[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.

2.6 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.

Intrinsic: $\exists(\boldsymbol{y}', \boldsymbol{y}'') \subseteq \boldsymbol{y}$ such that $\boldsymbol{y}' \not\Rightarrow \boldsymbol{y}''$ 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., 2020; 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., 2020) 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.

3.4 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 understanding 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., 2020) 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. (2020), 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* $I(\mathbf{x}_2; \mathbf{z}) = 0$*.*

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

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* $I(\mathbf{y}_2; \mathbf{z}) = 0$*.*

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.*

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 performance 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., 2020) 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. (2014) 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 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).

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

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: How Single Domain Training Removes Domain Features

This chapter covers on going research.

Chapter 6

Are Hallucinations Out of Distribution?

This chapter covers a proposal to research the nature of hallucinations.

Chapter 7

Research Timeline

Chapter 8

Discussion

Bibliography

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.

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

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.

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, 2020.

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.

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.

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.

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.

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.

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.

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.

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.

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 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

First Appendix

This is an appendix.

A.1 First Appendix Section

A.1.1 First Appendix Subsection