Wavelet Masked Autoencoder による Wavelet Transformer の精度向上

ER19069 福井 雅弥　　指導教授：藤吉 弘亘

1. はじめに

Wavelet Transformer [1] は，Wavelet 変換した画像を Vision Transformer (ViT) に入力する手法であり，明示的に物体の高域成分や低域成分，スケールの詳細な補助情報との関係性を考慮した特徴量の抽出を可能とした．ViT と比べて視覚的説明性が向上する一方で，精度が低下するという問題がある．そこで本研究では，Wavelet Transformer の事前学習に Masked Autoencoder (MAE) [2] を導入した Wavelet Masked Autoencoder (Wavelet MAE) を提案する．MAE は入力画像の一部にマスクを行い，マスクした領域のピクセルを予測することで学習を行う．そのため，Wavelet MAE は成分間やスケール間の対応関係を捉える事前学習が可能となる．

2. Wavelet Transformer

Wavelet Transformer は，Wavelet 変換により入力情報を予め人間が解釈可能な単位へ分割して ViT に入力する．このとき，ViT に入力する各パッチは Wavelet 変換によってスケール情報やテクスチャ情報に対応しているため，Self-Attention 機構の Attention から，スケールやテクスチャなどの詳細な情報と判断根拠の関係を人間へフィードバックできる．

3. 提案手法

提案する Wavelet MAE のネットワーク構造を図 1 に示す．Wavelet MAE は，入力画像に対して Wavelet 変換を低域成分に 4 回再帰的に適用する．Wavelet 変換後の画像をパッチ分割し，ランダムにマスク処理を適用する．マスクしたパッチのピクセルを予測することを pre-text タスクとして自己教師あり学習を行う．

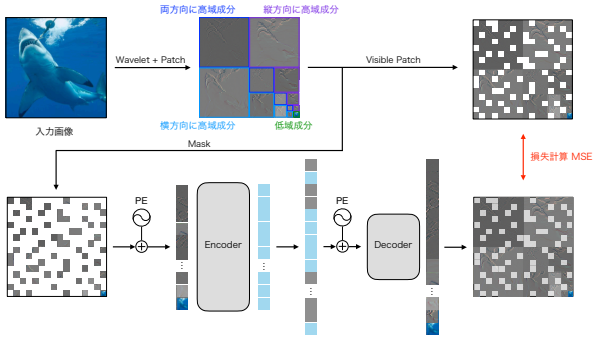

図 1：Wavelet MAE のネットワーク構造

Position Embedding は sin-cos 関数による固定のパラメータで位置情報を表現する．式 (1) に Position Embedding の計算式を示す．

$$PE_{(pos,2i)} = \sin(pos/10000^{4i/D})$$
$$PE_{(pos,2i+1)} = \cos(pos/10000^{4i/D}) \quad (1)$$

Encoder はマスクされていないパッチから特徴を抽出し，Decoder は抽出した特徴量とマスクしたパッチを表すマスクトークンから各パッチのピクセルを出力する．損失関数は，マスクトークンの出力と Wavelet 変換後の画像の平均二乗誤差である．Wavelet MAE で事前学習した重みを Wavelet Transformer の初期値として使用し，ファインチューニングを行う．ファインチューニングでは Decoder 部分を破棄し，Encoder のみを利用する．

4. 評価実験

事前学習において，マスク処理は画像全体の 75%をマスクする．事前学習とファインチューニングのデータセットには ImageNet-1k を使用する． Wavelet MAE のネットワーク構造は，Encoder に ViT-L/14，Decoder に Transformer block を 8 層，埋め込み次元数を 512 次元とした小規模な構造を使用する．

4.1 精度比較

学習結果を表 1 に示す．Wavelet Transformer の精度は 72.2%であるのに対し，提案手法では 79.5%と 7.3pt 精度向上した．

表 1：従来手法との精度比較（*は [2] より引用）

手法	事前学習	精度
ViT	-	82.5*
	✓	84.9*
Wavelet Transformer	-	72.2
	✓	79.5

図 2 に，Wavelet MAE の事前学習により得られた再構成画像を示す．ここでは Wavelet 変換後の画像に対して 75%の確率でランダムにマスク処理を適用する．図 2 より，Wavelet 変換した画像の各成分をみると，スケールやテクスチャ，色情報をある程度復元可能であることが確認できる．

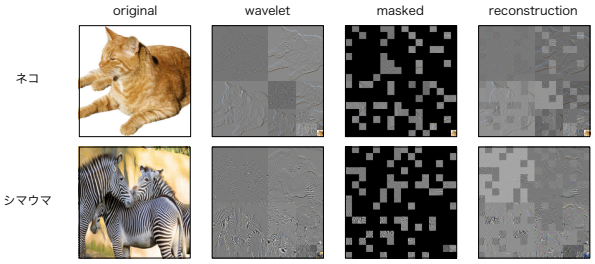

図 2：Wavelet MAE で事前学習した際の再構成画像

4.2 アテンションを用いた逆 Wavelet 変換

Wavelet Transformer から得られたアテンションを Wavelet 変換後のパッチ特徴に重み付けし，逆 Wavelet 変換を施すことで，Wavelet MAE がどのパッチ情報を重要としているかを確認する．図 3 にアテンションを用いた逆 Wavelet 変換画像を示す．アテンションの可視化には Attention Rollout を用いる．図 3 より，低域成分にアテンションが高くなる傾向にあることが分かる．これは，ImageNet で学習した場合，特定のパッチ情報を重視するような特徴量の捉え方をしていると考えられる．

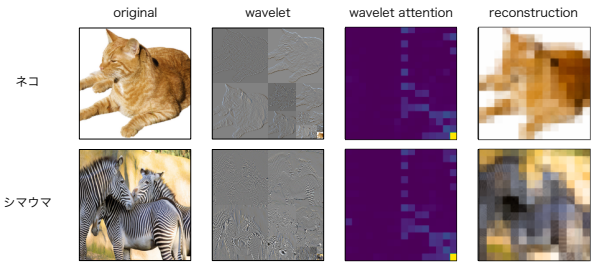

図 3：Attention を用いた逆 Wavelet 変換

5. おわりに

本研究では，Wavelet 変換による成分間やスケール間の対応関係を獲得する事前学習法である Wavelet MAE を提案した．提案手法では，Wavelet Transformer と比べて 7.3pt 精度が向上した．一方で，アテンションマップは偏りのある関係性となった．今後は，アテンションマップの改善を行う．

参考文献

[1] T. Ogata, *et al.*, “Wavelet Transformer: Exploit Interpretable Attentions by Splitting Input Component into Human-Explainable Information”, In MIRU, 2022.

[2] H. Kaiming, *et al.*, “Masked Autoencoders Are Scalable Vision Learners”, In CVPR, 2022.