

推論に適したハードウェアの選び方

望月 英輔

推論に適したハードウェアはどれか

ディープ・ラーニングの推論時には、主に行列積和演算を使います。推論には、行列積和演算を高速かつ効率的に行えるハードウェアが適していると言えるでしょう。

表1に、代表的なハードウェアと、それぞれの特徴、ディープ・ラーニング推論への適正について示します。

ただし、学習時の計算は、推論時と異なり、損失関数計算や誤差逆伝播といった処理を行う必要があります。そのため、行列積和演算ばかりというわけではなく、推論に適したハードウェアが学習にも適しているとは限りません。本稿では推論時に使うハードウェアについて、その適正や高速化手法、選ぶときのポイントなどを解説します。

● CPU…大量データの処理や並列演算は苦手

▶特徴…小さな処理を効率的に実行できる

CPU (Central Processing Unit, 中央処理装置) は、コンピュータの中心的な演算装置です。一般的なPCには必ず搭載されています。スマートフォンにおいてもCPUが動作の中心です。ハードウェア内部構造のイメージを図1に示します。

表1 ディープ・ラーニングの推論に使われる主要なハードウェア

ハードウェア	得意な処理	コア数	メモリ帯域
CPU	多様な小さな演算の効率的な処理	少	低～中
GPU	大量データへの並列演算処理	多	高
NPU	ディープ・ラーニング演算処理	中～多	中～高

CPUの特徴は、後述するGPUやNPUに比べると、バリエーションに富んださまざまな小さな処理を効率的に実行できる点です。PCやスマートフォンにはOSが搭載されていることがほとんどですが、OSはさまざまな小さな処理を常時実行しています。この処理を効率的に実行し、ユーザが快適にOS上で作業できるようにCPUは進化してきました。

▶ディープ・ラーニングへの適正

CPUのディープ・ラーニングにおける適正はどうか。ディープ・ラーニングの主要な計算は行列積和演算ですが、CPUで処理を行うことも当然可能です。しかし、高速に実行できるかどうかは別で、GPUやNPUと比較すると、大量のデータに対して同一の処理を並列に行うことは得意ではありません。それは、次のような事情があるためです。

- 並列演算に必要な演算器の絶対数が少ない
小さな処理を効率的に実行するためのハードウェア

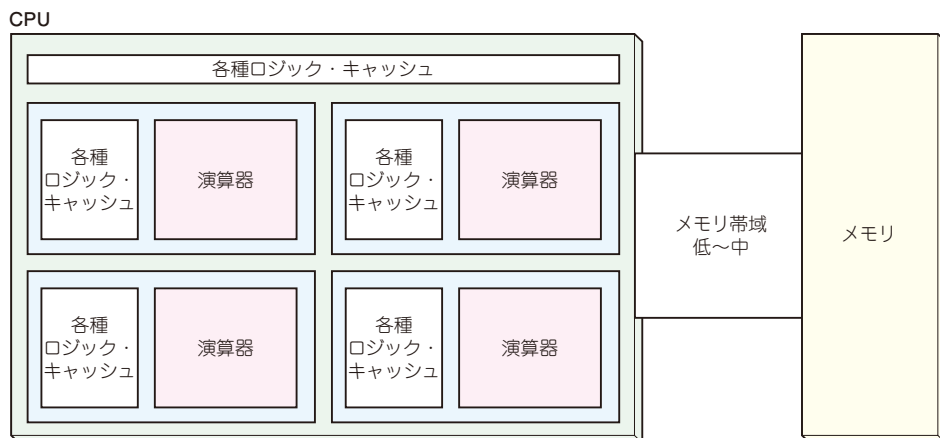


図1
CPUのハードウェア内部構造のイメージ
あくまでイメージであるため、正確でない部分を含む