IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Predictable Uncertainty-Aware Unsupervised Deep Anomaly Segmentation

Kazuki Sato, Kenta Hama, Takashi Matsubara, and Kuniaki Uehara
Graduate School of System Informatics, Kobe University,
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan.
Emails:{ksato@ai.cs., hamaken@ai.cs., matsubara@phoenix., uehara@}kobe-u.ac.jp

***Abstract*—Image-based anomaly segmentation is a fundamental topic for image analysis. For medical use, it supports treatments via refined diagnosis and growth rate evaluation of tumors and lesions. Especially, an unsupervised training is expected to generalize to unknown anomalies. Probabilistic models have been used for this purpose, whereby these models are trained to maximize the likelihood of known samples and detect anomalous samples by assigning low likelihoods. Recent studies have proposed a probabilistic model based on deep neural networks (DNNs) called AEs and they achieved significant performance thanks to their flexibility. However, AEs are sensitive to complex structure (e.g., ridges and grooves of a brain) rather than semantic anomalies (e.g., tumors and lesions). We decomposed the approximated log-likelihood into two terms; predictable uncertainty and normalized error. We found that the former represents the complexity of structure. Hence, we propose the normalized error as a novel uncertainty-sensitive score by removing the predictable uncertainty. We evaluated our score by experiments with head magnetic resonance imaging (MRI) datasets and demonstrate the robustness of the proposed normalized error to data complexity.**

***Index Terms*—deep learning, anomaly segmentation, unsupervised learning, uncertainty**

I. Introduction

Image-based anomaly segmentation is a fundamental topic for image analysis. For industrial use, anomaly segmentation detects subregions of products not coincident with defined standards such as cracks and deformations. Automatic detection of anomalous subregions reduces inspection costs and improves product reliability [1]. For medical use, the segmentation of anomalies such as tumors and stroke lesions in brains supports treatments via refined diagnosis and growth rate evaluation [2]–[4]. While supervised methods have achieved significant results for image segmentation tasks [5], they require a huge number of training samples labeled in a pixel-by-pixel manner for each known type of anomalies and often fail in detecting other anomaly types with different appearances. Here, a practical demand of unsupervised anomaly segmentation arises.

For unsupervised detection of sample-wise anomaly, an anomaly has often been defined as an outlier, whereby a machine learning method is trained to model the distribution or density of normal training samples and detects outlying test samples as anomalies (see [6] for a survey). For example, a probabilistic model $p_\theta(x_1, x_2, \dots)$ of pixels $\{x_1, x_2, \dots\}$ of an image x is trained to maximize the likelihoods of the given samples $x \in \mathcal{X}$ being generated from the probabilistic model $p_\theta(x)$. Since the probabilistic model $p_\theta(x)$ assigns high likelihoods to samples close to known samples, one can consider a pixel x of lower likelihoods $p_\theta(x)$ as an anomaly. A typical probabilistic model is a Gaussian mixture model (GMM). A GMM is highly sensitive to acceptable outliers due to the Gaussian base distributions and hence needs a modification of the base distributions for better results [7]. For unsupervised anomaly segmentation (i.e., detection of pixel-wise anomaly), the distribution is defined for each pixel individually or given other pixels (e.g., $p_\theta(x_i|x_{\setminus i})$ where $x_{\setminus i} = x \setminus \{x_i\}$).

Recent studies [8], [9] have employed a probabilistic model based on deeply-stacked artificial neural networks (deep neural networks; DNNs) called autoencoder (AE) [10]. Especially, an extended version of AEs, a variational autoencoder (VAE), is known to build a probabilistic model of given natural images with impressive accuracy [11]. DNNs have achieved remarkable results in various tasks by learning high-level features from a given dataset in a data-driven manner. However, subsequent studies have revealed that DNNs assign unreasonably high likelihoods to samples even if the DNNs mistakenly recognize the samples [12]. Hence, many studies have investigated more reliable measures of anomalies for DNNs in supervised segmentation and regression tasks [5], [13] and in sample-wise anomaly detection [14], [15].

Given the above, we propose a novel pixel-wise anomaly score for VAE. Since the VAE assigns a likelihood to each sample, some previous studies employed the pixel-wise mean-squared error between a real sample and an estimated mean as anomaly score [16]. Instead, we defined an approximation of a pixel-wise conditional likelihood like pseudo-likelihood. Then, we decomposed the approximated log-likelihood into two terms; predictable uncertainty (called aleatoric uncertainty [5], [13], [17]) and normalized error, and proposed the normalized error as a pixel-wise anomaly score of VAE, following the previous study on sample-wise anomaly detection [14].

We evaluated our proposed score using head magnetic resonance imaging (MRI) datasets. The purpose was to segment the tumor and stroke lesions in an unsupervised manner. Our proposed method and comparative methods were trained using IXI dataset[1], and evaluated using Anatomical Tracings of

This study was partially supported by the MIC/SCOPE #172107101.

