Fast Inference Services for Alternative Deep Learning Structures

Eduardo Romero, Christopher Stewart and Nathaniel Morris
The Ohio State University

ABSTRACT

AI inference services receive requests, classify data and respond quickly. These services underlie AI-driven Internet of Things, recommendation engines and video analytics. Neural networks are widely used because they provide accurate results and fast inference, but it is hard to explain their classifications. Tree-based deep learning models can provide accuracy and are innately explainable. However, it is hard to achieve high inference rates because branch misprediction and cache misses produce inefficient executions. My research seeks to produce low latency inference services based on tree-based models. I will exploit the emergence of large L3 caches to convert tree-based model inference from sequential branching toward fast, in-cache lookups. Our approach begins with fully trained, accurate tree-based models, compiles them for inference on target processors and executes inference efficiently. If successful, our approach will enable qualitative advances in AI services. Tree-based models can report the most significant features in a classification in a single pass. In contrast, neural networks require iterative approaches to explain their results. Consider interactive AI recommendation services where users seek to explicitly order their instantaneous preferences to attract preferred content. Tree-based models can provide user feedback much more quickly than neural networks. Tree-based models also have less prediction variance than neural networks. Given the same training data, neural networks require many inferences to quantify variances of borderline classifications. Fast tree-based inference can explain variance in seconds (versus minutes). Our approach shows that competing machine learning approaches can provide comparable accuracy but desire wholly different architectural and platform support.

ACM Reference Format:
Eduardo Romero, Christopher Stewart and Nathaniel Morris The Ohio State University. 2019. Fast Inference Services for Alternative Deep Learning Structures. In *Proceedings of SEC '19: ACM/IEEE Symposium on Edge Computing, Arlington, VA, USA, November 7–9, 2019 (SEC '19),* 3 pages.
https://doi.org/10.1145/3318216.3363331

1 MOTIVATION

Many modern Deep Learning models are based on neural networks, however, these models are not generally explainable. The lack of explainability of neural networks can be a concern for deployment of the model and even more so when an important decision is linked to the outcome of the model [10]. Some efforts have been directed at being able to explain these models, however, there is often a tradeoff with accuracy or inference throughput, as is the case with saliency maps [5] and causal inference [8]. On the other hand, decision trees are capable of making lightweight, fast inferences, and their inferences are explainable through linear models [12]. However, decision trees alone are not competitive in accuracy with neural networks. Random forests can achieve higher accuracy than individual decision trees, but they sacrifice inference throughput while still not achieving neural network levels of accuracy. Other more complex tree-based models, such as deep forests [15], are capable of maintaining the explainability of decision trees while achieving higher accuracy, but they still cannot achieve high inference throughput. The goal of our research is to support strict service level objectives (SLOs) on this type of model. A service level objective demands high throughput, scalability and low response time [2, 6, 7, 9]. For example, a strict SLO may require processing 1,000 queries per second while ensuring that the 95^{th} percentile response is less than 250 milliseconds [4, 11, 13, 14].

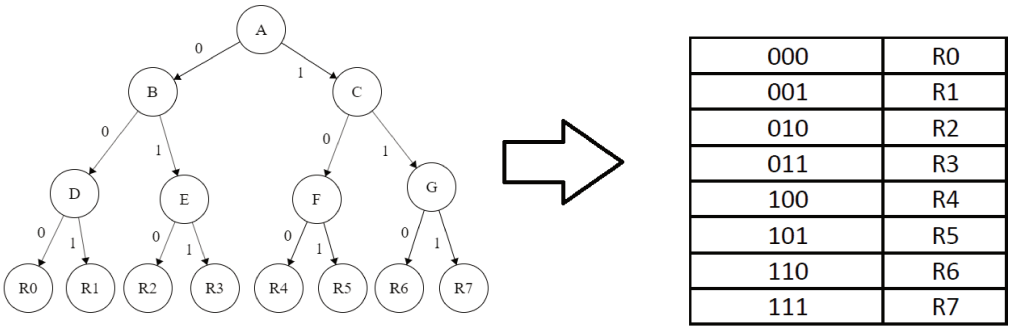
AI model	*Accuracy*	*Explainable*	*Inference Throughput*
CNN	Hi	Lo	Hi
CNN+Saliency map	Hi	Hi	Lo
Decision Tree	Lo	Hi	Hi
Random Forest	Lo	Hi	Lo
Deep Forest	Hi	Hi	Lo
Proposed Research	**Hi**	**Hi**	**Hi**

Table 1: Accuracy, inference and interpretability of competing AI models.

Inference on tree-based models struggles to achieve strict SLO because inference is often executed inefficiently. Tree-based models use branching to produce an inference, and therefore, can pay penalties for branch misprediction. These penalties can limit the time that a model takes to return an inference result. Since only one path from the root to a leaf of the tree is used on a given sample, and that path is determined by the values in the sample features, all the branches can be known before doing the first computation on the tree. Therefore, constant branching increases the probability of mispredicting a path while it is not necessary. Thus, an approach that reduces branching in a tree could increase the inference rate of a tree-based model. One such approach could involve computing the entire path and performing a single lookup to retrieve the inference result.

