Uncertainty Quantification using Variational Inference for Biomedical Image Segmentation

Abhinav Sagar*
Vellore Institute of Technology
Vellore, Tamil Nadu, India
abhinavsagar4@gmail.com

Abstract

Deep learning motivated by convolutional neural networks has been highly successful in a range of medical imaging problems like image classification, image segmentation, image synthesis etc. However for validation and interpretability, not only do we need the predictions made from the model but also how confident it is while making those predictions. This is important in safety critical applications for the people to accept it. In this work, we used an encoder decoder architecture based on variational inference techniques for segmenting brain tumour images. We compared different backbones architectures like U-Net, V-Net and FCN as sampling data from the conditional distribution for the encoder. We validated our work on BRATS dataset using Dice Similarity Coefficient and Intersection Over Union as the evaluation metrics. Our model achieves state of the art results while making use of a principled way of uncertainty quantification.

1 Introduction

Medical image segmentation is a challenging task for medical practitioners. It is costly, takes time and is prone to error. Hence there is a need to automate the manually done segmentation. Lately Neural Networks have shown great potential on a variety of medical image segmentation problems. The challenge with the approaches used in literature is that it doesn't predict the uncertainty associated with the model predictions. This is where Bayesian methods come into play as it gives a principled way of measuring uncertainty from the model predictions. Measuring uncertainty in the output predictions made by neural networks is important for interpretation and validation. Rather than learning the point estimates, Bayesian Neural Networks (BNN) learns the distribution over the weights. The training process of BNN involves first initializing the parameters of the neural network. Next the weights are sampled from some distribution (like gaussian with zero mean and unit variance) and both the forward pass and backward pass is done to update the weights using the conventional backpropagation algorithm.

Monte Carlo dropout networks (Kingma et al., 2015) use dropout layers to approximate deep Gaussian processes which still lack theoretical understanding. Bayesian Convolutional Neural Network (Gal et al., 2015) use variational inference to learn the posterior distribution over the weights given the dataset. The problem with this approach is that it requires a lot of computation involving a lot of parameters, making this technique not scalable in practice.

Variational Autoencoder (Kingma and Welling, 2014) which is based on generative models solves the above problems and has been successful in a number of tasks like generating images, texts, recommender systems etc. This approach comes with several challenges in its own right which have been successfully tackled in the literature. A random variable sampled from posterior distribution has

*Website of author - https://abhinavsagar.github.io/

Preprint. Under review.

no gradient so the conventional backpropagation techniques can't be applied to it. Local Reparameterization Trick (Kingma et al., 2015) was proposed to tackle this by converting the random variable to a deterministic one for computation. The second challenge was the huge computational requirement since it required weight updates in every iteration. Bayes by Backprop algorithm (Blundell et al., 2015) tackled this by calculating gradients in back-propagation using a scale and shift approach by updating the posterior distribution in the backward pass.

2 Related Work

2.1 Medical Image Segmentation

The problem of segmenting medical images have been successfully tackled in literature using mainly two techniques, first using a Fully Convolutional Network (FCN) (Long et al., 2014) and second those which are based on U-Net (Ronneberger et al., 2015). The main characteristic of FCN architectures is that it doesn't use fully connected layers at the end which have been used successfully for image classification problems. U-Net on the other hand uses an encoder-decoder architecture with pooling layers in the encoder and upsampling layers in the decoder. Skip connections connect the encoder layers to the decoder layer to create an additional path for the flow of gradients back in the backpropagation step. This helps in reducing overfitting due to many parameters involved while training the network.

2.2 Bayesian Neural Network

Lately, there has been a revival of interest in bayesian methods as some of the inherent problems with deep learning could be solved using it. It is a scalable approach of avoiding overfitting in neural networks and at the same time gives us a measure of uncertainty. This is very important in critical applications where not only do we require the predictions made from the model, but also how confident it is while making its predictions. BNN can be considered as an ensemble of neural networks (Gal et al., 2016). It has two advantages over the standard neural networks, first it avoids overfitting and second it gives a measure of uncertainty involved.

