Progress Report: Rationalizing Neural Predictions Replication in PyTorch and Robustness Analysis

Jason Wang ⋆
jw7383@nyu.edu

Chloe Zheng ⋆
cz1300@nyu.edu

Xu Han ⋆
xh852@nyu.edu

Completed Experiments

We have managed to both predict sentiment and select rationales on the BeerAdvocate dataset and evaluate each part with the metrics mean squared error (MSE) and precision, respectively, replicating the two core experiments from Rationalizing Neural Predictions. Our code is stored here. We trained, tested, and evaluated the model on the "appearance" aspect, although one issue we're currently working on is poor sentiment prediction performance that's not yet matching the paper's. In the upcoming weeks, we will debug the encoder, train and obtain results for the two other aspects that the paper experimented on, smell and palate, and complete our extension. Since the source code from the paper used theano for its implementation, to replicate the experiment we rewrote the model in PyTorch. We used another repository as a reference. The repository, owned by Adam Yala, also implements the model in PyTorch, but Yala's work mainly supports training on the sklearn NewsGroup dataset. So, we spent time adding components that Yala's repository didn't include, which included adapting Yala's work to correctly preprocess the BeerAdvocate dataset, debugging the generator model and training function, and adding precision calculations for the generated rationales.

Results

For our results, we performed hyperparameter tuning on the selection and continuity lambdas over the 3 pairs of values provided by the paper, 0.0002, 0.0003, and 0.0004 for the selection lambdas, and the continuity lambdas being twice those values. To cut down on training time, we chose the model with the best MSE and precision after 10 epochs. We obtained a MSE of 0.54 on the predicted sentiments and a precision of 0.802 on the generated rationales. The precision falls within the paper's precision range of 80-96%, although the MSE is not close to the paper's MSE of 0.0087.

Issues

We ran into issues getting the model to train successfully. Initially the rationales were only one word long and were often a common frequent token in the review such as the token "the". We adjusted the hyperparameter values for selection and continuity lambdas, which tune the length and continuity of the rationales, as well as debugged the generator model, so that our rationales make sense. Here is an example of a generated rationale for appearance: "burnished copper-brown topped by a large beige head". However, we are still having issues matching the paper's MSE on the predicted sentiment scores, which we believe is likely due to bugs in the encoder model. Our plan is to triage the errors and debug the encoder in the same way we did with the generator.

Planned experiments

Once we have a fully working model, we can proceed smoothly with training, testing, and evaluating the model with full epochs on the 3 aspects, appearance, smell, and palate. Finally, we will work on our extension with the Rotten Tomatoes reviews dataset. This will require slight modifications to our model in order to preprocess the additional dataset correctly. This will allow for sentiment prediction and evaluating on MSE. We are then going to manually annotate 100 reviews on which phrases contribute the most to the review's sentiment. This way we can also evaluate the generated rationales both qualitatively through inspection and quantitatively through precision.

References

Tao Lei, Regina Barzilay, Tommi Jaakkola. 2016. Rationalizing Neural Predictions. arXiv:1606.04155

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078.

B Kim, JA Shah, and F Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In Advances in Neural Information Processing Systems

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015. Molding cnns for text: non-linear, non-consecutive convolutions. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?": Explaining the predictions of any classifier. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Omar Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pages 260–267.

Ye Zhang, Iain James Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. CoRR, abs/1605.04469.

Michiel Hermans and Benjamin Schrauwen. 2013. Training and analysing deep recurrent neural networks. In Advances in Neural Information Processing Systems, pages 190–198.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv:1707.07250

M.D. Devika, C. Sunitha, Amal Ganesh. 2016. Sentiment Analysis: A Comparative Study on Different Approaches. In Procedia Computer Science, Volume 87, Pages 44-49, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2016.05.124.

Shi, Tian and Rakesh, Vineeth and Wang, Suhang and Reddy, Chandan K. 2019. Document-Level Multi-Aspect Sentiment Classification for Online Reviews of Medical Experts. In Association for Computing Machinery.

L. Stappen, A. Baird, L. Schumann and S. Bjorn. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. In IEEE Transactions on Affective Computing. doi: 10.1109/TAFFC.2021.3097002.

Finale Doshi-Velez, Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608.