

1 文献综述

1.1 背景和意义

现如今，随着移动互联网的不断普及与发展，网络上的数据量每时每刻都在以指数级增长，带来了信息的大爆炸。而视频作为信息的一大载体，同样为人们频繁使用。尤其是近几年类似抖音、快手等短视频移动应用的兴起，每天数以亿计的用户不仅观看视频，更是直接参与了视频的制作与分享，这更加大量地增长了视频在网络中传播的数量。

那么如何对海量的视频进行信息的处理与分析、如何在大量的视频中搜索信息、如何精准地为用户推荐视频等，都无疑成为热门的研究话题。如果使用人工的方式来处理视频中的信息，一般需要完整地观看每个视频，这势必会导致耗费人们大量的时间与精力。在大数据面前，人脑是显然无法高效处理的，而用机器来对视频中的信息进行处理分析，是对海量视频的处理上的有效选择。但是，要让机器自动地完成对视频内容的分析与理解，这也同样是一大难题。这需要机器不仅可以观察视频中的背景信息，识别视频中的各个目标以及分析各个目标之间的关系，还要能够识别时间维度下的动作特征以及具有一定的推理能力。而如今，随着深度学习领域的快速发展，其在一些代替人脑进行信息处理分析的任务上所展现的强大性能，为自动化高效处理海量视频数据带来了可能^[1-2]。

视频：



问题：what is a man playing while sitting down? 答案：guitar

图 1.1 视频问答任务的例子

视频问答^[3]（Video Question Answering, VQA），类似于较早的视觉问答^[4]（Visual Question Answering, VQA）和文本问答^[5]（Text Question Answering, TQA）等任务，通过问答的方式测试机器对视频的理解能力，是衡量机器高效处理视频信息能力的较好的切入点之一，其在提出后便受到了研究人员的广泛关注。视频问答任务通过给定一个视频以及针对该视频的一系列问题，需要机器可以在分析视频内容和理解问题内容后，

给出所需要的正确答案。视频问答任务的例子如图 1.1 所示，机器首先观看视频后理解该视频内容，从而在询问“坐着的男人在弹什么？”时，可以正确地回答“吉他”。

视频问答任务具有一般问答任务所不具有的困难与挑战，其难度也相对提升，对模型有着更高的要求。首先在理解复杂问题时，问题提问的方式、提问的意图以及问题所针对的视频重点等都是复杂多样的，这就需要机器对问题的各个方面有较好的理解。其次在处理视频上，视频语义的学习是机器理解视频信息的关键所在。而视频具有静态图像所不具有的时序性，对于类似“男人在做什么？”的动作类问题来说，有一些动作就不能仅凭分析单帧的静态图像而推理出来。这种动态序列的分析就需要机器不仅可以观察到视频中的静态图像的信息，识别图像中的各个目标以及分析各个目标之间的关系，还要能够识别动态序列下各个对象的动作特征，并且具有一定的推理能力。同时，视频问答任务是一种跨模态的任务，需要处理来自多个模态的信息，包括对视频和问题之间的综合理解。如何有效地融合各个模态的信息来获得答案，也是该任务所面临的挑战。

如果能够通过人工智能实现对视频的有效处理和分析，那么在这个短视频盛行的时代，我们可以通过该方法高效地完成对海量视频信息的语义理解，并通过提供视频搜索、视频分类、视频推荐等功能极大地提升用户的产品体验，这无疑具有巨大的理论意义和应用价值。

1.2 国内外研究现状

视频问答任务相对于一般的问答任务来说起步更晚，同时由于视频问答数据的采集麻烦以及视频语义分析复杂等难点，其发展相对缓慢。但随着近来数据集的不断构建完善以及深度学习技术的不断发展，对视频问答任务的研究也渐入佳境，各种建模方法层出不穷，受到了学术界的广泛关注。

视频问答任务的数据集主要分为影视类、生活类和生成类。影视类数据集中的视频取自影视作品，包括了从电影中取材的 MovieQA^[6]和从电视剧中取材的 TVQA^[7]等。生活类数据集中视频更加贴近于日常生活的场景，以更好的应用于实际中，比如数据集 LifeQA^[8]。而生成类中的视频则是通过程序自动化生成一些各异的虚拟几何物体形成的，例如数据集 SVQA^[9]中的视频便是通过 Unity3D 工具生成。视频问答数据集中问答形式主要包括了视频检索、选择、填空等方式，而一般通过分类方式来对答案进行预测。

近年来，视频问答任务的研究经过不断发展，其主要的求解方法可以大致抽象为一个简单的视频问答框架，如图 1.2 所示。其主要包括了视频特征的提取、问题特征的提取、多模态的融合以及最后的答案生成，而视频特征提取一般包括了静态的外观特征提

取和动态的动作特征提取。后来的研究均是通过改变各个模块的细节，来不断优化和提升模型性能。

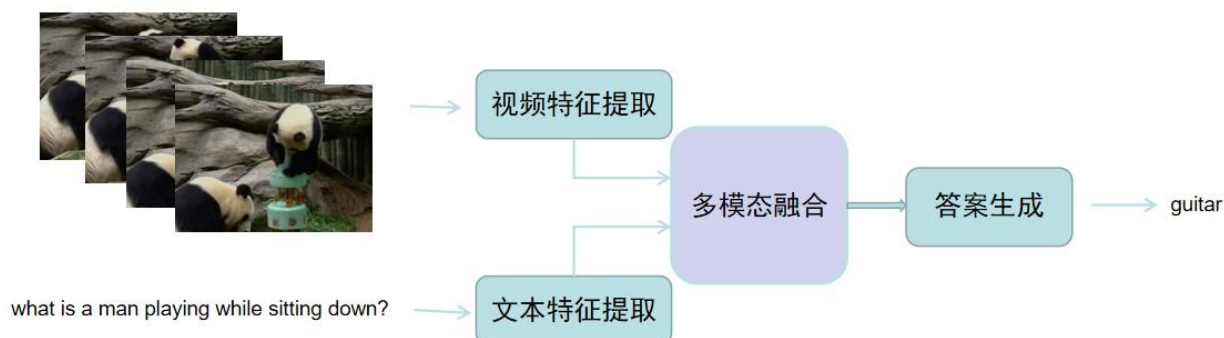


图 1.2 简单的视频问答框架

首先在视频特征提取时，一般通过在 ImageNet^[10]上预训练模型对静态的外观特征进行提取，其网络模型包括 VGG、ResNet^[11]等，使用 Kinetics^[12]上预训练的模型提取动态的动作特征，其网络模型如 C3D^[13]。对于问题文本特征的提取，主要使用预先训练好的词向量表来对每个单词进行编码，将单词表示为定长的向量，包括 Word2Vec、Glove^[14]等，接着通过循环神经网络对文本的语义特征等进一步提取，如 BiLSTM。对于视频问答任务来说，视频外观特征、视频动作特征和问题文本特征之间的交互与融合是研究的重点所在。对此，涌现出了各种各样的实现方式，而研究人员大多借鉴了在视觉问答中略有成效的经验方法，包括注意力机制、图卷积网络等。Jang^[15]等人通过时间和空间两个维度上的注意力，融合视频和问题特征，对视频的重要帧上的重要区域进行定位，对其结果特征进行分类。Kim^[16]等人引入记忆力机制以使模型学习到特征更深层的表征意义。Xu^[17]等人借鉴动态记忆网络（Dynamic Memory Network, DMN）的方法，提出了一个注意记忆单元（Attention Memory Unit, AMU），通过文本中的单词不断地提升在视频特征上的注意力。Gao^[18]等人考虑外观与动作特征之间的相关性，提出了共记忆网络的方法，使用动态注意力进行视频特征的学习。Zhang^[19]等人将卷积方法考虑进来代替循环神经网络，提出层级卷积自注意网络（Hierarchical Convolutional Self-Attention networks, HCSA），通过各个阶段的注意力机制不断的引入问题的注意力。这些方法主要通过注意力、记忆力等机制进行表征学习，但忽略了对对象之间的关系，在推理上仍有不足之处。Le^[20]等人对此提出了条件关系网络（Conditional Relation Network, CRN）来建模视觉对象之间的关系，但在处理多个对象关系时并不高效。随着图卷积网络的兴起，Wang^[21]等人在行为识别任务上通过该方法进行对象关系推理，有效地提升了对

关系之间的学习效果，之后便被广泛应用在视频问答任务中。Jiang^[22]等人提出异源图对齐网络（Heterogeneous Graph Alignment, HGA），将问题和视频的融合对齐特征作为图中的节点进行图卷积操作，通过无向异构图进一步推理模态内和模态间的关系表征。Huang^[23]等人提出位置意识图卷积网络（Location-aware Graph Convolutional Network, LGCN），结合了时间编码和位置编码进行时序的定位，并考虑了每帧下各个对象之间的交互。最后，在答案的生成上，一般采用分类的方式进行。根据在固定数量的答案集合中计算得到的各个候选答案的概率分布，在训练时使用交叉熵损失函数计算损失，在预测时概率最高的即为预测答案。

如今的视频问答任务虽然受到了学术界的广泛关注，但相较于更早的视觉问答和文本问答等，其研究现状仍有诸多不足。首先，视频问答的数据集中，问答形式仍以选择式为主，动态生成答案的方式缺失，这显然相较于一个真正智能的问答系统来说仍有一定距离。其次，在视频特征提取上，仍主要通过预训练模型提取外观和动作特征表征视频，方法较为单一，如何有效的提取视频信息并完整的表征视频语义仍然是一个值得研究的方向。最后，在视频和问题文本之间的交互与融合上，主要还是采用注意力、图卷积等方式进行特征的增强以及对象关系推理，对于对象级的静态特征和动态特征之间的交互与文本词级特征和视频帧级特征之间的协同仍需要更多的尝试。

1.3 主要研究内容

视频问答任务，就是通过给机器输入一段视频 V 以及针对该视频的一个问题 q ，需要机器可以在分析视频内容和理解问题内容后，正确地推断出答案。由于现有的数据集在求解上，一般都是通过从一个候选答案集合中选择出最有可能的答案，作为预测答案 a^* 。那么我们可以将这视为一个分类任务，即机器需要对答案集合 A 求解概率分布，将概率最高的答案作为预测答案。所以，预测答案 a^* 可以表示为：

$$a^* = \operatorname{argmax}_{a \in A} P_{\theta}(a | q, V) \quad (1.1)$$

其中， P 代表预测概率， θ 代表模型的参数。我们希望机器在通过训练后，其预测的答案可以接近于正确答案。

对此，本文结合注意力机制和图卷积网络等方法建立深度学习模型，综合利用了注意力机制在加强视觉和文本特定区域特征关注程度上的优势，以及图卷积网络在视频节点间关系推理上的有效性。同时，模型挖掘了视频中静态的外观特征与动态的动作特征之间的独有特征和关联关系，并通过应用知识蒸馏的方法，在对模型进行压缩的同时进一步加强了外观特征与动作特征之间的融合。本文主要研究内容包括了以下几个方面：

(1) 通过注意力机制加强特定区域特征关注。首先通过自注意力机制对问题特征中各个单词的注意力进行强化,以降低问题中一些不重要的单词的关注程度。同时,通过以问题为导向,计算视觉特征中各个片段上的注意力权重,以将机器的关注点放在与问题更加相关的视频片段中。并且在视频外观独有特征与关联特征融合和视频的各个片段特征融合时,同样使用了自注意力的机制来提升特征的融合效果。

(2) 通过图卷积网络进行视频节点间的关系推理。本文采用了多头图卷积的方式,并将每个图卷积头得到的特征进行连接,作为卷积结果。为了更充分的使用图卷积挖掘与推理视频节点关系,采用了多层图卷积的方式,并在图卷积过程中融入注意力的机制,以加强对视频各个节点之间关系的学习。同时,在该步骤中不仅推理了视频的外观特征和动作特征,还使用了损失函数约束的方式挖掘了在外观中与动作视觉信息相关联的特征与在动作中与外观视觉信息相关联的特征,进一步地提取了外观和动作之间潜在关联关系。

(3) 通过知识蒸馏进一步优化模型。本文在训练教师模型的基础上,构造并训练了一个相对轻量的学生模型来改善模型性能。而其在减少模型可训练参数,对模型进行轻量化的同时,还达到了加强多个模态之间的特征融合的效果。通过添加损失函数的方式,将教师模型中多模态融合后的知识蒸馏出来,用于学生模型中单模态的学习。这样学生模型在单模态训练过程中便可以得到丰富的多模态的信息,在训练初期更早地进行多模态之间的交互,以改善之后的多模态的融合效果。

1.4 论文的组织结构和内容安排

本文一共包括了五个章节,各个章节的内容概括如下:

第一章为绪论。首先对视频问答任务的研究背景及研究意义进行了讨论与分析,接着对国内外基于该任务的相关数据集及方法的现状进行了阐述与分析,最后概括了本文主要的研究内容。

