Universität Augsburg
Fakultät für Angewandte
Informatik

Calibrated and Uncertainty-aware Multimodal Emotion Recognition

Nicolas Kolbenschlag

Lukas Stappen, M.Sc.

Bachelor thesis, 2021

Agenda

1 Uncertainty in Deep Learning (Background)

2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research goals)

3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)

4 Evaluation and Discussion (Conclusion)

Uncertainty in Deep Learning

Introduction to Uncertainty

- Imagine a model (e.g. a neural network), which makes **a prediction** $\hat{y}$.
- But **how confident** (or the opposite uncertain) is the model about this prediction?
- Fundamental intuition:
  1. Define true uncertainty.
  2. Measure/predict uncertainty, alongside with the actual prediction $\hat{y}$.
  3. Model is referred to as ***well-calibrated***, if true uncertainty matches predicted uncertainty.
- Usually true uncertainty is defined as the prediction error, so **we want the models uncertainty to be high if its error (in terms of $\hat{y}$) is high** and vice-versa.

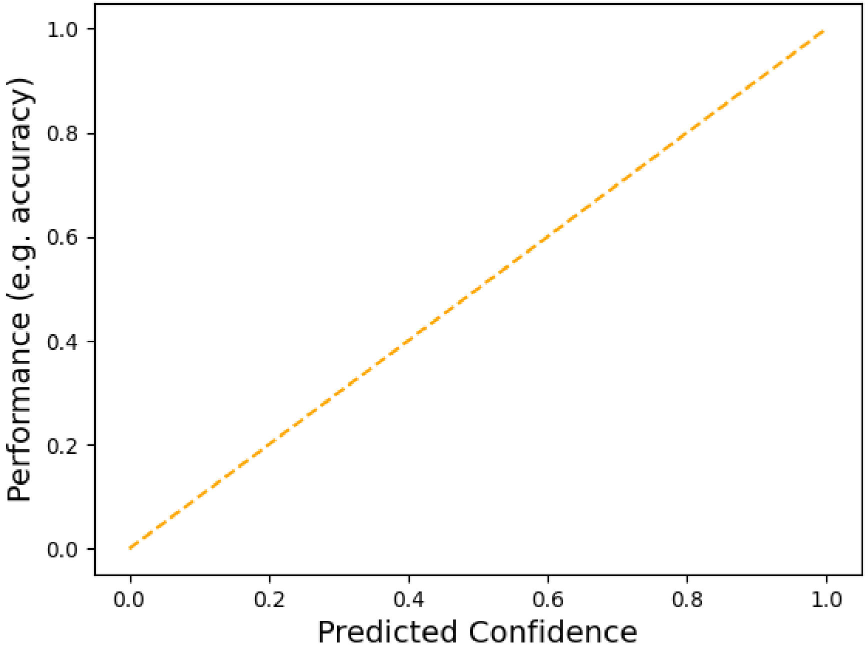


Fig.: *Well-calibrated* uncertainty measurements: the predicted confidence exactly matches the prediction error (here, represented by accuracy).

Uncertainty in Deep Learning

Types and Sources of Uncertainty

- **Aleatory** *(Data)* Uncertainty:
  - Caused by noise in the sample, as it leads to non-determination of an estimation problem.
  - Imagine the landing point of an arrow.
- **Epistemic** *(Model/Knowledge)* Uncertainty:
  - Appears if the model lacks in knowledge on a sample.
  - E.g., the image of a ship, fed into a cat-vs.-dog-classifier.

Uncertainty in Deep Learning

Measuring Predictive Uncertainty

- There are many approaches to obtain measurements of uncertainty from an neural network.
- For classification: most common and easiest approach is to apply softmax activation function on output layer.
- But what about regression?
  - Bayesian modelling:
    - Obtain multiple forecasts for $\hat{y}$ (with possibly different outcomes).
    - Use the mean of them as final prediction and the **variance** among them **as measurement for uncertainty**.
    - Approaches:
      - Monte Carlo Dropout: Enable dropout during inference.
      - Ensemble Averaging: Train multiple model from varying starting point (seeds).
      - …
  - Distributional parameter estimation:
    - Let model output parameters for probability distribution over $\hat{y}$, instead of $\hat{y}$.
- Note: approaches, mentioned for regression are possible for classification, too.

Uncertainty in Deep Learning

Calibration

- Often predicted uncertainty is not well-calibrated initially.
- *(Re-) calibration* describes the procedure of (attempting to) make non well-calibrated uncertainties well-calibrated.
- Therefore:
  1. Obtain non-well calibrated (*uncalibrated*) measurements of uncertainty $\hat{U}$ from a hold-out validation set.
  2. Calculate true uncertainty U of the validation set.
  3. Train a auxiliary model, the *calibrator*, to map $\hat{U}$ to U.
  4. Now, the calibrator can be applied on uncalibrated uncertainty quantification obtained from the test set.
- "Features → neural network → uncalibrated uncertainty → calibrator → calibrated uncertainty"
- Often used as calibrator, *Isotonic Regression*: fit non-decreasing function through validation points and interpolate to cover complete target space.

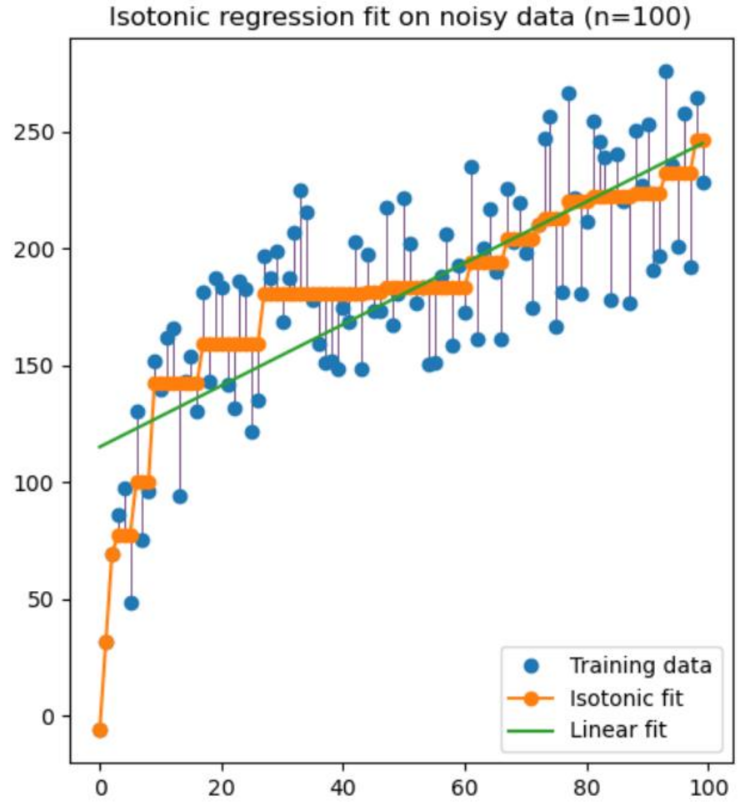


Fig.: Isotonic Regression. Image taken from: https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_isotonic_regression.html

Predictive Uncertainty, Subjectivity and Emotion Recognition

Emotion Recognition and the MuSe-CaR dataset

- **Emotion Recognition**: prediction of the emotional state, here described by valence (sentiment) and arousal (excitement).
- The **MuSe-CaR** dataset:
  - Contains YouTube videos of car reviews contiguously (one label per time step) annotated by valence and arousal.
  - For each sample, there are available **5 annotations by different human annotators**.
  - Further, there exists a *ground-truth* calculated, from the multiple annotations, by fusion techniques.
- Aims of this work:
  - **Measure predictive uncertainty** for the task of emotion recognition with MuSe-CaR.
  - Investigate **subjectivity among annotations** as source of predictive uncertainty.

Predictive Uncertainty, Subjectivity and Emotion Recognition

Before we start: Replacing Variance by Correlation

- We want to use approach like Monte Carlo Dropout to measure uncertainty, but such techniques use variance to obtain the final measurement of confidence, as mentioned before.
- We predict steps in time of a **contiguous** emotional state.
- Each point in time t is rated based on the evaluation of the signal at t and the annotation of the latest time step $t - 1$.
- Therefore, we replace variances by correlations:
  - We obtain multiple (differing) predictions, as usual.
  - We compute the **average correlation** (PCC or CCC) **among each pair** of them.
  - As higher the correlations, as higher the model's confidence and vice-versa.

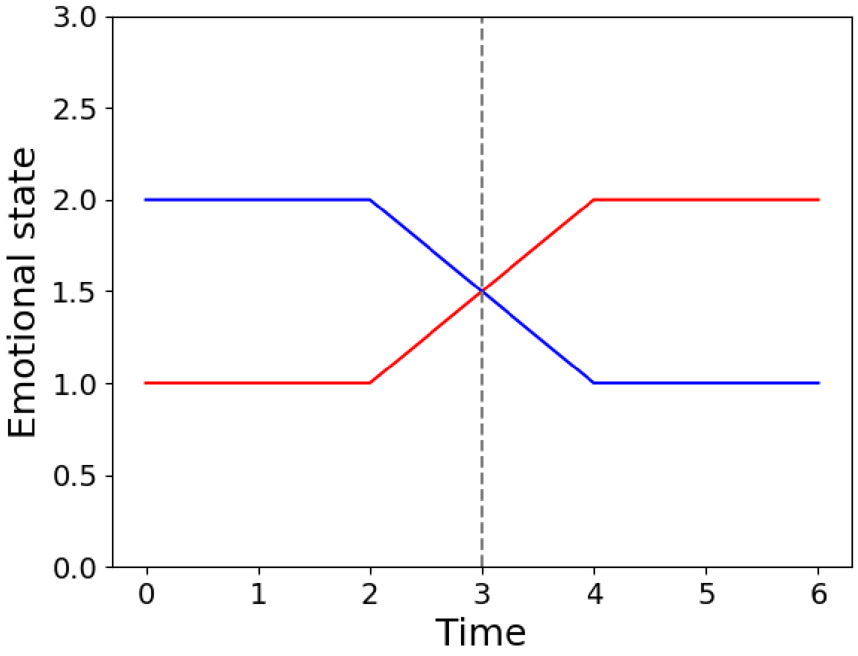


Fig.: At x = 3 both lines have the same value. But when it comes to measuring the disagreement of them, the correlation at this point is much more expressive, especially because the shown values at each step in time come from one contiguous measurement. Each point in time is not a standalone prediction.

Obtaining Estimates of Predictive Uncertainty

Overview over Concepts

- Why investigate subjectivity among annotations?
  - We want uncertainty to appear, where the prediction error is large (common definition).
  - But why should it indeed appear there? Or, why actually is there prediction error, if it is?
  - Whether model nor anybody know the prediction error during inference. So, there must be underlying sources that cause it (refers to *data* and *knowledge* uncertainty).
  - We investigate subjectivity among annotations as indicator for where we expect uncertainty.
  - So, we assume, it is a source of measured uncertainty.
- Two main conceptional approaches:
  - Local uncertainty quantification: for now, anything happens **per time step**, or *locally*.
  - Global uncertainty quantification: measurements are done for **multiple time steps** (sub-samples, respectively) **together**, or *globally*.

Obtaining Estimates of Predictive Uncertainty

Local Uncertainty Quantification: Formalization

True Uncertainty (at time step j)

- Predictive Performance (left figure)

$$P_j = PCC\big(y_{j-t} \dots y_j, \hat{y}_{j-t} \dots \hat{y}_j\big)\ with\ t = 2$$

  - With the prediction for the emotion $\hat{y}_i$ and the label y.

- Subjectivity among raters (right figure)

$$S_j = \frac{1}{10} \sum_{i=1}^{5} \sum_{k=i+1}^{5} PCC(x_{i,j-t} \dots x_{i,j}, x_{k,j-t} \dots x_{k,j-t})\ with\ t = 2$$

  - With x_i, the annotation of the i-th rater.

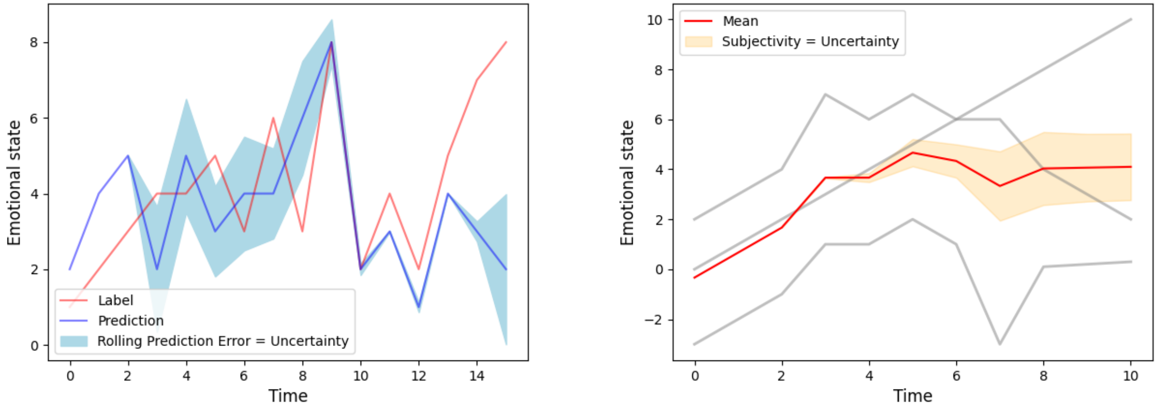


Fig.: Predictive Performance (left) and subjectivity among raters (right).

Predicted Uncertainty (at time step j)

- Monte Carlo Dropout

$$MCD_j = \frac{1}{10} \sum_{i=1}^{5} \sum_{k=i+1}^{5} PCC(\hat{y}_{i,j-t} \dots \hat{y}_{i,j}, \hat{y}_{i,k-t} \dots \hat{y}_{k,j})\ with\ t = 2$$

  - With the model's i-th prediction for the emotional state $\hat{y}_i$.

- *tilted* CCC (tCCC)
  - Network has three output nodes:
    - Mid node: usual CCC loss criterion (used for inference of emotional state).
    - Outer nodes: RCE_3 and RCE_{10} as criterion (n: sample length).

$$RCE_w(\hat{y}) = \frac{1}{n} \sum_{j=w}^{n} (1 - PCC\big(\hat{y}_{j-w+1} \dots \hat{y}_j, y_{j-w+1} \dots y_j\big))$$

    - Creates mini-ensemble with 2 nodes, focusing on different errors (rolling correlation error over the 3 and 10 latest time steps, respectively).
    - Used to obtain uncertainty quantification, by computing correlation between them (similar to P_j).

Obtaining Estimates of Predictive Uncertainty

Local Uncertainty Quantification: Results

- Quantified uncertainty did not match true uncertainty, whether subjectivity nor predictive performance.
- But the reason, that it did not work, was not the unfeasibility of the uncertainty quantification approaches themselves.
- We found a conceptional issue with the emotion recognition itself:
  - With Monte Carlo Dropout, the model is optimized using CCC loss, which evaluates the forecast per complete sample, not per time step.
  - Further, we observed, that during training with tCCC, the outer nodes did not converge significantly.
  - Indicates, that whole emotion recognition (here) works more globally. So the model is capable of predicting longer trends of the emotion state, but not locally (over a few points in time).
  - **If the model's (local) forecasts are prone to be wrong, so meaningless, they naturally lead to non-reliable confidence estimates.**
  - Therefore, we decided to measure uncertainty globally, for multiple time steps together, respectively.

Obtaining Estimates of Predictive Uncertainty

Global Uncertainty Quantification: Formalization

- Instead of measuring one value for uncertainty per time step, we now define uncertainty for complete samples.
- But, as we expect uncertainty to vary over time, we split up samples into non-overlapping sub-samples of length 20 (5 seconds).

True Uncertainty (for one sub-sample)

- Predictive Performance

$$P = CCC(y, \hat{y})$$

  - With label y and prediction $\hat{y}$.

- Subjectivity among raters

$$S = \frac{1}{10} \sum_{i=1}^{5} \sum_{k=i+1}^{5} CCC(x_i, x_k)$$

  - With x_i, the annotation of the i-th rater.

Predicted Uncertainty (for one sub-sample)

- Monte Carlo Dropout

$$MCD = \frac{1}{10} \sum_{i=1}^{5} \sum_{k=i+1}^{5} CCC(\hat{y}_i, \hat{y}_k)$$

  - With the model's i-th prediction for the emotional state $\hat{y}_i$.

- Ensemble Averaging
  - Exact the same, as Monte Carlo Dropout, only with forecasts $\hat{y}_i$ from five different models/seeds, instead of five forward runs.

