说明书

低空物联网的多模态数据特征转换与注意力融合方法

**技术领域**

本发明涉及低空物联网与人工智能领域，具体涉及一种低空物联网多模态数据特征转换与注意力融合方法，属于多模态数据处理、特征工程及深度学习融合技术范畴。所述方法广泛应用于无人机巡检、低空安防等低空智联网场景，能够针对分布在无人机、地面监控、边缘节点等多种终端设备上的视觉、声学、文本及多类型传感器等多模态数据，进行高效的特征转换、语义对齐与动态注意力融合，为多源异构数据环境下的智能分析与决策任务提供高质量的多模态数据融合基础

**背景技术**

随着人工智能技术的快速发展，多模态数据处理在多个领域中变得越来越重要。多模态数据通常包括文本、图像、音频和视频等多种类型的信息。传统的单模态处理方法往往无法充分利用这些不同类型数据之间的互补性，因此多模态融合技术应运而生。多模态融合技术旨在通过整合来自不同模态的信息，提高模型的整体性能和鲁棒性。然而，现有的多模态融合方法存在一些挑战和局限性。首先，特征转换的问题。不同模态的数据具有不同的特征空间和维度，如何有效地将这些特征进行转换和映射到一个统一的表示空间中是一个关键问题。其次，跨模态注意力机制的引入虽然在一定程度上解决了不同模态之间的交互问题，但在实际应用中，如何设计一个高效且鲁棒的跨模态注意力机制仍然是一个难点。

目前存在的问题是不同模态的数据具有不同的特征表示，如何将这些特征进行有效的转换和融合，仍然是一个技术难题。现有的方法往往依赖于手工设计的特征转换规则，这不仅耗时耗力，而且难以适应复杂的多模态数据；跨模态注意力机制的效率较低，虽然跨模态注意力机制在多模态融合中表现出色，但其计算复杂度较高，尤其是在处理大规模数据时，计算资源的消耗成为一个瓶颈。此外，如何设计一个能够自适应不同模态之间关系的注意力机制，仍然是一个开放性问题；最后是多模态数据的稀疏性和不平衡性，比如某些模态的数据可能非常丰富，而其他模态的数据则相对稀缺。如何在数据稀疏和不平衡的情况下进行有效的多模态融合，是一个亟待解决的问题。

中国专利（申请号202310012345.6）公开了一种基于神经网络的多模态融合方法，该方法通过简单拼接各模态特征后输入神经网络，未考虑模态间的特征对齐与动态关联，在复杂场景下融合效果有限。

随着低空智联网和人工智能技术的快速发展，面向无人机巡检、低空安防、应急响应等多元化业务场景，低空空间中涌现出大量多模态数据。\*\*这些数据不仅包括传统的文本、图像、音频和视频，还融合了来自多类型传感器（如温湿度、气体、磁场等）的时空异构信息，且常常分布在无人机、地面监控、边缘节点等多种终端设备上。\*\*在实际应用中，需要对来自不同空间、不同时间、不同设备的多模态数据进行高效的协同采集与初步预处理，如归一化、去噪、时空对齐等，以保证数据质量和后续处理的一致性。

传统的单模态处理方法往往无法充分利用多源异构数据之间的互补性，难以满足低空复杂环境下对实时性、鲁棒性和高效智能决策的需求。因此，多模态融合技术应运而生，其目标是通过整合来自不同模态的信息，提高模型的整体性能和鲁棒性。然而，现有的多模态融合方法仍存在诸多挑战和局限性。具体而言，低空场景下多源异构数据在分布式采集过程中，往往因终端差异导致采集时间、空间位置和数据格式等方面存在不一致，如何实现高效的数据同步和一致性预处理成为实现有效融合的前提。同时，不同模态的数据具有各自独立的特征空间和维度，如何将这些特征进行有效转换并映射到统一的表示空间，是当前多模态融合的关键难题。尽管跨模态注意力机制在一定程度上提升了模态间的交互能力，但在低空智联网高动态、强干扰、带宽受限等复杂环境下，设计高效且鲁棒的注意力机制依然面临巨大挑战。此外，多模态数据的稀疏性和不平衡性问题也尤为突出，部分模态数据丰富而其他模态数据则稀缺或质量不佳，如何在数据不完整、分布不均的情况下实现有效融合，是亟待解决的技术难题。

