entropy

MDPI

Review

To Compress or Not to Compress—Self-Supervised Learning and Information Theory: A Review

Ravid Shwartz Ziv [1,*] and Yann LeCun [1,2]

[1] Center of Data Science, New York University, New York, NY 10011, USA
[2] FAIR at Meta, Broadway, New York, NY 10003, USA
[*] Correspondence: ravid.shwartz.ziv@nyu.edu

Abstract: Deep neural networks excel in supervised learning tasks but are constrained by the need for extensive labeled data. Self-supervised learning emerges as a promising alternative, allowing models to learn without explicit labels. Information theory has shaped deep neural networks, particularly the information bottleneck principle. This principle optimizes the trade-off between compression and preserving relevant information, providing a foundation for efficient network design in supervised contexts. However, its precise role and adaptation in self-supervised learning remain unclear. In this work, we scrutinize various self-supervised learning approaches from an information-theoretic perspective, introducing a unified framework that encapsulates the self-supervised information-theoretic learning problem. This framework includes multiple encoders and decoders, suggesting that all existing work on self-supervised learning can be seen as specific instances. We aim to unify these approaches to understand their underlying principles better and address the main challenge: many works present different frameworks with differing theories that may seem contradictory. By weaving existing research into a cohesive narrative, we delve into contemporary self-supervised methodologies, spotlight potential research areas, and highlight inherent challenges. Moreover, we discuss how to estimate information-theoretic quantities and their associated empirical problems. Overall, this paper provides a comprehensive review of the intersection of information theory, self-supervised learning, and deep neural networks, aiming for a better understanding through our proposed unified approach.

Keywords: self-supervised learning; information theory; representation learning; deep neural networks

check for updates

Citation: Shwartz Ziv, R.; LeCun, Y. To Compress or Not to Compress—Self-Supervised Learning and Information Theory: A Review. *Entropy* **2024**, *26*, 252. https://doi.org/10.3390/e26030252

Academic Editors: Shuangming Yang, Shujian Yu, Luis Gonzalo Sánchez Giraldo, Badong Chen and Sotiris Kotsiantis

Received: 6 December 2023
Revised: 12 February 2024
Accepted: 26 February 2024
Published: 12 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep neural networks (DNNs) have revolutionized fields such as computer vision, natural language processing, and speech recognition due to their remarkable performance in supervised learning tasks [1–3]. However, the success of DNNs is often limited by the need for vast amounts of labeled data, which can be both time-consuming and expensive to acquire. By using unlabeled data, supervised learning costs can be reduced, especially in fields that require expensive annotations. As an example, biomedical task labels must be provided by domain experts, who are costly to hire. Besides the hiring cost, labeling tasks are often labor-intensive. For example, video data labels require the review of many frames. Self-supervised learning (SSL) emerges as a promising direction, enabling models to learn from data without explicit labels by leveraging the underlying structure and relationships within the data.

Recent advances in SSL have been driven by joint embedding architectures, such as Siamese Nets [4], DrLIM [5,6], and SimCLR [7]. These approaches define a loss function that encourages representations of different versions of the same image to be similar while pushing representations of distinct images apart. After optimizing the surrogate objective, the pre-trained model can be employed as a feature extractor, with the learned features serving as inputs for downstream supervised tasks, like image classification, object detection, instance segmentation, or pose estimation [7–10]. Although SSL methods have shown promising results in practice, the theoretical underpinnings behind their effectiveness remain an open question [11,12].

Information theory has played a crucial role in understanding and optimizing deep neural networks, from practical applications like the variational information bottleneck [13] to theoretical investigations of generalization bounds induced by mutual information [14,15]. Building upon these foundations, several researchers have attempted to enhance self-supervised and semi-supervised learning algorithms using information-theoretic principles, such as the Mutual Information Neural Estimator (MINE) [16] combined with the information maximization (InfoMax) principle [17]. However, the plethora of objective functions, contradicting assumptions, and various estimation techniques in the literature can make it challenging to grasp the underlying principles and their implications.

In this paper, we aim to achieve two objectives. First, we propose a unified framework that synthesizes existing research on self-supervised and semi-supervised learning from an information-theoretic standpoint. This framework allows us to present and compare current methods, analyze their assumptions and difficulties, and discuss the optimal representation for neural networks in general and self-supervised networks in particular. Second, we explore different methods and estimators for optimizing information-theoretic quantities in deep neural networks and investigate how recent models optimize various theoretical-information terms.

By reviewing the literature on various aspects of information-theoretic learning, we provide a comprehensive understanding of the interplay between information theory, self-supervised learning, and deep neural networks. We discuss the application of the information bottleneck principle [18], connections between information theory and generalization, and recent information-theoretic learning algorithms. Furthermore, we examine how the information-theoretic perspective can offer insights into the design of better self-supervised learning algorithms and the potential benefits of using information theory in SSL across a wide range of applications.

In addition to the main structure of this paper, we dedicate a section to the challenges and opportunities in extending the information-theoretic perspective to other learning paradigms, such as energy-based models. We highlight the potential advantages of incorporating these extensions into self-supervised learning algorithms and discuss the technical and conceptual challenges that must be addressed.

The structure of this paper is as follows. Section 2 introduces the key concepts in supervised, semi-supervised, and self-supervised learning, information theory, and representation learning. Section 3 presents a unified framework for multiview learning based on information theory. We first discuss what an optimal representation is and why compression is beneficial for learning. Next, we explore optimal representation in single-view supervised learning models and how they can be extended to unsupervised, semi-supervised, and multiview contexts. The focus then shifts to self-supervised learning, where the optimal representation remains an open question. Using the unified framework, we compare recent self-supervised algorithms and discuss their differences. We analyze the assumptions behind these models, their effects on the learned representation, and their varying perspectives on important information within the network.

Section 5 addresses several technical challenges, discussing both theoretical and practical issues in estimating theoretical information terms. We present recent methods for estimating these quantities, including variational bounds and estimators. As part of Section 6, we examine a wide range of review papers that cover information theory and self-supervised learning thoroughly. Section 7 concludes this paper by offering insights into potential future research directions at the intersection of information theory, self-supervised learning, and deep neural networks. Our aim is to stimulate further research that leverages information theory to advance our understanding of self-supervised learning and to develop more efficient and effective models for a broad range of applications.

2. Background and Fundamental Concepts

2.1. Multiview Representation Learning

Multiview learning, which utilizes complementary information from multiple features or modalities, has gained increasing attention and achieved great practical success. The multiview learning paradigm divides the input variable into multiple views from which the target variable should be predicted [19]. Using this paradigm, one can eliminate hypotheses that contradict predictions from other views and provide a natural semi-supervised and self-supervised learning setting. A multiview dataset consists of data captured from multiple sources, modalities, and forms but with similar high-level semantics [20]. This mechanism was initially used for natural-world data, combining image, text, audio, and video measurements. For example, photos of objects are taken from various angles, and our supervised task is to identify the objects. Another example is identifying a person by analyzing the video stream as one view and the audio stream as the other.

Although these views often provide different and complementary information about the same data, directly integrating them does not produce satisfactory results due to biases between multiple views [20]. Thus, multiview representation learning involves identifying the underlying data structure and integrating the different views into a common feature space, resulting in a high performance. In recent decades, multiview learning has been used for many machine learning tasks and influenced many algorithms, such as co-training mechanisms [21], subspace learning methods [22], and multiple kernel learning (MKL) [23]. Li et al. [24] proposed two categories for multiview representation learning: (i) multiview representation fusion, which combines different features from multiple views into a single compact representation, and (ii) the alignment of multiview representation, which attempts to capture the relationships among multiple different views through feature alignment. In this case, a learned mapping function embeds the data of each view, and the representations are regularized to form a multiview-aligned space. In this research direction, an early study is the canonical correlation analysis (CCA) [25] and its kernel extensions [23,26,27]. In addition to CCA, multiview representation learning has penetrated a variety of learning methods, such as dimensionality reduction [28], clustering analysis [29], multiview sparse coding [30–32], and multimodal topic learning [33]. However, despite their promising results, these methods use handcrafted features and linear embedding functions, which cannot capture the nonlinear properties of multiview data.

Deep learning provides a powerful way to learn complex, nonlinear, and hierarchical representations of data. By incorporating multiple hierarchical layers, deep learning algorithms can learn complex, subtle, and abstract representations of target data. The success of deep learning in various application domains has led to a growing interest in deep multiview methods, which have shown promising results. Examples of these methods include deep multiview canonical correlation analysis [34] as an extension of CCA, multiview clustering via deep matrix factorization [35], and the deep multiview spectral network [36]. Moreover, deep architectures have been employed to generate effective representations in methods such as multiview convolutional neural networks [37], multimodal deep Boltzmann machines [38], multimodal deep autoencoders [39,40], and multimodal recurrent neural networks [41–43].

2.2. Self-Supervised Learning

Self-supervised learning (SSL) is a powerful technique that leverages unlabeled data to learn useful representations. In contrast to supervised learning, which relies on labeled data, SSL employs self-defined signals to establish a proxy objective between the input and the signal. The model is initially trained using this proxy objective and subsequently fine-tuned on the target task. Self-supervised signals, derived from the inherent co-occurrence relationships in the data, serve as self-supervision. Various such signals have been used to learn representations, including generative and joint embedding architectures [7,44–47].

Two main categories of SSL architectures exist: (1) generative architectures based on reconstruction or prediction and (2) joint embedding architectures [48]. Both architecture classes can be trained using either contrastive or non-contrastive methods.

We begin by discussing these two main types of architectures:

1. **Generative architecture:** Generative architectures employ an objective function that measures the divergence between input data and predicted reconstructions, such as squared error. The architecture reconstructs data from a latent variable or a corrupted version, potentially with a latent variable's assistance. Notable examples of generative architectures include autoencoders, sparse coding, sparse autoencoders, and variational autoencoders [49–51]. As the reconstruction task lacks a single correct answer, most generative architectures utilize a latent variable, which, when varied, generates multiple reconstructions. The latent variable's information content requires regularization to ensure the system reconstructs regions of high data density while avoiding a collapse by reconstructing the entire space. PCA regularizes the latent variable by limiting its dimensions, while sparse coding and sparse autoencoders restrict the number of non-zero components. Variational autoencoders regularize the latent variable by rendering it stochastic and maximizing the entropy of the distribution relative to a prior. Vector quantized variational autoencoders (VQ-VAEs) employ binary stochastic variables to achieve similar results [52].
2. **Joint embedding architectures (JEAs):** These architectures process multiple views of an input signal through encoders, producing representations of the views. The system is trained to ensure that these representations are both informative and mutually predictable. Examples include Siamese networks, where two identical encoders share weights [7,53–55], and methods permitting encoders to differ [56]. A primary challenge with JEA is preventing informational collapse, in which the representations contain minimal information about the inputs, thereby facilitating their mutual prediction. JEA's advantage lies in the encoders' ability to eliminate noisy, unpredictable, or irrelevant information from the input within the representation space.

To train these architectures effectively, it is essential to ensure that the representations of different signals are distinct. This can be achieved through either contrastive or non-contrastive methods:

- **Contrastive methods:** Contrastive methods utilize data points from the training set as *positive samples* and generate points outside the region of high data density as *contrastive samples*. The energy (e.g., reconstruction error for generative architectures or representation predictive error for JEA) should be low for positive samples and higher for contrastive samples. Various loss functions involving the energies of pairs or sets of samples can be minimized to achieve this objective.
- **Non-contrastive methods:** Non-contrastive methods prevent the energy landscape's collapse by limiting the volume of space that can take low energy, either through architectural constraints or through a regularizer in the energy or training objective. In latent-variable generative architectures, preventing collapse is achieved by limiting or minimizing the information content of the latent variable. In JEA, collapse is prevented by maximizing the information content of the representations.

