A Tighter Bound on the Information Bottleneck with Applications to Deep Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Information Bottleneck (IB) provides a hypothetically optimal framework for data modeling, yet is often intractable. Recent efforts optimized supervised DNNs with a variational upper bound to the IB objective, resulting in improved robustness to adversarial attacks. However, when deriving the upper bound, the supervisor distribution $p^*(y)$ is assumed to be constant, where in practice it is optimized over. This work demonstrates that lifting this assumption not only results in a tighter bound on the IB and improved empirical performance, but also proposes a new motivation for conditional entropy regularization.

1 Introduction

Deep Neural Nets (DNNs) learn latent representations induced by their downstream task, objective function, and other parameters. The quality of the learned representations impacts the DNN's generalization ability and the coherence of the emerging latent space (Bengio 2009). A question emerges regarding the extraction of an optimal latent representation for all data points from a restricted set of training examples. Classic information theory provides rate-distortion (Shannon 1959) for optimal compression of data. However, rate-distortion regards all information as equal, not taking into account which information is more relevant to a specified downstream task, without constructing tailored distortion functions. The Information Bottleneck, IB, (Tishby, Pereira, and Bialek 1999) resolves this limitation by defining mutual information (MI) between the learned representation and a designated downstream task as a universal distortion function. Yet, learning representations using the IB method is possible given discrete distributions, and some continuous ones, but not in the general case (Chechik et al. 2003). Moreover, MI is either difficult or impossible to optimize over when considering deterministic models, such as MLPs (Saxe et al. 2018; Amjad and Geiger 2020). Nonetheless, the promise of the IB remains alluring, and recents works utilized VAE (Kingma and Welling 2014) inspired variational approximations to approximate upper bounds to the IB objective, allowing its utilization as a loss function for DNNs, where the underlying distributions are both continuous and unknown (Alemi et al. 2017; Fischer 2020; Cheng et al. 2020). These approaches learn representations in supervised settings, without knowledge of the underlying distribution $p^*(x, y)$, utilizing the learned variational conditional $p(y|x)$ to approximate MI. In contrast, non variational IB methods learn representations in unsupervised settings, where the stochastic process underlying the

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

observed data is known (Tishby, Pereira, and Bialek 1999; Chechik et al. 2003; Painsky and Tishby 2017). Nonetheless, when deriving the variational IB objectives, previous research (Alemi et al. 2017; Fischer 2020; Cheng et al. 2020) relax the problem by considering the learned representation as the only optimized parameter, when in practice the classifier is also optimized. We derive a new upper bound for the IB objective, and a subsequent variational approximation, by removing the relaxation. We show that our bound is tighter than previous bounds, and that our proposed loss function is a tighter variational approximation, when considering $p(y|x)$ as part of the optimization. We believe our new derivation is a better adaptation of the IB for supervised tasks, and show empirical evidence of improved performance across several challenging tasks over different modalities. We utilize previous studies on variational representation learning and regularization (Alemi et al. 2018; Pereyra et al. 2017; Achille and Soatto 2018) to interpret our findings, and conclude that our proposed derivation applies regularization over the variational classifier, preventing it from overfitting the learned representations and thus enabling greater MI between learned representation and the target Y.

The reader is encouraged to refer to the preliminaries provided in Appendix A before proceeding.

2 Related work

2.1 Deterministic Information Bottleneck

Classic information theory offers rate-distortion (Shannon 1959) to mitigate signal loss during compression: A source X is compressed to an encoding Z, such that maximal compression is achieved while keeping the encoding quality above a certain threshold. Encoding quality is measured by a task specific distortion function: $d : X \times Z \mapsto \mathbb{R}^+$. Rate-distortion suggests a mapping that minimizes the rate of bits to source sample, measured by $I(X;Z)$, that adheres to a chosen allowed expected distortion $D \geq 0$. The Information Bottleneck, IB, (Tishby, Pereira, and Bialek 1999) extends rate-distortion by replacing the tailored distortion functions with MI over a target distribution: Let Y be the target signal for some specific downstream task, such that the joint distribution $P^*(x, y)$ is known, and define the distortion function as MI between Z and Y. The IB is the solution to the optimization problem $Z : \underset{P(Z|X)}{min}\ I(X;Z)$ subject to $I(Z;Y) \geq D$, that can be optimized by minimizing the IB objective $\mathcal{L}_{IB} = I(X;Z) - \beta I(Z;Y)$ over $P(Z|X)$. The solution to the IB objective is a function of the Lagrange multiplier β, and is a theoretical limit for representation quality, given mutual information as an accepted metric, as elaborated in more detail in Appendix B. The IB is in fact an unsupervised soft clustering problem, where each data point x is assigned a probability to belong to the different clusters z, given the joint distribution of the input and target task $p^*(x, y)$ (Slonim 2002). Chechik et al. (2003) showed that computing the IB for continuous distributions is hard in the general case, and provided a method to optimize the IB objective in the case where X, Y are jointly Gaussian and known. Painsky and Tishby (2017) offered a limited linear approximation of the IB for any distribution by extracting the jointly Gaussian element of given distributions. Saxe et al. (2018) considered the application of the IB functional as an objective for DNNs, and concluded that computing mutual information in deterministic DNNs is problematic as the entropy term $H(Z|X)$ for a continuous Z is infinite. Amjad and Geiger (2020) extended this observation and pointed out that for a discrete Z MI becomes a piecewise constant function of its parameters, making gradient descent limiting and difficult.

2.2 Variational Information Bottleneck

Alemi et al. (2017) introduced the Variational Information Bottleneck (VIB) - a variational approximation for an upper bound to the IB objective for DNN optimization. Bounds for $I(X,Z)$ and $I(Z,Y)$ are derived from the non negativity of KL divergence, and are used to form an upper bound for the IB objective. A variational upper bound is derived by replacing intractable distributions with variational approximations. Using the 'reparameterization trick' (Kingma and Welling 2014) a discrete empirical estimation of the variational upper bound is then used as a loss function for classifier DNN optimization. The subsequent loss function is equivalent to the β-autoencoder loss (Higgins et al. 2017). VIB was evaluated over image classification tasks, and displayed substantial improvements in robustness to adversarial attacks, while causing a slight reduction in test set accuracy, comparing to equivalent deterministic models. (Achille and Soatto 2018) extended VIB with a total correlation term, designed to increase latent disentanglement. Fischer (2020) proposed an IB based loss function named Conditional Entropy Bottleneck (CEB), in which the conditional mutual information of X and Z given Y is minimized, instead of the unconditional mutual information. The CEB loss, $L_{CEB} = \min_{Z} I(X;Z|Y) - \gamma I(Y;Z)$, is designed to minimize all information in Z that is not relevant to the downstream task Y, by conditioning over Y. CEB is equivalent to IB for $\gamma = \beta - 1$ following the chain rule of mutual information (Cover 1999) and the IB Markov chain, as established in Appendix B. Similarly to VIB, a variational approximation for CEB was proposed as $L_{VCEB} = \mathbb{E}_{x,y} log\left(e(z|x)\right) - \mathbb{E}_{x,y} log\left(b(z|y)\right) - \gamma \mathbb{E}_{x,y} log\left(c(y|z)\right)$ and tested over the FMNIST (Xiao, Rasul, and Vollgraf 2017) and CIFAR10 (Krizhevsky 2009) datasets. Fischer (2020) shows that VIB is a special case of VCEB, where rate is approximated by the variational expression $\mathbb{E}_{x,y} log\left(e(z|x)\right) - \mathbb{E}_{x,y} log\left(b(z|y)\right)$ instead of $\mathbb{E}_{x,y} log\left(e(z|x)\right) - \mathbb{E}_{x,y} log\left(r(z)\right)$. Geiger and Fischer (2020) investigated wether VCEB is a tighter variational approximation to IB than VIB, and concluded that no ordering can be established in the general case, noting that any empirical improvement VCEB exhibits over VIB is not due to a tighter variational bound on the IB, but rather of VCEB being more amenable to optimization, or simply a successful loss function in its own regard. Cheng et al. (2020) proposed CLUB, an upper bound based MI estimator, that empirically outperformed the popular MINE estimator (Belghazi et al. 2018). Since CLUB is an upper bound for MI, it was evaluated as a replacement to the upper bound for the IB rate term, $I(X;Z)$, proposed in VIB (Alemi et al. 2017). CLUB based VIB was tested over the MNIST dataset (Deng 2012), resulting in a slight improvement in accuracy compared to VIB, without reporting adversarial robustness. We note that CLUB does not establish a tighter bound on the VIB rate term, and subsequently on the IB objective. We also note, that our current work derives a tighter bound on IB through the IB distortion term, $I(Z;Y)$, and that combining our suggested method with a CLUB bound on rate is an interesting avenue for future work.

2.3 Information theoretic regularization

Label smoothing (Szegedy et al. 2016) and entropy regularization (Pereyra et al. 2017) both regularize classifier DNNs by increasing the entropy of their output. This is achieved by either inserting a scaled conditional entropy term, $-\gamma \cdot H\left(p_\theta(y|x)\right)$, to the loss function, or by smoothing the training data labels. Applying either of these methods improved test accuracy and model calibration on various challenging classification tasks. Alemi et al. (2018) extends the information plane (Tishby, Pereira, and Bialek 1999) to VAE (Kingma and Welling 2014) settings, measuring distortion as MI between input and reconstructed images, and rate as MI between input and latent representation. The limits of representation quality in VAEs are looser than the theoretical IB limits, and heavily depend on the chosen variational families of the marginal and decoder distributions. The closer the families are

to the true distributions, the tighter the gap to the optimal IB limit. VAEs are susceptible to learn low quality representations, as the KL regularization term in the ELBO loss might induce a very uninformative representation, provided there's a strong enough decoder to compensate, by overfitting the encoder embeddings. This work is further elaborated on in Appendix B. In the current study, a conditional entropy term (Pereyra et al. 2017) emerges from our proposed derivation of a variational IB loss function, and we extend the theoretic framework proposed in (Alemi et al. 2018) to interpret why this new term facilitates a better variational approximation of the IB objective.

3 From VIB to VUB

As elaborated in Section 2.1, the IB objective, $\mathcal{L}_{IB} = I(X;Z) - \beta I(Z;Y)$, is computed over the joint distribution $p^*(x,y,z)$. When $p^*(x,y)$ is given, this expression is optimized over the distribution $p(z|x)$, as proposed by Tishby, Pereira, and Bialek (1999): $Z : \underset{p(z|x)}{min}\, I(X;Z)$ subject to $I(Z;Y) \geq D$. However, adapting IB to supervised tasks admits the learned classifier as a new RV to the optimization problem. Geiger and Fischer (2020) suggested the Markov chain $Y \leftrightarrow X \leftrightarrow Z \leftrightarrow \hat{Y}$ for supervised IB, distinguishing between the real unknown RV Y, and the learned classifier $\hat{Y}$. Following this logic, we argue that supervised IB optimization should be defined as: $Z, \tilde{Y} : \underset{p(z|x),p(\tilde{y}|z)}{min}\, I(X;Z)$ subject to $I(Z;\tilde{Y}) \geq D$.

The VIB loss (Alemi et al. 2017) consists of a cross entropy (CE) term, and a beta modulated KL regularization term, as in β-VAE loss (Higgins et al. 2017). The KL term is derived from a bound on the IB rate term $I(X;Z)$, while the CE term is derived from a bound on the IB distortion term $I(Z;Y) = H(Y) - H(Y|Z)$. During the derivation of the CE term, a relaxation is performed such that the term $H(Y)$ is assumed constant, and hence ignored, while the term $-H(Y|Z)$ is derived as CE between true and learned supervisor distributions $p^*(y), p(\tilde{y})$. We derive a new upper bound for the IB objective by not omitting $H(Y)$ from the distortion term. Subsequently, the variational approximation for our proposed bound is tighter when taking into account that the optimization process is done over $p(\tilde{y}|z)$ as well as $p(z|x)$. This modification attains a tighter variational bound on the IB objective for any Y with positive entropy, and a tighter empirical bound for all Y.

3.1 IB upper bound

We begin by establishing a new upper bound for the IB functional by bounding the mutual information terms, using the same method as in VIB.

Consider $I(Z;X)$:

$$I(Z;X) = \int\int p^*(x,z) log\left(p^*(z|x)\right) dxdz - \int p^*(z) log\left(p^*(z)\right) dz \tag{1}$$

For any probability distribution r we have that $D_{KL}\left(p^*(z)\middle|\middle|r(z)\right) \geq 0$, it follows that:

$$\int p^*(z) log\left(p^*(z)\right) dz \geq \int p^*(z) log\left(r(z)\right) dz \tag{2}$$

And so, by Equation 2:

$$I(Z;X) \leq \int\int p^*(x)p^*(z|x)log\left(\frac{p^*(z|x)}{r(z)}\right)dxdz \tag{3}$$

Consider $I(Z;Y)$:

For any probability distribution c we have that $D_{KL}\left(p^*(y|z)\middle||c(y|z)\right) \geq 0$, it follows that:

$$\int p^*(y|z)log\left(p^*(y|z)\right)dy \geq \int p^*(y|z)log\left(c(y|z)\right)dy \tag{4}$$

And so, by Equation 4:

$$\begin{aligned} I(Z;Y) &= \int\int p^*(y,z)log\left(\frac{p^*(y,z)}{p^*(y)p^*(z)}\right)dydz \geq \int\int p^*(y|z)p^*(z)log\left(\frac{c(y|z)}{p^*(y)}\right)dydz \\ &= \int\int p^*(y,z)log\left(c(y|z)\right)dydz + H_{p^*}(Y) \end{aligned} \tag{5}$$

We now diverge from the original VIB derivation by replacing $H_{p^*}(Y)$ with $H_c(Y|Z)$ instead of omitting it. In addition, we limit the new term to make sure that the inequality holds.

$$I(Z;Y) \geq \int\int p^*(y,z)log\left(c(y|z)\right)dydz + min\left\{H_{p^*}(Y), H_c(Y|Z)\right\} \tag{6}$$

We develop the first term in 6 using the IB Markov chain $Z \leftrightarrow X \leftrightarrow Y$ and total probability:

$$\begin{aligned} I(Z;Y) \geq &\int\int\int p^*(x)p^*(y|x)p^*(z|x)log\left(c(y|z)\right)dxdydz \\ &+min\left\{H_{p^*}(Y), -\int\int c(y,z)log\left(c(y|z)\right)dydz\right\} \end{aligned} \tag{7}$$

Denote $\tilde{Y}$ as a RV over the same support as Y, such that $Y \leftrightarrow X \leftrightarrow Z \leftrightarrow \tilde{Y}$ and $c(y,z) = c(\tilde{y},z), c(y|z) = c(\tilde{y}|z)$. We join Equation 3 with Equation 7 to establish a new upper bound for the IB objective:

$$\begin{aligned} L_{IB} \geq L_{UB} \equiv &\int\int p^*(x)p^*(z|x)log\left(\frac{p^*(z|x)}{r(z)}\right)dxdz \\ &-\int\int\int p^*(x)p^*(y|x)p^*(z|x)log\left(c(\tilde{y}|z)\right)dxdydz \\ &-min\left\{H_{p^*}(Y), -\int\int c(\tilde{y},z)log\left(c(\tilde{y}|z)\right)d\tilde{y}dz\right\} \end{aligned} \tag{8}$$

3.2 Variational approximation

Let $e(z|x)$ a variational encoder approximating $p^*(z|x)$ and $c(\tilde{y}|z)$ a variational classifier approximating $p^*(y|z)$. We define the variational approximation L_{VUB}:

$$
\begin{aligned}
L_{UB} \approx L_{VUB} \equiv & \beta \int\int p^*(x)e(z|x)log\left(\frac{e(z|x)}{r(z)}\right)dxdz \\
& - \int\int\int p^*(x)p^*(y|x)e(z|x)log\left(c(\tilde{y}|z)\right)dxdydz \\
& - min\left\{H_{p^*}(Y), -\int\int\int p^*(x)e(z|x)c(\tilde{y}|z)log\left(c(\tilde{y}|z)\right)dxd\tilde{y}dz\right\}
\end{aligned} \tag{9}
$$

3.3 Empirical estimation

The real and continuous distribution $p^*(x,y) = p^*(y|x)p^*(x)$ can be estimated by Monte Carlo sampling from a discrete dataset $\mathcal{S}$, and distributions featuring Z are sampled from a stochastic encoder. Let $e_\phi(z|x) \sim N(\mu, \Sigma)$ be a stochastic DNN encoder with parameters ϕ, and a final layer of dimension $2K$, such that for each forward pass, the first K entries are used to encode μ, and the last K entries to encode a diagonal Σ, after a soft-plus transformation. For each $x_n \in \mathcal{S}$ we generate a sample $\hat{z}_n$ from the encoder, using the reparameterization trick (Kingma and Welling 2014). Let C_λ be a discrete classifier neural net parameterized by λ, such that $C_\lambda(\tilde{y}|z) \sim Categorical$, let $\hat{H}_{\mathcal{S}}(Y)$ be the empirical entropy of the real RV Y as measured from the training dataset $\mathcal{S}$, and let $\tilde{Y}$ be the learned classifier. We chose a standard Gaussian as a variational approximation for the marginal $r(z)$.

$$
\hat{L}_{VUB} \equiv \frac{1}{N}\sum_{n=1}^{N}\left[\beta D_{KL}\left(e_\phi(z|x_n)\middle\|r(z)\right) - log\left(C_\lambda\left(y_n|\hat{z}_n\right)\right) - min\left\{\hat{H}_{\mathcal{S}}(Y), H_{C_\lambda}\left(\tilde{Y}|Z\right)\right\}\right] \tag{10}
$$

3.4 Intuition

Based on Equation 5 and our definition of $\tilde{Y}$ in Section 3.1 we give the following formulation of the Barber-Agakov bound and identity (Barber and Agakov 2003):

$$
I(Z;Y) \geq \int\int p^*(y,z)log\left(c(\tilde{y}|z)\right)dydz + H_{p^*}(Y) \equiv \tilde{I}(Z;Y)
$$

The following inequality holds for all distributions of Y with non negative entropy:

$$
H_{p^*}(Y) \geq I(Z;Y) \geq \tilde{I}(Z;Y) \geq \int\int p^*(y,z)log\left(c(\tilde{y}|z)\right)dydz + min\left\{H_{p^*}(Y), H_c(\tilde{Y}|Z)\right\}
$$

We have that the true MI $I(Z;Y)$ is squeezed by the Barber-Agakov MI $\tilde{I}(Z;Y)$ from below, which is in term squeezed by the VUB distortion term. All variational IB derivations (Alemi et al. 2017; Fischer 2020; Cheng et al. 2020) omit the entropy term $H_{p^*}(Y)$, and derive the conditional entropy term $\int\int p^*(y,z)log\left(c(\tilde{y}|z)\right)dydz$ into empirical cross entropy which is minimized during optimization. Assuming the conditional entropy term is minimized, intuition suggests that increasing the entropy term $H_c(\tilde{Y}|Z)$ as close as possible to $H_{p^*}(Y)$ will push the true MI $I(Z;Y)$ closer to its theoretical limit. Since the optimization cannot change the entropy of the true RV Y, the suggested increase in $I(Z;Y)$ can only be caused by a more informative representation Z. Figure 1 illustrates the possible effect of an increase in variational entropy: The left hand diagram suggest a model with low variational entropy, and a low variational mutual information, and the right hand diagram suggest

a model with increased variational entropy. Reduction in cross entropy pushes $I(Z;\hat{Y})$ up, and the real mutual information $I(Z;Y)$ increases as a result of the Barber-Agakov inequality (Barber and Agakov 2003).

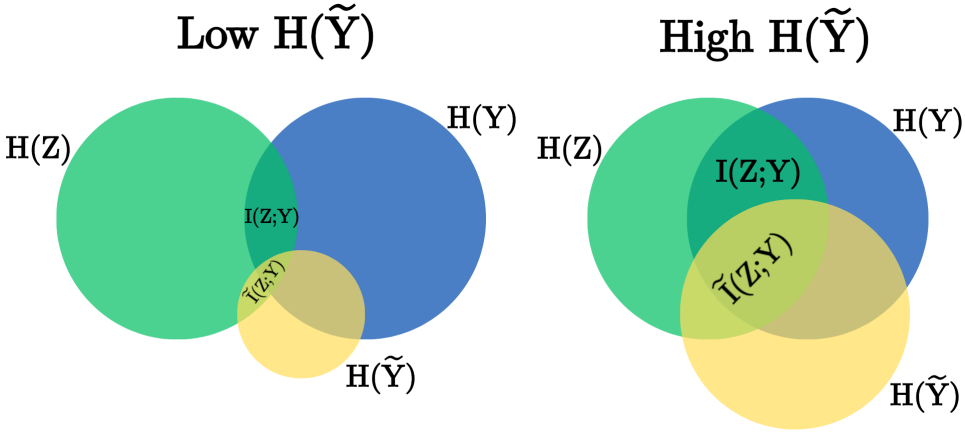

Figure 1: Venn diagrams illustrating the possible increase in true mutual information as a result of increased variational entropy. The right diagram features higher variational entropy, and higher variational mutual information that induce higher real mutual information as a result of the Barber-Agakov inequality (Barber and Agakov 2003).

4 Experiments

We follow the experimental setup proposed by Alemi et al. (2017), extending it to NLP tasks as well. We trained image classification models on the ImageNet 2012 dataset (Deng et al. 2009), and text classification on the IMDB sentiment analysis dataset (Maas et al. 2011). For each dataset, we compared a competitive Vanilla model with a VIB and a VUB model trained with beta values of $\beta = 10^{-i}$ for $i \in \{1, 2, 3\}$. Each model was trained and evaluated 5 times per β value, with consistent performance and statistical significance shown by a Wilcoxon rank sum test. Each model was evaluated using test set accuracy, and robustness to various adversarial attacks. For image classification, we employed the untargeted Fast Gradient Sign (FGS) attack (Goodfellow, Shlens, and Szegedy 2015) as well as the targeted CW L_2 attack (Carlini and Wagner 2017), (Kaiwen 2018). For text classification, we used the untargeted Deep Word Bug attack (Gao et al. 2018), (Morris et al. 2020) as well as the untargeted PWWS attack (Ren et al. 2019). Elaboration on the experimental setup, results and further insights from the experiments are available in Appendix C. Code to reconstruct the experiments is provided to the supplementary materials of this paper.

4.1 Image classification

A pre-trained inceptionV3 (Szegedy et al. 2016) base model was used and achieved a 77.21% accuracy on the ImageNet 2012 validation set (Test set for ImageNet is unavailable). Image classification evaluation results are shown in Table 1, examples of successful attacks are shown in Figures 5, 6 in Appendix C. The empirical results presented in Table 1 confirm that while VIB reduces performance on the validation set, it substantially improves robustness to adversarial attacks. Moreover, these results demonstrate that VUB significantly outperforms VIB in terms of validation accuracy, while providing competitive robustness to attacks similarly to VIB. A comparison of the best VIB and VUB models further substantiates these findings, with statistical significance confirmed by a p-value of less than 0.05 in a Wilcoxon rank sum test.

β	**Val** ↑	**FGS** ↓ $_{\epsilon=0.1}$	**FGS** ↓ $_{\epsilon=0.5}$	**CW**↑
		Vanilla model		
-	**77.2%**	68.9%	67.7%	788
		VIB models		
10^{-3}	73.7% ±.1%	**59.5%** ±.2%	**63.9%** ±.2%	**3917** ±291
10^{-2}	72.8% ±.1%	**53.5%** ±.2%	**62.0%** ±.1%	**3318** ±293
10^{-1}	72.1% ±.01%	58.4% ±.1%	**62.0%** ±.1%	**3318** ±293
		VUB models		
10^{-3}	**75.5%** ±.03%	62.8% ±.1%	66.4% ±.1%	2666 ±140
10^{-2}	**75.0%** ±.05%	57.6% ±.2%	64.3% ±.1%	1564 ±218
10^{-1}	**74.8%** ±0.09%	**57.9%** ±.5%	64.8% ±.5%	**3575** ±456

Table 1: ImageNet evaluation scores for vanilla, VIB and VUB models, average over 5 runs with standard deviation. First column is performance on the ImageNet validation set (higher is better ↑), second and third columns are the % of successful FGS attacks at $\epsilon = 0.1, 0.5$ (lower is better ↓), and the fourth column is the average L_2 distance for a successful Carlini Wagner L_2 targeted attack (higher is better ↑). VUB attains significantly higher accuracy over unseen data in all settings, while preserving competitive robustness to adversarial attacks.

4.2 Text classification

A fine tuned BERT uncased (Devlin et al. 2019) base model was used and achieved a 93.0% accuracy on the IMDB sentiment analysis test set. Text classification evaluation results are shown in Table 2, examples of successful attacks are shown in Figure 4 in Appendix C. In this modality, VUB significantly outperforms VIB in both test set accuracy and robustness to the two attacks. Moreover, VUB also outperformed the original model in terms of test set accuracy. A comparison of the best VIB and VUB models further substantiates these findings, with statistical significance confirmed by a p-value of less than 0.05 in a Wilcoxon rank sum test.

5 Discussion

The IB is a private case of rate distortion, and was initially designed as an optimization problem over compressed representations. Adapting the IB objective for supervised tasks results in optimization of a classifier distribution as well, and requires a reformulation of the initial optimization problem to include both representation and descriminator. Following this logic, assuming a constant $H(Y)$ relaxes the problem, and lifting this assumption lead us to derive a tighter variational bound over the optimized objective. When used as a loss function, our proposed bound produces significantly better classification accuracy, with equivalent or superior robustness to adversarial attacks, over high dimensional tasks of different modalities, with high statistical significance. On a practical level, the conditional entropy term that follows from our proposed derivation provides strong classifier regularization, as shown in (Pereyra et al. 2017). This type of regularization is a possible remedy to the inbalances inherit in the ELBO loss function, as they were described by Alemi et al. (2018),

β	**Test**$\uparrow$	**DWB**$\downarrow$	**PWWS**$\downarrow$
		Vanilla model	
-	93.0%	54.3%	100%
		VIB models	
10^{-3}	91.0% $\pm 1.0\%$	35.1% $\pm 4.4\%$	41.6% $\pm 6.6\%$
10^{-2}	90.8% $\pm 0.5\%$	41.0% $\pm 4.8\%$	62.9% $\pm 14.3\%$
10^{-1}	**89.4%** $\pm .9\%$	**90.0%** $\pm 8.0\%$	**99.1%** $\pm 0.9\%$
		VUB models	
10^{-3}	**93.2%** $\pm .5\%$	**27.5%** $\pm 2.0\%$	**28.4%** $\pm 1.3\%$
10^{-2}	**92.6%** $\pm .8\%$	**30.8%** $\pm 2.0\%$	**50.0%** $\pm 4.8\%$
10^{-1}	89.2% $\pm 2.0\%$	99.2% $\pm 0.5\%$	100% $\pm 0\%$

Table 2: IMDB evaluation scores for vanilla, VIB and VUB models, average over 5 runs with standard deviation. First column is performance over the test set (higher is better $\uparrow$), second is % of successful Deep Word Bug attacks (lower is better $\downarrow$) and third is % of successful PWWS attacks (lower is better $\downarrow$). In almost all cases VUB attains significantly higher accuracy over unseen data, as well as significantly higher robustness to adversarial attacks. For this modality, VUB also outperforms the vanilla model in terms of test set accuracy for $\beta = 10^{-3}$.

and inherintly to VIB also, as it is equivalent to a beta regulated ELBO loss. In addition, we propose an possible intuitive explnation to the effects of conditional entropy regularization on the quality of learned representaiton, by showing that in the extreme cases high variational entropy can skew the true mutual information $I(Z;Y)$ upwards, implying better representations learned. While other advancements have been done in recent years, (Fischer 2020; Cheng et al. 2020), none show a tighter bound than VIB, and all modify the derivation of the rate term, while our work derives an upper bound by modifying the distortion term.

While providing a complete framework for optimal data modeling, the IB, and its variational approximations, rely on three assumptions: (1) It suffices to optimize the mutual information metric to optimize a model's performance; (2) Forgetting more information about the input, while keeping relevant information about the output, induces better generalization; (3) Mutual information between the input, output and latent representation can be either computed or approximated to a desired level of accuracy. Our improved empirical results, induced by a tighter bound, suggest better data modeling, and strengthen the cause for the IB, and its variational approximations, as a learning framework for classifier DNNs. Conversely, one might argue that the improved adversarial robustness of both VIB and VUB is an artifact of latent geometry: Both methods promote a disentangled latent space by using a stochastic factorized prior, as suggested by Chen et al. (2018). In addition, both utilize KL regularization, enforcing clustering around a 0 mean, which might increase latent smoothness. These geometric traits of the latent space can make it difficult for minor perturbations to significantly alter latent semantics, possibly making the models more robust to attacks.

In conclusion, while the IB and its variational approximations do not provide a complete theoretical framework for DNN data modeling and regularization, they offer a strong, measurable, and theo-

retically grounded approach. VUB is presented as a tractable and tighter upper bound of the IB functional, that can be easily adapted to any classifier DNN to significantly increase robustness to various adversarial attacks, while inflicting minimal decrease in test set performance, and in some cases even increasing it.

This study opens many opportunities for further research: Further improvements to the upper bound, including combining VUB with proposed new bounds for the IB rate term such as CLUB (Cheng et al. 2020); Applying VUB in self-supervised learning, and in particular to measure whether representations learned with VUB capture better semantics than representations learned with non IB inspired loss functions; Finally, comparing the effects of latent geometry vs rate distortion on adversarial robustness is left to future work.

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.

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Google Research.

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J. V.; Saurous, R. A.; and Murphy, K. 2018. Fixing a Broken ELBO. In *Proceedings of Machine Learning Research*, volume 80, 159–168. PMLR.

Amjad, R. A.; and Geiger, B. C. 2020. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9): 2225–2239.

Barber, D.; and Agakov, F. V. 2003. The IM algorithm: a variational approach to Information Maximization. In *Neural Information Processing Systems*.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Hjelm, R. D.; and Courville, A. C. 2018. Mutual Information Neural Estimation. In *International Conference on Machine Learning*.

Bengio, Y. 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2(1): 1–127.

Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57. IEEE Computer Society.

Chechik, G.; et al. 2003. Gaussian information bottleneck. In *Advances in Neural Information Processing Systems*.

Chen, T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2615–2625.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1779–1788. PMLR.

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota.

Fischer, I. 2020. The Conditional Entropy Bottleneck. *Entropy*, 22(9): 999.

Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, 50–56. IEEE.

Geiger, B.; and Fischer, I. 2020. A Comparison of Variational Bounds for the Information Bottleneck Functional. *Entropy*.

Goldfeld, Z.; and Polyanskiy, Y. 2020. The Information Bottleneck Problem and its Applications in Machine Learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 19–38.

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

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

Kaiwen. 2018. pytorch-cw2. GitHub repository.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kingma, D. P.; and Welling, M. 2019. An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*, 12(4): 307–392.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. 32–33.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.

Painsky, A.; and Tishby, N. 2017. Gaussian Lower Bound for the Information Bottleneck Limit. *J. Mach. Learn. Res.*, 18: 213:1–213:29.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; and Hinton, G. E. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1085–1097.

Saxe, A. M.; Bansal, Y.; Dapello, J.; and Advani, M. 2018. On the information bottleneck theory of deep learning.

Shannon, C. E. 1959. Coding Theorems for a Discrete Source With a Fidelity Criterion. In *IRE National Convention*.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. 19 pages, 8 figures, arXiv:arXiv:1703.00810.

Slonim, N. 2002. *The information bottleneck: Theory and applications*. Ph.D. thesis, Hebrew University of Jerusalem Jerusalem, Israel.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The Information Bottleneck Method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*. Hebrew University, Jerusalem 91904, Israel.

Tishby, N.; and Zaslavsky, N. 2015. Deep Learning and the Information Bottleneck Principle. arXiv:1503.02406.

Vapnik, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc. ISBN 0-387-94559-8.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. Cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

Appendix A - Preliminaries

We denote random variables (RVs) with upper cased letters X, Y, and their realizations in lower case x, y. Denote discrete Probability Mass Functions (PMFs) with an upper case $P(x)$ and continuous Probability Density Functions (PDFs) with a lower case $p(x)$. Hat notation denotes empirical measurements.

Let X, Y be two observed random variables with an unknown joint distribution $p^*(x, y)$ and marginals $p^*(x)$, $p^*(y)$. We can attempt to approximate these distributions using a model p_θ with parameters θ, such that for generative tasks $p_\theta(x) \approx p^*(x)$, and for discriminative tasks $p_\theta(y|x) \approx p^*(y|x)$, using a dataset $\mathcal{S} = \{(x_1, y_1), ..., (x_N, y_N)\}$ to fit our model. One can also assume the existence of an additional unobserved RV $Z \sim p^*(z)$ that influences or generates the observed RVs X, Y. Since Z is unobserved, it is absent from the dataset $\mathcal{S}$, and so cannot be modeled directly. Denote $p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz = \int p_\theta(x, z)dz$ the marginal, $p_\theta(z)$ the prior as it is not conditioned over any other RV, and $p_\theta(z|x)$ the posterior following Bayes' rule.

When modeling an unobserved variable of an unknown distribution, we encounter a problem as the marginal $p_\theta(x) = \int p_\theta(x, z)dz$ doesn't have an analytic solution. This intractability can be overcome by choosing some tractable parametric variational distribution $q_\phi(z|x)$ to approximate the posterior $p_\theta(z|x)$, such that $q_\phi(z|x) \approx p_\theta(z|x)$, and estimate $p_\theta(x, z)$ or $p_\theta(x, z|y)$ by fitting the dataset $\mathcal{S}$ (Kingma and Welling 2019).

Vapnik (1995) defines *supervised* learning as follows:

- A generator of random vectors $x \in \mathbb{R}^d$, drawn independently from an unknown probability distribution $p^*(x)$.

- A supervisor who returns a scalar output value $y \in \mathbb{R}$, according to an unknown conditional probability distribution $p^*(y|x)$. We note that these probabilities can indeed be soft labels, where y is a continuous probability vector, rather the more commonly used hard labels.

- A learning machine capable of implementing a predefined set of functions, $f(x, \theta) : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}$, where Θ is a set of parameters.

The problem of learning is that of choosing from the given set of functions, the one that best approximates the supervisor's response. The choice is typically based on a training set of n independent and identically distributed pairs of observations drawn according to $p(x, y) = p(x)p(y|x) : \mathcal{S}$.

Given a set of unlabeled data points $\{x_1, ..., x_n\}, x_i \in \mathbb{R}^d$, Slonim (2002) defines *unsupervised* learning as the task of constructing a compact representation of these points, which in some sense reveals their hidden structure. This representation can be used further to achieve a variety of goals, including reasoning, prediction, communication etc. In particular, unsupervised clustering partitions the data points into exhaustive and mutually exclusive clusters, where each cluster can typically be represented by a centroid, typically a weighted average of the cluster's members. Soft clustering assigns cluster probabilities for each data point, and fits an assignment by minimizing the expected loss for these probabilities, usually a distance metric such as MSE.

In this work, information theoretic functions share the same notation for discrete and continuous settings and are denoted as follows:

	Notation	Differential	Discrete
Entropy	$H_p(X)$	$-\int p(x)log\left(p(x)\right)dx$	$-\sum_{x\in X} P(x)log\left(P(x)\right)$
Conditional entropy	$H_p(X\|Y)$	$-\int\int p(x,y)log\left(p(x\|y)\right)$ $dxdy$	$-\sum_{x\in X}\sum_{y\in Y}$ $P(x,y)log\left(P(x\|y)\right)$
Cross entropy	$CE(p,q)$	$-\int p(x)log\left(q(x)\right)dx$	$-\sum_{x\in X} P(x)log\left(Q(x)\right)$
Joint entropy	$H_p(X,Y)$	$-\int\int p(x,y)log\left(p(x,y)\right)$ $dxdy$	$-\sum_{x\in X}\sum_{y\in Y}$ $P(x,y)log\left(P(x,y)\right)$
KL divergence	$D_{KL}\left(p\|\|q\right)$	$\int p(x)log\left(\frac{p(x)}{q(x)}\right)dx$	$\sum_{x\in X} P(x)log\left(\frac{P(x)}{Q(x)}\right)$
Mutual information (MI)	$I(X;Y)$	$\int\int p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right)$ $dxdy$	$\sum_{x\in X}\sum_{y\in Y}$ $P(x,y)log\left(\frac{P(x,y)}{P(x)P(y)}\right)$

Appendix B - Related work elaboration

This appendix expands the related work Section 2, by providing a deeper review of the IB, the IB theory of deep learning, and variational approximations for the IB.

The Information Plane

As mentioned in Section 2.1, the solution to the IB objective, $\mathcal{L}_{IB} = I(X;Z) - \beta I(Z;Y)$, depends on the Lagrange multiplier β. Hence, the IB objective has no one unique solution, and can thus be plotted as a function of β and of Z's cardinality, over a Cartesian system composed of the axes $I(X;Z)$ (rate) and $I(Z;Y)$ (distortion). When plotted as a function of β the IB functional is called the 'information curve' and its Cartesian system the 'information plane' (Tishby, Pereira, and Bialek 1999), as illustrated in Figure 2. When β approaches 0 the distortion term is nullified and we learn a representation that has no information over the down stream task, and maximal compression (such a representation may be a null vector). When β approaches ∞ we learn a representation that has the maximal possible information over the downstream task, but that also holds the maximal information over the training data, hence overfitted. The region above the information curve is unreachable by any possible representation. The different bifurcation of the information curve, illustrated in Figure 2, correspond to the different possible cardinalities of the compressed representation.

