ISOLATING EFFECTS OF AGE WITH FAIR REPRESENTATION LEARNING WHEN ASSESSING DEMENTIA

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the most prevalent symptoms among the elderly population, dementia, can be detected by classifiers trained on linguistic features extracted from narrative transcripts. However, these linguistic features are impacted in a similar but different fashion by the normal aging process. Aging is therefore a confounding factor, whose effects have been hard for machine learning classifiers to isolate.

In this paper, we show that deep neural network (DNN) classifiers can infer ages from linguistic features, which is an entanglement that could lead to unfairness across age groups. We show this problem is caused by undesired activations of v-structures in causality diagrams, and it could be addressed with fair representation learning. We build neural network classifiers that learn low-dimensional representations reflecting the impacts of dementia yet discarding the effects of age. To evaluate these classifiers, we specify a model-agnostic score $\Delta_{eo}^{(N)}$ measuring how classifier results are disentangled from age. Our best models outperform baseline neural network classifiers in disentanglement, while compromising accuracy by as little as 2.56% and 2.25% on DementiaBank and the Famous People dataset respectively.

INTRODUCTION

One in three seniors die of Alzheimer's and other types of dementia in the United States (Association, 2018). Although its causes are not yet fully understood, dementia impacts people's cognitive abilities in a detectable manner. This includes different syntactic distributions in narrative descriptions (Roark et al., 2007), more pausing (Singh et al., 2001), higher levels of difficulty in recalling stories (Lunsford & Heeman, 2015), and impaired memory generally (Lehr et al., 2012). Fortunately, linguistic features can be used to train classifiers to detect various cognitive impairments. For example, Fraser et al. (2013) detected primary progressive aphasia with up to 100% accuracy, and classified subtypes of primary progressive aphasia with up to 79% accuracy on a set of 40 participants using lexical-syntactic and acoustic features. Fraser et al. (2015) classified dementia from control participants with 82% accuracy on narrative speech.

However, dementia is not the only factor causing such detectable changes in linguistic features of speech. Aging also impairs cognitive abilities (Harada et al., 2013), but in subtly different ways from dementia. For example, aging inhibits fluid cognitive abilities (e.g., cognitive processing speed) much more than the consolidated abilities (e.g., those related to cumulative skills and memories) (Deary et al., 2009). In other words, the detected changes of linguistic features, including more pauses and decreased short-term memories, could attribute to just normal aging process instead of dementia. Unfortunately, due to the high correlation between dementia and aging, it can be difficult to disentangle symptoms are caused by dementia or aging (Murman, 2015). Age is therefore a confounding factor in detecting dementia.

The effects of confounding factors are hard for traditional machine learning algorithms to isolate, and this is largely due to sampling biases in the data. For example, some algorithms predict higher risk of criminal recidivism for people with darker skin colors (Julia et al., 2016), others identify images of smiling Asians as blinking (Lee, 2009), and GloVe word embeddings can project European-American names significantly closer to the words like 'pleasant' than African-American names (Caliskan et al., 2017). It is preferable for classifiers to make decisions without biasing too heavily on demographic factors, and therefore to isolate the effects of confounding factors. However, as we will show in Experiments, traditional neural network classifiers bias on age to infer dementia; this can lead to otherwise avoidable false positives and false negatives that are especially important to avoid in the medical domain. Graphically, if both age A and dementia D cause changes in a feature X, the result is a *v-structure* (Koller & Friedman, 2009)

$$A \rightarrow X \leftarrow D$$

which is activated upon observing X. In other words, the confounder A affects $P(D|\mathbf{X})$ if we train the classifier in traditional ways, which is to collect data points $\{(\mathbf{X}, D)^{(i)}\}$ and to learn an inference model $P(\hat{D}|\mathbf{X})$ approximating the affected $P(D|\mathbf{X})$.

Traditionally, there are several ways to eliminate the effects of confounding factors A.

Controlling A gives a posterior distribution $P(D|\mathbf{X}, A)P(A)$. This is unfortunately unrealistic for small, imbalanced clinical datasets, in which sparsity may require stratification. However, the stratified distributions $P(D|\mathbf{X}, A)$ can be far from a meaningful representation of the real world (as we will show, e.g., in Figure 2). Moreover, a discrepancy in the sizes of age groups can skew the age prior $P(A)$, which would seriously inhibit the generalizability of a classifier.

Controlling X Conducting a randomized control trial (RCT) on X removes all causal paths leading "towards" the variable X, which gives a de-confounded dataset $P(D|do(\mathbf{X}))$ according to the notation in Pearl (2009). However, RCTs on $\mathbf{X}$ are even less practical because simultaneously controlling multiple features produces exponential number of scenarios, and doing this to more than 400 features require far more data points than any available dataset.

Pre-adjusting X according to a pre-trained model $X = f(A)$ per feature could also approximately generate the dataset $P(D|do(\mathbf{X}))$. However, such a model should consider participant differences, otherwise interpolating using a fixed age A would give exactly the same features for everybody. The participant differences, however, are best characterized via X, which are the values you want to predict.

To overcome the various problems with these methods, **we let our classifiers be aware of cognitive impairments while actively filtering out any information related to aging**. This is a fair representation learning framework that protects age as a “sensitive attribute”.

