Tumour Detection in Brain MRIs by Computing Dissimilarities in the Latent Space of a Variational AutoEncoder

Alexandra-Ioana Albu[1,*], Alina Enescu[1,2,*], and Luigi Malagò[1,*]

[1] Romanian Institute of Science and Technology Cluj-Napoca, Romania
[2] Babeş-Bolyai University, Cluj-Napoca, Romania

Abstract

The ability to automatically outline anomalies in brain Magnetic Resonance Images (MRIs) is of great importance in computer-aided diagnosis. Unsupervised anomaly detection methods work primarily by learning the distribution of healthy images and identifying abnormal tissues as outliers. In this paper, we propose a slice-wise detection method which first trains a pair of autoencoders on two different datasets, one with healthy individuals and the other one with images of both normal and tumoural tissues. Next, it classifies slices based on the distance in the latent space between the encoding of the image and the encoding of the reconstruction obtained through the autoencoder trained on healthy images only. We validate our approach with a series of preliminary experiments on the HCP and BRATS-2015 datasets, showing the capability of the proposed method to classify brain MRIs into healthy and unhealthy.

1 Introduction

Automatic analysis of medical images is of great relevance for developing reliable systems that can assist physicians in diagnosing pathologies. The importance of the task is given by the fact that an accurate diagnosis is time consuming and investigator-dependent. In this context, the study conducted by Drew et al. [5] showed the vulnerability to inattentional blindness which can lead to high miss rates of anomalies. Deep learning technologies have been extensively employed for analysing medical images, with impressive results [9]. However, annotations for large amounts of data are difficult to collect. For this reason, there is a need for designing unsupervised or semi-supervised methods that can outline anomalous regions in medical images.

In this paper, we propose a slice-wise tumour detection algorithm based on Variational AutoEncoders (VAEs) and we evaluate it on Magnetic Resonance Images (MRIs) of brains from two publicly available datasets: HCP [14] and BRATS-2015 [10, 7]. The characteristic of our proposed algorithm is that it discriminates slices based on the computation of a distance in the space of the approximate posteriors of a VAE trained on both healthy (or normal) and tumoural tissues. The distance is computed between the encoding of an original image and the encoding of its reconstruction through a VAE which has been trained only on healthy images. From this perspective we can describe our approach as semi-supervised, indeed even if the algorithm does not need to access any label from the dataset containing both normal and tumoural images, it needs to have access to a dataset of images which are guaranteed to be of healthy individuals.

VAEs [6, 12] are flexible generative models which can be used for performing inference on complex datasets. In the literature there are several applications of VAEs in the medical field, such as segmentation of tumours in brain MRIs [8, 11], estimation of the brain age from MRI scans [4, 15], or identification of mental disorders such as schizophrenia [13]. VAEs have been successfully used in numerous anomaly detection tasks, for instance to outline tumoural areas or other lesions in brain scans. The majority of these approaches are reconstruction-based, i.e., they detect abnormal pixels by quantifying the difference between the original image and its reconstruction [1, 2, 3]. Chen et al. [2, 3]

*{albu,enescu,malago}@rist.ro

https://doi.org/10.7557/18.5172

employed VAEs and Adversarial AutoEncoders to detect tumours and stroke lesions in MRIs. In [2] an additional penalty was added to the loss function of the autoencoders to obtain a better representation in the latent space for healthy and unhealthy images. Zimmerer et al. [17] proposed an alternative to the reconstruction error for detecting anomalous pixels, based on the derivative of the log-likelihood with respect to the inputs, which was shown to outperform reconstruction-based anomaly scores. In [16] a context-encoding mechanism was introduced in VAEs to improve the anomaly detection performance.

The paper is organized as follows. In Section 2 we present a brief overview of VAEs. Section 3 describes the proposed method, while in Section 4 we present the experimental setting for our experiments. In Section 5 we show that our method can discriminate between images containing tumours and healthy scans. Finally, Section 6 outlines the conclusions and future directions of research.

