Malicious URL Detection Algorithm Based on Multi Neural Network Series

Weirong Xiu (xiuweirong@vip.163.com)
Guangzhou College of Commerce https://orcid.org/0000-0001-9920-844X
Chen Bian
Guangdong University of Finance https://orcid.org/0000-0001-7990-5125
Chunhui Wu
Guangdong University of Finance

Research Article

Keywords: Malicious URL, CATIR, word vector feature, detection

Posted Date: May 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-442187/v1

Abstract

Abstract: Convolutional neural network based on attention mechanism and a bidirectional independent recurrent neural network tandem joint algorithm (CATIR) are proposed. In natural language processing related technologies, word vector features are extracted based on URLs, and the extracted URL information features and host information features are merged. The proposed CATIR algorithm uses CNN (Convolutional Neural Network) to obtain the deep local features in the data, uses the Attention mechanism to adjust the weights, and uses IndRNN (Independent Recurrent Neural Network) to obtain the global features in the data. The experimental results shows that the CATIR algorithm has significantly improved the accuracy of malicious URL detection based on traditional algorithms to 96.9%. Keywords: Malicious URL; CATIR; word vector feature; detection

Full Text

This preprint is available for download as a PDF.

Figures

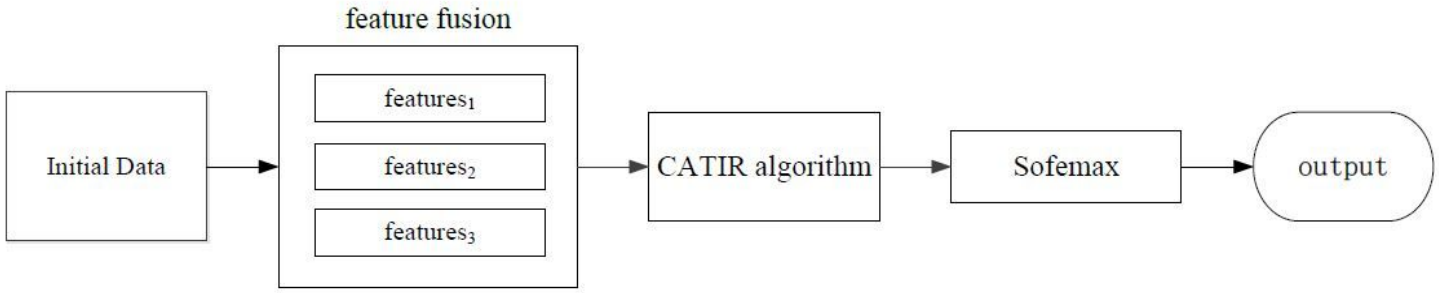

Figure 1

This is the overall schematic diagram, firstly, the host information features and URL information features are learned by the algorithm to ob tain block features, secondly, the face features are learned by the algorithm word vector features and block features , finally, the face features are input to the CATIR tandem joint algorithm in this paper for training, which is used for malic ious URL analysis and detection

