AI Explainability WG

April 2024

April 2024 updates

- TrustyAI core / service
  - 0.12.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.12.0
  - quay.io/trustyai/trustyai-service:v0.12.0
- TrustyAI operator
  - 1.18.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.18.0
  - quay.io/trustyai/trustyai-service-operator:v1.17.0

🚀 0.13.0 and 1.19.0 will be released this Friday (19/04/2024)

What’s new?

TrustyAI - What's new?

- **TrustyAI core/service 0.12.0**
  - Remove full endpoint and related methods [#514]
  - gRPC
    - Fix tensor name, add gRPC mock server. [#520]
    - Persist synthetic tag when tensor parameter present [#528]
    - Add rouge metric [#526]
    - Properly shutdown channel after prediction request [#531]
    - Get tensor data based on non-empty contents, instead of data type [#534]
  - Fix internal data being read whole when reading dataframe in batches [#538]
  - Service data batches for metrics calculations should account for synthetic data in its size [#540]
  - Security
    - Update Apache Commons Compress version [#541]
    - Update Quarkus from 3.2.9.Final to 3.2.11.Final [#529]
- **TrustyAI operator 1.18.0**
  - [CVE] Update sdk-go version

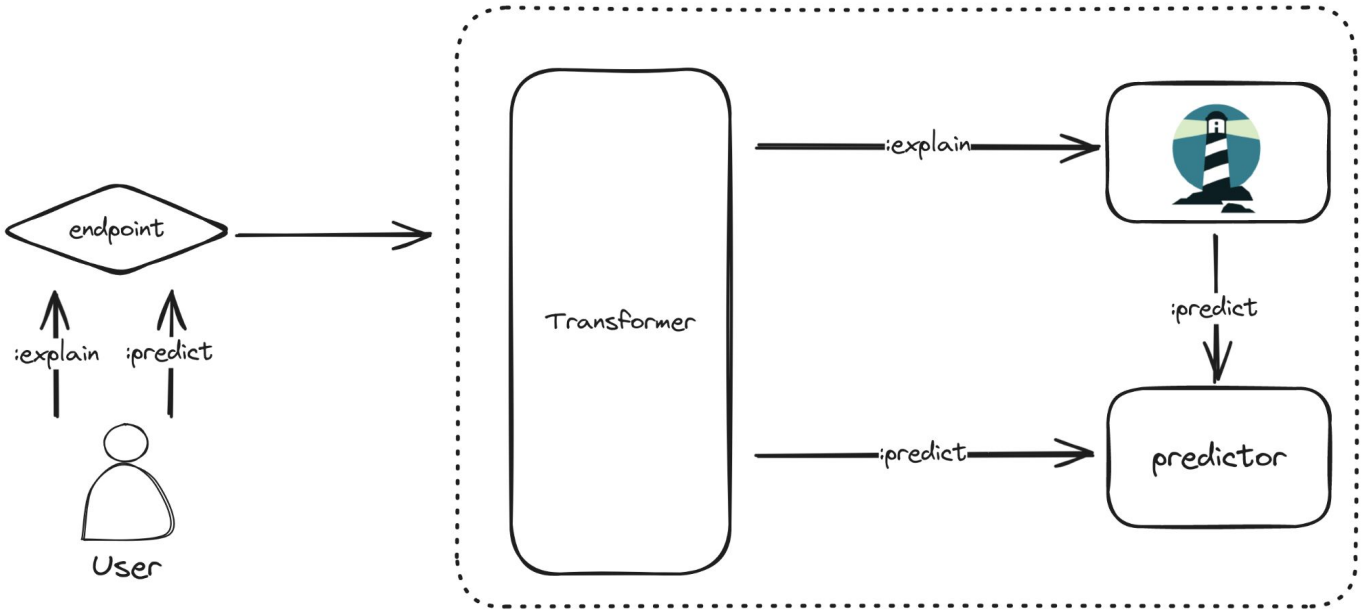
[1]: https://github.com/trustyai-explainability/trustyai-explainability-python-examples/blob/main/examples/DataDrift.ipynb

Current work

TrustyAI - current work

- **TrustyAI core/service**
  - Support for database storage
  - KServe v1 HTTP model abstraction
  - New Feature type: Tensor
- **TrustyAI operator**
  - TLS-enabled Kubernetes `Service` for payload consumer
  - Fixes to TrustyAIService finalizer
- **TrustyAI KServe explainer**
  - Expand PoC to add user configuration

KServe explainer



KServe explainer demo

Predictor only

```
apiVersion: "serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "explainer-test"
  annotations:
       sidecar.istio.io/inject: "true"
       sidecar.istio.io/rewriteAppHTTPProbers: "true"
       serving.knative.openshift.io/enablePassthrough:
"true"
spec:
  predictor:
       model:
         modelFormat:
         name: sklearn
       protocolVersion: v2
       runtime: kserve-sklearnserver
       storageUri:
https://github.com/ruivieira/model-collection/raw/main/c
redit-score/model.joblib
```

Predictor and explainer

```
apiVersion: "serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "explainer-test"
  annotations:
       sidecar.istio.io/inject: "true"
       sidecar.istio.io/rewriteAppHTTPProbers: "true"
       serving.knative.openshift.io/enablePassthrough:
"true"
spec:
  predictor:
       model:
         modelFormat:
         name: sklearn
       protocolVersion: v2
       runtime: kserve-sklearnserver
       storageUri:
https://github.com/ruivieira/model-collection/raw/main/c
redit-score/model.joblib
  explainer:
       containers:
       - imagePullPolicy: Always
       name: explainer
       image: quay.io/ruimvieira/trustyai-kserve:latest
       command:
       - /opt/jboss/container/java/run/run-java.sh
```

Community

TrustyAI - community

- **GitHub discussions**
  - https://github.com/orgs/trustyai-explainability/discussions

Roadmap

TrustyAI 2024 roadmap

- **December 2023 - March 2024 (proposal / discussion)**
  - Out-of-distribution (OOD) metrics completion
    - Finalise OOD metrics, namely data uploading aspect
      - Provide data connection for data upload
  - Explainability service endpoints completion
    - Formalise explainability payloads schema
    - Refining handling of synthetic payloads to avoid interference with metrics
    - Handle potential large computational times in a service setting
  - Detoxification at the library and service level
    - Integrate detoxification with Python TrustyAI (Jupyter main target)
    - Token scoring at the service level
  - Database backend
    - Address scalability
    - Replace PVC with DB?
  - Expand supported types (*eg* image data)
    - Metrics and explainability for non-tabular data
  - Model drift/data drift/anomaly detection
  - Improve handling of unsupported model serving runtimes

Legend

Not started

In progress

Completed

TrustyAI 2024 roadmap

- **March 2024 - May 2024 (proposal / discussion)**
  - KServe explainer integration
  - Detoxification fine-tuning

Legend
Not started
In progress
Completed