q-Space Novelty Detection with Variational Autoencoders

Aleksei Vasilev [1], Vladimir Golkov [1], Ilona Lipp [2], Eleonora Sgarlata [2,3],
Valentina Tomassini [2,4], Derek K. Jones [2], Daniel Cremers [1]

[1] Computer Vision Group, Technical University of Munich, Germany
[2] CUBRIC, Cardiff University, UK
[3] Department of Neurology and Psychiatry, Sapienza University of Rome, Italy
[4] Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK
`{alex.vasilev, vladimir.golkov, cremers}@tum.de`,
`elesgarlata@hotmail.it, {lippi, tomassiniv, jonesd27}@cardiff.ac.uk`

Abstract

In machine learning, novelty detection is the task of identifying novel unseen data. During training, only samples from the normal class are available. Test samples are classified as normal or abnormal by assignment of a novelty score. Here we propose novelty detection methods based on training variational autoencoders (VAEs) on normal data. Since abnormal samples are not used during training, we define novelty metrics based on the (partially complementary) assumptions that the VAE is less capable of reconstructing abnormal samples well; that abnormal samples more strongly violate the VAE regularizer; and that abnormal samples differ from normal samples not only in input-feature space, but also in the VAE latent space and VAE output. These approaches, combined with various possibilities of using (e.g. sampling) the probabilistic VAE to obtain scalar novelty scores, yield a large family of methods. We apply these methods to magnetic resonance imaging, namely to the detection of diffusion-space (q-space) abnormalities in diffusion MRI scans of multiple sclerosis patients, i.e. to detect multiple sclerosis lesions without using any lesion labels for training. Many of our methods outperform previously proposed q-space novelty detection methods.

1 Introduction

The purpose of novelty detection is to score how dissimilar each test sample is from a "normal" training set. This problem can also be seen as "one-class classification", where the model is trained to represent the normal class only. Application domains include medical diagnostic problems, fraud detection, and failure detection in industrial systems.

Novelty detection techniques can be classified into probabilistic, distance-based, reconstruction-based, domain-based, and information-theoretic [13]. We here focus on the first three categories.

The quality of novelty detection results depends on the algorithm and data distribution. Deep generative neural networks can be applied to reveal internal structure of the data and learn a better data representation. In this paper we design a set of novelty detection methods based on variational autoencoders (VAEs). We apply them to directly detect abnormalities such as multiple sclerosis lesions in diffusion magnetic resonance imaging (diffusion MRI). Non-deep novelty detection methods for diffusion space (q-space) have been used to tackle this problem [4; 5], and achieved promising AUC scores between 0.82 and 0.89 in multiple sclerosis lesion segmentation on various datasets. We compare the proposed VAE-based novelty detection methods with the original q-space novelty detection, and show that some of them are able to perform better.

1.1 Related work on novelty detection with generative models

With a dramatic increase of research in deep generative models during the last years, several new methods for novelty detection were proposed. These methods try to learn the normal patterns in the data using variational autoencoders, adversarial autoencoders [11], or generative adversarial networks (GANs) [2].

In [1] the VAE trained on a normal class only is used to detect abnormalities in the test data. The novelty score of the test sample is computed in the following way: the encoder of the VAE infers the distribution of the test sample in the latent space; several samples from this distribution are passed to the stochastic decoder of the VAE to infer a number of distributions of the test sample in the original feature space; these distributions are used to compute the average probability density value of the test sample, which is used as a novelty score.

A novelty score metric based on the adversarial autoencoder network is proposed in [10]. The model is trained to match the posterior distribution of the normal data in a latent space with a chosen prior distribution. The novelty metric in this case is based on the likelihood of the test sample according to the prior: low likelihood indicates that the test sample unlikely belongs to the normal data class.

Another approach [18], uses a framework consisting of an autoencoder network and a network that estimates a Gaussian mixture model of a normal class distribution in the latent space. For the test sample, both the reconstruction error of the autoencoder and the likelihood of this sample according to the estimated mixture model are used to define abnormality.

In [15; 16], a GAN is trained to learn the distribution of the normal data. During test time a search over the latent space is performed to find the closest generated sample to the test sample. Then this closest generated sample is fed into the discriminator and the overall novelty score is computed as a weighted sum of the discriminator loss and of the difference between the input and the generated sample.

Another line of work [17] tries to unify classification and novelty detection into a single framework. The main concept is to train a separate class-conditional generative model (VAE or GAN) for each normal class in the training dataset. During test time, for each of these generative models a search over latent space is performed to find the distance between the test sample and the closest generated sample. The minimal distance across all class-conditional generators is a novelty score. If this score is below a certain threshold, then the test sample is classified with the class of the corresponding class-conditional generator. One of the methods we employ utilizes this idea for novelty detection, using *only one* generative model for all the normal data, see Section 2.2.3.

