Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Survey Track

Information-Theoretic Methods in Deep Neural Networks: Recent Advances and Emerging Opportunities

Shujian Yu[1] , **Luis Sanchez Giraldo**[2] , **Jose Principe**[3]
[1]NEC Laboratories Europe, 69115 Heidelberg, Germany
[2]Department of ECE, University of Kentucky, Lexington, KY 40506, USA
[3]Department of ECE, University of Florida, Gainesville, FL 32611, USA
yusj9011@gmail.com, luis.sanchez@uky.edu, principe@cnel.ufl.edu

Abstract

We present a review on the recent advances and emerging opportunities around the theme of analyzing deep neural networks (DNNs) with information-theoretic methods. We first discuss popular information-theoretic quantities and their estimators. We then introduce recent developments on information-theoretic learning principles (e.g., loss functions, regularizers and objectives) and their parameterization with DNNs. We finally briefly review current usages of information-theoretic concepts in a few modern machine learning problems and list a few emerging opportunities.

1 Introduction

Information-theoretic methods have become the workhorse of several impressive deep learning achievements over the past years, ranging from practical applications (e.g., the variational information bottleneck in representation learning [Alemi *et al.*, 2017]) to theoretical investigations (e.g., the generalization bound induced by mutual information [Xu and Raginsky, 2017; Steinke and Zakynthinou, 2020]). Further, a few novel information-theoretic quantities (e.g., mutual information) estimators and learning principles have been developed and applied to different deep learning problems in fruitful ways. A recent example is the mutual information neural estimator (MINE) [Belghazi *et al.*, 2018] and its application in representation learning [Hjelm *et al.*, 2018] with the renowned information maximization (InfoMax) principle [Linsker, 1988].

Information theory requires the knowledge of the data probability density function (PDF), and in machine learning this is normally unknown. Therefore, the application of information theory to machine learning is predicated to the selection of PDF estimators or mathematical alternatives to compute from data the statistical quantities of entropy, mutual information or divergence. The theory of information-theoretic estimators is vast [Pardo, 2018; Gao *et al.*, 2018], so it is often hard for practitioners to make a quick decision on suitable estimators. On the other hand, the rapid development of computer vision and natural language processing applications is likely to drive our attention to a particular class of estimators, neglecting other possible choices that may enjoy other favorable properties. Such restrictions will, in turn, impose constraints on the performance limit of those applications.

In this work, we first discuss popular information-theoretic quantities (i.e., entropy, mutual information and divergence) and their estimators, aiming at illustrating their inner connections and specific properties. We then introduce common information-theoretic learning principles (e.g., InfoMax and the information bottleneck (IB) approach [Tishby *et al.*, 1999]) and their practical usages in the understanding and the design of DNNs. We finally demonstrate how information theory can be brought to bear on several challenging deep learning problems in unorthodox and fruitful ways. To conclude this survey, we provide a list of future directions that we consider, have potential to move the field forward.

2 Information-Theoretic Quantities and Estimators

Information-theoretic quantities provide useful descriptions of the underlying behavior of random variable (or process) and that this behavior is a key factor in developing and analyzing deep models. Shannon entropy and relative entropy (*a.k.a.*, Kullback-Leibler or KL divergence) evidence a long track record of usefulness in information theory and machine learning. However, there is no reason to restrict ourselves to Shannon's measures of entropy and relative entropy as alternative quantities may have other properties advantageous for machine learning [Kapur, 1994].

Hence, we first introduce definitions and estimators of popular information-theoretic quantities. For conciseness, we focus our discussion on continuous random variables for which it is assumed the probability density function exits.

2.1 Shannon's Entropy, Mutual Information and KL divergence

For a continuous random variable X with probability density function (PDF) $f(x)$ and support $\mathcal{X}$, Shannon's differential entropy $H(X)$ is given by:

$$H(X) = -\int_{\mathcal{X}} f(x)\log f(x)dx = \mathbb{E}\left(-\log f(x)\right) \qquad (1)$$

Similarly, the joint entropy for a pair of random variables (X, Y), with joint PDF $f(x, y)$, is defined as:

$$H(X,Y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)\log f(x,y)dxdy \qquad (2)$$

A fundamental measure of dependence (or correlation) between two random variables is the mutual information (MI):

$$I(X;Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y) \log \left(\frac{f(x,y)}{f(x)f(y)} \right) dxdy. \quad (3)$$

MI is symmetric and reduces to zero iff X and Y are statistically independent. Another important problem in information theory is measuring the dissimilarity between two probability distributions (or in our case PDFs) $f(x)$ and $g(x)$. The first principled approach that is based on the probability measure is the KL divergence (*a.k.a.*, relative entropy), defined as:

$$D_{KL}(f(x)\|g(x)) = \int_{\mathcal{X}} f(x) \log \left(\frac{f(x)}{g(x)} \right) dx. \quad (4)$$

MI can be expressed as the KL divergence between the joint distribution $f(x,y)$ and the product of marginals $f(x)f(y)$:

$$I(X;Y) = D_{KL}(f(x,y)\|f(x)f(y)). \quad (5)$$

A natural way to estimate MI from samples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ is the plug-in approach that uses a non-parametric density estimator such as kernel density estimator (KDE) [Parzen, 1962] to approximate $f(x,y)$, $f(x)$ and $f(y)$, and then uses the estimated densities to compute MI. However, density estimation is notoriously hard in high-dimensional space. Therefore, most of the commonly employed estimators for MI are based on nearest neighbour and graph properties [Kraskov *et al.*, 2004]. One of the most recent proposals of this kind is ensemble dependency graph estimator (EDGE) [Noshad *et al.*, 2019]. Despite fast convergence rate and low computational complexity of EDGE, one downside of all estimators based on neighborhoods or graphs is their lack of differentiabilty, and thus cannot be used for gradient-based optimization that is so prevalent in DNNs.

In an effort to devise estimators that scale to present-day machine learning problems, most recent work on estimating MI has focused on variational lower bounds that can be parameterized, for instance using neural networks [Belghazi *et al.*, 2018]. These approaches do solve the differentiability issue; however, theoretical results have shown that such high confidence estimators based on the lower bound on MI require a sample size that is exponential in the MI of the data, making reliable estimation impractical in high entropy, high MI scenarios [McAllester and Stratos, 2020].

