Robust outlier detection by de-biasing VAE likelihoods

Kushal Chauhan [1], Barath Mohan U [2], Pradeep Shenoy [1], Manish Gupta [1], Devarajan Sridharan [2*]

[1] Google Research

{kushalchauhan, shenoypradeep, manishgupt}@google.com

[2] Center for Neuroscience, and Computer Science and Automation, Indian Institute of Science

{barathu, sridhar}@iisc.ac.in

Abstract

Deep networks often make confident, yet, incorrect, predictions when tested with outlier data that is far removed from their training distributions. Likelihoods computed by deep generative models (DGMs) are a candidate metric for outlier detection with unlabeled data. Yet, previous studies have shown that DGM likelihoods are unreliable and can be easily biased by simple transformations to input data. Here, we examine outlier detection with variational autoencoders (VAEs), among the simplest of DGMs. We propose novel analytical and algorithmic approaches to ameliorate key biases with VAE likelihoods. Our bias corrections are sample-specific, computationally inexpensive, and readily computed for various decoder visible distributions. Next, we show that a well-known image preprocessing technique – contrast stretching – extends the effectiveness of bias correction to further improve outlier detection. Our approach achieves state-of-the-art accuracies with nine grayscale and natural image datasets, and demonstrates significant advantages – both with speed and performance – over four recent, competing approaches. In summary, lightweight remedies suffice to achieve robust outlier detection with VAEs.[1]

1. Introduction

Deep neural networks are increasingly deployed in real-world computer vision applications. A key issue with such deployments is over-confident predictions: the tendency of these networks to make confident, yet incorrect, predictions when tested with images whose statistics are far removed from the training data distribution [21]. Developing robust methods for outlier detection is, therefore, an important challenge with critical real-world implications.

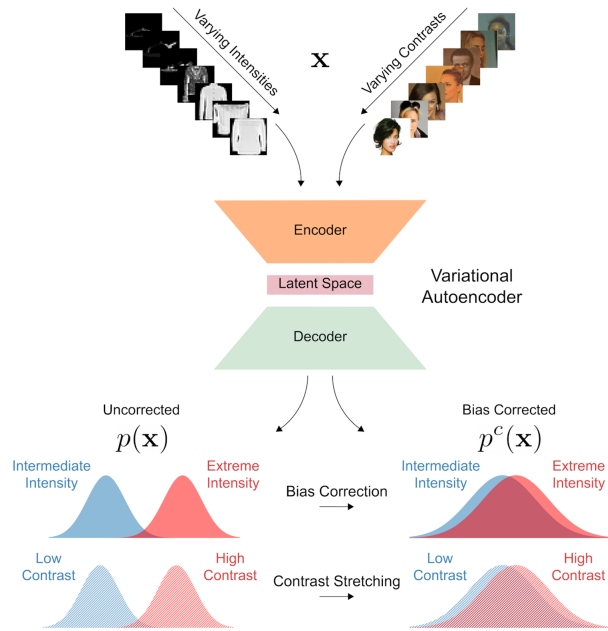A popular class of approaches for outlier detection, espe-

---

Figure 1. **Schematic of the proposed approach.** Variational autoencoders (VAEs) are routinely employed for unsupervised outlier detection with real-world images. Yet, variations in low-level image features, like overall intensity or contrast, can readily bias VAE likelihoods ($p(\mathbf{x})$). We analyze these biases, and propose lightweight analytical and algorithmic remedies (Bias Correction and Contrast Stretching) for de-biasing VAE likelihoods ($p^c(\mathbf{x})$), to achieve robust outlier detection.

cially with label-free data, involves computing sample likelihoods with deep generative models like variational autoencoders (VAEs [8]) or flow-based models (e.g. Glow [7]). Here, we explore outlier detection with VAEs, arguably among the simplest of deep generative models. In VAEs, both the generative model $p_\theta(\mathbf{x}|\mathbf{z})$ and the inference model $q_\phi(\mathbf{z}|\mathbf{x})$ are parameterized by deep neural networks with parameters θ (decoder) and ϕ (encoder) respectively; the stochastic latent representation $\mathbf{z}$ is typically of considerably lower dimensionality than the input data $\mathbf{x}$. VAEs dif-

fer from other kinds of deep generative models, like flow models, in that rather than directly optimizing the marginal likelihood $p(\mathbf{x})$, they seek to optimize the evidence lower bound (ELBO) as a proxy for maximizing $p(\mathbf{x})$.

Several previous studies have shown that likelihoods computed by deep generative models, including VAEs, are unreliable for outlier detection [3, 16, 19, 25]. For example, these likelihoods are readily biased by differences in low-level image statistics, like the number of zeros in a sample [19] or the variance in the distribution of pixel intensities [16]. A few solutions have been proposed to overcome these challenges. However, these suffer from computational bottlenecks [19, 25] or do not work well in low-sample regimes [15] (see Section 2, Related Work).

In this context, we propose and test efficient remedies that achieve or approach state-of-the-art for outlier detection. Our key contributions are as follows:

- We present a careful analysis of bias in VAE likelihoods and propose analytical and algorithmic approaches to correct this bias. These corrections can be computed inexpensively, *post hoc*, during evaluation time.
- We show that a standard image pre-processing step (contrast normalization) enables competitive outlier detection.
- We present a comprehensive evaluation of outlier detection with VAEs, with nine datasets (four grayscale and five natural image datasets), in all.
- We demonstrate key advantages in terms of both speed and accuracy over multiple competing, state-of-the-art approaches [3, 19, 20, 25].

In sum, lightweight and computationally efficient remedies suffice to achieve robust outlier detection with VAEs.

2. The challenge of outlier detection with VAE likelihoods

We briefly recapitulate a well-known challenge of outlier detection with VAE likelihoods, by taking a fresh look at two previously reported sources of bias [16, 19].

Bias arising from pixel intensity. As a first example, we train a VAE on grayscale FMNIST images [24] and compute the likelihoods for in-distribution (ID) FMNIST and out-of-distribution (OOD) MNIST [12] test samples. We employ the continuous Bernoulli visible distribution [14] for the VAE decoder (model details, Appendix A. We replicate the well-known issue with VAE likelihoods: FMNIST VAE likelihoods are higher for OOD (MNIST) samples as compared to ID (FMNIST) samples (Fig. 2a). The highest likelihoods are assigned to FMNIST samples with a large number of black pixels (Fig. 2b, top row), whereas the lowest likelihoods are assigned to samples with many intermediate (gray) pixel values (Fig. 2b, bottom row), consistent with previous reports [19]. On simulated images with different (constant) pixel intensities, we find a U-shape trend in likelihood bias (Fig. 3a, gray line).

Bias arising from channel variance or image contrast. Next, we train a VAE on the CelebA dataset [13], and compute ID (CelebA) and OOD (GTSRB [22]) likelihoods. Again, the VAE assigns high likelihoods to OOD samples (Fig. 2f). Faces with dark backgrounds and high contrast
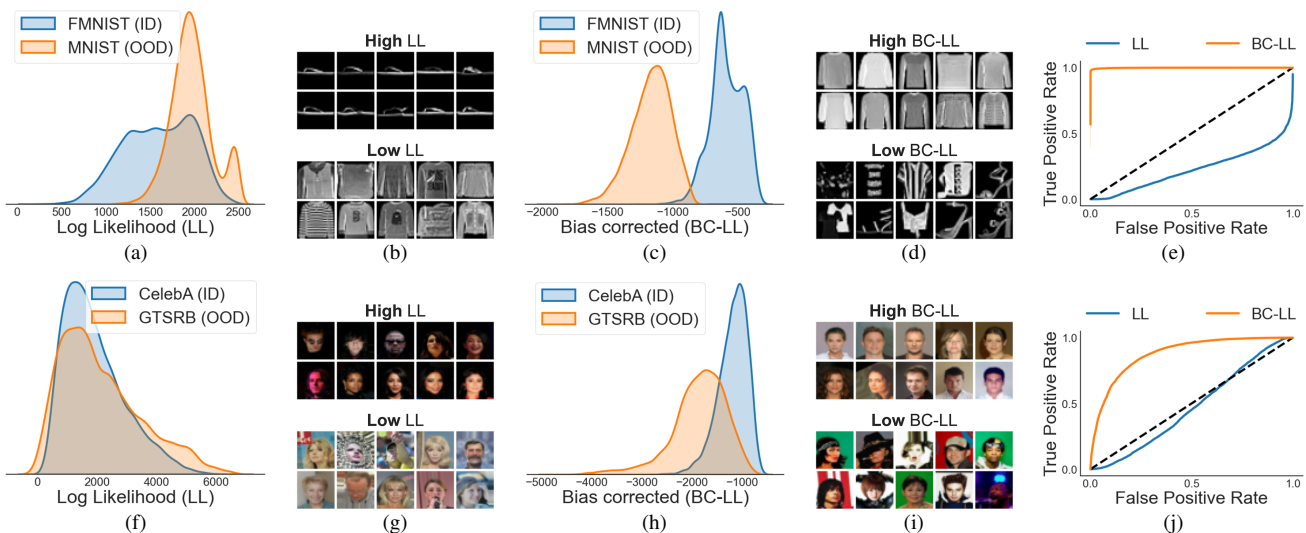

Figure 2. **Bias correction improves outlier detection.** **(a)** Kernel density estimate of Fashion-MNIST (FMNIST) VAE log likelihoods. Out-of-distribution (OOD) MNIST samples get higher log likelihoods as compared to in-distribution (ID) FMNIST samples. **(b)** Top and bottom rows: FMNIST images assigned highest and lowest log-likelihoods, respectively, by the FMNIST VAE. **(c)** Same as in panel (a), but with bias correction **(d)** Same as in panel (b), but with bias correction. **(e)** AUROC curves for outlier detection with the FMNIST VAE likelihoods using MNIST (OOD) test samples without (blue) bias correction and with bias correction (orange) **(f-j)** Same as in panels (a-e), but for the CelebA VAE with CelebA (ID) and GTSRB (OOD) samples.

between the face and background are assigned the highest likelihoods (Fig. 2g, top row), and vice versa for low contrast faces (Fig. 2g, bottom row). With simulated images, we observe that VAE likelihoods are strongly biased by contrast (Fig. 3b, gray line).

In this context, we develop efficient remedies for outlier detection with VAE likelihoods. First, we correct for pixel intensity bias using both analytical and empirical approaches. Second, we correct for the image contrast bias using a standard image pre-processing step (contrast stretching). Finally, we evaluate our remedies with multiple (nine) grayscale and natural image datasets, to show state-of-the-art outlier detection.

Related work

Our bias correction is most closely related to the work of Serra et al. (2019) [20] who proposed a correction for Glow and PixelCNN++ model likelihoods based on "input complexity" (IC). Their out-of-distribution score is computed by subtracting a sample-specific complexity estimate $L(\mathbf{x})$ from the negative log-likelihood. Yet, IC depends on an *ad hoc* choice of a compression algorithm (e.g. PNG, JPEG2000, FLIF), whereas our correction is analytically derived from the VAE decoder visible distribution. We also demonstrate superior performance over IC for multiple grayscale and natural image datasets.

