

BAYES-TREx: Model Transparency by Example

Serena Booth*, Yilun Zhou*, Ankit Shah, and Julie Shah

MIT CSAIL

*equal contribution

{serenabooth, yilun, ajshah, julie_a_shah}@csail.mit.edu

Abstract

Post-hoc explanation methods are gaining popularity as tools for interpreting, understanding, and debugging neural networks. Most post-hoc methods explain decisions in response to individual inputs. These individual inputs are typically drawn from the test set; however, the test set may be biased or may only sparsely invoke some model behaviours. To address these challenges, we introduce BAYES-TREx, a model-agnostic method for generating distribution-conforming examples of known prediction confidence. Using a classifier prediction and a data generator, BAYES-TREx can be used to visualize class boundaries; to find *in-distribution* adversarial examples; to understand novel-class extrapolation; and to expose neural network overconfidence. We demonstrate BAYES-TREx with rendered data (CLEVR) and organic data (MNIST, Fashion-MNIST). Code: github.com/serenabooth/Bayes-TrEx.

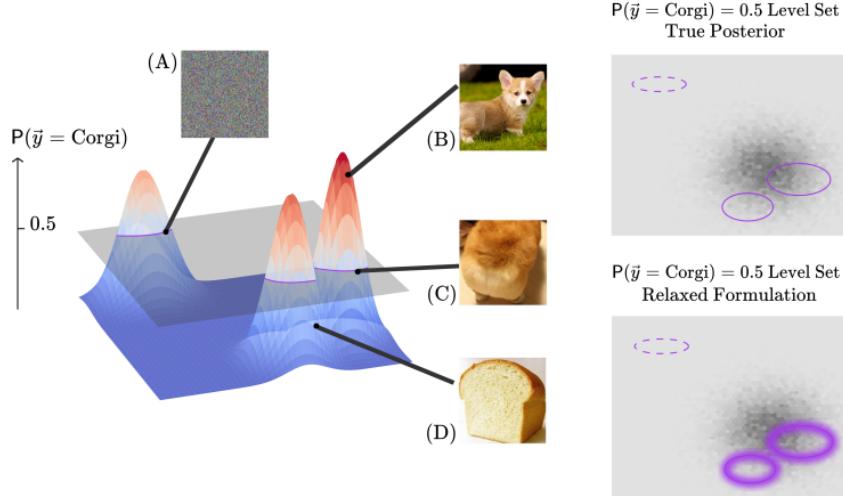


Figure 1: Given a Corgi/Bread classifier and a generative model for the data, we generate *prediction level sets*, or sets of examples of a target prediction confidence. (A) Existing adversarial approaches can perturb a given image to the target prediction confidence (e.g. $P_{\text{Corgi}} = P_{\text{Bread}} = 0.5$); however, such examples are “out-of-distribution.” (B, C, D) BAYES-TREx, on the other hand, takes a specified data distribution to sample “in-distribution” examples in the target prediction level set (e.g., likely Corgi, likely Bread, or ambivalent between Corgi and Bread). Top right: the classifier level set of ($P_{\text{Corgi}} = P_{\text{Bread}} = 0.5$) overlaid on the data distribution heatmap. Example (A) would not be sampled because its likelihood is low under the generative model, while example (C) would be sampled. Bottom right: As sampling directly from the true posterior is infeasible, we relax the formulation by widening the level set. By specifying different data distributions (Figure 3), we can uncover inputs that invoke various model behaviors to improve model transparency by example (Section 4.3 - 4.7).

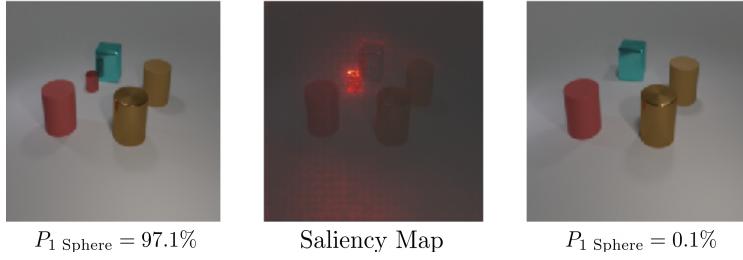


Figure 2: A preprocessed CLEVR example and its SmoothGrad saliency map [55]. This example is generated, *not* drawn from the test set. BAYES-TREX is tasked with finding a scene which contains no spheres but is misclassified as containing a sphere. The inferred example is composed of only cylinders and cubes, but the classifier is 97.1% confident this scene contains 1 sphere. The saliency map highlights the small red cylinder, indicating this object was confused for a sphere. Indeed, without the small red cylinder, the classifier confidence of containing 1 sphere drops to 0.1%.

1 Introduction

Debugging, interpreting, and understanding neural networks can be challenging [48, 39, 12]. Existing interpretability methods include visualizing filters [64, 49], saliency maps [64, 53], input perturbations [51, 41], prototype anchoring [27, 36, 9], tracing with influence functions [31], and concept quantification [17, 28]. While some post-hoc interpretability methods analyze intermediary network components such as convolutional layers [5, 49], most methods provide explanations which justify decisions for specific input examples. Finding and analyzing myriad inputs which invoke the gamut of model behaviours facilitates *transparency by example*. However, the inputs used by local explanation methods are typically drawn from the test set, which may be biased or may only sparsely invoke some model behaviours. This can make it challenging to extract meaningful insights with only test inputs.

Expanding the test set without new labels is impossible, but generative models can represent the underlying data distributions. We propose a tool to help provide transparency by example that varies the underlying input distributions via generative models. BAYES-TREX samples inputs at a known prediction confidence: for example, $P_{\text{Corgi}} = 0.7$, $P_{\text{Bread}} = 0.3$ for a Corgi/Bread classifier (Fig. 1). We call the set of all input examples that meet this prediction confidence the *p-level set*. Given a specified or learned data distribution and a classifier, BAYES-TREX generates level-set examples from the posterior of a hierarchical Bayesian model by applying Markov Chain Monte-Carlo (MCMC) inference methods. This applies across generative models: we show its use with VAEs and GANs (MNIST and Fashion-MNIST [63]), and with a scene-graph-rendered dataset (CLEVR [26]).

BAYES-TREX can aid transparency by example in different contexts. By specifying an ambiguous prediction target, BAYES-TREX can generate examples to visualize class boundaries. By restricting the generator to not produce certain classes, BAYES-TREX can generate “in-distribution” adversarial examples—these cause model failures, but are better suited for human assessment than typical adversarial examples. By introducing novel classes for generation, BAYES-TREX can help model designers anticipate how their models will behave. Further, BAYES-TREX examples can be used as input to existing explanation techniques (Fig. 2). Lastly, we show how BAYES-TREX can be used to expose network overconfidence, as we demonstrate on the domain-adapted model ADDA [60].

2 Related Work

2.1 Interpreting Neural Networks

A typical approach for interpreting a neural network is to view the model’s activation-maximizing ‘filters’ [49, 13]. Without regularization, these filters are unnatural images; to address this, Nguyen et al. used a generative adversarial network as a prior to ensure the generated filters appear realistic, balancing activation maximization with realism [45]. While these activation-maximizing filters can help diagnose individual units in a neural network, understanding pattern-matching between inputs and filters remains challenging [49, 47]. Instead of finding optimized images which maximize neuron activations, BAYES-TREX finds in-distribution inputs which invoke a specified model response—including arbitrary prediction confidence targets such as ambivalence between classes.

Saliency maps are the local counterpart to filters; saliency maps project feature activations to input pixel space to highlight regions of interest in inputs [64, 53]. While saliency maps have been used successfully for model diagnosis [64] and are now ubiquitous tools for interpreting neural networks, these visualizations can appear reasonable even when output from an untrained network [30, 1]. Other interpretability tools such as LIME [51] and SHAP [41] perturb a given input and score how these perturbations affect classification. Another approach, TCAV, scores the relative importance of concepts to a neural network’s classification for a given example [17, 28]. Notably, all these methods work on a *per-example* basis and assume that the example is given. Such examples are typically drawn from a test set or in-use failures; however, the test set may be biased or may only sparsely invoke some model responses, and failures may not be observed by model designers. Model explanations require a two-stage pipeline. First, we need a method for mining additional interesting examples, leveraging both test and train data distributions. We can then invoke local interpretation methods to explain these examples, as shown in Fig. 2 and Appendix K. To our knowledge, BAYES-TREX is the first work dedicated to mining interesting inputs for evaluation with local explanation methods.

2.2 Adversarial Examples for Neural Networks

Neural networks are known to be susceptible to adversarial examples, or perturbed inputs designed to cause network failures. Network predictions are sensitive to even small perturbations [56, 35, 4, 8, 46, 19]. A common approach to generating adversarial examples starts with an initial input—either random noise or a test image—and uses a gradient-based method to iteratively optimize a distance function, modifying the input to cause neural network failure. The examples generated through this process are unlikely given the training data and are therefore out-of-distribution [37]. Consequently, many defenses against adversarial attacks rely on examples being out-of-distribution [23, 14]. In contrast, BAYES-TREX exposes “in-distribution” adversarial examples [18] which cause failures but have high prior probability of being sampled from the data distribution. As traditional adversarial examples are either composed of random noise or imperceptible perturbations, analyzing these examples does not help humans understand neural network behaviours. The in-distribution adversarial examples generated by BAYES-TREX instead show systemic, human-interpretable network failures.

2.3 Confidence in Neural Networks

BAYES-TREX draws examples from confidence level sets in neural networks. Guo et al. showed that many neural networks are overconfident, with incorrect predictions often having high confidence [21]. While many approaches aim to address this network overconfidence problem [58, 34, 16, 7], our work is complementary to these efforts. Rather than altering the confidence of a neural network, BAYES-TREX instead infers examples of a particular confidence. If the model is overconfident, BAYES-TREX may return few, if any, samples with ambivalent predictions. Meanwhile, BAYES-TREX may find many misclassifications with high confidence. As such, BAYES-TREX can be used to assist in the diagnosis of overconfident networks. In our experiments (Section 4.7), we find that a popular domain adaptation technique produces a more overconfident model than a baseline model.

3 Methodology

Given a classifier $f : X \rightarrow \Delta_K$ which maps a data point to the probability simplex of K classes, the goal is to find an input $\mathbf{x} \in X$ in a known distribution $p(x)$ such that $f(\mathbf{x}) = \mathbf{p}$ for some particular prediction confidence $\mathbf{p} \in \Delta_K$. We consider the inference problem of sampling from the posterior

$$p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x}) p(f(\mathbf{x}) = \mathbf{p}|\mathbf{x}) \quad (1)$$

When the measure of the level set $\{\mathbf{x} : f(\mathbf{x}) = \mathbf{p}\}$ is small or even zero, sampling directly from this posterior using MCMC methods is infeasible: the posterior being zero everywhere outside of the level set means that it is unlikely for a Random-Walk Metropolis sampler to land on \mathbf{x} with non-zero posterior, and the gradient being zero everywhere outside of this level set means that a Hamiltonian Monte Carlo sampler does not have the necessary guidance toward the level set either.

To enable inference, we relax the formulation by accepting \mathbf{x} when $f(\mathbf{x})$ is close to \mathbf{p} (Fig. 1). Let u be a random variable (or vector), with conditional distribution $u|\mathbf{x} \sim q(f(\mathbf{x}))$ and target u^* , such that

$$f(\mathbf{x}) = \mathbf{p} \iff \mathbb{E}_q[u|\mathbf{x}] = u^*. \quad (2)$$

We now need to sample from the new posterior

$$p(\mathbf{x}|u = u^*) \propto p(\mathbf{x})p(u = u^*|\mathbf{x}). \quad (3)$$

Under expectation, when the sampled data has high probability, the classifier prediction matches \mathbf{p} .

For a given prediction confidence, assign $\mathbf{p} = \mathbf{p}^* = [p_1, \dots, p_K]^T$. Let $\mathbf{u} = [u_1, \dots, u_K]^T$.

$$u_i|\mathbf{x} \sim \mathcal{N}(f(\mathbf{x})_i, \sigma^2), \quad u_i^* = p_i, \quad (4)$$

where σ is a hyper-parameter determining the level-set relaxation (e.g., see Figure 1).

Since \mathbf{u} is sampled from a normal distribution, this formulation is asymptotically exact. Formally:

Lemma 3.1 $\lim_{\sigma \rightarrow 0} p(\mathbf{x}|\mathbf{u} = \mathbf{u}^*) = p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p})$. As σ goes to 0, the approximate posterior distribution approaches the true posterior distribution of interest.

