Computers in Biology and Medicine 156 (2023) 106700

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Class activation attention transfer neural networks for MCI conversion prediction

Min Luo [a], Zhen He [a,*], Hui Cui [a], Yi-Ping Phoebe Chen [a], Phillip Ward [b,c,d], 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

ARTICLE INFO

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

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 (task we are transferring information from) to automatically learn regions of interest (ROI) from a given image. Next we train a model to simultaneously classify progressive MCI (pMCI) and stable MCI (sMCI) (the target task we want to solve) and the ROIs learned 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.

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 ageing 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 using Magnetic Resonance Imaging (MRI) is of significant importance. This information is critical for clinical trials, decisions for early interventions, and to maximize 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. 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.

We use convolutional neural networks (CNN) to solve this problem by leveraging data labelled for related tasks. Datasets for classifying pMCI versus sMCI are typically small (593 subjects in our dataset) since it requires repeated MRI scans to compare baseline versus later diagnosis. In contrast, datasets with AD versus CN labels or cognitive subscale labels such as ADAS-cog and CDR-SB are typically larger (in our case we have 1587 subjects) since only a single MRI scan is needed to measure these values. An interesting research question is how can

* Corresponding author.
E-mail address: z.he@latrobe.edu.au (Z. 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.

https://doi.org/10.1016/j.compbiomed.2023.106700
Received 17 November 2021; Received in revised form 24 August 2022; Accepted 9 December 2022
Available online 21 February 2023

we best use the images from the entire 1587 subjects? A traditional method for achieving this is to use transfer learning [7,8], where model weights learned from a source classification task (e.g. Alzheimer's Disease/Cognitively normal (AD/CN), high/low Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-cog) score, high/low Clinical Dementia Rating scale Sum of Boxes (CDR-SB) score) are transferred to the target classification task (pMCI/sMCI). In this paper the term source task is used to define the task we want to transfer information from and the term target task refers to the actual task we want to solve. 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) learned from previous tasks to a new task. Existing ROI based pMCI versus sMCI classification approaches [7,9] 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 pMCI versus 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) [10] 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 discriminating 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 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 [11] that only uses whole 3D brain scans without using transfer learning or attention only achieved an accuracy of 70.84 in our experiments. 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 [8,12]. Mathotaarachchi et al. [12] 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. [8], each subject image had 93 manually-labelled 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. [13] 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. [7] 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 centred at the same point. These features were combined with the features obtained from FreeSurfer [14]. 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 [13] we only use the MRI images and do not use PET images. We compare to [7] we transfer attention from the source to the target task instead of the weights of the neural network.

Basaia et al. [15] used data augmentation techniques like flips, rotations, cropping to increase training set size and trained the data using a VGG-like network. Liu and Zhang et al. [16] proposed a landmark based sMCI versus pMCI classification method which firstly uses a landmark prediction application [17][2] to obtain the top 50 landmarks pretrained on an AD versus CN classification task. After obtaining the 50 landmarks, each landmark is used to generate multiple 3D patches by shifting the centre point for each landmark a few pixels multiple times. This is a form of data augmentation. After creating a bag of patches for each subject, they fed each patch into a 3D CNN to create learned features for each patch. Then concatenate the features for each patch together and feed that into an MLP to perform sMCI versus pMCI classification. In Lian and Liu et al.'s work [18], they first used the landmark prediction application to choose the top 120 landmarks to create proposal locations. After that, they feed all those patches into a 3D CNN to create a patch-level feature representation. These patch-level features are then used as the input to the subsequent region-level sub-networks to get the discriminative capacity of the corresponding region. Finally, the region-level feature representations and the classification scores of the region-level sub-networks are concatenated and processed by the subject-level sub-network. In contrast to [15] our method transfers attention from a source task instead of performing

[2] https://github.com/zhangjun001/AD-Landmark-Prediction.

data augmentation to improve performance. When compared to [16, 18] our method uses CAM heatmaps from the source task to determine where to focus attention instead of ROI-based patch selection.

There has been many existing works in computer vision [19–21], and medical imaging [22–24] that use class activation maps (CAMs) for visualizing the impact of input pixels to the predicted class. In contrast, we use CAMs to transfer attention from the source task to the target task. Hence, CAMs is used in a novel way to improve final classification performance on the target task instead of using CAMs to analyse model decision making.

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.

3.1. Subjects and data acquisition

This paper uses the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset[3]. 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.

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[4]. Then, we first perform linear (affine) registration using the FLIRT function from the FSL software package to align the image to the template MNI125. We then perform non-linear registration using FNIRT on the output of FLIRT. Doing the linear registration first helps us to reduce the distortion caused by the subsequent nonlinear registration. Finally, all nonlinearly registered images were cropped to an identical size of 169 x 208 x 179 with 1 mm^3 isotropic voxels for computational efficiency. The FSL software package for brain extraction and registration can be acquired on the website (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki)[5].

We found the class activation map (CAM) [10] 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. Fig. 1 shows examples of CAM heatmaps for the AD class. The following brains are highlighted by the CAM heatmaps: hippocampus, entorhinal cortex, and ventricles, etc. These are consistent with traditional analysis of the brain anatomy of AD disease [25].

We compared our approach against the landmark-based deep multi-instance learning (LDMIL) method [16]. We preprocessed the MRI images based on the approach used in [16]. We first performed N4 intensity non-uniformity correction and then applied linear registration to align the image to the template MNI125 using FSL. We then performed linear alignment for all images to the provided template image *Img.nii.gz*[6]. Finally, we used a mask to remove the cerebrum, pons, spinal cord, etc regions from each MRI image.

3.3. Class Activation Maps (CAM)

Here we explain how classification activation maps (CAM) developed by B Zhou et al. [10] can be obtained from a trained classification model. Zhou et al. [10] 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, CAM allows us to visualize 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 ***m-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 classification (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. The MMSE scores are not used as their value distribution is highly skewed.

Figs. 2 and 3 show CAM images for CDR-SB and ADAS-cog binary classification. These images allow us to verify that the CAMs are indeed highlighting the known regions of interest that are pertinent to diagnosing Alzheimer's disease.

From the CDR-SB CAM images from Fig. 2 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 judgement and problem planning problems. We can also see the part of the parietal lobe highlighted in the examples. The parietal lobe is related to attention, body awareness, sensations, and movement coordination, etc.

From the ADAS-cog CAM images from Fig. 3 we can see the parts related to short-term memory have been highlighted, such as the

[3] http://adni.loni.usc.edu.
[4] http://stnava.github.io/ANTs/.
[5] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki.
[6] https://github.com/zhangjun001/AD-Landmark-Prediction.

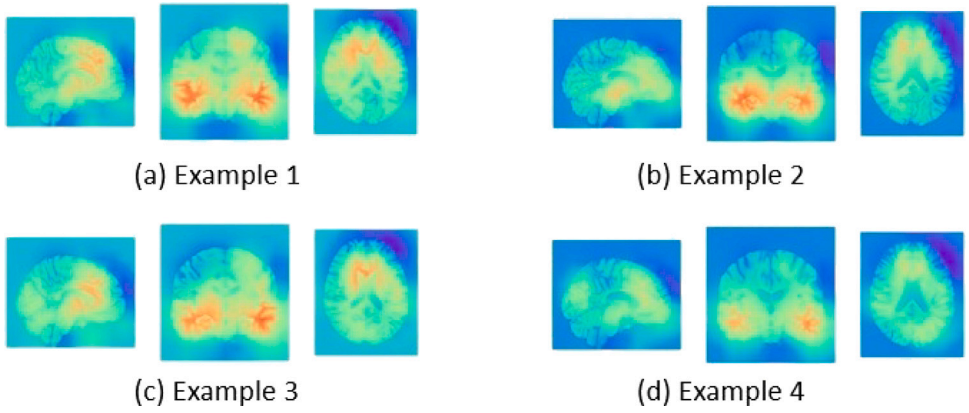

Fig. 1. The generated CAMs associated with the AD category, for four different AD examples from the ADNI1 and ADNI2 datasets: the highlighted regions of the brain correspond to the known regions of the brain: hippocampus, entorhinal cortex.

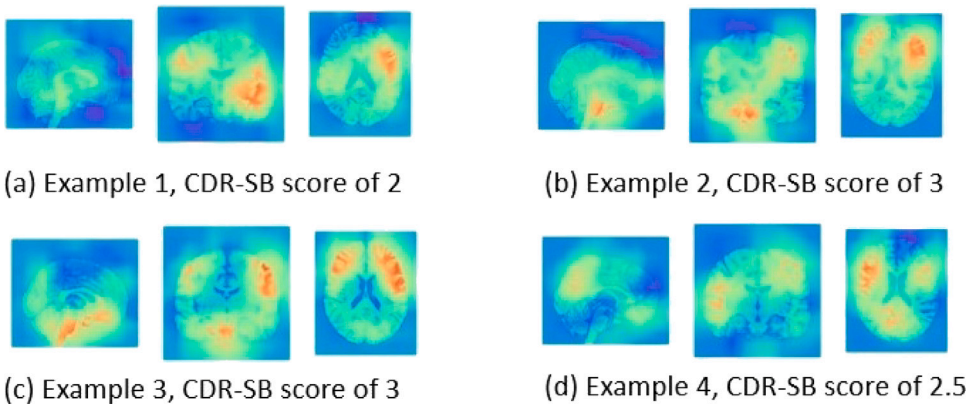

Fig. 2. The CAM results of the binary classification of CDR-SB scores, created on a 5-layer 3D CNN model. The CDR-SB scores for these four examples are 2, 3, 3, and 2.5 respectively.

