Explainable Artificial Intelligence for the NDE4.0

Ahmad OSMAN¹, Ziang WEI²
¹ Fraunhofer-Institut für Zerstörungsfreie Prüfverfahren IZFP, Saarbrücken, Germany
² University of Applied Sciences, htw saar, Saarbrücken, Germany

Contact e-mail: ahmad.osman@izfp.fraunhofer.de

Abstract

As AI Deep Learning (DL) methods have achieved super human-level performance in many tasks, NDE should naturally profit from this advance. However, the implementation of DL in NDE is decelerated by the lack of transparency and explainability of the DL models. While the traditional AI systems are easily interpretable, the last years have witnessed the rise of opaque DL Networks. Opacity is the opposite of transparency. By making a DL-Model transparent and explainable, we aim at reducing its opacity and understanding the mechanism by which the model acts. The success and high capability of DL-models for approximating any function stems from their huge parameters space, complex architecture, activation functions, efficient initialization and optimization algorithms. It is safe to say that highly accurate DL-models are increasingly opaque which will increase the challenge to explain them. This has been an active research topic in the last 6 years. Nowadays there are different methods trying to explain some deep models regarding different aspects. One of the most popular methods is saliency map method. It generates a saliency map in the input data space based on their contribution to the model’s prediction. This saliency map can then be used to assess if the model has used the correct features or not. Another method is to generate an artificial input example that maximize a certain output unit, in this way an insight of what the deep model has really learned can be obtained. Within this contribution, we will tackle the advance in the explainable AI scene and present applications of it for NDE examples. For the NDE society, we want to move toward the concept of Responsible Artificial Intelligence, namely, a methodology for the large-scale application of AI methods in NDE with model explainability, reliability and generalization (especially regarding varying materials anatomy) at its core.
Explainable Artificial Intelligence for the NDE4.0

M. Sc. Ziang Wei, Prof. Dr.-Ing. Ahmad Osman
Fraunhofer-Institute for Nondestructive Testing IZFP
ahmad.osman@izfp.fraunhofer.de
Phone +49 681 / 9302 3628

AUTHOR PRESENTATION

- Professor for Intelligent Sensing Methods at the Saarland University of Applied Sciences, Germany
- Adjunct Professor at Laval University, Canada
- Vice-Department Manager and AI-Research Area Coordinator at Fraunhofer IZFP, Germany
- Co-Initiator and Manager of the AI Group in German Society for NDT – DGZFP / NDE4.0 Committee
- IEEE Senior Member
- Research Topics: Applied AI for NDE4.0 and I4.0, Computer Vision and Signal Processing
CONTENT

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - Output maximization
  - Model activation visualization
- Exemplary Application in NDE
  - Application case
  - Decision attribution approach in NDE
  - Model activation visualization approach in NDE
- Conclusion and outlook
INTRODUCTION

- Deep neural network can read, can talk, can see, can hear and can describe

Assisted and Automated Driving
Object detection: YOLO & YOLO9000 [1,2]
Echo Dot & Amazon Alexa Voice Service

Source: www.amazon.com

INTRODUCTION

- But what do the models really learn?
INTRODUCTION

- Clever Hans: is the horse really clever?

CONTENT

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - Output maximization
  - Model activation visualization
- Exemplary Application in NDE
  - Application case
  - Decision attribution approach in NDE
  - Model activation visualization approach in NDE
- Conclusion and outlook
DECISION ATTRIBUTION APPROACH

\[ f(x) \approx \sum_{d=1}^{v} R_d \]

\( x \): input data, \( x \in \mathbb{R}^v \)

\( f(x) \): Prediction

\( R_d \): Relevance of each input pixel \( x_d \)

\( R_d > 0 \) contributes evidence for the presence of a structure

\( R_d < 0 \) contributes evidence against the presence of a structure

Concept of layer-wise relevance propagation, \( l \) is the index of the layer:

\[ f(x) = \cdots = \sum_{d \in l+1} R_{d}^{(l+1)} = \sum_{d \in l} R_{d}^{(l)} = \cdots = \sum_{d} R_{d}^{(1)} \]
DECISION ATTRIBUTION APPROACH

- Worked as expected!