[1] https://brain-development.org/ixi-dataset/

978-1-7281-2009-6/$31.00 ©2019 IEEE

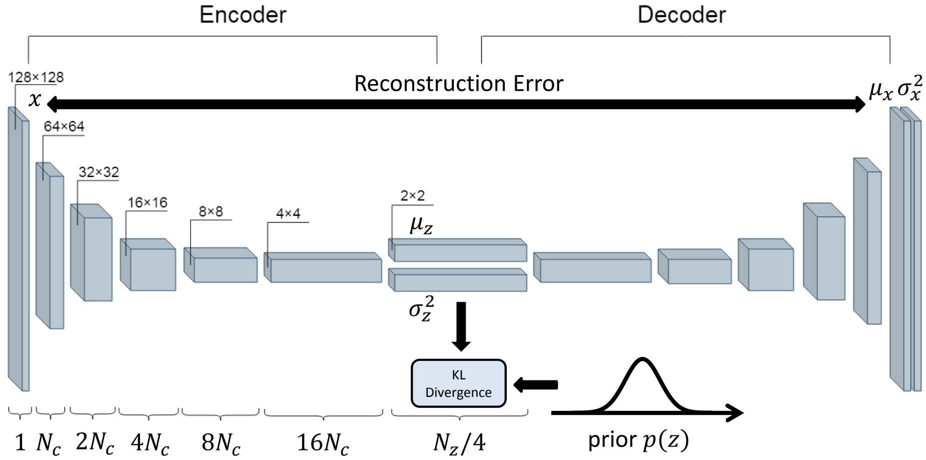

Fig. 1. A diagram of the variational autoencoder (VAE) implemented on convolutional neural networks (CNNs).

Lesions After Stroke (ATLAS) dataset [18] and Multimodal Brain Tumor Segmentation Challenge (BraTS) dataset [19], [20] The experimental results demonstrate that the proposed method outperforms comparative methods. The qualitative comparison suggests that the proposed method reduces false positives by improving signal-to-noise ratio as expected.

II. Related Works

A. Autoencoders

An autoencoder (AE) is a kind of DNN generally consisting of two components; the *encoder* and *decoder*, as depicted in Fig. 1. The encoder maps an input sample x into a low-dimensional intermediate variable z. The decoder accepts the intermediate variable z and then outputs a reconstruction $\tilde{x}$, which aims at reproducing the original sample x. The objective function to be minimized of an AE is typically a mean-squared-error (MSE) between the sample x and the reconstruction $\tilde{x}$. As an anomaly score, the mean squared error can be used for sample-wise anomaly detection [8], [9] and the pixel-wise squared error can be used for pixel-wise anomaly segmentation [16]. The AE can be viewed as a generative model, where they infer the latent variable z and then output x as a Gaussian posterior with an identity matrix as its covariance matrix.

A variational autoencoder (VAE) explicitly builds a model of the posteriors [11], as shown in Fig. 1. We consider a probabilistic model $p_\theta(x)$ of a sample x in the data space $\mathcal{X} \subset \mathbb{R}^{N_x}$ with a latent variable z in a latent space $\mathcal{Z} \subset \mathbb{R}^{N_z}$ for $N_z < N_x$:

$$p_\theta(x) = \int_z p_\theta(x|z)p(z),$$

Using the variational inference, the model evidence $\log p_\theta(x)$ is bounded as

$$\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{p_\theta(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] + D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \\
&\geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \\
&= -D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&=: -\mathcal{L}(x),
\end{aligned} \tag{1}$$

where $-\mathcal{L}(x)$ is called the evidence lower bound (ELBO). The ELBO $-\mathcal{L}(x)$ is the objective function to be maximized. The ELBO $-\mathcal{L}(x)$ is not equal to the exact log-likelihood $\log p(x)$, but one can expect that the ELBO $-\mathcal{L}(x)$ converges to the log-likelihood $\log p(x)$ [21].

For leveraging the flexibility of DNNs and the backpropagation algorithm, the VAE implements these probabilistic models $q_\phi(z|x)$ and $p_\theta(x|z)$ on the encoder and decoder using the reparameterization trick. Specifically, the encoder and decoder output the parameters of the corresponding probabilistic models instead of the point estimates. For a Bernoulli distribution, the parameter is a probability p that the variable takes the value of 1. For a Gaussian distribution, the parameters are a mean vector and covariance matrix. Due to the difficulty in calculating the expectation over a DNN in Eq. (1), it is approximated by Monte Carlo sampling. In the original implementation [11], the latent variable z is sampled once each iteration during training.

III. Methods

A. Predictable Uncertainty-aware Score for Variational Autoencoder

In this section, we propose the uncertainty-aware score for anomaly detection by the VAE (see also Fig. 1). We denote the elements of the visible variable $x = \{x_i\}$ and latent variable $z = \{z_j\}$ using the subscripts $i \in \{1, \ldots, N_x\}$

and $j \in \{1, \ldots, N_z\}$. In this study, we assume that each element (e.g., pixel) of a sample is a continuous variable and is modeled as a Gaussian distribution with a diagonal covariance matrix. Let $f(\cdot|\mu, \Sigma)$ be the probabilistic density function of a multivariate Gaussian distribution parameterized by the mean μ and covariance Σ. In this case, the output of the decoder is also a pair: a mean vector $\mu_x = \{\mu_{x_i}\}$ and a variance vector $\sigma_x = \{\sigma_{x_i}\}$ for the conditional probability $p_\theta(x|z) = f(x; \mu_x(z), \mathrm{diag}(\sigma_x(z)))$. The same goes for the encoder. Then, the negative ELBO $\mathcal{L}(x)$ can be rewritten as

$$\begin{aligned} \mathcal{L}(x) = &\int_z f(z; \mu_z, \mathrm{diag}(\sigma_z^2) \log \frac{f(z; \mu_z, \mathrm{diag}(\sigma_z^2))}{f(z; \mathbf{0}, I)} \\ &- \mathbb{E}_{q_\phi(z|x)} \left[-\log f(x; \mu_x, \mathrm{diag}(\sigma_x^2)) \right] \\ = &\sum_j \frac{1}{2}(-\log \sigma_{z_j}^2(x) - 1 + \sigma_{z_j}^2(x) + \mu_{z_j}^2(x)) \\ &+ \mathbb{E}_{q_\phi(z|x)} \left[\textstyle\sum_i \frac{1}{2} \log 2\pi\sigma_{x_i}^2(z) + \sum_i \frac{(\mu_{x_i}(z) - x_i)^2}{2\sigma_{x_i}^2(z)} \right]. \end{aligned} \tag{2}$$

The expectation over the posterior $q_\phi(z|x)$ is approximated by Monte Carlo sampling. This is the actual objective function used to train the VAE.

In the test phase, instead of the expectation over the posterior, the MAP estimate μ_z is often used for reconstruction in the test phase [11]. Since our purpose is to detect anomalous regions (pixels) in images, we focus on the posterior of the image x given a latent variable z. We can deal with individual pixels because of the diagonal covariance matrix. Then, the approximated negative log-likelihood $\mathcal{L}(x_i; x)$ of a pixel x_i given an image x is

$$\begin{aligned} \mathcal{L}(x_i; x) = &\log f(x_i; \mu_{x_i}(x), \sigma_{x_i}^2(x)) \\ = &\underbrace{\frac{1}{2} \log 2\pi\sigma_{x_i}^2(x)}_{\substack{\text{predictable uncertainty} \\ \mathcal{U}(x_i; x)}} + \underbrace{\frac{(\mu_{x_i}(x) - x_i)^2}{2\sigma_{x_i}^2(x)}}_{\substack{\text{normalized error} \\ \mathcal{E}(x_i; x)}} \end{aligned} \tag{3}$$

where the mean μ_{x_i} and variance $\sigma_{x_i}^2$ are determined by the input x via the encoder and decoder. First, we proposed this pixel-wise negative log-likelihood $\mathcal{L}(x_i; x)$ as an anomaly score.

The first term $\mathcal{U}(x_i; x)$ is the log-variance of the posterior and the second term $\mathcal{E}(x_i; x)$ is the squared error normalized by the variance. If a pixel x_i is blurred, at a border of two regions, suffering from noise, or obtained from a region with a complicated shape, the pixel x_i is almost unpredictable and the expected squared error $(\mu_{x_i}(x) - x_i)^2$ becomes large. Then, the VAE assumes a larger variance $\sigma_{x_i}^2$ to surpass the normalized error $\mathcal{E}(x_i; x)$ while increasing the log-variance $\mathcal{U}(x_i; x)$. The VAE balances the predictable uncertainty $\mathcal{U}(x_i; x)$ and the normalized error $\mathcal{E}(x_i; x)$ depending on the uncertainty of the pixel x_i given the image x, and the log-variance $\mathcal{U}(x_i; x)$ can be considered as predictable uncertainty. In other words, VAE can ignore a pixel x_i with a penalty $\mathcal{U}(x_i; x)$. Therefore, the log-likelihood of a pixel is lower-bounded by the uncertainty of the pixel even if the VAE builds a good model [5], [14]. Here, we propose the normalized error $\mathcal{E}(x_i; x)$ as an alternative anomaly score, which is the log-likelihood $\mathcal{L}(x_i; x)$ subtracted with the predictable uncertainty $\mathcal{U}(x_i; x)$.

