Reading Report

Log-based Anomaly Detection Without Log Parsing

汇报人：芮子淇

2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)

厚德弘毅 求是笃行

目录

CONTENT

Part One

背景介绍

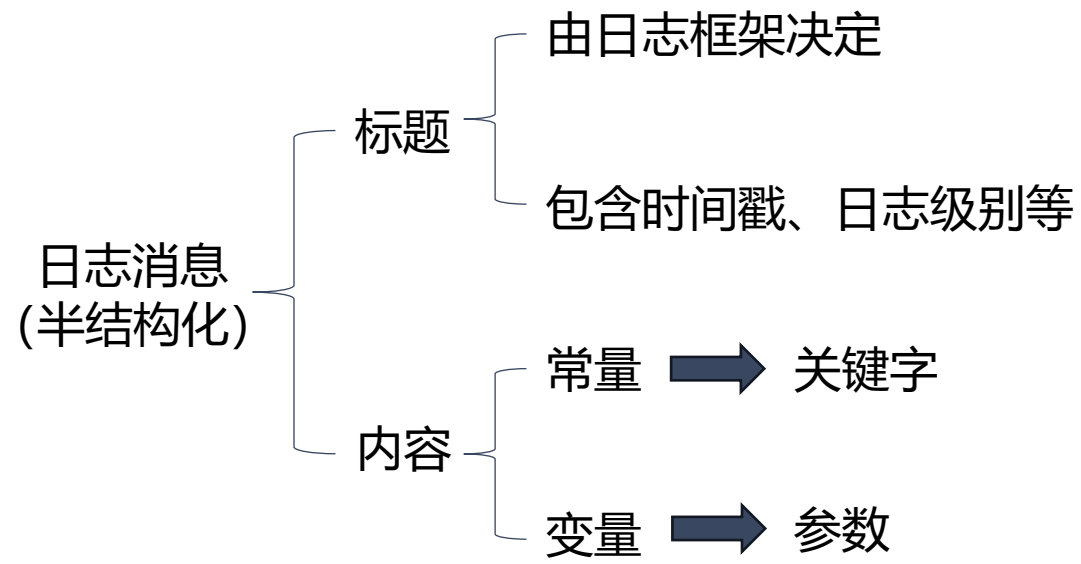
Background Introduction

01

背景介绍

日志数据

- 大型的软件系统通常在运行期间会产生大量日志数据用于记录系统的事件和运行状态。
- 通过分析日志，我们可以更好地理解系统状态，并在发生故障时及时进行系统诊断。

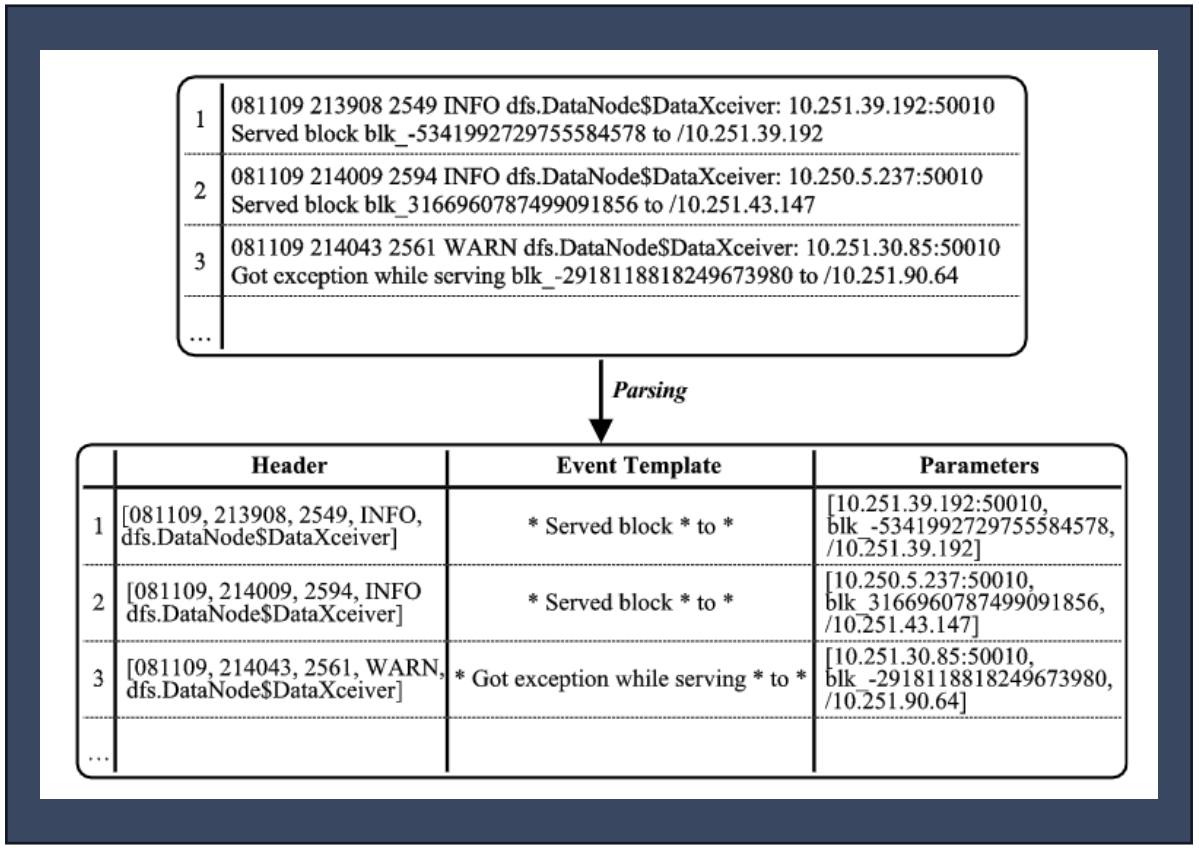


1	081109 213908 2549 INFO dfs.DataNode$DataXceiver: 10.251.39.192:50010 Served block blk_-5341992729755584578 to /10.251.39.192
2	081109 214009 2594 INFO dfs.DataNode$DataXceiver: 10.250.5.237:50010 Served block blk_3166960787499091856 to /10.251.43.147
3	081109 214043 2561 WARN dfs.DataNode$DataXceiver: 10.251.30.85:50010 Got exception while serving blk_-2918118818249673980 to /10.251.90.64
…	

