SOVEREIGN: What is the computational overhead of layer-wise score aggregation method compared to last-layer-only baseline

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

The recent breakthroughs in natural language processing for model pretraining on large quantities of data have opened the way for similar foundation models in computer vision. These models could greatly simplify the use of images in any system by producing all-purpose visual features, i.e., features that work across image distributions and tasks without finetuning. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit existing approaches and combine different techniques

1 Introduction

Analysis of: DINOv2: Learning Robust Visual Features without Supervision. Research goal: What is the computational overhead of layer-wise score aggregation method compared to last-layer-only baselines on SuperGLUE benchmark tasks?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 7.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Existing pretraining methods, especially self-supervised methods, can produce all-purpose visual features if trained on	✓	0.40
The authors propose an automatic pipeline to build a dedicated, diverse, and curated image dataset instead of uncurated	✓	0.29
The authors train a ViT model with 1B parameters and distill it into a series of smaller models.	✓	0.22
The distilled smaller models surpass the best available all-purpose features, OpenCLIP, on most benchmarks at image and	✓	0.28

References

- https://doi.org/10.1016/j.inffus.2023.101805
- https://doi.org/10.18653/v1/d19-1053
- https://doi.org/10.48550/arxiv.2304.07193