首发于**[诗品算法](https://www.zhihu.com/column/c_1282268278564810752)**



清华ChatGLM底层原理详解——ChatGPT国内最强开源平替，单卡可运行

[](https://www.zhihu.com/people/yang-ru-qi-3)

[**琦琦**](https://www.zhihu.com/people/yang-ru-qi-3)

[**​**](https://www.zhihu.com/question/48510028)

浙江大学 工学硕士

​关注

128 人赞同了该文章

​

目录

收起

引言

必备知识

介绍

一、背景

二、GLM预训练框架

2.1 自回归空格填充任务

2.2 多目标预训练

2.3 模型结构

2.4 GLM微调

2.5 各主流模型对比

三、实验

reference

欢迎关注公众号: 『诗品算法』，禁止一切未经本人@琦琦许可的转载。转载必须注明出处。

引言

之前笔者已经跟大家详细解析过OpenAI的GPT1～GPT3、InstructGPT、ChatGPT，Anthropic的Claude。随着算力的不断发展，模型容量也越来越大，但这些模型均未开源，走向了Close AI之路。不过即使开源，个体也很难玩转这些模型，计算资源（显卡）、数据集等都是困难。

在这样的背景下，国内外涌现出了一批开源模型，近期影响较大的有：Meta AI的LLama、斯坦福基于LLama的Alpaca、清华大学的GLM和ChatGLM等。笔者最近将对这些模型的细节和论文进行详细解析。

本文将对GLM结构进行深度剖析和底层原理解读。

必备知识

在阅读本系列文章之前，建议补充一些相关知识。若你之前未了解过GPT相关原理，可以参考以下的链接：

GPT全家桶系列——了解GPT的前世今生

[琦琦：[万字论文详解]媲美ChatGPT——Google支持的Anthropic推出”更理性的Cl“](https://zhuanlan.zhihu.com/p/629225773)

[琦琦：揭开大型语言模型ChatGPT的神秘面纱](https://zhuanlan.zhihu.com/p/627509839)

[琦琦：InstructGPT论文精读——ChatGPT前身，从人类反馈中学习](https://zhuanlan.zhihu.com/p/627312306)

[琦琦：[万字长文]ChatGPT系列论文精读——大模型经典论文GPT1、GPT2、GPT3](https://zhuanlan.zhihu.com/p/626494749)

[琦琦：一文读懂GPT家族和BERT的底层区别——自回归和自编码语言模型详解](https://zhuanlan.zhihu.com/p/625714067)

介绍

Github：[https://github.com/THUDM/ChatGLM-6B](https://link.zhihu.com/?target=https://github.com/THUDM/ChatGLM-6B" \t "https://zhuanlan.zhihu.com/p/_blank)

模型文件：[https://huggingface.co/THUDM/chatglm-6b](https://link.zhihu.com/?target=https://huggingface.co/THUDM/chatglm-6b" \t "https://zhuanlan.zhihu.com/p/_blank)

博客：[https://chatglm.cn/blog](https://link.zhihu.com/?target=https://chatglm.cn/blog" \t "https://zhuanlan.zhihu.com/p/_blank)

论文：[https://arxiv.org/pdf/2103.10360.pdf](https://link.zhihu.com/?target=https://arxiv.org/pdf/2103.10360.pdf" \t "https://zhuanlan.zhihu.com/p/_blank)

ChatGLM-6B 是一个开源的、支持中英双语的对话语言模型，基于 General Language Model (GLM) 架构，具有 62 亿参数。ChatGLM-6B 使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答。

为了方便下游开发者针对自己的应用场景定制模型，GLM同时实现了基于 P-Tuning v2 的高效参数微调方法，INT4 量化级别下最低只需 7GB 显存即可启动微调。下面进入正题。

一、背景

前文已经明确阐述了时下主流的预训练框架及其区别。主要有三种：

1、autoregressive自回归模型（AR模型）：代表作GPT。本质上是一个left-to-right的语言模型。通常用于生成式任务，在长文本生成方面取得了巨大的成功，比如自然语言生成（NLG）领域的任务：摘要、翻译或抽象问答。当扩展到十亿级别参数时，表现出了少样本学习能力。缺点是单向注意力机制，在NLU任务中，无法完全捕捉上下文的依赖关系。

2、autoencoding自编码模型（AE模型）：代表作BERT。是通过某个降噪目标（比如MLM）训练的双向文本编码器。编码器会产出适用于NLU任务的上下文表示，但无法直接用于文本生成。

3、encoder-decoder（Seq2seq模型）：代表作T5。采用双向注意力机制，通常用于条件生成任务，比如文本摘要、机器翻译等。

三种预训练框架各有利弊，没有一种框架在以下三种领域的表现最佳：自然语言理解（NLU）、无条件生成以及条件生成。T5曾经尝试使用MTL的方式统一上述框架，然而自编码和自回归目标天然存在差异，简单的融合自然无法继承各个框架的优点。

在这个天下三分的僵持局面下，GLM诞生了。

GLM模型基于autoregressive blank infilling方法，结合了上述三种预训练模型的思想。

二、GLM预训练框架

GLM有什么特点？又是如何将其他框架的优势巧妙融合的呢？

1、自编码思想：在输入文本中，随机删除连续的tokens。

2、自回归思想：顺序重建连续tokens。在使用自回归方式预测缺失tokens时，模型既可以访问corrupted文本，又可以访问之前已经被预测的spans。

3、span shuffling + 二维位置编码技术。

4、通过改变缺失spans的数量和长度，自回归空格填充目标可以为条件生成以及无条件生成任务预训练语言模型。

2.1 自回归空格填充任务

给定一个输入文本 x=[x\_1,...x\_n] ，可以采样得到多个文本spans \{s\_1,...s\_m\} 。为了充分捕捉各spans之间的相互依赖关系，可以对spans的顺序进行随机排列，得到所有可能的排列集合 Z\_m ，其中： S\_{z<i}=[s\_{z\_1}, ..., s\_{z{\_{i-1}}}] 。所以预训练目标很清晰：

