[Proposal Title **OR** Dissertation Title as it appears on the Dissertation Certificate]

by

[professional name of author]

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Computing and Information Sciences

B. Thomas Golisano College of Computing and
Information Sciences

Rochester Institute of Technology
Rochester, New York
[Month and year of Dissertation Acceptance was signed]

[Proposal Title **OR** Dissertation Title as it appears on the Dissertation Certificate]

by

[professional name of author]

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

[Advisor's name] Date
Dissertation Advisor

[Committee member's name] Date
Dissertation Committee Member

[Committee member's name] Date
Dissertation Committee Member

[Committee member's name] Date
Dissertation Committee Member

[External Chair's name] Date
Dissertation Defense Chairperson

Certified by:

[Ph.D. Program Director name] Date
Ph.D. Program Director, Computing and Information Sciences

[Proposal Title **OR** Dissertation Title as it appears on the Dissertation Certificate]

by

[professional name of author]

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences Ph.D. Program in Computing and Information Sciences
in partial fulfillment of the requirements for the
Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

This is the abstract text with no more than 350 words.

Acknowledgments

This is the acknowledgements text

This is the dedication text

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 2

Background and Definitions

2.1 Out of Distribution Detection

Out-of-distribution (OOD) detection addresses a critical challenge in modern machine learning: the ability to recognize when a model is presented with inputs that fall outside the scope of what it has been trained to understand. While supervised models excel at making predictions within the distribution of their training data, they are notoriously prone to overconfident predictions when confronted with novel or unexpected inputs — sometimes with dangerous consequences, especially in high-stakes applications such as healthcare, autonomous driving, or security systems.

The task of out-of-distribution detection is to identify a semantic shift in the data (Yang et al., 2021). This is determining when no predicted label could match the true label $\boldsymbol{y} \notin \mathbb{Y}_{in}$, where $\mathbb{Y}_{in}$ represents the set of in-distribution training labels. In this case, we would consider the semantic space of the sample and the training distribution to be different, representing a semantic shift. We can express the probability that a sample is out-of-distribution via $P(\boldsymbol{y} \notin \mathbb{Y}_{in}|\boldsymbol{x})$.

Definition 2.1.1. Out-of-Distribution (OOD) Detection: Given an input $\mathbf{x}$ and a set of in-distribution labels $\mathbb{Y}_{\text{in}}$, OOD detection is the task of identifying whether the true label $\mathbf{y}$ belongs outside the in-distribution set, i.e., $\mathbf{y} \notin \mathbb{Y}_{\text{in}}$, or equivalently estimating $P(\mathbf{y} \notin \mathbb{Y}_{\text{in}} \mid \mathbf{x})$.

One baseline approach to estimate this probability is the *Maximum Softmax Probability (MSP)* method. Here, a trained classifier outputs softmax probabilities across known classes, and the maximum probability $\texttt{MSP}(\mathbf{x})$ is interpreted as a confidence score. A low maximum confidence suggests that the input might not belong to any known class, leading to the simple baseline:

Method 2.1.2 (Maximum Softmax Probability (MSP))**.** Given an input $\mathbf{x}$, the MSP method estimates the probability of being out-of-distribution as $1 - \texttt{MSP}(\mathbf{x})$, where $\texttt{MSP}(\mathbf{x}) = \max_{y \in \mathcal{Y}_{\text{in}}} P(y \mid \mathbf{x})$.

Furthermore, we are only concerned with labels that can be generated using only $\mathbf{x}$, via function f which depends solely on $\mathbf{x}$ and no other information. $f_{\mathbf{y}}$ may represent human labelers that generate $\mathbf{y}$. If we consider $\mathbb{Y}_{all}$ as the set of all possible labels that can be generated from $f_{\mathbf{y}}(\mathbf{x} \in \mathbb{X}_{all})$, a subset of $\mathbb{X}_{all}$ considered as $\mathbb{X}_{training}$ may not contain all labels in $\mathbb{Y}_{all}$. For real world datasets, it is possible that $\mathbb{Y}_{in} \subsetneq \mathbb{Y}_{all}$.

While related to other concepts like *anomaly detection*, OOD detection is distinct in key ways. Anomaly detection usually focuses on rare or abnormal data points *within* the same distribution (e.g., detecting fraudulent transactions), whereas OOD detection focuses on recognizing inputs from *entirely different* distributions or classes, outside the model's prior knowledge. Furthermore, OOD detection primarily targets *epistemic uncertainty* — the model's uncertainty due to limited knowledge — rather than *aleatoric uncertainty*, which arises from inherent noise or variability in the data.

Practically, OOD detection methods fall into two broad categories:

- **Training-agnostic approaches**, like MSP or entropy-based scoring, which apply directly to existing classifiers without altering their training process.
- **Training-aware approaches**, which adapt model architectures, loss functions, or data augmentation strategies specifically to improve OOD recognition capabilities.

Evaluation typically involves exposing the model to benchmark OOD datasets designed to test its ability to reject or abstain from confident predictions on unfamiliar samples, while maintaining strong performance on in-distribution data. Additional information regarding methods and benchmarks will be provided in a later section.

2.2 Anomaly Detection

Although anomaly detection and out-of-distribution (OOD) detection are sometimes used interchangeably in the literature, they address fundamentally different problems, with distinct assumptions, goals, and evaluation settings.

2.2.1 Definiton and Scope

Anomaly detection focuses on identifying individual data points that deviate significantly from the expected patterns within a single dataset or distribution. These anomalies, or outliers, are typically rare and often correspond to noise, rare events, or fraudulent activities within the same domain as the training data. For example, anomaly detection might flag fraudulent transactions in a credit card dataset, where the system has only ever seen transaction records from that specific financial context.

OOD detection, on the other hand, aims to detect data that comes from a different distribution altogether — one that was not seen during training. It concerns the model's ability to recognize when a test input belongs to a class, domain, or environment that falls outside the model's learned distribution. For example, an image classifier trained only on animals should ideally flag an image of a car as OOD, even though the car is not necessarily anomalous within its own context.

2.2.2 Training Assumptions

Anomaly detection methods typically assume access only to normal (non-anomalous) data during training and must learn to recognize deviations without ever seeing examples of the anomalies. This is often referred to as a one-class learning problem.

In contrast, OOD detection typically operates in a supervised learning context where the model has been trained on multiple in-distribution classes, and the challenge is to detect test inputs that fall outside this known set. Here, the focus is on recognizing the model's epistemic uncertainty — i.e., knowing what the model doesn't know.

2.2.3 Evaluation Settings

Anomaly detection is typically evaluated using synthetic or labeled anomaly datasets, where the goal is to identify rare but known outlier patterns within the same dataset.

OOD detection is evaluated by exposing the model to entirely new datasets or domains and measuring its ability to correctly reject or abstain from making confident predictions on these unfamiliar inputs. This setting often requires curated OOD benchmark datasets distinct from the in-distribution training data.

2.3 Information Theory

Information theory provides a mathematical framework for quantifying uncertainty, information content, and the relationships between random variables. Originally developed by Shannon (1948), it has become foundational for many areas of machine learning, including uncertainty quantification, representation learning, and out-of-distribution (OOD) detection. The following are various definitions that are critical to understanding information theory.

Definition 2.3.1. Entropy: The entropy of a discrete random variable $\mathbf{x}$ with distribution $p(\boldsymbol{x})$ is defined as the expected amount of uncertainty or information contained in $\mathbf{x}$, denoted by $H(\mathbf{x})$.

Definition 2.3.2. Conditional Entropy: The conditional entropy $H(\mathbf{y} \mid \mathbf{x})$ measures the remaining uncertainty in a random variable $\mathbf{y}$ given knowledge of another random variable $\mathbf{x}$.

Definition 2.3.3. Mutual Information: The mutual information between random variables $\mathbf{x}$ and $\mathbf{y}$, denoted $I(\mathbf{x}; \mathbf{y})$, quantifies the amount of information shared between $\mathbf{x}$ and $\mathbf{y}$, or equivalently, the reduction in uncertainty about $\mathbf{x}$ given knowledge of $\mathbf{y}$.

Definition 2.3.4. Kullback-Leibler (KL) Divergence: The KL divergence between two distributions P and Q, denoted $D_{\mathrm{KL}}(P \parallel Q)$, measures how much the distribution P diverges from the reference distribution Q.

Definition 2.3.5. Chain Rule for Entropy: The joint entropy of two random variables $\mathbf{x}$ and $\mathbf{y}$ satisfies the chain rule:

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y} \mid \mathbf{x}),$$

expressing that the total uncertainty can be decomposed into the uncertainty of $\mathbf{x}$ plus the uncertainty of $\mathbf{y}$ given $\mathbf{x}$.

Definition 2.3.6. Mutual Information as KL Divergence: The mutual information between $\mathbf{x}$ and $\mathbf{y}$ can be equivalently defined as the KL divergence between the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ and the product of the marginals $p(\boldsymbol{x})p(\boldsymbol{y})$:

$$I(\mathbf{x}; \mathbf{y}) = D_{\mathrm{KL}}(p(\boldsymbol{x}, \boldsymbol{y}) \parallel p(\boldsymbol{x})p(\boldsymbol{y})).$$

Definition 2.3.7. Non-negativity of Mutual Information: Mutual information is always non-negative, that is, $I(\mathbf{x}; \mathbf{y}) \geq 0$, with equality if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent.

Definition 2.3.8. Chain Rule for Mutual Information: The mutual information between two random variables $\mathbf{x}$ and $\mathbf{y}$ can be decomposed using the chain rule as follows:

$$I(\mathbf{x}; \mathbf{y}) = I(\mathbf{x}; \mathbf{z}) + I(\mathbf{z}; \mathbf{y} \mid \mathbf{x}),$$

where $\mathbf{z}$ is a third variable, and $I(\mathbf{x}; \mathbf{y})$ represents the total mutual information between $\mathbf{x}$ and $\mathbf{y}$. This decomposition expresses the amount of shared information in terms of intermediate variables that mediate the relationship.

2.3.1 Entropy and Mutual Information

Entropy and mutual information are central concepts in information theory, and they play a key role in understanding uncertainty and information flow in machine learning models.

Entropy, as defined in Definition 2.3.1, measures the average uncertainty in a random variable $\mathbf{x}$. It quantifies how much unpredictability exists in the outcomes of $\mathbf{x}$. In machine learning, entropy is often used to assess the uncertainty in a model's predictions or to regularize a model by minimizing its uncertainty.

Mutual information, as defined in Definition 2.3.3, measures the amount of information shared between two random variables, $\mathbf{x}$ and $\mathbf{y}$. Specifically, it quantifies how much knowing one variable reduces uncertainty about the other. In machine learning, mutual information is used to gauge the relevance of features to the target variable, aiding in feature selection and representation learning. By maximizing mutual information between representations and labels, we can improve the expressiveness and usefulness of learned features.

2.4 Information Bottleneck and Minimal Sufficient Statistic

In this section, we introduce two important concepts in information theory that are relevant for understanding how to efficiently represent information while preserving relevant information: the **Information Bottleneck** and the **Minimal Sufficient Statistic**.

2.4.1 Minimal Sufficient Statistic

A **Minimal Sufficient Statistic** is a concept from statistics that defines a statistic that captures all the information about a parameter of interest in a dataset while being the most compact representation. In the context of information theory, it is the statistic that minimizes the loss of information and is often used in maximum likelihood estimation (MLE).

Definition 2.4.1. Minimal Sufficient Statistic: A statistic $T(\mathbf{x})$ is called a *minimal sufficient statistic* for a random variable $\mathbf{x}$ with respect to a parameter $\mathbf{y}$ if it satisfies the following two conditions:

- **Sufficiency**: $T(\mathbf{x})$ is sufficient for $\mathbf{y}$, meaning that it captures all the information about $\mathbf{y}$ contained in $\mathbf{x}$, i.e.,
$$I(\mathbf{x}; \mathbf{y} \mid T(\mathbf{x})) = 0.$$
- **Minimality**: There exists no other statistic s such that s is sufficient for $\mathbf{y}$ and s is a function of $T(\mathbf{x})$, i.e., there exists a function $f(s)$ such that $s = f(T(\mathbf{x}))$.

Formally, the minimal sufficient statistic $T(\mathbf{x})$ is the statistic that retains all the relevant information about $\mathbf{y}$ and satisfies the condition that for any other sufficient statistic s, there exists an $f(s)$ such that $s = f(T(\mathbf{x}))$, making $T(\mathbf{x})$ the minimal sufficient statistic.

The **Minimal Sufficient Statistic** is crucial in statistical inference because it ensures that no further information about the parameter θ can be extracted from the data, given the statistic $T(\mathbf{x})$. It is often used in the context of parameter estimation where the goal is to reduce the data to the smallest possible set that still retains all necessary information for accurate inference.

2.4.2 Information Bottleneck

The Information Bottleneck (IB) principle, introduced by Tishby et al. (2000), provides a framework for learning representations of data that capture the most relevant information while discarding unnecessary details. The core idea of the Information Bottleneck is to find a representation of a random variable $\mathbf{x}$ that preserves the information about a target variable $\mathbf{y}$, but minimizes the amount of information retained about irrelevant variables. Effectively, we expect a model's representation $\mathbf{z}$ to compress towards the minimal sufficient statistic, as per definition 2.4.1, under information bottleneck optimization.

Definition 2.4.2. Information Bottleneck: Given two random variables $\mathbf{x}$ and $\mathbf{y}$, the goal of the Information Bottleneck is to find a representation $\mathbf{z}$ of $\mathbf{x}$ such that the mutual information $I(\mathbf{z};\mathbf{y})$ is maximized while the mutual information $I(\mathbf{z};\mathbf{x})$ is minimized. Formally, the Information Bottleneck objective is:

$$\mathcal{L}_{\text{IB}} = I(\mathbf{z};\mathbf{y}) - \beta I(\mathbf{z};\mathbf{x}),$$

where β controls the trade-off between retaining information about $\mathbf{y}$ and compressing information about $\mathbf{x}$.

The **Information Bottleneck** principle has been used extensively in machine learning for unsupervised learning, feature selection, and representation learning. It provides a formalization of the idea that an optimal representation of data should balance between compressing the input and retaining sufficient information to predict the output.

2.5 Dataset Domain

In supervised learning, we can observe some datasets from a specific *domain*, which can be understood informally as the environment, context, or generating conditions under which the data was collected. For example, handwritten digit images from the MNIST dataset belong to a domain defined by grayscale digit images, whereas natural scene photographs from ImageNet belong to a much broader domain. Understanding domains is critical for tasks such as domain adaptation, transfer learning, and out-of-distribution (OOD) detection.

Formally, we define the **domain** of a dataset using a domain labeling function $f_{\mathbf{d}}$, which assigns a domain label $\mathbf{d}$ to each input sample $\mathbf{x}$:

$$\mathbf{d} = f_{\mathbf{d}}(\mathbf{x}).$$

For the purposes of this dissertation, we are particularly interested in certain cases where the training data belongs to a single domain $\mathbf{d}_1$, such that:

$$\forall \mathbf{x} \in \{f_{\mathbf{y}}(\mathbf{x}) \in \mathbb{Y}_{in}\}, \quad f_{\mathbf{d}}(\mathbf{x}) = \mathbf{d}_1,$$

where $f_{\mathbf{y}}$ is the labeling function producing the class label $\mathbf{y}$ and $\mathbb{Y}_{in}$ is the set of in-distribution labels. In such a setup, any data sample for which:

$$f_{\mathbf{d}}(\mathbf{x}) \neq \mathbf{d}_1$$

can be assumed to lie outside the in-distribution label set, i.e., $f_{\mathbf{y}}(\mathbf{x}) \notin \mathbb{Y}_{in}$.

Given that all elements of $\mathbb{Y}_{in}$ come from domain $\mathbf{d}_1$, any subset of features of $\mathbf{x}$ that is sufficient for determining the class label $f_{\mathbf{y}}$ is also sufficient for determining the domain label $f_{\mathbf{d}}$. We define the **domain features** $\mathbf{x_d}$ as the subset of features used to infer the domain, and the **class features** $\mathbf{x_y}$ as those used to infer the class label. For this work, we define class and domain features as separate, such that:

$$I(\mathbf{x_d} : \mathbf{x_y}) = 0,$$

meaning that domain features and class features are independent in the context of the training set. However, this independence does not necessarily hold in the full input space $\mathbb{X}_{all}$, where domain features can provide valuable additional information to determine the class. Importantly, the minimal set of features required for each task aligns with the notion of minimal sufficient statistics as defined earlier (Definition 2.4.1).

It is also important to recognize that domains are structured hierarchically. For example, the domain of *cats* is a subdomain of *mammals*, which is itself a subdomain of *animals*. At the top of the hierarchy, one could define an all-encompassing domain that includes everything, but in this case, the set of non-trivial domain features would be empty, i.e., $\{\mathbf{x_d}\} = \emptyset$.

Finally, some datasets might be labeled under a single domain $\mathbf{d}_1$ but effectively behave as multi-domain datasets because they have such a broad variety of classes. For instance, if we treat ImageNet as a single domain, the set of pure domain features (those not overlapping with class features) may approach zero, $|\{\mathbf{x_d}\}| \approx 0$. This indicates that the diversity of classes effectively spans multiple domains, and it is more appropriate to treat such datasets as multi-domain.

Definition 2.5.1. Single-Domain Dataset: A dataset is called a *single-domain dataset* if there exists a nontrivial set of domain features $\mathbf{x_d}$ that are:

- independent from the class features $\mathbf{x_y}$, i.e.,

$$I(\mathbf{x_d} : \mathbf{x_y}) = 0,$$

- non-vanishing in size, meaning

$$|\{\mathbf{x_d}\}| \gg 0.$$

This definition distinguishes single-domain datasets from multi-domain datasets, where the diversity of classes effectively collapses the set of independent domain features to approximately zero (i.e.,

$|\{\mathbf{x_d}\}| \approx 0$). In a single-domain dataset, domain features capture global properties shared across all samples (e.g., imaging modality, capture conditions), while class features capture the discriminative properties used for labeling.

The nature of domains their interaction with information theory is a topic of study in this dissertation.

2.5.1 Domain Features and Domain Feature Collapse

Building upon our understanding of dataset domains, we can further decompose the input features $\mathbf{x}$ into domain-specific and class-specific components. This decomposition is crucial for understanding certain failure modes in out-of-distribution detection, particularly in single-domain settings.

Definition 2.5.2. Domain Features: Given a dataset with domain $\mathbf{d}$ determined by the labeling function $f_{\mathbf{d}}(\mathbf{x})$, we define the domain features $\mathbf{x_d}$ as the minimal subset of features of $\mathbf{x}$ that is sufficient for $f_{\mathbf{d}}$, under the constraint that $\mathbf{x_d}$ is independent of the minimal sufficient class features $\mathbf{x_y}$, i.e., $I(\mathbf{x_d} : \mathbf{x_y}) = 0$.

For this work, we define domain features $\mathbf{x_d}$ such that they do not overlap with class features $\mathbf{x_y}$, implying $I(\mathbf{x_d} : \mathbf{x_y}) = 0$. The independence of domain and class features only applies to the training set, as domain features would provide useful information in the context of $\mathbb{X}_{all}$. Note that this also implies that $\neg(\forall \mathbf{x}, f_{\mathbf{y}}(\mathbf{x_y}) = f_{\mathbf{y}}(\mathbf{x}))$ and $\forall \mathbf{x}, f_{\mathbf{y}}(\mathbf{x_y}, \mathbf{x_d}) = f_{\mathbf{y}}(\mathbf{x})$. For both domain and class features, we refer to the minimal set of features, as per the minimal sufficient statistic definition.

Examples of domain features include:

- In medical imaging: imaging modality (X-ray vs. MRI vs. CT scan)
- In satellite imagery: sensor type, resolution, or atmospheric conditions
- In natural images: lighting conditions, camera characteristics, or background context

Definition 2.5.3. Domain Feature Collapse: A phenomenon where supervised learning models trained on single-domain datasets learn representations that discard domain-specific features, retaining only class-specific features. Formally, this occurs when $I(\mathbf{x_d}; \mathbf{z}) = 0$ for the learned representation $\mathbf{z}$, despite domain features being present in the input $\mathbf{x}$.

Domain feature collapse is particularly problematic for out-of-distribution detection because it means the model cannot distinguish between in-domain and out-of-domain samples that share similar class features. This leads to a critical safety gap where out-of-domain inputs containing in-distribution class features may be misclassified with high confidence.

2.6 Unlabeled OOD Detection

In most out-of-distribution (OOD) detection tasks, models are trained using labeled in-distribution (ID) data, where each training sample $\mathbf{x}$ is paired with its ground-truth label $\mathbf{y}$. These labels are typically used to train a supervised classifier whose outputs are then repurposed for OOD detection, such as through softmax confidence scores or logit-based methods.

However, not all OOD detection methods require labeled data. We define **unlabeled OOD detection** as any OOD detection approach where the model is trained solely on the ID data $\mathbf{x}$ without access to or use of the corresponding labels $\mathbf{y}$.

Formally, let the ID dataset be defined as:

$$\mathbb{D}_{in} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}.$$

An OOD detection method is considered *unlabeled* if its training process uses only the inputs $\mathbf{x}_i$ and does not depend on the labels $\mathbf{y}_i$. That is, the learned OOD detection function f_{OOD} is trained using:

$$f_{OOD} \leftarrow \text{Train}(\{\mathbf{x}_i\}_{i=1}^{N}),$$

and no supervision signal from $\{\mathbf{y}_i\}$ is involved.

Unlabeled OOD detection approaches often rely on unsupervised learning objectives, such as density estimation, reconstruction error, or self-supervised representations. We can also consider using a pretrained model (without fine tuning) as a form of unlabeled OOD detection, as one does not explicitly train on the in distribution labels. These methods are attractive in settings where label acquisition is expensive or infeasible, or where one desires OOD detection capabilities decoupled from any specific classification task.

Importantly, while unlabeled methods do not use labels during training, they still aim to solve the same core problem as labeled OOD detection: estimating the probability that a given test input $\mathbf{x}$ comes from a distribution different from the training data. Formally, both types of methods

estimate:

$$P(f_{\mathbf{y}}(\mathbf{x}) \notin \mathbb{Y}_{in}),$$

where $\mathbb{Y}_{in}$ is the set of in distribution class labels.

The nature of unlabeled OOD detection methods and their interaction with information theory is a topic of study in this dissertation.

2.7 Large Language Models

Definition (Large Language Model). A *Large Language Model* (LLM) is a parameterized probabilistic model

$$f_{\boldsymbol{\theta}} : \mathcal{X}^* \to [0, 1],$$

where $\mathcal{X}$ denotes a finite vocabulary and $\mathcal{X}^*$ the set of all finite-length sequences over $\mathcal{X}$. The model defines a distribution over sequences $\boldsymbol{x} = (x_1, \ldots, x_T)$ via an autoregressive factorization:

$$\Pr_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{t=1}^{T} \Pr_{\boldsymbol{\theta}}(x_t \mid x_{<t}),$$

where $x_t \in \mathcal{X}$, and $x_{<t} = (x_1, \ldots, x_{t-1})$. The conditional probabilities are parameterized by a deep neural architecture—typically a Transformer—with $\boldsymbol{\theta} \in \mathbb{R}^d$ and d in the order of billions.

LLMs are trained on large-scale text corpora $\mathcal{D} \subset \mathcal{X}^*$ by minimizing the empirical cross-entropy loss:

$$\min_{\boldsymbol{\theta}} \; \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[-\sum_{t=1}^{|\boldsymbol{x}|} \log \Pr_{\boldsymbol{\theta}}(x_t \mid x_{<t}) \right].$$

The qualifier "large" reflects both the scale of the model (e.g., $|\boldsymbol{\theta}| \geq 10^9$) and the training dataset (typically hundreds of billions of tokens). LLMs exhibit emergent behavior, in-context learning, and rich internal representations, motivating theoretical investigations into generalization, scaling laws, and the geometry of learned representations.

Definition: Hallucinations in Language Models

Definition (Hallucination). Let $\boldsymbol{x} \in \mathcal{X}^*$ be an input prompt and $\boldsymbol{y} \in \mathcal{X}^*$ a model-generated continuation sampled from $\Pr_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$. A *hallucination* occurs when the generated output $\boldsymbol{y}$ contains

content that is not grounded in verifiable facts, contextually entailed information, or externally available sources, relative to a defined reference world model or oracle $\mathcal{W}$.

Formally, $\boldsymbol{y}$ is said to hallucinate with respect to $\mathcal{W}$ if there exists a span $\boldsymbol{y}' \subseteq \boldsymbol{y}$ such that $\boldsymbol{y}'$ contradicts $\mathcal{W}$ or introduces unverifiable or fabricated content under the semantics of the task.

We further separate hallucinations in Extrinsic and Intrinsic. For the purpose of this work, we are primarily interested in Extrinsic Hallucinations.

Extrinsic Hallucination

An *extrinsic hallucination* refers to content in $\boldsymbol{y}$ that contradicts known facts or available reference data, i.e., $\boldsymbol{y}$ is inconsistent with $\mathcal{W}$. In this case, $\mathcal{W}$ corresponds to an external corpus or knowledge base. This type of hallucination typically occurs when an LLM must rely on its internal knowledge to complete a task. This could be something like answering a simple question or citing the correct sources when writing an article. Note that hallucinations are relative to $\mathcal{W}$, which may not align with current information, eg. asking what is the latest version of PyTorch will likely return an incorrect, but not hallucinated, answer.

$$\text{Extrinsic: } \exists \boldsymbol{y}' \subseteq \boldsymbol{y} \text{ such that } \boldsymbol{y}' \notin \mathcal{W} \quad \text{and} \quad \boldsymbol{y}' \text{ is asserted as fact.}$$

Example: A model generating "The capital of Canada is Toronto" when $\mathcal{W}$ (a knowledge base) correctly states that the capital is Ottawa.

Intrinsic Hallucination

An *intrinsic hallucination* arises when the generated content $\boldsymbol{y}$ violates internal logical consistency, coherence, or task-specific constraints—even without external knowledge. In this case, hallucinations are identifiable by contradiction, incoherence, or inconsistency relative to the prompt $\boldsymbol{x}$ or previously generated tokens.

$$\text{Intrinsic: } \exists (\boldsymbol{y}', \boldsymbol{y}'') \subseteq \boldsymbol{y} \text{ such that } \boldsymbol{y}' \nRightarrow \boldsymbol{y}'' \quad \text{under the task semantics.}$$

Example: A dialogue model stating "I was born in 1990" followed by "I am 20 years old" within

the same response, assuming current time is known or implied.

Chapter 3

Literature Review

3.1 Information Theory in Machine Learning

Information theory, originating from Shannon's foundational work (Shannon, 1948), provides a mathematical framework for quantifying uncertainty, dependence, and information flow. Its integration into machine learning has grown substantially in recent decades, offering both theoretical insight and practical methodologies for learning representations, optimizing communication-efficient models, and analyzing generalization. At the core of this intersection are measures such as entropy, mutual information (MI), and Kullback–Leibler (KL) divergence, which provide formal tools for characterizing uncertainty, dependencies, and divergences between distributions. These measures have been employed to interpret and regularize learning processes, as well as to derive principled algorithms from first principles.

One key area of application is in *representation learning*, where mutual information serves as both an objective and an interpretive lens. Methods such as InfoMax (Linsker, 1988) and its modern adaptations—including Deep InfoMax (Hjelm et al., 2019) and contrastive predictive coding (van den Oord et al., 2018)—maximize MI between inputs and learned representations to preserve task-relevant information while discarding noise. Conversely, the *information bottleneck* (IB) principle (Tishby et al., 2000) formalizes representation learning as an optimization trade-off between compression of input data and preservation of predictive information about the target variable. This has been extended to deep networks (Alemi et al., 2017), offering both training objectives and a theoretical framework for understanding the emergence of compressed representations.

Beyond representation learning, information-theoretic quantities play a central role in *regularization* and *generalization analysis*. PAC-Bayesian bounds (McAllester, 1999) and mutual-information-based generalization bounds (Xu & Raginsky, 2017) provide finite-sample guarantees on model performance, connecting overfitting behavior to the amount of information a learned model retains about its training set. These perspectives have informed methods such as noise injection, dropout, and stochastic weight averaging, which can be interpreted as constraining information flow between data and parameters.

In probabilistic modeling and generative learning, information theory provides the backbone for *variational inference* (Jordan et al., 1999; Kingma & Welling, 2014), where KL divergence measures guide the approximation of intractable posterior distributions. Variational autoencoders (VAEs) explicitly incorporate KL regularization to enforce compact, disentangled latent spaces. Similarly, generative adversarial networks (GANs) have been extended with MI-based terms, as in InfoGAN (Chen et al., 2016), to encourage interpretable latent factors.

Information-theoretic tools also influence *feature selection* and *causal inference*. Mutual information has been a longstanding criterion for selecting features with maximal relevance and minimal redundancy Peng et al. (2005), while recent advances use conditional MI to uncover causal structures in high-dimensional data Runge et al. (2019). Additionally, information flow measures—such as directed information and transfer entropy—are increasingly used to study temporal dependencies in time-series learning.

While the integration of information theory into machine learning is rich and diverse, challenges remain. Mutual information estimation in high dimensions is notoriously difficult, and the reliability of neural estimators Belghazi et al. (2018) has been questioned. Moreover, the precise role of compression in deep learning—whether it is a cause of generalization or a byproduct of optimization—remains debated Saxe et al. (2019). Nevertheless, ongoing work continues to refine both the theoretical foundations and practical estimators, reinforcing information theory as a powerful lens for designing, analyzing, and understanding modern machine learning systems.

3.2 Representation Learning

Representation learning aims to automatically discover useful features from raw data, learning transformations that map high-dimensional inputs to lower-dimensional representations capturing essential structure for downstream tasks. The theoretical foundations are deeply connected to

information theory through the information bottleneck framework (Tishby et al., 2000), which formalizes the trade-off between compression and prediction.

3.2.1 Unsupervised Representation Learning

Classical unsupervised methods include *Principal Component Analysis* (PCA) (Pearson, 1901), which learns linear projections maximizing variance, and *Independent Component Analysis* (ICA) (Hyvärinen & Oja, 2000), which seeks statistically independent components. These methods provide foundations for understanding how to decompose data into meaningful factors.

Deep unsupervised approaches revolutionized the field through *autoencoders* (Hinton & Salakhutdinov, 2006), which learn compact representations via reconstruction objectives. *Variational Autoencoders* (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014) introduced probabilistic frameworks combining neural networks with variational inference, using KL regularization to enforce structured latent spaces.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) learn representations through adversarial training between generator and discriminator networks. While primarily designed for generation, GANs implicitly learn rich data representations in their latent spaces, with variants like InfoGAN (Chen et al., 2016) explicitly encouraging disentangled factors through mutual information maximization. Similarly, *diffusion models* (Ho et al., 2020; Song et al., 2021) learn representations by modeling the gradual denoising process, capturing hierarchical data structure through their iterative generation procedure.

3.2.2 Self-Supervised Learning

Self-supervised learning leverages inherent data structure to create supervisory signals without manual labels. In computer vision, *contrastive learning* (Chen et al., 2020b; He et al., 2020) maximizes agreement between augmented views of the same image while minimizing agreement between different images. Frameworks like SimCLR (Chen et al., 2020b) and MoCo (He et al., 2020) have achieved remarkable success in learning transferable visual representations.

In natural language processing, self-supervised learning has been transformative through masked language modeling and autoregressive prediction. Early methods like *word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014) learn static word embeddings by predicting words

from contexts. Modern transformer-based models like *BERT* (Devlin et al., 2018) use bidirectional masked language modeling, randomly masking tokens and learning to predict them from surrounding context. Autoregressive models like *GPT* (Radford et al., 2018) and its successors learn representations by predicting the next token in a sequence, while encoder-decoder models like *T5* (Raffel et al., 2020) frame all tasks as text-to-text generation problems.

The success of self-supervised learning can be understood through mutual information maximization—contrastive methods implicitly maximize MI between representations of augmented views (Hjelm et al., 2019; van den Oord et al., 2018), while masked language models maximize MI between representations and missing tokens.

3.3 Out of Distribution Detection

Out-of-distribution (OOD) detection has emerged as a critical challenge in deploying machine learning systems safely in real-world environments. The field encompasses diverse methodologies ranging from simple confidence-based approaches to sophisticated training-aware techniques that modify model architectures and objectives specifically for OOD detection.

3.3.1 Classical and Training-Agnostic Approaches

Early OOD detection methods focused on post-hoc analysis of trained models without modifying the training process. The *Maximum Softmax Probability* (MSP) baseline (Hendrycks & Gimpel, 2016) uses the maximum predicted class probability as a confidence score, with low confidence indicating potential OOD samples. *ODIN* (Liang et al., 2017) enhances this approach through temperature scaling and input preprocessing to amplify the difference between in-distribution and OOD predictions.

Energy-based methods provide an alternative perspective, with *Energy Score* (Liu et al., 2020) interpreting the negative log-sum-exp of logits as an energy function, where OOD samples correspond to higher energy states. Distance-based approaches like *Mahalanobis distance* (Lee et al., 2018) measure similarity to class-conditional Gaussian distributions in feature space, while *KNN-based methods* (Sun et al., 2022) leverage nearest neighbor distances in learned representations.

3.3.2 Self-Supervised and Unsupervised OOD Detection

A significant advancement in OOD detection has come through leveraging self-supervised learning objectives that do not require explicit OOD data during training. *Contrastive learning* approaches have proven particularly effective, with methods like *CSI* (Tack et al., 2020) using contrastive learning on distributionally shifted instances to learn representations that naturally separate in-distribution from OOD data.

CADet (Guille-Escuret et al., 2024) represents a fully self-supervised approach that uses contrastive learning without requiring any labeled data, demonstrating that effective OOD detection can emerge from representation learning objectives alone. Similarly, *SSD* (Sehwag et al., 2021) provides a unified framework for self-supervised outlier detection by combining multiple self-supervised tasks.

Generative model approaches offer another unsupervised pathway, with *reconstruction-based methods* (Zhou, 2022) using autoencoders and VAEs to detect OOD samples through reconstruction error. Recent work has explored *diffusion models* (Liu et al., 2023) for unsupervised OOD detection, leveraging the inpainting capabilities of diffusion processes to identify distributional shifts.

The theoretical foundations of unsupervised OOD detection remain an active area of research, with recent work (Du et al., 2024a) investigating how unlabeled data provably helps OOD detection and exploring the fundamental limitations of label-agnostic approaches (Yang et al., 2025).

3.3.3 Benchmarking and Evaluation

The evaluation of OOD detection methods relies on carefully curated benchmark datasets that simulate realistic distribution shifts. The standard evaluation protocol involves training models on in-distribution data and testing their ability to distinguish between in-distribution test samples and out-of-distribution samples from different datasets or domains.

Computer vision benchmarks typically use datasets like CIFAR-10/100 (Krizhevsky et al., 2009) and ImageNet (Russakovsky et al., 2015) as in-distribution data, with various OOD datasets including SVHN, Textures, Places365, and LSUN (Yang et al., 2022). The *OpenOOD benchmark* (Yang et al., 2022; Zhang et al., 2023b) provides a comprehensive evaluation framework with standardized protocols, covering both near-OOD (semantically similar) and far-OOD (semantically distant) scenarios.

For *natural language processing*, benchmarks often use datasets like CLINC150 for intent classi-

fication, with OOD samples from different domains or artificially generated out-of-scope queries. Recent work has also explored OOD detection in large language models using datasets that test factual knowledge boundaries and domain-specific expertise.

Evaluation metrics typically include the Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUPR), and False Positive Rate at 95% True Positive Rate (FPR95). These metrics capture different aspects of detection performance, with AUROC providing overall discriminative ability and FPR95 focusing on practical deployment scenarios where high recall is essential.

The field has also developed specialized benchmarks for specific applications, including medical imaging (Zhang et al., 2021), autonomous driving (Ramanagopal et al., 2018), and earth observation (Ekim et al., 2024), reflecting the critical importance of reliable OOD detection in safety-critical domains.

3.3.4 Single-Domain OOD Detection

While most theoretical work in OOD detection focuses on multi-domain settings, applied research often occurs in single-domain contexts where models are deployed in narrowly scoped environments with highly consistent data characteristics. Single-domain OOD detection presents unique challenges that are not adequately captured by traditional multi-domain benchmarks.

Medical imaging represents a prominent application area for single-domain OOD detection. Zhang et al. (2021) investigate OOD detection in medical imaging contexts, while Cao et al. (2020) provide a comprehensive benchmark for medical out-of-distribution detection. Narayanaswamy et al. (2023) explore the specification of inliers and outliers for improved medical OOD detection, highlighting the domain-specific considerations required in healthcare applications.

Satellite imagery and remote sensing constitute another important single-domain application area. Ekim et al. (2024) examine distribution shifts at scale in earth observation, demonstrating the challenges of OOD detection when dealing with satellite data that shares consistent imaging characteristics but may contain novel land use patterns or environmental conditions.

Agricultural and biological applications also benefit from single-domain OOD detection methods. Saadati et al. (2024) develop OOD detection algorithms specifically for robust insect classification in agricultural settings, where the domain characteristics (imaging conditions, background, scale) remain consistent while the biological diversity creates classification challenges.

Industrial applications represent another critical area where single-domain OOD detection is essential. Kafunah et al. (2023) investigate out-of-distribution data generation for fault detection and diagnosis in industrial systems, while Kim et al. (2021) focus on wafer defect pattern classification with OOD detection in semiconductor manufacturing.

The common thread across these applications is that the in-distribution data comes from a narrow, well-defined domain with consistent characteristics (imaging modality, sensor type, environmental conditions), but the models must still detect when inputs fall outside the trained class distribution. This setting creates unique challenges that differ fundamentally from the multi-domain scenarios typically studied in general OOD detection research.

Recent work has begun to recognize the limitations of existing approaches in single-domain settings. The concept of *adjacent OOD detection* (Yang et al., 2025) specifically addresses scenarios where OOD samples come from the same domain as the training data but represent different classes, highlighting a critical gap in traditional evaluation protocols that focus primarily on far-OOD detection across different domains.

3.3.5 Domain Adaptation and Transfer Learning

The challenges of single-domain OOD detection are closely related to work in domain adaptation and transfer learning, though the objectives differ. Domain adaptation typically seeks to transfer knowledge from a source domain to a target domain, while single-domain OOD detection aims to identify when inputs fall outside the source domain entirely.

Katz-Samuels et al. (2022) investigate training OOD detectors in their natural habitats, emphasizing the importance of considering the deployment environment during model development. This work highlights the gap between laboratory evaluation on diverse benchmarks and real-world deployment in specialized domains.

The relationship between domain characteristics and OOD detection performance remains an active area of research, with implications for both theoretical understanding and practical deployment of OOD detection systems in specialized applications.

3.3.6 Multi-Stage and Ensemble Approaches

Recent advances in OOD detection have explored multi-stage and ensemble approaches that combine different detection mechanisms to improve overall performance. These methods recognize that no single OOD detection approach is optimal across all scenarios and seek to leverage the complementary strengths of different techniques.

Ensemble methods have been extensively studied in uncertainty estimation and OOD detection. Lakshminarayanan et al. (2017) demonstrate that deep ensembles provide simple and scalable predictive uncertainty estimation, while Xu et al. (2024) introduce deep multi-comprehension ensembles specifically for OOD detection. These approaches typically combine predictions from multiple models or detection mechanisms to improve robustness.

Two-stage detection frameworks represent a specific class of multi-stage approaches where different stages focus on different aspects of the detection problem. The first stage typically performs coarse-grained filtering or domain-level detection, while the second stage applies more sophisticated class-level OOD detection. This hierarchical approach allows for more targeted and efficient detection strategies.

The effectiveness of multi-stage approaches often depends on the careful design of each stage and the integration mechanism between stages. Domain filtering, as introduced in the context of single-domain OOD detection, represents a specific instantiation of this paradigm where the first stage focuses explicitly on domain-level detection before applying traditional OOD detection methods.

Hybrid supervised-unsupervised approaches combine the benefits of labeled and unlabeled learning objectives. While purely unsupervised methods may suffer from label blindness, and purely supervised methods may suffer from domain feature collapse, hybrid approaches attempt to capture both class-relevant and domain-relevant features in the learned representations.

3.4 Hallucination Detection

Hallucination detection in large language models has emerged as a critical challenge for deploying these systems in real-world applications where factual accuracy and reliability are paramount. Hallucinations—instances where models generate plausible-sounding but factually incorrect or unverifiable content—pose significant risks in domains such as healthcare, legal advice, and scientific research.

3.4.1 Taxonomy of Hallucination Detection Approaches

Hallucination detection methods can be broadly categorized into several approaches based on their underlying mechanisms and data requirements. *Confidence-based methods* leverage the model's own uncertainty estimates, using measures such as token-level probabilities, entropy, or attention patterns to identify potentially hallucinated content (Manakul et al., 2023; Zhang et al., 2023a). These approaches assume that hallucinated content often corresponds to regions of high model uncertainty.

Consistency-based approaches detect hallucinations by examining the consistency of model outputs across different prompting strategies or model variants. Methods like SelfCheckGPT (Manakul et al., 2023) generate multiple responses to the same query and flag inconsistencies as potential hallucinations, while other approaches use paraphrasing or different question formulations to probe consistency (Li et al., 2023).

External verification methods compare model outputs against external knowledge sources, databases, or retrieval systems to verify factual claims (Chern et al., 2023; Peng et al., 2023). These approaches often require access to structured knowledge bases or web search capabilities but can provide more definitive assessments of factual accuracy.

3.4.2 Information-Theoretic Perspectives

Recent work has begun exploring information-theoretic frameworks for understanding and detecting hallucinations. The connection between hallucinations and uncertainty quantification suggests that mutual information between model representations and factual knowledge may serve as a principled detection mechanism (Farquhar et al., 2024).

Some approaches frame hallucination detection as an out-of-distribution problem, where hallucinated content represents samples from outside the model's reliable knowledge distribution (Burns et al., 2023). This perspective opens possibilities for applying OOD detection techniques to hallucination identification, potentially unifying these two important safety challenges under a common theoretical framework.

3.4.3 Evaluation and Benchmarks

The evaluation of hallucination detection methods faces significant challenges due to the subjective nature of defining "hallucinations" and the difficulty of creating comprehensive ground truth datasets. Benchmarks such as HaluEval (Li et al., 2023) and TruthfulQA (Lin et al., 2022) provide standardized evaluation frameworks, though they often focus on specific types of factual errors rather than the full spectrum of hallucination phenomena.

The field continues to grapple with fundamental questions about the relationship between hallucinations, model uncertainty, and the broader challenge of ensuring reliable AI systems in high-stakes applications.

3.5 Model Architectures

The choice of model architecture significantly influences both representation learning capabilities and out-of-distribution detection performance. Different architectures exhibit varying inductive biases that affect how they encode information and handle distributional shifts.

3.5.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) remain fundamental for computer vision tasks, with architectures like ResNet (He et al., 2016) and DenseNet providing strong feature representations through hierarchical processing. CNNs' translation equivariance and local connectivity make them particularly effective for learning spatial representations, though their inductive biases can limit generalization to significantly different visual domains.

3.5.2 Transformers

The Transformer architecture (Vaswani et al., 2017) has revolutionized both natural language processing and computer vision through its self-attention mechanism. Vision Transformers (ViTs) (Dosovitskiy et al., 2020) demonstrate that attention-based models can achieve competitive performance on visual tasks, while their global receptive field may provide different robustness characteristics compared to CNNs. The attention mechanism also offers interpretability advantages for under-

standing model uncertainty and potential OOD behavior.

3.5.3 Foundation Models

Large-scale foundation models like CLIP (Radford et al., 2021), GPT (Brown et al., 2020), and BERT (Devlin et al., 2018) represent a paradigm shift toward general-purpose architectures trained on diverse data. These models exhibit emergent capabilities and transfer learning properties that significantly impact both representation quality and OOD detection performance. Their scale and training diversity often lead to more robust representations, though they also introduce new challenges for understanding and controlling their behavior on out-of-distribution inputs.

Chapter 4

Label Blindness in Unlabeled OOD Detection

This chapter is based on work published at ICLR 2025: "Can We Ignore Labels in Out-of-Distribution Detection?" by Hong Yang, Qi Yu, and Travis Desell.

4.1 Introduction

Safety-critical applications of deep neural networks have recently become an important area of investigation in the domain of artificial intelligence, ranging from autonomous driving (Ramanagopal et al., 2018) to biometric authentication (Wang & Deng, 2021) to medical diagnosis (Bakator & Radosav, 2018). In the setting of safety-critical systems, it is no longer possible to rely on the closed-world assumption (Krizhevsky et al., 2012), where test data is drawn i.i.d. from the same distribution as the training data, known as the in-distribution (ID). These models will be deployed in an open-world scenario (Drummond & Shearer, 2006), where test samples can be out-of-distribution (OOD) and therefore should be handled with caution. OOD detection seeks to identify inputs containing a label that was never present in the training distribution. The motivation for OOD detection is simple: we do not want safety-critical systems to act on an invalid prediction, where the predicted label cannot be correct because the label was never present in training.

There is significant interest in unlabeled OOD detection due to various factors. A method that does not rely on labels can save significant costs in labeling data, as proposed by Sehwag et al. (2021).

It is also possible to skip training on the in distribution data if such a model is generalizable, as proposed by Wang et al. (2023). Self supervised and unlabeled learning methods can also scale to much larger datasets and it is important for these models to be robust to OOD data. Recent work in unlabeled OOD detection methods, including Guille-Escuret et al. (2024); Liu et al. (2023); Sehwag et al. (2021); Tack et al. (2020); Wang et al. (2023), promise to improve safety using only unlabeled data. These methods can achieve even greater performance than a simple supervised baseline (Hendrycks & Gimpel, 2016), suggesting that one could replace supervised training with self-supervised learning (SSL) for a safety critical OOD detection task. This family of SSL OOD methods differ from traditional supervised OOD methods, including Fort et al. (2021), by the use of only unlabeled data. The importance of labels is an active area of research in OOD detection (Du et al., 2024a,b).

When we view SSL from an information-theoretic perspective, the selection of features depends solely on the SSL objective and not on the labels. This, however, provides no guarantee that any features relevant for label prediction will be retained. Figure 4.1 provides an example of how SSL features can be less effective for identifying a label. Our theory importantly shows that, when the label-relevant features are independent of the features relevant for the SSL algorithm's successful operation, OOD detection is guaranteed to fail due to what we call 'label blindness' and that this label blindness occurs regardless of how one selects the ID dataset from the population of all data. Our experiments also suggest that Zero Shot OOD methods (Esmaeilpour et al., 2022; Wang et al., 2023) may also suffer from this issue. We show that unsupervised OOD detection methods behave in the same way as SSL in the context of information theory.

However, one can unintentionally avoid label blindness problem via the selection of the OOD dataset when constructing OOD benchmarks. Existing methods generally consider ID and OOD data from different datasets, e.g., Fort et al. (2021), Sehwag et al. (2021), and Hendrycks et al. (2019). In these benchmarks, there is no significant overlap between the ID and OOD input data, allowing OOD detection algorithms to succeed on features independent of the label. To address this issue and to test for label blindness, we introduce the Adjacent OOD detection task to evaluate the performance on OOD detection algorithms when there is significant overlap between the OOD input data and ID input data. We also prove that it is impossible to guarantee that a real world system will never encounter OOD input data that significantly overlaps ID input data.

This work aims to answer the following question: *can we ignore labels when engaging in OOD detection?* Through numerous experiments and theoretical proofs, we show that it is not safe to ignore labels when performing OOD detection. This is contrary to the increasing recent efforts

Figure 4.1: An example failure case by visualizing the heatmaps of the gradient of a unlabeled SimCLR trained ResNet (Chen et al., 2020b) using the GradCAM method (Selvaraju et al., 2017). The OOD detection task is to detect OOD facial expressions. In this case, the OOD detection method fails as justified by our theoretical work, where the representations do not exhibit a strong gradient in regions commonly associated with facial expressions (i.e., eyebrows, mouth, etc.).

that propose new self supervised, unsupervised, and other unlabeled OOD detection methods. This work's key contributions include:

- **The Label Blindness Theorem.** We theoretically prove that any SSL or Unsupervised Learning algorithm will fail when its information required for the surrogate task is independent of the information required for predicting labels. Through this proof, we conclude that there cannot be a generally applicable SSL or Unsupervised learning OOD detection algorithm as there will always exist independent labels due to the no free generalization theorem.
- **Adjacent OOD detection benchmarks.** We introduce the concept of bootstrapping without replacement of the ID labels to create the Adjacent OOD detection task. To the authors' knowledge, this OOD detection task is novel to and absent from research in OOD detection. This task evaluates OOD detection when there is significant overlap in OOD data and ID. We also theoretically prove that overlapping OOD and ID data is possible in every real world dataset.
- **Impact on existing and future OOD methods.** We demonstrate that existing SSL and Unsupervised Learning OOD methods fail under the conditions suggested by our theory and that existing benchmarks do not capture such failures. We also evaluate zero shot OOD

detection methods, which fail in a similar manner to SSL and Unsupervised Learning OOD methods. We make recommendations on the development and testing of future OOD methods.

