Abst

ディープニューラルネットワーク（DNN）がコンピュータビジョンの分野で重要な技術として登場して以来、ImageNetの分類チャレンジは、最先端の技術を発展させる上で大きな役割を果たしてきました。精度は着実に向上していますが、受賞モデルのリソース利用については適切に考慮されていません。本研究では，実用上重要な指標である精度，メモリフットプリント，パラメータ，演算回数，推論時間，消費電力を総合的に分析しました。主な発見は以下の通りです。(1）消費電力は、バッチサイズやアーキテクチャに依存しない、（2）精度と推論時間は双曲線の関係にある、（3）エネルギー制約は、達成可能な最大精度とモデルの複雑さの上界である、（4）演算数は推論時間の信頼できる推定値である。私たちの分析は、効率的なDNNの設計やエンジニアリングに役立つ、説得力のある情報を提供してくれると信じています。

1. Intro

2012年のImageNetコンテストで、ディープニューラルネットワーク（DNN）を使用した最初のエントリーであるAlexNet（Krizhevskyら、2012年）が躍進して以来（Russakovskyら、2015年）、より良い性能を得るために、複雑さを増した複数のDNNがこの課題に提出されている。

ImageNetの分類問題では、実際の推論時間に関わらず、多クラス分類問題の枠組みの中で最高の精度を得ることが最終的な目標となっています。これがいくつかの問題を生んでいると考えています。まず、あるモデルの複数の学習済みインスタンスを、各検証画像の複数の類似したインスタンスに対して実行することが、現在では普通に行われています。これは、モデルの平均化やDNNのアンサンブルとも呼ばれていますが、公表されている精度を達成するために推論時に必要な計算量は劇的に増加します。第二に、モデルの選択は、投稿者によって検証画像でのモデル（アンサンブル）の評価回数が異なるため、報告された精度は特定のサンプリング技術（およびアンサンブルのサイズ）に偏ったものとなり、妨げとなります。第三に、これらのモデルの実用的なアプリケーションの重要な要素であり、リソースの利用、消費電力、およびレイテンシーに影響を与える推論時間を短縮するためのインセンティブが現在のところありません。

この記事は、過去4年間にImageNetチャレンジに提出された最先端のDNNアーキテクチャを、必要な計算量と精度の観点から比較することを目的としています。これらのアーキテクチャを、精度、メモリフットプリント、パラメータ、演算数、推論時間、消費電力といった、実際の導入時のリソース利用に関する複数の指標で比較します。この論文の目的は、これらの数値の重要性を強調することです。これらの数値は、実用的な展開やアプリケーションにおいて、これらのネットワークを最適化するための必須のハード制約となります。

2. 方法

異なるモデルの品質を比較するために，文献に報告されている精度の値を収集して分析した．その結果、サンプリング手法の違いによって、資源の利用状況を直接比較することができないことがすぐにわかりました。例えば、VGG-16 (Simonyan & Zisserman, 2014) と GoogLeNet (Szegedy et al., 2014) を1回実行したときのセントラル・クロップ（トップ5検証）誤差は、それぞれ8.70%と10.07%であり、VGG-16はGoogLeNetよりも性能が高いことが明らかになりました。しかし、10個のクロップでモデルを実行すると、誤差はそれぞれ9.33%、9.15%となり、1個のクロップではVGG-16はGoogLeNetよりも性能が劣ることになります。このような理由から、私たちは、単一のセントラル・クロップ・サンプリング手法を用いたすべてのネットワークのトップ1精度3の再評価に基づいて分析することにしました（Zagoruyko, 2016）。

推論時間とメモリ使用量の測定には，Torch7（Collobert et al.，2011）とcuDNN-v5（Chetlur et al.，2014）およびCUDA-v8バックエンドを使用した．すべての実験は、JetPack-2.3 NVIDIA Jetson TX1ボード（nVIDIA）上で行われました。このボードは、64 ビットの ARM R A57 CPU、1 T-Flop／s の 256 コア NVIDIA Maxwell GPU、4 GB LPDDR4 の共有 RAM を備えた、組み込み型ビジュアルコンピューティングシステムです。なお，ネットワークアーキテクチャの違いをより明確にするために，リソースが限られたデバイスを使用していますが，NVIDIA K40やTitan Xなどの最新のGPUでも同様の結果が得られています。操作回数は、私たちが開発したオープンソースのツールを使って取得しました（Paszke, 2016）。消費電力の測定には、Keysight 1146Bホール効果電流プローブを、サンプリング周期2秒、サンプルレート50 kSa／sのKeysight MSO-X 2024A 200MHzデジタルオシロスコープと組み合わせて使用しました。システムの電源は、Keysight E3645A GPIB制御のDC電源を使用しました。

