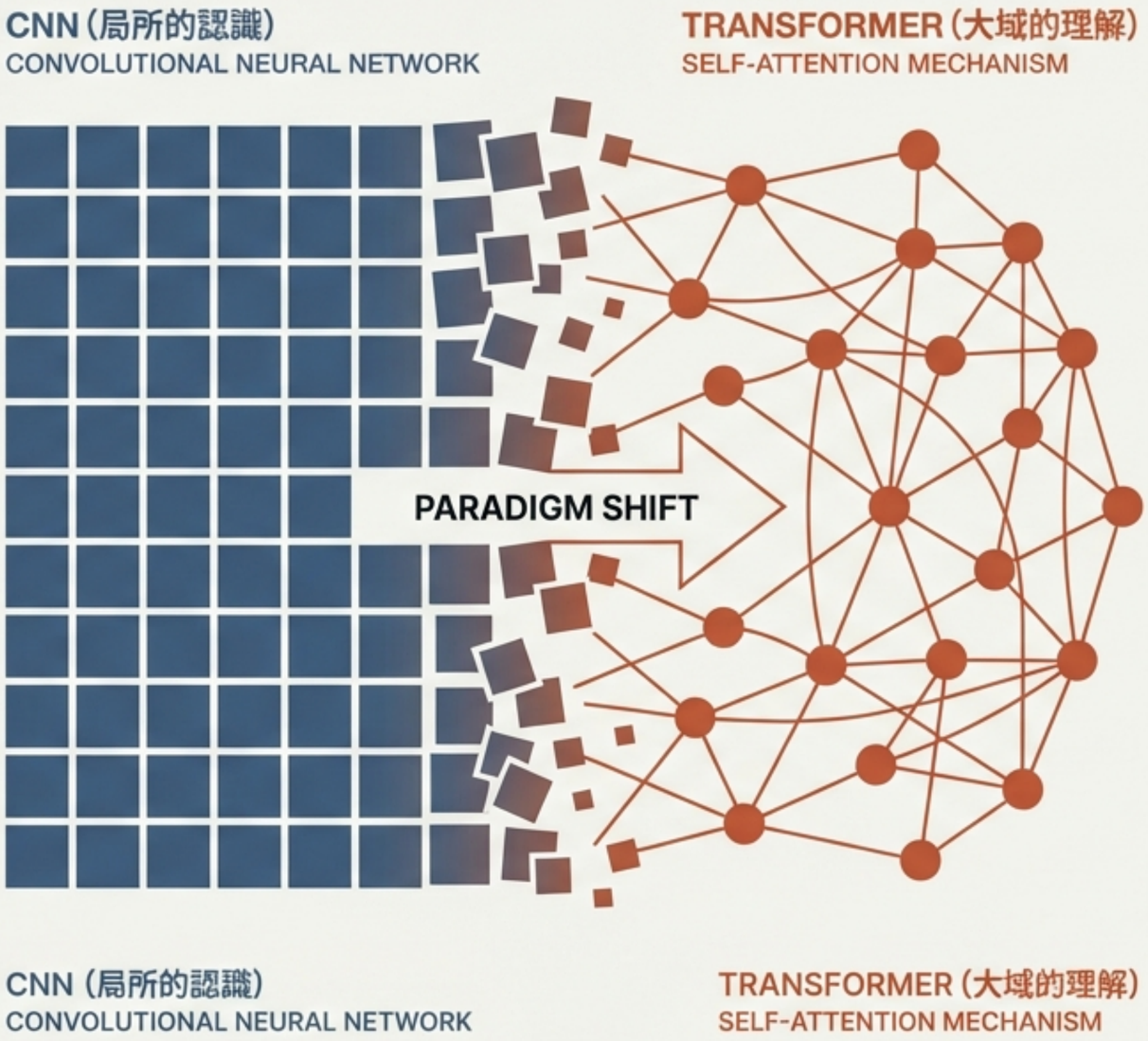
AI TREND REPORT 2024

画像認識AIのパラダイムシフト

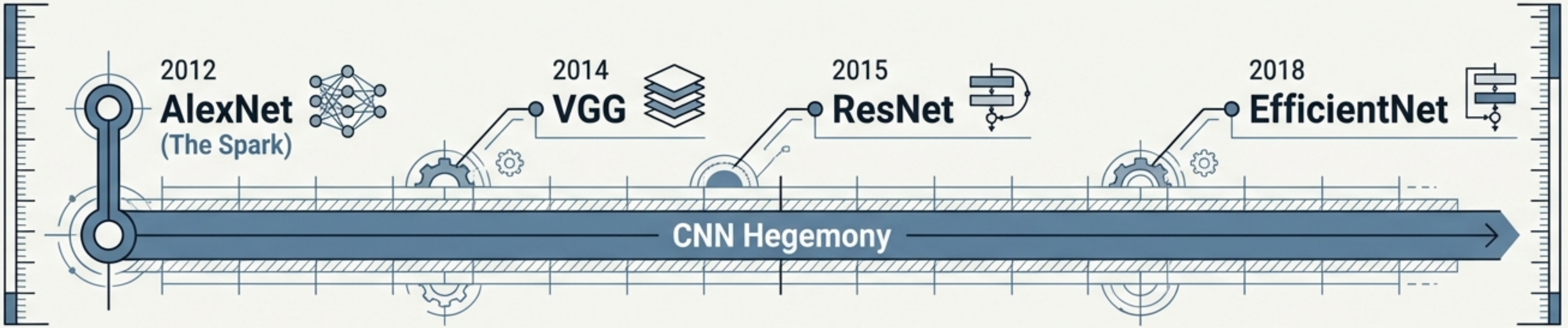
CNNの時代から、Transformerが支配する「全体を俯瞰する」AIへ

過去10年間の「常識」が覆されようとしています。自然言語処理の革命児であるTransformerが、画像認識の世界地図を塗り替え始めました。なぜ今、主役交代が起きているのか。その技術的本質とビジネスへのインパクトを紐解きます。



