Paper Critique: LLaVA-NeXT-Interleave

Dan Peng
Department of Computer Science
Oct 22, 2024
danpeng@unc.edu

1. Research Problem

1.1. What research problem does the paper address?

The paper addresses the research problems that existing LLMs largely focus on single-image tasks but not multi-image scenarios. Also, existing LLMs handles different scenarios separately leaving it impossible to generalize cross scenarios with new emerging capabilities.

1.2. What is the motivation of the research work?

The motivation is that training separate task-specific models for each application scenario is both labor-intensive and time-consuming. As a result, these fragmented methods are inefficient and often unscalable. The authors found that the image-text interleaved format can naturally serve as a general data template to unify different scenarios and solve the previous challenges.

2. Technical Novelty

2.1. What are the key technical challenges identified by the authors?

The key technical challenges are:

- How to extend the capability of LLM to handle multiple scenarios while in a single model?
- How to develop a unified data format of different tasks?
- How to train the model maintaining both efficiency and performance in each domain?
- How to evaluate the model's performance with high-quality benchmarks?

2.2. How significant is the technical contribution of the paper? If you think that the paper is incremental, please provide references to the most similar work

This paper offers the following contributions:

- converted multi scenarios into an interleaved training format, unifying different tasks in a single LLM
- compiled a high-quality training dataset, **M4-Intstruct**, spanning 4 primary domains (multi-image, video, 3D, and single-image) with 14 tasks and 41 datasets
- curated a new diverse set of benchmarks, **LLaVA-Interleave Bench**, to evaluate the model
- achieved SOTA results across different multi-image tasks with a single model
- showcased capabilities to transfer tasks across different settings and modalities

2.3. Identify 1-5 main strengths of the proposed approach.

- By applying an interleaved image-text format, LLaVA-NeXT-Interleave provides a flexible and scalable solution that can handle multiple real-world scenarios
- LLaVA-NeXT-Interleave achieves SOTA performance across a variety of tasks, including multi-image, video, and 3D benchmarks
- LLaVA-NeXT-Interleave performs pretty good at cross-task transfering
- It also curates a comprehensive benchmark and dataset while proposing this new method

2.4. Identify 1-5 main weaknesses of the proposed approach.

- Training and fine-tuning LLaVA-NeXT-Interleave across multiple modalities (multi-image, video, 3D) requires significant computational resources. Since this paper didn't mention its strength in computational costs, we might assume it failed to perform very well than other previous SOTA models.

- The performance of LLaVA-NeXT-Interleave heavily relies on large diverse and high-quality datasets like M4-Instruct. In this case, it may perform worse in those domains with unavailable high-quality dataset.

3. Empirical Results

3.1. Identify 1-5 key experimental results, and explain what they signify.

- LLaVA-NeXT-Interleave achieved leading results in multi-image and VQA tasks, outperforming previous SOTA models, e.g. outperforming models like GPT-4V and Mantis in several benchmarks across both in-domain and out-domain tasks. It indicates that LLaVA-NeXT-Interleave has the strong capability of handling diverse multimodal inputs.

- In ablation on the improvement of combined data scenarios for video tasks, the model was able to transfer learned knowledge across single-image tasks to multi-image or video tasks. For example, it successfully applied video reasoning from single-image to multi-image scenarios. And it achieved the highest results of combining both single-image and multi-image data for video tasks, highlighting its generalization ability.

3.2. Are there any weaknesses in the experimental section (i.e., unfair comparisons, missing ablations, etc)?

Yes, there is an unfair comparison in the table2 results on multi-frame benchmarks. They failed to demonstrate the results of most previous SOTA methods on NExTQA and MVBench, where these results showcases LLaVA-NeXT-Interleave achieves the best performance. While on other benchmarks, LLaVA-NeXT-Interleave is not always the best model. Even sometimes GPT-4V performs better than LLaVA-NeXT-Interleave 7B, they bold the results of their method. For example, GPT-4V achieves 4.09 at CI of VideoChat-GPT while LLaVA-NeXT-Interleave 7B has bolded 3.99.

4. Summary

The paper presents LLaVA-NeXT-Interleave which unifies multi-image, video, 3D, and single-image tasks to a single model. By applying an interleaved image-text input format, the model achieved the SOTA results on diverse tasks such as video understanding, multi-image visual reasoning, and 3D perception, all while maintaining high performance in single-image tasks. I like this paper because it addresses the pressing challenge of handling multiple modalities in a unified way, providing a powerful solution that outperforms in both general-purpose and complex multimodal tasks.

5. QA Prompt for a Paper Discussion

5.1. Discussion Question

What are the three additional techniques to improve performance in multi-image tasks compared to the LLaVA-NeXT model?

5.2. Your Answer

- Tech1: Continue training from single-image models
- Tech2: Mixed Interleaved data formats during training
- Tech3: Combining different data scenarios improves individual task performance