Contribution ID: **38**   Type: **not specified**

Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs

The driving force behind the recent success of LSTMs has been their ability to learn complex and non-linear relationships. Consequently, our inability to describe these relationships has led to LSTMs being characterized as black boxes. To this end, we introduce contextual decomposition (CD), an interpretation algorithm for analysing individual predictions made by standard LSTMs, without any changes to the underlying model. By decomposing the output of a LSTM, CD captures the contributions of combinations of words or variables to the final prediction of an LSTM. On the task of sentiment analysis with the Yelp and SST data sets, we show that CD is able to reliably identify words and phrases of contrasting sentiment, and how they are combined to yield the LSTM's final prediction. Using the phrase-level labels in SST, we also demonstrate that CD is able to successfully extract positive and negative negations from an LSTM, something which has not previously been done.

Author: MURDOCH, W. James

Presenter: MURDOCH, W. James