Detecting cognitive impairments by agreeing on interpretations of linguistic features

Anonymous

Abstract

Linguistic features have shown promising applications for detecting various cognitive impairments. To improve detection accuracies, increasing the amount of data or linguistic features have been two applicable approaches. However, acquiring additional clinical data could be expensive, and hand-carving features are hard. In this paper, we take a third approach, putting forward the scheme "diagnosis after reaching consensus", where non-overlapping subsets (modalities) of linguistic features are compressed into low-dimension *interpretation* vectors by neural networks ("ePhysicians"). By encouraging interpretation vectors from multiple modalities to be indistinguishable, the "ePhysicians" extract important information for classification. We show that with the same subsets of features, our models outperform baseline neural network classifiers on clinical data. Using all of the 413 linguistic features, our best models have accuracies in detecting cognitive impairments comparable to the state-of-the-art models on several balanced datasets (.82 on DementiaBank in detecting Alzheimer's Disease (AD) and .66 in detecting Mild Cognitive Impairment (MCI)).

1 Introduction

Alzheimer's Disease (AD) and its usual precursor, mild cognitive impairment (MCI), are neurodegerative conditions that inhibit cognitive ability, including language ability. For example, cognitively impaired subjects use more pronouns instead of nouns, and pause more often between sentences in narrative speeches (Roark et al., 2011).

Pronoun-noun-ratios, pauses, and other linguistic features have been used to build classifiers to detect cognitive diseases in many tasks. For example, Fraser et al. (2015) had up to 82% accuracy on DementiaBank[1], and Weissenbacher et al. (2016) achieved up to 86% accuracy on a corpus of 500 subjects. Yancheva et al. (2015) predicted Mini-Mental State Estimation score (MMSE), a score to characterize the extent of cognitive impairment.

To improve the accuracy of automated assessment using engineered linguistic features, there are usually two approaches: incorporating more data or calculating more features. Taking the first approach, Noorian et al. (2017) incorporated normative data from Talk2Me[2] and Wisconsin Longitudinal Study (Herd et al., 2014), which increased AD:control accuracy up to 93%, and moderateAD:mildAD:control three-way classification accuracy to 70% on DementiaBank. Taking the second approach, Yancheva and Rudzicz (2016) reached a .80 F-score using 12 features derived from vector space models. Santos et al. (2017) calculated features depicting characteristics of co-occurrence graphs of narrative transcripts (e.g: degree of each vertex in the graph). Their classifiers reached 65% accuracy on DementiaBank (MCI versus a subset of Control).

There are limitations in either of the two approaches. On one hand, additional clinical data from the same origin could be expensive to acquire (Berndt and Cockburn, 2013). Training data from different sources (e.g. those in Noorian et al. (2017)) should be similar enough to the existing training data, so as to enhance classifier accuracies. Acquiring additional data from either of the two origins is hard. On the other hand, carving new features require creativity and collaboration

[1] https://talkbank.org/DementiaBank
[2] https://www.cs.toronto.edu/talk2me/

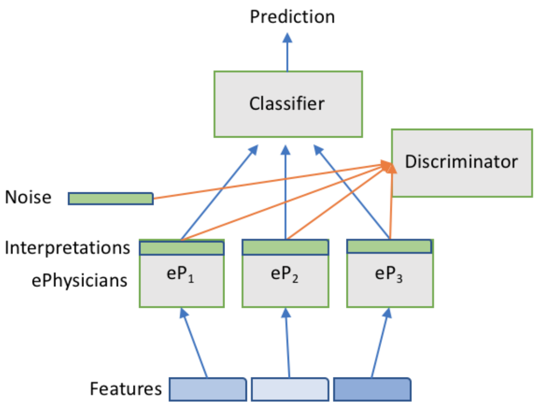


Figure 1: Overview of model structure when features (blue rectangles) are divided into three modalities (non-overlapping subsets). Each subset of features are passed into an "ePhysician" neural network whose outputs (green rectangles) are the *interpretation* vectors. The interpretation vectors are passed (one by one) into a "Discriminator" neural network and (after combined) into a "Classifier" network, respectively.

with subject matter experts. Besides, implementation and testing are time consuming.

These limitations motivate us to take a third approach. Instead of using new data or computing new features, we want to utilize precomputed features on existing dataset. Narrative description datasets contain multiple modalities, (audio and transcripts, to start with). Common information shared between multiple modalities have been applied to build good classifiers. Becker and Hinton (1992) predicted depths from multiple subsets of random-dot stereograms. de Sa (1994) divided linguistic features into two modalities, which are passed to two neural networks separately. The two neural networks supervised each other (i.e., output labels that are used to train the other) during alternative optimization steps to reach a consensus. Their self-supervised system reached 79$\pm$2% accuracy in Peterson-Barney vowel recognition dataset (Peterson and Barney, 1952). These examples illustrate the effectiveness of common information among different observations, but none of existing works apply adversarial networks to find these common information.

Goodfellow et al. (2014) proposed generative adversarial networks (GANs). In GANs, a "discriminator" network is trained to tell whether a vector is drawn from the real world or produced synthetically by a "generator" neural network, while the generator is trained to fool the discriminator. This setting have been used in multi-task classification from text (Liu et al., 2017), multilingual dialogue evaluation (Tong et al., 2018), audio voice conversion (Fang et al., 2018) and domain transfer (Taigman et al., 2017). However, to the knowledge of the authors, none of existing works apply GANs to discover knowledge shared among different aspects in data.

We propose a framework using adversarial training to utilize common information among modalities for classification. In this framework, several neural networks ("ePhysicians") are juxtaposed, each converting a partition of available linguistic features into a fixed-size vector ("interpretation") for each input document sample. Being trained towards producing indistinguishable interpretations, they should be increasingly able to capture common information contained across disparate subsets of linguistic features.

We show by experiments that neural network classifiers built and trained with the framework "reaching consensus among modalities" could outperform those without. Particularly, taking all 413 linguistic features we extract, our models have performances that align with the state-of-the-art results on balanced datasets (i.e., AD:Control, MCI:Control).

The novel contributions of this paper include:

- The "diagnosis by reaching consensus" scheme for neural network classifiers, where information shared between different modalities could be utilized.

- We improve on the methods to train the neural networks in iterative steps. Specifically, we train the ePhysicians to optimize both classification and discrimination loss, resulting in better performances of classifiers trained by the intuitive GAN approach (i.e. optimize only one type of network at a step).

- We show by experiment that an additional interpretation vector drawn from a Gaussian distribution (a.k.a, a *noise modality*) per data sample is beneficial to the classifier accuracy.

- We also visualize the interpretation vectors throughout several trials, and show the the interpretations have a trend towards symmetry in an *aggregate* manner.

2 Methods

2.1 Dataset

We use the DementiaBank dataset, which includes verbal descriptions (and associated transcripts) of the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Becker et al., 1994). The version we have access to contains 240 speech samples labeled Control (from 98 people), 234 with AD (from 148 people), and 43 with MCI (from 19 people). All participants have age greater than 44 years.

Note that the version of DementiaBank dataset we acquired contains different number of samples from what some previous works used. In Control:AD, Fraser et al. (2015) used 233 Control and 240 AD samples; Yancheva and Rudzicz (2016) had 241 Control and 255 AD samples; Hernández-Domínguez et al. (2018) had 242 Control and 257 AD samples (with 10% control samples excluded from the evaluation). In Control:MCI, Santos et al. (2017) used all 43 transcriptions from MCI and 43 sampled from Control group. With no clear mentions how the samples went, the constituents of Control group might differ from how we sample from the Control group. In this paper, we compare our model running on the same tasks (e.g: Control:AD) and compare to the best results reported in literature. The aforementioned slight difference in dataset should be noted.

2.2 Linguistic features

We pre-compute 413 linguistic features for each speech sample, and manually categorize them into four feature families as per below. These linguistic features are proposed by and identified as the most indicative of detecting cognitive impairments by various previous works including Roark et al. (2007); Chae and Nenkova (2009); Roark et al. (2011); Fraser et al. (2015); Hernández-Domínguez et al. (2018). After calculating these features, we use KNN imputation to replace the undefined values (resulting from divide-by-zero, for example), followed by a z-score normalization per feature.

Acoustic (185 features)

- Features related to speech fluency, including phonation rate, pause durations, and number and length of filled pauses (e.g., *'umm'*).
- Mean, variance, kurtosis, and skewness of the first 13 Mel-scaled cepstral coefficients, and their first- and second-order derivatives.

Syntactic and semantic (117 features)

- Average proportion of context-free grammar (CFG) phrase types[3], the rates of these phrase types[4], and the average phrase type length[5] (Chae and Nenkova, 2009)
- Average heights of the context-free grammar (CFG) parse trees, across all utterances in each transcript. Each tree comes from an utterance parsed by a context free grammar parser (`LexParser` implemented in Stanford CoreNLP (Manning et al., 2014))
- Number of occurrences of a set of 104 context-free production rules (e.g.,`S->VP`) in the CFG parse trees.
- Yngve scores statistics of CFG parse trees (Yngve, 1960; Roark et al., 2007). Yngve score is the degree of left-branching of each node in a parsed tree.

PoS-derived (80 features)

- The number of occurrences of part-of-speech (PoS) tags from Penn-treebank[6].
- The ratio of occurrences of several PoS tags, including noun-pronoun ratio.
- Number of occurrences of words in each of the five categories: subordinate (e.g: "because", "since", etc.), demonstratives (e.g: "this", "that"), function (e.g: words with PoS tag "CC", "DT", and "IN"), light verbs (e.g: "be", "have"), and inflected verbs (words with PoS tag "VBD", "VBG", "VBN", and "VBZ"), borrowing the categorization method in Kortmann and Szmrecsanyi (2004)

Lexical related (31 features)

- Lexical norms, including age of acquisition, familiarity, imageability, and frequency (Taler et al., 2009). They are averaged over the entire transcript and specific PoS categories, respectively.

[3] number of words in these types of phrases, divided by the total number of words in the transcript

[4] number of occurrences in a transcript, divided by the total number of words in the transcript

[5] number of words belonging to this phrase type in a transcript, divided by the occurrences of this phrase type in a transcript

[6] Using https://spacy.io

- Lexical richness, including moving-average type-token ratio over different window sizes (Covington and McFall, 2010), Brunet's index, and Honorés statistics (Guinn and Habash, 2012).
- Cosine similarity statistics (minimum, maximum, average, etc.) between pairs of utterances (represented as sparse vectors based on lemmatized words)
- Average word length, counts of total words, not-in-dictionary words, and fillers. The dictionary we use contains around 98,000 entries, including common words, plural forms of countable nouns, possessive forms of subjective nouns, different tenses of verbs, etc.

2.3 Model

Figure 1 is an example of our model structure (with M=3 modalities), and following is a formal formulation. First, each sample is converted into a vector $\mathbf{x}$ consisting of all available linguistic features. This vector is then divided into M partitions ('modalities') of approximately equal sizes [$\mathbf{x_1}$, $\mathbf{x_2}$, ..., $\mathbf{x_M}$], according to the families mentioned above. Unless specified otherwise, the modality assignments in our experiments are: (1) acoustic (185 features), (2) syntactic-semantic (117), and (3) pos-derived and lexical-related (111 here). In the rest of this paper, we will refer to them as Acoustic modality, SynSem modality, and LexPos modality[7]. These input vectors are then passed into respective ePhysician networks, each outputting an interpretation vector i_m consisting of distilled representation of a subject looking from a perspective (e.g: semantic-syntactic perspective). In other words, the m^{th} ePhysician can be written as a function, f_m:

$$\mathbf{i_m} = f_m(\mathbf{x_m})$$

To challenge how well the interpretations align, a discriminator network takes in the M interpretation vectors, and decides the likelihood from which ePhysician the interpretation vector comes:

$$\hat{m} = \text{softmax}(f_D(\mathbf{i_m}))$$

For each participant session, we add a "noise interpretation vector" $\mathbf{i_0}$ drawn from a normal distribution with the mean and variance identical to those of the interpretation vectors. To some extent, this noise works like a regularization mechanism to refrain the discriminator from making decisions based on superficial statistics. We will show in 3.1 that this addition empirically improves classifier performance.

$$\mathbf{i_0} \sim \mathcal{N}(\mu_{\mathbf{i_{1..M}}}, \sigma^2_{\mathbf{i_{1..M}}})$$