An Example of HDFS Logs

日志解析方法

- 日志解析是通过删除参数，保留关键字，来将每条日志转换为特定的事件模板。



- 图中展示了将日志消息解析为模板的过程，例如第一条日志经过解析得到的模板是** Served block * to **，这里用 * 代替了原始参数。

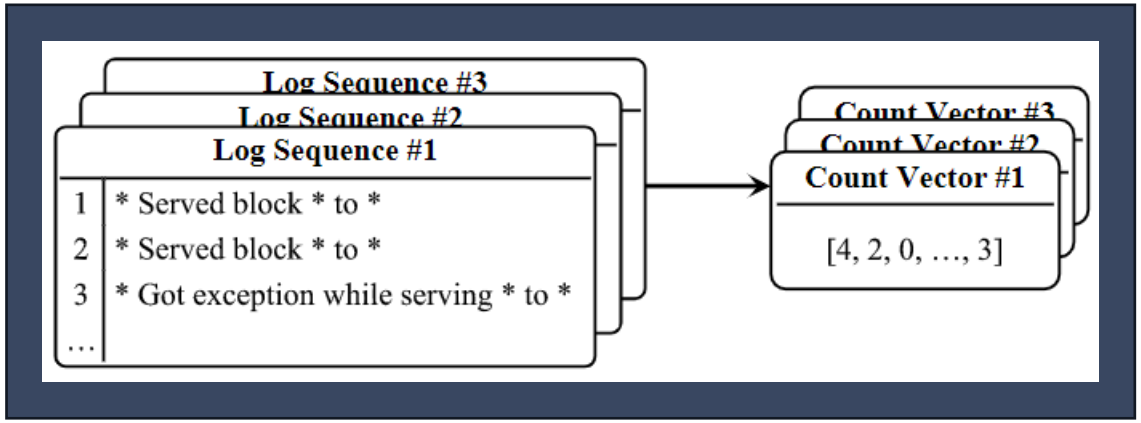
- 论文中评估了四个排名靠前的解析器，分别是Drain、AEL、Spell和 IPLoM。

背景介绍

日志异常检测

- 无监督的学习方法
- 有监督的学习方法

日志解析器



- **DeepLog**：它首先用 Spell 来提取日志模板，然后利用日志模板的索引将它们输入到 LSTM 模型中以预测下一个日志模板，根据下一条是否在预期内来检测异常。
- **LogAnomaly**：使用日志计数向量来检测异常日志事件编号反映的异常情况，并提出了一种基于同义词、反义词的方法来表示日志模板中的单词。
- **LogRobust**：结合了预训练的 Word2vec 模型，来学习由 Drain 生成的日志模板表示向量，并将这些向量输入基于注意力机制的双向 LSTM 模型来检测异常。

由于日志解析的缺陷，上述方法会丢失日志消息的语义，从而导致检测结果不够准确。

Part Two

相关研究

02

日志解析错误的实证研究

日志解析存在的问题

- 由未登录词（即OOV词）引起的日志解析错误
- 由语义误解引起的日志解析错误
- 日志解析错误对异常检测的影响

由未登录词引起的日志解析错误

关于“未登录词”

- 开发人员可以将新的日志语句添加到源代码中并修改现有日志语句的内容。此外，可以将运行信息添加到日志中作为记录系统状态的参数，所以未登录词（OOV词）经常出现在日志数据中。

研究思路

1. 通过日志的时间戳对日志消息进行排序

2. 利用前 P%（根据日志的时间戳）作为训练数据，其余作为测试数据。

3. 通过空格将每条训练日志消息拆分为一组标记，并用这些标记构建一个词汇表。
   未登录词是测试数据中不存在于词汇表的词。

4. 将训练数据的百分比从20%提高到80%，计算所有情况下的未登录词比例。

由未登录词引起的日志解析错误

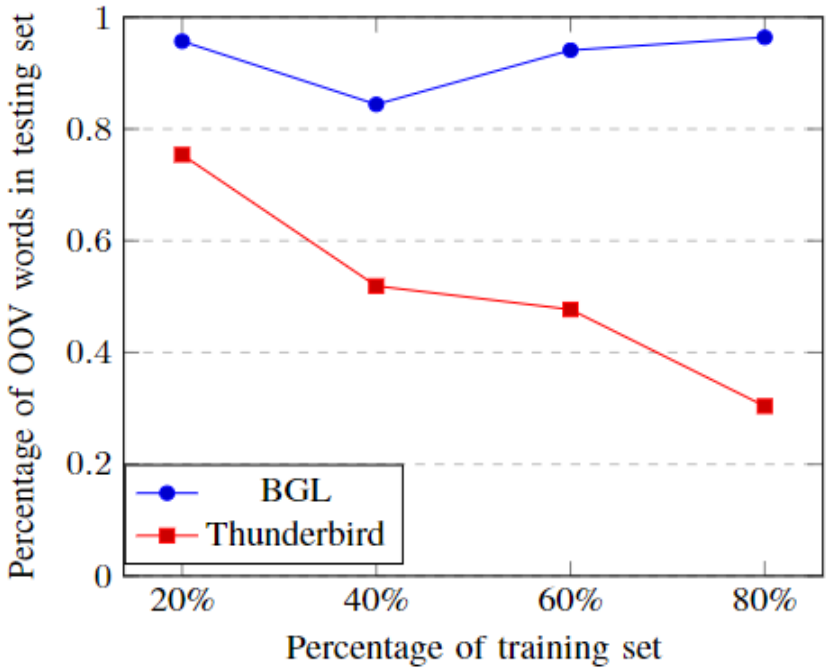
关于“唯一单词”

Tom likes to eat apple.

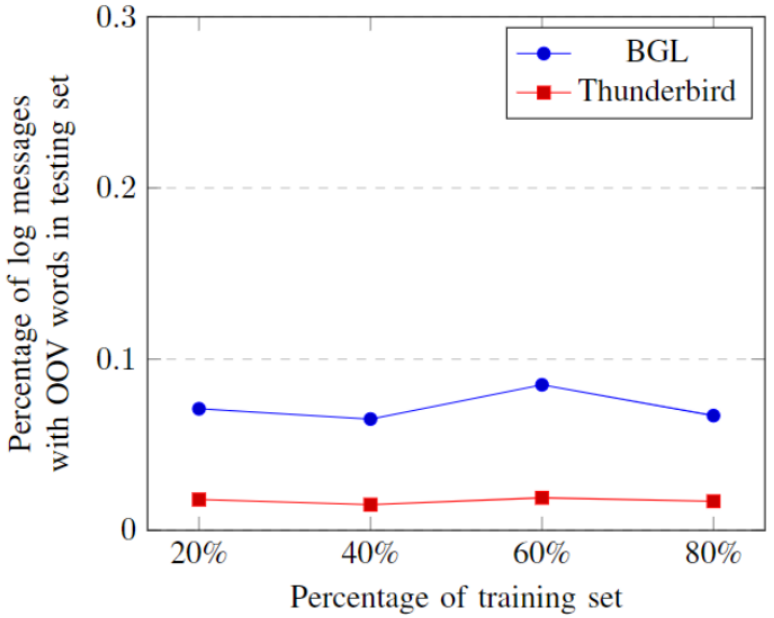
Jack likes apple, too.

