Parameterized Rate-Distortion Stochastic Encoder

Quan Hoang [1] **Trung Le** [1] **Dinh Phung** [1]

Abstract

We propose a novel gradient-based tractable approach for the Blahut-Arimoto (BA) algorithm to compute the rate-distortion function where the BA algorithm is fully parameterized. This results in a rich and flexible framework to learn a new class of stochastic encoders, termed PArameterized RAte-DIstortion Stochastic Encoder (PARADISE). The framework can be applied to a wide range of settings from semi-supervised, multi-task to supervised and robust learning. We show that the training objective of PARADISE can be seen as a form of regularization that helps improve generalization. With an emphasis on robust learning we further develop a novel posterior matching objective to encourage smoothness on the loss function and show that PARADISE can significantly improve interpretability as well as robustness to adversarial attacks on the CIFAR-10 and ImageNet datasets. In particular, on the CIFAR-10 dataset, our model reduces standard and adversarial error rates in comparison to the state-of-the-art by 50% and 41%, respectively without the expensive computational cost of adversarial training.

1. Introduction

The main objective of representation learning is to learn good representation that can be used for downstream tasks. From this standpoint, rate-distortion theory offers an attractive approach for representation learning where the goodness of a representation can be measured by an appropriately defined distortion function. Rate distortion theory is however often applied in machine learning in the form of the Information Bottleneck (IB) method (Tishby et al., 2000), which measures goodness of a representation by the mutual information with a *relevance variable*. Given the input random variable X and a relevance random variable Y, the objective is to compress X into a representation Z that captures as much information about Y as possible while retaining as little information about X as possible. Mathematically speaking, the IB objective finds the optimal encoding $q(\mathbf{z}|\mathbf{x})$ by minimizing the functional:

$$\mathcal{L}[q(\mathbf{z}|\mathbf{x})] = \mathbb{I}(Z;X) - \beta\mathbb{I}(Z;Y) \tag{1}$$

where β is the Lagrange multiplier. Unfortunately, the iterative algorithm proposed in (Tishby et al., 2000) to optimize the IB objective was intractable and therefore infeasible to apply in practice. Recently, Alemi et al. (2016) proposed Deep Variational Information Bottleneck (DVIB), a variational approximation to the IB objective by using variational lower-bound and upper-bound of mutual information, and claimed that DVIB improves robustness to adversarial attacks. However, DVIB is just an instance of gradient obfuscation (Athalye et al., 2018) as latter shown in Sec. 3.3, thus giving a false sense of robustness. Furthermore, by defining "goodness" of the representation as the mutual information with another variable, the Information Bottleneck method significantly limits the flexibility and potential applications of the rate distortion theory.

In this paper, we revisit the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972) and make a simple modification to make it feasible to numerically compute the rate-distortion function with gradient-based optimization. The result is an elegant and flexible framework for representation learning that can be applied to a wide range of settings from unsupervised, semi-supervised, multi-task to supervised and robust learning. The key component in our framework is a parameterized stochastic encoder that we term *PArameterized RAte-DIstortion Stochastic Encoder* or *PARADISE*. We investigate the behavior of the algorithm on the MNIST (LeCun et al., 1998) and CelebA (Liu et al., 2015) datasets. For supervised learning, we demonstrate that the derived objective can be seen as a form of regularization that helps improve generalization. For robust learning, we show that introducing inductive bias to the learning of PARADISE can significantly improve interpretability as well as robustness to adversarial attacks on the Cifar-10 (Krizhevsky et al., 2009) and ImageNet (Russakovsky et al., 2015) datasets. In particular, on the CIFAR-10 data set, our model reduces standard and adversarial error rates in comparison to the state-of-the-art (Qin et al., 2019) by 50%

[1]Department of DSAI, Faculty of Information Technology, Monash University, Australia. Correspondence to: Quan Hoang <qhoang.ai@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

and 41%, respectively without the expensive computational cost of adversarial training.

In short, our main contributions are: (i) a novel gradient-based tractable approach for the Blahut-Arimoto (BA) algorithm to compute the rate-distortion function where the BA algorithm is fully parameterized; (ii) a new class of stochastic encoders, termed PArameterized RAte- DIstortion Stochastic Encoder (PARADISE), that can be applied to a wide range of settings from semi-supervised, multi-task to supervised and robust learning; (iii) a novel posterior matching objective for robust learning; (iv) a comprehensive evaluation of PARADISE for supervised and robust learning; and (v) a new state-of-the-art result for adversarial accuracy against untargeted white-box attack for Cifar-10, reducing state-of-the-art adversarial errors by 41%.

2. Theoretical Framework

Rate distortion. We start with some key results in rate-distortion theory used in this work.[1] This theory was developed by (Shannon, 1948) in the context of transmitting information over noisy channels. Given an input (message) sequence $\mathbf{x}^n = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ where $\mathbf{x}_i$ is drawn i.i.d. from a source distribution $p(\mathbf{x})$, $\mathbf{x} \in X$, a communication channel receives the input sequence and outputs a sequence $\mathbf{z}^n = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$, $\mathbf{z}_i \in Z$. The sequence $\mathbf{z}^n$ is commonly referred to as the *codeword*, *reconstruction* or *representation.* In this work, we call $\mathbf{z}^n$ representation sequence and each element $\mathbf{z}_i$ representation to avoid confusion. To measure the quality of the output representation, a *distortion measure* (or function) is defined as a mapping from the set of input-representation pairs to the set of non-negative real numbers:

$$d : X \times Z \to \mathbb{R}^+$$

Given the distortion measure $d(\mathbf{x}, \mathbf{z})$, the distortion between sequences $\mathbf{x}^n$ and $\mathbf{z}^n$ is:

$$d(\mathbf{x}^n, \mathbf{z}^n) = \frac{1}{n} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{z}_i)$$

The rate-distortion function $R(D)$ is a mapping from the distortion budget D to the minimum *rate* of the communication channel for which the expected distortion stays within the budget as n goes to infinity. *Rate* can be interpreted as the average number of bits per representation required to specify each representation sequence without confusion. An important result in rate-distortion theory is that for bounded distortion function $d(\mathbf{x}, \mathbf{z})$, the rate-distortion function is equal to the *information rate-distortion function* $R^{(I)}(D)$[2]:

$$R^{(I)}(D) = \min_{p(\mathbf{z}|\mathbf{x}): \sum_{(\mathbf{x},\mathbf{z})} p(\mathbf{x}) p(\mathbf{z}|\mathbf{x}) d(\mathbf{x},\mathbf{z}) \leq D} \mathbb{I}(X; Z) \quad (2)$$

where the minimization is over all conditional distribution $p(\mathbf{z} \mid \mathbf{x})$ for which the expected distortion over the joint distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x})$ stays within the distortion budget D. For the rest of the paper, we will use $R(D)$ to denote both the canonical and information rate distortion function.