A GAN with a mixture generator is used for novelty detection in [9]. In contrast to most novelty detection approaches, here the model is trained on both normal and abnormal data.

1.2 Diffusion MRI

Diffusion MRI is a magnetic resonance imaging (MRI) technique that uses the diffusion of water molecules to generate contrast in MR images. Since this diffusion is not free and affected by obstacles, it can reveal microstructural details about the tissue. For each of several chosen diffusion-space (q-space) coordinates, a diffusion-weighted 3D image is acquired.

Classical data processing in diffusion MRI Traditional diffusion MRI processing methods fit a handcrafted mathematical or physical model/representation to the measurements, and interpret the estimated model parameters. However, these approaches have several limitations, since the model-fitting procedure is ill-conditioned, and interpreting its results requires prior knowledge about how disease-related microstructural changes affect the model parameters.

Supervised and weakly-supervised deep learning in diffusion MRI Recent research shows that deep learning can overcome said issues by learning a direct mapping between q-space measurements and diseases [3; 6]. In deep learning terminology, each diffusion-weighted 3D image corresponding to a certain q-space coordinate is treated as a *"channel"* of the overall multi-channel 3D volume. For *voxel-wise supervised learning*, for example to reconstruct missing q-space measurements from existing ones, or to predict handcrafted-model-based parameters more robustly and at a shorter scan time, or to directly estimate tissue types and properties, a "voxels-to-voxels" convolutional network can be used [3]. It can consist purely of convolutional layers with filter size $1 \times 1 \times 1$ if the risk

of spatial bias (overfitting) should be excluded [3]. On the other hand, *global supervised learning* (i.e. image-wise rather than voxel-wise prediction) and *voxel-wise weakly-supervised learning* can be performed with a convolutional network that reduces spatial resolution using pooling and/or fully-connected layers [6]. However, supervised and weakly-supervised disease detection requires disease-specific labels.

Novelty detection in diffusion MRI The aforementioned methods are complemented by q-space novelty detection methods, which do not require disease-related labels [4; 5]. In this line of work, each voxel is treated as a separate d-dimensional feature vector, where d is the number of measured diffusion directions. Voxels from scans of healthy volunteers are used as a reference dataset, and the Euclidean distance in feature space between the test datapoint and its nearest neighbor from the reference dataset is used as novelty score. A high novelty score thus indicates that the voxel is lesioned. This novelty score coincides with multiple sclerosis lesions at AUC scores between 0.82 and 0.89 on various datasets, and we use this method as a baseline to compare our methods with.

1.3   Our contributions

In this paper we show that the usage of a variational autoencoder can help to better understand the normal patterns in the data and thus improve the quality of novelty detection. We further explore the possibilities of applying novelty detection to diffusion MRI processing. The main contributions of the paper can be summarized as follows:

- We propose several new novelty detection methods in the VAE original and latent feature spaces. These methods can be applied to different novelty detection tasks. Our code is publicly available at `https://github.com/VAlex22/ND_VAE`

- We adopt the VAE network to q-space novelty detection. We show that this solution can beat the performance of the original q-space novelty detection algorithm.

2   VAE-based Novelty Detection Methods

A variatioal autoencoder is a deep neural network that models a relationship between a low-dimensional latent random variable z and a random variable x in the original data space. A VAE consists of an encoder and a decoder. The encoder is trained to approximate the posterior distribution $q_\phi(z|x)$, where ϕ are the network parameters, learned with backpropagation during training. The decoder performs the inverse task: it approximates the posterior distribution $p_\theta(x|z)$, where θ are learned parameters. In contrast to the traditional autoencoder, the VAE has a stochastic encoder and decoder: their outputs for a given input are not the variables z and x, but the parameters of the distributions of z and x.

The goal of the VAE is to learn a good approximation $q_\phi(z|x)$ of the intractable true posterior $p_\theta(z|x)$. The quality of the approximation can be evaluated with the Kullback–Leibler divergence: $D_{\mathrm{KL}}\big(q_\phi(z|x) \parallel p_\theta(z|x)\big)$. This divergence cannot be computed directly [7; 14], but it can be minimized by maximizing the sum of the Evidence Lower Bound (ELBO) on the marginal likelihood of datapoints x_i: $\mathrm{ELBO}_i = E_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] - D_{\mathrm{KL}}\big(q_\phi(z|x_i) \parallel p(z)\big)$, where $p(z)$ is a prior distribution of z which is usually chosen as a unit Gaussian. The loss function of the VAE is then:

$$\mathcal{L}_{\mathrm{VAE}} = -\sum_i \mathrm{ELBO}_i = -\sum_i \big[E_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] - D_{\mathrm{KL}}\big(q_\phi(z|x_i) \parallel p(z)\big)\big], \quad (1)$$

where the sum is calculated over all training data samples x_i.

