立教大学学術推進特別重点資金（立教ＳＦＲ）
大学院学生研究
２０２２年度研究成果報告書

研究科名	立教大学大学院　　人工知能科学研究科　人工知能科学専攻		
研究代表者 （2023年3月現在のものを記入）	在籍課程・学年 □博士前期課程　年 ■博士後期課程１年	氏　名 立浪　祐貴	
指導教員	所属部局・職名 人工知能科学研究科　准教授	氏　名 瀧　雅人	
自然・人文・社会の別	自然	**個人・共同の別**	個人
研究課題	画像認識・生成に有効な非畳み込みのニューラルネットワークの開発		
研究組織 （研究代表者・共同研究者） ※2023年3月現在のものを記入	在籍研究科・専攻・課程・学年 研究代表者 立教大学人工知能科学研究科 博士後期課程 1 年	氏　名 立浪　祐貴	
研究期間	２０２２　　年度		
研究経費 （1円単位）	（支出金額）　300,000円　／（採択金額）　300,000円		

研究の概要（200～300字で記入、図･グラフ等は使用しないこと。）

コンピュータービジョンでは、畳み込みニューラルネットワーク(CNN)が主流だったが、2020年以降、Transformerベースの手法への移行が起こっている。さらに、多層パーセプトロン(MLP)モデルでもTransformerに準ずる性能が報告されており、現在のアーキテクチャは優劣がつけがたい状況にある。今後の研究では、自己注意やMLPなどに代表されるToken-Mixerの新しいカテゴリの探索が必要であり、その有効性や特性が明らかにされることで画像認識のコミュニティが活性化される。本研究では、新しいToken-Mixerを3種類提案し、RNNのようなこれまで真剣に検討されてこなかったToken-Mixerであっても、最先端の性能を達成できることを示した。

キーワード（研究内容をよく表しているものを３項目以内で記入。）

〔　深層学習　〕〔　画像認識　〕〔　Transformer　〕

研究成果の概要（図･グラフ等は使用しないこと。）

1. RaftMLP

RaftMLP は、MLP-Mixer のマクロアーキテクチャをベースにしているニューラルネットワークで、いくつかの改良が加えられている。従来手法の MLP-Mixer は、Token 間の関係を捉えるために MLP を使用することで、空間的な情報について考慮できる。RaftMLP では、空間的な MLP に制約を加え、縦方向と横方向の関係を捉えるそれぞれの MLP に分割し、直列に繋げるモジュールに置き換えた。これは、画像には空間的な位置情報が存在するという帰納バイアスを利用した設計である。また、RaftMLP は通常のパッチ埋め込みに代わる、マルチスケールパッチ埋め込みを搭載している。Vision Transformer や MLP-Mixer などのアーキテクチャでは、パッチ埋め込みという手法が用いられている。この埋め込みには、通常、カーネルサイズとストライドが共に 16x16 の畳み込みが用いられる。それに対して、RaftMLP には、複数のカーネルを利用し、Unfold 関数で展開してから Point-wise 全結合層に入力する新しい埋め込み方法が採用されている。

ImageNet1K データセットのみで学習した RaftMLP は、Top-1 精度 79.4%を達成している。この精度は、MLP-Mixer に比べて 3.9%高く、FLOPs が 48.4%削減されている。効率的なアーキテクチャである。ただし、実測のスループットが悪いことや、入力画像の解像度に柔軟ではないなどの課題も存在する。

RaftMLP についての研究は、前年度までに研究の大枠は固まっていたが、細部を訂正し今年度投稿した。その結果、コンピュータービジョンの査読付き国際会議 The 16th Asian Conference on Computer Vision (ACCV 2022)に採択された。

2. Sequencer

Sequencer は、Vision Transformer やその類似のアーキテクチャにおいて、自己注意機構によって表現されていた長期的な依存関係を LSTM で表現したアーキテクチャである。かつて、RNN や LSTM を使って画像認識する研究もあったが、下火になっていた。このような RNN モジュールを Transformer に似た現代的なマクロアーキテクチャに当てはめ、リバイバルさせた。

