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 introduces 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 methods to approximate upper bounds on 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 this relaxation. We show that our bound is tighter than previous ones, 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, thus enabling greater MI between learned representation and 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 this 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 z to belong to different clusters, given the joint distribution of the input and target tasks $p^*(x, y)$ (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 objective as a loss function 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 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 inflicting a slight reduction in test set accuracy, when compared 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, $L_{CEB} = \underset{Z}{min} I(X; Z|Y) - \gamma I(Y; Z)$, 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) showed 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). CLUB 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 to the loss function, $-\gamma \cdot H\left(p_\theta(y|x)\right)$, 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) extended 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 KL divergence between variational representation and marginal. 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. Alemi et al. (2018) also showed that VAEs are susceptible to learn low quality representations, as the KL regularization term in the ELBO loss might induce very uninformative representations, provided there's a strong enough decoder to compensate for bad embeddings by overfitting them. 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 $Y \leftrightarrow X \leftrightarrow Z \leftrightarrow \tilde{Y}$ for supervised IB, distinguishing between the true unknown RV Y, and the learned classifier $\tilde{Y}$. 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 VIB 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 objective 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 $D_{KL}\left(p^*(y|z)\middle|\middle|c(y|z)\right) \geq 0$, 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 Equation 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 $Y \leftrightarrow X \leftrightarrow Z \leftrightarrow \tilde{Y}$. We join Equation 3 with Equation 7 to establish a new upper bound for the IB objective:

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

3.2 Variational approximation

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

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

3.3 Empirical estimation

The true and possibly continuous distribution $p^*(x, y) = p^*(y|x)p^*(x)$ can be estimated by Monte Carlo sampling from a discrete dataset $\mathcal{S}$. 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 true 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 $r(z)$.

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

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):

$$I(Z;Y) \geq \int\int p^*(y,z)log\left(c_{\tilde{y}|z}(y|z)\right)dydz + H_{p^*}(Y) \equiv \tilde{I}(Z;Y) \tag{11}$$

The following inequality holds for all distributions of Y with non negative entropy, assuming $H_{p^*}(Y) \gtrsim H_c(\tilde{Y}|Z)$, as should follow from a well fitted model:

$$H_{p^*}(Y) \geq I(Z;Y) \geq \tilde{I}(Z;Y) \gtrsim \int\int p^*(y,z)log\left(c_{\tilde{y}|z}(y|z)\right)dydz + H_c(\tilde{Y}|Z) \tag{12}$$

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. Previous variational IB derivations (Alemi et al. 2017; Fischer 2020; Cheng et al. 2020) omit the entropy term $H_{p^*}(Y)$, and derive the term $\int\int p^*(y,z)log\left(c_{\tilde{y}|z}(y|z)\right)dydz$ into cross entropy, which is minimized during optimization. Assuming that cross entropy is indeed minimized, intuition suggests that increasing the entropy term $H_c(\tilde{Y}|Z)$ as close as possible to $H_{p^*}(Y)$ will squeeze the true MI $I(Z;Y)$ closer to its theoretical limit. Since the optimization cannot change the entropy of the true RV Y, the potential increase in $I(Z;Y)$ can only be caused by learning a more informative representation Z. Figure 1 illustrates the possible effects of an increase in variational entropy: The left hand diagram suggests a model with low variational entropy, and low true MI, while the right hand diagram suggests a model with high variational entropy, and a higher true MI. Reduction in cross entropy increases $\tilde{I}(Z;Y)$, and the true mutual information $I(Z;Y)$ increases 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

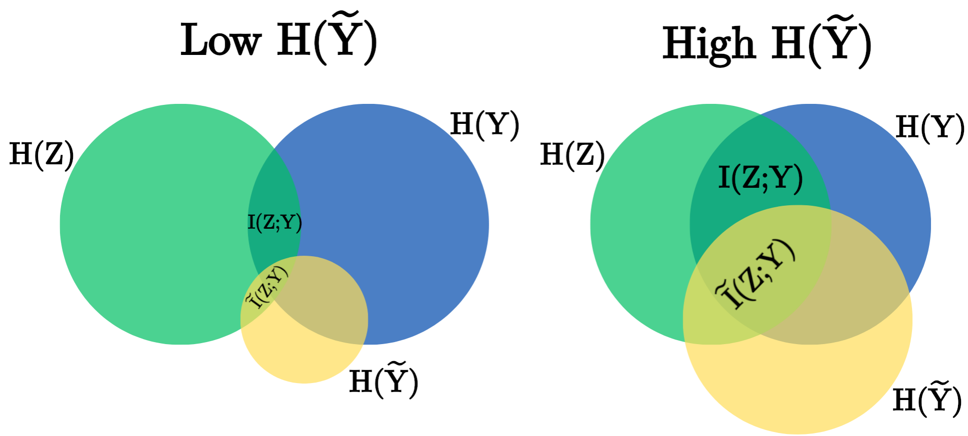

Figure 1: Venn diagrams illustrating a possible increase in true mutual information as a result of increased variational entropy. The right diagram features higher variational entropy, and higher variational mutual information, that induce higher true mutual information as a result of the Barber-Agakov inequality (Barber and Agakov 2003).We disregard change in the value of Z and the ratio between Z and Y to simplify the figure, and focus on the possible relation between true MI and variational entropy.

text classification models 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 on a Wilcoxon rank sum test.

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 Figures 3,4 in Appendix C. In this modality, VUB

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** $\pm$291
10^{-2}	72.8% $\pm$.1%	**53.5%** $\pm$.2%	**62.0%** $\pm$.1%	**3318** $\pm$293
10^{-1}	72.1% $\pm$.01%	58.4% $\pm$.1%	**62.0%** $\pm$.1%	**3318** $\pm$293
		VUB models		
10^{-3}	**75.5%** $\pm$.03%	62.8% $\pm$.1%	66.4% $\pm$.1%	2666 $\pm$140
10^{-2}	**75.0%** $\pm$.05%	57.6% $\pm$.2%	64.3% $\pm$.1%	1564 $\pm$218
10^{-1}	**74.8%** $\pm$0.09%	**57.9%** $\pm$.5%	64.8% $\pm$.5%	**3575** $\pm$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.

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 to optimize 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 problem to include both representation and discriminator. 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 imbalances inherit to the ELBO loss function, and correlatively to VIB, as described by Alemi et al. (2018). In addition, we propose a new intuition for conditional entropy regularization, by showing that in the extreme cases, high variational entropy can squeeze the true mutual information $I(Z;Y)$ higher, implying better representations learned. While other advancements have been done in recent years, (Fischer 2020;

β	**Test↑**	**DWB↓**	**PWWS↓**
		Vanilla model	
-	93.0%	54.3%	100%
		VIB models	
10^{-3}	91.0% ±1.0%	35.1% ±4.4%	41.6% ±6.6%
10^{-2}	90.8% ±0.5%	41.0% ±4.8%	62.9% ±14.3%
10^{-1}	**89.4%** ±.9%	**90.0%** ±8.0%	**99.1%** ±0.9%
		VUB models	
10^{-3}	**93.2%** ±.5%	**27.5%** ±2.0%	**28.4%** ±1.3%
10^{-2}	**92.6%** ±.8%	**30.8%** ±2.0%	**50.0%** ±4.8%
10^{-1}	89.2% ±2.0%	99.2% ±0.5%	100% ±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 ↑), second is % of successful Deep Word Bug attacks (lower is better ↓) and third is % of successful PWWS attacks (lower is better ↓). 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 $\beta = 10^{-3}$.

Cheng et al. 2020), they focus on modifications to the IB rate term, and none show a tighter bound than VIB. In contrast, our work derives a provable tighter bound by modifying the distortion term.

Applying variational IB as an objective to supervised learning relies 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 approximated to a desired level of accuracy. Our improved empirical results, induced by a tighter bound, suggest better data modeling, and hence strengthen the cause for variational IB as an objective for classifier DNNs. A possible counter argument is, that the improvements in adversarial robustness of variational IB DNNs is an artifact of their latent geometry, rather than the quality of their learned representations. As the KL regularization induces a smoother latent space, and the factorized reparameterization promotes disentanglement (Chen et al. 2018), minor perturbations might not cause a significant change in latent semantics, possibly making the models more robust to attacks. Nonetheless, VUB is presented as a tractable and tighter upper bound on the IB objective, 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 an increase.

