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

Ravid Shwartz-Ziv *New York University* ravid.shwartz.ziv@nyu.edu
Yann LeCun *New York University & Meta AI - FAIR*

Abstract

Deep neural networks have demonstrated remarkable performance in supervised learning tasks but require large amounts of labeled data. Self-supervised learning offers an alternative paradigm, enabling the model to learn from data without explicit labels. Information theory has been instrumental in understanding and optimizing deep neural networks. Specifically, the information bottleneck principle has been applied to optimize the trade-off between compression and relevant information preservation in supervised settings. However, the optimal information objective in self-supervised learning remains unclear. In this paper, we review various approaches to self-supervised learning from an information-theoretic standpoint and present a unified framework that formalizes the *self-supervised information-theoretic learning problem.* We integrate existing research into a coherent framework, examine recent self-supervised methods, and identify research opportunities and challenges. Moreover, we discuss empirical measurement of information-theoretic quantities and their estimators. This paper offers a comprehensive review of the intersection between information theory, self-supervised learning, and deep neural networks.

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 (Alam et al., 2020; He et al., 2015; LeCun et al., 2015). 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. Self-supervised learning (SSL) emerges as a promising alternative, enabling models to learn from data without explicit labels by leveraging the underlying structure and relationships within the data itself.

Recent advances in SSL have been driven by joint embedding architectures, such as Siamese Nets (Bromley et al., 1993), DrLIM (Chopra et al., 2005; Hadsell et al., 2006), and SimCLR (Chen et al., 2020a). 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 (Caron et al., 2021; Chen et al., 2020a; Misra and van der Maaten, 2020). Although SSL methods have shown promising results in practice, the theoretical underpinnings behind their effectiveness remain an open question (Arora et al., 2019; Lee et al., 2021a).

Information theory has played a crucial role in understanding and optimizing deep neural networks, from practical applications like the variational information bottleneck (Alemi et al., 2016) to theoretical investigations of generalization bounds induced by mutual information (Steinke and Zakynthinou, 2020; Xu and Raginsky, 2017). 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) (Belghazi et al., 2018b) combined with the information maximization (InfoMax) principle (Linsker, 1988). 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 (Tishby et al., 1999a), 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 the 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 need to be addressed.

The structure of the 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. Section 6 concludes the 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 inspire 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 has gained increasing attention and great practical success by using complementary information from multiple features or modalities. The multiview learning paradigm divides the input variable into multiple views from which the target variable should be predicted (Zhao et al., 2017b). 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 (Yan et al., 2021). 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 to identify 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 (Yan et al., 2021). Thus, multiview representation learning involves identifying the underlying data structure and attempting to integrate the different views into a common feature space, resulting in high performance. In recent decades, multiview learning has been used for many machine learning tasks and influenced many algorithms, such as co-training mechanisms (Kumar and Daumé, 2011), subspace learning methods (Xue et al., 2019), and multiple kernel learning (MKL) (Bach and Jordan, 2002). Li et al. (2018) 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) 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) (Hotelling, 1936) and its kernel extensions (Bach and Jordan, 2003; Hardoon et al., 2004; Sun, 2013). In addition to CCA, multiview representation learning has penetrated a variety of learning methods, such as dimensionality reduction (Sun et al., 2010), clustering analysis (Yan et al., 2015), multiview sparse coding (Cao et al., 2013; Jia et al., 2010; Liu et al., 2014), and multimodal topic learning (Pu et al., 2020). However, despite their promising results, these methods use handcrafted features and linear embedding functions, which cannot capture the nonlinear properties of multiview data.

The emergence of deep learning has provided a powerful way to learn complex, nonlinear, and hierarchical representations of data. By incorporating multiple hierarchical layers, deep learning algorithms are able to 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 (Andrew et al., 2013) as an extension of CCA, multiview clustering via deep matrix factorization (Zhao et al., 2017a), and the deep multiview spectral network (Huang et al., 2019). Moreover, deep architectures have been employed to generate effective representations in methods such as multiview convolutional neural networks (Liu et al., 2021a), multimodal deep Boltzmann machines (Srivastava and Salakhutdinov, 2014), multimodal deep autoencoders (Ngiam et al., 2011; Wang et al., 2015), and multimodal recurrent neural networks (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Mao et al., 2014).

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. A variety of such signals have been used to learn representations, including generative and joint embedding architectures (Bachman et al., 2019; Chen et al., 2020a,b).

Two main categories of SSL architectures exist: (1) generative architectures based on reconstruction or prediction, and (2) joint embedding architectures (Liu et al., 2021b). 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 thereof, potentially with the assistance of a latent variable. Notable examples of generative architectures include auto-encoders, sparse coding, sparse auto-encoders, and variational auto-encoders (Kingma and Welling, 2013; Lee et al., 2006; Ng et al., 2011). 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 auto-encoders restrict the number of non-zero components. Variational auto-encoders regularize the latent variable by rendering it stochastic and maximizing the entropy of the distribution relative to a prior. Vector quantized variational auto-encoders (VQ-VAE) employ binary stochastic variables to achieve similar results (Van Den Oord et al., 2017).

