

Embedding Arithmetic of Multimodal Queries for Image Retrieval Guillaume Couairon, Matthijs Douze, Matthieu Cord, Holger Schwenk

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.

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.





- Vanilla CLIP embeddings [2]
- Using geometric properties embeddings was not helpful