

Explainable AI for Computer Vision

Robert Haase

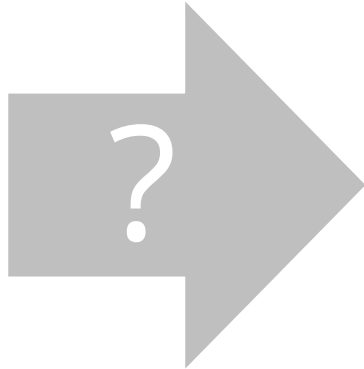


These slides can be reused under the terms of the [CC-BY4.0](https://creativecommons.org/licenses/by/4.0/) license.

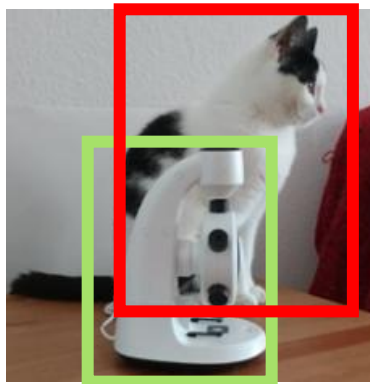
<https://doi.org/10.5281/zenodo.14996127>

Image Classification

Quiz: What could be the result of image classification?



Cat: 30%
Microscope: 20%
Dog: 5%
Car: 0%
...



"The picture shows a cat sitting next to a microscope."

Pixel Classification

Quiz: How is this task calledie bezeichnet man diesen Task?



Combinatorial
Segmentation



Semantic
Segmentation



Instance-
Segmentation



Connected-Component
Segmentation



Explainable Artificial Intelligence (XAI)

- Active field of research
- Key Points: Giving people the opportunity to
 - develop trust in AI systems,
 - interact effectively with AI systems, and
 - predict the outcomes of AI systems.

Explainable Artificial Intelligence (XAI)

What is explainable?

AI Algorithm

Decision-making in an
AI system

Contribution of
underlying data

Meaning of underlying
data

How is it explainable?

Reading+understanding of
equations and source code

Visualization of
intermediate results

Measurement of relation
between input and output data

Explainability

A coherent chain of argumentation that depicts a fact or an algorithm in a completely transparent way.

Intrinsically explainable AI-algorithms

- Example: Linear Regression

$$f(x_1, x_2) = w_1x_1 + w_2x_2$$

If w_1 is much larger than w_2 , the overall result is highly dependent on x_1 . The influence of x_2 is rather small.

Model
explainable

Results
predictable

Explainability

A coherent chain of argumentation that depicts a fact or an algorithm in a completely transparent way.

Intrinsically explainable AI-algorithms

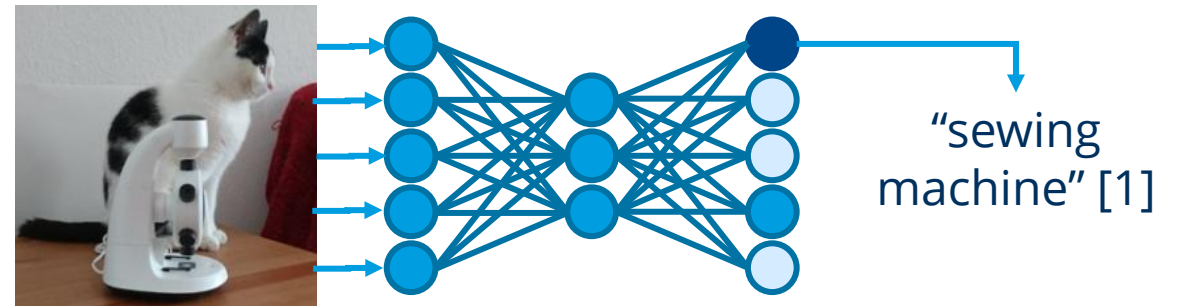
- Example: Linear Regression

$$f(x_1, x_2) = w_1x_1 + w_2x_2$$

If w_1 is much larger than w_2 , the overall result is highly dependent on x_1 . The influence of x_2 is rather small.

Black-Box AI-Algorithms

- Example : Deep Neural Networks (DNN)