第二章为相关技术的介绍。主要介绍了本文主要应用的相关深度学习技术方法,包括了预训练模型、注意力机制、图卷积网络 and 知识蒸馏。

第三章为本文所建立的深度学习模型。详细描述了模型的架构以及各个模块的具体设计思路。

第四章为实验分析。首先介绍了所使用的数据集,接着描述了本文所提出的方法在数据集上的实验,并对实验结果进行了分析与评价。

第五章为总结与展望。总结了本文的内容,并对视频问答任务未来的研究工作进行展望。

2.1 预训练模型

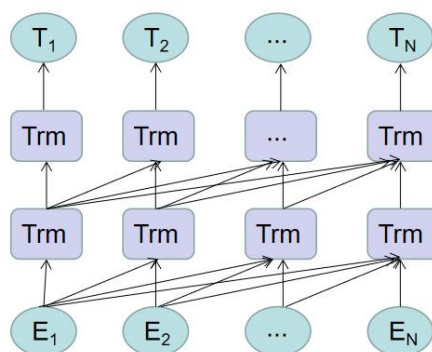


图 2.1 GPT 模型结构

建模型。其首次提出了无监督的预训练加上有监督的微调的训练模式，这使得预训练模型能够更好地适应下游任务，为未来预训练模型的发展带来了方向。BERT 作为 NLP 领域中的一个模型，是最经典的预训练模型之一，其模型结构如图 2.2 所示。其通过堆叠深层的双向 Transformer 来构建整个模型，将预训练加微调这一模式推广到更深层的双向结构中。该模型由谷歌在 2018 年发布，其参数超过了 340 万，所训练的数据量为 16GB 左右。其超出了单一任务的限制，在多种 NLP 测试中都取得了极好的成绩。更重要的是，在这一模型的出现后，越来越多基于此的模型如雨后春笋般层出不穷，并不断突破最高纪录，如 XLNET^[26]、RoBERTa 等模型。2020 年，由 GPT 发展而来的 GPT-3 腾空出世，其由于对各类任务均展现了强大性能，被科学界称为“最接近通用人工智能”的模型。其参数超出了 1750 亿的惊人数量，训练数据量的大小更是高达 45TB，这种大规模的训练下，造就了其在各种各样任务中的出色表现。

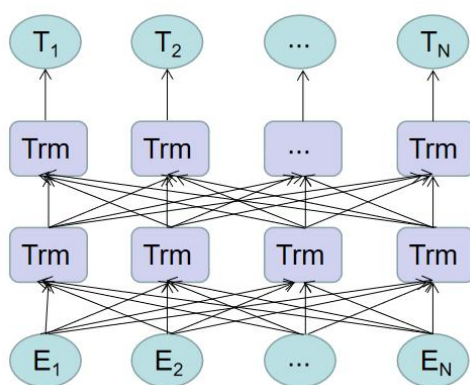


图 2.2 BERT 模型结构

预训练模型的出现促进了人工智能的发展，其意义重大，但目前在结合实际应用的场景下，预训练模型仍有许多需要改进之处。首先，大规模训练所带来的效果十分突出，但其在硬件的性能及训练算法的效率上，仍需要更进一步的优化与改善，以减少训练在算力和时间上的浪费。其次，预训练模型虽然通用性高，但在某些特定领域仍有不足之处，所以针对于特定任务的预训练模型仍有很大的发展空间。

2.2 注意力机制

日常生活中，我们的大脑无时无刻不在接收大量的信息，获取信息的渠道包括了视觉、听觉、味觉等等。而大脑想要时刻处理如此巨大的信息量，就需要选择性地忽略部分不重要的信息，将注意力放在更有用的信息上，才有有更高的处理效率。例如，我们在观察图 2.3 中的动物是猫还是狗时，我们只需要注意照片中的这个动物并识别它即可，

完全可以忽略那些背景图片区域。同样，当我们人类去观察一个场景时，一般不会将该场景所有的细节全部记住，往往是特别关注场景中的某一部分，这是一种高效快速的选择性获取信息的方法。

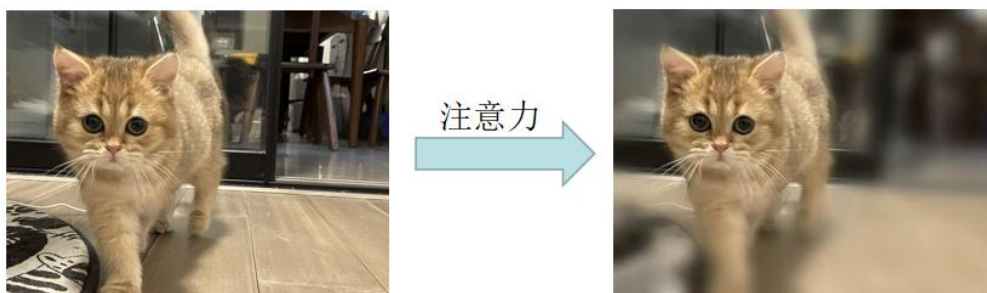


图 2.3 注意力的例子

注意力机制顾名思义，就是想通过该机制使得机器具有像人的注意力一样的能力，为了某种目的而将注意力更多地放在更加有利于达到目的的特征上，即从大量的特征中选择出与当前任务目标更相关、更有利的部分特征。

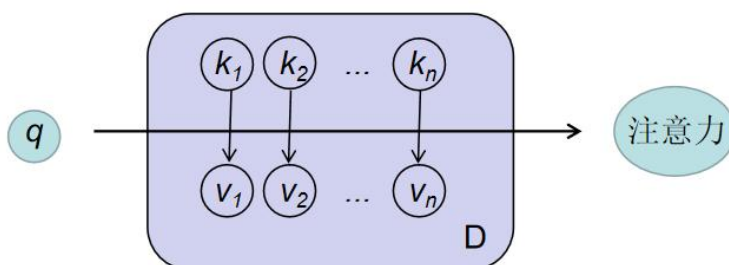


图 2.4 注意力机制

注意力机制本质上可以看作对目标数据集的一个查询，所得到的便是目标数据集在该查询上的注意力结果，基本结构如图 2.4 所示。在此将目标数据集记为 D ，其由若干键值对构成，将键和值分别记为 k 和 v ，将查询记为 q 。如果 q 也同样来自该数据集 D ，查询所得便是数据集的内部注意力，即自注意力。注意力计算过程中，首先计算 q 和各个 k 的相似性或者相关性，并使用 *Softmax* 激活函数归一化后，得到 k_i 对应的注意力得分 $score_i$ ，其公式如下：

$$score_i = \text{Soft max}(\text{Sim}(q, k_i)) \quad (2.1)$$

其中, Sim 代表相似度或相关度的计算, 其可以是点积、余弦函数、MLP 网络 (Multilayer Perceptron) 等。在得到注意力得分 $score_i$ 后, 将其作为 k 对应的 v 的权重, 对 v 进行加权求和, 得到了最终的注意力数值。其公式如下:

$$attention = \sum_{i \in N} score_i * v_i \quad (2.2)$$

其中 N 表示数据集 D 的大小。

如今, 注意力机制在深度学习中的应用十分广泛, 包括了 CV、NLP 等领域。注意力机制在 CV 领域中首次提出, 之后 Google mind 团队创新地将注意力机制与循环神经网络 (Recurrent Neural Network, RNN) 结合使用^[27], 通过注意力针对性地选择部分图像并提取简化特征, 避免了单独使用 RNN 所带来网络层数过深而产生的梯度爆炸的弊端。该机制在 CV 领域产生巨大成效后, 其也被用于在 NLP 领域。例如在机器翻译任务中, 将注意力机制应用在编-解码器 (encoder-decoder) 框架中, 实现模型在处理文本时翻译和对齐的同时进行, 并且解决了输出语句长度不固定的问题, 显著地提高了模型翻译的性能。长短期记忆网络 (Long Short-Term Memory, LSTM) 模型结构同样应用了注意力的机制^[28], 其通过遗忘门、记忆门、输出门将长序列数据转换为较短的序列, 实现了对重要信息的提取, 避免了长序列带来的梯度爆炸或梯度消失问题。而最近几年火热的 Transformer^[25]将自注意力计算作为基本单元, 同样效果显著。其通过添加位置编码和多头注意力机制, 使得数据可以并行输入, 提升速度、减少存储空间占用的同时, 有效提取了数据内部相关性。

注意力机制在 encoder-decoder、LSTM 中的应用可谓是初露锋芒, 但在近几年所出现的 Transformer 等模型结构中, 注意力机制更是大放异彩, 对深度学习的发展产生了重要影响。如今学术界对其研究改进的热度依旧, 注意力机制仍有极大的发展潜力。

2.3 图卷积网络

几年来, 图卷积网络 (Graph Convolutional Network, GCN) 由于在处理具有逻辑关系的拓扑结构上的巨大优势而引起了广泛关注, 并且在图像识别、自然语言处理等领域得到了有效应用。包括 RNN、CNN 等的传统模型只能有效地处理欧氏空间下的数据, 如图片、文本等。由于这些数据排列整齐规律, 具有平移不变性的特点, 普通的卷积网络可以通过全局共享的卷积核来建模图像的局部连接, 从而获得图像更加深层信息的特征表示。但是在对于非欧氏空间下的图数据, 如图 2.5 所示。其由节点和边构成, 比如社交网络图、通信网络图等。其各个节点的局部结构并不具有平移不变的性质, 普通的卷积网络无能为力。因此, 在卷积网络的基础之上, 图卷积网络被学者们提出。

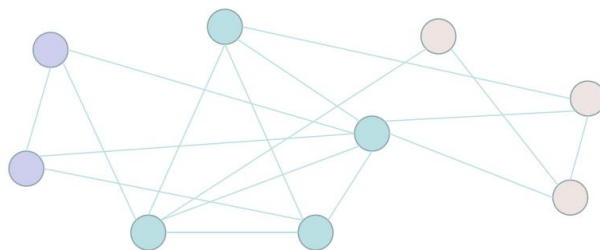


图 2.5 图数据

图卷积网络关键在于如何构建卷积算子，由此可以分为谱方法和空间方法。最开始出现的图卷积网络便是谱方法，其基于卷积定理在谱空间上定义图卷积。之后空间方法开始出现，其结合使用了注意力机制、序列化模型等建模方法计算节点间的权重。同时，卷积网络中所使用的池化算子也被考虑到图卷积中，可以通过该算子学习图的层级结构，主要被应用于图片分类任务。

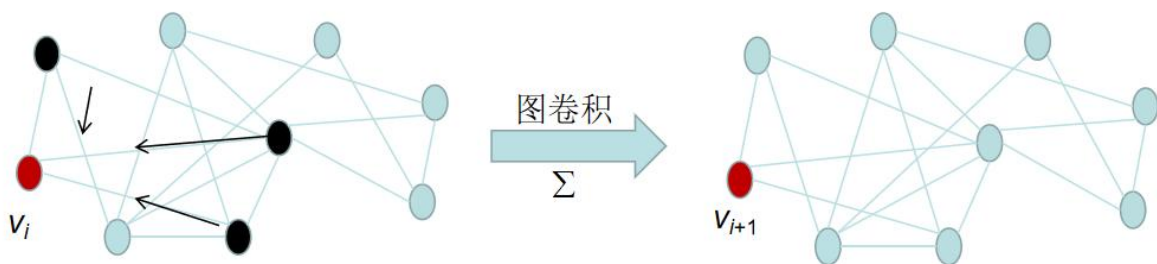


图 2.6 图卷积计算过程

图卷积网络和普通卷积网络一样，也使用了权值共享的思想。如图 2.6 所示，其一般计算过程是在对相邻节点的值加权求和后，更新每个节点的值。一个基本的图卷积不同层之间特征计算过程公式如下：

$$V_{i+1} = \sigma(AV_iW_i) \quad (2.3)$$

其中， A 代表邻接矩阵， V_i 代表第 i 层图卷积的输入特征， V_{i+1} 代表该层图卷积的输出特征， W_i 代表第 i 层图卷积的权重矩阵， σ 代表激活函数。在该计算公式中，权重矩阵和邻接矩阵是在各类图卷积的选择上所重点关注的。

近年来图卷积以其简单而强大的表示能力而被广泛应用于各种领域，其不仅在处理图数据上具有明显优势，并且在简单堆叠较少的层数时便可以达到较好的效果。但与此同时其仍然具有可以深入研究的地方。首先，在实际场景中数据网络的规模往往十分巨

大，如网络应用上的社交网络图，其涵盖了数以亿计的节点和边。对于该类大规模的网络图，目前的图卷积方法仍不太适用。其次，目前大多图卷积方法针对于静态的图结构进行建模，而如何处理动态的网络图也仍是一个值得研究的问题。还有图卷积网络成效如此显著，但在对其原因的理论解释上，在学术界还没有达成共识，仍待研究人员的探索。

