**Assignment 1: Report**

**Rohan Charudatt Salvi**

**Method:**

* I have used a pre-trained CLIP Model(Zero-shot Setting) for this assignment. The library used is transformers.
* Using the MSCOCO 2014 dataset I first segregated the data belonging to the TEST set.
* This was followed by encoding the images and captions using the CLIP model.
* For evaluation:
  + For Image to text retrieval, I used the encoding of an image and found its similarity across all caption encodings then picked top 10 sentence ids.
  + Then through a dictionary look if the most similar id is the truth value, if not then is one of the top 5, if not then in the remaining 5 is there one true value? This way I also calculated Recall@1, Recall@5, and Recall@10.
  + A similar method was used for text-to-image retrieval, just for every caption I found it’s similarity across all test images and used the same idea as above for Recall.

**Results**:

**For Text To Image Retrieval**

Recall 1 = 25.62

Recall 5 = 49.31

Recall 10 = 60.72

**For Image to Text Retrieval**

Recall 1 = 8.34

Recall 5 = 19.88

Recall 10 = 32.22

**Findings:**

One clear finding is the model is performing better on text-to-image retrieval as compared to image-to-text retrieval.

And on the error analysis I performed, I came up with two plausible reasons for it.

1. **With Image, the model is encoding the larger picture but missing smaller details thus it gets a poor recall score**

Example



Image Id : COCO\_val2014\_000000403385.jpg

For this image the top 5 captions retrieved were:

A small bathroom is pictured in this image.

A clean bathroom is pictured in this image.

A clean bathroom is seen in this image.

A small bathroom is featured in this photo.

THERE IS A TOILET IN THE BATHROOM AND A SHOWER IN THERE.

However, none of them is correct as in the dataset there is an important aspect that the tile is missing or bathroom is broken.

Likewise another example:



Image: COCO\_val2014\_000000384213.jpg

Top 5 cations retrieved:

A kitchen has a door, refrigerator, stove, sink, microwave, and dishwasher in it.

A kitchen has a refrigerator, stove, and sink.

The small kitchen has large cabinets and two stoves.

A microwave has been heavily used and is dirty in the kitchen.

The kitchen has a silver refrigerator, stove, and microwave.

Again here we can see the encoding captures kitchen aspect but in the embedding space it puts it closer to captions with details that are not in this image like refrigerator , microwave, door while it also misses other detail like windows which are in original caption.

Thus, in a zero-shot setting, the model is able to understand the overall image(context) but misses the finer details thus leading to a poor recall score.

1. **Certain captions align really(always high similarity irrespective of image) well when a model is unable to decipher the image and thus affect the recall score**

In our dataset, there are captions such as :

“There is no image here to provide a caption for” (has three entries with different sentence IDs) and “There is no image to provide a caption for”.

These two captions appear a quite a lot during retrieval which is reasonable considering if model is not able to understand or properly encode an image then these would be the one that will be most similar. Thus, due to this the model recall get’s affected a lot as these turn up in top 5 a lot and thus hamper Recall 1 and Recall 5 score.

Lastly, I would like to state the fact that since the model is still understanding the overall image hence during text to image retrieval they perform better. As there are not many images that fit the exact description while in case of captions there are multiple captions that in a way explain the image as we saw in the bathroom example above.

Possible Solution: I think finetuning the model on the dataset will vastly improve the recall scores as the model has shown it understand image and captions to an extend and fine tuning should give it that extra help needed to encode them even better.