Fixing a Broken Elbo

As mentioned in 2.3, Alemi et al. (2018) extends the information plane with an additional theoretical bound for rate distortion ratio, imposed by the usage of finite parametric families of variational approximations. Alemi et al. (2018) observed that the KL regularization term in the ELBO loss function isn't directly affected by the quality of the reconstructed image, and vice versa. Following this logic, given a powerful enough decoder, optimizing the ELBO loss can result in poor latent representation, com-

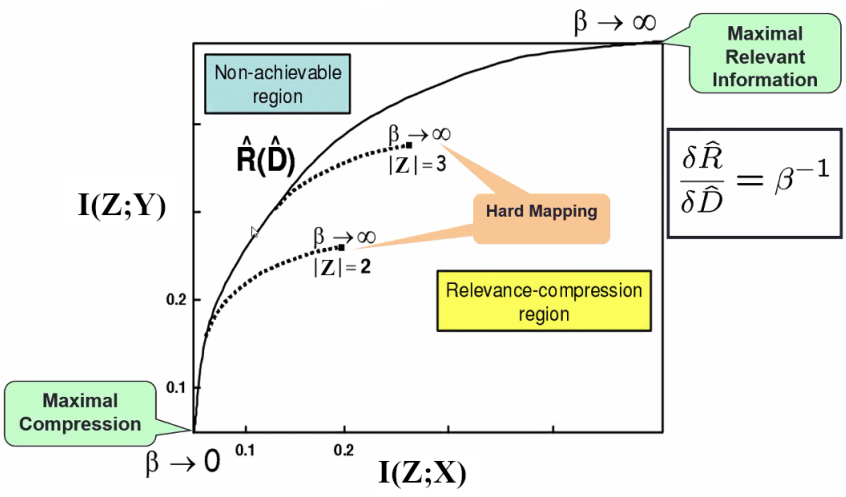

Figure 2: The information plane and curve: rate-distortion ratio over β. At $\beta = 0$ the representation is compressed but uninformative (maximal compression), at $\beta \to \infty$ the representation is informative but potentially overfitted (maximal information). Taken from (Slonim 2002).

pensated for by an overfitted decoder. Mutual information is used to measure representation quality: $I_e(X;Z) \equiv \int\int p_e(x,z) log\left(\frac{p_e(x,z)}{p^*(x)p_e(z)}\right) dxdz$, for a stochastic encoder e. I_e can be bounded from both sides as follows: $H_{p^*}(X) - I_d(X;\hat{X}) \leq I_e(X;Z) \leq D_{KL}\left(q_\phi(z|x)\middle||p_\theta(z)\right)$, where $H_{p^*}(X)$ is the true data entropy, I_d is the variational decoder's distortion. It is observed, that the KL term does not depend on the variational decoder distribution, and can be minimized regardless of reconstruction quality. Equivalently, good reconstruction does not directly depend on good representation. Figure 3 is suggested as an extension to the information plane (Tishby, Pereira, and Bialek 1999), by replacing real rate and distortion with their variational approximations. The new plane is divided into three sub planes: (1) Infeasible: This is the IB theoretical limit (As per Figure 2); (2) Feasible: Attainable given an infinite model family, and complete variety of: $e(z|x), d(x|z)$ and $p_\theta(z)$; (3) Realizable: Attainable given a finite parametric and tractable variational family.

The black diagonal line at the lower left of Figure 3 satisfies $H_{p^*}(X) - I_d(X;\hat{X}) = D_{KL}\left(q_\phi(z|x)\middle||p_\theta(z)\right)$, resulting in tight variational bounds on the mutual information: $R \leq I_e \leq R$. (Alemi et al. 2018) notes that a common degeneration of representations in VAEs is a decrease in rate and distortion over the training data, together with an increase in both over unseen data. This is caused by overpowerful decoders that overfit an uninformative learned representation. Low ELBO loss can be attained as both the distortion and rate terms are minimized during training. $D_{KL}\left(q_\phi(z|x)\middle||p_\theta(z)\right)$ approaches 0 iff $e(z|x) \to p_\theta(z)$. In this case, $e(z|x)$ is close to independence of x, and the latent representation fails to encode information about the input. However, a suitably powerful decoder could non the less learn to overfit encoded traces of the training examples, and reach a low distortion score during optimization. In the current study, we extend this theoretical framework to explain the advancements of our proposed loss function.

IB theory of deep learning

The following is a summary of work leveraging the IB framework for deterministic DNN optimization and interpretation. For a more comprehensive review of this opinion-splitting topic, the reader is advised to consult the work of Goldfeld and Polyanskiy (2020).

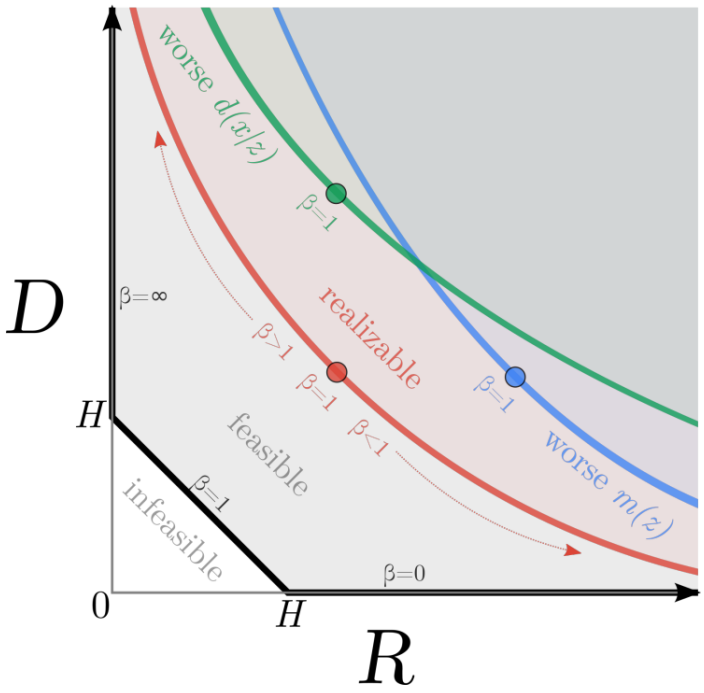

Figure 3: Phase diagram, a proposed information plane interpretation of VAEs. Axes are variational rate and distortion. The IB theoretical limit is extended by an additional limit induced by the constraint of a finite parametric variational family. Once a family is chosen, we seek to learn an optimal marginal $m(z)$ and decoder $d(x|z)$ in order to approach the new limit. β modulation controls the tradeoff between rate and distortion, regardless of the variational family. Note that this figure is inverted in orientation to Figure 2, i.e. low distortion corresponds to better performance, and not to lower MI. Taken from (Alemi et al. 2018).

Tishby and Zaslavsky (2015) proposed a representation-learning interpretation of DNNs using the IB framework, regarding DNNs as Markov cascades of intermediate representations between hidden layers. Under this notion, comparing the optimal and the achieved rate-distortion ratios between DNN layers will indicate if a model is too complex or too simple for a given task and training set. Shwartz-Ziv and Tishby (2017) visualized and analyzed the information plane behavior of DNNs over a toy problem with a known joint distribution. Mutual information of the different layers was estimated and used to analyze the training process. The learning process over Stochastic Gradient Descent (SGD) exhibited two separate and sequential behaviors: A short Empirical Error Minimization phase (ERM) characterized by a rapid decrease in distortion, followed by a long compression phase with an increase in rate until convergence to an optimal IB limit as demonstrated in Figure 4. Similar yet repetitive behavior was observed in the current study, as elaborated in Section **??**.