From the machine learning perspective, the first term in Eq. (1) can be considered as a reconstruction error forcing the model to reconstruct the input and the second term is a regularizer that prevents the model from giving each input a representation in a different region of the latent space.

In our case we train the VAE model to capture normal data only. Thereby, the VAE learns distributions of the normal dataset in latent and original feature space. Both of these distributions as well as their combination can be used to define novelty score metrics, thus we split our methods into three main categories. Novelty detection methods in latent space and original feature space are illustrated in Fig. 1.
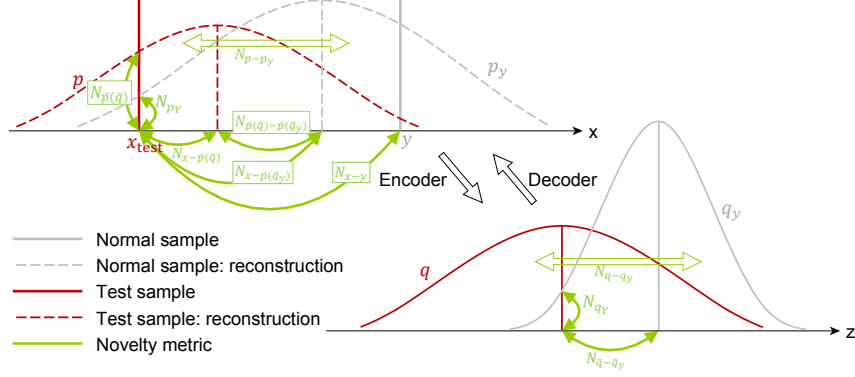
Figure 1: Proposed novelty detection methods for the simplified case of 1D original space x and 1D latent space z of the VAE with only one normal sample y and one test sample x_{test}. Arrows measure distances between points ($\hookrightarrow$) or distributions ($\Leftrightarrow$). The scores $N_{x-\bar{p}(\bar{q})}$ and $N_{\bar{p}(\bar{q})}$ quantify whether the VAE reconstructs x_{test} badly; N_{x-y}, $N_{x-\bar{p}(\bar{q}_y)}$ and $N_{\bar{p}(\bar{q})-\bar{p}(\bar{q}_y)}$ measure how dissimilar x_{test} (or its reconstruction "$\bar{p}(\bar{q})$") is from normal samples y (or their reconstruction "$\bar{p}(\bar{q}_y)$"); N_{q-q_y} and $N_{\bar{q}-\bar{q}_y}$ measure how dissimilar x_{test} is from y in latent space; N_{p_Y} and N_{q_Y} measure how likely x_{test} belongs to the modeled distribution of normal data. In this illustration, methods that use stochastic estimates "$\hat{q}$" or "$\hat{p}$" correspond to the methods using deterministic means "$\bar{q}$" or "$\bar{p}$". For example, $N_{x-\bar{p}(\bar{q}_y)}$, $N_{x-\bar{p}(\hat{q}_y)}$, $N_{x-\hat{p}(\bar{q}_y)}$, $N_{x-\hat{p}(\hat{q}_y)}$ all correspond to the same green arrow, but use distinct estimation methods.

2.1 Novelty in the latent space

The trained VAE maps each sample x to a distribution z in some lower-dimensional latent space. This representation can be used to define several novelty metrics.

2.1.1 Novelty as VAE regularizer

The VAE loss function includes the following regularizer term: $D_{\text{KL}}\big(q_\phi(z|x) \,\|\, \mathcal{N}(0, I)\big)$. This term forces the model to map inputs closely to the unit Gaussian distribution $\mathcal{N}(0, I)$ in the latent space. Without this regularizer the model could learn to give each input a representation in a different region of the latent space.

