Generating Attention Maps from Eye-gaze for the Diagnosis of Alzheimer's Disease

Carlos Antunes CARLOS.VALDES.ANTUNES@TECNICO.ULISBOA.PT
Margarida Silveira MSILVEIRA@ISR.TECNICO.ULISBOA.PT
Instituto Superior Técnico, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

Abstract

Convolutional neural networks (CNNs) are currently the best computational methods for the diagnosis of Alzheimer's disease (AD) from neuroimaging. CNNs are able to automatically learn a hierarchy of spatial features, but they are not optimized to incorporate domain knowledge.

In this work we study the generation of attention maps based on a human expert gaze of the brain scans (domain knowledge) to guide the deep model to focus on the more relevant regions for AD diagnosis. Two strategies to generate the maps from eye-gaze were investigated; the use of average class maps and supervising a network to generate the attention maps. These approaches were compared with masking (hard attention) with regions of interest (ROI) and CNNs with traditional attention mechanisms.

For our experiments, we used positron emission tomography (PET) scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. For the task of normal control (NC) vs Alzheimer's (AD), the best performing model was with insertion of regions of interest (ROI), which achieved 95.6% accuracy, 0.4% higher than the baseline CNN.

Keywords: Deep learning; Alzheimer's disease; Convolutional neural network; Attention mechanism; Eye tracking; Computer-aided diagnosis.

1. Introduction

Alzheimer's Disease (AD) is a chronic brain disorder that accounts for 60% to 80% of dementia cases worldwide (Gaugler et al., 2020) and affects predominantly the elderly.

Symptoms include forgetfulness, difficulty reasoning and mood changes like apathy, wandering, agitation and aggression. The brain presents atrophy due to death of neurons and lower metabolic activity. While there is still no cure for AD, its early detection is crucial, as an effective management of the disease may help prevent the progression to more severe stages. Clinical diagnosis is made by collecting medical and family history, asking relatives about changes in behaviour and conducting mental cognitive tests. Brain imaging, like magnetic resonance imaging (MRI) scans or positron emission tomography (PET) scans has also been recognized as a powerful biomarker, however their interpretation is difficult thus computer-aided diagnosis (CAD) has been requested by clinicians to amplify their diagnostic accuracy (Gauthier et al., 2021).

Currently, the best performing algorithms for AD classification from neuroimaging are convolutional neural networks (CNNs). In these networks, the features are automatically extracted rather than handcrafted, however it is not easy to incorporate medical knowledge.

A recent survey on deep models for medical image analysis concluded that integrating domain knowledge improved the performance of the networks in almost all tasks (Xie et al.,

2021). As an example, it states that the attention mechanism is a powerful technique to incorporate domain knowledge of radiologists, because the information about where medical doctors focus helped deep learning models yield better results (Li et al., 2020) (Mitsuhara et al., 2021) (Fang et al., 2019) (Cui et al., 2020) (Xie et al., 2022) (Zhang et al., 2021a). Inspired by these results, in this work we investigate whether the generation of attention maps based on eye-tracking data (physician gaze) can improve the performance of AD diagnosis, by directing the classification model to focus on important regions (determined by domain knowledge). The maps that are obtained are multiplied with CNN feature maps, thus certain locations are highlighted while others are attenuated. Two approaches were investigated for attention map generation. In the first approach, average maps are computed from the doctor's gaze maps. In the second approach, the eye-gaze data is used to supervise a CNN trained to generate attention maps. The inferred maps, like in the first approach, are then multiplied with the feature maps of the CNN that does classification, and whose parameters are trained with the class labels only. Finally, this CNN was also trained with regions of interest (ROI) to compare intuitive domain knowledge with pre-defined relevant regions for classification.

Therefore, the main contributions of this work are:

- Introduction of domain knowledge from eye-gaze data from an expert physician into a state-of-the-art CNN model to perform AD classification.
- Training a deep multiscale network and a U-Net with physician eye-gaze data to predict attention maps.

2. Related Work

2.1. AD detection models

In the last decade, there have been substantial developments in machine learning classification models for AD detection. CNNs are very effective for AD classification problems and ResNets are by far the most popular type of CNN applied (Korolev et al., 2017; Jin et al., 2019; Ullanat et al., 2021; Liang and Gu, 2021; Zhang et al., 2021c,d; Sun et al., 2021). Nonetheless, some authors used AlexNet (Zheng et al., 2018), Inception (Ding et al., 2019) and VGG (Lee et al., 2021; Turkan and Tek, 2021) or applied an ensemble of methods (Liu et al., 2018). Most studies train models with magnetic resonance imaging (MRI) scans (Korolev et al., 2017; Jin et al., 2019; Ullanat et al., 2021; Liang and Gu, 2021; Zhang et al., 2021c,d; Sun et al., 2021; Turkan and Tek, 2021; Zhang et al., 2021b; Basaia et al., 2019), although still a considerable number use other biomarkers, like PET scans (Zheng et al., 2018; Ding et al., 2019; Lee et al., 2021; Liu et al., 2018; Singh et al., 2017; Lu et al., 2018; Jo et al., 2020; Choi et al., 2020), largely from the Alzheimer's Disease Neuroimaging Initiative (ADNI) clinical datasets.

A recent in-depth study (Khojaste-Sarakhsi et al., 2022) about deep learning applications in AD diagnosis research analyzed about 100 published papers since 2019. Besides identifying many trending technologies, the study recognized the importance of the attention mechanism (AM) and suggested it should be further explored. The idea behind the attention mechanism comes from human visual attention, which illustrates that human vision typically does not scan the entire scene at once, but rather focuses on selective parts

of the whole visual field sequentially, according to the person's needs. The AM therefore can be interpreted as weighted values that represent the importance of each specific part of the image for classification. In CNN models there can be many types of attention, like spatial attention, channel attention, self-attention and layer attention, all of which were employed in the analyzed papers. As for examples of models, Dan J. et al. (Jin et al., 2019) trained a 3D ResNet with one layer of spatial attention (convolution and rectified linear unit (ReLU)), which led to an increase of 2% in accuracy. Ullanant et al. (Ullanat et al., 2021) inserted a residual attention block (Wang et al., 2017) to a vanilla ResNet. Liang S et al. (Liang and Gu, 2021) used one layer of channel attention per stage. Each attention block has global max-pooling for each channel, a convolution with 1x1 kernel, ReLU and dense layers. Zhang Y. et al. (Zhang et al., 2021d) created an attention mechanism inspired by the Squeeze-and-Excitation block (Hu et al., 2018) (channel attention) and got an increase of about 2% in accuracy. Regarding the location of the attention mechanism in the network, most studies place it in the middle of the network or throughout every residual block. However, one author (Zheng et al., 2022) concluded the AM was better placed at the head of the network.

All of the experiments mentioned that used AM were made with MRI scans. No studies that applied attention mechanisms to PET scans were found. Nonetheless, PET scans were chosen for this work, because they can show brain alterations before anatomical changes are observed in MRI scans, which is important for early diagnosis (Mayblyum et al., 2021).

2.2. Supervised attention

Since there were no studies on the effect of supervising attention mechanisms with human gaze (domain knowledge) for Alzheimer's disease, we looked at works in other fields.

Yu et al. (2017) showed that spatial attention guided by human eye-tracking data can, in fact, enhance performance, in their case, the performance of generating short text information about brief video clips. They created an AM block that predicts a gaze map per frame of the input video. The inclusion of this AM block improved the results by 3.2% for one language metric.

Li et al. (2020) proposed a CNN for glaucoma detection with an attention mechanism supervised by human attention, called AG-CNN. The human-generated attention maps were used to train the attention prediction subnet of their AG-CNN, which is comprised of a CNN with concatenated features of different layers passed through a deconvolution block at the end. Li's model has considerably better performance than other state-of-the-art methods in his field and increased accuracy by 3.4% when compared to the same model without attention.

Ma et al. (2022) proposed a vision transformer for the diagnosis of breast diseases. They infuse the human expert's prior knowledge to guide the network to focus on the patches with potential pathology. This design leads to higher performance (increased accuracy by almost 1% compared to a standard ResNet50). Moreover, the EG-ViT only introduces the mask operation and an additional residual connection to a vanilla vision transformer. This model has the limitation that it needs to be pre-trained with hundreds of millions of data samples in order to show better results than CNN. This is especially troublesome for 3D images.

Sheng Wang et al. (Wang et al., 2022) designed a supervised network to assess knee X-ray images for osteoarthritis. This model, called GA-Net, is composed of a ResNet classification network and the supervised attention consistency block. This last component is a CAM visualization/localization module (Zhou et al., 2016). Comparing the ResNet18 with ResNet18+Gaze, the accuracy increased by 2% to 62.8%.

3. Data

ADNI is a landmark partnership with the purpose of creating a longitudinal study intended to collect biomarkers of AD. From this database, we retrieved fludeoxyglucose (FDG) PET scans, which show the glucose metabolism in the brain, from participants with baseline and 6, 12 and 24-month follow-ups. 1393 scans from 406 subjects were used, 314 were from AD subjects, 714 were from mild cognitive impairment (MCI) subjects and 365 were normal controls (NC). Table 1 presents demographic and clinical information of the study subjects. All FDG-PET had been normalized, averaged and co-registered by ADNI, and were also further normalized to the [0,1] range.

Table 1: Clinical profile of the subjects in three categories (AD, MCI, NC) categories. Age and MMSE are average values with the standard deviation in parenthesis. MMSE refers to the Mini-Mental State Exam, a mental cognitive status assessment that evaluates memory, thinking and simple problem-solving abilities, where the maximum (best) score is 30.

Group	AD	MCI	NC	All
No. of subjects	95	207	104	406
Age	76.6 (7.1)	76.0 (7.3)	77.0 (4.8)	76.4 (6.7)
Sex (% M)	59.9	65.8	63.8	64.1
MMSE	21.1 (4.1)	26.7 (2.8)	29.1 (1.2)	26.1 (2.9)

Additionally, several PET scan images in this dataset have been complemented with records of the gaze of a medical doctor while performing a diagnosis, thus collecting areas of interest (domain knowledge). This was performed by Bicacro et al. (Bicacro et al., 2012), using a Tobii™ device. For their study, the gaze (a total of 4261 fixation points) for scans of 177 subjects (59 of each category - AD, MCI, NC) was collected. Table 2 presents the proportion of each type of scan within the overall dataset. It is noteworthy that the amount of scans with fixations is only 12.6% of the total scans available. Even though these eye-gaze data have been applied before in (Bicacro et al., 2012) and (Morgado et al., 2012), it was never employed in deep learning models. They were used for selecting and extracting features that were then fed to a support vector machine classifier.

For each scan, the eye-tracker provides discrete fixation points. However, the physician does not look at a particular pixel, but instead looks at a region centered in the fixation point and symmetrically spread out by the visual angle. Therefore, we convolve the fixation map $f(x)$ (image with the points where the doctor focused) with an isotropic bi-dimensional

Table 2: PET scans in various categories. Several different scans correspond to the same patients in different periods of the ADNI's longitudinal study, therefore there are more scans than subjects.

