Enhanced Referring Image Segmentation with Perceptual Loss

Pradeep P
2019MCB1227
Ashish Sharma
2019MCB1213

B.Tech Final Year
Mathematics & Computing
Indian Institute of Technology Ropar
Rupnagar, Punjab, India

Abstract

Multi modal problems involve more than one parallel or related sources of data. Effectively combining feature representations from different modalities is a challenging task in deep learning. We explore here one such task, Referring Image Segmentation, in which two modalities namely, images and text are involved and try improving the performance of existing model.

In this paper, we first present a review of ideas we explored in this domain, elaborate the baseline model and finally we detail the work done by us. We work with the **Language Aware Vision Transformer (LAVT)** as our baseline model. Our main contribution to the project is **incorporation of the perceptual loss with softmax activation** to the LAVT baseline and obtain **better results than the simple cross entropy loss.**

1 Introduction

Referring Image Segmentation (RIS) is one of the important multi-modal problems involving both vision and language, which was first introduced by Hu *et al.* [10]. Given an image and a natural language expression that describes the properties of a target object, the task is to ground the target object described by the language and generate a corresponding segmentation mask.

In RIS like any multi modal task, feature fusion is a challenge. Traditional ways involve obtaining independent feature representations from vision and language models and fusing them using a language decoder. This results in little correspondence between the two modalities.

Next, the task of RIS demands data in a very specific format. We need segmentation masks of images for natural language expressions. These masks, obviously need to be given by human data annotators, who manually outline the regions corresponding to the expression. This is practically not feasible always. To tackle this several weakly supervised learning approaches have been developed. In weakly supervised learning the ground truth values are known, but they are incomplete and not exactly what we want. [7] uses bounding boxes as weak labels, employ a maximization heuristic to obtain contours of the object in an image, given only the bounding box. Another challenge we observed from our experiments include noisy segmentation masks, as can be seen in Fig.2

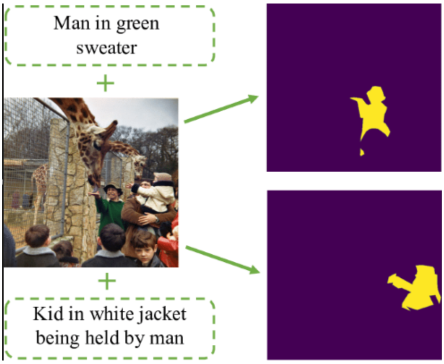

Figure 1: Referring Image Segmentation Example

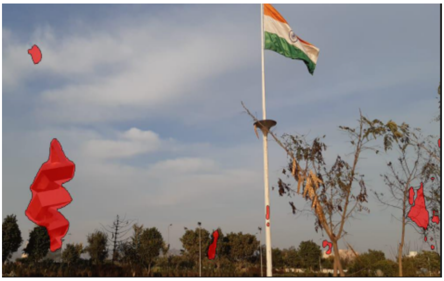

Figure 2: Noisy Output Segmentation mask for the Query "Silver Spiral Structure" on the LAVT Baseline

In the following section we discuss the various ideas in RIS we came across. It's worthy to mention out of all the below ideas, we mainly studied the idea of Locate then Segment [12], and employing bounding boxes as weak labels, a weakly supervised learning approach.

