Gaoxiang Huang
Applicant for a PhD in Computer Science

Research Interests Deep learning has achieved unprecedented success in fields such as image recognition and natural language processing, driving the rapid development of AI. However, its inherent black-box nature makes the decision-making process difficult to explain and trust. This opacity is particularly critical in high-stakes domains requiring rigorous credibility, such as medical diagnosis and legal contract review. While domain-specific models based on Large Language Models (LLMs) often yield intuitively correct conclusions, their internal mechanisms remain inscrutable. Currently, addressing errors relies heavily on reactive and superficial measures—such as collecting failure cases to construct preference pairs for alignment—rather than correcting the underlying reasoning logic. I aim to bridge this gap by contributing to the development of intuitive, truly intelligent, and efficient AI models that can seamlessly interpret, interact with, and learn from the multifaceted world through transparent and controllable representations.

The begin of my research journey My curiosity in understanding how closely intelligent machines can emulate human thinking and understanding led me to open the research journey myself in my bachelor. That time, my research focused on modelling human actions in emergency situations and simulating multi-agent behavior in the Anylogic environment. The classic Social Force Model (SFM), which comprehensively analyzes physical actions in real-time, allows multi-agent systems using this model to exhibit performance comparable to real human behavior in most social interaction scenarios. However, SFM overemphasizes social interactions and overlooks the significant changes in human behavior under dangerous circumstances. Such behaviors are not deliberate decisions but rather instinctive reactions driven by primal fears—of fire, cliffs, or creatures like snakes. Under fear, human responses are almost entirely subconscious, simple, and directly triggered, underscoring the urgent need for specialized simulation algorithms for emergency crowd evacuation, a field currently lacking sufficient research. Recognizing this gap, I collected real human evacuation data from the 2021 Beijing Jiaotong Research Institute subway drill. I then utilized the Entropy Weight Method to calculate the weight of each behavioral indicator. By introducing a regularizing vector, a "fear factor," derived from these weighted indicators, to improve the classic SFM, we enabled agents to respond to emergency environments and complex interactions. This innovation resulted in Anylogic simulation outcomes that more closely align with real-world data distributions. resulting in a publication at Journal of System Simulation (Chinese, EI).

Expanding Horizons: Venturing into Machine Learning Driven by my interest in Artificial General Intelligence, I transitioned my major from Statistics to Computer Science in 2024 during my Master's program. I subsequently collaborated with a teammate to develop a framework designed to defend against a new type of concept-level backdoor attack in deep neural networks. Unlike conventional backdoor attacks that directly embed triggers into the dataset's ground truth, concept-level attacks are far more insidious and severe. They specifically target highly sensitive datasets—such as those in the medical or financial domains—which rely on secondary annotations (or "concepts") between the sample and the final ground truth. This is done to ensure the final prediction is explainable by the antecedent concepts, thereby achieving higher trustworthiness. However, these secondary annotations are fragile; their contamination is less intuitive than polluting image samples and lacks a unified ground truth, making the compromise difficult for human experts to detect. Crucially, research in this area is severely lacking. Our defense method addresses this vulnerability by using, resulting in a publication[3]

under review in ICLR 2025 k-means clustering in the intermediate layers of the network to isolate the corrupted secondary annotations. The final prediction is then determined by a voting mechanism relying only on the remaining uncorrupted concepts. Tested on Concept Bottleneck Models based on ResNet18 and VGG architectures, using the CUB-200-2011 and AwA2 datasets, our framework significantly mitigated the impact of this novel backdoor attack. Specifically, the average attack success rate for groups with three or more concepts was dramatically reduced from over 90% to below 20%. This work provides a detailed methodology in this frontier domain, supported by extensive theoretical proofs.

Towards Holistic Concept-Based Interpretability Re-examining the field of AI interpretability, recent community work has increasingly deviated from the original goal of trustworthy AI—namely, exploring the internal mechanisms of black-box architectures and seeking alignment between human and AI reasoning. Concept-based interpretability methods, for instance, aim to build secondary annotation (concept) datasets so that networks can learn the relationships among visual patterns, intermediate concepts, and ground truth. Clearly, to ensure decision credibility, these intermediate concept annotations must be formulated jointly by human experts and domain-specific regulations.

