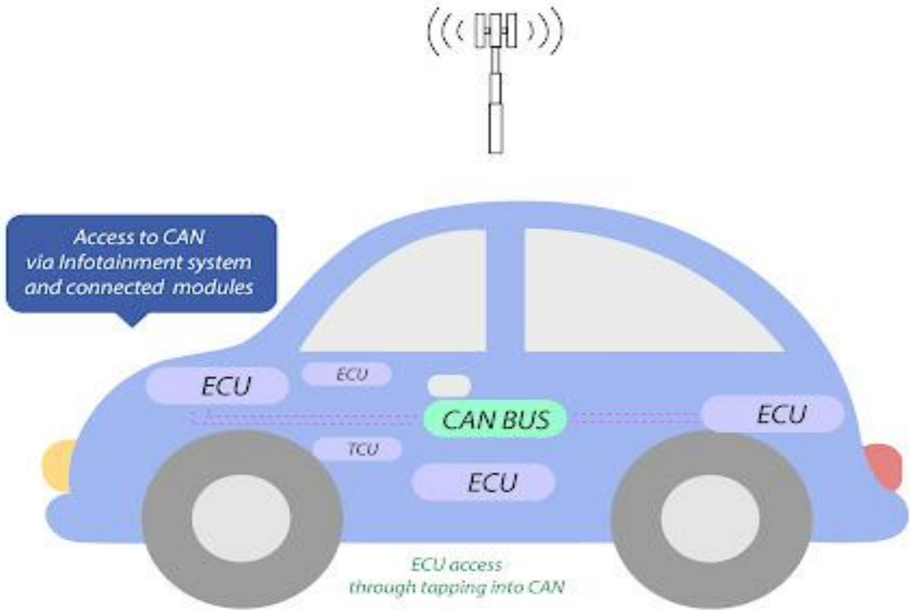
MASS 2021

A Transfer Learning based Abnormal CAN Bus Message Detection System

Liuwang Kang and **Haiying Shen**
Department of Computer Science, University of Virginia

Background

- Hundreds of electronic control units (ECUs) and devices communicate messages in a control area network (CAN) bus
- Modern vehicles become vulnerable to attacks when communicating with outside-vehicle environments



Background

- Vehicle under a CAN bus attack may fail to work and affect vehicle driving safety
- CAN bus message transmission behavior (time interval) is not the same for different vehicle types

Arbitration field						Data field			
SOF	ID	RTR	IDE	r0	DLC	Data	CRC	ACK	EOF
1 bit	11 bit	1 bit	1 bit	1 bit	4 bit	0 to 64 bit	16 bit	2 bit	7 bit

Accurately detecting abnormal CAN bus messages for different vehicle types become important

Related Work

- Some methods [PST'17, CISR'17, ICOIN'16] try to detect abnormal CAN bus messages by statistically analyzing message transmission behaviors

➢ Vehicle driving conditions (e.g., KEY on and KEY start) affect message transmission behaviors greatly

- Some methods [PST'18, PLOS'16] utilize Machine Learning (ML) technologies to detect abnormal CAN bus messages by capturing message features like data field values

➢ Detection performance highly depends on the training data size

Challenges

Propose a neural network based abnormal message detection system (NaDS), to detect abnormal messages in a CAN bus

