Popular science summary of the PhD thesis

PhD student	Hugo Senetaire
Title of the PhD thesis	Generative models and Applications for Safety in Machine Learning
PhD school/Department	DTU Compute

Science summary

The recent advances in machine learning have transformed many scientific domains (image recognition, protein folding, drug designs, etc.) and started widespread adoption in real-world applications. However, this expansion into critical areas (for instance, self-driving cars, medical imaging, etc.) emphasizes the need for safe and reliable deployment. As the complexity of the models grows, our ability to fully understand their predictions diminishes. For example, predictive models have been shown to rely on artefacts or undesirable features in the data, raising concerns about their robustness and trustworthiness. This thesis addresses these challenges by proposing methods to enhance the safe deployment of machine learning models. In particular, we want to leverage probabilistic generative models toward that goal.

Machine learning primarily encompasses two key paradigms: supervised learning (predicting a specific target for a given input) and unsupervised learning (modelling data to generate new samples). A significant subset of unsupervised learning is probabilistic generative modelling, which represents the underlying phenomena as probability distributions. This thesis explores these models and introduces a novel approach for training energy-based models to achieve maximum likelihood.

A notable limitation of deep probabilistic generative models is their poor performance in detecting out-of-distribution (OOD) data. To address this, we propose a robust OOD detection method that combines multiple statistical tests to deliver a single OOD prediction.

Additionally, we leverage the sampling capabilities of generative models to explain supervised learning models. In particular, we focus on finding the most essential features for a model to give its prediction. We introduce a framework, **LEX**, which formulates general feature attribution as a maximum likelihood objective. This framework is very modular and encompasses many different feature attribution methods.

Finally, we develop new feature attribution objectives to identify key features that differentiate between classes or instances. Using a generative adversarial network to produce counterfactuals, we extend **LEX** to **DupLEX** allowing us to generate these new feature importance scores.

Popular science summary of the PhD thesis

PhD student Hugo Senetaire

Please submit the summary to the department PhD coordinator together with your thesis