Abstrakt

Najnovšie pokroky v strojovom učení nás priviedli do stavu, v ktorom zvyšovanie presnosti modelov často nie je prvoradým cieľom. Úlohou môže byť aj vytvorenie modelu, ktorý okrem excelentnej úspešnosti bude mať aj iné potrebné kvality. Jednou zo súčasných výziev hlbokého učenia je nedostatočná robustnosť, t.j. nízka presnosť na dátach mimo trénovacej distribúcie. Toto otvára dvere pre rôzne útoky, proti ktorým sa modely nedokážu brániť. Ďalším problémom je nedostatočná úroveň transparentnosti a vysvetliteľnosti hlbokých neurónových sietí. Keďže hlboké učenie má v rôznych úlohách ľuďom pomáhať, pri interakcii často vyžadujeme nielen správnu a rýchlu odpoveď, ale aj jej zdôvodnenie. Zvýšená vysvetliteľnosť modelov by preto viedla k širšiemu nasadeniu umelej inteligencie v každodennom živote, a vo všeobecnosti k väčšej dôvere voči týmto modelom. Práve preto sa v tejto práci venujeme robustnosti a vysvetliteľnosti v hlbokom učení.

Jeden z kľúčových konceptov, ktorý v tejto práci využívame, je pozornosť. Mechanizmy pozornosti v ostatných rokoch priniesli obrovské úspechy, no podľa nášho názoru, tento koncept stále nie je dostatočne preskúmaný a ponúka mnohé smery v ktorých by vývoj umelej inteligencie mohol napredovať. Preto v tejto práci skúmame tri paralelné výskumné línie. Po prvé, analyzujeme skupinu škodlivých vstupov, taktiež nazývaných aj adverzariálne vstupy. Tieto dáta používame v štandardných modeloch hlbokého učenia a navrhujeme spôsoby, ako skúmať ich rozdiely od čistých dát. Po druhé, s cieľom zmiernenia účinkov adverzariálnych vstupov na klasifikáciu obrázkov, navrhujeme a testujeme model založený na pozornostnom mechanizme, RecViT. Po tretie, navrhujeme viacero prístupov riešenia odhadu adresáta pri interakcii robota s ľuďmi. Vďaka dizajnu s pozornostným mechanizmom, naše modely ponúkajú spôsoby, ako z nich extrahovať zrozumiteľné vysvetlenia ich rozhodnutí. Dúfame, že práve takáto práca umožní plynulejšiu a dôveryhodnejšiu interakciu človeka s robotom.

Kľúčové slová: pozornosť, vysvetliteľnosť, robustnosť, adverzariálne vstupy