Analyzing the Semantic Robustness of Vision-Language Models (VLMs)

Abhiram Dodda
Department of Computer Science and Engineering
Pennsylvania State University
University Park, PA 16802
amd8734@psu.edu

Abstract

Vision-Language models like CLIP demonstrated remarkable capabilities in cross-modal alignment and zero-shot retrieval. However, real-world scenarios can be different from the training data distribution and the robustness of the models to natural language decides the reliability. This project investigates CLIP(ViT-B/32) model aganist semantically similar text, generated by modifying original captions.

1 Methodology

1.1 Dataset and Model

Flickr30k dataset is chosen. 1000 images subset is chosen to complete the tasks. The project uses the pre-trained CLIP model, specifically the ViT-B/32 variant ("openai/clip-vit-base-patch32"). The model weights are frozen throughout the evaluation.

1.2 Baseline Metrics

CLIP image encoder and CLIP text encoder are used extract image and text features respectively for all 1000 image-caption pairs and are normalized to embedding space. A cosine similarity matrix is computed and finally Recall@K where $K \in 1, 5, 10$. This measures the percentage of queries where the correct image is ranked among the top K results.

1.3 Semantic Attacks

The attacks involve creating modified captions that maintain semantics but change the lexical structure. All attacks employ NLTK for part-of-speech (POS) tagging and WordNet lookups.

Synonyms Identifies nouns, verbs, and adjectives using NLTK POS tagging. Randomly replaces a portion of these words with a synonym retrieved from WordNet, while avoiding stopwords. The attack is performed using three percentages, 25, 50 and 75.

Hypernym and Hyponym Attacks Theses attacks focus on replacing nouns only. Hypernym generalizes the concept, while Hyponym details more. 25% and 50% are used for this attack.

Para-phrasing Attack T5 model is used to generate alternate captions for this attack. Complete sentence structure is changed while details are left untouched.

Preprint. Under review.

2 Results

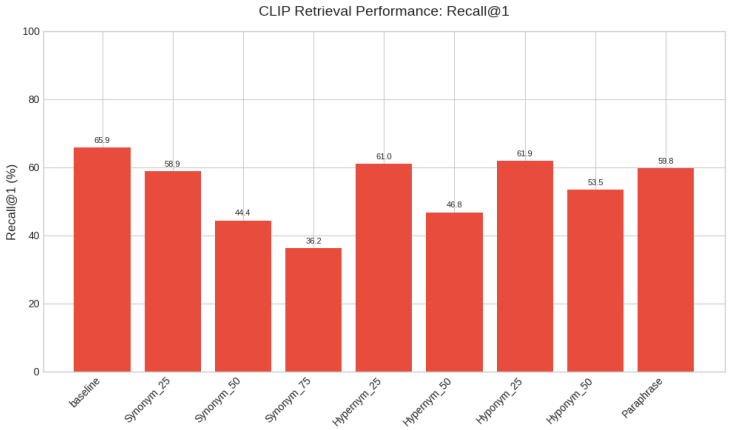

Figure 1: Recall@1 vs Baseline and Attacks

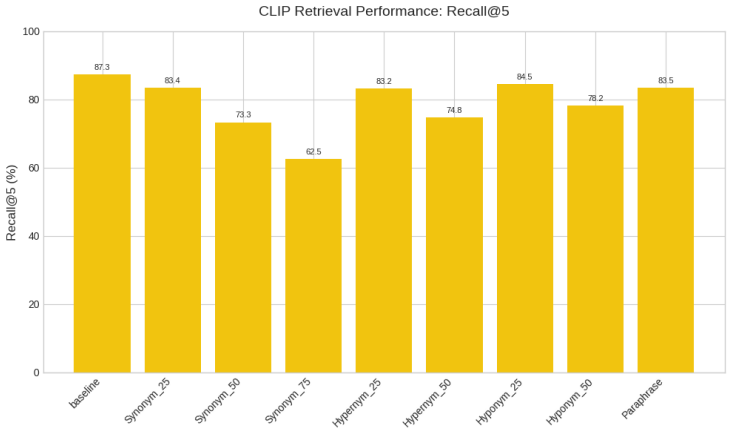

Figure 2: Recall@5 vs Baseline and Attacks

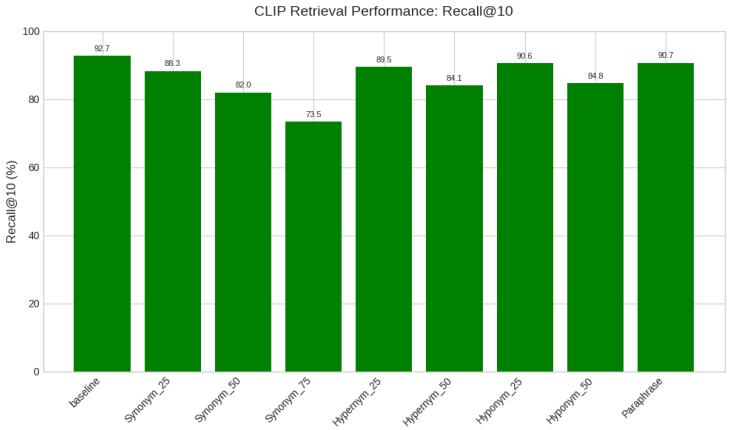

Figure 3: Recall@10 vs Baseline and Attacks

3 Examples

Original Caption: A person in gray stands alone on a structure outdoors in the dark. Correctly identified the original image as rank 1 in baseline
Modified image with synonym attack: A mortal in gray stands alone on a body structure outdoors in the darkness Identified wrong image as R1 and gave the original image R4

Original Caption: Some women are standing in front of a bus with buildings behind it . Correctly identified the original image as rank 1 in baseline
Modified image with synonym attack: Some adult female are standing in front of a omnibus with buildings behind it. Identified wrong image as R1 and gave the original image R2

Original Caption: A girl is on rollerskates talking on her cellphone standing in a parking lot. Correctly identified the original image as rank 1 in baseline
Modified image with synonym attack: A fille is on rollerskates talking on her cellphone standing in a parking heap. Identified wrong image as R1 and gave the original image R3

4 Conclusion

The results clearly depict the fall in recall values as the captions are changed which means as more changes are made to the data in the training distribution, the model is unable to retrieve the correct data. **Synonyms** effected the model the most. Drop in recall for paraphrasing is not significant and the difference kept reducing from Recall@1 to Recall@10. So the effect of paraphrasing is not as significant as synonyms. Hypernym and Hyponym performed almost similarly meaning the model is able to pay attention to both details and generalized features.