ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10445738

VISUAL PROMPT TUNING FOR WEAKLY SUPERVISED PHRASE GROUNDING

Pengyue Lin, Zhihan Yu, Mingcong Lu, Fangxiang Feng, Ruifan Li, Xiaojie Wang*

School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{linpengyue, yzh0, lmc8133, fxfeng, rfli, xjwang}@bupt.edu.cn

ABSTRACT

Previous works on the task of weakly supervised phrase grounding (WSG) rely heavily on object detectors providing RoIs for the localization. However, such methods cannot be applied effectively to real-world scenarios largely because that the detectors are trained with limited categories. In this paper, we propose a refinement-based approach to WSG through fine-tuning a detector-free phrase grounding model with a visual prompt. This visual prompt is extracted from the text-related representations in CLIP. Furthermore, we combine the visual prompt with learnable features and then fine-tune the grounding network. Our experimental results significantly outperform state-of-the-art methods on the WSG task and shows the effectiveness of our method.

***Index Terms*—** Weakly supervised, Phrase grounding, Visual prompt tuning, CLIP, Detector-free.

1. INTRODUCTION

Weakly supervised phrase grounding (WSG) is defined as the localization of an object described by a text phrase without additional annotations. The major approaches are to locate RoIs by a pre-trained supervised object detector and transform the grounding task into a retrieval task. However, there still exist countless unlabeled data and classes around the world that are not involved by the pre-trained object detector. To address this problem, a reasonable solution is to relabel the data and retrain the object detector. However, annotating the data is time-consuming and laborious. In addition, training the detector requires a lot of computational resources.

In contrast to the object detector, CLIP [1] trains a joint vision and language model using 400 million image-text pairs and demonstrates impressive results on directly transferring to over 30 datasets. Based on the CLIP's image-text alignment capability, the state-of-the-art detector-free approach [2] is to utilize the visual interpretation method GAE [3] to extract the pseudo labels from CLIP, and then apply the encoder-decoder backbone to generate text-related grounding results. This method is essentially a fully fine-tuning method based on pseudo labels. However, there is still a performance gap with the detector-based methods. We suppose that i) since CLIP is not a task designed for grounding, the fully fine-tuning methods with the pseudo label could destroy the image-text alignment in CLIP, which causes a huge performance loss; ii) fully fine-tuning is a straightforward way of training, and it misleads the model to fall into a local optimum.

To this end, we consider strengthening the relationship between CLIP and the grounding model, and focusing on the grounding image encoder and CLIP image encoder. The former encoder works together with the text encoder to create spatial representations that are aligned with the grounding. In other words, the tensors output by the grounding image encoder are all associated with feature vectors that are aligned with the corresponding text description representation. Thus, spatial locations aligned to the same textual description have similar encodings in both image encoders and vice versa.

Thus, we compute the cosine similarity of CLIP image embedding with all text tokens. These similarity tokens are text-related as they are output by CLIP's image encoder together with the text encoder. By reshaping these similarity tokens, we obtain a fairly excellent visual prompt, aggregating the relevant image patches and textual descriptions, which could be embedded in latent space of the grounding model. Note that this prompt provides not only an exemplar for these learnable fusion features (calculated by grounding image encoder and text encoder), but also an opportunity for further fine-tuning the network. Without changing the training objectives, we demonstrate the significant alterations brought by our method on multiple benchmarks.

Our contributions are: i) We observe that similarity tokens from CLIP's embeddings capture the spatial information of phrase-level concepts; ii) We use obtained prompts to fine-tune the detector-free grounding network; iii) We conduct experiments whose results achieve better performance than state-of-the-art methods.

