Brain Tumor Classification Using Deep Learning

Will Pratt
University of Colorado Boulder

ABSTRACT

Convolutional neural networks (CNN) have become the driving force behind developments in computer vision in recent years (1). Deep learning techniques have demonstrated extraordinary performance in the past, specifically on classification and segmentation of images. The aim of this project is to compare 2 of the well known CNN architectures against one another for the purpose of classifying and segmenting brain tumors.

Brain tumors are one of the most fatal cancers out there. It can happen in both adults and children. If a tumor is classified wrong that could be the difference between life or death for someone. The classification of a tumor in many cases is what leads to a precise treatment plan for the patient. For the purpose of this project, magnetic resonance imaging (MRI) images will be used. Generally, a radiologist will manually read the scans and make a clinical decision. As mentioned before, this manual examination has the potential to be quite error prone. Becasue of this, I am proposing to use 2 state of the art CNN models, specifically VGGNet and ResNet, to see if we can acheive superior accuracy.

Both algorithms performed exceptionally well with the ResNet having the best results for both classification and segmentation.

KEYWORDS

Convolutional Neural Network, Brain Tumor Classification, ResNet, VGGNet, Deep Learning

1 Introduction

Brain tumors are abnormal cell aggregations that grow inside the brain tissues. They can be divided into either benign tumors or malignant tumors. Benign brain tumors can be cured by surgery, while malignant brain tumors are one of the most deadly types of cancer and can lead directly to death (2)(3). Different symptoms appear with different lesion areas, such as headache, vomiting, visual decline, epilepsy, and confusion (3). In all cases, early detection of the tumor is of the utmost importance. In general, medical imaging and imaging techniques are used to identify the shape, size and location of the tumor. The most common of these techniques is megnetic resonance imaging (MRI). MRI is a non-invasive imaging technology that produces three dimensional detailed anatomical images. It is based on sophisticated technology that excites and detects the change in the direction of the rotational axis of protons found in the water that makes up living tissues. It has certain advantages to it such as, high soft tissue contrast as well.

Research in deep learning has progressed greatly over the past 10 years, with radiological research being no exception to that. Brain tumor classification includes two procedures: feature extraction and classification. In many of the previous studies involving medical image classification, researchers used manual extraction of features. The problem however, is that traditional feature extraction methods require expansive knowledge and experience in specific fields. Manual feature extraction will also reduce the efficiency of the system. Deep learning techniques overcome this disadvantage (4).

The goal of this project is to use 2 of the most highly regarded CNN architectures in the application of MRI classification and segmentation. The performance of CNNs are validated in terms of accuracy, recall, precision,

F1-score, loss, and Tversky score. The end goal is have the ability to vastly alter the outcomes of patients, if applied correctly, in a positive way.

Up to this point there has been limited studies comparing these 2 state-of-the-art models against each other in the field of radiology. Hopefully this project will illuminate which one has the best promise and will aid in future work of classification of medical images.

This project will walk through the implementation and results of applying both algorithms and discuss any challenges or future work that should be pursued.

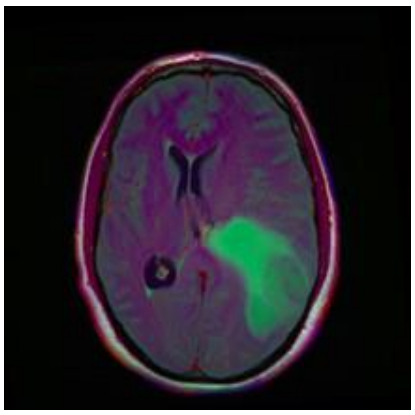

Figure 1: Example of an MRI image from dataset

2 Related Work

To this point the most common CNN architecture used for medical imaging classification has been the ResNet model. ResNet makes use of shortcut connections to solve the vanishing gradient problem. There has been other work published using various other CNNs to classify images as well.

This project builds upon prior work by using the same dataset of images and comparing it against a different CNN to clearly be able to identify a winner. Comparing the accuracy of various CNNs being used on different daatsets is like comparing apples and oranges. The goal here is to make it an apples to apples comparison.

The difference in simple terms between the more commonly used ResNet and the VGGNet is the numbers of parameters the VGGNet uses to train is larger, so it will be interesting to see if this leads to better results.

3 Proposed Work and Evaluation

Overview of Proposed Work and Evaluation Metrics

For this project we will be using tensorflow to implement the CNNs. The steps for implementation will be as follows. Images and labels for each image will be extracted from the dataset. After this we will standardize and preprocess all images. Next we will partition and create a training, test and validation set. From there, we will train and tune the models. The end goal here is to create a two part model. The first section will consist of classification of the images into either having a tumor or not having a tumor. Here, we will use 90% as our baseline for whether the MRI has a tumor. If it is identified as having one then it will be run through the segmentation model which will identify and highlight which part of the brain contains the tumor. Each CNN process will be completely separated from the other. For the ResNet, both the classification and segmentation will be trained on the same CNN structure albeit with different weights in the neural network, because they are being trained to execute different tasks. The same will be done for the VGGNet. For evaluation, we will use visualizations such as a confusion matrix and evaluation metrics that consists of precision, recall, f-1 score, support and overall accuracy. We will also be using the Tversky loss function as a metric. In terms of computing and storage, our data will be stored locally. This is due to tht dataset being relatively small (~4000 images). The computer I will be using has a Intel Core i7-10850H vPro processor with a Nvidia Quadro T2000 Max-Q GPU. When

comparing algorithms I will evaluate using all the evaluation metrics mentioned above along with total run time of each algorithm. An example of the brain MRI and the mask images that the model will be trained on are given below. Note the MRI with mask image was generated to help visualize.

