Ben, Bram, Kaiyu, Qingyuan, Yuzhe

embedding

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

If we have a good numeric representation of text data which captures even the context and semantics, we can use it for a wide variety of downstream real-world tasks like sentiment analysis, text classification, clustering, summarization, translation and so on. The fact of the matter is, machine learning or deep learning models run on numbers, and embeddings are the key to encoding text data that will be used by these models.



An embedding is a mapping of a discrete – categorical – variable to a vector of continuous numbers. In context of neural networks, embeddings are low-dimensional, learned continuous vector representations of discrete variables. Neural network embeddings are useful because they can reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space.

Neural network embeddings have 3 primary purposes:

* Finding nearest neighbors in the embedding space. These can be used to make recommendations based on user interests or cluster categories.
* As input to a machine learning model for a supervised task.
* For visualization of concepts and relations between categories.

Implementation

In the Wikipedia book project, the supervised learning task is set as predicting whether a given link to a Wikipedia page appears in the article for a book. We feed in pairs of training examples with a mix of positive – true – and negative – false – pairs. This set-up is based on the assumption that books which link to similar Wikipedia pages are similar to one another. The resulting embeddings should therefore place alike books closer together in vector space.

The network has two parallel embedding layers that map the book and wikilink to separate 50-dimensional vectors and a dot product layer that combines the embeddings into a single number for a prediction. The embeddings are the parameters, or weights, of the network that are adjusted during training to minimize the loss on the supervised task.







For this project, we are primarily interested in recommending books based on the nearest neighbors.

Application: LIWC (Language Inquiry and Word Count). The idea is creating thematic dictionaries, labeling words with dimensions depending of the subject of each dictionary.