疑う余地のない常識：CNNが一強だった10年間

2012年のAlexNet登場以来、画像認識といえばCNN（畳み込みニューラルネットワーク）が絶対的な「業界標準」でした。



きっかけ

2012年、AlexNetがImageNetコンペティションで他を圧倒。これを契機に「画像認識=CNN」という図式が確立されました。

研究者のマインドセット

関心は「どの手法を使うか」ではなく、「どのCNNアーキテクチャ（ResNet, VGGなど）を使うか」に集中していました。

実績

CNNは長年にわたり、画像分類や物体検出で最高精度（SOTA）を独占し続けました。

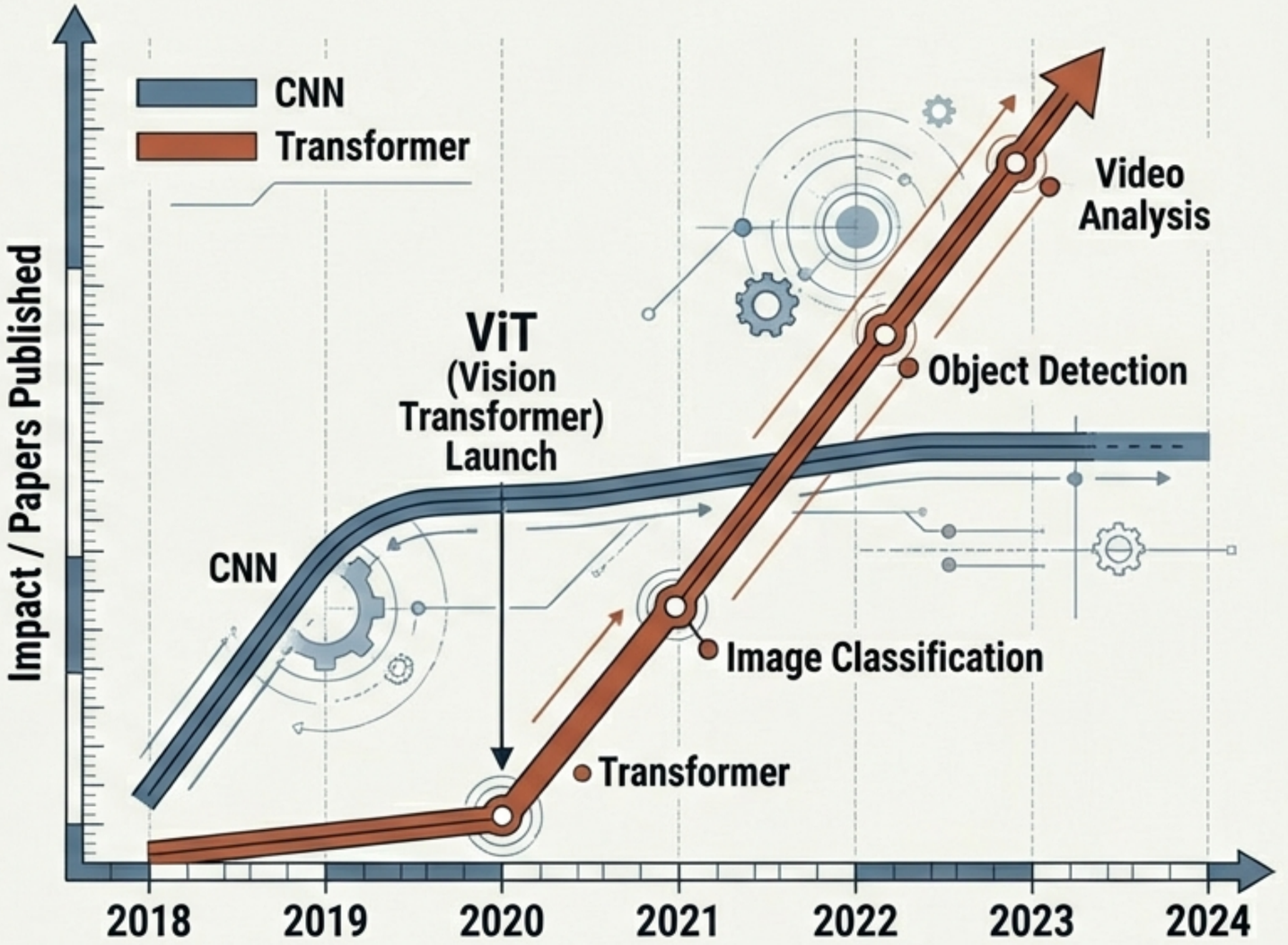
黒船来航：言語処理界からの侵略者、Transformer

自然言語処理（NLP）で革命を起こした技術が、2020年頃から画像の世界へ本格進出し、CNNの牙城を崩し始めました。

Vision Transformer (ViT) の登場により、画像認識ベンチマークでCNNに匹敵、あるいは凌駕する性能が示され始めました。

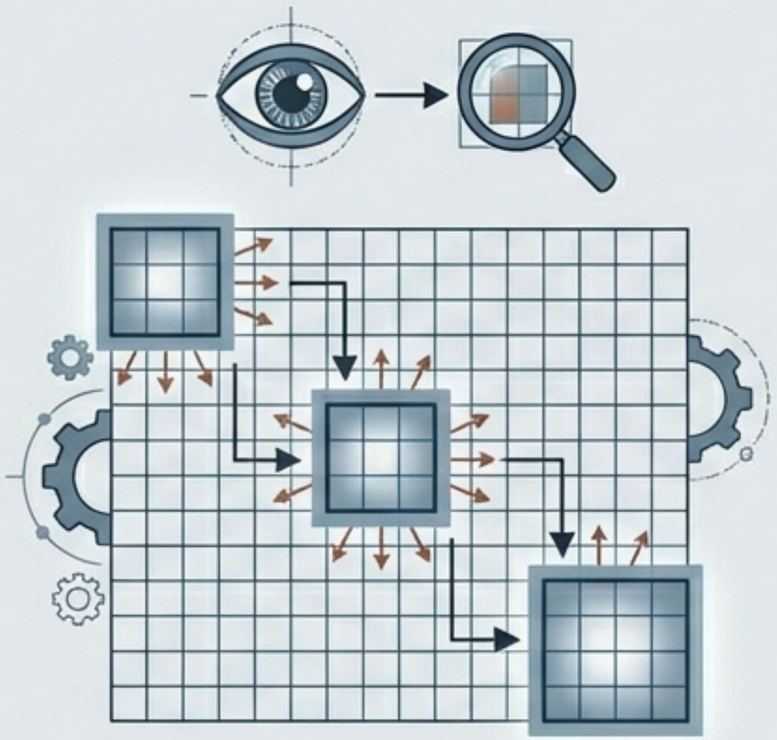
影響は画像分類だけにとどまりません。物体検出、セマンティックセグメンテーション、画像生成、動画解析に至るまで、あらゆる視覚タスクまで、あらゆる視覚タスクで記録を更新しています。

