RESEARCH

Transformer-based deep neural network language models for Alzheimer's disease detection from targeted speech

Alireza Roshanzamir[1], Hamid Aghajan[2] and Mahdieh Soleymani Baghshah[3*]

[*]Correspondence:
soleymani@sharif.edu
[3]Department of Computer Engineering, Sharif University of Technology, Azadi, 11365-11155 Tehran, Iran
Full list of author information is available at the end of the article

Abstract

Background: We developed transformer-based deep learning models based on natural language processing for early diagnosis of Alzheimer's disease from the picture description test.

Methods: The lack of large datasets poses the most important limitation for using complex models that do not require feature engineering. Transformer-based pre-trained deep language models have recently made a large leap in NLP research and application. These models are pre-trained on available large datasets to understand natural language texts appropriately, and are shown to subsequently perform well on classification tasks with small training sets. The overall classification model is a simple classifier on top of the pre-trained deep language model.

Results: The models are evaluated on picture description test transcripts of the Pitt corpus, which contains data of 170 AD patients with 257 interviews and 99 healthy controls with 243 interviews. The large bidirectional encoder representations from transformers ($\text{BERT}_{\text{Large}}$) embedding with logistic regression classifier achieves classification accuracy of 88.08%, which improves the state-of-the-art by 2.48%.

Conclusions: Using pre-trained language models can improve AD prediction. This not only solves the problem of lack of sufficiently large datasets, but also reduces the need for expert-defined features.

Keywords: Alzheimer's disease; Early diagnosis; Picture description test; Deep learning; Transformer; Natural language processing; Language model; Transfer learning

Background

Alzheimer's disease (AD) is the most common type of dementia which currently cannot be cured or reversed [1]. According to the World Alzheimer Report 2019, there were over 50 million people living with dementia in the world as estimated by Alzheimer's Disease International (ADI), while the projected estimates for 2050 reach above 150 millions [2]. The symptoms of AD include decreased awareness, disinterest in unfamiliar subjects, increased distraction, speech problems, and etc. [3]. However, if the disease is diagnosed in its early stage, a series of pharmacological and behavioral therapy approaches can be prescribed to reduce the pace or progression of the disease symptoms [4]. Clinical levels of cognitive impairment are categorized into 7 stages of: normal, normal ageing forgetfulness, mild cognitive impairment (MCI), mild AD, moderate AD, moderately severe AD, and severe AD [5]. If we want to enumerate only the observable linguistic symptoms, in the first three stages, the participants need more time to respond and find words, or have trouble to maintain focus on a conversation. In mild and moderate AD stages, patients have difficulty in understanding and explaining abstract concepts, completing sentences, and following long conversations. In the two most severe stages, patients cannot create grammatically correct sentences, almost lose the ability to understand words, and finally, become completely mute [5] [6] [7].

According to the recent increasing power of natural language processing (NLP) and deep learning techniques, employing these methods in medical text mining problems has seen increased interest in recent years. Given the importance of the impact of AD on speech abilities of the patients, this study aims to develop a technique for diagnosing AD from transcripts of targeted speech elicited from the participants.

The task for acquiring speech data from the patients is the Cookie-Theft picture description test[8]. Initially, the test was used as a part of the Boston Diagnostic Aphasia Examination [8] assessment tool which designed for diagnosing aphasia. Currently, the test is commonly used by speech-language pathologists to assess abnormal language performances in patients with disorders such as aphasia, AD, right hemisphere lesions, schizophrenia, and etc [9]. In this test, an image is shown to the participant and he/she is asked to describe what he/she sees in it. Generally, the Cookie-Theft image includes a mother washing the dishes in a sink while children try to steal cookies from a cookie jar.

Unlike most earlier studies, the features are extracted in our approach by the model itself in an unsupervised manner. As a result, more complex features are discovered and used for diagnosis. More precisely, the models are pre-trained on a large dataset to learn a good high dimensional (such as 1024 dimensions) vector representation for the input sentence or text, which will be used as input to AD versus healthy control (HC) classifiers. Another approach taken in this study to address the problem of insufficiently-sized datasets is text augmentation. Similar to most related works, the methods are evaluated on the Cookie-Theft picture description test transcripts of the Pitt corpus [10] from the DementiaBank [10] dataset. As mentioned earlier, the overall classification framework takes raw interview text as input. Our evaluation shows that pre-trained deep transformer-based language models with a simple logistic regression classifier work well in AD prediction and the results generally outperform those of the existing methods while the proposed method does not require any hand-crafted features for training the classifier.

Related Work

Feature-based approaches

For the first time, a computational approach to diagnosing Alzheimer's disease using speech in English was introduced by Bucks et al. [11]. In that study, 8 AD and 16 HC participants were asked to speak about themselves and their experiences in 20 to 45 minute sessions, and finally, some specific questions were also asked. Then, a number of linguistic features such as the noun rate, adjective rate, pronoun rate, and verb rate were extracted from the recorded speech and their distribution for the AD and control samples were used to train a classifier. Since then, many other studies [12, 13, 14, 15, 16, 17, 18, 19] haves been conducted on this topic to improve the accuracy of AD prediction and study the various dimensions of AD (and other types of dementia) effects on speech. In general, most of these methods propose improvements based on increasing the number of expert-defined features, increasing the number of participants, using acoustic features in addition to linguistic ones, involving AD severity and other types of dementia in classification, and changing the interviews structure.

