Disentangled Information Bottleneck

Ziqi Pan, Li Niu,* Jianfu Zhang, Liqing Zhang*
MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{panziqi_ai, ustcnewly, c.sis}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract

The information bottleneck (IB) method is a technique for extracting information that is relevant for predicting the target random variable from the source random variable, which is typically implemented by optimizing the IB Lagrangian that balances the compression and prediction terms. However, the IB Lagrangian is hard to optimize, and multiple trials for tuning values of Lagrangian multiplier are required. Moreover, we show that the prediction performance strictly decreases as the compression gets stronger during optimizing the IB Lagrangian. In this paper, we implement the IB method from the perspective of supervised disentangling. Specifically, we introduce *Disentangled Information Bottleneck* (DisenIB) that is consistent on compressing source maximally without target prediction performance loss (maximum compression). Theoretical and experimental results demonstrate that our method is consistent on maximum compression, and performs well in terms of generalization, robustness to adversarial attack, out-of-distribution detection, and supervised disentangling.

1 Introduction

Compression is a ubiquitous task in machine learning (Cover and Thomas 2012; MacKay and Mac Kay 2003). For example, over-parameterized deep networks are compressed with pruning for the sake of computational efficiency (Han, Mao, and Dally 2015; Dai et al. 2018), machines may transform complex data into compressed representations that generalizes well (Alemi et al. 2017; Shwartz-Ziv and Tishby 2017). It is important to determine which aspects of data should be preserved and which should be discarded. The *Information Bottleneck* (Tishby, Pereira, and Bialek 2000; Dimitrov and Miller 2001; Samengo 2002) (IB) method provides a principled approach to this problem, which compresses the source random variable to keep the information relevant for predicting the target random variable while discarding all irrelevant information. The IB method has been applied to various domains such as classification (Hecht, Noor, and Tishby 2009), clustering (Slonim et al. 2005), coding theory (Hassanpour, Wübben, and Dekorsy 2018; Zeitler et al. 2008), and quantization (Strouse and Schwab 2017; Cheng et al. 2019). Recent research also demonstrate that the IB method can produce well-generalized representations (Shamir, Sabato, and Tishby 2010; Vera, Piantanida, and Vega 2018; Amjad and Geiger 2019) and may be promising on explaining the learning behaviors of neural networks (Tishby and Zaslavsky 2015; Shwartz-Ziv and Tishby 2017; Saxe et al. 2019).

Given random variables X, Y with their joint distribution $p_{\text{data}}(X, Y)$, the IB method aims to compress X to a "bottleneck" random variable T keeping the information relevant for predicting Y. Namely, seeking a probabilistic mapping $q(T|X)$ such that the Mutual Information (MI) $I(X;T)$ is constrained while the information $I(T;Y)$ is maximized, which can be formally stated in terms of the constrained optimization problem

$$\underset{T\in\Delta}{\operatorname{argmax}} I(T;Y), \quad \text{s.t. } I(X;T) \leqslant r, \tag{1}$$

where a level of compression (*i.e.*, $I(X;T) \leqslant r$) is provided as the constraint, and Δ is the set of random variables T that obey the Markov chain $Y \leftrightarrow X \leftrightarrow T$ (Witsenhausen and Wyner 1975; Ahlswede and Korner 1975; Gilad-Bachrach, Navot, and Tishby 2003). Optimal solutions to Eq. (1) for $r \in [0, +\infty)$ form the *IB curve* (Tishby, Pereira, and Bialek 2000; Rodríguez Gálvez, Thobaben, and Skoglund 2020), which specifies the trade-off between the compression (*i.e.*, $I(X;T)$) and prediction (*i.e.*, $I(T;Y)$) terms. In practice, to avoid the non-linear constraint, Eq. (1) can be optimized by minimizing the so-called *IB Lagrangian* (Tishby, Pereira, and Bialek 2000; Gilad-Bachrach, Navot, and Tishby 2003; Shamir, Sabato, and Tishby 2010; Rodríguez Gálvez, Thobaben, and Skoglund 2020)

$$\mathcal{L}_{\text{IB}}[q(T|X);\beta] = -I(T;Y) + \beta I(X;T), \tag{2}$$

where β is the Lagrange multiplier which controls the trade-off and typically set in $[0, 1]$ (Rodríguez Gálvez, Thobaben, and Skoglund 2020).

Minimizing the IB Lagrangian encounters the following two problems: 1) It is hard to obtain the desired compression level r (Eq. (1)). In practice, a compression level is expected to obtain via minimizing the IB Lagrangian with certain β. However, recent works (Kolchinsky, Tracey, and Kuyk 2019; Rodríguez Gálvez, Thobaben, and Skoglund 2020) pointed out that β and the compression level are not causally related, and therefore multiple optimizations for different β value are required to achieve a specific compression level. 2) $I(T;Y)$ given by the optimal solution to the IB Lagrangian is strictly

*Corresponding author

decreasing as β increases, *i.e.*, the prediction performance is unavoidably reduced by the compression. We provide theoretical justification for this statement (see Theorem 1).

It is expected to extract the minimal sufficient (Friedman, Hastie, and Tibshirani 2001) part about Y from X into T, *i.e.*, compressing X maximally without reducing $I(T;Y)$, which is referred to as *maximum compression* in the remainder of this paper. However, such a case cannot be achieved through minimizing the IB Lagrangian, since compression always decreases $I(T;Y)$. Moreover, it is expected to eliminate the need for multiple optimizations and explore a consistent method for maximum compression with a single optimization. We start by realizing that *supervised disentangling* (Ridgeway 2016) is closely related to the idea behind the IB method. Supervised disentangling tackles the problem of identifying complementary data aspects and separating them from each other with supervision. Similarly, in the IB method, one must separate Y-relevant and Y-irrelevant data aspects. This inspires us to implement the IB method from the perspective of supervised disentangling, leading to our proposed *Disentangled Information Bottleneck* (DisenIB). To the best of our knowledge, we are the first to draw the connection between the IB method and supervised disentangling. Our contribution are threefold:

- We study the trade-off in the IB Lagrangian, showing that balancing compression and prediction terms can only decrease prediction performance, therefore the maximum compression can not be achieved.
- We propose a variant of IB, the *Disentangled Information Bottleneck* (DisenIB), which is proven to be consistent on maximum compression. Specifically, DisenIB eliminates the need for multiple optimizations and consistently performs maximum compression with a single optimization.
- Through experimental results, we justify our theoretical statements and show that DisenIB performs well in terms of generalization (Shamir, Sabato, and Tishby 2010), robustness to adversarial attack (Alemi et al. 2017) and out-of-distribution data detection (Alemi, Fischer, and Dillon 2018), and supervised disentangling.

The remainder of this paper is organized as follows. First, we analyze the trade-off in optimizing the IB Lagrangian in Section 2.1, showing that the prediction performance strictly decreases as the compression gets stronger. To overcome the trade-off problem, we firstly propose a formal definition on compressing source data maximally without prediction performance loss (maximum compression) in Section 2.2, followed by introducing our proposed DisenIB that is consistent on maximum compression in Section 2.3. All our experimental analyses are provided in Section 4.

2 Methodology

In this section, we first study the trade-off involved in the IB Lagrangian, showing that balancing compression and prediction terms can only decrease prediction performance and thus cannot achieve maximum compression. Then, we introduce our proposed DisenIB that is consistent on maximum compression.

2.1 The IB Lagrangian Trade-off

We first show that optimizing the IB Lagrangian leads to inevitable trade-off. Specifically, the optimal solutions to the compression and prediction objectives obtained by optimizing the IB Lagrangian are consistently inferior to that obtained by optimizing each objective independently. This can be formally stated by the following Theorem 1 (see supplementary for proof):

Theorem 1 *Consider the derivable IB Lagrangian,*

$$\mathcal{L}_{\mathrm{IB}}\left[q\left(T|X\right);\beta\right] = -I\left(T;Y\right) + \beta I\left(X;T\right), \tag{3}$$

to be minimized over $q(T|X)$ with $\beta \geqslant 0$. Let q^*_β optimize $\mathcal{L}_{\mathrm{IB}}\left[q\left(T|X\right);\beta\right]$. Assume that $I_{q^*_\beta}(X;T) \neq 0$,

$$\frac{\partial I_{q^*_\beta}\left(T;Y\right)}{\partial \beta} < 0 \text{ and } \frac{\partial I_{q^*_\beta}\left(X;T\right)}{\partial \beta} < 0. \tag{4}$$

We can learn that for every nontrivial solution q^*_β such that $I_{q^*_\beta}(X;T) \neq 0$, $I(T;Y)$ strictly decreases as β increases, and compression (*i.e.*, $\beta I(X;T)$) can only decrease prediction performance (*i.e.*, $I(T;Y)$), which is not expected.

2.2 Consistency Property

Optimizing the IB Lagrangian can not achieve maximum compression due to the aforementioned trade-off. It is expected to explore a method that is capable of performing maximum compression. Moreover, we also expect to eliminate the need for multiple optimizations. Namely, we expect to explore a method that consistently performs maximum compression with a single optimization, which is referred to as the *consistency* property on maximum compression.

We first specify r (Eq. (1)) that provides the maximum compression case. We analyze the case where Y is a deterministic function of X (Rodríguez Gálvez, Thobaben, and Skoglund 2020), which covers a wide range of application scenarios such as classification and regression. In such a case, $I(X;Y) = H(Y)$. However, our method is also applicable in general cases where $I(X;Y) < H(Y)$. According to basic properties of MI (Cover and Thomas 2012) (*i.e.*, $I(T;Y) \leqslant H(Y)$ when Y is discrete-valued), $I(T;Y) = H(Y)$ leads to the case without prediction loss. By leveraging the *data processing inequality* (DPI) (Cover and Thomas 2012) (*i.e.*, $I(T;Y) \leqslant I(X;T)$) and basic properties of MI (Cover and Thomas 2012) (*i.e.*, $I(X;T) \leqslant H(X)$ when X is discrete-valued), in the case without prediction loss, we have that

$$H\left(Y\right) = I\left(T;Y\right) \leqslant I\left(X;T\right) \leqslant H\left(X\right). \tag{5}$$

Hence $r = H(Y)$ provides the maximum compression, in which case

$$I\left(X;T\right) = I\left(T;Y\right) = H\left(Y\right). \tag{6}$$

In practice, we aim to design a cost function $\mathcal{L}$, such that the maximum compression case (Eq. (6)) is expected to be obtained via minimizing $\mathcal{L}$. Specifically, we expect that minimized $\mathcal{L}$ consistently satisfies Eq. (6). Hence the formal definition of *consistency* on maximum compression is given as

Definition 1 (Consistency) *The lower-bounded cost functional $\mathcal{L}$ is consistent on maximum compression, if*

$$\begin{aligned}&\forall\epsilon>0,\exists\delta>0,\quad \mathcal{L}-\mathcal{L}^{*}<\delta\Longrightarrow\\&\quad|I\left(X;T\right)-H\left(Y\right)|+|I\left(T;Y\right)-H\left(Y\right)|<\epsilon,\end{aligned}\tag{7}$$

where $\mathcal{L}^{*}$ is the global minimum of $\mathcal{L}$.

Satisfying Eq. (6) involves precise information amount control, *i.e.*, exactly constraining both $I\left(X;T\right)$ and $I\left(T;Y\right)$ at $H\left(Y\right)$. Several works (Alemi et al. 2017; Chen et al. 2016; Kolchinsky, Tracey, and Wolpert 2019) involve estimating MI. However, they can only maximize or minimize MI, but still struggle to constrain the MI at an exact value. For the examples (Alemi et al. 2017; Kolchinsky, Tracey, and Wolpert 2019) using variational bounds, the estimation only becomes accurate as the bound becomes tight. There also exist MI estimators like MINE (Belghazi et al. 2018) that can provide accurate estimations, but certain sample complexity (Belghazi et al. 2018) is required. Therefore, they exhibit high variances (Song and Ermon 2020), and it is non-trivial to achieve Eq. (6).

