自适应增强检索器作为通用插件提高了语言模型的泛化能力 [17]

Zichun Yu, Chenyan Xiong, Shi Yu, Zhiyuan Liu

摘要

检索增强可以通过向语言模型（LM）提供外部信息来帮助它们完成知识密集型任务。先前的检索增强工作通常会联合微调检索器和 LM，使它们紧密耦合。本文探讨了一种通用检索插件的方案：检索器旨在辅助可能事先不知道或无法一起微调的目标 LM。为了为未知的目标 LM 检索有用的文档，作者提出了增强适应型检索器（AAR），它从已知的源 LM 中学习得到 LM 的偏好。实验表明，使用小型源 LM 训练的 AAR，能够显著改善较大目标 LM 的零样本泛化能力。进一步的分析表明，不同 LMs 的偏好重叠，使得使用单个源 LM 训练的 AAR 可以作为各种目标 LM 的通用插件。

关键词：检索增强；语言模型

1 引言

具有数十亿参数的大型语言模型（LMs）能够捕捉大量人类知识，从而在各种下游任务中实现持续改进。然而，大型 LMs 的不可否认的缺点在于它们的高计算成本，这对其效率产生负面影响。此外，从预训练中记忆的知识以及 LMs 的隐式推理过程有时可能是不准确和棘手的，这妨碍了它们在知识密集型任务上的应用。

与利用嵌入在语言模型参数中的知识和推理能力不同，检索增强通过引入一个能够从外部语料库检索知识的检索器来提升 LM 的性能。另一方面，之前的检索增强方法需要对骨干 LM 进行微调，以适应检索器并解决特定的下游任务。当越来越多的独特需求出现时，这种微调可能是昂贵的。更重要的是，许多顶尖的 LM 只能通过黑盒 API 进行访问。这些 API 允许用户提交查询并接收响应，但通常不支持微调。

2 相关工作

使用从外部存储中检索到的信息增强 LMs 在各种知识密集型任务上表现出了有效性 [3]。先前的研究探索了以端到端方式训练整个检索器-语言模型系统的新方法 [1, 5, 7, 8, 10]，使用检索增强的序列对数似然、fusion-in-decoder 的注意力蒸馏或知识图谱。为了将检索器与 LM 分离，Rubin 等人 [14] 为上下文学习训练了一个独立的提示检索器，而 Lin 等人 [11] 仅通过与少量无监督样本相似的检索数据对 LM 进行微调。

最近的研究采用了零样本检索增强，不会对 LM 在 InstructGPT 上进行微调 [13]。它可以在以实体为中心的问答 [12]、思维链推理 [4] 和多跳问答 [9] 等方面受益。Shi 等 [15] 使用 LM 的似然性训练检索器以满足黑盒 LM 的偏好，并采用 GPT-3 Curie [2] 提供监督信号。

3 本文方法

3.1 本文方法概述

如下图所示，本文利用一个小型的编码器-解码器 LM 作为源 LM，通过其 Fusion-in-Decoder 注意力分数来注释 LM 首选文档。这些 LM 首选文档和人类首选文档结合，形成正文档集。负文档由检索器自己使用 ANCE 技术 [16] 进行挖掘。通过利用 LM 的偏好对检索器进行微调后，它可以作为插件直接辅助未见过的目标 LM 在零样本任务泛化上取得更好的性能。

