Curriculum learning for early Alzheimer's Disease diagnosis

Catarina Gracias[1] and Margarida Silveira[2]

Abstract— **The early and asymptomatic stages of Alzheimer's Disease (AD), such as mild cognitive impairment (MCI), are hard to classify, even by experienced physicians. Deep learning approaches, such as convolutional neural networks (CNNs), have been shown to help, achieving similar or even better results. Although these methods have the advantage that features are automatically extracted from images rather than handcrafted, they do not allow for incorporating medical knowledge. In this paper we propose curriculum learning (CL) strategies for CNNs designed to diagnose healthy subjects, MCI and AD, as a way to incorporate medical knowledge to boost the performance of the networks for early AD diagnosis. CL is a training strategy of the networks that tries to mimic the way humans, in this case doctors, learn. Several CL strategies were implemented and compared to commonly used baseline methods. The results show that they improve the performance, particularly that of MCI.**

Clinical relevance— **This work shows that the use of CL strategies improve the diagnosis of AD, particularly at an early stage.**

I. INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and one of the leading causes of death in developed countries, since there is not a cure available yet [1]. In the early and asymptomatic stages of AD, patients are classified as having mild cognitive impairment (MCI). The clinical research developed towards finding therapeutics and a cure for AD highly depends on the ability to diagnose patients accurately at this early stage of the disease, when it is still possible to delay the onset of AD. AD diagnosis is performed by medical doctors, who have access to patient's information: medical images, genetic data and cognitive tests, such as Mini Mental State Examination (MMSE) and Clinica Dementia Ratio (CDR). However, MCI stages are not easily identified solely by following these traditional diagnostic approaches. Consequently, AD research benefits from the use of deep learning methods to make faster, earlier and more accurate diagnoses [2], [3]. Currently, Convolutional Neural Networks (CNNs), which allow features being automatically extracted rather than handcrafted, have already been successful in AD diagnosis through the classification of medical images [3]. Nevertheless, these recent approaches still have some drawbacks, such as not being optimized to incorporate medical knowledge.

In this paper, as a way to overcome this bottleneck, we propose and evaluate novel curriculum learning (CL) strategies, which take medical knowledge into account to more accurately diagnose early AD. Our CL strategies incorporate medical knowledge from scores of patients cognitive tests and regions of interest (ROI) that doctors focus on when diagnosing AD. This strategies are the first ones to use patient's cognitive tests, which are consistently available as they are performed regularly, to build a learning curriculum.

II. BACKGROUND AND RELATED WORK

CL is a strategy of training machine learning models by mimicking the way humans learn. In this strategy, a curriculum is designed, which defines the order in which the data are presented to the model: the model is first trained with easier data (or tasks) and gradually more complex data (or tasks) are introduced, instead of being randomly presented [4]. Usually, the curriculum is predefined (manual strategies). However, since defining a good curriculum manually is not an easy task, some strategies rely on learning the curriculum from the data, simultaneously with network training (automatic strategies).

CL has recently been shown to improve the performance of CNNs for several medical image classification tasks [5], [6], [7]. Most approaches use a manual curriculum. For instance, Tang et al. [5] built a curriculum by categorizing the severity of patient injuries according to X-ray reports. By using it, they improved thoracic disease diagnosis from X-rays (AUC increased 3.19%). Haarburger et al. [6] used manually selected lesion-patch images for pre training the model and then fine tuned it with the whole MRI images, improving the AUC for breast cancer diagnosis by 27%. Automatic CL strategies have also been proposed. For example, Maicas et al. [7] proposed a meta learning approach for breast screening classification from DCE-MRI, which outperformed baseline approaches (AUC improved from 86% to 90%). Despite the recent success of CL strategies for medical image classification, they have still not been applied to networks for AD diagnosis.

III. METHODOLOGY

A. Data pre processing

The data used in the implemented strategies were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. FDG-PET images of 406 subjects at baseline and at 12 and 24 month follow-ups were used, labeled as Normal Control (NC), MCI or AD. The clinical profile of the groups studied is presented in Table I. All FDG-PET

[1] Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal catarina.gracias@hotmail.com

[2] Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal msilveira@isr.tecnico.ulisboa.pt

TABLE I

DEMOGRAPHIC AND CLINICAL PROFILE OF THE GROUPS STUDIED ($mean \pm standard\,deviation$).

