EXCESSIVE INVARIANCE CAUSES ADVERSARIAL VULNERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their impressive performance, deep neural networks exhibit striking failures on out-of-distribution inputs. One core idea of adversarial example research is to reveal neural network errors under such distribution shift. We decompose these errors into two complementary sources: sensitivity and invariance. We show deep networks are not only *too sensitive* to task-irrelevant changes of their input, as is well-known from ϵ-adversarial examples, but are also *too invariant* to a wide range of task-relevant changes, thus making vast regions in input space vulnerable to adversarial attacks.
After identifying this excessive invariance, we propose the usage of bijective deep networks to enable access to all variations. We introduce metameric sampling as an analytic attack for these networks, requiring no optimization, and show that it uncovers large subspaces of misclassified inputs. Then we apply these networks to MNIST and ImageNet and show that one can manipulate the class-specific content of almost any image without changing the hidden activations. Further, we extend the standard cross-entropy loss to strengthen the model against such manipulations via an information-theoretic analysis, providing the first approach tailored explicitly to overcome invariance-based vulnerability. We conclude by empirically illustrating its ability to control undesirable class-specific invariance, showing promise to overcome one major cause for adversarial examples.

1 INTRODUCTION

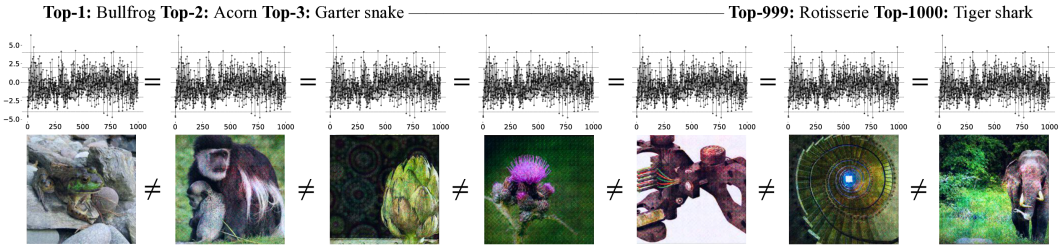

Figure 1: All images shown cause a competitive ImageNet-trained network to output the *exact same* probabilities over all 1000 classes (logits shown above each image). The leftmost image is from the ImageNet validation set; all other images are constructed such that they match the non-class related information of images taken from other classes (for details see section 2.2). The excessive invariance revealed by this set of adversarial examples demonstrates that the vector space of logits contains only a small fraction of the information that is perceptually relevant to humans for discrimination between the classes.

Adversarial vulnerability is one of the most iconic failure cases of modern machine learning models (Szegedy et al., 2013) and a prime example of their weakness in out-of-distribution generalization. It is particularly striking that under i.i.d. settings deep networks show superhuman performance on many tasks (LeCun et al., 2015), while tiny targeted shifts of the input distribution can cause them to make unintuitive mistakes.

The reasons for these failures and how they may be avoided or at least mitigated is a prime research area (Schmidt et al., 2018; Gilmer et al., 2018b; Bubeck et al., 2018).

So far, the study of adversarial examples has mostly been concerned with the setting of small perturbation, or ϵ-adversaries (Goodfellow et al., 2015; Madry et al., 2017; Raghunathan et al., 2018). Perturbation-based adversarial examples are appealing because they allow to quantitatively measure notions of adversarial robustness (Brendel et al., 2018). However, recent work argued that the perturbation-based approach is unrealistically restrictive and called for the need of generalizing the concept of adversarial examples to the unrestricted case, including any input crafted to be misinterpreted by the learned model (Song et al., 2018; Brown et al., 2018). Yet, settings beyond ϵ-robustness are hard to formalize (Gilmer et al., 2018a).

We argue here for an alternative, complementary viewpoint on the problem of adversarial examples. Instead of focusing on transformations erroneously crossing the decision-boundary of classifiers, we focus on excessive invariance as a major cause for adversarial vulnerability. To this end, we introduce the concept of invariance-based adversarial examples and show that class-specific content of almost any input can be changed arbitrarily without changing activations of the network, as illustrated in figure 1 for ImageNet. This viewpoint opens up new directions to analyze and control crucial aspects underlying vulnerability to unrestricted adversarial examples.
One hypothesis is that adversarial vulnerability is a consequence of narrow learning, yielding classifiers that rely only on a few highly predictive features in their decisions. It has indeed been shown that deep networks strongly rely on spectral statistical regularities (Jo & Bengio, 2017), or stationary statistics (Gatys et al., 2017) to make their decisions, rather than more abstract features like shape and appearance. We hypothesize that one reason for this can be understood from an information-theoretic viewpoint of cross-entropy, which maximizes a bound on the mutual information between labels and representation, giving no incentive to explain all class-dependent aspects of the input. This may be desirable in some cases, but to achieve truly general understanding of a scene or an object, machine learning models have to learn to successfully separate essence from nuisance and subsequently generalize even under shifted distributions.
Our contributions:

- We identify excessive invariance underlying striking failures in deep networks and formalize the connection to adversarial vulnerability.
- We propose an invertible network architecture that gives explicit access to the decision space, while preserving all information about the input. This enables us to apply almost any class-specific manipulation to images while leaving all dimensions of the representation seen by the final classifier invariant.
- From an information-theoretic viewpoint, we show that cross-entropy does not necessarily encourage explanation of all task-dependent variability. Leveraging bijective networks, we propose an extension that provably reduces invariance to semantically meaningful transformations and numerically demonstrate its promise to solve the problem.

2 Two Complementary Approaches to Adversarial Examples

In this section, we define pre-images and establish a link to adversarial examples.

Definition 1 (Pre-images / Invariance)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}^C$ be a neural network, $F = f_L \circ \cdots \circ f_1$ and layers f_i. Further, let $D : \mathbb{R}^d \to \{1, \dots, C\}$ be a classifier with $D = \arg\max_{k=1,\dots,C} softmax(F(x))_k$. Then, for input $x \in \mathbb{R}^d$, we define the following pre-images*

(i) i-th Layer pre-image: $\{x^* \in \mathbb{R}^d \mid F(x^*)_i = F(x)_i\}$

(ii) Logit pre-image: $\{x^* \in \mathbb{R}^d \mid F(x^*) = F(x)\}$

(iii) Argmax pre-image: $\{x^* \in \mathbb{R}^d \mid D(x^*) = D(x)\}$,

where (i) $\subset$ (ii) $\subset$ (iii) by the compositional nature of D.
*Moreover, the (sub-)network is **invariant** to perturbations Δx which satisfy $x^* = x + \Delta x$.*

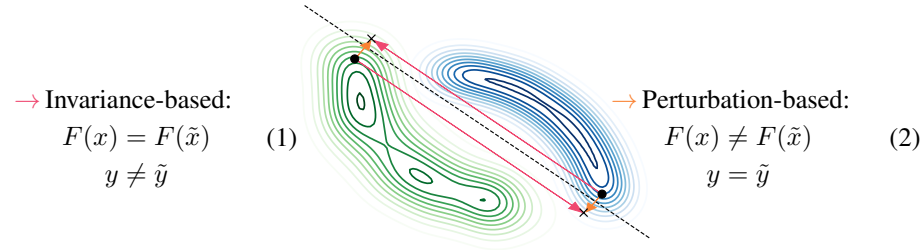

Figure 2: Connection between perturbation-based (short orange arrow) and invariance-based adversarial examples (long pink arrow). Class distributions are shown in green and blue; dashed line is the decision-boundary of a classifier. All adversarial examples can be reached either by crossing the decision-boundary of the classifier via ϵ-perturbations, or by moving within the pre-image of the classifier to mis-classified regions. The two viewpoints are *complementary* to one another and highlight that adversarial vulnerability is not only caused by excessive *sensitivity* to semantically meaningless perturbations, but also by excessive *insensitivity* to semantically meaningful transformations.

Non-trivial pre-images (pre-images containing more elements than input x) after the i-th layer occur if the chain $f_i \circ \cdots \circ f_1$ is not injective, for instance due to subsampling or non-injective activation functions like ReLU (Behrmann et al., 2018). This accumulated invariance can become problematic if not controlled properly, as we will show in the following.

We define perturbation-based adversarial examples by introducing the notion of an oracle (e.g., a human decision-maker or the unknown input-output function considered in learning theory):

Definition 2 (Perturbation-based Adversarial Examples)**.** *A **Perturbation-based adversarial example** $x^* \in \mathbb{R}^d$ of $x \in \mathbb{R}^d$ fulfills:*

*(i) Perturbation of decision: $D(x^*) \neq o(x^*)$, where $D : \mathbb{R}^d \to \{1, \dots, C\}$ is the classifier and $o : \mathbb{R}^d \to \{1, \dots, C\}$ is the **oracle**.*

(ii) Created by adversary: $x^* \in \mathbb{R}^d$ is created by an algorithm $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ with $x \mapsto x^*$.

*Further, **ϵ-bounded** adversarial ex. x^* of x fulfill $\|x - x^*\| < \epsilon$, $\|\cdot\|$ a norm on $\mathbb{R}^d$ and $\epsilon > 0$.*

Usually, such examples are constructed as ϵ-bounded adversarial examples (Goodfellow et al., 2015). However, as our goal is to characterize general invariances of the network, we do not restrict ourselves to bounded perturbations.

Definition 3 (Invariance-based Adversarial Examples)**.** *Let G denote the i-th layer, logits or the classifier (Definition 1) and let $x^* \neq x$ be in the G pre-image of x and and o an oracle (Definition 2). Then, an **invariance-based adversarial example** fulfills $o(x) \neq o(x^*)$, while $G(x) = G(x^*)$ (and hence $D(x) = D(x^*)$).*

Intuitively, adversarial perturbations cause the output of the classifier to change while the oracle would still consider the new input x^* as being from the original class. Hence in the context of ϵ-bounded perturbations, the classifier is *too sensitive to task-irrelevant changes*. On the other hand, movements in the pre-image leave the classifier invariant. If those movements induce a change in class as judged by the oracle, we call these invariance-based adversarial examples. In this case, however, the classifier is *too insensitive to task-relevant changes*. In conclusion, these two modes are complementary to each other, whereas both constitute failure modes of the learned classifier.
When not restricting to ϵ-perturbations, perturbation-based and invariance-based adversarial examples yield the same input x^* via

$$x^* = x_1 + \Delta x_1, \quad D(x^*) \neq D(x_1), \quad o(x^*) = o(x_1) \tag{3}$$
$$x^* = x_2 + \Delta x_2, \quad D(x^*) = D(x_2), \quad o(x^*) \neq o(x_2), \tag{4}$$

with different reference points x_1 and x_2, see Figure 2. Hence, the **key difference is the change of reference**, which allows us to approach these failure modes from different directions.

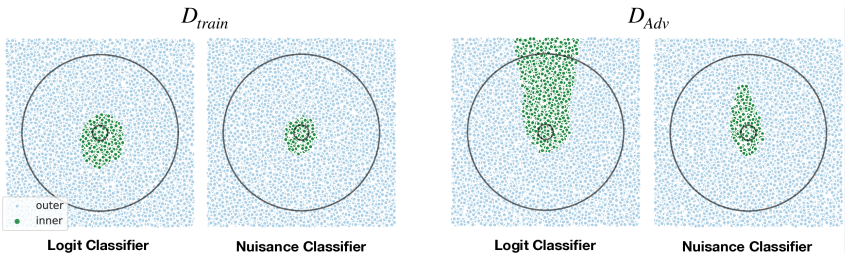

Figure 3: Left: Decision-boundaries in 2D subspace spanned by two random data points x_1, x_2. Right: Decision-boundaries in 2D subspace spanned by a random x and logit metamer x_{met}. For both subspaces we show the decision boundaries of classifier D trained on the logits z_s, and one trained on the nuisance variables z_n. In the adversarial distribution-shift case, it is possible to continuously move any point from the inner sphere to a large set of points on the outer sphere without changing the output of classifier D. However, this is not possible for the nuisance classifier.

