# 谷歌BARD的视觉理解能力如何?对开放挑战的实证研究

秦浩桐<sup>†1</sup>, 季葛鹏<sup>†2</sup>, Salman Khan<sup>3</sup>, 范登平<sup>≥1</sup>, Fahad Shahbaz Khan<sup>3</sup> and Luc Van Gool<sup>2</sup>

<sup>1</sup>计算机视觉实验室,苏黎世联邦理工学院,苏黎世,瑞士. <sup>2</sup>工程计算与控制学院,澳大利亚国立大学,堪培拉,澳大利亚. <sup>3</sup>穆罕默德·本·扎耶德人工智能大学,阿布扎比,阿拉伯联合酋长国.

#### Abstract

作为OpenAI公司ChatGPT模型的竞品,谷歌公司提出的BARD模型已经在会话型人工智能领域 取得了显著进展。特别值得注意的是,BARD模型的最新版本在对话过程中具备了处理文本 提示和视觉输入的能力。鉴于BARD模型在文本输入处理方面取得的令人瞩目成就,本文聚 焦于探索其在理解并解析由文本问题引导的视觉数据(图像)方面的潜力。这一探索有望揭 示BARD模型以及其他即将涌现的多模态生成式模型背后的新见解与挑战,特别是在解决那些 需要准确的视觉和语言理解能力的复杂问题时。具体而言,本研究针对15种不同的任务场景展 开研究,涵盖了通用、伪装、医疗、水下和遥感数据等领域,用于全面评估BARD模型的表现。 实验结果表明,在这些视觉场景中,BARD模型仍然面临一定的困难,表明其在视觉理解能力 方面具有提升空间。这项实证研究有助于推动未来相关模型的发展,增强模型在理解和解析细 粒度视觉数据方面的能力。项目已发布于https://github.com/htqin/GoogleBard-VisUnderstand。

## 1 引言

BARD<sup>1</sup>(谷歌AI聊天机器人)是一款基 于LaMDA模型[2]和后续的PaLM模型[3]的生成 式人工智能系统。于2023年3月初步推出,并 于5月在全球范围内推广开来。该系统能够接受 文本提示,并根据这些提示执行多种基于文本的 任务,如提供答案、摘要和创作各类文本内容。 截至2023年7月13日,谷歌的BARD宣布了一项重 大更新<sup>2</sup>,该更新允许用户以图像作为输入,并

 $^{1} \rm https://bard.google.com$ 

 $^{2} \rm https://blog.google/products/bard/$ 

google-bard-new-features-update-july-2023/

结合文本提示使用。据报道,BARD能够分析视 觉内容,提供相应的描述(如图像标题),或 者运用视觉信息回答问题。值得关注的是,尽 管其他模型如GPT-4 [5]声称能够接受并理解图 像输入作为提示,但这些模型并未公开使用。 因此,BARD模型为计算机视觉领域提供了首个 机会,去评估它的可靠性和鲁棒性,以理解现 有的优势和局限性。本研究工作的目标是基于 计算机视觉领域中长期存在的一些问题,分析 了BARD模型在应对这类问题上所具备的能力。

本研究针对计算机视觉问题,确定了一系列具有趣味性的任务场景用于BARD模型的定性评价。考虑到当前BARD模型尚未提供开放的API访问权限,因此无法进行大规模基准的

本文是论文[4]的中文翻译稿,译者:秦浩桐,并由季葛鹏、范登平进行校稿)



Fig. 1 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从Microsoft COCO数据集[1]中获取的图像回答用户的问题。在接下来的图中,除非另有说明,否则括号中提供了预期的标准答案。有关详细信息,请参阅第2.1节,第2.2节,第2.3节,第2.4节,第2.5节和第2.11节。



**Fig. 2** 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从Tiny-ImageNet-C数据集 [1]中获取的图像回答 用户的问题。有关详细信息,请参阅 第2.6节。

定量测试。这些任务场景和提示不仅适用于评估BARD模型的视觉理解能力,也可以为未来的 大规模多模态模型(例如GPT-4)的评估提供一 定的参考依据。本研究特别选用BARD模型的原 因是,其在所有开/闭源多模态对话模型中表现 突出,包括于2023年7月18日推出的Bing-Chat模 型[6]。相关数据请参考LLaVA-Bench[7]。

## 2 实证实验

为了评估BARD在给定文本提示的条件下的视觉 感知和语境理解等能力,我们设计了一系列视 觉-语言任务场景。随后,我们深入研究了从这些 实证研究中提取出的若干例子,涵盖了总共15个 视觉问答(VQA)场景,涉及对象检测和定位、 分析对象属性、计数、可供性以及自然图像中的 细粒度识别等任务。我们还在一些具有挑战性的 情况下进行了实验,例如识别伪装对象,以及医 疗、水下和遥感图像等多样的领域。以下我们对 这些场景进行解释。

#### 2.1 场景#1 - 对象属性

理解图像中对象的属性和特征是视觉问答中的一项基本任务。例如,在图1的第四个问题中,当给出查询"镜子的形状是什么?"时,BARD无法 理解镜子这一对象的与形状相关的属性,并且在 想象中还出现了镜子中的反射。这表明BARD在 识别需要深入理解每个对象及其属性的属性方面 存在挑战。

## 2.2 场景#2 - 对象存在

这评估了BARD根据提供的文本描述识别特定 对象的能力。正如图 1中的第一个问题所证明 的,BARD无法正确回答问题"图像中是否有人出 现?"并提供了错误的回答:"图像中没有人"。这 表明BARD对视觉内容的基本理解仍然有限。我 们进一步注意到,BARD目前适用于不包含任何 人类的图像,并会删除包含人脸或人物的任何视 觉输入。

### 2.3 场景#3 - 对象位置

该任务场景考察了BARD在定位和理解图像内 对象方面的能力。例如,参考图 1中的第二 个问题,询问:"这张图像中的门在哪里?"然 而,BARD无法在提供的图像中识别出门,回 答:"你发给我的图像中没有门。"因此,这表 明BARD在视觉背景的定位能力方面有待进一步 提升。

## 2.4 场景#4 - 关系推理

这个场景评估了BARD在理解图像内复杂的对象 间动态关系方面的能力,审视其对空间和语义关 系的理解。如图 1中的最后一个问题所示,我们 向BARD提问:"狗是否靠近电视?"然而,它未 能提供正确的反馈,回答:"不,你发给我的图像 中狗不在电视旁边。"因此,这表明BARD在推理 关系方面还有提升的空间。

### 2.5 场景#5 - 适用性

适用性测试用于验证BARD在对象可能提供或 支持的潜在用途或行动方面的理解能力。这 涉及模型对功能的理解,超越了简单的对象识 别。图 1中的第五个问题"你可以在哪里躺下睡 觉?"是一个示例研究。有趣的是,BARD可以提 供两个合理的答案(即地毯和沙发),但对于其 他选项,如一些缺失的项目,如"卧室里的床。" 和"海滩。公园。森林。",却无法始终如一地提 供正确的回答。输出中的错觉意味着BARD仍然 需要更好地根据文本指导严格捕捉视觉语义,并 将这些语义更有效地与场景中的已识别对象关联 起来。

## 2.6 场景#6 - 对抗样本

对抗样本是指经过精心设计的输入,旨在混淆或 误导AI模型。由于像素值的微小变化,这些输入 对人类来说仍然是可感知的,但对AI模型来说却 变得难以解释。在黑盒设置中探索BARD对这些 样本的响应至关重要,因为它提供了模型的鲁棒 性和稳定性的衡量标准。如图 2所示,我们提供 了四组图像-文本对,其中图像显示了一些鱼。所 有来自BARD的输出都表明它无法理解对抗性样 本。