Ren et al. (2019) [19] originally highlighted the problem of bias in deep generative model likelihoods, for samples with many zero-valued pixels. They proposed correcting for this bias by training a second generative model with noise-corrupted samples to capture background statistics; the "likelihood ratio" between the original and noisy VAEs provided a sensitive readout of foreground object statistics. Yet, this approach does not appear to work well with VAEs (see also [25]). Moreover, our bias correction obviates the need for training multiple, duplicate models. Similarly, Nalisnick et al. (2019) [16] originally identified the problem of bias in likelihoods arising from sample variance [15]. However, their solution – a "typicality" test – works best with batches of samples, and performs relatively poorly with single outlier samples.

Recent work by Yong et al. (2020) [27] proposed employing bias-free Gaussian likelihoods and their variances for outlier detection. Yet, Gaussian visible distributions are not theoretically appropriate to model the finite range of pixel values (0-255) encountered in images. We develop bias corrections for Bernoulli, continuous Bernoulli, truncated Gaussian and categorical visible distributions, which are appropriate for modeling pixel values in images. Moreover, Yong et al.'s approach does not work well with natural image datasets (their Appendix C). Our results show robust outlier detection, even with natural image datasets.

Xiao et al. (2020) [25] proposed a "likelihood re-

gret" metric that involves quantifying the improvement in marginal likelihood by retraining the encoder network to obtain optimal likelihood for each sample. Such sample-specific optimizations are computationally expensive, for example, when millions of samples need to be evaluated on-the-fly. In contrast, our proposed metrics are readily computed with a single forward pass through pre-trained VAEs, leading to a 50-100x speedup in evaluation times over likelihood regret.

3. De-biasing VAE likelihoods

To improve outlier detection with VAE likelihoods, we develop remedies for correcting for the two sources of bias discussed in the previous section.

3.1. Analytical correction for intensity bias

We develop an analytically-derived correction for VAE likelihoods, based on the recently developed continuous Bernoulli visible distribution for the decoder [14]. The procedure for computing the correction for a conventional Bernoulli visible distribution is presented in Appendix B.

The VAE marginal likelihood can be written as:

$$\log p_\theta(\mathbf{x}) = \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})/q_\phi(\mathbf{z}|\mathbf{x})]$$

We examine the negative reconstruction error term $p_\theta(\mathbf{x}|\mathbf{z})$, assuming perfect reconstruction of the input samples by the VAE. We denote this as $p_{cB}(\mathbf{x}; \boldsymbol{\lambda}^*)$ where p_{cB} denotes the continuous Bernoulli pdf, and $\boldsymbol{\lambda}^*$ are optimal parameters that correspond to perfect reconstruction ($\hat{\mathbf{x}} = \mathbf{x}$). We plot $\log p_{cB}(\mathbf{x}; \boldsymbol{\lambda}^*)$ for simulated images in Figure 3a (dashed green). $\log p_{cB}(\mathbf{x}; \boldsymbol{\lambda}^*)$ exhibits a bias that is nearly identical with the marginal likelihood (Fig. 3a, gray). Thus, even if two input samples are perfectly reconstructed by the VAE, these will be assigned different likelihoods, depending on the average pixel intensity in each sample; a bias that is largely driven by the reconstruction error term.

We eliminate this bias in the reconstruction error by dividing by the error for perfect reconstruction. For the continuous Bernoulli visible distribution, the negative reconstruction error is given by:

$$\begin{aligned} \log p_\theta(\mathbf{x}|\mathbf{z}) &= \log p_{cB}(\mathbf{x}; \boldsymbol{\lambda}_\theta(\mathbf{z})) \\ &= \sum_{i=1}^{D} \log C(\lambda_i) + x_i \log \lambda_i + (1 - x_i)\log(1 - \lambda_i) \end{aligned} \quad (1)$$

Note that the continuous Bernoulli decoder outputs the shape parameter (λ_i) for the ith pixel. The decoded pixel value itself is given by:

$$\begin{aligned} \hat{x}_i &= \frac{\lambda_i}{2\lambda_i - 1} + \frac{1}{2\tanh^{-1}(1 - 2\lambda_i)} \quad &\text{if} \quad \lambda_i \neq \frac{1}{2} \\ &= \frac{1}{2} \quad &\text{if} \quad \lambda_i = \frac{1}{2} \end{aligned}$$
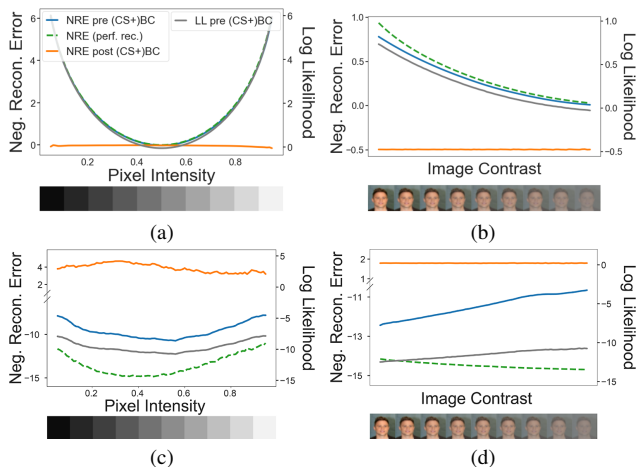
Figure 3. **Correcting for biases arising from pixel intensities and image contrasts.** **(a)** Negative reconstruction error for simulated images with uniform pixel intensities, ranging from full-black to full-white before (blue) and after (orange) analytical bias correction (BC) (y-axes units are kilonats). **(b)** Negative reconstruction error for simulated images with varying contrasts, from highest to lowest values before (blue) and after (orange) contrast stretching (CS) and analytical bias correction. (Both panels) Dashed green line: negative reconstruction error for perfect reconstruction with the continuous Bernoulli visible distribution. Gray line: log-likelihood before bias correction. **(c-d)** Same as in (a) and (b) but for VAE trained with categorical visible distribution, and algorithmic bias correction.

