frontiers in Public Health

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fpubh.2021.835960

Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network

Ning Liu[1,2,3], *Zhenming Yuan*[1,4]* *and Qingfeng Tang*[5]*

[1] School of Public Health, Hangzhou Normal University, Hangzhou, China, [2] Department of Mathematics and Computer Science, Quanzhou Normal University, Quanzhou, China, [3] Fujian Provincial Key Laboratory of Data-Intensive Computing, Quanzhou Normal University, Quanzhou, China, [4] School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China, [5] School of Computer and Information, Anqing Normal University, Anqing, China

Alzheimer's disease (AD) is a neurodegenerative disease involving the decline of cognitive ability with illness progresses. At present, the diagnosis of AD mainly depends on the interviews between patients and doctors, which is slow, expensive, and subjective, so it is not a better solution to recognize AD using the currently available neuropsychological examinations and clinical diagnostic criteria. A recent study has indicated the potential of language analysis for AD diagnosis. In this study, we proposed a novel feature purification network that can improve the representation learning of transformer model further. Though transformer has made great progress in generating discriminative features because of its long-distance reasoning ability, there is still room for improvement. There exist many common features that are not indicative of any specific class, and we rule out the influence of common features from traditional features extracted by transformer encoder and can get more discriminative features for classification. We apply this method to improve transformer's performance on three public dementia datasets and get improved classification results markedly. Specifically, the method on Pitt datasets gets state-of-the-art (SOTA) result.

Keywords: Alzheimer's disease, natural language processing, deep learning, transformer, machine learning, speech and language, mild cognitive impairment

OPEN ACCESS

Edited by:
Thippa Reddy Gadekallu, VIT University, India

Reviewed by:
Praveen Kumar, VIT University, India
Weizheng Wang, City University of Hong Kong, Hong Kong SAR, China

***Correspondence:**
Zhenming Yuan
zmyuan@hznu.edu.cn
Qingfeng Tang
tqf1013@sina.com

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *15 December 2021*
Accepted: *28 December 2021*
Published: *03 March 2022*

Citation:
Liu N, Yuan Z and Tang Q (2022) Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network. Front. Public Health 9:835960. doi: 10.3389/fpubh.2021.835960

INTRODUCTION

Alzheimer's disease (AD) is a nervous degenerative disease with an insidious and irreversible onset, which is difficult to be detected in every stage. AD can influence patients' daily living ability and social communicate ability and may even lead to disability (1, 2). Researchers have found that AD has a profound impact on patients' language function (3) in addition to mood, attention, memory, movement, and so on. Language is the representation of mental activities, which can clearly reflect the relationship among language, cognition, and communication (4). Language interference is a common manifestation of patients (5) with AD which may even earlier than orientation and memory difficulties (6, 7). Picture description task, taken from Boston Aphasia Diagnostic Test (8), has already been verified sensitive to subtle cognitive deficits (9); therefore, valuable clinical information can be obtained from spontaneous speech to recognize AD. The transcripts of speech can be used to detect AD effectively.

The problem of AD recognition can be regarded as text classification problem in natural language processing (NLP). Deep learning models manifest better in classification as they can extract deep semantic features by effective model architecture automatically. For example, RNN can capture long-term dependencies within sentence, but it may neglect some important local words which may important for classification (10), and CNN can capture local and position-related features (11) but cannot give enough weight to some discriminative or special words. To solve the problem, attention mechanism was introduced. Transformer gives different weights to different words using attention mechanism, the performance of which is better than CNN and RNN. Although transformer has made great progress in producing discriminative features by powerful representation learning, there is still room to improve. There are few studies nowadays in this area to improve representation learning of deep learning. Based on GRL (12–17) in extracting common features which are not discriminative for classification, this paper proposes a novel feature purification method to improve the representation learning of transformer to get a more discriminative feature vector to diagnose AD.

The original transcripts of speech are the description of a picture, which should be comprehensive and integrated for a normal individual. That is to say, the discriminative words or sentences, with relevant and less vague words, should be included. For example, accurate descriptive words, such as "mother", "tap", "the stool is tipping", etc., are usually a better cognitive sign. Words or sentences such as "I do not know", "um", and "pause" should be an indicative of a bad cognitive condition, and they are discriminative for AD recognition. But some equivocal, inconsequential, and even irrelevant descriptions are unhelpful and may even interfere with the final classification, such as "is not that enough?", "It is great", "there may be a little breeze coming in", et al. They can disturb the representation learning of deep learning by producing suboptimal representations. To solve this problem, transformer proposes a self-attention mechanism to give weight to words and usually can get better performance than RNN and CNN. Though attention may alleviate the influence by giving a higher or lower weight for those more or less relevant words, the classification problem cannot be solved properly with inaccurate attention mechanism or specificity of data. To solve the above problems, our study, inspired by the paper (17) which used feature projection method to purify the representation learning of deep learning, proposes a novel feature purification method to improve representation learning of transformer to get more discriminative features, which is GP-Net. It has two subnetworks, a common feature learning network called G-Net and a purification network called P-Net. G-Net uses gradient reverse layer (GRL) (12, 13, 18) to extract common features which are shared by classes and have no or few roles for classification. P-Net first uses transformer encoder to extract traditional feature vector for the sentence. Then, it rules out the common features from traditional feature vector to generate more purified features. It is clear that this operation gets rid of the effect of common features and makes the system only focus on discriminative features. We will explain the principle in Method Section.