2.2 Rényi's α-order Entropy and Related Quantities

A widely used generalization of Shannon entropy is Rényi's α-entropy. For a continuous random variable X with PDF $f(x)$ and support $\mathcal{X}$, the α-entropy $H_\alpha(X)$ is defined as:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f^\alpha(x) dx. \quad (6)$$

A big difference with respect to Shannon entropy is the interchange of the "log" with the integral, which has enormous implications for estimation [Principe, 2010]. Rényi also extended the concept to divergence. The α-relative entropy) between random variables with PDFs f and g is given by:

$$D_\alpha(f\|g) = \frac{1}{\alpha-1} \log \left(\int_{\mathcal{X}} f^\alpha(x) g^{1-\alpha}(x) dx \right). \quad (7)$$

In limit case where $\alpha \to 1$, both (6) and (7) reduce to Shannon's entropy and KL divergence, respectively. There are multiple ways to generalize MI in a similar way, as first done by Rényi [Rényi, 1961]. One of the most notable generalizations on α-MI was suggested by Arimoto [Arimoto, 1977]. For simplicity, one can also define the α-MI as the α-relative entropy between $f(x,y)$ and $f(x)f(y)$ [Pál *et al.*, 2010], i.e.,

$$I_\alpha(X;Y) = D_\alpha(f(x,y)\|f(x)f(y)) \quad (8)$$

k-NN graph-based estimators have been proposed for α-order generalizations (i.e., Eqs. (6)-(8)) [Leonenko *et al.*, 2008]. However, the consistency proof of these estimators (e.g., [Pál *et al.*, 2010; Póczos and Schneider, 2011]) are usually restricted to a narrow range of $\alpha \in (0,1)$. Additionally, [Singh and Póczos, 2014] established a rate of convergence for the simple KDE estimator of α-divergence.

A modified version of Rényi's definition of α-relative entropy is given by Lutwak [Lutwak *et al.*, 2005]:

$$D_\alpha(f\|g) = \log \frac{(\int g^{\alpha-1} f)^{\frac{1}{1-\alpha}} (\int g^\alpha)^{\frac{1}{\alpha}}}{(\int f^\alpha)^{\frac{1}{\alpha(1-\alpha)}}}. \quad (9)$$

For $\alpha = 2$, expressions for entropy of f and relative entropy (Lutwak's definition) between f and g can be obtained as functions of inner products between PDFs:

$$H_2(f) = -\log \int_{\mathcal{X}} f^2(x) dx, \quad (10)$$

$$D_2(f\|g) = -\frac{1}{2} \log \frac{(\int fg)^2}{(\int f^2)(\int g^2)}. \quad (11)$$

Eq. (11) is also called the Cauchy-Schwarz (CS) divergence [Jenssen *et al.*, 2006] and is defined based on the well-known CS inequality:

$$\left| \int f(x)g(x)dx \right|^2 \leq \int |f(x)|^2 dx \int |g(x)|^2 dx. \quad (12)$$

Given Eq. (11), the quadratic mutual information (QMI) between X and Y can be defined as the CS divergence between $f(x,y)$ and $f(x)f(y)$:

$$I_{CS} = D_{CS}(f(x,y)\|f(x)f(y)). \quad (13)$$

Alternatively, one can use the Euclidean distance $D_{ED}(f\|g) = \int_{\mathcal{X}} (f(x) - g(x))^2 dx$ to define QMI as:

$$I_{ED} = D_{ED}(f(x,y)\|f(x)f(y)). \quad (14)$$

Note that definitions in Eq. (13) and Eq. (14) are not equivalent. The case of $\alpha = 2$ is of special interest as it allows simple closed form expressions for KDE-based estimators [Principe, 2010, Chapter 2.10].

2.3 Matrix-based Entropy Functional

Information-theoretic quantities can be defined (or measured) over the eigenspectrum of symmetric positive semidefinite (SPS) matrix, avoiding the necessity of estimating the underlying density distributions of variables. For example, it is reasonable that the eigenvalues of sample covariance matrices

of two collections of samples manifest their distributional divergence [Moskvina and Zhigljavsky, 2003]. More formally, given a strictly convex, differentiable function $\varphi : \mathcal{S}_+^d \to \mathbb{R}$ that maps SPS matrices (of size $d \times d$) to a real number, the Bregman matrix divergence from matrix ρ to matrix σ is defined as [Tsuda *et al.*, 2005]:

$$D_{\varphi,B}(\sigma\|\rho) = \varphi(\sigma) - \varphi(\rho) - \mathrm{tr}((\nabla\varphi(\rho))^T(\sigma - \rho)), \quad (15)$$

where $\mathrm{tr}(A)$ denotes the trace of matrix A. Examples include $\varphi(\sigma) = \|\sigma\|_F^2$, which leads to the squared Frobenius norm $\|\sigma - \rho\|_F^2$.

When $\varphi(\sigma) = \mathrm{tr}(\sigma\log\sigma - \sigma)$, where $\log\sigma$ is the matrix logarithm, the resulting Bregman matrix divergence is:

$$D_{vN}(\sigma\|\rho) = \mathrm{tr}(\sigma\log\sigma - \sigma\log\rho - \sigma + \rho), \quad (16)$$

which is also referred to von Neumann divergence or quantum relative entropy [Nielsen and Chuang, 2010]. Another important matrix divergence arises by taking $\varphi(\sigma) = -\log\det\sigma$, in which the resulting Bregman matrix divergence reduces to:

$$D_{\ell D}(\sigma\|\rho) = \mathrm{tr}(\sigma\rho^{-1}) - \log\det(\sigma\rho^{-1}) - d, \quad (17)$$

and is commonly called the LogDet divergence.

Given two sets of observations $S_X = \{\mathbf{x}_i\}_{i=1}^n \sim f$ and $S_Y = \{\mathbf{y}_i\}_{i=1}^m \sim g$, the divergence from f to g can be simply measured by $D_{\varphi,B}$ on their respective sample covariance matrices Σ_x and Σ_y with $D_{\varphi,B}(\Sigma_x\|\Sigma_y)$. Taking the chain rule of KL divergence, $D_{\varphi,B}$ can also be used to quantify the divergence of two conditional distributions (i.e., $D(f(y|x)\|g(y|x))$) [Yu *et al.*, 2020b].

Similarly, one is able to non-parametrically obtain measures of entropy and MI on the eigenspectrum of a Gram matrix of the projected data in a reproducing kernel Hilbert space (RKHS) [Sanchez Giraldo *et al.*, 2014]. Specifically, the matrix-based Rényi's α-order entropy for samples $\{\mathbf{x}_i\}_{i=1}^n$ (each $\mathbf{x}_i$ can be a real-valued scalar or vector) is defined over the eigenspectrum of their normalized Gram matrix A of size $n \times n$ as follows:

$$H_\alpha(A) = \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}(A^\alpha)\right) \quad (18)$$

where $\mathrm{tr}(A^\alpha) = \sum_{i=1}^n \lambda_i(A)^\alpha$, and$A = K/\mathrm{tr}(K)$ in which $K = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ is sample Gram matrix[1] and tr denotes matrix trace. $\lambda_i(A)$ denotes the i-th eigenvalue of A.

Let B be another (normalized) Gram matrix from $\{\mathbf{y}_i\}_{i=1}^n$, the joint entropy between X and Y is defined as:

$$H_\alpha(A, B) = H_\alpha\left(\frac{A \circ B}{\mathrm{tr}(A \circ B)}\right), \quad (19)$$

where $A \circ B$ denotes the Hadamard product between A and B. Given Eqs. (18) and (19), one can define the matrix-based Rényi's α-order MI, in analogy to Shannon MI, as:

$$I_\alpha(A; B) = H_\alpha(A) + H_\alpha(B) - H_\alpha(A, B). \quad (20)$$

and the matrix-based Rényi's α-order conditional entropy,

$$H_\alpha(A|B) = H_\alpha(A, B) - H_\alpha(B). \quad (21)$$

[1] $\kappa : \chi \times \chi \mapsto \mathbb{R}$ is a real valued positive definite kernel that is also infinitely divisible [Bhatia, 2006]. Usually, a Gaussian kernel is chosen [Sanchez Giraldo *et al.*, 2014; Yu *et al.*, 2019].