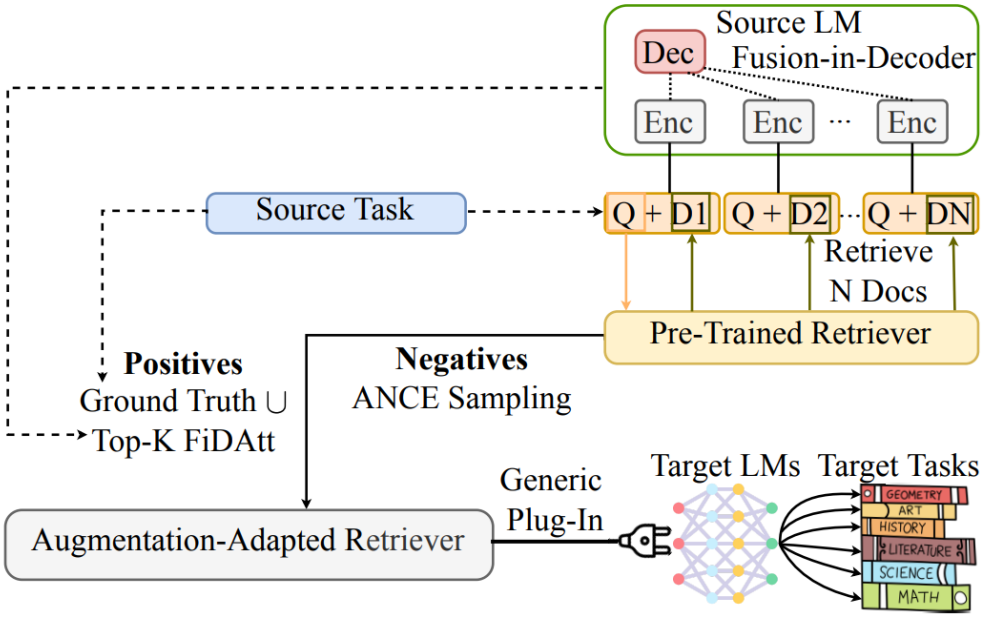

图 1. 方法示意图

3.2 基础知识

检索器的目标是从语料库 C 中找到一个增强文档集 D^a，帮助 LM 处理一个给定的查询 q。检索模型首先使用预训练的编码器 g 将查询 q 和文档 d 表示到嵌入空间，

$$\boldsymbol{q} = g(q); \boldsymbol{d} = g(d), d \in C \tag{1}$$

然后通过点积函数 f 匹配它们的嵌入，该函数支持快速近似最近邻搜索（ANN），得到包含前 N 个检索到的文档 D^a。

论文中使用的 Fusion-in-Decoder（FiD）是由 zacard 和 Grave [6] 提出的。如下图所示，它将检索到的每个 passage 都与 question 通过 encoder 分别编码，然后 concat 在一起输入 decoder 生成最终的回复。

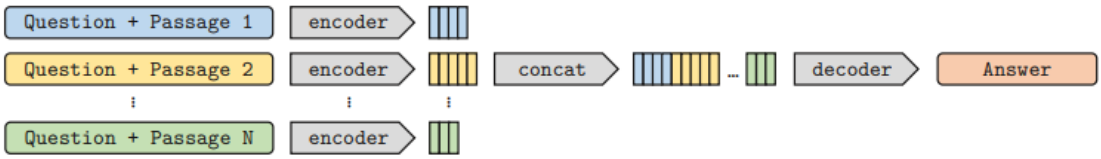

图 2. Fusion-in-Decoder 方法结构

即它首先分别对 (d_i^a, q) 对的每个连接分别进行编码，然后让解码器关注所有部分：

$$\mathrm{FiD}(q) = \mathrm{Dec}(\mathrm{Enc}(d_1^a \oplus q) \ldots \mathrm{Enc}(d_N^a \oplus q)) \tag{2}$$

3.3 自适应增强检索器

作者利用一个编码器-解码器 LM 作为源 LM（L_s），在源任务（T_s）上提供 LM 首选信号，以对预训练的检索器进行微调。然后，作者将微调后的检索器插入到未见过的目标语言模型（L_t）上，针对一组与源任务 T_s 不相交的目标任务（T_t）。

作者的训练方法始于源任务 T_s，作者汇总源语言模型 L_s 对应于文档 d_i^a 的第一个解码器标记的所有层、所有头以及 $d_i^a \oplus q$ 的所有输入标记 t 的平均 FiD 交叉注意力（FiDAtt）分数 S_i^a。

$$S_i^a = \frac{1}{\mathrm{ln} * \mathrm{hn} * \mathrm{tn}} \sum_{\text{layers}} \sum_{\text{heads}} \sum_{t \in d_i^a \oplus q} \mathrm{FiDAtt}(\mathrm{FiD}(q)) \tag{3}$$

其中 ln,hn,tn 是层数、头数和输入标记数。

正文档集由用 FiDAtt 分数注释的 lm 首选文档和人类首选文档结合生成。

$$D^{a+} = D^{h+} \cup \text{Top-}K_{S_i^a, D^a} \tag{4}$$

其中，D^{h+} 是源任务 T_s 上的人工首选正文档集（即 ground truth）。Top-$K_{S_i^a, D^a}$ 表示在检索到的文档集合 D_a 中，前 k 个平均 FiDAtt 分数 S_i^a 的文档。

然后，采用 ANCE 方法对难负样本进行采样，并构建检索器的训练损失 $\mathcal{L}$：

$$D^- = \mathrm{ANN}^M_{f(q,\circ)} \setminus D^{a+} \tag{5}$$

$$\mathcal{L} = \sum_q \sum_{d^+ \in D^{a+}} \sum_{d^- \in D^-} l(f(\boldsymbol{q}, \boldsymbol{d}^+), f(\boldsymbol{q}, \boldsymbol{d}^-)) \tag{6}$$