2.4 知识蒸馏

深度学习方法通过学习参数以解决目标任务，而高性能的深度学习模型往往需要大量的参数学习，这也意味着训练过程会耗费大量的算力资源。如此庞大的消耗导致了深度学习模型离更好地应用于实际工业生产仍有一段距离。因此，如何训练一个轻量并且高性能的深度学习模型成为了重要研究话题。知识蒸馏（Knowledge Distillation, KD）方法便是其中之一，在满足模型在低功耗与实时性的同时，达到较好的性能。但除了其最初模型压缩的目的之外，其应用空间得到了扩展，研究人员开始通过知识蒸馏以达到模型性能增强的目的。知识蒸馏以其独特的优势得到了学术界广泛的研究与应用。

知识蒸馏通常意义指一种教师-学生式的训练架构，在训练完成大规模复杂的教师模型后，将已训练的教师模型的知识蒸馏出来供相对简单的学生模型学习，而学生模型只需要以轻微的损失计算为代价便可学习到教师模型中丰富的知识。若是以模型压缩为目的，学生模型往往是一种轻量而高效的模型，其参数量远小于庞大复杂的教师模型，而最终性能相对于教师模型却没有下降很多。而模型增强则旨在通过教师模型丰富的知识指导学生模型，通过自学习和互学习等策略或利用跨模态等数据，进一步提高模型性能。其模型训练结构如图 2.7。

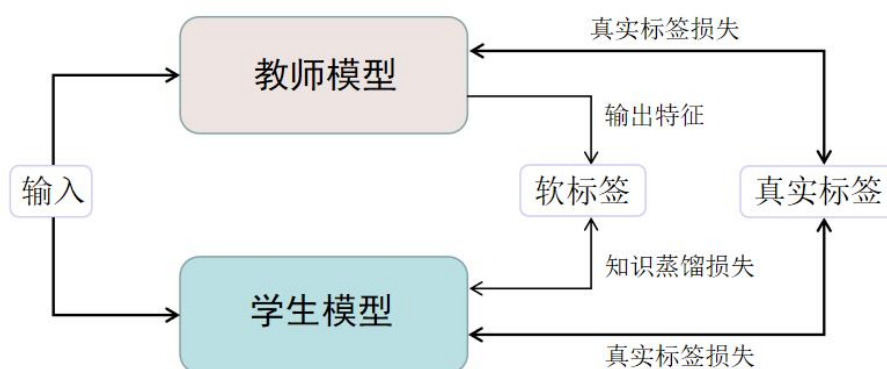


图 2.7 知识蒸馏训练结构

在分类任务中,传统的模型学习过程是将最后一层逻辑单元的输出值,通过 *Softmax* 激活函数进行归一化后,用所得到的类概率来与真实标签计算损失,从而进行迭代训练,其类概率公式为:

$$p(y_i) = \frac{\exp(y_i)}{\sum_l^n \exp(y_l)} \quad (2.4)$$

其中, y_i 表示为第 i 类逻辑单元的输出值, $p(y_i)$ 表示所预测的类别为第 i 类的概率, n 表示类别的个数。

对于知识蒸馏来说,将最后一层逻辑单元的输出用于学生模型学习时,其所包含的大量噪声信息可能会导致学生模型过拟合,影响其泛化能力;而如果使用类概率时,同样会造成信息丢失的问题。因此,在考虑这些问题下,学习软目标概念的知识蒸馏被提出,其知识表示公式如下:

$$p_{KD}(y_i, T) = \frac{\exp(y_i / T)}{\sum_l^n \exp(y_l / T)} \quad (2.5)$$

其中, T 代表所设置的蒸馏温度,为超参数,通过温度 T 的调节来改变软标签的软化程度。于此同时,在训练过程中加上真实标签的训练会使训练效果有效提升,因此,知识蒸馏时学生模型所计算的总损失可表示为:

$$L = \alpha L_{KD}(p(y_t, T), p(y_s, T)) + L_S(z, p(y_s, T)) \quad (2.6)$$

其中, y_t 和 y_s 分别表示教师模型和学生模型的输出特征, z 表示真实标签值, L_{kd} 和 L_s 分别表示知识蒸馏损失和真实标签损失的计算函数,一般为交叉熵损失函数, α 为超参数,通过调节该参数改变知识蒸馏损失在总损失中的权重大小。

由于知识蒸馏在模型压缩与模型增强上的显著优势,研究人员开始对知识蒸馏进行了多样化的应用,其中就不仅包括了通过输出特征学习软目标的知识,还有在模型计算的早期过程中便融入中间特征的知识进行学习,以及对关系特征知识和结构特征知识的学习。但在知识蒸馏机制应用如此广泛的同时,选择获取何种知识、获取何处的知识、学生结构如何定义等问题仍然需要研究人员共同努力解决。

3 基于知识蒸馏的视频问答模型

视频问答旨在模型需要对视频以及视频对应的问题进行分析与理解后,对该问题的正确答案进行分类,答案的类别即在一个固定数量的答案集中。为了解决该任务,本方法搭建了一个基于知识蒸馏的视频问答模型,一个以包括了视频和问题的多模态特征作为输入的答案分类模型。本文依据 DualVGR^[29]对模型整体框架进行了建模。在此基础上,为了实现压缩模型,以及使用大模型丰富的多模态知识来优化小模型的特征学习过程的目的,本文提出了一种多模态知识蒸馏的方式,来进一步提升模型性能。

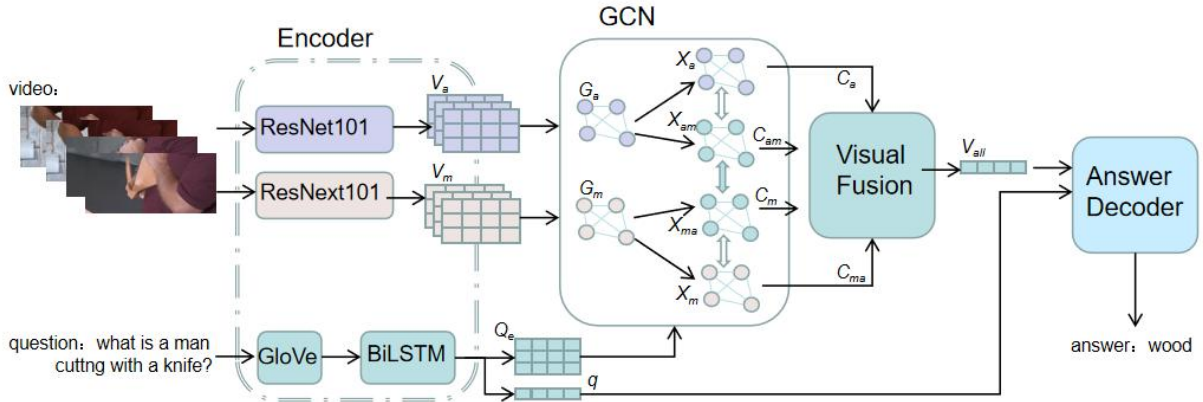


图 3.1 模型结构

在本方法中,分别训练了可训练参数较多的教师模型,以及在教师模型指导下的体量相对较小的学生模型。模型的整体结构如图 3.1 所示。其首先包括了视觉特征和文本特征的编码模块,通过预训练模型提取视频特征,得到外观特征和动作特征,同时使用预训练词向量表编码问题中的单词,之后通过 BiLSTM 处理来更好地提取了视频帧和问题单词的上下文信息。接着是视觉-文本交互模块,该模块主要采用了注意力的机制以及多头图卷积的方式,分别对视觉与文本信息进行了交互,减少了无关信息对模型的干扰,同时对视频各片段间关系进行推理。然后是视觉特征融合模块,该模块同样采用了注意力的机制对独有和关联特征融合及外观和动作特征进行融合,以及对各个片段的视觉特征进行了融合,得到了融合的视觉特征。最后是答案生成模块,其通过融合视觉特征以及文本语义特征,通过解码器实现了对最终答案的分类。

教师模型与学生模型在模型结构上完全相同,只是在图卷积部分略有差异。在教师模型中,为了更充分的使用图卷积挖掘信息,本方法采用了多层图卷积的方式。而在学生模型中,为了达到减少可训练参数的目的,仅使用了单层的图卷积进行训练。

3.1 编码模块

3.1.1 视觉编码

在视觉编码模块中，本方法使用特征向量来对视频进行表示，这里分别提取了视频的外观特征和动作特征。

由于视频的长短及每张图像的大小不一致，需要对视频特征的维度进行规范化，同时在回答视频对应的问题时，可以得出答案的视频内容可能不是整段视频，而是通过视频中几帧的片段便可以得出答案。所以本方法将一个视频进行均匀地切分，每个视频表示为 c 个连续的片段，每个片段包含了 f 帧的图像，图图像的像素大小固定为 $h \times w$ 。最终所得到的视觉特征表示为 $V \in \mathbb{R}^{c \times f \times h \times w \times \text{channel}}$ 。其中 c 表示片段数量， f 表示每个片段的帧数， h 表示图像像素的高度， w 表示图像像素的宽度， channel 表示图像像素的信道数。

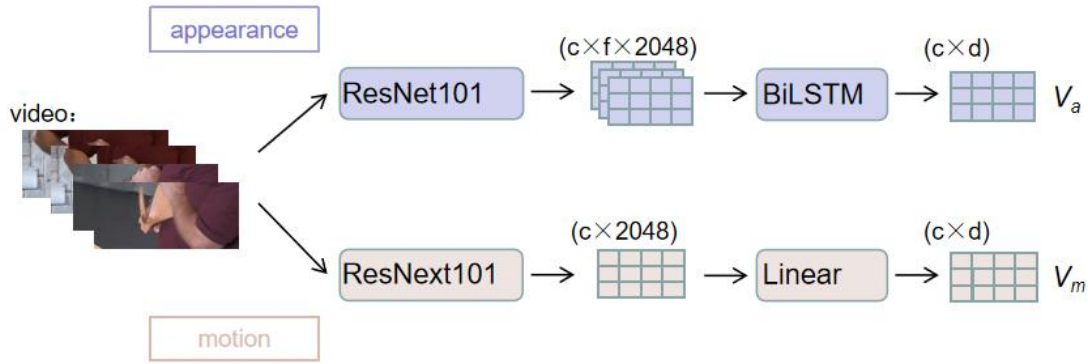


图 3.2 视觉编码

由于视频具有时空性，在特征提取时不仅要获得其静态的空间特征，也要得到其动态的时间上的特征，所以本方法通过预训练模型分别提取了视频静态的外观特征和动态的动作特征，如图 3.2 所示。对于外观特征本方法使用了预训练模型 ResNet-101 来提取，将该模型最后的全连接层除去后，通过该模型将每一帧图像表示为 2048 维的特征向量。接着，为了获取视频每个片段综合各帧的语义特征，使用了 BiLSTM 对每一个片段进行处理，得到了视频每个片段的语义特征表示。最终所得到的静态的外观特征表示为 $V_a \in \mathbb{R}^{c \times f \times d}$ ，其中 c 表示每个视频的片段数， f 表示每个片段的帧数， d 表示模型维度。对于动作特征，使用了预训练模型 ResNeXt-101 来提取，同样将该模型最后的全连接层除去后，使用该模型将每一个片段用 2048 维特征向量进行表示。接着使用了一层线性层对特征进行处理，以将特征维度映射为模型维度大小。最终所得到的动态的动作特征表示为 $V_m \in \mathbb{R}^{c \times d}$ ，其中 c 表示每个视频的片段数， d 表示模型维度大小。

3.1.2 文本编码

在文本编码模块中，需要对问题进行特征向量表示，这里分别考虑了问题的单词特征、嵌入特征及语义特征。

首先将训练集中所有问题出现的高频单词整理为词汇表，通过词汇表对单词进行唯一编码，以此每个句子可以表示为一串单词索引序列，便得到了对句子的编码表示，即 $Q \in \mathbb{R}^l$ ，其中 l 表示问题的长度。

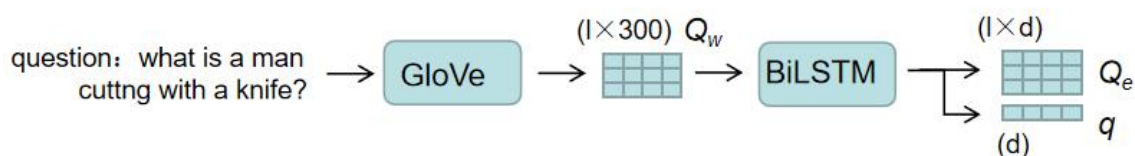


图 3.3 文本编码

为了更好地得到问题句子中的单词特征表示，本方法使用了预训练的 GloVe^[16]词向量表对问题中的每一个单词进行编码表示，如图 3.3 所示。将每一个单词表示为 300 维特征向量，便得到了问题的单词特征，即 $Q_w \in \mathbb{R}^{l \times 300}$ ，其中 l 表示问题的最大长度，300 表示 GloVe 嵌入特征的维度。同时，为了进一步提取单词的上下文特征以及整个句子的语义特征，分别使用了两个 BiLSTM 来提取句子的嵌入特征和语义特征。每个句子的嵌入特征表示为 $Q_e \in \mathbb{R}^{l \times d}$ ，语义特征表示为 $q \in \mathbb{R}^d$ ，其中 l 表示问题的最大长度， d 表示模型维度大小。