However, many current research efforts, often driven by the need to highlight novelty or chase trends, undermine interpretability. For example, some approaches use Large Language Models (LLMs) to assist with concept annotation [1, 2], but these annotations often lack strict formal standards and may bear no actual relevance to the ground truth. The resulting models achieve neither the performance of true black-box systems nor a highly understandable process mechanism, running contrary to the objectives of trustworthy AI.

To truly break the bottleneck of the performance-interpretability trade-off, I initiated a study from fundamental Convolutional Network (CNN) architectures and discovered an inherent conflict between neuron polysemy and disentangled concepts. Qualitatively, this manifests as a severe mismatch between the visual patterns in the input data and the intermediate layer concept vectors. Therefore, I proposed a gradient-based method that utilizes a specially designed loss function to separate visual features of different semantics. Furthermore, it automatically optimizes the alignment between diverse visual patterns and intermediate concepts by leveraging the incoming gradient signals at the intermediate layers. This approach is demonstrated to reduce the entropy due to polysemy judgment, ensure consistency between concepts and visual patterns, thereby pushing beyond the bottleneck trade-off of interpretability and performance. Resulting in a publication under review[**?**].

Future Research Agenda For my PhD, I intend to explore the frontiers of **Multimodal Learning** and **Efficient Knowledge Transfer**. A critical challenge in modern deep learning lies in the tension between increasing model capacity and achieving fine-grained control and efficient generalization in downstream tasks. Particularly in cross-modal scenarios involving complex instructions (e.g., image generation and editing), existing models often struggle with precise instruction adherence and frequently exhibit **Concept Hallucination**.

My research aims to establish a **unified concept-level knowledge representation framework**. By addressing the efficient transfer and alignment of knowledge across different modalities (e.g., from visual features to semantic concepts) and heterogeneous

architectures (e.g., from ViT to CBM), I seek to achieve controllable multimodal learning for complex tasks. My research will revolve around two primary questions:

RQ1: Learning Universal Knowledge Representations across Models and Modalities. I will address the core problem of extracting and aligning underlying shared knowledge from heterogeneous architectures (e.g., ViT and CBM) and associating it with human-interpretable semantic concepts to facilitate efficient knowledge transfer. My approach follows a multi-stage strategy. First, I will train a Universal Sparse Autoencoder (USAE) to establish a shared, sparse "machine language" dictionary between ViT visual patches and CBM convolutional features. This shared representation serves as the foundation for cross-model transfer. Second, leveraging the semantic labels provided by CBMs, I will train a semantic translator to map the $C_{machine}$ vector space to the C_{human} space. This "modal bridge" will link abstract low-level visual features with high-level semantic concepts, laying the groundwork for controllable multimodal learning.

RQ2: Enhancing Knowledge Transfer and Generalization via Concept Intervention. Building on the aligned concept framework, I will focus on leveraging distinct concept boundaries to achieve goal-driven knowledge transfer. I propose a concept-level causal intervention mechanism to address complex multimodal instruction adherence. Specifically, this mechanism allows for the precise localization of target concepts (e.g., color, shape) within the representation and enables "surgical" interventions in the intermediate layers of ViT inference via activation modification (e.g., concept ablation or injection). Through rigorous Concept Ablation and Activation Patching experiments, I will quantify the causal effect of specific semantic concepts on model decisions. This approach is key to selective knowledge transfer: by suppressing spurious correlations while enhancing critical discriminative concepts, I aim to significantly improve model robustness and generalization across datasets. Ultimately, this work will provide the theoretical and practical basis for interpretable multimodal generative models (e.g., solving fine-grained control in Diffusion Models) and next-generation controllable transfer learning frameworks.

Why the University of California, San Diego

References

[1] Panousis K P, Ienco D, Marcos D. Coarse-to-fine concept bottleneck models[J]. Advances in Neural Information Processing Systems, 2024, 37: 105171-105199.

[2] Tan A, Zhou F, Chen H. Explain via any concept: Concept bottleneck model with open vocabulary concepts[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 123-138.

[3] Songning Lai, Yu Huang, Jiayu Yang, **Gaoxiang Huang**, Wenshuo Chen, Yutao Yue. Guarding the Gate: ConceptGuard Battles Concept-Level Backdoors in Concept Bottleneck Models. Under Review, 2024.

[4] **Gaoxiang Huang**, Songning Lai, Yutao Yue. Towards more holistic interpretability: A lightweight disentangled Concept Bottleneck Model. Under Review, 2025.