
Training Normalizing Flows with the Information Bottleneck for Competitive Generative Classification

Lynton Ardizzone*

Visual Learning Lab, Heidelberg University

lynton.ardizzone@iwr.uni-heidelberg.de

Radek Mackowiak*

Bosch Center for Artificial Intelligence (CA)

radek.mackowiak@gmail.com

Carsten Rother

Visual Learning Lab, Heidelberg University

Ullrich Köthe

Visual Learning Lab, Heidelberg University

Abstract

The Information Bottleneck (IB) objective uses information theory to formulate a task-performance versus robustness trade-off. It has been successfully applied in the standard discriminative classification setting. We pose the question whether the IB can also be used to train generative likelihood models such as normalizing flows. Since normalizing flows use invertible network architectures (INNs), they are information-preserving by construction. This seems contradictory to the idea of a bottleneck. In this work, firstly, we develop the theory and methodology of IB-INNs, a class of conditional normalizing flows where INNs are trained using the IB objective: Introducing a small amount of *controlled* information loss allows for an asymptotically exact formulation of the IB, while keeping the INN’s generative capabilities intact. Secondly, we investigate the properties of these models experimentally, specifically used as generative classifiers. This model class offers advantages such as improved uncertainty quantification and out-of-distribution detection, but traditional generative classifier solutions suffer considerably in classification accuracy. We find the trade-off parameter in the IB controls a mix of generative capabilities and accuracy close to standard classifiers. Empirically, our uncertainty estimates in this mixed regime compare favourably to conventional generative and discriminative classifiers. Code: github.com/VLL-HD/IB-INN

1 Introduction

The Information Bottleneck (IB) objective (Tishby et al., 2000) allows for an information-theoretic view of neural networks, for the setting where we have some observed input variable X , and want to predict some Y from it. For simplicity, we limit the discussion to the common case of discrete Y (i.e. class labels), but results readily generalize. The IB postulates existence of a latent space Z , where all information flow between X and Y is channeled through (hence the method’s name). In order to optimize predictive performance, IB attempts to maximize the mutual information $I(Y, Z)$ between Y and Z . Simultaneously, it strives to minimize the mutual information $I(X, Z)$ between X and Z , forcing the model to ignore irrelevant aspects of X which do not contribute to classification performance and only increase the potential for overfitting. The objective can thus be expressed as

$$\mathcal{L}_{\text{IB}} = I(X, Z) - \beta I(Y, Z) . \quad (1)$$

The trade-off parameter β is crucial to balance the two aspects. The IB was successfully applied in a variational form (Alemi et al., 2017; Kolchinsky et al., 2017) to train feed-forward classification models $p(Y|X)$ with higher robustness to overfitting and adversarial attacks than standard ones.

In this work, we consider the relationship between X and Y from the opposite perspective – using the IB, we train an invertible neural network (INN) as a conditional generative likelihood model

$p(X|Y)$, i.e. as a specific type of conditional normalizing flow. In this case, X is the variable of which the likelihood is predicted, and Y is the class condition. It is a generative model because one can sample from the learned $p(X|Y)$ at test time to generate new examples from any class, although we here focus on optimal likelihood estimation for existing inputs, not the generating aspect.

We find that the IB, when applied to such a likelihood model $p(X|Y)$, has special implications for the use as a so-called generative classifier (GC). GCs stand in contrast to standard *discriminative* classifiers (DCs), which directly predict the class probabilities $p(Y|X)$. For a GC, the posterior class probabilities are indirectly inferred at test time by Bayes' rule, cf. Fig. 1: $p(Y|X) = p(X|Y)p(Y)/\mathbb{E}_{p(Y)}[p(X|Y)]$. Because DCs optimize prediction performance directly, they achieve better results in this respect. However, their models for $p(Y|X)$ tend to be most accurate near decision boundaries (where it matters), but deteriorate away from them (where deviations incur no noticeable loss). Consequently, they are poorly calibrated (Guo et al., 2017) and out-of-distribution data can not be easily recognized at test time (Ovadia et al., 2019). In contrast, GCs model full likelihoods $p(X|Y)$ and thus implicitly full posteriors $p(Y|X)$, which leads to the opposite behavior – better predictive uncertainty at the price of reduced accuracy. Fig. 2 illustrates the decision process in latent space Z .

In the past, deep learning models trained in a purely generative way, particularly flow-based models trained with maximum likelihood, achieved highly unsatisfactory accuracy, so that some recent effectiveness of GCs (Fetaya et al., 2019; Nalisnick et al., 2019b). In-depth studies of idealized settings (Bishop & Lasserre, 2007; Bishop, 2007) revealed the existence of a trade-off, controlling the balance between discriminative and generative performance. In this work, we find that the IB can represent this trade-off, when applied to generative likelihood models.

To summarize our contributions, we combine two concepts – the Information Bottleneck (IB) objective and Invertible Neural Networks (INNs). Firstly, we derive an asymptotically exact formulation of the IB for this setting, resulting in our IB-INN model, a special type of conditional normalizing flow. Secondly, we show that this model is especially suitable for the use as a GC: the trade-off parameter β in the IB-INN's loss smoothly interpolates between the advantages of GCs (accurate posterior calibration and outlier detection), and those of DCs (superior task performance). Empirically, at the right setting for β , our model only suffers a minor degradation in classification accuracy compared to DCs while exhibiting more accurate uncertainty quantification than pure DCs or GCs.

2 Related Work

Information Bottleneck: The IB was introduced by Tishby et al. (2000) as a tool for information-theoretic optimization of compression methods. This idea was expanded on by Chechik et al. (2005); Gilad-Bachrach et al. (2003); Shamir et al. (2010) and Friedman et al. (2013). A relationship between IB and deep learning was first proposed by Tishby & Zaslavsky (2015), and later experimentally examined by Shwartz-Ziv & Tishby (2017), who use IB for the understanding of neural network behavior and training dynamics. A close relation of IB to dropout, disentanglement, and variational autoencoding was discovered by Achille & Soatto (2018), which led them to introduce Information Dropout as a way to take advantage of IB in discriminative models. The approximation of IB in a variational setting was proposed independently by Kolchinsky et al. (2017) and Alemi et al. (2017), who especially demonstrate improved robustness against overfitting and adversarial attacks.

Generative Classification: An in-depth analysis of the trade-offs between discriminative and generative models was first performed by Ng & Jordan (2001) and was later extended by Bouchard &

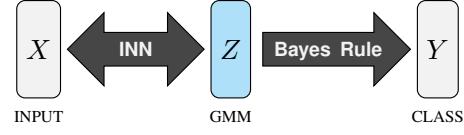


Figure 1: The Information Bottleneck Invertible Neural Network (IB-INN) as a generative classifier.

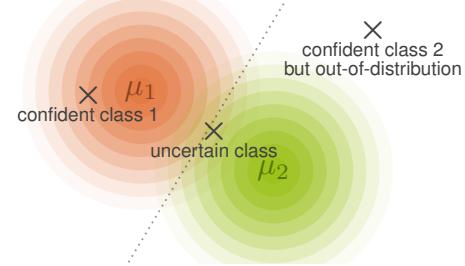


Figure 2: Illustration of the latent output space of a generative classifier. The two class likelihoods for $Y = \{1, 2\}$ are parameterized by their means $\mu_{\{1,2\}}$ in Z . The dotted line represents the decision boundary. A confident, an uncertain, and an out-of-distribution sample are illustrated.

work has called into question the overall effectiveness of GCs (Fetaya et al., 2019; Nalisnick et al., 2019b). In-depth studies of idealized settings (Bishop & Lasserre, 2007; Bishop, 2007) revealed the existence of a trade-off, controlling the balance between discriminative and generative performance. In this work, we find that the IB can represent this trade-off, when applied to generative likelihood models.

Triggs (2004); Bishop & Lasserre (2007); Xue & Titterington (2010), who investigated the possibility of balancing the strengths of both methods via a hyperparameter, albeit for very simple models. GCs have been used more rarely in the deep learning era, some exceptions being application to natural language processing (Yogatama et al., 2017), and adversarial attack robustness (Li et al., 2019; Schott et al., 2019). However, Fetaya et al. (2019) found that conditional normalizing flows have poor discriminative performance, making them unsuitable as GCs. GCs should be clearly distinguished from so-called hybrid models (Raina et al., 2004): these commonly only model the marginal $p(X)$ and jointly perform discriminate classification using shared features, with their main application being semi-supervised learning. Notable examples are Kingma et al. (2014); Chongxuan et al. (2017); Nalisnick et al. (2019c); Grathwohl et al. (2019).

3 Method

Below, upper case letters denote random variables (RVs) (e.g. X) and lower case letters their instances (e.g. x). The probability density function of a RV is written as $p(X)$, the evaluated density as $p(x)$ or $p(X=x)$, and all RVs are vector quantities. We distinguish true distributions from modeled ones by the letters p and q , respectively. The distributions q always depend on model parameters, but we do not make this explicit to avoid notation clutter. Assumption 1 in the appendix provides some weak assumptions about the domains of the RVs and their distributions. Full proofs for all results are also provided in the appendix.

Our models have two kinds of learnable parameters. Firstly, an invertible neural network (INN) with parameters θ maps inputs X to latent variables Z bijectively: $Z = g_\theta(X) \Leftrightarrow X = g_\theta^{-1}(Z)$. Assumption 2 in the Appendix provides some explicit assumptions about the network, its gradients, and the parameter space, which are largely fulfilled by standard invertible network architectures, including the affine coupling architecture we use in the experiments. Secondly, a Gaussian mixture model with class-dependent means μ_y , where y are the class labels, and unit covariance matrices is used as a reference distribution for the latent variables Z :

$$q(Z | Y) = \mathcal{N}(\mu_y, \mathbb{I}) \quad \text{and} \quad q(Z) = \sum_y p(y) \mathcal{N}(\mu_y, \mathbb{I}). \quad (2)$$

For simplicity, we assume that the label distribution is known, i.e. $q(Y) = p(Y)$. Our derivation rests on a quantity we call *mutual cross-information* CI (in analogy to the well-known cross-entropy):

$$CI(U, V) = \mathbb{E}_{u, v \sim p(U, V)} \left[\log \frac{q(u, v)}{q(u)q(v)} \right]. \quad (3)$$

Note that the expectation is taken over the true distribution p , whereas the logarithm involves model distributions q . In contrast, plain mutual information uses the same distribution in both places. Our definition is equivalent to the recently proposed predictive \mathcal{V} -information (Xu et al., 2020), whose authors provide additional intuition and guarantees. The following proposition (proof in Appendix) clarifies the relationship between mutual information I and CI :

Proposition 1. *Assume that $q(\cdot)$ can be chosen from a sufficiently rich model family (e.g. a universal density estimator, see Assumption 2). Then for every $\eta > 0$ there is a model such that $|I(U, V) - CI(U, V)| < \eta$ and $I(U, V) = CI(U, V)$ if $p(u, v) = q(u, v)$.*

We replace both mutual information terms $I(X, Z)$ and $I(Y, Z)$ in Eq. 1 with the mutual cross-information CI , and derive optimization procedures for each term in the following subsections.

3.1 INN-Based Formulation of the $I(X, Z)$ -Term in the IB Objective

Estimation of the mutual cross-information $CI(X, Z)$ between inputs and latents is problematic for deterministic mappings from X to Z (Amjad & Geiger, 2018), and specifically for INNs, which are bijective by construction. In this case, the joint distributions $q(X, Z)$ and $p(X, Z)$ are not valid Radon-Nikodym densities and both CI and I are undefined. Intuitively, I and CI become infinite, because p and q have an infinitely high delta-peak at $Z = g_\theta(X)$, and are otherwise 0. For the IB to be applicable, some information has to be discarded in the mapping to Z , making p and q valid Radon-Nikodym densities. In contrast, normalizing flows rely on all information to be retained for optimal generative capabilities and density estimation.

Our solution to this seeming contradiction comes from the practical use of normalizing flows. Here, a small amount of noise is commonly added to dequantize X (i.e. to turn discrete pixel values

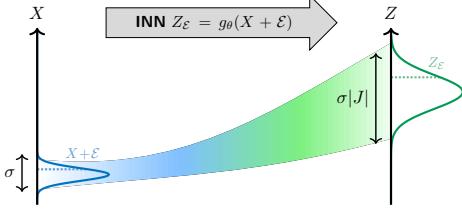


Figure 3: The more the noise is amplified in relation to the noise-free input, the lower the mutual cross-information between noisy latent vector Z_ε and noise-free input X .

into real numbers), to avoid numerical issues during training. We adopt this approach to artificially introduce a minimal amount of information loss: Instead of feeding X to the network, we input a noisy version $X' = X + \mathcal{E}$, where $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) = p(\mathcal{E})$ is Gaussian with mean zero and covariance $\sigma^2 \mathbb{I}$. For a quantization step size ΔX , the additional error on the estimated densities caused by the augmentation has a known bound decaying with $\exp(-\Delta X^2/2\sigma^2)$ (see Appendix). We are interested in the limit $\sigma \rightarrow 0$, so in practice, we choose a very small fixed σ , that is smaller than ΔX . This makes the error practically indistinguishable from zero. The INN then learns the bijective mapping $Z_\varepsilon = g_\theta(X + \mathcal{E})$, which guarantees $CI(X, Z_\varepsilon)$ to be well defined. Minimizing this CI according to the IB principle means that $g_\theta(X + \mathcal{E})$ is encouraged to amplify the noise \mathcal{E} , so that X can be recovered less accurately, see Fig. 3 for illustration. If the global minimum of the loss is achieved w.r.t. θ , I and CI coincide, as $CI(X, Z_\varepsilon)$ is an upper bound (also cf. Prop. 1):

Proposition 2. *For the specific case that $Z_\varepsilon = g_\theta(X + \mathcal{E})$, it holds that $I(X, Z_\varepsilon) \leq CI(X, Z_\varepsilon)$.*

Our approach should be clearly distinguished from applications of the IB to DCs, such as Alemi et al. (2017), which pursue a different goal. There, the model learns to ignore the vast majority of input information and keeps only enough to predict the class posterior $p(Y | X)$. In contrast, we induce only a small, explicitly adjustable loss of information to make the IB well-defined. As a result, the amount of retained information in our generative IB-INNs is orders of magnitude larger than in DC approaches, which is necessary to represent accurate class-conditional likelihoods $p(X | Y)$.

We now derive the loss function that allows optimizing θ and μ_y to minimize the noise-augmented $CI(X, Z_\varepsilon)$ in the limit of small noise $\sigma \rightarrow 0$. Full details are found in appendix. We decompose the mutual cross-information into two terms

$$CI(X, Z_\varepsilon) = \mathbb{E}_{p(X), p(\mathcal{E})} \left[-\log q(Z_\varepsilon = g_\theta(x + \varepsilon)) \right] + \underbrace{\mathbb{E}_{p(X), p(\mathcal{E})} \left[\log q(Z_\varepsilon = g_\theta(x + \varepsilon) | x) \right]}_{:=A}.$$

The first expectation can be approximated by the empirical mean over a finite dataset, because the Gaussian mixture distribution $q(Z_\varepsilon)$ is known analytically. To approximate the second term, we first note that the condition $X = x$ can be replaced with $Z = g_\theta(x)$, because g_θ is bijective and both conditions convey the same information

$$A = \mathbb{E}_{p(X), p(\mathcal{E})} \left[\log q(Z_\varepsilon = g_\theta(x + \varepsilon) | Z = g_\theta(x)) \right].$$

We now linearize g_θ by its first order Taylor expansion, $g_\theta(x + \varepsilon) = g_\theta(x) + J_x \varepsilon + O(\varepsilon^2)$, where $J_x = \frac{\partial g_\theta(X)}{\partial X} \Big|_x$ denotes the Jacobian at $X = x$. Going forward, we write $O(\sigma^2)$ instead of $O(\varepsilon^2)$ for clarity, noting that both are equivalent because we can write $\varepsilon = \sigma n$ with $n \sim \mathcal{N}(0, \mathbb{I})$, and $\|\varepsilon\| = \sigma \|n\|$. Inserting the expansion into A , the $O(\sigma^2)$ can be moved outside of the expression: It can be moved outside the log, because that has a Lipschitz constant of $1/\inf q(g_\theta(X + \mathcal{E}))$, which we show is uniformly bounded in the full proof. The $O(\sigma^2)$ can then be exchanged with the expectation because the expectation's argument is also uniformly bounded, finally leading to

$$A = \mathbb{E}_{p(X), p(\mathcal{E})} \left[\log q(g_\theta(x) + J_x \varepsilon | g_\theta(x)) \right] + O(\sigma^2).$$

Since ε is Gaussian with mean zero and covariance $\sigma^2 \mathbb{I}$, the conditional distribution is Gaussian with mean $g_\theta(x)$ and covariance $\sigma^2 J_x J_x^T$. The expectation with respect to $p(\mathcal{E})$ is thus the negative entropy of a multivariate Gaussian and can be computed analytically as well

$$\begin{aligned} A &= \mathbb{E}_{p(X)} \left[-\frac{1}{2} \log (\det(2\pi e \sigma^2 J_x J_x^T)) \right] + O(\sigma^2) \\ &= \mathbb{E}_{p(X)} \left[-\log |\det(J_x)| \right] - d \log(\sigma) - \frac{d}{2} \log(2\pi e) + O(\sigma^2) \end{aligned}$$

with d the dimension of X . To avoid running the model twice (for x and $x + \varepsilon$), we approximate the expectation of the Jacobian determinant by 0th-order Taylor expansion as

$$\mathbb{E}_{p(X)} [\log |\det(J_x)|] = \mathbb{E}_{p(X), p(\mathcal{E})} [\log |\det(J_\varepsilon)|] + O(\sigma),$$

where J_ε is the Jacobian evaluated at $x + \varepsilon$ instead of x . The residual can be moved outside of the log and the expectation because J_ε is uniformly bounded in our networks.

Putting everything together, we drop terms from $CI(X, Z_\varepsilon)$ that are independent of the model or vanish with rate at least $O(\sigma)$ as $\sigma \rightarrow 0$. The resulting loss \mathcal{L}_X becomes

$$\mathcal{L}_X = \mathbb{E}_{p(X), p(\varepsilon)} [-\log q(g_\theta(x+\varepsilon)) - \log |\det(J_\varepsilon)|]. \quad (4)$$

Since the change of variables formula defines the network's generative distribution as $q_X(x) = q(Z = g_\theta(x)) |\det(J_x)|$, \mathcal{L}_X is the negative log-likelihood of the perturbed data under q_X ,

$$\mathcal{L}_X = \mathbb{E}_{p(X), p(\varepsilon)} [-\log q_X(x + \varepsilon)]. \quad (5)$$

The crucial difference between $CI(X, Z_\varepsilon)$ and \mathcal{L}_X is the elimination of the term $-d \log(\sigma)$. It is huge for small σ and would dominate the model-dependent terms, making minimization of $CI(X, Z_\varepsilon)$ very hard. Intuitively, the fact that $CI(X, Z_\varepsilon)$ diverges for $\sigma \rightarrow 0$ highlights why $CI(X, Z)$ is undefined for bijectively related X and Z . In practice, we estimate \mathcal{L}_X by its empirical mean on a training set $\{(x_i, \varepsilon_i)\}_{i=1}^N$ of size N , denoted as $\mathcal{L}_X^{(N)}$.

It remains to be shown that replacing $I(X, Z_\varepsilon)$ with $\mathcal{L}_X^{(N)}$ in the IB loss Eq. 1 does not fundamentally change the solution of the learning problem in the limit of large N , small σ and sufficient model power. Sufficient model power here means that the family of generative distributions realizable by g_θ should be a universal density estimator (see Appendix, Assumption 2). This is the case if g_θ can represent increasing triangular maps (Bogachev et al., 2005), which has been proven for certain network architectures explicitly (e.g. Jaini et al., 2019; Huang et al., 2018), including the affine coupling networks we use for the experiments (Teshima et al., 2020). Propositions 1 & 2 then tell us that we may optimize $CI(X, Z_\varepsilon)$ as an estimator of $I(X, Z_\varepsilon)$. The above derivation of the loss can be strengthened into

Proposition 3. *Under Assumptions 1 and 2, for any $\epsilon, \eta > 0$ and $0 < \delta < 1$ there are $\sigma_0 > 0$ and $N_0 \in \mathbb{N}$, such that $\forall N \geq N_0$ and $\forall 0 < \sigma < \sigma_0$, the following holds uniformly for all model parameters θ :*

$$\begin{aligned} \Pr \left(\left| CI(X, Z_\varepsilon) + d \log \sqrt{2\pi e \sigma^2} - \mathcal{L}_X^{(N)} \right| > \epsilon \right) &< \delta \\ \text{and } \Pr \left(\left\| \frac{\partial}{\partial \theta} CI(X, Z_\varepsilon) - \frac{\partial}{\partial \theta} \mathcal{L}_X^{(N)} \right\| > \eta \right) &< \delta \end{aligned}$$

The first statement proves consistence of $\mathcal{L}_X^{(N)}$, and the second justifies gradient-descent optimization on the basis of $\mathcal{L}_X^{(N)}$. Proofs can be found in the appendix.

3.2 GMM-Based Formulation of the $I(Z, Y)$ -Term in the IB Objective

Similarly to the first term in the IB-loss in Eq. 1, we also replace the mutual information $I(Y, Z)$ with $CI(Y, Z_\varepsilon)$. Inserting the likelihood $q(z | y) = \mathcal{N}(z; \mu_y, \mathbb{I})$ of our latent Gaussian mixture model into the definition and recalling that $q(Y) = p(Y)$, this can be decomposed into

$$CI(Y, Z_\varepsilon) = \mathbb{E}_{p(Y)} [-\log p(y)] + \mathbb{E}_{p(X, Y), p(\varepsilon)} \left[\log \frac{q(g_\theta(x+\varepsilon) | y) p(y)}{\sum_{y'} q(g_\theta(x+\varepsilon) | y') p(y')} \right]. \quad (6)$$

In this case, $CI(Y, Z_\varepsilon)$ is a lower bound on the true mutual information $I(Y, Z_\varepsilon)$, allowing for its maximization in our objective. In fact, it corresponds to a bound originally proposed by Barber & Agakov (2003) (see their Eq. 3): The first term is simply the entropy $h(Y)$, because $p(Y)$ is known. The second term can be rewritten as the negative cross-entropy $-h_q(Y | Z_\varepsilon)$. For $I(Y, Z_\varepsilon)$, we would have the negative entropy $-h(Y | Z_\varepsilon)$ in its place, then Gibbs' inequality leads directly to $CI(Y, Z_\varepsilon) \leq I(Y, Z_\varepsilon)$.

