ChatGPT

AI業界の歴史的転換点: 画像認識AIはCNNからTransformerへ

背景: CNNが支配した10年と新たな潮流

過去10年ほど、**画像認識分野では畳み込みニューラルネットワーク（CNN）が事実上の標準**となり、画像分類や物体検出など様々なタスクで最高精度を叩き出してきました [1] [2]。2012年にAlexNet（CNNモデル）がImageNetコンペティションで他を圧倒したことを契機に、画像認識と言えばまずCNNを使うのが常識となったのです [2]。研究者やエンジニアは「どの手法を使うべきか？」ではなく**「どのCNNアーキテクチャを使うべきか？」**に注目してきたほどで、CNN一強時代が長く続きました。

しかし近年、この構図が大きく揺らぎ始めています。**自然言語処理（NLP）で革命を起こしたTransformerが、画像の世界にも本格的に進出してきたのです**。TransformerベースのVision Transformer（ViT）をはじめとするモデル群が2020年頃から登場し、画像認識ベンチマークでCNNに匹敵し時には凌駕する性能を示し始めました [3]。画像分類に限らず、物体検出やセマンティックセグメンテーション、画像生成、動画解析など**幅広い視覚タスクでTransformerモデルが次々と最先端の記録を更新**しています [4]。まさに**「業界標準が根底から覆った」パラダイムシフト**が進行しているのです。

CNNとTransformerの違い: 「木を見て森を見ず」 vs 「森全体を俯瞰」

なぜこれほど劇的な主役交代が起きたのでしょうか？理由の一つは、**画像の捉え方（視点）の根本的な違い**にあります。

- **従来型のCNN: 「木を見て森を見ず」** – CNNは画像を小さな局所領域（フィルタの受容野）に分割し、その**局所特徴（エッジやテクスチャなど）を積み上げていく**階層構造を持ちます [5] [6]。一度に見る範囲が限られているため、遠く離れた部分同士の関係性を直接捉えることが苦手です。実際、CNNは**狭い範囲のパターンに強く反応するバイアス（帰納的バイアス）**を持ち、局所的に強い手がかりがあると全体の文脈よりそちらを優先して判断してしまう傾向があります [7]。

- **新世代のTransformer: 「森全体を俯瞰する」** – Transformerは**自己注意機構（Self-Attention）**を用いて**画像内のあらゆる位置の特徴同士の関係を一度に計算し考慮する**ことができます [8]。つまり**最初の層からグローバルな文脈を捉える**ことが可能なのです。全体像を一度に眺めて各要素間の関連性を評価できるため、**画像全域にまたがるパターン認識**や**離れた部分にまたがる特徴の統合**が得意です [8]。この違いにより、Transformerは**画像全体の文脈や構造を把握してより的確な認識**ができると期待されています [9]。

この「局所 vs 全体」の違いを端的に示す例として、**「羽の生えた猫」の画像実験**が報告されています。 [10] ある研究者が「カラフルなインコの羽を持つ猫」の画像を生成し、従来のCNNモデル（ResNet）とVision Transformerに分類させたところ、**ResNetはその画像を「オウム（鳥）」と誤認**し、**Vision Transformerは正しく「エジプシャンマウ（猫の一種）」と識別**したといいます [11]。同じ画像を見せても両者が全く異なる予測をしたのは偶然ではなく、「局所テクスチャに着目するCNN」と「全体像を見るTransformer」の**構造上の違いによる必然的な結果**です [11]。羽毛の質感という**局所的な特徴にCNNは強く反応**して「鳥類」と判断しましたが、Transformerは**猫の形状や文脈を含む全体像を考慮**したため「猫」と認識できたのです [7]。このように、CNNは「木」（細部）に集中するあまり「森」（全体像）を見失う場合があるのに対し、Transformerは初めから森全体を見渡して木々の配置を把握できる点で有利だと言えます。

TransformerがCNNを凌駕した理由

Transformerが画像認識で急速に躍進した背景には、上述の**全体文脈把握力**に加えて、**モデルの表現力の高さ**があります。自己注意による柔軟な相関学習により、TransformerはCNNでは表現しきれない複雑なパターンも学習できる可能性があります。ただしその反面、**TransformerはCNNに比べ事前に組み込まれた知識（帰納バイアス）が少ない**ため、**学習に大量のデータを要する**というトレードオフがあります [12] [13]。

