発表概要

Pandasデータ解析ライブラリで記述された機械学習前処理の性能最適化に関する検討

仲池 卓也[1,a)]　川人 基弘[1]　小原 盛幹[1]

2020年10月30日発表

従来，機械学習においては，ロジスティック回帰分析等のモデルの実行性能が重要視され，GPU 等のハードウェアアクセラレータにより最適化されてきた．しかしながら，モデルの推論精度を向上させるためには，特徴量エンジニアリングを含めたデータの前処理が重要であり，それらの前処理の実行性能は十分に最適化されていない．本発表では，Pandas データ解析ライブラリで記述された機械学習前処理の性能を最適化する手法を提案する．Pandas は，Python で記述されたデータ解析ライブラリであり，その利便性のため，多くのデータサイエンティストに利用されている．しかしながら，すべてのライブラリが Python で実装されているため，高い性能を求めることが難しい．我々の提案手法は，Pandas で記述された機械学習前処理を ONNX 形式に変換し，高速な機械学習フレームワークを利用することにより性能向上を目指す．本発表では，我々が実装中の Pandas から ONNX の変換ツールの概要，および Pandas で記述された前処理と ONNX ランタイム上の前処理の性能比較について報告を行う．

Presentation Abstract

Performance Optimizations of Machine Learning Pre-Processing Written in Pandas Data Anlytics Library

Takuya Nakaike[1,a)]　Motohiro Kawahito[1]　Moriyoshi Ohara[1]

Presented: October 30, 2020

In machine learning, researchers and developers have been optimizing the performance of machine-learning models such as loggistic regression by using hardware accelerators such as GPU. However, data pre-processing was not the main forcus of the performance optimization even though it is very important to improve the inferencing accuracy of machine-learning models. This presentation proposes a method to optimize the performance of the data pre-processing code witten in Pandas which is a data analytics library. Pandas has been widely used by many data scientists due to its useful data anlytics APIs. However, Pandas is not so fast because it is written in Python which has type checking overhead and serializes the execution. Our proposed method aims to improve the performance of data pre-processing by converting the data pre-processing code written in Pandas into an ONNX graph, which is a standard formant to represent machine-learning models, and then running the graph on other high-performance machine learning platforms such as Tensorflow. This presentation overviews our tool to covert the Pandas code into an ONNX graph, and then show how the performance of data pre-processing is improved.

[1] 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research - Tokyo, Chuo, Tokyo 103–8510, Japan
[a)] nakaike@jp.ibm.com