Transfer Learning Methods for Alzheimer's Disease Diagnosis

Pedro Miguel Tavares Pereira
pedro.t.pereira@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

January 2021

Abstract

Alzheimer's disease (AD) is the most common cause of dementia among elderly people and one of the leading causes of death in developed countries, which is manifested by the loss of cognitive functioning and behavioral abilities, caused by progressive neuronal degeneration. Although there is still no practical diagnostic method available, a correct detection of the disease is crucial to slow down it's progression and cognitive decline. In recent years, Deep Learning methods gained popularity in AD detection, especially in dealing with brain scan images. Despite the success of these methods, the volume of medical images available is usually too small, which can easily lead to overfitting. At the same time, multimodal neuroimaging approaches have shown good results in AD diagnosis. In this work, a cross-modal Transfer Learning strategy was adopted using positron emission tomography (PET) and magnetic resonance imaging (MRI) brain scans: Deep Learning models based on Convolutional Neural Networks were pre-trained on one modality and fine-tuned using the other modality. The proposed approach obtained 86.4% accuracy for the classification task of AD vs normal controls (NC), showing improvement of around 2.5% of the classification system's accuracy with Transfer Learning, reducing overfitting, while taking advantage of the information provided by different neuroimaging modalities.

Keywords: Alzheimer's Disease; Transfer Learning; Convolutional Neural Network; Deep Learning; Medical Imaging; Multi-modality

1. Introduction

Alzheimer's Disease (AD) is the most common cause of dementia among people over the age of 65, affecting more than 5 million Americans [1]. It is defined as the loss of cognitive functioning and behavioral abilities, interfering with a person's daily life and activities, such as difficulty in communication, speaking or walking. Changes in the brain may begin a decade or more before memory and other cognitive problems appear. These changes are related to abnormal deposits of amyloid plaques and tau tangles throughout the brain, killing neuron cells and cutting connections with other neurons in an irreversible way, eventually leading to the person's death.

There are 3 main stages regarding AD: mild AD, moderate AD and severe AD. Preceding those stages, there is a phase called Mild Cognitive Impairment (MCI), in which people have more memory problems than normal for their age, but their symptoms do not interfere with their everyday lives. Some people with MCI may develop AD, but not all of them do. MCI patients that are likely to progress to AD are called MCI converters (MCI-C), while MCI patients that will not convert to AD are called MCI non-converters (MCI-NC). This rate of progression differs for each patient and each person may show different symptoms at each stage, which makes AD a challenging task for diagnosis and prognosis.

With the growth of popularity of machine learning methods and application of such strategies to AD diagnosis, it has been suggested a general adoption of computer-assisted methods for dementia diagnosis. Among these, it has been shown that Deep Learning methods obtain better results [21]. Deep learning is a specific sub field of machine learning, which consists on learning successive layers of increasingly meaningful representations, involving tens or even hundreds of successive layers of representations, all learned automatically, from exposure to training data, via neural networks [8].

The most popular deep learning model used for image analysis is the Convolutional Neural Network (CNN). When CNNs are trained on images, the first layers tend to learn generic features, such as edges, colors and textures, regardless of the cost function or dataset, as more deep layers tend to learn more abstract features related to particular dataset and task (specific).

Transfer Learning methods focus on solving one problem in the base domain and transferring the knowledge gained to a different but related problem, the target task. Usually, in Deep Learning, the method is based on training a base network and copying the first n layers to the first n layers of a target network. Some of the layers of the previous base task can then be fine-tuned to the new task or left frozen [27]. Fine-tuning works by unfreezing a few of the top layers of the frozen previously trained model and jointly training these layers with the top layers of the target network, slightly adjusting the more abstract representations of the model being reused to make them more relevant for the problem at hand. The use of a pre-trained CNN with adequate fine-tuning can outperform a CNN trained from scratch, using less amount of training data [24].

The identification of biomarkers for AD and it's combination with deep learning techniques that are able to identify patterns, features and hidden representations, contribute to the early detection of the disease and may accelerate the development of new therapies that can slow down the disease progression and cognitive decline, which can have huge impacts in patient's and caregivers' life quality.

In this work, a cross-modal Transfer Learning (TL) approach is investigated within a Deep Learning context, for AD diagnosis, using MRI and fluorodeoxyglucose positron emission tomography (FDG-PET) data, while studying the effects of fine-tuning on initial and deep layers. This approach is compared to other more common methods used to merge two types of neuroimaging data. Two types of CNN-based were compared and conclusions were taken based on the performance metrics of the applied methods and the visual representations of the features learned by the models developed, making possible to see into the "black box" of a deep learning model.

To date, from the information collected during this thesis, there are no studies that use cross-modal Transfer Learning for AD detection using PET and MRI modalities or using the deep learning networks that were implemented in this thesis, whereby this work might lead to a better understanding in this area.

2. State of the Art

In this section, different applications of Transfer Learning for AD detection within the last 10 years are summarized. In addition, several methods in which multimodal neuroimaging data has been combined are described.

2.1. Transfer Learning methods for AD detection

Transfer Learning methods have been proven to be robust even for very dissimilar domains, such as networks trained on a dataset containing natural images used with medical images [14, 26, 23, 13, 18, 19, 9], which is the most common application of TL in AD detection, but there are several other ways to apply Transfer Learning to Alzheimer's early detection problem, as detailed in this section.

[5] uses a Support Vector Machine (SVM) for classification of MCI-C vs MCI-NC patients, based on a related auxiliary domain, given by the task of classifying AD vs NC patients. The same authors extend their TL approach to use multiple auxiliary domains - Multi-Domain Transfer Learning [4, 6]. The classification was made by SVMs and obtained measures of performance for several tasks. Each task used the others as auxiliary domains, for example, in [4] the task of classifying MCI-C vs MCI-NC patients was based on the AD vs NC and MCI vs NC tasks, and [6] the target task MCI-C vs MCI-NC used also AD vs MCI in the auxiliary domain.

Filipovych & Davatzikos [12] had already been using, in 2011, AD vs NC domain to target the MCI-C vs MCI-NC classification, using a semi-supervised SVM to classify MCI subjects in the absence of certain diagnostic information for some patients in the ADNI database.

Because these methods used SVMs, there was a necessity of performing feature extraction manually. Regions of interest (ROIs) were labeled from each image and for each ROI the volume of gray-matter (GM) tissue was computed as a feature [4, 6, 12]. [5] also used average pixel intensity of each ROI for the PET images as features and three cerebrospinal fluid (CSF) biomarkers: CSF $A\beta_{42}$, CSF t-tau, CSF p-tau. [6] also used CSF $A\beta_{42}$, CSF t-tau, CSF p-tau as features.

Other methods used CNNs as learning algorithm, and therefore, no manual feature extraction was needed. [14] used MRI data from the OASIS dataset, using VGG16 and Inception V4 architectures, with pre-trained weights from ImageNet and fine-tuning. An accuracy of 96.25% was obtained for AD vs NC classification with the Inception architecture, also showing that this type of Transfer Learning achieves better results than only training the same model from scratch. Wu et al. [26] also compared the performance of GoogleNet and CaffeNet architectures using Transfer Learning from pre-trained ImageNet (and fine-tuning), obtaining accuracy measures of 87.78% for the three way classification of NC vs MCI-C vs MCI-NC for the CaffeNet architecture. Transfer Learning using pre-trained ResNet architetures and functional MRI scans from ADNI was used in [23] which achieved

an average accuracy of 97.92% in the multi-class classification of AD, NC, significant memory concern (SMC) and three MCI stages, including early MCI (EMCI), MCI, and late MCI (LMCI). The authors compared this architecture with the AlexNet architecture for Transfer Learning, concluding that the use of residual learning (ResNet) and Transfer Learning both improved the performance. In [13], a pre-trained ResNet model was fine-tuned from MRI slices, extracting slice-level features and a Long Short-Term Memory (LSTM) layer is then used to learn longitudinal-level features for each subject.

Lu et al. [18] also addressed this type of Transfer Learning for classification between individuals with brain pathology (AD among them) and NC. Using the AlexNet architecture, pre-trained on ImageNet, with fine-tuning, they achieved 100% accuracy for this task. [19] also fine-tuned a pre-trained AlexNet architecture to classify segmented GM, WM and CSF images and unsegmented MRI images from the OASIS dataset, in which the unsegmented images led to better results for the multi-class classification of the multiple stages of AD. Both of these studies modified the original AlexNet network for the target task, by replacing the last three layers with new layers with randomly initialized weights to learn class specific features in the target domain. Then, the weights of the remaining layers were adjusted during training jointly with the replaced layers in the case of [18], or remained fixed in [19] and only the replaced top layers had their weights updated.

Other studies analysed the current trend of using Transfer Learning from natural images to AD classification, specifically, using networks pre-trained on ImageNet, such as [9], which implemented several well-known 2D CNNs to extract discriminative features from MRI slices and a LSTM to incorporate spatial information across slices in the classification, showing better results when using the pre-trained SqueezeNet model.

A different Transfer Learning method was used by Hosseini-Asl et al. [15], in which a 3D CNN was built upon a stacked 3D convolutional autoencoder (CAE) network, which was pre-trained on CADDementia dataset in order to capture anatomical shape variations in structural MRI scans. The 3D CNN's layers were initialized by encoding the 3D-CAE weights and the upper layers were then fine-tuned for the specific task using data from the ADNI dataset (target domain), achieving 100% accuracy in AD vs MCI classification and more than 90% accuracy in the other evaluated tasks. Similarly, Payan et al. [21] used a 3D CNN for AD diagnosis based on pre-training by a 3D sparse autoencoder (SAE). This pre-training was performed by randomly selecting small 3D patches of MRI scans. The trained weights of the SAE are then used for pre-training of convolutional filters of 3D CNN. The fully connected layers of the 3D CNN are then fine-tuned. This method was improved by Vu et al. [25], to combine MRI and FDG-PET.

