

基于自监督图对比学习的视频问答方法^{*}

姚 暄^{1,2}, 高君宇^{1,2}, 徐常胜^{1,2,3}

¹(模式识别国家重点实验室(中国科学院 自动化研究所), 北京 100190)

²(中国科学院大学 人工智能学院, 北京 100190)

³(鹏城实验室, 广东 深圳 518055)

通信作者: 高君宇, E-mail: junyu.gao@nlpr.ia.ac.cn



摘 要: 视频问答作为一种跨模态理解任务, 在给定一段视频和与之相关的问题的条件下, 需要通过不同模态语义信息之间的交互来产生问题的答案. 近年来, 由于图神经网络在跨模态信息融合与推理方面强大的能力, 其在视频问答任务中取得了显著的进展. 但是, 大多数现有的图网络方法由于自身固有的过拟合或过平滑、弱鲁棒性和弱泛化性的缺陷使得视频问答模型的性能未能进一步提升. 鉴于预训练技术中自监督对比学习方法的有效性和鲁棒性, 在视频问答任务中利用图数据增强的思路提出了一种图网络自监督对比学习框架 GMC. 该框架使用针对节点和边的两种数据增强操作来生成相异子样本, 并通过提升原样本与生成子样本图数据预测分布之间的一致性来提高视频问答模型的准确率和鲁棒性. 在视频问答公开数据集上通过与现有先进的视频问答模型和不同 GMC 变体模型的实验对比验证了所提框架的有效性.

关键词: 图对比学习; 视频问答; 图数据增强; 预训练

中图法分类号: TP18

中文引用格式: 姚暄, 高君宇, 徐常胜. 基于自监督图对比学习的视频问答方法. 软件学报, 2023, 34(5): 2083–2100. <http://www.jos.org.cn/1000-9825/6775.htm>

英文引用格式: Yao X, Gao JY, Xu CS. Self-supervised Graph Contrastive Learning for Video Question Answering. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2083–2100 (in Chinese). <http://www.jos.org.cn/1000-9825/6775.htm>

Self-supervised Graph Contrastive Learning for Video Question Answering

YAO Xuan^{1,2}, GAO Jun-Yu^{1,2}, XU Chang-Sheng^{1,2,3}

¹(National Laboratory of Pattern Recognition (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

²(School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China)

³(Pengcheng Laboratory, Shenzhen 518055, China)

Abstract: As a cross-modal understanding task, video question answering (VideoQA) requires the interaction of semantic information with different modalities to generate answers to questions given a video and the questions associated with it. In recent years, graph neural networks (GNNs) have made remarkable progress in VideoQA tasks due to their powerful capabilities in cross-modal information fusion and inference. However, most existing GNN approaches fail to improve the performance of VideoQA models due to their inherent deficiencies of overfitting or over-smoothing, as well as weak robustness and generalization. In view of the effectiveness and robustness of self-supervised contrastive learning methods in pre-training techniques, this study proposes a self-supervised graph contrastive learning framework GMC based on the idea of graph data augmentation in VideoQA tasks. The framework uses two independent data augmentation operations for nodes and edges to generate dissimilar subsamples and improves the consistency between predicted graph data distributions

* 基金项目: 科技创新 2030-“新一代人工智能”重大项目 (2020AAA0106200); 国家自然科学基金 (62036012, U21B2044, 62102415, 62072286, 61721004); 之江实验室开放课题 (2022RC0AB02); CCF-海康威视“斑头雁”基金 (20210004)

本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.

收稿时间: 2022-04-18; 修改时间: 2022-05-29; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20

CNKI 网络首发时间: 2023-03-17

of the original samples and augmented subsamples for higher accuracy and robustness of the VideoQA models. The effectiveness of the proposed framework is verified by experimental comparisons with existing state-of-the-art VideoQA models and different GMC variants on the public dataset for VideoQA tasks.

Key words: graph contrastive learning; video question answering; graph data augmentation; pre-training

随着大数据时代的到来, 视频已成为信息密度最大的载体之一, 视频内容理解深受研究人员的关注. 而视频问答作为一种关键的多模态学习任务, 是视频内容理解领域的一个重要分支. 视频问答任务通常是指给计算机一个视频片段和若干与视频内容相关的自然语言问题, 令计算机根据提问的文本对视频内容进行针对性地理解, 从而自行推理出答案, 并用自然语言来回答问题. 相对于先前的文本问答^[1]和基于静态图像的视觉问答^[2], 视频问答需要在不同维度上对视频片段的空间位置、时序关系、因果逻辑进行建模^[3], 因此更具有挑战性, 也更能反映模型的智能化水平.

在早期的研究中, 视频问答任务主要延伸了图像视觉问答中广泛应用的模型与方法, 例如卷积神经网络 (CNN)^[4,5]、循环神经网络 (RNN)^[6,7]、注意力机制 (attention mechanism)^[8,9]等. 近期, 图神经网络 (GNN) 以其强大的学习表示和关系推理能力被成功应用于视频问答模型, 通过图的节点与边来表示物体特征和交互关系, 利用消息传递或等价的邻域聚合函数从节点及其邻域提取高级特征, 较好地提升了视频问答任务的性能^[10].

然而, 最近的一些研究表明, GNN 的特征传播过程存在一些固有的问题与缺陷, 这也阻碍了其在视频问答模型中的广泛应用与发展: 首先, 大多数的 GNN 存在过拟合 (over-fitting) 和过平滑 (over-smoothing) 的问题. 图数据 (graph data)^[11]是原始数据的抽象表示并具有丰富的结构化信息, 由于神经网络在学习图数据的过程中神经网络产生和使用的参数数量十分庞大, 训练促使模型对样本数据集实现了一对一的完美映射, 但在面对新的测试数据时, 模型却很难将提取后的新特征与自身学习得到的特征相匹配, 识别效果不佳^[12]. 也就是说, 当我们使用一个过度参数化的模型来拟合一个具有有限训练数据的分布时, 尽管学习后的模型很好地拟合了训练数据, 但很难推广到测试数据, 泛化能力有待提升^[13]. 同时, 由于图中节点之间存在边连接, 因此随着图网络消息传递的进行, 不同节点上的特征有可能趋于一致, 使得图结构的判别力降低, 导致过平滑现象. 其次, 由于大多数 GNN 采用确定性传播方式 (deterministic propagation), 其中图的每个节点信息都高度依赖于其多跳邻域 (multi-hop neighborhoods), 这使得节点更有可能被潜在的数据噪声所影响, 并且更容易受到对抗性扰动 (adversarial perturbations)^[14], 从而进一步造成视频问答模型的弱鲁棒性. 众所周知, 增大训练数据量来进行监督学习的深度学习方法可以有效解决这个问题, 但是获取高质量的数据标注需要花费大量的人力和计算资源, 并且难以保证样本的时间动态性^[15]. 在当前预训练技术显著发展的阶段, 通过增加高代价的标注数据以提升模型泛化性的方式已显得捉襟见肘. 事实上, 预训练技术可以通过对数据进行自监督增强的方式以实现低标注代价下的高泛化性和鲁棒性. 目前, 自监督的数据增强方法在众多计算机视觉任务中取得了显著的性能提升, 它是在不影响语义标签的情况下, 通过对已有数据做一系列随机改变, 来产生多样化、高质量的训练样本并自动构造这些样本间的对比关系, 从而扩大训练数据集的规模. 最近在计算机视觉领域提出的一些新方法, 例如 MixMatch^[16]、UDA^[17]等, 通过设计用于一致性正则化训练的数据增强方法来解决图上的半监督学习问题, 促进模型在训练期间提高对视频内容的理解和对问题的敏感性, 取得了很好的效果^[18].

然而, 在视频问答任务方面, 目前尚缺乏利用自监督图数据增强方法以提升模型在时空复杂、开放环境下问答鲁棒性及泛化性的探索. 针对上述问题, 受预训练技术中自监督对比学习方法^[19]的启发, 本文设计了用于自监督对比学习的图数据增强方法以提升视频问答模型的稳健性. 具体地, 针对一个输入视频-问题样本对, 我们可以构造一个异构图模型. 为了有效地增加图数据来提升模型泛化能力, 本文提出了一种基于图卷积网络 (GCN) 的掩码 (MASK) 机制, 包含节点丢弃 (Node-dropping) 和边丢弃 (Edge-dropping) 两种操作, 它的核心是在训练阶段从输入图中随机删除一定数量的节点或边. 该机制可以被视为一种数据增强技术, 通过生成不同的原始输入图的随机残缺子图, 来增加输入数据的多样性和丰富性, 从而更好地防止过拟合现象, 提升模型的泛化能力. 其次, 在每个训练阶段, 每个图数据样本经过前向传递, 利用这种 MASK 机制获得辅助图卷积网络的扩充训练样本, 本文通过最小化原样本与相异的子样本数据预测分布之间的 KL 散度 (Kullback-Leibler divergence), 使得相同数据的不同增

强输出相似的预测, 即对二者的输出进行正则化处理, 从而减小模型对原始数据的依赖, 以此提升自身的鲁棒性. 本文在最新的高挑战性的视频问答公开数据集 NExT-QA^[3]上完成了模型对开放式问答和多项式问答任务的评估并取得了良好的效果.

本文的主要贡献如下.

(1) 受预训练技术中的自监督对比学习方式的启发, 本文探索了面向视频问答任务的图数据增强及对比学习方法, 提出了一种新的基于自监督的图对比学习通用框架 GMC (graph contrastive learning with MASK module), 通过生成不同的原始输入图的随机残缺子图来提升模型的泛化能力.

(2) 本文提出了一种基于图卷积网络 (GCN) 的掩码 (MASK) 机制, 包含节点丢弃 (Node-dropping) 和边丢弃 (Edge-dropping) 两种操作, 并通过最小化原样本与生成子样本图数据预测分布之间的 KL 散度以高效构造图数据增强及自监督对比.

