ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Direct versus intermediate multi-task transfer learning for dementia detection from unstructured conversations

Daniel P. Kumpik[1], Yoav Ben-Shlomo[2], Elizabeth Coulthard[3], Alex Hepburn[1] and Raul Santos-Rodriguez[1] *

1- Department of Engineering Mathematics, University of Bristol, Bristol, UK

2- Department of Population Health Sciences, University of Bristol, Bristol, UK

3- Department of Translational Health Sciences, University of Bristol, Bristol, UK

Abstract. Leveraging unstructured conversations for detecting early dementia may be possible through information transfer from more systematically constrained representations. To explore whether cross-domain (from semi-structured to unstructured) transfer learning improves dementia classification from conversational speech, we fine-tuned a BERT-family model using semi-structured narratives. We further fine-tuned on naturalistic conversations recorded in the home, but found that direct transfer from BERT to conversations was more effective for improving generalization. These findings show scope to directly leverage unstructured language samples for in-the-wild dementia detection.

1 Introduction

Alzheimer's disease (AD) affects 42 million people globally [1]. Predicting future AD during preclinical mild cognitive impairment (MCI) could enable early intervention, but MCI is harder to detect [2]. However, language is a rich biomarker; AD and MCI disrupt lexical, syntactic, semantic and structural linguistic faculties [3]. Conversational speech may be more revealing of MCI, which is pronounced under cognitive load [4]. Unfortunately, few publicly available, large conversational datasets exist for training diagnostic machine learning (ML) models. Picture description datasets like Dementiabank – which contains semi-structured dyadic interactions [5] – can somewhat counter this lack of diversity, but are contextually constrained and do not reflect natural dialogue. Free conversation demands memory, theory of mind and executive functioning, but preparing conversational datasets with enough inter-individual variability to generalize across scenarios is time-consuming, and subject to availability of clinical populations.

Transfer learning (TL) leverages models pretrained on large datasets in related domains to improve performance. Large language models (LLMs) such as Bidirectional Encoder Representations from Transformers (BERT [6]) encode

*This research was funded by an MRC Momentum award (grant MC/PC/16029), the EPSRC SPHERE Interdisciplinary Research Collaboration (grant EP/K031910/1) and the EPSRC Centre for Doctoral Training in Digital Health and Care, University of Bristol (grant EP/S023704/1). RSR and AH are funded by the UKRI Turing AI Fellowship (EP/V024817/1). Analysis of behavioral data was partially funded by the BRACE charity.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

contextual linguistic and semantic patterns, which when fine-tuned on speech from AD patients can identify subtle cognitive decline [7]. Context-sensitive BERT embeddings yield excellent AD classification performance, sometimes performing better than engineered features fed to standard classifiers [8].

In this study, we pretrained BERT on a large semi-structured dyadic speech dataset from Dementiabank and hypothesized that further TL could improve dementia detection from a smaller dataset of naturalistic conversations acquired while participants watched TV at home. We saw a modest improvement at the tuned epoch number (4), but overall direct tuning of conversations produced better performance. These results suggest a promising diagnostic role for direct tuning of ML models for passive in-home monitoring of cognitive decline.

2 Methods

2.1 A dementia classifier for intermediate TL from semi-structured speech

Dataset We used the ADReSS subset of speech transcripts obtained from picture description (Dementiabank Pitt Corpus [5]; 78 control, 78 AD; 108 train, 48 test). These are monologues by healthy controls and dementia patients (denoted PAR), plus interjections from a clinical investigator (denoted INV). We used custom tokens for non-verbal cues: [SINGS], [WHISPERS], [SIGHS], [GASPS], [YAWNS], [LAUGHS], [FP] (filled pauses), and for turn changes ([TURNTOKEN_INV], [TURNTOKEN_PAR]). We segmented component utterances into contextualized sequences using each utterance plus the previous two from either speaker, separated by the standard SEP (`</s>`) token.