Focusing on the Hippocampal region and in two distinct imaging modalities: Diffusion Tensor Imaging (DTI), in particular Mean Diffusivity (MD) density maps derived from DTI and sMRI, [3] proposed a cross-modal Transfer Learning method for classification between AD, NC and MCI, in which a 2D CNN model is trained first on the sMRI dataset and then fine-tuned on the target MD dataset, with a limited amount of data, for each projection (Sagittal, Axial and Coronal). The results from each projection were fused using Majority Vote. This cross-modal Transfer Learning method showed a reduction of overfitting and improvement of learning performance, and encouraged a new perspective in Transfer Learning for Alzheimer's disease detection, in which each domain is represented by a different neuroimaging modality.

2.2. **Combination of neuroimaging modalities**

Apart Transfer Learning, there are several ways to fuse multimodal data, particularly neuroimaging data. Imaging data can be also combined with other available information such as cognitive measures or demographic information, and selecting the best modality combination and fusion method is a task that's been studied in-depth.

Multimodal classification was compared with the case of using only one biomarker by [28]. They combined MRI, PET and CSF biomarkers using a kernel combination method to train an SVM and obtaining an accuracy of 93.2% for the classification of AD vs NC and 76.4% for MCI vs NC. These results were better than using only one biomarker, emphasizing the benefits of having multimodal data. Similar methods based on SVMs that were mentioned in the previous section, reported better results when combining MRI and PET images and CSF biomarkers [5] and MRI and CSF biomarkers [6].

An early fusion method was used in [7], which combined two types of PET images, FDG-PET and ^{18}F-florbetapir (AV-45) PET, to train a 3D CNN for the AD vs NC task, exploring both glucose metabolism and amyloid deposit in the patient's brain at the same time. This network was then applied to predict between MCI-C and MCI-NC subjects, showing better results when both modalities were used simultaneously to train the model.

To predict conversion of MCI to AD, [16] used a multimodal gated recurrent unit (GRU) network, which integrated subject's demographic information, longitudinal CSF biomarkers, longitudinal cognitive performance and cross-sectional MRI images obtained from ADNI. This required two steps:

the training of a single GRU separately for each modality of data and merging the four networks into one. With the incorporation of several modalities into one prediction model, while using longitudinal data, the accuracy improved from 75% to 81% for the classification between MCI-C and MCI-NC. [10] also combined time series neuroimaging data from MRI and PET and subject's cognitive scores from 15 time steps and static background knowledge from the patients' first visits, (such as age, gender, CSF, symptoms, etc.) to predict disease progression and four cognitive scores at the time of progression in a multitask and multi-class deep learning framework which achieved 92.62% accuracy for the CN vs MCI-NC vs MCI-C vs AD task. Deep features extracted from each time series modality and fed into a separate stacked CNN-Bidirectional Long-short term memory (BiLSTM) pipeline and the learned representations are fused together with a set of dense layers. In a second fusing step, the common features from these modalities are fused with the baseline background data features and a final set of dense layers are used to learn task specific features.

Using a similar concatenation method, [22] designed a fusion model that obtained 92.34% accuracy for the AD vs NC classification problem using MRI and florbetapir PET images from ADNI. In this work, the authors built a 3D CNN for each modality with three convolutional layers and three fully connected layers. Then, to perform fusion, the output layer of both networks is replaced by a concatenation layer, which fuses the information from both modalities before the final classification is made. The improvements in performance due to fusion of both modalities indicates that the two modalities share complementary information useful in this task, although the authors revealed amyloid (AV-45) PET to be more discriminative in comparison to MRI in the first study that fused and compared these two modalities. A more common choice of modalities is MRI and FDG-PET which was fused in [25] and the authors reported improvements in comparison to the classification of each modality separately for the AD vs NC problem, reaching 91.14%. This approach fused the outputs of two 3D CNNs trained for each modality through a 3-layer fully connected layer neural network. This boosting of the performance is not only due to fusion, but also the pre-training of the CNN using a SAE trained on random 3D patches from the scans, similarly to [21]. [17] achieved an accuracy of 82.93% for discrimination between MCI-NC and MCI-NC subjects, by concatenating the representations learned from six Deep Neural Networks (DNNs), which corresponded to three different patch scales from FDG-PET and GM and using another DNN to fuse these representations.

