CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Key Takeaways for "Key Concepts in AI Safety: Reliable Uncertainty Quantification in Machine Learning"

Models are likely to be safer if they can appropriately express uncertainty. Good uncertainty quantification can reduce human overreliance on the correctness of AI outputs and reduce the chance of AI systems taking harmful actions in unusual circumstances. While progress has been made to improve uncertainty quantification, existing approaches are imperfect and more research remains to be done.

Uncertainty quantification focuses on getting machine learning (ML) models to "know what they know." For example, models should be **calibrated:** if a model reports 80% confidence in a particular output, that output should be correct 80% of the time. Models should also correctly quantify their uncertainty even when they encounter **distribution shifts:** inputs very different from those they encountered during training.

Several existing approaches aim to improve uncertainty quantification in ML models, each with its own benefits and limitations. Deterministic methods can be used to train models to have high uncertainty in select types of data different from those they were trained on but do not yield reliable uncertainty quantification across all possible kinds of data. **Model ensembling** compares predictions across several different models to arrive at an uncertainty estimate, but it can be computationally expensive and comes without mathematical guarantees that the estimates are reliable. **Conformal prediction** can provide mathematical guarantees of reliable uncertainty quantification, but only under assumptions that are unlikely to hold in the real world. Finally, **Bayesian inference** uses probability theory to express model uncertainty, but it is difficult to use for most large neural networks commonly used today.

It is possible to improve uncertainty quantification in ML models, but there is more work to be done. Applying the methods described above as add-ons to existing neural network models can improve uncertainty quantification, but human operators should be aware that none of the existing methods are perfect. Research is especially nascent in the large language models that power systems like Open AI's ChatGPT. Future research can and should continue to make progress in this important area.

For more information:

- Download the report: https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-reliable-uncertainty-quantification-in-machine-learning/
- Contact us: cset@georgetown.edu