目前已有的相关技术，例如中国专利（申请号202310012345.6）公开了一种基于神经网络的多模态融合方法，但该方法仅通过简单拼接各模态特征后输入神经网络，未充分考虑多模态数据在分布式环境下的采集同步、特征对齐与动态关联，因此在低空复杂场景下融合效果有限。

**发明内容**

本发明旨在解决当前多模态数据融合过程中普遍存在的特征维度灾难和异构特征融合难题，从而大幅提升多模态数据融合的效率和效果。具体而言，本发明通过有效降低特征维度，显著缓解了高维特征带来的计算压力和过拟合风险；同时，统一异构特征的语义表达，使得不同模态（如视觉、声学、文本等）的特征能够在同一语义空间内实现高效融合。该方法不仅提升了跨模态信息交互的灵活性和鲁棒性，还能更充分地捕捉不同模态之间的相关性，从而显著提升多模态融合模型的整体性能与实用价值，尤其适用于大规模、多源异构数据场景下的应用需求。

本发明的目的通过下述技术方案实现：

一种低空物联网的多模态数据特征转换与注意力融合方法，其方法包括：

S1、特征转换与统一语义空间构建

S11、数据特征收集与整理, 针对低空智联网应用场景，分别在无人机、地面监控、边缘节点等终端采集视觉（如摄像头图像）、声学（环境音、语音）、文本（任务指令、日志）等多模态异构数据的原始连续特征。对每个模态进行源数据格式和特征的转换，每个模态的数据通过特定的分词层（tokenization layer）转换为嵌入向量，然后输入到Transformer编码器中，Transformer编码器通过多头注意力机制（MHA）和前馈神经网络（MLP）提取模态特定的特征表示，保证输入数据质量。

S12、K-means聚类建模与肘部法则筛选，对视觉模态的特征，采用K-means聚类算法对特征空间进行聚类，对声学模态的特征，同样采用K-means聚类。利用肘部法则自动确定每个模态的最佳聚类中心数（如视觉 K=512，声学 M=256），以确保聚类既不过于稀疏也不过于密集。

S13、特征离散化与索引映射，将每个原始高维特征向量，通过计算与各聚类中心的距离，找到最近的聚类中心。用该聚类中心的索引替换原始特征，实现特征的离散化和降维。例如，视觉特征由2048维降至9维（log2(512)=9）。将每个模态的离散特征索引统一构建为离散语义词汇表，作为后续多模态融合的输入，这种方式在显著降低特征维度的同时，保证了特征的可区分性和表达能力，减少计算复杂度。

S2、三模态特征统一与嵌入

S21、针对文本模态的特征处理，对原始文本数据，采用分词、去除停用词等基础处理。通过预训练文本嵌入模型（如Word2Vec/BERT等）将文本转换为统一维度（如128维）的向量表示。

S22、视觉与声学特征嵌入将S1得到的离散化视觉、声学特征，分别输入各自的嵌入层（Embed Layer），将索引编码为统一维度（如128维）的向量，嵌入层参数可通过端到端训练优化，以适应下游任务需求。

S23、三种模态特征的对齐与归一化，对文本、视觉、声学三种特征向量进行归一化处理，确保三种模态的特征在统一语义空间中具有可比性和融合基础，最终得到每个样本对应的三组128维特征向量，为跨模态融合做准备。

S3、噪声抑制与模态置信度动态调整

S31、特征稳定性与置信度计算，对每个模态的嵌入特征，计算其分布的熵（Entropy），通过熵来反映特征的稳定性，采用公式 1−，其中D为特征维度，熵越低，特征越稳定，置信度越高。