2.4 Other Popular Divergence and Dependence Measures

Apart from the above-mentioned divergences and their matrix-based counterparts, Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012] and Wasserstein distance or optimal transport (OT) are perhaps the most two popular classes of distances in DNNs. MMD quantifies distances between distributions as the distance between mean embeddings of features. Specifically, given distributions f and g over $\mathcal{X}$, the MMD is defined by a feature map $\varphi : \mathcal{X} \mapsto \mathcal{H}$, where $\mathcal{H}$ is also called a reproducing kernel Hilbert space (RKHS). More formally, the MMD is defined as:

$$\mathrm{MMD}(f, g; \mathcal{H}) = \|\mathbb{E}_{X\sim f}\left[\varphi(X)\right] - \mathbb{E}_{Y\sim g}\left[\varphi(Y)\right]\|_{\mathcal{H}} \quad (22)$$

An unbiased U-statistic estimator for MMD on observations $S_X = \{\mathbf{x}_i\}_{i=1}^n$ and $S_Y = \{\mathbf{y}_i\}_{i=1}^m$ is given by:

$$\widehat{\mathrm{MMD}}_u^2(S_X, S_Y; \kappa) = \frac{1}{n(n-1)}\sum_{i\neq j} H_{ij}, \quad (23)$$

where $H_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) + \kappa(\mathbf{y}_i, \mathbf{y}_j) - \kappa(\mathbf{x}_i, \mathbf{y}_j) - \kappa(\mathbf{y}_i, \mathbf{x}_j)$.

On the other hand, let $\Pi(f, g)$ denotes the set of joint distributions π whose marginal distributions are f and g, the Wasserstein distance is defined as:

$$W_p(f, g) = \left(\inf_{\pi\in\Pi(f,g)}\int \|\mathbf{x} - \mathbf{y}\|^p d\Pi(\mathbf{x}, \mathbf{y})\right)^{1/p} \quad (24)$$

In recent years, alternative dependence measures, based on MMD and OT between probability joint and product of marginal distributions, have been proposed. Examples include the Hilbert Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2007] based MMD and the Wasserstein's dependence measure (WDM) [Ozair *et al.*, 2019]. Empirical results show that WDM overcomes to some extent the issue of measuring dependencies when MI is large. Nevertheless, the estimator of WDM requires optimizing the parameters of a Lipchitz-continuous neural network, a condition that remains challenging to be enforced in practice. A very recent development, is the concept of usable information under computational constraints [Xu *et al.*, 2020]. In this work, it is assumed that limiting the expressiveness of the quantity describing information might be beneficial not only in terms of estimation but also in the context of learning. A characteristic feature of these measures of mutual information and mutual information-like quantities, including the matrix-based entropy and mutual information, is the incorporation of some inductive bias in the estimation process, which can be advantageous for machine learning [Tschannen *et al.*, 2020].

Another popular variant of MI, also called the squared-loss MI (SMI) [Suzuki *et al.*, 2009], defines the dependence of two random variables X and Y as the Pearson's χ^2 divergence from $f(x, y)$ to $f(x)f(y)$:

$$\mathrm{SMI}(X; Y) = \iint_{\mathcal{X}\times\mathcal{Y}} f(x, y)\left(\frac{f(x, y)}{f(x)f(y)} - 1\right)^2 dxdy. \quad (25)$$

SMI is usually estimated by least-square approaches. A recent proposal is SMI with OT [Liu *et al.*, 2019].

3 Information-Theoretic Principles and Regularizations in DNNs

3.1 Information-Theoretic Loss Functions for Robust Deep Learning

Information-theoretic measures can be directly used as a loss function to train DNNs. The most notable example is the cross-entropy loss for classification. Another example is the aforementioned CS divergence, which has been used, as a loss function, in both deep clustering [Kampffmeyer *et al.*, 2019] and classification [Janocha and Czarnecki, 2017]. In terms of clustering, one could model each cluster by its PDF and optimize cluster assignments such that the divergence between their PDFs is maximized. On the other hand, the f-divergence (a generalization of KL divergence) can be used to train deep energy-based models [Yu *et al.*, 2020a].

However, recent research has demonstrated that DNNs trained with cross-entropy or mean squared error (MSE) can easily fit (corrupted or even randomly) labeled data [Zhang *et al.*, 2017; Jiang *et al.*, 2018; Feng *et al.*, 2020], but with very poor generalization capacity on held out data, indicating that loss functions like cross-entropy or MSE are not a reliable indicator of generalization and robustness.

To improve the learning performance in non-Gaussian noises and outliers, a variety of (information-theoretic) non-MSE criteria have been proposed in the literature. Particularly in recent years, the maximum correntropy criterion (MCC) have found many successful applications in domains of machine learning and signal processing, which is very useful for the case where the signals are contaminated by heavy-tailed impulsive noises [Chen *et al.*, 2019]. Under the MCC, an optimal model can be obtained by maximizing the correntropy between the desired output (e.g., class label) y and the estimated output $\hat{y}$ [Liu *et al.*, 2007]:

$$f^*_{MCC} = \arg\max_{f\in\mathcal{F}} V_\sigma(\hat{y}, y) = \arg\max_{f\in\mathcal{F}} \mathbb{E}(G_\sigma(e)), \quad (26)$$

where f^* is the optimal model, $\mathcal{F}$ stands for models hypothesis space, $e = y - \hat{y}$ is the prediction residual, $\mathbb{E}$ refers to mathematical expectation, and $V_\sigma(\hat{y}, y) = \mathbb{E}(G_\sigma(e))$ denotes the correntropy between $\hat{y}$ and y [Liu *et al.*, 2007], with $G_\sigma(e)$ being the Gaussian kernel function with width σ.

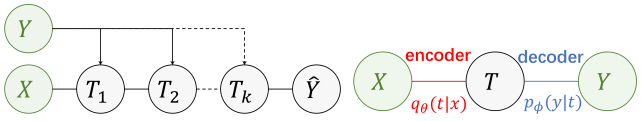
Compared to MSE (i.e., $\arg\min \mathbb{E}(e^2)$), MCC is robust to outliers because the correntropy criterion is closely related to M-estimators [Mandanas and Kotropoulos, 2016]. A popular alternative to MCC is the minimum error entropy (MEE) criterion $\arg\min H(e)$, in which $H(e)$ refers to the entropy of prediction residual e [Erdogmus and Principe, 2002]. The robustness of MEE over MSE is elaborated in [Chen *et al.*, 2016] and the references herein. As a replacement of MSE, both MCC and MEE have demonstrated improved robustness to non-Gaussian noises and outliers, when they are been deployed as a loss function to train DNNs [Qi *et al.*, 2014; Yu *et al.*, 2021a]. A most recent proposal along this line of research is the Determinant based Mutual Information (DMI) [Xu *et al.*, 2019], with a loss of $-\log[\text{DMI}(\hat{y}, y)]$.

