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

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

9 **1 Introduction**

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

30 observed data is known (Tishby, Pereira, and Bialek 1999; Chechik et al. 2003; Painsky and Tishby
31 2017). Nonetheless, when deriving the variational IB objectives, previous research (Alemi et al.
32 2017; Fischer 2020; Cheng et al. 2020) relax the problem by considering the learned representation
33 as the only optimized parameter, when in practice the classifier is also optimized. We derive a new
34 upper bound for the IB objective, and a subsequent variational approximation, by removing the
35 relaxation. We show that our bound is tighter than previous bounds, and that our proposed loss
36 function is a tighter variational approximation, when considering $p(y|x)$ as part of the optimization.
37 We believe our new derivation is a better adaptation of the IB for supervised tasks, and show empirical
38 evidence of improved performance across several challenging tasks over different modalities. We
39 utilize previous studies on variational representation learning and regularization (Alemi et al. 2018;
40 Pereyra et al. 2017; Achille and Soatto 2018) to interpret our findings, and conclude that our proposed
41 derivation applies regularization over the variational classifier, preventing it from overfitting the
42 learned representations and thus enabling greater MI between learned representation and the target Y .
43 The reader is encouraged to refer to the preliminaries provided in Appendix A before proceeding.

44 2 Related work

45 2.1 Deterministic Information Bottleneck

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

70 **2.2 Variational Information Bottleneck**

71 Alemi et al. (2017) introduced the Variational Information Bottleneck (VIB) - a variational approx-
72 imation for an upper bound to the IB objective for DNN optimization. Bounds for $I(X; Z)$ and
73 $I(Z; Y)$ are derived from the non negativity of KL divergence, and are used to form an upper bound
74 for the IB objective. A variational upper bound is derived by replacing intractable distributions
75 with variational approximations. Using the ‘reparameterization trick’ (Kingma and Welling 2014)
76 a discrete empirical estimation of the variational upper bound is then used as a loss function for
77 classifier DNN optimization. The subsequent loss function is equivalent to the β -autoencoder loss
78 (Higgins et al. 2017). VIB was evaluated over image classification tasks, and displayed substantial
79 improvements in robustness to adversarial attacks, while causing a slight reduction in test set accu-
80 racy, comparing to equivalent deterministic models. (Achille and Soatto 2018) extended VIB with a
81 total correlation term, designed to increase latent disentanglement. Fischer (2020) proposed an IB
82 based loss function named Conditional Entropy Bottleneck (CEB), in which the conditional mutual
83 information of X and Z given Y is minimized, instead of the unconditional mutual information.
84 The CEB loss, $L_{CEB} = \min_Z I(X; Z|Y) - \gamma I(Y; Z)$, is designed to minimize all information in
85 Z that is not relevant to the downstream task Y , by conditioning over Y . CEB is equivalent to IB
86 for $\gamma = \beta - 1$ following the chain rule of mutual information (Cover 1999) and the IB Markov
87 chain, as established in Appendix B. Similarly to VIB, a variational approximation for CEB was
88 proposed as $L_{VCEB} = \mathbb{E}_{x,y} \log(e(z|x)) - \mathbb{E}_{x,y} \log(b(z|y)) - \gamma \mathbb{E}_{x,y} \log(c(y|z))$ and tested over
89 the FMNIST (Xiao, Rasul, and Vollgraf 2017) and CIFAR10 (Krizhevsky 2009) datasets. Fischer
90 (2020) shows that VIB is a special case of VCEB, where rate is approximated by the variational
91 expression $\mathbb{E}_{x,y} \log(e(z|x)) - \mathbb{E}_{x,y} \log(b(z|y))$ instead of $\mathbb{E}_{x,y} \log(e(z|x)) - \mathbb{E}_{x,y} \log(r(z))$. Geiger
92 and Fischer (2020) investigated whether VCEB is a tighter variational approximation to IB than VIB,
93 and concluded that no ordering can be established in the general case, noting that any empirical
94 improvement VCEB exhibits over VIB is not due to a tighter variational bound on the IB, but rather
95 of VCEB being more amenable to optimization, or simply a successful loss function in its own
96 regard. Cheng et al. (2020) proposed CLUB, an upper bound based MI estimator, that empirically
97 outperformed the popular MINE estimator (Belghazi et al. 2018). Since CLUB is an upper bound for
98 MI, it was evaluated as a replacement to the upper bound for the IB rate term, $I(X; Z)$, proposed in
99 VIB (Alemi et al. 2017). CLUB based VIB was tested over the MNIST dataset (Deng 2012), resulting
100 in a slight improvement in accuracy compared to VIB, without reporting adversarial robustness. We
101 note that CLUB does not establish a tighter bound on the VIB rate term, and subsequently on the IB
102 objective. We also note, that our current work derives a tighter bound on IB through the IB distortion
103 term, $I(Z; Y)$, and that combining our suggested method with a CLUB bound on rate is an interesting
104 avenue for future work.