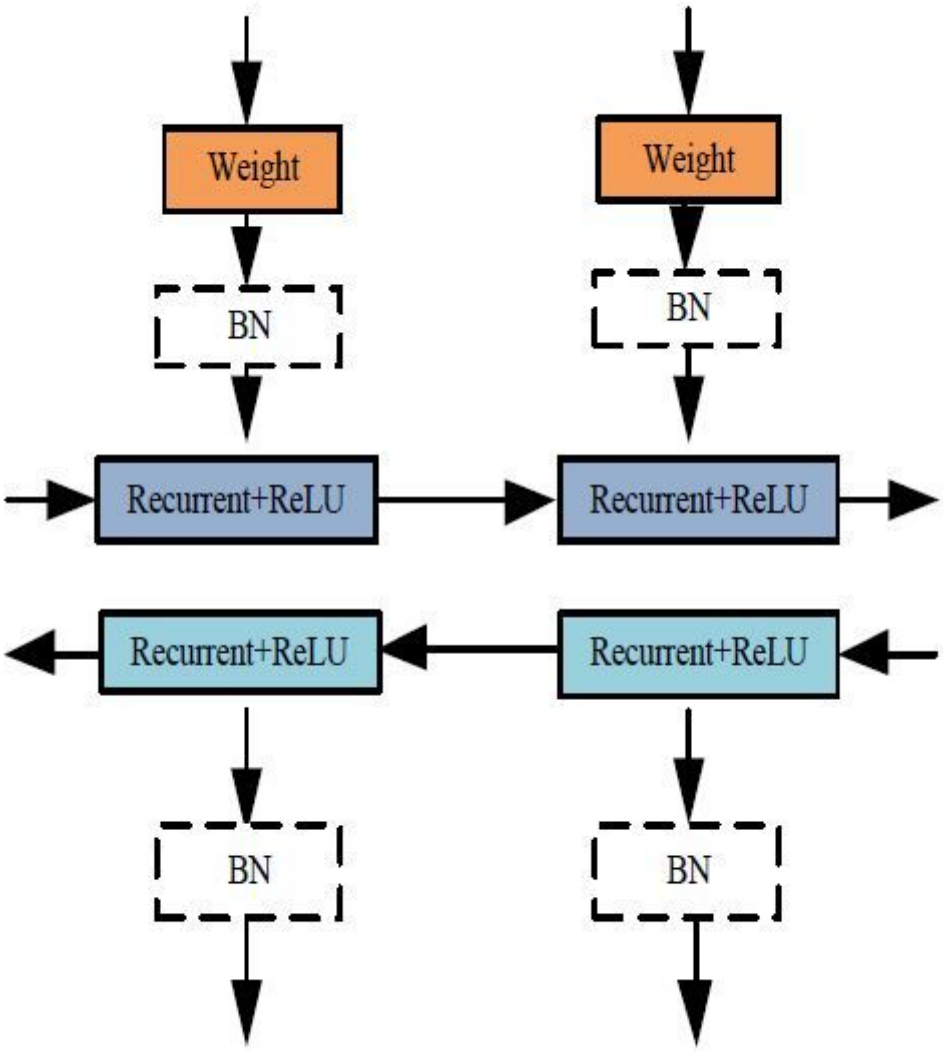

Figure 2

The structure of the IndRNN algorithm is illustrated, with the Weight and ReLU activation functions acting as an arithmetic and cyclic process for each step of the feature implementation, and the BN acting as a normaliser for each step of the feature imple mentation.

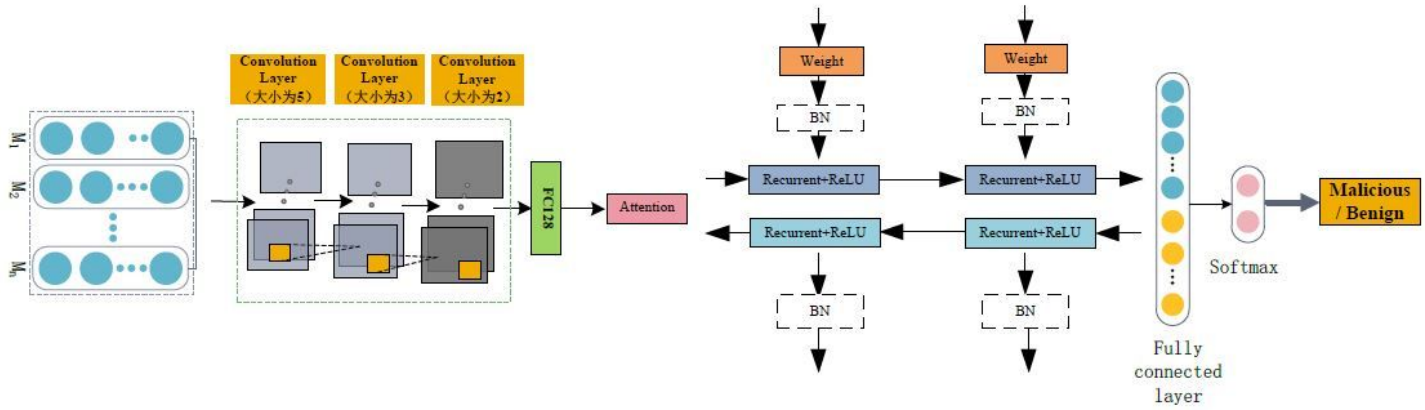

Figure 3

This is a diagram of the overall structure of the CATIR tandem union algorithm. The features are fed into the CNN algorithm, the A ttention mechanism, an d the IndRNN algorithm respectively, and then the CNN algorithm, the A ttention mechanism, and the IndRNN algorithm are combined in tandem to obtain the CATIR algorithm, which goes through the fully connected layer and finally the SoftMax classifier layer t o classify and obtain the malicious URL detection results.

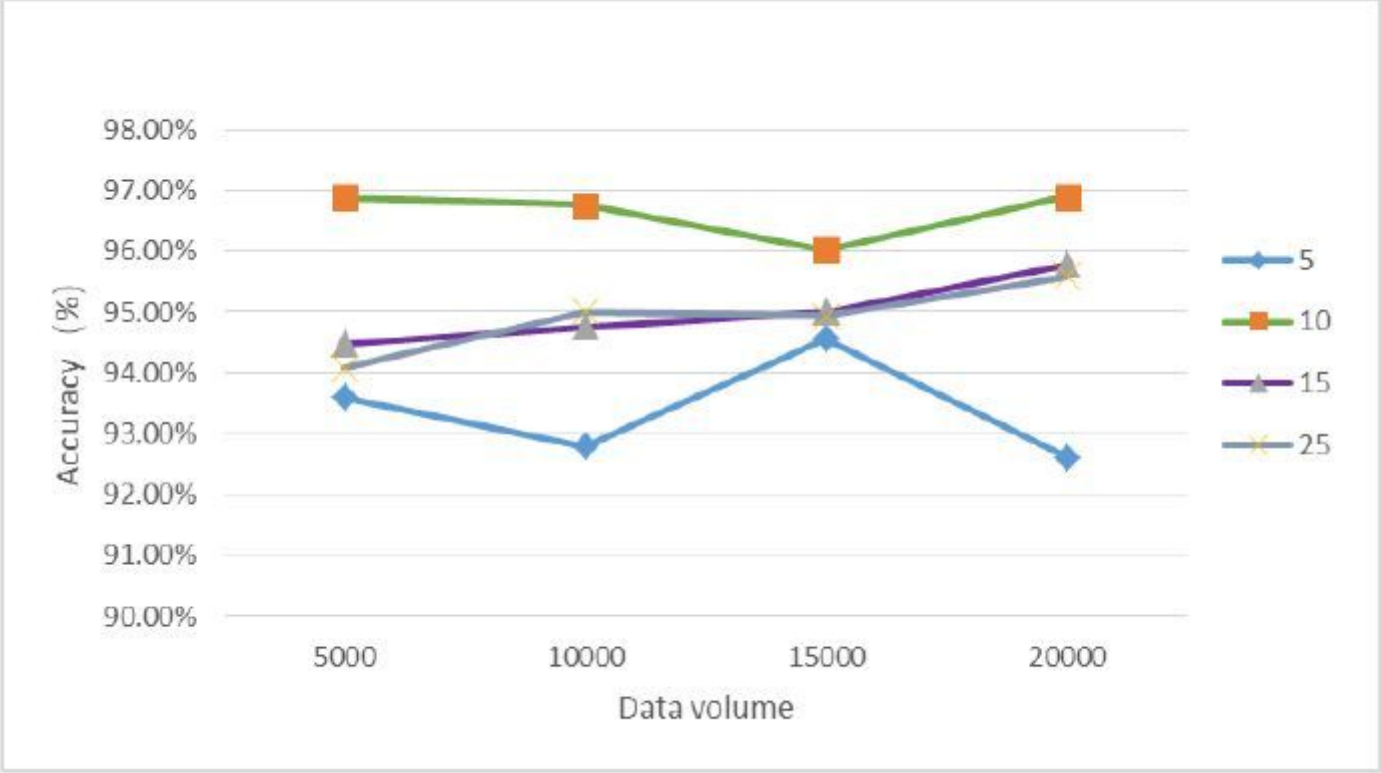

Figure 4

The horizontal axis is the amount of data, the vertical axis is the detection result, and the blue, green, purple and grey dashes represent the number of iterations parameter 5, 10, 15 and 25 respectively

