vitrivr-engine at the NTCIR-18 Lifelog-6 Task

Engaging Content
Engaging People

Taighde Éireann
Research Ireland

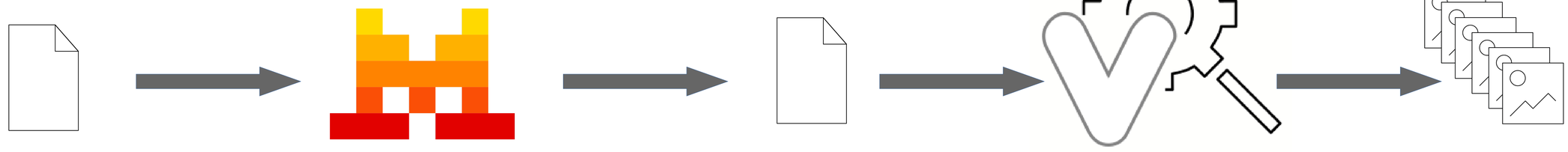
Luca Rossetto
School of Computing, Dublin City University

Challenge

- State-of-the-art text-based visual retrieval methods rely on visual-text co-embedding models such as CLIP
- These models are generally trained using large collection of image-caption pairs
- Image captions generally describe a scene from a third-person perspective
- Memories captured by ego-centrig lifelog images are generally described from a first-person perspective
- This perspective mismatch limits model applicablilty

Approach

- ‘shift perspective’ from an ego-centric scene description to an exo-centric scene description
- Use LLM to rewrite input query to me more aligned with typical image caption format CLIP models have been trained on
- Transformed query is directly used in downstream retrieval process without any additional



Results

Total Relevant Items	1995
Relevant Items Retrieved	0.208
Mean Precision @ 5	1385
Mean Recall @ 100	0.2203
Mean nDCG	0.1151

Insights

- LLMs can be used for zero-shot perspective shift to some degree
- Quality and fidelity of transformed image descriptions is inconsistent
- Transformed queries can be used with off-the-shelve visual-text co-embedders
- Performance is not competitive with purpose-built models or systems using additional result filtering
- Investigation on VLMs as superior option for perspective shift left for future work

DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University