Information-Theoretic measures also demonstrate great potential towards robust deep learning under domain shift [Quiñonero-Candela *et al.*, 2009], i.e., the joint distribution of samples $(\mathbf{x}, y)$ in the training (or source) data $P_{\text{source}}(\mathbf{x}, y)$ is different to that in the test (or target) data $P_{\text{target}}(\mathbf{x}, y)$. Taking the covariate shift (i.e., $P_{\text{target}}(y|\mathbf{x}) = P_{\text{source}}(y|\mathbf{x})$, but $P_{\text{target}}(\mathbf{x}) \neq P_{\text{source}}(\mathbf{x})$) as an example, [Greenfeld and Shalit, 2020] suggests that a DNN is robust to covariate shift iff the distribution of the prediction residuals $y - f(\mathbf{x})$ is statistically independent of the distribution of the input $\mathbf{x}$. The authors use HSIC to measure the degree of independence. [Yu *et al.*, 2021a] replaces HSIC with the matrix-based mutual information (i.e., Eq. (20)) and obtained favorable performance improvement. Note that, such independence criterion (for training regression models) was initially developed to identify causal direction among two variables. Feasible independence measures include HSIC [Mooij *et al.*, 2009] and least-square mutual information [Yamada *et al.*, 2014]. Interestingly, the independence criterion is also closely related to the aforementioned MEE [Yu *et al.*, 2021a], which provides new insight on its robustness.

3.2 InfoMax Principle and its Practical Usage

The information maximization (InfoMax) is well studied in statistics as exemplified by Burg's power spectrum algorithm [Burg, 1974]. For machine learning, the InfoMax principle dates back to [Linsker, 1988], which pointed out that a linear or nonlinear network can be treated as an information channel. A principle to self-organize such networks is to transfer as much information as possible of given data through the network. The InfoMax principle has recently been used in many computer vision and natural language processing studies on self-supervised representation learning, with the objective of maximizing the MI $I(X;T)$ between input X and latent representation T [Oord *et al.*, 2018; Hjelm *et al.*, 2018; Kong *et al.*, 2020]. Since MI is computational intractable in high-dimensional space, most of existing studies turn to optimize a lower bound of MI. However, recent studies show that InfoMax principle may introduce excessive and noisy information, which could be adversarial. The IB principle, as will be introduced later, can mitigate this issue [Wang *et al.*, 2021; Mahabadi *et al.*, 2021].

Alongside the representation learning, the InfoMax principle also offers an appealing strategy to train DNNs in a layer-wise manner without backpropagation. Traditionally, to learn a feedforward DNN in a supervised setting, one needs to train all components (or modules) of the network simultaneously using backpropagation (BP) since there is no explicit target for each hidden layer. By contrast, the InfoMax principle can help us to implement layer-wise training of DNNs. But to train a DNN, one needs to incorporate information about the desired output Y as well. Linsker's idea can be extended with the MI formalism to include not only the input information but also other sources of information which are usually characterized by desired outputs. From this perspective, different layers can be trained greedily [Hu and Principe, 2021]: the first hidden layer can be trained such that the MI $I(Y;T_1)$ between the output of this layer (denote T_1) and the desired output Y is maximized; then the second hidden layer can be trained in the same way by maximizing $I(Y;T_2)$. This procedure is repeated until the last hidden layer (see Fig. 1(a)). Interested readers can refer to [Duan *et al.*, 2020] for a comprehensive discussion on this topic.



(a) Information channel in InfoMax (b) Markov chain in deep IB

Figure 1: Given input X and desired output Y, denote T their latent representation. InfoMax can be implemented explicitly via $\max I(X;T)$ or implicitly with $\max I(Y;T(X))$. By contrast, deep IB approaches search T via $\max I(Y;T) - \beta I(X;T)$.

3.3 Information Bottleneck (IB) in Deep Learning

The IB principle was proposed in [Tishby *et al.*, 1999] as an information-theoretic framework for learning. It considers extracting information about a target signal Y through a correlated observable X. The extracted information is quantified by a variable T, which is (a possibly randomized) function of X, thus forming the Markov chain $Y \leftrightarrow X \leftrightarrow T$. Suppose we know the joint distribution $p(X,Y)$, the objective is to learn a representation T that maximizes its predictive power to Y subject to some constraints on the amount of information that it carries about X:

$$\mathcal{L}_{IB} = I(Y;T) - \beta I(X;T), \tag{27}$$

where $I(\cdot;\cdot)$ denotes the mutual information. β is a Lagrange multiplier that controls the trade-off between the **sufficiency** (the performance on the task, as quantified by $I(Y;T)$) and the **minimality** (the complexity of the representation, as measured by $I(X;T)$). In this sense, the IB principle also provides a natural approximation of *minimal sufficient statistic*.

The IB principle has both practical and theoretical impacts to DNNs. Practically, it can be formulated as a learning objective (or loss function) for deep models. When parameterizing IB with a DNN, X denotes input variable, Y denotes the desired output (e.g., class labels), T refers to the latent representation of one hidden layer. Usually, this was done by optimizing the IB Lagrangian (i.e., Eq. (27)) via a classic cross-entropy loss (which amounts to $\max I(Y;T)$ [Achille and Soatto, 2018; Amjad and Geiger, 2019]) regularized by a differentiable mutual information term $I(X;T)$. Depends on implementation details, $I(X;T)$ can be measured by variational approximation [Alemi *et al.*, 2017; Kolchinsky *et al.*, 2019], MINE [Elad *et al.*, 2019] and the matrix-based entropy functional [Yu *et al.*, 2021b]. According to [Kolchinsky *et al.*, 2019], the IB curve in classification scenario is piecewise linear and becomes a flat line at $I(Y;T) = H(Y)$ for $I(X;T) \geq H(Y)$. We obtain both theoretical and empirical IB curve by training a three layer MLP with 256 units in the bottleneck layer on MNIST dataset, as shown in Fig. 2(a). Empirically, the IB objective was observed to improve model generation performance and robustness to adversarial attack.

Theoretically, it was argued that, even though the IB objective is not explicitly optimized, DNNs trained with cross-entropy loss and stochastic gradient descent (SGD) inherently solve the IB compression-prediction trade-off [Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017]. The authors also posed the information plane (IP), i.e., the trajectory in $\mathbb{R}^2$ of the mutual information pair $\{I(X;T), I(Y;T)\}$ across

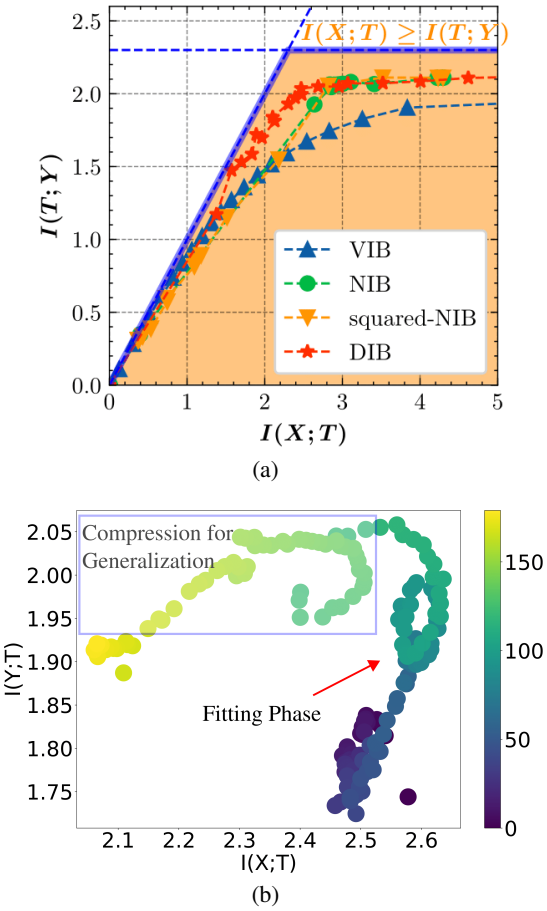


Figure 2: (a) Theoretical (the dashed light grey line) and empirical IB curve (of 4 different deep IB models) found by maximizing the IB Lagrangian with different values of β; different β leads to different trade-offs between compression and prediction; (b) A representative information plane for one DNN layer, different colors denote different training epochs. Two phases are visible: a short-term fitting phase in which both $I(X;T)$ and $I(Y;T)$ increase, a subsequent long-term compression phase manifested by decrease of $I(X;T)$.

training epochs, as a lens to analyze dynamics of learning of DNNs (see Fig. 2(b)). According to [Shwartz-Ziv and Tishby, 2017], there are two training phases in the common SGD optimization: an early "fitting" phase, in which both $I(X;T)$ and $I(T;Y)$ increase rapidly, and a later "compression" phase, in which there is a reversal such that $I(X;T)$ continually decreases. This work attracted significant attention, culminating in many follow-up works that tested the proclaimed narrative and its accompanying empirical observations. So far, the "fitting-and-compression" phenomena of the layered representation T have been observed in other types of DNNs, including the multilayer perceptrons (e.g., [Chelombiev *et al.*, 2019; Shwartz-Ziv and Tishby, 2017]), the AEs (e.g., [Yu and Principe, 2019]), and the CNNs (e.g., [Noshad *et al.*, 2019; Yu *et al.*, 2020c]). However, the IB theory is still a controversial issue, and different MI estimators may lead to different behaviors of curves in IP. We have to remember that not all the properties of the statistical definition of mutual informa-

tion are transferred to the estimators [Yu and Principe, 2019]. We recommend interested readers to [Goldfeld and Polyanskiy, 2020; Zaidi *et al.*, 2020] for comprehensive surveys on IB approach for deep learning.

3.4 Other Principles and Applications

There are other popular information-theoretic principles that have demonstrated great potential in their respective application fields. For example, in reinforcement learning, the maximum entropy regularization encourages exploration and avoids getting stuck in a local optima [Haarnoja *et al.*, 2017; Ahmed *et al.*, 2019]; the Stratonovich's value of information criterion [Stratonovich, 1965] can help agent to strike a balance between exploration and exploitation [Sledge and Príncipe, 2017]. Here, we additionally introduce two other learning principles that are relevant to InfoMax and IB.

CorEx and Representation Learning

The Cor-relation Ex-planation (CorEx) [Steeg and Galstyan, 2014] is an information-theoretic principle for learning representations that are maximally informative about the data. Specifically, let $\mathbf{x} = (x_1; x_2; \cdots; x_d)$ be a d-dimensional random variable with PDF $p(\mathbf{x})$, a measure of total dependence amongst each dimension of $\mathbf{x}$ is defined as (which is also known as total correlation [Watanabe, 1960]):

$$TC(\mathbf{x}) = \sum_{i=1}^{d} H(x_i) - H(\mathbf{x}) \tag{28}$$

Let $\mathbf{z}$ be the latent variable we want to infer from $\mathbf{x}$. The total dependence of $\mathbf{x}$, after conditioning on $\mathbf{z}$, becomes $TC(\mathbf{x}|\mathbf{z}) = \sum_{i=1}^{d} H(x_i|\mathbf{z}) - H(\mathbf{x}|\mathbf{z})$. A measure of informativeness of $\mathbf{z}$ about the dependence among the observed variables $\mathbf{x}$ can then be quantified as how total correlation is reduced after conditioning on $\mathbf{z}$, i.e.,

$$TC(\mathbf{x};\mathbf{z}) \equiv TC(\mathbf{x}) - TC(\mathbf{x}|\mathbf{z}) = \sum_{i=1}^{d} I(x_i;\mathbf{z}) - I(\mathbf{x};\mathbf{z}) \tag{29}$$

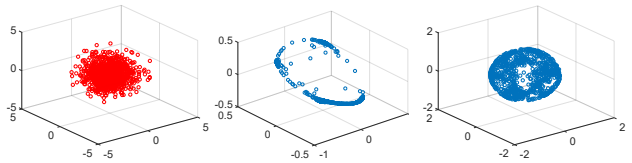
$TC(\mathbf{x};\mathbf{z})$ corresponds to the amount of dependence (in $\mathbf{x}$) that is explained by $\mathbf{z}$. Obviously, $TC(\mathbf{x};\mathbf{z})$ is maximized (or $TC(\mathbf{x}|\mathbf{z})$ is zero) iff the conditional distribution $p(\mathbf{x}|\mathbf{z})$ factorizes, in which case we can interpret $\mathbf{z}$ as capturing the information about common causes across all x_i. The CorEx principle is thus formulated as [Steeg and Galstyan, 2014]:

$$\max_{p(z|x)} TC(\mathbf{x};\mathbf{z}). \tag{30}$$

Recently, [Gao *et al.*, 2019] constructed a variational lower bound of CorEx and optimized the bound with DNNs. Interestingly, the resulting objective (under mild assumptions) is the same to the evidence lower bound (ELBO) in variational autoencoder (VAE) [Kingma and Welling, 2014].

Principle of Relevant Information

The fixed point update underlying the IB can de extended to a single random variable X, as demonstrated in the Principle of Relevant Information (PRI) [Principe, 2010, Chapter 8]. PRI is an unsupervised principle that aims to perform mode decomposition of a random variable X with a known and fixed probability distribution g. Suppose we aim to obtain (from X) a reduced statistical representation characterized by another random variable Z with probability distribution f.



(a) 3d Gaussian (b) $\beta = 2$ (c) $\beta = 5$

Figure 3: Illustration of the structures revealed by the PRI for (a) 3-dimensional isotropic Gaussian. As β increases, the solution identifies (b) principal curves and (c) principal surfaces.

The PRI casts this problem as a trade-off between the entropy $H(f)$ of Z and its descriptive power about X in terms of their divergence $D(f\|g)$:

$$J(f) = \arg\min_{f} H(f) + \beta D(f\|g), \tag{31}$$

where β is a hyper-parameter controlling the amount of relevant information that Z can extract from X. The minimization of entropy can be viewed as a means of reducing uncertainty (or redundancy) and finding the statistical regularities in the outcomes, whereas the minimization of information divergence ensures that such regularities are closely related to X. The PRI is similar in spirit to the IB approach, but the formulation is different because PRI does not require a relevant auxiliary variable Y and the optimization is done directly on the random variable X (rather than the joint distribution $p(X, Y)$ as in the IB).

The choice of entropy and divergence estimators is application-specific and depends on the simplicity of optimization. An interesting scenario arises when we use the Rényi's 2-order entropy (i.e., Eq. (10)) and the CS divergence (i.e., Eq. (11)), in which Eq. (31) has an elegant expression and provides multiscale representations of X controlled by increasing β, yielding clustering, principal curves (Fig. 3(b)) or surfaces (Fig. 3(c)), vector quantization and X itself in the limiting case ($\beta \to \infty$). Recent applications of PRI include the extraction of spectral-spatial features for hyperspectral image classification [Wei *et al.*, 2019] and the undersampling for imbalanced data classification [Hoyos-Osorio *et al.*, 2021]. The authors feel that PRI provides a solid mathematical foundation for data reduction using information theory.

4 Applications and Emerging Opportunities of Information Theory in DNNs

Information-theoretic measures (such as f-divergence, MMD, mutual information and HSIC) and principles (such as minimizing variational representations of f-divergence and maximizing an "evidence lower bound") have been extensively investigated to the design and the understanding of mainstream generative models [Goodfellow *et al.*, 2014; Kingma and Welling, 2014], with fruitful applications in both computer vision and natural language processing. Here, we additionally introduce two other emerging opportunities.

4.1 Understanding Deep Neural Networks

Information Theory is always a promising way to analyze and understand DNN behaviors in either training phase or decision process. Apart from the aforementioned IB theory that is still under debate (see Section 3.3), information-theoretic methods can also be used to understand the generalization capacity and construct generalization bound for DNNs. Examples in this category include [Xu and Raginsky, 2017; Bu *et al.*, 2020; Steinke and Zakynthinou, 2020]. On the other hand, information-theoretic measures can also quantify the importance (or redundancy) of individual neurons [Liu *et al.*, 2018] or filters [Yu *et al.*, 2020c], which in turn is beneficial to problems like pruning and generalization.

4.2 Learning with Multiple Tasks

A closely related topic to generalization is the joint learning of multiple tasks. Relevant problems include the *multi-task learning* (MTL) that aims to learn all tasks simultaneously and more proficiently than learning them independently; the *meta-learning* that aims to learn efficient learning model that can learn new tasks quickly; and the *continual learning* (CL), where the problem requires learning a sequence of tasks, avoiding negative transfer or catastrophic forgetting.

For all these three problems, information-theoretic tools have demonstrated great potential to strengthen theoretical discoveries and practical performance. For example, in CL, the elastic weight consolidation (EWC) [Kirkpatrick *et al.*, 2017] introduces penalty constructed by Fisher information to force important parameters of the network to remain close to the parameters of the network trained for the previous tasks. In MTL, [Yu *et al.*, 2020b] uses the Bregman-Correntropy (conditional) divergence to quantify the closeness of two tasks and penalize large model discrepancy for related tasks. In meta-learning, [Yin *et al.*, 2019] proposes an information-theoretic meta-regularizer to mitigate the memorization problem. In general, information theory is beneficial to construct generalization bound for these problems (e.g., the $\mathcal{H}\Delta\mathcal{H}$ divergence for both classification [Ben-David *et al.*, 2010] and regression [Cortes and Mohri, 2014]). These bounds can then be used to design different training or optimization algorithms. On the other hand, information theory also offers a new insight to analyze the fundamental trade-offs of existing algorithms [Vera *et al.*, 2018].

5 Topics of Future Interests & Conclusions

The intersection between information theory and deep neural networks (DNNs) a *challenging* and *promising* area due to its rapidly increasing prevalence in real-world applications. It is *challenging*, because most of expressions are defined on probability density functions (PDFs) that are hard to measure for real data. It is also *promising*, due to the remarkable performance gain and rigorous mathematical foundation. In this paper, we provided a survey of common information-theoretic estimators, learning principles and regularizations in DNNs, and recent developments of these estimators or principles in practical deep learning applications.

While there is now a significant body of work, there are still several open problems in applying information theory in DNNs. Some of these open problems - certainly not an exhaustive list - include the following.

- Information-theoretic measures beyond global PDF: The RKHS is a functional space that is very appropriate for statistical inference [Parzen, 1962], statistical embedding [Sriperumbudur *et al.*, 2010], and statistical modeling. Our current work is extending these methodologies to model local regions of the space of samples, i.e. working beyond the PDF for higher specificity and sensitivity. We have recently applied ideas from quantum mechanics, the famous Schroedinger equation based on the Laplacian of the wave function (here the estimator of the PDF in RKHS) to estimate model and data uncertaintity. Preliminary results show that estimators of uncertainty derived from this approach supplant conventional techniques [Singh and Principe, 2020]. This line of research can lead to better methods for transfer learning, and even causality.
- Information theory beyond $i.i.d.$ data: Most of existing information-theoretic estimators are limited to $i.i.d.$ data, a property that many problems do not meet (e.g., blind source separation on audio data and change detection on stream data). Therefore, it is crucial to extend information-theoretic measures to deal with structured and interdependent observations, including graphs [Han *et al.*, 2015; Wu *et al.*, 2020].
- Fruitful and unorthodox AI applications: It is not hard to foresee more successful applications of information-theoretic concepts on emerging AI topics. Taking the Explainable AI (XAI) as an example, information theory can help identify informative features to explain given example [Chen *et al.*, 2018] or generate interpretable representation to explain a black-box model [O'Shaughnessy *et al.*, 2020; Bang *et al.*, 2019]. Meanwhile, there is no doubt that information theory will contribute causal inference and offering fresh insights to its relevant modern deep learning topics include generalization and robustness [Schölkopf *et al.*, 2021].

References

[Achille and Soatto, 2018] A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE TPAMI*, 40(12):2897–2905, 2018.

[Ahmed *et al.*, 2019] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans. Understanding the impact of entropy on policy optimization. In *ICML*, pages 151–160. PMLR, 2019.

[Alemi *et al.*, 2017] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

[Amjad and Geiger, 2019] R. Amjad and B. Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE TPAMI*, 42(9):2225–2239, 2019.

[Arimoto, 1977] S. Arimoto. Information measures and capacity of order α for discrete memoryless channels. *Topics in information theory*, 1977.