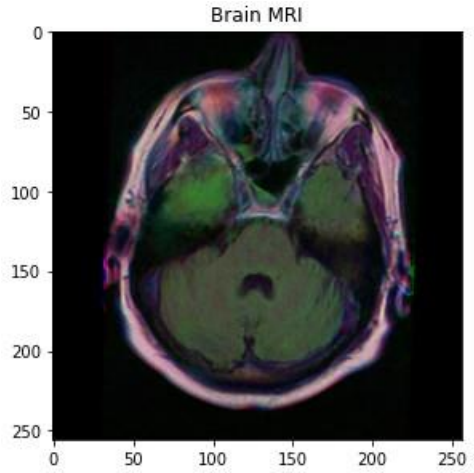

Figure 2: Example of an MRI image from dataset

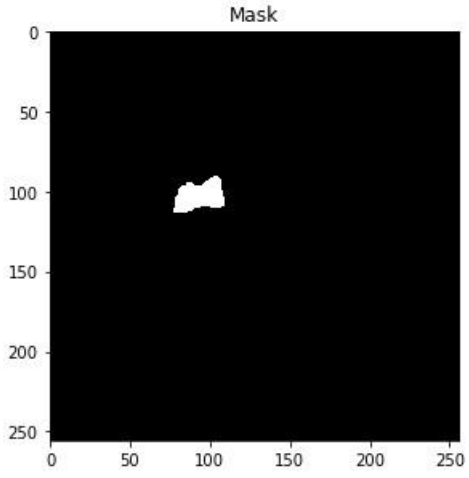

Figure 3: Example of an Mask image from dataset

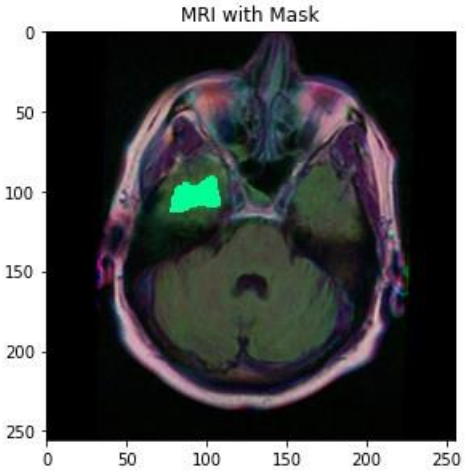

Figure 4: Example visualization from dataset

Results and Evaluation

The results for both algorithms fared decently well given all things considered, but in the end one model was clearly better than the other. That model being the ResNet. When looking at the two models, it was initially assumed that the VGGNet would perform better than the ResNet, because it was training on more parameters. The trade off here would obviously be it took a much longer time to run. In the end though the ResNet not only ran much faster then the VGGNet but it also had a higher overall precision, recall, f-1 score and overall accuracy on the same sized validation set which can be seen from the support number. We can see from figure 3, that the f-1 score, which is a blend of the precision and recall was .97 for non tumor images and .94 for tumor classification. Comparatively, past CNNs for image classification could only achieve up to 90% f-1 score, so this was a great result. Looking at the individual statistics of precision and recall it will be helpful to review both of the formulas before discussing the results. Recall is the number of true positives divided by true positives plus false negatives. Precision is the number of

true positives divided by true positives plus false positives. The equations for each are shown below.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

Figure 5 : Precision and Recall Formulas

When diagnosing brain tumors, recall is absolutly the more important metric. This is because we want to avoid false negatives at all costs. If you tell someone they do not have a tumor when they actually do this could have lethal consequences.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	390
1	0.98	0.90	0.94	200
accuracy			0.96	590
macro avg	0.97	0.94	0.95	590
weighted avg	0.96	0.96	0.96	590

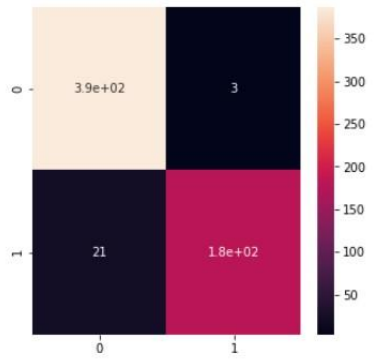

Figure 6: Classification Matrix for ResNet

Reviewing the same statistics below, it can be seen that on average all major indicators were approximately 1-8% lower for the VGGNet, depending on the evaluation metric. This is a huge dropoff, because these are not just evalaution metrics, but could mean the differnce between life and death for a patient. The recall for positive cases is 3% lower than that of the ResNet and the overall accuracy was 5% less.

0.9067796610169492

	precision	recall	f1-score	support
0	0.94	0.92	0.93	399
1	0.84	0.87	0.86	191
accuracy			0.91	590
macro avg	0.89	0.90	0.89	590
weighted avg	0.91	0.91	0.91	590

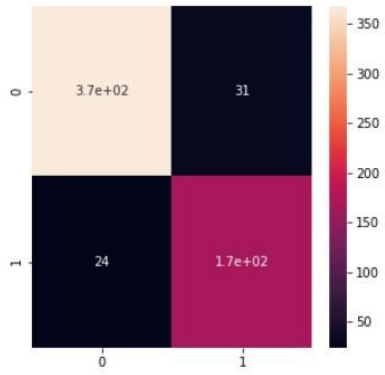

Figure 7: Classification Matrix for VGGNet

Next we will review the model loss for each. Loss values imply how poorly or well a model behaves after each iteration of an optimization. The lower the loss, the better the model. In other words, the loss function can quantify the difference between the expected outcome and the actual outcome. Accuracy on the other hand is a much more interpretable metric. It is simply a measure of the true positive plus the true negative divided by the total number of examples. The visualization below shows that for the ResNet, the loss stayed relatively low from 0 all the way through 50 epochs, without much variation, other than some noise around 15 epochs. On the other hand, the model accuracy greatly improved from

0-20 epochs, before flattening out, similar to a logrithmic curve.

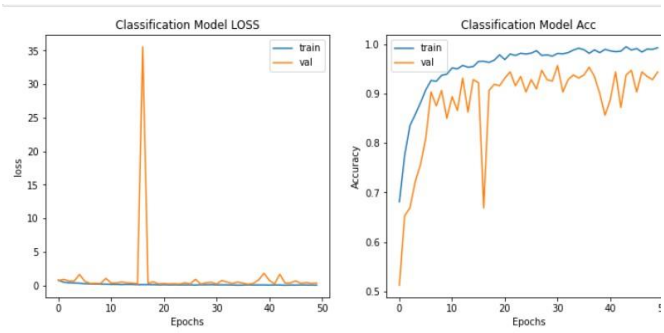

Figure 8: Loss and Accuracy for ResNet

The VGGNet model loss and model accuracy displayed below were quite different from the ResNet. The model began to slightly overfit the data around 6 epochs. This can be seen from the training set and validation set diverging. The loss and accuracy both performed worse as well.

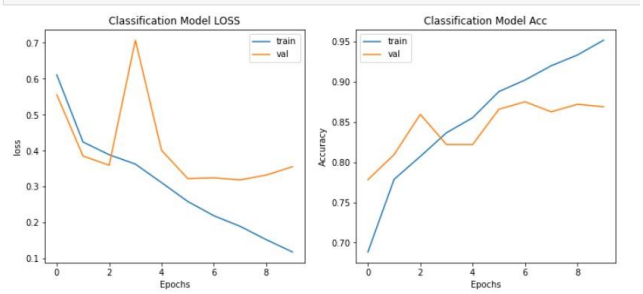

Figure 9: Loss and Accuracy for VGGNet

Lastly, we look at focal Tversky loss and the model Tversky score. These metrics were pulled from the segmentation portion of the models. A frequent problem of training segmentation networks is class imbalance in the dataset. To give an example, if you are segmenting an image and the pixels that are supposed to be segmented are only a small percentage of the total amount of pixels, then your accuracy metric becomes meaningless. This is becasue even if the model predicted an all black image it could still be rewarded with say 90% accuracy. The Tversky loss tries to solve this. It is an asymmetric similarity measure that is a generalisation of the dice coefficient and the Jaccard index. The equation is displayed below.

$$TI = \frac{TP}{TP + \alpha FN + \beta FP}$$

Figure 10: Tversky Index

The focal Tversky loss is just a generalisation of the Tversky loss. In the case of both models shown below, the focal Tversky loss and total Tversky score were similar and followed similar overall patterns. Segmentation for each was quite comparable.

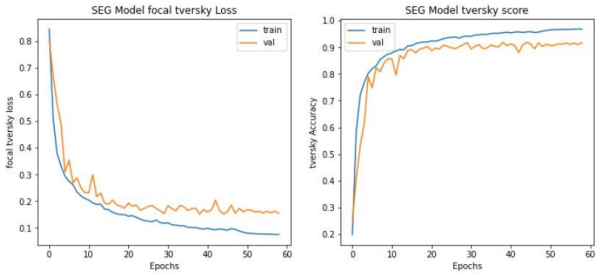

Figure 11: Focal Tversky Loss and Tversky Score for ResNet

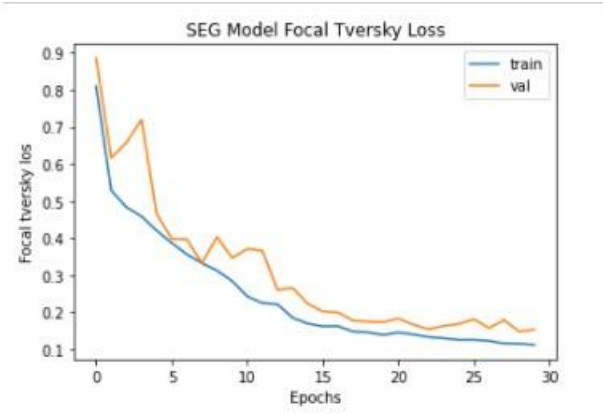

Figure 12: Focal Tversky Loss for VGGNet

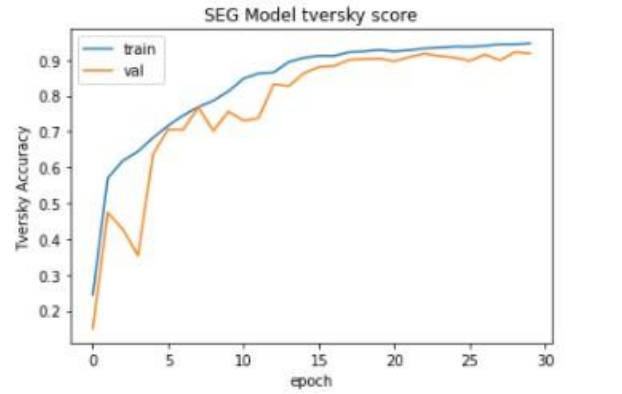

Figure 13: Tversky Score for VGGNet

For the segmentation of the brain tumors I created a pipeline that first ran the classification CNN to detect if there was a tumor or not. If the CNN reported 90% or higher certainty of the tumor, then it was run through a segmentation CNN. A sample images of this is shown below.

