arXiv:2301.03238v1 [cs.CL] 9 Jan 2023

MAQA：多模态否定问答基准

Judith Yue Li Google Research
judithyueli@google.com

Aren Jansen Google Research
arenjansen@google.com

Qingqing Huang Google Research
qqhuang@google.com

Joonseok Lee Google Research 和首尔国立大学
joonseok@google.com

Ravi Ganti Google Research
gmravi@google.com

Dima Kuzmin Google Research
gmravi@google.com

摘要

多模态学习能够受益于预训练大型语言模型（大语言模型）的表征能力。然而，最先进的基于 Transformer 的大语言模型常常忽略自然语言中的否定，而且目前还没有基准能够定量评估多模态 Transformer 是否也存在这一弱点。本研究提出一种新的多模态问答基准，该基准改编自 AudioSet 中带标签的音乐视频（Gemmeke 等，2017），旨在系统评估多模态 Transformer 能否进行复杂推理，将新概念识别为对先前已学习概念的否定。我们发现，采用标准微调方法时，无论模型规模大小，多模态 Transformer 仍无法正确理解否定。然而，我们的实验表明，通过在原始训练任务分布中加入否定式问答示例，可以使模型可靠地进行否定推理。为此，我们介绍了一种新颖的数据生成流程，利用 540B 参数的 PaLM 模型自动生成否定式问答示例，并将这些示例组合为易于获取的视频标签。与基于模板的任务增强方法相比，生成的示例包含更自然的语言模式，且性能提升显著。

1 引言

大语言模型（LLM）难以理解自然语言中的否定。预训练的大语言模型在完形填空问题中往往会忽略否定，并对被否定（“Birds cannot [MASK]”）和未被否定（“Birds can [MASK]”）的查询给出相同的预测（Kassner 和 Schütze，2019年；Hosseini 等人，2021年）。Hossain 等人（2022）分析了最先进的大语言模型的训练语料库，发现其中很少出现否定，导致模型在推理时对否定的处理效果不佳。

最先进的多模态学习利用预训练的大语言模型来融合不同模态（Jia 等人，2021年；Radford 等人，2021年；Oncescu 等人，2021年；Kilgour 等人，2022年；Nagrani 等人，2022年）。面向多模态应用进行微调的大语言模型是否也会继承否定问题？Huang 等人（2022年）表明，在基于文本查询的音频检索任务中，当文本查询包含否定线索（例如“无人声”）时，零样本性能会下降。Yu 等人（2022年）表明，即使文本提示暗示某个物体不存在，文本到图像模型仍会生成文本提示中提到的物体，尽管

NeurIPS 2022年赋能机器学习研究的合成数据研讨会。

文本提示暗示该物体不存在。然而，目前还没有基准能够定量评估这类多模态场景对文本中否定模式的处理效果。

在本研究中，我们创建了 MAQA，这是一个二元音乐音频问答基准，用于评估多模态 Transformer 对音乐相关问题中否定的理解能力。该基准由 AudioSet 音乐相关部分中带标签的视频构建而成（Gemmeke 等，2017）。原始基准包含 5000 小时的音频，标注了 527 个音频事件类别，但仅有少量标签涉及否定；相比之下，我们提出的 MAQA 基准包含大量否定问句，这些问句由原始音频标签以程序化方式生成。我们的目标是评估能否通过组合泛化对多模态 Transformer 进行微调，使其理解新概念，例如将“无人声”理解为对先前学到的概念（如“人声”）的否定。

本文的主要贡献如下：(1) 一项组合泛化实验，表明标准微调会阻碍我们的基线模型——基于多语言 T5（MT5）（Raffel 等，2019；Xue 等人，2021）改造的多模态 Transformer——泛化到对已学习概念进行否定的新概念。(2) 一种基于 PaLM 的数据生成方法，可根据易于获取的视频标签自动生成否定式问答示例。(3) 两种任务增强策略，显著提升了模型在 MAQA 中包含文本否定的部分数据上的表现。

本文其余部分安排如下。第2节介绍与否定、组合泛化和多模态学习相关的背景及研究工作。第3节概述 MAQA 数据集及其统计信息。第4节详细说明我们如何通过数据生成来创建该基准。第5节和第6节介绍模型与实验结果。最后，本文将讨论研究的局限性、结果的影响以及未来工作。

2 相关工作

否定。尽管大语言模型在自然语言理解、阅读理解和零样本文本生成等许多自然语言处理任务中取得了进步，但否定仍然是预训练大语言模型面临的一项挑战（Kassner 和 Schütze，2019年；Hosseini 等人，2021年）。自然语言处理文献中曾采用数据增强来应对否定。例如，Hossain 等人（2020年）提出，通过在动词或形容词前添加否定并反转标签来修改自然语言理解语料库；Hosseini 等人（2021年）则提出，在跨度破坏预训练任务中使用非似然损失。元学习文献也探讨了否定（Murty 等人，2021年），并将其视为一种推理类别，认为需要额外的少样本分类任务来扩充原始任务分布。

