BRAIN LESION DETECTION USING A ROBUST VARIATIONAL AUTOENCODER AND TRANSFER LEARNING

Haleh Akrami[⋆]*, Anand A. Joshi*[⋆]*, Jian Li*[⋆]*, Sergul Aydore*[†] *and Richard M. Leahy*[⋆]

[⋆] Signal and Image Processing Institute, University of Southern California, Los Angeles
[†] Electrical and Computer Engineering, Stevens Institute of Technology, NJ, USA

ABSTRACT

Automated brain lesion detection from multi-spectral MR images can assist clinicians by improving sensitivity as well as specificity. Supervised machine learning methods have been successful in lesion detection. However, these methods usually rely on a large number of manually delineated images for specific imaging protocols and parameters and often do not generalize well to other imaging parameters and demographics. Most recently, unsupervised models such as autoencoders have become attractive for lesion detection since they do not need access to manually delineated lesions. Despite the success of unsupervised models, using pre-trained models on an unseen dataset is still a challenge. This difficulty is because the new dataset may use different imaging parameters, demographics, and different pre-processing techniques. Additionally, using a clinical dataset that has anomalies and outliers can make unsupervised learning challenging since the outliers can unduly affect the performance of the learned models. These two difficulties make unsupervised lesion detection a particularly challenging task. The method proposed in this work addresses these issues using a two-prong strategy: (1) we use a robust variational autoencoder model that is based on robust statistics, specifically the β-divergence that can be trained with data that has outliers; (2) we use a transfer-learning method for learning models across datasets with different characteristics. Our results on MRI datasets demonstrate that we can improve the accuracy of lesion detection by adapting robust statistical models and transfer learning for a variational autoencoder model.

Index Terms— variational autoencoders, lesion detection, robust variational autoencoders, brain imaging, unsupervised machine learning, anomaly detection

1. INTRODUCTION

Accurate detection of lesions in the human brain is crucial for early diagnosis and treatment. Medical imaging techniques, such as MRI are now standard clinical tools for detecting and quantifying lesions. Humans excel in identifying lesions by visual inspection after extensive training, but the subjective and expensive nature of human detection and delineation makes the machine learning methods an attractive alternative or complement. Furthermore, machine learning might be able to achieve better-than-human performance for this specific task by leveraging multispectral MRI. Research based on supervised machine learning has already achieved significant success [1, 2, 3] with human-level or better performance. However, large numbers of manual lesion delineations are required for training supervised methods. Unsupervised approaches, on the other hand, do not require labeled data but generally are less accurate.

Unsupervised approaches such as the autoencoder and variational autoencoder (VAE) [4] and their variants [5] have shown that we can approximate the underlying distributions of high-dimensional data. A common application of unsupervised approaches is outlier detection [6], where the goal is to identify data samples whose representation deviates from the normal samples. For a population of brain images, assuming that lesions and other abnormalities occur rarely and in different locations across subjects, we conjecture that it is possible to learn the distribution that reflects a healthy brain structure using a VAE. Once this distribution is learned, we can measure the reconstruction error between a given image and the reconstructed image to identify and localize abnormalities in that image.

A VAE is a probabilistic autoencoder that uses the variational lower bound of the marginal likelihood of data as the objective function. It has been shown that VAEs achieve higher accuracy in lesion detection tasks than standard autoencoder [7, 8, 9]. VAEs are based on the assumption that the training dataset and the test dataset are sampled from the same distribution. However, this assumption may not hold in real-world settings such as medical imaging applications since different datasets can use different acquisition and pre-processing techniques. Ideally, we should still be able to leverage a pre-trained VAE model to develop a new model that adapts to our dataset. The topic of transfer learning focuses on addressing this problem [10]. With the aid of transfer learning, it is possible to store the knowledge gained while solving one problem and apply it to a different problem.

