PolyFormer: Referring Image Segmentation as Sequential Polygon Generation

摘要

在 PolyFormer 这项工作中，不直接预测像素级，而是将参考图像分割问题表述为连续多边形生成，再将预测出的多边形随后可以转换成分割掩码。它将图像补丁和文本查询标记序列作为输入，并以自回归方式输出多边形顶点序列。自回归输出多边形顶点序列。本工作在 PolyFormer 的基础上，使用更加轻量化的视觉特征提取骨干网络，再将多模态特征融合过程替换成线性层微调，并复现其可视化界面。

关键词：轻量化；微调；多模态

1 引言

以往的参考图像分割 (RIS) 结合了视觉语言理解和实例分割，目的是在给定自然语言的情况下定位物体的分割遮罩。它将传统的对象分割从固定数量的预定义类别扩展到由自由形式语言描述的任何概念。首先从图像和文本输入中提取特征，然后将多模态特征融合在一起，预测掩码。大多数实例分割模型都依赖于密集的二进制分类网络来确定物体的空间布局。这种像素间的预测是卷积运算的首选，但它忽略了输出预测之间的结构。例如，每个像素的预测都与其他像素无关。PolyFormer 采用了序列-序列（seq2seq）框架，并提出了用于指代图像分割的多边形转换器。将图像片段序列和文本查询标记作为输入、并自回归输出多边形顶点序列。由于每个顶点预测都是以前面所有预测顶点为条件的，因此输出的预测结果不再相互独立。由于文章中使用的视觉特征提取器为 swim—transformer，本工作采用更加轻量化的骨干网络，同时对 concat 的线性层进行微调，将图像特征映射到文本领域。

2 相关工作

2.1 参考图像分割（RIS）

参考图像分割（RIS）旨在提供图像中目标对象的像素级定位通过引用表达式来描述。以前的作品主要集中在两个方面：（1）视觉和语言特征提取；（2）多模态特征融合。对于在特征提取方面，已经有了丰富的工作，包括使用 CNNs [1] [2] 和 transformer 模型 [3]。在特征融合方面的努力已经探索了特征连接 [4]、连接机制 [5] 和多模态 transformer。SeqTR [6] 采用了 transformer 模型来顺序生成多边形顶点。但是，SeqTR 只能生成 18 个顶点的单个多边形分割掩码，无法用复杂的对象勾勒轮廓形状和遮挡。

2.2 参考表达理解（REC）

参考表达理解（REC）预测紧密包围中目标对象的边界框对应于参考表达的图像。现在的工作包括两种阶段性的方法 [7]，基于区域候选排名和直接预测目标的一阶段方法边界框。几篇论文 [8] 探讨了 REC 和 RIS 的多任务学习，因为它们是两个密切相关的任务。然而，MCN 和 RefTR 需要特定任务。尽管 SeqTR 在统一的框架中将这两个任务视为点预测问题，结果表明，与单任务变体相比，多任务监督会降低性能。

2.3 序列到序列（seq2seq）

序列到序列（seq2seq）建模已经实现在自然语言处理（NLP）方面取得了许多成功 [9]。Sutskever 等人提出了一个开创性的基于 LSTM [10] 的 seq2seq 模型。Raffel 等人开发了 T5 模型，以统一各种任务包括翻译、问题回答和文本到文本框架中的分类。进一步表明扩展语言模型显著提高了。受 NLP 成功的启发，最近计算机视觉和视觉语言的研究也开始了探索各种任务的 seq2seq 建模 [11] [12]。然而，他们执行几何定位任务作为一个分类问题，即将坐标量化为离散箱，并将坐标预测为框。这使他们能够将所有任务统一到一个简单的统一 seq2seq 框架中，但忽略了任务。在 PolyFormer 中，几何定位是公式化的作为一个更合适的回归任务，预测连续不带量化的坐标。

3 本文方法

3.1 本文方法概述

文章方法将图像数据和文本数据分别输入图像特征编码器和文本特征编码器，再将两个特征结合输入多模态特征编码器，再通过文章提出的基于回归的 Transformer 解码器，输出得到一系列点坐标围成的序列多边形。如图 1所示。

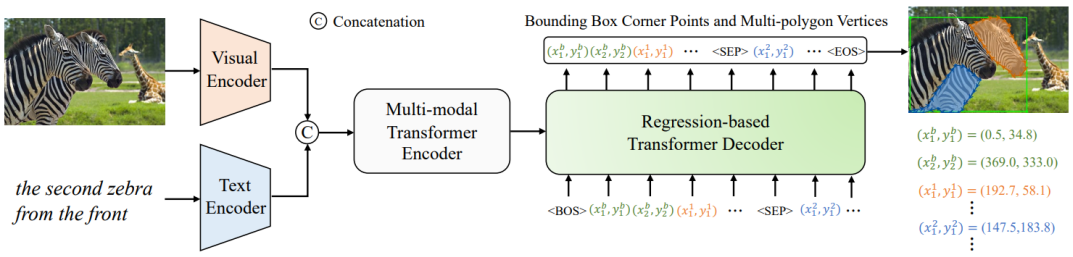


图 1. 方法示意图

3.2 特征提取模块

特征提取模块如图 2所示：
图像编码器：采用 Swin Transformer 来提取图像的第四阶段，作为视觉表示。
文本编码器：利用 BERT 来提取输入的带有 L 个单词的语言，得到单词特征。
多模态编码器：由 N 个 Transformer 层组成，输出多模态特征。

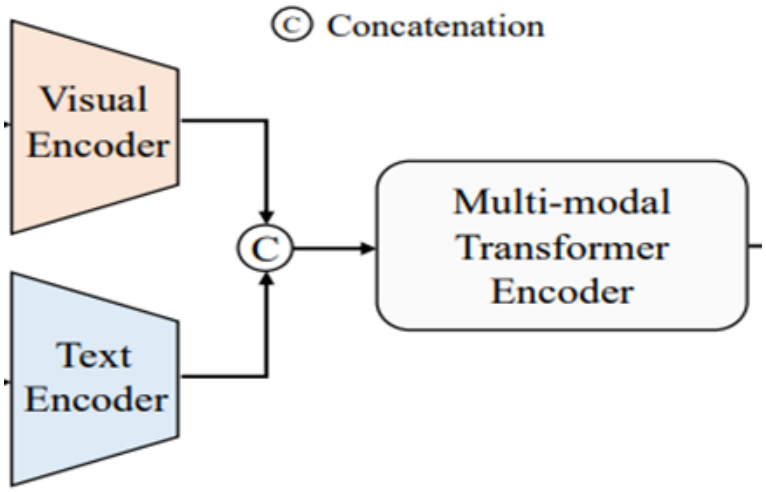


图 2. 特征提取模块

基于回归的 Transformer 编码器如图 3所示：之前的视觉 seq2seq 方法将连续坐标 x 量化为离散的 bin [x]，不可避免地引入了量化误差 |x - [x]|。他们将坐标定位表述为一个分类问题，这对于几何定位来说是次优的。为了解决这个问题，文章提出了一种基于回归的解码器，它不使用量化，而是直接预测连续坐标值（即 x 而不是 [x]）。2D 坐标嵌入抽象为图 4所示：计算过程可表示公式 (1) 所示：

$$\begin{aligned} e_{(x,y)} =&(\bar{x}-x)(\bar{y}-y)\cdot e_{(\underline{x},y)} + (x-\underline{x})(\bar{y}-y)\cdot e_{(\bar{x},y)} + \\ &(\bar{x}-x)(y-\underline{y})\cdot e_{(x,\bar{y})} + (x-\underline{x})(y-\underline{y})\cdot e_{(\bar{x},\bar{y})} \end{aligned} \tag{1}$$

