# Stichprobenbasiertes Testen eines CNN mithilfe von LRP

47. TAV, EIKE HANNES MEYER

# Aufbau des Vortrags



- Motivation und Fragestellung
- Methoden und Vorgehen
- Ergebnisse und Erkenntnisse

# Künstliche Intelligenz und ihre Untergebiete



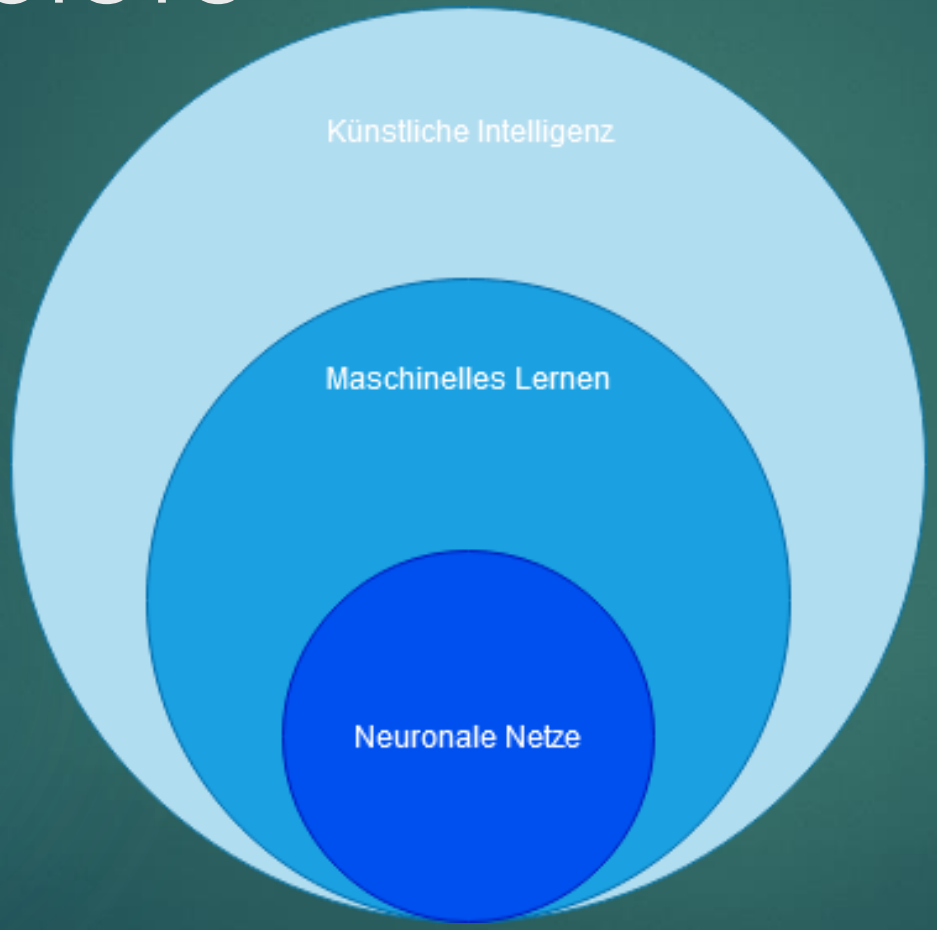



Abbildung 1: Vom Autor

Abbildung 1: Untergebiete der KI

# Motivation



- KI Komponenten werden in vielen Bereichen verwendet [8]
  - Autonomes Fahren, Beurteilung von Menschen, Medizinische Anwendungen
- Es existieren bekannte Probleme
  - Bilder werden anhand falscher Korrelationen klassifiziert
  - Datenlage gibt nicht die echte Welt wieder
  - Es ist Undurchsichtig, worauf Entscheidungen basieren
- Es muss also Methoden geben, Verhalten zu prüfen

