Note: MS-CLIP: MODALITY-SHARED CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

sgc

June 2, 2022

1 Introduction

This paper tends to built a modality-shared CLIP(MS-CLIP) architecture where parameters are shared in the vision encoder and the text encoder. Actually this method outperforms CLIP with less parameters.

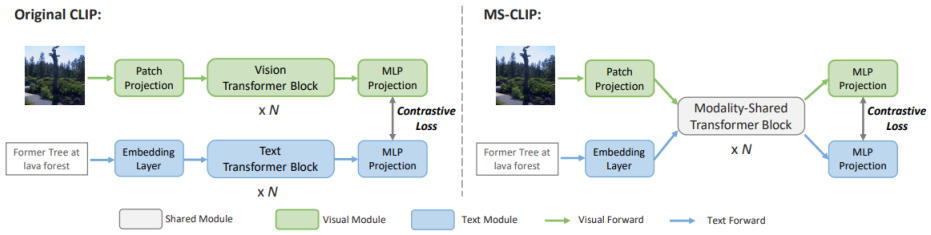

Figure 1: Overview of the original CLIP (left) and our proposed MS-CLIP (right).

2 Method

2.1 Shared Modules

In this work, the attention module, feedforward modules and LN layers are shared, the embedding layer and output projection layer re not shared.

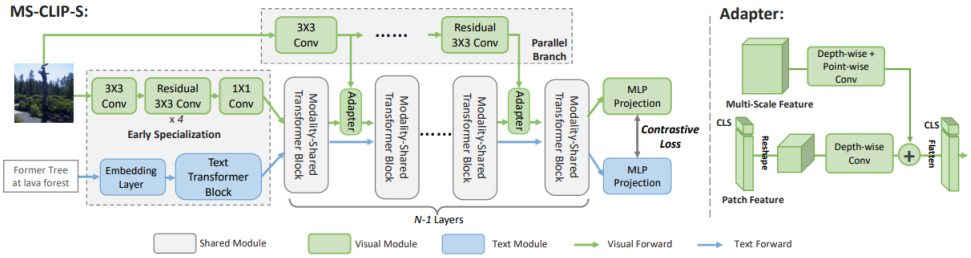

Figure 2: Overview of MS-CLIP-S.

2.2 Modality-Specific Auxiliary Module

Early Specialization Only the first layer is specialized for visual and text, leaving other layers shared. On vision side, CNN is used to downsample the image. On the text level, the de facto Transformer is implemented.

Efficient Parallel Branch Motivated by the work of SlowFast Networks for Video Recognition, this model used a auxiliary parallel branch because it is hard to share weights in ViT and language transformer since they have different number of channels.
The main function of parallel branch is to supplement the main branch with multi-scale feature when an image is taken as the input.
Therefore, they also employ one adapter after each parallel layer to integrate feature in different scales into different layer of shared Transformer.

$$H_p^{'} = bn(\text{PWConv}(\text{DWConv}(H_p)))$$
$$H^{'} = ln(bn(\text{DWConv}(H)) + H_p^{'})$$

where DWConv denotes the depth-wise convolutions and PWConv denotes the Point-wise convolutions. H' is the output.

3 Experiments Results

1. LNs need to be modality-specific.
2. Less is more: Sharing all layers is better than some.Share more layers makes the model performs better.
3. Shared model exhibits higher multi-modal fusion degree.

4 Summary

This paper proposed MS-CLIP sharing most parameter in the transformers on text and image. And a auxiliary parallel branch. This method actually outperforms the CLIP which have a separated architecture with more parameters. To explain that phenomenon researcher did several experiments and think that shared parameters gives a better fusion between the image and text information.

References