The first expectation can be dropped during training, as it is model-independent. Note how the the second term can also be written as the expectation of the GMM's log-posterior $\log q(y | z)$. Since all mixture components have unit covariance, the elements of Z are conditionally independent and the likelihood factorizes as $q(z | y) = \prod_j q(z_j | y)$. Thus, $q(y | z)$ can be interpreted as a naive Bayes classifier. In contrast to naive Bayes classifiers in data space, which typically perform badly because

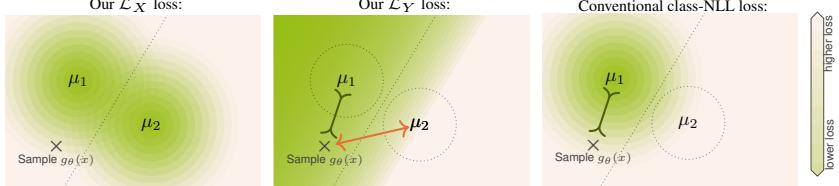


Figure 4: Illustration of the loss landscape for our IB formulation (*left, middle*) and standard class-conditional negative-log-likelihood (*right*). The loss is shown for an input x belonging to class $Y=1$, green areas correspond to low loss. The orange arrows and black inverted arrows indicate repulsive and attractive interactions with the cluster centers. Crucially, standard NNL exerts no repulsive force.

raw features are not conditionally independent, our training enforces this property in latent space and ensures accurate classification. Defining the loss $\mathcal{L}_Y^{(N)}$ as the empirical mean of the log-posterior in a training set $\{x_i, y_i, \varepsilon_i\}_{i=1}^N$ of size N , we get

$$\mathcal{L}_Y^{(N)} = \frac{1}{N} \sum_{i=1}^N \log \frac{\mathcal{N}(g_\theta(x_i + \varepsilon_i); \mu_{y_i}, \mathbb{I}) p(y_i)}{\sum_{y'} \mathcal{N}(g_\theta(x_i + \varepsilon_i); \mu_{y'}, \mathbb{I}) p(y')}. \quad (7)$$

3.3 The IB-INN-Loss and its Advantages

Replacing the mutual information terms in Eq. 1 with their empirical estimates $\mathcal{L}_X^{(N)}$ and $\mathcal{L}_Y^{(N)}$, our model parameters θ and $\{\mu_1, \dots, \mu_K\}$ are trained by gradient descent of the *IB-INN loss*

$$\mathcal{L}_{\text{IB-INN}}^{(N)} = \mathcal{L}_X^{(N)} - \beta \mathcal{L}_Y^{(N)} \quad (8)$$

In the following, we will interpret and discuss the nature of the loss function in Eq. 8 and form an intuitive understanding of why it is more suitable than the class-conditional negative-log-likelihood ('class-NLL') traditionally used for normalizing-flow type generative classifiers: $\mathcal{L}_{\text{class-NLL}} = -\mathbb{E} \log (q_\theta(x|y))$. The findings are represented graphically in Fig. 4.

\mathcal{L}_X -term: As shown by Eq. 5, the term is the (unconditional) negative-log-likelihood loss used for normalizing flows, with the difference that $q(Z)$ is a GMM rather than a unimodal Gaussian. We conclude that this loss term encourages the INN to become an accurate likelihood model under the marginalized latent distribution and to ignore any class information.

\mathcal{L}_Y -term: Examining Eq. 7, we see that for any pair $(g(x + \varepsilon), y)$, the cluster centers $(\mu_{Y \neq y})$ of the other classes are repulsed (by minimizing the denominator), while $g_\theta(x + \varepsilon)$ and the correct cluster center μ_y are drawn together. Note that the class-NLL loss only captures the second aspect and lacks repulsion, resulting in a much weaker training signal. We can also view this in a different way: by substituting $q(x|y) |\det(J_x)|^{-1}$ for $q(z|y)$, the second summand of Eq. 6 simplifies to $\log q(y|x)$, since the Jacobian cancels out. This means that our \mathcal{L}_Y loss directly maximizes the correct class probability, while ignoring the data likelihood. Again, this improves the training signal: as Fetaya et al. (2019) showed, the data likelihood will otherwise dominate the class-NLL loss, so that lack of classification accuracy is insufficiently penalized.

Classical class-NLL loss: The class-NLL loss or an approximation thereof is used to train standard GCs. The IB-INN loss reduces to this case for $\beta = 1$, because the first summand in \mathcal{L}_X (cf. Eq. 4) cancels with the denominator in Eq. 7. Then, the INN no longer receives a penalty when latent mixture components overlap, and the GMM loses its class discriminatory power, as Fig. 4 illustrates: Points are only drawn towards the correct class, but there is no loss component repulsing them from the incorrect classes. As a result, all cluster centers tend to collapse together, leading the INN to effectively just model the marginal data likelihood (as found by Fetaya et al., 2019). Similarly, Wu et al. (2019) found that $\beta = 1$ is the minimum possible value to perform classification with discriminative IB methods.

4 Experiments

In the following, we examine the properties of the IB-INN used as a GC, especially the quality of uncertainty estimates and OoD detection. We construct our IB-INN by combining the design efforts of various works on INNs and normalizing flows. In brief, we use a Real-NVP architecture consisting of affine coupling blocks (Dinh et al., 2017), with added improvements from recent works (Kingma & Dhariwal, 2018; Jacobsen et al., 2019, 2018; Ardzizzone et al., 2019). A detailed description of

the architecture is given in the appendix. We learn the set of means μ_Y as free parameters jointly with the remaining model parameters in an end-to-end fashion using the loss in Eq. 8. The practical implementation of the loss is explained in the appendix.

We apply two additional techniques while learning the model, label smoothing and loss rebalancing:

Label smoothing Hard labels force the Gaussian mixture components to be maximally separated, so they drift continually further apart during training, leading to instabilities. Label smoothing (Szegedy et al., 2016) with smoothing factor 0.05 prevents this, and we also apply it to all baseline models.

Loss rebalancing The following rebalancing scheme allows us to use the same hyperparameters when changing β between 5 orders of magnitude. Firstly, we divide the loss \mathcal{L}_X by the number of dimensions of X , which approximately matches its magnitude to the \mathcal{L}_Y loss. We define a corresponding $\gamma := \beta / \dim(X)$ to stay consistent with the IB definition. Secondly, we scale the entire loss by a factor $2/(1 + \gamma)$. This ensures that it keeps the same magnitude when changing γ .

$$\mathcal{L}_{\text{IB}}^{(N)} = \frac{2}{1 + \gamma} \left(\frac{\mathcal{L}_X^{(N)}}{\dim(X)} - \gamma \mathcal{L}_Y^{(N)} \right) \quad (9)$$

Finally, the noise amplitude σ should be chosen to satisfy two criteria: it should be small enough so that the Taylor expansions in the loss for $\sigma \rightarrow 0$ are sufficiently accurate, and it should also not hinder the model’s performance. Our ablation provided in the Appendix indicates that both criteria are satisfied when $\sigma \lesssim 0.25\Delta X$, with the quantization step size ΔX , so we fix $\sigma = 10^{-3}$ for the remaining experiments.

4.1 Comparison of Methods

In addition to the IB-**INN**, we train several alternative methods. For each, we use exactly the same INN model, or an equivalent feed-forward ResNet model. Every method has the exact same hyperparameters and training procedure, the only difference being the loss function and invertibility.

Class-NLL: As a standard generative classifier, we firstly train an INN with a GMM in latent space naively as a conditional generative model, using the class-conditional maximum likelihood loss. Secondly, we also train a regularized version, to increase the classification accuracy. The regularization consists of leaving the class centroids μ_Y fixed on a hyper-sphere, forcing some degree of class-separation.

Feed-forward As a DC baseline, we train a standard ResNet (He et al., 2016) with softmax cross entropy loss. We replace each affine coupling block by a ResNet block, leaving all other hyperparameters the same.

i-RevNet (Jacobsen et al., 2018): To rule out any differences stemming from the constraint of invertibility, we additionally train the INN as a standard softmax classifier, by projecting the outputs to class logits. While the architecture is invertible, it is not a generative model and trained just like a standard feed-forward classifier.

Variational Information Bottleneck (VIB): To examine which observed behaviours are due to the IB in general, and what is specific to GCs, we also train the VIB (Alemi et al., 2017), a feed-forward DC, using a ResNet. We convert the authors definition of β to our γ for consistency.

4.2 Quantitative measurements

In the following, we describe the scores used in Table 1.

Bits/dim: The bits/dim metric is common for objectively comparing the performance of density estimation models such as normalizing flows, and is closely related to the KL divergence between real and estimated distributions. Details can be found e.g. in Theis et al. (2015).

Calibration error: The calibration curve measures whether the confidence of a model agrees with its actual performance. All prediction outputs are binned according to their predicted probability P (‘confidence’), and it is recorded which fraction of predictions in each bin was correct, Q . For a perfectly calibrated model, we have $P = Q$, e.g. predictions with 70% confidence are correct 70% of the time. We use several metrics to measure deviations from this behaviour, largely in line with Guo et al. (2017). Specifically, we consider the expected calibration error (ECE, error weighted by bin count), the maximum calibration error (MCE, max error over all bins), and the integrated calibration error (ICE, summed error per bin), as well as the geometric

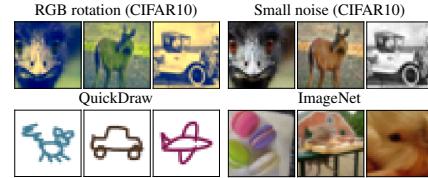


Figure 5: Examples from each OoD dataset used in the evaluation. The inlier data are original CIFAR10 images.

Table 1: Results on the CIFAR10 dataset. All models have the same number of parameters and were trained with the same hyperparameters. All values except entropy and overconfidence are given in percent. The arrows indicate whether a higher or lower value is better.