(3) 本文在视频问答公开数据集 NExT-QA^[3]上完成了模型对开放式问答和多项式问答任务的评估, 实验结果及大量的模型消融分析表明该方法可以有效提高视频问答模型的准确率和鲁棒性.

本文第 1 节介绍视频问答任务的研究背景与相关工作. 第 2 节结合图数据增强方法, 详细介绍本文构建的基于自监督图对比学习的视频问答模型. 第 3 节主要介绍实验数据、实验设置与细节, 通过对比实验验证了所提模型的可行性与有效性. 第 5 节总结全文的研究工作, 并展望未来视频问答模型的发展.

1 相关工作

1.1 视频问答

视频问答 (VideoQA) 是一项根据给定问题 Q 和视频片段 V 来预测正确答案 a^* 的任务, 通常有两种类型, 分别为多项式任务 (multi-choice task) 和开放式任务 (open-ended task). 对于多项式问答 (multi-choice QA) 中的每一个问题, 模型需要从所提供的几个候选答案 A_{mc} 中选择一个正确答案 $a^* = F(a|Q, V, A_{mc})$. 对于开放式问答 (open-ended QA), 模型需要进一步理解问题和视频内容, 并自动生成自然语言答案, 在之前的一些研究^[3,20,21]中它通常被设定为一个分类问题, 模型需要将视频问题对 (video-question pairs) 分类至预定义的答案集 A_{oe} , 即 $a^* = F(a|Q, V, a \in A_{oe})$ ^[22]. 在本文选用的 NExT-QA 数据集中, 开放式任务被设置为生成式问题, 答案多为简单的短语组合, 这种形式具有更高的实用价值, 最近也受到了广泛关注^[23-25].

常见的视频问答模型一般由 4 个模块构成. 视频特征提取模块通常使用卷积神经网络对视频片段进行处理, 语言特征提取模块使用词嵌入方法得到问题文本的单词向量空间表示, 跨模态融合模块需要对不同的语义空间下的各语义表示进行运算得到融合的特征信息, 答案推理与生成模块对融合特征作解码处理后, 根据每个位置的单词概率分布来选择单词, 直到生成结束字符为止, 从而得到模型对于问题的最终回答^[26].

在视频问答的技术方法方面, 早期的研究主要是在扩展基于图像的视觉问答方法的基础上、解决视频帧之间的时序建模问题. 基本方法是使用基于循环神经网络 (RNN) 的模型对序列化的视频帧特征进行建模, 模型的架构多表现为编码器-解码器架构 (encoder-decoder framework)^[27-29]. 但这些工作只是简单地利用 RNN 提取单独模态的信息, 而忽略了模态之间的交互. 记忆网络 (memory networks) 可以在内存插槽 (memory slot) 中缓存时序输入信息, 并显式地利用历史存储记忆. 它能够实现多步推理, 逐步完善答案, 相较于早期工作有了一定的进步, 但只具备这种能力的模型还不足以深入理解视频内容^[30-32]. Transformer 模型具有良好的长时序关系建模能力, 并且能够成功应用于视频问答等多模态视觉语言任务的建模. 尽管基于该方法的跨模态视频问答模型在多个事实问答数据集上实现了 SOTA 性能, 但它在推理问题的预测能力未得到有效的探索^[33-35]. 模块化网络 (modular networks) 具有较好的分层学习能力, 当视频问答模型的数据模式、视频长度或问题类型发生变化时, 该方法通过分层封装来生成输入特征之间的关系, 如 Le 等人^[36]设计了一种可重用的分层神经单元 CRN, 以不同的粒度 (包括 frame-level, clip-level 和 video-level) 嵌入以语言提示为条件的视频输入. 受 ImageQA^[37]中的神经符号方法的启发, NS-DR 模型旨在将用于模式识别和动态预测的神经网络与用于因果推理的符号逻辑相结合, 将问题转化为功能程序, 使用

动态预测器提取和预测视频的动态场景,并在动态场景上运行程序以实现因果推理^[38]。尽管神经符号在合成数据集上具有推理能力^[38,39],但在真实视频上的性能仍然未知。

图神经网络可以有效地捕捉局部与全局的关系,具有较好的特征抽取和融合能力,图结构推理方法相较其他方法可以更好地对关系信息进行建模,这对视频问答模型的推理能力至关重要。最近的一些工作尝试了应用图神经网络, Jiang 等人^[40]在提出的异构图对齐网络 (HGA) 中使用协同注意力机制 (co-attention) 来融合问题特征与视频特征,并将融合的多模态特征作为图中的节点进行答案推理,如图 1 所示, HGA 模型首先在视频中的单词 lady 与视觉区域之间建立语义关系,之后定位动作,与此同时还需要完成语义相似的模态间对齐,从而在时序推理中找到动作 laugh。与前者相同, B2A^[41]和 DualVGR^[42]等模型基于视频元素和单词所构建的图,实现了模态内与模态间的关系学习,实现了较好的模型性能。考虑到视频元素在语义空间上是分层的, Liu 等人^[43], Peng 等人^[44]和 Xiao 等人^[45]分别将分层学习思想融入到图网络中。具体而言, Liu 等人^[43]提出了一种图记忆机制 (HAIR),用于从对象级别到帧级别执行关系视觉语义推理; Peng 等人^[44]以渐进的方式连接不同级别的图,即对象级别 (object-level)、帧级别 (frame-level) 和片段级别 (clip-level),以学习视觉关系 (PGAT); 而 Xiao 等人^[45]提出了一种分层条件图模型 (HQGA),通过图聚合和池化,将低级实体到高级视频元素的视觉事实相对应,以实现多粒度级别的视觉文本匹配。图网络具有良好的关系建模能力,尤其在视频问答任务的推理环节中表现出色,而重点和难点在于如何巧妙地利用视频元素设计图。此外,目前的图网络仍然缺乏明确的逻辑形式推理^[22]。

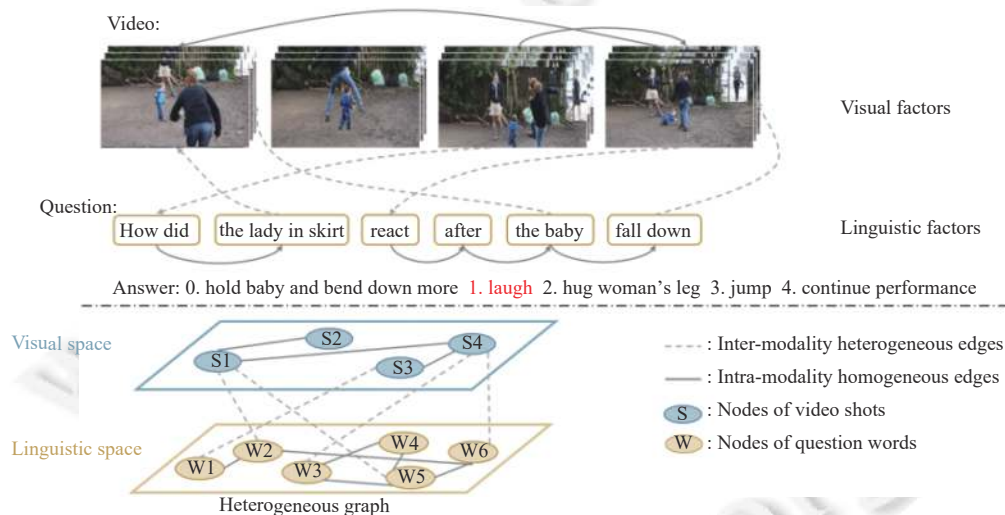


图 1 HGA 模型异构图的构建示例

1.2 图卷积网络

经典的卷积神经网络 (CNN) 在计算机视觉领域取得了巨大成功,将卷积核引入神经网络中,通过卷积操作实现了参数共享和加权平均,具有很强的提取和整合数据信息的能力。受此启发,图神经网络 (GNN) 将参数共享的思想引入,使得图卷积网络 (GCN) 通过卷积核参数的迭代更新来显式地学习包含较强逻辑关系的空间图谱结构,以此捕获不同的关系和空间结构特征^[8]。Bruna 等人^[46]首次提出了关于 GCN 的重要研究,该研究基于谱图理论 (spectral graph theory) 发展了图卷积运算。随后, Henaff 等人^[47], Defferrard 等人^[48]和 Levie 等人^[49]对基于频域方法 (spectral-based method) 的 GCN 进行了改进和推广。但频域方法具有较大的局限性,它的时间和空间复杂度会随着图规模的增加而剧烈提升^[50],为了解决 GCN 在大型图上的可扩展性,基于空间域 (spatial-based method) 的 GCN 得到了高度关注并迅速发展^[51-53]。

图卷积网络中的权值通常是一个集合,在计算某个节点的聚合特征值时,按一定规律将参与聚合的所有节点归为多个不同的子集,同个子集内的节点采用相同的权值以实现权值共享。GCN 中的前馈传播按如下方式递归执行:

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\Theta^{(l)}) \quad (1)$$

其中, $\mathbf{H}^{(l+1)} = \{h_1^{(l+1)}, \dots, h_N^{(l+1)}\}$ 是第 l 层隐含层 (the l^{th} hidden layer) 的特征矩阵, 且 $\mathbf{H}^{(0)} = \mathbf{X}$, $h_i^{(l)}$ 是节点 i 的隐含特征表示, $\hat{\mathbf{A}} = \hat{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\hat{\mathbf{D}}^{-\frac{1}{2}}$ 为邻接矩阵 $\mathbf{A}$ 重归一化后的矩阵表示, $\hat{\mathbf{D}}$ 为 $\mathbf{A} + \mathbf{I}$ 相应的度矩阵; $\sigma(\cdot)$ 为一种激活函数, 例如 ReLU 函数; $\Theta^{(l)}$ 为第 l 层隐含层的权值矩阵.

