Fusing Deep Learning and Fuzzy Logic: A Framework for Adaptive and Scalable Interpretability

Yumin Zhou

Nanyang Technological University
50 Nanyang Ave, Singapore 639798
S230038@e.ntu.edu.sg

Abstract

Deep learning models offer state-of-the-art performance but their inherent opacity is a major barrier to adoption in high-stakes domains. In contrast, Takagi-Sugeno-Kang (TSK) fuzzy systems provide rule-based transparency but often lack the predictive power of deep networks. My PhD research addresses this critical trade-off by developing the **Fuzzy-Modulated Linear Consequents (FMLC)** framework, a novel hybrid architecture that synergizes these two paradigms. The core of FMLC is a deep neural network that processes fuzzified input features to generate context-dependent "modulators". These modulators dynamically parameterize a TSK-style linear consequent layer, creating a model that is both highly performant and inherently interpretable. My latest work, **Learnable-FMLC (L-FMLC)**, advances this by introducing a regularized, adaptive fuzzification layer that autonomously learns the optimal fuzzy partitions from data, and a two-stage rule distillation framework to ensure interpretability remains scalable in high-dimensional problems. This research delivers a validated, theoretically-grounded, and scalable framework, contributing a significant step towards transparent and trustworthy AI.

Research Problem and Objectives

A central challenge in AI is the trade-off between predictive performance and interpretability. While deep neural networks (DNNs) are highly accurate, their "black-box" nature hinders their use in critical domains where decision-making must be transparent (Rudin 2019). Post-hoc explanations offer insights but may lack fidelity (Lundberg and Lee 2017). Conversely, transparent Takagi-Sugeno-Kang (TSK) fuzzy systems often cannot match the predictive power and scalability of deep models, even in neuro-fuzzy forms like ANFIS (Takagi and Sugeno 1985; Jang 1993).

My PhD research answers the question: *How can we fuse the representation power of deep learning with the structured transparency of TSK fuzzy systems?* The objectives are to: (1) design the hybrid L-FMLC architecture for state-of-the-art performance; (2) create mechanisms for scalable interpretability; (3) overcome the limitations of prior neuro-fuzzy systems; and (4) establish the framework's theoretical foundations.

The L-FMLC Framework and Contributions

This research introduces the Learnable Fuzzy-Modulated Linear Consequents (L-FMLC) framework (Figure 1).

Prediction Path: An input vector X is fed to a novel **Regularized Adaptive Fuzzification Layer**, which autonomously learns the optimal centers and widths of Gaussian membership functions. A `DeepModel` (e.g., Transformer) processes these fuzzified features to generate context-dependent "modulators" (N). These modulators then parameterize a **Dynamic Linear Consequent Layer**, which computes the prediction y_p as a precise, instance-specific linear function of the original inputs X:

$$y_p = \sum_{s=1} \left(\sum_{i=1} W_{s,i} N_i \right) X_s \tag{1}$$

where W are trainable weights, providing a clear linear explanation for each prediction.

Interpretation Path: To make interpretability scalable, I developed a **Two-Stage Rule Distillation Framework**.

1. **Architectural Pruning:** The adaptive fuzzification layer inherently prunes the rule space by learning a parsimonious set of active fuzzy sets for each feature, drastically reducing the number of empirically activated rules compared to a fixed-grid approach.
2. **Hierarchical Rule Clustering:** The remaining raw TSK-style rules are then vectorized and grouped using Hierarchical Agglomerative Clustering. This process distills thousands of raw rules into a handful of high-fidelity "meta-rules", each representing a core operational regime of the model. As a final step, these structured meta-rules can be translated into fluid natural language narratives using an LLM.

Research Progress to Date

As a third-year part-time student, my research has progressed significantly.

