medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

EXPLAINABLE CNN-ATTENTION NETWORKS (C-ATTENTION NETWORK) FOR AUTOMATED DETECTION OF ALZHEIMER'S DISEASE*

Ning Wang
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
nwang7@stevens.edu

Mingxuan Chen
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
mchen20@stevens.edu

K.P. Subbalakshmi
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
ksubbala@stevens.edu

June 25, 2020

ABSTRACT

In this work we propose three explainable deep learning architectures to automatically detect patients with Alzheimer's disease based on their language abilities. The architectures use: (1) only the part-of-speech features; (2) only language embedding features and (3) both of these feature classes via a unified architecture. We use self-attention mechanisms and interpretable 1-dimensional Convolutional Neural Network (CNN) to generate two types of explanations of the model's action: intra-class explanation and inter-class explanation. The inter-class explanation captures the relative importance of each of the different features in that class, while the inter-class explanation captures the relative importance between the classes. *Note that although we have considered two classes of features in this paper, the architecture is easily expandable to more classes because of its modularity*. Extensive experimentations and comparison with several recent models show that our method outperforms these methods with an accuracy of 92.2% and F1 score of 0.952 on the DementiaBank dataset while being able to generate explanations. We show by examples, how to generate these explanations using attention values.

1 Introduction

In 2019, Americans spent $244B in caring for patients with Alzheimer's Disease and Related Dementia (ADRD). The National Academy of Sciences, the National Plan to Address Alzheimer's Disease, and the Affordable Care Act through the Medicare Annual Wellness, all identify earlier detection of ADRD as a core aim for improving the brain health for millions of Americans. The success of disease modification and preventive therapeutics for ADRD requires the identification of the disease in very early stages, at least a decade before onset. Approaches to early identification have included the use of brief cognitive screening tests and biological markers (usually neuroimaging or cerebrospinal fluid examination [1]). Neuroimaging modalities often include magnetic resonance imaging(MRI) [2] or the evaluation of positron emission tomography (PET) [3] targeting amyloid, Tau or both. The traditional biological marker methods tend to be invasive, expensive and create patient compliance problems. Hence, there is a strong motivation to consider early detection schemes using non-invasive markers of the disease.

*This paper has been accepted for publication in BIOKDD 2020.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In practice, cognitive assessment tools like Practitioner assessment of Cognition (GPCOG) [4] Cambridge Cognitive Examination (CAMCOG) [5], Mini-Cog [6], Mini-Mental State Examination (MMSE) [7] etc. are used to classify dementia and mild cognitive impairment (MCI) Even in primary care settings, cognitive impairment is unrecognized in 27%–81% of the affected patients [8].

As language functions play an important role in the detection of cognitive deficiency across different stages of ADRD, speech transcripts can assist in early detection of the disease. Hence techniques at the nexus of natural language processing and deep learning offer an inexpensive solution to this early detection problem [9, 10].

A Convolutional Neural Network (CNN) and Long Short Term Memory Network (LSTM) model (CNN-LSTM) was proposed in [11, 12], by using part-of-speech (PoS) tags to get accuracy up to 89.7% on classification for AD disease on DementiaBank cookie sub-corpus. Deep neural networks and deep language models were combined [13] to classifying the disease. On a sparse clinical language dataset, the model could predict MCI and AD type dementia with accuracy of 83%.

One of the key criticisms of deep learning (DL) methods is that the decision process of the DL model is intractable thereby making it difficult to understand the reasoning behind its decisions. For applications such as AD detection, it is imperative that some form of reasoning be provided, because of the human angle involved. Hence in this work, *we develop explainable deep learning architectures using attention mechanism and 1-dimensional CNN (1-D CNN) to detect Alzheimer's disease (AD) from transcripts of an individual's speech*.

The main contributions of our work are:

- novel explainable deep learning architectures for early detection of Alzheimer's disease
- inter and intra-feature attentions that captures the relative importance between the different feature classes as well as the relative importance of features within a class. These can be used to provide explanation of the model's conclusions
- modular architecture that can be extended to as many feature classes as desired
- extensive testing on the Dementiabank corpus, a popular Alzheimer's disease dataset
- example explanations using the proposed models

The rest of the paper is organized as follows. Section 2, describes related work. Section 3, discusses the proposed explainable deep learning models. Section 4, details the datasets, the experimental set up, data bias compensation mechanism, example explanations and discussions of the results. Section 5 discuses the conclusions we draw from this work.

2 Related Work

Automatically detecting Alzheimer's disease using transcripts of conversations is not new. Some previous efforts [14, 15] used linguistic features, such as PoS tags and syntactic complexity; psycholinguistic features (e.g, lexical and morphological complexity, word information, and readability etc.) to detect Alzheimer's disease. [15] used regression models to achieve an accuracy of 81% on classifying between AD patients and healthy controls. Some researchers have combined latent features obtained via language embeddings like word2vec [16] 'GloVe' [17] sentence2vector [18] along with hand-crafted features in a [19, 11, 12] hybrid approach using a CNN and RNN architecture to achieve accuracy around 89%.

