arXiv:2301.04366v1 [cs.CL] 11 Jan 2023

用于基于知识的视觉问答的多模态逆向完形填空任务*

Paul Lerner[1],Olivier Ferret[2], 和 Camille Guinaudeau[1]

[1] 巴黎萨克雷大学，法国国家科学研究中心，LISN，91400，奥赛，法国{paul.lerner,camille.guinaudeau}@lisn.upsaclay.fr[2] 巴黎萨克雷大学，法国原子能和替代能源委员会，List，F-91120，帕莱索，法国 olivier.ferret@cea.fr

摘要。 我们提出了一种新的预训练方法——多模态逆向完形填空任务，用于针对命名实体的基于知识的视觉问答（KVQAE）。KVQAE 是一项近期提出的任务，要求利用知识库，结合视觉语境回答有关命名实体的问题。因此，模态之间的交互对于检索信息至关重要，必须通过复杂的融合模型来捕捉。鉴于这些模型需要大量训练数据，我们基于文本问答领域已有的研究设计了这一预训练任务。该任务将一个句子视为伪问题，将其语境视为伪相关段落，并进一步考虑多模态文档中文本附近的图像。我们的方法适用于不同的神经网络架构；与不使用预训练的基线相比，它使检索和阅读理解的相对平均倒数排名分别提升 9%，相对 F1 分别提升 15%。

关键词： 视觉问答 · 预训练 · 多模态融合。

1 引言

关于命名实体的基于知识的视觉问答（KVQAE）是一项近期在 [50]中提出的挑战性任务。该任务旨在利用知识库（KB）回答以视觉上下文为依据、有关命名实体的问题。图 1 展示了两个视觉问题及其在知识库中相关的视觉段落。因此，要完成这项任务，必须从知识库中检索 相关信息。这与标准视觉问答（VQA，[1]）不同，后者的问题针对图像内容（例如物体的颜色或物体数量）；它也不同于基于知识的视觉问答（针对粗粒度物体类别）[40], ，后者可以依赖现成的物体检测方法 [17]。在

* 本研究得到 ANR-19-CE23-0028MEERQAT 项目的资助。本研究获准使用 IDRIS 的高性能计算资源，所用资源由 GENCI 通过 2021年-AD011012846 配额提供。

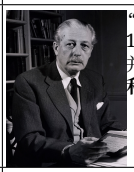
视觉问题（输入）	知识库中的相关视觉段落
 *“Which constituency did this man represent when he was Prime Minister?”*	 “麦克米伦确实在压倒性胜利中失去了斯托克顿1945年工党胜选，但后来重返议会——并于1945年11月在**布罗姆**补选中当选**利**。”
 *“In which year did this ocean liner make her maiden voyage?”*	 “伊丽莎白女王二号，通常简称为QE2，是一艘浮动酒店，也是一艘退役的远洋客轮由冠达邮轮公司建造，冠达邮轮公司由冠达邮轮运营，既是一艘跨大西洋客轮，也是一艘邮轮这艘船于**1969** 年至2008年间建造。”

图1：ViQuAE 数据集中有关命名实体的视觉问题示例，以及其知识库中的相关视觉段落 [32]。

KVQAE 中，文本和图像模态都能提供有用信息，且必须将二者结合。因此，这项任务更广泛地属于多模态信息检索 （IR）和多模态融合的范畴。

多模态信息检索以及更广泛意义上的多模态学习有两种范式：早期融合（数据级和特征级）与后期融合（分数级和决策级） [30]。一方面，后期融合更直接，因为自然语言处理和计算机视觉技术可以独立应用，但它忽略了模态之间的交互。另一方面，早期融合的丰富表现力往往以增加复杂性和模型参数为代价。这给 KVQAE 带来了额外挑战，因为现有的两个数据集要么规模较小（ViQuAE [32]），要么是自动生成的（KVQA [50]）。为应对这一挑战，我们提出使用多模态逆向完形填空任务（ICT）对融合模型进行预训练。ICT 最初在 [31] 中提出，用于预训练文本问答（QA）的神经检索器。该任务将一个句子视为伪问题，并将其上下文视为伪相关段落。它可以看作是跳字目标函数 [41] 的推广。我们通过考虑多模态文档中文本附近的图像对其进行了扩展。

我们的主要贡献包括：(i) 多模态逆向完形填空任务，这是一种新的预训练方法，可用于应对 ViQuAE 等规模较小的 KVQAE 数据集；(ii) 一个面向 KVQAE 的多模态信息检索框架；(iii) 对不同神经网络架构进行实验，其中包括近期提出的多模态 BERT。

2 相关工作

稠密检索。 稠密检索是一个快速发展的领域，[36,11],对其进行了综述，其中介绍了新兴的预训练任务、优化方法以及 Transformer 架构的变体 [47,23,15,14]。[31]率先在问答任务中使用稠密表示，超越了 BM25 等稀疏词袋表示。他们的方法依赖于三个组成部分：(i) BERT 等预训练语言模 [10],型，它们能够将句子的语义编码到一个

稠密向量中；(ii) 一种对比学习目标，用于优化问题与文本段落嵌入之间的相似度（见第 3 节）；(iii) 一项无监督训练任务，即反向完形填空任务（ICT，见第 1 节）。 [27] 批评后一种方法计算开销过大[3]，并认为普通句子并不是问题的良好替代物。相反，他们提出了 DPR，该方法利用：(i) 判断段落是否包含问题答案这一启发式规则来确定其相关性；(ii) 使用 BM25 等无监督信息检索方法挖掘困难负样本，这被证明是该方法成功的关键。我们旨在结合这两种方法的优势：(i) 像 DPR 一样，在文本问答数据集上预训练模型；(ii) 将 [31] 的 ICT 改造为适用于多模态文档，从而将多模态能力融入这个有望获得良好初始化的模型。

多模态融合与预训练。 BERT 在自然语言处理 [10],领域的成功，依托于易于并行化的 Transformer 架构 [58],、无监督训练目标以及与任务无关的架构，同时启发了视觉问答和跨模态检索领域的许多研究 [57,38,35,56,34,7]。这些模型在 [5] 中被统一到一个框架下，并在 [28]中得到部分综述。所有这些模型都依赖 Transformer 架构，通常以预训练的 BERT 初始化，以融合图像和文本。训练采用弱监督方式，基于 COCO [37] 或 Conceptual Captions [51], 等图像描述数据集，以及 Faster R-CNN [48]等预训练目标检测器。[22] 表明，这些模型能够为视觉问答学习到模态之间非平凡的交互。多模态 BERT 大致可分为单流和多流两类。单流模型将文本词元的嵌入和图像区域的嵌入输入同一个 Transformer模型，依靠自注意力机制将二者融合。相比之下，在多流架构中，文本和图像先分别由两个独立的 Transformer 处理，再通过交叉注意力融合不同模态。研究表明，这两种架构的表现同样出色 [5]。在本研究中，我们采用单流模型，以利用在纯文本数据（问答数据集上）的预训练。还需注意，尽管本研究受到这些工作的启发，但我们没有使用相同的训练目标或数据，因为正如下一节所述，这些目标或数据可能并不适用于命名实体的表示。

多模态信息检索与 KVQAE。 多模态信息检索主要采用后期融合技术来处理（综述见 [9] ），但本文主要关注早期融合技术。

[9] 回顾了早期融合的首次尝试。随后，人们通过将两种模态的特征串联成一个向量，系统地实现了早期融合，并重点关注特征加权方案。串联方法面临维数灾难，因为由此得到的特征空间维度等于各模态特征维度之和。

[44]和 [39]同时提出了一种与我们非常相似的基于知识的视觉问答方法。他们通过替换问题编码器，对 DPR [27] 进行了改造

[3] [31] 使用批次大小超过 4K 的问题。

将其替换为 LXMERT [57], , 从而能够融合问题和图像。然而，与我们不同的是，他们仍使用基于纯文本的段落编码器，并采用与 [57], 相同的预训练目标，即掩码语言建模、掩码区域建模和图文匹配。我们认为，这些目标适合学习粗粒度物体类别的表示，但不适合学习命名实体的表示。换言之，它们适用于标准视觉问答，却不适用于 KVQAE。例如，掩码区域建模依赖目标检测器，因此不适用于 KVQAE。尽管 [44] 和 [39] 都在 OK-VQA [40],上进行了实验，但结果并不一致： [44] 表明，他们的模型与 BM25 基线具有竞争力，该基线以问题和图像的人工撰写的 说明文字为输入；而 [39] 的模型则不及采用自动生成的 说明文字的 BM25。两项研究的结果之所以存在差异，是因为它们使用的 KB 和评估指标均不相同。[19]还尝试了不同的多模态 BERT，但不采用段落级标注，而是端到端地训练检索器和答案分类器[4]。

尽管他们在 KVQA 上进行了实验， [50], 但我们不考虑 [52,16,21] 的工作，因为他们的系统以人工撰写的 说明文字作为输入，这使得图像内容的作用变得不明确。 [50] 采用决策层面的后期融合方法。首先，他们检测并消歧问题中的命名实体。随后，由于他们的数据集 KVQA 仅包含关于人物命名实体的问题，因此会执行人脸识别步骤。从维基数据中检索文本和视觉检测到的实体的事实[5] ，并由记忆网络进行处理 [60]。相比之下，我们的工作与 [32], 一致，他们使用来自维基百科的非结构化文本作为知识库。不同于采用后期融合方法、分别检索问题和图像的 [32], ，我们旨在对文本和图像进行统一表示，既涵盖视觉问题，也涵盖知识库。

3 方法

本节首先对我们的 KVQAE 框架进行形式化描述，接着介绍模型，然后深入讲解三个训练阶段：(i) 用于文本问答的 DPR；(ii) 多模态逆向完形填空任务，这是我们的主要贡献；(iii) KVQAE 的微调。最后，我们讨论推理机制和实现细节。

3.1 信息检索框架

在我们的多模态设置中，视觉问题（来自数据集）和视觉段落（来自知识库）都由文本-图像对 $(t,\ i)$ 组成，如图 1 所示。我们的目标是找到最优模型 E ，以编码出充分的表示 $\mathbf{q} = E(t_q,\ i_q)$ 和 $\mathbf{p} = E(t_p,\ i_p)$，使得当 $(t_p,\ i_p)$ 与 $(t_q,\ i_q)$ 相关时，它们彼此接近（记作

[4] 标准的（基于知识的）视觉问答通常被视为分类任务。 [5] https://www.wikidata.org/

上标 (+) 和 (−)）。因此，搜索归结为检索与视觉问题最接近的 K 个视觉段落。在计算两个向量之间的相似度时，此处采用点积；在所有训练阶段（§3.3）中使用的目标，都是针对数据集中的所有视觉问题 [31,27]，最小化以下负对数似然损失：$-\log \frac{\exp(\mathbf{q}\cdot\mathbf{p}^+)}{\exp(\mathbf{q}\cdot\mathbf{p}^+)+\sum_j \exp(\mathbf{q}\cdot\mathbf{p}_j^-)}$。

这一对比目标能够高效地利用批次中与其他问题相关的段落作为批内负样本，因为与整个模型的前向传播相比，计算两个向量之间的相似度开销相当小。下一节将根据融合机制介绍两种不同的模型 E。

3.2 模型

我们的所有模型都利用 BERT[6] [10] 和 CLIP[7] [46] 作为构建模块，分别表示文本和图像。BERT 在维基百科和 BooksCorpus [62] 上通过掩码语言建模和下一句预测进行训练。CLIP 则通过对比目标，在 4 亿组图像和文本配对上以弱监督方式进行训练。正如 [32] 中的实验证明，与完全监督模型相比，它展现出更强的泛化能力，并且适用于 KVQAE。我们尝试了两种不同的融合技术：ECA 和 ILF。

早期交叉注意力融合（ECA）通过上述多模态 BERT 所采用的单流 Transformer模型实现（例如 UNITER [7]）。然而，我们没有依赖 Faster R-CNN 这类固定的目标检测器，而是按照前文的动机，利用 CLIP。为了实现早期融合，CLIP 生成的视觉嵌入会通过一个线性层投影到与文本相同的空间中，该线性层包含从头开始训练的 $\mathbf{W}_c \in \mathbb{R}^{c\times d}$ 个参数：$\mathbf{e}_c = \text{CLIP}(i)\cdot\mathbf{W}_c$。然后，所得嵌入会沿序列维度与词嵌入拼接，充当一个“视觉标记”。接着，这些嵌入被输入 Transformer模型，其中注意力机制应能促成不同模态之间的交互。最终嵌入对应于特殊的[CLS] 标记：$\text{ECA}(t,i) = \text{BERT}([t;\mathbf{e}_c])_{\texttt{[CLS]}}$。Transformer模型首先通过 BERT 初始化。

中间线性融合（ILF）引入额外的 $\mathbf{W}_t \in \mathbb{R}^{d\times d}$ 个参数，这些参数从头开始训练，用于将[CLS] 标记的表示简单投影到与 CLIP 嵌入相同的空间中，然后将二者相加：$\text{ILF}(t,i) = \text{BERT}(t)_{\texttt{[CLS]}}\cdot\mathbf{W}_t + \mathbf{e}_c$。

