Uncertainty Quantification of ChemCam Data with Bootstrap and Monte-Carlo Dropout

F. Holguin[1,2], N. Klein[2], D. A. Oyen[2], and P. Gasda[2].
[1]University of Michigan-Ann Arbor, Dept. of Astronomy,
[2]Los Alamos National Laboratory; opaco@umich.edu

Introduction: The ChemCam instrument, located on the Curiosity rover on Mars, measures rock chemical compositions at a distance via laser-induced breakdown spectroscopy (LIBS) spectra using analysis with traditional regression methods. Deep neural networks (Deep NNs) have recently[2] been employed to make predictions using the LIBS data calibrated on Earth; however, the models lack uncertainty estimates in their predictions, resulting in unknown confidence in model predictions of Mars data, where the true value is not available. In these distant and foreign environments, it is essential for terrestriallytrained NN models to contain uncertainty quantification (UQ). We examine two straightforward UQ methods, Bootstrap and Monte-Carlo Dropout, applied to NNs trained on LIBS calibration data.

Dense Neural Network and UQ Methods:

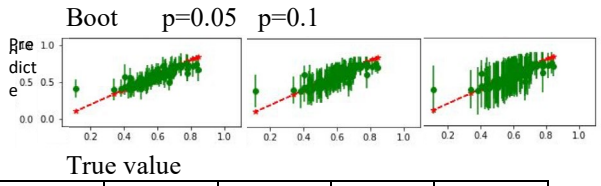
Bootstrap: Bootstrap uncertainty intervals for a trained model are produced by re-sampling the data, retraining the model, and producing a prediction from the new model. The distribution of these new data points defines the bootstrap uncertainty intervals. We only retrain the NN output layer, as this choice has been shown to produce satisfactory results[1].

Monte-Carlo Dropout (MCD): MCD refers to the random elimination (dropout) of NN nodes at either training or prediction time, which allows us to obtain a distribution over predictions. In these results, we apply MCD to a proportion p of neurons at prediction time to approximate Bayesian inference over the weight space[3].

NN model implementation: In order to efficiently explore the UQ method properties, we consider a simple, dense NN with single hidden layer: the input (LIBS spectra), hidden, and output (composition) layers are of size 5000, 512, and 1 respectively. The NN key parameters are the hidden layer size (fixed to 512) and activation function (fixed to ReLu), layer weight ℓ_2 regularization parameter λ, and dropout rate p.

ChemCam Data: We use ChemCam LIBS calibration measurements[5] made in a terrestrial Mars atmosphere chamber, corresponding to 378 unique materials. We compute the average LIBS spectrum for each material after discarding the first five shots, and then rescale by the maximum value across all wavelengths. The training, test, and validation sets are of size 200, 100, and 50 respectively. For simplicity, we consider a single composition value of SiO_2 as it is the most abundant in the samples.

Results: *Evaluation metrics:* In order to assess the mean predictions and uncertainty intervals we compute three metrics: (i) Root Mean Square Error (RMSE): square root mean of the difference squared between model prediction and LIBS reference value (ii) Coverage: percentage of predicted data points falling within model error bars, and (iii) Interval Score: a function[4] that penalizes both interval size and RMSE.



	Ideal	Boot1	MCD1	MCD2
RMSE	Lower	6%	6%	6%
Coverage	95%	96%	97%	98%
Int Score	Lower	0.19	0.38	0.55

Figure 1: Top: True vs. predicted composition, including 95% intervals for three modes. Bottom: Summary of metrics.

Summary: We train a series of NNs in the parameter space $\log \lambda \in [-5, -4, -3, -2]$ and $p \in [0.05, 0.1, 0.2, 0.3]$. In Figure 1, we present a model ($\log \lambda = -5$) analyzed with bootstrap and MCD: Boot, MCD1 (p=0.05) and MCD2 (p=0.1). These models have a RMSE comparable to previous works[2]. The bootstrap model performs the best on all three metrics. The lowest dropout rate also produced good results, although the interval score was worse. Increasing the dropout rate further to p=0.1 results in larger error bars. Our results demonstrate that bootstrap and MCD can potentially be useful for UQ applied to ChemCam data, although the parameter choices are important to consider.

Future Work: We plan to include additional compositions and examine the performance on Mars-collected data.

Acknowledgements: Research presented here was supported by the 2021 Los Alamos Applied Machine Learning workshop, and the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20210043DR.

References: [1] Brosse et al. (2020), [2] Castorena et al. (2021) *Spec. Acta Part B*. [3] Gal & Ghahramani. 2016. *PMLR.* [4] Gneiting et al. 2007. *J. Am. Stat. Assoc.* [5] Clegg et al. (2017) *Spectrochim. Acta B.*