BiST-Mamba: A Dual-branch Spatio-Temporal Mamba Network for Encrypted Traffic Classification (Student Abstract)

Tongle Zhao [1], Fang Fan[1], Huiqi Zhao[1], Xiaodu Liu[2],

[1]School of Intelligent Equipment, Shandong University of Science and Technology, Tai'an, Shandong, China
[2]Information Center, China Association for Science and Technology, Beijing, China
Zhao.TL@sdust.edu.cn, fangfan@sdust.edu.cn, zhaohq@sdust.edu.cn, liuxiaodu@cast.org.cn

Abstract

Encrypted traffic classification has become increasingly important in network security. To address the difficulty of existing architectures in collaboratively modeling spatio-temporal features, we propose **BiST-Mamba**, a novel dual-branch spatio-temporal Mamba network that enables simultaneous representation of spatio-temporal features. To the best of our knowledge, this is the first work to introduce VMamba into encrypted traffic classification. Preliminary experiments on a small-scale dataset show that our accuracy and F1 scores reach **94.13%** and **93.41%**, respectively. The method achieves promising classification performance, demonstrating the potential of the model for effective spatio-temporal modeling.

Introduction

With the increasing application of encryption technologies, encryption not only conceals the payload content of traffic information but also weakens the effectiveness of traditional feature extraction methods. Therefore, current research on encrypted traffic classification mainly focuses on directly learning effective representations from raw traffic data (Lotfollahi et al. 2017).However, raw encrypted traffic has dual attributes: on the one hand, it reflects the spatial structural features embedded within packets/flows; on the other hand, it exhibits long-range temporal dependencies across packet sequences.

Existing methods for encrypted traffic classification often struggle to achieve an effective balance between spatial modeling and temporal modeling. Spatial feature extractors are limited in capturing long-range temporal dependencies, while temporal models tend to ignore fine-grained spatial structures. Therefore, relying on a single type of architecture makes it difficult to simultaneously satisfy the dual requirements of encrypted traffic classification: sensitivity to spatial structures and the ability to model long-range temporal dependencies. With the further development of deep learning, an efficient sequence processing technique based on selective state space models (such as Mamba (Gu and Dao 2023) and VMamba (Liu et al. 2024)) has entered the research field. Mamba, with its linear computational complexity, demonstrates the ability to efficiently handle long-range dependencies in long-sequence modeling;while VMamba successfully introduces this advantage into the vision domain, enabling efficient unified modeling of both local details and global contexts in images.

Based on the above observations, to ensure that both spatial and temporal features of encrypted traffic can be effectively represented, we propose BiST-Mamba, a dual-branch spatio-temporal modeling architecture for encrypted traffic classification. Our main contributions are as follows: BiST-Mamba — We establish a unified spatio-temporal feature learning framework based on unidirectional Mamba, which can effectively capture both spatial structures and long-range temporal dependencies in a collaborative manner;For the first time, we introduce VMamba into the field of encrypted traffic classification, improving the ability to understand fine-grained spatial patterns.

Methodology

BiST-Mamba, whose core architecture is shown in Figure 1, consists of the following key components:

Preprocessing. The raw network traffic data (in PCAP format) first undergoes a preprocessing pipeline, which includes session segmentation based on the five-tuple, filtering to remove non-IP protocol packets, deduplication, anonymization (removal of MAC and IP address information), byte extraction, and length normalization. Each session is then transformed into two parallel representations: a two-dimensional grayscale image as the input to the spatial branch, and the original one-dimensional sequence as the input to the temporal branch. This process aims to prepare standardized and comparable input data for subsequent dual-branch collaborative modeling.

Dual-Branch. This module consists of a spatial branch and a temporal branch. **The spatial branch** takes a two-dimensional grayscale image as input, first applying Patch Partition to divide the image into small patches for local modeling. The input is then processed by the Vision Selective Scan (VSS) Block to extract basic local spatial features, followed by Downsampling to gradually reduce the feature dimension, while deeper VSS Blocks simultaneously capture local details and global context.In this study, the VMamba architecture was modified to adopt a three-stage design with an overall downsampling ratio of 1/16, providing better adaptation to the 64×64 grayscale input charac-