Instead of using fully connected layers to share information between modalities, [11] achieved better results using a Bidirectional Recurrent Neural Network (BiRNN), which took as inputs the features extracted from a 3D CNN trained on PET images and GM density maps segmented from anatomical MRI images.

Apart from combining modalities, different views from brain scans (Sagittal, Axial and Coronal) from the same modality can be combined to achieved a global prediction score, which is done frequently using Majority Vote, as in [9] and [3].

3. Methods

3.1. Data

The data used in the experiments came from the ADNI database [2]. A detailed description on how the MRI and PET datasets were acquired can be found in the public ADNI website [2].

In this study, 1.5T Magnetic Resonance images and FDG-PET images were acquired from subjects evaluated during a 24 month period. Evaluations of their mental state and collection of brain scans were performed at a baseline month, and 12 and 24 months after the first evaluation. The number of images acquired in each evaluation period are shown in table 1 for AD and NC subjects for each neuroimaging modality.

It's worth noting that there are subjects who don't go through the complete 24 months of observations. There are also subjects whose images are present in only one of the modalities at a given time. In total there are 383 PET volumes from 133 subjects (58 AD and 75 NC) and 648 GM volumes from 316 subjects (144 AD and 172 NC) in the dataset. Considering that some subjects have PET and GM volumes, the total number of subjects evaluated is 354 from which 152 are classified with AD and 202 are NC. All the patients that were diagnosed with AD in month 0 remain with that prognostic during the whole 24 months and the same happens with NC subjects.

Table 1: Number of images for different evaluation periods and corresponding number of subjects for both modalities.

Follow up period	PET		GM	
	AD	NC	AD	NC
0 months	58	75	107	124
12 months	54	74	98	135
24 months	53	69	69	115
Total	165	218	274	313
Subjects	58	75	144	172

3.2. **Data Division**

The training and evaluation of the models was performed using 5 fold. Furthermore, a validation set was used to perform early stopping. The data split is performed according to each subject: In a dataset containing PET or GM images, 20% of the subjects are used for test and from the remaining 80%, 20% are used for validation and 60% for the training set, while assuring that different images from the same subjects are stored in the same partition. Subjects' images from the first evaluation period were assigned to the corresponding training or test sets. Images from months 12 and 24 were then assigned to the corresponding partition to avoid an "information leak".

An important safeguard was in assuring that there wouldn't be any PET images from a patient used in training in the base domain, that could be used in testing in the target domain in the Transfer Learning approach, or vice-versa. This was done by using images from subjects that appear on both modalities in the same partition in each fold. The other subjects that appear on only one modality could be used on either set. This division of data allows the use of the same images in the same sets for the several experiments conducted, so that the results of each method could be compared between each other.

3.3. **Pre-Processing**

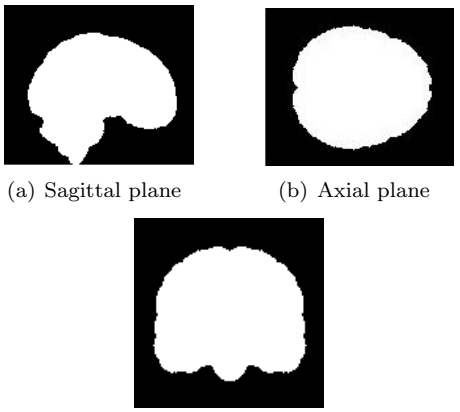
PET and MR images had already been subject to a series of preprocessing steps performed by the ADNI researchers [2]. Furthermore, the images retrieved from the ADNI database were warped into the MNI standard space as described in [20]. In this process, to the MR images was performed skull stripping, segmentation into GM and WM, producing gray and white-matter probability maps which were also smoothed with a Gaussian filter. In the same study, the resulting PET images were normalized using the Yakushev normalization procedure.

3.3.1 Crop

The images from the brain scans include the area surrounding the brain, which doesn't present any relevant information for the classification task. This area was cropped, resulting in a significant reduction in the size of the feature vectors. Only the area inside the brain was considered according to the MNI-152 template, represented by the area in white in figure 1. The volumes' dimension after the cropping was 104x122x98.

3.3.2 Feature Normalization

Besides the normalization process that had been previously applied, as described in [20], the voxel intensities in each volume slice were mapped to the range [-1, 1].



(a) Sagittal plane (b) Axial plane

(c) Coronal Plane

Figure 1: Representations of the MNI brain mask in the sagittal, axial and coronal sections of the brain.

In each fold, the maximum of the modulus of the volumes from PET and GM in the training set was computed. Then all the volumes in the of GM and PET training, validation and test set were divided by this value, which was computed separately for each modality and after the images from months 12 and 24 were added to the training set.

3.4. **Deep Learning Network**

Tables 2 and 3 present the final CNN-LSTM and 3D CNN architectures used for the AD vs NC classification problem. Before these configurations were achieved, others were tested, where parameters such as the number of layers, number of units in each layer, filter's shape, dropout or batch normalization were tested.

Each network was trained using all the available PET and GM data, displayed in table 1, including longitudinal data from the subjects, which achieved better results than using only data from the first follow up month, or only one image per subject.

3.4.1 CNN-LSTM network

As seen in table 2, each Convolution Block is made by a TimeDistributed 2D CNN and a max-pooling layer. The output of the three convolution blocks is flattened, goes into the LSTM and then fed into a densely connected classifier network, with softmax activation, corresponding the output to the binary classification of AD vs NC. The Keras TimeDistributed wrapper allows the distribution of CNN layers across every slice of the 3D input, applying the same instance of the convolution layer to every input slice. The LSTM layer is then used to detect interslice relationships.

Table 2: Architecture of the CNN-LSTM network.

Layer Type	Parameters	Filters/Units
Convolutional	Stride-2, ReLU	3x3x32
Max Pooling	2x2, Stride-2	-
Convolutional	ReLU	5x5x64
Max Pooling	2x2, Stride-2	-
Convolutional	ReLU	5x5x128
Max Pooling	2x2, Stride-2	-
Flatten	-	-
Dropout	50% Dropout	-
LSTM	Tanh, 50% Drop.	128
Dense	Softmax	2

3.4.2 3D CNN network

The 3D CNN model is also made of three convolution blocks, as shown in table 3: The first two convolution blocks are made by a 3D CNN layer, a 3D max-pooling layer and two batch normalization layers. The third convolution block only has a batch normalization layer. The output of the convolution blocks is then flattened and fed into a fully connected classifier network, which results in the classification of AD vs NC.

Table 3: Architecture of the 3D CNN network.

Layer Type	Parameters	Filters/Units
Convolutional	ReLU	3x3x3x8
Batch Norm.	-	-
Max Pooling	2x2x2, Stride-2	-
Batch Norm	-	-
Convolutional	ReLU	3x3x3x16
Batch Norm.	-	-
Max Pooling	2x2x2, Stride-2	-
Batch Norm.	-	-
Convolutional	ReLU	3x3x3x32
Batch Norm.	-	-
Max Pooling	2x2x2, Stride-2	-
Flatten	-	-
Dense	-	64
Batch Norm.	-	-
Dense	-	64
Batch Norm.	-	-
Dense	Softmax	2

3.5. Transfer Learning

Two methods were tested regarding Transfer Learning: Either pre-training a deep learning network with GM data as the base domain and fine-tuning using the PET dataset as the target domain (TL GM-PET), or instead pre-training a network with PET data and fine-tuning using the GM dataset (TL PET-GM).

Independently of using TL GM-PET or PET-GM, the number of fine-tuned layers and number of top densely connected layers replaced was also tested. For the CNN-LSTM, several configurations were tested: replacing the top four layers (Flatten, Dropout, LSTM and Dense) with the same layers having new randomly initialized weights, replacing the last Dense layer, or not replacing any layer. For the 3D CNN, the last six layers were replaced (Flatten, Dense, BatchNormalization, Dense, BatchNormalization and Dense), or the last Dense layer or no layer was replaced. Choosing one of these configurations, any number of the remaining layers can then be fine-tuned.

To perform these experiments, models trained separately with PET and GM data were used. Then the last layers of the pre-trained network were replaced by new layers with randomly initialized weights. Some number of the remaining layers were chosen to be fine-tuned and the others remained frozen (their weights could not be adjusted) and the unfrozen layers were trained jointly with the added part. The images used for training and testing were the same images used to train and test each network from scratch.

3.6. Joint training of both modalities in the same deep learning network

The input feature space was composed by volumes of both modalities, which were taken as input together for training a single model (CNN-LSTM or 3D CNN), thus concatenating modalities in an early fusion mode.

3.7. Concatenation of two deep learning models trained on separate modalities

To apply this method it was necessary to use the same number of input samples from each modality. From the complete set of images available, in each time window, only subjects with both modalities available were used, in order to concatenate information relative to the same person. The available number of images for each modality is shown in table 4.

Table 4: Number of input images for different evaluation periods for the concatenation model.

Follow up period	PET		GM	
	AD	NC	AD	NC
0 months	50	45	50	45
12 months	43	n 36	43	36
24 months	27	20	n 27	20
Total	120	101	120	101
Subjects	50	45	50	45

Several topologies were tested regarding the number of layers and units in each layer, from which the final model was chosen. For the concatenation of two CNN-LSTM networks, the resulting network concatenates the outputs of the last layer of both CNN-LSTM networks and adds two Fully Connected layers on top of the concatenation, to perform the final decision. The concatenation of two 3D CNN networks is made by concatenating the feature representations of the penultimate layer and adding three dense layers to perform the final decision. Only the added layers were trained, while the remaining layers remained frozen.