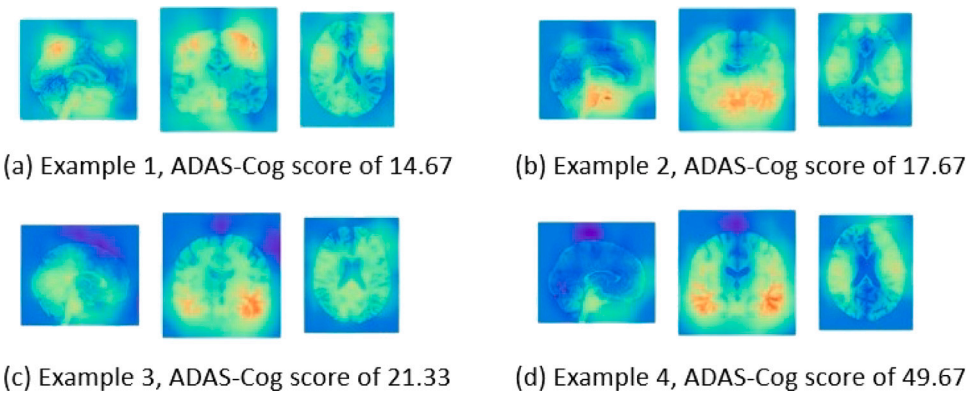

Fig. 3. The CAM results of the binary classification of ADAS-Cog scores, created on a 5-layer 3D CNN model. The ADAS-Cog scores reflect subject-completed tests and observer-based assessments. Note that higher score means more diseased. All four examples have above zero scores on the questions of Word Recall and Word Recognition.

hippocampus, entorhinal cortex, and prefrontal cortex, etc. Examples 2 and 3 have gotten 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.

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) method. As shown in Fig. 4, the proposed model consists of two parts: the source task and the target task. We employed the best performing subject-level architecture in [11], a five-layer 3D CNN network, for the target task of our model. Table 1 displays a precise description of the CNN architecture used in the target and source task phases. The source task CNN architecture is

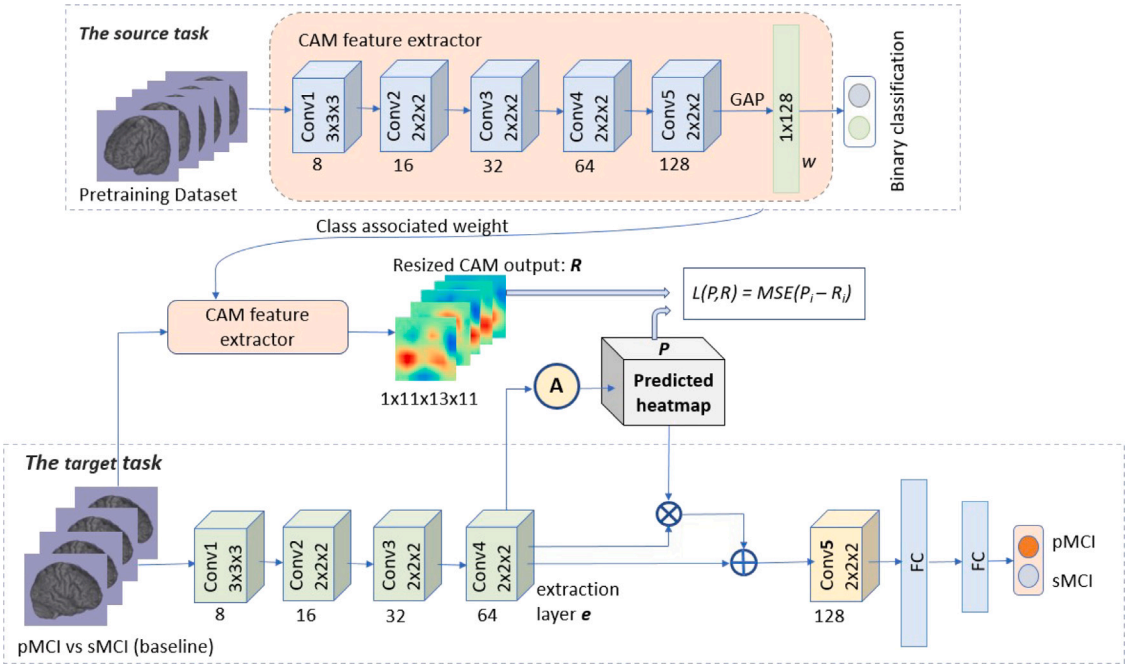

Fig. 4. Illustration showing our proposed Class Activation Attention Transfer Network architecture consisting of two parts: *the target task* used to predict pMCI vs. sMCI, and *the source task* used for producing the predicted CAM attention for the target task. The input 3D image size is [c = 1, w = 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 attention 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 size of the feature map for the 4th CNN layer of the model). The detailed model specifications for the target task and source task are presented in Table 1.

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 $1 \times 1 \times 1$ 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. This allows the model to separately learn distinct features that are more particular to the target task of classifying between pMCI and sMCI.

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 in Eq. (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.

Table 1
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 $2 \times 2 \times 2$ kennels, is set as 2 and padding of 1, except the kennel ($3 \times 3 \times 3$) for the Conv1 is set as 1. All convolutions had $3 \times 3 \times 3$ kernels, a stride of 1 and padding of 1. All convolutions had a padding of 1×1x1. 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.

Group	Target task layers	Source task layers
1	$3 \times 3 \times 3$ kennels, 8 output channels $3 \times 3 \times 3$ max pool, 1 stride 4 BatchNorm Relu activation	
2	$2 \times 2 \times 2$ kennels, 16 output channels $2 \times 2 \times 2$ max pool, 2 stride 4 BatchNorm Relu activation	
3	$2 \times 2 \times 2$ kennels, 32 output channels $2 \times 2 \times 2$ max pool, 2 stride 4 BatchNorm Relu activation	
4	$2 \times 2 \times 2$ kennels, 64 output channels $2 \times 2 \times 2$ max pool, 2 stride 4 BatchNorm Relu activation	
5	$2 \times 2 \times 2$ kennels, 128 output channels $2 \times 2 \times 2$ max pool, 2 stride 4 BatchNorm Relu activation	
6	nn.Linear (128 * 5 * 6 * 5, 1300), Relu nn.Linear (1300, 256), Relu Softmax activation	Global average pooling Softmax activation

4.1. Dataset splits

In our paper we have adopted the same ADNI 1 and ADNI 2 dataset splits used in Wen et al.'s review paper [11]. The patient ids for the splits can be downloaded from[7]. The reason is Wen et al. re-implemented most of the best performing Alzheimer's Disease classification methods and benchmarked their sMCI and pMCI classification performance using the ADNI dataset. By adopting the dataset splits of Wen et al. [11], we can compare our algorithms against the different methods implemented in [11]. The dataset consists of 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 summarized in Table 2.

4.2. Experimental setup

We performed all the experiments by using the stochastic gradient descent (SGD) optimizer 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.

We found the above training parameters gave best performance at the end of model tuning. We found the models converge before 65 epochs. Using the SGD optimizer with manual reduction of the learning rate after every 20 epochs helped the model converge to a more stable accuracy compared to using automatic optimizers such as ADAM. We used a batch size of 4 due to the large memory consumption of the large 3D CNN models.

Table 2
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]

4.3. 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 [11][8]. 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 [11] 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.
- ***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.

[7] https://github.com/aramis-lab/AD-DL.

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

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 [11]. 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. The numbers in parentheses report the variance value.

Model	pMCI vs. sMCI				
	SEN	SPE	F1	ACC(%)	AUC
Baseline 3D CNN	0.71 (0.005)	0.71 (0.003)	0.71 (0.003)	70.84 (0.001)	0.758 (0.003)
3D ROI-based CNN [7]	–	–	–	74.00	–
3D patch-level CNN [13]	–	–	–	70.00	-
LDMIL [16]	0.70 (0.004)	0.74 (0.004)	0.71 (0.002)	71.68 (0.002)	0.761 (0.004)
CAAT AD	0.70 (0.004)	**0.75** (0.004)	0.72 (0.001)	73.03 (0.002)	0.779 (0.002)
CAAT CDR-SB	**0.75** (0.003)	0.71 (0.006)	0.73 (0.002)	72.70 (0.002)	0.773 (0.002)
CAAT ADAS	0.73 (0.006)	0.74 (0.006)	0.73 (0.003)	73.03 (0.002)	0.777 (0.003)
CAAT ensemble	0.75 (0.004)	0.75 (0.006)	**0.75** (0.002)	**74.61** (0.002)	**0.780** (0.003)

- ***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.