105 **2.3 Information theoretic regularization**

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

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

122 3 From VIB to VUB

123 As elaborated in Section 2.1, the IB objective, $\mathcal{L}_{IB} = I(X; Z) - \beta I(Z; Y)$, is computed over the joint
 124 distribution $p^*(x, y, z)$. When $p^*(x, y)$ is given, this expression is optimized over the distribution
 125 $p(z|x)$, as proposed by Tishby, Pereira, and Bialek (1999): $Z : \min_{p(z|x)} I(X; Z)$ subject to $I(Z; Y) \geq D$.
 126 However, adapting IB to supervised tasks admits the learned classifier as a new RV to the optimization
 127 problem. Geiger and Fischer (2020) suggested the Markov chain $Y \leftrightarrow X \leftrightarrow Z \leftrightarrow \tilde{Y}$ for supervised
 128 IB, distinguishing between the real unknown RV Y , and the learned classifier \tilde{Y} . Following this logic,
 129 we argue that supervised IB optimization should be defined as: $Z, \tilde{Y} : \min_{p(z|x), p(\tilde{y}|z)} I(X; Z)$ subject to
 130 $I(Z; \tilde{Y}) \geq D$.

131 The VIB loss (Alemi et al. 2017) consists of a cross entropy (*CE*) term, and a beta modulated KL
 132 regularization term, as in β -VAE loss (Higgins et al. 2017). The KL term is derived from a bound
 133 on the IB rate term $I(X; Z)$, while the *CE* term is derived from a bound on the IB distortion term
 134 $I(Z; Y) = H(Y) - H(Y|Z)$. During the derivation of the CE term, a relaxation is performed such
 135 that the term $H(Y)$ is assumed constant, and hence ignored, while the term $-H(Y|Z)$ is derived
 136 as CE between true and learned supervisor distributions $p^*(y), p(\tilde{y})$. We derive a new upper bound
 137 for the IB objective by not omitting $H(Y)$ from the distortion term. Subsequently, the variational
 138 approximation for our proposed bound is tighter when taking into account that the optimization
 139 process is done over $p(\tilde{y}|z)$ as well as $p(z|x)$. This modification attains a tighter variational bound
 140 on the IB objective for any Y with positive entropy, and a tighter empirical bound for all Y .

141 3.1 IB upper bound

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

144 Consider $I(Z; X)$:

$$I(Z; X) = \int \int p^*(x, z) \log(p^*(z|x)) dx dz - \int p^*(z) \log(p^*(z)) dz \quad (1)$$

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

$$\int p^*(z) \log(p^*(z)) dz \geq \int p^*(z) \log(r(z)) dz \quad (2)$$

146 And so, by Equation 2:

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

¹⁴⁷ Consider $I(Z; Y)$:

¹⁴⁸ For any probability distribution c we have that $D_{KL}(p^*(y|z)||c(y|z)) \geq 0$, it follows that:

$$\int p^*(y|z)\log(p^*(y|z)) dy \geq \int p^*(y|z)\log(c(y|z)) dy \quad (4)$$

¹⁴⁹ And so, by Equation 4:

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

¹⁵⁰ We now diverge from the original VIB derivation by replacing $H_{p^*}(Y)$ with $H_c(Y|Z)$ instead of
¹⁵¹ omitting it. In addition, we limit the new term to make sure that the inequality holds.

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

¹⁵² We develop the first term in 6 using the IB Markov chain $Z \leftrightarrow X \leftrightarrow Y$ and total probability:

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

¹⁵³ 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
¹⁵⁴ IB objective:

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

¹⁵⁶ 3.2 Variational approximation

¹⁵⁷ 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} :

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

159 **3.3 Empirical estimation**

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

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

169 **3.4 Intuition**

170 Barber and Agakov (2003) define the following lower bound and identity:

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

171 All variational IB derivations (Alemi et al. 2017; Fischer 2020; Cheng et al. 2020) omit the entropy
 172 term $H_{p^*}(Y)$, and utilize the bound to optimize the negative IB distortion term. The following
 173 inequality holds for all distributions Y with non negative entropy:

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

174 We have that the true MI $I(Z; Y)$ is squeezed by the Barber-Agakov MI $\tilde{I}(Z; Y)$ from below, which
 175 is in turn squeezed by the VUB distortion term. It is possible that in the extreme case, increasing the
 176 entropy term $H_c(Y|Z)$ might skew the true mutual information $I(Z; Y)$ up. Since the real RV Y is
 177 constant, this increase can only be caused by learning a better representation Z . Figure 1 illustrates
 178 the possible effect of an increase in variational entropy: The left hand diagram suggest a model
 179 with low variational entropy, and a low variational mutual information, and the right hand diagram
 180 suggest a model with increased variational entropy. The cross entropy term pushes $I(Z; \hat{Y})$ up, and
 181 the real mutual information $I(Z; Y)$ increases as a result of the Barber-Agakov inequality (Barber
 182 and Agakov 2003).

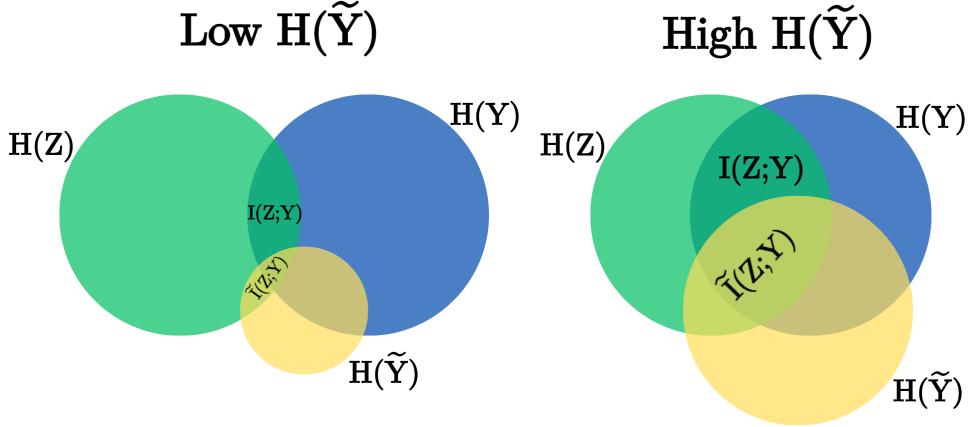


Figure 1: Venn diagrams illustrating the possible increase in true mutual information as a result of increased variational entropy. The right diagram features higher variational entropy, and higher variational mutual information that induce higher real mutual information as a result of the Barber-Agakov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

183 4 Experiments

184 We follow the experimental setup proposed by Alemi et al. (2017), extending it to NLP tasks as
 185 well. We trained image classification models on the ImageNet 2012 dataset (Deng et al. 2009), and
 186 text classification on the IMDB sentiment analysis dataset (Maas et al. 2011). For each dataset,
 187 we compared a competitive Vanilla model with a VIB and a VUB model trained with beta values
 188 of $\beta = 10^{-i}$ for $i \in \{1, 2, 3\}$. Each model was trained and evaluated 5 times per β value, with
 189 consistent performance and statistical significance shown by a Wilcoxon rank sum test. Each model
 190 was evaluated using test set accuracy, and robustness to various adversarial attacks. For image
 191 classification, we employed the untargeted Fast Gradient Sign (FGS) attack (Goodfellow, Shlens, and
 192 Szegedy 2015) as well as the targeted CW L_2 attack (Carlini and Wagner 2017), (Kaiwen 2018). For
 193 text classification, we used the untargeted Deep Word Bug attack (Gao et al. 2018), (Morris et al.
 194 2020) as well as the untargeted PWWS attack (Ren et al. 2019). Elaboration on the experimental setup,
 195 results and further insights from the experiments are available in Appendix C. Code to reconstruct
 196 the experiments is provided to the supplementary materials of this paper.

197 4.1 Image classification

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

β	Val \uparrow	FGS $\downarrow_{\epsilon=0.1}$	FGS $\downarrow_{\epsilon=0.5}$	CW \uparrow
Vanilla model				
-	77.2%	68.9%	67.7%	788
VIB models				
10^{-3}	73.7% $\pm .1\%$	59.5% $\pm .2\%$	63.9% $\pm .2\%$	3917 ± 291
10^{-2}	72.8% $\pm .1\%$	53.5% $\pm .2\%$	62.0% $\pm .1\%$	3318 ± 293
10^{-1}	72.1% $\pm .01\%$	58.4% $\pm .1\%$	62.0% $\pm .1\%$	3318 ± 293
VUB models				
10^{-3}	75.5% $\pm .03\%$	62.8% $\pm .1\%$	66.4% $\pm .1\%$	2666 ± 140
10^{-2}	75.0% $\pm .05\%$	57.6% $\pm .2\%$	64.3% $\pm .1\%$	1564 ± 218
10^{-1}	74.8% $\pm .09\%$	57.9% $\pm .5\%$	64.8% $\pm .5\%$	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 \uparrow), second and third columns are the % of successful FGS attacks at $\epsilon = 0.1, 0.5$ (lower is better \downarrow), and the fourth column is the average L_2 distance for a successful Carlini Wagner L_2 targeted attack (higher is better \uparrow). VUB attains significantly higher accuracy over unseen data in all settings, while preserving competitive robustness to adversarial attacks.