3.8. **Implementation Details**

The deep learning methods are implemented using Keras with a TensorFlow 2.3.0 backend. All the experiments were performed in Google Colab. The networks were trained using the Adam optimizer with a initial learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The accuracy metric was evaluated during training and binary cross-entropy was chosen as the loss function.

A weighted training strategy was applied, in which samples belonging to the class with the majority of data in the training set were given a weight equal to one and samples from the other class are given a weight equal to $\frac{N}{M}$, where N is the number of samples from the class with most data and N is the number of samples from the class with less samples. Heatmaps of intermediate activations were visualized using Jupyter notebooks.

4. Results and Discussion

In this section, the mean and standard deviation (SD) of the accuracy (ACC) and F1 Score results obtained using the architectures described in the previous section are detailed. Tables 5 and 6 present the results obtained for both Deep Learning architectures trained from scratch and the best Transfer Learning results obtained for TL GM-PET and TL PET-GM. The TL approach results are compared with the joint training approach in tables 7 and 8 and with the concatenation approach in tables 9 and 10.

4.1. **Transfer Learning vs training from scratch**

The results from tables 5 and 6, show that the CNN-LSTM model outperforms the 3D CNN model for the PET modality and the 3D CNN performs better for GM data. These results can be explained taking into account the number of training samples in each modality and the complexity and number of parameters of each model. The 3D CNN model requires 8 times more training parameters than the CNN-LSTM, which, on the other hand, has less ability to capture inter-slice information than the 3D CNN. Regarding the CNN-LSTM model, a notable performance is achieved when the model is trained on the PET modality, even having less training data than GM, thus being PET the most discriminative modality against AD changes in the brain. In the 3D CNN model, the best results are obtained by training the model using GM dataset, which has more available images.

Table 5: CNN-LSTM results using GM and PET modalities individually and using TL PET-GM and TL GM-PET.

Modalities	ACC (SD)	F1 Score (SD)
GM	0.679 (0.111)	0.695 (0.102)
PET	0.829 (0.034)	0.784 (0.023)
TL PET-GM	0.816 (0.060)	0.815 (0.065)
TL GM-PET	**0.861** (0.033)	**0.831** (0.025)

Table 6: 3D CNN results using GM and PET modalities individually and using TL PET-GM and TL GM-PET.

Modalities	ACC (SD)	F1 Score (SD)
GM	0.839 (0.031)	0.849 (0.034)
PET	0.761 (0.058)	0.717 (0.072)
TL PET-GM	**0.864** (0.023)	**0.870** (0.031)
TL GM-PET	0.851 (0.038)	0.814 (0.041)

Transfer Learning shows performance improvements compared with training each model from scratch. Using the CNN-LSTM model, fine-tuning with PET data was more effective than using GM data. These results were obtained by replacing the last layer (softmax) and unfreezing all the remaining layers for GM-PET or unfreezing 7 of the remaining layers (corresponding to the densely connected classifier and the second convolution block) for PET-GM.

Regarding the 3D CNN model, the results show fine-tuning with GM data to be more effective. The best results were obtained by replacing the whole densely connected classifier and unfreezing the last convolution block for PET-GM, while for GM-PET there were only fine-tuned 9 layers (the densely connected classifier and the top convolution block).

4.2. **Transfer Learning vs joint training both modalities in the same network**

The results relative to the early fusion model are expressed according to the nature of the testing set: whether it contained only PET or GM images or contained images from both modalities.

From the results shown in tables 7 and 8, it can be observed that the models obtained results close

Table 7: CNN-LSTM results for the early fusion and Transfer Learning methods.

Modalities	ACC (SD)	F1 Score (SD)
GM	0.778 (0.042)	0.764 (0.060)
PET	0.808 (0.051)	0.762 (0.053)
PET and GM	0.821 (0.031)	0.811 (0.020)
TL PET-GM	0.816 (0.060)	0.815 (0.065)
TL GM-PET	**0.861** (0.033)	**0.831** (0.025)

Table 8: 3D CNN results for the early fusion and Transfer Learning methods.

Modalities	ACC (SD)	F1 Score (SD)
GM	0.840 (0.028)	0.838 (0.042)
PET	0.832 (0.083)	0.806 (0.084)
PET and GM	0.852 (0.020)	0.848 (0.026)
TL PET-GM	**0.864** (0.023)	**0.870** (0.031)
TL GM-PET	0.851 (0.038)	0.814 (0.041)

to those of the Transfer Learning approaches. The fact that all the available data from both modalities was used in the training of the models, reduces the need for Transfer Learning, by reducing overfitting during training, although increasing in the average training time. In Transfer Learning, the training in the target domain doesn't require as many training samples and has less trainable parameters, which makes training faster, while enabling the final model to be well suited for the target task.

