SCIENCE CHINA
Information Sciences

• LETTER •

October 2025, Vol. 68, Iss. 10, 200111:1–200111:2
https://doi.org/10.1007/s11432-024-4312-9

Special Topic: Large Multimodal Models

Progressive language-aware encoding and decoding for referring expression comprehension

Yichen ZHAO[2], Yaxiong CHEN[2,3,4*], Yi RONG[1,2] & Shengwu XIONG[2,3*]

[1]*School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China*
[2]*Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China*
[3]*Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China*
[4]*Chongqing Research Institute, Wuhan University of Technology, Chongqing 401122, China*

Received 13 May 2024/Revised 2 September 2024/Accepted 23 November 2024/Published online 17 September 2025

Citation Zhao Y C, Chen Y X, Rong Y, et al. Progressive language-aware encoding and decoding for referring expression comprehension. Sci China Inf Sci, 2025, 68(10): 200111, https://doi.org/10.1007/s11432-024-4312-9

Referring expression comprehension (REC) [1] seeks to locate the visual object indicated by a referring expression, relying on multimodal fusion and reasoning for accurate interpretation and response to the referred object. This process builds a bridge between human language and the visual content of the physical world, with great potential for intelligent navigation and natural human-computer interaction.

Early research treated REC as an extension of conventional object detection, focusing on introducing multimodal interaction into prevailing two-stage [2] or one-stage [3] object detection frameworks. However, these methods rely heavily on the quality of the generated candidates (e.g., region proposals and predefined anchor boxes) and the heuristics used to assign targets to these candidates. Recently, transformer-based methods [1, 4, 5] have greatly advanced the development of REC, which exploits transformers to facilitate image-text alignment without the need for predefined candidates. A comprehensive overview of related work can be found in Appendix A. Despite promising results, the limitation arises from the prevailing practice of employing language-agnostic visual backbones for feature extraction in existing methods. This approach restricts the semantic representation of visual features, resulting in suboptimal performance in multimodal fusion and reasoning. Moreover, fusion and reasoning modules commonly wear multiple hats, being trained from scratch on limited data, which increases the difficulty of model training and optimization.

In this study, we propose an REC framework called PLAED, which is based on semantic-aware visual encoding to generate more reasonable visual features and progressive two-stage decoding to achieve efficient grounding. In addition, we implement informative sample learning to facilitate the model's understanding of complex multimodal relationships. Specifically, the semantic-aware visual encoding is implemented by dynamic semantic awareness gate (DSAG) and semantics-boost transformer (SBT). The two gradually improve the rationality and accuracy of visual features by focusing on crucial visual elements mentioned by expressions. Based on enhanced visual features, a progressive two-stage decoder is introduced to establish cross-modal deep associations, effectively capturing the grounded object information. Extensive experiments on four benchmarks confirm that the proposed model is effective and achieves state-of-the-art performance.

Methods. PLAED takes an image and a referring expression as input and generates a bounding box indicating the referred object. As depicted in Figure 1, the input images and expressions are encoded into visual and linguistic features using dedicated backbone networks. For the visual backbone, ResNet is adopted to extract the visual feature $\mathcal{I}_v \in \mathbb{R}^{H\times W\times D_v}$, where H and W denote height and width respectively, D_v is the channel number. Then, the DSAG and SBT are developed to actively perceive relevant semantic cues based on text priors, improving the rationality and accuracy of visual features. For the language backbone, BERT is selected to extract the linguistic feature $\mathcal{F}_e \in \mathbb{R}^{N_e\times D_e}$, where N_e and D_e represent the number of tokens and channels, respectively. The first token in $\mathcal{F}_e$ is the sentence token [CLS], defined as $f_e \in \mathbb{R}^{D_e}$, which reflects contextual information of the expression.

In addition, a progressive two-stage decoder built on the standard transformer is introduced for cross-modal decoding. This decoder adopts a two-stage strategy, progressing from "self-attention" to "cross-attention" to better capture the referred object information. More specifically, the visual and linguistic features are concatenated into a sequence and fed into the progressive two-stage decoder. In the first stage, the self-attention mechanism is utilized to self-associate each element in the sequence for integrating information within and between modalities. In the second stage, the cross-attention mechanism is deployed to guide the semantic aggregation of visual cues, with linguistic features acting as queries. Here, the second stage utilizes linguistic features to offer detailed guidance on the visual scene, capturing visual elements aligned with the text. Finally, the linguistic features are averaged and used directly to regress the bound-

* Corresponding author (email: chenyaxiong@whut.edu.cn, xiongsw@whut.edu.cn)

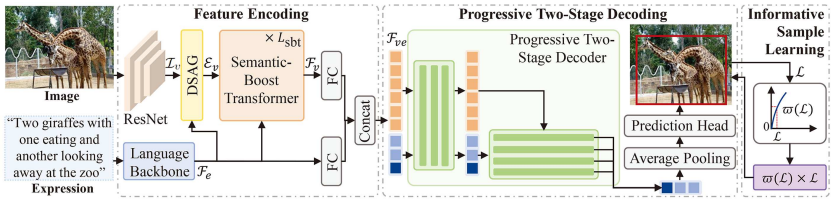

Figure 1 (Color online) Framework of PLAED involves feature encoding, progressive two-stage decoding, and informative sample learning. ResNet and the language backbone extract visual and linguistic features, respectively. DSAG and SBT are then employed to extract text-related discriminative features. These features are concatenated into a sequence, to which learnable positional embeddings are added. The sequence is then input into the two-stage decoder for multimodal fusion and reasoning. Lastly, the prediction head regresses a bounding box of the input expression. Informative sample learning is introduced to drive the model to adaptively prioritize informative samples.

ing box coordinates of the referred object. The details of our method are provided in Appendix B.

Experiments. We evaluate PLAED on four publicly available REC datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame. Consistent with the metric employed in prior studies [1–3], predictions are deemed accurate when the intersection over union (IoU) between the ground truth and predicted bounding boxes exceeds 0.5. The dataset and implementation details are summarized in Appendix C.

We report the comparative results of PLAED with state-of-the-art methods on four publicly available REC datasets. These methods are classified into two-stage and one-stage methods. Among them, one-stage methods are further subdivided into transformer-based methods and non-transformer-based methods. A more detailed comparative analysis can be found in Appendix C.3.

PLAED (ResNet50) with the default setting (3-layer first-stage decoder, 4-layer second-stage decoder) achieves the best performance compared to these two-stage and non-transformer-based methods. PLAED, employing ResNet50 as the visual network, achieves an 87.54% accuracy on the RefCOCO testA set. Upon switching to the more powerful ResNet101, the accuracy increased to 87.63%. This enhancement enables PLAED to outperform the top two-stage method, Ref-NMS [2] by 5.66% on the RefCOCO val set. PLAED brings an absolute improvement of 4.72 percentage points compared to the top non-transformer-based method, namely HFRN [3], showcasing its effectiveness.

Transformers exhibit high capacity and flexibility in feature modeling, enabling them to align images and text for cross-modal fusion and target grounding. TransVG [1] employs the transformer to process the concatenated sequence of images and expressions, with the goal of capturing their correlation and minimizing modal differences. While the transformer-based encoder is effective, it performs both intra-modal self-attention and inter-modal cross-attention, which raises computational and optimization costs. In contrast, our PLAED adopts an encoder-decoder paradigm, which effectively alleviates the computational pressure of the encoder by decoupling modal relationship modeling from target grounding. Notably, PLAED (ResNet101) outperforms TransVG (Resnet101) by absolute margins of 3.46% on the RefCOCO val set, 8.96% on the RefCOCO+ val set, and 7.63% on the RefCOCOg val-g set. QRNet dynamically computes visual attention dependent on the input expression at both the spatial and channel levels of the feature maps generated by the visual backbone. VGTR employs linguistic features to enhance the query tokens in the visual backbone, allowing for the aggregation of visual cues from the input expression. Despite their impressive performance, these guidance strategies depend on the overall information from the text, overlooking the in-depth exploration of fine-grained information between images and text. In this study, we propose the DSAG to intensify attention on key visual elements via fine-grained modal interactions, thus improving the semantic representation of visual features. Furthermore, we introduce SBT, enabling precise interpretation of key elements in images by seamlessly integrating global visual and linguistic features. In particular, our PLAED (ResNet101) achieves significant performance improvements on two more challenging datasets, such as RefCOCOg, achieving an accuracy of 76.25% on the RefCOCOg val-u set.

Conclusion. In this study, we propose a novel REC framework named PLAED with the objective of achieving more effective multimodal fusion and reasoning. The core of PLAED is centered around semantic-aware visual encoding, aiming to improve the accuracy and rationality of visual features, coupled with a progressive two-stage decoding to achieve efficient target grounding. Furthermore, the framework incorporates informative sample learning to bolster the model's comprehension of multimodal relationships. Extensive experiments conducted on four benchmarks have been carried out to validate the effectiveness of our method. In the future, we aim to enhance the transformer-based multimodal fusion method to improve multimodal alignment and address semantic mismatches between images and text.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2022ZD0160604), National Natural Science Foundation of China (Grant No. 62176194), and Hainan Province "Nanhai New Star" Technology Innovation Talent Platform Project (Grant No. NHXXRCXM202361).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Deng J, Yang Z, Chen T, et al. TransVG: end-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, 2021. 1769–1779
2 Chen L, Ma W, Xiao J, et al. Ref-NMS: breaking proposal bottlenecks in two-stage referring expression grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 1036–1044
3 Qiu H, Li H, Wu Q, et al. Language-aware fine-grained object representation for referring expression comprehension. In: Proceedings of the 28th ACM International Conference on Multimedia, Seattle, 2020. 4171–4180
4 Shi F, Gao R, Huang W, et al. Dynamic MDETR: a dynamic multimodal transformer decoder for visual grounding. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 1181–1198
5 Wu J, Wu C, Wang F, et al. Improving visual grounding with multi-scale discrepancy information and centralized-transformer. Expert Syst Appl, 2024, 247: 123223