For our CAAT technique, when training the 5 layered CNN model used for the source task of AD versus CN classification, we stopped the training early when the model reached 84.6% validation accuracy. We found if we trained the model to its highest validation accuracy of 92%, the resultant CAM was not as helpful when used as the attention map in the target task of pMCI versus sMCI classification. This may be due to the model overfitting the source task if we do not stop early.

4.4. Results comparing CAAT with existing methods

Experimental results in Table 3 indicate that our proposed CAAT ensemble method has the highest accuracy and AUC 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.

We re-implemented the landmark-based deep multi-instance learning (LDMIL) method [16]. Based on [16], we first used the landmark-detection application [17][9] to obtain the top 50 landmarks and then randomly shifted each centre point a few voxels to create the bag of patches for each subject. Finally, we fed the generated patches into the multi-instance based CNN framework specified by [16] to classify each image as sMCI or pMCI. The results in Table 3 show our CAAT ensemble outperforms LDMIL in all metrics measured. This can be attributed to CAAT's ability to leverage knowledge learned from the source AD versus CN classification to better focus its attention on regions of the brain that is better at discriminating between AD versus CN based on the characteristics specific to each brain image. In contrast, LDMIL uses the same landmarks for all brain images.

CAAT ensemble achieved an accuracy of 74.61 compared to 74 for the 3D ROI-based CNN method. Although this may seem like a small improvement, our attention transfer approach does not require any expert knowledge. Our model obtains all its information from the training data. In contrast the 3D ROI-based CNN method requires expert knowledge to first identify regions of interest. The model is then trained on these hand identified areas of interest. The ability to automatically identify regions of interest allows our approach to be applied to a wider range of applications where expert knowledge may not exist.

4.5. Model architecture comparison

In this section we investigate whether using different CNN architectures as backbone can all benefit from our attention transfer method. To this end we implemented a 3D version of the popular MobileNetV2 [26] network. MobileNetV2 is a 2D CNN architecture designed for small devices. Therefore it is a very efficient feature extractor that uses features such as lightweight depthwise separable convolutions, and thin bottleneck blocks without any non-linearity. These weight layers allow MobileNetV2 to go much deeper while still imposing a small compute and memory footprint thus making it ideal for inflating to 3D. We used the 3D inflated version of MobileNetV2 from [27][10]. We implemented CAAT with 3D Moblienetv2 as the CNN backbone to investigate if attention transfer can benefit this very different 3D CNN architecture compared to our default architecture.

We applied CAAT to 3D MobileNetV2 (3D MBNv2) in the same way as our default 3D CNN architecture. We use 3D MobileNetV2 both for source tasks and target tasks. In order to get the output CAM on the source task, we first trained three binary classifiers (AD versus CN, high versus low ADAS-cog, and high versus low CDR-SB) on 3D MBNv2. Then we extracted the weight matrices of the associated classes and created the predicted CAM for each subject. The output CAM size is 22 x 13 x 12. Since, MBNv2 has 19 residual bottleneck layers. We added attention at the end of the 13th bottleneck. We used 3D MBNv2 with width multiplier of 0.5 and trained the model using SGD for 200 epochs with the initial learning rate of $8e-4$ and a batch size of 2.

The results are displayed in Table 5. The results convincingly show using CAAT on top of the 3D MBNv2 as the backbone results in significant improvements in performance. Each of the variants of CAAT with 3D MBNv2 backbone outperform 3D MBNv2 without using attention transfer for almost all metrics. The CAAT ensemble (3D MBNv2) outperforms 3D MBNv2 without attention by at least 0.03 for all metrics tested. These results show our attention transfer method is beneficial across CNN architectures.

When comparing results between the default CNN backbone and the 3D MBNv2 backbone the results show without attention transfer 3D MBNv2 performs worse than the baseline 3D CNN. However, after adding the attention transfer, 3D MBNv2 performs about the same as the default 3D CNN with attention transfer. For example CAAT ensemble versus CAAT ensemble (3D MBNv2) perform about the same, with each model winning in some metrics while losing in others. This shows attention transfer is able to raise the performance of poorer performing models to the same level as higher performing models.

We also noticed the 3D MBNv2 model converges much faster during training when using CAAT (converges at the 60th epoch) versus without CAAT (converges at the 110th epoch). We attribute this faster convergence to CAAT's ability to identify the most important features and thereby focus the training on the most important regions.

