Embedding Arithmetic of Multimodal Queries for Image Retrieval

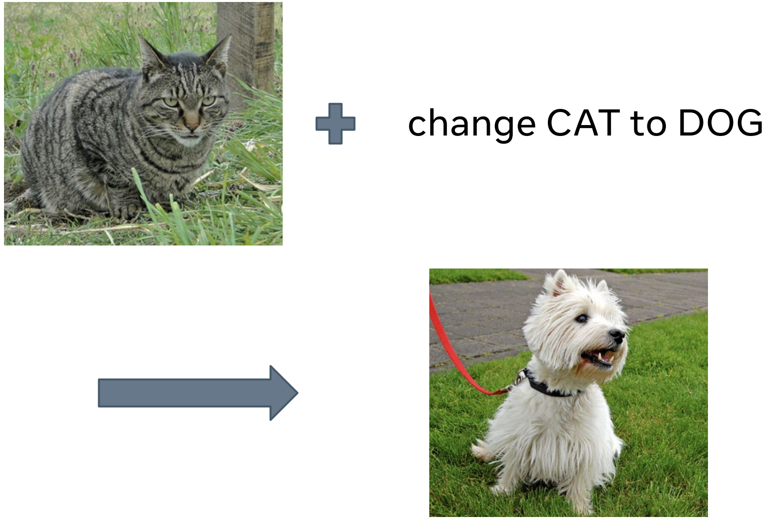
Guillaume Couairon, Matthieu Cord, Matthijs Douze, Holger Schwenk

O-DRUM 2022 CVPR Workshop

A task: Text-driven image transformation

change CAT to DOG

A task: Text-driven image transformation

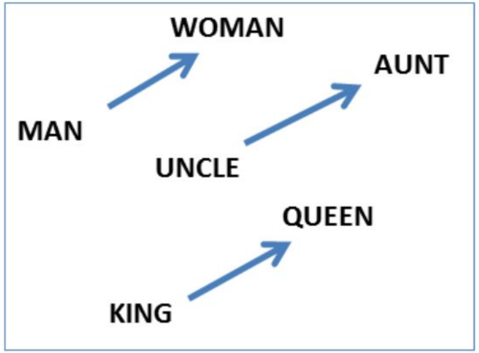


Contributions:

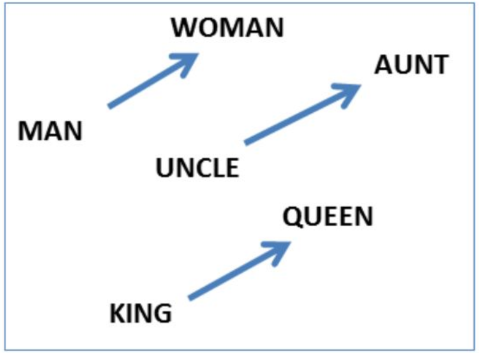
- Dataset and metrics to evaluate algorithms on this task

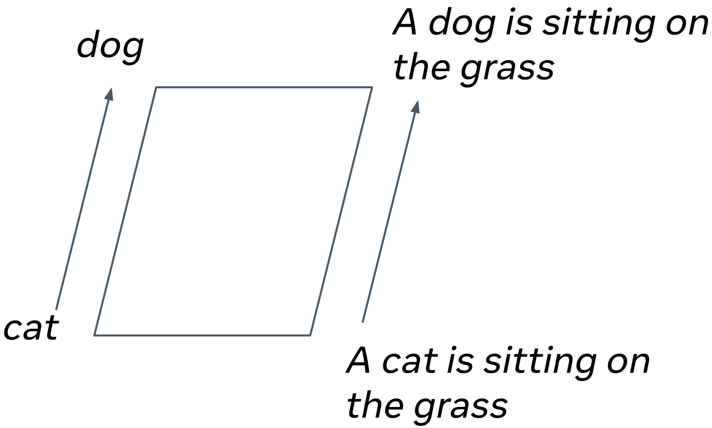
- We propose a simple zero-shot method and use it to assess geometric properties of multimodal embedding spaces