# Vergleich der Struktur



## Traditionelle Software

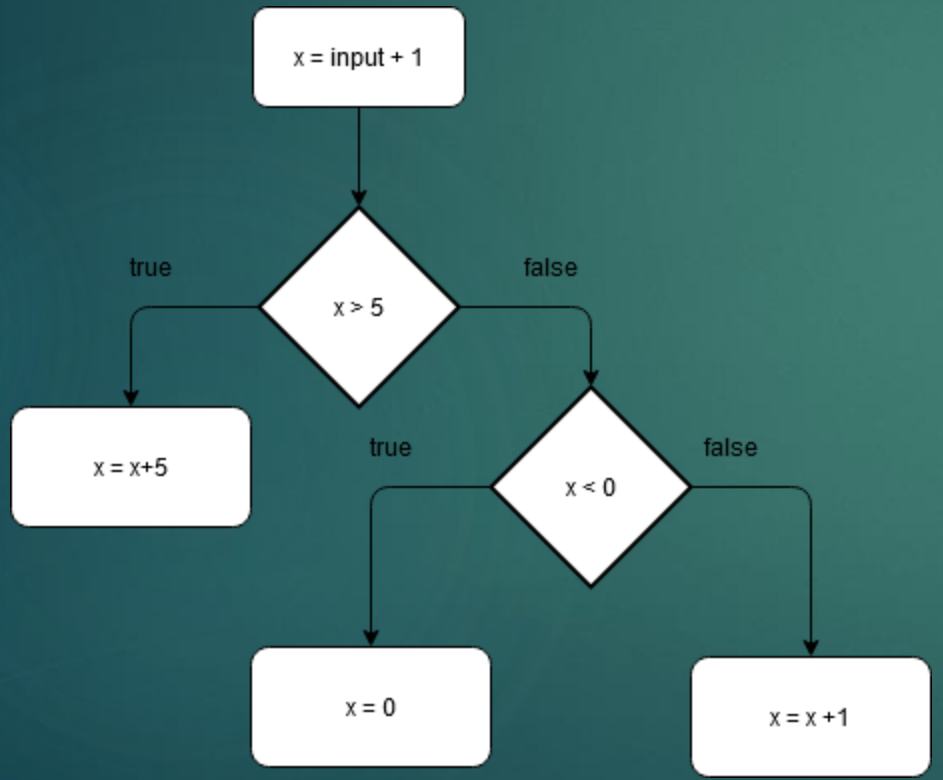



Abbildung 2: Flussdiagramm

## Simples Neuronales Netz

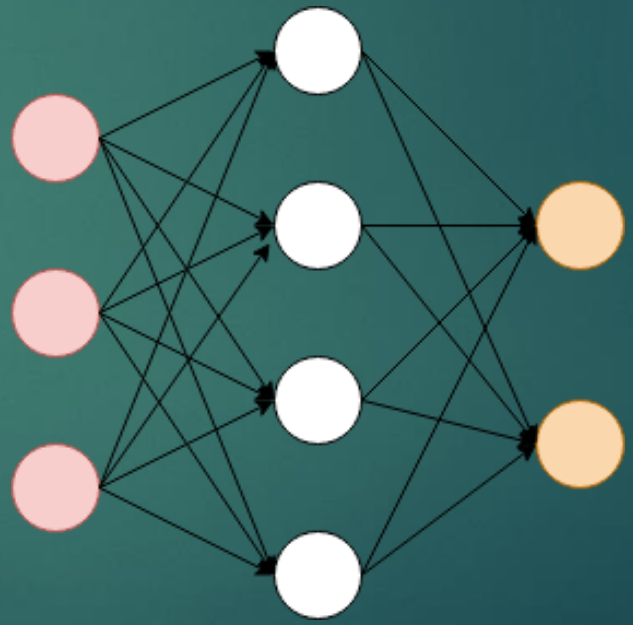


Abbildung 3: NN als gerichteter Graph

Abbildung 2: Vom Autor
Abbildung 3: Vom Autor

# Fragestellung und Thesen



- Neuronale Netze werden viel genutzt, entziehen sich aber traditionellen Testmethoden
- Große Frage: Wie kann Vertrauen in Methoden der AI geschaffen werden?
  - Diese Arbeit ist ein erster Schritt
  - Es gilt wie immer: Es kann nicht alles getestet werden

# Fragestellung



- Kann *Layer-wise relevance propagation* unerwünschtes Verhalten von neuronalen Netzen aufdecken?
  - Konkret: Kann mit Layer-Wise Relevance Propagation festgestellt werden, ob Netze auch Artefakte zur Klassifikation verwenden?

# Einordnung



- Visualisierungsmethoden werden aus Sicht des Software Testing betrachtet
- Visualisierungsmethoden und Black Box Methoden werden als Analogien betrachten
- Getestetes Objekt:
  - Neuronales Netz zur Bildklassifikation
  - Datensatz ist Teilmenge von ImageNet
- Anwendung von LRP auf das Netz

# Analogie zu Black Box Tests



- Analogie zwischen Visualisierungsmethoden und Black Box Methoden
  - Anwendung erfolgt anhand eines fertigen Netzes
  - Anhand einer Eingabe wird eine Ausgabe erzeugt
  - Ausgabe wird interpretiert
  - Es wird kein Wissen über interne Strukturen vorausgesetzt

# Abgrenzung Methoden



- Keine weiteren Methoden der künstlichen Intelligenz
- Keine weiteren Problemstellungen
- Keine weiteren explainable AI Methoden
- Insbesondere keine Betrachtung von Analogien zu White-Box Methoden
  - Betrachtung von Neuron und Kantenüberdeckung

# Abgrenzung in dieser Arbeit



- Keine rein theoretische Betrachtung der Methoden
  - Andere Autoren stellen Anforderungen an Methoden
- Visualisierungsmethoden werden nicht aus Optimierungssicht für Genauigkeit betrachtet
- Durch veränderte Daten sollen Ergebnisse bestärkt werden

# Layer-Wise Relevance Propagation

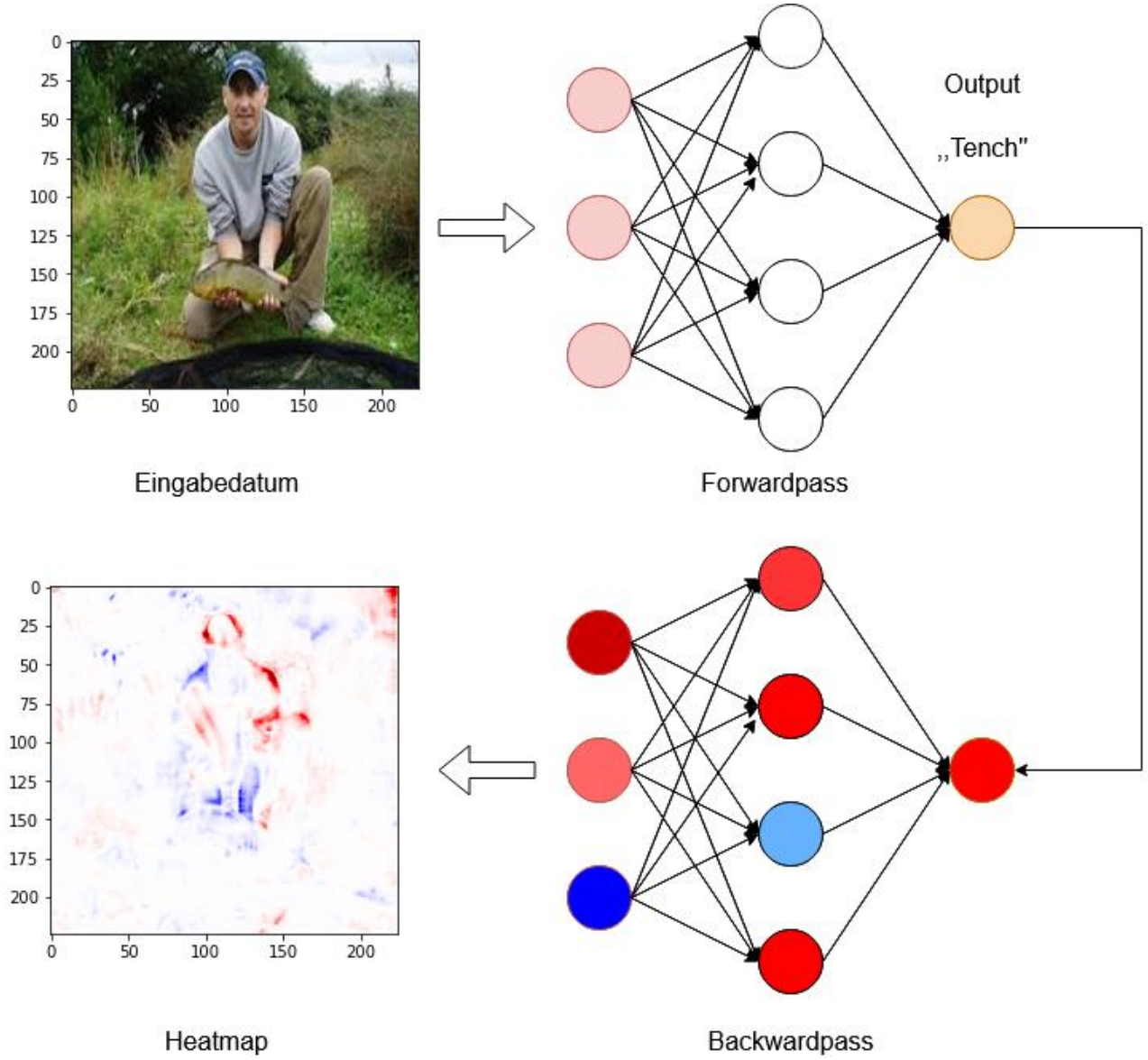



Abbildung 3: vom autoren nach [4]