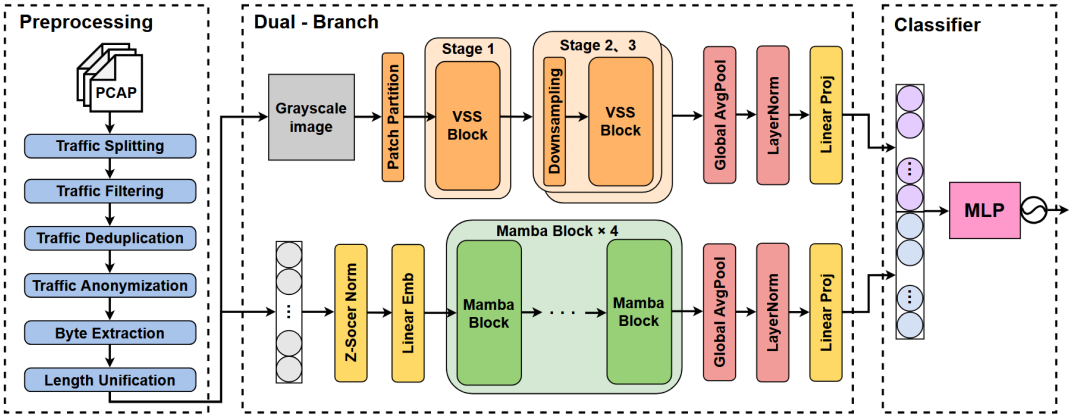

Figure 1: The overall architecture of BiST-Mamba.

Method	Acc	Pre	Rec	F1
FS-Net	92.94	91.73	91.15	91.44
TFE-GNN	93.38	92.91	91.76	92.33
BiST-Mamba	**94.13**	**93.85**	**92.97**	**93.41**

Table 1: Performance comparison on our dataset (%)

teristics of encrypted traffic. **The temporal branch** directly processes the raw one-dimensional byte sequence, where Z-Score Norm is first applied for standardization, and Linear Embedding maps the sequence into an appropriate feature space. Its core component is a stack of four Mamba Blocks, which leverage the Selective State Space Model (S6) to effectively capture long-range temporal dependencies. Finally, the outputs of both branches are separately passed through their respective feature aggregation and projection layers to generate spatial and temporal feature vectors. In addition, the two branches adopt independent pre-training and joint fine-tuning strategies to improve learning efficiency and generalization ability.

Classifier. We employ direct feature concatenation to integrate the representations from the two branches, forming a comprehensive representation that describes encrypted traffic. The fused vector is processed by a multilayer perceptron (MLP) head, which consists of dense layers with the GELU activation function, followed by a softmax output layer that generates the normalized probability distribution over the target traffic classes.

Experiments

To evaluate the effectiveness of the BiST-Mamba model, we conducted experiments on the ISCXVPN2016 dataset, which contains VPN encrypted traffic from 14 application categories. The data preprocessing procedure followed the same pipeline described in Section 2 to ensure consistency and comparability.

The model training followed a two-stage paradigm of independent pretraining and joint fine-tuning. In the first stage, the spatial (VMamba) and temporal (Mamba) branches were trained separately to obtain stable feature representations. In the second stage, the pretrained branches were jointly optimized end-to-end on paired samples to achieve cross-modal feature fusion and collaborative learning. This paradigm helped improved optimization stability and the complementarity of spatio-temporal features. All training used the AdamW optimizer with cosine-annealing learning rate scheduling (five-epoch warm-up), a batch size of 64, 50 training epochs, a weight decay coefficient of 0.01, and a dropout rate of 0.1. Experiments were conducted on two NVIDIA RTX 2080Ti GPUs using mixed-precision training to enhance computational efficiency. To accommodate differences between branches and training stages, the learning rates and freezing strategies were slightly adjusted to ensure sufficient convergence and feature alignment at each stage.

Two representative models were selected for comparison: FS-Net (Liu et al. 2019), which focuses on temporal sequence modeling, and TFE-GNN (Zhang et al. 2023), which emphasizes spatial structural modeling. Four standard metrics—Accuracy, Precision, Recall, and F1-score—were employed to comprehensively evaluate classification performance.

As shown in Table 1, BiST-Mamba outperforms both FS-Net and TFE-GNN across all metrics, achieving 94.13% accuracy and 93.41% F1-score. The two-stage training yields a more stable optimization process and stronger spatio-temporal representation capability. Furthermore, although not shown in the article, the actual ablation experiments also verified the complementary effect of the spatial and temporal branches.

Conclusion and Future Work

BiST-Mamba effectively extracts fine-grained spatial structures and long-range temporal dependencies in encrypted traffic through the collaborative modeling of spatial and temporal branches. Future work will validate the model's generalization ability by expanding the scale and diversity of the dataset, and will optimize BiST-Mamba's architecture and training strategies to enhance its representation capability and overall performance.

References

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Liu, C.; He, L.; Xiong, G.; Cao, Z.; and Li, Z. 2019. FS-Net: A Flow Sequence Network For Encrypted Traffic Classification. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 1171–1179.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. VMamba: Visual State Space Model.

Lotfollahi, M.; Zade, R. S. H.; Siavoshani, M. J.; and Saberian, M. 2017. Deep Packet: A Novel Approach For Encrypted Traffic Classification Using Deep Learning. *CoRR*, abs/1709.02656.

Zhang, H.; Yu, L.; Xiao, X.; Li, Q.; Mercaldo, F.; Luo, X.; and Liu, Q. 2023. Tfe-gnn: A temporal fusion encoder using graph neural networks for fine-grained encrypted traffic classification. In *Proceedings of the ACM web conference 2023*, 2066–2075.