To produce classification, a classifier network f_C takes in the M interpretations, combines them, and outputs a prediction:

$$p_\mathbf{x} = \text{softmax}(f_C([\mathbf{i_1}, \mathbf{i_2}, ..., \mathbf{i_M}]))$$

where the subscript $\mathbf{x}$ is a reminder that the classification probability is that of the data sample $\mathbf{x}$. As a note of implementation, all ePhysicians, classifiers, and discriminator networks are fully connected networks with Leaky ReLU activations (Nair and Hinton, 2010) and batch normalization (Ioffe and Szegedy, 2015). The hidden layer sizes are all 10 for the ePhysician network, and there are no hidden layers for the discriminator and classifier networks. Although modalities might contain different number of input dimensions, we do not scale the ePhysician sizes. Such choice comes from the intuition that the ePhysicians should extract into the interpretation as similar amount of information as possible.

2.4 Optimization

The ePhysician, discriminator, and the classifier networks have different objectives and are optimized in alternative steps. We now explain the steps.

P and D steps e**P**hysicians and **D**iscriminators are optimized in an adversarial manner:

$$\max_{P_{1..M}} \min_{D} \mathcal{L_D}$$

where the discriminator loss $\mathcal{L_D}$ is the cross entropy loss of the modality discrimination output. In the case where we divide features into M modalities, there are $M+1$ samples for j to iterate through, for each data point.

Similar to GAN (Goodfellow et al., 2014), we set up P step as $\max_{P_{1..M}} \mathcal{L_D}$ and D step as $\min_{D} \mathcal{L_D}$.

C step optimizes the **C**lassifier network to minimize the cross entropy loss of classification error: $\min_{C} \mathcal{L_C}$, where

$$\mathcal{L_C} = \mathbb{E}_\mathbf{x} \left\{ -\log p_\mathbf{x} \right\}$$

[7] Note: this is different from the conventional definition of modality (acoustic, text, facial expressions, body movements, etc.). But our method could potentially apply to modalities defined otherwise.

CP step is a variant of the C step in which we also allow the gradients to back propagate to optimize the parameters of the ePhysicians: $\min_{C,P_{1..M}} \mathcal{L}_{\mathcal{C}}$. If CP is applied, the ePhysicians should both work towards producing indistinguishable interpretations, and producing interpretations suitable for classification. We will show empirically in 3.2 that CP step produces better results than the C step.

Implementation The objective functions $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{C}}$ are not convex. We use three Adam optimizers (Kingma and Ba, 2014), each corresponding to P, D, C(or CP) steps, and optimize iteratively for no more than 100 steps. The optimization stops prior to step 100 if the classification loss $\mathcal{L}_{\mathcal{C}}$ converges (i.e., does not differ from the previous iteration by more than 1×10^{-4}).

3 Experiments

To evaluate whether our intuitions result in useful models, we analyze the importances of various components in the model. First, the effectiveness of the noise modality is tested. Second, models optimized with C and CP steps are compared. Then, we compare our model with neural network classifiers using the same subsets of data, to show the importance of reaching a consensus. After that, we evaluate our model against several supervised learning benchmarks and on representative cognitive impairment detection tasks. To understand the model further, we also visualize the principal components of the interpretation vectors throughout several runs.

3.1 Noise modality improves performance

We compare the classifier with one without the additional noise modality (while other details including hidden dimensions and initial learning rates are kept unchanged).

Table 1 shows that in the AD:MCI classification task, the model with additional noise modality is better than the one without ($p = 0.04$ on 2-tailed T test with 18 DoF). Here is a possible explanation. Without the noise modality, a very simple strategy for the discriminator is to tell apart the interpretations by superficial aspects, namely their means and variances, instead of their distributions. The discriminator taking this strategy fails to capture the detailed aspects that makes the modalities different. Adding in the noise modality penalizes this strategy, and trains better discriminators by forcing them to *study the details*.

In following experiments, all models contain the additional noise modality.

Model	F1 micro	F1 macro
Gaussian noise	$.7995 \pm .0450$	$.7998 \pm .0449$
Without noise	$.7572 \pm .0461$	$.7577 \pm .0456$

Table 1: Comparison of models with and without interpretations in noise modality. The models containing a Gaussian noise modality outperform those without.

3.2 CP step is better than C step

We compare the classifier trained with CP step ($\min_{C,P_{1..M}} \mathcal{L}_{\mathcal{C}}$) to the one with C step ($\min_{C} \mathcal{L}_{\mathcal{C}}$). As shown in Table 2, the optimization using CP step produces higher-score classifiers than that using C step ($p < 0.001$ on 2-tailed T test with 18 DoF). Using CP step, the ePhysicians are optimized towards producing interpretations that are both indistinguishable (by the discriminator) and beneficial (for the classifier). Although the interpretations might agree less to each other, they could contain more *complementary* information, leading to better overall classifier performances.

In other experiments, all of our models use CP steps.

Optimization	F1 micro	F1 macro
P-D-C	$.6696 \pm .0511$	$.6743 \pm .0493$
P-D-CP	$.7995 \pm .0450$	$.7998 \pm .0449$

Table 2: Comparison of models using C and CP steps. The models optimized with sequences containing CP steps outperforms those with only C steps.

3.3 Agreement among modalities is desirable

The reason for our model working might be attributed to the expressiveness of the extracted features themselves. To evaluate the effectiveness of the setting "letting multiple modalities agree", we compare our model with neural network classifiers taking only partial input features. The networks are all just multiple layer perceptrons containing the same total number of neurons as the 'classifier pipeline' of our models (a.k.a ePhysicians and the classifier)[8] with batch normalization between hidden layers. A few observations could be made from Table 3:

1. Some features from particular modalities are better than others. For example, acoustic features could be used for building better classifiers than those in the lexical-pos ($p = .005$ for 2-tailed T test with 18 DoF) or syntactic-semantic modality ($p < .001$ for 2-tailed T test with 18 DoF)
2. Combining features from different modalities usually result in better MLP classifiers. Syntactic-semantic features plus lexical and pos features is an exception. This might be because the large number of less expressive features in syntactic-semantic modality confuses the classifier.
3. Given the same number of features, training the networks to agree in interpretations between modalities improve the accuracy.

Models (Modality)	Accuracy
MLP (Acoustic)	.7519 ± .0245
MLP (SynSem)	.5222 ± .0180
MLP (LexPoS)	.6987 ± .0278
MLP (SynSem + LexPos)	.5819 ± .0216
Ours (SynSem + LexPos)	.7257 ± .0344
MLP (Acoustic + LexPos)	.7002 ± .1128
Ours (Acoustic + LexPos)	.7542 ± .0433
MLP (Acoustic + SynSem)	.6776 ± .0952
Ours (Acoustic + SynSem)	.7574 ± .0361
MLP (All 3 modalities)	.7528 ± .0520
Ours (All 3 modalities)	**.7995 ± .0450**

Table 3: Performance comparison between our model and neural network classifiers having partial modality information. Here `SynSem` is shorthand notation for Syntactic and Semantic related features, and `LexPos` for lexical related features.

3.4 Evaluation against benchmark algorithms

State-of-the-art papers use traditional classifiers with their features. To compare with theirs, we run traditional classifiers on our features and compare the performances. Several traditional supervised learning benchmark algorithms are tested in this paper: support vector machine (SVM), quadratic discriminant analysis (QDA), random forest (RF), and Gaussian process (GP). For completeness, multiple layer perceptrons (MLP) containing all features as inputs are also mentioned in Table 5. On the binary classification task, our model does better than them all.

3.5 Comparison to accuracies in literature

To illustrate the utility of our method against tasks other than AD:CTL, we train and run the "diagnosis after reaching consensus" model on major tasks in diagnosing cognitive diseases on DementiaBank. The best results (5-fold cross validation) are shown in Table 4. On binary AD:CTL and MCI:CTL (sampled a subset to make the dataset balanced, as in Santos et al. (2017), our best results are comparable to the best results reported in the literature on balanced datasets. However, on the ternary AD:MCI:CTL classification task, our model has limited performance. This is a limitation of the "diagnosis by reaching consensus" framework.

3.6 Visualizing the interpretations

To further understand what happens inside the models during training, we visualize the interpretation vectors with PCA. Figures 2, 3, 4 and 5 are drawn from four arbitrary runs of the model. Each interpretation is represented with a data point on the figure, with its color representing the modality it comes from (including the noise modality).

Several common themes could be observed:

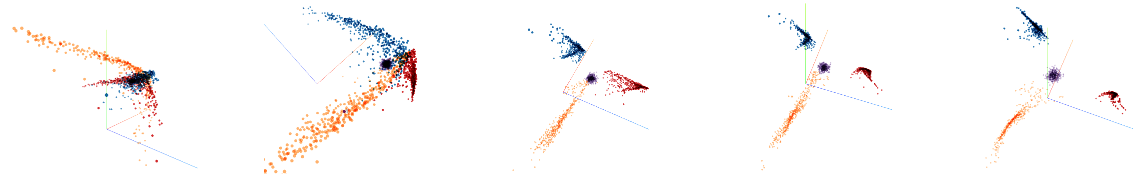
1. *Symmetric clustering*. Initially the configurations of interpretations are largely dependent on the initialization of network. Gradually the interpretations of the same modality tend to form clusters. Optimizing the ePhysicians towards both targets make they compress modalities into interpretation vectors which are symmetrical in an *aggregate* manner.

2. *The noise modality* lies at the center of the three petals. Its shape do not resemble any of the other three modalities. This indicates the distribution of interpretation vectors do not obey simple Gaussian distribution, which illustrates the importance of CP step (encour-

[8] For example, for models taking in two modalities, if our model contain ePhysicians with one layer of 20 hidden neurons, the interpretation vector dimension 10, and classifier 5 neurons, then the benchmarking neural network contains three hidden layers with [20×2, 10×2, 5] neurons.

Task	Statistics	Our model	Best in literature
AD vs Control	Accuracy, 10 folds CV	**.82**	**.82** (Fraser et al., 2015)
MCI vs. subset of Control	Accuracy, 5 folds CV	**.66**	.65 (Santos et al., 2017)
AD vs. MCI vs. Control	F micro / macro, 10 folds CV	.70/.73	**.78 / .82** (Hernández-Domínguez et al., 2018)

Table 4: Evaluation of top performance of our model on multiple tasks. The higher evaluations are marked bold. Fraser et al. (2015) used linear regressor on 50 carefully selected features. Santos et al. (2017) used SVM and ensembled traditional classifiers. Hernández-Domínguez et al. (2018) used SVM and Random Forest traditional classifiers when getting these results. In Table 5 we will compare our model to traditional classifiers on the dataset available to us.



(a) Step 5
$\mathcal{L}_\mathcal{D} = 1.30$
Val accr .52%
Variance 66.3%

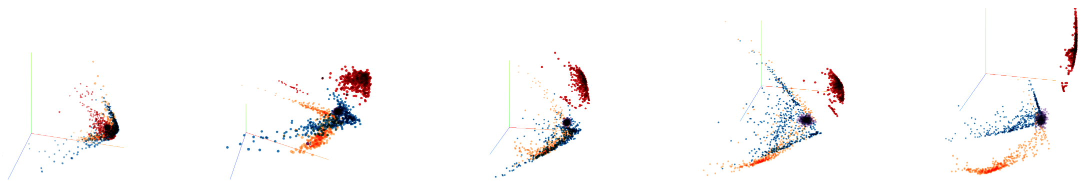
(b) Step 10
$\mathcal{L}_\mathcal{D} = 1.23$
Val accr .59%
Variance 70.2%

(c) Step 20
$\mathcal{L}_\mathcal{D} = 1.01$
Val accr .72%
Variance 76.0%

(d) Step 30
$\mathcal{L}_\mathcal{D} = 0.66$
Val accr .77%
Variance 76.7%

(e) Step 40
$\mathcal{L}_\mathcal{D} = 0.31$
Val accr .78%
Variance 87.7%

Figure 2: PCA visualizations in steps 5, 10, 20, 30 and 40 of a trial. The three clusters representing three modalities spread out like petals, while the noise modality remain in the center. “Variance” refers to the variance explained by the first *three* principal components.