This study opens many opportunities for further research: Further improvements to the upper bound, including combining VUB with the CLUB bound on rate (Cheng et al. 2020); Applying VUB in self-supervised learning, and in particular measuring whether representations learned with VUB capture better semantics than representations learned with non IB inspired loss functions; Finally, experiments with a full covariance matrix VUB, and studying the effects of latent geometry 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

Notation

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 $p(x)$. Subscripts are written where the RVs identities are not clear from the context, and hat notation denotes empirical measurements.

Let X, Y be two observed random variables with a true and unknown joint distribution $p^*(x, y)$, and true 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 of N i.i.d observation pairs $\mathcal{S} = \{(x_1, y_1), ..., (x_N, y_N)\}$ 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.

Variational approximations

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).

Learning tasks

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 supervised learning is that of choosing from the given set of functions, the one that best approximates the supervisor's response, based on observation pairs from the training set $\mathcal{S}$, drawn according to $p^*(x, y) = p^*(x)p^*(y|x)$.

Slonim (2002) defines *unsupervised* learning as the task of constructing a compact representation of a set of unlabeled data points $\{x_1, ..., x_N\}, x_i \in \mathbb{R}^d$, 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 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.

Information theoretic functions

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 supplements the related work presented in 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). We denote the resulting curve 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 maximal compression but no information over the down stream task (such a representation may be a null vector), and when β approaches ∞ we learn a representation that has the maximal possible information over the downstream task, but minimal compression. 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.

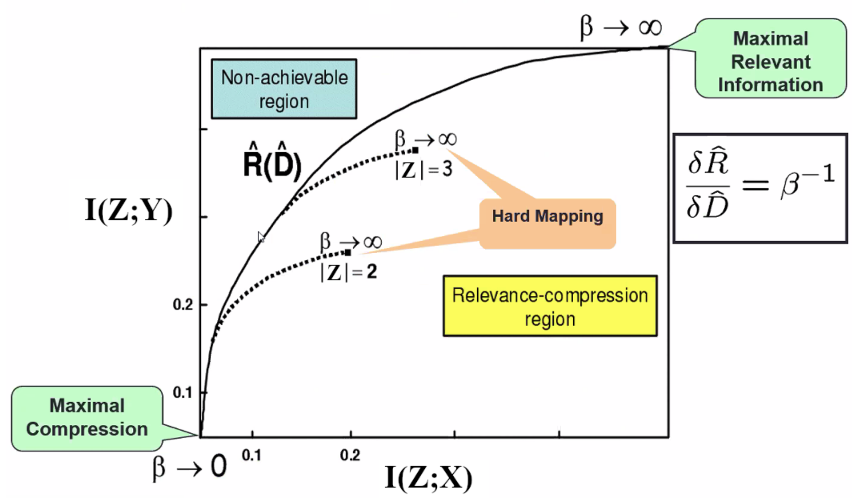

Figure 2: The information plane and curve: rate-distortion ratio over β. At $\beta = 0$ the representation is compressed but uninformative (maximal compression), at $\beta \to \infty$ the representation is informative but potentially overfitted (maximal information). Taken from (Slonim 2002).

Fixing a Broken ELBO

Kingma and Welling (2014) introduced variational auto encoders (VAEs) as a latent model based generative DNN architecture. In VAEs, an unobserved RV Z is assumed to generate evidence X, a variational DNN encoder $e(z|x)$ is used to approximate the intractable posterior $p^*(z|x)$, and a variational DNN decoder $d(\hat{x}|z)$ is used to reconstruct X. The log probability $log\,(p^*(x))$ is developed in to the tractable Evidence Lower Bound (ELBO) loss: $log\,(p^*(x)) \leq \mathcal{L}_{\text{ELBO}}(x) \equiv -\mathbb{E}_{e(z|x)}\left[log\,(d(x|z))\right] + D_{KL}\left(e(z|x)\middle|\middle|m(z)\right)$, consisting of a reconstruction error term (cross entropy), and a KL regularization term between encoder and variational marginal $m(z)$.

