INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Ensemble Approaches for Uncertainty in Spoken Language Assessment

Xixin Wu[1]*, Kate M. Knill*[1]*, Mark J.F. Gales*[1] *and Andrey Malinin*[2]

[1] ALTA Institute, Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK.
[2] Yandex, Moscow, Russia

{xw369,kate.knill,mjfg}@eng.cam.ac.uk,am969@yandex-team.ru

Abstract

Deep learning has dramatically improved the performance of automated systems on a range of tasks including spoken language assessment. One of the issues with these deep learning approaches is that they tend to be overconfident in the decisions that they make, with potentially serious implications for deployment of systems for high-stakes examinations. This paper examines the use of ensemble approaches to improve both the reliability of the scores that are generated, and the ability to detect where the system has made predictions beyond acceptable errors. In this work assessment is treated as a regression problem. Deep density networks, and ensembles of these models, are used as the predictive models. Given an ensemble of models measures of uncertainty, for example the variance of the predicted distributions, can be obtained and used for detecting outlier predictions. However, these ensemble approaches increase the computational and memory requirements of the system. To address this problem the ensemble is distilled into a single mixture density network. The performance of the systems is evaluated on a free speaking prompt-response style spoken language assessment test. Experiments show that the ensembles and the distilled model yield performance gains over a single model, and have the ability to detect outliers.

Index Terms: spoken language assessment, uncertainty estimation, computer-aided language learning, ensemble, distillation

1. Introduction

Access to jobs and study opportunities abroad is increasingly dependent upon an applicant demonstrating sufficient proficiency in the local or working language. For example, English language proficiency is assessed through standardised, universally accepted, examinations such as International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL). With more than 1.5 billion people predicted to be learning English as an additional language by 2020 [1], it will be difficult to train sufficient examiners so some level of automatic assessment of English skills is required. Hence, the demand for automatic assessment systems keeps growing. These automatic systems must be capable of yielding high quality predictions of a candidate's score.

Machine learning, and deep learning in particular, has dramatically improved the performance of automated assessment systems including spoken language learning and assessment. One issue with these deep learning approaches is they tend to be overconfident in their decisions. In speaking tests if a candidate has a skill level and/or first language not seen during training then it is likely that the system will yield an incorrect score which could have damaging consequences [2]. It is therefore crucial that the automated systems are able to detect when they "don't know", i.e. are uncertain in their predictions [3–6], in order to avoid errors. There are two sources of uncertainty in predictions: data and knowledge uncertainty. Data uncertainty arises from the natural complexity of the data and noisy observations. Knowledge uncertainty is caused by a mismatch between the training and test distributions, also known as dataset shift [7]. This situation often arises in real world applications. It can be reduced by providing more knowledge, in the form of more training data from regions associated with high knowledge uncertainty, to the model.

Van Dalen et al [8] proposed the use of a Gaussian Process (GP) for auto-marking free speaking English language tests. The predicted test score is given by the GP mean, whilst the GP variance is taken to indicate the level of uncertainty of the automarker. Whilst they produce good quality predictions, standard GPs scale poorly with data so increasing the amount of training data (required to reduce mismatches between training and test) results in impractical GP models. To address this [9–11] applied neural approaches to the problem. Neural networks do not suffer from the scaling problem allowing practical models to be trained on much larger quantities of data. In particular a deep density network (DDN) [12] was shown to outperform GPs and deep neural networks (DNNs) with Monte-Carlo dropout [13] in uncertainty based rejection for automatic assessment [9]. As with any DNN, the performance of individual DDNs is sensitive to a range of factors including the initial network parameters.

This paper considers the use of ensemble approaches to reduce model sensitivity to parameter initialization and increase the reliability of generated scores. Previously [8, 9] looked at detecting uncertainty to improve the overall performance by passing rejected scores to humans for grading. Here, the focus is on detecting the most unacceptable scores. Ensembles make use of the property that in an ensemble of independent models the in-domain predictions will be consistent whilst 'undefined' behaviour in each model yields a diverse set of predictions for out-of-distribution (OOD) inputs. By using this property, both data and knowledge uncertainty can be assessed within a single consistent probabilistic framework without the need for OOD training data. The latter, used in multi-task DDN models (DDN-MT) gave better rejection performance in [9] but is sensitive to the choice of OOD data. However, ensemble approaches do not scale well with the number of ensemble members. Computational cost significantly increases as more members are included, as required for sensitivity reduction and reliability enhancement. One possible solution is to distill the ensemble into a single model [14, 15]. Previous ensemble distillation ap-