组合泛化。组合泛化是指将新概念理解为由先前学习的概念或原子组合而成的能力。否定可以被视为一种组合形式。在语义解析领域，已有多个基准被提出，用于评估组合泛化（Lake and Baroni, 2018；Keysers 等人，2020；Kim and Linzen, 2020），这些基准推动了相关技术和架构的发展，使大语言模型更善于解决组合任务（Furrer et al., 2020；Ontanon et al., 2022；Csordás et al., 2021；Qiu et al., 2022）。多项多模态基准测试表明，在视觉推理任务（Johnson et al., 2017）、视觉接地指令跟随任务（Ruis et al., 2020）、文本到图像匹配（Zhang et al., 2021）等任务中，视觉接地大语言模型往往难以实现组合泛化。我们的研究旨在评估音频接地大语言模型在涉及否定的组合任务中的表现。

多模态问答。多模态问答基准测试用于探测多模态模型，评估其在不同模态上的感知与推理能力。视觉问答基准测试（Zhang 等，2015年；Agrawal 等，2015年）通常由三元组构成：（图像、关于该图像的自然语言问题、答案），任务是根据图像中的视觉线索回答问题。在音频感知领域，音频问答基准测试（Fayek 和 Johnson，2020）的应用不如音频分类基准测试（Gemmeke 等，2017年）常见。在音乐领域，大多数基准测试都是音乐信息检索任务（Law 和 Von Ahn，2009），其中文本标签通常采用简短音乐标签的形式。

表 1：MAQA 中生成的二元音频问答对示例。原始 AudioSet 示例是一段音乐音频片段，关联了以下标签：低音吉他、吉他、原声吉他和扫弦。基于 PaLM 的方法会根据采样属性生成问题及其被否定的对应问题。负属性钢吉他和滑棒吉他是从 AudioSet 本体的兄弟节点中采样得到的。

采样属性	问题	答案	被否定
低音吉他（+）	Q1）演奏的乐器	TRUE	No
低音吉他 (+)	歌曲中演奏的是低音吉他	TRUE	No
低音吉他 (+)	Q2）歌曲中没有演奏低音吉他	错误	Yes
钢吉他、滑棒吉他（–）	Q3）这首歌有钢吉他或滑棒吉他	错误	No
钢吉他、滑棒吉他（–）	Q4）这首歌没有滑棒吉他或钢吉他	TRUE	Yes

表 2：音乐音频问答（MAQA）基准的统计数据。音乐音频问答（MAQA）基准的统计数据。两个评估集 ASBaseEval 和 ASNegationEval 均由 PaLM 生成。每个训练集都有一个由模板生成的版本和一个由 PaLM 生成的版本。所有数据集的二元标签分布均衡。ASNegationEval 包含 ASBaseEval 及其否定版本。训练集和评估集划分中的音乐属性分布相似。

数据版本	标签统计 确实被否定	问答对数量 Non-被否定	被否定	提及项数量 流派	Mood	乐器角色	
ASBaseEval ASNegationEval	50% 0% 50% 50%	17,028 17,028	0 17,028	5574 （32.7%）	730 （4.3%）	9740 （57.2%）	984 (5.8%)
ASBaseTrain ASNegationTrain	50% 0% 50% 50%	1,263,004 1,263,004	0 1,263,004	439,904 (34.8%)	28 ,634 (2.3%)	740,126 (58.6%)	54,340 （4.3%）

多模态 Transformer。本研究将一系列基于 Transformer 的多模态模型（Sun 等，2019；Tan 和 Bansal，2019；Lu 等，2019）称为“多模态 Transformer”。这些研究探索了将 Transformer 编码器用作多模态融合的联合编码器，并在多种多模态问答任务上取得了最先进的成果。Changpinyo 等（2022年）提出了 T5 的多模态版本（Raffel 等，2019）。对于 VQA 示例中的图像和问题，多模态 T5 将预训练视觉编码器生成的全局和区域图像特征以及问题的文本词元作为输入，并针对 VQA 任务中的预定义答案类别解决分类问题。在 T5 微调期间，视觉编码器的参数保持冻结。我们沿用相同的方法，但使用预训练音频编码器（Gemmeke 等，2017；Huang 等，2022）提取音乐音频的全局表征。有关音频表征的详细综述，请参见 Huang 等（2022）。

3 音乐音频问答（MAQA）

为了评估多模态模型运用否定进行推理的能力，我们创建了一个音乐音频问答基准（MAQA），重点考察模型能否正确理解文本否定。音乐音频问答对是根据 AudioSet 中与音乐相关的部分以程序化方式生成的（Gemmeke 等，2017）。该部分包含带有音乐属性标注的音乐音频片段，以及描述这些属性之间关系的本体。训练集和评估划分中分别有 388、262 和 4,497 个唯一的音乐音频片段。每个片段都标注了 141 种独特音乐属性中的一种或多种，这些属性涵盖音乐流派、音乐角色、音乐乐器和音乐情绪。

