Contribution ID: **20** Type: **Oral presentation**

Preprocessing arbitrarily structured data for AI with Awkward Array

Thursday 3 October 2024 09:55 (35 minutes)

Processing heterogeneous multimodal data presents challenges. These datasets feature complex, irregular structures due to nested or variable-sized outputs from different sensors, or due to missing data values. The data are typically of mixed types, complicating the preprocessing steps required before they can be fed into algorithms like multimodal representation models. AI practitioners must manage these complexities effectively.
Awkward Array is a Python library designed to process arbitrarily structured data. Operating on an array-programming paradigm, it allows users to manipulate data using NumPy-like syntax. Awkward Array also includes GPU-accelerated kernels, enabling the preprocessing of complex data directly on modern hardware accelerators, which can significantly optimize the training process and reduce data transfer latency to the device.
We introduce the Awkward Array library and provide examples that demonstrate its usage, highlighting its potential as an AI preprocessor.

Contribution length

Middle

Primary authors: PIVARSKI, Jim (Princeton University); KOURLITIS, Vangelis (Technical University of Munich)

Presenter: KOURLITIS, Vangelis (Technical University of Munich)