A Comprehensive Testing Framework for Deep Learning Models

Author Name

***Abstract*—**

I. Introduction

II. Problem Statement

Deep learning models are being more widely used in a variety of applications, yet their reliability in practical applications remains a challenge.

III. Research Goal

This paper aims to develop a systematic framework for evaluating local and global robustness in deep learning models. The goal is to provide a comprehensive error summary to improve model design and training, ensuring their reliability for real-world applications.

IV. Contributions

This research makes the following key contributions to the field of deep learning robustness evaluation:

- We design an **end-to-end pipeline** for evaluating the robustness of system.
- We propose a **conceptual framework** that quantifies both local and global robustness,with formalized approach to verify system robustness.
- A novel **error summarization** approach which allows better identification of model weaknesses related to class and property.
- We perform all our **experiments** using publicly available deep learning models and MNIST dataset.

V. Research Questions

This paper addresses the following research questions applicable to various deep learning models and datasets:

- How can we design a comprehensive framework to test system robustness?
- How can we systematically evaluated the robustness both at local (property-specific) and global (overall system) levels within framework?
- How can error summarization be employed to quantify the impacts on model robustness?

VI. Methodology

This section presents an overview of the comprehensive testing framework, which is intended to test the specified properties according to given specification. The pipeline begins by precisely describing the properties to be evaluated. To comprehensively examine the model, test cases are created and tested. The error summary step next concentrating on the progression from local to global robustness to highlight the system systemic strengths and weaknesses. This systematic technique improves the robustness of deep learning models by tackling each essential aspect sequentially, from specification to complete error analysis.

A. Sampling

The sample selection process involves a random but balanced choice of samples from each class, focusing exclusively on instances that the model has correctly predicted. This method ensures a representative and fair distribution of data across all classes.

- Model Utilization:
  - A pre-trained CNN model is utilized to select samples.
  - Let $X = \{x_1, x_2, \ldots, x_N\}$ denote the set of MNIST images. The model function f predicts:
    $$f(x_i) \to y_i$$
  - The filter function g identifies accurate predictions, defined as:
    $$g(x_i) = \begin{cases} 1 & \text{if } f(x_i) = \text{true label of } x_i \\ 0 & \text{otherwise} \end{cases}$$
  - The subset S includes only correctly predicted images:
    $$S = \{x_i \in X \mid g(x_i) = 1\}$$
- Random Selection of Samples:
  - Randomly selects 200 samples from each class in S, totaling 2000 samples.
  - The random selection function R is defined to ensure:
    $$R(S_c, 200) \text{ for each class } c \text{ in } S$$
    where S_c represents the samples of class c within S.

B. Test Case Generation

This section outlines the generation of test cases to assess model robustness through properties such as noise, rotation, and brightness adjustments.

- **Noise Addition:**
  - Define the noise property $p_n(x_i, \sigma)$ for adding Gaussian noise to image x_i from subset S, where σ specifies the noise intensity:
    $$x_i' = p_n(x_i, \sigma)$$

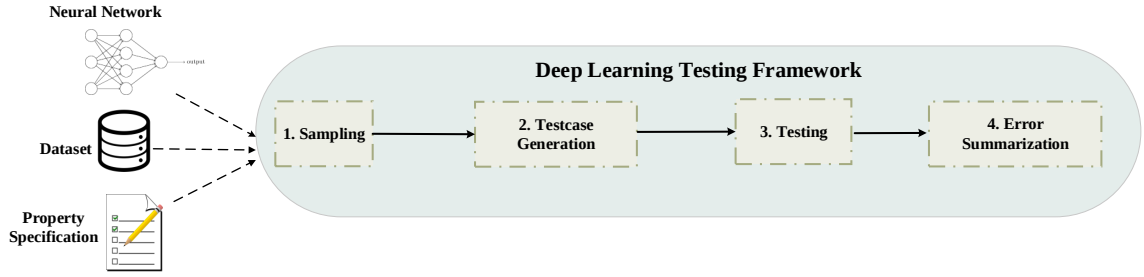

Fig. 1: Overview of Testing Framework

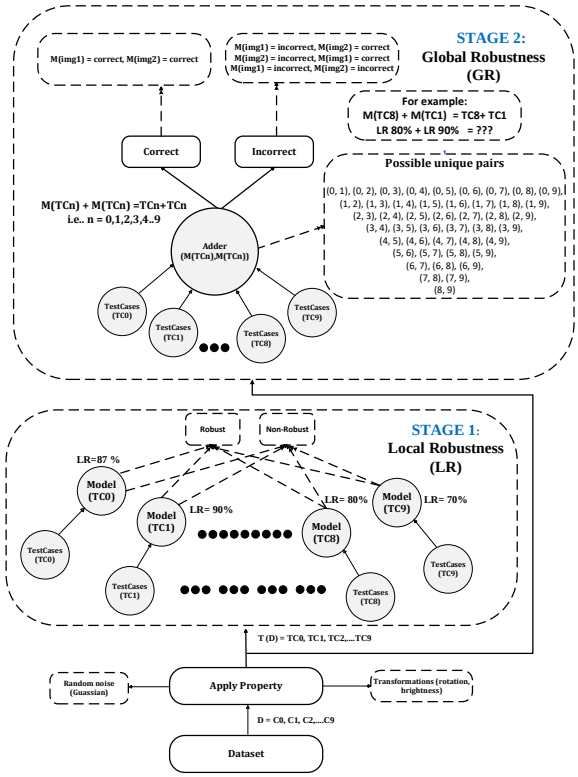

Fig. 2: Graphical View of Local and Global Robustness

- **Geometric Transformations:**
  - **Rotation:** Define $p_r(x_i, \theta)$ to rotate image x_i by angle θ:

$$x_i'' = p_r(x_i, \theta)$$

  - **Brightness Adjustment:** Define $p_b(x_i, b)$ to adjust the brightness of image x_i, where b modifies pixel intensity:

$$x_i''' = p_b(x_i, b)$$

C. Testing

The Testing section evaluates how accurately and confidently the model predicts under various properties (noise, rotation, brightness) applied to images from each class. This phase focuses on directly measuring and quantifying the robustness of the model.

- Confidence Level Assessment:
  - After generating test cases, measure the model's confidence for each class under each type of property.
  - Aggregate these measurements to assess the overall robustness of individual properties (noise, rotation, brightness).
- Local Robustness:
  - Each class c from the MNIST dataset undergoes individual tests where properties $p \in \{\text{noise}, \text{rotation}, \text{brightness}\}$ are applied:

$$x'_{c,p} = p(x_c, \theta_p)$$

    where θ_p represents the intensity or degree of property p applied to the image.
  - The model evaluates each modified input $x'_{c,p}$, and the outcome $\hat{y}_{c,p}$ is recorded to measure:

$$LR_{c,p} = P(\hat{Y}_{c,p} = Y_c | X_c = x'_{c,p})$$

    This probability is calculated over multiple trials to obtain a robust statistical measure of performance under each property.
- Global Robustness:
  - After applying the same property p to pairs of images (x_{c1}, x_{c2}) from different classes and testing them, the model's predictions $(\hat{y}_{c1,p}, \hat{y}_{c2,p})$ are compared against the actual class labels (y_{c1}, y_{c2}):

$$x'_{c1,p} = p(x_{c1}, \theta_p), \quad x'_{c2,p} = p(x_{c2}, \theta_p)$$

$$GR_{(c1,p),(c2,p)} = P(\hat{Y}_{c1,p} = Y_{c1} \wedge \hat{Y}_{c2,p} = Y_{c2} | X_{c1} = x'_{c1,p}, X_{c2} = x'_{c2,p})$$

D. Error Summarization

This section details the error summarization process following the global robustness testing, where we identify and analyze discrepancies in the model's predictions. By tracing errors from the combined outputs back to individual properties and classes, we systematically pinpoint and address underlying weaknesses.

1) Example Setup: We consider two classes from the MNIST dataset:

- Class A: Digit '5'
- Class B: Digit '0'

Both classes are tested under the properties of noise and rotation, with the following local robustness confidence levels:

- $LR_{A,p_1} = 85\%$ (Noise)
- $LR_{A,p_2} = 78\%$ (Rotation)
- $LR_{B,p_1} = 90\%$ (Noise)
- $LR_{B,p_2} = 88\%$ (Rotation)

2) Global Robustness Testing Scenario: The model is tasked with processing two images, one from each class, both subjected to noise. The ideal prediction should reflect the sum '5' (Class A) + '0' (Class B) = '5'. An incorrect sum indicates a prediction error.

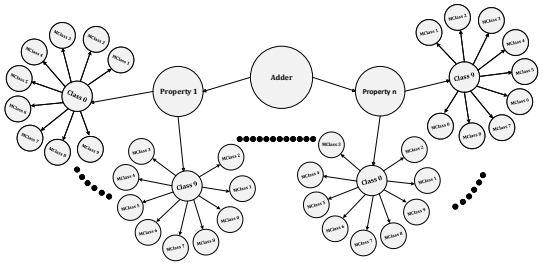

Fig. 3: Diagram of Error Summarization Highlighting Class-Property Impact

3) Error Summarization Process:

1) **Initial Error Detection:** The combined prediction incorrectly sums to '6' instead of '5'.
2) **Identify the Inaccurate Predictions:** Examination reveals the model predicts '6' for Class A and correctly predicts '0' for Class B.
3) **Drill Down to Property Level:** Both images were tested under noise. The noise robustness for each class is assessed:
   - $LR_{A,p_1} = 85\%$ — suggesting a 15% failure rate.
   - $LR_{B,p_1} = 90\%$
4) **Assess Individual Local Robustness:** Focus is placed on Class A's noise robustness, identifying potential misclassification trends under noisy conditions.
5) **Further Analysis:** Investigate if certain noise patterns consistently mislead the model concerning digit '5', potentially confusing it with a similar appearance to '6'.
6) **Systematic Error Identification:** Patterns of error involving Class A under noise are sought to determine if specific adjustments in model training or data handling can mitigate these issues.
7) **Propose Adjustments:** Modifications to the training process or data augmentation strategies are suggested to enhance noise robustness, especially for digits with appearances similar to '5'.

This error summarization framework enables a detailed understanding of the model's limitations and facilitates targeted improvements to enhance overall system robustness.

VII. Experiments

VIII. Threats to Validity

This section outlines significant limitations and assumptions in our study that may affect the validity and reliability of our findings.

- **Random Sampling:** Our current approach assumes a uniform distribution of samples across all classes, which may not represent the true complexity and variability within real-world data. This uniform sampling can lead to biased evaluations if the class distribution in practical applications is skewed or non-uniform. We plan to enhance our sampling techniques to better capture the diversity and distribution of data in realistic scenarios. Improved sampling strategies will help in developing more robust and generalizable error summarization methods.

IX. Related Work

X. Conclusion

References

[1] Reference details