[Bang *et al.*, 2019] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing. Explaining a black-box using deep variational information bottleneck approach. In *AAAI*, 2019.

[Belghazi *et al.*, 2018] M. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, et al. Mutual information neural estimation. In *ICML*, pages 531–540, 2018.

[Ben-David *et al.*, 2010] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[Bhatia, 2006] Rajendra Bhatia. Infinitely divisible matrices. *The American Mathematical Monthly*, 113(3):221–235, 2006.

[Bu *et al.*, 2020] Y. Bu, S. Zou, and V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE JSAIT*, 1(1):121–130, 2020.

[Burg, 1974] John Parker Burg. Maximum entropy spectral analysis. *Astronomy and Astrophysics Supplement*, 15:383, 1974.

[Chelombiev *et al.*, 2019] I. Chelombiev, C. Houghton, and C. O'Donnell. Adaptive estimators show information compression in deep neural networks. In *ICLR*, 2019.

[Chen *et al.*, 2016] B. Chen, L. Xing, B. Xu, H. Zhao, and J. C. Principe. Insights into the robustness of minimum error entropy estimation. *IEEE TNNLS*, 29(3):731–737, 2016.

[Chen *et al.*, 2018] J. Chen, L. Song, M. Wainwright, and M. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *ICML*, pages 883–892, 2018.

[Chen *et al.*, 2019] B. Chen, X. Wang, Y. Li, and J. C Principe. Maximum correntropy criterion with variable center. *IEEE SPL*, 26(8):1212–1216, 2019.

[Cortes and Mohri, 2014] C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[Duan *et al.*, 2020] S. Duan, S. Yu, Y. Chen, and J. C. Principe. On kernel method–based connectionist models and supervised deep learning without backpropagation. *Neural computation*, 32(1):97–135, 2020.

[Elad *et al.*, 2019] A. Elad, D. Haviv, Y. Blau, and T. Michaeli. Direct validation of the information bottleneck principle for deep nets. In *ICCV Workshops*, 2019.

[Erdogmus and Principe, 2002] D. Erdogmus and J. C. Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE TSP*, 50(7):1780–1786, 2002.

[Feng *et al.*, 2020] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, and B. An. Can cross entropy loss be robust to label noise. In *IJCAI*, pages 2206–2212, 2020.

[Gao *et al.*, 2018] W. Gao, S. Oh, and P. Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE TIT*, 64(8):5629–5661, 2018.

[Gao *et al.*, 2019] S. Gao, R. Brekelmans, G. Ver Steeg, and A. Galstyan. Auto-encoding total correlation explanation. In *AISTATS*, pages 1157–1166, 2019.

[Goldfeld and Polyanskiy, 2020] Z. Goldfeld and Y. Polyanskiy. The information bottleneck problem and its applications in machine learning. *IEEE JSAIT*, 1(1):19–38, 2020.

[Goodfellow *et al.*, 2014] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[Greenfeld and Shalit, 2020] D. Greenfeld and U. Shalit. Robust learning with the hilbert-schmidt independence criterion. In *ICML*, pages 3759–3768, 2020.

[Gretton *et al.*, 2007] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *NeurIPS*, volume 20, pages 585–592, 2007.

[Gretton *et al.*, 2012] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, 2012.

[Haarnoja *et al.*, 2017] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *ICML*, pages 1352–1361. PMLR, 2017.

[Han *et al.*, 2015] Lin Han, Richard C Wilson, and Edwin R Hancock. Generative graph prototypes from information theory. *IEEE TPAMI*, 37(10):2013–2027, 2015.

[Hjelm *et al.*, 2018] R. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2018.

[Hoyos-Osorio *et al.*, 2021] J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma, A. Orozco-Gutierrez, and G. Castellanos-Dominguez. Relevant information undersampling to support imbalanced data classification. *Neurocomputing*, 436:136–146, 2021.

[Hu and Principe, 2021] B. Hu and J. C. Principe. Training a bank of wiener models with a novel quadratic mutual information cost function. In *IEEE ICASSP*, 2021.

[Janocha and Czarnecki, 2017] K. Janocha and W. Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

[Jenssen *et al.*, 2006] R. Jenssen, J. C. Principe, D. Erdogmus, and T. Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

[Jiang *et al.*, 2018] Y. Jiang, D. Krishnan, H. Mobahi, and S. Bengio. Predicting the generalization gap in deep networks with margin distributions. In *ICLR*, 2018.

[Kampffmeyer *et al.*, 2019] M. Kampffmeyer, S. Løkse, F. Bianchi, L. Livi, A. Salberg, and R. Jenssen. Deep

divergence-based approach to clustering. *Neural Networks*, 113:91–101, 2019.

[Kapur, 1994] Jagat Narain Kapur. *Measures of information and their applications*. Wiley-Interscience, 1994.

[Kingma and Welling, 2014] D. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Kirkpatrick *et al.*, 2017] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

[Kolchinsky *et al.*, 2019] A. Kolchinsky, B. Tracey, and D. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

[Kong *et al.*, 2020] L. Kong, C. d'Autume, W. Ling, L. Yu, Z. Dai, and D. Yogatama. A mutual information maximization perspective of language representation learning. In *ICLR*, 2020.

[Kraskov *et al.*, 2004] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[Leonenko *et al.*, 2008] N. Leonenko, L. Pronzato, V. Savani, et al. A class of rényi information estimators for multidimensional densities. *Annals of statistics*, 36(5):2153–2182, 2008.

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

[Liu *et al.*, 2007] W. Liu, P. Pokharel, and J. C. Principe. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE TSP*, 55(11):5286–5298, 2007.

[Liu *et al.*, 2018] K. Liu, R. Amjad, and B. Geiger. Understanding individual neuron importance using information theory. *arXiv preprint arXiv:1804.06679*, 2018.

[Liu *et al.*, 2019] Y. Liu, M. Yamada, Y. Tsai, T. Le, R. Salakhutdinov, et al. Lsmi-sinkhorn: Semi-supervised squared-loss mutual information estimation with optimal transport. *arXiv preprint arXiv:1909.02373*, 2019.

[Lutwak *et al.*, 2005] E. Lutwak, D. Yang, and G. Zhang. Cramér-rao and moment-entropy inequalities for renyi entropy and generalized fisher information. *IEEE TIT*, 51(2):473–478, 2005.

[Mahabadi *et al.*, 2021] R. K. Mahabadi, Yonatan B., and James H. Variational information bottleneck for effective low-resource fine-tuning. In *ICLR*, 2021.

[Mandanas and Kotropoulos, 2016] F. Mandanas and C. Kotropoulos. Robust multidimensional scaling using a maximum correntropy criterion. *IEEE TSP*, 65(4):919–932, 2016.

[McAllester and Stratos, 2020] D. McAllester and K. Stratos. Formal limitations on the measurement of mutual information. In *AISTATS*, pages 875–884, 2020.

[Mooij *et al.*, 2009] J. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *ICML*, pages 745–752, 2009.