There is nice intuition as to why the average number of bits per transmission required is the mutual information $\mathbb{I}(X; Z)$. The conditional distribution $p(\mathbf{z} \mid \mathbf{x})$ induces a soft partitioning of X. The average volume of the elements of X mapped to the same representation is $2^{\mathbb{H}(X|Z)}$ where $\mathbb{H}(X \mid Z)$ is the conditional entropy of X given Z. Since the volume of X is $2^{\mathbb{H}(X)}$, the average cardinality of the partitioning of X is $2^{\mathbb{H}(X)} / 2^{\mathbb{H}(X|Z)} = 2^{\mathbb{I}(X;Z)}$.

To compute the rate-distortion function, the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972) makes use of the following lemma:

Lemma 1. *Let $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x})$ be a given joint distribution. The distribution $q^\star(\mathbf{z})$ that minimizes the Kullback-Leibler (KL) divergence $KL[p(\mathbf{x}, \mathbf{z}) \parallel p(\mathbf{x}) q(\mathbf{z})]$ is the marginal distribution $p(\mathbf{z})$ corresponding to $p(\mathbf{z} \mid \mathbf{x})$:*

$$q^\star(z) = \underset{q(\mathbf{z})}{argmin}\ KL[p(\mathbf{x}, \mathbf{z}) \parallel p(\mathbf{x}) q(\mathbf{z})] = p(\mathbf{z})$$

where $p(\mathbf{z}) = \sum_{\mathbf{x}} p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x})$.

Proof. See Sec. 1 of the supplementary material. □

Lemma 1 turns the problem of computing the rate-distortion function in Eq. (2) into a double minimization problem:

$$R(D) = \min_{q(\mathbf{z})} \min_{p(\mathbf{z}|\mathbf{x}): \mathbb{E}d(\mathbf{x},\mathbf{z}) \leq D} \sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x}) q(\mathbf{z})}$$

By introducing the Lagrange multiplier β for the distortion constraint, computing the rate-distortion function becomes minimizing the following functional:

$$\mathcal{F}[p(\mathbf{z} \mid \mathbf{x}), q(\mathbf{z})] = \sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{x}) q(\mathbf{z})} + \beta \sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x}) d(\mathbf{x}, \mathbf{z}) \quad (3)$$

The Blahut-Arimoto algorithm applies the process of alternating minimization: for a fixed conditional distribution $p^t(\mathbf{z} \mid \mathbf{x})$, the optimal distribution $q^t(\mathbf{z})$, by Lemma 1, is the marginal distribution $p^t(\mathbf{z}) = \sum_x p(\mathbf{x}) p^t(\mathbf{z} \mid \mathbf{x})$; for a fixed $q^t(\mathbf{z})$, the optimal distribution $p^{t+1}(\mathbf{z} \mid \mathbf{x})$ can be found analytically as[3]:

[1] Interested readers are encouraged to find more details in (Cover & Thomas, 2012)(ch 10).

[2] For the sake of simplicity, we slightly abuse the symbol $\sum$ to denote the integration for both continuous and discrete cases.

[3] For self-completeless, see Lemma 2 and the proof in the supplementary material.

$$p^{t+1}(\mathbf{z} \mid \mathbf{x}) = \frac{q^t(\mathbf{z}) \exp(-\beta d(\mathbf{x}, \mathbf{z}))}{\sum_{\mathbf{z}} q^t(\mathbf{z}) \exp(-\beta d(\mathbf{x}, \mathbf{z}))} \quad (4)$$

The algorithm results in a non-increasing sequence $\mathcal{F}(p^0, q^0) \geq \mathcal{F}(p^1, q^0) \geq \mathcal{F}(p^1, q^1) \geq \ldots$ and strictly decreasing unless reaching the limit point

$$p^\star(\mathbf{z} \mid \mathbf{x}) = \frac{p^\star(\mathbf{z}) \exp(-\beta d(\mathbf{x}, \mathbf{z}))}{\sum_{\mathbf{z}} p^\star(\mathbf{z}) \exp(-\beta d(\mathbf{x}, \mathbf{z}))}$$

Parameterised Rate Distortion. Theorem 6 in (Blahut, 1972) shows that this limit point satisfies the necessary and sufficient conditions to achieve the equality in Eq. (2). However, computing the analytical solution in Eq. (4) is intractable in practice. Therefore, we introduce a relaxation to the Blahut-Arimoto algorithm where rather than trying to find the optimal, but intractable, $p^{t+1}(\mathbf{z} \mid \mathbf{x})$ in Eq (4), we seek for the next *tractable* $p^{t+1}(\mathbf{z} \mid \mathbf{x})$ such that $\mathcal{F}(p^t, q^t) \geq \mathcal{F}(p^{t+1}, q^t)$ by taking a gradient step in the direction of $p(\mathbf{z} \mid \mathbf{x})$. With suitable choice of learning rate, this still results in a bounded and non-increasing sequence $\mathcal{F}(p, q)$ and converges to the optimal solution.

To realize this solution, we rewrite the functional objective $\mathcal{F}$ over $p(\mathbf{z} \mid \mathbf{x})$ and $q(\mathbf{z})$ in Eq. (3) with a single θ which parameterizes the conditional distribution $p_\theta(\mathbf{z} \mid \mathbf{x})$ and attains the optimal $q^*_\theta(\mathbf{z}) = \sum_x p_\theta(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x})$. This results in a new parameterized objective w.r.t θ:

$$\mathcal{L}(\theta) = \sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}) p_\theta(\mathbf{z} \mid \mathbf{x}) d(\mathbf{x}, \mathbf{z}) + \alpha \sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}) p_\theta(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{x}) p_\theta(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{x}) q^*_\theta(\mathbf{z})} \quad (5)$$

where $q^*_\theta(\mathbf{z}) = \sum_{\mathbf{x}} p(\mathbf{x}) p_\theta(\mathbf{z} \mid \mathbf{x})$ is the optimal *aggregate posterior* solution. We can now learn θ as follows.[4] At time t, the current solution θ^t parameterizes the conditional distribution $p^t(\mathbf{z} \mid \mathbf{x})$ and we take one gradient step of $\mathcal{L}(\theta)$ to obtain $p_{\theta^{t+1}}(\mathbf{z} \mid \mathbf{x})$ where $q^t(\mathbf{z}) = \text{StopGradient}[\sum_{\mathbf{x}} p(\mathbf{x}) p_{\theta^t}(\mathbf{z} \mid \mathbf{x})]$ is fixed. Here, we use the notation $\text{StopGradient}[\cdot]$ to emphasize that $q^t(\mathbf{z}) \leftarrow p^t(\mathbf{z}) = \sum_{\mathbf{x}} p(\mathbf{x}) p_{\theta_t}(\mathbf{z} \mid \mathbf{x})$, and we keep $q^t(\mathbf{z})$ constant when updating $p^{t+1}(\mathbf{z} \mid \mathbf{x})$ in a similar spirit to the BA algorithm.

The second term in Eq. (5) represents the mutual information and in alignment with rate-distortion theory we name it ***rate*** and denote by $\mathcal{R}(\theta)$. For convenience, the objective in Eq. (5) applies the weight α to the rate term instead of β to the distortion term. Assuming the function family $p_\theta(\mathbf{z} \mid \mathbf{x})$ is rich, optimizing Eq. (5) can learn the parameterized encoder $p_{\theta^\star}(\mathbf{z} \mid \mathbf{x})$ that well approximates the rate-distortion function in Eq (2). Therefore, we name it *PArameterized RAte-DIstortion Stochastic Encoder* (PARADISE).

[4] Please see the algorithm pseudo codes in the supplementary material for additional details.

In the context of deep learning, the distortion function is rarely fixed but desirable to be learned. For example, $d(\mathbf{x}, \mathbf{z})$ can be the mean squared error between $\mathbf{x}$ the reconstruction $\hat{\mathbf{x}}$ of $\mathbf{x}$ from $\mathbf{z}$ using a decoder; or $d(\mathbf{x}, \mathbf{z})$ can be the cross entropy loss of a softmax classifier using $\mathbf{z}$ to predict the label $\mathbf{y}$ associated with $\mathbf{x}$. If the distortion measure is parameterized by $d_\phi(\mathbf{x}, \mathbf{z})$, the parameter θ and ϕ can be learned through joint optimization similar to the EM procedure. Further decomposing the rate $\mathcal{R}(\theta)$ results in the following joint objective function:

$$\begin{aligned}\mathcal{L}(\theta, \phi) = &\mathbb{E}_{p_\theta(\mathbf{x},\mathbf{z})} d_\phi(\mathbf{x}, \mathbf{z}) \\ &- \alpha \sum_{\mathbf{x}} p(\mathbf{x}) \mathbb{H}[p_\theta(\mathbf{z} \mid \mathbf{x})] \\ &- \alpha \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{\mathbf{z}} p_\theta(\mathbf{z} \mid \mathbf{x}) \log q^*_\theta(\mathbf{z}) \end{aligned} \quad (6)$$

where $\mathbb{H}(\cdot)$ is the differential entropy. We call the first term in Eq. (6) the *expected distortion* and denote it as $\mathcal{D}(\theta, \phi)$. The objective $\mathcal{L}(\theta, \phi)$ minimizes the expected distortion while maximizing entropy and encouraging the encoder $p_\theta(\mathbf{z} \mid \mathbf{x})$ to map $\mathbf{x}$ to the representation $\mathbf{z}$ of high score $\log q^*_\theta(\mathbf{z})$.

The entropy term in $\mathcal{L}(\theta, \phi)$ can be analytically computed for certain parameterization choices of $p_\theta(\mathbf{z} \mid \mathbf{x})$ while the score term is more challenging. If the search space of the marginal $q(\mathbf{z})$ is limited to a prior distribution such as the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the rate $\mathcal{R}(\theta)$ becomes $\sum_{\mathbf{x}} p(\mathbf{x}) KL(p_\theta(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z}))$, which is exactly the regularization term used in β-VAE (Higgins et al., 2017) and DVIB (Alemi et al., 2016). Alternatively, we can avoid using a fixed prior $q(\mathbf{z})$ by approximating the score $\log q^*_\theta(\mathbf{z})$ using Mini-batch Weighted Sampling (MWS) proposed in (Chen et al., 2018) (see the supplementary material for details). For a batch size of B and the dimensionality of D for the representation $\mathbf{z}$, the time complexity of approximating the score is $O(B^2 D)$. This cost is small compared to the total cost of the forward and backpropagation passes, especially when the network architecture is deep and wide.

Intuitively, the encoder $p_\theta(\mathbf{z} \mid \mathbf{x})$, when parameterized as diagonal Gaussian, maps each $\mathbf{x}$ to an ellipse. The entropy term in the objective $\mathcal{L}(\theta, \phi)$ (Eq. 6) encourages the ellipses to be as big as possible while the score term pulls the ellipses together. In the mean time, the encoder $p_\theta(\mathbf{z} \mid \mathbf{x})$ must satisfy the distortion constraint. This can be achieved when points $\mathbf{x}$(s) that are, as implied by the distortion measure, similar are mapped to close neighborhoods in the $\mathbf{z}$-space.

