AI Explainability WG

November 2023

November 2023 updates

- TrustyAI service
  - 0.7.0 release
- TrustyAI core
  - Language performance metrics
- TrustyAI operator
  - 1.12.0 release
- Open Data Hub
  - 0.6.1 and 1.11.1 available in ODH 2.3

TrustyAI core / service

TrustyAI core / service

0.7.0 - What’s new?

- Available on GitHub and Quay.io
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.7.0
  - quay.io/trustyai/trustyai-service:v0.7.0
- Dependency updates
  - Major: Base Quarkus updated from 2.13 to 3.2
    - Numerous improvements and bug fixes
  - Minor: gRPC update
- Tensor conversion improvements
  - Support for multiple-tensor output and n-dimensional input
  - Fix upload of Nx1 tensors in upload endpoint
  - Raw-converted floats little-endian

TrustyAI core / service

Language performance metrics

- Word Error Rate (WER)
  - Percentage of errors at the word level compared to a reference text
- Bilingual Evaluation Understudy (BLEU)
  - Quantify similarity between model text output to a set of high-quality references
- Exact Match
  - Correctness metric for (fuzzy) exact matches

TrustyAI core / service

Language performance metrics

- Word Error Rate (WER)

  -
    ```
    reference =  "the quick brown fox  jumps over the lazy dog"
    hypothesis = "the quick brown dogs jump  over the lazy dog"
    WER(reference, hypothesis) # 0.22 or a 22% error rate.
    ```
- Bilingual Evaluation Understudy (BLEU)

  -
    ```
    references = [
         "The quick brown fox jumps over the lazy dog",
         "A fast brown fox leaps over the lazy dog",
         "Quick brown fox jumps over the lazy dog"]
    hypothesis = "The quick brown fox jumped over the lazy dog"
    BLEU(references, hypothesis) # 0.80 match to multiple references
    ```

TrustyAI operator

TrustyAI operator

1.12.0 - What’s new?

- Available on GitHub and Quay.io
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.12.0
  - quay.io/trustyai/trustyai-service-operator:v1.12.0
- Dependency updates
  - gRPC, HTTP
- Implement lifecycle events
  - Example emitted Kubernetes events
    - `PVCCreated`
    - `InferenceServiceConfigured`
    - `ServiceMonitorCreated`

TrustyAI operator

In progress

- OAuth authentication for external endpoints
    - `TOKEN=$(oc whoami -t)`
    - `curl -H “Authorization: Bearer ${TOKEN}” https://trustyai-service…`
- Support for self-signed certificates

TrustyAI operator

1.11.1 - Available now in ODH 2.3

- TrustyAI as a component of DataScienceCluster

```
apiVersion:
datasciencecluster.opendatahub.io/v1alpha1
kind: DataScienceCluster
metadata:
  name: default
spec:
  components:
    dashboard:
      managementState: Managed
    modelmeshserving:
      managementState: Managed
    trustyai:
      managementState: Managed
```

Community

Documentation

- New tutorials available
  - https://github.com/trustyai-explainability/odh-trustyai-demos
- Walkthrough available
  - https://opendatahub.io/docs/monitoring-data-science-models/#enabling-trustyai-service-cli_monitor-models

Roadmap

TrustyAI 2023 roadmap

July 2023

- *Explainers*
  - Support for explainers LIME, SHAP, CF at service level
- *Metrics*
  - Flexible scheduling/batching
  - Improve service metadata endpoints
    - Include available categories
- *Operator*
  - TrustyAI Operator v1
- *Explainers*
  - Support for external explainability libraries

September 2023

- *Storage*
  - Wider storage support (database backends)
- *Metrics*
  - Additional metrics
  - Metrics statistical tests
- KServe integration
- Drift detection
- ODH v2 onboarding

December 2023

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
  - Language metrics (WER, BLUE, EM)
- *Detection*
  - HAP/PII
- *Metrics*
  - Support for user-defined historical windows