Scaling Efficiency and Top-1 Accuracy of Vision Transformers vs. Five-Layer CNNs

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do modern attention-based vision transformers compare to the five-layer convolutional architecture in terms of scaling efficiency and top-1 accuracy when trained on high-resolution image. Modern computer vision offers a great variety of models to practitioners, and selecting a model from multiple options for specific applications can be challenging. Conventionally, competing model architectures and training protocols are compared by their classification accuracy. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ConvNet vs Transformer, Supervised vs CLIP: Beyond ImageNet Accuracy. Research question: How do modern attention-based vision transformers compare to the five-layer convolutional architecture in terms of scaling efficiency and top-1 accuracy when trained on high-resolution image classification tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

16 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
ViT and ConvNeXt have similar performance across both supervised and CLIP.	×	0.09
ConvNeXt is superior on almost every factor for both supervised and CLIP on synthetic data from PUG-ImageNet.	×	0.06
The transfer learning performance of a model indicates its ability to adapt to new tasks and datasets beyond its origina	×	0.04
Good transferability allows for rapid finetuning with minimal additional effort, making it easier to scale the model to	×	0.03
The ability of a model to adapt to shifts without significant degradation in performance serves as a valuable metric for	×	0.03
The VTAB benchmark comprises 19 diverse datasets grouped into three subcategories: natural, specialized, and structured.	×	0.04
High ImageNet accuracy does not guarantee good performance on diverse datasets.	×	0.07
Current robustification training techniques were found to overfit to ImageNet evaluations.	×	0.04
ImageNet suffers from dichotomous data difficulty, obscuring differences between models.	×	0.03
For domain shift, CLIP models are recommended.	×	0.04
ConvNeXt-sup has an accuracy of 85.5 and ECE of 0.028 on ImageNet-1K.	×	0.03
ViT-sup has an accuracy of 85.5 and ECE of 0.041 on ImageNet-1K.	×	0.03
ConvNeXt-clip has an accuracy of 66.3 and ECE of 0.141 on ImageNet-1K.	×	0.04
ViT-clip has an accuracy of 67.0 and ECE of 0.133 on ImageNet-1K.	×	0.04

References

- `http://arxiv.org/abs/1807.11583v1`
- `http://arxiv.org/abs/2205.04596v2`
- `http://arxiv.org/abs/2311.09215v3`