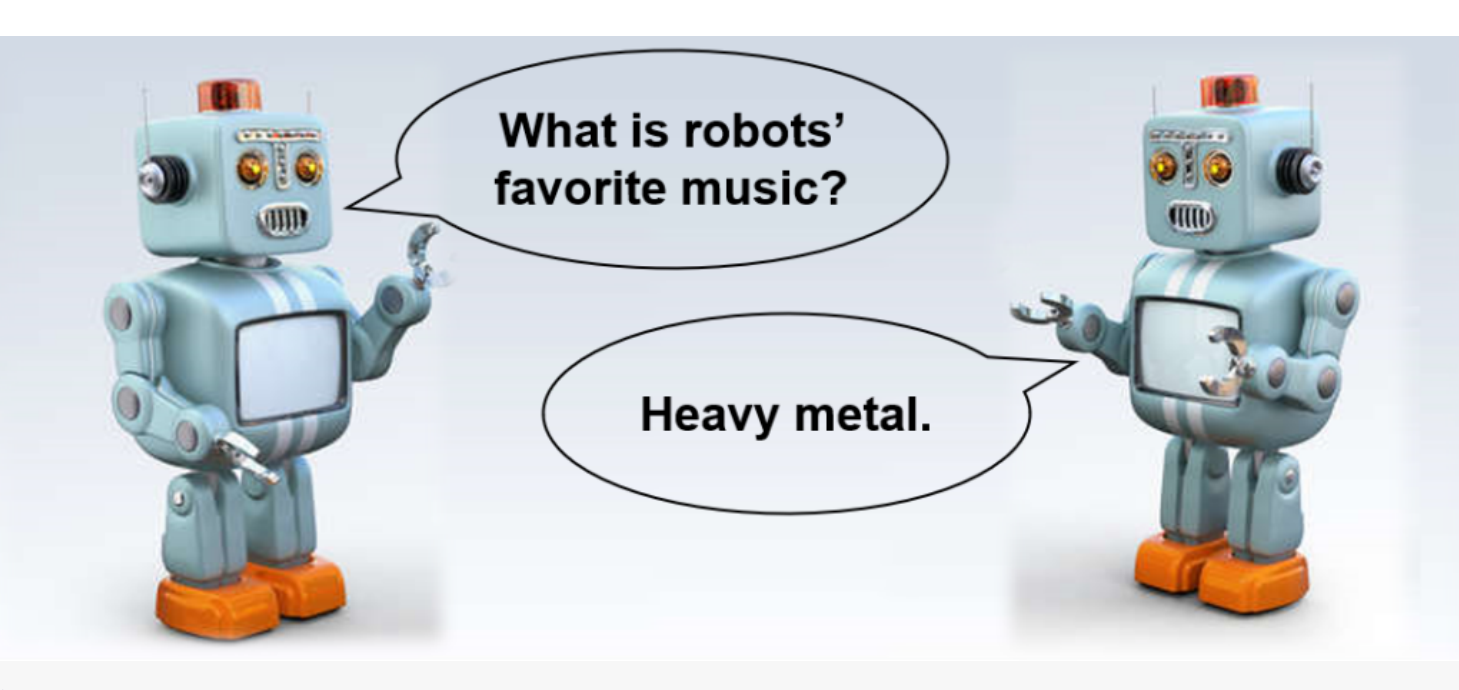
AllenAI 发布万能问答系统 MACAW！各类题型样样精通，性能大幅超越 GPT-3！

原创 python 夕小瑶的卖萌屋 2021-10-13 12:05



文 | python

前言

GPT-3 等超大规模预训练语言模型，在少监督任务（few-shot tasks）上取得了令人瞩目的成绩。而这篇文章中，AllenAI的研究员提出了大规模生成式问答模型，MACAW。基于多角度预训练，MACAW**可以用于包括段选取(span selection)、选择题、生成式问答在内的一切问答任务，以及包括问题生成、选项生成、解释生成等在内的多种问答相关任务**。MACAW在ARC、ARC-DA等多个问答基准上取得了业界最好的成绩，并且只用了GPT-3 十六分之一的参数规模，就**在无监督问答数据集 Challenge300 上，相较GPT-3取得了10%的绝对提升**。

论文题目：
General-Purpose Question-Answering with MACAW

论文链接：
https://arxiv.org/abs/2109.02593

项目地址：
https://github.com/allenai/macaw

概览

MACAW（Multi-Angle q(C)uestion-AnsWering），字面含义指一种多角度问答模型。在这篇文章中，作者扩展了之前自己在UnifiedQA[1] 中提出了统一问答框架，将不同的问答任务形式进一步扩展到不同的问答相关任务，从而实现一种多角度的预训练的方式，提升模型的通用性的同时，也提升模型的鲁棒性。

> 编者按：这篇文章也可以称为Unified-UnifiedQA。一方面，这篇文章两个作者是均为UnifiedQA文章的作者；另一方面，在UnifiedQA中，作者利用预训练语言模型，将所有生成、抽取、选择式的问答任务形式统一，而这篇文章中进一步统一了如问题生成、选项生成，回答解释生成等问答相关任务。

具体而言，MACAW基于预训练的T5模型[2]，并通过两阶段精调得到。在第一阶段中，采用包括BoolQ、 NarrativeQA、RACE在内的7个问答数据集，并通过问题生成、答案生成、选项生成、选项加答案生成等6种不同的任务范式，让模型充分地学到问答相关的一切技巧。而在第二阶段中，采用了两个标注有答案解释的数据集，ARC和ARC-DA，进一步引入了8种和解释相关的任务范式，让模型知其然的同时，也能知其所以然。

MACAW具有以下三点优势：

- 面向领域外的样本，MACAW具备出色的无监督迁移学习能力。在 Challenge300 数据集上，相较GPT-3取得10%的绝对提升。
- MACAW具有“多角度问答能力”，无论是问题生成，还是回答生成，亦或是选项生成，MACAW都能胜任。
- MACAW还能生成回答的解释，体现出知其然亦知其所以然的能力。

MACAW 模型

精调阶段1：会出题的问答模型，才是个好模型

在第一个精调阶段中，作者在7个问答数据集上，以6种不同的任务形式精调T5。这里选用的数据集有答案段选取形式的SQuAD 2.0，有是否类问题BoolQ，有描述类生成式回答的NarrativeQA，有多项选择题的RACE等等。

为了统一不同的任务形式，作者以slot的方式约定了任务的输入输出。例如下图展示的是一个给定问题（questions）和候选选项（mcoptions），让模型对答案（answer）做出预测的任务形式：

在7个数据集上，作者根据数据集特点，设计了6种任务作为第一阶段的训练目标。如下表所示。其中的符号，Q指问题、C指上下文（即阅读理解读的文本），A指答案，M指候选选项（选择题里才有）。例如QC→A指答案生成，AC→Q指问题生成，QAC→M指给定文章问题和答案的选项生成。可以看到，这里面除了在原本UnifiedQA中就包含的答案生成任务外，**还引入了大量问题生成、选项生成等任务。** 让模型在学会解题的同时，也学会出题。

这里有两个有意思的点。一方面，任务模式中可以有多种输出，而考虑到生成模型自回归解码，多种输出之间的顺序关系是有意义的。比如AC→QM，是先根据文章和答案，生成问题，再根据生成的问题，生成候选选项。另一方面，这里的任务设计考虑了数据特点，比如虽然QA→C，即给定问答对，生成阅读文章，理论上可行。但实际中，因为问答对中包含的信息过少，文章C中含有大量无关信息，导致这种任务没有太多实际意义。因此，这里也没有涉及这种没有意义的任务。

在实际训练过程中，所有数据集与所有任务范式混合在一起进行训练。以8的批处理大小，迭代训练了120k步。不同的数据集之间进行等概率采样。不同的任务之间也先验性的赋给了一个采样的权重。毕竟相对来说，答案生成比问题生成更重要一点，而这两者又都明显比选项生成等任务更重要。

第一阶段精调的模型，在精调任务上的表现如下表所示。其中，NarrativeQA以ROUGE-L作评价，SQuAD2.0以F1作评价，其余任务均以精度为评价指标。可以看到，引入多种不同的任务范式之后，模型在问答任务上的表现与单一问答任务的结果比是相当的，但具有了解决更多不同类型任务的能力。

> 编者按：实际上，由于训练时采用多任务混合训练，测试时使用单一任务测试，这一差异肯定会带来表现下降。个人感觉，如果在这一阶段训练后再引入单一问答任务的精调，或使用课程学习的方式，将这一阶段预训练逐渐转化为纯问答形式，在问答任务上的表现会更好。不过，这里作者主要是做一个初步的预训练，而非为了刷问答任务的指标，因此没有做这些尝试。

精调阶段2：成熟的问答模型，还能自我解释

作者进一步引入了解释类任务，让模型知其然的同时还能够知其所以然，使无监督问答任务上的回答更合理。作者使用了 WorldTree V2 explanation bank[3]中的几十标注，覆盖65% 的 ARC 数据集和 50% 的 ARC-DA 数据集。

