# A Case for the Score: Identifying Image Anomalies using Variational Autoencoder Gradients

**David Zimmerer, Jens Petersen, Simon A. A. Kohl** and **Klaus H. Maier-Hein**

Division of Medical Image Computing
German Cancer Research Center (DKFZ)
Heidelberg, Germany
{d.zimmerer,jens.petersen,simon.kohl,k.maier-hein}@dkfz.de

## Abstract

Through training on unlabeled data, anomaly detection has the potential to impact computer-aided diagnosis by outlining suspicious regions. Previous work on deep-learning-based anomaly detection has primarily focused on the reconstruction error. We argue instead, that pixel-wise anomaly ratings derived from a Variational Autoencoder based *score* approximation yield a theoretically better grounded and more faithful estimate. In our experiments, Variational Autoencoder gradient-based rating outperforms other approaches on unsupervised pixel-wise tumor detection on the BraTS-2017 dataset with a ROC-AUC of 0.94.

## 1 Introduction

In recent years several deep-learning-based methods have reported reaching comparable performance to trained medical physicians [11, 17]. One weakness of those approaches is that they still require a lot of annotated data for each condition to be trained on. Due to the time-intensive work of annotating medical images and the combinatorial number of cases for different modalities, image qualities, hardware devices, and different conditions, it is still infeasible to train an algorithm for each of the existing combinations. Anomaly detection can, while not determining the condition, highlight and identify suspicious regions for a closer inspection by a trained physician. By assigning each pixel an anomaly rating, it allows for an easy trade-off of specificity and sensitivity. While this may not be able to outperform supervised algorithms, it offers a way to make use of unlabeled data and aid physicians during the diagnosis.

Previous unsupervised anomaly detection approaches in the medical field were primarily based on a reconstruction error. Leemput et al. [19] use a statistical model to reconstruct the input tissue-wise, quantifying the discrepancies between the actual image and the model prediction to identify anomalies. Liu et al. [10] decompose the model into low-rank components which representing the normal parts of the image, and high-frequency parts which representing anatomical and pathological variations and are thus able to delineate suspicious areas. More recently multiple deep learning Autoencoder (AE) based methods have been proposed, all considering the reconstruction error. Chen et al. [4, 5] propose to use an adversarial latent loss in addition to a Variational Autoencoder (VAE) and compare it to different AE-based approaches. Baur et al. [3] use a VAE with an adversarial loss on the reconstruction to get a more realistic reconstruction. Pawlowski et al. [13] compare different AEs for CT based pixel-wise segmentation.

All those approaches use the reconstruction error to identify suspicious regions, based on the idea that models can not truthfully reproduce anomalies not seen during training. Despite showing good results, there are no formal guarantees for that assumption. In the next section we will describe how

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

to use the *score*, defined as the derivative of the log-density with respect to the input $\frac{\partial \log p(x)}{\partial x}$ [6], as an alternative anomaly rating.

## 2 Methods

Alain et al. [1] have shown that for AE-based models with a denoising criterion the reconstruction error approximates the *score*. It can be anticipated that most AE- and reconstruction-based models work due to an approximation of the *score*. Consequently and based on the following assumptions, we hypothesize that the *score* can give a good approximation for an abnormality rating:

- The *score* gives the directions towards the normal data samples, which for medical data is the data sample with abnormal anatomies and pathologies transformed into healthy parts,
- The magnitude of the *score* indicates how abnormal the pixel is.

In this work, we describe a way to directly estimate the *score* using VAEs, one of the best performing density-estimation models for images [5, 9]. The objective of VAEs is to learn a generative model of the data by maximizing the evidence lower bound (ELBO) for the given training data. The ELBO is defined as:

$$\log p(x) \geq -D_{KL}(q(z|x)||p(z)) + \mathbb{E}_{q(z|x)}[\log p(x|z)], \tag{1}$$

Where $q(z|x)$ is the inference model, $p(z)$ is the prior for the latent variables, $D_{KL}$ is the Kullback-Leibler divergence, and $p(x|z)$ is the generative model. Thus after training the VAE and maximizing the ELBO, an estimate of the log probability $\log p(x)$ of a data sample $x$ can be calculated by evaluating the rhs of Eq. 1 for the data sample $x$. The approximate *score* can consequently be calculated by taking the derivative of the ELBO with respect to the data sample:

$$\frac{\partial \log p(x)}{\partial x} \approx \frac{\partial(-D_{KL}(q(z|x)||p(z)) + \mathbb{E}_{q(z|x)}[\log p(x|z)])}{\partial x}, \tag{2}$$

Furthermore, the ELBO is fully differentiable [8, 14], when training a VAE using Gaussian distributions for $p(z)$ and $p(x|z)$, a parameterization by neural networks, the reparameterization trick, and MC sampling to approximate the expectation. This allows training of the VAE and the evaluation of Eq. 2 using the backpropagation algorithm.

We note that the above-mentioned assumptions can be violated in practice, especially in cases far away from the healthy sample data distribution. However, in the next section, we will present empirical evidence that our model can outperform reconstruction-based methods on an anomaly detection tasks and describe its benefits.

## 3 Experiments & Results

To learn the healthy data distribution we trained the VAE model on 1092 T2 MRI images of Human Connectome Project (HCP) dataset [18], with minor data augmentations, such as multiplicative color augmentations, random mirroring, and rotations. We evaluate the anomaly detection in the context of finding and outlining tumors on the BraTS-2107 dataset [2, 12]. Therefore we calculate a pixel-wise rating and then report the ROC-AUC. Both datasets were normalized and slice-wise resampled to a resolution of 64x64 pixels. As encoder and decoder for the AE-based models, we used a 5-layer fully convolutional neural network with LeakyReLUs and a latent size of 1024. To backpropagate onto the image and approximate the *score*, we used the Smoothgrad algorithm [16]. Due to checkerboard artifacts caused by the convolutions, we apply Gaussian smoothing to the gradients. The model was trained for 60 epochs with a batchsize of 64 and Adam as the optimizer with a learning rate of $0.0002$.

To evaluate the benefits of the *score*, we compare the model to a Denoising Autoencoder (DAE) [20] with the same architecture using the reconstruction error. Furthermore, we compare the *score* with the reconstruction error of the VAE, the smoothed reconstruction error, and the sampling deviations by determining the standard deviation of multiple MC samples. We further inspect the *score*, dividing it into the reconstruction-loss gradient and KL-loss gradient to get insights into the benefits of including the KL-term into the anomaly detection. The results can be seen in Fig. 1a (and Appendix Table 1), samples and the corresponding pixel-wise ratings for samples are presented in Fig. 1b (and Appendix Fig. 3 & 4).

