Eurographics Workshop on Visual Computing for Biology and Medicine (2019)
K. Lawonn and R. G. Raidou (Editors)

Short Paper

Semantic Segmentation of Brain Tumors in MRI data without any labels

Leon Weninger, Imke Krauhausen, Dorit Merhof

Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany

Abstract

Brain MR images are one of the most important instruments for diagnosing neurological disorders such as tumors, infections or trauma. In particular, grade I-IV brain tumors are a well-studied subject for supervised deep learning approaches. However, for a clinical use of these approaches, a very large annotated database that covers all of the occurring variance is necessary. As MR scanners are not quantitative, it is unclear how good supervised approaches, trained on a specific database, will actually perform on a new set of images that may stem from a yet other scanner. We propose a new method for brain tumor segmentation, that can not only identify abnormal regions, but can also delineate brain tumors into three characteristic radiological areas: The edema, the enhancing core, and the non-enhancing and necrotic tissue. Our concept is based on FLAIR and T1CE MRI sequences, where abnormalities are detected with a variational autoencoder trained on healthy examples. The detected areas are finally postprocessed via Gaussian Mixture Models and finally classified according to the three defined labels. We show results on the BraTS2018 dataset and compare these to previously published unsupervised segmentation results as well as to the results of the BraTS challenge 2018. Our developed unsupervised anomaly detection approach is on par with previously published methods. Meanwhile, the semantic segmentation - a new and unique model - shows encouraging results.

1. Introduction

Brain MR images are one of the most important instruments for detecting and diagnosing brain lesions. Currently, the analysis and examination of the scans is carried out by trained physicians. However, generating automated detections of anomalies could help to improve radiological studies and expedite the overall segmentation process. Automatic segmentation of these lesions has attracted the medical image analysis community and exceptional results have been accomplished with recent deep learning approaches, inter alia thanks to public challenges such as the Multi-modal Brain Tumor Segmentation (BraTS) or the Ischemic Stroke Lesion Segmentation (ISLES). However, most recent approaches with deep learning are supervised methods which require a vast amount of annotated data. Creating these annotations is a time-consuming task which can only be done by trained physicians. Simultaneously, large sets of unannotated MR data are available at hospitals, a yet untapped data source for deep learning.

By introducing unsupervised learning methods, the limitation of missing annotations or the lack of available medical experts can be overcome, and the existing data can be directly used. Recent approaches for unsupervised anomaly detection in brain MRI established a basis for detecting lesion contours. These unsupervised deep learning methods mostly rely on the use of variational autoencoders (VAEs) and generative adversarial networks (GANs). As healthy adult brains show similar structures, these architectures are trained on healthy brains only in order to learn their anatomy. Deviations from the norm, i.e., lesions, are then detected as they do not follow the learned structures.

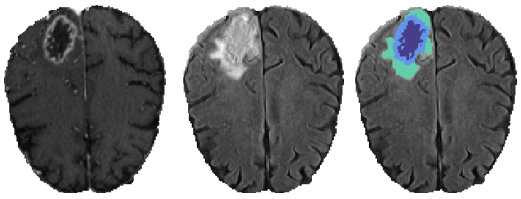

Figure 1: *MR imaging sequences. (a) T1-weighted image with contrast medium, (b) FLAIR image, (c) groundtruth labels.*

However, it is so far not possible to find and simultaneously identify the nature of outliers without training samples. By not only detecting lesions, but by further identifying the specific tissue types as shown in Figure 1, further clinical decision making is possible. The BraTS data set, which contains MRI scans of brain tumor patients, is suitable for this task, as different areas of the brain tumors need to be distinguished.

State-of-the-art anomaly detection architectures like VAEs and GANs build the basis of this work and are integrated in our model

DOI: 10.2312/vcbm.20191230

for the task of unsupervised multi-modal brain tumor semantic segmentation. The main contributions of this paper are:

- A comparison of different unsupervised deep learning techniques for anomaly detection in brain MRI scans.
- A new method built on anomaly detection techniques that incorporates medical knowledge in order to semantically segment different brain tumor areas without any training labels.

2. Related Work

In the last years, several papers on brain tumor segmentation in a supervised deep learning setting have been published. The BraTS challenge [MJBei15] provides a good overview and performance comparison of these approaches [BReIM18].

Concerning unsupervised lesion detection in MRI scans, several deep learning methods were also introduced in the last years. Baur et al. [BWAN19] examined different architectures of autoencoders and GANs. A combination of a spatial VAE and a GAN showed the best results on their in-house datasets of FLAIR and T1-weighted images of MS lesions. Finally, a dice score of up to 0.6 was reached. Chen and Konukoglu [CK18] implemented a ResNet based constrained adversarial autoencoders, on which a constriction to the latent space was added. Training on the Human Connectome Project, and testing on T2-weighted BRATS data, they reached AUC values for tumor segmentation between 0.897 and 0.923. Dice scores were not presented. Atlason et al. [ALS*19] used a convolutional autoencoder for segmentation of age-related lesions caused by ischemia or demyelination on T1, T2, and FLAIR MRI images, reaching dices score of up to 0.76. Zimmerer et al. [ZKP*18] added a new aspect to anomaly detection in brain MRI by combining a context encoding denoising autoencoder with a VAE. The networks were trained on the T2-weighted images of the HCP dataset and then tested on the BraTS and ISLES datasets. On the BraTS dataset, a dice score of 0.51 and an AUC value of 0.95 was reached for whole tumor segmentation.

3. Data

The BraTS2018 training dataset is used for training, validation and testing of the network. It consists of 285 patients from multiple institutions and is further divided into into 75 low-grade (LGG) and 210 high-grade (HGG) glioma. For each patient the T1, T1CE, T2 and FLAIR sequences as well as a segmentation map are available. The data shows a high variability between patients, as different scanners with varying degrees of quality are used. The original data was resampled to an isotropic resolution of 1 mm^2, with $240 \times 240 \times 150$ voxels. We crop this to $180 \times 180 \times 70$ voxels to get rid of excessive background and the brain stem. Finally, the data is scaled down to $128 \times 128 \times 70$ voxels, z-score normalized and subdivided into training and test set. The test set consists of 30 (20 HGG and 10 LGG) randomly chosen subjects, that remained completely unseen during network training, optimization and validation. During training, for each sample one augmented sample was created by randomly flipping, scaling, shearing or slightly rotating the sample. On top, for a stronger disturbance of the images, square patches of random size between 15 and 50 pixels were randomly added to the input images. In these patches, the original data was suppressed by Gaussian noise to make the autoencoder robust against deviations of the brain structure. This attempts to simulate disturbances that are later caused by the tumor.

4. Methods

First, VAEs, GANs, and combinations thereof are trained on 2D healthy MRI brain images to learn the composition and structure of a healthy brain. To ensure that only healthy structures are learnt, only slices without tumor annotations in the ground truth are used for training. Feeding unknown, anomalous data to the network then causes a desired failure: The network recreates healthy brain structures in de facto cancerous areas. Calculating the residual between input and output finally reveals anomalies.

The whole tumor contour seems to be best represented in the FLAIR image, and the enhancing tumor area is specifically highlighted in the T1CE image. Thus, as these two imaging sequences provide the best contrast for these areas, the anomaly detection models are trained on these acquisitions. After successful training, the network is tuned to recreate T1CE and FLAIR images of a healthy brain that are similar to its input. However, the network does not simply copy the input, but learns to recognize important features of healthy brain structures.

In most cases, the edema is the most visible part of the tumor and lies exterior to the tumor core. Since this is best seen in FLAIR sequences, tumor contour segmentation is based on FLAIR residuals. Based on the residuals, a neighborhood-weighted Gaussian mixture models is used for segmentation. This returns a binary mask, of which the smallest of the two clusters is chosen and smoothened by morphological operations. Using the T1CE residual in the same way as the FLAIR residual, the enhancing region can be detected.