3.2 视觉-文本交互模块

3.2.1 文本注意力

当对一个问题进行理解并做出回答时，其中某些单词的作用可能只是用来规范语法，而对于推理答案来说不那么重要。所以为了更好地突出文本的单词编码中某些单词的重要程度，而适当忽略某些不重要的单词，模型通过注意力的机制对特征进行了优化。

通过对嵌入特征 Q_e 计算各个单词的注意力得分以作为权重，对单词特征 Q_w 进行加权求和，以得到问题的整体表示，如图 3.4 所示。具体来说，首先将 Q_e 通过了一层线性层进行处理并使用 L2 归一化后，再通过一层线性层映射为得分。接着使用 Softmax 激活函数进行归一化处理，便得到了各个单词的注意力得分 α 。之后将 α 与单词编码特征进行加权求和，便得到了整个句子的注意力特征 Q_{att} ，公式表示为：

$$\begin{aligned}\alpha &= \text{Softmax}(\text{L2Norm}(Q_e W_1) W_2) \\ Q_{att} &= \alpha^T Q_w\end{aligned}\quad (3.1)$$

其中, W_1 和 W_2 为可学习参数, $W_1 \in \mathbb{R}^{d \times d}$, $W_2 \in \mathbb{R}^{d \times 1}$, 得到的注意力特征 $Q_{att} \in \mathbb{R}^{300}$ 。

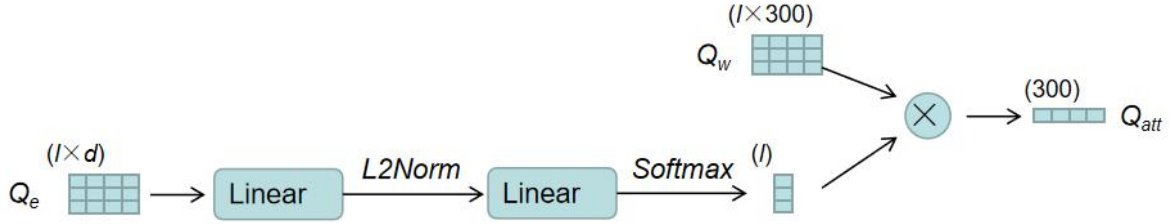


图 3.4 文本注意力

3.2.2 问题导向的视觉注意力

一段视频由多个片段构成, 而在回答某个问题时, 可能该问题只与视频的某几个片段相关。模型只需要重点理解这几个片段就可以对问题做出回答, 而不需要将所有片段进行考虑。所以在这里将以问题为导向的注意力作用在视频的各个片段上, 以获得视频各个片段的注意力得分, 来重点关注某些片段而适当忽略某些片段。

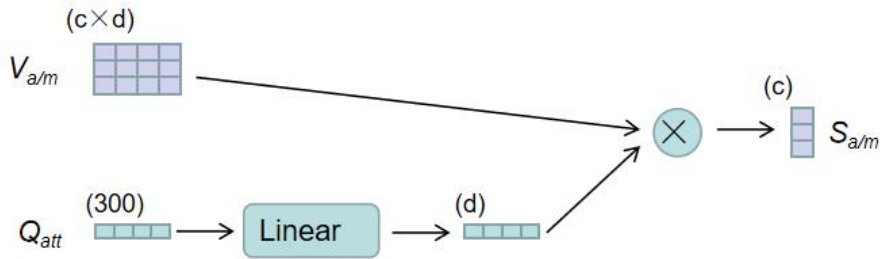


图 3.5 问题导向的视觉注意力

该模块对于视频的外观特征和动作特征处理的步骤相同, 接下通过 a/m 对外观特征和动作特征进行统一表示, 处理步骤如图 3.5 所示。首先将上一步得出的整个问题的注意特征通过一层线性层映射到模型维度, 使其特征维度与视频相同, 接着将其与视觉特征进行矩阵点积运算后, 经过 Sigmoid 函数激活便得到了视觉特征各个片段的得分 $S_{a/m}$, 外观和动作特征处理的统一的公式表示为:

$$S_{a/m} = \text{Sigmoid}(V_{a/m} Q_{att} W_{a/m}) \quad (3.2)$$

其中, $W_{a/m}$ 分别表示外观特征和动作特征的可训练参数。这里可以得到以问题为导向的视频各个片段注意力得分, 为 $S_{a/m} \in \mathbb{R}^c$ 。

3.2.3 图卷积

为了推理视频中各个片段间的关系, 同时挖掘更深层的特征表达, 本方法借鉴了 DualVGR^[29]的思想, 对图卷积网络进行建模, 其结构如图 3.6 所示。

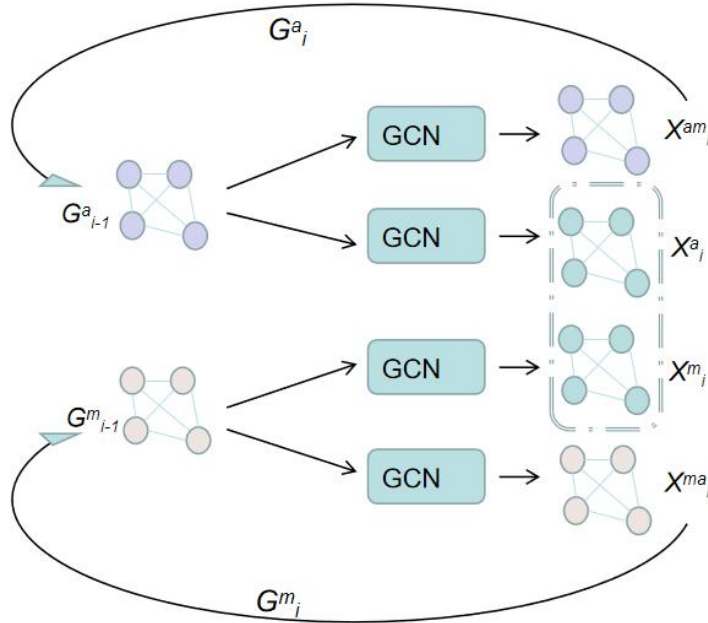


图 3.6 图卷积网络结构

DualVGR 在关系推理过程中, 综合采用了 GAT 的图卷积网络模型以及 AM-GCN^[30]的思想。GAT 采用多头图卷积（multi-head GCN）的策略, 并融合了注意力的机制在关系推理过程中, 以实现对节点之间关系更有效的表征。在教师模型中, 为了更充分的使用图卷积挖掘信息, 本方法采用了多层图卷积的方式, 而在学生模型中, 为了达到压缩模型, 减少可训练参数的目的, 仅使用了单层的图卷积进行训练。同时, 参考 AM-GCN 的思想, 为了提取外观特征和动作特征之间潜在的关联关系, 本方法不仅提取了外观的独有特征, 动作的独有特征, 同时通过损失函数约束的方式, 得到了在外观中融合动作信息的关联特征, 以及在动作中融合外观信息的关联特征。本方法通过损失函数约束来使独有特征和关联特征之间差异明显, 而关联特征之间表征相似。损失函数的详细计算过程将会在 3.6 节进行详细介绍。同样, 该模块对于外观和动作特征处理的步骤相同, 接下通过 a/m 对外观特征和动作特征进行统一表示。

首先，将视觉特征 $V_{a/m}$ 作为图卷积网络的输入，在通过 GCN 进行处理后分别得到了外观的独有特征 X_a 和关联特征 X_{am} 以及动作的独有特征 X_m 和关联特征 X_{ma} ，接着将视觉的独有特征 $X_{a/m}$ 作为下一层图卷积网络的输入，迭代计算 g 次， g 为图卷积的层数。其公式可以表示为：

$$\begin{aligned} G_0^{a/m} &= V^{a/m} \\ X_i^{a/am} &= GCN_i^{a/am}(G_{i-1}^a, S^a) \\ X_i^{m/ma} &= GCN_i^{m/ma}(G_{i-1}^m, S^m) \\ G_i^{a/m} &= X_i^{a/m} \end{aligned} \quad (3.3)$$

其中， G_{i-1} 表示第 i 层图卷积的输入特征， X_i 表示第 i 层图卷积的输出特征，下标 am 表示外观中融合动作信息的关联特征，下标 ma 表示动作中融合外观信息的关联特征， GCN_i 表示第 i 层图卷积操作， i 小于等于 g 。

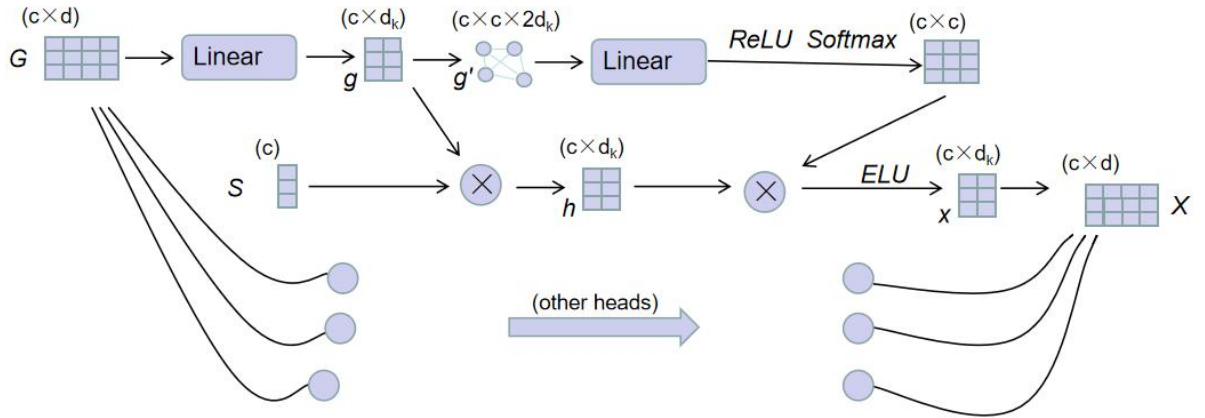


图 3.7 多头图卷积

在每一层的 GCN 中，其通过了 k 个图卷积头进行处理，最后将各个头的结果进行拼接。而这里对于外观或动作的独有或联合特征来说，处理步骤完全相同，简洁起见，忽略了对应特征的表示，其步骤如图 3.7 所示。具体来说，首先，将输入特征 G 通过一层线性层映射为 d_k 维后，将其作为该图卷积头所要处理的视觉特征 g ，其公式为：

$$g_j = GW_j \quad (3.4)$$

其中， G 为当前层图卷积的输入特征， $W_j \in R^{d \times d_k}$ 为可学习参数，且 j 小于等于 k ， k 为图卷积头的个数， d_k 大小为 d 和 k 的比值。接着通过将 g_j 与视觉注意分数 S 进行相乘的

方式，改变各个片段的特征的相对大小，达到对重要片段重点关注的目的，得到了注意力下的视觉特征 h_j ，其公式为：

$$h_j = g_j \otimes S \quad (3.5)$$

其中， S 为前文计算的视觉注意力分数， \otimes 表示矩阵对应位相乘的运算操作。同时，为了跟好地表示视频各个片段之间的关系，将各个片段作为节点，将两个片段之间的特征值拼接后组为边，把 g 构造成为了一个无向全连接的图 g' 。接着使用注意力的机制，将图特征 g' 经过一层线性层后，通过 ReLU 激活函数后再使用 Softmax 进行归一化，得到了每两个节点之间关系的权重大小 β_j ，其公式表示为：

$$\begin{aligned} g'_j &= \text{Graph}(g_j) \\ \beta_j &= \text{Softmax}(\text{ReLU}(g'_j W)) \end{aligned} \quad (3.6)$$

其中，Graph 表示构造全连接无向图的操作， $W \in R^{d_k \times l}$ 为可学习的参数。再接着，将该 β_j 与视觉特征 h_j 相乘，以通过各个片段之间的关系权重大小来得到每个片段基于全部片段的关系所得到的特征，并通过 ELU 激活函数得到该图卷积头的最终输出 x_j ，其公式表示为：

$$x_j = \text{ELU}(\beta_j h_j) \quad (3.7)$$

最后将多头图卷积得到的每个特征 x_j 进行拼接，就得到了该层图卷积的结果特征 X 。

在学生模型中，由于采用的是单层图卷积的方式，该结果便可以直接作为最终的关系特征；而对于教师模型，在多层图卷积策略下，还需要经过多轮的迭代计算，才可以得到最终的关系特征。而最终，可以得到了通过图卷积推理关系特征，而提取的外观的独有特征 C_a 和关联特征 C_{am} ，以及动作的独有特征 C_m 和关联特征 C_{ma} 。

3.3 视觉融合模块

3.3.1 独有-关联融合

在对视觉特征进行融合时，为了更好地突出视觉特征中独有特征和关联特征各自的重要性，类似地，本方法采取了注意力的机制来进行独有-关联特征融合。同样，该模块对于外观和动作特征处理的步骤相同，简洁起见，接下来只对外观特征的处理进行讨论，其过程如图 3.8 所示。