Instead of point estimates, the neural network learns posterior distribution over the weights given the dataset as given in the equation below.

$$p(\omega|\mathcal{D}) = \frac{p(\mathcal{D}|\omega)p(\omega)}{p(\mathcal{D})} = \frac{\prod_{i=1}^{N} p\left(y_i|x_i,\omega\right)p(\omega)}{p(\mathcal{D})} \tag{1}$$

The predictive distribution can be calculated by approximating the integral as shown in equation below.

$$p\left(y^*|x^*,\mathcal{D}\right) = \int_{\Omega} p\left(y^*|x^*,\omega\right)p(\omega|\mathcal{D})d\omega \tag{2}$$

The challenge is that the posterior is often intractable in nature. To combat this, (Neal et al., 1993) used Markov Chain Monte Carlo (MCMC) for learning the weights over the bayesian neural networks. Also [Graves, 2011, Blundell et al., 2015, Louizos and Welling, 2016, 2017a] proposed independently a technique using variational inference techniques for approximating the posterior distributions. KL Divergence between the posterior and the true distribution can be calculated using the equation below.

$$KL\left\{q_\theta(\omega)\|p(\omega|\mathcal{D})\right\} := \int_{\Omega} q_\theta(\omega)\log\frac{q_\theta(\omega)}{p(\omega|\mathcal{D})}d\omega \tag{3}$$

Alternatively minimizing the KL divergence can be written in another form by maximizing the Evidence Lower Bound (ELBO) which is tractable. This is shown in the equation below.

$$-\int_{\Omega} q_\theta(\omega)\log p(y|x,\omega)d\omega + KL\left\{q_\theta(\omega)\|p(\omega)\right\} \tag{4}$$

2.3 Variational Inference

Variational inference finds the parameters of the distribution by maximizing the Evidence Lower Bound. ELBO consists of sum of two terms Kullback-Leibler (KL) divergence between two distributions and the negative log-likelihood (NLL). The KL divergence term which has to be minimized is shown in the equation below.

$$\min \mathrm{KL}\left(q_\theta(w)\right) \| p(w|\mathcal{D})) \tag{5}$$

The KL divergence is defined as shown in the equation below.

$$\mathrm{KL}(q(x)) \| p(x)) = -\int q(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{6}$$

The posterior in the above equation contains an integral which is intractable in nature. The equation can be re written as shown in the equation below.

$$\begin{aligned} \mathrm{KL}\left(q_\theta(w)\right) \| p(w|\mathcal{D})) &= \mathbb{E}_{q_\theta(w)} \log \frac{q_\theta(w) p(\mathcal{D})}{p(\mathcal{D}|w) p(w)} = \\ &= \log p(\mathcal{D}) + \mathbb{E}_{q_\theta(w)} \log \frac{q_\theta(w)}{p(w)} - \mathbb{E}_{q_\theta(w)} \log p(\mathcal{D}|w) \\ &= \log p(\mathcal{D}) - \mathcal{L}(\theta) \end{aligned} \tag{7}$$

It can be decomposed into two parts one of which is the KL divergence between the exact posterior and its variational approximation which needs to be minimized and the second is ELBO which needs to be maximized. This is shown in the equation below.

$$\max_\theta \log p(\mathcal{D}) = \max_\theta \left[\mathrm{KL}\left(q_\theta(w)\right) \| p(w|\mathcal{D})) + \mathcal{L}(\theta)\right] \tag{8}$$

KL divergence is zero if exact posterior is equal to variational approximation. Since the KL divergence is always greater than zero hence the equation can be approximated by maximizing only the ELBO (Welling et al., 2015) as shown in equation below.

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(w)} \log p(\mathcal{D}|w) - \mathbb{E}_{q_\theta(w)} \log \frac{q_\theta(w)}{p(w)} = \mathcal{L}_\mathcal{D} - \mathrm{KL}\left(q_\theta(w) \| p(w)\right) \tag{9}$$

2.4 Aleatoric uncertainty and epistemic uncertainty

There are two types of uncertainty - aleatory and epistemic uncertainty where variance is the sum of both these. Bayesian Neural Networks can be considered an ensemble of neural networks initialized randomly which averages the test results in parallel (Gal et al., 2016). For final predictions, single mean and variance can be estimated as shown in the equation below.

$$\mu_c(x) = \frac{1}{M} \sum_{i=1}^{M} \hat{\mu}_i(x) \tag{10}$$

$$\hat{\sigma}_c^2(x) = \frac{1}{M} \sum_{i=1}^{M} \tilde{\sigma}_i^2(x) + \left[\frac{1}{M} \sum_{i=1}^{M} \hat{\mu}_i^2(x) - \hat{\mu}^2(x)\right] \tag{11}$$

The first term in variance denotes aleatoric uncertainty while the second denotes epistemic uncertainty. Bayesian Neural Network model for uncertainty estimation was done by (Kendall et al., 2017) with

the last layer representing the mean and variance of logits. The predictive distribution approximating the posterior distribution which gives a measure of uncertainty is shown in the equation below.

$$q_{\hat{\theta}}\left(y^{*}|x^{*}\right)=\int_{\Omega}p\left(y^{*}|x^{*},\omega\right)q_{\hat{\theta}}(\omega)d\omega \tag{12}$$

Aleatoric uncertainty is a measure of the variability of the predictions from the dataset hence it is inherent in the data present. Epistemic uncertainty on the other hand is a measure of the variability of predictions from the model which is tied to various metrics used for evaluation like accuracy, loss etc.

3 Proposed method

The prior distribution helps to incorporate learning of the weights over the network. Variational Autoencoder has been used successively as a kind of generative model by sampling from the prior distribution in the encoder. The decoder uses the mean vector and standard deviation vector from the latent space to reconstruct the input.

Our model uses a similar encoder decoder architecture as that used in VAEs with the input to the encoder coming from a pre trained image segmentation architecture. We tried different backbones which have enjoyed success and found original U-Net gave the best results. The input to the encoder only needs the mean the standard deviation vectors of the conditional distribution expressing the confidence with which the pixels are correctly predicted. After passing through the encoder, the parameters get converted to a latent representation which is again sampled in a mean and standard deviation vector. The decoder later recovers this back to the original distribution. The conventional backpropagation algorithm is used for training the model with gradient descent. The model architecture is shown in Fig 1.

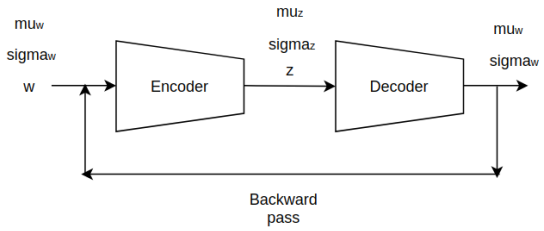

Figure 1: Our model architecture

Our model uses a similar approach which is trained using Gradient Descent is presented below:

Algorithm 1: Uncertainty Quantification using Variational Inference for Biomedical Image Segmentation

Input: Dataset $\mathcal{D}:\{(x_i,y_i)\}_{i=1}^{N}$
Input: Variational approximation of the posterior distribution $q_\theta(w)$
Input: encoder $r_\psi(z|w)$ and decoder $p_\phi(w)$
while *not converged do* **do**
  Sample minibatch: $\mathcal{D}^{*}:\{(x_i,y_i)\}_{i=1}^{M}$
  Sample weights with reparametrization: $\widehat{w}^{(i)}\sim q_{\theta_i}\left(w^{(i)}\right)$
  Sample latent variables with reparametrization: $\widehat{z}^{(i)}\sim r_{\psi^{(i)}}\left(z|\widehat{w}^{(i)}\right)$
  Compute stochastic gradients of the objective:
  $\mathcal{L}^{\text{approx}}=\mathcal{L}_{\mathcal{M}}+\sum_i\left[-\log q_{\theta_i}\left(\widehat{w}^{(i)}\right)\right)-\log r_{\psi^{(i)}}\left(\widehat{z}|\widehat{w}^{(i)}\right)+\log p(\widehat{z})+\log p_{\phi^{(i)}}\left(\widehat{w}^{(i)}|\widehat{z}\right)]$
  Update parameters $\theta=\theta+\alpha\nabla_\theta\mathcal{L}$ and $\psi=\psi+\beta\nabla_\psi\mathcal{L}$
  $g_\phi\leftarrow\frac{1}{m}\sum_{k=1}^{m}\nabla_\phi\log p_\phi\left(x^{(k)}|z_\theta\left(x^{(k)},\epsilon^{(k)}\right)\right)$
end
Output: $q_\theta(w)-$ posterior distribution of the model parameters

3.1 Datasets

To validate the performance of our proposed approach to generalization, publicly available datasets were used for brain tumour segmentation BRATS18 (Menze et al., 2015, Bakas et al., 2017). It contains MRI scans of 175 patients with glioblastoma and lower grade glioblastoma. The images were of resolution 240 * 240 * 155 pixels. The ground truth labels were created by expert neuroradiologists. The sample from the dataset is shown in Fig 2.

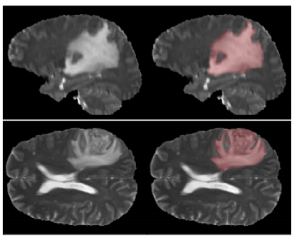

Figure 2: Example of MRI slices and ground truth segmentation

3.2 Hyperparameters

The hyperparameters used in our model are specified in Table 1.