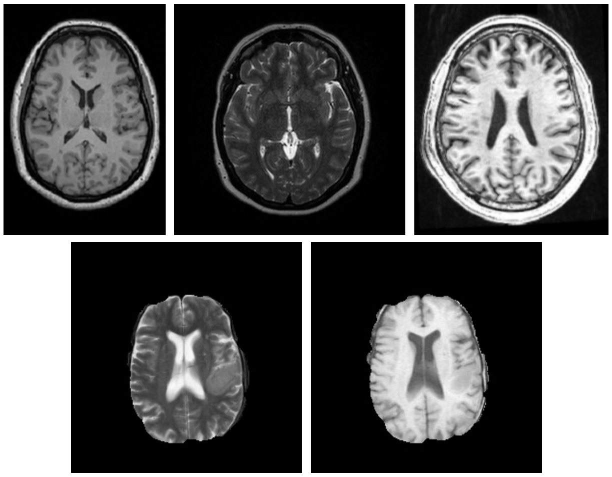

Fig. 2. Examples images in datasets. (upper row) IXI-T1w, IXI-T2w, and ATLAS-T1w. (bottom row) BraTS-T2w, and BraTS-T1w.

IV. Experiments

A. Data Acquisition

A head magnetic resonance imaging (MRI) is a technique imaging the anatomy of the brain by using powerful magnetic fields and radio wave pulses. The magnetic fields and radio wave pulses induce the state transitions and resonance of certain atomic nuclei. The response depends on the tissue properties. Varying the pulse repetition time and the echo time, we can emphasize some tissues and obtain several types of images (T1, T2, Flair, etc.). For example, cerebro-spinal fluid (CSF) is often visible as darker pixels in T1-weighted images while it is brighter in T2-weighted images.

We evaluated our proposed method using head MRI datasets. Example images are shown in Fig. 2.

- **IXI dataset**: We used the IXI dataset for training models. This dataset contains T1- and T2-weighted MRI images from 579 healthy subjects. Each T1-weighted image is composed of 256 horizontal slices and each slice has a size of 150×256. Each T2-weighted image is composed of 130 or 136 horizontal slices and each slice has a size of 256×256.
- **ATLAS-T1w** [18]: For evaluation, we used Anatomical Tracings of Lesions After Stroke (ATLAS) dataset, containing T1-weighted images of 220 stroke patients. Each MRI image is composed of 189 horizontal slices and each slice has a size of 197×233.
- **BraTS** [19], [20]: For additional evaluation, we used the T2- and T1-weighted images of 285 patients from the Brain Tumor Segmentation Challenge (BraTS). Each T2- and T1-weighted MRI image is composed of 155 horizontal slices and each slice has a size of 240×240.

While the original datasets are given as 3D images, we used the horizontal slices as 2D images separately because a comparative method, GMM, has a difficulty in modeling a high-dimensional image. We emphasize that our proposed method and other AE-based methods are applicable to 3D images directly using 3D convolutional neural networks (CNNs).

BraTS dataset is provided after brain tissue extraction. For other datasets, we performed Brain Extraction Tools (BET) [22] to remove non-brain tissues. After that, we cropped a brain tissue region and resized the image to the same scale, obtaining 128×128. We also scaled ground truth annotations to the same scale. Moreover, we normalized each image by dividing by the median intensity of brain pixels.

B. Proposed and Comparative Autoencoders

As comparative methods, we implemented the AE and VAE using PyTorch v1.0.0. In addition, we implemented the denoising autoencoder (DAE), which was an AE trained with input images x after a Gaussian noise of zero mean and a standard deviation of 0.5 was added [16]. The encoder consisted of 6 convolution layers with 4×4 kernels and a stride of 2. The feature map after n-th layer had $N_c \times 2^{n-1}$ channels and was followed by batch normalization and the ReLU activation function for $n < 6$. The last convolution layer had a 3×3 kernel and a stride of 1 to lead the feature map of $2 \times 2 \times \frac{N_z}{4}$. The last feature map was reshaped to a feature vector of N_z, which is the latent variable z.

In the case of VAE, the latent variable z was defined as a stochastic variable as mentioned in Section II. The output layer led the feature vector of N_z. Half of the vector elements were used as a mean vector μ_z, and the other half were used as a log-variance vector $\log \sigma_z^2$. The two vectors jointly represented the parameters of the Gaussian posterior $q_\phi(z|x) = \mathcal{N}(\mu_z, \mathrm{diag}(\sigma_z^2))$. The decoder had a structure paired up with the encoder.

In the case of VAE, the output layer led an image of 2-channels, each also representing the mean μ_x or log-variance $\log \sigma_x^2$ of the posterior $p_\theta(x|z) = \mathcal{N}(\mu_x, \mathrm{diag}(\sigma_x^2))$.

We trained the AE variations using Adam optimizer ($\alpha = 10^{-3}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$) with a weight decay of 10^{-4}. We selected the dimension number of the latent space from $N_z \in \{8, 16, 32, 64, 128, 256, 512\}$ and the number of channels from $N_c \in \{8, 16, 32\}$.

For the AE and DAE, we used mean squared error (MSE) as the objective function to be minimized and used pixel-wise squared error as anomaly score as previous studies [8], [9], [16]. For the VAE, we used the ordinary objective function introduced in Section II. For comparison, we used pixel-wise squared error, approximated pixel-wise log-likelihood $\mathcal{L}(x_i; x)$, and the proposed score $\mathcal{E}(x_i; x)$ as anomaly scores.

