AI Explainability WG

February 2024

February 2024 updates

- TrustyAI core / service
  - 0.10.1 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.10.1
  - quay.io/trustyai/trustyai-service:v0.10.1
- TrustyAI operator
  - 1.16.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.16.0
  - quay.io/trustyai/trustyai-service-operator:v1.16.0
- TrustyAI Python
  - 0.5.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability-python/releases/tag/0.5.0
  - https://pypi.org/project/trustyai/0.5.0/

What’s new?

TrustyAI - What's new?

- **TrustyAI core/service 0.10.1**
  - Expand missing metadata error message during metric creation.
  - [CI] Update TrustyAI service CR
- **TrustyAI operator 1.16.0**
  - Refactor deployment to use templates
  - Rename TrustyAI `apiGroup` to `trustyai.opendatahub.io`
  - Add missing labels to OAuth service
  - Add unique ClusterRoleBinding names
  - Change metrics service pod target to 8080
  - [CVE] CVE-2023-37788, CVE-2022-21698
- **Python TrustyAI 0.5.0**
  - Add HAP/detoxify module to Python TrustyAI
  - [Docs] Fix readthedocs generation

Python TrustyAI detoxify

- Install from PyPi
  - `pip install trustyai[detoxify]`
- Example Jupyter notebooks
  - https://github.com/trustyai-explainability/trustyai-explainability-python-examples/blob/main/examples/Detoxify.ipynb
- T-MaRCo is an extension of
  - *Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts*
    - makes it possible to use multiple combinations of experts and anti-experts to score and (incrementally) rephrase texts generated by LLMs.
  - *Towards Mitigating Hallucination in Large Language Models via Self-Reflection and N-Critics: Self-Refinement of Large Language Models with Ensemble of Critics*
    - integrate rephrasing with the base model self-reflection capabilities
- Features
  - content scoring: providing a disagreement score for each input token; high disagreement is often attached to toxic content.
  - content masking: providing a masked version of the input content, where all tokens that are considered toxic are replaced with the <mask> token.
  - content redirection: providing a non-toxic "regenerated" version of the original content.
- Pretrained models
  - https://huggingface.co/trustyai/gminus
  - https://huggingface.co/trustyai/gplus

Current work

TrustyAI - current work

- **TrustyAI core/service**
  - Support for database storage
- **TrustyAI operator**
  - TLS-enabled Kubernetes `Service` for payload consumer
- **Python TrustyAI**
  - Migrating core's bias metrics validation suite to Python
  - Decouple of visualisations to separate module
  - Support data transfer between service and Python
    - *e.g.* retrieving calculated metrics from service
- **CI**
  - Migration of CI to ODH v2