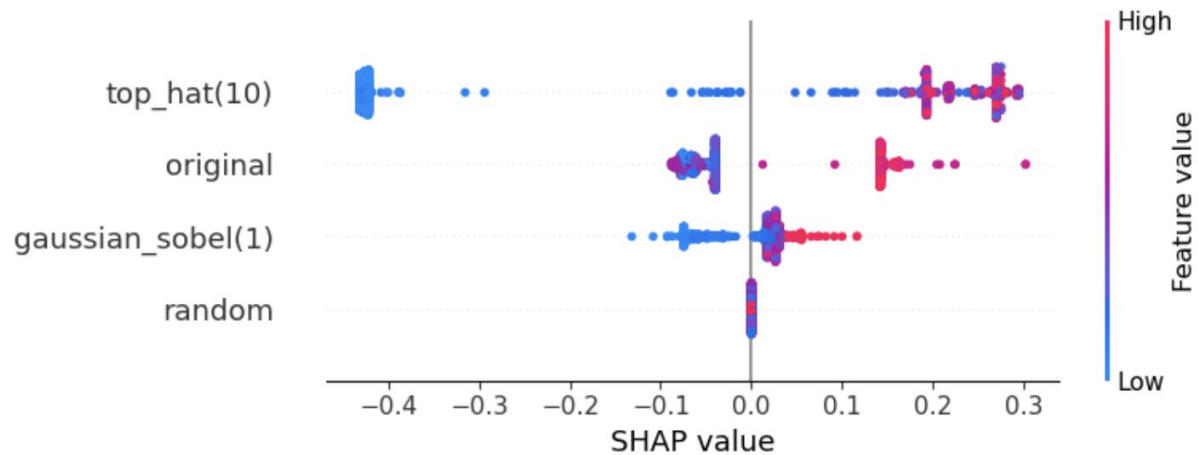
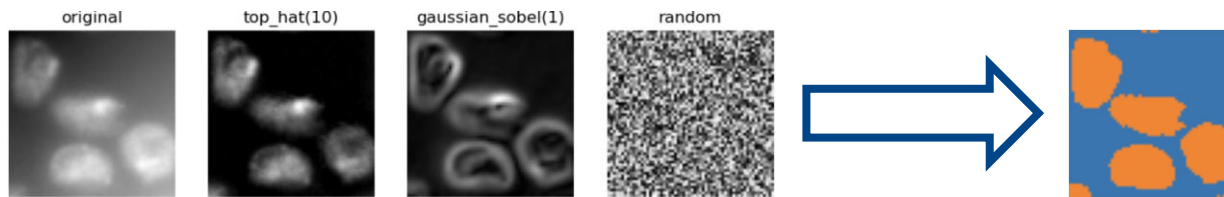
Nicht ohne Weiteres erklärbar und
Ergebnisse auch nicht solide vorhersagbar

Interpretability

Visualization of intermediate results and their influence on results

Model-agnostic methods

Example: Shapley's Additive exPlanations (SHAP)

