目录

[论文内容解析 1](#_Toc100135846)

[场景图生成 6](#_Toc100135847)

[2022.4.4 7](#_Toc100135848)

2022.3.8

基础内容：

在计算机视觉领域，目前神经网络的应用主要有图像识别，目标定位与检测，语义分割。图像识别就是告诉你图像是什么，目标定位与检测告诉你图像中目标在哪里，语义分割则是从像素级别回答上面两个问题。在图像领域，语义指的是图像的内容，对图片意思的理解

## 论文内容解析

**① Document-level Relation Extraction as Semantic Segmentation**

基于语义分割的文档级关系提取

**摘要：**

文档级关系提取的目的是从一个文档中提取多个实体对之间的关系。先前提出的基于图或基于转换器的模型独立地利用实体，而不考虑关系三元组之间的全局信息。本文通过预测一个实体级关系矩阵来捕获局部和全局信息的问题，并行于计算机视觉中的语义分割任务。在此，我们提出了一个文档级关系的文档u型网络提取。具体来说，我们利用一个编码器模块来捕获实体的上下文信息，并在图像风格的特征图上使用一个u型分割模块来捕获三元组之间的全局相互依赖性。

**正文：**

**Introduction**

关系提取是信息提取的一个重要组成部分，以前的工作专注于识别单个句子中的关系，而不能识别跨句子的实体之间的关系。然而，在现实应用中，许多关系是用多个句子来表达的。根据\_\_的研究超过40%关系只能在文档级别上进行标识。因此，对于模型能够提取文档级的关系是至关重要的。

与句子中一个只包含一个实体对进行分类的句子级RE相比，文档级的RE需要该模型同时对多个实体对的关系进行分类。此外，一个关系中涉及的主语和客体实体可能出现在不同的句子中。因此，不能仅仅根据一个句子来确定一个关系。

为了提取这些句子间实体对之间的关系，目前大多数研究都构建了基于启发式、结构化注意或依赖结构的文档级图模块。

具体地说，多三元组之间的相互依赖性是有利的，可以为许多实体的关系分类提供指导。例如，如果美国的内部关系（马里兰州，国家，美国）已经被确定，那么美国就不可能有任何其他的人与社会关系，比如“是……的父亲”。此外，根据埃尔德斯堡位于马里兰州和马里兰州属于美国的三组，我们可以推断埃尔德斯堡属于美国。如上所述，每个关系三重组可以为同一文本中的其他关系三元组提供信息。

为了捕获多个三组之间的相互依赖关系，我们将文档级RE任务重新定义为一个实体级分类问题。它类似于语义分割（一项众所周知的计算机视觉任务），其目标是通过卷积网络用相应的表示类来标记图像的每个像素。受此启发，我们提出了一种新的模型，称为Document U-shaped Network (DocuNet)。它用语义分割来阐述文档级关系提取。通过这种方式，给定实体对之间的相关特征作为一个图像，该模型预测每个实体对的关系类型作为一个像素级掩模。特别地，我们引入了一个编码器模块来捕获实体的上下文信息，以及一个在图像风格特征图上的u形分割模块来捕获三元组之间的全局相互依赖性。我们进一步提出了一种平衡软极大（balanced softmax method）的方法来处理不平衡关系分布。我们的贡献可以总结如下：

①据我们所知，这是第一个将文档级RE视为语义分割任务的方法。

②我们引入了DocuNet模型来捕获本地上下文信息和文档级RE的三元组之间的全局相互依赖关系。

③在三个基准数据集上的实验结果表明，与基线相比，我们的模型DocuNet可以达到最先进的性能

**Related Work：**

以往的关系提取方法主要集中于识别句子中两个实体之间的关系。多种方法已经提出了有效处理句子级重构任务的方法。然而，句子级的RE面临着一个不可避免的限制，因为许多现实世界的关系只能通过阅读多个句子来提取。因此，文档级的RE吸引了许多研究人员。

文档级RE的各种方法主要包括基于图的模型和基于转换器的模型。基于图形的方法现在被广泛采用，因为它们在关系推理中的有效性和强度。Jia等人[2019]提出了一个模型，该模型结合了跨整个文档和子关系层次的不同文本跨度学习的表示。Cristovoulou等人[2019]提出了一种用于文档级RE的面向边缘的图神经模型(EoG)。Li等人[2020]用图增强的双注意网络(GEDA)描述了句子和潜在关系实例之间的复杂交互作用。Zhang等人[2020c]提出了一种新的基于图形的双层异构图(DHG)模型，其中包含一个结构建模层，然后是一个关系推理层。Zhou等人[2020]提出了一种全局上下文增强的图卷积网络(GCGCN)，由实体作为节点，实体对的上下文作为节点之间的边组成。Wang等人[2020a]提出了一种新的模型(GLRE)，该模型根据全局和局部实体表示以及上下文关系表示对文档信息进行编码。Nan等人[2020]提出了一种新的模型(LSR)，该模型通过自动诱导一个潜在的文档级图来实现跨句子之间的关系推理。Zeng等人[2020]提出了基于双图的图聚合推理网络（文档级RE。Xu等人[2021]提出了一种编码器-分类器重构模型(HeterGSAN)，该模型设法从图表示重建地面真实路径依赖关系。显式图推理可以弥合出现在不同句子中的实体之间的差距，从而减轻长距离依赖性，并实现良好的性能。

相比之下，考虑到转换器架构可以隐式地建模长距离依赖关系，一些研究人员直接利用预先训练过的语言模型，而无需生成文档图。Wang等人[2019]提出了一种使用BERT作为预训练词嵌入的DocRED两步训练范式。他们观察到关系分布的不平衡，并解开了关系的识别和分类，以便更好地推断。Tang等人[2020]提出了一个层次推理网络(HIN)来充分利用来自实体的丰富信息，通过句子和文档级别来执行层次推理。Zhou等人[2021]提出了一种基于BERT的自适应阈值化和局部上下文池化模型(ATLOP)。然而，以往的研究大多集中在局部实体表示上，而不管三组之间的高级全局连接，这忽略了多重关系之间的相互依赖性。

一方面，我们的工作受到[Jinetal.，2020]的启发，这是第一个考虑全球关系互动的问题，关于RE的研究很少。另一方面，正如这些研究[阮和Grishman，2015；2015；沈和黄，2016]所做的，卷积神经网络长期应用于关系提取领域，这启发我们注意CNN在图像风格特征图信息提取中的作用。因此，我们的工作也与[Liuetal.，2020]的研究有关，他将不完全话语重写定义为语义分割任务，并激励我们从计算机视觉的角度来研究RE问题。在这项研究中，我们利用了U-Net[Ronneberger等人，2015]，它包括一个捕获上下文的收缩路径和一个能够实现精确定位的对称扩展路径。据我们所知，这是第一个将RE定义为语义分割任务的方法。

**Methodology：**

我们首先介绍问题的定义。对于包含一组实体{ei}的文档d，其任务是提取实体对(es，eo)之间的关系。在一个文档中，每个实体ei可能会出现多次。为了建模es和eo之间的关系提取，我们定义了一个N×N矩阵Y，其中条目表示es和eo之间的关系类型。我们得到了矩阵Y的输出，类似于语义分割的任务。Y中的实体根据它们在文档中的首次出现进行排列.我们通过实体到实体的相关性估计得到特征图，并将特征图作为一幅图像。需要注意的是，输出实体级关系矩阵Y在语义分割中与像素级掩码w并行，它连接了关系提取和语义分割。该方法也可应用于句子级关系的提取。由于文档具有相对较多的实体，因此实体级关系矩阵可以学习更多的全局信息，从而提高性能。

**② Spatial-Temporal Transformer for Dynamic Scene Graph Generation**

**摘要：**

动态场景图的生成旨在生成给定视频的场景图。与从图像中生成场景图的任务相比，由于对象之间的动态关系和帧之间的时间依赖关系，允许进行更丰富的语义解释，因此它更具挑战性。在本文中，我们提出时空transformer (STTran)神经网络，它包括两个核心模块：（1）空间编码器：需要输入帧提取空间背景和原因的视觉关系框架（2）时间解码器：将空间编码器的输出作为输入为了捕捉帧之间的时间依赖性和推断动态关系。此外，STTran可以灵活地将不同长度的视频作为输入，而不需要剪辑，这对于长视频尤其重要。我们的方法在基准数据集行动基因组(AG)上得到了验证。实验结果表明，该方法在动态场景图方面具有优越的性能。此外，还进行了一组烧蚀研究，并证明了每个所提出的模块的效果是合理的。

**Introduction：**

场景图是一种结构表示，它将感兴趣的对象总结为节点，其关系总结为边。近年来，场景图已成功应用于不同的视觉任务，如图像检索、目标检测、语义分割、人宾交互、图像合成，以及图像字幕或视觉问题回答(VQA)等高级视觉语言任务。它被视为一种很有前途的整体场景理解方法，也是连接视觉和自然语言领域之间的巨大差距的桥梁。因此，场景图的生成任务在社区中引起了越来越多的关注。

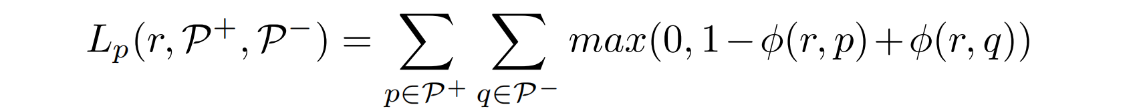
视频场景图生成任务（动态场景图生成）是一项更具有挑战性的新任务。静态场景图生成最流行的方法是建立在对象检测器上，该检测器生成对象建议，然后推断它们的关系类型和对象类。然而，物体在视频序列的每一帧都不确定一致，任何两个物体之间的关系可能因它们的运动而变化，这是动态的特征。在这种情况下，时间依赖性发挥了作用，因此，静态场景图的生成方法不能直接用于动态场景图的生成。

Spatial-Temporal Transformer (STTran)包含两个部分，spatial encoder和temporal decoder。其中spatial encoder负责处理单个帧中不同relationship的空间上下文，而在temporal decoder中不同帧之间的relationship可以共享信息并获取时间上下文。

Framework of Spatial-Temporal Transformer

我们利用对应的视觉信息 (visual vector), 主宾物体的空间信息 (bounding boxes坐标等)和语义信息来表征每个帧中的单独的relationship。同一帧中的所有relationships会作为一个batch被输入进spatial encoder，其中的self-attention module会对其进行处理。self-attention module的query和key即是原输入，因为relationship representation本身含有空间信息，且每个relationship之间并无先后顺序，所以我们在这里没有加任何encoding。在这点上，self-attention具有比RNN (比如Neural Motifs使用的LSTM)更强的解释性。然后我们通过sliding window对相邻帧采样，并将来自多个帧的relationships组成一个batch作为temporal decoder的输入。在temporal decoder中我们在query和key上加上一个训练得到的frame encoding用以区分来自不同帧的relationships，来自相同帧的relationships共享相同的encoding。不同于传统的Transformer结构，由于不需要cross attention我们仅在decoder中保留一个multihead attention module。

在现实中，同一个关系在语义上不是唯一的，在数据集中也存在很多object pairs被多个同义谓语标注 (例如person-holding-broom/person-touching-broom)。 我们引入Multi-label margin loss作为谓语分类的损失函数:



**③ Neural Motifs: Scene Graph Parsing with Global Context** (具有全局上下文的场景图解析)

**摘要：**

我们研究了产生视觉场景的结构化图表示的问题。我们的工作分析了主题的作用：在场景图中定期出现的子结构。我们在视觉基因组数据集中的这种重复结构提出了新的定量见解。我们的分析表明，对象标签可以高度预测关系标签，但反之亦然。我们还发现，即使是在更大的子图中，也有重复出现的模式：超过50%的图包含涉及至少两个关系的主题。我们的分析激发了一个新的基线：给定的对象检测，预测具有给定标签的对象对之间最频繁的关系，如在训练集中所示。这个基线在整个评估设置中平均相对提高了3.6%。然后，我们引入了堆叠主题网络，这是一种新的架构，旨在捕捉场景图中的高阶主题，它比我们的强基线进一步提高了平均7.1%的相对增益。

①超过大半的 images，实体对之间的可能关系高度依赖于实体对的标签，反之不是（object labels are highly predictive of relation labels but not vice-versa）

②同时当某一个 object label 出现时，另外一个 object label 的取值范围就会缩小（in general, the identity of edges involved in a relationship is not highly informative of other elements of the structure while the identities of head or tail provide significant information, both to each other and to edge labels）

**④ Graph R-CNN for Scene Graph Generation**

**摘要：**

我们提出了一种新的场景图生成模型，称为图R-CNN，该模型在检测图像中的目标及其关系方面非常有效。我们的模型包含了一个关系建议网络(RePN)，它能有效地处理图像中对象之间的潜在关系的二次数。我们还提出了一种注意图卷积网络(aGCN)，它可以有效地捕获对象和关系之间的上下文信息。最后，我们引入了一个新的评估指标，它比现有的指标更全面和现实。我们报告了使用现有的和我们提出的指标进行评估的场景图生成的最新性能。

给定一张image，首先从image中抓取出一组可见的物体，然后假定所有物体两两间都有联系，接着用学习出来的“relatedness”将那些不太可能的relation去掉，使得得到的graph变得稀疏一些，合理一些。最后，一个图卷积网络被用来整合global context，并且更新object node和relationship edge的标签。

**⑤ Scene Graph Generation by Iterative Message Passing** (通过迭代消息传递生成场景图)

**摘要：**

理解一个视觉场景，不仅仅是孤立地识别单个物体。对象之间的关系也构成了关于场景的丰富的语义信息。在这项工作中，我们明确地使用场景图来建模对象及其关系，这是一种基于视觉的图像图形结构。我们提出了一种新的端到端模型，从输入图像中生成这种结构化的场景表示。该模型利用标准rnn解决了场景图推理问题，并学习通过消息传递迭代地改进其预测。我们的联合推理模型可以利用上下文线索来更好地预测对象及其关系。实验表明，我们的模型在使用视觉基因组数据集生成场景图和使用NYU深度v2数据集推断支持关系方面明显优于以往的方法。

## 场景图生成

图片场景图生成任务（Image scene graph generation）目标是让计算机自动生成一种语义化的图结构（称为 scene graph，场景图），作为图像的表示。图像中的目标对应 graph node，目标间的关系对应 graph edge（目标的各种属性，如颜色，有时会在图中表示）。

形式化地，记关系集合为R，目标集为O, 目标位置为B （一般是 Bounding box），图像为 I，则图片场景图为G={ B, O, R}。根据给定条件多少，场景图生成任务可以由简单到复杂细分为以下几种：

关系分类(Predicate classification): 给定图中目标位置及类别，对关系进行归类，记为 P(R | O, B, I)。

场景图分类(Scene graph classification): 给定图中目标位置，对关系及目标关系进行归类，记为 P(R, O | B, I)。

场景图生成(Scene graph generation): 只给定图片，要求生成 scene graph，记为 P( G={R, O, B} | I)。

前一个任务可认为是后一个任务简化版。在评测模型能力时，一般需要考察模型在此三个任务的表现，以评价模型中关系分类模块、目标分类模块及目标定位模块的作用。

4.7 汇报需求

① Faster RCNN部分

② 关系预测部份

③ related work

④ 自己的思考

⑤ 论文内容

## 2022.4.4

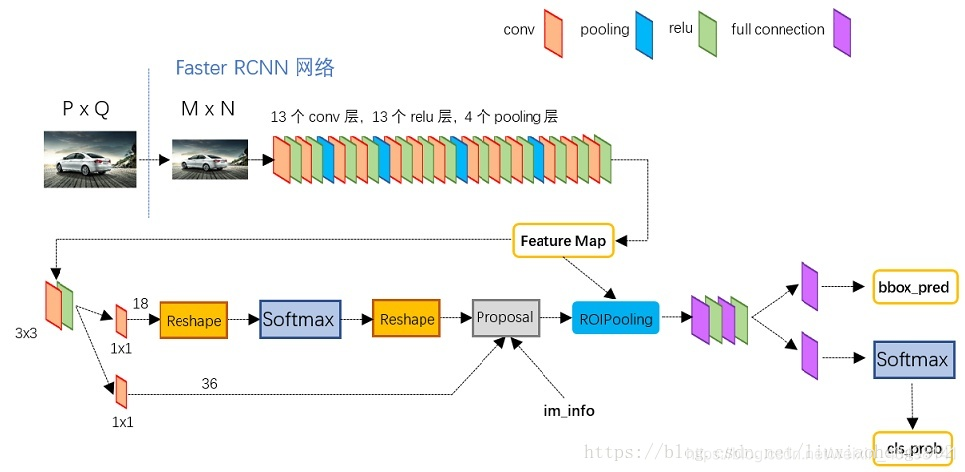
**主要工作内容：**

①提取图像中物体的方案；

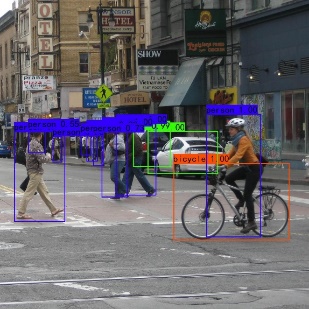
②针对提取到的物体有效地构建每对物体之间稀疏视觉关系的方案；

③将场景图生成看作实例分割任务，检测图像中物体之间存在的视觉关系并构建场景图；

对于任务①，我们采用Faster RCNN来进行目标检测。Faster RCNN是一个两阶段的目标检测算法，其原理图大致如下：



对于一幅输入的图片，其效果如下：可以看到标注了预测框、预测结果以及置信度。



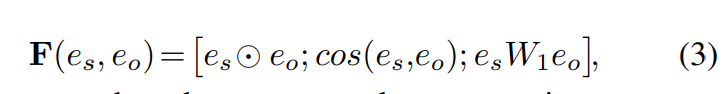
对于任务②来说，需要针对提取出的物体有效地构建每对物体之间的稀疏视觉关系。在任务①中，我们已经识别出了一些目标，每个目标都对应着一个特征向量，以及坐标信息。最简单的视觉关系肯定是将所有的目标进行全连接，但这样的复杂度会很高，因此必须想办法减少edge（relationship）的数量。

为了得到两个对象之间的关系，正常需要两两进入一个模型，输出两者之间的关系，此时的复杂度为O()。在Graph-RCNN这篇论文中用两个函数（个人理解为两个不同的MLP）将所有的对象分别映射到主语空间和宾语空间中的向量。经过矩阵乘法后即可得到邻接矩阵。其中的值即为不同对象的关系，由此可以删除掉一些边（关系）从而得到稀疏矩阵。论文的最后一步使用一个带注意力机制的GCN（个人理解这个注意力也是两层MLP）来预测关系，这一步相对比较简单。

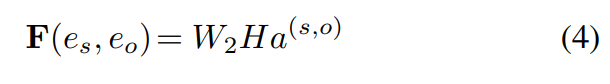
**接下来是我们的工作：**

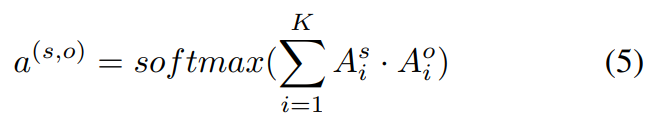
从Faster RCNN中我们已经得到了n个目标以及对应的特征向量。对于这n个实体（目标/对象）来说，他们的关系应该是一个n×n维度的矩阵。如果考虑到主客关系（例如 person wearing shirt），那么这还是一个不对称的矩阵。如何处理这个矩阵，实现矩阵的稀疏化以及关系的预测。我们想到了利用文档级语义关系提取。文档级语义关系提取将NLP的问题转化成了CV中的语义分割问题，从而得到语义关系；而我们当前也是有了一堆实体，需要预测他们的关系，那么这两个问题就存在一定的相似度。除此之外，原本将NLP问题转化为CV问题就是为了增加关系预测的全局性，这也是我们场景图生成中所需要的。

设共有n个实体，我们记第s个实体和第o个实体分别为和，接着，我们基于entity-entity相似度计算实体级的关系矩阵。这里最简单就是将两个实体的逐元相似度，余弦相似度，双线性相似度拼接起来，如公式：



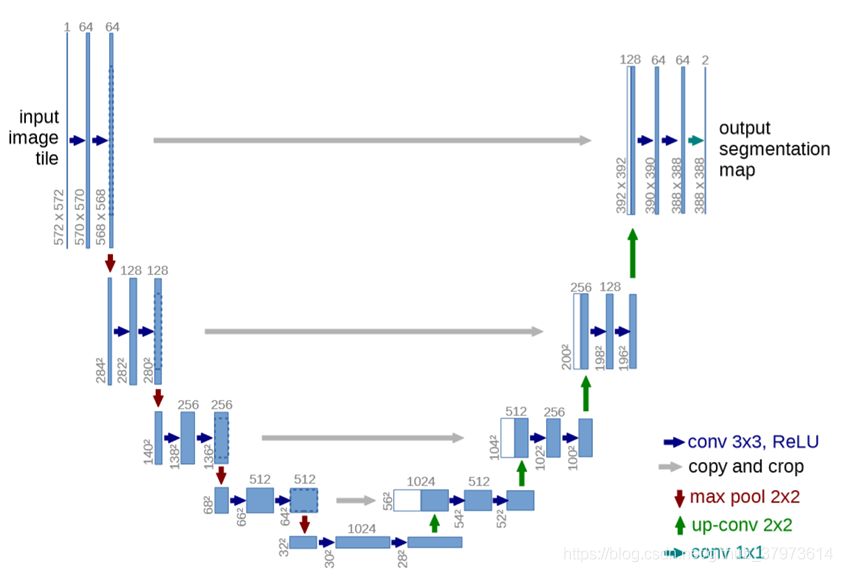
也可以使用带有一定注意力机制的变换：





由此我们得到了输入实体关系矩阵F。该矩阵为n×n×D维，D为第一步的特征向量经过相似性变换后的深度。

接下来，我们要将F输入进一个U-net网络，U-Net网络的结构大致如下



对于这个U-Net网络，我们输入的是实体关系矩阵F（n×n×D），而我们期望得到的输出是一个带有实体之间关系置信度的稀疏矩阵。首先我们要确定输出的维度。因为这是一个多分类问题，所以输出的应该是预测矩阵（n×n×c）其中c为实体关系的种类。输出矩阵的某一层的第（i，j）个元素代表第i个实体作为主语对第j个实体的关系相对于这一层对应关系的置信度。

由于两个实体之间的关系并不唯一，所以这是一个多分类问题，最后我们要经过softmax并对深度进行pooling的到最后的结果。因为大部分实体之间其实没有什么关系（即矩阵是稀疏的），所以我们在分类时使用一种“balanced softmax method”。首先引入一个格外的类别零，我们希望目标的分数都大于某个值，否则就视为这两个实体之间没有关系，将这两个实体之间的关系清零，以此来达到“剪枝“的目的。选取损失函数：

经过上述操作，我们就已经得到了所需的结果。

**存在的问题/创新点/相关工作：**

目前最大的问题是我们还没有实测U-Net网络是否有能力进行这个分类，因为在文档级关系提取中，实体的特征向量比较简单，实体之间的关系也比较单纯，甚至是对称阵，而在我们的第一步Faster RCNN中检测到的目标/实体的特征向量是1×1×m，这个m动则几千维（4096），U-Net网络是否可以能够达到效果还是个未知数，需要编程训练才知道结果。其次，在最后一步的balanced softmax method中的损失函数比较简单，是否能够达到效果也是一个未知数，可以考虑在损失函数中添加关于候选框的信息。

创新点在于将实体之间的关系提取和文档级实体关系提取的方法统一，将一个CV的问题简化为NLP的问题，再简化回到CV中的语义分割，将复杂的问题一步一步抽丝剥茧为相对简单的问题。同时语义分割能很好的吸收全局信息，这也正是我们场景图生成任务中所需要的。

目前应该还没有人用文档级关系提取的思路来做场景图生成。G-RCNN论文中第二步的思路虽然和我们类似，但是他的第二步仅用来去除掉不需要的枝叶，而关系预测是放到了第三步，使用了一个带注意力机制的图卷积神经网络。相比之下我们则是一步到位直接得到预测结果。