Validation of ML Models from the Field of XAI for Computer Vision in Autonomous Driving

Antonio Mastroianni[1] and Sibylle D. Sager-Müller[1,*]

[1] *Lucerne University of Applied Sciences and Arts (Hochschule Luzern), Suurstoffi 1, 6343 Rotkreuz, Switzerland*

Abstract

We describe Explainable Artificial Intelligence (XAI) methods for image segmentation for autonomous driving applications. The analysis is conducted using metrics such as efficiency, robustness, localization, and complexity. Four XAI methods, namely Gradient-weighted Class Activation Mapping (GradCAM), Local Interpretable Model Agnostic Explanations (LIME), Feature Ablation, and Saliency are applied and assessed on a dataset of street images.

Keywords

Validation, Metrics, Captum, XAI Methods, Segmentation, Computer Vision

1. Introduction

Computer vision through machine learning is a relevant topic in the field of autonomous driving. Within the field of computer vision, there are various tasks such as image classification, object detection, semantic segmentation, and instance segmentation. The focus of this paper lies in the evaluation of Explainable Artificial Intelligence (XAI) methods on segmentation models. Because we want to use the segmentation models for autonomous driving in the future, we focused on images showing street scenes. The current model quality indicates that there is a lot of further work necessary before their implementation in a real-world application.

The methods refer to algorithms that are included in the Captum package [1]. For the evaluation of XAI methods, four metrics were used, three of which originate from the categories of complexity, robustness, and localization as listed in Quantus [2]. Complexity is based on sparseness, robustness on average sensitivity, and localization on Relevance Rank Accuracy (RRA) and accordingly the False Positive Rate of RRA. The fourth metric, efficiency, was measured based on runtime.

To apply and evaluate XAI methods, it is necessary to have images and a model capable of segmenting images. A publicly available dataset that provides street images with ground

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

*Corresponding author

antonio.mastroianni@stud.hslu.ch (A. Mastroianni), sibylle.sager@hslu.ch (S. Sager-Müller)

https://orcid.org/0009-0000-3057-429X (A. Mastroianni), https://orcid.org/0009-0000-4857-5514 (S. Sager-Müller)

truth masks is Cityscapes [3]. Images from Zurich, Switzerland, were selected from this dataset. For the segmentation, pre-trained models from PyTorch were used [4].

2. Segmentation Models

To evaluate the segmentation of street elements, the pre-trained models DeeplabV3 and the Fully Convolutional Network (FCN) both with ResNet101 architecture from PyTorch were analyzed. These models were trained using part of the COCO val2017 dataset and the 20 segmentation classes from Pascal VOC. The decision to use pre-trained models was because these 20 classes include relevant street elements such as bicycles, buses, cars, motorbikes, people, and trains. The models were selected due to their high average Intersection over Union (IoU) scores, which are listed on the PyTorch website [4].

For the evaluation of the segmentation models, two metrics were employed: the IoU value and the Matching Area. The evaluation, shown in Table 1, involved calculating the average values of 122 images from Zurich for both metrics.

Table 1
Segmentation Performance Comparison of Deeplabv3 ResNet101 and FCN ResNet101 Models Across Road-Related Classes

Model	Class	Matching Area	IoU
Deeplabv3 ResNet101	Bicycle	0.084	0.052
	Bus	0.018	0.017
	Car	0.735	0.633
	Motorbike	0.032	0.025
	Person	0.360	0.232
	Train	0.009	0.008
FCN ResNet101	Bicycle	0.093	0.059
	Bus	0.020	0.018
	Car	0.738	0.628
	Motorbike	0.043	0.030
	Person	0.331	0.236
	Train	0.023	0.022

A conclusion drawn from the results of this table is that both the Deeplabv3 and FCN models demonstrate similar performances in segmenting road-related classes. Specifically, the "Car" and "Person" classes tend to exhibit the best results in segmentation for both models. Therefore, only these two classes will be used for further evaluations.

