

文章编号: 1006-3080(2026)03-0418-10

DOI: 10.14135/j.cnki.1006-3080.20250910003

多频思维链推理的情景式小样本视觉问答方法

魏一凡, 张 静

(华东理工大学信息科学与工程学院, 上海 200237)

摘要: 现有情景式小样本视觉问答方法多依赖于空间域视觉特征, 忽略了频域信息在解耦全局与局部语义中的潜力, 导致模型对图像内容的理解层次不足。此外, 传统推理机制难以实现跨频率的渐进式交互, 限制了跨模态推理能力的进一步提升。针对情景式小样本视觉问答任务, 提出了一种基于多频思维链的跨模态推理网络。该方法通过频率域信息对图像内容进行深度过滤, 设计自适应高低频分离模块解耦视觉模态中不同层次的语义表达, 并构建多模态高低频思维链推理机制, 利用认知银行存储跨模态推理中间结果, 结合双路径思维链充分挖掘过滤后的视觉信息, 以实现更准确的情景式小样本视觉问答。实验表明, 该方法显著增强了模型对小样本视觉问答任务的处理能力, 提升了跨模态推理的准确性和泛化性能, 在公开数据集上的实验结果优于同类算法。

关键词: 情景式小样本视觉问答; 跨模态推理网络; 自适应高低频分离模块; 思维链; 双路径

中图分类号: TP389.1

文献标志码: A

近年来, 深度学习技术极大程度地推动了视觉语言处理领域的发展, 使其成为人工智能研究中最具前景的方向之一。作为视觉语言处理领域的核心任务, 视觉问答因其独特的跨模态特性而备受学界关注——该任务不仅需要模型具备强大的计算机视觉能力来理解图像内容, 同时要求其拥有出色的自然语言处理能力来解析问题语义, 更重要的是必须建立有效的跨模态交互机制以实现视觉与语言模态的深度融合与推理。这种复杂的多模态特性使视觉问答任务成为评估模型跨模态理解能力的基准测试。

深度学习模型的数据依赖性是其在实际应用中的主要瓶颈之一。由于需要大量高质量标注数据进行训练, 在现实场景中这类模型往往面临显著的数据获取成本挑战, 构建大规模的图像-问题-答案三元组数据集通常需要耗费大量人力资源进行标注。自视觉问答任务提出以来^[1], 研究方法主要围绕特征增强与模态关系建模两个方向推进。在特征增强方

面, 从早期全局特征到区域特征^[2-4]、注意力机制^[5]及特征集成方法^[6-7]的演进, 显著提升了模型对关键区域的感知能力。在模态关系建模方面, 从基础注意力机制^[8-9]到双线性池化^[10]、Transformer 架构^[11-12]及多粒度建模^[13-14]的发展, 实现了跨模态特征的深度融合。这些系统性进展共同推动了视觉问答模型在复杂推理场景下的性能提升。然而, 尽管方法不断演进, 视觉问答模型仍受限于训练数据的质量与分布。本研究通过对主流 VQA v2 数据集的分析表明, 虽然该数据集具有多模态多样性, 但在许多答案类别上存在明显的长尾分布问题。具体而言, 大量低频答案类别的样本数量严重不足, 这种数据稀疏性直接制约了模型跨模态推理能力的提升。

针对这一挑战, 研究者们开始探索小样本学习在视觉问答任务中的应用。小样本学习作为模拟人类从少量示例中学习的关键技术, 在目标检测与识别等领域取得了显著进展^[15-17]。该领域研究主要分为 3 个方向: 元学习方法通过多任务训练学习更优的

收稿日期: 2025-09-10

作者简介: 魏一凡(2000—), 男, 山西大同人, 硕士生, 研究方向为计算机视觉、神经网络和小样本视觉问答

通信联系人: 张 静, E-mail: jingzhang@ecust.edu.cn

引用本文: 魏一凡, 张 静. 多频思维链推理的情景式小样本视觉问答方法 [J]. 华东理工大学学报(自然科学版), 2026, 52(3): 418-427.

Citation: WEI Yifan, ZHANG Jing. Episodic Few-Shot Visual Question Answering Method with Multi-Frequency Chain-of-Thought Reasoning[J]. Journal of East China University of Science and Technology, 2026, 52(3): 418-427.

初始化参数^[18];度量学习方法从基础嵌入空间优化^[19]发展到分布层面的距离度量,包括 EMD、DeepBDC 和 AMMD 等更精细的分布对齐策略^[20-22];特征增强方法则通过语义指导^[23]、任务信息利用^[24]、局部特征引入^[25]和多模态融合^[26]等多种途径丰富样本表示。近年来,研究者进一步提出了多阶段特征提取与自注意力机制^[27]、加权特征嵌入与自适应度量^[28]以及原型重合成与语义对齐^[29]等创新方法,持续推动着小样本学习方法的完善。由于小样本学习方法能够使模型从有限样本中快速学习有效特征表示,这使其成为解决视觉问答数据稀疏性问题的理想选择,情景式小样本视觉问答(Episodic Few-Shot Visual Question Answering, EFSVQA)应运而生。在相关研究中,Dong 等^[30]提出了基于度量学习的联合嵌入框架,通过优化视觉和语言模态的联合表示空间来提升模型性能。Zhang 等^[31]则从跨模态推理的层面出发,通过对 COCO 和 VQAv2 数据集的系统清洗和重构,构建了专门针对情景式小样本视觉问答任务的平衡数据集。同时,还创新性地引入了信息熵校准策略,通过优化图像分布的信息熵来增强多模态交互效果,显著提升了模型在该任务上的表现。Yin 等^[32]则提出了一个两阶段的网络结构,运用多层注意机制分别捕获模态内和模态间的关系以构建视觉特征图和语义关系图来完成多模态视觉语义学习。

