Vision Transformer の応用と今後の動向予想

箕浦大晃 †　平川翼 †　山下隆義 †　藤吉弘亘 †
† 中部大学
E-mail: himi1208@mprg.cs.chubu.ac.jp

1 概要

ImageNet による画像認識タスクにおいて Convolutional Neural Network を凌駕する性能を発揮した Vision Transformer (ViT) が登場した 2021 年以降，ViT ベースの手法が様々なコンピュータビジョンタスクで飛躍的発展を見せている．目覚ましい勢いで ViT が発展する一方で，その勢いのあまり類似研究が多数でどこにフォーカスするかが重要になる．このような背景のもと，本講演では ViT の応用例をまとめつつ今後の ViT 周辺に関する動向予想について紹介する．

2 はじめに

2012 年に一般物体画像認識コンペティション ILSVRC で深層学習モデルの畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) が圧倒的な成績を収めて以降，多くの画像認識タスクで CNN が利用されてきた．2021 年に自然言語処理で提案された Transformer をコンピュータビジョンに応用した Vision Transformer (ViT) [1] は，CNN の性能を凌駕することで新たなデファクトスタンダードモデルとして注目を浴びている．Transformer は，Self Attention と呼ばれるトークン間の相互作用を計算する機構により広域特徴を捉えることができる．ViT では，画像を均等なサイズのパッチに分解しトークンとして扱うことで，画像全体の特徴を Self Attention で捉える．CNN は，カーネルにより局所領域を畳み込み処理を繰り返すことで受容野が広がり徐々に広域特徴を捉えることができる．一方で，ViT は Self Attention により浅い層から局所領域を見つつ広域的特徴を捉えることができる．これが ViT の最大の利点と言える．

ViT はシンプルな構造で CNN より性能が高いことから，コンピュータビジョンの主要なタスクに ViT を応用した研究が多く提案されている．目覚ましい勢いで ViT が発展する一方で，その勢いのあまり類似研究が多数でどこにフォーカスするかが重要になる．このような背景のもと，本稿では ViT の基礎的構造を概説しつつ，コンピュータビジョンにおける ViT の改良手法や応用例を紹介する．また，ViT に関する近年の研究傾向から今後の ViT 周辺に関する動向予想を紹介する．

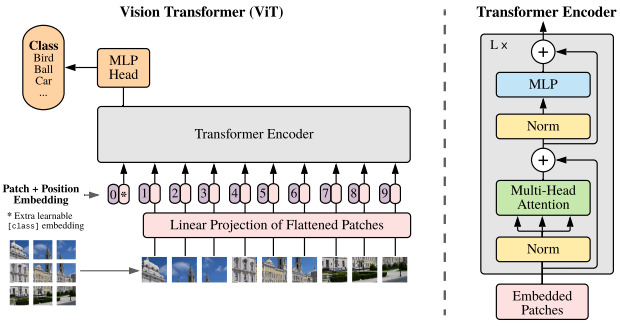

図 1　ViT の概略図．文献 [1] より引用．

3 Vision Transformer の基礎的解説

本章では，ViT の構造について基礎的な解説を行う．図 1 に ViT の概略図を示す．まず，ViT は入力画像を固定パッチに分解し linear projection でパッチ特徴量を得る．ここで，パッチ特徴量にはパッチの位置情報が含まれておらず，位置情報を付与するために学習可能な位置エンコードをパッチ特徴量に加算する．位置情報を付与したパッチ特徴量と物体クラスを識別するためのクラストークンと呼ばれる学習可能なベクトルを Transformer Encoder へ入力する．Transformer Encoder の内部には，Self Attention と呼ばれるパッチ間の対応関係を取得する機構と Feed Forword 層がある．前者はパッチ特徴量を空間方向に混ぜて変換し，後者は次元方向に個別に変換する役割を持つ．これらを繰り返し処理した後に，クラストークンの出力のみを利用してクラス分類を行う．

Self Attention: ここでは Self Attention について述べる．Self Attention はパッチ間のベクトルの内積を求めることでパッチ間の対応関係を取得できる．具体的には Self Attention にはクエリ，キー，バリューそれぞれの特徴ベクトルを求める linear があり，クエリとキーの内積がパッチ間の対応関係を表現している．[1] $\{\mathbf{x}_n\}_{n=1}^{N+1}$ をクラストークンを含む各パッチ特徴量として，クエ