C. Comparative Method

For additional evaluation, we implemented a method proposed by Leemput et al. [7]. The model is based on a Gaussian mixture model (GMM), which is a traditional generative model.

A GMM assumes that each pixel x belongs to one of the hidden classes $z \in \{1, 2, \ldots, M\}$. Each hidden class z has its base distribution expressed as a Gaussian distribution that has a mean μ_z and a covariance matrix Σ_z. A GMM is expressed as

$$\begin{aligned} p_\theta(x_i) &= \sum_{k=1}^{M} w_k^i p_\theta(x_i|z=k) \\ &= \sum_{k=1}^{M} w_k^i \frac{1}{\sqrt{(2\pi)^{N_x}|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right), \end{aligned}$$

where θ is a set of parameters. w_k^i denotes the mixture weight that the pixel i belongs to the hidden class k. A GMM is trained mainly using the Expectation-Maximization (EM) algorithm to maximize the model evidence $\log p_\theta(x)$. The EM algorithm iteratively updates the model parameters μ_k and Σ_k as

$$\mu_k \leftarrow \frac{\sum_i r_k^i x_i}{\sum_i r_k^i},$$

$$\Sigma_k \leftarrow \frac{\sum_i r_k^i (x_i-\mu_k)(x_i-\mu_k)^T}{\sum_i r_k^i}$$

where r_k^i represents the posterior probability that the pixel x_i belongs to the hidden class k given the current parameters;

$$r_k^i = \frac{p_\theta(x_i|z=k) w_k^i}{\sum_{q=1}^{M} p_\theta(x_i|z=q) w_q^i},$$

Generally, a Gaussian (mixture) model trained by EM algorithm is very sensitive to outliers; since $\lim_{x_i \to \pm\infty} \log p_\theta(x_i) = -\infty$, the contribution of an anomalous pixel to the log-likelihood is high. Therefore, Leemput et al. [7] proposed a mixture model of a Gaussian distribution of normal pixels and an uniform distribution of anomalous pixels;

$$p_\theta(x_i) = \sum_{k=1}^{M} w_k^i (f(x_i|\mu_k, \Sigma_k) + \lambda).$$

Then, the EM algorithm updates the parameter in each iteration as

$$\mu_k \leftarrow \frac{\sum_i r_k^i t_k^i x_i}{\sum_i r_k^i t_k^i},$$

$$\Sigma_k \leftarrow \frac{\sum_i r_k^i t_k^i (x_i-\mu_k)(x_i-\mu_k)^T}{\sum_i r_k^i t_k^i},$$

where t_k^i denotes the weight that reflects the typicality of the pixel i for the hidden class k:

$$t_k^i = \frac{f(x_i|\mu_k, \Sigma_k)}{f(x_i|\mu_k, \Sigma_k) + \lambda}.$$

Since $\lim_{x_i \to \pm\infty} \log(f(x_i|\mu_k, \Sigma_k) + \lambda) = \log \lambda$, the contribution of an anomalous pixel is bounded. In this model, the posterior

$$\sum_{k=1}^{M} r_k^i (1 - t_k^i) = \sum_{k=1}^{M} r_k^i \frac{\lambda}{f(x_i|\mu_k, \Sigma_k) + \lambda}$$

TABLE I
RESULTANT ROC-AUCS.

Model	Score	ATLAS-T1w		BraTS-T2w		BraTS-T1w	
		Hyper-Parameters	ROC-AUC	Hyper-Parameters	ROC-AUC	Hyper-Parameters	ROC-AUC
Baseline	pixel intensity	—	0.612	—	0.811	—	0.549
modified GMM [23]	$\lambda/(f(x_i)+\lambda)$	$\lambda = 0.001, N_z = 4$	0.635	$\lambda = 0.001, N_z = 4$	**0.798**	$\lambda = 0.001, N_z = 4$	0.550
AE	squared Error	$N_c = 16, N_z = 8$	0.588	$N_c = 16, N_z = 256$	0.566	$N_c = 16, N_z = 16$	0.545
DAE	squared Error	$N_c = 16, N_z = 8$	0.583	$N_c = 16, N_z = 256$	0.607	$N_c = 16, N_z = 16$	0.532
VAE	squared Error	$N_c = 32, N_z = 8$	0.611	$N_c = 16, N_z = 256$	0.733	$N_c = 32, N_z = 16$	0.554
	$\mathcal{L}(x_i; x)$	$N_c = 32, N_z = 8$	0.626	$N_c = 16, N_z = 256$	0.737	$N_c = 32, N_z = 16$	0.561
	$\mathcal{U}(x_i; x)$	$N_c = 32, N_z = 8$	0.404	$N_c = 16, N_z = 128$	0.520	$N_c = 16, N_z = 16$	0.436
VAE	$\mathcal{E}(x_i; x)$	$N_c = 32, N_z = 8$	**0.672**	$N_c = 16, N_z = 256$	0.788	$N_c = 32, N_z = 16$	**0.582**