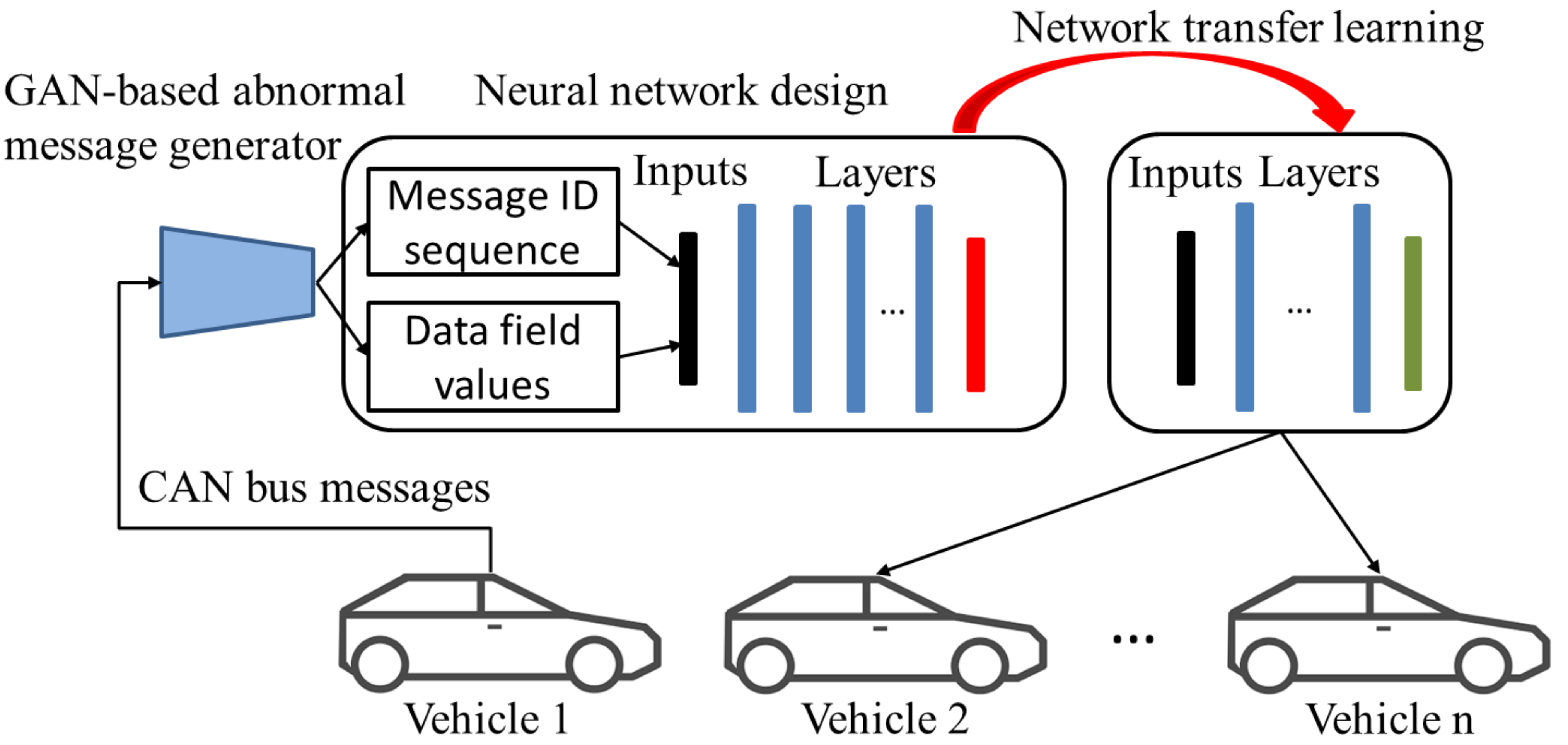
Challenge 1: How to increase accuracy for detecting both known and unknown abnormal messages?

- Difficult to collect sufficient training data including all kinds of attacks

Challenge 2: How to form a well-trained ML model for one vehicle type in spite of a small amount of training data?

