检索增强生成

播报 锁定 讨论 1 上传视频

大模型前沿技术之一 | 展开2个同名词条

收藏 | 11 | 3

同义词 RAG（大模型内化吸收知识的过程）一般指检索增强生成

本词条由中国科学院大学计算机科学与技术学院、中国科学院计算技术研究所 参与编辑并审核，经科普中国·科学百科认证 。

检索增强生成（Retrieval-augmented Generation），简称RAG，是当下热门的大模型前沿技术之一 [1]。

检索增强生成模型结合了语言模型和信息检索技术。具体来说，当模型需要生成文本或者回答问题时，它会先从一个庞大的文档集合中检索出相关的信息，然后利用这些检索到的信息来指导文本的生成，从而提高预测的质量和准确性 [2]。

中文名	检索增强生成	所属学科	人工智能
外文名	Retrieval-augmented Generation	简　称	RAG

目录

历史沿革

播报

2020年，Facebook AI Research(FAIR)团队发表名为《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》的论文，。该篇论文对RAG概念进行详细介绍和解释 [2]。

起初，Naive RAG 遵循传统过程 Indexing-Retrieval-Generation，也被称为 “Retrieve-Read” 框架。

之后，Advanced RAG 提高检索质量，采用了检索前和检索后策略（pre-retrieval and post-retrieval strategies）。为了解决索引问题，Advanced RAG 通过使用滑动窗口方法、细粒度分段和元数据的合并来改进其索引技术。

后来，Modular RAG 引入多个特定功能模块和替换现有模块， 总体上展示了更大的灵活性。其过程并不局限于顺序检索和生成，包括了迭代和自适应检索等方法 [11]。

技术定义

播报

检索增强生成（Retrieval-Augmented Generation，RAG）是一种结合检索和生成技术的模型。它通过引用外部知识库的信息来生成答案或内容，具有较强的可解释性和定制能力，适用于问答系统、文档生成、智能助手等多个自然语言处理任务中。RAG模型的优势在于通用性强、可实现即时的知识更新，以及通过端到端评估方法提供更高效和精准的信息服务 [1]。

检索增强生成的概述图

中国科学院大学
“春分工程·科学百科”

本词条认证专家为

顾佳 | 研究生
中国科学院大学计算机科学与技术学院
国科学院计算技术研究所

何苯 | 教授
中国科学院大学计算机科学与技术学

TA说 分享你的世界 查

RAG 技术：让 AI 从“书呆子”变“开卷小天才”！

Alter聊科技 · 知名科技自媒体

权威合作编辑

中国科学院大学计算机科学院
中国科学院大学计算机科学院是201...

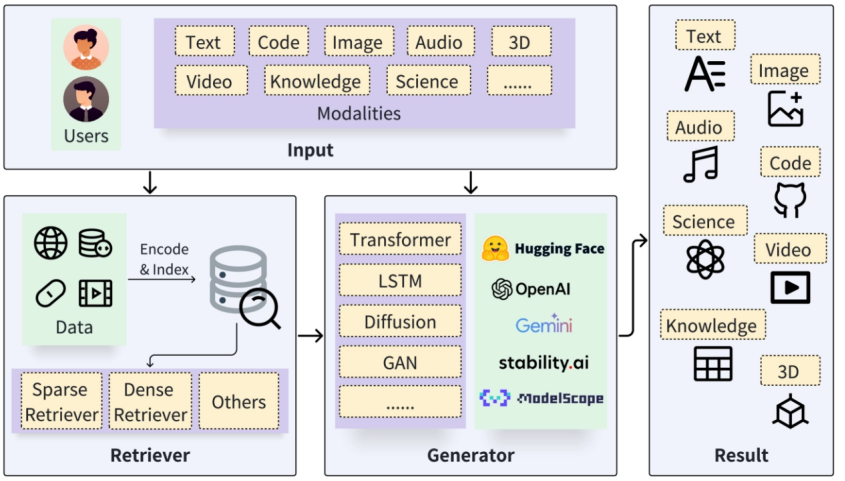
工作流程

播报

RAG的工作原理是通过检索大规模文档集合中的相关信息，然后利用这些信息来指导文本的生成，从而提高预测的质量和准确性。

RAG 包含三个主要过程：检索、增强和生成 [3]。

- 检索：根据用户的查询内容，从外部知识库获取相关信息。具体而言，将用户的查询通过嵌入模型转换为向量，以便与向量数据库中存储的相关知识进行比对。通过相似性搜索，找出与查询最匹配的前 K 个数据。
- 增强：将用户的查询内容和检索到的相关知识一起嵌入到一个预设的提示词模板中。
- 生成：将经过检索增强的提示词内容输入到大型语言模型中，以生成所需的输出。