This paper reports on research supported by Cambridge Assessment, University of Cambridge. The authors would also like to acknowledge helpful discussions with members of the ALTA Speech team and Machine Learning Group. Andrey Malinin completed this work while a member of the ALTA Institute.

http://dx.doi.org/10.21437/Interspeech.2020-2238

proaches have mainly focused on classification problems. In this work, the ensemble of regression models is distilled into a single model, a mixture density network (MDN), by minimizing the KL-divergence between the predictive distribution of the ensemble and that of the single model. Experiments show the distilled MDN achieves comparable, or better, performance than the ensemble on predicting grades and detecting unacceptable scores.

In the rest of this paper, Section 2 discusses the uncertainty measures for regression ensemble. The methods and experiments for spoken language assessment are presented in Sections 3 and 4, followed by conclusions.

2. Uncertainty for Regression

The aim of this work is to examine the uncertainty associated with the prediction of a (continuous) score for a particular candidate taking a spoken language assessment test. Given some training data $\mathcal{D}$ and the observation features $\boldsymbol{x}^\star$ from the candidate's test responses, this prediction of the score y can be expressed as

$$\begin{aligned}\mathtt{p}(y|\boldsymbol{x}^\star,\mathcal{D}) &= \mathbb{E}_{\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})}\left[\mathtt{p}(y|\boldsymbol{x}^\star,\boldsymbol{\theta})\right] = \int \mathtt{p}(y|\boldsymbol{x}^\star,\boldsymbol{\theta})\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\ &\approx \frac{1}{K}\sum_{i=1}^{K}\mathtt{p}(y|\boldsymbol{x}^\star,\boldsymbol{\theta}^{(i)}); \quad \boldsymbol{\theta}^{(i)} \sim \mathtt{p}(\boldsymbol{\theta}|\mathcal{D}). \end{aligned} \tag{1}$$

where the second expression is an ensemble, Monte-Carlo approximation to the full Bayesian integration, with $\boldsymbol{\theta}^{(i)}$ the ith component parameters. This ensemble approximation effectively yields a K-component mixture form for the score distribution.

Having derived a score distribution, a standard measure of the uncertainty associated with this score is given by the (differential) entropy of the distribution

$$\mathcal{H}\left[\mathtt{p}(y|\boldsymbol{x}^\star,\mathcal{D})\right] = -\int \mathtt{p}(y|\boldsymbol{x}^\star,\mathcal{D})\log\left(\mathtt{p}(y|\boldsymbol{x}^\star,\mathcal{D})\right)dy. \tag{2}$$

This expression gives a measure of the total uncertainty in the prediction of Monte-Carlo approximation. For Gaussian distribution the differential entropy is proportional to the log of the variance. However, this expression has no closed-form solution for mixture models.

With ensemble approaches it is possible to split the total uncertainty into data and knowledge uncertainty. Data uncertainty arises from the natural complexity of the data, corresponding to points where there is severe class overlap. Knowledge (also known as epistemic or distributional) uncertainty arises from a mismatch between the distributions of the training and test data. The ensemble approaches derive measures of the model consistency, as well as individual model uncertainty, based on the posterior distribution of model parameters from the training data, $\mathtt{p}(y|\boldsymbol{x},\boldsymbol{\theta})$. There are a range of options that can be used. In this paper the total variance is considered. Based on the law of total variance [16]

$$\underbrace{\mathbb{V}[y|\boldsymbol{x}^\star,\mathcal{D}]}_{\texttt{Total Variance}} = \underbrace{\mathbb{V}_{\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})}\left[\mathbb{E}_{\mathtt{p}(y|\boldsymbol{x}^\star,\boldsymbol{\theta})}[y]\right]}_{\texttt{Mean Variance}} + \underbrace{\mathbb{E}_{\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})}\left[\mathbb{V}_{\mathtt{p}(y|\boldsymbol{x}^\star,\boldsymbol{\theta})}[y]\right]}_{\texttt{Expected Data Variance}}. \tag{3}$$

This yields a partition of the uncertainty of a prediction into knowledge and data uncertainty.

