Quantitative Analysis of Explanation Methods

Silan He
Student
McGill University
silan.he@mail.mcgill.ca

Anirban Laha
Mentor
Université de Montréal
MILA
anirban.laha@umontreal.ca

Abstract

For more widespread adoption of machine learning models such as neural networks, models must be able to provide interpretable and robust explanations for their decisions. To this end, interpretability of machine learning models has gained traction in recent years. To improve explanation of novel machine learning approaches, we need to analyze if inaccuracies are due to model level inaccuracy or explanation model inaccuracy. Integrated Gradients by (Sundararajan et al., 2017) is an explanation technique that compares the direct relationship of the input with the labels. Contextual Decomposition by (Murdoch et al., 2018) relies on model parameters only, thus only indirectly explaining the relationship between the input and their respective output labels. Both of these interpretability techniques were showcased in an empirical manner. This project will start by reproducing the baselines of each of these papers. This project will then leverage the verification framework and evaluation criteria proposed by (Camburu et al., 2019). Thus, a standard recurrent neural network or RNN will be trained for text entailment for the SNLI dataset (Bowman et al., 2015). The neural model will be explained using Contextual Decomposition and Integrated Gradients where their explanations will be evaluated quantitatively on e-SNLI (Camburu et al., 2018). We hope to determine the efficacy of each approach while determining the correctness of the explanation models themselves.

1 Introduction

Neural networks, amongst other machine learning models, have been successfully applied to a plethora of NLP tasks. However, the vector based neural models generated are basically impossible to interpret meaningfully. . Neural networks easily match or outperform other state-of-the-art systems in many NLP tasks. Unfortunately, the underlying machinations of a neural model are still poorly understood. They are still popularly referred to as black boxes. Indeed, output labels of these models are not easily traced to their respective input texts. This makes these black boxes really hard to iterate or fix.

Interpretability of a model allows humans to validate and improve their work. Interpretability refers to explaining predictions produced by the model. As correlation often does not equal causality, a solid model understanding is needed when it comes to making decisions and explaining them. Proper interpretation of our state-of-the-art models could become a pillar to solving a slew of current and future neural model issues. The advancement of interpretability will allow machine learning models to be used on real world problems with much higher degrees of trust (Ribeiro et al., 2016) (Camburu et al., 2019).

Interpretability techniques are still in their infancy stages. Current post-hoc explainers have only been thoroughly validated on simple models, such as linear regression. When applied to real-world neural networks, explainers are commonly evaluated under the assumption that the learned models behave reasonably. However, neural networks often rely on unreasonable correlations, even when producing correct decisions (Camburu et al., 2019).

So far, many proposed explanation models have been evaluated empirically. Namely, Integrated Gradients by Sundararajan et al. (2017) and Contextual Decomposition by Murdoch et al. (2018). These explanation models have yet to be evaluated quantitatively on NLP tasks for accuracy of explanations.

Integrated Gradients estimates the integral of the gradients of the output labels with respect to the input by comparing with a suitable baseline. The method shows great qualitative results in a slew of

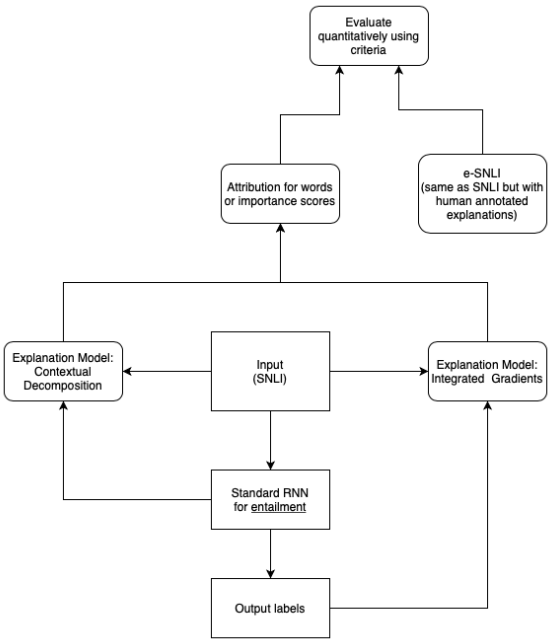

Figure 1: Methodology for Verification Framework

NLP and vision tasks (Sundararajan et al., 2017).

Contextual Decomposition was compared to Integrated Gradients by means of a correlation coefficient between logistic regression coefficients and extracted scores from the respective models. Indeed, authors of Contextual Decomposition claim that the quantitative score for a phrase's contribution to the LSTM's prediction corresponds to the input to a logistic regression (Murdoch et al., 2018). They show this relationship with all the tested models by graphing the relationship between the attribution score and the logistic regression inputs.

We will reproduce both Integrated Gradients and Contextual Decomposition on a standard neural network to do the SNLI task(Bowman et al., 2015). We aim to deeply understand the model at this stage of the process.

Then, we wish to leverage the verification framework and evaluation critera presented by Camburu et al. (2019) and explain a standard neural network trained for entailment on the popular Stanford Natural Language Entailment dataset or SNLI dataset (Bowman et al., 2015). e-SNLI (Camburu et al., 2018) or explained-SNLI is an extended SNLI with an additional layer of human-annotated natural language explanations of the entailment relations.