Modeling Utterance sequences were tokenized (maximum 128 tokens), then we tuned a multi-task DistilRoBERTa-base [9] pipeline to classify cognitive status and within-conversation **speaker_id**. We assumed a situation in which a clinician might recommend a passive dementia-detecting app to a patient and their conversational partner, for which the identities of the transcribed voices in a conversation were known (that is, person with suspected cognitive impairment (PAR) versus conversational partner (INV)). The model's primary task was to classify conversations as either between two cognitively healthy individuals or involving one person in cognitive decline. Thus, speaker detection within a conversation would be trivial, but could still contribute to learning the speaker-specific patterns that might emerge in dyadic conversations between either two people without dementia, or between one with and one without suspected dementia.

We routed to different speaker classifiers based on predicted dementia status per conversation (Figure 1). To link fragmentary utterances, we concatenated **conversation_id** to DistilRoBERTa's pooled output, feeding it to dementia and speaker classification branches. Minority speakers and conversations were upsampled in training and test sets for interpretability. Hyperparameters for MCI vs. AD and **speaker_id** classification were tuned with 5-fold cross-validation (CV) to minimize cross-entropy loss. To prioritize dementia classification, we

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

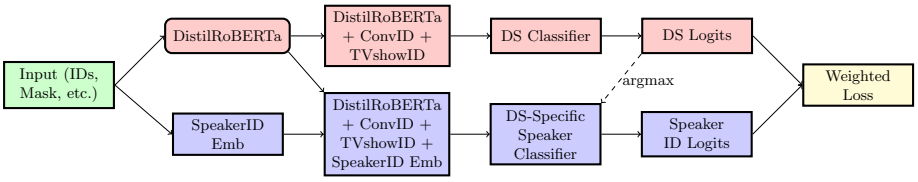

Fig. 1: Architecture of the Multi-Task Multi-Speaker DistilRoBERTa model. TVshowID was utilized in CUBOId only.

added $1 - \text{accuracy}_{\text{dementia}}$ to the loss and computed a weighted mean of dementia and speaker classification losses. A linear learning rate scheduler with 10% warmup, AdamW optimizer, and early stopping (patience = 1, threshold = 1e-2) were used. Hyperparameters were tuned via Tree-structured Parzen Estimator sampling, and the model with the best CV loss then trained on the full dataset.

2.2 A dementia classifier from unstructured conversations

Dataset The "TV task" of the ContinUous behavioral Biomarkers Of cognitive Impairment (CUBOId) study aimed to facilitate diagnosis from conversations. Over 18 months (2–3 intervals), 7 MCI/AD patients (5 MCI) and healthy partners recorded 30min TV-watching conversations across 5 days [10]. Conversations were transcribed and processed as in Section 2.1.

Modeling We included hand-crafted lexical, connected language, semantic and sentiment features as computed manually or using widely available tools [3], and concatenated these to the BERT output, **conversation_id** and **speaker_id**. Alongside the linguistic features, for this dataset we also included a **TVshow_id** to account for the different stimulation expected from different types of show. Patients were designated as PAR and partners as INV. Classifying CUBOId conversations via intermediate TL thus required adapting to natural conversational dynamics, and to domain shifts from healthy couples to pairs with one partner having MCI. We split the data into training and test sets by testing block at a 1:1 or 2:1 ratio based on block completion. We established baseline performance using the preprocessing and modeling protocol from Sec. 2.1, retaining ADReSS hyperparameters but training for 1-15 epochs in all subsequent experiments.

2.3 Transfer learning (TL)

We first established ADReSS and CUBOId baselines, including after training on ADReSS without and with the speaker embeddings, and also after testing the trained ADReSS speaker embedding model on CUBOId before TL. We also fine-tuned DistilRoBERTa on CUBOId alone using the optimal ADReSS hyperparameters. All subsequent TL was performed over 1-15 epochs with the optimal ADReSS hyperparameters, including the number of unfrozen DistilRoBERTa layers (i.e., 3). Prior to TL we retained the weights for the speaker embedding layer and the DistilRoBERTa portion of the fine-tuned model; the classifier heads

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

were re-initialized along with the optimizer. After TL we tested on CUBOId, and on ADReSS to evaluate positive transfer or catastrophic forgetting.