2. RELATED WORK

Weakly supervised phrase grounding (WSG). WSG is a challenging task that has attracted significant attention in recent years. Detector-based techniques assume the existence of a pre-trained object detector that performs RoI localization. These methods aim to create a joint visual-textual representation space, thereby converting the grounding task into a re-

* Corresponding Author.

trieval task [4, 5, 6, 7]. In contrast, detector-free WSG methods [8, 9, 10, 11, 12] perform dense localization on the given query phrase and generate attention-based heatmaps. Since such methods lack localization information, they define and optimize some auxiliary tasks (such as intra-modal classification and cross-modal alignment) on weakly supervised data. While the auxiliary tasks are not completely consistent with the grounding task, optimizing the tasks can achieve expected phase grounding results. The state-of-the-art method [2] designs a fully fine-tuning method to employ CLIP to ground entities. Subsequently, in a growing works [13, 3, 14, 15, 16, 17], CLIP is used for extracting visual-textual representations in vision encoder, which are applied to grounding-related tasks. Our approach is oriented towards the relationship between CLIP and the grounding model, thus we design visual prompt tuning to refine the training process.

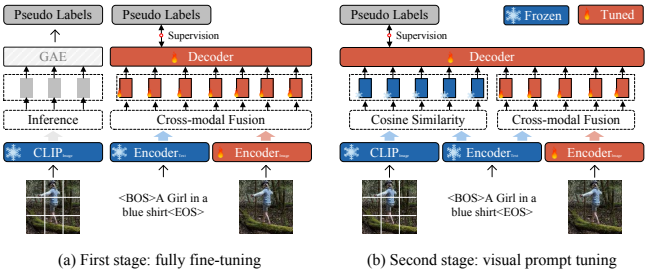
Prompt tuning. The field of visual recognition is rapidly adopting the concept of prompt learning, which has gained popularity in natural language processing tasks [18, 19]. Recent studies in prompt learning [20, 21] have shown promising results in various vision-language tasks, particularly in classification. However, our work differs from these studies in that we focus on WSG instead of classification, and we adopt prompt learning from a different perspective. Existing approaches mostly consider adding a learnable token and freeze pre-trained parameters to achieve prompt learning from a contextualized standpoint. We use input embeddings from the initial model as the learnable parameters, and then fine-tune the improved network to obtain better results.

3. METHOD

3.1. Prompt design

Given an RGB image $I \in R^{3\times W\times H}$ and an input text T, our method adopts the current state-of-the-art architecture for WSG [2]. Our proposed solution for phrase grounding refinement is based on calculating similarity tokens. The similarity tokens are obtained based on the visual embeddings of CLIP image encoder ($e_I = CLIP_{Image}(I)$) and text embeddings of CLIP text encoder ($e_T = CLIP_{Text}(T)$). These tokens are aggregated and serve as a visual prompt for fine-tuning WWbl [2] to obtain a fine-tuned network. Specifically, we first freeze the model parameters of CLIP, maintaining the semantic associations obtained from vision-language pretraining. Then, we resize the image I to 224×224 to fit the CLIP, and get 1×768 text embedding e_T and 256×768 visual embeddings e_I without [CLS] (we use ViT-B/14 of CLIP). By extending the text embedding to 256 copies, we also perform cosine similarity calculation between them and obtain 256 similarity tokens. Finally, we reshape them as a 16×16 similarity map $S_{I,T}$ as follows,

$$S_{I,T} = \text{Reshape}\left(\frac{{e_T}^{\mathrm{T}} e_I}{||e_T||||e_I||}\right) \quad (1)$$



(a) First stage: fully fine-tuning (b) Second stage: visual prompt tuning

Fig. 1. An illustration of our method. We first follow the fully fine-tuning in WWbl [2]: to pass the input image through CLIP image encoder and GAE [3] getting the pseudo label, which is used to train encoder-decoder grounding network. Then, we design a visual prompt based on similarity tokens (blue squares in (b)) to further fine-tune grounding network's parameters. The red squares in (b) represent fusion features.

$S_{I,T}$ is a text-related feature map as the textual embeddings e_T play an important role in its creation. Other downstream tasks [1] such as image classification has shown that such design in the CLIP inference stage can establish connections between encoders.