207 4.2 Text classification

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

215 5 Discussion

216 The IB is a private case of rate distortion, and was initially designed as an optimization problem over
 217 compressed representations. Adapting the IB objective for supervised tasks results in optimization
 218 of a classifier distribution as well, and requires a reformulation of the initial optimization problem
 219 to include both representation and discriminator. Following this logic, assuming a constant $H(Y)$
 220 relaxes the problem, and lifting this assumption lead us to derive a tighter variational bound over
 221 the optimized objective. When used as a loss function, our proposed bound produces significantly
 222 better classification accuracy, with equivalent or superior robustness to adversarial attacks, over high
 223 dimensional tasks of different modalities, with high statistical significance. On a practical level,
 224 the conditional entropy term that follows from our proposed derivation provides strong classifier
 225 regularization, as shown in (Pereyra et al. 2017). This type of regularization is a possible remedy
 226 to the imbalances 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 inherently to VIB also, as it is equivalent to a beta regulated ELBO loss. In addition, we propose an possible intuitive explanation to the effects of conditional entropy regularization on the quality of learned representation, 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.

- 288 Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale
289 hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,
290 248–255. Ieee.
- 291 Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE
292 Signal Processing Magazine*, 29(6): 141–142.
- 293 Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional
294 Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North
295 American Chapter of the Association for Computational Linguistics: Human Language Technologies,
296 Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota.
- 297 Fischer, I. 2020. The Conditional Entropy Bottleneck. *Entropy*, 22(9): 999.
- 298 Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-box generation of adversarial text
299 sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*,
300 50–56. IEEE.
- 301 Geiger, B.; and Fischer, I. 2020. A Comparison of Variational Bounds for the Information Bottleneck
302 Functional. *Entropy*.
- 303 Goldfeld, Z.; and Polyanskiy, Y. 2020. The Information Bottleneck Problem and its Applications in
304 Machine Learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 19–38.
- 305 Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples.
306 In *ICLR (Poster)*.
- 307 Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner,
308 A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In
309 *ICLR (Poster)*.
- 310 Kaiwen. 2018. pytorch-cw2. GitHub repository.
- 311 Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Con-
312 ference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference
313 Track Proceedings*.
- 314 Kingma, D. P.; and Welling, M. 2019. An Introduction to Variational Autoencoders. *Foundations
315 and Trends in Machine Learning*, 12(4): 307–392.
- 316 Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. 32–33.
- 317 Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word
318 vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for
319 computational linguistics: Human language technologies*, 142–150.
- 320 Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for
321 Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the
322 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*,
323 119–126.
- 324 Painsky, A.; and Tishby, N. 2017. Gaussian Lower Bound for the Information Bottleneck Limit. *J.
325 Mach. Learn. Res.*, 18: 213:1–213:29.

- 326 Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; and Hinton, G. E. 2017. Regularizing Neural Net-
327 works by Penalizing Confident Output Distributions. In *Proceedings of the International Conference*
328 *on Learning Representations*. OpenReview.net.
- 329 Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating natural language adversarial examples
330 through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the*
331 *association for computational linguistics*, 1085–1097.
- 332 Saxe, A. M.; Bansal, Y.; Dapello, J.; and Advani, M. 2018. On the information bottleneck theory of
333 deep learning.
- 334 Shannon, C. E. 1959. Coding Theorems for a Discrete Source With a Fidelity Criterion. In *IRE*
335 *National Convention*.
- 336 Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via
337 Information. 19 pages, 8 figures, arXiv:arXiv:1703.00810.
- 338 Slonim, N. 2002. *The information bottleneck: Theory and applications*. Ph.D. thesis, Hebrew
339 University of Jerusalem Jerusalem, Israel.
- 340 Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception
341 architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and*
342 *pattern recognition*, 2818–2826.
- 343 Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The Information Bottleneck Method. In *The*
344 *37th annual Allerton Conference on Communication, Control, and Computing*. Hebrew University,
345 Jerusalem 91904, Israel.
- 346 Tishby, N.; and Zaslavsky, N. 2015. Deep Learning and the Information Bottleneck Principle.
347 arXiv:1503.02406.
- 348 Vapnik, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc. ISBN
349 0-387-94559-8.
- 350 Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking
351 Machine Learning Algorithms. Cite arxiv:1708.07747Comment: Dataset is freely available at
352 <https://github.com/zalandoresearch/fashion-mnist> Benchmark is available at <http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/>.

