Out-of-Distribution Detection for Reliable Machine Learning

Abstract

Machine Learning models are typically trained under the assumption that training and deployment data follow the same distribution; however, this assumption often fails in real-world settings, leading to unreliable and overconfident predictions. This project investigates out-of-distribution (OOD) detection techniques to improve the reliability of machine learning systems. Classification models were trained on in-distribution data and evaluated using both in-distribution and out-of-distribution samples to assess their ability to detect distribution shifts. Confidence-based and distance-based OOD detection methods were explored, including softmax score analysis and feature-space representations. Experimental results indicate that models exhibit high confidence on unseen data when OOD detection is not applies, while incorporating OOD detection mechanisms significantly improves the identification of unfamiliar inputs. The study highlights the importance of OOD awareness for deploying dependable machine learning systems and emphasizes reliability evaluation beyond standard predictive accuracy.