2. **Joint Embedding Architectures (JEA):** 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 (Chen et al., 2020a; Chen and He, 2021; Grill et al., 2020; He et al., 2020), and methods permitting encoders to differ (Bardes et al., 2021). 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 effectively train these architectures, 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 (Devlin et al., 2018; He et al., 2022; Vincent et al., 2008) using 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 to be 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 examples of self-supervised visual representation learning include MoCo (He et al., 2020) and SimCLR (Chen et al., 2020a). The InfoNCE loss is a commonly used objective function in many contrastive learning methods:

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

where $x+$ is a sample similar to x, x^k are all the samples in the batch, and f is an encoder.

However, contrastive methods heavily depend on all other samples in the batch and require a large batch size. Additionally, recent studies (Jing et al., 2021) 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 (Chen and He, 2021; Grill et al., 2020), while Caron et al. (2020) employed an additional clustering step. VICReg (Bardes et al., 2021) is another non-contrastive method that regularizes the covariance matrix of representation. Consider two embedding batches $\boldsymbol{Z} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N)]$ and $\boldsymbol{Z'} = [f(\boldsymbol{x'}1), \dots, f(\boldsymbol{x'}N)]$, 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:

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

2.3 Semi-Supervised Learning

Semi-supervised learning employs both labeled and unlabeled data to enhance the model performance (Chapelle et al., 2009). Consistency regularization-based approaches (Laine and Aila, 2016; Miyato et al., 2018; Sohn et al., 2020) ensure that predictions remain stable under perturbations in input data and model parameters. Certain techniques, such as those proposed by Grandvalet and Bengio (2006) and Miyato et al. (2018), involve training a model by incorporating a regularization term into a supervised cross-entropy loss. In contrast, Xie et al. (2020) utilizes suitably weighted unsupervised regularization terms, while Zhai et al. (2019) 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 (Lee et al., 2013).

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 (Bengio et al., 2013). By extracting relevant information for classification and prediction tasks from the data, we can improve performance and reduce computational complexity (Goodfellow et al., 2016). 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 (LeCun et al., 2015). Bengio and LeCun (2007) 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. As a result, identifying relevant input features becomes challenging, as most input entropy is unrelated to the output (Shwartz-Ziv and Tishby, 2017).

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 captures all the information about Y in X. Cover (1999) proved this property:

Theorem 2 *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 3 *(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, MSS are the simplest sufficient statistics, inducing the coarsest sufficient partition on X. In MSS, the values of X are grouped into as few partitions as possible without sacrificing information. MSS are statistics with the maximum information about Y while retaining the least information about X as possible (Koopman, 1936).

2.4.2 The Information Bottleneck

The majority of distributions lack exact minimal sufficient statistics, leading Tishby et al. (1999b) 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 (Buesing and Maass, 2010; Palmer et al., 2015), slow feature analysis (Turner and Sahani, 2007), speech recognition (Hecht et al., 2009), and deep learning (Alemi et al., 2016; Shwartz-Ziv and Tishby, 2017).

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 $I\left(T;X\right)$ while maximizing $I\left(T;Y\right)$. Formally, the IB optimization involves minimizing the following objective function:

$$\mathcal{L} = \min_{P(t|x);p(y|t)} I(X;T) - \beta I(Y;T) \ , \tag{1}$$

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:

$$I(T:Y) = I(X:Y) - \mathbb{E}_{x\sim P(X),t\sim P(T|x)}\left[D\left[P(Y|x)||P(Y|t)\right]\right] \tag{2}$$

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 (Tishby et al., 1999b) 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.

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

Several issues arise with the 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 (Erdogmus, 2002). This loss has been employed in various algorithms, including the InfoMax principle (Linsker, 1988), tree-based algorithms (Quinlan, 2014), deep neural networks (Zhang and Sabuncu, 2018), and Bayesian modeling (Wenzel et al., 2020). Painsky and Wornell (2018) provided a rigorous justification for using the logarithmic loss and showed that it is an upper bound to any choice of 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:

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

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

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

Interestingly, the relationship between the true loss and the empirical loss can be bounded using the information bottleneck term. Shamir et al. (2010) 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 eq. (1) presents a trade-off between two opposing forces. On 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. (2018) 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 (2019) and Xu and Raginsky (2017) 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 (2018) 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. (2019) found that the generalization precision is positively correlated with the degree of compression of the last layer in the network. Shwartz-Ziv et al. (2019) 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.

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 performance, 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 better understand the process of learning optimal representations and explore recent methods working towards this goal.

The development of 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, IB approach, originally introduced by Tishby et al. (1999b), defines the optimal representation in supervised scenarios, enabling us to identify which terms to compress during learning. 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

The choice of 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 attempting to predict one representation based on the other, $\tilde{Z}_1|Z_2 = q_{\eta_1}(Z_1|Z_2)$ and $\tilde{Z}_2|Z_1 = q_{\eta_2}(Z_2|Z_1)$. Figure 1 illustrates this structure.

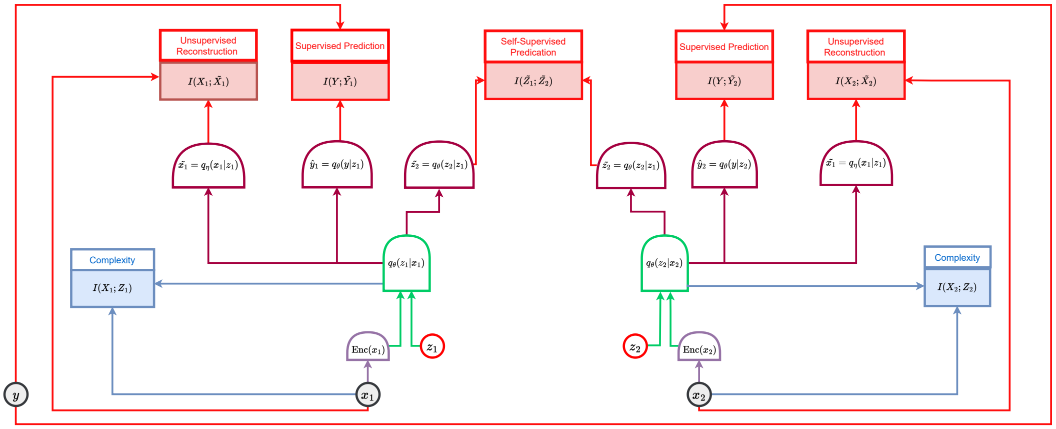
Unsupervised Reconstruction
$I(X_1;\tilde{X}_1)$
Supervised Prediction
$I(Y;\tilde{Y}_1)$
Self-Supervised Predication
$I(\tilde{Z}_1;\tilde{Z}_2)$
Supervised Prediction
$I(Y;\tilde{Y}_2)$
Unsupervised Reconstruction
$I(X_2;\tilde{X}_2)$
$\tilde{x}_1 = q_\eta(x_1|z_1)$
$\hat{y}_1 = q_\theta(y|z_1)$
$\tilde{z}_2 = q_\theta(z_2|z_1)$
$\tilde{z}_2 = q_\theta(z_2|z_1)$
$\hat{y}_2 = q_\theta(y|z_2)$
$\tilde{x}_1 = q_\eta(x_1|z_1)$
Complexity
$I(X_1;Z_1)$
$q_\theta(z_1|x_1)$
$q_\theta(z_2|x_2)$
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

The information-theoretic perspective of self-supervised networks has led to confusion in recent work regarding the information being optimized. 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 its corresponding information always stem from a single encoder and decoder.

However, in the self-supervised multiview scenario, we can construct our representation using various encoders and decoders. For instance, to define the information involved in $I(X_1; Z_1)$, we need to specify the associated random variable. 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 paths," 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 $P(X_1) \to P(Z_1|X_1) \to P(Z_2|Z_1)$. 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. The goal is to predict a label $y \in \mathcal{Y}$ using the learned representation. Achille and Soatto (2018) defined the sufficiency of Z for Y as the amount of label information retained after passing data through the encoder:

Definition 4 ***Sufficiency**: A representation Z of X is sufficient for Y if and only if $I(X;Y|Z) = 0$.*

Federici et al. (2020) showed that Z is sufficient for Y if and only if the amount of information regarding the task remains unchanged by the encoding procedure. 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:

$$\mathcal{L} = I(X;Z) - \beta I(Z;Y) \tag{3}$$

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. (2020) proposed splitting $I(X,Z)$ into two terms:

$$I(X;Z) = \underbrace{I(X;Z|Y)}_{\text{superfluous information}} + \underbrace{I(Z;Y)}_{\text{predictive information}} \tag{4}$$

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 (Shwartz-Ziv and Tishby, 2017): the fitting phase and the compression phase. The fitting phase involves extracting information from the input and converting it into learned representations, characterized by an increase in mutual information between inputs and hidden representations. Conversely, the compression phase, which is much longer, concentrates on discarding unnecessary information for target prediction, resulting in a decrease in mutual information between learned representations and inputs, while the mutual information between representations and targets increases. For more information, see Geiger (2020).

Despite the elegance and plausibility of the IB hypothesis, empirically investigating it remains challenging (Amjad and Geiger, 2018).

The study of representation compression in Deep Neural Networks (DNNs) for supervised learning has shown inconsistent results. For instance, Chelombiev et al. (2019) discovered a positive correlation between generalization accuracy and the compression level of the network's final layer. Shwartz-Ziv et al. (2018) also examined the relationship between generalization and compression, demonstrating that generalization error exponentially depends on mutual information, $I(X;Z)$. Furthermore, Achille et al. (2017) established that flat minima, known for their improved generalization properties, constrain the mutual information. However, Saxe et al. (2019) showed that compression was not necessary for generalization in deep linear networks. Basirat et al. (2021) 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 (Pensia et al., 2018; Xu and Raginsky, 2017). Lastly, Achille and Soatto (2018) explored the use of 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 but 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. (2014) introduced the large-margin multiview IB (LMIB) as an extension of the original IB problem. 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. By limiting Rademacher complexity, the solution's accuracy and generalization error bounds are improved. 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. (2019) 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:

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

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

In many practical scenarios, obtaining labeled data can be challenging or expensive, while a large number of unlabeled samples may be readily available. Semi-supervised learning aims to address 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. (2019) grouped regularization methods into three primary categories: entropy minimization, consistency regularization, and generic regularization.

Voloshynovskiy et al. (2020) 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 is 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:

Hand-Crafted 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. Hand-crafted priors in this context are similar to priors used in the Variational Information Bottleneck (VIB) formalism (Alemi et al., 2016; Wang et al., 2019).

Learnable Priors: Voloshynovskiy et al. (2020) 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 auto-encoder $q(m|z)$ 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:

$$\begin{aligned}\mathcal{L} &= D_{KL}(q(m|z)||p(m|z)) - \beta D_{KL}(q(x|m)||p(x|m)) - \beta_y D_{KL}(p(y|z)||p(y)) \\ &\Leftrightarrow I(M;Z) - \beta I(M;X) - \beta_y I(Y;Z)\end{aligned} \tag{5}$$

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. (2020) demonstrated that various popular semi-supervised methods can be considered special cases of the optimization problem described above. Notably, the semi-supervised AAE (Makhzani et al., 2015), CatGAN (Springenberg, 2015), SeGMA (Smieja et al., 2019), and VAE (Kingma et al., 2014) 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. (2020) defined 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 unsupervised IB can be defined as follows:

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

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, unsupervised IB is a special case of supervised IB where labels are replaced with the reconstruction performance of the training input. Alemi et al. (2016) showed that Variational Autoencoder (VAE) (Kingma and Welling, 2019) and β-VAE (Higgins et al., 2017) are special cases of unsupervised variational IB. Voloshynovskiy et al. (2020) extended their results and showed that many models, including adversarial autoencoders (Makhzani et al., 2015), InfoVAEs (Zhao et al., 2017c), and VAE/GANs (Larsen et al., 2016), could be viewed as special cases of unsupervised IB. The main difference between them is the bounds on the different mutual information of the IB. Furthermore, unsupervised IB was used by Uğur et al. (2020) to derive lower bounds for their unsupervised generative clustering framework, while Roy et al. (2018) used it to study vector-quantized autoencoders.

Voloshynovskiy et al. (2020) pointed out that for the classification task in supervised IB, the latent space Z should be 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? In this section, we analyze and discuss how to achieve 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 (Linsker, 1988; Wiskott and Sejnowski, 2002). As one of the earliest approaches, Linsker (1988) proposed maximizing 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 (1992) 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 (Bell and Sejnowski, 1995), 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 (Hjelm et al., 2018) 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 (Bachman et al., 2019; Henaff, 2020; Hjelm et al., 2018; Tian et al., 2020a), 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. (2019) 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 for compressing the representation to enhance generalizability.

To resolve this problem, Sridharan and Kakade (2008) proposed the *multiview IB framework*. 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 found in the input regarding downstream tasks. By compressing the information not shared between the two views, we improve generalization without affecting performance. 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 (Hjelm et al., 2018), 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. Image augmentations can be perceived as altering an image's style without changing its content. Thus, Tsai et al. (2020) 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 not necessary 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. (2020) used the relaxed Lagrangian objective to obtain the minimal sufficient representation Z_1 for X_2 as:

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

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

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

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:

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

where D_{SKL} represents the symmetrized KL divergence obtained by averaging the expected value 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. (2020) and Tsai et al. (2020) minimize the same objective. Another implementation of the same idea is based on the Conditional Entropy Bottleneck (CEB) algorithm (Fischer, 2020) and proposed by Lee et al. (2021b). 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 above-mentioned algorithms are based on the Multi-view 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. (2021) 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 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. (2021) 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:

$$\mathcal{L} = \min_{P(Z_1|X_1)} I_{P(Z_1|X_1)}(X_1; Z_1) + \beta H(Z_2 \mid X_1). \tag{6}$$

By 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 amount of 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 (Tian et al., 2020b; Wang et al., 2022). 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 (Linsker, 1988), 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:

$$I(Z_1; Z_2) = H(Z) - H(Z_1 \mid Z_2) \geq H(Z_1) + \mathbb{E}[\log q(z_1 \mid z_2)] \tag{7}$$

The bound is tight when $q(z_1|z_2) = p(z_1|z_2)$, in which case the first term equals the conditional entropy $H(Z_1|Z_2)$. The second term of eq. (7) can be thought of as a negative reconstruction error or distortion between Z_1 and 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

$$I(Y; Z) \geq H(Y) + \mathbb{E}\left[\log q(Y \mid Z)\right] \tag{8}$$

. Since Y is constant, optimizing the information $I(Z; Y)$ requires only minimizing the prediction term $\mathbb{E}\left[\log q(Y|Z)\right]$ 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 of the representation (large entropy) is significant only in the directions that have a high variance in the data, which is enforced by data augmentation (Jing et al., 2021). 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. (2020b) 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 the views. 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 performance.

Kahana and Hoshen (2022) 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. (2022) 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 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. (2022) and Kahana and Hoshen (2022) 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:

$$\mathcal{L} = \min_{P(Z_1|X_1),p(Z_2|X_2)} -I_{P(Z_1|X_1)}(X_1; Z_1) - I_{P(Z_2|X_2)}(X_2; Z_2) - \beta I_{P(Z_1|X_1),P(Z_2|Z_1)}(Z_1; Z_2). \tag{9}$$

Wang et al. (2022) demonstrated the effectiveness of their method for SimCLR (Chen et al., 2020a), BYOL (Grill et al., 2020), and Barlow Twins (Zbontar et al., 2021) across classification, detection, and segmentation tasks.

4.3 To Compress or Not to Compress?

As seen in Eq. 9, 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 the development and evaluation of 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 (Shwartz-Ziv and Tishby, 2017). Despite their elegance and plausibility, measuring and analyzing information in deep networks empirically presents challenges. Two critical problems are (1) information in deterministic networks and (2) estimating information in high-dimensional spaces.

Information in Deterministic Networks

Information-theoretic methods have made a significant impact on deep learning (Alemi et al., 2016; Shwartz-Ziv and Tishby, 2017; Steinke and Zakynthinou, 2020). 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 (Amjad and Geiger, 2019; Goldfeld et al., 2018). 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 (Shwartz-Ziv and Tishby, 2017).

However, binning and discretization are essentially equivalent to geometrical compression and serve as clustering measures (Goldfeld et al., 2018). Moreover, this discretization depends on the chosen bin size and does not track the mutual information across varying bin sizes Goldfeld et al. (2018); Ross (2014). To address these limitations, researchers have proposed alternative approaches such as interpreting binned information as a weight decay penalty Elad et al. (2019b), estimating mutual information based on lower bounds assuming a continuous input distribution without making assumptions about the network's output distribution properties (Wang and Isola, 2020; Zimmermann et al., 2021), injecting additive noise, and considering data augmentation as the source of noise (Dubois et al., 2021; Goldfeld et al., 2018; Lee et al., 2021b; Shwartz-Ziv and Tishby, 2017).

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 (Gao et al., 2015; Paninski, 2003), revealing the inefficiency of solutions for large dimensions and the limited scalability of known approximations with respect to sample size and dimension. Despite these difficulties, various entropy and mutual information estimation approaches have been developed, including classic methods like k-nearest neighbors (KNN) (Kozachenko and Leonenko, 1987) and kernel density estimation techniques (Hang et al., 2018), as well as more recent efficient methods.

Chelombiev et al. (2019) developed adaptive mutual information estimators based on entropies-equal bins and scaled noise kernel density estimator. Generative decoder networks, such as PixelCNN++ (Van den Oord et al., 2016), have been employed to estimate a lower bound on mutual information (Darlow and Storkey, 2020; Nash et al., 2018). Another strategy includes ensemble dependency graph estimators, adaptive mutual information estimation methods (EDGE) by merging randomized locality-sensitive hashing (LSH), dependency graphs, and ensemble bias reduction techniques (Noshad and Hero III, 2018). The Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018a) maximizes KL divergence using the dual representation of Donsker and Varadhan (1975) and has been employed for direct mutual information estimation (Elad et al., 2019a).

Improving mutual information estimation can be achieved by using larger batch sizes, although this may negatively impact 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 (Erdogan, 2022; Ozsoy et al., 2022), which reflects linear dependence. Goldfeld and Greenewald (2021) 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 and can be estimated with optimal parametric error rates in all dimensions by combining an MI estimator between scalar variables with an MC integrator (Goldfeld and Greenewald, 2021). The k-SMI, introduced by Goldfeld et al. (2022), extends the SMI by projecting to k-dimensional subspace, which relaxes the smoothness assumptions, improves scalability, and enhances performance.

In conclusion, estimating and optimizing information in deep neural networks presents 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. Future Research Directions

Despite the solid foundation established by existing self-supervised learning methods from an information theory perspective, there are several potential research directions that 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 how to differentiate 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 4.1. As a consequence, we must heavily rely on the process of 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 (Arik and Pfister, 2021; Ucar et al., 2021), further work is necessary from both theoretical and practical standpoints to achieve high performance with self-supervised learning for tabular data (Shwartz-Ziv and Armon, 2022). 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 (Chen and He, 2021; Grill et al., 2020). The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) 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 (2012) proposed an IB-inspired version of the EM, 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 (Cover, 1999). 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. (2020) 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 specific optimization of model parameters concerning the tractable log-likelihood (Dinh et al., 2016; Germain et al., 2015; Graves, 2013; Rezende and Mohamed, 2015) or a tractable lower bound of the likelihood (Alemi et al., 2016; Kingma and Welling, 2019). 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 (Huembeli et al., 2022; Song and Kingma, 2021).

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 (Younes, 1999), score matching (Hyvärinen, 2006), denoising score matching (Song et al., 2020; Vincent, 2011), and score-based generation models (Song and Ermon, 2019).

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) (Lou et al., 2013), 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 (Friedman et al., 2013) and setting Bayesian networks.

7. Conclusion

In this paper, we delved into the concept of optimal representation in self-supervised learning from an information theory perspective. We reviewed various approaches to the problem, emphasizing their assumptions and limitations, and integrated them into a cohesive framework. Additionally, we discussed several information-theoretic terms influencing optimal representation and methods for estimating them.

Despite existing challenges in defining and optimizing optimal representation in self-supervised learning, information theory furnishes a robust and versatile framework for analysis and algorithm development. It presents a flexible methodology applicable to a diverse array of learning models and facilitates understanding the implicit and explicit assumptions of data and model optimization.

Promising future research directions encompass expanding the multiview framework to accommodate more views and tasks, investigating energy-based models for self-supervised learning, and exploring information theory's role in other deep learning aspects, such as reinforcement learning and generative models.

In conclusion, information theory serves as a valuable resource for developing and comprehending self-supervised learning models. Employing information theory enables us to enhance our understanding of deep neural network learning processes and, ultimately, construct more effective models.

In this paper, we investigated the concept of optimal representation in self-supervised learning from an information theory perspective. We analyzed various approaches, highlighting their assumptions and limitations, and integrated them into a unified, comprehensive framework. Furthermore, we discussed several information-theoretic terms that influence optimal representations and explored methods for estimating them.

In supervised and unsupervised learning, we have direct access to relevant information, enabling us to separate and compress irrelevant information. However, self-supervised learning heavily relies on assumptions about the relationship between the data and downstream tasks to define and compress irrelevant information. When these assumptions are violated, separating information into relevant and irrelevant components becomes challenging, often leading to suboptimal performance.

Despite the challenges in defining and optimizing optimal representation in self-supervised learning, information theory offers a robust and versatile framework for analysis and algorithm development. This framework is applicable to a wide array of learning models and contributes to understanding the implicit and explicit assumptions of data and model optimization.

As datasets continue to grow in size and models are increasingly expected to serve as base models for various downstream tasks, reliance on the Multi-view assumption becomes less appropriate. Consequently, one of the most significant challenges in self-supervised learning is compressing irrelevant information when this assumption does not hold. Identifying new methods to separate relevant from irrelevant information based on alternative assumptions presents a promising avenue for research.

Moreover, potential future research directions include expanding the Multi-view framework to accommodate additional views and tasks, investigating energy-based models for self-supervised learning, and exploring the role of information theory in other deep learning aspects, such as reinforcement learning and generative models.

In conclusion, information theory serves as an invaluable resource for developing and understanding self-supervised learning models. By leveraging information theory, we can enhance our comprehension of deep neural network learning processes and ultimately construct more effective models which depend on our assumptions.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*, 2017.

Mahbubul Alam, Manar D Samad, Lasitha Vidyaratne, Alexander Glandon, and Khan M Iftekharuddin. Survey on deep neural networks in speech and vision systems. *Neurocomputing*, 417:302–321, 2020.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv:1612.00410*, 2016. URL http://arxiv.org/abs/1612.00410.

Rana Ali Amjad and Bernhard C Geiger. How (not) to train your neural network using the information bottleneck principle. *arXiv preprint arXiv:1802.09766*, 2018.

Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3(null):1–48, mar 2003. ISSN 1532-4435. doi: 10.1162/153244303768966085. URL https://doi.org/10.1162/153244303768966085.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mina Basirat, Bernhard C. Geiger, and Peter M. Roth. A geometric perspective on information plane analysis. *Entropy*, 23(6), 2021. ISSN 1099-4300. URL https://www.mdpi.com/1099-4300/23/6/711.

