Renxuan Li April 13, 2020

Hierarchical Transformer for Early Detection of Alzheimer's Disease

by

Renxuan Li

Jinho Choi

Adviser

Department of Computer Science

Jinho Choi

Adviser

Michelangelo Grigni

Committee Member

Hiram Maxim

Committee Member

2020

Hierarchical Transformer for Early Detection of Alzheimer's Disease

By

Renxuan Li

Jinho Choi

Adviser

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Department of Computer Science

2020

Abstract

Hierarchical Transformer for Early Detection of Alzheimer's Disease

By Renxuan Li

Alzheimer's disease is an irreversible disease that severely affect the brain functions and life quality of the patients. For now, there is no effective cure for the disease. Therefore, this unfortunate fact makes the early detection of Alzheimer's disease vital. The early stage of the Alzheimer's disease, Mild Cognitive Impairment (MCI), normally involve loss in memory, language ability, and object recognition ability. In this paper, we present a new dataset that includes the transcribed audio of the MCI patients and healthy subject. We also present a hierarchical transformer-based model and the corresponding analysis for the MCI/health classification task on our dataset

Hierarchical Transformer for Early Detection of Alzheimer’s Disease

By

Renxuan Li

Jinho Choi

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Department of Computer Science

2020

Acknowledgements

I would like to thank professor Jinho Choi for his consistent support and advising throughout the entire project. I would also like to thank Felicia Goldstein and Ihab Hajjar of Emory School of Medicine for collecting the data for this project. Moreover, I would like to give my thanks to the Phd students in Emory NLP labs who offered me help and guide in the experiment process. Last but not least, I would like to thank my friends and roommates for their help and support during the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Alzheimer's disease (AD) is a progressive neuro-degenerative disorder that is associated with loss of functional autonomy. The symptoms normally include memory loss and declines in several major brain functions including language and object recognition. [3][10] Current diagnosis methods are usually time-consuming and expensive since the methods requires often time lengthy clinical tests and usage of medical instruments with high precision such as MRI. As a result the fact that seniors are way more likely to develop Alzheimer's Disease[3] and the rapid increase in global life expectancy and aged population [11], the lengthy and expensive diagnosis process may cause increasing pressures on global public health system. In addition, Alzheimer's disease is known as irreversible and not curable [4]. With recent advances in the field of Natural Language Processing(NLP), attempts have been made to develop simpler and faster AD detection methods based on the language of the potential patients. Meanwhile, to help developing tools and models for AD detection, the speech of the patients and healthy people have been recorded, transcribed into texts, and packed into datasets. One of the best known datasets is the DementiaBank from University of Pittsburgh[1]. The DementiaBank collects recorded speech of patients with and without dementia, and also includes annotated grammatical features.

The irreverible nature of the disease makes the detection of the early stage of the disease vital. The early stage of the Alzheimer's disease, also known as Mild Cognitive Impairment (MCI), is usually associated with language change and impairment in reasoning ability. MCI is more difficult to identify, since it's the very beginning stage of AD and therefore the degradation of language may not be very obvious. In this study, we present the B-SHARP dataset that focuses on MCI detection. The dataset is consists of plain text transcripts of MCI and normal subjects' speech on three different tasks including events recall, environment description, and picture description. We also perform experiments on the dataset and build a fully automated hierarchical transformer-based model based on the hierarchy established by our three tasks in the dataset. The model achieves 74% accuracy on five-fold cross validation. In addition, we perform analysis to comprehend what the words are and are picked out by the model, and analyze which questions in the speech protocol is more important.

Chapter 2

Related Works

2.1 Detection of Alzheimer's Disease

There have been studies conducted to work on the detection of Alzheimer's disease (AD) or its early stage, known as Mild Cognitive Impairment(MCI), using the speech or speech transcripts of the patients with early stage Alzheimer's disease. In particular, the current widely studied dataset, the DementiaBank (Becker:1994) , is collected by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. The dataset is consists of audio recordings and transcripts of the recorded audios of elderly adults. Subjects are divided into two groups: 1. Dementias, who are clinically diagnosed with Alzheimer's disease or Dementia. 2. Controls, who are healthy elderly adults without Dementia related diseases. Each subject visits on a yearly basis and contributes a new transcript to the dataset. Any error that occurred during transcribing process is corrected. Moreover, for each sentence, corresponding grammar structures and information are also added to the transcriptions. Research has been conducted on the DementiaBank to help the detection of AD. Orimaye et al (Orimaye:2014) uses several Machine Learning Algorithms on the syntactic and lexical features of the transcripts to build predictive model that achieved F-measures over 74%. Pou-Prom and Rudzicz(multiview) use the linguistic features of the transcription to learn a multi-view embedding for AD and achieved F1 score of 0.82 in classification task.

While the approaches involving extracted language features do achieve good performance, the annotation process is time consuming, and potentially restricts our understanding of the language of patients with MCI. Many researchers have explored approaches of AD detection using neural networks and plain text/audio. Karlekar et al (Karlekar:2018) have proposed a CNN-LSTM network approach and achieved around 85% accuracy on DementiaBank. They have also done some important visualization and analysis to understand the reasons behind the conclusion of the neural network. Attention Mechanism, a recent significant breakthrough in NLP research, is also used to detect dementia in some studies. Di Palo and Parde(clstm+att) have proposed a model with both plain text and handcrafted features based on C-LSTM and attention mechanism. They have successfully pushed the F1 score of DementiaBank to 0.929.

While most of the previous work focuses on the DementiaBank dataset, there are also some studies that try to utilize new data. Choi et al (choi-etal-2019-meta) propose a new Meta-Semantic Representation to predict early stage of AD. The study uses 100 transcripts from audio recordings collected by Emory University School of Medicine. The 100 transcripts are also the subset of the dataset we use in this study.

2.2 Transformers for Natural Language Understanding

BERT(Bidirectional Encoder Representations from Transformers) (devlin-etal-2019-bert) is a variant of the transformer model that is consisted of 12 or 24 layers of the transformer encoders depends on the size of BERT model. BERT is pre-trained with English Wikipedia on several tasks, and can be fine-tuned to downstream NLP tasks. The BERT model achieved state-of-art performance in 11 tasks,including document classification, in the GLUE benchmark(glue). Therefore we would like to experiment using the BERT model on our dataset, and hope to get some new insights or inspirations. RoBERTa (liu-etal-2020-roberta) has the same structure with the BERT model, but RoBERTa uses different training approaches and outperforms BERT on GLUE benchmark test. While the BERT model and the RoBERTa model perform very well on many NLP tasks, the huge sizes of the models also entail strict memory constraints and longer and more difficult fine-tuning processes. The ALBERT (lan-etal-2019-albert) model, which adopt two parameter reducing techniques, reduces the model size to be 17 times smaller than regular BERT model while pushing the GLUE score to the new state-of-art level of 89.4. Both RoBERTa and ALBERT model make some improvement to the original BERT model, and we also hope that these two models can provide us with better result or some inspirations.

In addition, as we will explain in details later in the DataSet Section, our dataset is consists of several tasks, which makes our dataset inherently suitable to a more hierarchical model. Several models have been proposed for hierarchical document summarization from which we take some inspirations. Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization(HIBERT) (zhang-etal-2019-hibert) is a transformer-based model on two levels: the sentence level and the document level. The authors also perform large-scale pre-training on HIBERT models to achieve state-of-art performance on two major datasets in document summerization field. Although the main focus of our study is not document summarization, we can still use the ideas in this paper to build hierachical models using BERT, RoBERTa, and ALBERT models.

Chapter 3

Dataset

3.1 B-SHARP: Brain, Stress, Hypertension, and Aging Research Program

Data from 326 subjects collected as part of the **B**rain, **S**tress, **H**ypertension, and **A**ging **R**esearch **P**rogram, *B-SHARP*, are used for this work [1]. 185 cognitively normal controls and 141 patients with Mild Cognitive Impairment (MCI) were selected based on neuropsychological and clinical assessments. Every subject has been examined with two well-known standardized cognitive tests, the Montreal Cognitive Assessment (MoCA; nasreddine:05a) and the Boston Naming Test (BNT; kaplan:83a), and followed the speech task protocol for voice recording (Section 3.2).51.5% and 23.9% of the subjects thus far came back for their 2nd and 3rd visits to take new voice recording, respectively.

[1] B-SHARP: http://medicine.emory.edu/bsharp

This is an ongoing program; voice recordings of 20-25 subjects are taken every month such that the data is still growing.

Group	**Subjects**	**2nd Visits**	**3rd Visits**	**Recordings**	**MoCA**	**BNT**
Control	185	100	50	385	26.2 (±2.6)	14.2 (±1.2)
MCI	141	68	28	265	21.5 (±3.5)	13.4 (±1.5)
All	326	168	78	650	24.2 (±3.8)	13.9 (±1.4)

Table 3.1: Statistics of the control and MCI groups. Subjects: # of subjects; 2nd/3rd Visits: # of subjects who made 2nd/3rd visits; Recordings: # of voice recordings; MoCA/BNT: average scores and standard deviations from MoCA/BNT; Note that subjects with the 2nd/3rd visits take one/two additional voice recordings respectively; thus, Recordings = Subjects + 1·(2nd Visits) + 2·(3rd Visits).

Table 3.1 shows the statistics of the control and the MCI groups. The scores from both MoCA and BNT between these two groups show significant difference ($p < 0.0001$; Welch's t-test).

Note that when subjects make multiple visits, there is a year gap in between so that subjects do not necessarily remember much from their previous visits. Thus, recordings from the same subject are not any more similar than ones from the other subjects. In fact, most recordings across subjects, regardless of their groups, are very similar when they are transcribed into text since all subjects follow the same speech protocol.

