Paper Critique: Perception Test

Dan Peng
Department of Computer Science
Oct 30, 2024
danpeng@unc.edu

1. Research Problem

1.1. What research problem does the paper address?

The research problem addressed by the paper is the lack of comprehensive benchmarks that evaluate multimodal models' general perception and reasoning skills across different modalities (video, audio, text), in ofder to assess their ability to generalize to real-world scenarios beyond single computational tasks in a zero-shot or few-shot learning scenario.

1.2. What is the motivation of the research work?

The motivation of the research is to create a comprehensive diagnostic benchmark that evaluates multimodal models on their generalization and reasoning capabilities across different skills areas, modalities and reasoning types, while also evaluating their ability to generalize to real-world tasks in zero-shot or few-shot settings. This is able to accelerate the development of more versatile and robust multimodal models.

2. Technical Novelty

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

Key technical challenges are: limited scope of existing benchmarks and annotations; some rare data types are missing in public; the lack of multimodal and complex reasoning evaluation; difficulties in assessing generalization and transfer learning; overfitting and bias.

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

Compared to the established paradigm of multimodel benchmarks, the Perception Test is designed to evaluate the general perception capabilities of multimodal models across multiple cognitive skill domains, which are often overlooked in existing benchmarks. For example, benchmarks like Ego4D and EPIC are similar to Perception Test in their use of real-world video data collected from egocentric perspectives to assess action recognition. However, they focus mainly on action and object detection rather than the diagnostic assessment of diverse cognitive skills and reasoning types across video, audio, and text that Perception Test targets.

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

- The paper has comprehensive multimodal evaluation, including over 11,600 real-world videos densely annotated with 6 types of labels
- Able to assess models in a zero-shot, few-shot, or limited fine-tuning setting, emphasizing generalization and transferability over extensive training on the benchmark itself
- The results highlight current limitations in model capabilities between human cognition and provides clear targets for future improvement

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

- Provides a limited training set for fine-tuning though it's beneficial for testing zero-/few-shot abilities of models

3. Empirical Results

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

- The paper reports that SOTA video-language models like Flamingo and SeViLA, even in a fine-tuned setup, perform worse than humans on the multiple-choice VQA task. E.g. humans achieve an accuracy of 91.4%, while fine-tuned SeViLA only reaches 62%. It indicates that current models struggle with high-level reasoning and complex multimodal understanding compared to human cognitive abilities.

- Both Flamingo and SeViLA also perform worse than the 8-shot frequency dummy baseline on specific skills, particularly in physics-related tasks, such as understanding object collisions, stability, and conservation. It indicates models lack of physical reasoning and abstract conceptualization and suggests to modifying training strategies for physical and causal understanding.

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

Yes, the ablation studies are missing.

4. Summary

This paper showcases a novel benchmark designed to evaluate the perception and reasoning skills of multimodal models across multiple different real-world scenarios. Unlike traditional benchmarks that often focus on isolated tasks, the Perception Test is a comprehensive assessment that evaluates models' generalization capabilities in zero-shot, few-shot, or limited fine-tuning settings. I like this paper because it introduces human baselines and presents their performance on the benchmarks. The research team involved around 100 participants from multiple countries, ensuring a range of racial, ethnic, and gender representations in the video data. This diversity is reflected in the scenes, scripts, and objects within each video, making the dataset more representative of real-world variability.

5. QA Prompt for a Paper Discussion

5.1. Discussion Question

What are some potential challenges models might face when performing counterfactual reasoning in the Perception Test benchmark?

5.2. Your Answer

In Perception Test, models might struggle with counterfactual tasks that ask them "what-if" questions and face challenges such as lacking robust causal understanding and context sensitivity.