SnT 2019
CTBT: SCIENCE AND TECHNOLOGY CONFERENCE

Toward reliable certainty for seismic processing tasks with deep learning

Lisa Linville[1], Chris Young[1] and Nathan Downy[1]

1.Sandia National Labs

T3.5-P77

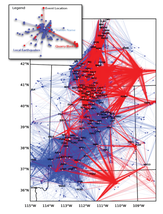
ABSTRACT

Machine learning is a powerful way to accomplish a variety of seismic processing tasks, outperforming state-of-the-art for many traditional discrimination techniques in terms of broad applicability and performance. However, self-reported certainty from learned models can be unreliable, especially for data with characteristics outside of training distributions. While these cases can be rare, or go unnoticed during controlled experiment set-up and testing, real-world situations, production settings, and monitoring objectives require reliable estimates of certainty for decision making. In this work the goal is to focus on understanding the current limitations of certainty from model output on a simple seismic event discrimination task using deep convolutional model architectures. With an understanding of the limitations on our current methods, can we improve upon the trustworthiness of a model's reported certainty through data augmentation, model adaptations, or external trust metrics?

BACKGROUND

You can use deep learning and seismic event catalogs to build models that predict attributes like event type at above 99% accuracy. We did this for Utah (Box 1), and it turns out to be an easy task for deep learning to accomplish. Not only does this work for event classification, but also many other things like phase identification (Ross et al, 2018) and event detection (Perol et al. 2017). **That's great news!!**

Box 1: Utah Classification

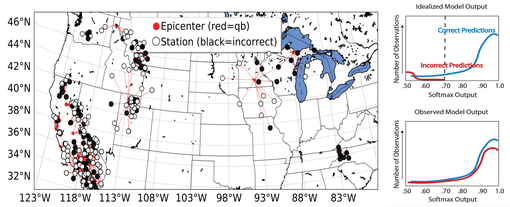


Linville et al. (2019) used a catalog of 100 thousand earthquakes (blue, LEFT) and quarry blasts (red, LEFT) from the University of Utah Seismograph network (stations in white, LEFT) to train deep neural network models (both convolutional neural networks or CNN's and recurrent or Long Short Term Memory networks) to identify the source type of events in Utah. These model could predict event class at 96% accuracy using any single observing station, and over 99% using all the stations for an observed event. The models turned out to also be very good at identifying where analysts misclassified events- these were often cases where in both training and test data the model disagreed with event labels, and all the model station predictions were in agreement.

Now we can use the new model to help overburdened analysts as they manually review events. We move the models into operations where they automatically report the event type for each new incoming event, thus improving the efficiency and accuracy of monitoring operations. **Except...**

Box 2: Unreliable Uncertainty

When a CNN model makes a prediction, it returns a score vector equal in length to the number options the model has for an answer. For the Utah data, the model has two options, either earthquake or quarry blast. The score vector reports how much the model thinks each input example is an earthquake and how much it believes the input is a quarry blast. Often normalized to be between 0 and 1, the score vector is the models confidence and interpreted as the probability of the input belonging to each of the possible options.



A common attribute of high-performing deep predictive models is that they are always think they are right. Take for example the map of quarry blast events (ABOVE) from outside of Utah. Utah models predict events slightly better than half of the time, yet instead of reporting low confidence for examples outside of the Utah, and in cases where it is wrong (i.e. the Top right idealized output gives a bimodal distribution for correct (blue) and incorrect (red) examples, the reported distributions are identical.

When the model reports a prediction it is generally highly confident that it's answer is correct. Even when the answer is incorrect. In short, the reported model confidence is untrustworthy. The problem gets worse when the model encouters examples unlike those it has ever seen before (Box 2 demonstrates these concepts for Utah data). This is problematic because it requires the analysts to scrutinize each decision they make, just as they would without the benefit of model decisions which are correct 99% of the time. **So where does this leave us?**

Option 1: The analysts can use the model predictions blindly, accepting that the model will get about the same number of the events wrong as human analysts would anyway. They benefit from increased processing efficiency and they expect that the 1% of events that are wrongly predicted will be less arbitrary than analyst errors (they aren't the result of fatigue, inexperience or lack of context), but it will also be difficult to isolate specific and systematic conditions which lead to bad model decisions. Comprehensive testing that spans the expected range of conditions the model will experience is the best way to anticipate model failure and credibly remain at acceptable peformance accuracies, and this will require peristent system testing, model retraining, and rigorous evaluation. Option 1 is a benefitial option for Utah becase trained models, despite their limitations, support Utah's monitoring directives at the same level or better than current practice. This option however will not be appropriate for operations who's directives are concerned primarily with the 1% of times the model fails or is uncertain... and without credible uncertainty these operations derive almost no benefit from the development of high performing predictive models with deep learning.

Option 2: Develop and test methods to obtain uncertainty estimates that capture meaningful sources of error and lead to actionable insight from deep predictive models. Although there are many approaches to understanding uncertainty inherent to datasets and target solutions, we seek solutions that produce a seperation in the distribution of events of correct predictions and incorrect predictions. With sufficient seperation of in-distribution examples (examples that are not rare, but are drawn from a dataset similar to what predictive models observed during training) thresholds on confidence allow analyst to use model descisions to prioritize their attention.

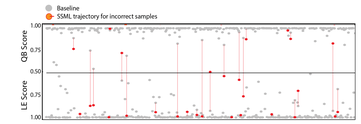
In this work we test three approaches to uncertainty quantification (Option 2) using the Utah data (Box 1).

APPROACH

We explore solutions to uncertainty quantification that balance the ease of implementation, interpretability, scalability and usability. For example, our first approach (Box 3) tests how different regularization strategies effect a models own reported confidence scores. If specific regularization strategies effect reported confidence adventageously, we gain a lot without significant investment in external methods or more sophisticated modelling. By comparison, state-of-the-art (SOTA) methods of uncertainty quantification for deep learning (Box 5) involve model architecture reframing that can then require significantly more computational resources and lead to difficult optimization. The payoff for these SOTA methods, however, is principled uncertainty that can be used for robust decision making. Intermediate approaches like external trust metrics (Box 4) seek the benefits of ease of use and interpretability while also offering solutions in support of our main objective: actionable insight from deep predictive models.

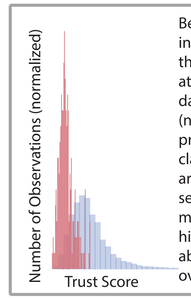
Box 3: Regularization

Observations from semisupervised machine learning (SSML) suggest that regularization techniques applied during learning may benefit confidence estimates. For example, if we observe model output for incorrect samples from baseline models (grey dots with red lines) compared to SSML models (red dots) these samples are often moved to the corrrect class with lower confidence.



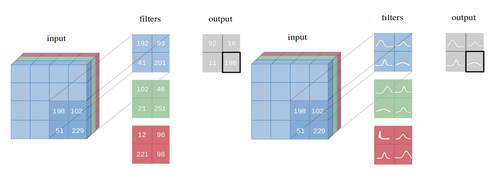
Another regularization strategy we might expect to influence reported confidence scores is through combined loss functions that are sentistive to physical constraints. For example, when a station makes a different prediction than other stations for the same event, we can enforce additional penalty on this example during learning by making the loss function sensitive to the predicitons of it's neighboring stations for a specific event. RESULT: Both consistency-based approaches to regularization fail to seperate correct and incorrect prediction distributions for Utah data. While both techniques increase the accuracy of final models, they often do so at the expense of calibration (the median reported confidence is higher than the average accuracy the model achieves). For the SSML case, we believe this is because while models do act on some incorrect examples in expected ways, there are often tradeoffs with other examples that reverse the average effect of SSML on overal reported confidence. Likewise, we demonstrate that enforcing station consistency can boost overall accuracy but exaggerate the problem of model confidence.