The instructions for manual segmentations to be followed by medical staff given by the BraTS organizers [BAS*17] gives an orientation on how the different areas can finally be detected. Specifically, the enhancing tumor core mostly encloses the necrotic tissue. Thus, using the convex hull around the enhancing core, an approximation of the non-enhancing and necrotic core of the glioma can be found.

4.1. Network architectures

Autoencoder are a special type of neural networks, and a promising approach for unsupervised learning [GBC16]. They consist of two parts building a simple encoder-decoder structure. The encoder translates the input information to a hidden layer often called latent space, which is significantly smaller than the input data. The decoder's task is to reconstruct the output from latent space z to resemble the input. The main idea behind this model is not to simply copy input to output, but to perform training in a way that the latent space holds indicative and conclusive information about the input.

VAEs and GANs represent further development of deep learning architectures for unsupervised learning tasks. VAEs follow the architecture of autoencoders, but the latent space z is sampled instead of being forwarded in a deterministic manner [KW13]. However, VAEs tend to create slightly blurred images which cause undesired differences in the residual.

In contrast, GANs are known to create highly defined outputs. A combination of both network types can lead to sharper outputs and improvements in performance [LSW15], as was also shown for brain lesions by Baur et al. [BWAN19]. GANs consist of two parts: a generator and a discriminator. The guiding idea is a constant competition between the generator and the discriminator (see Figure 2). While the generator creates fake images similar to the real input, the discriminator serves as a controlling instance and tries to identify counterfeit inputs. A forger and an inspector work against each other in steady competition until counterfeit and real data are indistinguishable [GPAM*14]. In case of adversarial autoencoders, the discriminator takes the original image and the reconstruction of the VAE as input and learns to recognize the reconstruction. Finally, as feedback on how good the reconstruction is, an adversarial loss is handed to the VAE.

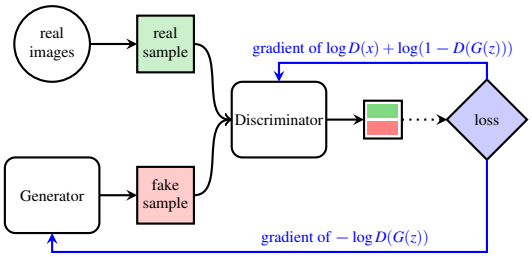

Figure 2: *Layout of generative adversarial networks.*

For image data, either dense networks or convolutional architectures can be utilized. Since Baur et al. [BWAN19] have shown that convolutional architectures are much more suited for the task of anomaly detection in MRI brain scans than dense networks, all presented networks use fully convolutional architectures.

The different autoencoders implemented and tested are denoted as follows: *cVAE* describes a common fully convolutional autoencoder adapted to the specific data shape, *dVAE* an autoencoder with dilated convolutional filters, and *VAE1px* an autoencoder with a very small latent space (1 pixel with 64 channels). Details of these architectures are revealed in Table 1. Our adversarial variational autoencoder *AAE*, uses the cVAE as autoencoder, and the architecture described in Table 1 as discriminator.

Finally, we also evaluate a bidirectional adversarial autoencoder *BiAAE*, which takes the latent space into account. While the adversarial autoencoder only compares the input scan and the reconstruction, the BiAAE also considers differences in the latent space and therefore forces the autoencoder to assimilate even further to the input images. The discriminator of the BiAAE has two inputs: one for the input scan and the encoder output, the second for the reconstructed image and the input of the decoder. The discriminator architecture differs only slightly compared to the AAE. The 6 layers of the AAE discriminator are assumed for the first input, the image/reconstruction. The structure only changes in the last layers, where the output increases from 64 to 128 channels. The latent space input is handled by a second smaller structure with only two convolutional blocks with 128 feature maps and 4x4 kernels. Those two networks fuse into a third network that consists of three 1x1 convolutional blocks.

