Information-Theoretic Approaches to Out-of-Distribution Detection and Hallucination Detection in Machine Learning Systems

PhD Dissertation Proposal Defense

Your Name

Your University

October 7, 2025

The Reliability Crisis in Modern ML

- **Safety-Critical Deployments**: ML systems in healthcare, autonomous driving, and financial services
- **Two Fundamental Failures**:
  - Overconfident predictions on unfamiliar inputs (OOD detection)
  - Plausible but false content generation (hallucination detection)
- **Real-World Consequences**:
  - Medical imaging models misclassifying rare conditions
  - Autonomous vehicles failing on novel scenarios
  - AI assistants providing incorrect medical/legal advice
- **Current Gap**: Lack of principled theoretical frameworks for reliability

Information Theory Foundations

- **Mutual Information**: $I(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y})$
  - Quantifies shared information between variables
  - Measures reduction in uncertainty about $\mathbf{X}$ given $\mathbf{Y}$
- **Information Bottleneck Principle**:

$$\mathcal{L}_{IB} = I(\mathbf{Z};\mathbf{Y}) - \beta I(\mathbf{Z};\mathbf{X})$$

  - Compress toward minimal sufficient statistics
  - Discard "irrelevant" information during learning
- **Why Information Theory for ML Reliability?**
  - Provides quantifiable, objective measures of uncertainty
  - Unifies OOD detection and hallucination detection under common framework

Presentation Roadmap

- **Three Interconnected Contributions**:
  1. Label Blindness: When unlabeled OOD detection ignores critical information
  2. Domain Feature Collapse: Why single-domain models fail at OOD detection
  3. Hallucination Detection: Information-theoretic framework for LLM reliability
- **Presentation Structure**:
  - High-level overview of all three contributions
  - Deep technical dive into each contribution
  - Research timeline and expected impact
- **Unifying Theme**: Information theory as principled framework for AI safety

Contribution 1: Label Blindness in Unlabeled OOD Detection

- **Core Problem**: Unlabeled OOD detection methods ignore critical label information
  - When $I(\mathbf{z}_{unsup}; \mathbf{y}) = 0$ (feature independence from labels)
  - Guaranteed failure when unsupervised features $\perp$ supervised features
- **Novel Insight**: Adjacent OOD evaluation paradigm
  - Example: Dog breeds dataset - 80% breeds as ID, 20% breeds as OOD
  - Reveals systematic failures hidden by traditional distant OOD benchmarks
- **Theoretical Contribution**: Label Blindness Theorem
  - Formal proof of when and why unlabeled methods fail
  - Information-theoretic conditions for detection success/failure
- **Practical Impact**: Guides method selection and hybrid approaches

Contribution 2: Domain Feature Collapse

- **Phenomenon**: Single-domain training discards domain-specific features
  - $I(\mathbf{x}_d; \mathbf{z}) = 0$ for learned representations $\mathbf{z}$
  - Example: X-ray model confidently classifying MRI scans
- **Theoretical Foundation**: Information Bottleneck drives inevitable collapse
  - $\mathcal{L}_{IB} = I(\mathbf{Z}; \mathbf{Y}) - \beta I(\mathbf{Z}; \mathbf{X})$
  - Domain features $\mathbf{x}_d$ discarded when $I(\mathbf{x}_d; \mathbf{Y}) = 0$
  - Mathematical proof of collapse under supervised learning
- **Solution**: Two-stage domain filtering framework
  - Stage 1: Domain-level detection (preserve $\mathbf{x}_d$ during training)
  - Stage 2: Class-level detection within correct domain
- **Impact**: First formal characterization + practical mitigation strategy

Contribution 3: Information-Theoretic Hallucination Detection

- **Central Hypothesis**: Hallucinations arise from insufficient mutual information
  - $I(\mathbf{x}; \mathbf{y}) < \tau_{critical}$ between queries and responses
  - Layer-wise information degradation in transformer architectures
- **Novel Method**: Contrastive mutual information estimation
  - Real-time detection without external knowledge bases
  - Scalable to large language models (GPT, BERT, T5, Mamba)
  - Question-answer consistency across transformer layers
- **System Architecture**: Two-stage detection framework
  - Primary model: Standard transformer inference
  - Secondary analysis: Contrastive MI estimation
- **Validation**: Natural Questions, TriviaQA, HaluEval, TruthfulQA, HalluLens

Formal Problem Definition

- **Out-of-Distribution Detection Task**:
  - Given: Training data $\mathcal{D}_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from distribution P_{ID}
  - Goal: Detect test samples $\mathbf{x}_{test} \sim P_{OOD}$ where $P_{OOD} \neq P_{ID}$
- **Unlabeled vs. Supervised Methods**:
  - Unlabeled: Use only $\{\mathbf{x}_i\}$ (ignore labels $\{y_i\}$)
  - Supervised: Use full training data $\{(\mathbf{x}_i, y_i)\}$
- **Label Blindness Definition**:
  - Unlabeled method fails when $I(\mathbf{z}_{unsup}; \mathbf{y}) = 0$
  - Where $\mathbf{z}_{unsup}$ are features learned without supervision
- **Research Question**: When do unlabeled methods systematically fail?