- The training data size affects ML model' performance greatly

Neural Network based Abnormal Message Detection System



Challenge 1

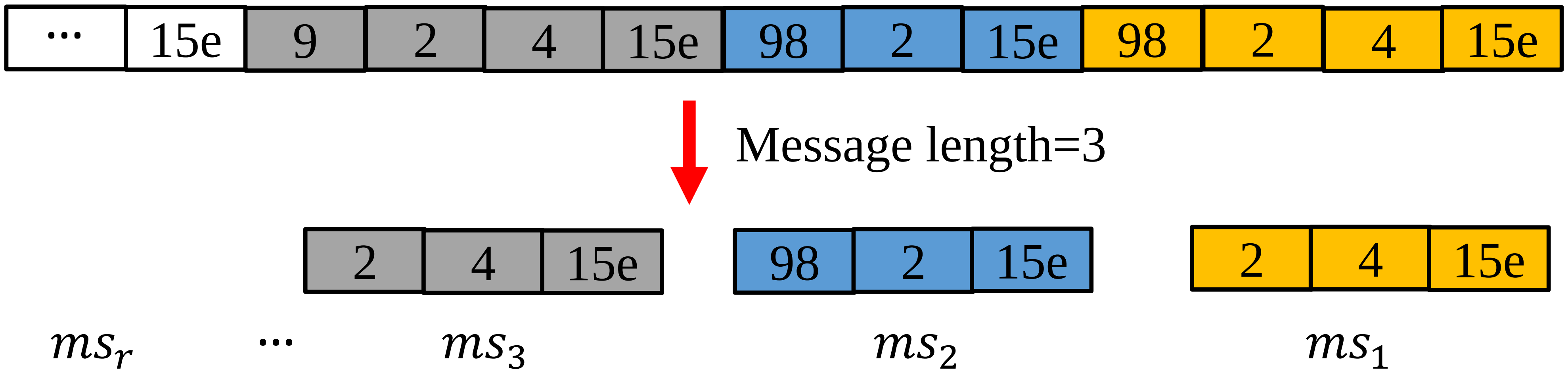
How to increase accuracy for detecting both known and unknown abnormal messages?

Abnormal Message Detection Using Neural Network

Observations from message ID sequence dissimilarity analysis results

Message ID sequence : A series of message IDs from itself to the previous message with the same message ID type

- Step 1: Determine message sequence ms_1 and previous message sequences (ms_2, ms_3, …, ms_r) for a message



Abnormal Message Detection Using Neural Network

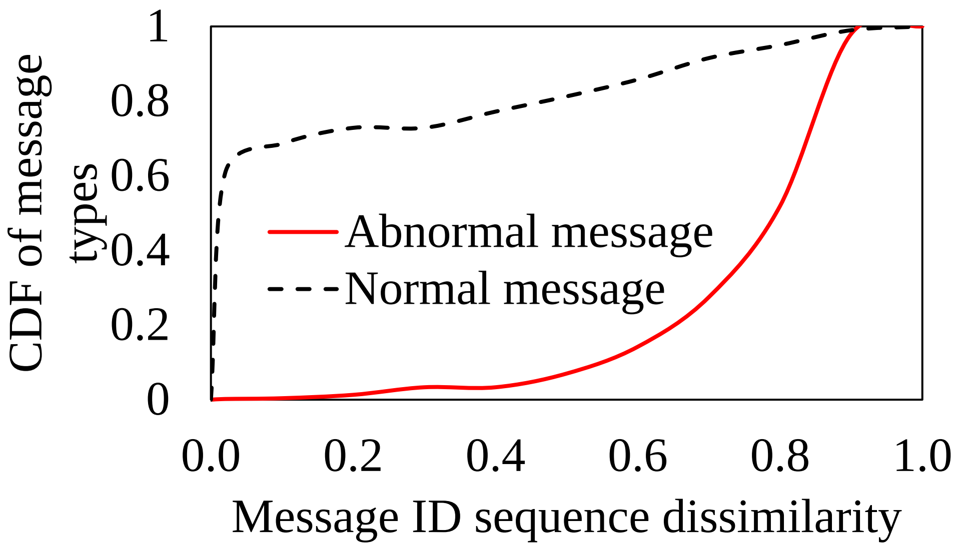
Observations from message ID sequence dissimilarity analysis results

- Step 2: Calculate Hamming distances between ms_1 and $(ms_2, ms_3, \ldots, ms_r)$

$$\mathrm{H}(ms_1, ms_i) = \frac{N_{min}(ms_1, ms_i)}{N_{total}(ms_1, ms_i)}$$

$N_{min}(ms_1, ms_i)$ – the minimum number of changed messages to ensure $ms_1 = ms_i$