One of the most comprehensive studies on this topic was conducted by Fraser et al. [20]. In that study, an extensive categorization of linguistic features was presented, in which linguistic features were categorized into POS (part-of-speech) tags, syntactic complexity, grammatical constituents, psycho-linguistics, vocabulary richness, information content, repetitiveness, and acoustics. Also, the study categorized all different kinds of language disorders into the four groups of semantic impairment, acoustic abnormality, syntactic impairment, and information impairment. The paper collected 370 linguistic features from the data and reported the topmost 35 of these features for AD prediction.

In all earlier works, in order to automatically diagnose the disease using speech, information content units were introduced by human experts, and a classifier used them in order to predict the participant's category. However, Yancheva et al. [21] and Sirts et al. [22] tried to enrich and enhance information content units of targeted speech by clustering pre-trained global vector (GloVe) [23] embedding of words used by AD and HC participants. Using the mentioned clusters, they introduced some cluster-based measures which were used along with a number of standard lexicosyntactic and acoustic features for AD prediction.

In languages other than English, Khodabakhsh et al. [24] and Weiner et al. [25] respectively examined the subject in Turkish and German. Also, Li et al. [26] and Fraser et al. [27] both focused on multilingual approach for diagnosing AD using targeted speech. They respectively tried to improve the AD prediction in Chinese and French languages (which the existing datasets were insufficient) using an English classifier trained on a larger English dataset.

Deep learning-based approaches

For the first time, Orimaye et al. [28] used a deep neural network to predict MCI using speech. Unlike most previous works, that study did not use any hand-crafted features and the raw transcripts were fed to the model. The dataset used in the study was part of the Pitt corpus of the DementiaBank dataset, comprising 19 MCI and 19 control transcripts of the Cookie-Theft picture description test. They trained a separate deep neural network language model for each category, and then calculated the likelihood of the text in both language models. Finally, the class of the model with higher probability was selected.

Karlekar et al. [29] also used a deep neural network model to diagnose AD using four types of interviews: the Cookie-Theft picture description, sentence construction, story recall, and vocabulary fluency which included an unbalanced 243 HC and 1017 AD transcripts. Three classifiers: a convolutional neural network (CNN), a long-short term memory recurrent neural network (LSTM-RNN), and a CNN-LSTM were trained, taking sentences as sequences of pre-trained word embedding. In addition to AD diagnosis, the authors interpret the models using activation clustering and first derivative saliency heat map techniques which cluster the most significant utterances. The research used a highly unbalanced dataset, rendering the results somewhat questionable as discussed in Section Why not using the entire Pitt corpus?.

Fritsch et al. [30] used two different auto-regressive LSTM-based neural network language models to classify AD and HC transcripts of the Pitt corpus from the DementiaBank dataset. After that, Pan et al. [31] worked on predicting AD using a stacked bidirectional LSTM and gated recurrent unit (GRU) layers equipped with a hierarchical attention mechanism. The overall model takes the GloVe word embedding sequence as input.

Methods

The most challenging problem in developing technique for recognizing Alzheimer's patients from speech transcripts is the lack of a large dataset. Currently, the largest available dataset is the Pitt corpus from the DementiaBank dataset, which contains 500 picture description interviews from the AD and control groups. For the mentioned reason, most of the earlier work was based on features designed by experts, as it was not possible to use models capable of learning informative features, by themselves. In this study, we simultaneously employ the two ideas of employing a highly pre-trained language model and dataset augmentation to address this issue and enhance the classification accuracy. Our implementation of these ideas is described next.

Pre-trained deep language model

Every model that defines a probability distribution over a sequence of words is called a language model. If a computational model wants to implement a language model, it is necessary to have a good understanding of the syntactic and semantic structures of that language. Therefore, using a model that has already learned a probabilistic distribution that correlates with these structures for classification almost eliminates the need for large target-specific datasets. The transfer of knowledge from one model to another with a similar purpose is called transfer learning. We use transformer-based language models that have offered a breakthrough in many language understanding tasks in recent years [32]. The general flow of using pre-trained language model for classification task consists of three steps:

1 Unsupervised training of the general language model on a large dataset (such as Wikitext).

2 Unsupervised fine-tuning of the pre-trained language model on the target dataset (such as the Cookie-Theft picture description transcripts).

3 Using (with or without supervised fine-tuning) the target-specific pre-trained language model for the classification task.

To address the problems facing recurrent models such as the issue of short-term memory and the challenges facing the parallelization of training, Vaswani et al. [33] introduced transformers which consist of an extreme use of the attention mechanism that underpins many NLP models. The paper argues that the attention mechanism allows the model to focus on certain parts of the text for decision making. This indicates its suitability for the diagnosis of AD as it can capture specific language markers related to the disease.

