

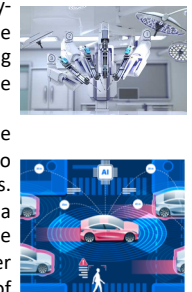
# Antonio Guerriero

Tutor: Prof. Stefano Russo – Co-Tutor Prof. Roberto Pietrantuono  
XXXIV Cycle – II year presentation

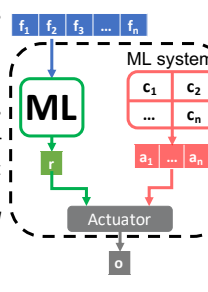
## Testing of ML systems, the oracle problem

### 1. Context

The increasing trend in adopting *Machine Learning* (ML) in safety-critical contexts, such as in the medical and autonomous driving domains, makes testing of these systems a great concern. *Testing* is a crucial activity in the development of such systems to avoid catastrophic failures. Currently, a tester has to front a strong effort to evaluate the performance of an ML system under test, due to the manual labeling of operational data.



An ML system takes a *feature vector* ( $f_i$ ) as input; an internal component uses an ML algorithm to compute a *response* ( $r$ ), while *other components* ( $c_i$ ), not based on ML, produce *additional outputs* ( $a_i$ ) which, combined with  $r$ , yield the ultimate *output*  $o$ .



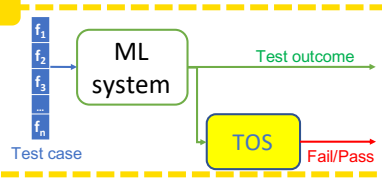
### 2. Open problem

- The “oracle problem”: there is no reliable “test oracle” to indicate what the correct output should be for arbitrary input
- Traditional software testing techniques are not adequate to test Machine Learning programs

We need a *test oracle surrogate (TOS)* for ML systems, able to automatically classify tests’ outcome (Fail/Pass), so as to obtain feedback about tests whose expected output is unknown.

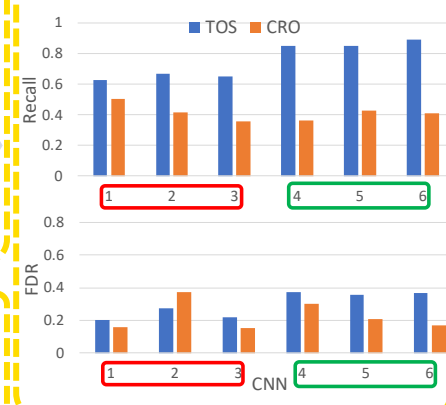
### 1. Proposal

Automatic detection of *failed tests*, based on three sets of *invariants* representing the expected behavior of the system under test, so that each time an *invariant* is violated a failure is said to have occurred.



### 4. Results

- Experiments with six Convolutional Neural Networks (CNNs) and two popular datasets:
  - MNIST (handwritten digits)
  - CIFAR 10 (color images)
- CNNs 1, 2, 3 trained on 28,000 images of the MNIST dataset.
- CNNs 4, 5, 6 trained on 24,000 images of the CIFAR10 dataset.
- For all CNNs, the test set consists of 2,500 images.
- The remaining images in the datasets are considered unlabeled arbitrary inputs.
- Evaluation: TOS is compared to a *Cross-Referencing Oracle (CRO)*, a majority oracle:
  - This oracle considers as the correct output, the most voted value.
  - When the three responses are different, the oracle refrains to make a decision.



### 2. Invariants definition

Input data invariants

Training data invariants

ML algorithm invariants

A set of invariants capturing both domain-related conditions that should never be violated and testing assumptions. Each output violating such rules is judged as a failure.

A set of invariants inferred from the prior knowledge in the training set. The invariants should be verifiable by the tester and able to identify a subset of failures preserving high accuracy.

A set of invariants about how the ML algorithm produces an incorrect response. The assumption is that failures have similar patterns that are not available in the training set, and that are specific to the adopted algorithm.

Implemented via manual coding

A decision tree is used to extract the invariants from the training set. All invariants with low confidence and low support are filtered out

Random Forest is used to extract patterns from a custom training set made with failure examples of the ML system under test

### 3. Evaluation Metrics

- **Recall**: defined as  $\text{True Positives} / (\text{True Positives} + \text{False Negatives})$ , namely the percentage of failures correctly detected with respect to the set of detections.
- **False Discovery Rate (FDR)**: defined as  $\text{False Positives} / (\text{True Positives} + \text{False Positives})$ , it represents the percentage of false positives with respect to the set of detections.

### Research Group

I am a member of DESSERT research group at DIETI – UNINA.



### Cooperation

Professor Michael R. Lyu, Chairman of Department of Computer Science and Engineering at Chinese University of Hong Kong.



### Future developments

- Define a complete testing strategy to evaluate ML systems.
- Fine-tune TOS in order to find the best trade-off between Recall and FDR.
- Investigate the effective interpretability of detections by TOS.
- Extend TOS to further ML application fields (beside image classification).