$N_{total}(ms_1, ms_i)$ – the total number of messages in ms_1

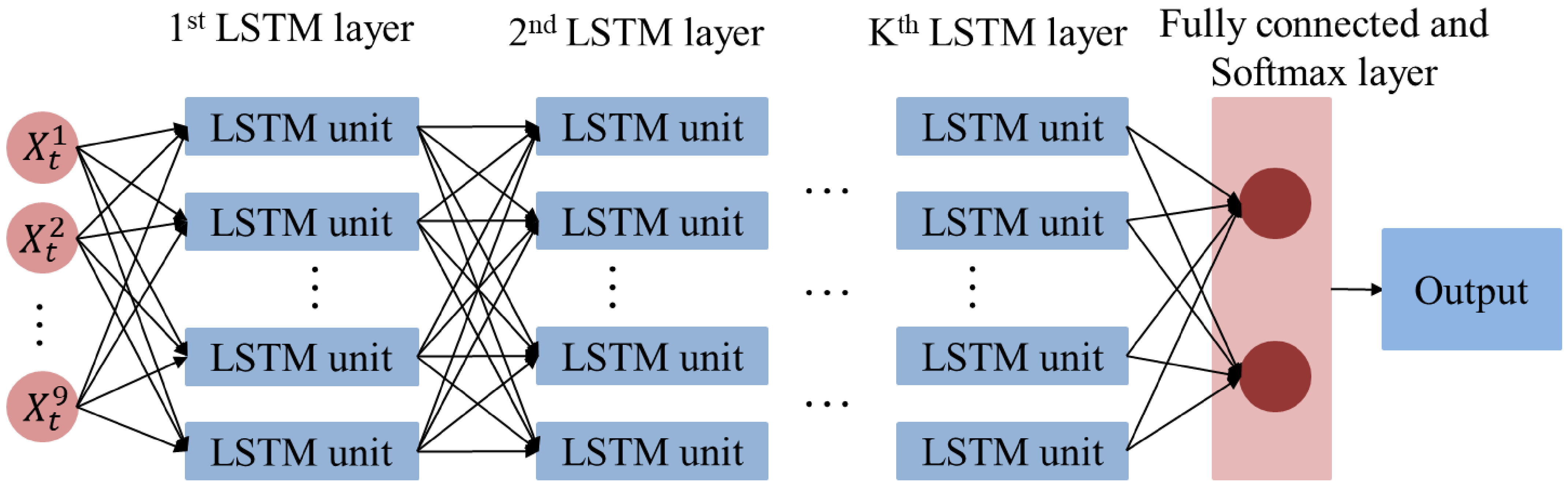


Abnormal messages have much larger message ID sequence dissimilarity value than normal messages

Abnormal Message Detection Using Neural Network

LSTM-NN based abnormal message detection method

Utilize a LSTM neural network to detect abnormal messages by inputting message ID sequence and values in the data filed