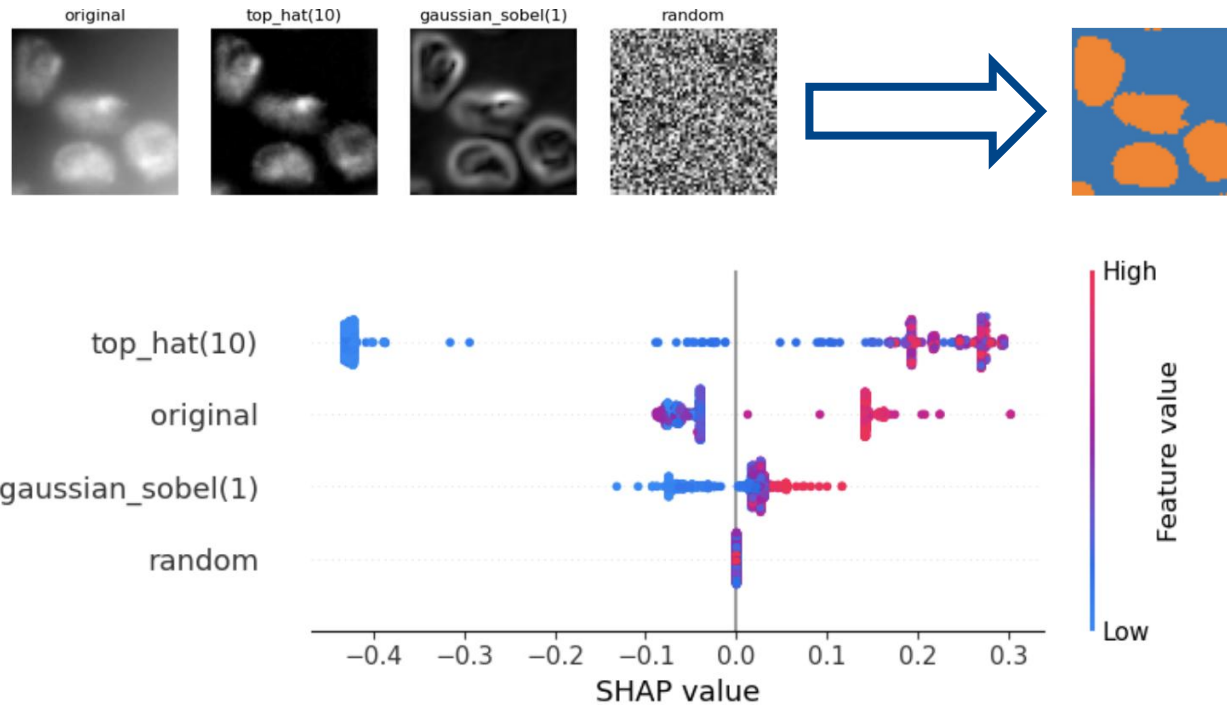


Interpretability

Visualization of intermediate results and their influence on results

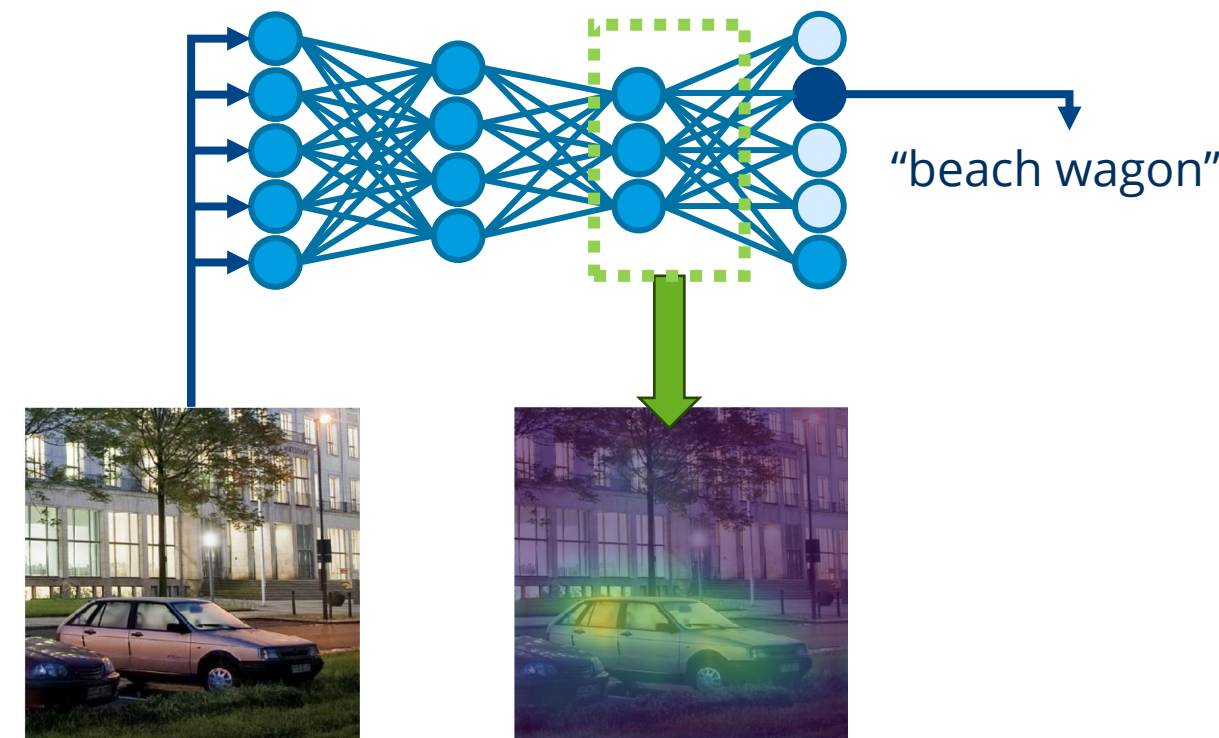
Model-agnostic methods

Example: Shapley's Additive exPlanations (SHAP)



Modell-specific Methods

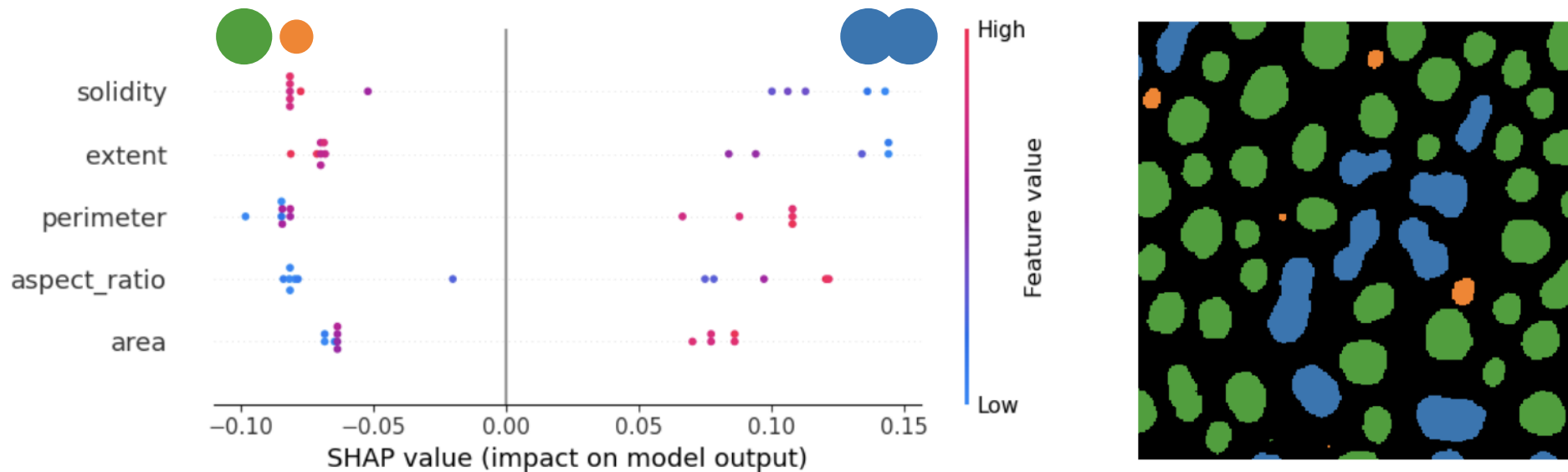
Example : Gradient Class Activation Maps (Grad-CAM)