{ “Tom” ,” likes” ,” to” ,” eat” ,” apple” ,” Jack” ,” too” }

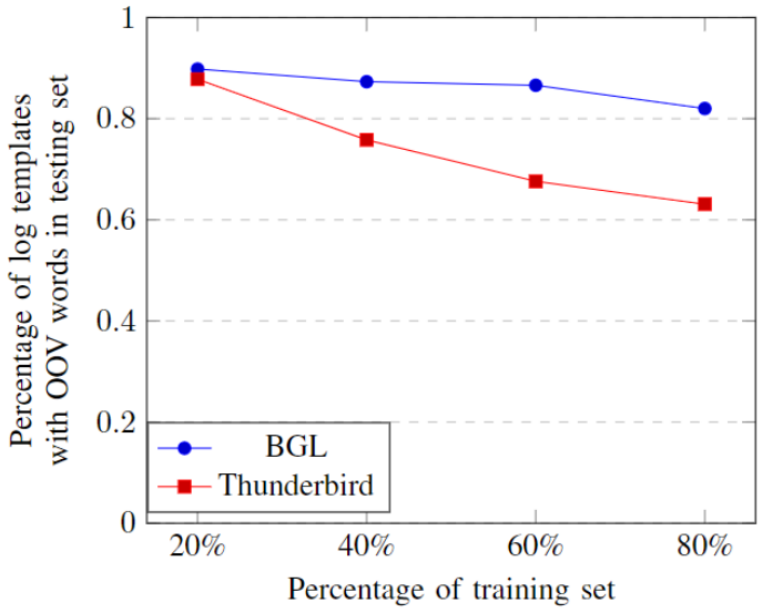
在BGL和雷鸟数据集上，当训练数据的百分比从20%增加到80%时，测试集中未登录词的百分比：



由未登录词引起的日志解析错误



(a) Log messages with OOV words



(b) Log templates with OOV words

包含未登录词的日志消息百分比

包含未登录词的日志模板百分比

由未登录词引起的日志解析错误

总结

少量包含未登录词的日志消息可以在测试集中产生大量未出现过的日志模板。

原因

1. 许多日志事件只出现在特定时段。
2. 日志的不平衡分布。
3. 未登录词会导致解析错误并产生许多额外的日志事件。

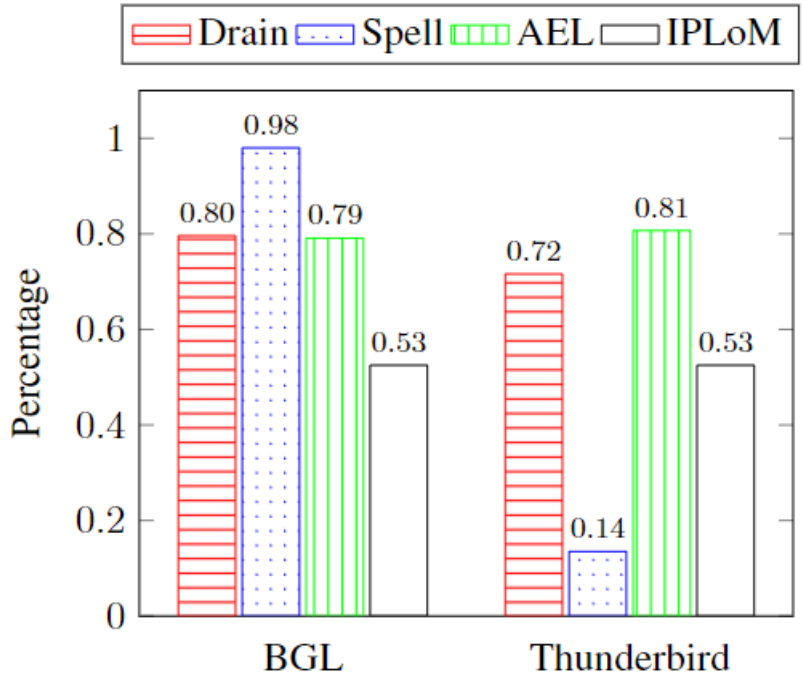
由语义误解引起的日志解析错误

Case1. 将参数误识别为关键字

Case 1:

- **Parsing results:**
L3 ecc status register: 00200000 → L3 ***ecc status*** register: *
L3 global control register: 0x001249f0 → L3 * ***control*** register: *
- **Ground truth Log Template:**
L3 * * register: *

两条日志消息原本是基于同一个日志模板，
但却被解析为两个不同的日志模板。



由语义误解引起的日志解析错误

Case2. 将关键字误识别为参数

Case 2:

- **Parsing results:**
machine check ***enable*** → machine check ∗
machine check ***interrupt*** → machine check ∗
- **Ground truth Log Templates:**
machine check enable
machine check interrupt

两个不同的日志消息被解析为相同的日志事件 *“machine check * ”*
第一个是正常行为，第二个是系统异常。

- **Anomaly:** ciod: LOGIN chdir(/p/gb1/stella/RAPTOR/2183) failed: ***Input/output error***
- **Normal:** ciod: LOGIN chdir(/home/bertsch2/src/bgl_hello) failed: ***Permission denied***
- **Parsed event:** ciod: LOGIN ∗ failed: ∗ ∗

(a) Errors introduced by Drain

- **Anomaly:** mptscsih: ioc0: ***attempting task abort!*** (sc=00000101bfc7a480)
- **Normal:** mptscsih: ioc0: ***task abort: SUCCESS*** (sc=00000101bfc7a480)
- **Parsed event:** mptscsih ioc0 ∗ ∗ ∗ sc ∗