Fair representation learning frameworks can be used to train classifiers to equally consider the subjects with different sensitive attributes. A sensitive attribute (or “protected attribute”) can be race, age, or other variables whose impact should be ignored. In the framework proposed by Zemel et al. (2013), classifiers were penalized for the differences in classification probabilities among different demographic groups. After training, the classifiers produced better demographic similarities while compromising only a little overall accuracy. To push the fair representation learning idea further, adversarial training can be incorporated. Goodfellow et al. (2014) introduced generative adversarial networks, in which a generator and a discriminator are iteratively optimized against each other. Incorporating adversarial training, Madras et al. (2018) proposed a framework to learn a latent representation of data in order to limit its adversary’s ability to classify based on the sensitive attributes.

However, these approaches to fair representation learning only handle binary attributes. E.g., Madras et al. (2018) binarized age. To apply to cognitive impairments detection, we want to represent age on a continuous scale (with some granularity if necessary). We formulate a fairness metric for evaluating the ability of a classifier to isolate a continuous-valued attribute. We also propose four models that compress high-dimensional feature vectors into low-dimensional representations which encrypt age from an adversary. We show empirically that our models achieve better fairness metrics than baseline deep neural network classifiers, while compromising accuracies by as little as 2.56% and 2.25% on our two empirical datasets, respectively.

MEASURING DISENTANGLEMENT

There are many measures of entanglement between classifier outcomes and specific variables. We briefly review some relevant metrics, and then propose ours.

TRADITIONAL METRICS

Correlation (Pearson, Spearman, etc.) is often used to compare classification outputs with component input features. To the extent that these variables are stochastic, several information theoretic measures could be applied, including Kullback-Leibler divergence and Jensen-Shannon divergence. These can be useful to depict characteristics of two distributions when no further information about available data is given.

Mutual information can depict the extent of entanglement of two random variables. If we treat age (A) and dementia (D) as two random variables, then adopting the approach of Kwak & Choi (2002) gives an estimation of $I(A, D)$. However, given the size of clinical datasets, it can be challenging to give precise estimations.

An alternative approach is to assume that these variables fit into some probabilistic models. For example, we might assume the age variable A, dementia indicator variable D, and multi-dimensional linguistic feature $\mathbf{X}$ fit into some *a priori* model (e.g., the v-structure mentioned above, $A \rightarrow \mathbf{X} \leftarrow D$), then the mutual information between A and D is:

$$I(A, D) = \mathbb{E}_{p(A,D)} \log \frac{p(A, D)}{p(A)p(D)} = \mathbb{H}_A + \mathbb{H}_D + \mathbb{E}_{p(A,D)} \left[\log p(A, D) \right]$$

where the entropy of age $\mathbb{H}_A$ and of cognitive impairment $\mathbb{H}_D$ remain constant with respect to the input data X, and $p(A, D) = \sum_{\mathbf{X}} p(A, \mathbf{X}, D) = \sum_{\mathbf{X}} p(A|\mathbf{X})p(D|\mathbf{X})p(\mathbf{X})$. However, this marginalized probability is difficult to approximate well, because (1) the accuracy of the term $p(A|\mathbf{X})$ relies on the ability of our model to infer age from features, and (2) it is hard to decide on a good prior distribution on linguistic features $p(\mathbf{X})$. We want to make the model agnostic to age, leading to a meaningless mutual information in the 'ideal' case.

In our frameworks, we do not assume specific graphical models that correlate confounds and outcomes, and we propose more explainable metrics than the traditional statistical ones.

FAIRNESS METRICS

The literature in fairness representation learning offers several metrics for evaluating the extent of bias in classifiers. Generally, the fairer the classifier is, the less entangled the results are with respect to some protected features.

Demographic parity Zemel et al. (2013) stated that the fairest scenario is reached when the composition of the classifier outcome for the protected group is equal to that of the whole population. While generally useful, this does not apply to our scenario, in which there really *are* more elderly people suffering from cognitive impairments than younger people (see Figure 2).

Cross-entropy loss Edwards & Storkey (2016) used the binary classification loss of an adversary that tried to predict sensitive data from latent representations, as a measure of fairness. This measure can only apply to those models containing an adversary component, not traditional classifiers. Moreover, this loss also depends on the ability of the adversary network. For example, a value of this loss could indicate confusing representations (so sensitive information are protected well), but it could also indicate a weak adversary.

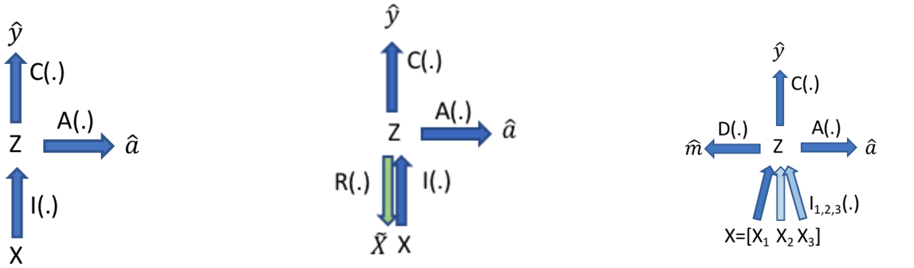
Equalized odds Hardt et al. (2016) proposed a method in which false positive rates should be equal across groups in the ideal case. Madras et al. (2018) defined fairness distance as the absolute difference in false positive rates between two groups, plus that of the false negative rates:

$$\Delta = \left|p_0 - p_1\right| + \left|n_0 - n_1\right|$$

where p_a and n_a correspond to the false positive rate and false negative rate, respectively, with sensitive attribute $a = 0$ ($a = 1$).