For the model trained on the normal class only, one could expect that abnormal samples will have distributions in the latent space that diverge more from the unit Gaussian than normal samples. We thus can use this term as a novelty score for test sample x_{test}:

$$N_{\text{VAE-reg}}(x_{\text{test}}) = D_{\text{KL}}\big(q_\phi(z|x_{\text{test}}) \,\|\, \mathcal{N}(0, I)\big). \tag{2}$$

2.1.2 Distance-based approaches in latent space

The latent space of a VAE trained on the normal class can be considered an effective representation of the distribution of normal data. Therefore, classical novelty detection approaches can be adapted to be used in this space. The algorithm here is to construct the reference dataset by capturing normal data in the latent space. The latent representation of each test sample should also be inferred. Then we can use nearest neighbour analysis to find the closest sample from the reference dataset to the test sample in a latent space using some distance measure. The distance to this closest sample will be a novelty score. A VAE maps each input point to a distribution (rather than a point) in latent space. We propose two distance measures:

1. **Euclidean distance between means of the distributions.** This approach uses only information about the means of the approximated posterior in the latent space for both normal and test datapoints. The novelty score in this case is computed as the distance between between the mean of the latent-space distribution of the test datapoint and the closest latent-space distribution mean of a normal sample:

$$N_{\bar{q}-\bar{q}_y}(x_{\text{test}}) = \min_{y \in Y} \|E[q_\phi(z|x_{\text{test}})] - E[q_\phi(z|y)]\|_2^2, \tag{3}$$

4

where the minimum is taken over all normal samples y from the normal dataset Y.

2. **Bhattacharyya distance between distributions.** The Bhattacharyya distance is a symmetric measure of dissimilarity of two probability distributions p and q. It is defined as $D_B(p, q) = -\ln(\text{BC}(p, q))$, where $\text{BC}(p, q) = \int \sqrt{p(z)q(z)}\, dz$ is the Bhattacharyya coefficient of distributions p and q. This approach utilizes information about the full learned distributions, computing the amount of the overlap between them. The proposed novelty score is defined as the Bhattacharyya distance between the latent-space distribution $q_\phi(z|x_{\text{test}})$ of the test sample and the most similar latent-space distribution $q_\phi(z|y)$ of a normal sample:

$$N_{q-q_y}(x_{\text{test}}) = \min_{y \in Y} D_B(q_\phi(z|x_{\text{test}}), q_\phi(z|y)). \tag{4}$$

2.1.3 Density-based approach in latent space

Another approach to novelty detection is to estimate the density of normal data in the latent space. Each normal datapoint is represented as a Gaussian distribution in the VAE latent space. Thus the distribution of the whole normal dataset can be estimated as an average of these Gaussians: $q_Y(z) = \frac{1}{|Y|} \sum_{y \in Y} q_\phi(z|y)$. Then, the novelty score for the test sample can be computed from the density estimate q_Y of the normal dataset, evaluated at the mean of the latent distribution $q_\phi(z|x_{\text{test}})$ of the test sample (see also Fig. 1):

$$N_{q_Y}(x_{\text{test}}) = -q_Y(E[q_\phi(z|x_{\text{test}})]). \tag{5}$$

2.2 Novelty in the original feature space

2.2.1 VAE reconstruction-based approaches

Like a traditional autoencoder network, a VAE is trained to reconstruct the input. If a VAE is trained on the normal class only, one could expect that it learns how to reconstruct data from the normal class well, but may not be particularly good at reconstructing unseen data, so the reconstruction error for the unseen class should be higher. Thus, the reconstruction error can be used as a novelty score.

The encoder and decoder are both stochastic. For each of them, we can either draw samples, or consider the mean. For the decoder it is also possible to consider the entire distribution and compute the density value of the test sample. We propose six possible reconstruction-based novelty scores:

1. **Deterministic reconstruction error**: Use means of both the encoder $q_\phi(z|x)$ and the decoder $p_\theta(x|z)$ to deterministically reconstruct the input and compute the reconstruction error:

$$N_{x-\bar{p}(\bar{q})}(x_{\text{test}}) = \left\| x_{\text{test}} - E\big[p_\theta\big(x|E[q(z|x_{\text{test}})]\big)\big] \right\|_2^2. \tag{6}$$

2. **Deterministic reconstruction likelihood**: Compute log likelihood of the input given the mean of the encoder:

$$N_{\bar{p}(\bar{q})}(x_{\text{test}}) = -\log p_\theta\big(x_{\text{test}}|E[q(z|x_{\text{test}})]\big). \tag{7}$$

3. **Encoder-stochastic reconstruction error**: Use samples from the encoder and mean of the decoder to compute several possible reconstructions of the input and calculate the average reconstruction error:

$$N_{x-\bar{p}(\hat{q})}(x_{\text{test}}) = \operatorname*{mean}_{z_i \sim q_\phi(z|x_{\text{test}})} \left\| x_{\text{test}} - E[p_\theta(x|z_i)] \right\|_2^2. \tag{8}$$