本文提出了一种基于多频思维链的跨模态推理网络,旨在解决情景式小样本视觉问答问题。该网络从频域视角出发,利用自适应高低频分离模块,通过参数化的快速傅里叶变换,自适应将图像解耦为低频图像分量和高频图像分量。在深度过滤视觉模态信息的基础上,本文提出了基于思维链的跨频率渐进式推理机制,通过认知银行保存基于多频信息进行跨模态推理的中间结果,并借助双路径思维链机制引导多频多模态特征的协同处理,在保持不同频率特征独立性的同时实现渐进式推理,最终得到更精准反应类别信息的链校准原型。

1 模型与算法

1.1 问题描述

情景式小样本视觉问答是一个 N -way K -shot 的分类问题。从答案不同的图像-问题对中采样 N 个类别,每个类别 K 个样本,对应的支持集 $\{S\}$ 中包含了 $N \times K$ 个样本。图像-问题对的答案集 $\{A\}$ 中则包含了样本相应的标签。如果从每个类别中采样 M 个查询样本,那么查询集 $\{Q\}$ 中就包含了 $N \times M$ 个样

本。由于情景式小样本视觉问答是一个多模态问题,可以将每个查询集依据模态进行划分,即视觉模态子集 $\{Q^I\}$,文本模态子集 $\{Q^T\}$ 和答案集 $\{Q^A\}$,支持集则被组成类似的模式,即 $\{S^I\}$ 、 $\{S^T\}$ 和 $\{S^A\}$ 。通过合并查询集和支持集的样本,可以得到视觉模态集合 $I = \{Q^I, S^I\}$,文本模态集合 $T = \{Q^T, S^T\}$ 以及答案集 $A = \{Q^A, S^A\}$ 。

本文中,视觉模态信息 $\{I\}^{N \times (K+M)}$ 和文本模态信息 $\{T\}^{N \times (K+M)}$ 的特征分别通过视觉嵌入 $\psi(\cdot; \theta_\psi)$ 和文本嵌入 $\phi(\cdot; \theta_\phi)$ 提取,在分类阶段,融合后的多模态特征将会被分成对应的支持集 $\{S^{\text{mul}}\}^{N \times K}$ 和查询集 $\{Q^{\text{mul}}\}^{N \times M}$ 。支持集为基于多频思维链的跨模态推理网络的输入,其通过最小化对应查询集上的损失来训练模型。

1.2 基于多频思维链的跨模态推理网络

本文主要研究在情景式小样本视觉问答任务中,如何利用频率域信息的特点对视觉模态信息进行过滤,以缓解空间域中不同层次的语义信息相互耦合的问题,从而帮助模型更深入地理解图像内容。利用过滤后的视觉、文本模态信息通过基于思维链的双分支架构进行渐进式推理,从而提高模型的跨模态推理能力,并提升模型在视觉问答任务上的表现效果。

如图 1 所示,在基于多频思维链的跨模态推理网络(Frequency-based Cross-modal Chain-of-Thought Inference Network, FC-CoT-IN)中,图像首先被输入到自适应高低频分离模块中,通过参数化的快速傅里叶变换,自适应将图像解耦为低频图像分量和高频图像分量:低频图像分量聚焦宏观语义结构以建立全局理解,高频图像分量捕获细粒度纹理细节以增强局部感知,以实现模型对视觉模态信息中多层次语义的精准建模。过滤后的图像将会被送入 SwinT 提取特征,而问题的令牌则由预训练的 GloVe 词嵌入提取获得。视觉和文本两种模态的特征将会作为多模态高低频思维链推理机制的输入,过程中产生的多频多模态特征将通过认知银行被保存下来,同时通过双路径思维链机制的引导,得到更精准反应类别信息的链校准原型,最后通过最小化多频思维链损失函数对模型的预测分布进行校准。通过对视觉模态信息过滤和解耦对思维链推理机制的引导,模型能够充分利用频率域信息弥补空间域信息的不足,加深模型对于视觉模态信息的理解,从而提升最终的跨模态推理效果。

1.3 自适应高低频分离模块

频率域信息在各种计算机视觉任务中往往能够

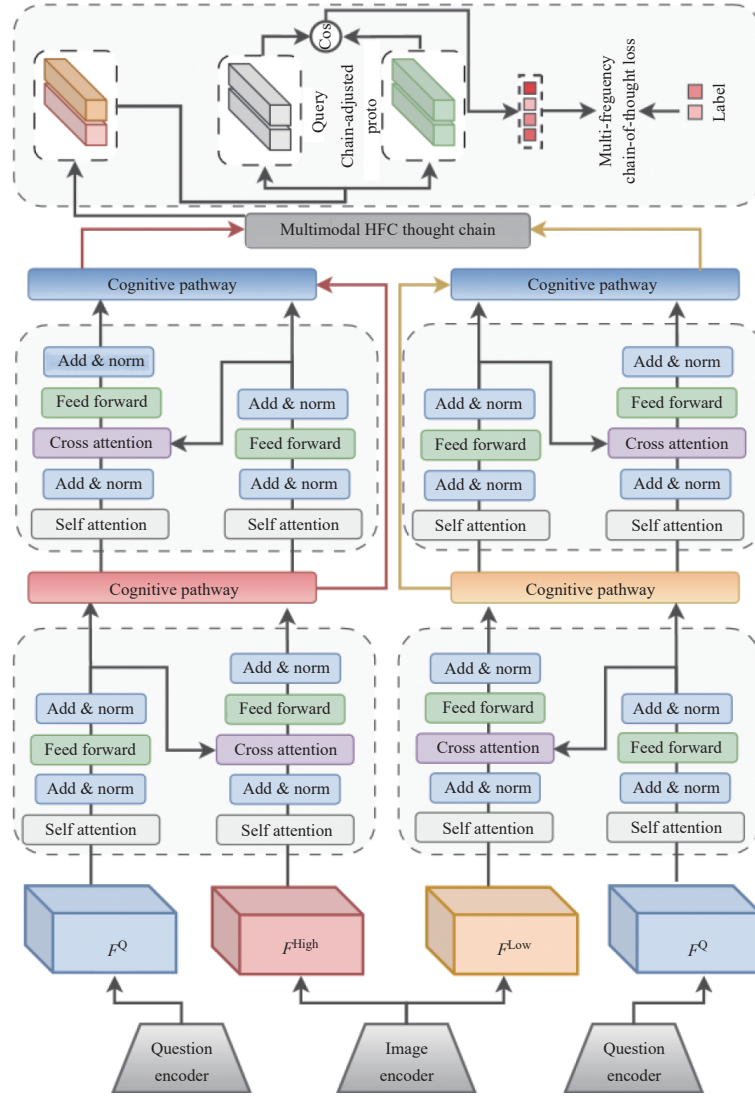


图 1 基于多频思维链的跨模态推理网络的框架图

Fig. 1 Framework of cross-modal reasoning network based on multi-frequency chain-of-thought

起到重要作用,不同的频率分量能够传达对于视觉模态理解的不同层次。然而,直接利用频域信息可能引入人类难以理解的频域成分,同时传统方法对高低频的划分往往过于绝对,难以自适应地提取图像中整体或细节的信息。针对这些问题,本文提出了自适应高低频分离模块(Adaptive High-Low Frequency Separation Module, AHLS)。

首先,对所有图像进行预处理,包括平移和裁剪。这一步处理依据不同的特征提取骨干网络进行,例如,针对 SwinT 所得到的预处理图像大小为 224×224 。对预处理后的图像进行如公式(1)所示的离散二维傅里叶变换:

$$I^f[u, v] = \sum_{m=0}^{N'-1} \sum_{n=0}^{N'-1} I^s[m, n] e^{-j2\pi \left(\frac{mu+nv}{N} \right)} \quad (1)$$

其中, u, v 表示离散二维傅里叶变换的频率域坐标,而 $I^s[m, n]$ 是输入图像的空间域值, N' 是图像大小,

如 84 或 224。对于计算过后的 $I^f[u, v]$, 通过频谱搬移操作将低频分量移至中心,同时定义了两个可参数化的掩码 $\text{Mask}^{\text{high}}$ 和 Mask^{low} , 掩码半径由可调节的比率参数控制。过滤操作如式(2)所示:

$$I_{\text{low}}^F = I^f[u, v] \odot \text{Mask}^{\text{low}}, \quad I_{\text{high}}^F = I^f[u, v] \odot \text{Mask}^{\text{high}} \quad (2)$$

低通滤波 I_{low}^F 保留了以频谱中心为原点、一定半径范围内的低频成分。而高通滤波 I_{high}^F 则抑制该中心区域,同时保留了高频边缘信息。滤波后的频谱经由逆傅里叶变换重建为空间域的图像,作为后续的多模态高低频思维链推理机制的输入。

通过上述操作,自适应高低频分离模块能够输出平滑后的宏观结构信息和细节增强的高频成分。频率域的参数化掩码会根据图像内容与任务需求进行自适应的优化,避免了传统固定阈值分频的局限性,增强了模型对不同频段信息的感知与利用能力。

1.4 多模态高低频思维链推理机制

思维链(Chain of Thought, Cot)是一种通过模拟人类逐步推理过程来提升模型解决复杂问题能力的技术,其本质是通过“过程透明化”弥补模型单步推理的不足,提升模型的推理能力。对于通过自适应高低频分离模块过滤的不同视觉模态信息,本研究将其分成高频分支和低频分支,这两部分分支分别捕捉图像中不同层次的语义信息,并分别结合文本模态做出对应的回答。

视觉模态信息和文本模态信息的特征分别通过视觉嵌入 $\psi(\cdot; \theta_\psi)$ 和文本嵌入 $\phi(\cdot; \theta_\phi)$ 提取,具体过程如式(3)所示:

$$\begin{aligned} F^{lh} &= \psi(I_{\text{high}}^F; \theta_\psi), F^{ll} = \psi(I_{\text{low}}^F; \theta_\psi) \\ F^Q &= \phi(T; \theta_\phi) \end{aligned} \quad (3)$$

式中, F^{lh} 、 F^{ll} 和 F^Q 分别为高频视觉特征、低频视觉特征和问题特征。首先,为了实现小样本情况下的推理过程,研究通过多头自注意力机制和交叉注意力机制,实现视觉模态和文本模态之间的细粒度交互,从而得到问题指导下的视觉模态特征。以高频过滤信息为例,多头交叉注意力通过不同的“头”独立计算,能够捕捉到图像中的不同区域与问题内容的相互依赖,将计算的权重对值特征进行加权,可以得到问题指导下的视觉模态特征,其表示为式(4):

$$\begin{aligned} F^{lh} &= \text{MHSA}(F^{lh}, F^Q, F^Q) = [h_1, h_2, \dots, h_t] W_h \\ h_i &= \text{softmax} \left(\frac{(W_i^Q F^{lh})(W_i^K F^Q)^T}{\sqrt{d}} \right) (W_i^V F^Q) \end{aligned} \quad (4)$$

其中查询特征为 $F^{lh} \in \mathbb{R}^{n \times d}$, 键特征和值特征均表示为 $F^Q \in \mathbb{R}^{n \times d}$, 而 W_i^Q, W_i^K, W_i^V 是其中的投影矩阵, $W_h \in \mathbb{R}^{h \times d_h \times d}$, d 是每个“头”的维度。问题指导下的低频视觉模态特征 F^{ll} 也可以以同样的方式得到。

然后,通过认知银行保存了不同分支的多模态特征和对应的后验分布,作为思维链的中间结果,为后续参与进一步的推理做准备。后验分布和对应的多模态特征计算式见式(5),它们共同构成了模型的初级推理状态:

$$\begin{aligned} P^{lh}(y_i = j | f_i^d) &= \frac{\exp(\psi(q_i^d, p_j^d))}{\sum_{j=1}^N \exp(\psi(q_i^d, p_j^d))} \\ F_{\text{step}}^{h_1} &= \text{GELU}(W_0^{(1)} \cdot \text{Concat}[F^{lh}, F^Q] + b_0^{(1)}) \end{aligned} \quad (5)$$

其中, $j \in \{1, \dots, N\}$ 为类别索引, $\psi(q_i^d, p_j^d)$ 表示 d 分量中第 i 个查询样本 q_i^d 和第 j 类域原型 p_j^d 之间的

余弦相似度, $W_0^{(1)}$ 和 $b_0^{(1)}$ 分别为初级推理中线性变换的权重矩阵和偏置。低频后验分布 $P^{ll}(y_i = j | f_i^d)$ 也可以通过类似的方式得到。

模型通过双向跨模态交互机制进一步深化视觉-文本特征的融合。具体来说,在初步交互之后,模型以问题作为注意力机制的查询特征,通过交换两种模态信息的注意力角色实现更深层次的协同推理,如式(6)所示:

$$F^{Qh} = \text{MHSA}(F^Q, F^{lh}, F^{lh}) \quad (6)$$

类似地,低频分支执行相同的双向交互操作。通过这种对称的跨模态交互模式,能够有效克服传统单向注意力的结构偏差。

完成二次交互之后,同样通过认知银行保存不同分支的后验分布作为思维链的中间结果,如式(7)所示:

$$P_2^{lh}(y_i = j | f_i^d) = \frac{\exp(\psi(q_i^d, p_j^d))}{\sum_{j=1}^N \exp(\psi(q_i^d, p_j^d))}$$

$$F_{\text{step}}^{h_2} = \text{GELU}(W_0^{(2)} \cdot \text{Concat}[F^{lh'}, F^{Qh}] + b_0^{(2)}) \quad (7)$$

其中, $F^{lh'}$ 为经过双向交互后的高频视觉特征, $W_0^{(2)}$ 和 $b_0^{(2)}$ 分别为高级推理中线性变换的权重矩阵和偏置。该计算过程与首次计算一致,但是基于更新后的特征进行计算,能够反映深度交互之后样本与类别原型之间的语义关联变化。二次计算得到的概率分布和特征表示共同构成了思维链中的高级推理状态。

作为思维链的核心,认知银行保存的多分支中间结果将会作为最终预测的辅助。首先,最终的预测分支需要综合多频多模态特征,如式(8)所示:

$$\begin{aligned} F_{\text{multi}}^{\text{final}} &= W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot \text{Concat}[F^{lh'}, F^{ll'}] + b_1) \oplus \\ &\quad F^{Qh} + b_2) \oplus F^{Ql} + b_3 \end{aligned} \quad (8)$$

其中, W_1, W_2, W_3 为3个线性层的权重矩阵,而 b_1, b_2, b_3 为3个线性层的偏置。接下来,模型通过渐进式思维链推理实现从局部到全局,从细节到整体的协同推理,通过多次迭代优化提升推理的准确性和鲁棒性。

高频分支作为思维链推理的开始阶段,专注于图像中的细节信息和局部特征,该阶段接收融合特征 $F_{\text{multi}}^{\text{final}}$ 和认知银行中保存的初级推理状态 $P^{lh}(y_i = j | f_i^d)$ 、 $F_{\text{step}}^{h_1}$ 以及高级推理状态 $P_2^{lh}(y_i = j | f_i^d)$ 、 $F_{\text{step}}^{h_2}$ 作为输入。在这一阶段,模型通过计算各特征的预测确定性,筛选出在高频细节层面具有一致性的样本子集。具体而言,模型分别从3个特征空间中提

取 Top-K 高确定性样本,然后通过集合交集操作获得共识样本集,如式 (9) 所示:

$$c_{\text{fuse}} = \max \phi(F_{\text{multi}}^{\text{final}}, P), c_{h1} = \max P^{h1}, c_{h2} = \max P^{h2}$$

$$I_{\text{con}}^j = \text{TopK}(c_{\text{fuse}}^j) \cap \text{TopK}(c_{h1}^j) \cap \text{TopK}(c_{h2}^j) \quad (9)$$

其中, $c_{\text{fuse}}, c_{h1}, c_{h2}$ 分别为基于多频、初级推理状态、高级推理状态中多模态特征对应的预测确定性, P 代表基于 $F_{\text{multi}}^{\text{final}}$ 计算得到的类原型, I_{con}^j 代表共识样本集。这些共识样本代表了需要通过细节进行问题回复的样本多模态特征,其加权组合形成了精炼后的高频原型表示,该阶段推理过程如式 (10) 所示:

$$P_h^{\text{new}} = \sum_{i \in I_{\text{con}}} \omega_{\text{multi}}(F_{\text{multi},i}^{\text{final}}) + \omega_{h1}(F_{h1,i}) + \omega_{h2}(F_{h2,i}) \quad (10)$$

其中, $\omega_{\text{multi}}, \omega_{h1}, \omega_{h2}$ 分别为基于多频、初级推理状态、高级推理状态中多模态特征的样本权重,由预测确定性加权平均得到,最终的高频分支原型 P_h^{refined} 则由 P_h^{new} 和 P 加权得到。

低频分支作为思维链推理的中间阶段,专注于图像的宏观结构和全局语义信息。该阶段采用和第一阶段类似的精炼策略,输入认知银行中保存的对应的低频分支的初级、高级推理状态。在这一阶段,模型关注的是通过图片全局结构进行问题回复的样本,通过类似的共识机制进行筛选。样本的多模态特征加权融合得到了最终的低频分支原型 P_l^{refined} ,与第一阶段形成互补。

最后,融合决策精炼作为最终的处理阶段以实现多频特征的协同整合。这一阶段的独特之处在于直接基于前两个阶段产生的由精炼原型所计算的后验分布进行整体优化。通过自适应权重机制动态调整高频和低频特征的贡献比例,最终产生融合多频段信息的链校准原型 P_{hl}^{refined} 作为最终的分类依据。

在训练阶段,则通过多频思维链损失 (Multi-Frequency Cot Loss) 函数对 P_{hl}^{refined} 进行约束以进行训练,让模型完成对于答案集 A 的预测。多频思维链损失函数在交叉熵损失函数的基础上设定阈值,在训练过程中,控制余弦相似度超过阈值的样本更加紧凑、而低于阈值的负样本更加远离,从而能够更好地对链校准原型进行优化。

2 实验方法与结果

2.1 数据集及实验设置

本文使用了 Toronto COCO-QA^[33]、FS COCO-QA^[31]、FS VG-QA^[31] 和 FS VQA^[31] 这 4 个数据集。

Toronto COCO-QA 为每个问题-答案 (QA) 对分配一种属性,包括对象,数量,颜色和位置,共计包含了 78 736 个训练问题, 38 948 个测试问题和 123 287 张图片。而 FS COCO-QA、FS VG-QA 和 FS VQA 分别是 Toronto COCO-QA、Visual Genome-QA 和 VQA v2 的小样本版本,其充分考虑到了数据集的不平衡问题,是为了解决情景式小样本视觉问答问题所构造的小样本数据集。

Visual Genome-QA 的图片来自 MSCOCO 和 YFCC100M 数据集,共计包含了 108 249 张图片和 170 万 QA 对,数据集集中的问题包括了“What”、“Where”、“How”、“When”、“Who”和“Why”等类型。FS VQA 包括 80 000 图片和 444 000 QA 对的训练集, 40 000 图片、214 000 QA 对的验证集和 80 000 图片、444 000 QA 的测试集。FS VG-QA 和 FS VQA 所选择的 QA 对不包含重复图像,且充分考虑到了数据平衡问题。3 个小样本数据集的数据规模如表 1 所示。

表 1 FS COCO-QA, FS VG-QA 和 FS VQA 数据集的数据规模
Table 1 Data scale of the FS COCO-QA, FS VG-QA, and FS VQA datasets

Data	Pair			Class		
	Train	Val	Test	Train	Val	Test
FS COCO-QA	10 575	2 968	2 968	195	57	55
FS VG-QA	22 537	6 304	6 304	568	153	165
FS VQA	36 835	10 375	10 375	749	205	207

在实验中,使用 Swin-Transformer 作为视觉特征提取骨干网络,视觉模态信息在经过视觉特征提取器后,被处理成 768 维的特征。在问题的处理上,则将问题长度限制为 15,并使用预训练的 GloVe 将单词的嵌入维度调整到 300 维。在元学习阶段,使用 SGD (Stochastic Gradient Descent) 优化器训练模型,初始学习率为 2.5×10^{-4} ,权重衰减设置为 5×10^{-4} 。

2.2 消融实验

本研究还进行了一系列的消融实验,以验证基于多频思维链的跨模态推理网络中模块有效性。表 2 展示了每个模块在 FS COCO-QA 数据集上消融结果。Case 1 是一个简单的跨模态模型,使用残差网络作为视觉模态的特征提取骨干网络以及全局向量词嵌入模型作为问题特征提取骨干网络,将提取后的特征通过交叉注意机制进行跨模态融合,最终送入分类器中进行答案分类。

从表中可以看出,所有提出的模块和方法都有

表 2 FS COCO-QA 数据集的消融实验结果

Table 2 Ablation experimental results on the FS COCO-QA dataset

Case	AHLC	CoT	CoT-loss	5-way accuracy/%	
				1shot	5shot
1				61.73	75.80
2	√			63.34	77.38
3	√	√		66.73	79.02
4	√	√	√	67.45	80.06

助于提升情景式小样本视觉问答问题的预测准确率。Case2 引入了频域信息,并通过自适应高低频分离模块、利用频率域信息对空间域的视觉模态信息进行过滤。因此,与 Case1 相比,Case2 准确率分别提升了 1.61、1.58 个百分点。同样的,Case3 也引入了频域信息,并在后续推理过程中引入了高低频思维链推理机制,相比于 Case1 分别取得了 5.00、3.22 个百分点的提升。在 Case3 的基础上,研究进一步引入了多频思维链损失函数,对基于思维链的迭代推理过程进行约束,分别实现了 0.72、1.04 个百分

点的提升。模型最终的准确率相比于 case1 分别实现了 5.72、4.26 个百分点的提升。

相比于简单的跨模态推理过程,基于多频思维链的跨模态推理网络中每个模块均有提升,这进一步证明了通过频率域信息对不同层次的空间域语义信息进行解耦以提升模型对于视觉模态的理解,同时通过基于思维链的迭代推理过程保持不同频率特征的独立性处理,并建立跨频率的渐进式推理路径的意义。

2.3 对比实验

为了证明所提出的方法的有效性,本文将其与现有的几种基于 Toronto COCO-QA, FS COCO-QA, FS VG-QA 和 FS VQA 的方法进行了比较,实验结果如表 3 和表 4 所示,其中 FC-CoT-IN-Res 表示以残差网络作为特征提取骨干网络、基于多频思维链的跨模态推理网络,而 FC-CoT-IN-Swint 表示以 swint 作为特征提取骨干网络、基于多频思维链的跨模态推理网络。本文所提出的方法部分设置上优于当前最先进的方法。

