Improving Global Representations with Captions for Remote Sensing Image Retrieval

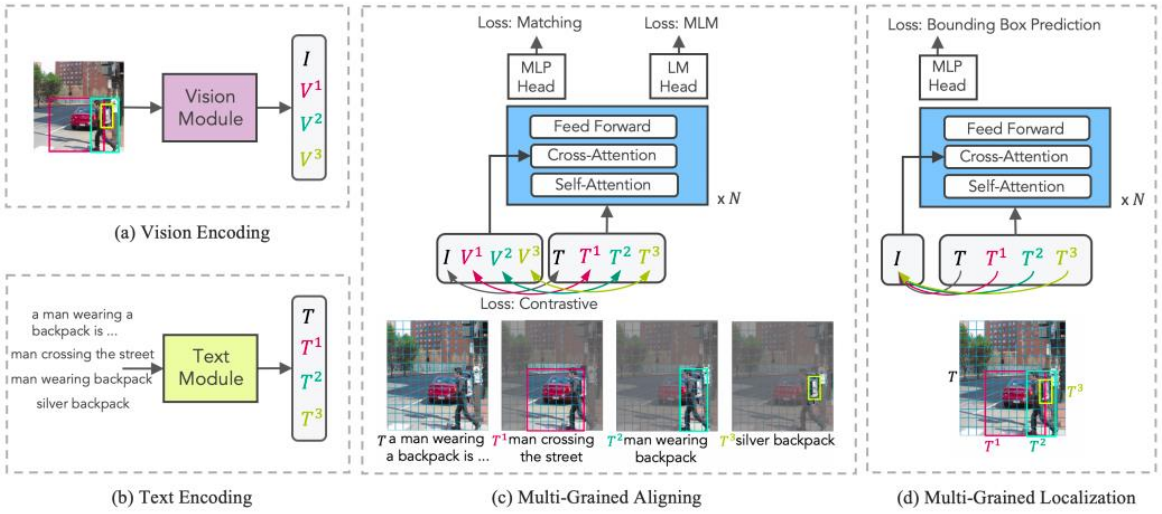
Midterm Project Presentation – Filippo Momentè (T3)

Language and Image Retrieval

- Language is fundamental for Image Retrieval when the query is textual
  - Text-to-image Retrieval
  - Composed Image Retrieval
- Often, pre-trained models are leveraged for this task

Language and Image Retrieval

- An example
  - X-VLM2 (Zeng et al., 2023)
  - Pre-trained for many vision-language tasks
  - Image representations fused with object-level and textual information
  - SoTA in Text-Image Retrieval (fine-tuned)



Zeng et al., 2023

An opportunity for Image-Image Retrieval

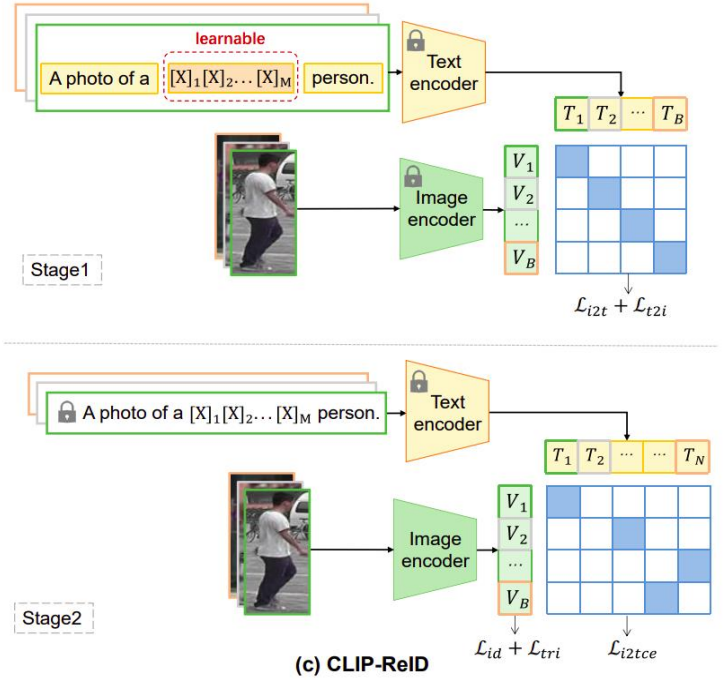
- Pre-training with textual and object-level information useful for Text-Image Retrieval
- Can we use these info for Image-Image Retrieval?
- Textual information limited in current approaches

An opportunity for Image-Image Retrieval

- Pre-training with textual and object-level information useful for Text-Image Retrieval
- Can we use these info for Image-Image Retrieval?
- Use of textual information limited in current approaches

Using Language to improve Image-Image Retrieval

- CLIP-ReID: Exploiting Vision-Language Model for Image Re-identifcation without Concrete Text Labels (Li et al., 2023)
- Used a VL model for Person ReID
  - Labels embedded into a prompt to improve learning
- Outperformed competitors on several benchmarks



An opportunity for Image-Image Retrieval

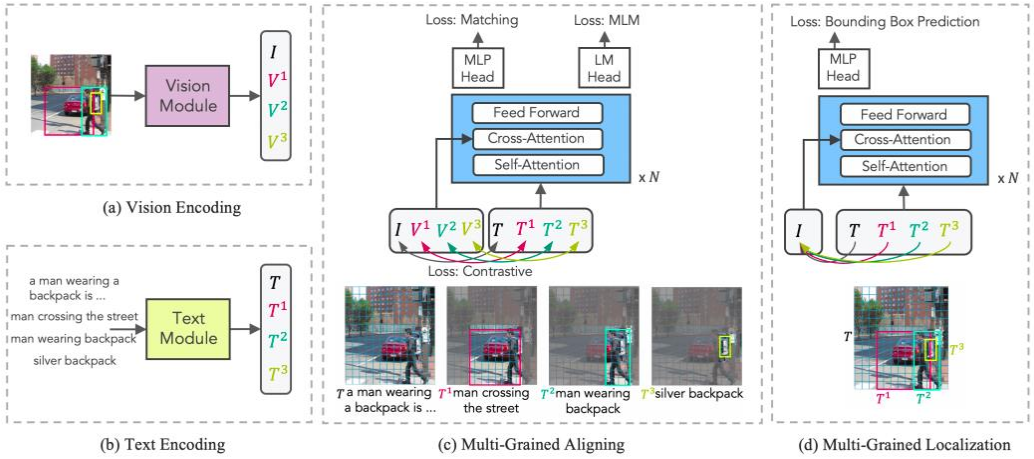
- In our scenario, captions present an interesting source of information
  - Represent overall semantic information of an image
  - Global representations do the same
  - Captions can help identifying the objects, as well as their relationships in a more straightforward way

An opportunity for Image-Image Retrieval

- Hypothesis: using captions during training can aid image-image retrieval
  - Using both may help learning better global representations for the task
- Attention: captions need to be carefully constructed

Improving Global Features

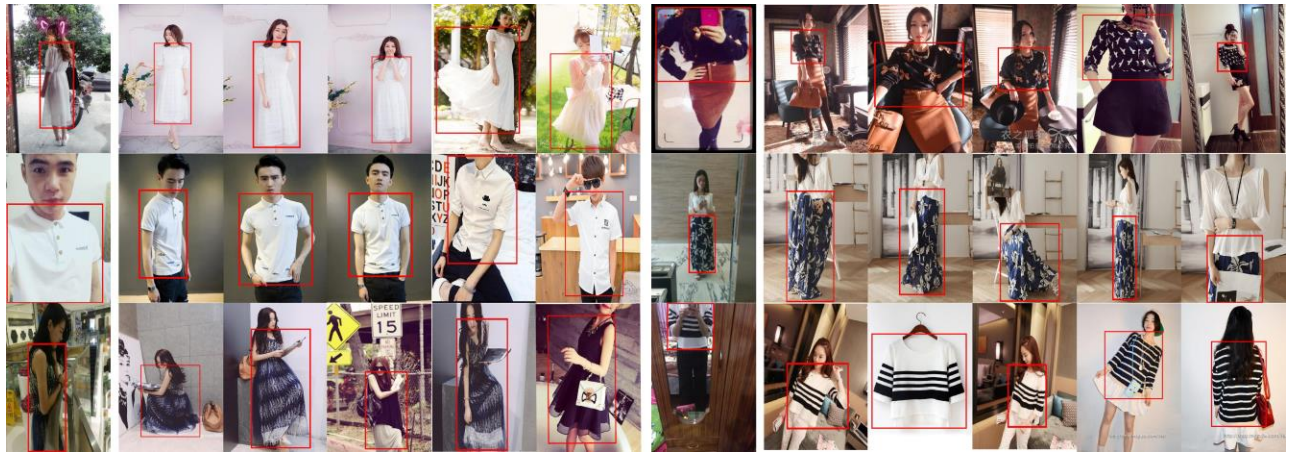
- Objective: see whether fusing info from carefully constructed captions with visual features can improve global representations
- Additionally, we can jointly train the model to detect the query object in order to improve the representation
- Idea: Let's adapt X-VLM2 for Image-Image Retrieval



Zeng et al., 2023

Dataset

- DeepFashion2 (Ge et al., 2019)
- Given a user-uploaded picture, find the correspondent commercial ones
- Challenging dataset
  - Different levels of occlusion, viewpoint change, scale
- It contains object-level annotations



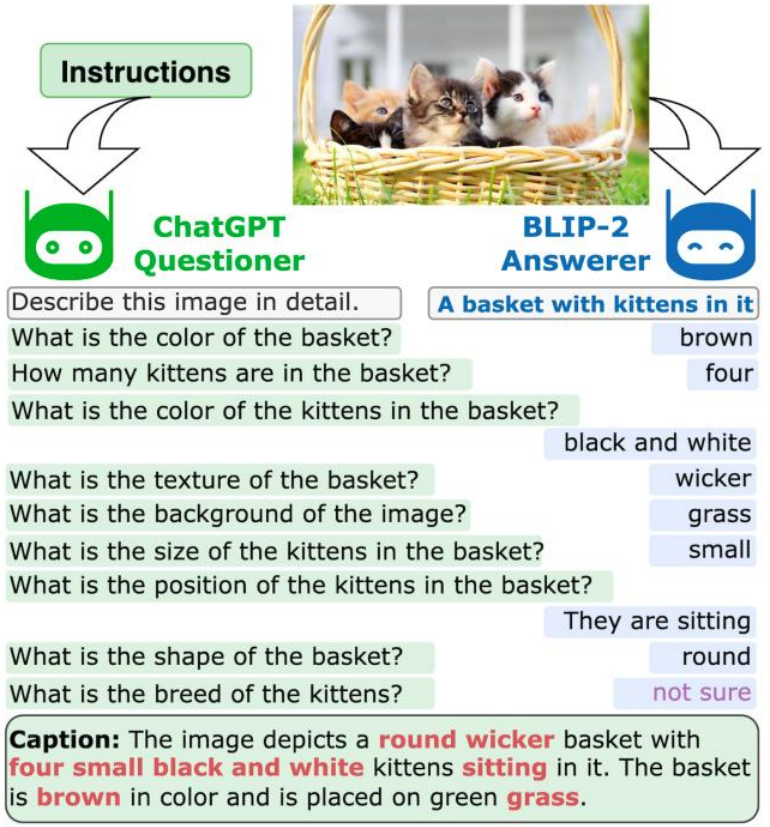
Ge et al., 2019

Improving Global Features

- Fine-tune X-VLM2
- Our dataset is annotated at object-level
- We need captions for our images
- How do we collect them?

Obtaining captions

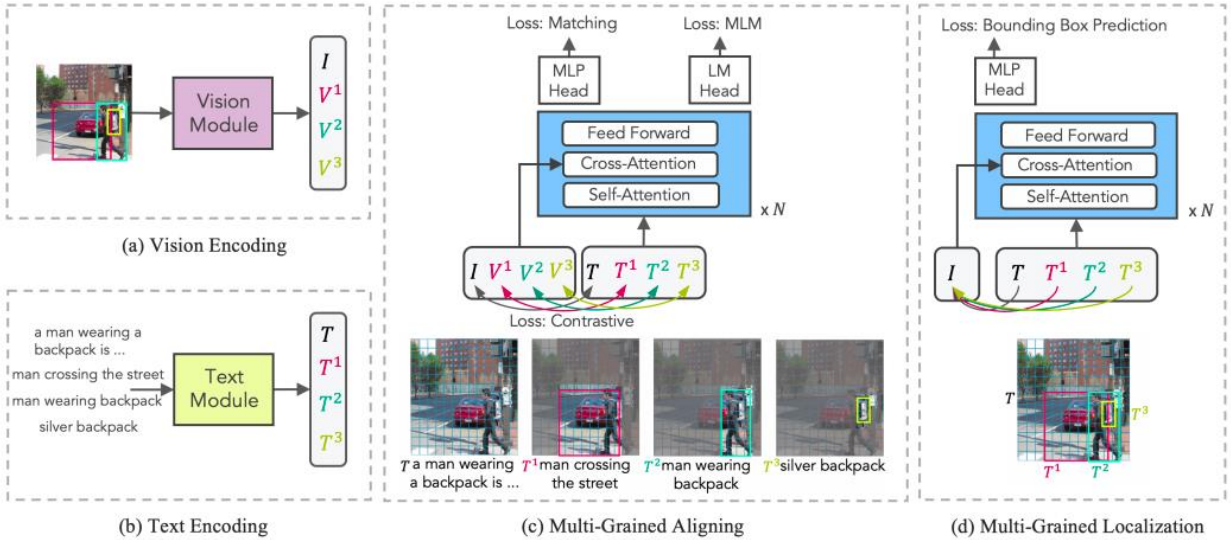
- Let's use a pre-trained VL model
  - LLaVa (Liu et al. 2023)
  - Specialized in VQA
- We can leverage an LLM in order to get more detailed captions
  - The LLM interacts with the VL model to obtain a refined caption
  - Zhu et al. 2023
  - BLIP-2 substituted with LlaVa
  - ChatGPT with an open-source model (e.g. LLama-3)
- Need to work on the prompt