Group	AD	MCI	NC	All
No. of scans	314	714	365	1393
Proportion of total (%)	22.5	51.2	26.3	100
Proportion of scans with fixations (%)	4.2	4.2	4.2	12.6

Gaussian function $G_\sigma(x)$, creating an attention map $S(x)$, like in Figure 1 ((*a*), (*b*), (*c*)) (image where the regions people's eyes focus are highlighted). The circular region is modeled by the isotropic Gaussian filter and the visual angle by the standard deviation ($\sigma = 3$). Some examples of the resulting maps are shown in Figure 1, where average maps are also shown ((*d*), (*e*), (*f*)), given the variability in attention maps.

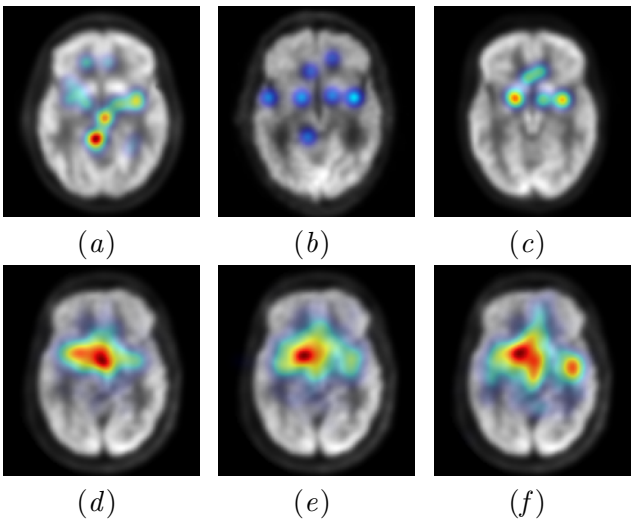
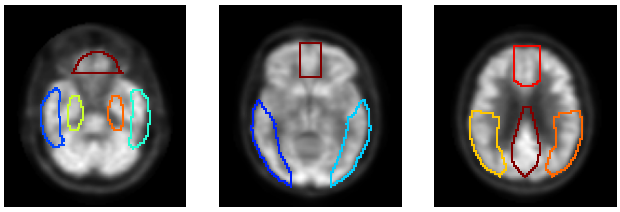


Figure 1: Examples of axial cut 25. The first row shows attention maps obtained by Gaussian filtering of the fixation points for three random patients, with NC (*a*), MCI (*b*) and AD (*c*). The second row shows average attention maps for NC (*d*), MCI (*e*) and AD (*f*).

The same expert physician has manually identified 12 regions of interest (ROI), as displayed in Figure 2. These regions include the lateral and mesial temporal, inferior frontal gyrus/orbitofrontal, inferior and superior anterior cingulate, dorsolateral parietal, posterior cingulate, and precuneus. These anatomical regions of the brain are considered by the doctor to be the most relevant for the task of AD diagnosis. If we compare the regions of interest with the regions where the doctor looked at, we discover that only 36.2% of fixations

fall inside the ROI. This might be concerning since it seems there is little coherence between the regions identified by the doctor and the regions where he focuses his gaze.



(a) (b) (c)

Figure 2: Examples of three axial slices with regions of interest (ROI) defined by the expert physician. (a) Red - Inferior frontal gyrus/Orbitofrontal; Dark and light blue - Lateral temporal; Light green and orange - Mesial temporal; (b) Dark and light blue - Lateral temporal; Red - Inferior and superiro anterior cingulate ; (c) Light red - Inferior and superior anterior cingulate ; Yellow and orange - Dorsolateral parietal; Dark red - Posterior cingulate and precuneus. Some slices do not contain any anatomical ROI.

4. Method

In this section, the different models studied are detailed. First, we present the two models investigated for attention mechanism supervision, then we present our approaches that use constant attention maps, either based on average eye-gaze data or from ROIs. Finally, we present our baselines which include a standard ResNet18 and the ResNet18 with attention mechanisms (either CBAM or Residual Attention).

4.1. Supervised attention mechanism

In this method, the model is composed of two sub-networks. The first network is used to predict the attention maps, and is supervised by the doctor's fixation maps. The second network is a standard ResNet18, where the created attention mechanism maps are inserted. Two alternatives for generating the attention maps from the doctors' eye-gaze were investigated. The first alternative is the deep multiscale network (Figure 3), which is similar to the glaucoma paper's (Li et al., 2020) attention prediction subnet, but adapted for 3D images and with resizing performed with average pooling and upsampling instead of bilinear interpolation. The encoder portion is a typical CNN, where the input passes through several residual blocks to extract hierarchical features. The decoder portion takes features from distinct basic blocks, normalizes them to the same dimensions, and concatenates them to perform convolutions four times, before applying convolution transpose twice.

The second alternative is a U-Net (Figure 4), which is also an encoder-decoder network. The encoder part performs feature extraction and learns abstract representations of the input image with convolutions. Here, the spatial dimensions decrease with max pooling operations. Furthermore, the network has two skip connections between the encoder and

decoder part, that concatenates two arrays, to be used in the next decoder stage. This helps to provide additional information to the decoder and assists in the flow of the gradient while backpropagating, since it is a shortcut. The decoder section takes the representations to generate the mask. It increases the size through upsampling.