表 3 FS COCO-QA 和 FS VQA 对比实验

Table 3 Comparative experiments on FS COCO-QA and FS VQA dataset

Method	Classification accuracy/%							
	FS COCO-QA				FS VQA			
	5 way		10 way		5 way		10 way	
	1shot	5shot	1shot	5shot	1shot	5shot	1shot	5shot
FEAT ^[25]	59.90	72.62	46.81	61.58	60.88	74.03	50.13	64.97
GCN ^[29]	61.84	74.72	47.69	62.26	63.20	76.25	51.81	67.05
MatchingNet ^[34]	61.50	75.36	47.32	64.05	63.07	80.74	51.01	72.68
ProtoNet ^[19]	61.64	75.76	47.52	64.83	65.80	80.51	54.59	73.41
DeepBDC ^[21]	56.39	72.70	45.83	63.07	68.77	83.05	60.44	77.28
CDCIN-Res ^[31]	66.11	79.43	52.36	68.83	78.75	88.85	69.50	82.43
CDCIN-Swint ^[31]	74.30	86.02	63.12	77.94	80.23	90.14	71.88	84.67
FC-CoT-IN-Res	67.45	80.06	54.82	68.50	79.35	89.04	71.14	82.59
FC-CoT-IN-Swint	74.73	86.10	62.98	76.85	81.33	90.21	71.62	83.53

从表 3 和表 4 中的实验结果可得,基于多频思维链的跨模态推理网络在一定设置下达到了性能最优。经典的小样本学习方法(如 FEAT, GCN, MatchingNet 和 ProtoNet)扩展到情景式小样本视觉问答领域,但这些方法并不是专门为了跨模态任务所设计的,因而不能在情景式小样本视觉问答领域取得最优。本研究尝试将 DeepBDC 扩展到情景式小样本视觉问答领域,在之前的小样本图像分类任务中,该方法通过度量计算的优化取得了性能提升,但缺乏专门设计的跨模态融合机制导致其性能受

限。相比之下,基于多频思维链的跨模态推理网络专门针对情景式小样本视觉问答任务所设计,通过频率域信息解耦图像中不同层次的语义信息,并通过多频思维链推理机制构建跨频率的迭代推理过程。在增强模型对视觉模态理解的同时提升其跨模态推理能力。在 FS COCO-QA 数据集的 5way、1shot 设置下,相较于传统小样本方法的性能最高者(GCN),本文提出的模型分类准确率提高了 5.61 个百分点。在 5way、5shot 下,相较于最优的 ProtoNet 提升了 4.30 个百分点。同样的,在 10way、1shot 设

表 4 FS VG-QA 数据集的分类准确率

Table 4 Classification accuracy on the FS VG-QA dataset

Method	Classification accuracy/%			
	5 way		10 way	
	1shot	5shot	1shot	5shot
FEAT ^[25]	69.65	80.39	57.33	70.70
GCN ^[29]	70.71	81.63	58.47	71.24
MatchingNet ^[34]	70.45	83.87	57.96	75.14
ProtoNet ^[19]	71.19	83.97	60.06	74.73
DeepBDC ^[21]	77.37	86.95	66.70	79.60
CDCIN-Res ^[31]	80.52	89.35	70.22	82.16
CDCIN-Swint ^[31]	83.43	91.59	73.79	85.38
FC-CoT-IN-Res	82.45	89.39	70.61	82.74
FC-CoT-IN-Swin	85.44	91.43	72.61	84.30

置下,相较于最优的 GCN 提升了 7.13 个百分点,而在 10way、5shot 的设置下,本文所提出的模型较之最优的 ProtoNet 提升了 3.67 个百分点。在其他数据集上亦取得了一致的提升。而相较于当前情景式小样本视觉问答方法中最先进的方法 CDCIN,本文提出的模型在 3 个数据集的 5way、1shot 设置下获得了一致的提升,最高在 FS VG-QA 数据集的 swint 特征提取骨干网中获得了 2.01 个百分点的性能提升。在 5way、5shot 的设置下,基于多频思维链的跨模态推理网络则基本超过了 CDCIN,只有在 FS VG-QA 数据集的 swint 特征提取骨干网中以 0.16 个百分点的差距略逊色于 CDCIN,最高则在 FS COCO-QA 数据集的残差网络特征提取骨干网中获得了 0.63 个百分点的性能提升。而在 10 way 设置中,本文的模型则和 CDCIN 互有胜负,总的来说,以残差网络作为特征提取骨干网络时,基于多频思维链的跨模态推理网络能展现出更稳定的竞争优势,这可能是由于 swint 作为特征提取骨干网时过于强大,其提取的深层视觉特征已经具备了高度丰富的语义信息,反而会削弱经过频率域过滤后的特征所起到的补充作用。

另外,本文还在 Dong 等^[30]清理的数据集上进行了对比实验,如表 5 所示。由结果可得,基于多频思维链的跨模态推理网络在 5way、1shot、5way、5shot、10way、1shot、10way、5shot 的设置下,均超过了专门为解决情景式小样本视觉问答问题所设计的方法 FPAIT-CNN, FPAIF+CLT 和 HGAT-Res。基于多频思维链的跨模态推理网络通过基于思维链的跨频率渐进式推理,相较 HGAT-Res 分别实现了 10.12、8.26、12.00、14.38 个百分点的分类准确率提升,同时

表 5 Toronto COCO-QA 数据集的分类准确率

Table 5 Classification accuracy on the Toronto COCO-QA dataset

Method	Classification accuracy/%			
	5 way		10 way	
	1shot	5shot	1shot	5shot
FPAIT-CNN ^[30]	59.38	71.92	45.11	60.20
FPAIT+CLT ^[30]	60.61	72.17	46.37	60.92
Relation Net ^[35]	61.75	71.89	45.60	60.13
EGNN ^[36]	62.21	73.41	46.99	60.01
HGAT-Res ^[32]	63.13	75.41	48.10	61.50
GCN-Res ^[37]	64.20	76.85	51.36	66.12
FETA-Res ^[25]	62.67	75.18	49.53	64.56
ProtoNet-Res ^[19]	63.08	78.40	50.74	67.63
CDCIN-Res ^[31]	72.26	84.18	59.70	75.65
CDCIN-Swint ^[31]	79.08	89.48	69.27	82.49
FC-CoT-IN-Res	73.25	83.67	60.10	75.88
FC-CoT-IN-Swint	79.23	88.51	69.13	81.79

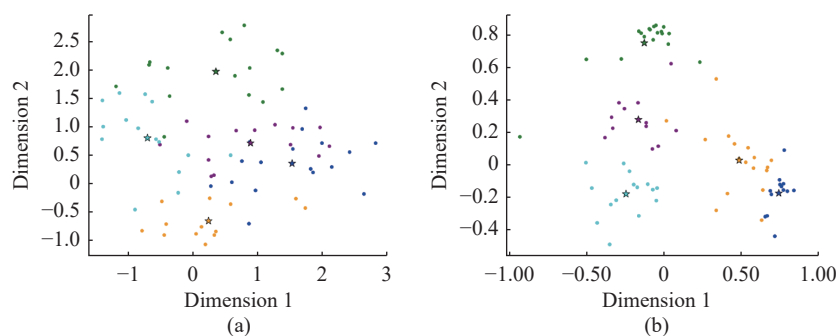
大幅度超过了 FPAIT-CNN、FPAIT+CLT。与最先进的 CDCIN 算法相比,模型仍然呈现和在上述 3 个数据集上类似的表现,和 CDCIN 互有胜负,在以残差网络作为特征提取骨干网络时展现出更稳定的竞争优势。值得注意的是,本文的方法并没有采用和 CDCIN 一致的预训练过程。

2.4 原理分析

图 2 可视化了引入频域信息进行基于思维链推理前后的多模态特征分布,图中展示了 FS COCO-QA 数据集 5way、1shot 设置下的特征分布示例。图 2(a) 表示没有引入频域信息进行推理的多模态特征分布,当没有引入频域信息时,模型只能捕获耦合的空域信息进行推理,缺乏对于视觉模态的深入理解和跨模态推理,从而使得特征分布之间的边界模糊。图 2(b) 所示为引入频域信息并进行基于思维链的跨频率迭代推理之后的特征分布。由于频域信息不同的过滤结果能体现对视觉模态不同层次的语义理解,因此,通过在认知银行保存基于多频信息进行跨模态推理的中间结果的方式和双路径思维链机制的使用,模型可以在保证不同频率特征独立性处理的同时进行跨频率的渐进式推理,更充分地利用深度过滤的视觉模态信息,以实现跨模态推理能力的提升,并提高分类性能。

2.5 样例研究

图 3 展示了从 FS COCO-QA 数据集中抽取的一些小样本任务示例,体现了本文提出的基于多频思



Note that “☆” means the prototype of each class and “o” means the feature distribution of each query sample

图2 在5way 1shot 设置下, 未引入频域信息(a)和引入频域信息并进行基于思维链的推理后(b)的多模态特征可视化图

Fig. 2 Under the 5way 1shot setting, multimodal feature visualization without incorporating frequency domain information (a) and with incorporating frequency domain information and performing chain-of-thought reasoning (b)



图3 模型在 FS COCO-QA 数据集上预测任务的典型案例

Fig. 3 Typical cases of the model's predictive tasks on the FS COCO-QA dataset

维链的跨模态推理网络的实验结果。第1行的所有图片及问题-答案对为支持集的样本, 第2行则包含了从查询集中获取的3个测试样本。在第1个测试样本中, 帽子的占比相对较大, 模型能够更准确地关注到回答问题所需的区域, 从而实现跨模态推理并预测答案。

与传统的小样本图像分类不同, FSVQA 任务要求对多个模态的理解。例如, 在第二行第2个测试样本中, 问题所询问的内容为桌上摆放着什么, 如果过度关注局部细节, 模型会将其错误地预测为“dish”或“plate”。而本文所提出的模型通过频率域对图像进行过滤, 平衡模型对问题引导下的全局视觉特征和局部视觉特征的关注, 并通过跨频率的迭代推理过程提升模型的跨模态推理能力, 从而产生了更精准的预测。

此外, 情景式小样本视觉问答还面临着对文本信息的深入理解和对模型准确回答问题能力的挑战。在支持集中, “plate”和“dish”之间存在高度语义

关联。尽管本文提出的模型在情景式小样本视觉问答任务的许多方面表现良好, 但在需要深度理解文本模态的第3个测试样本中, 仍然存在一定的局限性, 因此有时会产生错误的结果。

3 结束语

本文提出了一种基于多频思维链的跨模态推理网络, 用于解决情景式小样本视觉问答问题。该网络通过引入频率域信息对图像内容进行深度过滤, 利用不同的过滤结果解耦视觉模态不同层次的语义理解。同时, 本文还提出了多模态高低频思维链推理机制, 通过在认知银行保存基于多频信息进行跨模态推理的中间结果和双路径思维链机制的方式, 更充分地利用深度过滤的视觉模态信息, 以实现跨模态推理能力的提升。未来工作中, 研究者计划探索结合大模型的小样本推理, 以进一步提高情景式小样本视觉问答任务的准确性。

参考文献:

- [1] ANTOL S, AGRAWAL A, LU J, *et al.* Vqa: Visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 2425-2433.
- [2] SHIH K J, SINGH S, HOIEM D. Where to look: Focus regions for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Cancun, Mexico: IEEE, 2016: 4613-4621.
- [3] QIAO T, DONG J, XU D. Exploring human-like attention supervision in visual question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018: 2794-2800.
- [4] ZHOU Y, REN T, ZHU C, *et al.* Trar: Routing the attention spans in transformer for visual question answering[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 2074-2084.
- [5] ANDERSON P, HE X, BUEHLER C, *et al.* Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6077-6086.
- [6] CHOWDHURY S, SONI B. Envqa: Improving visual question answering model by enriching the visual feature[J]. Engineering Applications of Artificial Intelligence, 2025, 142: 109948.
- [7] HE K, CHEN X, XIE S, *et al.* Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022: 16000-16009.
- [8] 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. Cancun, Mexico: IEEE, 2016: 21-29.
- [9] XU H, SAENKO K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2016: 451-466.
- [10] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[C]// Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc, 2018.
- [11] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need[C]// Advances in neural information processing systems. Long Beach, CA, USA: Curran Associates Inc, 2017.
- [12] YU Z, YU J, CUI Y, *et al.* Deep modular co-attention networks for visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 6281-6290.
- [13] GU M, ZHAO Z, JIN W, *et al.* Video dialog via multi-grained convolutional self-attention context multimodal networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(12): 4453-4466.
- [14] NGUYEN B X, DO T, TRAN H, *et al.* Coarse-to-fine reasoning for visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 4558-4566.
- [15] 李耀, 李梅. 基于多尺度 WideResNet 的铁轨缺陷小样本检测算法 [J]. 华东理工大学学报 (自然科学版), 2025, 51(6): 843-849.
- [16] 全志楠, 林家骏. 文本无关的小样本手写汉字笔迹鉴别方法 [J]. 华东理工大学学报 (自然科学版), 2018, 44(6): 882-886.
- [17] 沈震宇, 朱昌明, 王喆. 基于 MAML 算法的 YOLOv3 目标检测模型 [J]. 华东理工大学学报 (自然科学版), 2022, 48(1): 112-119.
- [18] WANG W, BAO H, DONG L, *et al.* Image as a foreign language: Beit pretraining for vision and vision-language tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 19175-19186.
- [19] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc, 2017.
- [20] ZHANG C, CAI Y, LIN G, *et al.* DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 12203-12213.
- [21] XIE J, LONG F, LYU J, *et al.* Joint distribution matters: Deep brownian distance covariance for few-shot classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 7972-7981.
- [22] WU J, WANG S, SUN J. AMMD: Attentive maximum mean discrepancy for few-shot image classification[J]. Pattern Recognition, 2024, 155: 110680.
- [23] WU H, YE G, ZHOU Z, *et al.* Dual-view data hallucination with semantic relation guidance for few-shot image recognition[J]. IEEE Transactions on Multimedia, 2024, 26: 9332-9344.
- [24] ZHANG M, ZHANG J, LU Z, *et al.* IEPT: Instance-level and episode-level pretext tasks for few-shot learning[C]//International Conference on Learning Representations. Virtual: ICLR, 2021: 16482-16497.
- [25] YE H J, HU H, ZHAN D C, *et al.* Few-shot learning via embedding adaptation with set-to-set functions[C]//

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 8808-8817.
- [26] LI W, WANG L, XU J, *et al.* Revisiting local descriptor based image-to-class measure for few-shot learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 7260-7268.
- [27] ASKARI F, FATEH A, MOHAMMADI M R. Enhancing few-shot image classification through learnable multi-scale embedding and attention mechanisms[J]. Neural Networks, 2025, 187: 107339.
- [28] CHEN Y, XU Z, WANG J, *et al.* Adaptive set-level metric for few-Shot image classification[J]. Neural Networks, 2025: 107924.
- [29] DONG M, LI F, LI Z, *et al.* PRSN: Prototype resynthesis network with cross-image semantic alignment for few-shot image classification[J]. Pattern Recognition, 2025, 159: 111122.
- [30] DONG X, ZHU L, ZHANG D, *et al.* Fast parameter adaptation for few-shot image captioning and visual question answering[C]//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea: ACM, 2018: 54-62.
- [31] ZHANG J, LIU X, CHEN M, *et al.* Cross-modal feature distribution calibration for few-shot visual question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024, 38(7): 7151-7159.
- [32] YIN C, WU K, CHE Z, *et al.* Hierarchical graph attention network for few-shot visual-semantic learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 2177-2186.
- [33] REN M, KIROUS R, ZEMEL R. Exploring models and data for image question answering[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015). Montreal, Canada: MIT Press, 2015: 2953-2961.
- [34] VINALS O, BLUNDELL C, LILLICRAP T, *et al.* Matching networks for one shot learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016). Barcelona, Spain: Curran Associates Inc, 2016: 3637-3645.
- [35] SUNG F, YANG Y, ZHANG L, *et al.* Learning to compare: Relation network for few-shot learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1199-1208.
- [36] KIM J, KIM T, KIM S, *et al.* Edge-labeling graph neural network for few-shot learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 11-20.
- [37] GARCIA V, BRUNA J. Few-shot learning with graph neural networks [EB/OL]. (2017-11-10). <https://arxiv.org/abs/1711.04043>

Episodic Few-Shot Visual Question Answering Method with Multi-Frequency Chain-of-Thought Reasoning

WEI Yifan, ZHANG Jing

(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: To address the challenges of episodic few-shot visual question answering, we propose a Frequency-based Cross-modal Chain-of-Thought Inference Network (FC-CoT-In). The method adaptively filters image content through frequency-domain information, designs an Adaptive High-Low Frequency Separation Module (AHLS) to decouple semantic representations at different levels in the visual modality, and constructs a multimodal high-low frequency chain-of-thought reasoning mechanism. By leveraging a cognitive bank to store intermediate results of cross-modal reasoning and combining a dual-path chain of thought, the approach fully exploits filtered visual information to achieve more accurate episodic few-shot visual question answering. Experiments demonstrate that the proposed method significantly enhances the model's capability in handling Episodic Few-shot Visual Question Answering (EFSVQA) tasks, improves the accuracy and generalization performance of cross-modal reasoning, and outperforms comparable algorithms on public benchmarks.

Key words: EFSVQA; cross-modal reason network; AHLS; CoT; dual path

(责任编辑: 王晓丽)