Contribution ID: **19** Type: **not specified**

Tackling the Challenge of Uncertainty Estimation and Robustness to Distributional Shift in Real-World applications

Friday 26 November 2021 16:30 (40 minutes)

While much research has been done on developing methods for improving robustness to distributional shift and uncertainty estimation, most of these methods were developed only for small-scale regression or image classification tasks. Limited work has examined developing standard datasets and benchmarks for assessing these approaches. Furthermore, many tasks of practical interest have different modalities, such as tabular data, audio, text, or sensor data, which offer significant challenges involving regression and discrete or continuous structured prediction. In this work, we propose the Shifts Dataset for evaluation of uncertainty estimates and robustness to distributional shift. The dataset, which has been collected from industrial sources and services, is composed of three tasks, with each corresponding to a particular data modality: tabular weather prediction, machine translation, and self-driving car (SDC) vehicle motion prediction. All of these data modalities and tasks are affected by real, 'in-the-wild' distributional shifts and pose interesting challenges with respect to uncertainty estimation. We hope that this dataset will enable researchers to meaningfully evaluate the plethora of recently developed uncertainty quantification methods, assessment criteria and baselines, and accelerate the development of safe and reliable machine learning in real-world risk-critical applications.

An additional challenge to uncertainty estimation in real world tasks is that standard approaches, such as model ensembles, are computationally expensive. Ensemble Distribution Distillation (EnDD) is an approach that allows a single model to efficiently capture both the predictive performance and uncertainty estimates of an ensemble. Although theoretically principled, this work shows that the original Dirichlet log-likelihood criterion for EnDD exhibits poor convergence when applied to large-scale tasks where the number of classes is very high. Specifically, we show that in such conditions the original criterion focuses on the distribution of the ensemble tail-class probabilities rather than the probability of the correct and closely related classes. We propose a new training objective which resolves the gradient issues of EnDD and enables its application to tasks with many classes, as we demonstrate on the ImageNet, LibriSpeech, and WMT17 En-De datasets containing 1000, 5000, and 40,000 classes, respectively.

Presenter: ANDREY MALININ