图示 通用检索增强生成架构

通过这一过程，RAG模型能够在各种自然语言处理任务中发挥作用，如问答系统、文档生成和自动摘要、智能助手和虚拟代理、信息检索以及知识图谱填充等。同时，RAG模型具有及时更新、解释性强、高度定制能力、安全隐私管理以及减少训练成本等优点。与微调相比，RAG是通用的，适用于多种任务，并且能够实现即时的知识更新而无需重新训练模型 [1]。

重要组成成分

播报

整个 RAG 系统由两个核心模块组成：检索器和生成器。检索器从数据存储中搜索相关信息，生成器生成所需内容 [4]。

生成器

生成式人工智能在各种任务中的出色表现开启了 AIGC（人工智能生成内容）时代。生成模块在 RAG 系统中起着至关重要的作用。不同的生成模型适用于不同的场景，例如用于文本到文本任务的 Transformer 模型、用于图像到文本任务的 VisualGPT、用于文本到图像任务的 Stable Diffusion，以及用于文本到代码任务的 Codex 等 [4]。

生成器的基础技术包括 Transformer 模型、长短期记忆网络（LSTM）、扩散模型和生成对抗网络（GAN）。

Transformer模型是一种用于处理语言数据的神经网络模型，非常适合用于翻译、文本生成和理解等任务 [5]。由自注意力机制、前馈网络、层规范化模块和残差网络组成 [6]。

长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的RNN（循环神经网络）存在的长期依赖问题而专门设计出来的，所有的RNN都具有一种重复神经网络模块的链式形式 [7]。

扩散模型是一类深度生成模型，可以创建真实且多样化的数据样本（包括图像、文本、视频等） [8]。

生成对抗网络（GAN）是备受期待的深度学习模型，可以模拟和生成逼真的图像、音频和其他数据 [9]。

检索器

检索是在给定信息需求的情况下识别和获取相关信息。具体来说，让我们考虑可以概念化为键值存储的信息资源，其中每个键对应一个值（键和值可以相同）。给定一个查询，目标是使用相似度函数从数据库中选取最相似的键，并得到配对的值。根据不同的相似度函数，现有的检索方法可以分为稀疏检索、密集检索和其他方法 [4]。

稀疏检索方法通常用于文档检索，其中键/值表示要搜索的文档。这些方法利用术语匹配指标，例如 TF-IDF，查询可能性；BM2.5分析文本中的单词统计数据并构建倒排索引以实现高效搜索 [4]。

密集检索方法使用密集嵌入向量表示查询和键，并构建近似最近邻 (ANN) 索引以加快搜索速度 [4]。

什么是权威编辑

词条统计

浏览次数：158442次
编辑次数：8次历史版本
最近更新：simayi315（2025-02-27）

突出贡献榜

54hehezi

分类

播报

基于查询的 RAG

基于查询的 RAG 源于即时增强的理念，将用户的查询与从检索到的信息中获得的见解无缝集成，将其直接输入到生成器输入的初始阶段。这种方法在 RAG 应用中很普遍。检索后，获取的内容与用户的原始查询合并以形成复合输入，然后由生成器处理以创建响应。基于查询的 RAG 广泛应用于各种模态 [4]。

基于潜在表征的 RAG

在基于潜在表示的 RAG 框架中，检索到的对象作为潜在表示被纳入生成模型。这增强了模型的理解能力并提高了生成内容的质量 [4]。

基于 Logit 的 RAG

在基于 Logit 的 RAG 中，生成模型在解码过程中通过 Logit 整合检索信息。通常，通过简单的求和或模型将 Logit 组合起来，以计算分步生成的概率 [4]。

推测性 RAG

推测性 RAG 寻求使用检索而不是纯生成的方式，从文档中直接提取单词或短语，以节省资源和加快响应速度。它将生成器和检索器解耦，从而可以直接使用预训练模型作为组件。在这个范式中，我们可以探索更广泛的策略来有效利用检索到的内容 [4]。