Target groups

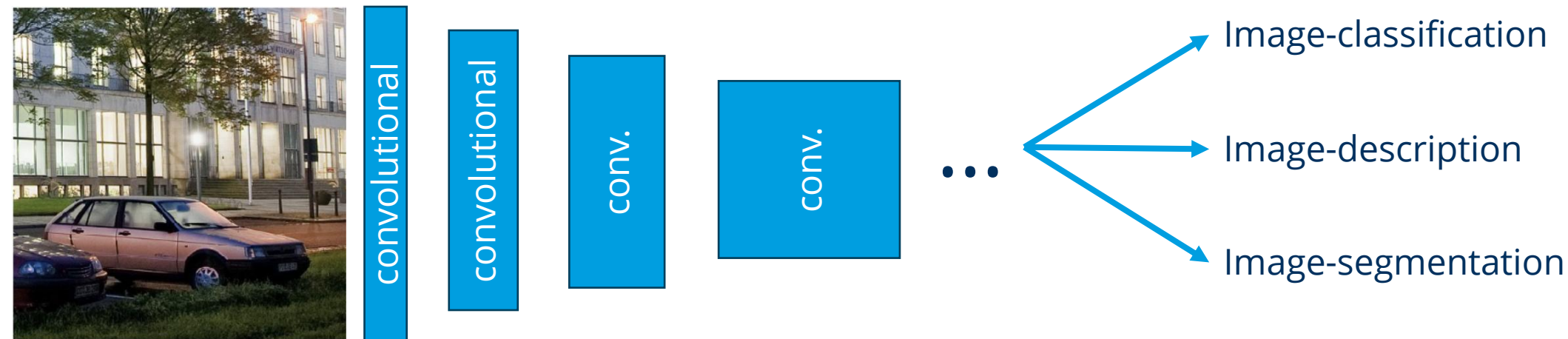
Depending on the target group [for the explanation], the influence of data is more important than how AI algorithms work.

- Many computer scientists want to explain and understand AI methods.
- Biologists use AI as a method to explain biological processes.
- Example: "What parameters distinguish **round objects** from **elongated ones**?"



Gradient Class-Activation Maps (Grad-CAM)

- Works only with NN algorithms that first process input data with convolutional layers. (model-specific)
- Independent of right half of the NN (model-agnostic)
- Visualizes intermediate results to make decision-making in the AI system interpretable



Gradient Class-Activation Maps (Grad-CAM)

Is applied to existing network ; no modification of the architecture necessary (post-hoc method).

Input image

Convolutional layers of a DNN such as ResNet

Output: a vector of probabilities.



convolutional

convolutional

conv.

conv.

0.7

Beach wagon

0.1

goldfish

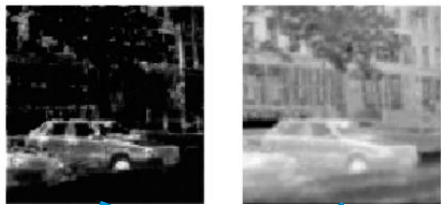
0.1

palace

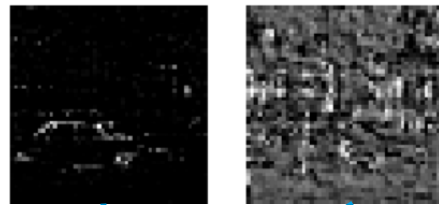
Gradient Class-Activation Maps (Grad-CAM)

Applied to existing network; no adaptation of the architecture necessary (post-hoc method).

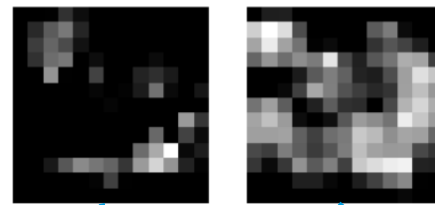
Layer 1 (256, 100, 100)



Layer 2 (512, 50, 50)



Layer 4 (2048, 13, 13)



“2028 feature images
with each 13x13
pixels”

400x400



convolutional

convolutional

conv.

conv.

- Beach wagon
- goldfish
- palace

Quiz

What is this part of a DNN typically called?

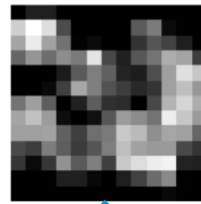
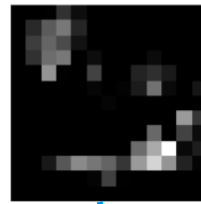
Layer 1 (256, 100, 100)



