Zoya Goel, Srikar Viswanatha

Project Proposal

One of the biggest areas of research in Artificial Intelligence is in teaching systems how to reason and observe relationships between two features in a given input. This has always been one of the main objectives in developing these systems, allowing them to develop their own intuition and dependencies given raw data. Understanding and modeling these interactions is crucial because real-world decisions—whether in autonomous driving, medical diagnosis, or natural language understanding—rarely depend on isolated factors but rather on complex interdependencies among multiple elements. As AI systems increasingly influence critical domains, the ability to not only make accurate predictions but also explain their reasoning through illuminating their understanding of these relationships becomes essential for trust, accountability, and further innovation. Thus, this area of study is not only extremely popular, but essential for the advancement of AI.

Much work has been done with respect to identifying and characterizing feature interactions, one such work being "Explaining Local, Global, And Higher-Order Interactions In Deep Learning" by Lerman et al. (2021). The authors discuss their method, Taylor-Neural Interaction Detection (T-NID) and Taylor-CAM, their method for visualizing feature interactions. T-NID utilizes cross derivatives to identify pairwise and higher order interactions between features. This in turn helps to measure how the influence of a feature on the output changes depending on the other feature (Lerman et al. 2021). Taylor-CAM is an extension of a popular program in computer vision that highlights multiple important regions in an image for a network's decision. These two innovations helped to develop a more human and accurate line for computer reasoning.

Another relevant work is "Explaining Explanations: Axiomatic Feature Interactions for Deep Networks (Janizek et al., 2020) " which introduces Integrated Hessians, a method to quantify pairwise feature interactions in neural networks. This method extends Integrated Gradients by applying it recursively to itself, capturing second-order interactions between features. Integrated Gradients (Sundarajan et al., 2017) is an earlier attribution method designed to explain the predictions of deep neural networks by assigning importance scores to each input feature. It is based on the idea of computing the gradient of the model's output with respect to its input along a path from a baseline input to the actual input. For Integrated Hessians (Janizek et al., 2020) , the key formula involves integrating the second-order partial derivatives of the model function along a path from a baseline input to the target input, ensuring that the method satisfies key interpretability axioms such as interaction completeness and symmetry. Unlike previous methods, Integrated Hessians generalizes across different neural network architectures and efficiently computes interactions even for large numbers of features. To handle ReLU-based networks, which lack second derivatives, the authors propose replacing ReLU activations with the smooth SoftPlus function during explanation without retraining. Empirical evaluations show that Integrated Hessians outperforms existing methods in accurately identifying interactions while maintaining computational efficiency.

Along with this, we will analyze the work Integrated Directional Gradients (IDG) (Sidkar et al., 2021), a feature interaction attribution method designed to explain neural network predictions in NLP models. It extends Integrated Gradients by incorporating cooperative game theory principles to measure the importance of groups of features rather than individual ones. IDG calculates attribution scores by computing directional derivatives along a straight-line path from a baseline input to the actual input, capturing both individual feature importance and interactions. The method ensures interpretability by satisfying a set of axioms, including monotonicity, superadditivity, and sensitivity. To efficiently compute attributions, IDG uses a hierarchical structure, such as a parse tree, to group features at different levels, from individual words to entire phrases.

Our inspiration for the goal of our proposed method is “Asymmetric feature interaction for interpreting model predictions” by Lu et al. (2023), showing the possibility of utilizing asymmetric feature interaction, using what is called an Asymmetric Shapley Interaction Value (ASIV) as the basis for measuring relations . Instead of looking for symmetry in relations, where we can assume a bidirectional arrow from thought A to B, we would only look at directed arrows, meaning directed relationships. With T-NID, a symmetric algorithm, we are not able to unearth these types of relationships, which are more common than one can imagine. One of the best examples of this is any buildup of logic where time is a factor, such as in video reasoning. Say that a video of a car driving, then stopping at a yield sign is provided, with a comparison of ASIV values, we could train our model to reason that since the car was previously driving, then stops when the light is red, that cars will stop when a light is red. In moments where action B depends on A specifically, and not the other way around, ASIV would be a great way to also process our data through (Lu et al. 2023). We will attempt to combine all of these methods together in order to obtain a higher benchmark than the Lerman et al. (2021)’s findings. Being able to provide higher or more significant benchmarks will not only provide immense value in the AI-reasoning space, but this technology can be essential in real-world decisions such as in autonomous driving.