由于 ECA 和 ILF 都会生成多模态表示$\mathbf{q}$ 和 $\mathbf{p}$，因此直接根据它们的相似度$\mathbf{q}\cdot\mathbf{p}$进行排序。作为基线，我们遵循 [32]，在进行零均值和单位方差

[6] 可在以下网址获取不区分大小写的“base”12层版本：https://huggingface.co。[7] 采用 ResNet-50 骨干网络 [20]。[8] 请注意，这等同于先将两者拼接，再像 $[\text{BERT}(t)_{\texttt{[CLS]}};\ \text{CLIP}(i)]\cdot[\mathbf{W}_t;\ \mathbf{W}_c]$。

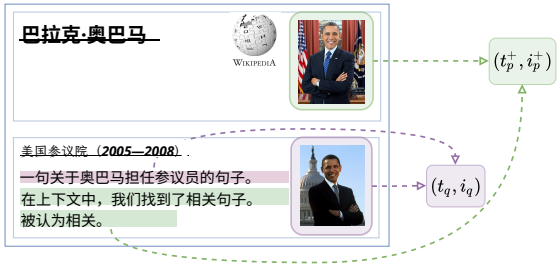


图 2：通过维基百科/WIT 实现多模态逆向完形填空任务概览。

归一化后，对文本和图像的相似度进行线性组合（以下公式中省略）：

$$\alpha \times \text{BERT}(t_q)_{\texttt{[CLS]}} \cdot \text{BERT}(t_p)_{\texttt{[CLS]}} + (1-\alpha) \times \cos(\text{CLIP}(i_q), \text{CLIP}(i_p)) \quad (1)$$

使用网格搜索在验证集上优化插值超参数 α，以最大化平均倒数排名。本文其余部分将左侧项（文本相似度）称为 DPR。

3.3 训练阶段

模型按顺序分三个阶段进行训练：

阶段 1：用于文本问答的 DPR。 先不考虑视觉表征，从 BERT 初始化 [27] 开始训练 DPR 模型。DPR 由两个 BERT 编码器组成：一个用于问题 t_q，另一个用于文本段落 t_p。我们采用 [32] 在 TriviaQA 上预训练的模型，并剔除了其数据集 ViQuAE 中使用过的所有问题。他们使用 KILT [43] 版本的 TriviaQA 和维基百科，后者在此阶段用作知识库。随后，将每篇文章拆分为互不重叠的 100 词段落，用于文本检索，同时保留句子边界，并将文章标题附加到每个段落的开头。由此得到 3200 万个段落，即每篇文章 ≈ 5.4 个段落。按照 [27]，使用 BM25 [49] 挖掘无关段落（即困难负样本）。

阶段 2：多模态逆向完形填空任务。 这是本文的主要贡献。我们提出将 [31] 的 ICT 扩展到多模态文档。ICT 的做法是将一个句子视为伪问题 t_q，并将其上下文视为伪相关段落 t_p^+。请注意，文章标题会附加到每个段落的开头 t_p（与阶段 1 中一样）。我们利用维基百科段落中的上下文图像来构造伪问题，并使用 信息框 图像来构造段落，从而扩展该方法（见图 2）。 [31] 通过实验证明，他们的方法取得成功的一个关键因素，是在 10% 的训练样本中将伪问题保留在相关段落中，使模型学会进行词语匹配，因为词汇重叠最终是检索中一个非常有用的特征。在我们的情形中，

然而，我们认为这样做既无必要——因为模型应通过在 TriviaQA 上进行阶段 1 训练而获得强有力的初始化——也无益处，因为模型可能因此忽略图像模态。我们使用在阶段 1 预训练的问题编码器和段落编码器，分别初始化视觉问题编码器和视觉段落编码器。

得益于 WIT 数据集 [54]，这一过程得以简化。WIT 包含数百万张图像及其对应的文本，这些内容来自维基百科和维基共享资源，涵盖 108 种不同语言。不过，本研究只关注英语。虽然 [54]有多种策略可以查找与特定维基百科图像相关的文本，例如其维基共享资源说明文字，但为了模拟下游 KVQAE 场景，我们仅将上下文段落作为文本来源。由此得到的 WIT 英语子集包含 40 万个信息框 图像／文章，对应 120 万个段落／图像。这 120 万个段落共包含 1360 万个句子，即潜在的伪问题，平均每个句子长 26 个词。因此，为了尽可能贴近阶段 1 和阶段 3（其中段落最长为 100个词 ），每个段落由四 个句子组成。这与 [31]略有不同，后者在伪问题遮蔽之前，考虑的是最长为 288 个词片段的段落。

由于 ViQuAE 和 WIT 的图像均取自 Wikimedia Commons[9]，，我们可以根据图像网址估计，ViQuAE 中有 14% 的图像与 WIT 重叠。这可能会导致偏差，我们将在第 4.1 节进行分析。

受 [2]，启发，为了防止灾难性遗忘 并确保实体的模态不变 表示，在此阶段冻结 BERT 的最后 l 层。这样，我们只调整 ECA 的前几层模态专属层，其思路是用图像中存在的“视觉”实体，替换阶段 1 学到的文本命名实体。ILF 在此阶段完全冻结 BERT，仅依靠$\mathbf{W}_t$ 参数来调整文本表示。此外，在所有阶段中，CLIP 都始终处于冻结状态。

在此阶段，我们没有直接的方法来挖掘无关的视觉段落。在早期实验中，我们尝试通过对批次中的图像进行置换来合成这些段落：$(t_p^+,\ i_p^+) \leftarrow (t_p^+,\ i_p^-)$，但这并未改善结果。

过滤掉损坏的图像、格式不合适的图像（例如 .svg）以及只有一个句子的段落后，我们最终得到 975K 个段落/图像。本文其余部分将其称为 WIT。它被划分为训练集（878K）、验证集（48K，用于调整超参数）和测试集（48K，用于合理性检查），各子集之间的文章没有重叠。

阶段 3：关于命名实体的基于知识的视觉问答。这一阶段包括在下游 KVQAE 数据集上对模型进行微调，该数据集提供视觉问题（$t_q,\ i_q$）和相关视觉段落（$t_p^+,\ i_p^+$）。按照 [2]，，这一阶段会调整模型的所有层。

这一阶段的一个细节是如何选择不相关的视觉段落（$t_p^-,\ i_p^-$）。如第 2 节所述，这已被证明对 DPR [27]，至关重要，而且通常对对比学习也很重要 [26]。在 [32]，中，不相关段落