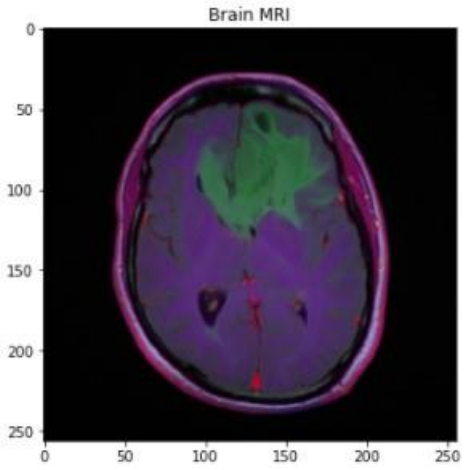

Figure 14 Segmentation Visualization for ResNet Part 1

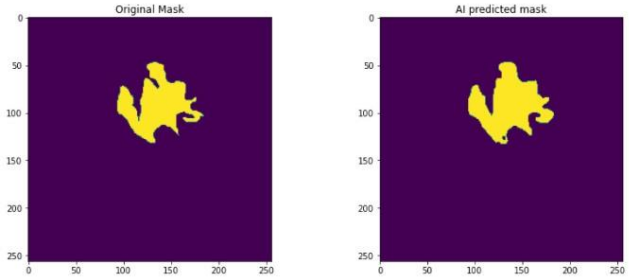

Figure 15 Segmentation Visualization for ResNet Part 2

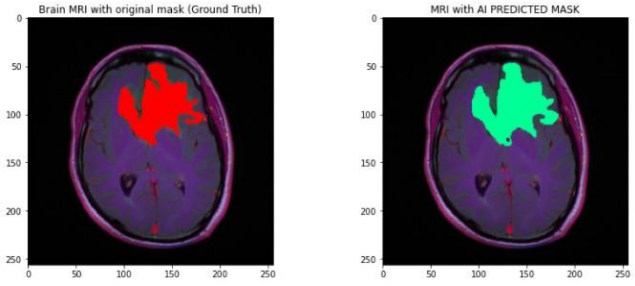

Figure 16 Segmentation Visualization for ResNet Part 3

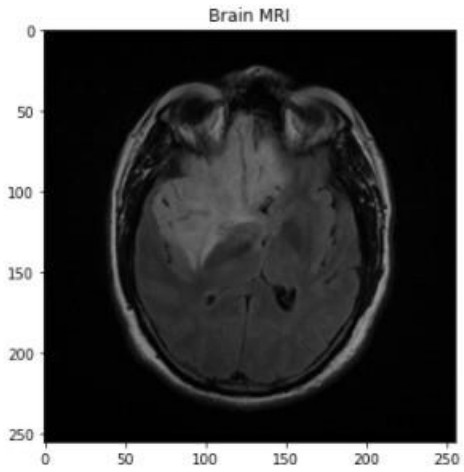

Figure 17 Segmentation Visualization for VGGNet Part 1

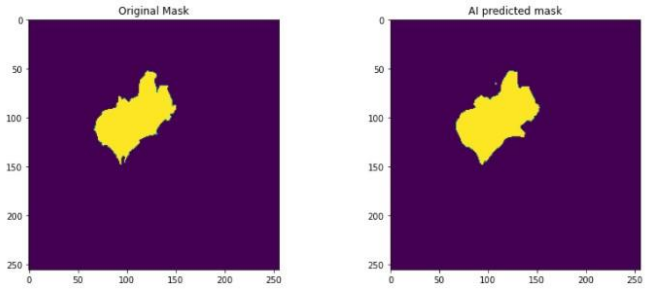

Figure 18 Segmentation Visualization for VGGNet Part 2

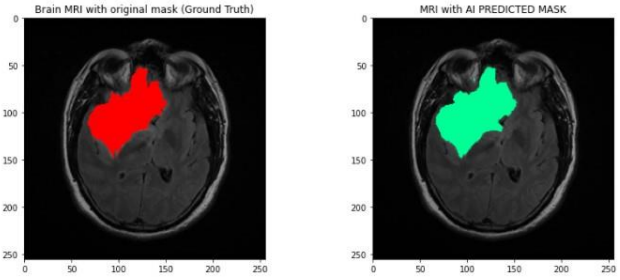

Figure 19 Segmentation Visualization for VGGNet Part 3

Based off both the ResNet and VGGNet labeled images you can see each was extremely accurate. In figures 15, 16, 18 and 19 the left image was the ground truth and the right image was the AI predicted segmentation. Although all surfaces and corners were not identical the overall shape and location was segmented to a high degree of accuracy.

The primary challenge for this project was the computing time for the algorithms. The ResNet CNN took approximately 10 hours to run just for the classification portion, which is extremely costly in both time and money. The VGGNet took approximately double that time for classification. Running the CNN through a cloud computing service such as AWS, will prove to be a huge time saver.

4 Discussion

This project has been a great learning experience and hopefully will shed some light on the differences in CNN architectures and how they can be used for medical imaging help. CNNs at their core consist of two main things. Those being feature learning and then classification. The feature learning is a 4-5 step process that generally involes an input, convolution and relu activation, pooling, another layer of convolution and relu and finally more pooling. The classification consists of a flatten layer, a fully connected layer and then finally an activation function for output that can best be described as probabalistic distribution. Both of these algorithms consisted of this, but with slightly different results in the end.

Interestingly enough all metrics leaned in favor of the ResNet, making it the easy winner, but judging on just visualization in the end with segmentation, it was quite hard to tell the

difference between one algorithm being better than the other.

The reason I mention this is because doctors in many cases would just care about the segmentation part of the model. This would aid in getting a conclusive answer as to where the tumor has spread, when the contrast in the image isn't always clear.