表 1 展示了 MAQA 中的一个示例，其中包含四组问答对，这些问答对根据 AudioSet 中一个带有音乐属性标签的音频片段示例生成。Q1 和 Q2 根据同一种子属性生成，本质上考查的是同一项音乐技能，即聆听音乐音频并判断其中是否演奏了低音吉他。Q2 是 Q1 的否定形式。如果模型正确回答了 Q1，却未能正确回答其否定对应项 Q2，这表明模型不理解问题中的否定逻辑，也无法实现组合泛化。

MAQA 包含两个评测集 ASBaseEval 和 ASNegationEval，以及两个训练集 AS- BaseTrain 和 ASNegationTrain，如表 2 所示，其二元标签分布均衡，涵盖关于音乐情绪、流派、乐器和角色的问答对。ASBaseTrain / ASBaseEval 包含

关于音乐音频录音的非否定式问答对。ASNegationTrain / ASNegationEval 是 ASBaseTrain / ASBaseEval 的超集，还包含这些问答对对应的否定版本。具备强大音乐音频理解能力的多模态模型应当在 AS- BaseEval 上取得高分。此外，要展现其对否定逻辑的推理能力，该模型还必须在 ASNegationEval 上取得高分。

4 数据生成

由于标准文本语料库中很少出现包含否定的音乐描述文本（Hossain 等人，2022年），我们提出以下三步方法，以程序化方式生成涉及文本否定的二元音频问答对：1. 对于原始数据集中的每个音乐音频-属性对，我们抽取一个与该音频片段无关的负属性。2. 将正向和负向音频-属性对转换为二元音频问答示例，格式为三元组（音频片段、关于该属性的问题、真/假标签）。3. 对问题进行文本否定，并同时翻转二元标签，从而创建被否定的音频问答对。作为初步尝试，我们采用此方法从 AudioSet 构建了 MAQA；不过，该方法也可应用于包含带注释音乐音频的其他数据集。接下来，我们将讨论如何遵循这三个步骤，从 AudioSet 创建 MAQA 的具体细节。

负属性采样。我们采用负采样来创建均衡的二元标签分布。具体而言，我们利用本体树中的兄弟节点采样困难负属性，并为由此得到的音频问答对赋予 False 标签。以表 1 为例，该音频片段标注了低音吉他和原声吉他，它们都属于父节点吉他。我们从兄弟节点中采样钢吉他和滑棒吉他作为困难负属性，以创建负音频-属性对。这种困难负采样方法有助于模型区分相关但不同的音乐概念。

问题生成。我们探索了以下两种方法，从第一步采样得到的音频-属性对生成问题。第一种方法基于模板，并利用 AudioSet 本体，其中每个音乐属性都对应四种属性类型之一：流派、角色、乐器和情绪。我们使用特定类型模板将属性转换为判断题，例如：“歌曲的 <属性类型> 是 <属性值>。”第二种方法利用 PaLM（Chowdhery 等，2022年）的少样本文本生成能力，以提高生成问题的多样性。与 GPT-3（Brown 等，2020）类似，在收到指令提示时，例如“根据给定的音乐属性生成一个关于音乐的句子”，PaLM 能从少量示例中学习，并针对未见属性生成问题。

通过否定进行任务增强。基于模板的方法通过在名词前插入修饰语 not，将问题转换为否定形式，即“歌曲的 <属性类型> 是 not<属性值>。”，并翻转二元标签。这种方法的局限之一是它依赖于特定属性类型，且只能修饰名词。基于 PaLM 的方法克服了这一局限：通过少样本学习，模型可以同时修改名词和动词，生成不同的否定模式。例如，与乐器属性“吉他”相关的否定模式包括“没有吉他”“没有演奏吉他”以及“歌曲中没有低音吉他”。对于每种音乐属性，我们都使用 PaLM 生成若干问题候选项，并手动挑选最佳问题。表 1 中第 2 行和第 4 行的问题就是以这种方式生成的示例。附录 8.3 展示了 PaLM 生成的更多示例问题及所用提示。

5 多模态建模

根据视觉问答领域的研究（Changpinyo 等人，2022；Zhang 等人，2015），我们将音频问答视为一个二元分类任务。我们采用与（Changpinyo 等人，2022）类似的多模态 T5 架构来融合音频和文本输入，并以其多语言版本 MT5 替换 T5。每个音乐音频片段输入都表示为一个 128 维嵌入，该嵌入取自 VGGish（Gemmeke 等人，2017）[1]，（采用 VGG 卷积网络架构），或基于 Transformer 的 MuLan 模型（Huang 等，2022）。微调多模态 T5 时，音频编码器保持冻结。音频嵌入通过线性投影层映射至文本词元嵌入

