

PIX2SEQ：一种面向目标检测的语言建模框架

陈挺、索拉布·萨克塞纳、拉拉·李、大卫·J·弗利特、杰弗里·辛顿 谷
歌研究院，Brain团队

摘要

我们提出了 *Pix2Seq*, 一个简单而通用的目标检测框架。与现有方法不同, 后者显式地整合了任务相关的先验知识, 而我们将目标检测视为一种基于观察到的像素输入的条件语言建模任务。目标描述(如边界框和类别标签)被表达为离散标记的序列, 我们训练神经网络来感知图像并生成所需的序列。我们的方法主要基于这样一种直觉: 如果一个神经网络已经知道目标的位置和内容, 我们只需教会它如何将这些信息读取出来。除了使用任务特定的数据增强外, 我们的方法对任务本身的假设极少, 但在具有挑战性的COCO数据集上, 与高度专业化且经过充分优化的检测算法相比, 仍取得了具有竞争力的结果。¹

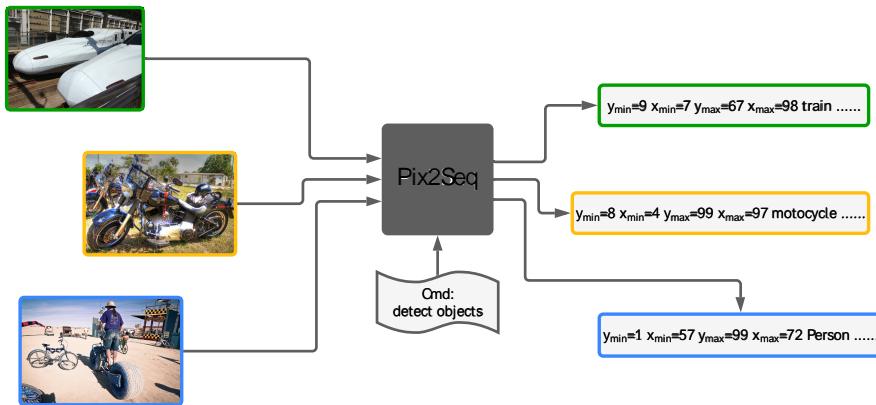


图1: Pix2Seq框架在目标检测中的示意图。神经网络感知图像并生成一系列对应于边界框和类别标签的标记。

1 引言

视觉目标检测系统旨在识别并定位图像中所有预定义类别的物体。检测到的物体通常由一组边界框及对应的类别标签描述。鉴于该任务的复杂性, 现有大多数方法(如Girshick, 2015; Ren et al., 2015; He et al., 2017; Lin et al., 2017b; Carion et al., 2020)都经过精心设计且高度定制化, 在架构选择和损失函数中融入了大量先验知识。例如, 许多架构专门针对边界框的使用进行优化(如采用区域提议(Girshick, 2015; Ren et al., 2015)和RoI池化(Girshick et al., 2014; He et al., 2017))。另一些方法则与用于物体绑定的对象查询机制紧密关联(Carion et al., 2020)。损失函数同样常针对边界框的使用进行专门设计, 例如框回归(Szegedy et al., 2013; Lin et al., 2017b)、基于集合的匹配(Erhan et al., 2014; Carion et al., 2020), 或通过整合{v*}

Correspondence to: iamtingchen@google.com

¹Code and checkpoints available at <https://github.com/google-research/pix2seq>.

特定性能指标，如边界框的交并比（Rezatofighi等人，2019年）。尽管现有系统已在众多领域找到应用，从自动驾驶汽车（Sun等人，2020年）到医学图像分析（Jaeger等人，2020年），再到农业（Sa等人，2016年），但其专业性和复杂性使得它们难以集成到更大的系统中，或泛化至与通用智能相关的更广泛任务阵列。

本文提出了一种新方法，其核心思想是：若神经网络已掌握物体位置与类别的识别能力，我们只需教会它如何输出这些信息。通过让模型学习“描述”物体，它能将“语言”与像素观察结果建立关联，从而形成有效的物体表征。我们通过Pix2Seq框架实现了这一构想（见图1）。给定一张图像，我们的模型会生成与物体描述（如物体边界框和类别标签）对应的离散标记序列，其原理类似于图像描述生成系统（Vinyals等人，2015b；Karpathy & Fei-Fei，2015；Xu等人，2015）。本质上，我们将物体检测任务转化为基于像素输入的语言建模任务，其模型架构与损失函数具有通用性且相对简单，无需专门针对检测任务进行设计。因此，该框架可轻松扩展至不同领域或应用场景，或整合到支持通用智能的认知系统中——通过语言接口为各类视觉任务提供服务。

为了利用Pix2Seq解决检测任务，我们首先提出了一种量化和序列化方案，将边界框和类别标签转化为离散标记序列。随后，我们采用编码器-解码器架构来感知像素输入并生成目标序列。目标函数简化为在像素输入及前序标记条件下的标记最大似然。尽管该架构与损失函数均与任务无关（无需预设目标检测的先验知识，如边界框），我们仍可通过下文提出的序列增强技术融入任务特定先验知识——该技术会在训练时同步修改输入与目标序列。大量实验表明，这一简洁的Pix2Seq框架在COCO数据集上能取得与高度定制化、成熟方法（如Faster R-CNN（Ren等人，2015）和DETR（Carion等人，2020））相竞争的结果。通过在更大规模的目标检测数据集上预训练模型，其性能还可进一步提升。

2 PIX2SEQ框架

在提出的Pix2Seq框架中，我们将目标检测任务转化为基于像素输入的语言建模任务（图1）。该系统包含四个主要组成部分（图2）：

- *Image Augmentation*在训练计算机视觉模型时，通常会采用图像增强技术来丰富固定的训练样本集（例如，通过随机缩放和裁剪等方式）。
- *Sequence construction & augmentation*由于图像中的对象标注通常以边界框和类别标签的set形式表示，我们将其转换为离散标记的sequence。
- *Architecture*我们采用编码器-解码器模型，其中编码器感知像素输入，解码器则逐步生成目标序列（每次一个标记）。
- *Objective/loss function*该模型训练的目标是最大化在给定图像及先前标记条件下标记的对数似然（采用softmax交叉熵损失函数）。

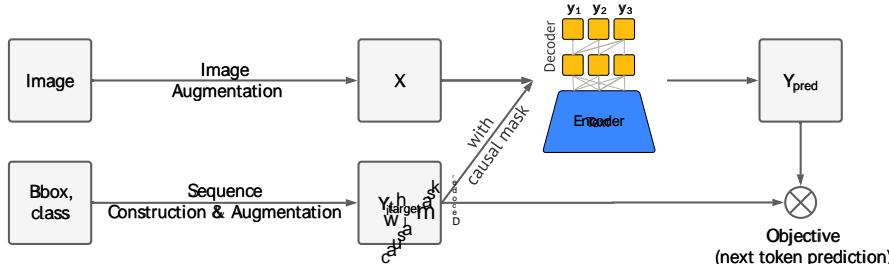


图2：Pix2Seq学习框架的主要组成部分。

2.1 从对象描述构建序列

在常见的物体检测数据集中，如Pascal VOC (Everingham等人, 2010)、COCO (Lin等人, 2014) 和OpenImages (Kuznetsova等人, 2020)，图像包含数量不定的物体，以边界框和类别标签的集合形式表示。在Pix2Seq中，我们将它们表达为离散标记的序列。

虽然类别标签自然以离散标记的形式表达，但边界框则不然。边界框由其两个角点（即左上角和右下角）确定，或通过中心点加高度和宽度来定义。我们提出将用于指定角点 x, y 坐标的连续数值离散化（若采用另一种边界框格式，则对高度和宽度同样处理）。具体而言，一个对象被表示为五个离散标记的序列，即 $[y_{\min}, x_{\min}, y_{\max}, x_{\max}, c]$ ，其中每个连续的角点坐标被均匀离散化为 $[1, n_{\text{bins}}]$ 之间的整数，而 c 为类别索引。所有标记共享同一词汇表，因此词汇表大小等于分箱数+加上类别数。这种边界框的量化方案使我们能够使用较小的词汇表，同时实现高精度。例如，一幅 600×600 的图像仅需600个分箱即可实现零量化误差。这远小于现代语言模型通常32K或更大的词汇量 (Radford等人, 2018; Devlin等人, 2018)。图3展示了不同量化级别对边界框定位的影响。