For perfect reconstruction we set $\hat{x}_i = x_i$. To find the optimal λ_i^* corresponding to perfect reconstruction, we used SciPy's implementation of Nelder-Mead simplex algorithm to iteratively maximize $\log p_{cB}(x_i; \lambda_i)$; the correction is then calculated by setting $\lambda_i = \lambda_i^*$ in equation (1) above.

Thus, the "bias-corrected" marginal likelihood (BC) evaluates to:

$$\log p_\theta^c(\mathbf{x}) = \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\frac{p_\theta(\mathbf{x}|\mathbf{z})}{p_{cB}(\mathbf{x}; \boldsymbol{\lambda}^*)} \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \quad (2)$$

$$= \log p_\theta(\mathbf{x}) - \log p_{cB}(\mathbf{x}; \boldsymbol{\lambda}^*) \quad (3)$$

Following this analytical bias correction, the bias in the negative reconstruction error is eliminated (Fig. 3a, orange). We note that this correction can be computed during evaluation time and does not require retraining the VAE.

3.2. Algorithmic correction for intensity bias

Next, we demonstrate an algorithmic approach for correcting biases with VAE likelihoods. Such an approach is relevant for decoder visible distributions for which the analytical correction is not tractable, or for which biases must be evaluated empirically. We illustrate this latter scenario with the "categorical" visible distribution" – a popular choice of visible distribution for various generative models,

including VAEs [19,25] and autoregressive models like PixelCNN [23].

In theory, the categorical distribution does not suffer from an analytical bias. As a result, perfectly reconstructed images should not suffer from any bias in the reconstruction error term, $p_\theta(\mathbf{x}|\mathbf{z})$. Yet, in practice, we find such a bias arises in a VAE trained with the categorical visible distribution for the decoder.

To illustrate this, we compute the empirical bias in the log likelihood for a CelebA VAE for images with various uniform pixel intensities (0-255) (Fig. 3c, gray). We observe a U-shaped profile, similar to that observed for VAEs trained with the continuous Bernoulli visible distribution (Fig. 3a). This bias can be explained as follows: For target pixel values close to full black or full white, the VAE decoder concentrates probability mass (PMF) at the target pixel value. On the other hand, for target values close to the middle of the gray range, the PMF is more dispersed around the target pixel value, resulting in lower PMF at the target pixel itself (Appendix F.1, for details).

We correct for this empirical bias with the following algorithmic approach. We compute the average categorical distribution output for every target pixel value (0-255) employing all of the training (inlier) samples for that VAE. Then, for each test sample, we compute the correction term for the negative reconstruction error as its average value across pixels, under the categorical visible distribution. The procedure is described in detail in Algorithm 1.

Again, following this algorithmic bias correction, the empirical bias with the categorical distribution is eliminated (Fig. 3c, orange). Such an algorithmic correction generalizes to other types of visible distributions as well: results for VAEs trained with the truncated Gaussian visible distribution are presented in Appendix F.2.

3.3. Correction for contrast bias: Normalization

Variation in image contrasts produces systematic biases in the likelihood and negative reconstruction error, both with the continuous Bernoulli visible distribution (Fig. 3b, blue), as well as with the categorical visible distribution (Fig. 3d, blue). This is particularly surprising because the categorical distribution is, in theory, bias-free. Nonetheless this bias arises, in practice, due to practical limitations with achieving perfect reconstructions with VAEs. Specifically, the VAE achieves systematically worse reconstructions – as measured by the categorical negative reconstruction error – for high-contrast, as compared to low-contrast, images (Fig. 3d, blue line). As a consequence, the contrast-related bias in the categorical distribution is in a direction opposite to that observed with the continuous Bernoulli distribution (compare Fig. 3d with Fig. 3b).

For eliminating this additional source of bias arising from image contrasts, we propose a standard image pre-

Algorithm 1: Algorithmic Bias correction

Data: Training Set $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ... \mathbf{x_n}\}$ with $\mathbf{x_p}$ of shape $32 \times 32 \times nc$ (no. of channels), Encoder Parameters ϕ, and Decoder Parameters θ

Result: Log Correction Factor $\mathbf{C} : (v, k) \rightarrow Float$ for $v = 0, 1, \ldots 255$ and $k = 1, 2, \ldots nc$

Init: Map $\mathbf{A} : (v, k) \rightarrow EmptyList$ for $v = 0, 1, \ldots 255$ and $k = 1, 2, \ldots nc$

for $\mathbf{x_p} \in \mathbf{X}$ **do**
 Init: Map $\mathbf{B} : (v, k) \rightarrow EmptyList$ for $v = 0, 1, \ldots 255$ and $k = 1, 2, \ldots nc$
 $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x_p})$
 for $i \leftarrow 1$ **to** 32, $j \leftarrow 1$ **to** 32, $k \leftarrow 1$ **to** nc **do**
  Append $p_\theta^{ijk}(x_p^{ijk}|\bar{\mathbf{z}})$ to $\mathbf{B}(x_p^{ijk}, k)$
 end
 for $v \leftarrow 0$ **to** 255, $k \leftarrow 1$ **to** nc **do**
  Append $\texttt{Mean}(\mathbf{B}(v, k))$ to $\mathbf{A}(v, k)$
 end
end

Init: Map $\mathbf{C} : (v, k) \rightarrow 0$ for $v = 0, 1, \ldots 255$ and $k = 1, 2, \ldots nc$

for $v \leftarrow 0$ **to** 255, $k \leftarrow 1$ **to** nc **do**
 $\mathbf{C}(v, k) \leftarrow \texttt{Log}(\texttt{Mean}(\mathbf{A}(v, k)))$
end

---

processing step – "contrast stretching" – which is known to improve the robustness of deep classifier models [5]. For our case, each image sample, both from the training and testing datasets, is contrast normalized with the following transformation: $x_i = \min(\max(0, [x_i - a]/r), 1)$, where x_i refers to the ith pixel of image $\mathbf{x}$, $r = P_{95}(\mathbf{x}) - P_5(\mathbf{x})$, $a = P_5(\mathbf{x})$, P_j refers to the jth percentile and $\mathbf{x}$ refers to the vectorized input sample tensor. Following contrast stretching, we observe more homogeneous distributions of per-channel variance across datasets (see Appendix E.4).