Group	NC	MCI	AD
Number of subjects	104	207	95
Number of images	365	714	314
Age	76.9 ± 4.8	76 ± 7.3	76.5 ± 7.1
Sex (% M)	63.8	66.2	59.9
MMSE	29.1 ± 1.1	26.6 ± 3.2	21.6 ± 4.4
CDR	0.02 ± 0.2	0.5 ± 0.2	0.95 ± 0.5

scans were normalized, averaged and co-registered by ADNI researchers, and were also further normalized in the range of [0,1] and cropped from 60x128x128 to 40x98x98, to remove most of the non-relevant area surrounding the brain.

B. Data division

To train and test the models a 5-fold cross-validation was performed. The subjects, and not the images, were separated into five folds, to guarantee that brain scans from the same subject were not present in different sets, avoiding data leakage. Five models were trained and each model used one of those folds for testing (around 20% of the dataset) and the remaining four for training (around 80% of the dataset). For each train, the subjects in the training set were further divided into subjects for training the model (80% out of the subjects of the original training test) and subjects for the validation of the model (20% out of the subjects of the original training test). In the end, to convert the subjects sets into image sets, all images from the same subject were added to the the corresponding set, originating the final training set (with 64% of the images), the final validation set (with 16% of the images) and the final test set (with 20% of the images).

C. Curriculum learning strategies

Five different CL strategies were implemented, three of which were manual and two were automatic.

MMSE strategy: This manual CL strategy consists on feeding the network with samples ordered by difficulty. Therefore, the network first trains with easier samples and afterwards is challenged with more difficult ones. The MMSE score was used to access sample difficulty. Using MMSE scores to build a curriculum is convenient, since cognitive tests are routinely performed in AD diagnosis, and has not been explored before. A sample was considered an easy if its label (NC, MCI and AD) and its corresponding MMSE score were in agreement. For example, according to the MMSE scale, a score between 24 and 30 is associated to no dementia (Figure 1), and all images labeled as NC with MMSE score in that range are considered easy samples. This strategy corresponds to 3 rounds of training, where at the end of the first two rounds, the last fully connected layer of the model (which contains the information about the predicted label) was replaced for a new randomly initialized layer. As depicted in Figure 1, in the first round only the easy samples of AD and NC (according to the MMSE) were included, to guarantee that the discriminative features of the AD and NC are well learned, without noisy information. Then, in the second round, the MCI samples are added to the training data. In the last round, the AD hard samples are added to the training data, which now comprises all training samples.

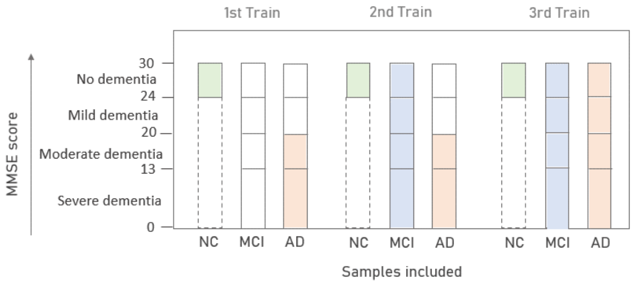

Fig. 1. Manually defined curriculum based on MMSE. The NC, MCI and AD samples included in each round of training are represented and their MMSE scores are presented in the vertical axis.

Task strategy: In this manual approach, the samples are fed into the network ordered by task complexity. It follows the transfer learning proposal of [8], yet it is adapted to a CL strategy consisting of two rounds of training: in the first one the model is trained with only AD and NC samples (samples from only two classes and easier to distinguish between them), and in the second round the MCI samples are added (samples from three classes and harder to distinguish between them).

ROI focused strategy: This manual strategy focus on progressively adding to the training set more complex regions of the images. The model was first trained with the dataset images multiplied by a ROI mask (1 inside the ROI, 0 outside), then it was retrained (fine-tuned) using the complete images. The ROIs used were manually delineated by an experienced physician and correspond to the gyrus, cingulate and precuneus, which match the most discriminative regions for AD [9].

