Do Vision Language Models Understand Compound Nouns?

Sonal Kumar[1], Sreyan Ghosh[1], S Sakshi[1], Utkarsh Tyagi[1], Dinesh Manocha[1]
[1]University of Maryland, College Park

NAACL 2024

Introduction

Compound nouns (CNs) combine two or more words to form a single noun with a new meaning (e.g., "paper towel", "full moon"). We focus on the type combining two nouns.

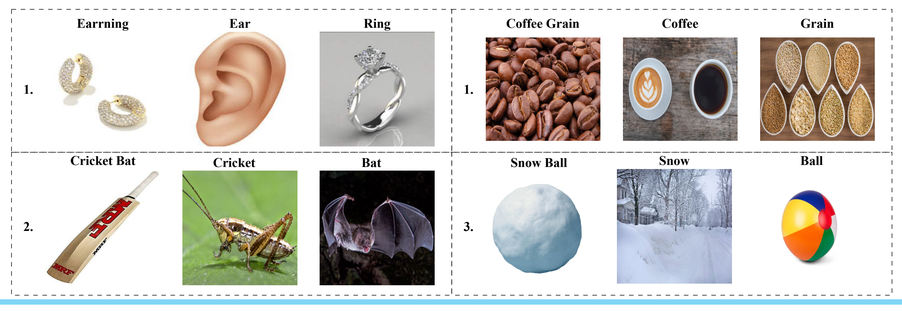
Understanding CNs involves deciphering the semantic relations between constituent nouns, a longstanding challenge in NLP.

Despite advances, the ability of VLMs like CLIP, trained with contrastive loss, to comprehend these semantic relations in CNs is poorly understood and underexplored.

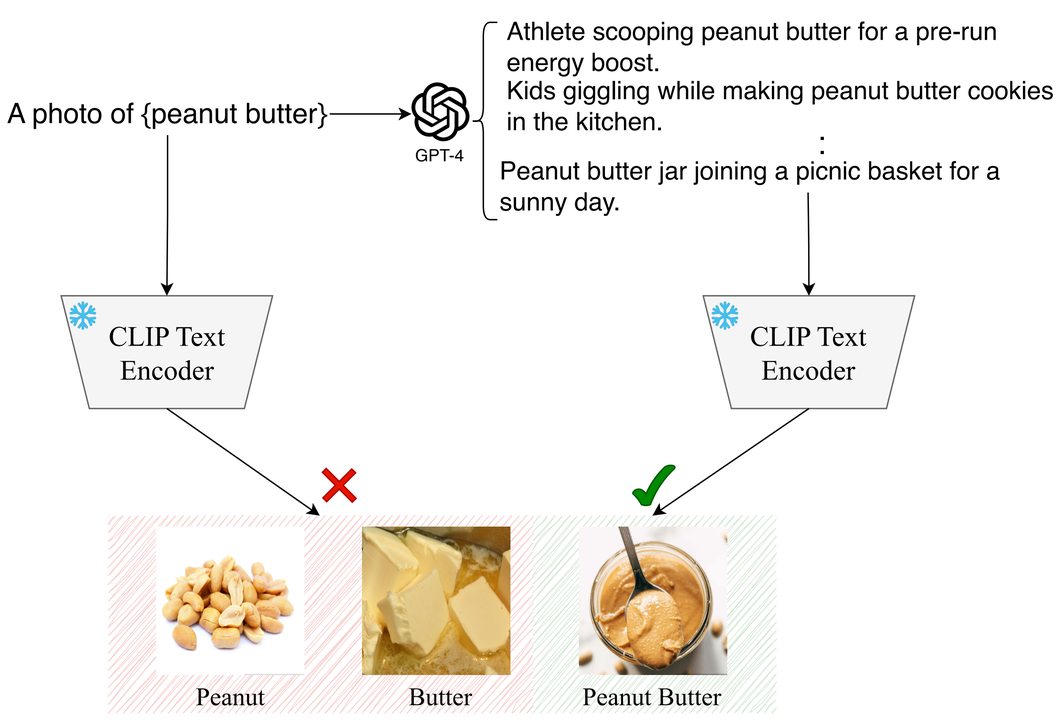
Main Contributions:

- **Compun Benchmark:** Introduced a new benchmark with 400 unique compound nouns paired with images to evaluate VLMs on text-to-image retrieval tasks. This tool challenges VLMs to distinguish between images of compound nouns and their constituent parts.
- **In-depth Analysis:** Conducted a comprehensive analysis of the CLIP model's performance on the Compun benchmark, providing insights into its limitations in interpreting compound nouns.
- **Novel Framework for Improvement:** Proposed an innovative framework that utilizes Large Language Models (LLMs) to generate diverse captions that include compound nouns. This method enhances VLMs' understanding of CNs, improving retrieval accuracy by incorporating contextually richer prompts.
- **Performance Enhancement:** Demonstrated significant improvements in the interpretation of compound nouns by VLMs, with an increase of 8.25% in performance on the Compun benchmark using our method compared to existing techniques.