3. Spoken Language Assessment

For this work the predicted score for a candidate is first modeled as a Gaussian distribution. Thus, for test observation $\boldsymbol{x}^\star$, the distribution can be expressed as

$$\mathtt{p}(y|\boldsymbol{x}^\star,\mathcal{D}) = \mathcal{N}(y; f_\mu(\boldsymbol{x}^\star,\mathcal{D}), f_{\sigma^2}(\boldsymbol{x}^\star,\mathcal{D})) \tag{4}$$

Two different forms of functions for generating the mean, $f_\mu(\boldsymbol{x}^\star,\mathcal{D})$, and variance, $f_{\sigma^2}(\boldsymbol{x}^\star,\mathcal{D})$, are considered: Gaussian Processes (GPs) and deep density networks (DDNs). An ensemble of DDNs is then built to approximate the model parameter distribution. However, the ensemble consumes significantly more computational resources in both training and testing stages. Therefore, the ensemble, considered as a K-component Gaussian mixture distribution, is distilled into one single mixture density network (MDN) model.

3.1. Deep Density Networks

Deep density networks (DDNs) are used to predict the parameters of a distribution, in this case a Gaussian distribution [12]. The training data, $\mathcal{D}$ is used to train a set of model parameters $\boldsymbol{\theta}$ by maximizing the likelihood of the training data

$$\mathcal{L}_{\mathtt{ml}}(\boldsymbol{\theta}) = \sum_{\{y,\boldsymbol{x}\}\in\mathcal{D}} \log\left(\mathtt{p}(y|\boldsymbol{x},\boldsymbol{\theta})\right). \tag{5}$$

One of the issues with using this criterion is that it is hoped that the variance will increase as the distance from the training data increases. To try and ensure that this happens a multi-task loss-function (DDN-MT) can be used where the variances for an out-of-domain training set $\tilde{\mathcal{D}}$ are large and the training is now based on KL-divergence [9, 17]

$$\begin{aligned}\mathcal{L}_{\mathtt{mt}}(\boldsymbol{\theta}) = &\sum_{\{y,\boldsymbol{x}\}\in\mathcal{D}} \mathcal{KL}[\mathtt{p}(y|\boldsymbol{x})||\mathtt{p}(y|\boldsymbol{x};\boldsymbol{\theta})] \\ &+ \sum_{\{y,\boldsymbol{x}\}\in\tilde{\mathcal{D}}} \mathcal{KL}[\mathtt{p}(y|\boldsymbol{x})||\mathtt{p}(y|\boldsymbol{x};\boldsymbol{\theta})]. \end{aligned} \tag{6}$$

For this work a factor analysis model trained on the in-domain training data is used to synthesize out-of-distribution training data $\tilde{\mathcal{D}}$ for the DDN-MT [9].

By using a Gaussian distribution as the predictive distribution, it is simple to derive the expressions used to measure uncertainty discussed in Section 2. As the differential entropy of a Gaussian distribution is only a function of the variance, using (2) for uncertainty will yield the same results as the total variance in (3).

3.2. Ensemble Generation

For the uncertainty measures described in Section 2 the scores all make use of the posterior distribution of the model parameters given the training data $\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})$. For deep learning models the dimensionality of the model parameters can be very large making it challenging to yield good estimates for this posterior distribution, or to generate samples from this distribution. For this work a simpler approach for generating the ensemble was adopted, where different seeds were used to initialise the DDN (or DDN-MT) model parameters [3]. This was found to yield sufficient diversity from the model predictions to form a good ensemble. It is possible to adopt more complex approaches to generate the ensemble [13, 18, 19], or to use approaches such as prior networks [20].