The VAE's objective function contains the KL-divergence

This work is supported by the following grants: R01 NS074980, W81XWH-18-1-0614, R01 NS089212, and R01 EB026299.

term which does not cope well with outliers and is therefore not robust. This may lead to unintended effects in applying transfer learning for adapting pre-trained VAE models when the characteristics of the new dataset differ significantly from that of the initial training dataset. To this end, we propose the use of robust VAE based on the notion of β-divergence from robust statistics [11] for applying transfer learning from pre-trained unsupervised lesion detection models. By varying the robustness hyperparameter β, we can control how much influence is granted to samples with low probability. We demonstrate the effectiveness of our approach on brain MRI datasets. Our results show that the combination of robust VAE and transfer learning allows us to use training data that has different imaging parameters and demographics than that of the test dataset. We demonstrate this using a quantitative comparison to VAE models.

2. MATHEMATICAL FORMULATION

In this section, we first present a summary of VAEs and robust variational inference. Then we formulate a robust VAE that can be trained on a mixture of normal and lesion images based on the assumption that the lesion-free images are drawn from a Gaussian distribution.

2.1. Variational Autoencoder

The VAE is a directed probabilistic graphical model whose posteriors are approximated by a neural network. Let $\vec{X}$ denote the input data, $\vec{x}^{(i)}$ denote the samples of $\vec{X}$, and $\vec{Z}$ denote its low-dimensional latent representation. The VAE consists of an encoder network that computes an approximate posterior $q_\phi(\vec{Z}|\vec{X})$, and a decoder network that computes $p_\theta(\vec{X}|\vec{Z})$ [4] and $p_\theta(\vec{Z})$ denotes the prior distribution which z is generated from. The model parameters ϕ and θ are found by maximizing the evidence lower bound (ELBO) function [4]:

$$\begin{aligned} L(\theta, \phi; \vec{x}^{(i)}) =& E_{q_\phi(\vec{Z}|\vec{x}^{(i)})}[\log(p_\theta(\vec{x}^{(i)}|\vec{Z}))] \\ & - D_{KL}(q_\phi(\vec{Z}|\vec{x}^{(i)})||p_\theta(\vec{Z})). \end{aligned} \tag{1}$$

The first term (log-likelihood) can be interpreted as the *reconstruction loss* and the second term (KL divergence) as the *regularizer*. Using empirical estimates of expectation, we form the Stochastic Gradient Variational Bayes cost [4]:

$$\begin{aligned} L(\theta, \phi; \vec{x}^{(i)}) \approx& \frac{1}{S}\sum_{j=1}^{S} \log(p_\theta(\vec{x}^{(i)}|\vec{z}^{(j)})) \\ & - D_{KL}(q_\phi(\vec{Z}|\vec{x}^{(i)})||p_\theta(\vec{Z})), \end{aligned} \tag{2}$$

where S is the number of samples drawn from $q_\phi(\vec{Z}|\vec{X})$. In practice, we can choose $S = 1$ as long as the minibatch size is large enough.

Assuming $p_\theta(\vec{X}|\vec{Z})$ is a Gaussian distribution and the output of the network is the mean of this distribution, the log-likelihood term simplifies to the mean-squared-error.

2.2. Robust Variational Autoencoder

Robust variational inference is based on the $\beta-$ELBO based loss function and replaces the log-likelihood term with β-divergence which is equivalent to minimizing β-cross entropy [11, 12]. The $\beta - ELBO$ function is given by:

$$\begin{aligned} L_\beta(q, \theta) =& - NE_{q_\phi(\vec{Z}|\vec{x}^{(i)})}[(H_\beta(\hat{p}(\vec{X})||p_\theta(\vec{X}|\vec{Z})))] \\ & - D_{KL}(q(\vec{Z})||p_\theta(\vec{Z})). \end{aligned} \tag{3}$$

where $p_\theta(\vec{Z}|\vec{X})$ is posterior distribution, the empirical distribution is $\hat{p}(\vec{X}) = \frac{1}{N}\sum_{i=1}^{N} \delta(\vec{X}, \mathbf{x}^{(i)})$ where δ is the Dirac delta function and $\vec{Z}$ represents the latent variable, N is the number of samples, and θ contains the generative model's parameters. The β-cross entropy is given by [12]:

$$\begin{aligned} & H_\beta(\hat{p}(\vec{X})||p_\theta(\vec{X}|\vec{Z})) = \\ & -\frac{\beta+1}{\beta}\int \hat{p}(\vec{X})(p_\theta(\vec{X}|\vec{Z})^\beta - 1)d\vec{X} + \int p_\theta(\vec{X}|\vec{Z})^{\beta+1}d\vec{X}. \end{aligned} \tag{4}$$