The experiments on three datasets with our method get an improved performance which prove that the purified features are more discriminative. To the best of our knowledge, there have been still no studies to recognize AD from spontaneous speech by purifying representation learning of deep learning up to now.

The key contributions we have made in this work include the following:

(1) A whole process of AD screening method, based on linguistic data, was designed and implemented.
(2) We propose a novel feature purification network to improve representation learning of transformer and get state-of-the-art (SOTA) result on Pitt dataset.
(3) The proposed method has the advantage of low cost, reliable, and convenience, which can provide a feasible solution for the screening of AD with a better performance.

RELATED WORK

Existing studies on AD diagnosis across spontaneous speech mainly focus on two aspects. One is feature extraction manually including acoustic features (19–21), linguistic features (22–25), or their combinations (21). This method is subjective and needs more professional knowledge. They are generally associated with a specific task scenario; once the scenario changes, these artificially designed features and prior settings cannot adapt to new scenarios and need to be redesigned, so the model has a low universality. The other is deep learning method which can extract deep semantic features automatically. Based on its powerful representation learning ability, the performance of deep learning is usually better than the first method. Additionally, deep learning improves the generalization ability of the classifiers which can be utilized further in different clinical environments. Deep neural network can process representation learning to extract deep semantic features using cascaded data of multilevel non-linear processing units without the need for feature engineering manually.

AD Detection Based on Deep Learning

There are many studies to detect AD from oral speech with deep learning methods (26–29), such as RNN, long–short-term memory (LSTM) networks [e.g., ELMo (30)] and CNN. Recurrent convolutional neural Networks (RCNN) (31) uses Bi-LSTM to get contextual information and then concedes the hidden output of Bi-LSTM and word embedding for classification. DPCNN (32) is a simple network with 15 layers which likes a deep CNN, and it increases network depth of CNN but does not increase the computational cost. Attention mechanism is used in many NLP tasks such as text classification (33–36). Transformer architecture [e.g., Bert (37)] uses attention mechanism to extract deep semantic features. Enhanced representation through knowledge integration (ERNIE) (38, 39), proposed in 2019 by Baidu corporation, is optimized further based on Bert model. They usually have better performance than CNN, RNN, and LSTM.

Public Dementiabank datasets or ADReSS challenge (40) datasets are often used to recognize AD. For example, Orimaye et al. (41) proposed the combination of deep language models and deep neural network to predict mild cognitive impairment (MCI) and AD. The datasets used were public Dementiabank transcript with 37 healthy elderly and 37 MCI transcripts. The study did not use any handcrafted features; just the original transcripts were fed to the model, and n-gram word embedding method combined with deep neural network (DNN) got a best AUC of 0.83. Different from our dataset and classification, there was no comparability with our method. Karlekar et al. (23) used four types of interviews: story recall, sentence construction, cookie-theft picture description, and vocabulary fluency; the dataset included 243 normal controls and 1,017 AD transcripts. Three classifiers were used for comparison, that is, LSTM-RNN, CNN, and CNN-LSTM, and achieved a best accuracy of 91.1%, but the results were somewhat questionable as mentioned in the Discussion. These methods used deep learning algorithms or their linear combinations to recognize MCI and AD. Our work is much different clearly as none of these existing studies improve representation learning of deep learning by feature purification method.

Studies Related to GRL

Our study is related to some former work. Ganin and Lempitsky (13) first introduced GRL to extract common features which were sentiment-sensitive and domain shared in domain adaptation (DA). It embeds DA to the process of representation learning in order that the final classification result is more discriminative for the domain changes. Though we use GRL to extract common features, we do not use it in the area of DA, and they also do not use for feature purification. Belinkov et al. (14) used adversarial learning to encourage the model to process representation learning on SNLI dataset. Combining with aspect attention and GRL, Kai Zhang et al. (16) studied cross domain text classification problem, and common features across domains were extracted from the aspects for text classifications. The idea of generative adversarial networks (GANs) (42) was used to ensure that the common feature space did not mix with private features and only contained pure task-independent common feature representation. In these studies, they all used GRL to extract common features inseparable for two domains, and domain-shared features were generated in the shared space according to adversarial training, whereas our study is different from them clearly as this existing work does not improve representation learning of the model. The study (17) proposed a feature projection method to further improve representation learning of deep learning from a novel angle. The method projected existing features into the orthogonal space of the common features, so the resulting projection is perpendicular to the space that common features located in and thus more discriminative for text classification. Different from this study (17) which only deletes a section of common features, we rule out the influence of whole common features, which we believe that a better classification performance should be achieved. Also, we did the experiment with the method of study (17) on Pitt dataset, and the performance is not better than our method, just as shown in **Table 1**.

