Towards Better Confidence Estimation for Neural Models

Vishal Thanvantri Vasudevan, Abhinav Sethy & Alireza Roshan Ghias

Alexa AI, Amazon

Introduction

- The confidence of a neural network classifier in it's output is typically computed as a function of the softmax posterior probability.
- We consider ensemble diversity and gradient measures to improve confidence calibration.
- We show that the proposed features and confidence prediction model produce a more calibrated confidence score by evaluating a number of metrics.

Ensemble and Gradient Uncertainty Features

Neural network models are high variance learners. Models with the same architecture trained on the same data with different initializations and data sampling order can be viewed as multiple experts, each with a different view of the data. Voting between these models is measure of confidence. Gradient based features are a sign of 're-learning-stress' and can be seen as a measure of the model's uncertainty.

Algorithm 1: Ensemble Features

Input: $P_{\Theta,x} = P_{\theta_1,x}, P_{\theta_2,x}, \ldots, P_{\theta_n,x}$, where $P_{\theta_i,x}$ is the probability distribution over output classes for model with parameters θ_i for input x

Output: $MeanKL_x, VarKL_x$

Procedure:
$meanPD_x \longleftarrow \text{mean}(P_{\Theta,x})$
$KLValues_x \longleftarrow \emptyset$

for i in $1, 2, \ldots n$ **do**
$KLValues_x[i] \longleftarrow \text{KLDivergence}(meanPD_x, P_{\theta_i,x})$
end for
$MeanKL_x \longleftarrow \text{mean}(KLValues_x)$
$VarKL_x \longleftarrow \text{variance}(KLValues_x)$
Return: $MeanKL_x, VarKL_x$

Algorithm 2: Gradient Features

Input: M_θ, x where M_θ is Model with parameters θ and x is the data-point

Output: $GradStats_{\theta,x}$

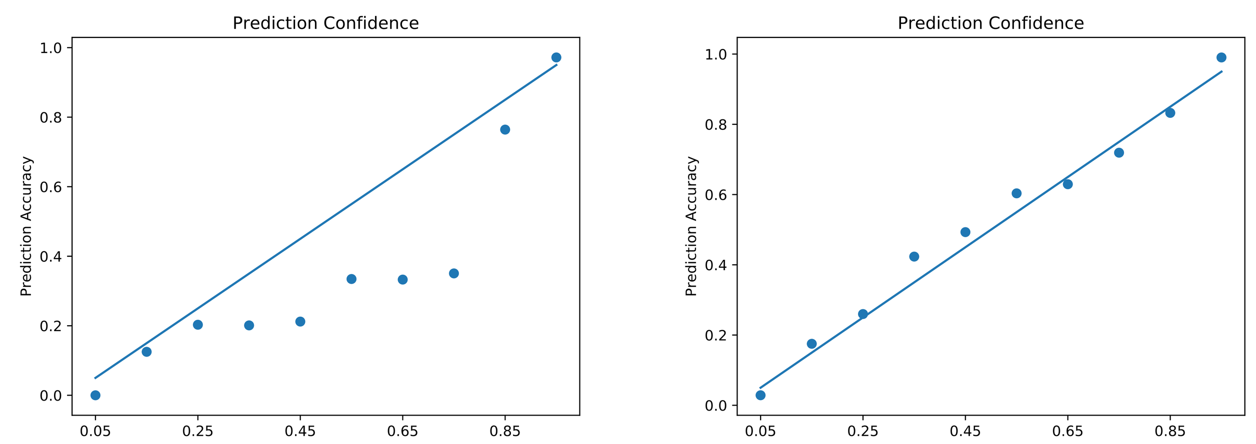
Procedure:
$outputPred_{\theta,x} \longleftarrow M_\theta(x)$
$predClass_{\theta,x} \longleftarrow \text{argmax}(outputPred_{\theta,x})$
$target_{\theta,x} \longleftarrow \text{OneHotEnc}(outputPred_{\theta,x}.size, predClass_{\theta,x})$
$loss \longleftarrow \text{CrossEntropy}(target_{\theta,x}, outputPred_{\theta,x})$
$Grad_{\theta,x} \longleftarrow \text{Gradient}(\theta, loss)$

for $pool$ in $\{max, min, mean, var, sum\}$ **do**
$GradStats_{\theta,x}[pool] \longleftarrow pool(Grad_{\theta,x})$
end for
Return: $GradStats_{\theta,x}$

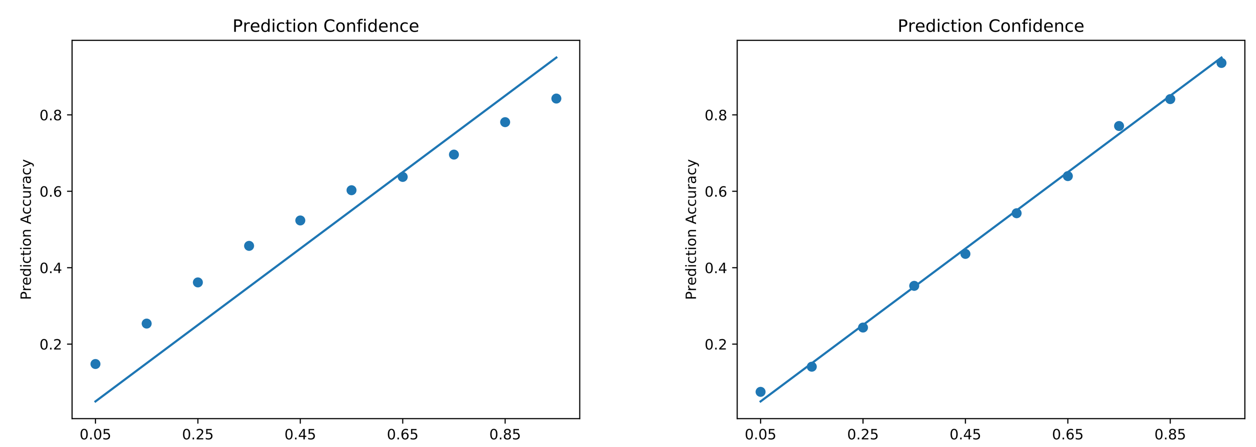
Experimental Setup and Dataset Description

- We tested our approach on three sentence classification tasks and a query rewriting task on subsets of Alexa NLU datasets collected from random users.
- A gradient boosting decision tree (GBDT) regressor model is trained with these features as inputs and instance prediction error as the target.
- For classification tasks, the datasets used were for intent (first party skills) classification, domain classification and skill (third party skills) classification.
- The query rewriting task is a sequence prediction task where we predict a good rewrite of an unsuccessful utterance.

Results



(a): Intent Classification - Uncalibrated Baseline Model (left) and Calibrated GBDT Model (right)



(b): Query Rewriting - Uncalibrated Baseline Model (left) and Calibrated GBDT Model (right)

Figure 1: Reliability diagrams for intent classification and query rewriting tasks. The solid line represents the reliability plot for a perfectly calibrated model

- We compare our proposed approach to the baseline which is the posterior probability.
- The following tables compare the Pearson correlation coefficients and Probability Alignment Score for the confidence scores generated by our confidence models and the baseline model.
- As can be seen from Figure 1, our confidence model is highly calibrated.
- The relative importance of the features used can be observed in Figure 2.

Task	Baseline	GBDT
Intent Classification	0.6500	**0.7782**
Domain Classification	0.6910	**0.7752**
Skill Classification	0.6013	**0.6616**
Query Rewriting	0.4277	**0.5425**

Table 1: Correlation with instance-level accuracy of Baseline vs GBDT model

Task	Baseline	GBDT
Intent Classification	0.8271	**0.8626**
Domain Classification	0.8253	**0.8772**
Skill Classification	**0.7023**	0.6938
Query Rewriting	**0.4006**	0.3967

Table 2: Probability Alignment Score of Baseline vs GBDT model

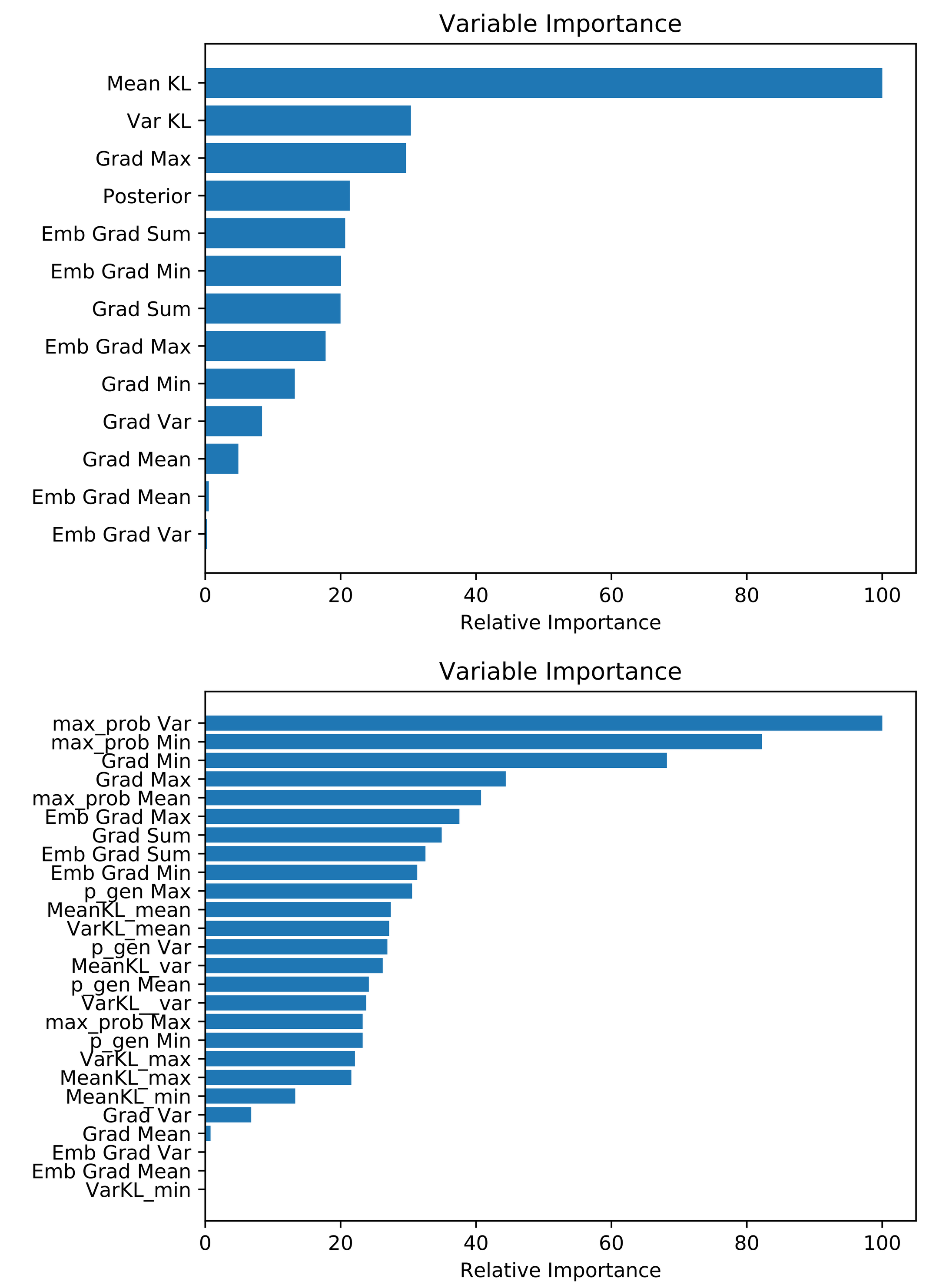


Figure 2: Variable importance as per the regression model for Intent Classification and Query Rewriting tasks

Conclusions and Future Work

- By using ensemble and gradient features to represent uncertainty, our proposed confidence model outperforms the baseline in almost all cases with respect to the evaluation metrics used.
- With minor adaptations, the proposed technique provided improvements on a sequence to sequence query rewriting task as well.
- The ensemble features can be computed much faster by parallelizing the forward pass of each of the models to speed up the algorithm.
- A different avenue to explore would be to alter training schedules and architectures with an additional loss that calibrates posterior probabilities implicitly.