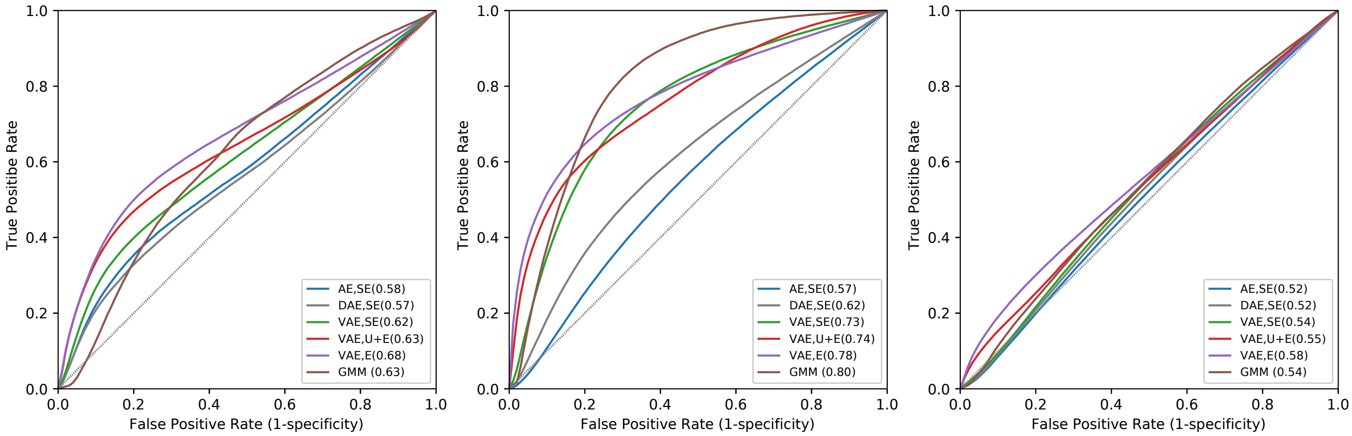

Fig. 3. Receiver operating characteristic (ROC) curves of the models with the best hyper-parameters. The datasets are ATLAS-T1w, BraTS-T2w, and BraTS-T1w from left to right.

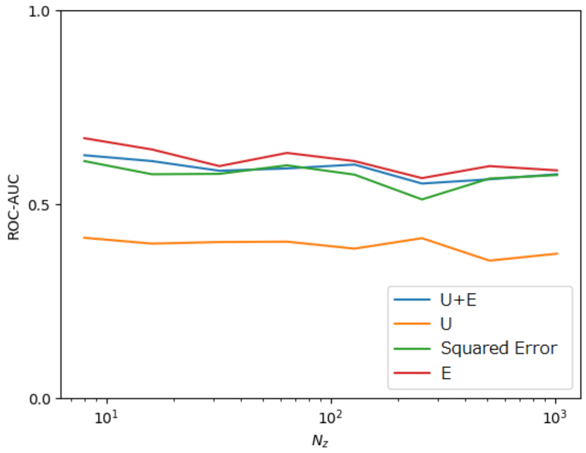

Fig. 4. ROC-AUCs on the ATLAS-T1w dataset with the varying N_z of the VAE.

that the pixel x_i is generated from the uniform distributions can be used as anomaly score. We call this model a modified GMM, hereafter. We implemented the modified GMM by using scikit-learn v0.19.1 [24]. We selected λ from $\lambda \in \{0.1, 0.01, 0.001, 0.0001\}$.

V. RESULTS AND DISCUSSION

We depicted the receiver operating characteristic (ROC) curves, which show the relationship between the true positive rate (or sensitivity) and the false positive rate (or 1-specificity) with a varying threshold. We also calculated the areas under the ROC curves (ROC-AUCs). We summarized the best ROC-AUCs and the corresponding hyper-parameters in Table I, and the corresponding ROCs in Fig. 3. The VAE with the proposed score $\mathcal{E}(x_i; x)$ outperformed the VAE with an ordinary score $\mathcal{L}(x_i; x)$ in any cases. Especially, it outperformed all other methods for the ATLAS-T1w dataset. On one hand, for the BraTS-T2w dataset, the modified GMM outperformed the comparative methods and the pixel intensity was the best. The T2-weighted MRI shows much bright pixels to visualize water-rich tissues, especially tumors, and then diagnostic models easily detect pixels brighter than a certain threshold as tumors. We consider it and GMM worked well thanks to its simplicity. Meanwhile, we also evaluated the comparative models on the BraTS-T1w dataset and confirmed that the VAE with the proposed score $\mathcal{E}(x_i; x)$ achieved the best result. While the T2-weighted MRI is practically useful, the T1-weighted MRI visualizes the brain structure and requires a more semantic and structural analysis. In this sense, we conclude that the VAE with the proposed score $\mathcal{E}(x_i; x)$ are superior to ordinary AE variations and the comparative model in semantic anomaly segmentation.