First, we aim to carry out benchmark testing for the methods described above. We will use a variety of metrics that the current literature uses. Lerman uses Area Under the Curve (AUC) to measure pairwise interactions. They further utilize AUC for multi-order interactions by assessing AUC scores in an order-by-order fashion (Lerman et al. 2021). There are also metrics that evaluate how removing the influential features identified by these methods impacts the model’s prediction confidence. Lu et al. (2023) uses Area Over the Perturbation Curve (AOPC) , which quantifies the average drop in the model’s output probability for the predicted class after the most important words/features are removed. Similarly, they also use Log-Odds Ratio (LOR), where instead of removing features, it replaces them with a padding token, measuring the log-probability change of the predicted class before and after feature modification. In slight contrast, Janizek re-trains the model after removing important features to see if the model still performs well without them, checking whether the detected features are truly important for generalization (Janizek et al. 2020). However, Sikdar highlights issues with these kinds of methods, specifically AOPC and LOR (Sikdar et al. 2021). These methods often involve removing features from the input and measuring performance drops. which creates out-of-distribution inputs. These can be misleading as neural networks are not trained to handle these perturbations. For this reason, Sikdar et al. (2021) argues for a qualitative approach. In their work, IDG is applied to state-of-the-art text classifiers on three benchmark sentiment analysis datasets. The authors present visualizations of IDG's attribution scores, demonstrating how it captures interactions between words, especially in negations and conjunctions. Therefore, in our benchmark, we aim to use both quantitative and qualitative metrics to ensure thoroughness.

Once we establish our benchmarks, we hope to create a new method, utilizing insights from the methods that perform the best on the benchmarks, in addition to potentially applying an asymmetric paradigm similar to the one introduced in Lu et al. (2023). We are particularly interested in considering asymmetry as all of these methods (excluding ASIV) assume that relationships between features are symmetric. However, Lu et al. (2023) shows how this might not be necessarily true in an NLP context. Specifically, they show how the meaning of words in a sentence is contextual, and one word can modify another without equal influence in return . Methods that illustrate asymmetric relationships between features might allow us a richer understanding of NLP model inference.

Overall, the remainder of the paper will follow an organized structure, beginning with an explanation of the original benchmarks and the criteria we'll use to evaluate our success. This will be followed by an exploration into the alternative methods, such as asymmetric features, integrated hessians, integrated directional gradients, and Shapley values, analyzing their approaches and comparing their performance against the benchmarks established in the initial paper .The discussion will flow logically, offering a comprehensive assessment of how our proposed method stands in relation to existing techniques, culminating in a clear measure of their relative strengths and limitations.

References

Janizek, J. D., Sturmfels, P., & Lee, S.-I. (2020). Explaining explanations: Axiomatic feature interactions for deep networks. *arXiv preprint arXiv:2002.04138*. https://arxiv.org/abs/2002.04138

Grabisch, M., & Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory, 28*(4), 547–565. https://doi.org/10.1007/s001820050125

Lu, X., Ma, J., & Zhang, H. (2023). Asymmetric feature interaction for interpreting model predictions. *arXiv preprint arXiv:2305.07224*. https://arxiv.org/abs/2305.07224

Lerman, S., Venuto, C., Kautz, H., & Xu, C. (2021). Explaining local, global, and higher-order interactions in deep learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1224–1233). IEEE. https://arxiv.org/abs/2006.08601

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 (ICML'17). JMLR.org, 3319–3328.

Sikdar, S., Bhattacharya, P., & Heese, K. (2021). Integrated Directional Gradients: Feature Interaction Attribution for Neural NLP Models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 865–878). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.71