CS230

A Preclinical Neurodegenerative Disease Classifier

James Chao and Alexander Karr
Department of Computer Science
Stanford University
jbchao@stanford.edu | ajkarr1@stanford.edu

Abstract

We have created a multi-label image classifier, whose neural network architecture bases itself off of vision transformers. It takes in inputs of 3D, axial, fMRI brain images, and categorizes them as either healthy, or members of various neurodegenerative disease category. Neurodegenerative disease classifiers, working from MRI data, are a well-explored space. However, not only are extant methods still dominated by generally-less-performant Convolutional Neural Networks (as compared to the pattern-finding abilities of transformers), *far* fewer classifiers have taken on longitudinal studies and attempted to diagnose preclinical neurodegenerative conditions. Our hope is that earlier diagnoses of these assorted neuropathologies can lead to earlier interventions, and longer, healthier lives for a global population whose age distribution is shifting older and older.

1 Introduction

Our work focuses on neurodegenerative disease. We are developing a diagnostic tool that works from a dataset of clinically labeled structural brain imaging to attack a biomedical problem that we see as both unresolved and computationally tractable for a discerning enough neural network.

Neocortical degeneration in various brain structures may begin to occur upwards of a decade[1] before a patient could be diagnosed with full-on clinical Alzheimer's Dementia. As a disease with no known cure, but many effective prophylactic measures, diagnosing AD early on allows us to preserve cognitive function and quality of life where otherwise it could have declined precipitously. Most MRI-driven Alzheimer's research has compared three categories of cortical health through brain images – healthy brains (no neurodegeneration), patients with Mild Cognitive Impairment (MCI), and patients with full-on Alzheimer's Dementia where their symptoms disrupt their daily lives. Despite the National Institute on Aging (include footnote) including it within their recognized stages of Alzheimer's Dementia, far less research has considered an intermediate category lying between healthy and MCI patients. This is the Preclinical stage, where patients exhibit no symptoms, but their neuronal structure already shows signs of deterioration. These are the patients with the most therapeutic promise, yet we are under-diagnosing them.

The damages to Preclinical brains are subtle, so while existing MRI-based deep learning models capable of prognosticating Preclinical AD boast impressive detective abilities, there is always room for refinement. This is especially true in a literature that is still very much dominated by Convolutional Neural Networks, that fail to take advantage of Transformer's attention features. Using a longitudinal neuro-imaging, clinical, and cognitive dataset named OASIS-3, we are therefore aiming to create a multi-label, vision transformer image classifying tool.

[1] https://www.nia.nih.gov/health/alzheimers/causes

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

It takes in MRI imaging from 1,098 individuals collected over the course of 15 years. The dataset contains 2,000+ sessions of MRI data, and we will be using images from the axial plane. Patient ages range from 42 to 95 years old. Our tool will classify brains into five categories:

1. Healthy Brains (No dementia)
2. Preclinical Alzheimer's dementia
3. Preclinical otherwise Dementia (Vascular Dementia, Normal Pressure Hydrocephalus, etc.
4. Alzheimer's dementia
5. Otherwise dementia

To quote the paper that was our primary inspiration for our project, which also only used brain imaging data to try and diagnose preclinical stage Alzheimer's, inability to "adapt the models to different factors that could be critical when predicting AD, such as the age or the sex of the patient... makes the classification task more challenging." Yet the ubiquity of the MRI in data collection of the potentially brain damaged, as compared to other rarer measures like genetic sequencing of patients, or PET scans, or answers to the TADPOLE Challenge questionnaire, emphasizes the value of predicting neurodegenerative disease from *exclusively* MRI data.

