Wavelet Components ViT によるテクスチャデータの性能及び解釈性の向上に関する研究

TR23007 福井 雅弥　　指導教授：藤吉 弘亘

1. はじめに

画像認識モデルとして，入力全体を考慮して広範な依存関係を捉えることができる Attention 機構を用いた Vision Transformer (ViT) [1] が注目されている．ViT は一般的な画像分類タスクにおいて高い性能を発揮し，Attention を可視化することでモデルの判断根拠を示すことが可能である．文献 [2] の調査により ViT は CNN と比較してテクスチャよりも形状を捉える傾向があると報告されている．そのため，テクスチャデータに対して ViT の性能は低い．そこで，本研究ではテクスチャ分類における ViT の性能と解釈性の向上を目的として，Wavelet 変換を導入した Wavelet Components ViT を提案する．

2. Vision Transformer (ViT)

ViT は入力画像をいくつかのパッチに分割したパッチトークンと分類用のクラストークンを Transformer Encoder に入力する．Transformer Encoder は，パッチ間の関係に基づいて特徴抽出を行う Multi-Head Attention，パッチ内の特徴量を混合することで特徴抽出を行う Multi-Layer Perceptron，正規化を行う Layer Normalization で構成されている．ViT では Multi-Head Attention で捉えたパッチ間の関連度に対して計算し，Attention rollout [4] を適用することで Attention Map を可視化できる．これにより，クラス分類時に重要なパッチをモデルの判断根拠として視覚的に示すことができる．

3. 提案手法

本研究では，ViT の解釈性向上を目的に Wavelet 変換を活用することで空間的に重要な情報に加えて重要なテクスチャ成分を可視化する Wavelet Components ViT を提案する．また，Wavelet Components ViT の事前学習に MIM を用いた成分間の対応関係を促す事前学習を行うことで性能向上を図る．

3.1 Wavelet Components ViT

Wavelet Components ViT は，入力画像を Wavelet 変換し，Wavelet 変換で得られた成分情報を学習する．図 1 に Wavelet Components ViT のネットワーク構造を示す．入力画像に対して Wavelet 変換を 1 回適用して，両方向の高周波成分（HH），縦方向の高周波成分（LH），横方向の高周波成分（HL），低周波成分（LL）の 4 つの成分情報を保持した画像に変換した後，パッチに分割する．低周波成分（LL）のパッチトークンは式 (1)，高周波成分（LH, HL, HH）のパッチトークンは式 (2) で導出できる．

$$Y_l[p] = \sum_k \sum_{m,n} x[m,n] \cdot \phi_k[m,n] \tag{1}$$

$$Y_h^d[p] = \sum_k \sum_{m,n} x[m,n] \cdot \psi_k^d[m,n] \tag{2}$$

ここで，添字 l は低周波成分，添字 h は高周波成分，p は Wavelet 変換後のパッチトークン，k はパッチのインデックス，$x[m,n]$ は入力画像のピクセル値，$\psi_k[m,n]$ は低周波成分を抽出する基底関数，$\phi_k[m,n]$ は高周波成分を抽出する基底関数，d は $d \in \{LH, HL, HH\}$ である．分割したパッチは，それぞれ 1 次元のベクトル p_l，p_h^d に変換し，成分ごとに Linear projection を適用することで，各成分に特化した特徴抽出を可能にする．p_l は式 (3)，p_h^d は式 (4) で導出できる．

$$p_l = \text{Flatten}\left(\{Y_l[p]\}_{p=1}^{P}\right) \tag{3}$$

$$p_h^d = \text{Flatten}\left(\{Y_h^d[p]\}_{p=1}^{P}\right) \tag{4}$$

ここで，Flatten はパッチを 1 次元のベクトルに変換する処理，P はパッチの総数である．さらに，各成分のパッチトークン p_l，p_h^d に同一の位置情報を付与する．式 (5) に Positional Embedding の計算式を示す．

$$\begin{aligned} \text{PE}_{(p,2i)} &= \sin(p/10000^{2i/D}) \\ \text{PE}_{(p,2i+1)} &= \cos(p/10000^{2i/D}) \end{aligned} \tag{5}$$

ここで，p は $p \in \{p_l, p_h^d\}$，i は PE の次元，D は PE の次元数である．Transformer Encoder には Wavelet 変換後のパッチトークンに加えて，Attention map におけるアーチファクトを防止するためにレジスタトークン [3] を入力する．レジスタトークンは特徴抽出時のみ使用し，推論時には破棄する．最後に，Transformer Encoder の出力として得られたクラストークンを MLP Head に入力し，確率分布からクラス分類を行う．