Table 1: *Encoder and latent space architecture. In the encoder, conv designates a convolutional block consisting of one convolutional layer, batch-norm normalization and LeakyReLU activation function. The conv-layer is denoted as "conv⟨receptive field size⟩, ⟨stride⟩, ⟨dilation⟩-⟨number of channels⟩. The decoder is directly transposed to this architecture, using fractionally strided convolutional layers. The two layers of the latent space describe $\mu(X)$ and $\Sigma^{0.5}(X)$. No normalization or activation functions are used here.*

VAE	dVAE	VAE1px	AAE/BiAAE-Discriminator
	Encoder		
conv4, 2,1-16	conv6, 1,1-8	conv4, 2,1-16	conv4, 2,1-16
conv4, 2,1-32	conv5, 2,2-16	conv4, 2,1-32	conv4, 2,1-32
conv4, 2,1-64	conv6, 1,2-16	conv4, 2,1-64	conv4, 2,1-64
conv4, 2,1-64	conv5, 3,2-16	conv4, 2,1-64	conv4, 2,1-64
			conv4, 2,1-64
			conv4, 2,1-64
	Latent space		
conv3, 2,1-64	conv3, 1,1-16	conv8, 1,1-64	
conv3, 2,1-64	conv3, 1,1-16	conv8, 1,1-64	

We also include the architecture proposed by Zimmerer et al. [ZKP*18] by adding a context-encoding autoencoder to the VAE. This architecture is designated as *VAE-CE* in the following.

4.2. Anomaly score

During training, the network constantly provides information on the features of the input with the mean and standard deviation inside the latent space. As the network is trained only on healthy brain structures, the means and standard deviations should correlate to healthy brains - cancerous inputs should show up as outliers. Thus, the mean and variance of the latent space are saved during training. During prediction, the latent representation of the new data can be compared to the stored latent space of healthy examples with the Mahalanobis distance. If the deviation is higher than two times the standard deviation of the Mahalanobis distance, the sample is classified as pathological.

As a second possibility, the mean and variance of the residuals can serve straightforwardly as an anomaly score, as they should be close to zero for brain slices similar to those seen during training. In contrast, the mean of the residual should be significantly lower for scans with brain tumors, as these show up as negative values in the residual. Thus, similarly to the Mahalanobis distance based latent space classification, a maximum threshold can be determined based on exemplary healthy slices.

4.3. Postprocessing

The number of components in the residual brain images is expected to be two: healthy brain mass and tumor. The clusters are calculated with an expectation maximization approach. Since the tumor is never bigger than the actual brain, the smallest of the two cluster is selected as tumor. As it is very plausible to make assumptions about the surrounding tissue if a first hypothesis is already given, a Gaussian mixture model with a neighborhood-based weight is introduced, as in [TDL08]. The neighborhood weighted probability can be calculated as follows:

$$p(k|x_i,\theta^t) = \frac{\alpha^t W_{ik} p_i(x_i|\theta^t)}{\sum_{j=1}^{K} \alpha^t W_{ik} p_i(x_i|\theta^t)} \quad (1)$$

with

$$W_{ik} = \frac{\sum_{n=1}^{N} p(k|x_{ni},\theta^t)}{|N_i|}, \quad (2)$$

where the neighborhood of directly surrounding 8 pixel is described by N, and x_{ni} indicates the nth neighbor's pixel value. This weighted probability can be used inside the expectation maximization of a Gaussian mixture model instead of the unweighted probability.

Subsequently, morphological operations help to improve results by removing small pixel groups from the segmentation masks. One binary opening operation removes small groups of false positives, and one binary closing removes small holes in the segmentation mask.

An overview over the postprocessing, with which we obtain the different areas, can be seen in Figure 3.

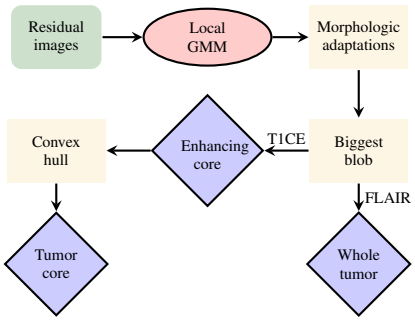

Figure 3: *Segmentation pipeline for tumor regions.*

5. Results

The final reconstructions of the network as well as the obtained segmentation masks are obtained during the prediction process. The quality of the segmentations masks is contrasted in the following. The different architectures and loss functions are compared against each other. A visual impression of the results is given in Figure 4.