2 Related work

Preclinical Stage Alzheimer's Disease Detection Using Magnetic Resonance Image Scans

This paper will be our primary source as it proposes a neural network architecture employing MRI data to predict preclinical AD. Their approach clearly has virtues, as they were able to obtain an accuracy score of over 90 percent. The paper has very few citations, however and is very recent, which speaks to the small volume of research currently funneled into the ambitious task of preclinical dementia diagnosis. Although we are employing the same Oasis-3 longitudinal dataset, we are trying to better appreciate the range of its clinical labeling – this original study's tool simply binarizes preclinical Alzheimer's Dementia brains versus otherwise brains. We are training a tool that can better diagnose a variety of neurodegenerative conditions, and is forced to learn features that enforce strict borders between them. This is essential, because the treatment paths for preclinical Alzheimer's dementia, versus the early stages of other neurodegenerative conditions (vascular dementia for instance), can vary significantly.

Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation

This paper proposes a model architecture using CNNs to classify AD amongst MRIs. It does not use the transformer architecture with self-attention, opting rather to use CNNs all the way through. While its approach can no longer be considered state-of-the-art, learning from different model architectures may give us insights into how we can tweak our architecture to fit out needs. Our attention units and queries are better modeling internal relationships amongst our parameters and data, but this paper got us started. Its discussion of the distinctions amongst 3D versus 2D processing of the image

data were particularly useful. The paper also provides a systematic literature review of the extent MRI-analyzing CNN models in current academic circulation for diagnosing AD.

Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images

This article describes the multi-modal employment of both MRIs and PET data to capture both structural and functional data. We wanted to evaluate the diagnostic efficacy of this joint approach, pre-processing the MRI data and then co-registering the PET images, which have the exact same dimensions and are obviously of the same brain, with a rigid transformation using FSL-FLIRT. We chose to forego this multi-modal approach for two reasons: first, its reported accuracy is not superior to the first paper we were citing (our greatest inspiration). Secondly, it requires both metabolic and structural imaging of each brain, and we want our tool to be utilizable in as many situations as possible (i.e. when just MRI images are available).

Alzheimer's Disease Diagnostic Guidelines

This National Institute of Health paper did a good job of giving us the lay of the land regarding how clinicians label their patients' cognitive abilities. This is a higher level paper, presenting no technical tool, but at least gave us a greater sense of the behavioral changes doctors were identifying, that produced the labels for the images we are computing on.

A Deep Siamese Convolution Neural Network for Multi-Class Classification of Alzheimer Disease

This is yet another CNN-based model. Its unique utility to our project, however, was its strategies to avoid overfitting, training its model as it did on a small dataset (one of the OASIS datsets, but not OASIS-3 like we used). They had greater need to data augment, due to their only having 382 images in their data set. The below image details their augmentation techniques, per image, that they used to enhance the generalizability of their model to a test set (and actual patient diagnoses). We were originally rotating our images more significantly than ten degrees in our own augmentation testing and it was drastically reducing our accuracy, so their augmentaiton guidelines were of great help. This CNN also classifies multiple states, with No Dementia (ND), Very Mild Dementia (VMD), Mild Dementia (MD), and Moderate AD (MAD), so it is exhibiting the more-general-neurodegenerative-diagnosability that we are looking for in a tool. Again, however, it fails to exhibit the Transformer architecture we are aiming for.

Table 3. Data augmentation.

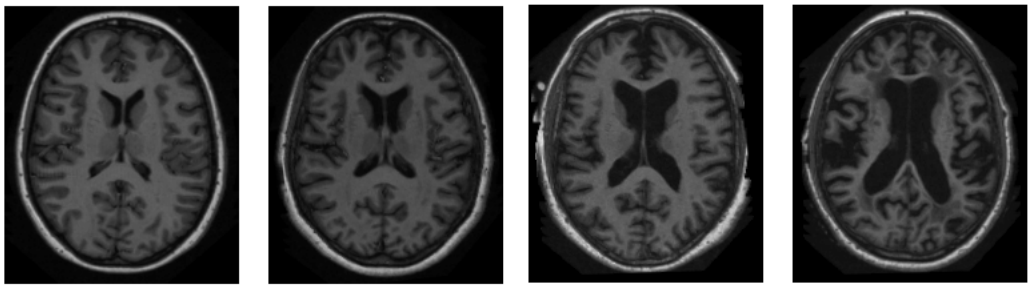
Rotation Range	**10 Degree**
Width shift range	0.1 Degree
Height shift range	0.1 Degree
Shear range	0.15 Degree
Zoom range	0.5, 1.5
Channel shift range	150.0

