CASC
Center for Applied
Scientific Computing

Uncertainty Quantification in Scientific ML

Project: ASCR SciML Project

Author(s):	*Role:*
Jayaraman J. Thiagarajan	PI
Rushil Anirudh	Team Member
Peer-Timo Bremer	Team Member
Bindya Venkatesh	Student Intern
Uday Shanthamallu	Student Intern

Lawrence Livermore
National Laboratory

Contents

1 Project Overview

The intricate interactions between data sampling, model selection and the inherent randomness in complex systems strongly emphasize the need for a rigorous characterization of ML algorithms. In conventional statistics, uncertainty quantification (UQ) provides this characterization by measuring how accurately a model reflects the physical reality and by studying the impact of different error sources on the prediction. Consequently, there is a strong need to utilize prediction uncertainties in deep models to shed light onto when and how much to trust the predictions. These uncertainty estimates can also be used for enabling safe ML practice, e.g., identifying out-of-distribution samples, detecting anomalies/outliers, delegating high-risk predictions to experts, defending against adversarial attacks etc.

Broadly, a rigorous statistical characterization of ML systems will enable us to:

- build reliable models – consistency between predictions and our understanding of the world.
- incorporate real-world priors.
- avoid machines from being overly confident even when making mistakes.
- identify regimes of strengths and weaknesses.
- design human-in-the-loop systems.

In this project, we have made crucial advances to the fundamental problem of reliably and scalably estimating uncertainties in deep neural networks [1, 2]. In addition to designing state-of-the-art UQ estimation methods, we also made an important finding that the notion of *prediction calibration* can be used to design new loss functions for optimizing deep neural networks and obtained significantly improved models in scientific problems [3, 4, 5, 6]. Finally, we explored novel applications of uncertainties in inverse modeling [7], active learning [8], transferring models under distribution shifts [9] and achieving robustness to adversarial attacks [10]. Results from this project were recently covered as a feature article [11] and discussed in a DataSkeptic podcast [12].

2 Designing Deep Uncertainty Estimators

A natural strategy to produce calibrated predictors is to directly optimize for prediction intervals that satisfy the calibration objective. For example, in the heteroscedastic regression approach, the variance estimates are obtained using the Gaussian likelihood objective, under a heteroscedastic prior assumption. However, by not explicitly constructing the intervals based on epistemic (model variability) or aleatoric (inherent stochasticity) uncertainties, it is not straightforward to interpret the variances from a heteroscedastic model, even when they are well calibrated. On the other hand, approaches designed to capture specific sources of uncertainties, e.g. Monte Carlo dropout for epistemic or conditional quantile based aleatoric uncertainties, are found to be poorly calibrated in practice. Hence, a typical workaround is to employ a separate recalibration step that adjusts the estimates from a trained model to achieve calibration. However, it has been found that even uninformative (random) interval estimates can be effectively recalibrated, thus rendering the estimates meaningless for subsequent analysis.

2.1 Obtaining Calibrated Uncertainties

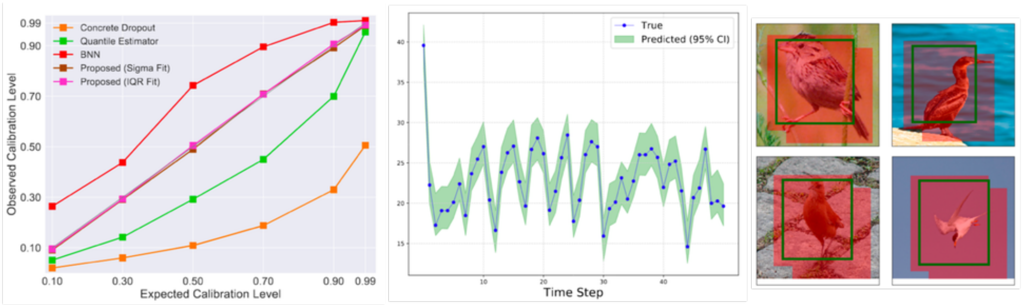


Figure 1: (a) **Regression**: While dropout-based methods are highly optimistic (very sharp intervals), Bayesian neural nets are often pessimistic and both are poorly calibrated. In comparison, our approach is well calibrated. (b) **Time-series forecasting**: Prediction intervals are critical in forecasting and our uncertainty matching technique produces meaningful intervals. (c) **Object Localization**: In vision tasks, the intervals need to be semantically meaningful for easy interpretation.

