计算机工程与应用
*Computer Engineering and Applications*
ISSN 1002-8331,CN 11-2127/TP

# 《计算机工程与应用》网络首发论文

题目：多模态数据融合综述
作者：任泽裕，王振超，柯尊旺，李哲，吾守尔·斯拉木
网络首发日期：2021-07-20
引用格式：任泽裕，王振超，柯尊旺，李哲，吾守尔·斯拉木．多模态数据融合综述．计算机工程与应用．
https://kns.cnki.net/kcms/detail/11.2127.tp.20210720.1125.002.html

# 多模态数据融合综述

任泽裕[1,2]，王振超[1,2]，柯尊旺[1,3]*，李哲[1,3]，吾守尔.斯拉木[1,2]
1. 新疆多语种信息技术实验室，新疆多语种信息技术研究中心，乌鲁木齐 830046
2. 新疆大学 信息科学与工程学院，乌鲁木齐 830046
3. 新疆大学 软件学院，乌鲁木齐 830046

**摘　要：**随着当今信息技术的飞速发展，信息的存在形式多种多样，来源也十分广泛。不同的存在形式或信息来源均可被称之为一种模态，由两种或两种以上模态组成的数据称之为多模态数据。多模态数据融合负责将多个模态的信息进行有效的整合，汲取不同模态的优点，完成对信息的整合。自然现象具有十分丰富的特征，单一模态很难提供某个现象的完整信息。面对保持融合后具有各个模态信息的多样性以及完整性、使各个模态的优点最大化、减少融合过程造成的信息损失等方面的融合要求，如何对各个模态的信息进行融合成为了多个领域广泛存在的一个新挑战。该文简要阐述了常见的多模态融合方法、融合架构，总结了三个常见的融合模型，简要分析协同、联合、编解码器三大架构的优缺点以及多核学习、图像模型等具体融合方法。在多模态的应用方面，对多模态视频片段检索、综合多模态信息生成内容摘要、多模态情感分析、多模态人机对话系统进行了分析与总结。该文还指出了当前多模态融合出现的问题，并提出未来的研究方向。

**关键词：**多模态；多模态融合；多模态融合架构；机器学习；神经网络
文献标志码：A　　doi：10.3778/j.issn.1002-8331.2104-0237

## A survey of Multimodal data fusion

REN Zeyu[1,2], WANG Zhenchao[1,2], KE Zunwang[1,3*], LI Zhe[1,3], WUSHOUR Silamu[1,2]

1. Xinjiang Multilingual Information Technology Laboratory, Xinjiang Multilingual Information Technology Research Center, Urumqi 830046, China

2. School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

3.School of Software, Xinjiang University, Urumqi 830046, China

**Abstract:** With the rapid development of today's information technology, information exists in various forms and sources. Different forms of existence or information sources can be referred to as one modal, and data composed of two or more modalities is called multi-modal data. Multi-modal data fusion is responsible for effectively integrating the information of multiple modalities, absorbing the advantages of different modalities, and completing the integration of information. Natural phenomena have very rich characteristics, and it is difficult for a single mode to provide complete information about a certain phenomenon. Faced with the fusion requirements of maintaining the diversity and completeness of the modal information after fusion, maximizing the advantages of each modal, and reducing the information loss caused by the fusion process, how to integrate the information of each modal has become A new challenge that exists in many fields. This paper briefly describes common multimodal fusion methods and fusion architectures, summarizes three common fusion models, and briefly analyzes the advantages and disadvantages of the three architectures of collaboration, joint, and codec, as well as

**基金项目：**国家重点研发计划资助(2017YFC0820700);多尺度十亿像素视频生成关键技术研究(2020D01C026);社会安全风险感知与防控大数据应用国家工程实验室，中国电子科学研究院资助。

**作者简介：**任泽裕(1998-)，男，硕士研究生，主研方向为自然语言处理、舆情分析；王振超(1995-)，男，硕士研究生，主研方向为自然语言处理、舆情分析；柯尊旺(1984-)，通信作者，讲师、博士，主研方向为自然语言处理、舆情分析；李哲，硕士研究生；吾守尔.斯拉木，中国工程院院士、博士生导师。

specific fusion methods such as multi-core learning and image models. In the application of multi-modality, it analyzes and summarizes multi-modal video clip retrieval, comprehensive multi-modal information generation content summary, multi-modal sentiment analysis, and multi-modal man-machine dialogue system. The paper also pointed out the current problems of multi-modal fusion and proposed future research directions.

**Key words**：multimodal; multimodal fusion; multimodal fusion architecture; machine learning; neural network

在数据领域，多模态用来表示不同形态的数据形式，或者同种形态不同的格式，一般表示文本、图片、音频、视频、混合数据[1]。多模态数据是指对于同一个描述对象，通过不同领域或视角获取到的数据，并且把描述这些数据的每一个领域或视角叫做一个模态[2]。而多模态数据融合，主要是指利用计算机进行多模态数据的综合处理[3]，负责融合各个模态的信息来执行目标预测。数据融合是一项具有挑战性的任务。首先，数据是由非常复杂的系统生成的；其次，由于数据多样性的增多，可以提出的新的可以进行研究的类型、数量以及规模都变得越来越大；第三，为使得各个数据集自身的优势得以最大程度的利用，使用异构数据集，使得缺点得到一定程度的抑制并不是一项简单的任务[4]。常见的机器学习算法等均可尝试应用于多模态数据融合中。

## 1 多模态融合分类法

关于多模态的融合方法，大致可分为模型无关的融合方法和基于模型的融合方法两大类。其中，模型无关的方法较简单但实用性低，融合过程容易产生损失；基于模型的融合方法较复杂但准确率高、实用性强，也是目前运用的主流方法。

### 1.1 模型无关的融合方法

在多模态融合的过程中，融合发生的时间是一个重要的考虑因素。针对不同的融合时期或融合水平，模型无关的融合方法共有三种，每种融合方法都有各自的特点。在不同的实验中，可以尝试使用不同的融合方法去得到更好的结果[5]。模态的一些特性，如不同的数据采集速率，对如何同步整个融合过程提出了新的挑战。以下对三种融合方法做详细的概述。表 1 对三种融合方法进行比较。

**表 1 三种模型无关融合方法性能比较**

**Table 1 Performance comparison of three model-independent fusion methods**

| | 信息损失 | 融合难度 | 容错性 | 融合水平 | 融合阶段 |
|---|---|---|---|---|---|
| 早期融合 | 中 | 难 | 差 | 低 | 提取特征后 |
| 后期融合 | 大 | 中 | 中 | 中 | 做出决策后 |
| 混合融合 | 小 | 易 | 好 | 高 | 同时 |

#### 1.1.1 早期融合

早期融合，又称为特征融合，是指对模态进行特征提取之后立刻进行的一种融合方式。特征融合的优势在于可以在早期利用来自不同模态的多个特征之间的相关性，适用于模态之间高度相关的情况。例如，在结合语音识别的音频和视频特征时采用早期融合[6]。但对于特征的提取难度较大[7]，并不是最理想的融合方法。

这种方法很难表示多模态特征之间的时间同步[8]。由于各种模态的表征、分布和密度可能有所不同，只进行简单的属性之间的连接可能会忽视各个模态独有的属性和相关性，并可能会产生数据之间的冗余和数据依赖[9]。并要求需要融合的特征在融合之前以相同的格式进行表示。随着特征数量的增加，很难获得这些特征之间的交叉相关性。图 1 所示为早期融合方法。

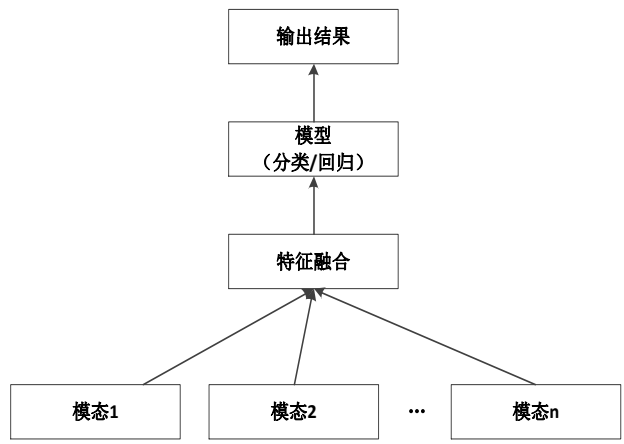



**图 1 早期融合方法**

**Fig.1 Early fusion methods**

#### 1.1.2 后期融合

后期融合，也称为决策层融合，指的是在每种模态都做出决策（分类或回归）之后才进行的融合。进行后期融合，需要使用相应的模型对不相同的模态进行训练，再对这些模型输出的结果进行融合。与之前的早期融合

作比较，该融合方式可以处理简单的数据异步性。另一个优势是允许使用最适合分析每种单一模态的方法，如音频使用隐马尔可夫模型(Hidden Markov Model，HMM)、图像使用可支持向量机(Support Vector Machines，SVM)。

但后期融合忽视了多个模态之间的低水平的相互作用，并且融合起来难度较高。由于不同的分类器需要不同的决策，学习过程变得既耗时又费力。图 2 所示为后期融合方法的结构。

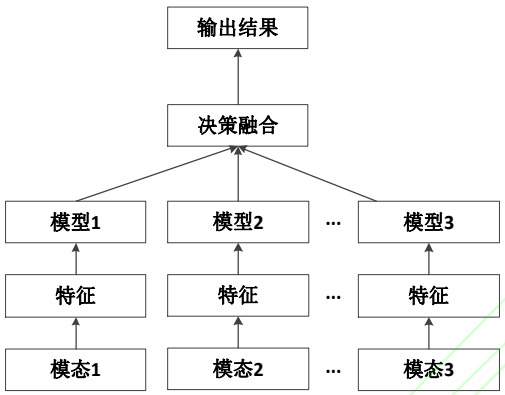



图 2 后期融合方法

Fig.2 Post fusion method

### 1.1.3 混合融合

混合融合综合了早期融合与后期融合的优点，但也使得模型的结构变得复杂并加大了训练的难度。由于深度学习模型所具有的灵活性和多样性的结构特点，比较适合使用混合融合方法。例如，倪建军、马晓萍等人[10]将混合融合方式应用于多媒体分析任务中，提出基于多重 BP(Back propagation)网络的图像融合方法，充分利用了各网络的精度。图 3 所示为混合融合方法的结构。

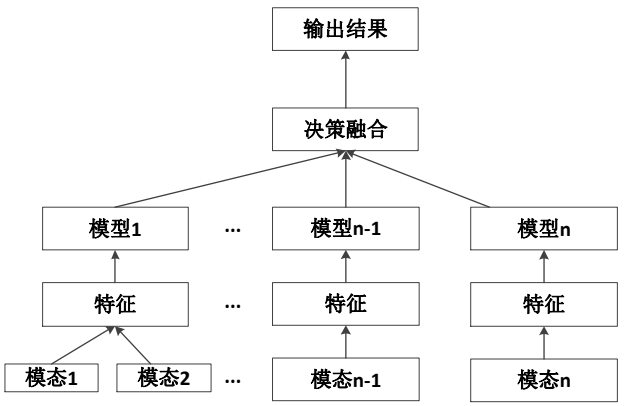



图 3 混合融合方法

Fig.3 Hybrid fusion method

综上所述，三种融合方法各有优势和劣势。早期融合可以较容易的找到各个特征之间的关系，却容易造成过拟合；后期融合可以解决过拟合问题，但不允许分类器同时训练所有数据[11]；混合融合方法较前两者方法灵活，但是需要针对具体体系结构，根据具体问题与研究内容去选择较适宜的融合方法。

## 1.2 基于模型的融合方法

基于模型的融合方法较模型无关的方法应用范围更广且效果更好，现在的研究更倾向用此类方法。常用方法包括多核学习方法、图像模型方法、神经网络方法等。

### 1.2.1 多核学习方法

多核学习(Multi-kernel Learning，MKL)方法是内核支持向量机(SVM)方法的扩展，是深度学习之前最常用的方法，其允许使用不同的核对应数据的不同视图[12-13]。由于核可以看作各数据点之间的相似函数，因此该方法能更好地融合异构数据且使用灵活[14]。McFee B，Lanckriet G 等人[15]使用 MKL 从声学、语义和艺术家的社会观三方面进行音乐艺术家相似性排序，提出的新的多内核学习(MKL)算法，它可以学习相似的空间项目来产生相似的空间，以最佳方式将所有特征空间组合到一个统一的嵌入空间中。图 4 为多核学习的过程。

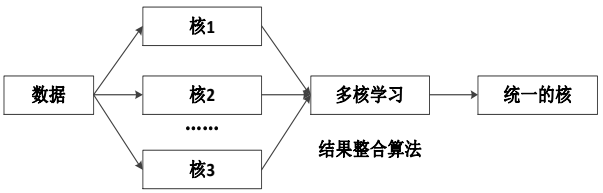



图 4 多核学习过程

Fig.4 Multi-kernel learning process

在某些应用中，可能有来自不同的模态或对应于不同实验方法的结果的不同信息源，而且每个信息源都有自己的一个或多个内核[16]。该方法的优点是核选择灵活，损失函数为凸函数(极小值即为最小值)，可使用全局最优解训练模型，提升模型性能。可以设计更好的 MKL 算法提高精度，减少复杂性和训练时间。

由于在许多应用中，人们提出许多可能的核函数，不是选其中一个而是将它们结合使用，导致在多核学习方法中存在大量的工作。较高的时间复杂度和空间复杂度是导致多核学习方法不能广泛应用的主要原因。另一个缺点是占用内存大，对训练数据有一点的依赖性。

表 2 生成模型与判别模型比较

**Table 2 Comparison of generation model and discrimination model**

| | 生成模型 | 判别模型 |
|---|---|---|
| 特点 | 寻找不同类别之间的最优分类面，反映的是异类数据之间的差异 | 对后验概率建模，从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度 |
| 区别 | 对联合分布进行建模 | 对条件分布进行建模 |
| 联系 | 由产生式模型可以得到判别式模型，但由判别式模型得不到产生式模型。 | |
| 样本数量 | 多 | 较生成模型少 |
| 准确率 | 低 | 较生成模型高 |

### 1.2.2 图像模型方法

图像模型方法也是一种常见的融合方法，主要通过对图像进行分割、拼接、预测的操作将浅层或深度图形进行融合，从而得到最终的融合结果[14]。

常见的图像模型分为生成式(联合概率)模型和判别式(条件概率)模型。许多研究中使用图像模型，尤其是在统计自然语言处理方面，集中在生成模型上，这些模型试图对输入和输出的联合概率分布进行建模[17]。早期主要使用生成模型，如动态贝叶斯网络(Dynamic Bayesian Networks)[18]、隐马尔可夫模型。后来的研究中，判别模型更受欢迎，比生成模型更简单、更容易学习。常见的判别模型如条件随机场(Conditional Random Field，CRF)[19]，对图像的组成成分进行分类标记[20]。上表 2 对生成模型和判别模型进行比较。

图像模型的优势主要是它们容易发掘数据中的空间结构和时间结构，通过将专家知识嵌入到模型中，使得模型的可解释性增强。缺点是特征之间具有复杂的依赖关系，并且模型的泛化性不强。

### 1.2.3 神经网络方法

神经网络方法是目前应用最广泛的方法之一[21]。常使用长短期记忆网络(Long Short-Term Memory，LSTM)和循环神经网络(Recurrent Neural Network，RNN)来融合多模态信息。例如利用双向长短期记忆网络进行多模态情感识别[22]；利用多模态循环神经网络(Multimodal Recurrent Neural Networks，m-RNN)，直接将图像表示和词向量以及隐向量作为多模判断的输入，在图像字幕处理等任务中表现出良好的效果[23]。

一些研究者通过模型拼凑达到了比多核学习和图像模型更好的效果。将神经网络方法应用于多模态融合中具有较强的学习能力、较好的可扩展性。缺陷是随模态数量的增加，深度学习可解释性变差，并需要依赖大量的训练数据。下表 3 对三种基于模型的融合方法进行比较。

表 3 基于模型的融合方法比较

**Table 3 Comparison of model-based fusion methods**

| 融合方法 | 工作量 | 模型可解释性 | 特征依赖性 | 应用举例 |
|---|---|---|---|---|
| 多核学习 | 较大 | 弱 | 简单 | EasyMKL[24] |
| 图像模型 | 较小 | 强 | 复杂 | 多模态数据分类[25] |
| 神经网络 | 较大 | 弱 | 简单 | 多模态情感识别[26] |

基于遗传算法（Genetic algorithm，GA）的神经网络结构优化是最早用于神经网络结构搜索和优化的元启发式搜索算法之一[27]。在 21 世纪初，一种称为增强拓扑的神经进化(NEIT)[28]的算法也使用 GAs 来进化越来越复杂的神经网络结构，受到了广泛关注。Shinozaki 和 Watanabe[29]应用 GAs 和协方差矩阵进化策略来优化 DNN 的结构，将 DNN 的结构参数化为基于有向无环图表示的简单二进制向量。由于遗传算法搜索空间可能非常大，并且搜索空间中的每个模型评估都很昂贵，所以使用大型 GPU 集群的并行搜索来加速该过程。如果设计了网络体系结构的合适表示，并且在搜索过程中训练和测试多个体系结构的成本不是非常昂贵，那么这些神经网络结构搜索和优化技术可以容易地扩展到多模态设置[30]。

贝叶斯优化(Bayesian optimization，BO)[31]是超参数优化的一种流行选择，常被用于多模态融合优化[32]。

## 2 背景知识

## 2.1 单一模态表示

### 2.1.1 图片特征提取

Dalal 等人提出的方向梯度直方图(Histogram of oriented gradient，HOG)特征提取算法[33]主要是通过计算图像局部区域梯度，并将每个局部区域中各像素点梯度的方向直方图级联。HOG 特征提取算法的基本流程图如图 5 所示：

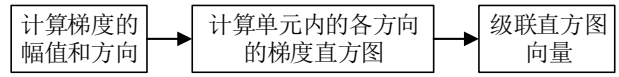



**图 5 HOG 特征提取算法的基本流程图**

**Fig.5 The basic flow chart of HOG feature extraction algorithm**

具体步骤如下[34]：

(1)对图像进行灰度化操作和 γ 标准化处理。

(2)用中心对称算子 k=[-1,0,1]及转置计算横纵坐标的方向梯度

(3)将图片分割为多个小方块，并且每个小方块由 4 个单元所组成，每个单元由 8×8 像素组成。方块的滑动步长为 1 个单元。$\theta(x,y)$把$[0,\pi]$分为 9 个小区间。单元中的每一个像素点都为直方图通道进行权重为$g(x,y)$的加权投票，从而得到每个单元内 9 个方向的梯度直方图。

(4)按照顺序级联 9 个单元的梯度直方图，得到图像的 HOG 特征$\mu_{HOG}$。

### 2.1.2 文本特征提取

林敏鸿、蒙祖强[35]采用双向门控循环网络(Bidirectional Gate Recurrent Unit，Bi-GRU)构建文本特征提取网络，并对 Bi-GRU 层的输出进行加权以突出关键部分，从而获得更精确的文本特征表达。该网络结构如图 6 所示。

在 Bi-GRU 神经网络中，将词向量$\{\omega_{i1},\omega_{i2},\cdots,\omega_{it}\}$按正向输入方式得到相应的前向隐藏层输出$\{\overrightarrow{h_{i1}},\overrightarrow{h_{i2}},\cdots,\overrightarrow{h_{it}}\}$。$\overrightarrow{h_{it}}$的计算如下式(1)：

$$\overrightarrow{h_{it}}=GRU(\overrightarrow{h_{i(t-1)}},\omega_{it}) \tag{1}$$

同理，得到相应的后向隐藏层输出$\{\overleftarrow{h_{i1}},\overleftarrow{h_{i2}},\cdots,\overleftarrow{h_{it}}\}$。将$\overrightarrow{h_{it}}$和$\overleftarrow{h_{it}}$拼接得到第 $t$ 个单词上下文信息的表示如式(2)：

$$h_{it}=[\overrightarrow{h_{it}},\overleftarrow{h_{it}}] \tag{2}$$

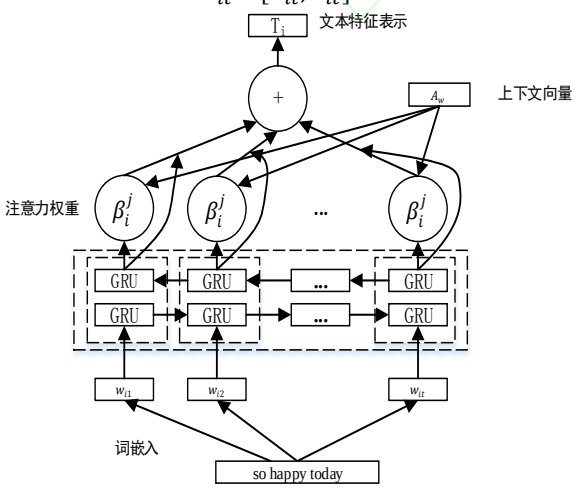



**图 6 文本特征提取网络结构**

**Fig.6 Text feature extraction network structure**

将$h_{it}$输入一层隐藏层，用 tanh 激活得到$y_{it}$，接而得到标准注意力权重。如式(3)~(4)。

$$y_{it}=tanh(W_0\cdot h_{it}) \tag{3}$$

$$\beta_{it}=\frac{exp(y_{it}^T A_\omega)}{\sum_t exp(y_{it}^T A_\omega)} \tag{4}$$

文本特征表示$T_i$，计算公式如式(5)：

$$T_i=\sum_t \beta_{it}h_{it} \tag{5}$$

### 2.1.3 语音特征提取

语音特征提取是以帧为单位进行提取的。在语音特征提取任务中，一般采用 openSMILE 工具[36]对语音数据进行特征提取。还有一种就是采用梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient，MFCC)，基于人类听觉感知(不能感知超过 1KHZ 的频率)进行特征提取[37]。MFCC 的整个过程如图 7 所示：

