



Deep Representation Classification of Vulnerabilities via Source Code

Dennis Law [Georgia Tech]

Pengchu Zhang, Org. 09371

Problem Statement:

Verification of the findings by R. Russel *et al.* (<https://arxiv.org/abs/1807.04320>) in using deep representation learning to identify vulnerabilities in source code.

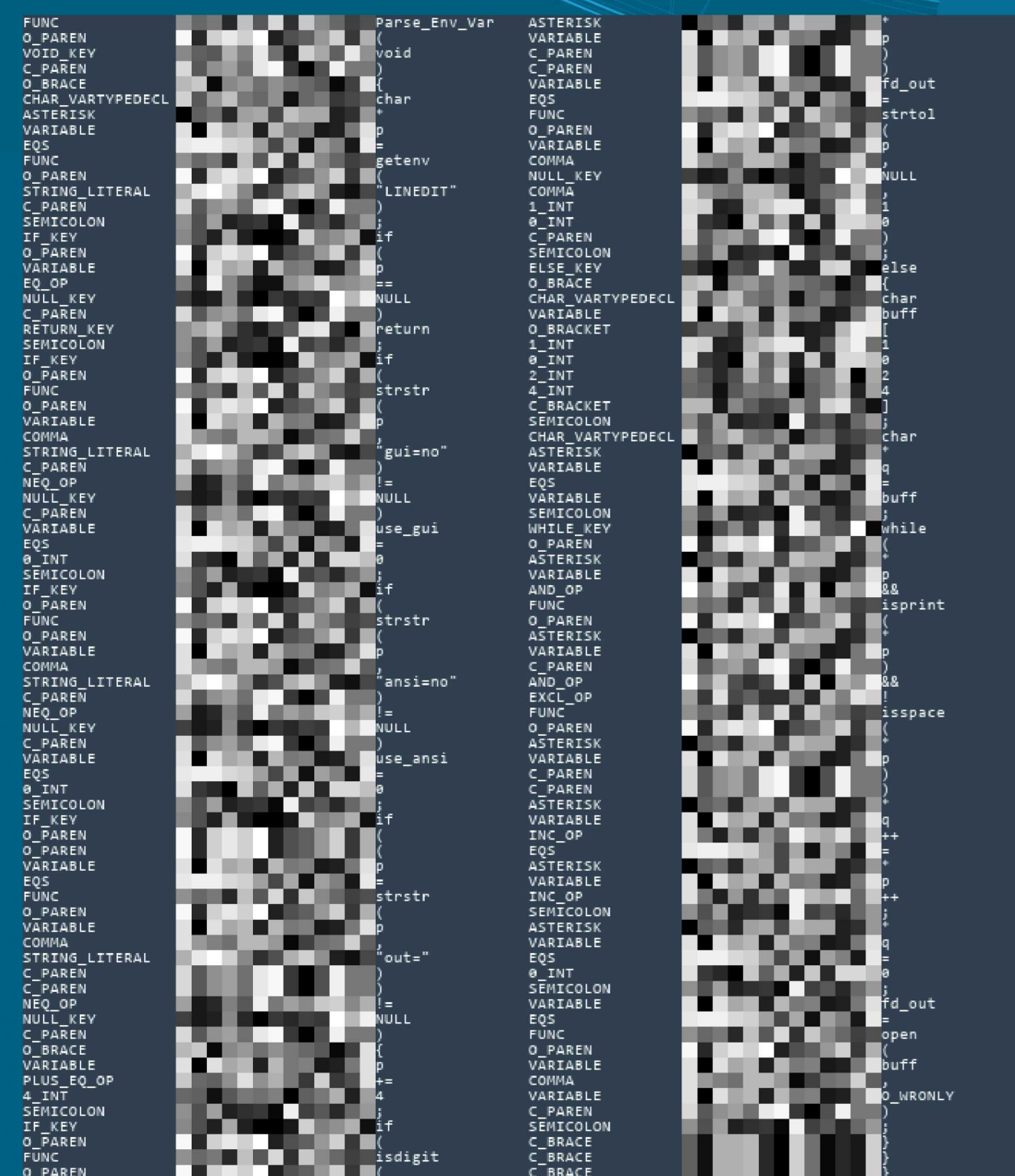
```

1 Parse_Env_Var(void)
2 {
3     char p = getenv("LINEDIT");
4
5     if (p == NULL)
6         return;
7
8     if (strstr(p, "gui=no") != NULL)
9         use_gui = 0;
10
11     if (strstr(p, "ansi=no") != NULL)
12         use_ansi = 0;
13
14     if ((p = strstr(p, "out=")) != NULL)
15     {
16         p += 4;
17         if (isdigit(*p))
18             fd_out = strtol(p, NULL, 10);
19         else
20         {
21             char buff[1024];
22             char q = buff;
23             while(*p && isprint(*p) && !isspace(*p))
24                 *q++ = *p++;
25             *q = '\0';
26             fd_out = open(buff, O_WRONLY); /* on error fd_out = -1 */
27         }
28     }
29 }
30
31
32
33

```

Highlighted feature activations from CNN overlaid on the example vulnerable function.

Results:



Token Embedding Representation

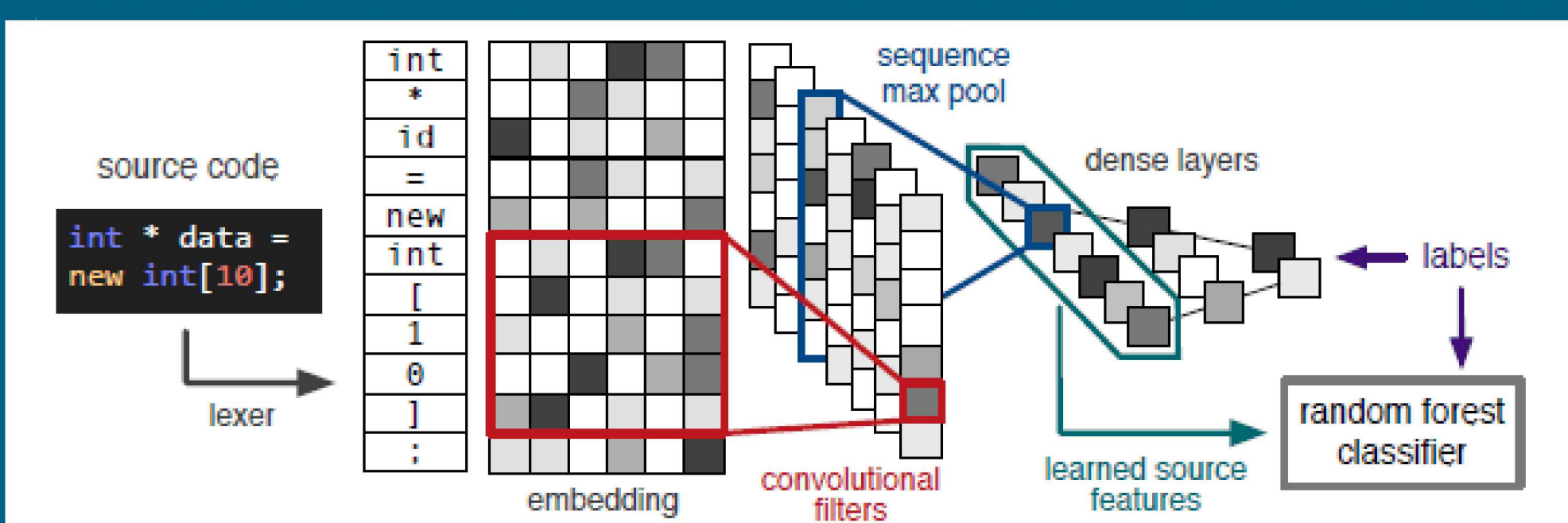


Diagram by Russel of the process used for classification.



Activations of the example vulnerable function through CNN feature-learning. Each row represents a single token.

Objectives and Approach:

- A tokenizer was created to convert C/C++ code into a tokenized representation comprised of 151 total tokens
- These tokens were fed into a word2vec embedding algorithm to obtain a 13-dimensional embedding for each token
- These were packaged up before being used to train the CNN model
- The CNN model could be cut off at the global-maxpooling stage and fed into a Random Forest Classifier

Impact and Benefits:

Successful automation of vulnerability detection will both alert developers to possible problems as they develop and allow analysts to prune large function sets.

Below is the current stage of replication. Red is the training set. Blue is the validation set.

