语言空间中的社交媒体跨模态话语理解

Chunpu Xu[1], **Hanzhuo Tan**[1],**JingLi**[1*], **Piji Li**[2] 香港理工大学计算学系[2] 南京航空航天大学计算机科学与技术学院[1]
{chun-pu.xu,han-zhuo.tan}@connect.polyu.hk[1]
jing-amelia.li@polyu.edu.hk; [2]pjli@nuaa.edu.cn

arXiv:2302.13311v1 [cs.MM] 26 Feb 2023

摘要

社交媒体上包含文本和图像的多媒体通信十分普遍。然而，关于图像如何与文本组织在一起，并在人类认知中形成连贯意义的研究却十分有限。为弥补这一空白，我们提出了跨模态话语这一新概念，以反映人类读者如何结合对图像和文本的理解。在多媒体语境中，首先从图像中提取文本描述（称为**字幕**）。随后，我们进一步采用五种标签——实体级插入、投射和具体化，以及场景级重述和扩展——来刻画字幕和文本的结构，并呈现二者共同表达的意义。作为一项初步研究，我们还构建了首个包含 1.6 万条多媒体推文的数据集，并对话语标签进行了人工标注。实验结果表明，基于多头注意力并结合图像描述的多媒体编码器能够取得最先进的结果。

1 引言

多媒体日益普及，正在彻底改变社交媒体上的交流方式。传统的纯文本形式已拓展为涉及文本和图像的跨模态信息交流。对于多媒体消息，语言理解不仅需要理解视觉语义和文本语义，还需要弄清楚是什么将二者联系起来，使其在人类心智中呈现连贯的意义。

然而，社交媒体语言理解领域的大多数进展都依赖文本来学习消息级语义（Shen 等人，2018年；Nguyen 等人，2020年），在很大程度上忽略了图像所传达的丰富含义（Cai 等人，2019a；Wang 等人，2020b）。其他近期的多模态研究则侧重于设计模型，以结合视觉和文本信号（Park 等人，2019年；Li 等人，2020年；Yu 等人，2021年），忽略了人们如何理解多媒体帖子所隐含结构的相关见解。

鉴于这些问题，我们将图像视为社交媒体语言不可或缺的一部分，并提出一个全新的跨模态话语概念，用以定义人类读者如何从图像和文本模态中构建连贯的意义。我们的研究受到Vempala 和 Preotiuc-Pietro（2019年）关于图像与文本之间信息重叠的研究启发；与此同时，我们更进一步，刻画多媒体消息如何为人类所理解，这超越了对是否观察到新事物进行简单的是或否预测。据我们所知，我们首次将话语这一纯语言学概念拓展至图像，以界定图像所发挥的语言学角色，以及图像与文本之间的语用关系如何共同塑造连贯的意义。

一般而言，跨模态话语由人类感知中用于关联图像与文本语义的操作来定义。读者可以首先从获取的图像中提取信息，以完成跨模态理解，信息可以是局部对象（实体），也可以是全局场景（Rayner，2009）。随后，提取出的实体或场景会以文本形式呈现，称为图像的**字幕**，并可进一步与多媒体语境中相匹配的文本共同构成实体级或场景级话语。具体而言，实体级话语可根据实体是被省略、描述还是映射，细分为插入、投射和具体化；类似地，场景级重述和扩展则用于体现一种模态中的故事是否在另一种模态中重现或延续。

为说明上述定义，图 1展示了五条多媒体推文。如 (a) 所示，读者可能会关注“草莓”这一对象，并 将 其名称插入文本中，而省略该实体。至于 (b)，“咖啡”这一对象

* 通讯作者



图1：五种跨模态话语标签及其示例。从上到下，各行依次展示其文本、图像、Vempala 和 Preotiuc-Pietro （2019年）中的图文关系标签，以及我们的跨模态话语类别。Vempala 和 Preotiuc-Pietro （2019年）中的标签关注图像是否为文本增添了新的含义，而我们的标签则定义了图像的语言学角色及其与文本之间促进连贯性的语用关系。

应从图像中提取，以具体说明文本中的“咖啡”一词。在 (c) 中，文本里的“法院”一词与“法槌”这一对象相关联。(d) 中的图像有助于复述文本所描绘的场景（狗在车里牵着手）。在 (e) 中，整体场景作为文本的扩展，补全了故事：“沿着通往海滩的轨道前行，我们距离夏天又近了一步。”。

相反，Vempala 和 Preotiuc-Pietro （2019年）中的图文关系仅限于判断图像是否为文本增添了新的含义，但这仍不足以反映人们如何在多媒体语境中理解语言。

作为跨模态话语的一项试点研究，我们还提出了首个用于探索这一任务的数据集。该数据集从 Twitter 收集，包含 1.6 万条高质量多媒体帖子，并对其话语标签进行了人工标注。[1] 我们相信，我们的任务及其配套数据集作为同类研究中的首个成果，将有望帮助机器获得理解包含多模态元素的社交媒体语言的能力。

为此，我们提出了一个用于学习文本与图像之间话语结构的框架。受多模态学习领域最新进展的启发（Wang 等人，2020b；Yuet al.，2020年），我们采用多头注意力机制（Vaswani 等人，2017年）来探索反映跨模态交互的视觉文本表示。此外，为了表征用于话语学习的字幕，我们利用由 COCO 图像描述数据集（Lin 等人，2014b）训练的模型生成的图像说明作为附加特征。

针对跨模态话语的实证研究，我们在自己的数据集上进行了全面实验。分类对比结果表明，机器推断话语结构面临诸多挑战，先进的多模态编码器也无法很好地处理我们的任务。尽管如此，探索文本、图像描述和视觉—文本交互之间的相关性，有助于在类内评估和总体评估中展现最先进的性能。我们还进一步考察了不同模态和文本长度的影响，发现文本信号对话语推断至关重要，而文本、图像和图像描述的联合效应带来了最佳结果。最后，定性分析展示了模型中的多头注意力如何解读话语结构。

2 相关工作

我们的论文涉及多媒体学习和自然语言处理中的话语分析。下面将进一步详细介绍。

多媒体学习。我们的论文属于跨媒体研究这一领域，旨在融合文本和视觉特征。已有多种深度学习方法被提出，用于利用跨模态特征，这些方法或基于协同注意力（Xu and Saenko, 2016; Lu et al., 2016）和多头注意力（Vaswani 等人, 2017; Wang 等人 , 2020a ）等先进神经网络架构，或基于预训练视觉语言表征（Lu et al., 2019; Su et al., 2020; Zhang et al., 2021）。这些方法在传统视觉语言任务（如图像描述（Park et al., 2019; Zhou et al., 2020; Shi et al., 2021）和视觉问答（VQA）（Yu et al., 2019; Tan and Bansal, 2019; Si et al., 2021））以及社交媒体

[1]数据集和代码已发布于https://github.com/cpaaax/Multimodal_话语。

应用中也展现出有效性，例如讽刺检测（Caiet al.，2019a）、事件追踪（Li et al.，2020年；Abav-isanietal.，2020b）、关键词短语预测（Zhanget al.，2019年；Wangetal.，2020b）。

可以看出，迄今为止，这一研究方向的大多数进展都致力于改进通用（Su 等人，2020年；Zhou 等人，2020年）或特定应用（Wang 等人，2020b）的方法设计，以更好地捕捉不同模态之间匹配的语义。然而，由于图文之间复杂的交互作用（Vempala 和 Preotiuc-Pietro，2019年），这些方法在社交媒体数据上的有效性难免会受到影响。因此，我们借鉴人类感知的见解，从语言学视角解读图文关系，并提出在多媒体语境中学习话语结构的任务。这是一项基础性研究，有望帮助模型获得跨模态理解能力，并惠及各种下游应用。

我们也与先前为理解社交媒体中图文关系而开展的分类任务相关，例如信息重叠（Vempala 和 Preotiuc-Pietro，2019年）、兴趣点类型（Villegas 和 Aletras，2021年）、作者意图（Kruk 等人，2019年）、物体所有关系（Chinnappa 等人，2019年）等。此外，有趣的是，“话语”这一概念也被用于考察烹饪食谱中的图文关系（Alikhani 等人，2019年）。与这些将视觉和文本嵌入拼接到一个“共同”空间中的研究相比，我们设计了文本形式的字幕来传达视觉故事，并探究它们如何在语言空间中与帖子文本共同构成连贯的意义。这将使深层语义学习能够捕捉图像和文本模态所承载的隐含结构，而现有模型可能无法通过简单的特征拼接获得语言理解能力。

话语分析。 本研究与文本级话语结构的既有研究相关。采用修辞结构理论（RST）（Mann 和 Thompson，1988；Liu 等人，2019年）或宾夕法尼亚话语树库（PDTB）（Prasad 等人，2008；Xu 等人，2018年）的常见任务，探究能够将话语单元（如短语或句子）连贯连接起来、形成连贯性的修辞关系。这些研究已证明其对各种自然语言处理应用大有助益（Choubey 等人，2020年），例如情感分析（Bhatia 等人，2015年）、文本分类（Ji 和 Smith，2017年）以及微博摘要（Li 等人，2018年）。然而，鲜有研究将社交媒体图像视为多媒体语境中语用结构的话语单元；本研究旨在填补这一空白。

3 研究设计

在本节中，我们首先定义预测 §3.1中的跨模态话语任务。然后，我们介绍如何构建 §3.2中的数据集，随后在 §3.3 中进行数据分析，并在 §3.4中探讨潜在应用。

3.1 任务定义

在我们的任务中，输入是来自社交媒体多媒体帖子的图文对，遵循以往的做法（Vempala 和 Preotiuc-Pietro，2019年）。对于每个图文对，目标是从预定义的标签集合中输出一个标签，该集合涵盖社交媒体上跨模态话语的主要类别。我们的直觉是，图像相对更吸引注意力，且可能会在文本之前被处理。关于图像理解，心理学实验的既有发现（Rayner，2009）指出，人类可能会先识别并提取全局场景中的意义，以填补语境中的信息缺口；如果缺口仍然存在，他们可能会回过头来捕捉局部对象。基于此，我们首先根据是提取对象还是场景来理解图像和文本的联合含义，将话语标签集合粗略划分为实体（对象）层面和场景层面。

为进一步阐释标签设计，从图像（作为物体或场景）中提取的信息会映射到文本模态，形成**字幕**，这使我们能够说明人类感官如何借助字幕和帖子文本构建连贯的意义。

对于实体级话语，我们考察三种情况：实体在文本中被省略、提及或关联。对于缺失实体（例如图 1(a)），应将以实体名称形式呈现的字幕插入帖子文本中，以补全消息的含义；而图 1(b) 中的实体则由图像中的物体具体化。图1(c) 中的实体则被隐含地投射到相关物体上。因此，我们将实体级插入、**具体化**和**投射**分别用于描述上述三种情况。

同样，场景级话语可分为**重述** 和**扩展**两类。

前者指图像充当文本描述（例如，图 1(d)）；后者则指帖子呈现图像场景，以阐述文本中留白的故事（例如，图 1(e)）。

3.2 数据收集与标注

我们的数据集取自 Twitter[2]，，Twitter 正日益受到数字通信研究的关注（Mozafari 等人，2019年；Nikolov 和 Radi- vchev，2019年；Müller 等人，2020年），且多媒体帖子的使用尤为普遍（Vempala 和 Preotiuc-Pietro，2019年；Wang 等人，2020b）。我们首先使用 Twitter 流式 API 抓取原始数据[3]，并移除非英语帖子、仅含文本的帖子以及包含多张图像的帖子。随后，为了更好地对噪声较多的 Twitter 数据中的话语进行建模（Vempala 和 Preotiuc-Pietro，2019年），我们删除了可能妨碍学习有意义话语信号的样本。此处有四类“糟糕”的图文对可能在学习过程中引入大量噪声，如图2所示。

第一类是配有一些引言、用于分享人生感悟的图像肖像（以下简称**肖像**），其中图像与文本之间缺乏连贯关系（从语言学角度来看），因此无法为其定义话语结构。此外，其中许多包含作者的自拍照，可能引发隐私问题。第二类帖子，即**背景**，依赖外部知识来理解其含义（例如图2（b）），而仅凭图像和匹配文本进行语言理解无法做到这一点。第三类是**低质量图像**（例如低分辨率或模糊的图像，如图2（c）），很难从中捕捉视觉含义。最后一类是**OCR 字幕**（图2（d）），字幕以光学字符的形式出现在图像中。这可能会使跨模态话语退化为文本层面话语，并导致模型学习到无关紧要的特征。

在数据标注过程中，我们首先为每个话语标签选取了25个典型示例，并将其与标注指南（其中详细描述了每个标签）一同提供，以进行质量控制。随后，我们招募了两名语言学专业的研究生，在给定图文对的情况下，手动标注话语类别。在标注过程

[2]twitter.com[3]developer.twitter.com/en/docs/ 教程/实时推文流式传输

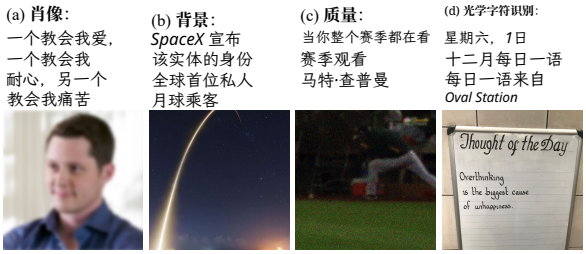


图2：四种“糟糕”类型的推文示例。(a)**肖像** 图像，其中文本中包含引言。(b)**背景** 是理解内容所必需的外部信息（此处为火箭轨迹场景）。(c)**低质量** 图像，其中物体几乎无法辨认。(d)**OCR字幕** （“每日思考”）以光学字符的形式出现在图像中。

	总计	Ins	Con	Pro	Res	Ext
Num	16,000	839	10,558	690	1,826	2,087
Len	10.69	9.11	10.85	10.98	11.24	9.92

表1：总体数据及各标签数据的统计信息：Ins：插入；Con：具体化；Pro：投射；Res：重述；Ext：扩展。Len：文本的平均词数。Num：推文数量。

中，还应标出属于上述四种类型的“差”样本。标注者间一致性为79.8%，为确保噪声数据中的特征学习质量，我们仅保留了两位标注者意见一致的数据。最后，剔除“差”类型的帖子后，最终数据集包含1.6万条带有人工标注的多媒体推文，涵盖五种话语类别。

3.3 数据分析

我们在此对数据集进行初步分析，并在表1中展示统计信息。标签分布不均衡，其中具体化和扩展标签相较于其他三种标签更为常见。这表明推特用户在选择如何组织文本和图像时偏好各异，也凸显了模型完成这项任务时面临的潜在挑战。

就文本长度而言，大多数推文包含的词数有限，这给模型从文本信号中捕捉关键特征带来了挑战。有趣的是，我们将统计结果与先前研究中的其他纯文本推特数据集（Wang 等人，2019年）进行比较，发现我们的多媒体推文平均词数少了30%。这意味着作者可能倾向于在多媒体帖子文本中放入较少内容，并通过图像呈现缺失的信息作为补充。我们还注意到，插入和扩展话语的文本平均相对较短。这可能是因为它们通过图像呈现了文本中省略、但实际存在的内容。

为了进一步描述我们数据集中的文本长度，

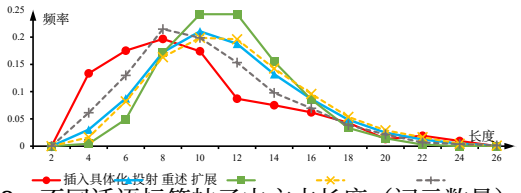


图3：不同话语标签帖子中文本长度（词元数量）的分布。

图 3 展示了不同标签的推文文本词数分布。所有曲线都显示出文本长度分布稀疏，这归因于社交媒体写作形式自由。插入和扩展曲线首先在 8 个词处达到峰值，而其他曲线则在 10–12 个词处达到峰值，随后都呈现长尾分布。这再次表明，多媒体帖子中的文本可能提供的内容有限，而插入和扩展中的文本词数更少。

3.4 潜在应用

在本小节中，我们进一步探讨这项任务和数据集潜在的下游应用，这或许能为未来工作的设计提供启发。一项直接的应用是微博摘要——这是一项重要任务，旨在从海量社交媒体数据中提炼显著内容。许多先进的摘要模型仅支持文本输入，而多媒体帖子在社交媒体上十分常见，因此可能需要将这些帖子压缩为文本，以便于处理。这不同于传统的图像描述任务（Anderson 等人，2018年；Rennie 等人，2017年；Huang 等人，2019年），在该任务中，生成的图像描述是从图像翻译而来的。对于社交媒体帖子而言，文本不能简单地被视为图像的“翻译”，因为图像与文本之间可能存在模糊的互动。鉴于话语分析在摘要中发挥着至关重要的作用（Xu 等人，2020年），不难设想，我们所研究的跨模态话语——描述图像与文本如何构建连贯性——将有助于多媒体摘要研究。

此外，跨模态话语可以被视为一项基础任务，并可能有助于社交媒体上的其他下游任务（例如，多模态命名实体识别（Yu 等人，2020年）、多模态危机事件分类（Abavisani 等人，2020a）、多模态讽刺检测（Cai 等人，2019b）、多模态情感分析（Truong 和 Lauw，2019年）以及多模态话题标签预测（Wang 等人，2020c））。然而，先前的大多数研究都侧重于利用视觉和语言表征，却忽视了作为连接这两种模态的语言本质。近年来，一些研究提出采用多任务学习，在多模态学习中考虑图文关系。例如，Sun 等人（2021年）研究文本与图像之间的关系传播，以提高推文中命名实体识别的准确率。Ju 等人（2021年）利用多模态关系类型作为辅助标签，探索多模态方面级情感分析。这些研究取得的积极结果表明，跨模态话语（作为对图文关系的语言描述）有潜力惠及广泛的多模态应用。此外，Sun 等人，2021年；Ju 等人，2021年）所使用的图文关系训练数据集是由Vempala 和 Preotiuc-Pietro（2019年）提出的 TRC数据集。与 TRC数据集相比，我们提出的话语数据集规模大得多（即 16K 对 4.5K），并且具有更细粒度的图文关系标签，如图 1所示。因此，我们相信，我们的数据集也将有助于提升各种多模态模型的性能。

4 话语学习框架

本节将介绍我们的框架，该框架结合图像和文本中的信号来预测其话语标签。如图 4所示，模型架构利用从文本、图像和图像描述（用于反映字幕）中学习到的表示，这部分内容将在§4.1中介绍。接下来，我们将讨论如何结合多模态表示 §4.2。最后，§4.3将介绍我们如何预测话语标签并设计训练流程。

4.1 文本、图像和图像描述编码

文本编码。 这里介绍如何学习文本特征。文本编码器基于预训练 Bertweet 的底部 6 层（Nguyen 等人，2020年）。它接收一个长度为 L的词元序列，并将其表示嵌入为一系列隐藏状态 $\mathbf{H}_{text} = (\mathbf{h}_1, ..., \mathbf{h}_L)$，其中每个元素都反映一个词元嵌入。$\mathbf{H}_{text}$ 随后经过最大池化层，生成$\bar{\mathbf{H}}_{text}$ 来表示文本。

图像编码。 为探索视觉信号，使用基于 CNN 的 ResNet-101 对图像进行编码（He 等，2016 年），该模型在 ImageNet 上预训练（Russakovsky 等，2015年）。提取 ResNet-101 最后一个卷积层的输出，作为输入图像的表示。首先将特征图的大小缩减至 $M \times M \times 2048$，然后

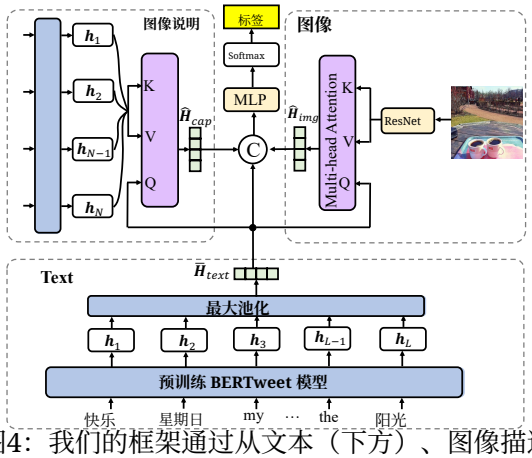


图4：我们的框架通过从文本（下方）、图像描述（左上方）和图像（右上方）编码得到的表示来学习跨模态话语。编码后的图像描述和文本在视觉文本语境中的输出层进行比较。

重塑为 $M^2 \times 2048$。每个 1×2048 向量表示相应图像区域中的视觉特征，并通过线性层投影到与文本特征**h** 相同的维度。帖子级视觉特征记为$\mathbf{H}^{img} = (\mathbf{v}_1, ...\mathbf{v}_{M^2})$，其中$\mathbf{v}_i$ 指表示图像中某个区域特征的 1×2048 向量。

图像描述编码 为了从图像中捕获更多语义信息，我们进一步利用图像描述（下文简称图像描述）作为一种附加模态。我们的直觉是，图像描述可以将图像所蕴含的关键视觉语义注入文本中的描述性语言（Xu 等人，2015年）。它们可能有助于反映图像对象之间丰富的交互，并发现作为关键话语指标的字幕式线索。我们首先采用 Anderson 等人（2018年）提出的模型，预测每幅图像的图像描述。该图像描述模型在 COCO 图像描述数据集（Lin 等人，2014年b）上进行了预训练，该数据集主要由社交媒体领域之外的自然图像组成。随后，我们按照文本编码的相同流程（上文所述）对图像描述的词元序列进行编码，并得到图像描述表示：$\mathbf{H}^{cap} = (\mathbf{h}_1, ..., \mathbf{h}_N)$。其中，$N$表示图像描述中的词元数量，$\mathbf{h}_i$ 表示 Bertweet 编码器的第 i个隐藏状态。

4.2 融合多模态表示

正如先前研究所指出的（Wang 等人，2020b），与广受研究的视觉语言数据集相比（Lin 等人，2014a；Young 等人，2014年），社交媒体数据中的模态呈现出更为复杂的交互。为使该框架能够关注各种类型的跨模态交互，我们采用多头注意力（Vaswani 等人，2017年）来全面探索编码后的图像特征（$\mathbf{H}^{img}$）¯与最大池化文本表示（$\mathbf{H}_{text}$）之间的交互。

具体而言，我们将文本特征设为查询**Q**，将图像特征设为键和值**K**，**V**，并按如下方式计算多头注意力 $MA(\cdot)$：

$$MA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [hd_1; ...; hd_n]\mathbf{W}^O \quad (1)$$

其中，n 是头的数量，$[\cdot]$ 表示拼接操作，j 号头的注意力为：

$$hd_j = A(\mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V) \quad (2)$$

$$A(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \theta(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \quad (3)$$

其中，d_k是归一化因子，$\theta(\cdot)$表示Softmax函数。$\mathbf{W}^O$，$\mathbf{W}_j^Q$，$\mathbf{W}_j^K$，$\mathbf{W}_j^V$是可学习参数。注意力图像（感知到ˆ文本）记为$\mathbf{H}^{img}$，它进一步作为上下文，帮助从图像描述和文本中探索话语线索。

在话语建模中，我们将编码文本¯$(\mathbf{H}_{text})$与图像描述（带有字幕风格特征）进行比较，以推断字幕如何与文本共同组织。为此，我们首先采用多头注意力机制对文本感知注意力字幕$\mathbf{H}^{cap}$进行编码，从图像描述中提取显著内容，以指示ˆ话语类别。此外，$\mathbf{H}^{cap}$ 与¯拼接，并与 $\mathbf{H}_{text}$ 共同建模其结构；同时，将注意力图像 $\mathbf{H}^{img}$ 拼接为图文交互上下文，用于跨模态话语学习。

4.3 话语预测与模型训练

话语标签由一个多层感知机（MLP）预测，其输入为 $\mathbf{H} = \hat{} [\mathbf{H}^{cap}; \mathbf{H}_{text}; \mathbf{H}_{img}]$，这些是整合后的特征向量，并进一步通过Softmax函数激活，以预测四种话语标签的概率。关于训练集，回顾表 1，我们观察到该任务存在严重的标签不平衡。为了解决这一问题，我们采用加权交叉熵损失，其权重根据训练数据中各标签的比例设定。

5 实验设置

模型设置。推文文本（L）和图像描述（N）的长度均通过截断限制为 20。批量大小设置为 100，学习率设置为 5×10^{-5}。所有多头注意力

方法	插入	具体化	投射	重述	扩展	F1
基线						
Qin 等人（2016年）	41.13	69.91	26.13	39.67	41.15	61.67
Rutherford 和 Xue（2016年）	43.17	70.78	32.62	42.31	40.82	62.73
Nam 等人（2017年）	46.49	74.83	33.33	39.33	42.39	65.76
文本+图像						
CONCATFUSE	52.86	81.62	34.78	39.19	42.93	71.09
ATTENTION	54.30	82.64	33.71	39.23	39.41	71.48
CO-ATTENTION	51.90	83.31	36.36	42.57	40.59	72.37
多头注意力	53.69	84.33	36.96	42.11	42.01	73.33
文本+图像说明						
CONCATFUSE	52.00	81.11	33.33	41.18	43.02	70.82
ATTENTION	54.79	81.26	36.78	39.72	42.55	70.97
CO-ATTENTION	53.73	82.13	37.20	41.78	39.16	71.38
多头注意力	53.79	82.27	34.55	43.96	43.46	72.08
图像+文本+图像说明						
CONCATFUSE	52.48	82.41	32.97	43.01	42.39	71.88
ATTENTION	53.24	83.01	34.95	43.45	43.65	72.58
CO-ATTENTION	54.81	83.98	36.96	45.24	39.76	73.15
多头注意力(完整模型)	**57.75***	**84.88***	**37.36**	**46.15***	**44.19***	**74.51***

表2：基线模型与我们各变体的比较结果。带 * 的分数表示我们完整模型相较于基线模型的显著性检验，p 值为 <0.05。

层的头数均设置为 6。对于图像编码，图像特征图大小 M 设置为 14。对于文本和评论编码，从 Bertweet 模型底部的 6 层中提取表示，并在训练集中进一步进行微调。在该设置中，我们随机划分 80%、10% 和 10% 的数据，分别用于训练集、验证集和测试集。在评估中，我们报告每个标签预测的 F1分数，并使用加权 F1 衡量整体结果。

基线模型与比较。 我们首先考虑 Qin 等人(2016年) 和 Rutherford 和 Xue(2016年) 提出的两个文本层面话语解析器，并将其文本编码器扩展为多模态编码器，以适配图文对。随后，我们将其与一种广受欢迎的多模态分类器进行比较（Nam 等人，2017年），该分类器采用双重注意力网络融合视觉和文本特征。

此外，我们评估了模型中不同的特征组合：测试集 + 图像、文本 +图像说明，以及文本 + 图像 + 图像说明（完整组合）。请注意，我们的框架采用多头注意力来整合从不同模态中学习到的特征。在实验中，我们还测试了其他模态融合方案的性能，包括基于简单特征拼接的方法（CONCATFUSE）、传统注意力机制（ATTENTION）以及协同注意力机制（CO-ATTENTION）。

6 实验讨论

本节首先展示主要比较结果（§6.1）。随后，我们讨论模型对以下因素的敏感性：在 §6.2 中变化的模态和文本长度。最后，§6.3 通过一个案例研究提供更多见解。

6.1 主要比较结果

表2 展示了各种多模态编码器的主要比较结果。由此可以得出以下观察结果。

首先，所有模型的 F1 表现都不理想。这表明跨模态话语预测是一项具有挑战性的任务。仅仅将话语解析器简单地适配到多模态设置，或应用现有的视觉语言编码器，都无法获得对该任务的良好理解。其次，两个实体级话语标签（即插入和具体化）的结果相对优于场景级标签，这表明局部对象比全局场景更容易被捕捉。在所有标签中，模型在具体化方面表现最佳，这可能归因于其拥有更丰富的特征学习数据样本（如表1所示）。模型在投射方面的结果则最差。原因可能是模型需要额外的知识来学习对象与实体之间的隐含关系。

最后，图像、文本和图像描述都有助于构建自动话语理解。联合建模这三种模态，使相应模型的表现优于其文本+图像和文本+图像描述的对应模型。

6.2 对模态和文本长度的敏感性

不同模态。 为进一步考察模态变化的影响，我们比较了 F1

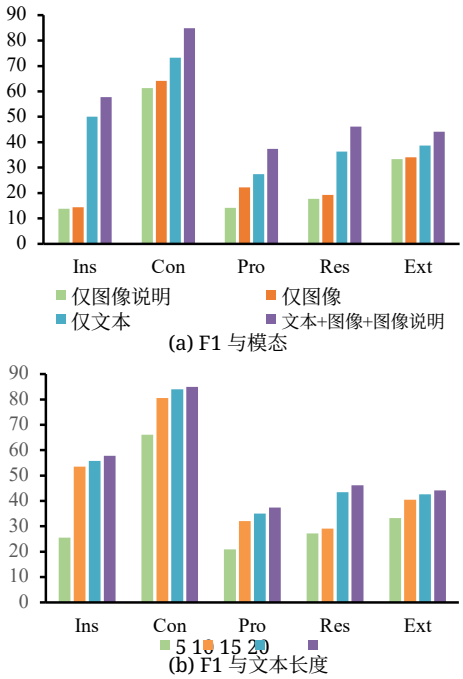


图5：完整模型与不同模态消融的性能对比见（a），其在不同文本长度下的结果见（b）。X轴：插入、具体化、投影、重述和扩展；Y轴：F1分数。对于每个标签，从左到右的柱形分别表示仅使用图像说明、仅使用图像、仅使用文本的消融结果，以及（a）中的完整模型；（b）中则表示推文文本长度上限分别为5、10、15和20时的结果。

分数，即完整模型与仅使用图像描述、仅使用图像和仅使用文本的消融模型在图中的分数。5(a)。可以看出，从所有标签来看，文本模态对话语建模的贡献相对更大，尤其是在插入任务中，此时命名实体被省略，使文本风格更易于识别。尽管如此，图像、文本和图像描述共同发挥的作用，在所有标签上都取得了最佳表现。

不同的文本长度。 如上所述，文本特征对于预测跨模态话语至关重要。这里我们进一步考察文本长度对模型性能的影响，完整模型的结果见图 5（b）。文本越长，得分越高，因为较长文本能够呈现更丰富的内容。这再次证明，文本为推断跨模态话语提供了至关重要的信号。

6.3 定性分析

以上讨论主要涉及图像说明和文本模态。这里我们通过一个案例研究，探究模型如何在视觉信号中体现话语指标。



图6：过采样图像的多头注意力热力图可视化。

案例研究。 通过热力图分析视觉特征（见图 6），热力图展示了图像上感知文本的注意力权重（公式 3），这些权重来自图文交互。可以看出，注意力能够突出显著区域，这些区域体现了与文本之间的重要语义联系，例如 (a) 和 (b) 中的实体（狗和吉普车）。还可以观察到，注意力在不同区域的关注重点会有所不同：对于实体层级话语，它往往集中于显著对象（实体）的某些部分；而对于场景层级话语，注意力还会考察背景，以获取全局视图。

7 结论

我们提出了一项学习跨模态话语的新任务，使模型能够在多媒体语境中获得社交媒体语言理解能力。为处理复杂的图文交互，我们首先将视觉语义转换为文本形式的字幕，然后将其与帖子文本进行比较，以探索语言空间中的深层句法关系。在实证研究中，我们进一步贡献了首个数据集，其中包含 1.6 万条经过人工标注、带有图文对话语标签的推文。数据集上的主要比较结果表明，多头注意力在探索文本、图像和图像说明模态之间的交互方面具有有效性。进一步讨论表明，我们有潜力生成能够揭示隐含图文结构的有意义表示。这些话语特征传达了与人类感知一致的关键语言线索，可能会极大促进社交媒体自动跨模态理解的未来发展。

局限性

类别不平衡是本研究的主要局限之一。如表1所示，具体化是占比最大的类别，占数据集的66.0%；而占比较小的类别（例如投射和插入）分别仅占 4.3% 和 5.2%。尽管这种不均衡分布反映了推文中真实的图文关系情形，但未来的研究应获取更多少数类别样本，以更好地解释图文关系。

后续研究也应考虑跨语言和多平台研究。探究不同语言中跨模态话语类别的分布将颇具趣味且富有启发性。是否存在会影响图像和文本使用方式的文化特征？与此同时，不同社交媒体平台也可能呈现出对图像和文本使用方式的偏好。例如，与 Twitter 用户相比，Instagram 用户是否更倾向于省略命名实体（插入类别）？

还可以采用更具体的模型，例如视觉语言 Transformer，将文本、图像说明和图像联合编码。目前的模型可在单张 NVIDIA RTX3080Ti GPU 上高效运行，而视觉语言 Transformer的训练成本可能较高，且需要更大的数据集。未来研究可以探讨计算成本与分类性能之间的权衡。

伦理考量

我们声明，本数据集不会引发伦理问题。首先，我们遵循 Twitter API 规范的标准数据获取流程。我们下载数据用于学术研究，且符合Twitter使用条款。随后，我们对数据进行了全面检查，并确保其中不含任何可能引发伦理问题的内容，例如有毒语言、人脸图像和审查图像。接着，我们对数据进行匿名化，以保护用户隐私。在语言使用方面，我们仅保留包含英文文本的帖子。对于人工标注，我们聘请标注者担任兼职研究助理，报酬为每小时 16 美元。

致谢

本论文得到国家自然科学基金青年科学基金项目（编号：62006203、62106105）以及研究资助局（香港特别行政区，中国）资助项目（编号：PolyU/25200821）、香港理工大学内部经费（1-BE2W、4-ZZKM 和 1-ZVRH）以及 CCF-百度开放基金（编号：2021PP15002000）的有力支持。

参考文献

Mahdi Abavisani、Liwei Wu、Shengli Hu、Joel Tetreault 和 Alejandro Jaimes。2020a。社交媒体中危机事件的多模态分类。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14679–14689页。

Mahdi Abavisani、Liwei Wu、Shengli Hu、Joel R. Tetreault 和 Alejandro Jaimes。2020b。社交媒体中危机事件的多模态分类。载于*2020 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，第14667–14677页。IEEE。

Malihe Alikhani、Sreyasi Nag Chowdhury、Gerard de Melo 和 Matthew Stone。2019年。CITE：图文话语关系语料库。载于*2019*年北美计算语言学协会分会会议：人类语言技术论文集，*NAACL-HLT 2019*，美国明尼苏达州明尼阿波利斯，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），第570–575页。计算语言学协会。

Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、Stephen Gould 和 Lei Zhang。2018年。用于图像描述和视觉问答的自底向上与自顶向下注意力。载于*IEEE*计算机视觉与模式识别会议论文集，第6077–6086页。

Parminder Bhatia、Yangfeng Ji 和 Jacob Eisenstein。2015年。通过RST话语解析改进文档级情感分析。载于*2015*年自然语言处理实证方法会议论文集，*EMNLP 2015*，里斯本，葡萄牙，*2015*年*9*月*17*日至*21*日，第2212–2218页。计算语言学协会。

Yitao Cai、Huiyu Cai 和 Xiaojun Wan。2019a。采用层次融合模型进行Twitter多模态讽刺检测。载于计算语言学协会第*57*届年会论文集，*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，第2506–2515页。计算语言学协会。

Yitao Cai、Huiyu Cai 和 Xiaojun Wan。2019b。基于层次融合模型的 Twitter 多模态讽刺检测。载于计算语言学协会第 *57* 届年会论文集，第 2506–2515 页。

Dhivya Chinnappa、Srikala Murugan 和 Eduardo Blanco。2019年。从社交媒体中提取所有物：图像补充语言。载于*2019*年实证方法会议论文集

自然语言处理与第九届国际自然语言处理联合会议论文集，*EMNLP-IJCNLP 2019*，香港，中国，*2019*年*11*月*3*日至*7*日，第663–672页。计算语言学协会。

Prafulla Kumar Choubey、AaronLee、Ruihong Huang 和 Lu Wang。2020年。话语作为事件的函数：围绕主要事件分析新闻文章中的话语结构。载于计算语言学协会第*58*届年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第5374–5386页。计算语言学协会。

何恺明、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016年。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页。

Lun Huang、Wenmin Wang、Yaxian Xia 和 Jie Chen。2019年。通过自适应注意力实现自适应对齐的图像描述。载于神经信息处理系统进展，第8942–8951页。

Yangfeng Ji 和 Noah A. Smith。2017年。用于文本分类的神经话语结构。载于计算语言学协会第*55*届年会论文集，*ACL 2017*，加拿大温哥华，*7*月*30*日至*8*月*4*日，第*1*卷：长篇论文，第996–1005页。计算语言学协会。

Xincheng Ju、Dong Zhang、Rong Xiao、Junhui Li、Shoushan Li、Min Zhang 和 Guodong Zhou。2021年。结合辅助跨模态关系检测的联合多模态方面情感分析。载于*2021*年自然语言处理实证方法会议论文集，*EMNLP 2021*，线上会议／多米尼加共和国蓬塔卡纳，*2021*年*11*月*7*日至*11*日，第4395–4405页。计算语言学协会。

Julia Kruk、Jonah Lubin、Karan Sikka、Xiao Lin、Dan Jurafsky 和 Ajay Divakaran。2019年。整合文本与图像：确定 Instagram 帖子中的多模态文档意图。载于 *2019*年自然语言处理实证方法会议及第*9*届自然语言处理国际联合会议论文集，*EMNLP-IJCNLP 2019*，中国香港，*2019*年*11*月*3*日至*7*日，第4621–4631页。计算语言学协会。

Jing Li、Yan Song、Zhongyu Wei 和 Kam-Fai Wong。2018年。微博上的会话话语与潜在主题联合模型。计算语言学，44(4)。

Manling Li、Alireza Zareian、Qi Zeng、Spencer White- head、Di Lu、Heng Ji 和 Shih-Fu Chang。2020年。用于多媒体事件抽取的跨媒体结构化共享空间。载于计算语言学协会第*58*届年会论文集语言学，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第2557–2568页。计算语言学协会。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014年a。Microsoft COCO：上下文中的常见物体。载于欧洲计算机视觉会议，第740–755页。Springer。

Tsung-Yi Lin、Michael Maire、Serge J. Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C. Lawrence Zitnick。2014年b。Microsoft COCO：上下文中的常见物体。载于计算机视觉——*ECCV 2014*——第*13*届欧洲计算机视觉会议，瑞士苏黎世，*2014*年*9*月*6*日至*12*日，论文集，第*V*部分，《计算机科学讲义》*Lecture Notes in ComputerScience*第8693卷，第740–755页。Springer。

Linlin Liu、Xiang Lin、Shafiq R. Joty、Simeng Han 和 Lidong Bing。2019年。层次指针网络解析。载于《*2019*年自然语言处理经验方法会议暨第*9*届自然语言处理国际联合会议论文集》，*EMNLP-IJCNLP 2019*，香港，中国，*2019*年*11*月*3*日至*7*日，第1007–1017页。计算语言学协会。

Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019年。ViLBERT：为视觉语言任务预训练任务无关的视觉语言表示。载于《神经信息处理系统进展》第*32*卷：*2019*年神经信息处理系统年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，温哥华，加拿大不列颠哥伦比亚省，第13–23页。

Jiasen Lu、Jianwei Yang、Dhruv Batra 和Devi Parikh。2016年。用于视觉问答的层次化问题-图像协同注意力。收录于《神经信息处理系统进展》第*29*卷：神经信息处理系统*2016*年年会，*2016*年*12*月*5*日至*10*日，西班牙巴塞罗那，第289–297页。

William C Mann 和 Sandra A Thompson。1988年。修辞结构理论：迈向文本组织的功能理论。文本，8(3)：243–281。

Marzieh Mozafari、RezaFarahbakhsh 和 Noel Crespi。2019年。一种用于在线社交媒体仇恨言论检测的基于BERT的迁移学习方法。收录于复杂网络及其应用国际会议，第928–940页。Springer。

Martin Müller、Marcel Salathé 和 Per E Kummervold。2020年。Covid-Twitter-BERT：一种用于分析 Twitter 上新冠肺炎内容的自然语言处理模型。*arXiv* 预印本 *arXiv:2005.07503*。

Hyeonseob Nam、Jung-Woo Ha 和 Jeonghee Kim。2017年。用于多模态推理和匹配的双重注意力网络。载于*IEEE*计算机视觉与模式识别会议论文集，第299–307页。

Dat Quoc Nguyen、Thanh Vu 和 Anh Tuan Nguyen。2020年。Bertweet：一种用于英语推文的预训练语言模型。载于*2020*年自然语言处理实证方法会议：系统演示，*EMNLP 2020* - 演示，线上，*2020*年*11*月*16*日至*20*日，第9–14页。计算语言学协会。

Alex Nikolov 和 Victor Radivchev。2019年。Nikolov-Radivchev 在 SemEval-2019 任务6中的工作：使用 BERT 和集成模型进行攻击性推文分类。载于第*13*届语义评测国际研讨会论文集，第691–695页。

Cesc Chunseong Park、Byeongchang Kim 和 Gunhee Kim。2019年。通过多模态记忆网络实现个性化图像描述。*IEEE* 模式分析与机器智能汇刊，41(4)：999–1012。

Rashmi Prasad、Nikhil Dinesh、Alan Lee、Eleni Miltsakaki、Livio Robaldo、Aravind K. Joshi 和 Bonnie L. Webber。2008年。宾州篇章树库2.0。载于语言资源与评测国际会议论文集，*LREC 2008*，*2008*年*5*月*26*日至*6*月*1*日，马拉喀什，摩洛哥。欧洲语言资源协会。

Lianhui Qin、Zhisong Zhang 和 Hai Zhao。2016年。一种用于隐式话语关系分类的堆叠门控神经架构。载于*2016*年自然语言处理经验方法会议论文集，第2263–2270页，得克萨斯州奥斯汀。计算语言学协会。

Keith Rayner。2009年。第35届弗雷德里克·巴特利特爵士讲座：阅读中的眼动与注意、场景知觉和视觉搜索。实验心理学季刊，62(8):1457–1506。

Steven J Rennie、Etienne Marcheret、Youssef Mroueh、Jerret Ross 和 Vaibhava Goel。2017年。用于图像描述的自批评序列训练。载于*IEEE*计算机视觉与模式识别会议论文集，第7008–7024页。

Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein 等。2015年。ImageNet大规模视觉识别挑战赛。国际计算机视觉杂志，115(3):211–252。

Attapol Rutherford 和 Nianwen Xue。2016年。面向英语和中文的稳健非显式神经篇章分析器。载于*CoNLL-16*共享任务论文集，第55–59页，柏林，德国。计算语言学协会。

Dinghan Shen、Xinyuan Zhang、Ricardo Henao 和 Lawrence Carin。2018年。通过细粒度词语对齐改进语义感知网络嵌入。载于《*2018*年自然语言处理实证方法会议论文集》，比利时布鲁塞尔，*2018*年*10*月*31*日—*11*月*4*日，第1829–1838页。计算语言学协会。

Zhan Shi、Hui Liu 和 Xiaodan Zhu。2021年。利用自然语言推理增强描述性图像字幕生成。载于《计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*2*卷：短篇论文）》，第269–277页，线上。计算语言学协会。

Qingyi Si、Zheng Lin、Ming yu Zheng、Peng Fu 和 Weiping Wang。2021年。再检查一次：通过视觉蕴含实现渐进式视觉问答。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第4101–4110页，线上。计算语言学协会。

Weijie Su、Xizhou Zhu、Yue Cao、Bin Li、Lewei Lu、Furu Wei 和 Jifeng Dai。2020年。VL-BERT：通用视觉语言表征的预训练。载于第*8*届国际学习表征会议，*ICLR2020*，亚的斯亚贝巴，埃塞俄比亚，*2020*年*4*月*26*日至*30*日。OpenReview.net。

Lin Sun、Jiquan Wang、Kai Zhang、Yindu Su 和 Fangsheng Weng。2021年。Rpbert：一种基于文本—图像关系传播的 BERT 多模态命名实体识别模型。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*，虚拟会议，*2021*年*2*月*2*日至*9*日，第13860–13868页。AAAI出版社。

Hao Tan 和 Mohit Bansal。2019年。LXMERT：从变换器中学习跨模态编码器表征。载于*2019* 年自然语言处理实证方法会议和第 *9* 届国际自然语言处理联合会议论文集，*EMNLP-IJCNLP 2019*，中国香港，*2019*年*11*月*3*日至*7*日，第5099–5110页。计算语言学协会。

Quoc-Tuan Truong 和 Hady W Lauw。2019年。Vistanet：用于多模态情感分析的视觉方面注意力网络。载于*AAAI* 人工智能会议论文集，第 33 卷，第 305–312 页。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。Attention Is All You Need。载于神经信息处理系统进展第 *30* 卷：*2017*年神经信息处理系统年度会议，*2017* 年 *12* 月 *4* 日至 *9* 日，美国加利福尼亚州长滩，第 5998–6008 页。

Alakananda Vempala和 Daniel Preotiuc-Pietro。2019年。对推文文本与图像之间关系进行分类和推断。载于论文集

第*57*届计算语言学协会年会论文集，*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，第2830–2840页。计算语言学协会。

Danae Sánchez Villegas 和 Nikolaos Aletras。2021年。基于文本和图像的兴趣点类型预测。载于《*2021*年自然语言处理实证方法会议论文集》（*EMNLP 2021*），线上会议／多米尼加共和国蓬塔卡纳，*2021*年*11*月*7*日至*11*日，第7785–7797页。计算语言学协会。

Yue Wang、Shafiq R. Joty、Michael R. Lyu、Irwin King、Caiming Xiong 和 Steven C. H. Hoi。2020a。VD-BERT：一种结合 BERT 的统一视觉与对话变换器。载于《*2020*年自然语言处理实证方法会议论文集》（*EMNLP 2020*），线上，*2020*年*11*月*16*日至*20*日，第3325–3338页。计算语言学协会。

Yue Wang、Jing Li、Irwin King、Michael R. Lyu 和 Shuming Shi。2019年。通过编码对话上下文生成微博话题标签。载于《*2019*年北美计算语言学协会分会：人类语言技术会议论文集》（*NAACL-HLT 2019*），明尼阿波利斯，明尼苏达州，美国，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），第1624–1633页。计算语言学协会。

Yue Wang、Jing Li、Michael R. Lyu 和 Irwin King。2020b。跨媒体关键词短语预测：一种结合多模态多头注意力和图像措辞的统一框架。载于《*2020*年自然语言处理经验方法会议论文集》（*EMNLP 2020*，线上，*2020*年*11*月*16*日至*20*日），第3311–3324页。计算语言学协会。

Yue Wang、Jing Li、Michael R Lyu 和 Irwin King。2020c。跨媒体关键词短语预测：一种结合多模态多头注意力和图像措辞的统一框架。*arXiv*预印本 *arXiv:2011.01565*。

Huijuan Xu 和 Kate Saenko。2016。提问、关注并作答：探索用于视觉问答的问题引导的空间注意力。载于《计算机视觉—*ECCV 2016*：第*14*届欧洲会议》，阿姆斯特丹，荷兰，*2016*年*10*月*11*日至*14*日，论文集第*VII*部分，《计算机科学讲义》第9911卷，第451–466页。Springer。

Jiacheng Xu、Zhe Gan、Yu Cheng 和 Jingjing Liu。2020年。面向话语的神经抽取式文本摘要。载于第*58*届计算语言学协会年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日，第5021–5031页。计算语言学协会。

Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、RuslanSalakhudinov、Rich Zemel，以及 Yoshua Bengio。2015年。《展示、关注与讲述：结合视觉注意力的神经图像描述生成》。载于国际机器学习会议，第2048–2057页。PMLR。

Yang Xu、Yu Hong、Huibin Ruan、Jianmin Yao、Min Zhang 和 Guodong Zhou。2018年。运用主动学习扩展隐式话语关系识别的训练数据。载于*2018*年自然语言处理实证方法会议论文集，比利时布鲁塞尔，*2018*年*10*月*31*日至*11*月*4*日，第725–731页。计算语言学协会。

PeterYoung、Alice Lai、Micah Hodosh 和 JuliaHock- enmaier。2014年。从图像描述到视觉指称：用于对事件描述进行语义推理的新型相似度度量。计算语言学协会会刊，2:67–78。

Jianfei Yu、Jing Jiang、Li Yang 和 Rui Xia。2020年。通过统一的多模态 Transformer，利用实体跨度检测改进多模态命名实体识别。计算语言学协会。

Tiezheng Yu、Wenliang Dai、Zihan Liu 和 Pascale Fung。2021年。用于多模态抽象摘要的视觉引导生成式预训练语言模型。载于*2021*年自然语言处理实证方法会议论文集，第3995–4007页，线上和蓬塔卡纳，多米尼加共和国。计算语言学协会。

Zhou Yu、Jun Yu、Yuhao Cui、Dacheng Tao 和 Qi Tian。2019年。用于视觉问答的深度模块化协同注意力网络。载于*IEEE*计算机视觉与模式识别会议，*CVPR 2019*，美国加利福尼亚州长滩，*2019*年*6*月*16*日至*20*日，第6281–6290页。计算机视觉基金会 / IEEE。

Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、LeiZhang、Lijuan Wang、YejinChoi 和 Jianfeng Gao。2021年。Vinvl：重新审视视觉语言模型中的视觉表征。载于*IEEE*计算机视觉与模式识别会议（*CVPR 2021*），线上，*2021*年*6*月*19*日至*25*日，第5579–5588页。计算机视觉基金会 / IEEE。

Suwei Zhang、Yuan Yao、Feng Xu、Hanghang Tong、Xiaohui Yan 和 Jian Lu。2019年。面向照片分享服务的主题标签推荐。载于第三十三届*AAAI*人工智能会议（*AAAI 2019*）、第三十一届人工智能创新应用会议（*IAAI 2019*）、第九届*AAAI*人工智能教育进展研讨会（*EAAI 2019*），美国夏威夷州檀香山，*2019*年*1*月*27*日至*2*月*1*日，第5805–5812页。AAAI出版社。

Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason J. Corso 和 Jianfeng Gao。2020年。用于图像说明生成和视觉问答的统一视觉-语言预训练。载于 第三十四届人工智能会议，*AAAI 2020*，

第三十二届人工智能创新应用会议，*IAAI 2020*，第十届 *AAAI* 人工智能教育进展研讨会，*EAAI 2020*，美国纽约州纽约，*2020*年*2*月*7*日至*12*日，第 13041–13049 页。AAAI出版社。