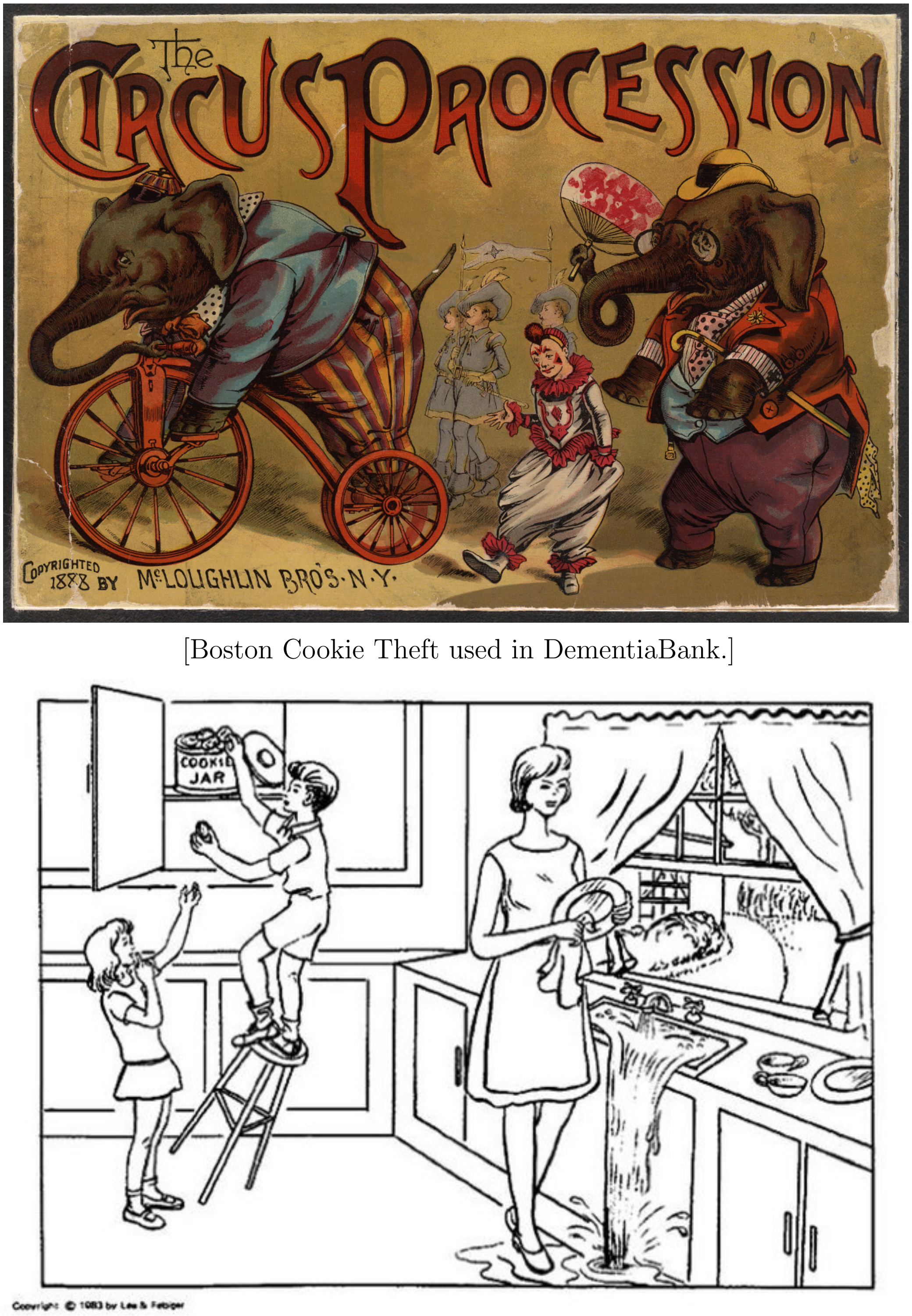
3.2 Speech Task Protocol

A speech task protocol has been conducted to collect voice recordings of the subjects who are asked to speak about Q1: today's activity, mostly different psychological and medical tests done on subjects; Q2: environment description, in this question subjects are asked to describe the item they see in the examination room; Q3: picture description. The picture is shown in figure3.1. For each task, subjects are asked to speak for 1-2 minutes. All subjects are provided with the same instructions in Table 3.2, and visual abilities of the subjects are confirmed before recording. To reduce potential variance, the subjects are guided to follow similar activities before Q1, located to similar room settings before Q2, and shown the same picture in Figure3.1, "The Circus Procession" copyrighted by McLoughlin Brothers as part of the Juvenile Collection, for Q3.

Type	1c**Instruction**
3*1	I would like you to describe to me everything we did from the moment we met today until now.
	Please try to recall as many details as possible in the order the events actually happened where we met,
	what we did, what we saw, where we went, and what you felt or thought during each of these events.
2	I would like you to describe everything that you see in this room.
2*3	I am going to show you a picture and ask you to describe what you see in as much detail as possible.
	You can describe the activities, characters, and colors of things you see in this picture.

Table 3.2: Instructions of Q1, Q2, and Q3 provided to the subjects.

The collected voice recordings are automatically transcribed by the online


The Circus Procession
Copyrighted 1888 by McLoughlin Bro's. N.Y.
Cookie Jar
Copyright © 1983 by Lea & Febiger

[The Circus Procession used in B-SHARP.]

[Boston Cookie Theft used in DementiaBank.]

Figure 3.1: The pictures used in (on the top) and DementiaBank (on the bottom).

tool, Temi.[2] The transcripts are then processed by the open-source NLP toolkit called ELIT[3]. We use ELIT to analyze various linguistic features about this dataset. As shown in Table 3.3, transcripts from the control group show significantly higher numbers of tokens, nouns, and complex structures while ones from the MCI group show a significantly higher number of discourse elements on average, implying that the control subjects give more expressive descriptions while the MCI subjects include more disfluency in their speeches.

2c——	1c—**Tokens**	1c—**Sentences**	1c—**Nouns**	1c—**Verbs**	1c—**Conjuncts**	1c—**Complex**	1c**Discourse**	
2*1	Control	186.6 (±60.4)	10.4 (±4.5)	28.1 (±9.6)	30.4 (±11.5)	8.5 (±4.5)	2.3 (±1.7)	8.1 (±5.4)
	MCI	175.6 (±54.5)	9.8 (±4.1)	23.7 (±8.3)	29.3 (±10.4)	8.5 (±4.2)	2.0 (±1.6)	9.2 (±6.0)
2*2	Control	191.5 (±11.8)	11.7 (±4.7)	41.1 (±13.3)	24.3 (±11.2)	6.6 (±4.5)	3.6 (±2.7)	7.1 (±4.8)
	MCI	178.6 (±11.7)	11.6 (±4.7)	36.7 (±12.1)	23.2 (±10.6)	6.4 (±4.4)	2.9 (±2.3)	8.4 (±5.3)
2*3	Control	193.4 (±63.4)	12.6 (±5.4)	39.5 (±13.5)	28.4 (±10.1)	8.0 (±4.8)	3.3 (±2.1)	6.1 (±5.5)
	MCI	187.8 (±63.4)	12.7 (±5.1)	36.2 (±13.2)	27.7 (±10.9)	7.2 (±4.2)	2.6 (±2.0)	7.3 (±5.5)
3*All	Control	**578.1** (±149.8)	34.5 (±10.7)	**110.5** (±27.9)	84.2 (±25.4)	23.5 (±10.1)	**9.3** (±4.5)	21.4 (±13.0)
	MCI	548.7 (±140.6)	34.0 (±10.5)	98.1 (±26.1)	81.2 (±24.1)	22.5 (±9.7)	7.7 (±4.2)	**25.3** (±15.0)
	p	1c—0.0110	1c—0.5541	1c—< 0.0001	1c—0.1277	1c—0.2046	1c—< 0.0001	1c0.0006

Table 3.3: Average counts and standard deviations of linguistic features per transcript in the B-SHARP dataset. Complex: the occurrences of complex structures such as relative clauses or non-finite clauses, Discourse: the occurrences of discourse elements such as interjections or disfluency.

3.3 Comparison to DementiaBank

DementiaBank is currently the largest public dataset that comprises audio recordings and their transcripts for four language tasks, picture description, verbal fluency, story recall, and sentence construction, from a large longitudinal

[2] Temi (Transcriber): https://www.temi.com

[3] (NLP Toolkit): https://github.com/elitcloud/elit

study [1]. Subjects in this study are divided into two groups, normal controls and dementia patients. Among the four tasks, data from only the picture description task can be used for classification since the other tasks give data of dementia patients only.[4] The design of this task is similar to 1 in (Section 3.2); each subject is shown the "Boston Cookie Theft" picture in Figure 3.2 to describe for 1-2 minutes.

Group	**Subjects**	**2nd Visits**	**3rd Visits**	**4th Visits**	**5th Visits**	**Recordings**
Control	99	29	28	9	8	243
Dementia	194	53	13	8	3	309
All	293	82	41	17	11	552

Table 3.4: Statistics of the control and the dementia groups in DementiaBank. Note that subject with i'th visits take $(i-1)$ additional recordings; thus, Recording = Subjects + $\sum_{i=2}^{5}(i-1)'thVisit$.

Table 3.4 shows the statistics of DementiaBank in comparison to Table 3.1. Subjects in this study made up to 5 visits compared to 3 in B-SHARP although the number of subjects in each visit is larger in B-SHARP. B-SHARP has ≈100 more recordings than DementiaBank, more importantly, B-SHARP is still growing, which makes it the largest dataset for NLP research related to the detection of Alzheimer's Disease. Unlike DementiaBank where 66.2% of the subjects are dementia patients, 43.3% of the subjects belong to the MCI group in B-SHARP; this makes sense because MCI is closer to the pre-clinical

[4] The verbal fluency task gives 1 audio recording of a normal control, that is still not enough to train classification models.

phase that has a much fewer number of patients reported in general.