1.3 图数据增强

近年来, 数据增强技术的应用显著提升了依托数据来推理的模型的泛化能力和性能改进^[54]. 而图结构数据的数据增强仍在探索中, 尽管 GNN 与数据增强具有一定的互补性, 但是关于二者的结合工作很少, 其中的一个难点是, 相较于其他数据是由位置编码的, 图结构通过节点连接进行编码, 具有不规则性和复杂性, 计算机视觉中常见的人工制作、结构化的数据增强方法无法直接应用到图上^[55]. 传统的自我训练方法^[56,57]利用训练过的模型对未标记的数据进行注释, 最近的一些工作成功实现了图数据增强, 但需要花费大量的人力和计算成本: Li 等人^[58]提出在对抗性学习环境中训练生成器分类器网络, 以生成假节点; Deng 等人^[59]和 Feng 等人^[60]对图结构上的节点特征产生对抗性扰动. 虽然图数据增强方面取得了一定的研究进展, 但其在视频问答任务上的研究还鲜有探索.

1.4 自监督学习

自监督学习是无监督学习 (unsupervised learning) 的一个分支, 通常指使用自动生成的伪标签显式训练神经网络的学习方法, 主要分为生成式 (generative SSL) 和对比式 (contrastive SSL) 两种方法, 目前多为设计各种辅助任务 (pretext task) 帮助模型从无标注的数据中学习特征.

生成对抗网络 (GAN)^[61]的诞生使得生成式方法获得热烈关注, 基于该网络研究人员相继提出了 CycleGAN^[62]、StyleGAN^[63]、DiscoGAN^[64]等模型并大获成功, 但 GAN 自身存在一定的复杂性使得模型训练困难, 如模型参数经常振荡、难以收敛, 判别器和生成器网络之间不同步使得学习难以继续等^[65]. 对比式方法是一种鉴别性模型, 通过噪声对比评估指标 (NCE) 驱使相似的样本更加接近、不同的样本相互远离:

$$L = E_{x, x^+, x^-} \left[-\log \left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}} \right) \right] \quad (2)$$

其中, x 与 x^+ 相似、与 x^- 不相似, $f(\cdot)$ 为一个编码器表示函数. 最近提出的 SimCLR^[66]、BYOL^[67]等模型在特征提取方面的效果能够与最先进的监督学习方法相当.

最近对比式方法在 CNN 的视觉表征学习任务中取得了巨大成功^[68], 而图数据由于具有不规则的抽象结构, 很难直接迁移使用视觉表征中的对比学习方法, 故图对比学习的探索仍较少. 一些工作使用图的不同部分来构建对比学习的样本对, 如节点与节点^[69]、节点与全局图^[70,71]、节点与子图^[72]等, 另一些工作尝试采用图数据增强的方式来生成对比样本对, GCA^[73]通过删除边缘和掩蔽特征的方式进行图数据扩充, GraphCL^[18]对不同组合的数据增强方式进行了广泛研究. 相比于之前的大多数方法选择计算 L_2 距离作为正则化方法, GMC 使用视频问答模型输出概率分布之间的 KL 散度作为对比学习损失函数. 由于实际的预测分布与图网络隐含状态的距离并不在同一空间, 基于 KL 散度的方式更易实现最大化模型预测结果一致性的训练目标, 能够更有效地实现优化.

2 基于自监督图对比学习的方法 GMC

本文提出了一个基于图自监督数据增强框架 GMC 的视频问答方法, 用于提升视频问答任务中 GCN 推理问题答案的性能. 总体框架如后文图 2 所示, 引入所构造的掩码 (MASK) 模块作为数据增强方法, 从而实现对输入图数据集的扩充, 之后通过潜在空间中的对比学习最大化同一图结构的增强图与原始图之间的一致性来提升视频问答推理的鲁棒性.

2.1 基于图卷积网络的视频问答模型

本文所采用的视频问答模型框架可以分为 4 个部分, 如图 2 所示. 首先, 将给定的视频片段和问题的集合 $S_{\text{input}} = \{(v_1, q_1), \dots, (v_n, q_n)\}$ 以及候选答案作为输入, 将视频片段 v_i 拆分成 N 个视频帧, 视频片段和帧通过视觉编码

器提取视频的外观特征和运动特征,并将二者拼接得到视觉联合特征,问题语句 q_i 通过语言编码器提取语言特征,同时将问题的答案标签 a_i 作为监督信息.其中特征提取模块由 ResNet 和 C3D 组成,其输出维度为 $N \times 2048$ (N frames and 2 048 features). GloVe 的输出是嵌入长度 (embedding size) 为 300 的固定长度向量.视觉编码器 (visual encoder) 和语言编码器 (linguistic encoder) 具有类似的结构,前者的输出是长度为 N 的隐含状态 (hidden states) 序列,而后者的输出为单个隐含状态^[74].

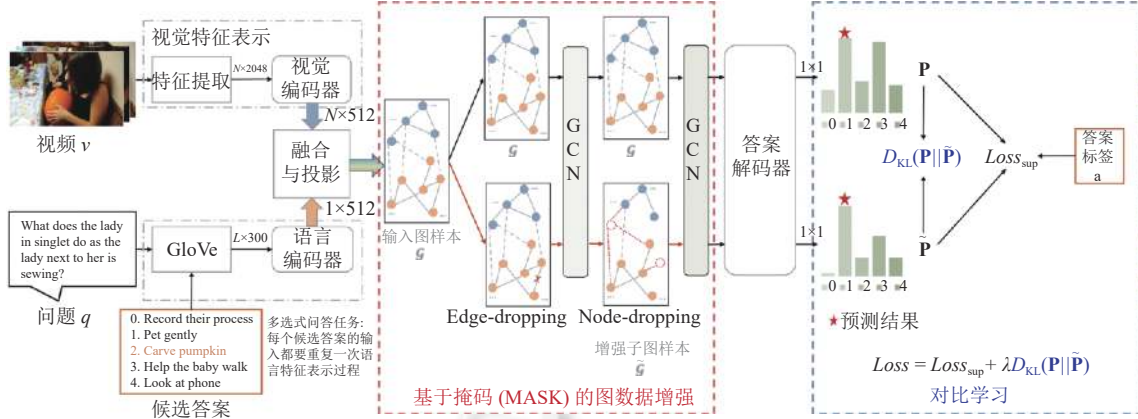


图2 总体框架图

之后不同时段的视觉联合特征 (维度为 $N \times 512$) 和问题特征 (维度为 1×512) 作为注意力融合模块的输入,得到跨模态融合特征.将融合特征投影到交互空间后得到输入邻接矩阵 $\mathbf{G}$,即可生成模型推理模块中 GCN 的输入图 $g = (N, E)$,其中 N 表示节点, E 表示边.最后由答案解码器可以得到归一化的预测答案权重,其输出为每个候选答案的实值分数,输出权重最大值者作为最终答案.模型中所使用的图自监督学习方法将在第 3.2 节中进一步阐述.

Jiang 等人^[40]提出的视频问答 HGA 模型,通过引入一个异构图推理模块和一个协同注意力嵌入操作来捕获视频片段、问题语句以及其跨模态之间的局部与全局关系,更适用于因果和时序问题的推理,是当前性能较好的基于图网络的视频问答模型.故本文选用 HGA 模型来开展下一步实验,以此更好地验证所提出的自监督学习方法.需要注意的是,本文提出的图自监督数据增强的视频问答学习策略可以被无缝嵌入到任何使用图结构推理的视觉问答模型中,本文仅选用 HGA 测试所提出方法的效果.

2.2 基于掩码 (MASK) 的图数据增强模块

在前述模型的特征表示和融合的基础上,利用视觉与语言两模态的交叉嵌入特征来构造无向异构图 g 作为图卷积网络的输入图,其特征矩阵定义为 $\mathbf{X}$,它由语言模态和视觉模态的两个交叉嵌入特征串联拼接所得, $\mathbf{X}$ 中的每个特征向量 $\{\mathbf{x}_i\}_{i=1}^N$ 作为异构图的节点,邻接矩阵定义为 $\mathbf{A}$, A_{ij} 表示节点 n_i 与 n_j 之间归一化后的对齐权重,随机 MASK 模块生成输入图的相异子图作为增强图.对于每个增强图 $\tilde{g}$,它都将被输入模型推理模块 (一个两层 GCN) 进行答案分类预测.本文设计了面向图数据中节点 (node) 和边 (edge) 两种类型的掩盖操作:

(1) 节点丢弃 (Node-dropping)

在训练阶段,对于每个输入数据,Node-dropping 方法会以一定的概率随机丢弃输入图的节点以及它的邻接边信息.本文通过随机利用 MASK 来掩盖一些节点,即随机将 $\mathbf{X}$ 中某些节点对应的特征置 0.

令图 g 中每个节点 $n_i \in N$ 对应的二元掩码 (MASK) 为 ε_i ,它的数值通过独立随机采样得到,服从参数为 δ 的伯努利分布 (Bernoulli distribution):

$$\varepsilon_i \sim \text{Bernoulli}(1 - \delta) \quad (3)$$

其中, δ 为 ε_i 取值 0 的概率, $i \in (1, 2, \dots, N)$.

随后将每个节点的特征向量与所对应的掩码相乘,得到扰动特征矩阵 $\tilde{\mathbf{X}}$,即 $\tilde{\mathbf{x}}_i = \varepsilon_i \cdot \mathbf{x}_i$,其中 $\mathbf{x}_i$ 表示矩阵 $\mathbf{X}$ 的

第 i 行向量.

图卷积网络的信息传递方式如公式 (1), 经过 Node-dropping 操作后, 网络的传播公式变化如下:

$$\begin{cases} \tilde{\mathbf{H}}^{(0)} = \tilde{\mathbf{X}} = \{\varepsilon_i \mathbf{x}_i\}_{i=1}^N \\ \tilde{\mathbf{H}}^{(l+1)} = \sigma(\hat{\mathbf{A}} \tilde{\mathbf{H}}^{(l)} \Theta^{(l)}) \end{cases} \quad (4)$$

(2) 边丢弃 (Edge-dropping)

在训练过程中, Edge-dropping 方法会以一定的概率随机删除输入图的边. 实际运算过程中, 它随机地将图邻接矩阵 $\mathbf{A}$ 中的 ξE 个非零元素强制置零, 其中 E 为边的总数, ξ 为删除率 (dropping rate). 此时新生成的增强子图的邻接矩阵 $\tilde{\mathbf{A}}$ 可表示为:

$$\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{A}_{\text{drop}} \quad (5)$$

其中, $\mathbf{A}_{\text{drop}}$ 可以理解为由原始图边 E 中 ξE 个元素随机组成的稀疏子集所构造的矩阵.