S32、特征质量预测，利用两层全连接神经网络（FC），输入模态特征的均值，输出该模态的质量评分，通过端到端训练，使网络能自动学习判断特征的内在质量。

联合特征熵与MLP质量评分：

S33、对特征稳定性得分与特征质量评分进行加权融合，可以设置权重λ=0.6，得到最终模态置信度Cm，计算得到的置信度反映当前模态在样本中的可靠性，后续可以根据置信度对三种模态各自的权重进行调整；

S34、根据每个模态的置信度Cm，对原有模态权重Wm(其中m表示三种模态中的一种)进行动态调整，调整公式为：

置信度高的模态权重被放大，置信度低的模态权重被抑制，实现对噪声模态的自动抑制和高质量信息的增强。

S35、优化后的模态特征输出用调整后的对原始模态特征加权，得到“加权优化后的模态特征”（文本、视觉、声学）。

S4、三元组动态跨模态注意力机制设计

S41、分别对优化后的文本、视觉、声学三种特征向量，利用线性变换（全连接层）得到三组向量，Q（Query），由文本特征通过权重矩阵 WQ变换获得，K（Key），由视觉特征通过权重矩阵 Wk变换获得，V（Value），由声学特征通过权重矩阵 Wv变换获得，其中文本特征，视觉特征，声学特征：

其中，，，显然，上述线性变换是对各个模态的维度进行调整，根据模态压缩定理，，视觉特征降噪需求，声学唯独扩散规则。

S42、计算Q与K的点积，得到各模态间的相关性得分，对相关性得分进行缩放和归一化，输入Softmax层，获得注意力权重分布，表征文本与视觉之间的依赖强度。

S43、注意力加权与特征融合，用上述注意力权重对V（声学特征）进行加权求和，得到融合后的特征向量。该融合特征能够动态地关注和整合文本、视觉、声学三者之间最相关的信息，提升多模态表达能力:

其中为注意力权重，表示文本特征T的第i个位置与视觉特征V的第j个位置的关联强度，为声学特征，新增残差连接项，为哈达玛积（即对应元素相乘），为可学习的缩放系数，初设置为0.3，残差增强项的作用，保留文本与视觉的原始交互模式，防止注意力机制丢失重要细节

S44、多模态输出与下游任务适配，将融合后的特征输入至后续分类器或回归模型，用于具体任务（如情感分析，行为识别，目标监控等），并通过端到端训练优化参数。

本发明和现有技术相比，有益效果为在多模态数据融合任务中，通过构建特征转换器和动态跨模态注意力机制，首先对视觉、声学等异构数据类型转换成对应的特征数据，接着对转换完的特征采用聚类方法，映射至统一的离散语义空间，显著降低了特征维度，缓解了高维特征带来的计算压力和过拟合风险，同时，将不同模态的数据进行特征维度对齐。随后，利用三元组注意力机制对文本、视觉、声学三种模态信息进行深度融合，有效挖掘不同模态间的关联关系。在此基础上，结合噪声抑制模块，动态评估各模态的置信度，并自适应调整融合权重，抑制噪声模态对结果的干扰，增强模型的鲁棒性和适应性。整体方案实现了多模态特征的高效、统一表达与智能融合，提升了融合模型在大规模、多源异构数据场景下的性能和实用性，能够为多模态智能任务提供更准确、可靠的融合结果。

**附图说明**

图1为本发明低空物联网的多模态数据特征转换与注意力融合方法的方法流程图；

图2为实施例中的原始数据通过分词层转换为嵌入向量过程示意图；

图3为实施例中Transformer编码器通过多头注意力机制提取模态的特征示意图；

图4为实施例中的特征离散化与聚类过程示意图；

图5为实施例中的多模态特征统一与嵌入结构图；

图6为实施例中噪声抑制与模态置信度动态调整流程图；

图7为实施例中的三元组动态跨模态注意力机制结构图；

**具体实施方式**

构建低空智联网多模态数据集并进行源数据格式和特征的转换，对于视频模态，输入为3通道RGB像素的视频帧。将整个视频片段（大小为T×H×W）划分为dT/te × dH/he × dW/we个patch，每个patch包含t × h × w × 3个像素体。对每个patch应用线性投影，将像素体映射为d维向量表示。使用特定于视频的可学习嵌入（如时间、水平和垂直位置嵌入）来编码patch的位置信息。对于音频模态，输入为1维的音频波形，长度为T′。将音频波形划分为dT′/t′e个段，每个段包含t′个波形幅度。对每个段应用线性投影，将波形幅度映射为d维向量表示，使用特定于音频的可学习嵌入来编码每个波形段的位置信息。对于文本模态，输入为文本序列，首先构建一个词汇表，大小为v。将每个单词映射为v维的一热向量，然后通过线性投影将一热向量映射为d维向量表示。

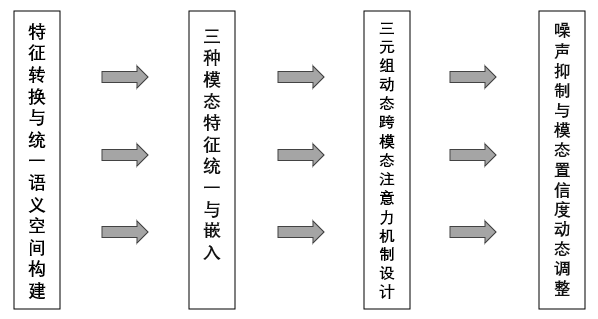


图1

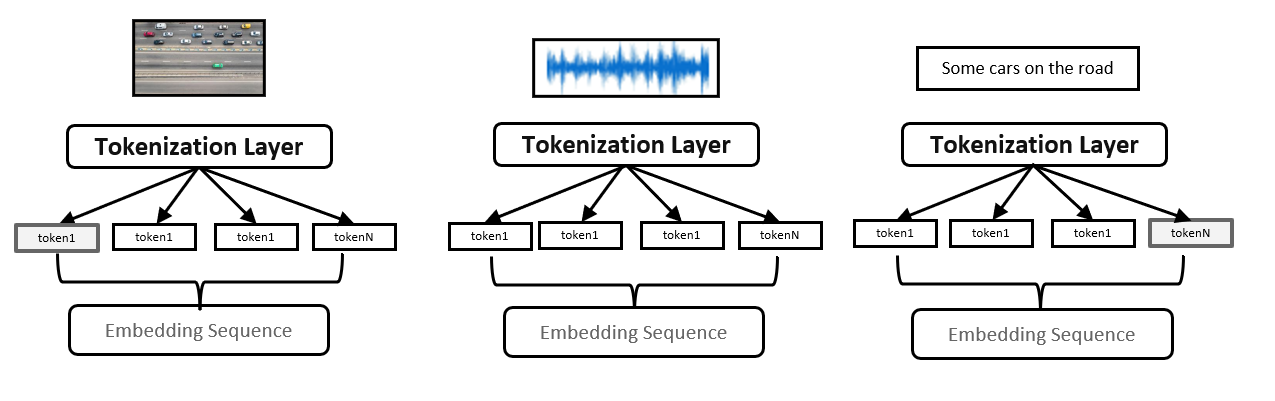


图2