# Versuchsaufbau

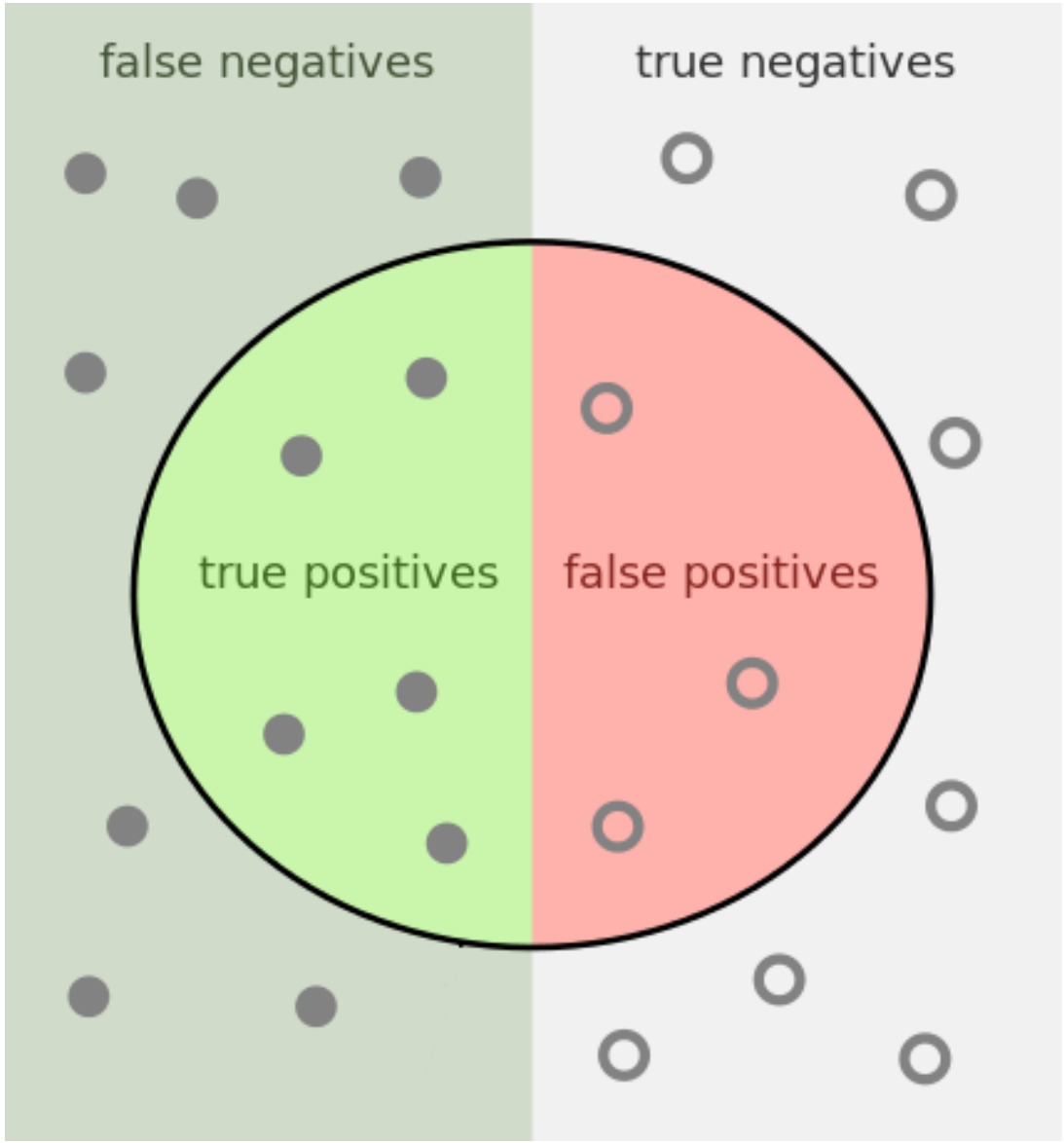


- Datensatz: ImageNette (Teilmenge von Imagenet)
  - Gute Auflösung, 10 stark unterschiedliche Klassen
- CNN zur Bildklassifikation
  - Mit Keras erstellt
  - Feste Struktur
  - 47 Millionen Parameter
- Werkzeug für LRP „iNNvestigate!“
- Bilder mithilfe von „Gimp“ verändert

# Auswahl der Testdaten



- Bilder der Klasse „Tench“
  - Bei einer erster händischer Durchsicht wurden bereits viele andere Objekte in den Daten gesehen
- Anhand der Erkennungssicherheit:
  - 10 Bilder mit höchster Sicherheit
  - 20 Bilder mit niedrigster Sicherheit




https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

# Ausgewählte Testfälle

ANHAND BINÄRER KLASSIFIKATION

# LRP-E1



- Unveränderte Bilder der Klasse „tench“ klassifizieren lassen
- Danach erstellen einer Heatmap für Bilder

ILSVRC2012_val_00006697.JPEG aus [1]

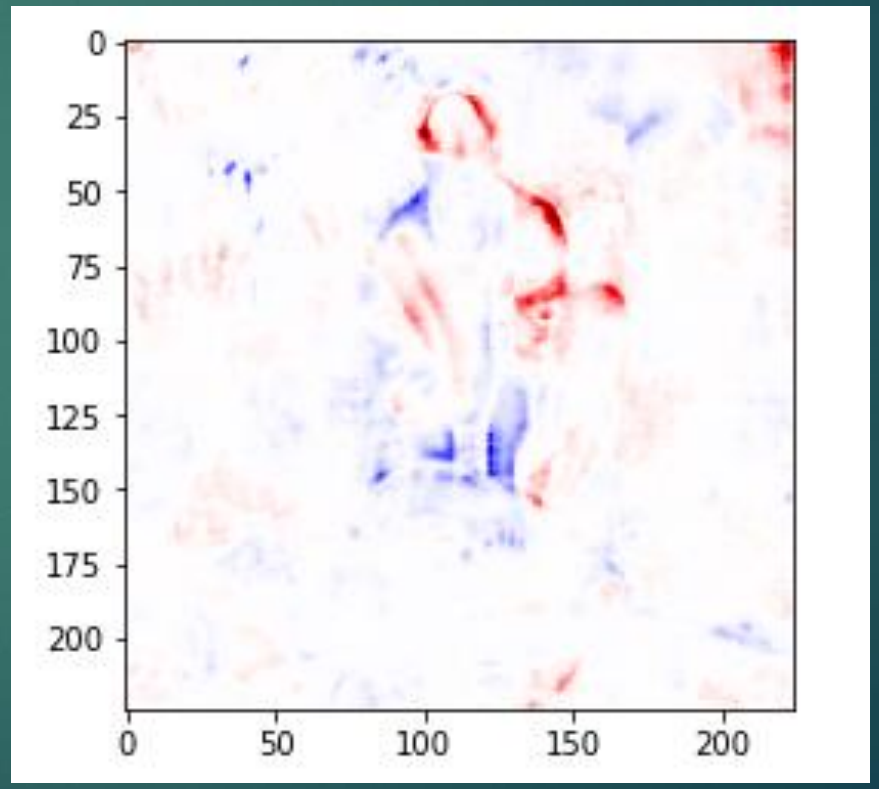
ILSVRC2012_val_00009379.JPEGaus [1]

