The Center for Cyber Defenders

Expanding Computer Security Knowledge

SAND2015-5869C

Analyzing Distributed Word Representations

Jacek Skryzalin, Stanford University

Project Mentor: Jeremy Wendt, 5632

Problem Statement:

Given a large collection of text, we would like to map each word or phrase into a low-dimensional vector space such that the vector space representation of each word captures its syntax and semantics.

Objective and Approach:

For words $\{w_i\}$ such that w_i co-occurs in the context of w_j with frequency $X_{i,j}$, the GloVe algorithm constructs vectors $\{v(w_i), \tilde{v}(w_i)\}$ to minimize:

$$J = \sum_{i,j} \left(\frac{X_{i,j}}{M}\right)^{\alpha} \left(v(w_i) \cdot \tilde{v}(w_j) - log\ X_{i,j}\right)^2$$

Once we obtain word vectors, we would like to exploit the vector space structure to glean insight into the words occurring in the corpus.

For example, we would expect $v(cat)$ to be very similar to $v(kitten)$ but very dissimilar to $v(sandia)$. We can use this structure to complete analogies.

Impact and Benefits:

By analyzing the distributed word representations constructed by various corpora, we gain information about the possible contexts in which a word is used; The word “won” will mean very different things when used in the contexts of sports and foreign finance (won is the currency of South Korea).

Future Work:

Given a large corpora of separate documents, we would like to perform unsupervised clustering to construct many smaller corpora. We would like to train word vectors on each corpus individually, and then experiment with ensemble algorithms to attain greater accuracy at our tasks.