2 Literature Review

[10] fuses the linguistic features extracted by Long Short-Term Memory (LSTM) networks and the visual features extracted by Convolutional Neural Networks (CNN), which are then sent to a fully convolutional network (FCN) [19] to generate the target segmentation mask. Liu *et al.* [17] propose convolutional multimodal LSTM which models each word in every recurrent stage to fuse the word feature with vision features in order to utilize the information of each word in the expression better. Li *et al.* [15] utilize the feature pyramids inherently existing in convolutional neural networks to capture the semantics at different scales and refine the segmentation mask progressively. Edgar *et al.* [21] propose Dynamic Multimodal Network (DMN) which introduces the usage of Simple Recurrent Units (SRUs) for efficient segmentation based on referring expressions. Yu *et al.* [30] propose Modular Attention Network (MAttNet) which extracts instances using Mask R-CNN [8], and then adopts word features to choose the target from those instances. Chen *et al.* [2] employ a caption generation network that takes features shared across both language and visual modules as input, and improves both representations via a consistency that enforces the generated sentence to be similar to the given referring expression. In [20], Luo *et al.* propose a novel Multi-task Collaborative Network (MCN) to jointly learn jointly learn referring expression comprehension (REC) and segmentation (RES) as they are highly related. Bottom-Up Shift (BUS) proposed by Yang *et al.* [27] progressively locates the target object along hierarchical reasoning steps implied by the expression. Due to the effectiveness of attention mechanism [25], many research works are adopting attention mechanism for referring segmentation. It is found to be powerful to extract the visual contents corresponding to the referring expression. Shi *et al.* [24] propose key-word-aware network (KWAN) to extract key words by a query attention model which suppresses the noise in the query and helps in highlighting the desired objects.Ye *et al.* [29] propose cross-modal self-attention (CMSA) module to dynamically capture long-range correlations between informative words in the expression

and important regions in the image. In order to realize the mutual guidance [31] between linguistic and visual features, Hu *et al.* [11] propose a bi-directional cross-modal attention module (BCAM). Jing *et al.* [12] decouple the referring segmentation to localization and segmentation and propose a Locate-Then-Segment (LTS) scheme to locate the target object first and then generate a fine-grained segmentation mask. This method considers the notion that people usually perform locate the corresponding target image regions, and then generate a segmentation mask about the object based on its context. Feng *et al.* [6] propose an encoder fusion network (EFN) which uses co-attention mechanism to refine the multi-modal features progressively. Most of these works are built on FCN-like networks, and only use the attention as auxiliary modules.

Contrastive learning is a method used for unsupervised representation learning. Wang *et al* propose a text-to-pixel contrastive learning based approach derived from CLIP (explained later), CLIP driven Referring Image Segmentation (CRIS) [26]. In [16], Li *et al* emphasise on two issues in Language-Vision models. Usually only high level features produced by unimodal language and vision encoders are fused. Also, the encoders are trained independently. They propose Mask Image-Language trimodal encoder (MAIL) with instance masks as third modality. Yang *et al* propose Language Aware Vision Transformer (LAVT)[28]. They use the idea of fusing the language and vision features in the intermediate layers of encoder transformer network. Ding *et al* focus on the issue of learned queries being fixed after training and not enough to handle the diverse and random natural language expressions. They propose a Query Generation Module (QGM) and a Query Balance Module (QBM). QGM dynamically produces input specific queries and QBM selectively fuses the responses of the queries[4].

3 Background: Transformers

Transformers are a type of neural networks first introduced first in 2017 by Vaswani *et al* [25]. They find vast application in the language and vision-language domain, and in multimodal tasks[9]. They leverage ideas from attention models, RNNs and CNNs. Attention models are slow to train as they are not able to handle large parallel input. On the other hand RNNs and CNNs might not be able to handle all dependencies (RNNs proved ineffective for long range dependencies) in data. Self-attention mechanism allows the network to pay attention to different parts of input, just as we humans do, while CNNs allow to handle parallel input, allowing effective GPU usage.

Transformers are made up of a series of designed similar encoders and decoders. Transformers have a two stage training process. First they are pre-trained on large-scale unlabelled datasets, helping avoiding annotation costs. Then the weights are later fine tuned on downstream tasks like image classification, object detection, action recognition, zero shot classification, etc[14].

There are many use specific versions of transformer based models. Bidirectional Encoder Representations from Transformers (BERT) [3] introduced in 2018 by Google to specialize on NLP tasks. It takes into account context from both sides of the word in a sentence. Image data has spatial coherence and so requires different network designs and training schemes. For vision tasks, we have Vision Transformer (ViT) [5] introduced in 2020, used for image classification, treats each image as a sequence of 16 x 16 patches. For object detection, Detection Transformer (DETR) [1] was introduced in 2020 by Facebook.

In vision language domain, the challenge is how we train a model that develops a correspondence between the pixels and text. Transformers have been successfully used in multi-modal vision language tasks. Ramesh *et al* propose an autoregressive transformer based model for zero shot text to image generation[23], modelling text and image tokens as single stream of data. *Radford et al* use contrastive learning in their Contrastive Language Image Pre-training (CLIP)[22] model to learn the features learned by two separate language and vision transformers.

4 Baseline Model

We use the **Language-Aware Vision Transformer (LAVT)** as baseline for the Referring Image Segmentation task. LAVT leverages a hierarchical Vision Transformer to jointly embed language and vision feature maps, facilitating correspondence between information from different modalities. The LAVT model takes in a pair of image and text embedding and outputs a pixel wise mask highlighting the object from the image.

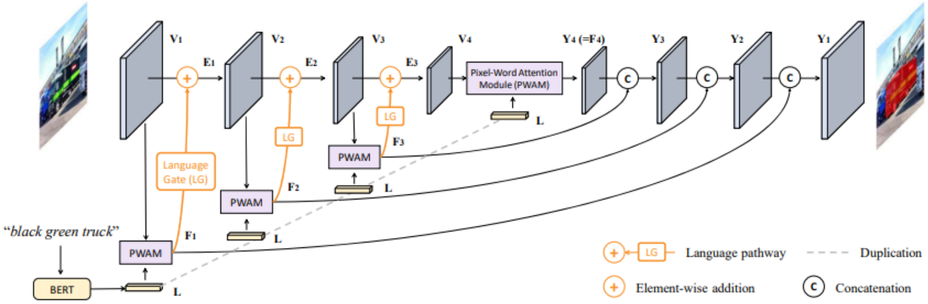

Figure 3: Pipeline of the Baseline LAVT

The Language features are extracted using BERT, a deep language representation model. Joint visual feature encoding and the feature fusion of vision and language features is carried out now in 4 stages of the **Swin Transformer**. LAVT employs the Swin Transformer[18] as the backbone. Three steps (components) are involved in each stage: getting the encoding of the vision features from the transformer layers, feature fusion with the multi-modal feature fusion module and the gating unit.

First, the Swin transformer layers take input as features from previous stage, outputting enriched visual features. Next, the output visual features are fused with the language embeddings with the multi-modal feature fusion module, called the **Pixel Word Attention Module (PWAM)** to obtain multi-modal features.

Finally, these fused features are weighed by the gating unit, called the **Language Gate**, before being added element-wise to the transformer's visual features to obtain enhanced visual features embedded with linguistic information.

5 Methodology

Traditionally in error terms of deep neural networks, pixel values are used as is in pixel-wise loss functions like MSE. Pixel wise loss may not always be able to capture the perceptual

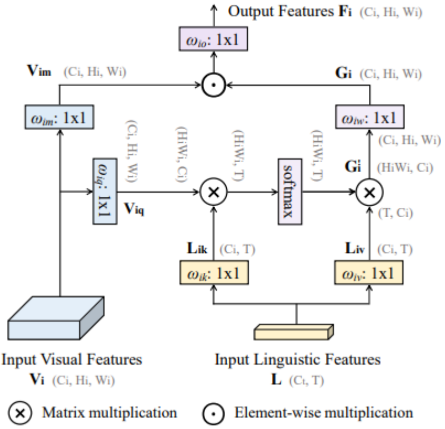

Figure 4: Pixel Word Attention Module Pipeline

similarity between two images. **Perceptual Loss**[13] was introduced by Johnson et. al in 2017 to address this issue. They use pretrained CNNs trained on large-scale image recognition tasks to help get high level feature mappings corresponding to pixel values. Different representations can be obtained from different layer outputs of the CNN, each representing different level of abstraction, the earlier layers capturing low level features like edges, corners and later or deeper layers capturing high level features like specific objects.

We incorporated the **perceptual loss term** in the baseline LAVT and obtained results close enough but inferior to the cross entropy loss. Next we added the **softmax** activation, to the two length output vector probabilities from the segmentation model. This helped in the model's performance to improve beyond the cross entropy loss. The model outputs two values for every pixel, first corresponding to probability of pixel not belonging to mask, second to probability of pixel belonging to the mask. Initially we were taking argmax of the probabilities, but adding softmax helped to improve performance as model tried to make the pixels with say 0.8 probability of mask converge to 1.