OUR METRIC

We propose an extension of the metric used by Madras et al. (2018) to continuous sensitive attributes, suitable for evaluating an arbitrary two-class classifier.



(a) age-indep-simple

(b) age-indep-autoencoder and age-indep-entropy

(c) age-indep-consensus-net

Figure 1: Model structures. Each colored arrow denotes a neural network. The common components are interpreters $I(.)$, adversary $A(.)$, and classifier $C(.)$. In age-indep-autoencoder and age-indep-entropy (Figure 1b), a reconstructor $R(.)$ tries to reconstruct input data from the hidden representation. In age-indep-consensus-nets (Figure 1c), a discriminator $D(.)$ tells apart from which modality the representation originates.

First, groups of age along a scale are divided so that each group has multiple participants with both positive and negative diagnoses, respectively. Let a be the age group each participant is in.

Then, we aim for the expected false positive (FP) rates of the classifier to be as constant as possible across age groups. This applies likewise to the false negative (FN) rates. To measure their variability, we use their sum of differences against the mean.

$$\Delta_{eo}^{(N_a)} = \sum_{a=1}^{N_a} \big|p_a - \hat{p}\big| + \sum_{a=1}^{N_a} \big|n_a - \hat{n}\big|,$$

where $\hat{x}$ represents the mean of variable x.

Analysis of metric

Special cases To illustrate the nature of our metric, we apply it to several special cases, i.e.:

1. When there is only one age group, our fairness metric has its best possible value: $\Delta_{eo} = 0$.
2. When there are only two age groups, our metric equals that of Madras et al. (2018).
3. In the extreme case where there are as many age groups as there are sample points (assuming there are no two people with identical ages but with different diagnoses), our metric becomes less informative, because the empirical expected false positive rates of that group is either 0 or 1. This is a limitation of our metric, and is the reason that we limit the number of age groups to accommodate the size of the training dataset.

Bounds Our metric is bounded. The lower bound, 0, is reached when all false positive rates are equal and when all false negative rates are equal across age groups. Letting N_a be the number of age groups divided, an upper bound for $\Delta_{eo}^{(N_a)}$ is N_a for any better-than-trivial binary classifier. The detailed proof is included in the Appendix.

Disentanglement Our fairness metric illustrates disentanglement. A higher $\Delta_{eo}^{(N)}$ corresponds to a higher variation of incorrect predictions by the classifier across different age groups. Therefore, a lower value of $\Delta_{eo}^{(N)}$ is desired for classifiers isolating the effects of age to a better extent. Throughout this paper, we use the terms 'fairness', 'disentanglement', and 'isolation' interchangeably.

Design choices We explain a few design choices here, namely linearity and indirect optimization.

Linearity. We encourage $\Delta_{eo}^{(N)}$ to be as linear as possible, for explainability of the fairness score itself. This eliminates possible scores consisting of higher order terms of FP / FN rates.

Indirect optimization. We avoid directly optimizing the fairness score $\Delta_{eo}^{(N)}$. The reasons are twofold. On one hand, although $\Delta_{eo}^{(N)}$ is correlated to the disentanglement between age and classification, it is based on FP / FN rates and hence bears their limitations – FP / FN rates do not capture all aspects of classifiers. Instead of making the representations beneficial for $\Delta_{eo}^{(N)}$, we encourage the hidden representations to be age-agnostic (we will explain how to set up age agnostic models in the following section). On the other hand, FP / FN rates are not differentiable after all.

MODELS

In this section, we describe four different ways of building representation learning models, which we call age-indep-simple, age-indep-autoencoder, age-indep-consensus-net, and age-indep-entropy.

AGE-INDEP-SIMPLE

The simplest model consists of an interpreter network $I(.)$ to compress high-dimensional input data, $\mathbf{x}$, to low-dimensional representations:

$$\mathbf{z} = I(\mathbf{x})$$

An adversary $A(.)$ tries to predict the *exact* age from the representation:

$$\hat{a} = A(\mathbf{z})$$

A classifier $C(.)$ estimated the probability of label (diagnosis) based on the representation:

$$P(\hat{y}) = \text{softmax}(C(\mathbf{z}))$$

For optimization, we set up two losses: the classification negative log likelihood loss $\mathcal{L}_c$ and the adversarial (L2) loss $\mathcal{L}_a$, where:

$$\mathcal{L}_c = \mathbb{E}_x \text{ -log} P(y) \qquad \mathcal{L}_a = \mathbb{E}_x ||\hat{a} - a||^2.$$

We want to train the adversary to minimize the L2 loss, to train the interpreter to maximize it, and to train the classifier (and interpreter) to minimize classification loss. Overall,

$$\min_{C,I} \mathcal{L}_c \text{ and}$$

$$\max_{I} \min_{A} \mathcal{L}_a.$$

Algorithm 1 Training age-indep-simple

```
Initialize $I$, $A$, $C$
for step := 1 to N do    ▷ N is a hyper-param
    for minibatch $\mathbf{x}$ in training data $\mathcal{X}$ do
        $\mathbf{z} = I(\mathbf{x})$, $a = A(\mathbf{z})$, $c = C(\mathbf{z})$
        Calculate $\mathcal{L}_a$, $\mathcal{L}_c$
        $\min_{I,C} \mathcal{L}_c - \mathcal{L}_a$    ▷ backprop gradients
        for k:=1 to K do
            $\min_{A} \mathcal{L}_a$    ▷ backprop gradients
```

The training steps are taken iteratively, as in previous work (Goodfellow et al., 2014).