[9] https://github.com/zhangjun001/AD-Landmark-Prediction.

[10] https://github.com/okankop/Efficient-3DCNNs.

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. The numbers in parentheses report the variance value. Note that in the table Model TL refers to Transfer learning and CAAT by itself refers to CAAT AD. For example TL AD/CN + CAAT refers to Transfer learning AD/CN + CAAT AD.

Model	pMCI vs. sMCI				
	SEN	SPE	F1	ACC(%)	AUC
Baseline 3D CNN	0.71 (.005)	0.71 (.003)	0.71 (.003)	70.84 (.001)	0.758 (.003)
TL AD/CN	0.71 (.005)	0.71 (.003)	0.71 (.003)	71.35 (.003)	0.772 (.002)
TL CDR-SB	0.67 (.007)	0.75 (.003)	0.70 (.003)	71.00 (.003)	0.768 (.002)
TL ADAS	0.74 (.005)	0.67 (.007)	0.72 (.002)	72.06 (.001)	0.768 (.002)
TL ensemble	0.73 (.005)	0.72 (.004)	0.72 (.002)	72.18 (.002)	0.773 (.002)
6-Conv TL AD/CN	0.71 (.002)	**0.76** (.007)	0.73 (.002)	73.03 (.003)	0.777 (.002)
6-Conv TL CDR-SB	0.72 (.007)	0.71 (.003)	0.72 (.003)	71.85 (.002)	0.768 (.003)
6-Conv TL ADAS	0.70 (.005)	0.72 (.002)	0.71 (.003)	71.34 (.002)	0.773 (.003)
6-Conv TL ensemble	0.72 (.005)	0.75 (.004)	0.73 (.003)	73.37 (.003)	**0.782** (.003)
CAAT AD	0.70 (.004)	0.75 (.004)	0.72 (.001)	73.03 (.002)	0.779 (.002)
CAAT CDR-SB	**0.75** (.003)	0.71 (.006)	0.73 (.002)	72.70 (.002)	0.773 (.002)
CAAT ADAS	0.73(.006)	0.74 (.006)	0.73 (.003)	73.03 (.002)	0.777 (.003)
CAAT ensemble	**0.75** (.004)	0.75 (.006)	**0.75** (.002)	**74.61** (.002)	0.780 (.003)
TL AD/CN+CAAT	0.72 (.004)	0.74 (.007)	0.73 (.002)	73.03 (.002)	0.774 (.002)
TL CDR-SB+CAAT	0.72 (.012)	0.75 (.002)	0.73 (.004)	73.52 (.002)	0.772 (.002)
TL ADAS+CAAT	**0.75** (.002)	0.71(.003)	0.73 (.001)	72.86 (.001)	0.775 (.003)
TL+CAAT ensemble	**0.75** (.006)	0.73 (.007)	0.74 (.003)	74.04 (.002)	0.776 (.003)

Table 5
Experimental results comparing CAAT using MobileNetV2 backbone versus CAAT using the default backbone. Note CAAT AD (3D MBNv2) refers to using CAAT on top of the 3D MobileNetV2 backbone for both the source and target tasks. The source task is pretrained on AD versus CN classification. The best results for each evaluation metric is highlighted in bold font text. SEN, SPE, F1 and ACC refer to the sensitivity, specificity, F1 score and accuracy metrics respectively. The numbers in parentheses report the variance value.

Model	pMCI vs. sMCI				
	SEN	SPE	F1	ACC(%)	AUC
Baseline 3D CNN	0.71 (.005)	0.71 (.003)	0.71 (.003)	70.84 (.001)	0.758 (.003)
CAAT AD	0.70 (.004)	**0.75** (.004)	0.72 (.001)	73.03 (.002)	0.779 (.002)
CAAT CDR-SB	**0.75** (.003)	0.71 (.006)	0.73 (.002)	72.70 (.002)	0.773 (.002)
CAAT ADAS	0.73(.006)	0.74 (.006)	0.73 (.003)	73.03 (.002)	0.777 (.003)
CAAT ensemble	**0.75** (.004)	**0.75** (.006)	**0.75** (.002)	**74.61** (.002)	0.780 (.003)
3D MBNv2	0.70 (.008)	0.70 (.007)	0.69 (.002)	69.48 (.002)	0.749 (.004)
CAAT AD (3D MBNv2)	0.74 (.002)	0.70 (.000)	0.72 (.001)	71.84 (.001)	0.773 (.001)
CAAT ADAS (3D MBNv2)	0.68 (.005)	0.72 (.006)	0.70 (.002)	70.67 (.002)	0.765 (.003)
CAAT CDR-SB (3D MBNv2)	0.73 (.004)	**0.75** (.003)	0.74 (.002)	73.54 (.002)	**0.800** (.003)
CAAT ensemble (3D MBNv2)	0.73 (.004)	**0.75** (.004)	0.74 (.002)	73.86 (.002)	0.799 (.003)

4.6. Impact of transfer learning

We further conducted a series of experiments to investigate the impact of traditional transfer learning methods. The 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 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 and AUC where it only performs 0.01 and 0.002 worse for SPE and AUC respectively 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.

The results for the 6 layer CNN transfer learning using AD/NC as the source task performs slightly better than CAAT AD. However, the results for the other source tasks of CDR-SB and ADAS show that CAAT works better than the 6 layer transfer learning approach. This is likely the reason why the ensemble version of CAAT performs significantly better than the ensemble version of the 6 layer CNN transfer learning method.

CAAT ensemble outperforms the 6 layer CNN transfer learning ensemble for sensitivity, F1 score and accuracy, but gives similar performance for specificity and AUC. It is important to perform well for the sensitivity metric since higher sensitivity means CAAT ensemble is able to better catch pMCI cases earlier and therefore give doctors time to intervene earlier. CAAT ensemble is able to achieve higher sensitivity without sacrificing specificity and AUC compared to the 6 layer CNN transfer learning ensemble.

4.7. Hyper-parameter search for transfer learning

To ensure we provide a fair comparison of CAAT versus transfer learning, we conducted a thorough hyper-parameter search to train the best model using transfer learning. Specifically we tried different initial learning rates and freezing different early layers in the model. We tried the learning rate of 0.001, 0.0001, 0.0003, 0.0005, and 0.0008.

Table 6
Experimental results of hyper-parameter search for transfer learning. The results are for the 6-Conv model with pretrained weights obtained from classifying CN versus AD. (lr.0001) refers to using the learning rate of 0.0001, and (fix-l1) refers to freezing the first layer of the model, (fix-l1l2) refers to freezing the first and the second layer of the model. 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. The numbers in parentheses report the variance value.

Model	pMCI vs. sMCI				
	SEN	SPE	F1	ACC(%)	AUC
6-Conv TL (lr.0001)	0.73 (.014)	0.69 (.004)	0.71 (.005)	71.01 (.003)	0.767 (.003)
6-Conv TL (lr.0003)	**0.74** (.006)	0.72 (.003)	0.71 (.004)	71.34 (.003)	0.769 (.003)
6-Conv TL (lr.0005)	**0.74** (.006)	0.71 (.005)	**0.73** (.002)	72.56 (.002)	**0.783** (.003)
6-Conv TL (lr.0008)	0.71(.002)	**0.76** (.007)	**0.73** (.002)	**73.03** (.003)	0.777 (.002)
6-Conv TL (fix-l1)	0.68 (.004)	0.74 (.004)	0.70 (.003)	70.66 (.003)	0.778 (.002)
6-Conv TL (fix-l1l2)	0.70 (.011)	0.70 (.004)	0.70 (.004)	69.58 (.003)	0.773 (.003)

Table 7
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. The numbers in parentheses report the variance value.

Model	pMCI vs. sMCI				
	SEN	SPE	F1	ACC(%)	AUC
CAAT AD-att-Conv3	0.705 (.006)	0.715 (.005)	0.710 (.002)	71.01 (.002)	0.768 (.002)
CAAT AD-intra-att	**0.709** (.002)	0.715 (.004)	0.709 (.003)	70.68 (.003)	0.752 (.004)
CAAT AD-no signal	0.705 (.005)	0.722 (.005)	0.711 (.002)	71.34 (.001)	0.766 (.003)
CAAT AD	0.705 (.004)	**0.753** (.004)	**0.723** (.001)	**73.03** (.002)	**0.779** (.002)

The results for learning rate of 0.001 were omitted due to very poor performance. We also tried freezing the weight of the first layer and both the first and second layers. For this experiment we used the 6-Conv model with pretrained weights obtained from classifying CN versus AD.

The results of the hyper-parameter search for transfer learning are shown in Table 6. The results show that using a learning rate of 0.0008 has the best specificity, F1 score (equal best), and accuracy and a learning rate of 0.0005 gives the best sensitivity, F1 score (equal best), and AUC value. Hence in our experiments we used a default initial learning rate of 0.0008 for transfer learning. In terms of freezing layers, 6-Conv TL (fix-l1) and 6-Conv TL (fix-l1l2) shows the results for freezing the first layer, first and second layers, respectively. The results show freezing the tested combination of layers results in poorer performance compared to fine tuning the entire model. Hence in our experiments we perform transfer learning by fine tuning the entire model.