将每个物体描述表达为简短的离散序列后，我们接下来需要将多个物体描述序列化，以形成给定图像的单一序列。由于物体顺序对检测任务本身并不重要，我们采用了随机排序策略（每次展示图像时随机打乱物体顺序）。我们还探索了其他确定性排序策略，但我们假设，只要具备强大的神经网络和自回归建模能力（网络能够学习根据已观察到的物体来建模剩余物体的分布），随机排序的效果将与任何确定性排序策略同样出色。

最后，由于不同图像中的物体数量往往不同，生成的序列也会有不同的长度。因此，我们引入了一个EOS标记来表示序列的结束。图4展示了采用不同排序策略时的序列构建过程。

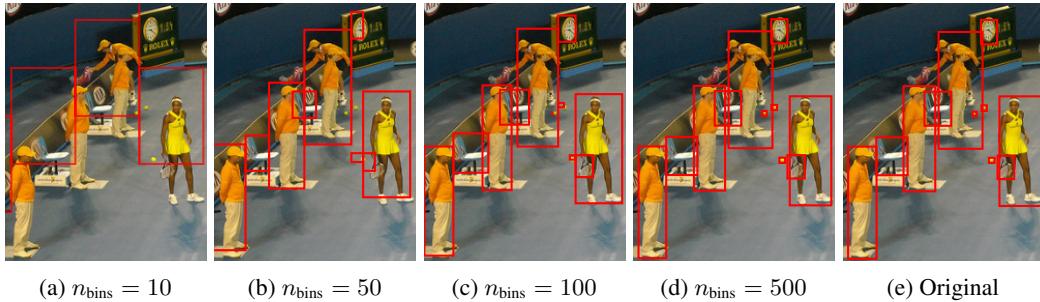


图3：在 480×640 的图像上应用所提出的边界框离散化方法。为更清晰展示，仅显示图像的四分之一区域。即使采用较少的区间数量，例如500个区间（ ~ 1 像素/区间），该方法对小物体也能实现高精度检测。



图4：使用 $n_{\text{bins}} = 1000$ 构建序列的示例，0为EOS标记。

2.2 架构、目标与推理

将我们从物体描述中构建的序列视为一种“方言”，我们转而采用在语言建模中已被证明有效的通用架构和目标函数。

架构 我们采用编码器解码器架构。编码器可以是通用的图像编码器，负责感知像素并将其编码为隐藏表示，例如ConvNet (LeCun等人, 1989; Krizhevsky等人, 2012; He等人, 2016)、Transformer (Vaswani等人, 2017; Dosovitskiy等人, 2020) 或二者的结合 (Carion等人, 2020)。在生成部分，我们使用Transformer解码器，该结构在现代语言建模中广泛应用 (Radford等人, 2018; Raffel等人, 2019)。它每次生成一个标记，基于先前生成的标记和编码后的图像表示进行条件生成。由于所有标记均通过单一词汇表的softmax生成，这消除了现代目标检测器架构中的复杂性和定制化需求，例如边界框提议和回归过程。

目标与语言建模类似，Pix2Seq的训练目标是在给定图像及先前标记的条件下预测标记，采用最大似然损失函数，即

$$\text{maximize} \sum_{j=1}^L \mathbf{w}_j \log P(\tilde{\mathbf{y}}_j | \mathbf{x}, \mathbf{y}_{1:j-1}), \quad (1)$$

其中 \mathbf{x} 是给定的图像， \mathbf{y} 和 $\tilde{\mathbf{y}}$ 是与 \mathbf{x} 相关联的输入和目标序列， L 是目标序列的长度。在标准的语言建模设置中， \mathbf{y} 和 $\tilde{\mathbf{y}}$ 是相同的，但它们也可以不同（如我们后续增强序列构建中的情况）。此外， \mathbf{w}_j 是序列中第 j 个令牌的预设权重。我们设定 $\mathbf{w}_j = 1, \forall j$ ，但也可以根据令牌类型（如坐标令牌与类别令牌）或对应对象的大小来加权。

在推理阶段，我们从模型似然中采样标记，即 $P(\mathbf{y}_j | \mathbf{x}, \mathbf{y}_{1:j-1})$ 。这可以通过选择具有最大似然的标记 (arg max采样) 实现，或采用其他随机采样技术。我们发现，使用核心采样 (Holtzman等人, 2019年) 相比arg max采样能带来更高的召回率 (附录C)。当生成EOS标记时，序列终止。序列生成后，直接提取并反量化对象描述（即获得预测的边界框和类别标签）即可。

2.3 序列增强以整合任务先验

EOS令牌允许模型决定何时终止生成，但在实践中我们发现，模型倾向于在不预测所有对象的情况下结束。这可能是由于：1) 标注噪声（例如，标注者未识别出所有对象），以及2) 某些对象识别或定位的不确定性。虽然这对整体性能的影响较小（例如，平均精度下降1-2%），但对召回率的影响更为显著。为提高召回率，一种技巧是通过人为降低EOS令牌的采样概率来延迟其采样。然而，这往往会导致预测结果出现噪声和重复。部分而言，这种精确度与召回率之间的艰难权衡，源于我们的模型对任务本身无感知，不了解检测任务的具体要求。

为解决这一问题，我们简单地引入了一种序列增强技术，从而融入了关于该任务的先验知识。在传统的自回归语言建模中（即未采用序列增强时），目标序列 $\tilde{\mathbf{y}}$ 与输入序列 \mathbf{y} 相同，且序列中的所有标记均为真实标记（例如，由人工标注转换而来）。而通过序列增强，我们在训练过程中对输入序列进行扩充，使其同时包含真实标记和合成的噪声标记。同时，我们调整目标序列，使模型学会识别而非模仿这些噪声标记。这增强了模型对噪声和重复预测的鲁棒性（尤其是在延迟EOS标记以提高召回率的情况下）。图5展示了序列增强所引入的修改，具体细节如下。

序列构造的修改 我们首先创建*}以下两种方式增强输入序列：1) 对现有真实标注对象添加噪声（例如，随机缩放或平移其边界框），2) 生成完全随机的框（附带随机关联的类别标签）。值得注意的是，部分噪声对象可能与某些真实标注对象完全相同或存在重叠，以此模拟噪声和重复预测的情况，如示例所示

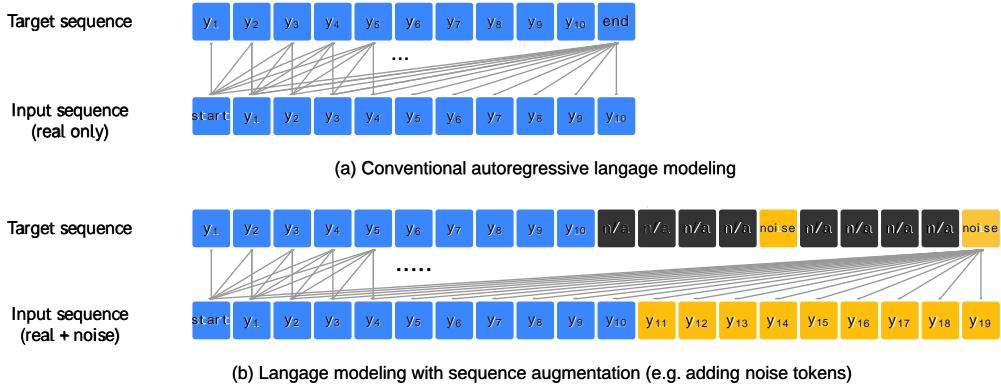


图5：序列增强与无序增强下的语言建模示意图。在序列增强的情况下，输入令牌被构建为包含真实对象（蓝色）和合成噪声对象（橙色）。对于噪声对象，模型被训练将其识别为“噪声”类，并且我们将“n/a”令牌（对应噪声对象的坐标）的损失权重设为零，因为我们不希望模型模仿它们。

