基于边缘增强的图谱排列网络和词对关系标签的多模态实体关系联合提取

李渊[1, 2], 蔡毅[1, 2, 3]*, 王进[4], 李青[5]
[1]华南理工大学软件工程学院，广州，中国
[2]中国教育部大数据与智能机器人重点实验室（SCUT）。
[3]彭程实验室，中国深圳。
[4]云南大学信息科学与工程学院，云南，中国。
[5]香港理工大学计算机系，香港，中国 seyuanli@mail.scut.edu.cn, ycai@scut.edu.cn, wangjin@ynu.edu.cn, qing-prof.li@polyu.edu.hk

摘要

多模态命名实体识别（MNER）和多模态关系提取（MRE）是多模态知识图谱构建任务中的两个基本子任务。然而，现有的方法通常单独处理两个任务，忽略了它们之间的双向互动。本文首次提出将MNER和MRE联合起来作为一个联合多模态实体-关系外延任务（JMERE）。此外，目前的MNER和MRE模型只考虑将视觉对象与文本对象进行对齐。

视觉和文本图中的实体，但忽略了实体-实体关系和物体-物体关系。为了解决上述挑战，我们为JMERE任务提出了一个边缘增强的图形对齐网络和一个词对关系标签（EEGA）。具体来说，我们首先设计了一个词对关系标签，以利用MNER和MRE之间的双向互动，避免错误传播。然后，我们提出了一个边缘增强的图形对齐网络，通过对齐交叉图中的节点和边缘来增强JMERE任务。与以前的方法相比，所提出的方法可以利用边缘信息来辅助对象和实体之间的对齐，并找到实体-实体关系和对象-对象关系之间的相关性。实验显示了我们的模型的有效性[1]。

简介

多模态命名实体识别（MNER）和多模态关系提取（MRE）是多模态知识图谱构建的两个基本子任务（Liu et al. 2019; Chen, Jia, and Xiang 2020），其目的是通过将图像作为额外输入来扩展基于文本的模型。以前的工作通常认为MNER和MRE是两个相互依赖的任务（Lu等人，2018；Moon，Neves和Carvalho，2018；Wu等人，2020b；Yu等人，2020；Zheng等人，2021c；Zhang等人，2021a），忽略了这两个任务之间的互动。最近，将NER和RE联合起来作为联合实体-关系提取任务在文本场景中引起了很多关注，这可以利用任务之间的双向作用并提高它们的性能（Wei et al. 2020; Yuan et al. 2020a,b）。如图1所示，如果我们提取到（**库里**，**NBA**）的实体类型是*Per*和*Org*，那么他们的关系就不应该是*peer*。否则，如果我们知道实体对（**Curry**, **Thompson**）的关系是*peer*，那么他们的实体类型应该是*Per*和*Per*。因此，NER可以促进RE。同时，RE对NER也是有益的。

多模式实体关系联合提取(JMERE)

图1：联合多模态实体关系提取（JMERE）任务的说明性例子，其中Per、Org和Misc表示为人、组织和杂项的实体类型。

然而，据我们所知，将MNER和MRE联合起来作为多模态实体关系提取任务（JMERE）在多模态情况下还没有被研究。与单独的任务相比，JMERE任务需要从视觉中提取不同的特征信息。如图1所示，对于MNER任务，如果模型能够从图像中捕捉到人的对象，例如多个人的轮廓（蓝框），它有助于识别文本中的人的实体。同时，MRE任务需要提取物与物之间的关系，例如，如果我们知道*持有的*是**人0**和**奖杯**之间的关系，我们可以理解实体**Thompson**和**O'Brien Trophy**之间*授予的*关系。因此，我们认为JMERE任务应该将实体与物体以及实体与实体之间的关系（在文本中）与物体与物体之间的关系（在图像中）统一起来。最近的MNER和MRE研究（Zhang等人，2021a；Zheng等人，2021a）在由物体和词的潜在关系构建的虚拟和文本图中，将实体与物

*对应的作者：Yi Cai (ycai@scut.edu.cn)

[1]代码和附录可在https://github.com/YuanLi95/EEGA-for-JMERE。

体对齐，如图中的红线所示。

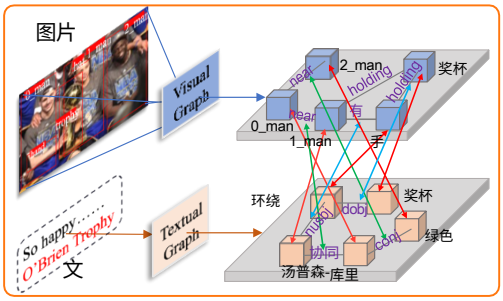


图2：说明交叉图中的节点（红线）和边缘排列（蓝线和绿线）的例子。

2.然而，这种方法只考虑到节点的排列，在忽略了交叉图的边缘对齐。如图2中的蓝线和绿线所示，交叉图中的边缘信息可以辅助对齐节点，并包含关于文本实体之间关系的线索。

此外，管道框架方法是解决JMERE任务的一种直观的方法。它用MNER方法提取实体，然后用MRE方法对其关系进行分类。然而，流水线框架只是通过MNER的结果使MRE受益，并且受到错误预测的影响（Ju等人，2021）。如图1所示，如果MNER提取的（**Curry**, **NBA Stars**）的实体类型是*Per*和*Misc*，结果应该是不正确的。受基于方面的情感三联体提取任务中的网格标记方案的启发（Wu等人，2020a），我们首先为JMERE任务采用了一个词对（w_i,w_j）分类方案，即词对关系标记法。这个方案同时训练MNER和MRE任务，利用它们之间的互动，避免管道框架造成的错误传播。如图3所示，词对（**Curry**, **Curry**）和（**Thompson**, **Curry**）的词对关系标签是*Per*和*Peer*，表示**Curry**属于一个人，*Peer*表示这个**Curry**和**Thompson**之间的关系。

为了应对上述挑战，我们提出了一个边缘增强图对齐网络（EEGA）和词对关系标签，通过在交叉图中同时对齐对象与实体（如**0人**与**Curry**和**tro-phy**与**trophy**）以及对象-对象关系与实体-实体关系（如*near*与*conj*和*holding*与*dobj*）来增强JMERE。EEGA的整体框架如图4所示。具体来说，我们使用图形编码器层，利用预先训练好的模型从输入的文本-图像中构建文本和视觉图形。然后，我们提出一个边缘增强的图形对齐模块，用Wasserstein距离来对齐交叉图中的节点和边缘。同时，该模块可以利用边缘信息来辅助物体和实体之间的对齐，并找到实体-实体关系和物体-物体关系之间的相关性。最后，我们设计了一个多通道层，从不同的角度挖掘词与词的关系，得到最终的词对表示。

我们的主要贡献可以概括为以下几点：

- 我们首次提出了联合多模态实体再提取（JMERE）任务，以处理多模态的NER和RE任务。同时，我们设计了一个词对对JMERE进行关系标记。这个方案可以利用MNER和MRE之间的双向互动，避免流水线框架造成的错误传播。
- 我们提出了一个边缘增强的图形对齐网络（EEGA），通过同时对齐交叉图中的节点和边缘来增强JMERE任务。与以前的方法相比，EEGA可以利用边缘信息来辅助对象和实体之间的对齐，并找到实体-实体关系和对象-对象关系之间的相关性。
- 我们对收集到的JMERE数据集进行了广泛的实验，实验结果证明了我们提出的模型的有效性。

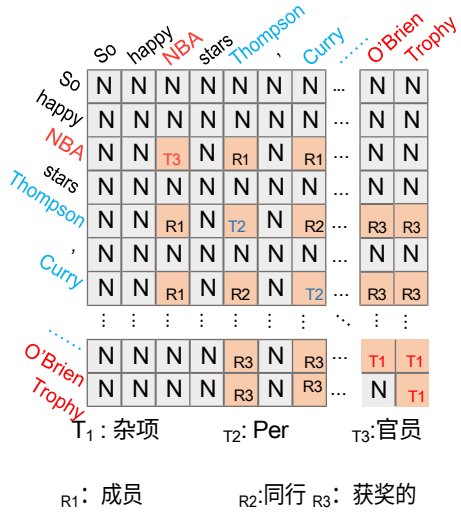


图3：JMERE任务中的一个词对关系标签的例子。

相关作品

知识图谱构建任务的关键部分（Chen, Jia, and Xiang 2020; Chen et al. 2022b,a），命名实体识别（NER）和关系提取（RE），引起了研究人员的广泛关注（Vashishth, Joshi, and Suman 2018; Wen et al. 2020; Li et al. 2020; Ren et al. 2020; Nasar, Jaffry, and Malik 2021; Zhao et al. 2021）。以前的研究主要集中在单一模式上。随着社交平台上多模态数据的日益普及，一些研究开始关注多模态NER（MNER）和多模态RE（MRE），其目的是将图像作为文本的补充，更好地识别实体及其关系。根据图像-文本对齐的对象，目前MNER和MRE的方法可以分为图像对齐方法、对象对齐方法和节点对齐方法。

图像对准方法

以前的研究通常使用RNN（递归神经网络）来编码文本，而CNN（卷积神经网络）则将图像编码为矢量。然后，设计一个隐性的交互模块来模拟文本和图像之间的信息。

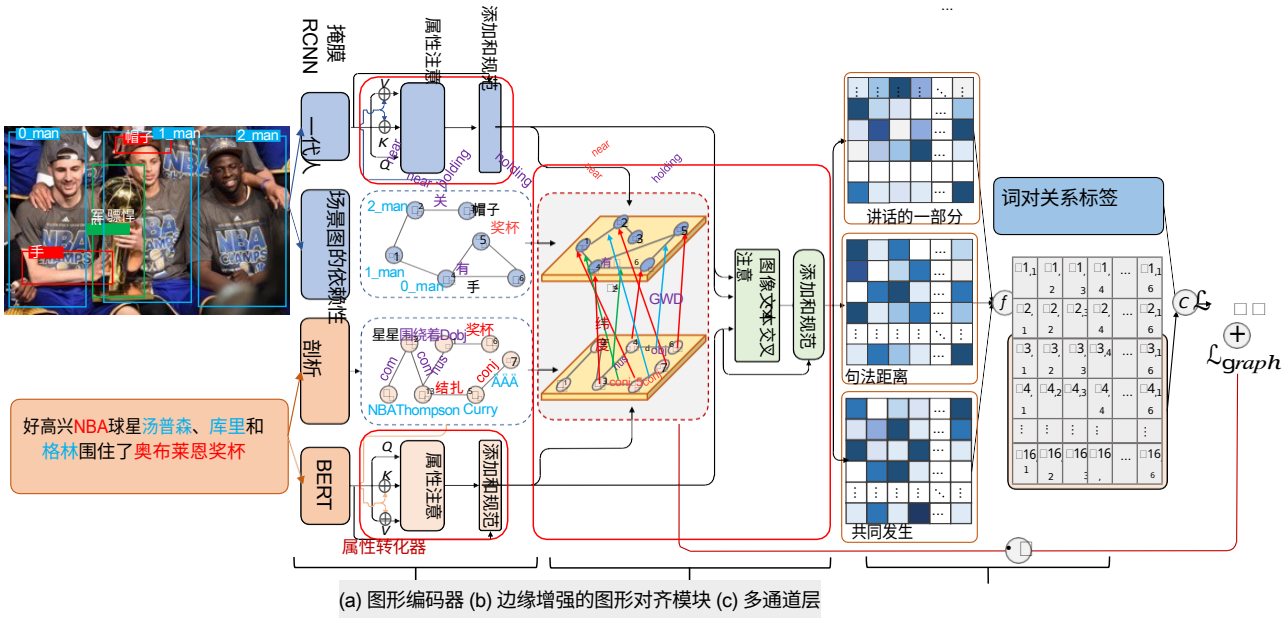


图4：JMERE的EEGA整体框架，其中com和nus表示句法依赖关系compound和nsubj。

MNER任务的模式（Zhang等人，2018；Lu等人，2018；Moon, Neves, and Carvalho，2018）。例如，Zhang等人（2018）构建了一个MNER数据集，并提出了一个基于双向长期记忆网络的基线模型，使用注意力机制将图像表示与文本对齐。然而，将图像编码为一个vec- tor不能有利于提取不同类型的实体，例如，**Curry**（*Per*）和**O'Brien Trophy**（*Misc*）。

对象排列方法

为了解决图像对齐方法的局限性，之前的模型使用Mask-RNN或Fast-RNN（He等人，2017）提取不同的视觉对象，并将视觉对象与文本表示法对齐（Wu等人，2020b；Yu等人，2020；Zheng等人，2021c；Zhan等人，2021a；Xu等人，2022）。吴等人（2020b）提出了一个交互式注意力结构，以将文本与视觉对象对齐。此外，Zheng等人(2021c)设计了一个具有对抗策略的门控双线性注意力网络（Kim, Jun, and Zhang 2018），以更好地从图像中提取细粒度的物体。然而，对象对齐方法没有考虑实体-实体和对象-对象的关系，该模型将无效地匹配重叠的视觉对象与文本实体。例如，文本中的**奖杯**可能与**1个人**对齐，因为**1个人**包含**奖杯**的区域。

节点对齐方法

为了解决上述限制，目前最多研究将实体与对象在由对象和词的潜在关系构建的视觉和文本图中进行对齐（Zhang等人，2021a，b；Zheng等人，2021a）。Zhang等人（2021a）提出了一个基于图的多模态融合模型，该模型基于句法依赖的文本图和全连接的视觉图，以利用模块来对齐文本图和视觉图中的节点。然而，这些节点对齐方法只考虑交叉图中的节点，而忽略了边缘信息。交叉图中的边缘信息可以有效地提高节点的匹配精度，并包含一些关于实体之间分类关系的线索。

任务定义和词对关系标记

任务定义

多模态实体关系的联合提取任务是这样的：给定一个输入文本$w = w_1, w_2, , w_n$和一个相应的图像I，模型需要提取的是一组五元组$y = \{(e_1, t_1, e_2, t_2, r)\}$，其中有一个五元组。，其中

(e, t, e, t, r) 代表C -第五个五元组，包括$_{1122}$ $_{c}$ $_{c=1}^{c}$

不同模态的细粒度语义对齐。在MRE任务中（Zheng等人，2021b），Zheng等人（2021a）设计了一个图的对齐方式

两个实体e_1和e_2与相应的实体类型t_1和t_2，其中$e_1 = e_2$。此外，r表示实体e_1和e_2之间的关系。图1给出了一个例子来更好地理解JMERE任务，其目的是提取所有的
五元组，例如，（**Thompson**, *Per*, **NBA**, *Org*, *Member of*），其中*Per*和*Org*表示**Thompson**和**NBA**的实体类型，*Member of*表示它们的关系类型。

词对关系标签

受基于方面的情感三联体提取中的网格标记方案的启发（Wu等人，2020a），我们设计了一个词对关系标记，以在一个步骤中提取JMERE的所有元素。通过词对关系标签，JMERE的任务被转换为提取每个词对（w_i, w_j）之间的关系Y，避免了管道框架造成的错误传播。这些关系可以解释如下、
而我们也在图3中给出了一个例子，以更好地理解词对关系标签。

- *N*表示该词对没有任何关系。

- r表示该词对属于一个实体类型，在以前的工作中包含4个定义的类型（Zheng等人，2021c）。
- R表示该词对属于定义的关系（Zheng et al. 2021a），每个词是一个实体。

边缘增强的图形对齐网络 图4显示了拟议模型的整体结构，由三个部分组成：图形编码器、边缘增强的图形对齐模块，以及多语言通道层。图形编码器层使用预先训练好的模型来构筑输入的文本和视觉图形。为了提高捕获边缘信息的能力，我们不直接将文本和视觉表示送到下一个模块，而是送到一个属性转化器。然后，为了更精确地匹配对象与实体，并从视觉图中捕捉实体与实体的关系线索，我们使用跨图最优转换方法（Chen等人，2020）与Wasserstein距离和Gromov-Wasserstein距离同时对齐。

交叉图中的节点和边。最后，我们提出了一个多通道层，使用加权图卷积
网络(W-GCN)来挖掘单词的潜在关系。
从多角度看，对。下文对每个组成部分进行了详细描述。手稿的代码
将在最终版本中公布。

图形编码器

文本图。 在形式上，我们首先使用依赖性parse toolkit[2] 来构建文本图。如图4（a）所示，在解析之后，给定的句子被转换为
成一个文本图$G_T = \{V_T, E_T\}$，其中$V_T \in \mathrm{R}^n$ 和$E_T \in \mathrm{R}^{n\times n}$ 表示句法依赖性的节点和边、
分别是，G_T是一个无方向的自环图。平均值-
同时，我们用$A_T \in \mathrm{R}^{n\times n}$ 来表示相邻的掩码马
trix为文本图，其中 $A_{T_{i,j}} \in \{0,1\}$ 说明是否
之间有一条边 w_i^T和 w_j 或不。此外、

节点V_T被送入BERT，得到$X_T \in \mathrm{R}^{n\times d_T}$。同时，一个边缘转换矩阵被用来映射边缘
类型E_T成一个可训练的向量，并获得边缘可训练的矩阵$Z_T \in \mathrm{R}^{n\times n\times d_Z}$、

$$X_T = BERT(w) \tag{1}$$

其中$BERT$表示作为文本编码器的BERT，d_T是BERT输出的尺寸。

视觉图。在以前的多模态任务中，物体被认为是图像的语义信息。如图4（a）所示，我们通过使用场景图生成模型（Tang等人，2020）将图像转换为视觉图$G_I = \{V_I, E_I\}$（Mask-RCNN作为后骨）。此外，我们只考虑具有最高物体分类分数的前k个突出的目标作为有效的视觉对象，并利用充分的视觉信息而忽略不相关的信息。因此，最后一个节点代表$V_I \in \mathrm{R}^k$ 由Mask-RCNN检测到的突出对象，$E_I \in \mathrm{R}^{k\times k}$ 表示视觉关系集，如位置关系（例如，*靠近*和*在前面*）和附属关系。

[2]https://spacy.io/models。

关系（例如，*持有*和*佩戴*）。我们用$A_I \in \mathrm{R}^{k\times k}$来表示视觉图的相邻矩阵。因此，视觉图中的最终节点向量$X_I \in \mathrm{R}^{k\times d_I}$被定义为、

$$X_I = Mask\text{-}RCNN(I) \tag{2}$$

其中，d_I是Mask-RCNN的隐藏维度，最终的边缘向量$Z_I \in \mathrm{R}^{k\times k\times d_Z}$是以与文本图相同的方式获得。

属性转化器。我们提出了属性关注（*At-Att*），通过将边缘类型纳入转化器的自我关注中的键和值，作为属性转化器，它可以更新模型间的节点状态，同时有效地纳入交叉图中的关系边缘（如文本图中的*subj*和*comp*，以及视觉图中的*holding*和*wear*）。

由于视觉图和文本图是两个包含不同模态信息的语义单元，我们使用类似的操作对它们进行建模，但有不同的派别。
参数。因此，*第*i*个*令牌的隐藏表征

$H_T^i \in \mathrm{R}^{n\times d_T}$在文本模式中被定义为、

$$\begin{aligned} H_T^i &= At\text{-}Att(X_T^i, Z_T^i, A_T^i) \\ &= \mathrm{Softmax}\left(\frac{Q_T^i (K_T^i)^\top}{\sqrt{d_T}}\right) V_T^i \end{aligned} \tag{3}$$

其中，$A_T^i \in \mathrm{R}^{1\times n}$ 是*第*i*个*节点的邻接掩码集、
$Q_T^i \in \mathrm{R}^{1\times d_T}$, $K_T^i \in \mathrm{R}^{n\times d_T}$，和$V_T^i \in \mathrm{R}^{n\times d_T}$是矩阵
将文本中*第*i*个*词的查询、键和值相应打包，其定义为：、

$$\begin{aligned} Q_T^i &= W_Q X_T^i \\ K_T^i &= W_K X_T^i + W_K^Z Z_T^i \\ V_T^i &= W_V X_T^i + W_V^Z Z_T^i \end{aligned} \tag{4}$$

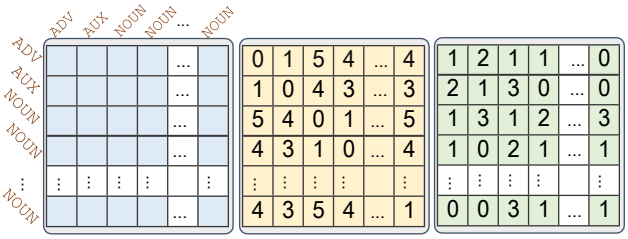
属性转换器的其他操作是一致的。

帐篷与香草转化器：H_T使用前馈网络（FFN）和层规范化（Layer- Norm）添加到X_T，得到文本表示$H_T \in \mathrm{R}^{n\times d_T}$。
我们使用类似的操作来获得视觉代表。
命名。特别是，我们使用一个变维的FFN来匹配对象H_I和标记H_T的维度。因此，图编码器的图像表示可以表示为$H_I \in \mathrm{R}^{k\times d_T}$。

边缘增强的图形对齐模块

给出文本图$G_T = \{H_T, Z_T\}$，以及视觉图$G_I = \{H_I, Z_I\}$。我们的目标是同时对齐交叉图的节点和边，并将匹配的语义信息从对象转移到实体。对于-
首先，我们使用最优传输方法（Chen et al. 2020）来匹配交叉图中的节点和边。此外，我们使用image2text attention将匹配的语义信息从视觉对象转移到文本模式，并获得精炼的文本表示。

边缘增强图的最优传输。为了明确地鼓励同时对准交叉图中的节点和边，我们最初应用了最优传输方法



(a) 语篇 (b) 句法距离 (c) 共同出现

图5：给定句子的多通道矩阵。

转移学习中提出的。如图4（b）所示，与考虑文本和图像为全连接图的原始最优传输方法不同，我们只考虑交叉图中具有相邻关系的节点和边。特别是，交叉图的匹配采用了两种类型的距离：（1）用于节点匹配的Wasserstein距离（WD）（Peyre′, Cuturi等人，2019）（红线）；

(2）用于边缘匹配的Gromov-Wasserstein距离（GWD）（Peyre′, Cuturi, and Solomon 2016）（蓝色和绿色线条）。形式上，$D_{wd}(H_I, H_T)$被测量为匹配节点H_I到H_T的最佳运输距离，其定义为：：

$$D_{wd}(H_I, H_T) = \min_{k} \sum_{i=1}^{n} \sum_{j=1} T_{i,j} \cdot c(H^i, H^j) \quad (5)$$

其中$c(H^i, H^j)$表示x的余弦距离^{i}到x^i_T，其定义为$c(H^i, H^j) = 1 - \frac{(H^I)^{H_T}}{\|H^i_I\|\|H^j_T\|}$。矩阵$T$是运输信息流，其中$T_{i,j}$代表从节点$x^i$转移到$x^j$的成本量。

然后，我们使用Gromov-Wasserstein距离（Peyre、Cuturi, and Solomon 2016）来衡量相似性分数D_{gwd}交叉图中的边缘通过计算节点对之间的距离，其定义为、

$$d_{gwd}(h_I, h_T, h_{i'}, h_{t'}) = \sum_{i=1}^{n} \sum_{j=1} T_{i,j} T_{i',j'} \cdot L(h_i, h_{i'}, h_t, h_{t'}) \quad (6)$$

其中H^{ij}_T和H^{ij}_I是文本中的相邻节点集。分别是$_{HT}$和$_{HI}$的视觉图和可视化图，以及$L(\cdot)$被认为是交叉图的距离成本。边缘$(H_i, H_{i'})$到$(H_j, H_{j'})$，即$L(H_i, H_{i'}, H_j, H_{j'}) = c(H_i, H_{i'}) - c(H_j, H_{j'})$。现在的学习矩阵$T$为注意到一个运输计划，该计划有助于对准交叉的边缘。

图。

我们使用一个统一的求解器，并使用Sinkhorn算法（Cuturi 2013）与熵规整器（Benamou et al. 2015）来迭代优化成本D_{wd}和D_{gwd}.因此，优化交叉图的对象损失函数是、

其中α是用于平衡成本重要性的超参数。然后，我们使用image2text关注，将视觉语义信息有效地转化为文本表示，表示为：、

$$\tilde{o} = att_{cross}(h_T, h_I, h_I) \quad (8)$$

其中ATT_{cross}表示跨模态多头关注（Ju等人，2020）。然后，将O与H_T，并发送一层归一化，得到最终的语境表征O。

多通道层

在本小节中，我们旨在挖掘w_i和w_j之间的不同依赖性特征，以帮助检测它们之间的关系。如图5所示：（a）我们认为语篇（Pos）可以为词对提供词汇信息。例如，大多数实体的Pos属于NOUN和PEROPN，例如**NBA**、**Curry**和**Thompson**；（b）对词对之间的句法距离（Sd）进行编码可以提高模型捕获长距离句法信息的能力；（c）词的共现矩阵（Co）可以提供词对之间的语料库级信息。例如，**库里**和**NBA**在语料库中出现了一些次。关于构建每个特征矩阵的细节被添加到了

在附录A中。

经过数据预处理后，三个特征矩阵被视为$M^I \in R^{n \times n}, I \in \{Pos, Sd, Co\}$.我们提出了一个W- GCN模块来模拟每个矩阵，获得每个通道表示。每个矩阵M^I，首先发送一个嵌入的层产生一个可训练的表示$R^I \in R^{n \times n \times d}$和$d_I$是代表的维度。计算W-

第i个字在第I个矩阵中的GCN过程显示为：、

$$s^I_i = \text{w-gcn}_I(r^I_i, o) = \text{Softmax}(\text{ReLU}(W_{I,R} + b_I)) - (W_I(O)) \quad (9)$$

其中，$R^I_i \in R^{n \times d}$ I是第I个语言矩阵中的第i个词和$W^r_I 1^{R1 \times d}$，和$W^r 2^{浦 \times d_T}$是共享权重用来进行线性层学习语言特征和表征能力。我们把表征结合起来并将它们发送到MLP（多层感知）层，以进行获得最终的文字表述、

$$S_i = \text{MLP}[S_{i\ 位置}; S_{i\ 证券}; S_{i\ 钻}] \quad (10)$$

其中$S_i \in R^{d\ T}$是第i个字的表示。因此，在多通道层的输出表示被表示为作为$S = [S_1, S_2 - -, S_n]$。最后，我们将恩S_i和S_j的高级表示法，以表示单词------。对(w_i, w_j)，即$r_{i,j} = [S_i; S_j]$。然后，将$r_{i,j}$发送给线性预测层并获得概率分布、

$$p_{i,j} = \text{Softmax}(W_{pi,j} r + b)_p \quad (11)$$

其中$W_p \in R^{dy \times 2dT}$和$b_p \in R^{y^d}$是可训练的参数和$d_y$是标签的数量。然后，我们用交叉熵误差来衡量地面真实分布和预测的标签分布、

$$L_{主要}(\vartheta) = -\sum^{S} \sum^{n} Y_s \text{ 对数 } (p^s$$

ϑ) (12) *i,j* *i,j*

$L_{graph} = \alpha - D_{wd}(H_I, H_T) + (1 - \alpha) -$

$d_{gwd}(h_{I,hi}, h_{T,ht})$ (7)

其中S和ϑ表示训练样本的数量，所有的 $_{s=1\ i=1\ j=1}$

分别是可训练的参数。

方法		JMERE			#MNER		
		#P	#R	F1	#P	#R	F1
管线方法	适应性强+MEGA	48.44	47.06	47.74	74.32	72.11	73.20
	OCSGA+MEGA	48.21	47.99	48.10	75.27	72.32	73.77
	AGBAN+MEGA	47.87	48.28	48.57	74.78	73.69	74.23
	UMGF+MEGA	49.28	50.76	50.01	75.02	76.77	75.88
词对关系标签方法	适应性强att*	50.22	47.67	48.91	77.32	73.28	75.25
	OCSGA*	52.11	47.41	49.64	77.13	75.03	76.07
	AGBAN*	51.07	48.89	49.95	76.57	75.82	76.19
	UMGF*	52.76	50.22	51.45	77.51	76.01	76.75
	巨大的*	55.08	51.40	53.18	77.78	76.67	77.22
	MAF*	52.56	**54.73**	53.62	76.07	77.57	76.81
	EEGA(*我们的*)	**58.**26†	52.61	**55.**29†	**78.27**	**78.**91†	**78.**59†

表1：JMERE任务的实验结果（%），#MNER表示由JMERE结果计算的MNER结果。AGBAN* 指使用AGBAN模型中的词对关系标记。标记†指的是显著的测试*P-值*<0.05。最佳结果用粗体表示，#P、#R和F1表示精度、召回率和F1分数。

方法	#P	#R	F1
EEGA(*全部*)	58.26	52.61	55.29
不含边缘增强型	51.85	50.31	51.07
不含属性变压器	55.69	51.07	53.28
不含多通道	55.48	52.13	53.75

表2：JMERE任务的消融研究结果。

加入培训

最终目标是主要任务和优化交叉图的结合，具体如下、

$$L = L_{main} + \lambda \cdot L_{graph} \quad (13)$$

其中λ是控制优化交叉图的贡献的交易性超参数。

实验

为了评估和比较EEGA方法与之前的七项工作的性能，我们进行了比较性实验。此外，更详细的实验（例如，数据集、设置和参数敏感性）在附录B-D中预发了。

比较结果

比较的方法。我们总结了MNER和MRE的研究，并结合最先进的方法作为我们强大的JMERE基线，如表1所示。它们包括AdapCoAtt（Zhang等人，2018）、OCSGA（Wu等人，2020b）、AGBAN（Zhang等人，2021a）和MAF（Xu等人，2022）用于实体和相应类型的提取，以及UMGF（Zheng等人，2021a）用于实体的关系提取。此外，我们将词对关系标记应用于上述基线模型，以考察词对关系标记和所提方法的有效性，如AdapCoAtt*，OCSGA*，MEGA*。

总体结果。观察管道方法，我们发现UMGF+MEGA的表现比其他管道方法更好，这表明将交叉图中的节点对齐可以有利于实体与对象的匹配。在JMERE和#MNER中，词对关系标记方法的表现优于管道方法，如OCSGA、AGBAN和UMGF，表明词对关系标记可以通过利用任务关系和减少管道框架引起的错误传播问题来提高per- formance。

此外，EEGA超过了所有的基线。与现有基线的最佳结果相比，EEGA在JMERE和#MNER上仍然取得了1.67%和1.37%的绝对F1分数增长。实验结果有力地证明，对交叉图中的节点和边进行多重对齐，可以有效地提高对象与实体的匹配精度，捕捉更多的实体之间的关系。此外，所提出的属性变换器通过将边缘信息纳入变换器中的键和值，增强了挖掘节点间关系的能力。同时，多通道层可以利用词对之间的语言关系来完善最终的表述，提高预测性能。

消减研究。为了研究EEGA中不同组件的有效性，边缘增强图最优传输（边缘增强）、属性变换器和多通道层，我们对表2中的JMERE任务进行了消减研究。*W/o属性变换器*是指用香草变换器代替属性变换器。F1分数下降了2.01%，表明将边缘信息整合到转化器的键和值中，可以提高捕捉节点间关系的能力，并有利于交叉图中的边缘对齐。去掉多通道层（*无多通道*）后，性能有所下降，说明多通道层可以从不同角度挖掘词对的关系，完善最终的表示。

*W/o edge-enhanced*是指从EEGA中去除边缘增强图的最优传输。去掉边缘增强后，模型的性能高度下降，说明同时对准交叉图中的节点和边缘，有利于更精确地匹配视觉对象和文本实体，从对象之间的关系中找到实体分类线索。

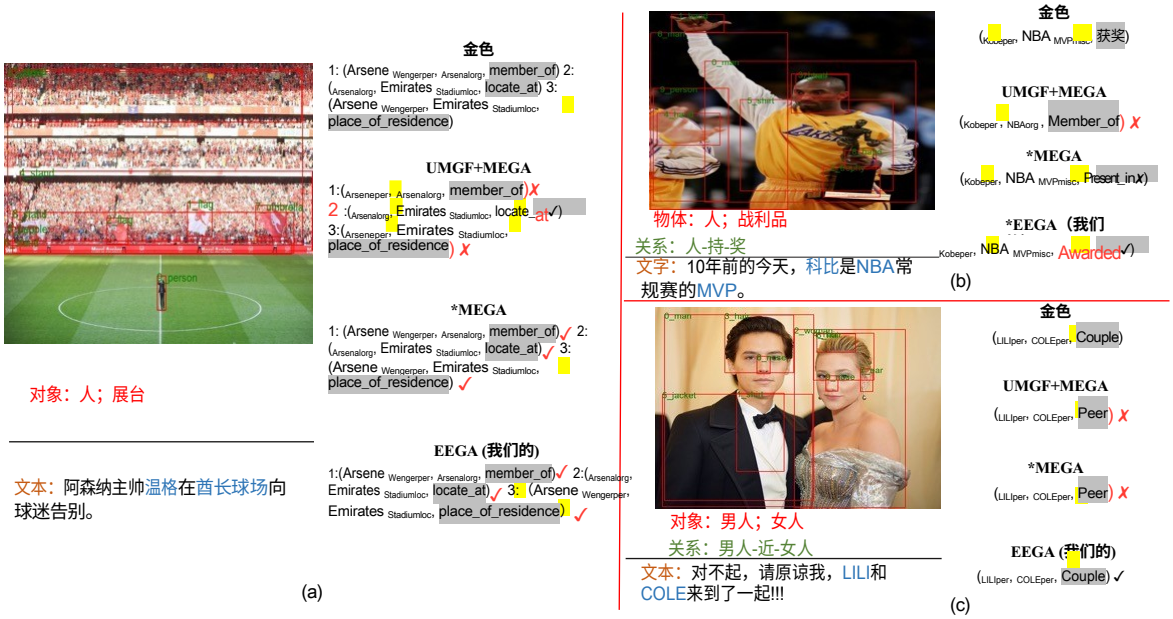
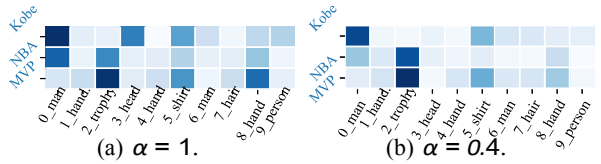


图6：通过UMGF+MEGA、*MEGA和EEGA进行预测的三种情况。



(a) α = 1. (b) α = 0.4.

图7：在公式（7）中，从α=1（只有节点对齐）和α=0.4（最佳性能设置）对实体-对象对的注意力可视化的比较。

案例研究

为了了解我们提出的模型的有效性，图6展示了三个例子的预测结果。同时，重要的物体和关系被从图像中检测出来。在例子(a)中，只有基于管道的模型UMGF+MEGA的提取是不正确的，因为管道模型很容易受到错误传播的影响，也就是说，UMGF的提取能力**Arsene**是不完整的，最终模型UMGF+MEGA提取的是不正确的五元组。在例子（b）中，UMGF+MEGA不精确地提取了**NBA**作为一个实体，而*MEGA错误地预测了**科比**和**NBA MVP**之间的关系作为*Present in*。在例子（c）中，情况与例子（b）类似。由于缺乏有效的方法将对象*Man-near- Woman*的语义关系映射到实体（*LILI-COLE*），UMGF+MEGA和
*MEGA错误地预测了实体之间的关系*同行*。

对于这三个例子，提议的EEGA做出了准确的判断。受益于边缘增强的图形优化传输模块，EEGA可以将交叉图中的节点和边缘对齐，以更精确地匹配实体和对象。同时，EEGA还有效地捕捉了例（b）和（c）中显示的从视觉图到文本图的关系线索。此外，属性转换器和多通道层可以进一步提高建立对象和词对关系模型的能力。

可视化分析

在这一节中，我们将图6中的例子（b）在α=1和α=0.4时可视化，以测试我们的边缘对齐策略是否有助于学习细粒度的实体-物体匹配。

如图7（a）所示，当只有节点对齐意味着α=1时，由于提出的模型缺乏边缘约束，注意力权重相对分散，影响了实体与对象的匹配精度。特别是，该模型很容易将实体**NBA**的类型分类为*Per*。平均而言，如图7(b)中的**NBA**和**Kobe**所示，从边缘对齐中获益，可以找到NBA和**Kobe**之间的映射关系。

EEGA有效地减少了模糊性，并将目标与实体更精确地匹配。

总结

在本文中，我们首次提出了一个联合多模态实体关系提取（JMERE）任务，以处理多模态的NER和RE任务。为了处理这个任务，我们提出了一个边缘增强的图对齐网络和一个词对关系标签。具体来说，我们设计了一个词对关系标签来避免管道框架引起的错误传播。然后，我们提出了一个边缘增强图对准网络（EEGA），通过同时对准交叉图中的节点和边缘来增强JMERE任务。EEGA可以利用边缘信息来辅助对象和实体之间的对齐，并找到实体-实体关系和对象-对象关系之间的相关性。详细的评估表明，我们提出的模型明显优于几个最先进的基线。我们将在未来的工作中把我们的方法扩展到多标签多模态任务，并研究其他方法（如自监督模型）以更好地建立JMERE模型。

鸣谢

这项工作得到了国家自然科学基金（62076100，61966038）、华南理工大学中央高校研究基金（x2rjD2220050）、广东省科技计划项目（2020B0101100002）、CAAI-华为MindSpore开放基金、香港再教育资助委员会（项目编号：PolyU 11204919和C1031-18G）的支持。理大11204919和C1031-18G）和香港理工大学的内部研究资助（项目1.9B0V）。

参考文献

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyre´、
G. 2015.迭代布雷格曼投影的正则化反转问题.*SIAM Journal on Scientific Computing*, 37(2)：1111-1138.

Bouma, G. 2009.Collocation Extraction中的归一化（Pointwise）Mutual Informa- tion.*2009年GSCL双年会论文集*, 31-40.

Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020。用于跨域对齐的图式最优传输。在*ICML 2020会议上*，1520-1531。

Chen, X.; Jia, S.; and Xiang, Y. 2020。一个回顾：知识图谱上的知识-边缘推理.*Expert Systems with Applications*, 141：112948-112966.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022a。用于多模态知识图谱梳理的多层次融合的混合转化器。In *Proceedings of the SIGIR '22*, 904-915.

Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022b。好的视觉指导使一个更好的提取器：用于多模态实体和关系提取的层次化视觉前缀。In *Findings of the NAACL 2022*, 1607-1618.

Cuturi, M. 2013.Sinkhorn Distances：最佳运输的光速计算。在*NIPS 2013会议上*，1-9。

He, K.; Gkioxari, G.; Dolla´r, P.; and Girshick, R. 2017.Mask r-cnn.In *Proceedings of the ICCV 2017*, 2961-2969.

Ju, X.; Zhang, D.; Li, J.; and Zhou, G. 2020。基于变换器的标签集生成，用于多模态多标签情绪检测。In *Proceedings of the ACM MM 2020*, 512-520.

Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; and Zhou, G. 2021。带有辅助性跨模态关系检测的多模态方面-情绪联合分析。在*EMNLP 2021会议上*，4395-4405。

Kim, J. H.; Jun, J.; and Zhang, B. T. 2018.双线性注意力网络。In *Proceedings of the NIPS 2018*, 1564-1574.

Kingma, D. P.; and Ba, J. L. 2015.Adam: A method for stochastic optimization.In *Proceedings of the ICLR 2015*, 1-15.

Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; and Li, J. 2020。
一个统一的命名实体识别的MRC框架。在*ACL 2020论文集*, 5849-5859.

Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; and Rosenblum, D. S. 2019.MMKG：多模式knowl-边缘图。In *Proceedings of the ESWC 2019*, 459-474.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018。用于多模态社交媒体中姓名标签的视觉注意力模型。In *Proceedings of the ACL 2018*, 1990-1999.

Moon, S.; Neves, L.; and Carvalho, V. 2018.针对社交媒体短文的多模态命名实体识别。In *Proceedings of the NAACL 2018*, 852-860.

Nasar, Z.; Jaffry, S. W.; and Malik, M. K. 2021.命名实体识别和关系提取.*ACM计算调查*, 54: 1-39.

Peyre´, G.; Cuturi, M.; and Solomon, J. 2016。核和距离矩阵的Gromov- wasserstein平均化。在*ICML 2016会议上*，2664-2672。

Peyre´, G.; Cuturi, M.; et al. 2019。计算的最佳运输：与数据科学的应用。*机器学习的基础和趋势*，11（5-6）：355-607.

Ren, H.; Cai, Y.; Chen, X.; Wang, G.; and Li, Q. 2020。一个用于增量少拍关系分类的两阶段原型网络模型。In *Proceedings of the COLING 2020*, 1618-1629.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020。从有偏见的训练中生成无偏见的场景图。在*CVPR 2020会议记录*中，3713-3722。

Vashishth, S.; Joshi, R.; and Suman, S. 2018。RESIDE: Im- proving distantly-supervised neural relation extraction using side information.In *Proceedings of EMNLP 2018*, 1257- 1266.

Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; and Chang, Y. 2020.一种用于关系型三重提取的新型级联二元标签框架。In *Proceedings of the ACL 2020*, 1476-1488.

Wen, Y.; Fan, C.; Chen, G.; Chen, X.; and Chen, M. 2020.命名实体识别的调查.*电气工程讲义*, 571: 1803-1810.

Wu, Z.; Ying, C.; Zhao, F.; Fan, Z.; Dai, X.; and Xia, R. 2020a.面向方面的细粒度意见提取的网格标签方案。In *Findings of the EMNLP 2020*, 2576-2585.

Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H. F.; and Li、Q.2020b.带有嵌入式视觉引导对象的多模态表示，用于社交媒体帖子中的命名实体识别。在*ACM MM 2020会议记录*中，1038-1046。

Xu, B.; Huang, S.; Sha, C.; and Wang, H. 2022.MAF：一个用于多模态命名实体识别的通用匹配和对齐框架。In *Proceedings of the WSDM 2022*, 1215-1223.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020.通过统一的多模态变换器的实体跨度检测提高多模态命名实体识别。In *Proceedings of the ACL 2020*, 3342-3352.

Yuan, L.; Wang, J.; Yu, L.-C.; and Zhang, X. 2020a.图注意网络与记忆融合用于方面级情感分析。在*AACL 2020会议上*，27-36。

Yuan, Y.; Zhou, X.; Pan, S.; Zhu, Q.; Song, Z.; and Guo, L. 2020b。用于联合实体和关系提取的特定关系注意网络。In *Proceedings of the IJCAI 2020*, 4054-4060.

Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; and Zhou, G. 2021a.多模式图融合的命名实体识别与目标视觉引导。在*AAAI 2021会议上*，14347-14355。

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018。用于推文中命名实体识别的自适应共同关注网络。In *Proceedings of the AAAI 2018*, 5674-5681.

Zhang, Y.; Wang, J.; Yu, L.-C.; and Zhang, X. 2021b.MA-BERT: 通过纳入变压器中的多属性知识来学习表示。In *Findings of the ACL 2021*, 2338-2343.

Zhao, T.; Yan, Z.; Cao, Y.; and Li, Z. 2021.提出有效和多样化的问题：一个基于机器阅读理解的联合实体关系提取框架。In *Pro- ceedings of the IJCAI 2021*, 3948-3954.

Zheng, C.; Feng, J.; Fu, Z.; Cai, Y.; Li, Q.; and Wang, T. 2021a。多模态关系提取与高效图谱对齐。在*ACM MM 21会议上*，5298-5306。

Zheng, C.; Wu, Z.; Feng, J.; Fu, Z.; and Yi, C. 2021b。MNRE：
一个用于社交媒体帖子中视觉证据的神经关系提取的多模态数据集的挑战。在*ICME 2021的会议上*，3-8。

Zheng, C.; Wu, Z.; Wang, T.; Cai, Y.; and Li, Q. 2021c.利用对抗性学习在社交媒体帖子中进行对象感知的多模态命名实体识别。*IEEE Transactions on Multimedia*, 23: 2520-2532.

附录

附录A：关于多通道层的细节 在这一节中，我们介绍了三个特征矩阵的构建过程：语音部分（Pos），句法距离（Sd）、
和词的共现矩阵（Co）。

语篇（Pos）。为了获得词对（w_i, w_j）之间的词汇级信息，我们用spaCy来获得
输入句子的Pos序列，如图8所示。
然后，用一个嵌入层将Pos序列嵌入到可训练的向量$R^{Pos} \in R^{n\times d}$，。我们将相应元素的*第i至j个*Pos向量作为最终的
词法层面的表示 $R^{Pos}_{ij} \in R^{d}$。

句法距离（Sd）。为了建立句法层面的信息模型
我们使用句法上的相对距离来表示词对。
$M^{Sd}_{i,j} \in R^{1}$，其定义为路径上的跳数
从标记x_i到标记x_j的依赖树。如图所示
在图8中，**汤普森**和阿姆斯特朗之间的绝对距离是多少？

Trophy是7，而句法上的相对距离是2。

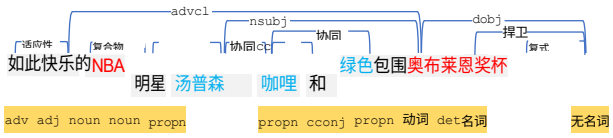


图8：给定例子的依赖性解析树。

统计数据	JMERE			
	#S	实体	词条	#AL
火车	3,618	9,006	15,981	16.31
拓展	496	1,248	3,535	16.57
测试	475	1,280	4,678	16.28

表3：JMERE数据集的统计数据。这里#S和Entity分别表示句子和实体的数量。#AL是指句子的平均长度。

共现矩阵（Co）。为了将语料库层面的信息纳入词对，如**库里**和**NBA**在语料库中多次出现，我们使用词频共现矩阵与点式互信息（PMI）（Bouma 2009）来计算词对的相关性。因此，*第i个令牌与第j个令牌*的PMI的相关性表示为$PMI_{i,j}$。为了将$PMI_{i,j}$编码为可训练的向量，我们将$PMI_{i,j}$四舍五入，得到最终的相关值$M^{Co} \in R^{1}$。尤其是
对于负相关的值，我们统一将其设置为-1。

附录B：数据集

为了验证所提出的模型的有效性，我们通过MNER（Zheng等人，2021a）和MRE数据集（Zheng等人，2021c）对联合多模态实体-关系提取数据集（JMERE）进行col- lect。此外，我们合并了同一句子的样本，但对实体和关系有不同的注释。每个样本包含原始句子、相应的图像和五元组。如果语料库包含许多未标记的数据，模型将学习一些无意义的信息。因此，我们消除了没有实体类型或实体之间关系的原始数据集样本。表3显示了我们整理的JMERE的静态。

为了评价不同方法的性能，我们使用精度、召回率和F1分数作为评价指标。只有当预测的五元组与地面实况的跨度精确匹配时，提取的五元组才被认为是正确的。

附录C：实施细节

我们将 spaCy[3] 与 *en core web trf* 版本应用于
将给定的句子解析为依赖树，然后从依赖树中建立文本图和相邻矩阵。我们通过基于`BERT-Uncased`的[4]，并设置维度D_T，初始化文本表示法。
768.此外，视觉对象提取的维度$D_$
通过场景图生成模型（Tang等人，2020年）（。
用作骨干的Mask-RCNN）是4096。此外，其他特征的维度，例如，边缘特征d_{zT}
和d_z，和语言学特征d_l，被初始化为100。语料库

Sinkhorn算法的迭代数为20，令牌序列和对象的最大数量分别为70和10。Adam（Kingma和Ba 2015）优化器，学习率为2e-5，衰减因子为0.5。早期停止策略也被应用于确定耐心为5的epochs的数量。我们实现了我们的模型

[3]https:github.com/explosion/spaCy
[4]https:huggingface.co/bert-base-uncased

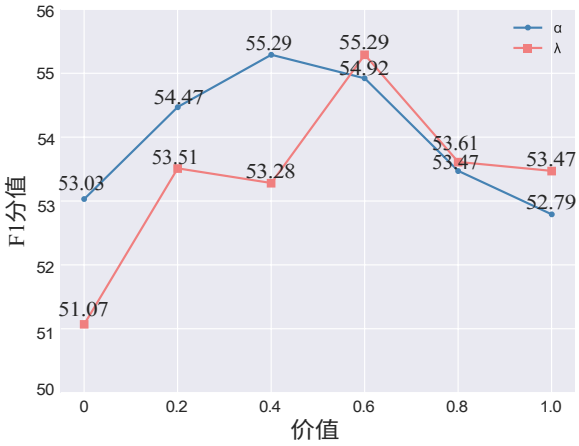


图9：不同参数对平衡系数的影响，其中公式（7）的α和公式（13）的λ。

用PyTorch框架，在装有NVIDIA RTX 3090的机器上进行实验。

附录D：参数敏感度

在本节中，我们进一步讨论参数的不同设置。我们关注的是公式（7）中的两个平衡系数α和公式（13）中的λ的影响。λ=0意味着*在没有边缘增强的情况*下，模型不能有效地将物体与实体相匹配，取得最差的性能。当λ=0.6时，所提出的EEGA取得了最好的性能；当λ超过0.6时，交叉图的对齐已经干扰了主要任务的训练过程，导致性能略低。此外，α=0和α=1意味着在交叉图谱对齐中只有对齐的节点和边，取得了较低的性能，尤其是α=1时。这表明，边对齐可以有效地提高实体与对象的匹配精度，以及捕捉对象之间潜在语义关系的能力。当α=0.4时，EEGA的结果要比其他设置好。