Word2Vec in Program Analysis

Haoran Cai[1]

[1] Department of Statistics
University of Washington

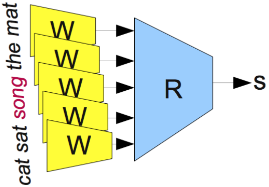
CSE504 Project Proposal
January 13, 2016

Representation is important in machine learning

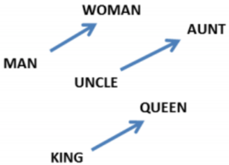
- Word embedding. W: words $\rightarrow \mathcal{R}^n$

$$\begin{aligned} W(\text{“}cat\text{”}) &= (0.2, -0.1, 0.3 \ldots) \\ W(\text{“}song\text{”}) &= (0.1, -0.2, 0.3 \ldots) \end{aligned}$$

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

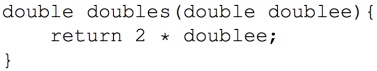


- Words with similar meanings have similar vectors [Mikolov et al.]



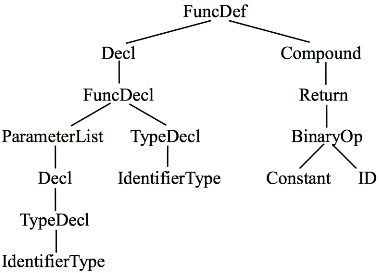
What is the “word” in the code?

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



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

(A) A C code snippet



What will this distributed representation can do?

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

Reference

Léon Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014.

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

Lili Mou, Ge Li, Yuxuan Liu, Hao Peng, Zhi Jin, Yan Xu, and Lu Zhang. Building program vector representations for deep learning. *arXiv preprint arXiv:1409.3358*, 2014.