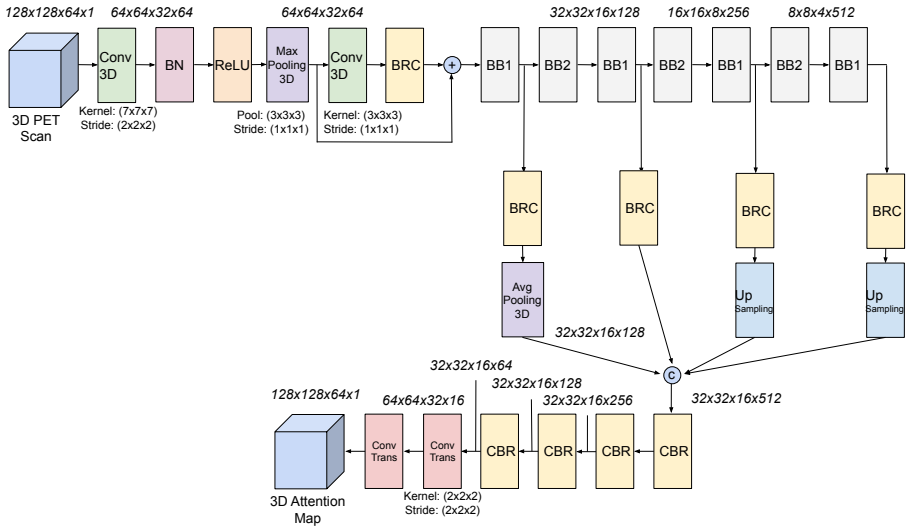


Figure 3: Representation of deep multiscale network that was chosen to learn the attention maps. BRC means batch normalization, ReLU and convolution layers.

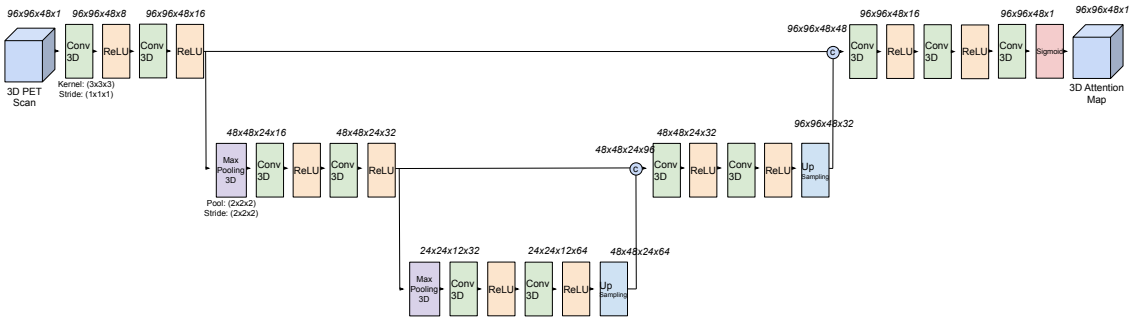


Figure 4: Representation of a 3D U-Net, the second network investigated for AM generation with eye-gaze supervision.

4.2. Constant average maps and ROI

In this approach, the attention maps are not created by layers with learned weights. Instead, the doctor's constant average attention map (based on the eye-tracking data) and the ROI maps (hard attention) are introduced into the network, without learning. These maps are inserted in the ResNet18 in the same place as the CBAM module.

4.3. Baseline CNNs

The simplest baseline is a vanilla 3D ResNet18. This is an appropriate model since residual networks are considered state-of-the-art and have been widely applied for AD classification. In fact, 38% of the 74 papers that used CNNs for AD diagnosis analyzed by Khojaste-Sarakhsi et al. used ResNets (Khojaste-Sarakhsi et al., 2022). Although this network does not include attention, we can visualize the regions of the input scans that the model considers more important with guided back-propagation (Springenberg et al., 2014) or Grad-CAM (Selvaraju et al., 2017).

Two additional baselines were tested, which integrated attention mechanisms into the ResNet, but that do not incorporate domain knowledge. One attention mechanism is CBAM (Woo et al., 2018), a commonly used attention module that can be integrated into any CNN.

CBAM sequentially infers attention maps along two separate dimensions, channel and spatial, which are multiplied by the input of the respective layer creating a refined feature map. For this study, CBAM was adapted for three dimensions, the same as the scans. To better understand the importance of the spatial attention component, the experiments were also done with the spatial attention sub-module only. The CBAM block was inserted in three different locations (one per trial): at the start of the network before any operation, in the middle basic block, and throughout the basic blocks of the ResNet.

Another attention mechanism tested is residual attention (Wang et al., 2017). This is another type of spatial and channel attention. It uses a bottom-up top-down structure to learn the mask. It collects global information and later guides input features in each position.

4.4. Experimental setup

The baseline CNN, the ResNets with CBAM and residual attention and the networks with constant maps/ROI were trained with categorical cross-entropy as the loss function, which was minimized with stochastic gradient descent optimizer for a maximum of 50 epochs. The learning rate was 1×10^{-2}. Train and testing were done using stratified 5-fold cross-validation. Since we have multiple scans of the same subject at different times, the subjects, and not the images, were separated into five folds. This methodology guarantees that brain scans from the same subject are not present in different sets, thus avoiding data leakage. About 15% of the available samples for training in each fold were used for validation. The model of the epoch with the lowest validation loss was selected as the best model to be tested. The supervised attention mechanism networks (deep multiscale network and U-Net) were trained like the aforementioned models but with Dice coefficient as loss. All models were created with the keras/Tensorflow package on Google Colab notebooks. The main components can be found in this link: https://tinyurl.com/GitHubPaperCode. The classification tasks performed were NC vs AD and NC vs MCI vs AD.