While the above methods show varying degrees of accuracy, they do not provide adequate explainability in their models. Some researches have introduced visualizations to help move the research in the direction of explainability. Activation clusters and first derivative saliency heatmap were used in [11]. K-means clustering was used [20] to find the most frequently used topic in the text data. Note that explainable AD models have been developed for MRI based AD detection methods, such as [21] which proposed a gradient-based visualization method to detect the most important biomarkers related to AD and progressive mild cognitive impairment (pMCI). However, these are not directly applicable to the language based detection problem that we are considering here.

Meanwhile, work on explainable AI has started to emerge in importance specifically to address the problem of trustablity of AI systems. Multiple surveys have analyzed those works [22, 23, 24] which focus on different areas of explainable AI (XAI). [23] revisited visualization of CNNs and discussed trends in explainable artificial intelligence. [24] categorized prior work into model transparency and model functionality on multitude dimensions for model interpretability and analyzed the insufficiency of current work.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

An interpretable convolutional neural network (ICNN) was developed in [23] for use on images and tested on three benchmark image datasets. The structure of LSTM was explored and contribution of variables to the prediction in multi-variable time series was captured in [25]. [26] developed local interpretable model-agnostic explanations (LIME) to explain both the predictions and models. [27] presented a mechanism they call "deep learning important features(DeepLIFT)" to specify the contributions of all neurons in the network to every feature of the input. Meijer G-functions were used to disclose the functional forms learned by a model without much apriori assumptions [28]. However, none of these approaches provided precise explainable results and/or worked specifically in NLP domain.

In this work, we adapt the multi-head self-attention (MHA) proposed in [29] and use 1-D CNN [30] to define two types of explanations: intra and inter-feature explanations. These capture the relative importance between features within a set as well as between different class of features respectively.

3 Model and Method

We propose explainable AI models for detecting AD from speech transcripts of patients using attention mechanisms [29] and the 1-D CNN described in[30] to interpreting the decisions that the AI models make. We use two types of features in the model, the self-attention mechanism and a 1-D CNN to understand the relative importance of the features in the final outcome. There is a debate on whether attention mechanisms are good for interpretation [31, 32]. However, this debate was settled in favor of using self-attention mechanisms as a viable method for interpretation for classification tasks [33]. Moreover, 1-D CNN's can be used to interpret an AI's decision as demonstrated in [30, 34] for NLP tasks.

3.1 Proposed Architectures: C-Attention Networks

We propose three architectures: one that uses only PoS features, one that uses only the latent features (language embeddings) and a unified architecture, which uses both features.

3.1.1 C-Attention-FT Network

The architecture of the model proposed for exploiting PoS features (C-Attention-FT Network) is depicted on the left hand side of Figure 1. This architecture comprises of a self-attention module that captures the intra-feature relationships; an attention layer together with a following 1-D CNN layer that can be used to generate feature level explanations followed by a softmax layer. The MHA module is the same as that proposed in [29] for the popular transformer architecture and is presented in Sec 3.1.4. Let $R = \{r_1, r_2, ..., r_n\}$ be the set of records, then r_i is the i^{th} record in the dataset. We compute PoS tags for each record using NLTK [15]. Let $P = \{p_1, p_2, ..., p_n\}$ be the set of PoS feature vectors and p_i be the i^{th} vector in the PoS matrix. We use h Multi-Head-Attention (MHA) layers on $P = \{p_1, p_2, ..., p_n\}$ to capture the relationship between the PoS features. The MHA transforms $P = \{p_1, p_2, ..., p_n\}$ to another matrix of n-dimensional vectors $A = \{a_1, a_2, ..., a_n\}$. The MHA module is followed by a 1-layer CNN and a softmax layer to get the final classification.

3.1.2 C-Attention-Embedding Network

The architecture of the proposed C-Attention-Embedding Network is shown on the right hand side of Figure 1. We propose this architecture as a means of capturing latent feature information implicit in language embeddings. Specifically we use the universal sentence embedding (USE) [35] to represent each sentence in a record. This architecture is similar to the proposed C-Attention (Sec 3.1.1 except for the addition of a positional encoding module. The positional encoding module is used to maintain the relative positions of the sentences and is the same as that used in the transformer [29] architecture.

Let u_i be the USE vector corresponding to the i^{th} sentence in the record. The positional encoding is applied to each vector and the resulting vectors, u'_i, are used to construct the matrix $U = \{u'_1, u'_2, ..., u'_n\}$. An h-layer MHA module is used to extract the intra-feature relationships in this architecture. This is followed by an attention layer that captures interpretation at the embedding feature level. The output of the attention layer is fed to a 1-layer CNN and a softmax to get the final prediction.