Posterior matching for robust learning. In practical deployment, the encoder $p_\theta(\mathbf{z} \mid \mathbf{x})$ often has to deal with "out-of-distribution" inputs. Recall that $p_\theta(\mathbf{z} \mid \mathbf{x})$ induces a soft partitioning of X, and rate minimization is intuitively equivalent to minimizing the cardinality of the partitioning. For some tasks, we may have prior knowledge of what an effi-

cient partitioning should be. For example, one might expect that images that are pixel-wise close to each other should be mapped to similar representation because human are invariant to tiny changes in pixel values. Let $\mathcal{S}(\mathbf{x})$ be a sampling process that draws data points that are considered *similar* to $\mathbf{x}$ based on our prior knowledge. We can introduce inductive bias to the rate minimization procedure by adding a posterior matching (PM) term:

$$\mathcal{PM}(\theta, \mathcal{S}) = \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{\mathbf{x}', \mathbf{x}'' \overset{\text{iid}}{\sim} \mathcal{S}(\mathbf{x})} D\left(p_\theta(\mathbf{z} \mid \mathbf{x}'), p_\theta(\mathbf{z} \mid \mathbf{x}'')\right)$$

where $D(\cdot, \cdot)$ is a discrepancy measure between the posteriors. Overall, the *robust learning* objective function for PARADISE with posterior matching (PARADISE-PM) is:

$$\mathcal{L}_{PM}(\theta, \phi) = \mathcal{D}(\theta, \phi) + \alpha\mathcal{R}(\theta) + \gamma\mathcal{PM}(\theta, \mathcal{S}) \quad (7)$$

There are two main considerations when applying this framework. Firstly, an appropriate distortion measure needs to be chosen according to the downstream tasks. Rate-distortion theory requires the distortion function to be bounded, but we empirically we found that this condition can be relaxed, especially when the distortion function is also jointly learned. Therefore, users have great flexibility in defining the distortion measure. Sec. 3.1 will demonstrate how different distortion measures result in different representation. An interesting scenario is when an encoder is trained using multiple distortion functions, which can be useful for multi-task learning or semi-supervised learning. Fully investigating this direction is, however, out of scope of this paper and left for future work. Sec. 3 instead focuses on applications of the proposed framework to supervised and robust learning.

The second consideration is the inductive biases to be included in the learning. The sampling process $\mathcal{S}(\mathbf{x})$ can be designed depending on the task and the prior knowledge. For example, when $\mathbf{x}$ is an image, $\mathcal{S}(\mathbf{x})$ can be defined using appropriate transformations such as cropping, rotation or pixel perturbations. Sec. 3.3 will demonstrate that introducing such inductive bias can significantly improve the learned model's robustness to adversarial attacks.

3. Experimental Results

We demonstrate several aspects of our proposed model in Sec. 3.1. Then, we report the results on supervised learning setting in Sec. 3.2. Our major results on robust learning are presented in Sec. 3.3. And finally, for additional details and to encourage reproducibility, the supplementary material contains more extensive information on the experiments as well as additional results. We will also release our source codes in public domain.

3.1. Model Behavior

Here, we conduct experiments to investigate the behaviors of the proposed algorithm. We first describe how we apply PARADISE to the supervised and unsupervised setting in our experiment. Recall that the objective function of PARADISE is:

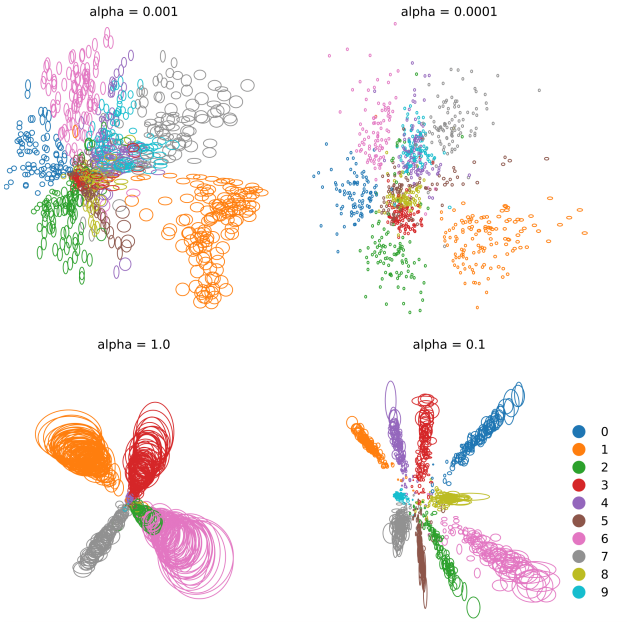
$$\mathcal{L}(\theta, \phi) = \mathcal{D}(\theta, \phi) + \alpha\mathcal{R}(\theta) \quad (8)$$

where $\mathcal{R}(\theta)$ is the rate defined in Sec. 2 and $\mathcal{D}(\theta, \phi) = \mathbb{E}_{p_\theta(\mathbf{x},\mathbf{z})} d_\phi(\mathbf{x}, \mathbf{z})$ is the expected distortion. For the unsupervised setting, we define $d_\phi(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - g_\phi(\mathbf{z})\|_2^2$ or the reconstruction error of $\mathbf{x}$ using a decoder g_ϕ. When $\mathbf{x}$ comes with a label $\mathbf{y} \in [1, C]$, we define $d_\phi(\mathbf{x}, \mathbf{z}) = -\log p_\phi(y \mid \mathbf{z})$ where $p_\phi(y \mid \mathbf{z})$ is the conditional probability corresponding to a softmax classifier c_ϕ. We parameterize $p(\mathbf{z} \mid \mathbf{x})$ as a Gaussian distribution $\mathcal{N}\left(\mu_\theta(\mathbf{x}), \text{diag}\left(\sigma_\theta^2(\mathbf{x})\right)\right)$ with μ_θ and σ_θ being the Neural Network with parameter θ. Both θ and ϕ are learned jointly using gradient descent and the reparameterization trick similar to VAE (Kingma & Welling, 2013). Pseudo-code of the learning algorithms are described in Sec. 1.6 of the supplementary material.

We conduct experiment on MNIST both in both supervised and unsupervised settings to see the impact of the distortion function and the parameter α. For ease of visualization, the dimensionality of $\mathbf{z}$ is set to 2. Details about the architecture and hyperparameters are in Sec. 2 of the supplementary material. Fig. 1 plots the posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ as Gaussian ellipse representing the 95% confidence region for $2,000$ images from the test set. We observe that $\mathbf{z}$ must retain adequate information about $\mathbf{x}$ to reconstruct well. Therefore, a smaller value of α is necessary, and the size of the learned posteriors as well as the level of overlapping among them is much less for the unsupervised setting. The impact of α is observed in both settings, with higher value of α leading to larger ellipses, greater level of overlapping and higher distortion loss. In both cases, harder examples are mapped to smaller posteriors near the center. Finally, when the distortion is based on class label, the posteriors form clusters well separated by color. In the unsupervised case, however, there is significant level of overlapping among class 3, 5, and 8, and between 4 and 9. This reflects the fact that these numbers, often having large blocks of similar pixel values, are considered similar when the distortion is based on the mean squared error (MSE) of reconstruction.