	1c—**Tokens**	1c—**Sentences**	1c—**Nouns**	1c—**Verbs**	1c—**Conjuncts**	1c—**Complex**	1c**Discourse**
Control	124.0 (±59.7)	12.6 (±5.1)	**23.7** (±11.8)	**27.1** (±11.9)	2.8 (±2.8)	1.6 (±1.6)	1.5 (±1.6)
Dementia	114.3 (±61.3)	12.1 (±6.4)	18.7 (±10.4)	23.9 (±12.9)	2.4 (±2.4)	1.4 (±1.4)	**2.8** (±2.9)
p	0.0625	0.3204	< 0.0001	0.0029	0.0715	0.1184	< 0.0001

Table 3.5: Average counts and standard deviations of linguistic features per transcript in DementiaBank.

Table 3.5 shows the statistics of linguistic features in comparison to Table 3.3. The same tool, ELIT (Section 3.2) is used to measure these statistics. Unlike B-SHARP, the control group in DementiaBank does not reveal a significantly greater number of tokens than the dementia group. The document size in DementiaBank is 4.9 times smaller than B-SHARP on average. In both datasets, the noun and discourse counts are significantly different between the control and the other groups. It is interesting that a significant difference is found in verbs whereas it is not the case for complex structures in DementiaBank, which is opposite in B-SHARP. This may indicate that the verb usage deteriorates as it progresses from MCI to dementia, but more thorough research is needed for further verification.

Chapter 4

Approaches

4.1 Baselines Approaches

4.1.1 Convolutional Neural Network(CNN)

The CNN for document classification we experiment comes from the paper Convolutional Neural Networks for Sentence Classification[6]. In addition to the original design, we add filters of larger sizes. In this approach, we use FastText Embeddding and keep the word embedding constant over the training process. Figure 4.1 from Kim's paper provide a good visualization of the architecture of the model.

4.1.2 CNN-LSTM

CNN-LSTM hybrid model is initially proposed by [5] on the DementiaBank dataset and achieves state-of-art performance on DementiaBank. The input text is embedded before going to the network, then goes into convolutional

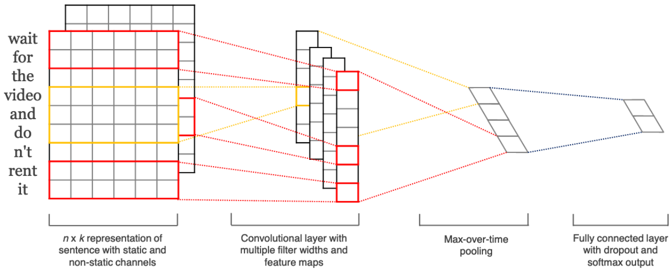

wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 4.1: CNN for sequence classification

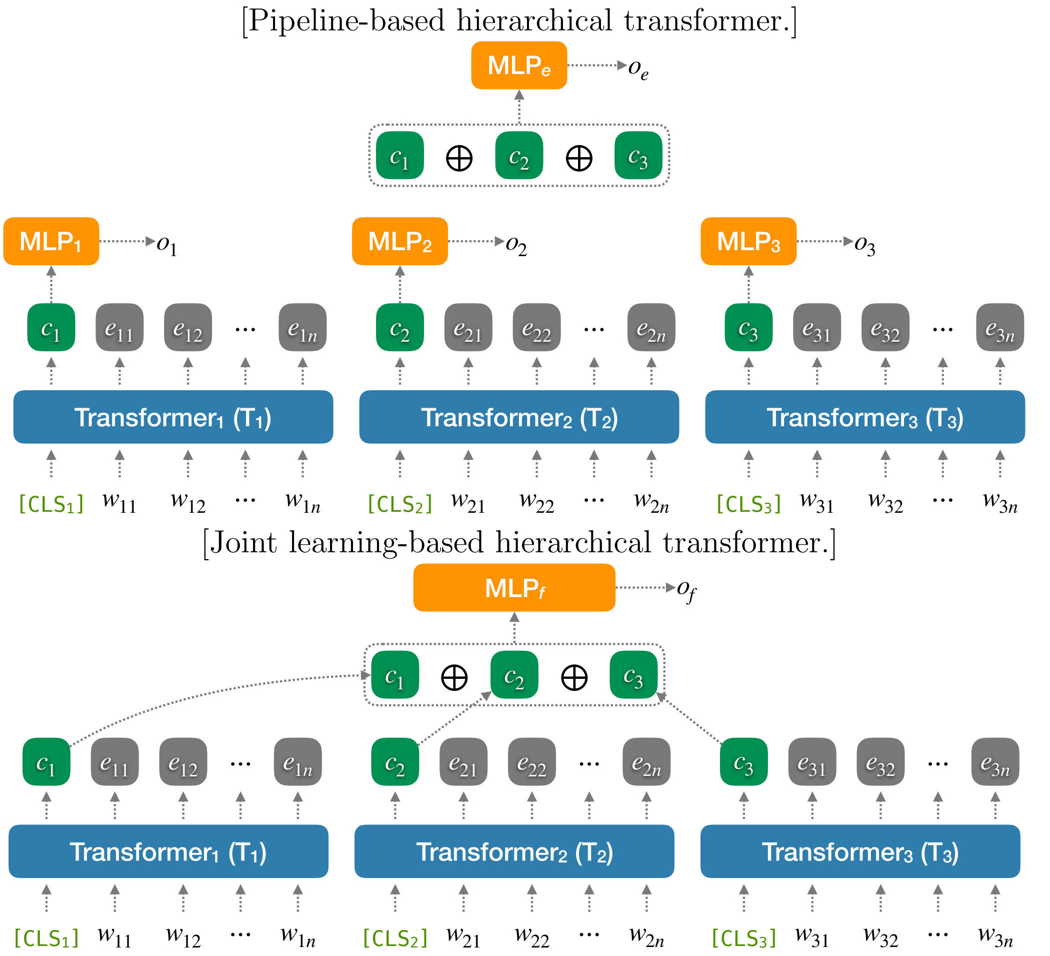
layers to extract local features. After the convolutional layers, the input is sent to a Long-Short-Term Memory(LSTM) network to capture relationship of longer terms.

4.2 Hierarchical Transformer

Although transformers have established the state-of-the-art results on most document classification tasks (Section 2.2), they have a limitation on the input size as a result of their huge memory requirement. As shown in Table 3.1, the average number of tokens in our input documents well-exceeds 512 when combining transcripts from the three tasks (Section 3.2), which is the maximum number of tokens that the pre-trained models of these transformers generally recommended.[1]This makes it difficult to simply join all transcripts

[1] As a matter of fact, their internal tokenizers such as WordPiece [12] or SentencePiece [7] make further segmentation; thus, the maximum number of input tokens is actually smaller than 512.

together and feed into a transformer. Thus, this section presents two types of hierarchical transformers to handle this long-document issue.



[Pipeline-based hierarchical transformer.]

[Joint learning-based hierarchical transformer.]

MLPₑ
o_e
c_1 ⊕ c_2 ⊕ c_3
MLP₁
o_1
MLP₂
o_2
MLP₃
o_3
c_1 e_{11} e_{12} ⋯ e_{1n}
c_2 e_{21} e_{22} ⋯ e_{2n}
c_3 e_{31} e_{32} ⋯ e_{3n}
Transformer₁ (T₁)
Transformer₂ (T₂)
Transformer₃ (T₃)
[CLS₁] w_{11} w_{12} ⋯ w_{1n}
[CLS₂] w_{21} w_{22} ⋯ w_{2n}
[CLS₃] w_{31} w_{32} ⋯ w_{3n}
MLP_f
o_f
c_1 ⊕ c_2 ⊕ c_3
c_1 e_{11} e_{12} ⋯ e_{1n}
c_2 e_{21} e_{22} ⋯ e_{2n}
c_3 e_{31} e_{32} ⋯ e_{3n}
Transformer₁ (T₁)
Transformer₂ (T₂)
Transformer₃ (T₃)
[CLS₁] w_{11} w_{12} ⋯ w_{1n}
[CLS₂] w_{21} w_{22} ⋯ w_{2n}
[CLS₃] w_{31} w_{32} ⋯ w_{3n}

Figure 4.2: Hierarchical transformers to combine the three types of transcripts in ensemble.

Figure 4.2 describes the pipeline-based hierarchical transformer, that is useful when no computational resource is available to fit three transformer models. Let $W_i = \{w_{i1}, \ldots, w_{in}\}$ be the input document where w_{ij} represents the j'th token in the transcript from the i'th task Qi (in our case, $i = \{1, 2, 3\}$). W_i is prepended by the special token $\texttt{[CLS}_i\texttt{]}$ that is used to learn the document embedding, and fed into the transformer T_i. The transformer then generates $E_i = \{c_i, e_{i1}, \ldots, e_{in}\}$, where c_i and e_{ij} are the contextualized embeddings for $\texttt{[CLS}_i\texttt{]}$ and w_{ij}, respectively. Finally, $c_i \in R^d$ is used to make two types of predictions, where d is the dimension of embeddings generated by T_i. First, c_i is directly fed into a multilayer perceptron layer, $\texttt{MLP}_i$, that generates the output vector $o_i \in R^2$ to predict whether or not the subject has MCI based on the transcript from Qi alone. Second, c_i is concatenated with the document embeddings generated by the transformers trained for the other tasks such that $c_e = c_1 \oplus c_2 \oplus c_3 \in R^{3d}$, which gets fed into another $\texttt{MLP}_e$ to generate the output vector $o_e \in R^2$ and makes the binary decision based on the transcripts from all three tasks, Q1, Q2 and Q3.