3 Results

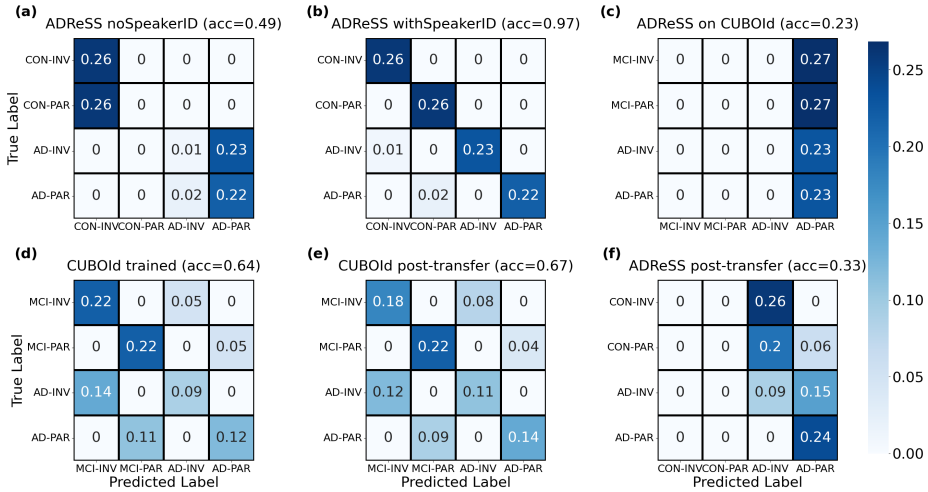

Fig. 2: Confusion matrices: a) tuned and evaluated on ADReSS semi-structured narratives, no speakerID, b) tuned and evaluated on ADReSS with speakerID, c) tuned on ADReSS and evaluated on CUBOId natural conversations, d) tuned and evaluated on CUBOId, e) tuned on ADReSS, further trained on CUBOId, evaluated on CUBOId, f) tuned on ADReSS, further trained on CUBOId, evaluated on ADReSS. *INV,* conversational partner. *PAR,* patient.

When the ADReSS model was trained without **speaker_id** (Figure 2a) it was required to learn **speaker_id** from the speaker labels and failed to do so, although it still managed to correctly classify approximately 97% of utterances by house diagnosis. Figure 2b shows the difference in performance with **speaker_id** provided. It was then able to tell the speakers apart very well without compromising conversational-level performance. In the absence of any trained exposure to the CUBOId naturalistic conversations, this model performed poorly on the CUBOId test set (Figure 2c). Training a model solely on CUBOId conversations produced mixed performance, yielding an accuracy of 64% which was better than chance at both the conversation (50%) and sub-conversation (25%) level (Figure 2d). However, the presence of many off-diagonal predictions indicates that the model struggled to distinguish both between conversations containing a speaker with or without dementia and between speakers within those conversations. TL with the tuned number of epochs produced an improvement in overall classification accuracy relative to CUBOId-only training of about 3% (Figure 2e), although when tested over epochs it was clear that direct fine-tuning of conversations produced better performance overall (Figure 3). Finally, we found evidence for catastrophic forgetting of the previously well-learned ADReSS dataset

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933.
Available from http://www.i6doc.com/en/.

after TL with CUBOId, with the model again showing bias for classifying all conversations as containing an individual with AD (Figure 2f).

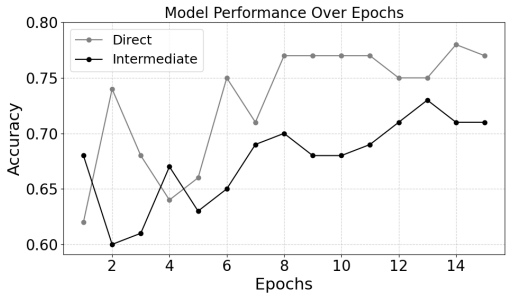

Fig. 3: Accuracy for the direct and intermediate-transfer learning strategies as a function of number of epochs used for transfer.