2 Variational AutoEncoders

VAEs [6, 12] are a specific type of autoencoders which consist of two neural networks. The first one is the encoder, which maps the input to the parameters θ of a probability density function q_θ over the latent space, the second one is the decoder, which maps the latent representation to a probability density function p_ϕ over the space of the observations. VAEs are usually trained using principles from variational inference, by maximizing a lower bound for the log-likelihood, which consists of two terms. The first one is the reconstruction error, measured as the expected log-likelihood with respect to samples obtained from the approximate posterior, while the second one is a Kullback-Leibler penalty term which forces the approximate posterior to be close to the prior, i.e.,

$$\mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] - \mathrm{KL}(q_\theta(z|x)||p(z)) \ . \quad (1)$$

For each input x, a VAE returns an approximate posterior $q_\theta(z|x)$ from which the latent variables are sampled. This implies that differently from a regular autoencoder, it is possible to compare the encodings of two different inputs also by computing a dissimilarity or a distance function between probability density functions in the space of the approximate posteriors.

Another important property of VAEs is related to the presence of the KL term which acts as a regularization. This is a characteristic property of VAEs for which the latent representations are more well-behaved, differently from an AutoEncoder (AE), in which no assumption can be made a priori on the distribution of the latent representations.

Since we are interested in defining robust measures in the space of the latent representations (either the space of the approximate posteriors or its sample space), due to the intrinsic features we have briefly highlighted, we consider VAEs to be more robust with respect to traditional AEs.

3 Proposed Methodology

Most of the unsupervised anomaly detection approaches for brain scans require the training of a single AE on healthy individuals, then use a dissimilarity between the original image and its reconstruction in order to detect possible anomalies [1, 2, 3, 16]. Nowadays, in medical imaging we have access to large (possibly unlabelled) datasets, which include individuals which may or may not have tumoural tissues. However such datasets cannot be directly used in training if we follow this classical approach.

In this section we introduce an alternative procedure based on training two different VAEs on two different datasets, the first one containing only healthy subjects and the second one containing brain scans which may or may not contain tumoural regions. Our proposed method is presented in Algorithm 1. While we expect the first autoencoder, `VAE-H`, to learn structural patterns characteristic of healthy brain tissues, the second autoencoder, `VAE`, is trained to learn more variegate representations of both tumoural and non-tumoural tissues. To detect an anomaly at test time we propose to reconstruct a given image through the model trained on healthy data and compute the distance between the original image and the reconstructed one, in the latent space of the encodings associated to the second autoencoder. Given the fact that `VAE-H` is trained on the healthy images to reconstruct the normal structure of the brain, we expect that the distance between the two encodings will be larger

for images containing abnormal regions.

Unlike other approaches in the literature that consider the discrepancy between reconstructions in the input space, our algorithm relies on the computation of a distance in the space of the latent representations of the model trained on healthy and unhealthy data.

We expect that by learning specific representations for tumoural tissues as done by `VAE`, we can better discriminate healthy and unhealthy images.

Algorithm 1: Classification of brain MRIs

Input: Let d be a distance defined over the latent space of a VAE
Data: `MRI-H`: dataset of MRI slices of healthy individuals
Data: `MRI`: dataset of MRI slices of individuals which may have tumoural tissues
Let `VAE-H` be a VAE trained on `MRI-H`
Let `VAE` be a VAE trained on `MRI`
Let x_h = `VAE-H.rec`(x) be the reconstruction of x through `VAE-H`
Let `VAE.enc`(x) be the encoding of x through `VAE`
Let $d_x = d($`VAE.enc`(x),`VAE.enc`$(x_h))$
Compute the distribution of the distances d_x, with x in the validation set of `MRI-H` and let d_* be a threshold selected based on a percentile
Compute d_x, with x in the test set of `MRI` and classify the slice healthy if $d_x < d_*$

4 Experimental Setting

In this section we provide a description of the two datasets we have used for our experiments, including aspects related to data preprocessing. Moreover, we describe the network topologies used for the two VAEs and details about the training process.