Motivation: Word and Sentence Embeddings



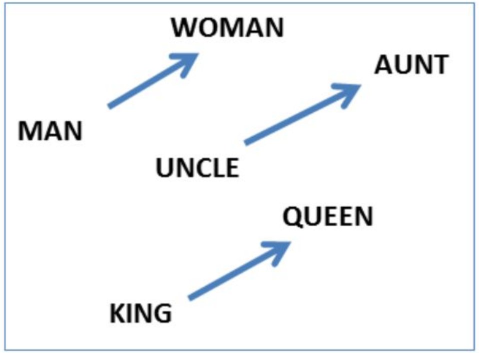
Motivation: Word and Sentence Embeddings

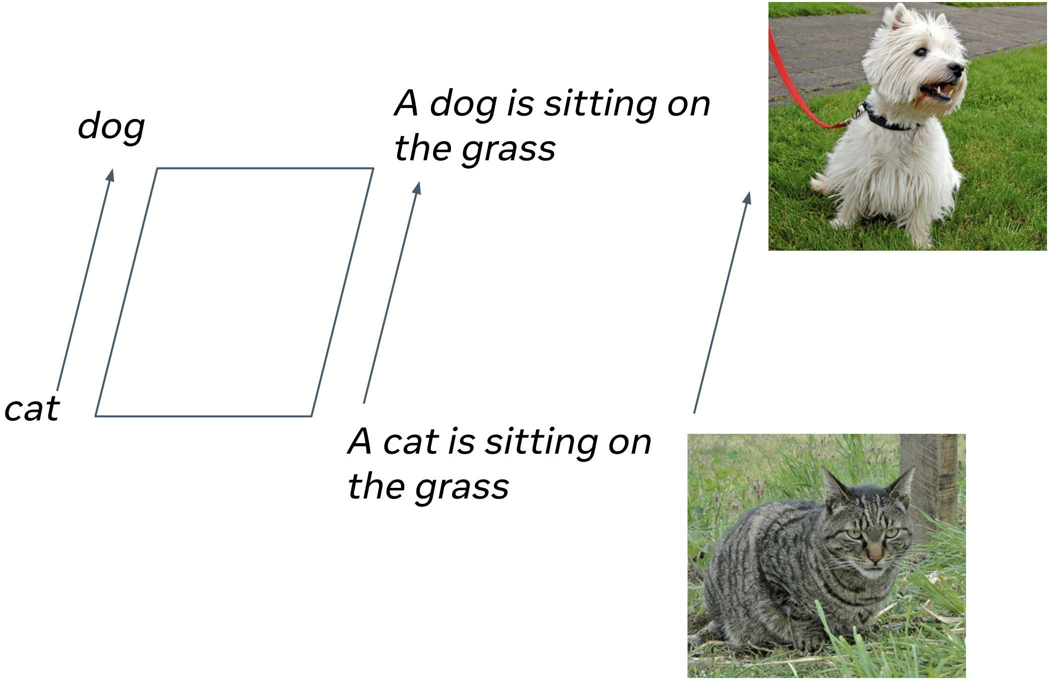




A dog is sitting on the grass ~ A cat is sitting on the grass + (dog - cat)