Self-paced Learning: Self-paced learning (SPL) is an automatic CL strategy where data are sorted while training, based on sample training loss [10]. A threshold, λ, is defined and the samples with loss below (above) λ are considered easy (hard). During training the threshold is updated, according to a growing factor (= 1.5), from including only the lower loss samples, to including all samples in the final epochs. This strategy does not take prior medical knowledge into account.

Self-paced Curriculum Learning (SPCL): SPCL results from the merge of manual CL with SPL, taking into account both prior knowledge and the learning progress of the model during training. In this strategy, the predetermined curriculum, where prior knowledge is encoded, is given as input and updated at each epoch.

In this paper, a SPCL algorithm was implemented (Algorithm 1), inspired by the implementation in [11], yet adapted to the current classification problem.

Algorithm 1 Self-paced curriculum learning algorithm

```
$training\_samples = [s_1, s_2, ..., s_N]$
$\gamma = [\gamma_{s_1}, \gamma_{s_2}, ..., \gamma s_{s_N}]$          ▷ Predetermined curriculum
$\lambda(t)$                                   ▷ Growing function
for $t$ in $[0, E]$ do:
    Train the model using $training\_samples$
    $losses = [ls_1, ls_2, ..., l_N]$          ▷ Normalized loss
    $\gamma = \gamma \odot losses$             ▷ Update curriculum
    $threshold$ = $\lambda(t)$                 ▷ Update threshold
    $updated = [\,]$
    foreach $x \in [0, ..., N]$ do:
        if $\gamma_{s_x} <= threshold$ then:
            $updated = updated + [s_x]$
        end if
    end for
    $training\_samples = updated$
end for
```

In Algorithm 1, $training\_samples$ contains the samples the model should train with, at each epoch, and $losses$ contains the loss for each sample, normalized to [0,1]. Moreover, γ consists on the curriculum, which is updated during training through element wise multiplication ($\odot$) with the $losses$ vector. N represents the total number of training samples and E represents the total number of epochs.

The predefined curriculum, γ, and the growing function, $\lambda(t)$, given as input, were defined according to:

- γ is an array with values in [0,1], where each instance γ_{si}, corresponds to the weight of each training sample, s_i. The easier samples have lower γ_{si} values, since they are the ones that should be learnt first in the training process. Two SPCL strategies were implemented, differing only on the predetermined curriculum. In SPCL 1, each entry of the predefined curriculum vector was defined as: $\gamma_{s_i} = 0.33$ if s_i is an easy AD or NC sample; $\gamma_{s_i} = 0.66$ if s_i is a MCI sample or $\gamma_{s_i} = 0.99$ if s_i is a hard AD sample. This follows the same curriculum used in the manual CL strategy described in Figure 1: first training with easy AD and NC samples, then MCI samples are added and afterwards hard AD samples are also added. In the other strategy, SPCL 2, the predetermined curriculum follows the curriculum of the task strategy and γ is defined as: $\gamma_{s_i} = 0.33$ if s_i is an AD or NC sample and $\gamma_{s_i} = 0.99$ if s_i is a MCI sample.
- The growing function, $\lambda(t)$, dictates how the threshold grows. Similarly to [12], $\lambda(t)$ was defined so training would start with only 2% of samples at the first iteration, t=0, and then exponentially increase to include all samples in 3/4 of the maximum epoch, in epoch t=75.

D. Architecture and experimental design

The CL strategies were applied to a 3D-CNN. Its architecture consists of three convolutional blocks where the 3D convolutional layers are composed of 8, 16 and 32 filters, respectively, with ReLU activation function. Each convolutional layer is followed by a 3D max-pooling layer and a batch normalization layer. The output of the last convolution block is then flattened and fed into a fully connected classifier network, with 64 units and a softmax layer in the end, allowing the classification into 3 classes: NC, MCI and AD.