4.3. Transfer Learning vs model concatenation

For a better comparison, the CNN-LSTM and 3D CNN models trained on separate modalities were tested using the same number of images from each modality as this concatenated model, meaning that the GM and PET images used in the testing set of the concatenation model were used for testing the models trained from scratch.

Table 9: Results for the concatenation of two CNN-LSTM networks and Transfer Learning methods. (1) Results for the CNN-LSTM model evaluated on the reduced dataset.

Modalities	ACC (SD)	F1 Score (SD)
PET and GM	0.762 (0.075)	0.670 (0.120)
TL PET-GM	0.816 (0.060)	0.815 (0.065)
TL GM-PET	**0.861** (0.033)	**0.831** (0.025)
GM (1)	0.610 (0.128)	0.496 (0.162)
PET (1)	0.730 (0.075)	0.631 (0.106)

Table 10: Results for the concatenation of two 3D CNN networks and Transfer Learning methods. (1) Results for the 3D CNN model evaluated on the reduced dataset.

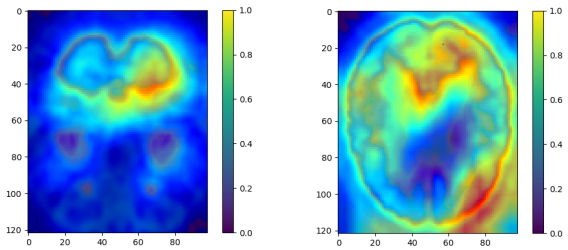
Modalities	ACC (SD)	F1 Score (SD)
PET and GM	0.755 (0.049)	0.681 (0.088)
TL PET-GM	**0.864** (0.023)	**0.870** (0.031)
TL GM-PET	0.851 (0.038)	0.814 (0.041)
GM (1)	0.716 (0.051)	0.644 (0.092)
PET (1)	0.729 (0.163)	0.695 (0.180)

From the analysis of tables 9 and 10, it can be concluded that the concatenation models achieve a better performance than the models trained from scratch, when comparing with data from the same subjects in the test set. A downside of this approach is that the model only accepts the same number of images from both modalities, corresponding to the same number of subjects, discarding relevant data in the datasets that can be used to improve the models, which doesn't happen in TL.

Contrarily to the early fusion method, this concatenation approach allows the use of two specific deep learning models for each neuroimaging modality, but the information shared among modalities is reduced. Furthermore, the concatenation of two deep learning models doesn't solve the problem of lack of training data: although the combination of learned features from both modalities can improve classification by exploring commonalities and differences between both types of data, the pre-trained networks can suffer from overfitting, which affects the performance of the concatenation model.

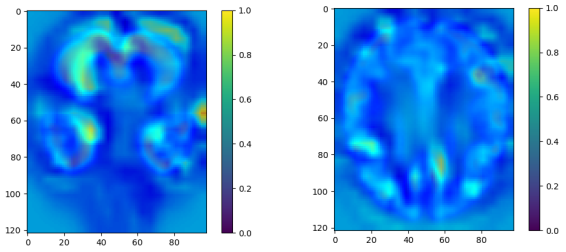
4.4. Visual comparison with biological changes in the brain

The output filters of the last convolutional layer of the CNN-LSTM model with TL GM-PET and the 3D CNN model with TL PET-GM are displayed in figures 2 and 3 for two selected slices in the axial plane.



(a) (b)

Figure 2: Heatmaps of intermediate activations for TL GM-PET using the CNN-LSTM model.



(a) (b)

Figure 3: Heatmaps of intermediate activations for TL PET-GM using the 3D CNN model.

In the heatmaps relative to the PET modality, high intensity values are located in the temporal region, as shown in figure 2 (a). In figure 2 (b), high intensity values are located in the parietal and posterior cingulate areas, which correspond to relevant ROIs. Regarding the GM images, high intensity values are observed in the temporal region in figure 3 (a) and the superior anterior cingulate region in figure 3 (b).

5. Conclusions and Future Work

Regarding the achievements of this work, it can be considered that the main objectives were accomplished, since the proposed Transfer Learning method achieves classification accuracies of 86.4% using a 3D CNN fine-tuned on GM data and 86.1% using a CNN-LSTM network fine-tuned on PET data, for the classification between AD vs NC subjects, outperforming the other studied approaches. Despite the fact that the 3D CNN generally outperformed the CNN-LSTM network, this model could still achieve an accuracy of 86.1% with Transfer Learning, which are satisfactory results considering this to be the first approach that applies this particular CNN-LSTM model in AD classification.

