Michael Bai
Discussion for 2/26/26

"Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation" explores the topic of local interpretability in deep learning, focusing on attribution methods used to explain individual decisions made by Deep Neural Networks (DNNs). The study focuses on the challenge of approximating Shapley values, a set of numerical scores derived from cooperative game theory that assigns importance scores to input features. More specifically, it is the average marginal contribution of a player to all possible "coalitions" or subsets of the total player count that can be formed without it [1].

While Shapley values are theoretically ideal because they are the only method satisfying important axioms like completeness, symmetry, and continuity of attributions needed to be mathematically consistent, calculating Shapley values is NP-Hard [1]. It is only feasible to calculate Shapley values for models with a few dozen features, making their calculation prohibitively expensive for larger deep-learning tasks. By overcoming this computational bottleneck, the authors aim to make formally resource-intensive, game-theoretic explanations accessible for complex, real world models. Providing a practical way to calculate these values is a step towards shifting AI away from being a "black-box" into transparent systems that users can trust.

The research question is "Does there exist a polynomial-time algorithm that can accurately approximate Shapley values in deep neural networks better or comparable than existing attribution methods?" The authors claim that their proposed algorithm, Deep Approximate Shapley Propagation (DASP), can reliably approximate Shapley values in polynomial time by representing the contributions of different feature coalitions as probability distributions and propagating these distributions through the network's layers. By treating the presence or absence of features as a Bernoulli distribution, DASP estimates how the mean and variance of the network's activations change at each layer [1]. This allows the model to account for complex, multi-linear feature interactions in poly-time which cannot be accomplished using existing "fast" methods such as DeepLift that rely on linear assumptions and heuristics [1].

DASP was tested against established benchmarks such as Integrated Gradients, DeepLIFT Rescale, Shapley Sampling and KernelSHAP in three distinct ML domains: predicting disease progression in the Parkinson's Telemonitoring dataset, identifying genetic patterns in DNA sequences, and recognizing handwritten digits in the MNIST collection [1]. The authors validate DASP by calculating exact Shapley values for small networks to serve as a "ground truth" for statistical comparison. They then calculate the Root Mean Squared Error (RMSE) to measure numerical precision and the Spearman rank correlation of all established benchmarks against the values to demonstrate the accuracy and efficiency of DASP.

At a low number of network evaluations (1-10 evaluations), DASP significantly outperforms all other methods in error reduction and feature ranking accuracy [1]. While NP-Hard methods like KernelSHAP and Shapley Sampling eventually achieve lower RMSE and higher Spearman scores at near-exhaustive iterations, DASP reaches a high accuracy score almost immediately, showing that it provides a reliable, polynomial-time approximation that closely matches the quality of exponential-time calculations far more efficiently than traditional sampling-based methods [1].

References: [1] Ancona, M., Öztireli, C., & Gross, M. (2019, May). Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International conference on machine learning* (pp. 272-281). PMLR.