5. Results and discussion

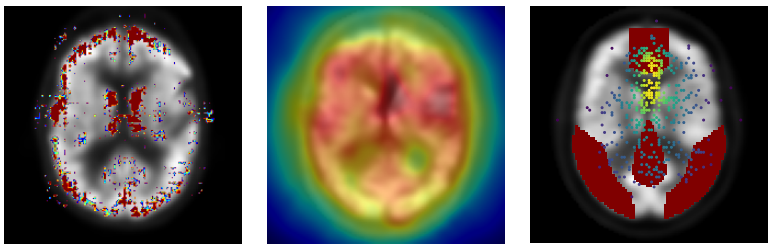
The results (accuracy, sensitivity, specificity and F_1-score) for the task NC vs AD and NC vs MCI vs AD are displayed in Table 3 and Table 4, respectively. All the models include a ResNet18. The tables only show the results for the best location of the attention mechanism (start, middle or throughout the network), as specified in the 'AM Location' column. The statistical significance of the differences between the results of each AM strategy and the baseline Resnet were evaluated with paired t-tests.

Table 3: Results of 5 fold cross validation for the task NC vs AD. Format: Mean (standard deviation), best result in bold. The lower section consists of models with domain knowledge, while the upper section does not. All the models are composed of a ResNet18 and the AM module specified in the first column.

Models	**AM Location**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**F_1-score (%)**
Standard ResNet18	-	95.2 (1.7)	95.0 (2.4)	95.3 (2.4)	94.8 (1.9)
CBAM	middle	94.9 (2.0)	94.7 (2.7)	95.3 (2.7)	94.6 (2.4)
CBAM spatial module	middle	95.0 (1.7)	94.8 (3.1)	95.2 (3.4)	94.6 (2.4)
Residual attention	throughout	95.5 (2.2)	94.7 (3.8)	**96.2 (2.3)**	94.8 (2.4)
Constant average map	middle	94.8 (1.5)	93.7 (3.4)	95.9 (2.4)	94.4 (1.5)
ROI	start	**95.6 (2.6)**	**95.1 (2.5)**	96.1 (2.8)	**95.2 (2.7)**
Deep multiscale network	start	94.0 (1.9)	92.9 (4.0)	94.9 (2.4)	93.4 (2.8)
U-Net	middle	95.2 (2.1)	94.7 (4.0)	95.6 (2.2)	94.6 (2.1)

For NC vs AD, the model with the highest accuracy was ResNet18 with ROI inserted in the start, achieving 95.6% accuracy. This was a 0.4% rise compared to the standard ResNet18, which is statistically significant (p-value$<$0.05), and the best performing model with domain knowledge.

Figure 5 displays a brain scan overlapped with heatmaps generated by guided backpropagation (*a*) and Grad-CAM (*b*) techniques of the standard ResNet18, as well as a scan with fixation points and ROI (*c*) for comparison. The red areas mean these regions are more important for the classification task. The most important regions for the guided backpropagation mode are slightly different than the ones activated by the Grad-CAM method, except for the center of the brain, which has some red regions for both types of images. The Grad-CAM maps are more similar to the doctor fixations than to the ROI. Nonetheless, from these types of images, no indisputable pattern stands out as a determinate location of the disease.



(a) (b) (c)

Figure 5: Examples of localization maps using Guided Backpropagation (a) and Grad-CAM (b) techniques for the standard ResNet18 for the task NC vs AD. Each of the maps presented is an average of the generated guided backpropagation and Grad-CAM output for all AD scans available. These maps highlight the regions of the brain image more activated in the network to make the prediction. (c) ROI and fixation points are displayed for comparison.

Examples of the generated attention maps are presented in Figure 6. We computed the Pearson correlation between these maps and the original fixation maps (results not shown) and concluded that the deep multiscale net created maps more similar to the original than the U-Net. Despite this, the U-net obtained slightly better performance and was the best method that incorporated the doctor's attention. Nevertheless, it was not able to obtain better performance than the baselines (p-values$<$0.05). Some reasons can be hypothesized: the eye-gaze dataset was too small, specially for deep learning which needs a lot of data; the methods of incorporating the eye-gaze were not the most suitable (other approaches were suggested, for example, a supervised CAM module (Wang et al., 2022) or a vision transformer with domain data (Ma et al., 2022)); the assumption that the doctor relies only on the intensity of the voxels to make decisions may be very simplistic, perhaps the doctor is comparing different regions' average intensity, performing basic computations or the mental process of information is different according to the region being analyzed.

For the task NC vs MCI vs AD, the best performing model is the ResNet18 with a constant average duration map in the middle, with 87.4% accuracy (+0.3% than standard ResNet18 and p-value$<$0.05). This means a different conclusion than for the task NC vs AD, for which the best performing model was with ROI. Therefore, perhaps the ROIs are optimized for AD regions and do not take into account MCI, while the eye-gaze was retrieved when the doctor was performing a classification task that included MCI (NC vs MCI vs AD), thus the constant average maps include this information.

The accuracy results of incorporating the CBAM spatial model and residual attention were not statistically different from those in the baseline ResNet for the binary task, but were statistically significant for the ternary task.

Figure 7 shows the accuracy of our models (in green) juxtaposed with the state-of-the-art networks for better comparison (in gray and blue). This figure shows that our deep models outperformed many of the studies found in the literature. Yet, these comparisons need to be taken lightly because different models were trained, with different biomarkers

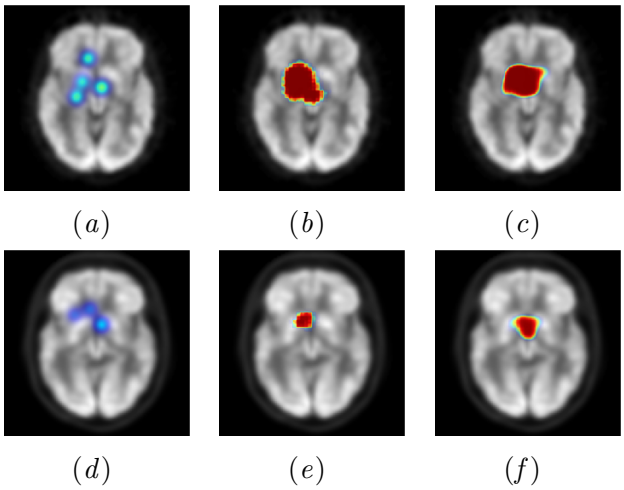


Figure 6: Examples of generated attention maps for axial cut 25 (first row - AD; second row - MCI). The doctor fixation maps are on the left, the deep multiscale network generated attention maps are in the middle and the U-Net maps on the right.

Table 4: Results of 5 fold cross validation for the task NC vs MCI vs AD. Format: Mean (standard deviation), best result in bold. All the models are composed of a ResNet18 and the AM module specified in the first column.

Models	**AM Location**	**ACC (%)**	F_1**-score (%)**
Standard ResNet18	-	87.1 (1.6)	86.7 (1.8)
CBAM	middle	85.9 (1.6)	85.3 (1.5)
CBAM spatial module	middle	86.9 (2.0)	86.5 (2.1)
Residual attention	throughout	85.5 (0.9)	84.5 (3.8)
Constant average map	middle	**87.4 (1.5)**	**86.9 (1.1)**
ROI	start	86.2 (1.4)	85.9 (1.7)
Deep multiscale network	start	86.6 (2.3)	86.3 (2.0)
U-Net	middle	86.0 (2.3)	85.5 (2.1)

and with a different number of scans. The figure also highlights that incorporating domain knowledge helped increase accuracy with ROI for the binary task and constant average maps for the multiclass task.

Our methods also performed better than most expert physicians in NC vs AD classification, who correctly predict 85.7% of scans on average (Klöppel et al., 2008).

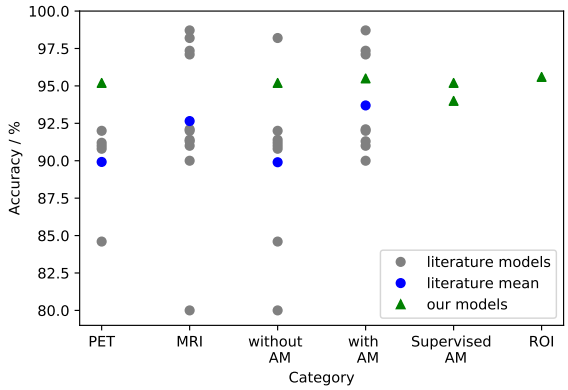


Figure 7: Comparison with state-of-the-art, for the task NC vs AD. Accuracy of our models (in green) contrasted with the models reviewed (in gray) and with the average value of the state-of-the-art models (in blue). 'PET' denotes models trained with PET scans only; 'MRI' denotes models trained with MRI scans only; 'without AM' denotes models without attention mechanism; 'with AM' denotes models that use attention mechanism; 'Supervised AM' refers to our models with supervised attention mechanism (there are no other state-of-the-art models in this category). Finally, on the right, the model with ROI, which was the best performing model.

6. Conclusion

In this work we investigated methods to integrate physician attention patterns obtained from eye-tracking data into CNNs for Alzheimer's Disease diagnosis. We explored the use of average gaze-maps and the supervision of a CNN to predict attention maps. We also compared these approaches with the use of ROI hard attention maps.

Our methods performed better than most CAD systems for AD working with FDG-PET images found in the literature. The ResNet18 with the ROI yielded the best results for NC vs AD, with an accuracy of 95.6% and the ResNet18 with constant average maps (Gaussian filtered eye gaze) achieved 87.4% for NC vs MCI vs AD task. These outcomes motivate further work like the creation of a bigger dataset, with more gaze data, following other approaches of introducing domain knowledge, like the visual transformer (Ma et al., 2022) or a CAM module (Wang et al., 2022) and extracting more information from the data besides just the voxel intensity of the "looked at" regions.

Acknowledgments

This work was supported by LARSyS - FCT Project UIDB/50009/2020.

The PET scans and subjects' data used in the preparation of this article were obtained from the ADNI database (https://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 the analysis or writing of this report.

The source of the highly experienced medical input (eye gaze and ROI) was Dr. Durval Campos Costa, a nuclear medicine expert, from the Champalimaud Foundation. While the method of acquisition and treatment of the eye-tracking data was performed by Eduardo Bicacro at Instituto Superior Técnico.

References

Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, and Massimo Filippi. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, 21:101645, 2019. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2018.101645. URL https://www.sciencedirect.com/science/article/pii/S2213158218303930.

Eduardo Bicacro, Margarida Silveira, Jorge S. Marques, and Durval C. Costa. 3D brain image-based diagnosis of Alzheimer's disease: Bringing medical vision into feature selection. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 134–137, 2012. doi: 10.1109/ISBI.2012.6235502.

Hongyoon Choi, Yu Kyeong Kim, Eun Jin Yoon, Jee-Young Lee, Dong Soo Lee, and Alzheimer's Disease Neuroimaging Initiative. Cognitive signature of brain FDG-PET based on deep learning: domain transfer from Alzheimer's disease to Parkinson's disease. *Eur. J. Nucl. Med. Mol. Imaging*, 47(2):403–412, February 2020.

Hui Cui, Yiyue Xu, Wanlong Li, Linlin Wang, and Henry Duh. Collaborative Learning of Cross-Channel Clinical Attention for Radiotherapy-Related Esophageal Fistula Prediction from CT. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I*, page 212–220, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-59709-2. doi: 10.1007/978-3-030-59710-8_21. URL https://doi.org/10.1007/978-3-030-59710-8_21.

Yiming Ding, Jae Ho Sohn, Michael G Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W Jenkins, Dmytro Lituiev, Timothy P Copeland, Mariam S Aboian, Carina Mari Aparici, Spencer C Behr, Robert R Flavell, Shih-Ying Huang, Kelly A Zalocusky, Lorenzo Nardo, Youngho Seo, Randall A Hawkins, Miguel Hernandez Pampaloni, Dexter Hadley, and Benjamin L Franc. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology*, 290(2):456–464, February 2019.

Leyuan Fang, Chong Wang, Shutao Li, Hossein Rabbani, Xiangdong Chen, and Zhimin Liu. Attention to Lesion: Lesion-Aware Convolutional Neural Network for Retinal Optical

Coherence Tomography Image Classification. *IEEE Transactions on Medical Imaging*, 38 (8):1959–1970, 2019. doi: 10.1109/TMI.2019.2898414.

Joseph Gaugler, Bryan James, Tricia Johnson, Allison Marin, and Jennifer Weuve. 2020 Alzheimer's Disease facts and figures. *Alzheimer's and Dementia*, 16:391–460, 2020. ISSN 15525279. doi: 10.1002/alz.12068.

Serge Gauthier, Pedro Rosa-Neto, José A. Morais, and Claire Webster. World Alzheimer Report 2021: Journey through the diagnosis of dementia. Technical report, Alzheimer's Disease International, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Dan Jin, Jian Xu, Kun Zhao, Fangzhou Hu, Zhengyi Yang, Bing Liu, Tianzi Jiang, and Yong Liu. Attention-based 3D Convolutional Network for Alzheimer's Disease Diagnosis and Biomarkers Exploration. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1047–1051, 2019. doi: 10.1109/ISBI.2019.8759455.

Taeho Jo, , Kwangsik Nho, Shannon L. Risacher, and Andrew J. Saykin. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. *BMC Bioinformatics*, 21(S21), December 2020. doi: 10.1186/s12859-020-03848-0. URL https://doi.org/10.1186/s12859-020-03848-0.

M. Khojaste-Sarakhsi, Seyedhamidreza Shahabi Haghighi, S.M.T. Fatemi Ghomi, and Elena Marchiori. Deep learning for Alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, 130:102332, 2022. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2022.102332. URL https://www.sciencedirect.com/science/article/pii/S0933365722000975.

Stefan Klöppel, Cynthia M. Stonnington, Josephine Barnes, Frederick Chen, Carlton Chu, Catriona D. Good, Irina Mader, L. Anne Mitchell, Ameet C. Patel, Catherine C. Roberts, Nick C. Fox, Jr Jack, Clifford R., John Ashburner, and Richard S. J. Frackowiak. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. *Brain*, 131(11):2969–2974, 10 2008. ISSN 0006-8950. doi: 10.1093/brain/awn239. URL https://doi.org/10.1093/brain/awn239.

Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3D brain MRI classification. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 835–838, 2017. doi: 10.1109/ISBI.2017.7950647.

Seung-Yeon Lee, Hyeon Kang, Jong-Hun Jeong, and Do-young Kang. Performance evaluation in [18F]Florbetaben brain PET images classification using 3D Convolutional Neural Network. *PLOS ONE*, 16(10):1–16, 10 2021. doi: 10.1371/journal.pone.0258214. URL https://doi.org/10.1371/journal.pone.0258214.

Liu Li, Mai Xu, Hanruo Liu, Yang Li, Xiaofei Wang, Lai Jiang, Zulin Wang, Xiang Fan, and Ningli Wang. A Large-Scale Database and a CNN Model for Attention-Based Glaucoma Detection. *IEEE Transactions on Medical Imaging*, 39(2):413–424, 2020. doi: 10.1109/TMI.2019.2927226.

Shuang Liang and Yu Gu. Computer-Aided Diagnosis of Alzheimer's Disease through Weak Supervision Deep Learning Framework with Attention Mechanism. *Sensors*, 21(1), 2021. ISSN 1424-8220. doi: 10.3390/s21010220. URL https://www.mdpi.com/1424-8220/21/1/220.

Manhua Liu, Danni Cheng, and Weiwu Yan. Classification of Alzheimer's Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET Images. *Frontiers in Neuroinformatics*, 12, 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00035. URL https://www.frontiersin.org/article/10.3389/fninf.2018.00035.

Donghuan Lu, , Karteek Popuri, Gavin Weiguang Ding, Rakesh Balachandar, and Mirza Faisal 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), April 2018. doi: 10.1038/s41598-018-22871-z. URL https://doi.org/10.1038/s41598-018-22871-z.

Chong Ma, Lin Zhao, Yuzhong Chen, Lu Zhang, Zhenxiang Xiao, Haixing Dai, David Liu, Zihao Wu, Zhengliang Liu, Sheng Wang, Jiaxing Gao, Changhe Li, Xi Jiang, Tuo Zhang, Qian Wang, Dinggang Shen, Dajiang Zhu, and Tianming Liu. Eye-gaze-guided Vision Transformer for Rectifying Shortcut Learning, 2022. URL https://arxiv.org/abs/2205.12466.

