Application of Reliability and Resilience Models to Machine Learning

Zakaria Faddi[1], Karen da Mata[1] , Priscila Silva[1], Vidhyashree Nagaraju[2], Susmita Ghosh[3], and Lance Fiondella[1]

[1]University of Massachusetts Dartmouth, [2]Stonehill College, [3]Jadavpur University

Introduction

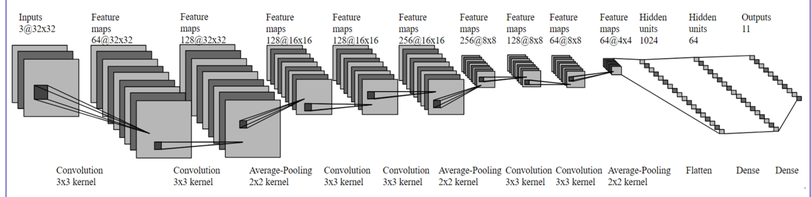
Machine learning (ML) has become integral to numerous domains due to its potential to automate decision-making processes and handle complex tasks. Despite the growing popularity of ML systems, they are susceptible to adversarial scenarios. This research demonstrates the applicability of software reliability and resilience tools to ML algorithms providing an objective approach to assess recovery after a degradation from known adversarial attacks.

Contribution Objectives

- Apply Software Reliability Growth Models (SRGM) to characterize defects discovery in ML-enabled applications
- Apply resilience models to track and predict ML-systems performance under adversarial scenarios
- Develop data collection technique to promote risk quantification that can provide best practices in order to conserve resources

Theory of Operation

Machine Learning Architecture



Generative Adversarial Attacks

Fast Gradient Sign Method (FGSM)
a likelihood-based model to generate adversarial examples for a neural network model

$$x_{adv} = x + \epsilon \times \text{sign}(\nabla_x \mathcal{L}(\theta, x, y))$$

Projected Gradient Descent (PGD)
uses the magnitude of the gradient, taking multiple steps in the direction of the gradient of the output of the classification model with respect to the input image.

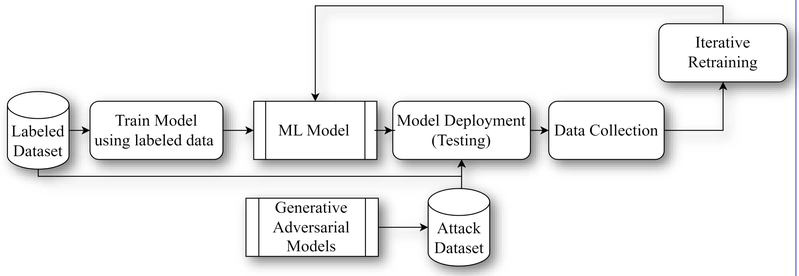
$$x_{adv_{t+1}} = \prod_{\Delta}(x_{adv} + \alpha \text{sign}(\nabla_x \mathcal{L}(\theta, x, y)))$$

Defensive Measures

Adaptive Adversarial Training (AAT): combines Adversarial and Online training to achieve an even more robust model. The model is continuously trained with newly available data while retaining previously learned knowledge

$$\min_{\theta}[\max_{\delta \in \Delta}[\mathcal{L}(x + \delta, y, \theta)]]$$

Test Setup



Modeling

Reliability

NHPP SRGM

Mean Value Function $m(t) = a \times F(t)$

- Goel–Okumoto (GO) $m(t) = a(1 - e^{-bt})$
- Weibull Model $m(t) = a(1 - e^{-bt^c})$
- Delayed S-shaped (DSS) $m(t) = a(1 - (1 + bt)e^{-bt})$

Failure Rate $\lambda(t) = \frac{dm(t)}{dt}$

Reliability Function $R(t|t_0) = e^{-(m(t+t_0)+m(t_0))}$

Discrete Cox Proportional Hazard NHPP SRGM

Mean Value Function $m(X) = \omega \sum_{i=1}^{n} p_{i,X_i}$

Probability

$$p_{i,X_i} = (1 - (1 - h(i))^{g(X_i;\beta)}) \prod_{k=1}^{i-1} (1 - h(k))^{g(X_k;\beta)}$$

