Journal of Machine Learning for Biomedical Imaging. 2022:008. pp 1-23 Submitted 04/2022; Published 04/2022
Special Issue: IPMI 2021
Guest Editors: Aasa Feragen, Stefan Sommer, Julia Schnabel, Mads Nielsen

Deep Quantile Regression for Uncertainty Estimation in Unsupervised and Supervised Lesion Detection

Haleh Akrami Akrami@usc.edu
Department of Biomedical Engineering, University of Southern California, Los Angeles, USA
Anand A. Joshi ajoshi@usc.edu
Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA
Sergül Aydöre sergulaydore@gmail.com
Amazon Web Services, New York, USA
Richard M. Leahy leahy@sipi.usc.edu
Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA

Abstract

Despite impressive state-of-the-art performance on a wide variety of machine learning tasks in multiple applications, deep learning methods can produce over-confident predictions, particularly with limited training data. Therefore, quantifying uncertainty is particularly important in critical applications such as lesion detection and clinical diagnosis, where a realistic assessment of uncertainty is essential in determining surgical margins, disease status and appropriate treatment. In this work, we propose a novel approach that uses quantile regression for quantifying aleatoric uncertainty in both supervised and unsupervised lesion detection problems. The resulting confidence intervals can be used for lesion detection and segmentation. In the unsupervised setting, we combine quantile regression with the Variational AutoEncoder (VAE). The VAE is trained on lesion-free data, so when presented with an image with a lesion, it tends to reconstruct a lesion-free version of the image. To detect the lesion, we then compare the input (lesion) and output (lesion-free) images. Here we address the problem of quantifying uncertainty in the images that are reconstructed by the VAE as the basis for principled outlier or lesion detection. The VAE models the output as a conditionally independent Gaussian characterized by its mean and variance. Unfortunately, joint optimization of both mean and variance in the VAE leads to the well-known problem of shrinkage or underestimation of variance. Here we describe an alternative Quantile-Regression VAE (QR-VAE) that avoids this variance shrinkage problem by directly estimating conditional quantiles for the input image. Using the estimated quantiles, we compute the conditional mean and variance for the input image from which we then detect outliers by thresholding at a false-discovery-rate corrected p-value. In the supervised setting, we develop binary quantile regression (BQR) for the supervised lesion segmentation task. We show how BQR can be used to capture uncertainty in lesion boundaries in a manner that characterizes expert disagreement.

1. Introduction

Inference based on deep learning methods that do not take uncertainty into account can lead to over-confident predictions, particularly with limited training data (Reinhold et al., 2020). Quantifying uncertainty is particularly important in critical applications such as clinical diagnosis, where a realistic assessment of uncertainty is important in determining disease

https://www.melba-journal.org/papers/2022:008.html.

status and appropriate treatment. For example, in the lesion detection task, knowing the uncertainty in detected boundaries may help in defining tumor margins. In the literature, predictive uncertainty is categorized into two types based on the source of uncertainty: (i) aleatoric uncertainty (Lakshminarayanan et al., 2016) that is the result of uncertainty inherent in the data, and (ii) epistemic uncertainty, which is often referred to as model uncertainty, as it is due to model limitations. Access to unlimited training data does not reduce the former uncertainty in contrast to the latter. Here we focus on aleatoric uncertainty and its estimation using quantile regression (QR) (Koenker and Bassett Jr, 1978).

A recent novel approach proposed using conditional QR to estimate aleatoric uncertainty in neural networks (Romano et al., 2019; Tagasovska and Lopez-Paz, 2019). QR can be used to estimate the conditional median (0.5 quantiles) or other quantiles of the response variable, conditioned on the input feature variable (Yu and Moyeed, 2001). QR is most commonly applied in cases where the parametric likelihood cannot be specified (Yu and Moyeed, 2001), here we apply develop QR methods for Gaussian (supervised) and binary (unsupervised) applications.

Lesion detection is an important application of deep learning in medical image processing. Here we address the important problem of learning uncertainty in order to perform statistically-informed inference for this application. Lesion detection can be applied either in a supervised framework when labels are available or with an unsupervised framework using a generative model such as the VAE. We describe how aleatoric uncertainty can be quantified in both of these settings using quantile regression to define confidence intervals, which are then used to identify lesions. In both supervised and unsupervised frameworks, we apply quantile regression by changing the loss function of the network. For quantile regression in the unsupervised setting, we use the formulation developed by He (1997). Our goal is to learn the characteristics of the input distribution to separate inliers from outliers. The quantiles help us to define confidence intervals from which we can identify outliers. For the supervised setting we use binary quantile regression (Koenker and Hallock, 2001) in order to capture uncertainty in binary classification problems. In both scenarios, our goal is to estimate aleatoric uncertainty in segmentation and calculate a confidence interval associated with each segmentation.

Unsupervised Lesion Detection: Generative models, including autoencoders, can be used for unsupervised lesion detection. Once the distribution of anomaly-free samples is learned during the training, during inference we can compute the reconstruction error between a given image and its reconstruction to identify abnormalities (Aggarwal, 2015; Akrami et al., 2021, 2020, 2019). Decisions on the presence of outliers are often performed based on empirically chosen thresholds. Here we use quantile regression to define a principled-approach to thresholding.

Variational autoencoder (VAE) (Kingma and Welling, 2013) and its variants can approximate the underlying distribution of high-dimensional data. VAEs are trained using the variational lower bound of the marginal likelihood of data as the objective function. They can then be used to generate samples from the data distribution, where probabilities at the output are modeled as parametric distributions such as Gaussian or Bernoulli that are conditionally independent across output dimensions (Kingma and Welling, 2013). By using VAE, An and Cho (2015) proposed to use reconstruction probability rather than the

reconstruction error to detect outliers. This allows a more principled approach to anomaly detection since inference is based on quantitative statistical measures and can include corrections for multiple comparisons.

For determining the reconstruction probability, we need to predict both conditional mean and variance using VAEs for each of the output dimensions. The estimated variance represents an aleatoric uncertainty associated with the conditional estimates given the data (Reinhold et al., 2020). Estimating variance is more challenging than estimating the mean in generative networks due to the unbounded likelihood (Skafte et al., 2019). In the case of VAEs, if the conditional mean network prediction is nearly perfect (zero reconstruction error), then maximizing the log-likelihood pushes the estimated variance towards zero in order to maximize the likelihood. This also makes VAEs susceptible to overfitting the training data giving a near-perfect reconstruction on the training data and very small uncertainty. This near-zero variance does not reflect the true performance of the VAE on the test data. For this reason, near-zero variance estimates, with the log-likelihood approaching an infinite supremum, do not lead to a good generative model. It has been shown that there is a strong link between this likelihood blow-up and the mode-collapse phenomenon (Mattei and Frellsen, 2018; Reinhold et al., 2020). In fact, in this case, the VAE behaves much like a deterministic autoencoder (Blaauw and Bonada, 2016).

While the classical formulation of VAEs allows both mean and variance estimates (Kingma and Welling, 2013), because of the variance shrinkage problem, most if not all implementations of VAE, including the standard implementations in PyTorch and Tensorflow libraries, estimate only the mean with a fixed value of variance (Skafte et al., 2019). Here we describe an approach that overcomes the variance shrinkage problem in VAEs using quantile regression (QR) in place of variance estimation. We then demonstrate the application of this new QR-VAE by computing reconstruction probabilities for a brain lesion detection task.

Supervised Lesion Detection: Labelled training data are preferable, if available, as they lead to better performance compared to unsupervised models (You et al., 2019). One approach to estimating uncertainty is to use the soft-max probability of the cross-entropy loss (DeVries and Taylor, 2018). Softmax probabilities are known to be poorly calibrated, and imperceptible perturbations to the input image can change the deep network's softmax output significantly (Guo et al., 2017). Softmax confidence also conflates two different sources of uncertainty (aleatoric and epistemic). Bayesian neural networks (Neal, 2012) can be used for estimating aleatoric uncertainty by measuring conditional entropy; however, these models are not able to capture multimodal uncertainty profiles (Tagasovska and Lopez-Paz, 2019). An alternative method to capture aleatoric uncertainty is using quantile regression (Tagasovska and Lopez-Paz, 2018). Here we used binary quantile regression (Kordas, 2006; Manski, 1985) to capture the quantiles of labels which can be used to define multiple nested segmentation masks with increasing uncertainty. Our goal is to capture the source of uncertainty within the data distribution where there is more than one plausible answer to the segmentation problem due to disagreement between the specialists who labeled the data. Binary quantile regression is also robust to label noise (Oh et al., 2016). Finally, Kordas et al. (2006) showed that binary quantile regression can be useful for unbalanced data and leads to a more comprehensive view on how the predictor variables influence the response.

Related Work: A few recent papers have targeted the variance shrinkage problem. Among these, Detlefsen et al. (2019) describe reliable estimation of the variance using Comb-VAE, a locally aware mini-batching framework that includes a scheme for unbiased weight updates for the variance network. In an alternative approach Stirn and Knowles 2020, (Stirn and Knowles, 2020) suggest treating variance variationally, assuming a Student's t-likelihood for the posterior to prevent optimization instabilities and a Gamma prior for the precision parameter of this distribution. The resulting Kullback–Leibler (KL) divergence induces gradients that prevent the variance from approaching zero (Stirn and Knowles, 2020).

In the supervised framework, several papers estimate uncertainty for segmentation; however, only a few separately consider aleatoric uncertainty and focus on multi-rated labels (Czolbe et al., 2021; Hu et al., 2019; Islam and Glocker, 2021; Kohl et al., 2018). Czolbe et al. (2021) compared these methods to investigate whether they are helpful in an assessment of segmentation quality and active learning. Recently, Monteiro et al. (2020) used a stochastic segmentation network for modeling spatially correlated uncertainty in image segmentation. They applied a multivariate Normal distribution over the softmax logits and used low-rank approximation to estimate the full covariance matrix across all the pixels in the image (Monteiro et al., 2020).

Our Contribution: In the unsupervised setting, to obtain a probabilistic threshold and address the variance shrinkage problem, we suggest an alternative and attractively simple solution. Assuming the output of the VAE has a Gaussian distribution, we quantify uncertainty in VAE estimates using conditional quantile regression (QR-VAE). The aim of conditional quantile regression (Koenker and Bassett Jr, 1978) is to estimate a quantile of interest. Here we use these quantiles to compute variance, thus sidestepping the shrinkage problem. It has been shown that quantile regression is able to capture aleatoric uncertainty (Tagasovska and Lopez-Paz, 2019). We demonstrate the effectiveness of our method quantitatively and qualitatively on simulated and brain MRI datasets. Our approach is computationally efficient and does not add any complications to training or sampling procedures. In contrast to the VAE loss function, the QR-VAE loss function does not have an interaction term between quantiles, and therefore, shrinkage does not happen. Since quantile regression does not satisfy finite-sample coverage guarantees, we applied conformalized quantile regression (Romano et al., 2019), which combines conformal prediction with classical quantile regression, to have the theoretical guarantee of valid coverage.

We also use binary quantile regression in a supervised framework in order to capture the uncertainty of lesion annotations. We demonstrate estimation of multiple quantiles in imaging data in which each lesion is delineated by four human observers and compare to human-rater ground truth and a binary cross-entropy formulation.

A preliminary version of these results was presented in Akrami et al. (2021). The novel extensions presented in the current work include: (1) application of conformalized quantile regression to unsupervised learning (section 3.1); (2) extension of unsupervised approach to the supervised approach using binary quantile regression (section 3.2), (3) application of binary quantile regression for lesion detection and uncertainty estimation, (section 4.3), and (4) additional results and validation that extend those in the earlier paper. We provide a public version of our code at `https://github.com/ajoshiusc/QRSegment` and `https://github.com/ajoshiusc/QRVAE`.

2. Background

2.1 Variance Shrinkage Problem in Variational Autoencoders

Let $x_i \in \mathbb{R}^D$ be an observed sample of random variable X where $i \in \{1, \cdots, N\}$, D is the number of features and N is the number of samples; and let z_i be an observed sample for latent variable Z. Given a sample x_i representing the input data, the VAE is a probabilistic graphical model that estimates the posterior distribution $p_\theta(Z|X)$ as well as the model evidence $p_\theta(X)$, where θ are the generative model parameters (Kingma and Welling, 2013). The VAE approximates the posterior distribution of Z given X by a tractable parametric distribution and minimizes the ELBO (evidence lower bound) loss (An and Cho, 2015). It consists of the encoder network that computes $q_\phi(Z|X)$, and the decoder network that computes $p_\theta(X|Z)$ (Wingate and Weber, 2013), where ϕ and θ are model parameters. Since we use the neural network for learning the distributions $q_\phi(Z|X)$ and $p_\theta(X|Z)$, the parameters θ and ϕ are modeled by the weights of the encoder and decoder networks and will be learned from the data during training. The marginal likelihood of an individual data point can be rewritten as follows:

$$\log p_\theta(x_i) = D_{KL}(q_\phi(Z|x_i), p_\theta(Z|x_i)) + L(\theta, \phi; x_i), \tag{1}$$

where

$$L(\theta, \phi; x_i) = \mathbb{E}_{q_\phi(Z|x_i)}[\log(p_\theta(x_i|Z))] - D_{KL}(q_\phi(Z|x_i)||p_\theta(Z)). \tag{2}$$

The first term (log-likelihood) in equation 2 can be interpreted as the *reconstruction loss* and the second term (KL divergence) as the *regularizer*. The total loss over all samples can be written as:

$$L(\theta, \phi, X) = L_{REC} + L_{KL} \tag{3}$$

where $L_{REC} \coloneqq \mathbb{E}_{q_\phi(Z|X)}[\log(p_\theta(X|Z))]$ and $L_{KL} \coloneqq D_{KL}(q_\phi(Z|X)||p_\theta(Z))$.

Assuming the posterior distribution is Gaussian and using a 1-sample approximation (Skafte et al., 2019), the likelihood term simplifies to:

$$L_{REC} = \sum_i \frac{-1}{2} \log(\sigma_\theta^2(z_i)) - \frac{(x_i - \mu_\theta(z_i))^2}{2\sigma_\theta^2(z_i)} \tag{4}$$

where $Z \sim p(Z) = N(0, I)$ (I is identity matrix), $X|Z \sim p_\theta(X|Z) = N(X|\mu_\theta(Z), \sigma_\theta(Z))$, and $Z|X \sim q_\phi(Z|X) = N(Z|\mu_\phi(X), \sigma_\phi(X))$. $\mu_\theta(Z), \sigma_\theta(Z)$ are posterior mean and variance; $\mu_\phi(X)$, and $\sigma_\phi(X)$ are encoder mean and variance. Optimizing VAEs over mean and variance with a Gaussian posterior is challenging (Skafte et al., 2019). If the model has sufficient capacity that there exists (ϕ, θ) for which $\mu_\theta(z)$ provides a sufficiently good reconstruction, then the second term pushes the variance to zero before the term $\frac{-1}{2}\log(\sigma_\theta^2(x_i)))$ catches up (Blaauw and Bonada, 2016; Skafte et al., 2019).

One practical example of this behavior is in speech processing applications (Blaauw and Bonada, 2016). The input is a spectral envelope which is a relatively smooth 1D curve. Representing this as a 2D image produces highly structured and simple training images. As

a result, the model quickly learns how to accurately reconstruct the input. Consequently, reconstruction errors are small and the estimated variance becomes vanishingly small. Another example is 2D reconstruction of MRI images where the images from neighbouring 2D slices are highly correlated leading again to variance shrinkage (Volokitin et al., 2020). To overcome this problem, variance estimation networks can be avoided using a Bernoulli distribution or by simply setting variance to a constant value (Skafte et al., 2019).

2.2 Conditional Quantile Regression

In contrast to classical parameter estimation where the goal is to estimate the conditional mean of the response variable given the feature variable, the goal of quantile regression is to estimate conditional quantiles based on the data (Yu and Moyeed, 2001). The most common application of quantile regression models is in cases in which a parametric likelihood cannot be specified (Yu and Moyeed, 2001). Another motivation for quantile regression is that quantiles are robust to outliers (John, 2015).

Quantile regression can be used to estimate the conditional median (0.5 quantile) or other quantiles of the response variable conditioned on the input data. The α-th conditional quantile function is defined as $q_\alpha(x) \coloneqq \inf\{y \in \mathbb{R} : F(y|X = x) \geq \alpha\}$ where $F = P(Y \leq y)$ is a strictly monotonic cumulative distribution function. Similar to classical regression analysis which estimates the conditional mean, the α-th quantile regression ($0 < \alpha < 1$) seeks a solution to the following minimization problem for input x and output y (Koenker and Bassett Jr, 1978; Yu and Moyeed, 2001):

$$\arg\min_{\theta} \sum_{i} \rho_\alpha(y_i - f_\theta(x_i)) \tag{5}$$

where x_i are the inputs, y_i are the responses, ρ_α is the *check function* or *pinball loss* (Koenker and Bassett Jr, 1978) and f is the model paramaterized by θ. The goal is to estimate the parameter θ of the model f. The *pinball loss* is defined as:

$$\rho_\alpha(y, f_\theta(x_i)) \coloneqq \begin{cases} \alpha(y - f_\theta(x_i)) & \text{if } (y - f_\theta(x_i)) > 0 \\ (1-\alpha)(y - f_\theta(x_i)) & \text{otherwise.} \end{cases} \tag{6}$$

Due to its simplicity and generality, quantile regression is widely applicable in classical regression and machine learning to obtain a conditional prediction interval (Rodrigues and Pereira, 2020). It can be shown that minimization of the loss function in equation 5 is equivalent to maximization of the likelihood function formed by combining independently distributed asymmetric Laplace densities (Yu and Moyeed, 2001):

$$\arg\max_{\theta} L(\theta) = \frac{\alpha(1-\alpha)}{\sigma} \exp\left\{\frac{-\sum_i \rho_\alpha(y_i - f_\theta(x_i))}{\sigma}\right\}$$

where σ is the scale parameter. Individual quantiles can be shown to be maximum likelihood estimates of Laplacian density. In this paper we estimate two quantiles jointly and therefore our loss function can be seen as a summation of two Laplacian likelihoods.

3. Deep Uncertainty Estimation with Quantile Regression

3.1 Quantile Regression Varational Autoencoder (QR-VAE)

Instead of estimating the conditional mean and conditional variance directly at each pixel (or feature), the outputs of our QR-VAE are multiple quantiles of the output distributions at each pixel. This is achieved by replacing the Gaussian likelihood term in the VAE loss function with the quantile loss (check or pinball loss). For the QR-VAE, if we assume a Gaussian output, then only two quantiles are needed to fully characterize the Gaussian distribution. Specifically, we estimate the median and 0.15-th quantile, which corresponds to approximately one standard deviation (more precisely 1.036 std dev.) from the mean under the Gaussian model. Our QR-VAE ouputs, Q_L (low quantile) and Q_H (high quantile), are then used to calculate the mean and the variance. To find these conditional quantiles, fitting is achieved by minimization of the pinball loss for each quantile. The resulting reconstruction loss for the proposed model can be calculated as:

$$L_{REC} = \sum_i \rho_L(x_i - f_{\theta_L}(x_i)) + \sum \rho_H(x_i - f_{\theta_H}(x_i))$$

where θ_L and θ_H are the parameters of the models corresponding to the quantiles Q_L and Q_H, respectively. The minimization of this loss results in the desired quantile estimates for each output pixel.

We reduce the chance of quantile crossing (consistency in the quantiles defined by: $Q_{\tau_1} \subset Q_{\tau_2}$ when $\tau_1 > \tau_2$) by limiting the flexibility of independent quantile regression. This is done by simultaneous estimation of both quantiles with one neural network, rather than training separate networks for each quantile (Rodrigues and Pereira, 2020). Note that the estimated quantiles share network parameters except for the last layer.

While quantile regression guarantees coverage of data (based on the quantiles chosen) in the training set, performance on a held-out validation data is not guaranteed. In order to have a coverage guarantee for finite samples on unseen data, we deployed conformalized quantile regression using a calibration set as explained in Romano et al. (2019). Conformal predictions provide a non-asymptotic, distribution-free coverage guarantee (Shafer and Vovk, 2008). The main idea of conformal prediction is to fit a model on the training data, then use the residuals on held-out calibration data to quantify the uncertainty in future predictions. This offers finite sample distribution-free performance guarantees. The conformalized quantile regression approach combines conformal prediction with quantile regression (Sesia and Candès, 2020). For this, we use the approach presented in Romano et al. (2019) that combines conformal prediction with quantile regression. This approach inherits both the finite sample, distribution-free validity of conformal prediction and the statistical efficiency of quantile regression.

3.2 Binary Quantile Regression U-Net (BQR U-Net)

For calculating quantiles of a binary response, consider the following model:

$$Y^* = h(x) + \epsilon, \quad Y = I\{Y^* \geq 0\}$$

where Y^* is a hidden variable, $h(x)$ is the true model, and ϵ is the noise. No distribution ϵ is assumed for the model (Kordas, 2006). Since the indicator function is monotone increasing,

and since quantiles are invariant under monotone transform, we have:

$$Q_\tau(Y|X) = I(Q_\tau(Y^*|X)).$$

$Q_\tau(Y|X)$ is the τth conditional quantile of Y given X. By modeling $Q_\tau(Y^*|X) = f_\tau(X, \beta)$ with β as the parameter, the β parameter can be estimated by:

$$\arg\min_\beta \sum_i \rho_\tau(y_i - I(f_\tau(x_i, \beta) \geq 0))$$

which can be shown to be equivalent to a maximization problem (Kordas, 2006):

$$\arg\max_\beta \sum_i [y_i - (1-\tau)] I(f_\tau(x_i, \beta) \geq 0)$$

However, the function is not differentiable, because of the use of the indicator function. To apply gradient based optimization methods for training the neural network, we use the smoothed approximation (Kordas, 2006):

$$\arg\max_\beta \sum_i [y_i - (1-\tau)] K(f_\tau(x_i, \beta)) \tag{7}$$

where $K(t)$ is smoothed version of the indicator function, with the following properties:

$$K(t) \geq 0, \forall t \in R, \lim_{t\to+\infty} K(t) = 1, \lim_{t\to-\infty} K(t) = 0.$$

Specifically, to train the neural networks, we choose $K(t) = \frac{1}{1+e^{-t}}$, the sigmoid function, which has the desired properties.

In this paper we us the BQR loss to solve the lesion detection and segmentation task. We use a U-Net architecture with multiple heads (output branches), where each head estimates a specific quantile for the labels at the pixel level. We observed that joint estimation of multiple quantiles is computationally faster than solving for each separately, and also avoids the quantile crossing problem.

A standard U-Net would use a cross-entropy loss for this segmentation task. Here we replace this with the BQR loss. To estimate the n-th quantile, the BQR loss is given by:

$$\text{Loss} = \sum_n \sum_i [y_i - (1-\tau_n)] K(f_{\tau_n}(x_i, \beta_n)) \tag{8}$$

where each f correspond to a head of U-Net, $\tau_1 ... \tau_n$ shows different quantiles and $\beta_1 ... \beta_n$ are estimated parameter for each quantile respectively. A single network is used to estimate all quantiles. We chose to train a single network with output branches for each quantile since there is a common network across quantiles except for the last layer. This makes training of the quantiles consistent avoiding crossing of the estimated quantiles. We choose $K(t)$ to be the sigmoid function.

4. Experiments and Results

We evaluate our proposed approaches for supervised and unsupervised deep quantile regression on (i) A simulated dataset for density estimation, (ii) Unsupervised lesion detection in a brain imaging dataset, and (iii) Supervised lesion detection in a lung cancer dataset. For the simulated data, we compare our results qualitatively and quantitatively, using KL divergence between the learned distribution and the original distribution, with Comb-VAE (Skafte et al., 2019) and VAE as baselines. For unsupervised lesion detection, we compare our lesion detection results with the VAE, which estimates both mean and variance. The area under the receiver operating characteristic curve (AUC) and dice coefficients are used as performance metrics. We also performed the unsupervised lesion detection task nonparametrically, estimating upper and lower quantiles of the images and then assigning voxel lesion labels if their intensities are outside those quantiles. Using the BQR U-Net, we estimated the thresholded probability of the labels for a dataset with multiple (4) annotators per image. We compared the dice coefficient of these thresholded areas obtained using the BQR U-Net with their corresponding counterparts calculated both using the softmax probability of a deterministic U-Net and the ground truth (as determined by the four human raters).

4.1 Simulations for VAE

Following Skafte et al. (2019), we first evaluate variance estimation using VAE, Comb-VAE, and QR-VAE on a simulated dataset. The two moon dataset inspires the data generation process for this simulation[1]. First, we generate 500 points in $\mathbb{R}^2$ in a two-moon manner to generate a known two-dimensional latent space. These are then mapped to four dimensions (v_1, v_2, v_3, v_4) using the following equations:

$$
\begin{aligned}
v_1(z_1, z_2) &= z_1 - z_2 + \epsilon\sqrt{0.03 + 0.05(3 + z_1)} \\
v_2(z_1, z_2) &= z_1^2 - \frac{1}{2}z_2 + \epsilon\sqrt{0.03 + 0.03||z_1||_2} \\
v_3(z_1, z_2) &= z_1 z_2 - z_1 + \epsilon\sqrt{0.03 + 0.05||z_1||_2} \\
v_4(z_1, z_2) &= z_1 + z_2 + \epsilon\sqrt{0.03 + \frac{0.03}{0.02 + ||z_1||_2}}
\end{aligned}
$$

where ϵ is sampled from a normal distribution. For more details about the simulation, please refer to Skafte et al. (2019)[2]. After training the models, we first sample from z using a Gaussian prior, and input that sample to the decoder to generate parameters of the posterior $p_\theta(x|z)$, and then sample again from the posteriors using the estimated means and variances from the decoder. The distribution of these generated samples represents the learned distribution in the generative model.

In Figure 1, we plot the pairwise joint distribution for the input data as well as the generated samples using various models. We used Gaussian kernel density estimation to model the distributions from 1000 samples in each case. We observe that the standard VAE

1. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons
2. https://github.com/SkafteNicki/john/blob/master/toy_vae.py

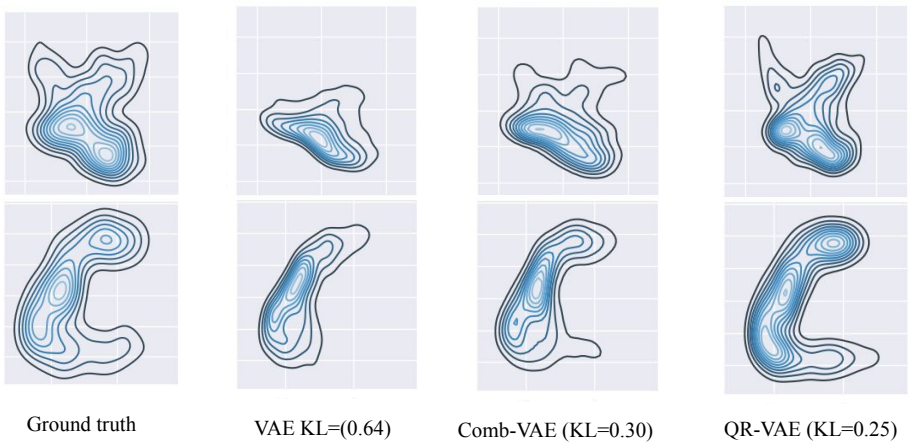

Figure 1: Pairwise joint distribution of the ground truth and generated distributions. Top: v_1 vs. v_2 dimensions. Bottom: v_2 vs v_3 dimensions. From left to right: original distribution and distributions computed using VAE, Comb-VAE and QR-VAE, respectively. We also list the KL divergence between the learned distribution and the original distribution in each case.

underestimates the variance resulting in insufficient learning of the data distribution. The samples from our QR-VAE model capture a data distribution more similar to the ground truth than either standard VAE or Comb-VAE. Our model also outperforms VAE and Comb-VAE in terms of KL divergence between input samples and generated samples as can be seen in Figure 1. The KL-divergence is calculated using universal-divergence, which estimates the KL divergence based on k-nearest-neighbor (k-NN) distance (Wang et al., 2009)[3].

4.2 Unsupervised Lesion Detection

4.2.1 Network Architecture

Next, we investigate utility of the proposed QR-VAE for the medical imaging application of detecting brain lesions. Multiple automatic lesion detection approaches have been developed to assist clinicians in identifying and delineating lesions caused by congenital malformations, tumors, stroke or brain injury. The VAE is a popular framework among the class of unsupervised methods (Chen and Konukoglu, 2018; Baur et al., 2018; Pawlowski et al., 2018). After training a VAE on a lesion free dataset, presentation of a lesioned brain to the VAE will typically result in reconstruction of a lesion-free equivalent. The error between input and output images can therefore be used to detect and localize lesions. However, selecting an appropriate threshold that differentiates lesion from noise is a difficult task. Furthermore, using a single global threshold across the entire image will inevitably lead to a poor trade-off between true and false positive rates. Using the QR-VAE, we can compute the conditional mean and variance of each output pixel. This allows a more reliable and sta-

3. `https://pypi.org/project/universal-divergence`

tistically principled approach for detecting anomalies by thresholding based on computed p-values. Further, this approach also allows us to correct for multiple comparisons.

The network architectures of the VAE and QR-VAE are chosen based on previously established architectures (Larsen et al., 2015). Both the VAE and QR-VAE consist of three consecutive blocks of convolutional layer, a batch normalization layer, a rectified linear unit (ReLU) activation function and a fully-connected layer in the bottleneck for the encoder. The decoder includes three consecutive blocks of deconvolutional layers, a batch normalization layer and ReLU followed by the output layer that has 2 separate deconvolution layers (for each output) with sigmoid activations. For the VAE, the outputs represent mean and variance while for QR-VAE the outputs represent two quantiles from which the conditional mean and variance are computed at each voxel. The size of the input layer is $3 \times 64 \times 64$ where the first dimension represents three different MRI contrasts: T1-weighted, T2-weighted, and FLAIR for each image.

4.2.2 Training, Validation, and Testing Data

For training we use 20 central axial slices of brain MRI datasets from a combination of 119 subjects from the Maryland MagNeTS study (Gullapalli, 2011) of neurotrauma and 112 subjects from the TrackTBI-Pilot (Yue et al., 2013) dataset, both available for download from `https://fitbir.nih.gov`. We use 2D slices rather than 3D images to make sure we had a large enough dataset for training the VAE. These datasets contain T1, T2, and FLAIR images for each subject, and have sparse lesions. We have found that in practice both VAEs have some robustness to lesions in these training data so that they are sufficient for the network to learn to reconstruct lesion-free images as required for our anomaly detection task. The three imaging modalities (T1, T2, FLAIR) were rigidly co-registered within subject and to the MNI brain atlas reference and re-sampled to 1mm isotropic resolution. Skull and other non-brain tissue were removed using BrainSuite (`https://brainsuite.org`). Subsequently, we reshaped each sample into 64×64 dimensional images and performed histogram equalization to a lesion free reference. We separated 40 subjects as the validation/calibration set.

We evaluated the performance of our model on a test set consisting of 20 central axial slices of 28 subjects from the ISLES (The Ischemic Stroke Lesion Segmentation) database (Maier et al., 2017) for which ground truth, in the form of manually-segmented lesions, is also provided. We performed similar pre-processing as for the training set.

4.2.3 Model-free Anomaly Detection

For simplicity, we first performed the lesion detection task using the QR-VAE without the Gaussian assumption as shown in Figure 2. We trained the QR-VAE to estimate the $Q_{0.025}$ and $Q_{0.975}$ quantiles. We then used these quantiles directly to threshold the input images for anomalies. This leads to a nominal 5% (per pixel) false positive rate. This method is simple and avoids the need for validation data to determine an appropriate threshold. However, without access to p-values we are unable to determine a threshold that can be used to correct for multiple comparisons by controlling the false-discovery or family-wise error rate. A model needs to be assumed to obtain p-value as we do in 4.2.4. The Dice coefficient for this model was 0.37 and 0.32 with and without conformalization respectively

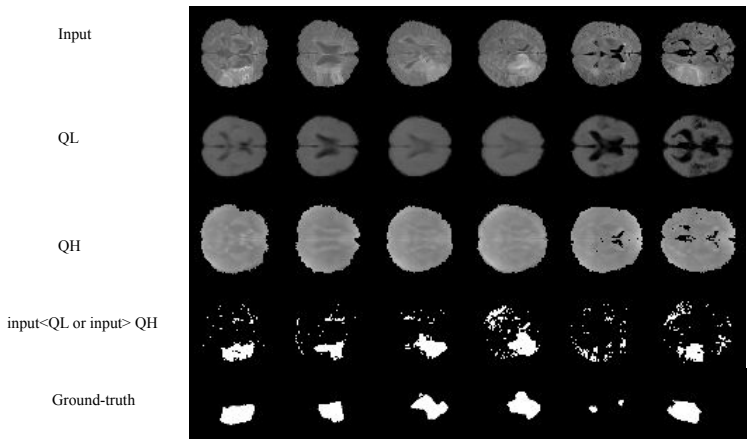

Figure 2: Model-free lesion detection for ISLES dataset using $Q_L = Q_{0.025}$ and $Q_H = Q_{0.975}$. Pixels outside the $[Q_L, Q_H]$ interval are marked outliers. Estimated quantiles are the outputs of QR-VAE.

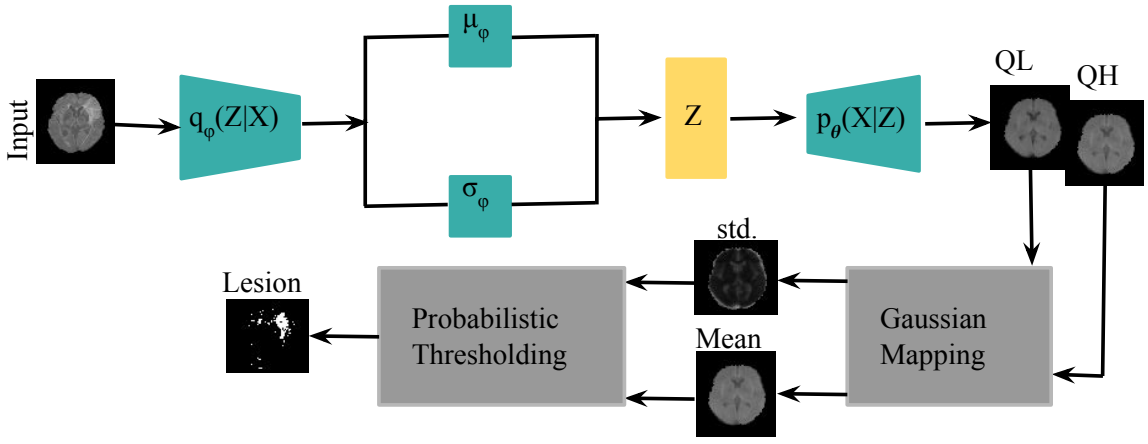

Figure 3: Estimating two quantiles in the ISLES dataset using QR-VAE. Using the Gaussian assumption for the posterior, there is 1-1 mapping from these quantiles to mean and standard deviation.

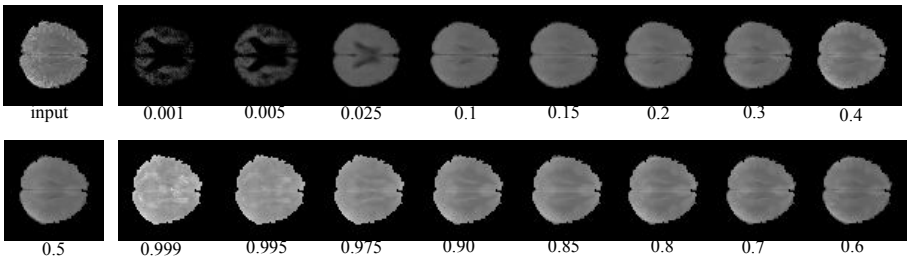

Figure 4: Pixel-wise quantile image thresholds for a single test image as a function of quantile computed using the QR-VAE.

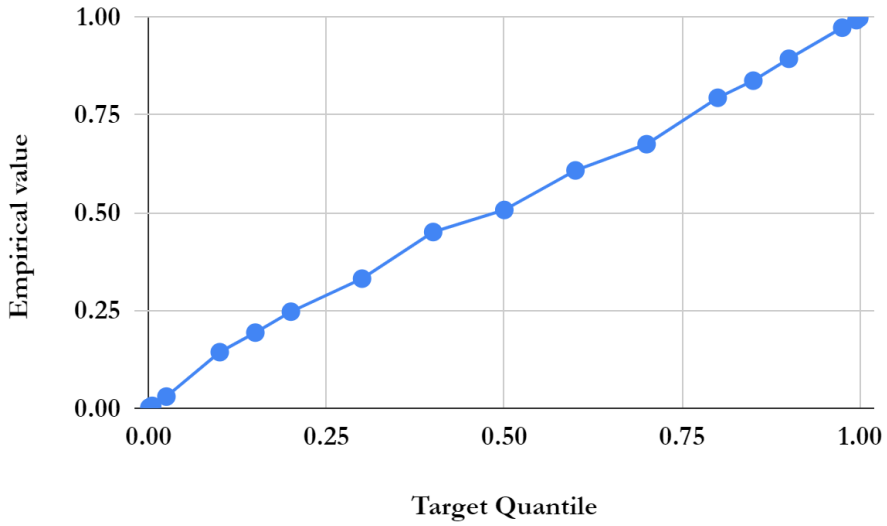

Figure 5: Vertical axis indicates the fraction of pixels in the entire testing set whose intensity is below the corresponding quantile for that pixel as computed using the QR-VAE. Note that as aggregated over the entire test set, the computed pixel-wise quantiles closely match the true distribution assuming anomaly-free data (in practice the fraction of anomalous pixels is a very small fraction of the total, so the presence of lesions in the data should not substantially affect this plot).

(see Table 1). For validating the accuracy of the computed quantiles we calculated the the percentage of pixels that lie below the estimated quantiles. Even in the extreme quantiles the percentage of pixels with lower intensity than the threshold predicted by each quantile was very close to each of the estimated quantile values (Figures 4 and 5).

4.2.4 Gaussian model for anomaly detection

In a second experiment, we trained a VAE with a Gaussian posterior and the QR-VAE as illustrated in Figure 3, in both cases estimating conditional mean and variance. Specifically, we estimated the $Q_{0.15}$ and $Q_{0.5}$ quantiles for the QR-VAE and used these values to compute pixel-wise mean and variance assuming a Gaussian model. By comparing the pixel intensity to the Gaussian model values, we can compute p-values for each pixel. In order to identify or segment lesions, we threshold the pixel-wise $p-$values. Naively applying a threshold separately at each pixel will result in a large number of false positives because of the multiple comparisons problem (Shaffer, 1995). For example, if all pixels are independent, and follow the null distribution, then thresholding at an $\alpha = 0.05$ significance level value would lead to 5% of all pixels being identified as lesion, even though none were present. While in practice this number is much lower because of spatial correlation in the image, it is still important to account for multiple comparisons.

The best known such adjustment is the Bonferroni correction (Bland and Altman, 1995). In medical imaging applications, this correction tends to be too conservative since pixels are correlated. Other methods for multiple comparison correction are designed to control the Family-Wise Error Rate (FWER, probability of making one or more false discoveries) (Tukey, 1953) or the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). The FDR is the expected ratio of the number of false positives to the total number of positives (rejections of the null). In other words, in an FDR-corrected thresholding at an $\alpha = 0.05$ significance level, we would expect 5% of the detected lesion pixels to be false positives. Here we use the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) with $\alpha = 0.05$. As shown in Figure 6, the VAE underestimates the variance, so that most of the brain shows significant p-values, even with FDR correction. On the other hand, the QR-VAE's thresholded results detect anomalies that reasonably match the ground truth. To produce a quantitative measure of performance, we also computed the area under the ROC curve (AUC) for VAE and QR-VAE. To do this we first computed z-score images by subtracting the mean and normalizing by standard deviation. We then applied a median filtering with a 7×7 window. By varying the threshold on the resulting images and comparing it to ground truth, we obtained AUC values of 0.52 for the VAE and 0.92 for the QR-VAE. We also obtained Dice coefficient values of 0.006 for the VAE and 0.37 for the QR-VAE. All quantitative measurements were computed on a voxel-wise basis. Also note that with the conformalized formulation, the Dice coefficient further increased to 0.41 (Table 1). Results using the conformalized formulation improved performance of the QR-models by calibrating the quantiles, resulting in an increase in the Dice coefficients in both Gaussian and model-free cases.

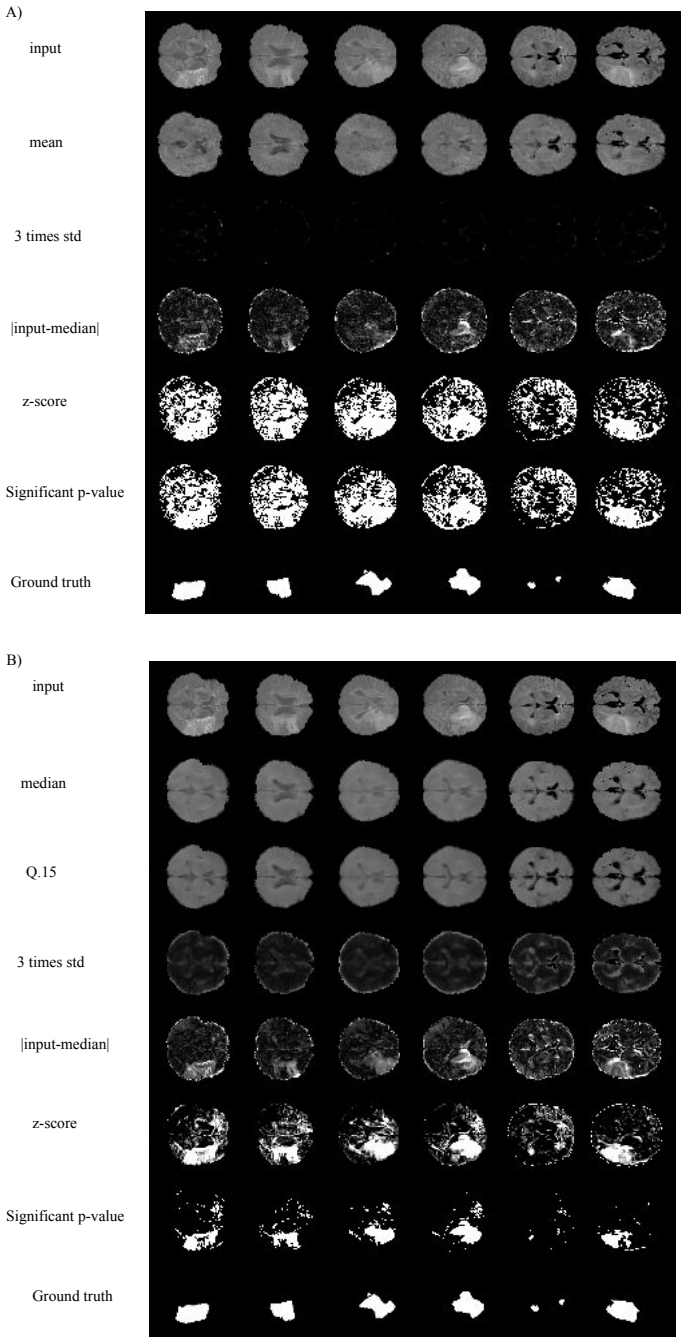

Figure 6: Lesion detection for the ISLES dataset. A) VAE with mean and variance estimation B) QR-VAE. First, we normalize the error value using the pixel-wise model's estimates of mean and variance. The resulting z-score is then converted to an FDR-corrected p-value and the images are thresholded at a significance level of 0.05. The bottom rows represent ground truth based on expert manual segmentation of lesions.

Table 1: Compariaon of the performance of unsupervised lesion detection for VAE and QR-VAE, with and without conformalization. QR-VAE-conf: conformalized QR-AVE; QR-VAE-GS: Gaussian QR-VAE; QR-VAE-GS-conf: Gaussian conformalized QR-VAE.