Box 4: External Metrics



Because deep predictive models learn to operate within a world that is entirely based on the data they experience during training, models may simply have limited ability to reliably know what they don't know. We turn to external trust metrics to assess if these can offer solutions to a model's self-reported uncertainty. We use the trust score developed by Jiang et al., 2018 to evaluate how well the distributions of correct and incorrect predictions are seperable. The trust score has two steps. First, we use the data to define a high-density subset for each class (we reject outliers from each class). Second, the trust score is ratio of the (nearest-neighbor) distance to the high-density set of the class that was not predicted, and the distance of the sample to the predicted class. Intuitively, when samples are far away from samples from the opposite class and close to sample of the same class we expect the predictions to be right and they are given a high trust score. Although the range of trust scores falls on an arbitrary axis defined by the relationships in the data, (so it will not be meaningful to compare absolute values from one dataset to another), for a given dataset the scores reflect the relationships between the confidence of predictions for any predictive model architecture. RESULT: We find that compared to softmax outputs for the same models, which yield nearly identical and highly confident estimates, the distributions of incorrect (red LEFT)/correct (blue LEFT) predictions using the trust score are able to seperate distribution peaks. Despite the promising peak seperation, this method leaves significant portions of the tails overlapping. Ideal seperation occurs when the tail of the red distrubution is well left of the blue peak distribution.

Box 5: Bayesian DNN



Bayesian neural networks (BNNs) provide a framework to represent uncertainty in output distributions from neural networks (Neal et al., 2012). In principle, BNNs provide reliable estimates of the uncertainties inherent to both the data and model. RESULT: We are still working on BNN model training for Utah data. Both the computation and the optimization are more difficult with BNNs than a traditional CNN and in order to be more valuable than external trust metrics (Box 4) uncertainty estimates from BNN's should outperform these methods in significant ways.

In Utah CNN models, our filters are made of discrete weights learned through model training (TOP LEFT). In the bayesian framework, weights become distributions and during inference (TOP RIGHT) and each example is sampled many times to generate an output distribution instead of a single answer. The variation in the output distribution is what we use to quanitfy how well our model understands and predicts each input example. Image from Shridhar et al., 2018.

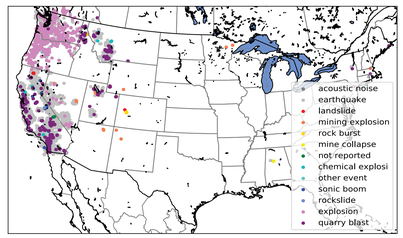
CONCLUSIONS

The main goals of this presentation are:

To convey a concrete understanding of the impact of missing uncertainty estimates from otherwise high performing predictive models (see: BACKGROUND).

To outline several approaches to uncertainty from deep predictive models and share our experiences with these approaches (see: APPROACH).
Currently, external trust estimates provide significant improvements over the self reported certainty from trained models. It remains to be seen if BNN's will outperform external trust metrics sufficently well to make them attractive in spite of their high computation and optimization cost.

Future: Here we test uncertainty with Utah data where we expect the uncertainty to be reported reliably for in-domain samples. Future work incorporates out of domain examples from a the continental United states for a variety of source types (shown in figure on the RIGHT).



Sandia National Laboratories

REFERENCES

Jiang, H., Kim, B., Guan, M., & Gupta, M. (2018). To trust or not to trust a classifier. In Advances in Neural Information Processing Systems (pp. 5541-5552).

Linville, L., Pankow, K., & Draelos, T. (2019). Deep Learning Models Augment Analyst Decisions for Event Discrimination. Geophysical Research Letters, 46(7), 3643-3651.

Neal, R. M. (2012). Bayesian learning for neural networks (Vol. 118). Springer Science & Business Media.

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. Science Advances, 4(2), e1700578.

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. Bulletin of the Seismological Society of America, 108(5A), 2894-2901.

Shridhar, K., Laumann, F., Llopart Maurin, A., & Liwicki, M. (2018). Bayesian Convolutional Neural Networks. arXiv preprint arXiv:1806.05978.

SAND2018-11166

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Disclaimer: The views expressed on this poster are those of the author and do not necessarily reflect the view of the CTBTO