Note: ReSTR: Convolution-free Referring Image Segmentation Using Transformers[1]

Sinkoo

April 16, 2022

1 Introduction

Referring image segmentation segment an image region corresponding to a natural language expression given as query. Most precious methods use CNNs and RNNs to process visual and text information respectively, but this is limited by locality of CNN and RNN. To handle that, this paper proposed ReSTR(Referring image Segmentation using TRansformers) that achieve convolution-free by using transformers.

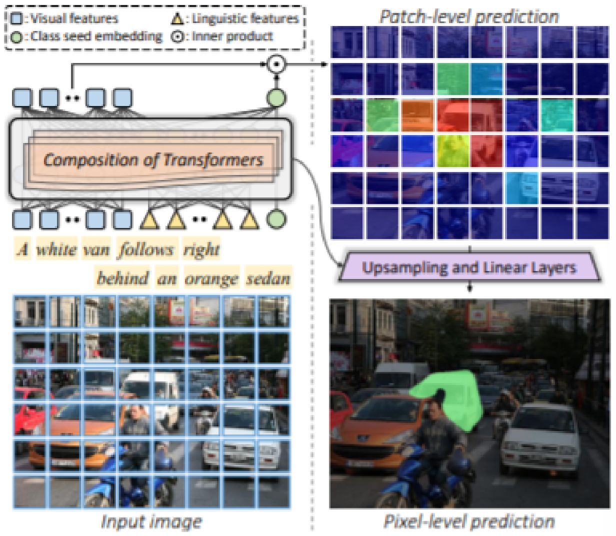

Figure 1. Our convolution-free architecture for Referring image Segmentation using TRansformer (ReSTR) takes a set of non-overlapped image patches and that of word embeddings, and captures intra- and inter-modality interactions by transformers. Then, ReSTR takes a class seed embedding to produce an adaptive classifier which examines whether each image patch contains a part of target entity. Finally, a series of upsampling and linear layers computes a pixel-level prediction in a coarse-to-fine manner.

The contribution:
1, This networks is the first architecture for Referring image segmentation that get rid of convolution and get long distance information by Transformers.
2, Authors designed a multimodal fusion encoder with the class seed embedding which is transformed to an adaptive classifier for referring image segmentation.
3, ReSTR achieves the state of the art on four public benchmarks without bells and whistles.

2 Method

2.1 Visual and Linguistic Feature Extraction

Conventional Transformer blocks are applied to both visual and linguistic procession.

2.2 Multimodal Fusion Encoder

The Multimodal Fusion Encoder consist of two transformer encoders: visual-linguistic and linguistic-seed. The input is visual features z_v and linguistic features z_l and a learnable class seed embedding e_s(initial randomly).
First normalize z_v and z_l and feed each of them into a different linear layer to adjust their channel dimension to be the same as D. Then, the visual-linguistic encoder takes the visual and linguistic features as inputs to produce patch-wise multimodal features $z'_v \in R^{N_v * D}$(N_v is the number of patches).

$$[z'_v, z'_l] = Transformer([z_v, z_l] : \theta_{vl})$$

where z'_l denotes visual-attended linguistic features.
The class seed e_s and visual-attended linguistic features z'_l are feed into linguistic-seed encoder:

$$e'_s = Transformer([z'_l, e_s]; \theta_{ls})$$

where e'_s is an adaptive classifier examines if each patch contains a part of a target entity. The adaptive classifier is used to deal with 2 major demands: 1, Comprehend fine relations of language expression; 2, irrelevant areas.

3 Coarse-to-Fine Segmentation Decoder

The patch level prediction is calculated as:

$$\hat{y}_p = \sigma(\frac{z'_v e'^T_s}{\sqrt{D}})$$

where $\sqrt{D}$ is a normalization factor.
This paper designed a segmentation decoder to compensate for the low-resolution patch-level prediction: First, the decoder produces masked multimodal features:

$$z_{masked} = z'_v \bigotimes \hat{y}_p$$

Then concatenate z_{masked} and z_v. The decoder consists of $K(= logP)$ blocks(illustrated in the Figure). Finally the output is reshaped into $\hat{Y}m \in R^{H*W*1}$.
The total Loss:

$$L(\hat{y}_p, y_p, \hat{Y}_m, Y_m) = \lambda L_n(\hat{y}_p, y_p) + L_b(\hat{Y}_m, Y_m)$$

where y^i_p is defined as:

$$y^i_p = \begin{cases} 1 & \text{if } h(p_{ij} > \tau) \\ 0 & \text{otherwise} \end{cases}$$