Alemi et al. (2018) adapt the information plane (Tishby, Pereira, and Bialek 1999) to VAEs by defining an additional theoretical bound for the ratio between rate and distortion, imposed by the limits of finite parametric families of variational approximations. Instead of true rate and distortion, the proposed information plane features variational rate as $R \equiv D_{KL}\left(e(z|x)\middle|\middle|m(z)\right)$, and variational distortion as $D \equiv -\int\int p^{*}(x)e(z|x)log\left(d(x|z)\right)dxdz$. Figure 3 illustrates the suggested information plane, which 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 $m(z)$; (3) Realizable: Attainable given a finite parametric and tractable variational family. The black diagonal line at the lower left satisfies $H_{p^{*}}(X) - D = R$, resulting in tight variational bounds on the mutual information.

Alemi et al. (2018) observe that the variational rate R does not depend on the variational decoder distribution $d(x|z)$. As R is used as the ELBO KL regularizer, high variational compression rates can be attained regardless of MI between decoder and learned representation. Equivalently, good reconstruction does not directly depend on good representation. Empirical evidence suggest that VAEs are prone to learn uninformative representations while still achieving low ELBO loss, a degeneration made possible by overpowerful decoders that are able to overfit the little information captured by the encoder. $D_{KL}\left(e(z|x)\middle|\middle|m(z)\right)$ approaches 0 iff $e(z|x) \rightarrow m(z)$, making $e(z|x)$ close to independence from x, resulting in a latent representation that fails to encode information about the input. However, a suitably powerful decoder could possibly 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.

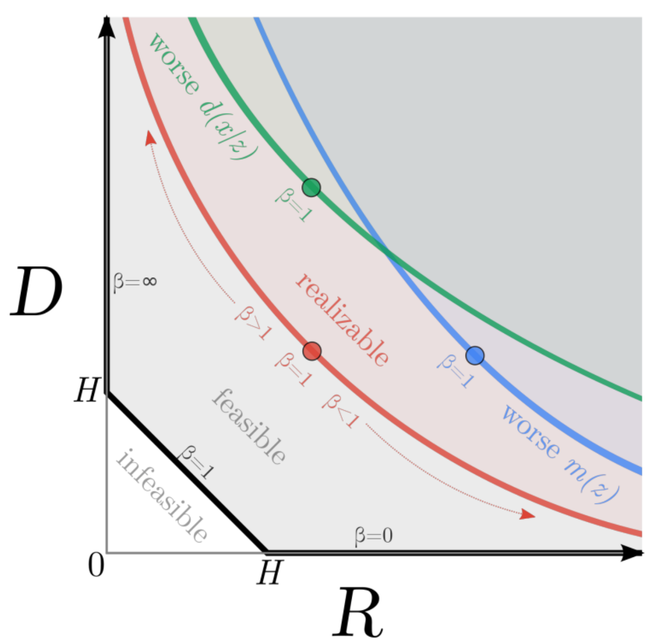

Figure 3: Phase diagram, a proposed information plane interpretation of VAEs. Axes are variational rate and distortion. The IB theoretical limit is extended by an additional limit induced by the constraint of a finite parametric variational family. Once a family is chosen, we seek to learn an optimal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

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).

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 5.

