ANALYSING BIAS IN SPOKEN LANGUAGE ASSESSMENT USING CONCEPT ACTIVATION VECTORS

Xizi Wei[1,2]*, Mark J.F. Gales*[2] *and Kate M. Knill*[2]

[1]School of Engineering, The University of Birmingham, UK
[2]ALTA Institute, Cambridge University Engineering Department, UK
xxw395@bham.ac.uk; {mjfg,kate.knill}@eng.cam.ac.uk

ABSTRACT

Deep learning based approaches yield the state-of-the-art performance on a wide range of tasks. A significant concern with these approaches is that they are difficult to interpret. This means that detecting bias in network predictions, for example due to bias in the training data, can be challenging. Concept Activation Vectors (CAVs) have been proposed to address this problem. These use representations, perturbations of activation function outputs, of interpretable concepts to analyse how the network is influenced by the concept. This work applies CAVs to assess bias in a spoken language assessment (SLA) system, a regression task. One of the challenges with SLA is the wide-range of concepts that can introduce bias, for example L1, age, acoustic conditions, and particular human graders, or the grading instructions. Simply generating large quantities of expert marked data to check for all forms of bias is impractical. This paper uses CAVs applied to the training data to identify concepts that might be of concern, allowing a more targeted data set to be collected to assess bias. The ability of CAVs to detect bias is assessed on the BULATS speaking test using both a standard system and a system to which bias was artificially introduced.

Index Terms— spoken language assessment, bias in deep learning, concept activation vectors.

1. INTRODUCTION

The demand for foreign language learning is growing, so is the need for standardised, universally accepted examination processes to evaluate candidates' language proficiency. This high demand leads to a shortage of capable examiners, especially for spoken language assessment. Significant progress has been made in applying various techniques from speech processing, machine learning (ML) and Artificial Intelligence (AI) to automate the language assessment process [1, 2, 3, 4]. With ML/AI based systems a concern, however, arises about potential bias within the system. For example, systematic biases with respect to gender, race and age have been found in analyses of facial recognition [5] and gender classification [6] computer vision systems, and social biases detected in nature language processing (NLP) models [7, 8].

It is crucial for trust in an exam's results that any automated spoken language assessment (SLA) system is unbiased. In other words, the system should be insensitive to factors that should not affect the candidate's exam score (e.g. first language (L1), gender, age, etc.). Bias typically arises in ML/AI systems due to uneven training data and/or human biases in the training set. The data used to train SLA systems is susceptible to both of these. The training data generally consists of recorded speech responses to questions and scores awarded by human examiners. There are likely to be variations such as the number of candidates available for each L1 (or accent) and the proficiency distribution of candidates per L1. In addition, the recording conditions and levels of background noise can vary considerably across countries and test centres. This affects the accuracy of automatic speech recognition (ASR) used to hypothesise what the candidate said. There will typically be a range of examiners used. Whereas the scores awarded by expert examiners have over 95% agreement, this drops for examiners in the field. Humans may also have unconscious biases to particular groups of candidates.

Analysis of bias in deep learning-based systems can be used to help understand a system's behaviour. If bias is detected, actions can be made to compensate, then to improve the fairness of the system. With a working high-performance system, it is more effective to analyse the behaviour of the system without retraining or modifying the model. Instead the network is perturbed in some way and the impact of the perturbations on the network's performance checked. One way to achieve this is to use data points [9] or perturbed features [10]. With these methods, however, there is a concern that the analysis is only true for a particular set of data

Researches have shown that linear classifiers can learn meaningful directions from the latent space of the networks [11, 12]. Kim et al. [13] proposed Test with Concept Activation Vector (TCAV) to quantify the sensitivity of model predictions to a human-defined high-level concept using directional derivatives and applied this to image classification. For each class, TCAV computes the fraction of inputs whose activation vectors are positively related to a concept. It measures how important a concept (e.g 'striped') is for the prediction of a given class (e.g 'zebras'). [13] also showed that TCAV can reveal biases e.g. it found the 'female' concept highly relevant to the 'apron' class. This technique (referred to as Attribute Vectors in [13]) was extended in [14] to measure the sensitivity to a concept directly from the model predictions for bias detection. Bias is detected in their smiling classifier model by assessing how a prediction changes if characteristics of an image are altered in a specific targeted manner.

