Class Activation Attention Transfer Neural Networks for MCI Conversion Prediction

Min Luo[a], Zhen He[a,*], Hui Cui[a], Phillip Ward[b,c,d], Yi-Ping Phoebe Chen[a], the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *Department of Computer Science and Information Technology, La Trobe University, Melbourne Vic, 3086 Australia*
[b] *Monash Biomedical Imaging, Melbourne, Vic, 3800 Australia*
[c] *Turner Institute for Brain and Mental Health, Monash University, Melbourne, Vic, 3800 Australia*
[d] *Australian Research Council Centre of Excellence for Integrative Brain Function, Melbourne 3800, Australia*

Abstract

Accurate prediction of the trajectory of Alzheimer's disease (AD) from an early stage is of substantial value for treatment and planning to delay the onset of AD. We propose a novel attention transfer method to train a 3D convolutional neural network to predict which patients with mild cognitive impairment (MCI) will progress to AD within 3 years. A model is first trained on a separate but related source task to automatically learn regions of interest (ROI) from a given image. Next we train a model to simultaneously classify pMCI and sMCI (our target task) and the ROIs learnt from the source task. The predicted ROIs are then used to focus the model's attention on certain areas of the brain when classifying pMCI versus sMCI. Thus, in contrast to traditional transfer learning, we transfer attention maps instead of transferring model weights from a source task to the target classification task. Our Method outperformed all methods tested including traditional transfer learning and methods that used expert knowledge to define ROI. Furthermore, the attention map transferred from the source task highlights known Alzheimer's pathology.

Keywords: Alzheimer's disease, Mild Cognitive impairment, Prediction, Class activation maps, Convolutional neural networks, Attention mechanism

[*] Corresponding author
Email address: z.he@latrobe.edu.au (Zhen He)

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disease in the elderly [1]. It is characterized by the progressive decline of memory functions and significant difficulties with retaining independence in simple daily activities [2], [3]. In this paper we focus our research on Mild Cognitive Impairment (MCI). MCI is known as an intermediate stage for individuals between the normal cognitive change of aging and early dementia. It is reported that 12% to 15% of patients who have MCI will progress to AD annually[4]. However, AD is very challenging to diagnose as the symptoms can be similar to other diseases and the cause of AD is not well understood [3], [5]. Unfortunately, AD is not curable and the decline of cognitive impairment is irreversible [6].

Accurately predicting whether an MCI patient will convert to AD is of significant importance. This information is critical for clinical trials, decisions for early interventions, and to maximise the chances of delaying onset. It also gives patients and their families time to draw a plan in advance for the management of treatment, care, and cost.

Magnetic Resonance Imaging (MRI) is a valuable and complementary tool for assessing and monitoring brain changes such as volume and tissue characteristics. MRI imaging can help to detect brain abnormality during the conversion to AD from MCI [7]. For example, in the early stage of AD, the brain areas associated with MCI may look normal [8].

In this paper, we are focused on predicting progressive MCI (pMCI) versus stable MCI (sMCI) trajectories from MRI images. pMCI (sMCI) is defined as (not) being diagnosed with AD following a previous MCI diagnosis. Specifically, our goal is to take a single MRI image of a patient diagnosed with MCI at a given time and accurately predict whether they will be diagnosed with AD within 3 years. This is a very challenging task since the brain may undergo a lot of change within the 3 year period. So taking an MRI image at the baseline time to predict what will happen in 3 years time is very difficult.

We use convolutional neural networks (CNN) to solve this problem in a data efficient way. We use the Alzheimer's disease neuroimaging initiative (ADNI) dataset 1 and 2 where there are a total of 1587 subjects but only 593 subjects are classified as MCI at the baseline. Therefore only images from the 593 subjects are directly applicable to our task. An interesting research question is how can we best use the images from the entire 1587 subjects? A traditional method for achieving this is to use transfer learning [9], [10], where model weights learnt from a source classification task (e.g AD/CN , high/low ADAS-cog score, high/low CDR-SB score) are transferred to the target classification task (pMCI/sMCI). We propose a novel alternative method where an **attention** map from the source task is transferred to the target task instead of model **weights**. This mimics how a radiologist would transfer their knowledge of the important regions of interest (ROI) learnt from previous tasks to a new task. Existing ROI based pMCI versus sMCI classification approaches [11], [9] directly identify ROI from prior expert knowledge. In contrast, our method automatically learns the ROI via attention maps derived from the source task. Furthermore we found our way of learning ROI from the source task outperforms methods that assign ROI based on prior expert knowledge. This may be attributable to the fact that the attention maps generated by our model are tailored to each image rather than the same ROI assigned to all images as is the case for traditional ROI based solution that use expert knowledge.

We propose a novel method called Class Activation Attention Transfer (CAAT) to solve the progressive MCI (pMCI) versus stable MCI (sMCI) problem using only baseline MRI images. CAAT classifies between pMCI and sMCI by transferring attention from a related source classification task to our target classification task. It learns the discriminative brain areas created from a source task via the output of class activation maps (CAMs) [12] without using prior expert knowledge to determine the ROIs. The CAMs identify parts of the brain that were salient for a related task, such as descriminating AD from CN and predicting cognitive performance, and uses this information to inform our model of which brain regions to pay particular attention to. We then train a 3D CNN model to simultaneously predict the source CAM for the target task images and use the predicted CAM as an attention map for solving the target classification task of pMCI and sMCI. Visualizations of the attention maps predicted by our CAAT approach show that the model is able to place attention on parts of the brain that are known to be important for diagnosing Alzheimer's disease. The highlighted areas are also coherent with cognitive test scores.

Experimental results on the ADNI dataset show that CAAT achieves state-of-the-art accuracy of 74.61 for classifying between pMCI and sMCI using only whole 3D images of the brain and no other ancillary information. Traditional transfer learning performs worse than CAAT by achieving 73.03 classification accuracy. Finally, a baseline method [13] that only uses whole 3D brain scans without using transfer learning or attention only achieved an accuracy of 70.84. Furthermore, compared to the other methods, our CAAT ensemble method achieves more balanced results of F1 score, the sensitivity, and specificity of 0.75, 0.75, and 0.75 respectively. Our innovations and major contributions include:

1. We developed a novel method called CAAT for transferring attention information from a source task to a target task that provides an alternative to traditional transfer learning. This general methodology can be applied to any existing task where the source and target tasks share similar regions of interest.
2. We applied CAAT to the problem of pMCI versus sMCI classification using the three different source classification tasks of CN versus AD, high versus low ADAS-cog score, high versus low CDR-SB score.
3. Experimental results for the ADNI dataset show CAAT achieves state-of-the-art performance for pMCSI versus sMCI classification. Even outperforming ROI methods which require prior human expert knowledge to identify areas of interest.

2. Related Works

As mentioned in the introduction this paper is focused on solving the problem of sMCI versus pMCI classification. The most common methods for solving this problem use biomarkers in combination with machine-learning [14], [10]. Mathotaarachchi et al. [14] employed a voxel-wise logistic regression method to extract the most discriminative features (dimensionality reduction) from amyloid PET images and matching T1-weighted MRI imaging. They also used demographic and APOE4 genotype data. Finally, MMSE scores and CDR values were also used. These features were fed into a random forest classifier. In the works of B. Cheng et al. [10], each subject image had 93 manually-labeled regions-of-interest (ROIs) (a 93- dimensional feature vector) based on the GM tissue volume. These features were concatenated with the baseline MRI and cerebrospinal fluid (CSF) data. First, they were trained via SVMs to get a list of source domain labels (AD vs. CN, MCI vs. CN, AD vs. MCI, and pMCI vs. sMCI). Secondly, they combined these labels and created a multi-source domain feature matrix. The similarity was measured between the residual vectors to get an estimated domain label. Finally, after using dimensionality reduction on the selected features, they fed the most informative features to an SVM for classification. This method required prior knowledge about brain structure as it needs to define ROIs as the first step. Our proposed

CAAT method automatically learns the important ROI from the source task.

There are many methods that have used CNNs to help solve the sMCI versus pMCI classification problem. Liu and Cheng et al. [15] proposed a 3D patch-level CNN model. They used a 3D CNN model to extract features from MRI and PET images and then concatenated the features to feed into 2D CNN layers for classification. Lin et al. [9] designed an ROI-based approach that first used 2.5D patch-based CNNs to extract features while performing AD and CN classification. They then used the pre-trained AD/CN feature extractor to extract features for pMCI/sMCI classification. After that, a 2.5D image was created from transverse, coronal, and sagittal plane centered at the same point. These features were combined with the features obtained from FreeSurfer. Both feature vectors used PCA for dimensionality reduction and then were concatenated into one feature vector. Finally, the feature representations were fed into an extreme learning machine (ELM) to perform the classification. In contrast to [15] we only use the MRI images and do not use PET images. In contrast to [9] we transfer attention from the source to the target task instead of the weights of the neural network.

Basaia et al. [16] used data augmentation techniques like flips, rotations, cropping to increase training set size and trained the data using a VGG-like network. Lian et al. [17] first divided the 3D brain images into patches and then used location proposals to select the patches at the most discriminative locations. The final classification result was obtained by feeding those ROI-based patches to a patch-level subnetwork. In contrast to [16] our method transfers attention from a source task instead of performing data augmentation to improve performance. In contrast to [17] our method uses CAM heatmaps from the source task to determine where to focus attention instead of ROI-based patch selection.

3. Materials and Methods

In this section, we first introduce how we set up the experimental datasets. We show and explain the predicted heatmaps (CAM images) generated from the different pretraining datasets such as AD versus CN, high versus low ADAS-cog, and high versus low CDR-SB. We then describe in detail our class activation attention transfer method.

Table 1: The Demographic and clinical characteristics of the subjects included in this study. SD: Standard Deviation.

	sMCI(298)	pMCI(295)
Female/male	123/175	119/176
Age (SD)	72.3 (7.4) [55-88.4]	73.78 (6.9) [55.1-88.3]
MMSE (SD)	28.0 (1.7) [23-30]	26.8 (1.8) [19-30]
ADAS11 (SD)	8.5 (3.5) [2-21.3]	13.0 (4.5) [0-27.67]
CDRSB (SD)	1.2 (0.6) [0.5-3.5]	2.0 (1.0) [0.5-5]

3.1. Subjects and data acquisition

This paper uses the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [2]. The primary goal of ADNI is to detect AD at the earliest possible stage and track the data trajectory of AD via studying patients' clinical, imaging, genetic, and biochemical biomarkers. To evaluate the performance, we performed 5-fold cross-validation of the dataset.

In this study, the subjects used were categorized into two groups: progressive MCI (pMCI) and stable MCI (sMCI), based on the diagnosis of their follow-up visits within 36 months. At the start (baseline time), all selected subjects were diagnosed with MCI, early MCI (EMCI), or late MCI (LMCI). However, if a subject was diagnosed with Dementia within the following 36 months, he/she was grouped into pMCI; and if the patient's diagnosis remained as MCI, we categorized him/her as sMCI.

Wen et al.'s review paper [13] reimplemented most of the best performing Alzheimer's Disease classification methods and benchmarked their sMCI and pMCI classification performance using the ADNI dataset. This allowed the different methods to be compared using the same dataset and same experimental setup. Hence we have based our sMCI versus pMCI classification study on the same data splits as that used in [13]. It includes 298 sMCI and 295 pMCI participants retrieved from datasets ANDI1 and ADNI2. Each subject had one structural T1 weighted MRI scan taken at the baseline. The corresponding neuropsychological data such as MMSE, CDR-SB, and ADAS-cog were also recorded in the dataset. The demographic information of the participants used in this paper is summarised in Table 1.

[2] [http://adni.loni.usc.edu]

In our experiments we performed pre-training on three tasks using ADNI1 and ADNI2 datasets. The first task was AD versus CN classification with 508 AD versus 508 CN images. The second and third tasks were high versus low CDR-SB and ADAS11 cognitive score classification using 1243 training images of which 382, 460, 401 were classified as MCI, CN and AD respectively and 310 testing images comprising 93 MCI, 109 CN and 108 AD.