当图卷积网络经过 Edge-dropping 处理后, 由公式 (1) 得到网络的传播公式变化如下:

$$\begin{cases} \tilde{\mathbf{H}}^{(0)} = \mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \\ \tilde{\mathbf{H}}^{(l+1)} = \sigma(\hat{\mathbf{A}} \mathbf{H}^{(l)} \Theta^{(l)}) \end{cases} \quad (6)$$

Edge-dropping 同前种方法一样生成了原图 G 的随机变形子样本, 增加了输入图数据的随机性和多样性, 能够有效抑制训练时神经网络的过拟合, 由于它删去了一部分图的信息传播路径, 在一定程度上减缓了 GCN 过平滑的收敛速度.

本研究提出的 MASK 模块可以作为一种辅助图卷积网络的训练样本扩充策略, 它对能够很好地促使模型对于视频内容和问题文本的深度理解, 减小模型对原始数据的依赖, 通过现有所学习到的特征来推测新的特征表示, 从而提升自身的鲁棒性. 该机制也在一定程度上缩减了信息传递的规模, 删除某些边会使节点连接更加稀疏^[11], 删除某些节点可以有效减少图卷积的运算成本^[75], 这使得模型能够更高效地传播高阶特征, 有助于降低 GCN 过平滑的风险.

2.3 自监督图对比学习

在每个训练阶段, 原异构图 g 将被输入 MASK 模块. 由于 Node-dropping 和 Edge-dropping 方法均为随机删除图中元素, 通过正向传递将产生不同的子图, 因此可以得到增强后的图特征矩阵 $\tilde{\mathbf{X}}$ 和邻接矩阵 $\tilde{\mathbf{A}}$, 原样本和它分别预测得到相应的输出:

$$\begin{cases} \mathbf{P}(a|v, q) = f_{\text{predict}}(\mathbf{X}, \mathbf{A}, \Theta) \\ \tilde{\mathbf{P}}(a|v, q) = f_{\text{predict}}(\tilde{\mathbf{X}}, \tilde{\mathbf{A}}, \Theta) \end{cases} \quad (7)$$

其中, Θ 为模型自身的参数, $\mathbf{P}, \tilde{\mathbf{P}} \in [0, 1]^{K \times 1}$ 分别表示原样本和增强子样本所预测的答案分布.

在原模型中, 图卷积网络通过数据的标签信息使得损失函数 $Loss_{\text{sup}}$ 最小化来实现监督学习. 我们增加定义了对比损失函数 $Loss_{\text{con}}$, 并以合理的方式对二者进行组合.

我们采用相对熵函数 (relative-entropy loss) 作为对比学习损失函数 (contrastive loss), 即 KL 散度 (Kullback-Leibler divergence, KLD), 来进行两个图结构预测之间的对比. KL 散度是两个概率分布之间差别的非对称性度量, 当两个分布越相似, KL 散度越小.

对于增强图和原图的预测结果, 我们希望二者的概率分布相近, 从而减小模型对于噪声的抗扰能力, 定义对比学习损失函数如下:

$$Loss_{\text{con}} = D_{\text{KL}}(\mathbf{P} \parallel \tilde{\mathbf{P}}) \quad (8)$$

其中, $D_{\text{KL}}(\mathbf{P} \parallel \tilde{\mathbf{P}}) = \sum_{x \in X} P(x) \log \frac{P(x)}{\tilde{P}(x)}$.

故在每个训练阶段, 我们同时使用公式 (8) 中的监督损失和对比损失, 本模型的最终损失函数为:

$$Loss = Loss_{\text{sup}} + \lambda Loss_{\text{con}} \quad (9)$$

其中, λ 为平衡两种损失的超参数 (hyper-parameter), $\lambda \in (0, 1)$. $Loss_{\text{sup}}$ 将在第 2.4 节中进行介绍.

2.4 模型训练与测试

在模型训练中, 我们定义监督学习损失函数 (supervised loss) 如下.

对于多选题问答任务, 输入 $S_{\text{input}} = \{(v, q), \text{candidate answers}\}$ 中包含 K 个候选答案, 重复输入 K 次后可得到最终的输出预测向量 $\mathbf{P}(a|v, q)$ (即 K 个候选答案的分数). 本文通过最小化成对铰链函数 (hinge loss) 来训练神经网络, 即每个训练阶段图的目标函数定义为原异构图和增强图之间的平均成对铰链损失:

$$Loss_{\text{sup}}^g = \sum_{m=1}^{K-1} \max(0, 1 + p_m^w - p^r) \quad (10)$$

$$Loss_{\text{sup}}^{\tilde{g}} = \sum_{m=1}^{K-1} \max(0, 1 + \tilde{p}_m^w - \tilde{p}^r) \quad (11)$$

$$Loss_{\text{sup}} = \frac{1}{2} \sum_{m=1}^{K-1} [\max(0, 1 + p_m^w - p^r) + \max(0, 1 + \tilde{p}_m^w - \tilde{p}^r)] \quad (12)$$

其中, 原样本预测结果为 $\mathbf{P}(a|v, q) = \{p_m^w, p^r\}$, 增强样本的预测结果为 $\tilde{\mathbf{P}}(a|v, q) = \{\tilde{p}_m^w, \tilde{p}^r\}$, p_m^w 、 $\tilde{p}_m^w$ ($1 \leq m \leq K-1$) 分别表示原样本与增强样本中不正确选项的得分, p^r 、 $\tilde{p}^r$ 分别表示原样本与增强样本中正确选项的得分.

对于开放式问答任务, 模型通过一个线性分类器和 *Softmax* 函数来评估预定义的答案集 A_{oe} 中所有答案的得分, J 表示样本中的总分类数. 本文通过最小化交叉熵函数 (cross-entropy loss) 来训练模型, 即每个训练阶段图的目标函数定义为原异构图和增强图之间的平均交叉熵损失:

$$Loss_{\text{sup}}^g = - \sum_{j=1}^J a_j \cdot \log p_j \quad (13)$$

$$Loss_{\text{sup}}^{\tilde{g}} = - \sum_{j=1}^J a_j \cdot \log \tilde{p}_j \quad (14)$$

$$Loss_{\text{sup}} = - \frac{1}{2} \sum_{j=1}^J (a_j \cdot \log p_j + a_j \cdot \log \tilde{p}_j) \quad (15)$$

其中, 原样本预测结果分别为 $\mathbf{P}(a|v, q) = \{p_j | 1 \leq j \leq J\}$, 增强样本的预测结果为 $\tilde{\mathbf{P}}(a|v, q) = \{\tilde{p}_j | 1 \leq j \leq J\}$, $a = \{a_j, 1 \leq j \leq J\}$ 为标注的预期结果.

在对训练好的模型进行评估测试时, 我们不对原样本进行数据增强, 由原样本的预测结果得到最终输出分布 $\mathbf{P}_{\text{eval}}$ 定义如下:

$$\mathbf{P}_{\text{eval}} = \mathbf{P} \quad (16)$$

本文所提出图卷积网络的自监督对比学习框架 GMC 的训练流程如算法 1 所示.

算法 1. GMC.

输入: 输入图数据 $g = (N, E)$, 超参数 λ , MASK 模块节点和边的掩蔽概率 ε 、 ξ ;

输出: 模型的权值参数.

1. 初始化模型参数
 2. **for** 每一个训练阶段 **do**
 3. 原图 g 输入 MASK 模块, 随机删除输入图节点和边
 4. 计算得到增强图 $\tilde{g}$ 和对应的邻接矩阵 $\tilde{\mathbf{A}}$ 、特征矩阵 $\tilde{\mathbf{X}}$
 5. 将原图与得到的增强图分别输入到答案推理模块, 计算输出向量 $\mathbf{P}$ 、 $\tilde{\mathbf{P}}$
 6. 由公式 (12) 或公式 (15) 计算监督学习损失函数值 $Loss_{\text{sup}}$
 7. 由公式 (8) 计算对比学习损失函数值 $Loss_{\text{con}}$
-

8. 由公式 (9) 通过优化损失函数以更新模型的权值参数

9. end

3 实验分析

3.1 实验数据集

本文采用 NExT-QA 数据集^[3]作为实验数据集. NExT-QA 数据集是一个以人们日常生活中的自然视频为主的视频问答数据集, 它来源于视频关系数据集 VidOR^[76], 视频内容主要关于社交聚会、儿童玩耍、户外活动、宠物和音乐等. 其中多选题问答数据的训练集、包含 34 132 个视频-问答对 (video-qa pair), 测试集包含 4 996 个视频-问答对; 开放式问答数据的训练集包含 37 523 个视频-问答对, 测试集包含 5 343 个视频问答对, 并且因果问题 (casual questions)、时序问题 (temporal questions) 和描述性问题 (descriptive questions) 分别占比 48%、29% 和 23%, 训练集/验证集/测试集的各部分比例为 7:1:2. 表 1 展示了 NExT-QA 数据集的具体数据统计.

表 1 NExT-QA 数据集划分

Tasks	Videos				Questions			
	Train	Val	Test	Total	Train	Val	Test	Total
Multi-choice QA	3 870	570	1 000	5 440	34 132	4 996	8 564	47 692
Open-ended QA					37 523	5 343	9 178	52 044

3.2 参数设置

(1) 评估准则

对于多选题问答任务, 本文采用最大化铰链函数 (hinge loss) 来优化模型, 并通过模型回答问题的准确率或百分比来评估模型性能:

$$ACC_{MC} = \frac{1}{|Q_t|} \sum_{q \in Q_t} \left(1 - \prod_{i=1}^M I[a_i \neq p_i] \right) \quad (17)$$

其中, Q_t 表示问答对的数量, M 表示问题的大小, a_i 表示正确答案, p_i 表示预测答案. 而 $I[\cdot]$ 表示一个指示函数, 当两个数相等时其输出值为 1, 反之为 0, 即当预测答案与正确答案完全相同时准确率等于 1, 否则为 0.

对于开放式问答任务, 本文采用最小化交叉熵函数 (cross-entropy loss) 来优化模型, 并通过计算 Wu-Palmer 相似度分数 ($WUPS$)^[77]来评估模型生成答案的质量, 但对于描述性问题中的二进制和计数问题我们选择准确率作为代替.

$WUPS$ 通过寻找最长的公共子序列来衡量语句相似性, 同时参考 WordNet^[78]中的同义词来衡量单词之间的相似度. 如果候选答案与标注答案之间的相似度小于设定的阈值, 则候选答案的得分为 0. 假设给定问题集 Q_t 中的某个问题 q , 模型预测的答案为 $P = \{p_1, p_2, \dots, p_M\}$, 标注的答案为 $A = \{a_1, a_2, \dots, a_M\}$, $WUPS$ 由公式 (18) 可计算二者之间的相似度分数:

$$WUPS(P, A) = \frac{1}{|Q_t|} \sum_{q \in Q_t} \min \left\{ \prod_{p \in P} \max_{a \in A} WUP(p, a), \prod_{a \in A} \max_{p \in P} WUP(a, p) \right\} \times 100 \quad (18)$$

其中, $WUP(p, a)$ 是指根据两个单词在所定义分类目录^[78,79]中的深度 ($depth$) 来计算它们之间的 Wu-Palmer 相似度, 符号化表达如公式 (19) 所示:

$$WUP(p, a) = 2 \times depth(lcs) / (depth(p) + depth(r)) \quad (19)$$

其中, lcs 表示单词 p 、 r 关系最疏远、最不常见的公共祖先单词 (least common ancestor of the words). 显然, 如果两个单词语义更为接近, 它们在分类目录中共享的公共单词也会更多, 深度也会增加, 从而获得更高的 $WUPS$.

(2) 实现细节

实验基于 PyTorch 深度学习框架, 采用现有的视频问答模型 HGA^[40]作为主干网络以实现视频-文本编码器、跨模态融合、答案预测等模块, 并在图卷积层后使用了批归一化层. 需要注意的是, 本文所提出的 GMC 框架并不依赖于特定的基准方法, 而是可以作为一种通用策略无缝嵌入到任何基于图的问答方法中以提升模型的鲁棒性. 本文利用 Adam optimizer^[18]来优化目标函数, 初始化学率为 0.000 1, batch 大小设置为 64, 多选题问答任务迭代次数设置为 50, 开放式问答任务迭代次数为 100, 所有实验均在 NExT-QA 数据集^[3]上进行. 对于 GMC 框架, 在多选题问答任务中其 MASK 模块节点和边的掩蔽概率 ε 、 ξ 均取 0.1, 而开放式问答任务中其 MASK 模块节点和边的掩蔽概率 ε 、 ξ 分别取 0.1 和 0.3, 对比学习模块损失函数的平衡超参数 λ 在多选题问答任务中取 5, 在开放式问答任务中取 8.

3.3 与先进方法的对比实验

为了证明本文所提出的自监督对比学习框架 GMC 的有效性, 我们分析并测试了当前几种性能较好的视频问答模型, 它们涵盖了不同的网络架构和视觉推理技术.

EVQA 模型^[3]使用两个 LSTM 网络分别对问题中的所有单词和视频中的帧进行编码, 然后将问题表示和视频表示融合为统一的融合特征表示, 用于解码的答案, 模型中不包含任何推理模块.

STVQA 模型^[20]通过两个双层 LSTM 对视频和问题建模, 将时间-空间注意力机制整合到 encoder 编码阶段中, 根据问题来决定关注视频中的某一帧及其中的某些区域.

Co-Mem^[31]和 HME 模型^[32]均采用与 STVQA 类似的视频和问题编码器, 但增加了记忆模块 (memory modules), 以多周期的方式对视觉外观、运动特征和语言特征进行推理.

HCRN 模型^[36]是一个以条件关系网络 (CRN) 为构件的分层模型, 它可以根据视觉或文本线索对处理可变长度的视频帧或片段集进行调整, 从而在多粒度上对视频进行推理.

UATT 模型^[23]提出了序列视频注意机制和时间问题注意机制, 并将两种注意力结合应用到开放式视频问答任务中.

在我们的实验中, 模型由神经网络构成, 其中有多项超参数 ε 、 ξ 和 λ , 前面两个参数的搜索范围都为 [0, 1]. 在实验结果中汇报的值是多个源项目进行验证后的平均值. 其中, 每次验证的结果是在最优的参数配置下模型收敛时的平均性能. 我们采用的策略是对于每个目标项目的同一种参数配置, 以其能在所有的源项目取得最优平均值的参数组合作为最佳的参数配置 (代码已发布在 <https://github.com/Feliciayao/GMC.git>).

为便于叙述, 本文将面向多选题问答任务的含 GMC 框架的视频问答模型记作 GMC(MC), 面向开放式问答任务的含 GMC 框架的视频问答模型记作 GMC(OE). 同时在数据增强模块, 在第 j 层 GCN 上使用 N -dropping 方法记作 layer j N -dropping, 其中 $j=1$ 或 2 , N 可以从 Node、Edge 或 None (即不使用数据增强) 中进行选择.

Xiao 等人^[3]的相关实验结果表明, GloVe 和 BERT 在多选题问答和开放式问答任务应用的效果有着显著差距. 因此, 本文在后续 NExT-QA 数据集上开展实验时, 模型针对多选题任务采用微调后的 BERT 模型 (BERT-FT) 提取文本特征, 针对开放式任务采用 GloVe 提取文本特征.

(1) 多选题问答任务实验分析

本实验中, GMC(MC) 模型采用 layer 1_Edge-dropping 和 layer 2_Node-dropping 的数据增强方法, 其中掩蔽概率 ε 、 ξ 分别为 0.1 和 0.1.

表 2 显示了各模型在多选题问答任务的评测分数. 其中 ACC_C 、 ACC_T 、 ACC_D 分别表示模型在回答因果问题、时序问题、描述性问题的正确率, ACC 表示总正确率. STVQA 由于引入了时间注意力模型, 能够捕获到基于时间的视觉线索, 性能相较于 EVQA 有所提高. Co-Mem、HME 中的记忆网络能够有效减少视频信息的丢失, HCRN 在时序问题中的推理具有较高的准确率. 本文提出的 GMC 模型在 NExT-QA 数据集的多选题问答任务中取得了最先进的实验结果, 由于引入基于图自监督的对比学习方法, 提升了回答问题的准确率, 将生成答案的准确率从 49.66% 提升至 50.46%.

表 2 多选题问答任务 (MC-QA) 实验结果 (%)

Model	Text Rep	ACC_C	ACC_T	ACC_D	ACC
EVQA	BERT-FT	42.46	46.34	45.82	44.24
STVQA	BERT-FT	44.76	49.26	55.86	47.94
Co-Mem	BERT-FT	45.22	49.07	55.34	48.04
HCRN	BERT-FT	45.91	49.26	53.67	48.20
HME	BERT-FT	46.18	48.20	58.30	48.72
HGA	BERT-FT	46.14	50.68	59.33	49.66
GMC(MC)	BERT-FT	47.99	50.81	58.69	50.46

(2) 开放式问答任务实验分析

本实验中, GMC(OE) 模型采用 layer 1_Edge-dropping 和 layer 2_Node-dropping 的数据增强方法, 其中掩蔽概率 ε 、 ξ 分别为 0.1 和 0.3。

表 3 总结了在 NEX-T-QA 数据集的开放式问答任务上不同模型声称答案相似度分数的比较结果, 其中 $WUPS_C$ 、 $WUPS_T$ 、 $WUPS_D$ 分别表示模型在因果问题、时序问题、描述性问题所生成答案与标注答案的相似度, $WUPS$ 表示总相似度分数。尽管 Co-Mem 拥有记忆存储模块, 但是 UATT 模型包含的两种时序注意力机制较完整地留存了视频的时序特征和问题的顺序结构信息, 使得最终相似度高于 Co-Mem。HGA 模型在多选题和开放式问答任务中均有着不俗的表现, 这充分体现了基于图卷积网络的关系推理在视频问答任务中的有效性和重要性。

表 3 开放式问答任务 (OE-QA) 实验结果

Model	Text Rep	$WUPS_C$	$WUPS_T$	$WUPS_D$	$WUPS$
Co-Mem	GloVe	10.75	14.06	40.91	18.07
UATT	GloVe	11.80	15.22	45.17	19.83
EVQA	GloVe	12.82	16.18	45.64	20.71
HGA	GloVe	13.60	15.77	45.12	20.86
GMC(OE)	GloVe	14.21	16.90	46.40	21.77

由数据可以看到, 本文提出的 GMC(OE) 模型在所有类型的开放式问题中均得到了先进的结果, 将生成答案的相似度分数从 20.86 提升至 21.77。GMC 框架的嵌入一方面对视觉-语言特征异构图进行数据增强、提高了模型的泛化能力, 另一方面通过增设对比学习的损失函数迫使增强子样本与原样本的预测分布相一致, 提升了视频问答模型的稳健性。

3.4 消去实验及模型分析

本研究单独对所提出的基于图的数据增强方法开展了消去实验, 不引入对比学习损失函数, 只使用监督损失函数, 以观察 GMC 中 MASK 模块对于模型性能提升的贡献, 并且对所设计的数据增强方法中的单个或联合性能进行一些效果分析。同时通过对比实验调节损失函数超参数 λ 使得模型预测性能最佳。

(1) 多选题问答任务实验分析

由于 HGA 模型推理模块由两层 GCN 组成, 我们通过应用不同的数据增强组合方式, 并适当调节 MASK 模块中 Node-dropping 与 Edge-dropping 的概率 ε 、 ξ , 以寻求模型的最佳性能。通过对比实验, 在多选题中 BERT-FT 词嵌入方法相较于 GloVe 效果更好, 故进一步实验中本文均选择使用 BERT-FT。

实验结果如表 4 所示, 第 1 列展示了两种数据增强方法在两层 GCN 上的不同顺序组合, 其中 MASK 部分“+”前表示图网络 layer 1 采用的数据增强方法, “+”后表示 layer 2 采用的数据增强方法, None 表示不对该层 GCN 做数据增强, 即保持原样本。可以看到: Edge(0.1) + Node(0.1) 组合的 MASK 模块效果最佳, 即推理模块中的第 1 层 GCN 添加 Edge-dropping, 第 2 层 GCN 添加 Node-dropping, 此时节点和边的掩蔽概率 ε 、 ξ 均取 0.1, 将生成答案

的准确率由 49.66% 提升到了 50.18%. 另外, 数据显示当单独对推理模块中第 1 层 GCN 的输入图数据进行 Node-dropping 或 Edge-dropping 操作时, 效果不升反降, 这可能是由于图数据中的边和节点信息丢失, 使得 GCN 推理获取的信息依据受到限制、使得模型回答问题的准确率略有下降, 同时也从侧面反映了组合式数据增强方法能够有效提升图网络的泛化和判别能力.

表 4 不同组合的数据增强技术在 HGA 模型上的实验结果 (MC-QA) (%)

MASK	Text Rep	ACC_C	ACC_T	ACC_D	ACC
None+None	GloVe	35.71	38.40	55.60	39.67
None+None	BERT-FT	46.14	50.68	59.33	49.66
Node(0.1)+None	BERT-FT	46.30	49.07	58.30	49.06
Edge(0.1)+None	BERT-FT	46.22	50.31	59.46	49.60
Edge(0.1)+Node(0.2)	BERT-FT	46.87	49.81	58.69	49.66
Edge(0.2)+Node(0.2)	BERT-FT	46.68	50.37	58.30	49.68
Node(0.1)+Node(0.1)	BERT-FT	46.87	50.50	58.30	49.82
Edge(0.1)+Node(0.3)	BERT-FT	47.14	50.37	58.56	49.96
Node(0.1)+Edge(0.1)	BERT-FT	46.57	51.49	58.69	50.04
Edge(0.1)+Edge(0.1)	BERT-FT	47.41	50.19	58.94	50.10
Edge(0.1)+Node(0.1)	BERT-FT	47.95	49.88	58.30	50.18

为进一步观察对比学习部分对模型性能的影响, 基于上述最佳组合的数据增强, 通过调节损失函数中平衡超参数 λ 的数值开展对比实验. 表 5 展示了 λ 不同数值时基于自监督图对比学习的视频问答模型在多选题问答任务中的实验结果, 当 λ 为 5 时, 能够有效提升模型回答问题的准确率.

表 5 不同 λ 设置下的实验结果 (MC-QA)

λ	Text Rep	ACC_C (%)	ACC_T (%)	ACC_D (%)	ACC (%)
0	BERT-FT	47.95	49.88	58.30	50.18
0.5	BERT-FT	46.68	50.06	59.46	49.76
1	BERT-FT	46.14	50.81	58.82	49.62
3	BERT-FT	47.37	50.37	58.30	50.04
5	BERT-FT	47.99	50.81	58.69	50.46
7	BERT-FT	46.53	49.69	58.17	49.36
8	BERT-FT	47.49	49.57	61.13	50.28

(2) 开放式问答任务实验分析

本实验首先将多选题问答任务中表现较好的改进模型迁移到开放式问答任务, 但模型的表现并不是太好, 通过尝试其他不同的数据增强组合方式, 并适当调节 MASK 模块中 Node-dropping 与 Edge-dropping 的概率 ε 、 ξ . 通过对比实验, 在多选题任务中 GloVe 词嵌入方法相较于 BERT-FT 效果更好, 故进一步实验中本文均选择使用 GloVe.

表 6 展示了数据增强方法在开放式问答任务的应用效果, 可以看到 Edge(0.1) + Node(0.3) 组合的 MASK 模块效果最佳. 并且由表 6 可知, 与当前一些先进的视频问答模型相比, 即推理模块中的第 1 层 GCN 添加 Edge-dropping, 第 2 层 GCN 添加 Node-dropping, 此时节点和边的掩蔽概率 ε 、 ξ 分别取 0.1 和 0.3, 将开放式问答任务中生成答案的相似度分数由 20.86 提升到了 21.41. 但在因果推理问题上, 模型的性能虽比原模型有所提升, 但对比其他实验效果不够明显, 我们猜测可能是因为因果问题侧重于考察图网络对多节点之间关系的检索和推理能力, 而此时节点掩盖率的增大导致信息丢失也随之增加.

同样, 本研究也在开放式问答任务中改变损失函数中平衡超参数 λ 的数值, 观察视频问答模型改进后的性能变化. 表 7 展示了不同数值 λ 时基于自监督图对比学习的视频问答模型在多选题问答任务中的实验结果, 当 λ 为 8 时, 能够有效提升模型回答问题的准确率.

表 6 不同组合的数据增强技术在 HGA 模型上的实验结果 (OE-QA)

MASK	Text Rep	$WUPS_C$	$WUPS_T$	$WUPS_D$	$WUPS$
None+None	GloVe	13.60	15.77	45.12	20.86
Node(0.1)+None	GloVe	14.49	15.77	44.03	21.07
Edge(0.2)+Node(0.2)	GloVe	14.26	16.34	44.12	21.15
Node(0.1)+Edge(0.1)	GloVe	14.04	16.07	45.23	21.19
Edge(0.1)+Node(0.2)	GloVe	13.96	16.81	44.58	21.24
Node(0.1)+Node(0.1)	GloVe	14.23	16.36	44.75	21.27
Edge(0.1)+Node(0.1)	GloVe	14.17	16.13	45.28	21.28
Edge(0.1)+Edge(0.1)	GloVe	14.27	16.21	45.21	21.34
Edge(0.1)+None	GloVe	14.33	16.04	45.26	21.37
Edge(0.1)+Node(0.3)	GloVe	13.78	16.63	46.06	21.41

表 7 不同 λ 设置下的实验结果 (OE-QA)

λ	Text Rep	$WUPS_C$	$WUPS_T$	$WUPS_D$	$WUPS$
0	GloVe	13.78	16.63	46.06	21.41
0.5	GloVe	14.43	16.04	44.55	21.23
1	GloVe	14.28	16.04	44.99	21.25
3	GloVe	14.46	16.33	43.63	21.14
6	GloVe	14.05	16.12	46.28	21.43
7	GloVe	14.07	16.40	45.92	21.45
8	GloVe	14.21	16.90	46.40	21.77
10	GloVe	14.51	16.12	43.02	20.97

通过上述实验,对于本文所提出的两种图数据增强方法,两种方法叠加使用的效果可能会比单独使用要好,能够带来较大的性能提升,这也侧面反映了 Node-dropping 和 Edge-dropping 之间存在着一定的互补性.并且如果在第 1 层 GCN 对图节点进行 MASK 的效果普遍不如对边,这可能是由于在删除节点的同时也会丢失与节点相连接的边的信息,而删除边能够保留图样本中所有节点的特征,具有更大的灵活性,使得此时 Node-dropping 的策略更为低效.

3.5 定性分析

在图 3-图 5 中,我们分别给出了采用自监督学习策略后的 GMC 模型在 NExT-QA 数据集上的实验结果展示图.图中加粗标黑部分表示正确答案,标红部分表示模型与正确答案不完全一致的预测结果.对于多选题问答任务和开放式问答任务,本文提出的自监督学习方法能够促进模型很好地得到预测答案.图 4 中的失败案例表明本文提出的方法在背景杂乱和低光照的情况下表现不佳,同时对计数任务的鲁棒性不足,一种可能的改进方式是采用自适应的方式进行图增强,根据不同的视频及问答对的情况动态调整图增强的策略,我们将这一点作为未来的研究方向.



图 3 NExT-QA 数据集多选题问答任务的成功实验案例

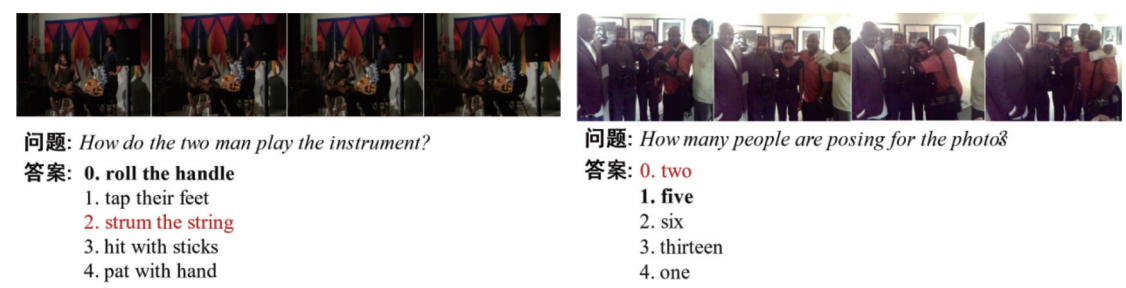


图 4 NExT-QA 数据集多选题问答任务的失败实验案例



图 5 NExT-QA 数据集开放式问答任务的实验案例

4 总 结

本文提出了一种基于 GCN 的 MASK 机制作为视频问答模型的数据增强方法. 针对图数据样本不足的问题, 采用 Node-dropping 和 Edge-dropping 方法的组合进行数据增强, 有效抑制了过拟合, 增强了网络的泛化能力. 之后采用效果较好的 MASK 机制, 通过最小化模型对新旧样本数据预测分布之间的 KL 散度来实现图网络的自监督学习, 进一步提升模型的抗扰动能力. 在 NExT-QA 数据集的实验结果表明, 该方法能够有效提升模型预测答案的能力, 有效抑制了过拟合, 增强了视频问答系统的鲁棒性.

视频问答研究是一个重要且极具挑战性的任务, 现阶段的研究工作仍具有很大的提升空间. 当前多模态的交互中只实现了视觉与语言模态的融合, 没有较多的涉及视频中的音频模态, 未来可以在实现视觉、语言、音频跨模态特征融合方面做更进一步的探索. 同时对于开放式问答任务中自然语言生成答案的评测标准仍然较少, 具有规范性和可解释性的权威数据集能够更好地推动视频问答技术的发展.

References:

[1] Gupta P, Gupta V. A survey of text question answering techniques. Int'l Journal of Computer Applications, 2012, 53(4): 1–8. [doi: 10.5120/8406-2030]

[2] Agrawal A, Lu JS, Antol S, Mitchell M, Zitnick CL, Parikh D, Batra D. VQA: Visual question answering. Int'l Journal of Computer Vision, 2017, 123(1): 4–31. [doi: 10.1007/s11263-016-0966-6]

[3] Xiao JB, Shang XD, Yao A, Chua TS. NExT-QA: Next phase of question-answering to explaining temporal actions. In: Proc. of the 34th IEEE Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9772–9781. [doi: 10.1109/CVPR46437.2021.00965]

[4] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 16th IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[5] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

[6] Chung J, Gulcehre C, Cho KH, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling.

- arXiv:1412.3555, 2014.
- [7] Zhang LJ, Zhou YQ, Duan XY, Chen RQ. A hierarchical multi-input and output bi-GRU model for sentiment analysis on customer reviews. *IOP Conf. Series: Materials Science and Engineering*, 2018, 322(6): 062007. [doi: [10.1088/1757-899X/322/6/062007](https://doi.org/10.1088/1757-899X/322/6/062007)]
 - [8] Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*. Lisbon: The Association for Computational Linguistics, 2015. 1412–1421.
 - [9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: *Proc. of the 31st Int'l Conf. on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017. 6000–6010.
 - [10] Zhang BL. Video question answering based on attention mechanism and graph convolutional network [MS. Thesis]. Harbin: Harbin University of Science and Technology, 2021 (in Chinese with English abstract). [doi: [10.27063/d.cnki.ghlgu.2021.001115](https://doi.org/10.27063/d.cnki.ghlgu.2021.001115)]
 - [11] Sun FY, Hoffmann J, Verma V, Tang J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In: *Proc. of the 8th Int'l Conf. on Learning Representations*. Addis Ababa: OpenReview.net, 2019.
 - [12] Xue DH. Research and application of road risk target detection algorithm based on convolutional neural network [MS. Thesis]. Nanjing: Nanjing University of Posts and Telecommunications, 2021 (in Chinese with English abstract). [doi: [10.27251/d.cnki.gnjdc.2021.001063](https://doi.org/10.27251/d.cnki.gnjdc.2021.001063)]
 - [13] Rong Y, Huang WB, Xu TY, Huang JZ. DropEdge: Towards deep graph convolutional networks on node classification. In: *Proc. of the 8th Int'l Conf. on Learning Representations*. Addis Ababa: OpenReview.net, 2020.
 - [14] Feng WZ, Zhang J, Dong YX, Han Y, Luan HB, Yang Q, Kharlamov E, Tang J. Graph random neural networks for semi-supervised learning on graphs. In: *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2020. 1853.
 - [15] Tao C, Yin ZW, Zhu Q, Li HF. Remote sensing image intelligent interpretation: From supervised learning to self-supervised learning. *Acta Geodaetica et Cartographica Sinica*, 2021, 50(8): 1122–1134 (in Chinese with English abstract). [doi: [10.11947/j.AGCS.2021.20210089](https://doi.org/10.11947/j.AGCS.2021.20210089)]
 - [16] Berthelot D, Carlini N, Goodfellow I, Oliver A, Papernot N, Raffel C. MixMatch: A holistic approach to semi-supervised learning. In: *Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2019. 454.
 - [17] Xie QZ, Dai ZH, Hovy E, Luong MT, Le QV. Unsupervised data augmentation for consistency training. In: *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2020. 525.
 - [18] Quan HB, Yang Y. Survey on language prior research of visual question answering. *China Computer & Communication*, 2022, 34(1): 55–58 (in Chinese with English abstract).
 - [19] You YL, Chen TL, Sui YD, Chen T, Wang ZY, Shen Y. Graph contrastive learning with augmentations. In: *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems*. 2020. 5812–5823.
 - [20] Jang Y, Song YL, Yu Y, Kim Y, Kim G. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In: *Proc. of the 30th IEEE Conf. on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 1359–1367. [doi: [10.1109/CVPR.2017.149](https://doi.org/10.1109/CVPR.2017.149)]
 - [21] Yu Z, Xu DJ, Yu J, Yu T, Zhao Z, Huang YT, Tao DC. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In: *Proc. of the 33rd AAAI Conf. on Artificial Intelligence*. Honolulu: AAAI Press, 2019. 9127–9134. [doi: [10.1609/aaai.v33i01.33019127](https://doi.org/10.1609/aaai.v33i01.33019127)]
 - [22] Zhong YY, Ji W, Xiao JB, Li YC, Deng WH, Chua TS. Video question answering: Datasets, algorithms and challenges. *arXiv:2203.01225*, 2022.
 - [23] Xue HY, Zhao Z, Cai D. Unifying the video and question attentions for open-ended video question answering. *IEEE Trans. on Image Processing*, 2017, 26(12): 5656–5666. [doi: [10.1109/TIP.2017.2746267](https://doi.org/10.1109/TIP.2017.2746267)]
 - [24] Zhao Z, Lin JH, Jiang XH, Cai D, He XF, Zhuang YT. Video question answering via hierarchical dual-level attention network learning. In: *Proc. of the 25th ACM Int'l Conf. on Multimedia*. Mountain View: ACM, 2017. 1050–1058. [doi: [10.1145/3123266.3123364](https://doi.org/10.1145/3123266.3123364)]
 - [25] Zhao Z, Zhang Z, Xiao SW, Yu Z, Yu J, Cai D, Wu F, Zhuang YT. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In: *Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence*. Stockholm: IJCAI.org, 2018. 3683–3689. [doi: [10.24963/ijcai.2018/512](https://doi.org/10.24963/ijcai.2018/512)]
 - [26] Wu M. Video question answering based on deep memory fusion method [MS. Thesis]. Harbin: Harbin University of Science and Technology, 2021 (in Chinese with English abstract). [doi: [10.27063/d.cnki.ghlgu.2021.000211](https://doi.org/10.27063/d.cnki.ghlgu.2021.000211)]
 - [27] Zhu LC, Xu ZW, Yang Y, Hauptmann AG. Uncovering the temporal context for video question answering. *Int'l Journal of Computer Vision*, 2017, 124(3): 409–421. [doi: [10.1007/s11263-017-1033-7](https://doi.org/10.1007/s11263-017-1033-7)]
 - [28] Maharaj T, Ballas N, Rohrbach A, Courville A, Pal C. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In: *Proc. of the 30th IEEE Conf. on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 7359–7368. [doi: [10.1109/CVPR.2017.778](https://doi.org/10.1109/CVPR.2017.778)]

- [29] Lei J, Yu LC, Bansal M, Berg TL. TVQA: Localized, compositional video question answering. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 1369–1379. [doi: [10.18653/v1/D18-1167](https://doi.org/10.18653/v1/D18-1167)]
- [30] Tapaswi M, Zhu YK, Stiefelhagen R, Torralba A, Urtasun R, Fidler S. MovieQA: Understanding stories in movies through question-answering. In: Proc. of the 29th IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4631–4640. [doi: [10.1109/CVPR.2016.501](https://doi.org/10.1109/CVPR.2016.501)]
- [31] Gao JY, Ge RZ, Chen K, Nevatia R. Motion-appearance co-memory networks for video question answering. In: Proc. of the 31st IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6576–6585. [doi: [10.1109/CVPR.2018.00688](https://doi.org/10.1109/CVPR.2018.00688)]
- [32] Fan CY, Zhang XF, Zhang S, Wang WS, Zhang C, Huang H. Heterogeneous memory enhanced multimodal attention model for video question answering. In: Proc. of the 32nd IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1999–2007. [doi: [10.1109/CVPR.2019.00210](https://doi.org/10.1109/CVPR.2019.00210)]
- [33] Li XP, Song JK, Gao LL, Lu XL, Huang WB, He XN, Gan C. Beyond RNNs: Positional self-attention with co-attention for video question answering. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 8658–8665. [doi: [10.1609/aaai.v33i01.33018658](https://doi.org/10.1609/aaai.v33i01.33018658)]
- [34] Yang A, Miech A, Sivic J, Laptev I, Schmid C. Just Ask: Learning to answer questions from millions of narrated videos. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 1666–1677. [doi: [10.1109/ICCV48922.2021.00171](https://doi.org/10.1109/ICCV48922.2021.00171)]
- [35] Zellers R, Lu XM, Hessel J, Yu Y, Park JS, Cao JZ, Farhadi A, Choi Y, Merlot. Multimodal neural script knowledge models. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. 2021. 23634–23651.
- [36] Le TM, Le V, Venkatesh S, Tran T. Hierarchical conditional relation networks for video question answering. In: Proc. of the 33rd IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9969–9978. [doi: [10.1109/CVPR42600.2020.00999](https://doi.org/10.1109/CVPR42600.2020.00999)]
- [37] Yi KX, Wu JJ, Gan C, Torralba A, Kohli P, Tenenbaum JB. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 1039–1050.
- [38] Yi KX, Gan C, Li YZ, Kohli P, WU JJ, Torralba A, Tenenbaum JB. CLEVRER: Collision events for video representation and reasoning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.
- [39] Chen ZF, Mao JY, Wu JJ, Wong KYK, Tenenbaum JB, Gan C. Grounding physical concepts of objects and events through dynamic visual reasoning. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.
- [40] Jiang P, Han YH. Reasoning with heterogeneous graph alignment for video question answering. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 11109–11116. [doi: [10.1609/aaai.v34i07.6767](https://doi.org/10.1609/aaai.v34i07.6767)]
- [41] Park J, Lee J, Sohn K. Bridge to answer: Structure-aware graph interaction network for video question answering. In: Proc. of the 34th IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15521–15530. [doi: [10.1109/CVPR46437.2021.01527](https://doi.org/10.1109/CVPR46437.2021.01527)]
- [42] Wang JY, Bao BK, Xu CS. DualVGR: A dual-visual graph reasoning unit for video question answering. IEEE Trans. on Multimedia, 2021, 24: 3369–3380. [doi: [10.1109/TMM.2021.3097171](https://doi.org/10.1109/TMM.2021.3097171)]
- [43] Liu F, Liu J, Wang WN, Lu HQ. HAIR: Hierarchical visual-semantic relational reasoning for video question answering. In: Proc. of the 34th IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1678–1687. [doi: [10.1109/ICCV48922.2021.00172](https://doi.org/10.1109/ICCV48922.2021.00172)]
- [44] Peng M, Wang C, Gao Y, Shi Y, Zhou XD. Multilevel hierarchical network with multiscale sampling for video question answering. arXiv:2205.04061, 2022.
- [45] Xiao JB, Yao A, Liu ZY, Li YC, Ji W, Chua TS. Video as conditional graph hierarchy for multi-granular question answering. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 2804–2812. [doi: [10.1609/aaai.v36i3.20184](https://doi.org/10.1609/aaai.v36i3.20184)]
- [46] Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: ICLR, 2014.
- [47] Henaff M, Bruna J, LeCun Y. Deep convolutional networks on graph-structured data. arXiv:1506.05163, 2015.
- [48] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: ACM, 2016. 3844–3852.
- [49] Levie R, Monti F, Bresson X, Bronstein MM. CayleyNets: Graph convolutional neural networks with complex rational spectral filters. IEEE Trans. on Signal Processing, 2018, 67(1): 97–109. [doi: [10.1109/TSP.2018.2879624](https://doi.org/10.1109/TSP.2018.2879624)]
- [50] Wu ZH, Pan SR, Chen FW, Long GD, Zhang CQ, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(1): 4–24. [doi: [10.1109/TNNLS.2020.2978386](https://doi.org/10.1109/TNNLS.2020.2978386)]

- [51] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: PMLR, 2016. 2014–2023.
- [52] Hamilton WL, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1025–1035.
- [53] Gao HY, Wang ZY, Ji SW. Large-scale learnable graph convolutional networks. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 1416–1424. [doi: [10.1145/3219819.3219947](https://doi.org/10.1145/3219819.3219947)]
- [54] Chen XX. Graph convolutional neural network algorithm for link prediction [MS. Thesis]. Guangzhou: Guangdong University of Technology, 2021 (in Chinese with English abstract). [doi: [10.27029/d.cnki.ggdgu.2021.000538](https://doi.org/10.27029/d.cnki.ggdgu.2021.000538)]
- [55] Li YJ, Tarlow D, Brockschmidt M, Zemel RS. Gated graph sequence neural networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan: ICLR, 2016.
- [56] Verma V, Qu M, Lamb A, Bengio Y, Kannala J, Tang J. GraphMix: Regularized training of graph neural networks for semi-supervised learning. arXiv:1909.11715, 2019.
- [57] Ding M, Tang J, Zhang J. Semi-supervised learning on graphs with generative adversarial nets. In: Proc. of the 27th ACM Int'l Conf. on Information and Knowledge Management. Torino: ACM, 2018. 913–922. [doi: [10.1145/3269206.3271768](https://doi.org/10.1145/3269206.3271768)]
- [58] Li QM, Han ZC, Wu XM. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 3538–3545.
- [59] Deng ZJ, Dong YP, Zhu J. Batch virtual adversarial training for graph convolutional networks. arXiv:1902.09192, 2019.
- [60] Feng FL, He XN, Tang J, Chua TS. Graph adversarial training: Dynamically regularizing based on graph structure. IEEE Trans. on Knowledge and Data Engineering, 2021, 33(6): 2493–2504. [doi: [10.1109/TKDE.2019.2957786](https://doi.org/10.1109/TKDE.2019.2957786)]
- [61] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.
- [62] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2242–2251. [doi: [10.1109/ICCV.2017.244](https://doi.org/10.1109/ICCV.2017.244)]
- [63] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proc. of the 32nd IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4396–4405. [doi: [10.1109/CVPR.2019.00453](https://doi.org/10.1109/CVPR.2019.00453)]
- [64] Kim T, Cha M, Kim H, Lee JK, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1857–1865.
- [65] Wu LR, Lin HT, Tan C, Gao ZY, Li SZ. Self-supervised learning on graphs: Contrastive, generative, or predictive. IEEE Trans. on Knowledge and Data Engineering, 2021: 1–20. [doi: [10.1109/TKDE.2021.3131584](https://doi.org/10.1109/TKDE.2021.3131584)]
- [66] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 149.
- [67] Grill JB, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, Doersch C, Pires BÁ, Guo ZD, Azar MG, Piot B, Kavukcuoglu K, Munos R, Valko M. Bootstrap your own latent: a new approach to self-supervised learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1786.
- [68] Suresh S, Li P, Hao C, Neville J. Adversarial graph augmentation to improve graph contrastive learning. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. 2021. 15920–15933.
- [69] Peng Z, Huang WB, Luo MN, Zheng QH, Rong Y, Xu TY. Graph representation learning via graphical mutual information maximization. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 259–270. [doi: [10.1145/3366423.3380112](https://doi.org/10.1145/3366423.3380112)]
- [70] Veličković P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD. Deep graph infomax. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.
- [71] Li P, Wang YB, Wang HW, Leskovec J. Distance encoding: Design provably more powerful neural networks for graph representation learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. 2020. 4465–4478.
- [72] Jiao YY, Xiong Y, Zhang JW, Zhang Y, Zhang TQ, Zhu YY. Sub-graph contrast for scalable self-supervised graph representation learning. In: Proc. of the 2020 IEEE Int'l Conf. on Data Mining (ICDM). Sorrento: IEEE, 2020. 222–231. [doi: [10.1109/ICDM50108.2020.00031](https://doi.org/10.1109/ICDM50108.2020.00031)]
- [73] Qiu JZ, Chen QB, Dong YX, Zhang J, Yang HX, Ding M. GCC: Graph contrastive coding for graph neural network pre-training. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM, 2020. 1150–1160. [doi: [10.1145/3394486.3403168](https://doi.org/10.1145/3394486.3403168)]
- [74] Falcon A, Lanz O, Serra G. Data augmentation techniques for the video question answering task. In: Proc. of the 2020 European Conf. on Computer Vision. Glasgow: Springer, 2020. 511–525. [doi: [10.1007/978-3-030-66415-2_33](https://doi.org/10.1007/978-3-030-66415-2_33)]

- [75] Chen J, Ma TF, Xiao C. FastGCN: Fast learning with graph convolutional networks via importance sampling. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.
- [76] Shang XD, Di DL, Xiao JB, Cao Y, Yang X, Chua TS. Annotating objects and relations in user-generated videos. In: Proc. of the 2019 Int'l Conf. on Multimedia Retrieval. Ottawa: ACM, 2019. 279–287. [doi: [10.1145/3323873.3325056](https://doi.org/10.1145/3323873.3325056)]
- [77] Wu Z, Palmer M. Verb semantics and lexical selection. arXiv:cmp-lg/9406033, 1994.
- [78] Miller GA. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39–41.
- [79] Chapelle CA. Vocabulary and language for specific purposes. In: The Encyclopedia of Applied Linguistics. Wiley Online Library. 2012.

附中文参考文献:

- [10] 张博伦. 基于注意力机制与图卷积网络的视频问答研究 [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2021. [doi: [10.27063/d.cnki.ghlgu.2021.001115](https://doi.org/10.27063/d.cnki.ghlgu.2021.001115)]
- [12] 薛东辉. 基于卷积神经网络的道路风险目标检测模型研究与应用 [硕士学位论文]. 南京: 南京邮电大学, 2021. [doi: [10.27251/d.cnki.gnjdc.2021.001063](https://doi.org/10.27251/d.cnki.gnjdc.2021.001063)]
- [15] 陶超, 阴紫薇, 朱庆, 李海峰. 遥感影像智能解译: 从监督学习到自监督学习. 测绘学报, 2021, 50(8): 1122–1134. [doi: [10.11947/j.AGCS.2021.20210089](https://doi.org/10.11947/j.AGCS.2021.20210089)]
- [18] 权海波, 杨颖. 视觉问答语言先验性研究综述. 信息与电脑(理论版), 2022, 34(1): 55–58.
- [26] 吴猛. 基于深度记忆融合方法的视频问答研究 [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2021. [doi: [10.27063/d.cnki.ghlgu.2021.000211](https://doi.org/10.27063/d.cnki.ghlgu.2021.000211)]
- [54] 陈学信. 面向链接预测的图卷积神经网络算法研究 [硕士学位论文]. 广州: 广东工业大学, 2021. [doi: [10.27029/d.cnki.ggdgu.2021.000538](https://doi.org/10.27029/d.cnki.ggdgu.2021.000538)]



姚暄(2000—), 女, 硕士生, 主要研究领域为计算机视觉, 多媒体计算.



徐常胜(1969—), 男, 博士, 研究员, CCF 杰出会员, 主要研究领域为多媒体分析/索引/检索, 模式识别, 计算机视觉.



高君宇(1994—), 男, 博士, 助理研究员, CCF 专业会员, 主要研究领域为计算机视觉, 多媒体计算.