2.3 Disentangled IB

We introduce our proposed DisenIB that is consistent on maximum compression. After realizing the relation between IB and supervised disentangling, we implement IB from the perspective of supervised disentangling by introducing another random variable S as the complementary aspect to T and further encoding information relevant (*resp.*, irrelevant) to Y into T (*resp.*, S). Formally, the objective functional to be minimized is stated as

$$\begin{aligned}&\mathcal{L}_{\mathrm{DisenIB}}\left[q\left(S|X\right),q\left(T|X\right)\right]\\&\quad=-I\left(T;Y\right)-I\left(X;S,Y\right)+I\left(S;T\right).\end{aligned}\tag{8}$$

Specifically, we encourage (S,Y) to represent the overall information of X by maximizing $I\left(X;S,Y\right)$, so that S at least covers the information of Y-irrelevant data aspect. We encourage that Y can be accurately decoded from T by maximizing $I\left(T;Y\right)$, so that T at least covers the information of Y-relevant data aspect. Hence, the amount of information stored in S and T are both lower bounded. In such a case, forcing S to be disentangled from T by minimizing $I\left(S;T\right)$ eliminates the overlapping information between them and thus tightens both bounds, leaving the exact information relevant (*resp.*, irrelevant) to Y in T (*resp.*, S).

Moreover, maximum compression can be consistently achieved via optimizing $\mathcal{L}_{\mathrm{DisenIB}}$, as stated in the following Theorem 2 (see supplementary for proof):

Theorem 2 *$\mathcal{L}_{\mathrm{DisenIB}}$ is consistent on maximum compression.*

We now introduce how to implement DisenIB in principle, leaving practical implementation in supplementary due to space limitation. Same as prior works (Alemi et al. 2017; Chalk, Marre, and Tkacik 2016; Achille and Soatto 2018; Kolchinsky, Tracey, and Wolpert 2019), we derive variational approximations to $I\left(T;Y\right)$ and $I\left(X;S,Y\right)$ terms. By introducing variational probabilistic mappings $p\left(y|t\right)$ and $r\left(x|s,y\right)$, the tractable variational lower bounds can be formulated as

$$\begin{aligned}I\left(T;Y\right)&=\mathbb{E}_{q(y,t)}\log q\left(y|t\right)-\mathbb{E}_{q(y)}\log q\left(y\right)\\&\geqslant\mathbb{E}_{q(y,t)}\log p\left(y|t\right)+H\left(Y\right),\end{aligned}\tag{9}$$

$$\begin{aligned}I\left(X;S,Y\right)&=\mathbb{E}_{q(x,s,y)}\log q\left(x|s,y\right)-\mathbb{E}_{q(x)}\log q\left(x\right)\\&\geqslant\mathbb{E}_{q(x,s,y)}\log r\left(x|s,y\right)+H\left(X\right),\end{aligned}\tag{10}$$

where the inequalities follow from

$$\begin{aligned}&\mathbb{E}_{q(y|t)}\log q\left(y|t\right)-\mathbb{E}_{q(y|t)}\log p\left(y|t\right)\\&\quad=D_{\mathrm{KL}}\left[q\left(Y|t\right)\parallel p\left(Y|t\right)\right]\geqslant 0,\end{aligned}\tag{11}$$

$$\begin{aligned}&\mathbb{E}_{q(x|s,y)}\log q\left(x|s,y\right)-\mathbb{E}_{q(x|s,y)}\log r\left(x|s,y\right)\\&\quad=D_{\mathrm{KL}}\left[q\left(X|s,y\right)\parallel r\left(X|s,y\right)\right]\geqslant 0.\end{aligned}\tag{12}$$

The lower bounds become tight as $D_{\mathrm{KL}}\left[q\left(Y|t\right)\parallel p\left(Y|t\right)\right]$ and $D_{\mathrm{KL}}\left[q\left(X|s,y\right)\parallel r\left(X|s,y\right)\right]$ approximate 0. By rewriting (leveraging Markov chains $Y\leftrightarrow X\leftrightarrow T$ and $Y\leftrightarrow X\leftrightarrow S$)

$$q\left(y,t\right)=\mathbb{E}_{p_{\mathrm{data}}(x)}p_{\mathrm{data}}\left(y|x\right)q\left(t|x\right),\tag{13}$$

$$q\left(x,s,y\right)=p_{\mathrm{data}}\left(x\right)p_{\mathrm{data}}\left(y|x\right)q\left(s|x\right),\tag{14}$$

we see that lower bounds in Eq (9)-(10) only require samples from joint data distribution $p_{\mathrm{data}}\left(x,y\right)$, probabilistic mappings $q\left(t|x\right),q\left(s|x\right)$ and variational approximations $p\left(y|t\right),r\left(x|s,y\right)$, therefore is tractable.

Minimizing the $I\left(S;T\right)=D_{\mathrm{KL}}\left[q\left(S,T\right)\parallel q\left(S\right)q\left(T\right)\right]$ term is intractable since both $q\left(s,t\right)$ and $q\left(s\right)q\left(t\right)$ involve mixtures with a large number of components. However, we observe that we can sample from joint distribution $q\left(s,t\right)$ efficiently by first sampling x from dataset uniformly at random and then sampling from $q\left(s,t|x\right)=q\left(s|x\right)q\left(t|x\right)$ due to the Markov chain $S\leftrightarrow X\leftrightarrow T$ (Kim and Mnih 2018). We can also sample from product of marginal distributions $q\left(s\right)q\left(t\right)$ by shuffling the samples from the joint distribution $q\left(s,t\right)$ along the batch axis (Belghazi et al. 2018). Then, we use the *density-ratio-trick* (Nguyen, Wainwright, and Jordan 2008; Sugiyama, Suzuki, and Kanamori 2012; Kim and Mnih 2018) by involving a discriminator d which estimates the probability that its input is a sample from $q\left(s,t\right)$ rather than from $q\left(s\right)q\left(t\right)$. Adversarial training is involved to train the discriminator,

$$\min_{q}\max_{d}\mathbb{E}_{q(s)q(t)}\log d\left(s,t\right)+\mathbb{E}_{q(s,t)}\log\left(1-d\left(s,t\right)\right).\tag{15}$$

As shown by (Goodfellow et al. 2014), $q\left(s,t\right)=q\left(s\right)q\left(t\right)$ when the Nash equilibrium is achieved, thus minimizing the $I\left(S;T\right)$ term.

