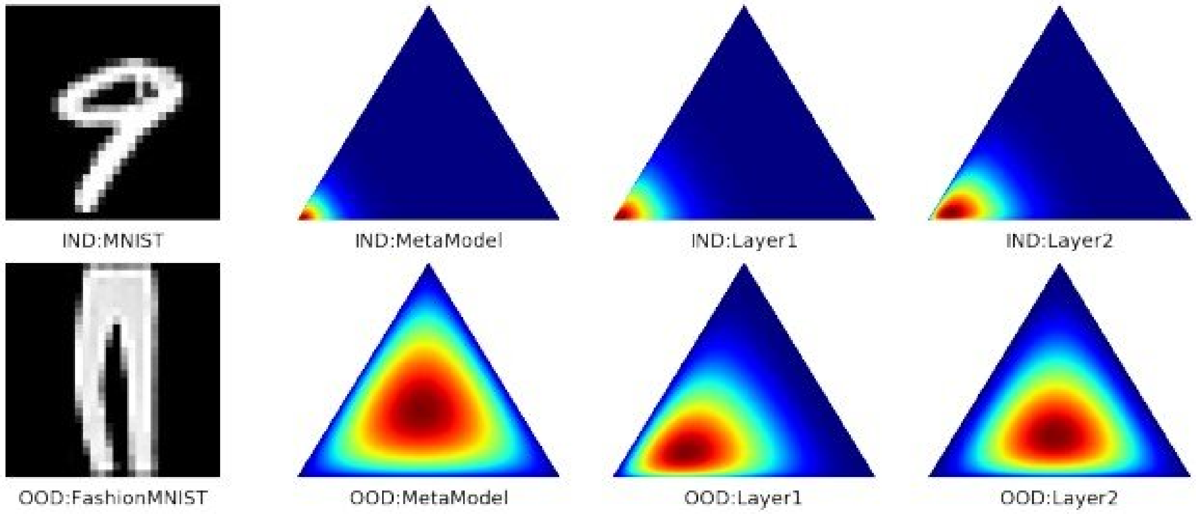
Efficient technique improves machine-learning models' reliability

February 13 2023, by Adam Zewe



A toy example of our proposed meta-model method in OOD detection application shows the diversity of features in different layers. MetaModel utilizes two intermediate features, while Layer1 and Layer2 only are trained with one individual feature. Credit: *arXiv* (2022). DOI: 10.48550/arxiv.2212.07359

Powerful machine-learning models are being used to help people tackle tough problems such as identifying disease in medical images or detecting road obstacles for autonomous vehicles. But machine-learning models can make mistakes, so in high-stakes settings it's critical that humans know when to trust a model's predictions.

Uncertainty quantification is one tool that improves a model's reliability; the model produces a score along with the prediction that expresses a confidence level that the prediction is correct. While uncertainty quantification can be useful, existing methods typically require retraining the entire model to give it that ability. Training involves showing a model millions of examples so it can learn a task. Retraining then requires millions of new data inputs, which can be expensive and difficult to obtain, and also uses huge amounts of computing resources.

Researchers at MIT and the MIT-IBM Watson AI Lab have now developed a technique that enables a model to perform more effective uncertainty quantification, while using far fewer computing resources than other methods, and no additional data. Their technique, which does not require a user to retrain or modify a model, is flexible enough for many applications.

The technique involves creating a simpler companion model that assists the original machine-learning model in estimating uncertainty. This smaller model is designed to identify different types of uncertainty, which can help researchers drill down on the root cause of inaccurate predictions.

"Uncertainty quantification is essential for both developers and users of machine-learning models. Developers can utilize uncertainty measurements to help develop more robust models, while for users, it can add another layer of trust and reliability when deploying models in the real world. Our work leads to a more flexible and practical solution for uncertainty quantification," says Maohao Shen, an electrical engineering and computer science graduate student and lead author of a paper on this technique.

Shen wrote the paper with Yuheng Bu, a former postdoc in the Research Laboratory of Electronics (RLE) who is now an assistant professor at the

University of Florida; Prasanna Sattigeri, Soumya Ghosh, and Subhro Das, research staff members at the MIT-IBM Watson AI Lab; and senior author Gregory Wornell, the Sumitomo Professor in Engineering who leads the Signals, Information, and Algorithms Laboratory RLE and is a member of the MIT-IBM Watson AI Lab. The research will be presented at the AAAI Conference on Artificial Intelligence, and the paper is available on the *arXiv* preprint server.

Quantifying uncertainty

In uncertainty quantification, a machine-learning model generates a numerical score with each output to reflect its confidence in that prediction's accuracy. Incorporating uncertainty quantification by building a new model from scratch or retraining an existing model typically requires a large amount of data and expensive computation, which is often impractical. What's more, existing methods sometimes have the unintended consequence of degrading the quality of the model's predictions.

The MIT and MIT-IBM Watson AI Lab researchers have thus zeroed in on the following problem: Given a pretrained model, how can they enable it to perform effective uncertainty quantification?

They solve this by creating a smaller and simpler model, known as a metamodel, that attaches to the larger, pretrained model and uses the features that larger model has already learned to help it make uncertainty quantification assessments.

"The metamodel can be applied to any pretrained model. It is better to have access to the internals of the model, because we can get much more information about the base model, but it will also work if you just have a final output. It can still predict a confidence score," Sattigeri says.

They design the metamodel to produce the uncertainty quantification output using a technique that includes both types of uncertainty: data uncertainty and model uncertainty. Data uncertainty is caused by corrupted data or inaccurate labels and can only be reduced by fixing the dataset or gathering new data. In model uncertainty, the model is not sure how to explain the newly observed data and might make incorrect predictions, most likely because it hasn't seen enough similar training examples. This issue is an especially challenging but common problem when models are deployed. In real-world settings, they often encounter data that are different from the training dataset.

"Has the reliability of your decisions changed when you use the model in a new setting? You want some way to have confidence in whether it is working in this new regime or whether you need to collect training data for this particular new setting," Wornell says.

Validating the quantification

Once a model produces an uncertainty quantification score, the user still needs some assurance that the score itself is accurate. Researchers often validate accuracy by creating a smaller dataset, held out from the original training data, and then testing the model on the held-out data. However, this technique does not work well in measuring uncertainty quantification because the model can achieve good prediction accuracy while still being over-confident, Shen says.

They created a new validation technique by adding noise to the data in the validation set—this noisy data is more like out-of-distribution data that can cause model uncertainty. The researchers use this noisy dataset to evaluate uncertainty quantifications.

They tested their approach by seeing how well a meta-model could capture different types of uncertainty for various downstream tasks, including out-of-distribution detection and misclassification detection. Their method not only outperformed all the baselines in each downstream task but also required less training time to achieve those results.

This technique could help researchers enable more machine-learning models to effectively perform uncertainty quantification, ultimately aiding users in making better decisions about when to trust predictions.

Moving forward, the researchers want to adapt their technique for newer classes of models, such as large language models that have a different structure than a traditional neural network, Shen says.

More information: Maohao Shen et al, Post-hoc Uncertainty Learning using a Dirichlet Meta-Model, *arXiv* (2022). DOI: 10.48550/arxiv.2212.07359

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Efficient technique improves machine-learning models' reliability (2023, February 13) retrieved 26 April 2024 from https://techxplore.com/news/2023-02-efficient-technique-machine-learning-reliability.html