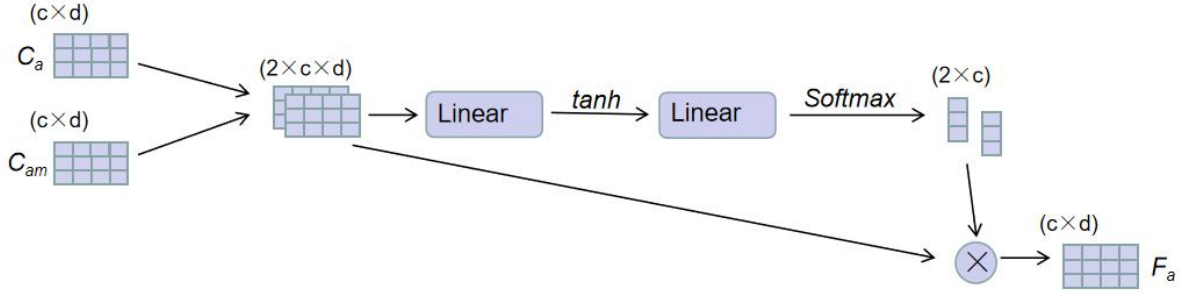


图 3.8 独有-关联融合

首先将外观的独有特征 C_a 和联合动作 C_{am} 拼接后为 $[C_a, C_{am}]$ ，通过一层线性层以及 \tanh 激活函数后，再通过一层线性层将其映射为对于两类特征的注意力分数。接着通过 Softmax 激活函数进行归一化处理后，将其作为权重对两类特征进行加权求和，便得到了融合的外观特征 F_a ，其公式表示为：

$$\begin{aligned} [\alpha_a, \alpha_{am}] &= \text{Soft max}(\tanh([C_a, C_{am}]W_1)W_2) \\ F_a &= \alpha_a C_a + \alpha_{am} C_{am} \end{aligned} \quad (3.8)$$

其中， $W_1 \in \mathbb{R}^{d \times d}$ 和 $W_2 \in \mathbb{R}^{d \times 1}$ 均为可学习参数，得到的最终的外观特征 $F_a \in \mathbb{R}^{c \times d}$ 。同理，将动作的独有特征和联合特征进行融合，可以得到了融合的动作特征 F_m ，其公式表示为：

$$\begin{aligned} [\alpha_m, \alpha_{ma}] &= \text{Soft max}(\tanh([C_m, C_{ma}]W_3)W_4) \\ F_m &= \alpha_m C_m + \alpha_{ma} C_{ma} \end{aligned} \quad (3.9)$$

其中， $W_3 \in \mathbb{R}^{d \times d}$ 和 $W_4 \in \mathbb{R}^{d \times 1}$ 均为可学习参数，得到的最终的动作特征 $F_m \in \mathbb{R}^{c \times d}$ 。

至此，模型得到了表征视频各个片段之间的关系信息的特征 F_a 和 F_m 。所以为了得到综合的包含关系信息的视觉特征，同时防止训练时反向传播过程中梯度消失的问题，本方法采用了残差连接^[31]的策略。将基于关系推理过程的结果特征与最初提取的视觉特征进行相加，得到了最终的融合了关系信息的视觉特征 $V_{a/m}$ ，其公式为：

$$V_{a/m} = V_{a/m} + F_{a/m} \quad (3.10)$$

3.3.2 外观-动作融合

该模块将视觉外观特征和视觉动作特征进行了融合，本方法采取了多模态因子双线性池化^[32] (Multimodal Factorized Bilinear pooling, MFB) 的方法，得到了视频每个片段地融合视觉特征，其结构如图 3.9 所示。

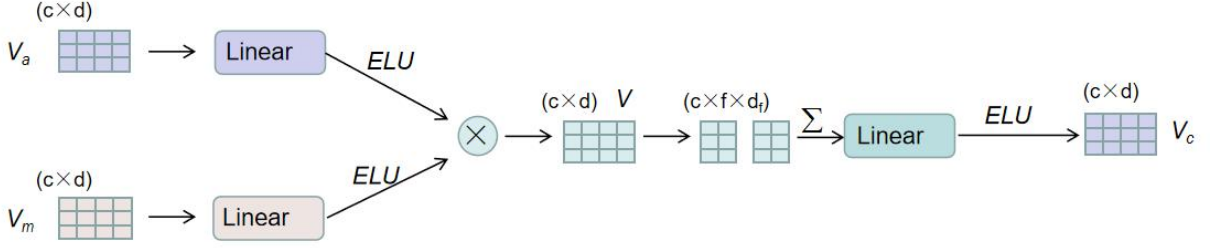


图 3.9 外观-动作融合

首先将外观特征 V_a 和动作特征 V_m 分别通过一层线性层处理后,再经过 ELU 激活函数处理,并将两个特征通过对应位乘积以得到整个的视觉特征表示 V 。接着为了挖掘特征之间潜在的联系,将 V 通过因子分解得到 f 个因子矩阵 V_i 。最后通过将各个因子矩阵求和以及一层线性层映射后,得到了表示每个片段的外观-动作融合视觉特征 $V_c \in \mathbb{R}^{c \times d}$, 其公式表示为:

$$\begin{aligned}
 V &= ELU(V_a W_a) \otimes ELU(V_m W_m) \\
 [V_1, V_2, \dots, V_f] &= V \\
 V_c &= (\sum_i^f V_i) W
 \end{aligned} \tag{3.11}$$

其中, $W_a \in \mathbb{R}^{d \times d}$ 、 $W_m \in \mathbb{R}^{d \times d}$ 和 $W \in \mathbb{R}^{d_f \times d}$ 均为可学习的参数, \otimes 表示矩阵对应位相乘的运算操作, 且 i 小于等于 f , d_f 大小为 d 和 f 的比值。

3.3.3 片段融合

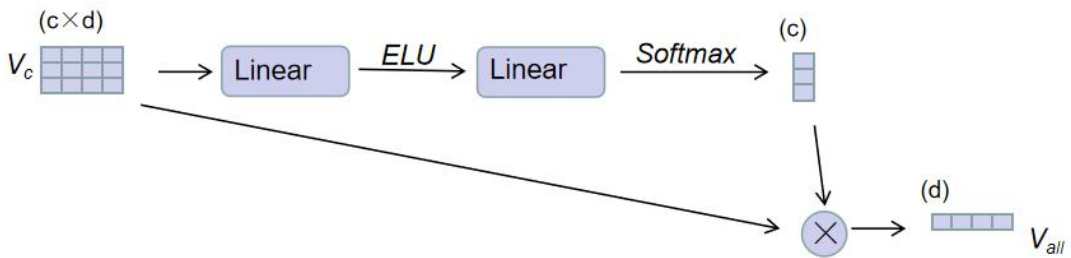


图 3.10 片段融合

该模块将视频中的每一个片段特征进行了融合,本方法借鉴了图读出操作^[33](Graph Readout Operation)中的方法,通过自注意力的机制得到整个视频的融合特征表示,其过程如图 3.10 所示。

具体来说，首先将每个片段的视觉特征 V_c 通过一层线性层以及 ELU 激活函数后，再通过一层线性层将其映射为自身对于各个片段的注意力分数，并通过 Softmax 激活函数进行归一化处理。接着将其作为权重将各个片段特征进行加权求和，得到了每个视频的融合片段特征 V_{all} ，其公式表示为：

$$\begin{aligned}\beta &= \text{Soft max}(ELU(V_c W_1) W_2) \\ V_{all} &= \beta^T V_c\end{aligned}\quad (3.12)$$

其中， $W_1 \in \mathbb{R}^{d \times d}$ 和 $W_2 \in \mathbb{R}^{d \times 1}$ 均为可学习的参数。

3.4 答案生成模块

至此，模型得到了与文本交互后的融合视觉特征，该特征便是在机器在该问题下所理解的视频信息，现在只需要通过该视频信息回答问题。这里本方法通过融合视觉特征和问题语义特征，解码最终用于生成答案的特征。模型采用了一种比较通用的解码器来对答案进行解码，其结构如图 3.11 所示。

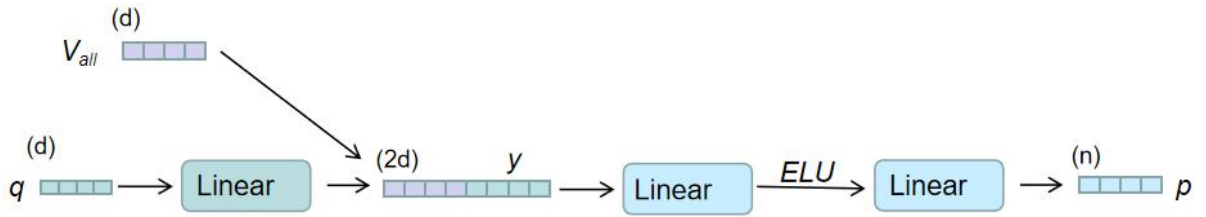


图 3.11 答案生成

首先，将语义特征 q 通过一层线性层后，与视觉特征 V_{all} 进行拼接，得到融合特征 y 。接着使用分类器对融合特征进行处理，其先经过了一层线性层映射为模型维度后，使用 ELU 激活函数处理，再经过一层线性层将特征维度映射为答案分类数目，得到了最终用于生成答案的特征 p ，其公式表示为：

$$\begin{aligned}y &= [V_{all}, q W_1] \\ p &= ELU(y W_2) W_3\end{aligned}\quad (3.13)$$

其中， $W_1 \in \mathbb{R}^{d \times d}$ 、 $W_2 \in \mathbb{R}^{2d \times d}$ 和 $W_3 \in \mathbb{R}^{d \times n}$ 均为可学习的参数， n 为答案分类数目。

3.5 多模态知识蒸馏架构

知识蒸馏的机制目前来主要应用于模型的压缩方面。该机制首先训练一个超大型的教师模型，其通过大量的参数学习来达到更好的效果。之后通过构建一个轻量级的学生

模型，其不仅学习训练集真实标签的知识，同时还学习大型的教师模型训练后所蒸馏出来的知识，以更小的参数体量来逼近大型的教师模型的预测效果，达到模型压缩的目的。

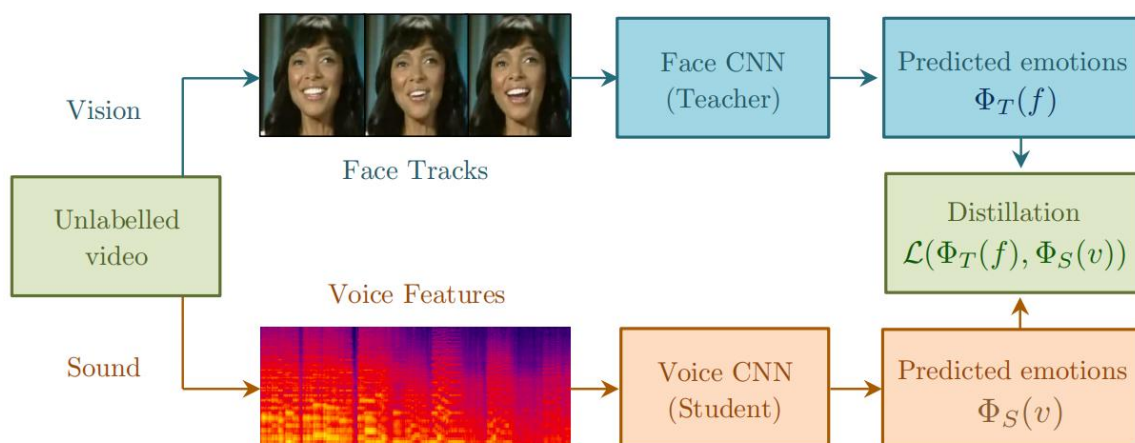


图 3.12 情感识别中的跨模态蒸馏^[46]

而除了模型压缩的效果之外，知识蒸馏的机制还可以实现用性能较差的教师模型教导出性能更好的学生模型^[34]，以此达到了模型增强的目的。**Wang**^[35]等人便是在动作识别任务中，通过完整的视频知识对早期的动作特征进行指导，以改善初期的特征学习效率。知识可以对特征的学习过程进行优化，使得学生模型具有更强的学习能力并获得更好的性能^[36]。不仅如此，在面对多模态的数据时，可以通过各个模态的信息来实现多模态蒸馏。例如 **Albanie**^[37]等人所提出的跨模态情感识别方法，如图 3.12 所示。该方法先训练了强大的教师模型用于图像模态的面部情绪识别，接着用教师模型指导学生模型对音频模态的演讲情绪识别的表征学习。通过对同步的图片模态信息和音频模态信息之间的知识蒸馏，来实现跨模态的知识传递。