理论依据

播报

RAG模型的理论依据主要围绕在利用检索、知识填充、智能助手、信息检索、数据更新、定制能力、安全管理等方面，以提供准确、及时、解释性强、高度定制、安全保障的服务。

基于检索的知识填充和自动生成：RAG利用大规模文档集合进行检索，填充文本以生成准确的答案和内容。

智能助手和虚拟代理：RAG可用于构建智能助手或虚拟代理，通过结合聊天记录回答用户问题，提供信息和执行任务。

信息检索和知识图谱填充：RAG能改进信息检索系统，使其更准确深刻，并用于填充知识图谱中的实体关系，提高生成文本的可靠性。

数据更新及时性和解释性回复：RAG模型具备检索库的即时更新机制，回复具有强解释性，由检索库直接提供答案，用户可核实准确性。

高度定制能力和安全隐私管理：RAG可根据特定领域的知识库和prompt定制，通过限制知识库权限实现安全控制。

减少训练成本和通用性：RAG模型具有可拓展性，通过大量数据直接更新知识库，不需重新训练模型，适用于多种任务 [2]。

应用场景

播报

1.问答系统（QA Systems）：RAG可以用于构建强大的问答系统，能够回答用户提出的各种问题。它能够通过检索大规模文档集合来提供准确的答案，无需针对每个问题进行特定训练。

2.文档生成和自动摘要（Document Generation and Automatic Summarization）：RAG可用于自动生成文章段落、文档或自动摘要，基于检索的知识来填充文本，使得生成的内容更具信息价值。

3.智能助手和虚拟代理（Intelligent Assistants and Virtual Agents）：RAG可以用于构建智能助手或虚拟代理，结合聊天记录回答用户的问题、提供信息和执行任务，无需进行特定任务微调。

4.信息检索（Information Retrieval）：RAG可以改进信息检索系统，使其更准确深刻。用户可以提出更具体的查询，不再局限于关键词匹配。

5.知识图谱填充（Knowledge Graph Population）：RAG可以用于填充知识图谱中的实体关系，通过检索文档来识别和添加新的知识点 [2]。

技术优势

播报

1.外部知识的利用：RAG模型可以有效地利用外部知识库，它可以引用大量的信息，以提供更深入、准确且有价值的答案，这提高了生成文本的可靠性。

2.数据更新及时性：RAG模型具备检索库的更新机制，可以实现知识的即时更新，无需重新训练模型。说明RAG模型可以提供与最新信息相关的回答，高度适配要求及时性的应用。

3.回复具有解释性：由于RAG模型的答案直接来自检索库，它的回复具有很强的可解释性，减少大模型的幻觉。用户可以核实答案的准确性，从信息来源中获取支持。

4.高度定制能力：RAG模型可以根据特定领域的知识库和prompt进行定制，使其快速具备该领域的能力。说明RAG模型广泛适用于的领域和应用，比如虚拟伴侣、虚拟宠物等应用。

5.安全和隐私管理：RAG模型可以通过限制知识库的权限来实现安全控制，确保敏感信息不被泄露，提高了数据安全性。

6.减少训练成本：RAG模型在数据上具有很强的可拓展性，可以将大量数据直接更新到知识库，以实现模型的知识更新。这一过程的实现不需要重新训练模型，更经济实惠 [2]。

限制

播报

检索结果中的噪声

信息检索不可避免的噪声（表现为不相关内容或误导性信息）可能会在 RAG 系统中造成故障点。检索噪声的影响仍然不清楚，导致在实际使用中对度量选择和检索器-生成器交互产生混淆 [4]。

额外开销

随着检索源规模的扩大，存储和访问复杂度也将增加 [10]。这样的开销阻碍了RAG在对延迟敏感的实时服务中的实用性。

检索器和生成器之间的差距

由于检索器和生成器的目标可能不一致，并且它们的潜在空间可能不同，因此设计它们的交互需要精心设计和优化。当前的方法要么将检索和生成分开，要么在中间阶段将它们集成在一起。虽然前者更加模块化，但后者可以从联合训练中受益，但会阻碍通用性。选择一种经济有效的交互方法来弥补差距是一项挑战，需要在实践中深思熟虑 [4]。

