计算机科学
COMPUTER SCIENCE

基于反事实思考的视觉问答方法

袁德森, 刘修敬, 吴庆波, 李宏亮, 孟凡满, 颜庆义, 许林峰

引用本文

袁德森, 刘修敬, 吴庆波, 李宏亮, 孟凡满, 颜庆义, 许林峰基于反事实思考的视觉问答方法[J]. 计算机科学, 2022, 49(12): 229-235.

YUAN De-sen, LIU Xiu-jing, WU Qing-bo, LI Hong-liang, MENG Fan-man, NGAN King-ngi, XU Lin-feng. Visual Question Answering Method Based on Counterfactual Thinking[J]. Computer Science, 2022, 49(12): 229-235.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600038

基于反事实思考的视觉问答方法

袁德森　刘修敬　吴庆波　李宏亮　孟凡满　颜庆义　许林峰
电子科技大学信息与通信工程学院　成都 611730
(desenyuan97@163.com)

摘　要　视觉问答是一项结合计算机视觉和自然语言处理的多模态任务,具有极大的挑战性。然而,目前的视觉问答模型存在着严重的语言偏见问题,对其鲁棒性有负面影响。以往的研究主要集中在利用生成反事实样本来辅助模型解决语言偏见。然而,这些研究忽略了分析反事实样本与原始样本的预测差异以及关键特征与非关键特征之间的两两差异。文中通过建立反事实思考流程,结合因果推理与对比学习,使模型能够区分原始样本、事实样本和反事实样本。基于此,提出了一种基于反事实样本的对比学习范式。通过对比3类样本对的特征差异和预测差异,减小了模型的语言偏见。在VQA-CP v2等数据集上的实验证明了所提方法的有效性。与CL-VQA方法相比,所提方法的整体精度提高了0.19%,平均精度提高了0.89%,尤其是Num精度提高了2.6%。相比CSSVQA方法,所提方法的鲁棒性辅助指标Gap从0.96提高到了0.45。

关键词:视觉问答;因果推理;反事实思考;对比学习;深度学习
中图法分类号　TP391

Visual Question Answering Method Based on Counterfactual Thinking

YUAN De-sen,LIU Xiu-jing,WU Qing-bo,LI Hong-liang,MENG Fan-man,NGAN King-ngi and XU Lin-feng
School of Information and Communication Engineering,University of Electronic Science and Technology of China,Chengdu 611730,China

Abstract　Visual question answering(VQA) is a multi-modal task that combines computer vision and natural language processing,which is extremely challenging. However,the current VQA model is often misled by the apparent correlation in the data,and the output of the model is directly guided by language bias. Many previous researches focus on solving language bias and assisting the model via counterfactual sample methods. These studies,however,ignore the prediction information and the difference between key features and non-key features in counterfactual samples. The proposed model can distinguish the difference between the original sample,the factual sample and the counterfactual sample. In view of this,this paper proposes a paradigm of contrastive learning based on counterfactual samples. By comparing these three samples in terms of feature gaps and prediction gaps,the VQA model has been significantly improved in its robustness. Compared with CL-VQA method,the overall precision,average precision and Num index of this method improves by 0.19%,0.89% and 2.6% respectively. Compared with the CSSVQA method,the Gap of the proposed method decrease to 0.45 from 0.96.

Keywords　Visual question answering,Causal inference,Counterfactual thinking,Contrastive learning,Deep learning

1　引言

视觉问答[1-9](Visual Question Answering,VQA)是计算机视觉和自然语言处理的跨领域任务。在多模态机器学习的应用和研究中,视觉问答问题变得越来越重要。在过去的几十年里,计算机视觉和自然语言处理取得了重大的研究进展,可获取和处理的视觉数据和文本数据呈爆炸式增长。在最常见的视觉问答(VQA)形式中,数据包括一张图片和一个问题,需要机器对其中的问题给出相应的答案。与其他计算机视觉任务相比,模型需要回答的问题是实时变化的,并非提前给出;同时,视觉问答任务要求模型理解图像和文本的多模态信息的内容,更符合人工智能的真实形式,可以帮助模型更深入地理解视觉和语言。

目前,视觉问答任务仍然是一个具有挑战性和开放性的研究课题。在视觉问答领域,如何解决语言偏见问题是近年来的研究热点[10-13]。语言偏见给视觉问答的落地带来了较大的负面影响,也说明目前的视觉问答模型在多模态信息的理解上是不足的。如图1所示,对于许多视觉问答模型而言,答案可直接通过语言数据推断出来,如模型会倾向性地直接回答是或者否。另一个经典的例子是,对于“图中香蕉是什么颜色”,虽然图中香蕉未成熟为“绿色”,但模型仍倾向于预测为“黄色”。VQA中的这种回答偏差即为语言偏见,它使得

到稿日期:2022-06-06　返修日期:2022-08-16
基金项目:国家自然科学基金(61831005,61971095)
This work was supported by the National Natural Science Foundation of China(61831005,61971095).
通信作者:吴庆波(qbwu@uestc.edu.cn)

模型在回答问题时依赖于问题与答案之间的表面相关性，而忽略了图像信息。

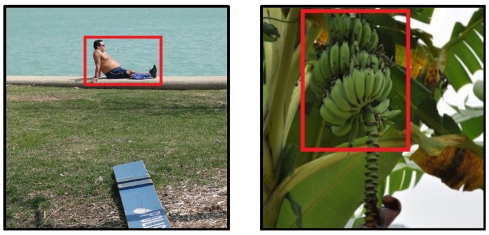


图 1 VQA 中语言偏见的例子(电子版为彩图)

Fig. 1 Example of language bias in VQA

语言偏见问题直接影响模型的输出。例如，对于所有关于香蕉颜色的问题，模型都回答“黄色”，但忽略了图像中实际上是一个绿色的香蕉。如果问题类型是“是”或“否”，则模型倾向于简单地回答“是”。

语言偏见的产生是由于训练集中数据分布的极端不平衡，如数据中 95% 的香蕉均为黄色，模型在测试时对绿色香蕉也倾向性地回答为黄色，从而导致模型的泛化性不佳。为了解决语言偏见问题，近年来出现了大量关于如何解决语言偏见的研究。利用数据增广来解决语言偏见的方法具有集成便捷且简单有效的特点，因而得到了广泛的使用。例如，Chen 等[14]提出了一种制造反事实样本的方法，并把反事实样本作为数据增广添加到训练中，但并未在特征和预测层次中构建区分反事实样本的过程，未深入挖掘样本之间的关系。然而，由于增广的反事实样本均是合成的，有近似外观的特点(如反事实样本的局部 mask)，同时事实样本与同类的自然图像样本又有差别，这会导致增广后正负样本可能在特征空间混淆，同时未进行区分的混淆样本会提高模型的复杂度，不利于模型的泛化。Liang 等[15]使用反事实样本结合对比学习来增强视觉问答的鲁棒性，但其仅从特征层面考虑了对比关系，且只分析了原始样本的对比关系，其对比关系不完整，也未考虑预测层面的特征映射关系，易产生预测混淆。因此，如何分析原始样本和反事实样本之间的全面关系，并构建多层次的思考步骤极为重要。

为了解决特征和预测的混淆，提高模型的泛化能力，本文依据神经科学和因果理论的发现提出了反事实思考的视觉问答方法。神经科学研究表明，人类通过对事物进行不同层次的比较来了解和认识世界。受 Judea Pearl 所建立的因果分析理论中的因果之梯启发，本文构建了三元的特征层面反事实对比关系以及预测层面的反事实预测结果之间的对比关系，从而构建出了一个反事实思考模块。该模块通过多层次的对比学习可以降低增广样本和原始样本之间的特征和预测混淆，从而实现降低模型复杂度和提高模型泛化能力的目的。从概念上说，该方法类似于因果之梯，其中包含的想象、反思以及理解 3 个步骤缺一不可。本文在 VQA-CP v2[16]和 VQA v2[13]数据集上进行了实验，并验证了所提方法的有效性。

2 相关工作

2.1 视觉问答中的语言偏见问题

VQA 中的语言偏见在实际场景的应用中有负面影响。导致语言偏见的一个直接原因是数据集中的训练数据的问题和答案之间通常有很强的相关性。此外，这些问题往往与图像中最为明显的物体有关。在 VQA v1[8]和 v2 中，一个肯定的答案或一个与问题相关的答案往往具有较高的准确性。当训练集和测试集的问题和答案分布不一致时，这种语言偏见是显而易见的。为了评估该现象，研究者们提出使用 VQA-CP v2 数据集来评估语言偏见问题。VQA-CP v2 在训练集和测试集中有着不同的问答分布。目前，根据 Yuan[7]的研究，针对语言偏见的方法可依据其原理分为三大类，分别是：1)强化视觉信息[17-18]，通过强化模型中视觉信息的重要性，建模视觉目标之间的关系，如 HINT[17]等，但其效果一般；2)弱化语言先验[15,19-21]，通过对语言偏见进行建模并使用集成的方法对偏见进行剔除，如 LMH[21]和 Rubi[20]等，其优势是可明显减缓语言偏见，但该类方法损失了语言输入中有用的先验信息；3)使用各种数据增强方法[14,22]，利用数据增强来平衡数据的分布，此类方法简单有效，如 CSSVQA[14]。

2.2 视觉问答中的因果推理

Goodfellow 等[23]尝试使用 GAN 网络合成反事实的图片，并用于视觉问答任务。Teney 等[24]则通过设置阈值掩盖图像中的关键特征来生成反事实样本。Chen 等[14]使用 Grad Cam 方法，通过梯度计算视觉特征和文本特征的关键贡献，并选择 Top-k 目标进行掩盖，从而获得最终的反事实样本。此外，Liang 等[15]使用反事实样本结合对比学习来增强视觉问答的鲁棒性，但其仅从特征层面考虑了对比关系，且只分析了原始样本的对比关系，没有构建完成整体的反事实思考流程。

3 基于反事实的视觉问答方法

本文提出的基于反事实的视觉问答方法分为 3 部分：1)基础视觉问答模块，主要基于 UpDn[25]和 LMH[21]完成，实现基础的视觉问答功能；2)反事实样本生成，利用 Grad Cam 方法生成反事实样本，用于数据增强和后续的反事实思考；3)反事实思考模块，计算原始样本与反事实样本在特征和预测层面的区别，增强模型的理解能力。本文方法框图如图 2 所示。

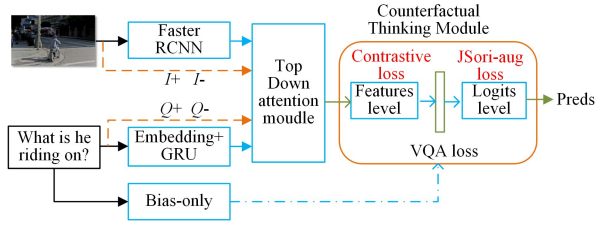


图 2 所提模型的整体示意图

Fig. 2 Overall schematic diagram of the proposed model

3.1 基础视觉问答模块

目前，视觉问答模型一般可建模为多分类问题，为了解决

该多分类问题，研究者们提出了许多用于 VQA 的基本模型。视觉问答问题的一般形式为：给定数据集 $D=\{I_i,Q_i,a_i\}^N$，其包含 N 对图像 I_i、问题 Q_i、回答 a_i。

视觉问答任务的目标是学习映射函数 $f_{vqa}:I\times Q\rightarrow[0,1]^{|A|}$，它可以生成任何给定的图像-问题对的回答分布。在下面的叙述中，本文省略下标 i。

UpDn[25]模型是目前主流的视觉问答模型，是研究者们所使用的最多的模型之一。该模型继承了一种自顶向下和自底向上的注意力机制，并将其应用于视觉场景理解和视觉问答系统的相关问题。采用 Faster-rcnn[26]提取图像中的感兴趣区域，获取目标特征；使用 LSTM[27]和自顶向下的注意力模型学习特征对应的权值，从而实现对视觉图像的深入理解。图 2 的左半部分给出了该模型的网络结构，同时本文采用了经典的去偏方法 LMH[21]来建模语言偏见，即图 2 中的偏差分支，其仅通过输入语言信息来建模语言偏见，并利用集成的方法来减小语言偏见。该方法可被集成于各类去偏方法，如 CSSVQA 等。本文模型包括 3 部分：基础视觉问答网络、反事实样本生成和反事实思考模块。

对于每个问题 Q 和图像 I，UpDn 模型分别使用一个问题编码器 eq 和一个物体检测器 iq 来提取一组单词嵌入 Q 和一组视觉嵌入 V。然后将 V 和 Q 同时输入模型，得到混合特征 $F(V,Q)$，再将特征输入分类器 C，得到最终的预测结果。

$$P_{vqa}(a|I,Q)=f_{vqa}(V_e,Q_e)=C(F(V_e,Q_e)) \tag{1}$$

3.2 反事实样本生成

因果推理[28-30]和数据增强方法近年来得到了广泛的应用，因此有许多反事实样本生成方法可供本文选择。在视觉问答领域中，Chen 等[14]提出的 CSSVQA 方法使用最广泛，其样本生成不需要事先存储，计算效率更高，可直接集成在其他网络中。因此，本文的反事实对比学习范式中采用了 Chen 等提出的反事实样本合成方法。其中 (I^+,I^-,Q^+,Q^-) 分别为生成的反事实样本，CSS 为方法函数。

$$(I^+,I^-,Q^+,Q^-)=CSS(f_{vqa},(I,Q,a)) \tag{2}$$

3.3 反事实思考模块

3.3.1 反事实思考流程

反事实思考是根据人类的思考过程所提出的一种思考模型，近年来研究火热。本文结合对比学习来完整建模这一反事实思考过程，从而帮助视觉问答模型深入理解多模态数据。

本文构建了三元的特征层面反事实对比关系，并进一步构建了预测层面的反事实预测结果对比关系，从而构建出了一个反事实思考模块。本文受到 Pearl 等[28-30]所构建的因果理论的启发，本文模型遵循了人类认知的第三个层次——反事实思考，其内容包括想象、反思和理解。本文方法正对应于这样一个反事实思考流程，各个步骤缺一不可。这种反事实思考机制参考了婴儿对于因果关系的理解，本文参考了 Harari[31]、Weisberg 等[32]、Roese 等[33]和 Brigard 等[34]相关的研究成果。3 部分对应内容如下。

(1)想象：对应于模型利用原始样本和梯度来生成反事实样本，从而完成反事实思考中的想象部分，即模型的输入。

(2)反思：对应于在特征层面进行对比学习，从而学习去混淆的特征关系。其可以类比于因果之梯中所提到的反思过程，即假设、行动与结果。

(3)理解：对应于在预测层面利用 JS 散度进行对比学习，从而学习去混淆的特征关系，帮助模型理解不同输入的预测结果之间的差异。特征的无混淆并非意味着预测的无混淆。由于预测阶段存在非线性因素，在预测阶段进行对比学习非常关键。

如图 3 所示，本文通过构建因果图来分析如何减小语言偏见。其核心思路为增强视觉信息对结果的影响，并减弱文本对结果的影响。

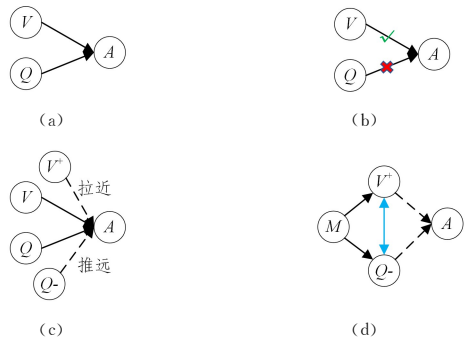


图 3 视觉问答中语言偏见问题的因果图

Fig. 3 Casual graph of language bias in VQA

图 3(a)为本文构建的视觉问答因果图模型，图 3(b)表示减缓语言偏见的核心思想，即增强视觉信息重要性并减缓文本信息的重要性。因此，如图 3(c)所示，考虑使用反事实样本来实现这一过程，通过包含核心视觉特征的事实样本来增强视觉信息重要性，通过反事实样本来减缓文本信息对模型的偏见引导。图 3(c)中，V^+ 表示与答案建立正向关系，特征与原始特征 V 拉近；Q^- 表示与答案建立反向关系，特征与原始特征 Q 拉远。

其数学表达为：图 3(a)中的总体因果效应 $TEa=Y_{v,q}(a;V=v,Q=q)$，而图 3(c)中的总体因果效应的目标为 $TEc=Y_{v,q}+Y_{v^+}-Y_{q^-}$，其中 Y_{v^+} 表示视觉样本为正，Y_{q^-} 表示文本为负所对应的情况。在因果效应相同的情况下，模型可以得到增强视觉信息和减弱文本信息的效果。

其次，由于反事实样本是由原始样本生成的，因此所生成的正负样本之间不可避免地存在混淆因子 M。在因果理论中，混淆(Confounder)是不可避免的，且混淆因子会影响特征的学习，降低模型的泛化能力。为了消除增广样本之间的混淆，如图 3(d)所示，需要拉远生成样本之间的距离。

下文将使用本文提出的反事实对比学习来实现对输入的原始样本和反事实样本进行特征和预测层面的反事实思考，从而实现反思与理解。

3.3.2 特征层面的反事实思考

为了实现特征级的反事实思考，本文使用了由 CSS 算法生成的反事实样本三联体，如 (I^+,I^-,Q^+,Q^-)。图 4 和图 5 给出了通过反事实样本示例对进行特征对比学习的方法。图 4 给出了通过建模原始样本与反事实样本三者之间的两两关系来完整实现特征层面的反事实思考。图 5 给出了在考虑 A 与 B 和 C 之间的对比损失情况下，B 可能存在 B_1 或 B_2 的情况，其在特征圆上。这会导致 B 和 C 之间的混淆，其距离

并非最佳。因此，拉远 B 和 C 可以降低模型的特征混淆。

本文通过缩小原始特征与包含关键视觉信息和关键语言信息的特征之间的距离，扩大了这些特征与非关键信息特征之间的距离。值得一提的是，需要建模三者之间的两两关系才能完整实现特征层面的反事实思考过程。

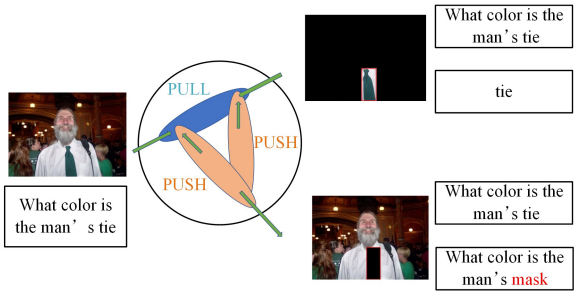
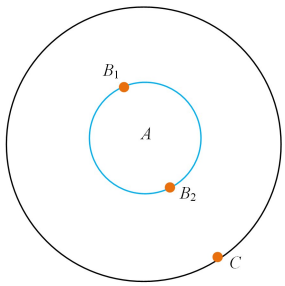


图 4 特征层次对比学习示例

Fig. 4 Some examples of features level contrastive learning



注：A 为原始样本，B 为事实样本，C 为反事实样本

图 5 样本之间的特征等高线

Fig. 5 Feature contour lines between samples

为了学习原始样本与反事实样本之间的表征差异，并训练模型找出这些样本在特征层面上的差异，本文采取了对比学习的经典范式，其表达式如下：

$$\cos(x,x^+)=\frac{x^{\mathrm{T}}\cdot x^+}{\|x\|\cdot\|x^+\|} \tag{3}$$

$$f_c(x,x^+,x^-)=E_{xx^+x^-}\left[-\log\left(\frac{e^{\cos(x,x^+)}}{e^{\cos(x,x^+)}+e^{\cos(x,x^-)}}\right)\right] \tag{4}$$

其中，$\cos(\cdot)$ 表示余弦函数，$\boldsymbol{x}$ 表示输入的向量，x^+ 表示正例，x^{T} 表示转置。

本文使用以下输入来缩短原始样本和正样本对之间的距离，增加原始样本和负样本对之间的距离，以及正样本对和负样本对之间的距离。将 3 对特征输入对比损失函数中得到反事实的对比损失结果，并累加这 3 项。

$$L_{c1}=f_c(F(V_e,Q_e),F(V_e^+,Q_e),F(V_e^-,Q_e)) \tag{5}$$

$$L_{c2}=f_c(F(V_e,Q_e),F(V_e,Q_e^+),F(V_e^-,Q_e^-)) \tag{6}$$

$$L_{c3}=f_c(F(V_e,Q_e),F(V_e^+,Q_e^+),F(V_e^-,Q_e^-)) \tag{7}$$

$$L_c=L_{c1}+L_{c2}+L_{c3} \tag{8}$$

通过上式可得到特征层面的对比学习损失函数，通过优化该损失函数，可以实现特征层面的反事实思考。

3.3.3 预测层面的反事实思考

为了实现预测层面的反思与理解，本文提出了一种针对预测的反事实思考方法。本文通过对模型输出的 Logits 进行对比学习训练来预测不同样本之间的差异。为了在预测层面训练模型的反事实思考，本文使用 Jensen-Shannon Divergence[35-37]来度量预测结果之间的分布差异。其表达式如下：

$$L_{JS(p\|q)}=\frac{1}{2}\left(KL(p\|\frac{p+q}{2})+KL(q\|\frac{p+q}{2})\right) \tag{9}$$

其中，p，q 为模型输入，KL 为 KL 散度。

本文对原始特征和正样本增强后的特征建立损失函数，并将这两者与负样本预测结果之间的距离拉远，从而实现预测层面的反事实思考。其表达式如下：

$$L_{JS_{ori}}=L_{JS(p\|p(V^-))}+L_{JS(p\|p(Q^-))}+L_{JS(p\|p(V^-,Q^-))} \tag{10}$$

$$L_{JS_{aug}}=L_{JS(p^+\|p(V^-))}+L_{JS(p^+\|p(Q^-))}+L_{JS(p^+\|p(V^-,Q^-))} \tag{11}$$

其中，p^+ 表示对应的正样本组合，L_{ori} 为针对原始样本的对比损失，L_{aug} 为针对反事实样本的对比损失。

综上所述，本文构建了完整的反事实思考流程，通过累加这两项和分类损失，得到如下所示的反事实模块的损失函数 L_{CFT}。

$$L_{CFT}=\lambda_{vqa}L_{vqa}+\lambda_cL_c-\lambda(L_{JS_{ori}}+L_{JS_{aug}}) \tag{12}$$

其中，λ 为各项的系数。通过优化该函数，可以帮助视觉问答模型实现反事实思考。

3.4 反事实思考机制算法

本文的核心方法步骤如算法 1 所示。

算法 1 反事实思考机制

输入：训练集中的图像 I、问题 Q 及答案 A，cond 是用于控制生成的反事实图片或者文本数量比例的参数，数值范围为[0,1]，以及反事实的样本对(I^+，I^-，Q^+，Q^-)

输出：反事实思考损失

Function CFT：

$pred_{ori}$，$loss_{ori}$←VQA(I,Q,a,cond)

$pred_{aug}$，F_{aug}←VQA(I,Q^+,None,cond)

$pred_{neg}$，F_{neg}←VQA(I,Q^-,a^-,cond)

L_c←f_c(F_{aug}，F_{neg})

$L_{JS_{ori}}$←L_{JS}($pred_{ori}$，$pred_{neg}$)

$L_{JS_{aug}}$←L_{JS}($pred_{aug}$，$pred_{neg}$)

L_{CFT}←L_{vqa}+L_c−$L_{JS_{ori}}$−$L_{JS_{aug}}$

return L_{CFT}

4 实验及结果分析

4.1 实验环境

本文使用标准的视觉问答评估指标[7]来评估所提模型在 VQA-CP v2 和 VQA v2 数据集上的结果。为了公平比较，对比的方法都以 UpDn 模型作为网络结构，并与每种方法在其论文中所记录的最佳性能进行比较。本实验在两个 Titan Xp GPU 上进行训练与测试。

4.1.1 数据集介绍

目前，针对视觉问答中的语言偏见问题，研究人员常用 VQA-CP v2 数据集来评估所提模型的性能，并在数据集 VQA v1 或 VQA v2 进行辅助验证。现有的研究成果大多在 VQA-CP v2 和 VQA v2 上进行测试并计算 Gap 指标来辅助验证模型的鲁棒性。

VQA v2 是 VQA 数据集的第二个版本。训练集包含 443 757 个图像问题对，验证集包含 214 354 个图像问题

对，测试集包含 447 793 个图像问题对，其数据集大小是 VQA v1 的两倍。

为了度量语言偏见，研究者们提出了 VQA-CP v2 数据集，该数据集是对 VQA v2 数据集进行样本重划分后得到的，该数据集与其原始版本是目前唯一的用于评价语言偏见的开源数据集。此外，该数据集的训练集和测试集的问题和答案分布有很大的差异，即对于同一类型的问题，训练集和测试集的答案分布差异较大。因此，该数据集非常适合用来衡量模型是否存在语言偏见。VQA-CP v2 数据集的训练集包含图片 121 000 张，问题 438 000 条，答案 440×10^6 条；测试集包含图片 98 000 张，问题 220 000 条，答案 220×10^6 条。

4.1.2 评价方法

在评价句子的正确性时，需要考虑句法的正确性以及句子的语义是否正确。为了简化问题，视觉问答的大多数据集将生成的答案限制为单词或短语，长度为 1～3 个单词。目前所采用的通用评估方法如式(13)所示，数据集对问题收集 10 个不同的真实答案。在评估时，需要将生成的句子(答案)与 10 个人工答案进行比较，从而得到准确率：

$$accuracy=\min\left(\frac{\#humans\cdot provided\cdot answers}{3},1\right) \tag{13}$$

4.2 实验结果及对比

表 1 列出了本文方法和近年来提出的其他方法在 VQA-CP v2 上的性能对比。选择的对比方法包括：UpDn[25]，RUBi[20]，LMH[21]，RMFE[38]，RandImg[39]，CSSVQA[14]，CFVQA[40]，Greedy[41]以及 CLVQA[15]。为公平起见，统一选择 UpDn 模型作为骨干网络。可以看出，与最佳方法相比，本文方法的总体精度提高了 0.19%，平均精度提高了 0.89%(各项指标平均)，尤其是 Num 精度提高了 2.6%。结果表明，本文方法优于目前 VQA-CP v2 上的对比方法。

对于单个指标，如 Num，yes/no 等，本文提出的方法并非全部优于其他方法。1)CFVQA 在数字指标(Num)方面略高于本文方法，这是因为 CFVQA 是一种基于因果推理的集成方法。与 Boosting 类似，CFVQA 使用了比本文方法更多的网络作为附加信息，因此，直接比较小指标是不合适的。2)Greedy 方法在其他指标(Others)方面略高于本文方法，这是因为 Greedy 也是一种 Boosting 方法，因此其在其他指标上略高于本文方法。

表 1 在 VQA-CP v2 上的最新精度的性能比较

Table 1 Performance comparison on VQA-CP v2

(单位：%)

Algorithm	Overall	Yes/No	Num	Others	Avg
UpDn	39.74	42.27	11.93	46.05	33.42
RUBi	47.11	68.65	20.28	43.18	44.03
LMH	52.05	69.81	44.46	45.54	53.27
RMFE	54.55	74.03	49.16	45.82	56.43
RandImg	55.37	83.89	41.60	44.20	55.56
CSSVQA	58.95	84.37	49.42	48.21	60.60
CFVQA	53.69	**91.25**	12.80	45.23	49.76
Greedy	57.32	87.04	27.75	**49.59**	54.79
CLVQA	59.18	86.99	49.89	47.16	61.34
CFTVQA(本文)	**59.37**	87.95	**52.42**	46.30	**62.23**

表 2 列出了本文方法和近年来提出的其他方法在 VQA v2 上的性能比较，该差距表示 VQA-CP v2 与 VQA v2 之间的总体精度差异。选择的对比方法包括：UpDn[25]，AReg[19]，SCR[18]，GRL[42]，RUBi[20]，LMH[21]，CSSVQA[14]，CFVQA[40]，Greedy[41]以及 CLVQA[15]。用于比较的指标是 Gap，它表示 VQA- CP v2 上的模型和 VQA v2 上的模型之间的性能差距，是一种辅助验证的指标，被广泛应用于视觉问答的鲁棒性研究中。Gap 差距越小，模型的鲁棒性越强。结果表明，该模型具有较强的鲁棒性。与 CSSVQA 方法相比，Gap 指标提高了 50%以上。

表 2 VQA v2 上的性能比较

Table 2 Performance comparison on VQA v2

(单位：%)

Algorithm	Overall	Yes/No	Num	Others	Gap
UpDn	63.48	81.18	42.14	55.66	23.74
AReg	62.75	79.84	42.35	55.16	21.58
SCR	62.30	77.40	40.90	56.50	13.83
GRL	51.92	—	—	—	9.59
RUBi	50.56	49.45	41.02	53.95	5.33
LMH	61.64	77.85	40.03	55.04	9.19
CSSVQA	59.91	73.25	39.77	55.11	0.96
CFVQA	63.65	82.63	44.01	54.38	9.96
Greedy	56.25	85.08	48.56	24.78	1.07
CLVQA	57.29	67.27	38.40	54.71	1.89
CFTVQA(本文)	59.82	74.91	38.64	53.97	**0.45**

从单独的指标上看，本文方法在 VQA v2 上的性能略低于 CSSVQA。这是因为本文提出的方法会更强烈地削弱语言先验，以减小语言偏见。因此，在具有相对平衡数据分布的 VQA v2 数据集上，本文方法对于依赖于语言偏见的样本，准确度将大幅下降，因为模型有时只能依赖语言先验来获得正确的答案。查看数据集可发现数字指标中的样本更依赖于语言偏见。因此，在更好地削弱语言先验之后，所提方法的数量指标略大于 CSSVQA，但总体 Gap 指标小于 CSSVQA。另外，图 4 很好地说明了这种情况。

4.3 消融实验结果

通过消融实验研究了影响所提方法性能的因素。如表 3 所列，消融实验分别在特征层面和预测层面进行。对预测层面的反事实思考进行消融可以发现，通过添加预测的反事实思考模块，模型的总精度指标得到了提升。相比特征层面的反事实思考，模型在 Yes/No 和 Other 的精度上都得到了提高，验证了模型的有效性。经过消融实验可以发现，所提出的两步反事实思考对于总的模型精度都有提升，且模型在不同数据分布下的 Gap 有所减小。

表 3 特征层面和预测层面的消融实验

Table 3 Ablation experiments at feature level and prediction level

(单位：%)

Algorithm	Overall	Yes/No	Num	Others
UpDn	39.74	42.27	11.93	46.05
+CSS	58.95	84.37	49.42	48.21
+Features CFT	59.22	87.85	53.41	45.81
+Prediction CFT	59.37	87.95	52.42	46.30

表 4 列出了损失函数中各种超参数对模型性能的影响。可以发现，Lambda 为 0.2 且 Lambda-c 为 1 时，模型精度最优。

表 4 VQA-CP v2 消融实验

Table 4 Ablation experiments at VQA-CP v2

Algorithm	Lambda	Lambda-c	Overall/%
UpDn	—	—	39.74
+LMH	—	—	52.05
+CFT	0.1	1	58.63
+CFT	0.2	1	**59.37**
+CFT	0.3	1	58.71
+CFT	0.1	2	58.84
+CFT	0.2	2	59.32
+CFT	0.3	2	58.85

为了更好地展示结果，本文从定性分析的角度对模型的代表性结果进行了可视化分析，并与其他方法进行了比较。图 6中的绿色单词表示该单词对最终预测的贡献更大。图7给出了在降维后在特征空间中 3 类样本的特征分布，可以看出，原始样本(蓝色)和事实样本(红色)的距离更近，它们与反事实样本(黑色)的距离更远，数字表示分类结果所对应的类别。

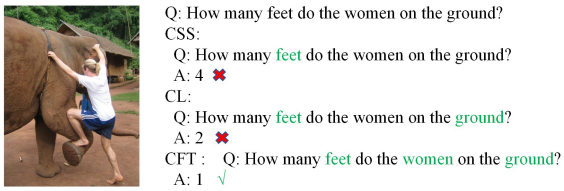


图 6 模型结果的可视化分析(电子版为彩图)

Fig. 6 Visualization analysis of model results

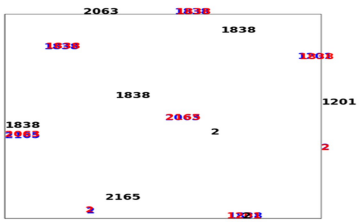


图 7 在训练过程中，随机选取的部分样本在降维后在特征空间中 3 类样本的特征分布(电子版为彩图)

Fig. 7 Feature distribution of randomly selected samples in feature space after dimensionality reduction

结束语 为了实现鲁棒的视觉问答，本文提出了一个反事实对比学习范式。将多层次特征比较和预测比较相结合，提出了一种反事实思考模型，使模型可以理解原始样本、事实样本和反事实样本之间的关系。结果表明，该方法提高了现有 VQA 模型的推理分析能力，减少了语言偏见对模型的误导性。

参 考 文 献

[1] NIU Y L, ZHANG H W. Survey on Visual Question Answering and Dialogue [J]. Computer Science, 2021, 48(3): 87-96.

[2] FU P C, YANG G, LIU X M, et al. Visual Question Answering Network Method Based on Spatial Relationship and Frequency [J]. Computer Engineering, 2022, 48(9): 96-104.

[3] ZOU P R, XIAO F, ZHANG W J, et al. Multi-Module Co-Attention Model for Visual Question Answering[J]. Computer Engineering, 2022, 48(2): 250-260.

[4] WU A M, JIANG P, HAN Y H. Survey of Cross-media Question Answering and Reasoning Based on Vision and Language [J]. Computer Science, 2021, 48(3): 71-78.

[5] XU S, ZHU Y X. Study on Question Processing Algorithms in Visual Question Answering [J]. Computer Science, 2020, 47(11): 226-230.

[6] WANG S H, YAN X, HUANG Q M. Overview of Research on Cross-media Analysis and Reasoning Technology [J]. Computer Science, 2021, 48(3): 79-86.

[7] YUAN D. Language bias in Visual Question Answering: A Survey and Taxonomy [J]. arXiv: 2111.08531, 2021.

[8] 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.

[9] ANTOL S, AGRAWAL A, LU J, et al. Vqa: Visual question answering[J]. International Journal of Computer Vision, 2017, 123(1): 4-31.

[10] AGRAWAL A, BATRA D, PARIKH D. Analyzing the behavior of visual question answering models[C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, USA: ACL, 2016: 1955-1960.

[11] PENG Z, GOYAL Y, SUMMERS-STAY D, et al. Yin and yang: Balancing and answering binary visual question[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 5014-5022.

[12] JUSTIN J, HARIHARAN B, MAATEN L, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning[C] // Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 2901-2910.

[13] YASH G, TEJAS K, SUMMERS-STAY D, et al, Making the v in vqa matter: Elevating the role of image understanding in visual question answering[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 6904-6913.

[14] CHEN L, YAN X, XIAO J, et al. Counterfactual samples synthesizing for robust visual question answering[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10800-10809.

[15] LIANG Z, JIANG W, HU H, et al. Learning to contrast the counterfactual samples for robust visual question answering [C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2020: 3285-3292.

[16] AGRAWAL A, BATRA D, PARIKH D, et al. Don't just assume; look and answer: Overcoming priors for visual question answering[C] // Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4971-4980.

[17] SELVARAJU R R, LEE S, SHEN Y, et al. Taking a hint: Leveraging explanations to make vision and language models more grounded[C] // Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korean, IEEE, 2019: 2591-2600.

[18] WU J L MOONEY R J. Self-critical reasoning for robust visual

question answering[J]. arXiv:1905.09998,2019.
[19] RAMAKRISHNAN S,AGRAWAL A,LEE S. Overcoming language priors in visual question answering with adversarial regularization[J]. arXiv:1810.03649,2018.
[20] REMI C,CORENTIN D. Rubi:Reducing unimodal biases in visual question answering[C]// Advances in Neural Information Processing Systems. 2019.
[21] CLARK C,YATSKAR M,ZETTLEMOYER L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hongkong, China: IEEE,2019:4069-4082.
[22] ABBASNEJAD E,TENEY D,PARVANEH A,et al. Counterfactual vision and language learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA:IEEE,2020:10044-10054.
[23] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[J]. arXiv:1406.2661v1,2014.
[24] TENEY D,ABBASNEDJAD E,VAN DEN HENGEL A. Learning what makes a difference from counterfactual examples and gradient supervision[C]// Proceedings of European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 580-599.
[25] 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 /CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE,2018:6077-6086.
[26] REN S,HE K,GIRSHICK R,et al. Faster r-cnn:Towards real-time object detection with region proposal networks[C]// NIPS. 2016.
[27] GREFF K,SRIVASTAVA R K,KOUTNÍK J,et al. LSTM:A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems,2016,28(10):2222-2232.
[28] PEARL J. Causality:models,reasoning and inference[M]. Cambridge:Cambridge University Press,2000.
[29] PEARL J. Direct and indirect effects[C]// Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc.,2001:411-420.
[30] PEARL J,MACKENZIE D. The book of why:the new science of cause and effect,Basic Books[J]. Science,2018,361:47-54.
[31] HARARI,Y. A brief history of humankind[M]. Beijing:CITIC Publishing House,2014.
[32] WEISBERG D S,GOPNIK A. Pretense,counterfactuals,and Bayesian causal models:Why what is not real really matters[J]. Cognitive Science,2013,37(7):1368-1381.
[33] ROESE N J,EPSTUDE K. The functional theory of counterfactual thinking:New evidence,new challenges,new insights[J]. Advances in Experimental Social Psychology,2017,56:1-79.
[34] BRIGARD F D,ADDIS D R,FORD J H,et al. Remembering what could have happened:Neural correlates of episodic counterfactual thinking[J]. Neuropsychologia,2013,51(12):2401-2414.
[35] OORD A,LI Y,VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv:1807.03748,2019.
[36] BENT F,FLEMMING T. Jensen-shannon divergence and hilbert space embedding[C]// Proceedings of International Symposium on Information Theory. Chicago,USA,IEEE,2004.
[37] LIN J. Divergence measures based on the Shannon entropy[J]. IEEE Transactions on Information theory,1991,37(1):145-151.
[38] GAT I,SCHWARTZ I,SCHWING A,et al. Removing bias in multi-modal classifiers:Regularization by maximizing functional entropies[J]. Advances in Neural Information Processing Systems,2020,33:3197-3208.
[39] TENEY D,KAFLE K,SHRESTHA R,et al. On the value of out-of-distribution testing:An example of goodhart's law[J]. Advances in Neural Information Processing Systems,2020,33:407-417.
[40] NIU Y,TANG K,ZHANG H,et al. Counterfactual vqa:A cause-effect look at language bias[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online,IEEE,2021:12700-12710.
[41] HAN X,WANG S,SU C,et al. Greedy gradient ensemble for robust visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. Online, IEEE, 2021:1584-1593.

YUAN De-sen, born in 1997, postgraduate. His main research interests include multi-modal learning and deep learning.

WU Qing-bo, born in 1985, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include image and video coding, image and video quality assessment and visual perception model.

（责任编辑：何杨）