[1] この内積結果は注意重み (Attention weight) と呼ばれ，これを利用すると ViT における判断根拠 (Attention map) を示すことができる．

リ，キー，バリューを以下のように示す．

$$Q = \phi_q(\{\mathbf{x}_n\}_{n=1}^{N+1}) \in \mathbb{R}^{(N+1)\times d} \quad (1)$$
$$K = \phi_k(\{\mathbf{x}_n\}_{n=1}^{N+1}) \in \mathbb{R}^{(N+1)\times d} \quad (2)$$
$$V = \phi_v(\{\mathbf{x}_n\}_{n=1}^{N+1}) \in \mathbb{R}^{(N+1)\times d} \quad (3)$$

ここで，N はパッチ数，d はクエリの次元数，$\phi_\cdot(\cdot)$ は linear，クエリ，キー，バリューのベクトルをひとまとまりの行列としてそれぞれ Q, K, V と表す．式 (4) より，Self Attention はクエリとキー間の内積をクエリの次元数で正規化した値を Softmax 関数に通し，バリューの加重和を行っている．

$$Z = \mathrm{Softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V \quad (4)$$

クエリとキー間の内積からわかるように，各パッチと類似する別のパッチの特徴量を多く取り込み，類似しないパッチ特徴量を少なく取り込む処理をしている．したがって，クエリの各パッチに対応するバリューの重要度をキーで求める，これが Self Attention の本質であると言える．

Multi-Head Attention: 式 (4) より，クエリとキー間の内積計算はそれぞれのベクトル全体で計算し Softmax 関数を通すため，注意重みが 1 つ (少数) のみのピークを持つことになり，小さな関係性が潰れた結果，ネットワーク全体の表現力が落ちる可能性がある．1 つのピークを捉えるより複数のピークを捉えた方がネットワークの表現力は向上し効率的に学習できる．そこで，Self Attention に複数の Attention weight を求めるための Multi-Head Attention が一般的な Transformer モデルに用いられている．これにより，1 つのネットワークだけでアンサンブル効果が期待できる．

ViT における課題: ViT では 3 つの主要な課題があげられる．

- **学習に膨大なデータが必要**．ViT で CNN の性能を凌駕するには 3 億枚の膨大な画像とラベルを保有するデータセットで事前学習し，ImageNet などの下流タスクに転移学習する必要がある．約 128 万枚の ImageNet で学習した場合では，学習データ不足が原因で ViT は CNN の認識精度より劣ることが知られている．これは，画像の上下左右といった局所領域に強い帰納バイアスを習得する CNN とは異なり，ViT はパッチ間の関係を広域で見る帰納バイアスを持つため，膨大なデータで学習しなければ性能向上しないのではないか，と ViT の著者らは考察を述べている．
- **膨大な正解ラベルが必要**．ViT の事前学習では「教師あり学習」をしており，すべてのサンプルに人が正解ラベルを付与するアノテーションが必要になる．それを回避するために Web から自動収集したラベルを用いて膨大なデータセットを構築できるが，ノイズの発生により完璧なデータセットに近づけるためには人の介入が必要不可欠になる．
- **リッチな計算資源が必要**．ViT に限らず CNN でも性能向上するための事前学習モデルの多くは，モデルサイズが大きく膨大なパラメータを有している．そのため，大規模モデルになれば性能が良い GPU や TPU 搭載マシンを複数用意する必要があること，学習に膨大なデータを必要とすることから，金銭コストと時間コストの両方がネックになる．