- **帰納バイアスと表現力のトレードオフ**: CNNは局所性や平行移動不変といった**画像に関する先験的な前提（バイアス）を持つぶん、少ないデータでも効率よく学習**できます [14]。一方、Transformerは柔軟で強力な表現力を持つ反面、画像の構造に関する先入観がほぼ無いため、**大量のデータを見せて学習させないと真価を発揮できない**のです [14] [15]。実際、Googleが提案した初期のViT論文では、ImageNet程度のデータ（数百万画像）ではCNNより精度が劣りましたが、より大規模なデータセット（数億画像規模）で事前学習することで**CNNの持つ人為的バイアスを凌駕する性能**を引き出せることが示されました [16]。例えば**約3億件の画像で事前学習したViTは、ImageNet分類でトップ1精度88.36%を達成し、従来の最先端CNNモデルを上回る結果**を出しています [17]。

- **各タスクへの適応**: Transformerが性能を発揮するまでには各タスクに合わせた工夫も必要でした。初期のViTをそのまま物体検出に応用した場合、**CNNベースの検出器より良い結果を出せないケースも報告されています** [18]。しかし、Transformerの長所を活かす専用アーキテクチャが開発され、現在では**物体検出でもTransformerベースの手法がCNNの精度記録を塗り替える**ようになりました。例えば、Transformerを用いた物体検出モデルの先駆けである**DETR**は、従来のFaster R-CNN＋ResNetといった手法を追い抜く検出精度を示し、特に改良版のDeformable DETRでは学習効率も大きく改善されています [19] [20]。また**Swin Transformer**のように階層的特徴を組み込んだVision Transformerは、同程度のパラメータ規模のResNeXt（高性能なCNN）と比較してCOCO物体検出で平均適合率(AP)を約4ポイント改善し、画像セグメンテーションでも3ポイント以上改善するという報告があります [21]。このように各分野でTransformerモデルが**従来のCNNベースモデルの記録を次々と更新**しているのです [22]。

以上の理由から、大規模データと計算資源さえ投入できれば**「Transformerで従来以上の精度を出せる」ことが明確になった**ため、研究コミュニティおよび産業界は急速にTransformerへのシフトを進めています [22]。特に画像分類では純粋なTransformerモデルが既に多数提案され、画像認識コンペの上位を席巻し始めています。物体検出やセグメンテーション、動画解析においても、Transformerベースの新手法が毎年のように登場しては性能の記録を更新している状況です。

ビジネス・応用への示唆: 精度向上とマルチモーダル、そして課題

単なる研究上のブレークスルーに留まらず、**このCNNからTransformerへの交代劇は実務やビジネスへのいくつかの重要な示唆**を与えています。

- ☒ **精度向上の新たな余地**: 既存の画像認識システムをCNNからTransformerベースに置き換えるだけで、**認識精度が一段階向上する可能性**があります。実際、前述の通りImageNet分類やCOCO検出といった主要ベンチマークでTransformerモデルが最高性能を更新しています。例えばVision Transformer系列はImageNet精度で従来比数ポイントの改善を示し [17]、Swin Transformerも物体検出・セグメンテーションでCNNバックボーンを用いた場合よりAPを数ポイント引き上げています [21]。精度向上は直訳すればサービスやプロダクトの品質向上に繋がるため、**競合優位性を得るためにもTransformerへの刷新は検討すべき**でしょう。

- ☒ **マルチモーダルAIの加速**: **Transformerはテキスト・画像・音声など異なる種類のデータを共通のアーキテクチャで扱える**ため、視覚と言語を統合したマルチモーダルAIの開発が飛躍的に進みつつあります。実際、OpenAIの**CLIP**のように**画像とテキストを同じモデルで結び付け、テキストから対応す**