Model	Classif. err. (↓)	Bits/dim (↓)	Calibration error (↓)			Incr. OoD prediction entropy (↑)			OoD detection score (↑)								
			Geo. mean	ECE	MCE	ICE	Average	RGB-rot.	Draw	Noise	ImgNet	Average	RGB-rot	Draw	Noise	ImgNet	
IB-INN (ours)	$\gamma = 1$	10.27	5.25	1.26	0.54	3.25	1.13	0.38	0.43	0.40	0.10	0.61	68.76	78.80	67.30	77.19	54.59
	only \mathcal{L}_X ($\gamma = 0$)	—	4.80	—	—	—	—	—	—	—	—	—	74.51	70.68	85.74	91.14	55.82
Stand. GC	Class-NLL	61.75	4.81	12.61	4.17	30.58	15.70	0.03	0.02	-0.06	0.02	0.12	73.92	70.65	83.31	90.97	55.76
	Class-NLL + regul.	40.04	4.83	24.75	7.13	70.63	30.11	0.01	0.00	-0.01	0.01	0.02	74.02	69.33	85.13	91.04	55.88
Pure DC	VIB ($\gamma = 1$)	6.83	—	6.66	0.81	26.56	13.75	0.17	0.14	0.23	0.00	0.32	—	—	—	—	—
	ResNet	6.51	—	6.23	0.76	29.29	10.92	0.18	0.16	0.20	0.00	0.34	—	—	—	—	—
	i-RevNet	9.22	—	4.19	0.79	16.68	5.54	0.24	0.09	0.38	0.00	0.51	—	—	—	—	—

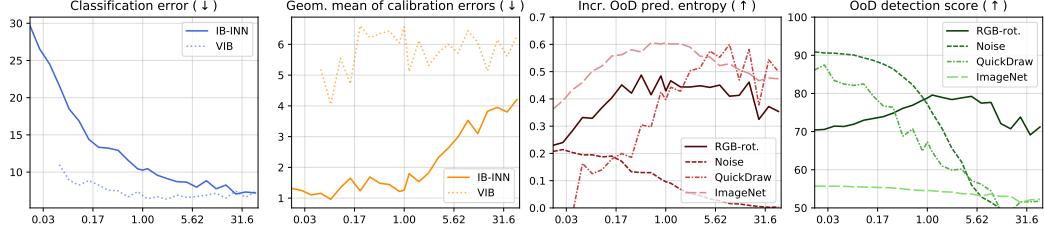


Figure 6: Effect of changing the parameter γ between 0.02 and 50 (logarithmic x -axis) on the different performance measures (y -axis). The left two plots show the IB-INN and VIB, the right two plots only show the IB-INN. The VIB does not converge for $\gamma < 0.05$. The arrows indicate if a larger or smaller score is better. While classification accuracy improves with γ , the uncertainty measures generally grow worse. The trend of OoD detection and OoD entropy is less clear, and depends on the OoD dataset. The special case $\beta = 1$ (class-NLL) translates to $\gamma \approx 3 \cdot 10^{-4}$ (cf. Table 1).

mean of all three: $\sqrt[3]{\text{ECE} \cdot \text{MCE} \cdot \text{ICE}}$. The geometric mean is used because it properly accounts for the different magnitudes of the metrics. Exact definitions found in appendix.

Increased out-of-distribution (OoD) prediction entropy: For data that is OoD, we expect from a model that it returns uncertain class predictions, as it has not been trained on such data. In the ideal case, each class is assigned the same probability of $1/(\text{nr. classes})$. Ovadia et al. (2019) quantify this through the discrete entropy of the class prediction outputs $H(Y|X_{\text{Ood}})$. To counteract the effect of less accurate models having higher prediction entropy overall, we report the difference between OoD and in-distribution test set $H(Y|X_{\text{Ood}}) - H(Y|X_{\text{In distrib.}})$.

OoD detection score: We use OoD detection capabilities intrinsically built in to GCs. For this, we apply the recently proposed typicality test (Nalisnick et al., 2019a). This is a hypothesis test that sets an upper and lower threshold on the estimated likelihood, beyond which batches of inputs are classified as OoD. We apply the test to single input images (i.e. batch size 1). For quantification, we vary the detection threshold to produce a receiver operator characteristic (ROC), and compute the area under this curve (ROC-AUC) in percent. For short, we call this the *OoD detection score*. It will be 100 for perfectly separated in- and outliers, and 50 if each point is assigned a random likelihood.

OoD datasets: The inlier dataset consist of CIFAR10/100 images, i.e. 32×32 colour images showing 10/100 object classes. Additionally, we created four different OoD datasets, that cover different aspects, see Fig. 5. Firstly, we create a random 3D rotation matrix with a rotation angle of $\alpha = 0.3\pi$, and apply it to the RGB color vectors of each pixel of CIFAR10 images. Secondly, we add random uniform noise with a small amplitude to CIFAR10 images, as an alteration of the image statistics. Thirdly, we use the QuickDraw dataset of hand drawn objects (Ha & Eck, 2018), and filter only the categories corresponding to CIFAR10 classes and color each grayscale line drawing randomly. Therefore the semantic content is the same, but the image modality is different. Lastly, we down-scale the ImageNet validation set to 32×32 pixels. In this case, the semantic content is different, but the image statistics are very similar to CIFAR10.

4.3 Results

Quantitative Model Comparison A comparison of all models is performed in Table 1 for CIFAR10, and in the appendix for CIFAR100. At the extreme $\gamma \rightarrow \infty$, the model behaves almost identically to a standard feed forward classifier using the same architecture (i-RevNet), and for $\gamma = 0$, it closely mirrors a conventionally trained GC, as the bits/dim are the same. We find the most favourable setting to be at $\gamma = 1$: Here, the classification error and the bits/dim each only suffer a 10% penalty compared to the extremes. The uncertainty quantification for IB-INN at this setting (calibration and

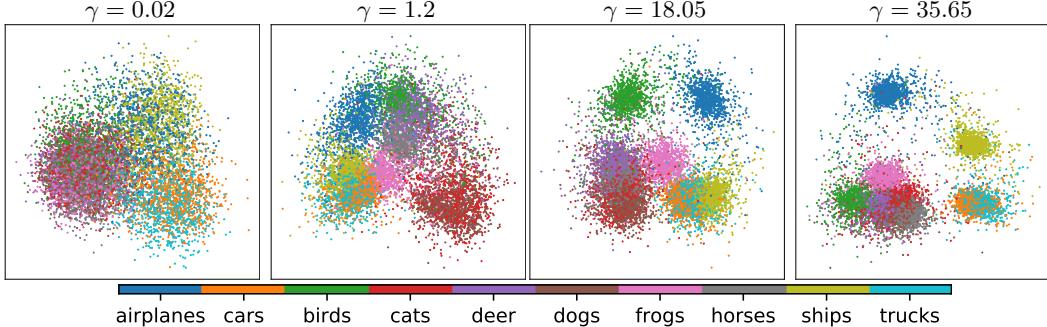


Figure 7: GMM Latent space behaviour by increasing γ . The class separation increases with larger γ . Note that ambiguous classes (e.g. truck and car) remain connected to account for uncertainty.

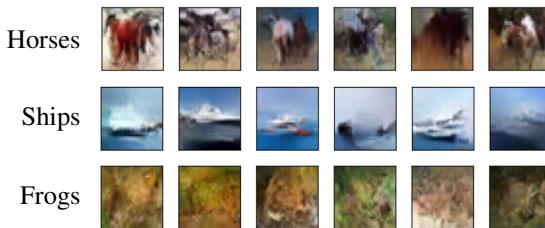


Figure 8: Images are generated for three different classes, by sampling from the respective mixture component in latent space, and inverting the network (more examples in Appendix). This gives insight what happens during classification, see text: only textures are generated for the frog class, indicating that this is the only aspect used for classification.



Figure 9: The columns show a latent space interpolation between two images (leftmost and rightmost). Each row shows a model with a different γ .

OoD prediction entropy) is far better than for pure DCs. Against expectations, standard GCs have worse calibration error. Our hypothesis is that their predictions are too noisy and inaccurate for a positive effect to be visible. For OoD detection, the IB-INN and standard GCs are all comparable, as we would expect from the similar bits/dim. Fig. 6 shows the trade-off between the two extremes in more detail: at low γ , the OoD detection and uncertainty quantification are improved, at the cost of classification accuracy. The VIB behaves in agreement with the other DCs: it has consistently lower classification error but higher calibration error than the IB-INN. This confirms that the IB-INN’s behaviour is due to the application of IB to GCs exclusively. This does not mean that the IB-INN should be preferred over VIB, or vice versa. The main advantages of the VIB are the increased robustness to overfitting and adversarial attacks, aspects that we do not examine in this work.

Latent Space Exploration To better understand what the IB-INN learns, we analyze the latent space in different ways. Firstly, Fig. 7 shows the layout of the latent space GMM through a linear projection. We find that the clusters of ambiguous classes, e.g. truck and car, are connected in latent space, to account for uncertainty. Secondly, Fig. 9 shows interpolations in latent space between two test set images, using models trained with different values of γ . We observe that for low γ , the IB-INN has a well structured latent space, leading to good generative capabilities and plausible interpolations. For larger γ , class separation increases and interpolation quality continually degrades. Finally, generated images can give insight into the classification process, visualizing how the model understands each class. If a certain feature is not generated, this means it does not contribute positively to the likelihood, and in turn will be ignored for classification. Examples for this are shown in Fig. 8.

5 Conclusions

We addressed the application of the Information Bottleneck (IB) as a loss function to Invertible Neural Networks (INNs) trained as generative models. We find that we can formulate an asymptotically exact version of the IB, which results in an INN that is a generative classifier. From our experiments, we conclude that the IB-INN provides high quality uncertainties and out-of-distribution detection, while reaching almost the same classification accuracy as standard feed-forward methods on CIFAR10 and CIFAR100.

Acknowledgements

LA received funding by the Federal Ministry of Education and Research of Germany project High Performance Deep Learning Framework (No 01IH17002). RM received funding from the Robert Bosch PhD scholarship. UK and CR received financial support from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (grant agreement No 647769). We thank the Center for Information Services and High Performance Computing (ZIH) at Dresden University of Technology for generous allocations of computation time. Furthermore we thank our colleagues (in alphabetical order) Tim Adler, Felix Draxler, Clemens Fruböse, Jakob Kruse, Titus Leistner, Jens Müller and Peter Sorrenson for their help and fruitful discussions.

References

- Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2897–2905, 2018. doi: 10.1109/TPAMI.2017.2784440. URL <https://doi.org/10.1109/TPAMI.2017.2784440>.
- Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=HyxQzBceg>.
- Amjad, R. A. and Geiger, B. C. How (not) to train your neural network using the information bottleneck principle. *arXiv preprint arXiv:1802.09766v3*, 2018. URL <http://arxiv.org/abs/1802.09766v3>.
- Ardizzone, L., Lüth, C., Kruse, J., Rother, C., and Köthe, U. Guided image generation with conditional invertible neural networks. *CoRR*, abs/1907.02392, 2019. URL <http://arxiv.org/abs/1907.02392>.
- Barber, D. and Agakov, F. V. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.
- Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R., and Jacobsen, J.-H. Understanding and mitigating exploding inverses in invertible neural networks. *arXiv preprint arXiv:2006.09347*, 2020.
- Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. C. MINE: mutual information neural estimation. *CoRR*, abs/1801.04062, 2018. URL <http://arxiv.org/abs/1801.04062>.
- Bishop, C. and Lasserre, J. Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, 8, January 2007. URL <https://www.microsoft.com/en-us/research/publication/generative-discriminative-getting-best-worlds/>.
- Bishop, C. M. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007. ISBN 9780387310732.
- Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.
- Bouchard, G. and Triggs, B. The tradeoff between generative and discriminative classifiers. In *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, pp. 721–728, 2004.
- Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for gaussian variables. *J. Mach. Learn. Res.*, 6:165–188, 2005. URL <http://jmlr.org/papers/v6/chechik05a.html>.
- Chongxuan, L., Xu, T., Zhu, J., and Zhang, B. Triple generative adversarial nets. In *Advances in neural information processing systems*, pp. 4088–4098, 2017.
- Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.
- De Bruijn, N. G. *Asymptotic methods in analysis*, volume 4. Courier Corporation, 1981.

- Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=HkpbnH9lx>.
- Fetaya, E., Jacobsen, J., and Zemel, R. S. Conditional generative models are not robust. *CoRR*, abs/1906.01171v1, 2019. URL <http://arxiv.org/abs/1906.01171v1>.
- Friedman, N., Mosenzon, O., Slonim, N., and Tishby, N. Multivariate information bottleneck. *CoRR*, abs/1301.2270, 2013. URL <http://arxiv.org/abs/1301.2270>.
- Gilad-Bachrach, R., Navot, A., and Tishby, N. An information theoretic tradeoff between complexity and accuracy. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, pp. 595–609, 2003. doi: 10.1007/978-3-540-45167-9_43. URL https://doi.org/10.1007/978-3-540-45167-9_43.
- Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017. URL <http://proceedings.mlr.press/v70/guo17a.html>.
- Ha, D. and Eck, D. A neural representation of sketch drawings. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL <https://openreview.net/forum?id=Hy6GHpkCW>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL <https://doi.org/10.1109/CVPR.2016.90>.
- Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*, 2018.
- Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.
- Jacobsen, J., Smeulders, A. W. M., and Oyallon, E. i-revnet: Deep invertible networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL <https://openreview.net/forum?id=HJsjkMb0Z>.
- Jacobsen, J., Behrmann, J., Zemel, R. S., and Bethge, M. Excessive invariance causes adversarial vulnerability. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL <https://openreview.net/forum?id=BkfbpsAcF7>.
- Jaini, P., Selby, K. A., and Yu, Y. Sum-of-squares polynomial flow. *arXiv preprint arXiv:1905.02325*, 2019.
- Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 10236–10245, 2018. URL <http://papers.nips.cc/paper/8224-glow-generative-flow-with-invertible-1x1-convolutions>.
- Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.
- Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. Nonlinear information bottleneck. *CoRR*, abs/1705.02436, 2017. URL <http://arxiv.org/abs/1705.02436>.

- L'Ecuyer, P. Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41(4):738–747, 1995.
- Li, Y., Bradshaw, J., and Sharma, Y. Are generative classifiers more robust to adversarial attacks? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 3804–3814, 2019. URL <http://proceedings.mlr.press/v97/li19a.html>.
- Nalisnick, E., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 2019a.
- Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019b. URL <https://openreview.net/forum?id=H1xwNhCcYm>.
- Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. Hybrid models with deep and invertible features. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 4723–4732, 2019c. URL <http://proceedings.mlr.press/v97/nalisnick19b.html>.
- Newey, W. K. and McFadden, D. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.
- Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pp. 841–848, 2001. URL <http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes>.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *CoRR*, abs/1906.02530, 2019. URL <http://arxiv.org/abs/1906.02530>.
- Raina, R., Shen, Y., Mccallum, A., and Ng, A. Y. Classification with hybrid generative/discriminative models. In *Advances in neural information processing systems*, pp. 545–552, 2004.
- Schott, L., Rauber, J., Bethge, M., and Brendel, W. Towards the first adversarially robust neural network model on MNIST. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL <https://openreview.net/forum?id=S1EH0sC9tX>.
- Serfling, R. J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.
- Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theor. Comput. Sci.*, 411(29-30):2696–2711, 2010. doi: 10.1016/j.tcs.2010.04.006. URL <https://doi.org/10.1016/j.tcs.2010.04.006>.
- Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017. URL <http://arxiv.org/abs/1703.00810>.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308. URL <https://doi.org/10.1109/CVPR.2016.308>.
- Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M. Coupling-based invertible neural networks are universal diffeomorphism approximators. *arXiv preprint arXiv:2006.11469*, 2020.

- Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.
- Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, pp. 1–5, 2015. doi: 10.1109/ITW.2015.7133169. URL <https://doi.org/10.1109/ITW.2015.7133169>.
- Tishby, N., Pereira, F. C. N., and Bialek, W. The information bottleneck method. *CoRR*, physics/0004057, 2000. URL <http://arxiv.org/abs/physics/0004057>.
- Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pp. 3835–3844, 2018.
- Wu, T., Fischer, I., Chuang, I., and Tegmark, M. Learnability for the information bottleneck. In *ICLR 2019 Workshop LLD*, 2019.
- Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.
- Xue, J. and Titterington, D. M. On the generative-discriminative tradeoff approach: Interpretation, asymptotic efficiency and classification performance. *Computational Statistics & Data Analysis*, 54(2):438–451, 2010. doi: 10.1016/j.csda.2009.09.011. URL <https://doi.org/10.1016/j.csda.2009.09.011>.
- Yogatama, D., Dyer, C., Ling, W., and Blunsom, P. Generative and discriminative text classification with recurrent neural networks. *CoRR*, abs/1703.01898, 2017. URL <http://arxiv.org/abs/1703.01898>.

– Appendix –

Contents

A Proofs and Derivations	14
A.1 Assumptions	14
A.2 Mutual Cross-Information as Estimator for MI	14
A.3 Loss Function \mathcal{L}_X	15
A.4 Density Error through Noise Augmentation	18
B Practical Loss Implementation	19
C Calibration Error Measures	20
D Additional Experiments	20
D.1 Choice of σ	20
D.2 CIFAR100	21
D.3 Further experiments	21
E Network Architecture	22

A Proofs and Derivations

A.1 Assumptions

Assumption 1. We assume that the sample space \mathcal{X} belonging to the input RV $X : \mathcal{X} \rightarrow \mathbb{R}$ is a compact domain in \mathbb{R}^d , and that $p(X | y)$ is absolutely continuous $\forall y \in \mathcal{Y}$, where \mathcal{Y} is the set of available classes.

The compactness of \mathcal{X} is the major aspect here. However, this is always fulfilled for image data, as the pixels can only take certain range of values, and equally fulfilled for most other real-world datasets, as data representations, measurement devices, etc. only have a finite range.

Assumption 2. We assume g_θ is from a family of universal density estimators, as defined by Definition 3 in Teshima et al. (2020). Moreover, we assume the network parameter space Θ is a compact subdomain of \mathbb{R}^n , g_θ and J_θ are uniformly bounded, and that the lower bound of $|\det J_\theta|$ is > 0 . We also assume that g_θ and J_θ are continuous and differentiable in both X and θ .

This is a fairly mild set of assumptions, as it is fulfilled by construction with most existing INN architectures using standard multi-layer subnetworks. See e.g. Behrmann et al. (2020); Virmaux & Scaman (2018) for details. Specifically, it holds for our tanh-clamped coupling block design (see Appendix E). Note that some properties directly follow from Assumption 2: Firstly, as J_θ is uniformly bounded, this implies that g_θ is uniformly Lipschitz-continuous. Second, using Assumption 1, the domain of $Z = g_\theta(X)$ is compact, and $p(Z)$ is absolutely continuous.

A.2 Mutual Cross-Information as Estimator for MI

In our case, we only require $CI(X, Z_{\mathcal{E}})$ and $CI(Y, Z_{\mathcal{E}})$, but we show the correspondence for two unspecified random variables U, V , as it may be of general interest. However, note that our estimator will likely not be particularly useful outside of our specific use-case, and other methods should be preferred (e.g. MINE, Belghazi et al., 2018). Our approach has the specific advantage, that we estimate the MI of the model using the model itself. For e.g. MINE, we would require three models,

one generative model, and two models that only serve to estimate the MI. Secondly, it is not clear how the large constant $d \log(\sigma)$ can be cancelled out using other approaches.

For the joint input space $\Omega = \mathcal{U} \times \mathcal{V}$, we assume that \mathcal{U} is a compact domain in \mathbb{R}^d , and \mathcal{V} is either also a compact domain in \mathbb{R}^l (Case 1), or discrete, i.e. a finite subset of \mathbb{N} (Case 2). In Case 1, we assume that $p(U, V)$ is absolutely continuous with respect to the Lebesgue measure, and in Case 2, $p(U|v)$ is absolutely continuous for all values of $v \in \mathcal{V}$. This is in agreement with Assumption 1.

In Case 1, $q(U)$, $q(V)$, $q(U, V)$, the densities can all be modeled separately, by three flow networks $g_\theta^{(U)}(u)$, $g_\theta^{(V)}(v)$, $g_\theta^{(UV)}(u, v)$. Although in our formulation, we are later able to approximate the latter two through the first.

In Case 2, we only model $q(U|V)$, and assume that $q(V)$ is either known beforehand and set to $p(V)$ (e.g. label distribution), or the probabilities are parametrized directly. Either way, $q(U, V) = q(U|V)q(V)$ and $q(U) = \sum_{v \in \mathcal{V}} q(U, v)$.

Proposition 1. *Assume that the $q(\cdot)$ densities can be chosen from a sufficiently rich model family (e.g. a universal density estimator). Then for every $\eta > 0$ there is a model such that*

$$|I(U, V) - CI(U, V)| < \eta \quad (10)$$

and $I(U, V) = CI(U, V)$ if $p(U, V) = q(U, V)$.

Proof. Writing out the definitions explicitly, and rearranging, we find

$$\begin{aligned} CI(U, V) &= I(U, V) + D_{\text{KL}}(p(U, V) \| q(U, V)) \\ &\quad - D_{\text{KL}}(p(U) \| q(U)) - D_{\text{KL}}(p(V) \| q(V)) \end{aligned} \quad (11)$$

Shortening the KL terms to D_1 , D_2 and D_3 for convenience:

$$|CI(U, V) - I(U, V)| = |D_1 - D_2 - D_3| \quad (12)$$

$$\leq D_1 + D_2 + D_3 \quad (13)$$

$$\leq 3 \max(D_1, D_2, D_3) \quad (14)$$

At this point, we can simply apply results from measure transport: if the g_θ are from a family of universal density estimators, we can choose θ^* to make $\max(D_1, D_2, D_3)$ arbitrarily small by matching p and q . This was shown in general for increasing triangular maps, e.g. in Hyvärinen & Pajunen (1999), Theorem 1 for an accessible proof, or Bogachev et al. (2005) for a more in-depth approach (specifically Corollary 4.2). Generality was also proven for several concrete architectures, e.g. Teshima et al. (2020); Jaini et al. (2019); Huang et al. (2018).

For the second part of the Proposition, we note the following: if $p(U, V) = q(U, V)$, we have $D_1 = D_2 = D_3 = 0$, and therefore $CI(U, V) = I(U, V)$. \square

A.3 Loss Function \mathcal{L}_X

In the following, we use the subscript-notation for the cross entropy:

$$h_q(U) = \mathbb{E}_{u \sim p(U)} [-\log q(u)], \quad (15)$$

to avoid confusion with the joint entropy that arises with the usual notation ($h(p(U), q(U))$).

Proposition 2. *For the case given in the paper, that $Z_\mathcal{E} = g_\theta(X + \mathcal{E})$, it holds that $I(X, Z_\mathcal{E}) \leq CI(X, Z_\mathcal{E})$.*

Proof. In the following, we first use the invariance of the (cross-)information to homeomorphic transforms (see e.g. Cover & Thomas (2012) Sec. 8.6). Then, we use $p(X + \mathcal{E}|X) = q(X + \mathcal{E}|X) = p(\mathcal{E})$ (known exactly) and write out all the terms, most of which cancel. Finally, we use the inequality that the cross entropy is larger than the entropy, $h_q(U) \geq h(U)$ regardless of q . The equality holds iff the two distributions are the same.

$$CI(X, Z_\mathcal{E}) - I(X, Z_\mathcal{E}) = CI(X, X + \mathcal{E}) - I(X, X + \mathcal{E}) \quad (16)$$

$$= h_q(X) - h(X) + 0 \quad (17)$$

$$\geq 0 \quad (18)$$

With equality iff $p(X) = q(X)$. \square

We now want to show that the network optimization procedure that arises from the empirical loss, in particular the gradients w.r.t. network parameters θ , are consistent with those of $CI(X, Z_{\mathcal{E}})$:

Proposition 3. *The defined loss is a consistent estimator for $CI(X, Z_{\mathcal{E}})$ up to a known constant, and a consistent estimator for the gradients. Specifically, for any $\epsilon_1, \epsilon_2 > 0$ and $0 < \delta < 1$ there are $\sigma_0 > 0$ and $N_0 \in \mathbb{N}$, such that $\forall N \geq N_0$ and $\forall \sigma < \sigma_0$,*

$$\Pr \left(\left| CI(X, Z_{\mathcal{E}}) + d \log \sqrt{2\pi e \sigma^2} - \mathcal{L}_X^{(N)} \right| < \epsilon_1 \right) > 1 - \delta$$

and

$$\Pr \left(\left\| \frac{\partial}{\partial \theta} CI(X, Z_{\mathcal{E}}) - \frac{\partial}{\partial \theta} \mathcal{L}_X^{(N)} \right\| < \epsilon_2 \right) > 1 - \delta$$

holds uniformly for all model parameters θ .

The loss function is as defined in the paper:

$$\mathcal{L}_X = h_q(Z_{\mathcal{E}}) - \mathbb{E}_{x \sim p(X+\mathcal{E})} \left[\log |\det J_{\theta}(x)| \right] \quad (19)$$

as well as its empirical estimate using N samples, $\mathcal{L}_X^{(N)}$.

We split the proof into two Lemmas, which we will later combine.

Lemma 1. *For any $\eta_1, \eta_2 > 0$ and $\delta > 0$ there is an $N_0 \in \mathbb{N}$ so that*

$$\Pr \left(|\mathcal{L}_X^{(N)} - \mathcal{L}_X| < \eta_1 \right) > 1 - \delta \quad (20)$$

$$\Pr \left(\left| \frac{\partial}{\partial \theta} \mathcal{L}_X^{(N)} - \frac{\partial}{\partial \theta} \mathcal{L}_X \right| < \eta_2 \right) > 1 - \delta \quad (21)$$

$$\forall N \geq N_0$$

Proof. For the first part (Eq. 20), we simply have to show that the uniform law of large numbers applies, specifically that all expressions in the expectations are bounded and change continuously with θ . For the Jacobian term in the loss, this is the case by definition. For the $h_q(Z_{\mathcal{E}})$ -term, we can show the boundedness of $\log q$ occurring in the expectation by inserting the GMM explicitly. We find

$$-\log(q(z)) \leq \max_y [(z - \mu_y)^2 / 2] + \text{const.} \quad (22)$$

while we know that $z = g_{\theta}(x)$ is bounded. Therefore, the uniform law of large numbers (Newey & McFadden, 1994, Lemma 2.4) guarantees existence of an N_1 to satisfy the condition for all $\theta \in \Theta$.

For the second part (Eq. 21), we will show that the gradient w.r.t. θ and the expectation can be exchanged, as the gradient is also bounded by the same arguments as before. We find that the conditions for exchanging expectation and gradient are trivially satisfied, again due to the bounded gradients (see L'Ecuyer (1995), assumption A1, with Γ set to the upper bound). This results in an $N_2 \in \mathbb{N}$ for which Eq. 21 is satisfied. As a last step, we simply define $N_0 := \max(N_1, N_2)$. \square

Lemma 2. *For any $\eta_1, \eta_2 > 0$ there is an $\sigma_0 > 0$, so that*

$$\left\| CI_{\theta}(X, Z_{\mathcal{E}}) + d \log \sqrt{2\pi e \sigma^2} - \mathcal{L}_X \right\| < \eta_2 \quad (23)$$

$$\left\| \frac{\partial}{\partial \theta} (CI_{\theta}(X, Z_{\mathcal{E}}) - \mathcal{L}_X) \right\| < \eta_2 \quad (24)$$

$$\forall \sigma < \sigma_0$$

Proof. In the following proof, we make use of the $O(\cdot)$ notation, see e.g. De Bruijn (1981):

We write $f(\sigma) = O(g(\sigma))$ ($\sigma \rightarrow 0$) iff there exists a σ_0 and an $M \in \mathbb{R}$, $M > 0$ so that

$$\|f(\sigma)\| < M g(\sigma) \quad \forall \sigma \leq \sigma_0. \quad (25)$$

Furthermore, to discuss the limit case, it is necessary we reparametrize the noise variable \mathcal{E} in terms of noise S with a fixed standard normal distribution:

$$\mathcal{E} = \sigma S \quad \text{with} \quad p(S) = \mathcal{N}(0, 1) \quad (26)$$

To begin with, we use the invariance of CI under the homeomorphic transform g_θ . This can be easily verified by inserting the change-of-variables formula into the definition. See e.g. Cover & Thomas (2012) Sec. 8.6. This results in

$$CI(X, Z_{\mathcal{E}}) = CI(Z, Z_{\mathcal{E}}) = h_q(Z_{\mathcal{E}}) - h_q(Z_{\mathcal{E}}|Z) \quad (27)$$

Next, we series expand $Z_{\mathcal{E}}$ around $\sigma = 0$. We can use Taylor's theorem to write

$$Z_{\mathcal{E}} = Z + J_\theta(Z)\mathcal{E} + O(\sigma^2) \quad (28)$$

We have written the Jacobian dependent on Z , but note that it is still $\partial g_\theta / \partial X$, and we simply substituted the argument. We put this into the second entropy term $h_q(Z_{\mathcal{E}}|Z)$ in Eq. 27, and then perform a zero-order von Mises expansion of h_q . In general, the identity is

$$h_q(W + \xi) = h_q(W) + O(\|\xi\|) \quad (\|\xi\| \rightarrow 0), \quad (29)$$

and we simply put $\xi = O(\sigma^2)$ (the identity applies in the same way to the *conditional* cross-entropy). Intuitively, this is what we would expect: the entropy of an RV with a small perturbation should be approximately the same without the perturbation. See e.g. Serfling (2009), Sec. 6 for details. Effectively, this allows us to write the residual outside the entropy:

$$h_q(Z_{\mathcal{E}}|Z) = h_q(Z + J_\theta(Z)\mathcal{E} + O(\sigma^2)|Z) \quad (30)$$

$$= h_q(Z + J_\theta(Z)\mathcal{E}|Z) + O(\sigma^2) \quad (31)$$

$$= h_q(J_\theta(Z)\mathcal{E}|Z) + O(\sigma^2) \quad (32)$$

At this point, note that $q_\theta(J_\theta(Z)\mathcal{E}|Z)$ is simply a multivariate normal distribution, due to the conditioning on Z . In this case, we can use the entropy of a multivariate normal distribution, and simplify to obtain the following:

$$-h_q(J_\theta\mathcal{E}|Z) = \mathbb{E} \left[\frac{1}{2} \log (\det(2\pi\sigma^2 J_\theta J_\theta^T)) \right] \quad (33)$$

$$= \mathbb{E} \left[\frac{1}{2} \log ((2\pi\sigma^2)^d \det(J_\theta)^2) \right] \quad (34)$$

$$= d \log \sqrt{2\pi e\sigma^2} + \mathbb{E} [\log |\det J_\theta|]. \quad (35)$$

Here, we exploited the fact that $J_\theta(Z)$ is an invertible matrix, and used $d = \dim(Z)$. Finally, as in practice we only want to evaluate the model once, we use the differentiability of J_θ to replace

$$\mathbb{E} [\log |\det J_\theta(Z)|] = \mathbb{E} [\log |\det J_\theta(Z_{\mathcal{E}})|] + O(\sigma). \quad (36)$$

The residual can be written outside of the expectation as we know it is bounded from our assumptions about g_θ and J_θ (Dominated Convergence theorem).

Putting the terms together, we obtain

$$CI(X, Z_{\mathcal{E}}) = h_q(Z_{\mathcal{E}}) - d \log \sqrt{2\pi e\sigma^2} - \mathbb{E} [\log |\det J_\theta|] + O(\sigma) \quad (37)$$

$$= \mathcal{L}_X - d \log \sqrt{2\pi e\sigma^2} + O(\sigma) \quad (38)$$

Through the definition of $O(\cdot)$, Eq. 23 is satisfied. To show that the gradients also agree (Eq. 24), we must ensure that the $O(\sigma)$ term is uniformly convergent to 0 over θ , i.e. there is a single constant M in the definition of $O(\cdot)$ that applies for all $\theta \in \Theta$. This is directly the case, as g_θ is Lipschitz continuous and the outputs are bounded (Arzela - Ascoli theorem). \square

We can now combine the two Lemmas 1 and 2, to show Proposition 3.

Proposition 3 - Proof.

Proof. The Proposition follows directly from Lemmas 1 and 2: for a given ϵ_1, ϵ_2 and δ , we choose each $\eta_i = \epsilon_i/2$, and apply the triangle inequality, meaning there exists an N_0 and σ_0 so that

$$\begin{aligned} & \left| CI(X, Z_{\mathcal{E}}) + d \log \sqrt{2\pi e \sigma^2} - \mathcal{L}_X^{(N)} \right| \\ & \leq \left| CI(X, Z_{\mathcal{E}}) + d \log \sqrt{2\pi e \sigma^2} - \mathcal{L}_X \right| + \left| \mathcal{L}_X - \mathcal{L}_X^{(N)} \right| \\ & \quad < \frac{\epsilon_1}{2} + \frac{\epsilon_1}{2} \end{aligned}$$

And therefore $\Pr(\dots) > 1 - \delta$. Equivalently for the gradients.