2 Related Works

The work presented is closely related to work done by Camburu et al. (2019). However, we will be evaluating the correctness of the explanation models on their e-SNLI dataset (Camburu et al., 2018). The paper uses their metrics to evaluate SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016) and L2X (Chen et al., 2018). They explained an RCNN described by Lei et al. (2016). We will also be evaluating on an entailment task defined by SNLI instead of sentiment analysis tasks defined by SST and Yelp dataset (Murdoch et al., 2018). Camburu et al. (2019) evaluated explanation models based on the the following error metrics.

- Percentage of instances for which the most important token provided by the explainer is among the non-selected tokens
- Average number of non-selected tokens ranked higher than any clearly relevant token
- Percentage of instances for which at least one non-selected token is ranked higher than a clearly relevant token

Our work extends theirs as we will additionally evaluate with the following success criteria:

- Percentage of exact matches of top scored tokens with explanation labels
- Percentage of partial matches of top scored tokens with explanation labels
- Average overlap with explanation labels using some similarity measure such as Jaccard distance

Many explanation models will do their own analysis of their model vs various others. Namely, Integrated Gradients by Sundararajan et al. (2017) and Contextual Decomposition by Murdoch et al. (2018). The Contextual Decomposition paper by Murdoch et al. (2018) was compared to Integrated Gradients by means of a correlation coefficient between logistic regression coefficients and extracted scores from the respective models. We have reasonable doubt with respect to what this correlation really shows and hope to inspect the impact of this assumption of the model's performance using a more fundamentally sound criteria. For these reasons, we will reproduce these models on the SST (Socher et al., 2013) and spend significant time trying to understand them at a baseline level. Then, we will test using our expanded quantitative verification criteria.

Ribeiro et al. (2016) quantitatively evaluates heuristic human and simulated user simulations built to score trustworthiness by their own metrics. Our approach differs in that we will directly score similarity of top explained labels with annotated gold standard explanation labels from the e-SNLI(Camburu et al., 2018).

Acknowledgments

Grateful to Anirban Laha for proofreading the proposal and mentoring the project. Thanks to Aanika Rahman for inspring the project.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can i trust the explainer? verifying post-hoc explanatory methods.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations.

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

A Appendices

A.1 SST

A.2 e-SNLI

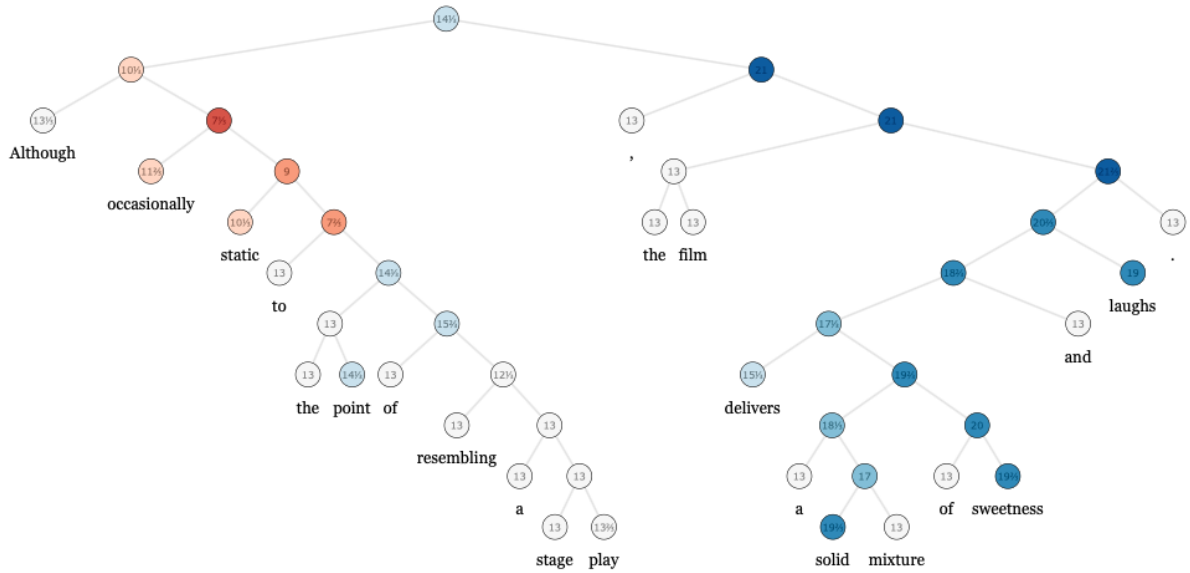

Figure 2: Example from SST.

Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother is playing with her daughter in a swing. Label: neutral Explanation: Child does not imply daughter and woman does not imply mother.
Premise: A man in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over a pickup truck implies that he is touching it.

Table 1: Excerpts from e-SNLI. Annotators were given the premise, hypothesis and label. They highlighted the words that they considered essential for the label and provided explanations. (Camburu et al., 2018)