Deep Density Networks with Uncertainty for spontaneous spoken language assessment

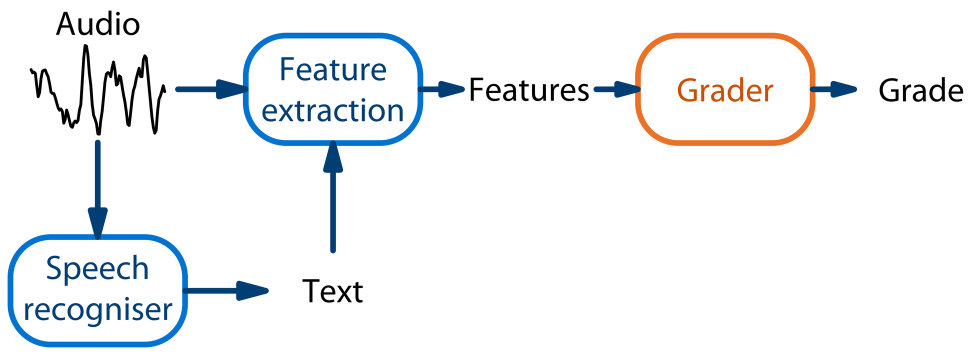
Andrey Malinin, Yu Wang, Kate Knill and Mark Gales
{am969,yw396,kate.knill,mjfg}@eng.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

ALTA Institute / Department of Engineering, University of Cambridge

Introduction

- Many people are learning English → want official qualifications
- To help meet this demand: Automatic assessment of spoken English



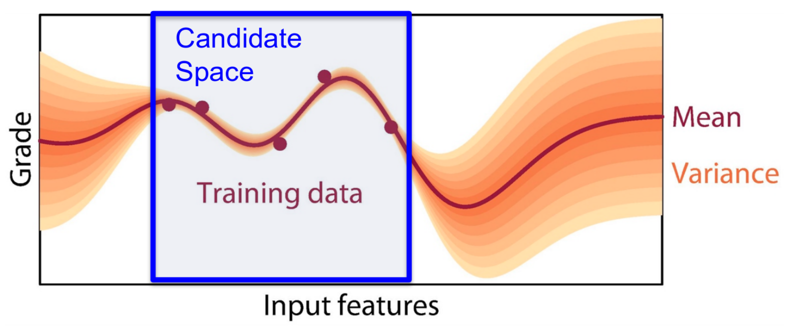
- An automatic grader:
  - is more consistent than human graders
  - has significantly higher throughput
- How to deal with difficult to grade speakers?
- Estimate uncertainty in prediction →
  - Reject speakers with greatest uncertainty to human graders

Uncertainty in Gaussian Processes and DNNs

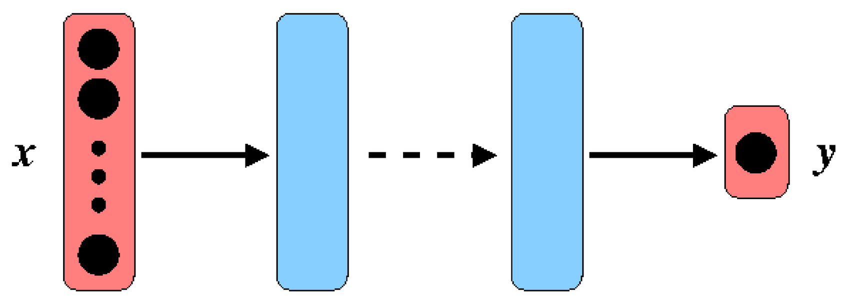
- Principled way of deriving model uncertainty:

$$\mathrm{p}(g|x,\mathcal{D}) = \int \mathrm{p}(g|x,\mathcal{M})\mathrm{p}(\mathcal{M}|\mathcal{D})\mathrm{d}\mathcal{M}$$

