Hidden Cues: Deep Learning for Alzheimer's Disease Classification
CS331B project final report

Tanya Glozman
Electrical Engineering
Stanford University
tanyagl@stanford.edu

Orly Liba
Electrical Engineering
Stanford University
orlyliba@stanford.edu

Abstract

Alzheimers disease is the most common form of dementia in adults aged 65 or older. While many neuro-imaging based biomarkers have been proposed over the years for detection of AD, these have mostly been hand-crafted and utilized domain-specific clinical knowledge. In this project, we are interested in automatically discovering such biomarkers (hidden cues) by using deep learning methods to classify Alzheimer's patients and normal controls.

1. Introduction

Alzheimer's disease (AD) is a progressive, degenerative brain disorder that leads to nerve cell death and tissue loss in the brain. It is the most common form of dementia in adults aged 65 and older. The worldwide prevalence of AD was reported 26.6 million in 2006 and is expected to rise to over 100 million by 2050 [3]. Another form of milder dementia is Mild Cognitive Impairment (MCI) - it is a prodromal stage of Alzheimer's disease. MCI patients may or may not convert to Alzheimer's disease, and it is not yet known why some MCI subjects convert to AD while the symptoms for others remain stable and do not deteriorate over time. Current diagnosis of MCI or AD relies mostly on cognitive testing and documenting mental decline by a set of standardized tests as well as clinician's subjective evaluation. Research efforts are focused on discovering an accurate and objective way of identifying the disease. Neuroimaging is often part of the standard diagnosis routine, used primarily to rule out other conditions that may cause similar symptoms of mental decline (e.g. stroke, head injury, etc.). Three neuroimaging data types are in the focus of this paper: (1) structural-MRI (sMRI), which provides good contrast for gray matter and subcortical brain structures, providing information on the structural integrity of the brain tissue; (2) Positron Emitting Tomography (PET) with FDG (a radiolabeled glucose analogue) , which measures glucose metabolism in the brain; (3) PET with AV45 radioactive tracer, which binds to the amyloid plaques in the brain, thus highlighting areas with high amyloid burden (one of the hallmarks of AD). PET imaging provides insight into the functional integrity of the brain tissues, emphasizing the areas with increased/reduced activity or amyloid deposition in the brain. Figure 1 shows a comparison of coronal views between an Alzheimer's patient (left on all panels), MCI patient (middle on all panels) and Normal Control subject (right on all panels. Top panel shows a structural MRI scan, middle panel shows PET-FDG scan and bottom panel shows PET-AV45 scan. The subjects we chose to show here are representative of the pathological changes occurring in the brain during AD, MCI and normal aging. In a clinical setting, a trained neuroradiologist will assess the 'shrinkage' of a the brain tissue on the sMRI data and determine whether it corresponds to normal aging or tissue degeneration as a result of dementia. On the PET-FDG images, a neuroradiologist will examine the intensity of the pixels in the different areas in the brain - lower pixel intensity corresponds to lower glucose consumption by the brain, indicating a lower activity. Similarly, on PET-AV45 images, a neuroradiologist is examining the relative pixel intensity in different brain structures- a proxy of the amount of amyloid plaques in the brain. It is important to note that there is a large inter- and intra- subject variability in neuroimaging data. Normal subjects may have larger structural atrophy in the brain, which the brain of some AD patients may appear normal. The subjects shown in figure 1 were chosen specifically to highlight the 'best-case' expected changes. Though subtle, these changes provide important diagnostic information to clinicians. For many subjects in this dataset, the changes are not so pronounced. In this project, we are interested in answering the following questions: (1) whether a deep neural network can be trained to perform this difficult diagnostic task of classification between three classes of subjects: Normal Controls (NC), Mild Cognitive Impairment (MCI) patients, and Alzheimer's Disease (AD) patients; and (2) which of these neuroimaging data sources contains more diagnostic information;

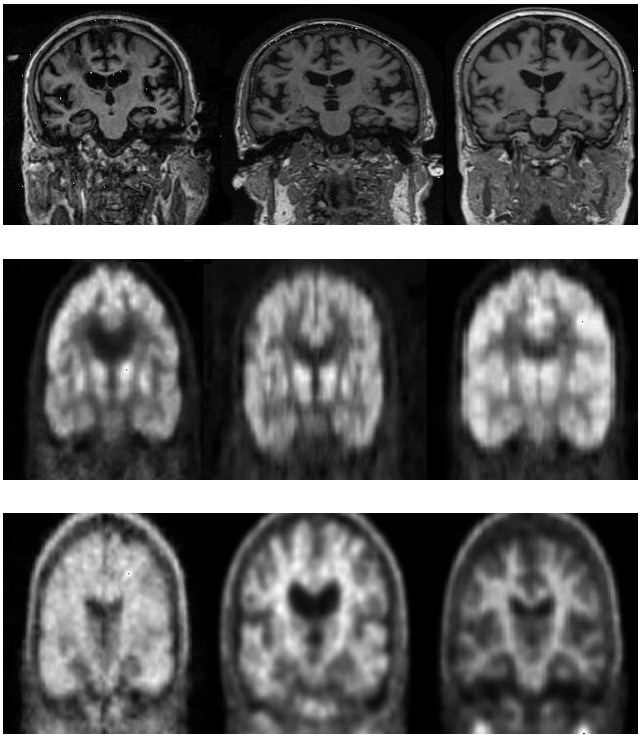

Figure 1. sMRI (top), PET-FDG (middle) and PET-AV45 (bottom) examples of three groups of subjects: AD (left on all panels), MCI (middle on all panels), NC (right on all panels). Top panel shows the structural changes occurring in the brain - tissue shrinkage. Middle panel shows the decreased metabolic activity in AD and MCI patient relative to the NC subject. Lower panel shows the increased amount of amyloid plaques detected in AD and MCI patients relative to the NC subject.

2. Previous Work

Medical image analysis has benefited from the development of deep neural networks, which are used for various tasks of classification and segmentation. Training a deep convolutional neural network from scratch is usually challenging owing to the limited amount of labeled medical data. A promising alternative is to fine-tune the weights of a network that was trained using a large set of labeled natural images. However, many times the appearance of the images and the classification tasks of medical images differ greatly from commonly used benchmarks (such as ImageNet [4]). The use of pre-trained networks versus full training for medical images has been explored in [18]. This work considered four distinct medical imaging applications and investigated how the performance of deep CNNs (Convolutional Neural Networks) trained from scratch compared with the pre-trained CNNs fine-tuned in a layer-wise manner. Their experiments demonstrated that the use of a pre-trained CNN with adequate fine-tuning performed as well as a CNN trained from scratch and were more robust to the size of training sets.

Another recent work [15] explored the utilization of deep CNNs to problems of computer aided detection (CADe) in the medical realm. The authors compared the performance of CifarNet[12], AlexNet [13], and GoogLeNet [17] with different model training paradigms on the problems of detection or classification of lymph nodules and several types of lung diseases from CT images. They were able to augment their dataset significantly since they were using patches for training rather than full images. They concluded that transfer learning performed significantly better than training from scratch (similarly to [18]) and that in most tasks GooglLeNet architecture proved superior, since a more complex network is able to better learn hidden structure from data.

There have also been several works using deep networks for Alzheimer's classification. In [16] and [7] the authors used stacked AutoEncoder for AD/MCI diagnosis. They are motivated by a belief that latent high-level information may exist in the low-level features, which may benefit the diagnostic model. [16] used both the imaging data (MRI and PET) as well as other types of data (different types of cognitive tests, CSF biomarkers) to train the autoencoder. They then concatenate the original low-level features with the SAE-learned latent feature representation, do feature selection and perform a multi-kernel SVM for diagnosis. This work is the current state-of-the-art for classification between AD, NC, MCI-converters and MCI-non-converters (see section 3.1 for further details). [7] aims to extract features related to AD-related variations of anatomical brain structures, such as, ventricles size, hippocampus shape, cortical thickness, and brain volume, using a three-dimensional convolutional autoencoder (figure 2). The autoencoder is pre-trained to capture anatomical shape variations in structural brain MRI scans. The encoder is fed into fully connected layers which are then trained for each task-specific AD classification task. Their experiments on the ADNI dataset have shown better accuracy compared to several conventional classifiers.

3. Methods

3.1. Data

The data for this project was obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI is a global effort with the primary goal of testing whether neuroimaging data, biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). In this work, we are interested in the following three types of neuroimaging data: (1) structural-MRI (sMRI), which provides good contrast for gray matter and subcortical brain structures, providing information on the

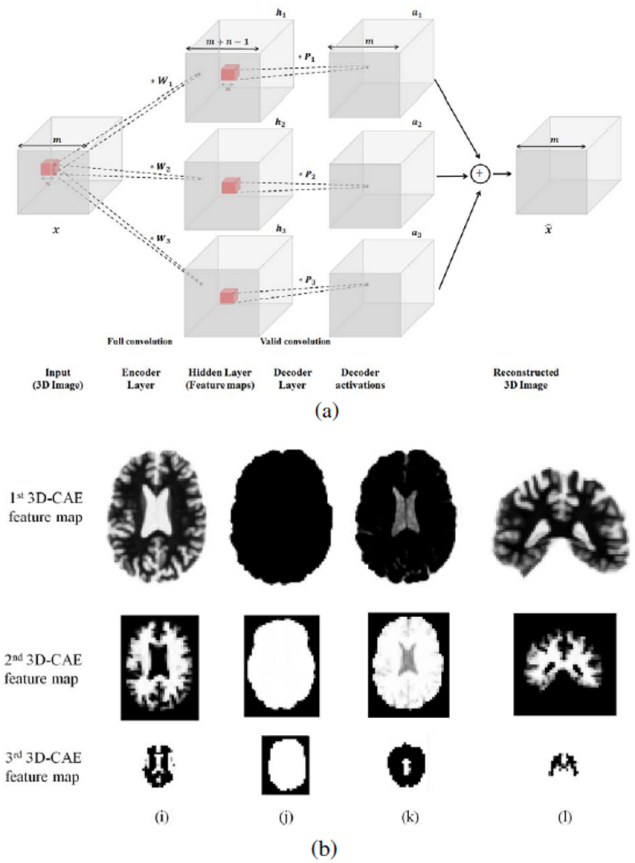

Figure 2. **3D autoencoder concept for classifying MRI brain volumes, from [7]. This image shows the autoencoder and the learned features. The feature maps are later used for classification.** (a) Schematic diagram of a convolutional autoencoder which works on three-dimensional volumes. (b) Selected feature maps extracted at 3 layers of the stacked autoencoder. The feature maps show (from left to right) indications of cortext volume, brain size, ventricle size and a model of the hipocampus.

structural integrity of the tissues; (2) and Positron Emitting Tomography (PET) with FDG, which provides contrast for glucose metabolism in the brain (tissue degeneration causes a decrease in this signal); (3) PET with AV45, which provides contrast for Amyloid deposits in brain tissue. Details about the acquisition protocol and the initial processing steps can be found in [8]. We downloaded all data currently available on ADNI database for PET and sMRI scans for AD, MCI and NC subjects. As ADNI is a longitudinal study (i.e. each subject undergoes a series of clinical, cognitive and imaging tests every 6 months), the diagnosis for some subjects changes with time. If the diagnosis changes from NC to MCI/AD or from MCI to AD (i.e. worse symptoms over time), the diagnosis change is considered 'conversion'. If the change is in the opposite direction (symptoms improve over time), the diagnosis change is considered 'reversion'. In this paper we decided to exclude the

Modality	AD	MCI	NC
PET-FDG	241 (482)	130 (554)	339 (895)
PET-AV45	140 (169)	454 (540)	229 (952)
sMRI	200 (864)	132 (651)	221 (1417)

Table 1. **Number of subjects per class.** Number in parentheses indicates the total number of scans (including all longitudinal scans)

'converted'/'reverted' subjects from our study data, and focus on subjects whose diagnosis is stable over time. For most subjects data at several time-points is available - some subjects have been tested every 6 months for up to 3 years. Table 1 shows the total number of subjects in each class. The number in parentheses indicates the total number of scans (including all longitudinal scans available).

3.2. Data Preprocessing

The raw data is a volume scan in NIfTI format [1]. We used the Nibabel software package [6] to read the raw data. Files that were corrupted in some way and were not readable by NiBabel software were excluded from the study. Following this, the data for each scan is a three-dimensional volume of size KxLxM, where K,L,M vary between different imaging centers (depending on the resolution of the MRI/PET scanner) and modalities. Due to the modest size of our dataset, since training the network from scratch would probably have led to overfitting, we experimented with utilizing a neural network pre-trained on natural images (AlexNet trained on ImageNet) and finetuning the last few layers. As the expected input is 2D images, we create the following images from the volume data:

1. *ax_key* - axial middle slice in all three color channels. An example resulting image is shown in figure 3(top-left). This orientation represents a horizontal cut through the brain.
2. *cor_key* - coronal middle slice in all three color channels. An example resulting image is shown in figure 3(top-middle). This orientation represents a vertical cut through the brain from left to right.
3. *sag_key* - saggital middle slice in all three color channels. An example resulting image is shown in figure 3(top-right). This orientation represents a vertical cut through the brain from front to back.
4. *ax3* - mid-axial slice in the R channel, and 10 slices above and below the mid-axial slice in the G and B channels.An example resulting image is shown in figure 3(2nd row -left).
5. *cor3* - coronal middle slice in the R channel, and 10 slices before and behind the mid-coronal slice in the G and B channels.An example resulting image is shown in figure 3(2nd row-right).
6. *sag3* - mid-saggital slice in the R channel, and 10 slices

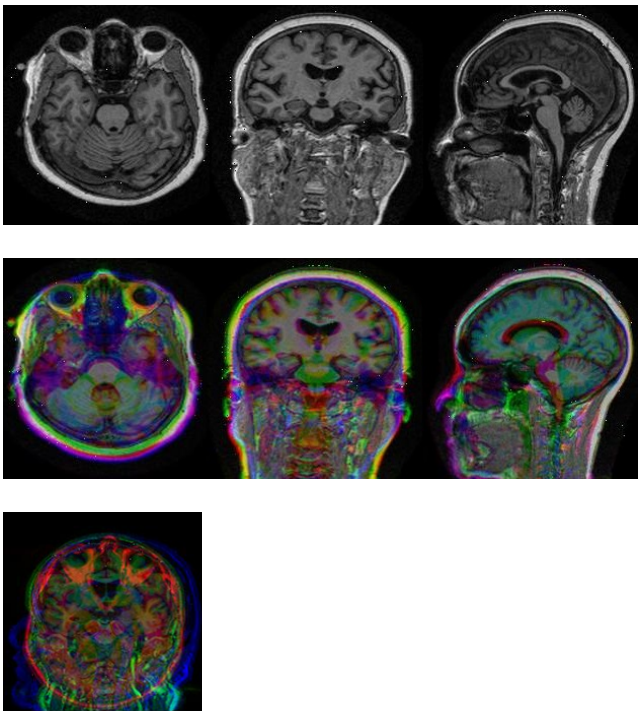

Figure 3. **Example of different images created from sMRI data of a single subject** Top-left: axial view (ax_key image), top-middle: coronal view (cor_key image), top-right: saggital view (sag_key image). Bottom-left: ax3 image, bottom-middle: cor3 image, bottom-right: sag3 image

before and behind the mid-saggital slice in the G and B channels.An example resulting image is shown in figure 3(2nd row-middle).

7. *axcosag* - mid-axial slice in the R channel, mid-coronal slice in the G channel and mid-saggital slice in the B channel. An example resulting image is shown in figure 3(bottom row).

These images were normalized to the range of [-128,128], and resized to size 227x227 (compatible with AlexNet expected input size and range) using 2nd order spline interpolation.