(b) Errors introduced by Spell

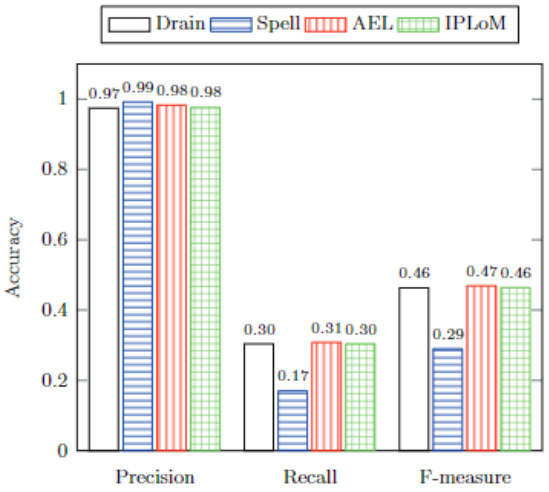
- **Anomaly:** floating point ***unavailable interrupt***
- **Normal:** floating point ***instr. enabled.....1***
- **Parsed event:** floating point ∗ ∗

(c) Errors introduced by AEL

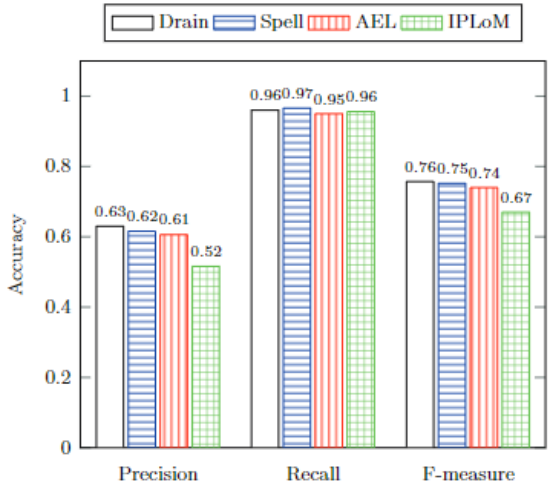
- **Anomaly:** ciod: Error creating node map from file ***map.dat: No child processes***
- **Normal:** ciod: Error creating node map from file ***map.dat: Bad file descriptor***
- **Parsed event:** ciod Error creating node map from file ∗ ∗ ∗ ∗

(d) Errors introduced by IPLoM

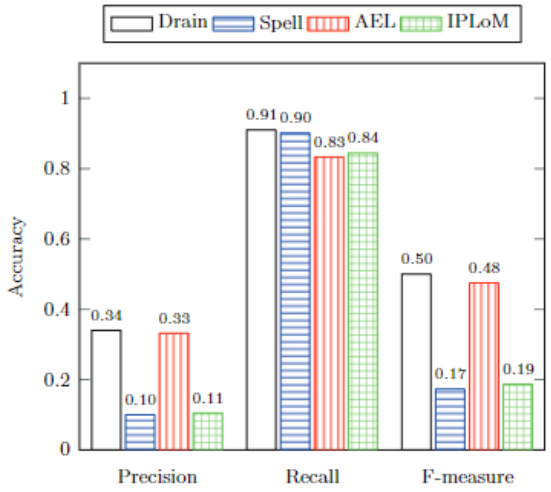
日志解析错误对异常检测的影响



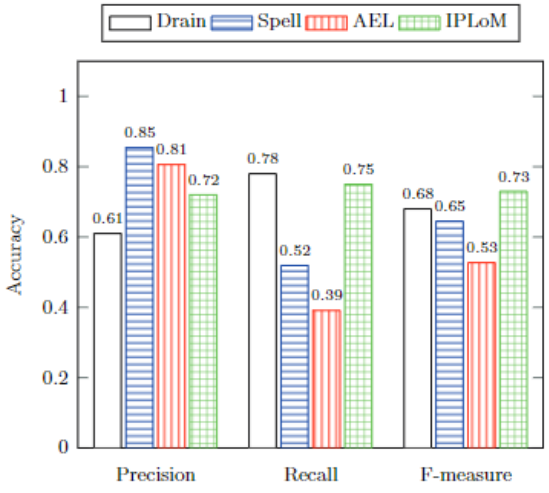
(a) SVM on BGL



(b) LogRobust on BGL



(c) SVM on Thunderbird



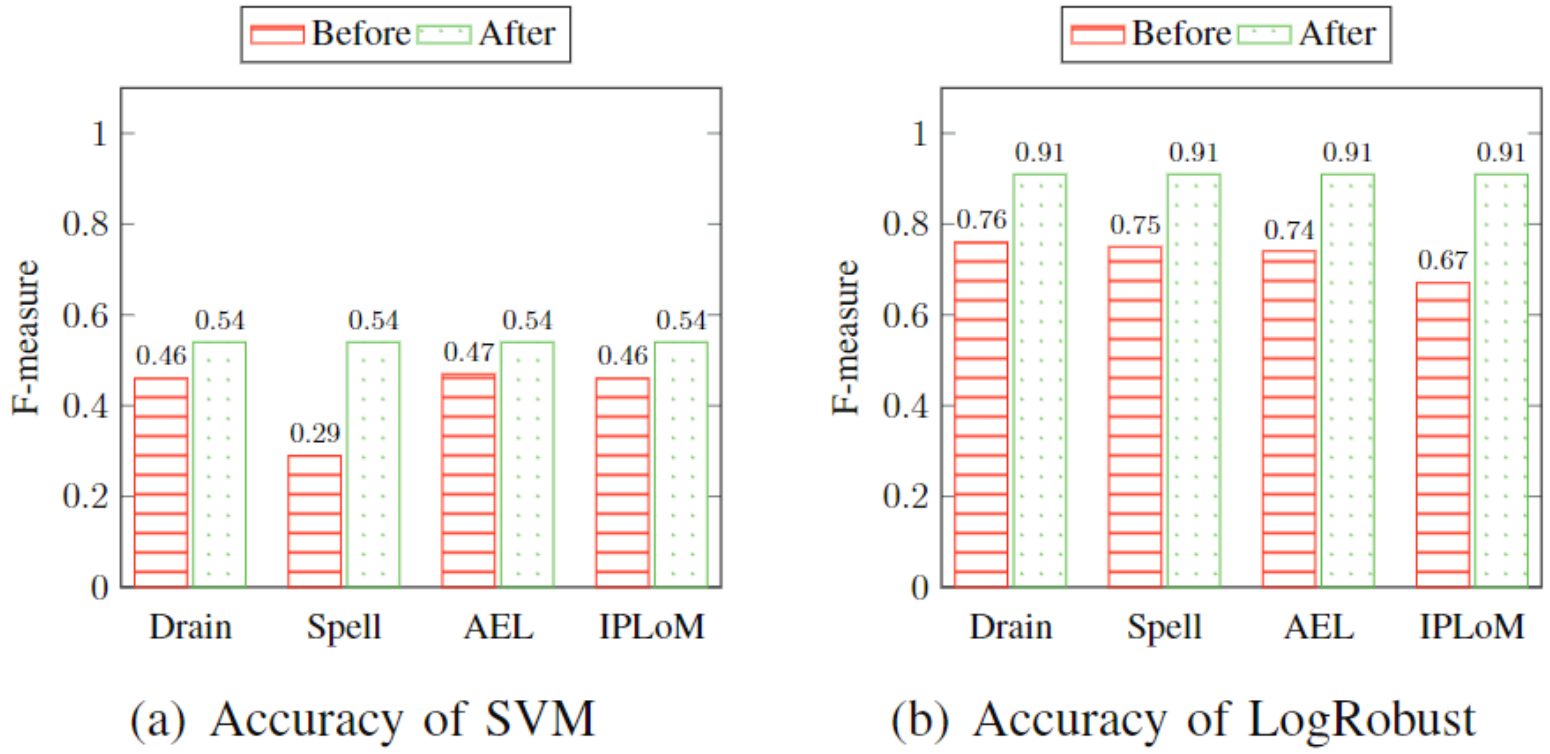
(d) LogRobust on Thunderbird

评价指标

实际类别		预测类别		
		Yes	No	总计
	Yes	TP	FN	P（实际为Yes）
	No	FP	TN	N（实际为No）
	总计	P'（被分为Yes）	N'（被分为No）	P+N

- 精度：$Precision = \frac{TP}{TP+FP}$
- 召回率：$Recall = \frac{TP}{TP+FN}$
- F值：$F1-score = \frac{2*Precision*Recall}{Precision+Recall}$