これは単なる技術流行ではなく、業界標準が根底から覆る歴史的転換点です。



最大の違いは「視点」にある：局所処理 vs 全体俯瞰

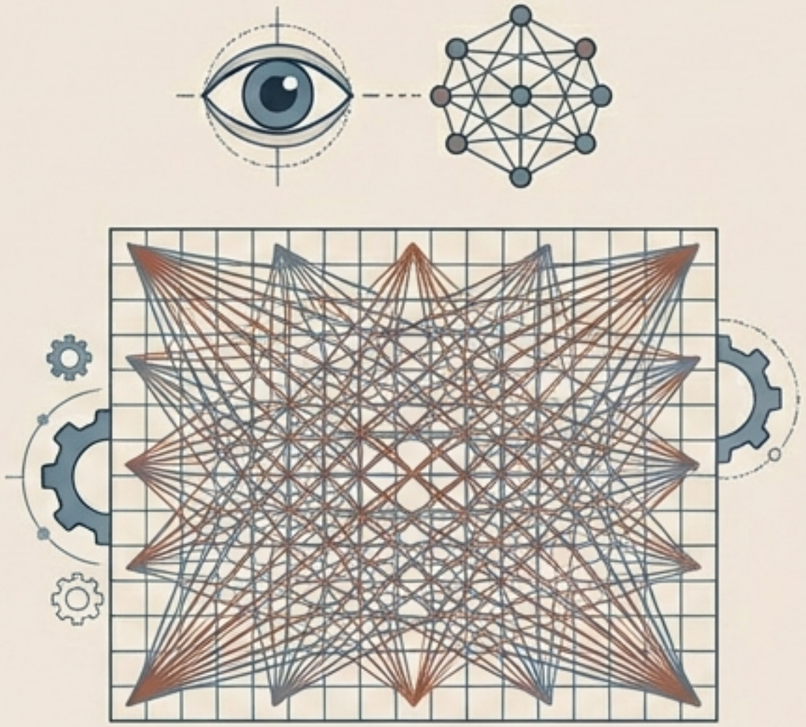
CNN（従来型）



「木を見て森を見ず」

画像を小さな局所領域（フィルタ受容野）に分割し、特徴を積み上げる。一度に見る範囲が限られ、遠く離れた部分同士の関係性を捉えるのが苦手。

Transformer（新世代）



「森全体を俯瞰する」

自己注意機構（Self-Attention）により、画像内のあらゆる位置の特徴同士の関係を一度に計算。最初初の層からグローバルな文脈を捉える。

ケーススタディ：「羽の生えた猫」の実験

CNNはテクスチャ（質感）に騙され、Transformerは形状と文脈を理解しました。

CNN（ResNet）の判定

❌ **「オウム（鳥）」と誤認**