Motivation: Word and Sentence Embeddings



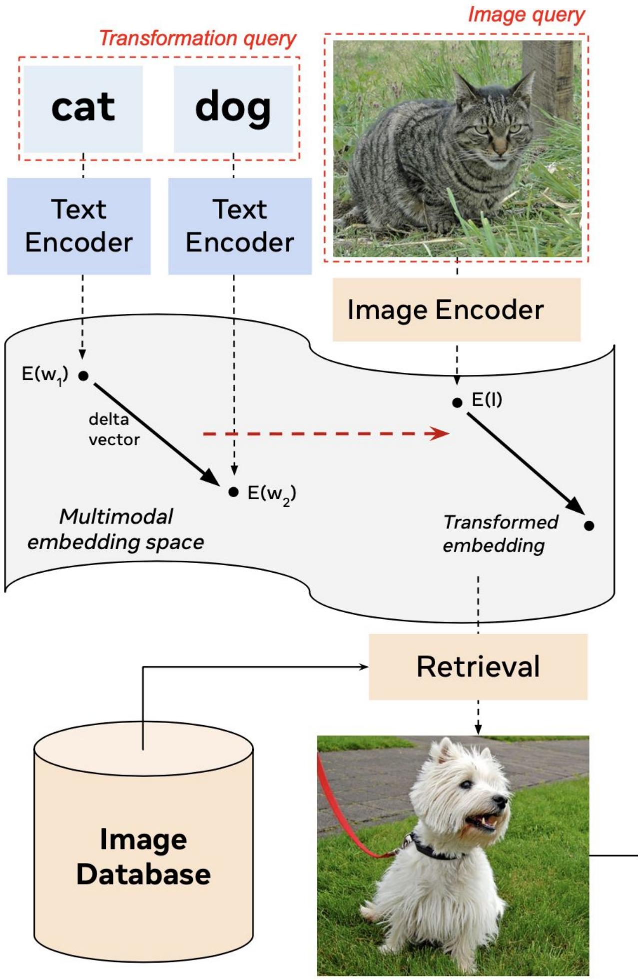


A dog is sitting on the grass ~ A cat is sitting on the grass + (dog - cat)

Method Overview

$$x = E_{img}(I) + \lambda \cdot (E_{txt}(w_2) - E_{txt}(w_1))$$

Lambda is the scaling factor



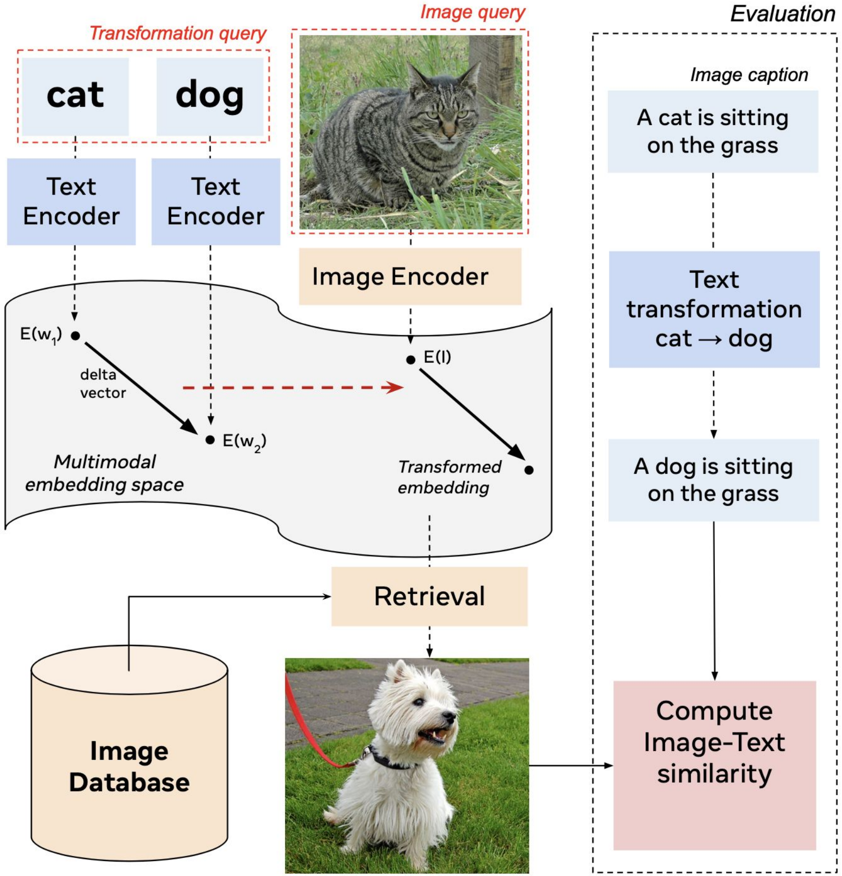
Evaluation (1)

- How to check if the transformation was successful ?
- How to check if the context has not been changed ?