Integrated Gradients

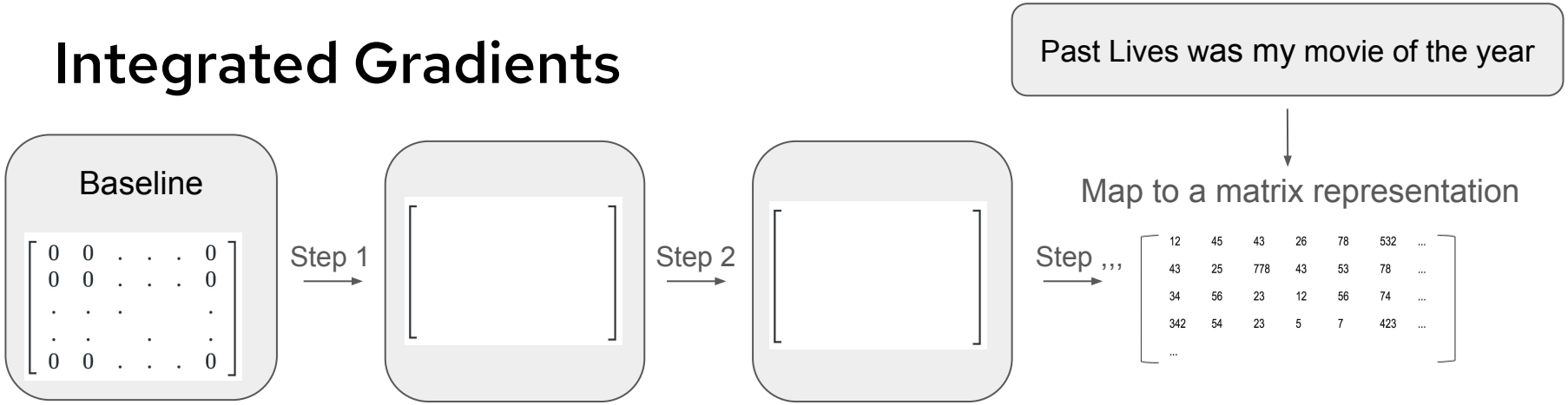
Integrated Gradients

Given an LLM, how can we measure the contribution of each of the input tokens on its output ?

ex) Sentiment Classification

```
Review: Wow. What a wonderful film. The script is nearly perfect it appears this is the only film written by Mingl
un Wei,I hope he has more stories in him.<br /><br />The acting is sublime. Renying Zhou as Doggie was amazing --
very natural talent, and Xu Zhu was a delight - very believable as the jaded old traditionalist. <br /><br />The s
oundtrack was very effective, guiding without being overwhelming. <br /><br />If only more movies like this were m
ade whether in Hollywood or Hong Kong- a family friendly, well acted, well written, well directed, near perfect ge
m.
```

Integrated Gradients



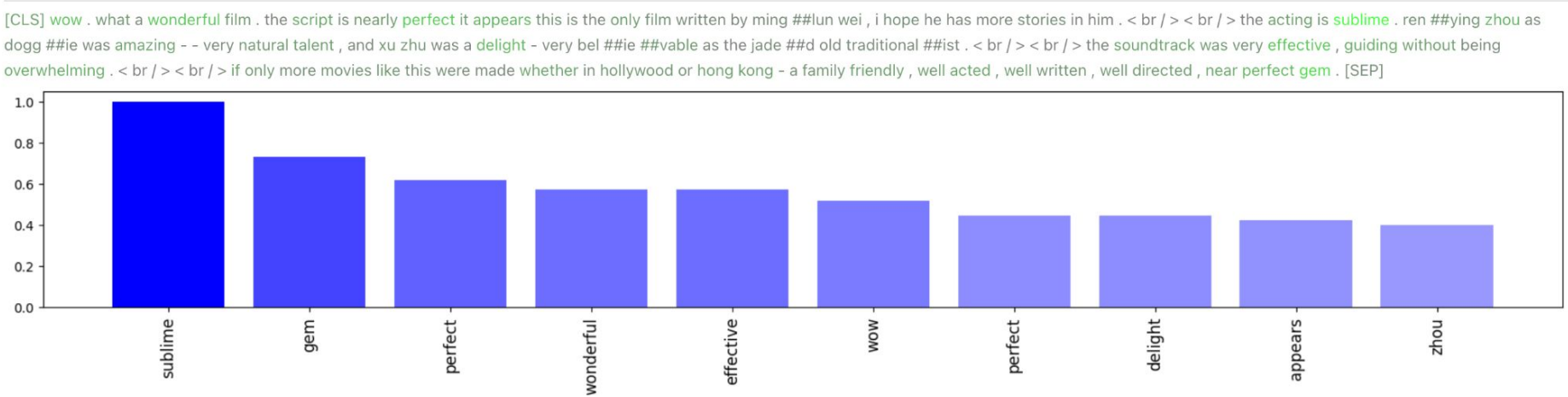
1. Start with a baseline i.e an input that has no effect on the model's prediction

2. At each step, linearly interpolate the baseline so it gradually gets closer to the original input. Get the difference between the model prediction at current input and baseline

3. Sum up the differences at each step to get integrated gradients. Results are interpreted as the feature importances of each token

Current Progress

- Implemented IG to be compatible with any LLMs under HF's AutoModelforSequenceClassification and AutoModelforMaskedLM classes
- Added tools to visualize the feature importance of each token and plot the top k tokens



Next steps

- IG has been shown to violate important attribution properties
  - Assigns two features that have different effects on the model prediction the same score
  - Scores features that have no effect on the model prediction positively
- Find a better path from the baseline to the original input such that these properties are satisfied

Community

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