Advanced Topic 3: Explainable AI (XAI) for Neural Networks

Contents

Introductions

- **Definition**: Explainable AI (XAI) for neural networks focuses on interpreting the complex, non-linear patterns that these models learn, often in a "**black-box**" manner.
- **Importance**: Given their complexity, neural networks are powerful but often obscure in how they make decisions. Explainability is essential for trust and accountability, especially in high-stakes fields.

Introductions

- The concept that an AI model and its output can be explained in a way that **“makes sense”** to an average person, i.e.
    - Data scientist => Explain to improve
    - Domain expert => Explain to discover
    - End user => Explain to build trust
    - Regulator/Lawyer => Explain to control
- XAI not only explains the output, but can also provide **counterfactual reasonings**, e.g.
    - If I change this feature value, does the prediction change?

Challenges

- **Complexity**: Neural networks contain numerous interconnected layers and neurons, making it difficult to track how each feature affects the outcome.
- **High-dimensionality**: Neural networks often handle vast amounts of data with many features, further complicating interpretations.
- **Non-linearity**: Hidden layers in neural networks capture non-linear relationships, making traditional interpretation methods insufficient.

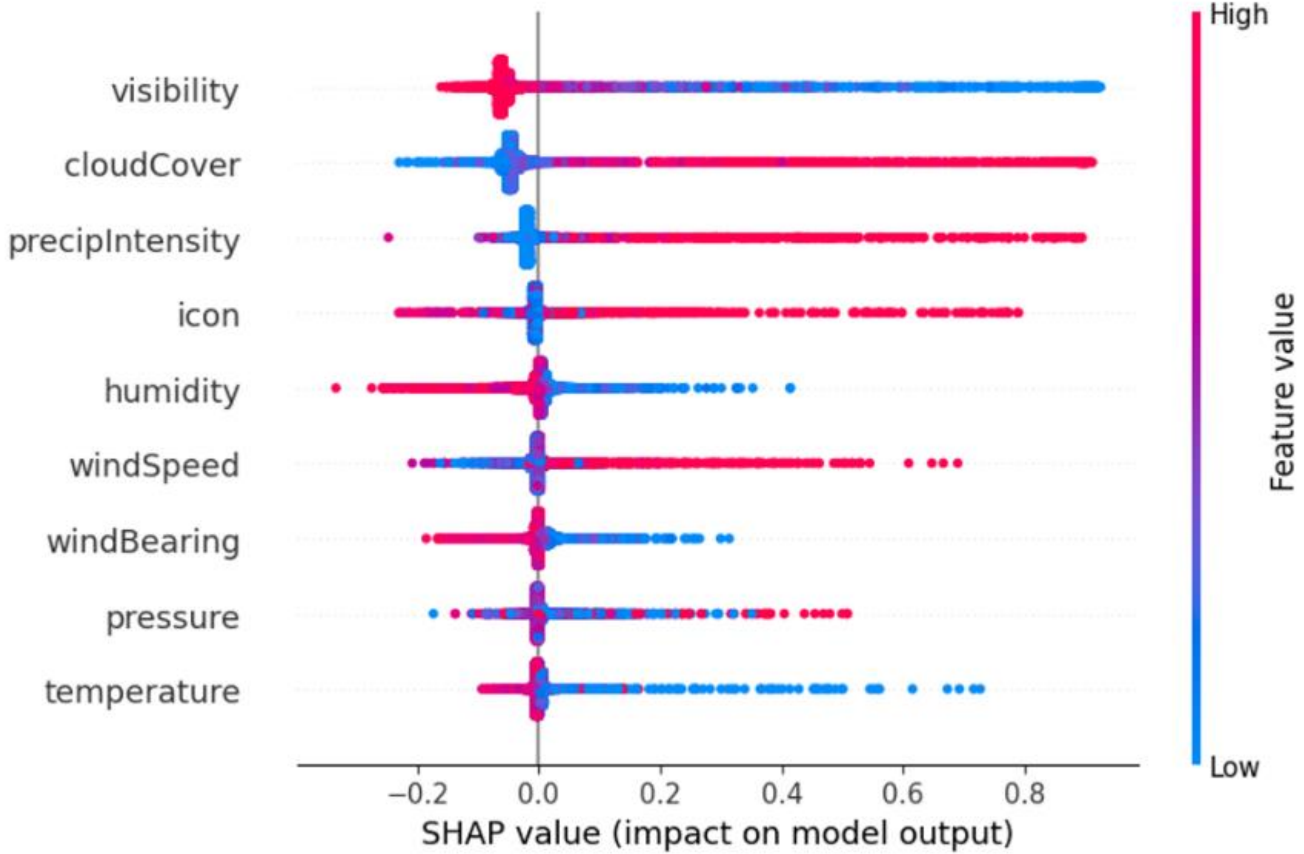
Types of explainability in neural networks

