ISSN 2186-5647



2-44

# Hybrid Kernel Convolution Neural Network

日大生産工(院)　○中村　匠海　　日大生産工　山内　ゆかり

## 1　まえがき

現在のコンピュータビジョンの分野では、Vision Transformer[1]が優秀な成績を収めている。特に、Swin Transformer[2]のような階層型Transformerでは、従来型の畳み込みニューラルネットワーク(CNN)[3]の畳み込み層のようなSliding Window戦略を採用することで、よりよい成績を収めた。これは、未だ畳み込み層のような仕組みが有用であることを示している。

そこで、Zhuang Liuらは、CNNの限界を確認しようとした。その理由は、Transformerは優秀であるが、畳み込み層を利用しようとすると、計算コストが大きくなる。しかし、CNNであれば、低コストでこれを利用することができる。また、既存のCNN手法には、見直すべき設計が多くあると考えたためだ。彼らは、ResNet[4]を元に、様々な設計を試した。その結果見つけた有用な設計を導入したものをConvNeXt[5]として発表した。しかし、その際に導入されたカーネルサイズ7の畳み込みは、性能向上に貢献する一方、従来のカーネルサイズ3の畳み込みに比べて、計算量が多くなる。

そこで、本研究では、同一の層内でカーネルサイズ7の畳み込みと、カーネルサイズ3の畳み込みを行う2種のノードを持つ手法を提案する。これにより、精度を維持しつつ、計算量の削減を図る。また、MNISTデータセットを用いた計算機実験により提案手法との精度、処理時間を比較し、計算効率が向上したかどうかについて報告する。

## 2　従来研究

### 2.1　ResNet

ResNet[4]は、He らが提案したCNNの代表的な手法の一つである。このCNNとは、カーネルを用いた畳み込み処理を行うことにより、コンピュータビジョンの分野で良く用いられるNeural Networkのことである。

ResNetの特徴は、残差接続を用いることである。Figure 1に、ResNetの残差接続を示す。残差接続とは、畳み込み層の入力値を出力値に加算させる手法のことである。つまり、畳み込み層は、入力値と加算値の残差を学習することになる。

この手法により、ResNetはそれ以前のものより、深層化することに成功した。深層化したことにより、それ以前のものより良い成績を収めた。

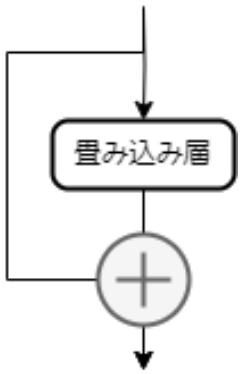


Figure 1　残差接続

### 2.2 ConvNeXt

Liu らのConvNeXt[5] とは、ResNetにSwin Transformerなどの階層型Transformerの設計を取り入れたものである。階層型Transformerには、Transformer由来のSelf-Attention以外にも、様々な設計が取り入れられている。対して、ResNetの設計には古いものも多い。ConvNeXtでは、ResNetの古い設計を階層型Transformerの新しい設計で置き換えたものである。

主な改善点として、入力層の変更、Depthwise Convolutionと逆ボトルネック層の採用、畳み込み層のカーネルサイズを7×7に変更などがある。Figure 2に、ResNetとConvNeXtのブロックを示す。ResNetでは3×3畳み込み層を使用している。対して、ConvNeXtでは7×7 Depthwise Convolutionを使用している。

これは、Swin Transformerのカーネルサイズを参考にしたものである。Liuらは、ResNetに対して従来の3×3より大きいカーネルサイズを用いて、実験を行った。そして、最終的に7×7のカーネルを採用した。

これらの改良により、ConvNeXtは、コンピュータビジョンの複数のタスクにおいて、良い成績を収めた。具体的には、既存の階層型Transformerに対して、複数のタスクで並ぶことができた。

---

Hybrid Kernel Convolution Neural Network

Takumi NAKAMURA and Yukari YAMAUCHI

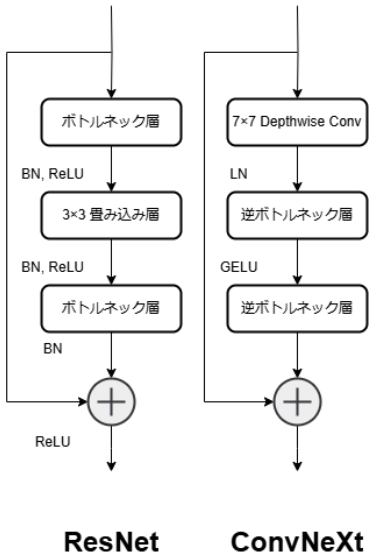


Figure 2　ResNet と ConvNeXt

## 3　提案手法

本研究では、ConvNeXtの7×7畳み込みの一部を3×3畳み込みに戻す手法を提案する。Figure 3に、ConvNeXtと提案手法のイメージを示す。

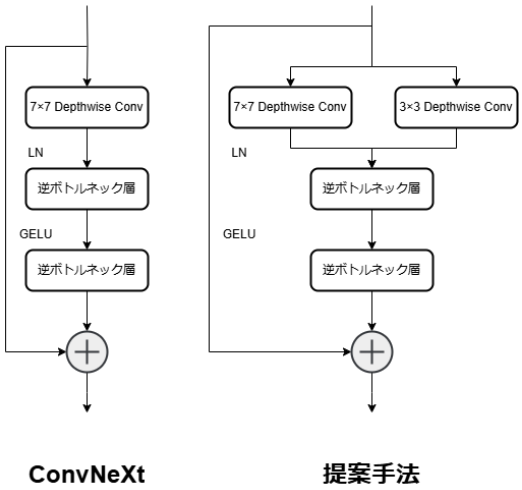


Figure 3　ConvNeXt と提案手法

このように、提案手法ではConvNeXtの7×7 Depthwise Convolutionの一部を、3×3 Depthwise Convolutionに変更する。

このような変更を行った理由は、7×7カーネルと3×3カーネルの両者の利点を両取りしようと考えたためである。7×7のカーネルは、3×3と比べ、より良い成績を収めることができる。しかし、7×7カーネルは、3×3より計算量が多くなってしまう。ここで、7×7と3×3を混合して使用し、両者の利点を両取りできるのではないかと考えた。

提案手法のその他細かな設計はConvNeXtを参考する。例として、逆ボトルネック層の採用、入力層として4×4カーネル、ストライド4の畳み込み層の採用、Layer正規化とGELUの採用などである。

## 4　実験および検討

MNISTデータセットを用い、ConvNeXtと提案手法の精度を比較する。また、処理に要した時間を比較し、計算量削減の効果があったかどうか検証する。

## 5　まとめ

本研究では、ConvNeXtの7×7 Depthwise Convolutionによる計算量の増加に対して、同一層内の一部の7×7カーネルを3×3カーネルに置き換える手法を提案した。これにより、7×7カーネルのみを使うConvNeXtより、計算量を削減しつつも、精度を維持することが可能であると考えられる。

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale.", arXiv preprint arXiv :2010.11929 (2020)

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows.", Proc. the IEEE/CVF international conference on computer vision (2021)

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Proc. NIPS (2012)

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, Proc. IEEE conference on CVPR (2016)

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, “A ConvNet for the 2020s”, Proc. IEEE/CVF conference on CVPR (2022)