基于前人的研究与应用，本方法使用了一种多模态的知识蒸馏方法。将教师模型中多模态融合后的知识蒸馏出来，用于学生模态中单模态的学习，使得学生模型可以在单模态训练过程中得到教师模型丰富的多模态知识，在训练初期更早地进行多模态之间的交互与融合，以改善之后多模态融合的效果。该方法不仅实现了模型体量的轻量化，而且优化了多个模态特征学习的能力，提升了模型性能。

正如前文所说，在教师模型中，模型使用了多层的图卷积对视觉特征进行了迭代计算，这是因为模型只有在迭代进行多次的图卷积计算后，才可以更好地实现视频关系的多步推理，提取关系特征。而在学生模型的构造上，只使用了单层的图卷积来处理视觉特征，以轻量化模型的体量，压缩模型。而在训练时，本方法使用了教师模型多模态融

合后的知识用于指导学生模型单模态的训练，以优化学习过程，提升性能。本方法的教师-学生的训练架构如图 3.13 所示。

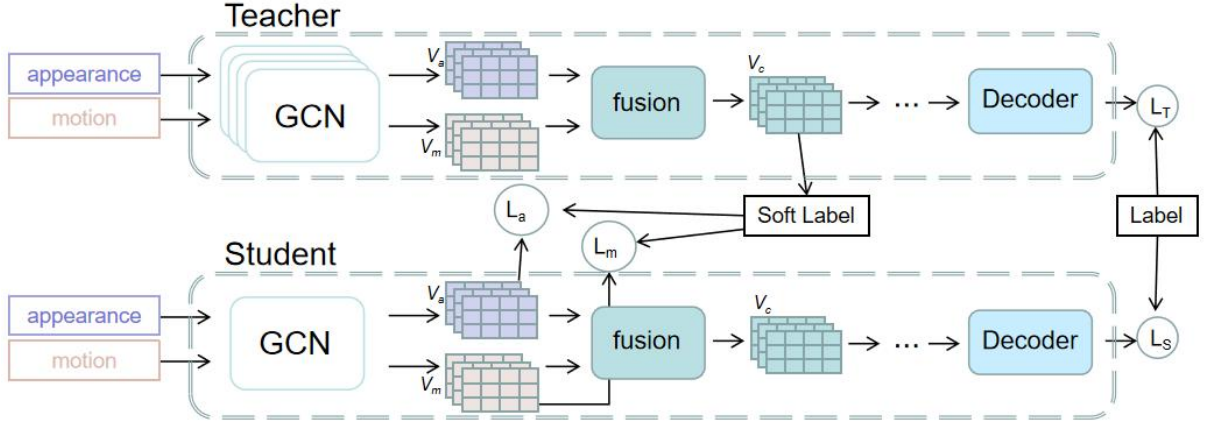


图 3.13 知识蒸馏结构

首先本方法完整训练一个教师模型，并通过实验进行参数调节，得到了训练得较好的教师模型。接着在学生模型训练时，将教师模型中外观视觉特征与动作视觉特征融合后的融合视觉特征，分别用于学生模型的外观和动作的视觉特征的学习。具体来说，通过添加损失函数的方式，将教师模型融合特征 V_c 中的知识蒸馏出来指导学生模型的外观特征 V_a 和动作特征 V_m 的学习。损失函数的计算过程将会在下节进行详细介绍。这样使得学生模型在外观或动作的单模态处理过程中便可以学习到其它模态的知识，以更好的进行模态间的交互与融合，同时以此来弥补图卷积迭代次数减少所带来的信息提取不充分的不足。

3.6 损失函数

3.6.1 教师模型

在模型的视觉-文本交互模块中，正如前文所提到的，对外观或动作视觉模态的学习时，为了其不仅能学习到单模态自身独有的知识，同时也希望可以学习到与另一个模态相关联的知识。于是在 3.2.3 小节处理外观视觉特征和动作视觉特征时，在每一层的图卷积中分别生成了四种特征，即包括外观的独有特征 G_a 和关联特征 C_{am} ，以及动作的独有特征 G_m 和关联特征 G_{ma} 。显然，独有和关联特征是通过相同的图卷积操作而得到的，那么所学到的特征有一定的重复。模型需要外观或动作模态上所学习到的独有特征和关联特征之间相差较大，而关联特征之间相差较小。所以本方法参考了 DualVGR 中

所提到的方法，对每一层的特征使用了差异约束和相似约束的机制，以此充分地提取视觉特征的自身信息和与其它模态相关联信息。

(1) 差异约束。为了使外观或动作的独有特征和关联特征之间存在明显差异性，以提取特征中不同的信息，更好地对两类特征进行表示，本方法使用了希尔伯特·施密特独立准则^[38] (Hilbert-Schmidt Independence Criterion, HSIC) 来进行约束。结合在这里的应用，该约束可以定义如下：

$$H = I - \frac{1}{n}ee^T$$

$$HSIC(X, Y) = (n-1)^{-2} \text{tr}(H(XX^T)H(YY^T)) \quad (3.14)$$

其中， H 为常量，由单位矩阵 I 和值全为 1 的行向量 e 计算得到， $X \in \mathbb{R}^{n \times d}$ 和 $Y \in \mathbb{R}^{n \times d}$ 为输入的两个分布。通过该公式，模型分别对每一层图卷积的 G_a 和 G_{am} 以及 G_m 和 G_{ma} 进行了差异约束，以此得到的损失计算公式为：

$$L_l = \frac{1}{g} \sum_l^g (HSIC(G_i^a, G_i^{am}) + HSIC(G_i^m, G_i^{ma})) \quad (3.15)$$

(2) 相似约束。同样，挖掘外观和动作之间的关联特征，那么所提取的关联特征之间一定是相似的，所以有必要对在两类特征空间下所挖掘的关联特征进行相似约束。这里本方法采取了一种矩阵距离的方法，首先将两个输入分布进行标准化处理，接着计算输入分布的点积运算，将两个输入的运算结果进行差值求模后，得到的数值便作为两个输入分布之间相似的损失。显然，如果两个输入分布相同，其结果为 0。该损失的计算公式如下：

$$L_2 = \frac{1}{g} \sum_l^g \|G_i^{am}(G_i^{am})^T - G_i^{ma}(G_i^{ma})^T\|_2 \quad (3.16)$$

同时，在模型的输出预测部分，先将输出使用 Softmax 激活函数归一化处理后，得到在答案集上的概率分布。使用交叉熵损失函数来对预测的概率分布和真实标签进行损失计算，其公式为：

$$L_T = -\sum z \ln y \quad (3.17)$$

其中， y 表示预测的概率分布， z 表示真实的概率分布。

综上所述，教师模型的损失函数最终可以表示为：

$$L_{total}^T = L_T + \gamma L_l + \eta L_2 \quad (3.18)$$

其中系数 γ 和 η 代表超参数，通过调节对模型进行优化。

3.6.2 学生模型

学生模型除了包含教师模型中所提到的损失 L_0 , L_1 和 L_2 之外，还包括了模型学习教师模型的蒸馏知识的损失。本小节只对知识蒸馏损失进行介绍，其它损失详见上一小节。

正如前文所提及的，为了在单模态学习的早期，希望可以学习到其他模态的知识，本方法将教师模型融合特征 V_c 中的知识蒸馏出来指导学生模型的外观特征 V_a 和动作特征 V_m 的学习，以达到改善各个模态间的交互与融合的目的。其先在适当温度 T 下通过 Softmax 激活函数进行归一化处理后，再使用交叉熵进行损失计算。该计算过程对于外观和动作特征的步骤相同，接下通过 a/m 对外观特征和动作特征进行统一表示。其损失函数为：

$$L_{a/m} = L_0(\text{Soft max}(V_c^t / T_{a/m}), \text{Soft max}(V_{a/m}^s / T_{a/m})) \quad (3.19)$$

其中， L_0 表示交叉熵损失计算操作， V_c^t 为教师模型的融合特征， $V_{a/m}^s$ 为学生模型的外观和动作特征， $T_{a/m}$ 为外观和动作进行知识蒸馏时的温度。

所以，学生模型的总损失可以表示为：

$$L_{total}^S = L_T + \gamma L_1 + \eta L_2 + \lambda_a L_a + \lambda_m L_m \quad (3.20)$$

其中系数 γ 和 η 直接使用教师模型的系数， λ_a 和 λ_m 为超参数，通过调节对模型进行优化。

4 实验

4.1 数据集

本方法选取了 MSVD-QA^[39]数据集进行实验，其视频内容是从微软研究视频描述语料库（Microsoft Research Video Description, MSVD）中所提取的，视频以日常生活的短视频为主，而问题则是通过程序根据视频的描述内容所生成的。



图 4.1 MSVD-QA 示例

该数据集由视频和对应的问答对组成，以小尺度的短视频和短问题为主，视频的平均时长为 10 秒左右，问题的平均长度大约为 6 个单词，每个视频平均有 25 个问题左右，该数据集的示例可见图 4.1。

表 4.1 MSVD-QA 数据集统计表

	视频	问题	What	Who	How	When	Where
训练集	1200	30933	19485	10469	736	161	72
验证集	250	4615	3995	2168	185	51	16
测试集	520	13157	8149	4552	370	58	28
合计	1970	50505	31629	17199	1291	270	116

表 4.2 MSVD-QA 数据集问题占比

	What	Who	How	When	Where	总计
数目	31629	17199	1291	270	116	50505
占比	62.6%	34.1%	2.6%	0.5%	0.2%	100%

问题的提问为开放式类型，其包括了 what、where、how、who 和 when 五种提问形式。MSVD-QA 数据集总共包含了 1970 个视频和 50505 个问答对，其训练集、验证集和测试集的详细数量见表 4.1，各个提问形式的占比可见表 4.2。可以看出五类问题的占比差别较大，极不均衡。主要为 what 和 who 类型的问题，总占比超 96%，what 类型超过 62%，平均每一个视频有 16 个 what 问题。而 how、when 和 where 类型极少，加起来一共才占 3.3% 左右，很多视频都不存在这三种提问形式。

该数据集中视频均取材于日常生活，与我们平时所接触的大部分视频类似，同时提问方式为开放式的类型，同样与人类自然的提问方式相接近。所以该视频问答数据集可以有效地考量一个机器对视频及问题的理解能力，并应用到我们日常生活中。但是，该数据集仍有很多不足，其问题是通过程序的方式自动生成，难免会造成语义错误等不正确的回答。同时从图 4.1 中的例子中，问题 3、4 和 5 意思均是想表达“what is a cat doing?” 的问题，然而答案却给出了三个不同的回答，分别是“imitate”、“watch”和“mimic”，这三种答案都是该问题的正确答案，但是在最后评价时中却只能对一个，这一定程度上降低了对模型评价的精准程度。

4.2 评价指标

准确率 Accuracy 通常是分类任务的评价指标，所以使用该指标来对该视频问答模型的预测性能进行评估，其为模型对测试集的预测正确的数量与总的测试集大小的比率，计算公式为：

$$Accuracy = \frac{1}{n} \sum_{i=1}^n (a_i \circ p_i) \quad (4.1)$$

其中， n 为数据集问答对的个数， a 表示为正确答案， p 表示为预测答案， \circ 为运算符，当两个值相等结果为 1，否则为 0。

同时，本方法中通过知识蒸馏的方式实现了对模型的压缩，而模型的大小通常使用该模型需要训练的参数来进行衡量，所以模型可训练参数数量 Parameter 也考虑在评价指标当中。

4.3 训练细节

本方法使用了 pytorch 深度学习框架对模型进行建模以及实验。在模型的各种超参数的设置上，首先确定超参数选取的数值范围，接着通过大量实验，使用了网格搜索的方法，寻找最优的参数配置，模型最终的参数设置如下。

4.3.1 教师模型

在视频的预处理阶段，对于每一个视频，将其等间隔切分的片段数 c 为 8，而每个片段包含的图像帧数 f 为 16。对于向前或向后不够 8 帧时，将最开始的帧或最末尾的帧作为填补。

模型的维度大小 d 为 768。在编码模块中，视觉编码和文本编码用来提取上下文信息和语义信息的 BiLSTM 均为单层。在视觉-文本交互模块中，图卷积的层数 g 为 4，图卷积的图卷积头数 k 为 4。在视觉融合模块中，多模态因子双线性池化中的因子数 f 为 4。

损失函数中，独有和关联特征的约束损失的系数 γ 和 η 分别设置为 100 和 $1e-6$ 。训练过程中所使用的优化器为 Adam optimizer，训练的学习率设置为 $1e-4$ ，分批训练数据的大小为 256，训练迭代次数为 25。

4.3.2 学生模型

在学生模型中，为了跟好得对比分析知识蒸馏所带来得效果，模型设置了除了图卷积层数之外与教师模型完全相同得参数配置。使用单层图卷积处理视觉特征，以轻量化模型的体量而压缩模型。

