Improving Global Representations with Captions and Object-level Information

Final Project Presentation – Filippo Momentè (T3)

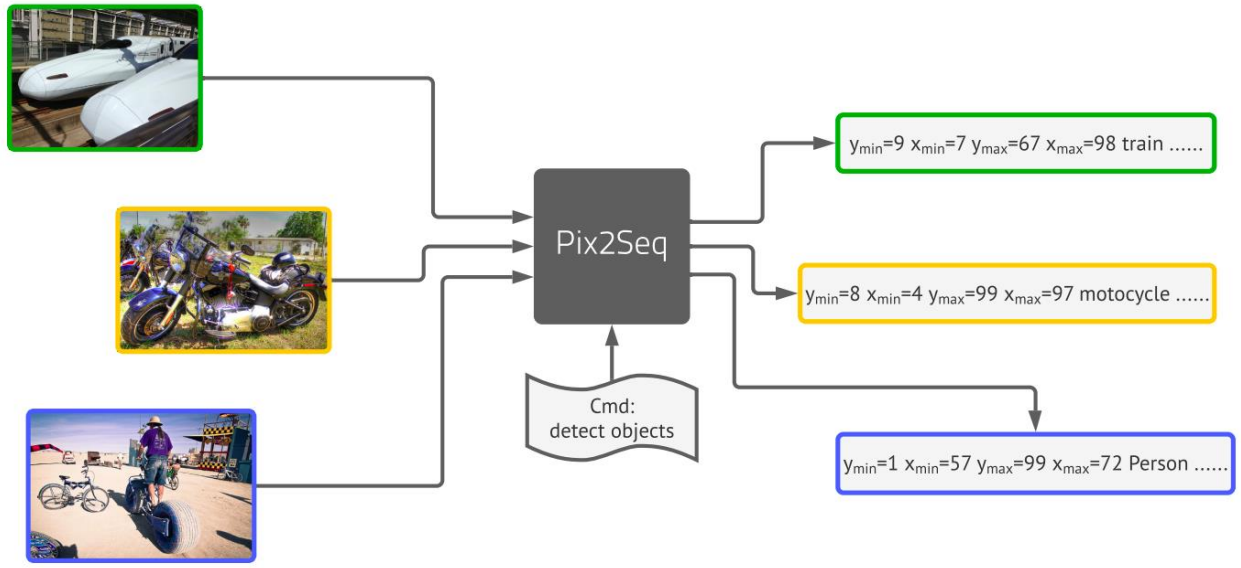
Introduction

Background Recap

- Main assumption: language can be beneficial to classical Computer Vision tasks
- It can reduce the complexity of models while retaining competitive performances

Background Recap

- Relevant example: Pix2Seq (Chen et al., 2022)
- Teaching a model to write the location of objects
- It enables efficient object detection
- Can we use language information to improve Image Retrieval?



Credits: Chen et al., 2022

Idea

- Let's use captions to improve global representations for content-based image retrieval
- Similarly to typical global features, they encode information about the general meaning of an image
  - We can modulate the granularity of the description by enriching it with details
- Fuse together text embeddings and image embeddings and train jointly

Idea

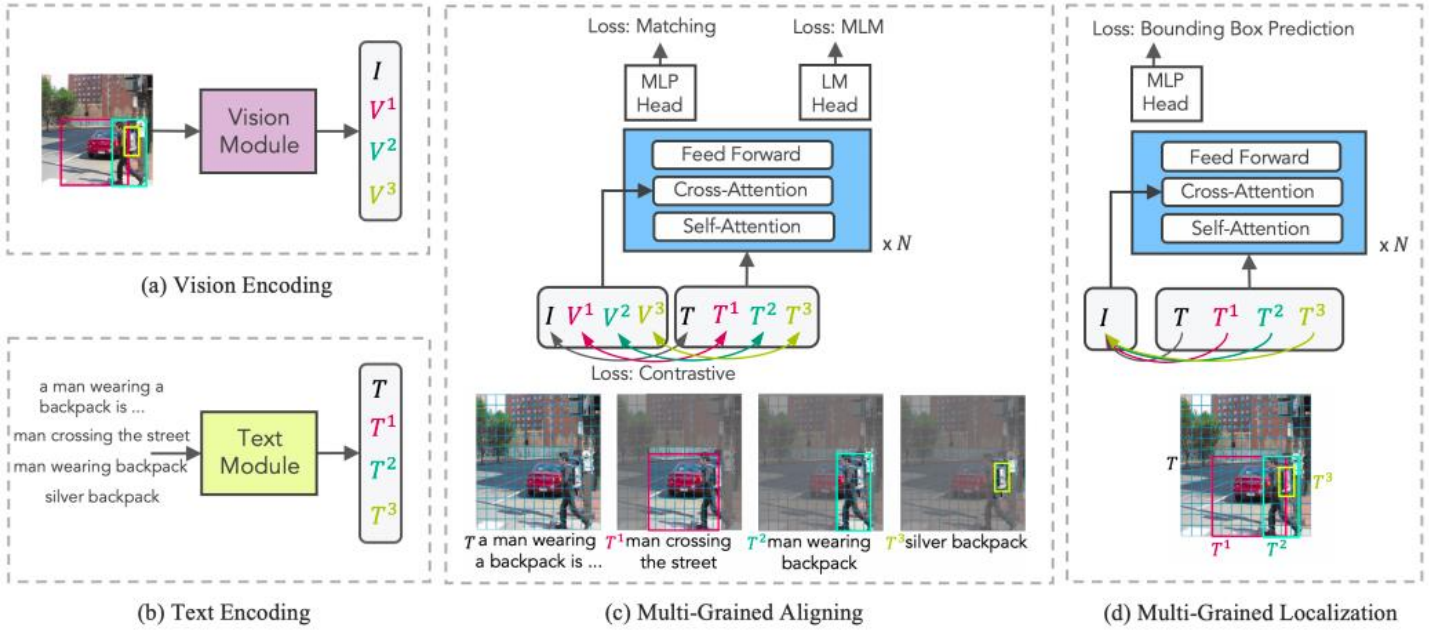
- This is typical when we have Image-to-Text or Text-to-Image retrieval
- Not when our queries and our database are both composed by images
- In this case, captions helpful to build a better representation space

Another idea

- Additionally, we can jointly train our model to detect relevant objects
- Assumption: learning to detect objects can improve retrieval accuracy
- Already attempted

A backbone

- 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)



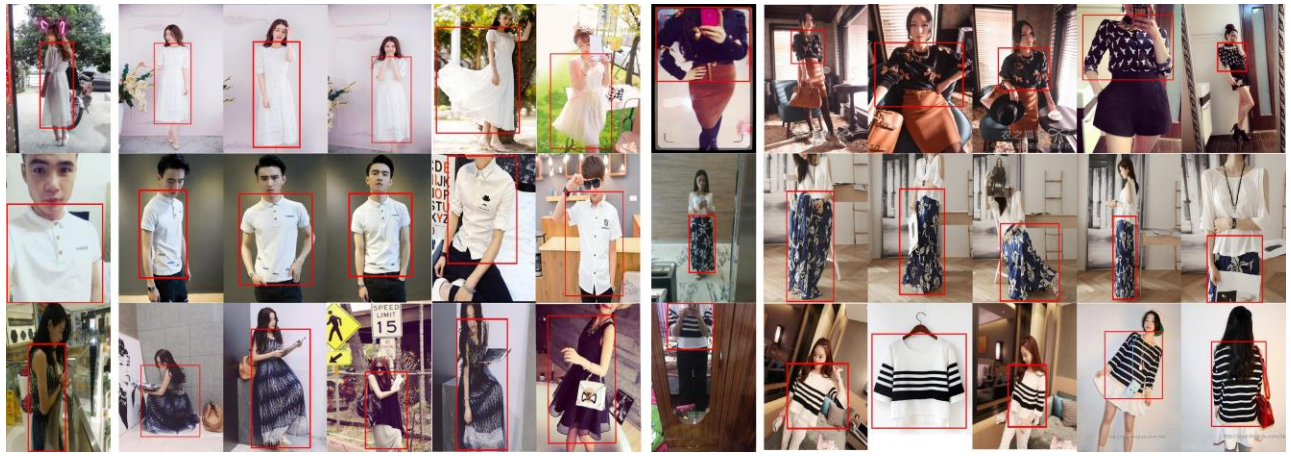
Zeng et al., 2023

X2-VLM

- Perfect for this idea
- Already pre-trained with language and object-level info
  - But not fine-tuned or tested on Content-Based Image Retrieval