3.3. Ensemble Distillation

The ensemble of parallel models is distilled into the single model of mixture density network (MDN) [12], by minimizing the KL-divergence between the ensemble and the MDN

$$\mathcal{L}_{\mathsf{ed}}(\boldsymbol{\phi}) = \sum_{\{y,\boldsymbol{x}\}\in\mathcal{D}} \mathcal{KL}[\mathtt{p}(y|\boldsymbol{x};\mathcal{D})||\mathtt{p}(y|\boldsymbol{x};\boldsymbol{\phi})], \quad (7)$$

$$\mathtt{p}(y|\boldsymbol{x};\mathcal{D}) = \frac{1}{K}\sum_{i=1}^{K}\mathtt{p}(y|\boldsymbol{x},\boldsymbol{\theta}^{(i)}),$$

$$\mathtt{p}(y|\boldsymbol{x};\boldsymbol{\phi}) = \sum_{c=1}^{C} f_{\pi_c}(\boldsymbol{x},\boldsymbol{\phi})\mathcal{N}\left(y; f_{\mu_c}(\boldsymbol{x},\boldsymbol{\phi}), f_{\sigma_c^2}(\boldsymbol{x},\boldsymbol{\phi})\right).$$

Since there is no closed-form solution when $C > 1$, Monte-Carlo sampling is adopted for the minimization [21]. When $C = K$, each of the output Gaussian distributions in the MDN is considered to correspond to one of the ensemble models. The mixture factors π_c can be fixed to equal value $1/K$ and a closed-form solution can be used.

A possible alternative is to use a reverse KL-divergence $\mathcal{KL}[\mathtt{p}(y|\boldsymbol{x};\boldsymbol{\phi})||\mathtt{p}(y|\boldsymbol{x};\mathcal{D})]$, rather than the KL-divergence in (7), for ensemble distillation [22]. Whilst KL-divergence focuses on the mean of target distribution, reverse KL-divergence emphasizes the distribution modes. Since the goal in this work is to utilize the diversity of ensemble models' output Gaussian distributions, the KL-divergence is adopted to fit an approximation that covers most of the probability mass.

Given the predictive distribution of Gaussian mixture, the total variance can be derived for measuring uncertainty as

$$\mathbb{V}[y|\boldsymbol{x}^\star,\boldsymbol{\phi}] = \sum_{c=1}^{C}\pi_c(\boldsymbol{x}^\star)\sigma_c^2(\boldsymbol{x}^\star) + \sum_{c=1}^{C}\pi_c(\boldsymbol{x}^\star)\left\|\mu_c(\boldsymbol{x}^\star) - \sum_{j=1}^{C}\pi_j(\boldsymbol{x}^\star)\mu_j(\boldsymbol{x}^\star)\right\|^2. \quad (8)$$

If the component number is equal to the model number in the ensemble, i.e. $C = K$, the differential entropy of the Gaussian mixture doesn't have a closed-form solution. However, with the equal mixture factors π_c, (8) can be considered as an approximation to the ensemble total variance in (3). When $C = 1$, the MDN degenerates into a DDN and the corresponding differential entropy is a function of the variance.

4. Experiments

4.1. Data and Features

The data used for these experiments was from candidates taking the BULATS, Use of Business English, test [23]. It comprises five sections: an initial short answer section; a read-aloud section; and three more general free speaking prompt-response answers. To train the models, data from 4303 candidates was used over a range of L1s and candidate grades. The evaluation data comprised 225 candidates from 6 L1s (all L1s were seen in training): Arabic, Dutch, French, Polish, Thai and Vietnamese. For this test set candidates were approximately evenly spread over CEFR levels A1, A2, B1, B2 and C (C1 and C2 merged) [24]. The test data was graded by experts to ensure scoring reliability. The experts graded each section, and then the scores were averaged to yield a scoring range of 0-6.

The features used for the grader were based on ASR output from the system described in [25] and the baseline features described in [26], using a succeeding-word RNN Language model (WER 19.5%) [27]. The standard grader features were extended to include grade dependent language model probabilities.

For the ensemble approaches 10 models were trained and evaluated. Each model is a feed-forward neural network consisting of 2 hidden layers of 180 units activated with leaky rectified linear activation (LReLU), dropped out with rate of 0.8. Variation was introduced by initialising each model with a different random seed. The networks weights were initialised from $\mathcal{N}(0,\ 0.001)$, whilst the biases were initialised at zero. They were trained using Adam [28] with a learning rate of $1e-2$ and batch size of 50 for 100 epochs. The distilled single model has the hidden layer structure as the ensemble models. Monte-Carlo simulation with 50 samples was used for distillation.