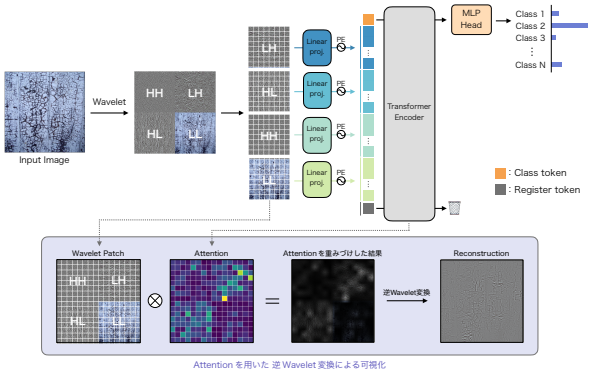

図 1: Wavelet Components ViT のネットワーク構造

3.2 Wavelet Components ViT の事前学習

Wavelet Components ViT の性能向上を目的として Masked Image Modeling (MIM) による事前学習を行う．MIM は，入力画像の一部にマスク処理を行い Encoder へ入力した後，Decoder でマスクした領域を予測することで学習を行う．そのため，Wavelet 変換後の画像において，特定の成分がマスクされた場合でも，残存する他の成分からマスクされたパッチのを予測することができるため，Wavelet 変換との相性が良いと考える．また，提案手法では Cross-Attention を導入することで，Wavelet 変換で得られた成分間の対応関係を捉えた学習を可能にする．Cross-Attention では，Query として Encoder 後の各成分のトークンを用い，Key と Value として Encoder 後の全成分のトークンを用いて，Attention を計算する．入力する．損失関数にはマスクトークンの出力と Wavelet 変換後の画像の Gradient Matching Loss を使用する．図 2 に Wavelet Components ViT の事前学習方法を示す．

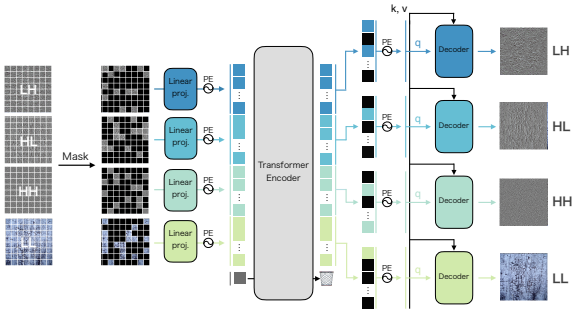

図 2: Wavelet Components ViT の事前学習

3.3 Attention を用いた逆 Wavelet 変換による可視化

推論時に得られた Attention は Wavelet 変換後の各パッチに対応しているため，Attention を解析することで空間

的な情報や各周波数成分に対する重要度を確認することができる．また，Wavelet 変換後の画像に逆 Wavelet 変換することで Wavelet 変換前の原画像を復元できることが知られている．逆 Wavelet 変換の計算式を式 (6) に示す．

$$x[m,n] = \sum_k Y_l[k] \cdot \phi_k[m,n] + \sum_d \sum_k Y_h^d[k] \cdot \psi_k^d[m,n] \tag{6}$$

これらの特性を活用し，モデルの判断根拠を可視化する新たな方法を提案する．Attention を用いた逆 Wavelet 変換による可視化方法を図 1 の紫枠内に示す．具体的には，推論時に得られた Attention を Wavelet 変換後のパッチ特徴に重み付けし，逆 Wavelet 変換を施すことで可視化する．Attention の可視化には Attention rollout [4] を使用する．

4. 評価実験

MIM による事前学習した Wavelet Components ViT をファインチューニングした際の性能評価を行う．事前学習のデータセットには ImageNet-1K，ファインチューニングには代表的な 4 つのテクスチャデータセットである DTD，CUReT，GTOS，KTH-TIPS2 を使用する．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．モデルサイズは，ViT-T/14 を使用する．比較対象には，CNN ベースの手法として ResNet18，ViT, MIM による事前学習をした ViT（MAE），スクラッチ学習した Wavelet Components ViT とする．

4.1 性能比較

表 1 に性能評価の結果を示す．表 1 より，全てのデータセットにおいて，MIM による事前学習を行うことで性能が向上し，提案手法が最も高性能であることが確認できる．

表 1: テクスチャデータセットにおける性能比較結果

Dataset	Method	Pretrain (MIM)	Params	Accuracy [%]
DTD	ResNet18	–	11M	53.1
	ViT	–	5M	42.9
		✓		51.0
	Ours	–	5M	33.4
		✓		**53.4**
CUReT	ResNet18	–	11M	94.0
	ViT	–	5M	93.2
		✓		93.6
	Ours	–	5M	93.7
		✓		**97.3**
GTOS	ResNet18	–	11M	74.9
	ViT	–	5M	75.1
		✓		75.6
	Ours	–	5M	64.6
		✓		**76.0**
KTH-TIPS2	ResNet18	–	11M	75.3
	ViT	–	5M	71.4
		✓		79.1
	Ours	–	5M	70.5
		✓		**80.4**