Sequencer アーキテクチャは、パッチ埋め込みののち、4 つの Sequencer2D ブロックのスタックが含まれている。Sequencer2D ブロックには、著者らが提案した BiLSTM2D 層が主要構成要素の一つである。BiLSTM2D 層は、2 つの双方向 LSTM を使用し、垂直方向と水平方向に対応させている。入力されたデータは、列方向と行方向に分割され、それぞれの列や行は、同じ双方向 LSTM に入力される。これにより、2 次元的なデータに対しても LSTM を適用できるようになっている。最後に、グローバル平均プーリング層を通じて線形分類器で分類される。Sequencer のアーキテクチャは、パラメータ数や層の数が異なる Sequencer2D-S、Sequencer2D-M、Sequencer2D-L の 3 つのバージョンがある。

ImageNet の分類では、Sequencer が最大で 83.4%の Top-1 精度を達成し、高い解像度の画像でファインチューニングすることで、84.6%の Top-1 精度を達成できることが判明した。これまで、RNN を用いた画像分類は ImageNet のような大規模なデータセットで実験をされたことはないため、Sequencer の成功は画像分類の研究に新しい視点を与えた。さらに、Sequencer はアーキテクチャ自体が、入力画像の解像度に柔軟である。それだけでなく、入力画像の解像度が変わっても精度が維持される興味深い特性があることが判明した。

Sequencer についての研究は、今年度投稿し、機械学習の査読付き国際会議 NeurIPS (Neural Information Processing Systems) 2022 に採択された。

研究成果の概要（つづき）

3.DFFormer/CDFFormer
DFFormer と CDFFormer は、提案手法である Dynamic Filter を使用したアーキテクチャである。Dynamic Filter は、GFNet の Global Filter を発展させた Token-Mixer であり、静的なフィルターではなく、サンプルごとに最適なフィルターを決定するための MLP が組み込まれている。DFFormer は Dynamic Filter に焦点を当てたアーキテクチャであり、CDFFormer は Dynamic Filter と畳み込みのハイブリッド型のアーキテクチャである。両方のモデルは、新しい MetaFormer ののマクロアーキテクチャを採用している。

ImageNet の一般的な設定では、DFFormer は 84.8%、CDFFormer は 85.0%の精度を達成している。実験・分析の結果、先行研究の ConvFormer や CAFormer と比較して、DFFormer や CDFFormer は、より人間に近い形状に依存した特徴を持っているが判明した。高い精度だけでなく、モデルの表現の比較分析なども実施しており、DFFormer は自己注意よりは畳み込みに近い表現を学んでいることが判明している。

DFFormer/CDFFormer についての研究成果は、2023 年 3 月に査読付き国際会議に投稿している（投稿済・査読前）。

※この（様式２）に記入の成果の公表を見合わせる必要がある場合は、その理由及び差控え期間等を記入した調書（Ａ４縦型横書き１枚・自由様式）を添付すること。

研究発表（研究によって得られた研究成果を発表した①～④について、該当するものを記入してください。該当するものが多い場合は主要なものを抜粋してください。なお、成果発表を確認できる資料を合わせて研究成果報告書提出フォームより提出してください（紙媒体等、研究成果報告書提出フォームから提出できない場合は、別途リサーチ・イニシアティブセンターへ提出してください）。

①雑誌論文（著者名、論文標題、雑誌名、巻号、発行年、ページ）
②図書（著者名、出版社、書名、発行年、総ページ数）
③シンポジウム・公開講演会等の開催（会名、開催日、開催場所）
④その他（学会発表、研究報告書の印刷等）

※修士論文・博士論文は含みません。

④ 査読付き国際会議

[発表済]

Yuki Tatsunami, Masato Taki, ″Sequencer: Deep LSTM for Image Classification″ NeurIPS (Neural Information Processing Systems on New Orleans, Poster presentation, December 2022.

Yuki Tatsunami, Masato Taki, ″RaftMLP: How Much Can Be Done Without Attention and with Less Spatial Locality?″, The 16th Asian Conference on Computer Vision on Macau (hybrid-online), Poster presentation, December 2022.

[投稿中]

Yuki Tatsunami, Masato Taki, ″FFT-Based Dynamic Token Mixer for Vision″, Submitted to International Conference on Computer Vision 2023.