	VAE	QR-VAE	QR-VAE-conf	QR-VAE-GS	QR-VAE-GS-conf
AUC	0.52	N/A	N/A	0.92	0.92
Dice coefficient	0.006	0.32	0.37	0.37	0.41

4.3 Supervised Lesion Detection

We evaluated our BQR supervised approach on the LIDC-IDRI dataset (Armato III et al., 2011). This data consists of 1018 3D thorax CT scans annotated by four radiologists tasked with finding multiple lung nodules in each scan. The data is ideal for capturing inherent uncertainty in the data labels that comes from disagreement between experts. The data was preprocessed as described by Kohl et al. (2018). They extracted 2D slices centred around the annotated nodules and generated 180 x 180 images when at least one expert has segmented a nodule. This process resulted in a dataset of 8882 images in the training set, 1996 images in the validation set, and 1992 images in the test set. We reshaped the data into 128 x 128 images for input to the neural network. We used a U-Net architecture (Ronneberger et al., 2015) with 2D convolutional layers. The output layer was modified to generate four quantiles with four output branches using softmax activation. We compared the performance of BQR U-Net with the deterministic U-Net (Ronneberger et al., 2015). As ground truth for the comparison, first we estimated agreement maps for each test image by combining the lesion annotations of the four raters. This generates, for each image, an annotation with $P(Y = 1|X)$ values greater than or equal to 0, 0.25, 0.5, 0.75,1. X here represents the input image. Given an input image, the BQR U-Net generates output regions where probabilities of the label $Y = 1$ are at or above the given quantile thresholds. The BQR U-Net was trained with the loss function in eq. 8. For comparison, the deterministic U-Net was trained using the binary cross-entropy loss. The BQR U-Net was trained to output 0.125, 0.375, 0.625, 0.875 quantiles. These quantile values represent the centroids of the intervals between the test data quantiles (0, 0.25, 0.5, 0.75,1) representing 0-4 rater agreements respectively. We used these thresholds rather than the same quantiles as the test data to avoid operating at the boundary points between operators resulting from combining data from four raters only. To generate the corresponding estimated quantiles for the deterministic U-Net we thresholded the softmax probability at 0.125, 0.375, 0.625, 0.875.

Both deep BQR and deterministic U-Net diverged to the trivial solution of predicting all labels as zero due to the extreme class imbalance when initialized with a cold start. We therefore warmed up both models using a weighted cross-entropy loss for one epoch weighting samples by $1 \div 166$, the ratio of the zero to one labels in the training set. We trained both models for 5 epochs. Our results show no significant improvement for deep BQR compared to the cross-entropy loss in terms of Dice coefficients for different agreement areas. The Dice coefficients between estimated probability areas for deterministic U-Net and BQR U-Net are plotted in Figure 8 and summarized in Table 2. In the figure we show

Table 2: The mean (std dev) of the Dice coefficients between estimated probability regions ($P(Y = 1|X) \geq \alpha$, where α is $(0.25, 0.5, 0.75, 1)$; GT: ground truth, DT: deterministic, $P = P(Y = 1|X)$

	$P \geq 0.25$	$P \geq 0.5$	$P \geq 0.75$	$P = 1$
BQR U-Net vs GT	0.68(0.27)	0.60(0.34)	0.50(0.40)	0.27(0.35)
DT U-Net vs GT	0.67(0.27)	0.59(0.34)	0.50(0.38)	0.32(0.37)
DT U-Net vs BQR U-Net	0.81(0.27)	0.63(0.40)	0.40(0.43)	0.16(0.33)
Prob. U-Net vs GT	0.60(0.26)	0.60(0.39)	0.51(0.43)	0.31(0.36)

the distribution of the Dice coefficients between the BQR and ground truth quantiles, DT (deterministic U-Net) and ground truth, and also between DT and BQR. In some cases, particularly for higher quantiles, there was low agreement between human raters in the ground-truth test data. For example, 64 percent of the test data showed zero pixels in common between all four human raters, leaving the 0.875 quantile empty for those images. We therefore computed the Dice coefficients only over those regions in which the test data had a non-empty data set for that quantile. The results show reasonable agreement for the 0.125, 0.375 and 0.625 quantiles, but relatively poor results for the 0.875 quantile. Surprisingly, results for BRQ and the determinsic U-Net (DT U-Net) are very similar, even though the actual degree of overlap between the two is no better than between each of them and the ground truth labels. The fact that both methods perform poorly for the 0.875 quantile reflects both that the data are of relatively poor quality in this region and also that performance is likely limited by the imbalance in the training data between non-lesional areas and lesions confidently identified by all four raters. Here we used the deterministic U-Net as a backbone since it is arguably the most commonly used network for medical image segmentation task. Other networks could also be used as a backbone. To investigate whether we would expect further improvements using a probabilistic backbone, we also implemented the Probabilistic U-Net (Prob. U-Net) (Kohl et al., 2018) to capture rater uncertainty. Our results show that for the quantile estimation task, the performance of Prob. U-Net and U-Net are comparable. As a result, it appears unlikely that replacing the U-Net with its proabilistic form in the backbone would lead to significant improvements.

5. Conclusion

Quantile regression is a simple yet powerful method for estimating uncertainty both in supervised and unsupervised lesion detection. We proposed novel cost functions to apply quantile regression and capture confidence intervals for lesion segmentation.

In the unsupervised framework we used the VAE, a popular model for unsupervised lesion detection (Chen and Konukoglu, 2018; Baur et al., 2018; Pawlowski et al., 2018). VAEs can be used to estimate reconstruction probability instead of reconstruction error for anomaly detection tasks. For calculating reconstruction probability, VAE models the output as a conditionally independent Gaussian characterized by means and variances for each output dimension. Simultaneous estimation of the mean and the variance in VAE underestimates the true variance leading to instabilities in optimization (Skafte et al., 2019).

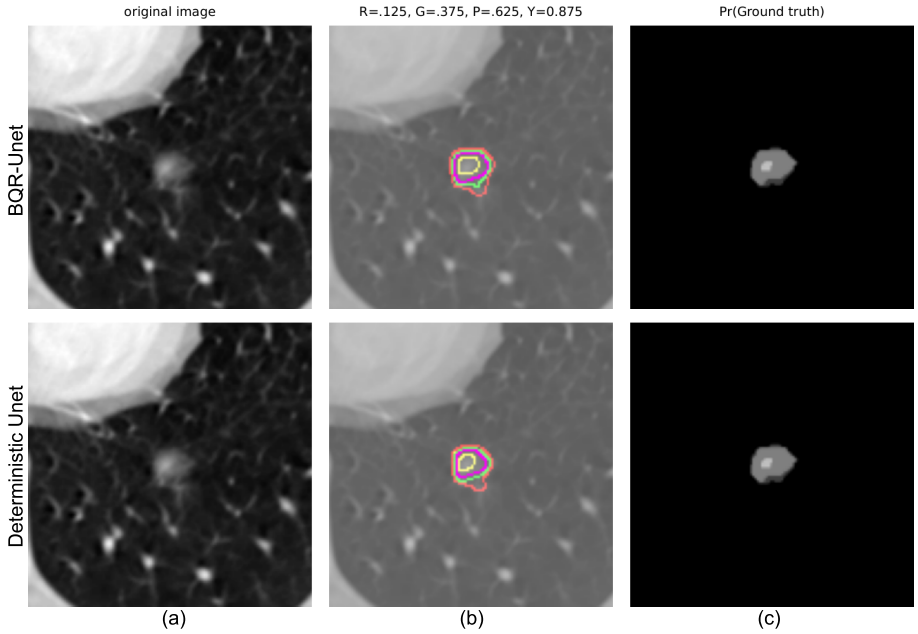

Figure 7: Top row: results of U-Net delineation of lesion boundaries. Bottom row: results of deterministic cross-entropy U-Net. (a) The original slice of the lung image; (b) estimated probability regions corresponding to 0.125,0.375,0.625,0.875 quantile levels shown with Red, green, purple and yellow colors respectively; (c) the estimate of thresholded lesion boundary from human raters corresponding to agreement between 1,2,3 and 4 raters.