3.1.3 Unified C-Attention Network

The third architecture we propose uses both the PoS and latent features of the sentences and is depicted in Figure 2. This architecture uses the proposed C-Attention-FT network and the C-Attention-Embedding network as two legs and combines them with another attention layer followed by a dense layer and the softmax layer. The dense linear layer is

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

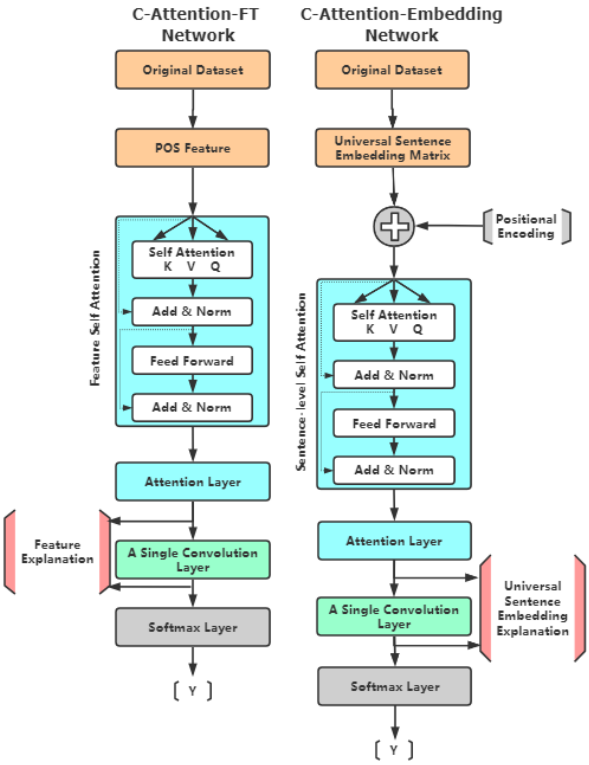


Figure 1: The proposed architecture of C-Attention-FT Network and C-Attention-Embedding Network. The C-Attention-FT network uses the PoS features and the C-Attention-Embedding network uses the sentence embeddings of the patient/control's description.

the same as that proposed in the transformer [29]. The attention layer captures the relative importance between the PoS and the USE features and helps in providing an interpretation at the feature class level.

3.1.4 Attention Mechanisms

Attention mechanisms have proved to be efficient in capturing global dependencies without any restrictions on the distance between the input and output sequences [36, 37]. Vaswani et al. [29] use self-attention [38] mechanisms along with positional coding in the design of the transformer which has become very popular in language models like BERT [39]. In this paper, we use the attention mechanisms and MHA mechanism proposed in [29]. These use a scaled dot product attention, which is given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K and V are the query, key and value matrices and d_k is dimension of the query and key vectors.

3.1.5 1-D Convolution Layer

We use a single layer 1-D CNN [40] as the penultimate layer, followed by a maxpooling layer which empowers the CNN to compress the information and extract global features. It was shown in [30] that 1-D CNNs filters essentially are n-gram detectors with each filter specializing in a closely-related family of n-grams. This feature can be used to interpret the action of the CNN layer in the architecture. We show how to trace back through these filters and the MHA layer to derive explanations for the classifiers outcomes.

4 Experiment and Result

We evaluate the proposed C-Attention Network architectures on the DementiaBank dataset and compare the performances of these architectures with each other as well as some recently published results [11, 12].

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

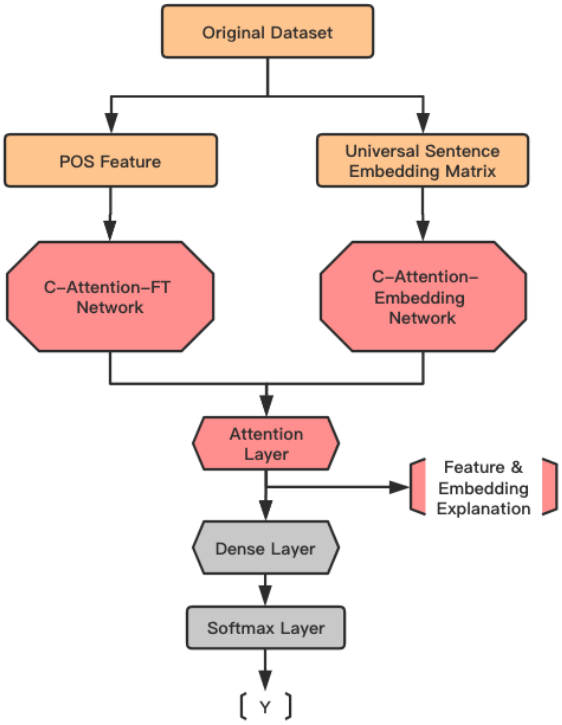


Figure 2: The Architecture of Unified C-Attention Network for Feature and Embedding

4.1 Data and Pre-processing

DementiaBank [41] is a database of multimedia interactions for the study of communications between people with dementia. Specifically the dataset contains transcripts of the description of a picture showing the theft of a cookie. This cookie theft corpus contains 1049 transcripts from 208 AD patients and 243 transcripts from 104 elderly control individuals for a total of 1229 transcripts.