We now present a few concrete examples of popular models that employ various combinations of generative architectures, joint embedding architectures, contrastive training, and non-contrastive training:

The **denoising autoencoder** approach in generative architectures [57–59] uses a triplet loss, which utilizes a positive sample, which is a vector from the training set that should be reconstructed perfectly, and a contrastive sample consisting of data vectors, one from the training set and the other being a corrupted version of it. In SSL, the combination of *JEA* models with *contrastive learning* has proven highly effective. In contrastive learning, the objective is to attract different augmented views of the same image (positive points) while repelling dissimilar augmented views (negative points). Recent self-supervised visual representation learning examples include MoCo [54] and SimCLR [7]. The InfoNCE loss is a commonly used objective function in many contrastive learning methods:

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{\sum k=1^K e^{f(x)^T f(x^k)}}\right)\right] \tag{1}$$

where $x+$ is a sample similar to x, x^k are all the samples in the batch, and f is an encoder. This loss, inspired by NCE [60], uses categorical cross-entropy loss to distinguish the positive sample amongst a set of unrelated noise samples in the batch. In this formulation, the numerator represents the output of a positive pair, while the denominator sums the values of both positive and negative pairs. This straightforward loss function aims to increase the value of positive pairs (driving the logarithmic term towards 1, thereby reducing the loss towards 0) and separate the negative pairs further.

However, contrastive methods heavily depend on all other samples in the batch and require a large batch size. Additionally, Jing et al. [61] have shown that contrastive learning can lead to dimensional collapse, where the embedding vectors span a lower-dimensional subspace instead of the entire embedding space. Although positive and negative pairs should repel each other to prevent dimensional collapse, augmentation along feature dimensions and implicit regularization cause the embedding vectors to fall into a lower-dimensional subspace, resulting in low-rank solutions.

To address these problems, recent works have introduced *JEA* models with *non-contrastive methods*. Unlike contrastive methods, these methods employ regularization to prevent the collapse of the representation and do not explicitly rely on negative samples. For example, several papers use stop gradients and extra predictors to avoid collapse [53,55], while Caron et al. [62] employed an additional clustering step. VICReg [56] is another non-contrastive method that regularizes the covariance matrix of representation. Consider two embedding batches $\boldsymbol{Z} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]$ and $\boldsymbol{Z}' = \left[f(\boldsymbol{x}'_1), \ldots, f(\boldsymbol{x}'_N)\right]$, each of size $(N \times K)$. Denote by $\boldsymbol{C}$ the $(K \times K)$ covariance matrix obtained from $[\boldsymbol{Z}, \boldsymbol{Z}']$. The VICReg triplet loss is defined by the following:

$$\mathcal{L}= \frac{1}{K}\sum_{k=1}^{K}\left(\alpha \max\left(0, \gamma - \sqrt{\boldsymbol{C}_{k,k} + \epsilon}\right)+\beta \sum_{k' \neq k}\left(\boldsymbol{C}_{k,k'}\right)^2\right) \quad + \gamma\|\boldsymbol{Z} - \boldsymbol{Z}'\|_F^2/N. \tag{2}$$

The variance loss (the diagonal terms) encourages high variance in the learned representations, thereby promoting the learning of a wide range of features. The covariance loss (the off-diagonal terms), however, aims to minimize redundancy in the learned features by reducing the overlap in information captured by different dimensions of the representation.

2.3. Semi-Supervised Learning

Semi-supervised learning employs both labeled and unlabeled data to enhance the model performance [63]. Consistency regularization-based approaches [64–66] ensure that predictions remain stable under perturbations in input data and model parameters. Certain techniques, such as those proposed by Grandvalet and Bengio [67] and Miyato et al. [65], involve training a model by incorporating a regularization term into a supervised cross-entropy loss. In contrast, Xie et al. [68] utilizes suitably weighted unsupervised regularization terms, while Zhai et al. [69] adopts a combination of self-supervised pretext loss terms. Moreover, pseudo-labeling can generate synthetic labels based on network uncertainty to further aid model training [70].

2.4. Representation Learning

Representation learning is an essential aspect of various computer vision, natural language processing, and machine learning tasks, as it uncovers the underlying structures in data [71]. Extracting relevant information for classification and prediction tasks from the data improves the performance and reduces computational complexity [72]. However, defining an effective representation remains a challenging task. In probabilistic models, a useful representation often captures the posterior distribution of explanatory factors beneath the observed input [2]. Bengio and LeCun [73] introduced the idea of learning highly structured yet complex dependencies for AI tasks, which require transforming high-dimensional input structures into low-dimensional output structures or learning low-level representations. Consequently, identifying relevant input features is challenging because most input entropy does not relate to the output. [74]. Ben-Shaul et al. [75] demonstrated that self-supervised learning inherently promotes the clustering of samples based on semantic labels. Intriguingly, this clustering is driven by the objective's regularization term and aligns with semantic classes across multiple hierarchical levels.

2.4.1. Minimal Sufficient Statistic

A possible definition of an effective representation is based on *minimal sufficient statistics.*

Definition 1. *Given $(X, Y) \sim P(X, Y)$, let $T := t(X)$, where t is a deterministic function. We define T as a sufficient statistic of X for Y if $Y - T - X$ forms a Markov chain.*

A sufficient statistic is defined relative to the statistics of the data, or a probabilistic function, and provides all the information in the data about that model or the parameters of that model.

Intuitively, a sufficient statistic captures all the information about Y in X. Cover [76] proved this property:

Theorem 1. *Let T be a probabilistic function of X. Then, T is a sufficient statistic for Y if and only if $I(T(X); Y) = I(X; Y)$.*

However, the sufficiency definition also encompasses trivial identity statistics that only "copy" rather than "extract" essential information. To prevent statistics from inefficiently utilizing observations, the concept of minimal sufficient statistics was introduced:

Definition 2. *(Minimal sufficient statistic (MSS).) A sufficient statistic T is minimal if, for any other sufficient statistic S, there exists a function f such that $T = f(S)$ almost surely (a.s.).*

In essence, MSSs are the simplest sufficient statistics, inducing the coarsest sufficient partition on X. In MSSs, the values of X are grouped into as few partitions as possible without sacrificing information. MSSs are statistics with the maximum information about Y while retaining the least information about X as possible [77].

2.4.2. The Information Bottleneck

The majority of distributions lack exact minimal sufficient statistics, leading Tishby et al. [18] to relax the optimization problem in two ways: (i) allowing the map to be stochastic, defined as an encoder $P(T|X)$, and (ii) permitting the capture of only a small amount of $I(X; Y)$. The information bottleneck (IB) was introduced as a principled method to extract relevant information from observed signals related to a target. This framework finds the optimal trade-off between the accuracy and complexity of a random variable $y \in \mathcal{Y}$ with a joint distribution for a random variable $x \in \mathcal{X}$. The IB has been employed in various fields, such as neuroscience [78,79], slow feature analysis [80], speech recognition [81], molecular relational learning [82], and deep learning [13,74].

Let X be an input random variable, Y a target variable, and $P(X, Y)$ their joint distribution. A representation T is a stochastic function of X defined by a mapping $P(T \mid X)$. This mapping transforms $X \sim P(X)$ into a representation of $T \sim P(T) := \int P_{T|X}(\cdot \mid x) dP_X(x)$. The triple $Y - X - T$ forms a Markov chain in that order with respect to the joint probability measure $P_{X,Y,T} = P_{X,Y} P_{T|X}$ and the mutual information terms $I(X; T)$ and $I(Y; T)$.

Within the IB framework, our goal is to find a representation $P(T \mid X)$ that extracts as much information as possible about Y (high performance) while compressing X maximally (keeping $I(X;T)$ small). This can also be interpreted as extracting only the relevant information that X contains about Y.

The data processing inequality (DPI) implies that $I(Y;T) \leq I(X;Y)$, so the compressed representation T cannot convey more information than the original signal. Consequently, there is a trade-off between compressed representation and the preservation of relevant information about Y. The construction of an efficient representation variable is characterized by its encoder and decoder distributions, $P(T \mid X)$ and $P(Y \mid T)$, respectively. The efficient representation of X involves minimizing the complexity of the representation $I(T;X)$ while maximizing $I(T;Y)$. Formally, the IB optimization involves minimizing the following objective function:

$$\mathcal{L} = \min_{P(t|x);p(y|t)} I(X;T) - \beta I(Y;T)\,, \tag{3}$$

where β is the trade-off parameter controlling the complexity of T and the amount of relevant information it preserves. Intuitively, we pass the information that X contains about Y through a "bottleneck" via the representation T. It has been shown that

$$I(T;Y) = I(X;Y) - \mathbb{E}_{x\sim P(X),t\sim P(T|x)}[D[P(Y|x)||P(Y|t)]]. \tag{4}$$

2.5. Representation Learning and the Information Bottleneck

Information theory traditionally assumes that underlying probabilities are known and do not require learning. For instance, the optimality of the initial IB work [18] relied on the assumption that the joint distribution of input and labels is known. However, a significant challenge in machine learning algorithms is inferring an accurate predictor for the unknown target variable from observed realizations. This discrepancy raises questions about the practical optimality of the IB and its relevance in modern learning algorithms. The following section delves into the relationship between the IB framework and learning, inference, and generalization.

Let $X \in \mathcal{X}$ and a target variable $Y \in \mathcal{Y}$ be random variables with an unknown joint distribution $P(X,Y)$. For a given class of predictors $f : \mathcal{X} \to \hat{\mathcal{Y}}$ and a loss function $\ell : \mathcal{Y} \to \hat{\mathcal{Y}}$ measuring discrepancies between true values and model predictions, our objective is to find the predictor f that minimizes the *expected population risk*:

$$\mathcal{L}_{P(X,Y)}(f,\ell) = \mathbb{E}_{P(X,Y)}[\ell(Y,f(X))]. \tag{5}$$

Several issues arise with the expected population risk. Firstly, it remains unclear which loss function is optimal. A popular choice is the logarithmic loss (or error's entropy), which has been numerically demonstrated to yield better results [83]. This loss has been employed in various algorithms, including the InfoMax principle [17], tree-based algorithms [84], deep neural networks [85], and Bayesian modeling [86]. Painsky and Wornell [87] provided a rigorous justification for using the logarithmic loss and showed that it is an upper bound to any choice of the loss function that is smooth, proper, and convex for binary classification problems.

In most cases, the joint distribution $P(X,Y)$ is unknown, and we have access to only n samples from it, denoted by $\mathcal{D}_n := (x_i, y_i) \mid i = 1, \ldots, n$. Consequently, the population risk cannot be computed directly. Instead, we typically choose the predictor that minimizes the empirical population risk on a training dataset:

$$\hat{\mathcal{L}}_{P(X,Y)}(f,\ell,\mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n}[\ell(y_i,f(x_i))]. \tag{6}$$

The generalization gap, defined as the difference between empirical and population risks, is given by

$$Gen_{P(X,Y)}(f,\ell,\mathcal{D}_n) := \mathcal{L}_{P(X,Y)}(f,\ell) - \hat{\mathcal{L}}_{P(X,Y)}(f,\ell,\mathcal{D}_n). \tag{7}$$

Interestingly, the relationship between the population risk and the empirical risk can be bounded using the information bottleneck term. Shamir et al. [88] developed several finite sample bounds for the generalization gap. According to their study, the IB framework exhibited good generalizability even with small sample sizes. In particular, they developed non-uniform bounds adaptive to the model's complexity. They demonstrated that for the discrete case, the error in estimating mutual information from finite samples is bounded by $O\left(\frac{|X|\log n}{\sqrt{n}}\right)$, where $|X|$ is the cardinality of X (the number of possible values that the random variable X can take). The results support the intuition that simpler models generalize better, and we would like to compress our model. Therefore, optimizing Equation (3) presents a trade-off between two opposing forces. On the one hand, we want to increase our prediction accuracy in our training data (high β).

On the other hand, we would like to decrease β to narrow the generalization gap. Vera et al. [89] extended their work and showed that the generalization gap is bounded by the square root of mutual information between training input and model representation times $\frac{\log n}{n}$. Furthermore, Russo and Zou [90] and Xu and Raginsky [14] demonstrated that the square root of the mutual information between the training input and the parameters inferred from the training algorithm provides a concise bound on the generalization gap. However, these bounds critically depend on the Markov operator that maps the training set to the network parameters, whose characterization is not trivial.

Achille and Soatto [91] explored how applying the IB objective to the network's parameters may reduce overfitting while maintaining invariant representations. Their work showed that flat minima, which have better generalization properties, bound the information with the weights, and the information in the weights bound the information in the activations. Chelombiev et al. [92] found that the generalization precision is positively correlated with the degree of compression of the last layer in the network. Shwartz-Ziv et al. [93] showed that the generalization error depends exponentially on the mutual information between the model and the input once it is smaller than $\log 2n$—the query sample complexity. Moreover, they demonstrated that M bits of compression of X are equivalent to an exponential factor of 2^M training examples. Piran et al. [94] extended the original IB to the dual form, which offers several advantages in terms of compression. As an example, when the data can be modeled in a parametric form, the dual IB preserves this structure and obtains the representation based on the original parameters, resulting in a more efficient compression.

These studies illustrate that the IB leads to a trade-off between prediction and complexity, even for the empirical distribution. With the IB objective, we can design estimators to find optimal solutions for different regimes with varying performances, complexity, and generalization.

3. Information-Theoretic Objectives

Before delving into the details, this section aims to provide an overview of the information-theoretic objectives in various learning scenarios, including supervised, unsupervised, and self-supervised settings. We will also introduce a general framework to understand better the process of learning optimal representations and explore recent methods working towards this goal.

Developing a novel algorithm entails numerous aspects, such as architecture, initialization parameters, learning algorithms, and pre-processing techniques. A crucial element, however, is the objective function. As demonstrated in Section 2.4.2, the IB approach, originally introduced by Tishby et al. [18], defines the optimal representation in supervised scenarios, enabling us to identify which terms to compress during learning by explicitly defining the relevant information $I(T;Y)$ that we want to optimize. However, determining the optimal representation and deriving information-based objective functions in self-supervised settings are more challenging. In this section, we introduce a general framework to understand the process of learning optimal representations and explore recent methods striving to achieve this goal.

3.1. Setup and Methodology

Using a two-channel input allows us to model complex multiview learning problems. In many real-world situations, data can be observed from multiple perspectives or modalities, making it essential to develop learning algorithms capable of handling such multiview data.

Consider a two-channel input, X_1 and X_2, and a single-channel label Y for a downstream task, all possessing a joint distribution $P(X_1, X_2, Y)$. We assume the availability of n labeled examples $S = (x_1^i, x_2^i, y^i)_{i=1}^{n}$ and t unlabeled examples $U = (x_1^i, x_2^i)_{i=n+1}^{n+t}$, both independently and identically distributed. Our objective is to predict Y using a loss function.

In our model, we use a learned encoder with a prior $P(Z)$ to generate a conditional representation (which may be deterministic or stochastic) $Z_i|X_i = P_{\theta_i}(Z_i|X_i)$, where $i = 1, 2$ represents the two views. Subsequently, we utilize various decoders to "decode" distinct aspects of the representation.

For the supervised scenario, we have a joint embedding of the label classifiers from both views, $\hat{Y}_{1,2} = Q_\rho(Y|Z_1, Z_2)$, and two decoders predicting the labels of the downstream task based on each individual view, $\hat{Y}_i = Q_{\rho_i}(Y|Z_i)$ for $i = 1, 2$.

For the unsupervised case, we have direct decoders for input reconstruction from the representation, $\bar{X}_i = Q_{\psi_i}(X_i|Z_i)$ for $i = 1, 2$.

For self-supervised learning, we utilize two cross-decoders, $\tilde{Z}_1|Z_2 = q_{\eta_1}(Z_1|Z_2)$ and $\tilde{Z}_2|Z_1 = q_{\eta_2}(Z_2|Z_1)$, attempting to predict one representation based on the other. Figure 1 illustrates this structure.

The information-theoretic perspective of self-supervised networks has led to confusion regarding the information being optimized in recent work. In supervised and unsupervised learning, only one "information path" exists when optimizing information-theoretic terms: the input is encoded through the network, and then the representation is decoded and compared to the targets. As a result, the representation and corresponding information always stem from a single encoder and decoder.

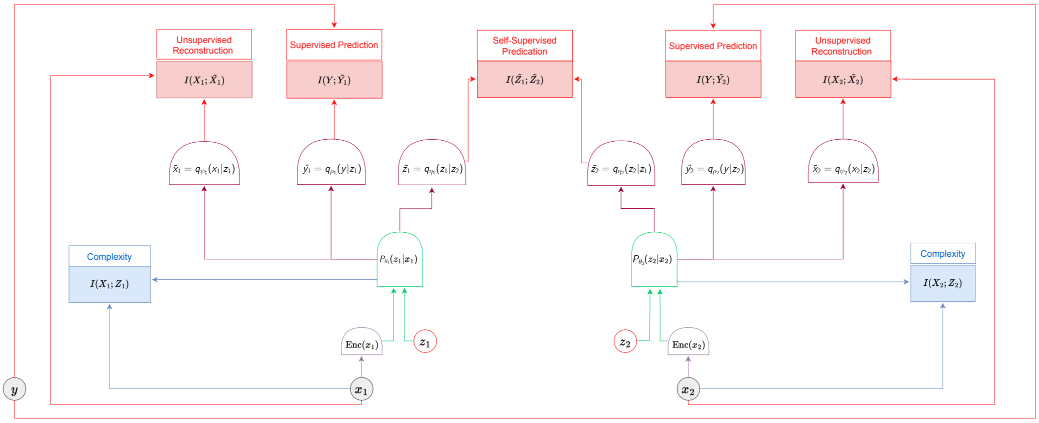
Unsupervised Reconstruction
$I(X_1;\tilde{X}_1)$
Supervised Prediction
$I(Y;\tilde{Y}_1)$
Self-Supervised Predication
$I(\tilde{Z}_1;\tilde{Z}_2)$
Supervised Prediction
$I(Y;\tilde{Y}_2)$
Unsupervised Reconstruction
$I(X_2;\tilde{X}_2)$
$\tilde{x}_1 = q_{\psi_1}(x_1|z_1)$
$\hat{y}_1 = q_{\rho_1}(y|z_1)$
$\tilde{z}_1 = q_{\eta_1}(z_1|z_2)$
$\tilde{z}_2 = q_{\eta_2}(z_2|z_1)$
$\hat{y}_2 = q_{\rho_2}(y|z_2)$
$\tilde{x}_2 = q_{\psi_2}(x_2|z_2)$
Complexity
$I(X_1;Z_1)$
$P_{\theta_1}(z_1|x_1)$
$P_{\theta_2}(z_2|x_2)$
Complexity
$I(X_2;Z_2)$
$\mathrm{Enc}(x_1)$
z_1
z_2
$\mathrm{Enc}(x_2)$
y
x_1
x_2

Figure 1. Multiview information bottleneck diagram for self-supervised, unsupervised, and supervised learning.

However, in the self-supervised multiview scenario, we can construct our representation using various encoders and decoders. For instance, we need to specify the associated random variable to define the information involved in $I(X_1; Z_1)$. This variable could either be based on the encoder of $X_1 - P_{\theta_1}(Z_1|X_1)$ or based on the encoder of $X_2 - P_{\theta_2}(Z_2|X_2)$, which is subsequently passed to the cross-decoder $Q_{\eta_1}(Z_1|Z_2)$ and then to the direct decoder $Q_{\psi_1}(X_1|Z_1)$.

To fully understand the information terms, we aim to optimize and distinguish between various "information path"; we marked each information path differently. For example, $I_{,P(X_1),P(Z_1|X_1),P(Z_2|Z_1)}(X_1, Z_2)$ is based on the path $P(X_1) \rightarrow P(Z_1|X_1) \rightarrow P(Z_2|Z_1)$.

In the following section, we will "translate" previous work into our present framework and examine the loss function.

3.2. Optimization with Labels

After establishing our framework, we can now incorporate various learning algorithms. We begin by examining classical single-view supervised information bottleneck algorithms for deep networks that utilize labeled data during training and extend them to the multiview scenario. Next, we broaden our perspective to include unsupervised learning, where input reconstruction replaces labels, and semi-supervised learning, where information-based regularization is applied to improve predictions.

3.2.1. Single-View Supervised Learning

In classical single-view supervised learning, the task of representation learning involves finding a distribution $p(z|x)$ that maps data observations $x \in \mathcal{X}$ to a representation $z \in \mathcal{Z}$, capturing only the relevant features of the input [95]. The goal is to predict a label $y \in \mathcal{Y}$ using the learned representation. Achille and Soatto [91] defined the sufficiency of Z for Y as the amount of label information retained after passing data through the encoder:

Definition 3. ***Sufficiency****: A representation Z of X is sufficient for Y if and only if $I(X;Y|Z) = 0$.*

Federici et al. [96] showed that Z is sufficient for Y if and only if the amount of information regarding the task remains unchanged by the encoding procedure.

$$I(X;Y|Z) = 0 \Leftrightarrow I(X;Y) = I(Y;Z). \tag{8}$$

A sufficient representation can predict Y as accurately as the original data X. In Section 2.4, we saw a trade-off between prediction and generalization when there is a finite amount of data. To reduce the generalization gap, we aim to compress X while retaining as much predicate information on the labels as possible. Thus, we relax the sufficiency definition and minimize the following objective:

$$\mathcal{L} = I(X;Z) - \beta I(Z;Y). \tag{9}$$

The mutual information $I(Y;Z)$ determines how much label information is accessible and reflects the model's ability to predict performance on the target task. $I(X;Z)$ represents the information that Z carries about the input, which we aim to compress. However, $I(X;Z)$ contains both relevant and irrelevant information about Y. Therefore, using the chain rule of information, Federici et al. [96] proposed splitting $I(X,Z)$ into two terms:

$$I(X;Z) = \underbrace{I(X;Z|Y)}_{\text{superfluous information}} + \underbrace{I(Z;Y)}_{\text{predictive information}}. \tag{10}$$

The conditional information $I(X,Z|Y)$ represents information in Z that is not predictive of Y, i.e., superfluous information. The decomposition of input information enables us to compress only irrelevant information while preserving the relevant information for predicting Y. Several methods are available for evaluating and estimating these information-theoretic terms in the supervised case (see Section 5 for details).

3.2.2. The Information Bottleneck Theory of Deep Learning

The IB hypothesis for deep learning proposes two distinct phases of training neural networks [74]: the fitting and compression phases. The fitting phase involves extracting information from the input and converting it into learned representations, characterized by increased mutual information between inputs and hidden representations. Conversely, the compression phase, which is much longer, concentrates on discarding unnecessary information for target prediction, decreasing mutual information between learned representations and inputs. In contrast, the mutual information between representations and targets increases. For more information, see Geiger [97]. Despite the elegance and plausibility of the IB hypothesis, empirically investigating it remains challenging [98].

The study of representation compression in deep neural networks (DNNs) for supervised learning has shown inconsistent results. For instance, Chelombiev et al. [92] discovered a positive correlation between generalization accuracy and the compression level of the network's final layer. Shwartz-Ziv et al. [93] also examined the relationship between generalization and compression, demonstrating that generalization error exponentially depends on mutual information, $I(X;Z)$. Furthermore, Achille et al. [99] established that flat minima, known for their improved generalization properties, constrain the mutual information. However, Saxe et al. [100] showed that compression was not necessary for generalization in deep linear networks. Basirat et al. [101] revealed that the decrease in mutual information is essentially equivalent to geometrical compression. Other studies have found that the mutual information between training inputs and inferred parameters provides a concise bound on the generalization gap [14,102]. Lastly, Achille and Soatto [91] explored using an information bottleneck objective on network parameters to prevent overfitting and promote invariant representations.

3.2.3. Multiview IB Learning

The IB principle offers a rigorous method for learning encoders and decoders in supervised single-view problems. However, it is not directly applicable to multiview learning problems, as it assumes only one information source as the input. A common solution is to concatenate multiple views, though this neglects the unique characteristics of each view. To address this issue, Xu et al. [103] introduced the large-margin multiview IB (LMIB) as an extension of the original IB problem. The LMIB employs a communication system where multiple senders represent various views of examples. The system extracts specific components from different senders by compressing examples through a "bottleneck", and the linear projectors for each view are combined to create a shared representation. The large-margin principle replaces the maximization of mutual information in prediction, emphasizing the separation of samples from different classes. For the complexity, they used the Rademacher complexity, which is defined as follows:

Definition 4. ***Rademacher complexity*** *Given a sample $S = \{X_1, \dots, X_n\} \in \mathcal{X}^n$ and a real-valued function class $\mathcal{F}$ defined on a space $\mathcal{X}$, the empirical Rademacher complexity of $\mathcal{F}$ is defined as*

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{j=1}^{n} \sigma_j f(X_j) \right| \middle| X_1, \dots, X_n \right],$$

where $\sigma = (\sigma_1, \dots, \sigma_n)$ are i.i.d. Rademacher variables (taking values $+1$ or -1 with equal probability).

In the LMIB framework, limiting the Rademacher complexity improves the solution's accuracy and generalization error bounds. Moreover, the algorithm's robustness is enhanced when accurate views counterbalance noisy views.

However, the LMIB method has a significant limitation: it utilizes linear projections for each view, which can restrict the combined representation when the relationship between different views is complex. To overcome this limitation, Wang et al. [104] proposed using deep neural networks to replace linear projectors. Their model first extracts concise latent representations from each view using deep networks and then learns the joint representation of all views using neural networks. They minimize the objective:

$$\mathcal{L} = \alpha I_{P(X_1),P(Z_1|X_1)}(X_1;Z_1) + \beta I_{P(X_2),P(Z_2|X_2)}(X_2;Z_2) - I_{P(Z_2|X_2),P(Z_2|X_1)}(Z_{1,2};Y). \tag{11}$$

Here, α and β are trade-off parameters, Z_1 and Z_2 are the two neural networks' representations, and $Z_{1,2}$ is the joint embedding of Z_1 and Z_2. The first two terms decrease the mutual information between a view's latent representation and its original data representation, resulting in a simpler and more generalizable model. The final term forces the joint representation to maximize the discrimination ability for the downstream task.

3.2.4. Semi-Supervised IB Learning: Leveraging Unlabeled Data

Obtaining labeled data can be challenging or expensive in many practical scenarios, while many unlabeled samples may be readily available. Semi-supervised learning addresses this issue by leveraging the vast amount of unlabeled data during training in conjunction with a small set of labeled samples. Common strategies to achieve this involve adding regularization terms or adopting mechanisms that promote better generalization. Berthelot et al. [105] grouped regularization methods into three primary categories: entropy minimization, consistency regularization, and generic regularization.

Voloshynovskiy et al. [106] introduced an information-theoretic framework for semi-supervised learning based on the IB principle. In this context, the semi-supervised classification problem involves encoding input X into the latent space Z while preserving only **class-relevant information**. A supervised classifier can achieve this if there are sufficient labeled data. However, when the number of labeled examples is limited, the standard label classifier $p(y|z)$ becomes unreliable and requires regularization.

To tackle this issue, the authors assumed a prior on the class label distribution $p(y)$. They introduced a term to minimize the D_{KL} between the assumed marginal prior and the empirical marginal prior, effectively regularizing the conditional label classifier with the labels' marginal distribution. This approach reduces the classifier's sensitivity to the scarcity of labeled examples. They proposed two variational IB semi-supervised extensions for the priors:

Handcrafted priors: These priors are predefined for regularization and can be based on domain knowledge or statistical properties of the data. Alternatively, they can be learned using other networks. Handcrafted priors in this context are similar to priors used in the variational information bottleneck (VIB) formalism [13,104].

Learnable priors: Voloshynovskiy et al. [106] also suggests using learnable priors as an alternative to handcrafted regularization priors on the latent representation. This method involves regularizing Z through another IB-based regularization with two components: (i) latent space regularization and (ii) observation space regularization. In this case, an additional hidden variable M is introduced after the representation to regulate the information flow between Z and Y. An autoencoder $q(m|z)$ is employed, and the optimization process aims to compress the information flowing from Z to M while retaining only label-relevant information. The IB objective is defined as follows:

$$\begin{aligned}\mathcal{L} &= D_{KL}(q(m|z)||p(m|z)) - \beta D_{KL}(q(x|m)||p(x|m)) - \beta_y D_{KL}(p(y|z)||p(y)) \\ &\Leftrightarrow I(M;Z) - \beta I(M;X) - \beta_y I(Y;Z).\end{aligned} \tag{12}$$

Here, β and β_y are hyperparameters that balance the trade-off between the relevance of M to the labels and the compression of Z into M.

Furthermore, Voloshynovskiy et al. [106] demonstrated that various popular semi-supervised methods can be considered special cases of the optimization problem described above. Notably, the semi-supervised AAE [107], CatGAN [108], SeGMA [109], and VAE [110] can all be viewed as specific instantiations of this framework.

3.2.5. Unsupervised IB Learning

In the unsupervised setting, data samples are not directly labeled by classes. Voloshynovskiy et al. [106] defined the unsupervised IB as a "compressed" parameterized mapping of X to Z, which preserves some information in Z about X through the reverse decoder $\bar{X} = Q(X|Z)$. Therefore, the Lagrangian of the unsupervised IB can be defined as follows:

$$I_{P(X),P(Z|X)}(X;Z) - \beta I_{P(Z),Q(X|Z)}(Z;\bar{X}), \tag{13}$$

where $I(X;Z)$ is the information determined by the encoder $q(z|x)$ and $I(Z;\bar{X})$ is the information determined by the decoder $q(x|z)$, i.e., the reconstruction error. In other words, the unsupervised IB is a special case of the supervised IB, where labels are replaced with the reconstruction performance of the training input. Alemi et al. [13] showed that the variational autoencoder (VAE) [111] and β-VAE [112] are special cases of the unsupervised variational IB. Voloshynovskiy et al. [106] extended their results and showed that many models, including adversarial autoencoders [107], InfoVAEs [113], and VAE/GANs [114], could be viewed as special cases of the unsupervised IB. The main difference between them is the bounds on the different mutual information of the IB. Furthermore, the unsupervised IB was used by Uğur et al. [115] to derive lower bounds for their unsupervised generative clustering framework, while Roy et al. [116] used it to study vector-quantized autoencoders.

Voloshynovskiy et al. [106] pointed out that for the classification task in the supervised IB, the latent space Z should have sufficient statistics for Y, whose entropy is much lower than X. This results in a highly compressed representation where sequences close in the input space might be close in the latent space, and the less significant features will be compressed. In contrast, in the unsupervised setup, the IB suggests compressing the input to the encoded representation so that each input sequence can be decoded uniquely. In this case, the latent space's entropy should correspond to the input space's entropy, and compression is much more difficult.

4. Self-Supervised Multiview Information Bottleneck Learning

How can we learn without labels and still achieve good predictive power? Is compression necessary to obtain an optimal representation? This section analyzes and discusses how to achieve an optimal representation for self-supervised learning when labels are not available during training. We review recent methods for self-supervised learning and show how they can be integrated into a single framework. We compare their objective functions, implicit assumptions, and theoretical challenges. Finally, we consider the information-theoretic properties of these representations, their optimality, and different ways of learning them.

One approach to enhance deep learning methods is to apply the *InfoMax principle* in a multiview setting [17,117]. As one of the earliest approaches, Linsker [17] proposed maximizing the information transfer from input data to its latent representation, showing its equivalence to maximizing the determinant of the output covariance under the Gaussian distribution assumption. Becker and Hinton [118] introduced a representation learning approach based on maximizing an approximation of the mutual information between alternative latent vectors obtained from the same image. The most well-known application is the Independent Component Analysis (ICA) InfoMax algorithm [119], designed to separate independent sources from their linear combinations. The ICA-InfoMax algorithm aims to maximize the mutual information between mixtures and source estimates while imposing statistical independence among outputs. The Deep InfoMax approach [120] extends this idea to unsupervised feature learning by maximizing the mutual information between input and output while matching a prior distribution for the representations. Recent work has applied this principle to a self-supervised multiview setting [46,120–122], wherein these works maximize the mutual information between the views Z_1 and Z_2 using the classifier $q(z_1|z_2)$, which attempts to predict one representation from the other.

However, Tschannen et al. [123] demonstrated that the effectiveness of InfoMax models is more attributable to the inductive biases introduced by the architecture and estimators than to the training objectives themselves, as the InfoMax objectives can be trivially maximized using invertible encoders. Moreover, a fundamental issue with the *InfoMax principle* is that it retains irrelevant information about the labels, contradicting the core concept of the IB principle, which advocates compressing the representation to enhance generalizability.

Going beyond the InfoMax principle requires us to tackle the crucial question: How do we discern between relevant and irrelevant information? A foundational concept in addressing this challenge is partial information decomposition (PID), as outlined by Williams and Beer [124] and further explored in Gutknecht et al. [125]. PID provides an elegant framework for categorizing the information provided by a set of source variables about a target variable into distinct types: unique, shared (redundant), and synergistic information.

PID enables us to leverage the complexity of information interactions. For example, shared information refers to common information across multiple sources, valuable for tasks that aggregate information from multiple sources or utilize redundancy in the data. Unique information, conversely, is knowledge exclusive to a specific source, enhancing the diversity of representations, and is particularly useful for tasks requiring specialized knowledge. Synergistic information arises from the combination of sources, unveiling insights unattainable when sources are considered individually.

For example, Sridharan and Kakade [126] proposed the *multiview IB framework*, which uses the shared information as the way to compress. According to this framework, in the multiview without labels setting, the IB principle of preserving relevant data while compressing irrelevant data requires assumptions regarding the relationship between views and labels. They presented the *multiview assumption*, which asserts that either view (approximately) would be sufficient for downstream tasks. By this assumption, they define the relevant information as the shared information between the views. Therefore, augmentations (such as changing the image style) should not affect the labels.

Additionally, the views will provide most of the information in the input regarding downstream tasks. We improve generalization without affecting the performance by compressing the information not shared between the two views. Their formulation is as follows:

Assumption 1. *The* ***multiview assumption:*** *There exists a ϵ_{info} (which is assumed to be small) such that*

$$I(Y; X_2|X_1) \leq \epsilon_{info}$$
$$I(Y; X_1|X_2) \leq \epsilon_{info}.$$

As a result, when the information sharing parameter, ϵ_{info}, is small, the information shared between views includes task-relevant details. For instance, in self-supervised contrastive learning for visual data [120], views represent various augmentations of the same image. In this scenario, the *multiview* assumption is considered mild if the downstream task remains unaffected by the augmentation [127]. Image augmentations can be perceived as altering an image's style without changing its content. Thus, Tsai et al. [128] contends that the information required for downstream tasks should be preserved in the content rather than the style. This assumption allows us to separate the information into relevant (shared information) and irrelevant (not shared) components and to compress only the unimportant details that do not contain information about downstream tasks. Based on this assumption, we aim to maximize the relevant information $I(X_2; Z_1)$ and minimize $I(X_1; Z_1 \mid X_2)$—the exclusive information that Z_1 contains about X_1, which cannot be predicted by observing X_2. This irrelevant information is unnecessary for the prediction task and can be discarded. In the extreme case, where X_1 and X_2 share only label information, this approach recovers the supervised IB method without labels. Conversely, if X_1 and X_2 are identical, this method collapses into the InfoMax principle, as no information can be accurately discarded.