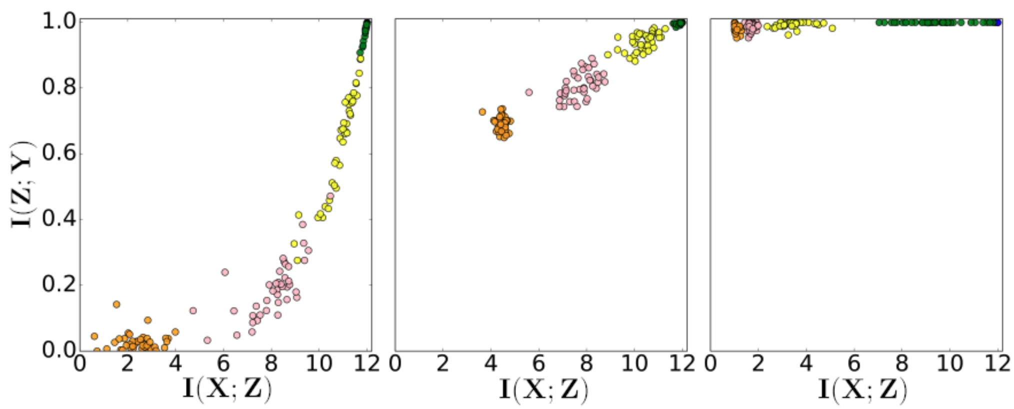

Figure 4: Information plane scatters of different DNN layers (colors) in 50 randomized networks. Left are initial weights, center are at 400 epochs, and right at 9000 epochs. Our study reproduced similar, yet repetitive behavior on complicated high dimensional tasks, as elaborated in Section 5, and in Figure 7. Taken from Shwartz-Ziv and Tishby (2017).

