Vulnerability Detection via Topological Analysis of Attention Maps

Pavel Snopov[1]

Institute for Information Transmission Problems of Russian Academy of Sciences

Recently, DL-based approaches to the vulnerability detection task became popular. They show promising results, outperforming traditional static code analysis tools. In this work, we test a novel approach to the vulnerability detection using topological data analysis of the attention matrices of pretrained on code LLMs. We choose `microsoft/codebert-base` as our base model.

For each code sample, we calculate the persistent homology in dimension 0 and 1 of the symmetrized attention matrices, obtaining the persistence diagram on each attention head of the BERT model. We compute the following features in each dimension from the diagrams:

- The mean lifespan of points on the diagram
- The variance lifespan of points on the diagram
- The max lifespan of points on the diagram
- The overall number of points on the diagram
- The persistence entropy

We symmetrize attention matrices in the following manner:

$$\forall i, j : W_{ij}^{\text{sym}} = \max\left(W_{ij}^{\text{attn}}, W_{ji}^{\text{attn}}\right). \tag{1}$$

We consider these features as the numerical characteristic of the semantic evolution processes in the attention heads. The features with «significant» persistence (i.e. those with large lifespan) correspond to the stable processes, whileas the features with short lifespan are highly influenced to noise and doesn't reflect the stable topological attributes.

Table 1. The results of the vulnerability detection experiments.

Model	F1 score	Accuracy
Logistic Regression	0.22	0.54
LightGBM	**0.55**	0.63
SVM	0.54	**0.65**
CodeBERTa (pre-trained)	0.28	0.45
CodeBERTa (fine-tuned)	**0.55**	0.55

Table 1 outlines the results of the vulnerability detection experiments on the *Devign* dataset. The results reveal that the proposed topology-based classifiers ourperform chosen LLM without fine-tuning and perform on par with the fine-tuned version.