Saxe et al. (2018) reproduced the experiments described in (Shwartz-Ziv and Tishby 2017), expanding them to different activation functions, different datasets and different methods to estimate mutual information. It was found that double-sided saturated nonlinear activations, such as the tanh, produced a distinct compressions stage when mutual information was measured by binning, as performed in (Shwartz-Ziv and Tishby 2017), while other activations did not. It was also shown that DNN generalization did not depend on a distinct compression stage, and that DNNs do forget task irrelevant information, but this happens concurrently to the learning of task relevant information, and not necessarily in a distinct phase. Amjad and Geiger (2020) also showed that equivalent representations might yield the same IB loss while one achieved better classification rate than the other. These

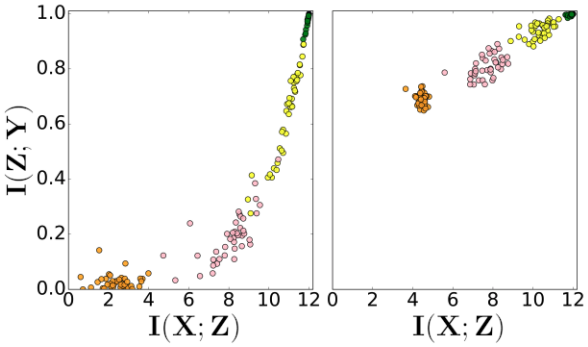

Figure 4: Information plane scatters of different DNN layers (colors) in 50 randomized networks. From Shwartz-Ziv and Tishby (2017). Left are initial weights, Right are at 400 epochs. Our study reproduced similar yet repetitive behavior on complicated high dimensional tasks, as elaborated in Section **??** and in Figure 7.

discrepancies are caused because mutual information in deterministic DNNs is either infinite or step like, because of mutual information's invariance to invertible transformations and because of the absence of a decision function in the objective.

When examining the information plane behavior in the current study, we notice recurring patterns of distortion reduction followed by rate increase, resembling the ERM and representation compression stages described by Shwartz-Ziv and Tishby (2017), as elaborated in Appendix 5.

Conditional Entropy Bottleneck

As mentioned in Section 2.2, Fischer (2020) showed that the conditional entropy bottleneck is equivalent to IB for $\gamma = \beta - 1$ following the chain rule of mutual information and the IB Markov chain. We develop this equivalence in detail:

$$
\begin{aligned}
CEB =& I(X;Z|Y) - \gamma I(Z;Y) \\
\overset{\text{MI chain rule}}{=}& H(Z|Y) - H(Z|X,Y) - \gamma I(Z;Y) \\
\overset{Z \leftarrow X \leftrightarrow Y}{=}& H(Z|Y) - H(Z|X) - \gamma I(Z;Y) \\
\overset{\gamma := \beta - 1}{\Longrightarrow}& H(Z|Y) - H(Z|X) - (\beta - 1) I(Z;Y) \\
=& H(Z|Y) - H(Z|X) - \beta I(Z;Y) + I(Z;Y) \\
=& H(Z|Y) - H(Z|X) - \beta I(Z;Y) + H(Z) - H(Z|Y) \\
=& H(Z) - H(Z|X) + H(Z|Y) - H(Z|Y) - \beta I(Z;Y) \\
=& I(X;Z) - \beta I(Z;Y)
\end{aligned}
$$

Appendix C - Experiments elaboration

Image classification models were trained on the first 500,000 samples of the ImageNet 2012 dataset (Deng et al. 2009), and text classification over the entire IMDB sentiment analysis dataset (Maas et al. 2011). For each dataset, a competitive pre-trained model (Vanilla model) was evaluated and then used to encode embeddings. These embeddings were then used as a dataset for a new stochastic classifier net with either a VIB or a VUB loss function. Stochastic classifiers consisted of two ReLU

activated linear layers of the same dimensions as the pre-trained model's logits (2048 for image and 768 for text classification), followed by reparameterization and a final softmax activated FC layer. Learning rate was 10^{-4} and decaying exponentially with a factor of 0.97 every two epochs. Batch sizes were 32 for ImageNet and 16 for IMDB. All models were trained using an Nvidia RTX3080 GPU with approximately 1-2 days per a single experiment run. Beta values of $\beta = 10^{-i}$ for $i \in \{1, 2, 3\}$ were tested, and we used a single forward pass per sample for inference, since previous studies indicated that these are the best range and sample rate for VIB (Alemi et al. 2017, 2018). Each model was trained and evaluated 5 times per β value, with consistent performance. Statistical significance was demonstrated in all comparisons using the Wilcoxon rank sum test as follows: For each compared metric a sorted vector of results was prepared, where each entry featured the attained result in each of the 5 i.i.d. experiments per algorithm, and a boolean indicator value for the algorithm type. All metrics compared attained a p-value of less than 0.05. For example:$r = ((0.94, 1)\ (0.935, 1)\ (0.93, 1)\ (0.93, 1)\ (0.925, 1)\ (0.92, 0)\ (0.915, 0)\ (0.915, 0)\ (0.91, 0)\ (0.89, 0))^{-1}$ be a sorted vector of (test accuracy, algorithm) tuples, 1 being VUB, 0 VIB. We computed the rank-sum as follows:

$$\mu_T = \frac{5 \cdot 11}{2} = 27.5,\ \sigma_T = \sqrt{\frac{5 \cdot 5 \cdot 11}{12}} \approx 4.78,\ Z(T) = \frac{15 - 27.5}{4.78} \approx -2.61$$

$$\Phi^{-1}(pval) = -2.61,\ pval = 0.0045 \leq 0.05$$

In practice, these were computed with the Python Scipy library:

```
import scipy.stats as stats
vib_scores = [0.915, 0.915, 0.91, 0.92, 0.89]
vub_scores = [0.93, 0.935, 0.925, 0.93, 0.94]
pvalue = stats.ranksums(vub_scores, vib_scores, 'greater').pvalue
assert pvalue < 0.05
```

Image classification

The ImageNet 2012 validation set was used for evaluation as the test set for ImageNet is unavailable. InceptionV3 yields a slightly worse single shot accuracy than inceptionV2 (80.4%) when run in a single model and single crop setting, however we've used InceptionV3 over V2 for simplicity. Each model was trained for 100 epochs. The entire validation set was used to measure accuracy and robustness to FGS attacks, while only 1% of it was used for CW attacks, as they are computationally expensive.

Text classification

Each model was trained for 150 epochs. The entire test set was used to measure accuracy, while only the first 200 entries in the test set were used for adversarial attacks as they are computationally expensive.