We consider two specific instantiations for sampling high confidence and ambivalent examples. These instantiations correspond to important BAYES-TREX use cases, including sampling of in-distribution adversarial examples, novel class extrapolation examples, and class boundary examples.

1. High confidence: $\mathbf{p}_i = 1, \mathbf{p}_{\neg i} = 0$. The classifier should be as confident in its class i prediction as possible—irrespective of the correctness of the classification. For this case:

$$u|\mathbf{x} \sim \mathcal{N}(f(\mathbf{x})_i, \sigma^2), \quad u^* = 1. \quad (5)$$

2. Binary ambivalence: $\mathbf{p}_i = \mathbf{p}_j = 0.5, \mathbf{p}_{\neg i,j} = 0$. The classifier should be equally ambivalent between class i and class j , while not predicting any other classes. For this case:

$$u_1|\mathbf{x} \sim \mathcal{N}(|f(\mathbf{x})_i - f(\mathbf{x})_j|, \sigma_1^2), \quad u_1^* = 0, \quad (6)$$

$$u_2|\mathbf{x} \sim \mathcal{N}(\min(f(\mathbf{x})_i, f(\mathbf{x})_j) - \max_{k \neq i,j} f(\mathbf{x})_k, \sigma_2^2), \quad u_2^* = 0.5, \quad (7)$$

with $\mathbf{u} = [u_1, u_2]^T, \mathbf{u}^* = [0, 0.5]^T$. σ_1 and σ_2 are hyperparameters.

While the formulation in Eqn. 4 is applicable to arbitrary confidence assignments, it requires the number of auxiliary variables in \mathbf{u} to be equal to the number of classes, posing scalability issues for large numbers of classes. Eqns. 5–7 exemplify alternate formulations using fewer auxiliary variables.

Rather than sampling directly from the data distribution, we sample in the latent factor space of a generative model, Z . Z is mapped to X via a deterministic reconstruction function $g : Z \rightarrow X$.

To summarize, given

$$\mathbf{x} = g(\mathbf{z}) \quad (8)$$

$$u|\mathbf{z} \sim q(f(g(\mathbf{z}))), \quad (9)$$

$$p(\mathbf{z}|u = u^*) \propto p(\mathbf{z})p(u = u^*|\mathbf{z}), \quad (10)$$

we wish to sample \mathbf{z} according to Eqn. 10 and reconstruct the example $\mathbf{x} = g(\mathbf{z})$ for model inspection.

3.1 Inference Details

While there are many inference methods which may be used to sample from the posterior, we implement in particular two Markov Chain Monte Carlo (MCMC) methods: Random-Walk Metropolis and Hamiltonian Monte Carlo. In our experiments, we apply Random-Walk Metropolis when the prediction is not differentiable with respect to \mathbf{z} (i.e., when rendering is non-differentiable). When a gradient is available, we instead use Hamiltonian Monte Carlo. For the Hamiltonian Monte Carlo method, we use the No-U-Turn sampler [25, 43] and the probabilistic programming language Pyro for implementation [6]. All hyperparameters σ are arbitrarily chosen to be 0.05 for all experiments.

Selecting appropriate stopping criteria for MCMC algorithms is an open problem. State-of-the-art approaches require a gold standard inference algorithm [10] or specific properties of the posterior distribution, such as log-concavity [20]. As neither of these requirements are met for our general use cases, we select stopping criteria based on heuristic performance and cost of compute. As CLEVR scenes require GPU-intensive rendering, we typically select a stopping criteria of 500 samples. For MNIST and Fashion-MNIST, we select a stopping criteria of 2000 samples.

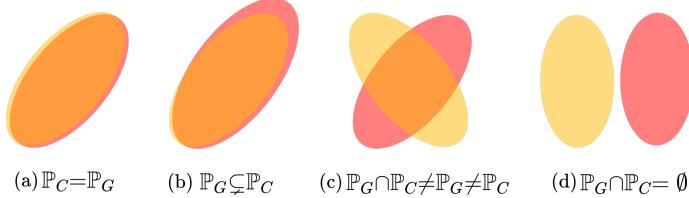


Figure 3: Relations of classifier training data (\mathbb{P}_C , red) and generator training data (\mathbb{P}_G , yellow). (a) \mathbb{P}_C and \mathbb{P}_G are equal. (b) \mathbb{P}_G is a subset of \mathbb{P}_C . (c) \mathbb{P}_G and \mathbb{P}_C overlap. (d) \mathbb{P}_C and \mathbb{P}_G are disjoint.

4 Experiments

BAYES-TREX enables the evaluation of a classifier on a target generative distribution \mathbb{P}_G irrespective of the distribution of the classifier training set \mathbb{P}_C . We demonstrate the versatility of BAYES-TREX on four relationships between \mathbb{P}_G and \mathbb{P}_C , as illustrated in Figure 3.

In Section 4.3, we consider $\mathbb{P}_C = \mathbb{P}_G$ (Figure 3(a)). This case finds in-distribution examples. In Section 4.5, we consider \mathbb{P}_G with narrower support than \mathbb{P}_C (Figure 3(b)). Specifically, \mathbb{P}_G cannot generate data from a particular class. In this case, high-confidence samples—as judged by the classifier C —with P_G as the generative distribution reveals in-distribution adversarial examples of C . In Section 4.6 and 4.7, we analyze the classifier C for out-of-distribution extrapolation and domain adaptation behaviours by considering \mathbb{P}_C and \mathbb{P}_G with overlapping or disjoint supports (Figure 3(c) and (d)). Representative results are in the main text; further results are in Appendices D - J.

4.1 Datasets

We evaluate BAYES-TREX on rendered data (CLEVR) and organic data (MNIST and Fashion-MNIST). In all CLEVR experiments, we use the pre-trained classifier distributed by the original authors [26]. Since CLEVR rendering is not differentiable, we use Random-Walk Metropolis for sampling. For probabilistic programming, we use a custom prior, consisting of a Gaussian proposal for the continuous variables (e.g., x -position) and categorical proposal for the discrete variables (e.g., color), with a high probability for the current value and uniform low probabilities for other values. For MNIST and Fashion-MNIST experiments, we train the classifiers; architectures in Appendix A. For domain adaptation analysis, we use the ADDA model and training code provided by the authors [60].

For CLEVR, we render scene graphs as the data generator. For (Fashion-)MNIST, we train and use both VAE and GAN models as the data generators. The latent distribution for these GANs and VAEs— $p(\mathbf{z})$ —is a unit Gaussian. VAEs and GANs are known to be imperfect representations of underlying data distributions [3]. To evaluate how well these generative models reproduce their training distributions, we report Fréchet Inception Distance scores [24]; see Appendix B.

4.2 Quantitative Evaluation

We evaluate how well BAYES-TREX-generated examples match the specified prediction target; a subset of these results are presented in Table 1 (full summary in Appendix C). This analysis confirms the inferred samples have predicted confidence closely matching the specified targets, indicating that the MCMC methods used by BAYES-TREX are successful for the tested domains and scenarios.

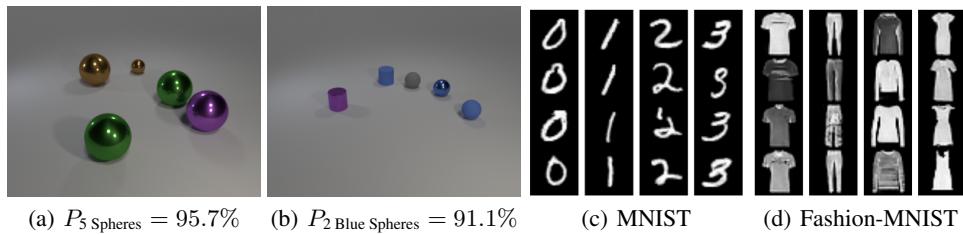


Figure 4: High-confidence samples. (a, b): CLEVR, with target class and predicted confidence. (c): MNIST Digit 0-3. (d): Fashion-MNIST classes of T-shirt, trousers, pullover and dress.

Table 1: Mean and standard deviation of the predicted confidence of the samples over 10 trials. Prediction confidence is for the target class, or two target classes in binary ambivalent and graded confidence cases. Appendix C presents the comprehensive extension of this table for all experiments.

Evaluation	Dataset	Target	Prediction Confidence
High Confidence	MNIST	$P_4 = 1$	0.998 ± 0.008
	Fashion	$P_{\text{Coat}} = 1$	0.983 ± 0.021
	CLEVR	$P_2 \text{ Blue Spheres} = 1$	0.892 ± 0.245
Binary Ambivalent	MNIST	$(P_1 = 0.5, P_7 = 0.5)$	$(0.490 \pm 0.024, 0.489 \pm 0.025)$
	Fashion	$(P_{\text{T-shirt}}, P_{\text{Dress}})$	$(0.480 \pm 0.022, 0.482 \pm 0.021)$
Graded Confidence	MNIST	$(P_8 = 0.2, P_9 = 0.8)$	$(0.170 \pm 0.039, 0.788 \pm 0.040)$
	Fashion	$(P_{\text{T-shirt}} = 0.7, P_{\text{Trousers}} = 0.3)$	$(0.685 \pm 0.039, 0.281 \pm 0.040)$
In-Distribution Adv.	MNIST	$P_8 = 1$	0.982 ± 0.024
	Fashion	$P_{\text{Bag}} = 1$	0.967 ± 0.026
	CLEVR	$P_1 \text{ Cube} = 1$	0.929 ± 0.062
OOD Extrapolation	MNIST	$P_6 = 1$	0.998 ± 0.007
	Fashion	$P_{\text{Sandal}} = 1$	0.995 ± 0.013
	CLEVR	$P_1 \text{ Cylinder} = 1$	0.959 ± 0.031
Domain Adaptation	MNIST	$P_5 = 1$	0.998 ± 0.007

4.3 High, Ambivalent, and Graded Confidence Examples

We first evaluate BAYES-TREX by finding highly confident examples. Figure 4 depicts samples on all three datasets; additional high confidence examples are in Appendix D. Next, we find examples where the neural network is ambivalent between two classes. We use a VAE generative model for (Fashion-)MNIST. Figure 5 shows a matrix of ambivalent examples from each pair of classes (e.g. 0v1, 0v2, ..., 8v9). As is expected, not all pairs result in successful sampling: for example, we were unable to find an ambivalent example with equal prediction confidence between a 7 and a 0. Note these examples are ambivalent from the classifier’s perspective; some may be readily classified by a human. Appendix E contains more examples and a latent space visualization of class boundaries. In addition to fully ambivalent examples, BAYES-TREX can also sample from level sets which interpolate between classes (Figure 5(d)–5(e)). Appendix F contains a complete interpolation.

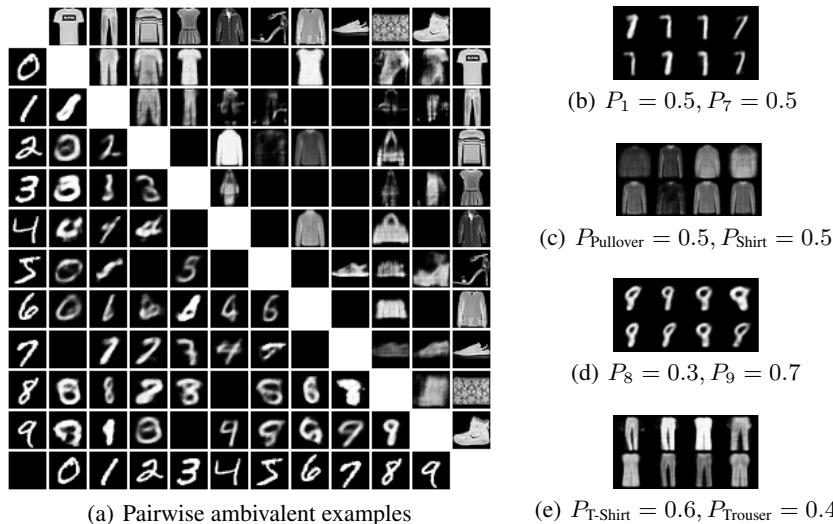


Figure 5: Ambivalent and graded confidence examples. 5(a): Each entry of the matrix is an ambivalent example for the classes on its row and column. Blacked-out cells indicate sampling failures. 5(b) and 5(c): Additional pairwise ambivalent examples. 5(d) and 5(e): Graded confidence examples.