Top: input images, bottom: input heat map

Reference: [3]
DECISION ATTRIBUTION APPROACH

- Something went wrong...

Top: input images, bottom: input heat map

Reference: [3]

CONTENT

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - **Output maximization**
  - Model activation visualization
- Exemplary Application in NDE
  - Application case
  - Decision attribution approach in NDE
  - Model activation visualization approach in NDE
- Conclusion and outlook
Example of a classification model inference:

- The idea is to generate a synthetic image so that the score of a certain class to be maximized.

<table>
<thead>
<tr>
<th>Score</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>Cat</td>
</tr>
<tr>
<td>0.91</td>
<td>Husky</td>
</tr>
<tr>
<td>0.01</td>
<td>Fish</td>
</tr>
</tbody>
</table>

\[ x^* = \arg \max_x (a_i(x) - R_\theta(x)) \]

- \( x \): input image, and \( x \in \mathbb{R}^{C \times H \times W} \), where \( C = 3 \) color channels and height(\( H \)) and width(\( W \))
- \( a_i(x) \): activation for some unit \( i \); \( R_\theta(x) \): regularization function

Reference: [5]
More examples:

- dumbbell
- cup
- dalmatian
- bell pepper
- lemon
- husky

Reference: [5]
In a pretrained-CNN network, if a certain filter that has strong correlation with for example “face”, then it is a good “face” feature detector.

Input image and their corresponding activation maps of the fifth convolutional layer

Reference: [6]
**CONTENT**

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - Output maximization
  - Model activation visualization
- Exemplary Application in NDE
  - **Application case**
    - Decision attribution approach in NDE
    - Model activation visualization approach in NDE
- Conclusion and outlook

**APPLICATION CASE**

- Automated quality assurance of fuel cell components

on material and product changes.
APPLICATION CASE

- U-Net defect segmentation for fuel cells stack

CONTENT

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - Output maximization
  - Model activation visualization
- Exemplary Application in NDE
  - Application case
  - Decision attribution approach in NDE
  - Model activation visualization approach in NDE
- Conclusion and outlook
DECISION ATTRIBUTION APPROACH IN NDE

Output image

3D Heat map

Contribution along frame index

CONTENT

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - Output maximization
  - Model activation visualization
- Exemplary Application in NDE
  - Application case
  - Decision attribution approach in NDE
    - Model activation visualization approach in NDE
- Conclusion and outlook
MODEL ACTIVATION VISUALIZATION APPROACH IN NDE

- U-Net defect segmentation for fuel cells stack

Exemplary selected activations maps of the last layer

PPT image of Input

Ground truth
MODEL ACTIVATION VISUALIZATION APPROACH IN NDE

- U-Net defect segmentation for fuel cells stack
CONTENT

- Introduction
- Approaches to explain the deep model
  - Decision attribution
  - Output maximization
  - Model activation visualization
- Exemplary Application in NDE
  - Application case
  - Decision attribution approach in NDE
  - Model activation visualization approach in NDE
- Conclusion and outlook

CONCLUSION AND OUTLOOK

- NDE data, compared to visual images, are not explainable for non-qualified personal
  - Explainability of data
  - Explainability of AI-Models

- The explainability of AI-DL-Models is very important and need more attention in order to ease the application and acceptance of AI in NDE

- Explainable AI methods for data segmentation is still not well explored

- Need to universal methods to explain NDE data and AI-Models while considering the physical knowledge about materials, methods and (sensitive) applications
OUTLOOK

Hybrid Responsible and Explainable Artificial Intelligence for NDE4.0

**Hybrid**: physical knowledge and AI

**Responsible**: AI as a disruptive force in our societies, ethical frameworks and regulation, expected levels of technical performance (accuracy, robustness, autonomy, etc.)

**Explainable**: trackable and explicable decisions and reasons for decision making

REFERENCE

Thank you for your attention!

Questions?

Prof. Dr.-Ing. Ahmad Osman
ahmad.osman@izfp.fraunhofer.de
Fon +49 681 9302 / 3628
Fraunhofer IZFP
www.izfp.fraunhofer.de