The experiments were performed on a single NVIDIA GeForce GTX 1070 GPU with 8GB of memory, in a machine with an Intel Core i7-6800HQ @ 3.40GHz CPU. To train and test the models a 5-fold cross-validation was performed. The subjects, and not the images, were separated into five folds, to guarantee that brain scans from the same subject were not present in different sets. The categorical cross-entropy was chosen as the loss function and ADAM optimizer was used with $lr = 0.001$. The models, except for focal loss and automatic, were trained with a weighted training strategy, where the weight of the class was inversely proportional to the class frequencies in the train set. Also, a batch size of 16 was used, for a total number of 100 epochs, using an early stop criterion monitoring the validation loss with patience of 50 epochs.

E. RESULTS

The CL strategies were compared to two baseline methods, simple model and Focal loss.

Simple model: The same architecture trained without CL.

Focal loss: In this method the model was trained like the simple model, but the loss function used was the balanced focal loss (FL) [13]. The FL function not only deals with class imbalance but also estimates sample complexity and takes it into account during training. It is described by equation 1:

$$FL(y, \hat{p}_y) = -\alpha(1 - \hat{p}_y)^{\delta} * log(\hat{p}_y) \qquad (1)$$

where $y = [0, ..., K - 1]$ is an integer class label (K denotes the number of classes), $\hat{p}_y = [\hat{p}_0, ..., \hat{p}_{K-1}]$ is a vector representing an estimated probability distribution over the K classes and α represents the balance factor. FL, according to δ, smoothly adjusts the rate at which easy examples (correctly classified) are down weighted. In our implementation we used $\alpha = 0.25$ and $\delta = 2$.

The overall results are presented in Table II. They show that the use of the CL strategies improves the overall accuracy and F1-score of the classifications, up to 4.5% and 4.3%, respectively. The simple baseline model presents the poorest performance. The FL model, although it takes the models feedback into account, it is still worse than all the CL strategies and does not allow for the incorporation of medical knowledge. The results per class are summarized in Figure 2, where we can observe that, even though the CL strategies decrease the accuracy of the classification of AD and CN individually, all of them improve the MCI classification. Although baseline methods are faster than

TABLE II

OVERALL RESULTS OF ACCURACY (ACC), F1-SCORE (F1) AND TRAINING TIME AS $(mean \pm standard\,deviation)$.

Model		**Acc (%)**	**F1 (%)**	**Time (min)**
Baseline	**Simple**	82.7 ± 0.8	83.0 ± 0.8	52
	FL	84.0 ± 0.5	83.7 ± 0.3	55
Manual CL	**MMSE**	86.2 ± 1.2	86.3 ± 1.3	174
	Task	86.6 ± 0.8	86.6 ± 0.8	97
	ROI	86.7 ± 1.8	86.8 ± 1.8	101
Automatic CL	**SPL**	85.9 ± 0.9	86 ± 0.9	**43**
	SPCL 1	$\mathbf{87.2 \pm 1.3}$	$\mathbf{87.3 \pm 1.3}$	48
	SPCL 2	86.4 ± 1.7	86.5 ± 1.8	48

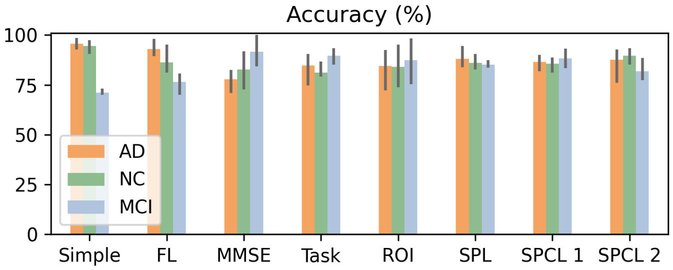

Fig. 2. Bar plots representing the accuracy per class (AD, NC and MCI), for all the implemented models.

manual CL, the automatic CL methods are even faster and yield better results. SPCL, in comparison with SPL, requires the extra work of building the curriculum. Nevertheless, it complements SPL, improving its accuracy by 1.3%, in the case of SPCL 1, achieving the best overall performance. Regarding the manual CL methods, the ROI strategy was the one that yielded better overall results. Still the MMSE strategy has shown to be the best at classifying MCI, with an MCI accuracy of $91.7(\pm5.1)\%$. The use of CL has proven to be advantageous in all cases. Additionally, the incorporation of medical knowledge into the process of building the curriculum has also proven to be advantageous, since all strategies that incorporate it yield better results than SPL, which does not take it into account.