3. Results

このセクションでは、我々の結果と比較を報告します。以下のDDNを分析しました。AlexNet (Krizhevsky et al., 2012), batch normalised AlexNet (Zagoruyko, 2016), batch normalised Network In Network (NIN) (Lin et al., 2013), ENet (Paszke et al., 2016) for ImageNet (Culurciello, 2016), GoogLeNet (Szegedy et al, 2014）、VGG-16 and -19（Simonyan & Zisserman, 2014）、ResNet-18, -34, -50, -101 and -152（He et al., 2015）、Inception-v3（Szegedy et al., 2015）、Inception-v4（Szegedy et al., 2016）は、ImageNet（Russakovsky et al., 2015）の課題で、この4年間で、最高の性能を得たからです。

3.1. Accuracy

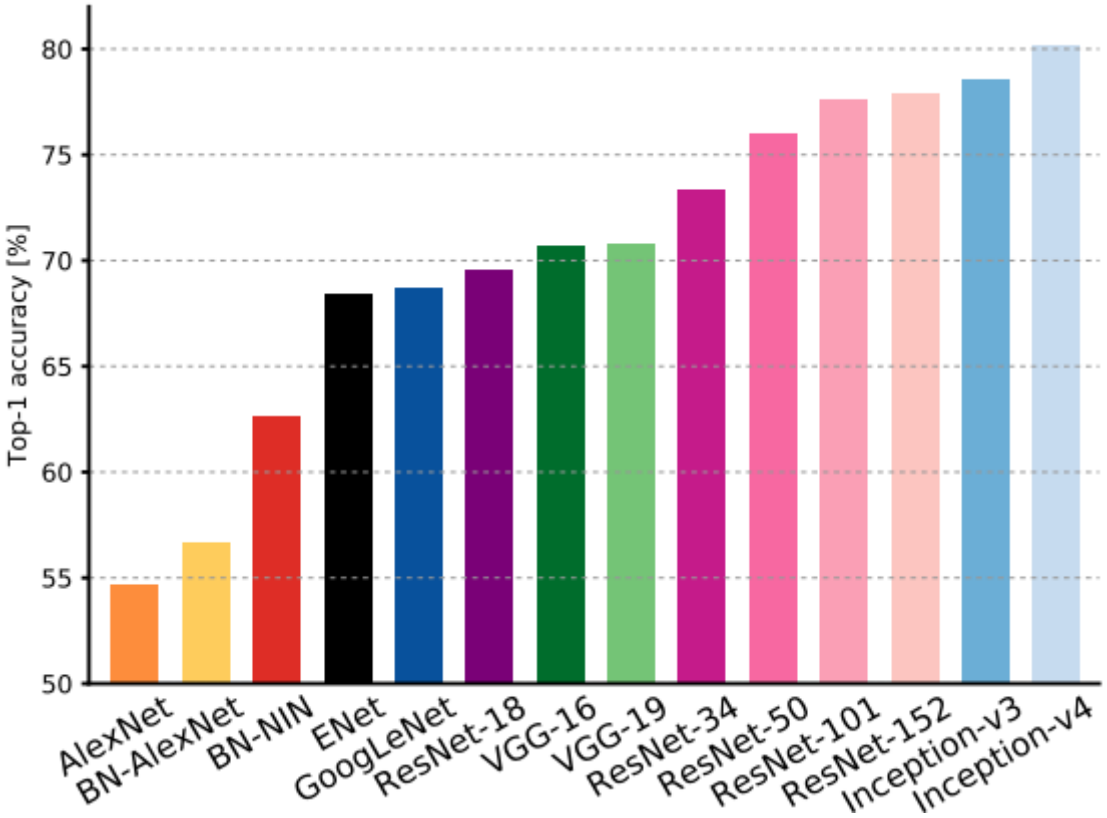

図1: **Top1 vs ネットワークの比較.** トップスコアを獲得したシングルモデル・アーキテクチャーのシングルクロップ・トップ1検証精度。この図では、異なるアーキテクチャとそれに対応する著者を効果的に区別するために、この出版物全体で使用されるカラースキームの選択を紹介しています。同じグループのネットワークは同じ色調を持っていることに注意してください。例えば、ResNetはすべてピンクのバリエーションです。