- **Global Explainability**: Understanding the overall behavior of the network, such as which input features generally have the most impact on predictions.
- **Local Explainability**: Focusing on individual predictions to understand why a network classified or predicted an input in a specific way.
- **Layer-wise vs. Neuron-level Interpretations**: Techniques vary from analyzing specific neurons to interpreting whole layers, depending on the desired level of detail.

Popular XAI Techniques: SHAP

- **SH**apley **A**dditive ex**P**lanations (SHAP)
  - **Global** explanation technique
  - Assigns **importance scores** to features based on cooperative game theory.
  - SHAP can be adapted for neural networks, although it is **computationally** demanding

Popular XAI Techniques: SHAP



Popular XAI Techniques: SHAP

Fig. 3 Series of product images of the class “good”

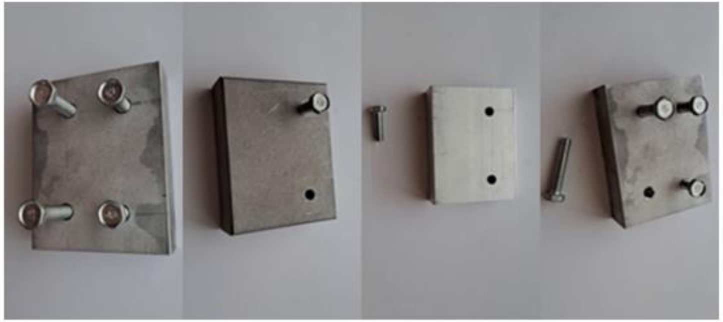


Fig. 4 Series of product images of the class “defect”

Approach to provide interpretability in machine learning models for image classification

https://doi.org/10.1007/s44244-023-00009-z

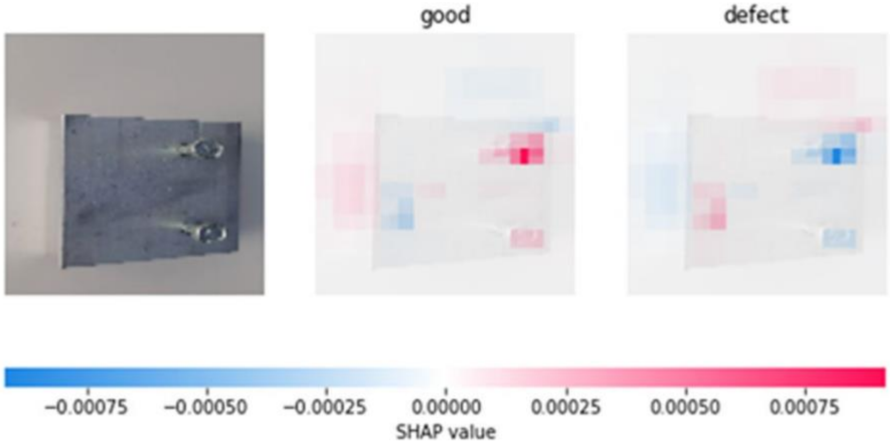


Fig. 11 SHAP interpretation image

Popular XAI Techniques: LIME

- **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations (LIME)
  - A local explainer to extract explanations on the **instance** level
  - Generates **synthetic** data randomly and replaces some of the original data with synthetic data.
  - Repeating the process provides insight into the importance of different features and their contribution towards the final outcome.

