Learning multiview embeddings for assessing dementia

Chloé Pou-Prom [1,2,3], **Frank Rudzicz**[1,2,3]
[1] Toronto Rehabilitation Institute - UHN, Toronto, Canada
[2]Vector Institute, Toronto, Canada
[3] Department of Computer Science, University of Toronto, Canada
{ chloe, frank }@cs.toronto.edu

Abstract

As the incidence of Alzheimer's Disease (AD) increases, early detection becomes crucial. Unfortunately, datasets for AD assessment are often sparse and incomplete. In this work, we leverage the multiview nature of a small AD dataset, DementiaBank, to learn an embedding that captures different modes of cognitive impairment. We apply generalized canonical correlation analysis (GCCA) to our dataset and demonstrate the added benefit of using multiview embeddings in two downstream tasks: identifying AD and predicting clinical scores. By including multiview embeddings, we obtain an F1 score of 0.82 in the classification task and a mean absolute error of 3.42 in the regression task. Furthermore, we show that multiview embeddings can be obtained from other datasets as well.

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative progressive disease whose symptoms include memory loss, disorientation, and behavioral issues (Ballard et al., 2011). In 2017, 5.7 million Americans were living with AD, and the disease accounted for $11.4 billion in healthcare costs in the United States (Alzheimer's Association, 2018). AD is diagnosed through clinician-administered questionnaires, such as the Mini-Mental State Examination (MMSE), which assigns a score between 0 and 30 based on responses to questions testing memory, recall, and orientation (Folstein et al., 1975). For context, a MMSE score of 23 and below is associated with cognitive decline.

AD affects language and some of its symptoms include difficulties in word-finding and changes in the voice. Detecting these subtle changes can help identify AD at an early stage. Indeed, many studies have applied a combination of natural language processing and machine learning techniques to detect AD. On the DementiaBank (DB) dataset, which includes audio files and corresponding transcripts of participants completing a picture description task, Wankerl et al. (2017) employed an n-gram based approach to classify between participants with and without AD. On the same dataset, Fraser et al. (2015) extracted an extensive list of lexicosyntactic features from the transcripts and identified participants with AD with an accuracy of 81%. More recently, Hernández-Domínguez et al. (2018) looked at the information content units of the pictures and compared them to healthy population-specific references to achieve an F-score of 0.81.

Predicting clinical scores is a harder task and is more common in image processing, where researchers make use of brain scans. For example, Huang et al. (2016) used MRI scans from 805 subjects and relied on the longitudinal aspect of their dataset to predict MMSE scores. Specific to the DB dataset, Yancheva et al. (2015) extracted linguistic features and used a bivariate dynamic Bayes net to represent the longitudinal nature of the data, and obtained a mean absolute error (MAE) of 3.83. Focusing on subjects with larger samples of data yielded a MAE of 2.91.

In instances where multiple views of the same data are available, it makes sense to learn a vector representation (an embedding) that encapsulates the different sources of information. Benton et al. (2016) used different representations of their data (e.g., bag-of-words, word vectors) to learn multiview embeddings for Twitter users, and obtained promising results when evaluating their embeddings in downstream prediction tasks.

In this work, we leverage the multiview nature of DB to learn an embedding for each user. We evaluate the utility of the multiview embedding in two downstream tasks: classification of AD vs non-AD participants, and clinical score prediction.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2812–2817
Brussels, Belgium, October 31 - November 4, 2018.

2 Methods

2.1 Dataset

We use the DementiaBank (DB) corpus (Becker et al., 1994), which consists of adults aged 44 and older, assigned to either the *'Dementia'* ($N = 167$) or *'Healthy'* ($N = 97$) group based on a battery of neuropsychological tests and on their medical histories. In DB, participants performed the "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983), in which they verbally describe the contents of a picture. Additionally, participants in the *'Dementia'* group completed the category fluency (i.e., naming words belonging to a given category), letter fluency (i.e., naming words that start with a given letter), sentence construction, and story recall tasks. The picture description and both fluency tasks were professionally transcribed and annotated with instances of filled pauses. Previous experiments in the literature on DB have been limited to the picture description task, most likely because the other tasks are not available for all participants.

2.2 Linguistic features

From transcripts of the picture description, category fluency and letter fluency tasks, we extract 565 linguistic features[1]. We compute lexical features (e.g., the mean number of syllables per word, the vocabulary richness as measured by the type-token-ratio[2]), semantic features (e.g., the mean specificity of words as measured by their depth in WordNet[3]), and syntactic features (e.g., the proportion of various parts-of-speech tags, such as nouns and adjectives). We also automatically extract various subjective measures, such as the mean imageability (i.e., a word's ability to evoke a mental image) and the mean age-of-acquisition of words using norms derived from the Bristol (Stadthagen-Gonzalez and Davis, 2006) and Gilhoolie-Logie (Gilhooly and Logie, 1980) norms. Finally, we train an LDA model of 100 topics (Blei et al., 2003) using a Wikipedia snapshot, and compute the topic probabilities for each transcript.

2.3 Learning a multiview embedding

We apply generalized canonical correlation analysis (GCCA) to our dataset to obtain a multiview embedding. We use GCCA as described by Benton et al. 2016 to learn linear transformations U_j which project different views of our data into the embedding G. In our experiments, we consider the following views of DB: linguistic features of the picture description, category fluency and letter fluency tasks, and demographic information.

Given $X \in \mathbb{R}^{d \times N}$, $X' \in \mathbb{R}^{d' \times N'}$, where N is the total number of data points, N' is the total number of data points for which all views J are available, and d and d' are the dimensions of X and X'; let X_j and X'_j denote views j of X and X'. Here, $j \in \{PD, CAT, LET, DEM\}$, which correspond to the picture description, category fluency, and letter fluency linguistic features, and demographic information, respectively.

1. We use GCCA to learn U_j from X'_{PD}, X'_{CAT}, X'_{LET}, X'_{DEM}, such that:

$$\underset{U_j, G'}{\text{minimize}} \sum_{j \in J} ||G' - U_j^T X'_j||_F^2$$

   $U_j \in \mathbb{R}^{d_j \times k}$, $G' \in \mathbb{R}^{k \times N'}$.

2. We compute $G = U_{PD}^T X_{PD}$. Since $U_{PD} \in \mathbb{R}^{d_{PD} \times k}$ and $X_{PD} \in \mathbb{R}^{d_{PD} \times N}$, then $G \in \mathbb{R}^{k \times N}$.

3. We concatenate G to a subset of the picture description linguistic features, X^*_{PD}, to obtain $C = (X^*_{PD}, G)$, where $C \in \mathbb{R}^{(k + d^*_{PD}) \times N}$.

4. We use the augmented set of features C for two downstream tasks: AD classification and clinical score prediction.

3 Results

We run all experiments with 10-fold cross validation and test various settings of k, the dimension of the multiview embedding.

3.1 GCCA and classification

We select the top n linguistic features, ordered through a one-way ANOVA and concatenate them with multiview embeddings of size k. The $n + k$ features are then given as input to a random forest

[1] The code to extract these is being made available at https://github.com/SPOClab-ca/COVFEFE.

[2] The type-token ratio is obtained by dividing the number of types (i.e., the total number of *different* words) by the number of tokens (i.e., the total number of words).

[3] WordNet (Miller, 1995) is a lexical database which groups English words into collections of synonyms. The database is ordered from most generic (e.g., "plant") to most specific (e.g., "rose").

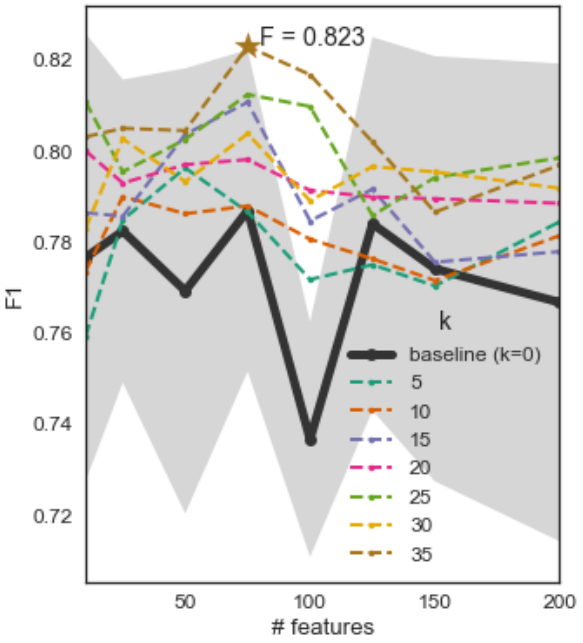

Figure 1: Classification of *'Dementia'* vs *'Healthy'* participants in DB. We report the F1 score as we increase the number of significant features used in our random forest classifier. The dark black line denotes the baseline (i.e., no GCCA) with the shaded grey region corresponding to the standard deviations, and the colored dotted lines denote experiments with multiview embeddings of size k.

classifier with 100 decision trees, and we report the F1 scores in Figure 1. Our best classification result ($F1 = 0.823 \pm 0.032$) is achieved with a multiview embedding of size $k = 35$ using the best $n = 75$ linguistic features. Adding GCCA embeddings improves classification results: an ANOVA test reveals a significant difference between F1 results with and without GCCA ($F = 15.85$, $p = 0.00018$), and a post-hoc Tukey's honest significant difference test reveals that F1 scores are significantly higher in experiments using GCCA embeddings ($p = 0.00018$).

Next, we look at multiview embeddings generated from different combinations of DB views, and report our F1 scores in Table 1 for embeddings of size $k = 35$ and using the top $n = 75$ features. Adding multiview embeddings always improves classification, and we obtain our best results by learning an embedding from the picture description and category fluency views.

3.2 GCCA and regression

To predict MMSE scores, we select the top n best features, ordered through a continuous one-way ANOVA, and concatenate them with our multiview embedding of size k. The $n + k$ features are then given as input to a linear regression model and we report the mean absolute error (MAE). 10-fold cross-validation results are given in Figure 2. Our lowest MAE of 3.412 ± 0.300 was obtained using a GCCA embedding of size $k = 5$ and retaining the top $n = 75$ linguistics features. Adding multiview embeddings yields the best results, but an ANOVA test reveals no significant difference ($F = 0.41, p = 0.53$).

Views	F1
None (baseline)	*0.782* (0.042)
$X'_{PD}, X'_{CAT}, X'_{LET}, X'_{DEM}$	0.811 (0.045)
$X'_{PD}, X'_{CAT}, X'_{LET}$	0.817 (0.037)
$X'_{PD}, X'_{LET}, X'_{DEM}$	0.815 (0.043)
$X'_{PD}, X'_{CAT}, X'_{DEM}$	0.818 (0.042)
X'_{PD}, X'_{CAT}	**0.824** (0.052)
X'_{PD}, X'_{LET}	0.805 (0.055)
X'_{PD}, X'_{DEM}	0.816 (0.057)

Table 1: Classification results with GCCA applied on different views of DB. We report the different views used to learn our multiview embedding and the resulting F1 scores (with standard deviation in parenthesis) on 10-fold cross-validation experiments. X' denotes the data points for which all views are present (i.e., the data used to learn multiview embeddings), and the subscripts PD, CAT, LET, DEM are used to represent the following views: picture description text features, category fluency text features, letter fluency text features, and demographic information.

