Evaluating Machine Learned Models for their Predictive Uncertainty in Context

Eric Heim, Senior Machine Learning Research Scientist
AI Division, Software Engineering Institute, Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM22-0248

Evaluating Machine Learned Models for their Predictive Uncertainty in Context

Eric Heim, Senior Machine Learning Research Scientist
AI Division, Software Engineering Institute, Carnegie Mellon University

An Argument for Uncertainty in Machine Learned Models

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

An Argument for Uncertainty in Machine Learned Models

Predicted Class
Forest

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

An Argument for Uncertainty in Machine Learned Models

Predicted Class	Confidence
Forest	0.95

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

An Argument for Uncertainty in Machine Learned Models

Predicted Class	Confidence
Forest	0.50

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Forest	0.50
Flooded Forest	0.46
Flooded Plain	0.02
Plain	0.01
…	…

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Forest	0.50
Flooded Forest	0.46
Flooded Plain	0.02
Plain	0.01
…	…

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

Confidence/Uncertainty can lead to more informed decision-making.

An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Forest	0.50
Flooded Forest	0.46
Flooded Plain	0.02
Plain	0.01
…	…

```
if conf("Flooded Forest") > 0.25
```

Alert!

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Forest	0.50
Flooded Forest	0.46
Flooded Plain	0.02
Plain	0.01
…	…

Why is the model uncertain about this instance?

Uncertainty can be used to identify how to improve models.

https://www.star.nesdis.noaa.gov/jpss/images/Paraguay/Paraguay_FalsecolorImage_June30_2014_Match_BWImage.png

What is Context-Specific Evaluation and Why?

Context-specific evaluation of a machine learned model quantitatively measures a performance characteristic of a model in a way that is meant to reflect important conditions of the desired deployment environment of the model, as well as cases that have been identified as important

Evaluate the model according to where it's deployed, how it's used, and against important cases.

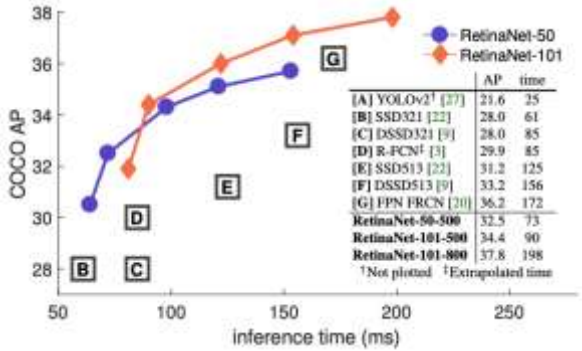
Why evaluate machine learned models in context?

1. Strong theoretical guarantees for modern ML models are rare and hard to directly apply to your application [Kawaguchi, Kaelbling, Bengio; 2020]

It is very difficult to make a detailed characterization of how well a specific hypotheses generated by a certain learning algorithm will generalize, in the absence of detailed information about the given problem instance

2. Most evaluations from the ML literature are on benchmark data sets and intentionally general metrics [Lin et al; 2018]

Without strong theoretic guarantees or directly-applicable empirical evidence provided by the academic community, it is difficult to tell how a model will perform for a specific application.



(Some) ML Contexts for Uncertainty

Uncertainty…

1. …as a component of analysis [Garg et al; 2020]
2. …that enables (efficient/safe/etc.) learning
   - Active Learning [Mukhoti et al; 2021]
   - Reinforcement Learning [Malik et al; 2019]
3. …that feeds other software in a larger AI System

Computer vision is just the first step in the autonomous vehicle data pipeline. The car incorporates data from its many cameras, identifies important elements of its surrounding environment, and fuses these elements with data from radar and lidar.

David Silver, Senior Software Engineer, Cruise
https://www.linkedin.com/pulse/how-computer-vision-works-self-driving-cars-david-silver

4. …that enables better decision-making
5. …that identifies potential model failures in important cases

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Classifier Calibration: The task of ensuring the probabilistic outputs of classifiers match the true likelihood of events.

Top-1 Reliability (Calibration) Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Classifier Calibration: The task of ensuring the probabilistic outputs of classifiers match the true likelihood of events.

Top-1 Reliability (Calibration) Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

For all possible inputs X such that the classifier outputs α as the highest probability for a class…

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Classifier Calibration: The task of ensuring the probabilistic outputs of classifiers match the true likelihood of events.

Top-1 Reliability (Calibration) Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

For all possible inputs X such that the classifier outputs α as the highest probability for a class…

…the probability of that class…

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Classifier Calibration: The task of ensuring the probabilistic outputs of classifiers match the true likelihood of events.

Top-1 Reliability (Calibration) Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

For all possible inputs X such that the classifier outputs α as the highest probability for a class…

…the probability of that class…

…should be equal to α

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Classifier Calibration: The task of ensuring the probabilistic outputs of classifiers match the true likelihood of events.

Top-1 Reliability (Calibration) Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

For all possible inputs X such that the classifier outputs 0.6 as the highest probability for a class…

…60% of the time the classifier should be correct

$$g(x_1) = [0.25, 0.10, 0.60, 0.05]$$
$$g(x_2) = [0.60, 0.30, 0.10, 0.10]$$
$$\ldots$$

} 60% of these should be correct

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Classifier Calibration: The task of ensuring the probabilistic outputs of classifiers match the true likelihood of events.

Top-1 Reliability (Calibration) Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

Other interpretations of classifier outputs are more appropriate for more focused evaluation or to facilitate decision making in different contexts.

Class	Confidence
Forest	0.50
Flooded Forest	0.46
…	…

VS.

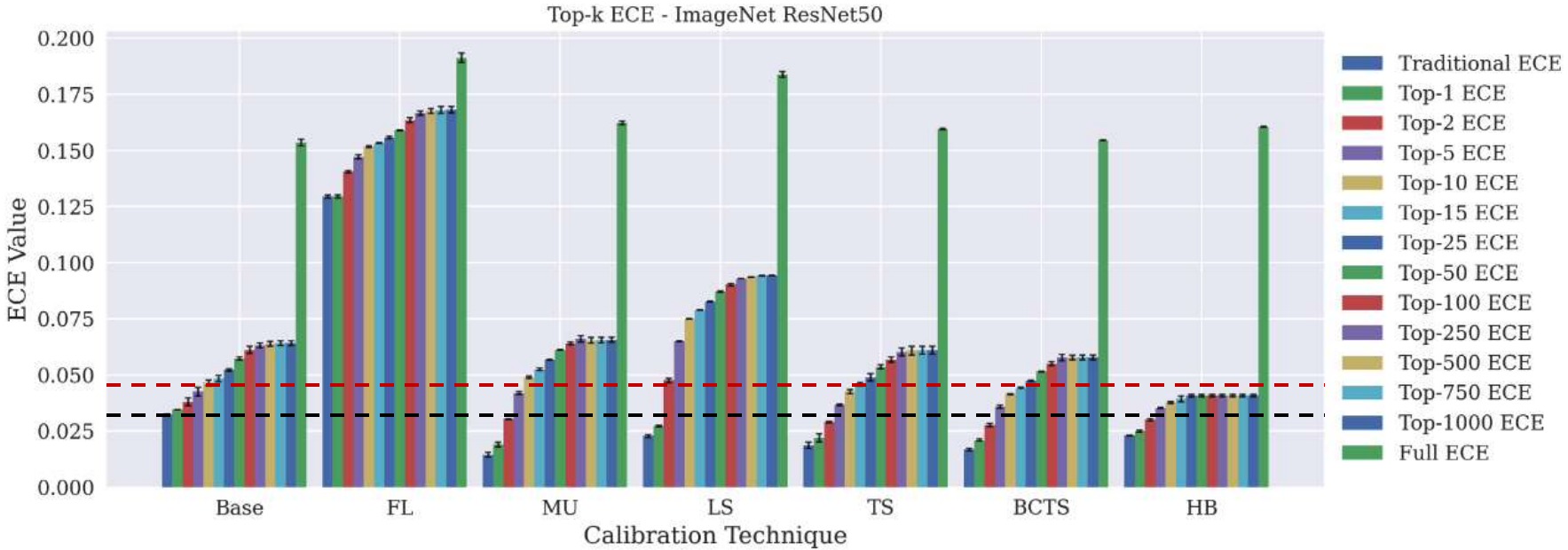
Confidence in disease
Very Low
Low
Medium
High
Very High

Evaluation Metrics for Context-Specific Calibration

(Kirchenbauer, Oaks, and Heim; Under Review)

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

Experiment #1: Top-k ECE



Main Takeaways

- Because of the lack of directly applicable theoretical and empirical results currently available, ML models must be **evaluated in the context** in which they are going to be used.
- **Uncertainty** is an important characteristic of ML models that can facilitate **reliable and robust** usage.
- Despite the number of contexts in which uncertainty is used, most work focuses on a standard definition of reliability (calibration) as a basis for evaluation metrics.
- Our work:
  - Developed a modular framework for measuring context-specific calibration
  - Showed that many popular calibration techniques behave inconsistently depending on the metric used to evaluate them.

etheim@sei.cmu.edu