Explaining Compositional Semantics in Neural Networks via Neighborhood Sampling and Decomposition

Anonymous ACL submission

1 Introduction

Deep neural networks have achieved impressive performance on multiple natural language processing tasks by learning complicated composition rules of words and phrases. LSTM (Hochreiter and Schmidhuber, 1997) and Transformers (Vaswani et al., 2017) are popular networks for modeling language, capturing human-like semantics especially when pretrained on large corpora (Devlin et al., 2018; Peters et al., 2018). However, it is non-trivial to understand how atomic words and their compositions contribute to the final results, leaving these models as "black boxes".

Recently, researchers study post-hoc explanation methods to explain neural networks without modifying the inner structure. Additive feature attribution methods (Lundberg and Lee, 2017; Ribeiro et al., 2016; Binder et al., 2016; Shrikumar et al., 2017) regard the model prediction as a weighted sum of contributions of input words and divide the effort of the final prediction to each atomic word. Another line is based on input occlusion (Kádár et al., 2017), which masks a word or phrase in an example and observes the change in the prediction. Although the algorithms explain which words and phrases are important to one specific prediction, these explanations provide limited insights on how the model handles complicated semantics like stress or negation, and how the atomic words and phrases interact and compose into high-level semantics. Contextual decomposition (Murdoch et al., 2018) is a recently proposed explanation method which tackles the challenge above. The algorithm computes *individual* contributions of words and phrases by decomposing outputs of each layer in the neural network. With the help of extracted individual contributions of phrases, it is possible to explain compositionality in semantics with simple strategies. For example, by calculating individual contributions for each node on a parsing tree of the input sentence, it can be explained how the model composes semantics of the root from subtrees and leaves.

However, contextual decomposition actually follows heuristics on calculating individual contributions of phrases. A formal definition of individual contributions, which is a crucial concept in the algorithm, is not provided mathematically. This leads to heuristic designs in some critical decomposition steps in the algorithm. (Singh et al., 2018) find that original contextual decomposition does not perform well on deeper neural networks, and modified it again with heuristics. In contrast, we will provide a formal way to quantify individual contributions of phrases.

2 Related Works

In this section, we discuss related works on post-hoc neural network explanations. (Guidotti et al., 2018) categorize explanation methods into global explanation methods, local explanation methods, and model inspection methods. We focus our discussion on global and local explanation methods.

Global explanation methods include fitting target black boxes with self-interpretable models such as trees (Craven and Shavlik, 1996; Krishnan et al., 1999), or assessing what training features the model regards as most significant (Vidovic et al., 2016; Doshi-Velez and Kim, 2017; Sonnenburg et al., 2008). (Zien et al., 2009) proposed Feature Importance Ranking Measure (FIRM), which was later generalized by (Vidovic et al., 2016) into Measure of Feature Importance (MFI) score to identify important pixels and k-mers for image and genome classification tasks. Given a subset of the dataset, the feature importance is calculated as the average prediction score for each example containing that feature. Unfortunately, the sparsity of an expression in natural language makes it infeasible for natural language processing tasks.

Local explanation methods provide explanations that are specific to an example. Input occlusion based methods calculate the contribution of a phrase as the difference between the prediction of the original input and that of the masked input. The phrases are either omitted (Kádár et al., 2017), or padded to a reference value (Li et al., 2016). Another family of local explanation methods is additive feature attribution methods (Lundberg and Lee, 2017), where the final prediction is divided additively to each atomic word. LIME (Ribeiro et al., 2016) fits a local linear model directly around a data point. Layer-wise relevance back-propagation (LRP) (Binder et al., 2016) and DeepLIFT (Shrikumar et al., 2017) back-propagates activation differences from outputs layer to input layers to assign contribution scores for inputs. Gradient based (Simonyan et al., 2013; Hechtlinger, 2016; Denil et al., 2014; Ancona et al., 2017) and integrated gradient based (Sundararajan et al., 2017) methods evaluate feature importance with output gradients or the integrated gradients from a reference input with respect to input features. (Lundberg and Lee, 2017) unified the additive feature attribution approaches above with a Shapley value assignment based framework.

References

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer.

Mark Craven and Jude W Shavlik. 1996. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30.

Misha Denil, Alban Demiraj, and Nando De Freitas. 2014. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. *arXiv preprint arXiv:1611.07634*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

R Krishnan, G Sivakumar, and P Bhattacharya. 1999. Extracting decision trees from trained neural networks. *Pattern recognition*, 32(12).

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Chandan Singh, W James Murdoch, and Bin Yu. 2018. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*.

Sören Sonnenburg, Alexander Zien, Petra Philips, and Gunnar Rätsch. 2008. Poims: positional oligomer importance matricesunderstanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–i14.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Marina M-C Vidovic, Nico Görnitz, Klaus-Robert Müller, and Marius Kloft. 2016. Feature importance measure for non-linear learning algorithms. *arXiv preprint arXiv:1611.07567*.

Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. 2009. The feature importance ranking measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 694–709. Springer.