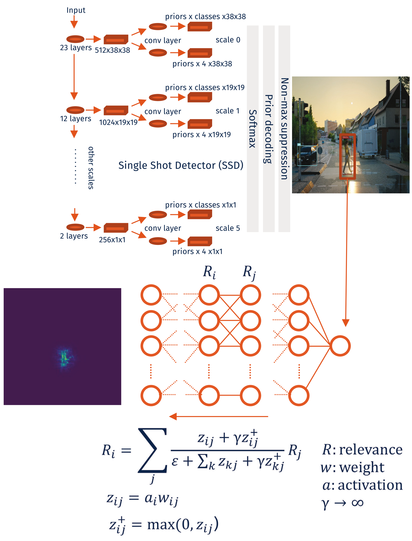
KI ABSICHERUNG
Safe AI for Automated Driving

Extension of Deep Taylor Decomposition to Object Detection

Firas Mualla

DNN specific Safety Concern: Incomprehensible Behavior



$$R_i = \sum_j \frac{z_{ij} + \gamma z_{ij}^+}{\varepsilon + \sum_k z_{kj} + \gamma z_{kj}^+} R_j$$

$$z_{ij} = a_i w_{ij}$$

$$z_{ij}^+ = \max(0, z_{ij})$$

R: relevance
w: weight
a: activation
$\gamma \to \infty$

- Deep Taylor Decomposition (DTD) is a heat map method inspired by decomposing a function value (e. g. class probability) as a sum of feature (e .g. pixel intensity) contributions based on Taylor series.
- We extended the concept from classification [1] to object detection (SSD).
- As there is no ground truth, the evaluation was done based on temporal stability analysis. The correlation coefficient between a heatmap bounding box (bbox) of an object at frame t and the heatmap bbox of the same object at frame t+1 is computed and then averaged over time frames and objects. The result for dynamic sequences of Tranche 5 is 0.63.
- In principle, the DTD can be applied to both label score and bbox's center and/or size. However, due to the prior-based design of SSD, the label information implicitly conveys (rough) object localization.
- In order to investigate the method's potential for adversarial defense, we applied the fast gradient sign method (FGSM) to attack the SSD, so that the bbox's class is flipped to another class. We then generated DTD heatmaps for both benign and adversarial bboxes.
- As shown in the figure, the heatmap energy distribution of the benign images tends to be different from the heatmap energy distribution of the adversarial images. There is, however, no easily describable distinctive visual appearance that characterizes adversarial heatmaps.

Safety Hypothesis:
The method addresses the safety concern incomprehensible behavior. It delivers some insights, as to which pixels contribute to the model's decision. Compared to other heatmap methods, it also claims kind of theoretical soundness based on Taylor decomposition.

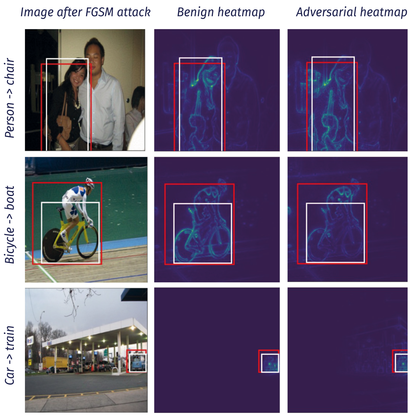
Track the non-max suppression output to a prior index

Find the relevant scale

Accordingly, define the subnet which is involved in the detection

Convert the subnet to a list of layers and activations

Apply DTD: For relu networks, DTD is equivalent to Layerwise Relevance Propagation (LRP) γ-rule when γ approaches infinity



Columns from left to right: 1) Attacked image, on which a benign bbox is marked white and an adversarial bbox is marked red. Adversarial noise energy is lower than 5%. 2) DTD applied to the benign bbox. 3) DTD applied to the adversarial bbox.

References:

[1] Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition, 65:211 –222, 2017.

For more information contact:
firas.mualla@zf.com

Image sources:
Fig1: ZF Friedrichshafen AG, input image from KIA dataset
Fig2: ZF Friedrichshafen AG, input images from Pascal VOC

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag