arXiv:2304.02328v1 [cs.MM] 5 Apr 2023

利用变分信息瓶颈增强多模态实体与关系抽取

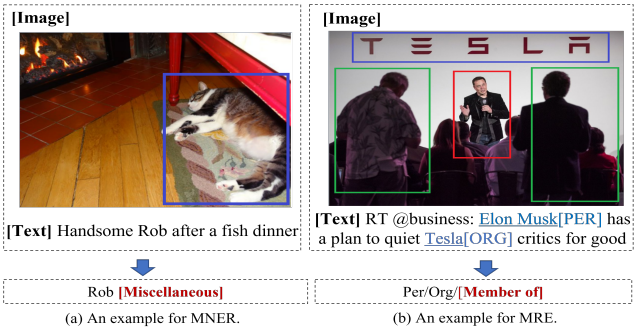
Shiyao Cui、Jiangxia Cao、Xin Cong、Jiawei Sheng、Quangang Li、Tingwen Liu、Jinqiao Shi

摘要——本文研究多模态命名实体识别（MNER）和多模态关系抽取（MRE），这两项任务对于多媒体社交平台分析至关重要。MNER 和 MRE 的核心在于融入显著的视觉信息以增强文本语义，其中有两个问题亟待研究。第一个问题是模态噪声，即每种模态中的任务无关信息可能成为误导任务预测的噪声。第二个问题是模态差距，即不同模态的表征并不一致，从而妨碍文本与图像之间建立语义对齐。为解决这些问题，我们提出一种用于 MNER 和 MRE 的新方法——基于信息瓶颈（MMIB）的多模态表示学习。针对第一个问题，我们采用一种精炼正则器，利用信息瓶颈原理平衡预测证据与噪声信息，从而生成具有表达力的预测表征。针对第二个问题，我们提出一种对齐正则器，其中基于互信息的项以对比方式发挥作用，以约束文本-图像表征的一致性。据我们所知，我们是首个探索用于 MNER 和 MRE 的变分信息瓶颈估计的研究。实验表明，MMIB 在三个公开基准上取得了最先进的性能。

索引词—**多模态命名实体识别，多模态关系抽取，信息瓶颈**

I. 引言

命名实体识别（NER） [1] 和关系抽取（RE） [2] 是社交媒体分析中的两项主要任务，惠及各种应用 [3], [4]。具体而言，NER旨在从非结构化文本中提取感兴趣的实体，而RE则寻求分别确定实体之间的语义关系。与新闻领域中的NER和RE相比，社交媒体中的文本通常较为简短，并伴有俚语 [5], [6] ，，其中不充分且含糊的语义增加了内容分析的难度。幸运的是，随着社交媒体中多模态帖子的普及，近年来，社交帖子中的图像被用于促进面向社交媒体的命名实体识别和关系抽取，即**多模态命名实体识别（多模态命名实体识别）** [7] 和**多模态关系抽取（多模态关系抽取）** [8]。



(a) An example for MNER. (b) An example for MRE.

图 1. Twitter 数据集中用于多模态命名实体识别和多模态关系抽取的示例。

多模态命名实体识别和多模态关系抽取的核心在于利用图像中显著的视觉信息，增强社交媒体文本的语义。为实现这一目标，研究者以不同方式融合视觉信息。早期研究者直接将整幅图像作为全局视觉线索进行探索。这些方法要么将原始图像编码为一个特征向量 [9] ，以增强语义；要么粗略地将图像划分为多个网格，以进行跨模态交互 [1],[10]。为了获取更细粒度的视觉线索，后续研究者提取图像中的显著对象，并将其编码为局部视觉提示 [5], [6], [8], [11], [12], [13], [14], [15] ，以指导任务预测。尽管这些方法取得了显著成效，仍有两个问题尚未得到充分解决。

第一个问题是**模态噪声**，其中每种模态都包含任务无关信息，这些信息不仅无法对最终任务作出贡献，甚至还可能成为误导预测的噪声。以图 1(a) 中的多模态命名实体识别为例。对于文本模态，由于“handsome”和“dinner”等词语产生的文本噪声将人物特征赋予了“Rob”，模型倾向于将“Rob”预测为人物类型（PER）的实体。幸运的是，从文本和图像的整体概览来看，图像中的“Cat”对象（例如蓝色框中的区域）显然能够引导出正确的预测。然而，对于图像模态而言，噪声可能体现在两个方面：1）在全局层面，图像中的大多数区域都无法为目标实体的识别提供有效信息；2）在局部层面，相应的显著区域所表达的视觉语义也比我们的实际需求（恰好只是一个“Cat”）更加复杂（“一只猫躺在地毯上”）。在这里，冗余信息可能会干扰模型为图像中的各个区域分配注意力权重，从而阻碍最终的任务预测。

稿件创建于2023年4月；本工作由IEEE出版技术部开发。本工作根据LaTeX项目公共许可证（LPPL）（http://www.latex-project.org/）1.3版分发。LPPL 1.3版的副本包含在2003/12/01或之后发布的所有LaTeX发行版的基础LaTeX文档中。此处表达的观点完全属于作者。不提供任何明示或暗示的保证。用户承担全部风险。Shiyao Cui、Jiangxia Cao、Xin Cong、Jiawei Sheng、Quangang Li和Tingwen Liu就职于中国科学院信息工程研究所（北京100193，中国）和中国科学院大学（北京，中国）（电子邮件：{cuishiyao,caojiangxia,congxin,shengjiawei,liquangang,liutingwen}@iie.ac.cn）

Jinqiao Shi 任职于北京邮电大学，中国北京 100876。（电子邮箱：shijinqiao@bupt.edu.cn）本文已提交 IEEE，供其考虑发表。版权可能在不另行通知的情况下转让，届时此版本可能无法再访问。

类似地，在图1(b)中，绿色框中的冗余噪声区域可能会驱使模型预测人物之间的关系，而不是组织与人物之间的关系。因此，精炼信息以减少冗余噪声并保持对任务的预测能力至关重要。

尽管冗余噪声已被消除，**模态差距**仍然存在，即输入文本和图像的表征并不一致。具体而言，由于文本和视觉表征分别由不同的编码器获得，它们具有不同的特征空间和分布。这种模态差距使文本和图像难以相互“理解”并捕捉跨模态语义关联。对于图 1(a) 中的多模态命名实体识别，理想情况下，文本实体“Rob”与蓝色框中的“Cat”应比与其他区域具有更强的语义相关性。然而，文本与视觉表征之间的差异妨碍了这种对齐的建立，阻碍了对预测性视觉线索的探索。图 1(b) 中的多模态关系抽取也是如此，语义表达上的不一致使得难以将标注的文本实体与视觉人物/组织对象对齐以进行关系判定。尽管先前的研究 [6], [16] 将视觉对象转换为相应的文本标签，以实现语义表达的一致性，但仍容易受到外部现成工具生成的文本标签的错误影响。

本文提出一种基于变分信息瓶颈的多模态表示学习方法，用于多模态命名实体识别和多模态关系抽取，称为**MMIB**。我们的方法基于变分自编码器 [17]框架对数据分布进行正则化，从而解决上述问题。具体而言，针对**模态噪声**，我们设计了精炼正则器（*RR*），利用信息瓶颈（IB）原理进行文本-图像表示学习。IB 原理旨在在获得简洁且保留自身信息的表征与具有普遍预测能力的表征之间进行权衡 [18]。在我们的任务中，IB 对文本/图像表示进行精炼，去除其中的噪声信息，同时保留对最终任务的预测能力。我们设计的 RR 包含两个 IB 项，分别作用于输入文本和图像的表示学习，从而为每种模态的预测生成稳健的表征。针对**模态差距**，我们提出了对齐正则器（*AR*）。由于互信息（MI）能够衡量两个变量之间的关联，AR 通过最大化配对文本-图像表征之间的互信息，同时最小化非配对文本-图像表征之间的互信息，促使文本-图像表征保持一致。基于上述正则化项，我们能够生成具有表达力的文本-图像表征，以促进最终任务预测。

总体而言，我们将贡献总结如下：

- 我们通过遵循变分信息瓶颈原理的表示学习，探索解决多模态命名实体识别和多模态关系抽取的新视角。
- 我们提出了一种新颖的方法 MMIB，其分别针对模态噪声和模态差距采用精炼正则器和对齐正则器。
- 实验结果表明，MMIB 在三个多模态命名实体识别和关系抽取公共基准上取得了新的最先进性能。

标记，以及广泛的分析验证了我们所提出方法的有效性。

II. 预备知识

本节从三个方面回顾所研究任务和所提出方法的相关基础知识。

A. 任务定义

1) 多模态命名实体识别：给定一个句子及其关联图像，多模态命名实体识别旨在识别文本中的命名实体，并将识别出的实体划分到预定义的实体类型中。我们将该任务形式化为序列标注范式。具体而言，令 $(x_1^t, x_2^t, ..., x_n^t)$ 表示包含 n 个词元的输入句子，我们旨在预测相应的标签序列 $Y = (y_1, y_2, ..., y_n)$，其中 y_i 是遵循 BIO标注体系 [10], [19] 的预定义标签，由此可提取所关注的实体。

*2)*多模态关系抽取：给定一个句子及其关联图像，多模态关系抽取的目标是检测两个标注实体之间的语义关系。形式化地，令 $(x_1^t, x_2^t, ..., x_n^t)$ 表示包含两个实体 $E_1 = (x_i^t, ..., x_{i+|E_1|-1}^t)$ 和 $E_2 = (x_j^t, ..., x_{j+|E_2|-1}^t)$ 的句子，该任务被表述为一个分类问题，用于从预定义的关系类型中确定 Y 在E_1 和 E_2 之间的关系类型。

B. 互信息

继续之前，我们首先介绍互信息（MI），因为它是信息瓶颈的一个基本概念。MI 是信息论中的一个通用度量， [20],用于衡量随机量之间关联的强度，其定义如下：

$$I(\boldsymbol{X};\boldsymbol{Y}) = \sum_{y\in\boldsymbol{Y}}\sum_{x\in\boldsymbol{X}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{1}$$

其中， $I(;)$表示两个随机变量$\boldsymbol{X}$ 和 $\boldsymbol{Y}$之间的互信息， $p(\cdot)/p(\cdot,\cdot)$分别表示样本 x 和 y的边缘/联合**概率**。同时，互信息也可以通过两个分布之间的库尔贝克-莱布勒散度 [21], [22] 来衡量，具体如下：

$$I(\boldsymbol{X};\boldsymbol{Y}) = \mathbb{D}_{KL}(\boldsymbol{p}(\boldsymbol{X}|\boldsymbol{Y})||\boldsymbol{p}(\boldsymbol{X})), \tag{2}$$

其中， $\boldsymbol{p}(\cdot)/\boldsymbol{p}(\cdot|\cdot)$表示变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$的先验/后验**分布**。

C. 信息瓶颈原理

信息瓶颈（IB），由 Tishby 等人提出 [23]，旨在在拥有显式表示和具有普遍预测能力的表示之间进行权衡，以获得有效的潜在特征 [18]。假设我们拥有任务输入信息 $\boldsymbol{X}$ 和任务标签信息 $\boldsymbol{R}$，IB 表示为：

$$\mathcal{L}_{IB} = \beta I(\boldsymbol{Z};\boldsymbol{X}) - I(\boldsymbol{Z};\boldsymbol{R}), \tag{3}$$

特别是，公式 3 的目标是推导出一个有限的潜变量 $\boldsymbol{Z}$，用于过滤由

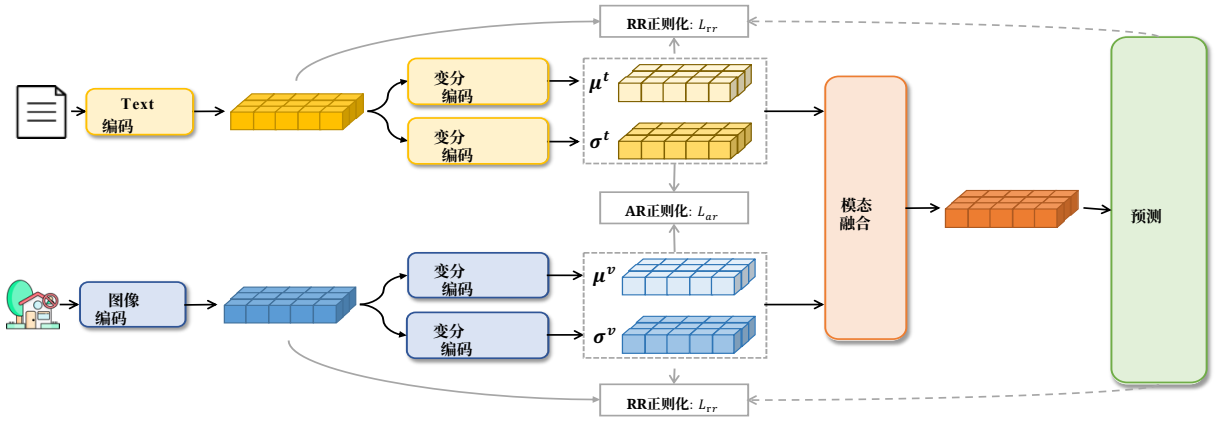


图 2. 模型架构。

输入 $\boldsymbol{X}$ 提供的冗余信息，同时最大限度地保留与重构预期信息 $\boldsymbol{R}$相关的信息。目标函数可解释如下：1）**最小化** $I(\boldsymbol{Z};\boldsymbol{X})$，使 $\boldsymbol{Z}$ 能够丢弃与任务预测无关的部分；2）**最大化**$I(\boldsymbol{Z};\boldsymbol{R})$，以强制 $\boldsymbol{Z}$ 保留预测信息。 $\beta > 0$ 是一个拉格朗日乘子，用于平衡这两个约束之间的权衡。在实现中，由于直接优化基于互信息的项是不可行的，因此通常采用变分近似 [24] 来优化相应的目标函数 [22]。

III. 方法

如图2所示，本节介绍MMIB。具体而言，MMIB包含三个模块：1）编码模块将文本和图像转换为基本的实值嵌入。2）表示学习模块利用我们提出的正则化项进行变分编码。3）预测模块执行模态融合，并进行最终的任务预测。

A. 编码模块

此模块对输入句子和图像进行编码，将它们转换为分布式表征。

*1)*文本编码： 为了得到输入句子的表征，我们采用广泛使用的BERT [25] 作为文本编码器。遵循 Devlin等人 [25], 的做法，在每个句子中插入两个特殊标记，其中 [CLS] 和[SEP] 分别添加在句子的开头和结尾。形式化地，令 $X^T = (x_0^t,\ x_1^t,\ x_2^t,\ .....x_n^t,\ x_{n+1}^t)$表示每个经过处理的句子，其中 x_0^t 和 x_{n+1}^t 分别表示插入的 [CLS] 和 [SEP]。我们将X^T 输入 BERT，并得到输出表征$\boldsymbol{X}^T = (\boldsymbol{x}_0^t,\ \boldsymbol{x}_1^t,\ \boldsymbol{x}_2^t,\ .....\boldsymbol{x}_n^t,\ \boldsymbol{x}_{n+1}^t)$，其中 $\boldsymbol{x}_i^t \in \mathbb{R}^d$是 i_{th} 标记的上下文化表征。

*2)*图像编码： 给定图像 X^V，它包含两个方面的视觉信息：(i)整幅图像提供的全局层面的抽象概念；(ii) 视觉对象提供的局部层面的语义单元。我们旨在探索这两方面的视觉信息，因此需要两个步骤参与生成图像表示。**步骤1。对象提取。** FollowingZhang 等人 [11], 我们采用视觉定位工具包 [26]来提取局部对象，这些对象表示为 $(x_1^v,\ x_2^v,\ .....x_m^v)$。**步骤2。图像表示。**鉴于ResNet [27] 在计算机视觉领域取得的巨大成功，我们将其用作视觉编码器。具体而言，我们首先将整幅图像及其对象图像重新缩放为$224 * 224$ 像素，然后将它们输入 ResNet。由于ResNet 生成维度为 2048 的视觉表示，因此线性变换矩阵 $\boldsymbol{W}_v \in \mathbb{R}^{2048\times d}$ 用于将图像表示投影到与文本表示相同的维度。最后，我们将图像表示连接起来，得到 $\boldsymbol{X}^V = (\boldsymbol{x}_0^v,\ \boldsymbol{x}_1^v,\ .....\boldsymbol{x}_m^v)$，其中$\boldsymbol{x}_0^v \in \mathbb{R}^{1\times d}$ 和 $\boldsymbol{x}_i^v \in \mathbb{R}^{1\times d}(i > 0)$分别表示整幅图像和 i_{th} 局部对象图像的表示。

B. 表示学习模块

该模块旨在通过我们提出的正则化项生成抗噪且一致的文本-图像表征。由于这些基于互信息的正则化项难以处理，因此采用变分编码进行表示学习。我们将首先介绍变分编码的工作原理，然后详细说明这些正则化项。

1) 变分编码： 我们以变分方式设计了一个编码器，用于学习具有表达力的文本/图像表示。由于潜在文本/图像表示在每个模态内进行压缩，编码器应探索模态内的信息传播。鉴于 Transformer [28] 对传播建模的充分能力，我们按照 Transformer 层的设计来构建编码器。具体而言，**我们将该编码器命名为"注意力传播编码器"**，记为 APEnc($\boldsymbol{Q}$, $\boldsymbol{K}$)，以查询 $\boldsymbol{Q}$ 和键值对 $\boldsymbol{K}$ 作为输入，其中首先对两个变量应用多头注意力机制，如下所示：

$$\begin{aligned}
&\mathrm{MultiHeadAtt}(\boldsymbol{Q},\boldsymbol{K}) = \boldsymbol{W}^{'}[\mathrm{CA}_1(\boldsymbol{Q},\boldsymbol{K}),...,\mathrm{CA}_h(\boldsymbol{Q},\boldsymbol{K})]^{\mathrm{T}},\\
&\mathrm{CA}_j(\boldsymbol{Q},\boldsymbol{K}) = \mathrm{Softmax}\Big(\frac{[\boldsymbol{Q}\boldsymbol{W}_{qj}][\boldsymbol{K}\boldsymbol{W}_{kj}]^{\mathbf{T}}}{\sqrt{d/h}}\Big)[\boldsymbol{K}\boldsymbol{W}_{vj}],
\end{aligned} \tag{4}$$

其中 CAj 表示这种多头注意力的 j_{th} 头，h 是头的数量。$\{\boldsymbol{W}_{qj},\boldsymbol{W}_{kj},\boldsymbol{W}_{vj}\} \in \mathbb{R}^{d\times(d/h)}$ 和

$\mathbf{W}' \in \mathbb{R}^{d\times d}$ 均为投影参数矩阵。随后，进一步堆叠一个全连接前馈网络和一个带层归一化的残差层，如下所示：

$$\begin{aligned}\widetilde{\boldsymbol{F}} &= \mathrm{LayerNorm}(\boldsymbol{Q} + \mathrm{MultiHeadAtt}(\boldsymbol{Q},\boldsymbol{K}))\\ \boldsymbol{F} &= \mathrm{LayerNorm}(\widetilde{\boldsymbol{F}} + \mathrm{FeedForward}(\widetilde{\boldsymbol{F}})),\end{aligned} \tag{5}$$

其中， $\boldsymbol{F}$ 是编码器 APEnc$(\boldsymbol{Q},\boldsymbol{K})$ 的最终输出表示，包含从公式 4 到公式 5 的计算。

由于基于互信息的正则化项难以处理，编码器以变分方式工作。为简单起见，**我们以图像表示的变分编码过程 APEnc$^V(\boldsymbol{X}^V,\ \boldsymbol{X}^V)$ 为例**。为促进图像之间的信息传播，每个图像 $\boldsymbol{x}_i^v \in \mathbb{R}^{1\times d}$ 分别作为查询，与所有图像表示 $\boldsymbol{X}^V \in \mathbb{R}^{(m+1)\times d}$ 作为键值对进行交互。具体而言，每个图像的潜在高斯分布变量 $\boldsymbol{z}_i^v$生成如下：

$$\boldsymbol{\mu}_i^v = \mathrm{APEnc}_\mu^V(\boldsymbol{x}_i^v, \boldsymbol{X}^V), \tag{6}$$

$$(\boldsymbol{\sigma}_i^t)^2 = \exp(\mathrm{APEnc}_\sigma^V(\boldsymbol{x}_i^v, \boldsymbol{X}^V)) \tag{7}$$

$$\boldsymbol{z}_i^v \sim \mathcal{N}\Big(\boldsymbol{\mu}_i^v, (\boldsymbol{\sigma}_i^v)^2\Big), \tag{8}$$

其中 $\boldsymbol{\mu}_i^v \in \mathbb{R}^{1\times d}$ 和 $\boldsymbol{\sigma}_i^v \in \mathbb{R}^{1\times d}$ 分别是高斯分布 $\mathcal{N}(\boldsymbol{\mu}_i^v, (\boldsymbol{\sigma}_i^v)^2)$ 的均值和方差向量，并且从该高斯分布中采样得到 i_{th} 图像的表示 $\boldsymbol{z}_i^v$。然而，这种直接采样过程对于反向传播来说难以处理，因此我们采用广泛使用的重参数化技巧 [17] 作为解决方案。在数学上，我们首先从标准高斯分布中采样 $\boldsymbol{\epsilon}$，然后执行等价采样以推导出 $\boldsymbol{z}_i^v$，如下所示：

$$\boldsymbol{z}_i^v = \boldsymbol{\mu}_i^v + \boldsymbol{\sigma}_i^v \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{diag}(\boldsymbol{I})) \tag{9}$$

其中， $\odot$ 表示逐元素乘积运算。类似地，我们可以获得所有图像的表征，并相应地获取视觉表示 $\boldsymbol{Z}^V \in \mathbb{R}^{(m+1)\times d}$，如下：

$$\boldsymbol{Z}^V = (\boldsymbol{z}_0^v, \boldsymbol{z}_1^v, ..., \boldsymbol{z}_m^v). \tag{10}$$

上述过程也适用于通过$\mathrm{APEnc}^T(\boldsymbol{X}^T,\boldsymbol{X}^T)$ **进行文本编码的情况，我们可以得到**：

$$\boldsymbol{Z}^T = (\boldsymbol{z}_0^t, \boldsymbol{z}_1^t, ..., \boldsymbol{z}_{n+1}^t). \tag{11}$$

2）精炼正则器（*RR*）：对于模态噪声，文本/图像的表征应**尽可能从任务无关信息中减少，同时最大限度地保持对最终任务的预测能力**。考虑到信息瓶颈（IB）可以平衡两个变量中的信息，我们旨在使用基于信息瓶颈的正则化项来实现这一目标。具体而言，我们提出了一种精炼正则器 L_{rr}，它由两个基于信息瓶颈的项组成，分别作用于文本 $\boldsymbol{Z}^T$ 和图像$\boldsymbol{Z}^V$ 的表示学习，如下所示：

$$\begin{aligned}L_{rr} =&\beta_1 I(\boldsymbol{Z}^T;\boldsymbol{X}^T) - I(\boldsymbol{Z}^T;\boldsymbol{R})\\ &+ \beta_2 I(\boldsymbol{Z}^V;\boldsymbol{X}^V) - I(\boldsymbol{Z}^V;\boldsymbol{R}),\end{aligned} \tag{12}$$

其中， $\boldsymbol{X}^T$、 $\boldsymbol{X}^V$ 分别表示来自输入文本和图像的原始信息，$\boldsymbol{R}$ 指的是理想的任务期望

用于预测的信息。考虑到 $\boldsymbol{Z}^T$ 和 $\boldsymbol{Z}^V$ [22] 之间的独立性，互信息链式法则可以重写为：

$$\begin{aligned}I(\boldsymbol{Z}^T;\boldsymbol{R}) + I(\boldsymbol{Z}^V;\boldsymbol{R}) =& I(\boldsymbol{Z}^T;\boldsymbol{R}|\boldsymbol{Z}^V) + I(\boldsymbol{Z}^V;\boldsymbol{R})\\ =& I(\boldsymbol{Z}^T,\boldsymbol{Z}^V;\boldsymbol{R}).\end{aligned} \tag{13}$$

因此， L_{rr} 在公式12中可简化如下：

$$L_{rr} = \underbrace{\beta_1 I(\boldsymbol{Z}^T;\boldsymbol{X}^T) + \beta_2 I(\boldsymbol{Z}^V;\boldsymbol{X}^V)}_{\text{Minimality}} - \underbrace{I(\boldsymbol{Z}^T,\boldsymbol{Z}^V;\boldsymbol{R})}_{\text{Reconstruction}}, \tag{14}$$

其中， L_{rr} 可以通过以下方式进行优化：(i)最小化前两个项，使文本/图像表示**从其原始模态中压缩而来，同时排除噪声**；(ii) 最大化重建项，以**促使预测表征趋向于任务预测**。更具体地说， L_{rr} 细化文本/图像表示，以保留显著信息，同时限制干扰噪声。**关于 L_{rr} 的可处理目标函数，请参阅第III-D1节**。

3) 对齐正则器（*AR*）：针对模态差距问题，文本/图像的表征应彼此一致。我们通过**增强 $\boldsymbol{Z}^T$ 与 $\boldsymbol{Z}^V$ 之间针对每个输入文本-图像对的互信息（MI）**来实现这一目标。这样一来，文本-图像表征被鼓励最大限度地彼此一致，因为互信息本质上可以度量两个变量之间的关联。具体而言，我们设计了一个基于互信息的对齐正则器（AR） L_{ar}，如下所示：

$$L_{ar} = -\underbrace{I(\boldsymbol{Z}^T;\boldsymbol{Z}^V)}_{\text{Maximality}}, \tag{15}$$

其中， L_{ar}通过最大化 $I(\boldsymbol{Z}^T;\ \boldsymbol{Z}^V)$进行优化，以使成对文本/图像的表征达成一致。**关于公式15的可处理目标函数，请参阅第III-D2节**。

C. 预测模块

在该模块中，我们首先进行模态融合，以纳入来自两种模态的信息，然后分别执行特定的任务预测。

1) 模态融合：为了得到用于最终任务预测的图像感知文本表征，我们探究模态之间的信息交互。**此处再次利用上述注意力传播编码器（APEnc），但关键区别在于查询和键-值对取自不同的模态**。为了利用模态之间的双向交互，APEnc以耦合方式工作，其中两个编码器分别以图像和文本作为查询。具体而言，对于 $\boldsymbol{Z}^V$ 和$\boldsymbol{Z}^T$，第一个 $\mathrm{APEnc}^{\mathrm{V2T}}$ 聚合来自文本的信息，以得到文本关注图像表示 $\boldsymbol{B} \in \mathbb{R}^{(m+1)\times d}$，如下所示：

$$\boldsymbol{B} = \mathrm{APEnc}^{\mathrm{V2T}}(\boldsymbol{Z}^V, \boldsymbol{Z}^T). \tag{16}$$

然后，为了利用视觉信息增强文本表示，第二个 $\mathrm{APEnc}^{\mathrm{T2V}}$ 用于推导图像感知的文本表示，如下所示：

$$\boldsymbol{C} = \mathrm{APEnc}^{\mathrm{T2V}}(\boldsymbol{Z}^T, \boldsymbol{B}), \tag{17}$$

其中，$\boldsymbol{C}=(\boldsymbol{c}_0,\ \boldsymbol{c}_1,\ .....\boldsymbol{c}_n,\ \boldsymbol{c}_{n+1})$ 是用于最终任务预测的图像感知文本表示。

由于我们将多模态命名实体识别建模为文本级序列标注问题，因此采用标准的条件随机场（CRF）推导出相应的标签序列$Y=(y_0,\ y_1,\ y_2,\ .....y_n,\ y_{n+1})$，以识别感兴趣的实体。给定输入句子的最终隐藏表示后，我们将其输入 CRF，得到标签序列Y 的概率：

$$p(Y|X^T)=\frac{\exp(\sum_{i=0}^{n+1}(\boldsymbol{w}_{y_i}^{crf}\boldsymbol{c}_i+\mathrm{Trans}(y_{i-1},y_i)))}{\sum_{Y'\in\boldsymbol{Y}}\exp(\sum_{i=0}^{n+1}(\boldsymbol{w}_{y_i}^{crf}\boldsymbol{c}_i'+\mathrm{Trans}(y_{i-1}',y_i')))}, \tag{18}$$

其中 $\boldsymbol{w}_{y_i}^{crf}$是从令牌 $\boldsymbol{c}_i$ 计算到标签 y_i的发射分数的参数向量，Trans(y_{i-1},y_i) 是从 y_{i-1} 到 y_i的可学习转移函数，$\boldsymbol{Y}$是所有可能的标签序列集合。

2）多模态关系抽取预测：多模态关系抽取被表述为一个分类问题。我们通过对 $\boldsymbol{C}$中包含的词元进行池化，得到每个标注实体的实体表示，并进一步将给定实体 E_1 和 E_2（详见§ II-A2）的表征分别记为 $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$。假设 Gold 标签 Y 是所有关系类型中的 k_{th} 关系，其对应的概率为：

$$p(Y|X^T,E_1,E_2)=\mathrm{Softmax}(\boldsymbol{W}([\boldsymbol{E}_1\oplus\boldsymbol{E}_2])+\boldsymbol{b})[k], \tag{19}$$

其中，$\oplus$ 是拼接结果，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是可学习参数。

D. 训练目标

MMIB的训练目标由三个组成部分构成：1) 精炼正则器的目标函数；2)对齐正则器的目标函数；3) 任务预测的目标函数。下文将分别对它们进行详细说明。

1) 精炼正则器目标函数：对 L_{rr} 的优化（式14）可通过最小化基于互信息的两项并最大化重建项来实现，如下所示。

将最小化应用于基于互信息的最小化项。考 [29],虑到精确计算互信息的挑战，我们通过 Kullback-Leibler 得出的**变分上界**来估计这两项。由于这两项 $I(\boldsymbol{Z}^T;\ \boldsymbol{X}^T)$ $I(\boldsymbol{Z}^V;\ \boldsymbol{X}^V)$ 形式相似，我们以 $I(\boldsymbol{Z}^V;\boldsymbol{X}^V)$ 为例说明这一过程。具体而言，根据公式 2，我们首先如下计算 $I(\boldsymbol{Z}^V;\ \boldsymbol{X}^V)$:

$$I(\boldsymbol{Z}^V;\boldsymbol{X}^V)=\mathbb{D}_{KL}(\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)||\boldsymbol{p}(\boldsymbol{Z}^V)). \tag{20}$$

然后，遵循先前的变分工作 [22], [30], [31],先验分布 $\boldsymbol{p}(\boldsymbol{Z}^V)$可被估计为标准高斯分布$\mathcal{N}(0,\mathrm{diag}(\boldsymbol{I}))$。对于后验分布 $\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$，其可被近似为一个变分后验分布 $\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$。因此，我们可以得到 $I(\boldsymbol{Z}^V;\ \boldsymbol{X}^V)$ 的上界为：

$$\begin{aligned}I(\boldsymbol{Z}^V;\boldsymbol{X}^V)&\le\mathbb{D}_{KL}(\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)||\boldsymbol{p}(\boldsymbol{Z}^V))\\&=\sum_{i=0}^{m+1}\mathbb{D}_{KL}\Big(\mathcal{N}(\boldsymbol{\mu}_i^v,[\mathrm{diag}(\boldsymbol{\sigma}_i^v)]^2)||\mathcal{N}\big(0,\mathrm{diag}(\boldsymbol{I})\big)\Big),\end{aligned} \tag{21}$$

其中 $\boldsymbol{q}_\phi^v$ 指式6中变分编码的参数。 $I(\boldsymbol{Z}^V;\boldsymbol{X}^V)$通过最小化其在式21中的上界作为目标函数进行优化。类似地， $I(\boldsymbol{Z}^T;\ \boldsymbol{X}^T)$的目标函数也可通过遵循从式20到式21的过程得到。

对重建项的最大化。重建项的目标是利用潜变量生成理想的任务期望信息 $\boldsymbol{R}$ 。我们通过所学文本/图像表示能够预测真实标签的概率来衡量重建信息的质量，如下所示：

$$I(\boldsymbol{Z}^T,\boldsymbol{Z}^V;\boldsymbol{R})\cong\mathbb{E}_{\boldsymbol{p}_\theta^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log p(Y|\boldsymbol{Z}^T,\boldsymbol{Z}^V)], \tag{22}$$

其中 Y 指目标（即多模态命名实体识别的金标准标签序列和多模态关系抽取的金标准关系）。$\boldsymbol{p}_\theta^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)$ $\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)$通过变分编码 $\boldsymbol{q}_\phi^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\mathbf{X}^V)$进行估计，我们可以得到 $I(\boldsymbol{Z}^T,\boldsymbol{Z}^V;\ \boldsymbol{R})$的**变分下界**，如下所示：

$$\begin{aligned}I(\boldsymbol{Z}^T,\boldsymbol{Z}^V;\boldsymbol{R})&\ge\mathbb{E}_{\boldsymbol{q}_\phi^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log p(Y|\boldsymbol{Z}^T,\boldsymbol{Z}^V)],\\&=\log(p(Y|\boldsymbol{Z}^T,\boldsymbol{Z}^V))+\sum_{Y'\in\boldsymbol{Y},Y'\neq Y}\log(1-p(Y'|\boldsymbol{Z}^T,\boldsymbol{Z}^V))\end{aligned} \tag{23}$$

其中 $\boldsymbol{Y}$ 指任务预测的标签空间， Y指真实标签，而 Y' 作为错误标签。为此，我们分别采用式18和式19中的 $p(\cdot|\cdot)$作为多模态命名实体识别和多模态关系抽取的$p(Y|\boldsymbol{Z}^T,\boldsymbol{Z}^V)$。因此，重构项可以使用相应的**第三节D3中的任务目标函数**进行优化。

2）对齐正则化器目标函数：优化 L_{ar} 的核心在于公式15中的**对基于 MI 的最大化项进行最大化**。我们遵循信息最大化 [32]，优化 L_{ar}，其中神经网络以对比方式度量 MI。相应地，我们构建判别器 $\mathcal{D}$ ，以如下方式度量文本 - 图像表征之间的一致程度：

$$I(\boldsymbol{Z}^T;\boldsymbol{Z}^V)=\mathbb{E}_{\boldsymbol{p}_\theta^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{p}_\theta^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log\mathcal{D}(\boldsymbol{Z}^T;\boldsymbol{Z}^V)]. \tag{24}$$

由于直接优化基于互信息的 L_{ar} 是不可行的，我们利用其**通过变分编码估计的下界**作为目标函数，如下所示：

$$\begin{aligned}I(\boldsymbol{Z}^T;\boldsymbol{Z}^V)&\ge\mathbb{E}_{\boldsymbol{q}_\phi^T(\boldsymbol{Z}^T|\boldsymbol{X}^T)\boldsymbol{q}_\phi^V(\boldsymbol{Z}^V|\boldsymbol{X}^V)}[\log\mathcal{D}(\boldsymbol{Z}^T;\boldsymbol{Z}^V)]\\&=\log\left(\mathcal{D}(\boldsymbol{Z}^T,\boldsymbol{Z}^{V+})\right)+\log\left(1-\mathcal{D}(\boldsymbol{Z}^T,\boldsymbol{Z}^{V-})\right),\end{aligned} \tag{25}$$

其中 $\boldsymbol{q}_\phi^T/\ \boldsymbol{q}_\phi^V$表示第三节B1中变分编码的参数，$(\boldsymbol{Z}^T,\ \boldsymbol{Z}^{V+})$ 指属于同一输入对的采样的文本-图像表示，否则为对比互信息中的批内负样本 $(\boldsymbol{Z}^T,\boldsymbol{Z}^{V-})$。由于 $\boldsymbol{Z}^T\in\mathbb{R}^{(n+2)\times d}$ 的形状以及

$\boldsymbol{Z}^V \in \mathbb{R}^{(m+1)\times d}$ 不匹配，我们对它们进行均值池化以便计算，即：

$$\mathcal{D}(\boldsymbol{Z}^T, \boldsymbol{Z}^V) = \text{Sigmoid}(\text{MLP}(\text{Pooling}(\boldsymbol{Z}^T) \oplus \text{Pooling}(\boldsymbol{Z}^V))), \tag{26}$$

其中，Pooling($\boldsymbol{Z}^T$) $\in \mathbb{R}^d$、Pooling($\boldsymbol{Z}^V$) $\in \mathbb{R}^d$和 $\oplus$ 表示向量之间的拼接操作。可使用标准的二元交叉熵优化该易处理函数。

3）任务目标函数：我们分别介绍多模态命名实体识别和多模态关系抽取的任务特定目标函数。

多模态命名实体识别。对于多模态命名实体识别，给定句子 X^T 及其 Gold 序列标签 Y，我们可以根据公式 18 得到 Y 的概率$p(Y|X^T)$，并计算相应的损失L_{ner} ，如下所示：

$$L_{task}^{ner} = -\log(p(Y|X^T)). \tag{27}$$

多模态关系抽取。对于多模态关系抽取，给定句子 X^T 和标注的头实体/尾实体 E_1/E_2，我们可以根据公式 19 得到概率$p(Y|X^T,\ E_1,\ E_2)$，并可推导出目标函数，如下所示：

$$L_{task}^{re} = -\log(p(Y|X^T, E_1, E_2)). \tag{28}$$

4）总体目标函数：基于上述函数，我们推导出 MMIB 的最终目标函数 L ，如下所示：

$$L = L_{rr} + L_{ar} + L_{task}, \tag{29}$$

其中， L_{rr} 中的所有极小性项和重建项由式21和式23进行可处理优化， L_{ar} 中的极大性项由式26进行可处理优化。L_{task}分别针对多模态命名实体识别和多模态关系抽取由式27和式28进行优化。

IV. 实验

A. 实验设置

1) 数据集：我们在*Twitter-2015* [33]和*Twitter-2017* [7] 上进行多模态命名实体识别实验，并在*MNRE* [8] 数据集上进行多模态关系抽取实验。这些数据集收集自Twitter上的多模态帖子，每条Twitter帖子由一段文本和一张图像组成。*Twitter-2015* 和 *Twitter-2017* 均包含四类实体：人物（**PER**）、地点（**地点**）、组织（**组织**）和其他类别（**MISC**）。表 I 展示了两个数据集的训练集、开发集和测试集中各类型实体的数量以及多模态推文的划分情况。*MNRE* 数据集包含9,201个句子-图像对，涵盖23个关系类别。表 II 给出了详细统计数据，并将其与广泛使用的关系抽取数据集SemEval-2010任务 8 [34], 进行比较，以体现MNRE在任务评估中的有效性。

2) 指标： 我们采用与基线方法相同的评估指标。对于**多模态命名实体识别**，当实体的跨度和实体类型均与 Gold答案匹配时，认为该实体被正确识别。对于 **多模态关系抽取**，当预测的关系类型符合 Gold 答案时，认为一对实体之间的关系被正确抽取。我们使用 Chen 等人发布的评估代码 [12], ，其中使用精确率（**精确率**）、召回率（**召回率**）和 F1分数（**F1 分数**）来评估性能。

表 I TWITTER-2015 和 TWITTER-2017 的基本统计信息

实体类型	Twitter15			Twitter17		
	训练集	Dev	Test	训练集	Dev	Test
PER	2217	552	1816	2943	626	621
LOC	2091	522	1697	731	173	178
ORG	928	247	939	1674	375	395
MISC	940	225	726	701	150	157
总计	7176	1546	5078	6049	1324	1351
推文	4000	1000 3257		3373	723	723

表 II MNRE 与 SemEval-2010 数据集的统计数据

数据集	句子	实体	关系	图像
MNRE	9,201	30,970	23	9,201
SemEval-2010	10,717	21,434	9	0

3) 超参数：为与以往研究 [12]进行公平比较，我们采用 BERT_{base} 和 ResNet50 分别作为文本和视觉编码器。因此，所有潜变量和表征的隐藏维度均为 768。我们使用 AdamW [35], 进行参数优化，其中衰减率为 0.01，学习率为 3e-5，批次大小为 8。我们手动调整拉格朗日乘子 β_1,β_2 ，并在 $\beta_1 = 0.1$和 $\beta_2 = 0.1$时取得最佳结果。通过截断较长的句子并填充较短的句子，MNER 的输入句子最大长度设为 128，MRE 的输入句子最大长度设为 80。模型实现代码已打包并作为补充材料上传，以供复现检查。

B. 基线方法

为了进行全面比较，我们将我们的方法与四组基线方法进行比较。

基于文本的基线方法：为验证引入视觉图像的有效性，我们选择了不融入视觉信息、同时执行两项任务的基线方法。**命名实体识别** 基线方法包括：1）**CNNBiLSTM-CRF** [36]通过双向长短期记忆网络和卷积神经网络，利用词级和字符级表征进行命名实体识别；2）**HBiLSTMCRF** [37] 将 CNNBiLSTM-CRF 中的卷积神经网络替换为长短期记忆网络；3）**T-NER** [38]是一种配备多种特征以提升性能的 Twitter专用命名实体识别系统，我们采用 Xu等人 [10]在所用数据集上的结果。**关系抽取** 基线方法包括：1）**PCNN** [39]使用卷积网络和分段池化；2）**MTB** [40] 是一种基于 BERT 的面向关系抽取的预训练模型。

朴素多模态基线方法：为确认解决模态噪声和模态差距的必要性，我们选择忽略这两个问题的朴素多模态基线方法。具体- 而言，**多模态命名实体识别** 基线方法包括 **1)GVATT** [7] 利用注意力机制结合图像级和文本级信息；**2) AdapCoAtt** [33] 设计了一个自适应协同注意力网络，以探索视觉信息；**3)UMT** [41]提出了一个多模态交互模块，以获得图像感知的词表征；**4)FMIT** [14] 设计了一个扁平多模态交互Transformer（FMIT）层，其中一个

表 III 总体多模态命名实体识别结果。

方法	Twitter15			Twitter17		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
T-NER	69.54	68.65	69.09	-	-	-
GVATT	73.96	67.90	70.80	83.41	80.38	81.87
AdapCoAtt	72.75	68.74	70.69	84.16	80.24	82.15
UMT	71.67	75.23	73.41	85.28	85.34	85.31
FMIT	74.18	75.03	74.60	85.55	85.29	85.42
UMGF	**74.49**	75.21	74.85	86.54	84.50	85.51
MAF	71.86	75.10	73.42	86.13	86.38	86.25
HVPNeT	73.87	76.82	75.32	85.84	87.93	86.87
MMIB	74.44	**77.68**	**76.02**	**87.34**	87.86	**87.60**

表 IV 多模态关系抽取总体性能。

方法	MRE		
	P	**R**	**F1**
PCNN	62.85	49.69	55.49
MTB	64.46	57.81	60.86
AdapCoAtt	64.67	57.98	61.14
VisualBERT	57.15	59.48	58.30
BERT+SG	62.95	62.65	62.80
UMT	62.93	63.88	63.46
UMGF	64.38	66.23	65.29
MEGA	64.51	68.44	66.41
HVPNeT	**83.64**	80.78	81.85
MKGformer	82.67	81.25	81.95
MMIB	83.49	**82.97**	**83.23**

统一格结构用于跨模态交互。为与 MMIB 进行公平比较，我们引用了仅包含一个 FMIT 层的 FMIT 性能。对于**多模态关系抽取**，其中包含 1）**VisualBERT** [42], , Chen 等人 [12] 利用预训练的多模态模型进行文本-图像编码与融合。2）**BERT-SG** [8] 直接整合了使用场景图（SG）工具提取的视觉特征 [43]。

考虑模态差距的多模态基线方法。我们选择倾向于推导一致的文本-图像表示的基线方法。对于**多模态命名实体识别**，包含：1）**UMGF** [11] 构建多模态图以实现语义对齐。2）**MAF** [10]采用对比学习方法以获得一致的表示。对于**多模态关系抽取**，**MEGA** [8] 构建双图以实现语义一致性。

考虑模态噪声的多模态基线方法。为验证 MMIB 应对模态噪声的优越性，我们选择了试图缓解该问题的基线方法。对于**多模态命名实体识别**，**HVPNeT** [12]通过探索层次化视觉特征，将其作为可插拔视觉前缀，从而缓解无关视觉对象带来的噪声。对于 **多模态关系抽取**，**MKGformer** [44]利用相关性感知融合模块来缓解噪声信息。

请注意，UMT、UMGF 和 MEGA 的原始结果仅涉及一项任务，我们引用了 Chen 等人 [12]的结果。

C. 主要结果

表 III和表 IV 分别展示了 MNER和MRE上的最终模型性能，MMIB 与所有基线方法之间存在显著差异（$p<0.05$）。从结果来看，我们有以下观察。

1)**视觉信息确实有助于提升性能**。由于多模态方法通常在两个任务上优于基于文本的方法，我们可以看出，引入视觉信息对于语义增强是必要的。然而，性能提升仍然有限，这需要对多模态方法进行进一步探索。

2)**缓解模态噪声并弥合模态差距可以为多模态方法带来性能提升**。在这两个表中，后两组基线方法表现优于普通的多模态方法。这证明了一致的图像-文本表示和抗噪模态信息的重要性。尽管如此，两项任务的最终性能仍难以令人满意，因此需要更好的策略来解决这两个问题。

3)**我们提出的MMIB在两项任务上取得了最佳结果**。如两张表所示，考虑到模态差距和模态噪声问题，MMIB 的表现优于此前的多模态基线方法。这证明了我们的方法在处理这两个问题方面的有效性。我们将这种性能优势归因于精炼正则器和对齐正则器能够有效地为最终任务预测导出“良好”的文本-图像表征。

V. 分析与讨论

A. 消融研究

为探究每个所提出组件对最终性能的贡献，我们在三个基准测试中的两个任务上开展消融研究。遵循以往研究 [10], [11], [12],，我们在表V的测试集中报告消融性能。具体而言，我们分别移除所提出的正则化项以确认其有效性，并给出相应分析如下。

1) – 精炼正则器 (L_{rr})：移除精炼正则器会导致在三个基准测试上的最终性能下降，揭示了限制任务相关噪声信息的重要性。没有 L_{rr}，两种模态中的噪声信息可能会干扰对明显视觉信息的融合，从而对最终性能产生负面影响。

2)– 对齐正则器 (L_{ar})：移除对齐正则器后，模型性能在所有基准测试上均有所下降。该现象表明弥合模态差距的必要性，并验证了L_{ar} 驱动一致的文本-图像表征的有效性。

3) – L_{rr} & L_{ar}：当我们同时移除所有正则化项时，最终性能会大幅下降。这表明这些正则化项不仅各自发挥作用，还会协同工作，从而提升最终性能。

表 V 消融研究。

消融	数据集	评估指标 P	R	F1	Δ F1
MMIB	Twitter15	74.44	77.68	76.02	-
	Twitter17	87.34	87.86	87.60	-
	MRE	83.49	82.97	83.23	-
$-\ L_{rr}$	Twitter15	73.36	76.61	74.95	-1.07
	Twitter17	86.67	87.56	87.11	-0.49
	MRE	82.10	81.71	81.91	-1.32
$-\ L_{ar}$	Twitter15	74.05	75.75	74.89	-1.13
	Twitter17	88.07	85.79	86.91	-0.69
	MRE	82.48	80.93	81.70	-1.53
$-\ L_{rr}\ \&\ L_{ar}$	Twitter15	74.02	74.81	74.42	-1.60
	Twitter17	87.15	85.86	86.50	-1.10
	MRE	80.62	80.00	80.31	-2.92

B. 模态噪声分析

为确认 MMIB 能够生成去除冗余噪声且保留任务预测性的文本/图像表示，我们展示了两个案例的可视化结果。为了量化原始表示$\mathbf{X}^T/\mathbf{X}^V$对学习得到的表示$\mathbf{Z}^T/\mathbf{Z}^V$的关注程度，我们定义了一个贡献分数来衡量这一点。我们以视觉表示为例，说明贡献分数的计算过程。具体而言，我们首先通过$\mathbf{X}^V/\mathbf{Z}^V$之间的点积得到矩阵$\mathbf{A}^V \in \mathbb{R}^{(m+1)\times(m+1)}$。然后，分别对$\mathbf{A}^V$的每一行求和，得到$\hat{}\mathbf{A}^V \in \mathbb{R}^{m+1}$，其中$\hat{}\mathbf{A}_i^V$是一个标量，表示该$i_{th}$图像对最终视觉表示$\mathbf{Z}^V$的贡献。按照这一过程，我们计算$\hat{}\mathbf{A}^V$、$\hat{}\mathbf{A}^T$并将其可视化。MMIB 和不含 L_{rr} 的 MMIB 的可视化结果分别以红色和绿色显示在色图中。如图 3.(a) 所示，对于多模态关系抽取，任务无关信息（“RT @Del: Tune in to”）在 MMIB 中得到弱化。对于视觉模态，在不含 L_{rr} 的 MMIB 中，每张图像对最终视觉表示的关注程度相同，导致模态融合中的注意力机制错误地融合了两位歌手的视觉信息，从而预测为 per/per/对等。与此同时，MMIB 会去除这类任务无关的视觉信息，从而更好地促进任务预测。对于图 3(b) 中的多模态命名实体识别，我们观察到任务无关信息（即文本“RT @FP”和图像中的视觉背景）对最终图像表示的贡献很小，各模态中的冗余噪声因此得到去除。凭借 L_{rr}，MMIB 对这两个案例均作出了正确预测。

此外，我们将 MMIB 与一些基线方法进行比较，以分析性能。具体而言，UMT 是一种基础多模态基线方法，它将 MNRE 中人物之间的关系预测为对等关系，并为实体分配实体类型

MNER 中从 PER 到“卡路里”。由于UMT通过注意力机制与图像中的所有网格进行跨模态交互，噪声网格可能会干扰注意力分配，从而导致错误的预测。对于MEGA而言，噪声信息可能会干扰文本实体与视觉对象之间的语义对齐。MAF 考虑了相关性

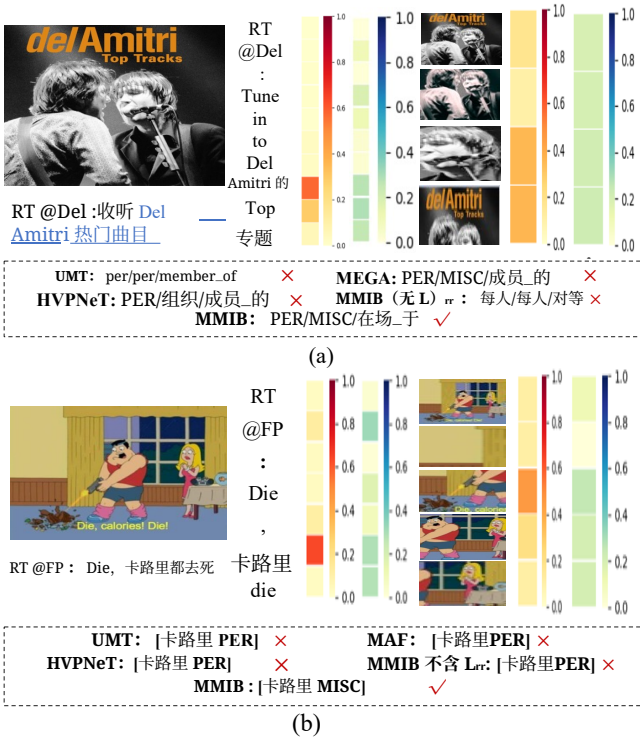


图 3。模态噪声案例研究。

在图像和文本之间进行全局处理，但缺乏细粒度的噪声过滤仍会产生错误的预测。尽管HVPNeT 试图通过视觉前缀来缓解视觉噪声，但它忽略了文本噪声语义。例如，在图 3(b) 中，词元“die”在内部仍保留着与人物相关的语义，因此视觉前缀可能会受到显著人物目标的误导。受这一现象影响，HVPNeT 无法进行正确的预测。

C. 信息瓶颈分析

为了分析信息瓶颈原理的作用方式，我们探究拉格朗日乘子 β_1, β_2在 L_{rr}中的取值如何影响最终性能。我们采用$\{0, 0, 0.01, 0.1, 0.5, 1.0, 1.5, 2.0\}$ 作为β_1, β_2的候选值。基于这些候选值，我们开展了三组实验：1）固定 $\beta_1 = 1$，并使用候选值调整 β_2；2）固定 $\beta_2 = 1$，并使用候选值调整 β_1；3）同时调整 β_1, β_2，并将它们的值保持为相同值。上述实验中 $F1$ 性能的波动分别以图4中的橙色、绿色和蓝色曲线表示。

从图 4 中可以看出，当 β_1 和 β_2 均设置为0.1时，能够取得最佳性能。具体而言，我们的观察和分析如下。1）同时调节 β_1, β_2 通常比分别调节它们取得更好的性能，这表明 L_{rr} 中的这两个IB 项相互协同工作，从而提升最终性能。2）当 β_1 或 β_2 设置为 0.0 时，性能会严重下降。这一现象体现了舍弃

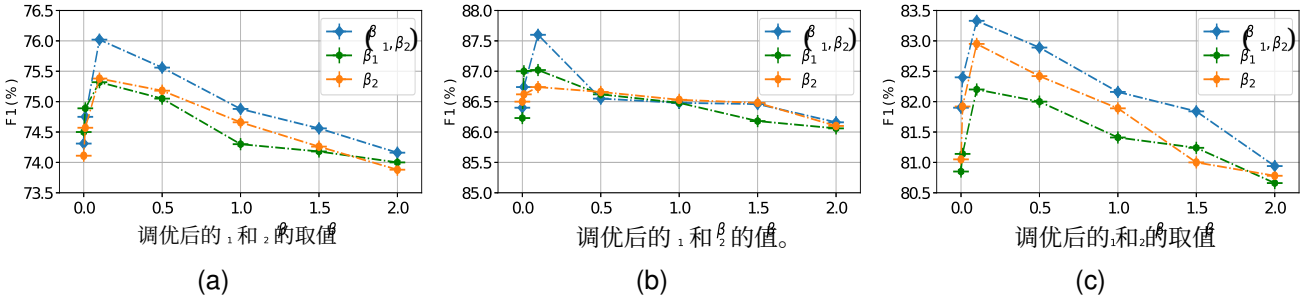


图4. 我们的模型在不同数据集上随 β_1 和 β_2 不同取值的 F1 分数变化：(a) Twitter15 数据集上的 F1 分数变化；(b) Twitter17 上的 F1 分数变化；(c) MRE 数据集上的 F1 分数变化

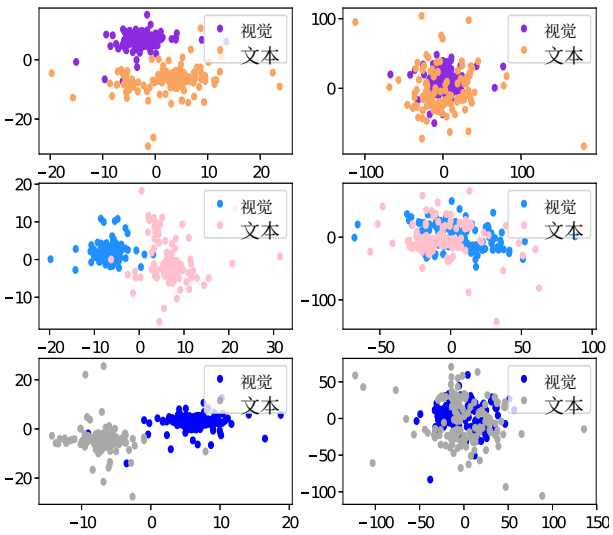


图 5. 模态差距的表示可视化。

来自原始文本和视觉信息的噪声信息。3) 当 β_1 或β_2 设置为 2.0 时，性能最差。原因可能是模型过度压缩了原始文本和视觉信息，导致仅保留了很少的有效信息来重构任务所需的信息。4) 当 $\beta_1 = 0.1$，$\beta_2 = 0.1$时，性能最佳，这表明两种模态中的噪声并未超过有效信息。即便如此，对于预测而言，丢弃噪声并控制噪声信息与有效信息之间的权衡仍然十分重要。

D. 模态差距可视化

为确认 MMIB 能够生成一致的文本-图像表征，我们对文本-图像表征进行可视化。具体而言，我们首先从输入的文本-图像对中，随机选取 150 组涉及 PER、地点和组织类型的实体-对象，并收集使用带有或不带对齐正则化器的 MMIB 训练后生成的表征。然后，使用 t-SNE [45] 对这些表征进行可视化，其训练迭代 500 次，困惑度设为 10。我们以 Twitter17 数据集上的结果为例，并在图 5 中展示结果，其中三行分别是 PER、地点和组织对应实体/对象的可视化。



(a) 多模态命名实体识别的一个实例。 (b) 多模态关系抽取的一个实例。

图 6. 所提出的 MMIB 模型的错误分析。

组织类型。在没有对齐正则器的情况下，实体/对象表征分别散布在各自的空间中，如图5 左列所示。相反，右列展示了从完整 MMIB 中获得的表示可视化结果。我们可以看到，对齐正则器能够实现文本和视觉表示之间的语义一致性。这一现象体现了对齐正则器对于一致的文本-图像表征学习的有效性。

E. 错误分析

为分析 MMIB 的潜在弱点，我们进行了错误分析。具体而言，我们分别从三个数据集中随机选取了 100 个预测错误的实例，并观察到，当文本与图像内容明显无关时，MMIB 的表现会受到影响。图 6 展示了一些典型实例。对于图 6.(a) 中的多模态命名实体识别，感兴趣的实体“凯尔特人”指的是一个组织类足球队。然而，图像中的显著对象是一块广告牌，并没有提供关于该足球队所属组织的明显线索。受到广告牌这一视觉线索的影响，MMIB 将“凯尔特人”识别为杂项类。图 6.(b) 中的多模态关系抽取也出现了类似情况。具体而言，句子描述的是在“纽约市”的“行程”，而图像中呈现的对象则与一个人物和一栋房屋有关。这里，文本与

图像内容之间没有明显的语义关联。因此，在这类视觉信息的引导下，MMIB 将居住地关系预测为给定实体之间的关系。

我们尝试推断出现这种现象的原因。尽管MMIB可以丢弃噪声信息，但当文本与图像的内容明显不相关时，仍不可避免地会引入一些噪声信息。我们认为，这指出了一个有趣的方向，并可能启发未来更多改进 MRE 和多模态命名实体识别的工作。

VI. 相关工作

A. 多模态实体与关系抽取

多模态命名实体识别（**MNER**）和多模态关系抽取（**MRE**）由于用户生成的多模态内容数量不断增加而引起了广泛的研究兴趣。现有研究可以分为两条路线。**第**一类研究直接将完整图像作为全局视觉线索，以指导文本任务预测。视觉信息通过将图像编码为一个特征向量 [9], [7], [33]或对应于网格的多个特征向量 [10], [41], [46],[47], [48], [49]来表示。相应地，研究者设计了基于注意力的机制，以促进文本与图像交互，从而为最终任务生成具有表现力的文本表示。然而，这些方法难以在视觉对象与命名实体之间建立细粒度的映射。**第**二类研究 [5], [6] 探索细粒度的对象级视觉信息，以提升模型性能。具体而言，这些研究采用预训练的目标检测模型来提取图像中的显著对象，其中视觉信息同时来自全局图像和局部对象图像。除了注意力机制 [5],[6],之外，研究者还提出了各种策略，以充分探索文本与图像交互。例如，Zheng 等人 [8]和 Zhang等人 [11] 利用多模态图结构来捕捉文本与图像之间的语义关联。Lu等人 [50]设计了用于跨模态交互的统一格结构。Wang等人 [51]将场景图开发为视觉内容的结构化表示，以进行语义交互。为了提供有关实体类型和图像区域的先验信息，Jia 等人 [15] 和 Jia 等人 [13]提出了基于机器阅读理解（MRC）的方法，其中MRC 中的查询经过精心设计，用于指导跨模态交互以进行任务预测。

尽管上述研究已取得巨大成功，但模态噪声和模态差距是两个尚未得到充分解决的问题。虽然大多数先前的工作忽略了模态噪声问题，但 Chen 等人 [12]尝试通过层次化视觉特征，防止无关对象中的视觉信息产生干扰。然而，我们关注两种模态中的噪声，旨在推导出不仅具有抗噪能力，而且对下游任务具有足够预测能力的表征。对于模态差距问题，尽管可以将图像或对象转换为文本描述 [6], [8], [16], 来弥合差距，但这种操作会受到外部工具误差传播的影响。不同于先前的工作，我们尝试通过表征同时解决这两个问题。从信息论视角进行学习，这种方法具有更强的泛化能力和适应能力，能够适用于其他任务。

B. 信息瓶颈

信息瓶颈（IB） [23] 是信息论中的重要概念，受到了广泛关注。IB 结合变分推断应用于深度神经网络，因此称为变分信息瓶颈（VIB） [18]。VIB 可作为表示学习的正则化技术，并已在计算机视觉 [31], [52], [53], 、自然语言处理 [21], [54], [55], [56]和推荐系统 [22]等领域展现出优势。尽管取得了上述诸多成果，VIB 尚未应用于多模态命名实体识别和多模态关系抽取，这仍是一个有待探索且值得关注的研究方向。本文采用 VIB 为多模态命名实体识别和多模态关系抽取提取任务所需且具备噪声鲁棒性的表征，并在三个公开基准上取得了出色的模型性能。

VII. 结论

在本文中，我们提出了一种用于多模态命名实体识别（MNER）和多模态关系抽取（MRE）的新方法 MMIB。MMIB旨在通过结合信息论的表示学习，解决这两个任务中的模态噪声和模态差距问题。具体而言，针对第一个问题，我们设计了一种探索信息瓶颈原理的精炼正则器，用于提取经过冗余噪声削弱且对最终任务保有预测性的文本-图像表示。针对第二个问题，我们提出了一种包含互信息项的对齐正则器，以获得一致的跨模态表示。在三个基准测试上的实验表明，MMIB 的性能显著优于当前最先进的基线方法。未来，我们将把 MMIB 应用于多媒体分析中的其他多模态任务。

参考文献

[1]J. Yu、B. Bohnet 和 M. Poesio，《将命名实体识别视为依存句法分析》，载于第*58*届计算语言学协会年会论文集，*ACL 2020*，线上会议，*2020*年*7*月*5*日至*10*日，D. Jurafsky、J. Chai、N. Schluter 和 J. R. Tetreault 编。计算语言学协会，2020年，第6470–6476页。

[ ] 10 18653 1 2020 577 线上会议。可访问：https://doi.org/。/v /。acl-main。Z. Zhang、B. Yu、X. Shu、M. Xue、T. Liu 和 L. Guo，《从“是什么”到“为什么”：利用依据图改进关系抽取》，载于计算语言学协会研究成果：*ACL/IJCNLP 2021*，线上会议，*2021*年*8*月*1*日至*6*日，ACL 研究成果系列，C. Zong、F. Xia、W. Li 和 R. Navigli 编，第 ACL/IJCNLP 2021 卷。计算语言学协会，2021年，第86–95页。 [线上会议]。可访问：https://doi.org/

[2] 10 18653 1 2021 8 。/v /。findings-acl。Q. Guo、F. Zhuang、C. Qin、H. Zhu、X. Xie、H. Xiong 和 Q. He，《基于知识图谱的推荐系统综述》，*IEEE* 知识与数据工程汇刊，第34卷，第8期，第3549–3568页，2022年。 [线上]。可访问：https://doi.org/

[3] 10 1109 2020 3028705 。/TKDE。。Y. Shen、X. Wang、Z. Tan、G. Xu、P. Xie、F. Huang、W. Lu 和 Y. Zhuang，《用于命名实体识别的并行实例查询网络》，载于第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），*ACL 2022*，爱尔兰都柏林，*2022*年*5*月*22*日至*27*日，S. Muresan、P. Nakov 和 A. Villavicencio 编。计算语言学协会，2022年，第947–961页。 [线上]。可访问：https://aclanthology.org/2022。acl-long.67

[4]

[5] Z. Wu、C. Zheng、Y. Cai、J. Chen、H. Leung 和 Q. Li，《在社交媒体帖子中使用嵌入式视觉引导对象进行命名实体识别的多模态表示》，载于*MM '20*：第*28*届 *ACM* 多媒体国际会议，线上会议／美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日，C. W. Chen、R. Cucchiara、X. Hua、G. Qi、E. Ricci、Z. Zhang 和 R. Zimmermann 编。ACM，2020年，第1038–1046 [ ] 10 1145 3394171 3413650 页。线上。可访问：https://doi.org/。/。

[6] C. Zheng、Z. Wu、T. Wang、Y. Cai 和 Q. Li，《通过对抗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* 多媒体汇刊，第 23， 2520 2532, 页–。线上。可访问：https://doi.org/。/TMM。。 2021 [ ] 10 1109 2020 3013398

[7] D. Lu、L. Neves、V. Carvalho、N. Zhang 和 H. Ji，《用于多模态社交媒体名称标注的视觉注意力模型》，载于第*56*届计算语言学协会年会论文集，*ACL 2018*，澳大利亚墨尔本，*2018*年*7*月*15*日至*20*日，第*1*卷：长篇论文，I. Gurevych 和 Y. Miyao 编。计算语言学协会，2018年，第1990–1999页。 [线上]。可访问：https://aclanthology.org/P18-1185/

[8] C. Zheng、J. Feng、Z. Fu、Y. Cai、Q. Li 和 T. Wang，《通过高效图对齐进行多模态关系抽取》，载于 *MM '21*：*ACM* 多媒体会议，线上会议，中国，*2021*年*10*月*20*日至*24*日，H. T. Shen、Y. Zhuang、J. R. Smith、Y. Yang、P. Cesar、F. Metze 和 B. Prabhakaran 编。ACM，2021年，第5298–5306页。 [线上]。可访问：https://doi.org/ 10 1145 3474085 3476968 。/。

[9] S. Moon、L. Neves 和 V. Carvalho，《面向短社交媒体帖子的多模态命名实体识别》，载于*2018*年计算语言学协会北美分会：人类语言技术会议论文集，*NAACL-HLT 2018*，美国路易斯安那州新奥尔良，*2018*年*6*月*1*日至*6*日，第*1*卷（长篇论文），M. A. Walker、H. Ji 和 A. Stent 编。计算语言学协会，2018年，第852–860页。 [线上]。可访问：https://doi.org/ 10 18653 1 18 1078 。/v /n - B. Xu、S. Huang、C. Sha 和 H. Wang，

[10] 《MAF：用于多模态命名实体识别的通用匹配与对齐框架》，载于*WSDM '22*：第*15*届 *ACM* 网络搜索与数据挖掘国际会议，线上会议／美国亚利桑那州坦佩，*2022*年*2*月*21*日至*25*日，K. S. Candan、H. Liu、L. Akoglu、X. L. Dong 和 J. Tang 编。ACM，2022年，第1215–1223页。 [线上]。可访问：https://doi.org/ 10 1145 3488560 3498475 。/。

[11] D. Zhang、S. Wei、S. Li、H. Wu、Q. Zhu 和 G. Zhou，《通过目标视觉引导进行命名实体识别的多模态图融合》，载于第*35*届 *AAAI* 人工智能会议，*AAAI 2021*；第*33*届人工智能创新应用会议，*IAAI 2021*；第*11*届人工智能教育进展研讨会，*EAAI 2021*；线上会议，*2021*年*2*月*2*日至*9*日。AAAI出版社，2021年，第14 347–14 355页。 [线上]。可访问：https://ojs.aaai.org/index.php/AAAI/article/view/17687

[12] X. Chen、N. Zhang、L. Li、Y. Yao、S. Deng、C. Tan、F. Huang、L. Si 和 H. Chen，《良好的视觉引导造就更好的抽取器：用于多模态实体与关系抽取的层次化视觉前缀》，载于计算语言学协会研究成果：*NAACL 2022*。美国西雅图：计算语言学协会，2022年7月，第1607–1618页。 [线上]。可访问：https://aclanthology.org/2022 121 。findings-naacl。

[13] M. Jia、X. Shen、L. Shen、J. Pang、L. Liao、Y. Song、M. Chen 和 X. He，《查询先验至关重要：用于多模态命名实体识别的机器阅读理解框架》，载于*MM '22*：第*30*届 *ACM* 多媒体国际会议，葡萄牙里斯本，*2022*年*10*月*10*日至*14*日，J. Magalh˜aes、A. D. Bimbo、S. Satoh、N. Sebe、X. Alameda-Pineda、Q. Jin、V. Oria 和 L. Toni 编。ACM，2022年，第3549–3558页。 [ ] 10 1145 3503161 3548427 线上。可访问：https://doi.org/。/。

[14] J. Lu、D. Zhang、J. Zhang 和 P. Zhang，《用于命名实体识别的扁平多模态交互Transformer》，载于第*29*届计算语言学国际会议论文集，*COLING 2022*，韩国庆州，*2022*年*10*月*12*日至*17*日。国际计算语言学委员会，2022年，第2055–2064页。 [ ] 2022 1 179 线上。可访问：https://aclanthology.org/。coling-。

[15] M. Jia、L. Shen、X. Shen、L. Liao、M. Chen、X. He、Z. Chen 和 J. Li，《MNER-QG：一种结合查询定位的多模态命名实体识别端到端机器阅读理解框架》，*CoRR*，第 abs/2211 14739，卷。。线上。可访问：https://doi.org/。/arXiv。。 2022 [ ] 10 48550 2211 14739

[16] X. Wang、M. Gui、Y. Jiang、Z. Jia、N. Bach、T. Wang、Z. Huang 和 K. Tu，《ITA：用于多模态命名实体识别的图像-文本对齐》，载于计算语言学协会北美分会*2022*年会议：人类语言技术论文集。美国西雅图：计算语言学协会 计算语言学，2022年7月，第3176–3189页。 [在线]。获取地址：https://aclanthology.org/ 2022 232 。

[17] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，载于第2届国际学习表征会议，ICLR 2014，班夫，加拿大艾伯塔省，2014年4月14日至16日，会议论文集，Y. Bengio 和 Y. LeCun 编，2014年。 [在线]。获取地址：http://arxiv.org/abs/1312 6114 。

[18] A. A. Alemi、I. Fischer、J. V. Dillon 和 K. Murphy，“深度变分信息瓶颈”，载于第5届国际学习表征会议，ICLR *2017*， *24 26*, 土伦，法国，2017 [ ] *2017*, 4月—会议论文集。OpenReview.net，。在线。获取地址：https://openreview.net/forum?id=HyxQzBceg

[19] E. F. T. K. Sang 和 J. Veenstra，“文本块的表示”，载于 EACL 1999，即计算语言学协会欧洲分会第9届会议，1999年6月8日至12日，卑尔根大学，挪威卑尔根。计算机语言学协会，1999年，第173–179页。 [在线]。获取地址：https://aclanthology.org/E99-1023/

[20] C. E. Shannon，“通信的数学理论”，《贝尔系统技术期刊》，第27卷，第3期，第379–423页，1948年。 [在线]。获取地址：https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[21] C. Zhang、X. Zhou、Y. Wan、X. Zheng、K. Chang 和 C. Hsieh，“利用信息瓶颈提升自然语言处理模型的对抗鲁棒性”，载于计算语言学协会论文集：ACL 2022，爱尔兰都柏林，2022年5月22日至27日，S. Muresan、P. Nakov 和 A. Villavicencio 编。计算语言学协会，2022年，第3588–3598页。 [在线]。获取地址：https://doi.org/10 18653 1 2022 284 . /v / .findings-acl.

[22] J. Cao、J. Sheng、X. Cong、T. Liu 和 B. Wang，“通过变分信息瓶颈向冷启动用户提供跨域推荐”，载于第38届 IEEE 国际数据工程会议，ICDE 2022，马来西亚吉隆坡，2022年5月9日至12日。IEEE，2022年，第2209–2223页。 [在线]。获取地址：https://doi.org/10 1109 53745 2022 00211 . /ICDE . .

[23] N. Tishby 和 N. Zaslavsky，“深度学习与信息瓶颈原理”，载于2015年 IEEE 信息论研讨会，ITW 2015，以色列耶路撒冷，2015年4月26日至5月1日。IEEE，2015年，第1 5 [ ] 10 1109 2015 7133169 –页。在线。获取地址：https://doi.org/ . /ITW. .

[24] S. Gershman 和 N. D. Goodman，“概率推理中的摊销推断”，载于认知科学学会第36届年会论文集，CogSci 2014，加拿大魁北克市，2014年7月23日至26日，P. Bello、M. Guarini、M. McShane 和 B. Scassellati 编。cognitivesciencesociety.org，2014年。 [在线]。获取地址：https://mindmodeling.org/cogsci2014/papers/098/

[25] J. Devlin、M. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，载于计算语言学协会北美分会：人类语言技术会议2019论文集，NAACL-HLT 2019，7，2019，2 1 美国明尼苏达州明尼阿波利斯，6月—第1卷（长篇和短篇论文），J. Burstein、C. Doran 和 T. Solorio 编。计算语言学协会，2019年，第4171 4186 [ ] 10 18653 1 19 1423 –页。在线。获取地址：https://doi.org/ . /v /n - Z. Yang、B. Gong、L. Wang、W. Huang、D. Yu 和 J. Luo，“一种快速准确的视觉定位单阶段方法”，载于2019年 IEEE/CVF 国际计算机视觉会议，ICCV 2019，韩国首尔，2019年10月27日至11月2日。IEEE，2019年，第4682–4692页。 [ ] 10 1109 2019 00478 在线。获取地址：https://doi.org/ . /ICCV. .

[26] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于2016年 IEEE计算机视觉与模式识别会议，CVPR 2016，美国内华达州拉斯维加斯，2016年6月27日至30日。IEEE计算机学会，2016年，第770–778页。 [在线]。获取地址：https://doi.org/10 1109 2016 90 . /CVPR. .

[27] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，“注意力就是一切”，载于神经信息处理系统进展第30卷：神经信息处理系统2017年度会议，2017年12月4日至9日，美国加利福尼亚州长滩，I. Guyon、U. von Luxburg、S. Bengio、H. M. Wallach、R. Fergus、S. V. N. Vishwanathan 和 R. Garnett 编，2017年，第5998–6008页。 [在线]。获取地址：https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[28] P. Colombo、P. Piantanida 和 C. Clavel，“一种用于学习解缠文本表示的互信息新型估计器”，载于计算语言学协会第59届年会暨第11届国际自然语言处理联合会议论文集，ACL/IJCNLP 2021，（第1卷：长篇论文），线上会议，2021年8月1日至6日，C. Zong、F. Xia、W. Li 和 R. Navigli 编。计算语言学协会，2021年，

第6539–6550页。[51] [在线]。获取地址：https://doi.org/10.18653/v1/2021.acl-long.[30] I. Higgins、L. Matthey、A. Pal、C. P. Burgess、X. Glorot、M. M. Botvinick、S. Mohamed 和 A. Lerchner，“β-VAE：通过受约束变分框架学习基本视觉概念”，载于第5届国际学习表征会议，ICLR 2017，法国土伦，2017年4月24日至26日，会议论文集。OpenReview.net，2017年。 [在线]。获取地址：https://openreview.net/ forum?id=Sy2fzU9gl [31] T. Q. Chen、X. Li、R. B. Grosse 和 D. Duvenaud，“分离变分自编码器中的解缠因素”，载于神经信息处理系统进展第31卷：神经信息处理系统年度会议 *2018, 2018, 3 8,* NeurIPS 12 月—加拿大蒙特利尔，S. Bengio、H. M. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编，2018年，第2615–2625页。 [在线]。获取地址：https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html[32] P. Velickovic、W. Fedus、W. L. Hamilton、P. Li`o、Y. Bengio 和 R. D. Hjelm，“深度图信息最大化”，载于第7届国际学习表征会议，ICLR 2019，美国路易斯安那州新奥尔良，2019年5月6日至9日。OpenReview.net，2019年。 [在线]。获取地址：https://openreview.net/forum?id=rklz9iAcKQ [33] Q. Zhang、J. Fu、X. Liu 和 X. Huang，“用于推文命名实体识别的自适应协同注意力网络”，载于第三十二届 AAAI人工智能会议论文集（AAAI-18）、第30届人工智能创新应用会议（IAAI-18）和第8届 AAAI人工智能教育进展研讨会（EAAI-18），美国路易斯安那州新奥尔良，2018年2月2日至7日，S. A. McIlraith 和 K. Q. Weinberger 编。AAAI出版社，2018年，第5674–5681页。 [在线]。获取地址：https://www.aaai.org/ocs/index.php/AAAI/AAAI 18 16432 / [34] paper/view/ I. Hendrickx、S. N. Kim、Z. Kozareva、P. Nakov、D. O. Seaghdha、S. Pad´o、M. Pennacchiotti、L. Romano 和 S. Szpakowicz，“Semeval-2010 任务8：名词性成分对之间语义关系的多路分类”，CoRR，第 abs/1911.10422卷，2019年。 [在线]。获取地址：http://arxiv.org/abs/1911 10422 。I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，载于 ICLR 论文集，2019年。[36] X. Ma 和 E. H. Hovy，“通过双向长短期记忆网络—卷积神经网络—条件随机场实现端到端序列标注”，载于计算语言学协会第54届年会论文集，ACL 2016，2016年8月7日至12日，德国柏林，第1卷：长篇论文。计算机语言学协会，2016年。 [在线]。获取地址：https://doi.org/ 10 18653 1 16 1101 . /v /p - G. Lample、M. Ballesteros、[37] S. Subramanian、K. Kawakami 和 C. Dyer，“用于命名实体识别的神经网络架构”，载于 NAACL HLT 2016，即计算语言学协会北美分会2016年人类语言技术会议，美国加利福尼亚州圣迭戈，2016年6月12日至17日，K. Knight、A. Nenkova 和 O. Rambow 编。计算语言学协会，2016年，第260–270页。 [在线]。获取地址：https://doi.org/ 10 18653 1 16 1030 . /v /n - A. Ritter、S. Clark、Mausam [38] 和 O. Etzioni，“推文中的命名实体识别：一项实验研究”，载于2011年自然语言处理实证方法会议论文集，EMNLP 2011，2011年7月27日至31日，约翰·麦金太尔会议中心，英国爱丁堡，ACL 特别兴趣小组 SIGDAT 会议。ACL，2011年，第1524–1534页。 [在线]。获取地址：https://aclanthology.org/D11-1141/ [39] D. Zeng、K. Liu、Y. Chen 和 J. Zhao，“通过分段卷积神经网络进行远程监督关系抽取”，载于2015年自然语言处理实证方法会议论文集，EMNLP 2015，葡萄牙里斯本，2015年9月17日至21日，L. M`arquez、C. Callison-Burch、J. Su、D. Pighin 和 Y. Marton 编。计算语言学协会，2015年，第1753–1762页。 [ ] 10 18653 1 15 1203 在线。获取地址：[40] https://doi.org/ . /v /d - L. B. Soares、N. FitzGerald、J. Ling 和 T. Kwiatkowski，“填补空白：关系学习中的分布相似性”，载于计算语言学协会第57届会议论文集，ACL 2019，意大利佛罗伦萨，2019年7月28日至8月2日，第1卷：长篇论文，A. Korhonen、D. R. Traum 和 L. M`arquez 编。计算语言学协会，2019年，第2895–2905页。 [ ] 10 18653 1 19 1279 在线。获取地址：[41] https://doi.org/ . /v /p - J. Yu、J. Jiang、L. Yang 和 R. Xia，“通过统一多模态Transformer中的实体跨度检测提升多模态命名实体识别计算语言学协会，ACL 2020，线上，2020年7月5日至10日，D. Jurafsky、J. Chai、N. Schluter与J. R. Tetreault编。计算语言学协会，2020年，第3342–3352页。 [ ] 10 18653 1 2020 306 在线。可获取：[42] https://doi.org/ . /v / .acl-main. L. H. Li、M. Yatskar、D. Yin、C. Hsieh与K. Chang，“VisualBERT：一种简单且高性能的视觉与语言基线”，CoRR，第 abs/1908.03557卷，2019年。 [在线]。可获取：http://arxiv.org/abs/1908 03557 。K. Tang、Y. Niu、J. Huang、J. Shi与H. [43] Zhang，“从有偏训练中进行无偏场景图生成”，见2020年IEEE/CVF计算机视觉与模式识别会议，CVPR 2020，美国华盛顿州西雅图，2020年6月13日至19日。计算机视觉基金会 / IEEE，2020年，第3713–3722页。 [在线]。可获取：https://openaccess.thecvf.com/ content CVPR 2020/html/Tang Unbiased Scene Graph Generation From Biased Training CVPR 2020 paper.html[44] X. Chen、N. Zhang、L. Li、S. Deng、C. Tan、C. Xu、F. Huang、L. Si与H. Chen，“用于多模态知识图谱补全的多层级融合混合Transformer”，见SIGIR ’22：第45届国际ACM信息检索研究与发展会议，西班牙马德里，2022年7月11日至15日。ACM，2022年，第904 915 [ ] 10 1145 3477495 3531992 –页。在线。[45] 可获取：https://doi.org/ . / . G. E. Hinton与S. T. Roweis，“随机邻居嵌入”，见神经信息处理系统进展15/神经信息处理系统，NIPS 2002，2002年12月9日至14日，加拿大不列颠哥伦比亚省温哥华/，S. Becker、S. Thrun与K. Obermayer编。麻省理工学院出版社，2002年，第833–840页。 [在线]。可获取：https://proceedings.neurips.cc/paper/2002/hash/6150ccc6069bea6b5716254057a194ef-Abstract.html[46] L. Sun、J. Wang、K. Zhang、Y. Su与F. Weng，“Rpbert：一种基于文本-图像关系传播的BERT模型，用于多模态命名实体识别”，见第35届 AAAI 人工智能会议，AAAI 2021，第33届人工智能创新应用会议，IAAI 2021，第11届人工智能教育进展研讨会，EAAI 2021，线上会议，2021年2月2日至9日。AAAI出版社，2021年，第13 860–13 868页。 [在线]。可获取：https://ojs.aaai.org/index.php/AAAI/article/view/17635 [47] S. Chen、G. Aguilar、L. Neves与T. Solorio，“图像能否帮助识别实体？图像在多模态命名实体识别中作用的研究”，见第七届噪声用户生成文本研讨会论文集，W-NUT 2021，线上，2021年11月11日，W. Xu、A. Ritter、T. Baldwin与A. Rahimi编。计算语言学协会，2021年，第87–96页。 [在线]。可获取：https://doi.org/ 10 18653 1 2021 1 11 . /v / .wnut- . C. Zheng、Z. Wu、J. [48] Feng、Z. Fu与Y. Cai，“MNRE：一个具有挑战性的多模态数据集，用于社交媒体帖子中带视觉证据的神经关系抽取”，见2021年IEEE多媒体与博览会国际会议，ICME 2021，中国深圳，2021年7月5日至9日。IEEE，2021年，第1–6页。 [在线]。可获取：https://doi.org/ 10 1109 51207 2021 9428274 . /ICME . . B. Xu、S. Huang、[49] M. Du、H. Wang、H. Song、C. Sha与Y. Xiao，“不同的数据，不同的模态！强化数据划分以实现从社交媒体帖子中有效进行多模态信息抽取”，见第29届计算语言学国际会议论文集，COLING 2022，韩国庆州，2022年10月12日至17日。国际计算语言学委员会，2022年，第1855–1864页。 [在线]。可获取：https://aclanthology.org/ 2022 1 160 .coling- . J. Lu、D. Zhang、J. Zhang与[50] P. Zhang，“用于命名实体识别的扁平多模态交互Transformer”，见第29届计算语言学国际会议论文集。韩国庆州：国际计算语言学委员会，2022年10月，第2055–2064页。 [在线]。可获取：https: // aclanthology.org/ 2022 1 179 .coling- . J. Wang、Y. Yang、K. Liu、Z. Zhu与[51] X. Liu，“M3S：场景图驱动的多粒度多任务学习，用于多模态命名实体识别”，IEEE ACM音频、语音与语言处理汇刊，第 31, 111 120, 卷，第–页。在线。2023 [ ] 10 1109 2022 可获取：https://doi.org/ . /TASLP. . 3221017 [52] A. Bardera、J. Rigau、I. Boada、M. Feixas与M. Sbert，“使用信息瓶颈方法的图像分割”，IEEE图像处理汇刊，第18卷，第7期，第1601–1612页，2009年。 [在线]。可获取：https://doi.org/ 10 1109 2009 2017823 . /TIP. . S. Motiian与G. [53] Doretto，“面向视觉识别的带特权信息的信息瓶颈域适应”，见计算机视觉 - ECCV 2016 - 第14届欧洲会议，荷兰阿姆斯特丹，2016年10月11日至14日，论文集，第七部分，计算机科学讲义丛书，B. Leibe、J. Matas、N. Sebe与

M. Welling，编，卷9911。Springer，2016，第630–647页。 [在线]。链接：https://doi.org/ 10 1007 978 3 319 46478 7 39 。

[54] R. K. Mahabadi、 Y. Belinkov 和 J. Henderson，“ 面向有效低资源微调的变分信息瓶颈”，载于第*9*届国际学习表征会议（*ICLR 2021*），线上会议，奥地利，*2021*年*5*月*3*日至7日。OpenReview.net，2021年。 [在线]。链接：https://openreview.net/forum?id=kvhzKz- DMF

[55] J. Zhou、Y. Wu、Q. Chen、X. Huang 和 L. He，“通过微调和压缩进行注意”，载于计算语言学协会论文集：*ACL/IJCNLP 2021*，线上会议，*2021*年*8*月*1*日至*6*日，ACL论文集系列，C. Zong、F. Xia、W. Li 和 R. Navigli 编，卷ACL/IJCNLP 2021。计算语言学协会，2021年，第2152–2161页。 [在线]。链接：https://doi.org/ 10 18653 1 2021 189 。 /v / .findings-acl。J.

[56] Zhou、Q. Zhang、Q. Chen、L. He 和 X. Huang，“通过变分信息瓶颈进行事件论元抽取的多格式迁移学习模型”，载于第*29*届计算语言学国际会议论文集，*COLING 2022*，韩国庆州，*2022*年*10*月*12*日至*17*日。国际计算语言学委员会，2022年，第1990–2000页。 [在线]。链接：https://aclanthology.org/2022.coling-1.173