In our recent paper [1], we conjectured that one can reliably build calibrated deep models by posing calibration as an auxiliary task and utilizing a novel uncertainty matching strategy. To this end, our approach employs two separate models – one for predicting the target and the other for estimating the prediction intervals, and pose a bi-level optimization formulation that allows the mean estimator to identify prediction uncertainties that are the most informative for matching the intervals from the interval estimator. Experiments with different use-cases and model architectures, including regression with FCNs, time-series forecasting with LSTMs and object localization with CNNs, show that our approach consistently produces well calibrated uncertainties (both epistemic/aleatoric) and improved generalization (Figure 1).

2.2 Heteroscedastic Calibration

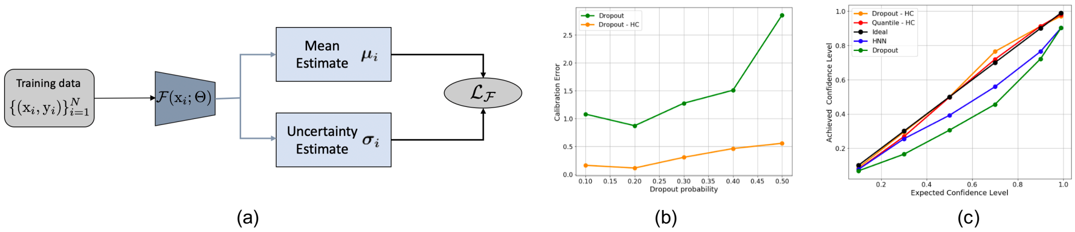
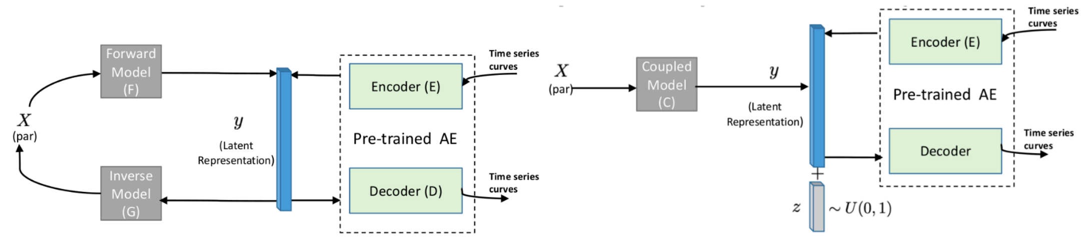


Figure 2: (a) **Heteroscedastic calibration** is a generic technique that can be applied to any existing uncertainty estimation to produce calibrated prediction intervals; (b) In the case of MC dropout, the resulting estimators are fairly well calibrated under a wide-range of dropout rates; (c) With both epistemic and aleatoric uncertainty estimation, our approach ourperforms existing methods.

Though a large class of methods exists for measuring deep uncertainties, in practice, the resulting estimates are found to be poorly calibrated, thus making it challenging to translate them into actionable insights. In [2] we proposed to repurpose the heteroscedastic regression objective as a surrogate for calibration, and enable any existing uncertainty estimator to produce inherently calibrated intervals. In other words, with this single-shot calibration approach, the uncertainty estimates are used in lieu of the heteroscedastic variances to compute the Gaussian likelihood. By performing calibration automatically in the training process based on an explicit uncertainty estimator, our approach does not suffer the limitations of recalibration methods and can be associated to specific error sources unlike classical heteroscedastic networks. Surprisingly, as showed in Figure 2, our approach is able to achieve significantly improved calibration with both an epistemic (MC dropout) and an aleatoric (quantile-based) uncertainty estimator, though they are known to be produce miscalibrated intervals.

2.3 Application: History Matching with Black-box Simulators

In a wide-range of applications in science and engineering, one often faces the need to learn complex mappings between independent parameters and dependent/measured quantities, i.e. the *forward* and *inverse* mappings. Building reliable inverse maps characterizing the conditional posteriors is challenging since in practice the mapping is seldom bijective, and it is challenging to incorporate scientific priors into the learning process. In [7], we showed that enforcing self-consistency between forward and inverse models is an effective regularizer for learning predictive models in



(a) Cyclical Training (b) Coupled Training

Figure 3: Proposed training strategies for building deep inverse models in history matching. In both approaches, we incorporate epistemic uncertainty estimation and optimize using a heteroscedastic calibration strategy.

scientific applications. In particular, we developed two different strategies to enforce self-consistency, namely cylical and coupled training methods (Figure 3). While implementing both approaches, we incorporated epistemic uncertainty estimation and optimized the model parameters using the heteroscedastic calibration strategy from the previous section.

