Tumor Segmentation and Survival Prediction with Generative Networks

Ty Trusty
Dept. of Computer Science
University of Texas at Austin
tytrusty@utexas.edu

Sanat Sharma
Dept. of Computer Science
University of Texas at Austin
sanatsharma@utexas.edu

Daniel Noble
Dept. of Computer Science
University of Texas at Austin
dnoble@cs.utexas.edu

ABSTRACT

Tumor recognition and segmentation is a critical tool in helping doctors analyze and treat patients. In this work, we look at solving the tumor segmentation problem by creating an efficient segmentation network and then predicting survival rate for operated patients. Multimodal Brain Tumor Segmentation Challenge (BraTS 2019)[9][1] aims to provide a harness to evaluate brain tumor segmentation methods and provides a 3D MRI dataset with ground truth tumor segmentation provided, which we utilize for our tasks. Previous top models have routinely utilized Deep networks to achieve state-of-the-art segmentation results[10] and we continue in that tradition. Our work makes use of a convolutional segmentation network UNet[13], which has been shown to do well on medical segmentation tasks, and we primarily focus on the second task in the BraTS challenge on patient survival prediction. Various methods, including linear regression and artificial neural networks, have been applied to this task, but none have achieved reasonable performance due to limited data availability. We explore ways to combat this lack of data by experimenting with the more recent machine learning development of Generative Adversarial Networks (GAN) [4] as well as traditional methods of Support Vector Machines [2] and Cross-validation. Experimental results demonstrate promise of our GAN approach with comparable results to state of the art, and the value of traditional methods by achieving higher validations scores than our baseline.

The source code, preprocessed datasets and additional model information are available at:

github.com/tytrusty/mri-survival-prediction

KEYWORDS

tumor, segmentation, GAN, unet, tumor segmentation

1 INTRODUCTION

Artificial Neural Networks have shown incredible performance on traditional computer vision tasks; however, they depend on large quantities of data to achieve this performance. In medical image processing, this is a huge constraint given the limited availability of data. For this reason, recent work has aimed to use generative modeling techniques, such as Generative Adversarial Networks (GAN) [4] to synthesize data to augment datasets. For example, Shin et al. [16] trained a GAN to synthesize MRI data and found improved generalization on the Brats segmentation challenge when the training data was augmented with synthetic data. More recently, Kwon et al. [8] proposed a GAN model to combat unstable training and mode collapse problems encountered in GAN training. Most of the medical GAN papers focus on synthesis and image-to-image translation tasks. For the Brats challenge, we have yet to see a paper focused on improving the survival prediction performance by making use of a generative model. In our work we aim to do this by treating the discriminator in our GAN as a feature extractor for training survival predictions. Previous work makes use of hand-crafted features extracted from segmentation masks, but we instead aim to use features learned through adversarial training in an attempt to better model the features associated with brain tumors.

In addition to exploring generative modeling, we also aimed to experiment with methods closer to previous work on the survival prediction task. Like these other methods, we use hand-crafted features from predicted segmentation masks, and train a simple model on this small feature set. We further sook to implement cross-validation techniques to compare accuracy of the data when using different data splits. Additionally, we utilized a support vector machine to provide a comparison with the neural network as a classifier, on the premonition that it could be better suited to the task.