4.2. Experimental Results

The performance of the systems was evaluated using five different metrics [6, 29]. The first three are related to ensuring that the overall performance of the system was consistent with the expert graders: Pearson Correlation Coefficient (PCC); Mean Squared Error (MSE); and Mean Absolute Error (MAE). The final two criteria are related to detecting inputs with unacceptably high prediction errors, which are defined as being outside 0.5 (i.e. over half a grade out), or outside 1.0 (i.e. over a full grade out) of the expert grade.

Model	GP	DDN		DDN-MT	
		Single	Ensbl	Single	Ensbl
PCC	88.8	$88.8_{\pm 0.3}$	88.9	$88.0_{\pm 0.6}$	88.4
MSE	0.32	$0.31_{\pm 0.01}$	0.31	$0.33_{\pm 0.02}$	0.32
MAE	0.44	$0.43_{\pm 0.01}$	0.43	$0.44_{\pm 0.02}$	0.43
$\%< 0.5$	63.5	$65.4_{\pm 2.8}$	65.8	$66.6_{\pm 2.5}$	66.7
$\%< 1.0$	94.1	$94.6_{\pm 1.2}$	95.5	$93.5_{\pm 2.5}$	94.1

Table 1: *Grading performance using single auto-markers (*`Single`*) and ensembles (*`Ensbl`*) for GP, DDN and DDN-MT models. Range indicates* $\pm 2\sigma$.

Table 1 shows the grading results for the various systems. It can be seen that the GP performs well, however, as previously mentioned this approach does not scale well as the number of training candidates increases. The ensemble approaches, both for DDN and DDN-MT, perform as well as, or better than, the average system performance but without the sometimes high variability depending on the criterion. Note it is hard to predict which members of an ensemble will perform best on a particular test set. Thus, it is not reasonable to consider simply the member of the ensemble that performs best on this subset.

Model	GP	Ensbl
$\%<0.5$ Rej. 10%	67.6	72.1
$\%<1.0$ Rej. 10%	95.9	96.4
$\mathtt{AUC}_{\mathtt{RR}}$ $\%>0.5$	7.5	31.8
$\mathtt{AUC}_{\mathtt{RR}}$ $\%>1.0$	52.4	54.4

Table 2: *Rejection performance for GPs and Ensembles of DDN-MT models based on total variance.*

One concern for spoken language assessment is to ensure that the number of candidates incorrectly classified as greater than 0.5 or 1.0 from the expert grade is minimised. From Table 1 it can be seen that even when the PCC values are in the high eighties over 30% of candidates have a prediction error

Criterion		PCC ↑	MSE ↓	MAE ↓	% <0.5 ↑		% <1.0 ↑		$\texttt{AUC}_{\texttt{RR}}$ ↑	
System	#Comp.				Rej. 0%	Rej. 10%	Rej. 0%	Rej. 10%	% >0.5	% >1.0
Ensbl	—	88.4	0.322	0.432	66.7	72.1	94.1	96.4	31.8	54.4
En-D	1⋆	88.1±0.6	0.321±0.01	0.429±0.01	67.1±1.6	73.3±1.6	93.2±1.1	95.5±1.2	28.5±4.7	54.1±13.0
	3	88.3±0.3	0.316±0.01	0.424±0.01	67.7±0.7	73.5±0.8	93.9±0.6	95.9±0.6	26.9±4.1	52.5±5.5
	5	88.5±0.4	0.310±0.01	0.420±0.01	68.6±2.4	74.0±2.5	94.2±0.7	95.9±0.6	26.5±4.7	50.4±9.0
	7	88.5±0.4	0.312±0.01	0.421±0.01	68.4±1.6	74.0±1.2	94.3±0.1	96.0±0.8	24.8±5.9	49.6±6.0
	10⋆	88.1±0.4	0.321±0.01	0.428±0.01	67.9±1.3	73.8±1.3	93.7±1.7	95.5±1.2	28.5±7.4	51.5±13.0

