进阶题

◌ 在完成了先前的所有题目之后，相信你对机器学习（深度学习）有了一定的理解，下面我们将在进阶题中对能力进行更深一步的考察

1.论文阅读

难度

> ：高

1.1 Attention is all you need!

大语言模型（large language model，LLM）是一种语言模型，由具有许多参数（通常数十亿个权重或更多）的人工神经网络组成，使用自监督学习或半监督学习对大量未标记文本进行训练。大型语言模型在2018年左右出现，并在各种任务中表现出色。目前，不仅仅在自然语言处理（NLP）领域，大语言模型在诸如计算机视觉（CV）和图深度学习（Graph）中都表现出了不俗的水平与潜力。

几乎无一例外，现有的大模型都是以 Transformer 模型及其变种作为基本组件搭建的。请阅读论文《Attention Is All You Need》https://arxiv.org/pdf/1706.03762，查阅相关资料，完成相关的学习笔记：

值得注意的是，本篇文章行文简洁，含义深刻，对大家不是非常友好，可查阅有关资料协助理解。但**不可不读原文。**

需要提交部分

（若有必要请结合公式说明）

1. Transformer模型的特点是什么？主要是为了解决什么问题？缺点是什么？（找出原文位置）
2. Layer Norm和Batch Norm有什么区别
3. 什么是位置编码（Positional Encoding），请详细解释。在 Transformer 模型中，使用了怎样的位置编码，该位置编码的具体作用与优势在哪里？
4. Transformer 模型中可以说是由 Encoder 与 Decoder 两个模块构成的，那么，这两个模块的大体结构又是怎样的，请花时间搞懂细节的意义和数据的传递方式（图片+文字描述，面试可能出现）？
5. 注意力机制是什么（回答包括：K,Q,V分别代表什么，是如何变化的；在Scaled Dot-Product Attention的公式中注意Q*K^T的内涵；Scaled Dot-Product Attention和 additive attention区别）
6. 在 Transformer 模型中，自注意力机制（Self-Attention）中的“自”体现在什么地方？
7. 注意力机制中有需要学习的参数吗？多头注意力机制中呢？如果有的话，我们希望该参数学习到什么呢？
8. 考虑一个情况：在进行自然语言处理任务时，对于输入的文本，Transformer 是如何进行处理的，数据的具体维度变化又是怎样的？
9. 什么是 BLEU ，在自然语言处理任务中它扮演了怎样的角色？

拓展

1. 对于某些难以理解的部分，有时可以尝试阅读一下模型的相关源代码，也许可以获得一些启发。
2. 对于学有余力的同学（在完成了下一题的比赛之后），不妨利用 Transformer 进行一个简单的训练任务进行体验（对于电脑没有GPU的同学，十分不建议利用本机环境这样做）

1.2 BERT

随着 Transformer 模型获得的巨大成功，自然语言处理领域涌现出大量的采用 Transformer 模型及其相关架构的工作，BERT 便是其中的一员。在有了 Transformer 模型的相关基础后，这里请大家阅读《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》https://arxiv.org/pdf/1810.04805，并对下述问题做出回答。

作为LLM芝麻街系列的一篇著名文章，作者通过融合前人的成果，取得了很惊艳的效果，这不失为一种**值得我们学习的做法。**

需要提交部分

（若有必要请结合公式说明）

1. 大致归纳一下论文分为几个部分，以及每个部分主要解决的问题是什么。
2. BERT 的input embedding 采用了什么方法？说说你认为这么做的好处。
3. 附录A.1中有这一个例子：

其中为什么是"flight ##less"而不直接写成"flightless"?

4. BERT 模型的基本架构是怎样的？它与 Transformer 模型的关系如何？
5. 关于预训练（Pre-train）和微调（Fine-tune），其含义是什么？在 BERT 中，是如何利用的？
6. GLUE 分数是什么，它评估了模型怎样的能力？
7. （自行搜集资料）BERT 衍生出了大批的以 BERT 为基础的衍生模型，举出例子简要分析这些模型发展的趋势。

拓展

1. 有兴趣的同学可以详细了解一下论文中有关实验设计、模型评估与消融实验（Ablation Study）相关的内容。在这里，作者是怎样证明自己的方法的有效性的？
2. 目前由 OpenAI 团队开发的 GPT 系列产品已经成为了高性能大模型的代名词，这里推荐大家阅读《Improving Language Understanding by Generative Pre-Training》https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf，该文章提出了 GPT 系列模型的开山鼻祖。它在思路上与 BERT 有何区别，又是怎样实现功能的呢？
3. 说说读完几篇文章之后你的感想

注意

1. **不允许**直接对他人的总结内容进行照搬，这里希望大家有自己的思考与理解
2. （加分）公式使用 LaTeX 格式

提交方式

将题目中要求的提交的总结内容利用 Markdowm 格式进行编辑，并存为 PDF 文件，提交至邮箱：gimmerml@163.com

文件名要求：姓名-学号-进阶题.pdf

出题人

Jason（学长）

QQ: 2725411278

皇家饼干（学长）