4.2 Preliminaries

4.2.1 Labeled and Unlabeled Out-of-Distribution Detection

The task of out-of-distribution detection is to identify a semantic shift in the data (Yang et al., 2021). This is determining when no predicted label could match the true label $\boldsymbol{y} \notin \mathbb{Y}_{in}$, where $\mathbb{Y}_{in}$ represents the set of in-distribution training labels. In this case, we would consider the semantic space of the sample and the training distribution to be different, representing a semantic shift. We can express the probability that a sample is out-of-distribution via $P(\boldsymbol{y} \notin \mathbb{Y}_{in}|\boldsymbol{x})$. One baseline method to calculate $P(\boldsymbol{y} \notin \mathbb{Y}_{in}|\boldsymbol{x})$ is to take $1 - \texttt{MSP}(\boldsymbol{x})$, where $\texttt{MSP}$ is the maximum softmax probability from a classifier for a particular datapoint.

Furthermore, we are only concerned with labels that can be generated using only $\boldsymbol{x}$, via function f which depends solely on $\boldsymbol{x}$ and no other information. f may represent human labelers that generate $\boldsymbol{y}$. If we consider $\mathbb{Y}_{all}$ as the set of all possible labels that can be generated from $f(\boldsymbol{x} \in \mathbb{X}_{all})$, a subset of $\mathbb{X}_{all}$ considered as $\mathbb{X}_{training}$ may not contain all labels in $\mathbb{Y}_{all}$. For real world datasets, it is possible that $\mathbb{Y}_{in} \subsetneq \mathbb{Y}_{all}$.

We can also approach the problem of OOD detection without the use of labels. One can train a model on ID data using a surrogate task for the purposes of computing a metric. For example, Sehwag et al. (2021) trains a resnet with SimCLR and computes the Mahalanobis distance between the training representations and the test sample representations to compute the OOD score. Alternatively, one could utilize a pretrained model with broad knowledge to compute a metric to use as the OOD score, such as in Wang et al. (2023).

4.2.2 Self-Supervised and Unsupervised Learning

This section covers representation learning and its implications for SSL and unsupervised learning. If there is no mutual information between two random variables, neither can be used to reduce uncertainty about the other (Shannon, 1948). In both self-supervised and unsupervised OOD detection, if there is no mutual information between the intermediate representations and the

OOD detection task, the OOD detection system cannot reduce uncertainty with respect to the OOD detection task using the intermediate representations.

Representation learning can be formulated as finding a distribution $p(\mathbf{z}|\mathbf{x})$ that maps the observations from $\boldsymbol{x} \in \mathbb{X}$ to $\boldsymbol{z} \in \mathbb{Z}$, while capturing relevant information for some primary task. When $\mathbf{y}$ represents some primary task, we consider only $\mathbf{z}$ that is sufficiently discriminative for accomplishing the task $\mathbf{y}$. For simplicity, we consider $\mathbf{y}$ as a classification label, but $\mathbf{y}$ can represent any objective or task. Federici et al. (2020) show that this sufficiency is met when the information relevant for predicting $\mathbf{y}$ is unchanged when encoding $\mathbf{x} \to \mathbf{z}$.

Definition 4.2.1. Sufficiency: A representation $\mathbf{z}$ of $\mathbf{x}$ is sufficient for $\mathbf{y}$ if and only if $I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) = 0$.

Since there exists the sufficient statistic $\mathbf{x} = \mathbf{z}$, we must consider the minimal sufficient statistic which conveys only relevant information for predicting $\mathbf{y}$. An SSL algorithm seeks to learn the minimal sufficient statistic via the information bottleneck framework (Shwartz-Ziv & LeCun, 2023).

Definition 4.2.2. Minimal Sufficient Statistic. A sufficient statistic $\mathbf{z}$ is minimal if, for any other sufficient statistic $\mathbf{s}$, there exists a function f such that $\mathbf{z} = f(\mathbf{s})$.

Information bottleneck optimization can be expressed as the minimization of the representation's complexity via $I(\mathbf{x}; \mathbf{z})$ and maximizing its utility $I(\mathbf{z}; \mathbf{y})$. This results in the information theoretic loss function below, where β is a trade-off between complexity and utility (Shwartz-Ziv & LeCun, 2023). In practice, learning $\mathbf{z}$ without $\mathbf{y}$ requires a surrogate task $\mathbf{y}_s$, e.g., Chen et al. (2020b), with the loss defined as:

$$\mathcal{L} = I(\mathbf{x}; \mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y}). \tag{4.1}$$

It should be noted that the primary task $\mathbf{y}$ may be equal to the SSL task $\mathbf{y}_s$. In such a case, compression towards the minimal sufficient statistic still occurs. This is important because unsupervised methods for deep neural networks (DNNs) will use a surrogate task $\mathbf{y}_u$ to train the DNN's weights. Thus, if we assign the primary task for an unsupervised learning method to be equal to its surrogate task, it will behave identically to SSL from the perspective of information theory.

When $\mathbf{x}$ has higher information content than $\mathbf{y}$, there exists information in $\mathbf{x}$ that is not relevant for predicting $\mathbf{y}$. This can be better understood by dividing $I(\mathbf{x}; \mathbf{z})$ into two terms (Federici et al., 2020) as follows:

$$I(\mathbf{x}; \mathbf{z}) = \underbrace{I(\mathbf{x}; \mathbf{z} \mid \mathbf{y})}_{\text{superfluous information}} + \underbrace{I(\mathbf{z}; \mathbf{y})}_{\text{predictive information}}. \tag{4.2}$$

However, superfluous information is not affected by the labels of primary task, only by $\mathbf{x}$ and $\mathbf{y}_s$. Using information theory, we can show that any SSL OOD detection algorithm will fail when the surrogate task $\mathbf{y}_s$ is independent of the labels in the in-distribution dataset. This applies to unsupervised OOD detection algorithms that also use a surrogate task.

4.3 Guaranteed OOD Detection Failure

This section introduces the concept of **Label Blindness**, with one key supporting theorem and one key supporting lemma. Note that $R_{\mathbf{x}}$ represents the support of random variable $\mathbf{x}$ such that $R_{\mathbf{x}} = \{\boldsymbol{x} \in \mathbb{R} : P(\boldsymbol{x}) > 0\}$. For clarity, we refer to cases where $I(\mathbf{x}_1; \mathbf{x}_2) = 0$ as **Strict Label Blindness** and discuss **Approximate Label Blindness** $I(\mathbf{x}_1; \mathbf{x}_2) \approx 0$ later in this section.

4.3.1 Label Blindness Theorem (Strict Label Blindness)

We identify a guarantee of OOD detection failure for any information bottleneck-based optimization process if the unlabeled learning objective is independent from labels used to determine the ID set, described by Corollary 4.3.3. This corollary is derived from two concepts: strict label blindness in the minimal sufficient statistic and the independence of filtered distributions. We first consider the minimal sufficient statistic and how it leads to strict label blindness; see Theorem 4.3.1.

Theorem 4.3.1 (Strict Label Blindness in the Minimal Sufficient Statistic)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. Let* $\mathbf{y}_1$ *be a surrogate task such that* $H(\mathbf{y}_1|\mathbf{x}_1) = 0$*. Let* $\mathbf{z}$ *be any sufficient representation of* $\mathbf{x}$ *for* $\mathbf{y}_1$ *that satisfies the sufficiency definition 4.2.1 and minimizes the loss function* $\mathcal{L} = I(\mathbf{x}_1\mathbf{x}_2; \mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y}_1)$*. The possible* $\mathbf{z}$ *that minimizes* $\mathcal{L}$ *and is sufficient must meet the condition* $I(\mathbf{x}_2; \mathbf{z}) = 0$*.*

Proof: See Appendix A.3.1 for the complete proof.

Intuitively, the minimal sufficient representation cannot encode any information independent of the surrogate learning objective, otherwise it would not be minimal. This means that the representation will be blind to any label built upon the independent information.

However, Theorem 4.3.1 is not sufficient to guarantee OOD failure. This is because the selection of the ID training set could change the learned representation $\mathbf{z}$, possibly improving OOD detection

performance by increasing mutual information, $I(\mathbf{x}_2; \mathbf{z}) > 0$. We formally disprove this possibility through Lemma 4.3.2.

Lemma 4.3.2 (Independence of Filtered Distributions)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. For* $\mathbf{x}'_2$ *where* $R_{\boldsymbol{x}'_2} \subset R_{\boldsymbol{x}_2}$*, there exists no* $\mathbf{x}'_2$ *such that* $H(\mathbf{x}_1|\mathbf{x}'_2) < H(\mathbf{x}_1)$*.*

Proof: See Appendix A.3.2 for the complete proof.

Lemma 4.3.2 states that filtering on a label generated on one of two independent variables cannot provide information about the other. This applies to the selection of ID data from the population, if the selection criteria is independent of the learning objective. This means that the strict label blindness properties predicted by Theorem 4.3.1 will apply to ID training data. These two concepts bring us to our main result – strict label blindness in filtered distributions; see Corollary 4.3.3.

Corollary 4.3.3 (Strict Label Blindness in Filtered Distributions)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. Let* $\mathbf{y}_1$ *be a surrogate task such generated by* $\boldsymbol{y}_1 = f_1(\boldsymbol{x}_1)$ $H(\mathbf{y}_1|\mathbf{x}_1) = 0$*. Let* $\mathbf{y}_2$ *be a label such that* $H(\mathbf{y}_2|\mathbf{x}_2) = 0$ *and* $\boldsymbol{y}_2 = f_2(\boldsymbol{x}_2)$*. Let* $\mathbb{Y}_{in}$ *be as subset of labels* $\mathbb{Y}_{in} \subset R_{\mathbf{y}_2}$*. Let* $\mathbf{x}'$ *be a subset of* $\mathbf{x}$ *where* $R_{\mathbf{x}'} = R_{\mathbf{x}} \cap \{\boldsymbol{x} \in \mathbb{R} : f_2(\boldsymbol{x}_2) \in \mathbb{Y}_{in}\}$ *such that* $\mathbf{x}'$ *is composed of independent variables* $\mathbf{x}'_1$ *and* $\mathbf{x}'_2$ *and* $\boldsymbol{y}'_1 = f_1(\boldsymbol{x}'_1)$*. The sufficient representation* $\mathbf{z}$ *learned by minimizing* $\mathcal{L} = I(\mathbf{x}'_1\mathbf{x}'_2; \mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y}'_1)$ *must have* $I(\mathbf{x}_2; \mathbf{z}) = 0$ *and* $I(\mathbf{y}_2; \mathbf{z}) = 0$*.*

Proof: See Appendix A.3.3 for the complete proof.

This means that, when we select the ID training data, if the selection criteria and labels are independent of the surrogate learning objective, then we can guarantee failure in OOD detection due to the absence of any information in the learned representation $\mathbf{z}$. For simplicity, we refer to this concept, supported by Corollary 4.3.3, as the problem of **Strict Label Blindness**.

In summary, when the surrogate learning task can be achieved without learning about features relevant for the label, it will not learn any features relevant for the label. If the SSL or unsupervised learning method fails to learn any label-relevant features, then any OOD detection algorithm built from those representations cannot differentiate between the labels selected as ID and those not selected as ID. This guarantees failure in OOD detection because no label information passes through the information bottleneck.

4.3.2 Implications of Strict Label Blindness in Real World Situations

We can utilize Fano Inequality to extend our understanding of strict label blindness to consider situations where the variables are not fully independent. The lower bound for prediction error is defined by the entropy of the target label $\mathbf{y}$ less the mutual information between the input $\mathbf{x}$ and target label, as shown in Theorem 4.3.4. Under strict label blindness, when $I(\mathbf{x};\mathbf{y}) = 0$, the lower bound for error is at its maximum. When $I(\mathbf{x};\mathbf{y}) \approx 0$, the lower bound for error is large enough to be unreliable. We refer to this condition as **Approximate Label Blindness** and we conduct experiments to evaluate this condition. Unless specified as strict, label blindness refers to the approximate case.

Theorem 4.3.4 (Fano's Inequality)**.** *Let* $\mathbf{y}$ *be a discrete random variable representing the true label with* $\mathcal{Y}$ *possible values and cardinality of* $|\mathcal{Y}|$ *and* $\mathbf{x}$ *be a random variable used to predict* $\mathbf{y}$*. Let* e *be the occurrence of an error such that* $\mathbf{y} \neq \hat{\mathbf{y}}$ *where* $\hat{\mathbf{y}} = f(\mathbf{x})$*. Let* H_b *represent the binary entropy function such that* $H_b(e) = -P(e)\log P(e) - (1 - P(e))\log(1 - P(e))$*. The lower bound for* $P(e)$ *increases with lower* $I(\mathbf{x};\mathbf{y})$*.*

$$H_b(e) + P(e)\log(|\mathcal{Y}| - 1) \geq H(\mathbf{y}) - I(\mathbf{x};\mathbf{y}). \tag{4.3}$$

4.3.3 Theoretical Implications

Our work applies to deep neural networks (DNNs) trained without labels for the purpose of OOD detection. The key assumption of information bottleneck compression is generally applicable to DNNs (Shwartz-Ziv & Tishby, 2017). Regardless of other assumptions, such as the multi-view assumption, an information bottleneck DNN trained without labels will still compress data irrelevant to its loss objective, even if that data is relevant for its intended task. It does not matter what task the training process was originally designed for because the unlabeled training process ultimately generates/adheres to its own learning objective. For any learning objective, there will exist an independent feature unless compression is not possible, as in $I(\mathbf{x};\mathbf{y}) = I(\mathbf{x};\mathbf{x})$. If there is compression, then there exists labels for which OOD detection failure is guaranteed.

Our work predicts a guarantee of failure only when we consider the OOD set of all non-ID data. In our own experiments and in work by Hendrycks et al. (2019); Liu et al. (2023); Sehwag et al. (2021), purely self-supervised and unsupervised OOD methods can perform well against common benchmark OOD sets. This suggests that the choice of ID set and OOD set pairs can unintentionally

hide label blindness failure. Alternatively, we can also construct a test to identify if the OOD detection algorithm suffers from label blindness. To construct such a test, we rely on the insight from Corollary 4.3.3 and the use of a simple statistical method.

4.4 Benchmarking for Label Blindness Failure

4.4.1 Bootstrapping and the Adjacent OOD Benchmark

One logical consequence of Corollary 4.3.3 is that one cannot avoid failure due to label blindness by selecting different labels for one's ID set, so long as the label selection is independent from the learning objective. To test any OOD detection algorithm for label blindness failure, this simply entails selecting different labels for one's ID set. To construct this benchmark, we randomly sample labels to be considered as ID and other labels to be consider OOD. This is similar to bootstrapping, but without replacement. If an OOD detection algorithm is 'approximately label blind', its average OOD detection performance across the samples should be poor. We refer to this as the **Adjacent OOD Detection Benchmark**.

4.4.2 Why Adjacent OOD is Safety-Critical to Almost All Real World Systems

The Adjacent OOD detection benchmark evaluates the performance of OOD detection algorithms when there may be a significant overlap between the ID data and OOD data. This condition applies to all systems where it is impossible to guarantee that there will be no significant overlap in the feature space between ID and OOD data. This is true for almost all real world systems and is theoretically proven below.

Theorem 4.4.1 (Unavoidable Risk of Overlapping OOD Data)**.** *Let* $\mathbf{x}$ *come from a distribution. Let* f *be some labeling function to generate labels* $\mathbf{y}$ *such that* $\boldsymbol{y} = f(\boldsymbol{x})$*, where there are at least two unique labels* $|R_{\mathbf{y}}| > 1$*. Let* $\mathbf{x}_{in}$ *be a random subset of* $\mathbf{x}$ *where* $R_{\mathbf{x}_{in}} \subsetneq R_{\mathbf{x}}$ *and* $|R_{\mathbf{x}_{in}}| < \infty$*. Let* $\mathbf{y}_{in}$ *be labels generated from* $\boldsymbol{y}_{in} = f(\boldsymbol{x}_{in})$*. The probability that a randomly selected* $\boldsymbol{x}$ *contains* $\boldsymbol{y}$ *not present in* $R_{\mathbf{y}_{in}}$ *is always greater than 0.*

Proof: See Appendix A.3.4 for the complete proof.

In theory, this risk can be reduced to an acceptable level by adding more data to the training dataset. However, this reduction in risk requires the assumption that the collected data is randomly

sampled. This is almost never true for real world datasets and often the opposite is true, where the nature of sampling can significantly increase this risk.

One risk factor present in every real world dataset is the dataset creation date. By creating the dataset at any specific point in time, the dataset cannot be randomly sampled with respect to time because it is impossible to collect data from the future. For example, if one were to create a dataset of diseases today, it would not contain any future diseases. In this example, the probability that the training dataset is incomplete is 100%, which guarantees that there will be OOD data that significantly overlaps with ID data. For most real world systems, the only safe assumption is that there may be OOD data that overlaps with ID data and it is necessary to plan accordingly.

The failure predicted by the label blindness theory is easiest to detect in the adjacent OOD situation. Where there is a likelihood of adjacent data, Theorem 4.3.3 predicts OOD detection failure. Where there is no adjacent data, features independent of the label can still be used to distinguish between ID data and non adjacent OOD data, as shown in various experiments in this paper and others (Hendrycks et al., 2019; Liu et al., 2023; Sehwag et al., 2021).

4.4.3 Comparing Adjacent, Near, and Far OOD Benchmarks

Many unlabeled OOD methods generally perform quite well on far and near OOD tasks. Far OOD is often defined by ID and OOD sets with different semantic labels and styles (Fang et al., 2022). One such far OOD benchmark is MNIST as ID data and CIFAR10 as OOD data. Near OOD contains ID and OOD sets with similar semantic labels and styles (Fang et al., 2022). These tasks tend to be more difficult for existing OOD detection methods than far OOD detection tasks. One such near OOD benchmark is CIFAR10 as ID and CIFAR100 as OOD. However, the overlap in the near OOD detection benchmarks is significantly less than the adjacent OOD detection benchmark, which evaluates the maximum possible feature overlap. For example, an Adjacent OOD benchmark on the ICML Facial Expressions dataset may contain the same face with different expressions, resulting in significant feature overlap. These existing benchmarks do not provide sufficient safety guarantees in applications where there may be significant overlap between ID and OOD data.

4.4.4 Implications for OOD from Unlabeled Data

While methods that utilize only unlabeled data, such as Guille-Escuret et al. (2024); Liu et al. (2023); Sehwag et al. (2021), show promising results on both near and far OOD tasks, their per-

formance in the adjacent OOD detection tasks depends on the mutual information between the learned representation and the ID labels. Our theoretical work suggests that such methods will perform poorly, if the surrogate task is independent of the labels.

The adjacent OOD detection benchmark can also evaluate the performance of zero shot OOD detection methods. While our theoretical work does not extend to pretraining due to the use of labels, it is also still important to consider the performance when OOD data overlaps ID data.

4.5 Experimental Results

We conduct the following experiments to verify the existence of label blindness in unlabeled OOD detection methods. All hyperparameters and configurations were the best performing from their respective original paper implementations, unless noted otherwise. Experiments are repeated 3 times.

4.5.1 Experimental Setup

Supervised Baseline. We use Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016) as our baseline supervised method for comparison. We augment the training data using random rotation, horizontal flip, random crop, gray scale, and color jitter. Images are resized to 64×64. We train using stochastic gradient descent with momentum and a cosine annealing learning schedule. We train for 10 warm up epochs followed by 150 regular epochs, selecting the weights with the highest validation accuracy. We use a standard ResNet50 architecture.

Self-supervised Baselines. We use two SSL methods to evaluate how representations are learned, SimCLR (Chen et al., 2020b) and Rotation Loss (RotLoss) (Hendrycks et al., 2019). Images are resized to 64×64 for both cases. For SimCLR, we augment the training data using random rotation, horizontal flip, random crop, gray scale, and color jitter. For Rotation Loss, we use only random crop and horizontal flip. We train using stochastic gradient descent with momentum (and a cosine annealing learning schedule) and employ a standard ResNet50 architecture and train for 10 warm up epochs followed by 500 regular epochs, selecting the weights with the best-learned representations. We use a KNN classifier to determine the best representations during validation at the end of each epoch.

To evaluate OOD performance, we use two methods to generate the OOD score of each sample, SSD (Sehwag et al., 2021) and KNN, similar to Sun et al. (2022). SSD considers the OOD score as the Mahalanobis distance of the sample from the center of all in-distribution training data samples. The KNN method considers the OOD score as the Euclidean distance from the Nth nearest neighbor of the test sample to all in-distribution training samples. Both methods are distance based OOD detection and are commonly used with representation learning. We use the same representation mentioned in the previous paragraph.

Unsupervised Baseline. To consider how an unsupervised OOD detection method functions, we evaluate the diffusion impainting OOD detection method proposed by Liu et al. (2023) using code provided in their paper's linked repository. We utilize the training configuration that generated the paper's main results, which involved an alternating checkerboard mask 8×8, an LPIPS distance metric to calculate the OOD score, and 10 reconstructions per image. We modify only the input image size to be 64×64 for all datasets and run additional experiments to evaluate performance on their alternative MSE distance metric. This method is representative of other generative methods, such as Xiao et al. (2020).

Zero-shot Baseline. To consider how well zero shot learning algorithms perform, we evaluate the CLIPN model presented by Wang et al. (2023). We utilize their pretrained weights provided in their paper's repository and perform zero shot OOD detection on our adjacent OOD detection benchmark. We evaluate CLIPNs performance using 3 of their paper's algorithms, Maximum Softmax Probability, Compete to Win (CTW), and Agree to Differ (ATD).

4.5.2 Adjacent OOD Datasets

To create the Adjacent OOD detection task, we randomly split 25% of all classes into the OOD set and retain 75% as the ID set. We also repeat our experiments three times with different seeds to account for different splits of the ID and OOD set. Only ICML Facial expressions has a major class imbalance for one of its seven classes.

The ICML Facial Expressions dataset (Erhan et al., 2013) contains seven facial expressions split across $28,709$ faces in the train set and $7,178$ in the test set. The expressions include anger, disgust, fear, happiness, sadness, surprise, and neutral. Self-supervised algorithms may not learn relevant features for distinguishing expressions and instead learn features relevant for distinguishing faces.

The Stanford Cars dataset (Krause et al., 2013) contains 16,185 images taken from 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, with each class being split roughly 50-50. Classes are typically very fine-grained, at the level of Make, Model, Year, e.g., 2012 Tesla Model S or 2012 BMW M3 coupe. This creates a particularly challenging Adjacent OOD task because of the reliance on more subtle features to differentiate cars.

The Food 101 dataset by Bossard et al. (2014a) consists of 101 food categories and 101,000 images. There are 250 manually reviewed test images and 750 training images for each class. Note that training images were not cleaned to the same standard as the test images and will contain some mislabeled samples. We believe that this should not significantly detract from the Adjacent OOD nature of the dataset.

4.5.3 Experimental Results

Experimental results for Adjacent OOD are presented in Table 4.1. It is apparent that the baseline supervised method performs better than most unlabeled methods on the Adjacent OOD detection task. In cases where the unlabeled methods exhibits performance as good as random guessing, it is likely that the learned representation contains little information about the semantic label. This is contrary to the reported performance improvements presented in unlabeled OOD papers (Hendrycks et al., 2019; Liu et al., 2023; Sehwag et al., 2021), as our experimental results suggest unlabeled OOD is significantly worse than a simple MSP baseline.

It is important to note that the zero shot CLIPN method performs well when the label text's usage in pretraining is similar to the label text's usage in the ID data. In the case of the Cars dataset, the pretraining dataset CC3M (Sharma et al., 2018) contains many images captioned with the make and model of various cars, resulting in good performance. The Food dataset also sees similar label usage in the pretraining set. However, the Faces dataset's labels are not aligned. For example, there are multiple images associated with the emotion angry that do not contain a human face, such as an image of a angry fist. When there is little or no mutual information between the pretraining data and the ID labels, zero shot methods will perform poorly in OOD detection tasks.