る画像を検索・分類できるシステムが登場しています [23]（例えば「犬」と文章で指示すると犬の画像を見つけ出すゼロショット画像分類が可能）。また**画像を入力して文章で説明文を生成する画像キャプショニング**や、逆に**文章から関連する動画シーンを検索する**ようなタスクにもTransformerが活用されています [24]。Transformerのおかげで**一つのモデルが異なるモダリティ間の対応関係を学習できる**ため、**「見る」と「読む」を組み合わせた新しいAI応用**（例: 視覚質問応答VQA、画像からの自動レポート作成など）が加速しています [25]。テキストも画像も同じ枠組みで処理できる強みは、今後のマルチモーダルAI分野の発展を強力に後押しするでしょう。

- 💻 **コストと効率の課題**: 注意すべき点として、Transformerモデルは**圧倒的なデータ量と計算資源（GPU時間）を要求する大食漢**でもあります [26] [20]。例えば基本的なVision Transformer (ViT-B)で1枚の画像を処理するには約180億回の演算(FLOPs)が必要と試算されており、**同等精度の軽量CNN（GhostNet）と比較して30倍以上も計算量が多い**という報告があります [27]。また、ある研究ではCOCO物体検出のモデル学習に、**CNNベースのFaster R-CNNが380 GPU時間で済んだのに対し、TransformerベースのDETRは最適化前は2000 GPU時間を要した**とされています [20]（改良版のDeformable DETRでは約325 GPU時間まで短縮）。さらにデータ量の面でも、Transformerは**数百万枚規模では充分に汎化性能を発揮できず、数千万〜数億枚規模の学習データが必要**との知見があります [26]。このため、**モデルの精度向上と計算コスト・開発コストとのバランスを慎重に検討する必要**があります。しかし幸いなことに、近年は効率的なTransformerアーキテクチャ（例えば局所的注意機構を導入したSwin Transformerなど）や蒸留・圧縮手法の研究が進み、**実用面でのハードルは徐々に下がりつつある**のも事実です [20] [27]。

以上を踏まえると、**画像AIの世界は新たな章に突入した**と言えます。現在のところ大規模データを持つ一部のプレイヤー（大企業や研究機関）がTransformerの恩恵を最大限に享受していますが、今後はより効率化されたモデルや事前学習済みモデルの公開によって、広範な分野でTransformerが使われるのは確実でしょう。**この流れはほぼ不可逆的**であり、画像認識のスタンダードがCNNからTransformerへ移行しつつあることは、多くの専門家が認めるところです [22]。もっとも、データが限られた状況では依然としてCNNが有利な場合もあり、両者は今後も併存して使い分けられるでしょう [28]。しかし「**AIの目**」とも呼ぶべきコンピュータビジョン分野において、**Transformerという新たな覇者が台頭した事実**は揺るがず、これからの研究開発の主役であり続けると考えられます。

参考文献: Transformerを用いた画像処理の包括的な動向についてはHuawei Noah's Ark Labによる詳細なサーベイ論文 [3] や、実応用でのCNNとViTの比較分析を行ったブログ記事 [22] [29] などが詳しいです。さらに、Vision Transformerの特性を分析した研究では、CNNと比較して**遮蔽やノイズに対する頑健性が高い**ことなども報告されています [22]。これらの資料も併せて参照すると、今回のパラダイムシフトの全貌をより深く理解できるでしょう。

---

[1] [3] [4] [8] [12] [14] [16] [17] [18] [25] [27] [2012.12556] A Survey on Visual Transformer
https://ar5iv.labs.arxiv.org/html/2012.12556

[2] Vision Transformers or CNNs? Who Wins the Vision Race
https://blog.scribeofai.com/the-future-of-computer-vision-vision-transformers-or-cnns/

[5] [6] [7] [10] [11] [13] When a Cat Becomes a Macaw: Why Vision Transformers Beat CNNs at Their Own Game | by Omer | Dec, 2025 | Medium
https://medium.com/@omer389/when-a-cat-becomes-a-macaw-why-vision-transformers-beat-cnns-at-their-own-game-dc6e14236d26

[9] [15] [19] [20] [21] [22] [26] [28] [29] Are Transformers replacing CNNs in Object Detection? — Picsellia
https://www.picsellia.com/post/are-transformers-replacing-cnns-in-object-detection

23 24 What are multimodal transformers and how do they work?
https://milvus.io/ai-quick-reference/what-are-multimodal-transformers-and-how-do-they-work