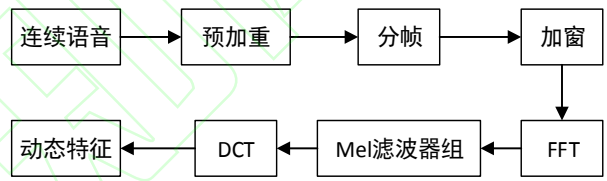



**图 7 MFCC 特征提取过程**

**Fig.7 MFCC feature extraction process**

在 Mel 滤波器组处理阶段，通过式(6)将普通频率转化到 Mel 频率：

$$mel(f)=2595*log_{10}(1+f/700) \tag{6}$$

在动态特征阶段，需要增加与倒频谱特征随时间变化相关的特征。从时间样本 $t_1$ 到时间样本 $t_2$ 的窗口中的信号 X 在帧中的能量由下式(7)表示：

$$Energy=\sum X^2[t] \tag{7}$$

分帧提取的特征信息只反应了本帧语音的特性，为了使特征更能体现时域连续性，可以在特征维度增加前后帧信息的维度。常采用一阶差分和二阶差分。$c(t)$表示第 t 帧的 MFCC 特征。一阶差分的计算方法如式(8)所示：

$$d(t)=\frac{c(1+t)-c(t-1)}{2} \tag{8}$$

## 2.2 多模态信息表示

利用多模态信息进行自然语言处理，要明确语音信息、文本信息和视觉模态信息如何进行融合。尤其是利用同源多模态信息或异源多模态信息时的语义融合范式是否相同。

根据具体融合操作不同，可以大致的划分为三种主要的方法：基于拼接和线性组合等简单融合操作的方法、基于注意力机制的融合方法和基于双线性池化的融合方法。这三种方法均是通过对特征向量进行相关操作

达到多模态信息的融合及表达。

### 2.2.1 简单融合操作的方法

深度学习可以通过简单的操作将来自不同信息源的向量化特征进行融合，如连接或加权求和。这些操作通常有很少或没有关联参数，因为深度模型的联合训练可以调整高层的特征提取层次以适应相应的操作。方法如下：

(1) 连接可以组合低级输入特征[38-40]或由预先训练的模型[41-42]提取的高级特征。

(2) 对于具有权重的加权求和，利用一种迭代方法实现，该方法要求预先训练的向量表示具有相同数量的元素，并按照适合元素相加的顺序排列[43]。这可以通过训练一个全连接层来进行维度控制并为每个模态重新排序来实现。

研究表明[44]可以利用渐进探索的神经结构搜索[45]来寻找一些融合功能的合适设置。并且每个融合功能都可以根据需要融合的层以及使用连接或加权和作为融合操作进行配置。

### 2.2.2 基于注意力机制的方法

目前注意力机制被广泛用于融合操作。注意力机制指的是由小型“注意力”模型在每个时间步长动态生成的一组标量权重向量的加权和[46-47]。通常使用多个输出来生成多组动态权重以进行求和。这组注意力的多个输出可以动态产生求和时要用到的权重，因此最终在拼接时候可以保存额外的权重信息。在将注意力机制应用于图像时，对不同区域的图像特征向量进行不同的加权，得到一个最终整体的图像向量。

**(1) 图注意力机制**

将用于文本问题处理的 LSTM 模型进行扩展，得到了一个以 LSTM 隐藏状态为条件的图像注意力模型，该模型的输入是当前嵌入单词和参与的图像特征的拼接[48]。最终利用 LSTM 的隐藏状态进行多模态融合的表征，进而可以被应用于视觉问答任务之中。这种基于 RNN 的编码-解码器模型的注意力模型可以用来帮助图像字幕问题分配注意力权重[49]，并且可以通过文本查询来找到图像对应的位置。堆叠注意力网络(Stacked Attention Networks, SANs)同样也可以使用多层注意力模型对图像进行多次查询，逐步推断出答案，模拟多步骤的推理过程[50]。在每一层中，通过将前一层根据图像特征和文本特征生成的查询向量添加到当前注意力模型生成的图像向量中，生成一个细化的查询向量并发送到下一层。将这一过程多次迭代，从而得到问题的答案。图 8 为视觉问答的堆叠注意力网络模型图。

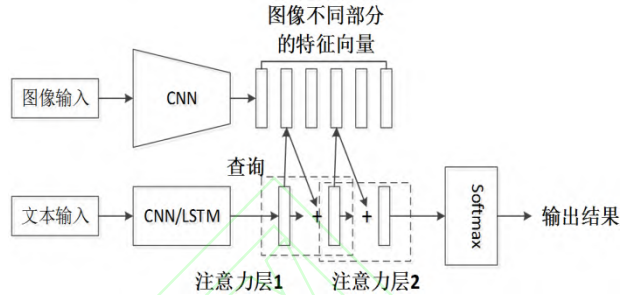



**图 8 视觉问答的堆叠注意力网络**

**Fig.8 Stacked Attention Network for Visual Question Answers**

**(2) 图和文本的对称注意力机制**

与图像注意力机制不同。共同注意力机制使用对称的注意力结构来生成注意力图像特征向量和注意力语言向量[51]。平行共注意力机制是利用一种联合表征方法，推导出图像和语言的注意力分布。交替共注意力机制则具有级联结构，首先利用语言特征生成含有注意力的图像向量，然后利用含有注意力的图像向量生成出含注意力的语言向量。平行注意力机制和交替注意力机制模型图如图 9、图 10 所示。

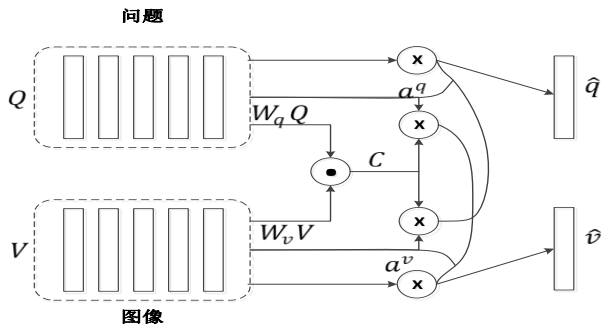



**图 9 平行共注意力机制**

**Fig.9 Parallel co-attention mechanism**

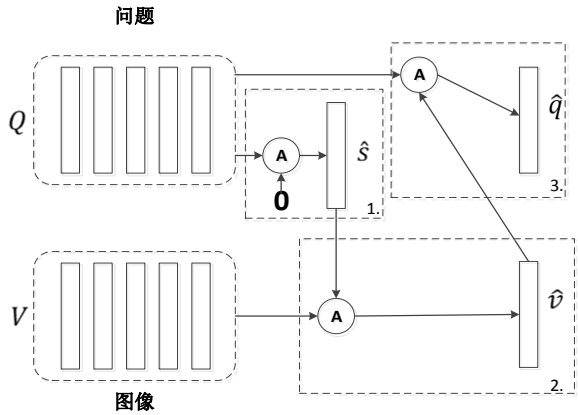



**图 10 交替共注意力机制**

**Fig.10 Alternating co-attention mechanism**

与平行共注意力网络类似，双重注意力网络(Dual Attention Network，DAN)同时估计图像和语言的注意力分布，从而获得注意力特征向量[52]。这种注意力模型以特征和相关的记忆向量为条件。与共同注意力相比，这是一个关键的区别，因为使用重复的 DAN 结构，记忆向量可以在每个推理步骤中迭代更新。

为了模拟模态之间的高阶交互作用，两个数据模态之间的高阶相关性可以表示为两个特征向量的内积，并用于导出两个模态的参与特征向量[53]。

**(3) 其他类似注意力机制**

门控多模态单元是一种基于门控，为图像和文本分配注意力权重的方法[54]。该方法基于门控机制动态生成的维度特定标量权重，计算视觉特征向量和文本特征向量的加权和。

### 2.2.3 基于双线性池化的融合方法

双线性池化通过计算外积的方式将视觉特征向量与文本特征向量进行融合，从而创建联合表示空间，这种方法可以充分利用向量元素间的交互作用。这种方法也被称为二阶池化[55]。与简单地向量组合操作(假设每个特征向量为 n 维)不一样的是，简单地向量组合操作(如连接、逐位相乘和加权求和)都会生成一个 n 或 2n 维的表征向量，而双线性池化则会产生一个 $n^2$ 维的表征向量。这意味着这种方法更有表现力。

双线性池化方法同样可以与注意力机制相结合。通过双线性池化相关方法，如多模态低秩双线性池，可以将融合的双模态表示作为注意力模型的输入特征，进而得到含有注意力的图像特征向量，再次使用该方法与文本特征向量融合，得到最终的联合表示[56]。

## 3 多模态深度学习模型

### 3.1 深层结构化语义模型

深度结构化语义模型[57](Deep Structured Semantic Model，DSSM)在 2013 年由微软的研究人员何晓东等人提出，是搜索领域的模型，属于后期融合。通过使用深度神经网络(DNN)把两种不同的模态数据表示为低维度的语义向量，并通过 cosine 距离计算两个语义向量之间的距离，最终训练出语义相似度模型。该模型既可以用来预测语义相似度，又可以获得某个模态的低维语义向量表达。该模型由输入层、表示层、匹配层三层结构构成，详细流程图如图 11 所示，模型图如图 12 所示。

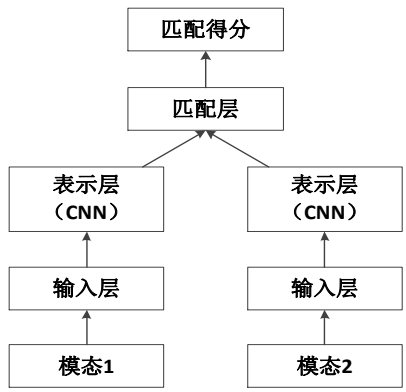



**图 11 DSSM 模型流程图**

**Fig.11 The flow chart of DSSM**

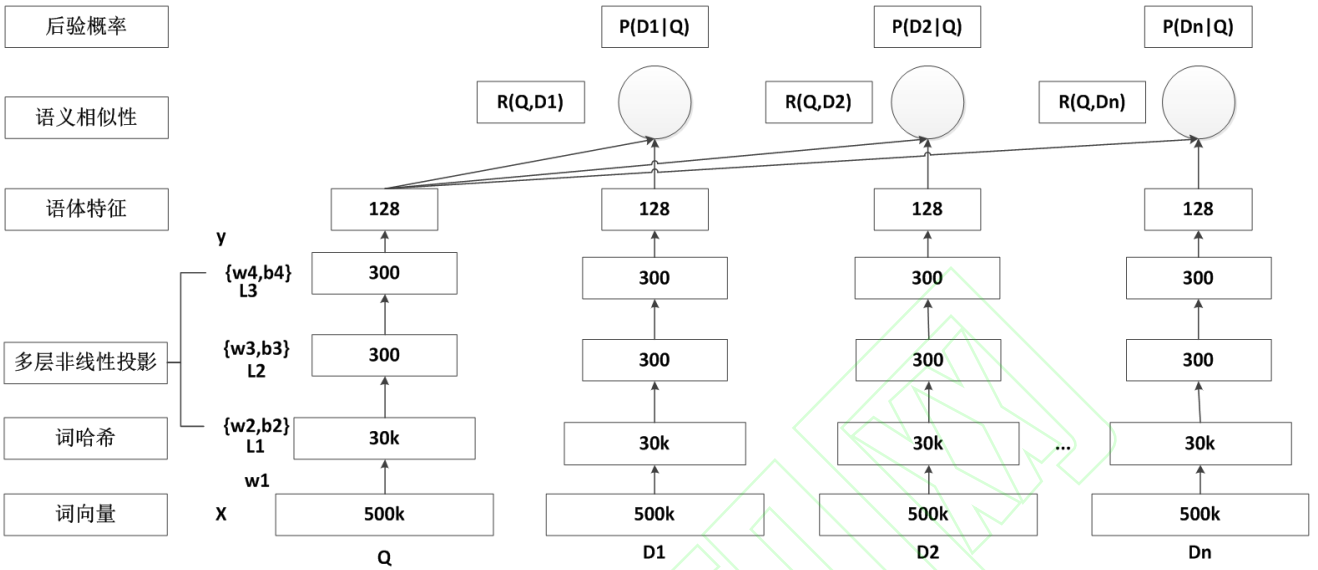



**图 12 DSSM 模型图**

**Fig.12 Illustration of the DSSM**

### 3.1.1 输入层

输入层的任务是将句子映射到一个向量空间里并将它输入到 DNN 中。

英文的输入层需要借助单词哈希表来实现，此类方法[57]旨在减少 BOW 向量的维数。它以字母 n-gram 为基础进行单词的切分，是专门为该任务开发的一种新方法。给定一个单词(如 good)，首先给该单词添加单词开始和结束标记(如#good#)，将单词分解成字母 n-grams，例如字母三元组:#go，goo，ood，od#。最后用字母 n-grams 的向量来表示这个单词。采用这种方法可以压缩空间，较为实用。

### 3.1.2 表示层

这层主要通过使用 DNN 将高维稀疏文本特征映射到语义空间中的低维密集特征，最终得到一个 128 维的低维语义向量。

将特征向量$X$映射到对应的语义概念向量$y$，如式(9)~(11)：

$$l_1 = w_1 x \tag{9}$$

$$l_i = f(w_i l_{i-1} + b_i), i = 2, \ldots, N-1 \tag{10}$$

$$y = f(w_N l_{N-1} + b_N) \tag{11}$$

用 tanh 作为隐藏层和输出层的激活函数。

### 3.1.3 匹配层

查询和文档的语义相似性可以用两个语义向量的余弦相似度来表示。计算余弦相似度的方法如式(12)所示：

$$R(Q,D) = \text{cosine}(\mathrm{y_Q}, \mathrm{y_D}) = \frac{y_Q^T y_D}{\|y_\mathbb{Q}\| \|y_D\|} \tag{12}$$

其中$y_Q$和$y_D$分别是查询和文档的概念向量。给定查询，文档按照它们的语义相关性分数排序。

通过 softmax 函数，根据文档之间的语义相关性得分，式(13)用来计算给定查询的文档的后验概率：

$$p(D|Q) = \frac{\exp(\gamma R(Q,D))}{\Sigma_{D' \in D} \exp(\gamma R(Q,D'))} \tag{13}$$

其中$\gamma$是 softmax 函数中的平滑因子。$D$为要排序的候选文档集，在理想条件下包含所有可能的文档。

此模型[57]的主要贡献是对之前提出的潜在语义模型在三方面进行了重大拓展。第一，通过直接针对文档排名的目标来优化所有版本模型的参数；其次，受最近在语音识别方面非常成功的深度学习框架的启发，使用多个隐藏表示层将线性语义模型扩展到它们对应的非线性模型。所采用的深层架构进一步增强了建模能力，从而可以捕获和表示查询和文档中更复杂的语义结构；第三，使用了一种基于字母 n-gram 的单词散列技术，这种技术被证明有助于扩大深度模型的训练，从而可以在实际的网络搜索中使用大量的词汇。DSSM 对文档排序任务的性能提升较为显著。

在大规模的真实世界数据集(验证数据集)上对该模型进行评估，评估的所有排名模型的表现均通过 NDCG[58]进行比较，表 4 中的结果表明，深度结构化语义模型表现最佳，以显著的优势击败了其他方法。其中，表格后四项为 DSSM 在不同环境中的结果。

**表 4 DSSM 与其他模型以及在不同环境下的比较结果**

**Table4 Comparative results with other models and in different environments of DSSM**

| 模型 | NDCG@1 | NDCG@3 | NDCG@9 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| TF-IDF | 0.319 | 0.382 | 0.462 |
| BM25 | 0.308 | 0.373 | 0.455 |
| WTM[59] | 0.332 | 0.400 | 0.478 |
| DPM[60] | 0.329 | 0.401 | 0.479 |
| DNN | 0.342 | 0.410 | 0.486 |
| L-WH linear | 0.357 | 0.422 | 0.495 |
| L-WH non-linear | 0.357 | 0.421 | 0.494 |
| **L-WH DNN** | **0.362** | **0.425** | **0.498** |

## 3.2 记忆融合网络

对于多模态序列学习而言，模态往往存在两种形式的交互：模态内关联与模态间关联。Amir Zadeh、Paul Pu Liang 等人提出的记忆融合网络模型(Memory Fusion Network，MFN)[61]用来处理多模态序列建模，对模态内与模态间进行不同的处理。

记忆融合网络由三部分组成，分别是：长短期记忆系统、增量记忆注意力网络和多模态门控存储器。模型图(Zadeh A, Liang P P, Mazumder N 等 2018)如图 13 所示。

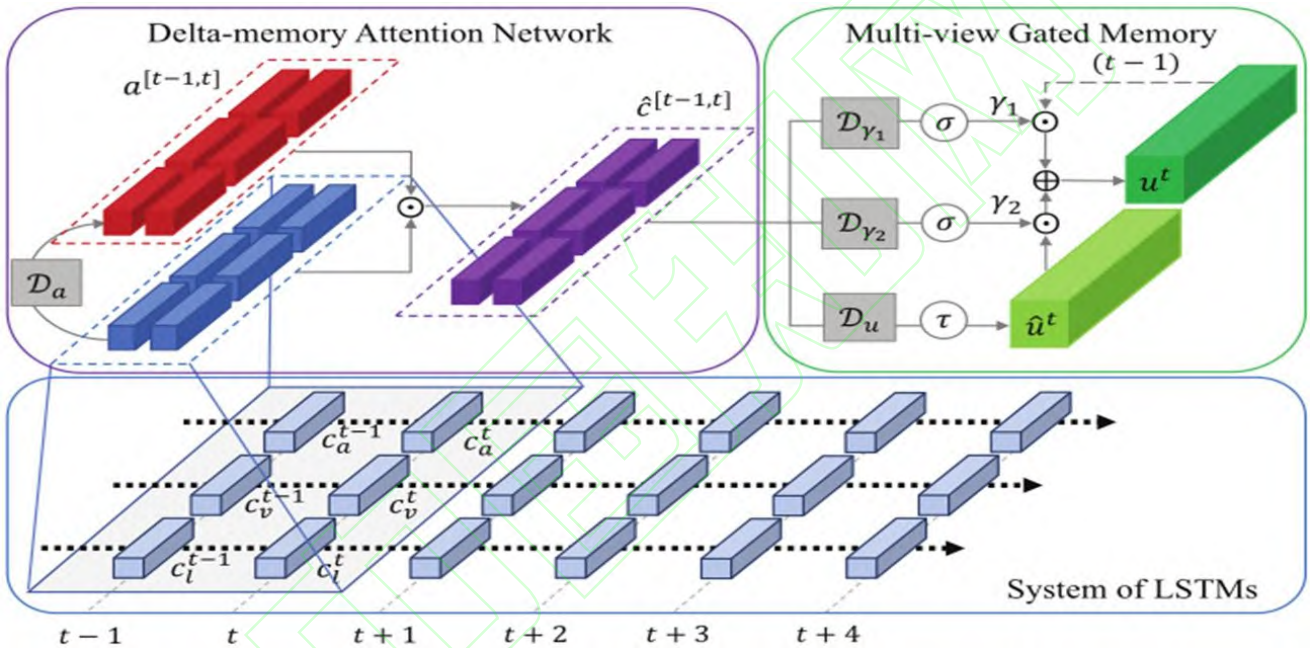



**图 13 记忆融合网络模型图**[61]

**Fig.13 Model diagram of memory fusion network**

图中，$\sigma$代表 sigmoid 激活函数，$\tau$代表 tanh 激活函数，⊙代表哈达玛积，⊕代表元素加法。每个 LSTM 从一个方面对信息进行编码，如语言。记忆融合网络输入的是一个多模态序列，其中包含 $N$ 个 $T$ 维的模态。

### 3.2.1 长短期记忆系统 LSTMs

在每一个模态序列中，一个 LSTM 随着时间对特定模态的交互进行编码。在每个时间点，各个模态的信息被输入到特定的 LSTM 中。对于第 n 个模态，$c_n$表示分配给该模态的 LSTM 的内存，并用$h_n$表示各个 LSTM 的输出，其中$d_{c_n}$为 LSTM 内存$c_n$的维度。不同序列的输入、内存和输出的规模有所不同。

式(14)~(19)为 LSTM 定义的更新规则(Hochreiter，Schmidhuber 1997)：

$$i_n^t = \sigma(W_n^i x_n^t + u_n^i h_n^{t-1} + b_n^i) \tag{14}$$

$$f_n^t = \sigma(W_n^f X_n^t + U_n^f h_n^{t-1} + b_n^f) \tag{15}$$

$$o_n^t = \sigma(W_n^o X_n^t + U_n^o h_n^{t-1} + b_n^o) \tag{16}$$

$$m_n^t = W_n^m x_n^t + U_n^m h_n^{t-1} + b_n^m \tag{17}$$

$$c_n^t = f_n^t \odot c_n^{t-1} + i_n^t \odot m_n^t \tag{18}$$

$$h_n^t = o_n^t \odot tanh(c_n^t) \tag{19}$$

$i_n$、$f_n$、$o_n$分别表示第 n 个 LSTM 的输入门、遗忘门和输出门，$m_n$为第 n 个 LSTM 在时间 t 下的内存更新。⊙代表哈达玛积，即元素乘积；$\sigma$为 sigmoid 激活函数。

### 3.2.2 增量记忆注意力网络 DMAN

此部分的任务是描绘长短期记忆系统中各个模态的内存在时间点 t 上的跨模态交互。使用系数分配技术来计算 t 时刻的 LSTM 的内存$c^t$。DMAN 的输入是时刻 t−1 和时刻 t 的内存串联，用$c^{[t-1,t]}$表示。这些内存需要被传递给神经网络$D_a: \mathbb{R}^{2d_c} \mapsto \mathbb{R}^{2d_c}$，用$d_c = \Sigma_n d_{c_n}$

来计算注意力系数。式(20)为 softmax 激活分数：

$$a^{[t-1,t]} = D_a\left(c^{[t-1,t]}\right) \tag{20}$$

其中，$a^{[t-1,t]}$是每个 LSTM 在时间$[t-1,t]$内的 softmax 激活分数。在$D_a$的输出层使用 softmax 来调整$c^{[t-1,t]}$的高值系数。$\hat{c}$的计算公式如式(21)：

$$\hat{c}^{[t-1,t]} = c^{[t\cdot 1,t]} \odot a^{[t-1,t]} \tag{21}$$

其中$\hat{c}^{[t-1,t]}$是长短期记忆系统的内存，利用这种逐个元素相乘的方法放大了$c^{[t-1,t]}$的维数并忽略了其余维数的影响。

### 3.2.3 多模态门控存储器

上一层的输出值直接传入该组件，用来标识长短期记忆系统的内存中哪些维度构成了跨模态交互。并将$\hat{c}^{[t-1,t]}$输入神经网络$D_u$: $\mathbb{R}^{2\times d_c} \mapsto \mathbb{R}^{d_{mem}}$来产生多模态门控存储器的跨模态更新规则$\hat{u}^t$，如式(22)所示。$d_{mem}$为多模态门控存储器的维度。

$$\hat{\mathrm{u}}^{\mathrm{t}} = \mathrm{D}_{\mathrm{u}}\left(\hat{\mathrm{c}}^{[\mathrm{t}-1,\mathrm{t}]}\right) \tag{22}$$

这个更新公式是在对 t 时刻跨模态交互的观察的基础上对多模态门控存储器进行修改的。

多模态储器分别由两组门电路构成，分别为维持门$\gamma_1$和更新门$\gamma_2$，并分别由不同的神经网络控制。$\gamma_1$负责记录多模态门控存储器当前有多少种状态；$\gamma_2$负责基于跨模态更新规则$\hat{u}^t$对多模态门控的内存进行更新。使用$\hat{c}^{[t-1,t]}$作为输入的多视角门控存储器门控机制的$\mathrm{D}_{\gamma_1}, \mathrm{D}_{\gamma_2}: \mathbb{R}^{2\times \mathrm{d_c}} \mapsto \mathbb{R}^{\mathrm{d_{mem}}}$的控制部分，式(23)为$\gamma_1^t$的计算公式：

$$\gamma_1^t = D_{\gamma_1}\left(\hat{c}^{[t-1,t]}\right) \tag{23}$$

在 MFN 递归的每一个时间点上，$u$利用维持门、更新门和当前的跨模态更新规则$\hat{u}^t$进行更新，公式(24)如下：

$$\mathrm{u}^{\mathrm{t}} = \gamma_1^{\mathrm{t}} \odot \mathrm{u}^{\mathrm{t}-1} + \gamma_2^{\mathrm{t}} \odot \tanh(\hat{\mathrm{u}}^{\mathrm{t}}) \tag{24}$$

通过用 tanh 函数来激活$\mathrm{u}^{\mathrm{t}}$，用以提高模型的稳定性。多模态门控存储器较 LSTM 存储器有两个优点：第一，多模态门控存储器具有更复杂的门控机制，两个门电路均由神经网络控制，所以性能更优。第二，多模态门控存储器的值在每次迭代中不会经历 sigmoid 激活，这样有利于加快收敛。

### 3.2.4 MFN 的输出

MFN 的输出包括多模态门控存储器的最终状态和每个长短期记忆系统的输出,计算方法如式(25)：

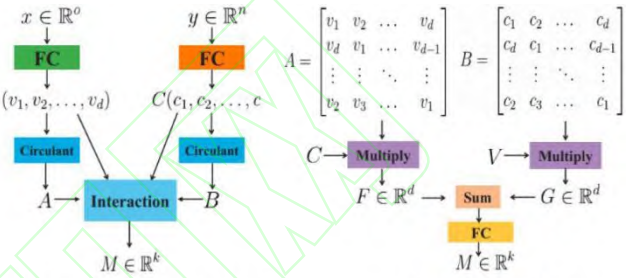
$$h^T = \oplus h_n^T \tag{25}$$

其中，$h^T$表示单个序列信息，$\oplus$表示向量的连接。

通过广泛的实验，将 MFN 与多个公开的基准数据集上提出的多模态序列学习的各种方法进行了比较。MFN 优于所有多模态方法，优于所有目前最前沿的模型。

## 3.3 多模态循环融合模型

Aming Wu 和 Yahong Han 提出的多模态循环融合模型(Multi-modal Circulant Fusion，MCF)[62]，是一种同时使用特征和矩阵的融合方法，通过此模型来发现多模态特征之间的相互作用。MCF 的模型图如图 14(a)、图 14(b)所示(Wu A, Han Y 2018)：




(a)多模套循环融合模型 (a) Multi-modal Circulant Fusion (MCF)

(b)模型交互细节图 (b) The detail of Interaction in MCF

图 14 多模态循环融合模型[62]

**Fig.14 The flowchart of Multimodal Circulant Fusion (MCF)**

给定两个不相同模态的特征向量：视觉特征$x \in \mathbb{R}^o$和文本特征$y \in \mathbb{R}^n$。以下式(26)~(27)是对投影向量的表示：

$$V = xW_1^T \tag{26}$$

$$C = yW_2^T \tag{27}$$

其中，$W_1 \in \mathbb{R}^{d\times o}$和$W_2 \in \mathbb{R}^{d\times n}$为投影矩阵，负责将两个输入要素投影到低维空间。

用投影向量$V \in \mathbb{R}^d$、$C \in \mathbb{R}^d$构造循环矩阵$A \in \mathbb{R}^{d\times d}$和$B \in \mathbb{R}^{d\times d}$,如下式(28)~(29)：

$$A = circ(V) \tag{28}$$

$$B = circ(C) \tag{29}$$

为了让投影向量和循环矩阵中的元素充分发挥作用，有以下两种不同的乘法运算：

第一种选择使用循环矩阵和投影向量相乘，如下式(30)~(31)：

$$F = CA \tag{30}$$

$$G = VB \tag{31}$$

第二种是让循环矩阵的投影向量与每个行向量作哈达玛积，如式(32)~(33)：

$$F = \frac{1}{d}\sum_{i=1}^{d} a_i \odot C \tag{32}$$

$$G = \frac{1}{d}\sum_{i=1}^{d} b_i \odot V \tag{33}$$

其中，$a_i \in \mathbb{R}^d$和$b_i \in \mathbb{R}^d$为循环矩阵 A 和 B 的行向量。

最后，通过一个投影矩阵$W_3 \in \mathbb{R}^{d\times k}$，将$F \in \mathbb{R}^d$和$G \in \mathbb{R}^d$的元素和向量转化为目标向量$M \in \mathbb{R}^k$。

在 MSVD 数据集上，将 MCF 模型与其他同类模

型进行比较，得到表 5 所示结果。

表 5 与其他模型在 MSVD 数据集上比较

Table 5 Comparison with other models on MSVD

| 方法 | BLEU@4 | METEOR | CIDEr |
|---|---|---|---|
| S2VT[63] | -- | 29.20 | -- |
| Tempor-attention[64] | 41.92 | 29.60 | 51.67 |
| MAM-RNN[65] | 41.40 | 32.20 | 53.90 |
| G+HRNE[66] | 43.8 | 33.10 | -- |
| Boundary[67] | 42.50 | 32.40 | 63.50 |
| G+MCNN+MCF-element-wise product | 45.65 | 33.56 | 73.86 |
| G+MCNN+MCF-matrix multiply | **46.46** | **33.72** | **75.46** |

## 4 多模态融合架构

多模态网络架构主要分为三种，即协同架构、联合架构和编解码器架构。

### 4.1 协同架构

协同架构的目标是查找协同子空间中各个模态之间的关联性。多模态协同架构是将各种单一模态在约束条件的作用下实现相互协同[68]。由于各个模态中所包含的信息有所差异，所以多模态协同架构有助于保留每个模态独特的特征。

此类架构[69]在跨模态学习中拥有较为广泛的应用，主流的协同方法是基于跨模态相似性方法，该方法旨在通过直接测量向量与不同模态的距离来学习公共子空间。基于跨模态相关性的方法[70]旨在学习一个共享子空间，从而使不同模态表示集的相关性最大化。图 15 为协同融合架构示意图。

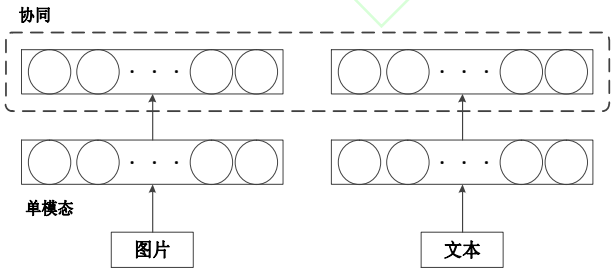



图 15 协同融合架构示意图

**Fig.15 Schematic diagram of collaborative integration architecture**

跨模态相似性方法在相似性度量的约束下保持模态间和模态内的相似性结构，使得相同语义或相关对象的跨模态相似距离尽可能小，不同语义的距离尽可能大[14]。

Ryan Kiros 等人提出的模态间排名方法[71]用来解决图像-文本融合任务，其中 x 为图像嵌入向量，v 为文本嵌入向量，$x_k$、$v_k$ 分别为用于文本嵌入的对比图像和用于图像嵌入的对比句子。定义分数函数 s( x,v )=xv，等价于余弦相似度。对排名的损失函数表示如公式(34)所示：

$$\min_{\theta} \sum_x \sum_k \max\{0, \alpha - S(x,v) + S(x,v_k)\} + \sum_v \sum_k max\{0, \alpha - S(v,k) + S(v,x_k)\} \quad (34)$$

此类方法较好的保存了各个模态之间的相似性结构。协同架构的优点是每个独立的模态都可以运行，这个优点有助于跨模式的迁移学习，目的是在各个模态之间传递信息。但此类架构的缺点是模态融合难度比较大，同时模型很难在多种（两种以上）模态之间实现迁移学习。

### 4.2 联合架构

联合模态是指将多模态空间映射到共享语义子空间中，从而融合多个模态特征[68]。每个独立模态通过各自单独的编码之后，就会被映射到共享子空间中，依据这样的方法，在情感分析、语音识别等多模态的分类和回归任务中都表现优异。图 16 为联合融合架构示意图。

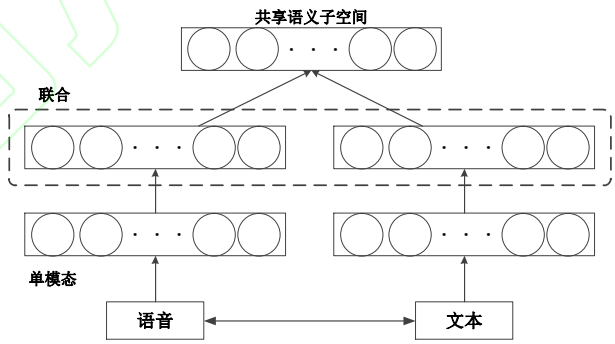



图 16 联合融合架构示意图

**Fig.16 Schematic diagram of joint fusion architecture**

联合架构的核心是实现特征“融合”，直接相加是一种最简单的方法。此方法在不同的隐藏层之间形成共享语义子空间，将经过转换的每个单模态特征向量的语义进行组合，从而完成多模态间的融合。

方法如下式(35)：

$$g = f(W_1^T X_1 + W_2^T x_2 + \cdots + W_n^T x_n) \quad (35)$$

其中，$w$为权重，$x$代表每个单模态，$f$将单个模态的语义映射到共享语义子空间上，$g$为最终的结果。

以上方法虽然实现简单，但容易造成语义丢失，“乘”方法优化了它的这一缺点，让特征语义得到充分的融合。表达式如下式(36)：

$$z = \begin{bmatrix} v^1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} v^2 \\ 1 \end{bmatrix} \otimes \ldots \otimes \begin{bmatrix} v^n \\ 1 \end{bmatrix} \quad (36)$$

其中，$v$表示各个模态，$\otimes$表示外积（即两个向量的向量积）。

此类架构对单个模态的语义的完整性有着比较高

的要求,对于数据的不完整在后期的融合之中会被解决。文献[72]通过各个模态的特征之间的相关性，来找出多个模态之间的关联，并对这些特征进行分类后使用，在视频分类任务中的使用效果显著。

联合架构较其他架构而言，具有融合方式简单的优点，其共享子空间拥有语义不变性，这有利于模型中将一种模态转化为另一种模态。缺点是每个单独的模态在早期较难处理和发现。

### 4.3 编解码器架构

此类架构一般在需要将一种模态映射到另一种模态的多模态转换时使用,由解码器与编码器两个部分组成。编码器将初始模态映射到向量中，解码器基于之前的向量生成一个新模态。编解码器架构在视频解码、图像标注、图像合成等研究领域具有十分广泛的应用。

此类架构的优点是可以在初始模态的基础上生成一个新的模态。缺点是每一个编码器和解码器只能唯一的编码一种模态。图 17 为编码器融合架构示意图。

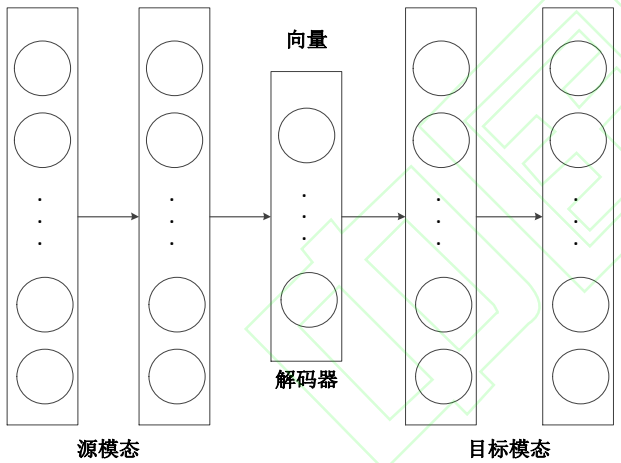



**图 17 编解码器融合架构示意图**

**Fig.17 Schematic diagram of codec fusion architecture**

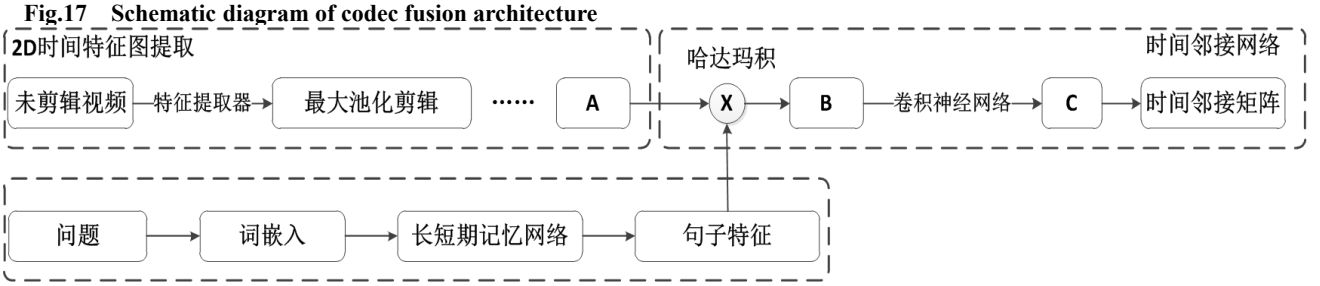



**图 18 2D 时间邻接网络框架图**

**Fig.18 2D time adjacency network frame diagram**

层作为输入句子的特征表示。提取出的语言和视频特征表示之后,从所有候选中预测句子所查询的最佳匹配时刻。它主要包括三个连续的过程:多模态融合、上下文建模和分数预测。$(B, C \in R^{N\times N\times D_H})$

## 5 多模态融合的应用

多模态融合技术，它融合了听觉、视觉、嗅觉、触觉等多种交互方式,使得表达信息的效率和表达信息的完整度更高。多模态以其描述对象的完全性，在多个领域有广泛的应用。以下列举几个比较常见的应用。

### 5.1 多模态视频片段检索

从不确定目标中检索特定时刻。以 **2D** 形式来表示不同的时间片段,为每个时间片段赋予预训练视频特征和语言特征的融合。关注的任务为时序动作检测，即需要在给定的长视频中,检测出其存在的动作片段类别,并定位出动作开始和结束的时间点。

Zhang S, Peng H 等人提出了一种新的 2D 时间邻接网络[73]，核心思想是在二维时间图上检索一个时刻，该时刻将相邻的候选时刻视为时间上下文,该模型可以扩展到其他时间定位任务，如时间动作定位、视频重定位等。图 18 为 2D 时间邻接网络。

提出的 2D 时间相邻网络的框架。它由用于语言表示的文本编码器、用于视频表示的 2D 时间特征映射提取器和用于矩定位的时间邻近网络组成。在模型中，给定一个未剪辑的视频和一句话做实验,来检索最佳匹配的临时段。2D 时间特征图部分主要负责提取输入的视频中的特征，并将这些特征编码成二维时间特征图。在该部分首先对将视频分割为多个视频剪辑,二维时间特征图由三个维数组成,前两维表示开始和结束片段索引,最后一维表示特征维度$(A \in R^{N\times N\times D_v})$在文本编辑器中,对于句子中的每一个单词通过 GloVe word2ve 模型

### 5.2 综合多模态信息生成内容摘要

此类应用是指在输入两种或多种模态信息( 通常包括文本、视频、图像、语音等信息 ) 之后，输出一段对多种模态信息综合之后的总结概括。如何使用相关文本、

音频和视频信息生成文本摘要。

Li H, Zhu J 等人提出了一种提取多模态摘要的方法[74]，可以自动生成一个文本摘要给定的一组文件、图像、音频和视频有关的一个特定的主题。关键思想是缩小多模态内容之间的语义差距。对于音频来使用图像作为对齐来指出文档中的重要句子。对于文本信息，设计了一种选择性使用其转录的方法。对于视觉信息，使用神经网络学习文本和图像的联合表示。最后，考虑所有的多模态方面，通过预算优化子模态函数，最大化显著性、非冗余性、可读性和图像覆盖范围，生成文本摘要。多模态模型的框架图如图 19。

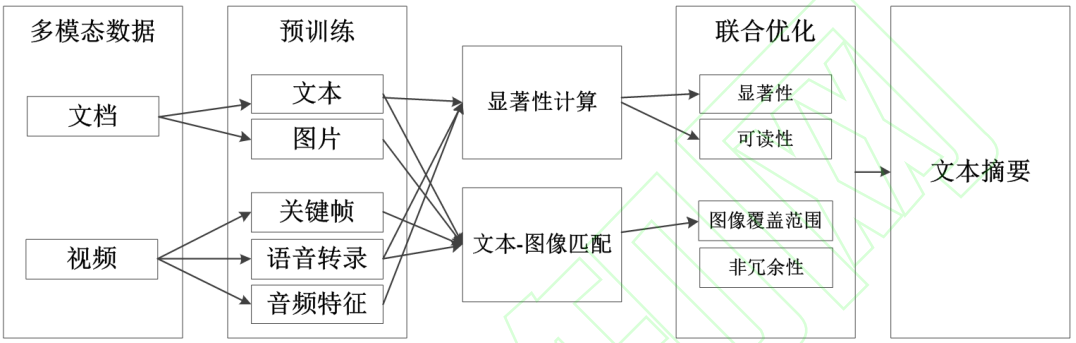



**图 19 多模态摘要模型框架图**

**Fig.19 Schematic diagram of multimodal abstract model**

## 5.3 多模态情感分析

情感分析作为近几年研究的一个热点问题，受到广大研究者的青睐。之前的情感分析大多指文本情感分析，是指利用自然语言处理和文本挖掘技术，对带有情感色彩的主观性文本进行分析、处理和抽取的过程[75]。

近一段时间研究者们开始探索视觉方面情感分析的新思路，并取得了一些进展，并将研究方向转向了多模态中的图像。

Truong Q T 等人提出了一种利用视觉信息进行情感分析的新方法[76]，称为视觉方面注意力网络。该模型有一个分层的三层架构，将表示从单词聚合到句子，然后聚合到特定于图像的文档表示，最后聚合到最终的文档表示。基于这样的观察，即一个句子倾向于集中在特定的东西上，就像每个图像一样，设计了一个模型。该模型的最底层是一个单词编码器，负责把单词转化成句子表示。中间层是句子编码层，借助于视觉方面的注意力，将句子表示转化为文档表示。顶层为分类层，负责为文档添加情感标签。模型图如图 20 所示。

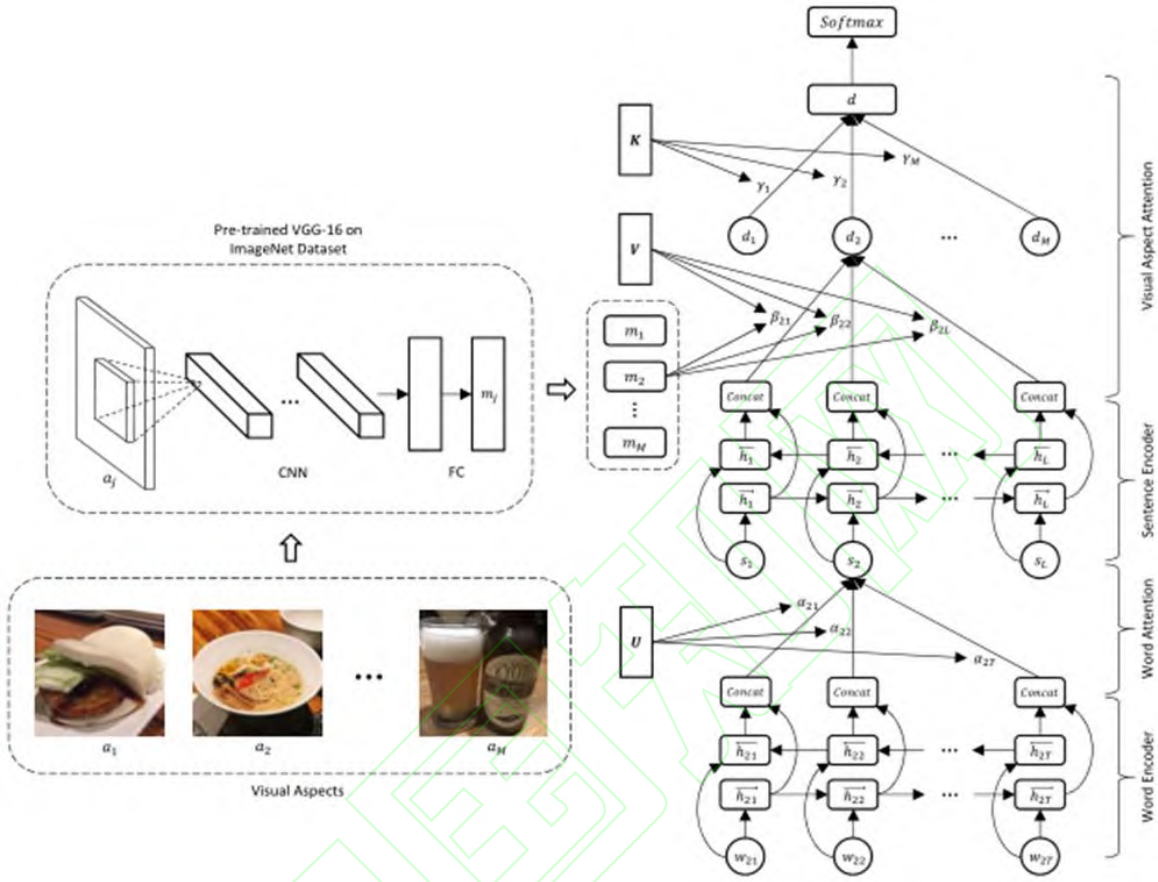



**图 20 视觉注意力网络**[76]

**Fig. 20 Visual attention network** [76]

## 5.4 多模态人机对话系统

对人机对话系统的研究一直以来都是人工智能研究领域中的一个重要的研究方向。人们希望能够与机器进行类似于人与人之间的自然的交流，然而由于自然语言本身的复杂性远高于人造语言，因此对自然语言的处理到目前仍十分具有挑战性，也是人工智能领域最为困难的问题之一。难点主要包括：内容的有效界定、语言的歧义性和瑕疵输入的处理。其中最难以处理的问题是如何消除在对话过程中广泛存在的歧义性。