Popular XAI Techniques: LIME

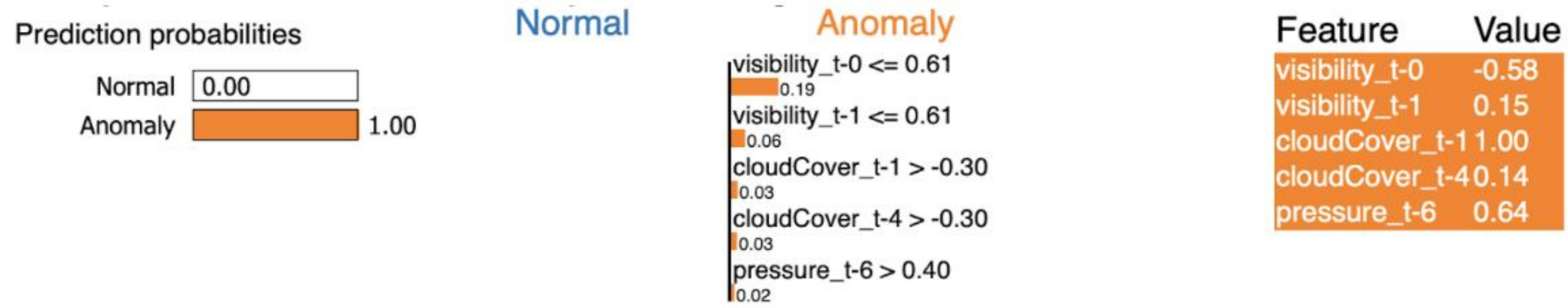


Figure 9: LIME explanations of ‘anomaly’ class for the instance no. 100

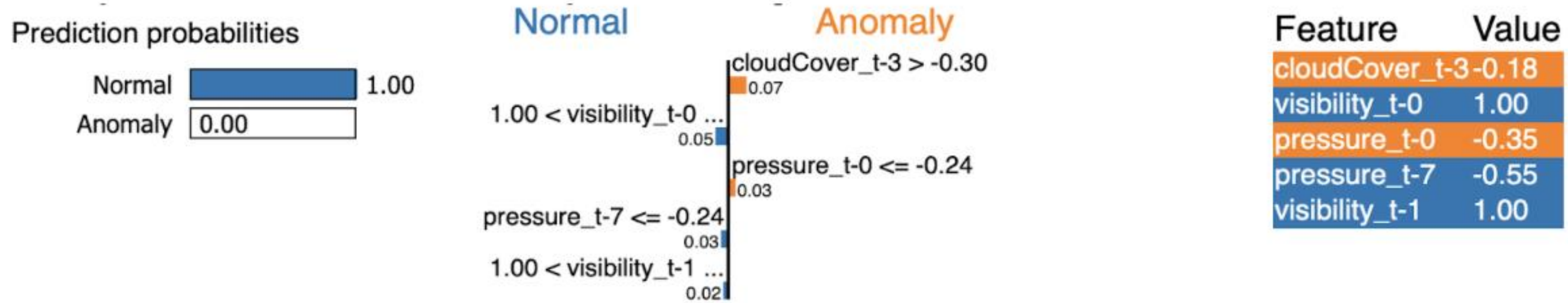
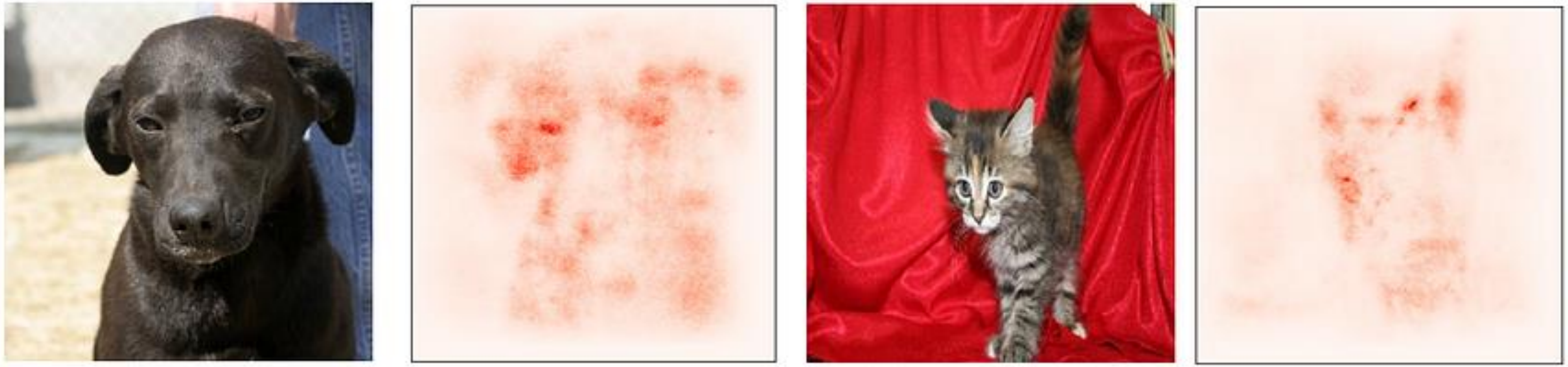


