SAILON Evaluation Metrics: Rigorously Defined

Sam Showalter

April 28, 2021

1 Experimental Setup

The SAILON evaluation system is abstractly defined as follows. *A priori*, a number of trials T is defined, where each trial $t \in T$ contains a single batch b of n samples. The total number of samples across the entire evaluation is denoted $Tn = N$.

Each sample $\mathbf{x}$ is a video of one of k possible actions, or an action unseen during training time as denoted the $k+1$ novel class. These samples are independently fed through a network and novelty detection system, considered jointly in this setup as $f_\theta(.)$. In this setting, the output for any sample $\mathbf{x_i}$ fed through $f_\theta(.)$ is a scalar label $y_{\theta i} \in \{0, 1, ...k, k+1\}$ where $k+1$ represents a novel input sample.

In addition, there is a third metric set *a-priori* on the trials of the evaluation. At an unspecified trial $t_a \in T$, the presence of novel instances begins and will remain for all subsequent trials. Before trial t_a, no novel samples are present, though the model $f_\theta(.)$ can at any time predict $y_{\theta i} = k+1$.

The evaluation is run sequentially in a batch-wise fashion over trials. Accordingly, $\mathbf{x}_{ti}$ represents the i^{th} sample of the t^{th} batch, and y_{ti} represents the corresponding label. In addition, a baseline model $f_\phi(.)$, whose predictions are denoted $\hat{y}_{\phi ti} \in \{0, 1, ..., k\}$, cannot predict novelty and is often referred to as the **baseline model**. All evaluation metrics are determined based on the relationship of predicted labels $\hat{y}_{\theta ti}$ to the ground-truth labels y_{ti} of the evaluation and to the baseline predictions $\hat{y}_{\phi ti}$.

Implicitly, if a given trial t at least one sample in its batch b_t was predicted to be novel ($\hat{y}_{\theta ti} = 1$), the trial itself is tagged with binary label 1 for its batch $\hat{l}_t = 1$, signifying the agent believes novelty has started. After the first trial possessing novelty occurs, all future trials will be given a novelty label = 1 regardless of the predictions in its batch. The ground truth labels across trials is denoted l_t

2 Trial-based Evaluation Metrics: Detection Performance

First, trial-based evaluation metrics are evaluated, with an optimal performance defined as predicting novelty for the first time in batch b_{ta} belonging to trial $t_a \in T$. Failure modes in terms of false positives and false negatives are defined as follows, where t represents the t^{th} trial and $\hat{l}_t$ represents the binary label defining if any samples in batch b_t were classified as novel.

$$\begin{aligned} TP &= \hat{l}_t = l_t \\ FP &= \hat{l}_t > l_t \\ FN &= \hat{l}_t < l_t \end{aligned} \tag{1}$$

These metrics are then utilized to generate slightly more sophisticated evaluation metrics across trials, including:

$$
\begin{aligned}
\text{Correctly Detected Trials (CDT)} &= \sum_{t=1}^{T} \mathbb{1}((FP_t = 1)\ \&\ (TP_t >= 1)) \\
\text{Trial Failures} &= (t_a - \arg\min_t(l_t = 1)) \\
\text{False Positive Trial \%} &= \frac{FP}{T} \\
\text{False Negative Trial \%} &= \frac{FN}{T}
\end{aligned}
\tag{2}
$$

In summary, the primary goal of this family of evaluation metrics is to minimize the number of false positives. This is far more important to PAR than the other metrics. Trial Failures exclusive looks at how many trials too early you declared novelty.

3 Sample-based Evaluation Metrics: Reaction Performance

The next family of evaluation metrics concerns sample-based evaluation metrics. These metrics conducted over batches (unless specified otherwise) and then averaged. The evaluation metrics are defined below.

$$
\begin{aligned}
FP &= (\hat{y}_{\theta ti} = k+1)\ \&\ (y_{ti} \neq k+1)) \\
FN &= (\hat{y}_{\theta ti} \neq k+1)\ \&\ (y_{ti} = k+1))
\end{aligned}
\tag{3}
$$

These can be computed in any way, but in particular there are several metrics that compare these FP, FN values in the pre- and post-novelty regime. The formulas are the same for the baseline model, but θ is replaced by ϕ.

$$
\begin{aligned}
\text{Accuracy} &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\hat{y}_{\theta ti} = y_{ti}) \\
\text{Baseline Accuracy} &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\hat{y}_{\phi ti} = y_{ti})
\end{aligned}
\tag{4}
$$

Per trial asymptotic performance is defined as the "measure in the change in accuracy over post-novelty rounds". No notation or formulae of any kind accompany this statement. However, some additional metrics include:

$$
\begin{aligned}
\text{Novelty Response Performance (NRP)} &= \frac{(\text{Accuracy}|t \geq t_a)}{(\text{Accuracy}|t < t_a)} \\
\text{Overall Performance Task Improvement (OPTI)} &= \frac{(\text{Accuracy}|t \geq t_a)}{(\text{Baseline Accuracy}|t \geq t_a)} \\
\text{Asymptotic Performance Task Improvement (APTI)} &= \frac{(\text{Accuracy per Trial}|t \geq t_a)}{(\text{Baseline Accuracy per trial}|t \geq t_a)} \\
\text{Area Under Activity Monitor Operating Characteristic (AUAMOC)} &= \text{AUROC of TP and FP}
\end{aligned}
\tag{5}
$$

In words, **NRP** is the ratio of the average batch accuracy in batches after novelty over the average batch accuracy in batches before novelty - we want this as high as possible. **OPTI** is the ratio of post-novelty accuracy in our model over post-novelty accuracy of the baseline. We also want this to be as high as possible. **APTI** is a more granular **OPTI** that makes use of per-trial differences. **AUAMOC** is a simple AUROC curve plotted for global TP and FP, **threshold based on anomaly score**

4 Oustanding items

1. Top-k classification with anomaly score not a probability - how do they intend to do this?