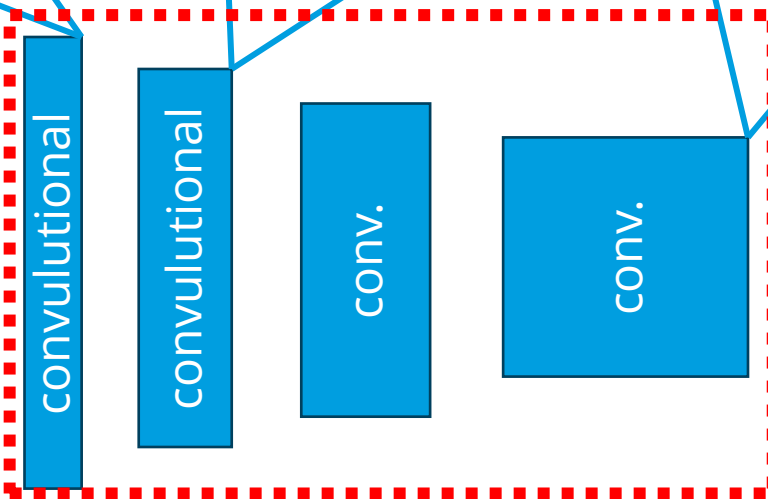
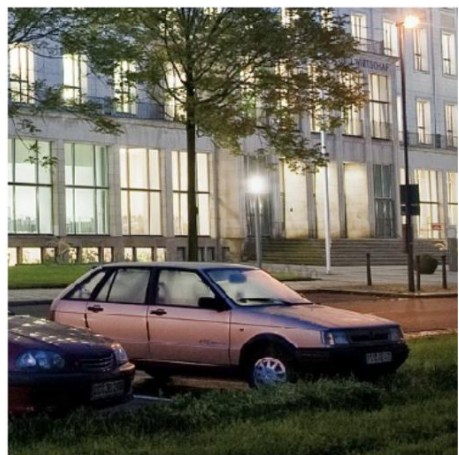
Layer 2 (512, 50, 50)



Layer 4 (2048, 13, 13)



400x400



- Beach wagon
- goldfish
- palace

Reducer



Increaser



Encoder



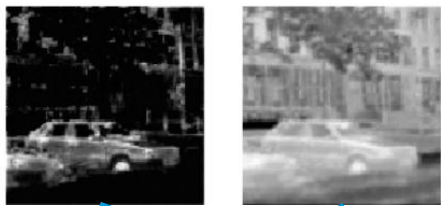
Decoder



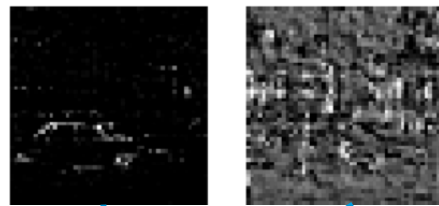
Gradient Class-Activation Maps (Grad-CAM)

Applied to existing network; no adaptation of the architecture necessary (post-hoc method).

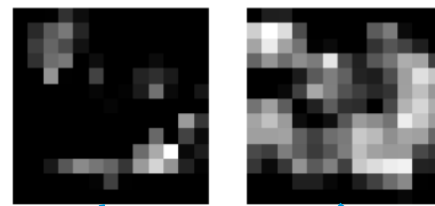
Layer 1 (256, 100, 100)



Layer 2 (512, 50, 50)



Layer 4 (2048, 13, 13)



None of these images directly says anything about image content. There is no feature image "Beach wagon"

400x400



convolutional

convolutional

conv.

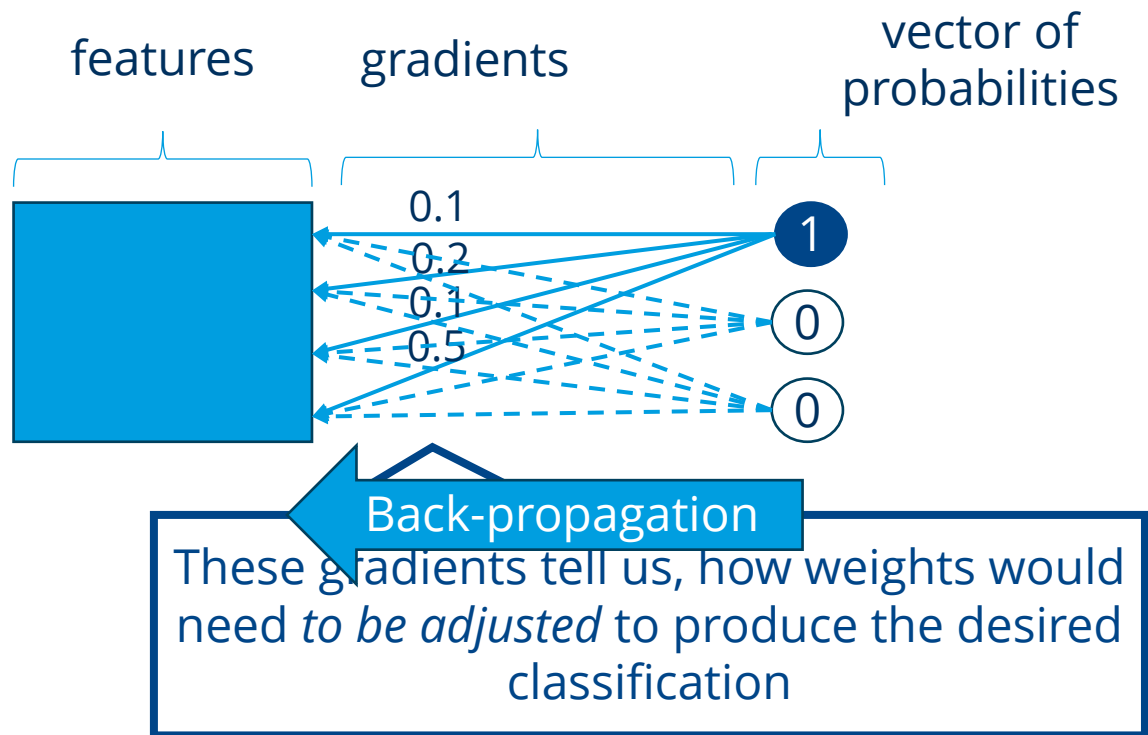
conv.

