Do Machine Learning Models Learn Statistical Rules Inferred from Data?

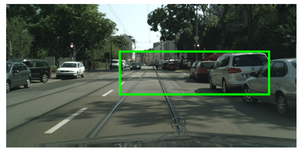
Aaditya Naik, Yinjun Wu, Mayur Naik, Eric Wong
asnaik@seas.upenn.edu

Try out our code!

Motivation

Models make mistakes!

Object Detection



Sentiment Analysis

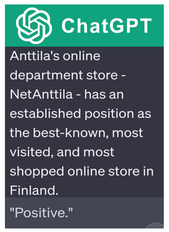


Image Classification

Sea Slug

Observation:

Fundamental errors defy rules based on human intuition

Challenges:

- Characterizing such rules
- Generating rules at scale
- Using the generated rules

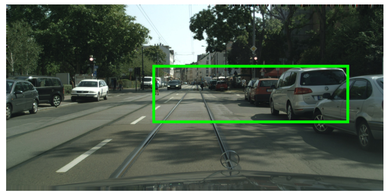
Statistical Quantile Rules (SQRs)

Rules for estimating errors must be

- **Valid** and hold over a large portion of the data,
- **Expressive** to capture complex and interesting patterns,
- **Scalable** to generate several rules without supervision.

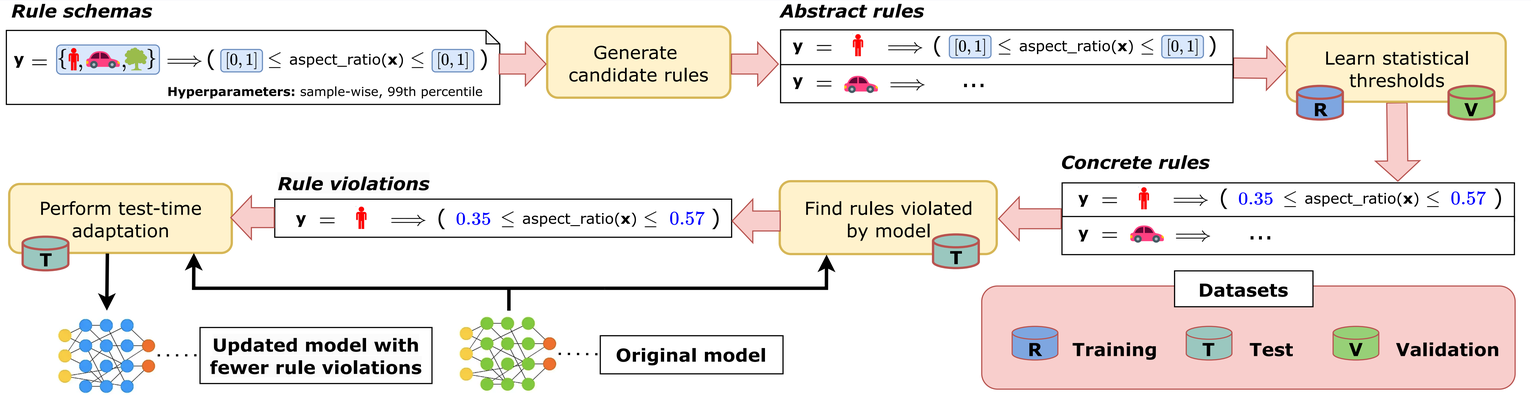
The answer is **Statistical Quantile Rules:**

$$\mathbb{P}(a \leq \phi(x) \leq b) = 1 - \delta$$



An example of a violation of the rule $0.07 \leq \text{aspect\_ratio}(car) \leq 2.77$. 98% of all the ground-truth cars satisfy this rule, making it valid, expressive and scalable.

SQRL: Synthesizing Statistical Quantile



The workflow of SQRL. SQRs are generated from the training and validation data and used to evaluate and improve models.

Examples of Rule Violations

Object Detection

Rule
$\text{car}(x) \Rightarrow 20.22 \leq \text{width}(x) < 1655.17 \wedge \text{aspect\_ratio}(x) < 0.81 \vee \ldots$
If an object is a **car**, then its aspect ratio and width must lie in one of the bounds defined by the rules above.
Original Prediction

Prediction after Test Time Adaptation


Sentiment Analysis

Rule
$\text{neutral}(x) \Leftarrow 0.02 \leq \text{fitness}(x) \leq 0.02 \wedge 0.04 \leq \text{news}(x) \leq 0.14$
If the probability that the sentence is about fitness and news is within the above bounds then the sentiment of this sentence is **neutral.**
Original Prediction

Anttila's online department store - NetAnttila - has an established position as the best-known, most visited, and most shopped online store in Finland.
`fitness_and_health: 0.0210`
`news_and_social_concern: 0.077`
`predicted label: positive (wrong)`

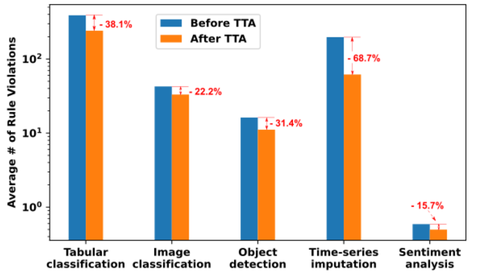
Prediction after Test Time Adaptation

Anttila's online department store - NetAnttila - has an established position as the best-known, most visited, and most shopped online store in Finland.
`fitness_and_health: 0.0210`
`news_and_social_concern: 0.077`
`predicted label: neutral (correct)`

Using SQRs

Task	# Rules		
	Total	Selected	Violations per sample
Tabular Classification	292,129	400	389.29
Image Classification	73,032	340	42.62
Object Detection	252	252	16.21
Time-Series Imputation	35	35	197.46
Sentiment Analysis	158	158	0.59

Number of SQRs generated by SQRL.



Results of Rule-based Test-Time Adaptation

Conclusions

Formalized SQRs to characterize and identify basic errors at scale and proposed the SQRL framework.

SQRL can find up to 300K rules and up to 158K violations.

We find that models do not always learn statistical rules but can be adapted to correct up to 68.7% rule violations.