Visualizing CNNs for Interpretable Alzheimer's Diagnosis Through Neuroimaging

Zi Ying Fan
Stanford University
zyfan@stanford.edu

Elissa Li
Stanford University
elissali@stanford.edu

Darian Martos
Stanford University
dtmartos@stanford.edu

1. Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative disease which starts in the temporal lobe and is the cause of approximately 70% of cases of dementia in seniors [6]. Diagnosis can be challenging due to the fact that neurological atrophy is common in elderly patients generally. A differential diagnosis therefore requires the ability to discriminate between typical neurological atrophy in elderly patients and AD-related atrophy. Machine learning models present a natural solution to this problem by virtue of their ability to recognize brain tissue alterations that humans may have a harder time discerning. Deep learning approaches in particular, such as convolutional neural network (CNN) or recurrent neural network (RNN) models, have the added benefit of being able to use neuroimaging data without relying on hand-crafted features and have proven to perform with good accuracy [4][1].

Research in the deep learning space as applied to AD diagnosis has tended to focus on classification accuracy, particularly when pushing neural networks as a diagnostic tool. However, due to the black-box nature of deep learning models, uptake in the medical community has been met with some reticence. Medical professionals are interested not only in accuracy but also in the model's decision process, especially given the possibility of confounding image features. For example, AD may co-occur with Frontal Lobe Dementia, which shares similar symptoms and patient profiles [8]. Similarly, because AD tends to occur in elderly patients, there is the concern that the model predicts for confounding features like age, instead of the disease itself. The opacity of deep learning models makes absolving these concerns difficult and also presents a challenge to medical professionals seeking to verify diagnoses. Clear interpretation of classification processes would increase the credibility of such models for use in the healthcare industry.

Towards that end, we are interested in exploring visualization techniques to elucidate the relationships and features that CNNs pick up on when diagnosing AD from neuroimaging. Specifically, we explore visualization methods which use heatmaps (relevance maps) to highlight regions in the brain scan that are particularly relevant for the model's diagnosis. We start by comparing four visualization methods in the tradition of [5]: basic sensitivity analysis, guided backpropagation, occlusion, and area occlusion, as wrapped around a 3D CNN trained on AD and control group (NC) brain scans. From there, our goal is to explore new backpropagation-based methods that can remove the effect of confounding variables, such as age, from the visualization process, in hopes of yielding a heatmap that is more robust to applications like AD diagnosis, where confounding features may be easily picked up by deep learning models.

2. Problem Statement

Our initial goal for the purposes of this milestone is to set up baseline classification and visualization methods, with the overarching purpose being the visualization of a high-performing CNN classifier. Given the success of deep learning classification models in this space, we expect our classifier to achieve approximately 85% accuracy on the test set.

From there, we compare four visualization methods in the tradition of [5], further described in the next section: (1) basic backpropagation (sensitivity map), (2) guided backpropagation, (3) basic occlusion, and (4) brain area occlusion. We evaluate the success of our visualization methods via visual comparison to the heatmaps supplied by [5], and also via comparison of relevancy breakdowns per brain area. We expect our results to approximately match those found in [5] with some variability, i.e. for similar brain slices, we expect our relevance maps to light up in roughly the same areas as the relevance maps in [5], which would indicate that our model is picking out the same image features for AD diagnosis as the model in [5]. For the relevancy breakdown per area, we expect our model to pick out the temporal lobe as one of the most relevant areas, as was the case in [5] and as supported by empirical studies [7][3].

However, we expect some divergence due to the fact that we preprocess the brain scans differently from [5]. We work with (64 x 64 x 64) 3D scans whereas [5] works with (193 x 229 x 193). The difference in representational sizing of the input image naturally has implications for the learned weights and corresponding relevance maps. We expect these differences to be especially prominent in the occlusion-based methods, where the difference in scaling across dimensions may affect the resulting occlusion ratio (the percentage of area covered by the patch).

3. Technical Approach

3.1. Baseline Classification and Visualization

The visualization techniques that we are interested in require a trained CNN model as input. We train a CNN model to perform baseline binary classification. The architecture consists of four convolutional layers, each with relu activation, batch normalization, and maxpool. We also use dropout on the middle two layers. Each layer uses the same convolutional filter size (3x3x3) but differs in the number of filters (16, 32, 64, 16.) Finally, we use three fully connected layers with a sigmoid activation to provide a single classification output. Our network uses an Adam optimizer.