By replacing log-likelihood with β-cross entropy in the VAE formulation, we obtain a new cost function which is robust to outliers [13]. For a Gaussian distribution, the $\beta-$ELBO-cost of RVAE for the j^{th} sample simplifies to [13]:

$$\begin{aligned} & L_\beta(\theta, \phi; \vec{x}^{(i)}) = \\ & \frac{\beta+1}{\beta}\left(\frac{1}{(2\pi\sigma^2)^{\beta D/2}} \exp\left(-\frac{\beta}{2\sigma^2}\sum_{d=1}^{D} ||\hat{\vec{x}}_d^{(j)} - \vec{x}_d^{(i)}||^2\right) - 1\right) \\ & \quad - D_{KL}(q_\phi(\vec{Z}|\vec{x}^{(i)})||p_\theta(\vec{Z})). \end{aligned} \tag{5}$$

Similar to the VAE, we use stochastic gradient variational bayes cost minimization using sampling to optimize β-ELBO to train the robust VAE.

Next, we describe the use of VAE and robust VAE in combination with transfer learning for lesion delineation tasks.

3. THE MODEL AND EXPERIMENTS

We used the VAE architecture proposed in [14] that consists of three consecutive blocks of convolutional layer, a batch normalization layer, a rectified linear unit (ReLU) activation function and two fully-connected layers in the bottleneck for the encoder and a fully-connected layer and three consecutive blocks of deconvolutional layers, a batch normalization layer and ReLU, and a final deconvolutional layers for the decoder. The size of the input layer is $3 \times 64 \times 64$.

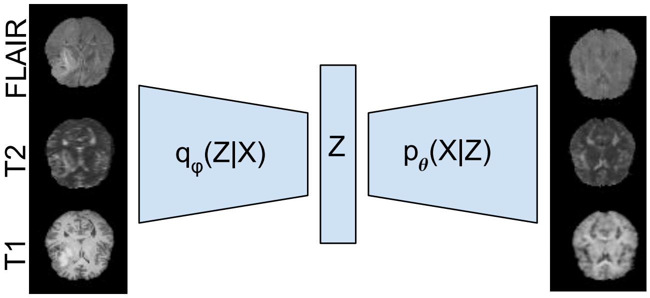

Fig. 1. VAE network and input, output sample for ISLES dataset

3.1. Data and Preprocessing

For the initial training, we used 20 central axial slices of brain MRI datasets from a combination of 119 subjects from the Maryland MagNeTS study [15] of neurotrauma and 112 subjects of TrackTBI-Pilot [16] dataset, both available for download from `https://fitbir.nih.gov`. We used 2D slices rather than 3D images to make sure we had a large enough dataset for training the VAE. These datasets contain T1, T2 and FLAIR images for each subject, and have sparse lesions. The three imaging modalities (T1, T2, FLAIR) were rigidly coregistered within subject and to the MNI atlas reference, and re-sampled to 1mm isotropic resolution. Skull and other non-brain tissue were removed using BrainSuite (`https://brainsuite.org`). Subsequently, we reshaped each sample into 64×64 dimensional images and performed histogram equalization to a lesion free subject that intensity-normalized by the value of the 99th percentile voxel. We used 191 subjects for training and 40 subjects for validation randomly sampled from MagNeTS and TrackTBI-Pilot datasets.

Experiments for pre-trained model: In this experiment, we evaluate the performance of a pre-trained model on a dataset that was pre-processed similarly to the training set. We used 20 central axial slices of 15 subjects from the ISLES (The Ischemic Stroke Lesion Segmentation) database [17] as a test set and performed similar pre-processing as for the training set.