Contrast stretching and bias correction ameliorate this bias for both types of visible distributions (Fig. 3b, orange, Fig. 3d, orange).

4. Experiments

We trained and tested VAEs with multiple grayscale and natural image datasets. These included four grayscale image datasets – MNIST, Fashion-MNIST, EMNIST (handwritten English letters [4]) and Sign Language-MNIST (hand gesture thumbnails [1]) – and tested them against four outlier datasets each. These outlier datasets comprised each of the other three datasets (the ones that the VAE was not trained on) and a grayscale, uniform noise dataset. Similarly, we trained VAEs with five natural im-

age datasets – SVHN (Street View House Numbers [17]), CelebA, ComprehensiveCars (traffic surveillance camera views of cars [26]), GTSRB (German Traffic Sign Recognition Benchmark) and CIFAR-10 [10] – and tested them against five outlier datasets each. Again, the outlier datasets comprised each of the other four datasets, and a colored, uniform noise dataset. Data sources and pre-processing details are provided in Appendix C. Each VAE was trained 6 times, with 3 different random initializations of the network weights across 2 train-validation splits; we report average performance measures across all 6 runs (e.g. Figs. 4, 5).

4.1. De-biased likelihoods improve outlier detection

We compared the outlier detection performance of the vanilla (uncorrected) VAE likelihood (LL) against the bias corrected likelihood (BC-LL). Both train and test images were contrast stretched prior to computing BC-LL scores. The results are shown as a grid (4×5 or 5×6) of area under the ROC curve (AUROC) values, separately for the grayscale (1 channel) and natural image (3 channel) datasets. Corresponding results for area under the precision-recall curve (AUPRC) and false-positive rate at 80% true-positive rate (FPR@80%TPR) are provided in Appendices E.2-E.3, and details regarding each measure are provided in Appendix D.

We evaluated the effectiveness of the analytical bias correction, first, with VAEs trained with the continuous Bernoulli visible distribution.

First, we evaluated our remedies for outlier detection with VAEs trained on grayscale image datasets. In nearly every case, we found that bias-corrected likelihoods (BC-LL, Fig. 4, unfilled orange squares) outperformed uncorrected likelihoods (LL, Fig. 4, unfilled blue circles); AUROC values for outlier detection approached ceiling levels, in many cases. For example, bias correction resolved the problematic case of the Fashion-MNIST VAE such that in-distribution Fashion-MNIST samples were assigned higher bias-corrected likelihoods than out-of-distribution MNIST samples (Fig. 2c), resulting in a perfect AUROC (Fig. 2e). Moreover, samples that were assigned the highest and lowest bias-corrected likelihoods were visually more typical and atypical, respectively (Fig. 2d), as compared to those based on uncorrected likelihoods (Fig. 2b). Across all outlier datasets tested with the Fashion-MNIST VAE, AUROC was typically at or near ceiling with the bias-corrected likelihoods (Fig. 4, second column). On average, we found accuracy improvements of up to $\sim$70% following bias correction, across all grayscale VAEs (Fig. 4, last row).

