Grounded Image Situation Recognition

Darshan Singh S
CVIT, IIIT Hyderabad

Darshana S
CVIT, IIIT Hyderabad

Puru Ojha
CVIT, IIIT Hyderabad

Abstract

A coherent understanding of a scene depicted in an image (e.g., what happens and who plays which roles) is challenging and relevant to real-world applications. In this regard, Grounded Situation Recognition is the task of predicting action (verbs) and localized entities (nouns) associated with a semantic role from a given image. This work improves upon the approach in [1] by introducing several design changes such as leveraging CLIP [5] features, object features [6], and contextualized role-aggregated image features. Further, we provide extensive quantitative results on five metrics in three settings and qualitative results on real-world images.

1 Introduction

The performance of deep learning models on basic vision tasks such as identifying objects [3], actions [7], and places [2] has achieved or even exceeded human capabilities. However, a coherent understanding of a scene depicted in an image (e.g., what happens and who plays which roles) is still challenging and relevant to real-world applications.

1.1 Problem Statement

Image Situation Recognition (ISR) [8] is the task of predicting salient actions, their participants, and their roles based on an image. As an addition to ISR, Grounded Situation Recognition (GSR) [4] addresses the localization of nouns in the image. In contrast with ISR, it is more challenging yet allows for a deeper understanding of the scene.

Let V, R, and N denote the sets of verbs, roles, and nouns defined in the task, respectively. For each verb, $v \in V$, a set of semantic roles, denoted by $R_v \subset R$, is predefined as its frame by FrameNet. GSR aims to predict a verb v of an input image and assign a grounded noun to each role in R_v.

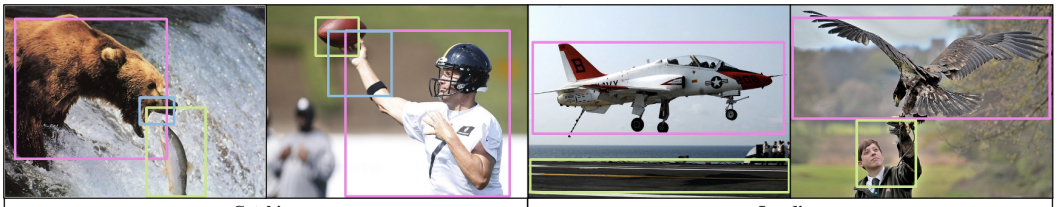
1.2 Challenges

Compared to standard Image Classification task, predicting verbs/actions is much more challenging as it requires modeling both semantic and visual similarity. For instance, although both images of Fig. 1 belong to the same action/verb class, they are visually very different.

Further, it is difficult to model complicated relations between entities. For improved noun prediction and localization, it is necessary to consider the relations between nouns since an action is performed by multiple entities related to one another. Such relationships, however, are latent and dependent on an input image, making modeling them difficult.

2 Contributions

- Reproduced the results of CoFormer [1]
- Improved the results using CLIP [5] Embeddings (ViT-B/32 and ViT-L/14@336px)
- Investigated the performance of using Object features extracted using Faster-RCNN [6]



Catching								Landing					
Agent	Caught Item	Tool	Place	Agent	Caught Item	Tool	Place	Agent	Destination	Place	Agent	Destination	Place
Bear	Fish	Mouth	River	Football Player	Football	Hand	Football Field	Airplane	Runway	Aircraft Carrier	Bird	Arm	Outdoors

Figure 1: An example from the SWiG dataset [4].

- Investigated the performance of contextualized role-aggregated image features
- Quantitative analysis on five metrics in three different settings
- Qualitative Analysis on real-world images (aspect-ratio, clutter, lighting, occlusion!)

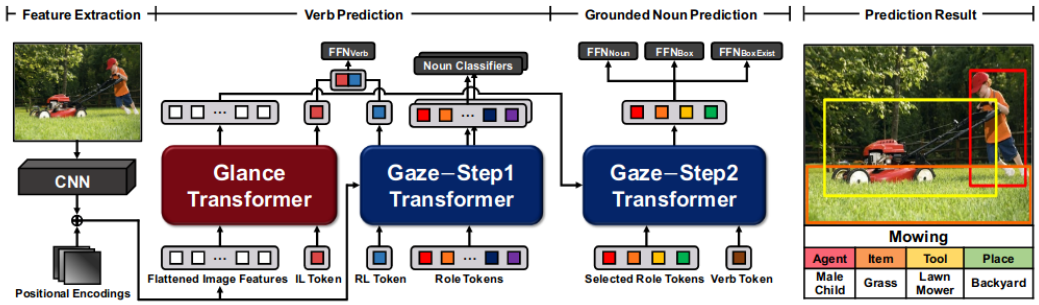


Figure 2: CoFormer [1] extracts flattened image features via a CNN backbone, which is fed as input to the Glance transformer and Gaze-S1 transformers. The output features corresponding to Image-Looking (IL) and Role-Looking (RL) tokens are used for verb prediction. Considering the predicted verb, the Gaze-S2 transformer estimates grounded nouns for the roles associated with the predicted verb by exploiting image features obtained by the Glance transformer.