We observe decent OOD performance on the unlabeled SimCLR compared to the labeled supervised MSP for CIFAR10 and CIFAR100 datasets. This is likely because the SimCLR algorithm is better at learning the relevant features in these datasets and that the classes are more visually dissimilar, resulting in less overlap of OOD and ID data. We also show strong results for far OOD performance for SimCLR based OOD detection, which confirms findings in papers that test unlabeled OOD

Table 4.1: Results from experiments across various datasets and methods. Unlabeled methods perform poorly in adjacent OOD detection. CLIPN performance is due to labels present in the pretraining dataset. Higher AUROC and lower FPR is better.

	Faces		Cars		Food	
Method	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95
Supervised MSP	70.8±0.3	88.2±0.2	69.2±0.9	88.8±0.8	78.8±1.2	81.1±1.6
SimCLR KNN	52.0±4.2	95.0±1.3	52.5±0.4	94.0±0.5	61.1±2.8	91.6±1.6
SimCLR SSD	55.0±4.5	95.1±2.0	52.7±0.7	93.7±1.1	64.4±0.8	89.3±0.5
RotLoss KNN	46.1±2.5	95.8±0.4	51.1±0.6	94.8±0.7	49.7±3.8	94.9±0.9
RotLoss SSD	46.6±3.0	95.7±0.5	50.7±1.9	95.0±1.2	50.7±3.6	94.9±0.9
Diffusion LPIPS	54.7±4.6	94.2±3.7	53.8±1.8	93.9±1.2	52.9±2.2	94.4±0.6
Diffusion MSE	55.3±2.2	94.2±1.4	51.6±1.6	94.4±0.5	52.5±3.4	94.2±0.6
CLIPN CTW	47.0±1.4	97.3±0.3	65.0±5.1	69.4±9.4	70.9±2.9	69.1±7.0
CLIPN ATD	44.2±1.4	97.5±0.2	81.1±4.3	56.6±10.4	84.9±0.2	53.9±4.5
CLIPN MSP	58.7±4.4	95.9±1.4	76.5±1.4	75.4±0.6	80.5±1.6	74.0±1.4

methods against a far OOD detection benchmark (Guille-Escuret et al., 2024; Liu et al., 2023; Sehwag et al., 2021; Tack et al., 2020; Wang et al., 2023).

4.6 Discussion

4.6.1 Impact of Label Blindness on Future Research

A consequence of the label blindness theorem is that there cannot exist a single unlabeled OOD detection algorithm for all unlabeled data. However, unlabeled learning methods, such as SimCLR, are vital for improving OOD detection. The model of Sun et al. (2022) learns representations using a supervised version of SimCLR, similar to Khosla et al. (2020). The combination of a multi-view information bottleneck with supervised classes produces a more robust representation of the in-distribution data than using only a supervised loss. Recent work by Du et al. (2024a) provides a strong theoretical basis for why unlabeled data can improve OOD detection performance.

The Adjacent OOD detection benchmark addresses a critical safety gap in existing OOD detection

research. Current benchmarks often use datasets with minimal feature overlap between ID and OOD data, which can mask the label blindness problem. In real-world applications, especially safety-critical ones, it is essential to evaluate OOD detection methods under conditions where ID and OOD data may have significant feature overlap.

Our findings suggest that practitioners should be cautious when deploying unlabeled OOD detection methods in scenarios where the training data may not capture all relevant semantic variations. The theoretical guarantees provided by our label blindness analysis indicate that such methods may fail catastrophically when encountering OOD data that shares features with ID data but differs in the semantic labels.

4.6.2 Recommendations for Future OOD Detection Research

Based on our theoretical and empirical findings, we recommend several directions for future research:

- **Hybrid approaches**: Combining supervised and self-supervised learning objectives may help mitigate label blindness by ensuring that label-relevant features are preserved during representation learning.
- **Adjacent OOD evaluation**: All OOD detection methods should be evaluated on Adjacent OOD benchmarks to assess their robustness to scenarios with high feature overlap between ID and OOD data.
- **Information-theoretic analysis**: Future unlabeled OOD detection methods should include theoretical analysis of the mutual information between their learning objectives and the target labels to identify potential label blindness issues.
- **Domain-aware methods**: Developing methods that can explicitly model and account for domain-specific features may help address some of the limitations identified in our work.

4.7 Conclusion

In this work we provide an answer to the question, can we ignore labels for OOD detection? Our theoretical work shows that the answer is no, unless the unlabeled method happens to capture the relevant features and does not need to work for different sets of labels. Due to the lack of existing

benchmarks that capture the theoretically expected failure, we introduce a novel type of OOD task, Adjacent OOD detection. This task addresses the critical safety gap caused by significant overlap of ID and OOD data. We show that the Adjacent OOD task accurately captures the failure in unlabeled OOD detection that is hypothesized by our theory.

The label blindness theorem demonstrates that when the surrogate learning task used in self-supervised or unsupervised learning is independent of the features relevant for label prediction, OOD detection is guaranteed to fail. This fundamental limitation cannot be overcome by simply selecting different ID datasets, as the independence property is preserved under filtering operations.

Our experimental results confirm the theoretical predictions, showing that unlabeled OOD detection methods perform poorly on Adjacent OOD benchmarks where there is significant feature overlap between ID and OOD data. In contrast, these same methods often perform well on traditional far OOD benchmarks, highlighting the importance of comprehensive evaluation.

The Adjacent OOD detection benchmark introduced in this work provides a crucial tool for evaluating the robustness of OOD detection methods in realistic scenarios. This benchmark addresses a previously ignored safety gap in OOD detection research and should be adopted as a standard evaluation protocol for future work.

We hope our work will help support more robust research into OOD detection and improve the safety of AI applications. The theoretical framework and empirical findings presented here provide important guidance for developing more reliable OOD detection methods that can handle the complexities of real-world deployment scenarios.

Chapter 5

Domain Feature Collapse in Single-Domain OOD Detection

This chapter is based on work submitted to AAAI 2026: "Domain Feature Collapse: Implications for Out-of-Distribution Detection and Solutions" by Hong Yang, Qi Yu, and Travis Desell.

5.1 Introduction

The deployment of deep neural networks (DNNs) in safety-critical domains – such as autonomous driving (Ramanagopal et al., 2018), biometric authentication (Wang & Deng, 2021), and medical diagnostics (Bakator & Radosav, 2018) – has spurred growing interest in ensuring their reliability. In these contexts, the traditional closed-world assumption (Krizhevsky et al., 2012), where training and test data are drawn i.i.d. from the same in-distribution (ID), is no longer valid. Instead, models must operate under open-world conditions (Drummond & Shearer, 2006), where inputs encountered at test time may stem from entirely different, out-of-distribution (OOD) sources.

Many state-of-the-art OOD detection methods demonstrate strong performance across established benchmarks (Zhang et al., 2023b). However, these benchmarks almost exclusively use in-distribution sets that contain samples and classes from a wide variety of domains, such as CIFAR10/100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009). While this approach provides a broad testbed for evaluating OOD robustness, it implicitly biases models and methods toward handling multi-domain in-distribution settings. As a result, there exists a gap in the literature: current OOD

detection techniques are largely tailored to scenarios where the ID data is inherently diverse, rather than narrow or homogeneous.

The single domain setting is understudied in bleeding edge OOD detection research, yet it has been heavily studied in the application of OOD methods for downstream tasks. Single-domain OOD detection is particularly important in application areas such as medical imaging (Zhang et al., 2021), satellite imagery (Ekim et al., 2024), and agriculture (Saadati et al., 2024), where models are often deployed in narrowly scoped environments with highly consistent data characteristics.

This chapter introduces the concept of and theoretically proves the existence of **domain feature collapse**. Through information theory and bottleneck compression, we show that artificial neural networks will remove domain specific features from their learned representations, under the single domain setting. This leads to a situation where OOD detection relies solely on class-specific features, while ignoring domain-specific features. Unfortunately, this failure results in higher OOD detection error rates when the ID set is single domain versus multi-domain.

5.2 Problem Formulation

5.2.1 Single-Domain Datasets and Domain Features

We define the dataset's domain $\mathbf{d}$ as a value generated from some domain labeling function $f_{\mathbf{d}}(\mathbf{x})$. For the purposes of this work, we are primarily concerned with cases where the data comes from a single domain $\mathbf{d}_1$, such that $\forall \mathbf{x} \in \{f_{\mathbf{y}}(\mathbf{x}) \in \mathbb{Y}_{in}\}, f_{\mathbf{d}}(\mathbf{x}) = \mathbf{d}_1$. In these situations, we can conclude that $\forall \mathbf{x} \in \{f_{\mathbf{d}}(\mathbf{x}) \neq \mathbf{d}_1\}, f_{\mathbf{y}}(\mathbf{x}) \notin \mathbb{Y}_{in}$, since any data outside of the domain cannot possibly have an in-distribution label.

For this work, we define domain features $\mathbf{x_d}$ such that they do not overlap with class features $\mathbf{x_y}$, implying $I(\mathbf{x_d} : \mathbf{x_y}) = 0$. The independence of domain and class features only applies to the training set, as domain features would provide useful information in the context of $\mathbb{X}_{all}$. Note that this also implies that $\neg(\forall \mathbf{x}, f_{\mathbf{y}}(\mathbf{x_y}) = f_{\mathbf{y}}(\mathbf{x}))$ and $\forall \mathbf{x}, f_{\mathbf{y}}(\mathbf{x_y}, \mathbf{x_d}) = f_{\mathbf{y}}(\mathbf{x})$. For both domain and class features, we refer to the minimal set of features, as per the minimal sufficient statistic definition.

Examples of single domain datasets could include a medical chest X-ray dataset (Yang et al., 2023), a geology dataset (Hossain et al., 2021), or a satellite imagery dataset (Helber et al., 2019). Further note that domains exist in a hierarchy; for instance, the domain of cats is a subdomain of mammals

which is itself a subdomain of animals. This means that there is a domain that includes all things, but such a domain would have $\{\mathbf{x_d}\} = \emptyset$. For a wide domain with a wide variety of classes, we expect fewer domain features and more class features.

There exist datasets that could be labeled as a single domain $\mathbf{d}_1$ yet contain $|\{\mathbf{x_d}\}| \approx 0$. For example, if one were to treat ImageNet as a single domain, the set of domain features that do not overlap with class features is likely to be zero or nearly zero. We refer to such datasets as multi-domain datasets, as their diversity of classes require multiple domains.

5.3 Theoretical Analysis: Domain Feature Collapse

This section theoretically proves that any supervised model under a label-based training objective will learn a representation that contains no information on domain features, such that $I(\mathbf{x_d}, \mathbf{z}) = 0$, if full bottleneck compression occurs. This is considered to be strict domain feature collapse and relies on full information bottleneck compression:

Theorem 5.3.1 (Strict Domain Feature Collapse in the Minimal Sufficient Statistic)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x_d}$ *and* $\mathbf{x_y}$*, where* $\mathbf{x_d}$ *is a set of domain features as per Definition 2.5.2. Let* $\mathbf{d}$ *be a domain label generated from* $f_{\mathbf{d}}(\mathbf{x_d}) = \mathbf{d}_1$*, where* $\mathbf{d}_1$ *is a constant value for all* $\mathbf{x}$*. Let* $\mathbf{y}$ *be a class label generated from* $f_{\mathbf{y}}(\mathbf{x_d}, \mathbf{x_y}) = \mathbf{y}$*. Let* $\mathbf{z}$ *be any sufficient representation of* $\mathbf{x}$ *for* $\mathbf{y}$ *that satisfies the sufficiency definition and minimizes the loss function* $\mathcal{L} = I(\mathbf{x_d x_y}; \mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y})$*. The possible* $\mathbf{z}$ *that minimizes* $\mathcal{L}$ *and is sufficient must meet the condition* $I(\mathbf{x_d}; \mathbf{z}) = 0$*.*

Proof. The proof follows from the information bottleneck principle and the independence of domain and class features in single-domain settings. Since $\mathbf{x_d}$ and $\mathbf{x_y}$ are independent ($I(\mathbf{x_d} : \mathbf{x_y}) = 0$) and the domain label $\mathbf{d}$ is constant for all training samples, the domain features $\mathbf{x_d}$ provide no information about the class label $\mathbf{y}$ within the training distribution.

Under the information bottleneck objective, the representation $\mathbf{z}$ seeks to minimize $I(\mathbf{x}; \mathbf{z})$ while maximizing $I(\mathbf{z}; \mathbf{y})$. Since $I(\mathbf{x_d}; \mathbf{y}) = 0$ in the training distribution, including domain features in $\mathbf{z}$ would increase the complexity term $I(\mathbf{x}; \mathbf{z})$ without contributing to the utility term $I(\mathbf{z}; \mathbf{y})$. Therefore, the optimal representation under bottleneck compression will satisfy $I(\mathbf{x_d}; \mathbf{z}) = 0$. □

Intuitively, the minimal sufficient representation cannot encode any information independent of

the learning objective, otherwise it would not be minimal. Due to the definition of $\mathbf{x_d}$ as domain features independent of class features, it is clear that compression results in the loss of domain features in the learned representation. This is contrary to the desired outcome, which is to learn $\hat{\mathbf{y}} = g(\mathbf{x_d}, \mathbf{x_y})$, as this would match the labeling function $\mathbf{y} = f_{\mathbf{y}}(\mathbf{x_d}, \mathbf{x_y})$. Instead, the model learns $\hat{\mathbf{y}} = g(\mathbf{x_y})$ because the domain features $\mathbf{x_d}$ are not predictive of the class in the context of the training data.

The lack of domain features is not problematic for safety purposes when $\forall \mathbf{x} \in \mathbb{X}_{all}, H(\mathbf{d}|\mathbf{x_y}) = 0$; it is safe when all out-of-domain data points contain no in-distribution class features. However, this is difficult to guarantee in an open world setting, as we do not possess information on the OOD distribution. For example, a model might learn that a Tyrannosaurus rex is a dinosaur that stands on two feet and proceed to classify "Barney" (a purple dinosaur character from a children's TV show) as a dinosaur, ignoring the fact that it is purple.

5.3.1 Implications for OOD Detection

Domain feature collapse creates a critical safety gap in OOD detection. When a model trained on single-domain data encounters out-of-domain samples that contain in-distribution class features, it may confidently misclassify them because it cannot recognize the domain shift. This is particularly dangerous in safety-critical applications where the cost of false negatives (failing to detect OOD samples) is high.

The problem is exacerbated by model overfitting, where a model may learn only a subset of $\mathbf{x_y}$ as opposed to the full set of features intended by the practitioner. Suppose we have a bird dataset made up of blue jays and cardinals. A model may only learn that blue jays are blue and assume that any blue object is a blue jay. Such a model would be safer if it could determine the domain of the blue object as a bird, before assuming it is a blue jay.

It should also be noted that full information bottleneck compression may not occur in real world scenarios, yet we can expect that some level of compression would still occur. In such cases, we can use Fano's Inequality to extend our theory of strict domain feature collapse onto partial compression cases. By Fano's Inequality, we would expect to observe unsafe and unreliable OOD detection conditions even with small $I(\mathbf{x_d}; \mathbf{z})$.

5.4 Domain Filtering: A Solution to Domain Feature Collapse

To address the risk of domain feature collapse in supervised networks, we propose a two-stage process that explicitly accounts for domain information before applying traditional OOD detection methods.

5.4.1 Two-Stage Detector: Domain Filtering + OOD Detection

The proposed solution utilizes a two-stage process. In the first stage, a pretrained network is used to determine if a data sample is in-domain. In the second stage, an OOD detector is used to determine if in-domain samples are also in-distribution. This requires the assumption that there exists no in-distribution data sample that is out-of-domain, which is consistent with our earlier definitions.

We evaluate a K-nearest neighbors (KNN)-based domain filter, similar to a KNN-based OOD detector proposed by Sun et al. (2022). To calibrate the domain filter, we calculate the domain threshold $\mathbf{t_d}$ such that $P(f_{knn}(\{\mathbf{x} \in \mathbb{X}_{train}\}) \leq \mathbf{t_d}) = p$, where f_{knn} is a KNN function considering the kth neighbor and p is a hyper parameter set to $p = 0.99$. Essentially, we select a distance such that 99% of the training data falls within that distance. The two stage process considers all samples with $f_{knn} > \mathbf{t}_d$ as OOD (due to it being out-of-domain) and uses the second stage detector to determine an OOD score for samples with $f_{knn} \leq \mathbf{t}_d$.

Algorithm 1 Two-Stage Domain Filtering for OOD Detection

Require: Training set $\mathbb{X}_{train}$, test sample $\mathbf{x}$, domain threshold $\mathbf{t_d}$, OOD detector f_{OOD}

Compute $d_{knn} = f_{knn}(\mathbf{x}, \mathbb{X}_{train})$ ▷ KNN distance to training data
if $d_{knn} > \mathbf{t_d}$ **then**
    return OOD (out-of-domain)
else
    return $f_{OOD}(\mathbf{x})$ ▷ Apply second-stage OOD detector
end if

This process ensures that only a small percentage (1%) of in-domain samples will be flagged as false positives in the first stage. While there are alternative distance calculation methods and percentile thresholds available, we find that a KNN filter at the 99th percentile with $K = 50$ works well as a first stage domain filter.

5.4.2 Relationship to Near, Far, and Adjacent OOD

In most recent work, such as Fort et al. (2021), and in the OpenOOD framework (Yang et al., 2022; Zhang et al., 2023b), there is a distinction between near and far OOD. Near OOD refers to out-of-distribution samples that are semantically different from the training data but visually or structurally similar. Far OOD refers to samples that are both semantically and visually dissimilar, often coming from completely unrelated domains.

However, by definition, both near and far OOD must be considered out-of-domain when the training data comes from a single domain. If an in-distribution dataset is composed of a single domain, e.g., X-rays, where $\{\mathbf{x_d}\} \neq \emptyset$, existing near and far OOD benchmarks will be considered out-of-domain, as they would not be considered in the same domain by $f_{\mathbf{d}}$. We observe that the domain filter is very capable at detecting both near and far OOD benchmark datasets as out-of-domain.

This is in contrast to the adjacent OOD benchmark (Yang et al., 2025), which explicitly tests OOD detection performance on in domain samples that are out-of-distribution. The adjacent OOD benchmark constructs a new in-distribution set using a random subset of the training set classes. It then evaluates the OOD performance against the remaining training set classes as if they were OOD, allowing us to consider the impact of in-domain yet OOD samples. When used alone, the domain filter often performs poorly on the adjacent OOD benchmark, as it is unlikely to contain any class features.

5.5 Experimental Validation

5.5.1 Domain Bench: Single-Domain Datasets

To empirically validate our theoretical findings, we introduce Domain Bench, a comprehensive benchmark consisting of 11 narrow domain datasets that exhibit the characteristics necessary for domain feature collapse. These datasets are specifically chosen because they contain substantial domain features that are independent of class features, making them ideal testbeds for studying domain feature collapse.

The datasets in Domain Bench include:

- **Butterfly** – A butterfly species classification dataset (AIPlanet, 2023)

- **Cards** – A playing card classification dataset by rank and suit (Soni, 2020)
- **Colon** – A colon pathology dataset with different diseases labeled (Yang et al., 2023)
- **Eurosat** – A satellite images dataset for classifying different types of land use (Helber et al., 2019)
- **Fashion** – The FashionMNIST dataset describing different articles of clothing (Xiao et al., 2017)
- **Food** – The Food101 datasets (Bossard et al., 2014b) with 101 classes of different types of food
- **Garbage** – A dataset to classify the material of different waste objects (Single et al., 2023)
- **Plant** – A plant leaves dataset detailing different types of disease (Hughes & Salathé , 2015)
- **Rock** – A dataset of different types of rocks and minerals (Hossain et al., 2021)
- **Tissue** – A kidney cortex microscope dataset with various types of tissue labeled (Yang et al., 2023)
- **Yoga** – A dataset of people performing different yoga poses from the internet (sumanthvrao, 2020)

5.5.2 Experimental Setup

For each narrow domain dataset, we generate ID train, ID validation, ID test, and OOD test datasets using unique seeds. We evaluate three different training approaches:

- **Cross Entropy ResNet50 (CE ResNet)**: Fine-tuned pretrained ResNet50 for 300 epochs using SGD optimizer with initial learning rate of 0.1
- **Cross Entropy DinoV2 (CE DinoV2)**: Fine-tuned pretrained DinoV2 ViTs14 for 75 epochs using Adam optimizer with initial learning rate of 0.0001
- **Supervised Contrastive Learning ResNet50 (SC ResNet)**: ResNet50 trained using supervised contrastive learning for 500 epochs using SGD optimizer with initial learning rate of 0.5 and temperature of 0.5

For OOD evaluation, we use both in-domain and out-of-domain benchmarks. The out-of-domain OOD benchmark includes datasets from OpenOOD (Zhang et al., 2023b): MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), Texture (Cimpoi et al., 2014), Places365 (Zhou et al., 2017), CIFAR10/100 (Krizhevsky et al., 2009), and Tiny ImageNet. For in-domain OOD evaluation, we use the adjacent OOD benchmark (Yang et al., 2025).

5.5.3 Results and Analysis

Our experimental results strongly support the theoretical predictions of domain feature collapse. Table 5.1 shows representative results demonstrating the effectiveness of domain filtering in addressing domain feature collapse.

Table 5.1: Summary OOD Performance Across All Datasets Reported As (In-Domain OOD Score)/(Out-of-Domain OOD Score). We exclude the Rock dataset from this summary as it is an outlier for reasons explained in Section 5.5.6. Best scores are in bold and second best are bold and italicized. The domain filter methods are italicized. SC Resnet is not compatible with OOD methods that use logits.

	FPR@95 (Lower is Better)			AUROC (Higher is Better)		
	CE DinoV2	CE Resnet	SC Resnet	CE DinoV2	CE Resnet	SC Resnet
Method						
PT KNN	79.7 / **0.9**	79.7 / **0.9**	79.7 / **0.9**	65.1 / **99.6**	65.1 / **99.6**	65.1 / **99.6**
MSP	65.4 / 43.0	61.8 / 38.9	NA	75.1 / 82.0	78.3 / 87.4	NA
Energy	65.0 / 37.3	65.3 / 41.4	NA	75.3 / 85.6	78.0 / 87.6	NA
Mahalanobis	***62.5*** / 18.5	**59.9** / 16.2	62.3 / 34.7	***75.9*** / 93.4	**78.4** / 94.4	**78.9** / 87.6
Scale	65.0 / 37.3	65.3 / 41.4	NA	75.3 / 85.6	78.0 / 87.6	NA
NCI	66.7 / 35.3	74.5 / 36.1	NA	74.1 / 86.6	73.3 / 88.5	NA
KNN	61.9 / 25.4	64.4 / 25.8	**61.5** / 32.9	75.8 / 91.0	76.1 / 91.1	78.0 / 87.8
ReAct	64.2 / 36.4	71.9 / 47.7	NA	75.9 / 86.3	74.4 / 84.9	NA
DF + KNN	65.2 / 3.2	64.3 / ***2.5***	63.8 / ***3.2***	73.9 / ***99.0***	75.8 / ***99.2***	76.2 / ***99.0***
DF + ReAct	64.3 / 3.7	72.3 / 4.1	NA	75.9 / ***99.0***	74.4 / ***99.0***	NA

Key findings from our experiments include:

1. **Domain feature collapse is real**: All methods show significantly worse performance on

out-of-domain OOD detection compared to in-domain detection, confirming our theoretical predictions.

2. **Domain filtering is effective**: Adding domain filtering consistently reduces FPR@95 for out-of-domain OOD detection by substantial margins, sometimes reducing error rates from over 40% to under 5%.

3. **Minimal impact on in-domain performance**: Domain filtering maintains comparable performance on in-domain OOD detection, showing that the solution does not compromise the primary OOD detection capability.

4. **Generalizability across methods**: The improvement from domain filtering is consistent across different OOD detection methods and model architectures.

For example, on the Colon dataset, ReAct achieves the best in-domain performance but suffers from extremely high FPR@95 of 61% on out-of-domain OOD samples. Adding domain filtering reduces this FPR rate to 0.7%, effectively eliminating the problem of out-of-domain OOD detection while maintaining the strong in-domain performance.

5.5.4 Detailed Results by Dataset

To provide additional insight into the performance characteristics across different single-domain datasets, Table 5.2 shows FPR@95 results for a representative subset of Domain Bench datasets. This table highlights the variability in both in-domain and out-of-domain performance across different application domains.

The results in Table 5.2 demonstrate several important patterns. First, the effectiveness of domain filtering varies across datasets, with some domains (like Colon and Tissue) showing near-perfect out-of-domain detection after filtering, while others (like Rock) remain more challenging. Second, the choice of base OOD detection method significantly impacts in-domain performance, with methods like ReAct and Mahalanobis often achieving the best in-domain results before domain filtering is applied.