We also tried to incorporate the Dice Loss, a popular loss function used in deep learning based segmentation task for imbalanced datasets, but the model did not seem to converge.

Another idea we could have explored was to feed the transformer,tokens of pair of original image with the Canny Edge output. This is kind of the opposite of perceptual loss in a superficial way in the sense edge maps are low level features, whereas the perceptual loss involved using high level representations of the images. This might have helped again in improving the performance, but we couldn't experiment with it due to time constraints.

6 Experiments

We used the **RefCOCO** dataset for carrying out the experiments. The RefCOCO dataset contains 19,994 images with 50,000 annotated objects and 142,209 annotated expressions. The natural language expressions in RefCOCO have an average length of 3.5, so we needed to make sure test or demo expressions too had similar short lengths.

Similar to the LaVT baseline, which uses the Swin Transformer, we used the Tiny Swin

Transformer, a lightweight version of the Swin Transformer to avoid long training times.

We carried out the experiments for 40 epochs each for the cross entropy and the perceptual loss with softmax activation. Each image was of size 224 x 224 and the batch size used for trainig was 32. We used the Adam optimizer with learning rate of 5 x 10^{-4}. The training required lots of computation power, we specifically used 4 units of 11GB NVIDIA GeForce GTX 1080 Ti in parallel.

7 Results

The performance of perceptual loss based model solely without softmax for 15 epochs was slightly inferior than cross entropy. The overall IoU for 15 epochs for the baseline with cross entropy (original paper) was 63.55, whereas for baseline with perceptual loss was 62.07. Similarly other performance metrics namely precision@0.5, precision@0.7, precision@0.9 and overall IoU were also worse for perceptual error based model when compared to baseline

When we also incorporated **softmax activation** the performance surpassed the Baseline model which used cross entropy loss, starting from first epoch The Average Object IoU after training both models for **40 epochs** for the baseline came out as **65.23** whereas for our modified model with both perceptual loss and softmax activation, we obtained Average Object IoU of **65.73**. Even the overall IoU improved from **64.87** for baseline to **65.09** for our modified pipeline. However it is a worthy observation to mention that the IoU of baseline was significantly better in the initial stages than our model. Example in first epoch baseline average object IoU was **26.5** whereas for our model it was **10.15**. So, it is expected that after training for more epochs we might obtain even better results than the baseline.

Model	Mean IoU	precision@0.5	precision@0.7	precision@0.9	Overall IoU
Baseline	65.23	75.18	60.69	17.53	64.87
Perceptual with Softmax	65.73	76.41	61.14	17.19	65.10

Table 1: Table Showing Performance Comparison on 40 epochs of Baseline and our modified model with Perceptual loss with SoftMax Activation

8 Conclusion

In this term paper we modify the baseline by incorporating the perceptual loss function along with using the softmax activation and obtain better results than the baseline proposed in [28], the LAVT paper. We also explored various issues in the field of Referring Image Segmentation like data scarcity which can be handled by weak labels, how to fuse the multi modal features efficiently and how to obtain noise free segmentation maps.

As part of future work we would like to experiment with feeding the transformer the image paired with it's canny edge map, and observe the performance. It is also important to develop data-efficient methods like from weak labels or semi-supervised learning based methods or even self-supervised learning methods in this field of multi-modal learning, which is very data expensive.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

[2] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

[4] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

[6] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15506–15515, June 2021.

[7] Guang Feng, Lihe Zhang, Zhiwei Hu, and Huchuan Lu. Learning from box annotations for referring image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2022. doi: 10.1109/TNNLS.2022.3201372.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1439–1449, October 2021.

[10] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016.

[11] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[14] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[15] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018.

[16] Zizhang Li, Mengmeng Wang, Jianbiao Mei, and Yong Liu. Mail: A unified mask-image-language trimodal network for referring image segmentation. *arXiv preprint arXiv:2111.10747*, 2021.

[17] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020.

[21] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

[24] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

[26] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[27] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021.

[28] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[29] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019.

[30] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[31] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Localize to classify and classify to localize: Mutual guidance in object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.