Backpropagation and occlusion-based methods are two common classes of visualization techniques. We adapt the visualization methods provided by Rieke et al., which include sensitivity analysis (backpropagation,) guided backpropagation, occlusion, and brain area occlusion.

Basic backpropagation (1) calculates the gradient of the model's output probability with respect to the input brain image, with the gradient value for each pixel representing the amount the classification probability changes when the pixel value changes. Guided backpropagation (2) sets negative gradients to 0 at ReLU layers in the backward pass, whereas (1) takes the absolute value of gradients. Occlusion (3) slides a black patch across the image and recalculates the output probability for each "occluded" image, then compares the probability of the resulting output class of the occluded image to the original probability of the raw image. Image area relevance is indicated by the "occluded" probability decreasing compared to the original raw probability. Finally, brain area occlusion (4) occludes an entire area of the brain using the Automated Anatomical Labeling atlas [2].

For the basic occlusion method, we use 10x10 occlusion patches with a stride of 5.

4. Dataset

Our dataset comes from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and consists of 1334 1.5T 3D MRI images collected across patients. Images of patients with Alzheimer's disease (AD) are given positive labels, while normal control (NC) patients serve as negative examples. Preprocessing steps were performed by Zhao et al. and included "denoising, bias field correction, skull striping, affine registration to the SRI24 template (which accounts for differences in head size), and re-scaling each image to a 64 64 64 volume" [9]. Finally, we apply data augmentation via rotations and shifts such that the number of positive and negative examples is balanced.

We split the data into train, validation, and test sets on the patient-level, in order to avoid overlap between images shown to the network at train and evaluation time. Summaries of the train, validation, and test datasets are provided below.

Training Set		
	NC	AD
original	538	276
augmented	486	748
total	1024	1024

Validation Set		
	NC	AD
original	180	89
augmented	76	167
total	256	256

Test Set		
	NC	AD
original	162	89
augmented	94	167
total	256	256

Table 1: Summary of augmentation and split of project dataset.

5. Preliminary/Intermediate Results

5.1. Baseline Classification

We train our model with a binary cross-entropy with logits loss function and L2 regularization, and we use a scheduler that reduces the learning rate on plateaus based on the validation accuracy from each epoch. Using an initial learning rate of 0.0005, we achieve 77% test classification accuracy over 30 epochs of training. This is consistent with comparable architectures in literature [5] and suffice for the purposes of our project, since we are interested in visualization and not optimizing classification accuracy.

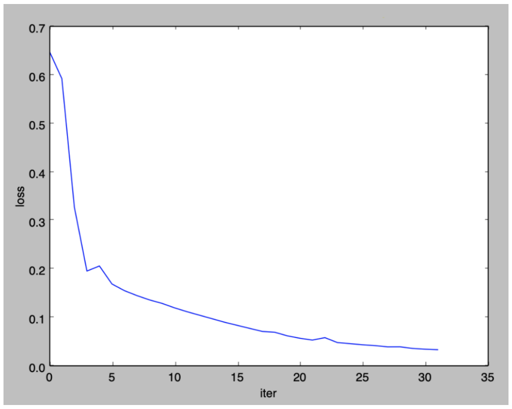

Figure 1. Training loss vs epoch for baseline classifier. Note that for each epoch, we only sample the loss of the last iteration.

5.2. Baseline Visualization

We apply the sensitivity analysis, guided backpropagation, occlusion, and brain area occlusion methods from [5] to our trained model. Red areas of the heatmaps produced by these methods indicate parts of the image that were especially relevant for the model's classification decision. The results produced by our network are shown in the figure below.