Table 1: Hyperparameters details

Parameter	Value
Batch Size	16
Optimizer	Adam
Learning Rate	0.001
Latent Variable Size	10

In addition to the above hyperparamaters, cyclical learning rate schedulers and ReduceLROnPlateau was used. In gradient descent, the value of momentum was taken as 0.9, gamma value of 0.1 and weight decay of 0.0005.

3.3 Evaluation

Evaluation metrics for semantic segmentation problems which have often been used in literature are Dice Similarity Coefficient (DSC) also known as F1-score and Intersection over union (IoU) as proposed by(Clèrigues et al., 2018, Kao et al., 2018, Myronenko, 2018, Deniz et al., 2018). The corresponding equations are shown below.

$$\mathrm{DSC} = \frac{2TP}{2TP + FN + FP} \tag{13}$$

$$\mathrm{IoU} = \frac{TP}{TP + FN + FP} \tag{14}$$

True positive (TR), false negative (FN) and false positive (FP) number of pixels is calculated separately for each image and averaged over the test set. The ground truth is labelled manually by experts which are compared against.

3.4 Loss

A combination of binary cross entropy and dice losses have been used to train the network. The first part binary cross entropy is a commonly used loss function for classification problems as shown by

(Goodfellow et al., 2016). Every pixel in the image needs to be classified and hence loss function can be written as shown in the equation below.

$$\mathcal{L}_{CE} = -\sum_{i,j} y_{i,j} \log \widehat{y}_{i,j} + (1 - y_{i,j}) \log (1 - \widehat{y}_{i,j}) \tag{15}$$

The problem with binary cross entropy loss is that it doesn't take into account the class imbalance as the background is the dominant class. This is one of fundamental challenges in semantic segmentation problems. Dice Loss is robust to the aforementioned problem which is based on Dice Similarity Coefficient as defined below.

$$\mathcal{L}_{DICE} = \sum_{i=1}^{N} \frac{FN_i + FP_i}{2TP_i + FN_i + FP_i} = \sum_{i=1}^{N} \left(1 - DSC^{(i)}\right) \tag{16}$$

Both the loss terms were combined in a single term with more weight given to the Dice Loss term since it handles the class imbalance problem better. This is defined using the equation below.

$$\mathcal{L} = 0.9 \cdot \mathcal{L}_{DICE} + 0.1 \cdot \mathcal{L}_{CE} \tag{17}$$

4 Results

The Mean Dice Similarity value for various backbone architectures compared against different train size values are shown in Table 2.

Table 2: Mean Dice Similarity metrics for the experiments

Train Size	UNet	VNet	FCN
5	53.1	50.6	50.2
10	56.6	51.3	52.1
15	60.8	53.8	54.4
20	64.3	56.5	58.9

The IOU value for various backbone architectures compared against different train size values are shown in Table 3.

Table 3: Intersection over Union metrics for the experiments

Train Size	UNet	VNet	FCN
5	48.4	46.7	47.2
10	50.6	48.8	50.4
15	53.1	50.8	52.6
20	55.8	52.8	54.3

The results of uncertainty involved in segmentation is shown in Fig 3.

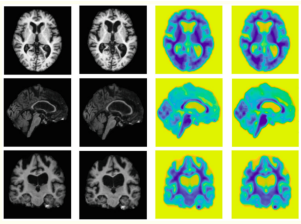

Figure 3: Uncertainty in Segmentation

5 Conclusions

In this work, we presented a way to quantify uncertainty in the context of medical image segmentation. Our model is based on an encoder decoder framework similar to that used by VAEs. The weights of the network represent distributions instead of point estimates and thus give a principled way of measuring uncertainty at the same time while making the predictions. Our model uses bayesian neural networks for both the encoder and decoder. The inputs to encoder come from backbones like U-Net, V-Net, FCN sampled from conditional distribution representing the confidence with which pixels are labelled correctly. We validated our results on publicly available BRATS dataset with our model achieving state of the art results on DSC and IOU metrics.

Acknowledgments

We would like to thank Dhruv Raj and Shreya Singh for their helpful comments on this work. We would also like to thank Nvidia for providing the GPUs.

References

1. Veredas, Francisco, Héctor Mesa, and Laura Morente. "Binary tissue classification on wound images with neural networks and bayesian classifiers." IEEE transactions on medical imaging 29.2 (2009): 410-427.

