Word2Vec in Program Analysis

Haoran Cai\textsuperscript{1}

\textsuperscript{1}Department of Statistics  
University of Washington

CSE504 Project Proposal  
January 13, 2016
Representation is important in machine learning

- Word embedding. $W$: words $\rightarrow \mathbb{R}^n$
  
  $W(\text{"cat"}) = (0.2, -0.1, 0.3 \ldots)$
  
  $W(\text{"song"}) = (0.1, -0.2, 0.3 \ldots)$

- The representation is useful for downstream machine learning tasks. [Bottou, 2014]

- Words with similar meanings have similar vectors [Mikolov et al.]

![Diagram](image-url)
What is the “word” in the code?

- Token-level
- [Mou et al., 2014] suggested nodes in abstract syntax trees (ASTs)

```c
double doubles(double doublee) {
    return 2 * doublee;
}
```

(A) A C code snippet

```
FuncDef
  Decl
  FuncDecl
    ParameterList
      Decl
      TypeDecl
    IdentifierType
  Compound
    Return
      BinaryOp
        Constant
        ID
  TypeDecl
  IdentifierType
```

References
What will this distributed representation can do?

- Clustering
- Error checking
- Generate code
- And many more . . .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality.