**Research on Panoptic Segmentation Algorithm Based on Transformer**

*Abstract*

Panoptic segmentation是一种统合了图像semantic segmentation和instance segmentation的图像分割任务，其不仅关注天空、道路等背景的stuff，还关注people、car等不同个体的things。为了能够学习更加全面的图像特征，在End-to-end Detection Transformer的基础上新增了Mask R-CNN中FPN-style的mask head，将Transformer Decoder的输出特征转换为实例分割图。最后在COCO 2017 Validation的PQ为42.8对于stuff的PQ为35.8、things的PQ为48.1，与同时期相比处于上游水平。

1. Introduction

随着计算机硬件和互联网技术发展，不仅计算机能够处理更复杂的任务，同时各大厂商、企业能够收集到更多类型和数量的数据，这些数据有时是文本，有时是语音，有时是图像。这些数据多起来的时候，更复杂的计算机任务需求逐渐提出，图像识别便是如此。早期图像识别是将图像读取为矩阵形式，每一个像素点当作一个特征，因此像素点数量即特征数量。受限于设备硬件早期的图像识别只能在像素特别小的图片上完成对图像特征进行分类的任务，大多数算法都是基于传统统计学和概率论进行的推理，并没有实质性突破。

随着21世纪初，LetNet-5[1]、AlexNet[2]、VGG[3]、ResNet[4]等CNN出现，图像识别领域不再局限对图像进行分类，提出了更复杂的目标检测、图像分割等任务。图像分割任务中早期是对图像进行像素级分类，J. Long, et al.[5]于2015年提出全卷积神经网络FCN，是第一个基于深度学习的语义分割模型。该模型可以分为编码器-解码器的结构，编码器即CNN的结构用于提取特征图；解码器由多个上采样层构成，通过特征图上采样至原始图像大小的分割图，因此后续提出的U-Net[6]、V-Net[7]和其他语义分割的深度学习模型都是基于编码器-解码器结构设计。随着图像语义分割的发展，学者认为天空、道路和行人、车辆不能混为一谈，B. Hariharan, et al.[8]认为应该存在一种方法能够对行人、车辆的目标物体Things进行分割，即图像分割发展中期的实例分割。实例分割与语义分割不同，其是在检测到的目标上形成目标物体的分割图，可以理解为目标检测生成对象预测框后，对物体框的对象进行语义分割的双阶段任务，因此学者们在目标检测模型上增加FCN解码器类似的结构以生成目标分割图。直到2017年K. He, et al.[9]在R-CNN框架的基础上提出Mask R-CNN更灵活简单的双阶段实例分割模型。该模型在MS COCO[10]数据集上取得较好的实验结果，采用了ResNet-FPN结构，即FPN中多层特征图融合，其多尺度特征图有利于多尺度物体以及小物体的检测，最后通过双线性插值上采样得到实例分割图，对于物体目标而言具有良好的分割质量。随着目标检测单阶段算法的发展，单阶段实例分割也逐渐兴起，YOLCAT[11]的代表，将掩码生成方式添加到单阶段的目标检测模型中。在实例分割和语义分割的基础上，衍生出统一Things和Stuff的全景分割任务，在实例分割的基础上加入了对后景的像素级分割。但是由于前景和后景的区分，会导致评价指标不好定义，因此A. Kirillov, et al.[12]提出全景分割的评价指标全景质量PQ，是统一了Things和Stuff对全景分割的评价指标，能够更加有效、全面的评价模型在全景分割的性能。

计算机视觉领域的发展随着卷积神经网络在2012年ImageNet比赛的胜出，加快了卷积神经网络为基础的各种领域发展，于此同时计算机视觉领域带来的深度学习，使得自然语言处理也在飞速发展。2017年Vaswani A.等人[13]于自然语言处理领域发表Transformer模型，该模型引入了注意力机制的编码器-解码器结构，用于机器翻译任务的序列到序列模型。相较于以往的序列到序列模型，Transformer完成了长短不一句子预测，体现了该模型强大的泛化能力，最大程度来源于注意力机制对全局特征感知的能力。介于Transformer出色的全局感知能力，A. Dosovitskiy, et al.[14]首次提出将Transformer编码器用于图像分类任务中，将图片分块后的每一块拉平后当成输入向量的单个Token，最后增加一个多层感知机实现分类任务，后续实验证实这种抛弃卷积的架构使得模型准确率远高于传统的卷积神经网络，至此Transformer正式进入计算机视觉领域学者的视线。2020年N. Carion, et al.[15]提出基于Transformer的端到端目标检测模型DETR，剔除了R-CNN、YOLCAT等架构中繁琐的Archer Box预设置和非极大值抑制后处理，通过Objects Query真正实现了端到端的目标检测深度模型。虽然在目标检测上其性能不及同时期其他架构，但是这种没有繁琐预/后处理的DETR架构成为了这种模式的开山之作，被后续学者广泛沿用，如近几年提出的分割架构SETR[16], MaskFormer[17]，均是基于Transformer的端到端架构，该架构基于卷积神经网络局部感受野和Transformer全局特征的特性，能够更全面的理解图像特征，从而完成更加复杂的视觉任务。

本文将基于DETR的框架，引入Mask R-CNN风格的Mask head，对Transformer Decoder输出进行处理以完成全景分割任务，在COCO 2017 Panoptic Validation上进行模型复现和算法剖析。由于Transformer架构本身的庞大参数量导致难以训练，因此在Backbone上采取ResNet于ImageNet比赛中已经完成的预训练模型作为basic，来减少模型参数收敛所需的训练时间。

1. Methods

本章将介绍基础的DETR架构和如何将Mask R-CNN的Mask Head引入到该结构中，并剖析训练时所需的损失函数和评价指标。

1. *DETR architecture*

DETR主要由Backbone、Transformer、共享参数的输出FFN三部分组成，原始结构如图1.

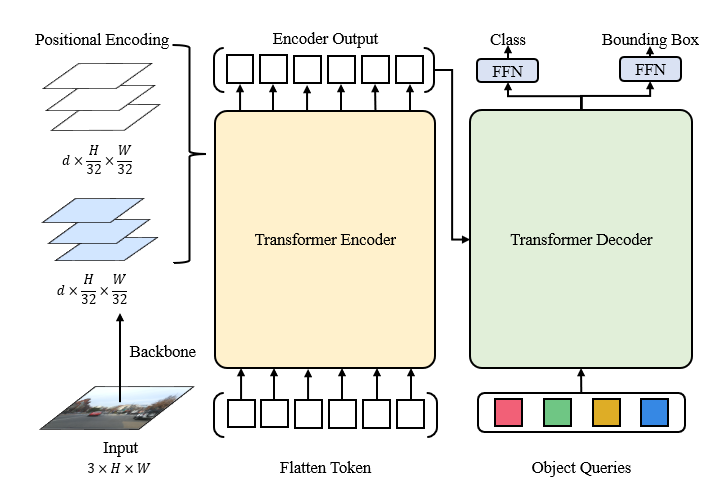


Figure 1. DETR original structure.

1. *Backbone*

该部分是使用CNN对原始图像进行特征提取，去掉CNN最后输出分类器的全局平均池化、全连接层，对输入的三通道彩色图像进行深度卷积运算，即三通道图像特征映射到更多通道数的特征，得到多通道特征图。假设输入图像的宽度和高度像素均为224，比较具有代表性的ResNet-50前向推理如图2.

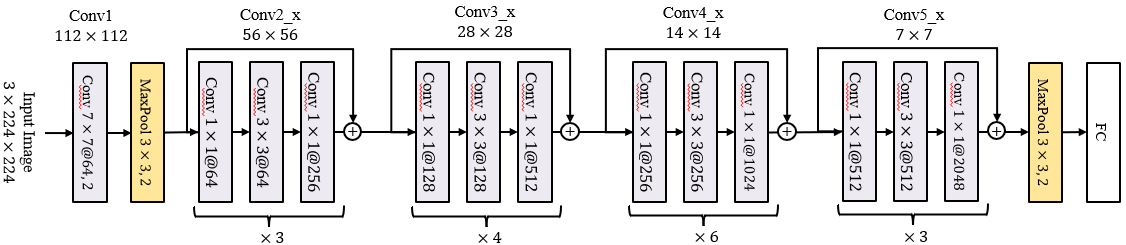


Figure 2. ResNet-50 forward calculation.

Backbone即去除最后黄色方块，仅剩2D卷积、池化层的结构。初始化每一层卷积层的kernel权重参数时，为了减少模型训练所需收敛时间，采取正态分布的形式初始化参数。在Backbone的选取上，考虑到整体模型的参数量，应该选取卷积层数尽可能少、性能尽可能优秀的结构，且后续还具有Transformer结构，因此选择了考虑更多原始特征的ResNet系列网络作为Backbone。每个层除了卷积结构还加入了残差连接，将尺寸相同的上一个模块输出累加到下一个输出特征图上，使模型能够学习到卷积运算前的部分特征。

1. *Positional Encoding*

CNN具有优秀的图像特征提取功能，但是提取的多通道特征图并不具备良好的相对位置信息，即特征之间不具有信息特征，而Transformer模型的输入是具有上下文关系的词向量，因此需要给多通道特征添加位置信息即位置编码。基于残差网络中瓶颈层的思想，将位置信息累加到特征向量上，特征向量具有了位置信息。在Transformer中具有两种位置编码方法：第一种是基于词索引向量Embedding的思想，给予一个可学习的权重矩阵给输入向量，在训练模型参数的同时位置编码也通过学习得到；第二种方法是根据经验公式生成的位置编码向量。

生成位置编码前需要使用Mask，由于数据集中每张图片尺寸大小不一，训练过程是批训练方式，Transformer的特性可以输入长度不同的特征，但是需要将迭代的当前批次内所有图片填充值为0的空白像素点至当前批次图像中最大的宽度和高度以形成统一大小。因为填充了空白像素点，训练过程为了防止空白像素点干扰Transformer参数学习，需要在输入过程添加一组feature Masks，为后续Multi-head Attention (MHA)进行Masking操作。When batch size is 2，feature Masks生成如图3.

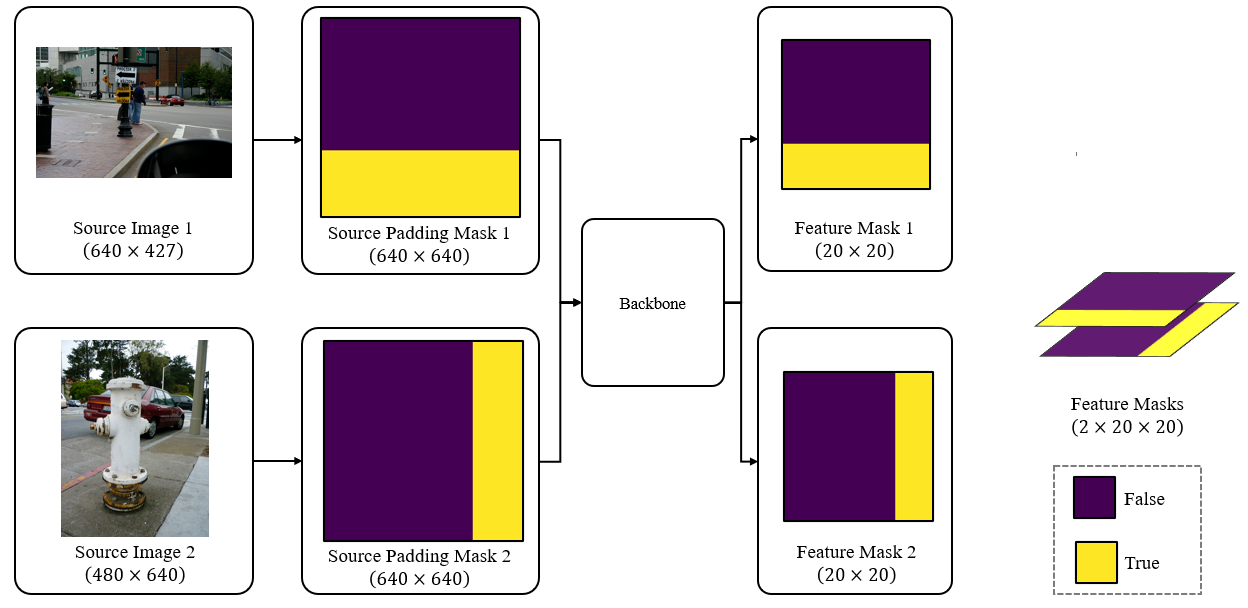


Figure 3. Generate batch feature mask.

其中Mask元素为False的位置表示该位置特征不需要进行屏蔽处理，相反当元素为True的位置在MHA的Masking中对特征向量的相应位置累加一个无穷小的数，从而使特征图在后续多头注意力的线性映射中相应数值也是无穷小，使注意力权重在对应位置上接近于0。

为了生成Positional Encoding，将采取正余弦函数的方式。根据Mask生成与其相反的二值化向量Not-Mask，即原来位置的元素由False变成True，True变成False，复制次Not-Mask以初始化于特征图尺寸相同的位置编码向量。随后对累计和横坐标方向的Cumulative sum，剩下的累计和纵坐标，将得到两组相对位置的向量如图4.

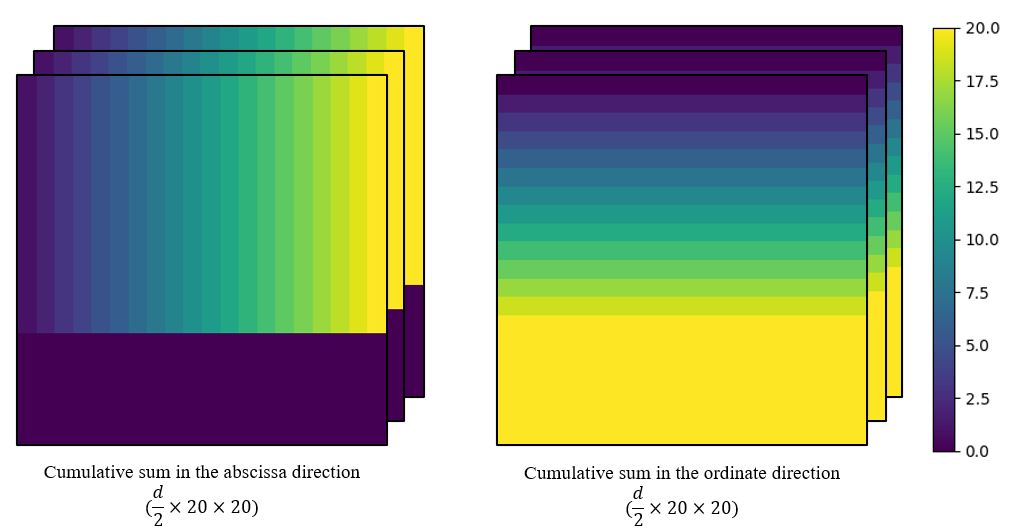


Figure 4. Two parts relative positions vector.

将嵌入维度均分两份进行不同方向的累积和，以达到同时考虑图像平面中横坐标和纵坐标之间的关系。随后使用正余弦函数对的元素进行非线性变换，从而生成具有相对位置信息的位置编码。对于位置、维度索引的值如式(1).

1. *Transformer*

DETR中Transformer结构与原始结构相似，输入向量调整为形状的3D特征，而原始结构的输入向量形状是2D，因此将拉平，即3D特征变成形状的Flatten Token。进入Transformer前需要另外经过一个 kernel大小的卷积层，以实现对多通道特征进行一定程度的降维。DETR中Transformer Encoder and Decoder结构如图5.

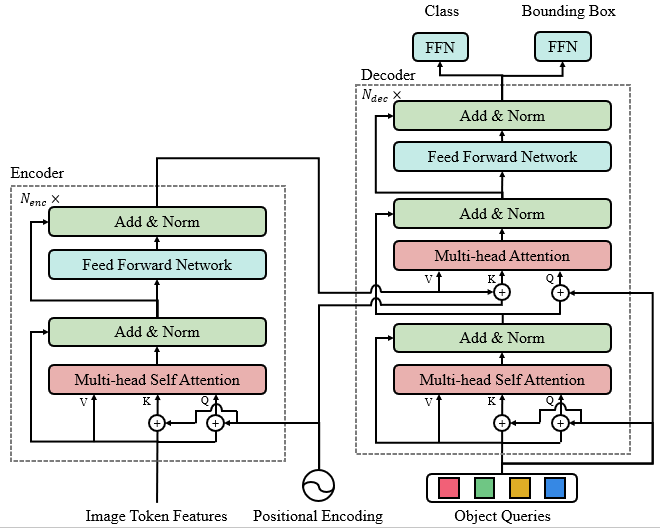


Figure 5. Detailed structure of Transformer in DETR.

Transformer的Encoder和Decoder分别由个Encoder blocks和个decoder blocks组成，每个encoder block和decoder block结构相似均有MHA、residual add、Layer Normalization (LN)、Feed Forward Network (FFN)，decoder block相比encoder block多了一个MHA去接受encoder的输出，而Multi-head Self Attention (MHSA)本质上与MHA相同，只是MHSA更侧重“自己对自己注意”，MHA的详细结构如图6.

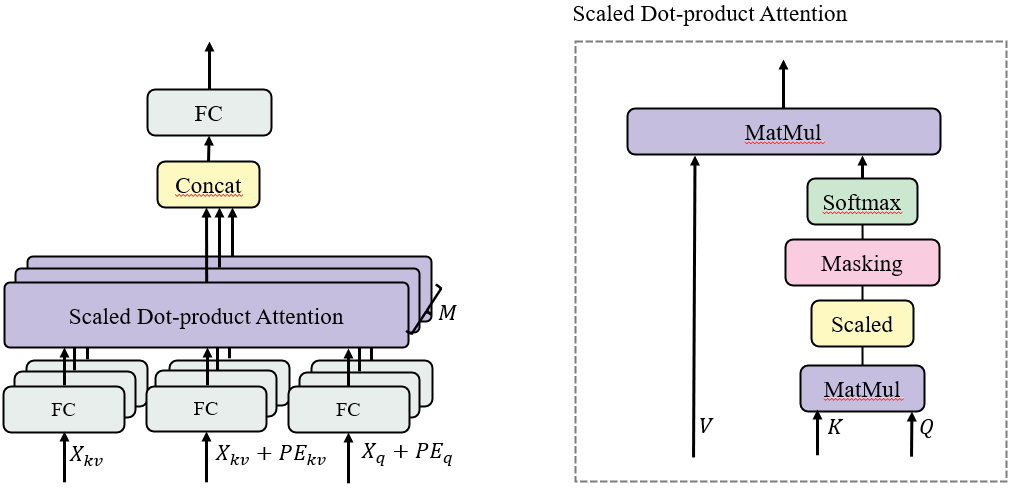


Figure 6. MHA and Scaled Dot-product Attention.

MHA本质上是一种软查询，是输入向量经过Fully Connected (FC)的线性映射后得到三个不同的矩阵，模型中和均来自同一个输入向量，从同一输入向量映射得到则为MHSA。在MHA中，将形状较大的输入向量拆分成了个头进行并向计算，最后拼接再一起后经过一个线性映射即得到与输入向量形状相同的输入。在单个头的注意力中，采取了Scaled Dot-product Attention运算方法，根据计算后得到一个方形矩阵，经过scaled缩小、Masking、Softmax后，变成了元素值在内的权重矩阵即注意力权重。注意力权重与相乘后得到的新向量便是根据生成的注意力权重处理后的特征向量，定义输入向量为，整个MHA计算公式如式(2)-(3).

其中，表示第个头中Scaled Dot-product Attention前的三个FC层权重矩阵。表示连接后FC层的权重。由此可得。经过MHA计算后，对特征进行残差累加和层标准化，以加强原始特征防止运算过程损失过多。在单个encoder or decoder block中，会包含一个FC-ReLU-Dropout-FC组成的FFN，将MHA运算后的特征进行经过先后两次线性映射完成向前传播，其中这两个FC层的权重维度为，即输入维度为的特征向量经过第一个FC层映射到维度空间中，经过第二个FC层映射回和输入向量相同形状的维度空间。

在原始用于Machine Translation的Transformer模型中，decoder的输入是从起始符号token的向量开始，预测下一个词并且添加到当前向量中得到的Token向量，再将该向量输入decoder中，以此类推直到预测的下一个词是结束符号token位置。但是对于DETR而言，在object detection上只需要预测class和bounding box即可，因此在decoder后通过共享参数的FFN即可。在decoder输入部分，原作者是定义了object queries，其本质是Embedding层的权重，以一种可学习的方式来完成以往object detection中繁琐的anchor box预处理，从而减少了对经验人士的依赖性，实现了端到端简易性，唯一缺点是增加了训练过程模型参数收敛时间。

1. *Mask head*

DETR是用于object detection的架构，使用此框架作为基础以完成panoptic segmentation，如同Mask R-CNN在共享参数FFN部分增加Mask head以生成分割图，mask head结构如图7.

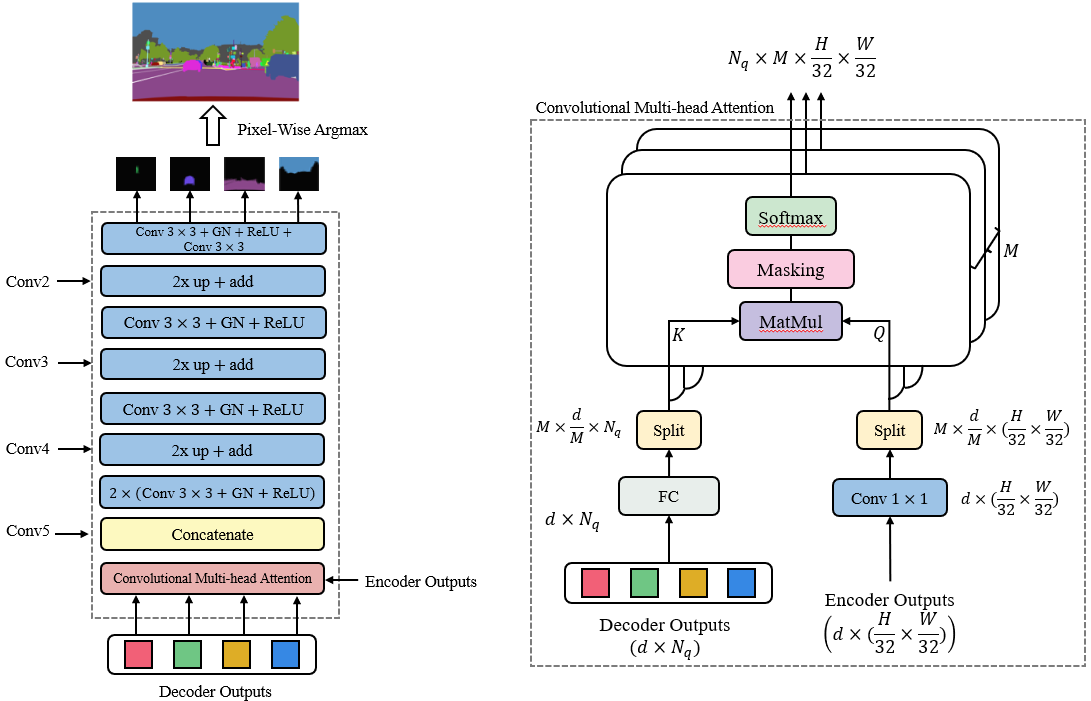


Figure 7. Mask head structure.

Convolutional Mult-head Attention是参考了MHA定义的一种处理encoder和decoder outputs方法，使用其软查询的特性将object queries的特征和encoder outputs的特征转换为对物体的“注意力”特征，通俗来说就是将stuff和things视为同一类型的对象，生成对个物体的特征图。 其中Masking和MHA的Masking类似，也是对填充空白像素处累加一个无限小的数，以此达到忽略这些空白像素的目的。当模型参数收敛后，这些“注意力”特征能够很好的表述物体在图像的位置、体积等信息，因此后续采取FPN-Style CNN的形式，根据Backbone的各层特征图通道数，对“注意力”特征图进行卷积运算和对应层特征图进行累加，学习原始图像中的细节。卷积运算后接着2倍的Nearest Interpolation上采样方式，逐渐将的实例特征图转换为的实例特征图，最后将实例特征图使用 kernel大小的卷积层，卷积运算后将得到每个物体的mask，根据这些mask可以得到物体在尺寸图中物体的相对位置和区域信息。

最后将生成的物体mask进行Pixel-wise Argmax筛选，融合出一张具有不同实例id的mask，最后采取Bilinear interpolation将实例id mask恢复至原始图像大小。上文提到mask head中是采用了类似MHA软查询机制的Convolutional Multi-head Attention以完成物体的像素点分割，因此decoder outputs同时可以根据共享参数的FPN以完成物体类别和bound box预测，即个特征图得到的个mask、class和bounding box是一一对应的，无需进行其他任何的后处理，一定程度上节省了很多工作量。

1. *Loss function*

由于DETR的共享参数FFN和mask head，因此使用segmentation、box、class三种类型损失以进行模型训练，如式(4).

1. *Hungarian matcher*

在模型推理阶段，会得到N个预测的集合，这些预测与Ground Truth (GT)进行匹配才能计算损失，因此需要使用Hungarian matcher将prediction和GT进行匹配。该算法会在预测对象和真实对象之间产生一个最佳的二分匹配，定义匹配的集合如式(5).

其中， 表示用最低损失在GT集合中搜索到个预测对象集合的最优配对； 表示个预测对象的集合；表示GT被（非物体类）填充数量至的集合；表示真实值和其匹配的预测值 之间计算的Pair-Wise Matching Cost，这个匹配代价同时考虑了类别、边界框与GT的相似性。对于的每个元素，定义；其中，表示对象的类别标签（包括）；表示真实框中心坐标、其相对于图像大小的宽度和高度的向量。对于的预测类别概率为 ，预测框为 的Pair-Wise Matching Cost如式(6).

根据计算每一个 和的每一个对象在类别、边界框的综合损失，损失最小的预测框 为真实框的最佳匹配框。接着计算上述所有配对的匈牙利损失如式(7).

1. *Bounding box loss*

上述的匈牙利匹配算法第二部分是对边界框进行评分，由于DETR是直接预测边界框，通过计算相对误差的范数损失和交并比对边界框评分，会因为尺度不同的大小边界框可能出现相对误差相似的情况，因此采取Generalized Intersection over Union (GIoU）损失和损失作为边界框评分如式(8-9)：

其中，是调整GIoU和范数损失比例的参数，。表示真实边界框$和预测边界框 之间交集的面积，表示并集的面积；表示最小闭包区域的面积，表示和 的最小闭包区域和并集的差集面积。

1. *Segmentation loss*

Segmentation mask可以看作二分类的像素级分类问题，该部分由Dice损失和Focal损失组成，即，如式(10-13)：

其中，同上是训练超参数；表示匹配的单个真实掩码和预测掩码概率。BCE指的是Binary Cross Entropy，用于二分类的损失函数。

1. *Classification loss*

分类损失由计数损失和类别损失组成：由于DETR直接生成对象预测框和分类，因此预测中会，统计非的数量和GT数量的误差将作为分类打分的计数损失；其次类别损失是根据匹配的预测框类别和GT类别进行交叉熵的计算，即回到多分类问题的损失计算，整个分类损失由计数损失和分类损失组成如式(14).

1. *Evaluation metric*

在传统的图像分割任务中，使用平均像素准确率作为评价指标，仅作用于像素级上；而实例分割任务中，使用Average Precision (AP)和IoU作为评价指标，仅作用于实例级上；因此全景分割的指标度量具有三大需求：(a)完整性：度量标准应以统一的方式处理Stuff和Things类，捕获所有方面；(b)可解释性：相较模型架构，评价指标应具有比较清晰的可解释性；(c)简易性：指标易于定义和实现，能够高效计算实现快速评估。因此便衍生出了统一处理Things和Stuff所有类的全景质量（Panoptic Quality, PG）作为全景分割的评价指标，如式(15).

其中，表示预测的单个对象；表示匹配的真实标注；TP (True Positives)，FP (False Positives)，FN (False Negatives)表示预测和GT匹配的分割片段、不匹配的预测片段和不匹配的GT集合。SQ是在IoU的基础上引入TP，综合考虑Things和Stuff的分割质量做出评分；RQ与二分类任务的F1分数相似，其性质和F1分数一样，综合了精度、召回率的评价指标。

1. Experiments
2. *Dataset*

MS COCO（Microsoft Common Object in Context）数据集是微软于2014年首次发布的一个大型数据集，该团队收集约33万张日常生活场景图片，并对其中的20万张图片进行150万个对象实例的人工标注，其包含80类的目标检测与分割、91类像素级的图像分割、人体关键点等任务的标注内容。在2017年提出的第二版Training和Validation新增了全景分割标注。

*B． Results*

模型训练和验证均在Linux系统的Intel(R) Xeon(R) Gold 6258R CPU @ 2.70GHz处理器，2张NVIDIA GeForce RTX 3080 Ti 12GB显卡的设备上完成。训练过程设置超参数如表1.

Table 1. Hyperparameters for training.

|  |  |  |  |
| --- | --- | --- | --- |
| Hyperparameter | Value | Hyperparameter | Value |
|  | 6 | Dropout | 0.1 |
|  | 6 |  | 1 |
|  | 100 |  | 1 |
|  | 8 |  | 5 |
|  | 256 |  | 1 |
|  | 2048 |  | 1 |
|  | 2 |  | 1 |

这里引入DETR base架构的已训练完成权重，只对mask head进行模型权重训练。采取AdamW作为优化器，训练100 epochs，得到的实验结果与其他模型的base对比如表2.

Table 2. Experimental results PQ comparison.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | PQ | SQ | RQ |  |  |  |  |  |  |
| PanopticFPN++ | 42.4 | 79.3 | 51.6 | 49.2 | 82.4 | 58.8 | 32.3 | 74.8 | 40.6 |
| UPSnet | 42.5 | 78.0 | 52.5 | 48.6 | 79.4 | 59.6 | 33.4 | 75.9 | 41.7 |
| DETR | **42.8** | 78.6 | **53.8** | 48.1 | 78.9 | 59.4 | **35.8** | **78.2** | **45.1** |

由上可得，新增mask head的DETR在COCO 2017 Validation实验中得到不错的结果，其综合实力超过同期其他专门用于panoptic segmentation的架构。在panoptic segmentation中，对于stuff的像素分割精度远超其他模型的，在things上略有不足。注意到无论是综合、things还是stuff，在RQ即对于实例分类上准确率都比以往的高，证明Transformer decoder的object queries得到的特征同时用于共享参数FFN和mask head时依然能保持特征信息的独立性和统一，生成的实例分割图和类别概率有着很强的相关性，与其他同用于panoptic segmentation的多分支任务架构而言，DETR基础的架构在完成task的同时获得更高质量的性能。可视化部分结果如图8.

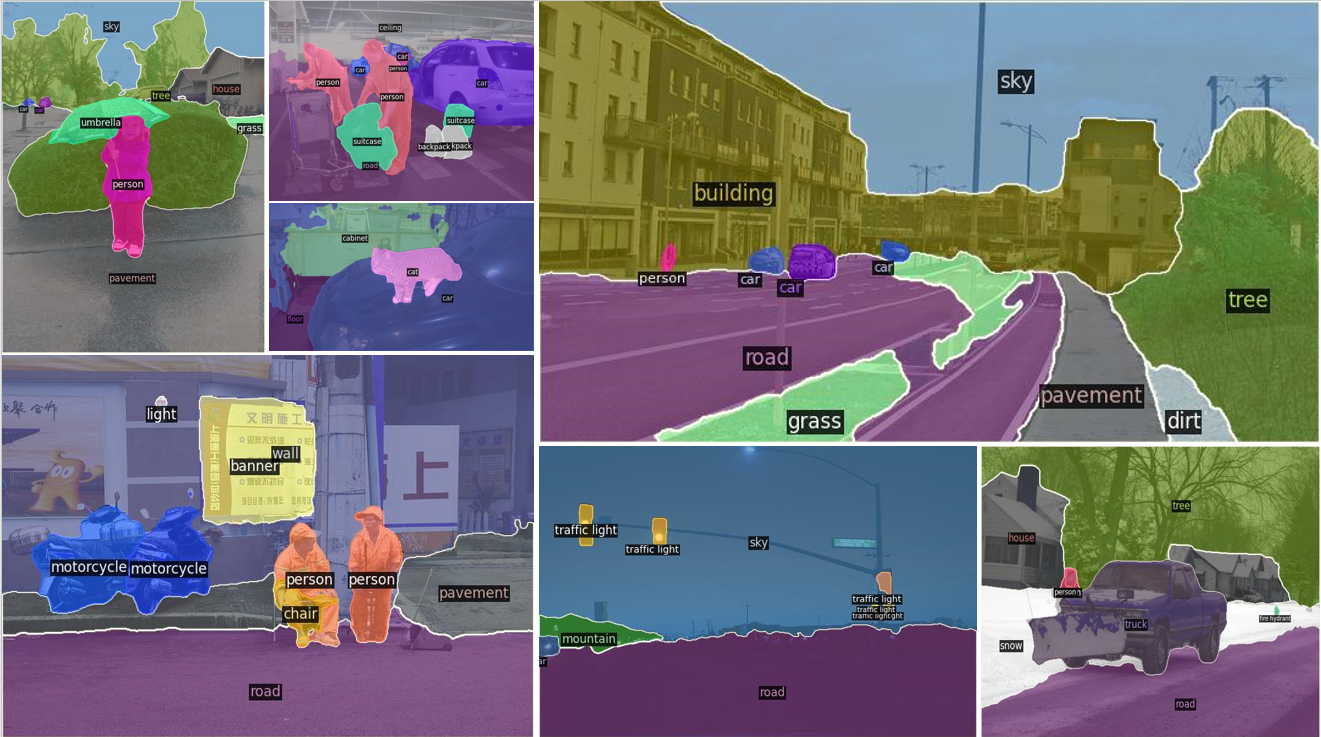


Figure 8. Visualization of small part segmentation figure results.

1. Conclusions

本文在端到端object detection架构的DETR基础上，新增了与Mask R-CNN相似style的mask head，以完成将特征转换成panoptic segmentation格式的分割图task。mask head中采取的FPN-style多尺度融合特征，损失函数增加的dice和focal loss，这两者对于panoptic segmentation任务中stuff和things样本数（像素点）极度不平衡的情况都具有一定程度的优化。对于最后输出部分的处理，共享参数FFN和基于MHA软查询的Convolutional Multi-head Attention mask head不仅妥善地处理了Transformer encoder和decoder的输出，还很好的解决了以往panoptic segmentation中实例分割像素区域和类别分类信息不相关的缺陷。以上仅仅是在DETR基础架构的基础上再训练mask head的参数，假如从0开始对整个架构训练可能能够达到更高精度。

Reference

1. Y. Lecun, L. Bottou, Y. Bengio, et al., “Gradient-Based Learning Applied to Document Recognition”, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-3324, 1998.
2. A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems*, vol. 25, pp. 1106-1114, 2012.
3. K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, *3rd International Conference on Learning Representations (ICLR 2015)*, 2015.
4. K. He, X. Zhang, S. Ren, et al., “Deep Residual Learning for Image Recognition”, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016*), pp. 770-778, 2016.
5. J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pp. 3431-3440, 2015.
6. O. Ronneberger, P. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich*, vol. 9351, pp. 234-241, 2015.
7. F. Milletari, N. Navab, S. A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”, *Fourth International Conference on 3D Vision (3DV 2016)*, pp. 565-571, 2016.
8. B. Hariharan, P. A. Arbeláez, R. B. Girshick, et al., “Simultaneous Detection and Segmentation”, *Computer Vision - ECCV 2014 - 13th European Conference*, vol. 8695, pp. 297-312, 2014.
9. K. He, G. Gkioxari, P. Dollár, et al., “Mask R-CNN”, *IEEE International Conference on Computer Vision* *(ICCV 2017)*, pp. 2980-2988, 2017.
10. T. Y. Lin, M. Maire, S. J. Belongie, et al., “Microsoft COCO: Common Objects in Context”, *Computer Vision - ECCV 2014 - 13th European Conference*, pp. 740-755, 2014.
11. D. Bolya, C. Zhou, F. Xiao, et al., “YOLACT: Real-Time Instance Segmentation”, *2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019*), pp. 9156-9165, 2019.
12. A. Kirillov, K. He, R. B. Girshick, et al., “Panoptic Segmentation”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 9404-9413, 2019.
13. A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is All You Need”, *Advances in Neural Information Processing Systems: Vol. 30*, pp. 5998-6008, 2017.
14. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, *9th International Conference on Learning Representations (ICLR 2021)*, 2021.
15. N. Carion, F. Massa, G. Synnaeve, et al., “End-to-End Object Detection with Transformers”, *Computer Vision - ECCV 2020 - 16th European Conference*, vol. 12346, pp. 213-229, 2020.
16. S. Zheng, J. Lu, H. Zhao, et al., “Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021*), pp. 6881-6890, 2021.
17. B. Cheng, A. G. Schwing, A. Kirillov, “Per-Pixel Classification is Not All You Need for Semantic Segmentation”, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021 (NeurIPS 2021)*, pp. 17864-17875, 2021.