3 Relation to Prior Work

In this section, we relate our proposed method to prior works on existing variants of IB Lagrangian and supervised disentangling methods.

3.1 Variants of IB Lagrangian

Optimizing the IB Lagrangian (Eq. (2)) involves integrals which are typically intractable. For this reason, only limited

cases (Tishby, Pereira, and Bialek 2000; Chechik et al. 2005) have been mainly developed until recently.

Recently, to optimize the IB Lagrangian on continuous and possibly non-Gaussian variables using neural networks, several works (Alemi et al. 2017; Chalk, Marre, and Tkacik 2016; Achille and Soatto 2018; Kolchinsky, Tracey, and Wolpert 2019) are proposed to derive variational approximations to the IB Lagrangian, which permits parameterizing the IB model using neural networks with gradient descent training. Specifically, they employed the same variational bound for the prediction term $I(T;Y)$ as our Eq. (9). However, they differ from ours in how to perform compression. While we compress via disentangling to avoid balancing compression and prediction terms, they all use variational upper bounds on the compression term $I(X;T)$. For VIB (Alemi et al. 2017) and similar works (Chalk, Marre, and Tkacik 2016; Achille and Soatto 2018), $I(X;T)$ is bounded as

$$\begin{aligned} I(X;T) &= \mathbb{E}_{q(x,t)} \log q(t|x) - \mathbb{E}_{q(t)} \log q(t) \\ &\leqslant \mathbb{E}_{q(x,t)} \log q(t|x) - \mathbb{E}_{q(t)} \log v(t), \end{aligned} \tag{16}$$

where v is some prior distribution. Differing from these three methods in upper bounding $I(X;T)$, the *Nonlinear Information Bottleneck* (Kolchinsky, Tracey, and Wolpert 2019) (NIB) uses a non-parametric upper bound based on *Kernel Density Entropy Estimates* (Kolchinsky and Tracey 2017),

$$I(X;T) \leqslant -\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{N}\sum_{j=1}^{N} e^{-D_{\mathrm{KL}}[q(t|x_i)\|q(t|x_j)]}, \tag{17}$$

where N is the total number of samples in the given dataset.

Recent research (Kolchinsky, Tracey, and Kuyk 2019) showed that optimizing the IB Lagrangian for different values of β cannot explore the IB curve in the case where Y is a deterministic function of X (Rodriguez Galvez 2019). To tackle this problem, they propose the *squared-IB Lagrangian* by squaring $I(X;T)$:

$$\mathcal{L}_{\mathrm{squared-IB}} = -I(T;Y) + \beta \left(I(X;T)\right)^2. \tag{18}$$

Gálvez *et al.* (Rodríguez Gálvez, Thobaben, and Skoglund 2020) then took a further step to extend the squared-IB Lagrangian, showing that applying any monotonically increasing and strictly convex functions on $I(X;T)$ is able to explore the IB curve.

All these methods balance the compression and prediction terms. As we show in Theorem 1, IB methods involving such trade-off cannot achieve maximum compression. Differing from them, we alter compression to disentangling, which is shown to be able to avoid trade-off and consistently perform maximum compression.

3.2 Supervised Disentangling

Supervised disentangling (Ridgeway 2016; Fletcher and Kasturi 1988) tackles the problem of identifying complementary aspects of data and separating them from each other with (partial) aspects label, which is a fundamental idea and various methods (Mathieu et al. 2016; Hadad, Wolf, and Shahar 2018; Jaiswal et al. 2018; Moyer et al. 2018; Zheng and Sun 2019; Song et al. 2019; Gabbay and Hoshen 2020; Jaiswal et al. 2020) are proposed. Though the IB method is closely related to supervised disentangling methods in the sense of separating target-relevant and target-irrelevant aspects of data, exactly controlling the amount of information stored in the respective data aspects is beyond the ability of these disentangling methods. Specifically, none of the optimality is guaranteed in the sense of information control for disentangling methods, which is referred to as *information leak* in (Gabbay and Hoshen 2020). For example (Hadad, Wolf, and Shahar 2018), the redundant information irrelevant to prediction is compressed via limiting the expressive capacity of the neural network model, making it tricky to exactly control the amount of information. In (Jaiswal et al. 2018), different terms are designed to compete with others, therefore tuning hyper-parameters that balance different terms is required, and exactly controlling information amount can not be achieved either. However, as we can see from the consistency property of our proposed DisenIB, our method can exactly control the amount of information stored in respective data aspects.

4 Experiments

We compare existing variants of IB Lagrangian with our proposed DisenIB method. Following previous works (Alemi et al. 2017; Kolchinsky, Tracey, and Wolpert 2019; Kolchinsky, Tracey, and Kuyk 2019; Rodríguez Gálvez, Thobaben, and Skoglund 2020), we optimize existing methods for different values of $\beta \in [0, 1]$, producing a series of models that explore the trade-off between compression and prediction, while our method does not involve such trade-off. Following previous works (Alemi et al. 2017; Alemi, Fischer, and Dillon 2018; Kolchinsky, Tracey, and Wolpert 2019), we evaluate our method in terms of generalization (Shamir, Sabato, and Tishby 2010; Vera, Piantanida, and Vega 2018), robustness to adversarial attack (Alemi et al. 2017), and out-of-distribution data detection (Alemi, Fischer, and Dillon 2018) on benchmark datasets: MNIST (LeCun et al. 1998), FashionMNIST (Xiao, Rasul, and Vollgraf 2017), and CIFAR10 (Krizhevsky, Hinton et al. 2009). We also provide results on more challenging natural image datasets: object-centric Tiny-ImageNet (Deng et al. 2009) and scene-centric SUN-RGBD (Song, Lichtenberg, and Xiao 2015). We also study the disentangling behavior of our method on MNIST (LeCun et al. 1998), Sprites (Reed et al. 2015) and dSprites (Matthey et al. 2017). Due to space limitation, the implementation details can be found in supplementary.

