1. Motivation

Context:
Known geometric properties in word embeddings (King is to Queen what Man is to Woman). Do multimodal embeddings also display such regularities?

Task:
Zero-shot Image Retrieval with multimodal queries. Given an input image and a text transformation query, find a “transformed image” in a database.

Contributions:
- We design a novel dataset for evaluating this task based on Visual Genome with queries focusing on subject-relation-object triplets.
- We use this dataset to assess geometric properties in multimodal embedding spaces.

2. Main Method

We fine-tune adaptation layers on COCO with infoNCE loss with different temperatures.

3. Qualitative Results

4. Evaluation

- We create a list of images annotated with subject-relation-object triplets from Visual Genome
- Each query asks to change one of these three elements
- An image-text matching algorithm, OSCAR [1], assesses whether or not the transformation is successful.

5. Analysis

- Vanilla CLIP embeddings [2] not well suited for delta-vector based transformation
- Best performance when fine-tuning with temperature $\tau = 0.1$
- Using geometric properties of pretrained sentence embeddings was not helpful

6. References

[1] Li et al., Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, ECCV 2020