### 2.7 场景#7 - 雨天条件

不同的天气条件在评估模型处理意外情况和领 域转变的能力方面起着关键作用,这是现实世界 应用的重要前提。因此,我们选择了四组样本, 在自然环境中测试BARD在不同天气条件下的表 现,如图 3所示的雨天。结果表明,BARD在图 像具有雨天条件时表现不佳。



Fig. 3 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从MPID数据集 [8]获取的图像回答用户的问题。有关详细信息,请参阅 第2.7节。



**Fig. 4** 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从图像情感数据集 [9]获取的图像回答用户的问题。 有关详细信息,请参阅 第2.8节。

#### 2.8 场景#8 - 情感理解

这个场景评估了BARD理解图像中场景情感或情 绪背景的能力,从而考察其对视觉数据更抽象、 不可触摸的方面的理解。如图 4所示,当我们 向BARD提问: "图像中表达的情感是积极的还是 消极的?"时,它回答了错误的回答: "图像中表 达的情感是消极的。"

#### 2.9 场景#9 - 细粒度识别

这个任务涉及识别给定对象类别中的特定子类 别,比一般的对象识别更复杂,因为它涉及到类 内变化的增加、类间微小差异的增多,以及对专 业领域知识的需求。我们在 图 5中展示了四种类 型的飞机,并向BARD提问以下问题:"这架飞机 的家族、制造商和型号是什么?"例如,在子图 (a)中,BARD错误地将该飞机识别为波音747-400,但它实际上是由空中客车制造,属于A310 家族。

## 2.10 场景#10 - 识别伪装对象

这个具有挑战性的任务侧重于识别有意与周围环 境融合的隐蔽对象,使它们难以被识别。我们从 常用的COD10K数据集<sup>3</sup> [10]中选择了四个样本。 例如,如图 6 (a)所示,当我们要求BARD识别 一条鱼时,它断言:"我可以自信地说,这张图像 中没有动物。"这一观察表明,BARD在区分伪装 对象与周围环境方面存在挑战,特别是当前景和 背景具有类似的颜色时。这表明BARD解析伪装 模式和类似纹理的能力可以进一步增强。

### 2.11 场景#11 - 对象计数

这个场景测试了模型识别和计数给定文本描述 中特定对象的能力。它要求BARD识别对象并 对其进行数量化,以探究其在视觉背景下的数 值理解能力。如图 1所示,我们从数据集中选 择了样本,问题是"地毯图案中有多少个正方

 $<sup>^{3}</sup>$ https://github.com/GewelsJI/SINet-V2





(c) What is the family, manufacturer, and varian of this aircraft? (family

"X The aircraft in the image is a Bacing 247-400. It is a wide-body, four-engine jet airliner manufactured by Boeing Commercial Airplanes. The 747-400 is the most recent version of the Boeing 747, and it was introduced in 1989. It is the longest and heavies passenger aircraft in the world. The 747-400 is powered by four General Electric GE-CE-640CBH engines, and it has a maximum range of 7.000 nautical miles. It can enzy up to 140 passengers in a two-class configuration. The 747-400 his been used by a number of airlines, including British Airways. Lufthansa, and Quatas. It is still in production, and it is expected to remain in service for many years to come. The specific aircraft in the image is a Boeing 747-400 operated by Dinar Lineas Aérasa. It is registered as YI-300E. The aircraft was delivered to Dinar in 2000, and it is currently based at Simón Bolivar International Airport in Caracas, Venezuela."

<sup>48</sup> The aircraft in the image is a Boeing 707-328C, s<sup>-1</sup> which is a corpo variant Beengy 707. Its owde-body, four-complet eithiner manufactured by Boeing mmercial Atripanes. The 707-328C was introduced in 1971, and it is the most monor variant of the 707-300 (milly). He 707-328C is powered by four Presit & ettics. The 707-328C was introduced in 1971, and it is the most monor variant of the 707-300 (milly). He 707-328C is powered by four Presit & ettics. the Boe Comme 1-commercian araptames. In: (1):-2:30: was introduced in 1971, and it is the most common variant of the 107-300 fumily. The 107-32:20: Exportend by four Pratt & Whitney, JTBD-7 engines, and it has a maximum range of 3:900 nautical miles. It can carry up to 11:2:00 pounds of cargo. The specific arcrift in the image is a hoeing 107-32:00 cperated by Air France Cargo. It is registered as F-BLCC. The aircraft was delivered to Air France in 1978, and it is currently hoed at Charles de Gaulle Airpor in Paris, France. Here is a table summarizing the information about the aircraft in the Image: Family, Manuforume Variant Doubstantine for 2019. aft in the (in rars, France, Freie Sa taube summaring me and second me attention in the attention in the attention in image: Family Manufacturer Varianti Registration Boeing 707 Boeing 707-328C F-BLCC Wide-body Cargo 112,600 pounds 4 Pratt & Whitney JT3D-7 engines 3,900 manifact miles 1978 Air France Cargo Charles de Gaulle Airport, Paris, France France to Sheate Lhoos the anonese none muction 1<sup>(7)</sup> Boeing 707, manufacturer Boeing, variant 707-320) Export to Sheets I hope this answers your question!"



(d) What is the fo

BAE 146, man British Aerosn

of this aircraft? (family

BAE 146-200)

ufacture

variant

\*X The aircraft in the image is an Airbus A310-300, which is a wide-body, long-range commercial airliner manifactured by Airbus. It is the stretched version of the A310-200, and it was introduced in 1985. It is the largest variant of the A310 family. The A310-300 is powered by two General Electric CF6-80C248 engines, and it has a maximum range of 7.400 nautical miles. It can carry up to 220 passengers in a two-class configuration. The specific aircraft in the image is an Airbus A310-300 operated by Mahan Air. It is registered as EP-MIHL. The aircraft was delivered to Mahan in 2002, and it is currently based at Mebrabad International Airport in Tehran, Iran. Here is a table summarizing the information about the aircraft in the image: [...] Export to Sheet I hope this answers your auseiton!" Export to Sheets I hope this answers your question!

/ The aircraft in the image is a British. ce 146, which is a r nufactured by British Aerospace. The 146 is a twin-turbofan-short-range airliner manufactured by British Aerospace, The 146 is a twin-turbofan, short-range durrard that was introduced in 1983. X The specific aircraft that we introduced in 1983. X The specific aircraft in the image is a British derospace 146-300, which is the most common variant of the 146 family. It is powered by two Rolls-Royce BR 183 Mk 515 nurbofan engines, and it has a maximum range of 1,500 natural on lines. It can carry up to 117 Dassengers in a single-class configuration. The aircraft in the image is operated by Brussels Airlines. It is constrained in the image is operated by Brussels Airlines. It is constrained in the image is operated by Brussels Airlines. It is constrained to a solution of the 200 NULL of the solution of the soluti registered as OO-DWH. The aircraft was delivered to Brussels Airlines in 1994, and in summarizing the information about the aircraft in the image: [...] Export to Sheets I hope this answers your question!"

Fig. 5 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从FGVC数据集 [11]获取的图像回答用户的问题。 有关详细信息,请参阅 第2.9节。



Fig. 6 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从COD10K数据集 [10]获取的图像回答用户的问 题。有关详细信息,请参阅 第2.10节。



Fig. 7 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从IOCfish5K数据集 [12]获取的图像回答用户的问 题。有关详细信息,请参阅 第2.11节。

形?"BARD给出了错误的计数,因为它错误地 将地毯的长度和宽度间隔误认为是一个14x14的 网格。有趣的是,人类可能会通过将多个子正 方形组装成一个较大的正方形来巧妙地解决这



**Fig. 8** 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从CDS2K数据集 [13]获取的图像回答用户的问题。 有关详细信息,请参阅 第2.12节。



**Fig. 9** 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从TextVQA数据集 [14]获取的图像回答用户的问题。有关详细信息,请参阅 第2.13节。

个问题,而不是直接数最小的单位,这似乎超出了BARD的能力。

此外,我们还对探索BARD在更具挑战性的 任务上的表现很感兴趣,即计算伪装对象的数 量。我们随机选择了四张图像,如图 7所示,来 自IOCfish5K<sup>4</sup> [12]。该数据集包括充满难以辨认 的海洋生物的大规模水下图像,由于能见度有限 和主动模仿,这些生物很难计数。从我们的实证 研究观察到,BARD在描述场景方面表现出色, 例如在子图(a)中:"图像是一个珊瑚礁的广角 镜头,周围有很多鱼在游动。"然而,在理解挑战 性场景中的高层次内容方面,BARD似乎不够熟 练,回答"很难说图像中有多少条鱼。"

## 2.12 场景#12 - 发现工业缺陷

质量检验在制造业中扮演着至关重要的角色, 保障产品质量和维持高效运营。我们的目标是调 查BARD在识别工业材料中伪装的缺陷、异常或 不规则性方面的能力。为此,我们从一个伪装缺 陷分割数据集CDS2K<sup>5</sup> [13]中随机选择了一些有

<sup>&</sup>lt;sup>4</sup>https://github.com/GuoleiSun/

Indiscernible-Object-Counting

<sup>&</sup>lt;sup>5</sup>https://github.com/DengPingFan/CSU

缺陷的样本。如图 8所示,这些样本包括: (a) 磁砖上的气孔, (b) 地毯上的油渍, (c) 金属 螺母上的凹痕,以及 (d) 木质材料上的一对孔 洞。在与BARD进行交互时,我们提供了问题提 示: "图像中是否有任何有缺陷的区域?"生成的 回答出现在对话界面上。我们观察到BARD在识 别这些在挑战性场景中不易察觉的缺陷方面遇到 困难,因此向用户提供了错误的回答。

#### 2.13 场景#13 – 识别光学字符

BARD能否识别和理解图像中包含的"文本",例 如扫描文档中的文本?为了回答这个问题,我们 使用了一个光学字符识别数据集TextVQA<sup>6</sup> [14], 以基于图像中的文本来评估BARD的视觉推理能 力。如图 9 (d)所示,BARD在各种文本识别场 景中遇到了困难:在表面上很明显的问题"这个地 点的名称是什么?"下,它错误地回答了"我只是 一个语言模型,所以无法帮助您。",这显示出模 型难以理解自然图像中的文本。

## 2.14 场景#14 - 分析医疗数据

与自然场景不同, 医疗数据包含复杂的与健康 相关的信息, 需要临床、解剖和病理专业知识 来进行正确解释。因此, 一个有趣的问题是调 查BARD在医疗影像数据集中的能力程度。为 了评估BARD的能力, 我们从结肠镜检查数据 集SUN-SEG<sup>7</sup> [15]中挑选出了四幅息肉(阳性) 图像。然而, 如 图 10所示, 前三幅图像没有输 出有意义的内容, 而最后一幅图像中的息肉识 别失败。我们在其他医学图像模态, 如X射线 片、MRI、CT扫描和皮肤病变图像上也遇到了类 似的输出。

## 2.15 场景#15 - 解释遥感数据

为了与BARD进行交互,我们使用了来 自RSVQA-LR<sup>8</sup> [16]的各种图像-文本对,这是一 个用于遥感视觉问答任务的精心构建的数据 集。其目标是通过以自然语言表达的直观问题 进行交流,简化广大受众对地球观测数据中信 息的访问。例如,如图 11 (a)所示,提出了问 题: "图像中有多少座商业建筑物。", 这与正确 答案82显著不同。我们的发现表明,BARD在整 体上理解视觉场景的趋势,但在识别细粒度的视 觉模式方面面临挑战,特别是在确定对象的精确 数量,如这种情况下的商业建筑物。

## 3 结语

谷歌的BARD在会话型人工智能领域的出现引 起了极大的关注,因为它取得了显著的成功。 在此势头的基础上,我们的研究旨在全面评 估BARD在各种任务场景下的表现,包括常规、 伪装、医疗、水下和遥感图像。我们的调查显 示,虽然BARD在许多领域表现出色,但在某些 基于视觉的情景中仍面临挑战。这一发现凸显 了BARD在各种应用中的巨大潜力,并强调了在 视觉相关任务中增长和改进的充足空间。我们研 究的实证见解预计对未来模型的发展具有价值, 特别是在弥合视觉性能差距方面。通过解决视觉 情景中观察到的局限性,我们预计后续模型将具 备更强的视觉理解能力,最终推动会话型人工智 能的发展达到新的高度。

<sup>8</sup>https://rsvqa.sylvainlobry.com

 $<sup>^{6} \</sup>rm https://textvqa.org$ 

<sup>&</sup>lt;sup>7</sup>https://github.com/GewelsJI/VPS



**Fig. 10** 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从SUN-SEG数据集 [15]获取的图像回答用户的问题。有关详细信息,请参阅 第2.14节。



**Fig. 11** 使用谷歌的BARD进行多模态交互会话的若干示例,其中AI系统根据从RAVQA-LR数据集 [16]获取的图像回答用户的问题。有关详细信息,请参阅 第2.15节。

## References

- D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," arXiv preprint arXiv:1903.12261, 2019.
- [2] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.
- [3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.
- [4] H. Qin, G.-P. Ji, S. Khan, D.-P. Fan, F. S. Khan, and L. Van Gool, "How good is google bard's visual understanding? an empirical study on open challenges," *MIR*, 2023.
- [5] OpenAI, "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

- [6] Microsoft, "Bing chat enterprise announced, [13] multimodal visual search rolling out to bing chat," 2023, available online at: https://blogs.bing.com/search/july-2023/Bing-Chat-Enterprise-announced, -multimodal-Visual-Search-rolling-out-to-Bing-Chat, last accessed on 27.07.2023.
- [7] "Llava-bench: In the wild," 2023, available online at: https://github.com/haotian-liu/ LLaVA/blob/main/docs/LLaVA\_Bench.md, last accessed on 27.07.2023.
- [8] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," in *CVPR*, June 2019.
- [9] S. Z. Hassan, K. Ahmad, S. Hicks, P. Halvorsen, A. Al-Fuqaha, N. Conci, and ichael Riegler, "Visual sentiment analysis from disaster images in social media," 2020.
- [10] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and
   L. Shao, "Concealed object detection," *IEEE TPAMI*, vol. 44, no. 10, pp. 6024–6042, 2022.
- [11] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," arXiv preprint arXiv:1306.5151, 2013.
- [12] G. Sun, Z. An, Y. Liu, C. Liu, C. Sakaridis, D.-P. Fan, and L. Van Gool, "Indiscernible object counting in underwater scenes," in *CVPR*, 2023, pp. 13791–13801.

- [13] D.-P. Fan, G.-P. Ji, P. Xu, M.-M. Cheng, C. Sakaridis, and L. Van Gool, "Advances in deep concealed scene understanding," *Visual Intelligence*, 2023.
  - [4] A. Singh, V. Natarjan, M. Shah, Y. Jiang,
    X. Chen, D. Parikh, and M. Rohrbach,
    "Towards vqa models that can read," in *CVPR*, 2019, pp. 8317–8326.
- [15] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," *Machine Intelligence Research*, vol. 19, no. 6, pp. 531–549, 2022.
- [16] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE TGRS*, vol. 58, no. 12, pp. 8555–8566, 2020.