Figure 10: LIME explanations of ‘anomaly’ class for the instance no. 200

Popular XAI Techniques: Saliency Maps

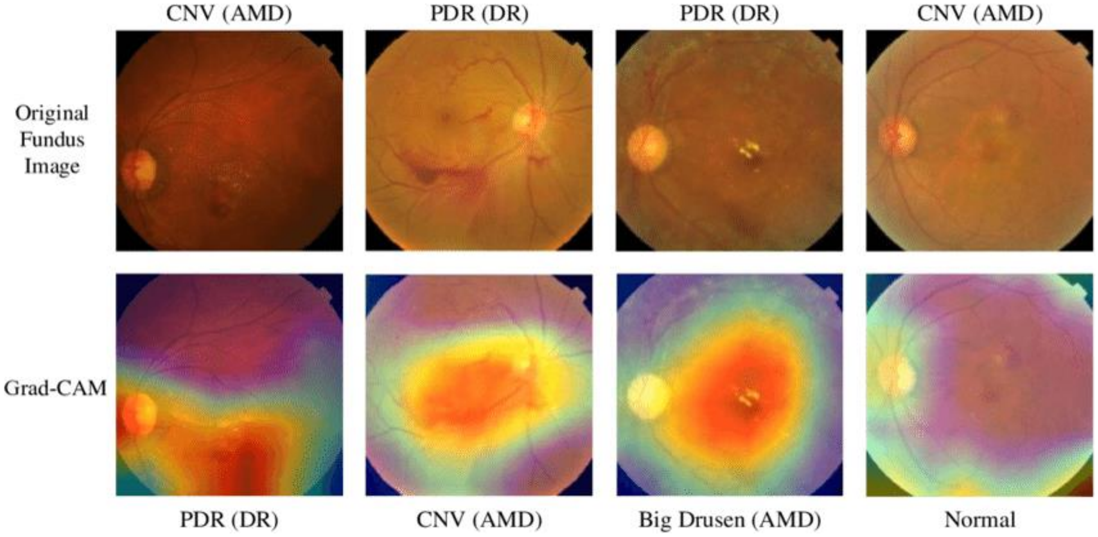
- Or "Pixel Attribution"
  - Highlight the input data (e.g. pixels) that were relevant for a certain image classification by a NN