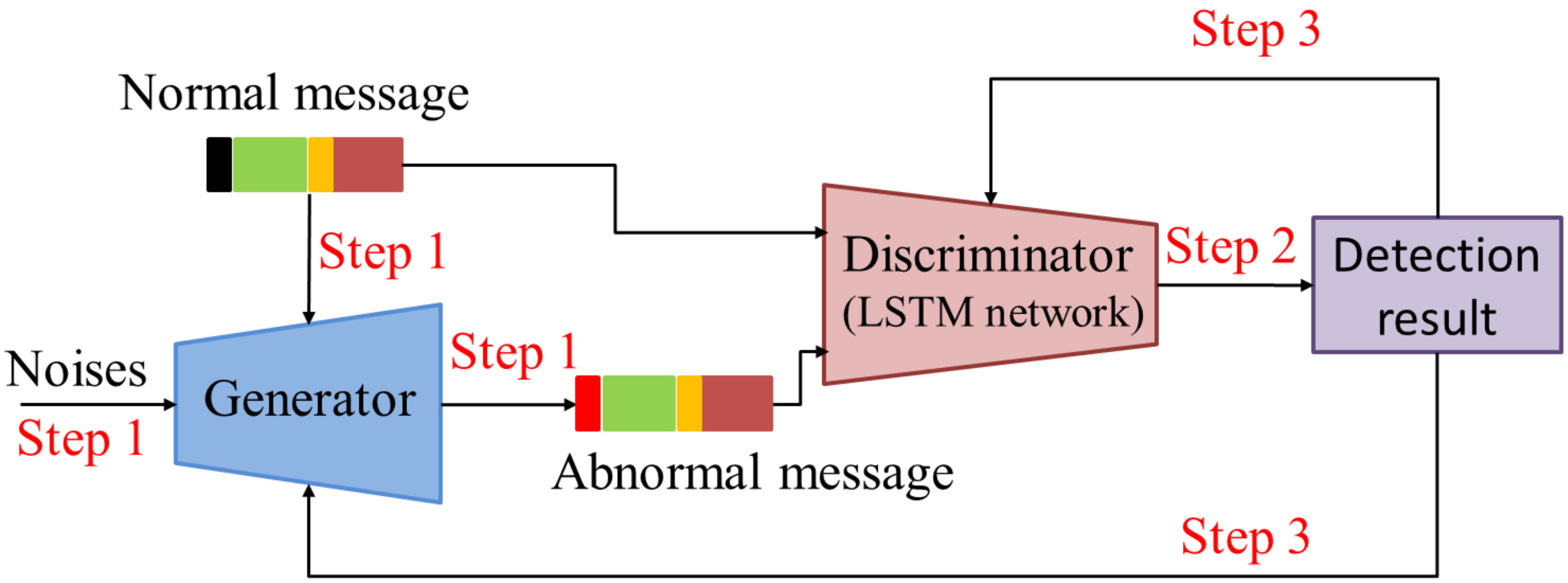


LSTM-NN: Long short-term memory based neural network

Abnormal Message Detection Using Neural Network

GAN-based abnormal message generator

Utilize a GAN to generate all possible abnormal messages for training the LSTM network



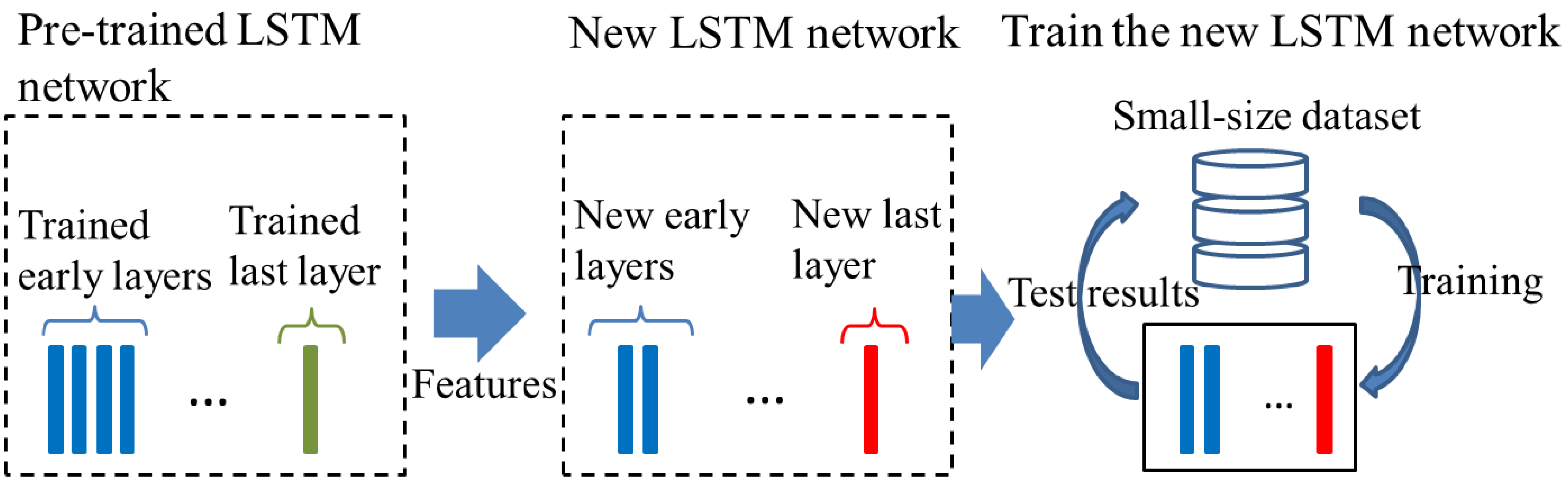
Challenge 2

How to create a well-trained ML model for one vehicle type in spite of a small amount of training data

Transferring a Pre-Trained Detection Model

Transfer learning for LSTM network

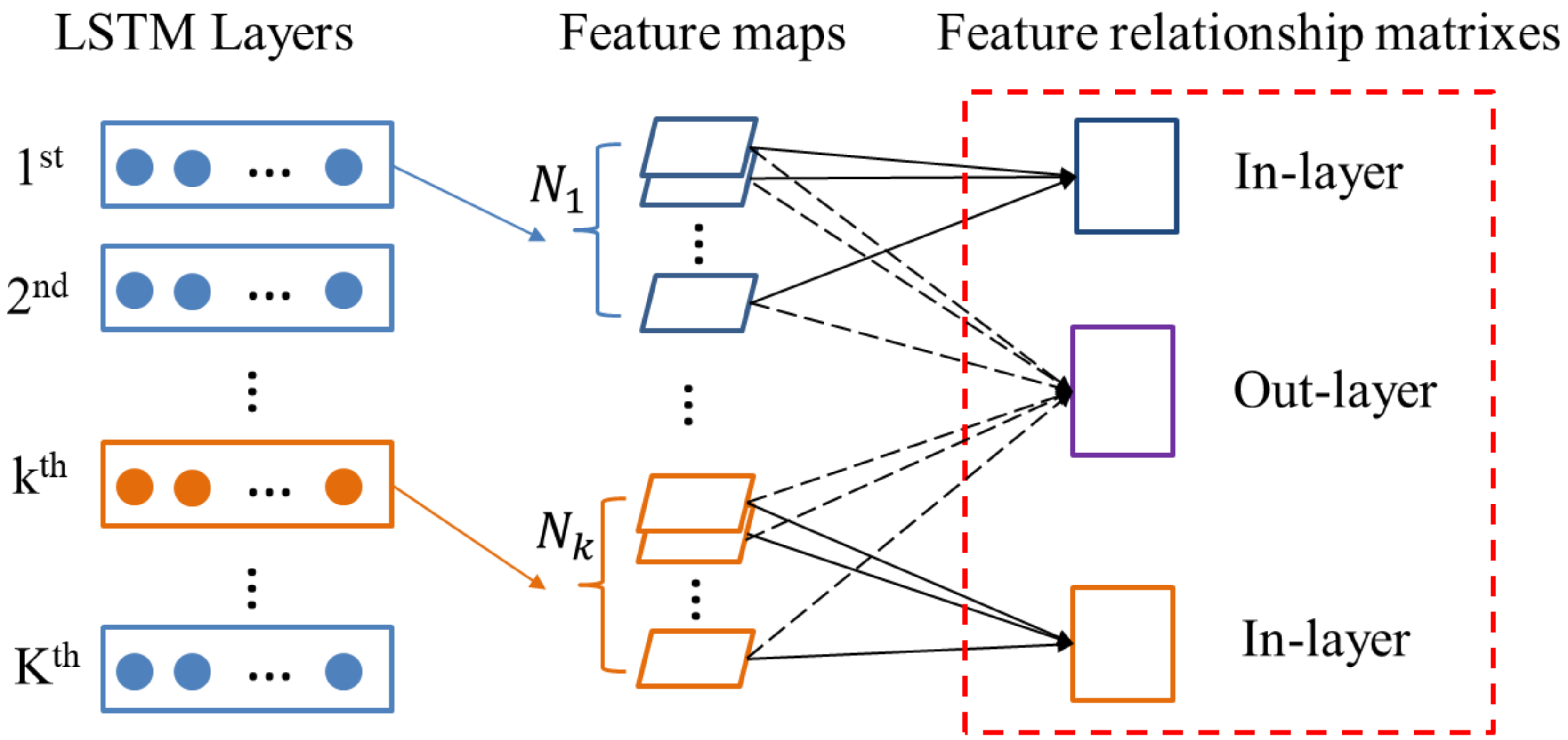
Transfers the pre-trained LSTM network of a vehicle into another LSTM network for detecting abnormal messages for a new vehicle type