Second, we tested these remedies with VAEs trained on natural image datasets. Again, in nearly every case the BC-LL score (Fig. 5, unfilled orange squares) outperformed uncorrected likelihoods (Fig. 5, unfilled blue circles). For example, for the CelebA VAE, AUROC values ranged be-
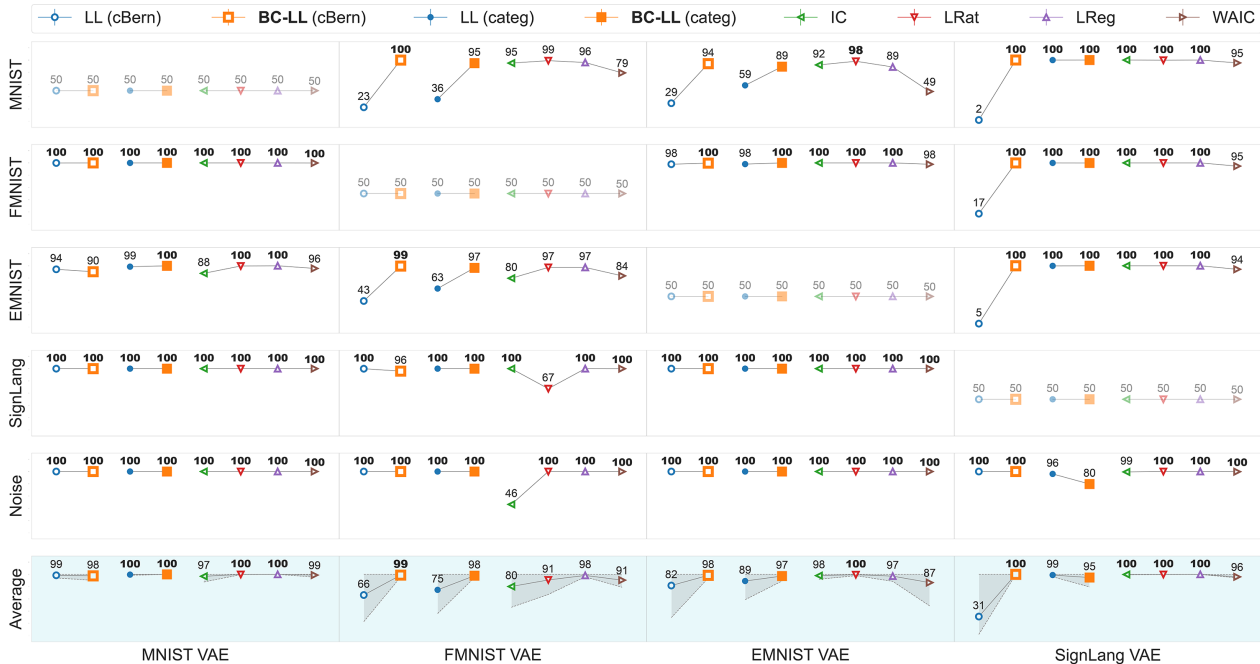
Figure 4. **Outlier detection with bias-corrected likelihoods, and comparison with competing approaches: Grayscale datasets.** Outlier detection AUROC values for VAEs trained with grayscale image datasets and tested with other grayscale datasets (or noise). Each column represents a training dataset, and each row represents a test dataset. Last row: average AUROC across all test datasets for each VAE. Unfilled blue and orange symbols: AUROC with uncorrected log likelihoods (LL), and bias-corrected log likelihoods (BC-LL) respectively for VAE trained with continuous Bernoulli visible distribution. Filled blue and orange symbols: Same as the unfilled symbols but for VAE trained categorical visible distribution. Green, red, purple, and brown symbols show AUROC scores of competing outlier detection approaches: Input Complexity (IC), Likelihood Ratio (LRat), Likelihood Regret (LReg), and Watanabe Akaike Information Criterion (WAIC), respectively. Higher values indicate better outlier detection performance. Numbers in bold indicate highest AUROC values across all scores for each train-test combination. Gray shading in the last row: range of AUROC values for each score across all test datasets.

tween 47-79 with the uncorrected likelihoods (across all out-of-distribution datasets, except Noise), whereas these values improved to 85-88 with bias correction (Fig. 5, second column, Fig. 2h, and Fig. 2j). As with the Fashion-MNIST dataset, CelebA samples assigned the highest bias-corrected likelihoods were visually more typical; containing face images on typical plain backgrounds, whereas the lowest likelihood samples comprised of face images with extraneous objects, such headgear, unconventional attire or unusual backgrounds (Fig. 2i). On average, we observed accuracy improvements of ∼10-40% following bias correction, across all natural image VAEs (Fig. 5, last row).

Ablation experiments revealed the relative contributions of contrast stretching and bias correction; the results are presented in Appendix E.4. First, contrast stretching was essential to achieving high accuracies, particularly with the natural image datasets. Moreover, contrast stretching, even at test time, sufficed to achieve robust outlier detection: With a VAE trained on the original images (without contrast stretching), outlier detection accuracies improved if contrast stretching was applied at test time alone. Lastly, contrast normalization with techniques other than contrast stretching (e.g. histogram equalization) also yielded comparable

outlier detection accuracies (see Appendix E.4).

Contrast stretching and bias correction generally improved outlier detection even with milder perturbations to the data, such as with adding various kinds of noise (Gaussian, impulse etc; see Appendix E.5). On the other hand, training the VAE by introducing intensity and contrast variations in the training samples was not as effective as contrast stretching and bias correction (Appendix E.6). Similarly, an alternative approach for normalizing per channel variance using image whitening (Zero-phase Component Analysis or ZCA) was also comparatively unsuccessful, especially with natural image datasets (Appendix E.7).

Finally, we replicated these results with the algorithmic bias correction with VAEs trained with the categorical visible distribution. Again, in nearly every case we observed significant improvements in outlier detection performance following bias correction (Fig. 4 and Fig. 5, filled blue circles/LL versus filled orange squares/BC-LL).

4.2. Comparison with competing approaches

We compared accuracies with the bias-corrected likelihood (BC-LL) scores against four competing, state-of-the-art approaches: i) input complexity (IC), ii) likelihood
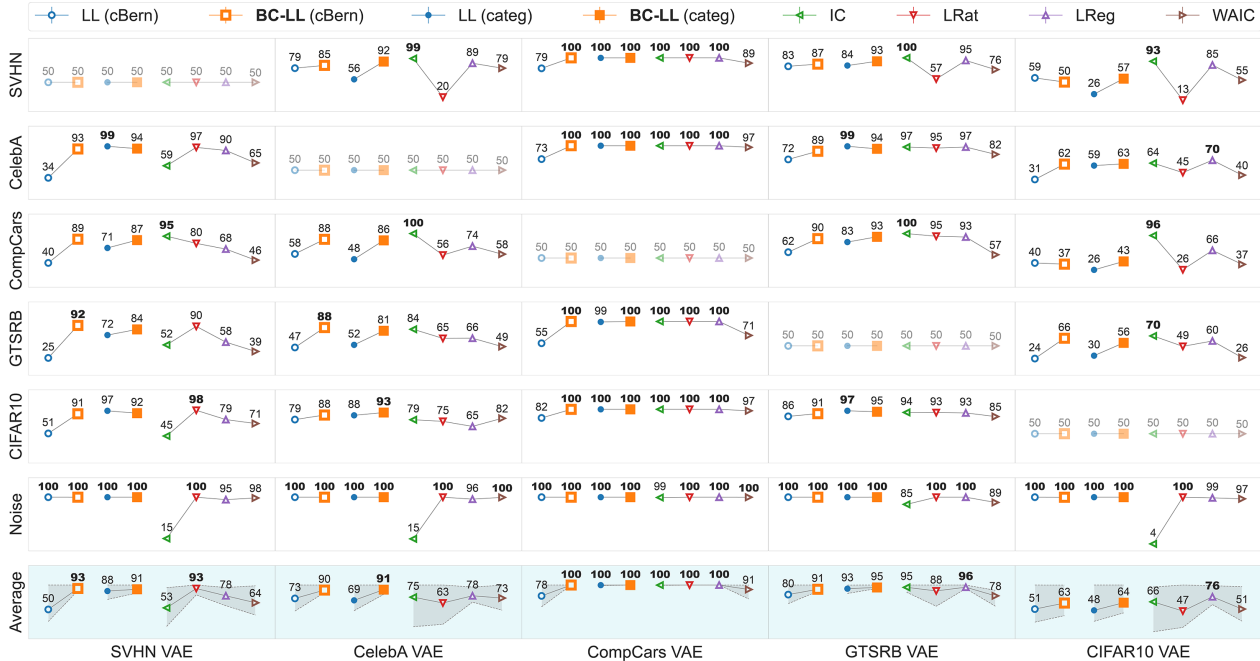
Figure 5. **Outlier detection with bias-corrected likelihoods, and comparison with competing approaches: Natural image datasets.**
Same as in Figure 4 but for VAEs trained with natural image datasets and tested with other natural image datasets (or noise). Other conventions are the same in Figure 4.

ratio (LRat), iii) likelihood regret (LReg) and iv) WAIC (Watanabe-Akaike Information Criterion). Briefly, IC employs a subtractive bias correction for each sample; the correction is computed based on sample complexity estimated with one of many compression algorithms (e.g. PNG) [20]. The likelihood ratio score computes the ratio between the sample likelihood obtained with a VAE trained with the original data, and one trained with noise-corrupted images [19]. Likelihood regret computes the improvement in the log-likelihood for a particular sample that can be achieved with a sample-specific optimization of the posterior $q_\phi(\mathbf{z}|\mathbf{x})$ [25]. Finally, WAIC is computed as $\mathbb{E}_\theta[\log p_\theta(\mathbf{x})] - \mathrm{Var}_\theta[\log p_\theta(\mathbf{x})]$, from an ensemble of deep generative models [3] (Appendix D, for details).

First, we performed an exhaustive (4×5) comparison of approaches with grayscale VAEs and datasets. For illustration, we discuss results for the problematic Fashion-MNIST VAE (Fig. 4, second column). The BC-LL score outperformed, or performed comparably with, other state-of-the-art methods for detecting outliers with the Fashion-MNIST VAE. Interestingly, IC performed relatively poorly for this VAE (Fig. 4, green symbols). Similar results were observed with the other grayscale VAEs: on average, BC-LL AU-ROC values were comparable with these state-of-the-art approaches (Fig. 4, bottom row, Average).

Next, we performed an exhaustive (5×6) comparison of all approaches with natural image VAEs and datasets. Again, we illustrate the results with the CelebA VAE

(Fig. 5, second column). As before, we noticed superlative outlier detection performance with the BC-LL score. Although, other methods (e.g. IC or LReg) performed better at detecting outliers belonging to specific datasets (e.g. SVHN, Fig. 5, top row, second column), BC-LL yielded the highest average AUROC values across all outlier datasets tested with the CelebA VAE (Fig. 5, bottom row, second column). Overall, across all combinations of VAEs and outlier datasets tested, BC-LL performed more consistently and yielded AUROC values on par with, or exceeding, state-of-the-art (Fig. 5, bottom row, Average).

Interestingly, outlier detection with BC-LL was relatively poor for the CIFAR-10 dataset, a failure shared by other approaches as well. Even IC that achieved high AU-ROC values for the challenging case of the CIFAR-10/ID versus SVHN/OOD case, performed sub-par when tested with the opposite pairing (SVHN/ID vs CIFAR-10/OOD) (Fig. 5, compare first row, last column with fifth row, first column). Moreover, IC's performance significantly deteriorated when tested against Noise OOD images for multiple ID datasets (e.g. SVHN, CelebA, CIFAR-10; Fig. 5, penultimate row, first, third and fifth columns). Similarly, the likelihood ratio metric that achieved superlative AUROCs for the SVHN/ID versus CelebA/OOD case failed when tested with the opposite pairing (CelebA/ID vs SVHN/OOD). The results indicate an important caveat with claiming superlative outlier detection by evaluating specific, challenging ID/OOD pairs (e.g. CIFAR-10/ID versus
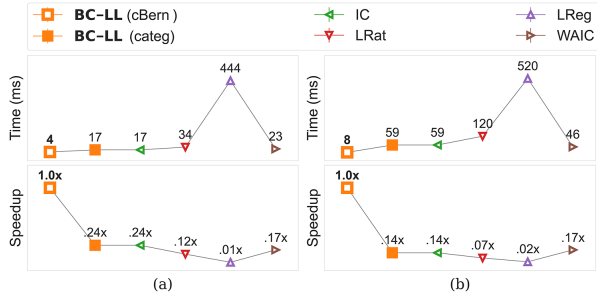
Figure 6. **Bias correction is significantly faster than competing approaches. (a)** (Top row) Compute times (ms) averaged across 500 test samples with each approach, for a grayscale image VAE (lower is better). (Bottom row) Speedup factor computed as the ratio of the compute time for each approach to the compute time for the BC-LL score (higher is better) **(b)** Same as in panel (a), but showing compute times (top row) and speedups (bottom row) using a natural image VAE. Other conventions are the same in Figures 4 and 5.