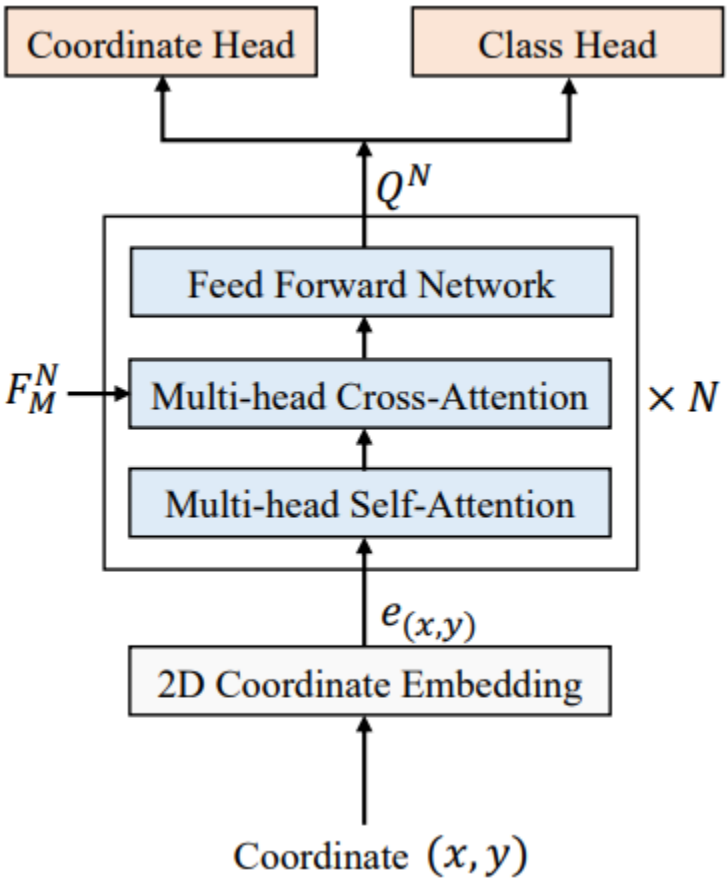


图 3. 基于回归的 Transformer 编码器

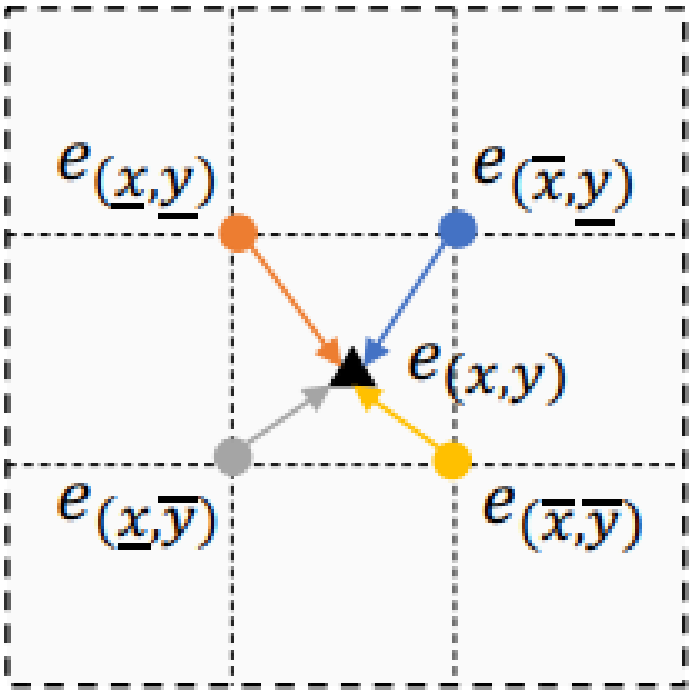


图 4. 2D 坐标嵌入

3.3 损失函数定义

$$L_t = \lambda_t L_{coo}\left(\left(x_t, y_t\right), \left(\hat{x}_t, \hat{y}_t\right) \mid I, T, \left(x_i, y_i\right)_{i=1:t-1}\right) \cdot \mathbb{I}\left[p_t == < \text{COO} >\right] + \lambda_{cls} L_{cls}\left(p_t, \hat{p}_t \mid I, T, p_{1:t-1}\right) \tag{2}$$

其中: $\lambda_t = \begin{cases} \lambda_{\text{box}}, & t \leq 2 \\ \lambda_{\text{poly}}, & \text{otherwise} \end{cases}$

L_{coo} 为 L_1 回归损失，L_{cls} 为 L_1 标签平滑 cross-entropy 损失。回归损失 L_1 仅用于计算坐标 tokens，λ_{box} 和 λ_{poly} 为相应的 tokens 权重。L_t 为一个序列中所有 tokens 的损失总和。

4 复现细节

4.1 与已有开源代码对比

这篇文章已有开源代码，由于文章主要贡献是设计一个基于回归的解码器，同时文章在做实验时，进行的实验对比对编码器都是采用较为复杂的网络结构。本文工作主要在此基础上替换掉图像的特征提取模块，使用更加轻量化的 Efficient-Net，加快推理的速度，将文章提供的预训练模型，将除了 visual encoder 和 concatenate 的权重冻结，使用新的骨干网络进行微调，同时改进了 concatenate 的方式。

4.2 实验环境搭建

4.3 界面分析与使用说明

如图 5所示：在左侧上传图片与文本描述，如 “the blue vase on the left”，点击提交进行推理。结果如图 6右侧所示，推理给出多边形序列掩码。

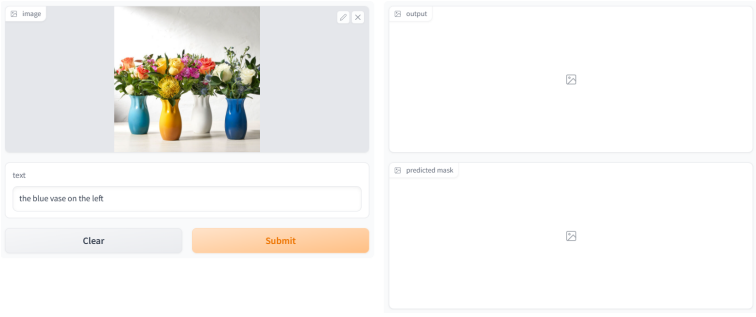


图 5. 操作界面示意

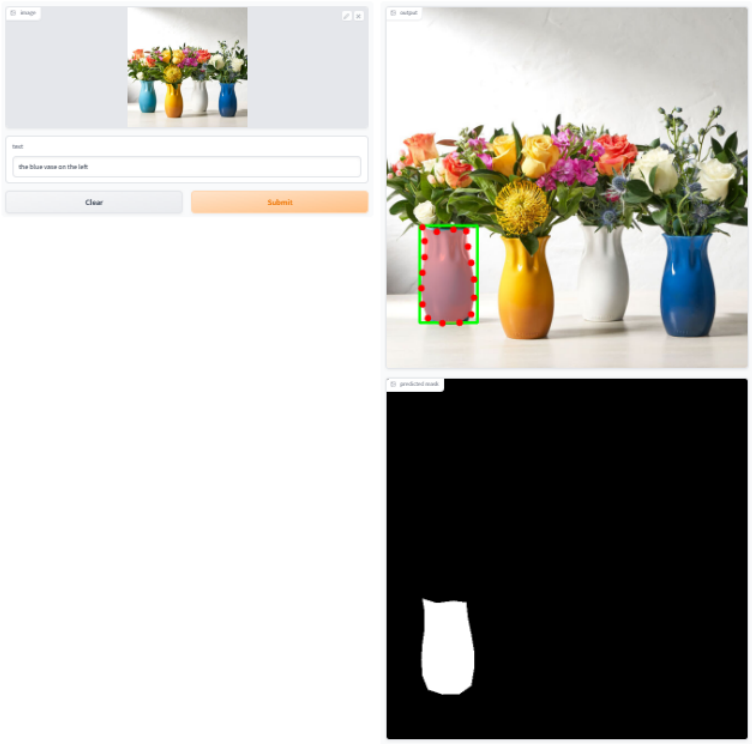


图 6. 操作界面示意

4.4 创新点

通过替换骨干网络将模型进行轻量化，验证基于回归的 transformer 是否具有优越性，并使用可训练的线性层将将视觉序列和单词序列投影到同一 embedding 空间，而不是简单的counstrncate。

5 实验结果分析

5.1 数据集与评价指标

5.1.1 数据集

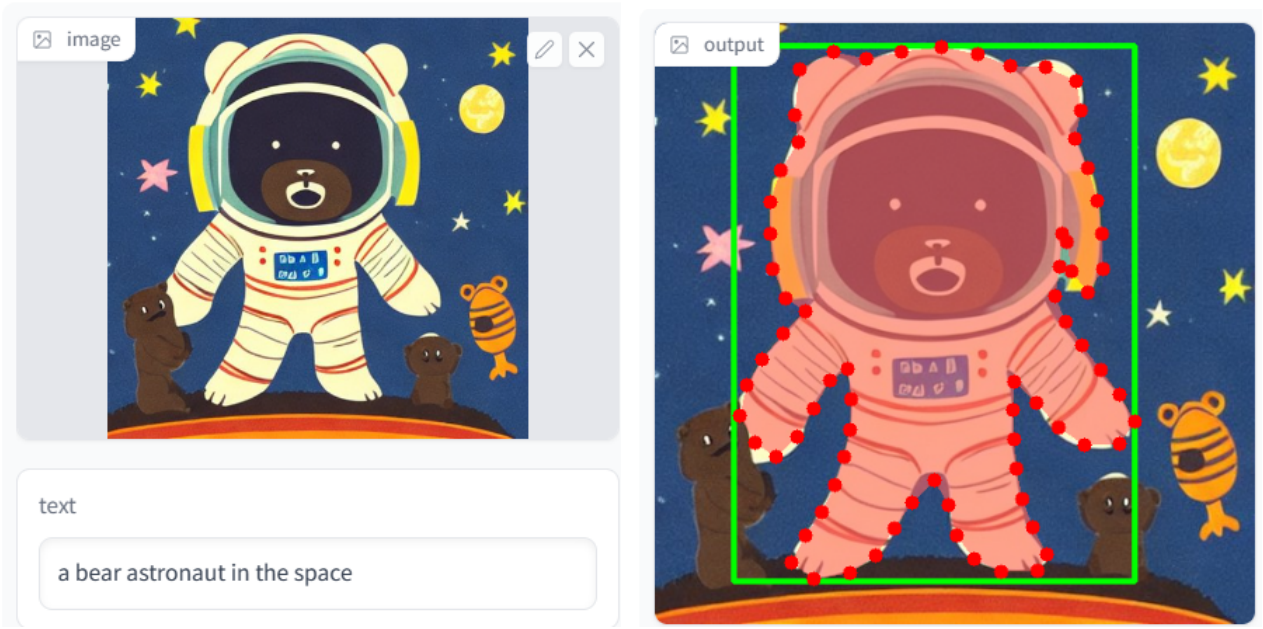
实验在 RIS 和 REC 的四个主要基准上进行：RefCOCO [13]、RefCOCO+、RefCOCOg [14] 和 ReferIt [15]。RefCOCO 有 142,209 个注释表达式，涉及 19,994 幅图像中的 50,000 个对象；RefCOCO+ 包含 141,564 个表达式，涉及 19,992 幅图像中的 49,856 个对象。与 RefCOCO 相比，RefCOCO+ 的引用表达式中没有位置词，因此更具挑战性。RefCOCOg 包含 85,474 个引用表达式，涉及 26,711 幅图像中的 54,822 个对象。这些引用表达式是在 Amazon Mechanical Turk 上收集的，因此描述更长、更复杂（平均 8.4 个单词，而 RefCOCO 和 RefCOCO+ 为 3.5 个单词）。ReferIt 包含 130,364 个表达式，涉及从 SAIAPR-12 数据集收集的 19,997 幅图像中的 99,296 个对象。对 RefCOCOg 采用 UMD 拆分法，对 ReferIt 采用伯克利拆分法。

5.1.2 评价指标

使用平均交叉点联合度（mIoU）作为 RIS 的评价指标。为了进行公平比较，在与只报告 oIoU 结果的论文进行比较时，还使用了总体交集-过联合（oIoU）。此外，还在 REC 任务中对 PolyFormer 进行了评估，因为它是 RIS 和 REC 任务的统一框架。采用了标准指标 Precision@0.5，即如果预测与地面实况框的交集（IoU）大于 0.5，则认为预测正确。

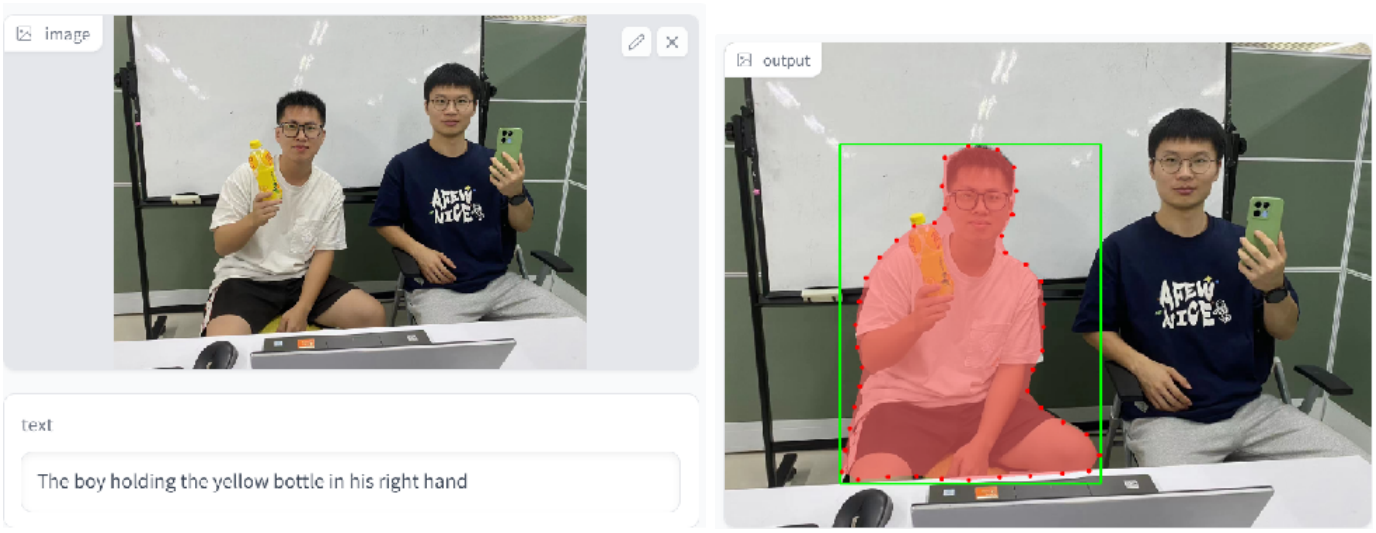
5.2 复现结果与对比分析

通过复现 PolyFormer 的整体框架，包括两种语义的编码器，特征映射，基于回归的解码器等，得出了以下可视化结果。



(a) 原始图像　　(b) 结果图像

图 7. 卡通图像实验



(a) 原始图像　　(b) 结果图像

图 8. 真实世界实验

由可视化结果可知，不管是卡通图像或者是真实世界图像，PolyFormer 都能通过理解文本语义，生成一系列坐标点，将目标掩码得出，且效果良好。

表 1. 在三个参考图像分割基准上与最先进方法的比较。

	Method	**Visual**	**Text**	**RefCOCO**	**RefCOCO+**	**RefCOCOg**
oIoU	STEP	RN101	Bi-LSTM	57.97	40.41	-
	BRINet	RN101	LSTM	59.21	42.11	-
	CMPC	RN101	LSTM	59.64	43.23	-
	LSCM	RN101	LSTM	59.55	43.5	-
	CMPC+	RN101	LSTM	60.82	43.47	-
	MCN	DN53	Bi-GRU	59.71	44.69	49.4
	EFN	WRN101	Bi-GRU	59.67	43.01	-
	BUSNet	RN101	Self-Att	61.39	44.13	-
	CGAN	DN53	Bi-GRU	62.07	44.06	51.69
	LTS	DN53	Bi-GRU	63.08	48.02	54.25
	ReSTR	ViT-B	Transformer	64.45	48.27	-
	PolyFormer-B	Swin-B	BERT-base	71.06	59.33	69.05
	PolyFormer-L	Swin-L	BERT-base	73.25	61.87	70.19
mIoU	VLT	DN53	Bi-GRU	62.73	49.36	56.65
	CRIS	RN101	GPT-2	66.1	53.68	60.36
	SeqTR	DN53	Bi-GRU	69.82	58.97	65.74
	RefTr	RN101	BERT-base	70.87	59.4	67.39
	LAVT	Swin-B	BERT-base	70.94	59.23	63.62
	PolyFormer-B	Swin-B	BERT-base	73.22	64.64	69.88

表 2. 更换多种骨干网络对比

Visual backbone	**Ref-DAVIS17**
ResNet-50	40.2
ResNet-50	51.5
ResNet-101	56.4
Swin-B	60.9
Swin-L	61.5

在多种数据集中, PolyFormer 都优于其他算法, 通过替换骨干网络将模型进行轻量化, 验证了基于回归的 transformer 具有优越性。

6 总结与展望

在这项工作中，提出了 PolyFormer，一个用于指代图像分割和指代表达理解的简单而统一的框架。它是一个序列-序列框架，可以自然融合多模态特征作为输入序列和多任务预测作为输出序列。此外，还设计了一种新颖的基于回归的解码器，以生成无量化误差的连续二维坐标。PolyFormer 在 RIS 和 REC 方面取得了具有竞争力的结果，并对未知场景显示出良好的泛化能力。相信这个简单的框架可以扩展到 RIS 和 REC 以外的其他任务。

参考文献

[1] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[3] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022.

[4] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017.

[5] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019.

[6] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

[7] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018.

[8] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[11] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.

[12] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022.

[13] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[14] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.