- For Gaussian Process can solve integral →



  - Non-parametric Bayesian model: $f_{GP}(\mathbf{x};\mathcal{D}) \to \mu_g(\mathbf{x}), \sigma^2_g(\mathbf{x})$
  - Uncertainty depends on proximity of test data to training data
  - Limitations - $O(n^2)$ memory, $O(n^3)$ compute → use DNNs



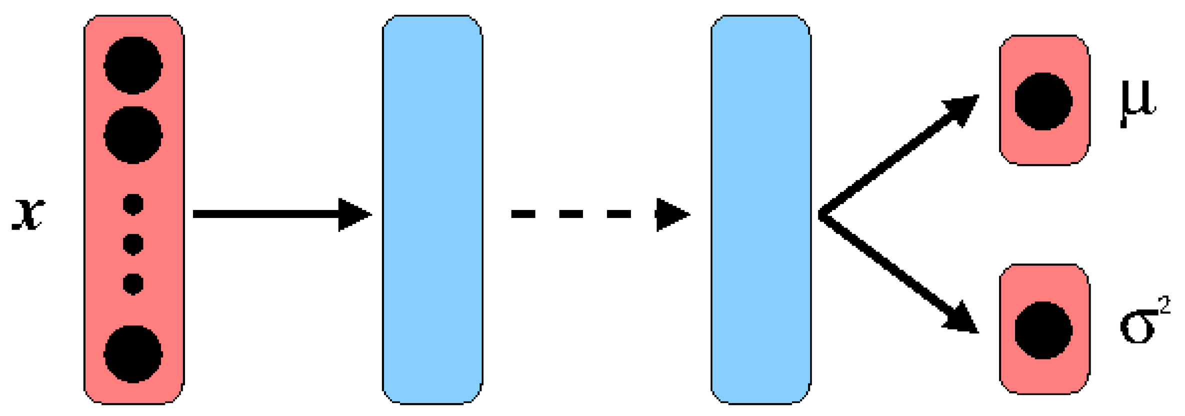
- Parametric model: $f_{DNN}(\mathbf{x};\mathcal{M}) \to \mu_g(\mathbf{x})$
  - Advantages - scalable and flexible architecture
  - Limitation - No natural uncertainty measure →
  - approximate via Monte-Carlo Dropout:

$$\hat{\mu}_g(x) = \frac{1}{N}\sum_{i=1}^{N} f(x;\mathcal{M}^{(i)})$$

$$\hat{\sigma}^2_g(x) = \frac{1}{N}\sum_{i=1}^{N} \left(f(x;\mathcal{M}^{(i)})\right)^2 - \hat{\mu}^2_g(x)$$

  - Prediction uncertainty depends on uncertainty in weights
- Uncertainty measures for both models are implicit

Deep Density Network



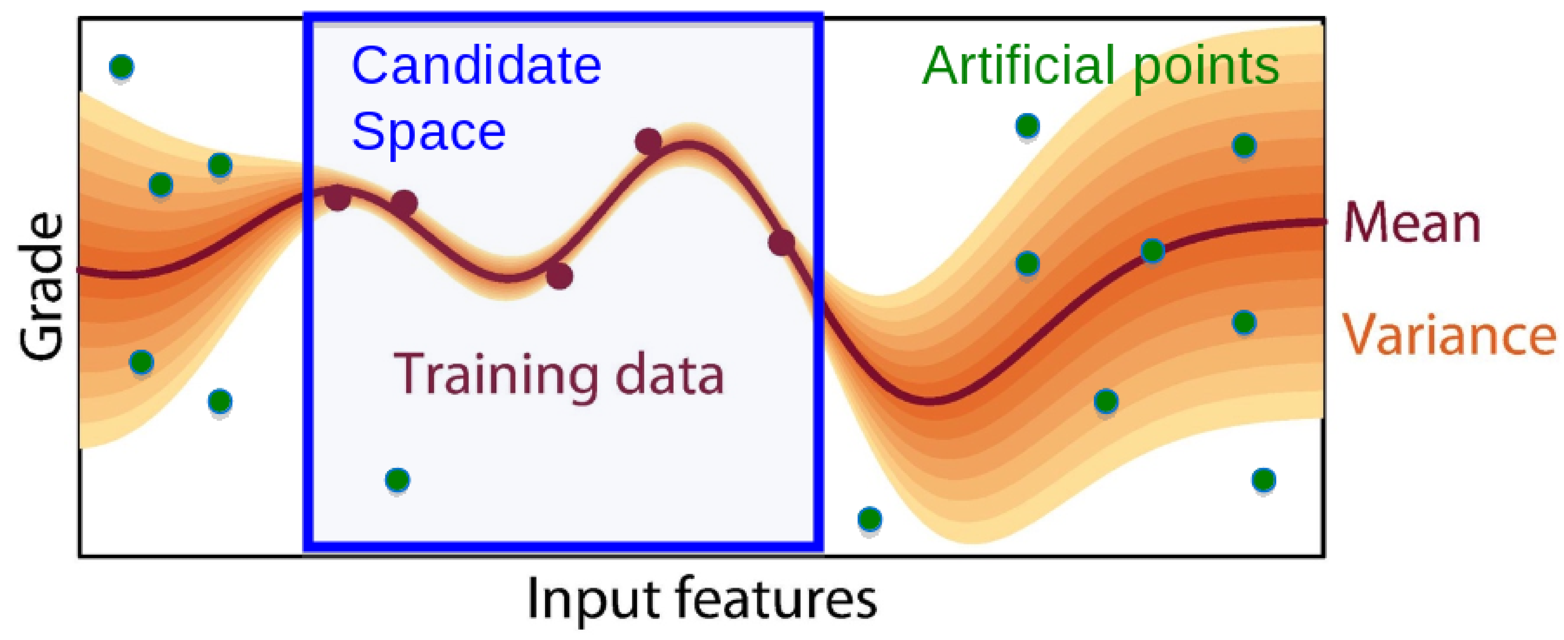
- DDN parametrises a normal distribution $\mathrm{p}(g|x;\mathcal{M})$ over grades.

$$f_{\mu}(\mathbf{x};\mathcal{M}) = \mu_g(\mathbf{x})$$
$$f_{\sigma^2}(\mathbf{x};\mathcal{M}) = \sigma^2_g(\mathbf{x})$$
$$\mathrm{p}(g|\mathbf{x};\mathcal{M}) = \mathcal{N}(g|\mu_g(\mathbf{x}),\sigma^2_g(\mathbf{x}))$$

- Train by maximizing likelihood
- DDN variance represents the spread in grade given input x →
  - Natural noise associated with the data → implicit uncertainty
- Want uncertainty based on similarity to training data
  - Assign uncertainty explicitly!

Deep Density Network with Noise

- Need variance to depend on distance of x from training data
  - Low/High variance near/far from training data
- Solution: Specify variance explicitly
  - Define a low variance empirical distribution p_{D} over real data
  - Define a high-variance artificial data distribution p_{N} (Factor Analysis)
  - Train DDN to model both distributions



- Two stage training process:
  1. Train standard DDN on real data
  2. Continue training DDN in multi-task fashion →
     - Minimize KL divergence of $\mathrm{p}(g|x;\mathcal{M})$ to p_{D} and p_{N}

$$\mathcal{L} = \mathbb{E}_{\hat{\mathbf{x}}}[\mathrm{KL}(\mathrm{p}_{\mathrm{D}}||\mathrm{p}(g|\hat{\mathbf{x}};\mathcal{M})] + \alpha\cdot\mathbb{E}_{\tilde{\mathbf{x}}}[\mathrm{KL}(\mathrm{p}_{\mathrm{N}}||\mathrm{p}(g|\tilde{\mathbf{x}};\mathcal{M})]$$

Evaluation Metrics, Data and Experiments

- Grader Performance Assessment:
  - Pearson Correlation Coefficient (PCC)
  - Mean Squared Error (MSE)
- Useful to have a single value to represent rejection performance
  - Assess using Area Under Curve Rejection Ratio $\mathrm{AUC}_{\mathrm{RR}}$