Transferring a Pre-Trained Detection Model

Extracting features in LSTM neural network

Uses a feature map to indicate features of units in each LSTM layer and feature relationship matrixes to describe relationships between feature maps in two layers or the same layer



Transferring a Pre-Trained Detection Model

Training the transferred LSTM neural network

Add a fully connected and softmax layer into the transferred LSTM network and train it as follows by minimizing the cross-entropy loss L

$$L = \frac{1}{CS} \sum_{i=1}^{S} \sum_{c=1}^{C} \mathrm{y}(o(s_i) \rightarrow \mathrm{c}) \log(p_c)$$

$\mathrm{y}(\mathrm{o}(s_i) \rightarrow \mathrm{c})$ – indicates a binary indicator and equals to 1 if detection result $\mathrm{o}(s_i)$ on sample s_i is the same as classification status c

p_c - the probability that c is the correct classification status of s_i

Performance Evaluation

Experiment settings

- Implement NaDS by running MATLAB on one laptop (Intel i5 CPU and 16 gigabyte memory)
- Contain 961,723 abnormal messages and 2,747,421 normal messages from three different vehicle types (KIA, SONATA, and SPARK)

Comparison methods

- Message time interval-based detection system (TIDS)[CISR’17], GAN based intrusion detection system (GIDS) [PST’18] and ML based detection system (RLD) [PLOS’16]

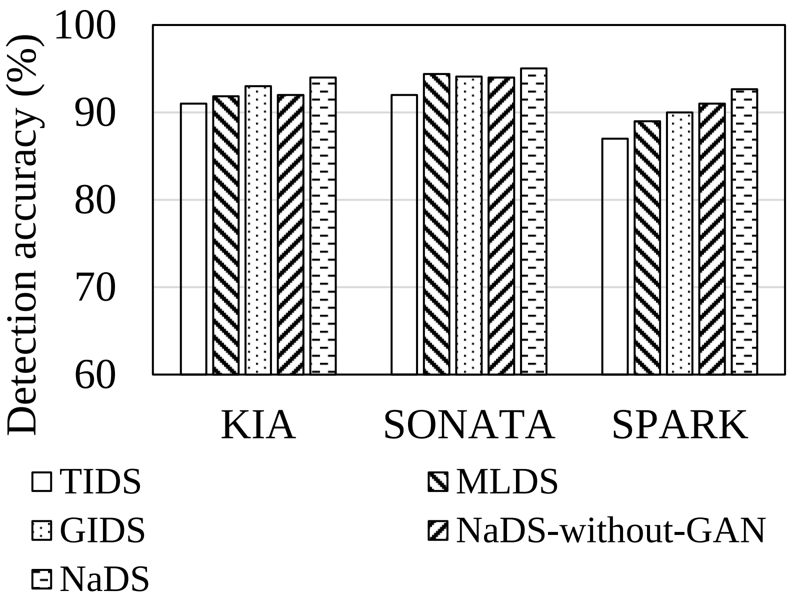
Evaluation metrics

- Abnormal message detection accuracy

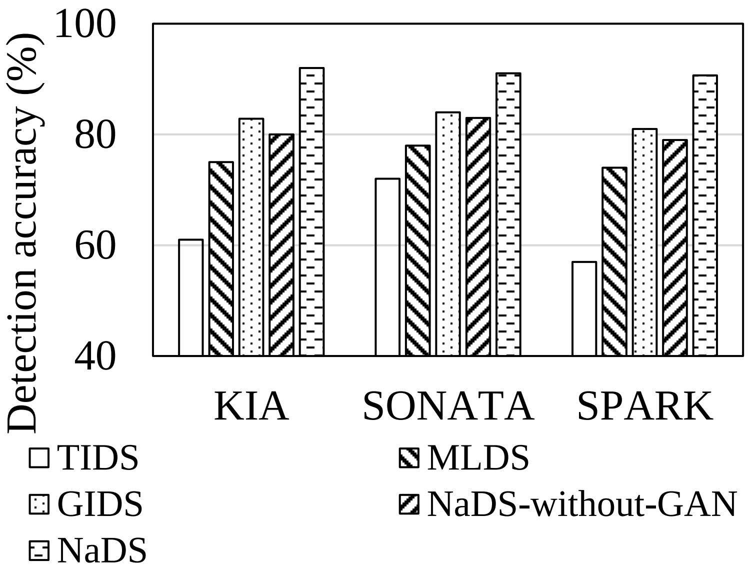
Performance Evaluation

Abnormal message detection accuracy comparisons

- NaDS has the highest detection accuracy on known abnormal messages
- Detection accuracy decreases for unknown abnormal messages and NaDS keeps the maximum detection accuracy value