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, 2018a.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018b.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards ai ,in l. bottou, o. chapelle, d. decoste, and j. weston, editors,. *Large Scale Kernel Machines,MIT Press.*, 2007.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Lars Buesing and Wolfgang Maass. A spiking neuron as information bottleneck. *Neural computation*, 22(8):1961–1992, 2010.

Tian Cao, Vladimir Jojic, Shannon Modla, Debbie Powell, Kirk Czymmek, and Marc Niethammer. Robust multimodal dictionary learning. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 259–266, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40811-3.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20 (3):542–542, 2009.

Ivan Chelombiev, Conor Houghton, and Cian O'Donnell. Adaptive estimators show information compression in deep neural networks. *arXiv preprint arXiv:1902.09037*, 2019.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

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

Luke Nicholas Darlow and Amos Storkey. What information does a resnet compress? *arXiv preprint arXiv:2003.06254*, 2020.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

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

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

Adar Elad, Doron Haviv, Yochai Blau, and Tomer Michaeli. Direct validation of the information bottleneck principle for deep nets. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019a.

Adar Elad, Doron Haviv, Yochai Blau, and Tomer Michaeli. The effectiveness of layer-by-layer training using the information bottleneck principle, 2019b. URL https://openreview.net/forum?id=r1Nb5i05tX.

Gal Elidan and Nir Friedman. The information bottleneck em algorithm. *arXiv preprint arXiv:1212.2460*, 2012.

Alper T Erdogan. An information maximization based blind source separation approach for dependent and independent sources. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4378–4382. IEEE, 2022.

Deniz Erdogmus. *Information theoretic learning: Renyi's entropy and its applications to adaptive system training*. University of Florida, 2002.

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

Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020.

Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck. *arXiv preprint arXiv:1301.2270*, 2013.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial Intelligence and Statistics*, pages 277–286, 2015.

Bernhard C Geiger. On information plane analyses of neural network classifiers–a review. *arXiv preprint arXiv:2003.09671*, 2020.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889. PMLR, 2015.

Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. Estimating Information Flow in Neural Networks. *ArXiv e-prints*, 2018.

Ziv Goldfeld and Kristjan Greenewald. Sliced mutual information: A scalable measure of statistical dependence. *Advances in Neural Information Processing Systems*, 34:17567–17578, 2021.

Ziv Goldfeld, Kristjan Greenewald, Theshani Nuradha, and Galen Reeves. k-sliced mutual information: A quantitative study of scalability with dimension. *arXiv preprint arXiv:2206.08526*, 2022.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016. URL http://www.deeplearningbook.org.

Yves Grandvalet and Yoshua Bengio. Entropy regularization., 2006.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

Hanyuan Hang, Ingo Steinwart, Yunlong Feng, and Johan AK Suykens. Kernel density estimation for dynamical systems. *The Journal of Machine Learning Research*, 19(1): 1260–1308, 2018.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. doi: 10.1162/0899766042321814.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Ron M Hecht, Elad Noor, and Naftali Tishby. Speaker recognition by gaussian information bottleneck. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 00063444. URL http://www.jstor.org/stable/2333955.

Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *IJCAI*, pages 2563–2569, 2019.

Patrick Huembeli, Juan Miguel Arrazola, Nathan Killoran, Masoud Mohseni, and Peter Wittek. The physics of energy-based models. *Quantum Machine Intelligence*, 4(1):1–13, 2022.

Aapo Hyvärinen. Some extensions of score matching, 2006.

Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Factorized latent spaces with structured sparsity. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/a49e9411d64ff53eccfdd09ad10a15b3-Paper.pdf.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Jonathan Kahana and Yedid Hoshen. A contrastive objective for learning disentangled representations. *arXiv preprint arXiv:2203.11284*, 2022.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Lyudmyla F Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Abhishek Kumar and Hal Daumé. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 393–400. Citeseer, 2011.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, ””, 2015.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/2d71b2ae158c7c5912cc0bbde2bb9d95-Paper.pdf.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34, 2021b.

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

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

Shiming Liu, Yifan Xia, Zhusheng Shi, Hui Yu, Zhiqiang Li, and Jianguo Lin. Deep learning in sheet metal bending with a novel theory-guided deep neural network. *IEEE/CAA Journal of Automatica Sinica*, 8(3):565–581, 2021a.

Weifeng Liu, Dacheng Tao, Jun Cheng, and Yuanyan Tang. Multiview hessian discriminative sparse coding for image annotation. *Computer Vision and Image Understanding*, 118: 50–60, 2014. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2013.03.007. URL https://www.sciencedirect.com/science/article/pii/S1077314213001550.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021b.