其中 M 是负采样深度的超参数，l 是标准交叉熵损失。微调检索器后，作者直接将其用于增强未见过的目标语言模型 L_t 的每个来自目标任务集 T_t 的任务。

4 复现细节

4.1 与已有开源代码对比

本工作引用了论文提供的源码，实现了增强适应型检索器（AAR），并使用 AAR 作为插件辅助更大的目标 LM 进行问答任务。

4.2 实验环境搭建

实验环境：Python 3.9

Python 依赖包：nltk 3.6.5；numpy 1.15.4；sentencepiece 0.1.8；tensorflow 1.12.0；regex 2021.8.3；deepspeed 0.3.9；torch 1.10.1

4.3 实验设置

源 LM：Flan − $T5_{Base}$

检索器：从 $T5_{Base}$ 初始化的 ANCE

检索语料库：MS MARCO

目标 LM：Flan − $T5_{Base}$ 和 Flan − $T5_{Large}$

目标任务：MMLU。MMLU 是一个多任务语言理解数据集，它包括 57 个选择题回答子任务。这些子任务一般可以分为四类: 人文、社会科学、STEM 和其他。我们对每个类别中的子任务的准确率进行平均，从而得到最终的分数。

在训练中设置总文档数 N=10,LM 首选文档数 K=2, 负挖掘深度 M=100。

5 实验结果分析

本部分主要分为两个方面，包括对不同大小的两个 LM 作为目标 LM 的实验结果与分析，实验主要在 MMLU 任务上进行测试。

5.1 评价指标

我们对 MMLU 每个类别的子任务的准确率进行平均，从而得到最终的分数。即我们使用准确率来进行评估。

5.2 Flan − $T5_{Base}$ 结果

使用 Flan − $T5_{Base}$ 作为源 LM，Flan − $T5_{Base}$ 作为目标 LM，实验结果如下表所示。可以看出，使用检索增强方法可以提高大语言模型的性能，并且作者提出的 AAR 可以有效学习到语言模型的偏好信号，进一步提高模型性能。

表 1. Flan − $T5_{Base}$ 结果

模型	MMLU
Flan − $T5_{Base}$	36.1
Flan − $T5_{Base}$ AAR（使用初始的 ANCE）	41.7
Flan − $T5_{Base}$ AAR	44.2
Flan − $T5_{Base}$ AAR（原）	44.8

5.3 $\text{Flan}-\text{T5}_{\text{Large}}$ 结果

使用 $\text{Flan}-\text{T5}_{\text{Base}}$ 作为源 LM,$\text{Flan}-\text{T5}_{\text{Large}}$ 作为目标 LM，实验结果如下表所示。可以看出，使用较小的 LM 学习到的偏好信号微调检索器后，在较大 LM 上可以表现出不错的性能。

表 2. $\text{Flan}-\text{T5}_{\text{Large}}$ 结果

模型	MMLU
$\text{Flan}-\text{T5}_{\text{Large}}$	44.8
$\text{Flan}-\text{T5}_{\text{Large}}$ AAR	49.3
$\text{Flan}-\text{T5}_{\text{Large}}$ AAR（原）	50.4

5.4 实验结果总结

实验结果表明，在 AAR 的帮助下，不同大小的目标 lm 在零样本设置中可以显著提高模型性能表现。

6 总结与展望

总的来说，自适应增强检索器（AAR）可作为插件，用以增强可能事先未知或难以进行联合微调的目标语言模型。此外，AAR 能够直接支持黑盒语言模型，无需对其进行任何微调。实现这一目标的方法是通过将来自小型源语言模型的首选文档与实际情况相结合，构建 AAR 的训练数据。或许我们还可以寻找其他方法，更合理地选择语言模型的首选文档。

参考文献

[1] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2021.

[2] 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, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[3] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. 2020.

[4] Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *ArXiv*, abs/2301.00303, 2022.

[5] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *ArXiv*, abs/2012.04584, 2020.

[6] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *ArXiv*, abs/2007.01282, 2020.

[7] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299, 2022.

[8] Mingxuan Ju, W. Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. Grape: Knowledge graph enhanced passage reader for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

[9] O. Khattab, Keshav Santhanam, Xiang Lisa Li, David Leo Wright Hall, Percy Liang, Christopher Potts, and Matei A. Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *ArXiv*, abs/2212.14024, 2022.

[10] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020.

[11] Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. Unsupervised cross-task generalization via retrieval augmentation. *ArXiv*, abs/2204.07937, 2022.

[12] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[13] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

[14] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *ArXiv*, abs/2112.08633, 2021.

[15] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Replug: Retrieval-augmented black-box language models. *ArXiv*, abs/2301.12652, 2023.

[16] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *ArXiv*, abs/2007.00808, 2020.

[17] Zichun Yu, Chenyan Xiong, Shih Yuan Yu, and Zhiyuan Liu. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Annual Meeting of the Association for Computational Linguistics*, 2023.