Information-Theoretic Analysis

- **Information Bottleneck in Unsupervised Learning**:

$$\mathcal{L}_{unsup} = I(\mathbf{Z}; \mathbf{X}) - \beta I(\mathbf{Z}; \mathbf{Y}) \tag{1}$$

- **Bottleneck Compression Effect**:
  - Unsupervised methods minimize $I(\mathbf{Z}; \mathbf{X})$ without label guidance
  - Compression discards features that correlate with labels $\mathbf{Y}$
  - Result: $I(\mathbf{z}_{unsup}; \mathbf{y}) \to 0$ as compression increases
- **Label Blindness Theorem**: When bottleneck compression removes label-relevant features:

$$I(\mathbf{z}_{unsup}; \mathbf{y}) = 0 \Rightarrow \text{AUC}_{f_{unsup}} \leq 0.5 + \epsilon \tag{2}$$

- **Critical Insight**: Unsupervised compression inherently conflicts with label preservation

Adjacent OOD Evaluation Paradigm

- **Traditional OOD Benchmarks** (hide label blindness):
  - CIFAR-10 (ID) vs. SVHN (OOD) - different domains, easy to distinguish
  - ImageNet vs. Textures - unsupervised features sufficient
- **Adjacent OOD Protocol**:
  - Split single dataset: 80% classes as ID, 20% classes as OOD
  - Examples: Dog breeds, Bird species, Fine-grained categories
  - Forces reliance on label-dependent features
- **Key Insight**: Adjacent OOD reveals when $I(\mathbf{z}_{unsup}; \mathbf{y}) \approx 0$
- **Experimental Validation**:
  - Unlabeled methods: 50-60% AUC (random performance)
  - Supervised methods: 80-90% AUC (strong performance)

Empirical Validation Results

- **Adjacent OOD Benchmark**:
  - Faces, Cars, Food datasets (1/3 classes held out as OOD)
  - Repeated 5 times with different random seeds
- **Methods Compared**:
  - Unlabeled: SimCLR KNN, SimCLR SSD
  - Supervised: MSP (Maximum Softmax Probability)
- **Results Summary** (AUROC scores):
  - **Faces**: Supervised MSP 70.8$\pm$0.3, SimCLR KNN 52.0$\pm$4.2
  - **Cars**: Supervised MSP 69.2$\pm$0.9, SimCLR KNN 52.5$\pm$0.4
  - **Food**: Supervised MSP 78.8$\pm$1.2, SimCLR KNN 61.1$\pm$2.8
- **Key Finding**: Unlabeled methods perform near-random ($\approx$50% AUROC)

Hybrid Approach Solutions

- **Motivation**: Combine strengths of both approaches
  - Unlabeled: Good for distant OOD (domain shift)
  - Supervised: Essential for adjacent OOD (within-domain)
- **Hybrid Architecture**:

$$\text{Score}_{hybrid} = \alpha \cdot \text{Score}_{unsup} + (1 - \alpha) \cdot \text{Score}_{sup} \tag{3}$$

- **Adaptive Weighting Strategy**:
  - Estimate $I(\mathbf{z}_{unsup}; \mathbf{y})$ during training
  - High MI $\Rightarrow$ increase α (trust unsupervised)
  - Low MI $\Rightarrow$ decrease α (trust supervised)
- **Performance**: Achieves best of both worlds across OOD types

Label Blindness: Key Takeaways

- **Theoretical Contribution**:
  - First formal characterization of when unlabeled OOD detection fails
  - Information-theoretic conditions: $I(\mathbf{z}_{unsup}; \mathbf{y}) = 0$
  - Rigorous proof connecting mutual information to detection performance
- **Methodological Innovation**:
  - Adjacent OOD evaluation paradigm reveals hidden failures
  - Exposes limitations of current benchmarking practices
- **Practical Impact**:
  - Guides method selection based on OOD type
  - Hybrid approaches for robust detection across scenarios
- **Future Directions**: Extend to other unsupervised learning tasks

Mathematical Formalization

- **Single-Domain Dataset Definition**:
  - Input $\mathbf{x} = [\mathbf{x}_d, \mathbf{x}_y]$ where $\mathbf{x}_d$ are domain features, $\mathbf{x}_y$ are class features
  - Domain features: imaging modality, sensor type, capture conditions
  - All samples share same domain: $f_d(\mathbf{x}_d) = d_1$ (constant)
- **Information Bottleneck Objective**:

$$\mathcal{L}_{IB} = I(\mathbf{Z}; \mathbf{Y}) - \beta I(\mathbf{Z}; \mathbf{X}) \tag{4}$$

