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

- change CAT to DOG

**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 \sim 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

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

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
<table>
<thead>
<tr>
<th>Image Query</th>
<th>Transformation Query</th>
<th>Target Caption</th>
<th>Retrieved Image</th>
<th>Success (OSCAR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woman → Man</td>
<td>Woman → Man</td>
<td>A man balancing on a surfboard.</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Leaning on → Jumping over</td>
<td>Leaning on → Jumping over</td>
<td>A horse jumping over a fence.</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Toilet → Suitcase</td>
<td>Toilet → Suitcase</td>
<td>A cat sitting on a suitcase.</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Kite → Rail</td>
<td>Kite → Rail</td>
<td>A man leaning on a rail.</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Boat → Bed</td>
<td>Boat → Bed</td>
<td>A woman sitting in a bed.</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Tennis racket → Skateboard</td>
<td>Tennis racket → Skateboard</td>
<td>A man playing with a skateboard.</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>
Fine-tuning

- Finetune on MSCOCO (500k text/image pairs)

- We study the importance of the temperature parameter

\[
\begin{align*}
C(I, T) &= - \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\exp(I_i \cdot T_i / \tau)}{\sum_{j=1}^{n} \exp(I_i \cdot T_j / \tau)} \right) \\
\mathcal{L} &= \frac{1}{2} C(I, T) + \frac{1}{2} C(T, I)
\end{align*}
\]
Findings (1) : fine-tuning CLIP embeddings

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

- Best performance when fine-tuning with temperature $\tau=0.1$
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