Pre-training End-to-End Vision-Language Transformers

Zi-Yi Dou
Advisor: Nanyun (Violet) Peng

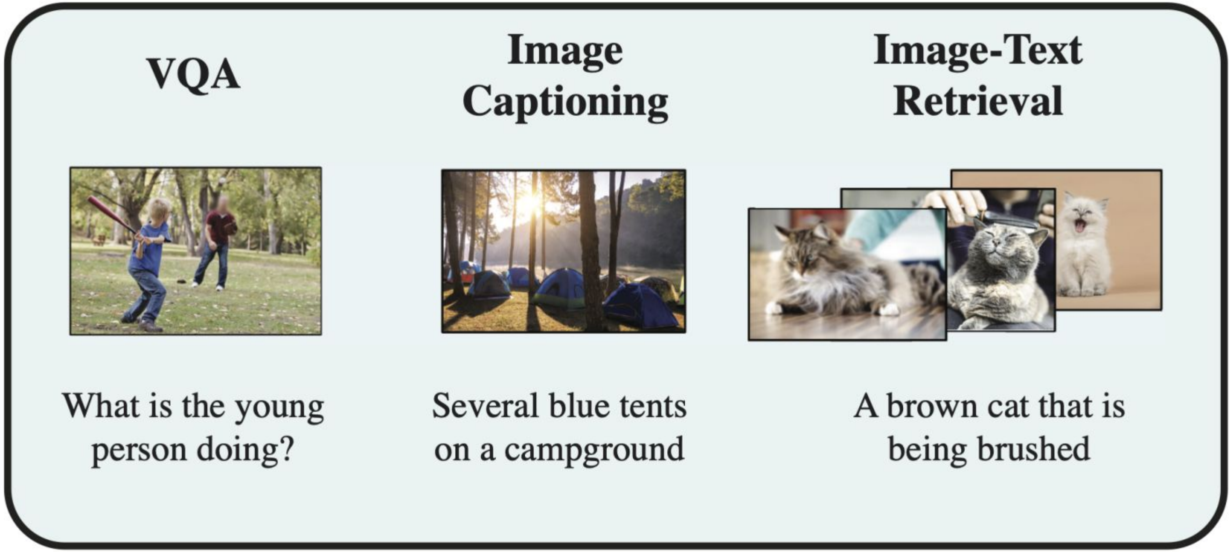
UCLA Samueli
Computer Science

Plus lab

Vision-Language Models

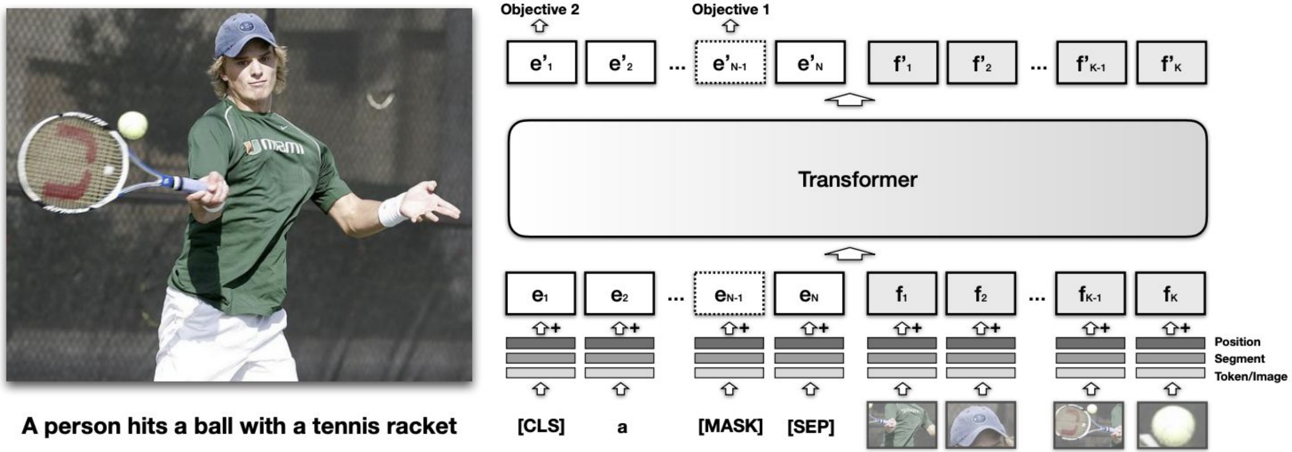
Vision-language models combine information from both visual and textual modalities to perform various tasks.