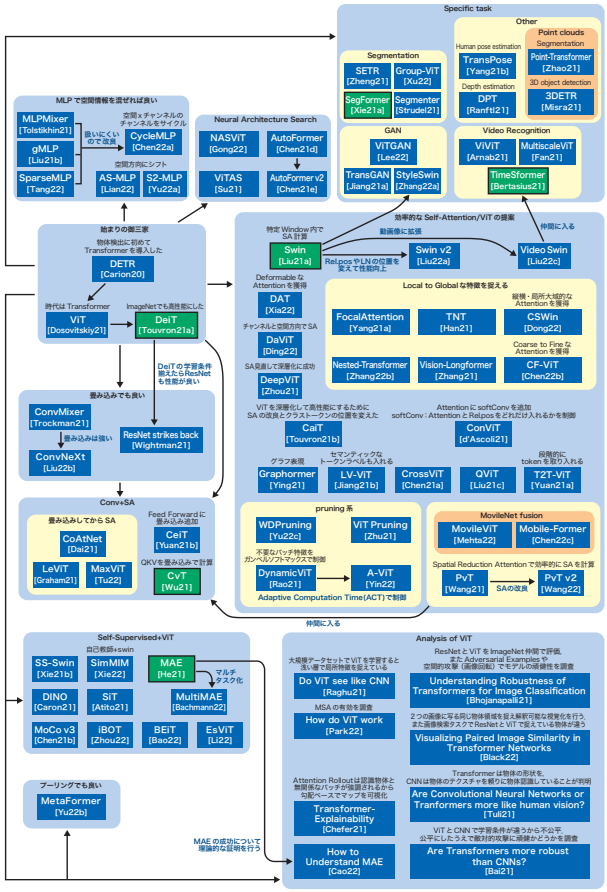

図 2　ViT の派生手法の分類．文献 [2] より引用．

4　Vision Transformer の応用

図 2 に ViT の派生手法の分類を示す．図 2 より，ViT の効率化を筆頭に ViT を各タスクに特化，畳み込みと複合させた ViT など様々な派生手法が提案されている．また，近年では Self Attention を畳み込みや空間方向を計算する全結合，畳み込みも MLP も不要な Pooling などに置換することで ViT と同程度以上の性能を発揮されるなど，ViT 周辺の研究は数多くされている．本章では，第 3 章で述べた ViT における課題という側面

から応用事例を紹介する．なお，緑色でハイライトされている手法は ViT の派生手法の中でも代表的なものであり，文献 [2] で詳細を紹介している．

4.1 ViT に適したデータ拡張による高性能化

膨大なデータで事前学習する必要がある ViT だが，Data-efficient Image Transformers (DeiT) [3] では，CNN を教師モデルとし ViT を生徒モデルとして，教師モデルで獲得した知識を生徒モデルに伝達する蒸留に加え，疑似的に学習用データ数をかさ増しするデータ拡張を利用することで高性能化を実現している．DeiT 以降の文献では，DeiT で設定されたデータ拡張がデファクトスタンダードに用いられている．データ拡張には，学習データを自動最適化させる AutoAugment [4] や RandAugment [5]，コンピュータビジョン分野で一般的に用いられる RandomResizedCrop，RandomHorizontalFlip，ColorJitter，Random Erasing，Mixup [6] 及び CutMix [7] が用いられる．特に，入力画像の一部を mix させる対象に置換させて学習サンプルを作成する CutMix が ViT のデータ拡張で最も認識精度に貢献していることが判明している．一方で，CutMix は ViT が入力画像を固定パッチに分解する際，2 つの画像が mix された領域を含むパッチが作成されノイズとして学習を妨げる可能性がある．そこで，パッチ内に 2 つの画像の mix 領域を含めない MixToken [8] が提案されており，パッチ内にノイズが含まれないことから Mixup や CutMix より性能が向上している．他にも ViT の注意重みを利用してパッチレベルで mix する領域を決定する TokenMix [9]，TransMix [10], Mixpro [11] が提案されている．

また，ViT は画像分類において物体形状を重視して特徴表現を獲得していることが判明している [12]．物体形状を捉えることが重要という仮説に基づき，DeiT III [13] では物体形状を捉えやすくするために Grayscale，Solarization 及び Gaussian Blur を用いたデータ拡張を利用している．さらに，フラクタル幾何学に基づきアルゴリズムに従って自動生成したデータセットを作成した研究もある [14]．このように，ViT に適したデータ拡張を選択することで更に性能向上が期待できるとともに，ViT の理解につながると考えられる．

4.2 自己教師あり学習と ViT の組み合わせ

自己教師あり学習はラベルのないデータに対して擬似的な問題を適用することで，データから自動的にラベルを付与して学習する方法である．正解ラベルを必要としないため，アノテーションにかかる人的コストを削減することができる．ViT と自己教師あり学習の組み合わせには，画像間の関係に基づいて画像から抽出した特徴量を埋め込む対照学習 [15, 16]，入力画像のパッチをランダムにマスクしたトークンから元画像を復元するように学習する Masked Image Modeling (MIM) [17, 18, 19] が提案されている．中でも MIM の代表手法である Masked Autoencoder (MAE) [18] は，入力画像を 75%ランダムにマスクしたのにもかかわらずほぼ正確に入力画像と同レベルに復元できたことからコンピュータビジョンのブレイクスルーとなった．その後，MAE の応用 [21] や対照学習と組み合わせる [22] など様々な派生手法が提案されている．

自己教師あり学習と ViT の組み合わせは年々増加傾向にある．この自己教師あり学習と ViT (または Transformer) は，大量かつ広範なデータで学習して様々な下流タスクに転移できる Foundation Model (FM) [20] としても注目されている．FM では，コンピュータビジョン，自然言語処理及び音声等の異なる領域との融合によるマルチモーダル性を高めた研究がされている．そのため，自己教師あり学習と ViT の組み合わせは FM として今後発展すると考えられる．FM については第 5 章で説明する．

4.3 計算コスト削減のための ViT モデル

モデルサイズが大きく膨大なパラメータを有する ViT 構造をそのままコンピュータビジョンの主要なタスクに流用することは，メモリ不足を引き起こし処理不可能になることが懸念された．これは，Self Attention の計算量が入力の系列長の二乗に比例するため，画像のパッチ数を増やすとそれに伴い計算量が膨大になることが原因である．そこで，図 2 右中央に示す Self Attention を効率化する ViT が提案された．代表的手法は Swin Transformer [23] で，shifted window と呼ばれる局所窓内のみで Self Attention を計算する．また，CNN 同様に階層型構造を採用しており，入力層から出力層にかけ特徴マップサイズをダウンサンプリングすることで計算コストを抑制している．同時期に報告された Pyramid Vision Transformer [24] では，キーとバリューを縮小させて特徴マップのスケールを階層毎に縮小する Spatial Reduction Attention（SRA）を導入することで計算コストを抑制している．SRA ではキーとバリュー特徴量を求める際，元の特徴マップサイズと同じになるようにリサイズしてから畳み込み処理を行っている．SRA はセグメンテーションタスク [25] や階層型構造の浅い層 [26] で用いられていたりと汎用性が高い．SRA の畳み込みを Average Pooling に変更したモデル [27] も提案されており，キーとバリューをどう効率的に計算するかが議論されている．

モバイル端末に利用可能な軽量の ViT として MobileViT [28] や MobileFormer [29] がある．これらは共通して MobileNet と ViT を組み合わせた手法である．MobileViT は，mobilenetv2 ブロックを通じて局所特徴を捉えつつ特徴マップをダウンサンプリングし，そのブロック間で提案している MobileViT ブロックで局

所特徴と広域特徴の両方を捉えている．MobileViT ブロックには，入力特徴から局所特徴を捉える local，local で抽出した特徴から広域特徴を捉える Transformer 及び，入力特徴と Transformer で抽出された特徴を連結し point-wise convolution で元の特徴次元サイズまで畳み込む Fusion で構成されている．従って，MobileViT ブロック 1 つで CNN の利点である局所領域に強い帰納バイアスと Transformer の利点である広域特徴を抽出可能で，それぞれの利点を活かしつつ計算コスト削減に寄与している．一方で，CNN の畳み込みとバッチ正規化などデバイスレベルの最適化があるが，Transformer にはそれがないことから計算速度では CNN に劣っていることが報告されている．MobileFormer には MobileNet と ViT それぞれに触発された Mobile と Former，これらのモジュール間を相互に計算する 2 つの Attention 機構の合計 4 つのモジュールがある．Mobile は画像の局所特徴，Former は Transformer Block の Self Attention と Feed Forward，Mobile から Former の Attention は Cross Attention を通じて局所特徴と広域特徴を組み合わせ，最後に Former から Mobile の Attention は広域特徴から局所特徴のパスとして利用される．MobileFormer では Attention 機構の計算量増加を抑制するために，Mobile から Former の Attention はクエリのみ計算，Former から Mobile はキーとバリューのみを計算している．MobileFormer は MobileNetv3 と比較しても同程度の速度で画像が大きければ MobileFormer の方が性能が良いことが報告されている．一方，MobileViT 同様にデバイスレベルでは CNN の方が効率よく計算されるため，低解像度では MobileNetv3 が高速と報告されている．そのため，Vision Transformer に最適化されたデバイスによる高速化が期待される．[2]

