Что нового в интерпретируемости ML-моделей

Дмитрий Колодезев, Промсофт
Data Fest 3.0 — Reliable ML
05.06.2022

Терминология

- Interpretability — характеризует нашу способность понять, как модель работает
  - Крайний случай — карты активации нейронов
- Explainability — характеризует нашу способность объяснить, как был получен конкретный результат
  - Крайний случай — самообъяснения модели
- Обычно используют вперемешку, не вдумываясь
- Не буду нарушать традицию

Шпаргалка

- Таблички:
  - Shap: значимость признаков для предсказания
  - Shap: значимость признаков для функции потерь
  - InterpretML: швейцарский нож
  - Нарисуйте сначала Mean Target Plot
  - Загляните в https://christophm.github.io/interpretable-ml-book/
- Картинки:
  - GradCAM и его друзья
  - Captum: все для PyTorch
- Трансформеры
  - BertViz например (но я не настоящий сварщик)

Основные сценарии

- Отладка / разбор инцидентов
  - SHAP
  - Влиятельные сэмплы
- Дымовой тест / верификация
  - SHAP
  - Контрпримеры
  - Лучшие и худшие точки
- Социализация (кооперативность)
  - Глобальные суррогатные модели
  - Якорные примеры

Основные тренды

- Ученые изобрели интерпетируемые модели! Используем только их (нет)
- Ученые выяснили, что интерпетируемость бесполезна (нет)
- Регуляторы требуют объяснений (нарисуем)
- Библиотека X все объясняет и думать не надо (нет)
- Модель научилась объяснять сама себя

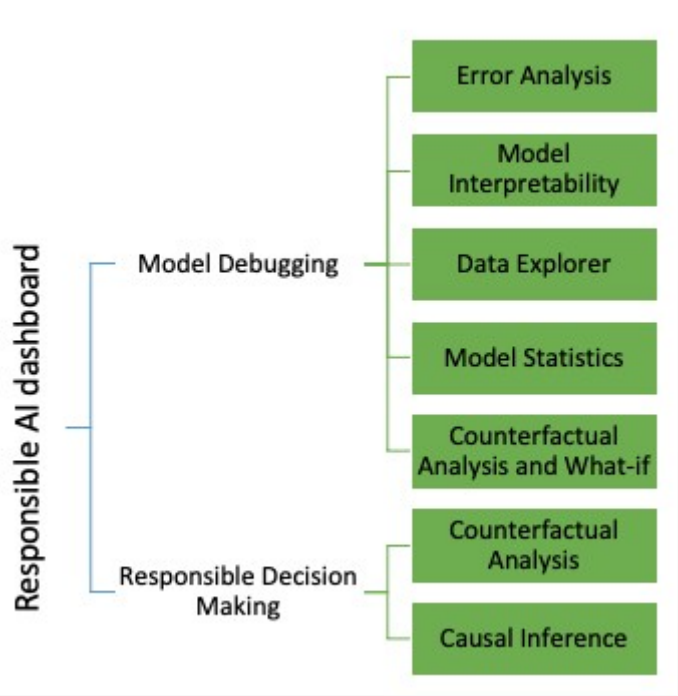
  (сочинять правдоподобное объяснение, как и люди)

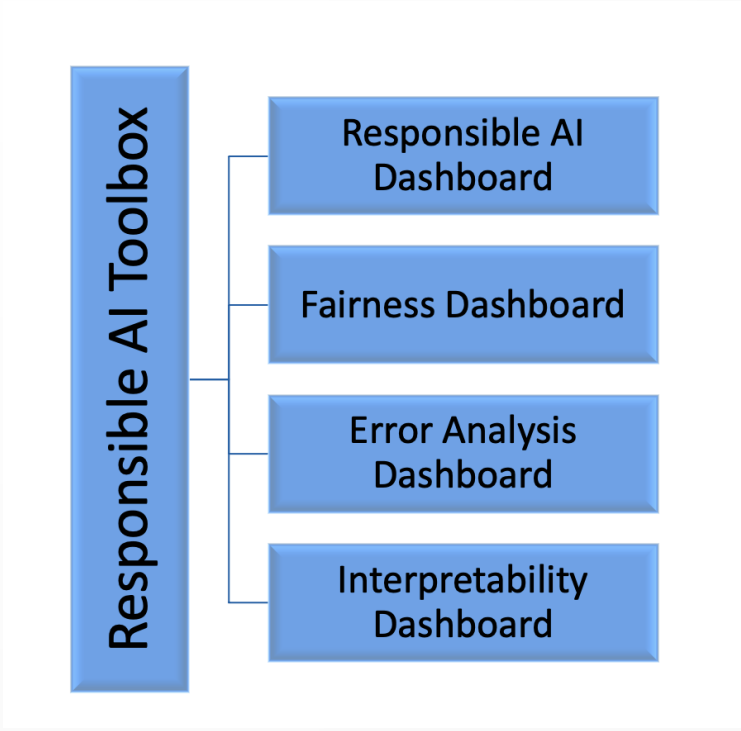
RAI Toolbox - вечно новое

Identify	Diagnose	Mitigate
What kinds of issues does my model have? In what areas are errors most prevalent?	Why does my model have these issues? What model decisions create the errors? Where should I focus my resources to improve my model?	How can I improve my model? What social or technical solutions exist for these issues?

- https://github.com/microsoft/responsible-ai-toolbox
- Для табличек — лучшее
- Публичная часть внутренних сервисов Azure

