***Summary***

***AIMMX: Artificial Intelligence Model Metadata Extractor***

***Authors:***

*Jason Tsay Jason*

*Alan Braz*

*Martin Hizrail*

***Universities:***

*IBM Research Yorktown Heights, New York, USA*

*IBM Research São Paulo, Brazi*

*IBM Research Yorktown Heights, New York, USA*

***Date:*** *MSR ’20, October 5–6, 2020, Seoul, Republic of Korea*

***INTRODUCTION:***

*The combination of suﬃcient hardware resources, the availability of large amounts of data, and innovations in artiﬁcial intelligence (AI) models has brought about a renaissance in AI research and practice. For this paper, we deﬁne an AI model as all the software and data artifacts needed to deﬁne the statistical model for a given task, train the weights of the statistical model, and/or deploy the trained model weights for prediction in a service or application. Our deﬁnition of model includes both traditional machine learning (ML) and deep learning models. One avenue for standardization is that software and AI development share many of the same languages and tools, such as version control systems. Existing software repository tools and services, such as GitHub, are popular with AI developers to store model deﬁnition code and development artifacts such as conﬁgurations and training logs. These connections may also enable automated training or deployment in future tools. This paper makes the following contributions:*

*• Tool for extracting AI model-speciﬁc metadata from software repositories with currently ﬁve extraction modules (Section 2).*

*• Evaluation of our tools against a dataset of 7,998 models (Section 3). This AI model metadata dataset is also available as part of a replication package.*

*• Preliminary usage of extracted metadata via an exploratory analysis of the data and method reproducibility of AI models in our dataset and implementation of a cataloging tool (Section 4).*

***ABSTRACT:***

*Despite all of the power that machine learning and artiﬁcial intelligence (AI) models bring to applications, much of AI development is currently a fairly ad hoc process. Software engineering and AI developments hare many of the same languages and tools, but AI development as an engineering practice is still in early stages. The extractors have ﬁve modules for extracting AI model-speciﬁc metadata: model name, associated datasets, references, AI frameworks used, and model domain. Our collected models are searchable in a catalog that uses existing metadata to enable advanced discovery features for eﬃciently ﬁnding models.*

***CONCLUSIONS:***

*This paper describes AIMMX which we intend as a step towards furthering engineering support for AI development through providing standardized metadata for existing AI models. We envision that generating analyzable metadata for disparate models is both the ﬁrst step towards managing models at scale and adapting existing mining software repositories techniques to AI models.*