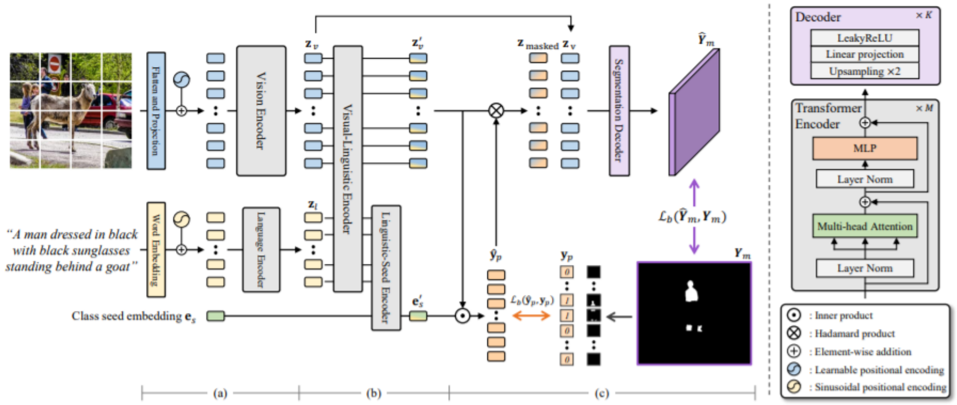

Figure 2. (*Left*) Overall architecture of ReSTR. (a) The feature extractors for the two modalities are composed on transformer encoders, independently. (b) The multimodal fusion encoder consists of the two transformer encoders: *the visual-linguistic encoder* and *the linguistic-seed encoder*. (c) The coarse-to-fine segmentation decoder transforms a patch-level prediction to a pixel-level prediction. (*Right*) Transformer encoder used in all encoders and the composition of the coarse-to-fine segmentation decoder.

4 Results

ReSTR clearly outperforms previous methods and can capture the long range interactions between two modalities(thanks to transformer).

5 Ablation Study

Multimodal encoder structure The classifier shows bias to imbalance of the length of features between visual and linguistic features. This paper considers disconnecting interactions between the visual features and the class seed embedding(illustrated (b),the class seed embedding interacts with only the linguistic features). In the end,this paper proposed a structure that indirectly conjugates the class seed embedding and the visual features with the linguistic features as medium(illustrated (c)).

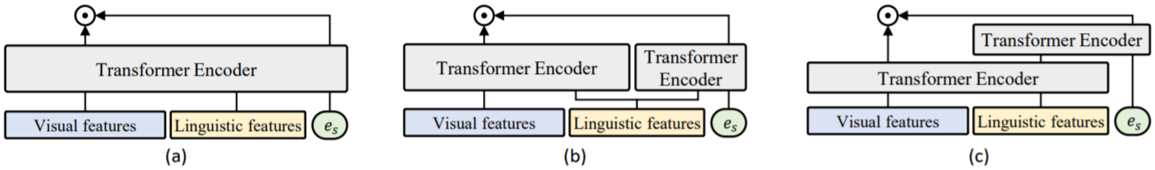

Figure 3. The variants of the multimodal fusion encoder based on transformer architecture. (a) Self-attention fusion encoder on all sequences in parallel. (b) Independent fusion encoder between the visual features and the class seed embedding. (c) Indirect conjugating fusion encoder between the visual features and the class seed embedding.

Number of layers in multimodal fusion encoder Authors test the performance on 2,4,6 layers and find performance improved a lot when number of layers increases from 2 to 4 and improved marginally from 4 to 6.

Segmentation decoder When coupled with the shallow fusion encoder that produces relatively larger potion of false patchlevel predictions, the effect of the segmentation decoder is marginal since it is trained to refine the mask of the positive patches.

Weight sharing The results show that the performance degradation incurred by weight sharing is marginal.

6 Summary

This paper proposed ReSTR that is the first architecture without CNN and RNN to deal with a RSS task, it uses Transformers. And Authors designed a Multimodal Fusion Encoder to concatenate the linguistic and visual information. This ReSTR model exploits the long distance relations using transformer encoders and outperform the previous methods.

References

[1] N. Kim, D. Kim, C. Lan, W. Zeng, and S. Kwak, "Restr: Convolution-free referring image segmentation using transformers," *arXiv preprint arXiv:2203.16768*, 2022. (document)