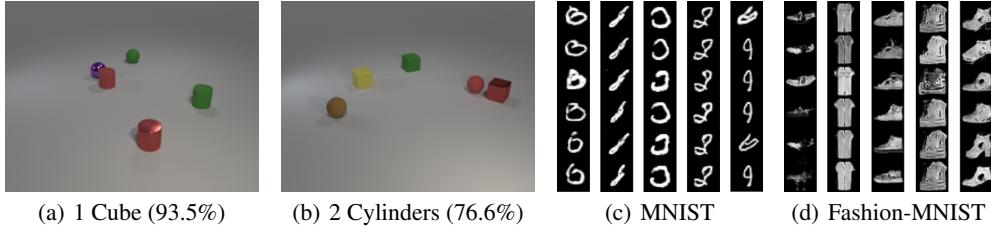


Figure 6: In-distribution adversarial examples, using a generator which is unable to produce examples of a target class. (a): CLEVR, 1 Cube. Note that no cube is present in the sample. (b): CLEVR, 2 Cylinders. (c) MNIST, adversarial examples for 0-4. 0s are composed of 6s; 1s of 8s; 2s of 0s, and so on. (d) Fashion-MNIST, adversarial examples for sandal, shirt, sneaker, bag, and ankle boot.

4.4 Ambivalent Examples and Overconfidence

Using VAEs, BAYES-TREX can infer ambivalent examples for (Fashion-)MNIST. However, inference fails when using GANs. The FID scores show the (Fashion-)MNIST GANs are better generators than their VAE counterparts (see App. B); the GAN-generated images are generally sharper and more visually realistic. We assert the failure to infer ambivalent examples with a GAN is due to the network being consistently confident when evaluating realistic, sharp images. We experimentally verify this: when we explicitly train a classifier to have ambivalent confidence using a KL-divergence loss, using a GAN to sample ambivalent targets succeeds. This experiment is described in detail in Appendix G.

BAYES-TREX is also unable to generate ambiguous examples for CLEVR. While (Fashion-)MNIST networks are high performing, with accuracies of $\approx 99\%$, CLEVR accuracy is $\approx 60\%$ [26]. By definition, a high-performing network is not necessarily overconfident, even when highly confident on all examples [21]. However, BAYES-TREX’s inability to sample ambivalent CLEVR examples is likely due to overconfidence. This problem has previously been observed in CLEVR-style settings [29].

4.5 In-Distribution Adversarial Examples

By revoking the generator’s ability to create objects of specific classes, BAYES-TREX can also find “in-distribution” adversarial examples. Like traditional adversarial examples, these cause network failures. However, these are better suited for human assessment as they are not composed of apparent noise or imperceptible perturbations. For a classifier trained on the entire dataset, we can expose class- c adversarial examples by sampling using a data distribution that excludes class- c examples. The generator distribution support is a subset of the classifier’s support (Fig. 3(b)). For CLEVR, we revoke the generative model’s ability to produce objects of a target class and sample high confidence images for that target (Fig. 6(a) and 6(b)). In Fig. 6(a), the classifier is highly confident there is one cube. As confirmed by the saliency map (App. K), the classifier mistakes the shiny red cylinder for a cube. Collating such examples can enable data augmentation to increase network reliability [15].

For (Fashion-)MNIST, for each class c , we train a GAN on the dataset with class- c removed. Fig. 6(c) and 6(d) depict adversarial examples for digits 0-4 in MNIST and sandal, shirt, sneaker, bag, and ankle boot in Fashion-MNIST, respectively. We observe several interesting phenomena. For MNIST, some thin 8s are classified as 1s and particular styles of 6s and 9s are classified as 4s. For Fashion-MNIST, most adversarial examples come from semantically similar classes, e.g. sneaker \leftrightarrow ankle boot. Less intuitively, chunky shoes are likely to be classified as bags. Additional experiments in Appendix H.

4.6 Novel Class Extrapolation Examples

BAYES-TREX can also be used to understand model extrapolation behaviours on novel classes. Specifically, the generator can produce examples of a novel class but cannot produce examples of a target class. Novel classes are not present in the classifier’s training data (Fig. 3(c, d)). For a CLEVR instance, the generator can produce a novel object—a cone—but cannot generate cubes. We observe that the classifier confidently mistakes cones for cubes (Fig. 7(a), 7(b); App. K). When an MNIST classifier does not see digits 7 and 8 during training, it confidently mistakes these for digits 0, 1, and 9 (Fig. 7(c)). A Fashion-MNIST classifier confuses sneakers for sandals and ankle boots. While confusing sandals and ankle boots may be reasonable extrapolation, the classifier also mistakes different bags as sandals, shirts, and ankle boots (Fig. 7(d)). Additional experiments in Appendix I.

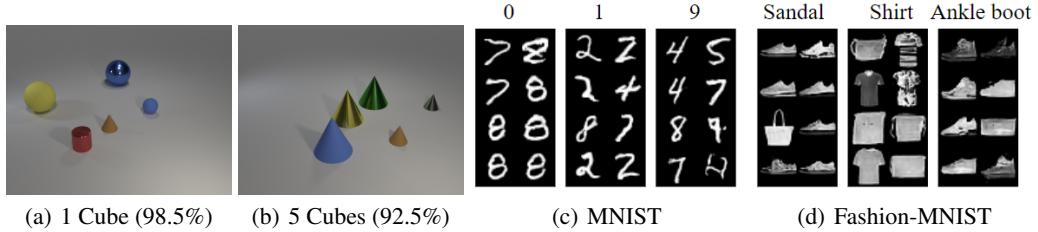


Figure 7: OOD extrapolation samples. (a, b): For CLEVR, we introduce a new shape—a cone—to the generative model, revoke the generative model’s ability to produce target class objects, and infer high-confidence examples. (c, d): For (Fashion-)MNIST, we train classifiers on subsets of the data (digits 0, 1, 3, 6, 9 for MNIST and Pullover, Dress, Sandal, Shirt, and Ankle boot for Fashion-MNIST), and train the data generators with the excluded data. The incorrect predicted class label is shown above the images. Average prediction confidence is $\approx 99\%$. Additional details in Appendix I.



Figure 8: Highly confident examples for each class (0 to 9) of the baseline model and ADDA model. A manual labeling reveals *more* misclassified high-confidence examples with ADDA (Appendix J).

4.7 Domain Adaptation

BAYES-TREX can also be used to analyze domain adaptation. Specifically, consider the SVHN \rightarrow MNIST domain adaptation problem. We train two classifiers, a baseline classifier on labeled SVHN data only, and the adversarial discriminative domain adaptation (ADDA) classifier [60] on labeled SVHN data and unlabeled MNIST data. Indeed, domain adaptation improves classification accuracy: the baseline classifier achieves 61% accuracy on MNIST while the ADDA classifier achieves 71%.

But is this the whole story? Using BAYES-TREX, we sample high confidence examples for each classifier using an MNIST GAN (Fig. 8). We re-label those inferred high-confidence examples and compute the accuracy. We find the baseline model achieves higher accuracy on these high confidence examples (80% vs 71%; see Supp. Table 8). This fine-grained analysis suggests the ADDA model is more overconfident than the baseline. Experiment protocol and analysis in Appendix J.

5 Discussion

We propose a model inspection method of sampling level set examples of known prediction confidence. BAYES-TREX selects examples such that the sampled data has high likelihood under a known data distribution. To achieve this, we use a Bayesian inference formulation and use probabilistic programming to sample from the posterior. On both procedurally-defined data (CLEVR) and organic data (MNIST and Fashion-MNIST), we demonstrate that BAYES-TREX can sample close to the specified prediction confidence target while remaining in the known distribution. In the experiments, we show how BAYES-TREX can be used to synthesize ambivalent predictions, uncover in-distribution adversarial examples, and understand novel-class extrapolation and domain adaptation behaviours.

By specifying the underlying distributions, BAYES-TREX can generate examples beyond the scope of the test set. These generated examples can facilitate transparency by example. Still, extracting insights from individual examples remains challenging, even with the aid of local explanation methods (e.g., Fig 2). We instead need to cluster and visualize trends in these generated examples, to increase model legibility. To accompany clustering, we further need to estimate coverage of the level set.

The current BAYES-TREX formulation only works for classification. Beyond classification, structured prediction is necessary for safety-critical robotics applications such as motion planning. As future work, to adapt BAYES-TREX for such models, we must incorporate dependency among outputs.

Broader Impact

BAYES-TREX can allow humans to build more accurate mental models of how neural networks make decisions. Further, BAYES-TREX can be useful for debugging, interpreting, and understanding networks—all of which can help us build *better*, less biased, increasingly human-aligned models.

However, BAYES-TREX is subject to the same caveats as typical software testing approaches: the absence of exposed bad samples does not mean the system is free from defects. One concern is how system designers and users will interact with BAYES-TREX in practice. If BAYES-TREX does not reveal degenerate examples, these stakeholders might develop inordinate trust [33] in their models.

Additionally, the examples which BAYES-TREX generates may be used as inputs to downstream local explanation methods. As a community, we know that many of these local explanations can be challenging to understand [49, 47], misleading [1, 30, 52], or susceptible to adversarial attacks [54]. In human-human interaction, even nonsensical explanations can increase compliance [32]. As we build post-hoc explanation techniques, we must evaluate whether the produced explanations help humans moderate trust and act appropriately—for example, by overriding the model’s decisions.

References

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.
- [2] Martin Arjovsky, Léon Bottou, Ishaaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- [3] Sanjeev Arora and Yi Zhang. Do GANs actually learn the distribution? An empirical study. *arXiv preprint arXiv:1706.08224*, 2017.
- [4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning*, 2018.
- [5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.
- [6] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.
- [7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1613–1622, 2015.
- [8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.
- [9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019.
- [10] Marco Cusumano-Towner and Vikash K Mansinghka. AIDE: an algorithm for measuring the accuracy of probabilistic inference algorithms. In *Advances in Neural Information Processing Systems*, pages 3000–3010, 2017.
- [11] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.
- [12] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [13] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.
- [14] Christian Ettmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. In *ICML*, 2019.
- [15] Daniel J Fremont, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. Scenic: a language for scenario specification and scene generation. In *ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2019.
- [16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

- [17] Amirata Ghorbani, James Wexler, and Been Kim. Automating interpretability: Discovering and testing visual concepts learned by neural networks. *arXiv preprint arXiv:1902.03129*, 2019.
- [18] 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*, 2018.
- [19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [20] Jackson Gorham and Lester Mackey. Measuring sample quality with stein’s method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.
- [21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.
- [22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- [23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.
- [25] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
- [26] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
- [27] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, 2014.
- [28] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*, 2017.
- [29] Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-So-CLEVR: learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4):20180011, 2018.
- [30] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.
- [31] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.
- [32] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of “placebic” information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.
- [33] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.
- [34] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples, 2017.
- [35] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. *arXiv preprint arXiv:1809.03113*, 2018.
- [36] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [37] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *arXiv preprint arXiv:1905.00441*, 2019.
- [38] Shiyu Liang, Yixuan Li, and R Srikant. Principled detection of out-of-distribution examples in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- [39] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.
- [40] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

- [41] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.
- [42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [43] Radford M Neal et al. MCMC using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.
- [44] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.
- [45] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, pages 3387–3395, 2016.
- [46] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.
- [47] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019.
- [48] Augustus Odena, Catherine Olsson, David Andersen, and Ian Goodfellow. TensorFuzz: Debugging neural networks with coverage-guided fuzzing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4901–4911, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- [49] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. <https://distill.pub/2017/feature-visualization>.
- [50] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1–8. IEEE, 2018.
- [51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [52] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [53] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.
- [54] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. How can we fool LIME and SHAP? adversarial attacks on post hoc explanation methods. *arXiv preprint arXiv:1911.02508*, 2019.
- [55] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.
- [56] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.
- [57] Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.
- [58] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 13888–13899, 2019.
- [59] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.
- [60] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [61] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.
- [62] Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection in classifiers via generation. *arXiv preprint arXiv:1910.04241*, 2019.
- [63] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- [64] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

A Network Architecture for MNIST & Fashion-MNIST

For all experiments on MNIST and Fashion-MNIST, the VAE architecture is shown in Table 2 (left), and the GAN architecture is shown in Table 2 (right). For all experiments on MNIST and Fashion-MNIST except for the domain adaptation analysis, the classifier architecture is shown in Table 3 (left). The classifier used in the domain adaptation analysis is the LeNet architecture, following the provided source code, shown in Table 3 (right). All experiments are performed with a single NVIDIA GeForce 1080 GPU.

Table 2: Left: VAE architecture; right: GAN architecture.