Table 5.2: Summary FPR@95 OOD Performance Across Selected ID Datasets Reported As (In-Domain OOD Score)/(Out-of-Domain OOD Score). Best scores are in bold and second best are bold and italicized. Domain filtering methods are italicized.

method	Colon	Eurosat	Food	Garbage	Rock	Tissue
PT KNN	67.4 / **0.0**	69.1 / **0.3**	80.0 / **0.6**	87.2 / ***0.4***	91.9 / **6.6**	89.3 / **0.0**
MSP	59.1 / 53.0	**41.3** / 49.8	74.9 / 63.7	68.0 / 42.0	85.8 / 71.8	84.2 / 76.6
Energy	61.0 / 70.7	***42.5*** / 50.1	75.2 / 62.9	78.7 / 54.3	86.7 / 71.2	84.4 / 79.2
Mahalanobis	40.8 / 12.5	51.4 / 13.7	76.8 / 52.1	**59.8** / 13.9	***83.1*** / 44.2	91.4 / 3.8
Scale	61.0 / 70.7	***42.5*** / 50.1	75.2 / 62.9	78.7 / 54.3	86.7 / 71.2	84.4 / 79.2
NCI	74.5 / 24.8	72.7 / 57.1	80.4 / 65.4	74.2 / 31.4	75.9 / 64.0	84.5 / 35.7
KNN	40.0 / 13.2	48.4 / 31.3	**73.3** / 62.7	77.9 / 33.3	77.3 / 61.8	92.6 / 31.6
ReAct	**39.0** / 61.2	55.5 / 54.4	85.9 / 71.1	82.9 / 58.6	84.7 / 75.0	**81.7** / 48.0
DF + KNN	41.5 / ***0.2***	49.6 / ***1.5***	***73.5*** / 2.3	76.5 / 2.1	75.1 / 52.5	92.2 / ***0.4***
DF + ReAct	40.6 / 0.7	65.2 / 4.3	86.4 / ***2.2***	82.9 / ***1.8***	84.9 / 61.0	***81.9*** / 0.7
DF + MDS	***40.4*** / 6.9	51.4 / 10.4	76.6 / 39.1	**61.6** / 12.7	**82.0** / ***39.8***	91.3 / 0.9

5.5.5 Case Study: Colon Dataset

To illustrate the detailed impact of domain feature collapse and the effectiveness of domain filtering, Table 5.3 presents comprehensive results for the Colon dataset across all out-of-domain test sets.

The Colon dataset results in Table 5.3 provide several key insights. Most notably, the dramatic difference between in-domain (adjacent) OOD detection and out-of-domain OOD detection clearly demonstrates domain feature collapse. For instance, ReAct achieves excellent in-domain performance (39.0% FPR@95) but struggles significantly with out-of-domain samples, with FPR@95 ranging from 41.0% to 98.5% depending on the specific OOD dataset.

The domain filtering approach (DF + KNN) virtually eliminates this problem, achieving consistently low FPR@95 (0.1-0.3%) across all out-of-domain test sets while maintaining competitive in-domain performance. This consistency across diverse out-of-domain datasets (from natural images like CIFAR to medical images like Chest X-rays) demonstrates the robustness of the domain filtering approach.

Table 5.3: Detailed FPR@95 OOD Detection Performance for the Colon Dataset. Results show performance across different out-of-domain test sets. Domain filtering methods are italicized.

OOD Dataset Method	In Domain (Adjacent)	Chest	Cifar10	Cifar100	Mnist	Place365	Svhn	Texture	Tin
PT KNN	67.4	**0.1**	**0.0**	**0.0**	**0.0**	**0.0**	**0.1**	**0.0**	**0.0**
MSP	59.1	3.3	46.8	66.9	38.2	36.9	63.5	96.1	72.4
Energy	61.0	3.1	89.1	92.1	41.1	75.5	77.4	98.5	89.3
Mahalanobis	40.8	16.7	8.7	9.0	27.7	6.1	15.6	13.2	3.1
Scale	61.0	3.1	89.1	92.1	41.1	75.5	77.4	98.5	89.3
NCI	74.5	11.0	24.2	24.5	14.4	29.2	42.4	28.5	24.4
KNN	40.0	12.4	11.9	11.6	26.7	4.3	16.0	17.8	5.0
ReAct	**39.0**	41.0	62.6	64.2	45.7	52.4	77.2	74.2	72.3
DF + KNN	41.5	***0.2***	***0.2***	***0.2***	***0.3***	***0.1***	***0.2***	***0.3***	***0.1***
DF + ReAct	40.6	0.4	0.8	0.8	0.5	0.6	0.9	0.9	0.8
DF + MDS	***40.4***	8.9	4.6	4.7	15.5	3.3	8.6	7.3	1.8

5.5.6 Discussion

Rock Dataset as an Outlier

The Rock dataset (Hossain et al., 2021) represents an interesting case study in the limitations of domain filtering. Unlike other datasets in Domain Bench, the Rock dataset exhibits characteristics that challenge our single-domain assumptions. The dataset contains images ranging from close-up shots of rock patterns to rock formations in natural settings, and even includes what appears to be a marble countertop as a member of the marble class.

This diversity within the Rock dataset results in a higher domain threshold ($t_{\mathbf{d}} \approx 1.78$) compared to more homogeneous datasets like Colon ($t_{\mathbf{d}} \approx 0.47$) or Food ($t_{\mathbf{d}} \approx 1.08$). The effectiveness of domain filtering can be improved by adjusting the percentile threshold from $p = 0.99$ to $p = 0.98$, which reduces the FPR@95 from 52.5% to 27.9% for out-of-domain OOD detection. However, this adjustment comes at the cost of increased false positive rejections for in-domain data.

This case highlights the importance of carefully evaluating whether a dataset truly represents a narrow domain and whether outliers within the in-distribution data may have an outsized influence on the domain threshold calculation. The Rock dataset serves as a reminder that domain filtering

works best when the underlying assumption of domain homogeneity is satisfied.

5.6 Limitations and Future Work

While our theoretical analysis and experimental validation provide strong evidence for domain feature collapse and the effectiveness of domain filtering, several limitations should be acknowledged:

5.6.1 Assumptions and Scope

Our theoretical analysis relies on the assumption of perfect information bottleneck compression and the independence of domain and class features within the training distribution. In practice, these conditions may not be perfectly satisfied, though our experimental results suggest that the core phenomenon persists even under relaxed conditions.

The domain filtering solution requires access to a pretrained model that can effectively distinguish between domains. In some cases, this may not be readily available or may require additional computational resources.

5.6.2 Generalization to Other Domains

While Domain Bench covers a diverse range of single-domain applications, further validation across additional domains and modalities would strengthen the generalizability of our findings. Future work should explore domain feature collapse in other critical applications such as autonomous driving, financial fraud detection, and cybersecurity.

5.6.3 Alternative Solutions

Domain filtering represents one approach to addressing domain feature collapse. Future research should investigate alternative solutions, such as:

- **Multi-task learning**: Training models to explicitly predict both class and domain labels

- **Domain-aware architectures**: Developing neural network architectures that naturally preserve domain information
- **Regularization techniques**: Designing loss functions that encourage retention of domain features
- **Data augmentation**: Creating synthetic domain variations to make single-domain datasets more robust

5.7 Conclusion

This chapter has identified and theoretically characterized domain feature collapse, a critical failure mode in out-of-distribution detection that occurs specifically in single-domain settings. Through information-theoretic analysis, we proved that supervised learning models trained on single-domain data will inevitably discard domain-specific features in favor of class-specific features, leading to dangerous blind spots in OOD detection.

Our key contributions include:

1. **Theoretical foundation**: We provided the first formal analysis of domain feature collapse using information bottleneck theory, proving that this phenomenon is inevitable under standard supervised learning objectives in single-domain settings.
2. **Practical solution**: We introduced domain filtering, a simple yet effective two-stage approach that addresses domain feature collapse while maintaining strong in-domain OOD detection performance.
3. **Comprehensive evaluation**: Domain Bench provides a new benchmark specifically designed to evaluate OOD detection methods in single-domain settings, filling a critical gap in existing evaluation protocols.
4. **Empirical validation**: Our experiments across 11 diverse single-domain datasets confirm the theoretical predictions and demonstrate the effectiveness of the proposed solution.

The implications of this work extend beyond academic research to practical deployment of machine learning systems in safety-critical applications. Medical imaging, satellite monitoring, industrial inspection, and many other domains rely on single-domain datasets where domain feature collapse

poses real risks. Our findings suggest that practitioners in these domains should carefully consider the potential for domain feature collapse and implement appropriate mitigation strategies.

Furthermore, this work highlights the importance of evaluation protocols that accurately reflect real-world deployment scenarios. The widespread use of multi-domain benchmarks in OOD detection research has inadvertently masked this critical failure mode, emphasizing the need for more diverse and realistic evaluation frameworks.

As machine learning systems become increasingly deployed in specialized, narrow domains, understanding and addressing domain feature collapse will become ever more critical for ensuring the safety and reliability of these systems. The theoretical framework and practical solutions presented in this chapter provide a foundation for future research in this important area.

Chapter 6

Hallucinations Through the Lens of Mutual Information and Representation Learning

This chapter proposes to investigate hallucinations in large language models from the perspective of mutual information theory and representation learning. We hypothesize that intrinsic hallucinations arise from a loss of mutual information between the question and answer during the generation process, particularly in the intermediate layers of foundation models.

6.1 Introduction and Motivation

Hallucinations in large language models represent a fundamental challenge where models generate plausible-sounding but factually incorrect or unverifiable content. While existing work has focused primarily on detection and mitigation strategies, there remains a significant gap in our theoretical understanding of why hallucinations occur from an information-theoretic perspective.

Our central hypothesis is that intrinsic hallucinations can be understood as a breakdown in the mutual information flow between input queries and generated responses within the model's intermediate representations. This perspective offers several advantages:

- **Theoretical Foundation**: Provides a principled framework for understanding hallucination

mechanisms

- **Measurable Quantities**: Mutual information offers quantifiable metrics for hallucination analysis
- **Intervention Strategies**: Information-theoretic insights can guide targeted mitigation approaches
- **Generalizability**: Framework applies across different model architectures and domains

6.2 Theoretical Framework

6.2.1 Mutual Information in Language Generation

We formalize the language generation process in terms of mutual information between input queries $\mathbf{x}$ and generated responses $\mathbf{y}$. For a well-calibrated model, we expect high mutual information $I(\mathbf{x}; \mathbf{y})$, indicating that the response contains substantial information about the input query.

Definition 6.2.1. Information-Preserving Generation: A language model exhibits information-preserving generation when the mutual information between input $\mathbf{x}$ and output $\mathbf{y}$ satisfies:

$$I(\mathbf{x}; \mathbf{y}) \geq \tau$$

for some threshold $\tau > 0$ that depends on the task complexity and expected response informativeness.

6.2.2 Hallucination as Information Loss

We propose that intrinsic hallucinations occur when there is insufficient mutual information between the input query and the generated response, particularly in the model's intermediate representations.

Definition 6.2.2. Information-Theoretic Hallucination: An intrinsic hallucination occurs when the mutual information between input $\mathbf{x}$ and output $\mathbf{y}$ falls below a critical threshold:

$$I(\mathbf{x}; \mathbf{y}) < \tau_{critical}$$

where $\tau_{critical}$ represents the minimum information required for factually grounded generation.

6.3 Proposed Research Methodology

6.3.1 Mutual Information Estimation in Foundation Models

Estimating mutual information in the high-dimensional intermediate representations of foundation models presents significant theoretical and computational challenges. We investigate several complementary approaches, each offering distinct advantages and limitations for understanding information flow in transformer architectures.

Neural Mutual Information Estimation

The Mutual Information Neural Estimation (MINE) framework (Belghazi et al., 2018) represents a significant advancement in MI estimation for high-dimensional data. MINE leverages the Donsker-Varadhan representation of the KL divergence to provide a tractable lower bound on mutual information through neural network optimization.

The method works by training a neural network T_θ to distinguish between samples from the joint distribution $p(x, y)$ and the product of marginals $p(x)p(y)$. The MI estimate is obtained as:

$$\hat{I}_{\text{MINE}}(X;Y) = \sup_{\theta} \mathbb{E}_{p(x,y)}[T_\theta(x,y)] - \log \mathbb{E}_{p(x)p(y)}[e^{T_\theta(x,y)}]$$

For our application to transformer layers, MINE offers several compelling advantages. The method scales naturally to the high-dimensional hidden states typical in modern language models, often ranging from 768 to several thousand dimensions. The differentiable nature of the estimation process allows for end-to-end optimization and integration with existing training pipelines. Furthermore, MINE can handle the complex, non-linear dependencies that characterize the relationship between different transformer layers.

However, MINE also presents notable challenges for our specific use case. The method is known to suffer from estimation bias, particularly when the true mutual information is high, which may be the case for adjacent transformer layers. The computational overhead can be substantial, requiring additional forward passes through the discriminator network during training. Additionally, MINE's performance is highly sensitive to hyperparameter choices, including the architecture of the discriminator network, learning rates, and batch sizes, necessitating careful tuning for each model architecture and scale.

Variational Bounds

InfoNCE (Information Noise Contrastive Estimation) (van den Oord et al., 2018) provides an alternative approach to MI estimation through contrastive learning principles. This method estimates a lower bound on mutual information by maximizing the agreement between positive pairs while minimizing agreement with negative samples.

The InfoNCE objective can be expressed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{f(x,y)}{\sum_{y' \in \mathcal{N}} f(x,y')}\right]$$

where $f(x,y)$ represents a learned similarity function and $\mathcal{N}$ denotes the set of negative samples.

InfoNCE demonstrates particular strength in providing stable training dynamics, making it well-suited for the iterative optimization required in our layer-wise analysis. The method benefits from well-established theoretical properties, including proven convergence guarantees under certain conditions. The contrastive framework naturally aligns with our goal of understanding how information about question-answer pairs is preserved or lost across transformer layers.

The primary limitation of InfoNCE lies in its provision of only a lower bound on the true mutual information, which may underestimate the actual information content in cases where the bound is loose. The quality of the MI estimate is critically dependent on the negative sampling strategy, requiring careful consideration of how to select informative negative examples that provide meaningful contrast without introducing bias. In the context of transformer layers, this translates to decisions about which layer representations to use as negatives and how to ensure they provide sufficient diversity for accurate estimation.

Kernel-Based Methods

Kernel density estimation approaches offer a non-parametric alternative for MI estimation that makes minimal assumptions about the underlying data distribution. These methods estimate the probability densities $p(x)$, $p(y)$, and $p(x,y)$ using kernel functions, then compute mutual information through numerical integration.

The kernel-based MI estimate takes the form:

$$\hat{I}_{\text{kernel}}(X;Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

where each density is estimated using kernel methods such as Gaussian kernels with adaptive bandwidth selection.

The theoretical foundation of kernel methods provides strong guarantees about estimation consistency and convergence properties. Unlike neural approaches, kernel methods do not require distributional assumptions about the data, making them particularly robust for the diverse range of representations that emerge across different transformer layers and model architectures. The non-parametric nature ensures that the method can capture complex, multimodal distributions that may characterize the relationship between layer representations.

However, kernel-based approaches face significant practical limitations when applied to high-dimensional transformer representations. The curse of dimensionality severely impacts both the accuracy and computational feasibility of density estimation in spaces with hundreds or thousands of dimensions. The computational complexity grows exponentially with dimensionality, making direct application to full transformer hidden states computationally prohibitive. Additionally, the choice of kernel bandwidth becomes increasingly critical and difficult to optimize in high-dimensional spaces, often requiring problem-specific tuning that may not generalize across different model architectures.

Discrete Approximations

Quantization-based approaches provide an alternative pathway to MI estimation by discretizing continuous representations and computing empirical mutual information on the resulting discrete distributions. This method involves partitioning the continuous space of layer representations into discrete bins and estimating MI using the standard discrete formula.

The discrete MI estimate is computed as:

$$\hat{I}_{\text{discrete}}(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where the probabilities are estimated from the empirical frequencies in the discretized space.

Discrete approximation methods offer the significant advantage of enabling exact computation of mutual information once the discretization is established, eliminating the approximation errors inherent in other estimation approaches. The resulting estimates are highly interpretable, allowing for direct analysis of which discrete states contribute most to the mutual information between layers. This interpretability can provide valuable insights into the specific types of information that are preserved or lost during the forward pass through transformer layers.

The primary challenge with discrete approximation lies in the information loss introduced by the quantization process itself. The choice of discretization scheme—including the number of bins, binning strategy, and handling of outliers—can significantly impact the quality of the MI estimate. Too few bins may fail to capture important distributional structure, while too many bins can lead to sparse empirical distributions and unreliable probability estimates. Furthermore, the optimal discretization strategy may vary across different layers and model architectures, requiring careful validation and potentially limiting the generalizability of findings across different experimental settings.

Contrastive Mutual Information Estimation (Proposed)

We propose a novel contrastive learning approach specifically designed for estimating mutual information between intermediate layer representations in language models during question-answering tasks. This method addresses the unique challenges of analyzing information flow in transformer architectures while providing interpretable insights into how question-answer relationships are preserved across model layers.

Method Overview and Motivation Our approach builds on the fundamental insight that mutual information between two random variables can be understood through their ability to predict each other. In the context of transformer layers processing question-answer pairs, we hypothesize that layers with high mutual information should contain representations that maintain consistent relationships for the same QA pair while exhibiting distinct patterns for different QA pairs.

Given two layers l_i and l_j in a transformer model, our method learns contrastive representations that maximize similarity for the same question-answer pair across these layers while minimizing similarity for different QA pairs. This approach directly targets the preservation of QA-specific information, making it particularly well-suited for understanding hallucination mechanisms where the loss of question-answer coherence is a primary concern.

Formal Mathematical Framework For a question-answer pair $(\mathbf{x}, \mathbf{y})$, let $\mathbf{z}_{l_i}$ and $\mathbf{z}_{l_j}$ represent the hidden states at layers l_i and l_j respectively. We learn projection functions $f_i : \mathbf{z}_{l_i} \to \mathbb{R}^d$ and $f_j : \mathbf{z}_{l_j} \to \mathbb{R}^d$ that map layer representations to a common embedding space where similarity can be meaningfully compared.

The contrastive learning objective is formulated as:

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

Here, $\mathbf{z}_{l_j}^{(k)}$ represents representations from different QA pairs serving as negative examples, $\text{sim}(\cdot, \cdot)$ denotes a similarity function such as cosine similarity, and τ is a temperature parameter that controls the sharpness of the distribution. The temperature parameter plays a crucial role in balancing between hard and soft assignments, with lower values creating sharper distinctions between positive and negative pairs.

The mutual information between layers is then estimated through the learned contrastive representations:

$$\hat{I}(\mathbf{z}_{l_i}; \mathbf{z}_{l_j}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})} \left[\log \frac{\exp(\text{sim}(f_i(\mathbf{z}_{l_i}), f_j(\mathbf{z}_{l_j}))/\tau)}{\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}[\exp(\text{sim}(f_i(\mathbf{z}_{l_i}), f_j(\mathbf{z}'_{l_j}))/\tau)]} \right]$$

This formulation provides a principled connection between contrastive learning objectives and mutual information estimation, grounded in the theoretical framework established by Poole et al. (2019).

Advantages and Theoretical Justification The proposed contrastive approach offers several significant advantages for analyzing information flow in language models. Most importantly, it directly optimizes for question-answer pair consistency across layers, ensuring that the MI estimate captures the specific type of information most relevant to hallucination analysis. This task-specific focus distinguishes our method from general-purpose MI estimators that may not prioritize the preservation of QA relationships.

The method demonstrates excellent scalability to large transformer models, as the contrastive learning framework can efficiently handle the high-dimensional representations typical in modern language models. Unlike kernel-based methods that suffer from the curse of dimensionality, our approach leverages learned projections to map representations to manageable embedding spaces while preserving the essential relational structure.

The contrastive framework provides interpretable similarity scores that can be analyzed to understand which types of information are preserved or lost between layers. These scores offer direct insights into the mechanisms of information degradation that may lead to hallucinations, enabling both detection and potential intervention strategies.

Additionally, our method can detect layer-specific information degradation patterns, identifying particular layers where QA consistency begins to break down. This capability is crucial for understanding the temporal dynamics of hallucination emergence during the forward pass through transformer layers.

The approach naturally handles variable sequence lengths common in question-answering tasks, as the projection functions can accommodate different input dimensions through appropriate pooling strategies.

Limitations and Challenges Despite its advantages, the contrastive MI estimation method faces several important limitations that must be carefully addressed in implementation. The quality of MI estimates is critically dependent on the negative sampling strategy, requiring thoughtful selection of negative examples that provide meaningful contrast without introducing systematic bias. Poor negative sampling can lead to either overestimation (if negatives are too easy) or underestimation (if negatives are too similar to positives) of the true mutual information.

The method exhibits sensitivity to the architecture and initialization of projection functions f_i and f_j. The choice of projection dimensionality, activation functions, and regularization strategies can significantly impact the quality of the learned representations and, consequently, the accuracy of MI estimates. This sensitivity necessitates careful hyperparameter tuning and validation across different model architectures.

While our method captures important aspects of mutual information related to QA consistency, it may not capture all forms of mutual information present between layer representations. The contrastive objective focuses specifically on preserving QA relationships, potentially missing other types of information dependencies that contribute to the overall mutual information between layers.

The computational overhead can become substantial for large batch sizes, as the method requires computing similarities between all positive pairs and their corresponding negative sets. This scaling challenge may limit the practical applicability to very large datasets or real-time applications without careful optimization.

Implementation Considerations Successful implementation of the contrastive MI estimation method requires careful attention to several technical details. We employ pooled representations such as CLS tokens or mean pooling to obtain fixed-size embeddings from variable-length sequences, ensuring consistent input dimensions for the projection functions while preserving the essential

semantic content.

Hard negative mining strategies can significantly improve the quality of contrastive learning by focusing on the most informative negative examples. This involves selecting negative samples that are semantically similar but factually distinct from the positive pairs, providing stronger learning signals for the contrastive objective.

Layer normalization applied before the projection functions helps stabilize training and ensures that the learned similarities are not dominated by magnitude differences between layer representations. This normalization is particularly important when comparing representations from layers at different depths, which may have different activation scales.

Momentum-based updates for the projection functions can provide more stable training dynamics, particularly important when dealing with the high-dimensional and potentially noisy representations typical in large language models. This approach helps prevent oscillations and ensures convergent learning of the similarity functions.

6.3.2 Representation Learning Analysis

We will analyze how different representation learning objectives affect the mutual information flow and hallucination propensity:

Layer-wise Information Flow

Investigate how mutual information $I(\mathbf{x}; \mathbf{z}_l)$ evolves across layers l in transformer architectures, where $\mathbf{z}_l$ represents the hidden state at layer l.

Attention Mechanism Analysis

Examine the role of attention patterns in preserving or degrading mutual information between input and intermediate representations.

Information Bottleneck Dynamics

Study how the information bottleneck principle applies to language generation and its relationship to hallucination emergence.

6.4 Experimental Design

Our experimental design follows a systematic approach to validate the proposed contrastive mutual information estimation method and its effectiveness for hallucination detection. The experiments are structured to address three primary research questions: (1) How accurately can our contrastive method estimate mutual information compared to existing approaches? (2) What is the relationship between MI estimates and hallucination occurrence in large language models? (3) How effectively can MI-based metrics detect hallucinations in real-world scenarios?

6.4.1 Datasets and Benchmarks

We employ a diverse collection of datasets spanning factual question answering, hallucination detection, and synthetic validation scenarios. Each dataset serves specific purposes in our experimental pipeline, from method validation to real-world performance assessment.

Factual Question Answering Datasets

Natural Questions The Natural Questions dataset (Kwiatkowski et al., 2019) provides a large-scale collection of real user questions paired with Wikipedia articles containing answers. We utilize the open-domain variant, which contains over 300,000 question-answer pairs derived from actual Google search queries. For our experiments, we construct training and test sets of 50,000 and 15,000 samples respectively, ensuring sufficient statistical power for reliable MI estimation and hallucination detection evaluation.

The dataset's strength lies in its naturalistic question formulation, reflecting the types of queries users actually pose to search engines. This authenticity makes it particularly valuable for evaluating hallucination detection in realistic scenarios. We preprocess the data to extract clean question-answer pairs, filtering out questions with ambiguous or incomplete answers to ensure clear ground

truth for hallucination assessment.