Danielle V Mayblyum, J Alex Becker, Heidi I L Jacobs, Rachel F Buckley, Aaron P Schultz, Jorge Sepulcre, Justin S Sanchez, Zoe B Rubinstein, Samantha R Katz, Kirsten A Moody, Patrizia Vannini, Kathryn V Papp, Dorene M Rentz, Julie C Price, Reisa A Sperling, Keith A Johnson, and Bernard J Hanseeuw. Comparing PET and MRI biomarkers predicting cognitive decline in preclinical Alzheimer's disease. *Neurology*, 96(24):e2933–e2943, May 2021.

Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding Human Knowledge in Deep Neural Network via Attention Map. In *VISIGRAPP*, 2021.

Pedro Morgado, Margarida C. da Silveira, and Jorge S. Marques. Automated Diagnosis of Alzheimer's Disease using PET Images: A study of alternative procedures for feature extraction and selection. Master's thesis, Instituto Superior Técnico, 9 2012.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Shibani Singh, Anant Srivastava, Liang Mi, Kewei Chen, Yalin Wang, Richard J Caselli, Dhruman Goradia, and Eric M Reiman. Deep-learning-based classification of FDG-PET

data for Alzheimer's disease categories. In Jorge Brieva, Juan David García, Natasha Lepore, and Eduardo Romero, editors, *13th International Conference on Medical Information Processing and Analysis*. SPIE, November 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, 2014. URL https://arxiv.org/abs/1412.6806.

Haijing Sun, Anna Wang, Wenhui Wang, and Chen Liu. An improved deep residual network prediction model for the early diagnosis of Alzheimer's disease. *Sensors*, 21, 6 2021. ISSN 14248220. doi: 10.3390/s21124182.

Yasemin Turkan and F. Boray Tek. Convolutional Attention Network for MRI-based Alzheimer's Disease Classification and its Interpretability Analysis. *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 1–6, 9 2021. doi: 10.1109/UBMK52708.2021.9558882. URL https://ieeexplore.ieee.org/document/9558882/.

Varun Ullanat, Vinay Balamurali, and Ananya Rao. A Novel Residual 3-D Convolutional Network for Alzheimer's disease diagnosis based on raw MRI scans. In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 82–87, 2021. doi: 10.1109/IECBES48179.2021.9398800.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual Attention Network for Image Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017. doi: 10.1109/CVPR.2017.683.

Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow My Eye: Using Gaze to Supervise Computer-Aided Diagnosis. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022. doi: 10.1109/TMI.2022.3146973.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–19, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Xiao-Zheng Xie, Jian-Wei Niu, Xue-Feng Liu, Qing-Feng Li, Yong Wang, Jie Han, and Shaojie Tang. DG-CNN: Introducing Margin Information into Convolutional Neural Networks for Breast Cancer Diagnosis in Ultrasound Images. *Journal of Computer Science and Technology*, 37(2):277, 2022. doi: 10.1007/s11390-020-0192-0. URL https://jcst.ict.ac.cn/EN/abstract/article_2863.shtml.

Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.101985. URL https://www.sciencedirect.com/science/article/pii/S1361841521000311.

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising Neural Attention Models for Video Captioning by Human Gaze Data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6119–6127, 2017. doi: 10.1109/CVPR.2017.648.

Boyan Zhang, Zhiyong Wang, Junbin Gao, Chantal Rutjes, Kaitlin Nufer, Dacheng Tao, David Dagan Feng, and Scott W. Menzies. Short-Term Lesion Change Detection for Melanoma Screening With Novel Siamese Neural Network. *IEEE Transactions on Medical Imaging*, 40(3):840–851, 2021a. doi: 10.1109/TMI.2020.3037761.

Jie Zhang, Bowen Zheng, Ang Gao, Xin Feng, Dong Liang, and Xiaojing Long. A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magnetic Resonance Imaging*, 78:119–126, 5 2021b. ISSN 18735894. doi: 10.1016/j.mri.2021.02.001.

Xin Zhang, Liangxiu Han, Wenyong Zhu, Liang Sun, and Daoqiang Zhang. An Explainable 3D Residual Self-Attention Deep Neural Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis using Structural MRI. *IEEE Journal of Biomedical and Health Informatics*, 2021c. ISSN 2168-2208. doi: 10.1109/jbhi.2021.3066832. URL http://dx.doi.org/10.1109/JBHI.2021.3066832.

Yanteng Zhang, Qizhi Teng, Yuyang Liu, Yan Liu, and Xiaohai He. Diagnosis of Alzheimer's Disease Based on Regional Attention with sMRI Gray Matter Slices. *Journal of Neuroscience Methods*, page 109376, 2021d. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2021.109376. URL https://www.sciencedirect.com/science/article/pii/S0165027021003113.

Chuanchuan Zheng, Yong Xia, Yuanyuan Chen, Xiaoxia Yin, and Yanchun Zhang. Early Diagnosis of Alzheimer's Disease by Ensemble Deep Learning Using FDG-PET. In *Lecture Notes in Computer Science*, pages 614–622. Springer International Publishing, 2018. doi: 10.1007/978-3-030-02698-1_53. URL https://doi.org/10.1007/978-3-030-02698-1_53.

Menghua Zheng, Jiayu Xu, Yinjie Shen, Chunwei Tian, Jian Li, Lunke Fei, Ming Zong, and Xiaoyang Liu. Attention-based CNNs for Image Classification: A Survey. *Journal of Physics: Conference Series*, 2171(1):012068, jan 2022. doi: 10.1088/1742-6596/2171/1/012068. URL https://doi.org/10.1088/1742-6596/2171/1/012068.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. doi: 10.1109/CVPR.2016.319.