354 **Appendix A - Preliminaries**

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

359 Let X, Y be two observed random variables with an unknown joint distribution $p^*(x, y)$ and marginals
 360 $p^*(x), p^*(y)$. We can attempt to approximate these distributions using a model p_θ with parameters
 361 θ , such that for generative tasks $p_\theta(x) \approx p^*(x)$, and for discriminative tasks $p_\theta(y|x) \approx p^*(y|x)$,
 362 using a dataset $\mathcal{S} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ to fit our model. One can also assume the existence
 363 of an additional unobserved RV $Z \sim p^*(z)$ that influences or generates the observed RVs X, Y .
 364 Since Z is unobserved, it is absent from the dataset \mathcal{S} , and so cannot be modeled directly. Denote
 365 $p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz = \int p_\theta(x, z)dz$ the marginal, $p_\theta(z)$ the prior as it is not conditioned over
 366 any other RV, and $p_\theta(z|x)$ the posterior following Bayes' rule.

367 When modeling an unobserved variable of an unknown distribution, we encounter a problem as the
 368 marginal $p_\theta(x) = \int p_\theta(x, z)dz$ doesn't have an analytic solution. This intractability can be overcome
 369 by choosing some tractable parametric variational distribution $q_\phi(z|x)$ to approximate the posterior
 370 $p_\theta(z|x)$, 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}
 371 (Kingma and Welling 2019).

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

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

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

378 • A learning machine capable of implementing a predefined set of functions, $f(x, \theta) : \mathbb{R}^d \times$
 379 $\Theta \mapsto \mathbb{R}$, where Θ is a set of parameters.

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

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

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

393

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

394

395 Appendix B - Related work elaboration

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

398 The Information Plane

399 As mentioned in Section 2.1, the solution to the IB objective, $\mathcal{L}_{IB} = I(X; Z) - \beta I(Z; Y)$, depends
 400 on the Lagrange multiplier β . Hence, the IB objective has no one unique solution, and can thus
 401 be plotted as a function of β and of Z 's cardinality, over a Cartesian system composed of the axes
 402 $I(X; Z)$ (rate) and $I(Z; Y)$ (distortion). When plotted as a function of β the IB functional is called
 403 the 'information curve' and its Cartesian system the 'information plane' (Tishby, Pereira, and Bialek
 404 1999), as illustrated in Figure 2. When β approaches 0 the distortion term is nullified and we learn a
 405 representation that has no information over the down stream task, and maximal compression (such a
 406 representation may be a null vector). When β approaches ∞ we learn a representation that has the
 407 maximal possible information over the downstream task, but that also holds the maximal information
 408 over the training data, hence overfitted. The region above the information curve is unreachable by
 409 any possible representation. The different bifurcation of the information curve, illustrated in Figure 2,
 410 correspond to the different possible cardinalities of the compressed representation.

411 Fixing a Broken Elbo

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

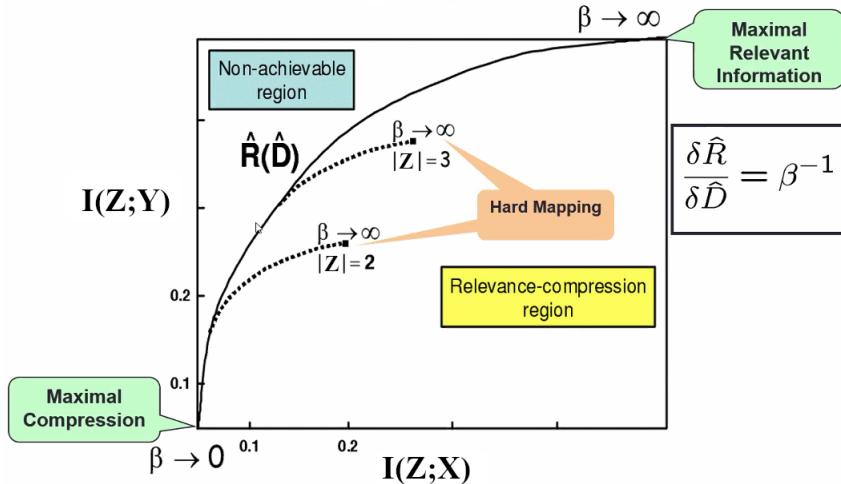


Figure 2: The information plane and curve: rate-distortion ratio over β . At $\beta = 0$ the representation is compressed but uninformative (maximal compression), at $\beta \rightarrow \infty$ the representation is informative but potentially overfitted (maximal information). Taken from (Slonim 2002).

