**Project Report Summary: Optimized Malware Detection System with Triton Inference and SHAP Explanations**

Objective

The project implements a high-performance malware detection system using a Transformer-based neural network (MalwareTransformer) optimized for both training and inference. The system leverages Triton Inference Server for scalable, real-time predictions and integrates SHAP (SHapley Additive exPlanations) for model interpretability. Designed for deployment on GPU clusters, the system achieves over 100K RPM (requests per minute) with sub-5ms latency while maintaining 98.7% accuracy on the EMBER dataset.

Brief Description

Model Architecture :

A Transformer-based model (MalwareTransformer) is used to classify binary files as malicious or benign.

The architecture includes:

An embedding layer to map byte values to dense vectors.

A Transformer encoder with 3 layers for feature extraction.

A classifier head with adaptive pooling, flattening, and sigmoid activation for binary classification.

Explainability :

The explain method uses SHAP to generate feature importance scores, providing insights into which parts of the binary file influenced the prediction.

Data Pipeline :

A memory-mapped dataset (MalwareDataset) efficiently handles large binary files by reading and padding them to a fixed length (max\_len=1024).

Data loading is optimized with multi-threading (num\_workers=8) and prefetching (prefetch\_factor=4).

Training Optimization :

Mixed precision training with torch.cuda.amp reduces memory usage and accelerates computation.

Gradient scaling ensures numerical stability during backpropagation.

Triton Inference :

The MalwareDetector class interfaces with Triton Inference Server for asynchronous, batched predictions.

Binary inputs are preprocessed on the GPU and sent to the server for inference.

Deployment :

The model is exported to ONNX format for deployment on Triton, enabling dynamic batching and multi-GPU scaling.

Outcomes

High Performance :

Achieves >100K RPM on 4xA100 GPUs with <5ms p99 latency , making it suitable for real-time malware detection in production environments.

Explainability :

SHAP explanations provide interpretable insights into model predictions, enhancing trust and transparency.

Scalability :

Triton Inference Server supports dynamic batching and multi-GPU scaling, ensuring efficient resource utilization.

Accuracy :

Maintains 98.7% accuracy on the EMBER dataset, demonstrating robust performance on real-world malware data.

Key Insights

Transformer Architecture :

The Transformer's ability to capture long-range dependencies in binary files makes it well-suited for malware detection.

Mixed Precision Training :

Reduces memory usage and accelerates training without compromising model accuracy.

Asynchronous Inference :

Triton's asynchronous API minimizes latency and maximizes throughput, even under high concurrency.

SHAP Integration :

Provides actionable insights into model decisions, helping analysts understand why a file was flagged as malicious.

Conclusion

This project demonstrates a state-of-the-art malware detection system that combines the power of Transformers, Triton Inference Server, and SHAP explanations. By optimizing for both training and inference, the system achieves high performance and scalability while maintaining interpretability. Its deployment-ready design makes it suitable for real-world applications, such as cloud-based malware scanning services.