Pre-training models on large image-text corpora is highly effective.


VQA
Image Captioning
Image-Text Retrieval
What is the young person doing?
Several blue tents on a campground
A brown cat that is being brushed

Previous Work

Most previous methods rely on CNN-based object detectors.



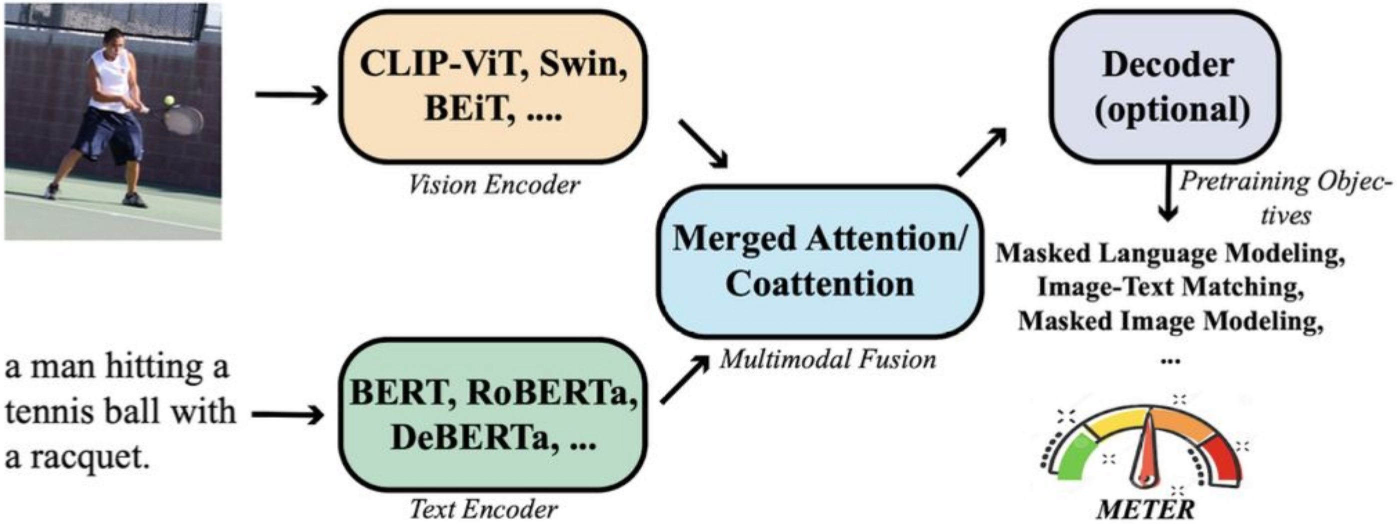
[Li et al., 2019]

Transformers have shown promising performance in both NLP and CV:

- Potential of having a unified architecture for vision and language.
- End-to-end training of both cross-modal and uni-modal modules during pre-training.

Empirical Studies of Training Vision-Language Transformers

We dissect the model designs along multiple dimensions and perform investigations on each of the modules:



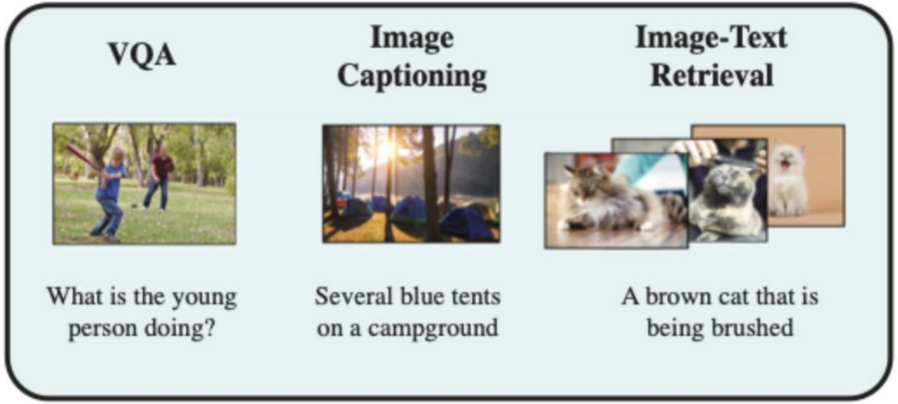
CLIP-ViT, Swin, BEiT, ...
Vision Encoder
a man hitting a tennis ball with a racquet.
BERT, RoBERTa, DeBERTa, ...
Text Encoder
Merged Attention/ Coattention
Multimodal Fusion
Decoder (optional)
Pretraining Objectives
Masked Language Modeling,
Image-Text Matching,
Masked Image Modeling,
...
METER

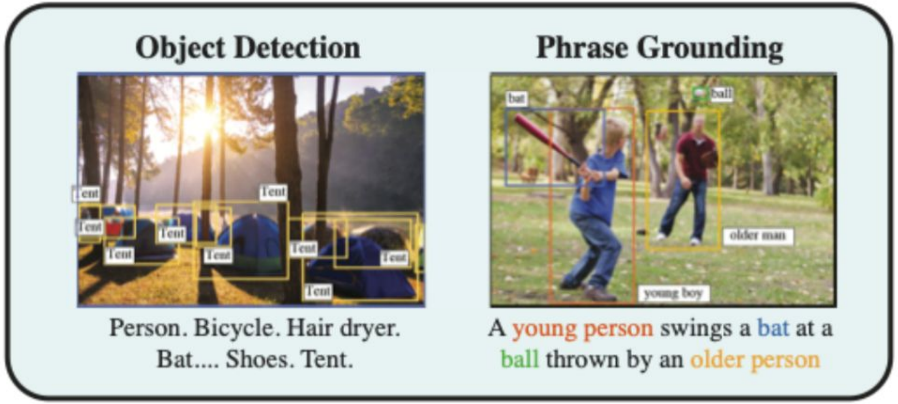
[Dou et al., CVPR 2022]

Region-Level Vision-Language Tasks

In addition to image-level tasks such as VQA, there are also region-level tasks like object detection and phrase grounding.