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

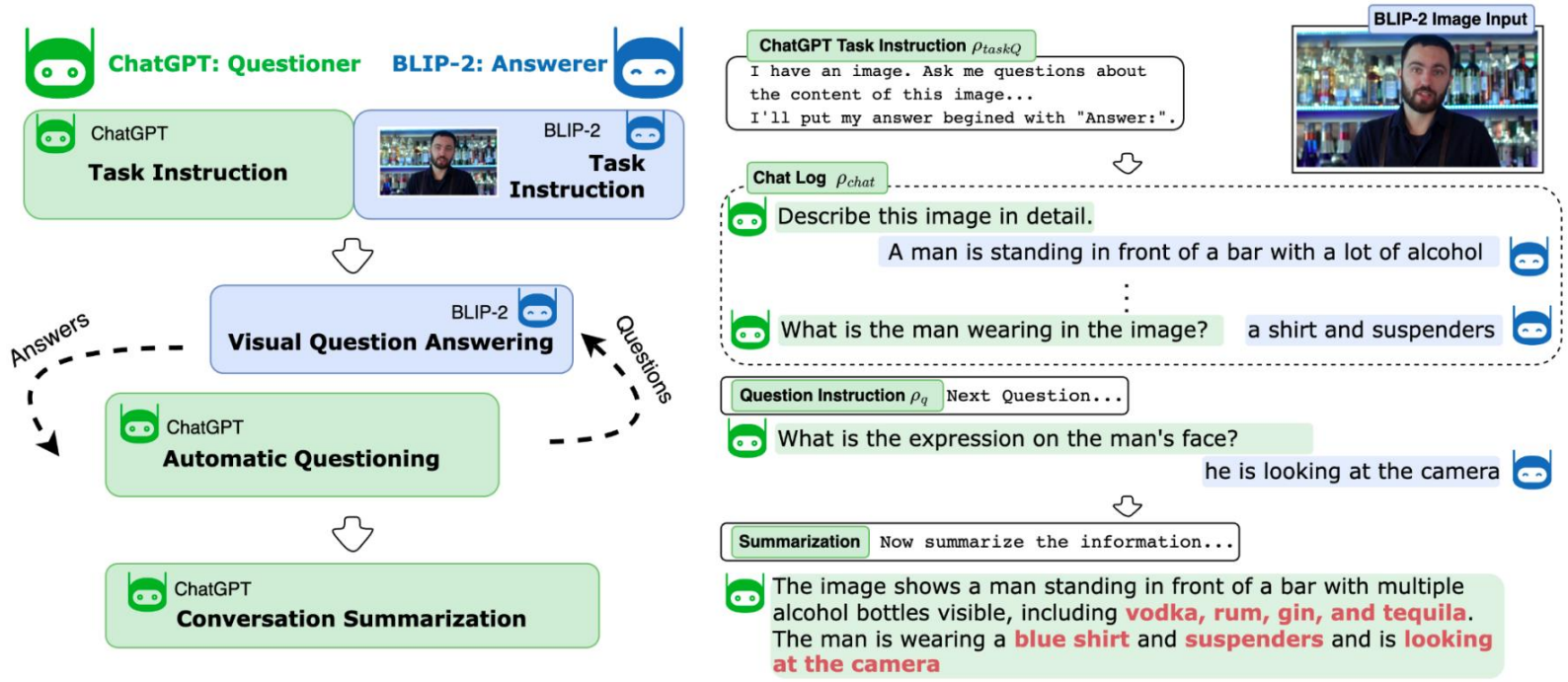


Ge et al., 2019

Methodology

Collecting Captions

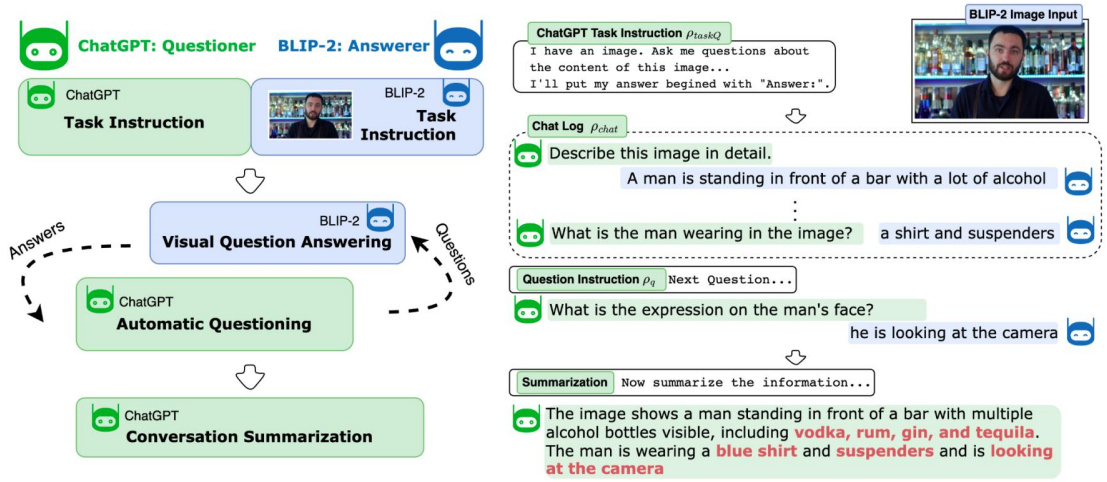
- Expensive
- They need to be highly detailed to be relevant
- Solution:



Zhu et al., 2023

Collecting Captions - 2

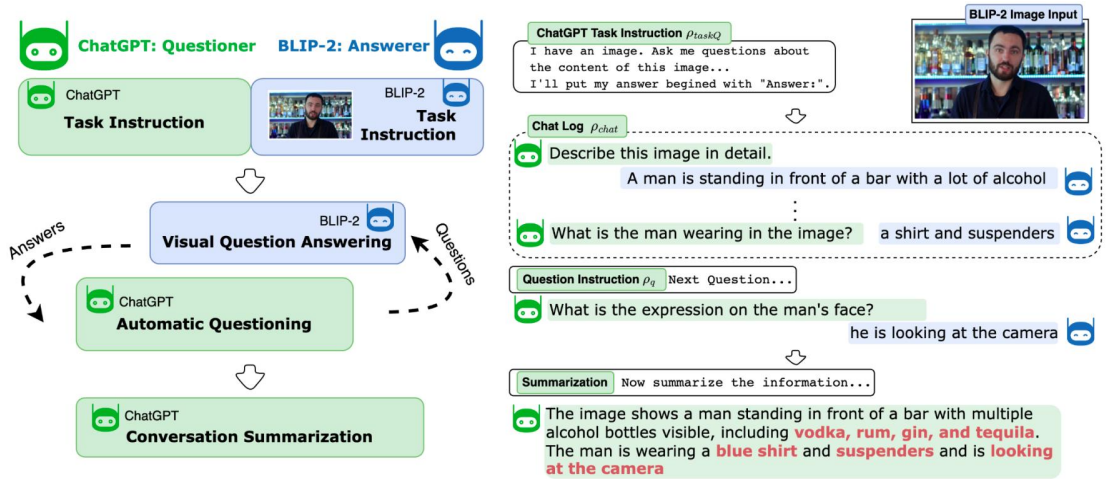
- An LLM and a VQA model interact
- The LLM asks questions
- VQA answers
- At the end, summarized caption is more detailed than the original caption



Zhu et al., 2023

Collecting Captions - 3

- Original work: GPT-3
  - Highly expensive
- Idea: let's try with Phi-3 (Abdin et al., 2024)
- X2-VLM for VQA



Zhu et al., 2023

Collecting Captions - Limitations

- My implementation doesn't do this
- Still too expensive
- Caption format: An image of a dress with viewpoint {}, occlusion level {}, scale {}, and zoom in {}

Finetuning X2-VLM

- Pre-training losses: MLM + ITC + ITM + (L1 + IoU)
- Fine-tuning losses: ITC + ITM + (L1 + IoU) + ArcFace
- Arcface for classification
  - Margin loss: you add a margin to logits and scale
  - Margin: 0.15, Scaling: 30.0

Inference

- ❑ Input: query images about clothes from consumers
- ❑ Need to find the corresponding images from the shop

Experiments

Overview of the Experiments

- Performed
  - Zero-shot Image Retrieval
  - X2-VLM already has been trained jointly with language and object info
  - Fine-tuned with object information and captions (not generated automatically)
- Planned
  - Fine-tuning with synthetic captions from Deepfashion2 and more samples

Zero-shot IR

- Zero-shot results really not satisying
  - MAP very close to zero (2.7%)
  - It is possible that mistakes in the implementation were made
    - A Transformer as backbone may also be the issue

Finetuning

- Fine-tuned with captions and object-level information
  - 1 epoch with 1000 samples
- Small improvement, but still mAP close to 0 (3.4%)
- The errors made however appears interesting

Error Analysis

- Interestingly, it seems like the type of clothes are somewhat identified, however not the exact one from the query

Query

Error Analysis

Query

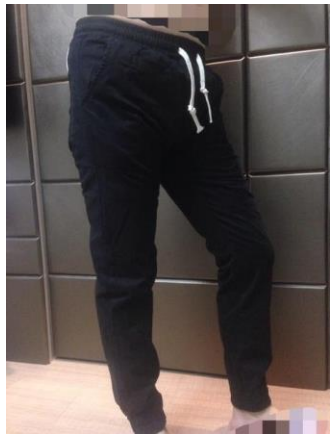
Error Analysis

Query

Error Analysis

Query

Error Analysis



Query

Conclusion

Observations and Limitations

- The network didn't really catch the point
- However, it is somehow reasoning
- Object detection and captions would likely help
  - Model doesn't get exactly what it should be looking for in the picture
- Only global features were considered
  - This likely shows in the results

Conclusion

- Objective: seeing whether textual information can be used to improve Content Based Image Retrieval
- Additionally include Object-level information
- Zero-shot + limited fine-tuning attempted
- Quantitative performances very low but qualitative ones suggests the model is somewhat reasoning

Thank you!

Final Project Presentation – Filippo Momentè (T3)