To further investigate the latent space of PARADISE in a more complicated data set, we train PARADISE on the CelebA data set using MSE as the distortion measure. Due to limited space, details about the architecture, hyperparameters and visualization are presented in Sec. 2 and 3 of the supplementary material. After training PARADISE, we fit a multivariate Gaussian distribution to the aggregate posterior of the unseen samples. Then, random $\mathbf{z}$(s) are sampled to generate face images, from which a series of reconstructions are made. The idea is to explore what each neighborhood of $\mathbf{z}$ represents. Fig. 2 shows this serial reconstructions result in images of similar-looking faces with transformations such

Figure 1: Visualization of the posterior $p(\mathbf{z} \mid \mathbf{x})$ as Gaussian ellipse representing the 95% confidence region for $2,000$ images from the test set. Top: unsupervised setting where the objective is to reconstruct $\mathbf{x}$; bottom: supervised setting where the objective is to predict the class associated with $\mathbf{x}$.



as smile, orientation, pose, face shape, eyeglasses, gender, beard and hair style. Interestingly, we also observe linearization of these semantic transformations. For example, Fig. 3 shows that adding the difference between the $\mathbf{z}$-vector of a smiling face and that of a neutral face to the $\mathbf{z}$-vector of another person's neutral face generates the smiling face of that person.

3.2. Supervised Learning

For the experiments in this and the next subsection, we apply the framework for supervised setting as described in the experiment on MNIST in Sec. 3.1. We conduct experiment on the Cifar-10 data set using the pre-activation ResNet (He et al., 2016b) with different number of layers, including 20, 32, 56 and 110 (He et al., 2016a). For the ImageNet data set, we try only the pre-activation ResNet with 34 layers, due to limited computational resources. For all settings, we train models using 10 random seeds and take the average test accuracy, except for ImageNet due to limited resources. For Cifar-10, PARADISE consistently improve test accuracy about 0.3% across architectures (see Tab. 1). On ImageNet, however, top-1 and top-5 accuracy drops about 0.8% and 0.5%, respectively (see Tab. 2). We make a mild conclusion that the rate $\mathcal{R}(\theta)$ acts as a regularizer that can be useful in certain settings but did not help for ImageNet where the base ResNet-34 model shows little sign of overfitting.

Figure 2: CelebA image generation. Images in column 1 are generated from random noise. Images in columns 2 to 5 are successive reconstructions of the previous column. In each rows, one can observe similar-looking faces with transformations such as smile, face orientation and shape, gender and hairstyle.

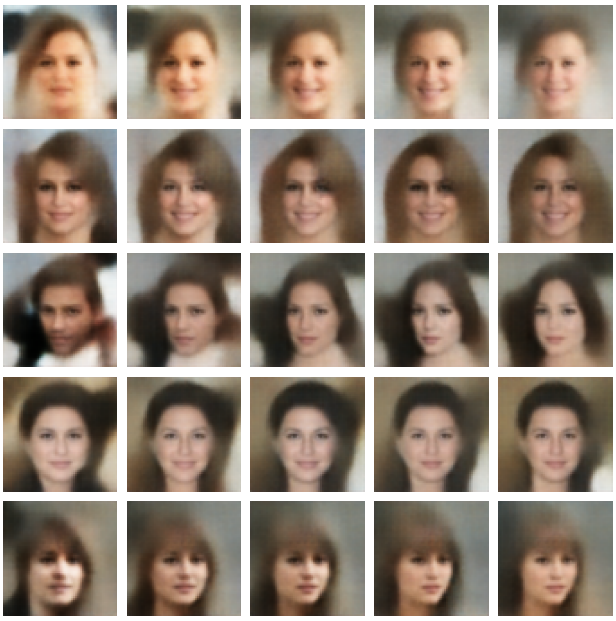


Figure 3: Examples of semantic transformation represented by linear operation in the latent space. In each sub-figure, the difference between the $\mathbf{z}$-vector corresponding to the images on the right and left in the first row is added to the left images of the other rows to generate the right images. The captions describe the transformation observed.

(a) Smile. (b) Gender, Beard.

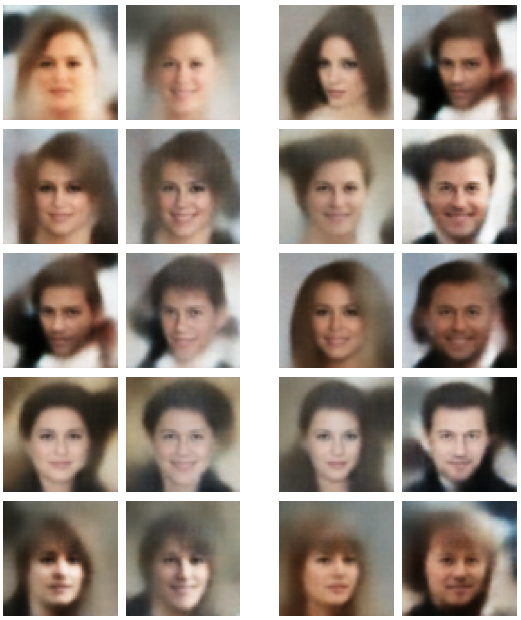


Table 1: Test accuracy (in %) on Cifar-10. To compute test accuracy, we take the average results of models trained using 10 different random seeds.

	RN-20	RN-32	RN-56	RN-110
Standard	91.20	92.16	92.81	93.22
PARADISE	**91.56**	**92.51**	**92.95**	**93.48**

Table 2: Test accuracy (in %) on ImageNet.

	Top-1	Top-5
Standard	72.54	90.73
PARADISE	71.73	90.27

3.3. Robust Learning

The main focus of this section is on improving robustness to adversarial examples, which are carefully perturbed samples that cause AI models to misclassify although the perturbation is not perceptible to humans. The perturbation can be any kind of transformation such as changing pixel value, translating or rotating images. It is challenging to mathematically define perceptibility of perturbation to humans. Most research so far has focused on attacks within a neighborhood of the data point $\mathcal{B}(\mathbf{x}, \epsilon) = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_\infty \leq \epsilon\}$. Attacks can be targeted, e.g. causing the model to predict a random target class, or untargeted, e.g. simply causing the model to misclassify. In terms of method, attacks can be white-box, where the attacker has full access to the model architecture and parameters, or black-box, which does not have such information and often exploits tranferability of white-box attacks (Liu et al., 2016; Tramèr et al., 2017b).