最後に，ViT の隠れ層 (Transformer block) の次元数は Base で 764，Large で 1,024，Huge で 1,280 と膨大でニューロンは密に繋がっている．文献 [30] では，事前学習された ViT は不要なニューロンが多くほとんどの層で 0 以上の値を持つニューロンが 10%未満であることが判明している．従って，ViT はネットワークの枝刈りを行うことで，計算回数が削減されメモリ使用量が少なくなる．例えば，DynamicViT [31] や A-ViT [32] は事前学習済みモデルを利用し層毎に不必要なパッチトークンを削除することで計算コストを抑制している．また，ViT Pruning [33] ではパッチトークンを削除するのでなく，次元方向で不要なニューロンを削除している．深層学習モデルを実機等に搭載する上でモデルサイズ削減は有効な手段であるため，枝刈りに限らず量子化や蒸留と組み合わせた ViT モデルの発展に期待できる．

[2]NVIDIA 社の H100 Transformer Engine で処理するとどこまで性能・速度が出るか興味深い．

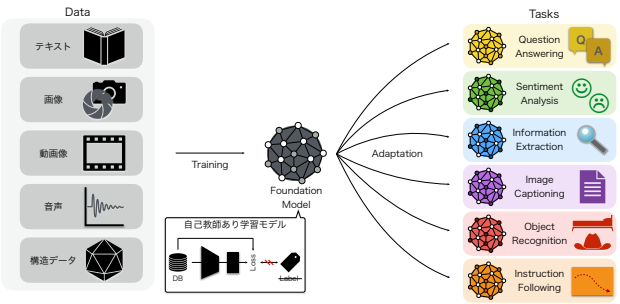

図 3　FM の概略図．文献 [20] より引用及び改変．

5　大規模モデルに関する今後の動向予想

第 3 章でも述べたように，Transformer モデルにおける認識性能の向上には大規模なモデル構築や膨大なデータセットに加えて膨大な学習ステップ数の 3 つが必要であることが知られている [34]．この知見を基に，2021 年にスタンフォード大学の研究者らは大量のテキストや画像等の広範なデータで学習することで質疑応答や感情分析等の様々な下流タスクに転移できるモデルを Foundation Model (FM) [20] と命名した．FM の概略図を図 3 に示す．FM は，ラベルのない大量のデータを用意して自己教師あり学習することで大規模かつ汎用型人工知能モデル構築を目指している．コンピュータビジョン分野における FM は，画像を入力とする場合 [18, 21, 19, 35] 及び，画像と言語 [37, 38, 39]，画像と音声 [40, 41] 等のマルチモーダルデータを入力とする場合の 2 種類に大別できる．本章ではこれらについてまとめつつ，今後の動向を筆者なりに予想する．

5.1　画像データを用いた FM

画像データを用いた FM は自然言語処理の Masked Language Modeling に触発された Masked Image Modeling (MIM) が代表的方法である．MIM は，入力画像のパッチをランダムにマスクしたトークンから元画像を復元できるように学習する自己教師あり学習手法である．MIM では，ViT から出力される各パッチトークンと入力パッチ間で損失計算することで元の入力画像に復元できる．損失計算は，マスクトークンと入力画像のマスクされた真値との平均二乗誤差をピクセルレベルで計算する方法と，パッチ毎に自然言語の語彙に相当する意味のあるトークンに置換し分類問題として扱う方法の 2 種類がある．前者は Masked Autoencoder (MAE) [18] や Multi-MAE [21]，後者は BEiT [19] や BEiT v2 [35] があり，前者について MIM を畳み込み処理で実現した ConvNeXt v2 [36] も提案されている．どちらが優れているか結論は出ていないが，セグメンテーションや物体検出等のコンピュータビジョンの主要なタスクで ImageNet 事前学習済みの教師あり学習

図 4　プロンプトテンプレートの例.

モデルより高性能なことが判明している.