[1] Russakovsky et al 15

# LRP-E1.1

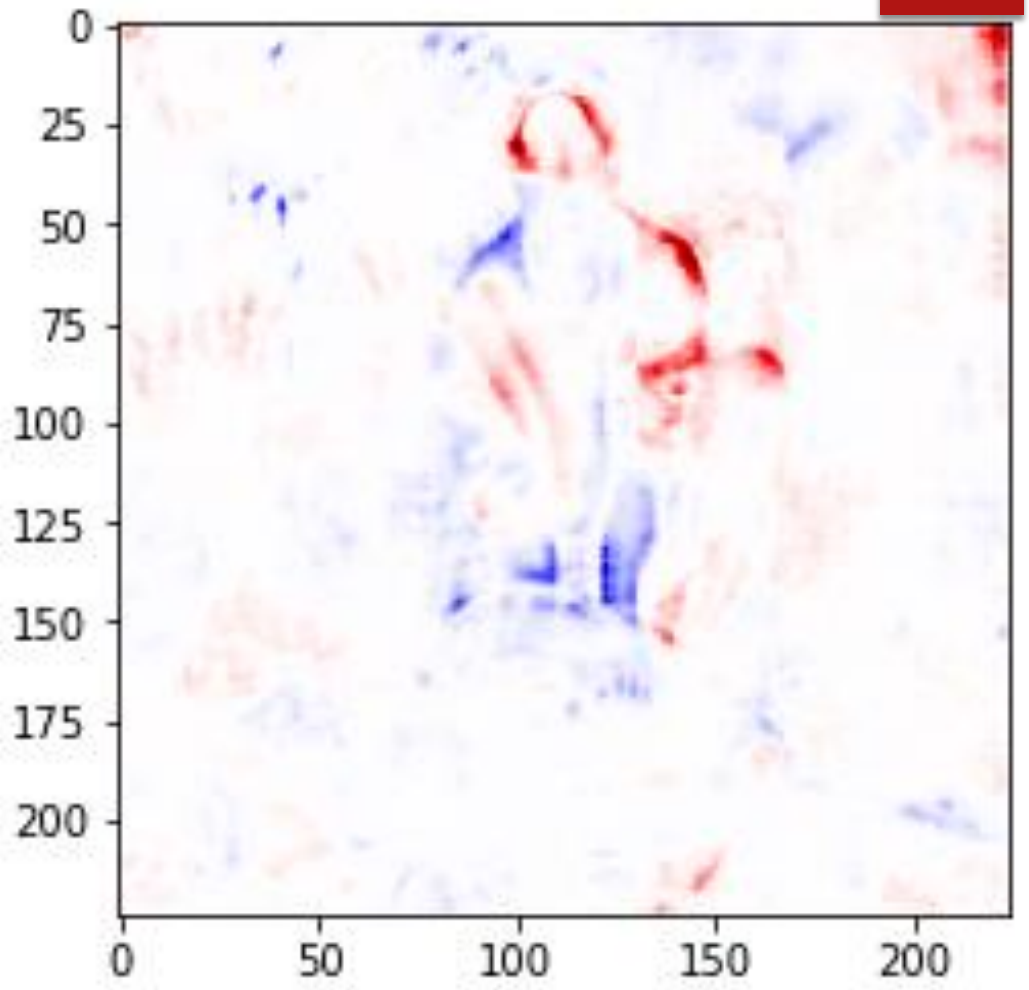


```
Predicted class is tench
with a score of
0.45526782
```