4.1 Datasets

For the dataset of healthy individuals, we have chosen the HCP dataset [14], while for the dataset of both normal and tumoural tissues we considered the BRATS-2015 dataset [10, 7].

The HCP dataset represents a mixture of several imaging modalities along with behavioural and genetic data gathered from 1,200 subjects [14]. We used in our experiments one of the subsets available for this dataset, which contains 100 T2-weighted MRI scans of unrelated subjects. For each scan, we have removed the black slices and we have kept 190 slices containing brain tissue. The training, validation and test datasets contain 70, 15, and 15 scans respectively, so that the total number of 2D images used as training data is 13,300, the total number of images used as validation, and test is 2,850 each.

The BRATS dataset is composed of a mix of pre-therapy and post-therapy multi-contrast magnetic resonance scans from glioma patients [10, 7]. Since we had access only to the training set, we have split it into train, validation, and test, with 192, 41, and 41 patients, respectively and selected the middle 130 slices from each scan. The total number of images used in train, validation, and test is therefore 24,960, 5,330, and 5,330, respectively.

We have cropped and resized the images from both datasets to 200×200 and down-sampled them to 64×64 pixels. To avoid overfitting during training the left and right hemispheres have been flipped with probability 0.5. Data augmentation such as adjustment of brightness and injection of Gaussian white noise with standard deviation equal to 0.01 has proved to be useful to further improve generalization from train to validation.

4.2 Network Architectures

We have trained two VAEs with similar network architectures and training parameters. The encoders are convolutional neural networks with channels $[64, 128, 256, 512]$ for the model trained only on healthy data and $[16, 64, 256, 1024]$ for the second model, kernel size 4×4 and strides 2, followed by a stochastic layer of independent Gaussian distributions with size 128. The decoders have output channels $[256, 128, 64]$ and $[256, 64, 16]$, respectively, output shapes $[8, 16, 32]$, kernel size 4×4 and strides 2, followed by a logit-normal distribution with a clip value for the mean equal to 0.01 and a covariance (scalar for HCP and vectorial for BRATS) with a minimum value of 0.001. For each input, 3 samples are generated in the latent space during training. The models have been trained with Adam for 200 epochs, learning rate 0.0001, and default values for the β parameters. Batch size has been set to 32. The global norm of the gradient has been clipped to 1,000 to avoid numerical instabilities.

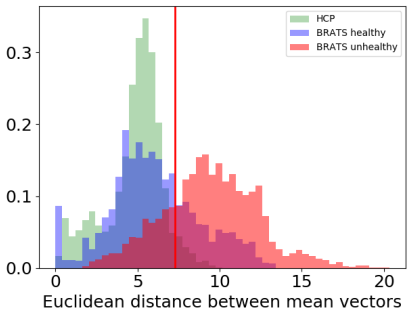

Figure 1: Histogram of the L^2 distances between the means in the space of the approximate posteriors between the two encodings on the test sets. The red line corresponds to the threshold associated to the 99% percentile of the distances computed for the HCP validation dataset.

5 Results

In this section, we present our preliminary results related to the evaluation of our proposed algorithm for the classification of brain MRI slices as being healthy or unhealthy.

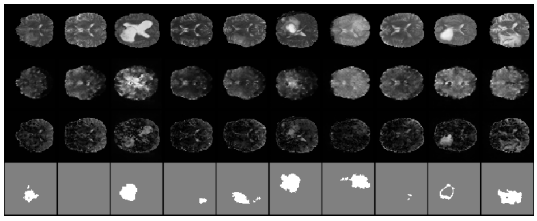

Figure 2: First row: original BRATS images from the test set. Second row: reconstruction through `VAE-H` of the original images. Third row: residual between original and reconstructed images. Fourth row: mask of the tumour.