羽毛のテクスチャという局所的な特徴に強く反応し、全体の文脈を無視したため。

Transformer（ViT）の判定

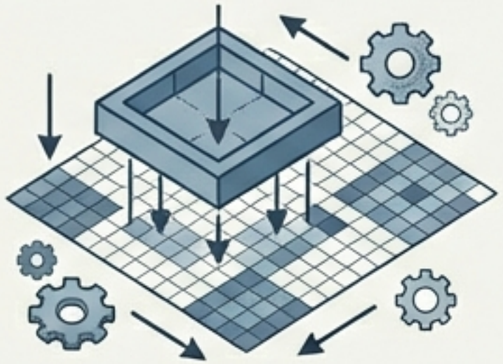
✅ **「エジプシャンマウ（猫）」と正解**

形状や文脈を含む全体像を考慮し、「羽はあるが、これは猫である」と判断できたため。

これは偶然ではなく、構造上の違い（局所 vs 全体）による必然的な結果です。

圧倒的な表現力の代償：帰納的バイアスとデータ量

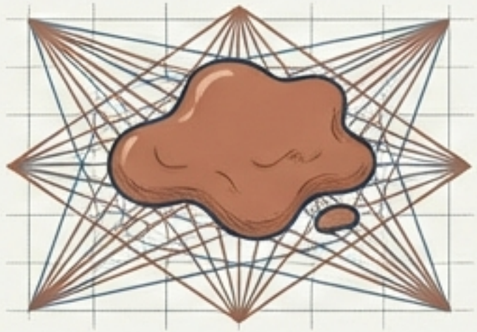
CNN（High Bias）



「画像とは局所的で、平行移動しても同じものでのである」という人間が与えた前提（バイアス）を持つ。

☑ メリット：少ないデータでも効率よく学習可能。

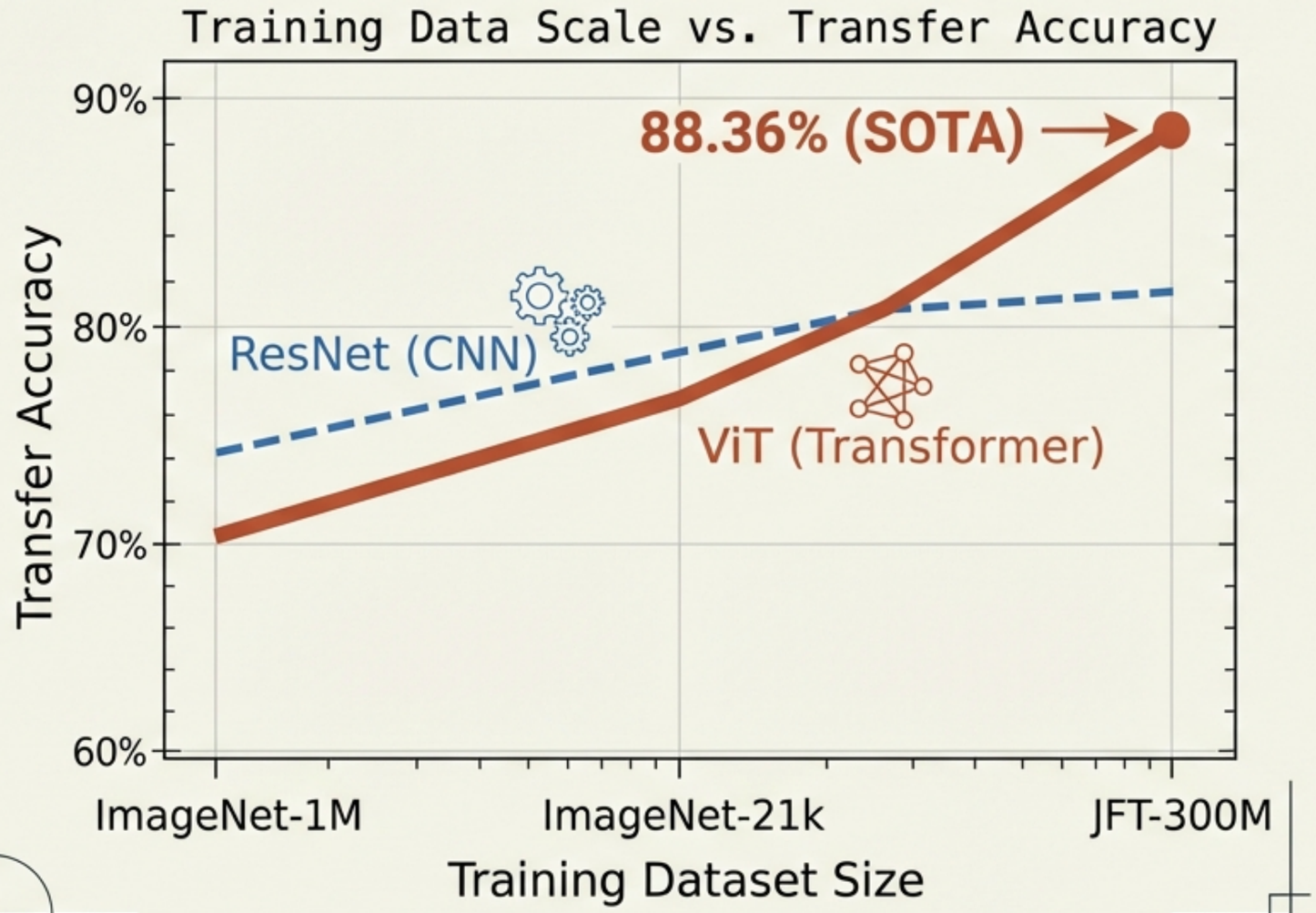