- We use (subject, relation, object) annotations from the Visual Genome dataset.
- Transformation queries : Change Subject / Change Relation / Change Object.
- We ensure that each transformation query has a valid solution in the dataset

Evaluation (2)

- Compute Image-Text similarity with OSCAR [1]
- “SIMAT score” : accuracy of transformation success



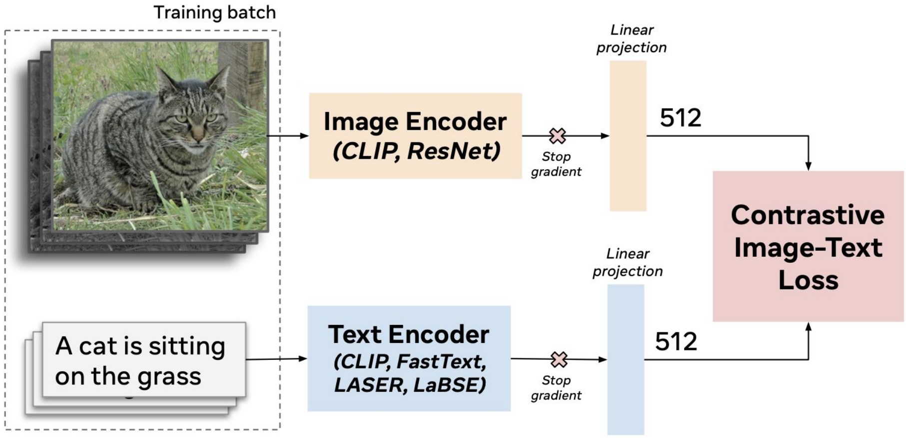
[1] Li et al., Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, ECCV 2020

Examples

Image Query						
Transformation Query	Woman → Man	Leaning on → Jumping over	Toilet → Suitcase	Kite → Rail	Boat → Bed	Tennis racket → Skateboard
Target Caption	A **man** balancing on a surfboard.	A horse **jumping over** a fence.	A cat sitting on a **suitcase**.	A man leaning on a **rail**.	A woman sitting in a **bed**.	A man playing with a **skateboard**.
Retrieved Image						
Success (OSCAR)	YES	YES	YES	NO	NO	NO

Fine-tuning

- Finetune on MSCOCO (500k text/image pairs)
- We study the importance of the temperature parameter