TABLE 1 | Relationship between predicted and true classes.

	True class	
Predicted class	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

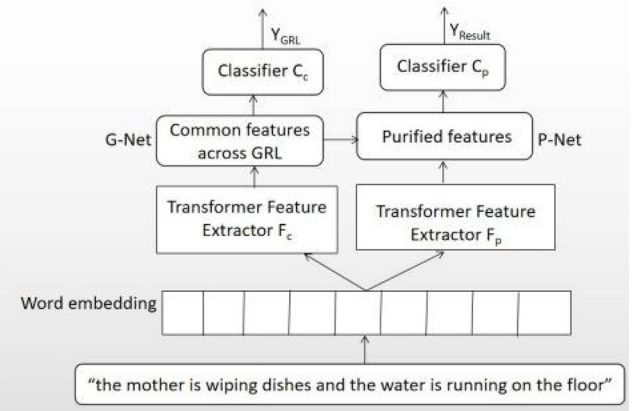

FIGURE 1 | The architecture of GP-Net.

METHODS

In this study, we propose a novel GP-Net framework to recognize AD from normal controls, which is indeed a binary classification problem in NLP.

Feature Purification Network: GP-Net

This paper proposed a novel architecture, named GP-Net, to recognize AD, the network structure of which is shown in **Figure 1**. The whole network includes two sections: G-Net and P-Net. The aim of G-Net is to extract common features by reversing the gradient direction in the training process, and these common features are shared by both classes and have no discriminative for classification. The aim of P-Net is to purify the features further by deleting the common features from traditional features extracted from transformer model. G-Net includes four sections, that is, the input layer X, feature extractor F_c, GRL, and classifier layer C_c. P-Net also includes four sections, which include the input layer X, feature extractor F_p (the features extracted by F_c and F_p have no share parameters), purification network, and classifier layer C_p. The main idea of proposed network is as follows: the feature vector, extracted by the feature extractor F_p, deletes the common features got from G-Net, and then, more discriminative purified features have got for the final classification. Two operations, including G-Net and P-Net, are required in order for feature purification operation.

Transformer Extractor

This study uses transformer encoder as the feature extractor. Transformer is a SOTA model which has a novel architecture to solve sequence to sequence tasks. The model can capture long-distance dependencies and learn global semantic features of input text thoroughly through multihead self-attention mechanism. As transformer has some mechanisms as self-attention and location code, it has excellent feature extraction and semantic abstract competence. Like most Seq2Seq model, transformer model also uses encoder–decoder structure, the encoder of which is a better feature extractor with multihead attention and feed forward neural network.

Supposing G-Net and P-Net have the same input X, the feature extractors of G-Net and P-Net are F_c and F_p, which can get the advanced features f_p and f_c from the input layer, respectively, but there are not any shared parameters between them. We refer to the features of P-Net and G-Net, respectively, as

$$f_p = Transformer_p(X), \quad (1)$$

$$f_c = Transformer_c(X), \quad (2)$$

Additional details of G-Net and P-Net will be introduced in G-Net and P-Net module.

G-Net Module

The main goal of G-Net module is to extract common features among datasets, which is not discriminative for the classification. As common features are those shared by all the classes, the classifier cannot use them to distinguish different classes effectively. To get common features, GRL (12, 13, 18) is added between the feature extractor F_c and the classifier to reverse the gradient direction. The common features that are shared among different classes are obtained after the training module. G_λ can be thought as two incompatible equations that describe the forward and back propagation behaviors:

$$G_\lambda(x) = x, \quad (3)$$

$$\frac{\partial G_\lambda}{\partial x} = -\lambda I, \quad (4)$$

where λ is a hyper parameter. We process feature vector f_c through GRL and get f_c', for example, G_λ (f_c) = f_c'. To make f_c' close to real common features, GRL acts as identity transform during the forward propagation and then takes the gradient from subsequent level and changes the value (i.e., multiplies it by -λ) before passing it to the next layer during back propagation, and this operation can ensure that the feature distributions are similar and as indistinguishable as possible for the classifier. Only in this way we can get the common features sharing among classes. Finally, f_c' is fed to classifier C_c.

$$Y_{GRL} = softmax(W_c^* f_c' + b_c), \quad (5)$$

$$Loss_c = CrossEntropy(Y_{True}, Y_{GRL}), \quad (6)$$

where W_c and b_c are the weight and bias of classifier C_c. By optimizing $Loss_c$, the feature extractor F_c can extract common features of different classes.

P-Net Model

The goal of P-Net is to extract the semantic information from input example first and then purify features for the classification. Supposing the traditional feature vector we extracted by transformer is f_p, the common feature vector is f_c. The final feature vector for classification is f_w.

$$f_w = f_p - f_c, \quad (7)$$

As f_c disturbs the classification result, we delete f_c from f_p to eliminate the influence of nondiscriminative feature vector (i.e., common features), so the feature vector f_w is more discriminative than f_p. Finally, the purification feature vector f_w is fed to classifier C_p.

$$Y_{Result} = soft\ max(W_p * f_w + b_p), \quad (8)$$

$$Loss_p = CrossEntropy(Y_{True}, Y_{Result}), \quad (9)$$

where W_p and b_p are the weight and bias of classifier C_p. By optimizing $Loss_p$, the feature extractor F_p can purify the features, $Loss_c$ and $Loss_p$ are trained simultaneously, but they use different optimizers. $Loss_c$ use moment SGD optimizer because Ganin and Lempitsky (13) also used moment SGD, and $Loss_p$ use Adam optimizer. We also conducted the experiments using Adam optimizer for both G-Net and P-Net and found that the results made no difference when using two different optimizers. In terms of optimization targets of feature extractor F_c, though the two losses are opposite to each other, a balance can be found to make the extracted feature f_c closer to real common features. The algorithm description of the whole training process is shown in **Algorithm 1**:

Algorithm 1 GP-Net

1: Input:
Supposing the datasets are $D = \{(x_i, y_i)\}_{i=1}^N$, x_i is the embedding matrix of deep learning, $X_i \in R^{Lk}$, y_i is the corresponding classes; randomly initialized the parameters of GP-Net.
2: **For** every iteration b=1, 2,...N, **do**:
3: Sample one batch x_b from D,
4: **G-Net section:**
5: Generate common features (CFs) (Equation 1)
6: CFs go through GRL and get the features closer to the common features (Equation 3)
7: Do the classification (Equation 5)
8: **P-Net section:**
9: Generate traditional features(tfs) (Equation 2)
10: Get purified features (Equation 7)
11: Do the classification (Equation 8)
12: **Update parameters:**
13: The parameters of P-Net and G-Net are updated together (Equation 6 and Equation 9)
14: **End for**

TABLE 2 | AD vs. CTRL classification scores on Pitt datasets.

Method	Embedding	Classifier	Precision	Recall	Accuracy	F1
Sweta Karlekar (23)	POS	CNN-RNN	-	-	91.1	-
Fritsch et al. (29)	n-gram	NNLM+LSTM	-	-	85.6	-
Orimaye et al. (41)	n-grams	D2NN	-	-	88.9	-
Fraser et al. (43)	35 Hand-Crafted Feature	LR	-	-	81.92	-
Yancheva et al. (45)	12 Cluster-Based Features+LS&A	Random Forest	80.00	80.00	80.00	80.00
Sirts et al. (46)	Cluster+PID+SID Features	LR	74.4 ± 1.5	72.5 ± 1.2	-	72.7 ± 1.2
Hernandez et al. (47)	105 Hand-Crafted Features	SVM	81.00	81.00	79.00	81.00
Roshanzamir et al. (48)	BERT $_{Base}$ (Sentence Level)	LR	90.31 ± 7.36	76.52 ± 8.06	84.46 ± 6.31	82.72 ± 7.21
Roshanzamir et al. (48)	Bert $_{Large}$	LR	90.57 ± 3.18	84.34 ± 7.58	88.08 ± 4.48	87.23 ± 5.20
Pan et al. (49)	GloVe Word Embedding Sequence	BiLSTM|GRU Hierarchical Attention	84.02	84.97	-	84.43
Li et al. (50)	185Hand-Craft Features	LR	-	-	77	-
Fraser et al. (51)	Info and LM Features	SVM	-	-	75	77
Transformer+FP^{25}	Transformer +Feature projection	Transformer	88	**91**	90.3	90.6
Transformer+GP	Transformer+Feature purification	Transformer	**94**	89	**93.5**	**91.19**

The study marked with bold is the best performances on Pitt dataset.

EXPERIMENT

Datasets

Three datasets are used to carry out the experiments which are all the dialogues of picture description task, including English and Chinese.