4. **Encoder-stochastic reconstruction likelihood**: Compute several possible log likelihood function values of the input given samples from the encoder:

$$N_{\bar{p}(\hat{q})}(x_{\text{test}}) = \operatorname*{mean}_{z_i \sim q_\phi(z|x_{\text{test}})} -\log p_\theta(x_{\text{test}}|z_i). \tag{9}$$

5. **Decoder-stochastic reconstruction error**: Use mean of the encoder and samples from the decoder to compute average reconstruction error:

$$N_{x-\hat{p}(\bar{q})}(x_{\text{test}}) = \operatorname*{mean}_{x_i \sim p_\theta(x|E[q_\phi(z|x_{\text{test}})])} \left\| x_{\text{test}} - x_i \right\|_2^2. \tag{10}$$

6. **Fully-stochastic reconstruction error**: Use samples from both the encoder and the decoder and compute average reconstruction error:

$$N_{x-\hat{p}(\hat{q})}(x_{\text{test}}) = \operatorname*{mean}_{\substack{x_i \sim p_\theta(x|z_i) \\ z_i \sim q_\phi(z|x_{\text{test}})}} \|x_{\text{test}} - x_i\|_2^2. \tag{11}$$

The novelty score defined in Eq. (9) is equal to "reconstruction probability" proposed by [1].

In metrics (8), (9), (10) and (11) the average operation can be replaced by the min operation, producing four more novelty metrics.

2.2.2 Distance- and density-based approaches

In addition to reconstruction-based approaches, it is possible to apply distance- and density-based approaches described in Sections 2.1.2 and 2.1.3 to the distributions produced by the decoder of the VAE. Therefore we can get three more novelty metrics:

$$N_{\bar{p}(\bar{q})-\bar{p}(\bar{q}_y)}(x_{\text{test}}) = \min_{y \in Y} \Big\| \underbrace{E\big[p_\theta\big(x|E[q_\phi(z|x_{\text{test}})]\big)\big]}_{\text{Deterministic reconstruction of } x_{\text{test}}} - \underbrace{E\big[p_\theta\big(x|E[q_\phi(z|y)]\big)\big]}_{\text{Deterministic reconstruction of } y} \Big\|_2^2, \tag{12}$$

$$N_{p(\bar{q})-p(\bar{q}_y)}(x_{\text{test}}) = \min_{y \in Y} D_B\Big(p_\theta\big(x|E[q_\phi(z|x_{\text{test}})]\big), p_\theta\big(x|E[q_\phi(z|y)]\big) \Big), \tag{13}$$

$$N_{p(\bar{q}_Y)}(x_{\text{test}}) = -p_Y(E[p_\theta(x|E[q_\phi(z|x_{\text{test}})])]), \tag{14}$$

where $p_Y(x) = \frac{1}{|Y|} \sum_{y \in Y} p_\theta(x|E[q_\phi(z|y)])$ is the average of the distributions reconstructed from the normal-class samples in the original feature space.

Metrics in Eqs. (12) and (14) can also be computed using the original test sample instead of its deterministic reconstruction produced by the VAE. Thus, two additional novelty scores are possible:

$$N_{x-\bar{p}(\bar{q}_y)}(x_{\text{test}}) = \min_{y \in Y} \Big\| x_{\text{test}} - E\big[p_\theta\big(x|E[q_\phi(z|y)]\big)\big] \Big\|_2^2, \tag{15}$$

$$N_{p(\bar{q}_Y)(x)}(x_{\text{test}}) = -p_Y(x_{\text{test}}). \tag{16}$$

It is also possible to apply the Euclidean distance-based approach to the reconstructed test sample and original (not reconstructed) normal datapoints:

$$N_{\bar{p}(\bar{q}-y)}(x_{\text{test}}) = \min_{y \in Y} \Big\| E\big[p_\theta\big(x|E[q_\phi(z|x_{\text{test}}) - y]\big)\big] \Big\|_2^2. \tag{17}$$

2.2.3 Distance to the closest generated sample

A VAE is a generative model. Trained on a normal class only, one could expect that it will not be able to generate the datapoints that do not belong to the normal class. Thus a novelty score can be computed as a distance between an input and the closest sample that the VAE decoder is able to produce from any latent vector z:

$$N_{x-\hat{p}}(x_{\text{test}}) = \min_z \|x_{\text{test}} - E[p_\theta(x|z)]\|_2^2. \tag{18}$$

This is an optimization problem over the latent space. It can be solved using a non-linear optimization method such as L-BFGS [12]. The encoder of the VAE with x_{test} as an input can be used to get an initial value for z for optimization.