4.2 解釈性の比較

評価実験で使用したテクスチャデータセットにおける各モデルの判断根拠を可視化し，解釈性の比較を行う．図 3 に解釈性の比較結果を示す．図 3 より，提案手法では全てのテクスチャデータセットに対して重要なエッジやテクスチャ成分を残した画像を表現できており，解釈性の向上を確認できる．

5. モデルが捉える特徴の分析

次に，モデルが形状とテクスチャのどちらを重視した学習をしているかを Stylized-ImageNet (SIN) を用いて評価する．SIN は ImageNet の画像にスタイル変換を適用し，テクスチャを変化させたデータセットである．例えば，猫の画像に対して象のクスチャを用いてスタイル変換した場合，その画像に対してモデルが“猫”と認識した場合は“形状”を重視したモデルであり“象”と認識した場合は“テクスチャ”を重視したモデルであると評価することができる．各モデルは ImageNet-1k で学習した重みを使用する．

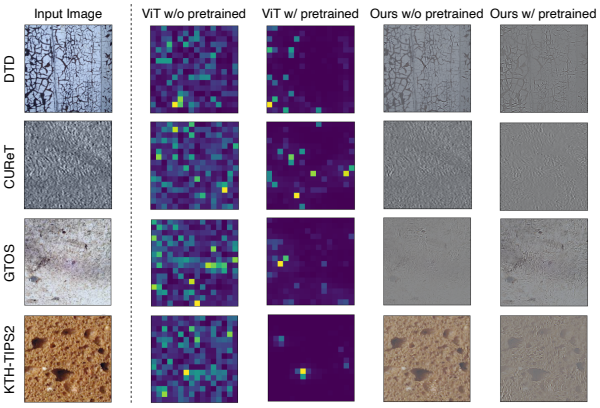

図 3: 解釈性の比較結果

図 4 に分析結果を示す．縦線は各モデルの平均値を表しており，線が左に位置するほど形状重視，右に位置するほどテクスチャ重視であることを表す．図 4 より，提案手法は ViT や ResNet18 と比較して右側に位置していることを確認できる．以上より，提案手法は Wavelet 変換で得られた各成分間の関係性を学習したことで ViT や ResNet18 と比較して，よりテクスチャを捉えたモデルに変化したといえる．

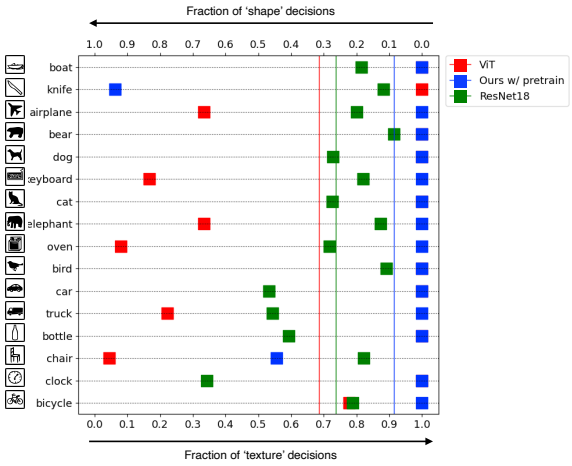

図 4: モデルが形状重視かテクスチャ重視かの分析結果

6. おわりに

本研究では，ViT のテクスチャデータに対する解釈性及び性能の向上を目的として Wavelet Components ViT と Wavelet Components MAE を提案した．代表的な 4 つのテクスチャデータセットで評価し，ViT より優れた性能と解釈性を獲得した．今後は，異常検知等に関するデータセットで評価を行う．

参考文献

[1] A. Dosovitskiy, *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, In ICLR, 2021.

[2] L. Scabini, *et al.*, “A Comparative Survey of Vision Transformers for Feature Extraction in Texture Analysis”, arXiv preprint arXiv:2406.06136, 2024.

[3] T. Darcet, *et al.*, “Vision Transformers Need Registers”, In ICLR, 2024.

[4] S. Abnar, *et al.*, “Quantifying Attention Flow in Transformers”, In ACL, 2020.

研究業績

[1] 福井雅弥，緒方貴紀，平川翼，山下隆義，藤吉弘亘，“Wavelet Components Transformer”, 画像の認識・理解シンポジウム, 2024.