日志解析错误对异常检测的影响



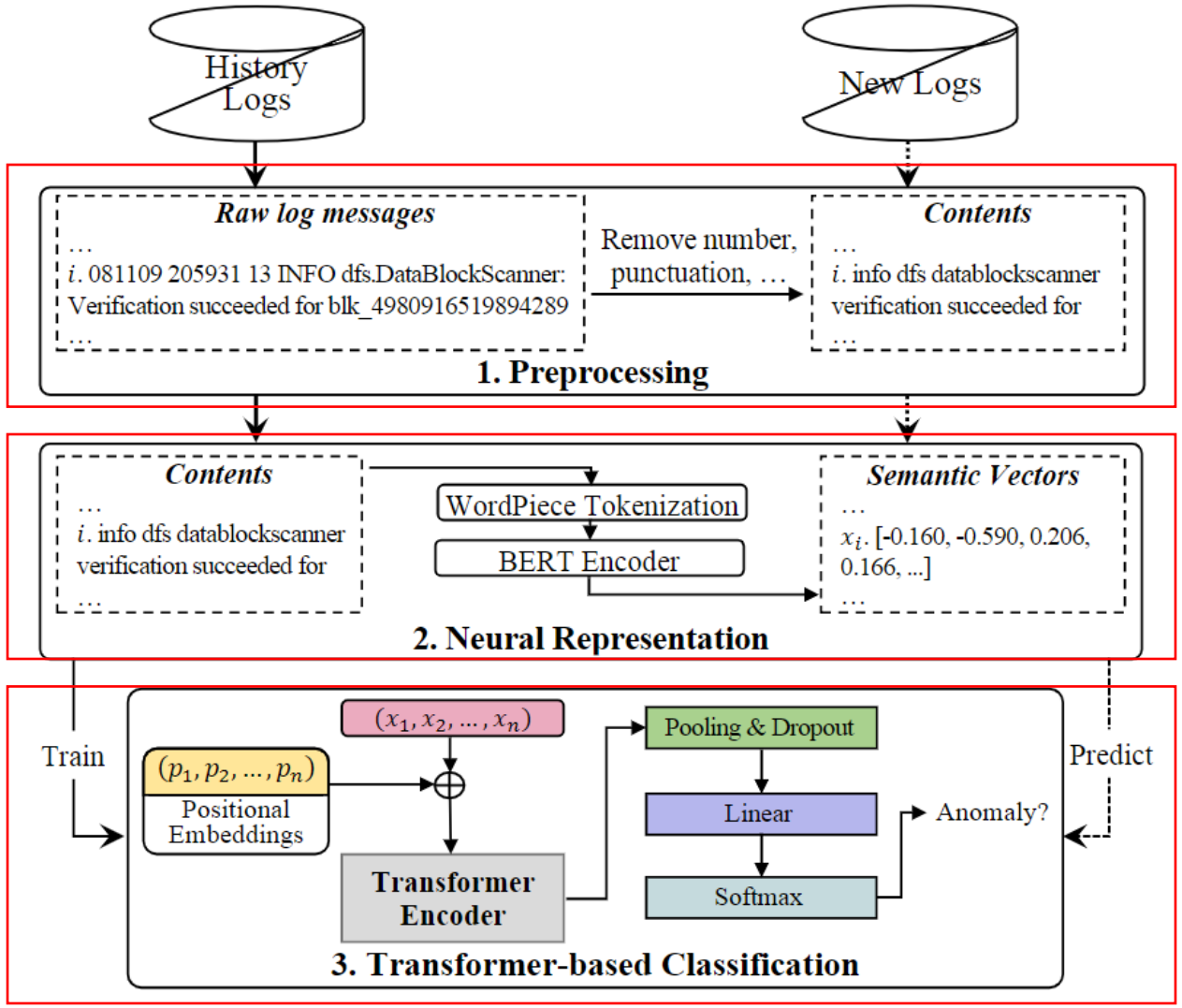
- 日志解析错误修复前后，SVM和LogRobust进行异常检测的准确性对比

Part Three

论文模型

03

模型设计



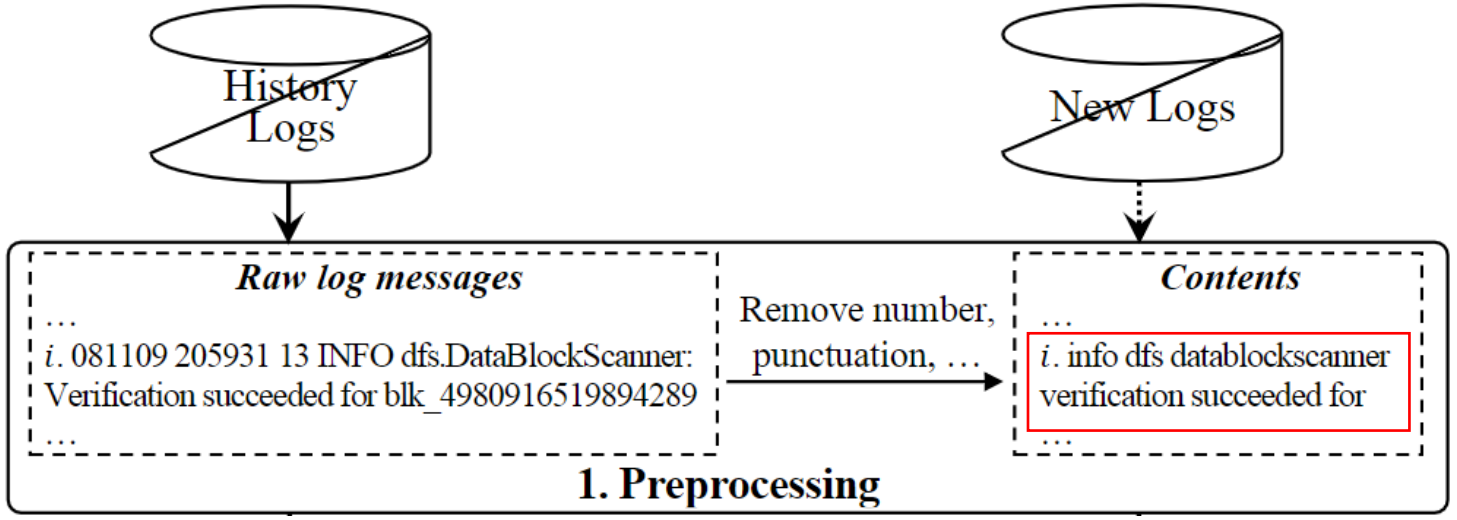
1. 日志预处理

2. 神经表示

3. 基于Transformer分类

第一步：日志预处理

1. 使用日志系统中常见的分隔符来切割日志消息。
2. 将每个大写字母转换为小写字母。
3. 从单词集中删除所有非字符标记，包括操作符、标点符号和数字。



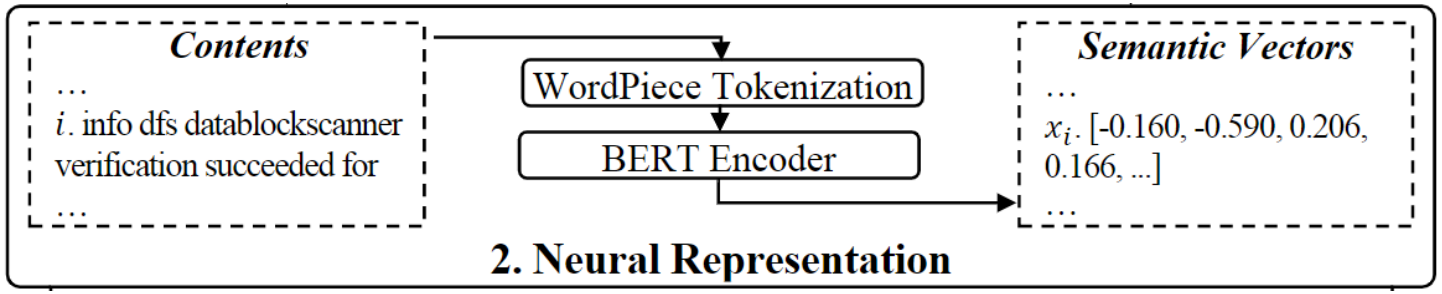
第二步：神经表示

Phase1. 子词标记化

WordPiece: datablockscanner ⟹ {"data", "block", "scan", "ner"}

Phase2. 日志消息表示

1. 将词集传入BERT，编码为固定维度向量
2. 生成日志消息整体的嵌入向量
3. 位置嵌入捕获单词上下文信息
4. 通过注意力机制衡量每个单词的重要性



第三步：基于Transformer分类

Phase1. 位置编码

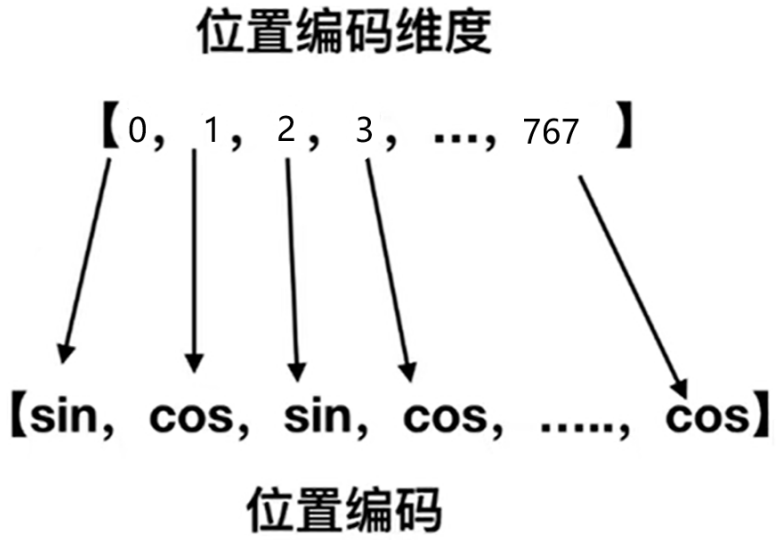
正弦编码器

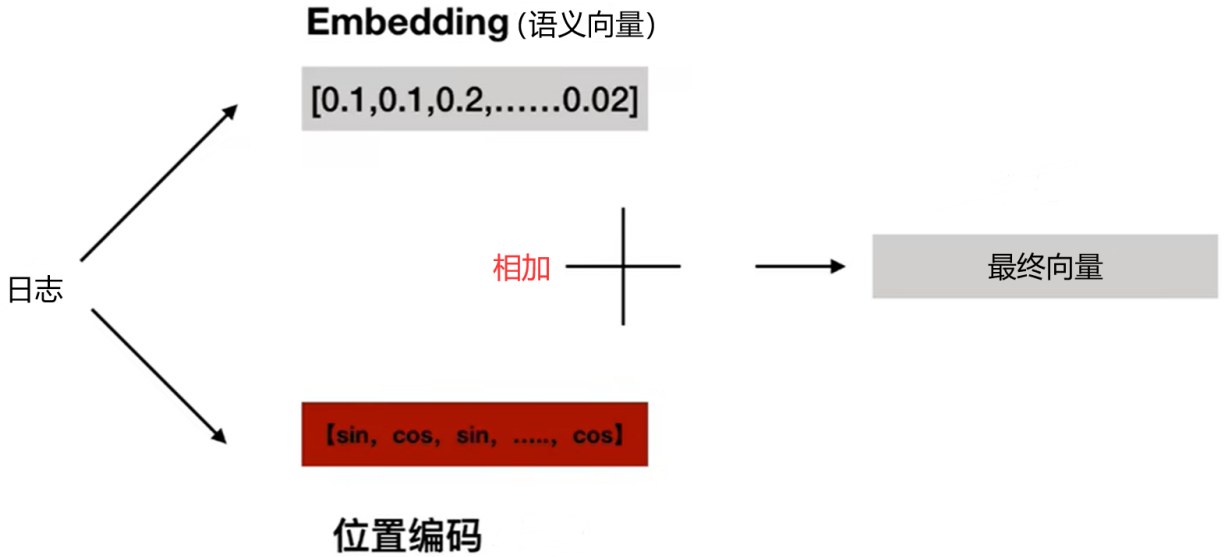