4.2 Experiment Setup

We implemented our proposed model by using Pytorch. The model is trained to minimize the cross-entropy loss function of predicting the class label of participants' speech records in the training set. As mentioned earlier, two types of features were extracted: part of speech (PoS) [15] and sentence embedding. We used the USE proposed in [35] for the embedding feature. For all models in our experiments, we have 6 layers for the multi-head attention (MHA) module. We used stochastic gradient descent + momentum (SGD + Momentum) [42] as the optimizer for training. Since the cookie theft sub-dataset is unbalanced we added a class weight correction by increasing the penalty for misclassifying the less frequent class during model training to reduce the affect of data bias, as in [12]. The class weight correction ratio used in this paper is $7:3$. The average number of utterances in this dataset is 17. In order to have a fixed number of utterances in our model, we set the number of utterances as 17. We truncated extra utterances for descriptions that had more than 17 utterances and added padding for those with less than 17 utterances. Note that changing this number to the median number of utterances or the maximum number of utterances did not give us better results. We randomly split the original data into 81% training, 9% validation and 10% testing.

4.3 The Benchmarks

We compare the performance of our architectures with recently published results in terms of accuracy, precision, recall, F1 score, area under the curve (AUC). These comparisons are shown in Table 1. We also include total number of true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP) for completion. In addition to these baseline architecture, we also compare performances with our architectures with a slight modification. All of these architectures are described below.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Approach	Accuracy	Precision	Recall	F1	AUC	TN	FP	FN	TP
C-LSTM	0.8384	0.8683	0.9497	0.9058	0.9057	6.3	15.6	5.3	102.6
C-LSTM-ATT	0.8333	0.8446	0.9778	0.9061	0.9126	2.6	19.3	2.3	105.6
C-LSTM-ATT-W	0.8512	0.9232	0.8949	0.9084	0.9139	14	8	11.3	96.6
C-BILSTM	0.8495	0.8508	0.9965	0.9178	0.9207	1	16.6	0.3	95
C-BILSTM-ATT	0.8466	0.8525	0.9895	0.9158	0.9503	1.3	16.3	1	94.3
C-BILSTM-ATT-W	0.882	0.9312	0.9298	0.9305	0.9498	11	6.6	6.6	88.6
C-BILSTM-ATT-W NO PSYCH.	0.879	0.887	0.9825	0.9319	0.9499	12	5.6	1.6	93.6
C-BILSTM-ATT-W NO SENT.	0.897	0.9239	0.9615	0.9321	0.9501	7.6	10	3.6	91.6
C-BILSTM-ATT-W NO DEMO.	0.8908	0.9005	0.9789	0.9308	0.9473	10.33	7.33	2	93.3
Attention-FT	0.868	0.895	0.94	0.917	0.924	18	11	6	94
Attention-Embedding	0.822	0.881	0.89	0.886	0.824	17	12	11	89
Attention-FT+Embedding	0.829	0.882	0.90	0.891	0.828	17	21	10	90
C-Attention-FT	**0.922**	0.935	0.971	**0.952**	0.971	19	7	3	100
C-Attention-Embedding	0.845	0.885	0.92	0.902	0.837	17	12	8	92
C-Attention-FT+Embedding	0.915	**0.969**	0.922	0.945	**0.977**	23	3	8	95

Table 1: The comparison of performance metric between our models and others: C-LSTM, C-LSTM-ATT and C-LSTM-ATT-W were referenced from [11], their standard model kernel is Convolutional-LSTM; C-BILSTM, C-BILSTM-ATT, C-BILSTM-ATT-W, C-BILSTM-ATT-NO PSYCH., C-BILSTM-ATT-W NO SENT. and C-BILSTM-ATT-W NO DEMO. were referenced from [12], the main kernel of these architectures is Convolutional-BILSTM. The rest six models are our models: Attention-FT is attention model with only PoS features; Attention-Embedding is the one with only universal sentence embedding; Attention-FT+Embedding is the combination of those two parts. C-Attention-FT, C-Attention-Embedding and C-Attention-FT+Embedding are the similar architectures but replaced the dense layer by a convolutional layer. All these six models are based on Attention Mechanism.

Label	Speech Record	Important Sentences
0	okay, well the mother is drying the dishes, the sink is overflowing, um the little girl's reaching for a cookie, and her brother's taking cookies out of the cookie jar, and the stool is going to f knock him on the floor laughs, he's going to fall on the floor because the stool's not uh what, with gravity, whatever, uh the uh curtains are blowing I think, that's all I can see	um the little girl's reaching for a cookie with gravity he's going to fall on the floor because the stool's not uh what
1	I would like to have a lead pencil, the tree is blossoming, I hope my child doesn't hafta go to the hospital , I hope my child doesn't hafta go to the hospital, I shouldn't say that because we have a daughter who's pregnant, and I do want her to go to the hospital, okay then, this winter has been a very cold one, the doctor said I, I sat in the chair by a the doctor, brief, I'm not, I forgot to try make them brief, the bureau drawer stands open, ,	I would like to have a lead pencil I shouldn't say that because we have a daughter who's pregnant , ,
1	uh the pencil is on the desk, , leafing, meaning the leaves are opening, , cold q exc, and winter q exc, last year was a cold winter	and winter q exc , , cold q exc

Table 2: Samples of explanation for correct predictions of three speech records. We sort embedding sentences based on the attention value and show the top three sentences. The Label is the ground truth: 0 represents the healthy control; 1 represents the patient. Attention values means how much importance of these sentences are for prediction.

- **Attention-FT:** The Attention-FT architecture is a slightly modified version of our proposed C-Attention-FT architecture. In this version, we replace the CNN with the dense linear network used in [29].
- **Attention-Embedding:** The Attention-Embedding architecture is a modification of the C-Attention-Embedding architecture. Just as in Attention-FT, we replace the CNN with a dense linear network.
- **Attention-FT+Embedding:** This is a variant of the proposed Unified C-Attention-FT+Embedding architecture, with the two "legs" of the original architecture replaced by the Attention-FT and Attention-Embedding architectures described above.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

- **C-LSTM:** The C-LSTM architecture consists of a CNN followed by an LSTM layer. PoS and word embedding features are used [12].
- **C-LSTM-ATT:** This model is the same as the CNN-LSTM, but followed by an attention layer. The additional attention layer was used to detect specific linguistic patterns related to dementia detection not for explanations. This work did not correct for data bias [12].
- **C-LSTM-ATT-W:** This model is the same as the C-LSTAM-ATT, except the bias in the data is corrected by assigning class weights [12].
- **C-BILSTM:** The C-BILSTM architecture is built upon the C-LSTM architecture by using bi-directional LSTM instead of LSTM. This work used PoS, word embedding and other handcrafted features includes psycho-linguistic, average sentiment and demographic features [12].
- **C-BILSTM-ATT:** This model is the same as the C-BILSTM but followed by an attention layer. This work did not correct for data bias [12].
- **C-BILSTM-ATT-W:** This model is the same as the C-BILSTM-ATT, except the bias in the data is corrected by assigning class weights [12].
- **C-BILSTM-ATT-W NO PSYCH:** This model is the same as the C-BILSTM-ATT-W but without the psycho-linguistic features [12].
- **C-BILSTM-ATT-W NO SENT:** This model is the same as the C-BILSTM-ATT-W but without the average sentiment feature [12].
- **C-BILSTM-ATT-W NO DEMO:** This model is the same as the C-BILSTM-ATT-W but without the demographic features [12].

4.4 Performance Analysis

From the Table 1 we see that the best overall performance is achieved by both the proposed C-Attention-FT and unified C-attention-FT+Embedding architecture. The C-Attention-FT architecture performs best in terms of accuracy and F1 scores, and the C-attention-FT+Embedding performs best in terms of precision and AUC. Attention-Embedding does the worst in terms of accuracy, although its performance is not too bad in terms of precision recall and F1 scores. We also note that using attention only without a convolutional layer does not seem to result in the best performance as evidenced by the fact that all three modifications: Attention-FT, Attention-Embedding and Attention-FT+Embedding are not near the top in terms of most scores. Comparing only the Attention-FT and Attention-Embedding results, we note that the Attention-FT model performs better on all metrics, which seems to suggest that the PoS features (used in Attention-FT) may have more value than the latent features with it comes to using only attentions and no convolutional layers.

4.5 Explainability Analysis

The advantage of our model is that we can interpret the classification process of the model. Specifically, for each case, we can explain whether the model considers the PoS features or the latent feature as more important in its decision. Similarly, we can also use the self-attention weights of to determine the relative importance of utterances within a single picture description and the relative importance of the different PoS features in arriving at the final decision.

4.5.1 Explaining the Universal Sentence Embedding Features:

Table 2 shows some sample sentences with their corresponding ground truth (labels). Label 0 indicates that the description was uttered by a healthy individual (control) and Label 1 indicates that the description was uttered by a patient. In each of these examples the unified C-Attention-FT+Embedding network classified the utterance correctly. The column “Important Sentences" refers to the sentences that are captured by the attention and 1-D CNN layers, which indicates higher importance of those sentences.

According to the analysis on entire testing dataset (129 speeches), we have following findings:

- the sentences that are considered most important by the attention layer (have highest attention values) is almost always captured by the filters in the 1-D CNN layer. We notice that 121 speeches out of 129 speeches show this pattern.
- We also note that the intra-feature attention value for a patient’s utterances seems to be more uniformly distributed compared to that of a healthy control which shows a definite higher value for some sentences compared to others. This might indicate that the AI is picking up on the “randomness" of the utterances of the patients compared to that of a health control.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4.5.2 Explaining the role of Part of Speech Features:

PoS features are extracted at speech level. A total of 36 PoS tags are defined. Like the latent features we notice that the PoS features with the highest attention weight are 100% captured by the the filters in the following 1-D CNN layer. The top PoS features captured by attention layer and 1-D convolutional layers are shown in Figure 3. This result is obtained by analysing the entire testing dataset. It would indicate that the PoS features NNPS (proper noun, plural), MD (modal), EX (existential) and PRP (personal pronoun) are the most important accounting for 80.9% of times for the PoS Features. Our findings are consistent with previous works [43] that compared showed that AD patients tend to use more pronouns instead of nouns compared with healthy controls.

4.5.3 Explaining the relative importance of feature classes (PoS vs latent features):

The final attention layer of the proposed unified C-Attention-FT+Embedding architecture (Figure 2), captures the relative importance between the PoS features and the latent features in a decision. Table 4 shows these attention values for three patients (with their record numbers indicated). Record number 1188 and 982 correspond to AD patients that were correctly identified as patients by the C-Attention-FT+Embedding model. Record 182 corresponds to a healthy control (indicated by a label value of 0) and also correctly identified by the model. For example, for speech record number 1188, the latent features played a bigger role in determining the overall decision with a weight of 0.703. By contrast, for record number 182, the PoS feature seems to have weighed slightly more (0.572) than the latent features (0.428). Figure 4 shows that the "Part of Speech " leg is assigned a higher attention value in 65.1% of the cases, while Universal Sentence Embedding is assigned a higher attention value in 34.9% of the cases, indicating that the PoS features seem to play a higher role in detecting AD. This fact is also demonstrated in Table 1.

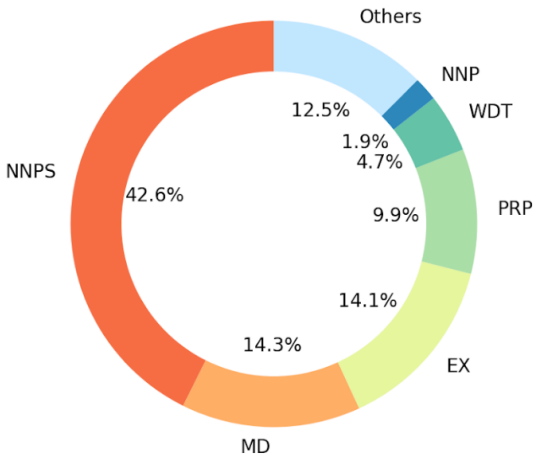


Figure 3: The Top PoS Features Captured by the C-Attention-FT+Embedding to make the final decision. The result is obtained by analysing on the entire testing dataset, including 129 speeches.

Tag	Meaning	Example
NNPS	proper noun, plural	Americans
MD	modal could	will
EX$	existential there	there is
PRP	personal pronoun	I
WDT	wh-determiner	which
NNP$	proper noun, singular	Harrisons

Table 3: Example of PoS tags

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Label	Record Number	PoS Weight	Embedding Weight
1	1188	0.297	0.703
0	182	0.572	0.428
1	982	0.157	0.843

Table 4: Attention Values for Part of Speech and Universal Sentence Embedding for three speech records numbered 1188, 182 and 982. Label 0 corresponds to a healthy control and Label 1 corresponds to an AD patient. In all cases, the C-Attention-FT+Embedding model identified the status of the individual correctly.

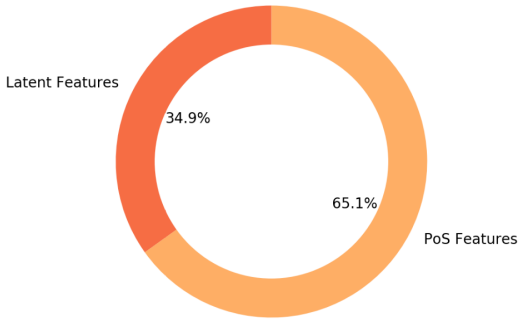


Figure 4: Attention value for Part of Speech leg is higher than Universal Sentence Embedding in 65.1% of cases over the entire dataset.

5 Conclusion

We proposed three explainable architectures using CNN and attention to detect Alzheimer's disease using two kinds of features: part-of-speech and language embeddings. One architecture uses only the PoS feature, one uses only the universal sentence embedding and the third is a unified architecture that uses both of these features. We propose the use of attention layers and 1-D CNN layer to capture explanations at 2 levels: one each at the intra-feature level and inter-feature-class level. The intra-feature level attention weights and 1-D CNN filters capture the relative importance the model places on the individual features in the category, whereas the inter-feature level attention weights gives us an idea of the relative importance that the model placed between the two classes of features.

Extensive testing on the popular DementiaBank datasets and comparisons with several recently published models as well as minor modifications of our own models show that **the C-Attention-FT architecture performs best in terms of accuracy and F1 scores, and the C-attention-FT+Embedding performs best in terms of precision and AUC while at the same time being able to generate explanations of the action of the AI**. We also show by examples how to generate explanations for the actions of the models. Our results agree with some of the previous work that shows that AD patient's tend to use more pronouns instead of nouns. Our work thus is an inexpensive, non-invasive, explainable AI model that can detect AD at good performance metric. Since it is based on only the spoken language, it can be potentially easily implemented in an app setting there by giving the option of taking it at home. This in turn can have a positive impact on patient compliance and therefore early detection of AD.

References

[1] M Pitschke, R Prior, M Haupt, and D Riesner. Detection of single amyloid β-protein aggregates in the cerebrospinal fluid of alzheimer's patients by fluorescence correlation spectroscopy. *Nature medicine*, 4(7):832–834, 1998.

[2] Ronald J Killiany, Teresa Gomez-Isla, Mark Moss, Ron Kikinis, Tamas Sandor, Ferenc Jolesz, Rudolph Tanzi, Kenneth Jones, Bradley T Hyman, and Marilyn S Albert. Use of structural magnetic resonance imaging to predict who will get alzheimer's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 47(4):430–439, 2000.

[3] Norman L Foster, Thomas N Chase, Paul Fedio, Nicholas J Patronas, Rodney A Brooks, and Giovanni Di Chiro. Alzheimer's disease: focal cortical changes shown by positron emission tomography. *Neurology*, 33(8):961–961,

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1983.

[4] Henry Brodaty, Dimity Pond, Nicola M Kemp, Georgina Luscombe, Louise Harding, Karen Berman, and Felicia A Huppert. The gpcog: a new screening test for dementia designed for general practice. *Journal of the American Geriatrics Society*, 50(3):530–534, 2002.

[5] B Schmand, G Walstra, J Lindeboom, S Teunisse, and C Jonker. Early detection of alzheimer's disease using the cambridge cognitive examination (camcog). *Psychological Medicine*, 30(3):619–627, 2000.

[6] Soo Borson, James Scanlan, Michael Brush, Peter Vitaliano, and Ahmed Dokmak. The mini-cog: a cognitive 'vital signs' measure for dementia screening in multi-lingual elderly. *International journal of geriatric psychiatry*, 15(11):1021–1027, 2000.

[7] Mary Galea and Michael Woodward. Mini-mental state examination (mmse). *Australian Journal of Physiotherapy*, 51(3):198, 2005.

[8] Cyndy B Cordell, Soo Borson, Malaz Boustani, Joshua Chodosh, David Reuben, Joe Verghese, William Thies, Leslie B Fried, Medicare Detection of Cognitive Impairment Workgroup, et al. Alzheimer's association recommendations for operationalizing the detection of cognitive impairment during the medicare annual wellness visit in a primary care setting. *Alzheimer's & Dementia*, 9(2):141–150, 2013.

[9] Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon K Wilcock. 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.

[10] Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, 53(1):1–19, 1996.

[11] Sweta Karlekar, Tong Niu, and Mohit Bansal. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*, 2018.

[12] Flavio Di Palo and Natalie Parde. Enriching neural models with targeted features for dementia detection. *arXiv preprint arXiv:1906.05483*, 2019.

[13] Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Chee Piau Wong. Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PloS one*, 13(11), 2018.

[14] Santosh Kumar Barnwal and Uma Shanker Tiwary. Using psycholinguistic features for the classification of comprehenders from summary speech transcripts. In *International Conference on Intelligent Human Computer Interaction*, pages 122–136. Springer, 2017.

[15] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[18] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[19] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. Detecting signs of dementia using word vector representations. In *INTERSPEECH*, pages 1893–1897, 2018.

[20] L JabaSheela, Shwetha Vasudevan, and VR Yazhini. A hybrid model for detecting linguistic cues in alzheimer's disease patients. *Journal of Information and Computational Science, 10(1), 85-90.*, 2020.

[21] Kanghan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1):1–16, 2019.

[22] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[23] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.24.20139592; this version posted June 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[24] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6. IEEE, 2017.

[25] Tian Guo, Tao Lin, and Nino Antulov-Fantulin. Exploring interpretable lstm neural networks over multi-variable data. *arXiv preprint arXiv:1905.12034*, 2019.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[27] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[28] Ahmed M Alaa and Mihaela van der Schaar. Demystifying black-box models with symbolic metamodels. In *Advances in Neural Information Processing Systems*, pages 11301–11311, 2019.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[30] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[31] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[32] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[33] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*, 2019.

[34] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. *arXiv preprint arXiv:1809.08037*, 2018.

[35] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[36] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[37] Maria Yancheva and Frank Rudzicz. 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)*, pages 2337–2346, 2016.

[38] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

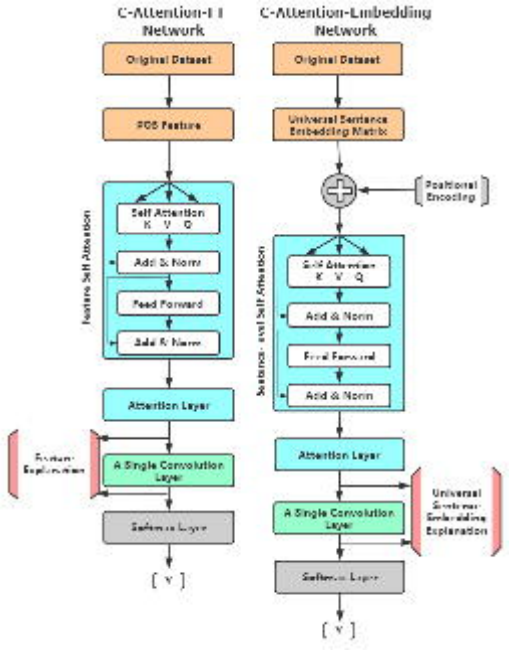
[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

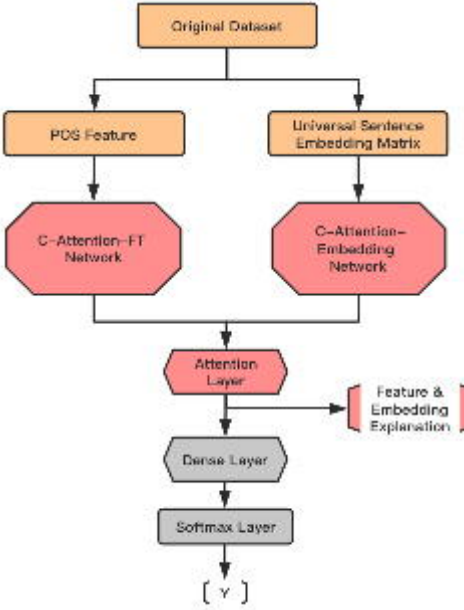
[40] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

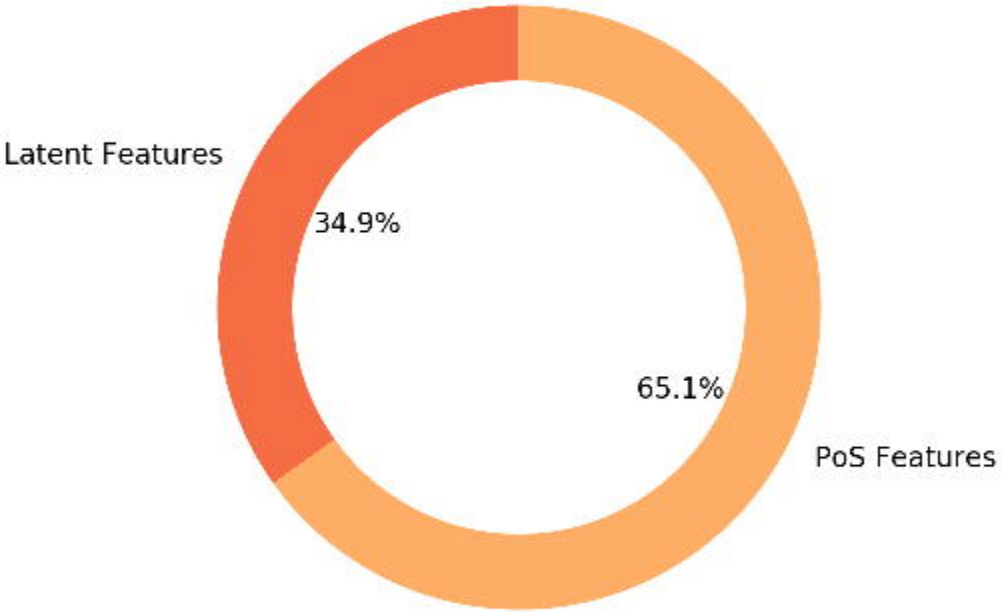
[41] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

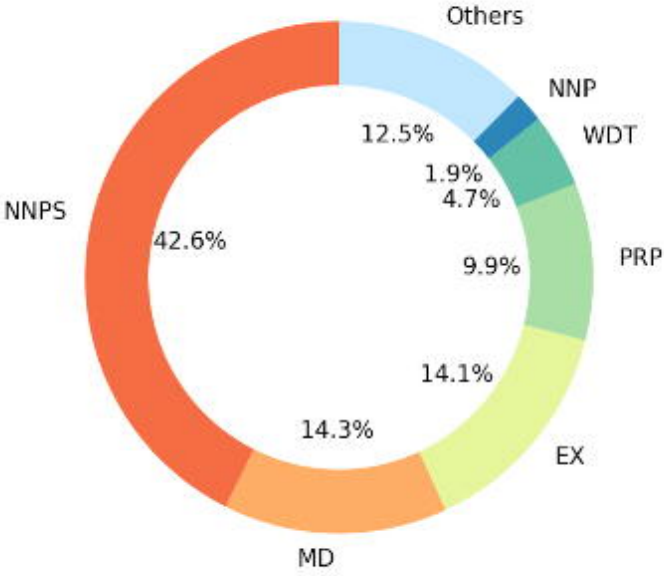
[42] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[43] William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 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*, pages 27–37, 2014.


C-Attention-I I Network
Original Dataset
POS Feature
Feature Self Attention
Self Attention
K V Q
Add & Norm
Feed Forward
Add & Norm
Attention Layer
A Single Convolution Layer
Softmax Layer
[Y]
C-Attention-Embedding Network
Original Dataset
Universal Sentence Embedding Matrix
Positional Encoding
Sentence-level Self Attention
Self Attention
K V Q
Add & Norm
Feed Forward
Add & Norm
Attention Layer
A Single Convolution Layer
Universal Sentence Embedding Explanation
Softmax Layer
[Y]


Original Dataset
POS Feature
Universal Sentence Embedding Matrix
C-Attention-FT Network
C-Attention-Embedding Network
Attention Layer
Feature & Embedding Explanation
Dense Layer
Softmax Layer
[Y]


Latent Features
34.9%
65.1%
PoS Features


Others
NNP
WDT
PRP
EX
MD
NNPS
12.5%
1.9%
4.7%
9.9%
14.1%
14.3%
42.6%