TSUBAME 共同利用　令和 4 年度　学術利用　成果報告書

人工画像を用いた大規模事前学習
Large Scale Pretraining Using Synthetic Images

横田 理央

産業技術総合研究所　デジタルアーキテクチャ研究センター

邦文抄録（300 字程度）

これまでの研究では学習中に tar ファイルから直接画像を読む方法により inode 数を 1/1000 に削減、パラメータ分散・再計算・オフロードの機構を導入することでメモリ消費をノードに反比例して削減する方法を開発してきた。本課題では、これらの技術を用いて初めて可能になる超巨大な vision transformer の事前学習を実施した。ただし、このような事前学習に必要な画像データは JFT-300M/3B 規模であり、それらのデータセットは非公開である。そこで、本申請課題では Visual Atom や Newton Fractal などの人工画像を用いて事前学習を行った。その結果、人工画像を用いた事前学習を行うことで ImageNet-21k を超える事前学習性能を達成した。

英文抄録（100 words 程度）

Previous studies have developed methods to reduce the number of inodes to 1/1000 by reading images directly from tar files during training, and to reduce memory consumption inversely proportional to nodes by introducing parameter distribution, recalculation, and offloading mechanisms. In this assignment, we conducted pre-training of a very large vision transformer, which is possible only by using these techniques. However, the image data required for such pre-training is on the scale of JFT-300M/3B, and these data sets are not publicly available. Therefore, in this proposal, pre-training was conducted using artificial images such as Visual Atom and Newton Fractal. As a result, we achieved a pre-training performance exceeding ImageNet-21k by pre-training with artificial images.

Keywords: 深層学習、事前学習、Vision Transformer、人工画像、大規模データ

背景と目的

深層学習技術の中で格段に高い精度を発揮しているのが超巨大な transformer の事前学習である。本課題はこのような巨大な transformer の事前学習に必要になる膨大なデータを数式から人工的に作った画像で代替する革新的な基盤技術を提案する。既に ImageNet-21k に関しては同等のデータ量で同じ精度を達成しており、数式から無限に生成できる人工画像で自然画像と同程度の事前学習効果が得られるという事実は、深層学習分野に革命をもたらすものであると予想される。

概要

本課題で用いる計算モデルは深層ニューラルネットの一種である vision transformer である。深層ニューラルネットはこの 20 年間で目覚ましい進歩を遂げてきた。図1に初期の LeNet や LSTM などの深層ニューラルネットから現在の Vision Transformer に至るまでの代表的な深層ニューラルネットの変遷を示す。画像処理分野では LeNet が畳み込みニューラルネットを採用することで精度を向上させ、AlexNet は GPU を用いることで高速にこれを処理できるようにし、ResNet ではスキップ接続とバッチ正規化により学習を容易にした。また、MobileNet の squeeze and excite 機構や EfficientNet の neural architecture search を用いて効率的なニューラルネットが設計できるようになり、精度を維持しつつ小型化できるようになった。一方、自然言語処理分野では LSTM などの再帰的ニューラルネットが長く用いられていたが、2017 年に登場した transformer により大幅な性能向上が実現され、現在では transformer が主流なニューラルネットとなっている。この transformer を画像処理分野で使えるようにしたものが本課題で用いる vision transformer である。LeNet から EfficientNet に至るまで画像処理分野で用いられてきたニューラルネットは、全て畳み込みニューラルネットであったが、vision transformer は畳み込

みのような機構を予め人手で組み込むことなくデータからそのような構造をも学習することができるため、大量のデータがある場合には優位になる。ただし、最も大きい実画像データセット JFT を Google が非公開としていることが画像処理分野全体の発展にとって大きな障壁となっている。本課題では、自動的に生成・ラベル付けできるフラクタルなどの人工画像を用いて vision transformer の事前学習を行い、ImageNet-21k や JFT-300M などの大規模自然画像での事前学習効果と比較する。