In Algorithm 1, we have chosen as a distance function d the norm of the difference between the means obtained with `VAE` for the encodings of x and x_h. After training the two VAEs, we have computed the distribution of d_x for all x in the validation set of HCP. This allowed us to fix a value for the threshold $d_* = 7.29$, based on a 99% percentile, see Fig. 1. Next, the threshold d_* has been used to classify images from the test set of BRATS based on the value of d_x. Since not all slices of the individuals in BRATS contain tumoural tissues, in Fig. 1 we have

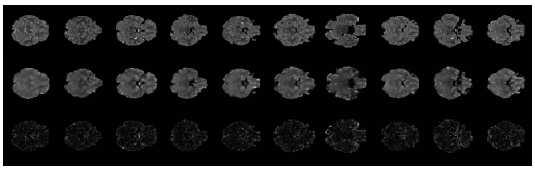

Figure 3: First row: original HCP images from the test set. Second row: reconstruction through `VAE-H` of the original images. Third row: residual between original and reconstructed images.

split the BRATS dataset into healthy and unhealthy individuals. The histograms show an overlapping for HCP and healthy BRATS, and a certain level of separability between normal and abnormal tissues. Notice that by resizing the images in BRATS to 64×64, the area of the tumour is reduced by a factor of 9.77. For this reason, to determine whether or not the slice contains tumoural tissues, we considered different thresholds for the number of pixels in the tumour mask (before the resize) to label the image as unhealthy. We evaluated the algorithm for different values of this threshold, however, we did not see a significant difference in the performance for values up to 50.

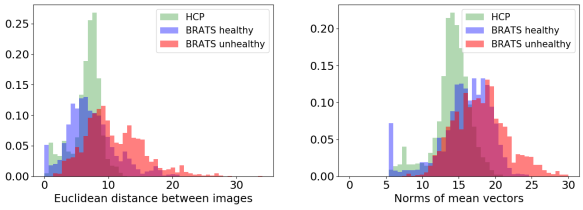

Figure 4: (left) Histogram of the L^2 distances between original and reconstructed images through `VAE-H` on test. (right) Histogram of the L^2 norms of the mean vectors of `VAE-H` on test.

In order to evaluate the quality of the reconstructions x_h, we report some examples in Fig. 2, while as a comparison we show in Fig. 3 the reconstructions of HCP through `VAE-H`. It is possible to observe that the reconstructions obtained with `VAE-H` for the HCP dataset are of acceptable quality, while for BRATS in certain cases we observe margins of improvements. The quality of the reconstructions for BRATS is determinant for our method. We expect that better reconstructions may lead to increased separability among healthy and unhealthy slices in

Fig. 1, and thus an improvement of the performance of our method. Moreover, as mentioned in [2], the overlap between healthy and unhealthy slices for BRATS can be partially explained also by the high variability present in the MRIs of healthy brains, which may be larger than the variability caused by tumoural tissues.

To validate our results, we compare with two different statistics computed using only the autoencoder trained on healthy images. The first one is given by the L^2 norm of the difference between the original image and its reconstruction using `VAE-H`. The distribution of the L^2 is represented in Fig. 4 (left) which shows a lower degree of separation between healthy and unhealthy slices for BRATS. The other statistics that we have computed is the norm of the mean vector obtained though the encoder of `VAE-H`. In Fig. 4 (right) we can see the distribution of the norms for each dataset. The BRATS dataset shows a higher variability compared to HCP, which is given by the fact that these images come from a different distribution than that of the images used during training. In particular, we can observe that unhealthy slices tend to have larger values for the norm of the mean vector, even if the difference is not significant. In both cases, our proposed method is able to better discriminate between the two classes. In Table 1 we computed the accuracy, F1 score, and Area Under the Curve for BRATS for our proposed approach as well as for the baseline computed in the input space.

Method	Accuracy	F1 score	AUC
L^2 input space	0.64	0.62	0.65
Our method	0.74	0.76	0.74

Table 1: Accuracy, F1 score, and Area Under the Curve for BRATS (test set) for L^2 distances computed in the input space versus in the latent space (our method).