2 METHODOLOGIES

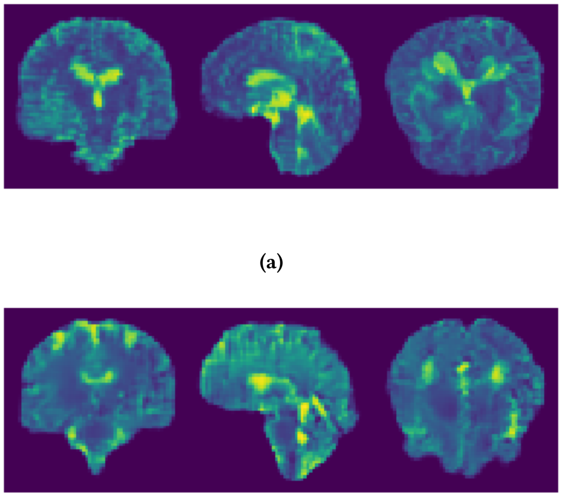
2.1 Dataset

The Brain Tumor Segmentation (BraTS) dataset utilizes MRI scans of the brain, focusing on the segmentation - that is, the separation of tumor from healthy brain tissue - of heterogeneous brain tumors. Features of the training and validation data include age of the patients, the survival - given by the number of days survived after resection by each of the patients - and the resection status; note that for our purposes, only patients who received a gross total resection, indicated by a status of GTR, were evaluated in our model.

2.2 GAN Survival Prediction

The majority of survival prediction approaches focus on using hand crafted features based on the geometry of tumors from predicted segmentation masks plus patient information such as age. Our motivation for using a GAN is to attempt an alternative approach by using features extracted from a GAN that learns the distribution of the MRI data. The discriminator of our network is treated as a feature extractor and is directly used to predict the survival prediction labels.

Our model is based on the GAN from [8]. This model focuses primarily on solving problems related to training stability, mode collapse, and blurriness. To do this they introduce a model that combines Wasserstein GAN with gradient penalty [5] as well as α-GAN [15]. Despite limited data availability, this model achieves very high high performance (see Figure 1 for an example synthesized MRI). We modify the discriminator architecture so that in



(a)

(b)

Figure 1: (a) A real MRI T2 image (b) An example scan synthesized by the GAN.

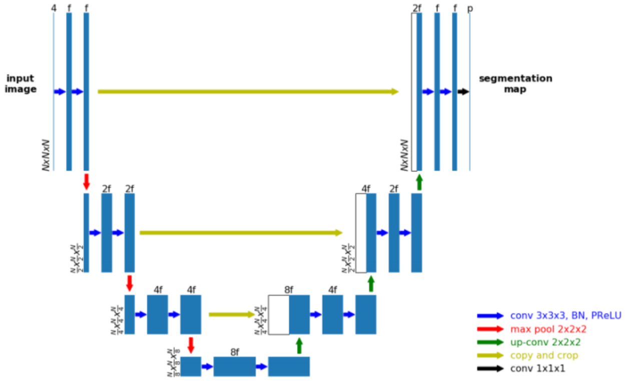

4
f
f
input image
NxNxN
2f
f
f
p
segmentation map
NxNxN
2f
2f
$\frac{N}{2}x\frac{N}{2}x\frac{N}{2}$
4f
2f
$\frac{N}{2}x\frac{N}{2}x\frac{N}{2}$
4f
4f
$\frac{N}{4}x\frac{N}{4}x\frac{N}{4}$
8f
4f
$\frac{N}{4}x\frac{N}{4}x\frac{N}{4}$
8f
$\frac{N}{8}x\frac{N}{8}x\frac{N}{8}$
conv 3x3x3, BN, PReLU
max pool 2x2x2
up-conv 2x2x2
copy and crop
conv 1x1x1

Figure 2: U-net Architecture

addition to classifying real or fake images, it also predicts the survival prediction classes. There are many methods of achieving this and our approach most closely follows that of AC-GAN [11]. The discriminator consists of a series of 3D convolutions with 4x4x4 filters, each of which is followed by Batch Normalization and a Leaky ReLU activation. The final layer outputs 4 values: 1 that predicts whether the input is real or fake, and 3 for predicting short, mid, and long survival class labels.

2.3 Tumor Segmentation

In our survival prediction experiments without GANs, we built upon previous work by using hand crafted features from segmentation masks. In order to produce these features, we first must be able to segment that MRI data. Much of the work on the BraTS challenge has been focused on this task and state of the art methods opt for a U-Net [14] architecture for segmentation. This is a simple 'U' shaped network (see Figure 2) with a series of convolutions followed by maxpool layers in the encoder, and mirrored in the decoder but with trilinear upsampling. Isensee et al. [6] present a general framework for training on BraTS data with a 3D U-Net and patch-based training. We make use of this architecture and use pretrained weights so that we could focus primarily on the survival prediction task.

