NDSU NORTH DAKOTA STATE UNIVERSITY

UNIVERSIDAD DE CHILE

Using Explainable AI on a Cyberattack Classification Model

By Will Ma

Introduction

When creating ML models, developers measure their performance with metrics such as accuracy that tend to hide bias against minority classes or metrics such as F1 score that are unintuitive to many. Explainable AI (XAI) [1] addresses this problem by creating intuitive explanations of models which reveal features that the model is using to make decisions. Xplique is an XAI toolkit with attribution methods, attribution metrics, concepts extraction capabilities, and feature visualization capabilities.

Related Work

Before Xplique was created, there were already some other XAI toolkits with different attribution methods and/or ML library interfaces. These include:

Captum
- Wide range of tools ☑
- TensorFlow compatible ✗

DeepExplain
- Wide range of tools ✗
- TensorFlow compatible ☑

Alibi
- Wide range of tools ☑*
- TensorFlow compatible ☑

*Xplique does however provide different tools

Selected attribution methods

•**Saliency:** This method computes the gradient of the target class with respect to the input image (or input features). This indicates the sensitivity of the output to changes in each pixel.
•**GradientInput:** This method computes the gradient of the model's output with respect to the input features, indicating how small changes in each input feature affect the output. This gradient is then multiplied by the input feature values to highlight the most influential features. This method is not as readable because it provides a lower-level explanation.
•**GuidedBackProp:** This method works similarly to GradientInput, except it does not include negative gradients; it only applies ReLU to all gradients. It is useful for understanding which pixels are most important. This method provides low resolution but is slightly easier to interpret than GradientInput.

References

1. Fel, T., Hervier, L., Vigouroux, D., Poche, A., Plakoo, J., Cadene, R., Chalvidal, M., Colin, J., Boissin, T., Bethune, L., Picard, A., Nicodeme, C., Gardes, L., Flandin, G., & Serre, T. (2022). Xplique: A Deep Learning Explainability Toolbox. *Workshop on Explainable Artificial Intelligence for Computer Vision (CVPR)*.

2. Dadkhah, S.; Carlos Pinto Neto, E.; Ferreira, R.; Chukwuka Molokwu, R.; Sadeghi, S.; Ghorbani, A. CICIoMT2024: Attack Vectors in Healthcare devices-A Multi-Protocol Dataset for Assessing IoMT Device Security. *Preprints* **2024**, 2024020898. https://doi.org/10.20944/preprints202402.0898.v1

CICIoMT2024 Dataset

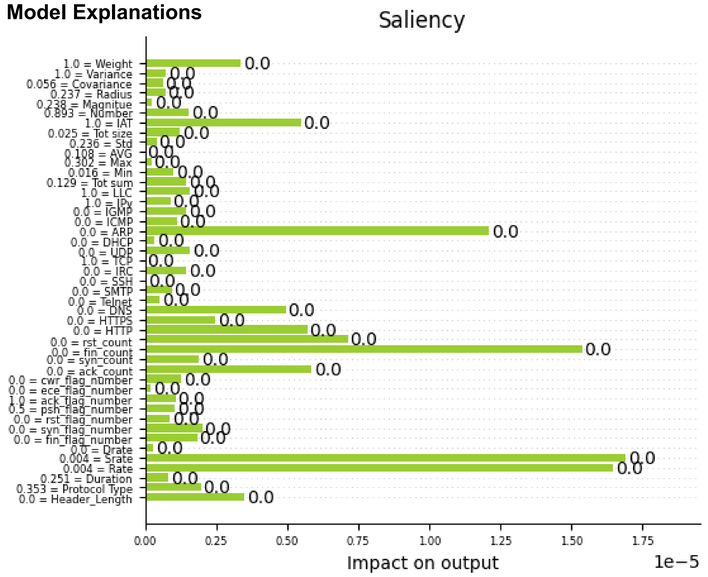
The CICIoMT2024 [2] dataset was developed by executing 18 different types of cyberattacks as well as benign traffic on a testbed of 40 IoMT devices, including both real and simulated devices. This dataset includes 44 features related to internet traffic. The researchers applied four machine learning algorithms (Logistic Regression, Adaboost, Random Forest, and Deep Neural Networks) to three classification tasks with varying levels of specificity (2 classes, 6 classes and 19 classes). The random forest algorithm performed the best with scores in the mid to high 0.90s across every task except the nineteen class recall for which it scored ~0.89 thus bringing down the F1 score for that task to ~0.91.