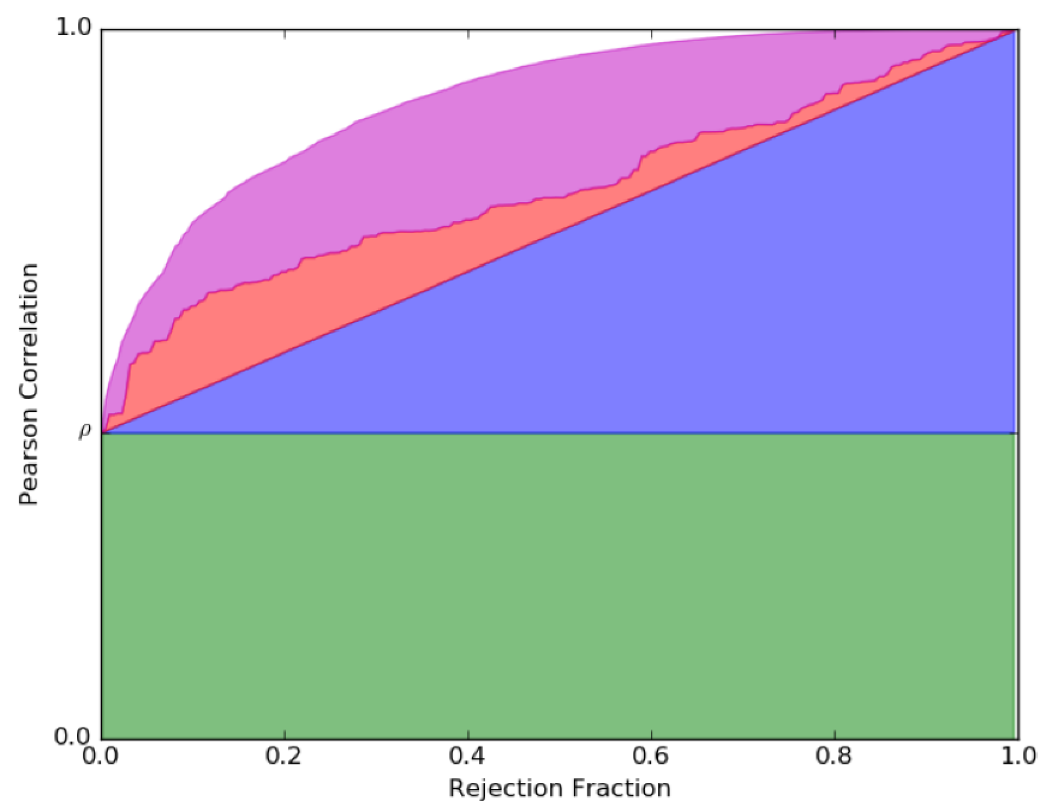
$$AUC_{RR} = \frac{AUC_{var}}{AUC_{max}}$$

$= AUC_{var}$

$+ \quad = AUC_{max}$

AUC_{RR} = 0 = random rejection

AUC_{RR} = 1 = optimal rejection



Experiments
- Acoustic and ASR-derived features from spontaneous responses
- 4300 training and 230 evaluation speakers
- Training on standard grades
- Evaluation on expert grades

Grader	PCC	10% Rej. PCC	AUC	$\mathrm{AUC}_{\mathrm{RR}}$
GP	0.876	0.897	0.942	0.233
$\mathrm{MCD}_{\mathrm{relu}}$	0.879	0.892	0.937	0.040
$\mathrm{MCD}_{\mathrm{tanh}}$	0.865	0.886	0.938	0.226
DDN	0.871	0.887	0.941	0.230
+MT	0.871	0.902	0.947	0.364

Table: Grading and rejection performance

Conclusions

- Novel method for explicitly training DDNs to yield uncertainty estimates
  - Has comparable grading performance as GP and DNN
  - Provides a better uncertainty measure for rejection.
  - Combines essence of GP uncertainty with scalability of DNN
- Future Work → consider advanced methods of generating artificial data