Federici et al. [96] used the relaxed Lagrangian objective to obtain the minimal sufficient representation Z_1 for X_2 as follows:

$$\mathcal{L}_1 = I_{P(Z_1|X_1)}(Z_1; X_1 \mid X_2) - \beta_1 I_{P(X_2|Z_2)P(Z_2|Z_1)P(Z_1|X_1)}(X_2; Z_1), \tag{14}$$

and the symmetric loss to obtain the minimal sufficient representation Z_2 for X_1:

$$\mathcal{L}_2 = I_{P(Z_2|X_2)}(Z_2; X_2 \mid X_1) - \beta_2 I_{P(X_1|_1),Q(Z_1|Z_2),P(Z_2|X_2))} I(X_1; Z_2), \tag{15}$$

where β_1 and β_2 are the Lagrangian multipliers introduced by the constraint optimization. By defining Z_1 and Z_2 on the same domain and re-parameterizing the Lagrangian multipliers, the average of the two loss functions can be upper-bounded as follows:

$$\mathcal{L} = -I_{P(Z_1|X_1),Q(Z_2|Z_1),P(Z_2|X_2),Q(Z_1|Z_2)}(Z_1; Z_2) + \beta D_{\mathrm{SKL}}[p(z_1 \mid x_1)||P(z_2 \mid x_2)], \tag{16}$$

where D_{SKL} represents the symmetrized KL divergence obtained by averaging the expected values of $D_{\mathrm{KL}}(p(z_1 \mid x_1)||p(z_2 \mid x_2))$ and $D_{\mathrm{KL}}(p(z_2 \mid x_2)||p(z_1 \mid x_1))$. Note that when the mapping from X_1 to Z_1 is deterministic, $I(Z_1; X_1 \mid X_2)$ minimization and $H(Z_1 \mid X_2)$ minimization are interchangeable and the algorithms of Federici et al. [96] and Tsai et al. [128] minimize the same objective. Another implementation of the same idea is based on the conditional entropy bottleneck (CEB) algorithm [129] and proposed by Lee et al. [130]. This algorithm adds the residual information as a compression term to the InfoMax objective using the reverse decoders $q(z_1 \mid x_2)$ and $q(z_2 \mid x_1)$.

In conclusion, all the algorithms mentioned above are based on the multiview assumption. Utilizing this assumption, they can distinguish relevant information from irrelevant information. As a result, all these algorithms aim to maximize the information (or the predictive ability) of one representation with respect to the other view while compressing the information between each representation and its corresponding view. The key differences between these algorithms lie in the decomposition and implementation of these information terms.

Dubois et al. [131] offers another theoretical analysis of the IB for self-supervised learning. Their work addresses the question of the minimum bit rate required to store the input but still achieve a high performance on a family of downstream tasks $Y \in \mathcal{Y}$. It is a rate-distortion problem, where the goal is to find a compressed representation that will give us a good prediction for every task. We require that the distortion measure is bounded:

$$D_{\mathcal{T}}(X, Z) = \sup_{Y \in \mathcal{Y}} H(Y \mid Z_1) - H(Y \mid X_1) \leq \delta.$$

Accessing the downstream task is necessary to find the solution during the learning process. As a result, Dubois et al. [131] considered only tasks invariant to some equivalence relation, which divides the input into disjoint equivalence classes. An example would be an image with labels that remain unchanged after augmentation. This is similar to the *multiview assumption* where $\epsilon_{info} \to 0$. By applying Shannon's rate-distortion theory, they concluded that the minimum achievable bit rate is the rate-distortion function with the above invariance distortion. Thus, the optimal rate can be determined by minimizing the following Lagrangian:

$$\mathcal{L} = \min_{P(Z_1|X_1)} I_{P(Z_1|X_1)}(X_1; Z_1) + \beta H(Z_2 \mid X_1). \tag{17}$$

Using this objective, the maximization of information with labels is replaced by maximizing the prediction ability of one view from the original input, regularized by direct information from the input. Similarly to the above results, we would like to find a representation Z_1 that compresses the input X_1 so that Z_1 has the maximum information about X_2.

4.1. Implicit Compression in Self-Supervised Learning Methods

While the optimal IB representation is based on the multiview assumption, most self-supervised learning models only use the InfoMax principle and maximize the mutual information $I(Z_1; Z_2)$ without an explicit regularization term. However, recent studies have shown that contrastive learning creates compressed representations that include only relevant information [132,133]. The question is why is the learned representation compressed? The maximization of $I(Z_1; Z_2)$ could theoretically be sufficient to retain all the information from both X_1 and X_2 by making the representations invertible. In this section, we attempt to explain this phenomenon.

We begin with the InfoMax principle [17], which maximizes the mutual information between the representations of random variables Z^1 and Z^2 of the two views. We can lower-bound it using the following:

$$I(Z_1; Z_2) = H(Z) - H(Z_1 \mid Z_2) \geq H(Z_1) + \mathbb{E}[\log q(z_1 \mid z_2)]. \tag{18}$$

The bound is tight when $q(z_1|z_2) = p(z_1|z_2)$, in which case $\mathbb{E}[\log q(z_1 \mid z_2)]$, or the negative reconstruction error, equals the conditional entropy $H(Z_1|Z_2)$.

In the supervised case, where Z is a learned stochastic representation of the input and Y is the label, we aim to optimize

$$I(Y; Z) \geq H(Y) + \mathbb{E}[\log q(Y \mid Z)]. \tag{19}$$

Since Y is constant, optimizing the information $I(Z; Y)$ requires only minimizing the prediction term $\mathbb{E}[\log q(Y|Z)]$ by making Z more informative about Y. This term is the cross-entropy loss for classification or the square loss for regressions. Thus, we can minimize the log loss without any other regularization on the representation.

In contrast, for the self-supervised case, we have a more straightforward option to minimize $H(Z_1|Z_2)$: making Z_1 easier to predict by Z_2, which can be achieved by reducing its variance along specific dimensions. If we do not regularize $H(Z_1)$, it will decrease to zero, and we will observe a collapse. This is why, in contrastive methods, the variance in the representation (large entropy) is significant only in the directions with a high variance in the data, which is enforced by data augmentation [61]. According to this analysis, the network benefits from making the representations "simple" (easier to predict). Hence, even though our representation does not have explicit information-theoretical constraints, the learning process will compress the representation.

4.2. Beyond the Multiview Assumption

According to the multiview IB analysis presented in Section 4, the optimal way to create a useful representation is to maximize the mutual information between the representations of different views while compressing irrelevant information in each representation. In fact, as discussed in Section 4.1, we can achieve this optimal compressed representation even without explicit regularization. However, this optimality is based on the *multiview assumption*, which states that the relevant information for downstream tasks comes from the information shared between views. Therefore, Tian et al. [133] concluded that when a minimal sufficient representation has been obtained, the optimal views for self-supervised learning are determined by downstream tasks.

However, the *multiview assumption* is highly constrained, as all relevant information must be shared between all views. In cases where this assumption is incorrect, such as with aggressive data augmentation or multiple downstream tasks or modalities, sharing all the necessary information can be challenging. For example, if one view is a video stream while the other is an audio stream, the shared information may be sufficient for object recognition but not for tracking. Furthermore, relevant information for downstream tasks may not be contained within the shared information between views, meaning that removing non-shared information can negatively impact the performance.

Kahana and Hoshen [134] identified a series of tasks that violate the *multiview assumption*. To accomplish these tasks, the learned representation must also be invariant to unwanted attributes, such as bias removal and cross-domain retrieval. In such cases, only some attributes have labels, and the objective is to learn an invariant representation for the domain for which labels are provided while also being informative for all other attributes without labels. For example, for face images, only the identity labels may be provided, and the goal is to learn a representation that captures the unlabeled pose attribute but contains no information about the identity attribute. The task can also be applied to fair decisions, cross-domain matching, model anonymization, and image translation.

Wang et al. [132] formalized another case where the *multiview assumption* does not hold when non-shared task-relevant information cannot be ignored. In such cases, the minimal sufficient representation contains less task-relevant information than other sufficient representations, resulting in an inferior performance. Furthermore, their analysis shows that in such cases, the learned representation in contrastive learning is insufficient for downstream tasks, which may overfit the shared information.

As a result of their analysis, Wang et al. [132] and Kahana and Hoshen [134] proposed explicitly increasing mutual information between the representation and input to preserve task-relevant information and prevent the compression of unshared information between views. In this case, the two regularization terms of the two views are incorporated into the original InfoMax objective, and the following objective is optimized:

$$\mathcal{L} = \min_{P(Z_1|X_1),p(Z_2|X_2)} -I_{P(Z_1|X_1)}(X_1;Z_1) - I_{P(Z_2|X_2)}(X_2;Z_2) - \beta I_{P(Z_1|X_1),P(Z_2|Z_1)}(Z_1;Z_2). \tag{20}$$

Wang et al. [132] demonstrated the effectiveness of their method for SimCLR [7], BYOL [55], and Barlow Twins [135] across classification, detection, and segmentation tasks.

4.3. To Compress or Not to Compress?

As seen in Equation (20), when the *multiview assumption* is violated, the objective for obtaining an optimal representation is to **maximize** the mutual information between each input and its representation. This contrasts with the situation in which the *multiview assumption* holds, or the supervised case, where the objective is to **minimize** the mutual information between the representation and the input. In both supervised and unsupervised cases, we have direct access to the relevant information, which we can use to separate and compress irrelevant information. However, in the self-supervised case, we depend heavily on the *multiview assumption*. If this assumption is violated due to unshared information between views that is relevant for the downstream task, we cannot separate relevant and irrelevant information. Furthermore, the learning algorithm's nature requires that this information be protected by explicitly maximizing it.

As datasets continue to expand in size and models are anticipated to serve as base models for various downstream tasks, the *multiview assumption* becomes less pertinent. Consequently, compressing irrelevant information when the *multiview assumption* does not hold presents one of the most significant challenges in self-supervised learning. Identifying new methods to separate relevant from irrelevant information based on alternative assumptions is a promising avenue for research. It is also essential to recognize that empirical measurement of information-theoretic quantities and their estimators plays a crucial role in developing and evaluating such methods.

5. Optimizing Information in Deep Neural Networks: Challenges and Approaches

Recent years have seen information-theoretic analyses employed to explain and optimize deep learning techniques [74]. Despite their elegance and plausibility, empirically measuring and analyzing information in deep networks presents challenges. Two critical problems are (1) information in deterministic networks and (2) estimating information in high-dimensional spaces.

5.1. Information in Deterministic Networks

Information-theoretic methods have significantly impacted deep learning [13,15,74]. However, a key challenge is addressing the source of randomness in deterministic DNNs.

The mutual information between the input and representation is infinite, leading to ill-posed optimization problems or piecewise constant outcomes [136,137]. To tackle this issue, researchers have proposed various solutions. One common approach is to discretize the input distribution and real-valued hidden representations by binning, which facilitates non-trivial measurements and prevents the mutual information from always taking the maximum value of the log of the dataset size, thus avoiding ill-posed optimization problems [74].

However, binning and discretization are essentially equivalent to geometrical compression and serve as clustering measures [137]. Moreover, this discretization depends on the chosen bin size and does not track the mutual information across varying bin sizes [137,138]. To address these limitations, researchers have proposed alternative approaches, such as interpreting binned information as a weight decay penalty [139], estimating mutual information based on lower bounds assuming a continuous input distribution without making assumptions about the network's output distribution properties [140–142], injecting additive noise, and considering data augmentation as the source of noise [74,130,131,137].

5.2. Measuring Information in High-Dimensional Spaces

Estimating mutual information in high-dimensional spaces presents a significant challenge when applying information-theoretic measures to real-world data. This problem has been extensively studied [143,144], revealing the inefficiency of solutions for large dimensions and the limited scalability of known approximations with respect to the sample size and dimension. Despite these difficulties, various entropy and mutual information estimation approaches have been developed, including classic methods like k-nearest neighbors (KNNs) [145] and kernel density estimation techniques [146], as well as more recent efficient methods.

Chelombiev et al. [92] developed adaptive mutual information estimators based on entropy equal bins and the scaled noise kernel density estimator. Generative decoder networks, such as PixelCNN++ [147], have been employed to estimate a lower bound on mutual information [148–150]. Another strategy includes ensemble dependency graph estimators (EDGEs), adaptive mutual information estimation methods by merging randomized locality-sensitive hashing (LSH), dependency graphs, and ensemble bias reduction techniques [151]. The Mutual Information Neural Estimator (MINE) [152] maximizes KL divergence using the dual representation of Donsker and Varadhan [153] and has been employed for direct mutual information estimation [154]. Shwartz-Ziv and Alemi [155] developed a controlled framework that utilized the neural tangent kernels [156], in order to obtain tractable information measures.

Recent work by Poole et al. [157] introduced a framework for variational bounds of mutual information (MI), addressing bias and variance in existing estimators. This approach unifies recent developments and proposes a continuum of lower bounds that flexibly trades off bias and variance. In contrast, McAllester and Stratos [158] highlighted the statistical limitations inherent in all MI measuring methods. They suggest a difference-of-entropies estimator as a feasible alternative for estimating large MI.

Improving mutual information estimation can be achieved using larger batch sizes, although this may negatively impact the generalization performance and memory requirements. Alternatively, researchers have suggested employing surrogate measures for mutual information, such as log-determinant mutual information (LDMI), based on second-order statistics [159,160], which reflects linear dependence. Goldfeld and Greenewald [161] proposed the Sliced Mutual Information (SMI), defined as an average of MI terms between one-dimensional projections of high-dimensional variables. SMI inherits many properties of its classic counterpart. It can be estimated with optimal parametric error rates in all dimensions by combining an MI estimator between scalar variables with an MC integrator [161]. The k-SMI, introduced by Goldfeld et al. [162], extends the SMI by projecting to a k-dimensional subspace, which relaxes the smoothness assumptions, improves the scalability, and enhances the performance.

In conclusion, estimating and optimizing information in deep neural networks present significant challenges, particularly in deterministic networks and high-dimensional spaces. Researchers have proposed various approaches to address these issues, including discretization, alternative estimators, and surrogate measures. As the field continues to evolve, it is expected that more advanced techniques will emerge to overcome these challenges and facilitate the understanding and optimization of deep learning models.

6. Related Work

This work lies at the intersection of information theory and SSL, aiming to enhance machine learning models through the principles of encoding, compression, and generalization.

6.1. Information Theory Reviews

Information theory has been crucial in machine learning's evolution, starting with Shannon [163], who introduced key concepts like entropy and mutual information. Further reviews by Cover and Thomas [164] and Yeung [165] extended these ideas, incorporating computational advances to address data transmission and decoding challenges. Recent studies, such as those by Wilde [166] and Dimitrov et al. [167], have explored information theory's application in quantum computing and neuroscience.

Significant review works on the IB principle include Slonim [168], which thoroughly reviewed the IB method and its extensions, including the multivariate IB. Recent research by Goldfeld and Polyanskiy [169] and Shwartz-Ziv and Tishby [74] has applied IB theory to deep learning, optimizing feature representations to balance informativeness and compression. This research underscores the theory's importance in advancing machine learning algorithms and deep learning, seeking to bridge theory and practice.

6.2. Self-Supervised Learning Reviews

SSL represents a significant shift, allowing for the use of unlabeled data to learn valuable representations. Jaiswal et al. [170] covers SSL's progress, especially in contrastive learning's application to computer vision, NLP, and beyond. It provides an overview of various methods, showcasing how SSL improves learning representations for diverse tasks, and evaluates the potential and limitations of current methods.

Liu et al. [48] and Gui et al. [171] give detailed analyses of SSL techniques across several domains, including computer vision, NLP, and graph learning. They details how SSL, using input data for supervision, overcomes supervised learning's limitations, enhancing representation learning without manual labeling. This survey classifies methods into generative, contrastive, and generative–contrastive (adversarial) categories, providing theoretical insights.

Patil and Gudivada [172] and Wang et al. [173] delve into SSL-enhanced language models and their application to non-sequential tabular data, respectively. Meanwhile, Xie et al. [174] highlights the parallels between graph neural networks and SSL algorithms, and Hojjati et al. [175] discusses SSL's impact on anomaly detection in fields such as cybersecurity, finance, and healthcare. Moreover, Schiappa et al. [176] and Yu et al. [177] reviewed SSL in videos and recommendation systems.

Our work compiles these insights, offering a comprehensive review that combines the theoretical rigor of information theory with the practical advancements of SSL. Our goal is to pave the way for future research that leverages this interdisciplinary approach to uncover new efficiencies and applications in machine learning.

7. Future Research Directions

Despite the solid foundation established by existing self-supervised learning methods from an information theory perspective, several potential research directions warrant exploration:

- **Self-supervised learning with non-shared information.** As discussed in Section 4, the separation of relevant (preserved) and irrelevant (compressed) information relies on the *multiview assumption*. This assumption, which states that only shared information is essential for downstream tasks, is rather restrictive. For example, situations may arise where each view contains distinct information relevant to a downstream task or multiple tasks necessitate different features. Some methods have been proposed to tackle this problem, but they mainly focus on maximizing the network's information without explicit constraints. Formalizing this scenario and exploring differentiating between relevant and irrelevant data based on non-shared information represents an intriguing research direction.

- **Self-supervised learning for tabular data.** At present, the internal compression of self-supervised learning methods may compress relevant information due to improper augmentation (Section 4.1). Consequently, we must heavily rely on generating the two views, which must accurately represent information related to the downstream process. Custom augmentation must be developed for each domain, taking into account extensive prior knowledge on data augmentation. While some papers have attempted to extend self-supervised learning to tabular data [178,179], further work is necessary from both theoretical and practical standpoints to achieve a high performance with self-supervised learning for tabular data [180]. The augmentation process is crucial for the performance of current vision and text models. In the case of tabular data, employing information-theoretic loss functions that do not require information compression may help harness the benefits of self-supervised learning.
- **Integrating other learning methods into the information-theoretic framework.** Prior works have investigated various supervised, unsupervised, semi-supervised, and self-supervised learning methods, demonstrating that they optimize information-theoretic quantities. However, state-of-the-art methods employ additional changes and engineering practices that may be related to information theory, such as the stop gradient operation utilized by many self-supervised learning methods today [53,55]. The Expectation–Maximization (EM) algorithm [181] can be employed to explain this operation when one path is the E-step and the other is the M-step. Additionally, Elidan and Friedman [182] proposed an IB-inspired version of the EM algorithm, which could help develop information-theoretic-based objectives using the stop gradient operation.
- **Expanding the analysis to usable information.** While information theory offers a rigorous conceptual framework for describing information, it neglects essential aspects of computation. Conditional entropy, for example, is directly related to the predictability of a random variable in a betting game where agents are rewarded for accurate guesses. However, the standard definition assumes that agents have no computational bounds and can employ arbitrarily complex prediction schemes [76]. In the context of deep learning, predictive information $H(Y|Z)$ measures the amount of information that can be extracted from Z about Y given access to all decoders $p(y|z)$ in the world. Recently, Xu et al. [183] introduced *predictive V-information* as an alternative formulation based on realistic computational constraints.
- **Extending self-supervised learning's information-based perspective to energy-based model optimization.** Until now, research combining self-supervised learning with information theory has focused on probabilistic models with tractable likelihoods. These models enable the specific optimization of model parameters concerning the tractable log-likelihood [184–187] or a tractable lower bound of the likelihood [13,111]. Although models with tractable likelihoods offer certain benefits, their scope is limited and necessitates a particular format. Energy-based models (EBMs) present a more flexible, unified framework. Rather than specifying a normalized probability, EBMs define inference as minimizing an unnormalized energy function and learning as minimizing a loss function. The energy function does not require integration and can be parameterized with any nonlinear regression function. Inference typically involves finding a low-energy configuration or sampling from all possible configurations such that the probability of selecting a specific configuration follows a Gibbs distribution [188,189].

  Investigating energy-based models for self-supervised learning from both theoretical and practical perspectives can open up numerous promising research directions. For instance, we could directly apply tools developed for energy-based models and statistical machines to optimize the model, such as Maximum Likelihood Training with

MCMC [190], score matching [191], denoising score matching [192,193], and score-based generation models [194].

- **Expanding the multiview framework to accommodate more views and tasks.** The multiview self-supervised IB framework can be extended to cases involving more than two views $(X_1, \cdots, X_n)$ and multiple downstream tasks $(Y_1, \cdots, Y_K)$. A simple extension of the multiview IB framework can be achieved by setting the objective function to maximize the joint mutual information of all views' representations $I(Z_1; \cdots Z_n)$ and compressing the individual information for each view $I(X_i; Z_i), \quad 1 \leq i \leq N$. However, to ensure the optimality of this objective, we must expand the *multiview assumption* to include more than two views. In this scenario, we need to assume that relevant information is shared among all different views and tasks, which might be overly restrictive. As a result, defining and analyzing a more refined version of this naive solution is essential. One potential approach involves utilizing the multi-feature information bottleneck (MfIB) [195], which extends the original IB. The MfIB processes multiple feature types simultaneously and analyzes data from various sources. This framework establishes a joint distribution between the multivariate data and the model. Rather than solely preserving the information of one feature variable maximally, the MfIB concurrently maintains multiple feature variables' information while compressing them. The MfIB characterizes the relationships between different sources and outputs by employing the multivariate information bottleneck [196] and setting Bayesian networks.

8. Conclusions

In this study, we delved deeply into the concept of optimal representation in self-supervised learning through the lens of information theory. We synthesized various approaches, highlighting their foundational assumptions and constraints, and integrated them into a unified framework. Additionally, we explored the key information-theoretic terms that influence these optimal representations and the methods for estimating them.

While supervised and unsupervised learning offer more direct access to relevant information, self-supervised learning depends heavily on assumptions about the relationship between data and downstream tasks. This reliance makes distinguishing between relevant and irrelevant information considerably more challenging, necessitating further assumptions.

Despite these challenges, information theory stands out as a robust and versatile framework for analysis and algorithmic development. This adaptable framework caters to a range of learning paradigms and elucidates the inherent assumptions underpinning data and model optimization.

With the rapid growth of datasets and the increasing expectations placed on models to handle multiple downstream tasks, the traditional multiview assumption might become less reliable. One significant challenge in self-supervised learning is the precise compression of irrelevant information, especially when these assumptions are compromised.

Future research avenues might involve expanding the multiview framework to include more views and tasks and deepening our understanding of information theory's impact on facets of deep learning, such as reinforcement learning and generative models.