417 compensated for by an overfitted decoder. Mutual information is used to measure representation quality:
 418 $I_e(X; Z) \equiv \int \int p_e(x, z) \log \left(\frac{p_e(x, z)}{p^*(x)p_e(z)} \right) dx dz$, for a stochastic encoder e . I_e can be bounded from
 419 both sides as follows: $H_{p^*}(X) - I_d(X; \hat{X}) \leq I_e(X; Z) \leq D_{KL}(q_\phi(z|x) || p_\theta(z))$, where $H_{p^*}(X)$
 420 is the true data entropy, I_d is the variational decoder's distortion. It is observed, that the KL term does
 421 not depend on the variational decoder distribution, and can be minimized regardless of reconstruction
 422 quality. Equivalently, good reconstruction does not directly depend on good representation. Figure 3
 423 is suggested as an extension to the information plane (Tishby, Pereira, and Bialek 1999), by replacing
 424 real rate and distortion with their variational approximations. The new plane is divided into three
 425 sub planes: (1) Infeasible: This is the IB theoretical limit (As per Figure 2); (2) Feasible: Attainable
 426 given an infinite model family, and complete variety of: $e(z|x)$, $d(x|z)$ and $p_\theta(z)$; (3) Realizable:
 427 Attainable given a finite parametric and tractable variational family.
 428 The black diagonal line at the lower left of Figure 3 satisfies $H_{p^*}(X) - I_d(X; \hat{X}) = D_{KL}(q_\phi(z|x) || p_\theta(z))$, resulting in tight variational bounds on the mutual information: $R \leq I_e \leq R$.
 429 (Alemi et al. 2018) notes that a common degeneration of representations in VAEs is a decrease in rate
 430 and distortion over the training data, together with an increase in both over unseen data. This is caused
 431 by overpowered decoders that overfit an uninformative learned representation. Low ELBO loss can be
 432 attained as both the distortion and rate terms are minimized during training. $D_{KL}(q_\phi(z|x) || p_\theta(z))$
 433 approaches 0 iff $e(z|x) \rightarrow p_\theta(z)$. In this case, $e(z|x)$ is close to independence of x , and the latent
 434 representation fails to encode information about the input. However, a suitably powerful decoder
 435 could non the less learn to overfit encoded traces of the training examples, and reach a low distortion
 436 score during optimization. In the current study, we extend this theoretical framework to explain the
 437 advancements of our proposed loss function.

439 IB theory of deep learning

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

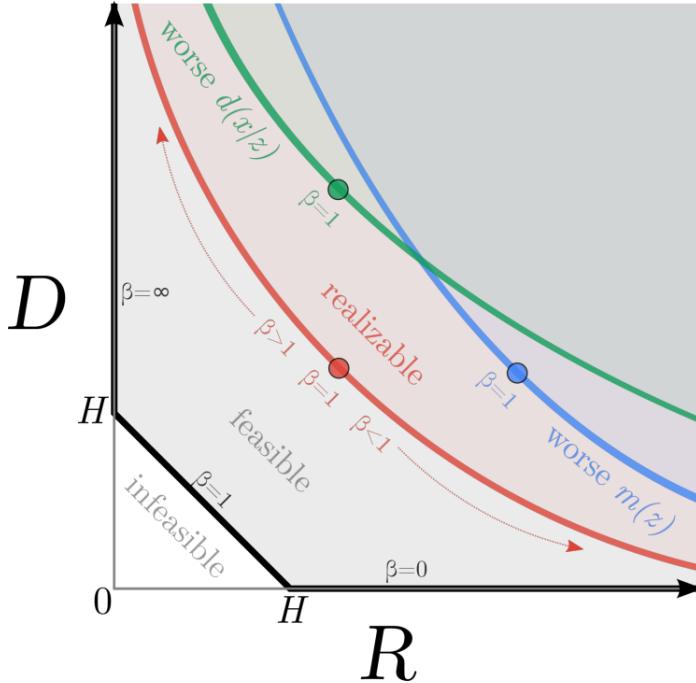


Figure 3: Phase diagram, a proposed information plane interpretation of VAEs. Axes are variational rate and distortion. The IB theoretical limit is extended by an additional limit induced by the constraint of a finite parametric variational family. Once a family is chosen, we seek to learn an optimal 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).

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

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

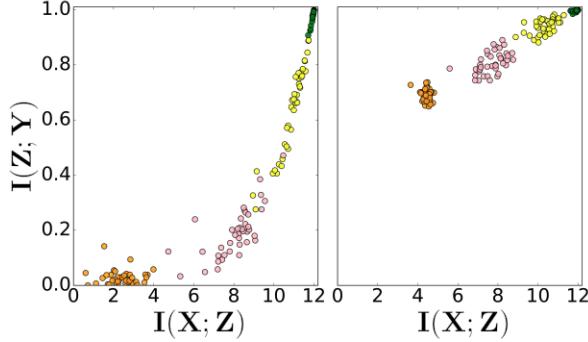


Figure 4: Information plane scatters of different DNN layers (colors) in 50 randomized networks. From Shwartz-Ziv and Tishby (2017). Left are initial weights, Right are at 400 epochs. Our study reproduced similar yet repetitive behavior on complicated high dimensional tasks, as elaborated in Section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