Zhengzheng Lou, Yangdong Ye, and Xiaoqiang Yan. The multi-feature information bottleneck with application to unsupervised image categorization. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Charlie Nash, Nate Kushman, and Christopher KI Williams. Inverting supervised representations with autoregressive neural density models. *arXiv preprint arXiv:1806.00400*, 2018.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Morteza Noshad and Alfred O Hero III. Scalable mutual information estimation using dependence graphs. *arXiv preprint arXiv:1801.09125*, 2018.

Serdar Ozsoy, Shadi Hamdan, Sercan Arik, Deniz Yuret, and Alper Erdogan. Self-supervised learning with an information maximization criterion. *Advances in Neural Information Processing Systems*, 35:35240–35253, 2022.

Amichai Painsky and Gregory W Wornell. On the universality of the logistic loss function. *arXiv preprint arXiv:1805.03804*, 2018.

Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6): 1191–1253, 2003. ISSN 0899-7667. doi: 10.1162/089976603321780272.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

Shi Pu, Yijiang He, Zheng Li, and Mao Zheng. Multimodal topic learning for video recommendation. *arXiv preprint arXiv:2010.13373*, 2020.

J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Brian C Ross. Mutual information between discrete and continuous data sets. *PLoS ONE*, 9 (2):e87357, 2014. doi: 10.1371/journal.pone.0087357. URL https://doi.org/10.1371/journal.pone.0087357.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*, 2018.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

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

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696 – 2711, 2010. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2010.04.006. URL http://www.sciencedirect.com/science/article/pii/S030439751000201X. Algorithmic Learning Theory (ALT 2008).

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

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

Ravid Shwartz-Ziv, Amichai Painsky, and Naftali Tishby. Representation compression and generalization in deep neural networks. 2018.

Ravid Shwartz-Ziv, Amichai Painsky, and Naftali Tishby. REPRESENTATION COMPRESSION AND GENERALIZATION IN DEEP NEURAL NETWORKS, 2019. URL https://openreview.net/forum?id=SkeL6sCqK7.

M Smieja, M Wolczyk, J Tabor, and B Geiger. Segma: Semi-supervised gaussian mixture auto-encoder. *arXiv preprint arXiv:1906.09333*, 2019.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

Karthik Sridharan and Sham Kakade. An information theoretic framework for multi-view learning. *SO*, 01 2008.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15(84):2949–2980, 2014. URL http://jmlr.org/papers/v15/srivastava14b.html.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Liang Sun, Betul Ceran, and Jieping Ye. 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*, pages 313–322, 2010.

Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23:2031–2038, 2013.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020b.

N. Tishby, F.C. Pereira, and W. Biale. The information bottleneck method. In *The 37th annual Allerton Conf. on Communication, Control, and Computing*, pages 368–377, 1999a. URL https://arxiv.org/abs/physics/0004057.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *In Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999b.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007.

Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

Yiğit Uğur, George Arvanitakis, and Abdellatif Zaidi. Variational information bottleneck for unsupervised clustering: Deep gaussian mixture embedding. *Entropy*, 22(2):213, 2020.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Matías Vera, Pablo Piantanida, and Leonardo Rey Vega. The role of information complexity and randomization in representation learning. *arXiv preprint arXiv:1802.05355*, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Slava Voloshynovskiy, Olga Taran, Mouad Kondah, Taras Holotyak, and Danilo Rezende. Variational information bottleneck for semi-supervised classification. *Entropy*, 22(9), 2020. ISSN 1099-4300. doi: 10.3390/e22090943. URL https://www.mdpi.com/1099-4300/22/9/943.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. *arXiv preprint arXiv:2203.07004*, 2022.

Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. *Deep Multi-view Information Bottleneck*, pages 37–45. A, 2019. doi: 10.1137/1.9781611975673.5. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611975673.5.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1083–1092. JMLR.org, 2015.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świkatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002. doi: 10.1162/089976602317318938.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Chang Xu, Dacheng Tao, and Chao Xu. Large-margin multi-viewinformation bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1559–1572, 2014.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.

Zhe Xue, Junping Du, Dawei Du, and Siwei Lyu. Deep low-rank subspace ensemble for multi-view clustering. *Information Sciences*, 482:210–227, 2019. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.01.018. URL https://www.sciencedirect.com/science/article/pii/S0020025519300271.

Xiaoqiang Yan, Yangdong Ye, and Zhengzheng Lou. Unsupervised video categorization based on multivariate information bottleneck method. *Knowledge-Based Systems*, 84: 34–45, 2015.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.03.090. URL https://www.sciencedirect.com/science/article/pii/S0925231221004768.

Laurent Younes. On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates. In *STOCHASTICS AND STOCHASTICS MODELS*, pages 177–228, 1999.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Thirty-first AAAI conference on artificial intelligence*, 2017a.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017b. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2017.02.007. URL https://www.sciencedirect.com/science/article/pii/S1566253516302032.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017c.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.