Local vs. Global interpretations for NLP

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, WordsWorth scores have been proposed for calculating feature importance in the context of traditional deep learning models trained for text classification tasks [Anonymous, 2021]. Here, we experiment with the idea behind these scores and present them as a global explanation for a trained model. Interpretability literature shows that delete one method acts as a good explanation for NLP tasks. Since WW scores act as a good proxy to delete one scores for text classification, we extend the argument and utilize them for interpretation. We provide local and global explanations for a CNN trained on the IMDB reviews dataset by comparing these scores with LIME. Similarly to LIME, the global representation is a bag of words representation. Overall, we argue that evaluating a trained neural network on single words, at all possible locations in the input text one by one, gives powerful and valid insights into the workings of these otherwise black box models. This is a work in progress and we are looking for further tests to evaluate the usefulness of our method.

1 Introduction

Deep learning models are black boxes for the end user. Interpretability is becoming an important concern, for users to trust the model overall as well as individual predictions. Global intepretations try to provide a summary of the model in some form, whereas local interpretations are concerned with explaining model prediction on a specific input. We use the idea of WordsWorth scores which have been recently introduced, to show how to generate global explanations for a CNN, and compare it to LIME. We further argue that evaluating a traditional deep learning model on all the words in the training vocabulary gives a faithful picture of the model. Our method consists of evaluating a neural network on all the words in the training vocabulary one by one, on all possible locations. For any single evaluation, only one location in the input text has a valid word and all other locations are set to zero. This gives us some estimate of how a neural network interprets each word.

2 Related Work

Interpretability is an exciting area of research in machine learning in general and deep learning in particular. In this section we provide a brief and by no means thorough overview of some existing techniques. This is meant to give the reader a brief idea of the different avenues that are being explored for the problem of interpretability in NLP. An explanation can be global, such that it explains the overall model, or local, such that it explains the decision made by the model on a particular instance. Jacovi et al. [2018] calculate n-gram and word level scores. They calculate word-level score with two methods that they name local and global, but both these methods use leave one out evaluations and require a specific input x. Xiong et al. [2018] compare saliency maps and LRP(Layer-wise Relevance Propagation), and conclude that LRP finds more relevant features. They argue that deleting a particular word does not give a true picture of the contribution of this particular word to output

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

since it could be a part of an important n-gram, for example. Rather than single words, they highlight high level features. Also this requires knowledge of the model parameters and architecture, so this is not a black box explanation.

Chen and Ji [2020] place a mask over the word embedding layer during training, and thus propose an interpretation friendly model architecture for text classification tasks. They compare LIME and SampleShapley [Strumbelj and Kononenko, 2010] over simple neural networks and modified networks and show that modified structure is more interpretable. LIME provides local explanations by sampling from the distribution near the example and fitting a local model to it.

Lai et al. [2019] compare different methods and conclude that for deep learning model, important features are often different from features for traditional models.

Chen et al. [2020] draw inspiration from Shapley values. They generate hierarachical explanations, which might consist of multiple features that interact and drive the classifier to a particular prediction, and argue that these perform better than other popular interpretation methods.

Arras et al. [2019] experiment on LSTMs. They compare contextual decomposition,Layer-wise relevance propagation, gradient based methods and occlusion based which involve deleting a particular word. They conclude that LRP performs best, but it is a white box method.

Nguyen [2018] examine sentiment analysis and topic classification tasks. They report that leave one out approach and first derivative based generates explanations that correlate with human judgement. For LIME to perform on a comparable level, a large number of samples are required typically around 5000. Further they argue that automatic evaluation matrices align reasonably well with human judgement. They examine a logistic regression model and a simple feed forward neural network. We argue that our technique of feature importance attribution aligns closely with omission technique.

For calculating word score matrices,Xu and Du [2020] propose a method which involves using test examples.

Global explanations in computer vision Wu et al. [2020] outline a global explanations for CNNs. First they find important features, such that occluding these features will flip the prediction on all training instances for a specific class. Then they run different tests, which they call evaluating these important features for semantic tasks.

3 Single word evaluations as global interpretation

3.1 Direct evaluation on a single word

In contrast to omission techniques, we make a case for directly calculating word influence. Leave one out methods, which seem to be the inspiration for all the feature ranking methods used in NLP [Li et al., 2016], have well grounded backing in statistics as well as intuition. In classical statistics, evaluating the function on just one datapoint or feature does not make much sense.

However, the inspiration for single word evaluations comes from the fact that human generated text is a highly compact form of data. Unlike images, where an individual pixel has no significance by itself and the surrounding pixels are always needed to give it context, words carry immense information within themselves. We might expect our deep learning classifier to be able to give meaningful insights for each individual word most of the time.

3.2 The case for position invariance for isolated words

Similarly, we argue that the effect of a word on the prediction would not be highly dependent on its exact position in the input sequence, since the inherent meaning of a word is quite independent of its position in a sentence in which it appears, particularly when the surrounding words are ignored. A well trained CNN, for example, should treat the word 'beautiful' roughly the same, whether it appears near the beginning of the review or the end. If this is the case, we do not need to specifically append words at a particular position to find out their importance.

Table 1: Words with highest WordsWorth scores

Highly positive words	Highly negative words
refreshing	awful
perfect	worst
excellent	waste
superb	poorly
unexpected	fails
perfectly	disappointing
rare	disappointment
enjoyable	forgettable
delightful	unfunny
blake	unwatchable

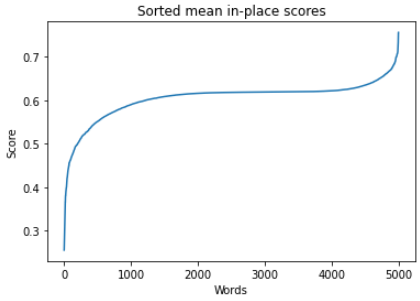

Figure 1: Mean values of in-place scores for IMDB Reviews with CNN

3.3 Experiments: IMDB Reviews with a CNN

We use the IMDB dataset [Maas et al., 2011], with 25000+25000 training+test examples of variable length. Each review in the training set has a positive/negative label attached to it. We use a simple CNN as the starting point of our experiments, with 32 dimensional embedding layer, 32 filters and 64 hidden ReLU units and a single output unit. Test accuracy is 86.7%. Output probabilities above 0.5 correspond to positive sentiment, and lower probabilities correspond to negative sentiment.Training vocabulary size is 5000 and each review is transformed to z 200 word piece of string by either appending zeros or removing the trailing aprt if it exceeds this length.

It has been shown in the original work that evaluating a trained model on a single word gives a score which is somewhat aligned with the sentiment behind that word. We go one step further and argue that a neural network trained for a sentiment analysis task on an input of size d and a training vocabulary of size v can be characterized by a $v * d$ matrix. Each row of this matrix corresponds to the scores for a particular word, and each column represents the score when this word is placed at that particular position in input. Recall that when a word is being evaluated at a particular position, all other inputs are set to zero.

Mean and standard deviation of word score matrix We compute this matrix and plot the mean and standard deviations for these scores along each row. These scores represent the trained network in a manner that is both interpretable and faithful to the representations the model has learned. Results in figures 1,2. The top words associated with each sentiment are presented in Table 2. Notice that the top 10 high standard deviation words have strong associations with the negative label.

For comparison, we present the top 10 words for each sentiment calculated through original WordsWorth scores in Table 1. Compare to mean scores and notice that the ordering for positive words is slightly different, and negative word list is exactly the same. This shows that the matrix we propose is a stable and reliable method of interpreting a neural network.

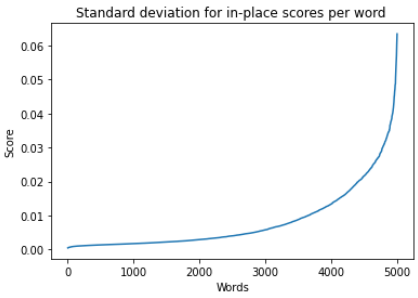

Figure 2: Standard deviation for in-place scores for IMDB Reviews with CNN

Table 2: Most important mean scores and corresponding words

Highly positive word	Scores	Highly negative words	Scores
perfect	0.7564884	awful	0.25465319
excellent	0.75603143	worst	0.2563593
refreshing	0.75512171	waste	0.26369784
superb	0.74893515	poorly	0.27800871
perfectly	0.73354795	fails	0.28250759
unexpected	0.73317402	disappointing	0.2965854
rare	0.72838456	disappointment	0.30099407
enjoyable	0.72051218	forgettable	0.30488086
delightful	0.71923713	unfunny	0.31003236
blake	0.71628927	unwatchable	0.32535568

4 LIME vs WordsWorth: A case of two reviews

Here we present two reviews and their local explanation given by LIME and global explanation given by WordsWorth scores.

Review 1: "Naturally in a film who's main themes are of mortality, nostalgia, and loss of innocence it is perhaps not surprising that it is rated more highly by older viewers than younger ones. However there is a craftsmanship and completeness to the film which anyone can enjoy. The pace is steady and constant, the characters full and engaging, the relationships and interactions natural showing that you do not need floods of tears to show emotion, screams to show fear, shouting to show dispute or violence to show anger. Naturally Joyce's short story lends the film a ready made structure as perfect as a polished diamond, but the small changes Huston makes such as the inclusion of the poem fit in neatly. It is truly a masterpiece of tact, subtlety and overwhelming beauty." Prediction: 0.99668723 Results in figures 3,4,5 and 6. The input review is overwhelmingly positive and this is depicted by

Table 3: Most and least stable words, picked by standard deviation of in-place scores

Most stable words	Least stable words
hugh	awful
face	worst
zombie	waste
thomas	poorly
walls	fails
dickens	disappointing
carter	disappointment
soderbergh	forgettable
clown	unfunny
fairy	unwatchable

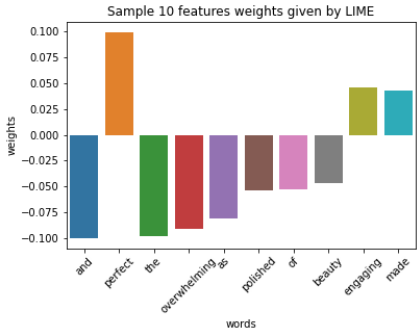

Figure 3: Top ten features highlighted by LIME in review 1

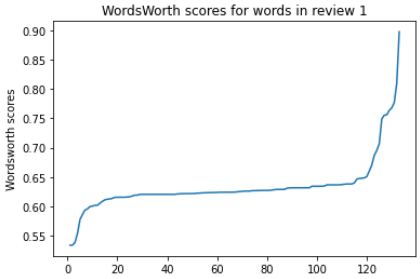

Figure 4: WordsWorth scores for all words in review 1

the WordsWorth score distribution. If a more faithful explanation is required, the scores from the word matrix we have introduced can be utilized for each word.

Review 2: "This movie is a disaster within a disaster film. It is full of great action scenes, which are only meaningful if you throw away all sense of reality. Let's see, word to the wise, lava burns you; steam burns you. You can't stand next to lava. Diverting a minor lava flow is difficult, let alone a significant one. Scares me to think that some might actually believe what they saw in this movie.

Even worse is the significant amount of talent that went into making this film. I mean the acting is actually very good. The effects are above average. Hard to believe somebody read the scripts for this and allowed all this talent to be wasted. I guess my suggestion would be that if this movie is about to start on TV ... look away! It is like a train wreck: it is so awful that once you know what is coming, you just have to watch. Look away and spend your time on more meaningful content." Prediction: 0.46570438 Results in figures 7,8,9 and 10. Notice the high number of negative words in figure 8.

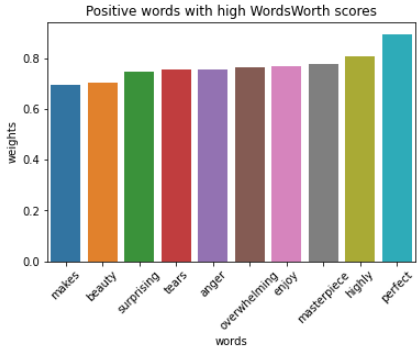

Figure 5: Most positive words for review 1

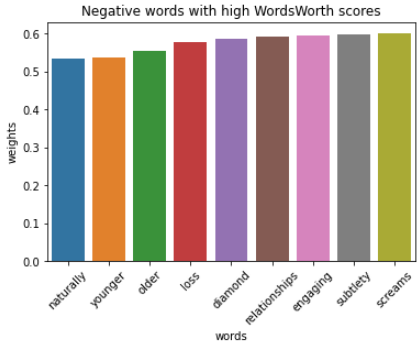

Figure 6: Most negative words for review 1

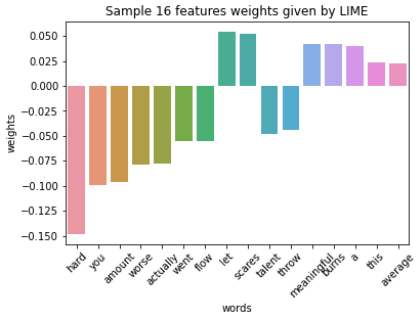

Figure 7: Top sixteen features highlighted by LIME in review 2

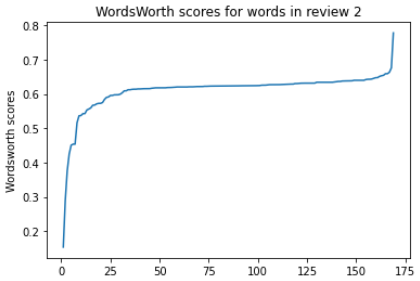

Figure 8: WordsWorth scores for all words in review 2

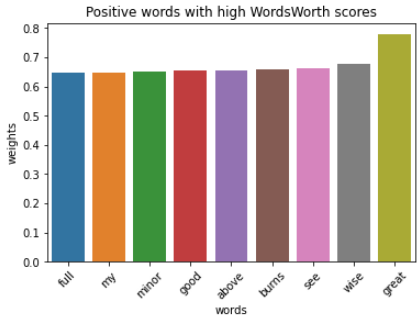

Figure 9: Most positive words for review 2

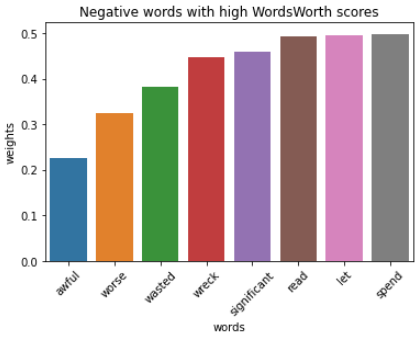

Figure 10: Most negative words for review 2

5 Conclusion

We provide a word matrix for evaluating a trained neural netowrk on a sentiment analysis task. Further we show how to provide explanations for a specific input using WordsWorth scores. These explanations are not entirely local but they still provide a useful summary of the input. We compare our results to LIME. Leave-one-out scores have been shown to be effective at explaining classifier decisions and WordsWorth scores provide a good proxy to these local explanations. Further, since these scores have been shown to be effective at attacking LSTMS as well as for topic classification taskd, they might serve as a faithful explanation in scenarios which have not been explored in this paper.

References

Anonymous. Wordsworth scores for attacking {cnn}s and {lstm}s for text classification. In *Submitted to International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=e6hMkY6MFcU`. under review.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. *CoRR*, abs/1904.11829, 2019. URL `http://arxiv.org/abs/1904.11829`.

Hanjie Chen and Yangfeng Ji. Learning variational word masks to improve the interpretability of neural text classifiers, 2020.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5578–5593. Association for Computational Linguistics, 2020. URL `https://www.aclweb.org/anthology/2020.acl-main.494/`.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. *CoRR*, abs/1809.08037, 2018. URL `http://arxiv.org/abs/1809.08037`.

Vivian Lai, Zheng Cai, and Chenhao Tan. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 486–495. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1046. URL `https://doi.org/10.18653/v1/D19-1046`.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL `http://arxiv.org/abs/1612.08220`.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P11-1015`.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 1069–1078. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1097. URL https://doi.org/10.18653/v1/n18-1097.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18, 2010. URL https://dl.acm.org/citation.cfm?id=1756007.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8649–8658. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00868. URL https://doi.org/10.1109/CVPR42600.2020.00868.

Wenting Xiong, Iftitahu Ni'mah, Juan M. G. Huesca, Werner van Ipenburg, Jan Veldsink, and Mykola Pechenizkiy. Looking deeper into deep learning model: Attribution-based explanations of textcnn. *CoRR*, abs/1811.03970, 2018. URL http://arxiv.org/abs/1811.03970.

Jincheng Xu and Qingfeng Du. On the interpretation of convolutional neural networks for text classification. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2252–2259. IOS Press, 2020. doi: 10.3233/FAIA200352. URL https://doi.org/10.3233/FAIA200352.