A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images

This paper proposes a transfer learning approach, to account for the small dataset environment often common to the biomedical imaging space, due to privacy concerns and differences in imaging techniques/patient populations. It transfers over a pre-trained model from the VGG family architecture, that was trained on the ImageNet Large Scale Visual Recognition Challenge benchmark to classify 1,000 different objects. The study authors then customize the last two fully connected layers, and obviously the final classification layer to the problem of diagnosing the same four stages of dementia from above (ND, VMD, MD, and MAD).

3 Dataset and Features

As we said in the introduction, we used the OASIS-3 dataset[2]. From OASIS-3, we processed 2,191 MRI scans from 1,172 different patients, aged 42 to 95. For each MRI scan, we used a script to cross-reference it with OASIS-3's database of Aging and Disability Resource Center (ADRC) clinical diagnoses. These diagnoses included the following classes: "Cognitively Normal," "AD Dementia," "Vascular Dementia," and "Uncertain Dementia." Of the 2,191 scans, only 103 were labelled as one of the dementia labels, and there was no preclinical labeling. As such, one of our major pre-processing tasks was to write a script that re-labled our data to consider preclinical neurodegenerative disease an actual labling class. The way we did that was to consider all "healthy brain" scans of a patient who later demonstrated dementia to be "preclinical dementia" images. Given that the upper-bound of time difference between any first and final image for a patient was around ten years, and neocortical degeneration is hypothesized to begin at least a decade before behavioral effects manifest, we consider this labeling judgment fair. It makes us confident that we are creating a tool whose loss function, with the preclinical classes, is actually leading the model to learn the very subtle necortical-degenerative image features that underlie preclinical dementia. After data relabelling, we ended up with 1590 patients who were "Cognitively Normal" and 601 who were either preclinical or had full-on dementia.



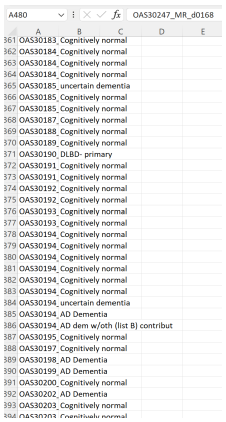
4 Methods

4.1 Data Processing

As mentioned in the Dataset section, we obtained 2191 MRI scans from the OASIS-3 database. We also obtained two CSV's, one matching MRI scans to patient IDs and one matching patient IDs to ADRC diagnoses. We then used a script to automatically match each MRI scan to a corresponding diagnosis within one year of the scan. This script gave us our labels file, matching each scan to an ADRC diagnosis.

However, since the ADRC diagnoses do not include prelinical data, and sometimes contained irrelevant information (e.g. AD Dementia w/ depression; non-contributing), we wrote another script to relabel the diagnoses into our desired five classes: Cognitively Normal (CN), Other Dementia (OD), Alzheimer's Dementia (AD), Preclinical OD (PO), and Preclinical AD (PA). We then wrote a script to populate the labels file with one-hot vectors corresponding to each class. We also wrote a script to rename all the files and the MRI entries in the labels file to match.

[2]https://www.oasis-brains.org/



Each MRI scan was given in .nii.gz medical imaging format and contained data for a 3D image of the whole brain. Since we did not have the capability to work with 3D data, we took axial slices of each image (slicing along the Z axis) and used those 2D representations as our input. Since our dataset was so large, we were only able to use the middle 6 slices of each image with the storage space and compute power that we had. We then resized each 2D slice into a 1x256x256 array, with 1 being the number of channels (grayscale).