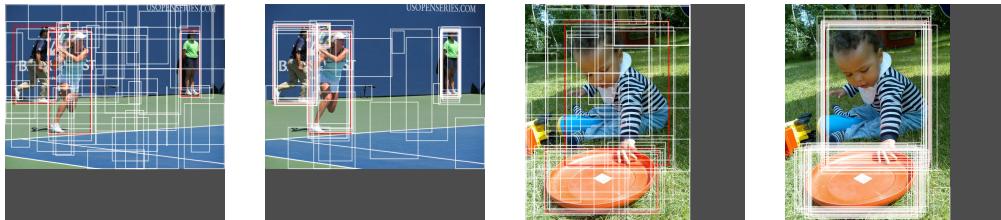


图6：随机采样的噪声对象（白色）与真实对象（红色）的对比示意图。

在图6中。合成并离散化噪声对象后，我们将其附加到原始输入序列的末尾。对于目标序列，我们将噪声对象的目标标记设为“噪声”类（不属于任何真实类别标签），并将噪声对象的坐标标记设为“n/a”，其损失权重设为零，即在公式1中设置 $w_j = 1_{[\tilde{y}_j \neq "n/a"]}$ 。

通过序列增强，我们能够显著延迟EOS（结束符）标记的出现，在不增加噪声和重复预测频率的情况下提升召回率。因此，我们让模型预测至最大长度，生成一个固定大小的对象列表。当从生成的序列中提取边界框和类别标签列表时，我们会将“噪声”类别标签替换为所有真实类别标签中具有最高似然值的实际类别标签，并将所选类别标记的似然值作为该对象的（排序）得分。

3 实验

3.1 实验设置

我们在MS-COCO 2017检测数据集（Lin等人，2014）上评估所提出的方法，该数据集包含18k训练图像和5k验证图像。为了与DETR和Faster R-CNN进行比较，我们在最后一个训练周期报告验证集上的平均精度（AP），这是一个跨多阈值的综合指标。我们采用两种训练策略：1) *training from scratch*在COCO上公平比较基线方法，以及2) *pretraining+finetuning*，即在更大的目标检测数据集Objects365（Shao等人，2019）上预训练Pix2Seq模型，然后在COCO上微调模型。由于我们的方法融入了零归纳偏差/目标检测任务的先验知识，我们预期第二种训练策略表现更优。

表1：在COCO验证集上，针对多阈值及不同物体尺寸的平均精度比较。各部分对比了采用相似ResNet“骨干”网络的不同方法。我们的模型在Faster R-CNN和DETR基线模型上均取得了具有竞争力的结果。

Method	Backbone	#params	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Faster R-CNN	R50-FPN	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster R-CNN+	R50-FPN	42M	42.0	62.1	45.5	26.6	45.4	53.4
DETR	R50	41M	42.0	62.4	44.2	20.5	45.8	61.1
Pix2seq (Ours)	R50	37M	43.0	61.0	45.6	25.1	46.9	59.4
Faster R-CNN	R101-FPN	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster R-CNN+	R101-FPN	60M	44.0	63.9	47.8	27.2	48.1	56.0
DETR	R101	60M	43.5	63.8	46.4	21.9	48.0	61.8
Pix2seq (Ours)	R101	56M	44.5	62.8	47.5	26.0	48.2	60.3
Faster R-CNN	R50-DC5	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster R-CNN+	R50-DC5	166M	41.1	61.4	44.3	22.9	45.9	55.0
DETR	R50-DC5	41M	43.3	63.1	45.9	22.5	47.3	61.1
Pix2seq (Ours)	R50-DC5	38M	43.2	61.0	46.1	26.6	47.0	58.6
DETR	R101-DC5	60M	44.9	64.7	47.7	23.7	49.5	62.3
Pix2seq (Ours)	R101-DC5	57M	45.0	63.2	48.6	28.2	48.9	60.4

在从头开始训练时，我们遵循（Carion等人，2020）的方法，采用ResNet主干网络（He等人，2016），后接6层Transformer编码器和6层（因果）Transformer解码器（Vaswani等人，2017）。我们将图像（保持固定宽高比）调整尺寸，使较长边为1333像素。在序列构建中，我们使用2000个量化区间，并在每次图像展示时随机化对象顺序。我们将噪声对象附加到真实对象上，使每张图像总共包含100个对象，因此序列长度为500。模型训练共进行300个周期，批次大小为128。

在Objects365数据集上进行预训练时，我们采用了与上述相似的设置，但存在几点差异。值得注意的是，我们没有使用较大的 1333×1333 图像尺寸，而是采用了较小的 640×640 尺寸，并以256的批量大小对模型进行了400K步的预训练。需要指出的是，由于使用了较小的图像尺寸，这一预训练过程甚至比从头开始训练还要快。在对COCO数据集进行微调时，仅需少量周期（例如20至60个周期）即可获得良好效果。此外，在微调阶段我们也可以使用更大的图像尺寸。得益于更大规模的预训练数据集，我们还尝试了基于Vision Transformers（Dosovitskiy等人，2020）的最大模型。

两种训练策略的更多细节可在附录B中找到。至于消融实验，我们采用ResNet-101主干网络，并缩小图像尺寸（较长边为640），从头开始训练模型200个周期。

3.2 主要比较

在COCO上从头训练 我们主要与两个广泛认可的基线方法进行比较DETR和Faster R-CNN。DETR与我们的模型架构相似，但我们的Transformer解码器无需学习“对象查询”或为边界框回归与分类设置独立头部，因为我们的模型通过单一softmax生成不同类型的令牌（如坐标令牌和类别令牌）。Faster R-CNN是一种成熟的方法，采用了如特征金字塔网络（FPN）（Lin等人，2017a）等优化架构。Faster R-CNN通常比DETR或我们的模型训练周期更短，这可能是因为其架构本身显式融入了任务先验知识。因此，我们还引入了一个改进的Faster R-CNN基线，记为Faster R-CNN^{v*}（源自Carion等人2020年的工作），其中Faster R-CNN模型采用GIoU损失（Rezatofighi等人，2019）、训练时随机裁剪增强以及长达9倍的训练计划进行训练。

结果如表1所示，其中各部分比较了同一ResNet“主干”网络的不同方法。总体而言，Pix2Seq取得了与两种基线方法相当的结果。我们的模型在中小型物体上表现与Faster R-CNN相当，但在较大物体上表现更优。相比

表2：不同主干架构和图像尺寸下，微调后的Pix2Seq模型在COCO上的平均精度。所有模型均在Objects365数据集上进行了预训练。作为对比，我们未经预训练的最佳模型在 1333×1333 的图像尺寸下取得了45.0 AP（见表1）。预训练采用 640×640 的图像尺寸，而微调（少量周期）可使用更大的图像尺寸。

Backbone	# params	Image size during finetuning		
		640×640	1024×1024	1333×1333
R50	37M	39.1	41.7	42.6
R50-C4	85M	44.7	46.9	47.3
ViT-B	115M	44.2	46.5	47.1
ViT-L	341M	47.6	49.0	50.0

使用DETR时，我们的模型在大型和中型物体上的表现相当或略逊一筹，但在小型物体上表现显著更优（提升4-5 AP）。

在Objects365上预训练并在COCO上微调 如表2所示， 经过Objects365预训练的Pix2Seq模型在不同模型规模和图像尺寸下均表现出色。最佳性能（1333图像尺寸）达到50 AP，比从头训练的顶级模型高出5%，且即便采用640图像尺寸，性能仍保持优异。值得注意的是，当预训练使用较小图像尺寸时，预训练+微调过程比从头训练更快，且泛化能力更强。这两点对于训练更大更优模型至关重要。

3.3 序列构建的消融研究

图7a探讨了坐标量化对性能的影响。在此消融实验中，我们考虑最长边为640像素的图像。图表显示，量化至500个或更多区间已足够；500个区间意味着每个区间约对应1.3个像素，这不会引入显著的近似误差。事实上，只要区间数量与图像最长边的像素数量相当，边界框坐标的量化就不会导致显著误差。

在训练过程中，我们还考虑了序列构建时的不同对象排序策略。这些策略包括：1) 随机排序，2) 按面积排序（即对象大小降序），3) 按dist2ori排序（即边界框左上角到原点的距离），4) 按类别（名称）排序，5) 按类别+面积排序（即首先按类别对对象进行排序，若同一类别有多个对象，则按面积排序），以及6) 按类别+dist2ori排序。图7b展示了前100个预测的平均精度（AP），图7c则展示了相应的平均召回率（AR）。无论是精度还是召回率，随机排序均表现出最佳性能。我们推测，确定性排序可能导致模型难以从早期遗漏对象的错误中恢复，而随机排序则仍有机会在后续阶段检索到这些对象。

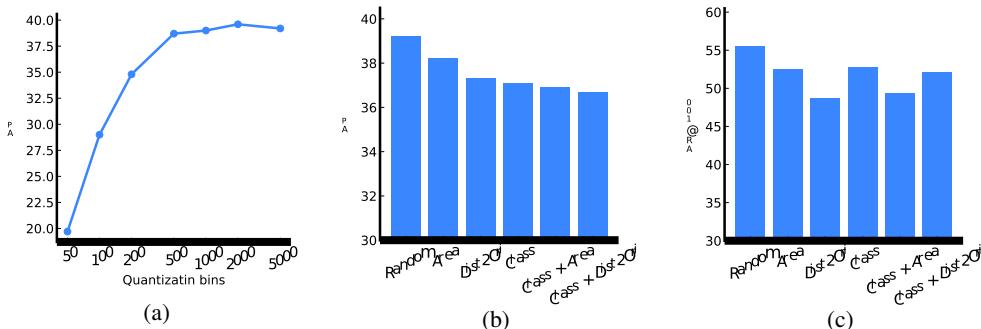


图7：序列构建的消融实验。(a) 量化分箱与性能关系。(b)和(c)展示了不同物体排序策略下的平均精度(AP)和召回率@100(AR@100)。

3.4 序列增强的消融研究

在此，我们研究了序列增强（即添加噪声对象）对两种模型训练策略的影响：1) 在COCO上从头开始训练，2) 在Objects365上预训练并在COCO上微调。图8展示了从头训练时使用与不使用序列增强的结果，我们发现，若不采用序列增强，在推理过程中延迟EOS令牌采样（通过似然偏移）时AP值会略微下降，但对于最优AP而言召回率则显著降低。表3显示了预训练+微调设置下的类似结果（此处我们对结束令牌设置了0.1的损失权重而非调整其似然偏移），可见序列增强缺失时AP未受显著影响，但召回率明显下降。值得注意的是，序列增强主要在微调阶段发挥显著作用。

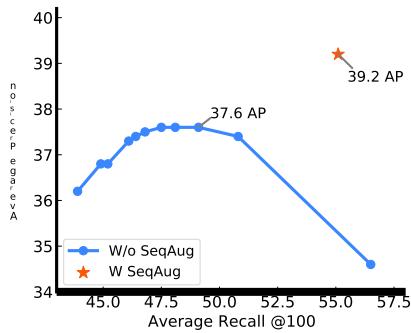


图8：在COCO上从头训练时序列增强的影响。

SeqAug in Pretrain	SeqAug in Finetune	AP	AR@100
✗	✗	43.7	55.4
✗	✓	44.5	61.6
✓	✓	44.7	61.7

表3：在Objects365上预训练并在COCO上微调时序列增强的影响。序列增强对平均召回率（@100）有显著提升，但对AP的影响较小。大部分改进可在微调阶段实现。

3.5 解码器交叉注意力图的可视化

在生成新标记时，Transformer解码器会对先前的标记进行自注意力计算，并对编码后的视觉特征图进行交叉注意力计算。这里我们将模型预测新标记时的交叉注意力（各层与注意力头的平均值）可视化。图9展示了生成最初几个标记时的交叉注意力分布图。可以看出，在预测第一个坐标标记（即 y_{\min} ）时，注意力分布非常分散，但随后迅速集中并锁定在目标物体上。

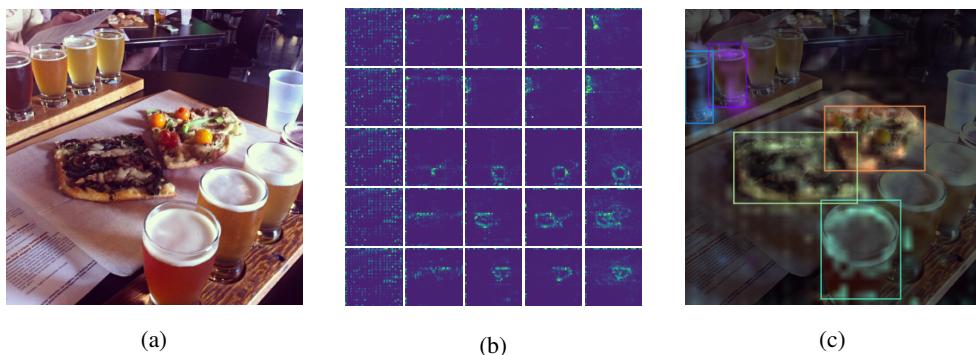


图9：解码器在预测前5个物体时对视觉特征图的交叉注意力。(b) 我们将25个预测序列重塑为 5×5 网格，因此每一行代表对5个标记 $[y_{\min}, x_{\min}, y_{\max}, x_{\max}, c]$ 的预测。在选择物体首个标记时注意力分布较为分散，随后迅速集中到该物体上。(c) 原始图像上叠加了（预测类别标记时的）交叉注意力分布。

4 相关工作

目标检测。现有的目标检测算法在其架构选择和损失函数设计中融入了关于任务的显式先验知识。为了预测一组边界框，现代检测器的架构专门设计用于生成大量提议（Girshick, 2015; Ren et al., 2015; Cai & Vasconcelos, 2018）、锚框（Lin et al., 2017b）或窗口中心（Tian et al., 2019; Zhou et al., 2019）。通常需要非极大值抑制（Bodla et al., 2017）来防止重复预测。尽管DETR (Carion et al., 2020) 避免了复杂的边界框提议和非极大值抑制，但它仍需要一组学习到的“对象查询”，专门用于对象绑定。这些检测器均需单独的子网络（或额外层）来回归边界框和类别标签。Pix2Seq通过采用通用的图像编码器和序列解码器，并仅使用单一softmax生成坐标标记和类别标签，从而避免了这些复杂性。

除了架构之外，现有检测器的损失函数也高度针对边界框匹配进行了定制。例如，损失函数通常基于边界框回归 (Szegedy等人, 2013; Lin等人, 2017b)、交并比 (Rezatofighi等人, 2019) 以及基于集合的匹配 (Erhan等人, 2014; Liu等人, 2016; Redmon等人, 2016; Stewart等人, 2016; Carion等人, 2020)。Pix2Seq避免了专门的损失函数，表明采用带softmax交叉熵的简单最大似然目标也能取得良好效果。

我们的工作还与目标检测中的循环模型相关 (Stewart等人, 2016; Park & Berg, 2015; Roura-Paredes & Torr, 2016; Salvador等人, 2017; Ren & Zemel, 2017)，这些模型通过学习一次预测一个目标来实现检测。如上所述，这些方法中的架构和损失函数通常针对检测任务进行了专门设计。此外，这些方法并非基于Transformer架构，也未在更大规模的数据集上与现代基准进行过对比评估。

语言建模。我们的工作受到现代语言建模近期成功的启发 (Radford等人, 2019; Raffel等人, 2019; Brown等人, 2020)。尽管最初是为自然语言设计的，但该方法论已被证明能够建模多种序列数据，如机器翻译 (Sutskever等人, 2014; Bahdanau等人, 2014)、图像描述生成 (Vinyals等人, 2015b; Karpathy & Fei-Fei, 2015; Xu等人, 2015) 以及许多其他领域 (Vinyals等人, 2015a; Huang等人, 2018; Ramesh等人, 2021; Chen等人, 2021)。我们的研究丰富了这一系列成果，并展示了该方法甚至适用于非序列数据（通过将一组对象转化为标记序列）。我们为模型增强了输入和目标序列，以融入任务特定的先验知识；类似的序列扰动策略已在语言模型中使用 (Devlin等人, 2018; Clark等人, 2020)，并与噪声对比学习 (Gutmann & Hyvärinen, 2010) 及生成对抗网络中的判别器 (Goodfellow等人, 2014) 存在一定相似性。

5 结论与未来工作

本文介绍了Pix2Seq，一个简单而通用的目标检测框架。通过将目标检测任务转化为语言建模问题，我们的方法极大地简化了检测流程，消除了现代检测算法中的大部分专业化设计。我们相信，该框架不仅适用于目标检测，还能应用于其他视觉任务——只要其输出可表示为相对简洁的离散标记序列（例如关键点检测、图像描述生成、视觉问答等）。为此，我们希望将Pix2Seq扩展为一个通用且统一的接口，用于解决各类视觉任务。

我们方法的一个主要局限在于，自回归建模对长序列而言计算成本高昂（尤其在模型推理阶段）。缓解这一问题的实际措施包括：1) 在生成结束标记时停止推理（例如，COCO数据集中平均每张图像包含7个目标，导致~35标记数量相对较少）；2) 将其应用于离线推理，或目标对象相对稀疏的在线场景（如通过语言描述定位特定对象）。然而，要实现实时目标检测应用的速度提升，仍需未来工作加以改进。另一局限是目前Pix2Seq的训练完全依赖人工标注，若能降低这种依赖性，模型将能从更多未标注数据中受益。

致谢

我们特别感谢顾秀业为准备Objects365数据集所做的贡献。同时，我们也感谢Mohammad Norouzi、Simon Kornblith、林忠毅(Tsung-Yi Lin)、Allan Jabri以及Kevin Swersky在有益讨论中提供的帮助。

参考文献 Dzmitry Bahdanau、Kyunghyun Cho和Yoshua Bengio。通过联合学习对齐与翻译的神经机器翻译。*arXiv preprint arXiv:1409.0473*, 2014年。 Navaneeth Bodla、Bharat Singh、Rama Chellappa和Larry S Davis。Soft-NMS——用一行代码提升目标检测。载于

Proceedings of the IEEE International Conference on Computer Vision, 第5561–5569页, 2017年。 Tom B Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell等。语言模型是小样本学习者。*arXiv preprint arXiv:2005.14165*, 2020年。 Zhaowei Cai和Nuno Vasconcelos。Cascade R-CNN：深入高质量目标检测。载于*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第6154–6162页, 2018年。 Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov和Sergey Zagoruyko。基于Transformer的端到端目标检测。载于*European Conference on Computer Vision*, 第213–229页。Springer, 2020年。 Lili Chen、Kevin Lu、Aravind Rajeswaran、Kimin Lee、Aditya Grover、Michael Laskin、Pieter Abbeel、Aravind Srinivas和Igor Mordatch。决策Transformer：通过序列建模实现强化学习。*arXiv preprint arXiv:2106.01345*, 2021年。 Ting Chen、Simon Kornblith、Mohammad Norouzi和Geoffrey Hinton。视觉表示对比学习的简单框架。载于*International Conference on Machine Learning*, 第1597–1607页。PMLR, 2020a。 Ting Chen、Simon Kornblith、Kevin Swersky、Mohammad Norouzi和Geoffrey E Hinton。大型自监督模型是强大的半监督学习者。

Advances in Neural Information Processing Systems, 33卷: 22243–22255页, 2020b。 Kevin Clark、Minh-Thang Luong、Quoc V. Le和Christopher D. Manning。ELECTRA：将文本编码器预训练为判别器而非生成器。载于*ICLR*, 2020年。 Jacob Devlin、Ming-Wei Chang、Kenton Lee和Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。

arXiv preprint arXiv:1810.04805, 2018年。 Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly等。一幅图像相当于16x16个词：大规模图像识别的Transformer。载于*International Conference on Learning Representations*, 2020年。 Dumitru Erhan、Christian Szegedy、Alexander Toshev和Dragomir Anguelov。使用深度神经网络的可扩展目标检测。载于

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 第2147–2154页, 2014年。 M. Everingham、L. Van Gool、C. K. I. Williams、J. Winn和A. Zisserman。PASCAL视觉对象分类(VOC)挑战赛。*International Journal of Computer Vision*, 88(2):303–338, 2010年6月。 Golnaz Ghiasi、Yin Cui、Aravind Srinivas、Rui Qian、Tsung-Yi Lin、Ekin D Cubuk、Quoc V Le和Barret Zoph。简单复制粘贴是实例分割的强大数据增强方法。载于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第2918–2928页, 2021年。 Ross Girshick。Fast R-CNN。载于*Proceedings of the IEEE International Conference on Computer Vision*, 第1440–1448页, 2015年。 Ross Girshick、Jeff Donahue、Trevor Darrell和Jitendra Malik。用于精确目标检测和语义分割的丰富特征层次结构。载于

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 第580–587页, 2014年。 Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville和Yoshua Bengio。生成对抗网络。*Advances in Neural Information Processing Systems*, 27卷, 2014年。

迈克尔·古特曼与阿波·许韦里宁。噪声对比估计：非归一化统计模型的新估计原理。载于 *Proceedings of the thirteenth International Conference on artificial intelligence and statistics*, 第297–304页。JMLR研讨会与会议论文集, 2010年。

何恺明、张翔宇、任少卿、孙剑。深度残差学习在图像识别中的应用。载于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第770–778页, 2016年。

何恺明、Georgia Gkioxari、Piotr Dollár和Ross Girshick。Mask R-CNN。载于 *Proceedings of the IEEE International Conference on Computer Vision*, 第2961–2969页, 2017年。

Elad Hoffer、Tal Ben-Nun、Itay Hubara、Niv Giladi、Torsten Hoefler与Daniel Soudry合著。通过实例重复增强批次：提升泛化能力的研究。载于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第8129–8138页, 2020年。

阿里·霍尔茨曼、简·布伊斯、杜丽、马克斯韦尔·福布斯与崔艺珍。神经文本退化的奇妙案例。*arXiv preprint arXiv:1904.09751*, 2019年。

安德鲁·G·霍华德。基于深度卷积神经网络的图像分类方法若干改进。*arXiv preprint arXiv:1312.5402*, 2013年。程安娜·黄、阿什什·瓦斯瓦尼、雅各布·乌兹科雷特、诺姆·沙泽尔、伊恩·西蒙、柯蒂斯·霍桑、安德鲁·M·戴、马修·D·霍夫曼、莫妮卡·丁库莱斯库与道格拉斯·埃克。音乐变换器。*arXiv preprint arXiv:1809.04281*, 2018年。高黄、孙宇、庄子刘、丹尼·塞德拉与基利安·Q·温伯格。随机深度深度网络。载于 *European Conference on Computer Vision*, 第646–661页。斯普林格, 2016年。保罗·F·耶格尔、西蒙·AA·科尔、塞巴斯蒂安·比克尔豪普特、法比安·伊森西、特里斯坦·安塞尔姆·库德尔、海因茨·彼得·施莱默与克劳斯·H·迈尔·海因。视网膜U-Net：医学目标检测中分割监督的极简利用。载于 *Machine Learning for Health Workshop*, 第171–183页。PMLR, 2020年。安德烈·帕西与李飞飞。生成图像描述的深度视觉-语义对齐方法。载于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第3128–3137页, 2015年。迪德里克·金马与吉米·巴。Adam：一种随机优化方法。*arXiv preprint arXiv:1412.6980*, 2014年。亚历克·斯里热夫斯基、伊利亚·苏茨克弗与杰弗里·E·辛顿。基于深度卷积神经网络的ImageNet分类。*Advances in Neural Information Processing Systems*, 25:1097–1105, 2012年。阿丽·布兹涅佐娃、哈桑·罗姆、尼尔·阿尔德林、贾斯珀·尤林斯、伊万·克拉辛、约尔迪·蓬特-图塞特、沙哈布·卡马利、斯特凡·波波夫、马泰奥·马洛奇、亚历山大·科列斯尼科夫等。Open Images数据集v4。*International Journal of Computer Vision*, 128(7):1956–1981, 2020年。勘昆、伯恩哈德·博瑟、约翰·S·登克、唐尼·亨德森、理查德·E·霍华德、韦恩·哈伯德与劳伦斯·D·杰克尔。反向传播算法在手写邮政编码识别中的应用。*Neural computation*, 1(4):541–551, 1989年。

李毅、齐浩志、戴继峰、纪翔、和魏亦忱。全卷积实例感知语义分割。载于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第2359–2367页, 2017年。

林宗一、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár与C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于 *European Conference on Computer Vision*, 第740–755页。Springer出版社, 2014年。

林宗仪、Piotr Dollár、Ross Girshick、何恺明、Bharath Hariharan和Serge Belongie。特征金字塔网络用于目标检测。收录于 *Proceedings of the IEEE Conference on Computer Vision and pattern recognition*, 第2117–2125页, 2017a。

林惊毅、Priya Goyal、Ross Girshick、何恺明与Piotr Dollár。密集目标检测中的焦点损失。《*Proceedings of the IEEE International Conference on Computer Vision*》, 第2980–2988页, 2017b年。

Wei Liu、Dragomir Anguelov、Dumitru Erhan、Christian Szegedy、Scott Reed、Cheng-Yang Fu和Alexander C Berg。SSD：单次多框检测器。载于 *European Conference on Computer Vision*, 第21–37页。Springer出版社, 2016年。

Ilya Loshchilov与Frank Hutter。解耦权重衰减正则化。见于*International Conference on Learning Representations*, 2018年。

Eunbyung Park 与 Alexander C Berg。学习分解以实现目标检测与实例分割。
arXiv preprint arXiv:1511.06449, 2015年。

亚历克·拉德福德、卡蒂克·纳拉辛汉、蒂姆·萨利曼斯与伊利亚·苏茨克维。通过生成式预训练提升语言理解能力。2018。

亚历克·拉德福德、杰弗里·吴、雷文·柴尔德、大卫·吕安、达里奥·阿莫迪、伊利亚·苏茨克弗等。语言模型是无监督多任务学习者。*OpenAI blog*, 1(8):9, 2019年。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li和Peter J Liu。探索统一文本到文本转换器在迁移学习中的极限。*arXiv preprint arXiv:1910.10683*, 2019年。

Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen与Ilya Sutskever。零样本文本到图像生成。*arXiv preprint arXiv:2102.12092*, 2021年。

约瑟夫·雷德蒙、桑托什·迪瓦拉、罗斯·吉尔希克与阿里·法哈迪。你只需看一次：统一、实时的目标检测。载于*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第779–788页, 2016年。

任梦野与理查德泽梅尔。基于循环注意力的端到端实例分割。载于*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第6656–6664页, 2017年。

任少卿、何恺明、罗斯·吉尔希克与孙剑。Faster R-CNN：利用区域提议网络实现实时目标检测。*Advances in Neural Information Processing Systems*, 28卷: 91–99页, 2015年。

哈米德扎托菲吉、内森·蔡、郭俊英、阿米尔·萨德吉安、伊恩·里德与西尔维奥·萨瓦雷斯。广义交并比。2019年6月。

伯纳迪诺·罗梅拉-帕雷德斯与菲利普·希莱尔·肖恩·托尔。循环实例分割。载于*European Conference on Computer Vision*, 第312–329页。斯普林格出版社, 2016年。

Inkyu Sa、葛宗元、Feras Dayoub、Ben Upcroft、Tristan Perez与Chris McCool。DeepFruits：基于深度神经网络的水果检测系统。

sensors, 16卷8期: 1222页, 2016年。

阿玛萨尔瓦多、米里亚姆·贝尔弗、维克多·坎波斯、马内尔·巴拉达德、费兰·马克斯、霍尔迪·托雷斯与泽维尔·吉罗·尼托。用于语义实例分割的循环神经网络。

arXiv preprint arXiv:1712.00617, 2017年。

邵帅、李泽明、张天元、彭超、余刚、张翔宇、李静与孙剑。

Objects365：面向目标检测的大规模高质量数据集。载于*Proceedings of the IEEE/CVF international conference on computer vision*, 第8430–8439页, 2019年。

拉塞斯图尔特、米哈伊洛·安德里留卡与吴恩达。拥挤场景中的端到端行人检测。载于

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 第2325–2333页, 2016年。

裴孙、亨里克·克雷茨施马尔、薛西斯·多蒂瓦拉、奥雷连·舒瓦德、维贾伊赛·帕特奈克、保罗·崔、詹姆斯·郭、尹舟、柴玉宁、本杰明·凯恩等。《自动驾驶感知的可扩展性：Waymo开放数据集》。载于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第2446–2454页, 2020年。

Ilya Sutskever、Oriol Vinyals和Quoc V Le。用神经网络进行序列到序列学习。载于*Advances in Neural Information Processing Systems*, 第3104–3112页, 2014年。

克里斯蒂安·塞格迪、亚历山大·托舍夫与杜米特鲁·埃尔汉。用于目标检测的深度神经网络。*Advances in neural information processing systems*, 26卷, 2013年。

支天、沈春华、陈浩与何通。FCOS：全卷积一阶段目标检测。载于*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 第9627–9636页, 2019年。

阿希什·瓦斯瓦尼、诺姆·沙泽尔、尼基·帕尔马、雅各布·乌兹科雷特、利昂·琼斯、艾丹·N·戈麦斯、卢卡什·凯泽和伊利亚·波洛苏欣。《注意力就是你需要的一切》。载于*Advances in Neural Information Processing Systems*, 第5998–6008页, 2017年。

奥里奥尔·维尼亞尔斯、卢卡什·凯泽、特里·库、斯拉夫·彼得罗夫、伊利亚·苏茨克维尔和杰弗里·辛顿。《语法作为一种外语》。*Advances in Neural Information Processing Systems*, 28:2773–2781, 2015a。

Oriol Vinyals、Alexander Toshev、Samy Bengio和Dumitru Erhan。展示与讲述：一种神经图像描述生成器。载于*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 第3156–3164页, 2015b。Yuxin Wu、Alexander Kirillov、Francisco Massa、Wan-Yen Lo和Ross Girshick。Detectron 2。<https://github.com/facebookresearch/detectron2>, 2019。Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel和Yoshua Bengio。展示、关注与讲述：基于视觉注意力的神经图像描述生成。载于*International Conference on Machine Learning*, 第2048–2057页。PMLR, 2015。Xingyi Zhou、Dequan Wang和Philipp Krähenbühl。以点代物。*arXiv preprint arXiv:1904.07850*, 2019。

坐标的量化与反量化

算法1和2展示了（归一化）坐标的量化与反量化过程。

算法1（归一化）坐标的量化

```
def quantize(x, bins=1000): # x 是一个介于 [0, 1] 之间的实数  
# 返回一个介于 [0, bins-1] 之间的整数 return int(x * (bins - 1))
```

算法2 坐标离散标记的反量化

```
def反量化(x, 分箱数=1000): # x 是一个介于 [0, 分箱数 - 1] 的整数 # 返回一个介于 [0, 1] 的实数 return float(x) / (分箱数 - 1)
```

B 训练细节

在COCO数据集上从头训练 对于基线架构，我们遵循arion等人，2020)的方法，采用ResNet主干网络(He等人，2016)，后接6层Transformer编码器和6层(因果)Transformer解码器(Vaswani等人，2017)。Transformer的主要维度设置为256，配备8个注意力头，前馈网络的维度设为1024。我们使用10%比率的随机深度(Huang等人，2016)来减少过拟合。根据(Carion等人，2020)，我们还尝试了ResNet的DC5变体(Li等人，2017)，该变体将其输出特征图的分辨率提高了一倍。²

在训练过程中进行图像增强时，我们采用随机裁剪的尺度抖动方法 (Ghiasi等人，2021；Wu等人，2019)，强度为[0.1, 3]。我们将图像（保持固定宽高比）调整大小，使较长边为1333像素。遵循 (Howard，2013；Chen等人，2020a；b) 的方法，我们还使用了强度为0.5的色彩失真。对于序列构建，我们使用2000个量化分箱，并在每次图像展示时随机化对象的顺序。我们将噪声对象附加到真实对象上，使每张图像总共包含100个对象，因此序列长度为500。

我们从头开始训练整个网络，共进行300个周期 (epoch)，每批 (batch) 大小为128。对于小批次中的每张图像，我们执行两次独立的增强操作，类似于(Hoffer等人，2020)的做法，从而得到256的有效批次大小，这有助于减少过拟合。我们采用AdamW优化器 (Kingma & Ba, 2014; Loshchilov & Hutter, 2018)，学习率设为0.003，权重衰减为0.05。训练初期使用10个周期的学习率预热 (warmup)，之后在整个训练过程中线性衰减学习率。

在Objects365上的预训练 我们探索了更广泛的架构变体，包括混合ResNet与Transformer模型 (Carion等人，2020)，以及基于图像块的纯Transformer架构 (Dosovitskiy等人，2020)。具体架构细节可在我们公开的代码中查阅。由于Objects365数据集规模远超COCO (170万张图像对比11.8万张图像)，我们采用了较弱的图像增强策略 (ViT骨干网络的尺度抖动范围为[0.3, 2]，ResNet骨干网络则为[0.9, 1.2])，且未使用色彩失真。在序列构建方面，我们采用1000个量化分箱。默认情况下，我们仍会通过添加采样噪声对象来实施序列增强。

我们采用较小的图像尺寸 640×640 ，并以256的批量大小对模型进行了400K步的预训练。与从头开始训练不同，我们未对每批次执行两次增强。同时，我们使用了较小的学习率0.001，并保持相同的权重衰减率0.05。学习率采用余弦衰减策略，初始预热阶段为20K步。

至于在COCO数据集上的微调，我们为ResNet主干网络使用128的批量大小，ViT主干网络则使用64。大多数模型以 $3e^{-5}$ 的学习率微调60个周期，但更少的周期也能得到相似的结果。我们仍采用范围在[0.3, 2]的尺度抖动进行图像增强。

²向th添加一个扩张 e last ResNet stage and removing the stride from the first convolution 该阶段的

C 推理消融实验 ($\arg \max$ 对比 核采样)

核采样 (Holtzman等人, 2019年) 已被应用于语言建模中, 以减少生成样本的重复性并增加多样性。在此, 我们研究了其对从训练模型中采样的影响。

给定分布 $P(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{1:j-1})$, 为了应用核采样, 我们首先将其top- p 词汇 $V^{(p)} \subset V$ 定义为满足以下条件的最小集合

$$\sum_{\mathbf{y}_j \in V^{(p)}} P(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{1:j-1}) \geq p. \quad (2)$$

令 $p' = \sum_{\mathbf{y}_j \in V^{(p)}} P(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{1:j-1})$, 我们可以如下重新校准条件似然以采样下一个标记。

$$P'(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{1:j-1}) = \begin{cases} P(\mathbf{y}_j|\mathbf{x}, \mathbf{y}_{1:j-1})/p' & \text{if } \mathbf{y}_j \in V^{(p)} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

我们调整了在生成输出序列 (推理阶段) 时使用的核采样超参数 p 。当 $p=0$ 时, 对应的是 $\arg \max$ 采样方式; 否则, 它会从一个累积和大于或等于 p 的截断排序令牌列表中进行采样。在图10中可以看到, 采用核采样 (当 $p>0$ 时) 能够提升物体召回率, 从而也带来更高的平均精度。在 0.2 至 0.5 之间存在一个相对平坦的平均精度区域, 因此我们选择 p 为 0.4 作为其他实验的默认值。

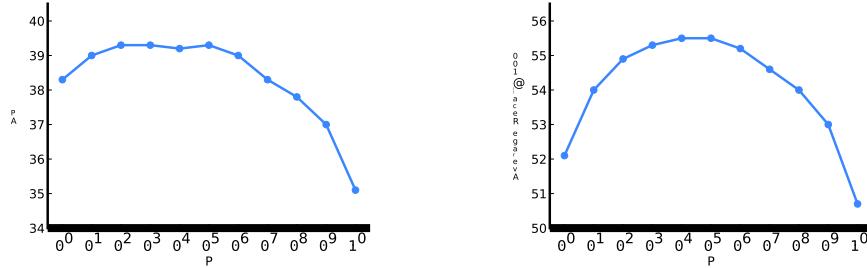


图10: 在推理过程中改变参数 p 进行核采样会导致不同的AP和AR结果。当 $p=0$ 时, 等同于 $\arg \max$ 采样。而采用 $p>0$ 的采样有助于提高召回率 (以及精确率)。

坐标令牌间相似性的D可视化

在我们的模型中, 边界框坐标并非以浮点数形式表示, 而是编码为离散的token。此处我们通过其嵌入向量研究这些坐标token间的相似性。需注意的是, 离散坐标token与类别名称 token 同属一个词汇表, 并共享相同的嵌入矩阵。具体而言, 我们首先切分出与坐标token对应的学习嵌入矩阵, 随后计算这些坐标token嵌入向量间的余弦相似度。

图11展示了坐标标记嵌入之间的余弦相似度。可以看出, 邻近坐标在标记嵌入中的相似度高于相距较远的坐标。我们模型表现出的这一特性, 很可能源于边界框标注中的噪声/不确定性 (即边界框标注是从潜在边界框分布中随机抽取的样本, 该分布编码了坐标的局部性)。

E THE ABI 以给定CO引导注意力的能力

纵坐标

我们探究模型通过坐标指定 *pay attention to a pointed region* 的能力。将图像均匀划分为 $N \times N$ 的矩形区域网格, 每个区域由一系列

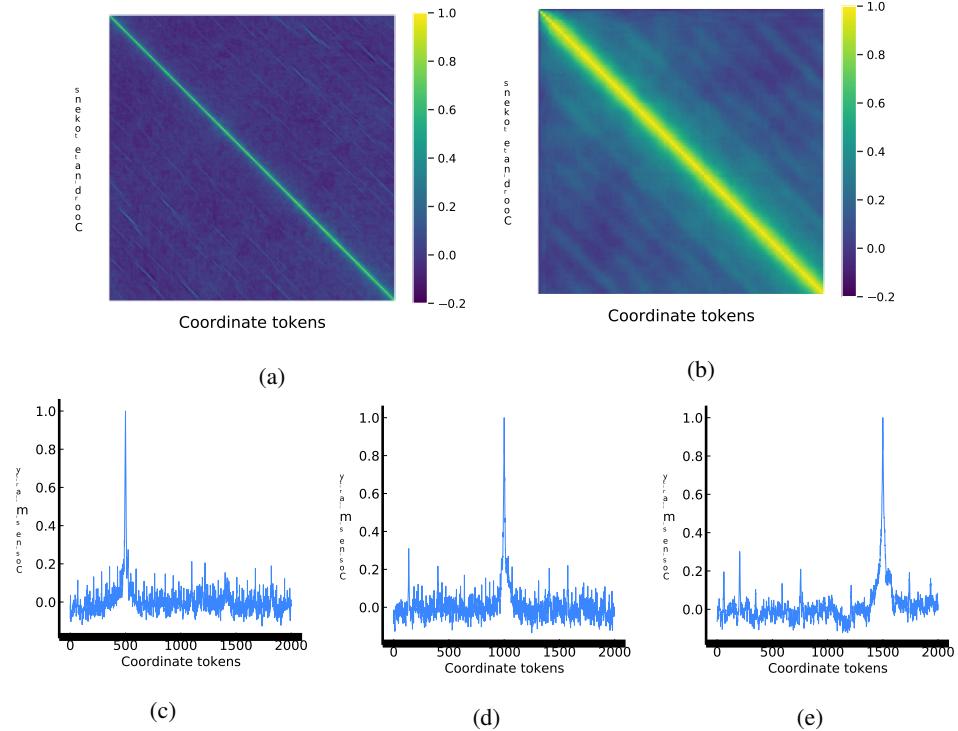


图11: (a) 坐标令牌嵌入间的余弦相似度。(b) 是(a)的一部分，仅覆盖前100个令牌。(c)、(d)和(e)分别是(a)的第500、1000和1500行。邻近坐标在其令牌嵌入中具有更高的相似性。

其边界框的坐标。随后，在解码器读取每个区域的坐标序列（即 $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ ）后，我们可视化其对视觉特征图的交叉注意力。为了消除现有对象的干扰，我们对图像中的像素进行了随机打乱，并移除了前2%的注意力权重以提高清晰度。有趣的是，如图12所示，模型似乎能够在不同尺度上关注到指定的区域。

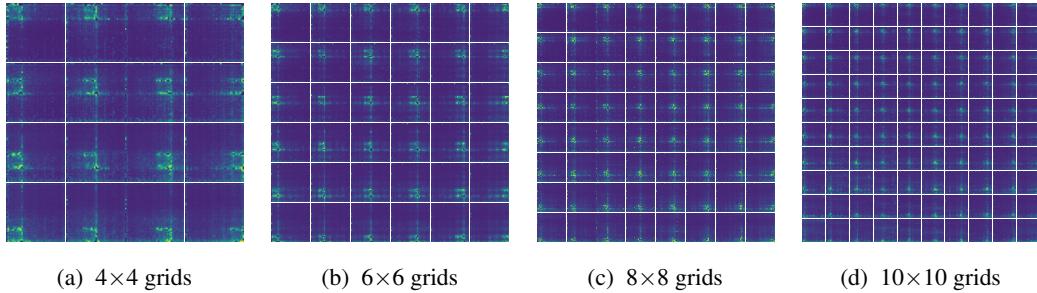


图12: 每个网格是对解码器在读取一小段坐标序列（即 $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ ）后注意力机制的可视化。针对不同尺寸的网格进行了可视化展示。网络学会在不同尺度上关注所指区域。

F 解码器交叉注意力更多可视化

在图13中，我们将交叉注意力（预测类别标记时）叠加到其他几张原始图像上，结果显示解码器在预测类别标记时最为关注目标对象。

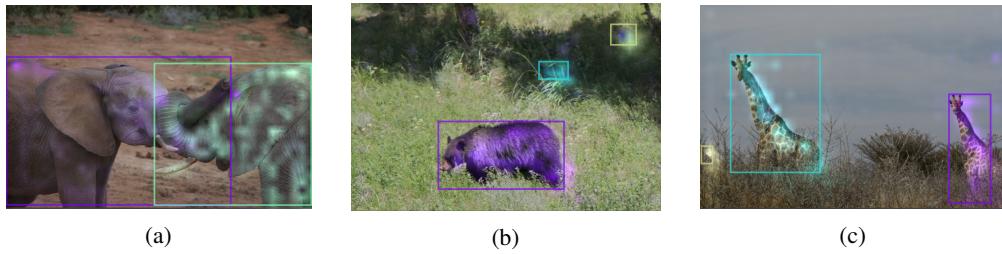


图13：Transformer解码器在给定边界框条件下（预测类别标记时）的交叉注意力可视化。

检测结果可视化

在图14中，我们可视化展示了Pix2seq模型（AP值为46）在COCO验证集部分图像上的检测结果，这些图像包含密集排列的物体。

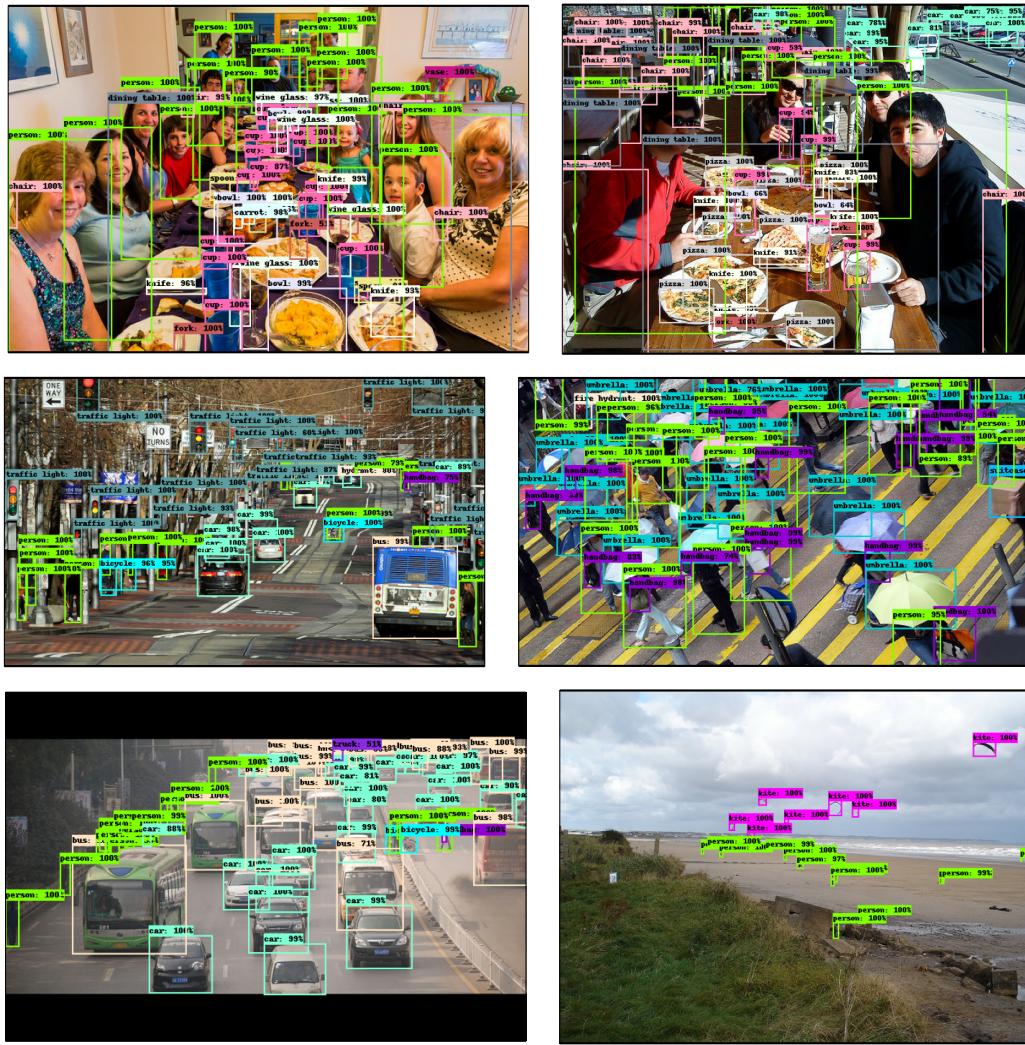


图14：模型预测示例（分数阈值为0.5）。通过点击支持的PDF阅读器中的图像可访问原始图像。