系统复杂性增加

引入检索不可避免地会增加系统复杂性和需要调整的超参数数量 [4]。

冗长的上下文

播报

RAG 的一个主要缺点，特别是基于查询的 RAG，是它极大地延长了上下文，这使得它对于上下文长度有限的生成器来说是不可行的。此外，加长的上下文也会普遍减慢生成过程 [4]。

未来的潜在方向

播报

灵活的 RAG 管道

RAG 系统正在逐步采用灵活的管道，例如递归、自适应和迭代 RAG。通过精确的调整和细致的工程设计，检索源、检索器、生成器和 RAG 子系统的独特组合有望解决复杂任务并提高整体性能。我们热切期待开创性的探索，这将推动更具创新性的 RAG 系统的发展 [4]。

融入长尾和实时知识

虽然 RAG 的一个主要动机是利用实时和长尾知识，但很少有研究探索知识更新和扩展的管道。许多现有研究仅使用生成器的训练数据作为检索源，而忽略了检索可以提供的动态和灵活信息。因此，越来越多的研究设计了具有不断更新的知识和灵活来源的 RAG 系统。我们还希望 RAG 能更进一步，适应当今网络服务中的个性化信息 [4]。

与其他技术相结合

RAG 与其他旨在提高 AIGC 效率的技术（如微调、强化学习、思路链和基于代理的生成）相互关联。这些方法的结合仍处于早期阶段，需要进一步研究，通过新颖的算法设计充分发挥其潜力 [4]。

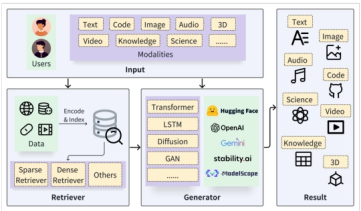
多模态RAG

RAG 已经超越了它最初基于文本的问答范围，包含了各种各样的模态数据。这种扩展催生了创新的多模态模型，将 RAG 概念整合到各个领域: 图像、音视频、Vid2Seq、code。 结语随着 LLM 及 RAG 技术的不断发展，Agent 的基础能力愈发强大，如何将底层能力整合，产生一个现象级的产品是当下最直接的诉求。

词条图册

更多图册 >

检索增强生成(1张)



词条图片(1张)

TA说 分享你的世界

查看更多 >

RAG 技术：让 AI 从 “书呆子” 变身 “开卷小天才”！
普通 AI 是 “闭卷死背党”，但 RAG 就是 “开卷活学王”。而且 RAG 可不只是会查，还能把知识说得让你忍不住拍大腿叫好。

Alter聊科技

参考资料

1 阿里云推出企业级大模型RAG解决方案 “最强外挂”可大幅提升语言模型表现 ．驱动之家 [引用日期2024-12-12]

2 RAG一文读懂！概念、场景、优势、对比微调与项目代码示例 ．飞桨PaddlePaddle [引用日期2024-12-12]

3 【RAG系统综述】|一文读懂RAG（检索增强生成） ．GPTSecurity | AIGC网络安全 [引用日期2024-12-12]

4 [2402.19473v6] Retrieval-Generated Content: A Survey ．arxiv [引用日期2024-12-12]

5 [1706.03762] Attention Is All You Need ．arxiv [引用日期2024-12-12]

6 [2009.06732] Efficient Transformers: A Survey ．arxiv [引用日期2024-12-12]

7 Long Short-Term Memory | MIT Press Journals & Magazine ．IEEE Xplore [引用日期2024-12-12]

8 [2209.00796] Diffusion Models: A Comprehensive Survey of Methods and Applications ．arxiv [引用日期2024-12-12]

9 [2001.06937v1] A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications ．arxiv [引用日期2024-12-12]

10 [2311.15578] Experimental Analysis of Large-scale Learnable Vector Storage Compression ．arxiv [引用日期2024-12-12]

展开全部

学术论文

内容来自 百度学术

- 陈卓. 图像检索数据库生成方法,增强现实的方法及装置. 2017

查看全部 >

相关搜索

宝宝奶粉质量排行榜　双电源转换开关　易贷通app官方下载　梦幻西游手游普陀怎么加点　vector机器人

monsti游戏下载

新手上路

成长任务　编辑入门
编辑规则　本人编辑 NEW

我有疑问

内容质疑　在线客服
官方贴吧　意见反馈

投诉建议

举报不良信息　未通过词条申诉
投诉侵权信息　封禁查询与解封