[9] https://维基共享资源/

通过 BM25 挖掘，用于训练 DPR。然而，我们认为这对 ECA 和 ILF 来说并非最优，因为 BM25 只能挖掘在文本上看似合理的段落，而无法挖掘在视觉上看似合理的段落。因此，我们使用 [32] 提供的系统来挖掘不相关段落。该系统将 DPR、ArcFace [8], CLIP 和 ImageNet-ResNet [20]进行后期融合。这导致 DPR（作为基线使用）与我们的模型采用了不同的训练设置。不过，我们也对 DPR 进行了两种设置的实验，发现差异并不显著[10]。

我们使用与 [32]，相同的知识库，该知识库基于 KILT 的维基百科和维基数据中对应实体的图像。它包含 150 万篇文章（因此也包含 150 万张图像/实体），并按照阶段 1 的方式划分为 1200 万个段落，每个段落最多 100 个词。

ViQuAE 中的视觉问题划分为训练集（1,190 个）、验证集（1,250 个）和测试集（1,257 个），各集合中的图像 URL 不重叠 [32]。我们没有在 KVQA [50] 上进行实验，原因如下：（i）该数据集由维基数据自动生成，因此基于文本的知识库对答案的覆盖率较低；（ii）其中包含是/否问题，无法自动评估段落相关性。

3.4 推理

为实现高效检索，视觉段落编码器会预先将知识库中的每个段落与其对应图像一同嵌入。给定一个以图像为依据的问题，视觉问题编码器会对二者进行嵌入。随后使用 Faiss [25] 通过最大内积搜索执行检索。

3.5 实现细节

我们的代码基于 PyTorch [42], Hugging Face 的 Transformer [61] 和 Datasets [33]（后者本身封装了 Faiss）构建。代码、数据和训练好的模型[11]均可免费获取。

训练 ECA 时，我们采用与 [32] 用于 DPR 的相同超参数；这些超参数本身基于 [27]。具体而言，我们使用学习率 2×10^{-5} 和 Adam优化器 [29]。学习率按线性方式调整，阶段 2 和阶段 3 分别使用 100 和 4 个预热步数。不过，对于 ILF，我们根据验证集发现，在阶段 2 使用学习率 2×10^{-3} 和常数调度器能更快收敛。我们认为这是因为 ILF 在阶段 2 完全冻结了 BERT，因此无需仔细调整学习率计划，也无需使用较小的学习率。Dropout [55] 应用于 BERT 中，以及将嵌入投影至$\mathbf{W}_c$ 和$\mathbf{W}_t$ 之后，概率为 0.1（与标准 BERT 配置相同）。同样，BERT 中以及 ILF 中将两个嵌入相加之后，均会应用层归一化 [3] 。梯度范数裁剪至 2。

阶段 2 和阶段 3 的模型训练时，批次大小分别为 512 和 298 个视觉问题。对比学习的成功部分依赖于大量批内负样本，因此也需要较大的批次大小 [45]。我们发现

[10] 评估方法详见第4节[11]

[11] https://github.com/PaulLerner/ViQuAE

表 1：ViQuAE 上的信息检索评估。l：多模态逆向完形填空任务期间冻结层的数量。上标表示在 Fisher随机化检验中与 $p \leq 0.01$ 存在显著差异。Hits@1 未列出，因为它与 P@1 等价。

# 模型	多模态逆向完形填空任务 平均倒数排名@100	P@1		P@20	Hits@20
a DPR	NA	32.8	22.8	16.4	61.2
b DPR +CLIP	NA	34.5^{a}	24.8^{a}	15.8	61.8
c ECA	✗	34.6	25.9^{a}	17.2^{ab}	61.6
d ECA ($l = 6$)	✓	**37.8**^{abce}	26.7^{a}	**19.5**^{abce}	**67.6**^{abce}
e ECA ($l = 0$)	✓	35.1	24.7	17.6^{b}	63.7
f ILF ($l = 12$)	✓	37.3^{a}	**26.8**^{a}	19.1^{abce}	66.9^{abc}

梯度检查点 [6] 使 ECA[12]能够使用大得多的批次大小。与 [32] 使用四 块各配备 32GB 内存的 NVIDIA V100 GPU 、总批次大小为 128 个问题不同，我们能够在一块单独的 V100 GPU 上容纳 298 个问题（如上所述）。阶段 2 占用大部分计算预算，大多数模型在 $\approx$ 8K 步后收敛，耗时约三天[13]。所有阶段的检查点均依据验证集批内 平均倒数排名进行选择。批内的含义是，只对批次中的其他视觉段落进行排序，并且每个视觉问题只与一个相关视觉段落配对（与训练期间相同）。

4 结果

检索模型通过两种不同方式进行评估：(i) 计算视觉段落检索的标准信息检索指标；(ii) 将检索到的视觉段落输入阅读器模块，由其提取问题的简洁答案，从而实现 KVQAE。换言之，要么评估系统能否为问题检索到相关段落，要么评估系统能否回答问题。我们发现这两种指标之间存在相关性。消融研究采用信息检索指标进行。

ViQuAE 基于 TriviaQA，因此仅接受远程监督：如果答案与真实标注的字符串匹配，就视为正确；同样，如果段落包含真实标注[14]，就视为相关。此外，真实标注在维基百科中的别名也被视为有效答案。

4.1 信息检索

由于 ViQuAE 的设置，无法完整覆盖相关段落。因此，我们不使用任何基于召回率的指标（例如 R-

[12] ILF 无需完全冻结 BERT。[13] 法国的 Jean Zay GPU 功耗为 0.482kW（热回收后为 0.259kW），该国的平均电网排放因子为 0.0569 kgCO2e/kWh，据https: // bilans-ges.ademe.fr/en。[14] 经过标准预处理（转换为小写、去除冠词和标点符号）后。