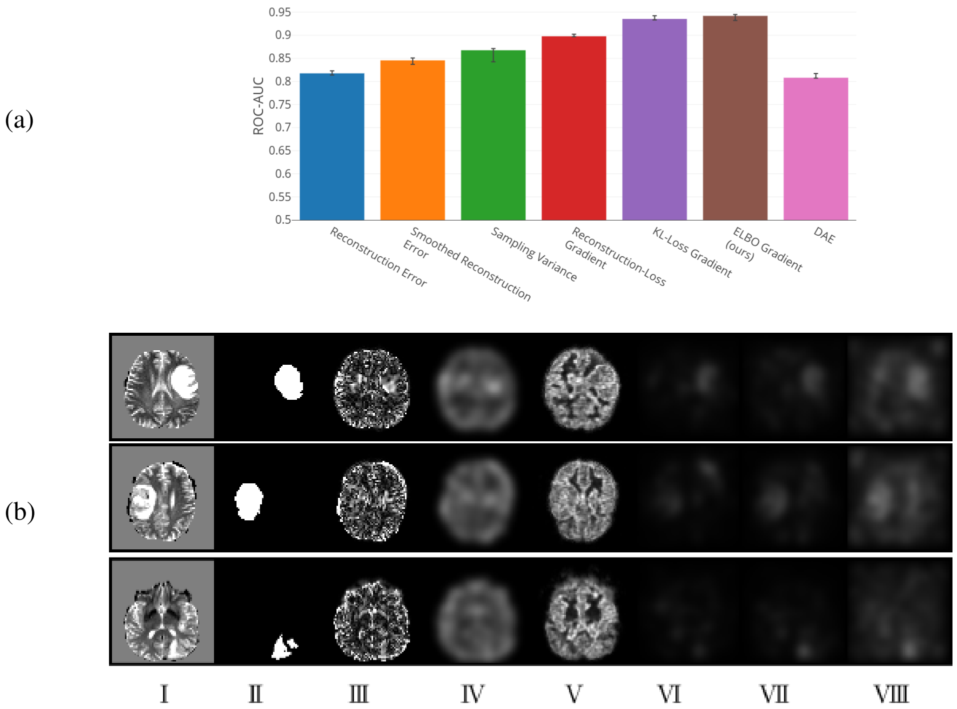

Figure 1: (a) Comparison of the pixel-wise tumor detection ROC-AUC on the BraTS-2017 dataset. (b) Samples from the dataset with the different pixel-wise rating schemes, showing the original sample (I), the annotation (II), the reconstruction error (III), the smoothed reconstruction error (IV), the sampling variances (V), the reconstruction-loss gradient (VI), the KL-loss gradient (VII), and the ELBO gradient which approximates the *score* (VIII).

The reconstruction error performs similarly for the VAE and the DAE, which was also reported in [5, 13]. Smoothing leads to slightly improved results, presumably by removing high-frequency detections, and performs on par with the usage of the sampling variances. The approximated *score* using the ELBO gradient (KL-loss + reconstruction-loss) performs best with a pixel-wise ROC-AUC of 0.94 (see Appendix Fig. 2) . It is interesting to see, that the addition of the reconstruction-loss to the KL-loss shows little benefit over the KL-loss gradient. Furthermore, the reconstruction-loss gradient performs worse than the KL-loss gradient but outperforms the reconstruction error.

In Fig. 1a, the reconstruction-loss gradient focuses on parts of poor reconstruction, and the combination of the KL-loss with the reconstruction-loss shows only marginal benefits over the KL-loss gradient. This might be an indication that for this model the KL-loss focuses primarily on the distance to the data distribution, while the reconstruction focuses more on the actual reconstruction task.

### 3.1 Discussion & Conclusion

We have presented a way to estimate the *score* using VAE gradients to detect anomalies on the BraTS-2017 tumor segmentation dataset. The results show competitive unsupervised segmentation performance, slightly outperforming the previously best reported ROC-AUC of 0.92 [4, 5]. The relative influence of the reconstruction loss can depend on the regularization of the latent variables. Using fewer latent variables or putting more importance on the KL-loss could, while potentially causing inferior overall performance, lead to a more competitive performance of the reconstruction error.

To the best of our knowledge, we are the first to use the gradients of a VAE, which approximate the *score*, to identify anomalies in images. The results suggest that the approximated *score*, including the often ignored KL-loss, can give a boost on the pixel-wise anomaly detection performance.

Furthermore, we want to stress the point that including the KL-loss for a pixel-wise anomaly detection and the *score* of a model can lead to an improvement in VAE-based methods for pixel-wise anomaly ratings.

This method should also be directly applicable to other state-of-the-art density estimation techniques, such as Grow [7] or Pixel-CNN++ [15], and it would be an interesting next step to see how different models perform.

## References

[1] G. Alain and Y. Bengio. What Regularized Auto-encoders Learn from the Data-generating Distribution. *J. Mach. Learn. Res.*, 15(1):3563–3593, Jan. 2014.

[2] 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. *Sci Data*, 4:170117, 2017.

[3] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. *CoRR*, abs/1804.04488, 2018.

[4] X. Chen and E. Konukoglu. Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *CoRR*, abs/1806.04972, 2018.

[5] X. Chen, N. Pawlowski, M. Rajchl, B. Glocker, and E. Konukoglu. Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging. *CoRR*, abs/1806.05452, 2018.

[6] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, Dec. 2005.

[7] D. P. Kingma and P. Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. *CoRR*, abs/1807.03039, 2018.

[8] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *CoRR*, abs/1312.6114, 2013.

[9] B. Kiran, D. Thomas, R. Parakkal, B. R. Kiran, D. M. Thomas, and R. Parakkal. An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos. *Journal of Imaging*, 4(2):36, Feb. 2018.

[10] X. Liu, M. Niethammer, R. Kwitt, M. McCormick, and S. Aylward. Low-Rank to the Rescue – Atlas-based Analyses in the Presence of Pathologies. *Med Image Comput Comput Assist Interv*, 17(Pt 3):97–104, 2014.

[11] Y. Liu, K. K. Gadepalli, M. Norouzi, G. Dahl, T. Kohlberger, S. Venugopalan, A. S. Boyko, A. Timofeev, P. Q. Nelson, G. Corrado, J. Hipp, L. Peng, and M. Stumpe. Detecting cancer metastases on gigapixel pathology images. Technical report, arXiv, 2017.

[12] 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, C. 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. S. 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 Trans Med Imaging*, 34(10):1993–2024, Oct. 2015.

[13] N. Pawlowski, M. C. H. Lee, M. Rajchl, S. McDonagh, E. Ferrante, K. Kamnitsas, S. Cooke, S. K. Stevenson, A. M. Khetani, T. Newman, F. A. Zeiler, R. J. Digby, J. P. Coles, D. Rueckert, D. K. Menon, V. F. J. Newcombe, and B. Glocker. Unsupervised Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders. 2018.

[14] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1278–II–1286, Beijing, China, 2014. JMLR.org.

[15] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *CoRR*, abs/1701.05517, 2017.

[16] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[17] G. V, P. L, C. M, and et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[18] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, E. Yacoub, and WU-Minn HCP Consortium. The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, Oct. 2012.

[19] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans Med Imaging*, 20(8):677–688, Aug. 2001.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

# 4 Appendix

## 4.1 Quantitative Results

| | ROC-AUC |
|---|---|
| DAE | $0.808 \pm 0.009$ |
| Reconstruction Error | $0.817 \pm 0.003$ |
| Smoothed Reconstruction Error | $0.843 \pm 0.008$ |
| Sampling Variance | $0.855 \pm 0.013$ |
| Reconstruction-Loss Gradient | $0.894 \pm 0.020$ |
| KL-Loss Gradient | $0.939 \pm 0.007$ |
| ELBO Gradient | $0.939 \pm 0.008$ |

Table 1: Pixel-wise ROC-AUC values of the compared approaches (see Fig. 1).

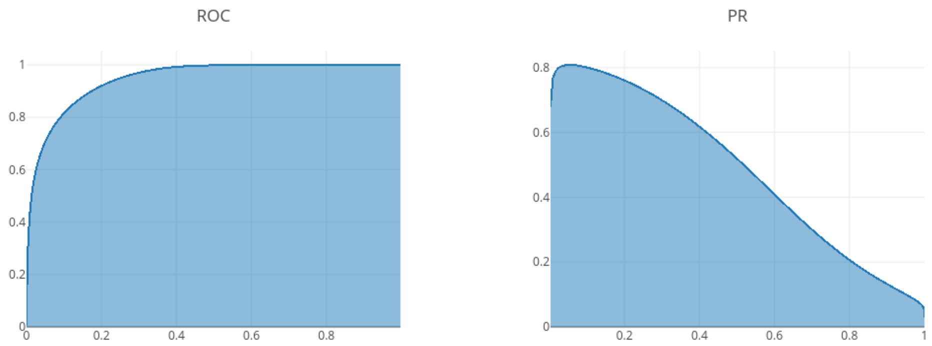

Figure 2: Pixel-wise Reciver Operator Curve (ROC) and Precision Recall (PR) Curve on the test set for the VAE ELBO-gradient with regard to the anomaly labels (all annotations are considered anomalies).