- Beach wagon
- goldfish
- palace

Grad-CAM happens here

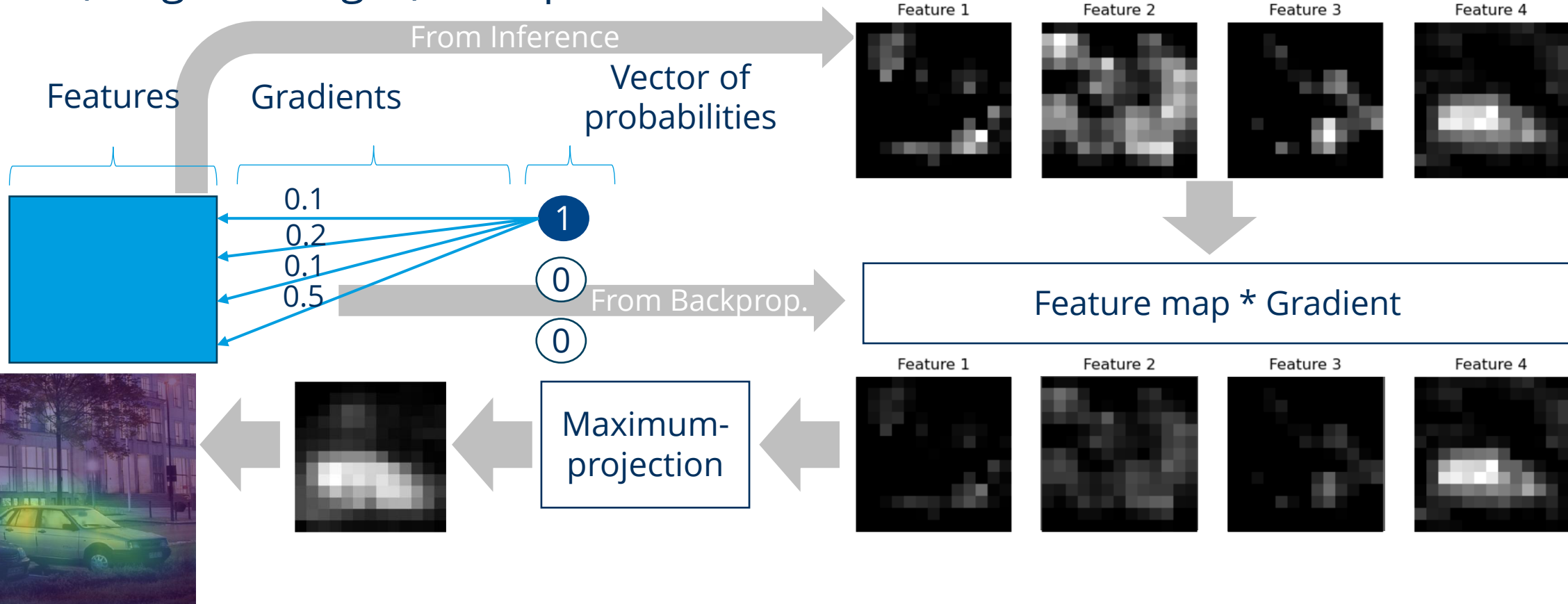
Gradient Class-Activation Maps (Grad-CAM)

Back-propagation of a perfect classification (1,0,0) gives us gradients (weight changes) to improve the classification.



Gradient Class-Activation Maps (Grad-CAM)

Back-propagation of a perfect classification (1,0,0) gives us gradients (weight changes) to improve the classification.



Gradient Class-Activation Maps (Grad-CAM)

Back-propagation of a perfect classification (1,0,0) gives us gradients (weight changes) to improve the classification.

This also works with other possible classifications, e.g. (0,1,0).

“beach waggon”



“palace”



“flagpole”

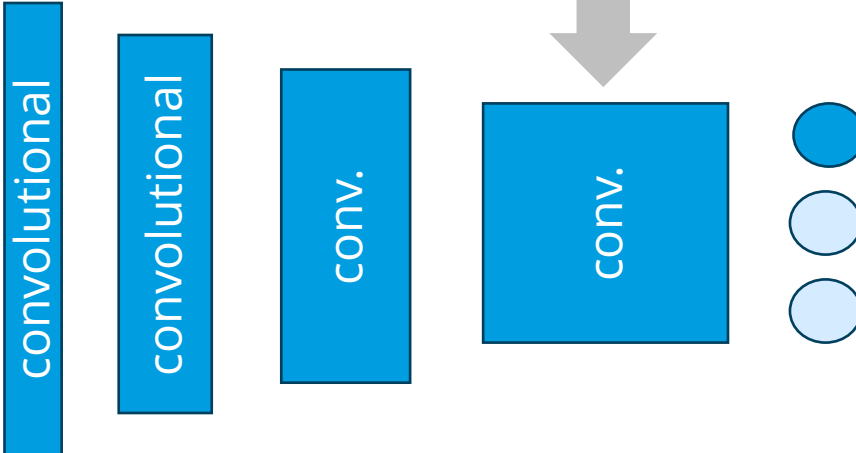


“great white shark”



Quiz

Assuming, this layer has $2048 \times 13 \times 13$ outputs. What does the 2048 stand for?



