医用画像に適したVision Transformer構成法

Vision Transformer Construction Method Suitable for Medical Image

清水　章人[1)]
指導教員　青木　輝勝[2)]

1）東京工科大学大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻　青木研究室
2）東京工科大学　コンピュータサイエンス学部　コンピュータサイエンス学科　青木研究室

本研究では，糖尿病網膜症の重症度分類における分類精度の向上と計算量削減のために，Vision Transformerにスーパーピクセルセグメンテーションを組み合わせた二つの手法を提案する．これにより，糖尿病網膜症の早期発見が可能になり，糖尿病患者の視力の保護につながる．

深層学習，医用画像，糖尿病網膜症

１．　はじめに

糖尿病網膜症(Diabetic Retinopathy:以下,DR)は，糖尿病患者における主要な合併症の一つであり，世界中で失明の原因となる重要な疾患である．厚生労働省の調査[1]では,2016年時点で国内の糖尿病患者は予備軍を含めると2000万人近くに及び,その中の約三分の一がDRに発症していると言われている.糖尿病患者の視力を保護するためには,早期発見と正確な病期分類が不可欠である．

近年，ディープラーニング技術の発展により，DRの自動診断が進展している．特に，画像認識分野におけるVision Transformer（以下，ViT）の登場は,従来の畳み込みニューラルネットワーク(以下,CNN)を超える性能を示し,注目を集めている．しかし，ViTの計算リソースの消費は大きく，特に微小な特徴を得るために高解像度画像を扱う場合には，その負荷が課題となる．本研究では，ViTにスーパーピクセルセグメンテーションを組み合わせることで，DRの5段階の重症度分類における計算量削減と精度向上を目指す．

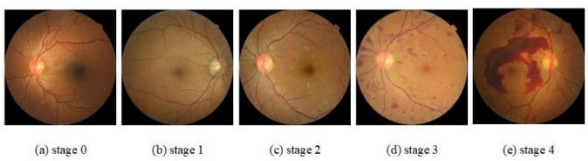

図1：糖尿病網膜症の5段階分類

２．関連研究

２．１．Vision Transformer

Vision Transformer [2]は，2020年にGoogle Researchによって提案されたモデルで，Transformerアーキテクチャを画像認識タスクに適用したものである．従来のCNNとは異なり，ViTは画像を固定サイズのパッチに分割し，それぞれのパッチをトークンとして扱う．これにより，Self-Attention機構を通じてパッチ間の関係性を学習することができる．

２．２．スーパーピクセルセグメンテーション

スーパーピクセルセグメンテーションは，画像を意味的に関連するピクセルの集合であるスーパーピクセルに分割する手法であり，画像解析の前処理として広く利用されている．スーパーピクセルは,画像の冗長性を低減し,局所的な構造情報を強調することで，後続の処理ステップを効率化することができる．

３．提案手法

本研究では，2つのアイデアをもとにVision Transformerにスーパーピクセルセグメンテーションを組み合わせることで，DRの重症度分類における計算量削減と精度向上を目指す．

３．１．パッチ内部のス―パーピクセル化

1 つ目がパッチの内部をスーパーピクセル化する手法である．従来の ViT と同様に画像を固定サイズのパッチに分割し，それぞれのパッチをスーパーピクセル化することで，パッチ内部の情報量を減らし，計算量の削減を行う．

3．2．スーパーピクセルのパッチ化

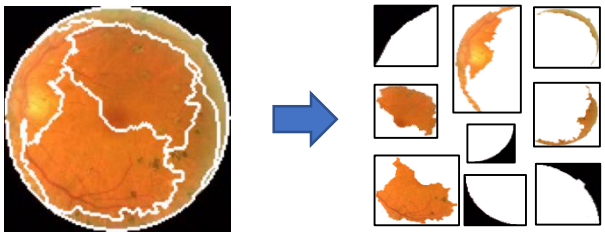

図３：パッチ内部のスーパーピクセル化

2 つ目がスーパーピクセル化したピクセルの集合を 1 つのトークンとして入力する手法である．画像全体に対してスーパーピクセルセグメンテーションを行い，これを画素値 1 で埋めることによって矩形で取り出したものをパッチとして扱う．これにより，従来よりも意味を持ったパッチ分割を行うことができ，精度を維持しつつ，計算量を削減することが可能になる．

4．実験と評価

現状 3.1 における提案手法の実験のみ完了しているため，本章では 3.1 の提案手法の有効性を示す．本研究では kaggle の Diabetic Retinopathy Detection データセットを用いて従来の ViT との分類精度と Multiply-Accumulate Operations（以下，MACs）の比較を行った．このデータセットは計 35,126 枚のカラー眼底画像を、複数の眼科医が疾患の程度に応じてステージ 0 から 4 に分類したものである．

比較するモデルはパッチサイズを 16×16 に設定した従来の ViT と提案手法であり，パッチ枚数と分類精度，計算量の関係を示すために ViT ではパッチサイズ 28×28 でも実験を行っている．

実験結果を表 1 に示す．従来の ViT では MACs が 16.21G であったのに対し，提案手法では 13.87G に計算量が削減され，分類精度に関しては 0.09%の減少に留まっている．一方で，パッチサイズ 28×28 の ViT では，MACs が 5.27G，分類精度が 68.20%と分類精度が 3.00%減少するが，非常に計算量が抑えられており，ViT の計算量がパッチの枚数に大きく依存していることが分かる．

表１：実験結果

	ViT(28x28)	ViT(16x16)	ours(16x16)
MACs	5.27G	16.21G	12.87G
acc	68.20%	71.20%	71.11%

5．おわりに

本研究では，ViT にスーパーピクセルセグメンテーションを適用する 2 つの手法を提案した．3.1 の手法では，DR の分類精度の低下を抑えつつ計算量が抑えられ，提案手法の有効性が示された．

しかし，表 1 の実験結果からわかる通り，ViT の計算量は特にパッチの枚数に依存しており，この点に関して改善の余地がある．3.2 の手法を実装することによって，より意味を持ったパッチに分割することで，少ないパッチ数で計算量を抑えながら分類精度を保つことができると考える．

参考文献

[1] 厚生労働省．「糖尿病患者数の状況」 https://www.mhlw.go.jp/stf/wp/hakusyo/kousei/18/backdata/01-01-02-08.html, (2024/07/26 参照)

[2] 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”, In International Conference on Learning Representations. (ICLR).

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S.Süsstrunk, “SLIC Superpixels Compared to State-of-the-art Superpixel Methods,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, pp. 2274–2282, 2012.