If the bottleneck of the VAE is not narrow enough, the VAE may still be able to reconstruct abnormal data. However, the values of z for abnormal datapoints might be far from the unit Gaussian, since the VAE loss includes the term $D_{\text{KL}}\big(q_\phi(z|x) \,\|\, \mathcal{N}(0, I)\big)$, and is trained on the normal data only. Thus the optimization can also be performed not over the whole latent space, but within some boundaries, defining a new novelty score metric $N_{x-\hat{p}_b}$. During the experiments, we found that the best results are achieved with $[-10, 10]$ boundaries for each of the dimensions of the latent space.

Table 1: AUC scores of multiple sclerosis lesion segmentation for different q-space novelty detection methods. Many of our methods are at par with or outperform (marked in **bold**) the original method N_{x-y}.

Method	dim z	AUC scores		
		scan 1	scan 2	scan 3
Baseline (N_{x-y})	–	0.859	0.838	0.884
$N_{\text{VAE-reg}}$	16	0.803	0.773	0.812
$N_{\bar{q}-\bar{q}_y}$	16	**0.893**	0.831	**0.890**
N_{q-q_y}	16	**0.888**	0.835	**0.893**
N_{q_Y}	24	**0.860**	0.815	0.880
$N_{x-\bar{p}(\bar{q})}$	24	0.857	**0.841**	0.882
$N_{\bar{p}(\bar{q})}$	16	0.843	**0.860**	0.865
$N_{x-\bar{p}(\hat{q})}$	24	0.857	0.826	0.860
$N_{\bar{p}(\hat{q})}$	16	0.850	**0.860**	0.865
$N_{x-\hat{p}(\bar{q})}$	24	0.851	**0.846**	0.881
$N_{x-\hat{p}(\hat{q})}$	16	0.854	0.826	0.859
$N_{\bar{p}(\bar{q})-\bar{p}(\bar{q}_y)}$	24	**0.895**	0.829	**0.891**
$N_{\bar{p}(\bar{q})-y}$	16	**0.893**	0.852	**0.897**
$N_{p(\bar{q}_Y)}$	16	0.753	0.587	0.649
$N_{x-\bar{p}(\bar{q}_y)}$	16	0.830	0.811	0.857
$N_{p(\bar{q}_Y)(x)}$	16	0.760	0.600	0.652
$N_{x-\hat{p}}$	8	0.857	**0.849**	0.878
$N_{x-\hat{p}_b}$	8	**0.859**	**0.849**	0.879
$N_{-\text{ELBO}}$	16	**0.895**	**0.856**	0.852
$N_{-\widehat{\text{ELBO}}}$	16	**0.897**	**0.867**	0.873

2.2.4 Novelty as additive inverse of highest generated likelihood

Another possibility is to not use Euclidean distance between the test sample and deterministic encoder output, but to compute the density value of $p_\theta(x_{\text{test}}|z)$. The novelty score in this case is equal to

$$N_{\hat{p}}(x_{\text{test}}) = \min_z -\log p_\theta(x_{\text{test}}|z). \tag{19}$$

In comparison to the $N_{\bar{p}(\hat{q})}$ metric (Eq. (9)), where z is sampled from the approximated posterior $q_\phi(z|x_{\text{test}})$, here z can take any value, or be restricted to lie within some boundaries for the same reasons as for the $N_{x-\hat{p}}$ score, Eq. (18).

2.3 Novelty as full VAE loss

If the VAE is trained on the "normal" class only, its loss function (Eq. (1)) can be considered as a novelty metric itself. From a machine learning perspective, novel data that were not available during training will cause a high value of the model's loss function. From a probabilistic perspective, the VAE loss function value from datapoint x_i is a reverse of the lower bound of the marginal probability $p(x_i)$ of this datapoint [7; 14]. Novel datapoints have a low probability according to the model trained on a "normal" class and thus high loss value. Thus we propose a novelty metric:

$$N_{-\text{ELBO}}(x_{\text{test}}) = L_{\text{VAE}} = -E_{q_\phi(z|x_{\text{test}})}[\log p_\theta(x_{\text{test}}|z)] + D_{\text{KL}}\big(q_\phi(z|x_{\text{test}}) \,\|\, \mathcal{N}(0,I)\big). \tag{20}$$

Like in the original VAE training algorithm [7], the Monte Carlo estimate of $E_{q_\phi(z|x_{\text{test}})}$ should be computed. We also propose to use several samples from the probabilistic decoder $q_\phi(z|x_{\text{test}})$ and define another novelty metric:

$$N_{-\widehat{\text{ELBO}}}(x_{\text{test}}) = \min_{z \sim q_\phi(z|x_{\text{test}})} \big[\log p_\theta(x_{\text{test}}|z)\big] + D_{\text{KL}}\big(q_\phi(z|x_{\text{test}}) \,\|\, \mathcal{N}(0,I)\big). \tag{21}$$