5.1. Segmentation without anomaly score

To compare the segmentation results of the different architectures, we first evaluate the results before applying the anomaly score. This means that a slice without pathogenic structures will have a dice score of zero. Thus, the overall dice score is significantly worse than for the final results. However, this approach allows a fair comparison of different architectures.

The results for the tumor outline segmentation are provided next to the semantic segmentation results with the Gaussian Mixture Model, so that this metric can be compared to the previously published approaches. All segmentations are created using the network with the lowest validation loss, that was chosen after training reached convergence. The complete results are presented in Table 2 for MSE and SSIM loss functions. The best value per loss

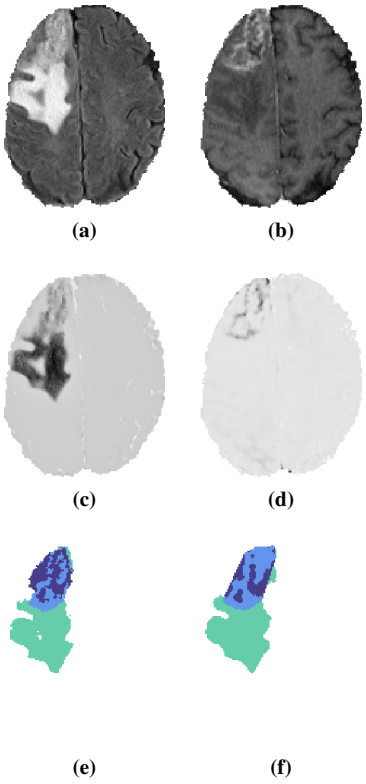

Figure 4: *Segmentation masks for exemplary segmentation results. (a) Flair input scan. (b) T1CE input scan. (c) Flair residual. (d) T1CE residual. (e) Ground truth. (f) Segmentation mask.*

for both outline segmentation and semantic segmentation is emphasized in the table. The SSIM loss achieves overall the best results. L1 loss functions were also tested but resulted in slightly worse results compared to the MSE loss functions.

Table 2: *Dice scores for all models using MSE loss and SSIM loss function*

Model:	Tumor outline		Semantic seg.	
Loss:	MSE	SSIM	MSE	SSIM
VAE	0.377	0.409	0.232	0.253
dVAE	0.379	**0.410**	0.234	**0.260**
VAE1px	0.371	0.350	0.221	0.201
VAE CE	0.383	0.408	0.236	**0.260**
AAE	0.377	0.370	0.231	0.216
BiAAE	0.383	0.407	0.235	0.248

The best settings reach dice score of 0.26 for the semantic segmentation before application of the anomaly score. This is achieved with VAEs using the SSIM loss function but without added GAN-loss.

Compared to the existing methods explained in Section 4.1, the dilated VAE shows a similar, or even slightly better performance as the VAE in combination with a denoising autoencoder proposed

by [ZKP*18]. All further results are demonstrated with this dilated convolutional autoencoder architecture in combination with the SSIM loss function.

5.2. Anomaly scores

By adding an anomaly score, it is compensated that the segmentation methods cannot handle cases in which no tumor is present. As explained in Section 4, two different anomaly scores are evaluated: Information based on the latent space, and information based on the mean residual. In Table 3, it can be seen that while the mean residual is good indicator of pathological slides with a ROC-AUC of 0.835, the latent space information provides a significantly lower ROC-AUC. Thus, the mean residual method is chosen for further experiments.

Table 3: *ROC-AUC: How well different anomaly scores predict pathological slices*

Anomaly score	AUC
Mean residual	0.835
Latent space	0.471

5.3. Overall results

In combination, the best network (dVAE) and the anomaly score reach the following accuracies on the held-out test dataset: Whole tumor: 0.64, tumor core: 0.61, and enhancing tumor: 0.62. In comparison, with a fully supervised setting in the BraTS challenge 2018, the top teams reached average dice scores between 0.75 and 0.9. [BReIM18].