TriviaQA TriviaQA (Joshi et al., 2017) offers a complementary perspective with its focus on trivia questions paired with evidence documents from Wikipedia and web sources. The dataset contains approximately 95,000 question-answer pairs, from which we sample 40,000 for training and 12,000 for testing, maintaining the minimum 10,000 sample requirement for robust evaluation.

The trivia format provides questions with well-defined factual answers, making it ideal for studying hallucinations where models generate plausible but incorrect information. The availability of evidence documents allows us to distinguish between cases where models hallucinate due to lack of knowledge versus cases where they fail to properly utilize available information.

WebQuestions WebQuestions (Berant et al., 2013) focuses on questions answerable from the Freebase knowledge base, providing a structured approach to factual question answering. With approximately 6,000 total questions, we use the entire dataset for testing while supplementing with synthetic variations to reach our minimum sample requirements.

This dataset is particularly valuable for studying hallucinations in structured knowledge domains, where the ground truth can be precisely verified against the knowledge base. The questions often require multi-hop reasoning, making them suitable for analyzing how information flows through multiple transformer layers.

Hallucination-Specific Benchmarks

HaluEval HaluEval (Li et al., 2023) represents the most comprehensive benchmark specifically designed for hallucination evaluation in large language models. The dataset encompasses multiple task types including question answering, dialogue, summarization, and text completion, with over 35,000 samples across all categories.

We focus primarily on the question-answering subset, which contains approximately 10,000 samples with carefully annotated hallucination labels. The dataset provides both binary hallucination labels and fine-grained categorizations of hallucination types, enabling detailed analysis of how our MI-based detection method performs across different hallucination categories.

The benchmark's strength lies in its systematic construction methodology, where hallucinations are

generated through controlled perturbations of factual content. This approach ensures a balanced distribution of hallucinated and non-hallucinated responses, crucial for training and evaluating detection systems.

TruthfulQA TruthfulQA (Lin et al., 2021) presents a unique challenge by focusing on questions designed to elicit false beliefs and misconceptions commonly held by humans. The dataset contains 817 questions across 38 categories, covering topics where models might generate plausible but incorrect answers based on common misconceptions.

While smaller than our preferred minimum of 10,000 samples, TruthfulQA provides invaluable insights into a specific type of hallucination where models reproduce human biases and false beliefs. We augment this dataset with paraphrased versions and related questions to increase the sample size while maintaining the essential characteristics of the original benchmark.

The dataset is particularly relevant for studying intrinsic hallucinations, where models generate content that contradicts established facts due to biases in training data or reasoning failures rather than simple knowledge gaps.

FEVER The Fact Extraction and VERification (FEVER) dataset (Thorne et al., 2018) provides a large-scale benchmark for fact-checking with over 185,000 claims paired with evidence from Wikipedia. We adapt FEVER for hallucination detection by treating unverifiable or contradicted claims as hallucinations and supported claims as factual content.

From the full dataset, we construct training and test sets of 80,000 and 20,000 samples respectively, ensuring robust statistical evaluation. The dataset's three-way classification (supported, refuted, not enough info) provides nuanced ground truth labels that allow for detailed analysis of different types of factual errors.

FEVER's strength lies in its systematic evidence-based verification process, providing clear criteria for distinguishing between factual and hallucinated content. The dataset's scale and rigorous annotation make it ideal for training and evaluating our contrastive MI estimation method.

Synthetic Validation Datasets

To validate the accuracy of our contrastive MI estimation method, we construct synthetic datasets where the ground truth mutual information can be computed analytically. These datasets serve as crucial benchmarks for method validation before application to real-world scenarios.

Gaussian Mixture Models We generate synthetic question-answer representations using Gaussian mixture models with known covariance structures. By controlling the overlap between mixture components, we can precisely control the mutual information between synthetic "layer" representations. These datasets range from 10,000 to 100,000 samples, allowing us to study the convergence properties of our estimation method.

Transformer-Based Synthetic Data We create synthetic datasets by extracting representations from small, controlled transformer models where we can compute or approximate the true mutual information through exhaustive sampling. These datasets provide more realistic validation scenarios while maintaining computational tractability for ground truth estimation.

6.4.2 Model Analysis

Our model analysis encompasses a comprehensive evaluation across different architectures, scales, and training paradigms to understand how mutual information dynamics vary across the landscape of modern language models.

Architecture Comparison

Transformer-Based Models We conduct extensive analysis across the transformer family, including both encoder-only and decoder-only architectures. The GPT family (GPT-2, GPT-3.5, GPT-4) serves as our primary focus for decoder-only models, given their widespread use in question-answering applications. We analyze models ranging from GPT-2 small (124M parameters) to GPT-3.5 (175B parameters), providing insights into how architectural scale affects information flow patterns.

For encoder-only models, we examine BERT variants including BERT-base (110M parameters), BERT-large (340M parameters), and RoBERTa in multiple sizes. These models provide comple-

mentary insights into bidirectional information processing, particularly relevant for understanding how question and answer information interact during encoding.

Encoder-decoder models such as T5 (ranging from T5-small to T5-11B) and BART offer additional perspectives on information flow in sequence-to-sequence architectures. These models are particularly valuable for studying how information is transferred from encoder representations to decoder states during answer generation.

State-Space Models Recent advances in state-space models, particularly Mamba and Structured State Space (S4) models, provide alternative architectures for sequence modeling that may exhibit different information flow characteristics compared to attention-based transformers. We analyze Mamba models in the 130M to 2.8B parameter range to understand how the selective state-space mechanism affects mutual information preservation.

The linear scaling properties of state-space models with sequence length make them particularly interesting for studying long-context question-answering scenarios where traditional transformers face computational limitations. Our analysis focuses on how information degrades over long sequences and whether the state-space mechanism provides better information preservation than attention mechanisms.

Hybrid Architectures We examine hybrid models that combine transformer attention with alternative mechanisms, such as retrieval-augmented generation (RAG) models and models incorporating external memory systems. These architectures provide insights into how external information sources affect internal information flow and hallucination patterns.

Mixture-of-experts (MoE) models represent another important hybrid category, where different expert networks may specialize in different types of information processing. We analyze how expert routing decisions correlate with mutual information patterns and hallucination emergence.

Scale Analysis

Parameter Count Effects We systematically investigate how model scale affects mutual information preservation and hallucination rates across parameter counts ranging from 100M to 175B parameters. This analysis reveals scaling laws for information flow, identifying whether larger models consistently preserve more mutual information between questions and answers or whether there

are optimal scales for different types of reasoning tasks.

The scale analysis includes both dense and sparse models, examining how parameter efficiency techniques such as pruning and quantization affect information flow patterns. We particularly focus on whether compressed models exhibit different hallucination characteristics due to altered information processing capabilities.

Training Data Scale Beyond parameter count, we analyze how training data scale affects information flow characteristics. Using models trained on datasets ranging from 1B to 1T tokens, we investigate whether exposure to more diverse training data improves information preservation or introduces additional sources of hallucination through conflicting information.

This analysis includes examination of domain-specific fine-tuning effects, studying how adaptation to particular domains (medical, legal, scientific) affects the mutual information patterns and hallucination rates in those domains versus general knowledge areas.

Context Length Analysis We conduct specialized experiments examining how context length affects information flow and hallucination patterns. Using models with varying context windows (from 512 to 32,768 tokens), we analyze how information degrades over long sequences and whether longer contexts provide better grounding for factual accuracy or introduce additional opportunities for hallucination.

6.4.3 Evaluation Metrics and Protocols

Our evaluation framework employs multiple complementary metrics to provide comprehensive assessment of both mutual information estimation accuracy and hallucination detection performance.

Mutual Information Estimation Metrics

Synthetic Data Validation For synthetic datasets where ground truth mutual information is known, we employ mean squared error (MSE) and mean absolute error (MAE) between estimated and true MI values. We also compute correlation coefficients to assess the ranking consistency of our estimates across different MI regimes.

Bias and variance decomposition provides insights into the systematic errors and estimation uncertainty of our contrastive method compared to baseline approaches. We conduct bootstrap sampling to estimate confidence intervals for MI estimates, ensuring robust statistical evaluation.

Cross-Method Consistency When ground truth MI is unavailable, we assess consistency across different estimation methods (MINE, InfoNCE, kernel-based, and our contrastive approach). High correlation between methods provides confidence in the reliability of estimates, while systematic differences reveal method-specific biases that must be accounted for in interpretation.

Hallucination Detection Metrics

Area Under the Receiver Operating Characteristic Curve (AUROC) AUROC serves as our primary metric for evaluating hallucination detection performance, providing a threshold-independent measure of discriminative ability. We compute AUROC scores for each dataset and model combination, with scores above 0.85 considered indicative of strong detection performance.

The AUROC metric is particularly valuable because it captures the trade-off between true positive and false positive rates across all possible decision thresholds. This comprehensive view is essential for understanding the practical utility of our MI-based detection approach across different deployment scenarios with varying tolerance for false alarms.

False Positive Rate at 95% True Positive Rate (FPR95) FPR95 provides a practically oriented metric that reflects real-world deployment constraints where high recall (95% true positive rate) is essential for safety-critical applications. This metric directly addresses the question: "If we want to catch 95% of all hallucinations, what percentage of non-hallucinated content will be incorrectly flagged?"

FPR95 is particularly relevant for applications where missing hallucinations carries high cost, such as medical or legal question-answering systems. We target FPR95 values below 20% as indicative of practical utility, though the specific threshold may vary by application domain.

Precision-Recall Analysis We conduct comprehensive precision-recall analysis to understand performance across different operating points. This analysis is particularly important for understanding how our method performs when hallucinations are rare (low base rate scenarios) versus

common (high base rate scenarios).

Area under the precision-recall curve (AUPR) provides a summary metric that is less sensitive to class imbalance than AUROC, making it valuable for datasets where hallucinations represent a small fraction of total samples.

Statistical Significance and Robustness

Cross-Validation and Bootstrap Sampling All experiments employ 5-fold cross-validation to ensure robust performance estimates and reduce dependence on particular train-test splits. Bootstrap sampling with 1,000 iterations provides confidence intervals for all reported metrics, enabling statistical significance testing between different methods and conditions.

Multiple Random Seeds We conduct all experiments across multiple random seeds (minimum 5 per condition) to account for initialization variability and ensure reproducible results. This is particularly important for contrastive learning methods, which can be sensitive to initialization and negative sampling randomness.

Ablation Studies Systematic ablation studies isolate the contribution of different components in our contrastive MI estimation method. These studies examine the effects of projection function architecture, temperature parameter values, negative sampling strategies, and pooling methods on both MI estimation accuracy and hallucination detection performance.

Contrastive MI Estimation Validation

Our proposed contrastive mutual information estimation method requires comprehensive validation to establish its accuracy, reliability, and practical utility for hallucination detection. This validation encompasses both theoretical verification on synthetic data and empirical assessment on real-world language model representations.

Synthetic Validation Protocol We conduct extensive validation using synthetic datasets where ground truth mutual information can be computed analytically or through exhaustive sampling.

The synthetic validation employs multiple data generation strategies to ensure robustness across different distributional assumptions.

Gaussian mixture models with controlled covariance structures provide the foundation for our synthetic validation. We generate pairs of random variables with mutual information values ranging from 0 (independence) to high dependence scenarios, testing our method's accuracy across this full spectrum. Each synthetic dataset contains at least 50,000 samples to ensure stable MI estimates and reliable assessment of method performance.

Non-Gaussian synthetic data tests the robustness of our method beyond standard distributional assumptions. We employ heavy-tailed distributions, multimodal distributions, and discrete-continuous mixtures to evaluate performance under realistic conditions that may arise in transformer representations.

The validation protocol includes systematic variation of key parameters including dimensionality (from 10 to 1,000 dimensions), sample size (from 1,000 to 100,000 samples), and noise levels to understand the operating characteristics of our method across different experimental conditions.

Cross-Method Comparison Framework We implement a comprehensive comparison framework that evaluates our contrastive method against established MI estimation approaches including MINE, InfoNCE, kernel density estimation, and discrete approximation methods. This comparison employs identical datasets and evaluation protocols to ensure fair assessment.

The comparison framework evaluates multiple performance dimensions including estimation accuracy (bias and variance), computational efficiency (time and memory requirements), and robustness to hyperparameter choices. We conduct systematic hyperparameter sweeps for all methods to ensure optimal performance in the comparison.

Statistical significance testing using paired t-tests and Wilcoxon signed-rank tests provides rigorous assessment of performance differences between methods. Effect size calculations complement significance tests to evaluate the practical importance of observed differences.

Layer Consistency Analysis A critical component of our validation examines how MI estimates change across transformer layers and their correlation with hallucination emergence patterns. This analysis employs layer-wise extraction of representations from multiple transformer models, computing MI estimates between all pairs of layers.

We analyze both adjacent layer pairs (consecutive layers) and distant layer pairs (layers separated by multiple intermediate layers) to understand how information flows and degrades through the transformer architecture. This analysis reveals critical layers where information loss occurs and identifies potential intervention points for hallucination mitigation.

Correlation analysis between layer-wise MI estimates and empirically observed hallucination rates provides direct validation of our theoretical framework linking information loss to hallucination emergence. We employ both Pearson and Spearman correlation coefficients to capture linear and monotonic relationships.

Comprehensive Ablation Studies Our ablation studies systematically isolate the contribution of different design choices in the contrastive MI estimation method, providing insights into optimal configurations and robustness to hyperparameter variations.

Negative sampling strategy ablation compares random negative sampling, hard negative mining, and stratified sampling approaches. Hard negative mining selects the most challenging negative examples that are semantically similar but factually distinct from positive pairs, potentially providing stronger learning signals for the contrastive objective.

Projection function architecture ablation examines linear projections, multi-layer perceptrons with varying depths and widths, and residual architectures. We evaluate how architectural complexity affects both MI estimation accuracy and computational efficiency, identifying optimal trade-offs for different application scenarios.

Temperature parameter sensitivity analysis systematically varies the temperature parameter τ from 0.01 to 10.0, examining its effect on both training dynamics and final MI estimation quality. This analysis reveals optimal temperature ranges and assesses the robustness of our method to this critical hyperparameter.

Pooling strategy comparison evaluates different approaches for converting variable-length sequences to fixed-size representations, including mean pooling, max pooling, attention-weighted pooling, and CLS token extraction. This analysis is crucial for understanding how sequence-level information is preserved in the contrastive learning process.

Computational Efficiency Benchmarking We conduct systematic benchmarking of computational costs compared to other MI estimation methods, measuring both training time and inference

time across different model scales and dataset sizes. This benchmarking employs standardized hardware configurations and implementation optimizations to ensure fair comparison.

Memory usage analysis examines the scalability of our method to large transformer models and datasets, identifying potential bottlenecks and optimization opportunities. We analyze both peak memory usage during training and steady-state memory requirements during inference.

Scalability analysis examines how computational costs grow with key problem dimensions including sequence length, batch size, model size, and dataset size. This analysis informs practical deployment considerations and identifies parameter regimes where our method remains computationally feasible.

Hallucination Correlation Validation The ultimate validation of our method lies in its ability to predict hallucination occurrence through MI estimates. We conduct extensive correlation analysis between contrastive MI estimates and empirically observed hallucination rates across multiple datasets and model architectures.

This validation employs both aggregate correlation analysis (correlation between average MI and hallucination rates across different conditions) and instance-level analysis (correlation between individual MI estimates and hallucination labels for specific question-answer pairs).

Temporal analysis examines how the correlation between MI estimates and hallucination rates evolves during model training, providing insights into the development of information processing capabilities and potential early stopping criteria for hallucination-aware training.

6.5 Expected Contributions

6.5.1 Theoretical Contributions

1. **Information-Theoretic Framework**: Formal characterization of hallucinations through mutual information theory
2. **Representation Learning Theory**: Understanding of how different learning objectives affect information preservation
3. **Critical Thresholds**: Identification of information-theoretic thresholds for hallucination emergence

6.5.2 Empirical Contributions

1. **Measurement Methodology**: Practical approaches for estimating mutual information in large language models, including our novel contrastive MI estimation method
2. **Hallucination Prediction**: Early detection of hallucinations through information-theoretic metrics and layer-wise consistency analysis
3. **Intervention Strategies**: Information-guided approaches for reducing hallucination rates
4. **Contrastive MI Validation**: Empirical validation of the proposed contrastive learning approach for MI estimation across different model architectures and scales

6.5.3 Practical Applications

1. **Model Design**: Architectural modifications to preserve information flow
2. **Training Objectives**: Information-theoretic regularization for hallucination reduction
3. **Inference-Time Detection**: Real-time hallucination detection using MI estimates

6.6 Challenges and Limitations

6.6.1 Technical Challenges

- **High-Dimensional MI Estimation**: Accurate estimation in transformer hidden spaces
- **Computational Complexity**: Scalability to large models and datasets
- **Ground Truth Definition**: Establishing reliable hallucination labels

6.6.2 Theoretical Limitations

- **Causality vs. Correlation**: Distinguishing causal relationships from correlations
- **Task Dependence**: Generalizability across different types of generation tasks
- **Model Specificity**: Applicability to different architectural paradigms

6.7 Related Work and Positioning

This work builds upon and extends several research directions, positioning itself at the intersection of information theory, representation learning, and hallucination detection in large language models.

6.7.1 Information Theory in Natural Language Processing

The application of information theory to natural language processing has a rich history, with recent advances making it increasingly relevant for understanding modern language models.

Classical Information-Theoretic Approaches

Early work by Shannon (1948) established the mathematical foundations that continue to influence NLP research. Cover & Thomas (1999) provided comprehensive theoretical frameworks that have been adapted for linguistic analysis. Classical applications include language modeling perplexity measures, which are fundamentally based on cross-entropy and information content.

Mutual Information in Representation Learning

The use of mutual information for representation learning has gained significant traction. Linsker (1988) introduced InfoMax principles that maximize mutual information between inputs and learned representations. This was later extended by Hjelm et al. (2019) with Deep InfoMax (DIM), which applies MI maximization to deep neural networks.

van den Oord et al. (2018) developed Contrastive Predictive Coding (CPC), which uses contrastive learning to estimate mutual information between different parts of a sequence. This work is particularly relevant to our proposed contrastive MI estimation method, though we extend it specifically to question-answering contexts and layer-wise analysis.

Information Bottleneck Theory

The Information Bottleneck principle (Tishby et al., 2000) provides a theoretical framework for understanding representation learning as a trade-off between compression and prediction. Alemi

et al. (2017) extended this to deep learning with the Deep Variational Information Bottleneck. Shwartz-Ziv & Tishby (2017) applied information bottleneck theory to understand deep neural networks, though their work has been subject to debate (Saxe et al., 2019).

Recent work by Federici et al. (2020) explores multi-view information bottleneck for robust representations, while Shwartz-Ziv & LeCun (2023) provides a comprehensive review of compression and information theory in self-supervised learning.

6.7.2 Hallucination Detection and Mitigation

Hallucination detection in large language models has emerged as a critical research area with diverse methodological approaches.

Confidence-Based Methods

Early approaches focused on using model confidence as a proxy for factual accuracy. Manakul et al. (2023) developed SelfCheckGPT, which uses consistency across multiple model generations to detect hallucinations. Zhang et al. (2023a) introduced SIRENS, which leverages uncertainty estimation for hallucination detection.

Farquhar et al. (2024) proposed using semantic entropy to detect hallucinations, measuring uncertainty in the semantic content rather than token-level probabilities. This approach is conceptually related to our information-theoretic framework but focuses on output uncertainty rather than internal information flow.

Consistency-Based Approaches

Several methods exploit consistency across different model behaviors. Li et al. (2023) developed comprehensive evaluation frameworks that test consistency across various prompting strategies. Peng et al. (2023) introduced iterative fact-checking with external knowledge bases.

Chern et al. (2023) created FacTool, which combines multiple detection strategies including consistency checking, knowledge base verification, and confidence estimation. While effective, these approaches are primarily post-hoc and do not provide insights into the underlying mechanisms of hallucination generation.

Mechanistic Approaches

Recent work has begun investigating the internal mechanisms of hallucination generation. Burns et al. (2023) explored latent knowledge in language models without supervision, providing insights into how models represent factual information internally.

Our work extends this mechanistic approach by using information theory to understand how factual information flows through model layers and where it may be lost or corrupted, leading to hallucinations.

6.7.3 Representation Learning in Language Models

Understanding how language models learn and utilize internal representations is crucial for our information-theoretic analysis of hallucinations.

Probing Studies

Extensive research has investigated what linguistic information is captured in different layers of transformer models. Rogers et al. (2020) provides a comprehensive survey of BERT probing studies, while Tenney et al. (2019) analyzed the hierarchical nature of linguistic representations in BERT.

Hewitt & Manning (2019) demonstrated that syntactic information can be extracted from BERT representations using simple linear probes. Voita et al. (2019) showed that different attention heads in transformers capture different types of linguistic phenomena.

Mechanistic Interpretability

The mechanistic interpretability community has made significant progress in understanding transformer internals. Olah et al. (2020) introduced the concept of "circuits" in neural networks, identifying specific computational pathways for different tasks.

Kim et al. (2018) developed Concept Activation Vectors (CAVs) for understanding high-level concepts in neural networks. While not specifically focused on language models, this work provides methodological foundations for our layer-wise analysis approach.

Information Flow Analysis

Several studies have investigated information flow in neural networks. Voita & Titov (2020) analyzed information flow in neural machine translation models, demonstrating how different types of information are processed at different layers.

Our work extends this line of research by specifically focusing on question-answering tasks and using contrastive learning to measure information preservation across layers, with direct applications to hallucination detection.

6.7.4 Contrastive Learning in NLP

Contrastive learning has become increasingly important in NLP, particularly for representation learning and similarity measurement.

Sentence and Document Representations

Gao et al. (2021) developed SimCSE for learning sentence embeddings through contrastive learning, demonstrating significant improvements over previous methods. Chen et al. (2020a) established foundational principles for contrastive learning that have been adapted across domains.

Khosla et al. (2020) introduced supervised contrastive learning, which incorporates label information into the contrastive objective. This work is relevant to our approach, though we focus on layer-wise consistency rather than classification performance.

Mutual Information Estimation via Contrastive Learning

Poole et al. (2019) provided theoretical foundations for using contrastive learning to estimate mutual information, establishing variational bounds that justify contrastive approaches. Belghazi et al. (2018) developed MINE (Mutual Information Neural Estimation), which uses neural networks for MI estimation.

Our proposed contrastive MI estimation method builds upon these foundations but is specifically designed for analyzing information flow in transformer layers during question-answering tasks.

6.7.5 Positioning and Novel Contributions

This work occupies a unique position at the intersection of several research areas, making several novel contributions:

Theoretical Contributions

- **Information-Theoretic Framework for Hallucinations**: First comprehensive framework linking hallucinations to mutual information loss between questions and answers
- **Layer-Wise Information Flow Analysis**: Novel application of MI estimation to understand information degradation in transformer layers
- **Contrastive MI Estimation**: New method specifically designed for QA contexts and transformer architectures

Methodological Innovations

- **QA-Specific Contrastive Learning**: Adaptation of contrastive learning principles to question-answering consistency across layers
- **Real-Time Hallucination Detection**: Practical system for inference-time hallucination detection using MI estimates
- **Cross-Architecture Analysis**: Systematic comparison of information flow patterns across different transformer variants

Bridging Theory and Practice

Unlike purely theoretical information-theoretic work or purely empirical hallucination detection methods, this research provides:

- Theoretical understanding of hallucination mechanisms through information theory
- Practical tools for real-world hallucination detection
- Interpretable insights into model behavior that can guide architecture design
- Scalable methods that work with large foundation models

Relationship to Existing Work

Our approach differs from existing hallucination detection methods in several key ways:

- **Mechanistic vs. Behavioral**: We analyze internal information flow rather than just output behavior
- **Predictive vs. Reactive**: Our method can potentially predict hallucinations before generation completion
- **Interpretable vs. Black-Box**: Provides insights into why hallucinations occur, not just detection
- **Architecture-Agnostic**: Works across different transformer architectures and scales

This positioning establishes our work as a novel contribution that advances both theoretical understanding and practical applications in the critical area of language model reliability and trustworthiness.

6.8 Conclusion

This chapter outlines a comprehensive research program for understanding hallucinations in large language models through the lens of mutual information and representation learning. By providing a theoretical framework grounded in information theory, we aim to advance both our understanding of why hallucinations occur and our ability to detect and mitigate them effectively.

The proposed research addresses a critical gap in our theoretical understanding of hallucinations while offering practical applications for improving the reliability and trustworthiness of large language models in real-world deployments.

