基于 Transformer 与 FasterRCNN 的多模态特征提取与融合

陈远露[1,2] 王 亮[1,2]
CHEN Yuanlu WANG Liang

摘 要 传统人机对话中仅使用文本单一模态存在对话中包含的信息量不够的问题，而使用文本和图片两种模态能丰富对话中的信息量，也更符合实际生活中的聊天场景。由于文本和图片两个模态的不同特点，对话模型只采用单一的传统 NLP 和 CV 领域的模型，不能同时很好地处理这两种模态。针对在多模态特征提取与融合模型上的问题，提出一种基于 Transformer 与 FasterRCNN 融合的多模态特征提取与融合模型，更好地进行两种模态的特征提取、融合，达到提高多模态对话的性能的目的。模型中，Transformer 对文本进行特征提取，FasterRCNN 对图片进行特征提取，然后通过 Late Fusion 融合技术将图片和文本两种模态的特征融合。实验结果表明，相较于传统的单一模型，提出的模型在多模态对话的几种性能评价指标上均取得了比较理想的效果。

关键词 多模态；对话系统；特征提取；特征融合；模型融合；Transformer；FasterRCNN

doi：10.3969/j.issn.1672-9528.2024.05.024

0 引言

在对话系统中同时利用多种模态如文本、图像、音频、视频等的信息进行交流和理解的方式叫作多模态对话。多模态对话系统可以提供更丰富的交互体验，拓展应用场景和领域，包括智能助理、智能客服、智能教育等。现实生活中，信息的传达不仅仅依赖于文本，还包括很多其他承载信息的载体，如音频、图像、视频和各种生理信息等。所以，仅仅依靠单一类型的数据要比多类型的数据所蕴含的信息量要少。然而研究多模态对话需要面对的一个重要问题是，纯文本对话的数据集比较容易找，但多模态对话的数据集却很少。随着网络的普及和技术水平的提升，人们在社交媒体上参与公共对话的程度大大提高了。这为对话系统的研究与发展提供了土壤。在语言和视觉领域，研究最广泛的领域之一是图像描述，即在给定输入图像的情况下输入单个语句。这通常涉及一个事实性的、描述性的句子来描述图像，而不是像在对话中那样产生一个会话语句。流行的数据集包括 COCO 和 Flickr30k。但在日常交流中并不是所有的对话都围绕着图片进行，会有内容偏离图片的对话出现。可以将表现出这种现象的对话称为稀疏的图像对话。Image-Chat[1] 图像聊天数据集是（图像、说话者 A 的风格特征、说话者 B 的风格特征、A & B 之间的对话）元组的大型集合，每个对话由说话者 A 和 B 的连续轮次组成。MMCHAT[2] 数据集则是将社交媒体上收集到的原始对话和图像精细化过滤，构建了一个大型的多模态对话数据集。现在拥有可以用来进行多模态对话系统研究的数据集，去训练计算机程序成为熟练的对话参与者。多模态对话系统的发展可以为用户提供更丰富的交互方式。用户可以更直观地表达自己的需求，提高交流效率。多模态对话系统可以拓展到各种领域，如智能家居、智能医疗等多个领域。

多模态将计算机视觉（CV）和自然语言处理（NLP）这两个不同领域结合起来，多模态对话系统弥合了视觉和语言之间的差距，整合文本、图像等不同模式的信息，可以为构建有效的端到端对话系统提供更多的细节。通过陈鑫等人 [3] 的研究汇总，可以了解在无限制主题及无明确对话目的的背景下，可基于检索或生成的方法进行人机对话的交互。同时，对话系统可以分为单轮对话和多轮对话。单轮对话主要考虑基于问题的回答，而多轮对话则更加注重上下文的整体信息，输出更加符合上下文语义的回复。Firdaus[4] 提出了一种针对多模态对话系统的方面感知响应生成方法。通过考虑对话历史和指定的方面信息，该方法能够生成与对话内容相关且语法正确的响应。实验结果表明，该方法在自动评估和人工评估中均优于基线模型，具有较高的生成质量和一致性。Yang 等人 [5] 提出了一种新颖的多轮对话生成模型，通过在编码过程中捕捉句子级别和话语级别的上下文信息，模型利用上下文的语义信息通过差异感知模块进行动态建模。此外，还设计了一个句子顺序预测训练任务，通过重新构建被打乱的句子的顺序来学习表示。实验结果表明，与基线模型相比，本文的模型在自动评估和人工评估指标上都取得了显著的改进，生成的回复更加流畅和信息丰富。Ke 等人 [6] 提出了一种基于 BERT 的人机上下文感知分层融合网络，用于多轮对

1. 沈阳化工大学计算机科学与技术学院 辽宁沈阳 110142
2. 辽宁省化工过程工业智能化技术重点实验室 辽宁沈阳 110142

话行为检测。该模型结合了上下文信息和语义表示，通过分层融合的方式提高了对话行为检测的准确性。

本文采用 Transformer 和 fastrcnn 融合模型，通过图片和文字两种模态实现多轮开放型多模态对话系统。采用 Late Fusion 融合技术将两种模态的特征进行融合。

1 Transformer 原理

Transformer 分为 Encoder 和 Decoder 两个部分。Encoder 部分负责把输入的语言序列映射成隐藏层输入 Dencoder 解码器映射成语言序列输出。Transformer 与 LSTM 最大的区别就是 LSTM 的训练是串行的，必须等当前字处理完，才能处理下一个字。而 Transformer 是并行，它是同时处理所有字，大大提高了效率。

1.1 Encoder

Encoder 主要由 Positional Encoding、Self Attention Mechanism、残差连接和 Layer Normalization 部分组成。

因为 Transformer 不是像 RNN 一样是迭代的，所以必须得给 Transformer 提供每个字的位置信息，也就是 Positional Encoding 来识别出字的顺序关系。将位置信息与字的信息进行相加而不是拼接，所以 Input Embedding 和 Positional Embedding 的纬度必须是一致的，都是 [Batch_size Sequence_length Embedding_dim]。使用了 Sin 和 Cos 函数的线性变换来提供给模型位置信息。位置嵌入函数的周期从 2π 到 10 000×2π 变化，而每一个位置在 Embedding_dimension 纬度会得到不同周期是 Sin 和 Cos 函数的取值组合，从而产生独一的纹理位置信息，使得模型学到位置之间的依赖关系和语句的时序特性。

将输入的矩阵 $\boldsymbol{X}$ 通过查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$。用查询矩阵 $\boldsymbol{Q}$ 乘以键矩阵 $\boldsymbol{K}$ 得到注意力权重矩阵，并将其进行 Softmax，使得其和为 1。再用权重矩阵乘以值矩阵 $\boldsymbol{V}$，最后将这些权重化后的值向量求和。

1.2 Decoder

Decoder 的主要结构是 Masked Multi-Head Self-Attention、Multi-Head Encoder-Decoder Attention 和 FeedForward Network 三个部分。和 Encoder 一样，每个部分通过残差连接再加上 Layer Normalization。Transformer 抛弃了 RNN 这种以时间驱动的模式改成 Self-Attemtion，导致整个 Ground Truth 暴露在 Decoder 中。因此，需要用 Mask 来解决这一问题。值得注意的是，Masked Encoder-Decoder Attention 里的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 为 Encoder 的输出，$\boldsymbol{Q}$ 为 Decoder 中 Masked Self-Attention 的输出。

2 FasterRCNN

FasterRCNN 的检测部分主要分别为四个模块：Conv Layers、RPN、RoI Pooling 和 Classification and Regression。

Conv Layers，特征提取网络，用于提取特征。通过一组 Conv+Relu+Pooling 层，共有 13 个 Conv 层，13 个 ReLU 层，4 个 Pooling 层来提取特征。根据池化公式，经过每一个 Pooling 层，特征图的大小变成之前宽高的一半。一个 $M\times N$ 大小的图片经过 Conv Layer 之后，特征图的大小为 $(M/16)\times(N/16)$。

通过预设不同大小的 Anchor，来覆盖图像上各个位置各种大小的目标。再通过后续淘汰一大批无用的 Anchor，并对预设的 Anchor 位置进行修正。$(M/16)\times(N/16)\times256$ 的特征通过 1*1 卷积得到 $(M/16)\times(N/16)\times18$ 的输出，通过 softmax 二分类判断是 Positive 目标还是 Negative 背景，Reshape 层用来纬度变换。之后再进行回归，$(M/16)\times(N/16)\times256$ 的特征通过 1*1 卷积得到 $(M/16)\times(N/16)\times36$ 的输出，生成每个 Anchor 的坐标偏移量，用来修正 Anchor。之所以预测偏移量而不是直接预测坐标，一方面是因为坐标数量级比较大，难以训练；另一方面是因为坐标偏移大小较小，且偏移求导方便。最后在 Proposal 层，通过输入 Cls 层生成的 $(M/16)\times(N/16)\times18$ 向量、reg 层生成的 $(M/16)\times(N/16)\times36$ 向量和 Im_info=[M, N,Scale_factor]，输出一堆 Proposals 左上角和右下角坐标值（$[x_1,y_1,x_2,y_2]$ 对应原图 $M\times N$ 尺度）。

RoI Pooling 层不是像 Crop 那种从图像中裁剪一部分，损失图像的完整信息，或者像 Wrap 将图像 Reszie 层需要的大小送入网络，从而破坏图像原始形状信息。RoI Pooling 会有一个预设的 Pooled_w 和 pooled_h。由于 Proposals 坐标是基于 $M\times N$ 尺度的，先映射回 $(M/16)\times(N/16)$ 尺度，再将每个 Proposal 对应的 Feature Map 区域分为 Pooled_w x Pooled_h 的网格，对网格的每一部分做 Max Pooling。这样处理后，即使大小不同的 Proposal 输出结果也是 Pooled_w x Pooled_h 固定大小，实现了固定长度输出。

Classification 分类和 RPN 中为了区分目标还是背景的二分类不同，这个分类是要对之前的所有 Positive Anchors 识别出属于哪一类。从 RoI Pooling 处获取到 Pooled_w x Pooled_h 大小的 Proposal Feature Map 后，通过全连接层和 Softmax 对所有的 Proposals 进行具体类别的分类，通常为多分类。再次对 Proposals 进行 Bounding Box Regression，获得更高精度的最终的 Predicted Box。

3 Transformer 与 FasterRCNN 融合的多模态对话模型

3.1 Transformer 与 FasterRCNN 融合的多模态对话模型的模型结构

多模态数据通常包括文本、图像、语音等不同类型的信息，这些数据在结构和表示上存在显著差异。传统的模型往往设计用于处理单一数据类型，难以有效地处理这种异构性。多模态数据之间可能存在非线性、时序、空间等多种复杂关系，传统模型可能无法很好地捕捉这些关系。传统的模型在整合来自不同模态的信息时可能丧失部分模态的特征，导致模型性能下降。对于一些复杂的多模态任务，例如自然语言处理和计算机视觉的结合，传统模型可能无法充分利用领域

知识，导致模型性能受限。

对于单个模型来说，单独的 Transformer 和 FasterRCNN 比较难处理文字和图片两个模态的数据，本文采用将两个模型进行融合的方式，让其处理各自适合的模态。可以实现更全面和准确的数据分析和处理，而且模型的融合可以是对于捕捉到那些在单个模型中难以发现的模式或规律。相比于 CNN 来说，Transformer 的多模态融合能力强 [7]。CNN 擅长使用卷积核来获取图像中的信息，但无法很好地处理声音、文字、时间这种非图片的模态。而 Transformer 的输入则不是像 CNN 一样必须保持二维图像，通常可以直接将其他模态的信息转换为向量就可以在输入端将这些特征融合。Transformer 模型在处理自然语言处理任务方面表现出色，它可以有效地捕获文本序列中的长期依赖关系，并且在处理大规模数据时具有良好的扩展性。FasterRCNN 是一种流行的目标检测算法，能够有效地检测图像中的物体并进行定位。它通过卷积神经网络提取图像特征，并结合区域建议网络（region proposal network）来实现高效的目标检测。

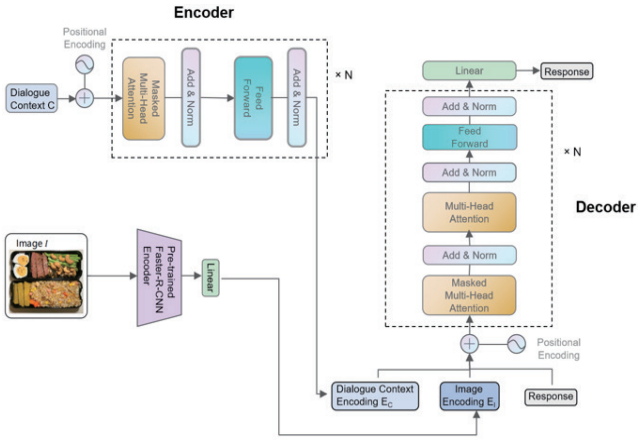

图 1 Transformer 与 FasterRCNN 融合的多模态对话模型的结构

3.2 Transformer 与 FasterRCNN 融合的多模态对话模型的特征融合

按融合层次来划分数据融合技术，主要可以分为三类：原始数据级融合、特征级数据融合和决策级数据融合 [8]。原始数据级融合是指直接融合底层数据，其中典型的方法如卡尔曼滤波。其优点是不存在数据丢失的问题，计算的结果较为准确，能够提供其他融合所没有的很细微的信息；缺点是计算量太大，处理的时间太长，从而实时性差。特征级融合属于中间层次融合，先从各个类型的数据中提取出特征向量，再把特征融合成单一的向量。在这三个融合层次中，特征级融合发展较为完善，对计算量和通信带的要求相对较低，可以提高数据的抽象层次、减少数据的冗余度，提高系统分类和识别的准确率。但由于部分数据的舍弃，准确率有所降低。林少娃等人 [9] 通过结合多源异构数据，实现了电力故障主动性预警。首先，针对交互式诉求文本进行主题挖掘，使用 LDA 模型对文本进行特征提取和主题分类。然后，将交互式文本的主题热点标签与电力公司收集到的多源异构数据集相结合，进行数据预处理和特征提取。通过这种方式，可以更准确地对电力故障和用户诉求进行分类和预测。决策级融合是高层次融合，在已经初步确定一个实体的状态之后，再将这些信息进行融合决策。胡新荣等人 [10] 提出了一种基于 SGD 的决策级融合维度情感识别方法。通过多模态情感识别框架和决策级融合方法，在 IEMOCAP 数据集上进行训练和测试，对多模态维度情感识别展开研究。该方法在语音单模态实验中相比基线模型提升了 1.64 个百分点，对于维度情感识别具有较好的效果。

本文采用的是特征融合的方法。用 Encoder 提取对话中文本的特征向量，用 FasterRCNN 提取对话中图片的特征向量，然后将两个特征向量融合。每个模态的特征都被提取出来，并通过特征融合网络进行处理。然后，这些特征被融合在一起，形成一个统一的特征表示。

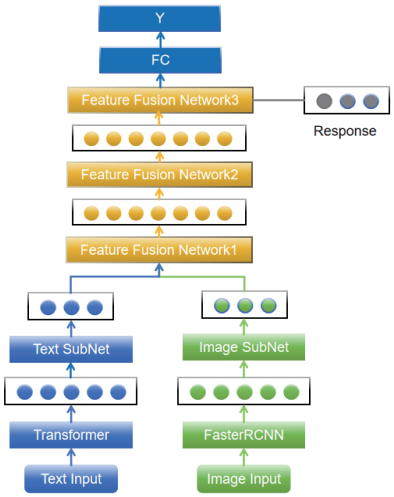

图 2 多模态融合框架

在第一级融合网络中，采用类似残差的设计思想，将 Transformer 提取的 768 维向量与融合后的 128 维向量拼接在一起。这种方法既可以避免由于网络加深而导致的过拟合，又可以更好地捕捉文本模态的原始语义空间和融合后的语义空间之间的不一致性，从而提高特征提取的准确性。第二级和第三级的融合方法与第一级类似，通过不断融合不同模态的特征来学习更深层次的语义。

其次，本研究还考虑了上下文特征对对话内容的影响。在第三级融合中，将上下文特征结合起来，进一步提高所提取特征的准确性。显然，通过最终话语的上下文可以更容易地理解对话内容。图像处理的方法与文本类似，不同之处在于图像特征提取采用了相应的图像处理方法 FasterRCNN。

4 实验的结果与分析

选择在 MMCHAT 数据集上进行训练本文的模型，用 4.0K 和 2.0K 对话会话进行测试和验证，并通过 MMCHAT 数据集提供的两个基线模型 Seq2Se 和 Seq2Seq+PIMG 来对比。

Seq2Se 建立在只有文本输入的情况下，Seq2Seq+PIMG 是一个基于图像的对话模型，使用一个单一的图像表示池来构建。

通过实验来评估 Transformer 和 FasterRcnn 的融合模型在数据集 MMCHAT 上的效果。主要通过 BLEU、Dist 和 Ent 这三个指标来评价模型的好坏。BLEU 在自然语言处理领域常用在机器翻译任务中，用于评估模型生成的句子和实际句子的差异，即用来判断两个句子的相似程度。BLEU 简单易懂，提供了一种相对简单而直观的评估方式。BLEU 根据 N-gram 可以划分成多种评价指标。N-gram 指的是连续的单词个数为 N，其公式为：

$$\begin{aligned} \mathrm{BLEU} &= B \times \mathrm{e}^{\sum_{n=1}^{N} w_n \log P_n} \\ B &= \min(1, \mathrm{e}^{\frac{1-r}{c}}) \\ P_n &= \frac{C}{N} \end{aligned} \tag{1}$$

式中：B 是一个惩罚因子，用于惩罚机器翻译结果长度过长的情况。P_n 是 N-gram 的精确匹配率，C 是生成语句中 N-gram 与参考语句中 N-gram 的最大匹配数，N 是生成语句中 N-gram 的总数。

Dist 用来评估文本的多样性，算出回复中不重复的 Ngram 数量 U 占回复中 Ngram 词语总数量 N 的比重，Dist-n 越大表示生成的多样性越高，有助于评估模型生成的多样性和创造性，是生成式任务评估的重要补充。其公式为：

$$\mathrm{Dist}-n = \frac{U}{N} \tag{2}$$

Ent 信息熵用来描述信息的混乱程度，度量样本集合纯度。信息熵越小，不确定度越高，纯度越高。其公式如下：

$$\mathrm{Ent} = -\sum_{i=1}^{n} p(x)\log p(x) \tag{3}$$

式中：$p(x)$ 表示文本中某个特定词或字符出现的概率。模型在 MMCHAT 上的对比评价结果，如表 1 所示。

表 1 模型在 MMCHAT 的测试结果对比

模型	BLEU-2	BLEU-3	BLEU-4	Dist-1	Dist-2	Ent-1	Ent-2
Seq2Seq	2.830	1.376	0.805	2.63	33.92	6.00	9.47
Seq2Seq+PIMG	2.928	1.469	0.888	2.73	34.34	6.01	9.45
Transformer	3.001	1.598	1.009	2.820	35.41	6.09	9.53
Transformer + FasterRcnn	3.142	1.832	1.287	3.132	36.53	6.23	9.66

Transformer+FasterRCNN 模型在各个评价指标上都比 Transformer 单一模型有所提升，BLEU-2 提高了 4.70%，BLEU-3 提高了 14.64%，BLEU-4 提高了 27.55%，Dist-1 提高了 11.06%，Dist-2 提高了 3.16%，Ent-1 提高了 2.30%，Ent-2 提高了 1.36%。基于以上结果，可以得出结论：Transformer+FasterRCNN 融合模型相比传统单一模型在对话中生成回复的效果更好。图片和文字的多模态特征也比单一的文字模态的效果要更好。

5 结语

针对纯文本对话信息量过于单一以及使用单一模型进行特征提取的情况，本文以 Transformer 作为一个整体框架，再融合 FasterRCNN 用来专门作为图片模块的特征提取，这使得生成回复的评价指标有所提升。后续可以研究如何增加个性化和情感表达，研究如何根据用户的特点、环境等因素，个性化地生成对话内容，以及在不同情境下如何适应性地进行多模态交互。可以考虑引入外部知识库来实现这一点，外部知识库可以帮助对话系统更好地理解对话语境，并进行推理和逻辑推断。

参考文献：

[1]KURT S, SAMUEL H, ANTOINE B, et al.Image-Chat: engaging grounded conversations[C]//58th Annual Meeting of the Association for Computational Linguistics,vol.5.Stroudsburg: Association for Computational Linguistics, 2020:2414-2429.

[2]ZHENG Y, CHEN G, LIU X, et al.MMChat: multi-modal chat dataset on social media[C]//Language Resources and Evaluation Conference,Vol. 8.Stroudsburg:Association for Computational Linguistics, 2021:5778-5786.

[3] 陈鑫 , 周强 . 开放型对话技术研究综述 [J]. 中文信息学报 , 2021, 35(11):1-12.

[4]MAUAJAMA F, NIDHI T, ASIF E.Aspect-aware response generation for multimodal dialogue system[J].ACM transactions on intelligent systems and technology, 2021, 12(2): 15.1-15.33.

[5]YANG Y, CAO J, WEN Y, et al.Multiturn dialogue generation by modeling sentence-level and discourse-level contexts[J]. Scientific reports, 2022, 12(1):20349.

[6]KE X, HU P, YANG C, et al.Human–machine multi-turn language dialogue interaction based on deep learning[J]. Micromachines, 2022, 13(3):355.

[7] 刘文婷 , 卢新明 . 基于计算机视觉的 Transformer 研究进展 [J]. 计算机工程与应用 , 2022,58(6):1-16.

[8] 金叶磊 , 古兰拜尔·吐尔洪 , 买日旦·吾守尔 . 情感分析中的多传感器数据融合研究综述 [J]. 计算机工程与应用 , 2023, 59(23): 1-14.

[9] 林少娃 , 陈奕汝 , 顾洁 , 等 . 基于隐含狄利克雷分布主题模型和特征级异构数据融合的电力故障主动性预警研究 [J]. 电子器件 , 2022, 45(2):432-438.

[10] 胡新荣 , 陈志恒 , 刘军平 , 等 . 基于 SGD 的决策级融合维度情感识别方法 [J]. 郑州大学学报 (理学版), 2022, 54(4): 49-54.

【作者简介】

陈远露（1999—），女，江苏苏州人，硕士研究生，研究方向：多模态对话。

王亮（1971—），男，辽宁沈阳人，博士，副教授，硕士生导师，研究方向：自然语言处理、推荐系统和多模态。

（收稿日期：2024-02-26）