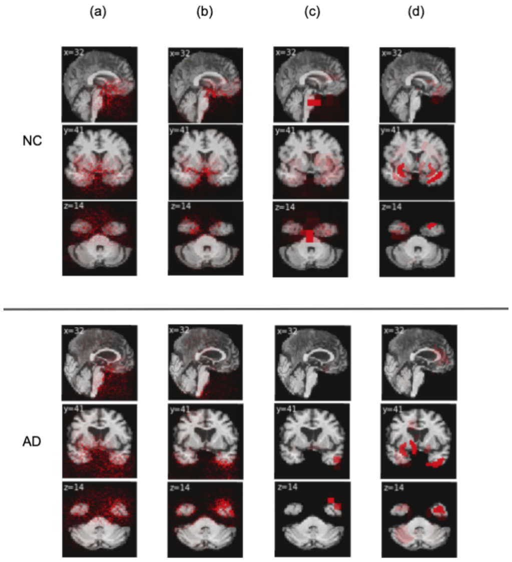

Figure 2: Sample visualization results of different slices along the x, y, and z axes, for one sample image per NC and AD conditions. (a) Sensitivity analysis (b) Guided backpropagation (c) Occlusion (d) Brain area occlusion.

In line with [5], we find that the relevance maps for AD and NC samples tend to illuminate similar areas of the brain. Intuitively, this is because regardless of the ultimate diagnosis, the model should focus on the same regions when determining the presence or absence of AD. We observe some differences between AD and NC samples for occlusion-based methods, but this instability is also in line with [5] and can be explained by the network's susceptibility to confusing the occlusion patch with brain atrophy and thereby increasing the AD-classification likelihood in those areas.

In general, we find that the relevance maps focus on the temporal lobe, in particular the middle and inferior temporal gyrus, as expected. The medial temporal lobe is affected in an early stage of AD and atrophy in this region has been proven to be a good predictor of AD in patients with mild cognitive impairment [7]. Unlike the results of [5], our model also identifies the insula as an area of interest, and it has also been shown that this region of the brain is affected since the early stages of AD [2]. The most relevant brain areas as identified by the visualization techniques on our model are shown in the table below.

	Sensitivity analysis	Guided backpropagation	Occlusion	Brain area occlusion

AD	TemporalInf (6.9%)
	FrontalInfOrb(5.6%)
	Insula(4.9%)
	TemporalMid (4.7%)
		TemporalInf (10.3%)
		TemporalPoleMid (10.1%)
		FrontalInfOrb (6.6%)
		TemporalPoleSup (6.3%)
			TemporalInf (9.0%)
			TemporalMid (8.9%)
			TemporalPoleSup (7.9%)
			TemporalPoleMid (7.4%)
				TemporalMid (25.5%)
				Insula (9.7%)
				TemporalPoleMid (8.3%)
				Caudate (6.4%)

NC	TemporalInf (5.4%) FrontalInfOrb (4.9%) FrontalInfTri (4.9%) Insula (4.9%)	Caudate (6.4%) FrontalInfOrb (5.5%) FrontalSupOrb (5.2%) FrontalInfTri (5.2%)	TemporalMid (9.8%) TemporalInf (9.1%) Fusiform (6.3%) TemporalSup (5.6%)	Insula (16.2%) TemporalMid (14.2%) TemporalSup (13.5%) TemporalPoleSup (10.1%)

Table 2: Most relevant brain areas per visualization method, for one sample image per AD and NC example.

6. Future Work

Post-milestone, we will be focusing on tuning our baseline classification model, gathering baseline visualization results averaged over the test set (instead of using a sample image), and implementing the confounder-aware visualization technique mentioned in the introduction section.

References

[1] S. Basaia. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks, 2018.

[2] A.L. Foundas. Atrophy of the hippocampus, parietal cortex, and insula in alzheimer's disease: A volumetric magnetic resonance imaging study, 1997.

[3] G. Frisoni. The clinical use of structural mri in alzheimer disease, 2010.

[4] W. Lin. Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment, 2018.

[5] J. Rieke. Visualizing convolutional networks for mri-based diagnosis of alzheimer's disease, 2018. Provided inspiration and a guide for much of our project and modeling. `1808.02874.pdf`.

[6] R. Tarawneh. The clinical problem of symptomatic alzheimer disease and mild cognitive impairment, 2012.

[7] P. Visser. Medial temporal lobe atrophy predicts alzheimer's disease in patients with minor cognitive impairment, 2002. Served as resource for background

resource on understanding brain activity with Alzheimer's. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1737837/`.

[8] Y. Zhang. Joint assessment of structural, perfusion, and diffusion mri in alzheimer's disease and frontotemporal dementia, 2011.

[9] Q. Zhao. Confounder-aware visualization of convnets, 2019. Main resource for confounder-aware visualizations and related functions. `12727.pdf`.