3. XAI Methods

Various XAI methods can be utilized. These methods are classified into two categories: model-specific and model-agnostic. Model-specific methods analyze how changes in the input features change the output, whereas model-agnostic methods work by manipulating input data and analyzing the respective model predictions. Within the subclasses of specific and agnostic, a further differentiation is made based on whether the method is local or global. Additionally, local methods explain the individual predictions of models, while global

methods explain the behavior of the model [5]. In this evaluation, a total of four XAI methods are applied. The selection of methods was partially based on Munn and Pitman [6]. The authors dedicated a chapter in their book to the topic of explainability for image data. In this chapter, the methods, GradCAM and LIME, were introduced, which was one reason for choosing these two methods. Two further methods were selected: Feature Ablation and Saliency.

Gradient-weighted Class Activation Mapping (GradCAM) [7] is a technique that analyzes gradient information for any convolutional layer of a model and generates a heatmap that highlights important regions in the image.

Local Interpretable Model Agnostic Explanations (LIME) [8] involves multiple iterations of removing specific regions of an image to determine which specific areas are more or less important. It is a model-agnostic, perturbation-based method.

Feature Ablation [9] is also a perturbation-based method. It calculates the difference of the attribution in the model output of each feature when it is active and when it is replaced. Multiple features can be turned off together instead of one at a time.

Saliency [10] is a method that follows the gradient of a class through the model using backpropagation. During this process, each pixel is minimally changed, and the resulting changes are observed in the prediction of the class score.

In the implementation of XAI methods, an image representing a street scene was selected. The following two segmentation classes were examined: cars and persons. The calculated attribution values from the methods were normalized and a heatmap was generated and overlaid on the original image, as shown in Figure 1. While the original image had dimensions of 2048 x 1024 pixels, it was resized to 512 x 256 pixels to enhance computational speed.

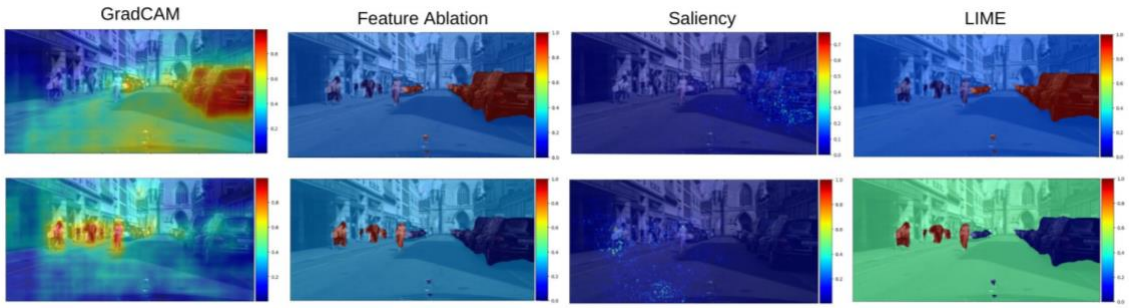

Figure 1: The Heatmaps from XAI Methods GradCAM, Feature Ablation, Saliency, and LIME Superimposed on the Original Street Images. The Top Row Shows the "Car" Segments while the Bottom Row Shows the "Person" Segments.

4. Metrics

Hedström et al. [2] compare various XAI libraries of evaluation metrics for XAI methods. Metrics were divided into six categories: Faithfulness, Robustness, Localisation, Complexity, Axiomatic and Randomisation. Three metrics based on these categories were used to evaluate XAI methods and an additional efficiency metric was employed. The selection of

the following metrics was guided by the idea to test and understand metrics from different categories but else chosen somewhat arbitrary.

To measure the efficiency of a method, the **runtime** was recorded. The time taken to compute the attributions for the "Car" and "Person" classes was measured.

To calculate robustness, the **average sensitivity** [11] was applied. Zhou et al. [12] describe that models usually do not adapt well to new environments when new factors such as weather or illumination conditions are introduced. For this reason, the original images were modified to various brightness levels while preserving the objects in the image. For calculating the average sensitivity, the following formula (1) was used:

$$AvgSensitivity = \frac{1}{N}\Sigma_{1=1}^{N}|AvgAttribution\ orig\ Image - AvgAttribution\ mod\ Image| \quad (1)$$

To obtain comparable values, the average sensitivity was calculated using the same method for the "Car" and "Person" classes. A low value of average sensitivity indicates good robustness, where 0 represents the lowest and 1 the highest possible value. Figure 2 illustrates an example of how the average attribution may appear at different brightness levels.

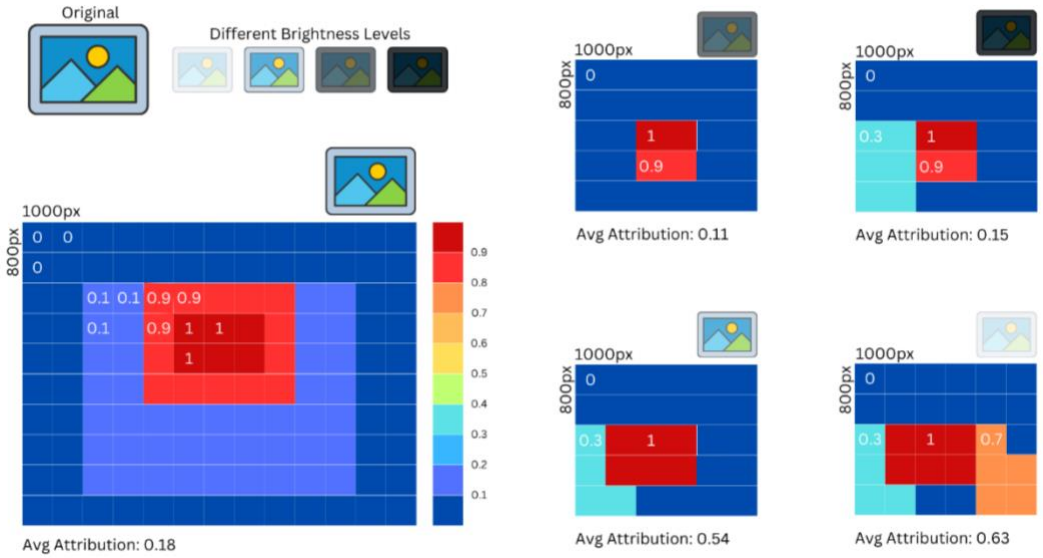

Figure 2: Example with average attribution at different brightness levels. Robustness is ensured when the average attribution remains equal to or only slightly deviates from the original.

Next, we consider the metric **Relevance Rank Accuracy** (RRA) [13], which falls in the category of Localization. This metric measures how many of the highly attributed pixels are located within the ground truth mask. In our evaluation, highly attributed pixels are defined as those falling within the top 20% range. This means when attributions range from 0 to 1, values from 0.8 to 1 are marked as highly attributed. An example of calculating the RRA is illustrated in Figure 3. Not only is the RRA an important indicator, but also the False Positive

(FP) Rate of the RRA. A method might represent all pixels in the image as highly important, resulting in an RRA of 1, which is the highest possible value. Therefore, to achieve more comparable results, the FP RRA was additionally examined.

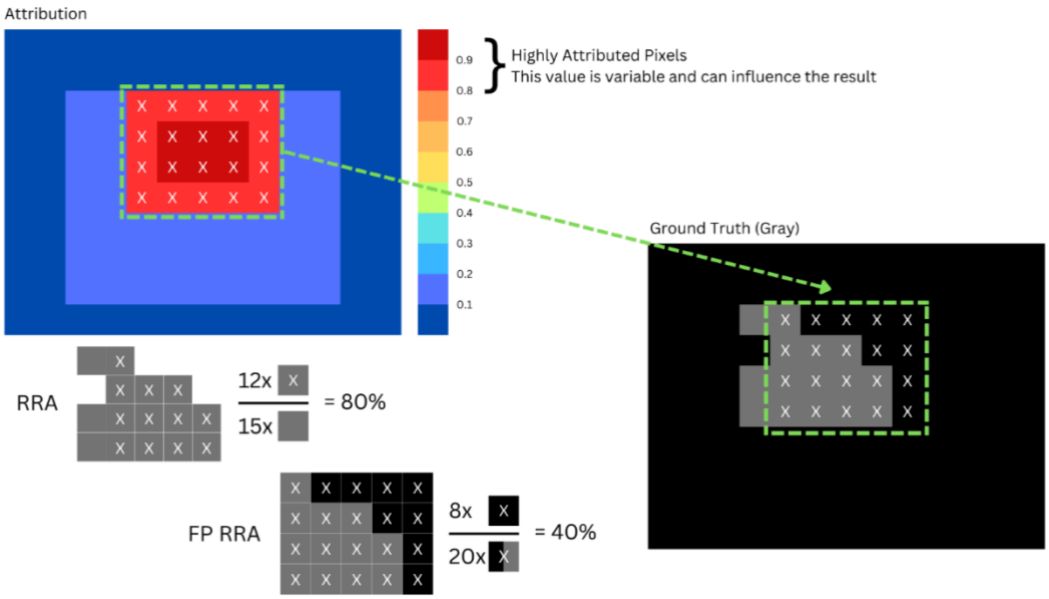

Figure 3: Example of calculating the RRA and FP RRA with the ground truth mask.

The **sparseness** [14] was measured as the last metric, falling in the category of Complexity. Here, the Gini index (G) is used for calculation. In our case, G indicates how scattered or concentrated the attributions are distributed in an image. A value of 0 (the smallest value) indicates a large dispersion, where each pixel is crucial for segmentation. A value of 1 (the largest value) indicates that the important attributions are concentrated. For calculating the Gini index, the following formula (2) was used [15]:

$$G = \frac{2\sum_{i=1}^{n} ix(i)}{n\sum_{i=1}^{n} x(i)} - \frac{n+1}{n} \qquad (2)$$

Let i be the indexing for the position of an attribution in an array, $x(i)$ be the value of the attribution at position i, and n be the number of pixels in the image. The higher the Gini index, the better. A lower value indicates that the importance of all pixels is equal. Our images consist of different objects and only the pixels representing the object should be marked as important, which favors a higher G value.

5. Evaluation of XAI methods

In the evaluation, the metrics of average sensitivity, RRA and sparseness were calculated for the targets Car and Person. The target is a parameter that can be selected in every

method. The first number in Table 2 represents the Car class, while the second number represents the Person class.

Table 2
Evaluation of the methods GradCAM, Feature Ablation, Saliency and LIME on the metrics. The arrow direction indicates whether higher or lower results are considered better.

	GradCAM	Feature Ablation	Saliency	LIME
Runtime [sec] (↓)	**4**	117	18	144
Avg-Sensitivity (↓)	0.043/0.022	0.199/0.351	**0.005/0.005**	0.194/0.389
RRA (↑)	0.472/0.287	**0.929/0.640**	0.001/0.003	**0.929/0.640**
FP RRA (↓)	**0.026/0.096**	0.079/0.334	0.125/0.167	0.079/0.334
Max(0,RRA – FP RRA) (↑)	0.446/0.191	**0.850/0.306**	0.000/0.000	**0.850/0.306**
Sparseness (↑)	0.302/0.242	0.258/0.159	**0.793/0.722**	0.327/0.128

