软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2024−2034 [doi: 10.13328/j.cnki.jos.006770]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向视觉语言理解与生成的多模态预训练方法*

刘天义[1,2,3], 吴祖煊[1,2,3], 陈静静[1,2,3], 姜育刚[1,2,3]

[1](复旦大学 计算机科学技术学院, 上海 200438)

[2](上海市智能信息处理重点实验室 (复旦大学), 上海 200438)

[3](上海市智能视觉计算协同创新中心 (复旦大学), 上海 200438)

通信作者: 吴祖煊, E-mail: zxwu@fudan.edu.cn; 姜育刚, E-mail: ygj@fudan.edu.cn

摘　要: 大多数现有的视觉语言预训练方法侧重于理解任务, 并在训练时使用类似于 BERT 的损失函数 (掩码语言建模和图像文本匹配). 尽管它们在许多理解类型的下游任务中表现良好, 例如视觉问答、图像文本检索和视觉蕴涵, 但它们不具备生成信息的能力. 为了解决这个问题, 提出了视觉语言理解和生成的统一多模态预训练 (unified multimodal pre-training for vision-language understanding and generation , UniVL). UniVL 能够处理理解任务和生成任务, 并扩展了现有的预训练范式, 同时使用随机掩码和因果掩码, 因果掩码即掩盖未来标记的三角形掩码, 这样预训练的模型可以具有自回归生成的能力. 将几种视觉语言理解任务规范为文本生成任务, 并使用基于模版提示的方法对不同的下游任务进行微调. 实验表明, 在使用同一个模型时, 理解任务和生成任务之间存在权衡, 而提升这两个任务的可行方法是使用更多的数据. UniVL 框架在理解任务和生成任务方面的性能与最近的视觉语言预训练方法相当. 此外, 实验还证明了基于模版提示的生成方法更有效, 甚至在少数场景中它优于判别方法.

关键词: 计算机视觉; 多模态学习; 预训练

中图法分类号: TP391

中文引用格式: 刘天义, 吴祖煊, 陈静静, 姜育刚. 面向视觉语言理解与生成的多模态预训练方法. 软件学报, 2023, 34(5): 2024–2034. http://www.jos.org.cn/1000-9825/6770.htm

英文引用格式: Liu TY, Wu ZX, Chen JJ, Jiang YG. Multimodal Pre-training Method for Vision-language Understanding and Generation. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2024–2034 (in Chinese). http://www.jos.org.cn/1000-9825/6770.htm

Multimodal Pre-training Method for Vision-language Understanding and Generation

LIU Tian-Yi[1,2,3], WU Zu-Xuan[1,2,3], CHEN Jing-Jing[1,2,3], JIANG Yu-Gang[1,2,3]

[1](School of Computer Science, Fudan University, Shanghai 200438, China)

[2](Shanghai Key Laboratory of Intelligent Information Processing (Fudan University), Shanghai 200438, China)

[3](Shanghai Collaborative Innovation Center of Intelligent Visual Computing (Fudan University), Shanghai 200438, China)

Abstract: Most existing vision-language pre-training methods focus on understanding tasks and use BERT-like loss functions (masked language modeling and image-text matching) during pre-training. Despite their good performance in the understanding of downstream tasks, such as visual question answering, image-text retrieval, and visual entailment, these methods cannot generate information. To tackle this problem, this study proposes unified multimodal pre-training for vision-language understanding and generation (UniVL). The proposed UniVL is capable of handling both understanding tasks and generation tasks. It expands existing pre-training paradigms and uses random masks and causal masks simultaneously, where causal masks are triangular masks that mask future tokens, and such pre-trained models can have autoregressive generation abilities. Moreover, several vision-language understanding tasks are turned into text generation tasks according to specifications, and the prompt-based method is employed for fine-tuning of different downstream tasks. The experiments show

* 基金项目: 科技创新 2030——“新一代人工智能”重大项目 (2021ZD0112805); 国家自然科学基金青年基金 (62102092)
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: 2022-04-17; 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: 2023-03-23

that there is a trade-off between understanding tasks and generation tasks when the same model is used, and a feasible way to improve both tasks is to use more data. The proposed UniVL framework attains comparable performance to recent vision-language pre-training methods in both understanding tasks and generation tasks. Moreover, the prompt-based generation method is more effective and even outperforms discriminative methods in few-shot scenarios.

Key words: computer vision; multimodal learning; pre-training

受自然语言处理中大规模预训练模型成功的启发, 人们提出了各种视觉语言预训练方法, 旨在从大规模图像-文本对中学习多模态表示. 一旦得到了预训练模型, 这些预先训练好的模型就可以进行微调, 以执行各种下游任务. 这种简单的预训练和微调范式最近在许多具有挑战性的视觉和语言任务中显示出巨大的潜力[1−5], 如视觉问答、图像文本检索、图像字幕和视觉蕴涵.

视觉语言下游任务分为两类: 理解任务和生成任务. 理解任务包括视觉问答、视觉蕴涵、图像分类和图像文本检索. 大多数现有的视觉语言预训练方法将此类任务规范为判别性任务, 它们要求模型从预定义的答案列表中选择答案, 例如, 对于视觉问答, 现有方法将其描述为多答案分类任务, 并将 CLS 标记输入额外的线性分类器来得到分类结果[3,4,6−8]. 这些任务通常要求模型粗略理解图像和文本的语义信息, 例如, 文本是否描述了图像的内容? 图像和文本之间的关系是什么, 蕴涵、中立还是矛盾? 相比之下, 生成任务通常要求模型生成描述特定图像的完整句子. 一个典型的例子是图像字幕生成, 它要求模型输出一个描述图像内容的句子. 现有的视觉语言方法采用类似于 BERT[9]的损失函数, 例如掩码语言建模和图像文本匹配来学习多模态表示. 它们在理解任务方面表现良好, 但不能直接应用于生成性任务.

我们提出了统一的多模态视觉语言理解和生成预训练 (unified multimodal pre-training for vision-language understanding and generation , UniVL), 该预训练模型通过共享参数可以同时处理理解任务和生成任务. 我们首先使用图像编码器和文本编码器分别对图像和文本进行编码. 然后, 我们使用一个多模态编码器来融合图像和文本特征, 并使用交叉注意力融合图像特征和文本特征. 与现有的视觉语言预训练方法一样, 该方法使用了两个共同的训练目标, 包括掩码语言建模和图像文本匹配. 然而, 与之前的方法不同, 我们在预训练时不仅使用了双向可见掩码, 还使用了因果掩码. 因果掩码的好处是允许模型进行自回归解码, 这对于完整句子的生成非常重要. 统一的理解和生成预训练得到了具有共享参数的统一模型, 从而减少了训练不同模型的需要.

典型的视觉语言预训练方法总是针对不同的下游任务训练多个任务特定的线性层. 这种策略是为了使预训练的模型适应不同的下游任务, 它需要设计特定于任务的目标函数. 基于模版提示的方法最近引起了人们的注意, 并被证明是简单有效的. 下游任务可以重新规范化为预训练模型在预训练期间学习到的任务模式. 以主题分类为例, 输入的是一个句子, “他在打篮球.”, 输出为多类标签, 包括健康、政治、体育或其他. 我们不需要添加另一个线性分类器来微调预先训练好的模型, 而是可以构造一个查询语句, “他在打篮球. 主题是关于__”, 并要求预训练的模型来填充空白. 填充空白任务对于预训练模型来说很熟悉, 因为它是预训练目标之一 (掩码语言建模). 与广泛使用的预训练和微调范式相比, 基于模版提示的方法将不同类型的下游任务的输入输出格式转换为预训练期间模型处理的输入输出格式, 可以更好地释放预训练模型的潜力. 我们将以前的一些分类任务规范为文本生成任务, 并使用语言模版对预训练好的模型进行微调.

本文第 1 节介绍大规模预训练模型的相关方法和研究现状. 第 2 节介绍本文构建的统一的多模态视觉语言理解和生成预训练模型. 第 3 节通过对比实验验证了所提模型的有效性. 第 4 节总结全文.

1 大规模预训练模型相关工作

随着大规模语言模型上预训练的成功, 视觉语言预训练最近得到了人们关注, 相关研究显示视觉语言预训练模型在多种下游任务上表现优异. 现有的大多数方法都基于 Transformer 架构, 并使用类似于 BERT 的训练目标: (1) 多模态掩码语言建模[2,5,7]: 根据输入图像和文本上下文预测掩码词或掩码视觉特征. (2) 图像文本匹配[1,7]: 预测输入图像是否与输入文本匹配. 他们在自注意力模块中使用双向可见的掩码, 这导致预训练任务形式和需要自回归生成的下游任务形式之间存在差异. 受 UniLM[10,11]的启发, 我们在训练前将因果掩码与双向可见掩码相结合, 这

样我们的预训练模型就可以同时具备理解和生成能力.

随着预训练模型的参数数量的增加, 针对某一个下游任务对预训练模型的所有参数进行微调会带来极大的代价. 模版提示方法的思想是将下游任务的数据输入输出格式转换为预训练期间模型已经学习到的格式. 例如, 在自然语言处理中, 情绪分类任务可以表述为带有掩码标记的自然语言句子, 模型需要填写单词 positive 或 negative. 由于数据输入的相似性, 预训练模型可以直接应用于下游任务, 而无需参数调整.

早期基于模版提示的方法通常采用手工制作的模板. 例如, Petroni 等人[12]为 LAMA 数据集中的知识探索任务手动设计的完形填空模板. GPT-3[13]为问答、翻译和探测任务设计前缀模板. 虽然这些手工设计的模版具有高度可解释性, 并且通常非常有效, 但它们需要大量的尝试, 不同的任务需要不同的领域经验. 我们使用手工制作的自然语言模版和参数化的模版进行实验.

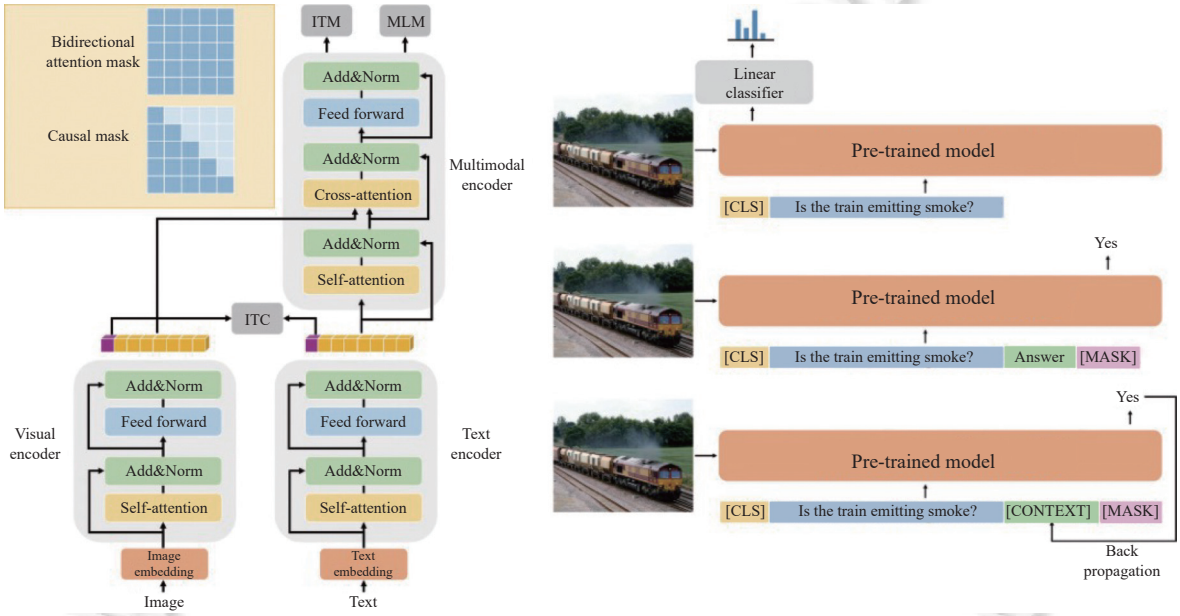

图 1 本文提出的模型结构和多模态提示模版

2 模型结构

本文所提出的预训练模型主要包括视觉编码器、文本编码器和多模态编码器 3 个部分, 下面就相关部分予以介绍.

2.1 视觉编码器

我们使用在 ImageNet-1k 上预训练的 ViT[14]作为视觉编码器来提取图像特征. 首先将输入图像 $I \in R^{C\times H\times W}$ 展开为 $N = HW/P^2$ 个图像块, 其中输入图像的分辨率为 $H\times W$, C 是通道数, 每个图像块的分辨率为 $P\times P$. 与 BERT 使用的 [CLS] 标记类似, ViT 为图像序列准备了一个参数化的可学习的标记 [CLS]. 视觉编码器由交替的多头自我注意模块 (MSA) 和多层感知机模块 (MLP) 组成, 其中多层感知机包含两个线性层和一个激活层. 视觉编码器在每一层中还使用了层归一化和残差连接.

$$z_0 = \left[v_{\text{CLS}}; v_p^1 V; v_p^2 V; \ldots; v_p^N V\right] + V_{\text{pos}},$$

$$z'_l = MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1},\ l = 1, \ldots, L_V,$$

$$z_l = MLP\left(LN\left(z'_l\right)\right) + z'_l,\ l = 1, \ldots, L_V,$$

其中, $v_p^1, \ldots, v_p^N$ 是展开的 2D 图像块, [CLS] 是可以学习的头部标记, z_l 是第 l 层的隐状态.

2.2 文本编码器

我们使用 BERT 作为文本编码器. 文本编码器和视觉编码器类似, 都包含了几层多头自我注意模块和多层感知机模块, 和视觉编码器的不同点在于层归一化作用于多头自我注意模块和多层感知机模块之后. 输入文本 $t \in R^{L\times O}$ 使用词嵌入矩阵 $T \in R^{O\times H}$ 和位置编码 $T_{\text{pos}} \in R^{(L+1)\times H}$ 被嵌入为 $t \in R^{L\times H}$.

$$p_0 = \left[t_{\text{CLS}}; t^1T; t^2T; \ldots; t^NT\right] + T_{\text{pos}},$$

$$p'_l = LN(MSA(p_{l-1})) + p_{l-1},\ l = 1,\ldots,L_T,$$

$$p_l = LN(MLP(p'_l)) + p'_l,\ l = 1,\ldots,L_T,$$

其中, $t^1,\ldots,t^N$ 是输入单词, p_l 是第 l 层输入序列的隐状态.

2.3 多模态编码器

多模态编码器与文本编码器类似, 只是需要额外进行一次交叉注意力的计算以融合图像特征和文本特征.

$$m_0 = p_{L_T},$$

$$m''_l = LN(MSA(m_{l-1})) + m_{l-1},\ l = 1,\ldots,L_M,$$

$$m'_l = LN(MCA(m''_l, z_{L_V})) + m''_l,\ l = 1,\ldots,L_M,$$

$$m_l = LN(MLP(m'_l)) + m'_l,\ l = 1,\ldots,L_M,$$

其中, p_{L_T} 是文本编码器的输出, z_{L_V} 是视觉编码器的输出, m_l 是第 l 层序列的隐状态.

多头自我注意模块中使用的注意力掩码是双向的, 每个标记都可以关注所有其他标记. 双向可见的掩码在判别性任务中表现良好, 但不适用于生成性任务. 通常, 生成性任务要求模型以自回归方式生成标记, 即从左到右. 为了在预训练时解决这个问题, 我们在文本编码器和多模态编码器的自我注意力模块中以不同的比例混合了两种注意力掩码.

2.4 多模态提示模版

之前的视觉语言预训练方法总是使用 [CLS] 标记作为图像-文本多模态表示, 并添加其他线性层, 以便对下游任务进行微调. 例如, 在 UNITER[4]中, 视觉问答被描述为一个多答案分类问题, 它将 [CLS] 标记作为线性层的输入并对线性层进行微调. 然而, 我们使用模版将视觉问答描述为一个文本生成问题. 例如, 当回答“他在做什么?”的时候, 我们可以继续输入提示“答案: ”, 因此模型接受的完整输入是“他在做什么? 答案: ”, 我们要求预训练的模型通过生成填补这些空白. 此外, 如图 1 所示, 我们用参数化的可学习的标记替换手工设计的语言模版. 这是因为设计一个合适的自然语言提示模版很难, 这需要领域专家的专业知识, 并且需要花费大量时间进行调整, 因为自然语言提示中每个单词的细微变化可能会对下游任务性能产生巨大影响. 我们使用标记器的 [UNUSED] 标记作为可学习的标记, 因为它们是参数化的, 可以通过反向传播进行更新.

2.5 训练目标

图像-文本对比缺失已被证明对视觉语言预训练有效. 我们利用图像-文本对比损失学习一个共同的低维空间来嵌入图像和文本. 我们将匹配的图像-文本对视为正样本, 将训练批次中的所有其他随机的图像-文本对视为负样本. 我们将两个损失之和最小化: 一个用于图像到文本, 另一个用于文本到图像.

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_i^B \log \frac{\exp\left(x_i^{\mathrm{T}} y_i/\sigma\right)}{\sum_{j=1}^B \exp\left(x_i^{\mathrm{T}} y_i/\sigma\right)},$$

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_i^B \log \frac{\exp\left(y_i^{\mathrm{T}} x_i/\sigma\right)}{\sum_{j=1}^B \exp\left(y_i^{\mathrm{T}} x_i/\sigma\right)},$$

$$\mathcal{L}_{itc} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i},$$

其中, x_i 是视觉编码器的第 1 层输出, y_i 是文本编码器的第 1 层输出. B 是批处理数据的大小, σ 是缩放数值的温度参数.

我们以 15% 的概率随机屏蔽文本标记, 并用一个特殊的 [MASK] 标记替换它们, 模型需要使用图像和上下文文本预测被屏蔽词.

$$\mathcal{L}_{mlm}=\sum H(p_{\text{mask}},y_{\text{mask}}),$$

其中, H 是交叉熵, p_{mask} 是模型对屏蔽标记的预测概率, y_{mask} 是词汇的概率分布, $\mathcal{L}_{mlm}$ 是每个被屏蔽词汇的交叉熵之和.

我们使用多模态编码器输出的第一个标记作为视觉语言两种模态的融合表示, 并附加一个完全连接层后使用 Softmax 来预测两类是否匹配的概率 p_{itm}. 图像-文本匹配损失函数预测图像和文本对是否匹配或不匹配. 通过将匹配样本中的图像或文本替换为从其他样本中随机选择的图像或文本, 可以创建负样本.

$$\mathcal{L}_{itm}=\sum H(p_{itm},y_{itm}),$$

其中, y_{itm} 是一个独热向量, 表示真值标签, 其中 1 表示匹配的图像文本对, 0 表示不匹配的图像文本对, $\mathcal{L}_{itm}$ 是所有正样本和负样本的交叉熵之和.

3 实验分析

3.1 实验数据

我们使用 GCC3M 和 COCO 作为预训练的数据集, 我们遵循 Karpathy 对于 COCO 的分割方式[15]. 最终训练集中总共的不同图像数量达到了 2.84M 张.

3.2 实现细节

我们使用 12 层 ViT-B/16[14]作为图像编码器, 使用在 ImageNet-1k 上预先训练的权重进行初始化. 文本编码器使用 BERT 模型的前 6 层初始化, 多模态编码器使用 BERT 模型的后 6 层初始化. 我们在 32 张 NVIDIA Tesla V100 32 GB GPU 上使用 2048 批量大小的数据预训练了 30 个周期. 我们使用的 AdamW 优化器的学习率为 1E–4, 权重衰减为 0.02.

3.3 下游任务

图像文本检索要求模型从候选图像集中选出符合给定描述的图像, 或者从候选描述集中选出符合图像内容的描述语句. 因此, 它包含两个子任务: 图像到文本检索和文本到图像检索. 我们使用图像-文本对比损失和图像-文本匹配损失来评估图像和文本之间的相似性. 在推理过程中, 我们首先使用视觉编码器和文本编码器计算所有图像-文本对的特征相似性. 然后, 我们选择评分最高的前 K 对, 并使用多模态编码器计算图像-文本匹配分数后进行排名. 我们在 Flickr30k[16]和 COCO[17]上评估了我们的模型.

图像字幕生成旨在生成描述图像内容的句子. 由于我们使用因果掩码对模型进行预训练, 我们的模型可以直接为图像生成一个句子. 在句子生成过程中, 我们首先使用 [CLS] 标记和 [MASK] 标记作为输入对图像编码器的输入进行编码. 特殊标记 [CLS] 是句子的开头, 我们的模型预测 [MASK] 位置的单词. 然后, 我们将另一个 [MASK] 标记追加到生成的标记序列中, 并预测下一个单词, 以此类推. 当模型输出 [SEP] 时, 生成过程终止. 我们在实验中使用了集束搜索, 设置集束大小为 5, 并在 COCO 图像字幕数据集上报告了实验结果.

视觉问答要求模型回答针对给定图像的给定问题. 现有的方法通常将视觉问答描述为一个多答案分类问题. 我们将视觉问答视为文本生成任务. 我们为视觉问答设计了两种提示模版: 自然语言提示模版和可学习上下文的参数化提示模版. 我们在 VQAv2[18]上评估我们的方法.

细粒度图像分类侧重于识别难以区分的图像类, 例如花卉的种类或动物的种类. 我们使用细粒度图像分类任务来评估模型的多模态理解能力. 我们将细粒度图像分类任务规范为文本生成, 并设计自然语言提示模版和可学习上下文的参数化提示模版. 与判别方法相比, 基于模版提示的方法具有更好的小样本学习能力. 我们在

Food101[19]、Flowers102[20]、DTD[21]上评估了我们的方法.

视觉蕴含[22]是一个细粒度的视觉推理任务, 用于预测图像和文本之间的关系是包含的、中性的还是矛盾的. 我们为视觉蕴含设计了自然语言提示模版和可学习上下文的参数化提示模版, 并将其与判别方法进行了比较.

3.4 实验结果

我们使用图像文本检索任务来评估预训练模型的视觉语言理解能力. 表 1 报告了 Flickr30k 上的零样本和微调图像文本检索的结果. 对于零样本检索, UniVL 用更少的数据达到了和 CLIP[23]和 ALIGN[24]相近的结果. 对于微调检索, UniVL 的召回率比 UNITER[4]高很多, 并且与 ALIGN[24]相似, 尽管它在更大的数据集上进行了预训练 (1.2B).

表 1 多模态检索实验结果

方法	预训练图像数量	Flickr30k多模态检索结果 (零样本/微调)					
		图像-文本检索			文本-图像检索		
		R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
UNITER	4M	83.6/87.3	95.7/98.0	97.7/99.2	68.7/75.6	89.2/94.1	93.9/96.8
CLIP	400M	88/—	98.7/—	99.4/—	68.7/—	90.6/—	95.2/—
ALIGN	1.2B	**88.6**/**95.3**	98.7/**99.8**	99.7/**100**	**75.7**/**84.9**	**93.8**/**97.4**	**96.8**/**98.6**
UniVL	3M	86.8/94.3	**98.7**/99.4	**99.7**/99.8	73.4/82.8	92.1/96.7	96.0/98.4

我们使用图像字幕生成任务来评估预训练模型的生成能力. 遵循 Karpath 的分割方式, 我们在 COCO 数据集上评估了自动图像字幕生成的性能, 该算法将训练图像和验证图像重新拆分为 113 287、5 000 和 5 000, 分别用于训练、验证和测试. 如表 2 所示, 我们报告了 4 个常用指标的结果: BLEU4[25], CIDEr[26], METEOR[27], SPICE[28]. 我们将我们的预训练模型与其他生成性视觉语言预训练方法进行了比较. 我们的 UniVL 模型与最近的生成性预训练方法相比, 具有相当的性能.

表 2 图像字幕生成实验结果

Method	BLEU4	CIDEr	METEOR	SPICE
Unified VLP	36.5	117.7	28.4	21.3
XGPT	**37.2**	**120.1**	28.6	21.8
VL-T5	34.6	116.1	**28.8**	**21.9**
VL-BART	34.2	114.1	28.4	21.3
UniVL	35.6	116.8	28.6	21.4

我们将视觉问答规范为文本生成任务, 而不是多答案分类任务. 我们为视觉问答设计了自然语言提示模版和可学习上下文的参数化提示模版. 自然语言提示模版为“[QUESTION] Answer: [ANSWER]”, 其中 [QUESTION] 表示问题文本, [ANSWER] 表示答案文本. 我们屏蔽了 [ANSWER] 中的单词标记, 并在微调过程中优化掩码语言建模损失. 在推理过程中, 输入文本是“[QUESTION] Answer: [MASK]”. 模型预测单词并将 [MASK] 反复添加到生成的序列中, 直到生成 [SEP] 标记. 可学习上下文的参数化提示模版将自然语言提示替换为可学习的标记, 而对于视觉问答, 则是“[QUESTION] [CTX] [ANSWER]”, [CTX] 是可学习标记的序列, 我们使用 BERT 标记器的 [UNUSED] 标记作为 [CTX]. [CTX] 的长度为 16. 与自然语言提示符相比, [CTX] 是一个参数化的标记提示, 可以与其他参数一起更新.

以前的判别方法将视觉问答规范为一个多答案分类问题, 并要求模型从预定义的答案列表中选择答案. 为了对判别方法和生成方法进行细粒度的比较, 我们将 Karpathy 的测试数据集拆分为两类: 答案在预定义答案列表中的问题 (域内样本) 和答案不在列表中的问题 (域外样本). 预定义答案列表的大小为 3 129, 域内样本和域外样本的数量分别为 25 750 和 530. 典型的判别方法无法回答域外问题, 因为它们的答案很少, 而且不在预定

义的答案列表中. 为了比较判别方法和生成方法的泛化能力, 我们将域外样本的答案附加到预定义的答案列表中, 并使用扩展的答案列表对判别法进行微调. 对于判别方法, 我们将多模态编码器输出的第 1 个隐状态输入到 1 个额外的线性分类器中进行答案预测. 如表 3 所示 (LC 为使用线性层微调, NLP 为使用自然语言模版, LCP 为使用可学习的参数化模版), 与判别性方法相比, 基于模版提示的生成方法在这两个类别中都表现得更好, 并且在域外样本中进行比较时, 提升更为显著. 为了评估基于模版提示的方法的小样本学习能力, 我们使用了不同数量的 Karpathy 训练数据. 如表 4 所示, 自然语言提示模版和可学习上下文的参数化提示模版的性能都优于判别方法.

表 3 COCO 数据集视觉问答实验结果

方法	域内	域外	平均
LC	70.8	3.7	69.4
NLP	68.4	13.9	67.3
LCP	**72.1**	**15.1**	**71.0**

表 4 不同训练样本数量下视觉问答实验结果

方法	训练样本数量 (域内/域外)			
	4k	22k	44k	88k
LC	0.5/0	5.6/0	10.9/0.5	15.4/0.9
NLP	0.9/**0.1**	11.9/**0.9**	14.8/1.1	18.3/1.6
LCP	**0.9**/0	**12.4**/0.7	**16.7**/**1.5**	**20.1**/**2.4**

为了与最新的视觉语言预训练方法进行公平比较, 我们参考之前的方法 UNITER[4], 使用 VQAv2 的训练集和验证集来微调预训练模型. 如表 5 所示, 我们的 UniVL 达到了与最先进的方法相当的性能.

与视觉问答类似, 我们将图像分类规范为文本生成任务, 并为图像分类设计自然语言提示模版和可学习上下文的参数化提示模版. 自然语言提示模版是“a photo of [CATEGORY]”, 相应的可学习上下文提示是“[CTX] [CATEGORY]”. [CATEGORY] 是图像的类名, 我们在微调过程中屏蔽了 [CATEGORY]. 如表 6 所示, 对于这个下游任务, 可学习上下文的参数化提示模版是一种有效的方法, 因为提示模版是一种更接近预训练任务的形式.

表 5 视觉问答与视觉蕴含测试集实验结果

方法	视觉问答		视觉蕴含	
	test-dev	test-std	val	test
VisualBERT	70.8	71	—	—
12-in-1	73.15	—	—	76.95
UNITER	72.7	72.91	78.59	78.28
ViLT	70.94	—	—	—
VILLA	**73.59**	**73.67**	79.47	79.03
UniVL	72.31	72.53	**79.70**	**80.00**

表 6 判别式方法和生成式方法图像分类实验结果

方法	微调模块	Food101	Flowers102	DTD
LC	VE	92.8	**93.8**	**65.4**
NLP	VE	88.4	90.1	53.3
LCP	VE	92.8	93.4	62.1
	TE	78.6	74.6	19.2
	ME	79.4	76.2	20.6
	VETE	92.8	93.5	63.3
	VEME	**93.3**	93.7	63.5

图 2 显示可学习上下文的参数化提示模版具有更好的小样本学习能力. 由于提示模版是预训练模型更熟悉的输入数据形式, 因此可学习上下文提示模版可以更好地利用从预训练模型中学习到的知识. 与视觉问答不同, 图像分类任务具有简单的文本输入, 并且训练样本的数量非常少, 因此更新预训练模型的所有参数是不明智的. 如表 6 所示, 视觉编码器是细粒度图像分类任务的关键, 因为文本很简单, 除了类名之外几乎不包含语义信息.

并非所有生成性方法都比判别性方法表现更好. 判别性方法更适合于其他一些类别更少的下游任务, 视觉蕴涵就是其中之一. 我们也设计了自然语言提示模版和可学习上下文的参数化提示模版, 将视觉蕴含规范为文本生成任务. 自然语言提示是“[SENTENCE] Relationship: [LABEL]”, 可学习上下文的参数化提示模版是“[SENTENCE] [CTX] [LABEL]”, [LABEL] 是图像和 [SENTENCE] 之间的关系, 它是蕴涵、中立和矛盾的关系之一. 我们在微调过程中屏蔽了 [LABEL]. 与根据每个单词的分数从词汇表中预测一个单词不同, 我们还对每个可能答案的分数进行排序, 并在推理过程中返回分数最高的答案. 它是一种判别性的方法, 与使用额外的线性分类器相比, 它更适用于预训练模型, 因为输入是带有文本的图像, 目标是掩码语言建模. 如表 7 所示, 1in3 意味着我们用 3 个词来预测 [MASK]: 蕴涵、中立和矛盾. 应该注意的是, 我们可以在视觉蕴涵中做到这一点, 因为视觉蕴涵的每个答案都是一个单词, 因此没有共同的前缀. 与生成式方法相比, 判别式方法更适合于视觉蕴涵, 因为候选答案的集合太小.

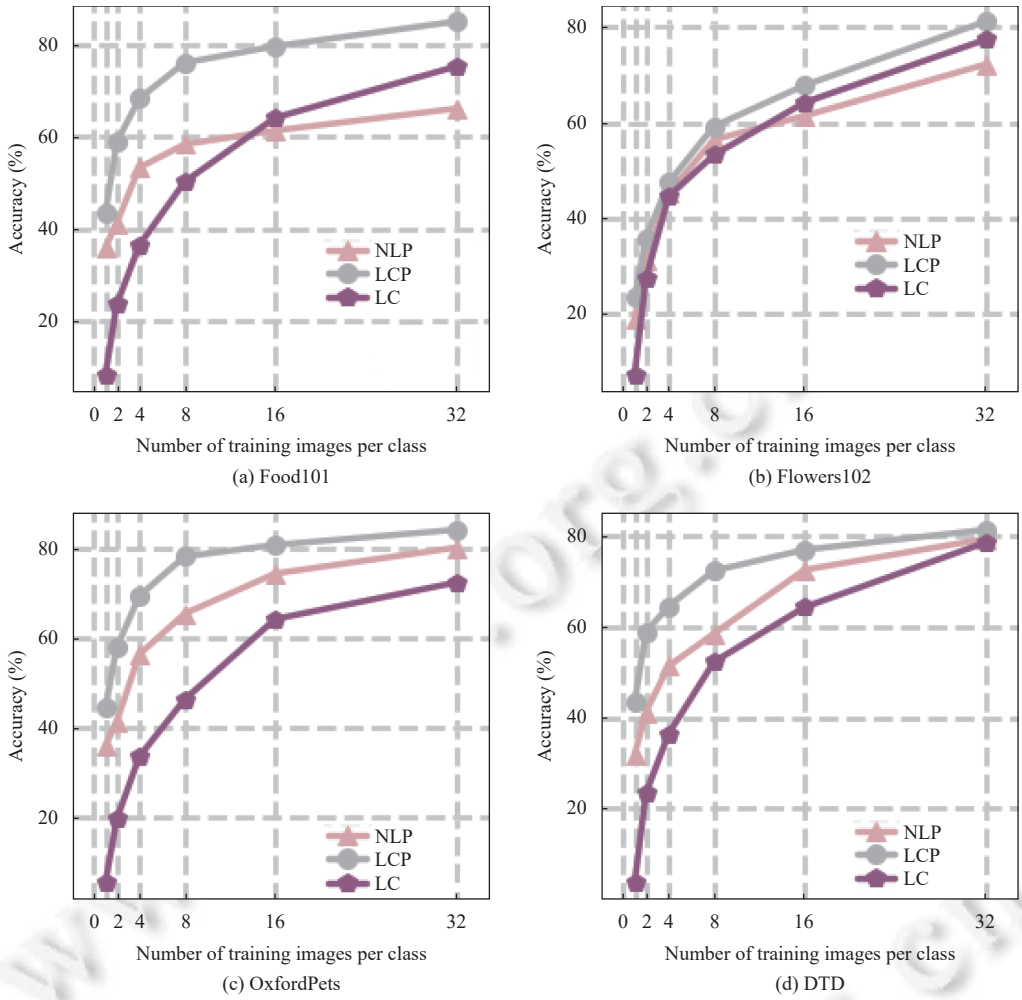

图 2 不同方法的小样本学习结果

表 7 视觉蕴含实验结果 (%)

Method	val	test
LC	78.4	78.1
NLP	65.4	65.7
NLP (1in3)	75.3	75.9
LCP	77.6	78.0
LCP (1in3)	**79.7**	**80.0**

3.5 消融实验

我们首先评估预训练中因果掩码矩阵的有效性. 表 8 显示了预训练模型的多模态生成能力和理解能力, 在预训练期间, 我们使用双向注意掩码矩阵和因果掩码矩阵的不同混合比例以及不同数量的图像-文本对. 我们使用图像字幕生成任务来评估模型的生成能力, 使用图像分类、视觉蕴涵和图像文本检索来评估模型的理解能力. 对于图像字幕生成, 我们将特殊标记 [MASK] 附加到序列中, 并迭代预测单词, 直到模型输出特殊标记 [SEP]. 对于理解任务, 我们将多模态编码器输出的第一个隐藏状态作为输入给额外的线性分类器, 以预测答案. 理解任务要求模

型判断图像和句子之间的关系, 这是一个封闭的任务, 模型需要从预定义的集合中选择答案. 生成任务更加困难, 因为模型需要生成开放式答案.

表 8 不同数据量和不同因果掩码矩阵比例实验结果

图像-文本数据对数量 (M)	因果掩码矩阵比例	生成任务				理解任务 (%)			
		图像字幕生成				视觉蕴含	图文检索	文图检索	图像分类
		B1	B4	R	C	Acc	Acc	Acc	Acc
0.75	0.0	15.7	3.5	10.1	11.7	50.9	56.4	43.3	64.3
	0.33	50.7	20.2	35.8	68.5	49.6	52.1	41.0	59.7
	0.66	58.7	23.4	37.5	78.4	47.5	51.8	39.8	66.9
	1.0	66.9	24.8	38.7	84.9	33.3	41.9	27.7	47.1
1.5	0.0	24.9	5.3	16.8	18.2	73.5	82.6	70.4	85.4
	0.33	59.8	22.9	38.0	73.7	72.4	79.4	67.9	79.9
	0.66	65.1	24.8	38.9	82.4	72.2	80.8	69.4	85.1
	1.0	69.4	26.0	39.4	89.7	61.9	70.2	61.3	61.7
3.4	0.5	**96.1**	**35.6**	**67.0**	**116.8**	**78.1**	**94.3**	**82.8**	**92.8**

如表 8 所示, 随着预训练中因果掩码矩阵的增加, 该模型在生成性任务中表现更好. 然而, 随着因果掩码的增加, 双向注意掩码的减少, 模型在理解任务方面表现更差. 我们发现, 一般来说, 因果掩码有利于生成性任务, 双向注意掩码有利于理解性任务, 而训练数据的增加对理解性任务和生成性任务都有更大的好处.

我们使用 BERT 标记器的 [UNUSED] 标记作为可学习上下文的参数化模版的组成元素. 对于视觉问答和视觉蕴含, 输入文本包含一个句子、一个提示模版和 [MASK] 标记, 并且提示模版可以在输入文本的开始处, 或者在输入文本的中间. 值得注意的是, 提示不应该出现在输入的末尾, 因为我们使用了因果掩码, [MASK] 标记不能处理右侧的标记. 对于图像分类, 输入文本仅包含 [CTX] 和 [MASK], 提示模版应该在 [MASK] 的左侧. 如表 9 所示, 我们发现只有一个 [CTX] 无法有效地提示模型, 并且随着提示长度的增加, 不同下游任务的准确性可以得到提升, 而过长的提示模版效率较低, 与 16 的长度相比, 32 的长度增加了一倍, 但下游任务的准确性几乎保持不变. 对于视觉问答和视觉蕴含, 提示模版放在中间优于放在开始, 因为中间的提示模版更接近 [MASK] 标记, 对于后面文本的生成是一个更有效的信号.

表 9 不同提示模版长度和位置实验结果

Task	Position	模版长度				
		1	4	8	16	32
VQA	begin	66.4	69.2	70.4	70.8	71.0
VQA	mid	67.9	69.5	70.4	71.0	71.1
VE	begin	77.3	78.2	78.8	79.4	79.9
VE	mid	77.5	78.5	78.6	80.1	80.1
IC (Food101)	begin	90.6	91.4	91.9	92.1	92.5
IC (Flowers102)	begin	93.0	93.6	94.2	94.4	94.0

4 总 结

在本文中, 我们提出了统一的多模态视觉语言理解和生成预训练, 它可以处理视觉语言理解和生成任务. 实验表明, 我们提出的方法在理解和生成任务上都达到了与当前视觉语言方法相当的性能. 我们还提出了基于提示的方法, 这是一种简单有效的方法, 可以对不同的下游任务进行微调.

References:

[1] Lu JS, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In:

Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 2.

[2] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–16.

[3] Lu JS, Goswami V, Rohrbach M, Parikh D, Lee S. 12-in-1: Multi-task vision and language representation learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10437–10446. [doi: 10.1109/CVPR42600.2020.01045]

[4] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: Universal image-text representation learning. arXiv:1909.11740, 2020.

[5] Li XJ, Yin X, Li CY, Zhang PC, Hu XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi Y, Gao JF. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 121–137. [doi: 10.1007/978-3-030-58577-8_8]

[6] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[7] Kim W, Son B, Kim I. ViLT: Vision-and-language transformer without convolution or region supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 5583–5594.

[8] Gan Z, Chen YC, Li LJ, Zhu C, Cheng Y, Liu JJ. Large-scale adversarial training for vision-and-language representation learning. In: Proc. of the 34th Advances in Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 6616–6628.

[9] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[10] Dong L, Yang N, Wang WH, Wei FR, Liu XD, Wang Y, Gao JF, Zhou M, Hon HW. Unified language model pre-training for natural language understanding and generation. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1170.

[11] Bao HB, Dong L, Wei FR, Wang WH, Yang N, Liu XD, Wang Y, Gao JF, Piao SH, Zhou M, Hon HW. UniLMv2: Pseudo-masked language models for unified language model pre-training. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 642–652.

[12] Petroni F, Rocktäschel T, Riedel S, Lewis PSH, Bakhtin A, Wu YX, Miller AH. Language models as knowledge bases? In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 2463–2473.

[13] Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: Proc. of the 34th Advances in Neural Information Processing Systems. Curran Associates Inc., 2020. 1877–1901.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–21.

[15] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3128–3137. [doi: 10.1109/CVPR.2015.7298932]

[16] Plummer BA, Wang LW, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2641–2649. [doi: 10.1109/ICCV.2015.303]

[17] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[18] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6904–6913. [doi: 10.1109/CVPR.2017.670]

[19] Bossard L, Guillaumin M, van Gool L. Food-101–Mining discriminative components with random forests. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 446–461. [doi: 10.1007/978-3-319-10599-4_29]

[20] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. In: Proc. of the 6th Indian Conf. on Computer Vision, Graphics & Image Processing. Bhubaneswar: IEEE, 2008. 722–729. [doi: 10.1109/ICVGIP.2008.47]

[21] Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 3606–3613. [doi: 10.1109/CVPR.2014.461]

[22] Xie N, Lai F, Doran D, Kadav A. Visual entailment: A novel task for fine-grained image understanding. arXiv:1901.06706, 2019.

[23] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[24] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.

[25] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[26] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

[27] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proc. of the 9th Workshop on Statistical Machine Translation. Baltimore: ACL, 2014. 376–380. [doi: 10.3115/v1/W14-3348]

[28] Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic propositional image caption evaluation. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 382–398. [doi: 10.1007/978-3-319-46454-1_24]

刘天义(1998－), 男, 硕士生, 主要研究领域为计算机视觉.

陈静静(1990－), 女, 博士, 副研究员, CCF 专业会员, 主要研究领域为多媒体内容分析, 计算机视觉, 鲁棒可信人工智能.

吴祖煊(1991－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为计算机视觉, 深度学习.

姜育刚(1981－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体信息处理, 计算机视觉, 鲁棒可信人工智能.