Accuracies for known abnormal messages

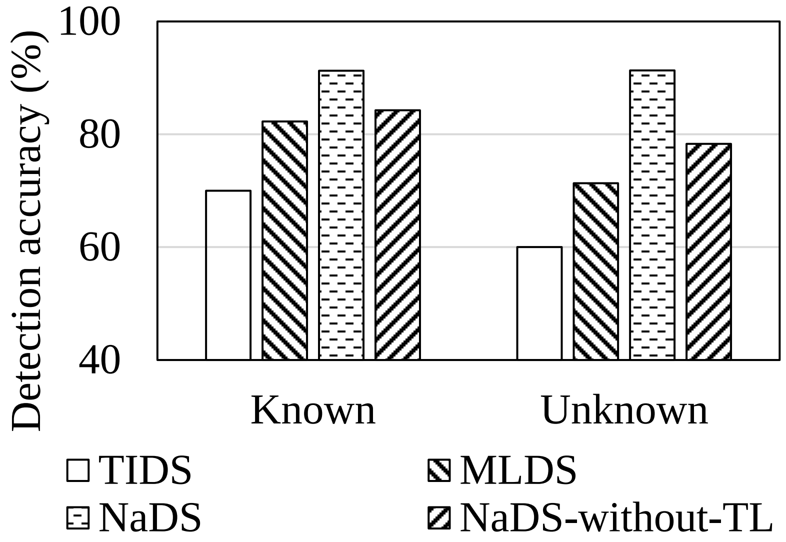


Accuracies for unknown abnormal messages

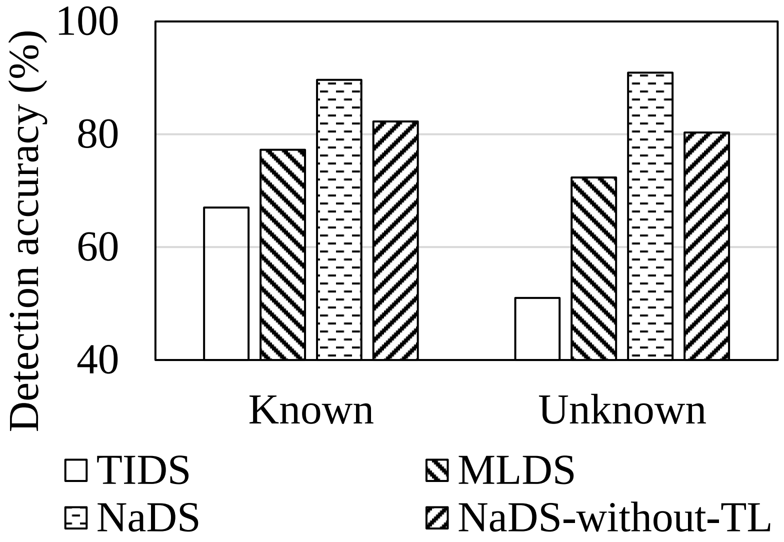
Performance Evaluation

Detection accuracy comparisons on new vehicle types

- Abnormal message detection accuracy of NaDS keeps stable because of transfer learning
- Abnormal message detection accuracy of other methods decrease greatly



Accuracies when LSTM network transfers from KIA to SONATA



Accuracies when LSTM network transfers from KIA to SPARK

Summary

Propose NaDS to detect abnormal messages in CAN bus on different vehicle types

- Built a LSTM-NN based abnormal message detection method
- Developed a network transfer method to transfer a pre-trained LSTM network
- Used real CAN bus message data to verify NaDS

Future work

- Consider more message related information

Thank you!