4 Discussion

In this study we used a hybrid multi-task TL approach to understand the extent to which the widely available ADReSS dataset of semi-structured speech from dementia patients can be leveraged via TL to improve diagnosis of dementia from unstructured conversations as recorded passively in the home. For semi-structured speech at baseline, we found that simply permitting a BERT-family model to separately classify speaker identity contingent upon a prior hard prediction of dementia status produced state-of-the-art classification performance at the conversation level. This was not adversely affected by explicitly tagging the speaker ID. When we used the latter configuration for TL with an unstructured conversational speech corpus of unpredictable verbal content, we saw an improvement in diagnostic prediction accuracy of 3%. However, this was not stable as a function of the number of epochs used for training, suggesting that direct transfer onto conversations is a better strategy and highlighting the need for careful epoch selection to avoid performance degradation. We also observed that the model exhibited catastrophic forgetting after TL with CUBOId when tested again on ADReSS, consistent with well-documented TL limitations. These findings imply a vital supporting role for directly leveraging unstructured speech to realize the potential clinical, diagnostic and prognostic benefits of natural speech as recorded at home for enhancing early dementia detection.

TL with LLMs has become ubiquitous for dementia detection due to their ability to leverage complex pretrained language patterns and semantic structures for detection of the subtle cognitive changes in mild cognitive impairment and early dementia [8]. These studies often fine-tune their models using semi-structured speech corpora such as Dementiabank and performance accuracies of above 90% correct are not uncommon, especially for well-balanced datasets like the ADReSS subset of the Dementiabank Pitt Corpus [11]. Our ADReSS accuracy of 97% indicates excellent performance, and while this did not transfer to CUBOId conversations, those conversations apparently contained enough in-

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

formation for direct TL from DistilRoBERTa to CUBOId to produce reasonable classification performance. This suggests that natural conversations as recorded in the home may be a viable diagnostic tool for clinicians to consider.

To outperform the intermediate TL pipeline, the direct-trained model adapted to diverse linguistic and interaction patterns, emphasizing the need for datasets closer to real-world use cases. While semi-structured datasets are valuable for pretraining, unstructured datasets like CUBOId provide crucial context for clinical applications. However, even within individuals natural conversations yield highly variable linguistic output over time, and are subject to both environmental and statistical noise [12]. This makes them challenging for ML, particularly end-to-end models incorporating automatic speech recognition. Future work should therefore focus on optimizing TL pipelines to balance the strengths of semi-structured and unstructured datasets, potentially through improved fine-tuning strategies or architectures designed for conversational data.

References

[1] Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H Michael Arrighi. Forecasting the global burden of alzheimerâs disease. *Alzheimer's & dementia*, 3(3):186–191, 2007.

[2] Jeremy Brown. The use and misuse of short cognitive tests in the diagnosis of dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(6):680–685, 2015.

[3] Natasha Clarke, Peter Foltz, and Peter Garrard. How to do things with (thousands of) words: Computational approaches to discourse analysis in alzheimer's disease. *Cortex*, 129:446–463, 2020.

[4] Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey A. Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:2081–2090, 2011.

[5] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech, 2021.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Ziming Liu, Eun Jin Paek, Si On Yoon, Devin Casenhiser, Wenjun Zhou, and Xiaopeng Zhao. Detecting alzheimer's disease using natural language processing of referential communication task transcripts. *Journal of Alzheimer's disease*, 86(3):1385–1398, 2022.

[8] Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*, 2020.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[10] Daniel Paul Kumpik, Raul Santos-Rodriguez, James Selwood, Elizabeth Coulthard, Niall Twomey, Ian Craddock, and Yoav Ben-Shlomo. A longitudinal observational study of home-based conversations for detecting early dementia: protocol for the CUBOId TV task. *BMJ Open*, 12(11), 2022.

[11] Lovro Matošević and Alan Jović. Accurate detection of dementia from speech transcripts using roberta model. In *2022 45th Jubilee International Convention on Information, Communication, and Electronic Technology (MIPRO)*, pages 443–447, 2022.

[12] Kathleen C. Fraser and Majid Komeili. Measuring cognitive status from speech in a smart home environment. *IEEE Instrumentation & Measurement Magazine*, 24(6):13–21, 2021.