5.2　マルチモーダルデータを用いた FM

本節ではマルチモーダルデータを入力する FM の中で，数多く提案されている画像と言語の方法にフォーカスする．画像と言語を入力とする FM による成功例の一つに Zero-shot でクラス分類した点である．Zero-shot クラス分類は，学習データにないクラスをモデルが正確に分類できるようにするタスクである．この Zero-shot クラス分類で重要となるのがプロンプトテンプレートである．プロンプトテンプレートは質問や指示等を表すもので，例えば図 4 のようにベッドの “位置はどれ?”や “画像は何を説明している？”というプロンプトテンプレートからそれに応じた出力が可能になる．Zero-shot クラス分類では，プロンプトテンプレートを用いることで認識精度が僅かながら向上する．これは Web から収集した画像と言語のペア情報は文章なことが多く，ラベルを直接予測するのでなくプロンプトと紐づけることで画像と言語のペア情報を自然な形で予測できるためと考えられる．代表的手法の CLIP [37] では，画像とプロンプトテンプレート “A photo of a クラス名”の特徴量の類似度をクラススコアとして用いる．CLIP は，画像と言語のペア情報を最大化し，それ以外を最小化するように学習することで Zero-shot クラス分類を実現している．CoCa [38] は 2023 年 2 月時点で ImageNet ベンチマークで state-of-the-art モデルで，画像と言語の特徴量間で Cross Attention することで画像と言語表現を学習している．画像と言語のマルチモーダルデータは画像分類，物体検出，セグメンテーション及びキャプション生成等のコンピュータビジョンの主要なタスクで非常に効果的である．加えて，画像生成可能で様々なタスクを実行できる汎用的モデルの Unified-IO [42] 等，画像と言語を入力とする FM の発展は目を見張るものがある.

5.3　今後の動向予想

FM はその汎用性の高さから今後確実に発展する技術であり，FM を用いたサービスが提供されると考えられる．実際に，Chat-GPT や Stable Diffusion といった技術が公開され全世界の人々に使用されている．一方で，高性能な FM は膨大なパラメータを有するため計算資源を必要としたり，使用データの都合上．著作権や肖像権の違反や個人情報漏洩の危険性，各国の法律に触れる恐れがあり安易にモデルやデータを公開できない問題が挙げられる．後者は，一部企業がモデルやデータの独占により研究の寡占化が浸透し，コンピュータビジョンの発展を妨げる恐れもある．これらは難しい問題ではあるが，AI を社会実装する上で必要不可欠な要素であり我々は問題に対処し続けなければならない.

このような問題に対し，著者は大規模モデルと同程度な精度を発揮する小規模モデルの構築とオープンでクリーンな研究がされると予想する．前者は 4.3 節で述べたように，大規模モデルから小規模モデルに知識転移する蒸留やモデルの不要な重みやニューロンを削除する枝刈りが発展すると予想する．後者は一般公開できるように研究機関等が各国の法律に則ったり，セキュリティ強化等をする必要があると考えられる．特に後者は，一般公開することで悪意ある攻撃，例えば敵対的攻撃を受ける事が想定されるため，敵対的攻撃から守る FM のための防御手法のような研究が発展していくことを予想している.

6　おわりに

本稿では，初学者向けに ViT の基礎的構造について概説しつつ，コンピュータビジョンにおける応用例及び，今後の動向予想を説明した．ViT はパッチ間の対応関係を取得する Self Attention がキーで，この機構により浅い層から画像の局所領域を見つつ広域領域を捉えることが可能になった．一方で，ViT の高精度化には膨大な学習データが必要な点，リッチな計算資源が必要な点といった ViT 特有の制限をあげた．そこで，ViT に適したデータ拡張による高精度化や計算コストを削減した ViT 等，ViT の制限の観点から応用例を紹介した．さらに，大量かつ広範なデータで学習して下流タスクに転移可能で主に Transformer を用いた大規模モデルの FM が，今後のコンピュータビジョンのブレイクスルーになる可能性を秘めていることを述べた．FM は今後確実に発展する技術である一方，ViT 同様にリッチな計算資源を有する必要があること，著作権や肖像権を違反する恐れがあるため安易にモデルやデータを公開できず研究の寡占化が浸透しコンピュータビジョンの発展が困難，といった問題が挙げられる．そのため，今後の ViT 含め大規模モデルは，大規模モデルと同程度の精度を発揮する小規模モデルの構築とオープンでクリーンな研究がされると予想する．本稿が，ViT の制限の観点から効率的なモデルの選定及び，ViT における学習テクニック，今後 ViT に関する研究をする上での参考になれば幸いである.

参考文献

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

[2] 片岡ら,Vision Transformer 入門, 技術評論社 ISBN 978-4-297-13058-9, 2022.