3 Experiments and Results

- CLIP: With CoFormer [1] as our base model, we use CLIP [5] embeddings of the image along with the Image-Looking (IL) token and Role-Looking (RL) token to enhance verb prediction as shown in Fig. 3. We show the result using CLIP ViT-B/32 and CLIP-ViT-L/14@336px for the same in Table. 6. We got a significant improvement on all the metrics, notably verb prediction accuracy, which improved by 14
- Object Features: We replace the Image Features extracted from ResNet in the original model with Object Features extracted from FasterRCNN [6] as shown in Fig. 4. This reduces the training time from 30 hours to 20 hours.
- Contextualized role-aggregated image features: We replace the input of the Gaze-Step-1 transformer with the contextualized role-aggregated image features instead of image features from ResNet-50 as shown in Fig. 5. We observed no significant improvements with this setting.

4 Qualitative Results

We visualize the predictions of our model in Fig. 7 and compare them with the predictions of the CoFormer model. It can be seen that our model predicts a much tighter bounding box and also predicts the correct verb, which is "Frisking," compared to the CoFormer model. We also show qualitative

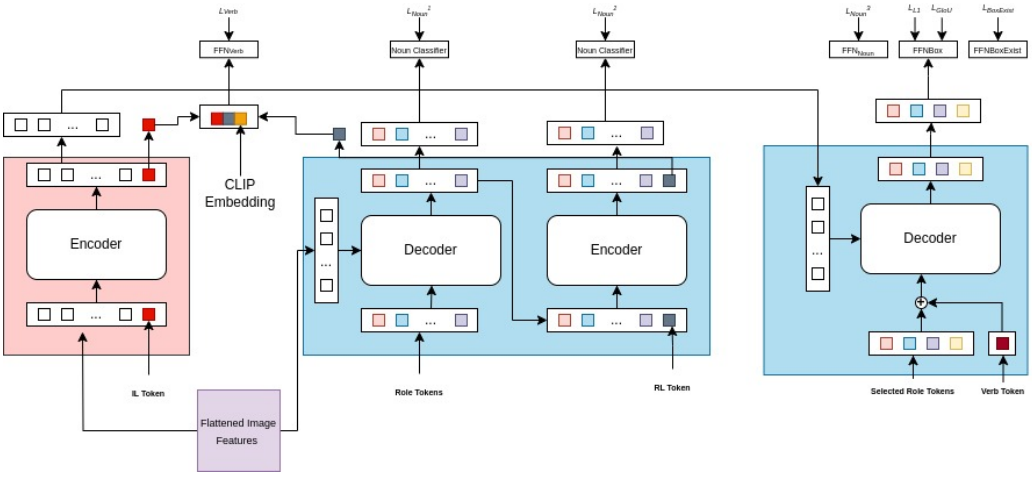


Figure 3: Using CLIP Embeddings.

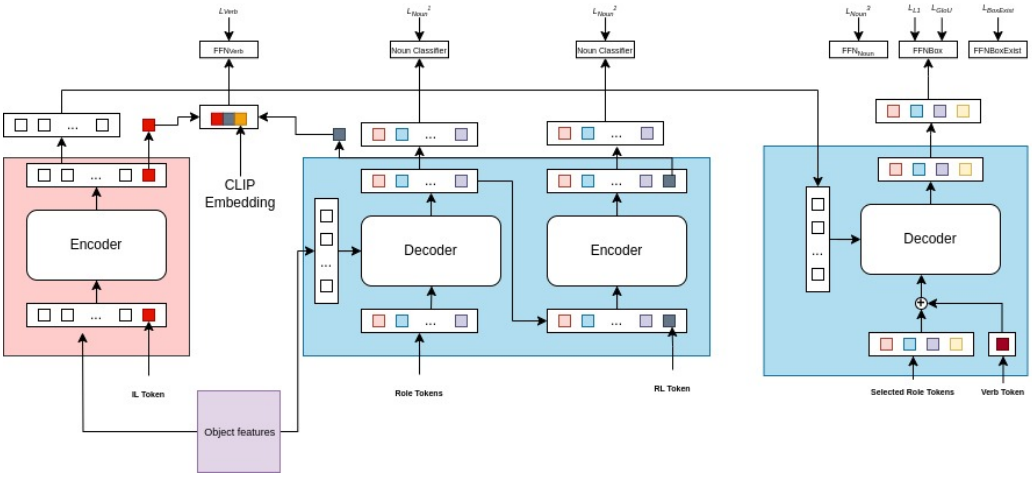


Figure 4: Using Object Features.

