Motivation
- Malware (malicious software) remains the most popular and damaging attack vector, costing hundreds of billions in damage.
- Malware evolves rapidly, with reports showing that 99% disappear after 58 seconds.
- Traditional machine learning models heavily depend on feature engineering and could be easily deceived by hackers.
- Anti-virus industry prefer to increase the recall (i.e., true positive rate) while maintaining a low false positive rate (usually less than 0.1%).

Contribution
- Proposed an end-to-end malware detection framework based on deep learning techniques, which achieves the best performance among existing deep learning based methods.
- Proposed an effective loss function for optimizing recall with a fixed tiny false positive rate.
- Conducted experiments on a real large dataset to confirm the effectiveness of the proposed feature learning framework and loss function for malware detection.

Model Architecture
- Feature Learning:
  - Two separate structures for processing PE header and PE sections
- Classification:
  - Neural Decision Trees and logistic regression
  - A special loss function for optimizing the recall given a fixed false positive rate

Header Feature Extraction
- Input:
  - Raw byte sequence of the PE header
- Embedding layer:
  - Embeds the raw bytes into a continuous and distributed representation
- Gated Convolution layer:
  - \( X_c \in \mathbb{R}^{C \times L} \) provides a mechanism to learn, select and pass along the important and relevant information
- Global Max-pooling layer:
  - Produces the header feature regardless of the location of the detected features

Neural Decision Trees
- A differentiable version of decision tree, enabling end-to-end training and reducing overfitting.
- Two ensemble techniques are designed for Neural Decision Trees:
  - Neural Random Forest
  - Neural Gradient Boosting Decision Tree
- Logistic regression is applied on all the outputs of the decision trees, providing a more flexible way of utilizing the generated trees.

Loss Function Optimization
- To maximize recall rate with the restriction that false positive rate \( \leq 0.1\% \).

\[
L = \sum_{i \in Y^{-}} \left( \alpha y_i y_i^* \right) + \max \left( 0, \lambda \frac{\sum_{i \in Y^{-}} f ( x, y_i ) - \alpha}{|Y^{-}|} \right)
\]

Data Summary
- SecureAge deployed 12 commercial antivirus engines that are continuously scanning data from the endpoints.
- Positive: num of engines > 4
- Negative: num of engines = 0

Experiment Results
- Training Dataset: Text Dataset: Model: Without optimized loss function: With optimized loss function: 
  - Loss for positive samples: Loss for negative samples:

ROC Curves Comparison
- ROC curves of MaConv
- ROC curves of ComNet
- ROC curves of EntropyNet
- ROC curves of Proposed Model

Experiments
- Comparison of different datasets on different models
- Evaluation of the proposed model against existing models