Encoder input: $28 \times 28 \times 1$	Input: 5 (latent dimension)
Flatten	Reshape $1 \times 1 \times 5$
Fully-connected 784×400	Conv-transpose: 512 filters, size= 4×4 , stride = 1
ReLU	Batch-norm, ReLU
Mean: Fully-connected 400×5	Conv-transpose: 256 filters, size= 4×4 , stride = 2
Log-variance: Fully-connected 400×5	Batch-norm, ReLU
Decoder input: 5 (latent dimension)	Conv-transpose: 128 filters, size= 4×4 , stride = 2
Fully-connected 5×400	Batch-norm, ReLU
ReLU	Conv-transpose: 64 filters, size= 4×4 , stride = 2
Fully-connected 400×784	Batch-norm, ReLU
Reshape $28 \times 28 \times 1$	Conv-transpose: 1 filters, size= 1×1 , stride = 1
Sigmoid	Sigmoid

Table 3: Left: classifier architecture in all experiments except domain adaptation analysis; right: LeNet classifier architecture in domain adaptation analysis (used in code released by ADDA authors).

Input: $28 \times 28 \times 1$	Input: $28 \times 28 \times 1$
Conv: 32 filters, size = 3×3 , stride = 1	Conv: 20 filters, size = 5×5 , stride = 1
ReLU	ReLU
Conv: 64 filters, size = 3×3 , stride = 1	Max-pool, size = 2×2
Drop-out, prob = 0.25	Conv: 50 filters, size = 5×5 , stride = 1
Max-pool, size = 2×2	ReLU
Flatten	Max-pool, size = 2×2
Fully-connected 9216×128	Flatten
ReLU	Fully-connected 800×500
Drop-out, prob = 0.5	ReLU
Fully-connected 128×10	Fully-connected 500×10
Soft-max	Soft-max

B Evaluating Learned Data Distribution Fit: Fréchet Inception Distance (FID) Scores

Fréchet Inception Distance (FID) scores are one measure for evaluating how well a learned data distribution fits the ground truth distribution. We report these FID scores for all learned generative models in Table 4. We note that GANs consistently outperform VAEs for all data distributions. We further note that all generative models perform better on MNIST data than on Fashion-MNIST. MNIST GANs are the best performing models; Fashion-MNIST VAEs are the worst performing models. We measure FID for each VAE and GAN. In particular, these include the generative models trained for exposing in-distribution adversarial examples by leaving out one class at a time. Further, these FID scores include generative models trained on select subsets of the data for novel class extrapolation studies. For example, one GAN is trained only on the MNIST digits {2, 4, 5, 7, 8}.

Table 4: Fréchet Inception Distance (FID) scores for all learned data distributions; a lower score indicates a better distribution fit. We note that GAN performance is markedly superior to VAE performance, and MNIST performance is superior to Fashion-MNIST performance across all models. Results are computed across 1000 samples. Fashion-MNIST class mapping: [0: T-shirt, 1: Trouser, 2: Pullover, 3: Dress, 4: Coat, 5: Sandal, 6: Shirt, 7: Sneaker, 8: Bag, 9: Ankle boot].

Model	Dataset	Subset	FID	Model	Dataset	Subset	FID
GAN	MNIST	All	11.83	MNIST	All	72.33	
		Without 0	12.10		Without 0	71.28	
		Without 1	12.08		Without 1	75.36	
		Without 2	13.57		Without 2	64.77	
		Without 3	12.71		Without 3	63.66	
		Without 4	12.25		Without 4	66.96	
		Without 5	12.21		Without 5	63.31	
		Without 6	11.86		Without 6	67.64	
		Without 7	11.64		Without 7	62.45	
		Without 8	12.31		Without 8	64.14	
		Without 9	12.34		Without 9	66.57	
		{2, 4, 5, 7, 8}	13.45	VAE	—	—	
GAN	Fashion	All	29.44		All	87.89	
		Without 0	28.91		Without 0	89.21	
		Without 1	31.18		Without 1	92.02	
		Without 2	30.11		Without 2	91.20	
		Without 3	28.95		Without 3	85.51	
		Without 4	30.43		Without 4	88.38	
		Without 5	27.67		Without 5	84.17	
		Without 6	29.68		Without 6	85.58	
		Without 7	28.56		Without 7	84.93	
		Without 8	30.87		Without 8	83.66	
		Without 9	29.22		Without 9	81.48	
		{0, 1, 4, 7, 8}	33.11	—	—	—	

C Full Quantitative Confidence Results

Tables 5, 6, and 7 present the full extension of Table 1 in the main text. These results show that the inferred samples have predicted confidence closely matching the specified confidence targets. This indicates the MCMC methods used by BAYES-TREX are successful for the tested domains and scenarios. Queries for 5 Cubes in the out-of-distribution CLEVR experiments use a stopping criterion of 1500 samples instead of the standard 500. Averages reported across 10 inference runs.

Table 5: Full table of measured prediction confidence values for BAYES-TREX samples on high-confidence examples (left) and in-distribution adversarial examples (right). Accompanying visual results in Appendices D and H.

Target	Prediction Confidence	Target	Prediction Confidence
$P_0 = 1$	0.999 ± 0.006	$P_0 = 1$	0.981 ± 0.027
$P_1 = 1$	0.999 ± 0.003	$P_1 = 1$	0.953 ± 0.028
$P_2 = 1$	0.999 ± 0.006	$P_2 = 1$	0.968 ± 0.028
$P_3 = 1$	0.999 ± 0.005	$P_3 = 1$	0.969 ± 0.027
$P_4 = 1$	0.998 ± 0.008	$P_4 = 1$	0.955 ± 0.030
$P_5 = 1$	0.999 ± 0.006	$P_5 = 1$	0.990 ± 0.018
$P_6 = 1$	0.998 ± 0.007	$P_6 = 1$	0.970 ± 0.026
$P_7 = 1$	0.998 ± 0.007	$P_7 = 1$	0.968 ± 0.029
$P_8 = 1$	0.999 ± 0.004	$P_8 = 1$	0.982 ± 0.024
$P_9 = 1$	0.998 ± 0.007	$P_9 = 1$	0.983 ± 0.022
$P_{\text{T-Shirt}} = 1$	0.991 ± 0.016	$P_{\text{T-Shirt}} = 1$	0.964 ± 0.029
$P_{\text{Trouser}} = 1$	0.999 ± 0.006	$P_{\text{Trouser}} = 1$	(sample failure)
$P_{\text{Pullover}} = 1$	0.984 ± 0.019	$P_{\text{Pullover}} = 1$	0.886 ± 0.027
$P_{\text{Dress}} = 1$	0.993 ± 0.008	$P_{\text{Dress}} = 1$	0.970 ± 0.026
$P_{\text{Coat}} = 1$	0.983 ± 0.021	$P_{\text{Coat}} = 1$	0.938 ± 0.030
$P_{\text{Sandal}} = 1$	0.998 ± 0.008	$P_{\text{Sandal}} = 1$	0.968 ± 0.030
$P_{\text{Shirt}} = 1$	0.987 ± 0.020	$P_{\text{Shirt}} = 1$	0.938 ± 0.032
$P_{\text{Sneaker}} = 1$	0.994 ± 0.016	$P_{\text{Sneaker}} = 1$	0.969 ± 0.028
$P_{\text{Bag}} = 1$	0.999 ± 0.006	$P_{\text{Bag}} = 1$	0.967 ± 0.026
$P_{\text{Ankle Boot}} = 1$	0.996 ± 0.012	$P_{\text{Ankle Boot}} = 1$	0.971 ± 0.027
$P_5 \text{ Spheres} = 1$	0.943 ± 0.020	$P_1 \text{ Cube} = 1$	0.929 ± 0.062
$P_2 \text{ Blue Spheres} = 1$	0.892 ± 0.245	$P_1 \text{ Cylinder} = 1$	0.972 ± 0.021
		$P_1 \text{ Sphere} = 1$	0.843 ± 0.266
		$P_2 \text{ Cylinders} = 1$	0.545 ± 0.230

Table 6: Full table for graded confidence examples for MNIST (left) and Fashion-MNIST (right); accompanying visual results in Appendix F.

Target	Prediction Confidence	Target	Prediction Confidence
$P_8 = 0.0, P_9 = 1.0$	$(0.002 \pm 0.006, 0.990 \pm 0.016)$	$P_{\text{T-Shirt}} = 0.0, P_{\text{Trousers}} = 1.0$	$(0.001 \pm 0.004, 0.995 \pm 0.012)$
$P_8 = 0.1, P_9 = 0.9$	$(0.030 \pm 0.039, 0.936 \pm 0.051)$	$P_{\text{T-Shirt}} = 0.1, P_{\text{Trousers}} = 0.9$	$(0.026 \pm 0.035, 0.950 \pm 0.050)$
$P_8 = 0.2, P_9 = 0.8$	$(0.170 \pm 0.039, 0.788 \pm 0.040)$	$P_{\text{T-Shirt}} = 0.2, P_{\text{Trousers}} = 0.8$	$(0.166 \pm 0.040, 0.791 \pm 0.041)$
$P_8 = 0.3, P_9 = 0.7$	$(0.275 \pm 0.041, 0.682 \pm 0.040)$	$P_{\text{T-Shirt}} = 0.3, P_{\text{Trousers}} = 0.7$	$(0.275 \pm 0.037, 0.686 \pm 0.038)$
$P_8 = 0.4, P_9 = 0.6$	$(0.378 \pm 0.040, 0.578 \pm 0.040)$	$P_{\text{T-Shirt}} = 0.4, P_{\text{Trousers}} = 0.6$	$(0.379 \pm 0.038, 0.586 \pm 0.038)$
$P_8 = 0.5, P_9 = 0.5$	$(0.477 \pm 0.039, 0.477 \pm 0.039)$	$P_{\text{T-Shirt}} = 0.5, P_{\text{Trousers}} = 0.5$	$(0.436 \pm 0.040, 0.459 \pm 0.040)$
$P_8 = 0.6, P_9 = 0.4$	$(0.581 \pm 0.038, 0.374 \pm 0.039)$	$P_{\text{T-Shirt}} = 0.6, P_{\text{Trousers}} = 0.4$	$(0.583 \pm 0.038, 0.382 \pm 0.037)$
$P_8 = 0.7, P_9 = 0.3$	$(0.680 \pm 0.041, 0.275 \pm 0.039)$	$P_{\text{T-Shirt}} = 0.7, P_{\text{Trousers}} = 0.3$	$(0.685 \pm 0.039, 0.281 \pm 0.040)$
$P_8 = 0.8, P_9 = 0.2$	$(0.788 \pm 0.040, 0.167 \pm 0.041)$	$P_{\text{T-Shirt}} = 0.8, P_{\text{Trousers}} = 0.2$	$(0.790 \pm 0.037, 0.177 \pm 0.037)$
$P_8 = 0.9, P_9 = 0.1$	$(0.926 \pm 0.050, 0.039 \pm 0.040)$	$P_{\text{T-Shirt}} = 0.9, P_{\text{Trousers}} = 0.1$	$(0.936 \pm 0.045, 0.029 \pm 0.041)$
$P_8 = 1.0, P_9 = 0.0$	$(0.989 \pm 0.016, 0.002 \pm 0.007)$	$P_{\text{T-Shirt}} = 1.0, P_{\text{Trousers}} = 0.0$	$(0.985 \pm 0.019, 0.000 \pm 0.003)$

Table 7: Full table for out-of-distribution, novel class extrapolation (left) and domain adaptation (right) samples (using the ADDA model). Accompanying visual results in Appendices I and J.

Target	Prediction Confidence	Target	Prediction Confidence
$P_0 = 1$	0.976 ± 0.025	$P_0 = 1$	0.996 ± 0.011
$P_1 = 1$	0.988 ± 0.186	$P_1 = 1$	0.994 ± 0.014
$P_3 = 1$	0.987 ± 0.020	$P_2 = 1$	0.998 ± 0.008
$P_6 = 1$	0.989 ± 0.018	$P_3 = 1$	0.994 ± 0.015
$P_9 = 1$	0.995 ± 0.013	$P_4 = 1$	0.997 ± 0.010
$P_{\text{Pullover}} = 1$	0.991 ± 0.016	$P_5 = 1$	0.998 ± 0.007
$P_{\text{Dress}} = 1$	0.994 ± 0.013	$P_6 = 1$	0.996 ± 0.011
$P_{\text{Sandal}} = 1$	0.995 ± 0.013	$P_7 = 1$	0.996 ± 0.011
$P_{\text{Shirt}} = 1$	0.994 ± 0.012	$P_8 = 1$	0.995 ± 0.013
$P_{\text{Ankle Boot}} = 1$	0.993 ± 0.015	$P_9 = 1$	0.996 ± 0.012
$P_{1 \text{ Cube}} = 1$	0.983 ± 0.014		
$P_{1 \text{ Cylinder}} = 1$	0.959 ± 0.031		
$P_{1 \text{ sphere}} = 1$	0.969 ± 0.022		
$P_{5 \text{ Cubes}} = 1$	0.921 ± 0.029		

D More Details on the High-Confidence Example Experiment

Figure 9 presents additional high-confidence examples for CLEVR. Figure 10 presents additional high-confidence examples for MNIST and Fashion-MNIST.