- 463 discrepancies are caused because mutual information in deterministic DNNs is either infinite or step
 464 like, because of mutual information's invariance to invertible transformations and because of the
 465 absence of a decision function in the objective.
 466 When examining the information plane behavior in the current study, we notice recurring patterns of
 467 distortion reduction followed by rate increase, resembling the ERM and representation compression
 468 stages described by Shwartz-Ziv and Tishby (2017), as elaborated in Appendix 5.

469 **Conditional Entropy Bottleneck**

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

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

473 **Appendix C - Experiments elaboration**

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

479 activated linear layers of the same dimensions as the pre-trained model's logits (2048 for image
 480 and 768 for text classification), followed by reparameterization and a final softmax activated FC
 481 layer. Learning rate was 10^{-4} and decaying exponentially with a factor of 0.97 every two epochs.
 482 Batch sizes were 32 for ImageNet and 16 for IMDB. All models were trained using an Nvidia
 483 RTX3080 GPU with approximately 1-2 days per a single experiment run. Beta values of $\beta = 10^{-i}$
 484 for $i \in \{1, 2, 3\}$ were tested, and we used a single forward pass per sample for inference, since
 485 previous studies indicated that these are the best range and sample rate for VIB (Alemi et al. 2017,
 486 2018). Each model was trained and evaluated 5 times per β value, with consistent performance.
 487 Statistical significance was demonstrated in all comparisons using the Wilcoxon rank sum test as
 488 follows: For each compared metric a sorted vector of results was prepared, where each entry featured
 489 the attained result in each of the 5 i.i.d. experiments per algorithm, and a boolean indicator value
 490 for the algorithm type. All metrics compared attained a p-value of less than 0.05. For example:
 491 $r = ((0.94, 1) (0.935, 1) (0.93, 1) (0.93, 1) (0.925, 1) (0.92, 0) (0.915, 0) (0.915, 0) (0.91, 0) (0.89, 0))^{-1}$
 492 be a sorted vector of (test accuracy, algorithm) tuples, 1 being VUB, 0 VIB. We computed the
 493 rank-sum as follows:

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

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

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

```

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

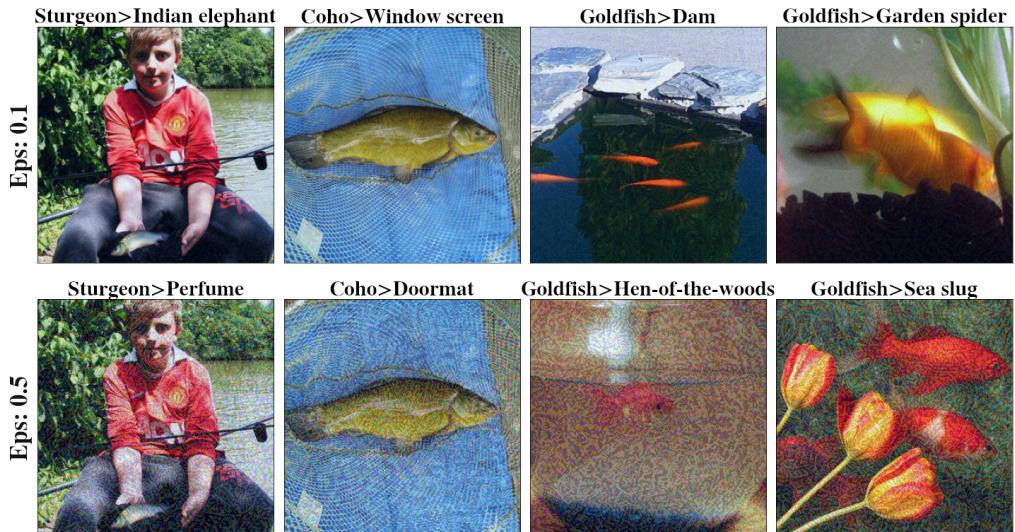
500 **Image classification**

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

507 **Text classification**

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

Untargeted FGS attacks for VIB $\beta=0.01$



Untargeted FGS attacks for VUB $\beta=0.01$

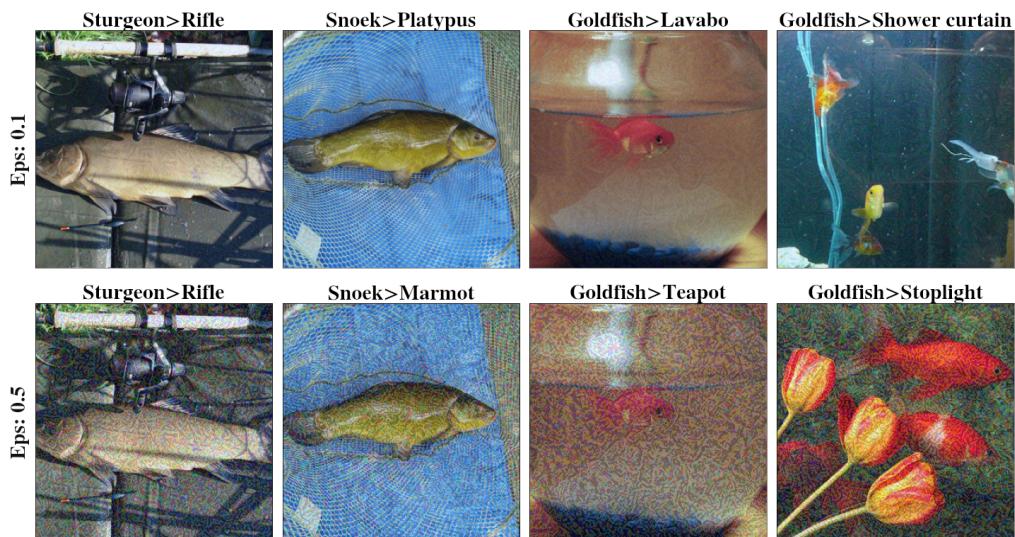


Figure 5: Successful untargeted FGS attack examples. Images are perturbations of previously successfully classified instances from the ImageNet validation set. Perturbation magnitude is determined by the parameter ϵ shown on the left, the higher, the more perturbed. Original and wrongly assigned labels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

Targeted CW attacks for VIB $\beta=0.01$. Target: Soccer ball



Targeted CW attacks for VUB $\beta=0.01$. Target: Soccer ball



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 <i>astounding</i> given the production 's austere locales .
Perturbed text
the acting , costumes , music , cinematography and sound are all <i>dumbfounding</i> 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.

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

Original text
<i>great</i> 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 <i>subject</i> . My only fault for this movie is it was photographed in black and white; wished it had been in color ... wow !
Perturbed text
<i>gnreat</i> 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 <i>sSbject</i> . 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.

Estimated information plane

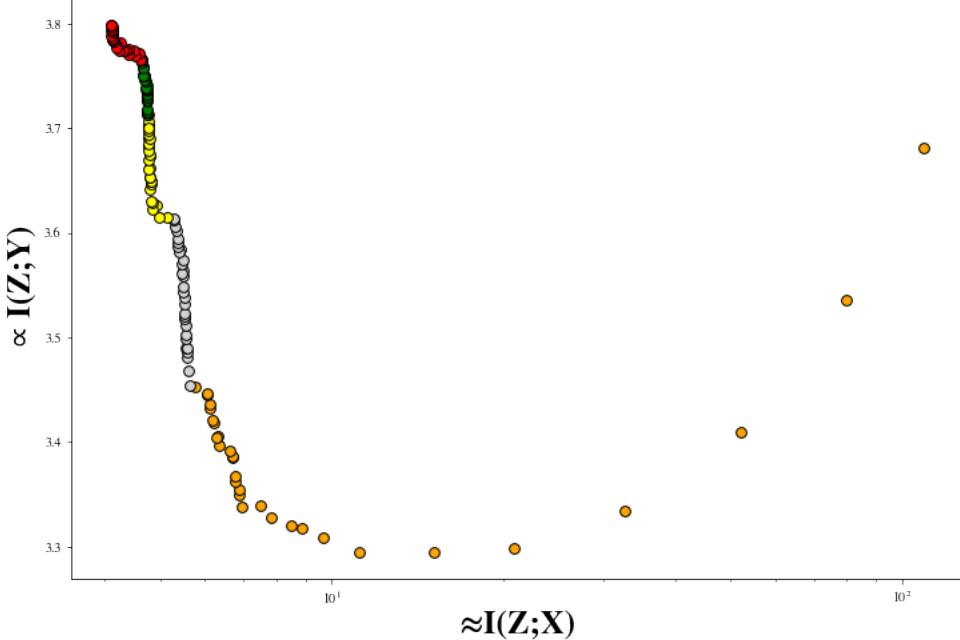


Figure 7: Estimated information plane metrics per epoch for VUB trained on IMDB with $\beta = 0.001$. $I(X; Z)$ is approximated by $H(R) - H(Z|X)$ and $\frac{1}{CE(Y; \hat{Y})}$ is used as an analog for $I(Z; Y)$. The epochs have been grouped and color-coded in intervals of 30 epochs in the order: Orange (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).

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