In summary, information theory is a crucial tool in our quest to better understand and optimize self-supervised learning models. By harnessing its principles, we can more adeptly navigate the intricacies of deep neural network development, paving the way for creating more effective models.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: Author Yann LeCun was employed by the company Meta AI-FAIR. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Alam, M.; Samad, M.D.; Vidyaratne, L.; Glandon, A.; Iftekharuddin, K.M. Survey on deep neural networks in speech and vision systems. *Neurocomputing* **2020**, *417*, 302–321. [CrossRef]
2. LeCun, Y.; Bengio, Y.; Hinton, G.E. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
4. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 1993; Volume 6.
5. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 539–546.
6. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
7. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
8. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
9. Misra, I.; van der Maaten, L. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6707–6717.
10. Shwartz-Ziv, R.; Goldblum, M.; Souri, H.; Kapoor, S.; Zhu, C.; LeCun, Y.; Wilson, A.G. Pre-train your loss: Easy bayesian transfer learning with informative priors. In *Advances in Neural Information Processing Systems;* Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 27706–27715.
11. Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv* **2019**, arXiv:1902.09229.
12. Lee, J.D.; Lei, Q.; Saunshi, N.; Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems;* Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 3.
13. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
14. Xu, A.; Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems;* Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
15. Steinke, T.; Zakynthinou, L. Reasoning about generalization via conditional mutual information. In Proceedings of the Conference on Learning Theory, PMLR, Graz, Austria, 9–12 July 2020; pp. 3437–3452.
16. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; Volume 80, pp. 531–540.
17. Linsker, R. Self-organization in a perceptual network. *Computer* **1988**, *21*, 105–117. [CrossRef]
18. Tishby, N.; Pereira, F.; Biale, W. The Information Bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
19. Zhao, J.; Xie, X.; Xu, X.; Sun, S. Multi-view learning overview: Recent progress and new challenges. *Inf. Fusion* **2017**, *38*, 43–54. [CrossRef]
20. Yan, X.; Hu, S.; Mao, Y.; Ye, Y.; Yu, H. Deep multi-view learning methods: A review. *Neurocomputing* **2021**, *448*, 106–129. [CrossRef]
21. Kumar, A.; Daumé, H. A co-training approach for multi-view spectral clustering. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Citeseer, Bellevue, DC, USA, 28 June–2 July 2011; pp. 393–400.
22. Xue, Z.; Du, J.; Du, D.; Lyu, S. Deep low-rank subspace ensemble for multi-view clustering. *Inf. Sci.* **2019**, *482*, 210–227. [CrossRef]
23. Bach, F.R.; Jordan, M.I. Kernel independent component analysis. *J. Mach. Learn. Res.* **2002**, *3*, 1–48.
24. Li, Y.; Yang, M.; Zhang, Z. A survey of multi-view representation learning. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1863–1883. [CrossRef]
25. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **1936**, *28*, 321–377. [CrossRef]
26. Hardoon, D.R.; Szedmak, S.; Shawe-Taylor, J. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Comput.* **2004**, *16*, 2639–2664. [CrossRef]
27. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **2013**, *23*, 2031–2038. [CrossRef]
28. Sun, L.; Ceran, B.; Ye, J. A scalable two-stage approach for a class of dimensionality reduction techniques. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 313–322.
29. Yan, X.; Ye, Y.; Lou, Z. Unsupervised video categorization based on multivariate information bottleneck method. *Knowl.-Based Syst.* **2015**, *84*, 34–45. [CrossRef]

30. Jia, Y.; Salzmann, M.; Darrell, T. Factorized Latent Spaces with Structured Sparsity. In *Advances in Neural Information Processing Systems*; Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2010; Volume 23.
31. Cao, T.; Jojic, V.; Modla, S.; Powell, D.; Czymmek, K.; Niethammer, M. Robust Multimodal Dictionary Learning. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013: 16th International Conference, Nagoya, Japan, 22–26 September 2013*; Proceedings, Part III; Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 259–266.
32. Liu, W.; Tao, D.; Cheng, J.; Tang, Y. Multiview Hessian discriminative sparse coding for image annotation. *Comput. Vis. Image Underst.* **2014**, *118*, 50–60. [CrossRef]
33. Pu, S.; He, Y.; Li, Z.; Zheng, M. Multimodal Topic Learning for Video Recommendation. *arXiv* **2020**, arXiv:2010.13373.
34. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, PMLR, Atlanta, GA, USA, 17–19 June 2013; pp. 1247–1255.
35. Zhao, H.; Ding, Z.; Fu, Y. Multi-view clustering via deep matrix factorization. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
36. Huang, Z.; Zhou, J.T.; Peng, X.; Zhang, C.; Zhu, H.; Lv, J. Multi-view Spectral Clustering Network. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 2563–2569.
37. Liu, S.; Xia, Y.; Shi, Z.; Yu, H.; Li, Z.; Lin, J. Deep learning in sheet metal bending with a novel theory-guided deep neural network. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 565–581. [CrossRef]
38. Srivastava, N.; Salakhutdinov, R. Multimodal Learning with Deep Boltzmann Machines. *J. Mach. Learn. Res.* **2014**, *15*, 2949–2980.
39. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal Deep Learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Madison, WI, USA, 28 June–2 July 2011; pp. 689–696.
40. Wang, W.; Arora, R.; Livescu, K.; Bilmes, J. On Deep Multi-View Representation Learning. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, ICML'15, Lille, France, 7–9 July 2015; Volume 37, pp. 1083–1092.
41. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
42. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv* **2014**, arXiv:1412.6632.
43. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
44. Zhu, J.; Shwartz-Ziv, R.; Chen, Y.; LeCun, Y. Variance-Covariance Regularization Improves Representation Learning. *arXiv* **2023**, arXiv:2306.13292.
45. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
46. Bachman, P.; Hjelm, R.D.; Buchwalter, W. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
47. Bar, A.; Wang, X.; Kantorov, V.; Reed, C.J.; Herzig, R.; Chechik, G.; Rohrbach, A.; Darrell, T.; Globerson, A. Detreg: Unsupervised pretraining with region priors for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14605–14615.
48. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, 35, 857–876. [CrossRef]
49. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2014**, arXiv:1312.6114.
50. Lee, H.; Battle, A.; Raina, R.; Ng, A. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*; Schölkopf, B., Platt, J., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA, 2006; Volume 19.
51. Ng, A. Sparse autoencoder. *CS294A Lect. Notes* **2011**, *72*, 1–19.
52. Van Den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
53. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15750–15758.
54. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
55. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 21271–21284.
56. Bardes, A.; Ponce, J.; LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv* **2021**, arXiv:2105.04906.
57. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.

58. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019.
59. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
60. Gutmann, M.; Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; Teh, Y.W., Titterington, M., Eds.; Volume 9, pp. 297–304.
61. Jing, L.; Vincent, P.; LeCun, Y.; Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv* **2021**, arXiv:2110.09348.
62. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 9912–9924.
63. Chapelle, O.; Scholkopf, B.; Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Trans. Neural Netw.* **2009**, *20*, 542. [CrossRef]
64. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
65. Miyato, T.; Maeda, S.i.; Koyama, M.; Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993. [CrossRef]
66. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 596–608.
67. Grandvalet, Y.; Bengio, Y. Entropy Regularization. 2006. Available online: https://www.researchgate.net/profile/Y-Bengio/publication/237619703_9_Entropy_Regularization/links/0f3175320aaecbde17000000/9-Entropy-Regularization.pdf (accessed on 8 May 2023).
68. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6256–6268.
69. Zhai, X.; Oliver, A.; Kolesnikov, A.; Beyer, L. S4l: Self-supervised semi-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1476–1485.
70. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the Workshop on Challenges in Representation Learning, ICML, Daegu, Republic of Korea, 3–7 November 2013; Volume 3, p. 896.
71. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
72. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning;* MIT Press: Cambridge, MA, USA, 2016.
73. Bengio, Y.; LeCun, Y. Scaling Learning Algorithms towards AI. In *Large Scale Kernel Machines*; Bottou, L., Chapelle, O., DeCoste, D., Weston, J., Eds.; MIT Press: Cambridge, MA, USA, 2007.
74. Shwartz-Ziv, R.; Tishby, N. Opening the black box of deep neural networks via information. *arXiv* **2017**, arXiv:1703.00810.
75. Ben-Shaul, I.; Shwartz-Ziv, R.; Galanti, T.; Dekel, S.; LeCun, Y. Reverse Engineering Self-Supervised Learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
76. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
77. Koopman, B.O. On distributions admitting a sufficient statistic. *Trans. Am. Math. Soc.* **1936**, *39*, 399–409. [CrossRef]
78. Buesing, L.; Maass, W. A spiking neuron as information bottleneck. *Neural Comput.* **2010**, *22*, 1961–1992. [CrossRef]
79. Palmer, S.E.; Marre, O.; Berry, M.J.; Bialek, W. Predictive information in a sensory population. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6908–6913. [CrossRef] [PubMed]
80. Turner, R.; Sahani, M. A maximum-likelihood interpretation for slow feature analysis. *Neural Comput.* **2007**, *19*, 1022–1038. [CrossRef]
81. Hecht, R.M.; Noor, E.; Tishby, N. Speaker recognition by Gaussian information bottleneck. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
82. Lee, N.; Hyun, D.; Na, G.S.; Kim, S.; Lee, J.; Park, C. Conditional Graph Information Bottleneck for Molecular Relational Learning. *arXiv* **2023**, arXiv:2305.01520.
83. Erdogmus, D. *Information Theoretic Learning: Renyi's Entropy and Its Applications to Adaptive System Training*; University of Florida: Gainesville, FL, USA, 2002.
84. Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
85. Zhang, Z.; Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
86. Wenzel, F.; Roth, K.; Veeling, B.S.; Świkatkowski, J.; Tran, L.; Mandt, S.; Snoek, J.; Salimans, T.; Jenatton, R.; Nowozin, S. How good is the bayes posterior in deep neural networks really? *arXiv* **2020**, arXiv:2002.02405.
87. Painsky, A.; Wornell, G.W. On the Universality of the Logistic Loss Function. *arXiv* **2018**, arXiv:1805.03804.
88. Shamir, O.; Sabato, S.; Tishby, N. Learning and generalization with the information bottleneck. *Theor. Comput. Sci.* **2010**, *411*, 2696–2711. [CrossRef]

89. Vera, M.; Piantanida, P.; Vega, L.R. The role of information complexity and randomization in representation learning. *arXiv* **2018**, arXiv:1802.05355.
90. Russo, D.; Zou, J. How much does your data exploration overfit? controlling bias via information usage. *IEEE Trans. Inf. Theory* **2019**, *66*, 302–323. [CrossRef]
91. Achille, A.; Soatto, S. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.* **2018**, *19*, 1947–1980.
92. Chelombiev, I.; Houghton, C.; O'Donnell, C. Adaptive estimators show information compression in deep neural networks. *arXiv* **2019**, arXiv:1902.09037.
93. Shwartz-Ziv, R.; Painsky, A.; Tishby, N. Representation Compression and Generalization in Deep Neural Networks. 2018. Available online: https://arxiv.org/pdf/2202.06749.pdf#page=56 (accessed on 12 December 2023).
94. Piran, Z.; Shwartz-Ziv, R.; Tishby, N. The dual information bottleneck. *arXiv* **2020**, arXiv:2006.04641.
95. Shwartz-Ziv, R. Information flow in deep neural networks. *arXiv* **2022**, arXiv:2202.06749.
96. Federici, M.; Dutta, A.; Forré, P.; Kushman, N.; Akata, Z. Learning robust representations via multi-view information bottleneck. *arXiv* **2020**, arXiv:2002.07017.
97. Geiger, B.C. On Information Plane Analyses of Neural Network Classifiers—A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 7039–7051. [CrossRef]
98. Amjad, R.A.; Geiger, B.C. How (Not) To Train Your Neural Network Using the Information Bottleneck Principle. *arXiv* **2018**, arXiv:1802.09766.
99. Achille, A.; Rovere, M.; Soatto, S. Critical learning periods in deep neural networks. *arXiv* **2019**, arXiv:1711.08856.
100. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech. Theory Exp.* **2019**, *2019*, 124020. [CrossRef]
101. Basirat, M.; Geiger, B.C.; Roth, P.M. A Geometric Perspective on Information Plane Analysis. *Entropy* **2021**, *23*, 711. [CrossRef] [PubMed]
102. Pensia, A.; Jog, V.; Loh, P.L. Generalization error bounds for noisy, iterative algorithms. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 546–550.
103. Xu, C.; Tao, D.; Xu, C. Large-Margin Multi-ViewInformation Bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1559–1572. [CrossRef] [PubMed]
104. Wang, Q.; Boudreau, C.; Luo, Q.; Tan, P.N.; Zhou, J. Deep Multi-view Information Bottleneck. In Proceedings of the 2019 SIAM International Conference on Data Mining (SDM), Calgary, AB, Canada, 2–4 May 2019; pp. 37–45. [CrossRef]
105. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
106. Voloshynovskiy, S.; Taran, O.; Kondah, M.; Holotyak, T.; Rezende, D. Variational Information Bottleneck for Semi-Supervised Classification. *Entropy* **2020**, *22*, 943. [CrossRef] [PubMed]
107. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. *arXiv* **2015**, arXiv:1511.05644.
108. Springenberg, J.T. Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06390.
109. Śmieja, M.; Wołczyk, M.; Tabor, J.; Geiger, B.C. SeGMA: Semi-Supervised Gaussian Mixture Autoencoder. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *32*, 3930–3941. [CrossRef]
110. Kingma, D.P.; Mohamed, S.; Jimenez Rezende, D.; Welling, M. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
111. Kingma, D.P.; Welling, M. An introduction to variational autoencoders. *Found. Trends® Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
112. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
113. Zhao, S.; Song, J.; Ermon, S. Infovae: Information maximizing variational autoencoders. *arXiv* **2019**, arXiv:1706.02262.
114. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 20–22 June 2016; pp. 1558–1566.
115. Uğur, Y.; Arvanitakis, G.; Zaidi, A. Variational information bottleneck for unsupervised clustering: Deep gaussian mixture embedding. *Entropy* **2020**, *22*, 213. [CrossRef]
116. Roy, A.; Vaswani, A.; Neelakantan, A.; Parmar, N. Theory and experiments on vector quantized autoencoders. *arXiv* **2018**, arXiv:1805.11063.
117. Wiskott, L.; Sejnowski, T.J. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Comput.* **2002**, *14*, 715–770. [CrossRef] [PubMed]
118. Becker, S.; Hinton, G.E. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* **1992**, *355*, 161–163. [CrossRef] [PubMed]
119. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159. [CrossRef] [PubMed]
120. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2019**, arXiv:1808.06670.