3.2. Image preprocessing

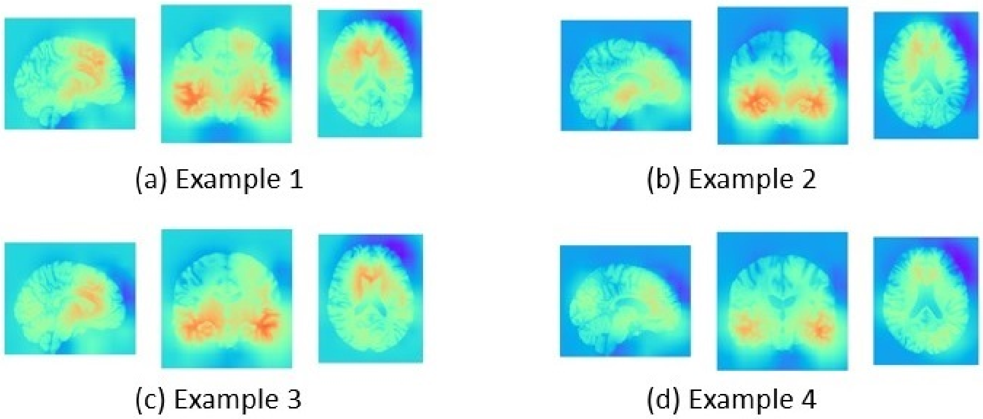
All the brain MR images acquired from ADNI1 and ADNI2 had undergone some steps of preprocessing such as N3 Intensity non-uniformity correction, B1 non-uniformity correction, and 3D Gradwarp correction for gradient nonlinearity if necessary. For better differentiating MRI images among subjects, a further preprocess was performed. First, N4ITK was used for intensity non-uniformity correction by the ANTS N4 BiasField Correction pipeline. The toolkit is available on the website [3]. Then, we performed an affine registration to standardize MRI data by nonlinearly aligning image data onto the template MNI125. Finally, all nonlinearly registered images were cropped to an identical size of 169×208×179 with 1 mm^3 isotropic voxels for computational efficiency. The tool FSL for brain extraction and registration can be acquired on the website (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) [4]. We found the class activation map (CAM) [12] for a classifier trained to classify between AD versus CN classification contained a lot of highly useful information. Since the trained model needed to focus its attention on the discriminative parts of the brain for separating the classes. We pretrained a 5-layer CNN model for the classification of AD and CN subjects. Figure 1 shows examples of CAM heatmaps for the AD class. The CAM heatmaps show highlighting of the following brain regions: hippocampus, entorhinal cortex, and ventricles, etc. These are consistent with traditional analysis of the brain anatomy of AD disease [18].

3.3. Class Activation Maps (CAM)

Here we explain how classification activation maps (CAM) developed by B Zhou et al. [12] can be obtained from a trained classification model. Zhou et al. [12] showed pretraining a CNN with a global average pooling (GAP) layer inserted between the final convolution layer and the output layer, can produce generic regional deep features for a particular class. Moreover, by using heatmaps,

[3] [http://stnava.github.io/ANTs/]

[4] [https://fsl.fmrib.ox.ac.uk/fsl/fslwiki]



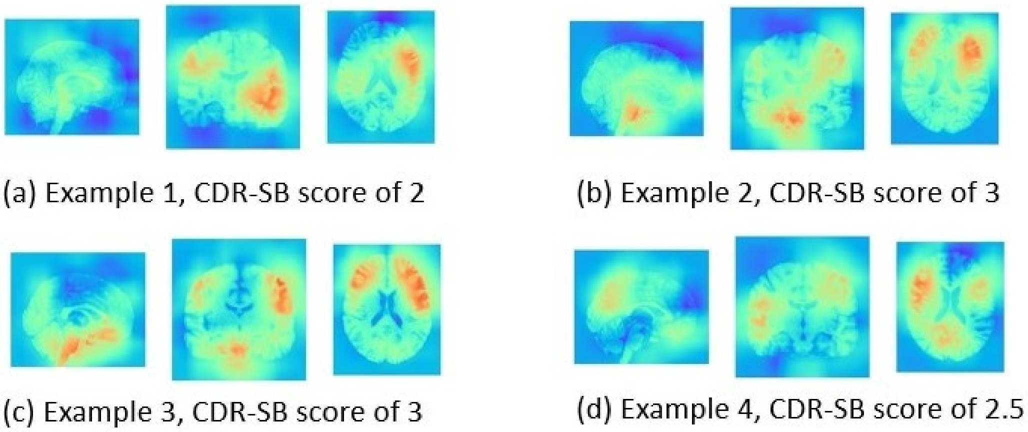
(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Figure 1: The generated CAMs associated with the AD category, for four different AD examples from the ADNI1 and ADNI2 datasets: the highlighted regions of the brain correspond to the known regions of the brain: hippocampus, entorhinal cortex.



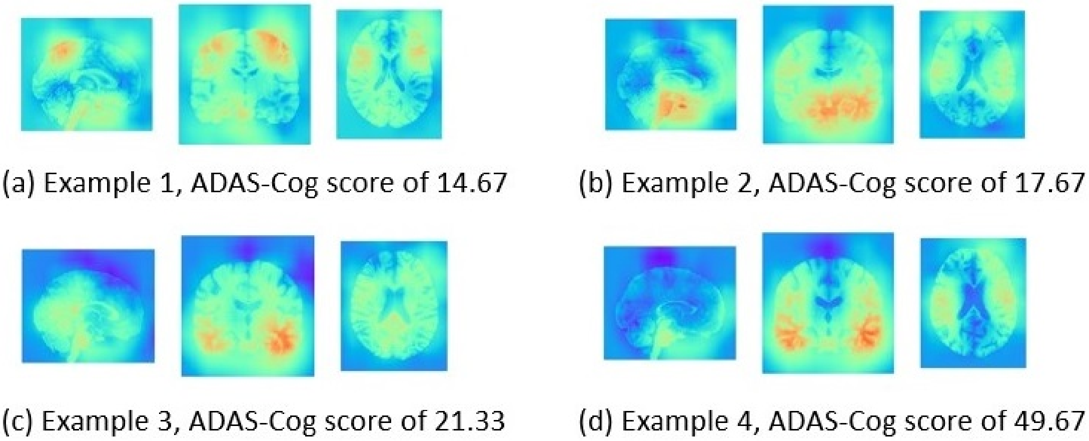
(a) Example 1, CDR-SB score of 2

(b) Example 2, CDR-SB score of 3

(c) Example 3, CDR-SB score of 3

(d) Example 4, CDR-SB score of 2.5

Figure 2: The CAM results of the binary classification of CDR-SB scores, created on a 5-layer 3D CNN model. The CDR-SB scores for these four examples are 2, 3, 3, and 2.5 respectively. From the CAM images, we can see all examples have some memory problems as the parts related to processing long-term memory have been highlighted, such as the hippocampus, entorhinal cortex, and prefrontal cortex, etc. Moreover, all these examples have some parts of their frontal lobe highlighted, which are associated with judgment and problem planning problems. We can also see the part of the parietal lobe in some examples is highlighted. The parietal lobe is related to attention, body awareness, sensations, and movement coordination, etc.