3.2.1 Data augmentation

To increase the amount of data available to the network, we added mirror-transformed images to the training set. As we are dealing with brain images, we believe that other methods of data augmentation (i.e. random crops, scales and color jittering) do not preserve the diagnostic value of the image.
The data was randomly divided into training and testing sets on the subject level (rather than on the scan level, to avoid possible bias of having data of the same subject at different time points in both the training and testing sets).

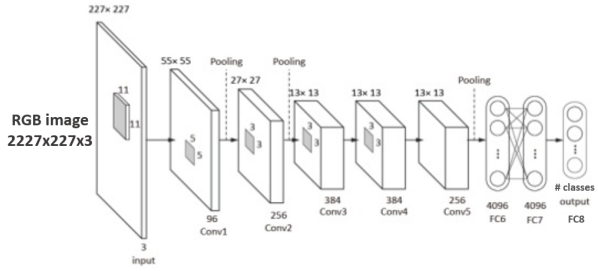

Figure 4. **AlexNet architecture.** Includes 5 convolutional layers and 3 fully-connected layers.

3.3. Network Architecture and Training

Due to the limited amount of training data in ADNI, we chose to work with pre-trained network. We fine-tuned AlexNet [14] (figure 4) which was pre-trained on the ImageNet benchmark [4]. Fine-tuning was performed on images of the 3 modalities (sMRI, PET-FDG and PET AV45) and 7 image types, as described in figure 3. The depth of fine-tuning (meaning, which layers are tuned) and the learning rate have a large impact on the accuracy results. We tested the accuracy of the fine tuned network which was trained with learning rates of 0.001, 0.005, 0.01 and 0.015. These learning rates were chosen based on the work of Tajbakhsh et al [18] which systematically explores fine-tuning of neural networks for medical images. The same learning rate was used for all of the fine-tuned layers, which are fc8, fc7-8, fc6-8 and conv5+fc6-8. All layers were initialized with ImageNet weights and biases, except for fc8, which was initialized with random values from a Gaussian distribution. Training was done with dropout of 0.1 (keep rate of 0.9), mini-batch size of 30 images, and Adams-optimizer [11]. We implemented the training and testing on TensorFlow [2].

Owing to the large impact of the learning rate and depth of fine-tuning, we explored a 2-step training scheme. In this approach, we first perform coarse-tuning for fc8, with a large learning rate, and then fine-tune all the layers down to conv4 with a small learning rate.

In order to combine information from different modalities (PET and sMRI) or from different view angles (sagital/coronal/axial), we implemented three-stream and two-stream networks, which combine three or two fine-tuned AlexNets at the final classifier layer (figure 5). The AlexNets up to the final classifier (fc8) were fine-tuned separately and the final classifier was trained from random weights on a combination of images from different views or different modalities.

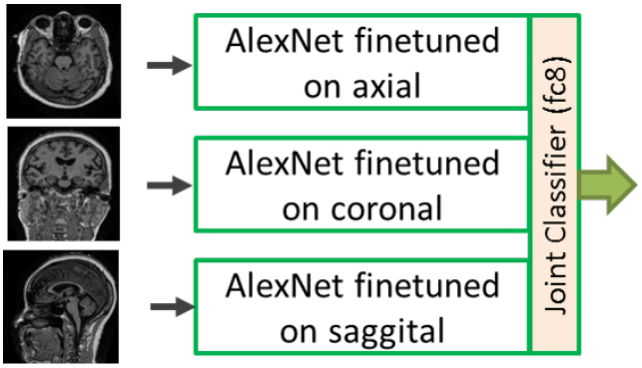

Figure 5. **Training and classification based on different view angles.** In order to take into consideration the volumetric data, we can fine tune the network of each view separately and then combine the fine-tuned network into a 3-stream network and train the final layer on the 3 views simultaneously. In the classification stage we input the three views to obtain the classification result. This was done on key-frames and on RGB images of 3 frames in the same view.

4. Experiments

4.1. Accuracy Measurements

We measured the classification accuracy from the following experiments:

4.1.1 Single-modality single-view classification.

In these experiments we were interested in understanding which of the three modalities and which of the three possible orientations (axial/coronal/saggital - referred to as ax_key,cor_key,sag_key in this paper) and the three aggregated orientations (referred in this paper as ax3, cor3, sag3) contain the most diagnostic information, thus achieving higher accuracy. We varied the number of layers that were fine-tuned and experimented with different learning rates, as previously described. The results of these experiments, with the hyperparameters that yielded the highest accuracy, are summarized in table 2.

4.1.2 Single-modality single-view 2-step fine-tuning.

In these experiments we set first tuned fc8, which was initialized with random weights and biases, with a high learning rate (α) of 0.015 and then fine tuned the network's deeper layers, from conv4 to fc8 with smaller learning rates, in a range of 0.0001 to 0.002. We found that the second step often improved upon the accuracy of the first step by a few percents, however, one-step multi-layer tuning with a medium learning rate achieved better results. The results of some of these experiments are shown in table 3.

4.1.3 Two- or Three-stream single-modality network.

In these experiments we train two or three separate networks to classify images of two or three different orientations - axial, coronal and saggital correspondingly. These networks are combined by concatenating their fine-tuned fc7 layers and then training the fc8 layer of the combined network (which now has 2 or 3 times more weights) on pairs or 3-tuples of images with corresponding views from volumes in the training set. Figure 5 shows a schematic of a three-stream network trained on key slices from three views. For testing, we used pairs or 3-tuples from volumes in the test-set.
Two types of images are considered in these experiments: (1) keyslices from each orientation - ax_key, cor_key, and sag_key; (2) aggregated slices from each orientation - ax3, cor3, sag3. We consider these representations as two different types of approximations to the 3D volumetric input. Results from these experiments are summarized in table 4.

4.1.4 Two-way classification.

We compare the performance of the network on two-way classification (AD vs NC) to three-way classification (AD vs MCI vs NC) in a single-view single-modality network. Table 5 shows the comparison. The results improved significantly - almost 23% for PET-AV45!). It is clear that including the MCI class in classification confuses the network, leading to the reduced performance, since the MCI images are very similar to both NC and AD classes.

4.2. Learning Hidden Cues (Visualizing the Learned Representation)

We now proceed to identify which image features contribute the most to the classification of AD/MCI/NC. We have considered the following visualization strategies:

4.2.1 t-SNE

t-SNE [20] is a dimensionality reduction method that allows visualization of high-dimensional datasets. In this work we used the Barnes-Hut approximation implementation of the technique [19]. We compute a two-dimensional embedding for the fc8 CNN features. Following this computation, the images with similar fc8 value will be nearby in the resulting embedding image. Figure 6 shows the 2D embedding of the fc8 features of a network trained on PET-FDG data. Although there is no clear separation between the classes, a trend can clearly be seen where the images belonging to NC class are in the top left corner whereas the MCI and AD images are clustered more towards bottom right.

	ax-key	cor-key	sag-key	ax3	cor3	sag3	axcosag
sMRI	48.3	48.7	47.9	45.9	45.7	49.1	48.5
PET-FDG	57.5	64.3	61.0	67.4	64.7	56.1	70.8
PET-AV45	57.5	58.7	60.6	57.2	56.6	60.3	57.2

Table 2. **Classification accuracy percentages for the different experiments.** Chance is 33%

α: fc8	0.015	0.015
α: conv4-fc8	0.0005	0.002
Accuracy, Step 1	50	51
Accuracy, Step 2	55.25	54.25

Table 3. **Accuracy of networks with 2-step fine-tuning, PET-FDG ax_key.**

	ax-cor(-sag)	ax3-cor3(-sag3)
sMRI	48.8	45.6
PET-FDG	57.4	67

Table 4. **Two/Three stream single-modality network accuracy percentages.** Three-way classification - Chance is 33%

	AD/NC	AD/MCI/NC
sMRI	66.51	48.79
PET-FDG	81.09	70.83
PET-AV45	83.57	60.66

Table 5. **Two-way vs. three-way classification accuracy percentages.** Including the MCI class reduces accuracy significantly.