Transformer（Low Bias）



画像構造に関する先入観がほぼ無い。柔軟な相関学習が可能。

✕ デメリット：真価を発揮するには、大量のデータでゼロから関係性を学ばせる必要がある。

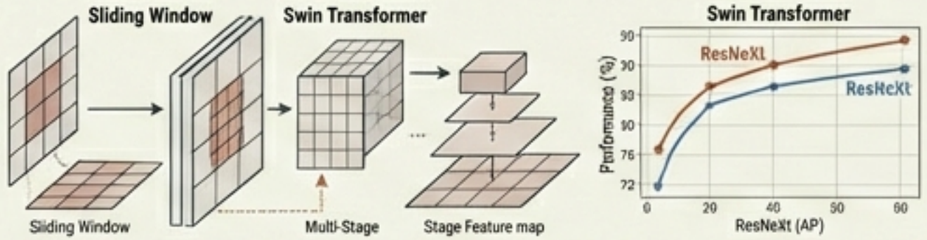
勝利の方程式：大規模データ × 事前学習

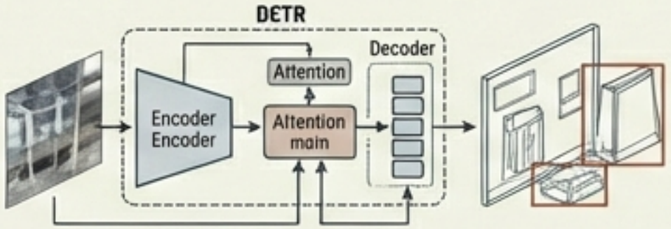


データ量が臨界点を超えた時、Transformerは無敵になります。

- **Small Scale:** 数百万枚レベルではCNNの方が高精度。Transformerは汎化しきれない。
- **Large Scale:** 3億枚（JFT-300M）で事前学習を行うと、ViTはCNNを圧倒的に凌駕する。
- **結論:** 大規模データと計算資源があれば、Transformerは従来以上の精度を出せることが証明されました。