Al-Rfou et al. [34] used transformers for the first time as essential elements of a character-level language model. After that, Dai et al. [35] extended the model using relative positional encoding and segment-level recurrence. As a turning point in the transformer-based language models, we can refer to the bidirectional encoder representations from transformers (BERT) model proposed by Devlin et al. [36] at Google. In the training phase, the input sentence is masked, which means 15% of tokens are replaced with the [MASK] token, and the model tries to learn such representation or embedding for the context that considers both syntax and semantics to predict the masked token using the context. On the other hand, in the test phase the model takes in a raw sentence from one or multiple languages and returns a 768- or 1024-dimensional vector representation of the input text to be used as input to other classifiers such as LR, MLP, etc. An enhanced version of BERT for multilingual language understanding tasks was introduced by Conneau et al. [37], called cross-lingual language model (XLM), which benefits from using the translated language model (TLM) as well as the masked language model (MLM). Unlike BERT, XLM takes two related masked sentences from two different languages and tries to predict masked tokens using the same and the other language input sentences. This allows XLM to understand multilingual texts better. Also, BERT suffers from the train and test phase discrepancy and independent prediction of masked tokens. To correct this, Yang et al. [38] introduced an extended large network (XLNet) model based on a language model called Permutation Language Model.

In the current study, we use pre-trained BERT, XLNet, and XLM as deep networks for text embedding which convert raw participant transcripts / sentences to 768- or 1024-dimensional vectors. These models are used in two ways described in Section Overall classification framework.

Baseline models

In this study, in addition to the transformer-based models, bidirectional-LSTM and convolutional neural networks over the GloVe [23] word embedding were also evaluated as baseline models to illustrate the advantages of pre-trained transformer-based deep language models over conventional deep models. In the CNN model, each transcript (truncated or padded to T number of words) is converted to a sequence of embedded words. Then the sequence is passed to a number of stacked convolutional and max-pooling layers followed by fully-connected layers and finally a sigmoid output layer that yields $P(AD|transcript)$. Also, in the bidirectional-LSTM model, the embedded word sequence is passed to a number of stacked forward and backward LSTM cells followed by fully-connected layers and a sigmoid output layer in a similar fashion. Structurally, if we move forward in the CNN layers, the model tries to conclude more semantic features using spatially close features in the previous layer. But in the LSTM model that considers long range dependencies, an attempt is made to learn new compound features from features of all previous steps (or from features of the whole sequence in the bidirectional LSTM). The main weakness of this model is the forgetting of distant features (spatially) to produce new compound features. In both of these models, there is no attention mechanism.

Dataset augmentation

Another approach to overcome the lack of access to large training input is dataset augmentation which means increasing the number of labeled samples of the dataset using some probabilistic or even heuristic algorithms. For example, the word *"beau-*

tiful" in a sentence such as *"What a beautiful car!"* can be replaced with the word *"nice"* without changing the meaning of the sentence a lot. Augmentation in NLP can be done at the character, word, and sentence levels, and in this study, the word and sentence levels are used for augmenting the dataset. The most crucial challenge of augmentation in the text classification task is preserving the text class during augmentation. For example, a probabilistic model can replace *"beautiful"* with *"dirty"* in the mentioned sentence, which is grammatically and semantically correct but changes the sentence category. Two general approaches to augmentation have been used in this study, which are described below.

Similar word substitution augmentation

In this approach, a similarity measure must first be defined. The most obvious definition of similarity for words is the synonym relation which was first used in the field of deep learning by Zhang et al. [39] using the WordNet database [40]. Another common similarity measure is the inverse of the Euclidean distance or the Cosine similarity between word embeddings which was first used by Wang et al. [41]. In the mentioned methods, there is no guarantee of the correct grammar in the output sentence. It is also possible that the output sentence category changes by augmentation. For example, one of the markers of Alzheimer's disease is the reduction in the vocabulary used in the conversation, so replacing a simple word like *"Delicious"* with its sophisticated synonym like *"Scrumptious"* can change the sentence category from patient to healthy and mislead the classifier. Another method that considers grammatical correctness along with the sentence context was introduced by Kobayashi [42] and is called contextual augmentation. In the contextual augmentation method, there is a language model which takes both the word's context (i.e. the sentence that contains the word) and the whole sentence's category and returns a probability distribution over all vocabulary. Augmentation is done by sampling from the returned probability distribution. Kobayashi [42] trained a Bi-Directional LSTM language model with this approach, and Wu et al. [43] enhanced the approach by using BERT as an underlying model.

All the mentioned methods were evaluated in this study, and the implementation was done using the NLPAug library [44] except for contextual augmentation for which the released code by the authors of [42] was used.

Sentence removal augmentation

Another ad-hoc approach which does not change the sentence category and also retains grammatical correctness is sentence removal. In this approach, one sentence is removed from the transcript, and it is expected that the output is still a valid transcript in the same category. Although it can be argued that the label may be changed by reducing the length of the text, considering the results of using or not using this idea, it is appropriate to use it in models that process the entire text at once (not sentence by sentence).

Overall classification framework

The overall process of classification is summarized in Figure 1. The process consists of five layers. The augmenter layer enriches the dataset using the methods introduced in Section Dataset augmentation. Note that this layer will be disabled in the test phase. The splitter layer splits the entire transcript text into its sentences when we want to work on sentences and could be disabled by being set to the identity function when we intend to work on the whole transcript. The embedder layer embeds each input element (i.e. the entire transcript or a sentence) to a high-dimensional representation vector, and the classifier layer predicts the label of each embedded input. In fact, the classifier layer learns which of (and to what extent) the features that BERT offers is suitable for diagnosing Alzheimer's disease. Finally, if the classifier layer outputs multiple labels (that may happen when working on sentences), the voter makes the final decision using a majority voting mechanism.

In this study, two different approaches for classifying a transcript are implemented. In the first approach, the entire transcript is passed to an embedder and then the embedded transcript is directly classified. In this approach, the splitter and voter layers are disabled. In the second approach, the transcript is first split into sentences, and then these sentences are embedded and are subsequently classified. Finally, the label of the entire transcript is decided by majority voting on the labels of all sentences in the transcript. The second approach is more compliant with pre-trained embedders since they are mostly pre-trained on single- or two-sentence inputs.

The embedding models (which used in this study as an embedder layer) are only passed through Phase 1 and 3 of the flow described in Section Pre-trained deep language model. The reason for this is that the dataset used is insufficient for unsupervised fine-tuning even when using vast augmentation methods. In practice, using unsupervised fine-tuning has no impact on the overall model's performance used in the current research. For the first phase, all embedding models are pre-trained with the corpus mentioned in their main article, and their implementation is taken from the HuggingFace transformers library [45].

Results

Dataset

The models are evaluated on the transcripts of the Cookie-Theft picture description test of the Pitt corpus from the DementiaBank dataset, which contains 170 possible or probable AD patients with 257 interviews and 99 healthy control (HC) participants with 243 interviews.

Most of the data were gathered as a part of the Alzheimer's and related dementias study at the University of Pittsburgh School of Medicine between 1983 and 1988. The interviewer shows the participant the Cookie-Theft picture and asks him/her to state everything he/she sees in it. The audio records of all interviews were manually transcribed and annotated with POS-tags in the CHAT [46] format.

Detailed demographics of the data is specified in Table 1.

Why not using the entire Pitt corpus?

Some earlier studies based on the Pitt corpus (such as Kerlekar et al. [29]) used all the tests of the corpus including the Cookie-Theft picture description, story recall, sentence construction, and categorical/verbal fluency for classification purposes. The first problem with using the entire corpus is that the corpus is highly unbalanced, and as a result, a naïve classifier that always outputs AD labels can achieve a classification accuracy of 80% on such a dataset.

The second problem is that except for the Cookie-Theft picture description test, the Pitt corpus only contains a single transcript for all the other tests, which means that the classifier might learn invalid features for AD detection. For example, a classifier may just output an AD label by checking if the input is not from the Cookie-Theft picture description test, and otherwise, work as normal. Using this approach, a normal classifier with 80% accuracy can achieve approximately 92% accuracy on the whole Pitt corpus. Figure 2 provides an example of this problem. The figure shows visualized two-dimensional tSNE [47] diagram for the $\mathrm{BERT}_{\mathrm{Base}}$

embedding of the entire transcripts of all tests in the Pitt corpus. According to the figure, the tests are completely differentiable, and as a result, the mentioned problem is quite probable to arise. Thus, in Section Results, studies based on the entire corpus were not included.

Evaluation measures

The most well-known measure to evaluate classification is the accuracy score which is the fraction of predictions the model performed correctly. Most related studies have reported accuracy as the quality of their classification models and tried to improve this measure as an important goal. As discussed in the previous section, the accuracy measure alone does not provide a complete interpretation of the model performance (for example, high accuracy can be achieved using the entire Pitt corpus, while the model performance is not sufficient for practical use). Two other practical measures are precision and recall (also called sensitivity). In this study, precision is the number of correct AD predicted samples over the total number of AD predicted samples and recall is the number of correct AD predicted samples over the total number of AD samples. These two measures should be examined together and for this reason, the F_1 score is defined. The F_1 score is the harmonic mean of the precision and recall measures. A combined high precision and recall results in a high F_1 score. In other words, highly imbalanced precision and recall indicates that the model has not an approximately equal performance for detecting all labels. All the aforementioned measures are in the range of zero to one, and can be reported as a percentage. Compared to the accuracy score, fewer previous studies have reported recall, precision, and F_1 measures. In this study, all the introduced measures are reported to make it possible to compare our work more comprehensively with previous works.

Compared methods

We compared the results of our models with all related studies that evaluated their models on the Cookie-Theft picture description test of the Pitt corpus. Therefore, the best models (according to the introduced performance measures) are selected for comparison. The first one is the method introduced in [20] which maintained the status of having the state-of-the-art accuracy score for several years. The second compared method was introduced by Yancheva et al. [21]. They tried to enrich and enhance human-supplied information content units by clustering GloVe embedding of frequent words of each category. After that, Sirts et al. [22] extended the idea of Yancheva et al. [21] by introducing propositional idea density features that work better on free-topic conversational speech. Hernández et al. [19] introduced 105 hand-crafted features and used them to train a support vector machine (SVM) classifier. They reported all the well-known and informative measures for the classification tasks and also achieved good results. Fritsch et al. [30] trained two different auto-regressive LSTM-based language models for each group and classified each transcript by calculating its perplexity on the models and selecting the model corresponding to the lowest perplexity. Currently, that study has the best recall and accuracy scores for AD versus HC classification on the target dataset. Pan et al. [31] utilized a stacked bidirectional LSTM and GRU recurrent units equipped with a hierarchical attention mechanism. Up to now, this study has the best precision and F_1 scores for AD versus HC classification on the target dataset. The last two studies by Li et al. [26] and Fraser et al. [27] were focused on multilingual AD prediction and hence their main goal was not to improve the unilingual classification. Li et al. [26] used 185 lexicosyntactic features for a logistic regression classifier and Fraser et al. [27] utilized class-based language modeling and information-theoretic features for an SVM classifier.