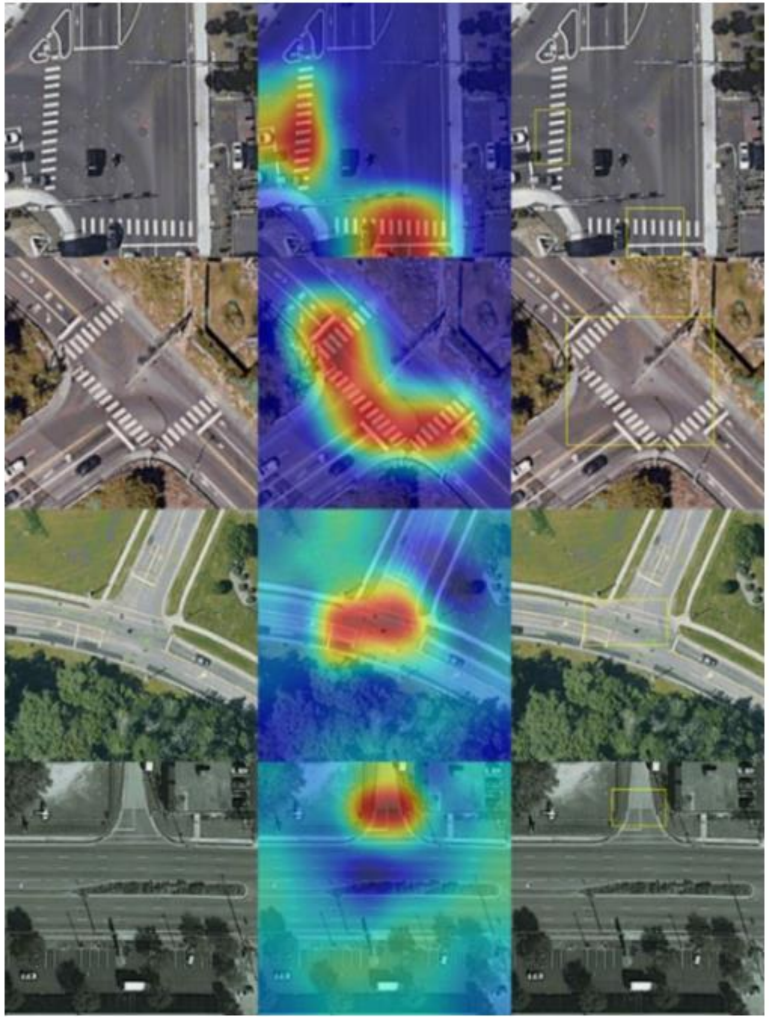
https://medium.com/@bijil.subhash/explainable-ai-saliency-maps-89098e230100
https://christophm.github.io/interpretable-ml-book/pixel-attribution.html

Popular XAI Techniques: Grad-CAM

- **Grad**ient-weighted **C**lass **A**ctivation **M**apping
  - Shows areas in images that strongly contribute to the network's prediction by highlighting the most important regions in the input



https://medium.com/@divakar1591/grad-cam-gradient-weighted-class-activation-mapping-8e1aeaf96d94



Popular XAI Techniques: Grad-CAM

Leveraging Gradient Weighted Class Activation Mapping to Improve Classification Effectiveness: Case Study in Transportation Infrastructure Characterization

[IS&T | Library]

Figure 1. Examples of different crosswalk markings [13].

Popular XAI Techniques: Grad-CAM

Visualization of Facial Attractiveness Factors Using Gradient-Weighted Class Activation Mapping

PREPRINT

Figure 2: Result of visualization of averaged male images using Grad-CAM.