4.1 Behavior on *IB Plane*

In this section, we compare existing variants of IB with our method in terms of the behavior on *IB Plane*, which is defined by the axes $I(X;T)$ (the x-axis) and $I(T;Y)$ (the y-axis), showing that our method can perform maximum compression. We report results on both training set and test set. Since all methods use stochastic encoder that produces Gaussian bottleneck, we use Monte Carlo sampling (Goldfeld 2019) to get accurate estimate of the $I(X;T)$ term. To estimate the $I(T;Y) = H(Y) - H(Y|T)$ term, we

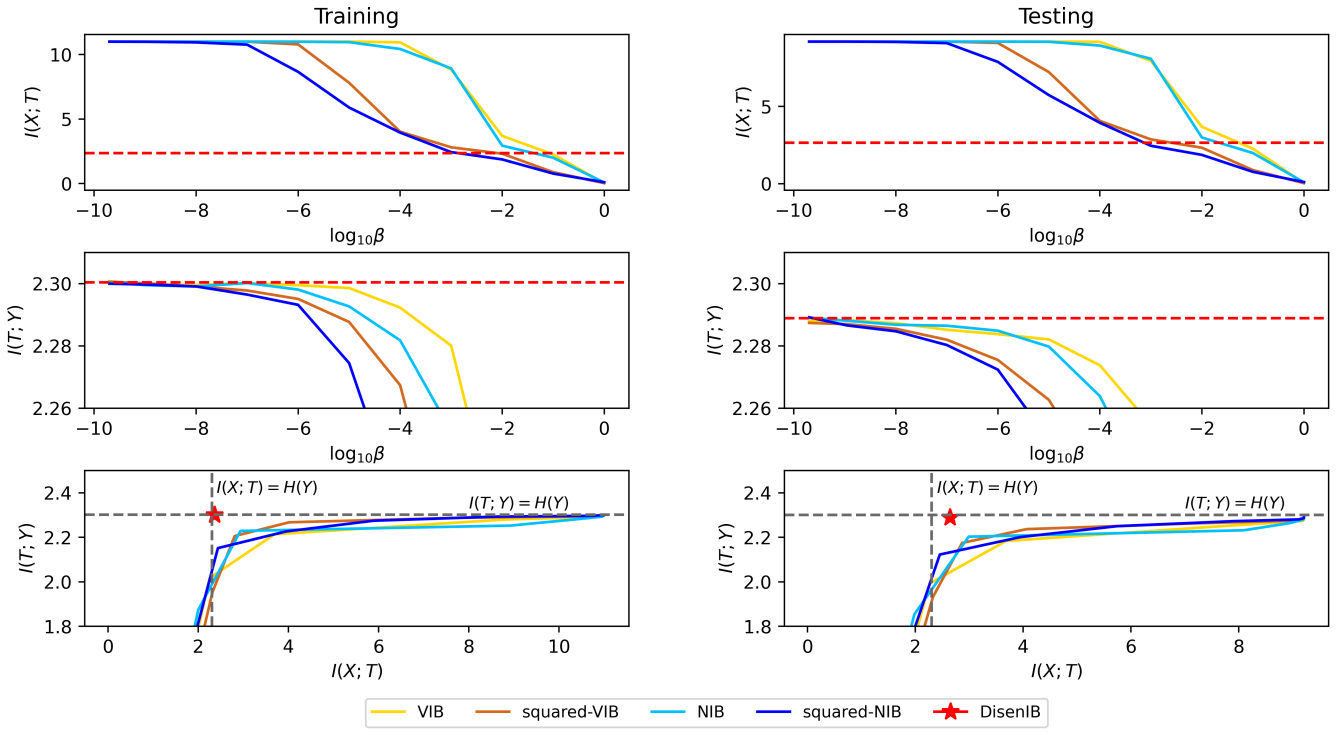


Figure 1: MNIST training (left) and testing (right) results. Our method only needs a single optimization to obtain $I(X;T)$ and $I(T;Y)$, so our method is represented by red doted line in the top and middle rows, and red star in the bottom row. Top row: $I(X;T)$ *vs.* β curve. Middle row: $I(T;Y)$ *vs.* β curve. Bottom row: IB-plane diagrams defined by $I(T;Y)$ *vs.* $I(X;T)$.

approximate the conditional entropy with the cross-entropy loss, which is employed in (Kolchinsky, Tracey, and Wolpert 2019). $H(Y)$ is a known constant specified by the dataset. Experimental results on MNIST are summarized in Figure 1, while the results on FashionMNIST, CIFAR10, Tiny-ImageNet, and SUN-RGBD can be found in supplementary.

From Figure 1, we learn that intensive β tuning is required for optimizing IB Lagrangian to obtain desired compression level. Specifically, as by (Kolchinsky, Tracey, and Kuyk 2019; Rodríguez Gálvez, Thobaben, and Skoglund 2020), β and the compression level are not causally related, and therefore multiple optimizations for different β value are required to reach a specific compression level. We can also observe that as $I(X;T)$ is compressed to $H(Y)$, $I(T;Y)$ is inferior to $H(Y)$, which makes maximum compression (*i.e.*, $I(X;T) = I(T;Y) = H(Y)$) impossible via minimizing the IB Lagrangian. Meanwhile our method can effectively compress X while maintaining $I(T;Y)$ at a high value. For clarity, we report numerical results at compression level $I(X;T) = H(Y)$ as in Table 1. From Table 1, we observe that $I(T;Y)$ obtained using our method is barely even reduced and much closer to $H(Y)$, while $I(T;Y)$ obtained using other methods are consistently inferior.

