2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS) | 978-1-6654-7735-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CCIS57298.2022.10016352

Multimodal Co-Attention Mechanism for One-stage Visual Grounding

Zhihan Yu[1], Mingcong Lu[1], Ruifan Li[1,2*]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Engineering Research Center of Information Networks, Ministry of Education, Beijing, 100876, China
{yzh0, lmc8133, rfli}@bupt.edu.cn

Abstract: Visual grounding aims to locate a specific region in a given image guided by a natural language query. It relies on the alignment of visual information and text semantics in a fine-grained fashion. We propose a one-stage visual grounding model based on cross-modal feature fusion, which regards the task as a coordinate regression problem and implement an end-to-end optimization. The coordinates of bounding box are directly predicted by the fusion features, but previous fusion methods such as element-wise product, summation, and concatenation are too simple to combine the deep information within feature vectors. In order to improve the quality of the fusion features, we incorporate co-attention mechanism to deeply transform the representations from two modalities. We evaluate our grounding model on publicly available datasets, including Flickr30k Entities, RefCOCO, RefCOCO+ and RefCOCOg. Quantitative evaluation results show that co-attention mechanism plays a positive role in multi-modal feature fusion for the task of visual grounding.

Keywords: Deep learning; Visual grounding; Multimodal Alignment; Attention mechanism

1 Introduction

Recent years, deep learning has driven rapid progress in the field of multimodal machine learning. As one of the multimodal research directions, multimodal alignment requires to identify direct relationships between sub-elements from different modal information, which leads to a more detailed and accurate multimodal representation and comprehension.

Visual grounding is such a multimodal alignment task which attempts to establish correspondence between phrase descriptions and image regions. Generally, the grounding process includes extracting image features, parsing text query, and leveraging correlation information between image features and textual representations to locate the specific object. The major challenge of multimodal machine learning is the high degree of data heterogeneity between different modalities. Thus, solving the visual grounding problem requires compound visual reasoning as well as quantitative comparation between visual information and language semantics. Therefore, visual grounding has attracted attention of many researchers, becoming a new hot topic in the field of multimodal research.

* Corresponding author.

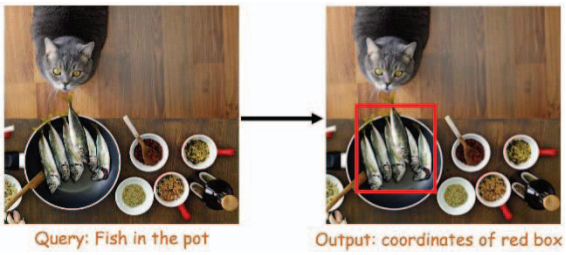

Figure 1 An Illustration on Visual Grounding.

The value of the research lies in image-text alignment and human-computer interaction. More specific application scenes include interactive photo editing, experiential visualization, and robot navigation. As shown in Figure 1, with a given query the model is required to produce the red box. The entertaining and educational aspects of this technology also allow it to be used in early education. Additionally, an effective visual grounding model will be enlightening for a range of other vision-language tasks, such as visual question answering [1], image captioning [2, 3], and cross-modal retrieval [4].

In this paper, we propose a one-stage visual grounding model based on feature fusion. The model achieves end-to-end optimization by directly predicting the coordinates of the target object through deep fusion of visual and textual features, which are encoded by CNNs and large-scale pre-trained text representation model, respectively. The multimodal feature fusion module specifically incorporates co-attention mechanism to deeply capture the hidden semantic and visual information. Finally, localization of the target object is gradually implemented and optimized taking advantages of an object detector.

Our major contributions are summarized as follows: 1) We propose a one-stage visual grounding model based on feature fusion and introduce co-attention mechanism to fuse features extracted from different modalities. 2) We evaluate our model on public datasets including Flickr30k Entities and RefCOCO and achieve good results, which shows that co-attention mechanism is effective in feature fusion for visual grounding.

2 Related Work

The mainstream methods of supervised visual grounding can be summarized into two categories: two-stage methods and one-stage methods.

978-1-6654-7735-2/22/$31.00 ©2022 IEEE

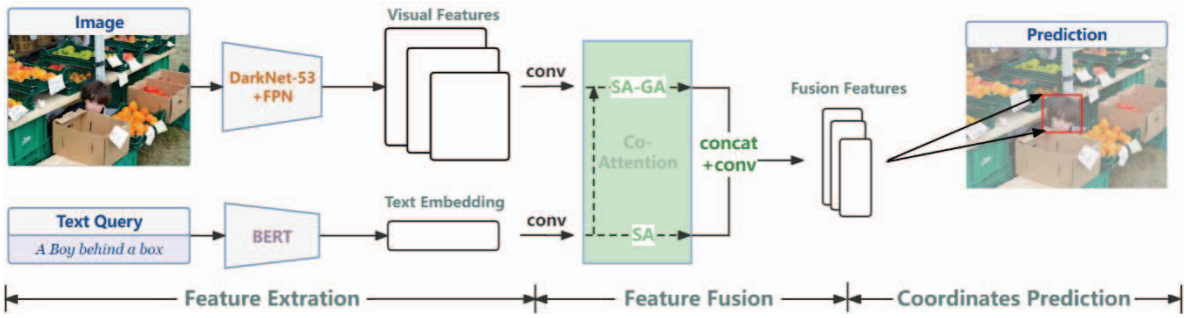

Figure 2 Structure of Our Model

Two-stage methods regard visual grounding as a multi-classification problem on several candidate regions. In the first stage, multiple candidate regions are proposed by an object detector, and the features of each candidate are extracted. Subsequently, all the region features and the textual representations are projected into a common vector space. In the second stage, those candidate regions are scored based on the similarity between text embeddings and themselves, and the region with the highest score is output as the prediction result. For example, Zhang et al. [5] propose a variational Bayesian approach to represent the interrelationship of denotational expressions and their contexts with a scoring function that has a variational lower bound consisting of multimodal modules of three specific cues. MattNet [6] is a modular attention network, which attempts to split the query expression from the semantic level into three modules and the sorting score is weighted summation of scores of the three modules. The advantages of two-stage methods are higher accuracy and better interpretability. However, those models generally have redundant candidate regions, leading to high computational effort and slow convergence.

One-stage methods discard the concept of candidates in two-stage methods and treats the task as a direct coordinate regression problem. Generally, the bounding box's coordinates are predicted by the feature mapping of the query text closely combined with visual features. In comparison, one-stage grounding models are simpler in structure and faster enough to locate objects in real time. SSG [7] is a one-stage model that leverages an attention-based multimodal interactor to fuse text features and image features extracted from Bi-directional Long Short Term-Memory (Bi-LSTM) and Convolutional Neural Network (CNN) respectively. Similarly, FAOA [8] concatenates image features and language vectors, and feeds them together to a single-stage object detection model to predict the bounding box of target object. On the other hand, RCCF [9] describes visual grounding as a correlation filtering process, where text expressions mapping from linguistic domain to visual domain is used as convolution kernels to correlation filter the image feature maps.

Latest works focus on pre-trained models. TransVG [10] leverages Transformer-based framework to establish multimodal correspondence, reasoning about intra-modal and inter-modal relationships by multiple stacked Transformer layers. MDETR [11] is also based on the Transformer structure, where text representations and image features are stitched together and introduced into the end-to-end object detection model DETR. Such Transformer-based approaches usually have higher performance and better generalization ability, with the disadvantage of large parameter size and huge computational resource consumption.

3 Proposed One-stage Grounding Model

The proposed visual grounding model is based on the classical one-stage visual grounding framework, which regards the grounding task as a coordinate regression problem, using CNN-based frameworks and pre-trained language models as encoders to generate a mapping of fused feature vectors to image regions.

The structure of our model is shown in Figure 2. The model encodes the image information by DarkNet-53 and encodes the text query by BERT [12], then combines them through multi-modal co-attention mechanism. The fused features are finally passed through an object detector YOLOv3 [13] to obtain the coordinates of prediction box, thus implementing the recognition and localization of the specific target object.

3.1 Feature Extraction

The feature extraction module consists of two parts: image feature extraction and text feature extraction. For the input image matrix, the module first resizes it to 256×256 dimensions, and extracts features initially by two convolutional blocks. Then the feature vector is mapped to deep features by several residual convolution blocks, which does not change the dimension of the image features. Down-sampling of the image information can be achieved gradually by setting the step size of the convolution operations to 2. The result of convolution is an 8×8×1024-dimensional feature vector, used as the initial image feature vector.

Similar to object detection, the problem of varying scales of target objects still exists in visual grounding. In the process of down-sampling layer by layer in convolution blocks, visual information of small targets is easily lost. To solve this problem, feature pyramid network [14] fuses shallow convolutional outputs with deep outputs through top-down pathway and lateral connections to obtain a pyramidal multi-scale image feature representation.

In order to simplify the subsequent feature fusion calculation, the image feature extraction module finally adjusts the number of channels to 512, followed by regularization to improve the generalization ability. The module finally outputs three different scales of image feature information.

The text feature extraction module, on the other hand, adopts pre-trained model Bidirectional Encoder Representations from Transformers (BERT) to obtain the textual representation vectors. This language model based on transformer [15] architecture consists of multiple 12-head attention blocks with stack layers of 12, and output a 768-dimensional textual representation vector. The text feature extraction module also sets up a multilayer perceptron (MLP) to adjust the vector's dimension. The MLP is composed of two fully-connected layers with activation functions, through which the dimension of text feature vector is set down to 512. Similar to the extraction of image features, text features are finally regularized to reduce the effect of overfitting.

3.2 Feature Fusion

The feature fusion module utilizes a modular co-attention mechanism [16] to extract deep information from both modalities. The module takes image feature vectors and text representation vectors with the same dimension as inputs, and outputs feature fusion vectors. Notably, the feature pyramid structure is extended in this module, i.e., the image features of three scales are fused with text features in parallel so that the fusion vector still maintains the three scales of information for subsequent modules.

The cross-modal co-attention mechanism defines two types of attention operations: Self-Attention (SA) and Guided-Attention (GA). The difference between the two is the source of Q, K, and V. In other words, Q, K, and V in Self-Attention come from the same input vector, while Q and K, and V in Guided-Attention come from feature vectors of two different modalities. The combination used in this fusion module is shown in Figure 3. Text features are processed by only one SA block, while image features are first processed by a SA block, and then the guided attention operation is realized with itself as Q and text features as K and V. Series of attention operations cause the weights of text and image features more inclined to valid information.

Output of above module are converted into a feature fusion vector by concatenation and convolution. After four convolutional blocks, the dimension of the concatenated vector is reduced to the original dimension of 512, and finally reduced to 15, which corresponds to five types of information of the bounding box at three different scales.

3.3 Coordinates Prediction

The coordinates prediction module mainly refers to the anchor-based object detection model YOLOv3 [13]. The original image is divided into several non-intersecting grids, and each grid is forward propagated to obtain a prediction vector, which is composed of the bounding box

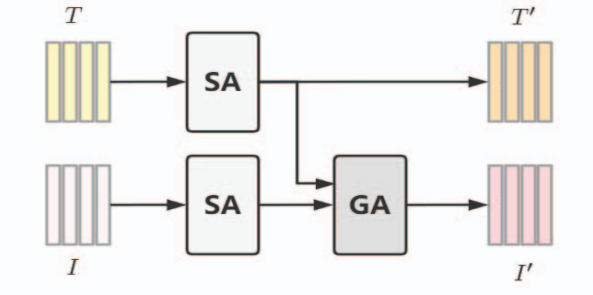

Figure 3 Co-attention Module in our model

coordinates and the confidence of the bounding box. After that, the module focuses on the grid where the center point of ground truth in the supervision information is located, which is called Grid Cell. Several anchors are then applied to calculate their intersection ratio with the ground truth area. The prior box, obtained by selecting the anchor box with the highest intersection ratio, is combined with the fusion feature vectors to predict the offset value of its coordinates. The priorly learned anchor box is proved very helpful for the model to fine-tune the coordinates of bounding box.

The three scales of fusion vector under the feature pyramid structure continue to be retained, and each scale corresponds to three anchor boxes, so that there are nine anchor boxes in total, which are computed by K-means clustering analysis of the dimensions of all the ground truth boxes in datasets in advance. Since the visual grounding task requires only one output, the confidence of one grid is computed by Softmax function. Among the bounding box vectors represented by the feature fusion vectors, the vector with the highest confidence is selected for calculating the bounding box coordinates' offset.

Loss function of the model is composed of two parts: cross entropy (CE) between the computed object confidence distribution and the true confidence distribution for all grids, as well as the mean squared error (MSE) between the four values of the predicted box coordinates and their corresponding true values.

$$\mathrm{L} = \mathrm{CE}(\mathrm{p}, \mathrm{p}^*) + \frac{1}{4}\sum_{\mathrm{i}=1}^{4}(x_i - x_i^*)^2 \qquad (1)$$

where p is the confidence distribution of all grids calculated by Softmax function, and p^* is a one-hot vector with the actual target grid set to 1 and other grids set to 0. (x_1, x_2, x_3, x_4) represents X-axis and Y-axis coordinate of center point, width and height of bounding box, respectively; while x_i^* is the corresponding value of ground truth.

4 Experiments

4.1 Setups

We evaluate our model on four public visual grounding datasets: Flickr30k Entities [17], RefCOCO [18], RefCOCO+ [18], and RefCOCOg [19]. Flickr30k Entities, as an extension of Flickr30k dataset, contains a total of 31,783 images and annotates 275,775 terms mentioned in all 158,915 referring expressions.

Table I Experimental Results of Our Model Compared with Previous Methods

Accuracy	Flickr	RefCOCO		RefCOCO+		RefCOCOg		
	test	testA	testB	testA	testB	val-g	val-u	test-u
MMI[19]	-	0.6490	0.5451	0.5403	0.4281	0.4585	-	-
Attr[20]	-	0.7208	0.5729	0.5797	0.4620	-	-	-
SSG[7]	-	0.7651	0.6750	0.6214	0.4927	-	0.4778	0.5880
FAOA[8]	0.6704	0.7481	0.6759	-	-	-	-	-
Ours	**0.6872**	0.7483	**0.6869**	0.5767	0.4876	**0.4700**	**0.5797**	0.5703

RefCOCO, RefCOCO+, and RefCOCOg are based on MS COCO dataset, and consists of nearly 20K images, 50K object bounding box, and referring expressions of each object. All objects in the dataset fall within the 80 categories of common objects defined in the MS COCO dataset.

We report accuracy as the evaluation metric for visual grounding task, which depends on the Intersection over Union (IoU) ratio between the predicted box and the ground truth box. Testing sample is considered as a positive case only when IoU is greater than the threshold which is set to 0.5. Accuracy is calculated by the proportion of positive cases.

Our model is trained by RMSProp optimizer with decay rate of 0.9 and learning rate of 1×10^{-4}. In the fusion module, the number of attention heads of SA and GA blocks is set to 8, and the dimension of each attention head is 64. We train our model on Flickr30k Entities with the batch size of 64 for 30 epochs, and on RefCOCO, RefCOCO+ and RefCOCOg with the batch size of 32 for 50 epochs.

4.2 Quantitative Results

We report our experimental results on four datasets and the metrics of some other visual grounding models, shown in Table I. It can be seen that our model achieves better performance than previous methods on certain datasets.

As for feature fusion, current one-stage grounding models usually use simple methods such as concatenation, element-wise addition or multiplication to fuse visual and textual features. Here, FAOA [8] is used as the baseline model, whose fusion method is direct concatenation. The innovation point of our one-stage model, however, is that the visual features are weighted and updated by co-attention mechanism before the concatenation and fusion: text features as the reference information achieve attentive transformation by self-attention, and image features after multi-layer convolution are interacted with text features by multimodal guided-attention operations.

Table I shows that compared to the baseline FAOA, the metrics of our model improve on both the test set of Flickr30k Entities and the test set A and B of RefCOCO, with a maximum improvement of 1.68%. FAOA does not report the metrics on RefCOCO+ and RefCOCOg datasets, on which our model is trained and tested, but the accuracies are not so high. The reason may be that the description statements in RefCOCO+ and RefCOCOg are long sentences with comparisons of attributes between objects, while it is difficult for BERT to focus on the differences between objects in the sentences when extracting the overall representations of the text.

4.3 Qualitative Results

We qualitatively analyze the strengths and weaknesses of our model, shown as Figure 4. Green boxes are ground truth and blue boxes are the model's predictions. The first row shows examples of correct prediction. Figure 4 (a) successfully identifies "bride" from two people. (b) provides a positive example of predicting small targets. (c) identifies the correct target from a group of trees, showing that the model is able to utilize information about the size and state of objects in the query. The second row shows the wrong examples. (d) does not capture the key attribute of the long sentence "left". (e) identify the hat, but it is not sensitive to the color attribute, which leads to error. (f) doesn't identify the important noun "sunglasses", but only the head noun "man" is identified, leading to error. Green boxes are ground truth and blue boxes are the model's predictions.

Overall, it can be seen that our model can already work well on recognizing the specific object in the image based on text descriptions. In particular, the grounding effect is better for short sentences that do not involve comparisons between objects. In contrast, the model is still prone to generate errors for long sentences that contain detailed information about object attributes. This indicates that our model needs further deep fusion of image and text features, especially the latter one.

5 Conclusions

In this work, we have proposed a one-stage visual grounding model based on feature fusion. To improve the quality of fusion features, cross-modal co-attention mechanism is incorporated to interact and fuse the information from two modalities. The model is evaluated on Flickr30k Entities, RefCOCO, RefCOCO+ and RefCOCOg datasets with quantitative analysis, which shows that our model achieves promising results on these datasets and the co-attention mechanism for multimodal feature fusion is effective in visual grounding model. However, our model still has difficulty in dealing with long and detailed text queries. Our future work is to make better use of the intrinsic connection among inter-modal information in semantic space.

Figure 4 Examples for Qualitative Analysis.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62076032. The authors would like to thank the editors and anonymous reviewers for their valuable comments on improving the final version of this paper.

References

[1] Wu C, Liu J, Wang X, et al. Differential networks for visual question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 8997-9004.

[2] Shi Y, Liu Y, Feng F, et al. S2TD: A tree-structured decoder for image paragraph captioning[M]//ACM Multimedia Asia. 2021: 1-7.

[3] Liu Y, Shi Y, Feng F, et al. Improving Image Paragraph Captioning with Dual Relations[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022: 1-6.

[4] Feng F, Wang X, Li R. Cross-modal retrieval with correspondence autoencoder[C]//Proceedings of the 22nd ACM international conference on Multimedia. 2014: 7-16.

[5] Zhang H, Niu Y, Chang S F. Grounding referring expressions in images by variational context[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4158-4166.

[6] Yu L, Lin Z, Shen X, et al. Mattnet: Modular attention network for referring expression comprehension[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1307-1315.

[7] Chen X, Ma L, Chen J, et al. Real-time referring expression comprehension by single-stage grounding network[J]. arXiv preprint arXiv:1812.03426, 2018.

[8] Yang Z, Gong B, Wang L, et al. A fast and accurate one-stage approach to visual grounding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4683-4693.

[9] Liao Y, Liu S, Li G, et al. A real-time cross-modality correlation filtering method for referring expression comprehension[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10880-10889.

[10] Deng J, Yang Z, Chen T, et al. Transvg: End-to-end visual grounding with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1769-1779.

[11] Kamath A, Singh M, LeCun Y, et al. MDETR-modulated detection for end-to-end multi-modal understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1780-1790.

[12] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[13] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[14] Lin T Y, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[J]. IEEE Computer Society, 2017.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[16] Yu Z, Yu J, Cui Y, et al. Deep modular co-attention networks for visual question answering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6281-6290.

[17] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2641-2649.

[18] Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions[C]//European Conference on Computer Vision. Springer, Cham, 2016: 69-85.

[19] Mao J, Huang J, Toshev A, et al. Generation and Comprehension of Unambiguous Object Descriptions[J]. IEEE, 2016.

[20] Liu J, Wang L, Yang M H. Referring expression generation and comprehension via attributes[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4856-4864.