Types of CN in Compun



Method



Prompt

Return a list of 5 diverse captions with a com- pound_noun in a photo. The captions should be a maximum of 10 words and one-liners. All 5 captions should describe the compound noun in diverse settings with different verbs and actions being performed with the com- pound noun. An example output for "chicken burger": ['Sizzling chicken burger grilling at a lively backyard BBQ.', 'Chef expertly flip- ping a juicy chicken burger in a diner.", Fam- ily enjoying homemade chicken burgers on a sunny picnic.', 'Athlete fueling up with a protein-packed chicken burger post-workout.', 'Friends sharing a chicken burger at a vibrant street festival.']. Only return a list of strings and nothing else.

Evaluation Metric

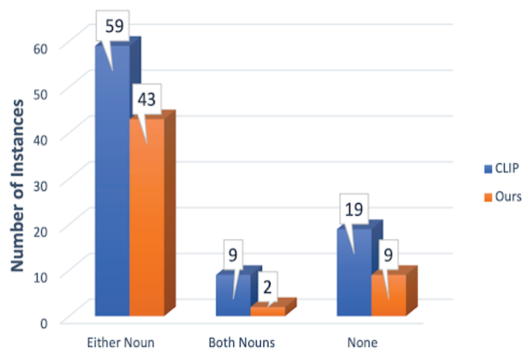
$$f(\mathcal{C},\mathcal{P},\mathcal{N}_1,\mathcal{N}_2)=\begin{cases}1 & \text{if } s(\mathcal{C},\mathcal{P})>s(\mathcal{C},\mathcal{N}_1)\\ & \text{and } s(\mathcal{C},\mathcal{P})>s(\mathcal{C},\mathcal{N}_2)\\ 0 & \text{otherwise}\end{cases}$$

$\mathcal{P}$: The image illustrating the compound noun as a positive
$\mathcal{N}_1$, $\mathcal{N}_2$: The other 2 distractor images illustrating the compound noun as as negatives
C : The natural language prompt for the compound noun
$s(.)$: The standard cosine similarity, widely used for retrieval

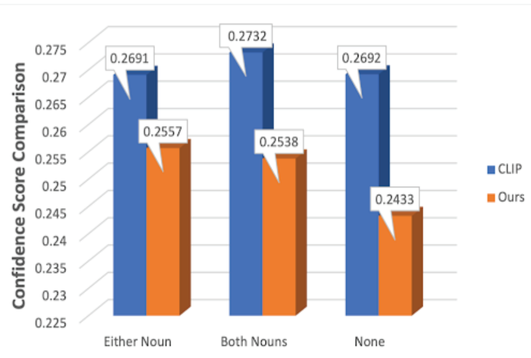
Result & Analysis

Model	Text-to-Image Acc.
Human	96.25
Random	33.33
ALBEF (Li et al., 2021)	80.55
BLIP (Li et al., 2022a)	79.85
MetaCLIP (Xu et al., 2023)	81.35
CLIP (Radford et al., 2019)	78.25
CLIP *rev.*	41.00
CLIP *w/ desc* (Menon and Vondrick, 2023)	81.15
CLIP *w/ examples (ours)*	86.50
OpenCLIP (Ilharco et al., 2021)	83.90
OpenCLIP *w/ examples (ours)*	86.25

Comparison of our proposed version of CLIP with other baselines on the Compun benchmark. Our proposed method outperforms CLIP by 8.25% and OpenCLIP by 2.35%.



Count of misclassified instances by CLIP on Com- pun for three settings, either, both, and none. Section 6 de- scribes these settings. CLIP is more likely to retrieve a neg- ative when the positive image shows either constituent noun, highlighting CLIP's limited understanding of attributed CNs.



Average CLIP similarity scores for correct predic- tions on Compun on three unique settings, either, both, and none. Section 6 describes these settings. High scores on the Compun benchmark are superficial, and CLIP often wins by low similarity scores.

https://arxiv.org/pdf/2404.00419

https://github.com/sonalkum/Compun