Experiments for re-training models (VAEbr, RVAEbr): In this experiment, we re-train VAE and RVAE models from scratch using a combination of the initial dataset and an additional 20 independent subjects from the BRATS dataset (`https://www.smir.ch/BRATS/Start2015`). We used 20 central axial slices from the rest of the 20 subjects of BRATS 2015 as test data.

Experiments for transfer learning (PreVAE, PreRVAE): In this final experiment, we assume that we only have access to the pre-trained models but the training datasets used for pre-trained models are not available. We updated the pre-trained models using 20 subjects from the BRATS 2015 dataset. Similar to the experiments for re-training the models, we tested the updated models on 20 central axial slices from 20 subjects of the BRATS 2015 dataset.

3.2. Results

The absolute error maps between reconstructed and original images were computed for segmentation of the lesions. A median filter of size 7x7 was applied to remove isolated pixels. The filtered lesion error maps were used to plot ROC (Receiver Operating Characteristic) curves from which we computed the AUC (Area Under The Curve) Hand-traced lesions were used to define ground truth. Only the pixels inside the brain mask were used for AUC computation. A example input image from the ISLES test dataset and its reconstruction using the pre-trained VAE model is shown in Figure 1. The AUC for this experiment was 0.93.

Experimental results of re-training the models and using transfer learning are illustrated in Figure 2 with the ROC curves and AUC values shown in Figure 3. Figure 2A shows that RVAE did not reconstruct the lesions while the lesions are more apparent in the reconstructed images from the VAE model. As a result, it can be concluded that the RVAE can capture the locations of the lesions more accurately by computing the error between original and reconstructed images. The AUC of the pre-trained VAE was 0.75. When the VAE is re-trained from scratch using the BRATS dataset (VAEbr), the AUC has increased to 0.9. However, the value of AUC decreased to 0.82 when transfer learning is applied to the pre-trained VAE model (PreVAE).

The AUC of the RVAE model that was re-trained using the initial and the BRATS datasets (RVAEbr) was 0.92. The AUC increased to 0.93 when transfer learning was applied to the RVAE model (PreVAE).

The values of beta for these experiments were chosen using the validation dataset. We chose a beta value that prevents RVAE from reconstructing lesions in validation dataset.

4. DISCUSSION AND CONCLUSION

After training the VAE using nominally normal (anomaly free) data, we can use it for anomaly detection and specifically for identification of abnormal structures in medical images. We focused on delineating lesions from MRI scans which might have differing characteristics and pre-processing. This causes degradation in the performance of VAE. Utilizing the robustness of RVAE, we described a framework that enables us to fine-tune the model for new test sets with differing specific attributes. We used a pre-trained model and re-trained it with the additional subjects from the new dataset for model refinement. The robustness of RVAE forces the model to only learn common features between these data samples instead of their anomalous features (lesions). We have shown quantitatively and qualitatively that RVAE outperforms VAE both before and after model refinement. A previous study on the BRATS 2015 dataset [7] reported AUC of 0.9 using VAE.