尽管此前在许多研究中提出了不同的解决方法并取得了不错的处理效果，但大多数是基于单一模态信息提出的解决方案，例如：文本处理和语音识别等。而在交流的过程中，信息的传递通常是通过多种形式进行的，如语音、肢体语言和面部表情等。而不同的模态信息在信息的表达性上具有不同层次的效果，因此在许多情况下难以通过某种单一模态的信息了解到信息传递者所要表达的完整意图。

而多模态人机对话系统则是充分利用了多模态信息之间的互补性，综合来自同一实例的音频、视频、图像、语义等信息进行识别工作，以获得更完整、更好的表达特征，对解决语言理解的歧义性具有很好的效果，如图 21 所示。例如，当用户询问“这本书的价格”时，对话系统需要通过视频根据用户的肢体动作来判断出用户所询问的书目信息进而做出相应反馈。

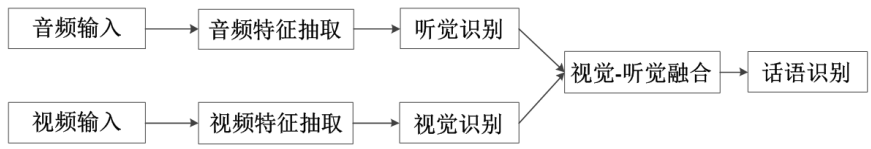



**图 21 基于视觉-音频的多模态识别模型图**

**Fig.21 Visual-Audio Multimodal Recognition Model Diagram**

Hung Le，Doyen Sahoo 等人[77]开发了一个基于视频的对话系统，在该系统中是基于给定视频的视觉和听觉方面进行对话，比传统的基于图像或文本的对话系统更具挑战性，因为视频的特征空间跨越多个图像帧，使得

难以获得语义信息；以及对话代理必须感知和处理来自不同模态(音频、视频、字幕等)的信息来获得全面的了解。而大多数现有的工作都是基于 RNNs 和序列到序列的架构，这对于捕获复杂的长期依赖关系(如在视频中)不是很有效。为了克服这一点，提出了多模式变压器网络(MTN，一个基于多脑注意力的神经网络，可以在多模态环境下产生良好的会话反应)，来编码视频和合并来自不同模式的信息。并且提出了通过自动编码器从非文本模式中提取查询感知特征的查询软件关注。并为此开发了一个训练程序来模拟令牌级解码，以提高推理过程中生成的响应的质量。我们的模型还推广到另一个多模态视觉对话任务，并获得了良好的性能。模型的整体框架如图 22 所示。

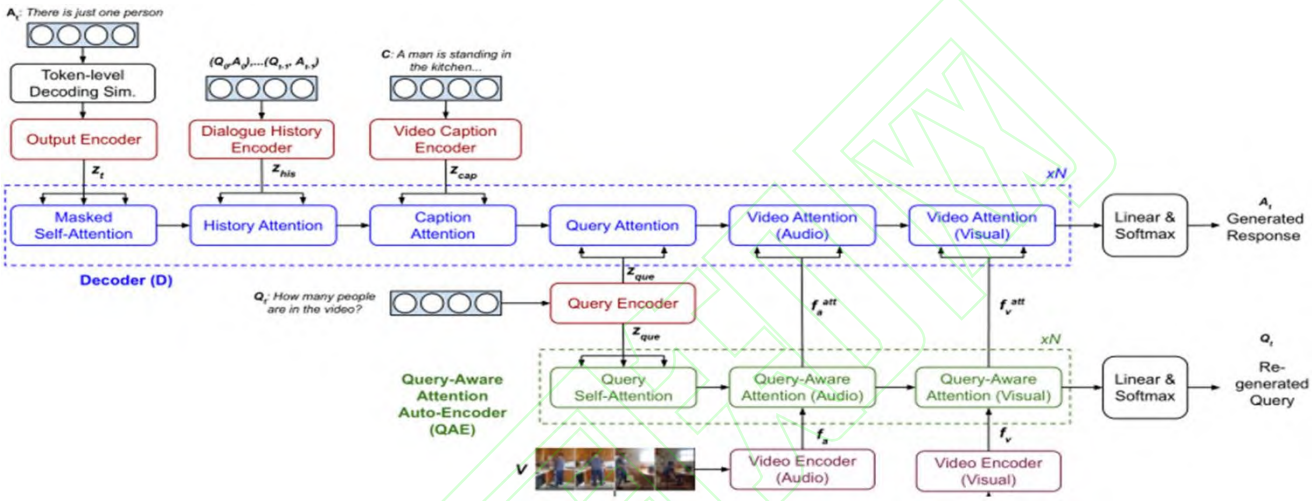



**图 22 多模态转换网络架构**[77]

**Fig.22 Multi-modal Conversion Network Architecture** [77]

Chen Cui，Minlie Huang 等人[78]提出了用户注意力指导的多模态对话系统。模型的整体框架图如下图所示，该框架的任务是生成回复文本和选择回复图像，希望通过利用多模态对话的形式，结合不同模态信息，以给用户更加直观的印象，同时能够更加清晰的了解用户的表达。从高层的角度来看，双向 RNN 模型被用于编码用户和聊天机器人之间的话语级交互。对于低层视角，多模态编码器和解码器能够分别编码多模态话语和生成多模态响应。多模态编码器在分类-属性组合树的帮助下学习图像的视觉呈现，然后视觉特征通过注意机制与文本特征交互；而多模式解码器根据对话历史选择所需的可视图像并生成文本响应。为了评估提出的模型，该文作者在零售领域的公共多模态对话数据集上进行了大量实验。实验结果表明，通过整合多模态话语和基于用户属性级注意力的视觉特征编码，模型效果优于现有的先进方法。模型的整体框架图如图 23。

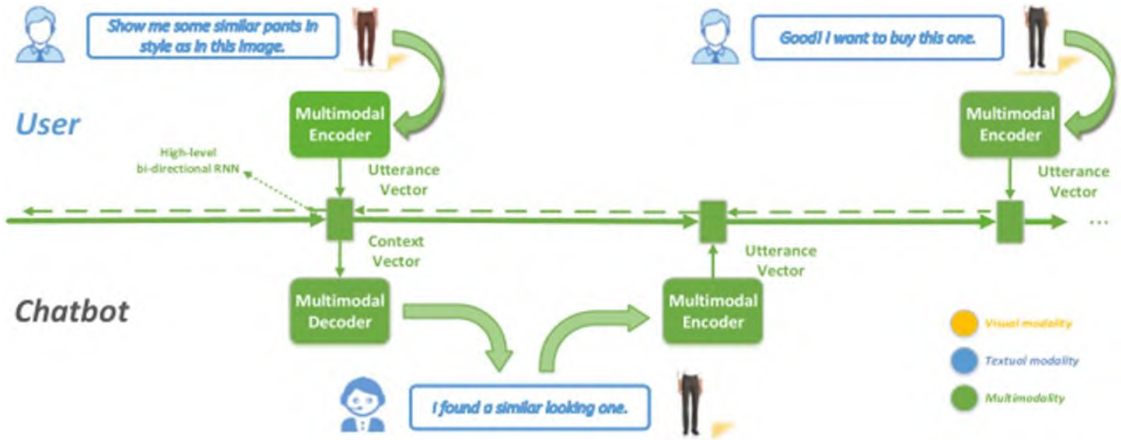



**图 23 用户注意力指导的多模态对话系统模型**[78]

**Fig.23 User Attention Guided Multimodal Dialog System Model** [78]

## 6 多模态融合有助于深度学习可解释

深度学习技术以数据驱动学习的特点，在自然语言处理、图像处理、语音识别等领域取得了巨大成就[79]。由于深度学习模型具有数据量大、神经网络层数较深、结构复杂等特点，使得可解释性变差，是人工智能领域中的一大挑战。因为在数据通过神经网络并得出相关预测结果时，人们往往难以解释其产生的原因，这样就会导致在很多时候研究人员不清楚如何修正和优化神经网络，进而提高其效率或避免其在应用过程中产生难以挽回的错误。

多模态融合技术通过充分利用多模态信息之间的互补性，能够获得更完整、更好的特征表达。使得在保证模型效果的情况下，在学习的过程中对不同的特征获得不同程度的强化，这对深度学习的可解释性有一定的帮助。在此基础之上可以进一步引入注意力机制，这种方法在保证模型效果的前提下，通过引入注意力向量，对信息特征及多模态深度学习网络中的隐藏层特征赋予不同的权重，并在训练过程中对该权重进行学习，进一步加强了其学习效果。通过利用各个特征对于模型学习的重要性程度对模型进行理解，进而达到解释模型的效果。

## 7 总结与展望

本文总结了多模态数据融合的研究现状，总结分析多模态融合方法、单一模态的表示、融合完成后信息的表示、多模态深度学习模型、融合架构。多模态融合方法分为模型无关的融合方法和模型相关方法两种。模型无关方法有早期、晚期、混合融合三种；模型相关方法包括多核学习方法、图像模型方法和神经网络方法三种。单一模态的表示、融合完成后信息的表示是融合过程的基础，确保特征提取及融合过程中信息的完整性是融合成功的关键。

由于单一模态缺乏多样性，现如今的研究者已经开始着手进行多模态的输入与输出，当输出信息包含多种模态时，可以尽可能达到研究者的期望。例如当给出一段外文视频，我们对其中的语言并不熟悉时，可以通过图片和视频大致领会重要内容。在未来的研究中，跨模态学习将会变成一个热点问题，在各个研究领域都会有广泛的应用。人工智能的最终目的是设计出完全与人类智能相媲美的智能计算机系统[80]。而单一的自然语言处理、计算机视觉和语音识别技术从一种模态对信息的理解与人类的行为之间有着较为明显的差异。所以多模态的应用比单一模态更接近人类的行为。作为一种能让机器更加贴近人类行为的技术，跨模态学习有望在未来获得全面的发展。

下一步可利用多模态交互在空间上可以更快、效率更高、交互识别准确率更高的优势，而且对于关键的任务完成率更高的特点，针对多模态人机交互融合进行研究。单一的交互模式有时候在交互过程当中会有一定的局限性，并且交互效率较低，而采用多模态融合交互模式会解决这些问题。并对模态间的语义冲突、多模态融合程度评价标准等研究不充分的问题进一步研究，推动该技术在机器学习的一些新的领域中的发展。

## 参考文献


[1] Liu J, Li T, Xie P, et al. Urban big data fusion based on deep learning: An overview[J]. Information Fusion, 2020, 53: 123-133.

[2] 赵亮.多模态数据融合算法研究[D].大连理工大学,2018.
ZHAO Liang. Research on Multi-modal Data Fusion Algorithm[D]. Dalian University of Technology,2018.

[3] 韩崇昭,朱洪艳, 段战胜.多源信息融合[M].清华大学出版社, 2010.
Han Chongzhao, Zhu Hongyan, Duan. Multi-source Information Fusion [M]. Tsinghua University Press, 2010.

[4] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects[J]. Proceedings of the IEEE, 2015, 103(9): 1449-1477.

[5] Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia systems, 2010, 16(6): 345-379.

[6] Nefian A V, Liang L, Pi X, et al. Dynamic Bayesian networks for audio-visual speech recognition[J]. EURASIP Journal on Advances in Signal Processing, 2002, 2002(11): 1-15.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.

[8] Wu Z, Cai L, Meng H. Multi-level fusion of audio and visual features for speaker identification[C]//International Conrence on Biometrics. Springer, Berlin, Heidelberg, 2006: 493-499.

[9] Martínez H P, Yannakakis G N. Deep multimodal fusion: Combining discrete events and continuous signals[C]// roceedings of the 16th International conference on multimodal interaction. 2014: 34-41.

[10] Ni J, Ma X, Xu L, et al. An image recognition method based on multiple bp neural networks fusion[C]//International Conference on Information Acquisition, 2004. Proceedings. IEEE, 2004: 323-326.

[11] Murphy R R. Computer vision and machine learning in science fiction[J]. Science Robotics, 2019, 4(30).

[12] Yeh Y R, Lin T C, Chung Y Y, et al. A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection[J]. IEEE Transactions on multimedia, 2012, 14(3): 563-574.

[13] 陈鹏, 李擎, 张德政, 等. 多模态学习方法综述[J]. 工程科学学报, 2020, 42(5): 557-569.
CHEN Peng, LI Qing, ZHANG Dezheng, et al. A review of multimodal learning methods[J]. Chinese Journal of Engineering Science, 2020, 42(5): 557-569.

[14] 何俊,张彩庆,李小珍,张德海.面向深度学习的多模态融合技术研究综述[J].计算机工程,2020,46(5):1-11.
He Jun, Zhang Caiqing, Li Xiaozhen, Zhang Dehai. A review on the research of multi-mode fusion technology for deep learning[J]. Computer Engineering,2020,46(5):1-11.

[15] McFee B, Lanckriet G, Jebara T. Learning Multi-modal Similarity[J]. Journal of machine learning research, 2011, 12(2).

[16] Gönen M, Alpaydın E. Multiple kernel learning algorims[J]. The Journal of Machine Learning Research, 2011, 12: 2211-2268.

[17] Sutton C, McCallum A. An introduction to conditional random fields for relational learning[J]. Introduction to statistical relational learning, 2006, 2: 93-128.

[18] Friedman N, Murphy K, Russell S. Learning the structure of dynamic probabilistic networks[J]. arXiv preprint arXiv:1301.7374, 2013.

[19] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proc. 18th International Conf. on Machine Learning. 2001.

[20] 薛俊欣. 条件随机场模型研究及应用[D]. 山东大学, 2014.
Xue Junxin. Research and Application of Conditional Random Field Model [D]. Shandong University, 2014.

[21] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]//Proceedings of the 28th International Conference on Machine Learning. Washington D. C. , USA: IEEE Press, 2011: 689-696

[22] Wöllmer M, Metallinou A, Eyben F, et al. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling[C]//Proc. INTERSPEECH 2010, Makuhari, Japan. 2010: 2362-2365.

[23] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014.

[24] Aiolli F, Donini M. EasyMKL: a scalable multiple kernel learning algorithm[J]. Neurocomputing, 2015, 169: 215-224.

[25] Jiang X, Wu F, Zhang Y, et al. The classification of multi-modal data with hidden conditional random field[J]. Pattern Recognition Letters, 2015, 51: 63-69.

[26] Wöllmer M, Metallinou A, Eyben F, et al. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling[C]//Proc. INTERSPEECH 2010, Makuhari, Japan. 2010: 2362-2365.

[27] Whitley D, Starkweather T, Bogart C. Genetic algorithms and neural networks: Optimizing connections and connectivity[J]. Parallel computing, 1990, 14(3): 347-361.

[28] Stanley K O, Miikkulainen R. Efficient reinforcement learning through evolving neural network topologies[C]// Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation. 2002: 569-577.

[29] Shinozaki T, Watanabe S. Structure discovery of deep neural network based on evolutionary algorithms[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4979-4983.

[30] Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends[J]. IEEE Signal Processing Magazine, 2017, 34(6): 96-108.

[31] Shahriari B, Swersky K, Wang Z, et al. Taking the Human Out of the Loop: A Review of Bayesian Optimization[J]. Proceedings of the IEEE, 2016, 104(1).

[32] Ramachandram D, Lisicki M, Shields T J, et al. Structure optimization for deep multimodal fusion networks using graph-induced kernels[J]. arXiv preprint arXiv:1707.00750, 2017.

[33] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[34] 陆莉霞,邹俊忠,郭玉成,张见,王蓓.多模态融合的膝关节损伤预测[J/OL].计算机工程与应用:1-14
Lu Lixia, Zou Junzhong, Guo Yucheng, Zhang Jian, Wang Bei. Prediction of knee joint injury by multi-modal fusion [J/OL]. Computer Engineering and Applications :1-14

[35] 林敏鸿,蒙祖强.基于注意力神经网络的多模态情感分析[J].计算机科学,2020,47(S2):508-514+548.
Lin Minhong, Meng Zuqiang. Multi-modal Sentiment Analysis Based on Attention Neural Network. Computer Science,2020,47(S2):508-514+548.

[36] Eyben F, Weninger F, Gross F, et al. Recent developments in opensmile, the munich open-source multimedia feature extractor[C]//Proceedings of the 21st ACM international conference on Multimedia. 2013: 835-838.

[37] Muda L, Begam M, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. arXiv preprint arXiv:1003.4083, 2010.

[38] Nojavanasghari B, Gopinath D, Koushik J, et al. Deep multimodal fusion for persuasiveness prediction[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 284-288.

[39] Wang H, Meghawat A, Morency L P, et al. Select-additive learning: Improving generalization in multimodal sentiment analysis[C]//2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017: 949-954.

[40] Anastasopoulos A, Kumar S, Liao H. Neural language modeling with visual features[J]. arXiv preprint arXiv:1903.

02930, 2019.
[41] Vielzeuf V, Lechervy A, Pateux S, et al. Centralnet: a multilayer approach for multimodal fusion[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 0-0.
[42] Zhou B, Tian Y, Sukhbaatar S, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015.
[43] Pérez-Rúa J M, Vielzeuf V, Pateux S, et al. Mfas: Multimodal fusion architecture search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6966-6975.
[44] Zoph B, Le Q V. Neural architecture search with reinforcement learning[J]. arXiv preprint arXiv:1611.01578, 2016.
[45] Liu C, Zoph B, Neumann M, et al. Progressive neural architecture search[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 19-34.
[46] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
[47]Graves A, Wayne G, Danihelka I. Neural turing machines[J]. arXiv preprint arXiv:1410.5401, 2014.
[48] Zhu Y, Groth O, Bernstein M, et al. Visual7w: Grounded question answering in images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4995-5004.
[49] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// International conference on machine learning. PMLR, 2015: 2048-2057.
[50] Yang Z, He X, Gao J, et al. Stacked attention networks for image question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 21-29.
[51] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[J]. arXiv preprint arXiv:1606.00061, 2016.
[52] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 299-307.
[53] Schwartz I, Schwing A G, Hazan T. High-order attention models for visual question answering[J]. arXiv preprint arXiv:1711.04323, 2017.
[54] Arevalo J, Solorio T, Montes-y-Gómez M, et al. Gated multimodal units for information fusion[J]. arXiv preprint arXiv:1702.01992, 2017.
[55] Tenenbaum J B, Freeman W T. Separating style and content with bilinear models[J]. Neural computation, 2000, 12(6): 1247-1283.
[56] Kim J H, On K W, Lim W, et al. Hadamard product for low-rank bilinear pooling[J]. arXiv preprint arXiv:1610.04325, 2016.
[57] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 2333-2338.
[58] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents[C]//ACM SIGIR Forum. New York, NY, USA: ACM, 2017, 51(2): 243-250.
[59] Gao J, He X, Nie J Y. Clickthrough-based translation models for web search: from word models to phrase models[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. 2010: 1139-1148.
[60] Gao J, Toutanova K, Yih W. Clickthrough-based latent semantic models for web search[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 675-684.
[61] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
[62] Wu A, Han Y. Multi-modal Circulant Fusion for Video-to-Language and Backward[C]//IJCAI. 2018, 3(4): 8.
[63] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4534-4542.
[64] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4507-4515.
[65] Li X, Zhao B, Lu X. MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning[C]//IJCAI. 2017: 2208-2214.
[66] Pan P, Xu Z, Yang Y, et al. Hierarchical recurrent neural encoder for video representation with application to captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1029-1038.
[67] Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1657-1666.
[68] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.
[69] He Y, Xiang S, Kang C, et al. Cross-modal retrieval via deep and bidirectional representation learning[J]. IEEE Transactions on Multimedia, 2016, 18(7): 1363-1377.
[70] Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval[C]//Pro-

ceedings of the 18th ACM international conference on Multimedia. 2010: 251-260.

[71] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models[J]. arXiv preprint arXiv:1411.2539, 2014.

[72] Jiang Y G, Wu Z, Wang J, et al. Exploiting feature and class relationships in video categorization with regularized deep neural networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(2): 352-364.

[73] Zhang S, Peng H, Fu J, et al. Learning 2d temporal adjacent networks for moment localization with natural language[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 12870-12877.

[74] Li H, Zhu J, Ma C, et al. Multi-modal summarization for asynchronous collection of text, image, audio and video[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1092-1102.

[75] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. Information Retrieval, 2008, 2(1-2): 1-135.

[76] Truong Q T, Lauw H W. Vistanet: Visual aspect attention network for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1): 305-312.

[77] Le H, Sahoo D, Chen N F, et al. Multimodal transformer networks for end-to-end video-grounded dialogue systems[J]. arXiv preprint arXiv:1907.01166, 2019.

[78] Cui C, Wang W, Song X, et al. [C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 445-454.

[79] 曾春艳,严康,王志锋,等.深度学习模型可解释性研究综述[J]. 计算机工程与应用,2021,57 (8):1-9.

Zeng Chunyan, Yan Kang, Wang Zhifeng, et al. A review on the interpretability of deep learning models [J]. Computer Engineering and Applications,2021,57 (8):1-9.

[80] 马宪民,人工智能的原理与方法[M],西北工业大学出版社,2002.

Ma Xianmin, Principles and Methods of Artificial Intelligence[M]. Northwestern Polytechnical University Press, 2002.