Deep Learning for Brain MRI Classification of Alzheimer's Disease

University of California San Diego
COGS 181
Tommy Shen
November 9, 2023

Abstract

Neuroimaging plays an integral role in diagnosing Alzheimer's disease (AD). Structural magnetic resonance imaging (MRI) provides indicators of the brain's progression from mild cognitive impairment (MCI) to AD. Identifying patients with AD early allows for the greatest impact of preventative measures and easiest lifestyle transitions. There has been a recent surge in popularity in deep learning and its ability to capture complex patterns. I aim to implement a convolutional neural network to autonomously diagnose AD within a brain MRI scan.

1 Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease commonly associated with deficits in memory and cognition. Patients usually begin with mild cognitive impairment (MCI), which is characterized by symptoms of forgetfulness and in more severe cases, diminished judgement and critical thinking skills. The disease is the leading cause of dementia with no universally effective treatment or medication.

Patterns in dendritic and neuronal losses, which are believed to contribute to cerebral atrophy, are distinct between diseases. Studies have observed that different diseases are highly dependent on specific neuronal vulnerabilities and a disease's regional expression. The typical onset of atrophy in AD begins at the entorhinal cortex, followed by the hippocampus, amygdala, and parahippocampus. In later stages, the degeneration can spread throughout the whole brain [1].

MRI provides a high-resolution spatial image which easily contrasts soft and hard tissues, making it an effective tool for analyzing the brain's anatomy. The procedure is non-invasive, posing fewer health risks than other common AD neuroimaging techniques such as positron emission tomography and computed tomography [2]. Though a combination of imaging tends to factor into a patient's diagnosis, MRI is chosen for the classification task due to its sensitivity to the atrophy indicative of AD [1].

The manual diagnosis of AD in large scale MRI datasets is difficult and time consuming, making deep learning an attractive solution to assist clinicians. Traditional machine learning methods related to AD have largely centered around MRI segmentation, in which the brain is partitioned into well-defined regions to highlight critical brain areas. This strategy however, relies elaborate feature engineering and specialized expertise [2]. This paper explores a broad application of convolutional neural networks as a potential MRI pre-screening tool.

2 Methods

2.1 Dataset and Data Preparation

The dataset used for this model is a publicly available compilation of pre-processed brain MRI scans comprised from various sources including hospital data, dedicated Alzheimer's disease research organizations, and publications. The authors of the aggregated dataset, filtered the scans to include only horizontal cross-section scans and standardized all images to a size of 128x128 pixels with the brain centered accordingly. A few samples from the dataset are shown in Figure 1. The images are separated into 4 classes of varying severity.

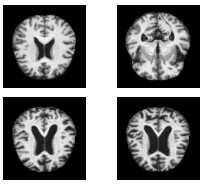

Figure 1: Example MRIs

There are an equal number of non-demented and demented cases. Of the demented MRI scans, 70% were classified as mild cases. I further simplified the dataset into a binary format, resulting in perfectly balanced positive and negative classes. Exact class counts are depicted in Table 1.

Class	Count
Non_Demented	3200
Very_Mild_Demented	2240
Mild_Demented	896
Moderate_Demented	64

Table 1: Dataset Distribution

With the image dimensions of 128x128 pixels the number of features totals to 128^2 with each feature vector entry $x_i \epsilon [0, 255]$ representing a grayscale intensity. Because pixels exterior to the brain have a 0 grayscale intensity, vector representations are sparse. Consequentially, certain pixels like those encoding the edges of an image are also expected to be uniformly 0, thus uninformative.

The dataset was shuffled and split into a 75/25 ratio for training and testing datasets.

2.2 Model Evaluation

Three classifier performance metrics are considered for this task: precision, recall, and F1. The formulas are as defined in Table 2.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 2: ResNet Architecture. From ResNet Publication [4]

Metric	Formula
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1	$\frac{2(precision*recall)}{precision+recall}$

Table 2: Performance Metrics

Precision explains how many positive case predictions were correct. This metric quantifies a model's susceptibility to false positives. A low precision score, in the context of autonomous medical diagnosis, is suggestive of a problematic tendency to over-predict positive cases.