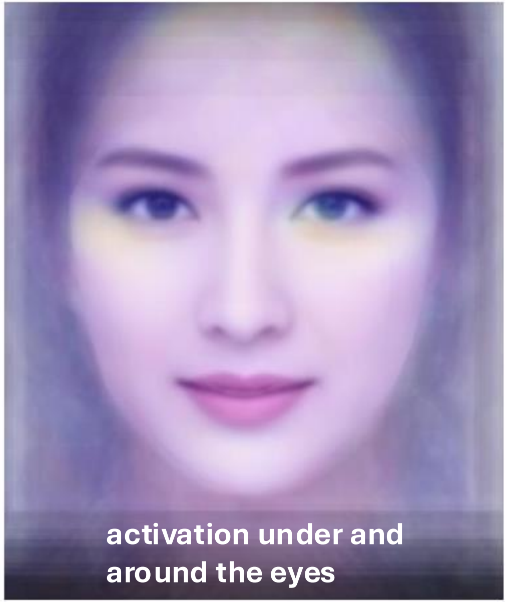
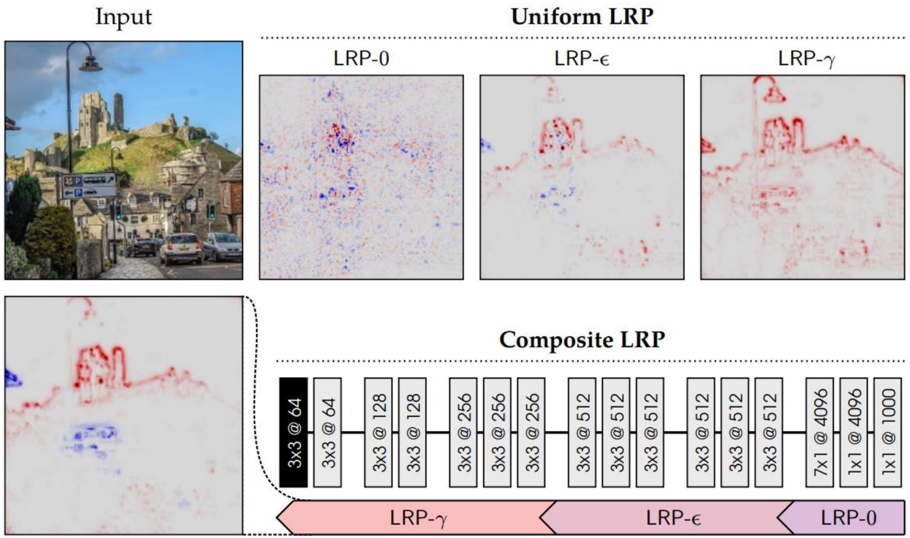


Figure 3: Result of visualization of averaged female images using Grad-CAM.

Popular XAI Techniques: LRP

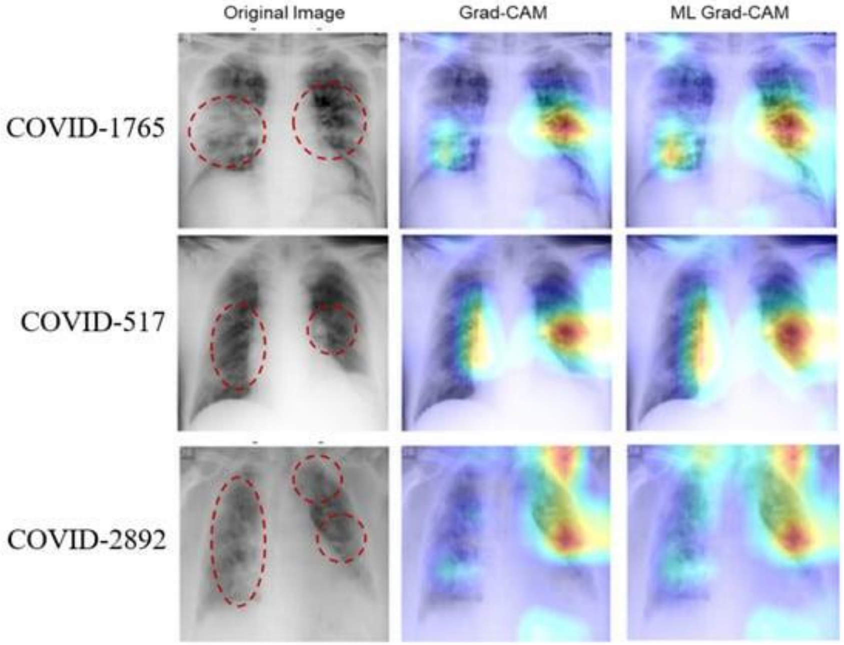
- **L**ayer-wise **R**elevance **P**ropagation
  - Method for NN
  - Traces back the relevance of the prediction to each input feature, layer by layer



https://iphome.hhi.de/samek/pdf/MonXAI19.pdf

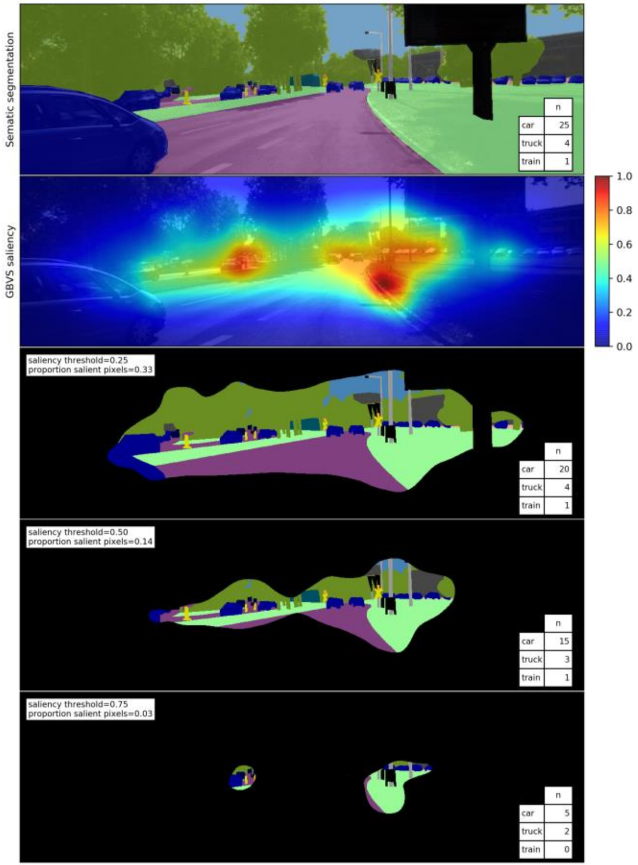
Application of XAI Techniques

- Healthcare – radiology
  - Identify areas in medical images, like X-rays, that influence diagnoses with Grad-CAM