### 4.2 Qualitative Results

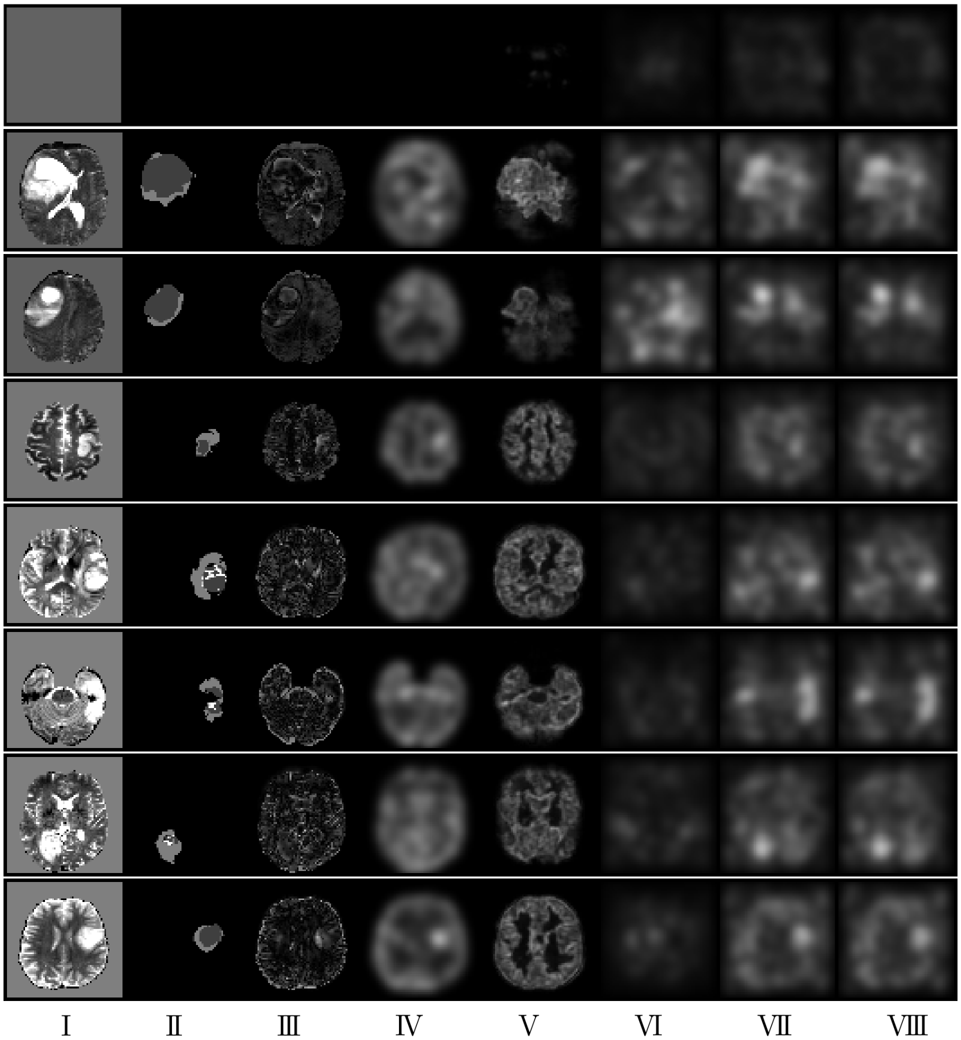

Figure 3: More samples as presented in Fig. 1, showing the original sample (I), the annotation (II), the reconstruction error (III), the smoothed reconstruction error (IV), the sampling variances (V), the reconstruction-loss gradient (VI), the KL-loss gradient (VII), and the ELBO gradient which approximates the *score* (VIII).