SVHN/OOD). It is possible that high outlier detection accuracies may occur for such specific pairings due to "over-correction" of the OOD likelihood by the respective methods. We discuss this case in the Limitations section and, in further detail, in Appendix G.

Finally, we compared compute times for our scores with previous outlier detection approaches, based on the inference time averaged across 500 randomly selected samples from each test dataset. Compute times for BC-LL scores outperformed competing approaches by a factor of up to 100x (Fig. 6), for both grayscale and natural image datasets. In summary, the bias corrected likelihood provides a simple, and computationally inexpensive approach for achieving state-of-the-art outlier detection.

5. Limitations and Conclusions

We have proposed, and extensively tested, a simple and efficient remedy to achieve robust outlier detection by de-biasing VAE likelihoods. We demonstrate the effectiveness of this remedy for outlier detection with multiple grayscale and natural image datasets (total of 9 datasets), using four different, popular VAE visible distributions (continuous Bernoulli, Bernoulli, categorical and truncated Gaussian), and comparing performance with four competing state-of-the-art, approaches [3, 19, 20, 25]. In nearly all cases, our correction approached or achieved state-of-the-art outlier detection performance. We also show that outlier detection is robust to the number of latent dimensions in the VAE (Appendix E.1). In addition, our remedy is computationally inexpensive and significantly outperforms all competing approaches in terms of evaluation times. In sum, lightweight remedies suffice to ameliorate biases arising from low-level image statistics (e.g. intensity or contrast variations), to

achieve robust outlier detection with VAE likelihoods.

Yet, we note that certain kinds of datasets pose a challenge for outlier detection with VAE likelihoods. For example, even with bias correction VAE likelihoods failed to achieve superlative outlier detection with the CIFAR-10 dataset (Fig. 5, last column). In fact, nearly all of the other competing approaches under-performed when tested with this dataset. These results suggest that the failure may be due to the unique nature of the CIFAR-10 dataset. Should we, or should we not, then discount VAE likelihoods as a generally applicable metric for outlier detection?

To answer this question we ask what, specifically, the VAE ELBO objective seeks to optimize. On the one hand, the VAE must learn a sparse representation of features in a low-dimensional latent space (the KL-divergence term). On the other hand, the VAE must also faithfully reconstruct the image, pixel-for-pixel (negative reconstruction error). In fact, minimizing the reconstruction error is key to achieving high VAE likelihoods. To achieve this latter objective, the VAE must also learn spatial relationships among its input features in a coordinate frame locked to the bounding frame (edges) of the image. Consequently, VAE likelihoods cannot (and do not) reflect solely high-level, semantic features of the image.

We propose that VAE likelihoods are ideally suited for outlier detection for datasets with relatively homogeneous images, in which features of the foreground objects (e.g. faces, cars) appear at consistent spatial locations in a coordinate frame locked to the edge of the image (e.g. Comp-Cars, GTSRB, CelebA, SVHN). In all of these cases, our bias correction sufficed to achieve excellent, and state-of-the-art outlier detection. Yet, when image features associated with the foreground object are heterogeneously and inconsistently distributed relative to edge coordinates (e.g. CIFAR-10), bias corrected likelihoods were not as effective. This reasoning also explains why conventional deep network models [11], or approaches based on optimizing VAE latent feature representations [25], or approaches that train models on deep classifier-based semantic feature representations [9] may outperform conventional VAE likelihood-based approaches for outlier detection, especially with heterogeneous datasets like CIFAR-10. We explore this hypothesis further in Appendix G.

Nonetheless, the bias correction we propose for VAEs will be relevant for outlier detection in key real-world applications. For example, medical imaging data such as retinal scans or MRI images are typically acquired in highly stereotyped spatial coordinates. In these cases, VAEs can be deployed as expert deep learning systems that are sensitive to fine-grained spatial relationships among image features (e.g. the morphology of pathological tissues) for accurate outlier detection. Future work will explore these applications.

References

[1] Sign Language MNIST. https://www.kaggle.com/datamunge/sign-language-mnist. Kaggle, 2017. 5

[2] Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations (Poster)*, 2016. 11

[3] Hyunsun Choi, Eric Jang, and Alexander A. Alemi. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. *arXiv e-prints*, page arXiv:1810.01392, Oct. 2018. 2, 7, 8, 12

[4] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017. 5

[5] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. 5, 15, 16, 19

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015. 11

[7] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1

[8] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, 2014. 1, 11

[9] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, volume 33, pages 20578–20589. Curran Associates, Inc., 2020. 8, 19

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 8, 19

[12] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. 2

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 2

[14] Gabriel Loaiza-Ganem and John P Cunningham. The continuous bernoulli: fixing a pervasive error in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 3

[15] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. *arXiv e-prints*, page arXiv:1906.02994, June 2019. 2, 3

[16] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. 2, 3

[17] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 5

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings*, 2016. 11

[19] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 3, 4, 7, 8, 11, 12

[20] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. 2, 3, 7, 8

[21] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[22] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. Selected Papers from IJCNN 2011. 2

[23] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1747–1756. JMLR.org, 2016. 4

[24] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 2

[25] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 20685–20696. Curran Associates, Inc., 2020. 2, 3, 4, 7, 8, 11, 12

[26] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3973–3981, 2015. 5

[27] Bang Xiang Yong, Tim Pearce, and Alexandra Melike Brintrup. Bayesian autoencoders : Analysing and fixing the

bernoulli likelihood for out-of-distribution detection. In *International Conference on Machine Learning Workshop on Uncertainty & Robustness in Deep Learning*, 2020. 3