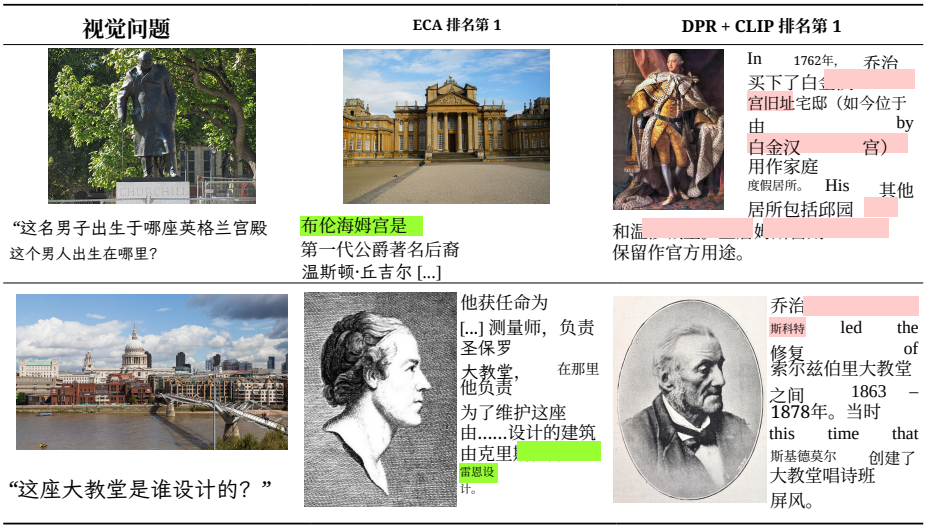


图 3：定性示例，其中 ECA（$l = 6$）在排名第 1 的结果中找到了相关视觉段落，而后期融合则表现不佳。

Precision、mAP 等）。相反，我们使用 Precision@K（P@K）、平均倒数排名（MRR）和 Hits@K 来评估模型。Hits@K 是信息检索在前 K 个结果中检索到至少一个相关段落的问题所占的比例。我们使用 Fisher随机化检验 [12,53]进行统计显著性检验。指标和统计检验均使用 ranx [4] 计算，并报告于表 1 中。

使用多模态逆向完形填空任务预训练的最佳模型（d 和 f）在所有指标上均优于纯文本（a）和后期融合（b）基线。图 3 展示了一些定性示例。在第一行中，我们可以看到，描绘温斯顿·丘吉尔的图像与提及他的段落之间存在跨模态交互（尽管该段落配有一张完全不同的图像）。相比之下，后期融合基线表现出文本偏见，返回了提及多座英国宫殿的段落（以红色突出显示）。第二行也能观察到同样的现象：圣保罗大教堂只在相关段落中被提及，并未出现在上下文图像中。在这种情况下，跨模态交互很有用，因为视觉描绘具有异质性：在视觉问题中，温斯顿·丘吉尔以雕像的形式出现，而在知识库中则以照片的形式出现。

我们可以看到，多模态逆向完形填空任务对 ECA 至关重要（c 与 d）。若没有这一任务，其表现与后期融合相当。我们认为，这是因为 ViQuAE 的小规模训练集导致了过拟合。然而，我们发现，对 ViQuAE 进行微调对 ECA 也至关重要；由于采用了顺序学习设置，ECA 出现了灾难性遗忘：事实上，经过阶段 2 后，其表现落后于 DPR。我们看到， [2] 的冻结技术在一定程度上有助于防止灾难性遗忘（d 与 e）。在上游 WIT 预训练中也能观察到这一点，其中

表2：在 ViQuAE 上进行的阅读理解评估结果，为 阅读器 运行 5 次的平均值。 l：多模态反向完形填空任务期间冻结的层数。

# 信息检索模型	多模态反向完形填空任务	完全匹配	F1
a DPR	NA	16.9 ± 0.4	20.1 ± 0.5
b DPR + CLIP	NA	19.0 ± 0.4	22.3 ± 0.4
c ECA	✗	17.7 ± 0.6	21.2 ± 0.8
d ECA ($l = 6$)	✓	20.6 ± 0.3	24.4 ± 0.2
e ECA ($l = 0$)	✓	20.8 ± 0.8	24.3 ± 0.9
f ILF ($l = 12$)	✓	**21.3** ± 0.6	**25.4** ± 0.3

使用 $l = 6$ 和$l = 0$时，ECA 在 WIT 测试集上的批内平均倒数排名分别为 91.6 和 92.9：对 WIT 的拟合效果越好，遗忘就越严重。

与相关工作（§2）的结论不同，我们发现，线性融合模型的表现与更早期的基于交叉注意力的融合模型不相上下（f 与 d）。这表明，相较于后期融合基线的提升确实来自多模态 ICT 预训练，而且对模型架构并不十分敏感。此外，ILF 的架构允许在阶段 2 完全冻结 BERT，从而避免灾难性遗忘[15]。其他训练策略（例如多任务学习、使用适配器 [24]）留待未来工作探索。

没有证据表明，ECA 在 ViQuAE 中与 WIT 重叠的 14% 图像上表现更好。ECA 在 WIT 之外子集上的表现更佳（MRR 为 38.0，对比 36.5），但 DPR 和后期融合的情况则相反。

4.2 阅读理解

为了从检索到的段落中提取答案，我们沿用 [32]的同一模型。该模型采用多段落 BERT 架构 [59] ，因此仅基于文本，因为一旦检索到相关段落，无需查看图像也能回答问题。为减少训练带来的变化并控制实验数量，我们直接使用 [32] 提供的现成模型，只更换其输入段落。该模型以排名前 24 的段落为输入。模型先在 TriviaQA 上训练（剔除了 ViQuAE 中使用的所有问题），再在 ViQuAE 上微调，与阶段 1 和阶段 3 的做法类似。作者提供了五个不同版本的模型，分别对应不同的随机种子。

我们使用完全匹配和 F1 分数（在词袋层面）评估提取的答案。在表2中，我们可以确认，更相关的段落确实能带来更好的下游答案。与信息检索评估相比，唯一的区别在于 [2] 的冻结技术所起的作用（d 与 e），这一点在此处并不那么明显。

[15] 由于采用冻结策略，ILF 在 WIT 的测试集上批内平均倒数排名仅达到 87.1。

5 通用与专用图像表征

上一节报告的数字实际上与 [32] 的最佳结果不相上下。这是因为后者除了使用 DPR 和 CLIP，还采用了 ArcFace 和 ImageNet-ResNet。具体而言， [32] 使用一种启发式方法来利用 ArcFace 提供的人脸表示：检测到人脸时使用 ArcFace，否则则结合使用 CLIP 和 ImageNet-ResNet。他们表明，这种方法能提高有关人物的问题的检索精度（这类问题与人脸表示相关）。然而，这种方法无法扩展，原因有二：（i）ViQuAE 中有近 1,000 种不同的实体类型（依据维基数据的本体），并非所有类型都能受益于专用表示；（ii）针对同一实体类型组合多种表示（例如 CLIP 和 ImageNet-ResNet）在计算上代价高昂，而且很快就会达到饱和。为了构建一个可与 [32], 的后期融合相媲美的系统，我们尝试在 ECA 中集成 ArcFace 和 ImageNet-ResNet。然而，我们未能胜过仅使用 CLIP 的 ECA 版本。直观而言，我们认为 ECA 会稀释 ArcFace 的专用表示，并且无法在 BERT 的全部十二层中保留这些表示。因此，在这种设置下，ECA 总体上与后期融合不相上下（MRR 为 37.7 对 37.9，差异不显著），但在有关非人物的问题上表现更好（MRR 为 39.3 对 35.7），这再次表明它无法利用 ArcFace 的表示。

6 结论与展望

我们提出了一种新的预训练方法——多模态逆向完形填空任务，用于关于命名实体的基于知识的视觉问答。多模态反向完形填空任务利用多模态文档中的上下文图像生成伪视觉问题。与此前提出的后期融合方法相比，该方法能够使用更复杂的多模态融合模型。因此，在近期推出的 ViQuAE 数据集上，我们的方法使检索准确率提高了 10% 的相对平均倒数排名，并使下游阅读理解（即答案抽取）的相对 F1 值提高了 9%。我们认为，这得益于后期融合所不允许的跨模态交互。更具体地说，我们通过定性观察发现，这些交互发生在视觉问题中的图像与知识库文本之间，从而缓解了视觉描绘的异质性。

我们使用两种不同的神经网络架构对预训练方法进行了实验：（i）ECA，该架构遵循近期提出的多模态 BERT，通过交叉注意力实现模态的早期融合；（ii）ILF，一种更为标准的模型，通过线性投影融合模态。我们发现，两者的表现同样出色，这与标准视觉问答和跨模态检索中的情况不同。我们认为，这可能是因为两者的训练设置不同，导致 ECA 出现灾难性遗忘。不过，仍需进一步研究。

在致力于为命名实体构建通用多模态表示时，我们发现，在模型中整合专门的表示（例如用于人脸的 ArcFace）

并无益处。我们推测，所研究的架构可能并不适合，但将这一问题留待未来研究探讨。

对于未来的工作，我们认为，将多模态逆向完形填空任务推广到重排序（同时处理 $(t_q,\ i_q)$ 和 $(t_p,\ i_p)$）以及阅读理解（从 $(t_p,\ i_p)$ 中生成或提取答案）是一个令人振奋的研究方向。事实上，有证据表明，对文本问答而言，共享同一模型进行信息检索和阅读理解有益；对跨模态检索而言，共享同一模型进行信息检索和重排序也有益：这两项任务都与 KVQAE 密切相关。 [13] [18],

参考文献

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: 视觉问答：视觉问答。在：2015年IEEE国际计算机视觉会议（ICCV）。第2425–2433页。IEEE，圣地亚哥，智利（2015年12月）。https://doi.org/10.1109/ICCV.2015.279, http://ieeexplore.ieee.org/document/7410636/ 2. Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A.: 跨模态场景网络。《IEEE模式分析与机器智能汇刊》40(10), 2303–2314（2017），出版商：IEEE 3. Ba, J.L., Kiros, J.R., Hinton, G.E.: 层归一化。arXiv:1607.06450 [cs, stat] 2016 1607 06450, 1607 06450 (7月), 4 http://arxiv.org/abs/ 。arXiv: 。 。 Bassani, E.: ranx：用于排序评估与比较的极速Python库。在：Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørv˚ag, K., Setty, V.（编）信息检索进展。第259–264页。计算机科学讲义，Springer国际出版，Cham（2022）。https://doi.org/10 7 30 5 1007 978 3 030 99739 。 / - - - - . Bugliarello, E., Cotterell, R., Okazaki, N., Elliott, D.: 揭开多模态预训练的面纱：视觉与语言BERT的元分析与统一框架。《计算语言学协会会刊》9，978–994（2021年9月）。https://doi.org/10.1162/tacl a 00408, https://doi.org/10.1162/tacl a 004086. Chen, T., Xu, B., Zhang, C., Guestrin, C.: 以次线性内存开销训练深度网络（2016年4月）。https://doi.org/10.48550/arXiv.1604.06174, http: //arxiv.org/abs/ 1604 06174, 1604 06174 1604 06174 [ ] 。编号：7 arXiv: 。arXiv: 。cs。Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER：通用图文表征学习。在：欧洲计算机视觉会议。第104–120页。Springer（2020）8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace：用于深度人脸识别的加性角度间隔损失。在：IEEE/CVF计算机视觉与模式识别会议论文集（CVPR）（2019年6月），https: //openaccess.thecvf.com/content CVPR 2019/html/Deng ArcFace Additive Angular Margin Loss for Deep Face Recognition CVPR 2019 paper.html 9. Depeursinge, A., M¨uller, H.: 融合技术：结合文本与视觉信息检索。在：M¨uller, H., Clough, P., Deselaers, T., Ca- puto, B.（编）ImageCLEF：视觉信息检索实验评估，第95–114页。信息检索系列，Springer，柏林，海德堡（2010）。https://doi.org/10.1007/978-3-642-15181-1 6, https://doi.org/10.1007/978-3-642-15181-1 6

10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT：用于语言理解的深度双向变换器预训练。在：2019年计算语言学协会北美分会会议论文集：人类语言技术，第1卷（长篇与短篇论文）。第4171–4186页。计算语言学协会，明尼阿波利斯，明尼苏达州（2019年6月）。https://doi.org/10.18653/v1/N19-1423，https://aclanthology.org/N19 1423 11 - 。Fan, Y., Xie, X., Cai, Y., Chen, J., Ma, X., Li, X., Zhang, R., Guo, J., Liu, Y.: 信息检索中的预训练方法。arXiv:2111.13853 [cs]（2021年11月），http://arxiv.org/abs/2111 13853 12 13853，2111。arXiv: 。 。Fisher, R.A.: 实验设计。《实验设计》。（第2版）（1937年），https://www.cabdirect.org/cabdirect/abstract/19371601600，出版商：Oliver & Boyd，爱丁堡与伦敦。13. Fun, H., Gandhi, S., Ravi, S.: 高效的检索优化多任务学习。arXiv:2104.10129 [cs]（2021年4月），http://arxiv.org/abs/2104.10129，arXiv: 2104.10129 14. Gao, L., Callan, J.: Condenser：面向稠密检索的预训练架构。在：2021年自然语言处理实证方法会议论文集。第981–993页。计算语言学协会，线上及蓬塔卡纳，多米尼加共和国（2021年11月），https://aclanthology.org/2021.emnlp-main.75 15. Gao, L., Callan, J.: 面向稠密段落检索的无监督语料库感知语言模型预训练。在：计算语言学协会第60届年会论文集（第1卷：长篇论文）。第2843–2853页。计算语言学协会，都柏林，爱尔兰（2022年5月）。https://doi.org/10.18653/v1/2022.acl-long.203，https://aclanthology.org/ 2022 203 16 .acl-long. 。Garcia-Olano, D., Onoe, Y., Ghosh, J.: 通过实体增强的知识注入改进并诊断基于知识的视觉问答。在：2022年万维网大会伴随论文集。第705–715页。WWW '22，计算机协会，纽约，纽约州，美国（2022年）。https://doi.org/10.1145/3487553.3524648，https://doi.org/10 1145 3487553 3524648 。/。。Garderes, F., Ziaeefard, M.: 17 ConceptBert：用于视觉问答的概念感知表征。《计算语言学协会论文集：EMNLP 2020》10页（2020），https://aclanthology.org/2020.findings-emnlp.44/ 18. Geigle, G., Pfeiffer, J., Reimers, N., Vuli´c, I., Gurevych, I.: 快速检索，智能重排序：改进跨模态检索的协作式联合方法。《计算语言学协会会刊》10，503 05 2022 521 10 1162 00473，10 1162 – （）。https://doi.org/ . /tacl a https://doi.org/ . /tacl a 00473 19. Guo, Y., Nie, L., Wong, Y., Liu, Y., Cheng, Z., Kankanhalli, M.: 面向基于知识的视觉问答的统一端到端检索器—阅读器框架（2022年6月），http: //arxiv.org/abs/2206 14989，2206 14989 2206 14989 [ ]。编20 号：arXiv: 。arXiv: 。cs。He, K., Zhang, X., Ren, S., Sun, J.: 用于图像识别的深度残差学习。在：IEEE计算机视觉与模式识别会议论文集。第770–778页（2016），https://openaccess.thecvf.com/content cvpr 2016/papers/He Deep Residual Learning CVPR 2016 paper.pdf 21. Heo, Y.J., Kim, E.S., Choi, W.S., Zhang, B.T.: 超图变换器：用于基于知识的视觉问答的弱监督多跳推理。在：计算语言学协会第60届年会论文集

《语言学》（第1卷：长篇论文）。第373–390页。计算语言学协会，都柏林，爱尔兰（2022年5月）。https://doi.org/10.18653/v1/2022.acl- long.29, 29 22 2022 https://aclanthology.org/ .acl-long. . Hessel, J., Lee, L.：我的多模态模型是否学会了跨模态交互？这比你想象的更难判断！载于：2020年自然语言处理实证方法会议论文集（EMNLP）。第861–877页。计算语言学协会，线上（2020年11月）。https://doi.org/10.18653/v1/2020.emnlp-main.62, https://aclanthology.org/2020.emnlp-main.23 ¨ . Hofstatter, S., Lin, S.C., Yang, J.H., Lin, J., Hanbury, A.：通过平衡主题感知采样，高效地训练高效的稠密检索器。载于：第44届国际 ACM 信息检索研究与发展 SIGIR 会议论文集。第113–122页。SIGIR '21，计算机协会，纽约，纽约州，美国（2021年）。https://doi.org/10.1145/3404835.3462891, https://doi.org/10 3462891 24 1145 3404835 . / . . Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Ges- mundo, A., Attariyan, M., Gelly, S.：面向自然语言处理的参数高效迁移学习。载于：第36届国际机器学习会议论文集。第2790–2799页。PMLR（2019年5月），https://proceedings.mlr.press/v97/houlsby19 3498 25 2640 a.html, iSSN: - . Johnson, J., Douze, M., J´egou, H.：利用图形处理器进行十亿规模的相似性搜索。《IEEE大数据汇刊》7(3)，535–547（2019年）。https://doi.org/10 2921572 26 1109 2019 . /TBDATA. . . Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.：对比学习的困难负样本混合。载于：《神经信息处理系统进展》。第33卷，第21798–21809页。Curran Associates, Inc.（2020年），https://proceedings.neurips.cc/paper/2020/hash/f7cade80b7cc92b991cf4d2806d6bd78-Abstract.html 27. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.：用于开放域问答的密集段落检索。载于：2020年自然语言处理实证方法会议论文集（EMNLP）。第6769–6781页。计算语言学协会，线上（2020年11月 550 28 2020 ），https://www.aclweb.org/anthology/ .emnlp-main. . Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.：视觉中的 Transformer：综述。《ACM计算评论》（2021年12月）。https://doi.org/10.1145/3505244, https://doi.org/10.1145/3505244, 刚刚接收 29. Kingma, D.P., Ba, J.：Adam：一种随机优化方法。载于：第3届国际学习表征会议论文集（2015年1月），http://arxiv.org/abs/1412 6980 30 . . Kludas, J., Bruno, E., Marchand-Maillet, S.：多媒体信息检索中的信息融合。载于：国际自适应多媒体检索研讨会。第147–159页。施普林格（2007年）31. Lee, K., Chang, M.W., Toutanova, K.：弱监督开放域问答的潜在检索。载于：第57届计算语言学协会年会论文集。第6086–6096页。计算语言学协会，佛罗伦萨，意大利（2019年7月）。https://doi.org/10 1612 32 18653 1 19 1612, 19 . /v / P - https://aclanthology.org/P - . Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besan¸con, R., Moreno, J.G., Lov´on Melgarejo, J.：ViQuAE：一个关于命名实体的基于知识的视觉问答数据集。载于：第45届国际

ACM SIGIR 信息检索研究与发展会议论文集。SIGIR'22，计算机协会，纽约，纽约州，美国（2022年）。https://doi.org/10.1145/3477495.3531753, https://hal.archives-ouvertes.fr/hal-03650618 33. Lhoest,Q., Villanova del Moral, A., Jernite, Y.,Thakur,A.,vonPlaten,P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., ˇSaˇsko, M., Chh- ablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussi`ere, T., De- but, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., Wolf, T.：Datasets：一个用于自然语言处理的社区库。载于：2021年自然语言处理实证方法会议：系统演示论文集。第175–184页。计算语言学协会，线上及多米尼加共和国蓬塔卡纳（2021年11月），https://aclanthology.org/202121 34.emnlp-demo.. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.：Unicoder-VL：通过跨模态预训练实现视觉与语言的通用编码器。《AAAI 人工智能会议论文集》**34**(07)，11336–11344（2020年4月）。https://doi.org/10.1609/aaai.v34i07.6795, https://ojs.aaai.org/index.php/AAAI/article/view/6795, 07 编号：. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.：VisualBERT：视觉与语言的简洁高效基线（2019年1月），https://openreview.net/forum?id=OiixWIb7JpAg 36. Lin, J., Nogueira, R., Yates, A.：用于文本排序的预训练变换器：BERT及其他。《人类语言技术综合讲座》**14**(4)，1–325（2021 053 37 10 2200 01123 1 01 202108 ）。https://doi.org/ . /S ED V Y HLT . Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L.：Microsoft COCO：上下文中的常见物体。载于：Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T.（编）《计算机视觉—ECCV 2014》。第740–755页。《计算机科学讲义》，Springer国际出版，Cham（2014 1 48 38 10 1007 978 3 319 10602 ）。https://doi.org/ . / - - - - . Lu, J., Batra, D., Parikh, D., Lee, S.：ViLBERT：用于视觉与语言任务的预训练任务无关视觉语言表征。《神经信息处理系统进展》**32**，13–23（2019年），https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html 39. Luo, M., Zeng, Y., Banerjee, P., Baral, C.：用于基于知识的问答的弱监督视觉检索器—阅读器。载于：2021年自然语言处理实证方法会议论文集。第6417–6431页。计算语言学协会，线上及多米尼加共和国蓬塔卡纳（2021年11月）。https://doi.org/10.18653/v1/2021.emnlp-main.517, https: // aclanthology.org/2021517 40.emnlp-main.. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.：OK-VQA：一个需要外部知识的视觉问答基准。载于：IEEE计算机视觉与模式识别会议论文集。第3195–3204页（2019年），https://ieeexplore.ieee.org/document/8953725/ 41. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.：词语和短语的分布式表征及其组合性。《神经信息处理系统进展》**26**（2013年），https://papers.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html 42. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z.,

Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Meng, L., Bai, J.：《PyTorch：一种命令式风格、高性能的深度学习库》。神经信息处理系统进展 **32** （2019年），https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

43. Petroni,F.,Piktus, A.,Fan, A.,Lewis,P., Yazdani,M.,De Cao,N.,Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rockt¨aschel, T., Riedel, S.：KILT：面向知识密集型语言任务的基准。在：2021年计算语言学协会北美分会会议论文集：人类语言技术。第2523–2544页。计算语言学协会，线上（2021年6月）。https://doi.org/10.18653/v1/2021.naacl-main.200，https://aclanthology.org/2021.naacl-main.200

44. Qu, C., Zamani, H., Yang, L., Croft, W.B., Learned-Miller, E.：面向外部知识视觉问答的段落检索。在：第44届美国计算机协会 SIGIR 信息检索研究与发展国际会议论文集。第1753–1757页。SIGIR ’21，计算机协会，美国纽约州纽约（2021年）。https://doi.org/10.1145/3404835.3462987，https://doi.org/10.1145/3404835.3462987

45. Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H., Wang, H.：RocketQA：一种用于开放域问答密集段落检索的优化训练方法。在：2021年计算语言学协会北美分会会议论文集：人类语言技术。第5835–5847页。计算语言学协会，线上（2021年6月）。https://doi.org/10.18653/v1/2021.naacl-main.466，https://aclanthology.org/2021.naacl-main.466

46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等：通过自然语言监督学习可迁移视觉模型。在：国际机器学习会议。第8748–8763页。PMLR（2021）

47. Ram, O., Shachaf, G., Levy, O., Berant, J., Globerson, A.：学习在无监督条件下检索段落。在：2022年计算语言学协会北美分会会议论文集：人类语言技术。第2687–2700页。计算语言学协会，美国西雅图（2022年7月），https://aclanthology.org/2022.naacl-main.193

48. Ren, S., He, K., Girshick, R., Sun, J.：Faster R-CNN：借助区域建议网络实现实时目标检测。《神经信息处理系统进展》**28**，91–99（2015年），https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

49. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.：TREC-3中的Okapi。在：Harman, D.K.（编）第三届文本检索会议（TREC-3）。NIST 特别出版物，第500-225卷，第109–126页。美国国家标准与技术研究院（NIST）（1995年），https://citeseerx.ist.psu.edu/ viewdoc/download?doi=10.1.1.32.9922&rep=rep1&type=pdf

50. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.：KVQA：知识感知视觉问答。在：AAAI 人工智能会议论文集。第33卷，第8876–8884页（2019年），https://144.208.67.177/ojs/index.php/AAAI/article/view/4915

51. Sharma, P., Ding, N., Goodman, S., Soricut, R.：Conceptual captions：一个经过清理和上位词化、用于自动图像说明生成的图像替代文本数据集。在：论文集——

计算语言学协会第56届年会论文集（第1卷：长篇论文）。第2556–2565页（2018年）52. Shevchenko, V., Teney, D., Dick, A., van den Hengel, A.：视觉与语言推理：探索补充知识的益处。在：第三届超越视觉与语言：整合现实世界知识（LANTERN）研讨会论文集。第1–18页。计算语言学协会，乌克兰基辅（2021年4月1 5320211），https://aclanthology.org/.lantern-.. Smucker, M.D., Allan, J., Carterette, B.：信息检索评估中统计显著性检验的比较。在：美国计算机协会第十六届信息与知识管理会议论文集。第623–632页。CIKM '07，计算机协会，美国纽约州纽约（2007年11月）。https://doi.org/10.1145/1321440.1321528，https://doi.org/10.1145/1321440.132152854. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.：WIT：基于维基百科的多模态多语言机器学习图像文本数据集。在：第44届美国计算机协会 SIGIR 信息检索研究与发展国际会议论文集。第2443–2449页。SIGIR '21，计算机协会，美国纽约州纽约（2021年）。https://doi.org/10.1145/3404835.3463257，https://doi.org/103463257 5511453404835./.. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.：Dropout：一种防止神经网络过拟合的简单方法。《机器学习研究杂志》15(1)，1929–1958（2014年），出版方：JMLR. org 56. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.：VL-BERT：通用视觉语言表征的预训练。在：ICLR 2020论文集（2020年2月 08530 57190808530，1908），http://arxiv.org/abs/.arXiv：. . Tan, H., Bansal, M.：LXMERT：从Transformer学习跨模态编码器表征。在：2019年自然语言处理实证方法会议与第9届国际神经网络联合会议（EMNLP-IJCNLP）论文集。第5100–5111页。计算语言学协会，中国香港（2019年11月）。https://doi.org/10.18653/v1/D19-1514，https://www.aclweb.org/anthology/D191514 58-. Vaswani, A., Shazeer, N., Parmar, N.,Uszkoreit, J., Jones, L., Gomez,A.N., Kaiser, -L., Polosukhin, I.：注意力就是你所需要的一切。在：神经信息处理系统进展。第5998–6008页（2017年）59. Wang, Z., Li, L., Li, Q., Zeng, D.：多模态数据增强的知识图谱表征学习。在：2019年国际神经网络联合会议（IJCNN）。第1–8页（2019年7月）。https://doi.org/10 4407 60 1109 2019 8852079, 2161 2015 . /IJCNN. . iSSN: - . Weston, J., Chopra, S., Bordes, A.：记忆网络（ ）61. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.：HuggingFace 的 Transformers：最先进的自然语言处理。arXiv：1910 03771 [ ] 2020 1910 03771 . cs（7月），62 http://arxiv.org/abs/.. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.：书籍与电影对齐：通过观看电影和阅读书籍实现类似故事的视觉解释。在：IEEE国际计算机视觉会议（ICCV）论文集（12月

2015年），https://www.cv-foundation.org/openaccess/content iccv 2015/html/Zhu_书籍对齐and ICCV 2015 paper.html