[1]https://github.com/tensorflow/models/tree/master/research/audioset/vggish

空间，该投影层在微调开始时随机初始化。随后，将音频令牌和文本令牌作为向量序列输入预训练多层 MT5（Xue 等人，2021）编码器，并将最终的多模态表示分类为二元类别。多模态代码基于 Flaxformer 框架[2]。训练详情见附录 8.1。

6 实验与结果

我们分别在表 3 和表 4 中报告了 ASBaseEval 和 ASNegationEval 评测基准上的实验结果。音频问答任务被定义为一个二元分类问题，我们报告正类上最佳的 AUC-ROC 分数及相应的准确率。为了评估模型的组合泛化能力，即理解“没有人声”这类涉及否定的组合音乐概念的能力，我们按照 (Keysers 等人，2020) 建议的设计，将数据划分为训练集和测试集。根据设计，音乐属性或原子在训练集和测试集中具有相似的表示，而测试集包含训练集中未见过的原子新组合。复合散度（CD）用于定量衡量训练集和测试集划分中的复合分布差异（Keysers 等人，2020）；而在我们的研究中，CD 用作定性指标（附录 8.2 中的表 5）。这里的复合项是指对原子应用组合规则（例如否定）后得到的问答对。对于每种划分方案，我们比较了采用不同音频特征提取器的微调多模态 Transformer，以及不同规模的预训练 MT5 模型的性能。此外，我们还改变了训练集划分中使用的问答对类型（基于模板或基于 PaLM），并研究复合发散度如何影响模型学习否定。

6.1 音乐理解

表 3(a) 展示了第一个划分场景的结果：模型使用 PaLM 生成的非否定式问答对进行训练和评估。这项低 CD 实验为基础音乐概念建立了一个微调基线。在 ASBaseEval 基准上，经过微调的多模态 MT5 的曲线下面积-ROC 得分超过 90%；该基准涵盖了有关音乐风格、情绪、流派、乐器等的音频问答任务。请注意，平衡二元分类任务的随机基线为 50%。这表明，多模态 Transformer 能够通过微调有效融合音频和文本信号，即使它是从纯文本检查点热启动的。对模型在不同音乐属性上的表现进行探测后发现，“恐怖音乐”“儿童音乐”等音乐概念，以及“牛铃”等常见打击乐器，都很容易被微调后的模型掌握；而对于“鼓打贝斯”“迷幻音乐”等电子音乐流派，模型则较难理解。

我们进一步将 PaLM 生成的训练样例替换为模板生成的 QA 样例，从而形成中等 CD 设置。与低 CD 设置相比，模型在中等 CD 设置中的得分低约 6%（表 3(b)）。这表明，模型仍能通过组合泛化，将在不同语言语境中学到的大部分音乐知识迁移过来。在这两种划分场景中，表现最佳的多模态模型都是采用 Mulan 嵌入作为音频特征的 MT5-XL。

6.2 使用否定进行推理

对于第三种划分情形（表 3(b)），我们采用与表 3(a) 相同的微调设置，但在 ASNegationEval 上进行评估；其中一半未被否定的数据来自 ASBaseEval，另一半则包含与之对应的否定版本。如表4(a) 所示，在此高 CD 设置下，使用非否定式音频问答对（ASBaseTrain）微调的多模态 MT5，在 ASNegationEval 基准上的得分仅为 50%。尽管模型在未被否定的问答上仍得分约 80%（与其在表 3 中 ASBaseEval 上的准确率相当），但在对应的否定版本上得分仅约 20%。微调后，模型在这些否定问句上的表现甚至不如 50% 的随机猜测基线。这表明，尽管模型经过训练，能够正确回答未被否定的问题，但它们也学会了“忽略”否定问句中的否定线索。我们还表明，增大模型规模并不能提高 AUC-ROC 分数，这说明即使是更大的模型，也无法通过标准微调方法实现组合泛化。

[2]https://github.com/google/flaxformer

表 3：在两种不同复合散度（CD）设置下，ASBaseEval 的准确率。

	微调详情		ASBaseEval		
	模型	训练数据 - 问答类型	CD 类型	AUC	Acc
a)	MT5-Base+VGGish	ASBaseTrain-PaLM	低 CD	0.905	0.821
	MT5-XL+VGGish	ASBaseTrain-PaLM	低 CD	0.911	0.827
	MT5-Base+MuLan	ASBaseTrain-PaLM	低 CD	0.913	0.828
	MT5-XL+MuLan	ASBaseTrain-PaLM	低 CD	**0.918**	0.832
b)	MT5-Base+VGGish	ASBaseTrain-Temp	中 CD	0.847	0.771
	MT5-XL+VGGish	ASBaseTrain-Temp	中 CD	0.850	0.765
	MT5-Base+MuLan	ASBaseTrain-Temp	中 CD	0.845	0.766
	MT5-XL+MuLan	ASBaseTrain-Temp	中 CD	**0.851**	0.767

表 4：三种不同复合散度（CD）设置下 ASNegationEval 的准确率。

	微调详情		ASNegationEval				
					Acc		
	模型	训练数据 - 问答类型	CD 类型	AUC	Avg	Neg	无负例
a)	MT5-Base+VGGish	ASBaseTrain-PaLM	高 CD	0.524	0.513	0.218	0.803
	MT5-XL+VGGish	ASBaseTrain-PaLM	高 CD	0.525	0.525	0.247	0.802
	MT5-Base+MuLan	ASBaseTrain-PaLM	高 CD	**0.553**	0.541	0.273	0.802
	MT5-XL+MuLan	ASBaseTrain-PaLM	高 CD	0.528	0.520	0.220	0.819
b)	MT5-Base+VGGish	ASNegationTrain-PaLM	低 CD	0.896	0.814	0.814	0.813
	MT5-XL+VGGish	ASNegationTrain-PaLM	低 CD	0.903	0.821	0.821	0.821
	MT5-Base+MuLan	ASNegationTrain-PaLM	低 CD	0.905	0.821	0.821	0.822
	MT5-XL+MuLan	ASNegationTrain-PaLM	低 CD	**0.907**	0.825	0.825	0.825
c)	MT5-Base+VGGish	ASNegationTrain-Temp	中 CD	0.784	0.715	0.690	0.741
	MT5-XL+VGGish	ASNegationTrain-Temp	中 CD	0.823	0.743	0.724	0.763
	MT5-Base+MuLan	ASNegationTrain-Temp	中 CD	0.805	0.739	0.723	0.755
	MT5-XL+MuLan	ASNegationTrain-Temp	中 CD	**0.828**	0.750	0.740	0.759

6.3 任务增强

随后，我们在训练期间采用任务增强，将 PaLM 生成的否定式问答示例加入 ASBaseTrain（AsNegationTrain-PaLM），从而降低复合发散度。事实证明，任务增强是应对否定问题的有效策略。如表4(b)所示，采用任务增强进行微调的多模态 MT5，在 ASNegationEval 上的表现较表4(a)所示的基线提高了近 40%，同时在非否定式问答对（ASBaseEval）上的表现相近。其 AUC-ROC 分数和准确率与非否定式音频问答对（ASBaseEval）上的得分相当，这表明任务增强确实有助于模型学会正确回答带有否定的问题。在 ASNegationEval 上的最佳结果由使用 MuLan 音频嵌入进行微调的 MT5-XL 取得。

6.4 模板与 PaLM

我们进一步探究不同的任务增强策略如何影响学习结果。如表 4(c) 所示，我们采用基于模板的方法来编写问答对并进行任务增强，并将其与使用 PaLM生成的问答对进行微调的结果进行比较。与基于 PaLM 的微调相比，基于模板的微调的 AUC分数低约 10%。这种差距可能是由于训练数据与评测数据之间的复合发散度增大所致。在非否定子集上，准确率相差约 7%；在否定子集上，相差约 10% 至 12%。请注意，基于模板的方法只会替换名词来实现否定，而基于 PaLM 的方法则融入了更多变体，这或许可以解释为何基于模板的微调在否定子集上的表现较差。这项实验凸显了编写增强任务时采用自然语言变体的重要性，这些变体应贴近生产环境中使用的人类语言。然而，即使是基于模板的任务增强，也能显著提升否定理解能力，平均表现比不进行任务增强的训练高出 30%（表4(a)）。

7 结论

在这项工作中，我们提出了一个面向音乐领域的新型二元音频问答基准 MAQA，用于探测最先进的多模态模型对否定的理解能力。MAQA 填补了多模态设置中缺乏专注于否定的评测基准这一空白。我们的实验表明，标准微调会阻碍多模态 Transformer 将已学概念泛化到这些概念的否定形式。增加模型规模或采用更好的音频编码器并不能提升模型处理否定的能力，而任务增强则通过提供更多包含否定的微调示例，使模型能够对否定进行推理。此外，PaLM 等大语言模型可以生成包含更自然语言变体的否定示例，并显著影响学习结果。借助 MAQA 基准，我们希望推动多模态研究社区开发新的建模框架或算法，以处理包含否定的复杂自然语言指令。我们计划在 GitHub 上发布 MAQA 数据集。

参考文献

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh 和 Dhruv Batra。2015年。视觉问答：视觉问答。International Journal of Computer Vision，123:4–31。

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever 和 Dario Amodei。2020。语言模型是少样本学习者。ArXiv，abs/2005.14165。

Soravit Changpinyo、Doron Kukliansky、Idan Szpektor、Xi Chen、Nan Ding 和 Radu Soricut。2022年。视觉问答所需的一切或许就是图像说明。ArXiv，abs/2205.01883。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vinodkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael Isard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。2022年。PaLM：借助路径扩展语言模型。ArXiv。