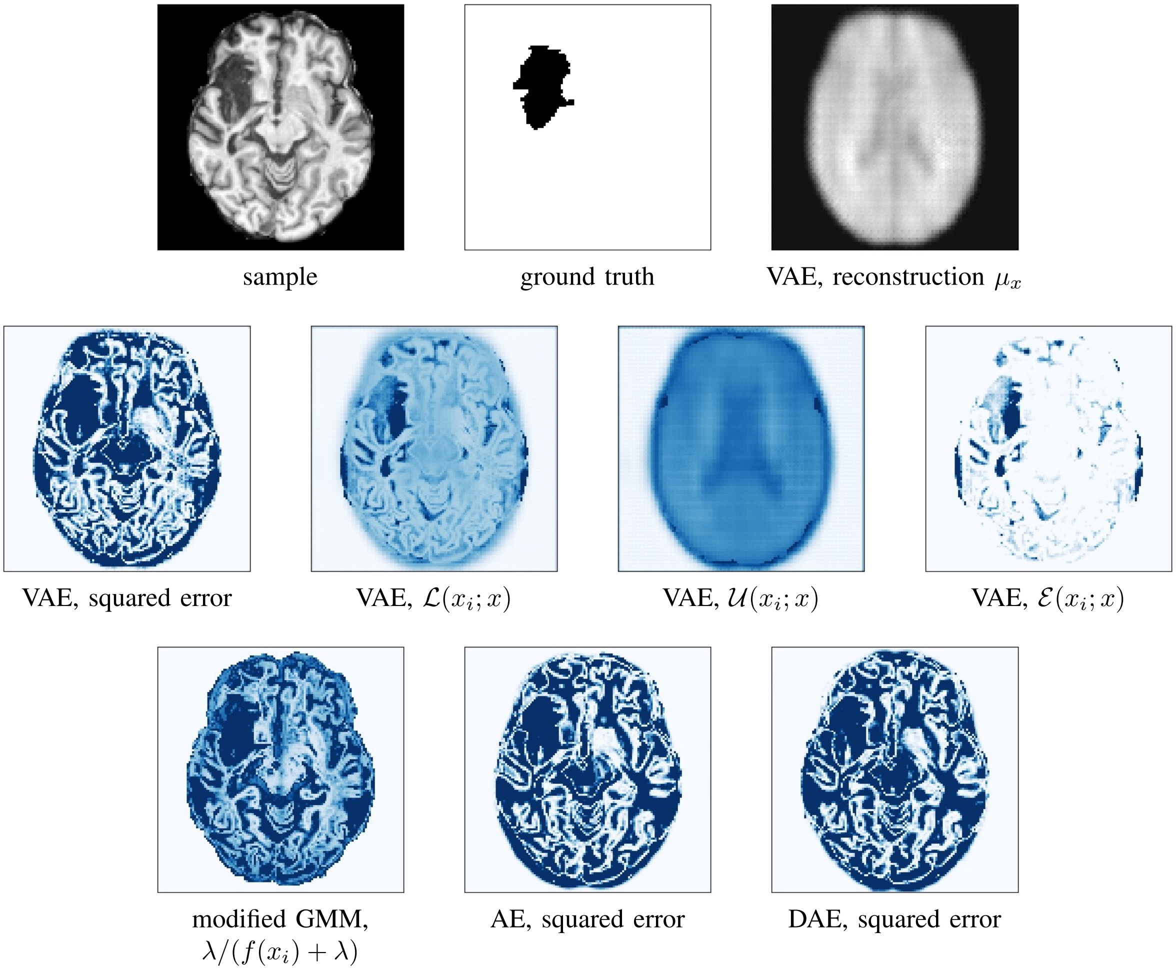
sample
ground truth
VAE, reconstruction μ_x
VAE, squared error
VAE, $\mathcal{L}(x_i; x)$
VAE, $\mathcal{U}(x_i; x)$
VAE, $\mathcal{E}(x_i; x)$
modified GMM, $\lambda/(f(x_i)+\lambda)$
AE, squared error
DAE, squared error

Fig. 5. A sample from the ATLAS-T1w dataset, the ground truth of anomalous regions, and the reconstruction μ_x by the VAE. The other panels depict the corresponding heat maps of the anomaly scores.

Fig. 4 depicts the resultant ROC-AUCs on the ATLAS-T1w dataset with the varying dimension number N_z for the VAE and with the fixed number $N_c = 32$ of channels of the first feature map. The ROC-AUCs worsened with the increase in the dimension number N_z of the latent space. With any dimension numbers N_z, the ROC-AUC obtained from the proposed score $\mathcal{E}(x_i; x)$ exceeded other anomaly scores. The figure also shows that the ROC-AUCs with $\mathcal{U}(x_i; x)$ never exceeded 0.5, which implies that the log-variance term $\mathcal{U}(x_i; x)$ did not react to the semantically anomalous regions in this dataset.