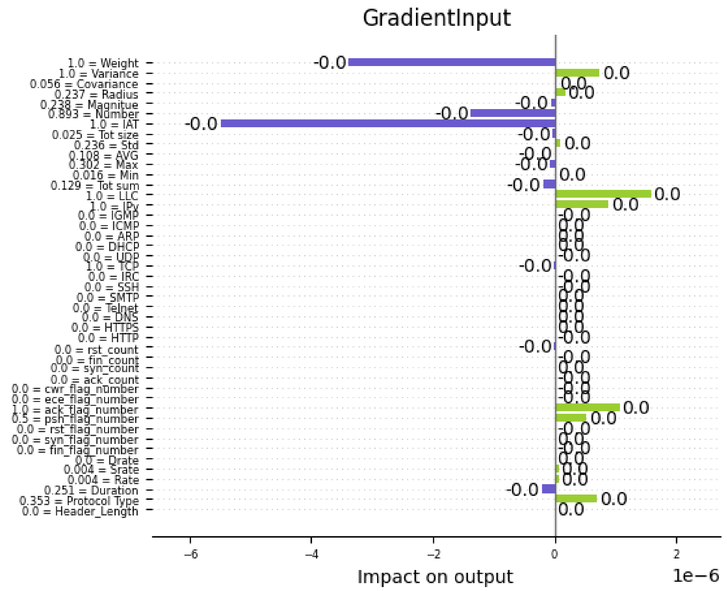
Explaining a Deep Neural Network

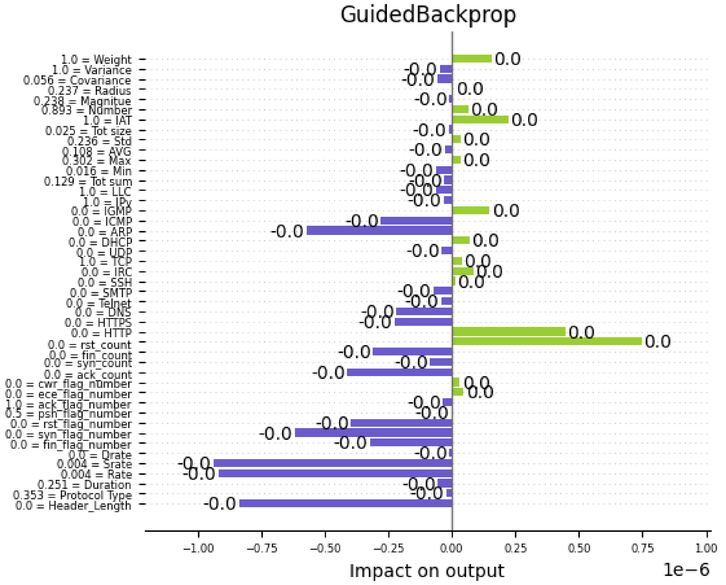
```python
# Build the DNN model
model = Sequential()
model.add(Dense(128, input_dim=len(feature_columns), activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(32, activation='relu'))
model.add(Dense(len(target_columns), activation='softmax'))

# Compile the model
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

Model Performance
- Accuracy - 0.73
- Precision - 0.55
- Recall - 0.54
- F1 - 0.48

Model Explanations







Conclusion

From any one of these graphs you might draw insights as to which features have the most impact on the model's predictions, however when comparing them, it becomes difficult to draw out any sort of pattern. For example, the saliency attribution method ranks Rate and Srate as the most impactful features while guided backpropagation ranks them having the highest negative impact on the prediction and gradient input ranks them as having a minimal positive input. Thus, for this model and the other models we created and received the same mixed explainability results on, the Xplique results do not provide a clear understanding of how the model is making its predictions.