**Title: A Novel Testing Framework for Vision Models Using Bayesian Network**

Deep learning (DL) models are critical in high-stake domains such as autonomous driving, medical diagnostics, and security systems. Real-world deployment of these models requires rigorous robustness testing due to diverse environmental conditions. Current testing approaches primarily focus on neuron coverage (i.e., neuron activation ratio on test inputs). Although this metric is critical, it alone does not ensure comprehensive coverage of all corner cases, which can lead to unexpected failures, thus leaving a gap in the overall evaluation of model’s robustness. My research aims to develop a comprehensive testing framework designed to enhance the evaluation of DL models through a five-stage process. The initial stage, *specification*, focuses on clearly defining all necessary properties of the system to guide the entire testing process and ensure comprehensive coverage. The second stage, *sampling*, gathers all relevant samples necessary for thorough model testing. In the third stage, *test case generation*, properties specified in the first stage are correlated with the collected samples, and test cases are generated accordingly. For example, in autonomous car testing, properties such as dust, noise, rain, and night conditions are considered to evaluate model performance under these conditions. The fourth stage, *probabilistic testing*, evaluates the generated test cases. Robustness tests are conducted both locally and globally. Locally, the robustness of the model is evaluated within individual categories or classes to identify weaknesses. In contrast, globally, the model’s performance is assessed across various categories to enhance its generalisation capabilities across different scenarios. This stage integrates a probabilistic approach using Bayesian network, combined with solid mathematical formulation, to provide a comprehensive visual and quantitative analysis of the model’s performance at both local and global levels. Errors are systematically recorded for later analysis in the final stage, *error summarization*, which produces actionable graphical error reports and recommendations for model refinement.