2.4 Neural Survival Prediction

In addition to a GAN-based survival prediction approach, we also looked at neural approaches that do not solely rely on physical features but also utilize additional features such as the patient's age.

In order to obtain spatial and geometric features, we extract multiple features, including the ratios of the volume of each tumor sub-region to the size of the whole brain. Further, we get the gradient of each tumor and sum up gradient values to approximate the total area of the tumor surface. In total, we utilize 6 features for our feature vector.

Based on the BraTS challenge, we aim to classify survival based on 3 classes: less than 300 days as short-survival, from 300 to 450 as mid-survivor, and more than 450 as long-survival.

2.4.1 Neural Network.

We utilize a simple sequential/fully-connected neural network for multi-class classification with 2 hidden layers and Adam Optimizer[7] and categorical crossentropy loss. Each hidden uses 64 hidden units, with a softmax activation on the output layer. We trained the network with 5 fold cross-validation, as well as early stopping.

2.4.2 SVM.

Due to the lack of data, we estimate that a Support Vector Machine might be better suited for the current task than a neural net. We utilize a simple SVM with a radial kernel.

2.5 Cross-validation

As part of the project, we wanted to test how splitting the training and validation datasets using cross-validation would affect the results. As such, our model implemented Leave-one-out cross-validation (LOOCV), k-fold cross-validation, and Monte Carlo cross-validation to compare the results of each of the methods.

LOOCV, a technique that works well for independent datasets, was implemented by iterating over all of the datapoints, and for each iteration, choosing one datapoint to be the validation set, whilst the remaining datapoints were grouped into the training set. K-fold cross-validation generalized the LOOCV method, splitting the data set into k folds and setting one of these folds of n / k datapoints as the validation set whilst using the remaining folds to train; again, this was done, iterating over all folds. Finally, we implemented Monte Carlo cross-validation - assigning some random fraction of the training set as the validation set on each iteration - with the idea that this would allow for a large number of folds whilst maintaining the stability of the estimate.

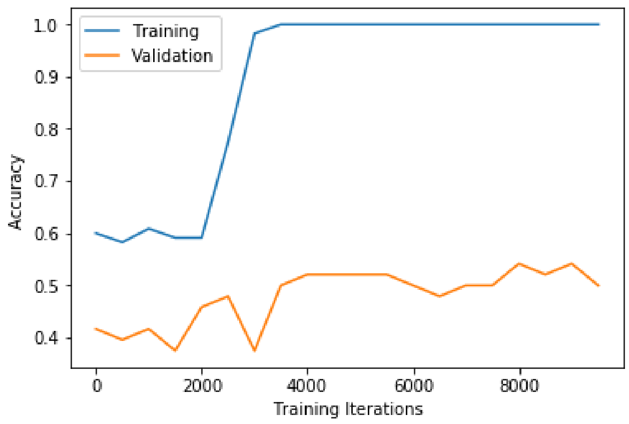


Figure 3: Training Plot for Classification Fine-Tuning

3 EXPERIMENTS

BraTS has had a history of poor results for survival days prediction, partly due to the small amount of data. State of the art methods in 2018 and 2019 typically report 50-55 percent accuracy on the validation data. For example, the third place winner [17] of the survival prediction in Brats 2019 reported a 0.448 validation accuracy and a 0.551 test accuracy. Also, the winner [3] of the Brats 2018 challenge reported a test accuracy of just 0.321, indicating the heavy amount of overfitting among submissions.

3.1 GAN Survival Prediction

For this task we exclusively work with T2 scans from the BraTS 2018 training set. We lack the ground truth validation data or the test data, so we did a 70:30 training, validation split on the data. We had 163 total data that which included survival days data. After the split this left us with only 115 training examples. The experiments and training of this model were conducted on an NVIDIA Gtx 1080 ti 12GB GPU. We initially trained the entire GAN model for 4000 iterations, and then trained an additional 10000 iterations fine tuning the discriminator classifier by only considering the classification loss. High memory requirements of the 3D network force us to use a batch size of just 4.