Table 3: *Distilling DDN-MT ensemble (*`Ensbl`*) to a single MDN model (*`En-D`*) with various numbers of Gaussian components.* ⋆`En-D` *with 1 and 10 components use a closed-form KL solution for distillation. Rejection performance is based on the total variance measure. 'Rej. 10%' stands for rejecting 10% of the test results to experts. Range indicates* $\pm 2\sigma$*, arrows inidcate desired direction.*

greater than 0.5. To address this problem it is possible to re-introduce humans into the loop to manually assess a subset of candidates; the task is now to optimally select the subset for manual grading. One approach to do this is to rank all the candidates by a measure of the confidence in the grade, and see how this reduces the number of candidates with errors greater than 0.5 or 1.0. Table 2 gives the percentage of prediction errors inside 0.5 and 1.0 after rejecting 10% of the evaluation set for manual grading, based on the uncertainty measure of total variance. It can be observed that the DDN-MT ensemble significantly outperforms the GP.

In this work the Area Under the Curve (AUC) based scheme described in [9] is also used, in order to get a measure of performance that is not highly sensitive to baseline performance. Here the following AUC rejection ratio is used

$$\texttt{AUC}_{\texttt{RR}} = \frac{\texttt{AUC}_{\texttt{rnd}} - \texttt{AUC}_{\texttt{mod}}}{\texttt{AUC}_{\texttt{rnd}} - \texttt{AUC}_{\texttt{opt}}} \times 100$$

where $\texttt{AUC}_{\texttt{rnd}}$ is the area under the random rejection curve, $\texttt{AUC}_{\texttt{opt}}$ the optimal rejection curve, and $\texttt{AUC}_{\texttt{mod}}$ the model rejection curve. As shown in Table 2, for grade error greater than 0.5 and 1.0, the ensemble still outperforms the GP on $\texttt{AUC}_{\texttt{RR}}$.

Table 3 shows that the distilled single models (En-D) with various numbers of components achieve better performance than the original ensemble in grading, and comparable rejection results. As expected, the variability of prediction errors outside 0.5 of the distilled model is reduced significantly, especially when 3 components are used, from 2.5 of DDN-MT (Table 1) to 0.7 of En-D (Table 3). The prediction errors outside 1.0 also show a similar pattern. Figure 1 shows the percentage of candidates with score error greater than 0.5 where the candidates are ranked by total variance based on an MDN model with 3 components. Candidates are then "rejected" (passed to human assessors) based on this ranking and the percentage greater than 0.5 absolute error plotted against the rejection fraction. In addition two straight lines are added: the optimal rejection where only candidates with scores greater than 0.5 are rejected; and the expected random rejection line. Figure 2 shows the same for scores with an error greater than 1.0. It can be seen that using the MDN variance for rejection performs better than random.

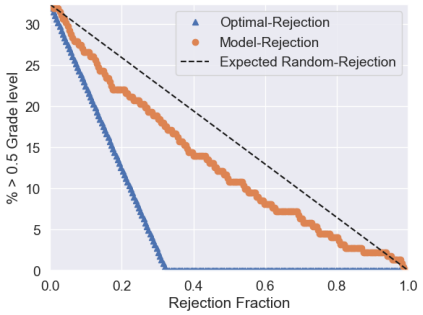

Figure 1: *Rejection performance, fraction >0.5 based on total variance of the distilled MDN with 3 components.*

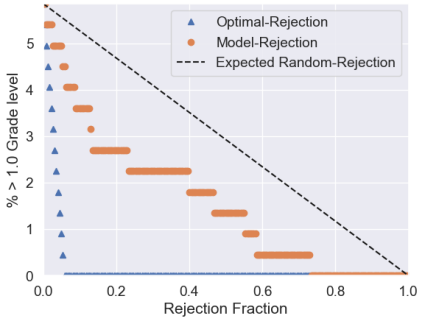

Figure 2: *Rejection performance, fraction >1.0 based on total variance of the distilled MDN with 3 components.*

5. Conclusions

For high stakes situations it is crucial that a spoken language automatic assessment system yields not only high quality score predictions but also knows when it is making predictions that are beyond acceptable bounds. In this paper ensembles of deep density networks (DDNs), with and without multi-task (DDN-MT) training were applied to assess non-native learner spoken English. The grading performance was seen to be equivalent to the average single-model performance without the inconsistency caused by individual models. Ensembles were shown to yield a better measure of the uncertainty of predicted scores than the baseline Gaussian Process based systems, leading to more accurate rejection of outlier scores. The distillation of an ensemble into a single model achieves better grading performance than the ensemble, and comparable rejection performance, whilst significantly reducing computational costs for deployment. As desired, the model sensitivity of a distilled single model is also much lower than that of the source ensemble.

In the future, the goal is to determine uncertainty within the same probabilistic framework as ensemble-based approaches, but with computational simplicity and ease of training of single-model approaches. A potential approach is to extend Prior Networks [20] to regression for this task [17].

6. References

[1] British Council, "The English Effect," Aug 2013, Research Report.

[2] A. A. Press, "Computer says no: Irish vet fails oral English test needed to stay in Australia." *The Guardian*, 2017. [Online]. Available: https://www.theguardian.com/australia-news/2017/aug/08/ computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia

[3] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *ArXiv*, vol. abs/1612.01474, 2016.

[4] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] S.-Y. Yoon and S. Xie, "Similarity-based non-scorable response detection for automated speech scoring," in *Proc. of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014, pp. 116–123.

[6] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[7] J. Quiñonero-Candela, *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[8] R. C. van Dalen, K. M. Knill, and M. J. F. Gales, "Automatically grading learners' English using a Gaussian process," in *Proc. ISCA International Workshop on Speech and Language Technology in Education (SLaTE)*, 2015, pp. 7–12. [Online]. Available: http://www.isca-speech.org/archive/slate_2015/sl15_007.html

[9] A. Malinin, A. Ragni, K. Knill, and M. J. F. Gales, "Incorporating uncertainty into deep learning for spoken language assessment," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics, ACL, Volume 2: Short Papers*, 2017, pp. 45–50. [Online]. Available: https://doi.org/10.18653/v1/P17-2008

[10] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)." in *Proc. INTERSPEECH*, 2013, pp. 1886–1890.

[11] Y. Qian, P. Lange, K. Evanini, R. Pugh, R. Ubale, M. Mulholland, and X. Wang, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8112–8116.

[12] C. M. Bishop, "Mixture density networks," Aston University, Tech. Rep. NCRG/94/004, 1994.

[13] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd International Conference on Machine Learning (ICML)*, 2016.

[14] A. Malinin, B. Mlodozeniec, and M. Gales, "Ensemble distribution distillation," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[15] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-End Speech Translation with Knowledge Distillation," in *Proc. INTERSPEECH*, 2019, pp. 1128–1132.

[16] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Proc. 35th International Conference on Machine Learning, ICML*, 2018, pp. 1192–1201. [Online]. Available: http://proceedings.mlr.press/v80/depeweg18a.html

[17] A. Malinin, "Uncertainty estimation in deep learning with application to spoken language assessment," Ph.D. dissertation, University of Cambridge, 2019.

[18] J. M. Hernández-Lobato and R. Adams, "Probabilistic backpropagation for scalable learning of Bayesian neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 1861–1869.

[19] W. Maddox, T. Garipov, P. Izmailov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for Bayesian uncertainty in deep learning," *CoRR*, vol. abs/1902.02476, 2019. [Online]. Available: http://arxiv.org/abs/1902.02476

[20] A. Malinin and M. J. F. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7047–7058. [Online]. Available: http://papers.nips.cc/paper/7936-predictive-uncertainty-estimation-via-prior-networks

[21] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. IV–317.

[22] A. Malinin and M. Gales, "Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness," in *Proc. Advances in Neural Information Processing Systems*, 2019, pp. 14 520–14 531.

[23] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

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

[25] Y. Wang, J. H. M. Wong, M. J. F. Gales, K. M. Knill, and A. Ragni, "Sequence teacher-student training of acoustic models for automatic free speaking language assessment," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 994–1000. [Online]. Available: https://doi.org/10.1109/SLT.2018.8639557

[26] Y. Wang, M. J. F. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, pp. 47–56, 2018. [Online]. Available: https://doi.org/10.1016/j.specom.2018.09.002

[27] X. Chen, X. Liu, Y. Wang, A. Ragni, J. H. M. Wong, and M. J. F. Gales, "Exploiting future word contexts in neural network language models for speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 9, pp. 1444–1454, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2922048

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2014.

[29] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech & Language*, vol. 25, no. 2, pp. 282–306, 2011.