AGE-INDEP-AUTOENCODER

The age-indep-autoencoder structure is similar to Madras et al. (2018), and can be seen as an extension from the age-indep-simple structure. Similar to age-indep-simple, there is an interpreter $I(.)$, an adversary $A(.)$, and a classifier $C(.)$ network. The difference is that there is a reconstructor network $R(.)$ that attempts to recover input data from hidden representations:

$$\hat{\mathbf{x}} = R(\mathbf{z})$$

The loss functions are set up as:

$$\mathcal{L}_c = \mathbb{E}_x \text{ -log} P(y) \qquad \mathcal{L}_a = \mathbb{E}_x ||\hat{a} - a||^2 \qquad \mathcal{L}_r = \mathbb{E}_x ||\hat{\mathbf{x}} - \mathbf{x}||^2$$

Overall, we want to train both the interpreter and the reconstructor to minimize the reconstruction loss term, in additional to all targets mentioned in the age-indep-simple network.

$$\min_{C,I,R} \mathcal{L} \text{ and } \max_{I} \min_{A} \mathcal{L}_a \qquad \text{where } \mathcal{L} = \mathcal{L}_c + \mathcal{L}_r$$

The detailed algorithm is similar to Algorithm 1 and is in the Appendix.

AGE-INDEP-CONSENSUS-NET

This is another extension from the age-indep-simple structure, borrowing an idea from consensus networks (Zhu et al., 2018a), i.e., that agreements between multiple modalities can result in representations that are beneficial for classification. By examining the performance of age-indep-consensus-net, we would like to see whether agreement between multiple modalities of data can be trained to be disentangled from age.

Similar to age-indep-simple structures, there is also an adversary $A(.)$ and a classifier $C(.)$. The interpreter, however, is replaced with several interpreters $I_{1..M}$, each compressing a subset of the input data ("modality") into a low-dimensional representation. The key of age-indep-consensus-network models is that these representations are encouraged to be indistinguishable. For simplicity, we randomly divide the input features into three modalities ($M = 3$) with equal (± 1) features. A discriminator $D(.)$ tries to identify the modality from which the representation comes:

$$\hat{m} = D(\mathbf{z})$$

The loss functions are set up as:

$$\mathcal{L}_c = \mathbb{E}_x\text{-log}P(y) \qquad \mathcal{L}_a = \mathbb{E}_x||\hat{a} - a||^2 \qquad \mathcal{L}_d = \mathbb{E}_x\text{-log}P(\hat{m})$$

Overall, we want to iteratively optimize the networks:

$$\min_{C,I} \mathcal{L}_c \text{ and } \max_I \min_A \mathcal{L}_a \text{ and } \max_I \min_D \mathcal{L}_d$$

The detailed algorithm is in the Appendix. Note that we do not combine the consensus network with the reconstructor because they do not work well with each other empirically. In one of the experiments by Zhu et al. (2018b), each interpreter $I_m(.)$ is paired with a reconstructor $R_m(.)$ and the performance decreases dramatically. The reconstructor encourages hidden representations to retain the fidelity of data, while the consensus networks urges hidden representations to keep only the information common among modalities, which prohibits the reconstructor and consensus mechanism to function together.

AGE-INDEP-ENTROPY

The fourth model we apply to fair representation learning is motivated by categorical GANs (Springenberg, 2016), where information theoretic metrics characterizing the confidences of predictions can be optimized. This motivates an additional loss function term; i.e., we want to encourage the interpreter to increase the uncertainty (i.e., to minimize the entropy) while letting the adversary become more confident in predicting ages from representations.

Age-indep-entropy models have the same network structures as age-indep-autoencoder, except that instead of predicting the exact age, the adversary network outputs the probability of the sample age being larger than the mean:

$$P(a|I, A, \mathbf{x}) = \text{softmax}(A(\mathbf{z}))$$

This enables us to define the empirical entropy $\mathbb{H}[p] = \mathbb{E}_x p\text{log}\frac{1}{p}$, which describes the uncertainty of predicting age.

Formally, the loss functions are set up as follows:

$$\begin{aligned}\mathcal{L}_c &= \mathbb{E}_x\text{-log}P(y)\\ \mathcal{L}_r &= \mathbb{E}_x||\hat{\mathbf{x}} - \mathbf{x}||^2\\ \mathcal{L}_a &= \mathbb{E}_x[\text{-log}P(a|I, A, \mathbf{x})] + \lambda_H \mathbb{H}[P(\hat{a}|I, A)]\end{aligned}$$

where λ_H is a hyper-parameter. For comparison, we also include two variants, namely the age-indep-entropy (binary) and age-indep-entropy (Honly) variants, each keeping only one of the two terms in $\mathcal{L}_a$. In our experiments, we show that these two terms in $\mathcal{L}_a$ are better applied together.