IV. CONCLUSIONS

This paper was, as far as we know, the first work investigating the use of curriculum learning for early AD diagnosis from neuroimaging. It is also the first work using the scores of cognitive tests to build a training curriculum. Five different CL strategies were implemented, three manual and two automatic, incorporating different kinds of medical knowledge into the process of building the curriculum (task complexity, cognitive test scores, ROI information).

The results obtained show that all the proposed CL strategies improve both overall and MCI classification (early AD) performances. Out of the manual strategies, the ROI focused strategy was the one to yield the best overall results and MMSE obtained the best MCI accuracy. The automatic strategies have shown to be the best ones, allowing to obtain the highest performances in lowest time. In fact, SPCL 1 has obtained the highest overall accuracy and F1-score. The incorporation of medical information (ROI information and MMSE scores) into the CL strategies has proven to be advantageous, improving the overall accuracy, F1-score and MCI accuracy.

The results obtained in this paper show that CL strategies, in which the curriculum is built based on medical knowledge, allow for better early AD diagnosis, which can contribute to the ongoing search for treatments to prevent or delay the onset of this devastating disease.

V. ACKNOWLEDGMENT

This work was supported by Portuguese funds through FCT (Fundação para a Ciência e Tecnologia) under LARSyS - FCT Project UIDB/50009/2020.

REFERENCES

[1] C. L. Masters, R. Bateman, K. Blennow, C. c. Rowe, R. A. Sperling, and J. L. Cummings, "Alzheimer's disease," *Nature Reviews Disease Primers*, vol. 1, pp. 15056, 2015.

[2] M.A. Myszczynska, P.N. Ojamies, A.M.B Lacoste, D. Neil, A. Saffari, R. Mead, G.M Hautbergue, J. Holbrook, and L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Reviews Neurology*, vol. 16, pp. 440–456, 2020.

[3] A. Ebrahimighahnavieh, , S. Luo, and R. Chiong, "Deep learning to detect Alzheimer's Disease from neuroimaging: A systematic literature review," *Computer Methods and Programs in Biomedicine*, vol. 187, 2020.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," *Springer*, pp. 41–48, 2009.

[5] Y. Tang, X. Wang, A.P. Harrison, L. Lu, J. Xiao, and R.M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," *Springer*, pp. 249–258, 2018.

[6] C. Haarburger, M. Baumgartner, D. Truhn, M. Broeckmann, H. Schneider, S. Schrading, C. Kuhl, and D. Merhof, "Multi scale curriculum CNN for context-aware breast MRI malignancy classification," *Medical Image Computing and Computer Assisted Intervention*, p. 495–503, 2019.

[7] G. Maicas, A. P. Bradley, J. C. Nascimento, I. Reid, and G. Carneiro, "Training medical image analysis systems like radiologists," *Springer*, pp. 546–554, 2018.

[8] M. Grassi, D. A. Loewenstein, D. Caldirola, K. Schruers, R. Duara, and G. Perna, "A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach," *International psychogeriatrics*, vol. 31, no. 7, pp. 937–945, 2019.

[9] J. Rondina, L. Ferreira, F. de Souza Duran, R. Kubo, C. R. Ono, C. C. Leite, J. Smid, R. Nitrini, C. A. Buchpiguel, and G. F. Busatto, "Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: A comparison across functional and structural imaging modalities and atlases," *NeuroImage: Clinical*, vol. 17, pp. 628–641, 2018.

[10] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *Advances in neural information processing systems*, vol. 23, pp. 1189–1197, 2010.

[11] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[12] K. Ghasedi, X. Wang, C. Deng, and H. Huang, "Balanced self-paced learning for generative adversarial clustering network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4391–4400. 2019.

[13] T. Lin, P. Goyal, R. Girshick, H. He, and P. Dollár, "Focal loss for dense object detection," *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.