(a) Step 5
$\mathcal{L}_\mathcal{D} = 1.32$
Val accr .74%
Variance 67.4%

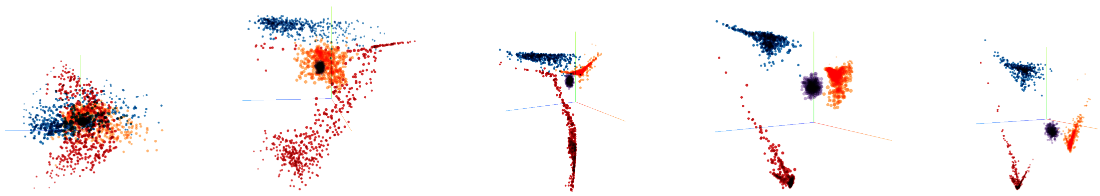
(b) Step 10
$\mathcal{L}_\mathcal{D} = 1.28$
Val accr .81%
Variance 68.0%

(c) Step 20
$\mathcal{L}_\mathcal{D} = 1.19$
Val accr .77%
Variance 70.5%

(d) Step 30
$\mathcal{L}_\mathcal{D} = 1.04$
Val accr .79%
Variance 76.9%

(e) Step 40
$\mathcal{L}_\mathcal{D} = 0.72$
Val accr .79%
Variance 76.9%

Figure 3: In this trial, the petals are wider than those of Figure 2.



(a) Step 5
$\mathcal{L}_\mathcal{D} = 1.34$
Val accr .74%
Variance 71.2%

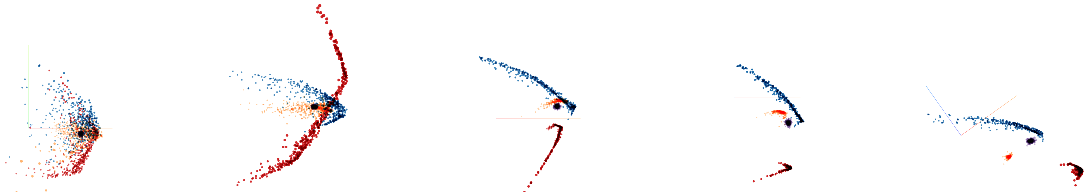
(b) Step 10
$\mathcal{L}_\mathcal{D} = 1.32$
Val accr .76%
Variance 72.9%

(c) Step 20
$\mathcal{L}_\mathcal{D} = 1.17$
Val accr .77%
Variance 77.1%

(d) Step 30
$\mathcal{L}_\mathcal{D} = 0.89$
Val accr .79%
Variance 77.0%

(e) Step 40
$\mathcal{L}_\mathcal{D} = 0.64$
Val accr .79%
Variance 80.8%

Figure 4: In this trial, both the blue and the orange cluster form wide petals. Interestingly, they gradually become tighter towards the noise modality, but still maintain clear gaps in between.



(a) Step 5
$\mathcal{L}_{\mathcal{D}} = 1.33$
Val accr $.65\%$
Variance 63.4%

(b) Step 10
$\mathcal{L}_{\mathcal{D}} = 1.29$
Val accr $.71\%$
Variance 69.2%

(c) Step 20
$\mathcal{L}_{\mathcal{D}} = 1.19$
Val accr $.72\%$
Variance 79.7%

(d) Step 30
$\mathcal{L}_{\mathcal{D}} = 0.98$
Val accr $.74\%$
Variance 81.6%

(e) Step 40
$\mathcal{L}_{\mathcal{D}} = 0.71$
Val accr $.78\%$
Variance 86.6%

Figure 5: Each petal here has the shape of a long hook from step 10 to 30, but gradually degenerates towards small points.

Classifier	Micro F1	Macro F1
SVM	.4810 ± .0383	.6488 ± .0329
QDA	.5243 ± .0886	.5147 ± .0904
RF	.6184 ± .0400	.6202 ± .0422
GP	.6775 ± .0892	.6873 ± .0819
MLP	.7528 ± .0520	.7561 ± .0444
Ours	**.7995 ± .0450**	**.7998 ± .0449**

Table 5: Comparison with different traditional classifiers in AD:Control classification task. Our model has higher accuracy than the best traditional classifier, MLP ($p = 0.046$ on 20DoF one-tailed t tests).

aging the discriminator to study the distributions of interpretations).

3. *The variances* explained by the first a few principal components usually increase as the optimizations proceed. This might indicate that by encouraging the interpretations to reach an agreement, *the consensus tend to be simple*.

4. *Accuracy* in validation set generally increases as the training proceeds, and as the interpretations demonstrate a clearer separation from each other visually. In other words, the interpretations do not need to be perfectly aligned (which should correspond to overlapping, indistinguishable dots from PCA visualizations). As long as they are working towards forming an indistinguishable interpretation, the classifier accuracy can be boosted.