GLM自回归空格填充任务的技术细节：

1、输入x可以被分成两部分：Part A是被损坏的文本 x\_{corrupt} ，Part B由masked spans组成。假设原始输入文本是[x1, x2, x3, x4, x5, x6]，采样的两个文本片段是[x3]以及[x5, x6]。那么mask后的文本序列是：x1, x2, [M], x4, [M]，即Part A；同时我们需要对Part B的片段进行shuffle。每个片段使用[S]填充在开头作为输入，使用[E]填充在末尾作为输出。

2、二维位置编码：Transformer使用位置编码来标记tokens中的绝对和相对位置。在GLM中，使用二维位置编码，第一个位置id用来标记Part A中的位置，第二个位置id用来表示跨度内部的相对位置。这两个位置id会通过embedding表被投影为两个向量，最终都会被加入到输入token的embedding表达中。

3、观察GLM中自定义attention mask的设计，非常巧妙：

（1）Part A中的tokens彼此可见，但是不可见B中的任意tokens。

（2）Part B tokens可见Part A。

（3）Part B tokens可见B中过去的tokens，不可见B中未来的tokens。

4、采样方式：文本片段的采样遵循泊松分布，重复采样，直到原始tokens中有15%被mask。

5、总结：模型可以自动学习双向encoder（Part A）以及单向decoder（Part B）。

2.2 多目标预训练

上述方法适合于NLU任务。作者希望可以训练一个既可以解决NLU任务，又具备文本生成能力的模型。因此除了空格填充目标之外，还需要增加一个生成长文本目标的任务。具体包含以下两个目标：

1、文档级别。从文档中采样一个文本片段进行mask，且片段长度为文档长度的50%～100%。这个目标用于长文本生成。

2、句子级别。限制被mask的片段必须是完整句子。多个片段需覆盖原始tokens的15%。这个目标是用于预测完整句子或者段落的seq2seq任务。

2.3 模型结构

GLM在原始single Transformer的基础上进行了一些修改：

1）重组了LN和残差连接的顺序；

2）使用单个线性层对输出token进行预测；

3）激活函数从ReLU换成了GeLUS。

但我觉得这部分的修改比较简单常见。核心和亮点还是空格填充任务的设计。

2.4 GLM微调

对于下游NLU任务来说，通常会将预训练模型产出的序列或tokens表达作为输入，使用线性分类器预测label。所以预训练与微调之间存在天然不一致。

作者按照PET的方式，将下游NLU任务重新表述为空白填充的生成任务。具体来说，比如给定一个已标注样本(x, y)，将输入的文本x转换成一个包含mask token的完形填空问题。比如，情感分类任务可以表述为："{SENTENCE}. It’s really [MASK]"。输出label y也同样会被映射到完形填空的答案中。“positive” 和 “negative” 对应的标签就是“good” 和 “bad。

其实，预训练时，对较长的文本片段进行mask，以确保GLM的文本生成能力。但是在微调的时候，相当于将NLU任务也转换成了生成任务，这样其实是为了适应预训练的目标。但难免有一些牵强。

2.5 各主流模型对比

|  |  |  |  |
| --- | --- | --- | --- |
| BERT | XLNet | T5 | UniLM |
| 1、无法捕捉mask tokens的相互依赖性。2、不能准确填充多个连续的tokens。为了推断长度为l的答案概率，BERT需要执行l次连续预测。 | 与GLM相同，使用自回归目标预训练。1、使用文本mask之前的原始位置编码，推理过程中，需要事先知晓或枚举答案长度，与BERT的问题相同。2、双流自注意力机制，使预训练时间成本增加了一倍。 | 使用类似的空格填充目标预训练encoder-decoder Transformer。在编码和解码阶段使用独立的位置编码，使用多个哨兵token来区分mask片段。而在下游任务中，仅使用一个哨兵token，造成模型能力的浪费以及预训练-微调的不一致。 | 通过改变双向、单向以及交叉注意力之间的注意力mask，统一不同的预训练目标。1、总是使用[mask] token替代mask片段，限制了它对mask片段及其上下文的依赖关系进行建模的能力。2、在下游任务微调时，自编码比自回归更加低效。 |

三、实验

预训练数据集：为了与BERT公平对比，使用与BERT相同的数据集训练——BooksCorpus和English Wikipedia。

1、GLM-Base和GLM-Large在相同数据集下的表现优于BERT-Base和BERT-Large。

2、评估GLM多目标任务的表现。短片段采样的GLM-Doc和长片段采样的GLM-Sent，评估其在NLU、seq2seq、空格填充以及零样本语言建模任务上的能力。

GLM-Doc和GLM-Sent仅使用一个预训练目标，因此在NLU上的表现均不如GLM-Large，但仍然优于BERT-Large和UniLM-Large。

3、条件生成任务（Seq2seq）。值得注意的是，GLM-Sent比GLM-Large的表现好，GLM-Doc的表现略差于GLM-Large。这表明教模型拓展文本的文档级别的目标，对条件生成任务的帮助不大，旨在从context中抽取有效信息。

4、消融实验。

GLM的本身训练范式（完形填空式微调）对于模型结果的提升最有效；shuffle spans也对性能有明显提升。

reference

[1] [https://arxiv.org/pdf/2103.10360.pdf](https://link.zhihu.com/?target=https://arxiv.org/pdf/2103.10360.pdf)

编辑于 2023-05-22 18:03・IP 属地浙江