Since the vast majority of images were cognitively normal, we used data augmentation to increase the number of samples of prelinical and demented patients. For each demented scan, we generated three rotated images (with the rotation not exceeding ten degrees), and for each preclinical scan, we generated ten rotated images. We then undersampled from the cognitively normal class to balance out our data distribution.

We then built our data into numpy arrays and split it into 70% train, 15% validation, and 15% test, making sure that no patients' multiple scans were split in the process.

4.2 Model

Since the original transformer was designed for sequence-to-sequence language models, it has to be adapted to work with image data. To convert each image into a sequence, we use a patch embedding layer similar to the idea mentioned in Figure 1:

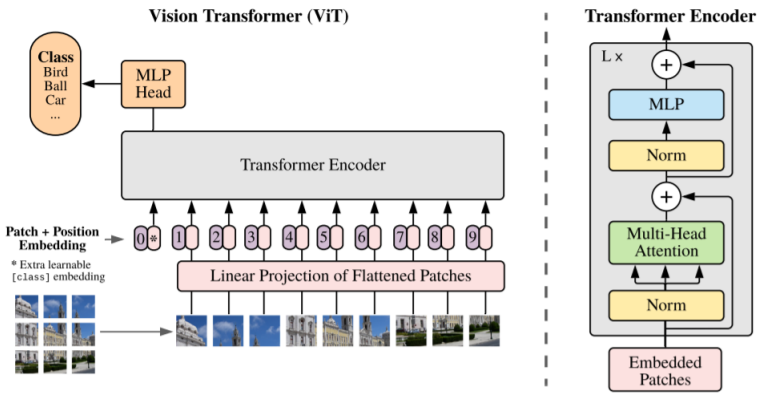


Figure 1: Vision Transformer architecture

Our model architecture consists of the following three layers: Patch Embedding, Transformer Encoder, and Output.

4.2.1 Patch Embedding

The input is first fed into the patch embedding layer. For each 1x256x256 image, we break it up into 16 patches. Then, each patch is flattened and linearly projected into the model's hidden size so that it can be fed as a sequence to the transformer encoder. However, before passing it to the encoder, we add an extra trainable parameter to each patch-sequence whose sole responsibility is predicting the class at the end. Furthermore, for each of the patch-sequences, we imbue a trainable relative positional embedding that encodes where the patch was in the original image.

4.2.2 Transformer Encoder

The transformer encoder consists of three (a hyperparameter we tuned) encoder blocks where the first block's output is fed into the second block's input and so forth. In each encoder block, the input is first fed into a layer norm. Then, Query, Key, and Value tensors are derived from the input by feeding the input through a linear projection layer that triples its size. The Q, K, and V are then fed through a multi-head attention layer with 2 heads (another hyperparameter). A residual connection is implemented before and after the attention layer. The output is then fed through a layer dropout layer (randomly dropping the residual connection). Then, the output is fed through two fully-connected layers, with a residual connection over them, using leaky ReLU and dropout after the first FC layer and a layer dropout after the second layer. Finally, the output of all three encoder blocks are fed through a layer norm layer.

4.2.3 Output

Before the transformer output is fed to the output layer, the class token we implemented in the patch embedding is extracted, and this singular token is fed to the output layer. The output layer is a standard multiclass classification layer, including a layer norm, a linear layer projecting to the number of output classes (5), and finally a softmax layer.

5 Experiments/Results/Discussion

We used Adam for our optimizer with standard Beta1, Beta2 and a learning rate of 0.005. We arrived at this learning rate after much experimentation, having both dramatically over and under-shot it. At this magnitude, the learning rate was just small enough to stop the model from overshooting and just big enough (not 0.001) to get out of plateaus/generally expedite learning. Our mini-batch size of 128 struck a good balance between the subtle regularization you get from batch sizes small enough to introduce general noise to the measurement, and one big enough to smooth out the parameters in any individual batch. We used Crossentropy loss. We found that training didn't improve past 50 or so epochs, and so set our number of epochs to be 50. For our metrics, we used Accuracy, Precision, Recall, and F1.