For the comparison in terms of generalization, robustness to adversarial attack and out-of-distribution data detection in Sections 4.2-4.3, since our results are at the compression level $I(X;T) = H(Y)$, we also report results of existing IB variants at the same compression level by tuning β.

Dataset	Training	Testing	β
$H(Y)$	2.30	2.30	
VIB	2.01	1.99	10^{-2}
squared-VIB	1.94	1.95	10^{-3}
NIB	1.95	1.97	10^{-2}
squared-NIB	1.98	1.99	10^{-3}
DisenIB	**2.25**	**2.17**	N/A

Table 1: $I(T;Y)$ and β value obtained by different methods at compression level $I(X;T) = H(Y)$ on MNIST training set and test set. $H(Y)$ is the amount of $I(T;Y)$ in the case of maximum compression (Eq. (6)).

4.2 Generalization and Adversarial Robustness

In this section, we compare our method with existing methods in terms of generalization performance and robustness to adversarial attack. We report the results on MNIST dataset, leaving those on FashionMNIST, CIFAR10, Tiny-ImageNet, and SUN-RGBD in supplementary.

We firstly introduce how to perform evaluation in terms of generalization and robustness to adversarial attack. Generalization performance in Table 2 is evaluated by the classification mean accuracy on MNIST test set after training the model on MNIST training set (Alemi et al. 2017; Kolchinsky, Tracey, and Wolpert 2019). Considering that deep neural networks can be easily "fooled" into making

		Training	Testing
		VIB / squared-VIB / NIB / squared-NIB / DisenIB	
Generalization		N/A	97.6 / 96.2 / 97.2 / 93.3 / **98.2**
Adversary Robustness	$\epsilon = 0.1$	74.1 / 42.1 / 75.2 / 61.3 / **94.3**	73.4 / 42.7 / 75.2 / 62.0 / **90.2**
	$\epsilon = 0.2$	19.1 / 8.7 / 21.8 / 24.1 / **81.5**	20.8 / 9.2 / 23.6 / 24.5 / **80.0**
	$\epsilon = 0.3$	3.5 / 5.9 / 3.2 / 9.3 / **68.4**	4.2 / 5.9 / 3.4 / 9.9 / **67.8**

Table 2: Generalization and adversarial robustness performance (%) on MNIST dataset.

Metric	SUN-RGBD (Song, Lichtenberg, and Xiao 2015)	Gaussian Noise
	VIB / squared-VIB / NIB / squared-NIB / DisenIB	
FPR (95% TPR) ↓	27.4 / 49.9 / 34.4 / 47.5 / **0.0**	4.5 / 12.7 / 13.4 / 5.3 / **0.0**
AUROC ↑	94.6 / 86.6 / 94.2 / 85.6 / **99.4**	98.8 / 95.5 / 97.4 / 90.8 / **99.7**
AUPR In ↑	94.8 / 83.5 / 95.2 / 83.3 / **99.6**	99.8 / 96.6 / 97.8 / 92.4 / **99.8**
AUPR Out ↑	93.7 / 83.2 / 91.8 / 83.1 / **98.9**	98.5 / 95.5 / 96.8 / 88.8 / **99.5**
Detection Error ↓	11.5 / 20.0 / 11.9 / 15.0 / **1.7**	4.7 / 4.7 / 7.6 / 15.7 / **1.0**

Table 3: Distinguishing in- and out-of-distribution test data for MNIST image classification (%). ↑ (*resp.*, ↓) indicates that larger (*resp.*, lower) value is better.

mistakes by changing their inputs by imperceptibly small amounts (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015), adversarial robustness exams how robust the model is to such adversarial examples. To exam adversarial robustness, we use the standard baseline attack method (Goodfellow, Shlens, and Szegedy 2015). Specifically, after training the model on MNIST training set, the training (*resp.*, test) adversary robustness in Table 2 are measured by the classification mean accuracy on the adversarial examples of the training (*resp.*, test) set, where the adversarial examples are generated by taking a single step in the gradient direction. We vary $\epsilon \in \{0.1, 0.2, 0.3\}$, which controls the magnitude of the perturbation at each pixel (Goodfellow, Shlens, and Szegedy 2015).

We see that our method can slightly outperform existing methods in terms of generalization performance. For adversary robustness, our method is significantly better than existing methods. Compared with our method, existing methods can be easily fooled by making perturbations, which is because a model with degenerated prediction performance (*i.e.*, $I(T;Y)$) will reduce its adversary robustness (Alemi et al. 2017). However, since our method avoids information reduction while compressing, it is more robust to adversarial examples than existing methods.

4.3 Out-of-distribution Detection

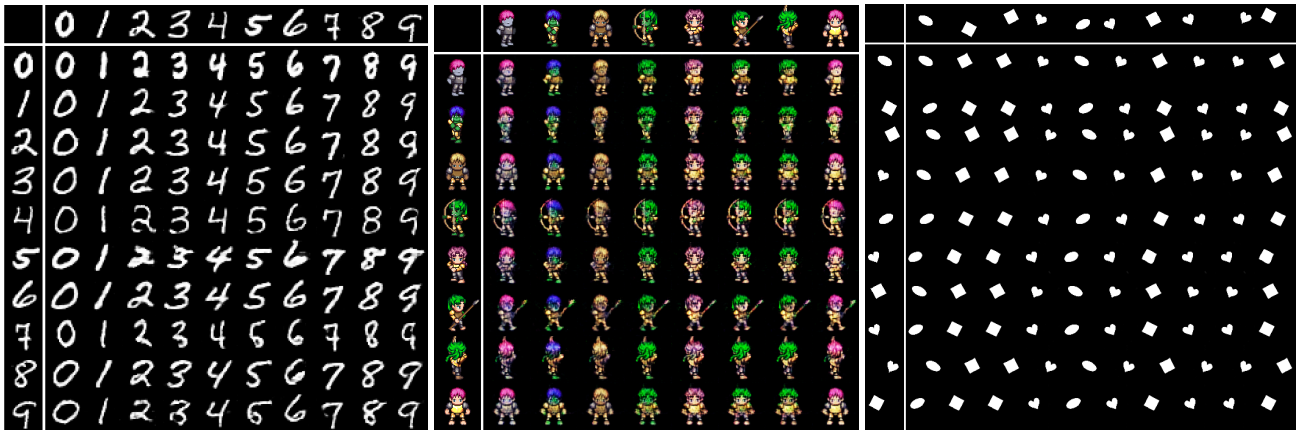
Modern neural networks are known to generalize well when the training and test data are sampled from the same distribution (Zhang et al. 2017). However, when deploying neural networks in real-world applications, there is often very little control over the test data distribution. It is important for classifiers to be aware of uncertainty when shown new types of inputs, *i.e.*, *out-of-distribution* examples. We report experimental results on MNIST dataset, leaving results on FashionMNIST, CIFAR10, Tiny-ImageNet, and SUN-RGBD in supplementary.