Recall measures a model's ability to correctly detect a positive class. In context, this score quantifies how likely a model is to identify AD when a patient truly has the disease. A low recall demonstrates that AD cases tend to go undetected by the model.

F1 score is the harmonic mean of recall and precision. Though recall is often an emphasized concern in medical settings, a model should not be flagging every patient as a positive case as it would have little clinical utility. Precision is capable of capturing this behavior thus this final metric is included. F1 score, which aggregates both recall and precision, enforces a balanced cost between the two and was the primary metric considered for tuning.

2.3 Baseline Models

Two baseline models were introduced to compare against the final neural network. The first is a Support Vector Machine (SVM) kernelized by the radial basis function (RBF).The RBF kernel introduces non-linearity to SVM in a manner that mimics a binary k-Nearest Neighbors classifier without as many memory storage drawbacks. This model is a more efficient implementation of a popular initial classification model and serves as a reference point to a traditional machine learning method.

The second baseline model introduced is the neural network known as ResNet-18. ResNet-18 is composed of 18 convolutional layers and employs softmax activation. Figure 2, from the original publication for ResNet, depicts the ResNet architecture for various layer counts. It has typically been used for multi-class image classification on miscellaneous everyday images. I retrained this model on the MRI dataset to compare my own custom network against another robust deep learning model. Since neuroimaging data is very different from the ImageNet database which ResNet has been trained on, the pretrained weights were retrained on this dataset. Deeper ResNet networks can take hours or even days to fully retrain thus the smallest 18-layer architecture was selected.

3 Experiment

3.1 SVM

In consideration of the numerous uninformative image pixels in the data's high dimension feature vectors, principal component analysis was used to reduce the feature count to 6500. After performing 5-fold cross validation on F1 score, the best slack parameter was determined to be $C = 5$.

3.2 ResNet-18

The ResNet architecture was retrained with cross entropy loss and Adam optimization. A learning rate of 0.001 and a weight decay of 0.0001 were used but not fine tuned due to the costly training time of a single epoch. On 10 epochs with a mini-batch size of 40, the entire training process took approximately 17-20 minutes to complete on my device's local processor (I9-9750H). Initially, the 50-layer

variant of ResNet was considered, however due to its immense training times, the benchmark model was changed to ResNet-18. With this architecture the training loss appeared to converge around 10 epochs.

3.3 Scratch Network Architecture

I defined a simple 5-layer model consisting of 4 convolutional layers and a single pooling layer to speed up the training process and allow for easier hyperparameter tuning. Cross entropy loss and stochastic gradient descent were the loss function and optimizer defined for the training loop. The learning rate and momentum values were set to 0.001 and 0.9 respectively. Of the various architectures tested, the best performing model was as defined in Table 3.

layer	kernel size	output channels	stride
conv1	1x1	12	1
conv2	2x2	18	1
conv3	3x3	36	1
conv4	1x1	42	1
avgpool	2x2	NA	2

Table 3: Final Network Architecture

Some of the hyperparameters considered during validation were, the number of convolutional layers, the number of filters, convolution kernel sizes, the pooling function, mini-batch size, and number of epochs. A single pooling layer yielded better performance than multiple pooling layers. Conversely, adding a fourth convolutional layer with a small filter saw better performance than just two and three convolutions. The mini-batch size of 24 was the smallest size implemented but increased the model's overall fit. The neural network from scratch also appeared to converge near 20 epochs after 25 minutes. In this experiment, hyperparameter tuning was critical in building an effective network. Different combinations of hyperparameters resulted in F1 validation scores ranging from 0.647-0.913.

3.4 Training Loss Curves

Figure 3 depicts the average training loss per epoch for the ResNet architecture. As mentioned prior, the training loss for ResNet appears to stagnate at 0.60 between epochs 8-10.

Likewise, Figure 4 shows the average training loss per epoch for the new model designed in this experiment. The training loss improved beyond 10 epochs but experienced a plateau near 19-20. The minimum mini-batch training loss was 0.213.