In terms of efficiency, GradCAM clearly outperforms all other XAI methods, followed by Saliency, then Feature Ablation and LIME. In terms of robustness, measured by average sensitivity, both GradCAM and Saliency show remarkable performance. In particular, Saliency shows almost no change in attribution values despite brightness variations. In the RRA and FP RRA category, Feature Ablation and LIME using Max(0,RRA - FP RRA) show identical top performance. The performance of Saliency in this category is considered significantly off target. The low, almost zero RRA means that Saliency only finds very few highly attributed pixels, present in the ground truth mask. In terms of the sparseness metric, the Saliency method emerges as the top performer. The other three methods show similar values that are significantly different from that of Saliency. Furthermore, it is observed that LIME and Feature Ablation exhibit similar or even identical values. The similarity may arise from the fact that both LIME and Feature Ablation are perturbation-based methods.

Additionally, it is notable that in most cases, the values from the category "Car" are better than those from the category "Person." One possible reason for the lower performance values in the "Person" class could be the downsizing of the image, as described in Chapter 3. Table 3 shows how the values of the Matching Area, which indicates how well the segmentation aligns with the ground truth mask, get worse when the image is resized. Another reason could be the general size of the objects in the image. A car, in comparison, has a significantly larger area than a person and is therefore easier to detect.

Table 3
Effects of Image Resizing on Segmentation Accuracy of Class “Person”

Matching Area	2048 x 1024	1024 x 512	512 x 256
Class “Person”	0.756	0.745	0.710

In Figure 4, the measured metrics are visualized in a radar chart with best scores on the outer circle and worse scores in the inner area.

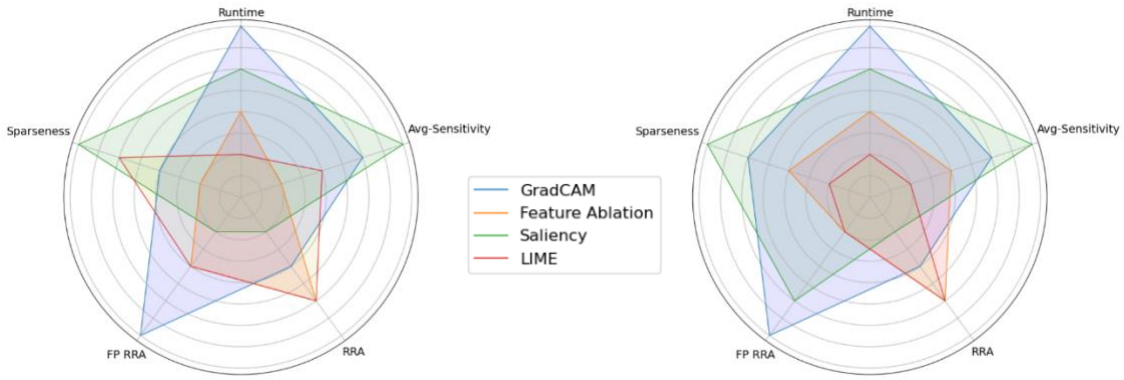

Figure 4: Spider plot comparing evaluation metrics of methods (Car left and Person right)

To evaluate the methods by the used metrics, we need to distinguish between their importance. Two of the most relevant metrics are shown in the lower part of the spider plot, namely the metrics RRA and FP RRA. They are a central aspect in the human interpretability, as they indicate in the heatmap (Figure 1) what is considered important. Due to the bad interpretability, we decide not to pursue Saliency further. Setting Saliency aside as a method, the three remaining methods can be ordered according to their performance: GradCAM demonstrates superiority, followed by Feature Ablation and LIME.

6. Conclusion

In summary, GradCAM emerges in this study as the first choice for evaluating XAI methods for image segmentation, not only because of its favorable metrics but also because of its interpretability in the heatmap. Figure 1 illustrates how GradCAM considers not only the objects themselves but also contextual features such as the road. The heatmap shows a gradient from orange (important) to red (very important), which is missing in other methods. The situation for Saliency was different: While it performed well in the metrics, its heatmap provided almost no information, making interpretation difficult. This highlights the importance of understanding and selecting a comprehensive set of metrics that can provide a clear understanding of the reliability of the methods. We acknowledge that the results are specifically tailored to this case and cannot be directly applied to other fields, such as medicine. Nevertheless, they represent a crucial first step toward future autonomous driving applications.