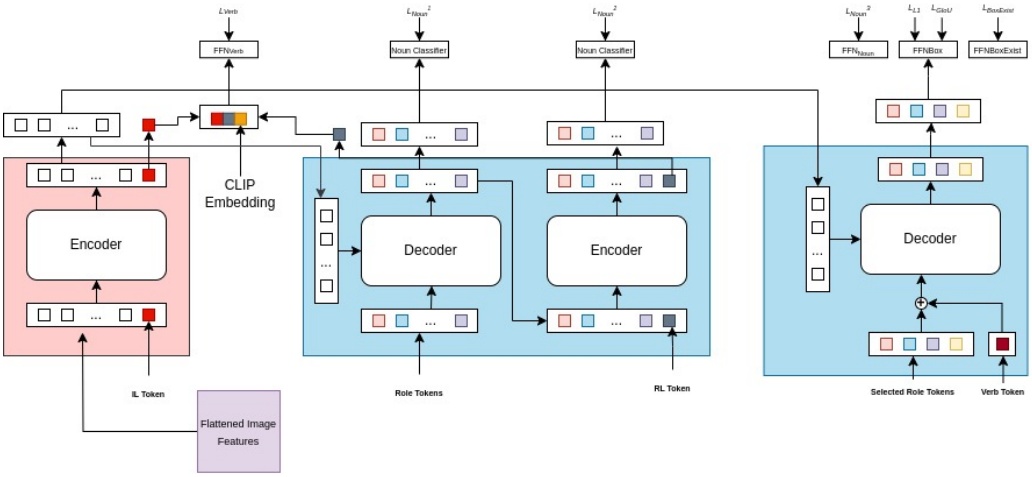


Figure 5: Contextualized role-aggregated image features.

Method	Top 1 Predicted Verb					Top 5 Predicted Verbs					Ground Truth Verb			
	verb	value	value-all	ground value	ground value all	verb	value	value-all	ground value	ground value all	value	value-all	ground value	ground value all
CoFormer	44.66	35.98	22.22	29.05	12.21	73.31	57.76	33.98	46.25	18.37	75.95	41.87	60.11	22.12
CoFormer (Reproduced)	43.79	35.42	22.26	29.39	13.01	72.14	57.01	34.04	46.94	19.54	76.21	**42.46**	62.07	23.96
With CLIP (ViT-B/32)	50.48	40.48	25.28	33.51	14.88	77.85	61.12	36.06	50.1	20.66	76.22	42.36	61.91	23.85
with CLIP (ViT-L/14@336px)	58.43	**46.39**	**28.07**	**38.15**	**16.38**	**84.38**	**65.69**	**38.09**	**53.77**	**21.75**	**76.29**	**42.46**	**62.11**	**24.11**
with CLIP + contextualized img. features	**58.52**	46.03	27.46	37.15	15.05	84.19	65.14	37.37	52.28	20.15	75.63	41.4	60.38	22.16
CLIP + Object feats - Img. feats	57.46	44.06	24.74	33.5	11.61	83.31	62.53	33.6	47.4	15.56	73.16	37.47	55.2	17.35

Figure 6: Results from various experiments.

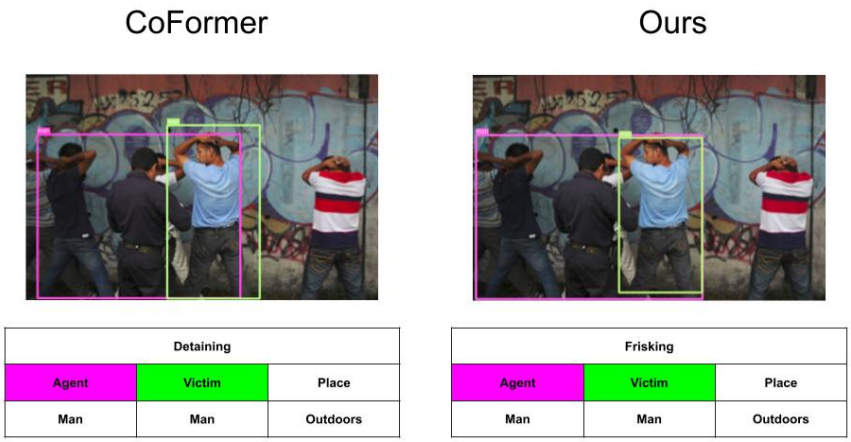


Figure 7: Comparison between CoFormer and Ours.

results on some real-world images captured using a mobile camera with no pre/post-processing of the images in Fig. 8. It can be seen that despite several challenges with the captured images, such as illumination, occlusion, etc., our model is able to make decent predictions.

(a) Painting (b) Sketching

Figure 8: In Fig. (a), all the predictions are correct, but the noun brush could not be grounded. In Fig. (b), the prediction for the role: image is "blank" since it is not clear the what the agent is sketching.

References

[1] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19659–19668, 2022.

[2] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 392–407. Springer, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[7] Marjaneh Safaei and Hassan Foroosh. Still image action recognition by predicting spatial-temporal pixel evolution. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 111–120. IEEE, 2019.

[8] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016.