Figure 4 shows the training and validation score during the fine-tuning phase. We see that after about 3000 iterations the training has converged and the validation score plateaus at about 0.5. The highest validation accuracy achieved was 0.542. These results were quite surprising and demonstrates the promise of our approach. Since we lack the validation or test ground truths it would not be fair to compare directly. However, we believe these positive results indicate that the auxiliary class labels on our discriminator forces the model to the important features for classifying survival based on the information in the T2.

3.2 Cross-validation

The results of using k-fold cross-validation can be seen in the MSE yielded for the validation and training sets in Figure 4. LOOCV and Monte Carlo cross-validation yielded higher MSEs than k-fold cross-validation did; this can perhaps be attributed to the related nature of the dataset.

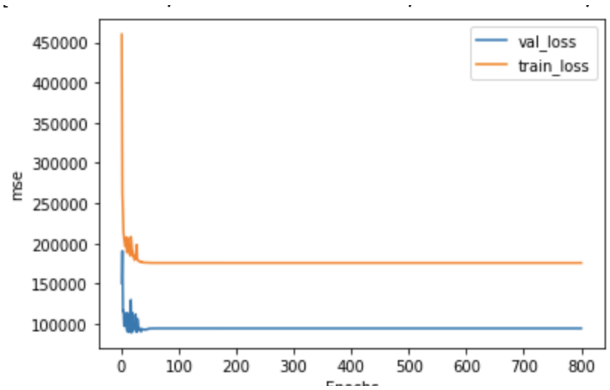


Figure 4: MSE using cross-validation

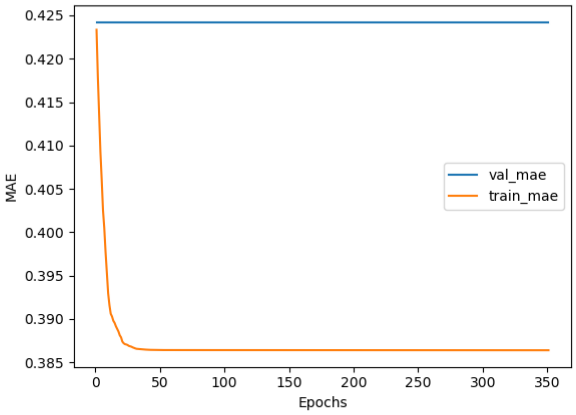


Figure 5: Mean Absolute Error Neural Net

3.3 Neural Survival Prediction

For survival days prediction, since we did not have a validation set, we utilized training data from 2018 BraTS challenge as the validation set (and part of it for our training set as well). Our total training set comprised of 190 training samples and 78 test samples.

For the neural net, we ran 2 models, one with 2 hidden while the other with a single hidden layer. We saw the single hidden layer network to perform better, however, the paucity of data led to overfitting. The highest validation accuracy achieved by the neural net was .395. This was expected, even with the rich features we were using. We expect accuracy to radically increase with increase in data. Figure 5 shows the Mean Absolute Error curve for the neural net.

The SVM performed radically better than the neural net. The SVM achieved an accuracy of .576 and an F1 score of .544, performing better than current state of the art methods, based on the training losses.

Table 1: Experiment Results.

Model	Train Accuracy	Val Accuracy
Wang et al. [17] 2019	0.515	0.448
GAN (ours)	–	0.542
Neural Net (ours)	0.601	0.395
SVM (ours)	–	**0.576**

4 FUTURE WORK AND CONCLUSION

In this work we studied survival prediction and generative techniques for tumor segmentation. Some approaches worked and some that did not fare as well. The biggest issue in training a good prediction and segmentation model is the lack of MRI training data. We wish to supplement this work by diving deeper into probabilistic models for creation of high quality tumor images. Recent work in latent Autoencoders[12] have shown significant improvement in generative tasks.