Note that the input to $\texttt{MLP}_e$ is the document embeddings generated by $T_{1,2,3}$ that are already trained so that what is learned from this ensemble does not get passed all the way to the transformer level. Figure 4.2 describes the joint learning-based hierarchical transformer that optimizes all three transformer models together with the predictions made by the ensemble. Transcripts from the three tasks $W_{1,2,3}$ are fed into the transformers to generate the document embeddings $c_{1,2,3}$, which get concatenated and fed into $\texttt{MLP}_f$ to generate $o_f \in R^2$ for the binary decision. Our hypothesis is that the joint learning approach tends to give more robust results than the pipeline approach since the transformers are optimized by the features from other documents as well as their owns; however, it requires more powerful computational resource that may be too costly.

Chapter 5

Experiments

5.1 Data Split

There is a total of 650 audio recordings in our dataset (Table 3.1), which is rather small to divide into training, development, and test sets. Thus, 5-fold cross-validation is used to evaluate the performance of our models. Table 5.1 shows the distributions of the 5 cross-validation sets used in our experiments, where the transcript of each recording is treated as an independent input document.

	Recordings						**Subjects**					
	$\mathtt{CV}_0$	$\mathtt{CV}_1$	$\mathtt{CV}_2$	$\mathtt{CV}_3$	$\mathtt{CV}_4$	ALL	$\mathtt{CV}_0$	$\mathtt{CV}_1$	$\mathtt{CV2}_2$	$\mathtt{CV}_3$	$\mathtt{CV}_4$	ALL
Control	77	77	77	77	77	**385**	37	37	37	37	37	**185**
MCI	53	53	53	53	53	**265**	27	28	28	29	29	**141**

Table 5.1: Distributions of the cross-validation sets for our experiments. $\mathtt{CV}_i$: the i'th set. ALL: $\sum_{i=0}^{4} \mathtt{CV}_i$.

It is worth mentioning that recordings from the same subject are never distributed across different cross-validation sets. In other words, all recordings from the same subject are assigned to the same set in our approach so that there is no overlap in terms of subjects between these cross-validation sets. This prevents potential inflation in accuracy due to the unique language patterns used by individual subjects.

5.2 Transformers

Three types of transformers are used to encode the input documents, BERT [2], RoBERTa [9], and ALBERT [8], which have shown the state-of-the-art performance in many natural language understanding tasks recently (Section 2.2). Table 5.2 shows the configurations of these transformers used in our experiments.

Transformer	Type	Layers	Attention Heads	Input Cells	Hidden Cells	Parameters
BERT	Base	12	12	768	768	108M
RoBERTa	Base	12	12	768	768	125M
ALBERT	Base	12	12	768	128	12M

Table 5.2: Configurations of BERT, RoBERTa, and ALBERT for our experiments.

BERT and RoBERTa are very similar in nature although RoBERTa uses a larger number of parameters.]These transformers are used to develop models for individual tasks (Section 5.3) as well as the ensemble approaches (Section 5.4). All three models are initially loaded with pre-trained parameters that are available publicly. The pre-trained parameters and implemented transformer are provided by huggingface[1]. Every model is trained three times and its average performance with the standard deviation is reported for robust evaluation.

5.3 Performance on Individual Tasks

Table 5.3 shows the model performance in terms of accuracy, sensitivity, and specificity from the three transformer models on the individual tasks. The performance on 2 shows the highest accuracy for all three models, achieving 69.9% with RoBERTa, implying that the environment descriptions involving many spatial relations in 2 are more effective in distinguishing the MCI group than the other two tasks. The highest sensitivity of 57.1% is achieved by BERT on 2 whereas the highest specificity of 86.8% is achieved by ALBERT on 3. Such a low sensitivity and a high specificity indicate that it is relatively easy to recognize the normal controls but not the MCI patients from short speeches. We will explore how to make this task conversational. Besides the pre-trained model, we also fine-tune the BERT model on the language model build on B-SHARP dataset, and train the fine-tuned BERT model. However, language model fine-tuning does not give us any significant improvement

[1] huggingface transformers: https://github.com/huggingface/transformers

on the performance of model. Language model fine-tuning usually allows transformer models to pick up new vocabularies from given datasets, and the fact that language model fine-tuning does not improve the model performance indicate that the vocabularies used in B-SHARP datasets are mostly common words.

	——**BERT**			——**RoBERTa**			**ALBERT**		
	1c—**1**	**2**	**3**	**1**	**2**	1c—**3**	**1**	**2**	**3**
`ACC`	67.6 (±0.4)	**69.0** (±1.2)	67.7 (±0.7)	69.0 (±1.5)	**69.9** (±0.2)	65.2 (±0.3)	67.6 (±1.5)	**69.5** (±0.3)	66.6 (±1.3)
`SEN`	48.9 (±1.8)	**57.1** (±2.5)	41.5 (±3.6)	44.3 (±4.5)	**55.3** (±1.2)	37.1 (±3.7)	45.9 (±1.9)	**52.2** (±0.6)	37.4 (±3.3)
`SPE`	80.4 (±1.2)	77.3 (±2.8)	**85.2** (±3.0)	**85.8** (±2.1)	79.7 (±0.7)	84.5 (±3.0)	82.6 (±3.7)	81.4 (±0.3)	**86.8** (±3.3)

Table 5.3: Model performance on the individual tasks. `ACC`: accuracy, `SEN`: sensitivity, `SPE`: specificity.

5.4 Performance with Ensemble Approaches

Table 5.4 shows the model performance of the ensemble models. We see that B+R+A ensemble model achieves the best performance overall, with highest accuracy of 74.07% and highest sensitivity of 60.88%. B+R ensemble achieves accuracy over 72%. We also perform experiment of different R+A ensemble, which achieves accuracy of 71.49%, which is not as good as B+R ensemble. We also notice that the joint learning approach does not bring better performance as we expected. The accuracy of ALBERT joint learning is only 68.34%, which is lower than Pipeline style ALBERT ensemble. This may be the result of the size of our datast is small compared to size of model parameters, and thus it's harder for the model to fit on the dataset nicely.

	CNN	**BERT**$_e$	**RoBERTa**$_e$	**ALBERT**$_e$	**ALBERT**$_f$	**B**$_e$ + **R**$_e$	**B**$_e$ + **R**$_e$ + **A**$_e$
`ACC`	69.49(±0.24)	69.90(±1.13)	71.60(±1.46)	69.75(±2.88)	68.34(±1.59)	72.21(±0.71)	74.07(±0.32)
`SEN`	49.18(±0.79)	57.61(±3.42)	48.55(±6.13)	46.16(±8.31)	44.28(±1.86)	56.48(±2.46)	60.88(±5.23)
`SPE`	83.46(±0.91)	77.36(±4.80)	87.48(±1.82)	85.39(±0.47)	86.11(±2.63)	83.09 (±0.93)	84.01(±2.43)

Table 5.4: Performance of ensemble models.

Given the limited GPU resource we have, only ALBERT is experimented with the joint-learning approach; meanwhile, all transformers are experimented with the pipleline-based hierarchical transformer The result for LSTM and CNN-LSTM is not included in the table, because the models do not fit on to the dataset very well, and thus their performances are not comparable to the models shown above.

Chapter 6

Analysis

6.1 Attention Analysis

In this section, we investigate how the transformer models (BERT, RoBERTa, ALBERT)are tuned to our MCI classification tasks. We modify the code from bertviz[1] project take the attention scores that corresponds to the [CLS] token. We use the following method to extract the more attend tokens:

6.1.1 Methods and Measures

First, for transcript T, we first tokenize the transcript into a list of tokens $[t_0, t_1 ... t_{|T|}]$ where t_0 is the [CLS] token added by the transformers tokenizer for classification tasks. Then we feed the encoded tokens into the pre-trained model and get Attention Matrix Att_p, notice the pre-trained model is not fine-tuned to our MCI classification task as we mention in 5.2. We also feed

[1] bertviz: https://github.com/jessevig/bertviz

the encoded tokens into out fine-tuned transformer model and get Attention Matrix Att_f. Notice the transformers have been trained and perform well on MCI classification task. For both attention matrix, we only take the last layer $Att_p[12]$ and $Att_f[12]$. The attention matrices have 4 dimensions. The first dimension represents the layer of attention, second dimension represents attention head in the model. The third and fourth layer are the attention vectors for each token. Each row represents the softmax normalized attention distribution of a token. For the attention matrix, we are only interested in the attention of [CLS] token, which is directly fed into MLP layer for classification. We define the following two attention score for one token t_i:

1. pre-trained attention score(PAS):

$$att_{t_i}^{pre-trained} = \frac{\sum_{h \in H} Att_p[12][h][0][t_i]}{|H|} \tag{6.1}$$

1. MCI-classification fine-tuned attention score (FAS):

$$att_{t_i}^{fine-tuned} = \frac{\sum_{h \in H} Att_f[12][h][0][t_i]}{|H|} \tag{6.2}$$

where H is the set of attention heads. In our case, RoBERTa model and BERT model both have 12 attention heads.

With the definition of the two type of attention score, we propose **Attention Change Ratio(ACR)** and use ACR scores to measure how much our fine-tune process change the attention to a token. The ACR score for token t_i is calculated the following way:

$$ACR_{t_i} = \frac{att_{t_i}^{fine-tuned} - att_{t_i}^{pre-trained}}{att_{t_i}^{pre-train}} \tag{6.3}$$

The reason we use ACR in this analysis instead of directly using FAS scores is that both FAS and PAS scores contain the information of the pre-trained transformer model, which is trained on language models. Therefore, the token with highest FAS or PAS may have higher attention scores just because the token is more important in the language model. While for ACR scores, we only measure how much the attention scores change after the fine-tuning process, and we can be certain that the tokens with higher ACR scores are directly related to how the transformer models classify the MCI/Normal subjects.