3.3 Learning a multiview embedding from another dataset

We then perform the same experiments as described in sections 3.1 and 3.2, but we learn our U_{PD} linear projection matrix with a different dataset. We use *Talk2Me*[4], an online language assessment from the University of Toronto, in which participants use the web to complete a variety of language tasks, including the picture description task, the vocabulary task, the Winograd task (Levesque et al., 2011), and the word fluency task (including both category and letter fluency). For all tasks in *Talk2Me*, we transcribe the audio recordings using the Kaldi open-source automatic speech recognition engine (Povey et al., 2011), and extract the same set of text features as in Section 2.2. Next, we apply GCCA to learn a

[4] https://www.cs.toronto.edu/talk2me

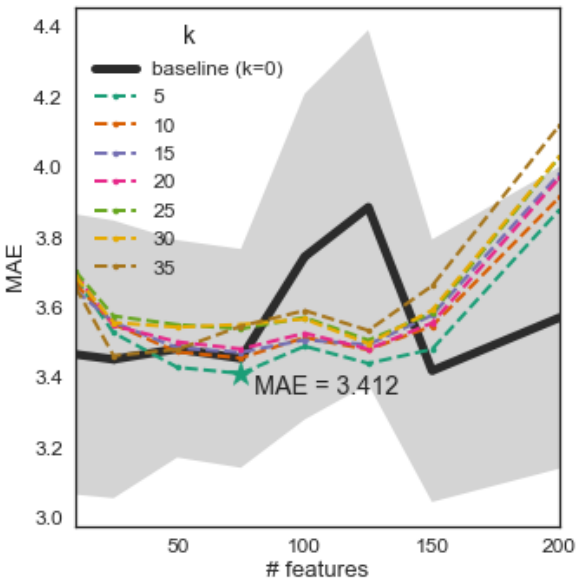

Figure 2: Clinical score prediction. We report the mean absolute error (lower is better) in predicting MMSE score as we increase the number of significant features used in our linear regression model. The black line denotes the baseline (i.e., no GCCA) with the shaded grey region corresponding to the standard deviations, and the colored dotted lines denote experiments with multiview embeddings of size k.

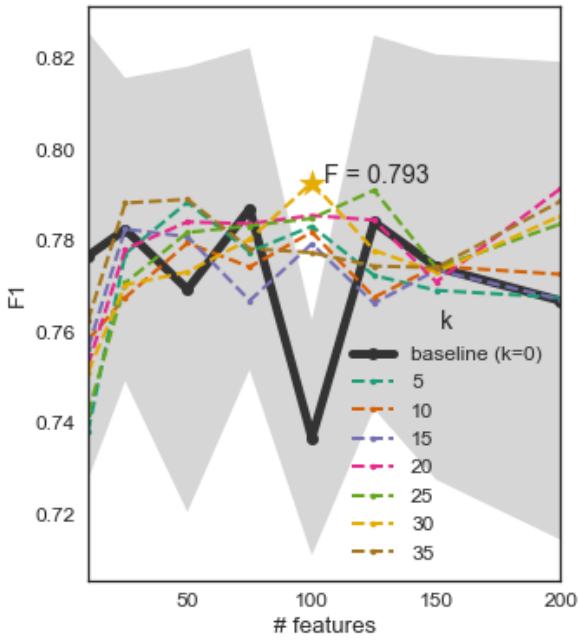

Figure 3: Classification task using multiview embeddings learned from the *Talk2Me* dataset. We report the F1 scores as we increase the number of significant features. The black line denotes the baseline (i.e., no GCCA) with the shaded grey region corresponding to the standard deviations, and the dotted colored lines denote experiments with multiview embeddings of size k obtained through GCCA on *Talk2Me* views.

multiview embedding from the following views: picture description, story recall, vocabulary, fluency, and image naming tasks, and demographics.

As in previous experiments, we concatenate the multiview embedding with the DB picture description linguistic features, and use these to classify AD participants and to predict MMSE scores. In the regression task, the GCCA features from *Talk2Me* greatly hinder performance. The best result we obtain with *Talk2Me* multiview embeddings is an MAE of 3.929 ± 1.37. In classification, we observe improvements, as shown in Figure 3, and obtain an F1 of 0.793 ± 0.052. However, an ANOVA test reveals no significant difference with multiview embeddings ($F = 0.45$, $p = 0.50$).

