Efficient Vision-Language Pretraining with Visual Concepts and Hierarchical Alignment (ViCHA)

Mustafa Shukor1, Guillaume Couairon1,2, Matthieu Cord1,3
1Sorbonne University & 2Meta AI & 3Valeo.ai

Introduction

- Current Vision-Language Models are trained on large datasets, which need huge computation infrastructure.
- Other paths are also promising: training objectives, model architectures and the quality of data.
- How can we learn from less data?

Method

ViCHA consists of an image encoder (ViT), text encoder (6-layer BERT) and multimodal decoder (6-layer BERT + CA), trained on 4 main objectives; H-ITC, ITM, MLM and MIM on Image-Text pairs datasets:

- Hierarchical Image-Text Contrastive (H-ITC) objective aligns the cross modal representation at different layers of the unimodal encoders.
- MAE-based Masked Image Modeling (MIM) objective that leverages MAE [1] for VLP, by reconstructing masked image tokens.
- Visual Concepts Module that leverages image-level annotations (Visual Concepts-VCs) using CLIP [2], to enrich the visual representation.

Contributions

- Contributions:
  ○ H-ITC
  ○ MAE-MIM
  ○ Visual Concepts Module

Results

Finetuning on VQA v2 NLVR2 and SNLI-VE

We pretrain on COCO, Visual Genome and SBU and then finetune on VQA v2, SNLI-VE, COCO and Flickr30K retrieval, and visual grounding.

Finetuning on COCO and Flickr30K Image-Text Retrieval

<table>
<thead>
<tr>
<th>Method</th>
<th># Pre-train Images</th>
<th>VQA: test-dev</th>
<th>VQA: test-std</th>
<th>NLVR2: dev</th>
<th>SNLI-VE: val</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViCHA</td>
<td>800K</td>
<td>73.23</td>
<td>-</td>
<td>78.14</td>
<td>79.72</td>
</tr>
</tbody>
</table>