Vision-and-Language Pretrained Models

Sravani Thota
Id: 016422406

Agenda

- Computer Vision(CV)
- Natural Language Processing (NLP)
- VLPM’s
- Architecture of VLPM’s
- Encoding Vision and Language inputs.
- Future Research Directions

Computer Vision

Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs.

CV tasks:

- Image classification
- Object detection
- Object tracking
- Content-based image retrieval

NLP

It is the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP tasks:

- Speech recognition
- Part of speech tagging
- Sentiment analysis
- Natural language generation

VLPM’s

- Pretrained models have been highly successful in both Computer Vision (CV) and Natural Language Processing (NLP).
- This success has led to the development of joint representations of vision and language through the use of Visual-Language Pretrained Models (VLPMs).
- VLPMs, have been designed and trained by feeding visual and linguistic contents into the multi-layer transformer.
- The VLPMs have become the de facto standard due to the ease of use and solid representational power of large, publicly available models trained on large-scaled data sources.

Architecture

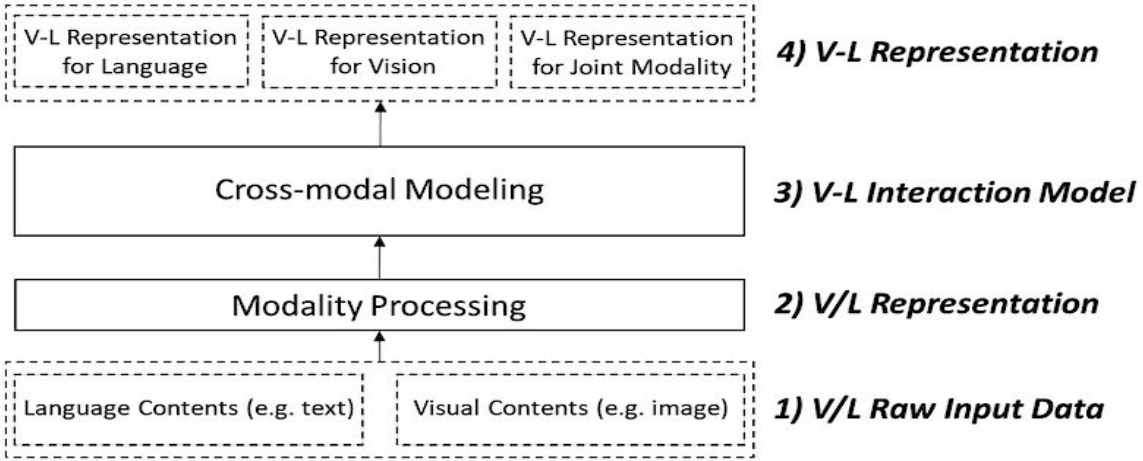

Figure 1: General architecture of VLPMs

Architecture

- **V or L Raw Input Data:** defines the representative raw data streams from the single modality contents, being either vision or language.
- **V or L Representation**: processes the raw data input into the desired format of modality representations.
- **V and L Interaction Model:** enforces the cross-modal modelling between the two modality representations.
- **V and L Representation**: defines the possible cross-modal representations derived from the cross-modal modelling.

Input Encoding

An example of **single image** aligned with **single sentence**

(MSCOCO 2014 dataset)

Old man on a horse walking it on rocks

- The pretraining process normally replies on pairwise image-text corpus
- Most of the downstream tasks entail image-text pair as input, e.g. Text Image Matching

An example of **single image** or **multiple sentences**

(VisDial v1.0 dataset)

Q:Where are they located
A: In city
Q: Any buildings
A: Yes
Q: Do they look like couple
A: They are
...

Visual Dialogue (VD) task entails an image with a dialogue of utterances as raw input.

An example of **multiple image** with **multiple sentences**

(R2R dataset)

Caption: bed, bedroom, family room with blinds, blinds, entry way with mirror, bathtub
Instruction: Exit column on window and pass lounge with bathtub
Instruction: Take left and then stop on entry with furniture
...

Vision-Language Navigation (VLN) task involves a set of panorama images along the navigation trajectory path with the textual caption and instructions.

Future Research Directions

- **VL Interaction Modeling:** Investigating how to explicitly align the embedding features between image and text so that it can learn fine-grained representations
- **VLPM Pre-Training Strategy**: Exploring how the VL multi-tasking can be applied for VLPM pre-training that can generate the best transfer performance on the specific targeted domain, or even more generalizable transfer performance across different domains.
- **Training Evaluation:** The trained VLPMs is only evaluated during the downstream tasks so far. It worth to explore some metrics during the training procedure.