Cox Exponential Function

$$g(X_i; \beta) = exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im})$$

Resilience

Regression Models

Discrete Resilience Curve

$$P(i) = P(i-1) + \Delta P(i)$$

Multiple Linear Regression (MLR)

$$\Delta P(i) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j(i)$$

MLR with Interaction (MLRI)

$$\Delta P(i) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j(i) + \sum_{j=1}^{m} \sum_{k=j+1}^{m} \beta_{j(m+k)} X_j(i) X_k(i)$$

Polynomial Regression (PR)

$$\Delta P(i) = \beta_0 + \sum_{j=1}^{\rho} \sum_{k=1}^{m} \beta_{j(j+k)} X_k(i)^j$$

Model Estimation

Likelihood Function

NHPP SRGM

$$L(\Theta|T) = e^{-m(t_n)} \prod_{i=1}^{n} \lambda(t_i)$$

Discrete Cox

$$L(\theta, \beta, \omega) = exp(-\omega \sum_{i=1}^{n} p_{i,X_i;\theta,\beta}) \omega^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \frac{p_{i,X_i;\theta,\beta}^{y_i}}{y_i!}$$

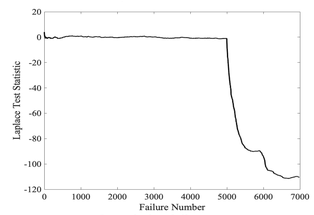
Resilience

$$L(\Delta P; \beta_0, \beta_i, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(n)/2}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[\Delta P(i) - \left(\beta_0 + \sum_{j=1}^{m} \beta_j X_j(i) \right) \right] \right\}$$

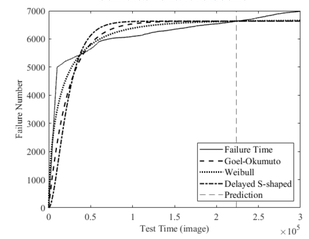
Results

Software Failure and Reliability Assessment Tool (SFRAT)

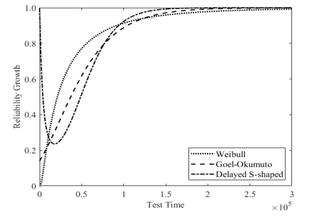
Laplace Trend test



Cumulative failure count

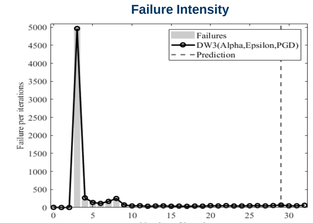


Reliability

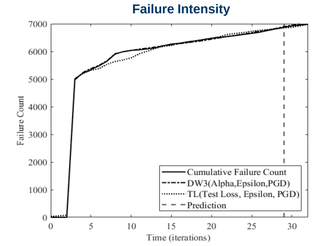


Covariate Software Failure and Reliability Assessment Tool (C-SFRAT)

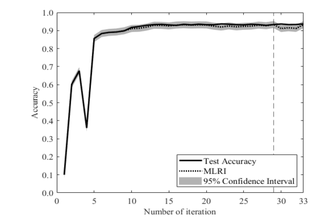
Failure Intensity



Failure Intensity



Predictive System Resilience Assessment Tool (PSRAT)



Conclusion & Future Work

- This research
  - Developed data collection technique to apply traditional NHPP models with and without covariates
  - Successfully characterized the deterioration and recovery of the ML-based model using resilience modeling approaches
- Future research will explore advanced statistical techniques, applications to cyber-physical systems, and pre-existing ML-based systems

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant Number (#1749635). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

Author Contact Information

Lance Fiondella, PhD
Department of Electrical and Computer Engineering
University of Massachusetts Dartmouth
285 Old Westport Road North Dartmouth, MA USA
Email: lfiondella@umassd.edu