4.8. Ablation study

We performed an ablation study to gain insights into our CAAT. The results are reported in Table 7.

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 learned 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 calibrate 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 labelled 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. A limitation of our experimental setup is the use of 5 fold cross validation for both tuning and selecting the models. This may lead to an optimistically biased evaluation of the model performance. Hence an important extension to our experiments is to use nested cross validation to overcome the shortcoming with the standard 5 fold cross validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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? Mol. Cell. Proteom. 15 (2) (2016) 409–425, http://dx.doi.org/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, Am. J. Geriatr. Psychiatry 6 (2, Supplement 1) (1998) S64–S78, http://dx.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, http://dx.doi.org/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, Arch. Gen. Psychiatry 63 (8) (2006) 916–924, http://dx.doi.org/10.1001/archpsyc.63.8.916, arXiv:https://jamanetwork.com/journals/jamapsychiatry/articlepdf/668194/yoa60002.pdf.

[5] M.A. DeTure, D.W. Dickson, The neuropathological diagnosis of Alzheimer's disease, Mol. Neurodegeneration 14 (32) (2019) http://dx.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, Front. Aging Neurosci. 11 (2020) 369, http://dx.doi.org/10.3389/fnagi.2019.00369, URL https://www.frontiersin.org/article/10.3389/fnagi.2019.00369.

[7] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, The Alzheimer's Disease Neuroimaging Initiative, Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment, Front. Neurosci. 12 (2018) 777, http://dx.doi.org/10.3389/fnins.2018.00777, URL https://www.frontiersin.org/article/10.3389/fnins.2018.00777.

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

[9] 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, http://dx.doi.org/10.1109/CBMS.2018.00067.

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

[11] 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, 2019, CoRR abs/1904.07773, URL http://arxiv.org/abs/1904.07773, arXiv:1904.07773.

[12] 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, Neurobiol. Aging 59 (2017) 80–90, http://dx.doi.org/10.1016/j.neurobiolaging.2017.06.027, URL https://www.sciencedirect.com/science/article/pii/S0197458017302294.

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

[14] B. Fischl, FreeSurfer, Neuroimage 62 (2) (2012) 774–781.

[15] 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: Clin. 21 (2019) 101645, http://dx.doi.org/10.1016/j.nicl.2018.101645, URL https://www.sciencedirect.com/science/article/pii/S2213158218303930.

[16] M. Liu, J. Zhang, E. Adeli, D. Shen, Landmark-based deep multi-instance learning for brain disease diagnosis, Med. Image Anal. 43 (2018) 157–168.

[17] J. Zhang, Y. Gao, Y. Gao, B.C. Munsell, D. Shen, Detecting anatomical landmarks for fast Alzheimer's disease diagnosis, IEEE Trans. Med. Imaging 35 (12) (2016) 2524–2533.

[18] 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 Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 880–893, http://dx.doi.org/10.1109/TPAMI.2018.2889096.

[19] B.N. Patro, M. Lunayach, S. Patel, V.P. Namboodiri, U-CAM: Visual explanation using uncertainty based class activation maps, in: Proceedings of the IEEE/CVF International Conference on Computer Vision.

[20] S. Yang, Y. Kim, Y. Kim, C. Kim, Combinational class activation maps for weakly supervised object localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2941–2949.

[21] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[22] K.H. Sun, H. Huh, B.A. Tama, S.Y. Lee, J.H. Jung, S. Lee, Vision-based fault diagnostics using explainable deep learning with class activation maps, IEEE Access 8 (2020) 129169–129179.

[23] H.-G. Nguyen, A. Pica, J. Hrbacek, D.C. Weber, F. La Rosa, A. Schalenbourg, R. Sznitman, M.B. Cuadra, A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2019, pp. 370–379.

[24] C. Yang, A. Rangarajan, S. Ranka, Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification, in: AMIA Annual Symposium Proceedings, vol. 2018, American Medical Informatics Association, 2018, p. 1571.

[25] 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, http://dx.doi.org/10.1038/nrneurol.2009.215.

[26] M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, L. Chen, Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation, 2018, CoRR abs/1801.04381, URL http://arxiv.org/abs/1801.04381, arXiv:1801.04381.

[27] O. Köpüklü, N. Kose, A. Gunduz, G. Rigoll, Resource efficient 3d convolutional neural networks, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, IEEE, 2019, pp. 1910–1919.