Zhu et al., 2023

Jointly learning Image Retrieval with Object and Caption information

- X-VLM2 loss: Lbbox + Litc + Lmatch + Lmlm
- Modular architecture
- Our fine-tuning task: Lbbox + Litc + ArcFace



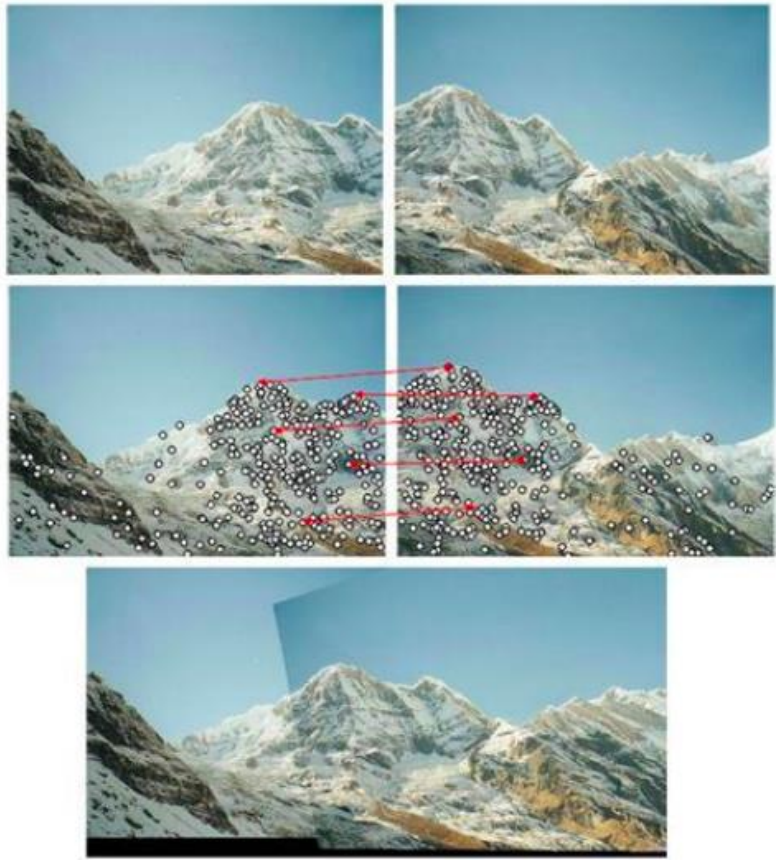
Zeng et al., 2023

Characteristics of the Global Representation

- Output: a global representation
  - Constucted with textual supervision
  - Guided by object detection
  - Image and Text information fused together

Reranking

- Geometric Verification



At Inference Time

- Input: a query image
  - Captions will be constructed behind the scene

Evaluation

- Metrics
  - Mean Average Precision (mAP)
- Compare against various Image Retrieval architectures
  - E.g. DELG (Cao, 2020), SuperGlobal (2023),..

Limitations

- Bad captions may worsen the model's performances
  - In a real scenario, ground-truth captions would be very important
- Captions are likely to slow down the retrieval
  - Important to quantify the delay caused by this procedure

Recap

- Fine-tuning a pre-trained VL model for building better global representations
  - Contrastive learning with captions
  - Jointly trained for object detection
- Reranking without Geometric Verification
- Trained and evaluated on Deepfashion2

Schedule

- Week 11: Basic setup
- Week 12: Adaptation of X-VLM2 to the task
- Week 13: Building the caption generation module and attaching it to X-VLM2
- Week 14: Performing the experiments
- Week 15: Wrap up and presentation

Thank you!

Midterm Project Presentation – Filippo Momentè (T3)