□

A.4 Density Error through Noise Augmentation

For the derivation of the losses, we only assumed that X and $X + \mathcal{E} =: X_{\mathcal{E}}$ are both RVs on a domain \mathcal{X} , and required no further assumptions about a possible quantization of X . However, if X is quantized, which is mostly the case in practice, we can exploit this fact to derive a bound on the additional modeling error caused by the augmentation. To demonstrate this, we introduce the discrete, quantized data W . This is essentially the same as X , but is only defined on a finite, discrete set \mathcal{W} . With F regular quantization steps in each of the d dimensions, spaced by the quantization step size ΔX , we write

$$\mathcal{W} = \{0, 1\Delta X, 2\Delta X, \dots, (F-1)\Delta X\}^d \subset \mathcal{X}, \quad (39)$$

We denote probabilities of this discrete variable as upper case P and Q for true and modeled probabilities, respectively. We index the finite number of elements in \mathcal{W} as w_i . For convenience, we also introduce the following notation:

$$P(w_i) =: P_i \quad Q(w_i) =: Q_i. \quad (40)$$

Furthermore, we denote the noise distribution used for augmentation as $r(\mathcal{E})$ in the following, as this simplifies the notation and avoids ambiguities (it was denoted $p(\mathcal{E})$ instead for the loss derivation). From this, we can see how the distribution $p(X_{\mathcal{E}})$, which is used to train the network, can be expressed in terms of $P(W)$ and $r(\mathcal{E})$:

$$p(X_{\mathcal{E}}) = \sum_i P_i r(X_{\mathcal{E}} - w_i) \quad (41)$$

At test time, we want to recover an estimate Q_i . For standard normalizing flows, this is generally computed as

$$\tilde{Q}_i := \frac{q(X_{\mathcal{E}} = w_i)}{r(0)} \quad (42)$$

Among other things, this is used to measure the bits/dim. In the most general case, \tilde{Q} will not sum to 1, so it is not guaranteed to be a valid probability, indicated by the tilde. Nevertheless, we can see why this definition is sensible by considering the noise distribution r used by most normalizing flows: hereby the support of r in each dimension is smaller or equal to the quantization step size. Then, only one term in the sum in Eq. 41 is $\neq 0$ at any point. As a result, we obtain

$$q(X_{\mathcal{E}}) = p(X_{\mathcal{E}}) \implies \tilde{Q}(W) = P(W). \quad (43)$$

This means that in principle a standard normalizing flow can learn the true underlying discrete distribution from the noisy augmented distribution. In other words, the augmentation process does not introduce an additional error to the density estimation.

We now apply these definitions to our setting of a Gaussian noise distribution, $r(\mathcal{E}) = \mathcal{N}(0, \sigma^2 \mathbb{I})$. We consider the case where the model learns the training data distribution perfectly, i.e. $q(X_{\mathcal{E}}) = p(X_{\mathcal{E}})$. We find that Eq. 43 no longer holds for the Gaussian case, but that the error between $\tilde{Q}(W)$ and $P(W)$ has a known bound that decreases exponentially for small σ . For convenience, we write

$A := \mathcal{N}(0; 0, \sigma^2 \mathbb{I}) = (2\pi\sigma^2)^{-d/2}$. From this, we get

$$\tilde{Q}_j = \frac{q(X_{\mathcal{E}} = w_j)}{A} = \frac{p(X_{\mathcal{E}} = w_j)}{A} \quad (44)$$

$$= \frac{1}{A} \sum_i P_i \mathcal{N}(w_j - w_i; 0, \sigma^2 \mathbb{I}) \quad (45)$$

$$= \frac{P_j \mathcal{N}(0; 0, \sigma^2 \mathbb{I})}{A} + \frac{1}{A} \sum_{i \neq j} P_i \mathcal{N}(w_j - w_i; 0, \sigma^2 \mathbb{I}) \quad (46)$$

$$= P_j + \underbrace{\frac{1}{A} \sum_{i \neq j} P_i \mathcal{N}(w_j - w_i; 0, \sigma^2 \mathbb{I})}_{:= \Delta P_j} \quad (47)$$

We are now interested in determining a bound for the error ΔP_j . Because $\|w_i - w_j\| \geq \Delta X$ for $i \neq j$, we know

$$\mathcal{N}(w_i - w_j; 0, \sigma^2 \mathbb{I}) \leq A \exp\left(-\frac{\Delta X^2}{2\sigma^2}\right). \quad (48)$$

From that, we obtain the following bound:

$$\Delta P_j \leq \left(\sum_{i \neq j} P_i\right) \frac{1}{A} A \exp\left(-\frac{\Delta X^2}{2\sigma^2}\right) \quad (49)$$

$$\leq \exp\left(-\frac{\Delta X^2}{2\sigma^2}\right) \quad (50)$$

B Practical Loss Implementation

In the following, we provide the explicit loss implementations, as there are some considerations to make with regards to numerical tractability. Specifically, we make use of the operations `softmax`, `log_softmax`, `logsumexp` provided by major deep learning frameworks, as they avoid the most common pitfalls.

The class probabilities $q(Y)$ can be characterized through a vector Φ , with

$$q(y) = \text{softmax}_y(\Phi), \quad (51)$$

where the subscript of the softmax operator denotes which index is selected for the enumerator. The use of the softmax ensures that w_y stay positive and sum to one. For our work, $q(Y) = p(Y)$ is known beforehand, so we leave Φ fixed to 0 (equal probability for each class). However, we also find it is possible to learn Φ as a free parameter during training. In this case, only the gradients of the \mathcal{L}_X loss w.r.t. Φ should be taken, as the \mathcal{L}_Y loss is no longer a lower bound, and can be exploited by sending $\Phi_y \rightarrow \infty$ for some fixed y , and $\Phi_k \rightarrow -\infty$ for all $k \neq y$. If only \mathcal{L}_X is backpropagated w.r.t. Φ , this is avoided and Φ converges to the correct class weights. We use the shorthand $w_y := \log p(y)$ in the following.

With $z := g_\theta(x + \varepsilon)$, we also have

$$\bullet \log q(y) = w_y = \text{logsoftmax}_y(\Phi) \quad (52)$$

$$\bullet \log q(z|y) = -\frac{1}{2} \|z - \mu_y\|^2 + \text{const.} \quad (53)$$

$$\bullet \log q(z) = \text{logsumexp}_{y'}\left(-\frac{\|z - \mu_{y'}\|^2}{2} + w_{y'}\right) + \text{const.} \quad (54)$$

With this, the loss functions are evaluated as

$$\mathcal{L}_X(x) = \text{logsumexp}_{y'}\left(\frac{\|z - \mu'_y\|^2}{2} - w_{y'}\right) - \log J(x) \quad (55)$$

$$\mathcal{L}_Y(x, y) = \text{logsoftmax}_y\left(-\frac{\|z - \mu_{y'}\|^2}{2} + w_{y'}\right) - w_y. \quad (56)$$

The constants have been dropped for convenience. The use of the `logsumexp` and `logsoftmax` operations above is especially important. Otherwise when explicitly performing the `exp` and `log` operations with 32 bit floating point numbers, the values become too large, and the loss numerically ill-defined (`NaN`).

C Calibration Error Measures

In the following, we make use of the Iverson bracket:

$$[C] := \begin{cases} 1 & \text{if } C \text{ is true;} \\ 0 & \text{otherwise,} \end{cases} \quad (57)$$

Firstly, we define the bin edges b_i , with $i \in \{1, \dots, K+1\}$, so that $b_1 = 0$, $b_{K+1} = 1$, and $b_{i+1} > b_i$. In practice, we choose the b_i be spaced more tightly near high and low confidences, as this is where the bulk of the predictions are made:

```
concatenate(range(0.00, 0.05, stepsize=0.01),
            range(0.05, 0.95, stepsize=0.1),
            range(0.95, 1.00, stepsize=0.01))
```

The bins themselves are then half-open intervals between the bin edges: $B_i = [b_i, b_{i+1})$ with $i \in \{1, \dots, K\}$. We now define $n^{(i)}$, the count of predictions within a confidence bin; as well as $n_c^{(i)}$, the count of *correct* predictions in that bin:

$$n^{(i)} := \sum_{x_j} \sum_{y'} [p(y'|x_j) \in B_i] \quad (58)$$

$$n_c^{(i)} := \sum_{(x_j, y_j)} \sum_{y'} [p(y'|x_j) \in B_i] \cdot [\arg \max_{y'} (p(y'|x_j)) = y_j] \quad (59)$$

where x_j and the (x_j, y_j) -pairs are from the test set.

We define the confidence P as the center of each bin, and the achieved accuracy in this bin as Q :

$$P_i = \frac{b_i + b_{i+1}}{2} \quad (60)$$

$$Q_i = \frac{n_c^{(i)}}{n^{(i)}} \quad (61)$$

Finally, using Q and P , we define the calibration error measures, in agreement with Guo et al. (2017):

$$\text{ECE} = \sum_i \frac{n^{(i)}}{n_{\text{tot}}} |P_i - Q_i| \quad (\text{Expected calib. err.}) \quad (62)$$

$$\text{MCE} = \max_i |P_i - Q_i| \quad (\text{Maximum calib. err.}) \quad (63)$$

$$\text{ICE} = \sum_i (b_{i+1} - b_i) |P_i - Q_i| \quad (\text{Integrated calib. err.}) \quad (64)$$

using the shorthand $n_{\text{tot}} := \sum_i n^{(i)}$.

D Additional Experiments

D.1 Choice of σ

Fig. 10 shows the behaviour for 25 different models trained with σ between 10^{-4} and 10^0 (x-axis), and fixed $\gamma = 0.2$. We find that the loss values (left) and performance characteristics (middle) do not depend on σ below a threshold that is about a factor 4 smaller than the quantization step size ΔX . Contrary to expectations from existing work on normalizing flows, the models performance does not

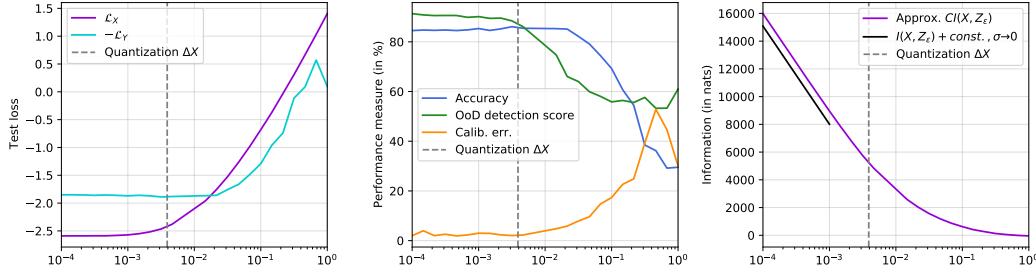


Figure 10: From left to right: Changes in test-loss, performance metrics, and a comparison between approximation and known slope of the true mutual information for varying values of σ (x-axis)

Table 2: Results on the CIFAR100 dataset. All models have the same number of parameters and were trained with the same hyperparameters. All values except entropy and overconfidence are given in percent. The arrows indicate whether a higher or lower value is better.