Alemi et al. (2016) claims that Deep Variational Information Bottleneck can improve adversarial robustness. However, the DVIB was evaluated only on white-box attacks. Our investigation (Sec. 3.3, supplementary material) shows that the DVIB model trained on Cifar-10 achieves significantly lower accuracy on black-box attacks than on the white-box FGSM attacks (Goodfellow et al., 2014), a sign of gradient obfuscation (Athalye et al., 2018). The issue is more serious with the higher value of α (equivalent to the β parameter in (Alemi et al., 2016)). Under stronger attacks, DVIB's adversarial accuracy degrades to 0.00% as shown in Tab. 3.

We argue that an inductive bias must be introduced to the learning for the encoder $p_\theta(\mathbf{z} \mid \mathbf{x})$ to handle out-of-distribution inputs well. Therefore, we investigate PARADISE with posterior matching (PARADISE-PM). The robust learning objective from Sec. 2 is:

$$\mathcal{L}_{PM}(\theta, \phi) = \mathcal{D}(\theta, \phi) + \alpha \mathcal{R}(\theta) + \gamma \mathcal{PM}(\theta, \mathcal{S}) \quad (9)$$

where $\mathcal{S}(\mathbf{x})$ is a sampling process to sample points $\mathbf{x}'$ that is *similar* to $\mathbf{x}$ based on our prior knowledge, and $\mathcal{PM}(\theta, \mathcal{S})$ is the posterior matching objective:

$$\mathcal{PM}(\theta, \mathcal{S}) = \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{\mathbf{x}', \mathbf{x}'' \overset{\text{iid}}{\sim} \mathcal{S}(\mathbf{x})} D(p_\theta(\mathbf{z} \mid \mathbf{x}'), p_\theta(\mathbf{z} \mid \mathbf{x}''))$$

where $D(\cdot, \cdot)$ is a discrepancy measure such as the Frechet distance (Dowson & Landau, 1982). We, however, found a simple discrepancy measure based on L_1 is effective in practice. The sampling process $\mathcal{S}(\mathbf{x})$ can be flexibly defined using appropriate transformations such as cropping, rotation or pixel perturbations. In this work, we consider only pixel perturbations. Details about the discrepancy measure and $\mathcal{S}(\mathbf{x})$ are described in Sec. 1 of the supplementary material.

The PM objective resembles the idea of logit matching proposed in (Kannan et al., 2018), which has been shown to learn a bumpier, depressed loss landscape that make it harder to attack but is still highly vulnerable (Engstrom et al., 2018). Our visual investigation, presented later, however shows that PARDADISE-PM has a smooth and highly linear loss surface. In addition, we emphasize that although PARADISE is stochastic, we make the model deterministic at inference time by feeding forward the mean of $p_\theta(\mathbf{z} \mid \mathbf{x})$ to the classifier, thus eliminating the possibility of gradient obfuscation due to the sampling operation.

For Cifar-10, we use the base wide ResNet WRN-28-10 architecture (Zagoruyko & Komodakis, 2016). To evaluate adversarial robustness, we craft strong untargeted and multi-targeted attacks (Gowal et al., 2019) following (Qin et al., 2019). Experiment details are presented in Sec. 2 of the supplementary material. We train DVIB and PARADISE with posterior matching and compare with the state-of-the-art baselines collected from (Qin et al., 2019), including ADV (Madry et al., 2017), TRADES (Zhang et al., 2019b) and LLR (Qin et al., 2019). It should be noted that these baselines are trained with 10-step PGD adversarial examples, thus costing about $11\times$ training time of a standard model. Our method only requires doubling the batch size for posterior matching. Tab. 3 compares the adversarial accuracy on Cifar-10 under the two attacks with the perturbation size ϵ of $8/255$. It can be observed that the PM objective significantly improves robustness of DVIB from 0.00% to 70.71%. Both DVIB-PM and PARADISE-PM outperform the baselines by a huge margin, reducing over 41% adversarial errors. Compared to DVIB-PM, PARADISE-PM achieves about 0.5% higher adversarial accuracy. Furthermore, both DVIB-PM and PARADISE-PM reduce about 50% errors on natural images in comparison to the baselines. Some recent work posits that there is a natural trade-off between standard accuracy and robustness to adversarial examples (Tsipras et al., 2018; Zhang et al., 2019b; Ilyas et al., 2019). Our result confirms this position but suggests that robustness can be achieved with a much smaller drop in standard accuracy.

To evaluate our proposed method on large-scale problems, we conduct experiment on the ImageNet data set. Defense

Table 3: Adversarial accuracy in (%) on Cifar-10 for perturbation size of $8/255$. Baselines include ADV (Madry et al., 2017), TRADES (Zhang et al., 2019b), LLR (Qin et al., 2019) and DVIB (Alemi et al., 2016).

	Natural	Untargeted	Multi-Targeted
ADV	85.11	53.96	48.79
TRADES	87.40	50.46	49.48
LLR	86.83	52.99	51.13
DVIB	**95.72**	0.00	0.00
DVIB-PM	93.53	71.54	70.71
PARADISE-PM	93.77	**72.04**	**71.17**

Table 4: Top-1 accuracy (in %) on white-box attacks crafted on 2,000 ImageNet validation images using 200 PGD steps. DENOISE refers to the ResNet 152 with denoising blocks trained with 30-step PGD attacks in Xie et al. (2019).