6. Discussion

It was demonstrated that unsupervised learning methods are capable of segmenting different radiological areas in brain tumors when incorporating medical knowledge, and the presented segmentation algorithm proved to be a suited choice for segmenting the three areas of the tumor. Critically examined must be the decision to choose the biggest blob during segmentation as well the assumption that three classes should always be distinguished. Not every tumor actually exhibits all mentioned areas, and there may be more than a single brain tumor. Especially in brain tumors without contrast enhancing tissue, our model reaches its limits. Further development is needed to adjust this model to more complex cases, and to reach the accuracy that state-of-the art supervised models offer.

Nevertheless, with FLAIR and T1CE images as input, a distinction between edema, tumor core and enhancing tumor was possible without any training labels. It was shown that it is possible to achieve good segmentation results by using unsupervised deep learning methods in combination with prior medical knowledge. Such models offer big advantages over fully supervised deep learning methods: First of all, they do not rely on vast annotated databases. Second, as they fail to reconstruct pathological data, and with medical knowledge incorporated in the postprocessing part, they are better explainable than straightforward supervised deep learning approaches.

7. Conclusion

This paper proposes an unsupervised semantic segmentation method for gliomas in brain MRI. Without using any labels, we devised a model that can identifying the three different tumor tissue classes as used in the BraTS dataset. The model includes state-of-the-art variational autoencoders and GANs to create a reconstruction of healthy tissue that fails on pathological tissue. While our model produces good results, it currently implements a simplified version of the complex nature of brain tumors. For future advances, the model incorporating medical knowledge into the segmentation process needs to be refined.

References

[ALS*19] ATLASON H. E., LOVE A., SIGURDSSON S., GUDNASON V., ELLINGSEN L. M.: Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder. *SPIE Medical Imaging* (2019). 2

[BAS*17] BAKAS S., AKBARI H., SOTIRAS A., BILELLO M., ROZYCKI M., KIRBY J. S., FREYMANN J. B., FARAHANI K., DAVATZIKOS C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data 4* (09 2017), 170117. 2

[BReIM18] BAKAS S., REYES M., ET INT, MENZE. B.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv e-prints* (2018). arXiv:1811.02629. 2, 5

[BWAN19] BAUR C., WIESTLER B., ALBARQOUNI S., NAVAB N.: Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (2019), Springer International Publishing, pp. 161–169. 2, 3

[CK18] CHEN X., KONUKOGLU E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. *arXiv e-prints* (2018). arXiv:1806.04972. 2

[GBC16] GOODFELLOW I., BENGIO Y., COURVILLE A.: *Deep Learning*. MIT Press, 2016. 2

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, Ghahramani Z., Welling M., Cortes C., Lawrence N. D., Weinberger K. Q., (Eds.). Curran Associates, Inc., 2014, pp. 2672–2680. 3

[KW13] KINGMA D. P., WELLING M.: Auto-Encoding Variational Bayes. *arXiv e-prints* (2013). arXiv:1312.6114. 2

[LSW15] LARSEN A. B. L., SØNDERBY S. K., WINTHER O.: Autoencoding beyond pixels using a learned similarity metric. *CoRR* (2015). arXiv:1512.09300. 3

[MJBei15] MENZE B. H., JAKAB A., BAUER S., ET INT.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging 34*, 10 (Oct 2015), 1993–2024. 2

[TDL08] TANG H., DILLENSEGER J.-L., LUO L. M.: A vectorial image classification method based on neighborhood weighted gaussian mixture model. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2008), IEEE. 3

[ZKP*18] ZIMMERER D., KOHL S. A. A., PETERSEN J., ISENSEE F., MAIER-HEIN K. H.: Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. *arXiv e-prints* (2018). arXiv:1812.05941. 2, 3, 5