(a) Example 1, ADAS-Cog score of 14.67

(b) Example 2, ADAS-Cog score of 17.67

(c) Example 3, ADAS-Cog score of 21.33

(d) Example 4, ADAS-Cog score of 49.67

Figure 3: The CAM results of the binary classification of ADAS-Cog scores, created on a 5-layer 3D CNN model. The ADAS-Cog scores reflect subject-completed tests and observer-based assessments. Note that higher score means more diseased. The scores for the four examples shown in this figure are 14.67, 17.67, 21.33, and 49.67. All these four examples have above zero scores on the questions of Word Recall and Word Recognition. From their CAM images, we can see the parts related to short-term memory has been highlighted, such as the hippocampus, entorhinal cortex, and prefrontal cortex, etc. Examples 2 and 3 have got above zero score for Question Constructional Praxis and Orientation meaning these examples performed poorly in this task. Accordingly, the part of the brain involved in processing information (parietal lobe) and the part associated with short-memory tasks (frontal lobe) such as planning and motivation are highlighted.

CAM allows us to visualise the discriminative object areas associated with a particular predicted class. By simply upsampling the CAM to the given image size, those areas associated with a particular class can be visualized by overlaying the acquired heatmaps on the given images. The process of generating CAMs can be described as follows:

For a given image, after training on a typical CNN, we get the feature maps $f_m(x, y, z)$ at spatial location *(x, y, z)* in the last convolutional layer, where $\boldsymbol{m}$ indicates the number of filters. The output CAM $M_c(x, y, z)$ is defined as:

$$M_c(x, y, z) = \sum_m w^c f_m(x, y, z) \tag{1}$$

where, w^c is the weight matrix of the $\boldsymbol{m\text{-}th}$ filter associated with class $\boldsymbol{c}$. By stacking up all $\boldsymbol{m}$ outputs, the most discriminative regions can be highlighted via a heatmap.

We found similar results when we visualized the class activation maps for CDR-SB binary clas-

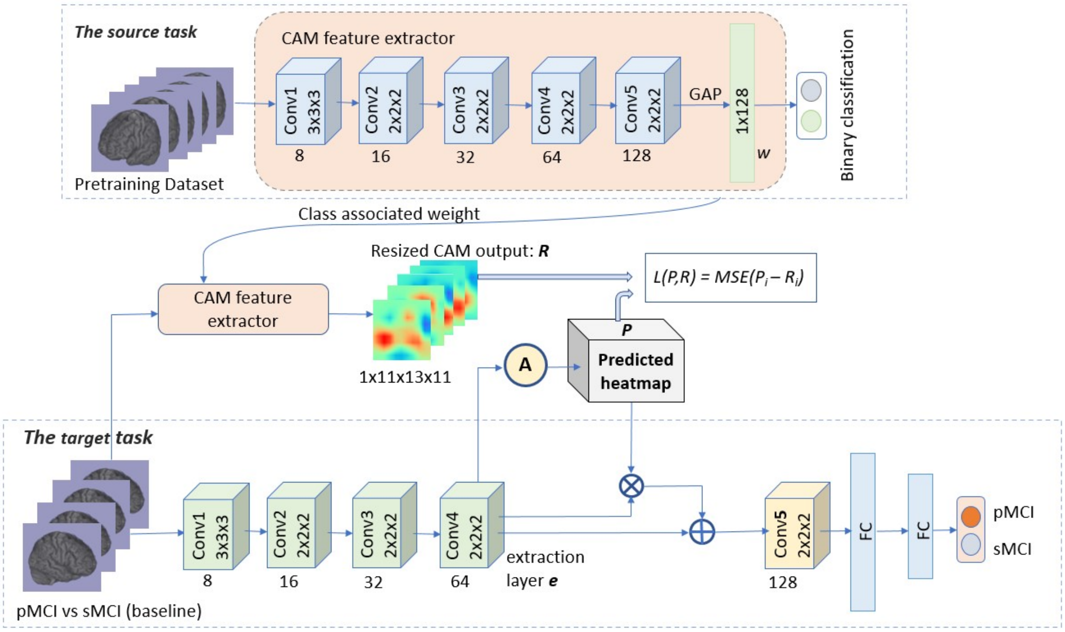


Figure 4: Illustration showing our proposed Class Activation Attention Transfer Network architecture consisting of two parts: *the target task* used to predict pMCI vs. sMCI, and *the source task* used for producing the predicted CAM attention for the target task. The input 3D image size is [c=1, w=169, h=208, d=179], c is the channel size. w is the CAM weight matrix which is the spatial average of the Conv5 feature map produced by global average pooling (GAP). A is the network attetion which calibrates the predicted heatmap. Note that the resized CAM outputs the 3D heatmap R with size [c=1, w=11, h=13, d=11]. The detailed model specifications for the target task and source task are presented in table 2

sification (high versus low CDR-SB score) and ADAS-cog binary classification (high versus low ADAS-cog classification). We use the median score as the threshold used to separate the low and high score classes for both ADAS-cog and CDR-SB. For ADAS-cog, the median is 10.33, and the median for CDR-SB is 1.5. Figure 2 and figure 3 show CAM images for CDR-SB and ADAS-cog binary classification. The MMSE scores are not used as their value distribution is highly skewed.

3.4. Class Activation Attention Transfer (CAAT)

Our aim is to use the information from the source task class activation maps described in the previous section to improve the accuracy of models trained for our target task of pMCI versus sMCI classification. Our model predicts the CAM produced by a model trained on the source task and use the resultant heatmap as an attention map when predicting pMCI versus sMCI.

In the rest of this section, we introduce our proposed Class Activation Attention Transfer (CAAT). As shown in figure 4, the proposed model consists of two parts: the source task and the target task. We employed the best performing subject-level architecture in [13], a five-layer 3D CNN network, for the target task of our model. Table 2 displays a precise description of the CNN architecture used in the target and source task phases. The source task CNN architecture is similar to the target task except for the last two FC layers (group 6) are replaced by a global average pooling (GAP) layer.

The source task was to output the CAM for the three binary classification tasks of AD vs CN, high versus low ADAS-cog and high versus low CDR-SB. We first train a model to perform each of these three binary classification tasks. We then extracted the weight matrices w^c of the associated more diseased classes c (AD, high ADAS-cog, and high CDR-SB). Then each pMCI or sMCI image I_i was fed into the 5-layer CNN to extract the feature maps f_i of the last CNN layer. Using the formula 1, we got the output that was the predicted CAM M_i^c for subject i. To use the predicted CAM M_i^c as attention for the target task, M_i^c need to be upsampled to the size of the predicted heatmap P_i in the target task and denoted it as $R_i = fn(M_i^c(x, y, z))$, where fn is an upsampling function (in our study, $fn = 1$ as the source task and the target task use the same CNN layer structure), $R_i \in \mathbf{R}^{W \times H \times L}$ ($W \times H \times L$ is the size of the predicted heatmaps). We call R_i the predicted CAM. Note that R_i represents a voxel-based vector, each element of the vector has its value constrained between [0, 1].

In the target task, each MRI image I_i was fed into the CNN model, the feature maps $f_m(x, y, z)$ of the extraction layer e (layer $Conv4$ in our experiments) were extracted. Here, m indicates the number of filters. To reduce the dimensionality and increase the nonlinearity of the predicted heatmap feature representation, the obtained feature maps $f_m(x, y, z)$ were then squeezed by using 3 Conv layers with 1x1x1 convolutions to create the predicted heatmap P for I_i. So the size of $f_m(x, y, z)$ was reduced to $P_i(x, y, z)$. $P_i(x, y, z)$ represents the voxel-wise feature vector. In order to make the output CAM R_i from the source task match with P_i and work as the attention for the whole network, we used MSE loss, which is formulated as:

$$L(P, R) = MSE(P_i - R_i) \tag{2}$$

where R_i is the upsampled CAM for the subject i. P_i is the squeezed feature representation (heatmap) from the extraction layer of the image for subject i. Both P_i and R_i are voxel-wise features with all values constrained within the range of [0,1].

We replicated P m times to create $\hat{P}$ and then performed element wise multiply with f_m to produce $Prd = \hat{P} \otimes f_m$. We concatenated Prd with f_m in order to pass both the original CNN features f_m and the features with attention Prd to the later classification layers. f_m acted like a skip connection to allow the later layers to directly use the original CNN features. The loss for the whole network was the sum of the loss from the target task network, and the loss between the predicted and target heatmap mentioned above 2. It can be formulated as:

$$L(Y_i, d_i, P_i, R_i) = aL(Y_i, d_i) + bL(P_i, R_i) \tag{3}$$

where $L(Y_i, d_i)$ is the Cross-Entropy Loss between Y_i and d_i. Y_i is the predicted diagnosis for subject i by the target task CNN, d_i is the true diagnosis for subject i. $L(P_i, R_i)$ is explained in 2. R_i indicated the output CAM by the source task network for subject i. P_i is the predicted heatmap from the extraction layer from the target task network. a and b were the coefficients for balancing the loss (in our experiment, $a = 0.8$, $b = 1.0$). We used three types of pretrained CAM outputs (AD, high ADAS-cog, and high-CDR-SB) as the attention for our proposed model. We also ensembled the predictions made by the three CAAT models (CAM of AD, high ADAS-cog and high CDR-SB) using majority voting to help reduce the effects of overfitting.

4. Experiments and Results

In this section, we explain how we set up the evaluation datasets of the experiment. We compare our proposed network against rival methods in terms of classification performance. We have also conducted an ablation study to determine how the attention part of CAAT contribute to the overall performance.

4.1. Experimental Setup

We performed all the experiments by using the stochastic gradient descent (SGD) optimiser for 65 epochs with the initial learning rate of $8e - 4$ and a batch size of 4. The learning rate was decreased by 0.5 after every 20 epochs. We trained our models on a GeForce RTX 2080 Ti GPU. We used the Pytorch deep learning framework to implement and train our CNN models.

Group	Target Task Layers	Source Task Layers
1	3x3x3 kennels, 8 output channels	
	3×3×3 max pool, 1 stride	
	4 BatchNorm	
	Relu activation	
2	2x2x2 kennels, 16 output channels	
	2×2×2 max pool, 2 stride	
	4 BatchNorm	
	Relu activation	
3	2x2x2 kennels, 32 output channels	
	2×2×2 max pool, 2 stride	
	4 BatchNorm	
	Relu activation	
4	2x2x2 kennels, 64 output channels	
	2×2×2 max pool, 2 stride	
	4 BatchNorm	
	Relu activation	
5	2x2x2 kennels, 128 output channels	
	2×2×2 max pool, 2 stride	
	4 BatchNorm	
	Relu activation	
6	nn.Linear (128 * 5 * 6 * 5, 1300), Relu	global average pooling
	nn.Linear (1300, 256), Relu	
	Softmax activation	Softmax activation

Table 2: 5-layer CNN architecture used for the source task and the target task for predicting pMCI vs. sMCI. The number of the channels from Conv1 to Conv5 are 8, 16, 32, 64, and 128 respectively. The stride used from Conv2 to Conv5, for the 2x2x2 kennels, is set as 2 and padding of 1, except the kennel (3x3x3) for the Conv1 is set as 1. All convolutions had 3x3x3 kernels, a stride of 1 and padding of 1. All convolutions had a padding of 1x1x1. The 2nd max pooling layers had a padding of (1, 0, 0). The 3rd max pooling layers had a padding of (1, 1, 0). The input image to the model is [c=1, w=169, h=208, d=179], here c is the channel size. Note that the difference between the source task and the target CNN architecture is the last two FC layers (group 6) in the target task are replaced by a global average pooling (GAP) layer in the source task.

4.2. Evaluation measures and comparison methods

Our dataset consists of 593 MRI images, consisting of 298 sMCIs and 295 pMCIs images. We performed 5 fold cross validation using the same data splits as that used in [13] [5]. In order to gain a comprehensive view of the performance of the algorithms, we used the following four evaluation metrics for the model performance include sensitivity (SEN), specificity (SPC), F1 score (F1) and accuracy (ACC).

Our experimental study included the following methods:

- ***Baseline 3D CNN:*** To evaluate the classification performance of our model, the 5-layer 3D CNN model used in [13] was implemented as the baseline model.
- ***Transfer learning AD/CN, CDR-SB, ADAS:*** We applied the traditional transfer learning on the Baseline 3D CNN model by using the pretrained network weights obtained from three different classification tasks: CN vs. AD, high versus low ADAS-cog score, and high vs. low CDR-SB score, respectively.
- ***6-Conv Transfer learning AD/CN, CDR-SB, ADAS:*** We added one more convolutional layer on the Baseline 3D CNN model and made a 6-layer 3D CNN model in order to provide a fairer comparison with CAAT in terms of the number of the network parameters and model depth. We also applied the traditional transfer learning method (pre-training on CN vs. AD, high vs. low ADAS-cog score, and high vs. low CDR-SB score) on this 6-layer 3D CNN model.
- ***6-Conv Transfer learning ensemble:*** The three predictions of *6-Conv Transfer learning AD/CN, 6-Conv Transfer learning CDR-SB, 6-Conv Transfer learning ADAS* were ensembled and the final result was decided by a majority voting method.
- ***CAAT AD, CAAT CDR-SB, CAAT ADAS:*** We report the results of three implementations of our CAAT model, each with one of the following source tasks: AD versus CN classification; high versus low ADAS cog score classification and; high versus low CDR SB Score classification.

[5] [`https://github.com/aramis-lab/AD-DL/tree/master/data/ADNI`]

- ***CAAT ensemble:*** In order to reduce the effects of overfitting, the prediction results of CAAT AD, CAAT CDR-SB, CAAT ADAS were ensembled using majority voting.

- ***Transfer Learning AD/CN + CAAT AD, Transfer Learning CDR-SB + CAAT AD, Transfer Learning ADAS + CAAT AD:*** We applied the traditional transfer learning method for the target network part on Conv1, Con2, and Conv3 layers by using pretrained weights of classification tasks for CN versus AD, high versus low ADAS-cog score, and high versus low CDR-SB score, respectively. Meanwhile, we passed the predicted CAM associated with AD from the source task to the target task network working with the predicted heatmap as the transferred attention as well. Hence these methods use both traditional transfer learning and also CAAT to transfer attention maps from the source task of AD versus CN classification.

- ***Transfer Learning + CAAT AD ensemble:*** This is similar to CAAT ensemble, we ensembled the three predictions of *Transfer Learning AD/CN + CAAT AD, Transfer Learning CDR-SB + CAAT AD, Transfer Learning ADAS + CAAT AD* using majority voting.

4.3. Results comparing CAAT with existing methods

Experimental results in Table 3 indicate that our proposed CAAT ensemble method has the highest accuracy among all methods tested. Compared with 3D ROI-based CNN, the CAAT ensemble model archives slightly higher accuracy without requiring expert knowledge. The results show that the source task in our CAAT method is able to detect the important brain areas via generating CAM and the attention mechanism enable the network focus on the important brain information, which are helpful for classifying pMCI and sMCI in the target task.

4.4. Impact of Tansfer learning

We further conducted a series of experiments to investigate the impact of the traditional transfer learning methods. Experiments results are reported in Table 4. We make the following observations from the experimental results. The results show traditional transfer learning consistently outperforms the baseline solution. This is likely due to transfer learning's ability to leverage the larger dataset used for the source tasks (AD versus CN, high versus low ADAS-cog and high versus low CDR-SB classification) to learn useful features for the target task in pMCI versus sMCI classification.

The results show traditional transfer learning using 6 Conv Layers generally perform better than traditional transfer learning using just 5 Conv layers. It verifies that using a deeper model can produce

	pMCI vs. sMCI			
Model	SEN	SPE	F1	ACC(%)
Baseline 3D CNN	0.71	0.71	0.71	70.84
3D ROI-based CNN [9]	-	-	-	74.00
3D patch-level CNN [15]	-	-	-	70.00
CAAT AD	0.70	0.75	0.72	73.03
CAAT CDR-SB	**0.75**	0.71	0.73	72.70
CAAT ADAS	0.73	0.74	0.73	73.03
CAAT ensemble	**0.75**	0.75	**0.75**	**74.61**

Table 3: Experimental results comparing existing CNN-based methods for pMCI versus sMCI classification against variants of our CAAT method. For fair comparison, all the existing methods reported in this table were trained and tested using the same train/validation splits as reported on the review paper [13]. The best results for each evaluation metric is highlighted in bold text font. SEN, SPE, F1 and ACC refer to the sensitivity, specificity, F1 score and accuracy metrics respectively.

better results. This maybe due to the extra hidden layer creating more abstract and discriminative features than a shallower model.

Compared to the other models, our proposed CAAT ensemble model achieves the highest performance for all metrics with the exception of specificity where it only performs 0.01 worse than the best performer. In contrast, none of the traditional transfer learning solutions consistently performs near the best for all metrics. This demonstrates that the prediction ability of the CAAT model is improved by using the attention mechanism. The heatmap from CAM (AD, high ADAS-cog, and high CDR-SB) helps the model to focus on the parts of the brain that was most discriminative for the source task. Since both the source and target tasks are very related, these attention heatmaps when applied to the target task helps the model to ignore unimportant regions of the brain and thereby help CAAT reduce the amount of overfitting. The results also show combining traditional transfer learning and CAAT performs slightly worse than using CAAT by itself.

4.5. Ablation Study

We performed an ablation study to gain insights into our CAAT. The results are reported in table 5.

	pMCI vs. sMCI			
Model	SEN	SPE	F1	ACC(%)
Baseline 3D CNN	0.71	0.71	0.71	70.84
Transfer Learning AD/CN	0.71	0.71	0.71	71.35
Transfer Learning CDR-SB	0.67	0.75	0.70	71.00
Transfer Learning ADAS	0.74	0.67	0.72	70.50
Transfer learning ensemble	0.73	0.72	0.72	72.18
6-Conv Transfer Learning AD/CN	0.71	**0.76**	0.73	73.03
6-Conv Transfer Learning CDR-SB	0.72	0.71	0.72	71.85
6-Conv Transfer Learning ADAS	0.70	0.72	0.71	71.34
6-Conv Transfer learning ensemble	0.72	0.75	0.73	73.37
CAAT AD	0.70	0.75	0.72	73.03
CAAT CDR-SB	**0.75**	0.71	0.73	72.70
CAAT ADAS	0.73	0.74	0.73	73.03
CAAT ensemble	**0.75**	0.75	**0.75**	**74.61**
Transfer Learning AD/CN + CAAT AD	0.72	0.74	0.73	73.03
Transfer Learning CDR-SB + CAAT AD	0.72	0.75	0.73	73.52
Transfer Learning ADAS + CAAT AD	**0.75**	0.71	0.73	72.86
Transfer Learning + CAAT AD ensemble	**0.75**	0.73	0.74	74.04

Table 4: Experimental results comparing traditional transfer learning against our CAAT variants. The best results for each evaluation metric is highlighted in bold text font. SEN, SPE, F1 and ACC refer to the sensitivity, specificity, F1 score and accuracy metrics respectively.

	pMCI vs. sMCI			
Model	SEN	SPE	F1	ACC(%)
CAAT AD-attention on Layer Conv3	0.705	0.715	0.710	71.01
CAAT AD-intra-task attention	**0.709**	0.715	0.709	70.68
CAAT AD-no signal	0.705	0.722	0.711	71.34
CAAT AD	0.705	**0.753**	**0.723**	**73.03**

Table 5: Results of an ablation study of our CAAT AD method. The best results for each evaluation metric is highlighted in bold text font. SEN, SPE, F1 and ACC refer to the sensitivity, specificity, F1 score and accuracy metrics respectively.

