Leveraging Pretrained Representations with Task-related Keywords for Alzheimer's Disease Detection

Jinchao Li[1], Kaitao Song[2], Junan Li[1], Bo Zheng[1], Dongsheng Li[2], Xixin Wu[1], Xunying Liu[1], Helen Meng[1]
[1]The Chinese University of Hong Kong, Hong Kong SAR, China; [2]Microsoft Research Asia, Shanghai, China

ICASSP 2023
Paper ID: 6063

Introduction

Motivation:

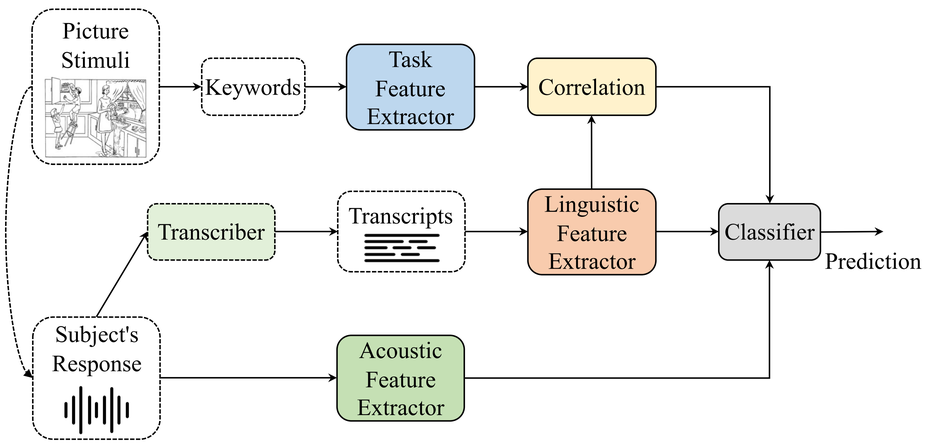
- Detection of Alzheimer's disease (AD) is crucial for timely intervention to slow down disease progression
- Speech offers non-intrusive, accessible, affordable and automatic means for detecting AD
- Transfer learning-based linguistic embeddings have proven to be valuable for AD detection, but current acoustic embeddings leave much to be desired
- Visual-spatial abilities are accessed in cognitive tasks, but have been overlooked in previous AD detection research

Objective & Contributions:

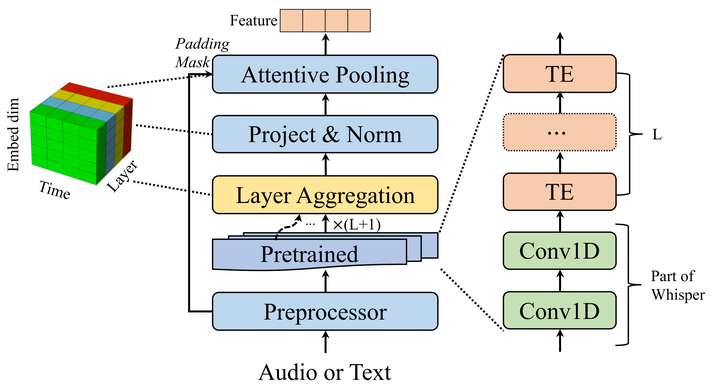
- Aggregation methods to leverage pretrained representations effectively
- Incorporate visual information by modeling relationship between speakers' utterances and picture stimuli

Approach

Overview:

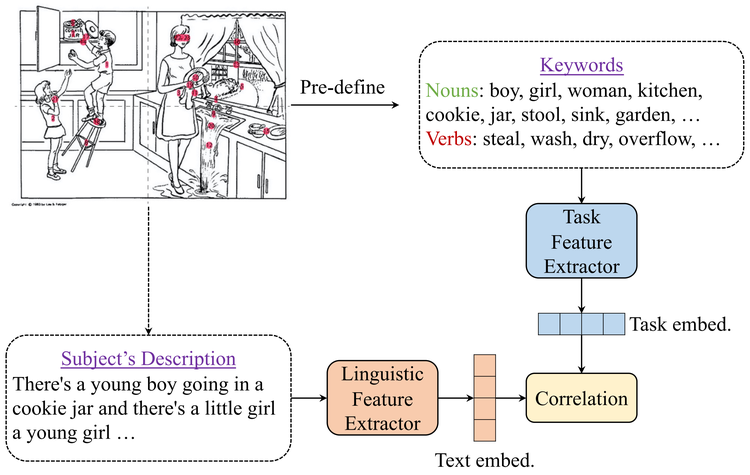


Feature Extractor:



Approach

Task-related text-visual correlation:



Experimental Setup

Dataset: ADReSS Corpus[Luz'20]

		AD	non-AD
Train	Male	24	24
	Female	30	30
Test	Male	11	11
	Female	13	13

Model & training details:

- Pretrained models: Whisper, BERT
- Binary cross-entropy loss, AdamW optimizer

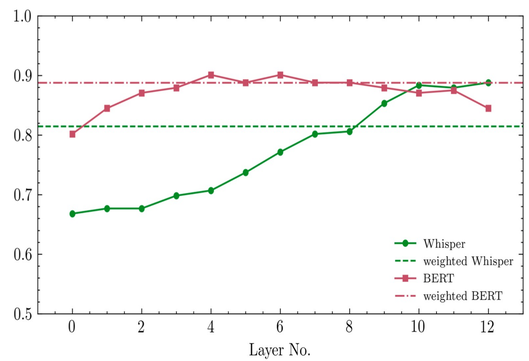
Evaluation Protocols:

- Binary classification accuracy and F1 scores

Results

Results using different single layers:

- Whisper: top layers (semantic)
- BERT: middle layers (syntactic)



Results

Comparison of different aggregation methods:

Feature	Layer AGG	Time AGG	Accuracy(%)	F1-score(%)
ComParE [1]	-	Mean	62	62
VGGish [2]	Top	Mean	72.92	72.62
OpenL3 [3]*	Top	Mean	81.25	81.20
Wav2vec 2.0	WS	Mean	77.50	76.69
HuBERT	WS	Mean	78.88	78.79
WavLM	WS	Mean	79.74	79.66
Whisper	WS	Mean	79.31	79.30
WavLM	WS	Attention	82.33	82.33
Whisper	WS	Attention	81.47	81.46
WavLM	MS	Attention	85.78	85.78
Whisper	MS	Attention	**88.79**	**88.79**

Comparison of different visual keyword sets:

Keyword Type	Accuracy(%)	F1-score(%)
None	52.34	34.36
Nouns	85.16	85.11
Verbs	83.59	83.58
Nouns + Verbs	**85.94**	**85.88**

Combination of multimodal features:

Feature	Modality	Accuracy(%)	F1-score(%)
ComParE [1]	A	62	62
VGGish [2]	A	72.92	72.62
OpenL3 [3]*	A	81.25	81.20
Linguistics [1]	T	75	71
ERNIE [4]	T	85.4	85.3
Glove [5]	T	89.6	-
BERT+Roberta [6]	T	**91.7**	**91.7**
Temporal + Glove [5]	A+T	**91.67**	-
Whisper	A	88.79	88.79
BERT	T	90.09	90.07
Correlation	V	85.94	85.88
Whisper + BERT	A+T	**91.19**	**91.19**
Whisper + Correlation	A+V	89.84	89.83
BERT + Correlation	T+V	90.62	90.60
Whisper+BERT+Correlation	A+T+V	**91.41**	**91.38**

Conclusion

- Performance gap between acoustic and linguistic models has significantly narrowed compared to the past
- Incorporate visual information to assess cognitive abilities in visual-spatial, attention domains
- Superior performance achieved on ADReSS corpus