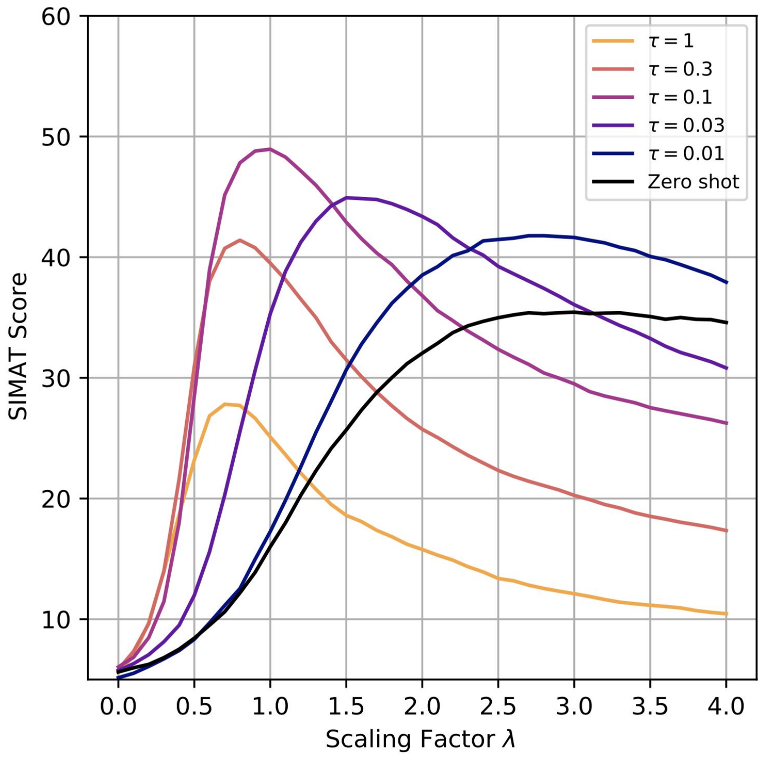
$$\mathcal{C}(I,T) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp(I_i \cdot T_i/\tau)}{\sum_{j=1}^{n}\exp(I_i \cdot T_j/\tau)}\right)$$

$$\mathcal{L} = \frac{1}{2}\mathcal{C}(I,T) + \frac{1}{2}\mathcal{C}(T,I) \quad (3)$$

Findings (1) : fine-tuning CLIP embeddings

- Vanilla CLIP embeddings not well suited for delta-vector based transformation

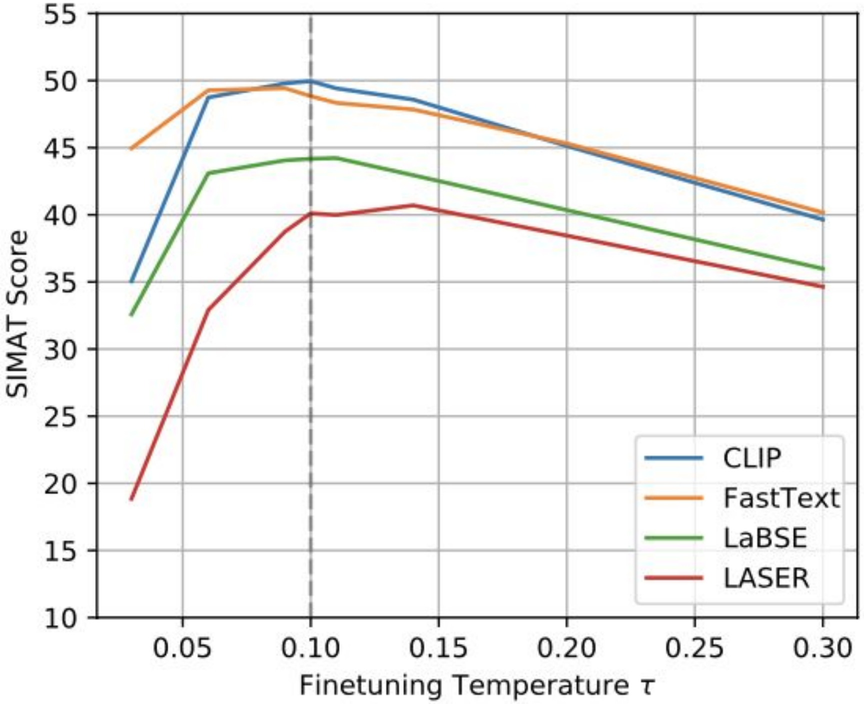
- Best performance when fine-tuning with temperature τ=0.1



Findings (2): leveraging properties of pretrained sentence encoders

- Best fine-tuning temperature does not depend on the text encoder

- Using geometric properties of pretrained sentence embeddings was not helpful



Embedding Arithmetic of Multimodal Queries for Image Retrieval

Guillaume Couairon, Matthieu Cord, Matthijs Douze, Holger Schwenk

O-DRUM 2022 CVPR Workshop