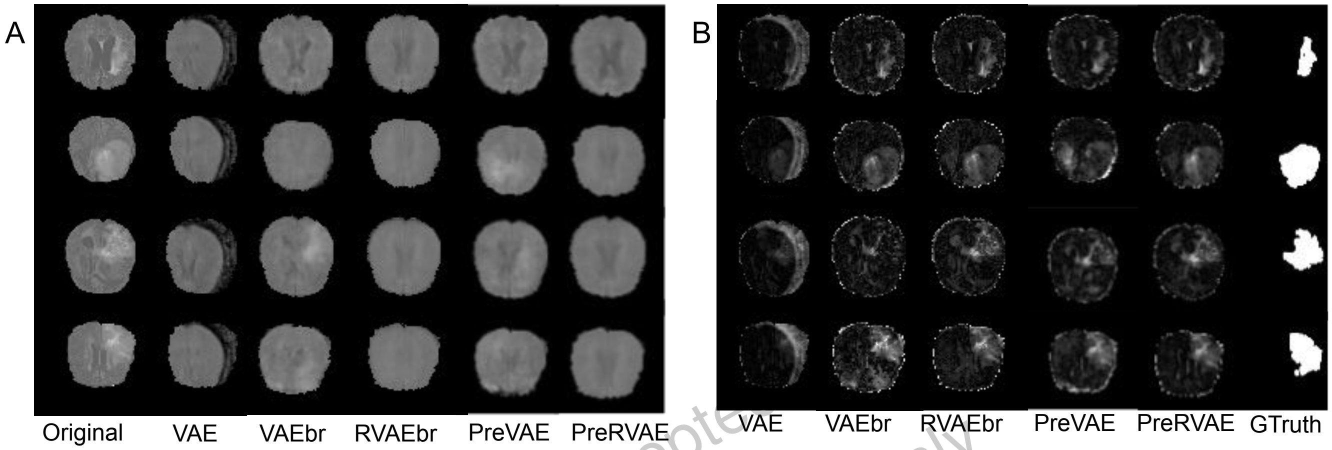

Fig. 2. (A) Original and reconstructed test images using different models. (B) Absolute reconstruction error of the test images and associated hand-delineated lesions (GTruth). VAEbr: VAE model re-trained from scratch using the initial data and the BRATS samples, RVAEbr: RVAE model re-trained from scratch using the initial data and BRATS samples, PreVAE: transfer learning of VAE from the pre-trained VAE model using additional BRATS samples, PreRVAE: transfer learning of RVAE from the pre-trained VAE model using additional BRATS samples.

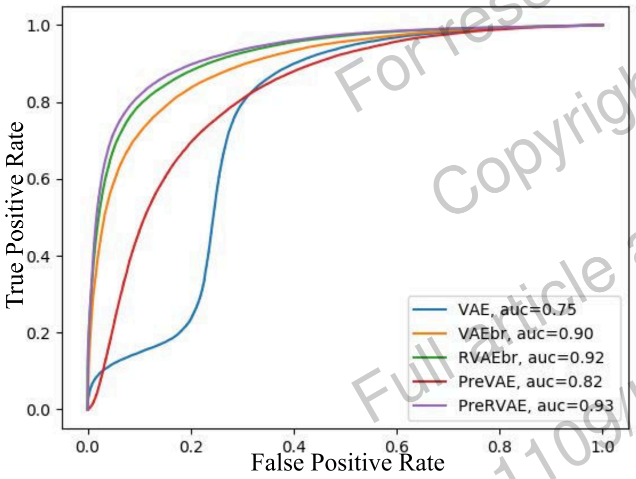

Fig. 3. ROC curves of different models. RVAE outperforms VAE both when trained from scratch using BRATS samples in addition to the initial data (RVAEbr vs VAEbr) and when updated using the pre-trained models (PreRVAE vs PreRVAE).

We achieved a similar level of performance by using only a subset of this dataset and a pre-trained model from a different dataset.

5. REFERENCES

[1] Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze, "Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images," *NeuroImage*, vol. 183, pp. 650–665, 2018.

[2] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[3] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[4] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[6] Charu C Aggarwal, "Outlier analysis," in *Data mining*. Springer, 2015, pp. 237–263.

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

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

[9] Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al., "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders," *OpenReview*, 2018.

[10] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010.

[11] Futoshi Futami, Issei Sato, and Masashi Sugiyama, "Variational inference based on robust divergences," *arXiv preprint arXiv:1710.06595*, 2017.

[12] Andrzej Cichocki and Shun-ichi Amari, "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[13] Haleh Akrami, Anand A Joshi, Jian Li, and Richard M Leahy, "Robust variational autoencoder," *arXiv preprint arXiv:1905.09961*, 2019.

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

[15] Rao P Gullapalli, "Investigation of prognostic ability of novel imaging markers for traumatic brain injury (tbi)," Tech. Rep., BALTIMORE UNIV MD, 2011.

[16] John K Yue, Mary J Vassar, Hester F Lingsma, Shelly R Cooper, David O Okonkwo, Alex B Valadka, Wayne A Gordon, Andrew IR Maas, Pratik Mukherjee, Esther L Yuh, et al., "Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury," *Journal of neurotrauma*, vol. 30, no. 22, pp. 1831–1844, 2013.

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