CROSS-LINGUAL TRANSFER LEARNING FOR ALZHEIMER'S DETECTION FROM SPONTANEOUS SPEECH

*Bastiaan Tamm, Rik Vandenberghe, Hugo Van hamme**

KU Leuven, Belgium

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disease most often associated with memory deficits and cognitive decline. With the aging population, there has been much interest in automated methods for cognitive impairment detection. One approach that has attracted attention in recent years is AD detection through spontaneous speech. While the results are promising, it is not certain whether the learned speech features can be generalized across languages. To fill this gap, the ADReSS-M challenge was organized. This paper presents our submission to this ICASSP-2023 Signal Processing Grand Challenge (SPGC). The model was trained on 228 English samples of a picture description task and was transferred to Greek using only 8 samples. We obtained an accuracy of 82.6% for AD detection, a root-mean-square error of 4.345 for cognitive score prediction, and ranked 2nd place in the competition out of 24 competitors.

***Index Terms*—** Alzheimer's disease, cross-lingual

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disease most often associated with memory deficits and cognitive decline. It is the most common form of dementia and the fifth-leading cause of death among people age 65 or older [1]. With an aging population, there has been much interest in automated methods for cognitive impairment detection, especially ones that are inexpensive and easily scalable. One possibility that has gained a lot of attention in recent years is to analyze spontaneous speech, a readily available medium that can provide insight into the working of the brain. However, most of the proposed approaches have not investigated which speech features can be transferred across languages for AD detection [2].

Hence, ADReSS-M, an ICASSP-2023 Signal Processing Grand Challenge (SPGC), was organized to investigate this matter [3]. The goal is to train a model on English speech from a picture description task and apply it to a different picture description task in Greek. The challenge has two tasks, first to predict the AD diagnosis and second to predict the Mini-Mental State Examination (MMSE) score of a participant. The latter is a questionnaire that is used extensively in clinical and research settings to measure cognitive impairment and is scored out of 30 points.

In this paper, we present our submission to the challenge.[1] The models use a sequence of acoustic features and covariates (age, gender, education) to make the predictions. They are first trained in English, and then they are transferred to Greek using mixed-language batches and parameter averaging. This approach obtained an accuracy of 82.6% for AD detection and 4.345 for cognitive score prediction, compared to 73.9% and 4.955 respectively by the baseline. With this submission, we ranked 2nd place in the competition out of 24 competitors.

2. METHODS

2.1. Datasets

Two datasets were used in this challenge, one English and one Greek, consisting of audio recordings of healthy controls and AD patients who were asked to describe a picture. The challenge organizers divided the data into three splits: an English training split (n=237, 122 AD), a Greek sample split (n=8, 4 AD), and a Greek test split (n=46, 22 AD). The test statistics were derived from the confusion matrix of the submission and were not known ahead of time. It was known that the splits were balanced for AD, age, and gender [3].

We removed one healthy control from the English training split (no cognitive score) and 8 AD patients for balancing (n=228, 114 AD). Finally, in 12 controls where education was not available, we assumed the missing value to be 12 years.

2.2. Preprocessing

Each audio file is split into ten equal segments. For each segment, a 25-D eGeMAPS [4] feature vector is calculated using openSMILE [5].

2.3. Model

Both the AD detection model and the cognitive score prediction model are based on the same architecture. They are

*This research was supported by KU Leuven Special Research Fund grant C24M/22/025.

[1] Our Code: https://github.com/lcn-kul/madress-2023

rather small: only 767 and 468 parameters respectively. Each model takes as input a sequence of ten eGeMAPS features and the covariates age, gender and education. The estimated AD probability is also included as a covariate for the cognitive score prediction model. For simplicity, the covariates are concatenated to the eGeMAPS feature sequence.

The architecture consists of four parts. First, batch normalization is applied to the input features. Next, the features are down-projected into a smaller hidden space (12- or 8-dimensional respectively), followed by dropout and ReLU activation.

Afterward, attention pooling is used to collapse the time dimension. The attention weights are calculated using a 2-layer feed-forward network with an intermediate space that is twice as large as the hidden space, followed by softmax such that the weights sum to 1. Between the two layers, dropout and ReLU activation are used.

The final step is a linear projection to map the vector to the output space (2- or 1-dimensional respectively). For the cognitive score prediction model, a sigmoid function is used to map the value to the range [0,1]. The MMSE labels are also normalized to the same range.

2.4. English pre-training

The English data is split into 80% train and 20% validation. The models are trained on the English training data and validated on the Greek sample set. To account for the effects of random initialization, training takes place five times with five different random seeds. The model with the lowest validation loss over the five runs is selected as the pre-trained model.

2.5. Mixed-batch transfer learning

The pre-trained model is finetuned on the English training data and 4 of the Greek samples (2 AD). Every fifth sample of the mini-batch is replaced by a Greek sample. The model is validated on the English validation data and the 4 held-out Greek samples, inserted in the same way.

To improve robustness, we repeat the procedure but swap the 4 training Greek samples with the 4 held-out samples. Finally, the parameters of these two models can be averaged since they are initialized from the same pre-trained model.

2.6. Training details

The models are implemented using the PyTorch (v.1.11.0) and PyTorch Lightning (v.1.8.6) libraries in Python 3.8. The network is trained using the AdamW optimizer with a weight decay of 1e-2. The learning rate is warmed up linearly for 100 steps and is fixed at 3e-3 afterward. Cross-entropy loss is used for the AD detection model, and mean-square-error loss is used for the cognitive score prediction model. Each model is trained with a batch size of 32 for a total of 30 epochs, and the model with the lowest validation loss is selected.

3. RESULTS

The challenge allowed for five submissions. So, to test the robustness of the procedure proposed above, the entire procedure was run five times with different random seeds. The test accuracies for the AD detection task in ascending order were 71.7%, 73.9%, 76.1%, 80.4% and **82.6%**. The best-performing model had a specificity of 91.7%, precision of 88.9%, sensitivity of 72.7%, and an F1-score of 80.0%.

The root-mean-square errors (RMSE) of the cognitive score prediction task in descending order were 4.837, 4.816, 4.716, 4.713, **4.345**. Note that these models use the probabilities of the AD detection model as input. Since the best AD detection model was not known ahead of time, the probabilities of the 5 submitted AD detection models were averaged.

4. CONCLUSIONS

In this paper, we present our submission for the ADReSS-M challenge. Our approach outperforms the best baseline model and is robust to model initialization. The innovation in our work lies in pre-training and mixed-batch fine-tuning procedure. The actual architecture is extremely simple and we believe with an improved feature extraction network, performance can be further improved.

5. REFERENCES

[1] "2022 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 18, no. 4, pp. 700–789, 2022.

[2] Sofia de la Fuente Garcia, Craig W. Ritchie, and Saturnino Luz, "Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review," *Journal of Alzheimer's Disease*, vol. 78, pp. 1547–1574, 2020, 4.

[3] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney, "Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," 2023.

[4] Florian Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[5] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, p. 1459–1462, Association for Computing Machinery.