Beyond Hard Decisions: Accounting for Uncertainty in Deep MIR Models

Qingyang (Tom) Xi, Brian McFee
Music and Audio Research Laboratory, New York University

{tom.xi, brian.mcfee}@nyu.edu

Abstract

Deep MIR Models typically produces a **confidence score** with their predictions, but do they really reflect the prediction's **probability of being correct**?

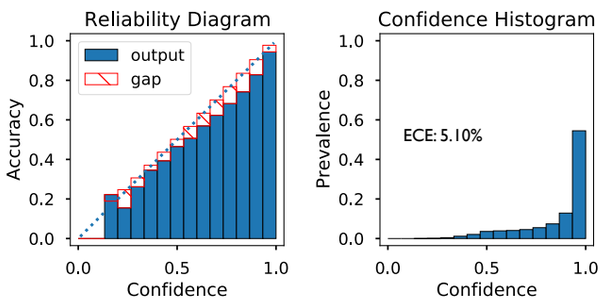
Why do we care?

- To go beyond using just the hard predictions from deep models.
- To interpret confidence scores as probabilities.
- To account for uncertainties in model predictions properly.

What can we do?

- Check confidence scores with **Reliability Diagrams**!
- **Calibrate** model confidence scores!

Measuring Model Reliability: Reliability Diagram and ECE



Reliability Diagram (left) and confidence histogram (right) of the CREMA root predictor on its test set.

Expected Calibration Error (ECE) = < |gap| , prevalence >

ECE is the average size of the reliability gaps per histogram bin, weighted by prevalence. It is a single number summary of the Reliability of the model.

Calibrating Model Confidence: Temperature Calibration

Typical deep MIR model use Soft-Max to turn the logit vector $\mathbf{z}$ into confidence scores $\hat{\mathbf{p}}$:

$$\hat{\mathbf{p}} = \sigma_{\text{SM}}(\mathbf{z}) = \frac{\exp(\mathbf{z})}{\sum_k \exp(\mathbf{z}^{(k)})}$$

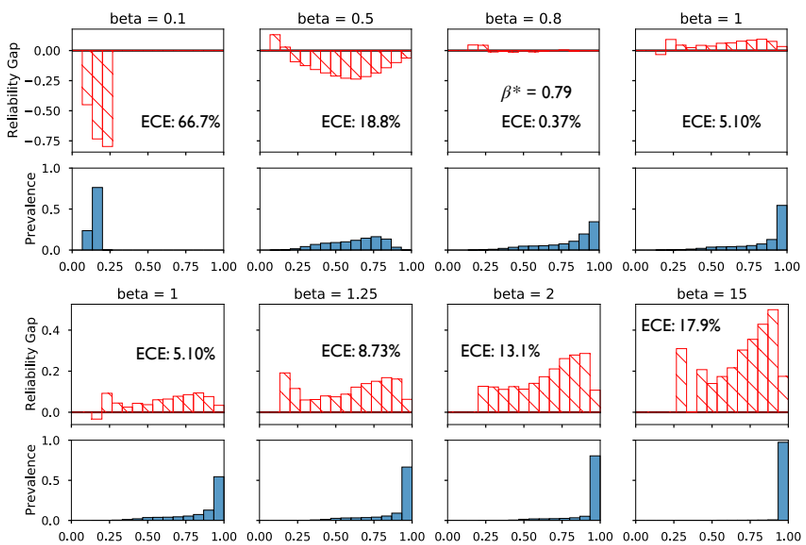
Temperature calibration produce calibrated confidence scores $\hat{\mathbf{q}}$ by scaling $\mathbf{z}$ with a positive constant β^*, determined by minimizing the NLL over a labeled calibration set $(\mathbf{x}_i, y_i)_{i=1}^N \sim \mathcal{D}$.

$$\hat{\mathbf{q}} = \sigma_{\text{SM}}(\beta^* \cdot \mathbf{z}) \qquad \beta^* = \arg\min_{\beta} -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\left(\hat{\mathbf{p}}^{\beta}\right)^{(y_i)}}{\sum_k \left(\hat{\mathbf{p}}^{\beta}\right)^{(k)}}$$

While the logits $\mathbf{z}$ are typically not accessible, we can also calibrate directly from the model confidence output $\hat{\mathbf{p}}$:

$$\hat{\mathbf{q}} = \sigma_{\text{SM}}(\beta \cdot \log \hat{\mathbf{p}}) = \sigma_{\text{SM}}(\log \hat{\mathbf{p}}^{\beta}) = \frac{\hat{\mathbf{p}}^{\beta}}{\Sigma_k \left(\hat{\mathbf{p}}^{\beta}\right)^{(k)}}$$

Effects of different β choices on model reliability:



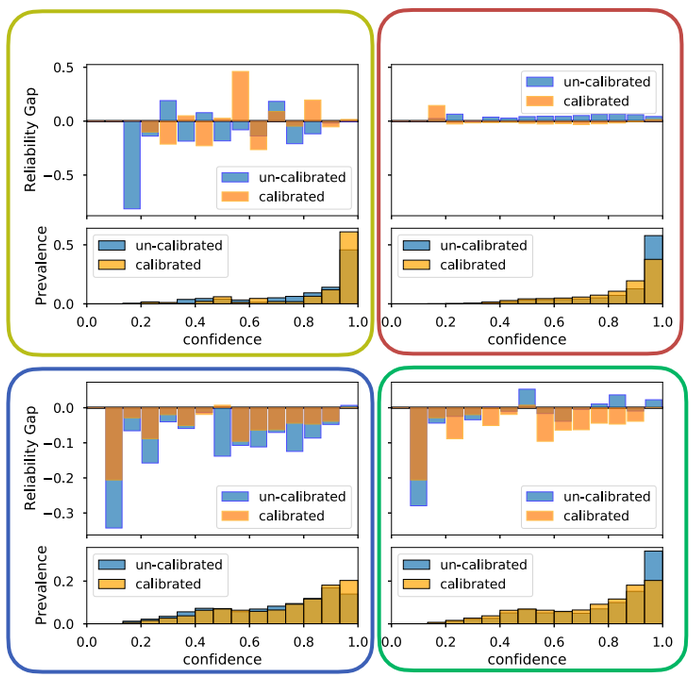
Pilot Experiment: Analyzing the Rock Corpus

By combing the probabilistic confidence outputs from a chord root predictor (CREMA) and a key predictor (MADMOM), we can generate a relative chord root analysis, roman numeral style.

We determine the calibration constants for either model using their labeled calibration set respectively, and evaluate the effect of calibration on this analysis pipeline using the Rock Corpus.

Key \ Root	H	U	C	A
H	23.66%	11.44%	6.18%	13.65%
U	10.16%	4.55%	7.07%	6.74%
C	13.03%	2.21%	3.64%	5.99%
A	17.00%	4.78%	1.59%	0%

Expected Calibration Error (ECE) of the realtive root analysis produced by using one of four outputs (C: calibrated, U: un-calibrated, H: hard decisions, A: annotation) for either the key or the chord root model.



NEW YORK UNIVERSITY