6.1.2 RoBERTa Analysis

In this part, we randomly choose one control transcript and one MCI transcript from each cross validation set. We only choose samples that all three individual

CV set	0	1	2	3	4
MCI transcript	V00205 YR1	V00058 YR3	V00189 YR1	V00427 BL	V00369 BL
Normal transcript	V00278 BL	V00114 YR1	V00196 BL	V00003 YR3	V00282 YR1

Table 6.1: RoBERTa attention analysis samples

RoBERTa models correctly classified. The sampled transcripts for RoBETRa analysis are shown in table 6.1.

Question 1 Model Analysis

In this part, we analyze the Individual RoBERTa model that is trained only on Q1 part of transcripts. We first calculate the ACR scores and rank the tokens their ACRs from highest to lowest. By simply looking at the tokens with highest ACR scores, we cannot recognize any very obvious pattern. However, we can still notice some words that appear to be important in several samples. The words include **room**, **question**,**cake**,**lady**. To explore the pattern that are not very obvious, we use V00003 YR3 question 1 to perform a more detailed analysis. To make three individual models directly comparable, we will continue to use V00003 YR3 to make heat chart for question 2 RoBERTa model and question 3 RoBERTa model. We take 10 tokens with highest ACR scores for each of 4 POS categories: Nouns, Verb, Adjective and Adverb if a category has less than 10 tokens selected, we will append with lowest scores.

Then we plot the tokens into a heat chart. In figure 6.1, we see the most attended POS category is nouns. We also notice Adverb category is also attended. In addition, we also perform visualization on MCI sample. We use V00427 BL as sample, and to be consistent, we also use this sample for Q2 and Q3. In figure 6.2, we can see the pattern is approximately same as the pattern in figure 6.1, excerpt that there are slightly fewer attended tokens for each category.

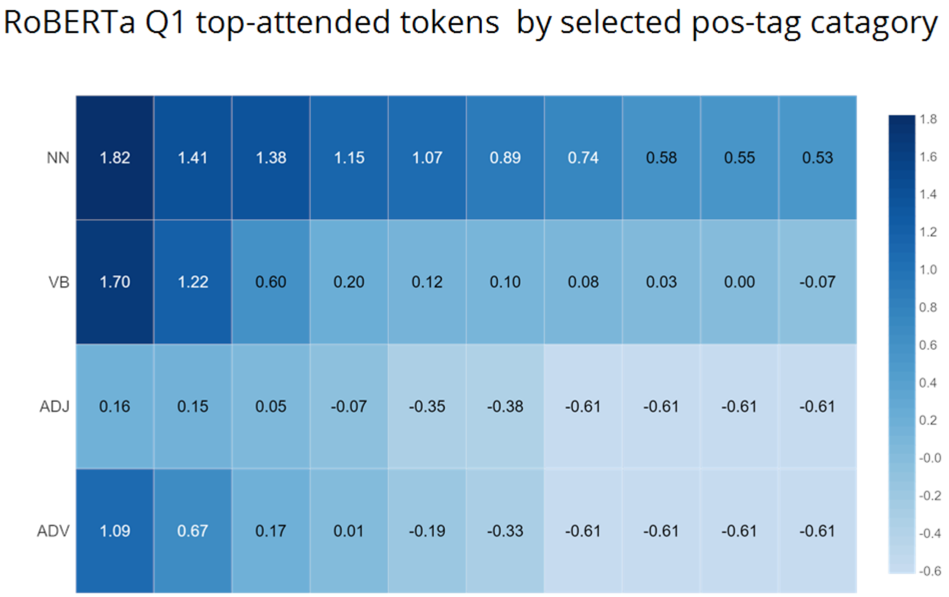


Figure 6.1: RoBERTa Q1 Model top 10 ACR tokens by POS tag

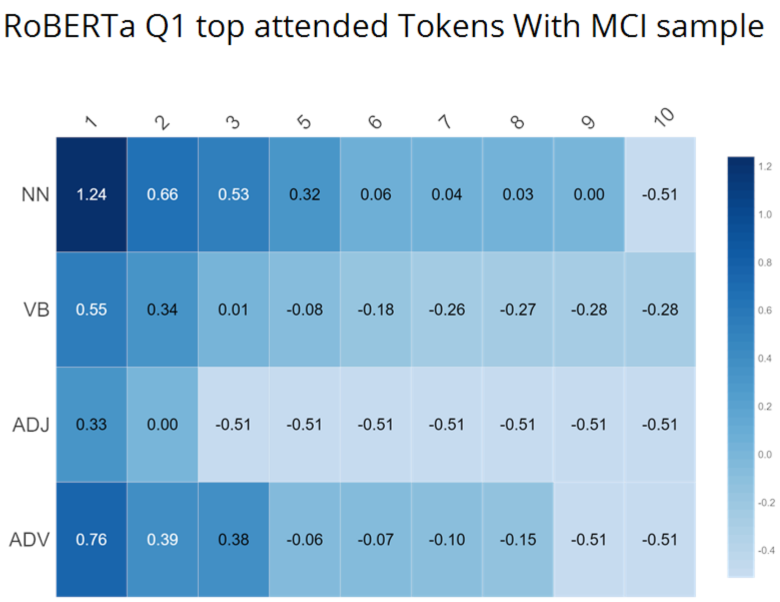


Figure 6.2: RoBERTa Q1 Model top 10 ACR tokens by POS tag

Question 2 Model Analysis

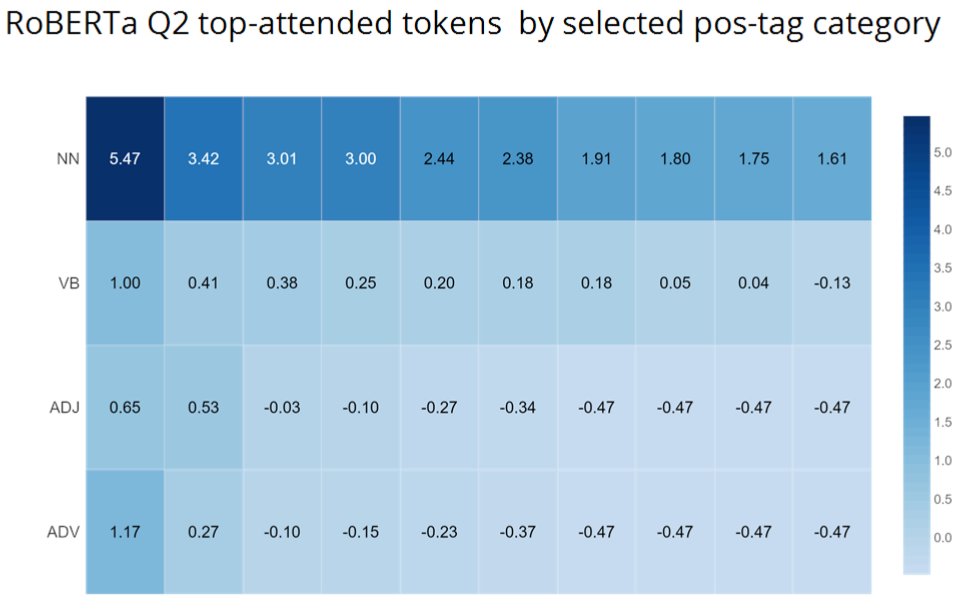
We notice that in question 2 Models, in both MCI transcripts and normal transcripts, the words with highest ACR scores are mostly objects in the examination room and words that describe these objects. Some important tokens includes **"sink"**, **"cabinets"**,**"floor"** , **"door"**, **"lighting"**, **"computer"**, **"monitor"**,**"picture"**. We use transcript V00003 YR3 to illustrate with details. We mark the tokens related to the objects in the room if the tokens are among the top 20 tokens with highest ACR scores. In the brackets after the token, the first item is the rank of ACR scores and the second item is the ACR score of the tokens. We see 15 out of 20 top tokens are either describing an objects in the room or the nouns of the objects.

V00003 YR3 Question 2:*The room we're sitting in is not exciting and it's rather neutral. And there's a* ***cabinet(10 — 180.24%)*** *with, um, a darker wood* ***finish(16 — 139.97%)****. And then there's, uh, that's above an aluminum or steel* ***sink(5 — 300.02%)****. And on the* ***counter(6 — 256.21%)*** *tops. I think that's for Micah. And then there's the darker wood* ***storage(13 — 161.47%) area(7 — 244.19%)*** *underneath. And to the right of the cabinet is a paper towel* ***holder(1 — 546.90%)****. To the right* *of that is the 10 ways to love your brain **chart(8 — 237.63%)**, which is neat. Then there is a four mica desk that bumps up to the **cabinet(10 — 180.24%)** and Tiffany's sits in that black **chair(11 — 175.25%)** across it or on the **stool(4 — 301.36%)**, uh, at the end of it. And I sit in a black chair opposite her and there's, um, a big computer **screen(3 — 342.11%)** that's not being used right now and there's some lap computer and a **blood(14 — 159.40%) pressure(17 — 139.82%) machine(20 — 129.26%)**.*

To show how important noun tokens are, we also visualize the top attended words by POS tag. In figure 6.3, we plot the same heatmap as we did for question 1. As we can see, the noun category, which includes mostly objects in the examining room, is the most dominant and maybe the only crucial category. Similarly, we also visualize the top attended words for MCI sample. In figure 6.4, we can see for both MCI sample and Control sample, noun is the dominant category, but we can also see for MCI sample, the ACR scores are smaller. This may indicate that this MCI subject is listing less important objects in the examination room.