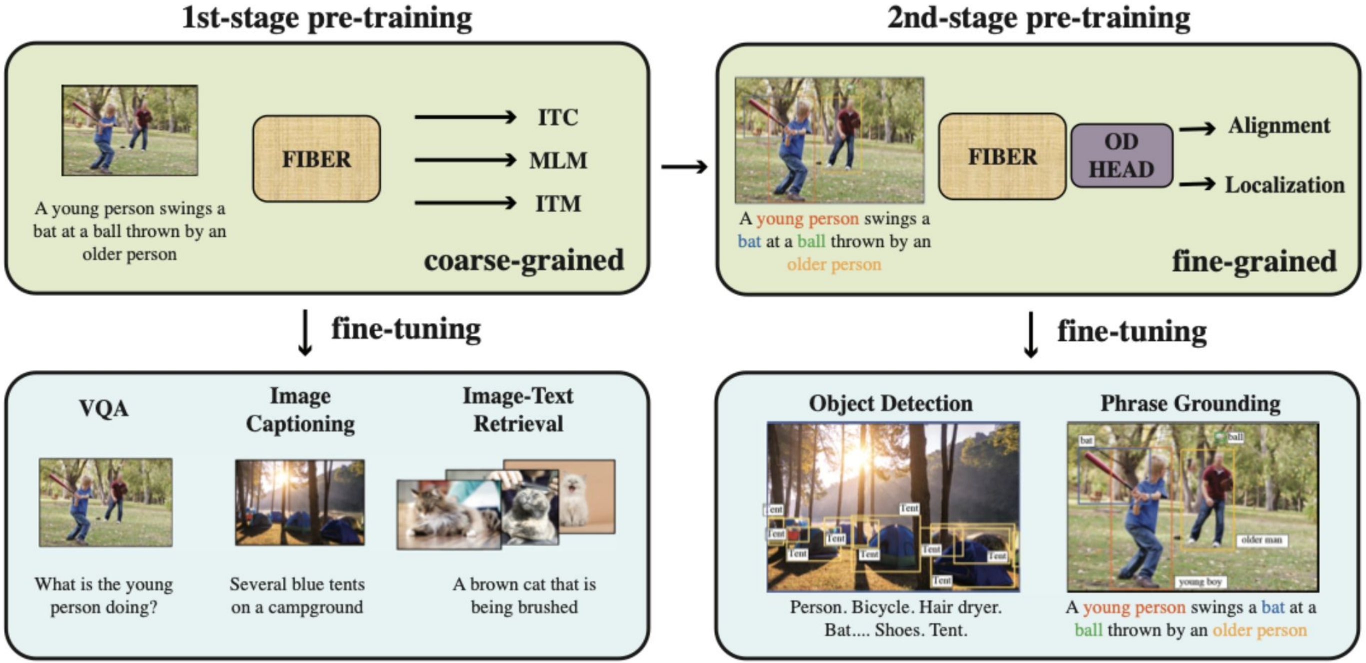
Collecting fine-grained annotations for region-level tasks is costly and non-scalable.


VQA
Image Captioning
Image-Text Retrieval
What is the young person doing?
Several blue tents on a campground
A brown cat that is being brushed


Object Detection
Phrase Grounding
Tent
Tent
Tent
Tent
Tent
Tent
Tent
Tent
bat
ball
older man
young boy
Person. Bicycle. Hair dryer. Bat.... Shoes. Tent.
A young person swings a bat at a ball thrown by an older person

Coarse-to-Fine Vision-Language Pre-training

We propose a coarse-to-fine pre-training paradigm that can support both image-level and region-level vision-language tasks.



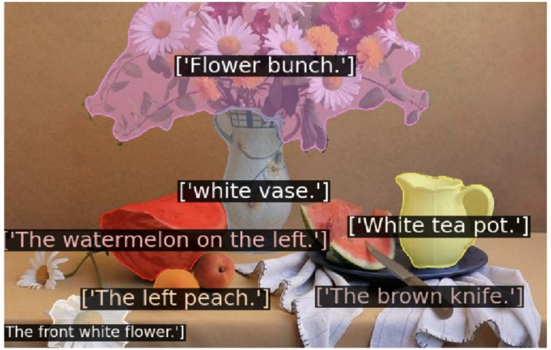
1st-stage pre-training
A young person swings a bat at a ball thrown by an older person
FIBER
ITC
MLM
ITM
coarse-grained
2nd-stage pre-training
A young person swings a bat at a ball thrown by an older person
FIBER
OD HEAD
Alignment
Localization
fine-grained
fine-tuning
fine-tuning
VQA
What is the young person doing?
Image Captioning
Several blue tents on a campground
Image-Text Retrieval
A brown cat that is being brushed
Object Detection
Tent
Person. Bicycle. Hair dryer. Bat.... Shoes. Tent.
Phrase Grounding
bat
ball
older man
young boy
A young person swings a bat at a ball thrown by an older person

[Dou & Kamath & Gan et al., NeurIPS 2022]

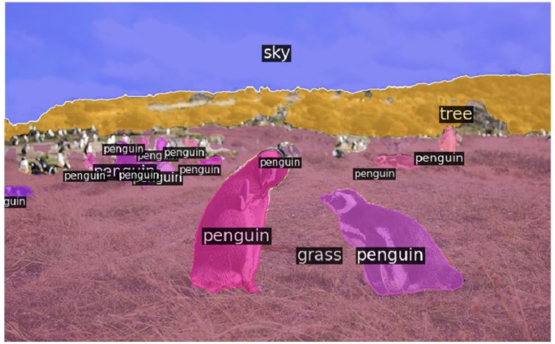
Pixel-Level Vision-Language Tasks

Tasks such as image segmentation require pixel-level outputs.

It is non-trivial to build models that support both traditional vision-language and segmentation tasks.

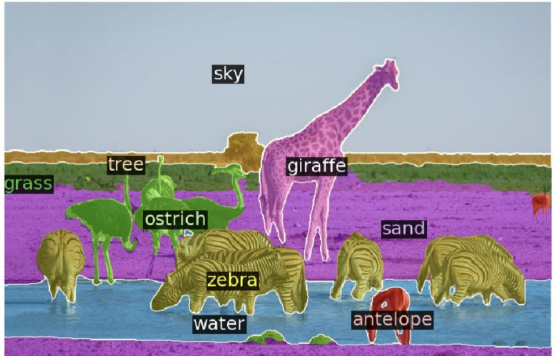


Referring Segmentation

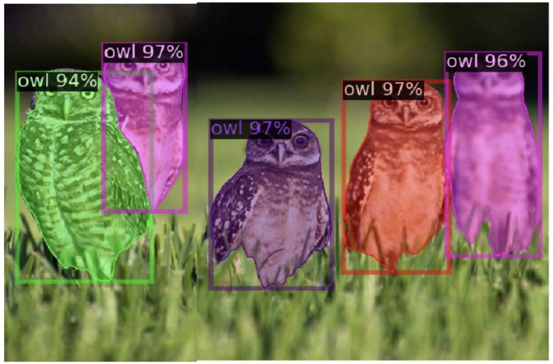


Open-Vocabulary Panoptic Segmentation

Open-Vocabulary Semantic Segmentation



Open-Vocabulary Instance Segmentation

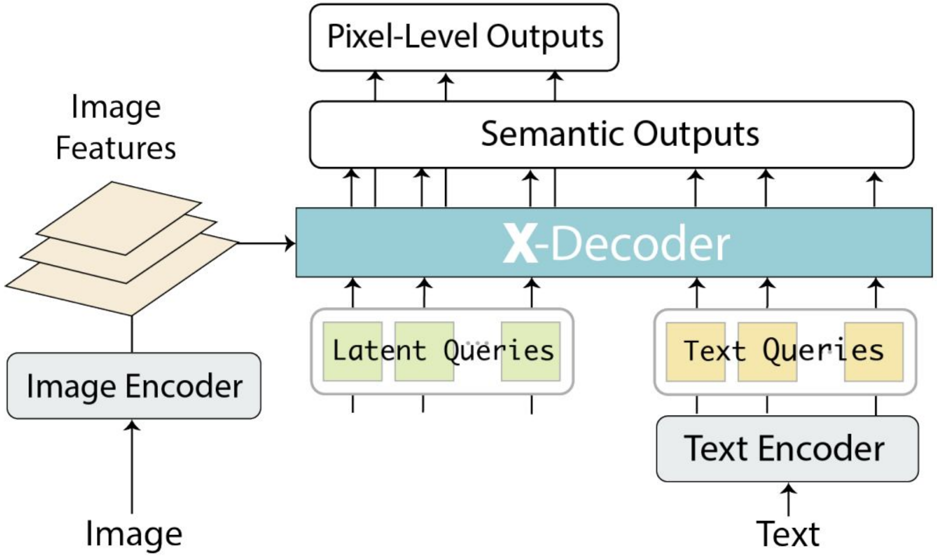


Generalized Decoding for Pixel and Language

We present X-Decoder, a generalized decoding pipeline that can predict pixel-level segmentation and language tokens seamlessly.

X-Decoder takes as inputs two types of queries:

- (i) generic non-semantic queries;
- (ii) semantic queries induced from text inputs.



[Zou & Dou & Yang et al., 2023]