H-055

MeMViT の時間方向へのマルチスケール拡張

神谷広大　志水秀熙　玉木徹
名古屋工業大学
E-mail: {k.kamiya.865,s.shimizu.562}@nitech.jp, tamaki.toru@nitech.ac.jp

1 はじめに

本研究では，近年様々なタスクで有効性が実証されている Vision Transformer（ViT）[1] を時間方向に改良を加えた Memory-Augmented Multiscale Vision Transformer (MeMViT) [5] に焦点を当てる．MeMViT はメモリをキャッシュすることで，従来の ViT よりも計算量を削減しつつ性能を保つことに成功している．しかし MeMViT はモデル内の全ての層で同じようにメモリを保存しているため，モデル内の層は階層が上がっていくにつれ時間的な受容野が大きくなり，1 つのメモリが対応するフレームが多くなる．そこで，階層が上がるにつれて保存するメモリの数を少なくする時間方向のマルチスケールを導入すれば，MeMViT の性能を維持しつつさらに計算量を減らすことが期待される．本稿では今回は MeMViT のメモリ部分に前述の改良を加えたモデルを検証し，その性能を従来の MeMViT と比較する．

2 関連研究

2.1 ViT

Vision Transformer (ViT) [1] は，画像から局所的な特徴を抽出する CNN とは異なり，画像を固定サイズのパッチに分割し，各パッチを線形に埋め込み，エンコーダに入力する．

2.2 MViT

Multiscare Vision Transformer (MViT) [2] は，複数の空間スケールで画像パッチを処理する階層構造を導入した．このマルチスケール特徴抽出により，異なるスケールのオブジェクトの認識を必要とするタスクでより優れた性能を達成することができる．

2.3 MeMViT

Memory augmented Multiscale Vision Transformer (MeMViT) [5] は，MViT の各レイヤーにメモリを導入することで，長期的な時間依存関係をモデル化するように改良したモデルである．このモデルでは動画像をオンラインで処理し，各勾配更新においてメモリをキャッシュする．これにより，少ない計算量で時間的に離れた依存関係を学習することができる．

3 手法

本研究では，動画データセットから逐次的にフレームを読み込み，複数フレームをまとめたクリップを用いた動作認識を検討する．そのために，本節では動画から逐次的にクリップを作成する処理 [3] と，マルチスケールに拡張した MeMViT のメモリ保存手法について述べる．

3.1 動画クリップ

データセット中のある動画像を $v \in \mathbb{R}^{F\times 3\times H\times W}$ とする．ここで F は動画像のフレーム数，H, W はフレームの高さと幅である．この動画に対して動作カテゴリがアノテーションされているとする．

この動画像 v から連続的に切り出した動画クリップを $x_t \in \mathbb{R}^{T\times 3\times H\times W}$ とする．

$$x_t(i) = v(s_t + is), \quad i = 0, \ldots, T-1 \tag{1}$$

ここで s_t は第 t クリップ $(t = 0, 1, \ldots)$ の開始フレームであり，T は動画クリップのフレーム数，s はストライド（フレーム間隔）である．

3.2 MeMViT

MeMViT [5] は，入力された動画クリップ x_t を各レイヤーで処理する際に，そこで計算されたキー $\overline{K}^{(t)}$ と値 $\overline{V}^{(t)}$ を以下のようにメモリにキャッシュする．

$$\overline{K}^{(t)} := \left[\hat{K}^{(t-M)}, \ldots, f(\mathrm{sg}(\overline{K}^{(t-1)})), \overline{K}^{(t)}\right], \tag{2}$$

$$\overline{V}^{(t)} := \left[\hat{V}^{(t-M)}, \ldots, f(\mathrm{sg}(\overline{V}^{(t-1)})), \overline{V}^{(t)}\right], \tag{3}$$

なお $\hat{V}^{(t-m)} = f(\mathrm{sg}(\overline{V}^{(t-m)}))$ である．つまり過去 M 時刻までキーと値をキャッシュし，これを用いて現在時刻のキー $K^{(t)}$ と値 $V^{(t)}$ を計算して，自己アテンションに利用する．また長い動画像の特徴を効率的にメモリに乗せるために，メモリを関数 f で圧縮し，2 時刻以前の値は圧縮された値を再利用している．

3.3 レイヤーと受容野

MeMViT の各レイヤーにおける時間的な受容野を図 1 に示す．各矩形はメモリ中のキーと値を表しており，矢印はメモリ間の依存関係を表している．上位レイヤーのメモリは時間的に広い受容野を獲得しており，した