Question 3 Model Analysis

We observe similar pattern in question3 as the pattern in question2. Several tokens have high ACR scores across the transcripts. Some important



NN	5.47	3.42	3.01	3.00	2.44	2.38	1.91	1.80	1.75	1.61
VB	1.00	0.41	0.38	0.25	0.20	0.18	0.18	0.05	0.04	-0.13
ADJ	0.65	0.53	-0.03	-0.10	-0.27	-0.34	-0.47	-0.47	-0.47	-0.47
ADV	1.17	0.27	-0.10	-0.15	-0.23	-0.37	-0.47	-0.47	-0.47	-0.47

Figure 6.3: RoBERTa Q2 Model top 10 ACR tokens by POS tag

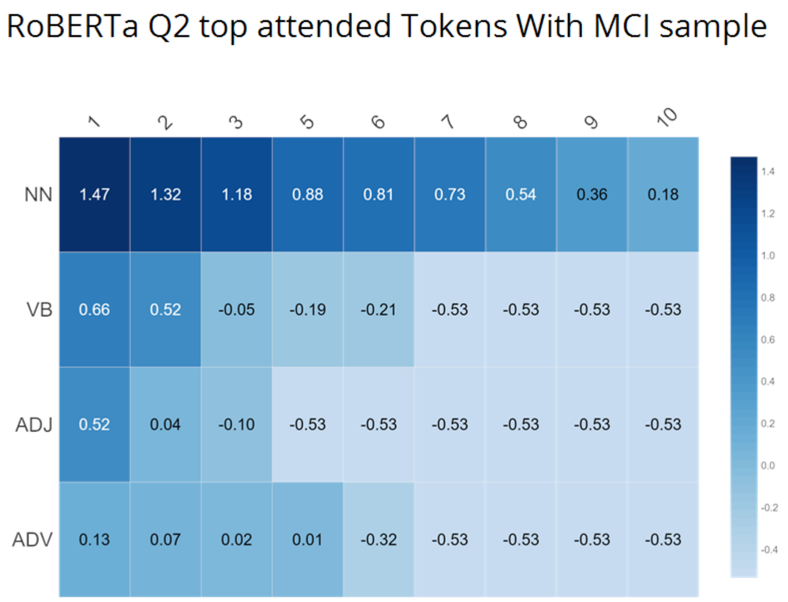


Figure 6.4: RoBERTa Q2 Model top 10 ACR tokens by POS tag

tokens include **"glasses"**,**"tri-cycle"**,**"cane"**,**"tie"**,**"grey"** , **"vest"**. The attended words are mostly nouns that appear in the picture, with some verbs that describe the movement or action of the characters in the picture and some adjectives describing the details on the dressing of the characters and the drawing style of the picture. We also plot the heatmap for top tokens by POS category for question3. In figure 6.5, we can see noun is still the most attended category. Compared to adjectives, there are more attended verbs, but there are some adjectives with higher ACR socres. Therefore we think for Q3 model, verbs and adjectives should be approximately at the same importance level. In addition, we plot the heatmap for MCI sample. in figure 6.6, we observe that noun is still dominating, which is consistent with our observation in normal sample. However, in this MCI sample, we notice that verb is more important than adjectives, while in the normal samples, we see some adjective token with higher ACR scores. The potential reason is that MCI subjects may tend to use less adjective to describe details in the picture.

6.1.3 BERT Analysis and Comparison to RoBERTa

The analysis on BERT model is similar to our analysis for RoBERTa model. Table 6.1.3 shows list of the chosen samples transcripts. In General we observe

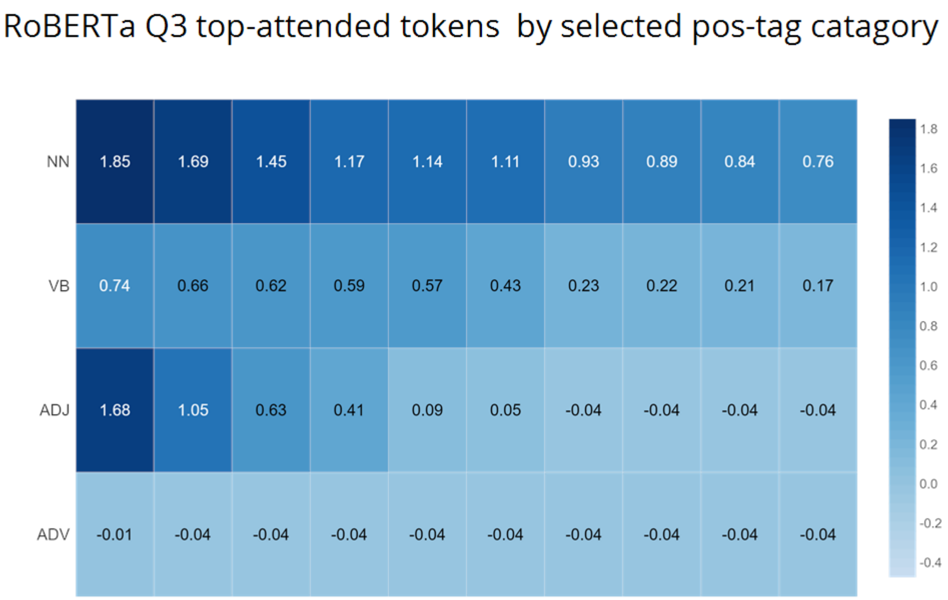


Figure 6.5: RoBERTa Q3 Model top 10 ACR tokens by POS tag

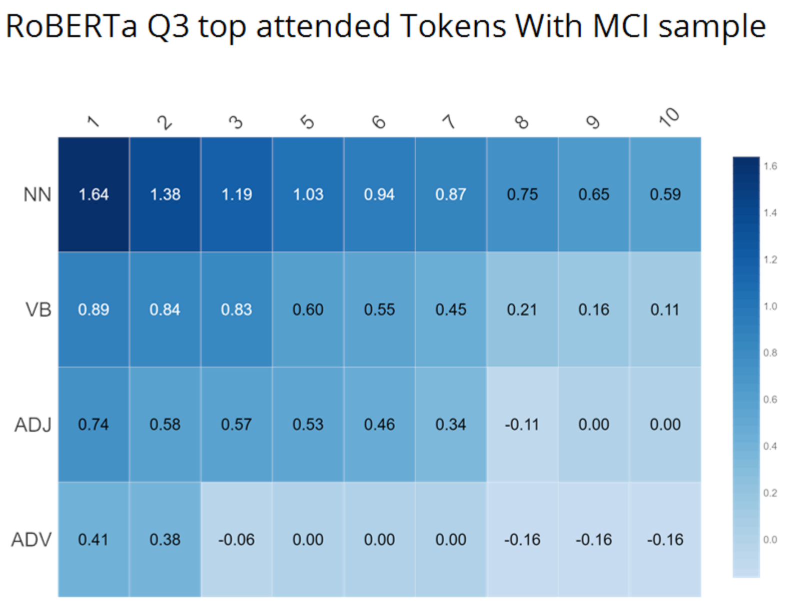


Figure 6.6: RoBERTa Q3 Model top 10 ACR tokens by POS tag

the similar pattern for three models, but in Q3 model, the pattern is slightly less noisy than the pattern in RoBERTa. We show this with a specific example. We choose transcript V00074 YR2, which both RoBERTa and BERT model correctly classified. We take the top 20 tokens for both model and mark the "noisy tokens"(transcription error or toknization error) in bold. We can see in RoBERTa top tokens, there are 7 "noisy tokens" while BERT only has two.

RoBERTa:V00074 YR2: *boots 199.42%, poles 144.38%,* ***tr(im)*** *137.38%,* ***(c)oller*** *129.71%, velvet 126.99%,* ***(Pol)ka*** *122.35%, waist 120.58%, dot 117.98%, belt 115.81%, emblem 115.47%, star 112.38%, shelf 111.27%,* ***(pro)cession*** *108.80%,* ***conf(ederate)*** *104.66%, lips 101.31%, fancy 98.38%,* ***(conf)eder(ate)*** *94.98%, white 91.38%,* ***(cor)nered*** *85.94%, feathers 81.54%*

BERT:V00074 YR2: *three 147.36%,elephant 90.74%,* ***(tri)##cycle*** *69.78%,um 58.93%,coat 57.09%,pants 56.58%,soldiers 52.76%,an 51.06%,* ***tri(cycle)*** *50.95%, boots 47.64%, confederate 47.42%, holding 47.27%,have 43.52%,one 42.11%, hat 38.95%, he 38.02%, their 37.90%, possession 37.19%,riding 34.59%,behind 30.70%*

CV set	0	1	2	3	4
MCI transcript	V00269 YR1	V00325 YR1	V00067 YR2	V00413 BL	V00099 YR2
Normal transcript	V00023 YR2	V00141 YR1	V00029 YR1	V00074 YR2	V00019 YR3

Table 6.2: BERT attention analysis samples

6.1.4 Finding validation

To Validate our finding on the samples, we perform the ACR token ranking on all transcripts using our fine-tuned RoBERTa models. However, we do not average the ACR score for one token appears in more than one transcripts; instead, we sum up its ACR score, because we believe tokens appearing in many documents should be more important and should rank higher. We choose the top 20 words and make chart 6.3 after taking out some very obvious stop words, and in the bracket is the ranking of the tokens.