2.1 Semantic and Nuisance Variation

In this section, we connect the notion of invariance to nuisance and semantic variations, see also (Achille & Soatto, 2018):

Definition 4 (Semantic/ Nuisance perturbation of an input)**.** *Let o be an oracle (Definition 2) and $x \in \mathbb{R}^d$. Then, a perturbation Δx of an input $x \in \mathbb{R}^d$ is called **semantic**, if $o(x) \neq o(x + \Delta x)$ and **nuisance** if $o(x) = o(x + \Delta x)$*

For example, such a nuisance perturbation could be a translation or occlusion in image classification.

Example 5 (Semantic and nuisance on Adversarial Spheres (Gilmer et al., 2018b))**.** *Consider classifying inputs x from two classes given by radii R_1 or R_2. Further, let (r, ϕ) denote the spherical coordinates of x. Then, any perturbation Δx, $x^* = x + \Delta x$ with $r^* \neq r$ is semantic. On the other hand, if $r^* = r$ the perturbation is a nuisance with respect to the task of discriminating two spheres.*

In this example, the max-margin classifier $D(x) = sign\left(\|x\| - \frac{R_1+R_2}{2}\right)$ is invariant to any nuisance perturbation, while being only sensitive to semantic perturbations. In summary, the transform to spherical coordinates allows to linearize semantic and nuisance perturbations. Using this notion, invariance-based adversarial examples can be attributed to perturbations of $x^* = x + \Delta x$ with following two properties

1. Perturbed sample x^* stays in the pre-image $\{x^* \in \mathbb{R}^d \mid D(x^*) = D(x)\}$ of the classifier
2. Perturbation Δx is semantic, as $o(x) \neq o(x + \Delta x)$.

Thus, the failure of the classifier D can be thought of a mis-alignment between its invariance (expressed through the pre-image) and the semantics of the data and task (expressed by the oracle).

Example 6 (Mis-aligned classifier on Adversarial Spheres)**.** *Consider the classifier*

$$D(x) = sign\big(\|x_{1,\ldots,d-1}\| - \frac{R_1 + R_2}{2}\big), \tag{5}$$

which computes the norm of x from its first $d-1$ cartesian-coordinates. Then, D is invariant to a semantic perturbation with $\Delta r = R_2 - R_1$ if only changes in the last coordinate x_d are made.

We empirically evaluate the classifier in equation 5 on the spheres problem (10M/2M samples setting (Gilmer et al., 2018b)) and validate that it can reach perfect classification accuracy. However, by construction, perturbing the invariant dimension $x_d^* = x_d + \Delta x_d$ allows us to move all samples from the inner sphere to the outer sphere. Thus, the accuracy of the classifier drops to chance level when evaluating its performance under such a distributional shift.
To conclude, this underlines how classifiers with optimal performance on finite samples can exhibit non-intuitive failure modes due to excessive invariance with respect to semantic variations.

2.2 Using Bijective Networks to Analyze Excessive Invariance

As invariance-based adversarial examples manifest themselves in changes which do not affect the network F, we need a generic approach that gives us access to the discarded nuisance variability.

While feature nuisances are intractable to access for general architectures (see comment after Definition 1), invertible classifiers only remove nuisance variability in their final projection (Jacobsen et al., 2018). For $C < d$, we denote the classifier as $D : \mathbb{R}^d \to \{1, ..., C\}$.
Our contributions in this section are: 1) we introduce an invertible architecture with a simplified readout structure, allowing to exactly visualize manipulations in the hidden-space, 2) we propose a sampling scheme based on this architecture allowing to analyze its decision-making.

We achieve this by removing the final linear mapping from bijective representation onto class-probes in iRevNet-classifiers and call these networks fiRevNets for fully-invertible. Further, we apply *actnorm* (Kingma & Dhariwal, 2018) and affine block structure (Dinh et al., 2017) to increase their flexibility, see Appendix B. The fiRevNet classifier can be written as $D_\theta = \arg\max_{k=1,...,C} softmax(F_\theta(x))_k$, where F_θ represents the bijective network. We denote $z = F_\theta(x)$, $z_s = z_{1,...,C}$ as the logits (semantic variables) and $z_n = z_{C+1,...,d}$ as the nuisance variables (z_n is not seen by classifier).

As we have relatively large feature maps in the final layer on ImageNet, we apply a DCT on them and read out the logits from the low-pass components, effectively giving us an invertible global average pooling and making our network more similar to ResNets in design (see appendix B for details). The resulting invertible network allows to qualitatively and quantitatively investigate the task-specific content in nuisance and logit variables. Despite this restriction, we achieve 0 % error on the spheres set, an error rate of 0.37 % on MNIST and performance on par with several solid baselines on ImageNet, see Table 1.
To analyze the trained models, we can sample elements from the logit pre-image by computing $x_{met} = F^{-1}(z_s, \tilde{z}_n)$. We choose a heuristic based on random combinations of nuisance and logit features from different images, which we term **metameric sampling**. These samples would be from the true data distribution if the joint would equal the factorial: $P(z_s, z_n) = P(z_s)P(z_n)$. Experimentally we find that logit metamers are indeed revealing adversarial subspaces on the spheres dataset and are visually close to natural images on ImageNet.

% Error	fiRevNet48	VGG19	ResNet18	ResNet50	iRevNet300
ILSVRC-2012 Val Top1	29.50	28.70	30.43	24.70	26.70
ILSVRC-2012 Val Top5	11.30	9.90	10.80	7.89	

Table 1: The table shows error rates on the ILSVRC-2012 Imagenet validation set of our proposed fully invertible fiRevNet48 compared to a VGG (Simonyan & Zisserman, 2014) and two ResNet (He et al., 2016) variants, as well as an iRevNet300 (Jacobsen et al., 2018) with a non-invertible final projection onto the logits. The results show that our proposed fiRevNet performs roughly on par with the other architectures.

First, we analyze the model trained on the synthetic spheres dataset and choose $d = 100$ and $R_1 = 1$, $R_2 = 10$. We use the same approach as Gilmer et al. (2018b) and visualize two cases: 1) the decision-boundary on a 2D plane spanned by two randomly chosen data points, 2) the decision-boundary spanned by x_{met} and reference point x. We visualize these decision-boundaries for the original classifier D and for a posthoc trained classifier on z_n (nuisance classifier), see Figure 3. Most notably, the visualized failure is not due to a lack of data seen during training but rather due to excessive invariance of D, as the nuisance classifier does not make the same errors.

After validating the potential to uncover adversarial subspaces, we apply metameric sampling to fiRevNets trained on MNIST and Imagenet, see Figure 4. The result is striking, as the nuisance variables z_n are dominating the visual appearance of the logit metamers, making it possible to attach any semantic content to any logit activation pattern. Note that the entire 1000-dimensional feature vector containing probabilities over all ImageNet classses remains unchanged by any of the transformations we apply. From a standpoint of sampling from a data distribution, metameric sampling induces a distributional shift which renders the features z_s (used by classifier D) visually uninformative about the classification task.

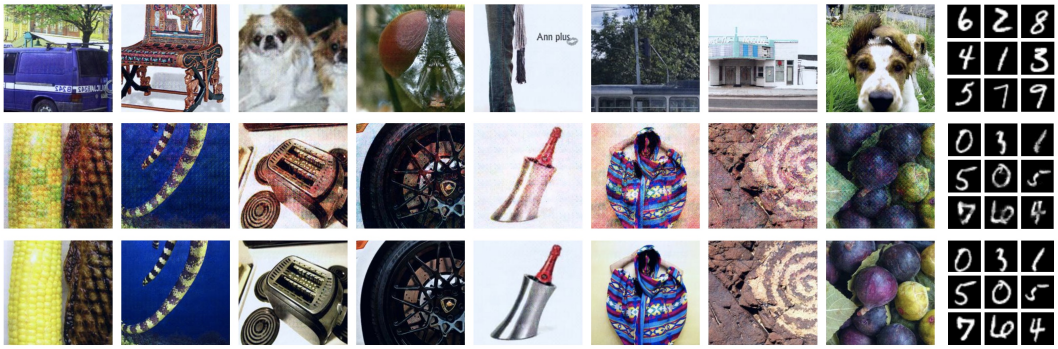

Figure 4: Each column shows three images belonging together. Top row are source images from which we sample the logit activations, middle row are logit metamers and bottom row images from which we sample the nuisance variabilities. It is possible to change the image content completely without changing the 10- and 1000-dimensional logit vectors respectively. This highlights the striking failure of the classifiers to capture all task-dependent variability and makes it straightforward to design invariance-based adversarial examples for arbitrary images.

3 Overcoming Insufficiency of Crossentropy-based Information-maximization

After revealing how the nuisance classifier used task-relevant information not captured by the logit classifier D (evident by its good performance in the adversarial subspace of D in Figure 3), we leverage the simple readout-structure of our invertible network and turn this observation into a principled objective using information theory: Let $(x, y) \sim \mathcal{D}$ with labels $y \in \{0, 1\}^C$. Now, our goal can be stated as maximizing the mutual information ((Cover & Thomas, 2006)) between semantic features z_s extracted by network F_θ and labels y, denoted by $I(y; z_s)$.

As the previously discussed failures required to modify input data from distribution $\mathcal{D}$, we introduce the concept of an **adversarial distribution shift $\boldsymbol{\mathcal{D}_{Adv} \neq \mathcal{D}}$** to account for these modifications. This distribution shift could be due to a re-weighting of probability mass while keeping the support; e.g., in Example 6 the samples still lie on the spheres but the unused coordinates of the classifier were used to decide its norm. Furthermore, shifts in the support of $\mathcal{D}$ could be possible, like changes in the noise level or other input statistics. Our only assumption for $\mathcal{D}_{Adv}$ is $I_{\mathcal{D}_{Adv}}(z_n; y) \leq I_{\mathcal{D}}(z_n; y)$. Intuitively, the nuisance variables of our network do not become more informative about y. Thus, the distribution shift may reduce the predictiveness of features encoded in z_s, but does not introduce or increase the predictive value of variations captured in z_n.
As bijective networks F_θ capture all variations by design which translates to information preservation $I(y; x) = I(y; F_\theta(x))$, see (Kraskov et al., 2004). Consider the reformulation

$$I(y; x) = I(y; F_\theta(x)) = I(y; z_s, z_n) = I(y; z_s) + I(y; z_n|z_s) = I(y; z_n) + I(y; z_s|z_n) \tag{6}$$

by the chain rule of mutual information (Cover & Thomas, 2006), where $I(y; z_n|z_s)$ denotes the conditional mutual information. Most strikingly, equation 6 offers three ways forward:

1. Direct increase of $I(y; z_s)$
2. Indirect increase of $I(y; z_s|z_n)$ via decreasing $I(y; z_n)$
3. Reducing dependence $I(z_s; z_n)$ to make maximizing $I(y; z_s)$ equivalent to minimizing $I(y; z_n)$, as $I(y; z_n|z_s) = I(y; z_s)$ under independence ($I(z_s; z_n) = 0$).

Usually in a classification task, only $I(y; z_s)$ is increased actively via training a classifier. While this approach is sufficient in most cases, expressed via high accuracies on training and test data, it may fail under $\mathcal{D}_{adv}$. However, by leveraging the bijection F_θ we can 1) minimize the unused information $I(y; z_n)$ using the intuition of a nuisance classifier, 2) control the dependence structure between z_s and z_n via maximum-likelihood under a factorial prior. Hence, we extend cross-entropy minimization by two terms that aim for independence.

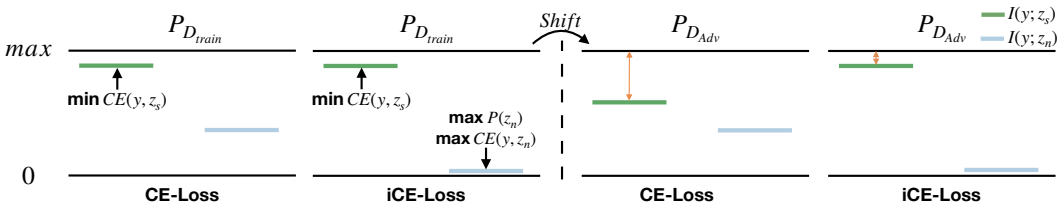

Figure 5: Left: Mutual information under distribution $\mathcal{D}_{train}$, Right: Effect of distributional shift to $\mathcal{D}_{Adv}$. Each case under training with cross-entropy (CE) and independence cross-entropy (iCE). Under distribution $\mathcal{D}$, the iCE-loss minimizes $I(y; z_n)$ (Lemma 10, Appendix A), but has no effect as the CE-loss already maximizes $I(y; z_s)$. However under the shift to $\mathcal{D}_{Adv}$, the information $I(y; z_s)$ decreases when training only under the CE-loss (orange arrow), while the iCE-loss induces $I(y; z_n) = 0$ and thus leaves $I(y; z_s)$ unchanged (Theorem 8).

Definition 7 (Independence cross-entropy loss)**.** *Let $F_\theta : \mathbb{R}^d \to \mathbb{R}^d$ a bijective network with parameters $\theta \in \mathbb{R}^{p_1}$ and $\tilde{F}_\theta(x) = softmax(F_\theta(x)_{1,\dots,C})$. Furthermore, let $D_{\theta_{nc}} : \mathbb{R}^{d-C} \to [0,1]^C$ be the nuisance classifier with $\theta_{nc} \in \mathbb{R}^{p_2}$. Then, the independence cross-entropy loss is defined as*

$$\min_{\theta} \max_{\theta_{nc}} \mathcal{L}_{iCE}(\theta, \theta_{nc}) = \underbrace{-\sum_{i=1}^{C} y_i \log \tilde{F}_\theta^{z_s}(x)_i}_{=:\mathcal{L}_{sCE}(\theta)} + \underbrace{\sum_{i=1}^{C} y_i \log D_{\theta_{nc}}(F_\theta^{z_n}(x))_i}_{=:\mathcal{L}_{nCE}(\theta,\theta_{nc})} - \underbrace{\prod_{k=1}^{d-C} p_k(F_\theta^{z_n}(x)_k)|det(J_\theta^x)|}_{=:\mathcal{L}_{MLE_n}(\theta)}$$

with $(x, y) \sim \mathcal{D}$. Furthermore, $det(J_\theta^x)$ denotes the determinant of the Jacobian of $F_\theta(x)$ and $p_k \sim \mathcal{N}(\beta_k, \gamma_k)$ with β_k, γ_k learned parameter.

The underlying principles of the nuisance classification loss $\mathcal{L}_{nCE}$ can be understood using a variational lower bound on mutual information from (Barber & Agakov, 2003). In summary, the minimization is with respect to a lower bound on $I_{\mathcal{D}}(y; z_n)$, while the maximization aims to tighten the bound. Furthermore, the term $\mathcal{L}_{MLE_n}$ on the nuisance variables together with $\mathcal{L}_{sCE}$ is maximum-likelihood under a factorial prior. See Lemma 10 and its proof in Appendix A. By using these results, we now state the main result under the assumed distribution shift and successful minimization (proof in Appendix A.1):

Theorem 8 (Information $\boldsymbol{I_{\mathcal{D}_{Adv}}(y; z_s)}$ maximal after distribution shift)**.** *Let $\mathcal{D}_{Adv}$ denote the adversarial distribution and $\mathcal{D}$ the training distribution. Assume $I_{\mathcal{D}}(y; z_n) = 0$ by minimizing $\mathcal{L}_{iCE}$ and the distribution shift satisfies $I_{\mathcal{D}_{Adv}}(z_n; y) \leq I_{\mathcal{D}}(z_n; y)$. Then,*

$$I_{\mathcal{D}_{Adv}}(y; z_s) = I_{\mathcal{D}}(y; x).$$

In summary, incorporating the nuisance classifier allows for the discussed indirect increase of $I_{\mathcal{D}_{Adv}}(y; z_s)$ under an adversarial distribution shift, visualized in Figure 5.

4 Applying Independence Cross-entropy

To test our derived loss, we apply it on the MNIST classification task as a test bed. While classifying MNIST digits is considered a toy example, it has been shown to be challenging from the perspective of ϵ-robustness (Schott et al., 2018). We make the same observation for invariance-based adversaries, as metameric sampling allows us to virtually attach any semantic image content to any logit configuration.

We compare our proposed independence cross-entropy training to vanilla cross-entropy training in three aspects: 1) error on train and test set, 2) effect under distribution shift: perturbing nuisances via metameric sampling and 3) accuracy of a classifier on the nuisance variables to quantify the class-specific information in them. For both experiments we use the exact same network architectures and settings, the only difference being that one network has an additional classifier on the nuisance variables providing us with a lower-bound on the mutual information $I(y, z_n)$ and the two additional weighted loss terms as explained in Definition 7. In terms of test error of the logit classifier, both losses perform approximately on par, whereas the gap between train and test error vanishes for our proposed loss function, indicating less overfitting, see Table 2.

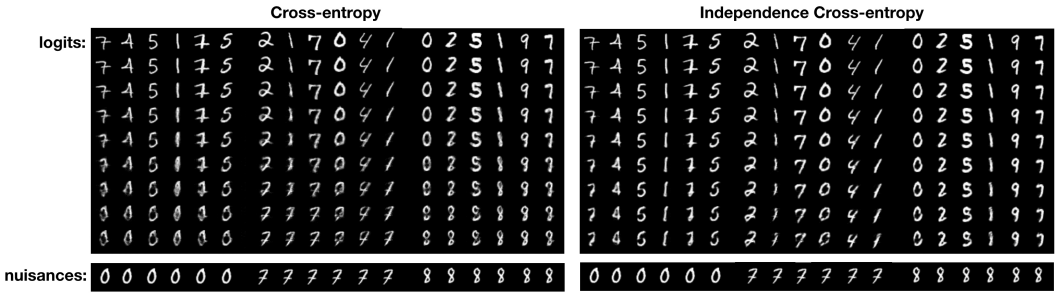

Figure 6: Interpolations between original image and metameric sample $x_{met} = F^{-1}(z_s, \tilde{z}_n)$. (one per column). Logit activations z_s are taken from the original image and kept constant, while nuisances z_n are interpolated between original image and target image (shown below each column). When training with cross-entropy, virtually any image can be turned into any class without changing the logits z_s, illustrating strong vulnerability to invariance-based adversaries. Yet, training with independence cross-entropy solves the problem and interpolations between different nuisances z_n leave the semantic content of the image invariant.

	SOTA	LeNet	CE	iCE (ours)	CE	iCE (ours)
MNIST	Logit	Logit	Logit	Logit	Nuisance	Nuisance
% Test Error	0.21	1.70	0.39	0.38	0.34	27.70
% Train Error			0.00	0.37	0.00	40.21

Table 2: Results comparing cross-entropy training (CE) with independence cross-entropy training (iCE) from Definition 7 and two architectures from the literature. The accuracy of the logit classifiers is on par for the CE and iCE networks, but the train error is higher for CE compared to test error, indicating less overfitting for iCE. Further, a classifier independently trained on the nuisance variables is able to reach even smaller error than on the logits for CE, but just 27.70% error for iCE, indicating that we have successfully removed most of the information of the label from the nuisance variables and fixed the problem of excessive invariance to semantically meaningful variability with no cost in test error.

In Figure 6 we show interpolations between images and logit metamers. In particular, we are holding the activations z_s constant, while interpolating nuisances z_n down the column, from the original to target values drawn from the image in the last row. This shows how excessive invariance underlies vulnerabilities towards visually small distributional shifts. However, when training with our proposed independence cross-entropy, the picture changes fundamentally and now the interpolation in the pre-image only changes the style of a digit, but not its semantic content. This is strong evidence for our loss having the ability to overcome the issue of excessive task-related invariance and encourages the model to explain and separate all task-related variability of the input from the nuisances of the task. Finally, a classifier trained on the nuisance variables of the cross-entropy trained model performs almost as well as the logit classifier. Yet, a classifier on the nuisances of the independence cross-entropy trained model is performing poorly (Table 2), indicating little class-specific information in the nuisances, as intended by our objective function.

5 Related Work

Adversarial examples often include ϵ-norm restrictions (Szegedy et al., 2013), while (Gilmer et al., 2018a) argue for a broader definition to fully capture the implications for security. Some works (Song et al., 2018; Fawzi et al., 2018) consider unrestricted adversarial examples, which are closely related to invariance-based adversarial vulnerability. The difference to human perception revealed by adversarial examples fundamentally questions which statistics the network use to base its decisions (Jo & Bengio, 2017; Tsipras et al., 2018).

The role of invariance has been studied extensively in the context of deep neural networks (Bruna & Mallat, 2013; Mallat, 2016; Achille & Soatto, 2018; Behrmann et al., 2018; Cadena et al., 2018). The invariance-based viewpoint also has a long history in neuroscience that started in the context of color vision. Metamerism refers to the cognitive equivalent to the preimage in neuroscience (Wandell, 1995; Freeman & Simoncelli, 2011; Wallis et al., 2016). Here we generate metameric samples in artificial neural networks to explore their invariance properties.

We also leverage recent advances in reversible (Gomez et al., 2017) and bijective networks (Jacobsen et al., 2018; Ardizzone et al., 2018; Kingma & Dhariwal, 2018) for our analysis.

Finally, the information-theoretic view has gained recent interest due to the information bottleneck (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017; Alemi et al., 2017) and usage in generative modelling (Chen et al., 2016; Hjelm et al., 2018). As a consequence, the estimation of mutual information (Barber & Agakov, 2003; Alemi et al., 2018; Achille & Soatto, 2018; Belghazi et al., 2018) has attracted growing attention. The concept of group-wise independence between latent variables goes back to classical independent subspace analysis (Hyvärinen & Hoyer, 2000) and received attention in learning unbiased representations, e.g. see the Fair Variational Autoencoder (Louizos et al., 2015). Furthermore, extended cross-entropy losses via entropy terms (Pereyra et al., 2017) or minimizing predictability of variables (Schmidhuber, 1991) has been introduced for other applications. Our proposed loss also shows similarity to the GAN loss (Goodfellow et al., 2014). However, in our case there is no notion of real or fake samples, but exploring similarities in the optimization are a promising avenue for future work.

6 Conclusion

The failures of deep networks under distribution shift and their difficulty in out-of-distribution generalization are prime examples of the limitations in current machine learning models. The field of adversarial example research aims to close this gap from a robustness point of view. While a lot of work has studied ϵ-adversarial examples, recent trends extend the efforts towards the unrestricted case. Adversarial examples with no restriction are hard to formalize beyond testing error. However, based on our reverse view we: 1) make the control of invariance tractable via fully-invertible networks; and 2) introduce a loss to explicitly combat excessive invariance, a major cause of unrestricted adversarial examples. Thus, we provide an approach tailored explicitly to overcome invariance-based adversarial vulnerability.

In summary, we demonstrated how a bijective network architecture enables us to identify large adversarial subspaces, due to excessive invariance on multiple datasets like the adversarial spheres, MNIST and ImageNet. We formalized the distribution shifts causing such undesirable behavior via information theory and find one of its major reasons in the insufficiency of the vanilla cross-entropy loss to learn semantic representations that capture all task-dependent variations in the input. We extend the loss function by components that explicitly encourage a split between semantically meaningful and nuisance features and show that this split can remove unwanted invariances.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 18, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *International Conference on Lerning Representations*, 2017.