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 separately. Amjad and Geiger (2020) argued against the use of the IB as an objective for deterministic DNNs, as 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. Using IB as an objective in stochastic DNNs, such as of the variational IB family, is suggested as a possible solution. 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 (Cover 1999), 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 with all metrics compared attaining a p-value of less than 0.05. Rank sum was computed as follows: A sorted vector of results was prepared for each compared metric, 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. For example, let $r :=$ $((0.94, 1)\ (0.935, 1)\ (0.93, 1)\ (0.93, 1)\ (0.925, 1)\ (0.92, 0)\ (0.915, 0)\ (0.915, 0)\ (0.91, 0)\ (0.89, 0))$ be a sorted vector of (test accuracy, algorithm) tuples, 1 being VUB, 0 VIB. We compute 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 as follows:

```
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. Examples of successful attacks are shown in Figures 5,6.

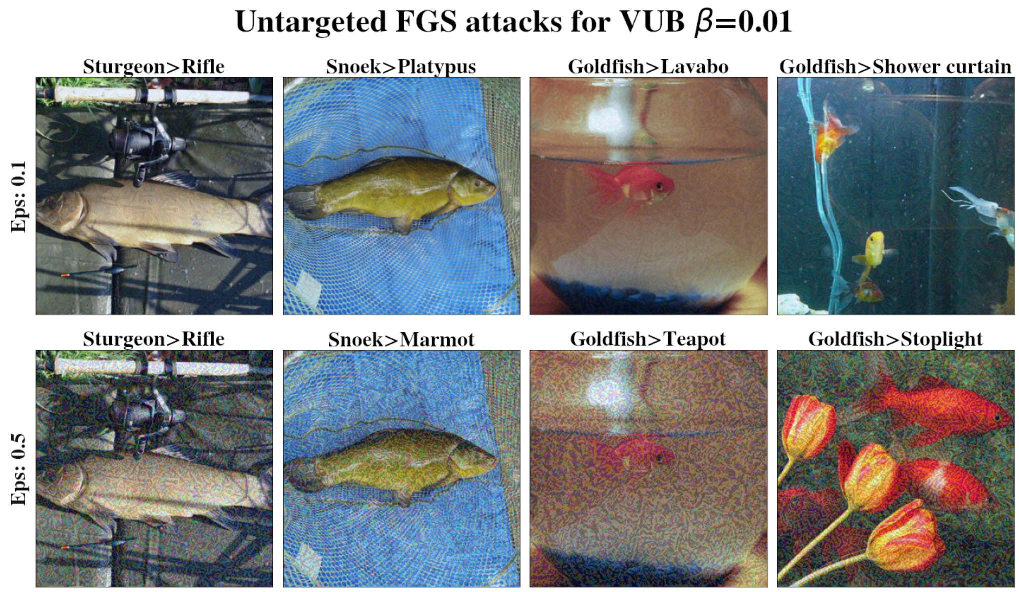

Figure 5: Successful untargeted FGS attack examples. Images are perturbations of previously successfully classified instances from the ImageNet validation set. Perturbation magnitude is determined by the parameter ϵ shown on the left, the higher, the more perturbed. Original and wrongly assigned labels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

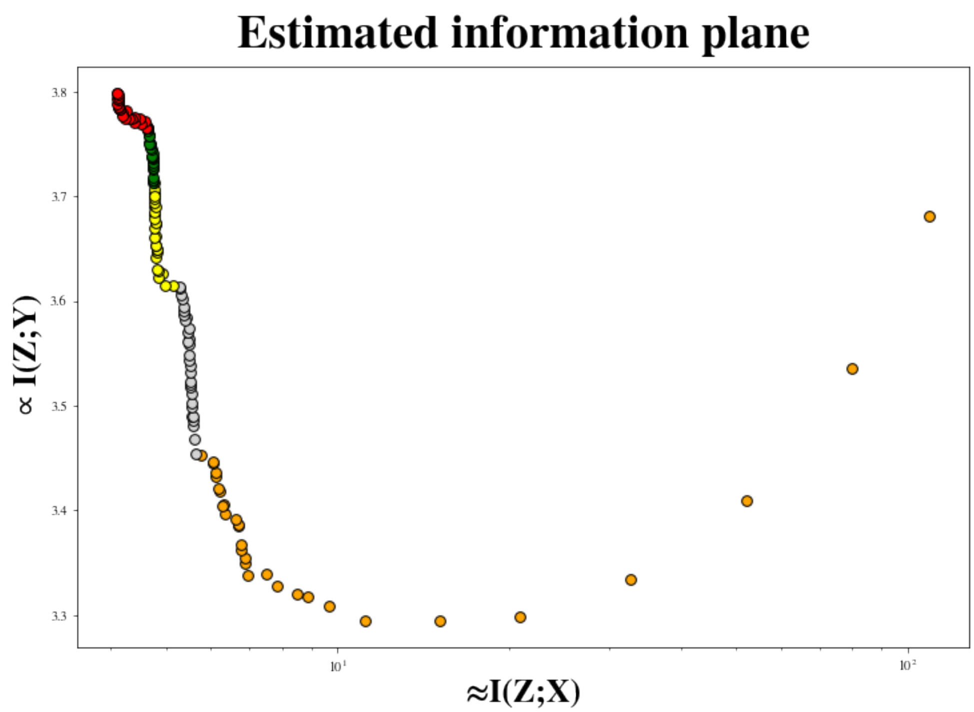

Figure 7: Estimated information plane metrics per epoch for VUB trained on IMDB with $\beta = 0.001$. $I(X;Z)$ is approximated by $H(R) - H(Z|X)$ and $\frac{1}{CE(Y;\hat{Y})}$ is used as an analog for $I(Z;Y)$. The epochs have been grouped and color-coded in intervals of 30 epochs in the order: Orange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

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. Examples of successful attacks are shown in Figures 3,4.

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 text classification training, and plotted them on the information plane as shown in Figure 7. Examining the resulting curve, we notice recurring patterns of distortion reduction followed by rate

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.

increase, resembling the ERM and representation compression stages described by Shwartz-Ziv and Tishby (2017), suggesting that interesting information plane patterns can occur in high dimensional tasks, opening the door to possible future research.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We believe our abstract and intro are well aligned with the main body of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly state experiments where we do not outperform previous work in Section 4, and provide alterative hypothesis for our claim for better generalization, and define the assumptions underlying our claim, in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide a clear proof for our claim of a tighter bound on the IB objective in Section 3, and accompany it with a rigorous analysis of prior work both in Section 2, and in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experiments are provided in full detail in Section 4 and in Appendix C. Provided code is easy to run and is well documented.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The provided code runs easily using a virtual Python environmet, with a complete README.md file and in code documentation. The datasets used are well known and public datasets.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C provides a thorough drill down of the experimental process, including handling of data, hyperparameters and architecture.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Each experiment was run 5 times with mean and std deviation reported in detail. Statistical significance was assured using a Wilcoxon rank sum test as elaborated in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The type of GPU and training time is elaborated in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We've reviewed the NeurIPS code of ethics and assured our research does not breach it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our research focuses on learning theory and does not have any societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not release any proprietary data or high risk model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Every dataset and code repository used are of an open license and are duely cited.

    Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code released with the paper is well documented both with a clear README.md file, and with in code documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.