At first, we obtained what seemed like promising results, with a val/test accuracy of **76/66%,** with our formula for accruacy being simply how many instances of **yhat** have the same argmax as the corresponding instances of **y**. However, we found that the model was simply predicting "Cognitively Normal" for everything since there was a large data imbalance. As such, we used data augmentation to produce roughly equal sample sizes for each class, ending up with around 1000 samples of each class. We also switched over to using the official Pytorch metrics. When we ran our model again, with both the augmented data and the original data, we found that accuracy stayed at 0.2 the entire time, effectively meaning that the model is just randomly guessing. We checked to make sure that the gradients were being properly backpropogated, and though they were, we found that all the gradients quickly approached zero as training continued. This looked like the vanishing gradient problem, which we attempted to solve with more regulation, but were unable to make any progress. If we were to have more time, we would continue debugging.

We have several possible explanations for why our model would not train correctly:

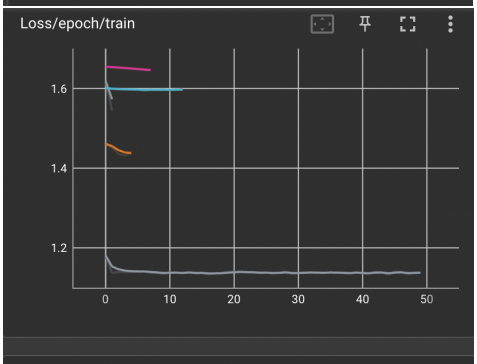
1. Our model architecture is not robust enough for image data.
   We assumed that by taking 2D axial slices of the MRI data, we would be able to treat each image like a normal image commonly used in image classification scenarios. As such, we implemented a version of the vision transformer. However, it is possible that this architecture is too simple to successfully model multiclass dementia prediction from MRI scans.

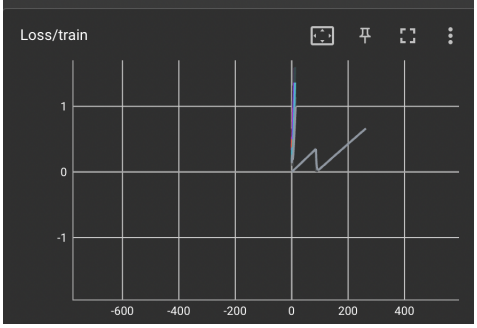
2. In order for complicated MRI data to be used with image classification, it first has to be fed through a massive pre-trained network.
   The authors of the binary preclinical classification paper first fed their input through VGG-16 before feeding that output into their own architecture. This is so that VGG-16 can handle most of the image classification parts, and the transformer can adapt to the more MRI-focused nuances. However, we did not have the compute power for VGG-16, and were unable to get our data onto AWS as the data files were too large.

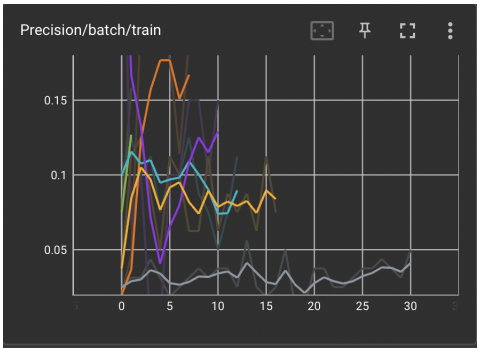
3. Not enough data
   Since only around 80 scans were considered preclinical and only around 400 scans were considered demented, there was a significant class imbalance. Even with data augmentation rotating these images, there are still very few unique preclinical images for the model to learn from. Perhaps we need a larger dataset with more balanced classes to accomplish significant training.

We are pretty sure the problem does not arise from our hyperparameter choices, as we experimented with batch sizes between 4 and 256, learning rates between 0.0001 and 0.01, different optimizers such as SGD, different number of heads, different hidden sizes, etc. As such, we have diagnosed our problem to be a fundamental incapability of our model to work with complicated MRI data on limited computational resources.