This paper proposes applying CAVs to spoken language assessment to look for concepts that may be of concern. They provide a way to analyse the trained model and training data to efficiently detect concepts to further analyse through held-out, expert graded test data. Since only concepts of concern need to be tested on the latter, this can reduce time and costs of expert grading. Since the SLA scoring is a regression task CAVs are here extended from their pre-

This report reports on research supported by Cambridge Assessment, University of Cambridge. Xizi Wei was employed part-time in the ALTA Institute for this work. The authors would like to thank the ALTA Speech Team for baseline sytems, and making the ASR system output available.

vious use in classification. Perturbations in the direction of a CAV are assessed to see if they impact the predicted SLA score of a well trained high performing SLA system. The impact of the CAV directly on the score is also considered. Concepts identified by CAVs on the training data as potentially a source of bias were verified on held-out test data. Both concepts expected to affect (e.g. grade) and be independent of the score (e.g. L1) were investigated.

In the rest of this paper, Section 2 discusses the CAV-based methods to assess the sensitivity of a network to a concept. The experiments and results are presented in Section 3, followed by conclusions.

2. CONCEPT ACTIVATION VECTORS (CAVs)

The aim of this work is to assess the sensitivity of a regression task network to a particular, human interpretable, concept. It is assumed that there is supervised training data, $\mathcal{D}$, comprising feature vectors, $\boldsymbol{x}^{(i)}$, and associated score, $y^{(i)}$, that is used to train the network parameters, $\boldsymbol{\theta}$, and assess predictions, $\hat{y}^{(i)}$,

$$\mathcal{D} = \left\{ \{\boldsymbol{x}^{(i)}, y^{(i)}\} \right\}_{i=1}^{N}; \quad \hat{y}^{(i)} = \mathcal{F}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \tag{1}$$

The predictions from the network can be split into two distinct stages, mapping from the input to the activation function output from a particular layer, $\boldsymbol{h}^{(i)}$, and then from that activation vector to the output

$$\boldsymbol{h}^{(i)} = \mathcal{F}_{\mathtt{h}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}); \quad \hat{y}^{(i)} = \mathcal{F}_{\mathtt{y}}(\boldsymbol{h}^{(i)}; \boldsymbol{\theta}) \tag{2}$$

In this work the perturbations are applied at the function activation vector level. Thus the perturbed output for a particular concept $\mathcal{C}_c$, $\tilde{y}_{\alpha}^{(ci)}$, can be expressed as

$$\tilde{y}_{\alpha}^{(ci)} = \mathcal{F}_{\mathtt{y}}(\boldsymbol{h}^{(i)} + \alpha \Delta \boldsymbol{h}^{(c)}; \boldsymbol{\theta}) \tag{3}$$

where α indicates the level of perturbation being applied for that concept. The aim is to derive the appropriate perturbation on the activation function vector related to a particular concept $\mathcal{C}_c$, $\Delta \boldsymbol{h}^{(c)}$, to examine whether that concept perturbation alters the score.

In this work CAVs will be assessed in terms of the direction of the CAV and also the assessment system will be analysed at two levels: the individual speaker; averaged over all speakers. The direction of the CAV, $\boldsymbol{d}^{(c)}$, is obtained by minimising the following hinge-loss function (with an L2-norm penalty on $\boldsymbol{d}^{(c)}$)

$$\mathcal{L}(\boldsymbol{d}^{(c)}, b) = \sum_{i=1}^{N} \max\left\{0, 1 - t^{(i)}(\boldsymbol{d}^{(c)\mathtt{T}}\boldsymbol{h}^{(i)} + b)\right\} \tag{4}$$

where $t^{(i)} \in \{1, -1\}$ is the target defined by splitting the training data $\mathcal{D}$ according to concept $\mathcal{C}_c$. When analysing the impact of CAVs on the system, especially when comparing over different network configurations, it is useful to have consistent magnitudes of the CAV vectors compared to the activation function outputs that they act on. In this work the average activation L2-norm for all the training data is used. The CAV for concept $\mathcal{C}_c$, $\Delta \boldsymbol{h}^{(c)}$, is

$$\Delta \boldsymbol{h}^{(c)} = h_{\mu} \boldsymbol{d}^{(c)} / ||\boldsymbol{d}^{(c)}||_2; \quad h_{\mu} = \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{h}^{(i)}||_2 \tag{5}$$