3.2. Visual prompt tuning

Recently, some visual prompt works [22, 23] have labeled some symbols on images and fed them into pre-trained models. The visual prompts perform better on downstream tasks. The major reason behind is the pre-trained models encountered these symbols and modeled the relationship between such symbols and text features during training. Similarly, we aim is to provide a symbol-like visual prompt for the fusion features of the grounding model. Therefore, we add the constructed similarity map to the fusion features in the latent space and perform inference. However, the grounding model does not show significant performance improvement because no such prompts were encountered during training. Thus, we do not choose to freeze model's parameters and perform a short training for semantic aggregation of the fusion features guided by the prompt.

In fact, our approach changes the structure and feature propagation of the pre-trained model. However, in terms of results, it is equivalent to freezing part of the model (or it could be called "CLIP's") parameters and using part of the learnable parameters to participate in fine-tuning, thus categorizing our approach equally as "visual prompt tuning". In addition, our use of visual prompt tuning after fully fine-tuning is consistent with the solution process from local to global optimization, whereas only visual prompt tuning consumes more computational resources (As shown in Figure 1). Note that during this process we do not change the training objective in WWbl [2].

4. EXPERIMENT

4.1. Datesets

We adopted four benchmark datasets for experimental evaluation. **Flickr30K Entities** [24] contains 224K phrases describing bounding boxes for 31K images, and each image includes 5 captions. Following MG [11], we select 1000 images from the test split for evaluation. **COCO** [25] contains 82,783 training images and 40,504 validation images. **Visual Genome** [26] consists of 77,398 training images, 5,000 test images and 5,000 validation images. **ReferIt** has 20,000 images and 99,535 segmenting regions from another two datasets [27, 28], respectively. There contain approximately 130K entity captions. Following MG [11], we use the same dataset construction strategy with 9K training, 1K validation and 10K test data.

4.2. Implementation Details and Metrics

We used VGG16 as the visual encoder in WWbl [2] for a fair comparison. The model accepts an image size of 224×224, i.e., the image input size of CLIP visual branch VIT-B/32, and generates a heatmap of the same size. We trained 150 epochs for the first stage with SGD optimizer (a batch size of 64 and an initial learning rate of 0.0003), where the optimizer momentum is 0.9 and the weight decay is 0.0001, and trained 1 epoch for the spatial prompt tuning. All methods were implemented on an NVIDIA RTX A6000.

In addition, the "pointing game" accuracy and the bounding box accuracy are used as the performance evaluation metrics. The former measure the percentage of predicted maximum points of the heatmap that lie within the bounding box ground truth. The latter measures the percentage of heatmap bounding boxes that have an IoU greater than 0.5 for the testing set of "image-query" pairs.

4.3. Evaluation Results

We follow the experimental protocol of MG, using the same experimental data and the splits of the train, test as well as validation, and train our method over two train schemes: using COCO train split, and VG train split, and evaluate the model on the test splits of Flickr30K, VG, and Referlt. Our results are compared with the state-of-the-art methods from quantitative and qualitative perspectives respectively.

Performance on standard datasets. The performance of our method and other state-of-the-art DF-WSG methods are shown in Table 1. Our methods are basically superior to the state-of-the-art methods on Visual Genome, Flickr30K, and Referlt. The results show that our method improves all metrics in Flickr30K and Referlt, with absolute improvements of 0.41% - 5.72% for pointing game accuracy and 0.14%- 9.52% for bounding box accuracy. Why can our method achieve performance improvement without changing other training objectives? This is because prompt tuning breaks the local optimal equilibrium formed by fully fine-tuning. Sample results are shown in Fig. 2, compared with WWbl [2]. The visualization results also show that our pipeline is superior to a unique process of fully fine-tuning.

Table 1. Comparison with SoTA DF-WSG methods evaluated using the "pointing game" accuracy and bounding box accuracy on VG, Flickr30K, and ReferIt. The best results are in boldface.

