Abstrakt

Pochopenie vnútorných reprezentácií neurónových sietí je základnou výzvou v oblasti umelej inteligencie. Napriek dosiahnutiu pozoruhodných výsledkov v rôznych oblastiach tieto modely často postrádajú interpretovateľnosť, transparentnosť a vysvetliteľnosť, čo bráni ich nasadeniu v kritických aplikáciách, kde je nevyhnutné pochopiť ich rozhodovacie procesy. Táto dizertačná práca skúma a rozširuje metódy na interpretáciu predikcií neurónových sietí transformáciou ich vnútorných reprezentácií do foriem zrozumiteľných pre človeka, pričom sa snaží minimalizovať zavádzanie zaujatostí alebo skreslení. Práca je štruktúrovaná okolo troch kľúčových príspevkov:

Po prvé, v oblasti počítačového videnia predstavujeme metódu Bi-Source Class Visualization (BSCV) na riešenie výziev vo vizualizácii čŕt a klasifikačných tried. BSCV využíva rámec adversariálnej neurónovej siete, pričom vstupná vrstva slúži ako zdieľané rozhranie medzi diskriminátorom a klasifikátorom. Umožnením prechodu gradientov cez toto rozhranie BSCV generuje realistické a interpretovateľné vizualizácie tried bez potreby ručne vytvorených regularizácií.

Po druhé, skúmame, ako predtrénovanie poskytuje morfosyntaktické znalosti jazykovým modelom prostredníctvom sondovania vnútorných reprezentácií viacjazyčných modelov BERT. Diagnostické klasifikátory, tzv. sondy, sú trénované na predpovedanie lingvistických vlastností z vnútorných stavov modelov, ktoré nie sú dotrénované na morfosyntaktické úlohy. Rozsiahle ablačné štúdie a kontroly sond zabezpečujú platnosť zistení, odhaľujúc, že predtrénovanie vedie k vysoko abstraktným vnútorným reprezentáciám kódujúcim bohaté lingvistické informácie.

Po tretie, v prípadovej štúdii zahŕňajúcej projekt MIMEDIS analyzujeme mediálny diskurz o migrácii pomocou viacjazyčných a jednojazyčných modelov. Vyvinutím špecializovanej aplikácie Shapleyho hodnôt pre úlohy spracovania prirodzeného jazyka sme systematicky dekomponovali predpovede modelov, aby sme kvantifikovali príspevok jednotlivých vstupných tokenov. Táto metodológia umožnila podrobnú analýzu vnútorných rozhodovacích procesov modelov, odhaľujúc významné rozdiely vo vnútorných reprezentáciách architektonicky podobných modelov trénovaných na odlišných dátach, čím demonštrujeme dopad tréningových dát na správanie modelu.