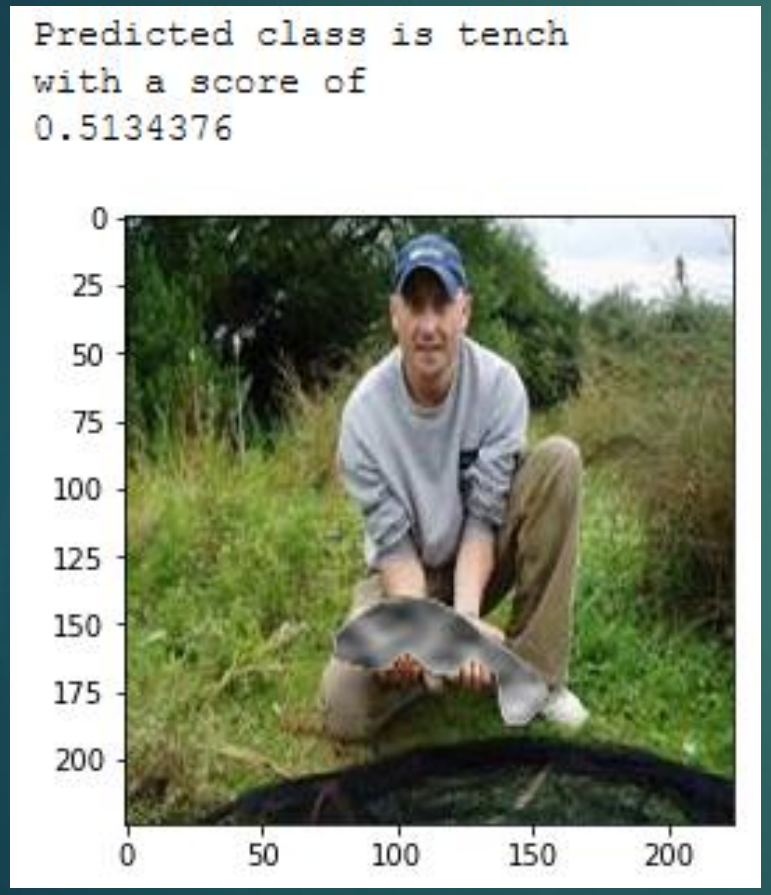
# LRP-E1.1 Beobachtungen

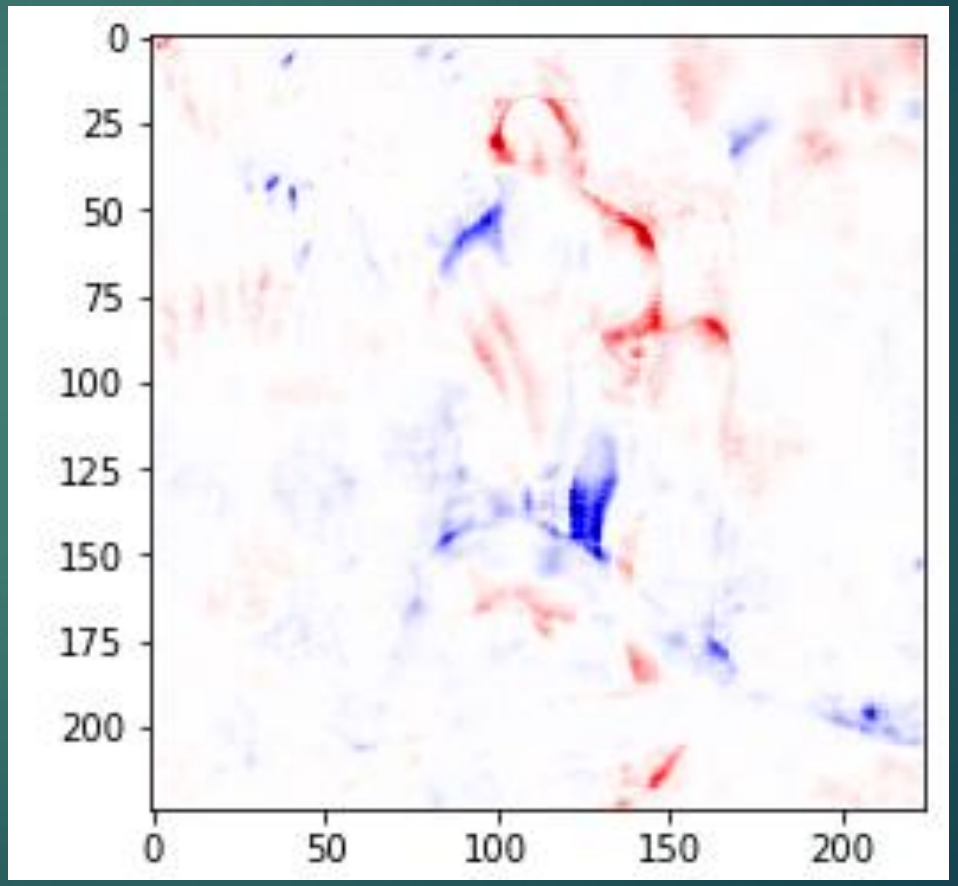
- Heatmap zeigt hohe Relevanz an Teilen des Menschen
- Fisch hat keinen bis negativen Einfluss