[3] H.Touvron *et al.*, Training data-efficient image transformers & distillation through attention, ICML, 2021.

[4] E.D.Cubuk *et al.*, AutoAugment: Learning Augmentation Policies from Data, CVPR, 2019.

[5] E.D.Cubuk *et al.*, RandAugment: Practical automated data augmentation with a reduced search space, CVPR Workshops, 2020.

[6] H.Zhang *et al.*, mixup: Beyond Empirical Risk Minimization, ICLR, 2018.

[7] S.Yun *et al.*, CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features, ICCV, 2019.

[8] Z.Jiang *et al.*, All Tokens Matter: Token Labeling for Training Better Vision Transformers, NeurIPS, 2021.

[9] J.Liu *et al.*, TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers, ECCV, 2022.

[10] J.N.Chen *et al.*, TransMix: Attend to Mix for Vision Transformers, CVPR, 2022.

[11] Q.Zhao *et al.*, MixPro: Data Augmentation with MaskMix and Progressive Attention Labeling for Vision Transformer, ICLR, 2023.

[12] S.Tuli *et al.*, Are Convolutional Neural Networks or Transformers more like human vision?, CogSci, 2021.

[13] H.Touvron *et al.*, DeiT III: Revenge of the ViT, arXiv, 2022.

[14] H.Kataoka *et al.*, Replacing Labeled Real-Image Datasets with Auto-Generated Contours, CVPR, 2022.

[15] M.Caron *et al.*, Emerging Properties in Self-Supervised Vision Transformers, ICCV, 2021.

[16] X.Chen *et al.*, An Empirical Study of Training Self-Supervised Vision Transformers, ICCV, 2021.

[17] S.Atito *et al.*, Self-supervised vIsion Transformer, arXiv, 2021.

[18] K.He *et al.*, Masked Autoencoders Are Scalable Vision Learners, CVPR, 2022.

[19] H.Bao *et al.*, BEiT: BERT Pre-Training of Image Transformers, ICLR, 2022.

[20] R.Bommasani *et al.*, On the Opportunities and Risks of Foundation Models, arXiv, 2021.

[21] R.Bachmann *et al.*, MultiMAE: Multi-modal Multi-task Masked Autoencoders, ECCV, 2022.

[22] Z.Huang *et al.*, Contrastive masked autoencoders are stronger vision learners, arXiv, 2022.

[23] Z.Liu *et al.*, Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, ICCV, 2021.

[24] W.Wang *et al.*, Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions, ICCV, 2021.

[25] E.Xie *et al.*, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, NeurIPS, 2021.

[26] Z.Xia *et al.*, Vision Transformer with Deformable Attention, CVPR, 2022.

[27] W.Wang *et al.*, PVT v2: Improved Baselines with Pyramid Vision Transformer, CVMJ, 2022.

[28] S.Mehta *et al.*, General-purpose, and Mobile-friendly Vision Transformer, ICLR, 2022.

[29] Y.Chen *et al.*, Mobile-Former: Bridging MobileNet and Transformer, CVPR, 2022.

[30] X.Chen *et al.*, When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations, ICLR, 2022.

[31] Y.Rao *et al.*, DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification, NeurIPS, 2021.

[32] H.Yin *et al.*, A-ViT: Adaptive Tokens for Efficient Vision Transformer, CVPR, 2022.

[33] M.Zhu *et al.*, Vision Transformer Pruning, KDD, 2021.

[34] J.Kaplan *et al.*, Language models are few-shot learners, NeurIPS, 2020.

[35] Z.Peng *et al.*, BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers, arXiv, 2022.

[36] S.Woo *et al.*, ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders, CVPR, 2023.

[37] A.Radford *et al.*, Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021.

[38] J.Yu *et al.*, CoCa: Contrastive Captioners are Image-Text Foundation Models, TMLR, 2022.

[39] T.Gupta *et al.*, Towards General Purpose Vision Systems, CVPR, 2022.

[40] A.Jaegle *et al.*, Perceiver: General Perception

with Iterative Attention, ICML, 2021.

[41] V.Likhosherstov *et al.*, PolyViT: Co-training Vision Transformers on Images, Videos and Audio, arXiv, 2021.

[42] J.Lu *et al.*, UNIFIED–IO: A Unified Model for Vision, Language, and Multi-modal Tasks, ICLR, 2023.