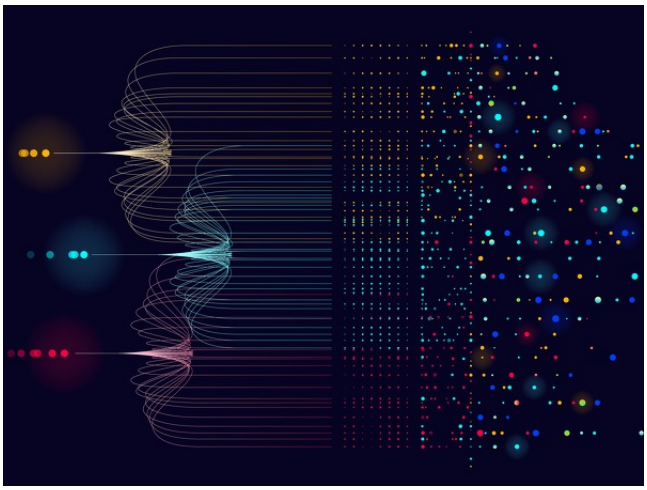
INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

Unified Complex Networks - Classifier Playground

TECHNOLOGY NUMBER: 2024-250



OVERVIEW

A comprehensive toolkit for safe, multi-method fine-tuning of pretrained AI models

- Enhances safety and evaluation across a wide range of fine-tuning techniques
- Research, enterprise AI deployment, and regulated industry model customization

BACKGROUND

As artificial intelligence systems advance, leveraging pretrained models has become standard practice in machine learning. Pretrained models are typically trained on vast general datasets, and are then fine-tuned on smaller, task-specific datasets to improve their relevance and performance for domain-specific applications. Historically, fine-tuning has been conducted using individual methods, often lacking broad interoperability and consistent safety evaluations. Existing tools seldom provide support for multiple fine-tuning paradigms within a unified code base, and typically underemphasize robust safety and bias testing. As AI models are increasingly applied in sensitive domains like healthcare, finance, and legal settings, insufficient safety validation can lead to biased, incorrect, or even harmful outputs. This has established a critical need for an integrated, extensible toolkit that supports a spectrum of fine-tuning strategies alongside robust, systematic safety evaluation mechanisms.

INNOVATION

Technology ID
2024-250

Category
Software
MOSS - Michigan Open Source Support

Inventor
Danai Koutra
Puja Trivedi
Jayaraman J. Thiagarajan

Further information
Ashwathi Iyer
ashwathi@umich.edu

View online page

The released code base introduces a modular platform that supports various fine-tuning methods—including supervised, reinforcement, and transfer learning—across diverse pretrained models. This approach empowers users to tailor model adaptation strategies to their unique needs while ensuring reliability and reproducibility. A defining feature is the built-in, safety-centric evaluation suite, which rigorously assesses downstream task performance and surfaces risks related to bias and harmful outputs. Technical advances include streamlined integration with popular deep learning libraries, support for custom evaluation protocols, and detailed reporting of safety metrics. This toolkit enables researchers to experiment efficiently with different fine-tuning modalities, organizations to safely customize models for industry requirements, and developers to rapidly deploy compliant AI solutions in regulated settings such as healthcare, finance, or education. The comprehensive, safety-first approach also encourages ethical best practices throughout the AI development lifecycle.

ADDITIONAL INFORMATION

PROJECT LINKS:

- Classifier Playground Github

DEPARTMENT/LAB:

- Danai Koutra, Electrical Engineering and Computer Science (EECS)

LICENSE: