Improving Visual Grounding by Encouraging Consistent Gradient-based Explanations

Ziyan Yang*, Kushal Kafle‡, Franck Dernoncourt‡, Vicente Ordóñez*
* Rice University ‡ Adobe Research

Motivation

- Weakly supervised visual grounding tasks rely on object detectors trained on images annotated with labeled bounding boxes.
- Our work demonstrates that using region and box annotations directly in the vision-language model pretraining stage can bypass the need for training object detectors and achieve much better performances on visual grounding and referring expression comprehension.

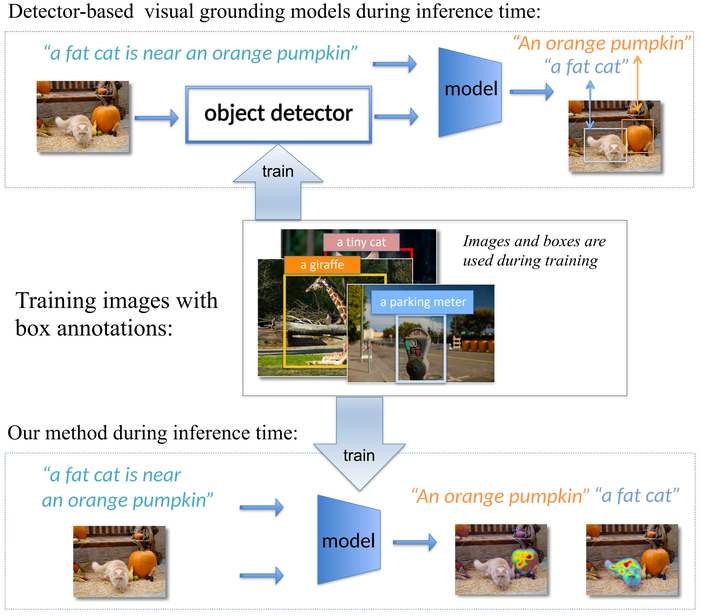

Figure 3: We compare our method with the detector-based framework. Detector-based methods use object detectors trained on images with box annotations to extract image features and generate box proposals. Our method gets rid of object detectors and generate heatmaps directly.

Method

- We inherit model structures and all loss functions proposed in ALBEF [1], including L_{mlm}, L_{itm} and L_{itc}.
- We propose Attention Map Consistency (AMC) loss. It relies on producing explanation heatmaps using the GradCAM [2] method.

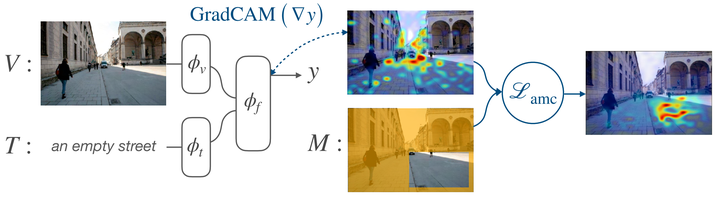

Figure 2: Overview of our method. Among other objectives, standard vision-language models are trained to produce a matching score y given an input image-text pair (V,T) as shown on the left. For inputs containing region annotations (e.g.~a triplet (V,T,M)), we optimize the GradCAM gradient-based explanations of the model so that the produced explanations are consistent with region annotations.

- Some images in our dataset have region annotations to construct a triplet (V,T,M) where $M \in \{0, 1\}$ is a binary mask such that $M_{i,j}$ is 1 if the location i, j is inside boxes or 0 otherwise. V and T are images and text descriptions. This heatmap A, constructed from an intermediate layer of the fusion encoder identifies which area in the image explains the model decision for L_{itm}. We leverage the region annotations M so that the model focuses its heatmap scores in A inside the region of interest indicated by M.

$$\mathcal{L}_{\text{mean}} = \max\left(0,\ \frac{1}{\sum_{i,j}(1-M_{i,j})}\sum_{i,j}\left((1-M_{i,j})\,A_{i,j}\right) - \frac{1}{\sum_{i,j}M_{i,j}}\sum_{i,j}M_{i,j}A_{i,j} + \Delta_1\right) \quad (1)$$

$$\mathcal{L}_{\text{max}} = \max\left(0,\ \max_{i,j}\left((1-M_{i,j})\,A_{i,j}\right) - \max_{i,j}M_{i,j}A_{i,j} + \Delta_2\right) \quad (2)$$

$$\mathcal{L}_{\text{amc}} = \mathbb{E}_{(V,T,M)\sim D}\left[\lambda_1 \cdot \mathcal{L}_{\text{mean}} + \lambda_2 \cdot \mathcal{L}_{\text{max}}\right] \quad (3)$$

[1] Li, Junnan, et al. "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in Neural Information Processing Systems* 34 (2021).
[2] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Results

- We apply our method on a pre-trained ALBEF model by finetuning it on the Visual Genome (VG) [3] dataset. The evaluation metric is pointing accuracy. We compare with VMRM [4] and ALBEF and evaluate our method on Flickr30k Entities [5] and RefCOCO+ [6].

Data	Flickr30k	RefCOCO+	
		test A	test B
VMRM	81.11	58.87	50.32
ALBEF	79.14	69.35	53.77
AMC	86.49	78.89	61.16

Table 1: We compare with previous state-of-the-art methods.

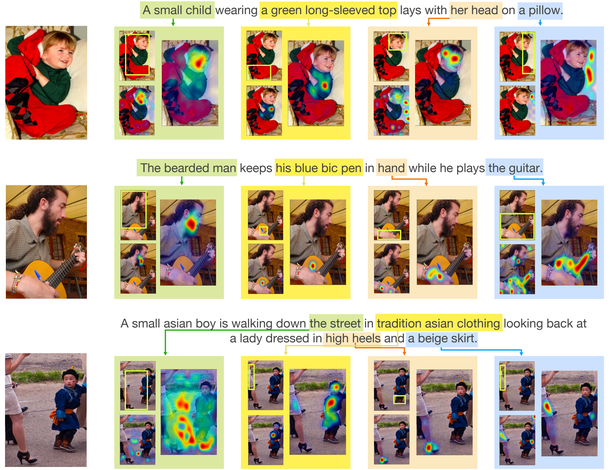

Figure 3: Qualitative comparison of the generated explanations for various images and input phrases. For each phrase, we show VMRM and ALBEF results on the left and our results on the right.

[3] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International journal of computer vision* 123.1 (2017): 32-73.
[4] Dou, Zi-Yi, and Nanyun Peng. "Improving Pre-trained Vision-and-Language Embeddings for Phrase Grounding." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.
[5] Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." *Proceedings of the IEEE international conference on computer vision*. 2015.
[6] Yu, Licheng, et al. "Modeling context in referring expressions." *European Conference on Computer Vision*. Springer, Cham, 2016.