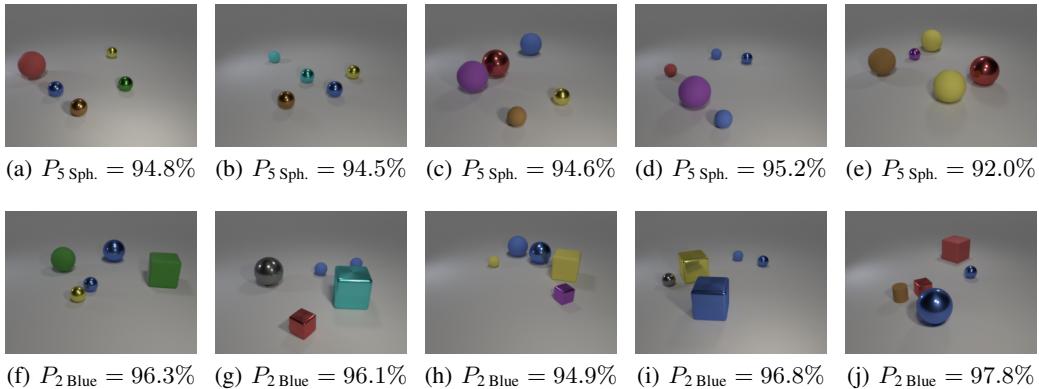


Figure 9: Above, 9(a)–9(e): selected examples classified as containing 5 spheres with high confidence. Below, 9(f)–9(j): selected examples classified as containing 2 blue spheres with high confidence.



Figure 10: High-confidence examples from MNIST and Fashion-MNIST. There are no misclassifications. MNIST columns represent digit 0 to 9, respectively. Fashion-MNIST columns represent t-shirt, trousers, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot, respectively.

E More Details on the Binary Ambivalent Example Experiments

Figure 5(a) in the main text presents a matrix showing one sampled image for each pair of classes from both MNIST and Fashion-MNIST (e.g., Digit 1 vs. Digit 7; T-shirt vs. Pullover). Figure 11 presents additional breakdowns for two pairs, Digit 1 vs. Digit 7 from MNIST and T-shirt vs. Pullover from Fashion-MNIST. The violin plot confirms that the neural network is indeed making the target predictions of ambivalence, and the latent visualization indicates that the samples lie around the class boundaries and are in-distribution (i.e., having close proximity to others from the prior in the latent visualization with T-SNE [42] dimensionality reduction).

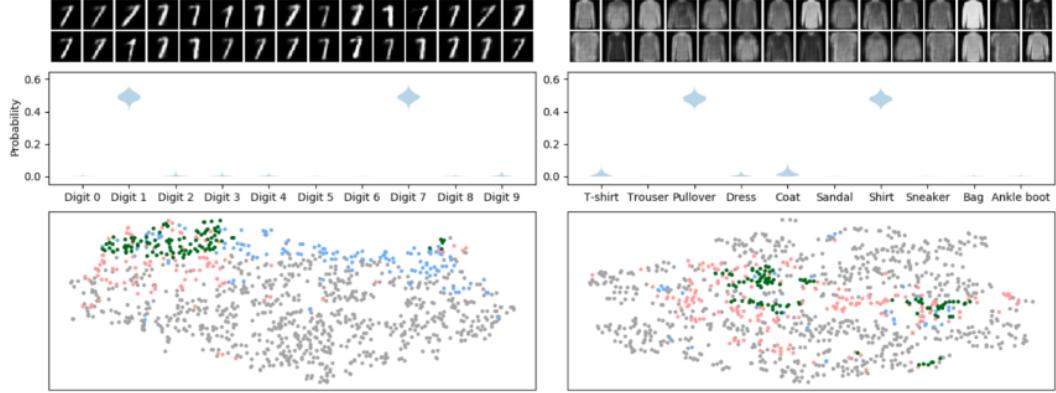


Figure 11: Left: ambivalent samples for digit 1 vs. 7 in MNIST. Right: ambivalent samples for pullover vs. shirt in Fashion-MNIST. Top: 30 sampled images. Middle: the ambivalent predictions made by the classifier. Bottom: latent space visualization. Green dots represent ambivalent samples from the posterior, red and blue dots represent samples from the prior that are predicted by the classifier to be either class of interest, and gray dots represent other samples from the prior. The ambivalent samples are on the class boundaries.

F More Details on the Graded-Confidence Example Experiments

BAYES-TREX can be used to sample confidence predictions which interpolate between classes. In Figure 12, we show MNIST samples which interpolate from $(P_8 = 1.0, P_9 = 0.0)$ to $(P_8 = 0.0, P_9 = 1.0)$ and Fashion-MNIST samples from $(P_{\text{T-shirt}} = 1.0, P_{\text{Trousers}} = 0.0)$ to $(P_{\text{T-shirt}} = 0.0, P_{\text{Trousers}} = 1.0)$ over intervals of 0.1, using a VAE as the generator. The target probability for other classes is 0.

By interpolating between two quite different classes, we can gain some insight into the model’s behaviour. For example, in Figure 12, we see that the interpolation from 8 to 9 generally shrinks the bottom circle toward a stroke, which is the key difference between digits 8 and 9 to be the width of the bottom circle. For Fashion-MNIST, we consider the case of T-shirt vs. Trousers. We uncover that the presence of two legs is important for trousers classification, even appearing in samples with $(p_{\text{T-shirt}} = 0.9, p_{\text{Trousers}} = 0.1)$ (second column); by contrast, a wider top and the appearance of sleeves are important properties for T-shirt classification: most of the interpolated samples have a short sleeve on the top but two distinct legs on the bottom.

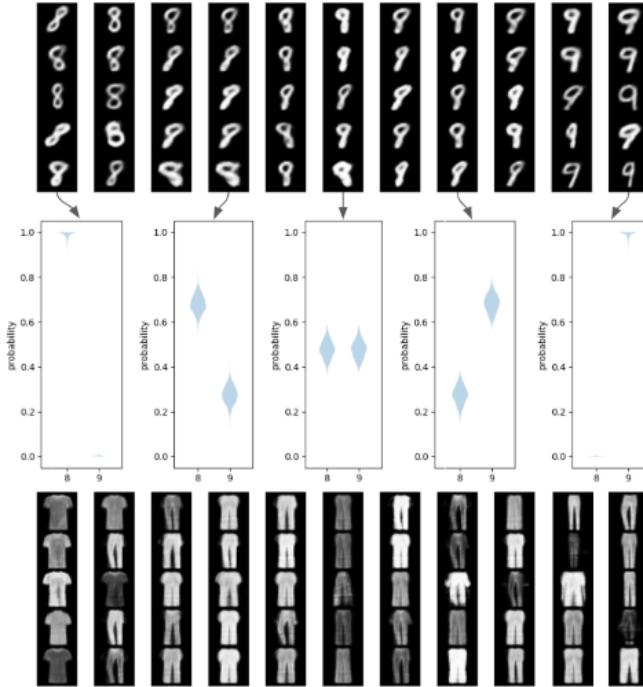


Figure 12: Confidence interpolation between digit 8 and 9 for MNIST and between T-shirt and trousers for Fashion-MNIST. Each of the 11 columns show samples of confidence ranging from $[P_{\text{class a}} = 1.0, P_{\text{class b}} = 0.0]$ (left) to $[P_{\text{class a}} = 0.0, P_{\text{class b}} = 1.0]$ (right), with an interval of 0.1. Select confidence plots for MNIST samples are shown in the 2nd row.

G More Details on the Investigation into GAN Sampling Failure

We failed to sample binary ambivalent and graded confidence (Fashion-)MNIST examples using a GAN as the generative distribution. There are two possible explanations:

1. the formulation of the posterior inference problem somehow is not suitable for a GAN-induced distribution, and/or a Hamiltonian Monte-Carlo sampler cannot easily sample from the posterior; or
2. there are no GAN-generated examples that match the ambivalent prediction targets.

Empirically, we found that the neural network confidences on GAN images are very extreme (i.e. probability very close to 1 for one class and very close to 0 for all other classes). In comparison, we found that the somewhat blurry images produced by VAE have less extreme prediction confidences. We further confirmed that GANs are better data generators than VAEs for our domains by computing FID scores (Appendix B). Therefore, we believe the GAN is largely unable to produce examples which result in ambivalent predictions. This suggests the classifier is consistently confident on high-quality examples.

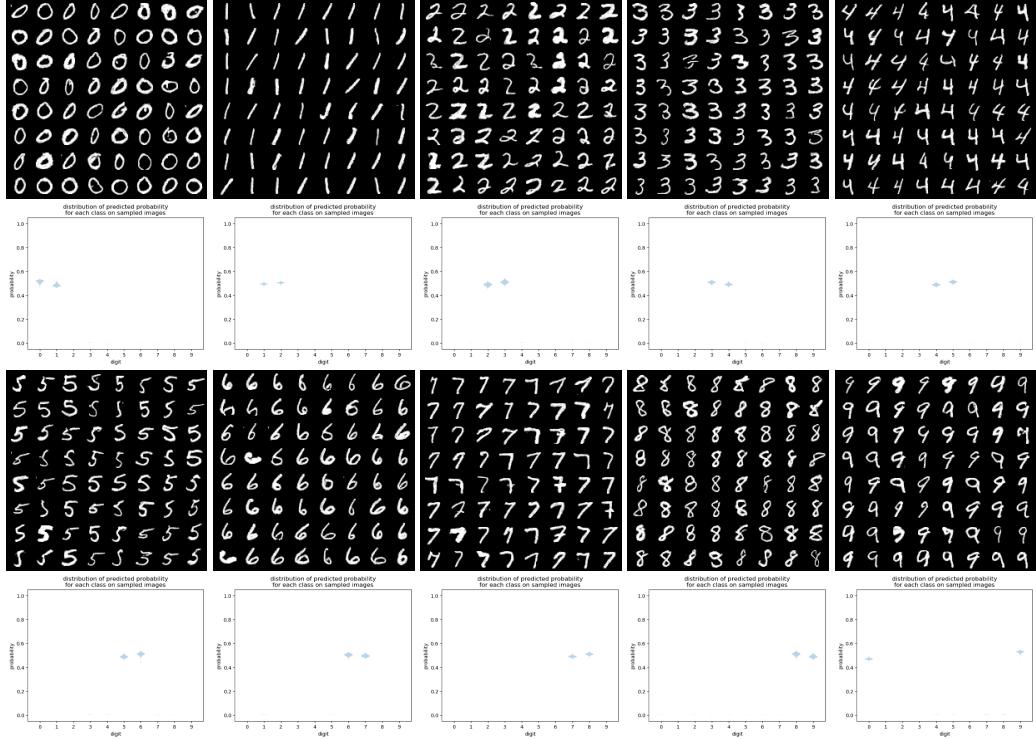


Figure 13: Sampling results with an explicitly ambivalent classifier and a GAN generator. Top 2 rows: digit i vs. $i + 1$ for $i \in \{0, 1, 2, 3, 4\}$. Bottom 2 rows: digit i vs. $i + 1 \pmod{10}$ for $i \in \{5, 6, 7, 8, 9\}$.

To verify this, we train a classifier on MNIST with intentionally binary ambivalent targets. Specifically, the loss on an image \mathbf{x} with groundtruth digit $y \in \{0, 1, \dots, 9\}$ is

$$l(y, f(\mathbf{x})) = \text{KL}(P_y, f(\mathbf{x})), \quad (11)$$

$$P_{y,i} = \begin{cases} 0.5 & i = y \text{ or } i = (y + 1) \pmod{10}, \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

In words, we use a KL-divergence loss between the network predicted probability $f(\mathbf{x})$ and a target that is *defined* to be binary ambivalent between digit i and digit $i + 1$ (wrapping around at $10 = 0$).

