増大するAIの計算量、専門知識なしで高速処理する世界初の技術とは？

FUJITSU JOURNAL / 2019年12月10日

AI処理が数年間で30万倍に増大、高速化技術開発が急務

AI技術の進化と普及により画像認識や音声翻訳などへの活用が進み、大容量データの処理需要が増加しています。2012年にわずか約50秒で処理していたデータ量は、2018年には「15,000,000秒（約6カ月）」へと急増。30万倍にまで膨らみました。

これに対し、コンピュータの計算性能は、十分に追いついていないのが現状です。例えば、現在のGPU（Graphics Processing Unit）で4Kカメラが1日に撮影した映像を学習させるには約10日間かかると言われています。このように、現在のコンピュータは、大量のデータを使う映像分析や機械学習などには性能不足というのが現状です。

こうした課題に対し、富士通では高効率な分散並列処理を使い、学習精度を低下させずに計算量を拡張する技術を開発。当時のディープラーニングによる処理速度の記録を30秒短縮し、世界最高速を達成しました。

しかし、コンピュータにはハードウェア技術の進化だけでなく、並列化や専用言語などの使いこなしなど、高度なソフトウェア技術が必要となり、高度なスキルを持った専門家にしか触れないのが現状でした。

使い勝手の良さと高速コンピューティングを両立

富士通研究所では、誰でも使える使い勝手の良さと、高速コンピューティングの両立を目指した新技術「Content-Aware Computing」を世界で初めて開発しました。

「Content-Aware Computing」は、計算の厳密性を自動調整し高速化する技術です。今回、AI処理向けに「ビット削減」と「同期緩和」の2つを開発しました。

ビット削減技術：データに合わせて自動的に削減

「ビット削減」とは、ニューラルネットワークにおける学習の進捗状況に注目し、データの分布に応じてビット幅を最適化する技術です。データが広く分布している学習初期にはより多くのビットを割り振り、学習が進むに連れて分布の範囲が狭くなったデータには低ビットを動的に割り振ることで、演算結果の劣化を抑えつつ、処理の効率化・高速化を図ります。これをディープラーニングに適用した結果、従来の3倍の高速化を実現しました。（図1）