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. An information-theoretic analysis of deep latent-variable models. *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W. Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Kthe. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:abs/1808.04730*, 2018.

David Barber and Felix Agakov. The im algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systemss*, 2003.

Jens Behrmann, Sören Dittmer, Pascal Fernsel, and Peter Maaß. Analysis of invariance and robustness via invertibility of relu-networks. 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P Mohanty, and Matthias Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018.

Tom B. Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples, 2018.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Santiago A. Cadena, Marissa A. Weis, Leon A. Gatys, Matthias Bethge, and Alexander S. Ecker. Diverse feature visualizations reveal invariances in early layers of deep neural networks. *arXiv preprint arXiv:abs/1807.10589*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2016.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *International Conference on Learning Representations*, 2017.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14 (9):1195, 2011.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture and art with deep neural networks. *Current opinion in neurobiology*, 46:178–186, 2017.

Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018a.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018b.

Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. *Advances in Neural Information Processing Systems*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. In *International Conference on Learning Representations*, 2018.

Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69, 2004.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:abs/1511.00830*, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2017.

Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065): 20150203, 2016.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. *International Conference on Lerning Representations (workshop)*, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *International Conference on Learning Representations*, 2018.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. Technical report, Dept. of Comp. Sci., University of Colorado at Boulder, 1991.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Lukas Schott, Jonas Rauber, Wieland Brendel, and Matthias Bethge. Robust perception through analysis by synthesis. *arXiv pre-print arXiv:1805.09190*, 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *arXiv preprint arXiv:1805.07894*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pp. 1–5. IEEE, 2015.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:abs/1711.11561*, 2018.

Thomas SA Wallis, Matthias Bethge, and Felix A Wichmann. Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of vision*, 16(2):4–4, 2016.

Brian A Wandell. *Foundations of vision*, volume 8. sinauer Associates Sunderland, MA, 1995.

A INFORMATION THEORY

Computing mutual information is often intractable as it requires the joint probability $p(x, y)$, see (Cover & Thomas, 2006) for an extensive treatment of information theory. However, following variational lower bound can be used for approximation, see (Barber & Agakov, 2003).

Lemma 9 (Variational lower bound on mutual information)**.** *Let X, Y be random variables with conditional density $p(y|x)$. Further, let $q_\theta(y|x)$ be a variational density depending on parameter θ. Then, the lower bound*

$$\begin{aligned} I(Y;X) = h(Y) - h(Y|X) &= h(Y) + \mathbb{E}_X \mathbb{E}_{Y|X} \log q_\theta(y|x) + \mathbb{E}_X (p(y|x) \parallel q_\theta(y|x)) \\ &\geq h(Y) + \mathbb{E}_X \mathbb{E}_{Y|X} \log q_\theta(y|x) \end{aligned}$$

holds with equality if $p(y|x) = q_\theta(y|x)$.

While above lower bound removes the need for the computation of $p(y|x)$, estimating the expectation $\mathbb{E}_{Y|X}$ still requires sampling from it. Using this bound, we can now state the effect of the nuisance classifiation loss.

Lemma 10 (Effect loss terms)**.** *Define semantics as $z_s = F_\theta(x)_{1,\dots,C}$ and nuisances as $z_n = F_\theta(x)_{C+1,\dots,d}$, where $(x, y) \sim \mathcal{D}$. Then, the nuisance classification loss yields*

(i) ***Minimization of $I(z_s; z_n)$:*** *$\theta^* = \arg\min_\theta \mathcal{L}_{sCE}(\theta) + \mathcal{L}_{MLE_n}(\theta)$ minimizes $I(z_s; z_n)$*

(ii) ***Minimization of lower bound on $I_\mathcal{D}(y; z_n)$:*** *$\theta^* = \arg\min_\theta \mathcal{L}_{nCE}(\theta, \theta^*_{nc})$ minimizes $I_{\theta^*_{nc}}(y; z_n)$, where $I_{\theta^*_{nc}}(y; z_n) \leq I_\mathcal{D}(y; z_n)$ and $\theta^*_{nc} = \arg\max_{\theta_2} \mathcal{L}_{nCE}(\theta, \theta_{nc})$.*

(iii) ***Maximization to tighten bound on $I_\mathcal{D}(y; z_n)$:*** *Under a perfect model of the conditional density, $D_{\theta^*_{nc}}(z_n) = p(y|z_n)$, it holds $I_{\theta^*_{nc}}(y; z_n) = I_\mathcal{D}(y; z_n)$.*

Proof. We begin by proving claim (i), which handles the dependence $I(z_s, z_n)$. Let $\tilde{z}_s = softmax(z_s)$. Then minimizing the loss terms $\mathcal{L}_{sCE}$ and $\mathcal{L}_{MLE_n}$ is a maximum likelihood estimation under the factorial prior

$$p(\tilde{z}_s, z_n = p(\tilde{z}_s)p(z_n) \tag{7}$$

$$= Cat((\tilde{z}_s)_1, \dots, (\tilde{z}_s)_C) \prod_{k=1}^{d-C} p_k(z_n)_k, \tag{8}$$

where Cat is a categorical distribution. As $softmax$ is shift-invariant, $softmax(x + c) = softmax(x)$, above factorial prior for $\tilde{z}_s$ and z_n yields independence between logits z_s and z_n

up to a constant c. Finally note, the log term and summation in $\mathcal{L}_{MLE_n}$ and $\mathcal{L}_{CE}$ is re-formulation for computational ease but does not change its minimizer as the logarithm is monotone.
For claim (ii) and (iii), we need to draw the connection to the variational lower bound on mutual information from Lemma 9. Let the nuisance classifier $D_{\theta_{nc}}(z_n)$ model the variational posterior $q_{\theta_{nc}}(y|z_n)$. Then we have the lower bound

$$I(y; z_n) \geq h(y) + \mathbb{E}_{z_n}\mathbb{E}_{y|z_n} \log D_{\theta_{nc}}(z_n) =: I_{\theta_{nc}}(y; z_n). \tag{9}$$

From Lemma 9 follows, that if $D_{\theta_{nc}}(z_n) = p(y|z_n)$, it holds $I(y; z_n) = I_{\theta_{nc}}(y; z_n)$. Hence, the nuisance classifier needs to model the conditional density perfectly.
Estimating this bound via Monte Carlo simulation requires sampling from the conditional density $p(y|z_n)$. Following (Alemi et al., 2017), we have the Markov property $y \leftrightarrow x \leftrightarrow z_n$ as labels y interact with inputs x and representation z_n interacts with inputs x. Hence,

$$\begin{aligned} p(y|z_n)p(z_n) &= p(y, z_n) \\ &= \int_{\mathcal{X}} p(x, y, z_n)dx \\ &= \int_{\mathcal{X}} p(z_n|x, y)p(y|x)p(x)dx \\ &= \int_{\mathcal{X}} p(z_n|x)p(y|x)p(x)dx \\ &= \mathbb{E}_x[p(z_n|x)p(y|x)]. \end{aligned}$$

Including above and assuming $F_\theta(x) = z_n$ to be a deterministic function, we have

$$\mathbb{E}_{z_n}\mathbb{E}_{y|z_n} \log D_{\theta_{nc}}(z_n) = \mathbb{E}_x\mathbb{E}_{y|x}\mathbb{E}_{z_n|x} \log D_{\theta_{nc}}(z_n) = \mathbb{E}_x\mathbb{E}_{y|x} \log D_{\theta_{nc}}(z_n).$$

□

A.1 Proof of Theorem 8

From the assumptions follows $I_{\mathcal{D}_{Adv}}(y; z_n) = 0$. Furthermore

$$I_{\mathcal{D}_{Adv}}(y; z_s|z_n) = I_{\mathcal{D}_{Adv}}(z_s; y|z_n) = I_{\mathcal{D}_{Adv}}(z_s; y),$$

where the first equality follows from the symmetry of mutual information and the last from the properties of conditional mutual information under independence $I_{\mathcal{D}_{Adv}}(y; z_n) = 0$. By the chain rule of mutual information (Cover & Thomas, 2006) and information preservation under homeomorphisms (Kraskov et al., 2004), we have

$$\begin{aligned} I_{\mathcal{D}_{Adv}}(x; y) &= I_{\mathcal{D}_{Adv}}(y; z_n) + I_{\mathcal{D}_{Adv}}(y; z_s|z_n) \\ &= I_{\mathcal{D}_{Adv}}(z_s; y). \end{aligned}$$

A.2 Mutual information bounded

Remark 11. *Since our goal is to maximize the mutual information $I(y; z_s)$ while minimizing $I(y; z_n)$, we need to ensure that this objective is well defined as mutual information can be unbounded from above for continuous random variables. However, due to the data processing inequality (Cover & Thomas, 2006) we have $I(y; z_n) = I(y; F_\theta(x) \leq I(y; x)$. Hence, we have a fixed upper bound given by our data (x, y). Compared to (Belghazi et al., 2018) there is thus no need for gradient clipping or a switch to the bounded Jensen-Shannon divergence as in (Hjelm et al., 2018) is not necessary.*

B Training and Architectural Details

All experiments were based on a fiRevNet model with different hyperparameters for each dataset. For the spheres experiment we used Pytorch (Paszke et al., 2017) and for MNIST, as well as Imagenet Tensorflow (Abadi et al., 2016).

B.1 SPHERES EXPERIMENTS

The network is a fully connected fiRevNet. It has 4 RevNet-type ReLU bottleneck blocks with additive couplings and uses no batchnorm. We train it via cross-entropy and use the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001 and otherwise default Pytorch settings. The nuisance classifier is a 3 layer ReLU network with 1000 hidden units per layer.

We choose the spheres to be 100-dimensional, with $R_1 = 1$ and $R_2 = 10$, train on 500k samples for 10 epochs and then validate on another 100k holdout set. We achieve 100% train and validation accuracy for logit and nuisance classifier.

B.2 MNIST EXPERIMENTS

We use a convolutional fiRevNet with additional actnorm and invertible 1x1 convolutions between each layer as introduced in Kingma & Dhariwal (2018). The network has 3 stages, after which half of the variables are factored out and an invertible downsampling, or squeezing (Dinh et al., 2017; Jacobsen et al., 2018) is applied. The network has 16 RevNet blocks with batch norm per stage and 128 filters per layer. We also dequantize the inputs as is typically done in flow-based generative models.

The network is trained via Adamax (Kingma & Ba, 2014) with a base learning rate of 0.001 for 100 epochs and we multiply the it with a factor of 0.2 every 30 epochs and use a batch size of 64 and l2 weight decay of 1e-4. For training we compare vanilla cross-entropy training with our proposed independence cross-entropy loss. To have a more balanced loss signal, we normalize $\mathcal{L}_{nCE}$ by the number of input dimensions it receives for the maximization step. The nuisance classifier is a fully connected 3 layer ReLU network with 512 units. As data-augmentation we use random shifts of 3 pixels.

B.3 IMAGENET EXPERIMENTS

We use a convolutional fiRevNet with 4 stages, 4 RevNet blocks per stage and invertible downsampling after each stage, as well as two invertible downsamplings on the input of the network. The first three stages consist of additive and the last of affine coupling layers. After the final layer we apply an orthogonal 2D DCT type-II to all feature maps and read out the classes in the low-pass components of the transformation. This effectively gives us an invertible global average pooling and makes our network even more similar to ResNets, that always apply global average pooling on their final feature maps. We train the network with momentum SGD for 128 epochs, a batch size of 480 (distributed to 6 GPUs), a base learning rate of 0.1, which is reduced by a factor of 0.1 every 32 epochs. We apply momentum of 0.9 and l2 weight decay of 1e-4.