121. Henaff, O. Data-efficient image recognition with contrastive predictive coding. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 4182–4192.
122. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 776–794.
123. Tschannen, M.; Djolonga, J.; Rubenstein, P.K.; Gelly, S.; Lucic, M. On mutual information maximization for representation learning. *arXiv* **2020**, arXiv:1907.13625.
124. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
125. Gutknecht, A.J.; Wibral, M.; Makkeh, A. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proc. R. Soc. A* **2021**, *477*, 20210110. [CrossRef]
126. Sridharan, K.; Kakade, S. An Information Theoretic Framework for Multi-View Learning. In Proceedings of the 21st Annual Conference on Learning Theory—COLT 2008, Helsinki, Finland, 9–12 July 2008.
127. Geiping, J.; Goldblum, M.; Somepalli, G.; Shwartz-Ziv, R.; Goldstein, T.; Wilson, A.G. How Much Data Are Augmentations Worth? An Investigation into Scaling Laws, Invariance, and Implicit Regularization. *arXiv* **2023**, arXiv:2210.06441.
128. Tsai, Y.H.H.; Wu, Y.; Salakhutdinov, R.; Morency, L.P. Self-supervised learning from a multi-view perspective. *ICLR 2021* **2020**.
129. Fischer, I. The conditional entropy bottleneck. *Entropy* **2020**, *22*, 999. [CrossRef]
130. Lee, K.H.; Arnab, A.; Guadarrama, S.; Canny, J.; Fischer, I. Compressive visual representations. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34.
131. Dubois, Y.; Bloem-Reddy, B.; Ullrich, K.; Maddison, C.J. Lossy compression for lossless prediction. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34.
132. Wang, H.; Guo, X.; Deng, Z.H.; Lu, Y. Rethinking Minimal Sufficient Representation in Contrastive Learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
133. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6827–6839.
134. Kahana, J.; Hoshen, Y. A Contrastive Objective for Learning Disentangled Representations. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022.
135. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12310–12320.
136. Amjad, R.A.; Geiger, B.C. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2225–2239. [CrossRef] [PubMed]
137. Goldfeld, Z.; van den Berg, E.; Greenewald, K.; Melnyk, I.; Nguyen, N.; Kingsbury, B.; Polyanskiy, Y. Estimating Information Flow in Neural Networks. *arXiv* **2018**, arXiv:1810.05728.
138. Ross, B.C. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef]
139. Elad, A.; Haviv, D.; Blau, Y.; Michaeli, T. The Effectiveness of Layer-by-Layer Training Using the Information Bottleneck Principle. 2019. Available online: https://openreview.net/forum?id=r1Nb5i05tX (accessed on 12 February 2024).
140. Wang, T.; Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 9929–9939.
141. Zimmermann, R.S.; Sharma, Y.; Schneider, S.; Bethge, M.; Brendel, W. Contrastive learning inverts the data generating process. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12979–12990.
142. Shwartz-Ziv, R.; Balestriero, R.; LeCun, Y. What Do We Maximize in Self-Supervised Learning? *arXiv* **2022**, arXiv:2207.10081.
143. Paninski, L. Estimation of Entropy and Mutual Information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
144. Gao, S.; Ver Steeg, G.; Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 277–286.
145. Kozachenko, L.F.; Leonenko, N.N. Sample estimate of the entropy of a random vector. *Probl. Peredachi Informatsii* **1987**, *23*, 9–16.
146. Hang, H.; Steinwart, I.; Feng, Y.; Suykens, J.A. Kernel density estimation for dynamical systems. *J. Mach. Learn. Res.* **2018**, *19*, 1260–1308.
147. Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
148. Darlow, L.N.; Storkey, A. What Information Does a ResNet Compress? *arXiv* **2020**, arXiv:2003.06254.
149. Nash, C.; Kushman, N.; Williams, C.K.I. Inverting Supervised Representations with Autoregressive Neural Density Models. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Playa Blanca, Lanzarote, 9–11 April 2018.
150. Shwartz-Ziv, R.; Balestriero, R.; Kawaguchi, K.; Rudner, T.G.; LeCun, Y. An Information-Theoretic Perspective on Variance-Invariance-Covariance Regularization. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
151. Noshad, M.; Zeng, Y.; Hero, A.O. Scalable Mutual Information Estimation Using Dependence Graphs. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2962–2966. [CrossRef]
152. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Hjelm, R.D.; Courville, A.C. Mutual Information Neural Estimation. In Proceedings of the ICML, Stockholm, Sweden, 10–15 July 2018.

153. Donsker, M.D.; Varadhan, S.S. Asymptotic evaluation of certain Markov process expectations for large time, I. *Commun. Pure Appl. Math.* **1975**, *28*, 1–47. [CrossRef]
154. Elad, A.; Haviv, D.; Blau, Y.; Michaeli, T. Direct validation of the information bottleneck principle for deep nets. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
155. Shwartz-Ziv, R.; Alemi, A.A. Information in infinite ensembles of infinitely-wide neural networks. In Proceedings of the Symposium on Advances in Approximate Bayesian Inference, PMLR. 2020; pp. 1–17. Available online: http://proceedings.mlr.press/v118/shwartz-ziv20a.html (accessed on 8 May 2023).
156. Jacot, A.; Gabriel, F.; Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
157. Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; Tucker, G. On variational bounds of mutual information. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 5171–5180.
158. McAllester, D.; Stratos, K. Formal limitations on the measurement of mutual information. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Online, 26–28 August 2020, pp. 875–884.
159. Ozsoy, S.; Hamdan, S.; Arik, S.; Yuret, D.; Erdogan, A. Self-supervised learning with an information maximization criterion. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 35240–35253.
160. Erdogan, A.T. An information maximization based blind source separation approach for dependent and independent sources. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 23–27 May 2022; pp. 4378–4382.
161. Goldfeld, Z.; Greenewald, K. Sliced mutual information: A scalable measure of statistical dependence. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 17567–17578.
162. Goldfeld, Z.; Greenewald, K.; Nuradha, T.; Reeves, G. k-sliced mutual information: A quantitative study of scalability with dimension. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022.
163. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
164. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
165. Yeung, R.W. Information Theory and Network Coding (Yeung, R.W.; 2008) [Book review]. *IEEE Trans. Inf. Theory* **2009**, *55*, 3409. [CrossRef]
166. Wilde, M.M. *Quantum Information Theory*; Cambridge University Press: Cambridge, UK, 2013.
167. Dimitrov, A.G.; Lazar, A.A.; Victor, J.D. Information theory in neuroscience. *J. Comput. Neurosci.* **2011**, *30*, 1–5. [CrossRef]
168. Slonim, N. The Information Bottleneck: Theory and Applications. Ph.D. Thesis, Hebrew University of Jerusalem, Jerusalem, Israel, 2002.
169. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38. [CrossRef]
170. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **2020**, *9*, 2. [CrossRef]
171. Gui, J.; Chen, T.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. *arXiv* **2023**, arXiv:2301.05712.
172. Patil, R.; Gudivada, A. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appli. Sci.* **2024**, *14*, 20. [CrossRef]
173. Wang, W.Y.; Du, W.W.; Xu, D.; Wang, W.; Peng, W.C. A Survey on Self-Supervised Learning for Non-Sequential Tabular Data. *arXiv* **2024**, arXiv:2402.01204.
174. Xie, Y.; Xu, Z.; Zhang, J.; Wang, Z.; Ji, S. Self-supervised learning of graph neural networks: A unified review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2412–2429. [CrossRef] [PubMed]
175. Hojjati, H.; Ho, T.K.K.; Armanfard, N. Self-supervised anomaly detection: A survey and outlook. *arXiv* **2023**, arXiv:2205.05173.
176. Schiappa, M.C.; Rawat, Y.S.; Shah, M. Self-supervised learning for videos: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
177. Yu, J.; Yin, H.; Xia, X.; Chen, T.; Li, J.; Huang, Z. Self-supervised learning for recommender systems: A survey. *IEEE Trans. Knowl. Data Eng.* **2023**, *36*, 335–355. [CrossRef]
178. Ucar, T.; Hajiramezanali, E.; Edwards, L. Subtab: Subsetting features of tabular data for self-supervised representation learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 18853–18865.
179. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. *AAAI Conf. Artif. Intell.* **2021**, *35*, 6679–6687. [CrossRef]
180. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
181. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–22. [CrossRef]
182. Elidan, G.; Friedman, N. The information bottleneck EM algorithm. *arXiv* **2012**, arXiv:1212.2460.
183. Xu, Y.; Zhao, S.; Song, J.; Stewart, R.; Ermon, S. A theory of usable information under computational constraints. *arXiv* **2020**, arXiv:2002.10689.
184. Graves, A. Generating sequences with recurrent neural networks. *arXiv* **2013**, arXiv:1308.0850.

185. Germain, M.; Gregor, K.; Murray, I.; Larochelle, H. Made: Masked autoencoder for distribution estimation. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 881–889.
186. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using real nvp. *arXiv* **2017**, arXiv:1605.08803.
187. Rezende, D.; Mohamed, S. Variational inference with normalizing flows. In Proceedings of the International conference on machine learning, PMLR, Lille, France, 7–9 July 2015; pp. 1530–1538.
188. Huembeli, P.; Arrazola, J.M.; Killoran, N.; Mohseni, M.; Wittek, P. The physics of energy-based models. *Quantum Mach. Intell.* **2022**, *4*, 1–13. [CrossRef]
189. Song, Y.; Kingma, D.P. How to train your energy-based models. *arXiv* **2021**, arXiv:2101.03288.
190. Younes, L. On The Convergence Of Markovian Stochastic Algorithms With Rapidly Decreasing Ergodicity Rates. *Stochastics Stochastics Model.* **1999**, *65*, 177–228. [CrossRef]
191. Hyvärinen, A. Some Extensions of Score Matching. 2006. Available online: https://www.sciencedirect.com/science/article/abs/pii/S0167947306003264 (accessed on 12 February 2024).
192. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2021**, arXiv:2011.13456.
193. Vincent, P. A Connection Between Score Matching and Denoising Autoencoders. *Neural Comput.* **2011**, *23*, 1661–1674. [CrossRef] [PubMed]
194. Song, Y.; Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
195. Lou, Z.; Ye, Y.; Yan, X. The multi-feature information bottleneck with application to unsupervised image categorization. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
196. Friedman, N.; Mosenzon, O.; Slonim, N.; Tishby, N. Multivariate information bottleneck. *arXiv* **2001**, arXiv:1301.2270.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.