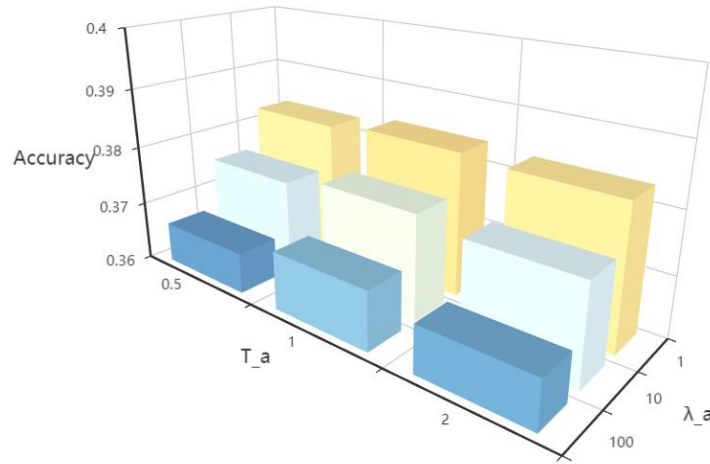


图 4.2 外观特征蒸馏参数搜索

而在知识蒸馏的损失函数中，先通过一些探索性的实验来确定参数大概的搜索范围。接着对参数进行网格搜索实验，确定相对较好的蒸馏参数配置。对于外观视觉特征的知识蒸馏损失，其调参实验如图 4.2 所示，最终温度 T_a 设置为 1，系数 λ_a 为 1；对于动作视觉特征的知识蒸馏损失，其调参实验如图 4.3 所示，温度 T_m 设置为 0.7，系数 λ_m 为 100。

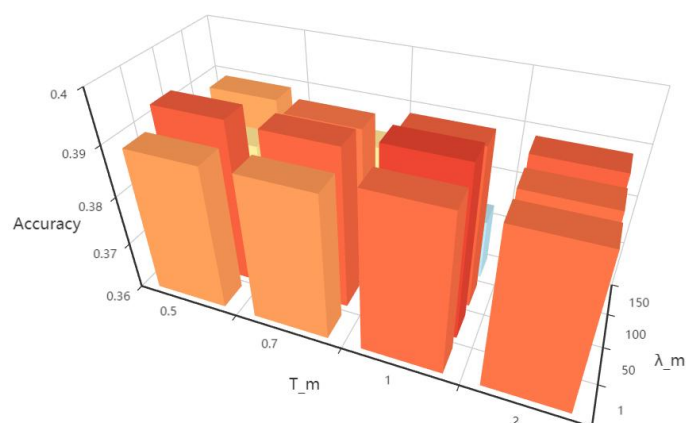


图 4.3 动作特征蒸馏参数搜索

4.4 实验结果与分析

4.4.1 可视化分析

在通过以上的参数设置进行实验时，对模型的训练过程中准确率和损失的变化进行了可视化分析。

训练过程中训练集和验证集的准确率变化如图 4.4 所示。可以看出，训练集在不断的参数迭代学习下，准确率增长迅速且升幅较大，其在第 25 轮训练时已经接近了 0.8 的准确率。而验证集上的准确率却上升缓慢且升幅较小，在第 5 轮左右准确率达到 0.35 左右后，上下波动，未呈现出平稳的上升趋势。

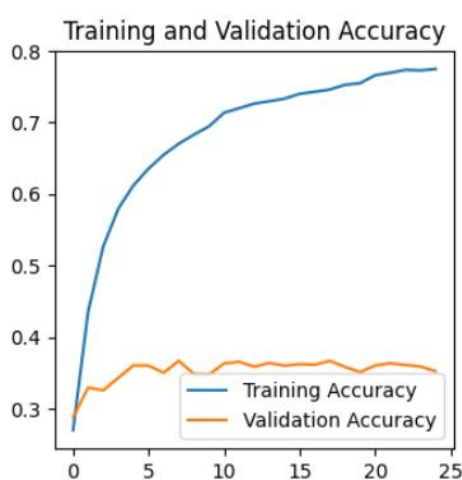


图 4.4 训练集与验证集的准确率变化

训练过程中训练集和验证集的损失变化变如图 4.5 及图 4.6 所示，其分别表示了真实标签损失、外观特征蒸馏损失和动作特征蒸馏损失。可见，在不断的参数迭代学习下，训练集和验证集的各个损失均在不断地下降收敛，可见训练的过程是有效的。但是，训练集各损失的下降相对更加迅速，而验证集各损失的下降相对更加缓慢，且具有一定波动性。尤其是验证集的真实标签损失下降过程，虽有下降趋势，但极其缓慢，且不断波动幅度较大。

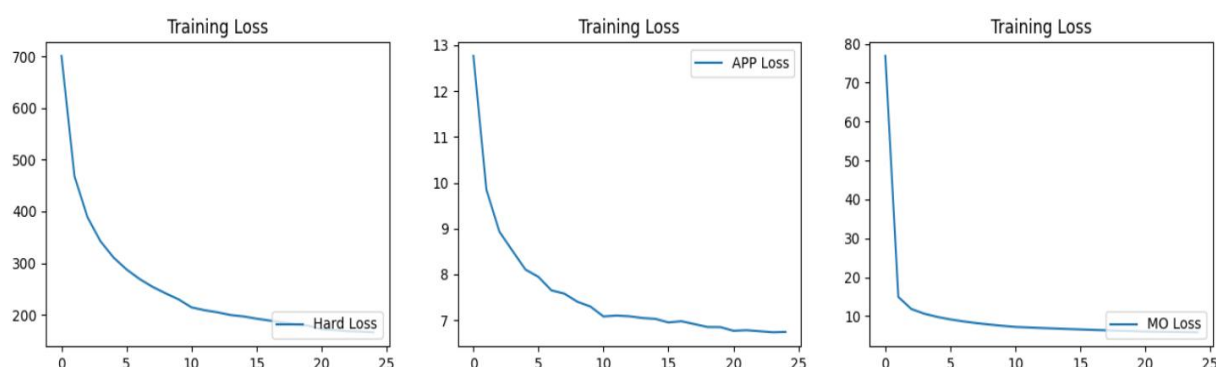


图 4.5 训练集损失变化

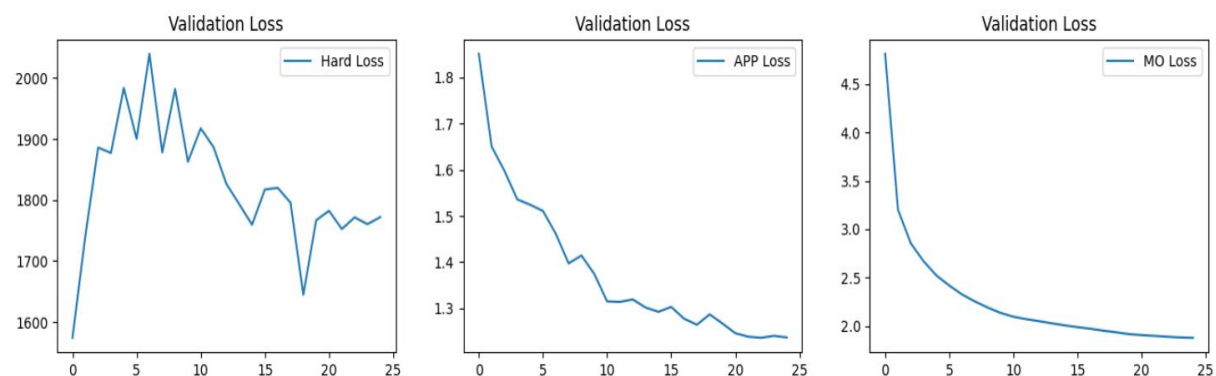



图 4.6 验证集损失变化

综合训练集和测试集的准确率和损失变化来看，在训练过程中，模型存在着一定的过拟合现象。训练集的准确率在 25 轮学习中不断攀升，而验证集的准确率过早地在第 5 轮达到一定高度后增长缓慢，反复波动，并且相对应的，其损失的下降过程同样非常缓慢且不断波动。

接着，为了进一步对模型的预测性能进行分析，从模型的预测结果中挑选了如图 4.7 所示的部分错误预测的样例。针对模型预测错误的情况来讨论模型在视频问答任务上的不足与改进。



input question	answer	prediction
1. what <UNK> down on a <UNK> running in a field?	<UNK>	bear
2. what did the bird of prey attack?	rabbit	bear
3. what attacked the white rabbit?	bird	chase
4. what does a rabbit flee from?	eagle	mud
5. what is an eagle doing?	try	chase

图 4.7 错误预测样例

问题 1 中可以看到，由于单词表中的单词数量有限，部分单词需要用<UNK>进行表示，这就导致了模型对某些单词无法理解其语义，从而无法对答案进行准确的预测。问题 2 中模型对“兔子”预测错误，误以为是“熊”，可能是棕色的背景与熊的颜色非常类似所导致的，这说明该模型在目标识别上仍然不够精确。问题 3 中在询问目标对象时，模型反而给出了一个动词“追逐”，这显然说明模型对句子语义的理解仍然不够准确。问题 4 和问题 5 所出现的错误可能是数据集不够完善的原因导致的，因为所预测的答案对我们人类来说是可以认为是正确的答案，但由于数据集答案的固定性，其无法将意思上正确的答案视为回答正确。

4.4.2 对比分析

为了分析验证本方法在视频问答任务中的有效性，对比了较为前沿的基于视频问答任务的模型算法。接下来首先对这些算法进行简单介绍，后将实验结果进行对比分析。

(1) ST-VQA^[40]。该算法主要提出了使用空间-时间注意力的机制对视频信息进行理解。其首先通过文本语义对视频进行空间上的注意力，即重点关注每帧图像中的某部分区域；接着通过文本编码再对视频进行时间上的注意力，即重点关注视频的某几帧图像，得到结合文本的空间-时间推理特征。

(2) Co-Mem^[18]。该方法由视觉问答中的动态记忆网络（Dynamic Memory Network, DMN）发展而来，基于视频问答任务进行了改进。其再情景记忆模块中引入外观-动作协同记忆力的注意力机制，同时运用了基于时序的卷积-反卷积网络以及动态的事实集成方法，充分挖掘视频信息。

(3) AMU^[3]。该算法为端到端的视频问答模型，应用了问题中的细粒度特征进行视频理解。其通过逐词的读取问题中的单词，通过注意力机制进行单词和外观特征及动作特征进行交互，不断地细化视频注意力特征，最终得到融合了问题不同尺度特征的视频理解。

(4) HME^[41]。该算法同样由记忆网络发展而来。其通过自更新的注意力来进行视觉和文本结合的多步推理。并在每一步中，使用问题记忆来学习问题的上下文信息，动态优化问题的表征。且提出了一种异构记忆方法，结合外观特征和动作特征进行时空上的注意力来优化视频的表征。

(5) HGA^[22]。该模型引入图网络进行推理学习。其将视频剪辑和问题单词均构造成图的形式，进行跨模态的图推理学习过程。

(6) HCRN^[20]。这是一种基于片段的关系网络模块的堆叠式模型。该关系网络将输入作为一组张量对象和一个条件特征，输出了一组包含它们的关系信息，接着通过分层地堆叠该网络模块，实现对关系信息的多步推理。

(7) TSN^[42]。该算法通过综合外观特征和动作特征进行同步推理，以挖掘外观和动作之间存在的关联信息。其首先通过混合模块来实现每时间片上对外观和动作时序上对齐的同步推理，并通过转换模块自适应地选择外观或动作特征作为主要特征指导推理过程。

(8) DualVGR^[29]。该模型是一种注意力图推理网络的堆叠式模型。注意力图推理网络模块中通过查询惩罚机制实现对视频关键片段的特征强化，接着通过结合注意力的多视图的图网络对关系进行建模，模型通过堆叠该网络模块进行关系信息的多步推理。

表 4.3 各模型准确率对比

模型	What	Who	How	When	Where	All
ST-VQA	18.1	50.0	83.8	72.4	28.6	31.3
Co-Mem	19.6	48.7	81.6	74.1	31.7	31.7
AMU	20.6	47.5	83.5	72.4	53.6	32.0
HME	22.4	50.1	73.0	70.7	42.9	33.7
HGA	23.5	50.4	83.0	7.4	46.4	34.7
HCRN	/	/	/	/	/	36.1
TSN	25.0	51.3	83.8	78.4	59.1	36.7
DualVGR	28.7	53.8	80.0	70.7	46.4	39.0
Ours	29.22	53.98	80.81	74.14	53.57	39.48

表 4.3 总结了各个模型在 MSVD-QA 数据集上的实验结果的对比。通过对比分析可以看出，本方法在最终的准确率上高于前文所列的对比模型。其中，what 和 who 形式的问题上，本方法准确率均为最高，而在 how、where 和 when 的提问形式上，准确率仍有差距。但考虑到该数据集五类问题类型分布极其不均匀，what 和 who 总占比超过 96%，可见在这些主要提问形式上，模型得到了充分的训练，准确率相对更高；而对于 how、where 和 when 的提问形式，其数量极少，模型未能充分地学习小样本的问答对，最终准确率略低与其它模型。并且，本方法的准确率在各个指标上都超过了所参考的 DualVGR 模型，可见在相同架构下，方法对模型进行了有效地优化。