- **Domain Feature Collapse Theorem**: For single-domain training:

$$I(\mathbf{x}_d; \mathbf{y}) = 0 \Rightarrow I(\mathbf{x}_d; \mathbf{z}) = 0 \tag{5}$$

- **Consequence**: Learned representations $\mathbf{z}$ contain no domain information

Theoretical Analysis of Collapse

- **Why Collapse Occurs**:
  - Domain features $\mathbf{x}_d$ are independent of labels: $I(\mathbf{x}_d; \mathbf{y}) = 0$
  - Including $\mathbf{x}_d$ in $\mathbf{z}$ increases complexity without improving prediction
  - Bottleneck compression discards "irrelevant" domain information
- **Formal Proof Sketch**:
  - Optimal $\mathbf{z}$ minimizes $\mathcal{L}_{IB} = I(\mathbf{Z}; \mathbf{Y}) - \beta I(\mathbf{Z}; \mathbf{X})$
  - Since $I(\mathbf{x}_d; \mathbf{y}) = 0$, domain features only contribute to complexity term
  - Therefore: $I(\mathbf{x}_d; \mathbf{z}) = 0$ in optimal representation
- **Real-World Implications**: Even partial compression leads to unsafe OOD detection
- **Fano's Inequality**: Small $I(\mathbf{x}_d; \mathbf{z})$ still causes unreliable detection

Empirical Demonstration

- **Domain Bench**: 11 single-domain datasets
  - Medical: Tissue (kidney cortex microscopy)
  - Agriculture: Plant (leaf disease classification)
  - Geology: Rock (mineral classification)
  - Waste Management: Garbage (material classification)
  - Fitness: Yoga (pose classification)
- **Experimental Setup**:
  - In-domain OOD: Adjacent OOD (25% classes held out)
  - Out-of-domain OOD: MNIST, SVHN, Textures, Places365, CIFAR-10/100
- **Key Finding**: All current SOTA methods perform worse on certain out-of-domain sets vs. their in-domain OOD performance
- **Evidence**: FPR@95 increases from $<$10% (in-domain) to $>$40% (out-of-domain)

Domain Filtering Methodology

- **Two-Stage Detection Framework**:
  - **Stage 1**: Domain filtering - Is sample in-domain?
  - **Stage 2**: OOD detection - Is in-domain sample in-distribution?
- **Domain Filter Implementation**:
  - Pretrained DinoV2 ViT-S/14 for domain-aware features
  - KNN distance at 99th percentile threshold ($K = 50$)
  - Preserves domain-specific information during training
- **Key Assumption**: No in-distribution samples are out-of-domain
  - Consistent with single-domain dataset definition
  - Allows clean separation of domain vs. class detection
- **Integration**: Compatible with any existing OOD detection method

Implementation and Validation

- **Experimental Results**:
  - **Domain filtering effectiveness**: FPR@95 reduced from >40% to <5%
  - **Consistent improvement**: Works across all 11 single-domain datasets
  - **Empirically validated**: Works with KNN, ReAct, and MDS methods
- **Performance Metrics**:
  - Out-of-domain OOD: Substantial FPR@95 reduction (8x improvement)
  - In-domain OOD: Minimal performance impact (maintains baseline)
  - AUROC improvements: 15-25 percentage points on out-of-domain
- **Validation Across Domains**:
  - Medical imaging, agriculture, geology, waste management
  - Confirms theoretical predictions empirically

Domain Collapse: Key Takeaways

- **Theoretical Breakthrough**:
  - First formal proof of domain feature collapse using information bottleneck theory
  - Explains why single-domain training creates dangerous OOD detection blind spots
  - Connects supervised learning objectives to systematic safety failures
- **Practical Solution**:
  - Domain filtering: Simple, effective, and method-agnostic approach
  - Two-stage framework preserves both domain and class detection capabilities
  - 8x improvement in out-of-domain OOD detection performance
- **Broader Impact**:
  - Domain Bench: New benchmark for single-domain OOD evaluation
  - Safety implications for medical imaging, autonomous systems
  - Guides deployment decisions in safety-critical applications

Information-Theoretic Hallucination Framework

- **Central Hypothesis**: Hallucinations arise from information degradation
  - $I(\mathbf{x}; \mathbf{y}) < \tau_{critical}$ between input queries and generated responses
  - Layer-wise information loss in transformer architectures
  - Critical threshold where reliable generation becomes impossible
- **Information Flow Analysis**:
  - Track $I(\mathbf{x}; \mathbf{z}_l)$ across transformer layers l
  - Identify bottleneck layers where information degrades
  - Attention mechanism role in preserving/destroying information
- **Theoretical Foundation**: Information Bottleneck Principle
- **Advantage**: No external knowledge bases required for detection