The current study is to the best of our knowledge the first one to evaluate a subset of XAI methods for image segmentation, an application area of computer vision that has not received as much attention as, e.g., image classification. Future research should involve a variety of data sets and segmentation models and prioritize the evaluation of a broader range of XAI methods for image segmentation. This research should include the development and application of various evaluation metrics, with a particular focus on interpretability. By optimizing these evaluation criteria, an understanding of the specific strengths and weaknesses of different XAI methods can be achieved, which eventually leads to the identification of the most suitable methods for specific applications.

References

[1] ‘Algorithm Descriptions’, Captum. Accessed: Nov. 25, 2023. [Online]. Available: https://captum.ai/docs/attribution_algorithms

[2] A. Hedström *et al.*, ‘Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond’. arXiv, Apr. 27, 2023. Accessed: Mar. 20, 2024. [Online]. Available: http://arxiv.org/abs/2202.06861

[3] ‘Dataset Overview’, Cityscapes Dataset. Accessed: Dec. 15, 2023. [Online]. Available: https://www.cityscapes-dataset.com/dataset-overview/

[4] ‘Models and pre-trained weights’, PyTorch. Accessed: Nov. 28, 2023. [Online]. Available: https://pytorch.org/vision/stable/models.html#semantic-segmentation

[5] C. Molnar, G. Casalicchio, and B. Bischl, ‘Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges’, vol. 1323, 2020, pp. 417–431. doi: 10.1007/978-3-030-65965-3_28.

[6] M. Munn and D. Pitman, *Explainable AI for practitioners: designing and implementing explainable ML solutions*. Beijing, Sebastopol, CA: O’Reilly, 2022.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization’, *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. arXiv, Feb. 26, 2016. Accessed: May 13, 2024. [Online]. Available: http://arxiv.org/abs/1602.04938.

[9] A. A. Ismail, M. Gunady, H. C. Bravo, and S. Feizi, ‘Benchmarking Deep Learning Interpretability in Time Series Predictions’. arXiv, Oct. 26, 2020. Accessed: Mar. 20, 2024. [Online]. Available: http://arxiv.org/abs/2010.13924

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, ‘Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps’. arXiv, Apr. 19, 2014. Accessed: Mar. 20, 2024. [Online]. Available: http://arxiv.org/abs/1312.6034

[11] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, ‘On the (In)fidelity and Sensitivity for Explanations’. arXiv, Nov. 03, 2019. Accessed: Mar. 20, 2024. [Online]. Available: http://arxiv.org/abs/1901.09392

[12] W. Zhou, J. S. Berrio, S. Worrall, and E. Nebot, ‘Automated Evaluation of Semantic Segmentation Robustness for Autonomous Driving’, *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1951–1963, May 2020, doi: 10.1109/TITS.2019.2909066.

[13] L. Arras, A. Osman, and W. Samek, ‘Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI’, *Inf. Fusion*, vol. 81, pp. 14–40, May 2022, doi: 10.1016/j.inffus.2021.11.008.

[14] P. Chalasani, J. Chen, A. R. Chowdhury, S. Jha, and X. Wu, ‘Concise Explanations of Neural Networks using Adversarial Training’. arXiv, Jul. 04, 2020. Accessed: Mar. 20, 2024. [Online]. Available: http://arxiv.org/abs/1810.06583

[15] ‘Gini Koeffizient Definition und Berechnung’, Studyflix. Accessed: Mar. 31, 2024. [Online]. Available: https://studyflix.de/wirtschaft/gini-koeffizient-898