Method	Training	Test Point Accuracy			Test Bbox Accuracy		
		VG	*Flickr*	*Referlt*	*VG*	*Flickr*	*Referlt*
FCVC [29]	COCO	14.03	29.03	33.52	-	-	-
	VG	-	-	-	-	-	-
VGLS [8]	COCO	24.40	-	-	-	-	-
	VG	-	-	-	-	-	-
TD [10]	COCO	-	-	-	-	-	-
	VG	19.31	42.40	31.97	-	-	-
SSS [9]	COCO	-	-	-	-	-	-
	VG	30.03	49.10	39.98	-	-	-
MG [11]	COCO	47.94	61.66	47.52	15.77	27.06	15.51
	VG	48.76	60.08	60.01	14.45	27.78	18.85
GbS [12]	COCO	52.00	72.60	56.10	-	-	-
	VG	53.40	70.48	59.44	-	-	-
WWbl [2]	COCO	59.09	75.43	61.03	27.22	35.75	30.08
	VG	62.31	75.63	65.95	27.26	36.35	32.25
Ours	COCO	**60.74**	**81.15**	**66.14**	**27.65**	**45.09**	**31.14**
	VG	**62.72**	**80.03**	**68.21**	**27.40**	**45.60**	**34.76**

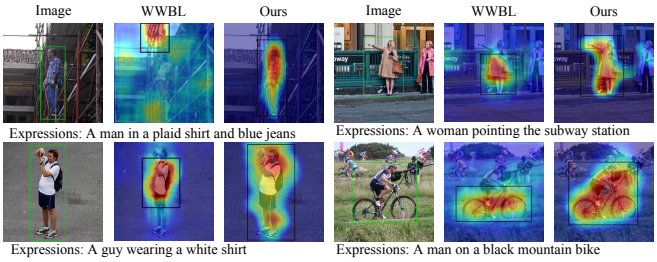


Fig. 2. Several results from phrase grounding models.

Category-wise performance. Interestingly, we find that our proposed method actually attenuates the effect brought about by task-gap between CLIP and the grounding model. As shown in Table 2, the heatmap generated by our model basically outperform that generated by the CLIP-based methods in terms of bounding box accuracy. It indicates that our method gets rid of the performance constraints brought by the CLIP structure and effectively improves its performance on the grounding task. Additionally, we report the point game accuracy under different categories and compare our method with the state-of-the-arts on Flickr30K Entities. Our method not only achieves higher performance than detector-free methods (include Gbs and WWbl), but also goes beyond detector-based methods (including Align2Ground, InfoGround, l2-in-1 and RIR) on almost categories. The training conditions of CLIP provide it with the ability to recognize more categories than the pre-trained detectors, and our visual prompt awakens the grounding model to perceive and

Table 2. Comparison of the first group methods by bounding box accuracy on Flickr30K Entities, and the second group methods by pointing game accuracy on Flickr30K Entities. The best results are in boldface. The suboptimal results are underlined.

Method	Overall	People	Animals	Vehicles	Instruments	Bodyparts	Clothing	Scene	Other
AdaptingCLIP [14]	29.47	29.23	40.15	45.00	24.69	**13.19**	27.23	41.86	24.92
GAE [3]	25.56	26.76	39.72	38.12	**36.76**	9.14	19.56	33.72	22.22
GradCAM [13]	23.18	24.77	35.98	36.30	34.25	8.73	18.24	31.98	20.09
MaskCLIP [15]	34.26	37.46	40.93	52.25	36.42	9.56	**29.36**	48.4	25.87
Ours	**45.60**	**56.63**	**60.31**	**57.83**	27.05	7.32	25.71	**70.29**	**32.71**
Gbs [12]	72.60	82.50	91.50	81.10	56.60	34.80	58.60	70.90	59.90
WWbl [2]	75.63	-	-	-	-	-	-	-	-
Align2Ground [7]	71.00	-	-	-	-	-	-	-	-
InfoGround [5]	76.74	83.20	89.70	87.00	69.70	**45.10**	**74.50**	80.60	67.30
l2-in-1 [6]	76.40	85.70	82.70	**95.50**	**77.40**	33.30	54.60	80.70	**70.60**
RIR [4]	78.60	-	-	-	-	-	-	-	-
Ours	**81.15**	**88.49**	**92.19**	89.37	55.72	25.94	55.47	**97.17**	59.31

Table 3. Visual prompt tuning with different prompts. The performance on VG, Flickr30K, and ReferIt is shown.