Pitt Datasets

This is a Pitt corpus (43) from Dementiabank dataset (43), which comes from a study at School of Medicine in Pittsburgh University and is gathered longitudinally every year. More detailed description about the dataset can refer to the study (43). After deleting some unqualified datasets such as unknown label, memory impairment, and other dementia diagnose, for example, vascular dementia, there are 498 participants enrolled in our study after data preprocessing, which is composed of 242 controls and 256 possible or probable AD. Both categories are balanced basically.

ADReSS Datasets

The datasets include 78 dementia patients and 78 normal controls from ADReSS challenge in 2020. The speech is segmented using a voice activity detect method based on signal energy value. All datasets have already been preprocessed by removing noise.

iFLY Datasets

The Chinese datasets include 111 CTRL and 68 AD, with 60 women and 51 men in CTRL group and 38 women and 30 men in AD group, respectively. More details can refer to the website: http://challenge.xfyun.cn/2019/gamedetail?blockId=978.

Feature Parameters

In the training stage of GP-Net module, a stochastic gradient of 0.9 is used as momentum, and annealing learning rate can be calculated by the following formula:

$$l_p = \frac{l_0}{(1 + \alpha^* p)^\beta} \tag{10}$$

where $l_0 = 0.01$, $\alpha = 10$, $\beta = 0.75$, p is training progress linearly changing from 0 to 1. In Equation (4), the parameter λ is set as [0.05, 0.1, 0.2, 0.4, 0.8, and 1.0]. Transformer encoder is used as the feature extractor, with three blocks and single head specifically.

Experiment Results

For our model, 5-fold cross validation was used for training dataset. The dataset was divided into five parts randomly, of which four parts were used for training, and one part was used for test. We repeat the process five times using different test dataset

TABLE 3 | The result of pre-trained models on Pitt dataset.

Model	Embedding	Classifier	Precision	Recall	Accuracy	F1
BertCNN	Bert	CNN	58.85	56.25	56.25	52.79
BertRCNN	Bert	RCNN	-	50.00	50.00	33.33
BertDPCNN	Bert	DPCNN	41.11	47.92	47.92	35.59
ERNIEDPCNN	ERNIE	DPCNN	-	50.00	50.00	33.33
BertLogistic	Bert	Logistic Regression	88	85	86.20	85.60
Transformer+GP	Transformer	Transformer	94.00	89.00	93.50	91.19

every time. Finally, the results of five times were summarized, and the average value was used as the estimation of model performance index. The classification of our model adopts the following indexes: accuracy, precision, recall, and F1 score are used as the final index (44). The relationship between the actual class and predicted class is shown in **Table 1**, and the evaluate metrics in this study are defined as Equations (11–14).

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall Rate} = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{14}$$

Table 2 is the classification scores for AD and CTRL on Pitt Dementiabank datasets, including handcrafted features extracted methods and deep learning methods. As far as we know, SOTA on Pitt corpus is the study of Roshanzamir et al. (48) in 2021, and our method in this paper performs better than SOTA. Also, transformer+FP[25] is the feature project method for text classification, and we did the experiment with this method on our Pitt datasets; the performance of our method is better than the project method with the same datasets. To further compare with some proposed popular pretrained models in recent years, including Bert (37), ERNIE (38), RCNN (31), and DPCNN (32), we do the comparative experiments with the combination models, including BertRCNN, BertDPCNN, BertLogistic, and ERNIEDPCNN models, which are the combination of Bert + CNN, Bert + RCNN, Bert + DPCNN, Bert + Logistic Regression, and ERNIE + DPCNN, respectively. The former is the feature extractor and the latter is the classifier. The evaluation index is shown in **Table 3**.

From **Table 3**, we can find that the performance of first four models is not better, with an accuracy of only 50% or so, BertLogistic model has a better accuracy of 86.2%, and our method gets the best result than these pretrained models. In the meanwhile, to prove the superior of our method, we also test the method on ADReSS and iFLY datasets. The result is shown in **Table 4**, accuracy is improved by 2.1, 4.3, and 2.1%, respectively, on Pitt, ADReSS, and iFLY dataset, which means that the purified features are more discriminative than the features extracted by transformer encoder. Though we do not get SOTA accuracy on ADReSS and iFLY dataset, the performance of our method is improved than transformer.

TABLE 4 | Accuracy on three datasets.