$$PE_{(t,2k)} = \sin\left(\frac{t}{10000^{2k/d_{\text{model}}}}\right)$$

偶数索引用正弦

$$PE_{(t,2k+1)} = \cos\left(\frac{t}{10000^{2k/d_{\text{model}}}}\right)$$

奇数索引用余弦





第三步：基于Transformer分类

Phase2. Transformer编码器

Attention is all you need

多头注意力：获得不同的注意力模式

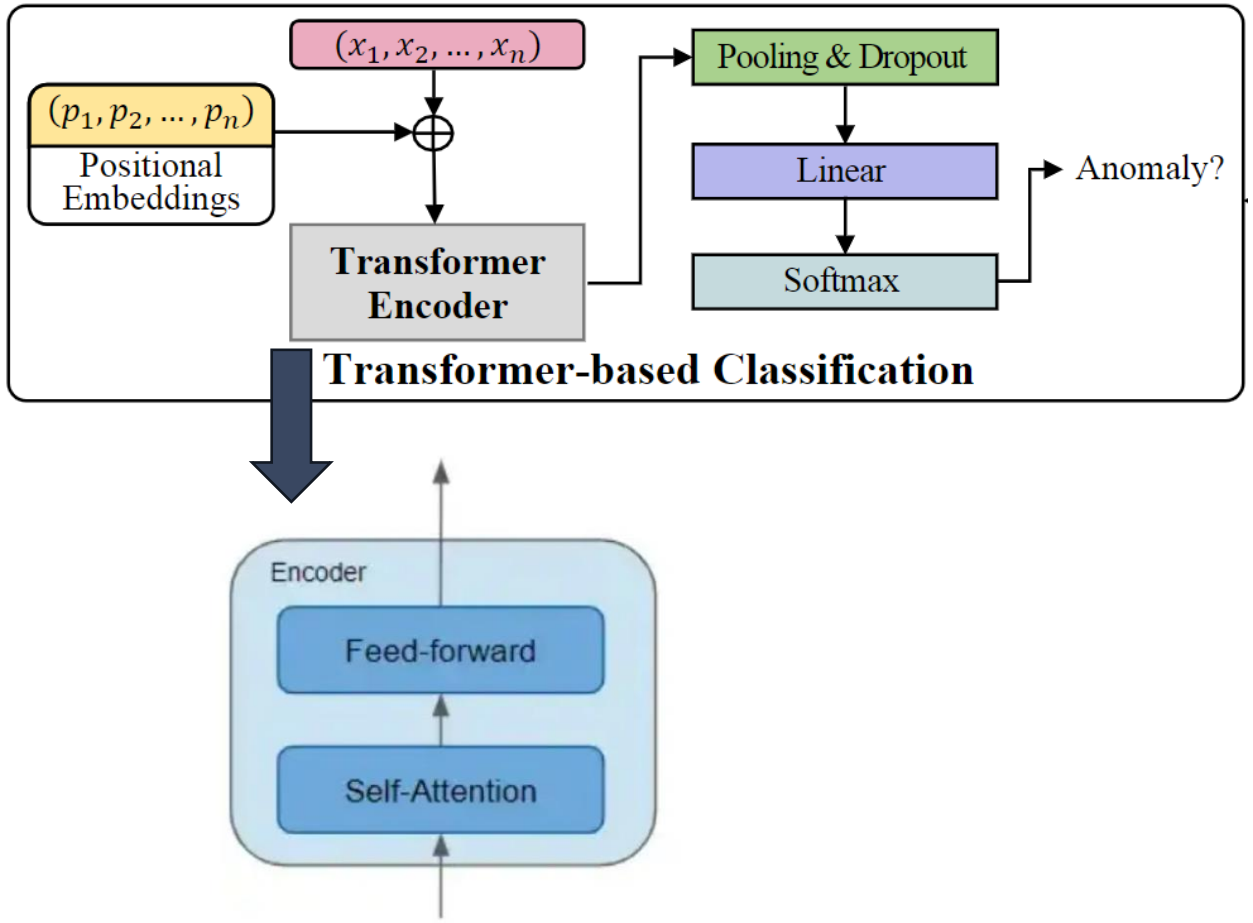
前馈网络层：将不同的注意力分数组合起来

池化：减少模型的复杂度，提取关键特征

Dropout：丢弃某些神经元，防止过拟合

Linear：用于特征提取，将高维数据转换为低维数据，同时保留重要的信息

Softmax：接收来自模型的分数，生成一个概率分布，进行异常分类



Part four

实验评估

04

实验设计

数据集

	Category	Size	#Messages	#Anomalies
HDFS	Distributed system	1.5 G	11,175,629	16,838
Blue Gene /L	Supercomputer	743 M	4,747,963	348,460
Thunderbird	Supercomputer	1.4 G	10,000,000	4,934
Spirit	Supercomputer	1.0 G	7,983,345	768,142

实验设置

- 注意力头：12
- 前馈网络大小：2048
- 优化器：Adam
- 初始学习率：3e-4
- mini-batch：64
- Dropout：0.1

NeuralLog的效果

实验结果

- 编码：14.3 min/数据集
- 训练模型：5.2 min
- 异常检测：3.1 ms/日志序列

Dataset		LR	SVM	IM	LogRobust	Log2Vec	NeuralLog
HDFS	P	0.99	0.99	1.00	0.98	0.94	0.96
	R	0.92	0.94	0.88	1.00	0.94	1.00
	F1	0.96	0.96	0.94	**0.99**	0.94	0.98
BGL	P	0.13	0.97	0.13	0.62	0.80	0.98
	R	0.93	0.30	0.30	0.96	0.98	0.98
	F1	0.23	0.46	0.18	0.75	0.88	**0.98**
Thunder-bird	P	0.46	0.34	-	0.61	0.74	0.93
	R	0.91	0.91	-	0.78	0.94	1.00
	F1	0.61	0.50	-	0.68	0.84	**0.96**
Spirit	P	0.89	0.88	-	0.97	0.91	0.98
	R	0.96	1.00	-	0.94	0.96	0.96
	F1	0.92	0.93	-	0.95	0.95	**0.97**

'-' denotes timeout (30 hours), P denotes Precision, R denotes Recall, and F1 is the F1-score.

NeuralLog获取日志语义的效果

将日志表示为语义向量的编码部分

两种变体
- NeuralLog-Index
- NeuralLog-Template

RESULTS OF DIFFERENT REPRESENTATION METHODS

Dataset	Metric	NeuralLog-Index	NeuralLog-Template	NeuralLog
HDFS	Precision	0.93	0.93	0.96
	Recall	1.00	1.00	1.00
	F1-Score	0.96	0.96	**0.98**
BGL	Precision	0.98	0.92	0.98
	Recall	0.30	0.88	0.98
	F1-Score	0.46	0.90	**0.98**
Thunderbird	Precision	0.58	0.89	0.93
	Recall	0.98	0.91	1.00
	F1-Score	0.73	0.90	**0.96**
Spirit	Precision	0.96	0.93	0.98
	Recall	0.95	0.95	0.96
	F1-Score	0.95	0.94	**0.97**

NeuralLog获取日志语义的效果

处理未登录词的子词标记化部分

两种变体
- NeuralLog-Word2Vec
- NeuralLog-NoWordPiece

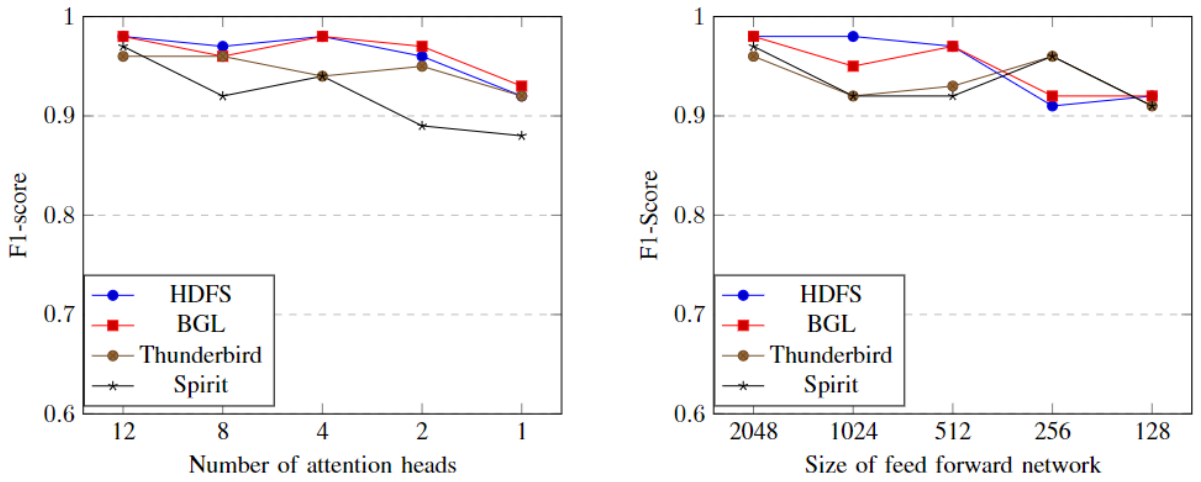
RESULTS OF HANDLING OOV WORDS

Dataset	Metric	NeuralLog-Word2Vec	NeuralLog-NoWordPiece	NeuralLog
HDFS	Precision	0.94	0.94	0.96
	Recall	0.93	1.00	1.00
	F1-Score	0.94	0.97	**0.98**
BGL	Precision	0.94	0.93	0.98
	Recall	0.88	0.96	0.98
	F1-Score	0.91	0.96	**0.98**
Thunderbird	Precision	0.80	0.90	0.93
	Recall	0.80	0.89	1.00
	F1-Score	0.80	0.90	**0.96**
Spirit	Precision	0.94	0.93	0.98
	Recall	0.92	0.80	0.96
	F1-Score	0.93	0.86	**0.97**

NeuralLog在不同设置下的有效性

RESULTS OF DIFFERENT PRE-TRAINED MODELS

Dataset	Metric	BERT [38]	GPT2 [53]	RoBERTa [54]
HDFS	Precision	0.96	0.95	0.85
	Recall	1.00	1.00	1.00
	F1-Score	**0.98**	0.97	0.92
BGL	Precision	0.98	0.95	0.95
	Recall	0.98	0.99	0.90
	F1-Score	**0.98**	0.97	0.93
Thunderbird	Precision	0.93	0.85	0.78
	Recall	1.00	0.91	1.00
	F1-Score	**0.96**	0.88	0.88
Spirit	Precision	0.98	0.88	0.84
	Recall	0.96	0.95	0.90
	F1-Score	**0.97**	0.91	0.87



(a) Results of NeuralLog with different number of attention heads

(b) Results of NeuralLog with different size of feed forward network

Fig. 11. Results of different hyperparameter settings

- 将BERT替换为不同模型的效果
- 使用不同参数设置的结果

Part Five

总结分析

05

论文贡献

- 直接处理原始日志消息，没有使用日志解析器。
- 使用BERT和WordPiece在子词级别捕捉未登录词的含义。
- 基于Transformer的分类模型提高了异常检测的性能。

不足与思考

- 预处理环节删除了包含数字和特殊字符的单词，但某些情况下，被删除的单词可能包含重要信息。
- 数据集主体系统的数量有限，不能覆盖所有领域。
- 公共数据集一般由工程师手工检查和标记，但是在人工标注的过程中可能会产生噪声。

THANKS FOR WATCHING

汇报完毕 谢谢观看

汇报人：芮子淇

厚德弘毅 求是笃行