[Moskvina and Zhigljavsky, 2003] V. Moskvina and A. Zhigljavsky. An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics-Simulation and Computation*, 32(2):319–352, 2003.

[Nielsen and Chuang, 2010] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.

[Noshad *et al.*, 2019] M. Noshad, Y. Zeng, and A. Hero. Scalable mutual information estimation using dependence graphs. In *IEEE ICASSP*, pages 2962–2966, 2019.

[Oord *et al.*, 2018] A. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[O'Shaughnessy *et al.*, 2020] M. O'Shaughnessy, G. Canal, M. Connor, M. Davenport, and C. Rozell. Generative causal explanations of black-box classifiers. In *NeurIPS*, 2020.

[Ozair *et al.*, 2019] S. Ozair, C. Lynch, Y. Bengio, A. Oord, S. Levine, and P. Sermanet. Wasserstein dependency measure for representation learning. In *NeurIPS*, 2019.

[Pál *et al.*, 2010] D. Pál, B. Póczos, and C. Szepesvári. Estimation of renyi entropy and mutual information based on generalized nearest-neighbor graphs. In *NeurIPS*, 2010.

[Pardo, 2018] Leandro Pardo. *Statistical inference based on divergence measures*. CRC press, 2018.

[Parzen, 1962] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[Póczos and Schneider, 2011] B. Póczos and J. Schneider. On the estimation of alpha-divergences. In *AISTATS*, pages 609–617, 2011.

[Principe, 2010] J. C. Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[Qi *et al.*, 2014] Y. Qi, Y. Wang, X. Zheng, and Z. Wu. Robust feature learning by stacked autoencoder with maximum correntropy criterion. In *IEEE ICASSP*, pages 6716–6720, 2014.

[Quiñonero-Candela *et al.*, 2009] J. Quiñonero-Candela, M. Sugiyama, N. Lawrence, and A. Schwaighofer. *Dataset shift in machine learning*. The MIT Press, 2009.

[Rényi, 1961] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

[Sanchez Giraldo *et al.*, 2014] L. Sanchez Giraldo, M. Rao, and J. C. Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE TIT*, 61(1):535–548, 2014.

[Schölkopf *et al.*, 2021] B. Schölkopf, F. Locatello, S. Bauer, N. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[Shwartz-Ziv and Tishby, 2017] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[Singh and Póczos, 2014] S. Singh and B. Póczos. Generalized exponential concentration inequality for rényi divergence estimation. In *ICML*, pages 333–341, 2014.

[Singh and Principe, 2020] R. Singh and J. Principe. Time series analysis using a kernel based multi-modal uncertainty decomposition framework. In *UAI*, pages 1368–1377, 2020.

[Sledge and Príncipe, 2017] I. Sledge and J. C. Príncipe. Balancing exploration and exploitation in reinforcement learning using a value of information criterion. In *ICASSP*, pages 2816–2820. IEEE, 2017.

[Sriperumbudur *et al.*, 2010] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

[Steeg and Galstyan, 2014] G. Steeg and A. Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *NeurIPS*, pages 577–585, 2014.

[Steinke and Zakynthinou, 2020] T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. In *COLT*, pages 3437–3452, 2020.

[Stratonovich, 1965] RL Stratonovich. On value of information. *Izvestiya of USSR Academy of Sciences, Technical Cybernetics*, 5:3–12, 1965.

[Suzuki *et al.*, 2009] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC bioinformatics*, 10(1):1–12, 2009.

[Tishby and Zaslavsky, 2015] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE ITW*, pages 1–5, 2015.

[Tishby *et al.*, 1999] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Allerton*, pages 368–377, 1999.

[Tschannen *et al.*, 2020] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *ICLR*, 2020.

[Tsuda *et al.*, 2005] K. Tsuda, G. Rätsch, and M. Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *JMLR*, 6(Jun):995–1018, 2005.

[Vera *et al.*, 2018] M. Vera, L. Vega, and P. Piantanida. Compression-based regularization with an application to multitask learning. *IEEE JSTSP*, 12(5):1063–1076, 2018.

[Wang *et al.*, 2021] B. Wang, S. Wang, Y. Cheng, Z. Gan, R. Jia, B. Li, and J. Liu. Infobert: Improving robustness of language models from an information theoretic perspective. In *ICLR*, 2021.

[Watanabe, 1960] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[Wei *et al.*, 2019] Yantao Wei, Shujian Yu, Luis Sanchez Giraldo, and Jose C Principe. Multiscale principle of relevant information for hyperspectral image classification. *arXiv preprint arXiv:1907.06022*, 2019.

[Wu *et al.*, 2020] T. Wu, H. Ren, P. Li, and J. Leskovec. Graph information bottleneck. In *NeurIPS*, 2020.

[Xu and Raginsky, 2017] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *NeurIPS*, pages 2521–2530, 2017.

[Xu *et al.*, 2019] Y. Xu, P. Cao, Y. Kong, and Y. Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pages 6222–6233, 2019.

[Xu *et al.*, 2020] Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A theory of usable information under computational constraints. In *ICLR*, 2020.

[Yamada *et al.*, 2014] M. Yamada, M. Sugiyama, and J. Sese. Least-squares independence regression for non-linear causal inference under non-gaussian noise. *Machine learning*, 96(3):249–267, 2014.

[Yin *et al.*, 2019] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn. Meta-learning without memorization. In *ICLR*, 2019.

[Yu and Principe, 2019] S. Yu and J. C. Principe. Understanding autoencoders with information theoretic concepts. *Neural Networks*, 117:104–123, 2019.

[Yu *et al.*, 2019] S. Yu, L. Sanchez Giraldo, R. Jenssen, and J. C. Principe. Multivariate extension of matrix-based renyi's α-order entropy functional. *IEEE TPAMI*, 2019.

[Yu *et al.*, 2020a] L. Yu, Y. Song, J. Song, and S. Ermon. Training deep energy-based models with f-divergence minimization. In *ICML*, pages 10957–10967, 2020.

[Yu *et al.*, 2020b] S. Yu, A. Shaker, F. Alesiani, and J. C. Principe. Measuring the discrepancy between conditional distributions: Methods, properties and applications. In *IJCAI*, pages 2777–2784, 2020.

[Yu *et al.*, 2020c] S. Yu, K. Wickstrøm, R. Jenssen, and J. C. Príncipe. Understanding convolutional neural networks with information theory: An initial exploration. *IEEE TNNLS*, 2020.

[Yu *et al.*, 2021a] S. Yu, F. Alesiani, X. Yu, R. Jenssen, and J. C. Principe. Measuring dependence with matrix-based entropy functional. In *AAAI*, 2021.

[Yu *et al.*, 2021b] X. Yu, S. Yu, and J. C. Principe. Deep deterministic information bottleneck with matrix-based entropy functional. In *IEEE ICASSP*, 2021.

[Zaidi *et al.*, 2020] A. Zaidi, I. Estella-Aguerri, et al. On the information bottleneck problems: Models, connections, applications and information theoretic views. *Entropy*, 22(2):151, 2020.

[Zhang *et al.*, 2017] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.