What I found to be the biggest challenge was computing power it took to run these models. CNNs are highly complex deep learning models that consist of millions of paramters, but never would I have thought that it would take 20 hours to run one model to get weights. This proved to be especially costly in that with the VGGNet I unfortunately could not produce a classification model loss or accuracy visualization the first time through, so I had to redo it.

For future applications I would absolutly put more focus on the efficiency of thr algorithm to cut down on the running time. Biggest is not always the best.

Conclusion

Applying machine learning methodologies to medical classification has the potential to save lives. Brain tumors are a fatal disease that can be extremely challenging to diagnose with high accuracy. My hope for this project is to aid in the clinical outcomes of patients who may have a brain tumor.

This project proved past research to be true in that the ResNet reigned supreme in terms of not only accuracy, but runtime as well. It beat the VGGNet by up to 8% in certain metrics and when pairing that with the fact that every sliver of a percent could save a human life, it is pertinent to use what is proven to be the most effective model.

In terms of future work I believe continued experimentation with CNNs is important. ResNet and VGGNet are not the only two well known CNNs out there. Trying out others against the ResNet would be extremely useful and could provide interesting insight to improvements in image classification.

References:

(1) Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9,611–629(2018). https://doi.org/10.1007/s13244-018-0639-9

(2) Ayadi, W., Elhamzi, W., Charfi, I. et al. Deep CNN for Brain Tumor Classification. Neural Process Lett 53, 671–700 (2021). https://doi.org/10.1007/s11063-020-10398-2

(3) Gu X, Shen Z, Xue J, Fan Y and Ni T (2021) Brain Tumor MR Image Classification Using Convolutional Dictionary Learning With Local Constraint. Front. Neurosci. 15:679847. doi: 10.3389/fnins.2021.679847

(4) Sajjad, Muhammad & Khan, Salman & Muhammad, Khan & Wu, Wanqing & Ullah, Amin & Baik, Sung. (2018). Multi-Grade Brain Tumor Classification using Deep CNN with Extensive Data Augmentation. Journal of Computational Science. 30. 10.1016/j.jocs.2018.12.003.

(5) Dataset used from Kaggle: https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation