特設

衝撃的な性能UPの可能性を秘めた注目テクノロジ速報

AIをサクサク動かすGoogle新機能 TensorFlow「XLA」を探る

@ Vengineer

Googleが提供する機械学習ライブラリ「TensorFlow」は，C++だけでなくPythonで記述でき，マルチCPU環境やマルチGPU環境や分散処理にも対応可能なディープ・ラーニングのフレームワークです．2017年2月にr1.0（TensorFlow 1.0）が発表になりました．

このTensorFlow 1.0の新機能にXLAがあります．

TensorFlow XLAは，処理の高速化やメモリ使用量の削減などを目的としたコンパイラです．現段階ではまだ実験的ですが，これから，クラウドではない装置/端末（スマホや組み込み装置）で人工知能を実現しようとするときに，従来と比べて衝撃的な性能を出せる可能性を秘めています．

本稿では，TensorFlow XLAとはどういうものか，基本的な使い方，内部動作などについて解説し，今後のTensorFlow XLAの可能性について探ってみたいと思います．

Google人工知能の新機能「TensorFlow XLA」

● XLAの構成

Googleは2017年2月15日に開催されたTensorFlow Summit 2017[3]にて，自社の機械学習ライブラリである「TensorFlow」のr1.0（TensorFlow 1.0）をリリースしました[4]．今回のリリースでCPUとGPUで利用可能なコンパイラ「XLA：Accelerated Linear Algebra」を実験的に採用しました．

TensorFlow XLAは，次の2つの機能に分けることができます．

(1) JITコンパイル（Just In Time Compilation）[8]
(2) AOTコンパイル（Ahead Of Time Compilation）[9]

ざっくり言うと，JITは学習向け，AOTは推論向けの機能だと思います（詳細は後述）．

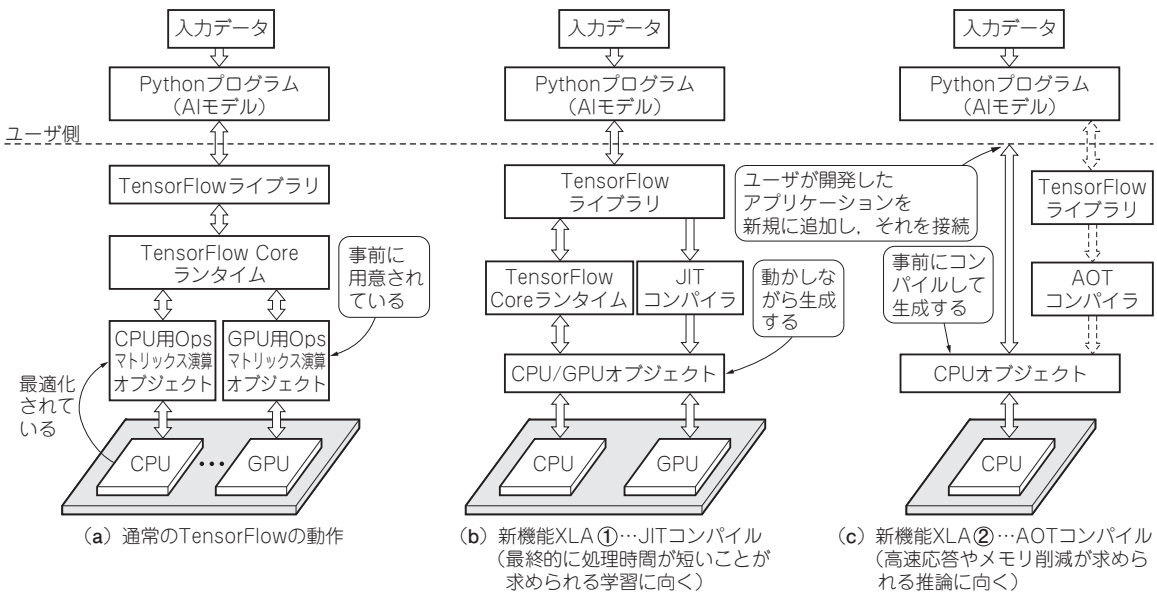

図1　Google人工知能の新機能TensorFlow＆新機能XLAの動作イメージ