Model	Pitt	ADReSS	iFLY
Transformer	91.4	74.3	81.6
Transformer+GP	93.5	78.6	83.7

DISCUSSION

Why the performance is better after purification? We know that transformer is superior to RNN, CNN on its long-distance reasoning ability, but it is not easy to understand the deep semantic feature vector extracted by transformer as deep learning is a "black-box." The common features in the study are the vector that cannot differentiate for classification in semantic space. They may be the words or sentences that are unimportant, unmeaningful, and irrelevant that may disturb the final classification. Our original dataset is a dialogue of description. It should include some important people, scenes, and ongoing events in the picture. The study (52) pointed out the seed words of the picture should include the following 23 words: boy, girl, woman, cookie, stool, sink, overflow, fall, window, curtain, plate, cloth, jar, water, cupboard, dish, kitchen, garden, take, wash, reach, attention, and see. The sentences including these words are helpful for the classification. Other unrelated words or sentences such as "Can you tell me", "look, there is no people outside" which are unhelpful words or sentences that cannot distinguish cognitive condition. They are, which we think maybe the common features, unhelpful and may even disturb the final classification. When we rule out these words or sentences (i.e., common features) that disturb the classification, the result can be improved correspondingly. The features extracted manually in this area usually include part-of-speech (POS), fluency, semantic feature, lexical richness, and so on. Now, there is an opinion that the features that deep learning extracted automatically maybe are much like the features that people extract manually, and deleting those unhelpful words for the classification can improve the classification performance.

We know that in transformer model, complexity per layer of self-attention is $O(n^2 * d)$, where d is the representation dimension, and n is the sequence length. Our model includes

two sections, one is transformer encoder, the other is feature purification layer which just multiply (-λ) when running back propagation. Both of them can run concurrently and have the same complexity, so the computational complexity of our model is the same as that of self-attention.

CONCLUSIONS

Nowadays, many medical problems used artificial intelligence method to solve (53, 54). which is low cost and convenient. Two methods, that is, feature extraction manually and automatically by deep learning, are usually used to recognize disease. Features extracted method manually based on machine learning does not generalize well, as it needs many special knowledge and annotation to extract features. Due to high cost of manual annotation, it is not feasible to procure numbers of annotated datasets for most clinical tasks. But deep learning does not need any annotation and can finish the process automatically. This paper combines transformer-based model with a feature purification network to improve the classification performance to a large extent. We pretrain transformer and then fine-tune the model on new datasets to transfer learned knowledge to our text classification task. Our work is obviously different from the former studies in AD recognition because none of the former studies improve representation learning of deep learning in this area, as far as we know. The common features extracted by GRL maybe the words that shared by different classifications, or nonimportant words that have small role for classification, ruling out them from traditional representation vector can improve the performance of the model. In addition, we can develop WeChat procedure or APP in mobile device further in order that the elderly can test their cognitive condition at home. So, large volumes of patient's datasets need to be transferred to central cloud server for data analysis, the safety of which is important, and blockchain technology is a better choice which may ensure the security of medical data (53, 55).

Transformer model is still the most widely used deep learning algorithm, but the time complexity of self-attention is higher, which hinders the development of the model, so the improvement of model efficiency is of great importance in the future. Transformer, as the feature extractor we used in this study, can also be replaced by other deep learning algorithms such as Bert, RNN, CNN, and so on; next, we will perfect the work further. In the meanwhile, we also believe that our feature purification method may predict other diseases that language and cognitive impairment related, such as Parkinson's disease, Aphasia, and Autism spectrum disorder. Aphasia is maybe more pronounced as Aphasia is a disease of the brain tissue associated with language function. Our method provides a feasible solution for detecting patients with AD at the doorsteps. Feature purification method for deep learning, as far as we think, is a promising direction to explore in the future.

DATA AVAILABILITY STATEMENT

The public datasets we used can get from the website: https://sla.talkbank.org/TBB/dementia/English/Pitt, or visit our github website: https://github.com/lzy1012/Public-Pitt-Dementiabank-Dataset.

AUTHOR CONTRIBUTIONS

ZY designed the research. QT analyzed the data and interpreted the analysis. NL and ZY wrote the main manuscript text and revised carefully. All authors reviewed and approved the final manuscript.

FUNDING

This research was funded by grants of Natural Science Foundation of Zhejiang Province (LGF20F020009), Anhui Provincial Natural Science Foundation (No. 2108085QF269), and the 4th Graduate Student Innovation and Entrepreneurship Competition of Hangzhou Normal University.