R. Csordás、Kazuki Irie 和 Jürgen Schmidhuber。2021年。《细节决定成败：一个简单技巧》 s i提升 Transformer 的系统性泛化。ArXiv，abs/2108.12284。

Haytham M. Fayek 和 Justin Johnson。2020。《通过音频问答进行时间推理》。IEEE/ACM 音频、语音与语言处理汇刊，28：2283–2294。

Daniel Furrer、Marc van Zee、Nathan Scales 和 Nathanael Scharli。2020。《语义解析中的组合泛化：预训练与专用架构》。ArXiv，abs/2007.08970。

Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。2017年。《AudioSet：用于音频事件的本体和人工标注数据集》。发表于 IEEE 声学、语音与信号处理国际会议（ICASSP）论文集。

Md Mosharaf Hossain、Dhivya Chinnappa 和 Eduardo Blanco。2022年。自然语言理解语料库中否定的分析。载于

第60届计算语言学协会年会论文集（第2卷：短论文），第716–723页，都柏林，爱尔兰。计算语言学协会。

Md Mosharaf Hossain、Venelin Kovatchev、Pranoy Dutta、Tiffany Kao、Elizabeth Wei 和 Eduardo Blanco。2020。透过否定视角分析自然语言推理基准测试。载于《自然语言处理实证方法会议》（EMNLP）论文集，第 9106–9118 页。

Arian Hosseini、Siva Reddy、Dzmitry Bahdanau、R Devon Hjelm、Alessandro Sordoni 和 Aaron Courville。2021年。通过理解「不」来理解：在语言模型中建模否定。载于《2021年计算语言学协会北美分会年会：人类语言技术》论文集，第 1301–1312 页，线上。计算语言学协会。

Qingqing Huang、Aren Jansen、Joonseok Lee、Ravi Ganti、Judith Yue Li 和 Daniel P. W. Ellis。2022年。Mulan：音乐音频与自然语言的联合嵌入 https://arxiv.org/abs/2208.12415。载于第 23 届国际音乐信息检索学会会议（ISMIR 2022）。

Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。2021年。通过噪声文本监督扩展视觉及视觉语言表征学习。发表于国际机器学习会议（ICML）论文集。

Justin Johnson、Bharath Hariharan、Laurens Van Der Maaten、Li Fei-Fei、C Lawrence Zitnick 和 Ross Girshick。2017年。CLEVR：用于组合语言和基础视觉推理的诊断数据集。载于《IEEE计算机视觉与模式识别会议论文集》，第2901–2910页。Nora Kassner 和 Hinrich Schütze。2019年。否定版LAMA：鸟不会飞。CoRR, abs/1911.03343。Daniel Keysers、Nathanael Schärli、Nathan Scales、Hylke Buisman、Daniel Furrer、Sergii Kashubin、Nikola Momchev、Danila Sinopalnikov、Lukasz Stafiniak、Tibor Tihon、Dmitry Tsarkov、Xiao Wang、Marc van Zee 和 Olivier Bousquet。2020。测量组合泛化：一种基于真实数据的综合方法。载于《国际学习表征会议》。Kevin Kilgour、Beat Gfeller、Qingqing Huang、Aren Jansen、Scott Wisdom 和 Marco Tagliasacchi。2022年。文本驱动的任意声音分离。arXiv:2204.05738。Najoung Kim 和 Tal Linzen。2020。COGS：基于语义解释的组合泛化挑战。arXiv预印本 arXiv:2010.05465。Brenden Lake 和 Marco Baroni。2018。无需系统性的泛化：序列到序列循环网络的组合能力。载于《国际机器学习会议论文集》，第2873–2882页。PMLR。Edith Law 和 Luis Von Ahn。2009。输入一致性：一种利用人类计算游戏收集数据的新机制。载于《ACM SIGCHI计算机系统人因会议论文集》。Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019年。ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表征。CoRR，abs/1908.02265。Shikhar Murty、Tatsunori B. Hashimoto 和 Christopher Manning。2021年。DReCa：一种用于少样本自然语言推理的通用任务增强策略。载于《2021年北美计算语言学协会分会会议：人类语言技术论文集》，第1113–1125页，线上。计算语言学协会。Arsha Nagrani、Paul Hongsuck Seo、Bryan Seybold、Anja Hauth、Santiago Manen、Chen Sun 和 Cordelia Schmid。2022年。从图像说明中学习音视频模态。arXiv:2204.00679。

Andreea-Maria Oncescu、A Koepke、João F Henriques、Zeynep Akata 和 Samuel Albanie。2021年。利用自然语言查询进行音频检索。arXiv:2105.02192。

Santiago Ontanon、Joshua Ainslie、Zachary Fisher 和 Vaclav Cvicek。2022年。让 Transformer 解决组合任务。载于第 60 届计算语言学协会年会论文集（第 1 卷：长篇论文），第 3591–3607 页，爱尔兰都柏林。计算语言学协会。

Linlu Qiu、Peter Shaw、Panupong Pasupat、Tianze Shi、Jonathan Herzig、Emily Pitler、Fei Sha 和 Kristina Toutanova。2022年。评估模型规模对语义解析中组合泛化的影响。ArXiv，abs/2205.12253。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议（ICML）论文集。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2019年。探索使用统一的文本到文本 Transformer 进行迁移学习的极限。ArXiv 预印本 ArXiv:1910.10683。

Laura Ruis、Jacob Andreas、Marco Baroni、Diane Bouchacourt 和 Brenden M Lake。2020。一项用于扎根语言理解中系统性泛化的基准。《神经信息处理系统进展》，33:19861–19872。

Chen Sun、Austin Myers、Carl Vondrick、Kevin Murphy 和 Cordelia Schmid。2019年。VideoBERT：一种用于视频与语言表征学习的联合模型。CoRR，abs/1904.01766。

Hao Tan 和 Mohit Bansal。2019年。LXMERT：从 Transformer 中学习跨模态编码器表征。CoRR，abs/1908.07490。

Linting Xue、Noah Constant、Adam Roberts、Mihir Kale、Rami Al- Rfou、Aditya Siddhant、Aditya Barua 和 Colin Raffel。2021年。mT5：一种大规模多语言预训练文本到文本 Transformer。载于《2021年北美计算语言学协会分会会议论文集：人类语言技术》，第483–498页，线上。计算语言学协会。

Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vi- jay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan、Ben Hutchinson、Wei Han、Zarana Parekh、Xin Li、Han Zhang、Jason Baldridge 和 Yonghui Wu。2022年。扩展自回归模型，以实现丰富内容的文本到图像生成。ArXiv。Bowen Zhang、Hexiang Hu、Linlu Qiu、Peter Shaw 和 Fei Sha。2021年。视觉接地的概念组合。载于《计算语言学协会研究成果：2021年自然语言处理经验方法会议》，第201–215页，蓬塔卡纳，多米尼加共和国。计算语言学协会。Peng Zhang、Yash Goyal、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。2015年。阴与阳：平衡并回答二元视觉问题。arXiv。

8 附录

8.1 训练详情

对于 MT5-base 编码器，有 12 层 Transformer 编码器层，输入嵌入维度为 768。对于 MT5-XL，有 24 层 Transformer 编码器层，输入嵌入维度为 2048。MT5-XL 有 37 亿个参数，MT5-base 有 5.8 亿个参数。MT5-base 模型的批量大小为 64，MT5-XL 模型的批量大小为 128。我们在 $\{10^{-3}, 10^{-4}, 5^{-5}\}$ 中采用固定学习率，并观察到 1×10^{-3} 通常效果最佳。由于音频问答任务被表述为二元分类，模型输出 2 维 logits。我们训练所有模型

属性	训练示例	测试示例	CD 类型
(a) 不使用任务增强			
班卓琴	这首歌曲由班卓琴演奏。 (ASBaseTrain-PaLM)	这首歌曲由班卓琴演奏。 (ASBaseEval-PaLM)	低 CD
布鲁斯	这首歌曲的 [流派] 是 [布鲁斯]。 (ASBaseTrain-Temp)	这是一首布鲁斯歌曲。 (ASBaseEval-PaLM)	中 CD
恐怖	音乐很恐怖。 (ASBaseTrain-PaLM)	音乐很恐怖。 音乐并不恐怖。 (ASNegEval-PaLM)	高 CD
(b) 使用任务增强（ASNegationTrain）			
圣歌	这段音乐是一首圣歌。 这段音乐不是一首圣歌。 (ASNegTrain-PaLM)	这段音乐是一首圣歌。 这段音乐不是圣歌。 (ASBaseTrain-PaLM)	低 CD
舞蹈音乐	这首歌曲的 [音乐角色] 是 [舞曲]。 这首 [歌曲的音乐角色] 并非 [舞曲]。 (ASNegTrain-Temp)	这段音乐适合跳舞。 这首音乐不适合跳舞。 (ASBaseTrain-PaLM)	中 CD

表5：MAQA 的训练集和测试集划分设计。对于 5 种划分场景中的每一种，我们列出了一些包含在训练集和测试集中的示例问题。所有问答示例或复合项均由种子音乐属性或原子通过基于模板（Temp）或基于 PaLM（PaLM）的方法组合而成。复合散度（CD）类型用于衡量训练集与测试集的复合项分布差异。对于 ASNegationEval-LowCD 和 ASNegationEval-MedCD 划分场景，训练期间会应用任务增强。

采用 16 个 Cloud TPU Pod 进行数据并行。所有实验均运行 50,000 步，并报告基于 AUCROC 衡量的最佳检查点所对应的 AUCROC 和准确率。对于每项实验，我们选取多次运行中的最高 AUCROC。MT5-Base 大约需要 1 小时，MT5-XL 大约需要 4 小时。

8.2 复合差异度

由于 PaLM 生成的组合不依赖规则来组合不同的原子，因此很难直接计算 CD。这里我们使用复合发散度作为定性而非定量指标，来衡量训练集和测试集划分中的组合差异，如表5所示。

8.3 用于 PaLM 的文本提示

经过几轮迭代，我们发现表6所示的文本提示最适合生成否定和非否定式问答对。提示开头那条单行指令对于引导 PaLM 生成所需句子至关重要。表7展示了更多 PaLM生成的问题。用于文本生成的种子属性总数为 130；对于每个属性，我们从 PaLM 生成的候选问题中各选出一个未被否定的问题和一个被否定的问题。

用户输入（音乐标签）：小号

模型输入：

根据音乐标签，生成一个关于音乐及其否定的句子
音乐标签：鼓
描述音乐：这段音乐是用鼓演奏的
否定该描述：歌曲中没有演奏鼓
音乐标签：摇滚
描述音乐：这是一首摇滚歌曲
否定描述：这首歌曲的流派不是摇滚
音乐标签：快乐
描述音乐：这段音乐让人感到快乐
否定描述：这首歌曲的情绪不是快乐
音乐标签：合唱团
描述音乐：这首歌曲中有合唱团演唱
否定描述：这首歌曲中没有合唱团演唱
音乐标签：塔布拉鼓
描述音乐：这首歌曲中演奏了塔布拉鼓
否定描述：这首歌曲中没有演奏塔布拉鼓
音乐标签：小号
描述音乐：<额外 _id_1>
否定该描述：<额外_id_2>

模型输出：

<extra_id_1>: 这首歌曲中有小号演奏<extra_id_2>:
这首歌曲中没有小号演奏

表6：用于 PaLM 根据种子音乐属性生成未被否定和被否定句子的文本提示。此处展示了属性“小号”的一个示例输出。

种子属性	生成的问题	生成的否定问题
无伴奏合唱	这首歌曲没有使用任何乐器伴奏	这首歌曲并非没有任何乐器
手风琴	这首歌曲的乐器是手风琴	这首歌曲中没有演奏手风琴
原声吉他	这首歌曲使用原声吉-tar	这首歌曲不是使用原声吉他演奏的
氛围音乐	这首歌曲是氛围音乐	这首歌曲不是氛围音乐
愤怒音乐	这首歌曲让人感到愤怒	这首歌曲的情绪并不愤怒
背景音乐	这首歌曲是背景音乐	这首歌曲不是背景音乐
风笛	这首歌曲使用的乐器是风笛	这首歌曲中没有演奏风笛
班卓琴	这首歌曲使用班卓琴演奏	这首歌曲不是用班卓琴演奏的
低音鼓	这首歌曲使用的乐器是低音鼓	这首歌曲中没有演奏低音鼓
低音吉他	这首歌曲中的乐器是低音吉他	这首歌曲中没有演奏低音吉他
节奏口技	这首歌曲中有节奏口技	这首歌曲中没有节奏口技
Bell	这首歌曲使用的乐器是bell	这首歌曲中没有演奏钟
蓝草音乐	这首歌曲的流派是蓝草音乐	蓝草音乐不是这首歌曲的流派
布鲁斯	这首歌曲是一首布鲁斯歌曲	这首歌曲不是布鲁斯歌曲
弓弦乐器	这首歌曲使用的乐器是弓弦乐器	这段音乐不是用弓弦乐器演奏的乐器
铜管乐器	这首歌曲使用的乐器是铜管乐器	这首歌曲中没有演奏铜管乐器
卡纳提克音乐	这种音乐属于卡纳蒂克音乐	这种音乐不属于卡纳蒂克音乐
大提琴	这首歌曲使用的乐器是大提琴	这首歌曲中没有演奏大提琴
圣歌	这段音乐是一首圣歌	这段音乐不是一首圣歌
合唱团	这首歌曲有合唱团参与演唱	这首歌曲没有合唱团参与演唱
基督教音乐	这是基督教音乐	这不是基督教音乐
圣诞音乐	这首音乐是圣诞音乐	这首音乐不是圣诞音乐
单簧管	这首歌曲由单簧管演奏	这首歌曲中没有单簧管演奏
古典音乐	这段音乐是古典音乐	这首歌曲的流派不是古典音乐
乡村音乐	这首歌曲的流派是乡村音乐	这不是一首乡村音乐歌曲
牛铃	这首歌曲中有牛铃	这首歌曲中没有牛铃
钹	歌曲中的乐器是钹	歌曲中没有演奏钹
舞曲	这首音乐适合跳舞	这首音乐不适合跳舞
迪吉里杜管	这首音乐由迪吉里杜管演奏	这首音乐不是由迪吉里杜管演奏的
迪斯科	这首歌曲属于迪斯科流派	这首歌曲不属于迪斯科流派
低音提琴	这首歌曲所使用的乐器是低音提琴	这首歌曲中没有演奏低音提琴
鼓打贝斯	这段音乐是鼓打贝斯	这首歌曲不是鼓打贝斯

表7：更多由 PaLM 生成的未被否定和被否定问句示例。