> 这里的“无监督”，其实也可以理解成是领域外数据，即没有和测试集同分布的训练数据，但有大量形式类似的相关任务可以用于训练。

任务形式如上图所示，其中E代表解释（explanation）。**除了部分在第一阶段也采用的任务外，作者引入了8个和解释相关的任务**，例如QM→AE，即给定问题和选项，生成答案后再生成解释，AQC→E，即给定文本、问题和答案，生成解释，E→QA，给定解释，生成问题并作出回答。第二阶段精调中，作者采用和第一阶段类似的训练策略，在第一阶段的结果上进一步训练了6k步。部分示例如下图所示。

实验分析

ARC数据集

作者在ARC数据集上检测了MACAW的表现，如下表所示。**MACAW在ARC、ARC-Easy和ARC- DA 上均达到了业界最优的表现**[4]。不过，但监督学习范式下，在生成答案之后引入解释的生成（即QM→AE），并没有让模型表现有明显的提升。作者分析表示，引入解释生成后，答案生成时条件依赖于生成的解释，会使得生成答案的确定性更高。而不够完美的解释可能会反而强化错误答案的概率。

Challenge300数据集

Challenge300 是一个手工构建的问答评价数据集，由300道问题组成，涵盖了22类不同的考察方面。在这里作为一个无监督（领域外）的评价基准。因这个数据集答案较为灵活，规模较小，评价时以人工评价为准。作者对比了MACAW与GPT-3等主流无监督问答模型，实验结果如下表所示。可以看到，和GPT-3相比，MACAW也可以取得10%的绝对提升，即使MACAW的11B的参数规模知识GPT-3 的175B的参数规模的十六分之一。

在不同类型的问题上的表现对比如上图所示。可以看出，MACAW在**很多问题类型上均表现出了明显的优势**。例如：

- 实体替换类问题（Entity Substitution）：挖掘实体关键属性并找出可替代实体。
- 在条件假设下进行推理（Hypotheticals）：
- 一些属性相关的暗指（Meta-reasoning）

不过，MACAW也在某些问题上表现不佳，比如：

- 数学题（Non-trivial Arithmetic）：其实在“5+7=?”这种简单的数学题上，MACAW的表现还是不错的。但数大一些之后就不行了。
- 物体追踪（Entity Tracking and State Changes）：这类问题在之前的bAbI数据集上比较常见。
- 空间推理（Spatial Reasoning）：

总结

这篇文章提出的MACAW，在预训练模型T5的基础上，整合了包括段选取(span selection)、选择题、生成式问答在内的一切问答范式，以及包括问题生成、选项生成、解释生成等在内的多种问答相关任务做联合精调。MACAW在多个问答基准上取得了业界最好的成绩，并只用了GPT-3 十六分之一的参数规模，就在无监督问答数据集 Challenge300 上，相较GPT-3取得了10%的绝对提升，展现了强大的无监督学习的能力。

多数据集多任务整合一直是问答任务的一大研究趋势。2016年SQuAD提出以来，大量的问答数据集涌现，为多数据集整合提供了有力的数据支撑。受到Dual learning的启发，MSRA的段楠老师等人在2017年EMNLP上提出联合问题生成与问答任务[5]，展现出多任务整合有利于问答表现。而2018年提出的BERT，因其适用于多种任务多种形式的包容性，给这一趋势提供了无限可能。近期的工作包括：MultiQA (ACL 2019) [6]整合6种大规模段选取（span selection）任务，并探讨了对小规模任务的迁移能力；UnifiedQA (EMNLP Findings 2020)，整合了多种不同的问答任务形式；以及这篇工作，进一步整合了问答任务及问题生成、选项生成、解释生成等更多的问答相关任务形式。问答数据集本身可能才是限制问答任务难以走向实际应用的一大难题，毕竟在业务场景中用户的问题千奇百怪，很难有足够的同分布数据用以训练。借助大规模预训练语言模型强大的通用性，与prompt方法的灵活性，更多数据、知识与任务形式可以整合在一起，有助于打破问答任务应用的壁垒。

[1] Khashabi, Daniel, et al. "UnifiedQA: Crossing Format Boundaries With a Single QA System." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020.

[2] Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning Research 21.140 (2020): 1-67.

[3] Jansen, Peter, et al. "WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[4] 数据集leaderboard：https://leaderboard.allenai.org/arc/submissions/public，https://leaderboard.allenai.org/arceasy/submissions/public，https://leaderboard.allenai.org/genie-arcda/submissions/publicarcda/submissions/public

[5] Duan, Nan, et al. "Question generation for question answering." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[6] Talmor, Alon, and Jonathan Berant. "MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.