On the survival prediction front, we are forced to rely to survival data which will improve and increase in the future. Given time constraints and limited compute resources, we were unable to fully evaluate the effectiveness of the GAN approach. The original Kwon et al. paper recommends training the GAN for 200,000 iterations, which is far more than we were able to. An immediate future work would be to train the GAN model fully and evaluate performance then. Also, since GAN-based survival prediction is largely unexplored, it would be valuable to experiment with different GAN architectures and parameters. Another GAN approach may be to learn to synthesize segmentation masks, which may produce better features for survival prediction.

5 INDIVIDUAL CONTRIBUTIONS

Ty: Initial implementation of survival prediction feature extraction based on previous work. GAN implementation. Write-up of GAN + Segmentation parts, and other bits and pieces.
Sanat: Implementation of survival prediction (neural net, SVM); write-up and editing
Daniel: Implementation and write-up of LOOCV, k-fold cross-validation, and Monte Carlo cross-validation; write-up of dataset; general editing of report

REFERENCES

[1] Spyridon Bakas, Mauricio Reyes, András Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, Marcel Prastawa, Esther Alberts, Jana Lipková, John B. Freymann, Justin S. Kirby, Michel Bilello, Hassan M. Fathallah-Shaykh, Roland Wiest, Jan Kirschke, Benedikt Wiestler, Rivka R. Colen, Aikaterini Kotrotsou, Pamela LaMontagne, Daniel S. Marcus, Mikhail Milchenko, Arash Nazeri, Marc-André Weber, Abhishek Mahajan, Ujjwal Baid, Dongjin Kwon, Manu Agarwal, Mahbubul Alam, Alberto Albiol, Antonio Albiol, Alex Varghese, Tran Anh Tuan, Tal Arbel, Aaron Avery, Pranjal B., Subhashis Banerjee, Thomas Batchelder, Kayhan N. Batmanghelich, Enzo Battistella, Martin Bendszus, Eze Benson, José Bernal, George Biros, Mariano Cabezas, Siddhartha Chandra, Yi-Ju Chang, and et al. 2018. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *CoRR* abs/1811.02629 (2018). arXiv:1811.02629 http://arxiv.org/abs/1811.02629

[2] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[3] Xue Feng, Nicholas Tustison, and Craig Meyer. 2018. Brain Tumor Segmentation using an Ensemble of 3D U-Nets and Overall Survival Prediction using Radiomic Features. arXiv:cs.CV/1812.01049

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 2672–2680.

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. arXiv:cs.LG/1704.00028

[6] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. 2018. No New-Net. arXiv:cs.CV/1809.10483

[7] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[8] Gihyun Kwon, Chihye Han, and Dae shik Kim. 2019. Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks. arXiv:eess.IV/1908.02498

[9] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* 34, 10 (2015), 1993–2024.

[10] Andriy Myronenko. 2018. 3D MRI brain tumor segmentation using autoencoder regularization. *CoRR* abs/1810.11654 (2018). arXiv:1810.11654 http://arxiv.org/abs/1810.11654

[11] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2016. Conditional Image Synthesis With Auxiliary Classifier GANs. arXiv:stat.ML/1610.09585

[12] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial Latent Autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. [to appear].

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* abs/1505.04597 (2015). arXiv:1505.04597 http://arxiv.org/abs/1505.04597

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:cs.CV/1505.04597

[15] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. 2017. Variational Approaches for Auto-Encoding Generative Adversarial Networks. arXiv:stat.ML/1706.04987

[16] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine P. Andriole, and Mark Michalski. 2018. Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks. *CoRR* abs/1807.10225 (2018). arXiv:1807.10225 http://arxiv.org/abs/1807.10225

[17] Feifan Wang, Runzhou Jiang, Liqin Zheng, Chun Meng, and Bharat Biswal. 2019. 3D U-Net Based Brain Tumor Segmentation and Survival Days Prediction. arXiv:eess.IV/1909.12901