Obtaining Estimates of Predictive Uncertainty

Global Uncertainty Quantification: Results

Evaluation and Discussion

Summary

Evaluation and Discussion

Conclusion

- TODO: subjectivity is aleatory; OOD features are epistemic

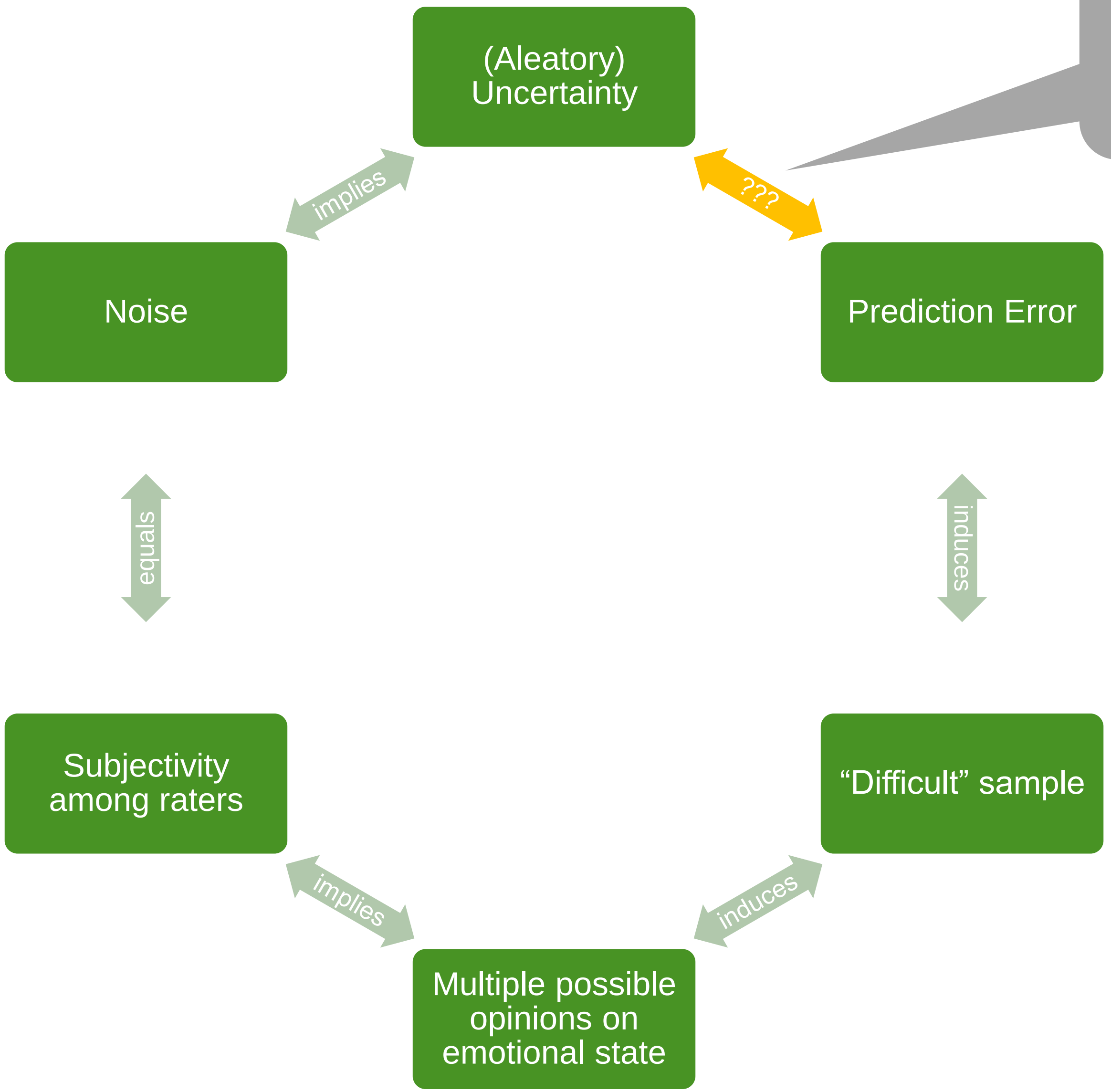
Universität Augsburg
Fakultät für Angewandte
Informatik

Thank‘s for your attention!

Any questions?

Nicolas Kolbenschlag
Department of Computer Science
Universität Augsburg
nicolas.kolbenschlag@student.uni-augsburg.de
www.uni-augsburg.de

Backlog



It is our research question to check, whether this equation holds or not.