Evaluation results

Table 3 reports precision, recall, accuracy, and F_1 scores of the compared methods as well as those of the proposed methods in the framework introduced in this paper. The reported scores are averaged on a 10-fold cross-validation procedure. Note that for the Fritsch et al. [30] method there is no such entity as a classifier and classification was performed by evaluating perplexity of input transcripts on the trained language models of both classes. As mentioned earlier, two different approaches have been implemented to use the pre-trained embedders, the first one is passing the entire text to the embedder (specified by a T- prefix in the method's name) and the second one is passing each sentence of the text to the embedder separately (specified by an S- prefix in the method's name). All the methods with the first approach have been enriched by the one-sentence-removal augmentation method. Furthermore, the CNN method is used with the synonym substitution augmentation

(SSA) method and the BiLSTM is used with the SSA and contextual augmentation (CA) methods separately. The CA and SSA augmentations had almost no effect on the methods which used pre-trained language models, so they are not reported in Table 3.

Moreover, Figure 3 illustrates the mean 10-fold cross-validation classification accuracy, true positive rate (the number of correct predicted AD samples over total number of AD samples, also called the sensitivity), and true negative rate (the number of correct predicted HC samples per total number of HC samples, also called the specificity) plotted versus the mini-mental state exam (MMSE) [48] scores of the participants. The figure helps us to see how the model works for detecting label of participants with different AD severity levels. The true positive rate for each MMSE score represents the model performance in detecting AD from actual AD patients in that score. Similarly, the true negative rate represents the model performance in detecting HC label from actual HC participants in that score. Totally, the accuracy score represents the model performance in detecting the correct label from both participant groups in the corresponding MMSE score. Numbers in the red bars are true positive rates and in the green bars are true negative rates. Also, the numbers on top of the bars are the total mean accuracy for that MMSE score. Note that all of the rates are scaled between 0 and 1. The MMSE scores were not reported in the dataset for some participants while their AD / HC labels were present. The results for these participants are grouped in the "Unspecified" bar in this figure.

In addition to classification, models such as logistic regression and neural networks with a sigmoidal final activation function can also output the AD probability (or 1 - health probability) of the current input. Referring to the continuity of linguistic impairments from perfect health to severe AD, this probability can be interpreted as a correlated variable to the severity of the AD condition of the participant. Therefore, another approach for interpreting the models and evaluating them is calculating the similarity between their predicted health probability and the MMSE score, scaled between 0 and 1. The results using two common similarity measures, the Pearson correlation and Spearman's rank correlation (which is the Pearson correlation on the samples' ranking), are reported in Table 2. Both mentioned correlation measures are reported between -1 and 1.

Discussion

Interpretation of results

According to Table 3, among the models that use only hand-crafted features, Fraser et al. [20] reports the best accuracy score, although it has not reported other evaluation measures. Among the baseline models introduced in our study (CNN + SSA, BiLSTM + SSA, and BiLSTM + CA), which are conventional deep neural network models, the contextual augmented version of bidirectional-LSTM achieved the highest accuracy score of 77.36%. However, even with the extreme use of augmentation methods these baseline models did not yield acceptable results compared to other methods. Overall, the sentence-level $\text{BERT}_{\text{Large}}$ embedding of sentences passed to logistic regression (S-$\text{BERT}_{\text{Large}}$-LR method) achieved the highest accuracy score (88.08%) among all the models introduced in this study as well as the models used in previous studies, and improved the accuracy score by 2.48% (equivalently 17.22% error-rate reduction). At the same time, this model achieved the best precision and F_1 scores with 6.55% and 2.80% improvements, respectively. Still, Fritsch et al.[30] showed the best recall score with 1.66% difference although they did not report F_1 measure . The first advantage of our proposed methods compared to Fritsch et al.[30] is that we train a single language model for both the AD and HC groups which helps the model to use samples from both classes for the desired task. The other advantage is that our models are highly pre-trained on large datasets which enables them to start training on new, smaller datasets with good initialization parameters and also avoid overfitting.

Among the methods evaluated in this study, on average, the models based on the BERT family of embedders worked better than the others. Although XLNet has historically been designed to address BERT problems, BERT and its derivatives still perform better in many activities [32]. One important point to note is that pre-trained deep language models are unaffected by augmentation because these models are highly pre-trained on a large dataset (and hence the evaluation of their versions with augmentation is not reported in Table 3).

Table 2 shows that the best model has a Pearson correlation of 0.78 and 0.70 for the train and validation phases, and a Spearman's rank correlation of 0.81 and 0.74 for these phases between the health score and the MMSE score, indicating that the model has learned useful features for classification. Based on the reported similarity measures, it can be concluded that on average the MMSE score and our model's health score are linearly correlated. This is indeed an advantage for the proposed model in that while the MMSE score [48] is obtained through a detailed interactive exam that evaluates visuospatial, executive, naming, memory, attention, language, abstraction, delayed recall, and orientation cognitive skills, the data collection task involved in the Cookie-Theft picture description test used in our model is a simple and short pseudo-conversational procedure.

In this study, neural network interpretation methods were not used but in Table 4, two false negative and false positive classification errors are reported. In comparison, it is almost clear that the first sample has less grammatical fluency but both samples refer to similar information elements. In the S-BERT$_{\mathrm{Large}}$-LR model, the predicted AD probability is the mean of logistic regression classifier outputs for each sentence of the transcript. The important point is that in both samples, the predicted AD probabilities are very close to 0.5 which can be interpreted as that the model has not learned a wrong feature, rather, it has not learned a proper feature to diagnose AD from the reported samples.

