**Explaining Explanations: An Overview of Interpretability of Machine Learning**

In order for humans to trust black-box methods, we need explainability – models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions.

Interpretovatelnost - chceme chapat co model robit
vysvetlitelnost - chceme pochopit preco sa model sprava tak ako sa sprava,

- interpretovatelnost sama o sebe nestaci
- Aby ľudia dôverovali metódam black box, potrebuju vysvetliteľne modely, ktoré sú schopné zhrnúť dôvody pre správanie neurónovej siete, tieto získaju dôveru používateľov alebo produkuju informácie o príčinách ich rozhodnutí
- Hoci pojmy interpretovatelnost a vysvetlitelnost sa zvyknu casto zamienat je velmi dolezite rozlisovat medzi nimi
- vysvetlitelne metody su standartne interpretovatelne ale naopak to nemusi platit

Zaujimavost:
- Niektoré existujúce zavedené systémy a nariadenia spôsobujú, že je potrebné, aby boli systémy vysvetlitelne surne a vcasne
- S blížiacimi sa nariadeniami, ako je „právo Európskej únie na vysvetlenie “[12], ktore pozaduje rozmanitosť a začlenenie do systémov AI [13], zistenia, že niektoré automatizované systémy môžu posilniť nerovnosť a zaujatosť [14], sposobili ze v poslednej dobe extremne narastol zaujem o vysvetlitelne systemy
- Tieto black box modely su implementovane na vela miestach a v roznych disciplinach kde je ich vysvetlitelnost potrebna
- Príklady všeobecných „vysvetliteľných systémov“ : interpretovateľné AI, vysvetliteľné ML, kauzalita, bezpečna AI, výpočtová sociálna

veda a automatický vedecký objav. Daľší výskum vo vysvetleniach a ich vyhodnotení sa nachádzajú v strojovom učeni, interakcii človeka s počítačom (HCI) a mnoho ďalších inych disciplín.

sekcia 2: zakladne pojmy - explanation, interepretability, explainability
- explanation: možno vyhodnotiť dvoma spôsobmi: podľa jeho interpretovateľnosť a podľa jej úplnosti.
- cielom interpretovatelnosti je opisat vnutro systemu tak aby bol zrozumitelny pre ludi
- Úspech tohto ciela je spojený s poznaním, znalosťami a zaujatosťou používateľa: na to, aby bol systém interpretovateľný, musí poskytnúť opisy, ktoré sú dostatočne jednoduché na to, aby ich človek pochopil pomocou

slovnej zásoby, ktorá je pre používateľa pochopitelna.
- Cielom uplnosti je presne popisat fungovanie systemu.
- vysvetlenie je uplnejsie ked umoznuje prepdpovedat spravanie sa systemu vo viacerych situaciach
- Pri vysvetlovani programu pocitaca ako napriklad DNN, dokonale kompletne vysvetlenie možno vždy získať odhalením všetkých matematických operácií a parametrov v systéme.
- Výzva, ktorej čelí vysvetliteľná AI, spočíva vo vytváraní vysvetlení, ktoré sú úplné a interpretovateľné: je ťažké dosiahnuť súbežnú interpretovateľnosť a úplnosť.
- Najpresnejšie vysvetlenia nie sú ľahko interpretovateľné ľuďom; a naopak najviac interpretovateľné popisy často neposkytujú prediktívnu silu.
- Herman [18] poznamenáva, že by sme mali byť opatrní pri vyhodnocovaní interpretovateľných systémov iba pomocou ľudských hodnotení interpretovateľnosti, pretože hodnotenia ľudí implikuju silne skrelsenie k jednoduchším opisom. Varuje že spoliehanie sa na hodnotenia ľudí môže viesť vedcov k vytváraniu presvedčivych systémov a nie transparentnych
- Dve dolezite eticke dilemy

Kedy je neetické manipulovať s vysvetlením na to aby sme lepšie presvedčili používateľov?
Ako vyvážime naše obavy týkajúce sa transparentnosti a etiky s našou túžbou po interpretovateľnosti?

- veria ze je neeticke prezentovat zjednoduseny opis komplexneho systemu len preto aby sme zvysili doveryhodnost, najma ak rozne limitacie a obmedzenia zjednodušeného opisu používatelia nie su schopni pochopit
  - a čo je horšie, ak je optimalizovane vysvetlenie skryva nežiaduce atribúty systému
- na to aby sme tomu predisli vysvetlenia by mali predstavovat akysi kompromis medzi interpretovatelnostou a uplnostou

Explainability of Deep Networks processing

- vysvetlitelnost neuronovych sieti sa zameriava bud na vysvetlenie spracovania udajov sietou alebo na vysevteleinie reprezentacie dat vo vnutri siete
- Bežne používané hlboké siete odvodzujú svoje rozhodnutia pomocou veľké množstvo základných operácií - vela parametrov a operacii
- Teda základným problémom, ktorému čelia vysvetlenia takéhoto spracovania, je nájsť spôsoby, ako znížiť zložitosť všetkých týchto operácií.
- To sa dá dosiahnuť vytvorením proxy modelu, ktorý sa správa podobne ako pôvodný model

Linear Proxy Models:

- LIME, Riberio
- Pri LIME sa systém čiernych skriniek vysvetľuje skúšaním správania poruchy vstupu a potom sa tieto údaje použijú na zostavenie lokalneho lineárneho modelu, ktorý slúži ako zjednodušený úplný model v blizkosti k

vstupu.
- Ribeiro ukazuje metódu, ktoru možno použiť na identifikáciu oblastí vstupu, ktoré majú najväčší vplyv na rozhodovanie v rôznych druhoch modelov a problémovych doménach
-

Decision trees
- obmedzena skalovatelnost
- Úsilie rozložiť neurónové siete na rozhodovacie stromy nedávno rozšírili prácu od 90. rokov, zameraná na plytké siete, na zovšeobecnenie procesu hlbokých neurónových sietí.
- jednou z takychto metod je DeepRed
- ANN DT
-

Automatic-Rule Extraction:
- Automatická extrakcia pravidiel je ďalším dobre preštudovaným prístupom k sumarizácii rozhodnutí.

- Dekompozičné prístupy pracujú na úrovni neurónov, extrahuju pravidlá napodobňujúce správanie jednotlivých jednotiek.

Salience Mapping
- sieť je opakovane testovaná so zakrytymi časťami vstupu aby sa vytvorila mapa ukazujúca, ktoré časti údajov skutočne majú vplyv na sieťový výstup.
  - Príkladmi sú LRP [40], DeepLIFT [41], CAM [42], GradCAM [43], integrované prechody [44] a SmoothGrad [45]. Každá technika vytvára rovnováhu medzi zobrazením oblastí , vysokou aktiváciou siete, kde neuróny sú najsilnejšie a oblastami s vysokou citlivosťou na sieť, kde by zmeny mohli najviac ovplyvniť výstup.
    - Porovnanie niektorých z týchto metód Ancona[46].

explainability of deep learning representation

Role of layers
Role of individual unitis
Role of representation vectors

Zaujimave zdroje: 46