3.5 Results

SVM was the best model, beating both of the other models in every metric when evaluated on a withheld testing

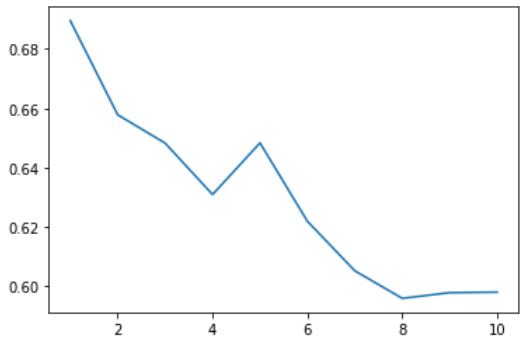

Figure 3: ResNet Avg Mini-Batch Training Loss

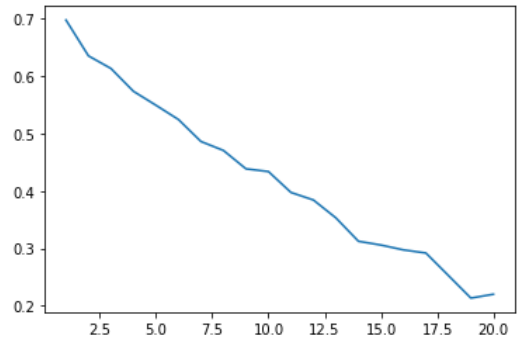

Figure 4: Scratch Net Avg Mini-Batch Training Loss

dataset. SVM boasted a perfect precision score paired with a high recall score of 0.997. An F1 score of 0.998 reflects nearly perfect classification.

Model	Precision	Recall	F1
SVM	1.0	0.997	0.998
ResNet-18	0.431	0.678	0.527
Scratch Net	0.896	0.930	0.913

Table 4: Model Performance

The next best model was the neural network from scratch, which held an F1 score of 0.913. It's precision and recall scores were 0.896 and 0.930 respectively. Unlike the SVM model, this model was weaker in precision rather than recall.

Finally, the ResNet-18 model achieved an F1 score of 0.527 with a precision of 0.431 and a recall of 0.678. This performance was significantly worse than the other two models.

A summary of each model's performance can be found in Table 4.

4 Discussion & Conclusion

While the stellar performance of SVM could potentially appear discouraging for deep learning methods for MRI classification, the dataset used in this experiment was pre-standardized. The pre-processing procedure, while highly beneficial, is tedious and could ultimately hurt the utility

of traditional machine learning algorithms, as they don't learn complex patterns in unprocessed MRI data quite as easily as convolutional neural networks.

Despite every model being trained and tested on the standardized dataset, the implementation of deeper networks could supersede a classification tool's need for inputs to be subjected to uniform reformatting by learning the complex MRI patterns in AD in different orientations. This would eliminate any time-consuming image processing steps and create a direct pipeline for autonomous AD diagnosis.

ResNet-18's early training loss convergence potentially indicates that an 18 layer network is too complex for the dataset. Though this may not be the case for unstandardized MRI data, a 5 layer convolutional network, like the one designed for this experiment, appears to be sufficient for AD classification when images all are cropped and centered. This model's architecture might serve as a starting point for future models in MRI classification.

Since AD tends to have a distinct atrophy pattern compared to other diseases, small filter sizes and fewer pooling steps are likely to improve classification as such methods might capture spatial information more effectively.

Even though the results of this paper may appear encouraging for the medical field, machine learning methods as a whole are susceptible to biases in the data they are trained on. Any biases in the collection methods could hurt a model's generalizability thus a disease classification decision should still be augmented by a clinical professional.

References

[1] Johnson, K. A., Fox, N. C., Sperling, R. A., & Klunk, W. E. (2012). Brain imaging in Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, 2(4), a006213. https://doi.org/10.1101/cshperspect.a006213

[2] Yamanakkanavar, N., Choi, J. Y., & Lee, B. (2020). MRI Segmentation and Classification of Human Brain Using Deep Learning for Diagnosis of Alzheimer's Disease: A Survey. *Sensors* (Basel, Switzerland), 20(11), 3243. https://doi.org/10.3390/s20113243

[3] Frisoni, G., Fox, N., Jack, C. *et al.* The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology* 6, 67–77 (2010). https://doi.org/10.1038/nrneurol.2009.215

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.48550/arXiv.1512.03385