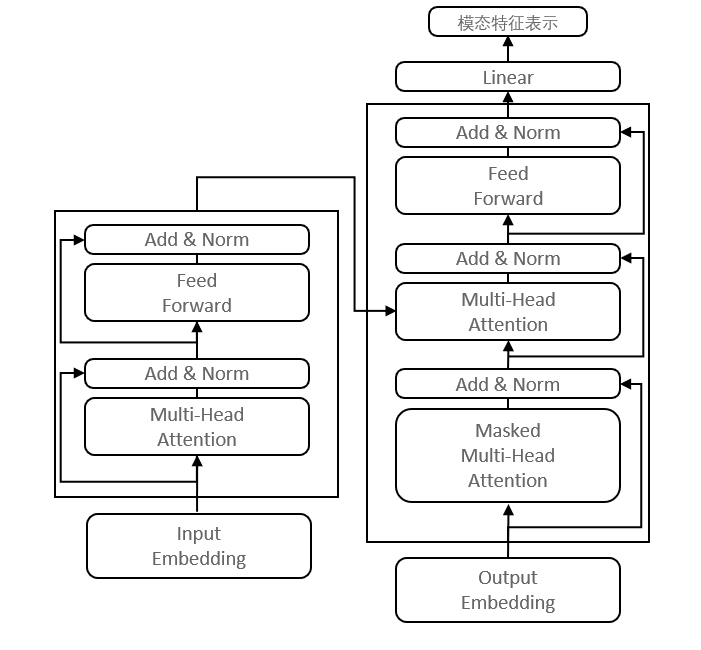


图3

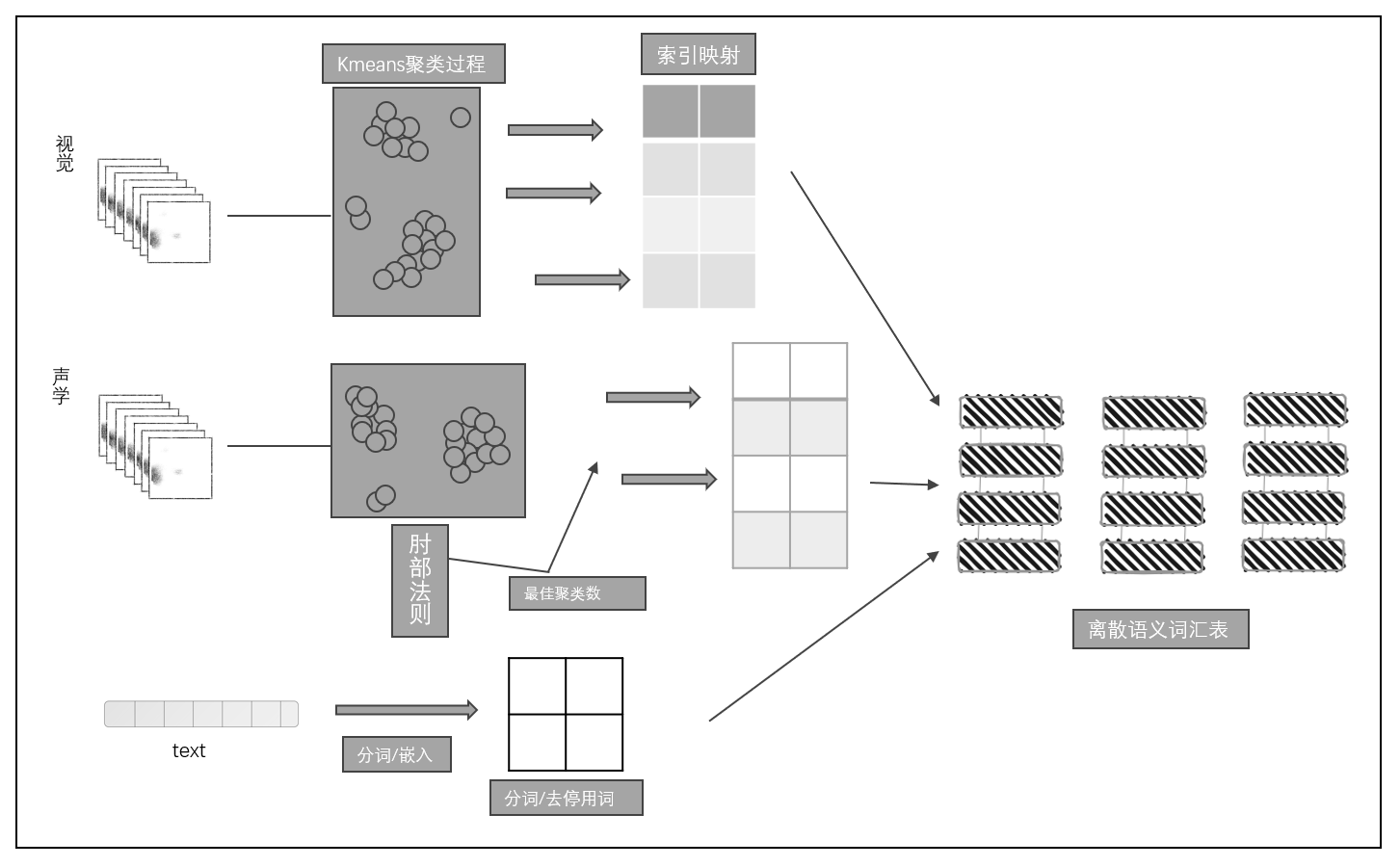


图4

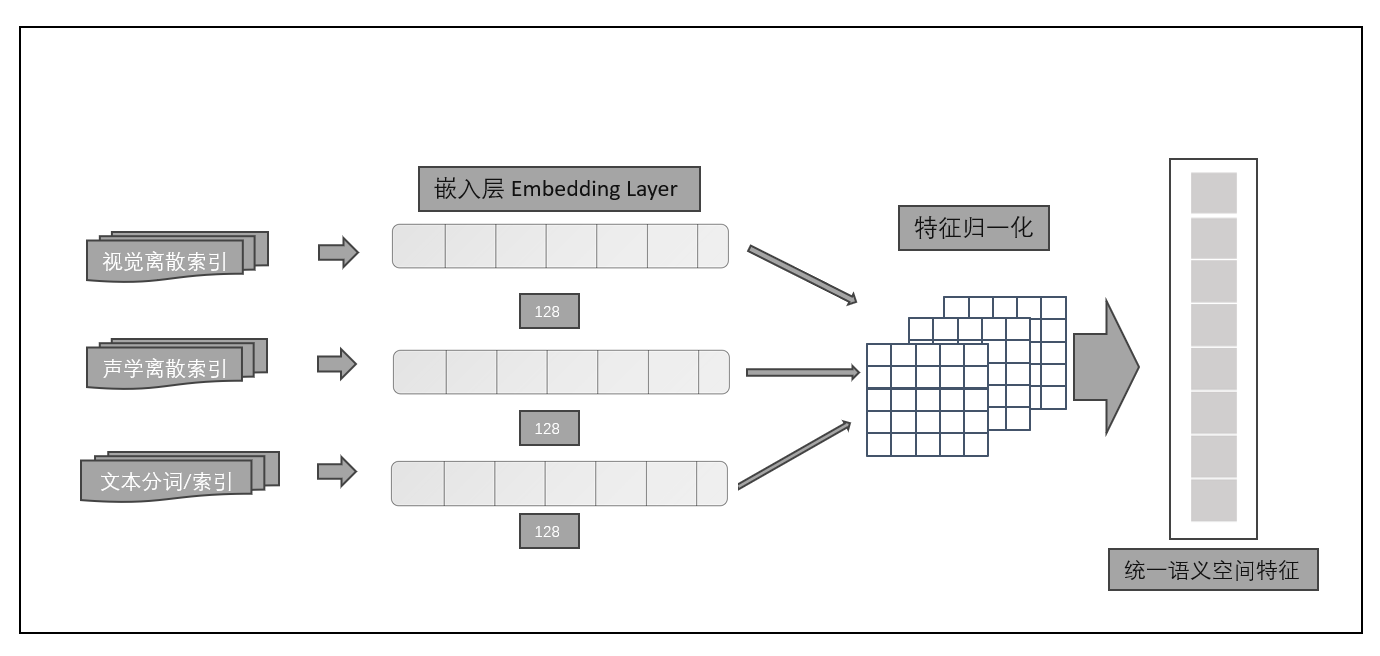


图5

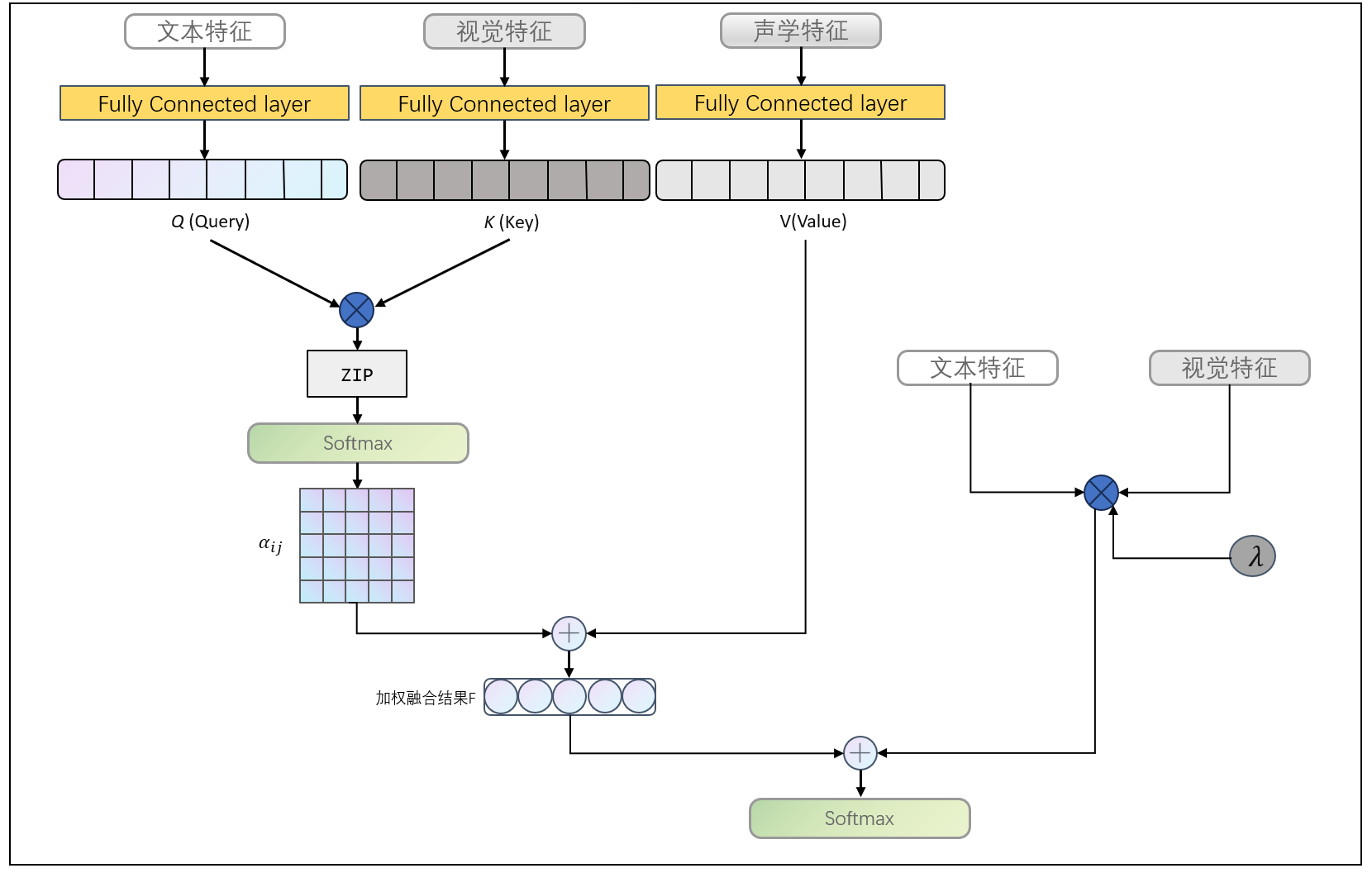


图7