- **Framework Development:** I have designed and implemented the foundational FMLC concept and its advanced successor, L-FMLC, featuring the adaptive fuzzification layer and scalable rule distillation pipeline.
- **Empirical Validation:** Extensive experiments show L-FMLC achieves competitive or SOTA accuracy against

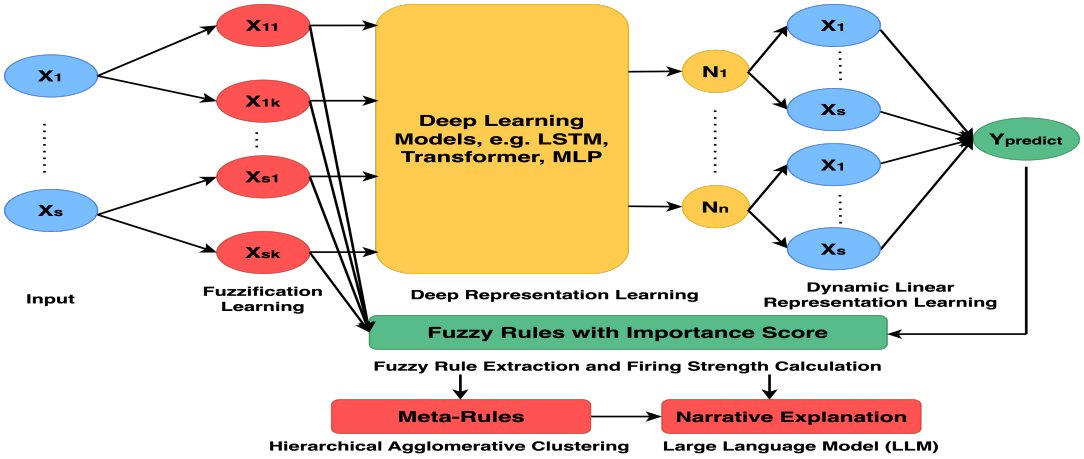

Figure 1: The L-FMLC framework, illustrating its dual pathways. The **prediction path (top)** uses a deep model on adaptive fuzzified inputs to generate modulators (N) for the final linear layer. The **interpretation pipeline (bottom)** extracts Type-I TSK rules, distills them into "meta-rules" via clustering, and finally leverages an Large Language Model to generate human-readable natural language explanations.

strong baselines like Transformers (Vaswani et al. 2023) and EBMs (Nori et al. 2019). Ablation studies confirm that our regularization is essential for both interpretability and superior generalization.

- **Scalable Interpretability:** I have demonstrated that L-FMLC solves the "rule explosion" problem. On a 20-feature dataset, the framework reduced over 19,000 potential rules to just 5 comprehensible meta-rules with $> 94\%$ fidelity.
- **Peer-Reviewed Dissemination:** My research has been accepted at the **the Interpreting Cognition in Deep Learning Models (CogInterp) Workshop at NeurIPS 2025**. A full journal paper is also under review by *IEEE Transactions on Fuzzy Systems*, and another work has been submitted to **ICLR-2026**.

Proposed Future Research

My remaining PhD work will extend the L-FMLC framework along three thrusts.

Thrust 1: Rigorous Theoretical Guarantees (Year 3). I will formalize and extend the theoretical underpinnings. This involves (a) refining the universal approximation theorems for the L-FMLC architecture under learnable MFs, and (b) analyzing the convergence properties of the training algorithm under its non-convex optimization landscape.

Thrust 2: Integrating Uncertainty Quantification (Year 3). To enhance reliability, I will integrate UQ by implementing Bayesian and ensemble-based variants of FMLC to provide calibrated uncertainty estimates for both predictions and rules.

Thrust 3: Expanding Applicability and Scope (Year 4). I will investigate the adaptation of the FMLC paradigm to more complex data modalities. This includes designing `Vision-FMLC` (using CNN backbones) and `NLP-FMLC` (using Transformer embeddings) to bring structured interpretability to image and text-based tasks.

Anticipated Thesis Contribution

This thesis will contribute the L-FMLC framework, a validated "glass-box" solution to the performance-interpretability dilemma that fuses deep learning and fuzzy logic. It delivers three key innovations: (1) a novel, self-structuring neuro-fuzzy architecture; (2) a scalable, multi-level interpretability framework that converts rules to natural language explanations; and (3) comprehensive empirical and theoretical validation. Ultimately, this work offers a robust pathway towards more powerful, transparent, and trustworthy AI systems.

References

Jang, J.-S. R. 1993. ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3): 665–685.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nori, H.; Jenkins, S.; Koch, P.; and Caruana, R. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154.

Takagi, T.; and Sugeno, M. 1985. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1): 116–132.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.