We make use of (Liang, Li, and Srikant 2018) for out-of-distribution data detection without re-training the model. To perform detection, the model is first trained on MNIST training set. At detection time, examples from MNIST test set can be viewed as in-distribution data, and we use examples from scene-centric natural image dataset SUN-RGBD (Song, Lichtenberg, and Xiao 2015) as out-distribution data, considering the large distribution gap between the two datasets. Following (Liang, Li, and Srikant 2018), we also use synthetic Gaussian noise dataset which consists of 10,000 random Gaussian noise samples from $\mathcal{N}(0, 1)$ as out-of-distribution dataset. A good model is expected to accurately detect whether the given example is an out-distribution data or not. In terms of metrics, we use *FPR at 95% TPR*, *Detection Error*, *AUROC*, *AUPR* following (Liang, Li, and Srikant 2018; Hendrycks and Gimpel 2017).

The results are summarized in Table 3. We achieve superior performance than all baselines. For SUN-RGBD (Song, Lichtenberg, and Xiao 2015) dataset, our method consistently outperforms the baselines by a large margin, showing that our method is sensitively aware of outliers when given new samples from considerably different data distributions. Since results of existing methods are obtained at the same compression level $I(X;T) = H(Y)$ as ours, the prediction performances (*i.e.*, $I(T;Y)$) are reduced due to the aforementioned trade-off. Because a reduced prediction performance will degenerate a model's capacity of detecting out-of-distribution data (Zhang et al. 2017), detection performances of existing methods are inferior to ours due to the prediction performance reduction.

4.4 Supervised Disentangling

We briefly study the disentangling behavior of our method by showing both qualitative and quantitative results. We visualize disentanglement by reconstructing images after *swapping* (Mathieu et al. 2016) data aspects (S or T) of



(a) MNIST (b) Sprites (c) dSprites

Figure 2: Visualization grids of image swapping generation. The top row and leftmost column images come from the dataset. The other images are generated using S from leftmost image and T from the image at the top of the column. The diagonal images show reconstructions.

Datasets	Classification Accuracy		Information Amount	
	S	T	$I(X;T)$	$I(X;S)$
	(Mathieu et al. 2016) / (Hadad, Wolf, and Shahar 2018) / DisenIB			
Training	10.0 / 10.1 / 10.0	99.2 / 99.7 / 99.6	7.92 / 9.86 / **2.32**	10.75 / 10.81 / 10.86
Testing	9.9 / 10.1 / 10.1	98.2 / 98.3 / 98.2	9.60 / 10.17 / **2.71**	9.12 / 9.46 / 9.15

Table 4: MNIST classification mean accuracy based on S or T (%) as well as $I(X;T)$ and $I(X;S)$.

different images, which is commonly used for qualitative evaluation in disentangling literature (Mathieu et al. 2016; Hadad, Wolf, and Shahar 2018; Higgins et al. 2017; Kim and Mnih 2018). In a good disentanglement, given S from one image I_1 and T from another image I_2, the image generated based on S and T should preserve the S-qualities of I_1 and the T-qualities of I_2. We visualize qualitative results via swapping data aspects on disentanglement benchmark datasets: MNIST, Sprites (Reed et al. 2015), and dSprites (Matthey et al. 2017), where Y represents the digit category, body color, and shape category, respectively. T is Y-relevant data aspect and S is Y-irrelevant data aspect (*e.g.*, thickness on MNIST, body type on Sprites, and orientation/positision on dSprites). From Figure 2, we learn that our method can separate data aspects S and T well, generating reasonable results when combining data aspects from two different images.

In terms of quantitative metrics, we perform *classification* (Mathieu et al. 2016; Hadad, Wolf, and Shahar 2018) based on both data aspects. To do so, we first train a classifier to predict Y labels based on S or T following (Mathieu et al. 2016; Hadad, Wolf, and Shahar 2018). For a good disentanglement, high accuracy (*resp.*, random results) should be obtained when applying the classifier on T (*resp.*, S). Moreover, to quantify the amount of information of X preserved in each data aspect, we estimate $I(X;T)$ and $I(X;S)$ terms. The quantitative evaluation results are summarized in Table 4. We see that the classification performance based on S or T are almost the same across all methods. However, the information controlling behaviors are quite different. While baseline methods leak too much information to the T aspect, our method can exactly control the amount of information stored in the respective aspects.

5 Conclusion

In this paper, to perform maximum compression without the need of multiple optimizations, we have implemented the IB method from the perspective of supervised disentangling, introducing the Disentangled Information Bottleneck (DisenIB). Theoretical and experimental results have demonstrated that our method is consistent on maximum compression, and performs well in terms of generalization, robustness to adversarial attack, out-of-distribution data detection, and supervised disentangling.

Acknowledgements

The work is supported by the National Key R&D Program of China (2018AAA0100704) and the Shanghai Science and Technology R&D Program of China (20511100300) and is partially sponsored by National Natural Science Foundation of China (Grant No.61902247) and Shanghai Sailing Program (19YF1424400). This work is also sponsored by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Achille, A.; and Soatto, S. 2018. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(12): 2897–2905.