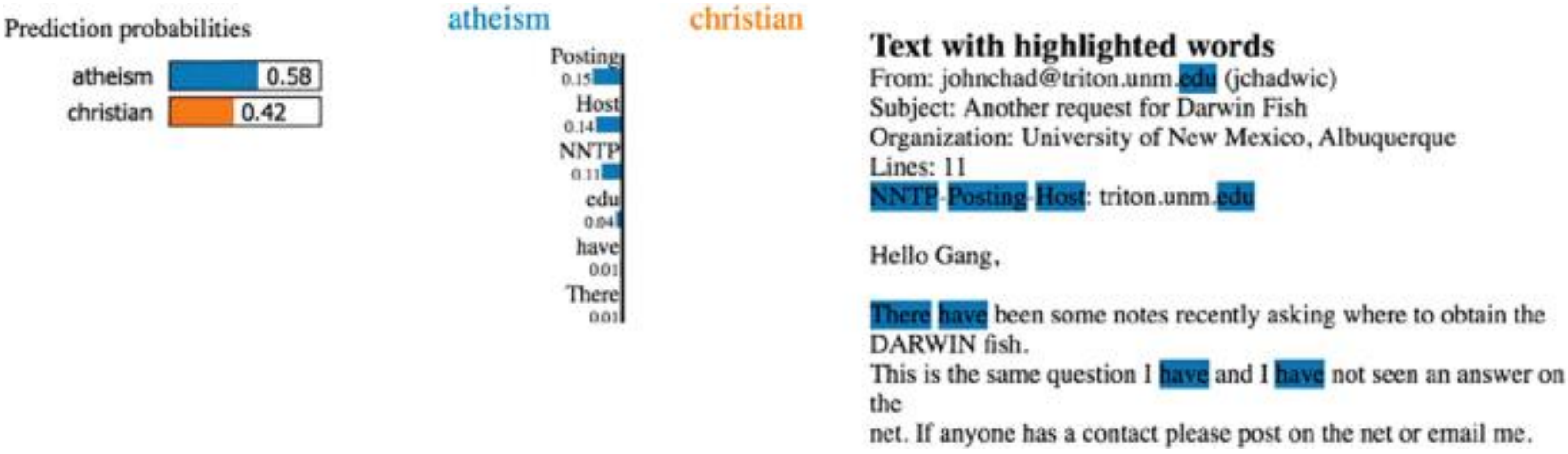
Application of XAI Techniques

- Autonomous driving
  - Saliency maps can reveal which areas of an image the model emphasizes to make safe driving decisions.



Application of XAI Techniques

- Natural Language Processing (NLP)
  - Highlight words or phrases that heavily influence the network's decision



Advantages

- Model debugging
  - Helps identify bias and overfitting
- Example: Husky classified as wolf due to snow in the background

(a) Husky classified as wolf (b) Explanation

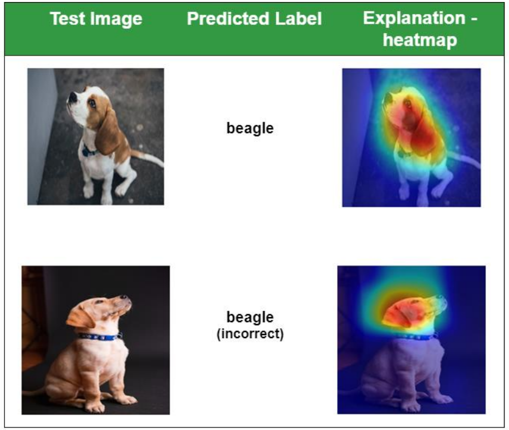
Source: Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Advantages

- Better decision-making
  - Gives an understanding of feature importance
- Faster optimization
  - Model transparency makes it easier to find areas of improvement
- Raise trust
  - With clear interpretations, users can better trust and understand model predictions

Limitations

- Model complexity
  - Complex models can be inherently difficult to explain
- Example: XAI sometimes give similar explanation for correct and incorrect guess



Limitations

- Interpretation bias
  - Interpretation can be biased by human preconceptions or expectations
- Computational demands
  - Some techniques are computationally expensive, making them less available for certain applications

Future Directions

- Interpretable Neural Network Architectures
  - Designing networks with built-in interpretability
  - Examples: Using attention layers or self-explaining models to ensure each layer produces interpretable outputs.
- Hybrid Models with XAI-Friendly Features
  - Combining neural networks with other models for improved interpretability without compromising accuracy.
- Real-time Explanation Methods
  - Focusing on rapid, real-time explanation capabilities.
  - Crucial for time-sensitive applications like real-time medical imaging and autonomous vehicles

Source: Ibrahim et al. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. 2023. https://doi.org/10.1145/3563691

Conclusions

- **Importance of Explainability**: XAI is crucial for ensuring transparency, trust, and accountability in neural networks, especially in high-stakes fields
- **Challenges and Solutions**: Although neural networks' complexity and non-linearity make interpretation difficult, techniques like SHAP, LIME, and Grad-CAM provide valuable insights into model behavior and decision-making processes.
- **Advantages and Limitations**: XAI also presents challenges, such as potential interpretation biases and high computational costs. Addressing these limitations is vital for broadening XAI's accessibility and reliability.