Learning Through Concepts: Hierarchies, Logic and Reasoning

Deepika SN Vemuri

Indian Institute of Technology, Hyderabad
ai22resch11001@iith.ac.in

Abstract

This thesis aims to bridge the gap between data-driven models and symbolic learning through the lens of Concept-Based Learning, a paradigm that guides model learning through high-level, human-understandable concepts. Here, models first learn a set of concepts, subsequently using them to perform a task of interest. Prior work on concept-based models has largely focused on relatively simple classification settings, where classes are linear combinations of pre-specified concepts; treating concepts largely as tools to increase interpretability, rather than as fundamental building blocks of the learning process itself. In contrast, this thesis explores the broader potential of concepts, as the core units of representation and reasoning in neural network models, capable of shaping how models learn and generalize.

Main Research Question: How to bridge the gap between data-driven models and symbolic learning?

Background and Context: Most state-of-the-art computer vision models are sub-symbolic and work with raw data like pixels. Two key problems with these models are that: 1) they are black-boxes, which limits their deployment in high-stakes applications, and 2) the pixels themselves lack semantic structure, making them inherently difficult to reason with. Humans, as opposed to this, learn conceptually; raising several questions. For instance, how do humans form concepts in the first place? How does the human brain organize and store these concepts? And how do we know which level of granularity of a concept to retrieve according to the task and context at hand? Replicating this paradigm of learning in AI models could help answer these questions while also making strides in giving these models the ability to plan, reason, and better understand the world around them.

This thesis centers around *Concept-Based Learning*, an approach that aims to guide the learning process of models through high-level, human-understandable concepts, with a focus on the interpretability of these models. Here, the concepts that the model learns are used to perform a task of interest. For example, in classification, the model could learn to look for concepts like *fur, whiskers* and *mammal* to classify an input as a *cat*. The concepts hence serve as a looking glass into the model, capturing semantic abstractions.

Related Work: Concept-based models (CBMs) started as a way to train ante-hoc interpretable models for classification tasks (Koh et al. 2020), due in part, to the inherent limitations of post-hoc interpretability. Several aspects of these models have been studied like addressing concept leakage, quantifying uncertainty and improving the robustness of these models. Other efforts have attempted the usage of LLMs and VLMs for concept guidance and annotations (Oikarinen et al. 2023). Finally, some very recent works have attempted to apply concept-based learning in foundation models (Choi et al. 2025). However, in all such works, concepts have largely been treated as tools to increase interpretability, rather than as fundamental building blocks of the learning process itself. Concepts can play a more central role as the core units of representations and reasoning in neural network architectures. This lies at the core of this thesis and is what we have been exploring through some threads like learning logical relations between concepts, how to formally structure concepts and how concepts can help with reasoning.

Contribution 1. Logic-Enhanced Concept-Based Learning (WACV 2026): In most current CBMs, the concept-to-class mapping is deliberately kept simple (usually just a linear layer). This is for interpretability - so that we can infer the importance of each concept present in a particular class. However, a linear mapping is inherently limiting and prevents the model from leveraging higher-order relations between concepts. For example, consider a model learning to recognize an arctic fox. Arctic foxes have either white fur or brown fur depending on the environmental conditions (like an exclusive OR operation). But a linear layer cannot capture such nuances. So the question we ask in this work is

> How can we go beyond a linear concept-to-class mapping while retaining interpretability?

Logic operations present a natural way to model such relations in a structured manner. To this end, we introduce a logic module comprising differentiable fuzzy logic gates that learn predicates (logical combinations of concepts) for each class. Importantly, logic gates are interpretable non-linearities and using their fuzzy versions (t-norm) enables end-to-end learnability. We experimentally observe that logic improves multiple aspects of a model's behav-

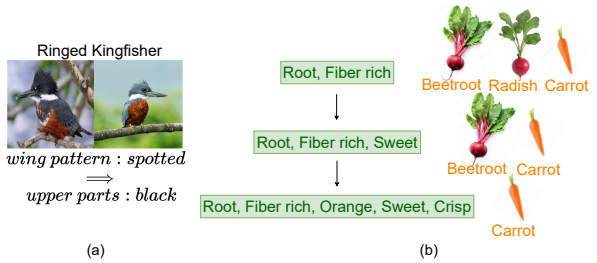

Figure 1: (a) Sample result from our method. A *Ringed Kingfisher* has black upper parts if it has a spotted wing pattern. (b) Attributes shared by more classes are more general.

ior leading to better accuracy, more effective interventions and improved interpretability. Fig 1(a) shows an example of a predicate captured for the class *Ringed Kingfisher* from the CUB200 dataset. The differentiable logic neurons in our method converge to the logic gates that minimizes the objective the most - not requiring any ground truth logic. So, to validate the goodness of the learned logic, we evaluate our method on synthetic datasets (we introduce a new dataset), where the class-level ground truth logic is known. Finally, we also introduce a new metric to evaluate concept-based models in a worst case setting. Specifically, we measure the change in confidence of a model when a misleading concept is corrected to its ground truth and observe that our logic-based models do the best in terms of this metric.

Contribution 2. Formal Concept-Based Models (CVPR-W 2025, Under Review at A* Conference): From another perspective, most current CBMs do not impose much structure over concepts. It is known that deep neural networks learn hierarchical representations - early layers capture general properties like texture and later layers more class specific information. The exact nature of these representations remains opaque, necessitating the need for more interpretable models. Existing CBMs, although interpretable to a certain degree, typically learn all concepts at a single layer, overlooking the inherent hierarchical structure in neural network representations across multiple layers. So the question we ask in this work is

> Can we align a neural network's depth-wise representation with a semantic hierarchy?

To this end, we explicitly guide the network to learn general concepts in early layers and specific ones in deeper layers. In our concept-based classification setting, we define a hierarchy using subset-superset relations. For example, a beetroot and carrot are *roots*, are *sweet* and *dense* and are *fiber-rich*; while a beetroot, carrot and radish are *roots* and are *fiber-rich* (Fig 1(b)). Here, the latter group is more general as it spans more classes. To model such a semantic structure, we draw on Formal Concept Analysis (FCA) to construct a concept lattice from binary concept-class associations. We then show how to provide supervisory signals from the lattice to learn attributes at different layers in the network according to their level of generality - each deeper group being more specific. For example, for the class *White Stork* in the Imagenet100 dataset, our model learns these attribute sets at different depths in the network: {*animal, vertibrate*}, {*animal, vertibrate, webbed feet, long beak*}, {*animal, vertibrate, a baby, webbed feet, long beak, white feathers*} (subsets of learned attributes provided for simplictiy).

We empirically observe that our models learn more interpretable embeddings, support more effective interventions, and learn concept representations that are hierarchically structured. Finally, we also define the notion of a *multi-level intervention* which is enabled by having access to learned attribute sets of varying granularities at different points in the network. We determine what set of attributes to intervene on based on the severity of the misclassification (we intervene on more general attributes the more severe the misclassification). We see that such interventions are more effective in general and enable a finer level of control when intervening.

Finally, this thesis has also contributed to complementary efforts on introducing multimodal concepts for continual learning (**AAAI 2025**).

Ongoing/ Future Work: Building on these contributions, my ongoing and future work aims to **extend the role of concepts to large-scale transformer models.** One immediate direction is to investigate how concepts could enhance the reasoning capabilities of transformers, in the context of **parameter efficient fine-tuning (PEFT)** methods such as LoRA, which add small trainable components while keeping most pretrained weights frozen. Although efficient, these fine-tuning strategies remain opaque. To address this, we propose a shift in perspective: finetune concepts instead of raw weights. A second ongoing direction we have been exploring is on the **generalization of reasoning in multimodal large language models (MLLMs).** Current MLLMs struggle to transfer reasoning strategies to unseen domains, similar to domain generalization in classification. We are exploring this in the context of **VQA (visual question answering)**, where the model is finetuned on several source VQA tasks and is expected to generalize to an unseen target VQA task. Our approach introduces concepts as abstract reasoning steps, using them as transferable units the model can use to adapt its reasoning strategy across domains. The manuscripts for these works are currently under preparation and are planned submissions to upcoming conferences.

To conclude, we envision this thesis to contribute to a concept-centered paradigm of learning to make models more generalizable, trustworthy and human-centered.

References

Choi, J.; Raghuram, J.; Li, Y.; and Jha, S. 2025. CONDA: Adaptive Concept Bottleneck for Foundation Models Under Distribution Shifts. In *The Thirteenth International Conference on Learning Representations*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Oikarinen, T.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.