# LRP-E2.1

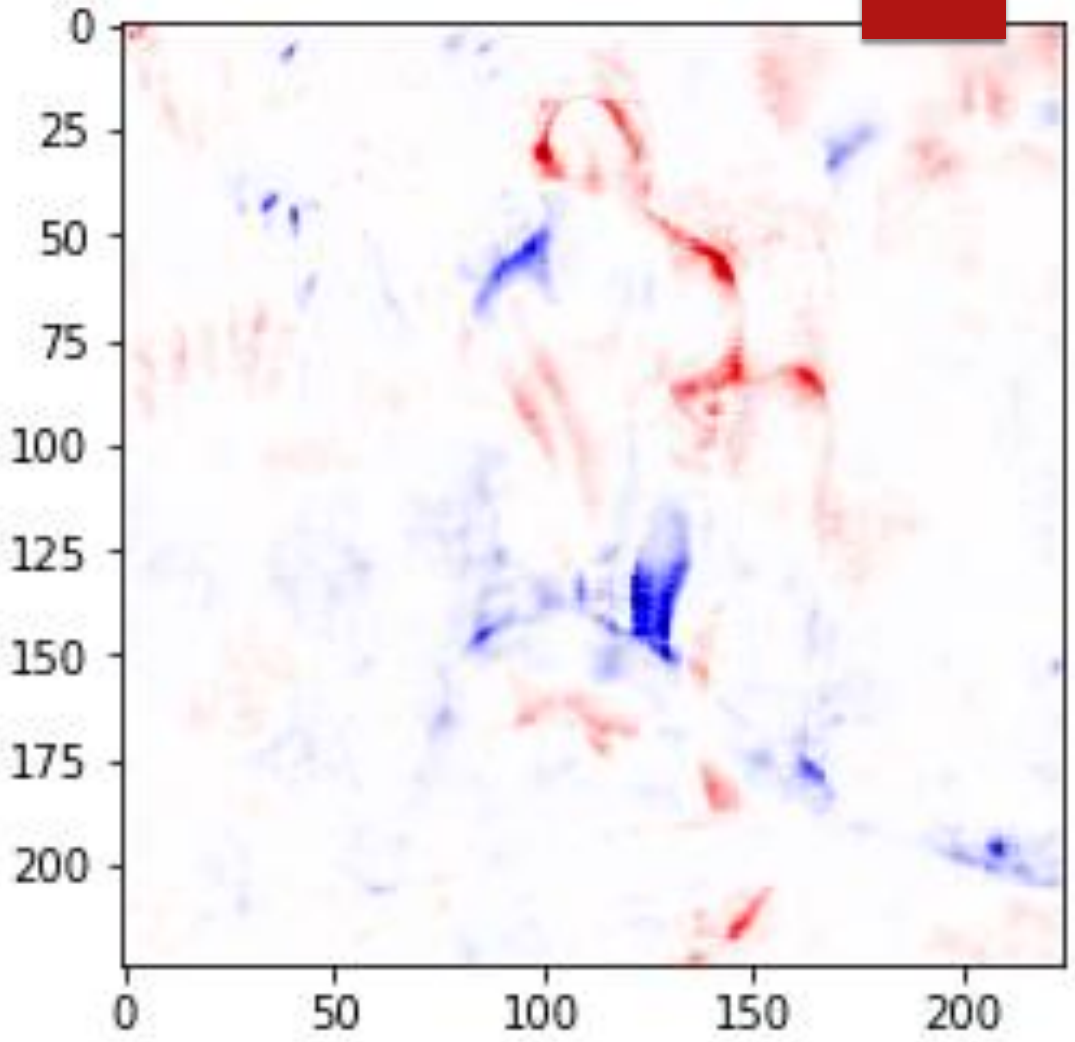
- Eigentliche Instanz wurde entfernt



# LRP-E2.1 Beobachtungen

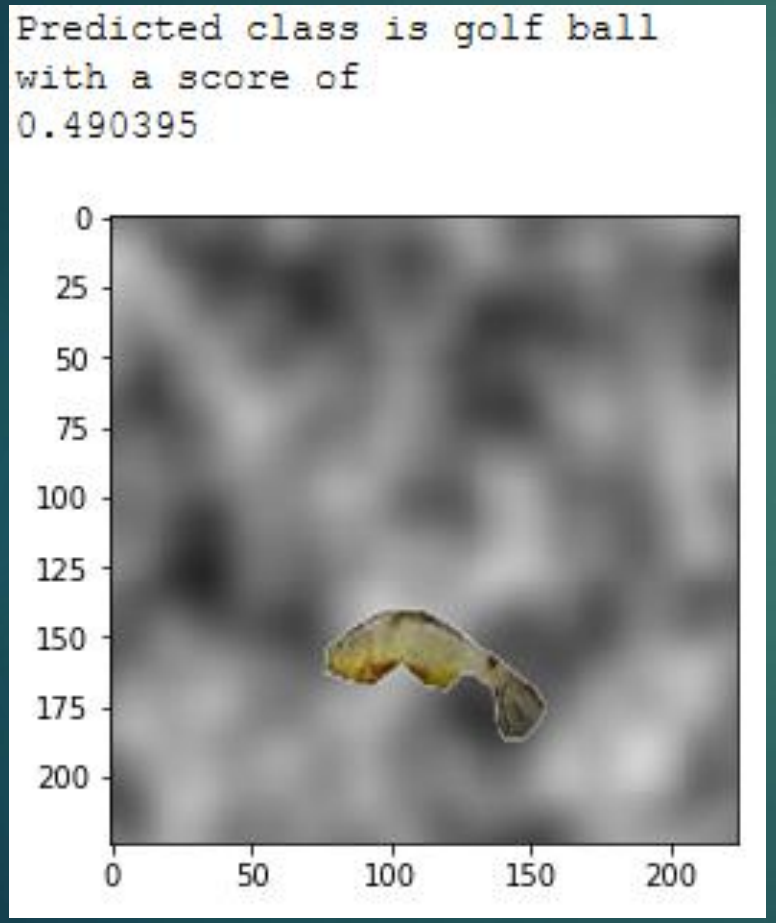
- Richtig klassifiziert (False-Positive)
- Heatmap zeigt hohe Relevanz an Teilen des Menschen
  - Wie LRP-E1
- Fehlen der Instanz erhöht Genauigkeit

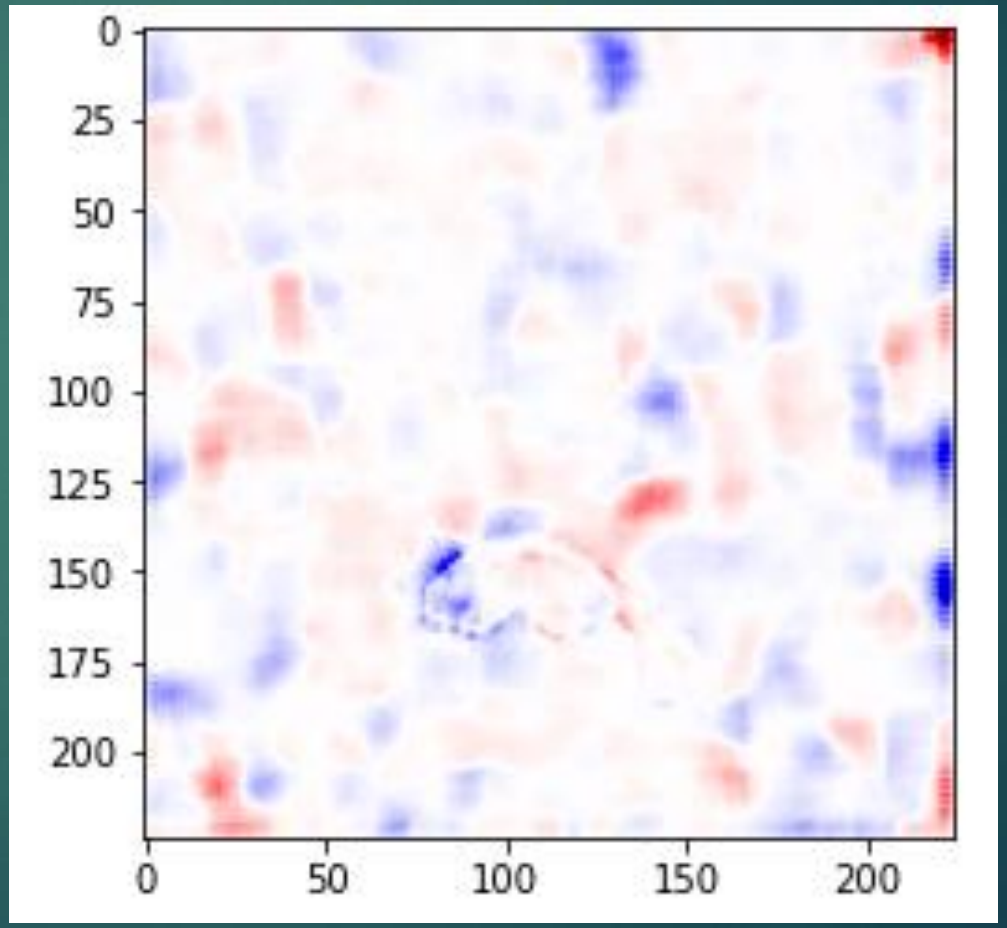


# LRP-E3.1



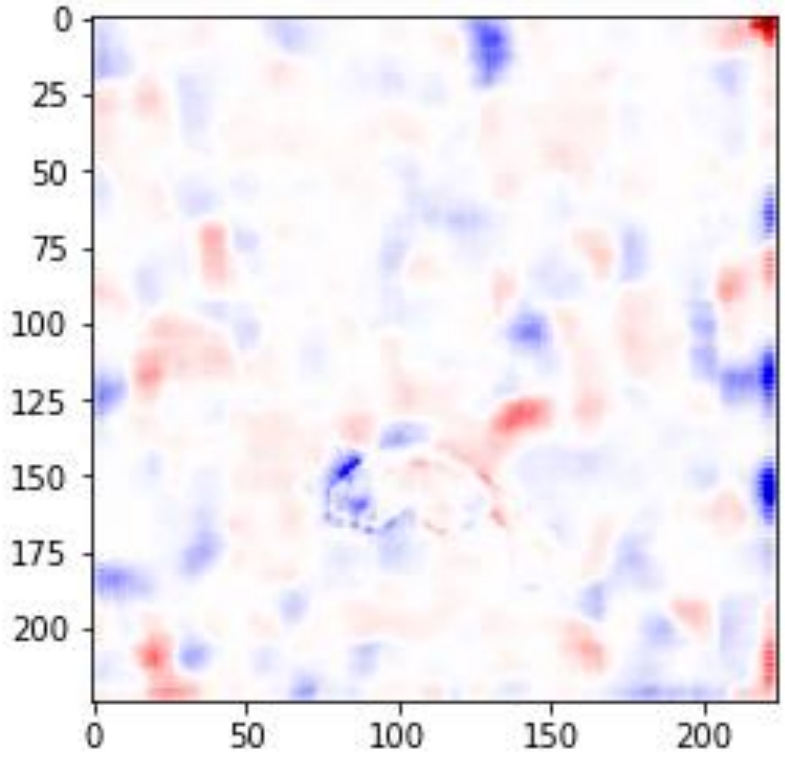
- Alles außer der Instanz wurde entfernt