Model	Prompt Type	Point Accuracy			Bbox Accuracy		
		VG	*Flickr*	*Referlt*	*VG*	*Flickr*	*Referlt*
GAE [3]	CLIP	**54.72**	**72.47**	**56.76**	16.70	25.56	19.10
GradCAM	CLIP	53.08	71.11	54.37	14.25	23.18	17.80
MaskCLIP [3]	CLIP	54.24	70.32	56.41	18.73	34.26	20.28
Ours	CLIP	50.27	63.83	51.34	**20.91**	**38.25**	**22.64**
GAE [3]	WWbl	56.75	79.21	67.50	18.76	25.78	31.81
GradCAM	WWbl	54.54	78.06	66.77	17.34	24.27	31.11
MaskCLIP [3]	WWbl	50.80	75.82	66.53	19.34	40.02	33.18
Ours	WWbl	**62.72**	**80.03**	**68.21**	**27.40**	**45.60**	**34.76**

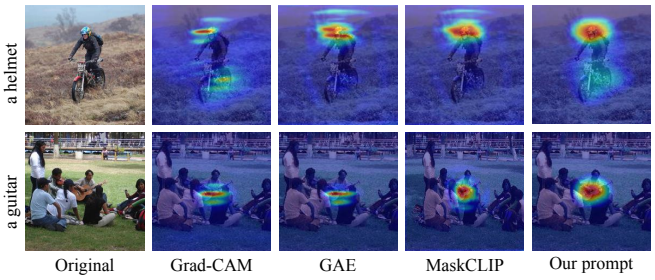


Fig. 3. Our method's qualitative results with various prompts.

attend to more categories. It worth noting that our method does not have the best performance in the categories of "Vehicles", "Instruments", "Bodyparts" and "Clothing". CLIP is pre-trained on images from the Internet, and the images depicting bodypart or clothing often show people rather than referred part. Therefore, whether encountering the phrase of bodypart or clothing during training the model tends to localize people. The poor performance of instruments might be attributed to small amount in pre-trained datasets.

Prompt variations. We also evaluate the proposed approach with different visual prompts designed by other introduced CLIP-based methods, including MaskCLIP [15], GradCAM [13], GAE [3]. AdaptingCLIP [14] was not designed for the relevant experiment because its too slow inference speed affects the training speed of the grounding model. From Table 3, we find that using the proposed similarity map in our pipeline results can achieve the best grounding performance. We conclude two explanations. First, some prompts tend to ground the most discriminative region related to the phrase description while ignoring the object boundary semantics. This reduces their IoU despite them also being able to capture the object of interest. The other reason lies in the design of CLIP-based methods. They design grounding schemes for the features in the CLIP image encoder. However, the CLIP model is trained in a contrastive learning framework on image-text pairs, which may make these visual features lacking semantic aggregation from text embeddings.

5. CONCLUSIONS

CLIP provides the ability to enrich the semantic representation of the input image. Although it has been widely used in recent years, we are unaware of other contributions of applying multimodal information processing techniques to dual-encoder embedding spaces to obtain spatial information. Our work suggests that this is an oversight because the information in the embeddings can easily be further exploited, by aggregating the semantic similarity between different modal features. It is natural to extract a visual prompt that is very useful for improving the performance of weakly-supervised phrase grounding method beyond the state-of-the-arts. In the future, we will study the application of our method to closely related multi-modal tasks, such as image captioning [30, 31].

6. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 61906018 and 62076032, BUPT Excellent Ph.D. Students Foundation CX2023113, and High-Performance Computing Platform of BUPT.

7. REFERENCES

[1] Alec Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[2] Tal Shaharabany et al., "What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs," in *NeurIPS*, 2022, pp. 28222–28237.

[3] Hila Chefer et al., "Transformer interpretability beyond attention visualization," in *CVPR*, 2021, pp. 782–791.

[4] Yongfei Liu et al., "Relation-aware instance refinement for weakly supervised visual grounding," in *CVPR*, 2021, pp. 5612–5621.

[5] Tanmay Gupta et al., "Contrastive learning for weakly supervised phrase grounding," in *ECCV*, 2020, pp. 752–768.

[6] Jiasen Lu et al., "12-in-1: Multi-task vision and language representation learning," in *CVPR*, 2020, pp. 10437–10446.

[7] Samyak Datta et al., "Align2ground: Weakly supervised phrase grounding guided by image-caption alignment," in *ICCV*, 2019, pp. 2601–2610.

[8] Fanyi Xiao et al., "Weakly-supervised visual grounding of phrases with linguistic structures," in *CVPR*, 2017, pp. 5945–5954.

[9] Syed Ashar Javed et al., "Learning unsupervised visual grounding through semantic self-supervision," in *IJCAI*, 2019, pp. 796–802.

[10] Jianming Zhang et al., "Top-down neural attention by excitation backprop," *IJCV*, vol. 126, no. 10, pp. 1084–1102, 2018.

[11] Hassan Akbari et al., "Multi-level multimodal common semantic space for image-phrase grounding," in *CVPR*, 2019, pp. 12476–12486.

[12] Assaf Arbelle et al., "Detector-free weakly supervised grounding by separation," in *ICCV*, 2021, pp. 1801–1812.

[13] Sanjay Subramanian et al., "Reclip: A strong zero-shot baseline for referring expression comprehension," in *ACL*, 2022, pp. 5198–5215.

[14] Jiahao Li et al., "Adapting clip for phrase localization without further training," *arXiv*, pp. 1–17, 2022.

[15] Chong Zhou et al., "Extract free dense labels from clip," in *ECCV*, 2022, pp. 696–712.

[16] Hsuan-An Hsia et al., "Clipcam: A simple baseline for zero-shot text-guided object and action localization," in *ICASSP*, 2022, pp. 4453–4457.

[17] Zeyi Sun et al., "Alpha-clip: A clip model focusing on wherever you want," *arXiv*, pp. 1–22, 2023.

[18] Pengfei Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[19] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL-IJCNLP*, 2021, pp. 4582–4597.

[20] Yu Du et al., "Learning to prompt for open-vocabulary object detection with vision-language model," in *CVPR*, 2022, pp. 14084–14093.

[21] Menglin Jia et al., "Visual prompt tuning," in *ECCV*, 2022, pp. 709–727.

[22] Aleksandar Shtedritski et al., "What does clip know about a red circle? visual prompt engineering for vlms," in *ICCV*, 2023.

[23] Lingfeng Yang et al., "Fine-grained visual prompting," in *NeurIPS*, 2023, pp. 1–20.

[24] Bryan A Plummer et al., "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015, pp. 2641–2649.

[25] Tsung-Yi Lin et al., "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[26] Ranjay Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, pp. 32–73, 2017.

[27] Kan Chen et al., "Query-guided regression network with context policy for phrase grounding," in *ICCV*, 2017, pp. 824–832.

[28] Michael Grubinger et al., "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International workshop ontoImage*, 2006, vol. 2, pp. 13–23.

[29] Hao Fang et al., "From captions to visual concepts and back," in *CVPR*, 2015, pp. 1473–1482.

[30] Yun Liu et al., "Improving image paragraph captioning with dual relations," in *ICME*, 2022, pp. 1–6.

[31] Yihui Shi et al., "S2td: A tree-structured decoder for image paragraph captioning," in *ACMMM Asia*, 2021, pp. 1–7.