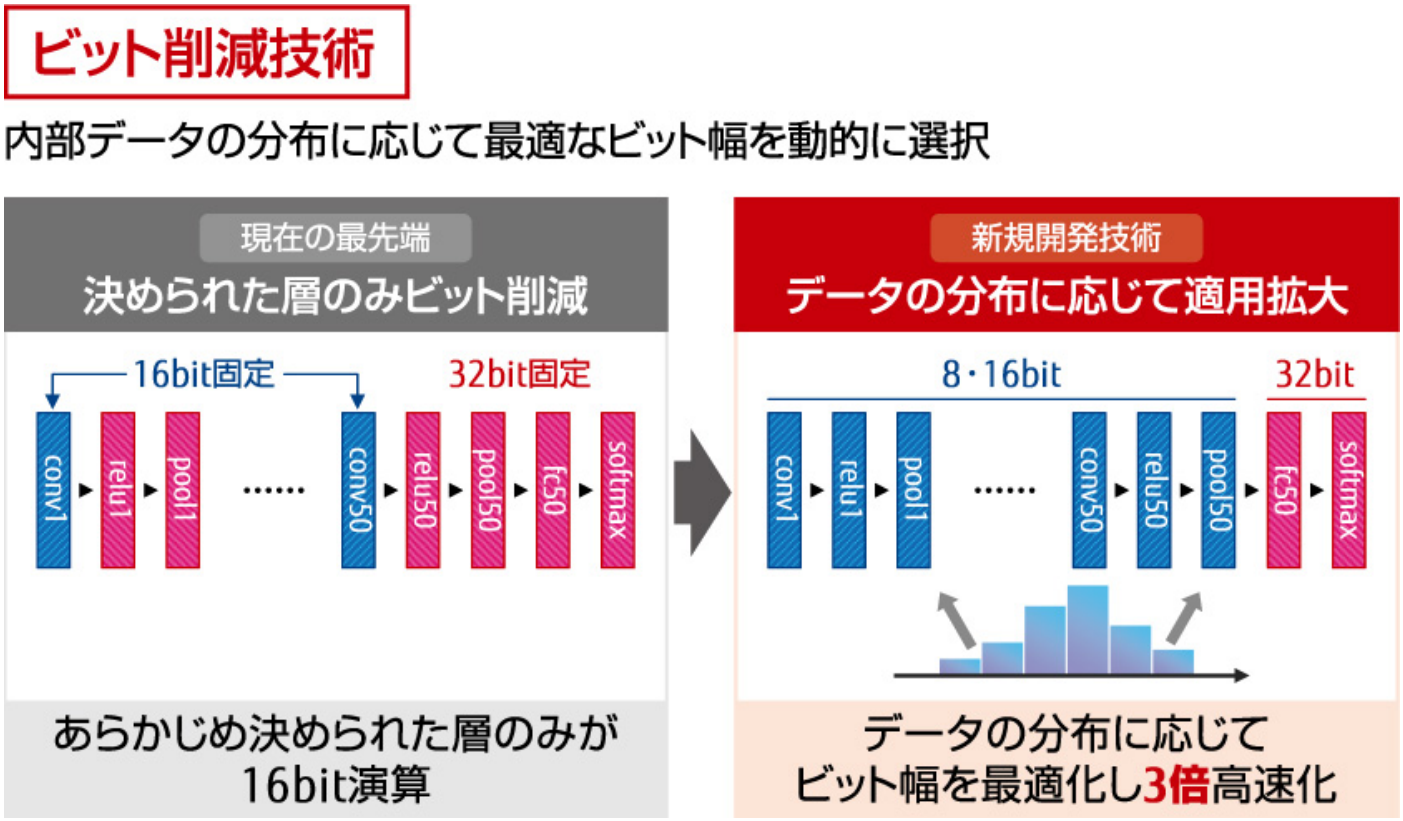

図1 学習の進捗に合わせて自動的に低ビット化し、処理を効率化・高速化

同期緩和技術：ばらつきのある並列環境で高速実行を可能

「同期緩和技術」とは、並列処理の各演算において処理を打ち切った場合に削減できる同期待ち時間と演算結果への影響を予測し、各演算の打ち切りを自動的に調整する技術です。例えば、多数のノードを使った演算やCPUを複数アプリケーションで共用する際には、競合や割り込みにより一部処理のレスポンスが遅れることにより無駄な同期待ちが発生します。当技術を使うことで、実行時間が最小になるよう処理を打ち切り、演算速度を安定化させることが可能です。ディープラーニングに適用した結果、最大で3.7倍の高速化を実現しました。（図2）