Overall, the training procedure is the same as age-indep-autoencoder and algorithm pseudocode is in the Appendix:

$$\min_{C,I,R} \mathcal{L}, \text{ and} \max_I \min_A \mathcal{L}_a, \text{ where } \mathcal{L} = \mathcal{L}_c + \mathcal{L}_r$$

Implementation

All models are implemented in PyTorch (Paszke et al., 2017), optimized with Adam (Kingma & Ba, 2014) with initial learning rate of 3×10^{-4}, and L2 weight decay 10. For simplicity, we use fully connected networks with ReLU activations (Nair & Hinton, 2010) and batch normalization (Ioffe & Szegedy, 2015) before output layers, for all interpreter, adversary, classifier, and discriminator networks. Our frameworks can be applied to other types of networks in the future.

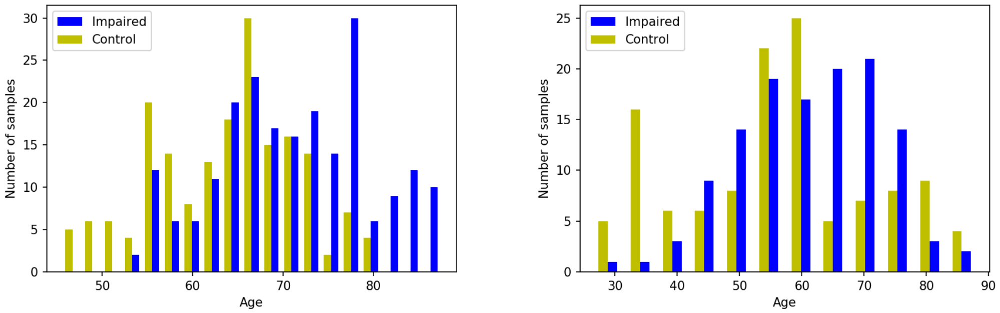
Experiments

Datasets

DementiaBank DementiaBank[1] is the largest available public dataset for assessing cognitive impairments using speech, containing 473 narrative picture descriptions from subjects aged between 45 and 90 (Becker et al., 1994). In each sample, a participant talks about what is happening in a clinically validated picture. There is no time limit in each session, but the average description lasts about a minute. 79 samples are excluded due to missing age information. In the remaining data samples, 182 are labeled 'control', and 213 are labeled 'dementia'. All participants have mini-mental state estimation (MMSE) scores (Folstein et al., 1975) between 1 and 30 [2]. Of all data samples containing age information, the mean is 68.26 and standard deviation is 9.00.

Famous People The Famous People dataset (Balagopalan et al., 2018) contains 252 transcripts from 17 people (8 with dementia including Gene Wilder, Ronald Reagan and Glen Campbell, and 9 healthy controls including Michael Bloomberg, Woody Allen, and Tara VanDerveer), collected and transcribed from publicly available speech data (e.g., press conferences, interviews, debatse, talk shows). Seven data samples are discarded due to missing age information. Among the remaining samples, there are 121 labeled as control and 124 as impaired. Note that the data samples were gathered across a wide range of ages (mean 59.25, standard deviation 13.60). For those people diagnosed with dementia, there are data samples gathered both before and after the diagnosis, and all of which are labeled as 'dementia'. The Famous People dataset permits for early detection several years before diagnosis, which is a more challenging classification task than DementiaBank.

Older participants in both DementiaBank (Figure 2a) and the Famous People dataset (Figure 2b) are more likely to have cognitive impairments.



(a) Histogram plot for DementiaBank

(b) Histogram plot for Famous People Dataset

Figure 2: Expository histogram plots for the ages of people in the impaired and control groups.

[1] https://dementia.talkbank.org/

[2] A higher MMSE score corresponds to a healthier estimated cognitive ability – scores 24 to 30 typically indicate a healthy state, 18-23 usually indicate mild cognitive impairment (MCI), and scores below 17 indicate dementia (or other type of cognitive impairment). To formulate a binary classification task, we label all of MCI and dementia as 'dementia'.

Classifier	DementiaBank			Famous People		
	Accuracy	$\Delta_{eo}^{(2)}$	$\Delta_{eo}^{(5)}$	Accuracy	$\Delta_{eo}^{(2)}$	$\Delta_{eo}^{(5)}$
Using raw features						
DNN	**.78±.05**	0.13±0.12	0.94±0.23	**.59±.05**	0.30±0.19	1.56±0.60
SVM	.77±.05	0.17±0.13	0.93±0.29	**.60±.04**	0.23±0.19	1.28±0.29
Random Forest	.74±.03	0.19±0.14	1.07±0.36	.56±.06	0.33±0.26	1.35±0.42
Adaboost	**.78±.07**	0.14±0.11	0.96±0.22	.54±.04	0.23±0.14	1.36±0.57

Table 1: Accuracy and fairness ($\Delta_{eo}^{(2)}$ and $\Delta_{eo}^{(5)}$) of several traditional classifiers. DNN is the baseline used to benchmark our neural network based representation learning models.

Preprocess and feature extraction

We extract 413 linguistic features from the narrative descriptions and their transcripts. These features were previously identified as the most useful for this task (Roark et al., 2007; Fraser et al., 2015; Lunsford & Heeman, 2015; Hernández-Domínguez et al., 2018). Each feature is z-score normalized. Relevant features include:

Acoustic: mean, variance, skewness, and kurtosis of the first 42 cepstral coefficients.

Speech fluency: pause-word ratio, utterance length, number and lengths of filled/unfilled pauses.

Lexical: cosine similarity between pairs of utterances, word lengths, lexical richness (moving-average type-token ratio, Brunet's index, and Honoré's statistics (Guinn & Habash, 2012)).

PoS: Number of occurrences of part-of-speech tags, tagged by SpaCy[3].

Syntactic and semantic: occurrences of context-free grammar phrase types, parsed by Stanford CoreNLP (Manning et al., 2014), and Yngve depth statistics (Yngve, 1960).

Linguistic features can predict age

As part of expository data analysis, we show that these linguistic features contain information indicating age. Simple fully connected neural networks can predict age with mean absolute error of 15.5 ± 1.3 years (on DementiaBank[4]) and 14.3 ± 2.5 years (on the Famous People dataset[5]). This indicates that even simple neural networks are able to infer information about age from linguistic features. Neural classifiers can therefore also easily bias on age, given the utility of age in downstream tasks.

Evaluating classical classifiers and disentanglement methods

We first set up benchmarks for our classifiers. We evaluate several traditional classifiers with our fairness metrics ($\Delta_{eo}^{(2)}$ and $\Delta_{eo}^{(5)}$, corresponding to dividing ages into $N = 2$ and $N = 5$ groups respectively). The results[6] are listed in Table 1. A DNN is used as the baseline because (1) all our models are based on neural networks, and (2) DNN classifiers have had the best (or statistically indistinguishable from the best) accuracy on the DementiaBank and Famous People datasets.

Performance and discussion

We evaluate the performances of our four proposed neural networks against the DNN baseline. As an additional ablation study, two variants of age-indep-entropy are also evaluated. Table 2 shows classification accuracies and fairness metrics, and the DNN baseline for comparison. Several observations emerge, as discussed below.

[3] `http://spacy.io`

[4] Hidden layer sizes 64, 32, 8. 5-fold cross validation.

[5] Hidden layer sizes 32, 20, 2. 5-fold cross validation

[6] All accuracy and fairness results in this paper are based on 5-fold cross validations, where no speaker occurs both in train and test data.

Model	DementiaBank			Famous People		
	Accuracy	$\Delta_{eo}^{(2)}$	$\Delta_{eo}^{(5)}$	Accuracy	$\Delta_{eo}^{(2)}$	$\Delta_{eo}^{(5)}$
DNN baseline	.78±.05	0.13±0.12	0.94±0.23	.59±.05	0.30±0.19	1.56±0.60
*-simple	.75±.00	**0.08±0.01**	**0.80±0.08**	.57±.05	0.24±1.90	1.47±0.57
*-autoencoder	**.76±.01**	0.11±0.00	0.88±0.24	.55±.07	**0.21±0.16**	**1.28±0.31**
*-consensus-nets	.72±.00	0.11±0.01	0.83±0.24	**.58±.05**	0.25±0.16	1.43±0.41
*-entropy	.75±.00	0.15±0.01	0.88±0.24	**.58±.06**	0.23±0.16	1.35±0.44
*-entropy (binary)	.72±.00	0.12±0.01	1.10±0.37	.55±.07	0.26±1.53	1.41±0.40
*-entropy (Honly)	.74±.00	0.17±0.02	1.27±0.54	.53±.06	**0.20±0.16**	1.39±0.49

Table 2: Evaluation results of our representation learning models. The ”age-indep” prefix are replaced with ”*” in model names. age-indep-simple and age-indep-autoencoder have better disentanglement scores, while the rest two models could have better accuracy.

Accuracy The fair representation learning models compromise accuracy, in comparison to DNN baselines. This confirms that part of the classification power of DNNs come from biasing with regards to age. On DementiaBank, the age-indep-autoencoder reduces accuracy the least (only 2.56% in comparison to the DNN baseline). On the Famous People data, age-indep-consensus and age-indep-entropy models compromise accuracies by only 2.25% and 2.75% respectively, which are not statistically different from the DNN baseline[7].

Disentanglement In comparison to DNN baselines, our fair representation learning models improve disentanglement/fairness[8], the improvements are mostly significant when measured by the two-group scores $\Delta_{eo}^{(2)}$. Also, the five-group scores $\Delta_{eo}^{(5)}$ are less stable for both datasets, and the scores in the Famous People have higher variances than in DementiaBank. Following is an explanation. DementiaBank has ∼400 data samples. In 5-fold cross validation, each of the five age groups has only ∼16 samples during evaluation. Famous People data contains ∼250 samples, which increases the variance. When the number of groups, N of $\Delta_{eo}^{(N)}$, is kept small (e.g., ∼100 samples per label per group, as in DementiaBank $N = 2$), the fairness metrics are stable.