SEC '19, November 7–9, 2019, Arlington, VA, USA

ACM ISBN 978-1-4503-6733-2/19/11…$15.00
https://doi.org/10.1145/3318216.3363331



000	R0
001	R1
010	R2
011	R3
100	R4
101	R5
110	R6
111	R7

Fig. 1: Example of binary decision tree processed as into an array of responses.

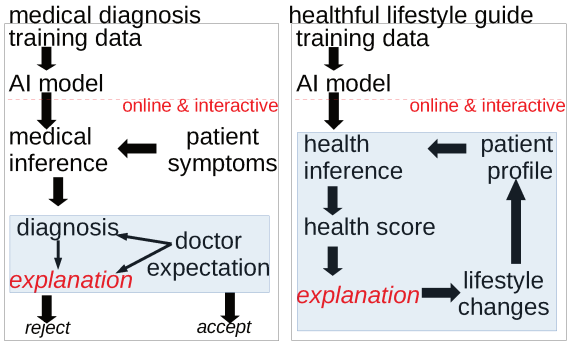


Fig. 2: Example work flows that require explainable AI models.

2 PROPOSED RESEARCH

Our approach is based on two key ideas, the realization that any decision tree that is not binary can be turned into a binary tree, and that a decision tree can be viewed as a Deterministic Finite Automaton (DFA). Both these ideas allow us to view the response of a decision tree on some input as an accepting state of a DFA that takes a sequence of Boolean values as input. Therefore, this determinism can be exploited by linking each of the possible outcomes of the decision tree to a binary string with length equal to the height of the corresponding branch (Fig.1). Further processing is required to guarantee all paths to leaves have the same length and that the different input features being considered across the tree appear in the same order. Once this processing is done, there is an injective function from the possible paths of the decision tree to the set of binary strings of length equal to the number of distinct nodes in the tree. Clearly these binary strings can be made to represent a lookup address. Thus, by this approach we can reduce cache misses and branch mispredictions, while doing only one lookup per input for one tree. Naturally, this approach can be extended to more complex tree-based models such as a Random Forest, although more complex models would require an extra step to make sure paths of different trees do not map to the same address.

3 POTENTIAL APPLICATIONS

Our goal is for this approach to increase the inference rate on existing explainable tree-based models. Such models have already been shown to perform competitively in terms of accuracy to neural networks [12, 15], and if successful, our approach could increase their throughput without sacrificing accuracy or explainability. Producing a fast, accurate and explainable model can be very important in several situations, for instance the medical field [1, 3]. As depicted in Fig.2, a model designed to assist in medical diagnosis requires explainability both from the doctor's and the patient's perspective. Similarly, a model suggesting changes in the lifestyle of a patient should provide an explanation to help convince them [8]. However, potential applications are not limited to the medical field, any application that requires an accurate model with certain level of explainability, can profit from the increased inference throughput our approach hopes to provide. Ultimately, fast inferences for tree-based models could benefit any AI-driven application requiring an alternative to neural networks.

REFERENCES

[1] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2016.

[2] J. Ding, R. Cao, I. Saravanan, N. Morris, and C. Stewart. Characterizing service level objectives for cloud services: Realities and myths. In *IEEE International Conference on Autonomic Computing*, 2019.

[3] Y. Hayashi. A neural expert system with automated extraction of fuzzy if-then rules and its application to medical diagnosis. In *Advances in neural information processing systems*, 1991.

[4] C. Holmes, D. Mawhirter, Y. He, F. Yan, and B. Wu. Grnn: Low-latency and scalable rnn inference on gpus. In *ACM EuroSys*, 2019.

[5] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *International conference on machine learning*, pages 597–606, 2015.

[6] J. Kelley, C. Stewart, N. Morris, D. Tiwari, Y. He, and S. Elnikety. Measuring and managing answer quality for online data-intensive services. In *ICAC*, 2015.

[7] J. Kelley, C. Stewart, N. Morris, D. Tiwari, Y. He, and S. Elnikety. Obtaining and managing answer quality for online data-intensive services. In *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 2017.

[8] E. Kıcıman and M. Richardson. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[9] N. Morris, S. M. Renganathan, C. Stewart, R. Birke, and L. Chen. Sprint ability: How well does your software exploit bursts in processing capacity? In *ICAC*, 2016.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin. 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*, pages 1135–1144. ACM, 2016.

[11] C. Stewart, A. Chakrabarti, and R. Griffith. Zoolander: Efficiently meeting very strict, low-latency slos. In *IEEE International Conference on Autonomic Computing*, 2013.

[12] R. Tanno, K. Arulkumaran, D. C. Alexander, A. Criminisi, and A. V. Nori. Adaptive neural trees. *CoRR*, abs/1807.06699, 2018.

[13] C. Zhang, M. Yu, W. Wang, and F. Yan. Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving. In *USENIX Annual Technical Conference*, 2019.

[14] M. Zhang, S. Rajbhandari, W. Wang, and Y. He. Deepcpu: Serving rnn-based deep learning models 10x faster. In *USENIX Annual Technical Conference*, 2018.

[15] Z. Zhou and J. Feng. Deep forest: Towards an alternative to deep neural networks. *CoRR*, abs/1702.08835, 2017.