisor: Peter von Niederhäusern
Expert: Prof. Dr. Andreas Spichiger (BFH Wirtschaft)
Industrial partner: University of Bern, Bern

Modern neural networks are achieving remarkable performance in many fields. In comparison with classic machine learning techniques, it is much harder to explain how these neural networks came to their conclusions, because they use millions of trained parameters. Interpretability methods for image classification tasks are already well established. The goal of this thesis is to provide new interpretability methods for image segmentation problems.

Interpretability

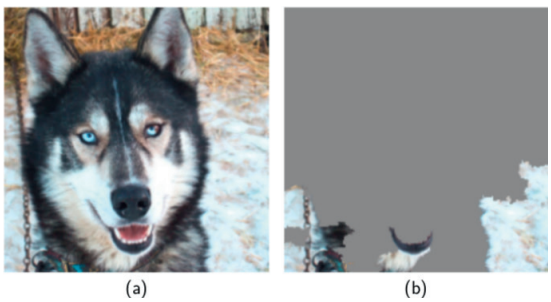
Figure 1a) shows a misclassification of a Husky as a wolf. Figure 1b) shows the output of the interpretability method LIME (Riberio et al.). The output explains that the neural network classified the Husky as a wolf because the background of the image shows snow. The neural network learned the fact that most images of wolfs contain snow in the background and that this fact is the best clue that an image shows a wolf.

Goals

For classification problems many interpretability methods such as LIME already exist and work well. The main goal of the thesis is to build interpretability methods for image segmentation models. Image segmentation is different from classification in the way that the algorithms do not detect what is visible on a picture, e.g. a Husky, but instead mark a region in an image where they think something is visible, e.g. a tumor.

Hausdorff Distance Masks

One of the developed methods we call Hausdorff Distance Masks (HDM). In essence, it works by first occluding parts of the image. Next, the segment generated by the neural network from the modified image is compared with the segment generated from the unmodified original image. If the two segments differ by a wide margin, the occluded region holds important information for the generation of the segment.



(a) (b)

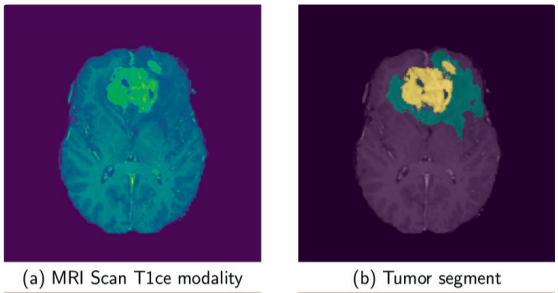
Fig. 1: a) Husky missclassified as a wolf, b) LIME method showing the importance of snow for the (mis)classification.

Example

An MRI scan produces four different images per slice, called modalities. Figure 2a) below shows the T1 contrast enhanced modality. Figure 2b) shows the tumor segment generated from the four images. Figures 2c) and 2d) show the results of the Hausdorff Distance Masks method on two modalities of an MRI scan. The method shows which parts of the image are important for the generation of the tumor segmentation.

Conclusion

Two methods have been successfully implemented: The new Hausdorff Distance Masks method discussed above and the method RISE (Vitali Petsiuk et al.), which has been modified to work on image segmentation tasks. Other methods were also evaluated but did not produce good results. Especially the HDM method looks promising and can be used by researchers to assess and enhance their neural networks.



(a) MRI Scan T1ce modality (b) Tumor segment

(c) Important parts of the T1ce modality generated with the HDM method (d) Important parts of the FLAIR modality generated with the HDM method

Fig. 2: a) T1ce modality, b) tumor segment, c) and d) two modalities interpreted with Hausdorff Distance Masks.

Fabio Anderegg
fabioanderegg@msn.com