Advantages and limitations

As mentioned in Section Pre-trained deep language model, the proposed approach takes advantage of the powerful pre-trained language models that attempt to learn the structure and features of the language from a large dataset, and only uses the target dataset to learn how to use these features for AD prediction. This not only reduces the need for expert-defined language features, but also makes it possible for more complex features to be extracted from the data. The next advantage of sentence embedding models is that they consider the entire raw text and there is no out-of-context word embedding layer that would convert each word to a representation vector without considering its context.

As mentioned earlier, even using augmentation methods, the largest currently available dataset for AD prediction is still insufficient in size for unsupervised fine-tuning (Second phase specified in Section Pre-trained deep language model) large transformer-based language models (e.g., BERT$_{\mathrm{Large}}$ has 340 million parameters). But if there is a large enough dataset, using language model fine-tuning, our approach can extract more complex and context-related features while the models based on expert-defined features can only choose from a limited set of predefined features.

The most important limitation of the current study that needs to be addressed in the future is that it is difficult to use common neural network interpretation methods due to the large number of model parameters. Using interpretation, we can understand why the model predicts a wrong label for a transcript. Also, in the case of a correct prediction, we can identify language features that the network has paid more attention to. This is particularly useful for studying Alzheimer's disease as such interpretation can reveal important attributes of the speech which can most effectively discriminate between the participant groups.

Future work

One of the most popular types of transformer-based language models is the class of multilingual models. With a proper use of multilingual models, similar to approaches by Li et al. [26] and Fraser et al. [27], the problem of lacking access to a large dataset in one language can be addressed by transferring the knowledge of AD prediction from another language in which a large dataset is available. Using such transfer, the need to define linguistic features by experts in the target language is also addressed. In future work, we aim to improve multilingual AD prediction using pre-trained multilingual transformer-based language models along with cross-lingual transfer learning.

Conclusions

According to the results of earler studies, Alzheimer's disease affects speech in the form of syntactic, semantic, information, and acoustic impairments. We employed a transfer-learning approach to improve automatic AD prediction using a relatively small targeted speech dataset without using the expert-defined linguistic features. We evaluated recently developed pre-trained transformer-based language models that we enriched with augmentation methods on the Cookie-Theft picture description test of the Pitt corpus. Using sentence level $\mathrm{BERT}_{\mathrm{Large}}$ with a simple logistic regression classifier, the accuracy and F_1 scores of 88.08% and 87.23% were achieved which improved the state-of-the-art results by 2.28% and 2.80%, respectively. Pre-trained language models are available in many languages. Hence, the approach in this paper can be examined in languages other than English as well. Also, with the multilingual versions of these models, the knowledge of AD prediction in one language can be transferred to another language in which a sufficiently large dataset does not exist.

Declarations

Abbreviations

AD: Alzheimer's Disease; ADI: Alzheimer's Disease International; Bidirectional Encoder Representations from Transformers (BERT); CA: Contextual Augmentation; CNN: Convolutional Neural Network; XLM: Cross-lingual Language Model; XLNet: Extended Large Network; GRU: Gated Recurrent Unit; GloVe: Global Vector; HC: Healthy Control; LSTM: Long-short Term Memory; MLM: Masked Language Model; MCI: Mild Cognitive Impairment; MMSE: Mini-mental State Exam; NLP: Natural Language Processing; POS: Part-of-speech; RNN: Recurrent Neural Network; TLM: Translated Language Model; SVM: Support Vector Machine; SSA: Synonym Substitution Augmentation;

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data (Pitt corpus from DementiaBank dataset) that support the findings of this study are available from TalkBank project but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of TalkBank project owners.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Author's contributions

AR and MSB analyzed the data. HA conceptualized the work. All authors wrote, edited, and approved the manuscript.

Acknowledgements

Not applicable.

Author details

[1]Department of Computer Engineering, Sharif University of Technology, Azadi, Tehran, Iran. [2]Department of Electrical Engineering, Sharif University of Technology, Azadi, Tehran, Iran. [3]Department of Computer Engineering, Sharif University of Technology, Azadi, 11365-11155 Tehran, Iran.

References

1. Glenner, G.G.: Alzheimers disease. Biomedical Advances in Aging, 51–62 (1990)
2. International, A.D.: World Alzheimer Report 2019: Attitudes to dementia. Alzheimer's Disease Internationals London (2019)
3. Blanken, G., Dittmann, J., Haas, J.-C., Wallesch, C.-W.: Spontaneous speech in senile dementia and aphasia: Implications for a neurolinguistic model of language production. Cognition **27**(3), 247–274 (1987)
4. Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack Jr, C.R., Kaye, J., Montine, T.J., *et al.*: Toward defining the preclinical stages of alzheimer's disease: Recommendations

from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & dementia **7**(3), 280–292 (2011)