Number of
features



Width of the
feature maps



Number of
classes

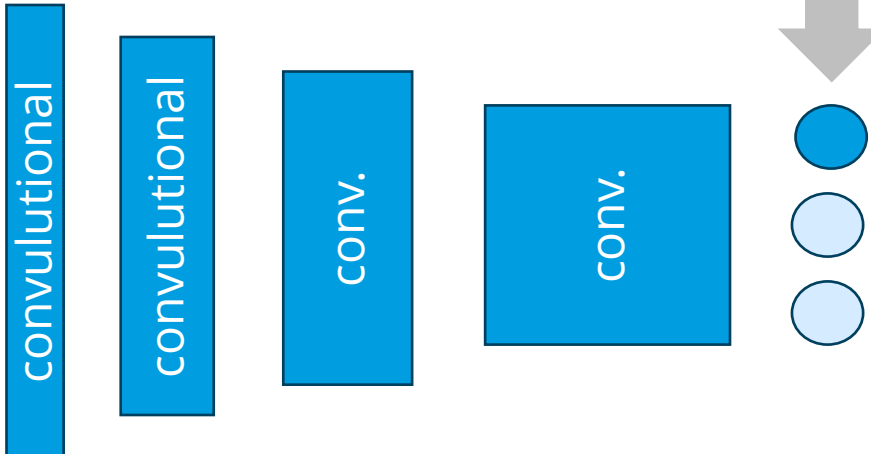


Number of
layers



Quiz

Assume this vector has 1000 elements.
What does the 1000 stand for?



Number of
features



Width of the
feature maps



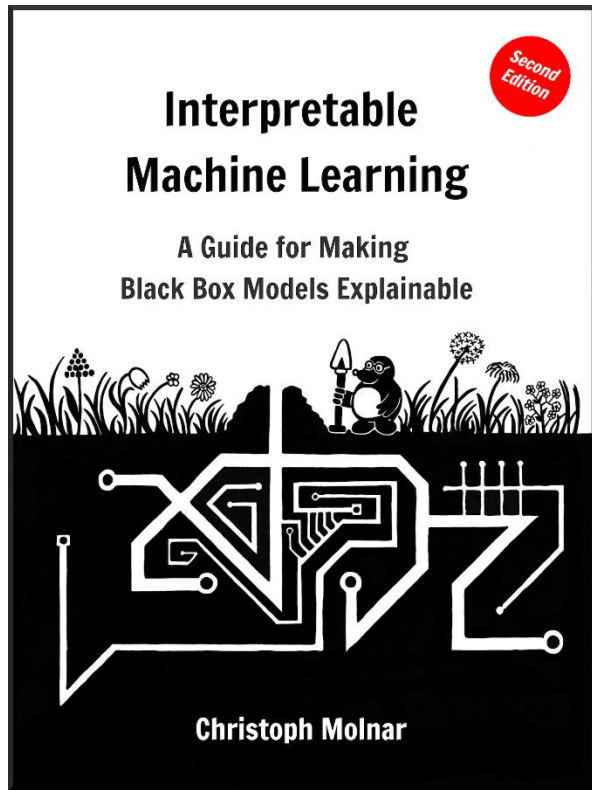
Number of
classes



Number of
layers



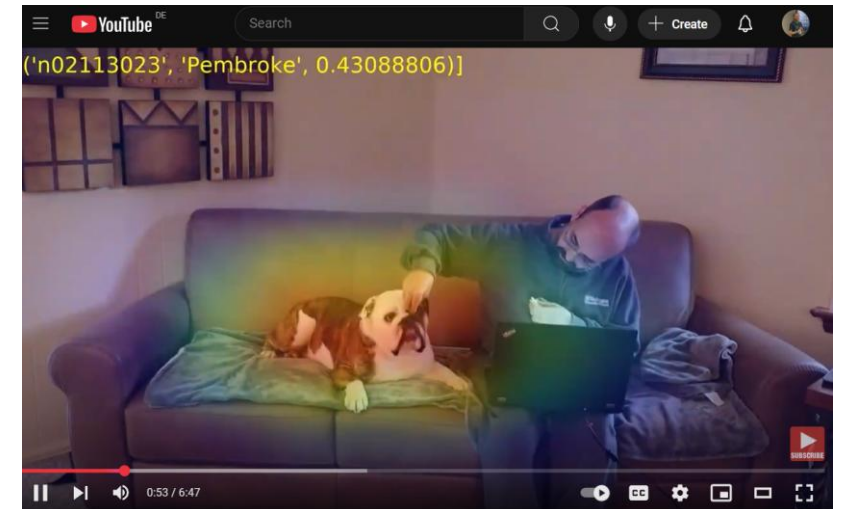
Read more...



<https://christophm.github.io/interpretable-ml-book/>



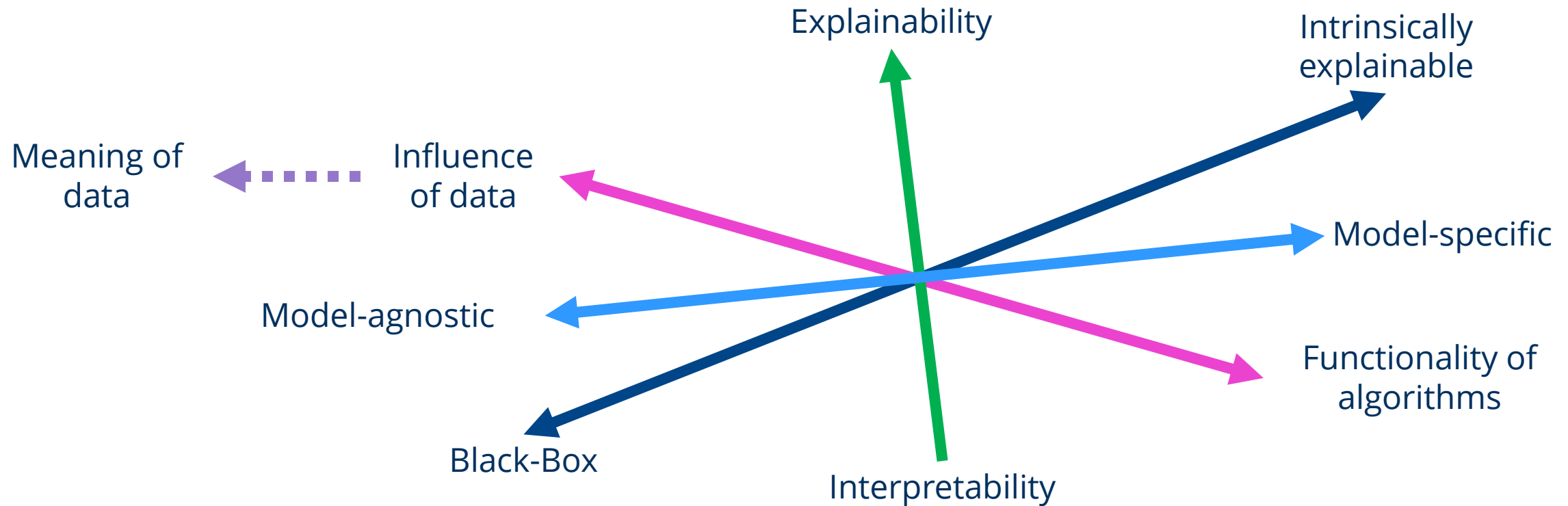
<https://www.amazon.de/dp/3030686396>



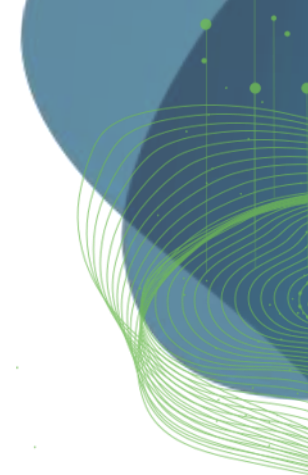
https://www.youtube.com/watch?v=dw63QH_b3Jo

Summary: Explainable AI

Methods of XAI can be classified on different scales



Exercises



Exercises

Explaining Object classification

haesleinhuepf.github.io/xai/30_shap/object_classification.html

ScaDS.AI
DRESDEN LEIPZIG

Search th [Ctrl] [K]

Explainable Artificial Intelligence Notebooks

Setup
Setting up your computer

SHAP Analysis
Pixel classification explained with SHAP
Explaining Object classification using SHAP

Grad-CAM
Gradient Class-Activation Maps (Grad-CAM)

Links
Imprint

Explain classification using SHAP values

```
# Import necessary Libraries
import shap

# Calculate SHAP values
explainer = shap.TreeExplainer(rf)
shap_values = explainer.shap_values(X)[...,0]

shap.summary_plot(shap_values, X) #, feature_names=feature_columns)
```

Exercise

Draw the SHAP summary plot for the shap values [..., 1]. Which object class was this SHAP plot drawn for?

SHAP-values

Gradient Class-Activation Maps

haesleinhuepf.github.io/xai/60_grad-cam/classification_resnet.html

ScaDS.AI
DRESDEN LEIPZIG

Search th [Ctrl] [K]

Explainable Artificial Intelligence Notebooks

Setup
Setting up your computer

SHAP Analysis
Pixel classification explained with SHAP
Explaining Object classification using SHAP

Grad-CAM
Gradient Class-Activation Maps (Grad-CAM)

Links
Imprint

```
show_cam_for_class("great white shark")
```

Exercise

What needs to be changed above to make sure the classification returns "car"?

Exercise

Write a Python function that takes an image filename as parameter and returns the class name as string and a corresponding CAM image. Call this function in a loop which iterates over all images in the folder 'data'.

Grad-CAM