mSUsIE: multimodal Search Using Image Encoders

Advanced Information Retrieval, Group 15
Benedikt Kantz: Model Search, Dataset preparation
Corinna Kindlhofer: Evaluation, Metrics
https://github.com/coki1405/mSUsIE

Introduction & Motivation

- Private database of ~10k Cliparts of various styles and contents, bad naming
- Hard to search through manually (deep folder structure)
- Our goal: make it searchable using multimodal search (image search + text embeddings of metadata)
- Should be fast and precise (not necessarily with a good recall)

Data & Methods

- Custom dataset of Cliparts
- Manually created ground truth of:
  - 36 queries
  - ~300 responses
- Methods:
  - Storing embeddings in Milvus (vector database)
  - Generate image embeddings using CLIP-ViT or CLIP-ResNet, textual using multilingual sentence-transformers
  - Text search based on generated description (TF-IDF+BLIP)
  - Baseline: TF-IDF on metadata (file path)

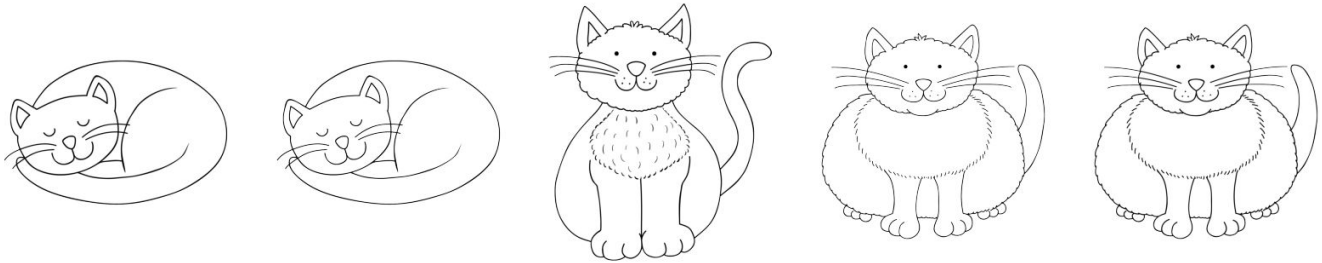
Results

- Quite good results in English Baseline
- BLIP does help a bit
- Vector search levels German and English results
- Query response times < 1 second
- Also implemented simple web interface

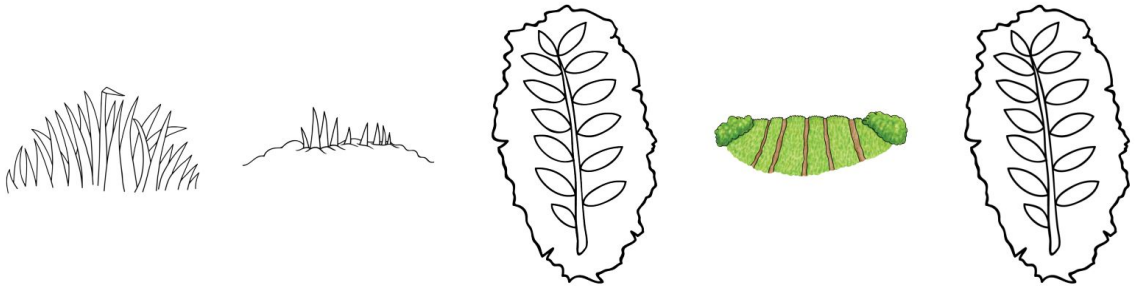
Precision	English	German
TF-IDF	0.3869	0.0278
TF-IDF+BLIP	**0.4692**	0.0278
CLIP-ViT+sentence-transformers	0.3160	**0.1326**
CLIP-ResNet	0.2871	0.0426

Examples (Textual queries)

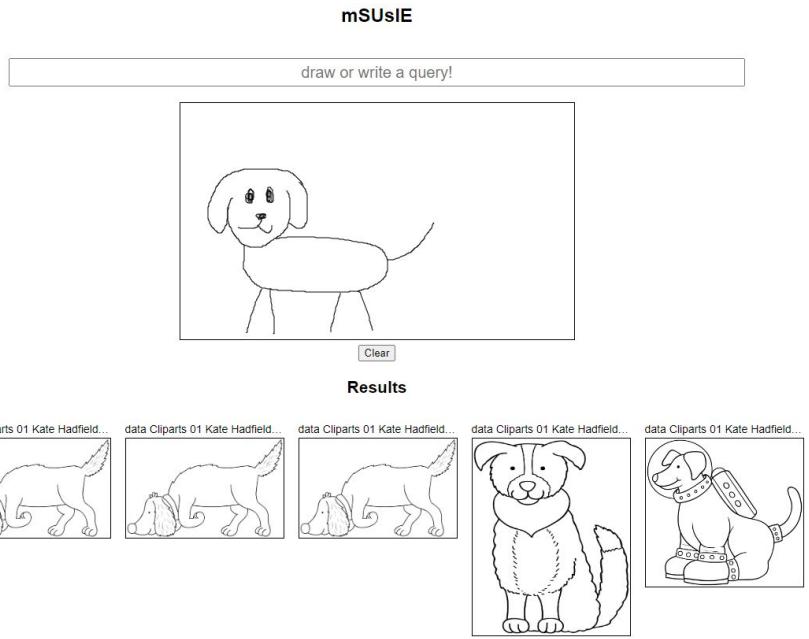
Query: “cat”

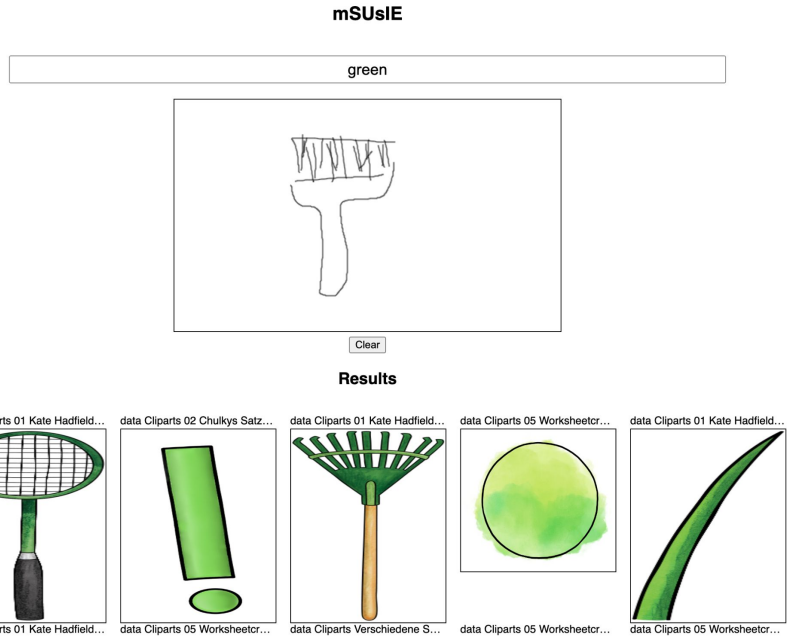


Query: “Natur”



Example (Visual query)





Conclusions

- Successfully implemented a retrieval system for clipart images
  - Including simple frontend
- Good usability: Querying is fast
- Quality of retrieved results is quite good (especially for English queries)
  - But: Depends heavily on the underlying ground truth
- User study (N = 1): Participant was happy with results
  - Relevant images were retrieved