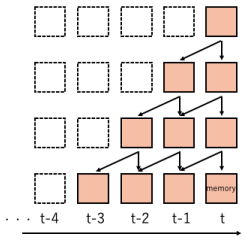

図 1: 最上位レイヤーのメモリの時間的受容野

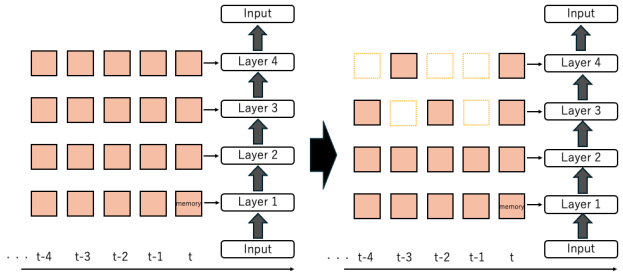

図 2: マルチスケール拡張型 MeMViT

がって上位のレイヤーほど隣接する時刻のメモリ間で時間的な受容野に重複が多くなる．

3.4 時間的なマルチスケール拡張

本研究では，MeMViT のメモリを時間的にマルチスケールに拡張する．つまり，上位のレイヤーほど，保持するメモリの数を削減する．これにより，少数ながらも広範囲の受容野に対応したメモリを実現する．

具体的には，従来の MeMViT の処理に，上位レイヤーのメモリを各時刻で保存するのではなく，複数時刻に 1 回保存するという操作を導入する．図 2（左）は従来の MeMViT のメモリの保存方法を表している．図 2（右）は提案手法のメモリ保存方法であり，最上位レイヤーでは 3 回に 1 回，その次のレイヤーでは 2 回に 1 回の保存を行った様子を表している．ここでレイヤー L での保存方法について，具体的な式を式 (4) (5) に示す．

ここで S は保存する頻度， mod は剰余演算を表し，$(t-(t \bmod S))$ は t が S の倍数でない場合の最後の保持される時刻を示している．

これにより，上層部での広範囲の需要野を保ちつつ，メモリ数を削減することができ，計算効率の向上が見込める．

4 実験

4.1 実験設定

実験には動作認識データセットである UCF101 [4], HMDB51 [6]，Kinetics400 [7] を使用する．比較するモデルは ViT [1]，MViT [2]，MeMViT [5]，提案手法であるマルチスケール拡張型 MeMViT の 4 つである．また，性能における評価指標には validation の top1 を，計算効率における評価指標には GFLOPS を単位として評価に用いる．

4.2 実験

実験では従来の MeMViT に対して，メモリ削減を行い，従来の ViT モデルとの計算効率の比較を行う．また，層ごとに保存方法を変え比較し，最も計算効率のよい保存方法の検証を行う．具体的には以下の内容で実験を予定している．

- 3 つのデータセットに対して，ViT,MVit,MeMViT, そして提案手法であるマルチスケール拡張型 MeMViT で動作認識を行い，性能，計算時間を比較する
- マルチスケール拡張型 MeMViT の最上層のレイヤーで保存方法を 2 回に 1 回，3 回に 1 回のように変え，計算効率の良いメモリの削減方法を検証する．さらに，上から 2 番目，3 番目の層にもメモリ削減を行い，最も効率の良いメモリ削減方法を検証する

5 おわりに

本稿では MeMViT に対して新しいメモリ保存方法の提案を行った．この手法によって従来の ViT モデルと比較して性能を落とすことなく計算効率を上げることが可能である．

謝辞

本研究の一部は，JSPS 科研費 JP22K12090 の助成を受けた．

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, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1, 2

[2] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers, 2021. 1, 2

[3] Kodai Kamiya and Toru Tamaki. Multi-model learning by sequential reading of untrimmed videos for action recognition. In *The 30th International Workshop on Frontiers of Computer Vision (IW-FCV2024)*. The Institute of Electrical Engineers of Japan, 2024. 1

$$\overline{K}_{(L)}^{(t)} := \left[\hat{K}^{(t-M)-(t \bmod S)}, \hat{K}^{(t-M+S)-(t \bmod S)}, \ldots, \overline{K}^{(t-(t \bmod S))}\right] \tag{4}$$

$$\overline{K}_{(L)}^{(t)} := \left[\hat{V}^{(t-M)-(t \bmod S)}, \hat{V}^{(t-M+S)-(t \bmod S)}, \ldots, \overline{V}^{(t-(t \bmod S))}\right] \tag{5}$$

[4] Amir Roshan Zamir Khurram Soomro and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. November 2012. 2

[5] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition, 2022. 1, 2

[6] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset, 2021. 2

[7] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. The kinetics human action video dataset. 2017. 2