Classification of Alzheimer's Disease from Speech Data

Sierra Rowley, Usha Bhalla, Ally Zhu

Abstract

It is imperative that Alzheimer's Disease is caught in its early stages to prevent rapid progression, but it is often difficult to be diagnosed both quickly and inexpensively. Given that speech degradation is one of the earliest symptoms of AD, it has been suggested that neural nets can be used to classify speech data for AD. In this study, a network with a bi-directional GRU and four dense layers was trained on a relatively limited dataset from DementiaBank with 243 samples in each category. The model was found to have a mean and maximum accuracy of 0.63 and 0.825 when randomly tested 40 times, and an AUC ROC score of 0.654 when cross-validated. While these values are not ideal, they prove that using RNNs for AD diagnosing is promising.

Introduction

Alzheimer's disease (AD) among other dementias is a leading cause of death worldwide with more than 3 million cases in the US per year. Despite minimal available and effective treatment options, diagnosing a patient in the early stages of AD can be key to treatment, allowing for lifestyle and other changes that can prevent, slow, or lessen early and rapid symptom exacerbation and improve overall quality of life for those afflicted. However, diagnoses are often slow, expensive, or invasive because they frequently require medical imaging, family background, medical history, and extensive testing.

Machine Learning has been proposed as a potential tool to make improvements on this issue, because many of the early signs of AD are prevalent in a patient's speech patterns, such as confusion, repetition, stuttering, pauses, and decreased vocabularies. Furthermore, patients often suffer from mood and personality changes, which can be reflected in speech patterns as well. Compiling datasets of AD patients' language is highly feasible and machine learning techniques such as support vector machines and random forests have proven to be moderately successful classifiers of AD when trained on quantitative characteristics of speech alone (number of pauses, stutters, etc). As such, the use of neural networks for diagnosis should allow for accurate classification, as they can take sentiment, word usage, partial meaning, and more into account.

In 2019, Yi-Wei Chien et al.[1] published a paper classifying patients with and without AD by taking Chinese speech data from individuals and

converting them into monosyllable feature sequences which were then fed into a Recurrent Neural Network, which output percentage probabilities of Alzheimer's corresponding to the input labels of patient's Mini Mental State Exam scores. Their model was found to have a maximum area under the receiver operating characteristic curve score of 0.838 with a sensitivity of 0.75. For this study, speech transcripts from TalkBank's Pitt Dementia cookie dataset[2], in which patients with and without AD were asked to describe in English the picture below, were used to train the neural network. If models like this are effective, they could be used for frequent, rapid, and economical testing and monitoring of symptom onset and progression. Furthermore, these models could relatively easily be entirely automated, if an audio to transcript network was used before the classifier.

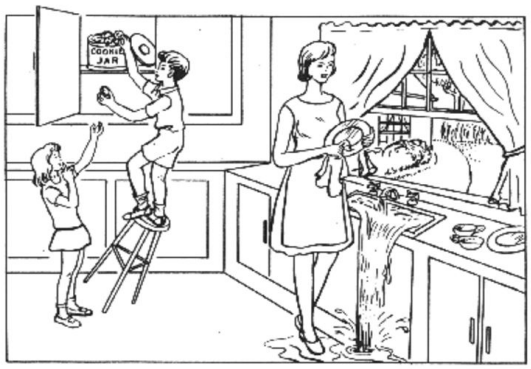
Preliminary testing of the model yielded an average categorical accuracy of 63.3% when evaluated 40 times, and a maximum accuracy of 82.5% on a testing set of 40 samples, as well as an accuracy of 0.654 when cross-validated with a specificity of 0.679 and a sensitivity of 0.625. While these values are lower than desired, they show promising results for the use of recurrent neural networks as AD classifiers.

Methodology

Data Collection:

The model was trained and tested on TalkBank's Pitt Dementia cookie dataset, which comprises 243 control samples and 309 dementia patient samples. Participants were asked to describe in English an image consisting of two children reaching for a cookie jar, and a woman to the right drying dishes in front of an overflowing sink. The audio files were transcribed by hand and contain specific characters representing pauses, stutters, and other speech patterns, all of which are documented by TalkBank. The transcripts also include the assessor's lines, parts of speech transcripts, and time stamps.

Image shown to patients:



Sample Fictional Transcript:*

*PAR: &um [\\] there's also water splashin(g) &=coughs onto &uh the floor. [+ exc] 000_6789
%mor: pro:exist|there~cop|be&3S adv|also n|water part|splash-PRESP prep|onto det:art|the n|floor .

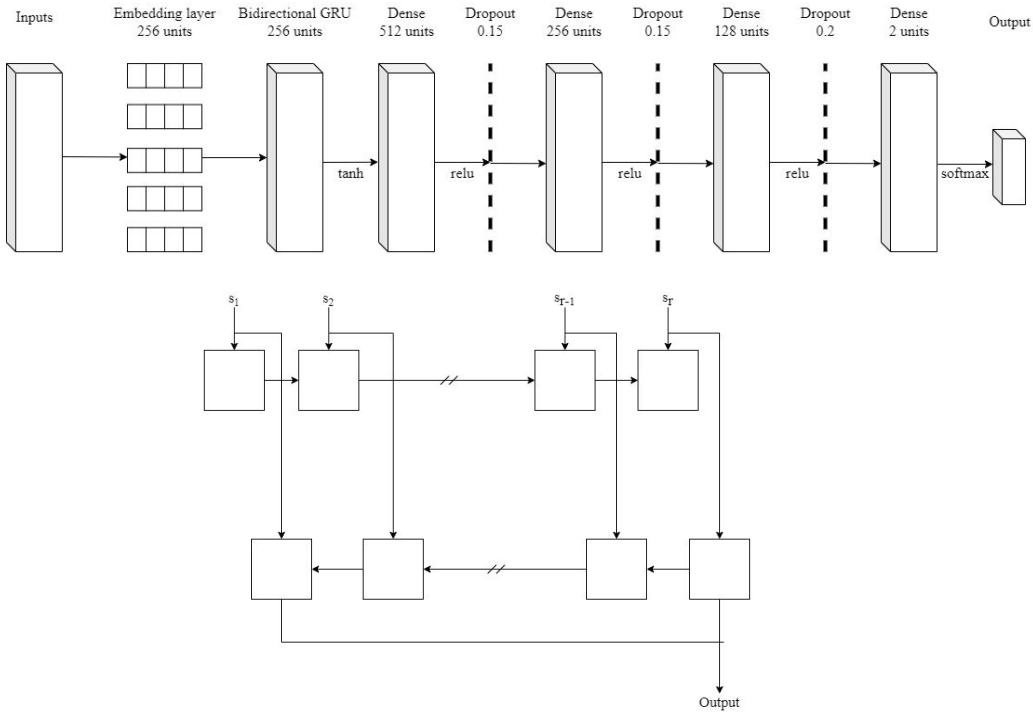
*Note: This transcript was not taken from the dataset, but was authored by this paper's authors.

***Data Preprocessing**:*
As seen above, the transcripts from the dataset contain very detailed information about patients' language, including notes about parts of speech, symbols representing speech patterns, and time stamps. Given that the dataset was very limited, there was a clear tradeoff between excess uncommon symbols and words in the training samples and loss of information of auditory symbols that could be associated with one category more than the other. As such, it was necessary to decide what information should be left in the transcripts to be tokenized, and what information should be removed to prevent too many uncommon tokens.
The time stamps, assessor's lines, and grammatical (denoted by lines beginning with %gra) and morphological transcripts (marked by %mor) were all stripped from the inputs, as well as any symbols or characters that did not represent repetition ([\], [\\]), common sounds (&=coughs, &=laughs, &um), or pauses ((.), (..), (...)). Following that, the data was tokenized and UNKed to remove any uncommon words; however, given that many words in the transcripts were spelled in accordance with the participant's pronunciation instead of their accurate spelling (i.e. 'hafta' instead of 'have to,' 'forget' instead of 'forget,' 'pictur' instead of 'picture'), this likely resulted in significant but unpreventable data loss. Finally, any bias in the data set was removed by removing excess Dementia samples, resulting in 263 control and treatment data points in total.

***Model Architecture**:*
The model consists of an embedding layer of embedding size 256 and vocabulary of 1700. The outputs of the embedding layer are fed into a bi-directional GRU with 256 units and a tanh activation. The outputs of the RNN layer are fed into three intermediate dense layers, the first of size 512, the second of size 256, and the third of size 128. All three layers had relu activation and dropout 0.15, except the third which had dropout 0.2. Finally, the output of the hidden layers is passed into a dense layer of size 2, which outputs the logits of the two classes, control and AD, and softmax is applied to yield the probabilities of the two categories.

Model Training:

The model was trained on the categorical accuracy of each training batch with an Adam optimizer and a learning rate of 0.001. The batch size was set to 100 and the model was trained for a total of ten epochs.