# LRP-E3.1 Beobachtungen

- Falsch Klassifiziert (False-Negative)
- Heatmap zeigt hohe Relevanz am Hintergrund
- Fehler der Artefakte sorgt für falsche Klassifikation



# Ergebnisse



- Kann mit Layer-Wise Relevance Propagation festgestellt werden, ob Netze auch Artefakte zur Klassifikation verwenden?
  - Ja!
  - Schon anhand vorhandener Validierungsdaten
  - Insbesondere durch manuell veränderte Datenbeispiele
    - False-Positives, False-Negatives

# Erkenntnisse



- Das untersuchte Netz nutzt Artefakte zur Klassifikation
  - Dies konnte Mit LRP durch wenige Stichproben nachgewiesen werden
- Veränderte Bilder untermauern Ergebnisse
  - Bild ohne Instanz wurden richtig klassifiziert (FP)
  - Bild nur mit Instanz wurden falsch klassifiziert (FN)
- LRP testet vorrangig Daten
  - Validierung erfolgte durch händische Zählung: ca. 52% der Bilder enthalten zusätzlich Menschen

# Handlungsempfehlungen



- Validierung von AI Komponenten sollte nicht nur anhand korrekter Klassifikation stattfinden
  - Nicht nur Genauigkeit oder andere Metriken verwenden
  - Auch die Entscheidungsgrundlage sollte betrachtet werden
- Trainingsdaten müssen sorgsam betrachtet werden
  - Hierbei hilft auch LRP indirekt

# Mehrwert der Erkenntnisse



- Trainings sowie Validierungsdaten sollten auch Sondersituationen enthalten
  - Synthetische Veränderungen sind hierbei hilfreich [7]
  - Konkret: Eigentliche Instanz mit unterschiedlichen Umgebungen
- Testfälle müssen von Domänenexperten mit entworfen werden
  - Synthetische Veränderungen sind hierbei hilfreich [7]

# Ausblick



- Verstärken der Ergebnisse
  - Andere Modelle
  - Andere Datensätze
- Weitere Beschäftigung mit CMV und GRAD-CAM

# Quellen




- [1] Bach, S. ; Binder, Alexander ; Montavon, Grégoire ; Klauschen, F. ; Müller, K. ; Samek, W.: On Pixel-Wise Explanations for Non-Linear Classier Decisions by Layer-Wise Relevance Propagation. In: PLoS ONE 10 (2015)
- [2] Jordon, James ; Yoon, Jinsung ; Schaar, Mihaela van der: Measuring the quality of Synthetic data for use in competitions. In: CoRR abs/1806.11345 (2018). URL http://arxiv.org/abs/1806.11345
- [3] Montavon, Grégoire ; Binder, Alexander ; Lapuschkin, Sebastian ; Samek, Wojciech; Müller, Klaus-Robert: Layer-Wise Relevance Propagation: An Overview. S. 193209. In: Samek, Wojciech (Hrsg.) ; Montavon, Grégoire (Hrsg.) ; Vedal-di, Andrea (Hrsg.) ; Hansen, Lars K. (Hrsg.) ; Müller, Klaus-Robert (Hrsg.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham : Springer International Publishing, 2019. URL https://doi.org/10.1007/978-3-030-28954-6_10. ISBN 978-3-030-28954-6

# Quellen




- [4] Russakovsky, Olga ; Deng, Jia ; Su, Hao ; Krause, Jonathan ; Satheesh, Sanjeev ; Ma, Sean ; Huang, Zhiheng ; Karpathy, Andrej ; Khosla, Aditya ; Bernstein, Michael ; Berg, Alexander C. ; Fei-Fei, Li: ImageNet Large Scale Visual Recognition Challenge. In: International Journal of Computer Vision (IJCV)115 (2015), Nr. 3, S. 211252
- [5] Simonyan, Karen ; Zisserman, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015
- [6] Sun, Youcheng ; Huang, Xiaowei ; Kroening, Daniel ; Sharp, James ; Hill, Matthew ; Ashmore, Rob: Testing Deep Neural Networks. 2019

