**Reproducing Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs Executive Summary**

Track: 3, Paper: T3-03 (<https://arxiv.org/abs/1801.05453>)

Olivier Simard-Morissette. ID: 260563480,

Email: olivier.morissette@gmail.com

Ruo Yu Tao. ID: 260706629,

Email: ruo.tao@mail.mcgill.ca

Irene Qizhen Zhang. ID: 260671588,

Email: [qizhen.zhang@mail.mcgill.ca](mailto:qizhen.zhang@mail.mcgill.ca)

This reproducibility study focused on the paper, *Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs.* The paper describes an interpretation algorithm for analysing individual predictions made by standard LSTMs, without any changes to the underlying model. For a given text, and a phrase or word, the algorithm returns a CD score, which is a quantitative measure of how the subtext contributes to the LSTM’s output. These scores are computed by decomposing the cell and hidden state in an LSTM into a sum of two contributions: those made by the given phrase or word, and those by elements outside of the phrase or word. More specifically, the cell and hidden state is decomposed using linearizing activation functions.

Using this method, three experiments were completed to confirm the main findings of the paper. The first experiment’s goal was to demonstrate CD’s interpretability. To show interpretability, a logistic classifier and an LSTM were trained on the SST dataset on the task of binary (positive or negative) phrase sentiment. The linear dependence of logistic regression coefficients and the outputs of the CD algorithm were evaluated by computing the correlation coefficient. By demonstrating a higher correlation coefficient, the paper suggests that this would justify interpreting CD scores in the intuitive way that the magnitude of coefficients are interpreted in a linear model in the task of word importance. The correlation produced in the paper is 0.76, and our score is 0.5. The discrepancy in results may be explained by the difference in accuracy between the LSTM model described in the paper and the model produced in the reproducibility study. The model in the reproducibility study had an accuracy of 82.7% while the model in the paper was 87%.

The second experiment involved interpreting the results of an LSTM in the task of classifying dissenting sub-phrases. The original paper was too ambiguous in parts of the filtering method to find dissenting sub-phrases (more specifically, filtering by an arbitrary “absolute score” of 1.5 in section 4.3 of the original paper). These parts were not reproducible and were not included. The report reproduced individual examples of CD scores providing some interpretability but does not include the visually appealing heat maps available in Table 2 of the original paper. The graphs were created by aggregating the phrases into their respective labels (positive/negative) and plotting the distribution of positively labelled phrases with negatively labelled phrases against CD score. The distributions showed little separation between the two classes unlike the separation depicted in Figure 3 of the original paper. The discrepancy may be due to the differences in model performance referenced in the previous paragraph, or due to the lack of filtering by “absolute score”. If it were the latter, the arbitrary “absolute score” chosen would have added bias in the results procured by the original paper.

The last experiment evaluates how well CD captures instances of negation. A negated phrase is a “negation term” (ie. not, never etc.) followed by a phrase that is negated by the negation term. Reviews that contain a negation were filtered from the SST dataset and the CD scores from the two distributions were plotted for both positively and negatively labeled phrases. The tree-structure of the SST dataset allowed to filter for reviews containing these terms, but the reasoning behind a lot of the decisions behind the design of the experiment seem biased. One of the biggest examples of bias is how the paper filters out all negation phrases that are less than 10 words, and only shows the distribution of all phrases less than 5 words. This gives a clear separation between the positively and negatively labelled phrases. Seeing as the paper seems to arbitrarily pick shorter phrases to test, the same experiment was conducted, but with phrases with more than 10 words. The distinction between the positively and negatively labelled distribution disappears with more than 10 phrases - this seems to imply that CD doesn’t seem to be able to interpret negation in longer phrases.

The reproducibility study shows that CD can be used to decompose and interpret the results of an LSTM, but the quantifiable performance described in the paper was not reproducible. The major challenges that made it not reproducible included ambiguities in the paper and cherry-picked data to demonstrate performance.