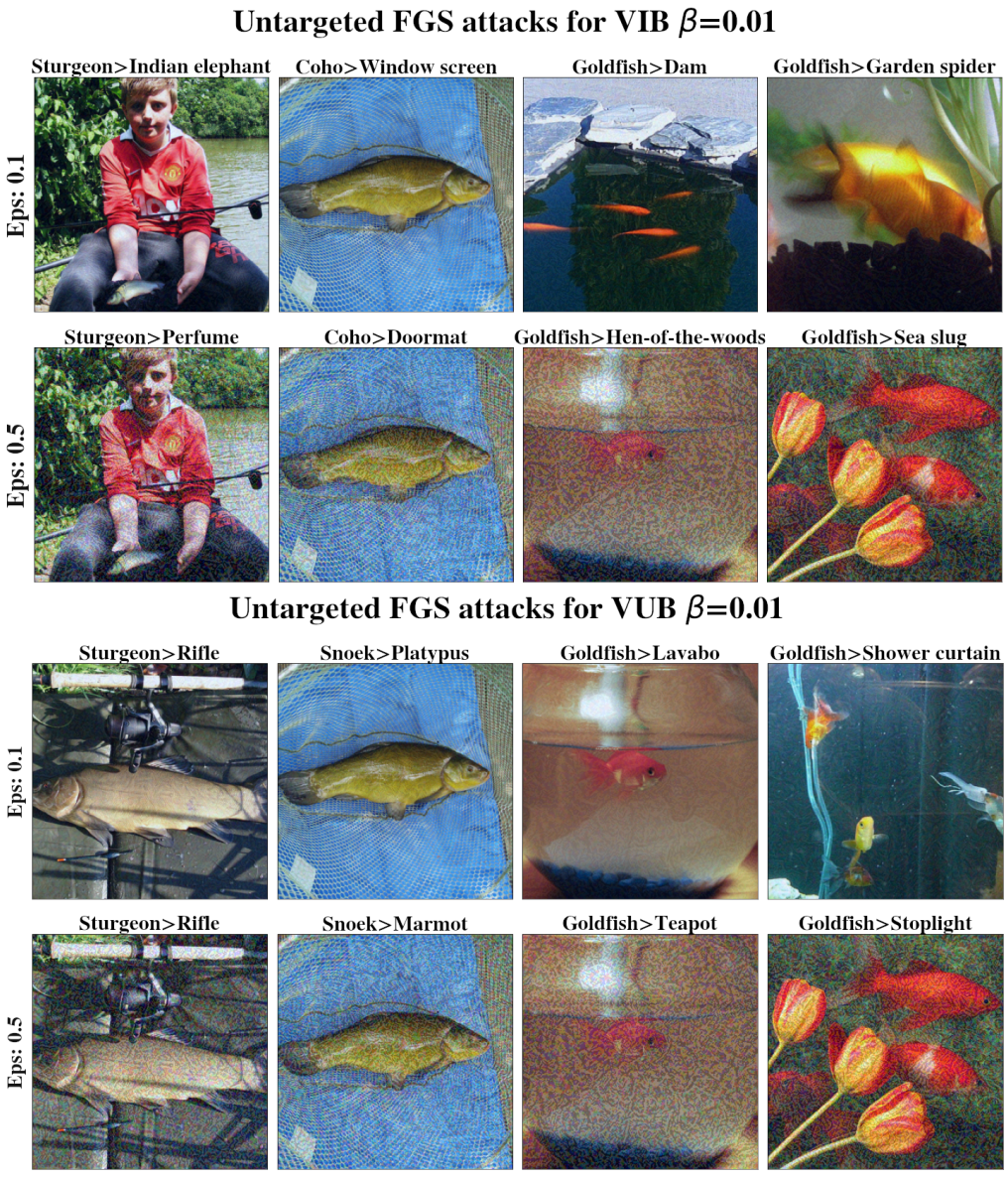

Figure 5: Successful untargeted FGS attack examples. Images are perturbations of previously successfully classified instances from the ImageNet validation set. Perturbation magnitude is determined by the parameter ϵ shown on the left, the higher, the more perturbed. Original and wrongly assigned labels are listed at the top of each image. Notice the deterioration of image quality as ϵ increases.

Figure 6: Successful targeted CW attack examples. Images are perturbations of previously successfully classified instances from the ImageNet validation set. The target label is 'Soccer ball'. Average L_2 distance required for a successful attack is shown on the left. The higher the required L_2 distance, the greater the visible change required to fool the model. Original and wrongly assigned labels are listed at the top of each image. Mind the difference in noticeable change as compared to the FGS perturbations presented in Figure 5.

Original text
the acting , costumes , music , cinematography and sound are all *astounding* given the production 's austere locales .
Perturbed text
the acting , costumes , music , cinematography and sound are all *dumbfounding* given the production 's austere locales .

Table 3: Example of a successful PWWS attack on a vanilla Bert model, fine tuned over the IMDB dataset. The original label is 'Positive sentiment'. The substituted word, marked in italic font, changed the classification to 'Negative sentiment'. VUB and VIB classifiers are far less susceptible to these perturbations as shown in Table 2.

In addition to the above evaluation metrics, we also measured approximated rate and distortion throughout training, and plotted them on the information plane as shown in Figure 7 in Appendix 5. Examining the resulting curve, we notice recurring patterns of distortion reduction followed by rate increase, resembling the ERM and representation compression stages described by Shwartz-Ziv and Tishby (2017). While previous research has documented the occurrence of error minimization and representation compression phases (Shwartz-Ziv and Tishby 2017), our work revealed that these phases can occur in cycles throughout training. This finding is particularly noteworthy, because previous studies observed this phenomenon in simple toy problems, whereas our research demonstrated it in complex tasks of high dimensionality, with unknown distributions. This suggests that this

Original text
great historical movie, will not allow a viewer to leave once you begin to watch. View is presented differently than displayed by most school books on this *subject*. My only fault for this movie is it was photographed in black and white; wished it had been in color ... wow !
Perturbed text
gnreat historical movie, will not allow a viewer to leave once you begin to watch. View is presented differently than displayed by most school books on this *sSbject*. My only fault for this movie is it was photographed in black and white; wished it had been in color ... wow !

Table 4: Example of a successful Deep Word Bug attack on a vanilla Bert model, fine tuned over the IMDB dataset. The original label is 'Positive sentiment'. Perturbations, marked in italic font, change the classification to 'Negative sentiment'. VUB and VIB classifiers are far less susceptible to these perturbations, as shown in Table 2.

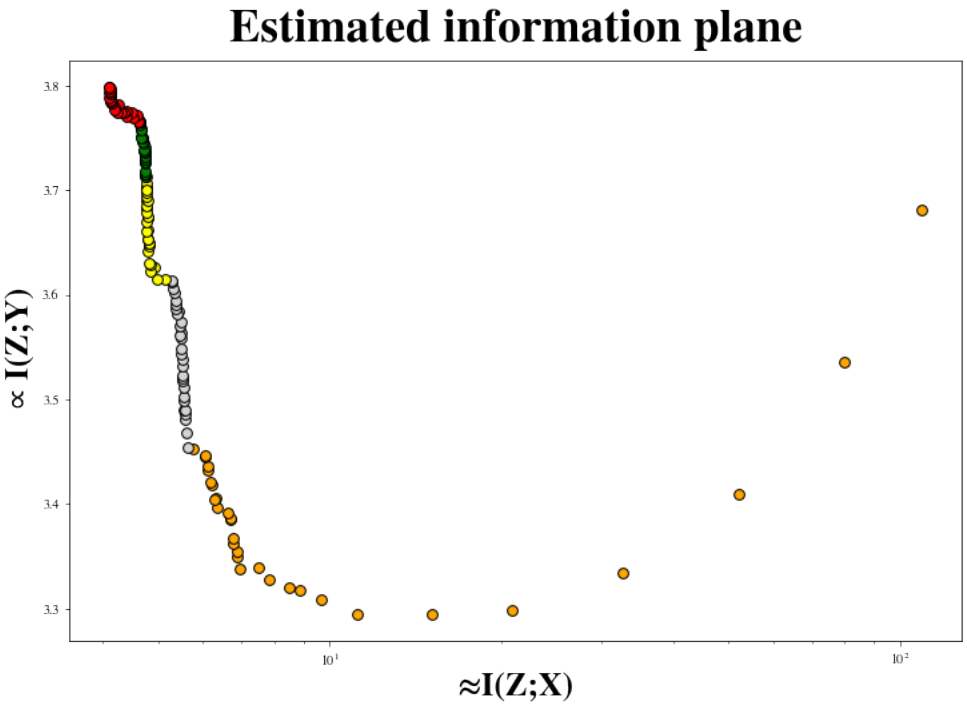

Figure 7: Estimated information plane metrics per epoch for VUB trained on IMDB with $\beta = 0.001$. $I(X;Z)$ is approximated by $H(R) - H(Z|X)$ and $\frac{1}{CE(Y;\hat{Y})}$ is used as an analog for $I(Z;Y)$. The epochs have been grouped and color-coded in intervals of 30 epochs in the order: Orange (0-30), gray (30-60), yellow (60-90), green (90-120) and red (120-150). We notice recurring patterns of distortion reduction followed by rate increase, resembling the ERM and representation compression stages described by Shwartz-Ziv and Tishby (2017).

information plane behavior is not limited to simplified scenarios, but is a possible characteristic of the learning process in more challenging tasks as well.