Model	Classif. err. (↓)	Bits/dim (↓)	Calibration error (↓)			Incr. OoD prediction entropy (↑)				OoD detection score (↑)							
			Geo. mean	ECE	MCE	ICE	Average	RGB-rot	Draw	Noise	ImgNet	Average	RGB-rot	Draw			
IB-ENN (ours)	only \mathcal{L}_X ($\gamma = 0$)	—	4.82	—	—	—	—	—	—	—	—	70.03	63.35	87.45	85.12	50.99	
	$\gamma = 0.1$	42.57	4.94	2.60	0.58	7.04	4.28	0.50	0.66	0.28	0.35	68.31	66.53	78.91	81.70	50.75	
	only \mathcal{L}_Y ($\gamma \rightarrow \infty$)	33.78	18.44	4.49	0.62	16.76	8.72	0.58	0.52	1.04	0.00	0.77	58.29	47.95	99.37	49.23	49.23
Stand. GC	Class-NLL	97.92	4.82	16.20	1.02	95.63	43.53	-0.04	-0.14	0.55	-0.53	-0.03	70.26	64.68	86.54	85.19	51.09
	Class-NLL + regul.	69.28	5.07	13.94	0.75	89.74	40.15	0.00	-0.00	-0.01	0.01	0.01	68.83	64.96	82.19	83.32	50.44
Stand. DC	ResNet	29.27	—	5.13	0.65	20.57	10.16	0.60	0.68	0.97	-0.00	0.74	—	—	—	—	—
	i-RevNet	37.54	—	5.18	0.63	19.85	11.09	0.51	0.32	1.00	-0.00	0.75	—	—	—	—	—

decrease even when σ is 50 times smaller than ΔX . Detrimental effects might occur more easily if the quantization steps are larger, e.g. $\Delta X = 1/32$ as used by Kingma & Dhariwal (2018), or if the model were more powerful or less regularized (e.g. from the tanh-clamping we employ). The rightmost plot compares our approximation of $CI(X, Z_\varepsilon)$ with the asymptotic $I(X, Z_\varepsilon) + \text{const.}$ for $\sigma \rightarrow 0$, where the constant is unknown. The slope of the approximation agrees well for small σ , but breaks down for larger values.

D.2 CIFAR100

Table 2 reports the performance of the models on CIFAR100. The general behaviour observed for CIFAR10 is repeated here: The IB-ENN model which balances both loss terms performs significantly better in terms of uncertainty calibration than both standard GCs and DCs. It also performs OoD detection almost as well as pure GCs, with a much better classification error.

There are two differences compared to the CIFAR10 case: Firstly, in terms of increase in predictive entropy on OoD data, there are much smaller differences between models (excluding the standard GCs). The standard ResNet has the best overall performance by a small margin. Note that the increase in prediction entropy is also influenced by the calibration and overall classification error of the model to some degree, so we are careful in drawing any conclusions from minor differences. Secondly, we find that the most advantageous trade-off regime is now at a lower value of γ . The only values trained for CIFAR100 were $\gamma \in \{0.1, 1, 10\}$, and we find that the models with γ set to 1 and 10 behave almost the same as the limit case $\gamma \rightarrow \infty$. The explanation for this is simple: due to the increased difficulty of the task, the \mathcal{L}_Y loss is higher than for CIFAR10. Therefore, it has a larger influence at the same setting for γ compared to the CIFAR10 models.

D.3 Further experiments

Figure 11 provides all the performance metrics discussed in the paper over the range of γ .

In Figure 12 we show the trajectory of a sample in latent space, when gradually increasing the angle α of the RGB-rotation augmentation used in the paper as an OoD dataset. It travels from in-distribution to out-of-distribution. Such images were never seen during training.

Figure 14 shows samples generated by the model, using different values of γ . In general, we find the quality of generated images degrades faster with γ than the interpolations between existing images. We see indications that the mass of points in latent space is offset from the learned μ_y , meaning

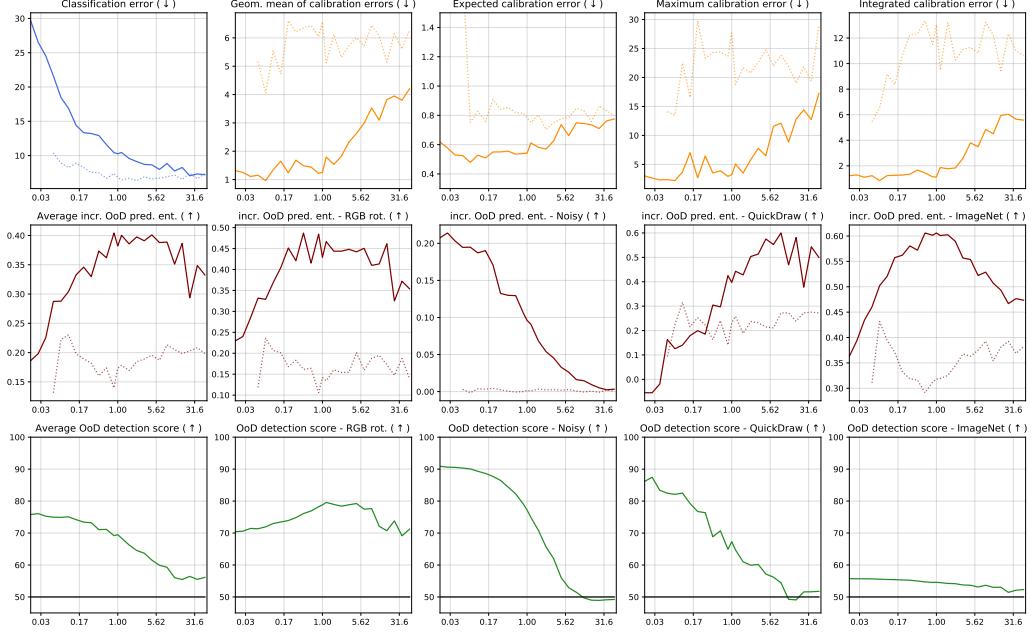


Figure 11: Effect of changing the parameter $\tilde{\beta}$ (x-axis) on the different performance measures (y-axis). The arrows indicate if a larger or smaller score is better. The black horizontal line in the last row indicates random performance. Details are explained in the paper. The VIB results are added as dotted lines. The VIB does not converge reliably for values of $\gamma < 0.2$, producing some outliers e.g. for expected calibration error. This is not to claim that the IB-INN is better than the VIB or vice versa. The comparison serves to show how the IB affects GCs and DCs differently.

the regions that are sampled from have not seen much training data. In contrast to the IB-INN, the standard class-NLL trained model generates fairly generic looking images for all classes, due to the collapse of class-components in latent space.

E Network Architecture

As in previous works, our INN architecture consists of so-called *coupling blocks*. In our case, each block consists of one affine coupling (Dinh et al., 2017), illustrated in Fig. 13, followed by random and fixed soft permutation of channels (Ardizzone et al., 2019), and a fixed scaling by a constant, similar to ActNorm layers introduced by Kingma & Dhariwal (2018). For the coupling coefficients, each subnetwork predicts multiplicative and additive components jointly, as done by Kingma & Dhariwal (2018). Furthermore, we adopt the soft clamping of multiplication coefficients used by Dinh et al. (2017).

For downsampling blocks, we introduce a new scheme, whereby we apply the i-RevNet downsampling (Jacobsen et al., 2018) only to the inputs to the affine transformation (u_2 branch in Fig. 13), while the affine coefficients are predicted from a higher resolution u_1 by using a strided convolution in the corresponding subnetwork. After this, i-RevNet downsampling is applied to the other half of the channels u_1 to produce v_1 , before concatenation and the soft permutation. We adopt this scheme as it more closely resembles the standard ResNet downsampling blocks, and makes the downsampling operation at least partly learnable.

We then stack sets of these blocks, with downsampling blocks in between, in the manner of [8, down, 25, down, 25]. Note, we use fewer blocks for the first resolution level, as the data only has three channels, limiting the expressive power of the blocks at this level. Finally, we apply a discrete cosine transform to replace the global average pooling in ResNets, as introduced by Jacobsen et al. (2019), followed by two blocks with fully connected subnetworks.

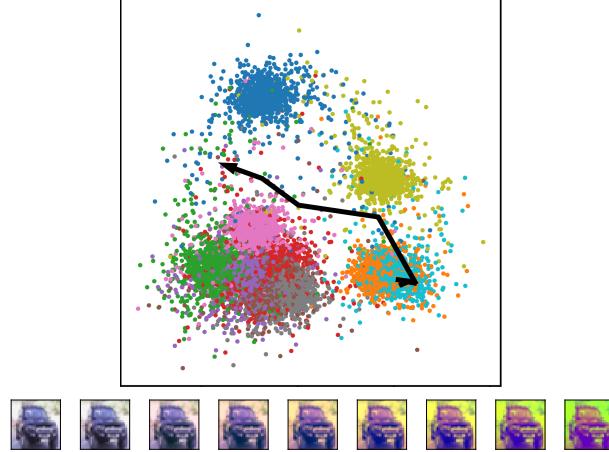
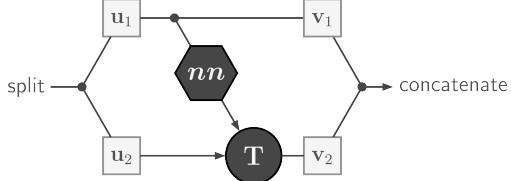


Figure 12: The scatter plot shows the location of test set data in latent space. A single sample is augmented by rotating the RGB color vector as described in the paper. The small images show the successive steps of augmentation, while the black arrow shows the position of each of these steps in latent space. We observe how the points in latent space travel further from the cluster center with increasing augmentation, causing them to be detected as OoD.

Forward computation (left to right):

$$v_1 = u_1, \quad v_2 = T(u_2; nn(v_1))$$



Inverse computation (right to left):

$$u_1 = v_1, \quad u_2 = T^{-1}(v_2; nn(u_1))$$

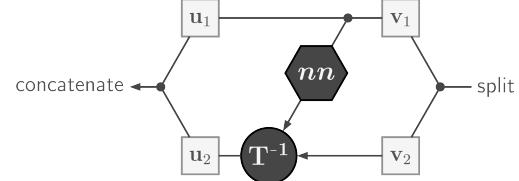


Figure 13: Illustration of a coupling block. T represents some invertible transformation, in our case an affine transformation. The transformation coefficients are predicted by a subnetwork (nn), which contains fully-connected or convolutional layers, nonlinear activations, batch normalization layers, etc., similar to the residual subnetwork in a ResNet (He et al., 2016). Note that how the subnetwork does not have to be inverted itself.

We perform training with SGD, learning rate 0.07, momentum 0.9, and batch size 128, as in the original ResNet publication (He et al., 2016). We train for 450 epochs, decaying the learning rate by a factor of 10 after 150, 250, and 350 epochs.



Figure 14: Samples from four different models, each using the same architecture but a different loss. From top to bottom: class-NLL, $\gamma = 0.02$, $\gamma = 0.2$, $\gamma = 1$. In each subfigure, each row corresponds to one class. The classes from top to bottom are: plane, car, bird, cat, deer, dog, frog, horse, boat, truck.