3 Calibration-Driven Learning

Building functional relationships between a collection of observed input variables $\mathbf{x} = \{x_1, \cdots, x_d\}$ and a response variable $\mathbf{y}$ is a central problem in scientific applications – examples range from estimating the future state of a molecular dynamics simulation to searching for exotic particles in high-energy physics and detecting the likelihood of disease progression in a patient. Emulating complex scientific processes using computationally efficient predictive models can achieve significant speed-ups over traditional numerical simulators or conducting actual experiments, and more importantly provides surrogates for improving the subsequent analysis steps such as inverse modeling, experiment design, etc. Commonly referred to as supervised learning in the machine learning literature, the goal here is to infer the function $f : \mathbf{x} \mapsto \mathbf{y}$ using a training sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, such that the expected discrepancy between $\mathbf{y}$ and $f(\mathbf{x})$, typically measured using a loss function $\mathcal{L}(\mathbf{y}, f(\mathbf{x}))$, is minimized over the joint distribution $p(\mathbf{x}, \mathbf{y})$. Despite the importance of $\mathcal{L}$ in determining the fidelity of f, in practice, simple metrics, such as the ℓ_2-metric, $||\mathbf{y} - f(\mathbf{x})||_2$, are used, mostly for convenience but also due to lack of priors on the distribution of residuals. However, this disregards the inherent characteristics of the training data and more importantly the fact that choosing a metric implicitly defines a prior for n. Yet appropriately accounting for noise is crucial to robustly estimate f and to create high-fidelity predictions for unseen data. However, this assumption can be easily violated in real-world data. For example, the ℓ_2 metric is known to be susceptible

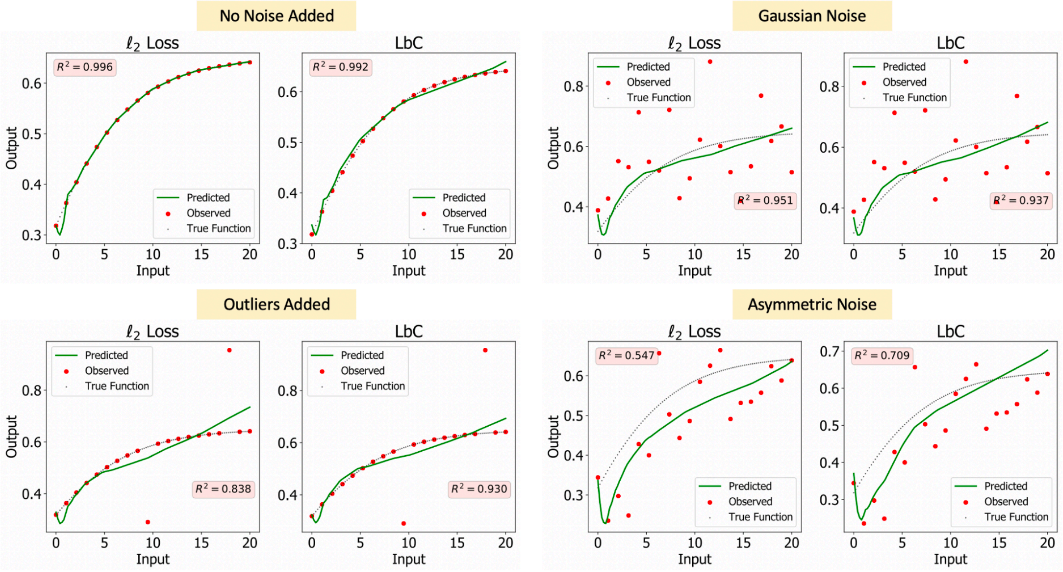


Figure 4: Comparing LbC with ℓ_2 using a synthetic example with a non-linear function. While ℓ_2 is found to be highly effective when there is no noise in the data and the underlying noise process in Gaussian, LbC is consistently superior when there are outliers or asymmetric noise components in the data.

to outliers.

3.1 Learn by Calibrating

As part of this project [3, 5], we have developed Learn-by-Calibrating (LbC), a non-parametric approach based on interval calibration for building emulators in scientific applications, that are effective even with heterogeneous data and are robust to outliers. Though calibration has been conventionally used for evaluating and correcting uncertainty estimators, this paper advocates for utilizing calibration as a training objective in regression models. More specifically, LbC uses two separate modules, implemented as neural networks, to produce point estimates and intervals respectively for the response variable, and poses a bi-level optimization problem to solve for the parameters of both the networks. This eliminates the need to construct priors on the expected residual structure and makes it applicable to both homogeneous and heterogeneous data. Figure 4 provides an illustration of a simple $1-$D regression experiment using a single layer neural network with 100 neurons and ReLU (rectified linear units) non-linear activation. We find that LbC is consistently superior to the widely adopted ℓ_2 loss function, under both symmetric and asymmetric noise models, as well as in the presence of outliers.

3.2 Application: Emulators for Scientific Processes

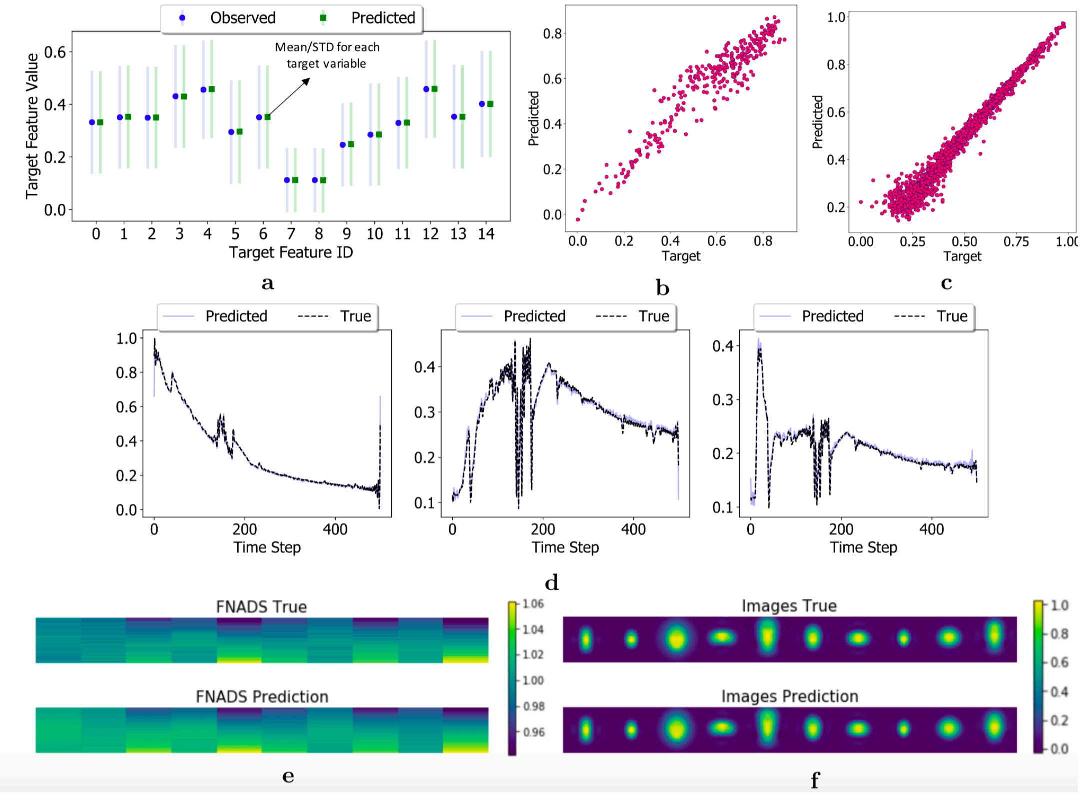


Figure 5: **Qualitative Evaluation of LbC Predictions.** Predictions obtained using LbC on different use-cases: **a** ICF JAG - We show the distribution of values for each of the target variables and the corresponding predictions; **b** Airfoil self-noise; **c** Electric grid; **d** Reservoir model - reconstructions from the decoder; **e-f** FNADS and image predictions from the decoder for ICF Hydra. Across benchmarks of varying dimensionality and complexity, LbC produces high-fidelity emulators that can be reliably used in scientific workflows.

We evaluated the proposed approach using a large suite of use-cases, which require the design of accurate emulators for the underlying scientific processes (see Figure 5. These benchmarks represent a broad range of real-world scenarios including different sample sizes, varying input dimensionality and the need to handle response variable types ranging from single/multiple scalar quantities and multi-variate time-series measurements to multi-modal outputs. Our empirical studies clearly demonstrate the effectiveness of calibration-based training in inferring high-fidelity functional approximations to complex scientific processes. We find that LbC is a simple, yet powerful, approach to design emulators that are robust, reflect the inherent data characteristics, generalize well to unseen samples and reliably replace accurate (expensive) simulators in scientific workflows.

3.3 Application: Clinical Predictive Models

Artificial intelligence (AI) techniques such as deep learning have achieved unprecedented success with critical decision-making, from diagnosing diseases to prescribing treatments, in healthcare. However, to prioritize patient safety, one must ensure such methods are accurate and reliable. For example, a neural network model can produce highly concentrated softmax probabilities – suggesting a reliable class assignment – even for out-of-distribution test samples, which indicates that the confidences are not well-calibrated. This strongly emphasizes the need to both reliably assess model's confidences, and enable rigorous introspection of model behavior.

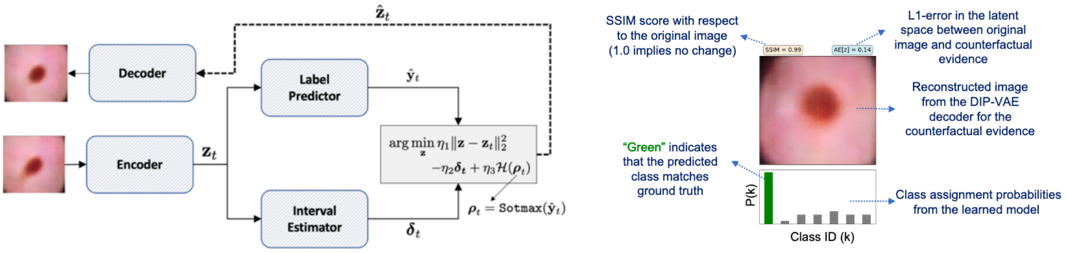

$\hat{\mathbf{z}}_t$
Decoder
Label Predictor
$\hat{\mathbf{y}}_t$
Encoder
$\mathbf{z}_t$
$\arg\min_{\mathbf{z}} \eta_1 \|\mathbf{z} - \mathbf{z}_t\|_2^2 - \eta_2 \boldsymbol{\delta}_t + \eta_3 \mathcal{H}(\boldsymbol{\rho}_t)$
Interval Estimator
$\boldsymbol{\delta}_t$
$\boldsymbol{\rho}_t = \texttt{Sotmax}(\hat{\mathbf{y}}_t)$
SSIM score with respect to the original image (1.0 implies no change)
SSIM = 0.99
AE[z] = 0.14
L1-error in the latent space between original image and counterfactual evidence
Reconstructed image from the DIP-VAE decoder for the counterfactual evidence
"Green" indicates that the predicted class matches ground truth
P(k)
Class assignment probabilities from the learned model
Class ID (k)

Figure 6: Proposed approach for model introspection via prediction calibration - Illustration of the optimization strategy for generating counterfactual evidences (left); Description of components in the visual layout used for showing our results.

We generalized the LbC method [4, 6], originally designed for regression problems, to classification tasks and successfully used it to build highly reliable models. In this process, we also introduced *reliability plots*, which quantify the trade-off between model autonomy and improved generalization by including experts in the loop during inference, as a holistic evaluation mechanism of model reliability. By deferring the "most uncertain" samples to the expert, we obtained a more realistic evaluation of clinical models. Furthermore, we developed a novel interpretability technique that enables us to rigorously explore model behavior (local) via counterfactual evidences generated in a disentangled latent space through prediction calibration (see Figure 6). This technique has been recent utilized in the study of CXR images from COVID-19 patients and interesting insights were obtained [11].

4 Active Learning using Calibrated Predictions

The superior performance of data-driven methods, including deep learning, comes at the price of requiring large amounts of labeled data. This can be a critical bottleneck in applications involving time-consuming data acquisition or high labeling costs. Furthermore, fully supervised methods assume access to samples representing the entire data distribution beforehand, thus making it challenging to handle changes in data distribution over time or adapt the learned model when diverse samples are incrementally included into the training process. This has motivated the use of *active* learning techniques that involve humans in the training loop to build predictive models that are more data-efficient.

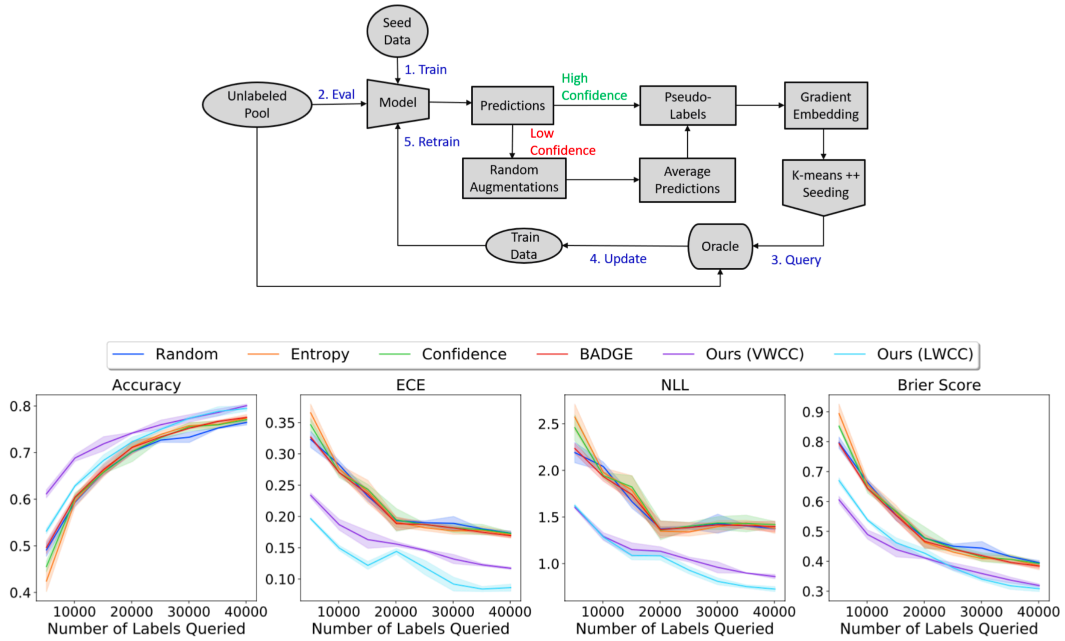


Figure 7: **Ask-n-Learn** utilizes reliable gradient representation obtained via calibrated classifier models and a data-augmentation strategy for reducing confirmation bias. In this example with the CIFAR-10 dataset, our approach provides significant performance gains at small sample regimes.

In this project, we developed a new active learning framework, *Ask-n-Learn*, which addresses the inherent limitations with existing uncertainty-based AL methods (see Figure 7). Our approach uses gradient embeddings for selecting samples, and utilizes calibrated uncertainties to produce reliable gradient embeddings, and a data augmentation strategy for avoiding confirmation bias during pseudo-labeling. Using benchmark image classification tasks in the vision community, we find that the proposed approach significantly outperforms existing active learning approaches.

5 Improving Model Transferability under Distribution Shifts

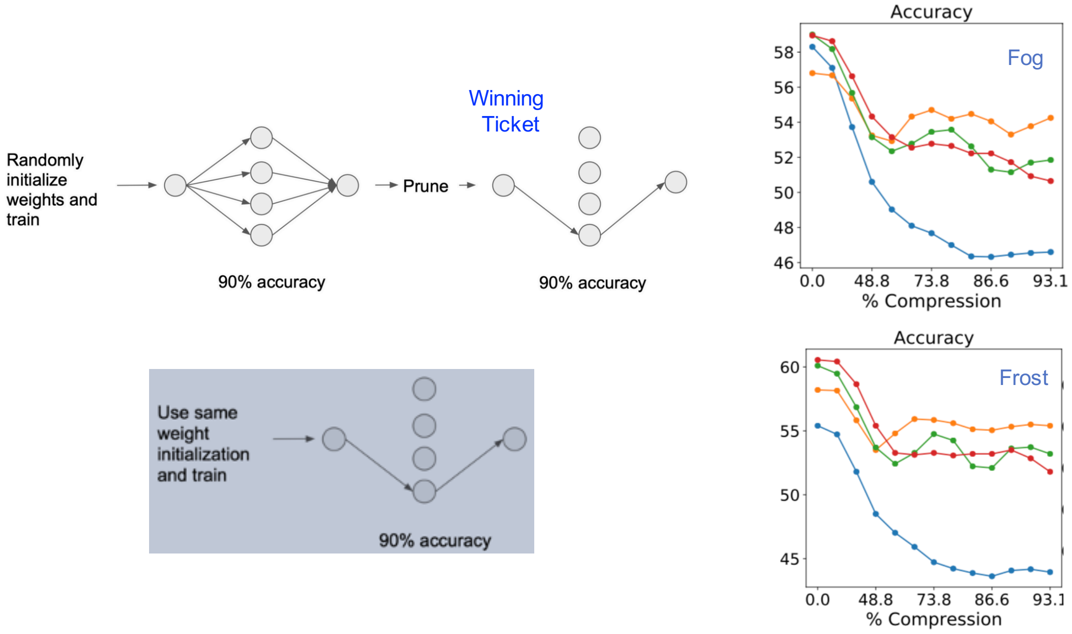


Figure 8: Lottery Ticket transfer performance on target datasets (CIFAR-10-C) that are characterized by distribution shifts when compared to the source data (standard CIFAR-10). The shifts were created using natural image corruptions, and we used ResNet-18 models for this experiment.

With an over-parameterized neural network, pruning or compressing its layers, while not compromising performance, can significantly improve the computational efficiency of the inference step. However, until recently, training such sparse networks directly from scratch has been challenging, and most often they have been found to be inferior to their dense counterparts. The recent work on lottery ticket hypothesis (LTH), showed that one can find sparse sub-networks embedded in over-parameterized networks, which when trained using the same initialization as the original model can achieve similar or sometimes even better performance. Surprisingly, even aggressively pruned networks ($>95\%$ weights pruned) were showed to be comparable to the original network, as long as they were initialized appropriately. Such a well-performing sub-network is often referred as a *winning lottery ticket* or simply a *winning ticket*.

We studied the impact of prediction calibration during model training on the inferred tickets and their generalization (see Figure 8). It is well known that supervised models with uncalibrated confidences tend to be overconfident even while making wrong predictions. This observation is highly relevant to LTH methods, where the most popular strategy used for selecting winning tickets is to rank the network weights based on their magnitudes. We hypothesize that, while neurons with the largest magnitude are the most useful for sub-network selection, they also present the highest risk for causing over-confidences in model predictions. Consequently, including confidence calibration as an explicit training objective will temper the influence of neurons that can eventually lead to miscalibration, as they continue to be updated in the gradient descent process. For the first time [9], we showed that pruned tickets obtained via confidence calibration, though retrained using the same initialization as the standard LTH, leads to improved performance.

6 Leveraging Uncertainties to Handle Structural Noise

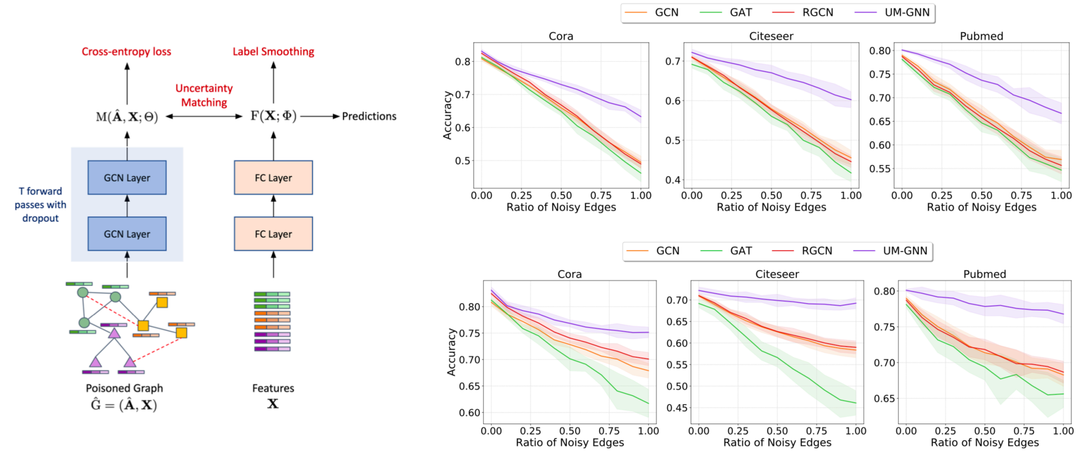


Figure 9: Our approach trains a surrogate model through an uncertainty matching strategy and provides immunity against a wide-range of corruptions.

Graph Neural Networks (GNNs), a generalization of neural networks to graph-structured data, are often implemented using message passes between entities of a graph. They are being adopted in a wide-range of science applications including material design, drug discovery and the study of neurogloical conditions. While GNNs are effective for node classification, link prediction and graph classification, they are vulnerable to structural noise, i.e., a small perturbation to the structure can lead to a non-trivial performance degradation. In a recent paper [10], we proposed Uncertainty Matching GNN, that is aimed at improving the robustness of GNN models, particularly against poisoning attacks to the graph structure, by leveraging epistemic uncertainties from the message passing framework. More specifically, we designed a surrogate predictor that does not directly access the graph structure, but systematically extracts reliable knowledge from a standard GNN through a novel uncertainty-matching strategy. Interestingly, this uncoupling makes GNNs immune to evasion attacks by design, and achieves significantly improved robustness against poisoning attacks. Using empirical studies with standard benchmarks and a suite of global and target attacks, we demonstrated the effectiveness of our method, when compared to existing baselines including the state-of-the-art robust GCN.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52- 07NA27344 and was supported by the ASCR UQ porject SCW1685.

Publications

[1] Jayaraman Thiagarajan, Bindya Venkatesh, Prasanna Sattigeri, and Peer-Timo Bremer. Building Calibrated Deep Models via Uncertainty Matching with Auxiliary Interval Predictors. In *AAAI Conference on Artificial Intelligence*, pages 6005–6012, 2020.

[2] Bindya Venkatesh and Jayaraman Thiagarajan. Heteroscedastic Calibration of Uncertainty Estimators in Deep Learning. *Preprint at https://arxiv.org/abs/1910.14179*, 2020.

[3] Jayaraman Thiagarajan, Bindya Venkatesh, Rushil Anirudh, Peer-Timo Bremer, Jim Gaffney, Gemma Anderson, and Brian Spears. Designing Accurate Emulators for Scientific Processes using Calibration-Driven Deep Models. *Nature Communications (to appear)*, 2020.

[4] Jayaraman Thiagarajan, Bindya Venkatesh, Deepta Rajan, and Prasanna Sattigeri. Improving Reliability of Clinical Models using Prediction Calibration. In *MICCAI UNSURE Workshop*, 2020.

[5] Jayaraman Thiagarajan, Bindya Venkatesh, and Deepta Rajan. Learn-By-Calibrating: Using Calibration as a Training Objective. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[6] Jayaraman Thiagarajan, Kowshik Thopalli, Deepta Rajan, and Pavan Turaga. Calibrating Healthcare AI: Towards Reliable and Interpretable Deep Predictive Models. *Preprint at https://arxiv.org/abs/2004.14480*, 2020.

[7] Vivek Sivaraman Narayanaswamy, Jayaraman Thiagarajan, Rushil Anirudh, Fahim Forouzanfar, Peer-Timo Bremer, and Xiao-Hui Wu. Designing Deep Inverse Models for History Matching in Reservoir Simulations. In *ML for Physical Sciences Workshop*. NeurIPS, 2019.

[8] Bindya Venkatesh and Jayaraman Thiagarajan. Ask-n-Learn: Active Learning via Reliable Gradient Representations for Image Classification. *Preprint at https://arxiv.org/abs/2009.14448*, 2020.

[9] Bindya Venkatesh, Jayaraman Thiagarajan, Kowshik Thopalli, and Prasanna Sattigeri. Calibrate and Prune: Improving Reliability of Lottery Tickets Through Prediction Calibration. *Preprint at https://arxiv.org/abs/2002.03875*, 2020.

[10] UdayShankar Shanthamallu, Jayaraman Thiagarajan, and Andreas Spanias. Uncertainty-Matching Graph Neural Networks to Defend Against Poisoning Attacks. *Preprint at https://arxiv.org/abs/2009.14455*, 2020.

[11] News feature. https://tinyurl.com/y32zyqmx.

[12] Dataskeptic podcast. https://t.co/VofvjWSaYL?amp=1.