Here are some of the failed training graphs that we had:







Our final metrics were train/val accuracy = 0.25, train/val F1 = 0.22/0.20, train/val precision = 0.17/0.2, and train/val recall = 0.2. Our test accuracy, precision, recall, and F1 were 0.2, 0.16, 0.2, 0.17.

6 Conclusion/Future Work

6.1 Conclusion

It is difficult to make any algorithmic comparison between our and other tools, primarily Convolutional Neural Networks, on account of differences in data composition, augmentation strategies, and numbers of classes to classify. We could not realistically expect the same predictive accuracy from our classifier as many of the tools circulating in the literature, because their usual classes of mild cognitive impairment and full-on dementia present as much more dramatic changes in neocortical volume than any preclinical conditions do. It is nonetheless reasonable to expect that our model be able to adapt to MRI data with a standard vision transformer architecture. Our implementation does train, so we are fairly certain that the majority of the problem comes from our data pre-processing, due to the complicated nature of MRI data. While our tool still has much to learn much about the relationship between preclinical dementia and structural brain shifts, *we* have learned much about the importance of rigorous data curation. Since we both believe in the importance of this topic, we will continue to work on the model after the class finishes to steer it towards better feature learning.

6.2 Future Work

Besides fixing our model/dataset, we would explore multi-modal approaches for neuro-degenerative disease classification – especially tools resembling our third reference, that jointly analyze both fMRI and PET scans to diagnose disease. In preclinical dementia, the shifts in neocortical volume are so small that learning to recognize them at the exceptional rate one would want from any medical tool, affecting peoples' future paths of treatment, probably requires a ton of data. However, adding the metabolic information that PET scans measure opens opens up a whole new channel of discernment, where you could train a model both on structural changes, and shifts in glucose concentrations throughout regions in the brain.

7 Contributions

The work distribution was fairly balanced in all of the sections – though James did more of the computational work and Alex more of the report. Alex has a biocomputational background and James had taken a course on transformer architecture before, so we had different expertises that we had to both explain to one another to get each other up to speed.

References

[1] Altay, Fatih, et al. "Preclinical Stage Alzheimer's Disease Detection Using Magnetic Resonance Image Scans." ArXiv.org, 28 Nov. 2020, https://arxiv.org/abs/2011.14139v1.

[2] "Alzheimer's Disease Diagnostic Guidelines." National Institute on Aging, U.S. Department of Health and Human Services, https://www.nia.nih.gov/health/alzheimers-disease-diagnostic-guidelines: :text=Preclinical

[3] Lu, Donghuan, et al. "Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease Using Structural MR and FDG-PET Images." Nature News, Nature Publishing Group, 9 Apr. 2018, https://www.nature.com/articles/s41598-018-22871-z.

[4] Mehmood, Atif, et al. "A Deep Siamese Convolution Neural Network for Multi-Class Classification of Alzheimer Disease." MDPI, Brain Sciences, 5 Feb. 2020, https://www.mdpi.com/2076-3425/10/2/84

[5] Mehmood, Atif, et al. "A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images." ELSEVIER, Science Direct, 15 Apr. 2021, https://www.sciencedirect.com/science/article/pii/S0306452221000075.

[6] Wen, Junhao, et al. "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation." ELSEVIER, Science Direct, July 2020, https://www.sciencedirect.com/science/article/pii/S1361841520300591?via%3Dihub.

[7] Zuppichini, Francesco. “Implementing a Visual Transformer in Pytorch.” Medium, Towards Data Science, 9 Nov. 2022, https://towardsdatascience.com/implementing-visualttransformer-in-pytorch-184f9f16f632.

[8] Paszke, Adam, et al. “Automatic Differentiation in Pytorch.” OpenReview, 28 Oct. 2017, https://openreview.net/forum?id=BJJsrmfCZ.