Chapter 7

Research Timeline

This chapter outlines a focused 12-month research timeline for investigating hallucinations in large language models using our proposed contrastive mutual information estimation method. The timeline emphasizes practical hallucination detection applications while building the necessary theoretical foundations.

7.1 Overview

The research program is structured around three main phases over 12 months, with each phase building upon the previous one to culminate in a robust hallucination detection system based on information-theoretic principles.

7.2 Phase 1: Foundation and Method Development (Months 1-4)

7.2.1 Month 1: Theoretical Framework

- Formalize information-theoretic framework for hallucinations
- Establish mathematical foundations for contrastive MI estimation
- Literature review of existing MI estimation methods
- Design synthetic datasets with known ground-truth MI for validation

7.2.2 Month 2: Contrastive MI Implementation

- Implement contrastive mutual information estimation method
- Develop projection functions and similarity metrics
- Create training pipeline for contrastive learning
- Validate method on synthetic data with known MI values

7.2.3 Month 3: Baseline Methods and Comparison

- Implement baseline MI estimation methods (MINE, InfoNCE)
- Establish comparison framework between methods
- Test all methods on small-scale language models (GPT-2, BERT-base)
- Initial correlation analysis between MI estimates and model confidence

7.2.4 Month 4: Method Refinement

- Conduct ablation studies on contrastive method design choices
- Optimize hyperparameters (temperature, projection architecture, negative sampling)
- Establish computational efficiency benchmarks
- Prepare for large-scale experiments

7.3 Phase 2: Large-Scale Validation and Hallucination Detection (Months 5-8)

7.3.1 Month 5: Foundation Model Analysis

- Apply contrastive MI estimation to large language models (GPT-3.5, LLaMA)
- Analyze layer-wise information flow patterns
- Identify critical layers where information degradation occurs

- Establish baseline hallucination rates on benchmark datasets

7.3.2 Month 6: Hallucination Correlation Studies

- Measure correlation between MI estimates and hallucination rates
- Test on factual QA datasets (Natural Questions, TriviaQA, WebQuestions)
- Analyze hallucination-specific benchmarks (HaluEval, TruthfulQA, FEVER)
- Develop MI-based hallucination detection thresholds

7.3.3 Month 7: Detection System Development

- Build real-time hallucination detection system using contrastive MI
- Implement inference-time MI estimation for new QA pairs
- Develop confidence calibration based on MI scores
- Create interpretable visualizations of information flow

7.3.4 Month 8: Cross-Architecture Validation

- Test detection system across different model architectures
- Validate on transformer variants (BERT, RoBERTa, T5)
- Analyze performance on different model scales
- Compare with existing hallucination detection methods

7.4 Phase 3: Applications and Deployment (Months 9-12)

7.4.1 Month 9: Domain-Specific Applications

- Apply hallucination detection to specialized domains (medical, legal, scientific)
- Test robustness across different question types and complexity levels

- Analyze performance on multi-turn conversations
- Develop domain-specific calibration strategies

7.4.2 Month 10: Intervention Strategies

- Develop MI-guided training objectives to reduce hallucinations
- Implement attention mechanism modifications based on MI insights
- Test intervention strategies on fine-tuned models
- Measure improvement in hallucination rates post-intervention

7.4.3 Month 11: Comprehensive Evaluation

- Conduct large-scale evaluation on diverse benchmarks
- Compare with state-of-the-art hallucination detection methods
- Analyze computational costs and scalability
- Perform human evaluation studies on detection accuracy

7.4.4 Month 12: Documentation and Dissemination

- Finalize research findings and prepare publications
- Create open-source implementation of contrastive MI method
- Develop user-friendly tools for hallucination detection
- Write comprehensive documentation and tutorials

7.5 Key Deliverables

7.5.1 Technical Deliverables

- Novel contrastive mutual information estimation method

- Real-time hallucination detection system
- Comprehensive benchmark evaluation results
- Open-source implementation and tools

7.5.2 Research Outputs

- Peer-reviewed publications on information-theoretic hallucination analysis
- Technical reports on contrastive MI estimation methodology
- Benchmark datasets and evaluation protocols
- Workshop presentations and conference talks

7.6 Risk Mitigation

7.6.1 Technical Risks

- **MI Estimation Accuracy**: Validate on synthetic data and cross-compare methods
- **Computational Scalability**: Implement efficient approximations and caching
- **Model Generalization**: Test across diverse architectures and scales

7.6.2 Timeline Risks

- **Implementation Delays**: Maintain parallel development tracks
- **Data Access Issues**: Establish multiple benchmark dataset sources
- **Computational Resources**: Secure cloud computing access and local GPU clusters

7.7 Success Metrics

7.7.1 Quantitative Metrics

- Hallucination detection AUROC > 0.85 on benchmark datasets

- MI estimation correlation > 0.7 with ground-truth synthetic data
- Computational overhead $< 20\%$ compared to baseline inference
- Cross-architecture performance variance $< 10\%$

7.7.2 Qualitative Metrics

- Interpretable insights into information flow in language models
- Practical applicability to real-world deployment scenarios
- Community adoption of open-source tools and methods
- Positive peer review feedback on research contributions

Chapter 8

Discussion

Bibliography

AIPlanet. Data Sprint 107 – Butterfly Image Classification [dataset]. `https://aiplanet.com/challenges/325/butterfly_identification/overview/about`, 2023. Accessed: 2025-05-09.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3):47, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, pp. 531–540, 2018.

Jonathan Berant, Andrew Ceccaldi, Antoine Fader, Evgeniy Gabrilovich, Percy Liang, and Luke Zettlemoyer. Semantic parsing on freebase from question-answer pairs. *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014a.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2023.

Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pp. 1597–1607, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2172–2180, 2016.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Qian, Kehua Wei, Chunting Zou, and Neubig Graham. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1999.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Nick Drummond and Rob Shearer. The open world assumption. In *eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web*, volume 15, pp. 1, 2006.

Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? *arXiv preprint arXiv:2402.03502*, 2024a.

Xuefeng Du, Yiyou Sun, and Yixuan Li. When and how does in-distribution label help out-of-distribution detection? *arXiv preprint arXiv:2405.18635*, 2024b.

Burak Ekim, Girmaw Abebe Tadesse, Caleb Robinson, Gilles Hacheme, Michael Schmitt, Rahul Dodhia, and Juan M Lavista Ferres. Distribution shifts at scale: Out-of-distribution detection in earth observation. *arXiv preprint arXiv:2412.13394*, 2024.

Dumitru Erhan, Ian Goodfellow, Will Cukierski, and Yoshua Bengio. Challenges in representation learning: Facial expression recognition challenge, 2013. URL `https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge`.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6568–6576, 2022.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems*, 35:37199–37213, 2022.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *arXiv preprint arXiv:2406.15012*, 2024.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Charles Guille-Escuret, Pau Rodriguez, David Vazquez, Ioannis Mitliagkas, and Joao Monteiro. Cadet: Fully self-supervised out-of-distribution detection with contrastive learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Shahriar Hossain, Jahir Uddin, Rakibul Nahin, and Salman Ibne Eunus. Rock classification dataset, 08 2021.

David P. Hughes and Marcel Salathé . An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *CoRR*, abs/1511.08060, 2015. URL `http://arxiv.org/abs/1511.08060`.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Jefkine Kafunah, Priyanka Verma, Muhammad Intizar Ali, and John G Breslin. Out-of-distribution data generation for fault detection and diagnosis in industrial systems. *IEEE Access*, 11:135061–135073, 2023.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *International conference on machine learning*, pp. 2668–2677, 2018.

Yusung Kim, Donghee Cho, and Jee-Hyong Lee. Wafer defect pattern classification with detecting out-of-distribution. *Microelectronics Reliability*, 122:114157, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets, 2009. URL `https://www.cs.toronto.edu/~kriz/cifar.html`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3214–3252, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pp. 22528–22538. PMLR, 2023.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Vivek Narayanaswamy, Yamen Mubarka, Rushil Anirudh, Deepta Rajan, and Jayaraman J Thiagarajan. Exploring inlier and outlier specification for improved medical ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4589–4598, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. *International Conference on Machine Learning*, pp. 5171–5180, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Manikandasriram Srinivasan Ramanagopal, Cyrus Anderson, Ram Vasudevan, and Matthew Johnson-Roberson. Failing to learn: Autonomously identifying perception failures for self-driving cars. *IEEE Robotics and Automation Letters*, 3(4):3860–3867, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 61:65–95, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Mojdeh Saadati, Aditya Balu, Shivani Chiranjeevi, Talukder Zaki Jubery, Asheesh K Singh, Soumik Sarkar, Arti Singh, and Baskar Ganapathysubramanian. Out-of-distribution detection algorithms for robust insect classification. *Plant Phenomics*, 6:0170, 2024.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu S Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. The information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint arXiv:2304.09355*, 2023.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Nikita Single, Harsh Jain, and Priyanka Jain. Realwaste: A novel real-life data set for landfill waste classification using deep learning. *IEEE Access*, 11:112562–112584, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Aditya Soni. Playing cards dataset, 2020. URL `https://www.kaggle.com/datasets/adityasoni04/playing-cards-dataset`.

sumanthvrao. Yoga poses, 2020. URL `https://www.kaggle.com/datasets/sumanthvrao/yoga-poses`. Version 6.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2019.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1802–1812, 2023.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL `http://arxiv.org/abs/1708.07747`.

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33:20685–20696, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *IEEE Transactions on Information Theory*, 63(9):5948–5964, 2017.

Chenhui Xu, Fuxun Yu, Zirui Xu, Nathan Inkawhich, and Xiang Chen. Out-of-distribution detection via deep multi-comprehension ensemble. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55465–55489. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/xu24ae.html`.

Hong Yang, Qi Yu, and Travis Desell. Can we ignore labels in out of distribution detection? In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Hanlin Zhang, Ziyang Li, Yuxin Zhao, Sheng Xu, et al. Sirens: Detecting hallucinations in large language models using uncertainty. *arXiv preprint arXiv:2310.13988*, 2023a.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023b.

Oliver Zhang, Jean-Benoit Delbrouck, and Daniel L Rubin. Out of distribution detection for medical images. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*, pp. 102–111. Springer, 2021.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

Yibo Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7379–7387, 2022.

Appendices

Appendix A

Theoretical Proofs for Label Blindness

This appendix contains the detailed theoretical proofs supporting the Label Blindness theory presented in Chapter 4. The proofs establish the mathematical foundations for understanding when and why unlabeled out-of-distribution detection methods are guaranteed to fail.

A.1 Properties of Mutual Information and Entropy

In this section we enumerate some of the properties of mutual information that are used to prove the theorems reported in this work. For any random variables $\mathbf{w}, \mathbf{x}, \mathbf{y}$ and $\mathbf{z}$:

(P_1) Positivity:

$$I(\mathbf{x}; \mathbf{y}) \geq 0, I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) \geq 0$$

(P_2) Chain rule:

$$I(\mathbf{xy}; \mathbf{z}) = I(\mathbf{y}; \mathbf{z}) + I(\mathbf{x}; \mathbf{z} \mid \mathbf{y})$$

(P_3) Chain rule (Multivariate Mutual Information):

$$I(\mathbf{x}; \mathbf{y}; \mathbf{z}) = I(\mathbf{y}; \mathbf{z}) - I(\mathbf{y}; \mathbf{z} \mid \mathbf{x})$$

(P_4) Positivity of discrete entropy: For discrete $\mathbf{x}$

$$H(\mathbf{x}) \geq 0, H(\mathbf{x} \mid \mathbf{y}) \geq 0$$

(P_5) Entropy and Mutual Information

$$H(\mathbf{x}) = H(\mathbf{x} \mid \mathbf{y}) + I(\mathbf{x}; \mathbf{y})$$

(P_6) Conditioning a variable cannot increase its entropy

$$H(\mathbf{y}|\mathbf{z}) \leq H(\mathbf{y})$$

(P_7) A variable knows about itself as much as any other variable can

$$I(\mathbf{x};\mathbf{x}) \geq I(\mathbf{x};\mathbf{y})$$

(P_8) Symmetry of Mutual Information

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{y};\mathbf{x})$$

(P_9) Entropy and Conditional Mutual Information (This is simply P_5 conditioned on $\mathbf{z}$)

$$I(\mathbf{x};\mathbf{y}|\mathbf{z}) = H(\mathbf{x}|\mathbf{z}) - H(\mathbf{x}|\mathbf{y}\mathbf{z})$$

(P_{10}) Functions of Independent Variables Remain Independent

$$I(\mathbf{x};\mathbf{y}) = 0 \rightarrow I(f(\mathbf{x});\mathbf{y}) = 0$$

A.2 Supporting Theorems and Proofs

This section contains supporting theorems and proofs that establish the foundation for the main Label Blindness results.

A.2.1 Sufficiency

Proposition A.2.1. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be random variables from joint distribution* $p(\mathbf{x},\mathbf{y})$*. Let* $\mathbf{z}$ *be a representation of* $\mathbf{x}$*, then* $\mathbf{z}$ *is sufficient for* $\mathbf{y}$ *if and only if* $I(\mathbf{x};\mathbf{y}) = I(\mathbf{y};\mathbf{z})$

Hypothesis:

$(H_1)\,\mathbf{z}$ *is a representation of* $\mathbf{x} : I(\mathbf{y}; \mathbf{z} \mid \mathbf{x}) = 0$

Thesis:

$(T_1)\, I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) = 0 \iff I(\mathbf{x}; \mathbf{y}) = I(\mathbf{y}; \mathbf{z})$

Proof.

$$I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) \overset{(P_3)}{=} I(\mathbf{x}; \mathbf{y}) - I(\mathbf{x}; \mathbf{y}; \mathbf{z}) \overset{(P_3)}{=} I(\mathbf{x}; \mathbf{y}) - I(\mathbf{y}; \mathbf{z}) + I(\mathbf{y}; \mathbf{z} \mid \mathbf{x})$$
$$\overset{(H_1)}{=} I(\mathbf{x}; \mathbf{y}) - I(\mathbf{y}; \mathbf{z})$$

Since both $I(\mathbf{x}; \mathbf{y})$ and $I(\mathbf{y}; \mathbf{z})$ are non-negative $(P_1)\,, I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) = 0 \iff I(\mathbf{y}; \mathbf{z}) = I(\mathbf{x}; \mathbf{y})$

□

A.2.2 Lower Bound of Mutual Information for Sufficiency

Lemma A.2.2. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be random variables with joint distribution* $p(\mathbf{x}, \mathbf{y})$*. Let* $\mathbf{z}$ *be a representation of* $\mathbf{x}$ *that is sufficient, as per definition 4.2.1. Then* $I(\mathbf{x}; \mathbf{z}) \geq I(\mathbf{z}; \mathbf{y})$ *and* $I(\mathbf{x}; \mathbf{z}) \geq I(\mathbf{x}; \mathbf{y})$*.*

Hypothesis:

(H_1) $\mathbf{z}$ *is a representation of* $\mathbf{x} : I(\mathbf{y}; \mathbf{z} \mid \mathbf{x}) = 0$

(H_2) $\mathbf{z}$ *is a sufficient representation of* $\mathbf{x} : I(\mathbf{x}; \mathbf{y}|\mathbf{z})) = 0$

Thesis:

(T_1) $\forall \mathbf{z}. I(\mathbf{x}; \mathbf{z}) \geq I(\mathbf{z}; \mathbf{y}), I(\mathbf{x}; \mathbf{z}) \geq I(\mathbf{x}; \mathbf{y})$

Proof. By Construction

$$
\begin{aligned}
I(\mathbf{xy}|\mathbf{z})) &\overset{(H_2)}{=} 0 \\
&\overset{(P_2)}{=} I(\mathbf{zy};\mathbf{x}) - I(\mathbf{z};\mathbf{x}) \\
&\overset{(P_2)}{=} I(\mathbf{x};\mathbf{y}) + I(\mathbf{x};\mathbf{z}|\mathbf{y}) - I(\mathbf{z};\mathbf{x}) \\
&\overset{(PropB1)}{=} I(\mathbf{z};\mathbf{y}) + I(\mathbf{x};\mathbf{z}|\mathbf{y}) - I(\mathbf{z};\mathbf{x}) \\
I(\mathbf{z};\mathbf{x}) &= I(\mathbf{z};\mathbf{y}) + I(\mathbf{x};\mathbf{z}|\mathbf{y}) \\
I(\mathbf{z};\mathbf{x}) &\overset{(P_1)}{\geq} I(\mathbf{z};\mathbf{y})
\end{aligned}
$$

Note that $I(\mathbf{z};\mathbf{y}) = I(\mathbf{x};\mathbf{y})$ for all sufficient representations, as per proposition A.2.1.

This supports our intuition that the information in the representation consists of relevant information $I(\mathbf{z};\mathbf{y})$ and irrelevant information $I(\mathbf{x};\mathbf{z}|\mathbf{y})$. By definition of sufficiency, there must be enough information for $I(\mathbf{z};\mathbf{y})$ in $I(\mathbf{x};\mathbf{z})$, which is to say that the size of the encoding cannot be smaller than the minimum size to encode all of $I(\mathbf{x};\mathbf{y})$.

□

A.2.3 Conditional Mutual Information of Noise

Lemma A.2.3. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be independent random variables and* $\mathbf{z}$ *be a function of* $\mathbf{x}$ *with joint distribution* $p(\mathbf{x},\mathbf{y},\mathbf{z})$*. The conditional mutual information* $I(\mathbf{x};\mathbf{z}|\mathbf{y})$ *is always equal to the mutual information* $I(\mathbf{x};\mathbf{z})$*. As in the information content is unchanged when adding noise.*

Hypothesis:

(H_1) *Independence of* $\mathbf{x}$ *and* $\mathbf{y}$ *:* $I(\mathbf{x};\mathbf{y}) = 0$

(H_2) $\mathbf{z}$ *is fully determined by* $\mathbf{x}$ *:* $H(\mathbf{z}|\mathbf{x}) = 0$

Thesis:

(T_1) $I(\mathbf{x};\mathbf{z}|\mathbf{y}) = I(\mathbf{x};\mathbf{z})$

Proof. By Construction.

(C_1) Demonstrates that $H(\mathbf{z}|\mathbf{xy}) = 0$

$$0 \overset{(P_4)}{\leq} H(\mathbf{z}|\mathbf{xy}) \overset{(P_6)}{\leq} H(\mathbf{z}|\mathbf{x})$$
$$H(\mathbf{z}|\mathbf{xy}) \overset{(H_2)}{\leq} 0$$

(C_2) Demonstrates that $I(\mathbf{z};\mathbf{y}) = 0$

$$\begin{aligned} I(\mathbf{z};\mathbf{y}) &\overset{(H_2)}{=} I(f(\mathbf{x});\mathbf{y}) \\ &\overset{(P_{10})}{=} I(\mathbf{x};\mathbf{y}) \\ &\overset{(H_1)}{=} 0 \end{aligned}$$

Thus

$$\begin{aligned} I(\mathbf{x};\mathbf{z}|\mathbf{y}) &\overset{(P_9)}{=} H(\mathbf{z}|\mathbf{y}) - H(\mathbf{z}|\mathbf{xy}) \\ &\overset{(C_1)}{=} H(\mathbf{z}|\mathbf{y}) - 0 \\ &\overset{(P_5)}{=} H(\mathbf{z}) - I(\mathbf{z};\mathbf{y}) \\ &\overset{(C_2)}{=} H(\mathbf{z}) - 0 \\ &\overset{(H_2)}{=} H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}) \\ &\overset{(P_5)}{=} I(\mathbf{x};\mathbf{z}) \end{aligned}$$

This supports the intuition that if one added a random noise channel it will not change the mutual information.

□

A.2.4 Factorization of Bottleneck Loss

Lemma A.2.4. *Let* $\mathbf{x}$ *be a random variable with label* $\mathbf{y}$ *such that* $H(\mathbf{y}|\mathbf{x}) = 0$ *and* $\mathbf{z}$ *is a sufficient representation of* $\mathbf{x}$ *for* $\mathbf{y}$*. The loss function* $\mathcal{L} = I(\mathbf{x};\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y})$ *is equivalent to* $\mathcal{L} = H(\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y})$*, with* β *as some constant.*

Hypothesis:

(H_1) $\mathbf{z}$ *is fully determined by* $\mathbf{x}$ *:* $H(\mathbf{z}|\mathbf{x}) = 0$

Thesis:

(T_1) $I(\mathbf{x};\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y}) = H(\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y})$

Proof. By Construction.

$$I(\mathbf{x};\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y}) \overset{(P_5)}{=} H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}) - \beta I(\mathbf{z};\mathbf{y})$$
$$\overset{(H_1)}{=} H(\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y})$$

Due to the relationship between $\mathbf{x}$ and $\mathbf{z}$, we can create an intuitive factorization of the bottleneck loss function. Effectively, we want to maximize $I(\mathbf{z};\mathbf{y})$ while minimizing the information content of $\mathbf{z}$

□

A.3 Main Theorems and Proofs

We ignore cases where the determined variable has an entropy of 0. Generally, if $H(\mathbf{y}|\mathbf{x}) = 0 \rightarrow H(\mathbf{y}) > 0$. Also, we only consider cases where the random variables have more than zero entropy.

Note that $R_{\mathbf{x}}$ represents the support of random variable $\mathbf{x}$ such that $R_{\mathbf{x}} = \{\boldsymbol{x} \in \mathbb{R} : P(\boldsymbol{x}) > 0\}$.

A.3.1 Strict Label Blindness in the Minimal Sufficient Statistic

Theorem A.3.1 (Strict Label Blindness in the Minimal Sufficient Statistic)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. Let* $\mathbf{y}_1$ *be a surrogate task such that* $H(\mathbf{y}_1|\mathbf{x}_1) = 0$*. Let* $\mathbf{z}$ *be any sufficient representation of* $\mathbf{x}$ *for* $\mathbf{y}_1$ *that satisfies the sufficiency definition 4.2.1 and minimizes the loss function* $\mathcal{L} = I(\mathbf{x}_1\mathbf{x}_2;\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y}_1)$*. The possible* $\mathbf{z}$ *that minimizes* $\mathcal{L}$ *and is sufficient must meet the condition* $I(\mathbf{x}_2;\mathbf{z}) = 0$*.*

Summary: *This proof uses the derivative of the loss function to establish the possible set of local minima that satisfies* $\mathcal{L}$*. For any possible minima of* $\mathcal{L}$*, the representation* $\mathbf{z}$ *must contain information of only* $\mathbf{x}_1 \rightarrow H(\mathbf{z}|\mathbf{x}_1) = 0$ *or only* $\mathbf{x}_2 \rightarrow H(\mathbf{z}|\mathbf{x}_2) = 0)$ *or both* $\mathbf{x}_1, \mathbf{x}_2 \rightarrow H(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2) = 0$*. We show that possible set of all local minima must satisfy* $H(\mathbf{z}|\mathbf{x}_1) = 0$ *by showing that the other*

two cases must always have greater $\mathcal{L}$. This proves the Theorem that the learned representation cannot contain information about $\mathbf{x}_2$.

Hypothesis:

(H_1) $\mathbf{z}$ *is fully determined by* $\mathbf{x}$ *:* $H(\mathbf{z}|\mathbf{x}) = 0$

(H_2) $\mathbf{z}$ *is a representation of* $\mathbf{x} : I(\mathbf{y}; \mathbf{z} \mid \mathbf{x}) = 0$

(H_3) $\mathbf{z}$ *is a sufficient representation of* $\mathbf{x} : I(\mathbf{x}; \mathbf{y}|\mathbf{z}) = 0$

(H_4) $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1, \mathbf{x}_2$ *:* $\mathbf{x} = \mathbf{x}_1, \mathbf{x}_2, I(\mathbf{x}_1; \mathbf{x}_2) = 0$

(H_5) $\mathbf{y}$ *is fully determined by* $\mathbf{x}_1$*:* $H(\mathbf{y}|\mathbf{x}_1) = 0$

Thesis:

(T_1) $\forall \mathbf{z}. I(\mathbf{x}_2, \mathbf{z}) = 0$

Proof. By Construction

(C_1) demonstrates that $\mathcal{L} = H(\mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y})$ via factoring $I(\mathbf{x}_1\mathbf{x}_2; \mathbf{z})$. Alternatively, Theorem A.2.4 creates the same result.

$$
\begin{aligned}
I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{z}) &\overset{(P_2)}{=} I(\mathbf{x}_1; \mathbf{z}) + I(\mathbf{x}_2; \mathbf{z}|\mathbf{x}_1) \\
&\overset{(P_5)}{=} H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}_1) + I(\mathbf{x}_2; \mathbf{z}|\mathbf{x}_1) \\
&\overset{(P_9)}{=} H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}_1) + H(\mathbf{z}|\mathbf{x}_1) - H(\mathbf{z}|\mathbf{x}_1\mathbf{x}_2) \\
&\overset{(H_1)}{=} H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}_1) + H(\mathbf{z}|\mathbf{x}_1) - 0 \\
\mathcal{L} &= H(\mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y})
\end{aligned}
$$

(C_2) Demonstrates that $I(\mathbf{z}; \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$ as per Theorem A.2.1.