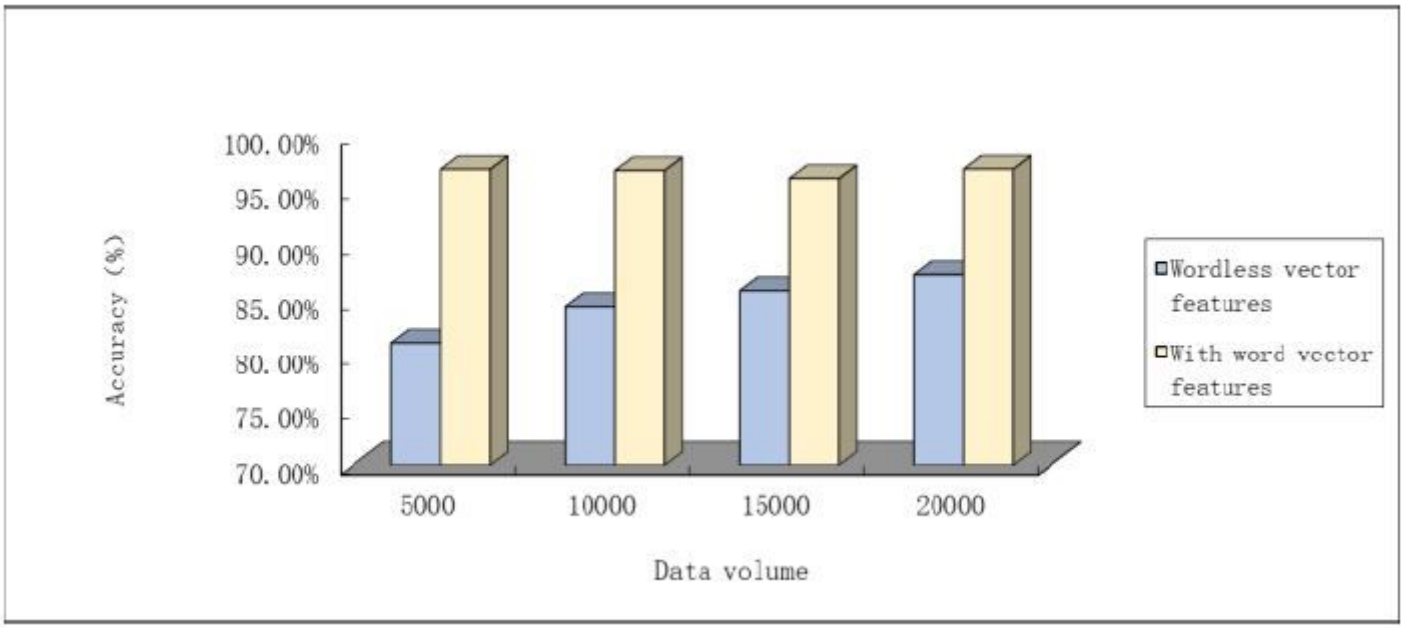

Figure 5

The horizontal axis is the amount of data, the vertical axis is the detection result, the blue bars represent the wordless vector features, the yellow bars represent the worded vector features

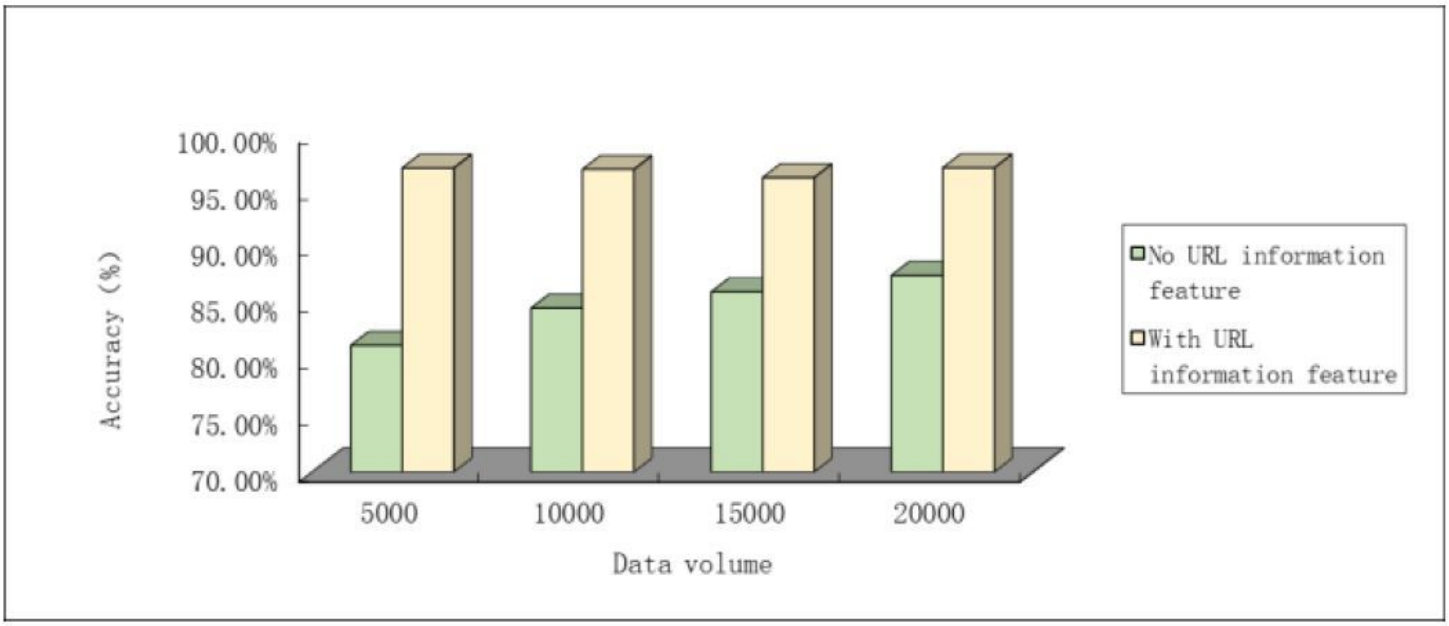

Figure 6

The horizontal axis is the amount of data, the vertical axis is the accuracy of the detection results, the green bars represent the features without URL information, the yellow bars represent the features with URL information

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Figure1Overallchematic.vsd
- Figure2IndRNNstructure.vsd
- Figure3CATIRstructure.vsd