3 Experiments and Discussion

In this section we evaluate the proposed methods on multiple sclerosis lesion detection, and compare them with the baseline q-space novelty detection.
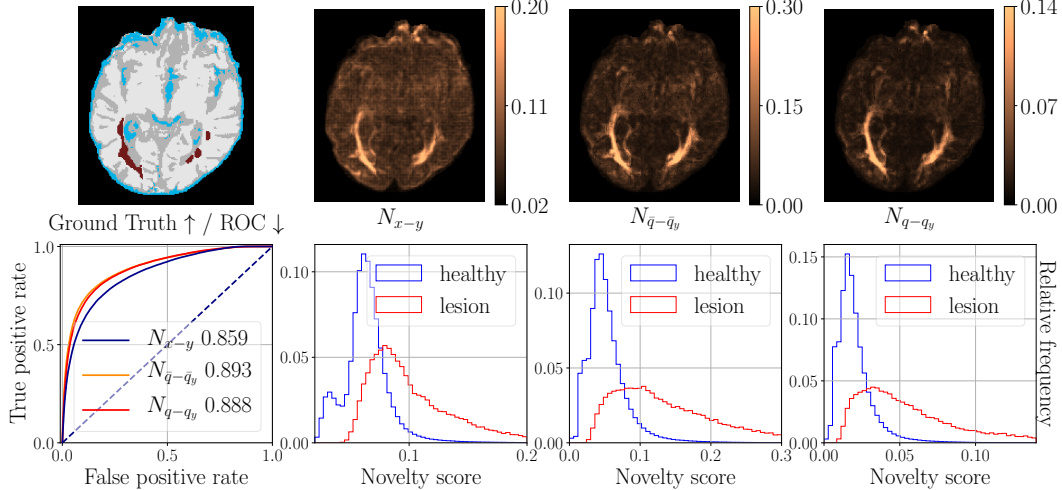
Figure 2: Feasibility of q-space novelty detection for multiple sclerosis lesion segmentation. Top row: manual lesion segmentation; novelty scores using the baseline method N_{x-y} and two VAE latent space distance-based novelty detection methods $N_{\bar{q}-\bar{q}_y}$ and N_{q-q_y}. Bottom row: ROC for one scan measuring coincidence of novelty detection with human-marked labels: proposed methods outperform the baseline; normalized histogram of the novelty score for lesion and non-lesion voxels for the methods above: many lesion voxels have considerably higher novelty scores than healthy voxels. In other words, disease-related microstructural tissue changes are detected in a data-driven way, without any prior knowledge about them.

We used the following data: As the the "normal" data to train the model on, we used 26 diffusion MRI scans of healthy volunteers. Each scan has six $b = 0$ images and 40 uniformly distributed diffusion directions ($b_{max} = 1200s/mm^2$, SE-EPI, voxel size 1.8mm × 1.8mm × 2.4mm, matrix 128 × 128, 57 slices, $T_E = 94.5ms$, $T_R = 16s$, distortion-corrected with elastix [8] and upsampled to 256 × 256 × 172). Using machine learning nomenclature, we refer to these 46 volumes (six $b = 0$ and 40 diffusion-weighted volumes) as *channels* or *voxel-wise features*. Test data consisted of three multiple sclerosis patients diffusion MRI scans with the same scan parameters as for healthy volunteers. To validate the results of proposed methods we used multiple sclerosis lesion labels created by human raters using additional structural T_2-weighted scans. As a performance metric for segmentation, we selected the AUC score. We compared the performance of proposed novelty detection algorithms with the distance-based q-space novelty detection algorithm [4; 5] (described in Section 1.2).

As a normalization to avoid potential discrepancies in image intensity across scans, each scan was divided by its mean intensity. In addition, to prevent some channels dominating over others, feature scaling was performed, i.e. each channel was divided by the channel-wise mean taken across all scans. Each voxel in the diffusion MRI scan was considered as a separate data sample. Thus the local neighborhood information was ignored and only q-space measurements were used to perform "voxel-to-voxel" inference. Trained in this way, the VAE is able to produce the latent-space and the original feature space distributions of the input data sample. We used a symmetric architecture, where the encoder and the decoder have the same number of hidden layers and the same number of units in each layer. We split the normal dataset into a training and validation set, and used early stopping to prevent overfitting.

