1. Vision Transformerのこれまでとこれから

立花　泰彦　量子科学技術研究開発機構QST病院画像診断課

近年の人工知能（AI）の進歩はとどまるところを知らない。なかでも，特に認知されているのは「ChatGPT」（OpenAI社）などの自然言語領域における大規模モデルであろう。かの領域は，2017年に発表されたTransformerという新技術によって，それこそ爆発的に進歩した。そして，その立役者は，2020年にはついに画像AIの分野にも進出し，Vision Transformer（ViT）と呼ばれ，急激に存在感を高めてきた。本稿では，このViTとはなにか，MRIなどの医用画像においてどう応用されるか，そして，最新の発展状況や今後の展望について簡単に解説する。

ViTとは何か，従来の画像AIと何が違うのか

1. 従来型の画像AI（CNN）

AIはいかなるものであれ，要はデジタルデータから別の有用なデジタルデータを作る機能である（画像AIであれば，入力と出力のどちらか，あるいは両方が画像）。であれば，極端な話，優秀なAIを作るには，入力データを成す「数字」をあらゆる方法で組み合わせ，目的をよく達成する出力を得るパターンを見つければよい。実際，初期のAIはそれに近いものであったが，必然的に効率は悪く，医用画像のような情報量の多いデータへの適用は難しかった。この限界を突破し，実用的な画像AIを実現したのが，畳み込み計算を軸としたAI設計である（convolutional neural network：CNN）。畳み込み計算とは，一定の範囲にある近接ピクセル同士の値を組み合わせて新しい値を作る方法であり（図1），ここでは，「画像であれば，近いピクセル同士の関係が深いに決まっているのだから，それらを組み合わせて計算する方が，何でもかんでも組み合わせる従来法より効率が良いはず」，という効率化の発想が根底にある。実際のAIでは，この畳み込みが幾重にも重ねられることで，超局所の特徴をつなぎ合わせて徐々に広域の特徴に迫っていく，という機能が実現されており，いわば「1本1本の木を見まくることで森を理解しようとするスタイル」のAI設計であったと理解してよい。

かつて大成功したこのスタイルは，多くのタスクに有効であったが，やや相性の悪い相手も存在した。例えば，k-spaceのように，ピクセル同士の「近さ」が情報の関連性と関連しない場合などである。あるいは，肺がんと副腎転移のように，物理的に遠い位置に関連の深い情報がある場合なども該当する。

2. ViTの登場

オリジナルViT[1]があまりにも斬新だったのは，「ピクセル同士の距離の近さ」という情報をまったく使わなかった，いわば伝家の宝刀を捨ててしまったことである。では，どうしたかというと，写真に写っているものが何かを判定するタスクにおいて，画像をまず16×16のパッチに分割し，さらに，それぞれのパッチのピクセルを一列に並べ替え，その1つ1つをトークン（自然言語AIが文章を処理する時の1単位。単語のようなもの）

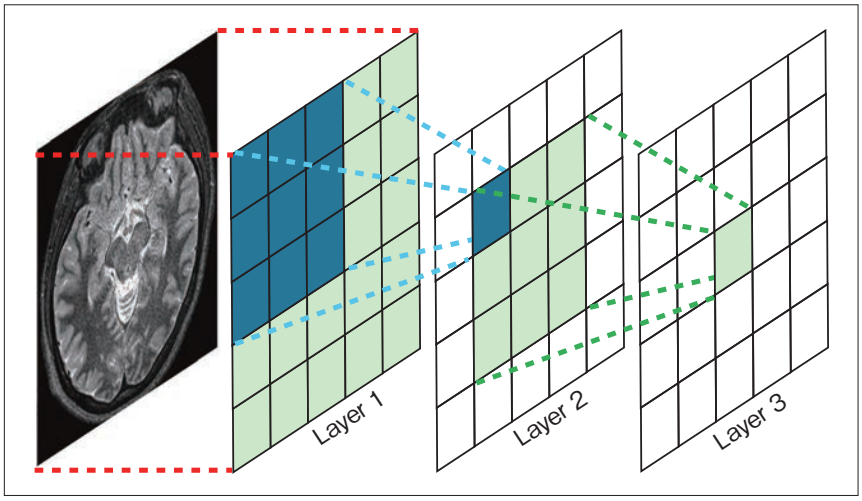

図1　畳み込みの仕組み
元画像の一定範囲のピクセルを組み合わせて次の値を作る（どのように組み合わせるかは学習により最適化される），ということを繰り返していくことで，局所の特徴抽出から徐々に高次の特徴の把握につなげていく。

〈0913-8919/25/￥300/論文/JCOPY〉