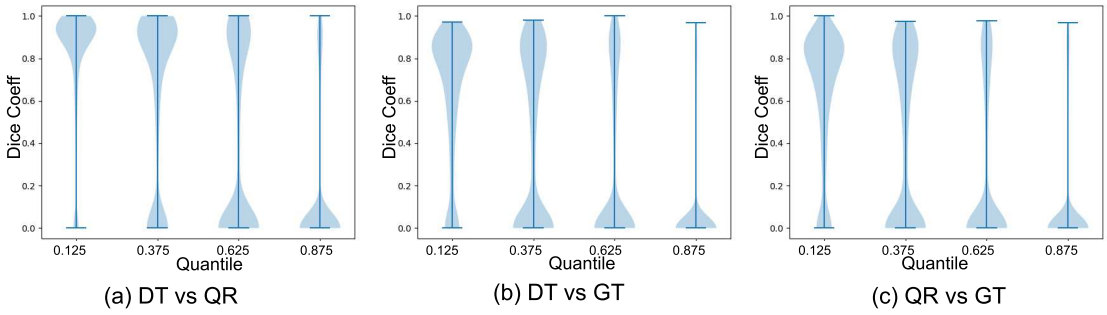

Figure 8: Violin plots of the Dice coefficients between quantiles $(0.125, 0.375, 0.625, 0.875)$) and rater agreement maps for the test datasets

, GT: ground truth, DT: binary cross-entropy (Deterministic U-Net), QR: quantile regression (BQR U-Net). The fraction of empty quantiles in the ground truth (excluded from Dice coefficient computations) were $0.07, 0.31, 0.45, 0.64$ respectively. The width of the violin indicates the fraction of the dataset as a function of the dice coefficient.

For this reason, classical VAE formulations that include both mean and variance estimates are rarely used in practice. Typically, only the mean is estimated with variance assumed constant (Skafte et al., 2019). To address this problem in variance estimation, we proposed an alternative quantile-regression model (QR-VAE) for improving the quality of variance estimation. We used quantile regression and leveraged the Guassian assumption to obtain the mean and variance by estimating two quantiles. We showed that our approach outperforms VAE as well as a Comb-VAE which is an alternative approach for addressing the same issue, in a synthetic as well as real world dataset. Our approach also has a more straightforward implementation compared to Comb-VAE. As a demonstrative application, we used our QR-VAE model to obtain a probabilistic heterogeneous threshold for a brain lesion detection task. This threshold results in a completely unsupervised lesion (or anomaly) detection method that avoids the need for a labeled validation dataset for principled selection of an appropriate threshold to control the false discovery rate. Beyond the current application we note that Quantile regression is applicable to deep learning models for medical imaging applications beyond the VAE and anomaly detection as the pinball loss function is easy to implement and optimize and can be applied to a broader range of network architectures and cost functions.

For supervised lesion detection, we present deep binary quantile regression to estimate label uncertainty. Specifically, we use this technique to estimate quantiles of the labels that represent uncertainty. The lesion segmentations generated for each quantile reflect this uncertainty. Using LIDC data with 4 annotations we aimed to estimate the disagreement between the annotators. Although it has been reported that deep binary QR has a better performance in imbalanced datasets in lesion segmentation task with extreme imbalance toward the class zero (normal), warming up the model was still needed in order to prevent it from converging to the trivial solution. Our result show no significant improvement in terms of dice coefficient between ground truth and estimated areas of agreement for deep binary QR compared to a deterministic U-Net. We found relatively small agreement for the 0.875 quantile region for these two estimations (row three, Table 2) demonstrating that although we obtain similar performance, these two estimators are not annotating the same region regions. This finding indicates the potential for further improvements in the performance of both methods. Based on the current results, while numerical results are similar, the fact that with fewer training samples QR is less likely to diverge than the deterministic U-Net indicates that the QR approach may be more robust and stable.

We investigated the advantages of using quantile regression in both supervised and unsupervised settings. In the unsupervised framework, the estimated confidence interval is used to capture uncertainty from which can identify outliers that represent our detected lesions. We demonstrated the advantage of using this quantile regression approach in the VAE setting. In the supervised framework, we used BQR to estimate the uncertainty of raters for the case where multi-rater data is available for training (and testing).

Acknowledgements

This work was supported by NIH grants R01 NS074980, R01 NS089212, and R01 EB026299, and by the DOD grant W81XWH-18-1-061.

References

Charu C Aggarwal. Outlier analysis. In *Data mining*, pages 237–263. Springer, 2015.

Haleh Akrami, Anand A Joshi, Jian Li, Sergul Aydore, and Richard M Leahy. Robust variational autoencoder. *arXiv preprint arXiv:1905.09961*, 2019.

Haleh Akrami, Anand A Joshi, Jian Li, Sergul Aydore, and Richard M Leahy. Brain lesion detection using a robust variational autoencoder and transfer learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 786–790. IEEE, 2020.

Haleh Akrami, Anand Joshi, Sergul Aydore, and Richard Leahy. Quantile regression for uncertainty estimation in vaes with applications to brain lesion detection. In *International Conference on Information Processing in Medical Imaging*, pages 689–700. Springer, 2021.

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Merlijn Blaauw and Jordi Bonada. Modeling and transforming speech using variational autoencoders. *Morgan N, editor. Interspeech 2016; 2016 Sep 8-12; San Francisco, CA.[place unknown]: ISCA; 2016. p. 1770-4.*, 2016.

J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170, 1995.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.

Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *International Conference on Information Processing in Medical Imaging*, pages 715–726. Springer, 2021.

Nicki S Detlefsen, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *arXiv preprint arXiv:1906.03260*, 2019.

Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018.

Rao P Gullapalli. Investigation of prognostic ability of novel imaging markers for traumatic brain injury (tbi). Technical report, BALTIMORE UNIV MD, 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *International Conference on Information Processing in Medical Imaging*, pages 677–688. Springer, 2021.

Onyedikachi O John. Robustness of quantile regression to outliers. *American Journal of Applied Mathematics and Statistics*, 3(2):86–88, 2015.

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

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

Simon AA Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus H Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *arXiv preprint arXiv:1806.05034*, 2018.

Gregory Kordas. Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407, 2006.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Oskar Maier, Bjoern H Menze, Janina von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. ISLES 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical image analysis*, 35:250–269, 2017.

Charles F Manski. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333, 1985.

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models. In *Advances in Neural Information Processing Systems*, pages 3855–3866, 2018.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *arXiv preprint arXiv:2006.06015*, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Man-Suk Oh, Eun Sug Park, and Beong-Soo So. Bayesian variable selection in binary quantile regression. *Statistics & Probability Letters*, 118:177–181, 2016.

Nick Pawlowski et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. *OpenReview*, 2018.

Jacob C Reinhold, Yufan He, Shizhong Han, Yunqiang Chen, Dashan Gao, Junghoon Lee, Jerry L Prince, and Aaron Carass. Validating uncertainty in medical image translation. *arXiv preprint arXiv:2002.04639*, 2020.

Filipe Rodrigues and Francisco C Pereira. Beyond expectation: deep joint mean and quantile regression for spatiotemporal problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1): 561–584, 1995.

Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. In *Advances in Neural Information Processing Systems*, pages 6323–6333, 2019.

Andrew Stirn and David A Knowles. Variational variance: Simple and reliable predictive variance parameterization. *arXiv preprint arXiv:2006.04910*, 2020.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *arXiv preprint arXiv:1811.00908*, 2018.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems*, pages 6417–6428, 2019.

John Wilder Tukey. The problem of multiple comparisons. *Multiple comparisons*, 1953.

Anna Volokitin, Ertunc Erdil, Neerav Karani, Kerem Can Tezcan, Xiaoran Chen, Luc Van Gool, and Ender Konukoglu. Modelling the distribution of 3d brain mri using a 2d slice vae. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 657–666. Springer, 2020.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

David Wingate and Theophane Weber. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.

Suhang You, Kerem C Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning*, pages 540–556. PMLR, 2019.

Keming Yu and Rana A Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.

John K Yue et al. Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury. *Journal of neurotrauma*, 30(22):1831–1844, 2013.