CompGPT: Probing the Compositional Understanding of Visual Language Models

Evan Wang[1], Advisor: Felix Heide[1],

1. Department of Computer Science, Princeton University

Summary

- Vision-language-models (VLMs) struggle to understand compositional relations (spatial and property concepts) [4]
- VLMs that are trained on spatially-relevant objectives, such as word patch alignment, still have this weakness.
- Existing datasets for benchmarking compositional reasoning are flawed: caption bias hampers true testing of compositional understanding. [1]
- State-of-the-art language models provide a solution for mitigating these issues.
- Composition-Aware Fine Tuning can improve compositonal understanding, but validity of the data used is crucial.

Motivation

VLMs are weak at compositional reasoning. For example, given the image in Figure 1, some VLMs may not understand that the dog is to the left of the cat, and the cat is to the right of the dog.

Figure 1: Sample image to test compositional reasoning.

- Dataset that evaluates is Attribution, Relation, and Order (**ARO**). [4] Each image in the dataset has two captions: one true caption, e.g. "dog to left of cat", and a deliberately false caption generated by a rule-based approach that swaps the entities, e.g. "cat to left of dog." VLMs are tasked with choosing correct caption. We benchmark on ARO and investigate its validity.
- VLMs we focus on:
  - OpenAI's **CLIP** for its foundational and high performance on ImageNet
  - Vision and Language Transformer (**ViLT**) for its simple, efficient transformer-based architecture [2]

Approach

- Investigate validity of ARO with manual inspections and language-model-based analyses of captions.
- Refine ARO captions with GPT models
- Evaluate ViLT and CLIP on original and new ARO datasets.

Implementation

Validity of ARO

- Quantify bias between true and false captions via perplexity:

$$\exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1,\ldots,w_{i-1})\right) \quad (1)$$

as measured by GPT-2. Higher perplexity indicates less sensical captions. False caption perplexity > true caption perplexity indicates a trivial benchmarking question.

Improving ARO

- Use GPT to create false captions that are coherent and logical. Tell model to make differing caption that maintains coherency and realism. Few-shot prompting with examples.
- Two approaches:
  - Chatbot Interface, GPT-4 model. Input and output JSON files.
  - OpenAI API, GPT-3.5 model. Input one caption per prompt.

Composition-Aware Fine-Tuning NegCLIP model: Maximize contrast between true/false caption embeddings.

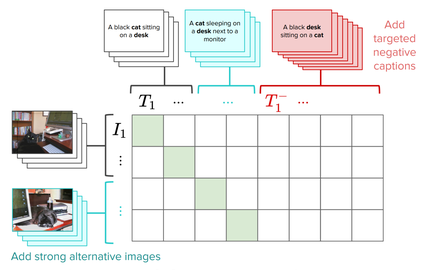

Figure 2: Visualization of Composition-Aware Fine-Tuning for CLIP (CITE)

Evaluation Evaluate CLIP, ViLT, and NegCLIP (CLIP with Composition-Aware Fine-Tuning) on original and refined ARO.

Results

- 59.4% of pairs of captions in ARO are trivial. Best GPT approach reduces triviality to 57.3% and gap to 16.0. Distribution shown in Figure 3.

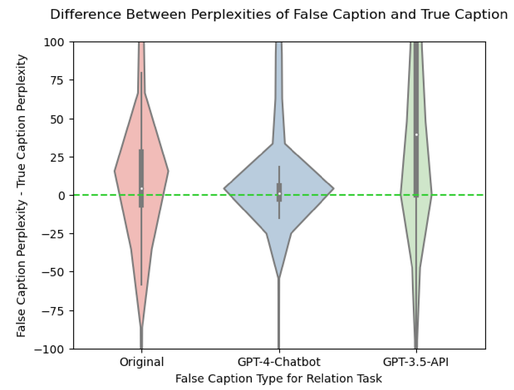

Figure 3: Violin plot and box plot showing perplexity gap distributions for original ARO, GPT4-Chatbot, and GPT-3.5 API.

- Composition-Aware Fine-Tuning improves performance to various degrees.

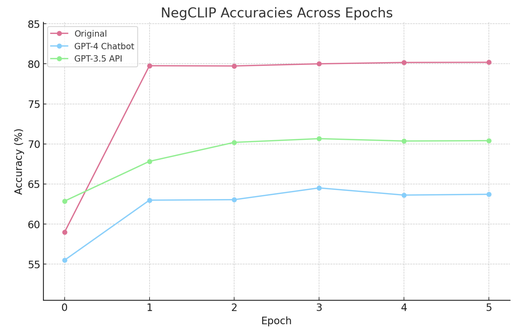

Figure 4: Performance of NegCLIP across training epochs

- Overall Accuracies

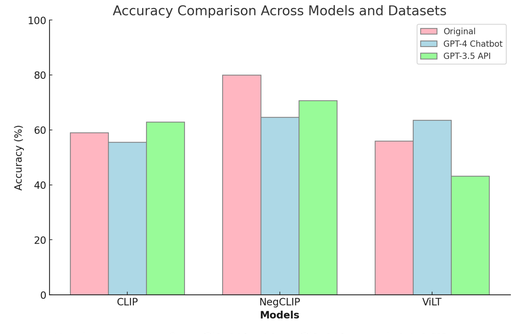

Figure 5: Accuracy for CLIP, NegCLIP, and ViLT, across datasets

Discussion

60% of questions in original ARO have less sensical and logical false captions, indicating a trivial task.

- **"Blind" language model that solely considers caption can pick correctly.**
- GPT can reduce this bias.

Figure 6: ARO image relating cat and sink. True caption is "the cat is on the sink" (perp. 232.0). False caption is "the sink is on the cat." (perp. 287.6). New false caption is "the cat is under the sink" (perp. 195.7)

- Run-Time Analysis for Caption-Generation Procedure. Linear complexities.

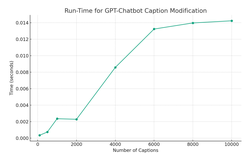

Figure 7: Run-Time for GPT Chatbot Interface

Figure 8: Run-time for GPT OpenAPI API

- GPT-4 can reduce bias. GPT-3.5 increased bias, which may be due to weakness in adhering to realistic alternative captions.
- Original ARO **overestimates** compositional understanding. Reasoning is very weak, even for ViLT that has a spatial training objective.
- Composition-Aware Fine Tuning does improve performance on ARO benchmark, but improvements are limited on refined datasets.

References & Acknowledgement

[1] C.-Y. Hsieh et al., "Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality," 2023.
[2] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," 2021.
[3] E. Wang, "Investigating clip's understanding of the vision-language space," 2023, final project for Computational Models of Cognition, Griffiths, Princeton University.
[4] M. Yuksekgonul et al., "When and why vision-language models behave like bags-of-words, and what to do about it?" 2023.

Acknowledgement: Thanks to Professor Heide for his seminar, Howard Chen for his inspiration, and the SML Center for their support.