弱点の克服と進化：物体検出への適応



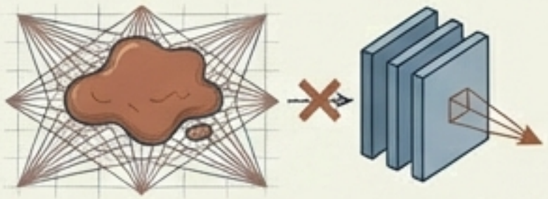


Phase 3: Swin Transformer

階層的特徴を組み込み、ResNeXtと比較して平均適合率(AP)を約4ポイント改善。

Phase 2: DETRの登場

専用アーキテクチャ「DETR」が登場。
従来のFaster R-CNN等を凌駕。



Phase 1: 苦戦

初期のViTは、そのまま物体検出に使ってもCNNに勝てなかった。

- [x] 現在、コンペの上位は既にTransformerベースの手法が席巻しています。

Swiss Editorial

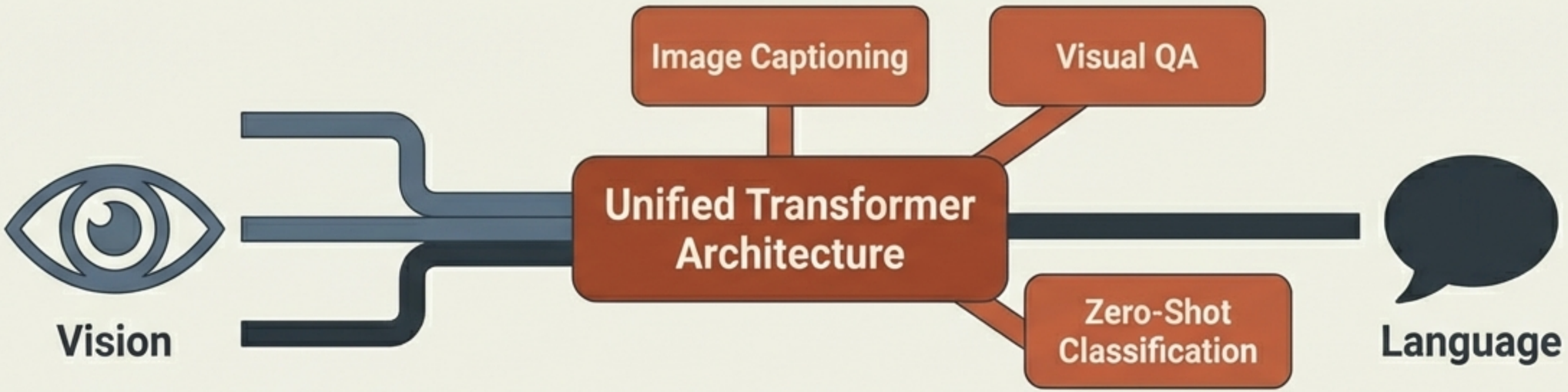
ビジネスへの示唆①：精度の純粋な向上

モデルを置き換えるだけで得られる「競合優位性」があります。

- 認識精度の向上は、直訳すればサービスやプロダクトの品質向上です。
- ImageNet分類やCOCO物体検出におけるSOTA（最高性能）更新は、Transformerによって成し遂げられています。

競合他社に対する優位性を維持するためには、既存のCNNベースのシステムからTransformerへの刷新を検討すべき段階にあります。

ビジネスへの示唆②：マルチモーダルAIの加速



- 「見る」機能と「読む」機能の壁が消滅しつつあります。
- Transformerはテキスト、画像、音声を共通のアーキテクチャで扱えるため、これらを統合した開発が容易になります。
- **OpenAI CLIP**: 「犬」という文字で犬の画像を検索するゼロショット分類が可能。
- **Future**: 異なるモダリティ（感覚）を行き来する新しいAIアプリケーションの爆発的な普及を後押しします。

実装における課題：コストと効率

Performance (Accuracy)

Higher accuracy often requires more complex models.