5. Reisberg, B., Sclan, S., Franssen, E., DeLeon, M., Kluger, A., Torossian, C., Shulman, E., Steinberg, G., Monteiro, I., McRae, T., *et al.*: Clinical stages of normal aging and alzheimers-disease-the gds staging system. Neuroscience Research Communications **13**, 51–54 (1993)
6. Mace, N.L., Rabins, P.V.: The 36-hour Day: A Family Guide to Caring for People Who Have Alzheimer Disease, Related Dementias, and Memory Loss. JHU Press, ??? (2011)
7. Ostuni, E., Santo Pietro, M.J.C.: Getting Through: Communicating When Someone You Care for Has Alzheimer's Disease. Speech Bin, ??? (1986)
8. Goodglass, H., Kaplan, E.: The assessment of aphasia and related disorders, 2nd edn lea & febiger: Philadelphia. Dictionary of Biological Psychology **230** (1983)
9. Mackenzie, C., Brady, M., Norrie, J., Poedjianto, N.: Picture description in neurologically normal adults: Concepts and topic coherence. Aphasiology **21**(3-4), 340–354 (2007)
10. Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J., McGonigle, K.L.: The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology **51**(6), 585–594 (1994)
11. Bucks, R.S., Singh, S., Cuerden, J.M., Wilcock, G.K.: Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology **14**(1), 71–91 (2000)
12. Thomas, C., Keselj, V., Cercone, N., Rockwood, K., Asp, E.: Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In: IEEE International Conference Mechatronics and Automation, 2005, vol. 3, pp. 1569–1574 (2005). IEEE
13. Guinn, C.I., Habash, A.: Language analysis of speakers with dementia of the alzheimer's type. In: 2012 AAAI Fall Symposium Series (2012)
14. Meilán, J.J.G., Martínez-Sánchez, F., Carro, J., López, D.E., Millian-Morell, L., Arana, J.M.: Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? Dementia and Geriatric Cognitive Disorders **37**(5-6), 327–334 (2014)
15. Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J.: Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 27–37 (2014)
16. Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C.A., Garrard, P.: Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology. Journal of Alzheimer's Disease **42**(s3), 3–17 (2014)
17. Orimaye, S.O., Wong, J.S.-M., Golden, K.J.: Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 78–87 (2014)
18. König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., *et al.*: Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **1**(1), 112–124 (2015)
19. Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., Roche-Bergua, A.: Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **10**, 260–268 (2018)
20. Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's Disease **49**(2), 407–422 (2016)
21. Yancheva, M., Rudzicz, F.: Vector-space topic models for detecting alzheimer's disease. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2337–2346 (2016)
22. Sirts, K., Piguet, O., Johnson, M.: Idea density for predicting alzheimer's disease from transcribed speech. arXiv preprint arXiv:1706.04473 (2017)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
24. Khodabakhsh, A., Demiroglu, C.: Analysis of speech-based measures for detecting and monitoring alzheimer's

disease. In: Data Mining in Clinical Medicine, pp. 159–173. Springer, ??? (2015)

25. Weiner, J., Herff, C., Schultz, T.: Speech-based detection of alzheimer's disease in conversational german. In: INTERSPEECH, pp. 1938–1942 (2016)
26. Li, B., Hsu, Y.-T., Rudzicz, F.: Detecting dementia in mandarin chinese using transfer learning from a parallel corpus. arXiv preprint arXiv:1903.00933 (2019)
27. Fraser, K.C., Linz, N., Li, B., Fors, K.L., Rudzicz, F., König, A., Alexandersson, J., Robert, P., Kokkinakis, D.: Multilingual prediction of alzheimer's disease through domain adaptation and concept-based language modelling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3659–3670 (2019)
28. Orimaye, S.O., Wong, J.S.-M., Fernandez, J.S.G.: Deep-deep neural network language models for predicting mild cognitive impairment. In: BAI@ IJCAI, pp. 14–20 (2016)
29. Karlekar, S., Niu, T., Bansal, M.: Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 701–707. Association for Computational Linguistics, New Orleans, Louisiana (2018)
30. Fritsch, J., Wankerl, S., Nöth, E.: Automatic diagnosis of alzheimer's disease using neural network language models. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5841–5845 (2019). IEEE
31. Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D., Christensen, H.: Automatic hierarchical attention neural network for detecting ad. Proc. Interspeech 2019, 4105–4109 (2019)
32. GLUE Benchmark. https://gluebenchmark.com/leaderboard. (Accessed on 03/14/2020)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
34. Al-Rfou, R., Choe, D., Constant, N., Guo, M., Jones, L.: Character-level language modeling with deeper self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3159–3166 (2019)
35. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
36. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
37. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems, pp. 7057–7067 (2019)
38. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, pp. 5754–5764 (2019)
39. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems, pp. 649–657 (2015)
40. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
41. Wang, W.Y., Yang, D.: That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2557–2563 (2015)
42. Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201 (2018)
43. Wu, X., Lv, S., Zang, L., Han, J., Hu, S.: Conditional bert contextual augmentation. In: International Conference on Computational Science, pp. 84–95 (2019). Springer
44. Ma, E.: NLP Augmentation. https://github.com/makcedward/nlpaug (2019)
45. Transformers — transformers 3.3.0 documentation. https://huggingface.co/transformers/index.html. (Accessed on 09/29/2020)
46. MacWhinney, B.: The CHILDES Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs. Psychology Press, ??? (2014)

47. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
48. Folstein, M., Folstein, S., McHugh, P.: Mini-mental state": a practical method for grading the cognitive state of patients for the clinician" j psychiatr res 12: 189–198. Find this article online (1975)

Figures

Figure 1 Overall classification procedure. The overall classification procedure contains the steps of augmentation, splitting, embedding, classification, and voting, where augmentation is only used in the training phase. Also, when passing the entire transcript to the embedding layer, the splitting and voting layers are disabled. The underlined models are trainable here, and the others are fixed.