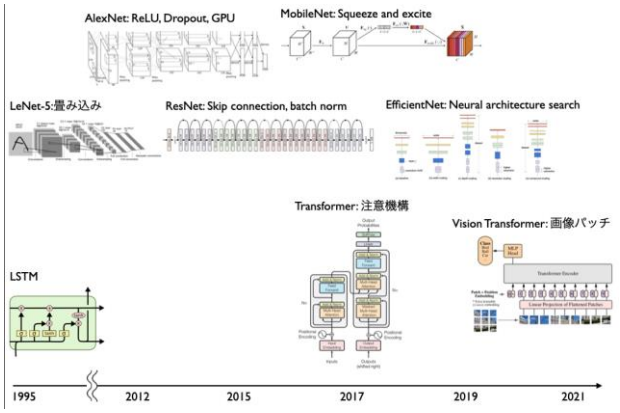

図1 主要な深層ニューラルネットモデルの変遷

結果および考察

本課題では、vision transformer に対して4種類の異なる分散並列化を行った。1つ目は、データをそれぞれのプロセスに分散させる**データ並列**。2つ目は、層内のテンソルを分散させる**テンソル並列**。3つ目は、パラメータを層ごとに分散し層間でパイプライン化を行う**パイプライン並列**。4つ目は、勾配を AllReduce する代わりに ReduceScatter しておき、forward や backward で必要になったときだけ AllGather する手法の **ZeRO** である。これらの手法を「富岳」上で動作させることは継続元課題で既に行っており、今年度はその技術を応用して vision transformer による大規模事前学習を行った。

図2にデータ並列数のみを変化させたときの強スケーリングの様子を示す。左図は更新ステップあたりの計算時間、右図はそのときの総 TFLOPs を示す。この実験では、ノード数によらずグローバルバッチサイズを固定するために、1024 ノード以外の実験ではパラメータの更新を行わずに勾配を累積するサブステップを設けている。左図に示す更新ステップあたりの計算時間がノード数に反比例して減少しているのはこのサブステップの数がノード数に反比例して減少しているためである。Original は最適化前の計算時間で Optimized は最適化後の計算時間を表す。右図に示すように 1024 ノードを用いたときの最適化後の演算性能は 368.6TFLOPs となっている。「富岳」のノードあたりの単精度の理論演算性能が 6.8 TFLOPs であることを考えると改善の余地が大いにあることが分かる。Transformer の演算の 9 割以上は行列積となっているため、高い理論演算性能比が期待できるが、現状では大きな正方行列でないと高い演算性能がでないため、transformer が必要とする行列の次元で性能最適化を行う必要がある。また、Python から行列積のカーネルを呼び出すオーバーヘッドも極力小さくする必要があるため、バッチ行列積のカーネルを用いたり、カーネルフュージョンを行う必要がある。

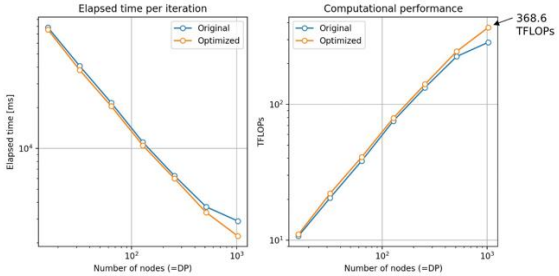

図2 データ並列による強スケーリング

まとめ、今後の課題

本課題では、これまでの継続元課題で培ってきた「富岳」上で大規模深層学習を行うための超分散並列・省メモリ・省 I/O の技術を集積し、画像処理分野の最大の課題となっているデータセットの不足・寡占の問題を、人工画像データセットによって解決する方法を提案することができた。これまでは、インターネットからスクレーピングされた写真などの画像を用いて学習することが最も有効とされてきたが、本課題では vision transformer のような最新のニューラルネットでは自然画像が必ずしも必要でないことを、このような大きな規模で世界で初めて示すことができた。ただし、図2の結

果からも分かるように並列化効率は良いものの、ノード単体性能は「富岳」の理論演算性能の 1/20 程度の実効性能しかでておらず、今後これを向上させることが最重要課題である。