2. Ma, Jingting, et al. "A novel Bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018.

3. Li, Yingzhen, and Yarin Gal. "Dropout inference in Bayesian neural networks with alpha-divergences." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

4. Kwon, Yongchan, et al. "Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation." (2018).

5. Bromiley, Paul A., et al. "Bayesian and non-Bayesian probabilistic models for medical image analysis." Image and Vision Computing 21.10 (2003): 851-864.

6. Ayhan, Murat Seckin, and Philipp Berens. "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks." (2018).

7. Leibig, Christian, et al. "Leveraging uncertainty information from deep neural networks for disease detection." Scientific reports 7.1 (2017): 1-14.

8. Jiang, Jianmin, P. Trundle, and Jinchang Ren. "Medical image analysis with artificial neural networks." Computerized Medical Imaging and Graphics 34.8 (2010): 617-631.

9. Saad, Ahmed, Torsten Möller, and Ghassan Hamarneh. "ProbExplorer: Uncertainty-guided Exploration and Editing of Probabilistic Medical Image Segmentation." Computer Graphics Forum. Vol. 29. No. 3. Oxford, UK: Blackwell Publishing Ltd, 2010.

10. Wang, Guotai, et al. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks." Neurocomputing 338 (2019): 34-45.

11. Nair, Tanya, et al. "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation." Medical image analysis 59 (2020): 101557.

12. Saad, Ahmed, Ghassan Hamarneh, and Torsten Moller. "Exploration and visualization of segmentation uncertainty using shape and appearance prior information." IEEE Transactions on Visualization and Computer Graphics 16.6 (2010): 1366-1375.

13. Lê, Matthieu, et al. "Sampling image segmentations for uncertainty quantification." Medical image analysis 34 (2016): 42-51.

14. Jungo, Alain, et al. "On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018.

15. Jungo, Alain, and Mauricio Reyes. "Assessing reliability and challenges of uncertainty estimations for medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.

16. Taha, Abdel Aziz, and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool." BMC medical imaging 15.1 (2015): 29.

17. Norouzi, Alireza, et al. "Medical image segmentation methods, algorithms, and applications." IETE Technical Review 31.3 (2014): 199-213.

18. Litjens, Geert, et al. "A survey on deep learning in medical image analysis." Medical image analysis 42 (2017): 60-88.

19. Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In International Conference on Machine Learning, pages 1613–1622, 2015.

20. Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via upper bound minimization. In Advances in Neural Information Processing Systems, pages 2729–2738, 2017.

21. Yarin Gal. Uncertainty in Deep Learning. PhD thesis, University of Cambridge, 2016.

22. Alex Graves. Practical variational inference for neural networks. In Advances in Neural Information Processing Systems, pages 2348–2356, 2011.

23. Mohammad Havaei, Axel Davy, David Warde-Farley, et al. Brain tumor segmentation with deep neural networks. Medical image analysis, 35:18–31, 2017.

24. Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems, pages 5580–5590, 2017.

25. Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. arXiv preprint arXiv:1703.02914, 2017.

26. Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In International Conference on Machine Learning, pages 1708–1716, 2016.

27. David JC MacKay. A practical bayesian framework for backpropagation networks. Neural computation, 4(3):448–472, 1992.

28. Oskar Maier, Bjoern H Menze, von der Gablentz, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Medical image analysis, 35:250–269, 2017.

29. Radford M Neal. Bayesian learning via stochastic dynamics. In Advances in neural information processing systems, pages 475–482, 1993.

30. Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In Artificial Intelligence and Statistics, pages 814–822, 2014.

31. K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. arXiv preprint arXiv:1603.05027v3,2016.

32. Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

34. Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

35. Tran, D., Dusenberry, M.W., van der Wilk, M., Hafner, D.: Bayesian layers: A module for neural network uncertainty (2019), arXiv preprint 1812.03973

36. Shridhar, K., Laumann, F., Liwicki, M.: Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference (2019), arXiv preprint 1806.05978

37. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

38. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision (2016).

39. Kingma, Durk P., Tim Salimans, and Max Welling. "Variational dropout and the local reparameterization trick." Advances in neural information processing systems. 2015.

40. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

41. Gal, Yarin, and Zoubin Ghahramani. "Bayesian convolutional neural networks with Bernoulli approximate variational inference." arXiv preprint arXiv:1506.02158 (2015).

42. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

43. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

44. Myronenko, Andriy. "3D MRI brain tumor segmentation using autoencoder regularization." International MICCAI Brainlesion Workshop. Springer, Cham, 2018.