Side notes The model age-indep-entropy is best used with a loss function containing both the binary classification term and the uncertainty minimization term. As shown in Table 2, although having similar fairness metrics[9], the two variants with only one term could have lower accuracy than age-indep-entropy.

In general, age-indep-simple and age-indep-autoencoder achieve the best fairness metrics. Noticeably, the better of them surpass traditional classifiers in both $\Delta_{eo}^{(2)}$ and $\Delta_{eo}^{(5)}$.

CONCLUSION

Here, we identify the problem of entangling age in the detection of cognitive impairments. After explaining this problem with causality diagrams, we formulate it into a fair representation learning task, and propose a fairness score to measure the extent of disentanglement. We put forward four fair representation learning models that learn low-dimensional representations of data samples containing as little age information as possible. Our best model improves upon the DNN baseline in our fairness metrics, while compromising as little accuracy as 2.56% (on DementiaBank) and 2.25% (on the Famous People dataset).

[7] $p = 0.20, 0.16$ on 38-DoF one-tailed t-tests, respectively.

[8] On DementiaBank, $p = 0.01$ and 0.03 for age-indep-simple and age-indep-entropy on $\Delta_{eo}^{(2)}$ respectively; these are significant. $p = 0.08$ and 0.09 on age-indep-autoencoder and age-indep-consensus-net on $\Delta_{eo}^{(2)}$ respectively; these are marginally significant. However, these differences are not as significant on $\Delta_{eo}^{(5)}$ (0.05, 0.31, 0.44, and 0.16.). On Famous People data, the p values for our four models are $0.15, 0.05, 0.17, 0.10$ on $\Delta_{eo}^{(2)}$ and $0.32, 0.03, 0.20, 0.10$ on $\Delta_{eo}^{(5)}$. These are all 38-DoF one-tailed t-tests.

[9] On DementiaBank, $p = 0.19, 0.06$ for $\Delta_{eo}^{(2)}$ and $\Delta_{eo}^{(5)}$ of age-indep-Honly against age-indep-entropy, $p = 0.24, 0.22$ for age-indep-binary. On Famous People, $p = 0.24, 0.39$ for age-indep-Honly, and $p = 0.33, 0.32$ for age-indep-binary. None of them are significant on 38-DoF one-tailed t-tests.

REFERENCES

Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimer's & dementia*, 2018.

Aparna Balagopalan, Jekaterina Novikova, and Frank Rudzicz. Early prediction of Alzheimer's disease from spontaneous speech. *Submitted to AAAI*, 2018.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. URL `http://science.sciencemag.org/content/356/6334/183`.

Ian J Deary, Janie Corley, Alan J Gow, Sarah E Harris, Lorna M Houlihan, Riccardo E Marioni, Lars Penke, Snorri B Rafnsson, and John M Starr. Age-associated cognitive decline. *British medical bulletin*, 92(1):135–152, 2009.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *ICLR*, 2016.

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

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease 49(2016)407-422*, 2015.

Katie Fraser, Frank Rudzicz, and Elizabeth Rochon. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Proc. Interspeech*, pp. 2177–2181, Lyon France, aug 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

Curry I Guinn and Anthony Habash. Language analysis of speakers with dementia of the Alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pp. 8–13. Menlo Park, CA, 2012.

Caroline N Harada, Marissa C Natelson Love, and Kristen L Triebel. Normal cognitive aging. In *Clinics in geriatric medicine*, volume 29, pp. 737–752. Elsevier, 2013.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *NIPS*, pp. 3315–3323, 2016.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. Computer-based evaluation of AD and MCI patients during a picture description task. In *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*. Elsevier, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.

Angwin Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Nojun Kwak and Chong Ho Choi. Input feature selection by mutual information based on Parzen window. In *IEEE Trans. Patt. Anal. Mach. Intell.*, volume 24, pp. 1667–1671, 2002.

Odella Lee. Camera misses the mark on racial sensitivity. *Gizmodo*, 2009.

Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Proc. Interspeech*, pp. 1039–1042, 2012.

Rebecca Lunsford and Peter A Heeman. Using linguistic indicators of difficulty to identify mild cognitive impairment. In *Proc. Interspeech*, pp. 658–662, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, pp. 3381–3390, 2018.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Daniel L Murman. The impact of age on cognition. In *Seminars in hearing*, volume 36, pp. 111. Thieme Medical Publishers, 2015.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted Boltzmann machines. In *ICML*, pp. 807–814, 2010.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. Syntactic complexity measures for detecting mild cognitive impairment. In *Workshop on BioNLP 2007*, pp. 1–8. Association for Computational Linguistics, 2007.

Sameer Singh, Romola S. Bucks, and Joanne M. Cuerden. Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech. In *Aphasiology*, volume 15, pp. 571–583. Routledge, 2001.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *ICLR*, 2016.

Victor H Yngve. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466, 1960.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *ICML*, pp. 325–333, Atlanta, Georgia, USA, 2013.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. Detecting cognitive impairments by agreeing on interpretations on linguistic features. *arxiv 1808.06570*, 2018a.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. Semi-supervised classification by reaching consensus among modalities. *arxiv 1805.09366*, 2018b.

Appendix 1: Proof of Upper bound of $\Delta_{eo}^{(N_a)}$

In this section, we detail the steps leading to an upper bound for the metric $\Delta_{eo}^{(N_a)}$.

Proposition The expectation of all false positive and false negative rates are bounded by $[0, 1]$.

This gives an upper bound to our metric $\Delta_{eq}^{(N_a)} \leq 2N_a$. If the classifier is not trivial, there is a tighter upper bound.

Definition A *trivial binary classifier* always predicts the majority class.

Lemma The expected error rate of a trivial binary classifier is no more than 0.5.

Proof of Lemma Let λ ($0 \leq \lambda \leq 1$) denote the composite of positive samples in the dataset. Table 3 shows the possible values of error rates. Regardless of whether the dataset has balanced classes, the error rate of a trivial binary classifier is no more than 0.5.

	$\lambda < 0.5$	$\lambda \geq 0.5$
Trivial prediction t	0	1
False positive rate (FP)	0	$1 - \lambda$
False negative rate (FN)	λ	0
Error rate (FP+FN)	$\lambda < 0.5$	$1 - \lambda \leq 0.5$

Table 3: Table of values showing statistics of a trivial binary classifier.

Theorem Our score $\Delta_{eo}^{N_a}$ is upper bounded by N_a for any non-trivial binary classifier:

$$\sum_{a=1}^{N_a} \{|p_a - \hat{p}| + |n_a - \hat{n}|\} \leq N_a$$

Proof of Theorem For each of the age groups:

$$\begin{aligned} &|p_a - \bar{p}| + |n_a - \bar{n}| \\ &\leq \max\{|p_a - 0| + |n_a - 0|, |p_a - 0.5| + |n_a - 0.5|\} \\ &\leq \max\{0.5, 1\} = 1 \end{aligned}$$

Summing up the N_a age groups results in our upper bound N_a for non-trivial classifiers.

Appendix 2: Algorithms for our models

Following are the pseudo-code algorithms for our remaining three models; age-indep-AutoEncoder, age-indep-ConsensusNetworks, and age-indep-Entropy.

Algorithm 2 Training age-indep-AutoEncoder

```
Initialize $I$, $A$, $C$, $R$
for step := 1 to N do                                          ▷ N is a hyper-parameter
    for minibatch $\mathbf{x}$ in training data $\mathcal{X}$ do
        $\mathbf{z} = I(\mathbf{x})$, $a = A(\mathbf{z})$, $c = C(\mathbf{z})$
        $\tilde{\mathbf{x}} = R(\mathbf{z})$                    ▷ Reconstructing the original feature vector.
        Calculate $\mathcal{L}_a$, $\mathcal{L}_c$, $\mathcal{L}_r$
        $\min_{I,C,R} \mathcal{L}_c - \mathcal{L}_a + \mathcal{L}_r$      ▷ backprop gradients
        for k:=1 to K do                                       ▷ K is a hyper-parameter
            $\min_{A} \mathcal{L}_a$                           ▷ backprop gradients
```

Algorithm 3 Training age-indep-consensus-net

```
Each data point are split into M modalities
Initialize $I_{1..M}$, $A$, $C$
for step := 1 to N do                                          ▷ N is a hyper-parameter
    for minibatch $\mathbf{x}$ in training data $\mathcal{X}$ do
        for m := 1 to M do
            $\mathbf{z_m} = I_m(\mathbf{x_m})$                 ▷ interpretation
            $\hat{m}_m = D(\mathbf{z_m})$                      ▷ predict modality
            $\hat{a}_m = A(\mathbf{z_m})$                      ▷ predict age group
        $P(\hat{y}) = \text{softmax}(C([\mathbf{z_1}, ..\mathbf{z_M}]))$
        Calculate $\mathcal{L}_a$, $\mathcal{L}_c$, $\mathcal{L}_d$
        $\min_{I,C} \mathcal{L}_c - \mathcal{L}_a - \mathcal{L}_d$           ▷ backprop gradients
        for k:=1 to $K_D$ do                                   ▷ $K_D$ is a hyper-parameter
            $\min_{D} \mathcal{L}_d$                           ▷ optimize modality discriminator
        for k:=1 to $K_A$ do                                   ▷ $K_A$ is a hyper-parameter
            $\min_{A} \mathcal{L}_a$                           ▷ optimize adversary
```

Algorithm 4 Training age-indep-Entropy

```
Initialize $I$, $A$, $C$, $R$
for step := 1 to N do                                          ▷ N is a hyper-parameter
    for minibatch $\mathbf{x}$ in training data $\mathcal{X}$ do
        $\mathbf{z} = I(\mathbf{x})$, $a = A(\mathbf{z})$, $c = C(\mathbf{z})$, $\mathbf{x} = R(\mathbf{z})$
        Calculate $\mathcal{L}_c$, $\mathcal{L}_r$
        Calculate $\mathcal{L}_a$
        $\min_{I,C,R} \mathcal{L}_c - \mathcal{L}_a + \mathcal{L}_r$             ▷ backprop gradients
        for  k:=1 to K do                                      ▷ K is a hyper-parameter
            $\min_{A} \mathcal{L}_a$                                 ▷ backprop gradients
```