Abstract

Mathijs J. Scheepers

Improving the Compositionality of Word Embeddings

We present an in-depth analysis of four popular word embeddings (Word2Vec, GloVe, fastText and Paragram) [84, 99, 17, 135] in terms of their semantic compositionality. In addition, we propose a method to tune these embeddings towards better compositionality. We find that training the existing embeddings to compose lexicographic definitions improves their performance in this task significantly, while also getting similar or better performance in both word similarity and sentence embedding evaluations.

Our method tunes word embeddings using a simple neural network architecture with definitions and lemmas from WordNet [86]. Since dictionary definitions are semantically similar to their associated lemmas, they are the ideal candidate for our tuning method, as well as evaluating for compositionality. Our architecture allows for the embeddings to be composed using simple arithmetic operations, which makes these embeddings specifically suitable for production applications such as web search and data mining. We also explore more elaborate and involved compositional models, such as recurrent composition and convolutional composition. We call our architecture: CompVec.

In our analysis, we evaluate original embeddings, as well as tuned embeddings, using existing word similarity and sentence embedding evaluation methods [30, 38]. Aside from these evaluation methods used in related work, we also evaluate embeddings using a ranking method which tests composed vectors using the lexicographic definitions already mentioned. In contrast to other evaluation methods, ours is not invariant to the magnitude of the embedding vector—which we show is important for composition. We consider this new evaluation method (CompVecEval) to be a key contribution.

Finally, we expand our research by training on a significantly larger dataset we constructed ourselves. This dataset contains pairs of titles and articles introductions from Wikipedia. Since the introduction of an encyclopedic article is definitional, we found them to be a logical progression from dictionaries. The creation of this dataset, by extracting it through the Wikipedia API, is another important contribution.