Figure 2 Visualized tSNE dimensionality reduction for the $\text{BERT}_{\text{Base}}$ embedding of the entire Pitt corpus.

Figure 3 Mean 10-fold cross-validation classification accuracy, true positive rate, and true negative rate.

Tables

Table 1 Demographics of Cookie-Theft picture description test of the Pitt corpus.

	AD	**HC**
Participants	170	99
Samples	257	243
Age (years)	71.7±8.5	64.2±7.9
Gender (male/female)	87/170	88/155
Mini-Mental State Exam	18.6±5.1	29.1±1.1
# of Words	100.9±58.3	111.5±57.2

Table 2 The similarity between predicted health scores of S-BERT$_{Large}$-LR model and MMSE [48] scores.

Measure / Phase	Pearson Correlation	Spearman's Rank Correlation
Train	0.78	0.81
Validation	0.70	0.74

Table 3 AD versus HC classification scores.

Method	Embedding	Classifier	Precision	Recall	Accuracy	F$_1$
Fraser et al. [20]	35 Hand-Crafted Features	LR	-	-	81.92	-
Yancheva et al. [21]	12 Cluster-Based Features + LS&A	Random Forest	80.00	80.00	80.00	80.00
Sirts et al. [22]	Cluster+PID+SID Features	LR	74.4 ±1.5	72.5 ±1.2	-	72.7 ±1.2
Hernández et al. [19]	105 Hand-Crafted Features	SVM	81.00	81.00	79.00	81.00
Fritsch et al. [30]	One-Hot Word Embedding Sequence	-	-	**86**	85.6	-
Pan et al. [31]	GloVe Word Embedding Sequence	Bi-LSTM|GRU Hierarchical Attention	84.02	84.97	-	84.43
Li et al. [26]	185 Hand-Crafted Features	LR	-	-	77	-
Fraser et al. [27]	Info and LM Features	SVM	-	-	75	77
CNN + SSA	GloVe Word Embedding Sequence	CNN	76.38 ±8.49	77.47 ±8.97	76.48 ±5.88	76.36 ±5.91
BiLSTM + SSA	GloVe Word Embedding Sequence	Bi-LSTM	74.71 ±1.92	75.00 ±14.82	75.51 ±5.77	74.22 ±8.71
BiLSTM + CA	GloVe Word Embedding Sequence	Bi-LSTM	78.40 ±6.60	73.95 ±12.96	77.36 ±6.19	75.43 ±7.83
T-BERT$_{Base}$-LR	BERT$_{Base}$ (Text Level)	LR	85.09 ±3.11	78.69 ±8.35	82.76 ±3.74	81.51 ±4.73
T-BERT$_{Large}$-LR	BERT$_{Large}$ (Text Level)	LR	88.21 ±5.33	80.86 ±7.58	85.10 ±3.43	84.04 ±3.93
T-XLNet$_{Base}$-LR	XLNet$_{Base}$ (Text Level)	LR	84.74 ±6.31	79.26 ±7.72	81.92 ±5.88	81.75 ±6.19
T-XLNet$_{Large}$-LR	XLNet$_{Large}$ (Text Level)	LR	82.30 ±5.15	83.83 ±4.34	82.87 ±3.14	82.86 ±2.60
T-XLM-LR	XLM (Text Level)	LR	80.31 ±5.29	79.13 ±8.43	80.21 ±4.94	79.49 ±5.76
S-BERT$_{Base}$-LR	BERT$_{Base}$ (Sentence Level)	LR	90.31 ±7.36	76.52 ±8.06	84.46 ±6.31	82.72 ±7.21
S-BERT$_{Large}$-LR	BERT$_{Large}$ (Sentence Level)	LR	**90.57** ±3.18	84.34 ±7.58	**88.08** ±4.48	**87.23** ±5.20
S-XLNet$_{Base}$-LR	XLNet$_{Base}$ (Sentence Level)	LR	83.19 ±6.39	74.34 ±8.12	80.00 ±5.48	78.32 ±6.16
S-XLNet$_{Large}$-LR	XLNet$_{Large}$ (Sentence Level)	LR	76.95 ±6.62	71.30 ±8.29	75.31 ±5.56	73.75 ±6.14
S-XLM-LR	XLM (Sentence Level)	LR	84.00 ±4.74	73.47 ±9.80	80.21 ±5.47	78.14 ±6.72

Other settings of the proposed framework with different classifiers or augmenters which did not have significant effects on the scores are not shown.

Table 4 Two invalid predicted transcripts by the model with the best accuracy score (S-$\text{BERT}_{\text{Large}}$-LR).

Transcript	**Actual Label**	**Predicted Label**	**Predicted AD Probability**
And the boy in the cookie jar. And the girl reaching up to him. The stool slanting ready to topple. And the cookie jar is open. And the lid's in there. And the door's open. And mother's drying the dishes and standing in a pool of water it looks water running down from the sink. ...	AD	HC	0.483
Okay. It was summertime and mother and the children were working in the kitchen. And the window was open and there was a slight breeze blowing in. Mother was daydreaming and forgot and left the water in the sink running and it was overflowing. The children were hungry and ...	HC	AD	0.532

Predicted AD probability ranges between 0 and 1.