We observed that adding the attention on the layer Conv4 of the CAAT model performs better than on the layer Con3. This is likely due to the fact the latter convolutional layer (Conv4) learn more high level features and patterns than the earlier layer (Conv3). The attention derived from the higher level features is more likely to highlight larger areas of importance than very detailed small regions. This coarser grained attention will be less likely cause overfitting.

We perform the following tests to determine how the attention impacts the performance of our proposed model. First, we turned off the loss function between the predicted heatmap P and the predicted CAM R, we describe this model as *CAAT AD-intra-task attention* because this means the model was no longer trying to train the attention to mimic the attention from the source task. Additionally, we stopped the model from using any attention by fixing each voxel value of the predicted heatmap P to a constant value of 1 / (11 x 13 x 11), where the denominator is the heatmap size. We denote this model as *CAAT AD-no signal*.

The results show that both *CAAT AD-intra-task attention* and *CAAT AD-no signal* perform worse that our normal *CAAT AD* model. This shows that attention learnt only from the target task is not as effective as attention transferred from the source task. Second, not using any attention is worse than using transferred attention.

The ablation study experiments show that the attention transfer mechanism in our proposed CAAT method is critical to the good performance of CAAT AD. The output CAM from the source task passed to callibrate the predicted attention heatmap enables the network to focus on the highly predictive parts of the brain based on knowledge gained from performing the source task.

5. Conclusion

In this paper, we presented our Class Activation Attention Transfer (CAAT) method which offers an alternative way of leveraging labeled data from a source classification task to enhance the classification accuracy of a target task. CAAT transfers attention from the source task to the target task instead of transferring the weights. Our experiments show transferring attention works better than transferring weights for the pMCI versus sMCI classification task. In addition, when we visualized the attention heatmaps (CAMs) that are transferred to the target task, we found the regions highlighted by the heatmap match known important regions for diagnosing Alzheimer's disease. Results also show that CAAT can outperform the previous state-of-the-art region of interest-based solutions that required expert domain knowledge to manually select regions of interest. In contrast, CAAT automatically selects the regions of interest via the CAM heatmaps.

For future work, we would like to explore predicting, ADAS, MMSE, CDR scores, or predicting brain age as the target task and using a source task such as AD versus CN classification.

Acknowledgement

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement List.pdf

References

[1] G. Moya-Alvarado, N. Gershoni-Emek, E. Perlson, F. C. Bronfman, Neurodegeneration and alzheimer's disease (ad). what can proteomics tell us about the alzheimer's brain?, Molecular & cellular proteomics 15 (2) (2016) 409–425. doi:https://doi:10.1074/mcp.R115.053330. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4739664/

[2] J. L. Cummings, C. Back, The cholinergic hypothesis of neuropsychiatric symptoms in alzheimer's disease, The American Journal of Geriatric Psychiatry 6 (2, Supplement 1) (1998)

S64–S78. doi:https://doi.org/10.1097/00019442-199821001-00009.
URL https://www.sciencedirect.com/science/article/pii/S106474811261063X

[3] S. Karantzoulis, J. E. Galvin, Distinguishing alzheimer's disease from other major forms of dementia, Expert Rev Neurother 11 (11) (2011) 1579–1591. doi:https://doi:10.1586/ern.11.155.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3225285/

[4] M. H. Tabert, J. J. Manly, X. Liu, G. H. Pelton, S. Rosenblum, M. Jacobs, D. Zamora, M. Goodkind, K. Bell, Y. Stern, D. P. Devanand, Neuropsychological prediction of conversion to alzheimer disease in patients with mild cognitive impairment, Archives of General Psychiatry 63 (8) (2006) 916–924. arXiv:https://jamanetwork.com/journals/jamapsychiatry/articlepdf/668194/yoa60002.pdf, doi:10.1001/archpsyc.63.8.916.
URL https://doi.org/10.1001/archpsyc.63.8.916

[5] M. A. DeTure, D. W. Dickson, The neuropathological diagnosis of alzheimer's disease, Molecular Neurodegeneration 14 (32) (2019). doi:https://doi.org/10.1186/s13024-019-0333-5.

[6] G. K. Bhatti, A. P. Reddy, P. H. Reddy, J. S. Bhatti, Lifestyle modifications and nutritional interventions in aging-associated cognitive decline and alzheimer's disease, Frontiers in Aging Neuroscience 11 (2020) 369. doi:10.3389/fnagi.2019.00369.
URL https://www.frontiersin.org/article/10.3389/fnagi.2019.00369

[7] M. J. Knight, B. McCann, R. A. Kauppinen, E. J. Coulthard, Magnetic resonance imaging to detect early molecular and cellular changes in alzheimer's disease, Frontiers in Aging Neuroscience 8 (2016) 139. doi:10.3389/fnagi.2016.00139.
URL https://www.frontiersin.org/article/10.3389/fnagi.2016.00139

[8] C. N. Harada, M. C. N. Love, K. Triebel, Normal cognitive aging, Clin Geriatr Med 29 (4) (2013) 737–52. doi:https://doi:10.1016/j.cger.2013.07.002.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015335/

[9] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, T. A. D. N. I. , Convolutional neural networks-based MRI image analysis for the alzheimer's disease prediction from mild cognitive impairment, Frontiers in Neuroscience 12 (2018) 777. doi:10.3389/fnins.2018.00777.
URL https://www.frontiersin.org/article/10.3389/fnins.2018.00777

[10] B. Cheng, M. Liu, D. Zhang, D. Shen, Robust multi-label transfer feature learning for early diagnosis of alzheimer's disease, Brain Imaging and Behavior 63 (2019) 138–153. doi:10.1007/s11682-018-9846-8.
URL https://doi.org/10.1007/s11682-018-9846-8

[11] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, G. Catheline, Classification of alzheimer disease on imaging modalities with deep cnns using cross-modal transfer learning, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 2018, pp. 345–350. doi:10.1109/CBMS.2018.00067.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Computer Vision and Pattern Recognition, 2016.