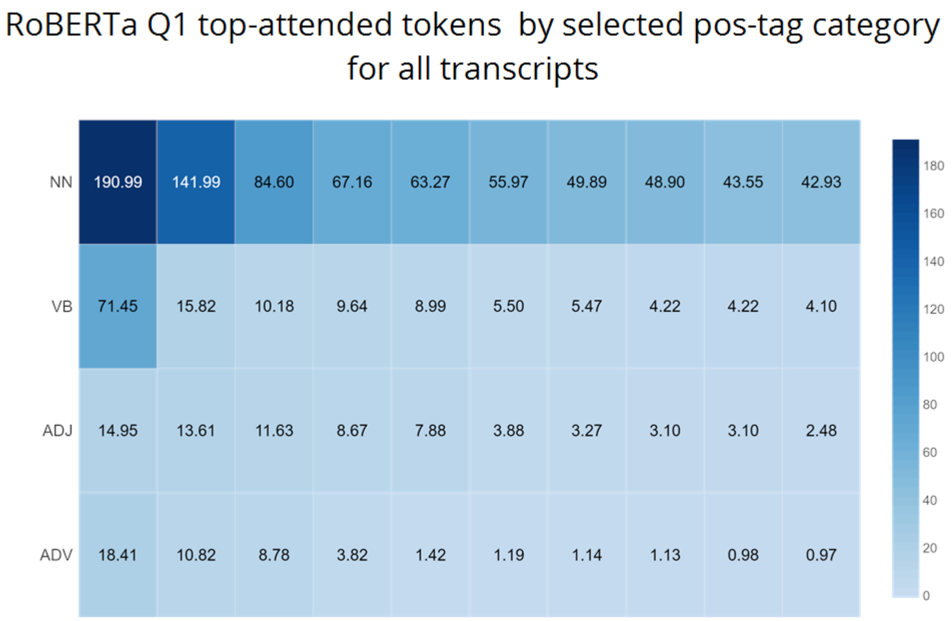
As a matter of fact, with the ranking of words from all the transcripts, we now can see a clearer pattern in the change of attention of the transformer models through fine-tuning process. In question 1 model, we can see the event-related nouns is still where most of the attentions go, and verb is also attended a few times. In question 2 model, the attended words are mostly nouns, and mostly the objects that are in the room. In question 3 model,

Q1	Q2	Q3
room(1)	sink(5)	tri(4)
pressure(3)	table(6)	pants(5)
um(5)	cabinet(8)	cycle(8)
tests(7)	Wall(9)	red(9)
met(10)	door(14)	cane(12)
blood(12)	room(15)	jacket(13)
test(13)	computer(16)	clown(16)
MRI(15)	gloves(17)	tie(17)
computer(16)	cabinets(20)	vest(20)
testing(17)	chair(21)	like(21)
Um(18)	trash (22)	wearing(22)
Uh(19)	pressure(23)	looks(24)
arm(20)	desk(24)	carrying(26)
lobby(21)	counter(26)	dressed(27)
floor(22)	blood(28)	yellow(28)
ultrasound(23)	see(30)	riding(30)
questions(25)	monitor(31)	striped(31)
morning(26)	can(32)	fan(32)
study(27)	dispenser(33)	holding(35)

Table 6.3: RoBERTa token ranking for all transcripts

verbs and descriptive adjective are also attended.

We also visualize the attended tokens by POS categories for three Individual Models. In this visualization, since we are ranking the sum of ACR scores for tokens, we filter out some stop words to make clearer maps. We see in figure 6.7, 6.8 and,6.9 our observation on the vitality of nouns is correct for all three models. For question 3 model, we see verb is the second most attended token category, and adjective is the third most attended token category.



NN	190.99	141.99	84.60	67.16	63.27	55.97	49.89	48.90	43.55	42.93
VB	71.45	15.82	10.18	9.64	8.99	5.50	5.47	4.22	4.22	4.10
ADJ	14.95	13.61	11.63	8.67	7.88	3.88	3.27	3.10	3.10	2.48
ADV	18.41	10.82	8.78	3.82	1.42	1.19	1.14	1.13	0.98	0.97

Figure 6.7: RoBERTa Q1 top 10 ACR tokens by POS tag for all transcripts

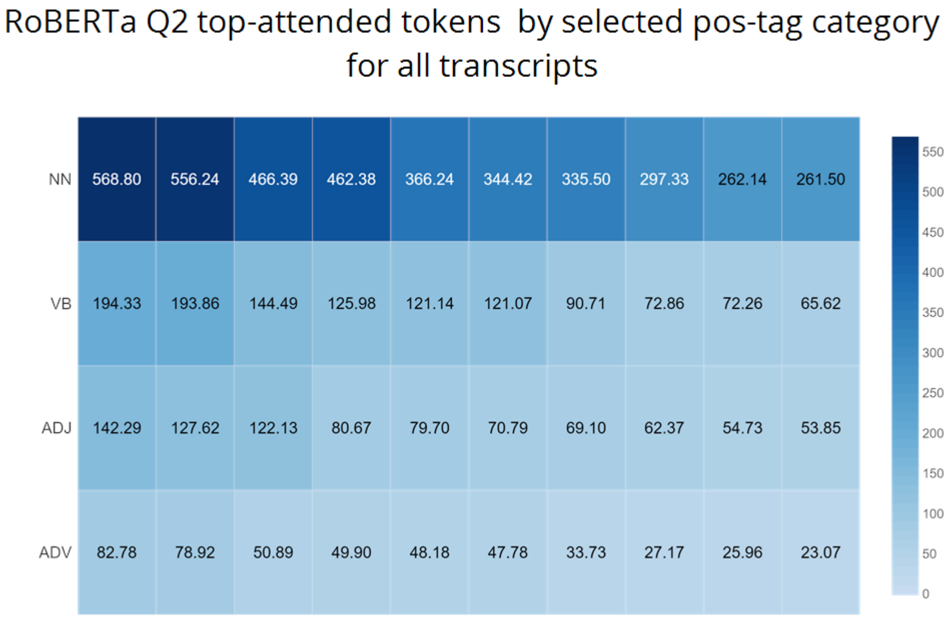
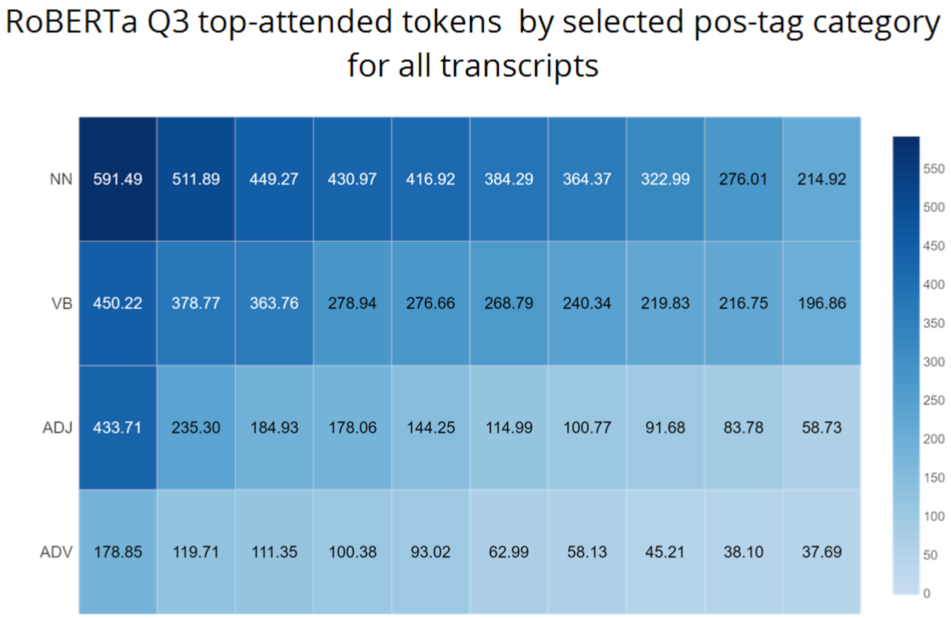


Figure 6.8: RoBERTa Q2 top 10 ACR tokens by POS tag for all transcripts



NN	591.49	511.89	449.27	430.97	416.92	384.29	364.37	322.99	276.01	214.92
VB	450.22	378.77	363.76	278.94	276.66	268.79	240.34	219.83	216.75	196.86
ADJ	433.71	235.30	184.93	178.06	144.25	114.99	100.77	91.68	83.78	58.73
ADV	178.85	119.71	111.35	100.38	93.02	62.99	58.13	45.21	38.10	37.69

Figure 6.9: RoBERTa Q3 top 10 ACR tokens by POS tag for all transcripts

6.2 Ensemble Analysis

In this section we analyze our ensemble models with our main focus on the final MLP classification layers of the ensemble models. we would like to find out which question in the speech protocol is playing a more effective role in revealing subjects' brain health condition. To get the most meaningful samples for analysis, we choose the best performing model from three runs. For RoBERTa Ensemble model, we use the first run model , which has the following statistics: Accuracy **72.72%** Sensitivity **53.96%**, Specificity: **85.67%**.

For B+R+A Ensemble Model, we use the model saved from the third run, which has the following performance: Accuracy **74.43%**, Sensitivity **65.28%**, and Specificity **83.27%**.

6.2.1 RoBERTa Ensemble Analysis

We analyze how the final decision of the Ensemble Model is made by slicing the weight of the final MLP layer and calculate the corresponding "vote" from each question. The last layer MLP works the following way. On input q_1,q_2, and q_3, we first concatenate three input to a long vector $q_{all} = q_1 \oplus \mathrm{q}_2 \oplus \mathrm{q}_3$. q_{all}

has size [768*3]. The final vector containing two scores for MCI and normal s_{all} is calculated the following way: $s_{all} = q_{all} * W_{all} + b$, where W_{all} is the weight matrix of size [768*3,2], and b is the bias term of size [1,2]. The final decision d_{all} is obtained by taking argmax of s_{all}. If $d_{all} = 1$, the transcript is classified as MCI, else the transcript is classified as normal. To investigate how the MLP layer handle the input from three transformer models. We first slice the weight of final MLP layer to three smaller matrix W_{q1}, W_{q2}, and W_{q3}, each of size [768,2], then we calculate the MCI/Normal scores of each input from transformer the following way: $s_i = q_i * W_i + \frac{1}{3} * b$, for $i \in \{1, 2, 3\}$. Then the final votes of each individual model $v_i = argmax(s_i)$. Notice we should have $s_{all} = \sum_{i \in \{1,2,3\}} s_i$. Therefore our voting analysis can properly reflect the how the final MLP layers handle text encoding of each transformer model.