总体上，本方法的模型在 MSVD-QA 数据集上相较于其它对比模型来说，总性能确实得到了适当的提升，得到了很好的预测结果。可见，知识蒸馏其不仅可以轻量化模型体量，同时在用于跨模态的信息传递与融合以增强对单模态的特征提取能力时，对模型的性能也是有一定提升的。

4.4.3 消融研究

在本方法中，主要提出了通过知识蒸馏的方式对模型进行压缩的同时，使强化跨模态的特征学习与融合，以达到改善模型的目的。为了验证知识蒸馏的有效性，通过了消融研究来进行分析。

表 4.4 消融研究

模型	准确率 (Accuracy)	可训练参数数量 (Parameter)
Teacher	39.03%	3119 万
Student	38.85%	2409 万
Student-kd	39.48%	2409 万

首先所构建的教师模型 Teacher 中可学习参数包含了大约 3119 万参数，本方法通过训练教师模型并调优参数，以达到教师模型的最优参数配置，最终其在测试集上所达到的准确率为 39.03%。接着，本方法通过减少图卷积的层数来构造相对轻量的学生模型，其可学习的参数降低至约 2409 万。在其它参数配置与教师模型完全相同情况下，本方法首先对学生模型进行单独训练，这里将其表示为 Student，其在没有教师模型的指导下训练完成后在测试集上所达到的准确率为 38.85%。接着，通过多模态知识蒸馏的方法，让教师模型对学生模型进行指导训练，并调优知识蒸馏参数。而通过知识蒸馏机制

进行训练的学生模型 Student-kd，在测试集上准确率达到了 39.48%。整个消融研究的结果见表 4.4。

Student 和 Student-kd 的测试结果之间进行对比可以发现，它们两个模型架构完全相同，可训练的参数量完全相同，但是在通过知识蒸馏学习了教师模型的跨模态特征下的 Student-kd 展现了更加高的准确率，可以推断知识蒸馏的方式有效地对模态间融合的效果进行了提升。而通过对比 Teacher 和 Student-kd 的测试结果可以发现，知识蒸馏的作用下，Student-kd 在准确率没有降低反而有少许增高的同时，大幅轻量化了模型的大小，减少了原来可训练参数数目近 23%的参数。由此我们可以看出，在本方法中，知识蒸馏用于跨模态的信息传递时，其不仅可以轻量化模型体量，减少可训练参数，同时能提升模态间的特征融合效果，达到特征增强的目的。

结 论

近几年随着短视频移动应用的兴起，大量以视频作为载体的信息在网络中传播。如何应用人工智能技术使机器可以对视频语义进行有效理解，以对海量了视频进行分析利用、信息搜索、视频推荐等，这些都无疑成为热门的研究话题。视频问答任务通过问答的方式测试机器对视频的理解能力，具有重大的理论意义和应用价值，一经提出后便受到了研究人员的广泛关注。

本文首先介绍视频问答任务的发展背景及意义，并简要分析了国内外基于此任务的相关研究工作的现状，接着在基于前人研究的基础上，本文通过结合注意力机制和图卷积网络等深度学习方法，提出了一种基于知识蒸馏的视频问答模型。综合利用了注意力机制在对视觉和文本特定区域特征的关注程度加强上的优势，以及图卷积网络在视频节点间关系推理上的有效性。同时，进一步挖掘了视频中静态的外观与动态的动作特征之间的独有特征和关联关系。并通过应用知识蒸馏的方法在对模型进行压缩的同时，进一步加强外观与动作特征之间的融合。该模型在 MSVD-QA 数据集上进行了消融实验，同时与现有其他方法的实验结果进行对比，取得了较好的结果，验证了本方法的有效性。

视频问答任务现发展迅速，但仍然面临着极大的挑战，在对其的研究上依然有着很大的发展空间。在结合全文工作的基础上，未来对视频问答任务的研究工作有以下几点改进：（1）视频问答的数据集上，问答形式仍以选择式为主，动态生成答案的方式缺失，这显然较一个真正智能的问答系统来说仍有一定距离。（2）在视频特征提取上，仍主要以通过预训练模型提取外观和动作特征表征视频，方法较为单一，如何有效的提取一个视频信息并完整的表征视频语义仍然是一个值得研究的方向。（3）在视频和问题文本之间的交互与融合上，主要还是采用注意力、图卷积等方式进行特征的增强以及对对象关系推理，缺少更多方法上的创新。而对于通过知识蒸馏的方式对模型多模态特征特征融合过程进行优化增强时，获取何处的融合特征、在什么时机指导学生模型的学习训练、学生模型的结构如何定义才能最优等问题仍需要更多的研究与尝试。

参 考 文 献

- [1] Poms A, Crichton W, Hanrahan P, et al. Scanner: Efficient video analysis at scale[J]. ACM Transactions on Graphics, 2018, 37(4): 1-13.
- [2] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]. Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea (South), 2019: 7083-7093.
- [3] Xu D, Zhao Z, Xiao J, et al. Video question answering via gradually refined attention over appearance and motion[C]. Proceedings of the 25th ACM international conference on Multimedia, San Francisco, CA, USA, 2017: 1645-1653.
- [4] Antol S, Agrawal A, Lu J, et al. Vqa: Visual question answering[C]. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 2425-2433.
- [5] Gupta P, Gupta V. A survey of text question answering techniques[J]. International Journal of Computer Applications, 2012, 53(4): 1-8.
- [6] Tapaswi M, Zhu Y, Stiefelhagen R, et al. MovieQA: Understanding Stories in Movies through Question-Answering[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.
- [7] Lei J, Yu L, Bansal M, et al. Tvqa: Localized, compositional video question answering[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018: 1369-1379.
- [8] Castro S, Azab M, Stroud J, et al. LifeQA: A Real-life Dataset for Video Question Answering[C]. Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, 2020: 4352-4358.
- [9] Song X, Shi Y, Chen X, et al. Explore multi-step reasoning in video question answering[C]. Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea (South) 2018: 239-247.
- [10] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA, 2009: 248-255.
- [11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 770-778.
- [12] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 6299-6308.

- [13] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4489–4497.
- [14] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014: 1532–1543.
- [15] Jang Y, Song Y, Yu Y, et al. Tgif-qa: Toward spatio-temporal reasoning in visual question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 2758–2766.
- [16] Kim K M, Heo M O, Choi S H, et al. Deepstory: Video story qa by deep embedded memory networks[C]. Proceedings of the 26 International Joint Conference on Artificial Intelligence, Melbourne, Australia, 2017: 2016–2022.
- [17] Xu D, Zhao Z, Xiao J, et al. Video question answering via gradually refined attention over appearance and motion[C]. Proceedings of the 25th ACM international conference on Multimedia, San Francisco, CA, USA, 2017: 1645–1653.
- [18] Gao J, Ge R, Chen K, et al. Motion-appearance co-memory networks for video question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 6576–6585.
- [19] Zhang Z, Zhao Z, Lin Z, et al. Open-ended long-form video question answering via hierarchical convolutional self-attention networks[C]. Proceedings of the 28 International Joint Conference on Artificial Intelligence, Macao, China, 2019: 4383–4389.
- [20] Thao M L, Vuong L, Svetha V, et al. Hierarchical conditional relation networks for video question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020: 9972 – 9981.
- [21] Wang X, Gupta A. Videos as space-time region graphs[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 399–417.
- [22] Jiang P, Han Y. Reasoning with Heterogeneous Graph Alignment for Video Question Answering[C]. Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 2020: 11109–11116.
- [23] Huang D, Chen P, Zeng R, et al. Location-Aware Graph Convolutional Networks for Video Question Answering[C]. Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 2020: 11021–11028.
- [24] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pretraining[EB/OL]. (2018-11-06) [2022-05-15].
https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

- [25] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C]. Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017: 5998–6008.
- [26] Yang Z, Dai Z, Yang Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]. Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2019: 5754–5764.
- [27] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]. Proceedings of the 27th International Conference on Neural Information Processing Systems, Kuching, Malaysia, 2014: 2204–2212.
- [28] Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 2016: 606–615.
- [29] Wang J, Bao B, Xu C. DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering[J]. arXiv preprint arXiv, 2021: 2107.04768.
- [30] Wang X, Zhu M, Bo D, et al. Am-gcn: Adaptive multi-channel graph convolutional networks[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, USA, 2020: 1243–1253.
- [31] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.
- [32] Yu Z, Yu J, Fan J, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 1821–1830.
- [33] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Trans. Neural Netw. Learn. Syst., 2020, 32: 4–24.
- [34] Cho J H, Hariharan B. On the efficacy of knowledge distillation[C]. Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea (South), 2019: 4794–4802.
- [35] Wang X, Hu J F, Lai J H, et al. Progressive teacher-student learning for early action prediction[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3556–3565.
- [36] Cheng X, Rao Z, Chen Y, et al. Explaining knowledge distillation by quantifying the knowledge[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12925–12935.
- [37] Albanie S, Nagrani A, Vedaldi A, Zisserman A. Emotion recognition in speech using cross-modal transfer in the wild[C]. Proceedings of the 26th ACM international Conference on Multimedia. Seoul, Korea (South), 2018: 292–301.

- [38] Song L, Smola A, Gretton A, et al. Supervised feature selection via dependence estimation[C]. The 24th Annual International Conference on Machine Learning, Corvallis, OR, USA, 2007: 823-830.
- [39] Chen D, Dolan W B. Collecting highly parallel data for paraphrase evaluation[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 2011: 190-200.
- [40] Singh G. Spatio-temporal relational reasoning for video question answering[D]. Vancouver, Canada: Univ. of British Columbia, 2019.
- [41] Fan C, Zhang X, Zhang S, et al. Heterogeneous memory enhanced multimodal attention model for video question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019: 1999-2007.
- [42] Yang T, Zha Z, Xie H, et al. Question-aware tube-switch network for video question answering[C]. Proceedings of the 25th ACM international conference on Multimedia, Nice, French, 2019: 1184-1192.

修改记录

(1) 毕业设计(论文)内容重要修改记录

第一次修改记录:

第 6 页第 2 章, **修改前:** 内容较少

修改后: 补充了相关技术的内容及插图

第 13 页第 3 章, **修改前:** 缺少各个模块的结构图

修改后: 添加了各个模块的结构图

第二次修改记录:


第 27 页第 4 章, **修改前:** 实验细节内容较少

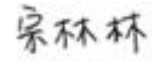
修改后: 补充了实验细节及可视化分析

第 37 页参考文献, **修改前:** 格式有误

修改后: 更正了参考文献的格式

(2) 毕业设计(论文)正式检测重复比: 3.2%

记录人(签字): 

指导教师(签字): 

致 谢

行文至此，不禁发觉我的大学生涯已然接近尾声。回首四年往昔，多姿多彩的校园生活历历在目，可惜韶光易逝，我们又将踏上新的旅程。遥想那年夏天，刚踏入校园的我，是懵懂与激动，是好奇与喜悦，憧憬的大学生活就要开始。而如今，四年时间已经给我带来了无数学识与成长，我想要感谢的有太多太多。

我要感谢美丽的大连理工大学，这一方育人的水土使我在求学道路上无所顾虑，让我在千里之外的求学生活如家一般温暖。

我要感谢我的所有敬爱的老师们。感谢指导老师宗林林老师一直以来对我的悉心指导。刚跟着宗老师研究时，基础薄弱，完全没有方向。但是宗老师会耐心地解答我所有困惑，并为我指明下一步的方向，一步步带着我求知进步。我开始慢慢步入了深度学习领域，并顺利完成了此次地研究课题。感谢我的辅导员房园老师，她对我们级队所有学生无论是学习还是生活，以及大大小小的事务上尽心尽责，关爱有加，对待我们像孩童一般呵护。感谢我的班主任刘馨月老师，经常与我们交流心得，分享经验，为我们指明未来的方向。感谢大连理工大学所有老师，你们辛勤工作，尽心尽责，教导有方，给我带来渊博知识与见识，使我在求真的道路上一路向前。

我要感谢我的所有可爱的同学们。感谢我的三个室友，和我共度了四年的同寝岁月，容忍我的缺点，陪伴我的生活，消除我的忧虑，关照我的学习。感谢我的全班同学，感谢班长、团支书及其他班委对我各种事务上的帮助与耐心，感谢所有同学，你们给我带来了欢笑，无条件地帮助我解决各种困难。感谢最近以来和我一起研究学习的两位同伴，在探索知识的旅程中有你们的陪伴毫不孤单，也不畏惧任何挑战，你们的帮助我牢记在心。感谢我的一位学长，你帮助了刚刚进入校园的我所面临的各种疑惑和困难，为我分享经验，提供建议。我要感谢大连理工大学所有同学，与你们一同在这一片蓝天下求知求学，我倍感温暖。

我要感谢我亲爱的家人。从小在你们的教育与呵护下成长，我今天才能够在一流的学府探索真知。是你们默默无闻的付出和坚实的臂膀，给了我勇往直前的决心。

感谢你们，一路有你们的陪伴，是我最大的幸运。