[13] J. Wen, E. Thibeau-Sutre, J. Samper-González, A. Routier, S. Bottani, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation, CoRR abs/1904.07773 (2019). arXiv:1904.07773.
URL http://arxiv.org/abs/1904.07773

[14] S. Mathotaarachchi, T. A. Pascoal, M. Shin, A. L. Benedet, M. S. Kang, T. Beaudry, V. S. Fonov, S. Gauthier, P. Rosa-Neto, Identifying incipient dementia individuals using machine learning and amyloid imaging, Neurobiology of Aging 59 (2017) 80–90. doi:https://doi.org/10.1016/j.neurobiolaging.2017.06.027.
URL https://www.sciencedirect.com/science/article/pii/S0197458017302294

[15] M. Liu, D. Cheng, K. Wang, Y. Wang, Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis, Neuroinformatics 16 (10 2018). doi:10.1007/s12021-018-9370-4.

[16] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, Automated classification of alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks, NeuroImage: Clinical 21 (2019) 101645. doi:https://doi.org/10.1016/j.nicl.2018.101645.
URL https://www.sciencedirect.com/science/article/pii/S2213158218303930

[17] C. Lian, M. Liu, J. Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural MRI, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (4) (2020) 880–893. doi:10.1109/TPAMI.2018.2889096.

[18] G. B. Frisoni, N. C. Fox, C. R. J. Jack, P. Scheltens, P. M. Thompson, The clinical use of structural MRI in alzheimer disease, Nat Rev Neurol 6(2) (2010) 66–67. doi:10.1038/nrneurol.2009.215.

[19] C. Lian, M. Liu, L. Wang, D. Shen, End-to-end dementia status prediction from brain MRI using multi-task weakly-supervised attention network, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part IV, Vol. 11767 of Lecture Notes in Computer Science, Springer, Cham, 2019, pp. 158–167. doi:10.1007/978-3-030-32251-9_18.
URL https://doi.org/10.1007/978-3-030-32251-9_18

[20] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images (2019). arXiv:1808.08114.

[21] S. Jetley, N. A. Lord, N. Lee, P. H. S. Torr, Learn to pay attention (2018). arXiv:1804.02391.

[22] J. Schlemper, O. Oktay, L. Chen, J. Matthew, C. L. Knight, B. Kainz, B. Glocker, D. Rueckert, Attention-gated networks for improving ultrasound scan plane detection, CoRR abs/1804.05338 (2018). arXiv:1804.05338.
URL http://arxiv.org/abs/1804.05338

[23] J. Huang, L. Zhou, L. Wang, D. Zhang, Attention-diffusion-bilinear neural network for brain network analysis, IEEE Transactions on Medical Imaging 39 (7) (2020) 2541–2552.

[24] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net (2015). arXiv:1412.6806.

[25] M. Lin, Q. Chen, S. Yan, Network in network (2014). arXiv:1312.4400.

[26] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks (2018). arXiv:1608.06993.

[27] J. K. Kueper, M. Speechley, M. Montero-Odasso, The alzheimer's disease assessment scale–cognitive subscale (adas-cog): Modifications and responsiveness in pre-dementia populations. a narrative review, Journal of Alzheimer's Disease 63 (2018) 423–444. doi:10.3233/

JAD-170991.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5929311/

[28] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, S. M. Smith, Accurate brain age prediction with lightweight deep neural networks, Medical Image Analysis 68 (2021) 101871. doi:https://doi.org/10.1016/j.media.2020.101871.
URL http://www.sciencedirect.com/science/article/pii/S1361841520302358

[29] S. Balsis, J. F. Benge, D. A. Lowe, L. Geraci, R. S. Doody, How do scores on the adas-cog, mmse, and cdr-sob correspond?, The Clinical Neuropsychologist 29 (7) (2015) 1002–1009, pMID: 26617181. arXiv:https://doi.org/10.1080/13854046.2015.1119312, doi:10.1080/13854046.2015.1119312.
URL https://doi.org/10.1080/13854046.2015.1119312

[30] J. H. Grochowalski, Y. Liu, K. L. Siedlecki, Examining the reliability of adas-cog change scores, Aging, Neuropsychology, and Cognition 23 (5) (2016) 513–529, pMID: 26708116. arXiv:https://doi.org/10.1080/13825585.2015.1127320, doi:10.1080/13825585.2015.1127320.
URL https://doi.org/10.1080/13825585.2015.1127320

[31] A. Nibali, Z. He, D. Wollersheim, Pulmonary nodule classification with deep residual networks, International journal of computer assisted radiology and surgery 12 (10) (2017) 1799—1808. doi:10.1007/s11548-017-1605-6.
URL https://doi.org/10.1007/s11548-017-1605-6

[32] M. Liu, J. Zhang, D. Nie, P. T. Yap, D. Shen, Anatomical landmark based deep feature representation for mr images in brain disease diagnosis, IEEE Journal of Biomedical and Health Informatics 22 (5) (2018) 1476–1485. doi:10.1109/JBHI.2018.2791863.

[33] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708. doi:10.1109/CVPR.2014.220.

[34] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[35] A. M. Alayba, V. Palade, M. England, R. Iqbal, A combined cnn and lstm model for arabic sentiment analysis, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2018, pp. 179–191.

[36] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing [review article], IEEE Computational Intelligence Magazine 13 (3) (2018) 55–75. doi:10.1109/MCI.2018.2840738.

[37] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, H. Radha, Deep learning algorithm for autonomous driving using googlenet, in: 2017 IEEE Intelligent Vehicles Symposium (IV), 2017, pp. 89–96. doi:10.1109/IVS.2017.7995703.

[38] C. Salvatore, A. Cerasa, I. Castiglioni, MRI characterizes the progressive course of ad and predicts conversion to alzheimer's dementia 24 months before probable diagnosis, Frontiers in Aging Neuroscience 10 (2018) 135. doi:10.3389/fnagi.2018.00135.
URL https://www.frontiersin.org/article/10.3389/fnagi.2018.00135