After training, we find that the classifier is indeed confused between the two respective classes, and the classification accuracy is very close to 50%, as expected. Then we use a GAN to sample image that are ambivalent between digit i and $i + 1 \pmod{10}$. Indeed, the sampler successfully returns images of class i from the posterior (Fig. 13). The violin plots also confirm that the examples are correctly at the target level set. Note that sampling examples that are binary ambivalent between other pairs of digits i and j still fails because the network simply does not make such predictions on any images drawn from the GAN distribution. Similarly, for this experiment, we cannot sample graded confidence examples with any confidence target other than (0.5, 0.5).

This raises the question of whether the typical MNIST and Fashion-MNIST are overconfident. By definition, a high-performing network is not necessarily overconfident, even when highly confident on all examples [21]. We conclude that these MNIST and Fashion-MNIST classifiers, with their average accuracies in the $\approx 99\%$ -range, are appropriately highly confident.

We similarly found BAYES-TREX is unable to sample ambivalent CLEVR examples. The data distribution for CLEVR is rendered from a uniform prior, not learned; as a consequence, the resulting examples are high quality. Unlike MNIST and Fashion-MNIST, the CLEVR network only achieves accuracies of $\approx 60\%$. The inability to sample ambiguous examples for the CLEVR network likely indicates overconfidence. Indeed, this problem has been previously observed in CLEVR-style settings [29].

H More Details on the In-Distribution Adversarial Examples Experiment

Many previous approaches reveal that neural network predictions are locally unstable; i.e. an input can be slightly perturbed to make the network produce confident, incorrect predictions [56, 35, 4, 8, 46, 19]. Such adversarial attacks produce examples that are out of distribution; indeed, a promising approach for detecting or defending against adversarial examples relies on these examples being out-of-distribution [23, 14]. While common adversarial examples highlight how a classifier can fail in malicious environments, BAYES-TREX exposes “in-distribution” adversarial examples [18] and highlights how realistic examples can also cause failures.

Specifically, if we train the classifier on the entire dataset, but we perform sampling with the data distribution induced by a generative model trained on data that are not from a particular class (Figure 3(b)), then the resulting samples can be considered adversarial examples for that class. Figure 14 shows in-distribution adversarial examples for CLEVR. For each target (e.g. “1 Cube”), we revoked the ability of the generator to produce objects belonging to the target class (e.g., no cubes).

For MNIST and Fashion-MNIST, we sample in-distribution adversarial examples for each class; these are presented in Fig. 15. Upon inspection, we observe several interesting phenomena. For MNIST, some of the very thin 8s are classified as 1, stretched 9s are classified as 4, and a particular style of 2s and 3s are classified as 7. For Fashion-MNIST, most samples are between semantically similar classes, such as T-shirt \leftrightarrow shirt, and sneaker \leftrightarrow ankle boot. However, we also see that chunky shoes are likely to be classified as bags. The sampling procedure failed to find any adversarial examples for trousers class.

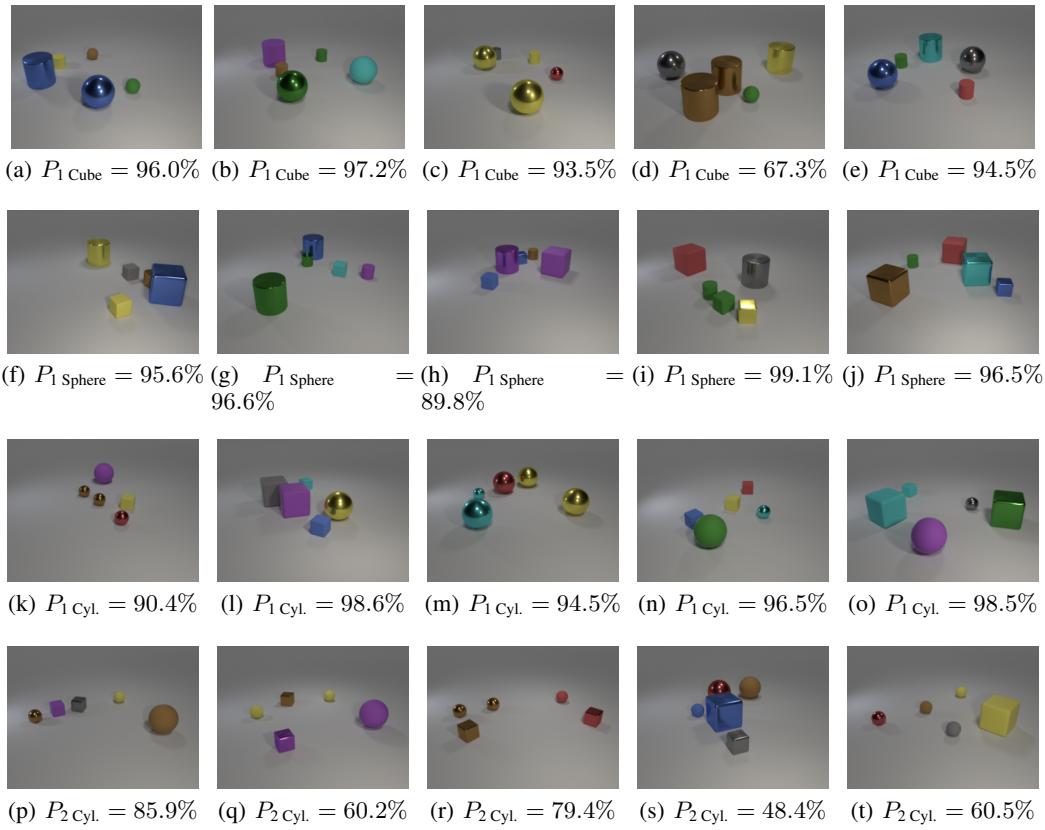


Figure 14: Sampled in-distribution adversarial examples and their associated prediction confidences. For each target constraint (e.g., “1 Cube”), the generator is unable to produce examples from the target class (e.g., cubes).

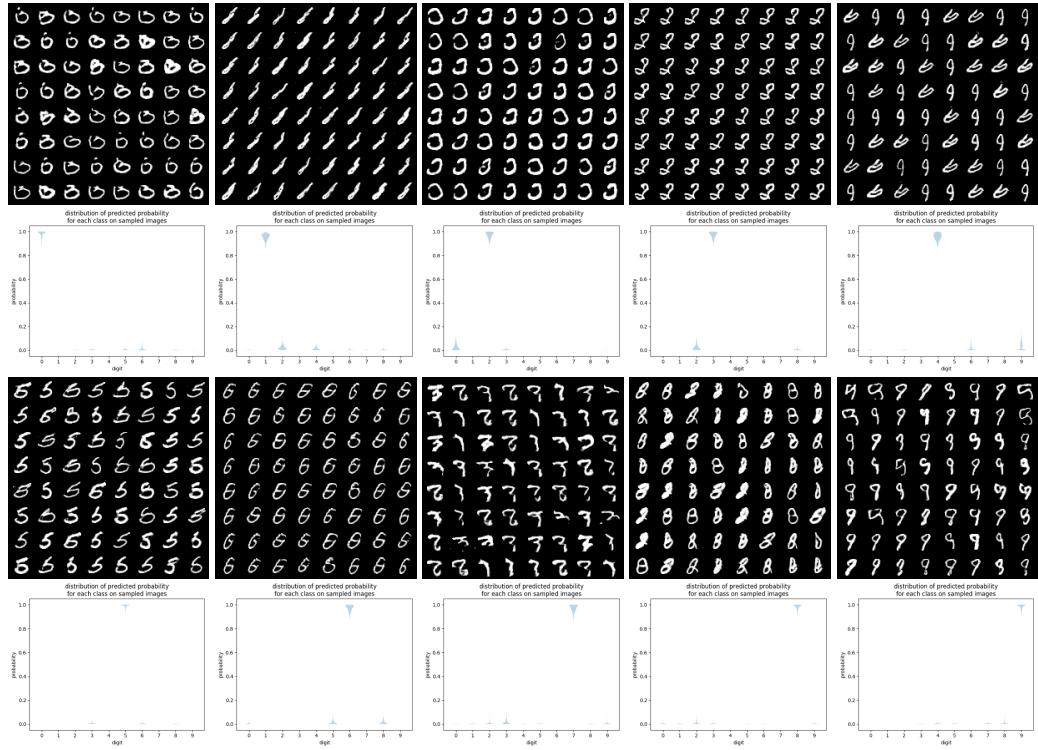


Figure 15: Samples and violin plots for in-distribution adversarial examples. Top two rows: digit 0-4; bottom two rows: digit 5-9.

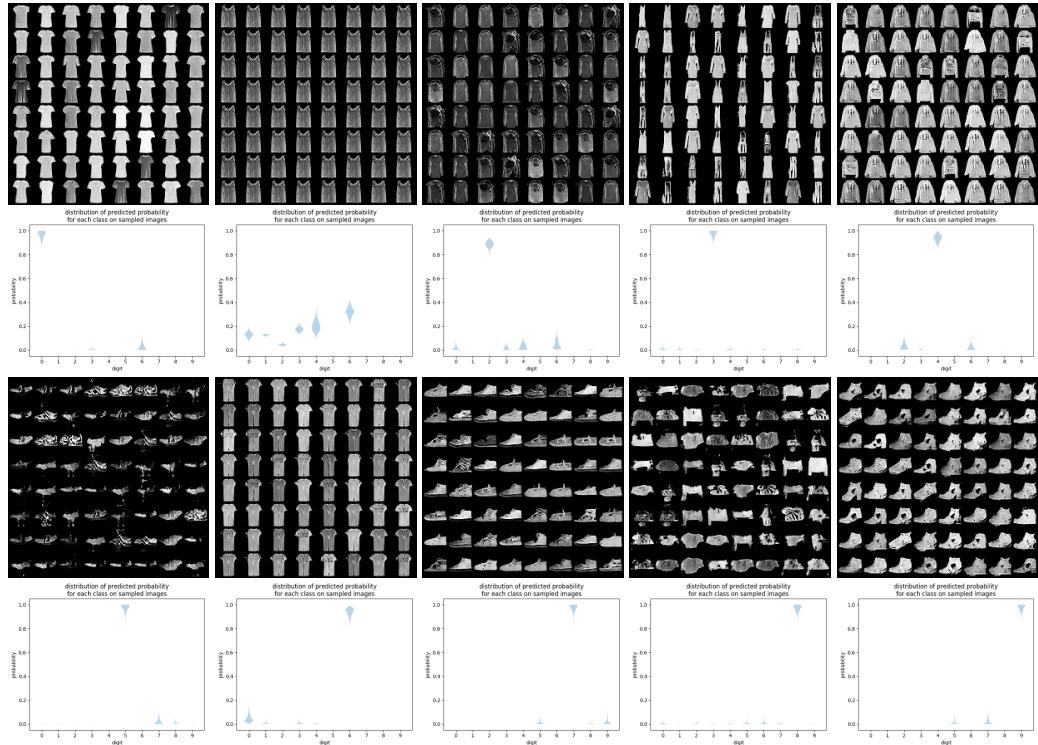


Figure 16: Samples and violin plots for in-distribution adversarial examples. Top row: T-shirt, trousers (sample failure), pullover, dress, coat. Bottom row: sandal, shirt, sneaker, bag, ankle boot.

I More Details on the OOD Extrapolation Experiment

While training and test distributions should ideally be the same, this assumption is often violated. One such violation occurs when a model encounters data from a new class. For example, if an autonomous driving system is trained on images of cars, pedestrians, and cyclists, it is desirable to know how such a model would behave when it sees a tandem bike on the road. While many methods have been proposed to detect out-of-distribution data [62, 11, 22, 38], BAYES-TREX facilitates model insight by finding examples from novel classes that are likely to cause incorrect detections, and can help further understand any inductive biases the model may have.

Figure 17 shows novel class extrapolation examples for CLEVR. Similar to the protocol for generating in-distribution adversarial examples, for each target (e.g. “1 Cube”), we revoked the ability of the generator to produce objects belonging to the target class (e.g., no cubes). We also introduced a new class to the generator, a cone. For the “5 Cubes” query, we increased the stopping criterion for Random-Walk Metropolis to 1500 examples; for all other experiments, we use the standard 500 samples criterion.