Compute Cost (FLOPs/GPU)

Increased computational demands and time.

Calculation Volume

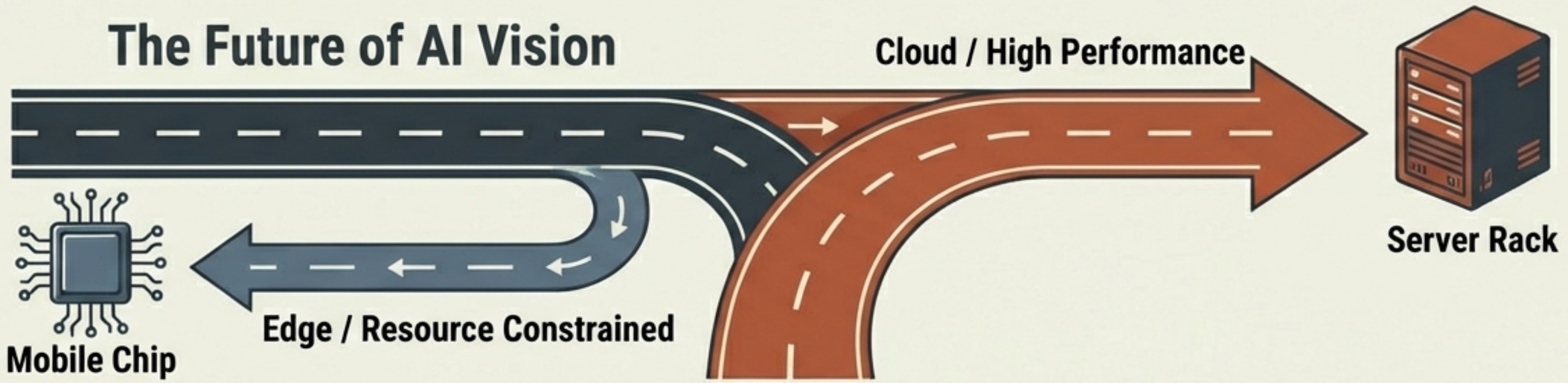
ViT-B = **18 billion** FLOPs/image (**30x** more than GhostNet CNN).

Training Time

DETR (Early) = **2000** GPU Hours
vs
Faster R-CNN = **380** GPU Hours.

Swin Transformerのような効率化モデルや、蒸留・圧縮技術の研究が進んでおり、ハードルは下がりつつありますが、導入時にはROI（投資対効果）の慎重な計算が必要です。

今後の展望：不可逆的な流れと共存



1. 画像AIの世界は新たな章（Transformer時代）に突入しました。この流れは不可逆的です。
2. 大規模データを持つプレイヤーが最大の恩恵を受けますが、公開モデルによって恩恵は民主化されつつあります。
3. 共存（Coexistence): データ量が限られる、あるいは極端な省電力が必要なエッジデバイスなどでは、依然としてCNNが有利な場合もあり、当面は適材適所での使い分けが続きます。

しかし、「AIの目」の進化の中心地は、間違いなくTransformerに移っています。

参考文献・Appendix

- [Survey] A Survey on Visual Transformer (Huawei Noah’s Ark Lab) - **包括的な動向サーベイ**
- [Blog] Vision Transformers or CNNs? Who Wins the Vision Race - **実応用での比較分析**
- [Case Study] When a Cat Becomes a Macaw (Medium) - **「羽の生えた猫」の実験詳細**
- [Detection] Are Transformers replacing CNNs in Object Detection? - **物体検出における比較**
- [Multimodal] What are multimodal transformers and how do they work? - **マルチモーダル技術の解説**