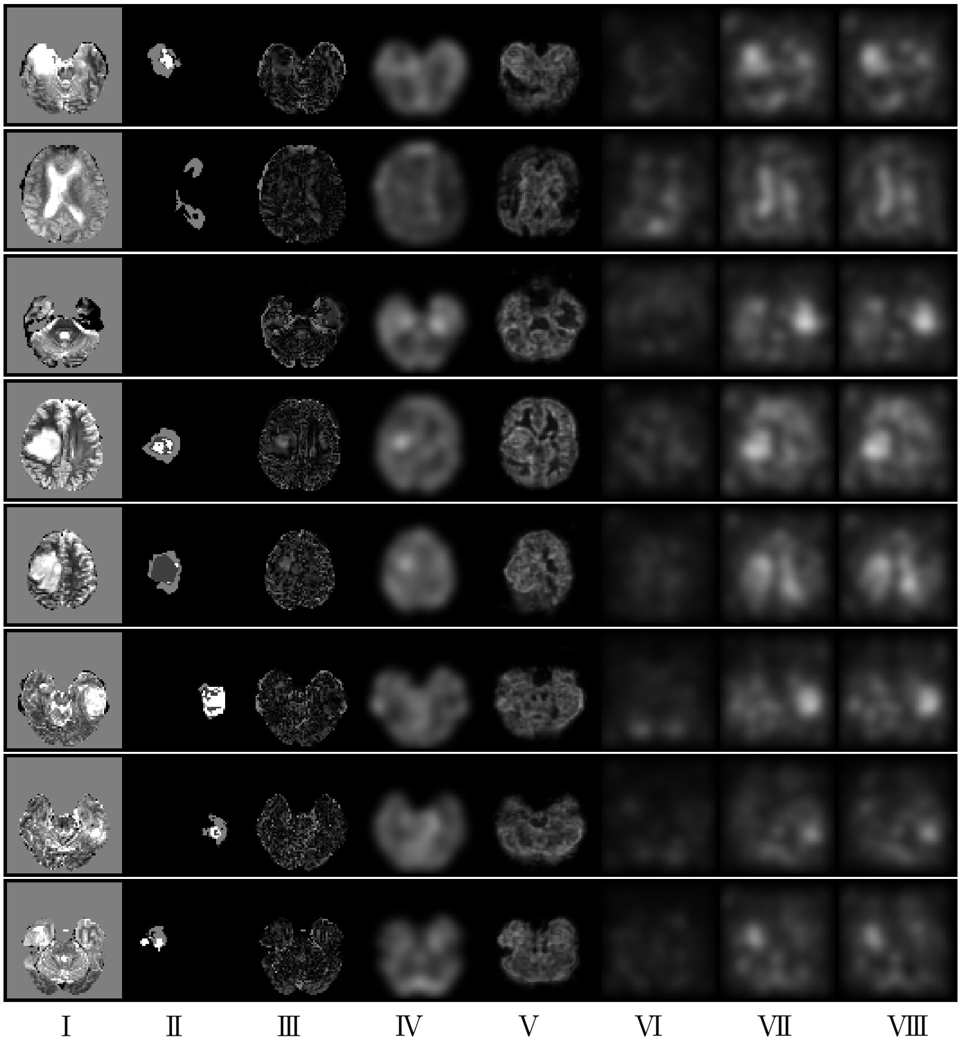

Figure 4: More samples as presented in Fig. 1, showing the original sample (I), the annotation (II), the reconstruction error (III), the smoothed reconstruction error (IV), the sampling variances (V), the reconstruction-loss gradient (VI), the KL-loss gradient (VII), and the ELBO gradient which approximates the *score* (VIII).