UDC 004.85:004.93

THE CASE FOR INFERENCE-TIME COMPUTE SCALING IN COMPUTER VISION MODELS

Lukashov D.S.
e-mail: dmytro.lukashov@nure.ua
Scientific Supervisor – Cand. Sc. (Tech.), Ass. Prof. Kozyrenko S.I.
Kharkiv National University of Radio Electronics,
Department of Applied Mathematics
Kharkiv, Ukraine

Нещодавня траєкторія розвитку штучного інтелекту значною мірою була сформована законами масштабування попереднього навчання. Однак, впровадження масштабування обчислень (так званий «ланцюг думок») під час інференсу започаткувало нову еру можливостей, головним чином в обробці природної мови. Новітні архітектури, що дозволяють моделям «думати» під час тестування, дають значне підвищення продуктивності. Однак, комп'ютерний зір залишається закріпленим у попередній парадигмі. У цій статті обґрунтовується необхідність дослідження переходу моделей комп’ютерного зору до архітектур, де обчислення під час інференсу масштабуються динамічно.

Historically, the dominant paradigm for improving the capabilities of deep neural networks has been to scale the size of the training dataset and the number of model parameters. While this approach has yielded remarkable advances, the computational and data-gathering costs of pre-training are beginning to encounter practical limits. In response, a new frontier has emerged in the form of inference-time compute scaling. By allowing a model to "think" for longer periods before producing an output, researchers have unlocked complex reasoning capabilities that were previously thought to be exclusive to vastly larger models.

The transition from static inference to dynamic, test-time compute allocation in language models has demonstrated that reasoning can be dramatically improved without altering the underlying model parameters. Crucially, scaling inference-time compute has been shown to yield significant benefits beyond mere accuracy. For example, studies have demonstrated that allocating additional inference-time compute significantly enhances a model’s adversarial robustness across diverse domains [1]. This capacity to iteratively refine responses, parse ambiguity, and reject adversarial manipulations highlights a fundamental shift: performance bottlenecks can be alleviated by extending the computation spent per query.

The theoretical groundwork for allowing models to adjust their behavior at test-time has been in development for years. Historically, attempts to modify model behavior dynamically during testing through reinforcement learning showed early promise by adjusting computation conditionally [2]. By formalizing the inference stage as a sequential decision-making process, models could

theoretically learn when to expend more effort. Today, these concepts have evolved into sophisticated search and reasoning trees that allow LLMs to backtrack, verify, and correct their own logic in real-time.

In computer vision, the utilization of computation at test-time has historically been relegated to Test-Time Adaptation (TTA). TTA aims to mitigate domain shifts, such as changes in lighting, weather, or camera sensors by updating model parameters or representations based on unlabeled test data at deployment.

For example, in the realm of vision-language models, recent approaches utilize lightweight dynamic adapters and key-value caches to refine pseudo-labels efficiently during inference without the need for backpropagation [3]. Similarly, in video processing, self-supervised adaptation techniques have been developed to handle severe covariate shifts in dense tracking applications by leveraging the temporal continuity of video frames to adapt the network to novel visual distributions [4].

While these methods are highly effective for domain generalization, they fundamentally differ from the concept of "thinking" or "reasoning." Adaptation adjusts the model to the statistical distribution of the input; it does not equip the model to iteratively solve a complex visual puzzle. The network remains a feed-forward mechanism that outputs a rapid, albeit better-calibrated, guess based on immediate feature extraction.

The limitations of single-pass inference become glaringly apparent when vision models are tasked with high-level cognitive functions. True object recognition requires more than mere pattern matching; it necessitates inference across causal dependencies, object attributes, and spatial-temporal relations [5]. Consider the task of causal reasoning in video-understanding that a glass shattered because it fell, or predicting the hidden trajectory of an occluded object. A feed-forward model attempts to map pixel patterns directly to these complex conclusions, an approach prone to catastrophic failure when confronted with novel compositional structures or adversarial perturbations.

If computer vision models were endowed with inference-time "thinking," they could actively construct and evaluate multiple scene hypotheses. This mimics the System 2 cognitive processing in humans. For example, a radiologist examining a complex MRI scan does not rely on a single, instantaneous glance. Instead, they trace anomalies, cross-reference different lobes, actively generate differential diagnoses, and physically adjust their focus to verify their hypotheses. A "thinking" vision model would operate similarly. When presented with a heavily occluded scene, it would not simply output a low-confidence classification. Instead, it would allocate additional compute to iteratively generate internal representations of the occluded geometries, verify these geometries against visible semantic cues, and deduce the most physically plausible scene configuration.

To bridge this gap, future CV architectures must decouple perception from reasoning. While the initial layers of a network can remain fast, feed-forward feature extractors (System 1), the deeper layers should operate as an iterative

workspace (System 2). Research must focus on developing visual verifiers – functions that can score the physical and semantic plausibility of a generated visual hypothesis. With a reliable verifier, vision models could employ tree-search algorithms over the latent space, testing multiple interpretations of a scene before committing to a final bounding box or segmentation mask. Furthermore, integrating test-time physical simulation could allow models to "run" a mental video of a scene to verify if their structural understanding holds up to the laws of physics.

The introduction of inference-time compute scaling has revolutionized artificial intelligence, yet its analog in computer vision remains largely unexplored. While the CV community has successfully harnessed test-time compute for domain adaptation, it has not yet leveraged it for dynamic, iterative visual reasoning. Transitioning to models capable of visual "thinking" will be crucial for solving complex, structurally rich tasks such as causal video analysis, robust medical imaging, and compositional scene understanding. As the field looks forward, developing mechanisms for inference-time search, hypothesis generation, and self-correction in the visual domain stands as one of the most pressing and promising frontiers in machine learning research.

References:

1. Trading inference-time compute for adversarial robustness / W. Zaremba [et al.] // arXiv. 2025. URL: https://doi.org/10.48550/arxiv.2501.18841 (accessed: 06.03.2026).

2. Odena A., Lawson D., Olah C. Changing model behavior at test-time using reinforcement learning // International Conference on Learning Representations (ICLR). 2017. URL: https://doi.org/10.48550/arXiv.1702.07780 (accessed: 06.03.2026).

3. Efficient test-time adaptation of vision-language models / A. Karmanov [et al.] // 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) : proceedings. 2024. P. 13842–13852. DOI: https://doi.org/10.1109/CVPR52733.2024.01314.

4. Self-supervised test-time adaptation on video data / F. Azimi [et al.] // 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV): proceedings. 2022. P. 2603–2612. DOI: https://doi.org/10.1109/WACV51458.2022.00266.

5. Sarkar A., Idris M. Y. I., Yu Z. Reasoning in computer vision: Taxonomy, models, tasks, and methodologies // IEEE Access. 2025. Vol. 13. P. 45210–45235. DOI: https://doi.org/10.1109/ACCESS.2025.10523.