Modelgo: A Tool for Machine Learning License Analysis

Moming Duan
National University of Singapore
Singapore
moming@nus.edu.sg

ABSTRACT

Productionizing machine learning projects is inherently complex, involving a multitude of interconnected components that are assembled like LEGO blocks and evolve throughout development lifecycle. These components encompass software, databases, and models, each subject to various licenses governing their reuse and redistribution. Therefore, existing license analysis approaches for Open Source Software (OSS) are not well-suited for this context. The intricate web of licenses often leads to conflicts. For instance, bundling the LGPL-LR corpus dataset with another corpus dataset licensed under CC-BY-SA 4.0 can result in a copyleft proliferation conflict. In contrast, bundling this corpus with a model licensed under copyleft GPL 3.0 does not lead to such conflict, as the derivative work is no longer considered a linguistic resource. Indeed, the analysis of licenses in machine learning projects grows significantly more intricate as it involves interactions among diverse types of licenses and licensed materials. To the best of our knowledge, no prior research has delved into the exploration of license proliferation and conflicts within this domain. In this paper, we propose a feasible tool called Modelgo for assessing license conflict risks in machine learning projects. [...]

CCS CONCEPTS

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

KEYWORDS

Software licenses, Software reuse, Open source software, Model mining

1 INTRODUCTION

Open source software license compliance [1]

The open source definition [2]

AFL [3]

Wudao2.0 1.75T MoE [FASTMOE: A FAST MIXTURE-OF-EXPERT TRAINING SYSTEM] [GLM-130B: AN OPEN BILINGUAL PRE-TRAINED MODEL]

Objectives and challenges associated with analyzing dataset license compliance? Getty Images (US), Inc. v. Stability AI, Inc. (1:23-cv-00135) Andersen et al v. Stability AI Ltd. et al (3:23-cv-00201) We are not aware of any copyright restrictions of the material

C4, Pile Common Crawl crowdsourced

COCO (CC-BY 4.0), CIFAR10 -> Flickr Unsplash License (custom): Compiling photos from Unsplash to replicate a similar or competing service. https://unsplash.com/license Pixabay License: Data mining, extraction, scraping and the use of programs or robots for automatic data collection and/or extraction of digital data on the Services and/or the content available therein is strictly prohibited for all purposes, including without limitation for machine learning purposes.

Google Street View (SVHN) https://about.google/brand-resource-center/products-and-services/geo-guidelines/

Software reuse is very simple from the legal point of view, if a company or an individual reuses software for which it has copyrights. However, things change dramatically if one wants to reuse software made by others, since software is protected by copyright and possibly by patents. Without explicit permission, no person other than the copyright holder is allowed to copy, distribute, or make derivative works from the original work.

ACKNOWLEDGMENTS

Ack.

REFERENCES

[1] Philippe Ombredanne. 2020. Free and open source software license compliance: tools for software composition analysis. *Computer* 53, 10 (2020), 105–109. https://doi.org/10.1109/MC.2020.3011082

[2] Bruce Perens. 1999. The open source definition. *Open sources: voices from the open source revolution* 1 (1999), 171–188.

[3] Lawrence Rosen. 2005. *Open Source Licensing: Software Freedom and Intellectual Property Law.* Prentice Hall Professional Technical Reference, New Jersey.