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

$$\arg\max_{T \in \Delta} I(T; Y), \quad \text{s.t. } I(X; T) \leq r,$$  

(1)

where a level of compression (i.e., $I(X; T) \leq r$) is provided as the constraint, and $\Delta$ 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; Rodriguez 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; Rodriguez Gálvez, Thobaben, and Skoglund 2020)

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

(2)

where $\beta$ is the Lagrange multiplier which controls the trade-off and typically set in $[0, 1]$ (Rodriguez 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 $\beta$. However, recent works (Kolchinsky, Tracey, and Kuyk 2019; Rodriguez Gálvez, Thobaben, and Skoglund 2020) pointed out that $\beta$ and the compression level are not causally related, and therefore multiple optimizations for different $\beta$ value are required to achieve a specific compression level. 2) $I(T; Y)$ given by the optimal solution to the IB Lagrangian is strictly
decreasing as $\beta$ 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,

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

(3)

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

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

We can learn that for every nontrivial solution $q^*_\beta$ such that $I_{q^*_\beta}(X;T) \neq 0$, $I(T;Y)$ strictly decreases as $\beta$ 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., $I(T;Y) \leq H(Y)$ 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., $I(T;Y) \leq I(X;T)$) and basic properties of MI (Cover and Thomas 2012) (i.e., $I(X;T) \leq H(X)$ when $X$ is discrete-valued), in the case without prediction loss, we have that

$$H(Y) = I(T;Y) \leq I(X;T) \leq H(X).$$ (5)

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

$$I(X;T) = I(T;Y) = H(Y).$$ (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
Theorem 2 (see supplementary for proof): 
Satisfying Eq. (6) involves precise information amount control, i.e., exactly constraining both \( I(X;T) \) and \( I(T;Y) \) at \( H(Y) \). 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

\[
\mathcal{L}_{\text{DisenIB}}[q(S|X), q(T|X)] = -I(T;Y) - I(X;S,Y) + I(S;T). \tag{8}
\]

Specifically, we encourage \((S,Y)\) to represent the overall information of \( X \) by maximizing \( I(X;S,Y) \), 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 \( I(T;Y) \), 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 \( I(S;T) \) 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 \( \mathcal{L}_{\text{DisenIB}} \), as stated in the following Theorem 2 (see supplementary for proof):

Theorem 2 \( \mathcal{L}_{\text{DisenIB}} \) 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 Tkcik 2016; Achille and Soatto 2018; Kolchinsky, Tracey, and Wolpert 2019), we derive variational approximations to \( I(T;Y) \) and \( I(X;S,Y) \) terms. By introducing variational probabilistic mappings \( p(y|t) \) and \( r(x|s,y) \), the tractable variational lower bounds can be formulated as

\[
I(T;Y) = E_{q(y|t)} \log q(y|t) - E_{q(y)} \log q(y) \geq E_{q(y|t)} \log p(y|t) + H(Y), \tag{9}
\]

\[
I(X;S,Y) = E_{q(x,s,y)} \log q(x|s,y) - E_{q(x)} \log q(x) \geq E_{q(x,s,y)} \log r(x|s,y) + H(X), \tag{10}
\]

where the inequalities follow from

\[
E_{q(y|t)} \log q(y|t) - E_{q(y)} \log p(y|t) = D_{KL}[q(Y|t) \| p(Y|t)] \geq 0, \tag{11}
\]

\[
E_{q(x,s,y)} \log q(x|s,y) - E_{q(x)} \log r(x|s,y) = D_{KL}[q(X|s,y) \| r(X|s,y)] \geq 0. \tag{12}
\]

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

\[
q(y,t) = \text{E}_{p_{\text{data}}(x)} p_{\text{data}}(y|x) q(t|x), \tag{13}
\]

\[
q(x,s,y) = p_{\text{data}}(x) p_{\text{data}}(y|x) q(s|x), \tag{14}
\]

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

Minimizing the \( I(S;T) = D_{KL}[q(S,T) \| q(S)q(T)] \) term is intractable since both \( q(s,t) \) and \( q(s)q(t) \) involve mixtures with a large number of components. However, we observe that we can sample from joint distribution \( q(s,t) \) efficiently by first sampling \( x \) from dataset uniformly at random and then sampling from \( q(s,t|x) = q(s|x)q(t|x) \) due to the Markov chain \( S \leftrightarrow X \leftrightarrow T \) (Kim and Mnih 2018). We can also sample from product of marginal distributions \( q(s)q(t) \) by shuffling the samples from the joint distribution \( q(s,t) \) 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 \( q(s,t) \) rather than from \( q(s)q(t) \). Adversarial training is involved to train the discriminator,

\[
\min_{q} \max_{d} \mathbb{E}_{q(s|q(t))} \log d(s,t) + \mathbb{E}_{q(s,t)} \log (1 - d(s,t)). \tag{15}
\]

As shown by (Goodfellow et al. 2014), \( q(s,t) = q(s)q(t) \) when the Nash equilibrium is achieved, thus minimizing the \( I(S;T) \) 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
Recent experiments have demonstrated that supervised disentangling methods can outperform their unsupervised counterparts in terms of maximum achievable performance. In this section, we compare existing variants of IB with our proposed DisenIB method. Following previous works (Alemi et al. 2017; Kolchinsky, Tracey, and Wolpert 2019; Kolchinsky, Tracey, and Kuyk 2019), 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; Rodriguez 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
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 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), and the compression level are not causally related, and therefore multiple optimizations for different 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 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 $\beta$.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training</th>
<th>Testing</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H(Y)$</td>
<td>2.30</td>
<td>2.30</td>
<td>N/A</td>
</tr>
<tr>
<td>VIB</td>
<td>2.01</td>
<td>1.99</td>
<td>$10^{-2}$</td>
</tr>
<tr>
<td>squared-VIB</td>
<td>1.94</td>
<td>1.95</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>NIB</td>
<td>1.95</td>
<td>1.97</td>
<td>$10^{-2}$</td>
</tr>
<tr>
<td>squared-NIB</td>
<td>1.98</td>
<td>1.99</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>DisenIB</td>
<td>2.25</td>
<td>2.17</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 1: $I(T; Y)$ and $\beta$ 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
inputs, control over the test data distribution. It is important for clas-
sification (Zhang et al. 2017). However, when deploying neural
networks in real-world applications, there is often very little

Modern neural networks are known to generalize well when
the training and test data are sampled from the same distri-
bution. Since results of existing methods are obtained at
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., $I(X; T)$) 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 per-
formance of existing methods is 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

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>SUN-RGBD (Song, Lichtenberg, and Xiao 2015)</th>
<th>Gaussian Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection Error</td>
<td>$11.5 / 11.5 / 11.5 / 11.5 / 11.5$</td>
<td>$4.7 / 4.7 / 4.7 / 4.7 / 4.7$</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Classification Accuracy</th>
<th>Information Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( S )</td>
<td>( T )</td>
</tr>
<tr>
<td>Training</td>
<td>10.0 / 10.1 / 10.0</td>
<td>99.2 / 99.7 / 99.6</td>
</tr>
</tbody>
</table>

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).


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*.