4 Discussion

In our experiments, we use GCCA to learn a multiview embedding and augment our existing set of features. The multiview embedding consists of a vector representation which encapsulates information from various sources of information (i.e., the picture description task, the category and letter fluency tasks, and demographic data). We hypothesize that the additional information contained in this embedding would be useful in downstream tasks such as classification and regression. Indeed, the multiview embedding obtained from DB improves AD detection and clinical score prediction. Similarly, we also observe an improvement in classification when using a multiview embedding learned from a normative dataset.

Our results are better in the classification task than in the regression task, since the MMSE score is mainly used as a screening tool (i.e., determining if a person has AD or not) and has restricted sensitivity, especially for identifying milder stages of AD (Trzepacz et al., 2015).

5 Conclusion

We have shown that we can make use of the multiview aspect of a small dataset such as DB to learn a multiview embedding. This embedding can subsequently be used to improve models for classification and regression. In our experiments, multiview embeddings allowed the use of both the category and letter fluency data in DB, even though they were only available for the *'Dementia'* participants. Benefits are also possible using secondary datasets to learn multiview embeddings.

Extracting acoustic features – such as pause ratio, pitch, and Mel-frequency cepstral coefficients (MFCCs) – and treating them as an additional view is part of our future work. Furthermore, we will look into other secondary datasets as well as different approaches of obtaining multiview embeddings. While GCCA allows for an arbitrary number of views, it is limited in that it only learns linear projections to the embedding space. A possible alternative is deep generalized canonical correlation analysis (DGCCA), which makes use of neural networks to learn non-linear mappings to the embedding space (Benton et al., 2017).

Acknowledgments

This work was partially funded by an NSERC Discovery grant (RGPIN 435874) and by a Young Investigator award by the Alzheimer Society of Canada, both held by Rudzicz.

References

Alzheimer's Association. 2018. 2018 Alzheimer's disease facts and figures. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 13(4):325–373.

Clive Ballard, Serge Gauthier, Anne Corbett, Carol Brayne, Dag Aarsland, and Emma Jones. 2011. Alzheimer's disease. *Lancet*, 377(9770):1019–31.

James T. Becker, Francois Boller, Oscar I. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology*, 51(6):585–594.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning Multiview Embeddings of Twitter Users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 14–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. 2017. Deep Generalized Canonical Correlation Analysis. *arXiv preprint, arXiv:1702.02519*.

David M. Blei, Andrew Y. Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. 1975. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric research*, 12(3):189–198.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2015. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

K. J. Gilhooly and R. H. Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427.

Harold Goodglass and Edith Kaplan. 1983. Boston Diagnostic Aphasia Examination.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. 2018. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268.

Lei Huang, Yan Jin, Yaozong Gao, Kim-Han Thung, and Dinggang Shen. 2016. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of Aging*, 46:180–191.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI spring symposium: Logical formalizations of commonsense reasoning*, volume 46, page 47. Association for the Advancement of Artificial Intelligence (AAAI).

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, Hawaii, USA. Institute of Electrical and Electronics Engineers (IEEE) Signal Processing Society.

Hans Stadthagen-Gonzalez and Colin J. Davis. 2006. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605.

Paula T Trzepacz, Helen Hochstetler, Shufang Wang, Brett Walker, and Andrew J. Saykin. 2015. Relationship between the Montreal Cognitive Assessment and Mini-mental State Examination for assessment of mild cognitive impairment in older adults. *BioMed Central Geriatrics*, 15(1):107.

Sebastian Wankerl, Elmar Nöth, and Stefan Evert. 2017. An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language. In *Proceedings of Interspeech 2017*, pages 3162–3166, Stockholm, Sweden. International Speech and Communication Association.

Maria Yancheva, Kathleen C. Fraser, and Frank Rudzicz. 2015. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies*, page 134, Dresden, Germany. Speech and Language Processing and Assistive Technologies (SLPAT).