REFERENCES

1. Sousa RM, Ferri CP, Acosta D, Albanese E, Guerra M, Huang YQ, et al. Contribution of chronic diseases to disability in elderly people in countries with low and middle incomes: a 10/66 dementia research group population-based survey. *Lancet*. (2009) 374:1821–30. doi: 10.1016/S0140-6736(09)61829-8
2. Zhou Y, Lu Y, Pei Z. Intelligent diagnosis of Alzheimer's disease based on internet of things monitoring system and deep learning classification method. *Microprocess Microsyst*. (2021) 83:104007. doi: 10.1016/j.micpro.2021.104007
3. Sabat SR. Language function in Alzheimer's disease: a critical review of selected literature. *Lang Commun*. (1994) 14:331–51. doi: 10.1016/0271-5309(94)90025-6
4. Bartha L, Benke H. Acute conduction aphasia: an analysis of 20 cases. *Brain Lang*. (2003) 85:93–108. doi: 10.1016/S0093-934X(02)00502-3
5. Appell J, Kertesz A, Fisman M. A study of language functioning in Alzheimer patients. *Brain Lang*. (1982) 17:73–91. doi: 10.1016/0093-934X(82)90006-2
6. Corkin S DK, Growdon J, Usdin E, Wurtman R. *Some Relationships Between Global Amnesias and the Memory Impairments in Alzheimer's Disease*. Raven Press (1982). p. 149–64.
7. Wechsler AF, Verity MA, Rosenschein S, Fried I, Scheibel AB. Pick's disease. A clinical, computed tomographic, and histologic study with Golgi impregnation observations. *JAMA Neurology*. (1982) 39:287–90. doi: 10.1001/archneur.1982.00510170029008
8. Goodglass H, Kaplan E, Barresi B. *Boston Diagnostic Aphasia Examination Record Booklet*. Lipppincott Williams and Wilkins (2001).
9. Taler V, Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *J Clin Exp Neuropsychol*. (2008) 30:501–56. doi: 10.1080/13803390701550128
10. Yin W, Kann K, Yu M, Hinrich S. *Comparative Study of CNN and RNN for Natural Language Processing. arXiv [Preprint] arXiv*:1702.01923 (2017).
11. Scherer D, Muller A, Behnke S. Evaluation of pooling operations in convolutional architectures for object recognition. *ICANN*. (2010) 19(Suppl 8):92–101. doi: 10.1007/978-3-642-15825-4_10

12. Ganin Y, Lempitsky V. Unsupervised domain adaptation by back propagation. In: *International Conference on Machine Learning*. Florida, FL: PMLR (2015). p. 1180–9.
13. Liu P, Qiu X, Huang X. Adversarial multi-task learning for text classification. *arXiv [Preprint]*. (2017). arXiv:1704.05742. doi: 10.18653/v1/p17-1001
14. Belinkov Y, Poliak A, Shieber SM, Durme BV, Alexander MR. On adversarial removal of hypothesis-only bias in natural language inference. *arXiv [Preprint] arXiv*:1907.04389 (2019). doi: 10.18653/v1/S19-1028
15. Qin L, Xu X, Che W, Zhang Y, Liu T. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). doi: 10.18653/v1/2020.acl-main.565
16. Zhang K, Zhang H, Liu Q, Zhu HS, Chen E. Interactive attention transfer network for cross-domain sentiment classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Hawaii, HI (2019). doi: 10.1609/aaai.v33i01.33015773
17. Qin Q, Hu W, Liu B. Feature projection for improved text classification. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). doi: 10.18653/v1/2020.acl-main.726
18. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F. Domain-adversarial training of neural networks. *J Mach Learn Res*. (2016) 17:2096–30. doi: 10.1007/978-3-319-58347-1_10
19. Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front Aging Neurosci*. (2018) 10:369. doi: 10.3389/fnagi.2018.00369
20. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis*. (2015) 49:407–22. doi: 10.3233/JAD-150520
21. Gosztolya G, Vincze V, Tóth L, Pákáski M, Hoffmann I. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Comput Speech Lang*. (2018) 53:181–97. doi: 10.1016/j.csl.2018.07.007
22. Fraser KC, Fors KL, Kokkinakis D. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Comput Speech Lang*. (2018) 53:121–39. doi: 10.1016/j.csl.2018.07.005
23. Karlekar S, Niu T, Bansal M. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana (2018). p. 701–7.
24. López-De-Ipiña K, Alonso JB, Solé-Casals J, Barroso N, Henriquez P, Faundez-Zanuy M. On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*. (2015) 7:44–55. doi: 10.1007/s12559-013-9229-9
25. Orimaye SO, Wong SM, Wong CP. Correction: deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *Plos ONE*. (2019) 14:e0214103. doi: 10.1371/journal.pone.0214103
26. Fritsch J, Wankerl S, Nöth E. Automatic diagnosis of Alzheimer's disease using neural network language models. In: *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton). (2019).
27. López-de-Ipiña K, Martinez-de-Lizarduy U, Calvo PM, Beitia B, Faundez ZM. Analysis of disfluencies for automatic detection of Mild Cognitive Impartment: a deep learning approach, 2017. In: *International Conference and Workshop on Bioinspired Intelligence (IWOBI)*. (2017). p. 1–4. doi: 10.1109/IWOBI.2017.7985526
28. Palo FD, Parde N. Enriching neural models with targeted features for dementia detection. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence (2019). doi: 10.18653/v1/P19-2042
29. Fritsch J, Wankerl S, Noth E. Automatic diagnosis of Alzheimer's disease using neural network language models. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton). (2019). p. 5841–5.
30. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. *arXiv [Preprint] arXiv*:1802.05365 (2018).
31. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Austin, TX (2015). p. 2267–73
32. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. (2017). p. 562–70. doi: 10.18653/v1/P17-1052
33. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv [Preprint]*. (2016). arXiv: 1409.0473v7. Available online at: https://arxiv.org/pdf/1409.0473.pdf (accessed May 19, 2016).
34. Yang Z, Yang D, Dyer C, He XD, Smola A, Hovy E. Hierarchical attention networks for document classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (2016). p. 1480–9. doi: 10.18653/v1/N16-1174
35. Lin Z, Feng M, Santos CN, Yu M, Xiang B, Zhou B. A structured self-attentive sentence embedding. *arXiv [Preprint] arXiv*:1703.03130 (2017).
36. Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: *Thirtysecond AAAI conference on artificial intelligence*. New Orleans, LA (2018).
37. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv [Preprint] arXiv*:1810.04805 (2018).
38. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence. (2019). p. 1441–51.
39. Liu N, Yuan ZM. Spontaneous language analysis in Alzheimer's disease: evaluation of natural language processing technique for analyzing lexical performance. *J Shanghai Jiaotong Univ*. (2021) 1–8. doi: 10.1007/s12204-021-2384-3
40. Luz S, Haider F, Fuente S, Fromm D, Macwhinney B. *Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge*. Shanghai, China: Interspeech.2020-2571 (2020). doi: 10.21437/Interspeech.2020-2571
41. Orimaye SO, Wong SM, Wong CP, Liang P. Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS ONE*. (2018) 13. doi: 10.1371/journal.pone.0205636
42. Zhao W, Gao H, Chen S, Wang N. Generative Multi-Task Learning for Text Classification. *IEEE Access*. (2020) 8:86380–7. doi: 10.1109/ACCESS.2020.2991337
43. Becker JT, Boiler F, Lopez OL, Saxton J, Mcgonigle KL. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Arch Neurol*. (1994) 51:585–94. doi: 10.1001/archneur.1994.00540180063015
44. Na KS, Cho SE, Zong WG, KimYK. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. *Neurosci Lett*. (2020) 721:134804. doi: 10.1016/j.neulet.2020.134804
45. Yancheva M, Rudzicz F. Vector-space topic models for detecting Alzheimer's disease. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. (2016). p. 2337–46. doi: 10.18653/v1/P16-1221
46. Sirts K, Piguet O, Johnson M. Idea density for predicting Alzheimer's disease from transcribed speech. *arXiv [Preprint]*. (2017). arXiv:1706.04473v1. doi: 10.18653/v1/K17-1033
47. Hernández-Domínguez L, Ratté S, Sierra-Martínez G, Roche-Berguac A. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *DADM*. (2018) 10:260–8. doi: 10.1016/j.dadm.2018.02.004
48. Roshanzamir A, Aghajan H, Baghshah MS. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Inform Decis Mak*. (2021) 21:1–14. doi: 10.1186/s12911-021-01456-3
49. Pan Y, Mirheidari B, Reuber M, Venneri A, Blackburn D, Christensen H. Automatic hierarchical attention neural network for detecting AD. *Proc Interspeech*. (2019) 2019:4105–9. doi: 10.21437/Interspeech.2019-1799
50. Li B, Hsu YT, Rudzicz F. Detecting dementia in mandarin Chinese using transfer learning from a parallel corpus. *arXiv [Preprint] arXiv:1903.00933*. (2019). doi: 10.18653/v1/N19-1199

51. Fraser KC, Linz N, Li B, Fors KL, Rudzicz F, König A, et al. Multilingual prediction of Alzheimer's disease through domain adaptation and conceptbased language modelling. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Florence. (2019). p. 3659–70.
52. Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology*. (2000) 14:71–91. doi: 10.1080/026870300401603
53. Dhanamjayulu C, Nizhal UN, Maddikunta PKR, Gadekallu TR, Lwendi C, Wei C. Identification of malnutrition and prediction of BMI from facial images using real-time image processing and machine learning. *IET Image Process*. (2021) 16:647–58. doi: 10.1049/ipr2.12222
54. Thippa RG, Bhattacharya S, Maddikunta P, Hakak S, Tariq U. Antlion re-sampling based deep neural network model for classifification of imbalanced multimodal stroke dataset. *Multimed Tools Appl*. (2020) 26:1–25. doi: 10.1007/s11042-020-09988-y
55. Wang W, Qiu C, Yin Z, Srivastava G, Gadekallu TR, Alsolami F. Blockchain and PUF-based lightweight authentication protocol for wireless medical sensor networks. *IEEE Internet Things J*. (2021). doi: 10.1109/JIOT.2021.3117762

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.