Table 1 shows AUC scores of the best-performing methods. For each of the novelty metrics described in Section 2 we performed a hyperparameter search varying the dimensionality of the latent space and the number of hidden layers to find the model that achieves the highest AUC score. The models with three hidden layers in both the encoder and the decoder perform better than others, so the only architecture difference between the VAEs presented in the Table 1 is in the dimensionality of the latent space.

The results show that the Euclidean-distance-based method applied to the latent space (N_{q-q_y} and $N_{\bar{q}-\bar{q}_y}$) can outperform the same method applied to the original data (N_{x-y}; see also Figure 2),

despite the fact that the model was not trained on abnormal data. This happens on one hand due to the fact that the VAE has a regularizer that keeps the latent representation of the normal data tightly clustered around the unit Gaussian, but at the same time, abnormal data that was not used during training can be mapped in a completely different region of the latent space (which is confirmed by the fact that $N_{\text{VAE-reg}}$ also produces meaningful novelty scores); and on the other hand due to the fact that the trained VAE simply happens to map abnormal data to slightly different latent-space regions.

Reconstruction-based methods also perform well, since the reconstruction of abnormal data (of which the model has not seen any during training) can be very imprecise. However, in order to be able to reconstruct the normal data, a VAE should have a sufficient number of latent dimensions to avoid significant information loss. The results of the methods $N_{x-\hat{p}}$ and $N_{x-\hat{p}_b}$ show the generative ability of the VAE. High AUC scores prove that the model is able to generate only the normal data it used during training. However, to achieve this ability, the bottleneck of the VAE (in contrast to reconstruction-based methods) should be narrow enough, otherwise the model may also be able to generate some "random" data that may be very close to the abnormal data in the original feature space. Finally, the VAE loss itself as the inverse of the lower bound of the likelihood of the data sample is a good novelty metric: abnormal data (none of which were used during training) will have low likelihood according to the model and thus high novelty score.

4 Conclusions

In this work, we presented a set of novelty detection algorithms that utilize the ability of variational autoencoders to reveal the internal structure of the normal class data. We found that high novelty scores produced by our methods coincide with multiple sclerosis lesions in diffusion MRI data, and many of our methods outperform the baseline q-space novelty detection method. However, the proposed methods are also applicable to other novelty detection problems, since they do not rely on any diffusion-MRI-specific properties.

References

[1] J. An and S. Cho. *Variational Autoencoder based Anomaly Detection using Reconstruction Probability.* Technical Report. SNU Data Mining Center, 2015.

[2] I.J. Goodfellow et al. *Generative Adversarial Nets* NIPS, 2014.

[3] V. Golkov et al. *q-Space Deep Learning: Twelve-fold Shorter and Model-Free Diffusion MRI Scans.* IEEE Trans Med Imag, 35(5):1344-1351, 2016.

[4] V. Golkov et al. *Model-Free Novelty-Based Diffusion MRI.* ISBI 1233-1236, 2016.

[5] V. Golkov et al. *q-Space Novelty Detection in Short Diffusion MRI Scans of Multiple Sclerosis.* ISMRM 5378, 2018.

[6] V. Golkov et al. *q-Space Deep Learning for Alzheimer's Disease Diagnosis: Global Prediction and Weakly-Supervised Localization.* ISMRM 1580, 2018.

[7] D.P. Kingma and M. Welling. *Auto-Encoding Variational Bayes.* ICLR, 2014.

[8] S. Klein et al. *elastix: A Toolbox for Intensity-based Medical Image Registration.* IEEE Trans Med Imag. 29(1):196–205, 2010.

[9] M. Kliger and S. Fleishman. *Novelty Detection with GAN.* arXiv:1802.10560, 2018.

[10] V. Leveau and A. Joly. *Adversarial Autoencoders for Novelty Detection.* Research Report. Inria - Sophia Antipolis, 2017.

[11] A. Makhzani et al. *Adversarial Autoencoders.* ICLR, 2016

[12] J. Nocedal. *Updating Quasi-Newton Matrices with Limited Storage.* Mathematics of computation 35(151):773–782, 1980.

[13] M.A. Pimentel, et al. *A Review of Novelty Detection.* Signal Processing 99:215-249, 2018.

[14] D.J. Rezende et al. *Stochastic Backpropagation and Approximate Inference in Deep Generative Models.* ICML 32(2):1278-1286, 2014.

[15] T. Schlegl et al. *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery.* IPMI, 2017.

[16] H. Wang et al. *Generative Adversarial Network Based Novelty Detection using Minimized Reconstruction Error.* Frontiers Inf Technol Electronic Eng 19:116, 2018.

[17] W. Wang et al. *Safer Classification by Synthesis.* arXiv:1711.08534, 2017.

[18] B. Zong et al. *Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection.* ICLR, 2018.