4 Conclusion and future works

We have put forward the "diagnosis after reaching consensus" scheme, in which neural networks are encouraged to compress various modalities into indistinguishable fixed-size *interpretation* vectors. We show this "agreement between modalities" mechanism, with the additional noise modality, improves performances of neural network classifiers to be higher than MLP baselines given the same features. With all 413 linguistic features, we show our best performing models have comparable results as state-of-the-art ones on balanced classification tasks.

In the future, the "agreement among modalities" idea could be applied to design objective functions for training classifiers in various tasks. It would also be meaningful to test models on other datasets than DementiaBank. In addition, the mechanisms making the clusters of interpretation vectors symmetric could be analyzed.

References

James T Becker, François Boiler, Oscar L 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.

Suzanna Becker and Geoffrey E Hinton. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* 355(6356):161.

Ernst R Berndt and Iain M Cockburn. 2013. Price indexes for clinical trial research: a feasibility study. Technical report, National Bureau of Economic Research.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 139–147.

Michael A Covington and Joe D McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). In *Journal of quantitative linguistics*. Taylor & Francis, volume 17, pages 94–100.

Virginia R de Sa. 1994. Learning classification with unlabeled data. In *Proceedings of Advances in neural information processing systems*. pages 112–119.

Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. 2018. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *Proceedings of International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*.

Kathleen C Fraser, Jed A Metlzer, and Frank Rudzicz. 2015. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease 49(2016)407-422* .

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of Advances in neural information processing systems*. pages 2672–2680.

Curry I Guinn and Anthony Habash. 2012. Language Analysis of Speakers with Dementia of the Alzheimer's Type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*. Menlo Park, CA, pages 8–13.

Pamela Herd, Deborah Carr, and Carol Roan. 2014. Wisconsin longitudinal study (WLS). In *International journal of epidemiology*. Oxford University Press, volume 43, pages 34–41.

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 and Dementia: Diagnosis, Assessment and Disease Monitoring* 10(3):260–268.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. pages 448–456.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Bernd Kortmann and Benedikt Szmrecsanyi. 2004. Global synopsis: morphological and syntactic variation in English. *A handbook of varieties of English* 2:1142–1202.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1–10.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. pages 807–814.

Zeinab Noorian, Chloé Pou-Prom, and Frank Rudzicz. 2017. On the importance of normative data in speech-based assessment. In *Proceedings of Machine Learning for Health Care Workshop (NIPS MLHC)*.

Gordon E Peterson and Harold L Barney. 1952. Control methods used in a study of the vowels. In *The Journal of the acoustical society of America*. ASA, volume 24, pages 175–184.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, pages 1–8.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. In *IEEE transactions on audio, speech, and language processing*. U.S. National Library of Medicine.

Leandro Santos, Edilson Anselmo Corrêa Júnior, Osvaldo Oliveira Jr, Diego Amancio, Letícia Mansur, and Sandra Aluísio. 2017. Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1284–1296.

Yaniv Taigman, Adam Polyak, and Lior Wolf. 2017. Unsupervised cross-domain image generation. In *Proceedings of International Conference on Learning and Representations (ICLR)*.

Vanessa Taler, Ekaterini Klepousniotou, and Natalie A Phillips. 2009. Comprehension of lexical ambiguity in healthy aging, mild cognitive impairment, and mild Alzheimer's disease. In *Neuropsychologia*. Elsevier, volume 47, pages 1332–1343.

X Tong, Z Fu, M Shang, D Zhao, and R Yan. 2018. One "Ruler" for All Languages: Multi-Lingual Dialogue Evaluation with Adversarial Multi-Task Learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

Davy Weissenbacher, Travis A Johnson, Laura Wojtulewicz, Amylou Dueck, Dona Locke, Richard Caselli, and Graciela Gonzalez. 2016. Automatic prediction of linguistic decline in writings of subjects with degenerative dementia. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1198–1207.

Maria Yancheva, Kathleen 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 (SLPAT 2015)*. Association for Computational Linguistics, Dresden, Germany, pages 134–139.

Maria Yancheva and Frank Rudzicz. 2016. 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)*. volume 1, pages 2337–2346.

Victor H Yngve. 1960. A model and an hypothesis for language structure. In *Proceedings of the American philosophical society*. JSTOR, volume 104, pages 444–466.