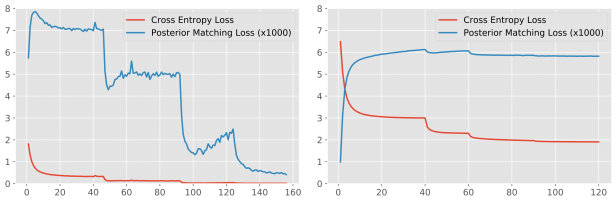
Model	Natural	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 16$
ResNet34	**72.54**	0.00	0.00	0.00
DENOISE	69.70	–	**38.90**	**7.50**
PARADISE-PM	61.57	34.90	16.80	0.60

methods have so far been shown to be unsuccessful on ImageNet. PGD training often incurs a significant drop in accuracy, e.g. 15% to 30% depending on the perturbation size used during training while adding enormous computation burden and achieving limited success. For example, Xie et al. (2019) trains a variant of ResNet-152 against 30-step PGD attacks using 128 GPUs with batch size of 4,096 and achieves the top-1 accuracy of just 7.5% on untargeted attacks with $\epsilon = 16/255$ (Qin et al., 2019). Even for $\epsilon = 2/255$, Uesato et al. (2018) broke three defense methods, degrading top-1 accuracy on untargeted attacks below 1%. In this work, we simply train a humble ResNet-34, as allowed by available resources, to evaluate the potential of the proposed method.

We did not manage to train DVIB-PM on ImageNet despite trying various values of α and γ so we only report the results for PARADISE-PM. Tab. 4 shows top-1 accuracy on white-box attacks for different perturbation sizes. PARADISE-PM significantly improves top-1 accuracy for $\epsilon = 2, 4$ from 0% of the base model to 34.90% and 16.80%, respectively. However, the top-1 accuracy degrades to 0.60% for $\epsilon = 16/255$. Although not comparable, results for the ResNet-152 with denoising blocks trained with 30-step PGD adversarial examples in (Xie et al., 2019) is included in Tab. 4 for reference.

Why did not PARADISE-PM repeat its success on ImageNet? Model capacity is a possible reason. Fig. 4 plots train cross entropy (in red) and PM loss (in blue) over epochs. The PM loss on ImageNet did not decrease during the training and is more than 10 times higher than that on

Figure 4: Train cross entropy (red) and posterior matching loss (blue) over epochs. Posterior matching loss is multiplied by $1,000$ for visibility.



(a) Cifar-10 (b) ImageNet

Table 5: Top-1 black-box accuracy (in %) on 2,000 ImageNet validation images for different perturbation size ϵ.

Model	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Standard	51.95	41.15	31.95	23.45
PARADISE-PM	**60.85**	**60.35**	**59.05**	**49.20**

Cifar-10. In our experiment, we observe that low PM loss is highly indicative of adversarial robustness. On Cifar-10, 70% of the reduction in the PM loss occurs after epoch 96 when the train cross entropy approaches zero, and train accuracy reaches almost 100%. The ImageNet model, however, incurs high cross entropy loss throughout the training. Previous work showed that model capacity plays a crucial role for adversarial robustness (Kurakin et al., 2016; Madry et al., 2017). We hypothesize that PARADISE-PM would perform much better when applied to a more powerful architecture.

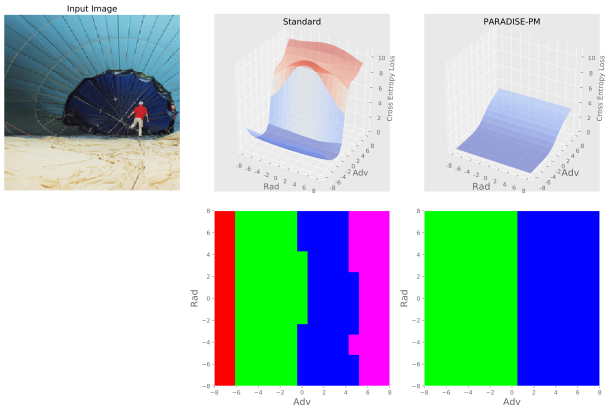
PARADISE-PM is much more robust to black-box attacks. We use attacks from (Goodfellow et al., 2014; Kurakin et al., 2016; Madry et al., 2017; Tramèr et al., 2017a), including FGSM, R+FGSM, Step-Rand, Iter-Rand, and PGD-Rand, to craft adversarial examples and report the min accuracy. Tab. 5 shows the accuracy on black-box attacks for different perturbation sizes. For $\epsilon = 16/255$, PARADISE-PM improves the standard model's black-box accuracy from 23.45% to 49.20%, which is just about 12% drop from PARADISE-PM's accuracy on natural images.

PARADISE-PM's robustness can be explained by Fig. 5, which visualizes the loss surface and decision boundary of PARADISE-PM and the standard model when moving an input image along the signed gradient (adversarial) direction and another random Rademacher vector orthogonal to the signed gradient. The loss surface of PARADISE-PM is much smoother than that of the standard model. Along the random orthogonal direction, the loss is virtually constant, and the class prediction does not change. The standard model however shows bumpy loss surface and changes prediction decision even along the random direction. More examples presented in Sec. 3 of the supplementary material shows the same pattern that PARADISE-PM is virtu-

Table 6: Attacker success rate (in %, lower is better) on 2,000 ImageNet validation images.

Model	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Standard	95.35	99.20	99.95	100
PARADISE-PM	**0.30**	**3.55**	**41.20**	**90.45**

Figure 5: Comparison of the loss surface and decision boundary of PARADISE-PM and the standard model. Top: loss plots generated by moving the input along the signed gradient (adversarial) direction and another random Rademacher vector orthogonal to the signed gradient. Bottom: the corresponding decision boundary; green represents the ground-truth class.



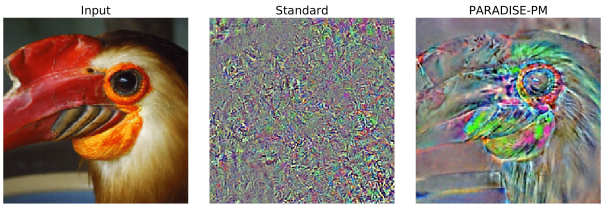
ally constant along random directions, highly linear along the adversarial direction, and has a much simpler decision boundary. Previous work observed that gradients of different models w.r.t the same input are orthogonal (Liu et al., 2016). Our investigation (Sec. 3, supplementary material) confirms this observation. Therefore, it is not surprising that PARADISE-PM, being virtually constant along random orthogonal directions, is robust to black-box attacks.

PARADISE-PM's simpler decision boundary also makes it harder for targeted attacks on PARADISE-PM. Tab. 6 reports the attacker success rate - the percentage of times an attacker successfully causes the model to predict a target class. Lower success rate is better. The success rate on the standard model reaches 95.35% for $\epsilon = 2$ and exceeds 99% for $\epsilon = 4$. The success rate on PARADISE-PM is only 3.55% for $\epsilon = 4$. PARADISE-PM is still vulnerable to targeted attacks for $\epsilon = 16$, but Fig. 6 shows that the attack on PARADISE-PM significantly changes the image of the baby to make him look like a leopard while the targeted attack on the standard model makes little change.

Overall, PARADISE-PM significantly improves adversarial robustness to black-box and targeted attacks but is still vulnerable to white-box attacks. Increasing model capacity as well as introducing adversarial examples during training

Figure 6: Comparison of successful targeted attacks ($\epsilon = 16/255$). Left: the original input; middle: the attack on the standard model; right: the attack on PARADISE-PM. More examples are shown in Sec. 3 of the supplementary material.

Figure 7: Visualization of the loss gradient w.r.t. input pixels. The gradients are clipped within ± 3 standard deviations of their mean and rescale to the range $[0, 1]$ and visualized in RGB mode.



might help optimize the PM loss better. Moreover, some recent work has focused on improving speed while achieving similar performance to PGD-adversarial training (Shafahi et al., 2019; Zhang et al., 2019a; Qin et al., 2019). In particular, Qin et al. (2019) introduces a regularization term to encourage the loss surface more linear and smooth around data points so that the inner maximization takes fewer steps to find hard adversarial examples, thus achieving a $5\times$ speed up for adversarial training on ImageNet. The visual investigation in this section shows that the loss surface of the PARADISE-PM is smooth and highly linear around the data points. Therefore, it can significantly improve efficiency of adversarial training, and combining the two approaches is a promising direction.

Lastly, Fig. 7 visualizes the loss gradient w.r.t to the input pixels to help investigate the input features that strongly affect the model's prediction. We observe that PARADISE-PM attention is significantly more aligned with human while the standard model's gradients look noisy and exhibit no coherent patterns. Similar behavior has been observed in PGD-trained models (Tsipras et al., 2018). More examples are presented in Sec. 3 of the supplementary material.

4. Related Work

Our work is closely related to the Information Bottleneck method (Tishby et al., 2000) which measures goodness of a representation through mutual information with another variable. We, however, define goodness of a representation in a more straightforward manner by using a distortion function

that directly evaluates the performance on downstream tasks. Furthermore, the algorithm proposed in (Tishby et al., 2000) follows the iterative procedure of the Blahut -Arimoto algorithm, which is infeasible to apply in practice. Our simple yet nontrivial modification of the Blahut-Arimoto algorithm makes it possible to compute the rate-distortion function with gradient-based optimization.

Deep Variational Information Bottleneck (DVIB) (Alemi et al., 2016) is a practical realization of the Information Bottleneck method. Sec. 2 shows that the objective function of DVIB can be recovered within our framework when the the search space of the marginal $q(\mathbf{z})$ is fixed. Alemi et al. (2016) claims that DVIB can improve robustness to adversarial attacks, but we show that DVIB is just an instance of gradient obfuscation (Athalye et al., 2018). We further demonstrate that introducing inductive bias to the learning can tremendously improve its robustness.

Wang et al. (2009) proposes a rate-distortion approach for semi-supervised learning by applying the information constraint objective to unlabeled data while computing distortion loss on labeled data. A straightforward application of the PARADISE framework to semi-supervised yields a slightly different formulation where the distortion is defined as the supervised loss approximated on the available labeled data, while the rate-minimization objective is optimized on both labeled and unlabeled data. We can also employ self-supervised objectives as the distortion measure. Semi-supervised learning however is not the focus of the experiments in this work and left for future work.

In the Adversarial Machine Learning literature, many defense methods have been proposed since the discovery of adversarial example phenomenon (Biggio et al., 2013; Szegedy et al., 2013) but Uesato et al. (2018); Athalye et al. (2018) proved that most did not actually improve adversarial robustness. One of the few reliable defense methods is adversarial training where a model is trained on adversarial examples crafted during the training through a few projected-gradient descent (PGD) steps of inner-maximization (Madry et al., 2017). However, three major drawbacks of PGD adversarial training are: i) using k inner-maximization steps adds approximately k times training time or requires a large number of GPUs for training; ii) adversarial training often causes a huge drop in standard accuracy; iii) PGD adversarial training still does not scale to ImageNet. PARADISE-PM adds virtually no additional cost to standard training except for doubling the batch size, experiences a much smaller drop standard accuracy and has a smooth, highly linear loss surface which may require fewer PGD steps to find good adversarial examples. Therefore, combining PARADISE-PM with adversarial training is a promising direction.

5. Discussion and Conclusion

We have presented an efficient and elegant framework for representation learning based on rate-distortion theory with extensive experimental evaluation to demonstrate its merits. The resulting model is a new class stochastic encoders whose key versatility lies in the great freedom in defining the distortion function. A particular interesting scenario is learning representation using multiple distortion functions, which is directly relevant to multi-task learning and semi-supervised learning. PARADISE might also extend to multiple data sources by computing the rate-distortion function for a *product source* (Shannon, 1959).

Under the PARADISE framework, one can introduce inductive bias to explicitly influence the partitioning of the input space. The idea of posterior matching is simple yet extremely effective for adversarial robustness. Combining PARADISE-PM with adversarial training is a promising direction to tackle the Adversarial Machine Learning problem at large scale.

In this work, we considered only robustness to pixel perturbation, but posterior matching can be applied to other kinds of transformation. Furthermore, robustness is not limited to adversarial attacks. Many AI systems today are quite sensitive. For example, the text recognizer in an OCR pipeline may output differently because of updates on the text detector causing changes in the cropped inputs to the text recognizer. The idea of posterior matching can help to deal with such nuisance.

Acknowledgments. This work was partially supported by the US Air Force grant FA2386-19-1-4040.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

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

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Engstrom, L., Ilyas, A., and Athalye, A. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Uesato, J., Qin, C., Huang, P.-S., Mann, T., and Kohli, P. An alternative surrogate loss for pgd-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

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

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pp. 13824–13833, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017a.

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017b.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Uesato, J., O'Donoghue, B., Oord, A. v. d., and Kohli, P. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

Wang, Y., Haffari, G., Wang, S., and Mori, G. A rate distortion approach for semi-supervised conditional random fields. In *Advances in Neural Information Processing Systems*, pp. 2008–2016, 2009.

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. *arXiv preprint arXiv:1905.00877*, 2019a.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.