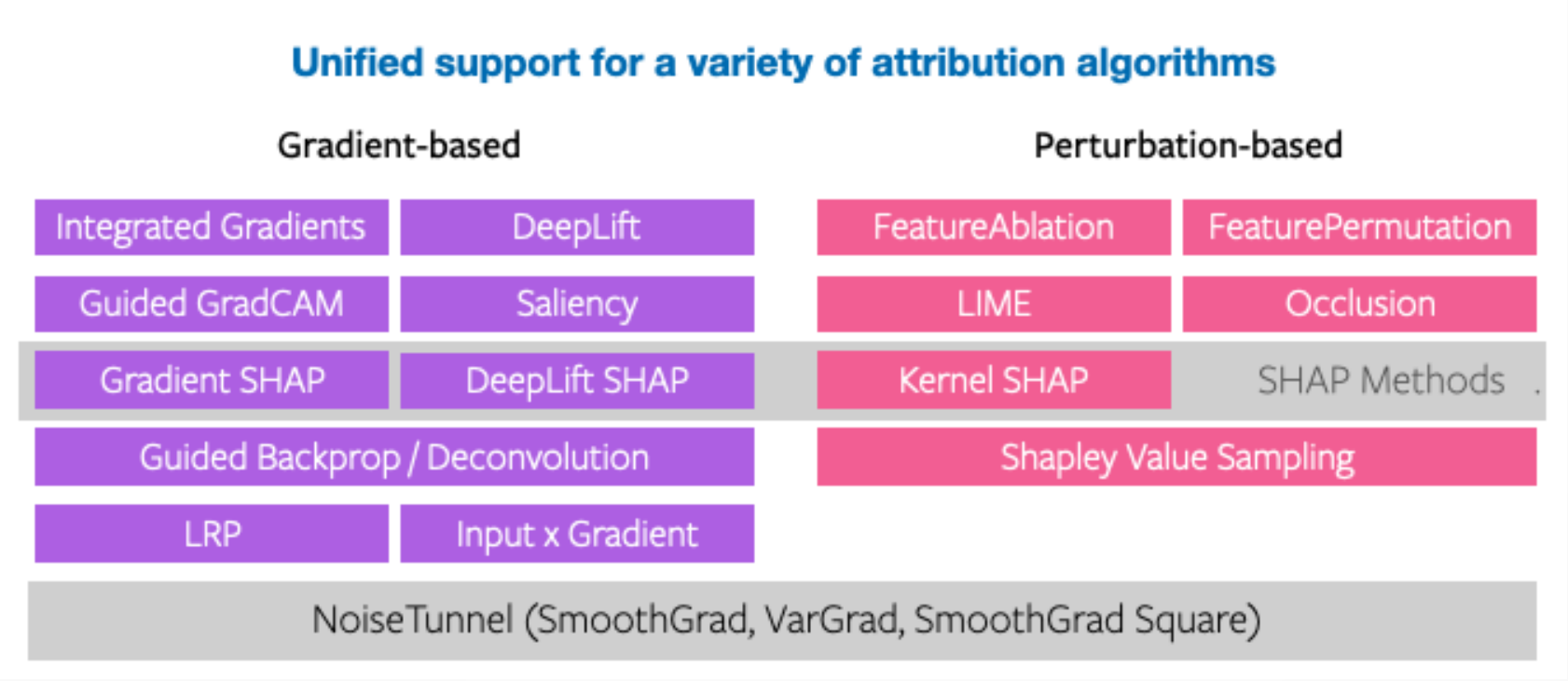
RAI Toolbox — весь ReliableML





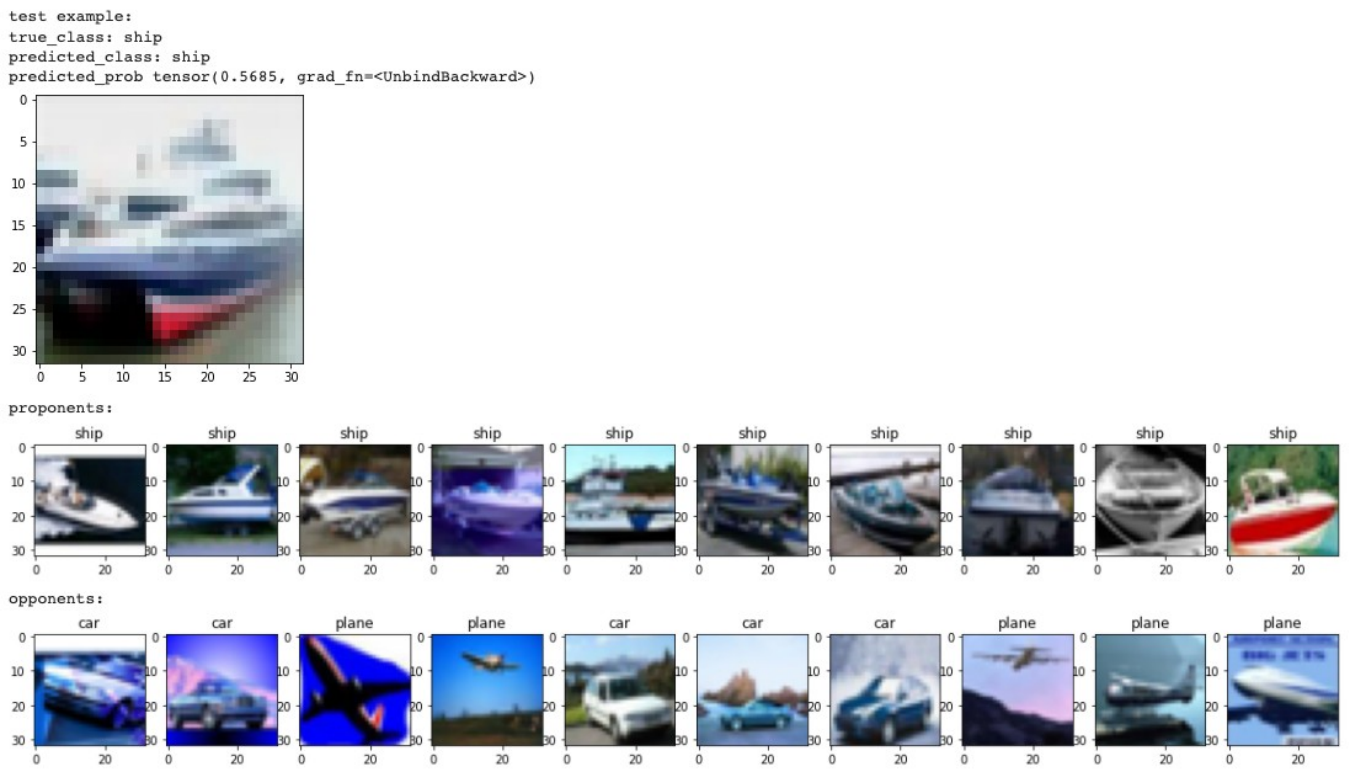
https://github.com/microsoft/responsible-ai-toolbox

Captum — всё для Pytorch



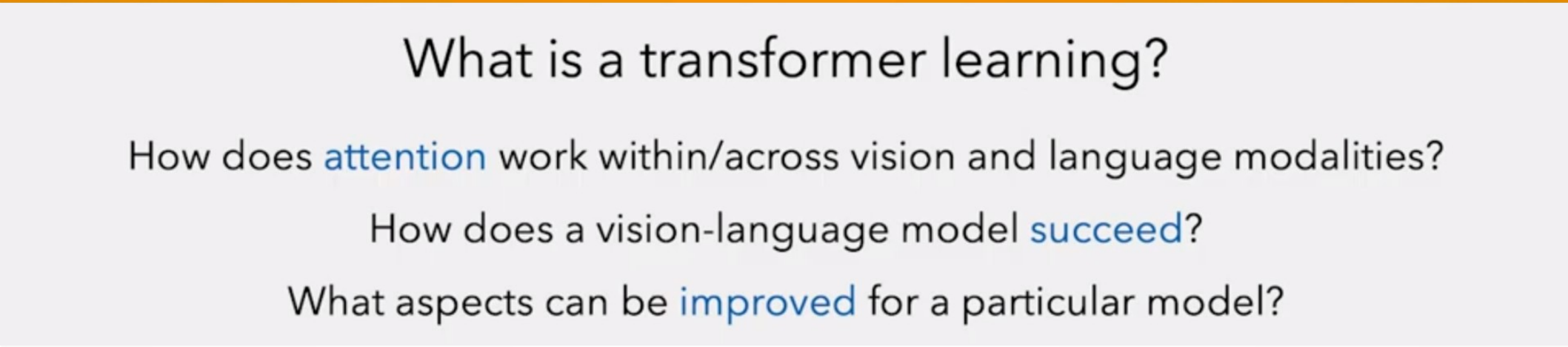
https://github.com/pytorch/captum

Captum - Influential Examples



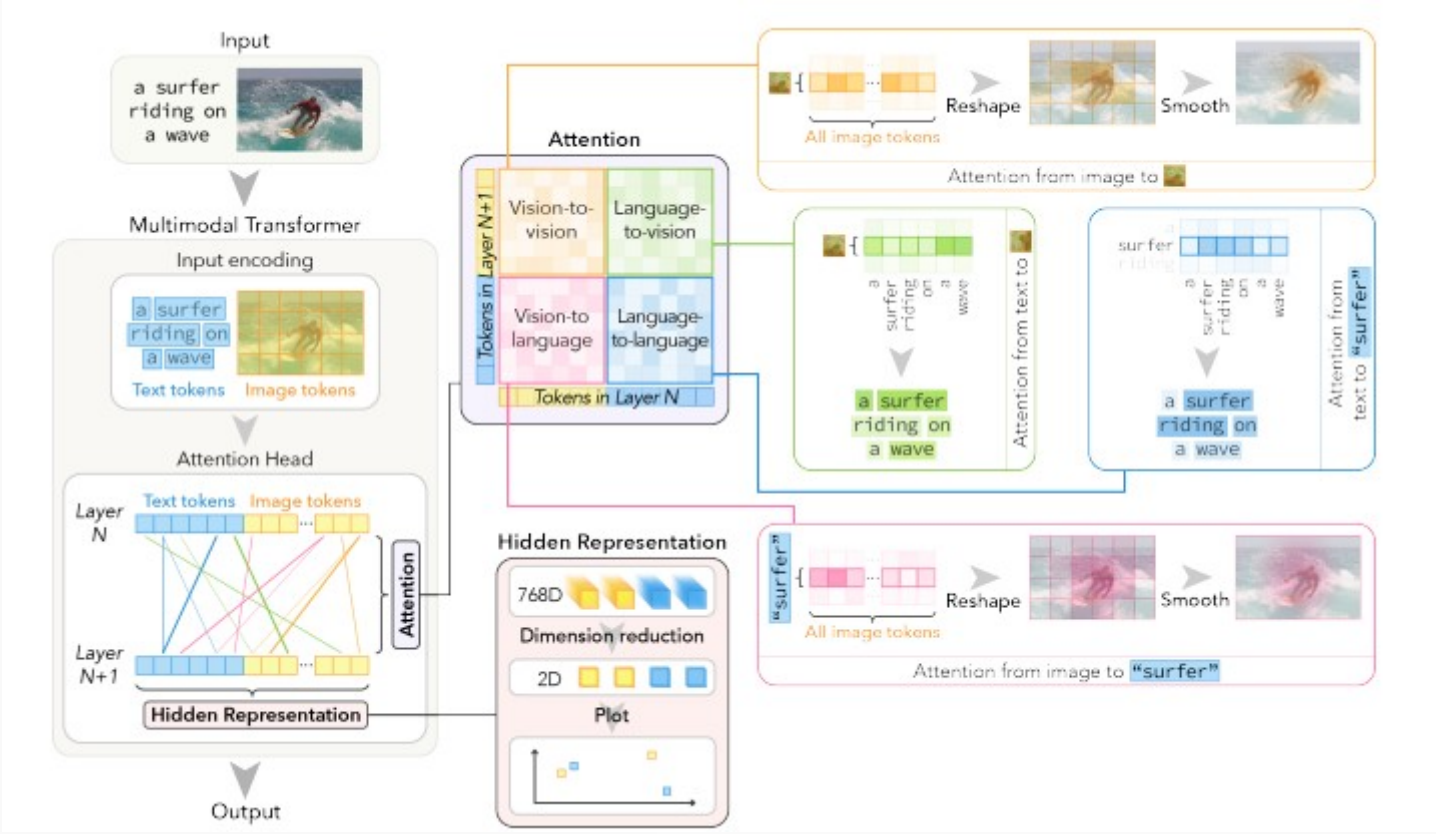
https://github.com/pytorch/captum/releases/tag/v0.5.0

VL-InterpreT от Emergent AI

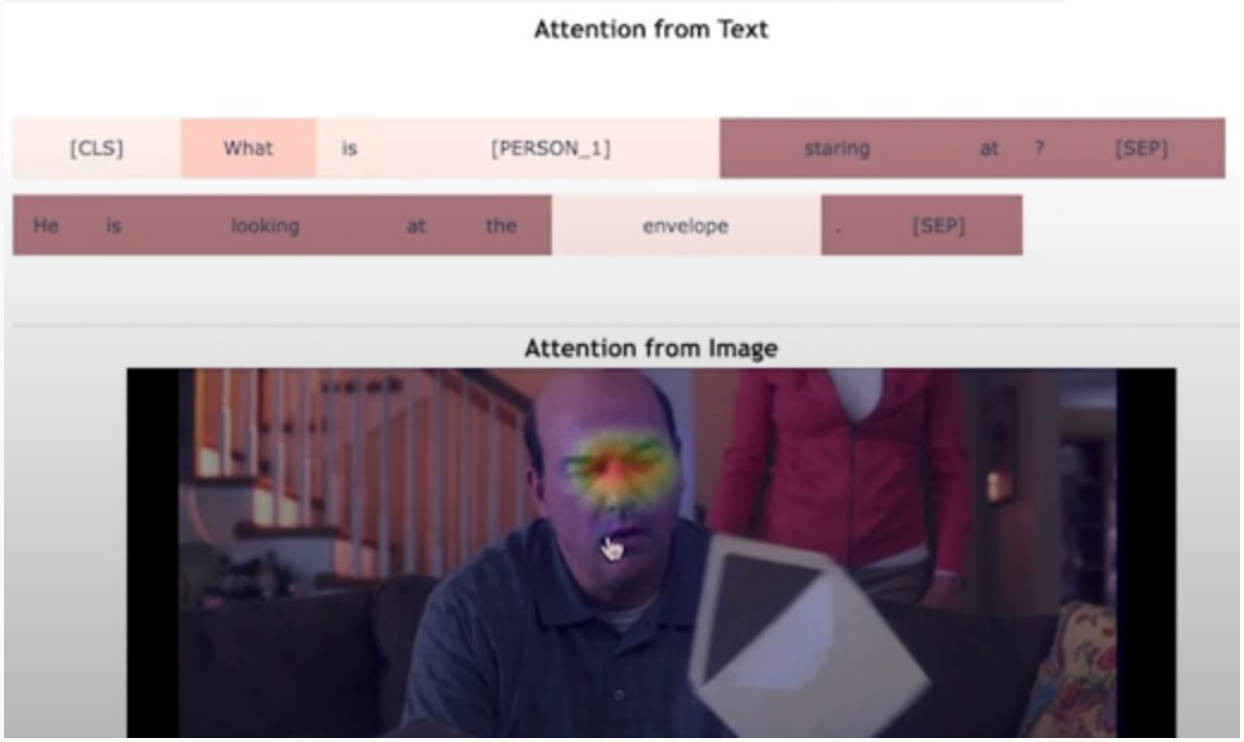


- Пока доступен в виде ролика на ютубе ;-)
- https://www.youtube.com/watch?v=4Rj15Hi_Pdo
- https://arxiv.org/abs/2203.17247

VL-InterpreT



VL-InterpreT

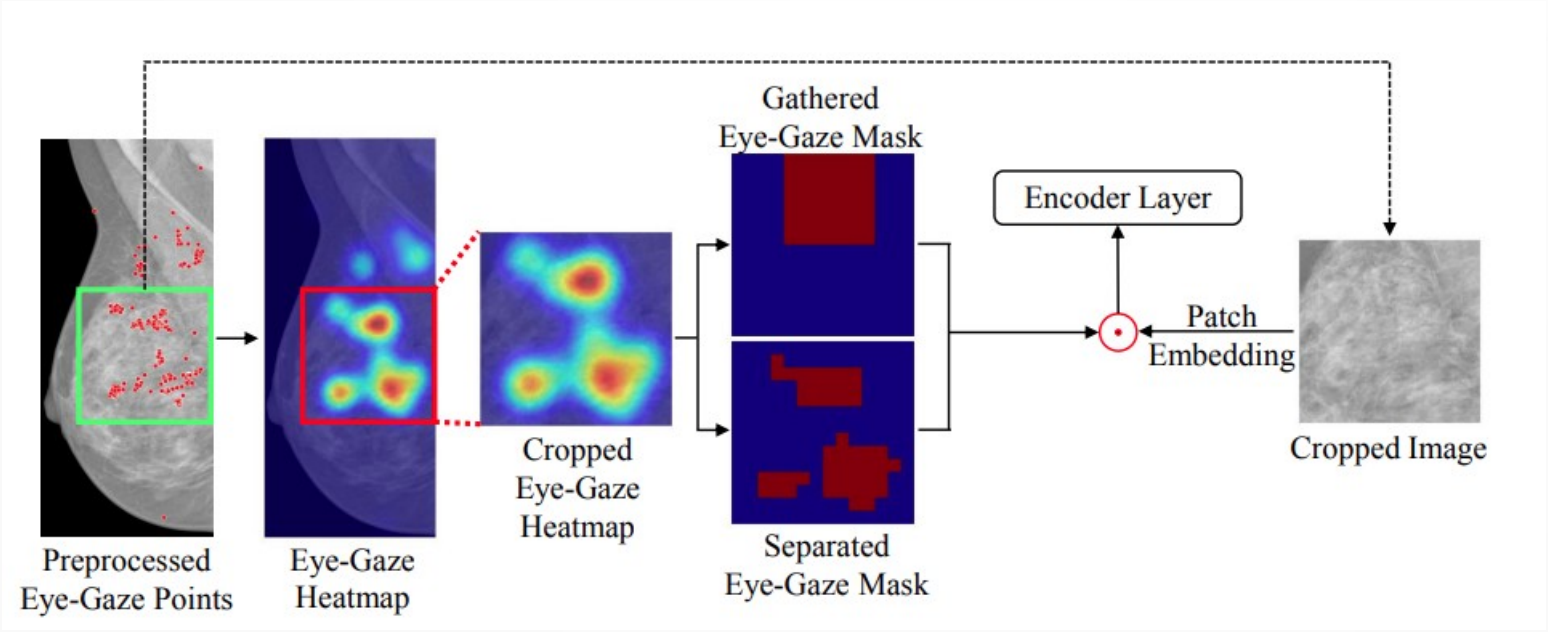


- Хороший
- Понятный
- Недоступен
- Ждем

Похожее https://arxiv.org/abs/2203.05922
Visualizing and Understanding Patch Interactions in Vision Transformer

Интерпретируемость наоборот

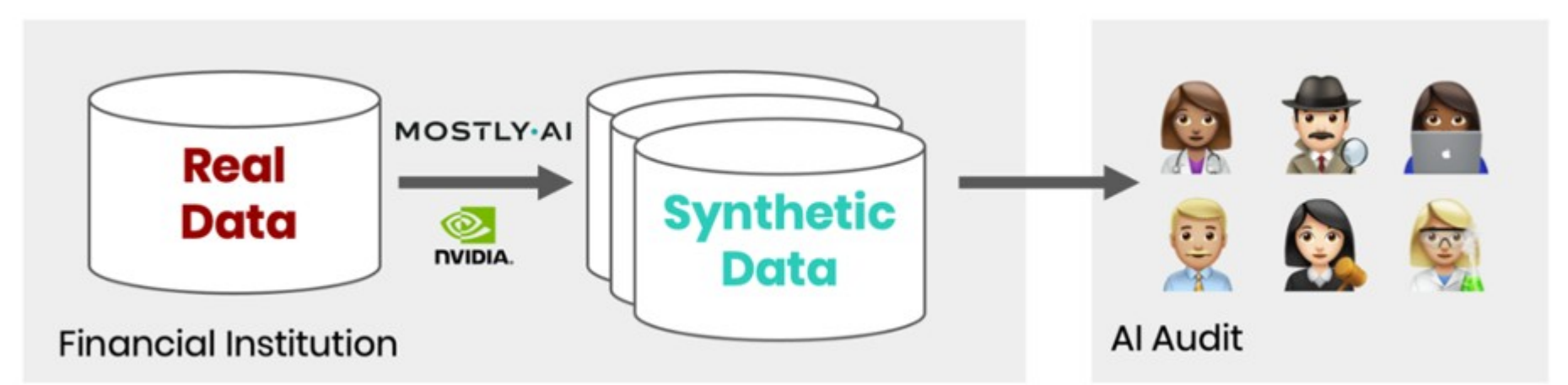
Eye-gaze-guided Vision Transformer for Rectifying Shortcut Learning



https://arxiv.org/abs/2205.12466

Объяснения на синтетике

- Когда объяснять нужно
- А данные показывать нельзя
- Практический кейс от Mostly.Ai



Кстати - объяснения на внешних данных

- Мы не обязаны строить интерпретацию на том же наборе признаков, на котором учили модель
- Признаки модели — часто трансформированные
- Люди их не понимают
- Как считать важность категориальных признаков?
- Учим модель на том, что работает
- Учим LIME на понятных людям признаках

https://christophm.github.io/interpretable-ml-book/lime.html#advantages-13

HIVE — что пользователь понял

Key findings

We conduct IRB-approved human studies with ~1000 participants across 4 different interpretability methods (e.g., post-hoc explanations, interpretable-by-design models, heatmaps, and prototype-based explanations): GradCAM [1], BagNet [2]. ProtoPNet [3], ProtoTree [4].

Key findings

- Participants tend to believe that a model prediction is correct, when given an explanation for it.
- Participants struggle to identify the correct prediction based on explanations.
- A gap exists between prototype-region similarity ratings of ProtoPNet [3] & ProtoTree [4] and those of human participants.
- To prefer a baseline model over a model that comes with explanations, participants require the baseline model to have higher accuracy in higher-risk settings.

https://arxiv.org/abs/2112.03184v3

HIVE: Evaluating the Human Interpretability of Visual Explanations

Разное интересное

- https://arxiv.org/abs/2112.13112v2
  A Survey on Interpretable Reinforcement Learning
  Кратко: Это возможно!
- https://arxiv.org/abs/2104.08782v2
  On the Sensitivity and Stability of Model Interpretations in NLP
  Кратко: давайте измерим comprehensiveness, sufficiency, sensitivity , stability и расстроимся, что все плохо.
- https://arxiv.org/pdf/2205.09971.pdf
  On Tackling Explanation Redundancy in Decision Trees
  Кратко: деревья решения тоже так себе интерпретируемы, много лишней информации

Вопросы

Слайды тут

dkolodezev

dmitry_kolodezev

promsoft

d_key

https://kolodezev.ru/download/interpretable2022.pdf