Kód tématu: 350-IT-CZ

Evoluční výpočetní techniky a vysvětlitelná umělá inteligence

Školitel: prof. Ing. Šenkeřík Roman, Ph.D.

Konzultant: Ing. Viktorin Adam, Ph.D., ---

Ústav fakulty: Ústav informatiky a umělé inteligence

Studijní program: Informační technologie

Anotace:

Tato doktorská práce zkoumá rozvíjející se oblast evolučních výpočtů se silným zaměřením na zvýšení jejich vysvětlitelnosti a důvěryhodnosti nebo na vysvětlitelnou umělou inteligenci (Explainable Artificial Intelligence, XAI) jako takovou. Existují dva možné směry. Zaprvé vysvětlitelné principy umělé inteligence pro evoluční výpočty a zadruhé evoluční výpočty pro vysvětlitelnou umělou inteligenci.

Evoluční výpočty (Evolutionary computation – EC), které čerpají inspiraci z přirozené biologické evoluce, využívají algoritmy, které iterativně vyvíjejí řešení složitých problémů. Zásadní výzvou v této oblasti je neprůhlednost těchto algoritmů, která ztěžuje pochopení a důvěryhodnost jejich rozhodovacích procesů. Tento výzkum se bude snažit tento problém řešit vývojem metod, které umožňují interpretaci a transparentnost evolučních algoritmů, čímž zvyšují jejich spolehlivost a přijatelnost. Významná část práce bude věnována začlenění konceptu důvěryhodné optimalizace do evolučních výpočtů.

Oblast vysvětlitelné umělé inteligence se snaží řešit rostoucí potřebu transparentnosti a interpretovatelnosti systémů umělé inteligence. Tento výzkum může k tomuto úsilí přispět zkoumáním aplikace metod EC v rámci XAI. Výzkum se zaměří na nové přístupy, které používají techniky EC k vývoji modelů umělé inteligence a současně zajišťují jejich vysvětlitelnost, tj. navrhují přístupy, které zjednodušují řešení generovaná EC algoritmy, aniž by se podstatně snižovala jejich účinnost. Disertační práce se může týkat několika směrů, například genetického programování pro generování pravidel. Pro XAI lze programy/modely vyvíjet tak, aby generovaly soubor pravidel nebo rozhodovacích stromů, které jsou snadno interpretovatelné člověkem. Dále výběr a redukce fetaures, kde lze algoritmy EC použít pro výběr features a redukci dimenzionality v souborech dat, což je klíčové pro vytváření vysvětlitelných modelů. Možné směry zahrnují také optimalizaci architektur neuronových sítí, post-hoc analýzu a vizualizaci, zjednodušování složitých řešení a další.

Literatura:

[1] BACARDIT, Jaume, et al. The intersection of evolutionary computation and explainable AI. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2022. p. 1757- 1762.

[2] XU, Feiyu, et al. Explainable AI: A brief survey on history, research areas, approaches and challenges. In: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. Springer International Publishing, 2019. p. 563-574.

[3] HOFFMAN, Robert R., et al. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608, 2018.

[4] HOLZINGER, Andreas. From machine learning to explainable AI. In: 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE, 2018. p. 55-66.

[5] SAMEK, Wojciech; MÜLLER, Klaus-Robert. Towards explainable artificial intelligence. Explainable AI: interpreting, explaining and visualizing deep learning, 2019, 5-22.

[6] BACARDIT, Jaume, et al. Evolutionary Computation and Explainable AI Towards ECXAI 2023: A Year in Review. ACM SIGEVOlution, 2023, 16.2: 1-6.