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

Zi-Yi Dou
Advisor: Nanyun (Violet) Peng

UCLA Samueli
Computer Science

PlusLab
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**
- What is the young person doing?

**Image Captioning**
- Several blue tents on a campground

**Image-Text Retrieval**
- A brown cat that is being brushed
Previous Work

Most previous methods rely on CNN-based object detectors.

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.

[Li et al., 2019]
Empirical Studies of Training Vision-Language Transformers

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

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

[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.
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.