4.2.2 Visualizing the activation of layers in the network.

One can visualize the activation of layers in the network as a result of an input image, however, the middle layers are usually difficult to interpret. The simplest activation to understand is the output of the last layer (fc8). The probabilities for classifying each image in a certain class can be estimated by applying the *softmax* function on the result of the last layer. Looking at the images with the highest probability in each class can help us understand what the network relies on when making its decision. In figure 7 we show the images with the highest probability in each class in PET-FDG key-slice images. From these images we can see that the overall signal in AD is lower in the brain compared to the skull and that the AD brains have the most shrinkage (most gap between skull and brain tissue). These features are expected in Alzheimers disease, due to the decreased metabolism and size of the brain.

Figure 6. **t-SNE embedding of fc8 features on PET-FDG data** The box surrounding each image corresponds to the true label of the image: red for AD, blue for MCI and green for NC.

5. Conclusions

Deep learning methods have become ubiquitous in computer vision applications. In medical imaging, due to a much smaller amount of labeled data, these techniques face many challenges. In this work, we explored whether a network pre-trained on natural images can be fine-tuned to classify neuroimaging data in which the difference between the different classes are very subtle, even for the human eye. Our results suggest that with the available data, the network can learn to classify the two extreme classes (NC vs AD), but when faced with a three-way classification task, it will not achieve good accuracy. The reason for this is not only the limited amount of data, but also the ambiguity in it - MCI images look very similar to both classes. Furthermore, neuroimaging data differs significantly from natural

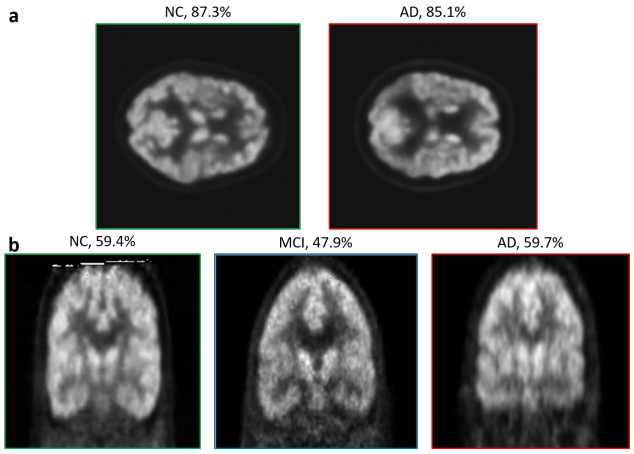

Figure 7. **Highest ranked images in each class.** PET-FDG images with the highest probability in each class in 2-way axial view (a) and 3-way coronal view (b) classification. The images show the expected shrinkage and lower signal in the AD brains, resulting from lower activity and metabolism due to the disease.

images on which our network was pretrained. It is conceivable, thus, that given much more data, training the network from scratch may increase its performance. Fine-tuning a network trained on ImageNet data is also problematic since in ImageNet the locations of the detected objects are not significant, whereas brain changes are localized in specific anatomical structures.

Classification accuracy on the functional PET images (PET-AV45 and PET-FDG), was significantly higher than on the structural MRI images. This is likely because functional changes in the brain, such as lower metabolism and higher amyloid plaques, are more pronounced than structural changes.

There was no significant difference between the classification accuracy results on the different views - this is because in AD tissue degeneration and functional changes affect the whole brain.

The visualization strategies we have used in this project, t-SNE and highest probability images, revealed minor differences between the classes which were along the expected functional changes from NC to AD brains.

6. Future Work

1. **Improve image pre-processing**. In this work we used the images as downloaded from the ADNI dataset. All volumes are in the same orientation, however, the sMRI volumes were not registered to a template in the pre-processing steps. A common practice when dealing with neuroimaging data is to register all data using a linear or non-linear transformation to some template (an "average" brain), such that all the images are aligned. One of the most popular tools for such registration is the FSL software suite [10, 9]. The steps for alignment included skull stripping, calculation of linear affine transformation between the skull-stripped volume and the MNI152 standard brain atlas (which was created by averaging 152 normal subjects), and applying the affine transformation on the original (non-skull-stripped) volume. The processing time for each scan is of the orders of 25 minutes. Due to limited time and computation resources to which we had access, in this project, we used unregistered data for sMRI. It is reasonable to assume that the results of sMRI data would improve if registered data is used. In the future, we are planning to run the registration on all sMRI data and repeat the experiments for the registered images. All PET data was pre-registered to a common template by the dataset acquisition team, thus there re-processing won't be necessary.

2. **Better representation of the volumetric data**. The neuroimaging data is volumetric, thus it may be beneficial to use all slices for classification. This may be achieved in one of the following two approaches: (1) Using a pre-trained fine-tuned 2D-CNN for each slice and concatenating fc7 outputs for a joint classification; (2) using RNNs to encode the volume, sequentially feeding each slice to the network. (3) Use a 3D-CNN, when a pre-trained network becomes available, or if a larger training database becomes available.

3. **Processing volumes using an autoencoder**. Considering the good results obtained by [7], it would be very interesting to train a classifier on features obtained with a volumetric autoencoder (an autoencoder that receives the entire 3D volume ad an input, rater than key slices). In addition to the benefit of using the entire volume for the classification, by using an autoencoder we can learn features from a larger dataset: the Human Connectome Project (HCP) [5], which includes MRI scans of 900 healthy individuals. This approach could help learn general features from the larger dataset which could be used for classification of AD in the ADNI dataset. In addition to improved classification, the autoencoder approach can also help us visualize the features of the hidden layers. This approach could provide a meaningful representation and reveal which features are characteristic of AD.

4. **Modality fusion**. In the ADNI dataset, the number of subjects for which scans were available from three, and even two, imaging modalities was very limited and insufficient for training a modality-fusion model. It would be interesting to develop an approach to fuse these modalities using a limited dataset (perhaps by using autoencoders and using the hidden layer activations) or use a conventional method (such as 3-stream)

if more data that combines the three modalities becomes available.

5. **Visualization using occlusion sensitivity** Following the publication by Zeiler and Fergus [21], we implemented occlusion sensitivity by hiding parts of the image and calculating the probability for correct classification. However, since the images of brains have a very defined structure, and features that are not necessarily local, the result of this approach did not produce meaningful results. A possibly better approach for brain images would be to change the intensity of the images or perform dilation and erosion, which artificially change the structure or functional-image of the brain. This perhaps would help show boundaries for brain size or activity which the network interprets as healthy versus diseased.

7. Acknowledgments

The authors would like to thank the course staff, Dr. Silvio Savarese, Dr. Amir Zamir and Kenji Hata for their valuable ideas and suggestions.

References

[1] Neuroimaging informatics technology initiative - national health institute. https://nifti.nimh.nih.gov/.

[2] M. Abadi, A. Agarwal, P. Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of alzheimers disease. *Alzheimer's amd Dementia*, 3(3):186 – 191, 2007.

[4] J. Deng, W. Dong, R. Socher, et al. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] D. C. V. Essen, S. M. Smith, D. M. Barch, et al. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013.

[6] K. Gorgolewski, C. D. Burns, C. Madison, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform*, 5, 08 2011.

[7] E. Hosseini-Asl, R. Keynton, and A. El-Baz. Alzheimer's disease diagnostics by adaptation of 3d convolutional network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 126–130, Sept 2016.

[8] C. R. Jack, M. A. Bernstein, N. C. Fox, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[9] M. Jenkinson, B. PR, B. JM, and S. M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 2(17):825–841, 2002.

[10] M. Jenkinson and S. M. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[12] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[15] H. Shin, H. Roth, M. Gao, et al. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning,. *IEEE transactions on medical imaging (TMI)*, 35(5):1285–1298, 2016.

[16] H.-I. Suk, S.-W. Lee, and D. Shen. Latent feature representation with stacked auto-encoder for ad/mci diagnosis. *Brain Structure and Function*, 220(2):841–859, 2015.

[17] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, May 2016.

[19] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15:1–21, 2014.

[20] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[21] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.