図1は、ImageNetチャレンジに提出された最も関連性の高いエントリーのワンクロップ精度を、左端のAlexNet（Krizhevsky et al, 2012）から、最も性能の高いInception-v4（Szegedy et al, 2016）まで示したものです。最新のResNetとInceptionのアーキテクチャは、他のすべてのアーキテクチャを少なくとも7%の大差で上回っています。

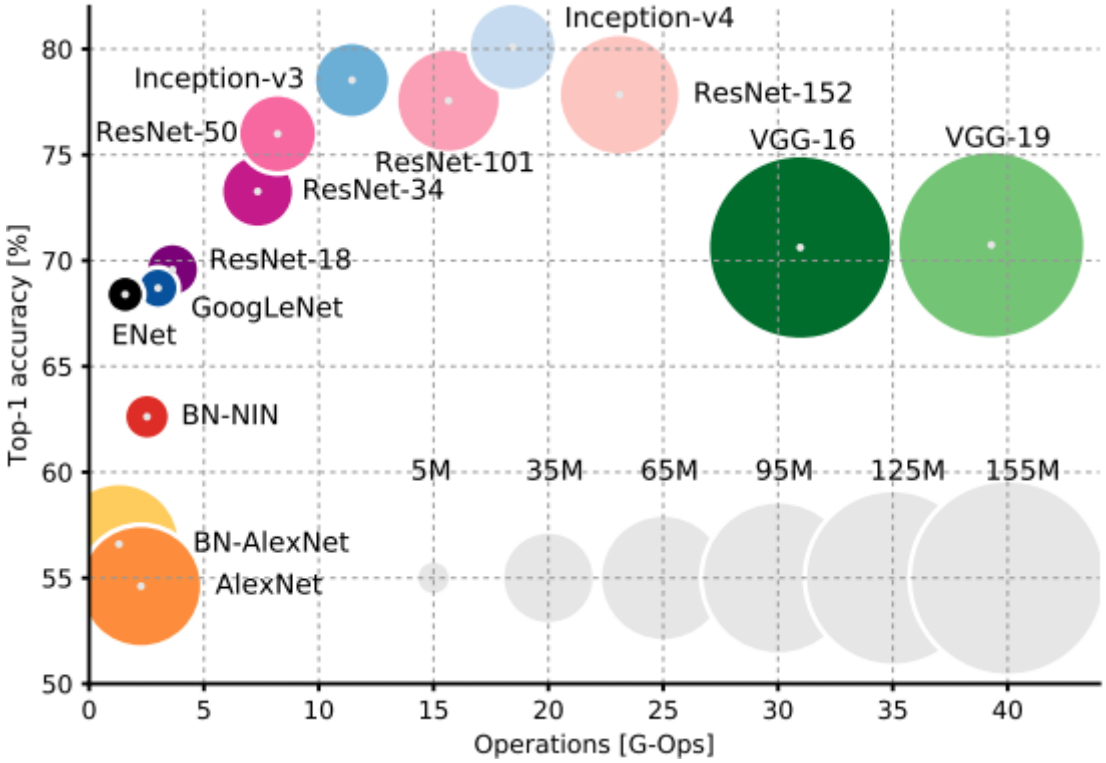

図2: **Top1対演算、サイズ∝パラメータ。** Top-1ワンクロップの精度と1回のフォワードパスに必要な演算量の関係。ブロブの大きさは、ネットワークパラメータの数に比例している。右下には、5×106から155×106個のパラメータにまたがる凡例が報告されている。これらの図はいずれも同じY軸を持ち，灰色のドットがblobの中心を示している.

図2では、計算コストとネットワークのパラメータ数も表示されているため、精度の高さがよりわかりやすくなっています。まず明らかなのは、VGGは多くのアプリケーションで広く使われているにもかかわらず、計算コストとパラメータ数の両方の点で、最もコストの高いアーキテクチャであるということです。その16層と19層の実装は、実際には他のすべてのネットワークから孤立しています。他のアーキテクチャは急峻な直線を描いていますが、InceptionとResNetの最新版では平坦になり始めているようです。これは、このデータセットにおいて、モデルが変曲点に達していることを示唆しているのかもしれません。この変曲点では、複雑さという点でコストが精度の向上を上回るようになります。この傾向が双曲線であることは後ほど説明します。