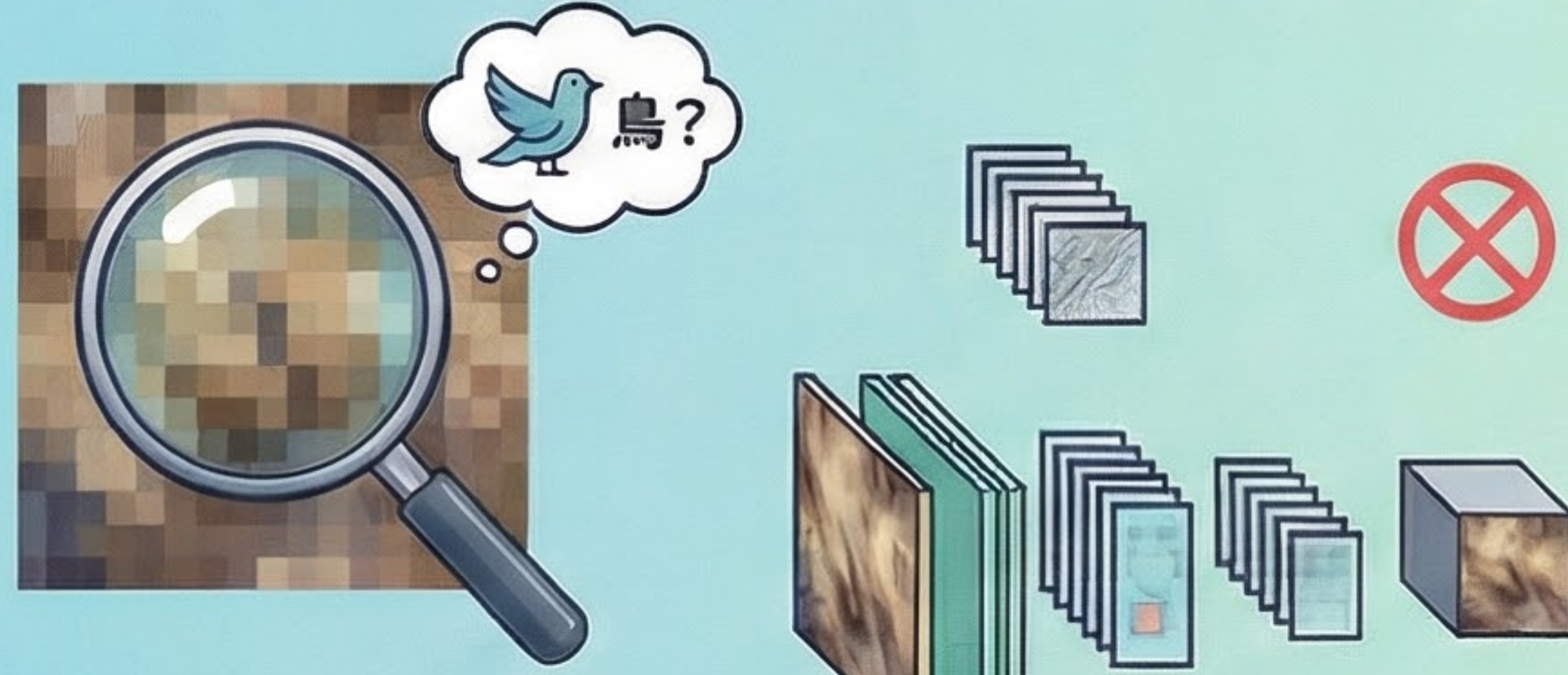


画像認識のパラダイムシフト：CNNからTransformerへ

2012年以降、画像認識はCNN（畳み込みニューラルネットワーク）が標準でしたが、2020年頃から自然言語処理で成功したTransformerが画像分野にも進出しました。
この「木を見るCNN」から「森を見るTransformer」への移行は、AIの認識精度やマルチモーダル対応に革命をもたらしています。

CNNは「木を見て森を見ず」



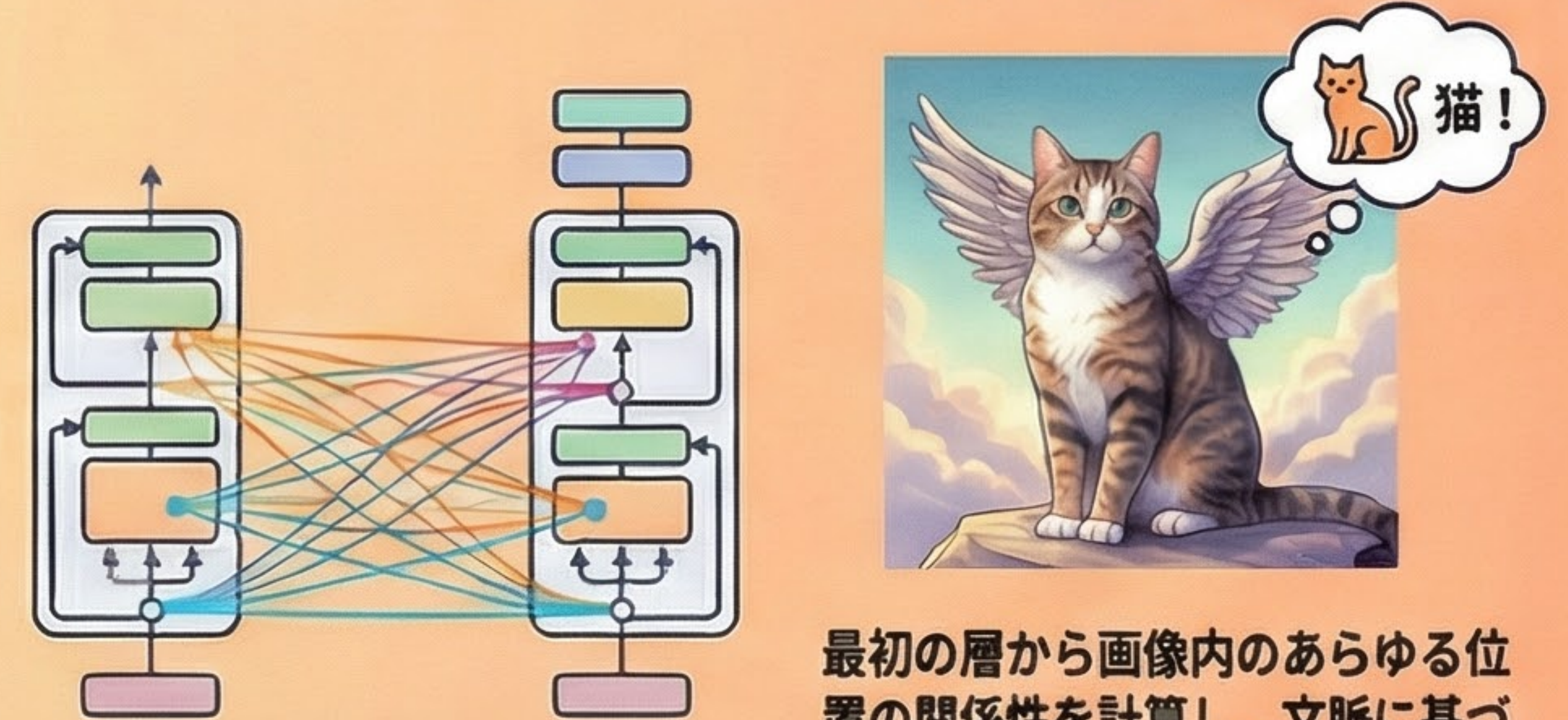
CNNは局所的な模様を積み上げますが、Transformerは画像全体の関連性を一層に捉えます。

自己注意機構（Self-Attention） によるグローバルな理解



最初の層から画像内のあらゆる位置の関係を

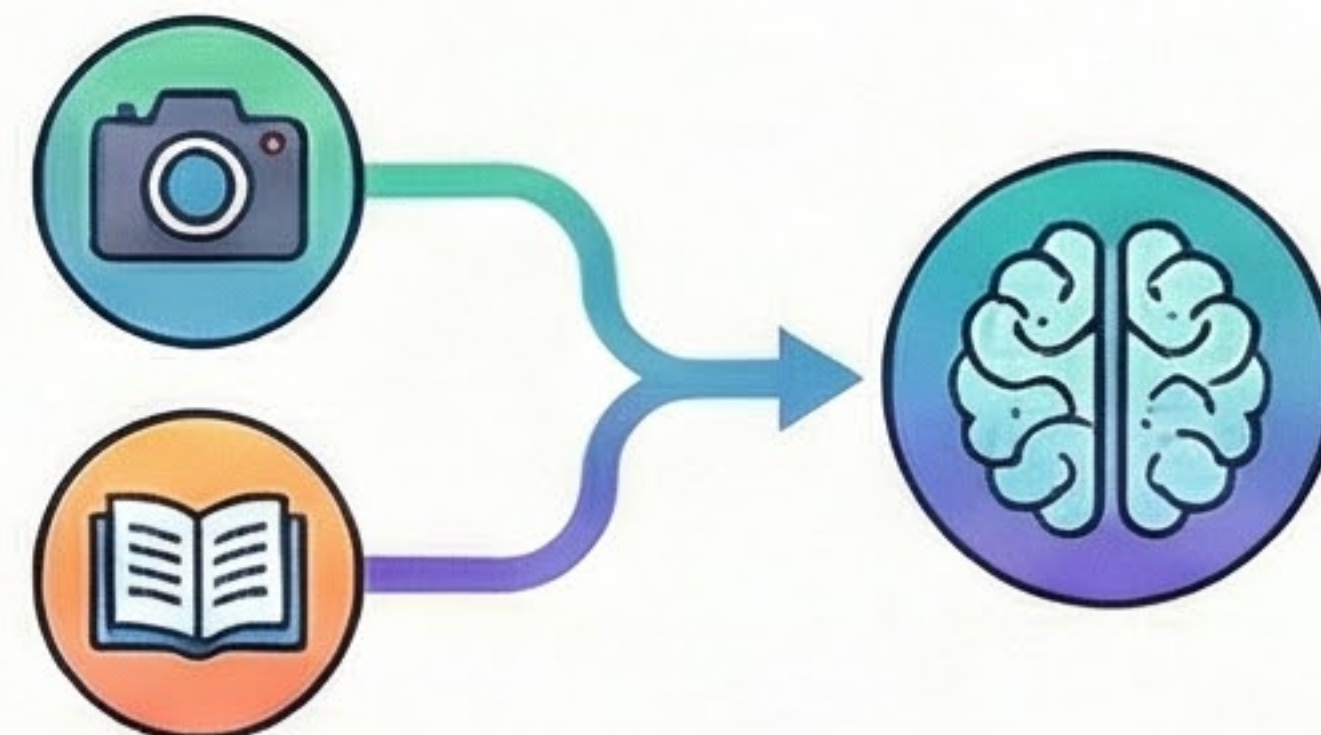
Transformerは「森を俯瞰」



最初の層から画像内のあらゆる位置の関係を計算し、文脈に基づいた的確な認識を可能にします。

実用上の特性とトレードオフ

「見る」と「読む」を繋ぐマルチモーダル化の加速



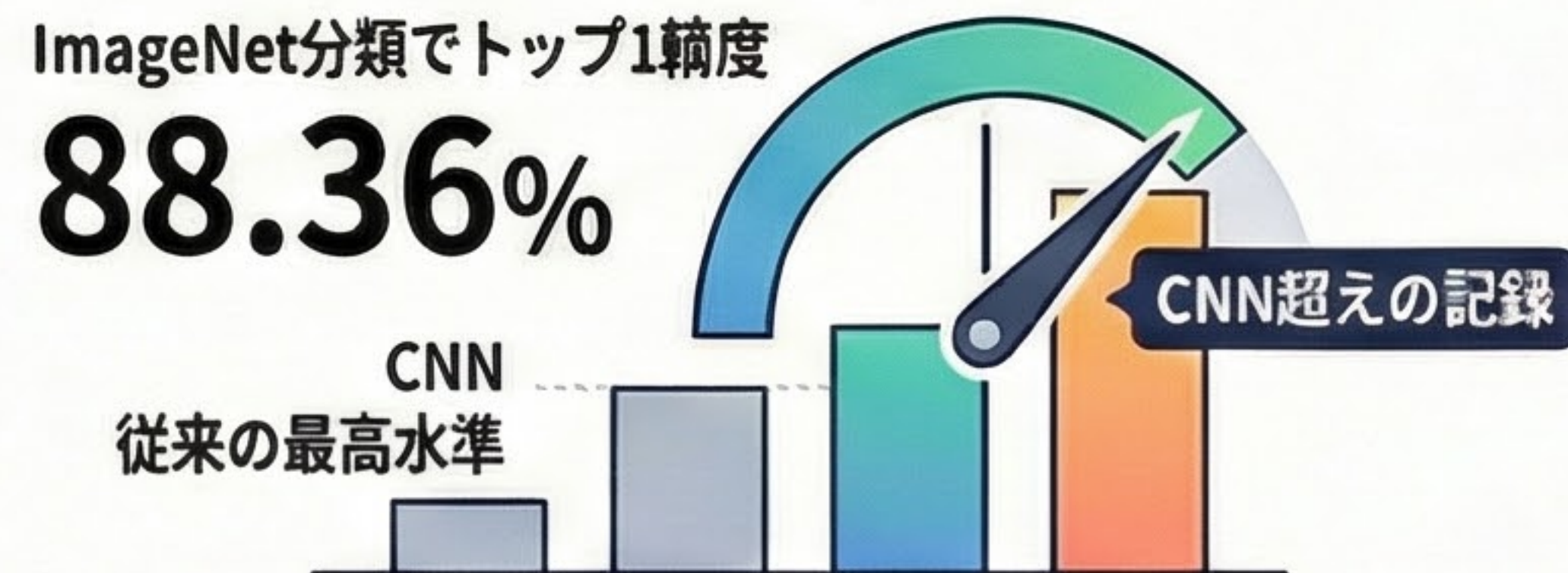
言語と同じアーキテクチャのため、画像とテキストを統合した高度なAI開発が容易になります。

課題は膨大な計算資源と学習コスト



モデルによっては軽量のCNNに比べ30倍以上の計算量を要するため、効率化が今後の普及の鍵です。

大規模データで真価を発揮する高い表現力



数億枚規模の学習により、CNN超えの記録を達成しました。