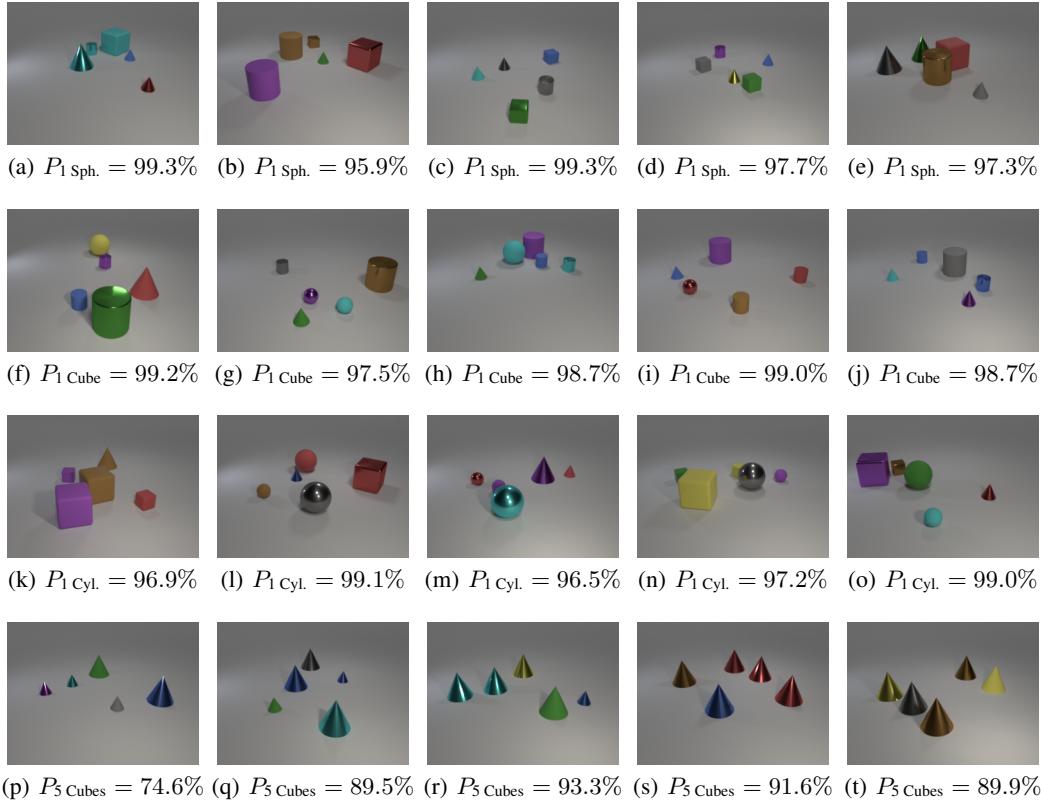


Figure 17: Sampled novel class extrapolation examples and their associated prediction confidences. Similar to in-distribution adversarial examples, for each target constraint (e.g., “1 Cube”), we deprive the generator of the ability to produce examples from the target class (e.g., cubes). For extrapolation, we equip the data generator with the ability to generate cones, a novel class not present in the training distribution. 17(n) is the only generated sample which does not include an object of the novel class.

To evaluate this extrapolation case on (Fashion-)MNIST, we consider the case of disjoint classifier and generator support as in Fig. 3(d). We split a dataset D into two disjoint parts, D_1 and D_2 by class label sets C_1 and C_2 respectively, with $C_1 \cap C_2 = \emptyset$. Then we train a classifier on D_1 and a GAN generator on D_2 . We then sample high-confidence examples for each class in C_1 from the GAN-induced distribution. For MNIST, we choose $C_1 = \{0, 1, 3, 6, 9\}$, and $C_2 = \{2, 4, 5, 7, 8\}$. For Fashion-MNIST, we choose $C_1 = \{\text{Pullover, Dress, Sandal, Shirt, Ankle boot}\}$, and $C_2 = \{\text{T-shirt, Trousers, Coat, Sneaker, Bag}\}$. Fig. 7(c, d) depicts samples from a GAN trained on D_2 for which the classifier trained on D_1 has high confidence for classes in C_1 .

Fig. 18 shows examples for novel-class extrapolation on MNIST. The classifier is trained on digit 0, 1, 3, 6 and 9, and tested on images generated by a GAN trained on digit 2, 4, 5, 7 and 8. Fig. 19 shows examples for novel-class extrapolation on Fashion-MNIST. The classifier is trained on pullover, dress, sandal, shirt and ankle boot, and tested on images generated by a GAN trained on T-shirt, trousers, coat, sneaker and bag.

For Fashion-MNIST, some results are intuitive. For example, the only shoe class in set C_2 is “sneaker,” which accounts for most of the “sandal” and “ankle boot” predictions. The same is true for “T-shirt” → “shirt.” Nevertheless, the classifier will also classify most of trousers as dresses, and bags of different styles will be classified as dresses, sandals, shirts, and ankle boots, despite visual dissimilarity. By comparison, for MNIST, it is hard to explain most of the samples. Several digits (e.g., 7) are inconsistently classified as different classes (i.e., split across 0, 1, and 9), likely because the digits are too visually distinct to allow for reasonable extrapolation.

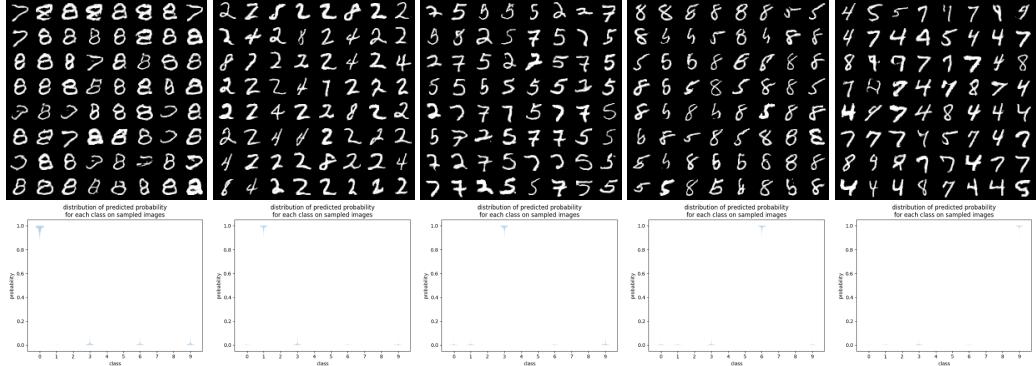


Figure 18: Samples and violin plots for novel-class extrapolation. Examples drawn from classifier classes (0, 1, 3, 6 and 9, in that order).

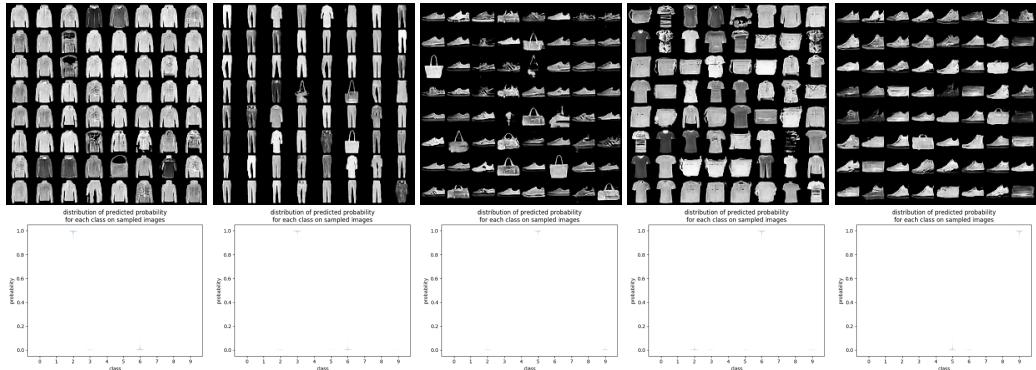


Figure 19: Samples and violin plots for novel-class extrapolation. Examples drawn from classifier classes (pullover, dress, sandal, shirt and ankle boot, in that order).

J More Details on the Domain Adaptation Experiment

In domain adaptation, we have a source domain \mathcal{X}_S and a target domain \mathcal{X}_T , with shared label set C and cross-domain consistent decision rule $p(y|x)$ [2, 40, 61]. The need for domain adaptation arises when the non-causal features vary across datasets. For example, for object recognition, lighting condition does not affect the object’s identity, and domain adaptation techniques can be applied to learn on bright images while being tested on dark ones. Domain adaptation is especially important in robotics, where algorithms trained in simulation often fail when using real sensor data in deployment [50, 59, 57].

To inspect a domain adaptation model, we sample high-confidence examples from the target domain for each class as predicted by the source-trained classifier. In our experiments, we train the adversarial discriminative domain adaptation (ADDA) model [60] using the code provided by the authors without modification¹, with Street View House Number (SVHN) [44] as the source domain and MNIST as the target domain. We also trained a baseline model with the same architecture on SVHN without the domain adaptation technique. Overall, we found that the baseline model achieves 61% target accuracy on MNIST, while the ADDA model achieves 71%.

To gain further insight into the domain adapted model, we sampled MNIST images with high prediction confidence for each class. Figure 20 shows additional samples and violin plots for the baseline model in the domain adaptation analysis. Top two rows are for digit 0-4, and bottom two rows are for digit 5-9. Figure 21 shows additional samples and violin plots for the adversarial discriminative domain adaptation (ADDA) model in the domain adaptation analysis. Top two rows are for digit 0-4, and bottom two rows are for digit 5-9.

Although the average confidence for the target class is over 99%, we can see both models have several wrong predictions. For example, some digit 6s are confidently classified as 4s by the baseline model, while some 0s are classified as 1s by the ADDA model. In addition, for each model and each target digit, we performed a human labeling on 10 images, and we check how many of those images are correctly labeled. The result is summarized in Table 8. Surprisingly, on high confidence samples, the baseline model is more accurate than the ADDA model. Although this does not fundamentally contradict the improved transfer performance for ADDA, this does highlight a concerning fact: ADDA seems to suffer from the overconfidence problem more severely, as a larger proportion of high confidence samples are incorrect. This fine-grained analysis suggests the need for more extensive investigation into and potential calibration of confidence for domain adaptation models.

Table 8: Per-digit and aggregate accuracy among high confidence MNIST samples for the baseline and domain adjusted (ADDA) models. While the ADDA model has overall higher accuracy than the baseline (0.71 vs. 0.61), on high-confidence samples it has lower accuracy than the baseline (0.72 vs. 0.80), suggesting an overconfident ADDA model.

Model \ Digit	0	1	2	3	4	5	6	7	8	9	All
Baseline	1.0	0.6	1.0	0.7	0.5	0.9	0.9	0.7	1.0	0.7	0.80
ADDA	0.9	0.0	0.8	0.9	0.2	1.0	0.8	1.0	1.0	0.6	0.72

¹<https://github.com/erictzeng/adda>. SVHN images are converted to gray-scale and 28×28 to match MNIST data.

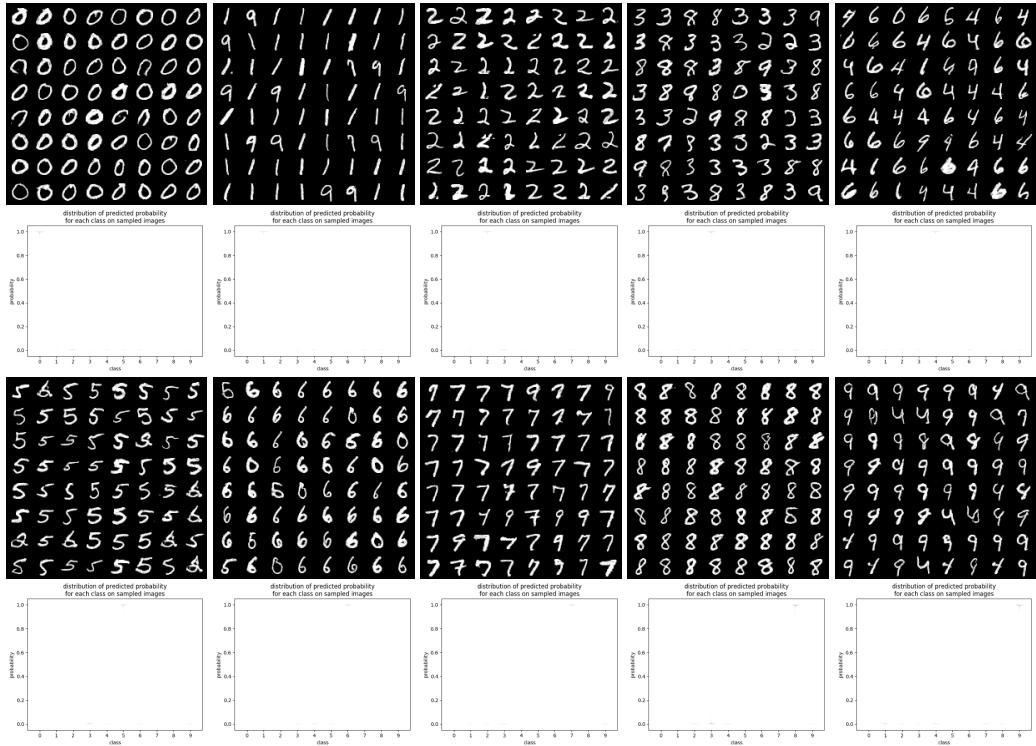


Figure 20: High confident MNIST samples generated for each class as predicted by the baseline model trained on SVHN dataset. Top row: digits 0-4. Bottom row: digits 5-9.

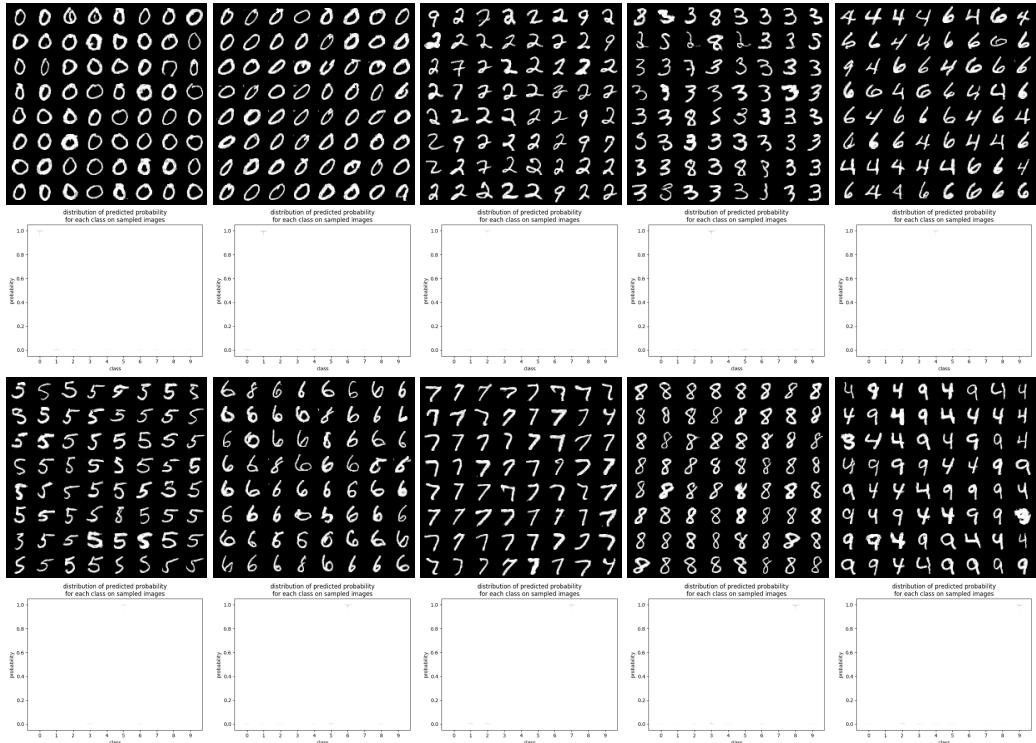


Figure 21: High confident MNIST samples generated for each class as predicted by the ADDA model trained on SVHN dataset. Top row: digits 0-4. Bottom row: digits 5-9.

K Case Study: Using BAYES-TREX with Saliency Maps

BAYES-TREX is a tool for mining interesting inputs to increase the breadth of evaluation beyond the data contained in the test set. By selecting appropriate distributions from the test and/or training sets, BAYES-TREX can expose additional interesting examples which communicate model behaviours. Local interpretability techniques, such as saliency maps, can in turn be applied to these examples to derive insights. We use SmoothGrad saliency maps for our evaluation [55], as this technique passes the saliency map sanity checks [1]. Fig. 24 shows several examples of CLEVR inputs and their accompanying SmoothGrad saliency maps.

K.1 In-distribution Adversarial Examples

Consider an example of an in-distribution adversarial example, uncovered by BAYES-TREX. Figure 22 shows one such example and its saliency map in which the classifier has high confidence contains 1 cube though the scene is composed of only cylinders and spheres. The scene is pre-processed for input to the CLEVR classifier. This pre-processing step includes resizing the image to 224×224 pixels. The accompanying saliency map suggests that the red metal cylinder is the cause of the classifier’s confusion.

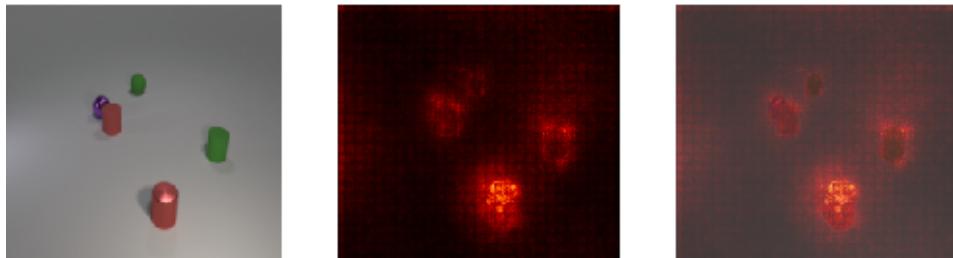


Figure 22: Left: the original image, preprocessed for classification by resizing and normalizing. The classifier is 93.5% confident this scene contains 1 cube, when in fact it is composed of 3 cylinders and 2 spheres. Middle: the SmoothGrad saliency map for this input. Right: the saliency map overlaid upon the original image. This saliency map most strongly highlights the red metal cylinder, indicating that this cylinder is likely the cause of the misclassification.

We re-render the scene, removing each object and re-classifying. As shown in Figure 23 and suggested by the original saliency map, we find the classifier mistook the red metal cylinder as a cube. Removing the red metal cylinder reduces the classifier’s confidence that the scene contained 1 Cube from 93.5% to 29.0%.

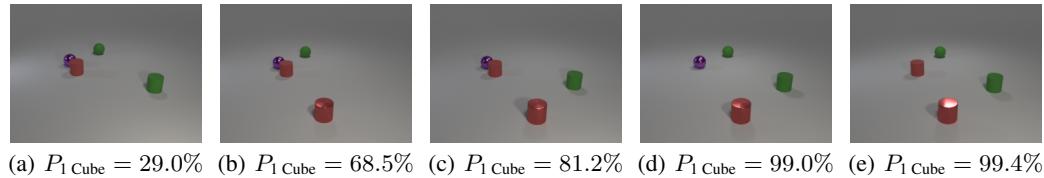
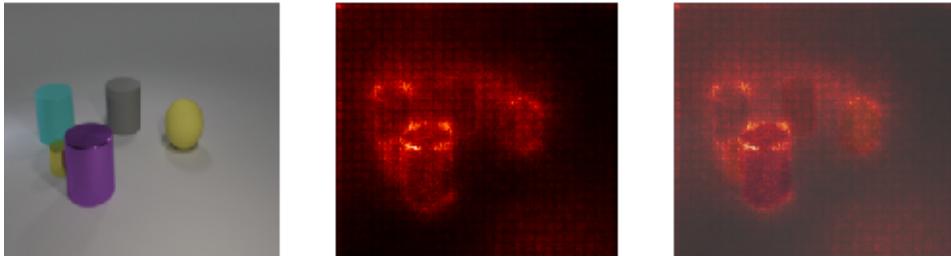
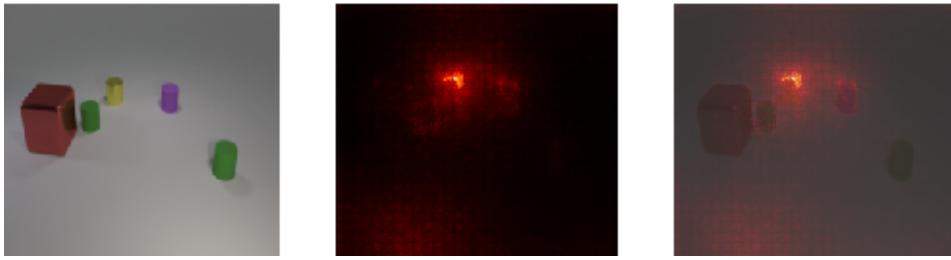


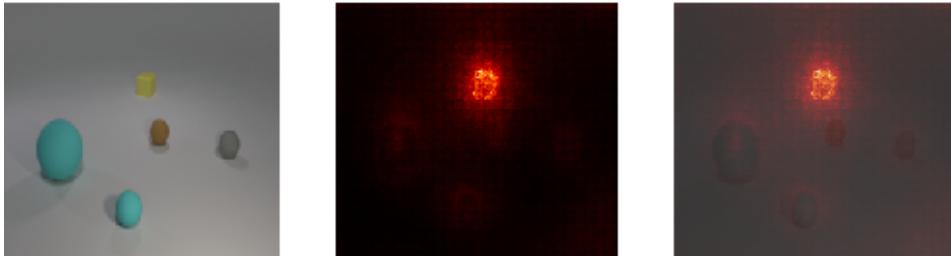
Figure 23: Re-rendering the scene, leaving out each object in turn. (a) The red metal cylinder is removed. (b) The green cylinder is removed. (c) The green sphere is removed. (d) The red rubber cylinder is removed. (e) The purple sphere is removed. As suggested by the saliency map, the removal of the red metal cylinder most prominently reduces the classification confidence, from 93.5% to 29.0%. The other scenes see a smaller confidence drop, and are still classified as containing one cube.



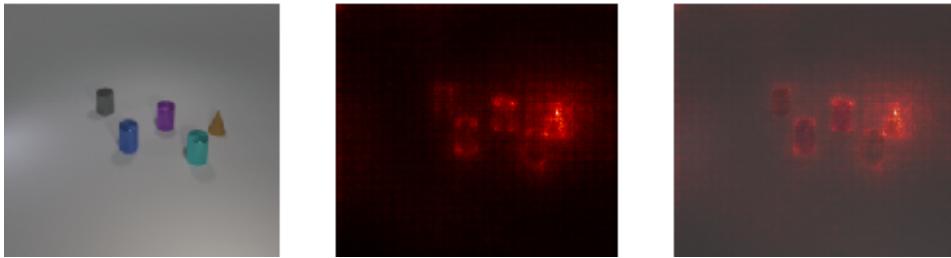
(a) Original image: $P_{1 \text{ Cube}} = 85.5\%$. Purple cylinder removed: $P_{1 \text{ Cube}} = 1.9\%$



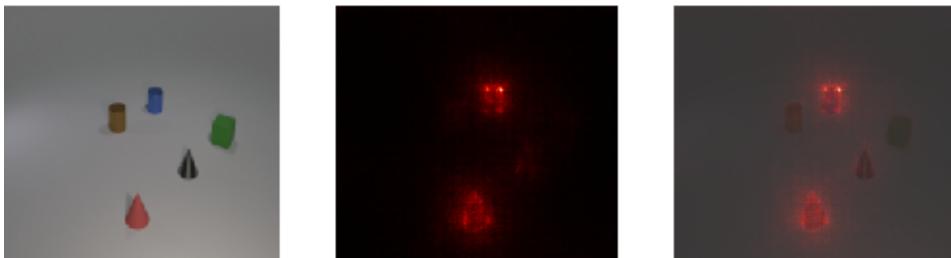
(b) Original image: $P_{1 \text{ Sphere}} = 97.9\%$. Yellow cylinder removed: $P_{1 \text{ Sphere}} = 5.2\%$



(c) Original image: $P_{1 \text{ Cylinder}} = 85.4\%$. Red sphere removed: $P_{1 \text{ Cylinder}} = 0.9\%$



(d) Original image: $P_{1 \text{ Cube}} = 99.7\%$. Cone removed: $P_{1 \text{ Cube}} = 0.4\%$



(e) Original image: $P_{1 \text{ Sphere}} = 98.0\%$. Gray cone removed: $P_{1 \text{ Sphere}} = 0.3\%$

Figure 24: Images sampled with BAYES-TREX and their saliency maps. 24(a)-24(c) are in-distribution adversarial examples; 24(d)-24(e) are novel class extrapolation examples. In 24(e), the saliency map primarily highlights two objects: the red cone and the blue cylinder. Removing either of these objects does not result in a change of prediction. Instead, the misclassification of 1 Sphere is due to the marginally-highlighted gray cone.