Possible approaches to be tested in the future could involve the combination of the Transfer Learning and concatenation approaches studied. This cross-modal Transfer Learning approach could also be extended to the classification of the several AD stages, or to the prediction of MCI conversion to AD. To do this, it would be interesting to compare between other biomarkers, such as AV-45 PET. Training a model on the AD vs NC task and fine-tuning the weights to classify also between MCI subjects can be an interesting way to apply Transfer Learning to the classification of several AD stages. Furthermore, better results could be achieved using ROI-based features, instead of voxel based features or slice-based features. Specially in the case of the CNN-LSTM model, improvements could be made using BiLSTM, and also exploring slices in the sagittal and coronal views besides the axial view. These views could be combined in order to improve performance by capturing complementary information.

Acknowledgements

I would like to thank Prof. Margarida Silveira, for the guidance provided during this thesis, and my friends and family for the encouragement and support.

References

[1] 2020 alzheimer's Disease Facts and Figures. https://www.alz.org/alzheimers-dementia/facts-figures, Accessed: December 2020.

[2] Alzheimer's Disease Neuroimaging Initiative - ADNI. http://adni.loni.usc.edu/, Accessed: December 2020.

[3] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and 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)*, pages 345–350. IEEE, 2018.

[4] B. Cheng, M. Liu, D. Shen, Z. Li, D. Zhang, A. D. N. Initiative, et al. Multi-domain transfer learning for early diagnosis of Alzheimer's Disease. *Neuroinformatics*, 15(2):115–132, 2017.

[5] B. Cheng, M. Liu, D. Zhang, B. C. Munsell, and D. Shen. Domain transfer learning for MCI conversion prediction. *IEEE Transactions on Biomedical Engineering*, 62(7):1805–1817, 2015.

[6] B. Cheng, M. Liu, D. Zhang, D. Shen, A. D. N. Initiative, et al. Robust multi-label transfer feature learning for early diagnosis of Alzheimer's Disease. *Brain imaging and behavior*, 13(1):138–153, 2019.

[7] H. Choi, K. H. Jin, A. D. N. Initiative, et al. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural brain research*, 344:103–109, 2018.

[8] F. Chollet. *Deep Learning with Python*. Manning Publications Co, 2018.

[9] A. Ebrahimi-Ghahnavieh, S. Luo, and R. Chiong. Transfer Learning for Alzheimer's Disease Detection on MRI Images. In *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 133–138. IEEE, 2019.

[10] S. El-Sappagh, T. Abuhmed, S. R. Islam, and K. S. Kwak. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing*, 412:197–215, 2020.

[11] C. Feng, A. Elazab, P. Yang, T. Wang, B. Lei, and X. Xiao. 3D convolutional neural network and stacked bidirectional recurrent neural network for Alzheimer's disease diagnosis. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 138–146. Springer, 2018.

[12] R. Filipovych, C. Davatzikos, A. D. N. Initiative, et al. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). *NeuroImage*, 55(3):1109–1119, 2011.

[13] L. Gao, H. Pan, F. Liu, X. Xie, Z. Zhang, J. Han, A. D. N. Initiative, et al. Brain disease diagnosis using deep learning features from longitudinal MR images. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 327–339. Springer, 2018.

[14] M. Hon and N. M. Khan. Towards Alzheimer's Disease classification through transfer learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1166–1169. IEEE, 2017.

[15] E. Hosseini-Asl, G. Gimel'farb, and A. El-Baz. Alzheimer's Disease diagnostics by a deeply supervised adaptable 3D convolutional network. *arXiv preprint arXiv:1607.00556*, 2016.

[16] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1):1–12, 2019.

[17] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Scientific reports*, 8(1):1–13, 2018.

[18] S. Lu, Z. Lu, and Y.-D. Zhang. Pathological brain detection based on AlexNet and transfer learning. *Journal of computational science*, 30:41–47, 2019.

[19] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, and O.-y. Song. Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans. *Sensors*, 19(11):2645, 2019.

[20] P. M. Morgado, M. Silveira, A. s Disease Neuroimaging Initiative, et al. Minimal neighborhood redundancy maximal relevance: Application to the diagnosis of Alzheimer s disease. *Neurocomputing*, 155:295–308, 2015.

[21] A. Payan and G. Montana. Predicting Alzheimer's Disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*, 2015.

[22] A. Punjabi, A. Martersteck, Y. Wang, T. B. Parrish, A. K. Katsaggelos, and A. D. N. Initiative. Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks. *Plos one*, 14(12):e0225759, 2019.

[23] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood. A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks. *Journal of Medical Systems*, 44(2):37, 2020.

[24] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[25] T. D. Vu, H.-J. Yang, V. Q. Nguyen, A.-R. Oh, and M.-S. Kim. Multimodal learning using convolution neural network and Sparse Autoencoder. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 309–312. IEEE, 2017.

[26] C. Wu, S. Guo, Y. Hong, B. Xiao, Y. Wu, Q. Zhang, A. D. N. Initiative, et al. Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks. *Quantitative Imaging in Medicine and Surgery*, 8(10):992, 2018.

[27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[28] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.