(C_3) Demonstrates that $I(\mathbf{z}; \mathbf{y})$ is a constant across all sufficient representations because Theorem A.2.1 applies.

(C_4) Demonstrates that for all possible $\mathbf{z}$ satisfying (H_3), their loss can be compared using only $\mathcal{L}_z = H(\mathbf{z})$ for comparing across $\mathbf{z}$

$$\frac{d\mathcal{L}}{d\mathbf{z}} \overset{(C_1)}{=} \frac{H(\mathbf{z})}{d\mathbf{z}} - \frac{\beta I(\mathbf{z};\mathbf{y})}{d\mathbf{z}}$$
$$\overset{(C_3)}{=} \frac{H(\mathbf{z})}{d\mathbf{z}} - 0$$

(C_5) Demonstrates that the value of $H(\mathbf{z})$ at all possible $\mathbf{z}$ that minimizes $\mathcal{L}$ is the same. Even for different minimal $\mathbf{z}$, they must have the same $H(\mathbf{z})$ to all be minimal. When comparing possible minimal solutions to $\mathcal{L}$, $H(\mathbf{z})$ is constant across all minimal solutions.

(C_6) Demonstrates that any $\mathbf{z}$ that satisfies sufficiency must satisfy $I(\mathbf{z};\mathbf{x}) \geq I(\mathbf{z};\mathbf{y})$ and $I(\mathbf{z};\mathbf{x}) \geq I(\mathbf{x};\mathbf{y})$ as per Theorem A.2.2.

(C_7) Demonstrates that minima(s) exists only where $H(\mathbf{z}) = I(\mathbf{z};\mathbf{y})$ and $H(\mathbf{z}|\mathbf{x}) = 0$. Note that $H(\mathbf{z}) = I(\mathbf{x};\mathbf{y}) = I(\mathbf{z};\mathbf{y})$ is the most compact representation size that is sufficient.

$$I(\mathbf{z};\mathbf{x}) \overset{(C_6)}{\geq} I(\mathbf{z};\mathbf{y})$$
$$H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}) \overset{(P_5)}{\geq} I(\mathbf{z};\mathbf{y})$$
$$\forall \mathbf{z} | C_6 \wedge H_3 \wedge I(\mathbf{z};\mathbf{x}) > I(\mathbf{z};\mathbf{y}). \exists \mathbf{z}' | \mathbf{z}' = f(\mathbf{z}) \wedge I(\mathbf{z};\mathbf{x}) > I(\mathbf{z}';\mathbf{x}) \wedge C_6 \wedge H_3$$

From (C_7) there exists only 3 types of minimas, separated by their dependence on the variables $\mathbf{x}_1, \mathbf{x}_2$. As per ($H_1$), any $\mathbf{z}$ must follow one of the 3 types.

1. Dependent only on $\mathbf{x}_1$: $\forall \mathbf{z} | H(\mathbf{z}|\mathbf{x}_1) = 0 \rightarrow I(\mathbf{x}_2;\mathbf{z}) = 0$
2. Dependent only on $\mathbf{x}_2$: $\forall \mathbf{z} | H(\mathbf{z}|\mathbf{x}_2) = 0 \rightarrow I(\mathbf{x}_2;\mathbf{z}) > 0$
3. Dependent on both $\mathbf{x}_1\mathbf{x}_2$: $\forall \mathbf{z} | H(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2) = 0 \wedge H(\mathbf{z}|\mathbf{x}_1) > 0 \wedge H(\mathbf{z}|\mathbf{x}_2) > 0 \rightarrow I(\mathbf{x}_2;\mathbf{z}) > 0$

From here we will show that all type 2 and type 3 minimas always fail (H_3) or have greater $\mathcal{L}$ than any type 1 minima.

Type 1 $\mathbf{x}_1$: $\forall \mathbf{z} | H(\mathbf{z}|\mathbf{x}_1) = 0 \rightarrow I(\mathbf{x}_2;\mathbf{z}) = 0$

(C_8) Demonstrates that there exists $H(\mathbf{z}) = I(\mathbf{z};\mathbf{y}) = I(\mathbf{x}_1;\mathbf{z})$ and it is a set of minimas satisfying (C_7). This also establishes an upper bound for solutions to $\mathcal{L}$ due to (C_5). Therefore, any solution for type 1, type 2, and type 3 must satisfy $I(\mathbf{z};\mathbf{y}) \leq I(\mathbf{x}_1;\mathbf{z})$ to be sufficient and $I(\mathbf{z};\mathbf{y}) = I(\mathbf{x}_1;\mathbf{z})$ to be minimal.

$$
\begin{aligned}
I(\mathbf{z};\mathbf{y}) &\overset{(C_6)}{\leq} I(\mathbf{z};\mathbf{x}) \\
&\overset{(H_4)}{\leq} I(\mathbf{x}_1,\mathbf{x}_2;\mathbf{z}) \\
&\overset{(P_2)}{\leq} I(\mathbf{x}_2;\mathbf{z}) + I(\mathbf{x}_1;\mathbf{z}|\mathbf{x}_2) \\
&\overset{(Type1)}{\leq} 0 + I(\mathbf{x}_1;\mathbf{z}|\mathbf{x}_2) \\
&\overset{(TheoremA.2.3)}{\leq} I(\mathbf{x}_1;\mathbf{z}) \\
&\overset{(P_5)}{\leq} H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}_1) \\
\exists \mathbf{z} | I(\mathbf{x}_1;\mathbf{z}) &= I(\mathbf{z};\mathbf{y}) = I(\mathbf{x};\mathbf{z}) = I(\mathbf{x};\mathbf{y})
\end{aligned}
$$

(C_9) Demonstrates that there exists no $H(\mathbf{z}') < H(\mathbf{z})$ that satisfies sufficiency if $\mathbf{z}$ satisfies (C_8) and is also $I(\mathbf{z};\mathbf{x}_2) = 0$.

$$
\begin{aligned}
C_8 &\rightarrow I(\mathbf{x}_1;\mathbf{z}) = I(\mathbf{x};\mathbf{y}) \\
H(\mathbf{z}') < H(\mathbf{z}) &\rightarrow I(\mathbf{x}_1;\mathbf{z}') < I(\mathbf{x}_1;\mathbf{z}) \\
&\rightarrow \neg(C_2) : I(\mathbf{x}_1;\mathbf{z}') < I(\mathbf{x}_1;\mathbf{z}) = I(\mathbf{y};\mathbf{z}) = I(\mathbf{x};\mathbf{y})
\end{aligned}
$$

Type 2 $\mathbf{x}_2$: $\forall \mathbf{z} | H(\mathbf{z}|\mathbf{x}_2) = 0 \rightarrow I(\mathbf{x}_2;\mathbf{z}) > 0$

(C_{10}) Demonstrates that no type 2 minima can exist, simply because it would contain no information regarding $\mathbf{x}_1$, thus failing to satisfy (H_3). This is because $\mathbf{z}$ cannot contain any information about $\mathbf{x}_1$, otherwise we would not satisfy $H(\mathbf{z}|\mathbf{x}_2) = 0$. If the representation $\mathbf{z}$ contains no information about $\mathbf{y}$, then it is not sufficient.

$$
\begin{aligned}
H(\mathbf{z}|\mathbf{x}_2) = 0 &\rightarrow \mathbf{z} = f(\mathbf{x}_2) \\
0 &\overset{(H_4)}{=} I(\mathbf{x}_1;\mathbf{x}_2) \\
&\overset{(P_{10})}{=} I(f(\mathbf{x}_1);\mathbf{x}_2) \\
&\overset{(H_5)}{=} I(\mathbf{y};\mathbf{x}_2) \\
&\overset{(P_{10})}{=} I(\mathbf{y};f(\mathbf{x}_2)) \\
0 &= I(\mathbf{y};\mathbf{z})
\end{aligned}
$$

Type 3 $\mathbf{x}_1, \mathbf{x}_2$: $\forall \mathbf{z} | H(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2) = 0 \wedge H(\mathbf{z}|\mathbf{x}_1) > 0 \wedge H(\mathbf{z}|\mathbf{x}_2) > 0 \rightarrow I(\mathbf{x}_2;\mathbf{z}) > 0$

(C_{11}) Demonstrates that any $\mathbf{z}$ that could be minimal must also satisfy (C_8) for sufficiency. Note that (C_8) implies that any $I(\mathbf{x}_1;\mathbf{z}) > I(\mathbf{z};\mathbf{y})$ is not minimal.

$$
\begin{aligned}
I(\mathbf{z};\mathbf{y}) &\overset{(C_6)}{\leq} I(\mathbf{z};\mathbf{x}) \\
&\overset{(H_4)}{\leq} I(\mathbf{x}_1\mathbf{x}_2;\mathbf{z}) \\
I(\mathbf{z};\mathbf{y}) &\overset{(P_2)}{\leq} I(\mathbf{x}_1;\mathbf{z}) + I(\mathbf{x}_2;\mathbf{z}|\mathbf{x}_1) \\
(C_8) &\rightarrow I(\mathbf{x}_1;\mathbf{z}) = I(\mathbf{z};\mathbf{y})
\end{aligned}
$$

(C_{12}) Demonstrates that any $\mathbf{z}'$ where $I(\mathbf{z}';\mathbf{x}_2) > I(\mathbf{z};\mathbf{x}_2)$ and $I(\mathbf{z};\mathbf{x}_2) = 0$ that maintains $H(\mathbf{z}') = H(\mathbf{z})$ results in solutions that are not sufficient as required by (H_3) because we know that the size of the representation must be at least $I(\mathbf{x};\mathbf{y})$ as defined in (C_6)

$$
\begin{aligned}
C_8 &\rightarrow H(\mathbf{z}) \text{ is constant across all minima} \\
C_8 &\rightarrow H(\mathbf{z}) = H(\mathbf{z}') \text{ for } \mathbf{z}' \text{ to be minimal} \\
C_8 &\rightarrow I(\mathbf{x}_1;\mathbf{z}) = I(\mathbf{x};\mathbf{y}) \\
I(\mathbf{x}_2;\mathbf{z}) = 0 &\rightarrow H(\mathbf{z}|\mathbf{x}_1) = 0 \\
\forall \mathbf{z}' | I(\mathbf{x}_2;\mathbf{z}') > 0 &: H(\mathbf{z}'|\mathbf{x}_1) > H(\mathbf{z}|\mathbf{x}_1) \\
H(\mathbf{z}'|\mathbf{x}_1) > H(\mathbf{z}|\mathbf{x}_1) &\rightarrow H(\mathbf{z}') - H(\mathbf{z}'|\mathbf{x}_1) < H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}_1) \\
&\overset{(P_5)}{\rightarrow} I(\mathbf{x}_1;\mathbf{z}') < I(\mathbf{x}_1;\mathbf{z}) \\
\rightarrow \neg(C_6) &: I(\mathbf{x}_1;\mathbf{z}') < I(\mathbf{x};\mathbf{y})
\end{aligned}
$$

(C_{13}) Demonstrates that combining (C_{11}) and (C_{12}), there is no type 3 solution that has an equal $\mathcal{L}$ to the minimal type 1 solution that also maintains sufficiency (H_3) and (C_6). This confirms the definition of entropy, in that encoding more independent information requires more bits or nats.

This means that only a type 1 solution can be both minimal and sufficient, which proves the thesis.

To summarize this proof, we can compare the losses of all sufficient solutions with $\mathcal{L} = H(\mathbf{z})$. Of those sufficient solutions, the one that minimizes $\mathcal{L}$ is the one with the smallest $H(\mathbf{z})$. The minimal sufficient representation is $\mathbf{z}$ that captures only all of $I(\mathbf{x}_1;\mathbf{y})$ and nothing else. Thus the minimal $\mathbf{z}$ cannot have $I(\mathbf{x}_2;\mathbf{z}) > 0$ because such $\mathbf{z}$ would encode information outside of $I(\mathbf{x}_1;\mathbf{y})$.

□

A.3.2 Independence of Filtered Distributions

Lemma A.3.2 (Independence of Filtered Distributions)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. For* $\mathbf{x}'_2$ *where* $R_{\boldsymbol{x}'_2} \subset R_{\boldsymbol{x}_2}$*, there exists no* $\mathbf{x}'_2$ *such that* $H(\mathbf{x}_1|\mathbf{x}'_2) < H(\mathbf{x}_1)$*.*

Summary: *This proof uses the chain rule of mutual information to show that contradiction arises if* $\mathbf{x}'_2$ *could filter* $\mathbf{x}_1$ *in a non random way.*

Hypothesis:

(H_1) $\mathbf{x}'_2$ *is fully determined by* $\mathbf{x}_2$ *:* $H(\mathbf{x}'_2|\mathbf{x}_2) = 0$ *where* $R_{\boldsymbol{x}'_2} \subset R_{\boldsymbol{x}_2}$

(H_2) *Independence of* $\mathbf{x}_1$ *and* $\mathbf{y}_2$ *:* $I(\mathbf{x}_1; \mathbf{x}_2) = 0$

Thesis:

(T_1) $\nexists \mathbf{x}'_2 . H(\mathbf{x}_1|\mathbf{x}_2) < H(\mathbf{x}_1)$

Proof. by contradiction $H(\mathbf{x}_1|\mathbf{x}_2) < H(\mathbf{x}_1)$

(C_1) Demonstrates $I(\mathbf{x}_2; \mathbf{x}'_2) = I(\mathbf{x}'_2; \mathbf{x}'_2)$

$$\begin{aligned} I(\mathbf{x}_2; \mathbf{x}'_2) &\overset{(P_4)}{=} H(\mathbf{x}'_2) - H(\mathbf{x}'_2|\mathbf{x}_2) \\ &\overset{(H_1)}{=} H(\mathbf{x}'_2) - 0 \\ &\overset{(P_3)}{=} H(\mathbf{x}'_2) - H(\mathbf{x}'_2|\mathbf{x}'_2) \\ &\overset{(P_3)}{=} I(\mathbf{x}'_2; \mathbf{x}'_2) \end{aligned}$$

(C_2) Demonstrates $I(\mathbf{x}'_2; \mathbf{x}_1|\mathbf{x}_2) = 0$

$$
\begin{aligned}
I(\mathbf{x}_2';\mathbf{x}_1|\mathbf{x}_2) &\overset{(P_2)}{=} I(\mathbf{x}_2';\mathbf{x}_2\mathbf{x}_1) - I(\mathbf{x}_2;\mathbf{x}_2') \\
&\overset{(C_1)}{=} I(\mathbf{x}_2';\mathbf{x}_2\mathbf{x}_1) - I(\mathbf{x}_2';\mathbf{x}_2') \\
I(\mathbf{x}_2';\mathbf{x}_1|\mathbf{x}_2) &\overset{(P_7)}{\leq} 0 \leftarrow I(\mathbf{x}_2';\mathbf{x}_2\mathbf{x}_1) \leq I(\mathbf{x}_2';\mathbf{x}_2') \\
&\overset{(P_1)}{\geq} 0 \\
&= 0
\end{aligned}
$$

(C_3) Demonstrates $I(\mathbf{x}_1;\mathbf{x}_2') > 0$ via non independence implied by $\neg T_1$

Contradiction arises when we consider symmetric applications of the chain rule to $I(\mathbf{x}_1;\mathbf{x}_2\mathbf{x}_2')$

$$
\begin{aligned}
I(\mathbf{x}_1;\mathbf{x}_2'\mathbf{x}_2) &\overset{(P_2)}{=} I(\mathbf{x}_1;\mathbf{x}_2') + I(\mathbf{x}_1;\mathbf{x}_2|\mathbf{x}_2') \\
I(\mathbf{x}_1;\mathbf{x}_2'\mathbf{x}_2) &\overset{(C_3)}{>} 0 \\
I(\mathbf{x}_1;\mathbf{x}_2\mathbf{x}_2') &\overset{(P_2)}{=} I(\mathbf{x}_1;\mathbf{x}_2) + I(\mathbf{x}_1;\mathbf{x}_2'|\mathbf{x}_2) \\
&\overset{(C_2)}{=} I(\mathbf{x}_1;\mathbf{x}_2) \\
&\overset{(H_2)}{=} 0
\end{aligned}
$$

Since $I(\mathbf{x}_1;\mathbf{x}_2\mathbf{x}_2')$ cannot be both zero and greater than zero, $\neg T_1$ creates a contradiction, which supports T_1.

It is easy to confuse this with the existence of a non independent subset $\mathbf{C} := \mathbf{A} \cap \mathbf{B}$, where $\mathbf{A}, \mathbf{B}$ are independent events. However, this example violates (H_1), since we cannot determine $\mathbf{C}$ using only $\mathbf{A}$ or only $\mathbf{B}$.

□

A.3.3 Strict Label Blindness in Filtered Distributions - Guaranteed OOD Failure

Corollary A.3.3 (Strict Label Blindness in Filtered Distributions)**.** *Let* $\mathbf{x}$ *come from a distribution.* $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. Let* $\mathbf{y}_1$ *be a surrogate task such generated by* $\boldsymbol{y}_1 = f_1(\boldsymbol{x}_1)$ $H(\mathbf{y}_1|\mathbf{x}_1) = 0$*. Let* $\mathbf{y}_2$ *be a label such that* $H(\mathbf{y}_2|\mathbf{x}_2) = 0$ *and* $\boldsymbol{y}_2 = f_2(\boldsymbol{x}_2)$*. Let* $\mathbb{Y}_{in}$

be as subset of labels $\mathbb{Y}_{in} \subset R_{\mathbf{y}_2}$. *Let* $\mathbf{x}'$ *be a subset of* $\mathbf{x}$ *where* $R_{\mathbf{x}'} = R_{\mathbf{x}} \cap \{\boldsymbol{x} \in \mathbb{R} : f_2(\boldsymbol{x}_2) \in \mathbb{Y}_{in}\}$ *such that* $\mathbf{x}'$ *is composed of independent variables* $\mathbf{x}'_1$ *and* $\mathbf{x}'_2$ *and* $\boldsymbol{y}'_1 = f_1(\boldsymbol{x}'_1)$. *The sufficient representation* $\mathbf{z}$ *learned by minimizing* $\mathcal{L} = I(\mathbf{x}'_1, \mathbf{x}'_2; \mathbf{z}) - \beta I(\mathbf{z}; \mathbf{y}'_1)$ *must have* $I(\mathbf{x}_2; \mathbf{z}) = 0$ *and* $I(\mathbf{y}_2; \mathbf{z}) = 0$.

Summary: *This proof combines Theorem A.3.2 and Theorem A.3.1.*

Hypothesis:

(H_1) $\mathbf{z}$ *is fully determined by* $\mathbf{x}$ *:* $H(\mathbf{z}|\mathbf{x}) = 0$

(H_2) $\mathbf{z}$ *is a representation of* $\mathbf{x} : I(\mathbf{y}; \mathbf{z} \mid \mathbf{x}) = 0$

(H_3) $\mathbf{z}$ *is a sufficient representation of* $\mathbf{x} : I(\mathbf{x}; \mathbf{y}|\mathbf{z}) = 0$

(H_4) $\mathbf{x}$ *is composed of two independent variables* $\mathbf{x}_1, \mathbf{x}_2$ *:* $\mathbf{x} = \mathbf{x}_1\mathbf{x}_2, I(\mathbf{x}_1; \mathbf{x}_2) = 0$

(H_5) $\mathbf{y}$ *is fully determined by* $\mathbf{x}_1$*:* $H(\mathbf{y}|\mathbf{x}_1) = 0$

(H_6) $\mathbf{x}'$ *is a subset of* $\mathbf{x}$ *filtered by* $\mathbb{Y}_{in}$ *:* $R_{\mathbf{x}'} = R_{\mathbf{x}} \cap \{\boldsymbol{x} \in \mathbb{R} : f_2(\boldsymbol{x}_2) \in \mathbb{Y}_{in}\}$

Thesis:

(T_1) $\forall \mathbf{z}. I(\mathbf{x}_2; \mathbf{z}) = 0, I(\mathbf{x}'_2; \mathbf{z}) = 0$

Proof. By Construction.

(C_1) Demonstrates that $I(\mathbf{x}'_1; \mathbf{x}'_2) = 0$ due to Lemma A.3.2

Using (P_{10}), we know that independent functions stay independent and thus $I(\mathbf{x}'_1; \mathbf{x}_2) = 0, I(\mathbf{x}'_1; \mathbf{x}_2) = 0$. From TheoremA.3.1 we know that encoding an variable independent of the target y results in a higher loss, therefore $I(\mathbf{x}'_2; \mathbf{z}) = 0$ and $I(\mathbf{x}_2; \mathbf{z}) = 0$ since both are independent of $\mathbf{x}'_1$.

By combining Lemma A.3.2 and TheoremA.3.1, we know that any surrogate learning objective independent of a downstream objective (say classifying labels) results in a representation containing no information for the downstream objective. If it contains no information for one objective, it contains no information for derivatives of that objective (eg. no label information means no OOD detection information).

□

A.3.4 Unavoidable Risk of Overlapping Out of Distribution Data

Theorem A.3.4 (Unavoidable Risk of Overlapping OOD Data)**.** *Let* $\mathbf{x}$ *come from a distribution. Let* f *be some labeling function to generate labels* $\mathbf{y}$ *such that* $\boldsymbol{y} = f(\boldsymbol{x})$*, where there are at least two unique labels* $|R_{\mathbf{y}}| > 1$*. Let* $\mathbf{x}_{in}$ *be a random subset of* $\mathbf{x}$ *where* $R_{\mathbf{x}_{in}} \subsetneq R_{\mathbf{x}}$ *and* $|R_{\mathbf{x}_{in}}| < \infty$*. Let* $\mathbf{y}_{in}$ *be labels generated from* $\boldsymbol{y}_{in} = f(\boldsymbol{x}_{in})$*. The probability that a randomly selected* $\boldsymbol{x}$ *contains* $\boldsymbol{y}$ *not present in* $R_{\mathbf{y}_{in}}$ *is always greater than 0.*

Hypothesis:

(H_1) $\mathbf{x}$ *comes from any distribution*

(H_2) $\mathbf{y}$ *is a label generated from function* $\boldsymbol{y} = f(\boldsymbol{x})$ *such that* $|R_{\mathbf{y}}| > 1$

(H_3) $\mathbf{x}_i$ *is a random subset of* $\mathbf{x}$ *where* $R_{\mathbf{x}_{in}} \subsetneq R_{\mathbf{x}}$ *and* $|R_{\mathbf{x}_{in}}| < \infty$ *and* $\boldsymbol{y}_{in} = f(\boldsymbol{x}_{in})$*.*

Thesis

(T_1) $\forall \mathbf{x}.P(f(\boldsymbol{x}) \notin R_{\mathbf{y}_{in}}) > 0)$

Proof. by contradiction $(\neg T_1)$ $P(f(\boldsymbol{x}) \notin R_{\mathbf{y}_{in}}) = 0\}$

(C_1) Demonstrates that $\forall \mathbf{x}_i.R_{\mathbf{y}_{in}} = R_{\mathbf{y}}$ because there must exist no sample $\boldsymbol{x}$ such that $f(\boldsymbol{x}) \notin R_{\mathbf{y}_{in}}$.

(C_2) Demonstrates that $\forall \boldsymbol{y}_n.P(f(\boldsymbol{x}) = \boldsymbol{y}_n) > 0$, where $\boldsymbol{y}_n \in R_{\mathbf{y}}$

Contradiction arises when we consider that it is possible to sample the same label $\boldsymbol{y}_n$ for any finite number of repetitions, as per (C_2). This would create a set of any finite size consisting only of the label $\boldsymbol{y}_n$. Thus, there always exists $R_{\mathbf{y}_{in}} \subsetneq R_{\mathbf{y}}$ which contradicts (C_1).

More realistically, $R_{\mathbf{y}_{in}}$ can consist of all elements of $R_{\mathbf{y}}$ except one and still guarantee $P(f(\boldsymbol{x}) \notin R_{\mathbf{y}_{in}}) > 0)$.

□