The concepts can be split into two distinct groups based on: predicted score values; more general concepts such as gender. For example for the spoken language assessment presented, the first set of CAVs will be based on grades, such as the concept, <B2, of having a score less than 4.0 corresponding to CEFR level [15] grades below B2. For the extracted CAVs it is then necessary to examine whether perturbations in the direction of the CAV impact the score. Three metrics are considered. The first two are gradient based

$$\mathcal{B}_{\nabla}^{(c)} = \mathtt{cos}\left(\Delta \boldsymbol{h}^{(c)}, \frac{1}{N} \sum_{i=1}^{N} \left. \frac{\partial \mathcal{F}_{\mathtt{y}}(\boldsymbol{h}; \boldsymbol{\theta})}{\partial \boldsymbol{h}} \right|_{\boldsymbol{h}^{(i)}}\right) \tag{6}$$

$$\mathcal{B}_{\mathtt{gr}}^{(c)} = \frac{1}{N} \sum_{i=1}^{N} \mathtt{cos}\left(\Delta \boldsymbol{h}^{(c)}, \left. \frac{\partial \mathcal{F}_{\mathtt{y}}(\boldsymbol{h}; \boldsymbol{\theta})}{\partial \boldsymbol{h}} \right|_{\boldsymbol{h}^{(i)}}\right) \tag{7}$$

where $\mathtt{cos}(\boldsymbol{a}, \boldsymbol{b}) = 1 - \boldsymbol{a}^{\mathtt{T}}\boldsymbol{b}/(||\boldsymbol{a}||_2||\boldsymbol{b}||_2)$. CAVs that are orthogonal to the gradients, with a cosine distance of 1.0, will not be expected to impact the scores. Otherwise the CAV may influence the score, unless that direction is removed by subsequent layers of the network. Eq (6) is efficient to compute and rapidly examine any CAV. Conversely Eq (7) requires examining all, or a sufficiently large subset, training examples but enables the distance to individual samples $\mathcal{B}_{\mathtt{gr}}^{(ci)}$ to also be computed.

The final metric is the impact that the CAV has on the score, relative to the predicted score with no CAV:

$$\mathcal{B}_{\mathtt{as}}^{(c)} = \frac{1}{N} \sum_{i=1}^{N} (\tilde{y}_{\alpha}^{(ci)} - \hat{y}^{(i)}); \tag{8}$$

This metric is sensitive to the absolute value of the CAV vector, in addition to the direction but directly assesses the impact on the score.

3. EXPERIMENTS

3.1. Spoken Language Assessment Systems

The data used for these experiments are from candidates taking the BULATS, Use of Business English test [16]. It comprises five sections: an initial short answer section; a read-aloud section; three general free speaking prompt-response answers. The models were trained on data from 4303 candidates over a range of L1s, genders, ages and candidate grades (the Training data). Scores for each section of the test were generated by trained graders, operating within local test centers. The scores were then averaged over all sections to yield a score in the range 0 to 6, which can then be mapped to the appropriate CEFR level. The Evaluation data comprises 225 held-out candidates from 6 L1s (all L1s were seen in training): Arabic, Dutch, French, Polish, Thai and Vietnamese. For this test set candidates are approximately evenly distributed over CEFR levels A1, A2, B1, B2 and C (C1 and C2 merged). The grades of the Evaluation data are given by expert graders with inter-grader agreement of 0.95 to 0.97, which are assumed to have no bias for assessment. These expert graders score each section, which are then averaged to again yield a score in the range 0 to 6.

A deep density network (DDN) based assessment system was built using the same architecture and features as [17]. The ASR system, used to derive the features, had an average WER on the evaluation data of 19.5%, for more details of the ASR architecture see [1, 17]. A Gaussian distribution is predicted by the DDN across all the test sections, with the Gaussian mean taken as the predicted score. The DDN comprised 2 hidden layers of 180 units activated with Leaky Rectified linear activation (LReLU) functions. In addition to the standard grader, a deliberately biased grader was constructed. This had an identical architecture, but all the training scores associated with the Thai L1 were increased by one CEFR grade, an

increase of 1.0, yielding a biased network. The standard grader will be denoted as Ge (general grader) and the Thai-biased grader as Ge_b.

For the final score prediction from both graders an ensemble of 5 models were combined, using random seeds to generate diversity in the ensemble [18]. The performance on the evaluation data of the individual models and ensembles are shown in Table 1. Standard performance metrics are shown: Pearson Coefficient (PCC), Root Mean Square Error (RMSE), percentage of the predictions that are within 0.5 (%<0.5) or 1.0 (%<1.0) grades of the expert score. Scatter plots for the evaluation data predictions for the two graders are shown in Figure 1. The biased system, scored against the expert grades, performs less well, with the Thai L1 speakers badly predicted.

Grader		PCC	RMSE	%< 0.5	%< 1.0
Ge	Ind	$0.89_{\pm 0.00}$	$0.55_{\pm 0.01}$	$65.7_{\pm 1.3}$	$95.0_{\pm 0.6}$
	Ens	0.89	0.55	65.8	96.4
Ge_b	Ind	$0.84_{\pm 0.00}$	$0.71_{\pm 0.01}$	$57.0_{\pm 1.2}$	$82.9_{\pm 1.2}$
	Ens	0.83	0.71	56.3	83.3

Table 1: Individual model (Ind) and ensemble (Ens) performance of standard (Ge) and Thai-biased (Ge_b) graders on Evaluation data against expert scores. Range indicates $\pm\sigma$.

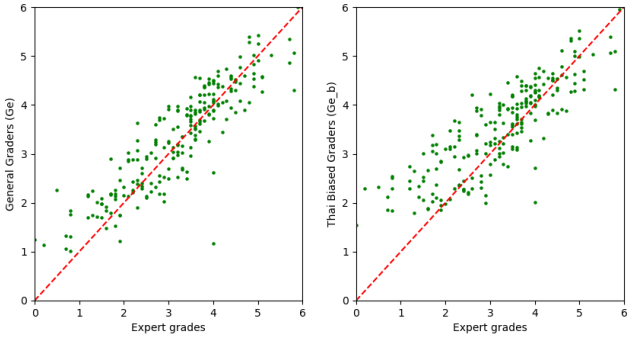


Fig. 1: Ensemble standard (Ge) and Thai-biased (Ge_b) grader scores against expert scores on Evaluation data.

3.2. Concept Activation Vector Analysis

CAVs for both the standard and biased system were constructed using the Training data and Biased Training data. Three distinct groups of CAVs were built: grade, based on <B1 (<3.0), <B2 (<4.0) and <C1 (<5.0); L1, for which there was a reasonable quantity in the training data; and gender. These CAVs were estimated using activation function outputs from the first hidden layer of the individual models for both graders. An important aspect of the CAVs is whether the decision that determines the CAV can appropriately partition the space to yield a high accuracy of concept classification. The classification performance on the training data is shown as Acc(%) in Table 2, it measures the performance of the linear classifier trained in Eq (4). High accuracy can be observed for all concepts.

Table 2 also shows the cosine distance for each of the CAVs to the average score gradient for each of the graders. A cosine distance of 1.0 would indicate a CAV is orthogonal to the average gradient and so the concept has no effect on predictions. For both the standard and biased graders the grade CAVs all indicate the CAVs direction is towards lower grades showing the grade concepts all have a negative impact on scores. In general the L1 and gender CAVs indicate that the scores are not sensitive to these concepts. The CAV cosine distance indicates that Dutch may have a bias for the standard grader and Thai for the biased grader, both indicating positive score biases.

Concept	Ge		Ge_b	
	Acc (%)	$\mathcal{B}_\nabla^{(c)}$	Acc (%)	$\mathcal{B}_\nabla^{(c)}$
<B1	$88.9_{\pm 0.2}$	$1.58_{\pm 0.02}$	$86.6_{\pm 0.4}$	$1.67_{\pm 0.02}$
<B2	$87.7_{\pm 0.2}$	$1.61_{\pm 0.03}$	$86.6_{\pm 0.3}$	$1.73_{\pm 0.02}$
<C1	$96.8_{\pm 0.1}$	$1.47_{\pm 0.06}$	$96.2_{\pm 0.1}$	$1.45_{\pm 0.02}$
Thai	$94.6_{\pm 0.2}$	$1.04_{\pm 0.02}$	$95.9_{\pm 0.2}$	$\mathbf{0.73}_{\pm 0.05}$
Spanish	$84.3_{\pm 0.2}$	$1.10_{\pm 0.04}$	$86.2_{\pm 0.6}$	$1.13_{\pm 0.02}$
Arabic	$88.1_{\pm 0.4}$	$0.96_{\pm 0.06}$	$90.1_{\pm 0.5}$	$1.13_{\pm 0.01}$
Viet.	$92.1_{\pm 0.2}$	$0.84_{\pm 0.04}$	$92.9_{\pm 0.3}$	$0.92_{\pm 0.04}$
Polish	$92.9_{\pm 0.2}$	$0.97_{\pm 0.02}$	$93.6_{\pm 0.1}$	$0.99_{\pm 0.07}$
Dutch	$95.5_{\pm 0.1}$	$\mathbf{0.78}_{\pm 0.03}$	$96.2_{\pm 0.2}$	$0.91_{\pm 0.04}$
Female	$96.1_{\pm 0.2}$	$1.04_{\pm 0.02}$	$96.4_{\pm 0.1}$	$0.94_{\pm 0.03}$

Table 2: CAV accuracy and cosine distance on Training data for standard (Ge) and Thai-biased (Ge_b) graders. Range indicates $\pm\sigma$.

The cosine distance of a subset of the concepts to the average gradient cosine distance ($\mathcal{B}_\text{gr}^{(c)}$) and average shift ($\mathcal{B}_\text{as}^{(c)}$) on the training data are shown in Table 3. These show the same trends as the cosine distance to the average gradient in Table 2, again indicating possible concerns about Dutch for the standard grader and Thai for the biased grader. Note for both gradient based measures only the direction of the CAV is of importance. This makes the measure invariant to the need to normalise the length of the CAV. Conversely the average shift in the score caused by the CAV depends on both the direction and the magnitude. Thus this gives a better concept of the impact of the CAV, but requires appropriate normalisation. The difference is illustrated in Table 3. The smallest cosine distance (Dutch for Ge and Thai for Ge_b) have similar values, but the average shift for Thai for the biased system is larger.

Concept	Ge		Ge_b	
	$\mathcal{B}_\text{gr}^{(c)}$	$\mathcal{B}_\text{as}^{(c)}$	$\mathcal{B}_\text{gr}^{(c)}$	$\mathcal{B}_\text{as}^{(c)}$
<B2	$1.57_{\pm 0.03}$	$-0.21_{\pm 0.01}$	$1.63_{\pm 0.02}$	$-0.31_{\pm 0.02}$
Thai	$1.05_{\pm 0.02}$	$-0.01_{\pm 0.01}$	$\mathbf{0.78}_{\pm 0.04}$	$\mathbf{0.13}_{\pm 0.03}$
Arabic	$0.96_{\pm 0.05}$	$0.02_{\pm 0.02}$	$1.12_{\pm 0.01}$	$-0.06_{\pm 0.00}$
Viet.	$0.85_{\pm 0.04}$	$0.06_{\pm 0.01}$	$0.93_{\pm 0.03}$	$0.04_{\pm 0.02}$
Dutch	$\mathbf{0.79}_{\pm 0.03}$	$\mathbf{0.08}_{\pm 0.01}$	$0.92_{\pm 0.03}$	$0.04_{\pm 0.02}$
Female	$1.03_{\pm 0.02}$	$-0.01_{\pm 0.01}$	$0.94_{\pm 0.03}$	$0.03_{\pm 0.01}$

Table 3: Standard (Ge) and Thai-biased (Ge_b) graders CAV impact on Training data, $\alpha = 0.1$. Range indicates $\pm\sigma$.

Tables 2 and 3 have assessed possible bias in terms of CAVs. It is also possible to look at the performance of the concepts on the Training data. Unfortunately as the networks themselves are trained on this data, they will reproduce any bias present in the Training data labels. For example examining the RMSE (between the network prediction and training data) for each of the concepts for the Thai biased grader, there's little difference between Thai (0.436) and Arabic (0.444). This is one of the reasons that CAVs are adopted as system performance on training data (or associated dev data) will not identify bias in that training data.

3.3. Concept Evaluation Data Performance

The previous section has used CAVs to analyse the network to detect concepts for which there may be unwanted bias, in this case Dutch for the standard grader and Thai for the biased grader. In practice evaluation data focused on these concepts would then be collected

and accurately scored to enable complete analysis of whether there is bias actually present. For this work the Evaluation data, with expert scores, can be used. Two metrics are used for this analysis; RMSE to give an overall assessment of the performance, and average error (AveE) to assess whether the system performance for a concept over or under-predicts on average. PCC is not used here as any concept bias, for example a shift of all the scores, does not impact PCC.

Table 4 shows the performance of the two graders, standard (Ge) and Thai biased (GE_b), on the Evaluation data, as well as the performance of the data associated with particular concepts [1]. For the standard grader it can be seen that Dutch as a concept does not seem to indicate significant bias compared to the overall system performance, either from the RMSE or average error. On this grader the predicted grades for Thai were slightly generous, shown by a higher average error, but the RMSE values were in-line with other concepts. For the biased grader (Ge_b) the story is clearer as Thai performance has both a large RMSE (1.076) and a large average error (0.927) compared to both the system in general and the other concepts.

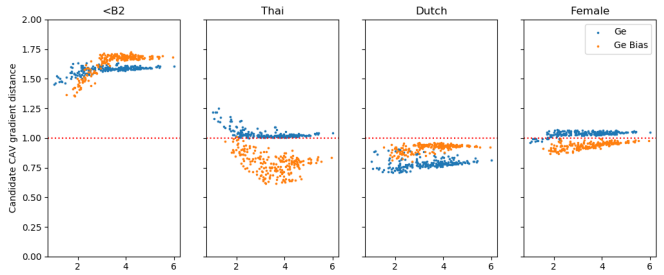
Concept	Ge		Ge_b	
	RMSE	AveE	RMSE	AveE
—	0.551	0.125	0.710	0.307
Thai	0.559	0.343	**1.076**	**0.927**
Arabic	0.561	0.071	0.571	0.105
Viet.	0.617	0.160	0.711	0.386
Polish	0.509	-0.006	0.496	0.041
Dutch	0.484	0.043	0.577	0.140
Female	0.503	0.172	0.726	0.411
Male	0.593	0.081	0.693	0.207

Table 4: Standard (Ge) and Thai-biased (Ge_b) graders ensemble performance against expert scores on Evaluation data.

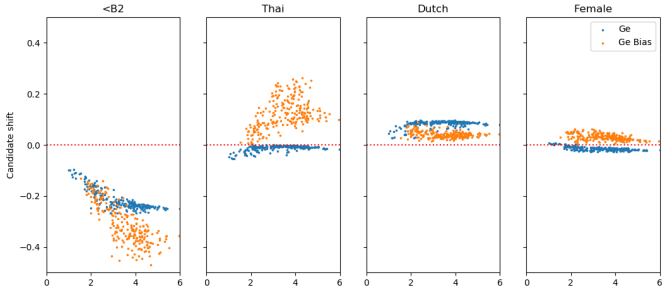
From the results in Table 4 it is clear that when the CAVs have indicated that the concept is unlikely to have bias, this is reflected in the evaluation scores. Similarly strong bias observed in the Evaluation data matches to CAVs indicating bias so limited evaluation data aimed at assessing the bias for a particular concept can be used. A CAV indicating bias may be present is not sufficient in itself. Given that only limited data are often available for a particular concept there will be noise on the results. A decision will need to be made as to whether more data is needed to determine if there is bias.

Having examined the overall performance of the system and the concept specific performance, it is interesting to examine the relationship between the CAVs and individual candidates. Figure 2 shows the individual CAV distance for the gradients ($\mathcal{B}_{\text{gr}}^{(ci)}$) and assessment score shifts ($\mathcal{B}_{\text{as}}^{(ci)}$) on the Evaluation data candidates. To make the figures clearer the x-axis is based on predicted scores, rather than expert scores. As expected the grade CAV has a large impact on both graders and using both distances. For Thai there is a clear difference between the standard and the Thai-biased graders. As expected the biased grader shows a distinct shift down for the cosine-distance, indicating the CAV is in-line with the gradient for those candidates, and a positive shift for the assessment score. It is interesting that for Dutch there is more of a difference in the gradient direction compared to the assessment shift. Finally for gender (the female CAV) it can be seen that both graders have an orthogonal direction to the gradient (cosine distance of 1.0) and minimal assessment shift.

[1]There is no Spanish data in the Evaluation data so this concept was not examined. From the previous section, however, this L1 does not appear to be a concept of concern.



(a) Candidate CAV gradient distance $\mathcal{B}_{\text{gr}}^{(ci)}$



(b) Candidate shift ($\mathcal{B}_{\text{as}}^{(ci)}$) with $\alpha = 0.1$

Fig. 2: Impact of CAVs on standard (Ge) and Thai-biased (Ge_b) ensemble graders on Evaluation data candidates, x-axis predicted score $\hat{y}^{(i)}$.

4. CONCLUSIONS

This paper has examined the problem of detecting bias in deep learning based systems for regression tasks, in particular for spoken language assessment. One of the issues with these deep learning approaches is that the networks are highly distributed and non-linear making network analysis very challenging. This means that it is hard to determine whether there may be bias in the network simply given the network parameters. Testing on held-out expert scored evaluation data can show if bias exists. This, however, requires sufficient examples of each concept to be checked. Given the large number of possible concepts, for example L1 and gender, that can possibly have biased output, it is a challenge to collect such evaluation data. This means that alternative approaches for network analysis are required.

This paper examines the use of Concept Activation Vectors (CAVs) for analysing possible bias in the network. A CAV is associated with a particular activation layer output of the network, and concept, such that moving in the direction of the CAV has the impact of increasing the "quantity" of that concept on the input. If the score is independent of the CAV then that concept is not an influence on the system performance, otherwise the concept might be a source of bias if its influence is unexpected. This approach, requiring only training data or labelled development data which may include bias, allows the analysis of a wide range of concepts. By using CAVs to determine selected concepts of interest from a much larger initial set, the amount of expert scored evaluation data required for full bias analysis can be significantly reduced.

Both gradient-based and score-based CAV distance measures are described and the performance evaluated on the BULATS speaking test. From the results when a CAV indicates that there is no bias, then this is reflected in the evaluation data results. When the CAV indicates for the training data that there may be bias, then evaluation data can be used to determine the impact of the concept on the score.

5. REFERENCES

[1] Y. Wang, M.J.F. Gales, Kate Knill, Konstantinos Kyriakopoulos, Andrey Malinin, Rogier van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, 09 2018.

[2] A. Malinin, A. Ragni, K. Knill, and M. Gales, "Incorporating uncertainty into deep learning for spoken language assessment," in *Proc. ACL*, 2017.

[3] Yao Qian, Patrick L. Lange, Keelan Evanini, Robert A. Pugh, Rutuja Ubale, Matthew David Mulholland, and X. Wang, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *Proc. ICASSP*, 2019.

[4] C. Baur, Andrew Caines, C. Chua, J. Gerlach, Mengjie Qian, Manny Rayner, M. Russell, H. Strik, and Xizi Wei, "Overview of the 2019 Spoken CALL Shared Task," in *Proc. ISCA International Workshop on Speech and Language Technology in Education (SLaTE)*, 2019.

[5] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[6] Inioluwa Deborah Raji and Joy Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[7] Su Lin Blodgett and Brendan O'Connor, "Racial disparity in natural language processing: A case study of social media African-American English," in *Proc. Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, 2016.

[9] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," in *Proc. International Conference on Machine Learning (ICML)*, 2017.

[10] Marco Túlio Ribeiro and Sameer Singh and Carlos Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016.

[11] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Network dissection: Quantifying interpretability of deep visual representations.," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Guillaume Alain and Yoshua Bengio, "Understanding intermediate layers using linear classifier probes," in *Proc. 5th International Conference on Learning Representations, (ICLR)*, 2018.

[13] Been Kim, M. Wattenberg, J. Gilmer, C. J. Cai, James Wexler, F. Viégas, and Rory Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. International Conference on Machine Learning (ICML)*, 2018.

[14] Emily L. Denton, B. Hutchinson, Margaret Mitchell, and Timnit Gebru, "Detecting bias with generative counterfactual face attribute augmentation," in *Proc. Fairness, Accountability, Transparency and Ethics in Computer Vision Workshop (in conjunction with CVPR)*, 2019.

[15] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge University Press, 2001.

[16] L. Chambers and K. Ingham, *The BULATS online speaking test*, 2011, [Online], Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf.

[17] X. Wu, K. Knill, and M. Gales, "Ensemble approaches for uncertainty in spoken language assessment," in *Proc. INTERSPEECH*, 2020.

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. 31st International Conference on Neural Information Processing Systems*, 2017.