Contrastive MI Estimation Method

- **Novel Approach**: Contrastive learning for MI estimation
  - Learn projections $f_i : \mathbf{z}_{l_i} \to \mathbb{R}^d$ and $f_j : \mathbf{z}_{l_j} \to \mathbb{R}^d$
  - Maximize similarity for same QA pairs across layers
  - Minimize similarity for different QA pairs
- **Contrastive Objective**:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(f_i(\mathbf{z}_{l_i}), f_j(\mathbf{z}_{l_j}))/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(f_i(\mathbf{z}_{l_i}), f_j(\mathbf{z}_{l_j}^{(k)}))/\tau)} \tag{6}$$

- **MI Estimation**: $\hat{I}(\mathbf{z}_{l_i}; \mathbf{z}_{l_j})$ from learned representations
- **Advantages**: Task-specific, scalable, differentiable

System Architecture Details

- **Two-Stage Detection Framework**:
    - **Stage 1**: Primary LM generates responses + extracts layer embeddings
    - **Stage 2**: Secondary analysis model estimates MI between layers
    - Real-time detection during inference
- **Training Process**:
    - Use QA pairs with hallucination labels for contrastive learning
    - Learn to distinguish faithful vs. hallucinated responses
    - Optimize projection functions for MI estimation
- **Detection Mechanism**: $\hat{I}(\mathbf{x};\mathbf{y}) < \tau_{critical}$ triggers hallucination alert
- **Cross-Architecture Compatibility**: GPT, BERT, T5, Mamba

Experimental Design and Datasets

- **Validation Strategy**:
  - **Synthetic datasets**: Ground truth MI for method validation
  - **Real-world benchmarks**: HaluEval, TruthfulQA, FEVER, HalluLens
  - Cross-method consistency analysis (MINE, InfoNCE, kernel-based)
- **Evaluation Metrics**:
  - MI estimation accuracy vs. ground truth
  - Hallucination detection: AUROC, precision, recall, F1
  - Bias-variance decomposition of MI estimates
- **Model Coverage**: GPT-3.5/4, BERT, T5, LLaMA, Mamba architectures
- **Ablation Studies**: Projection architecture, temperature, negative sampling

Results and Performance Analysis

- **MI Estimation Performance**:
  - Superior accuracy on QA-specific tasks vs. general MI methods
  - Computational efficiency: 10-100x faster than MINE
  - Robust performance across different model scales and architectures
- **Hallucination Detection Results**:
  - Target: $>$85% AUROC on major benchmarks (HaluEval, TruthfulQA)
  - Real-time detection with $<$50ms latency overhead
  - Cross-architecture generalization without retraining
- **Information Flow Insights**:
  - Identify critical layers where hallucinations emerge
  - Quantify attention mechanism role in information preservation

Hallucination Detection: Key Takeaways

- **Theoretical Innovation**:
  - First information-theoretic framework for hallucination detection
  - Novel contrastive MI estimation method for transformer architectures
  - Principled connection between information flow and generation reliability
- **Practical Advantages**:
  - Real-time detection without external knowledge bases
  - Cross-architecture compatibility (GPT, BERT, T5, Mamba)
  - Scalable to large language models with minimal overhead
- **Research Impact**:
  - Opens new research directions in information-theoretic AI safety
  - Enables targeted interventions at critical transformer layers
  - Foundation for next-generation trustworthy AI systems

Research Timeline

- **Phase 1: Foundation & Method Development (Months 1-4)**:
  - Theoretical framework refinement and prototype analysis
  - Implementation optimization and baseline comparisons
  - Infrastructure setup for large-scale experiments
- **Phase 2: Large-Scale Validation (Months 5-8)**:
  - Foundation model integration (GPT, BERT, T5, Mamba)
  - MI-hallucination correlation studies and detection system development
  - Cross-architecture validation and performance benchmarking
- **Phase 3: Applications & Deployment (Months 9-12)**:
  - Domain-specific applications and intervention strategies
  - Comprehensive evaluation and open-source implementation
  - Research dissemination and community adoption

Expected Impact and Significance

- **Theoretical Contributions**:
  - First comprehensive information-theoretic framework for AI safety
  - Novel understanding of failure modes in OOD detection and hallucination
  - Principled connection between information theory and model reliability
- **Practical Applications**:
  - Real-time detection systems for safety-critical deployments
  - Cross-architecture compatibility enabling broad adoption
  - Open-source tools for community use and further research
- **Broader Impact**:
  - Enhanced trustworthiness of AI systems in healthcare, finance, autonomous vehicles
  - New research directions in information-theoretic AI safety
  - Foundation for next-generation reliable machine learning systems

Questions and Discussion

Thank you!

Questions?