6 Conclusions

In this paper, we have introduced a novel approach to semi-supervised anomaly detection for brain MRIs based on the use of two autoencoders, the first one trained on healthy individuals and the second one trained on images which include both normal and tumoural tissues. We have defined a criterion for the detection of a tumour in a slice based on the computation of a distance, in the space of the approximate posteriors of the second autoencoder, computed between the encoding of the image and the encoding of its reconstruction through the first autoencoder. In our preliminary experiments, we used the HCP and BRATS datasets, respectively, and computed distances in the latent space by evaluating the L^2 norm of the difference of the means. The results validate the goodness of our method. As expected we have observed that the performance strongly depends on the quality of the reconstructions of the autoencoders. For this reason, we believe further research should be conducted in the direction of using more powerful autoencoders compared to vanilla VAE. Another direction which has proved to provide an advantage in the context of anomaly detection is given by the use of denoising techniques, cf. [16]. Denoising improves the quality of the reconstruction of tumoural tissues through autoencoders trained only with healthy individuals, being more robust to input perturbations. One more remark is that while HCP contains images of young subjects, BRATS is mainly formed of MRIs taken from older patients, thus healthy slices may have structural differences. This is an issue which we plan to tackle by combining several datasets, such as ISLES-2015, Cam-CAN, MIDAS, and IXI.

We are currently investigating the possibility to train a single VAE where the encoder and the decoder are conditioned on the type of dataset (i.e., healthy versus both normal and tumoural tissues). Not only this approach would be more efficient since we expect some of the CNN filters to be shared between the two VAEs, but also it would allow to easily cast learning in a more general semi-supervised setting, for instance in presence of limited available annotations of unhealthy individuals.

Acknowledgements

The authors are supported by the DeepRiemann project, co-funded by the European Regional Development Fund and the Romanian Government through the Competitiveness Operational Programme 2014-2020, project ID P_37_714, SMIS code 103321, contract no. 136/27.09.2016. Data

used in the preparation of this work were obtained from the Human Connectome Project (HCP) database (https://ida.loni.usc.edu/login.jsp) and the Multimodal Brain Tumor Segmentation Challenge (BRATS) database (www.med.upenn.edu/sbia/brats2018/data.html).

References

[1] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

[2] X. Chen and E. Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *Medical Imaging for Deep Learning (MIDL 2018)*, 2018.

[3] 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:1806.05452*, 2018.

[4] H. Choi, H. Kang, D. S. Lee, A. D. N. Initiative, et al. Predicting aging of brain metabolic topography using variational autoencoder. *Frontiers in aging neuroscience*, 10:212, 2018.

[5] T. Drew, M. L.-H. Võ, and J. M. Wolfe. The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological science*, 24(9):1848–1853, 2013.

[6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[7] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler. The virtual skeleton database: an open access repository for biomedical research and collaboration. *Journal of medical Internet research*, 15(11):e245, 2013.

[8] K. W. Lau. Representation learning on brain mr images for tumor segmentation (master thesis), 2018.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017.

[10] 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*, 34(10):1993–2024, 2014.

[11] A. Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[12] 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 Machine Learning*, pages 1278–1286, 2014.

[13] T. Tashiro, T. Matsubara, and K. Uehara. Deep neural generative model for fmri image based diagnosis of mental disorder. In *International Symposium on Nonlinear Theory and its Applications (NOLTA)*, 2017.

[14] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[15] Q. Zhao, E. Adeli, N. Honnorat, T. Leng, and K. M. Pohl. Variational autoencoder for regression: Application to brain aging analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 823–831. Springer International Publishing, 2019.

[16] D. Zimmerer, S. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *Medical Imaging with Deep Learning (MIDL 2019)*, 2019.

[17] D. Zimmerer, J. Petersen, S. A. Kohl, and K. H. Maier-Hein. A case for the score: Identifying image anomalies using variational autoencoder gradients. *arXiv preprint arXiv:1912.00003*, 2019.