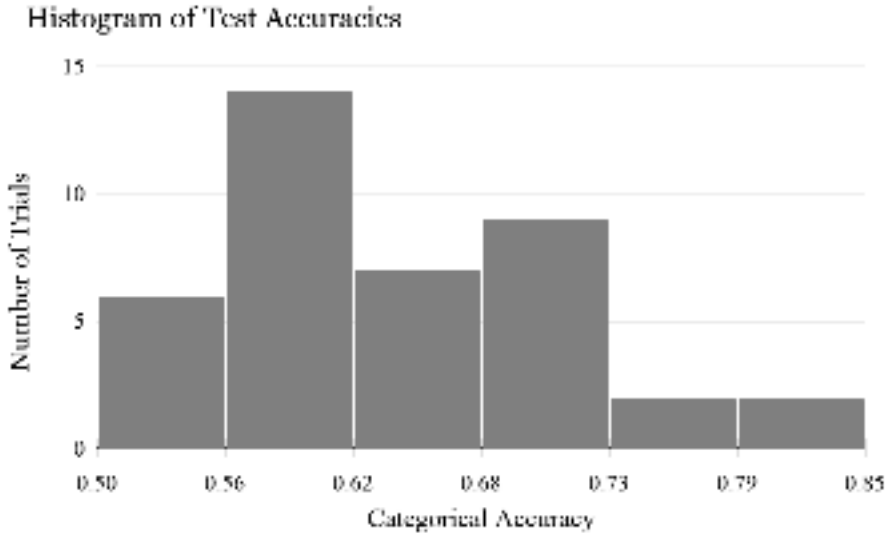


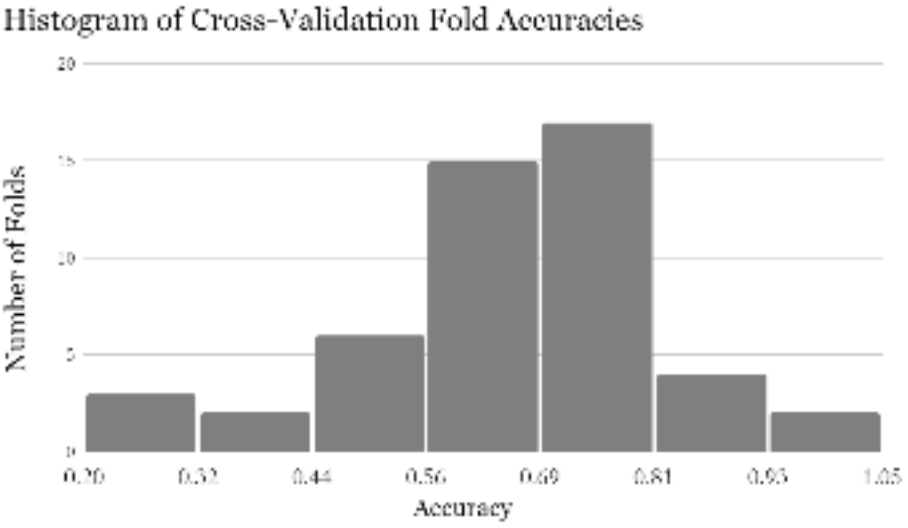
Results

The model was evaluated with two different methods, the first being 40 trials in which a random set of 40 samples were selected as the testing data and the rest were utilized as the training set. Accuracy was defined as the categorical accuracy between the model predictions and defined labels. Overall, the trials yielded an average accuracy of 0.633, a median of 0.61, and a maximum accuracy of 0.825. Training accuracies began at 0.5 and progressed to 0.85 over the ten training epochs, with categorical cross-entropy loss beginning at 0.7 and decreasing to 0.3.

The model was also evaluated with cross-validation. The data were initially shuffled and then 10 samples were selected sequentially as the testing data, and the rest were left for training. This process was repeated until all data points in the data set had been included in exactly one testing data set. The accuracy was calculated to be the average of these accuracies, equivalent to the accuracy of the model over the entire data set, at 0.654. Furthermore, the model was found to accurately predict 152 of the positive patients, and 165 of the control patients, with each set having a total of 243

data points. As such, the model had a sensitivity of 0.626 and a specificity of 0.679, with a false negative rate of 0.372 and false positive rate of 0.321.





Discussion

While the model results are much lower than necessary for clinical or commercial use, they prove that neural networks are successful in learning differences in audio transcripts of AD patients and control individuals. Given the limited size of the dataset, as well as the many inconsistencies within it, the results are lower than what can be expected from a larger, more standardized dataset. It is widely known that data is the most important requirement for deep learning models to accurately predict, and the increase in accuracy between the random trials and cross-validation evaluation further exemplify that. Given that the accuracy improved when the training dataset increased by only 30 samples, a large dataset would certainly perform much better. Furthermore, while the DementiaBank

dataset includes both audio files and transcripts, the audio files include lines from both the assessor and the participant, preventing the use of a Feature Sequence Generator model, which would allow for the process of physical data collection (i.e. recording individuals) to the output of the prediction to be entirely automated. Theoretically, a sequence generator would also increase the model's accuracy significantly, as the model could train on both syllables and word features. Finally, the DementiaBank data are labelled discretely as either 1 or 0; however, if a continuous labelling were to be used (i.e. based off of Mini Mental State Exam scores), the network would better be able to pick up on the progression of language degradation from non-afflicted patients to the different stages of the disease. A larger dataset with continuous labelling would also allow for training the model on minimally symptomatic or early stage AD patients, resulting in the model picking up on the nuanced changes in language in the beginning of the disease. This could potentially also yield a higher sensitivity than specificity, or at least a higher sensitivity than the model in this paper, which is important if diagnosis models were to be used in the future. Similarly, a dataset that tracked the progression of patients would allow for a model that could detect when exactly individuals exhibit symptoms. All in all, the results of this study are certainly promising, and prove that recurrent neural networks are able to distinguish between the speech data of patients with and without Alzheimer's Disease.

Challenges

Preprocessing: The DementiaBank data set came as individual files for each patient interview. Each file contained the patient's lines, the interviewer's lines, and other information such as metadata and specifics of the speech like physical actions and grammar. These files were transcripts of audio recordings and therefore contained a lot of information meant to help connect the transcripts to their audio files. For example, each spoken line ended in a time stamp and was followed by a set of post codes. We only wanted to use the patient's lines, so first we had to extract these and eliminate all other extra information. The formatting of each patient line was not the same, so it was unintuitive to find the time stamps and delete them. Another issue was that the data already had some preprocessing done to it, most likely to help match what actually occurred in the audio. Some examples are sounds were given their own symbol representation (ex. "(.)" represents a pause and "(..)" represents a longer pause) and babbling was written using a different alphabet. Also, shortened terms were rewritten as the full word with the ending in parentheses (ex. "havin" was turned into "hav(ing)") and certain phrases were encompassed by carrot symbols (< and >) to represent interruptions and repeating. Since we did not want this added punctuation to cause words to be treated differently in our vocabulary dictionary, we had to read through the data and strip any punctuation that was not used for standard practices. On the other hand, we wanted to keep

information about pauses and babbling since these could be important in distinguishing if a patient has AD. So, we created unique representations for each of these occurrences. Overall, there were a lot of anomalies in the punctuation of the transcripts and this caused our data to be difficult to work with.

Model Changes: Initially, we built the model based on a previous paper by Yi-Wei Chien which contained a 128 bi-directional GRU units and a single dense layer. However, that model only yielded an average accuracy of around 55%, only slightly better than a coin flip. The original implementation by Chien had arrived at 83% accuracy, and the disparity in our results was likely due to the difference in data content used, audio vs. transcript data formats, and inconsistent transcription in our own data.
We thus played around with different model specs, such as adding dropout layers, changing the activation between layers, and using single-direction GRU units. Training and validation loss had been hovering around the same level indicating potential underfitting, so we increased the network size of the GRU units and dense layers. This pushed accuracy to around 60% but was still highly unstable and had room for improvement.

Embedding matrix: Consequently, we turned to the vectorization of the language itself. We started by training the embedding matrix on the controlled data instead of on both dementia and control patients, and this was following the reasoning that control patients speech would generate more accurate and consistent vector representations. With that same reason in mind, we also experimented with external word vectorization databases, namely the Stanford GloVe database. Surprisingly, results between embeddings trained on control data vs. those trained on GloVe word vector database were not significant, and this is likely a result of transcription abnormality with potential for exploration. Overall, this ended up increasing mean accuracy to 63% and maximum accuracy to 82.5%, with training accuracy consistently reaching up to 85% after 10 epochs.

Reflection

While we did not achieve the exact results we hoped for, we are still happy with the outcome of our model. We started off with big goals of reaching above 90% test accuracy, however we ended with an average test accuracy of 63% and a maximum accuracy of 82.5%. Although this is not close to 90%, the 63% test accuracy indicates that the model was learning some information about the difference between a patient with and without AD. This shows that there is the possibility for a successful automatic AD detection model in the future. After we implemented the model from the original paper, the test accuracy was not any better than a coin flip. From there we changed our approach to building the model by trying things like changing hyper parameters and pre-training the embedding layer on

different datasets to raise the testing accuracy. If we could do the project over again, we would pick a different data set than the DementiaBank cookie dataset. We had a lot of problems with the formatting of the transcripts and that made training our model difficult. Also, this dataset was very small and this created large fluctuations in testing accuracy everytime we ran the model. If we had more time we would do more preprocessing to make the vocab dictionary more standardized. One of the biggest takeaways from this project is how applicable deep learning models can be on real world problems. This classification problem could be tremendously helpful to real people. Another takeaway is the difficulty of working with datasets. We didn't realize that the dataset we chose to use would have such complex formatting and this was a big obstacle in our project's process.

References

1. Chien, YW., Hong, SY., Cheah, WT. et al. An Automatic Assessment System for Alzheimer's Disease Based on Speech Using Feature Sequence Generator and Recurrent Neural Network. Sci Rep 9, 19597 (2019). https://doi.org/10.1038/s41598-019-56020-x
2. DementiaBank (Pitt corpus): Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology, 51(6), 585-594.