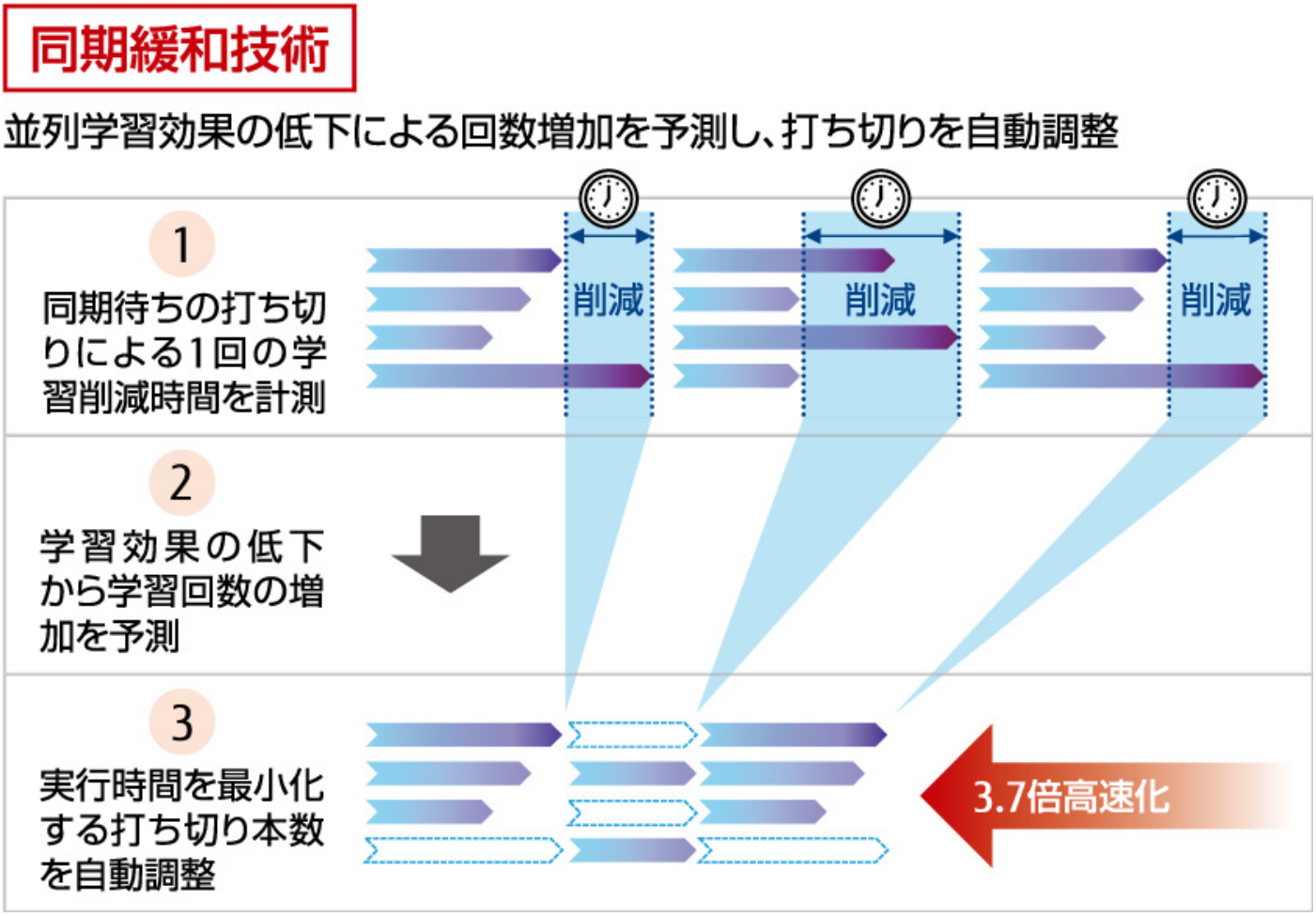

図2 並列処理の各演算において打ち切りを自動的に調整し、演算速度を高速化

今回の技術により、学習の度合いに合わせ最適な演算精度を自動制御し、高速化することが可能となります。また、クラウドなど実行時間にばらつきのある並列環境でも演算速度を安定化させることができます。

専門知識がなくても簡単に、AI処理を自動的に高速・最適化

「Content-Aware Computing」をディープラーニングに活用したところ、AI処理を最大10倍にまで高速化することに成功しました。AIフレームワークやライブラリに組み込むことによって、低ビット演算機能を搭載したGPU・CPU、あるいはそれらが使われているクラウドやデータセンターでのAI処理を自動的に高速化することが可能となります。これにより、専門知識がなくても簡単に利用することができます。

富士通は今後、「Content-Aware Computing」をサービスビジネス、プラットフォームビジネスへ展開し、さらなるDXビジネスへの活用を目指します。

FUJITSU JOURNAL / 2019年12月10日