In Fig. 5, we show a sample from ATLAS-T1w, the ground truth annotation, where black pixels indicate anomalous regions, and the reconstruction μ_x by the VAE on the top. Of course, a brain has many gyri and sulci (ridges and grooves), whose shapes depend on a subject. The VAE did not reconstruct them in details. We also summarize the anomaly scores obtained by the comparative models, where each anomaly score is normalized to the range from 0 to 1 with brighter pixels representing lower scores. The squared error obtained from the VAE could not deal with the complicated shapes of the gyri and sulci and mistakenly detected them as anomalies. The same goes for the modified GMM, AE, and DAE. In short, they tend to be sensitive to high/low pixel intensities rather than semantic anomalies. Using the log-likelihood $\mathcal{L}(x_i; x)$, the VAE assigned relatively higher scores to the anomalous regions but it assigned certain scores also to the remaining regions. This is because the VAE outputted the high predictable uncertainty $\mathcal{U}(x_i; x)$ to absorb the reconstruction error of the gyri, sulci, center region, and edges. Only the proposed normalized error $\mathcal{E}(x_i; x)$ selectively detected the anomalous region thanks to the predictable uncertainty $\mathcal{U}(x_i; x)$. We conclude that the proposed normalized error $\mathcal{E}(x_i; x)$ is robust to the complexity of given image structure and detects semantic anomalies well.

VI. Conclusion

In this study, we proposed a novel anomaly score of variational autoencoder for unsupervised anomaly segmentation task. We obtained the score by removing the log-variance term from the objective function of the variational autoencoder. As the log-variance term tends to react to the data complexity rather than semantic anomalies, the remaining term is more robust to the complexity. We examined our score by anomaly segmentation task with head-MRI datasets. Thanks to the robustness, the proposed score worked better for these datasets.

References

[1] D. B. Perng, S. H. Chen, and Y. S. Chang, "A novel internal thread defect auto-inspection system," *International Journal of Advanced Manufacturing Technology*, vol. 47, no. 5-8, pp. 731–743, 2010.

[2] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[3] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, no. March, pp. 1–13, 2017. [Online]. Available: http://dx.doi.org/10.1038/sdata.2017.117

[4] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with Deep Neural Networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.media.2016.05.004

[5] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" *Advances in Neural Information Processing Systems (NIPS)*, 2017. [Online]. Available: http://arxiv.org/abs/1703.04977

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," in *ACM Computing Surveys*, vol. 41, no. 3, 2009, pp. 1–58. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1541880.1541882

[7] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 677–688, 2001. [Online]. Available: papers2://publication/uuid/7F80CB3D-AB28-4D1B-BB2D-BDAC398AD98B http://ieeexplore.ieee.org/document/938237/

[8] C. Zhou and R. C. Paffenroth, "Anomaly Detection with Robust Deep Autoencoders," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2017, pp. 665–674. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3097983.3098052

[9] R. Chalapathy, A. K. Menon, and S. Chawla, "Robust, Deep and Inductive Anomaly Detection," *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pp. 1–16, apr 2017. [Online]. Available: http://arxiv.org/abs/1704.06743

[10] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning (ICML)*, 2008, pp. 1096–1103. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1390156.1390294

[11] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations (ICLR)*, 2014, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1312.6114

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–11. [Online]. Available: http://arxiv.org/abs/1412.6572

[13] A. Kendall, Vijay Badrinarayanan, R. Cipolla, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *arXiv*, 2015. [Online]. Available: http://arxiv.org/abs/1511.02680

[14] T. Matsubara, R. Tachibana, and K. Uehara, "Anomaly Machine Component Detection by Deep Generative Model with Unregularized Score," in *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[15] H. Choi and E. Jang, "Generative Ensembles for Robust Anomaly Detection," pp. 1–10, 2018. [Online]. Available: http://arxiv.org/abs/1810.01392

[16] X. Chen, N. Pawlowski, M. Rajchl, B. Glocker, and E. Konukoglu, "Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging," *arXiv*, pp. 1–10, 2018. [Online]. Available: https://arxiv.org/pdf/1806.05452.pdf

[17] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.strusafe.2008.06.020

[18] S.-L. Liew, J. M. Anglin, N. W. Banks, M. Sondag, K. L. Ito, H. Kim, J. Chan, J. Ito, C. Jung, N. Khoshab *et al.*, "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations," *Scientific data*, vol. 5, p. 180011, 2018.

[19] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, p. 1993, 2015.

[20] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 170117, 2017.

[21] C. C. M. C. Bishop, "Pattern recognition and machine learning," in *Springer*, vol. 4, no. 4, 2006, p. 738. [Online]. Available: http://www.library.wisc.edu/selectedtocs/bg0137.pdf

[22] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[23] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE transactions on medical imaging*, vol. 20, pp. 677–88, 09 2001.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490