Ahlswede, R.; and Korner, J. 1975. Source coding with side information and a converse for degraded broadcast channels. *IEEE Transactions on Information Theory* 21(6): 629–637.

Alemi, A. A.; Fischer, I.; and Dillon, J. V. 2018. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906* .

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *ICLR*.

Amjad, R. A.; and Geiger, B. C. 2019. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(9): 2225–2239.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *ICML*.

Chalk, M.; Marre, O.; and Tkacik, G. 2016. Relevant sparse codes with variational information bottleneck. In *NIPS*.

Chechik, G.; Globerson, A.; Tishby, N.; and Weiss, Y. 2005. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research* 6: 165–188.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*.

Cheng, D.; Tu, Y.; Ma, Z.-W.; Niu, Z.; and Zhang, L. 2019. Risk Assessment for Networked-guarantee Loans Using High-order Graph Attention Representation. In *IJCAI*.

Cover, T. M.; and Thomas, J. A. 2012. *Elements of information theory*. Wiley.

Dai, B.; Zhu, C.; Guo, B.; and Wipf, D. 2018. Compressing Neural Networks using the Variational Information Bottleneck. In *ICML*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dimitrov, A. G.; and Miller, J. P. 2001. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems* 12(4): 441–472.

Fletcher, L. A.; and Kasturi, R. 1988. A robust algorithm for text string separation from mixed text/graphics images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(6): 910–918.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The elements of statistical learning*. Springer New York Inc.

Gabbay, A.; and Hoshen, Y. 2020. Demystifying Inter-Class Disentanglement. In *ICLR*.

Gilad-Bachrach, R.; Navot, A.; and Tishby, N. 2003. An Information Theoretic Tradeoff between Complexity and Accuracy. In *COLT*.

Goldfeld, Z. 2019. Estimating Information Flow in Deep Neural Networks. In *ICML*.

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

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

Hadad, N.; Wolf, L.; and Shahar, M. 2018. A two-step disentanglement method. In *CVPR*.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* .

Hassanpour, S.; Wübben, D.; and Dekorsy, A. 2018. On the equivalence of double maxima and KL-means for information bottleneck-based source coding. In *WCNC*.

Hecht, R. M.; Noor, E.; and Tishby, N. 2009. Speaker recognition by Gaussian information bottleneck. In *ISCA*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

Jaiswal, A.; Moyer, D.; Ver Steeg, G.; AbdAlmageed, W.; and Natarajan, P. 2020. Invariant Representations through Adversarial Forgetting. In *AAAI*.

Jaiswal, A.; Wu, R. Y.; Abd-Almageed, W.; and Natarajan, P. 2018. Unsupervised adversarial invariance. In *NIPS*.

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In *ICML*.

Kolchinsky, A.; and Tracey, B. D. 2017. Estimating mixture entropy with pairwise distances. *Entropy* 19(7): 361.

Kolchinsky, A.; Tracey, B. D.; and Kuyk, S. V. 2019. Caveats for information bottleneck in deterministic scenarios. In *ICLR*.

Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2019. Nonlinear information bottleneck. *Entropy* 21(12): 1181.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical Report TR-2009, University of Toronto, Toronto* .

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR*.

MacKay, D. J.; and Mac Kay, D. J. 2003. *Information theory, inference and learning algorithms*. Cambridge University Press.

Mathieu, M. F.; Zhao, J. J.; Zhao, J.; Ramesh, A.; Sprechmann, P.; and LeCun, Y. 2016. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*.

Matthey, L.; Higgins, I.; Hassabis, D.; and Lerchner, A. 2017. dsprites: Disentanglement testing sprites dataset. *URL https://github. com/deepmind/dsprites-dataset/.[Accessed on: 2018-05-08]* .

Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Ver Steeg, G. 2018. Invariant representations without adversarial training. In *NIPS*.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2008. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS*.

Reed, S. E.; Zhang, Y.; Zhang, Y.; and Lee, H. 2015. Deep visual analogy-making. In *NIPS*.

Ridgeway, K. 2016. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299* .

Rodriguez Galvez, B. 2019. The Information Bottleneck: Connections to Other Problems, Learning and Exploration of the IB Curve. In *KTH Royal Institute of Technology*.

Rodríguez Gálvez, B.; Thobaben, R.; and Skoglund, M. 2020. The Convex Information Bottleneck Lagrangian. *Entropy* 22(1): 98.

Samengo, I. 2002. Information Loss in an Optimal Maximum Likelihood Decoding. *Neural Computation* 14(4): 771–779.

Saxe, A. M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B. D.; and Cox, D. D. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment* 2019(12): 124020.

Shamir, O.; Sabato, S.; and Tishby, N. 2010. Learning and generalization with the information bottleneck. *Theoretical Computer Science* 411(29-30): 2696–2711.

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

Slonim, N.; Atwal, G. S.; Tkačik, G.; and Bialek, W. 2005. Information-based clustering. *Proceedings of the National Academy of Sciences* 102(51): 18297–18302.

Song, J.; and Ermon, S. 2020. Understanding the Limitations of Variational Mutual Information Estimators. In *ICLR*.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2019. Learning controllable fair representations. In *AISTATS*.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*.

Strouse, D.; and Schwab, D. J. 2017. The deterministic information bottleneck. *Neural computation* 29(6): 1611–1630.

Sugiyama, M.; Suzuki, T.; and Kanamori, T. 2012. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* 64(5): 1009–1044.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057* .

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *ITW*.

Vera, M.; Piantanida, P.; and Vega, L. R. 2018. The role of the information bottleneck in representation learning. In *ISIT*.

Witsenhausen, H.; and Wyner, A. 1975. A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory* 21(5): 493–501.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* .

Zeitler, G.; Koetter, R.; Bauch, G.; and Widmer, J. 2008. Design of network coding functions in multihop relay networks. In *ISTC*.

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

Zheng, Z.; and Sun, L. 2019. Disentangling latent space for vae by label relevant/irrelevant dimensions. In *CVPR*.