# Quellen

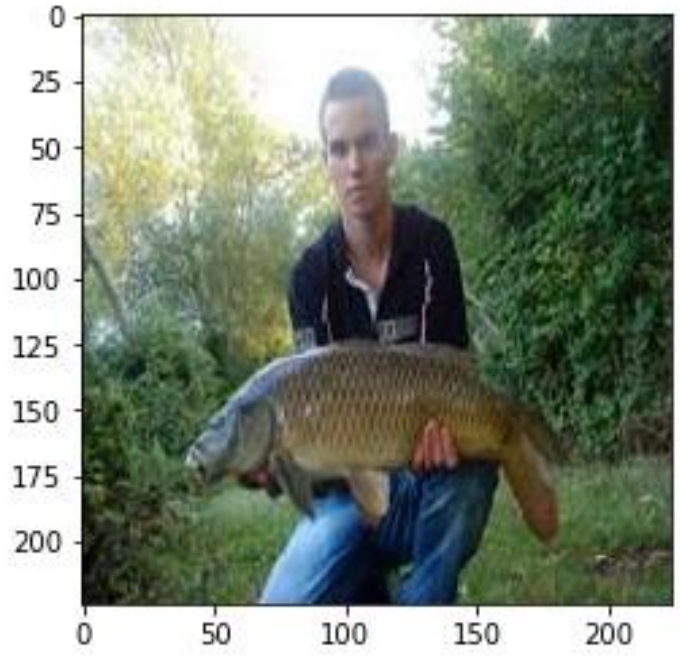


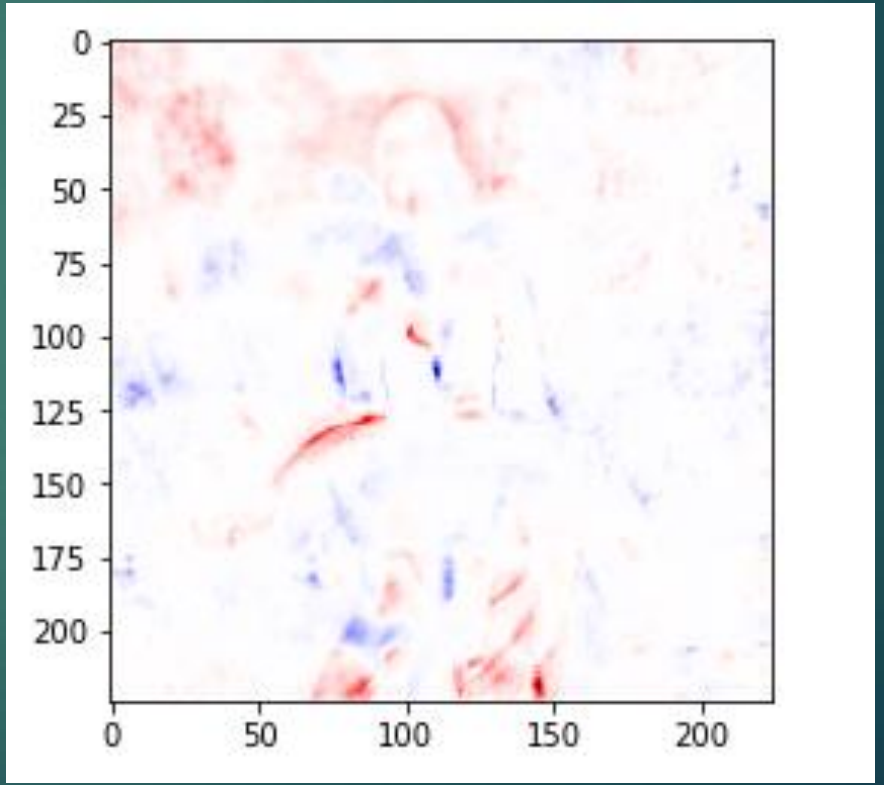

- [7] Tremblay, Jonathan ; Prakash, Aayush ; Acuna, David ; Brophy, Mark ; Jam-pani, Varun ; Anil, Cem ; To, Thang ; Cameracci, Eric ; Boochoon, Shaad ; Birchfield, Stan: Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. 2018
- [8] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated, 2020.

# LRP-E1.2



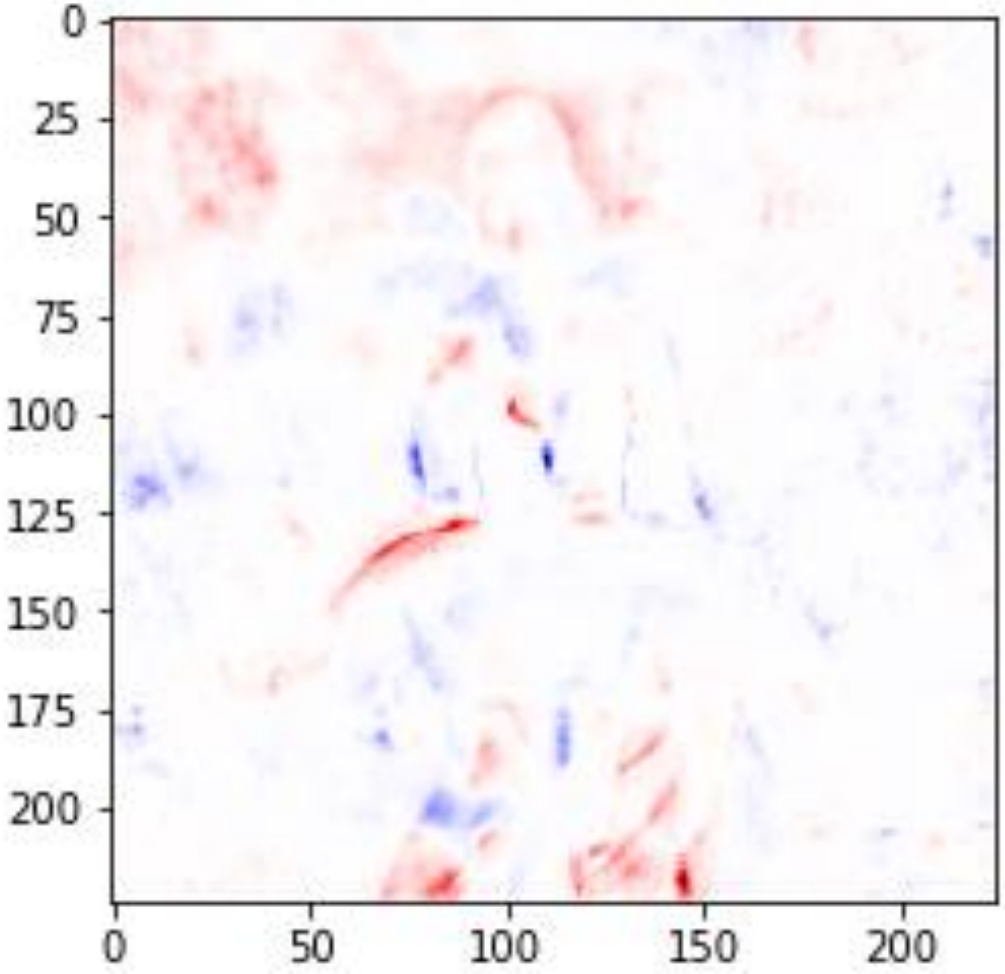
Predicted class is tench
with a score of
0.999841



# LRP-E1.2 Beobachtungen

- Heatmap zeigt hohe Relevanz an Teilen des Menschen
- Kopf des Fisches hat einen positiven Einfluss



- Eigene Architektur

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_12 (Conv2D) | (None, 224, 224, 64) | 1792 |
| conv2d_13 (Conv2D) | (None, 224, 224, 64) | 36928 |
| max_pooling2d_6 (MaxPooling2 | (None, 112, 112, 64) | 0 |
| batch_normalization_6 (Batch | (None, 112, 112, 64) | 256 |
| conv2d_14 (Conv2D) | (None, 112, 112, 128) | 73856 |
| conv2d_15 (Conv2D) | (None, 112, 112, 128) | 147584 |
| max_pooling2d_7 (MaxPooling2 | (None, 56, 56, 128) | 0 |
| batch_normalization_7 (Batch | (None, 56, 56, 128) | 512 |
| conv2d_16 (Conv2D) | (None, 28, 28, 256) | 295168 |
| conv2d_17 (Conv2D) | (None, 14, 14, 256) | 590080 |
| max_pooling2d_8 (MaxPooling2 | (None, 7, 7, 256) | 0 |
| batch_normalization_8 (Batch | (None, 7, 7, 256) | 1024 |
| conv2d_18 (Conv2D) | (None, 4, 4, 512) | 1180160 |
| conv2d_19 (Conv2D) | (None, 2, 2, 512) | 2359808 |
| max_pooling2d_9 (MaxPooling2 | (None, 1, 1, 512) | 0 |
| batch_normalization_9 (Batch | (None, 1, 1, 512) | 2048 |
| conv2d_20 (Conv2D) | (None, 1, 1, 512) | 2359808 |
| conv2d_21 (Conv2D) | (None, 1, 1, 512) | 2359808 |
| max_pooling2d_10 (MaxPooling | (None, 1, 1, 512) | 0 |
| batch_normalization_10 (Batc | (None, 1, 1, 512) | 2048 |
| flatten_2 (Flatten) | (None, 512) | 0 |
| dense_4 (Dense) | (None, 8192) | 4202496 |
| dropout_2 (Dropout) | (None, 8192) | 0 |
| dense_5 (Dense) | (None, 4096) | 33558528 |
| dropout_3 (Dropout) | (None, 4096) | 0 |
| dense_6 (Dense) | (None, 10) | 40970 |

Total params: 47,212,874
Trainable params: 47,209,930
Non-trainable params: 2,944