Figure 6.10 shows the proportions of different voting combinations among the 466 transcripts that the RoBERTa Ensemble Model correctly classifies. The label for each slice represents the voting distribution. For example, Q1-Q2 means that vote for Q1 model and Q2 model is correct while a vote for Q3 model is incorrect, and the proportion of Q1-Q2 is 19.5%, meaning that among 466 transcripts, there are 19.5% cases where the votes for Q1 and

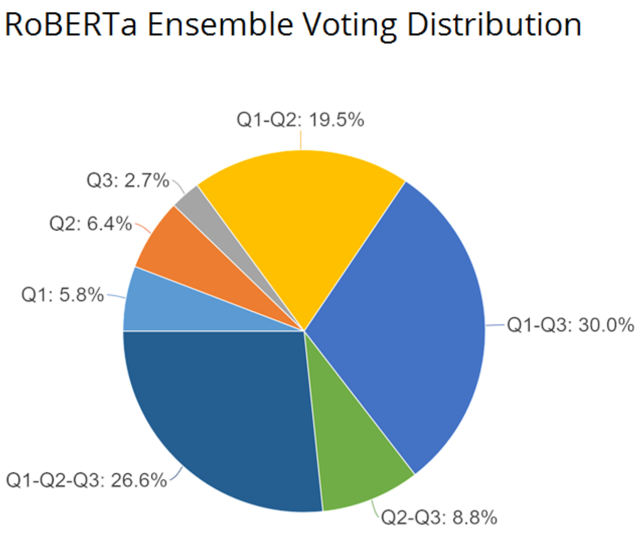


Figure 6.10: RoBERTa Ensmeble Model Votes Distribution

Q2 are correct. Looking at the graph, we noticed that the most common situation is that votes for Q1 and Q3 are correct, followed by situation where the votes for all three questions are correct. We also notice that the majority vote cases (two or more votes get correct) are the most common cases, that constitutes 85.1% of all correctly classified samples. However, we still notice that there is 14.9% of the correctly classified samples come from the situation where only the vote for one question is correct. We call the situation where the v_i is the only correct vote the **dominant case** of question i, and we call the ratio of such cases among all cases that the ensemble model correctly classifies **% of dominant cases** of quesion i

To show the importance of each individual input, we introduce the following two measures:

1.**Correct Vote Ratio(CV)** :

$$cv_i = \frac{|C_i \cap C_{all}|}{|C_{all}|} \tag{6.4}$$

where C_i is the set of transcripts where vote v_i is correct, C_{all} is the set of transcripts that the ensemble model correctly classifies. For cv_i, we essentially increment one when both the ensemble model and vote v_i are correct.

2. Normalized Correct Vote Ratio (NCV):

$$ncv_i = \frac{\sum_{c_t \in (C_i \cap C)} \frac{1}{|V_t|}}{|C_{all}|} \tag{6.5}$$

where the V_t is the set of votes that are correct for transcript t. In RoBERTa ensemble model, $|V_t| \in \{1, 2, 3\}$. For NCV_i, essentially we add 1 to the numerator of NCV_i for the dominant cases of question i, add 0.5 to the numerator of NCV_i if there are two correct votes and add $\frac{1}{3}$ to the numerator of NCV_i if all three votes are correct. Notice that the sum of all NCV scores should equal to 1. We apply this normalization, since we believe that if one vote is correct when fewer votes are right, then the vote should be considered more important.

We compute CV, NCV % of dominant cases in table 6.2.1. We also add the fourth row, which records the Individual transformer Model Accuracy(IMA). we use this score to make comparison between three models in ensemble and three models as individuals.

In table 6.2.1, the NCV scores indicate that it's unlikely that the model is dominated by the model of any single question. However, we can still see that Q1 is the most important question in the ensemble model, it has the highest

	Q1	Q2	Q3
Correct Votes	81.97%	61.37%	69.09%
Normalized Correct Votes	39.45%	29.47%	31.08%
% Dominant Cases	5.79%	6.43%	2.79%
Individual Accuracy	68.25%	70.11%	64.87%

Table 6.4: Importance comparison of 3 inputs in RoBERTa Ensemble

scores for Correct Votes and Normalized Correct Votes, the high importance may be the result of the fact that Q1 is a memory test and can more directly reflect the subjects' brain health condition. We notice Q2, although it achieves the best individual model accuracy, is actually not the most important in the ensemble model. Q3, although it does not do well as an individual model, does better than Q2 in the ensemble model. This observation indicates that our ensemble model is not simply piling up the individual models; instead, the ensemble model focuses on the text representation and comes up with its own logic for classification. In addition, we notice that the ensemble model still reflects the performance of three models as individuals to a certain degree. The % of Dominant Cases matches the Individual Accuracy pretty well.

BERT-RoBERTa-ALBERT Ensemble (B-R-A Ensemble)

Similar to the analysis we perform on RoBERTa ensemble and B-R Ensemble, we slice the weight of the final MLP layer into 9 equal-size matrices corresponding to each of 9 inputs, perform matrix multiplication and add $\frac{1}{9}$ of the bias vectors from MLP weight to get 9 deciding votes corresponding to 9 inputs. Then for each transcript, we record the 9 deciding votes, the final ensemble prediction and the label, and then count the distribution of votes and influencing power of each input.

Due to the fact that B+R+A Ensemble create 2^9 combinations of votes, we obviously cannot list the percentage for every single case; therefore, we only record the number of correct votes every time the ensemble model is making a correct prediction. Among 484 transcripts that the model correctly classified, **86.16%** are derived from a majority vote, meaning that at least 5 out of 9 votes match with the label. vote of 6 and 5 are the biggest groups, with proportion of **34.50%** and**28.51%**. We also notice that cases of under 3 correct votes do not exist, indicating that there is not likely to be any single input or small input group that dominates the final prediction. Moreover, we also notice that only on very few samples (**0.21%**), all 9 votes agree together.

To show which input takes a more important role in the model, we calculate

	B-Q1	B-Q2	B-Q3	R-Q1	R-Q2	R-Q3	A-Q1	A-Q2	A-Q3
CV	62.19%	74.38%	74.38%	**80.99%**	55.58%	65.91%	41.53%	60.12%	55.17%
NCV	10.87%	13.32%	12.81%	**13.98%**	9.43%	11.31%	7.90%	10.84%	9.53%
IMA	67.80%	68.41%	66.87%	**70.72%**	69.80%	65.18%	69.18%	69.19%	66.87%

Table 6.5: Importance comparison of 9 transformer models in Ensemble Model

the CV, NCV scores for each input and include its IMA score for reference.

First, the NCV result in the table 6.2.1 validates our previous observation that there is not a single very dominant transformer model. We notice that the three most important transformer models of the Ensemble Model are RoBERTa-Q1, BERT-Q2, and BERT-Q3, which form the components of a complete transcript; therefore a possible reason is that the ensemble model picks the best representation of the questions in a transcript and performs classification based on the chosen representations. Moreover, we notice that the importance of input does not necessarily correspond to individual model performance, which is also seen in the RoBERTa Ensemble model. This observation further signals that our ensemble model is not a simple summation of the Individual models, instead, it focuses more on the representations produced by the individual transformers models.

Chapter 7

Conclusion

In this work, we mainly do three things.

1. We process and introduce the B-SHARP dataset, which is currently the largest dataset for MCI classification task.

2. We experiment current existing approaches on MCI classification task using B-SHARP dataset, and propose a hierarchical transformer model and use the hierachical transformer model to achieve 74% accuracy on MCI classification task.

3. We perform analysis on individual transformer models, find out that different categories of words are attended for different questions in the speech protocol. Also, we perform ensemble model analysis and explain how the final MLP classification layer handles encoding from individual models.

Bibliography

[1] James T. Becker, Francois Boller, 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.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. URL `https://www.aclweb.org/anthology/N19-1423`.

[3] Serge Gauthier, Michel Panisset, Josephine Nalban-toglu, and Judes Poirie. Alzheimer's dis-ease: current knowledge, management and research. *Canadian Medical Association Journal*, 157(8):1047–1052, 1997.

[4] George G Glenner. *Alzheimers disease.* Springer, 1990.

[5] Sweta Karleka, Tong Niu, and Mohit Bansal. 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)*, pages 701–707, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2110. URL `https://www.aclweb.org/anthology/N18-2110`.

[6] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. URL `https://www.aclweb.org/anthology/D14-1181`.

[7] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018. URL `https://www.aclweb.org/anthology/D18-2012`.

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel,

Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv*, 11942(1909), 2019.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Proceedings of the International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SyxS0T4tvS`.

[10] Frank Rudzicz, Rosalie Wang, Momotaz Begum, and Alex Mihailidis. Speech recognition in alzheimer's disease with personal assistive robots. In *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies)*, page 20–28, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1904. URL `https://www.aclweb.org/anthology/W14-1904`.

[11] Richard Suzman and John Beard. Global health and aging., 2011.

[12] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaob-

ing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv*, 1609(08144), 2016. URL `http://arxiv.org/abs/1609.08144`.